CLUSTERING IN TRANSFORMERS

Tristan Peat
Georgia Institute of Technology
Atlanta, GA 30313, USA
tpeat@gatech.edu

ABSTRACT

Transformers have revolutionized deep learning, yet the mechanisms behind their exceptional performance remain poorly understood. In this work, we propose and validate the hypothesis that Transformers function as clustering machines, organizing input tokens into distinct clusters through their attention mechanisms. Through empirical analysis of pretrained ALBERT models, we demonstrate the emergence of token clusters across network layers and reveal correlations between attention heads that suggest structured information processing. To explain these phenomena, we model both self-attention and causal-attention dynamics using the framework of Wasserstein gradient flows within an interacting particle system. Our simulations reveal how eigenvalue structure and attention temperature control cluster formation, while also providing a mathematical explanation for the emergence of attention sinks - tokens that consistently receive high attention weights regardless of semantic content. These findings suggest that clustering behavior is not merely a byproduct of training but an inherent feature of the Transformer architecture's mathematical structure.

1 INTRODUCTION

Transformers have transformed deep learning by introducing a powerful mechanism for processing sequential data and capturing relationships across long distances (Vaswani et al., 2017). When analyzed through optimal transport theory, the self-attention mechanism reveals an unexpected property: it naturally forms clusters of similar tokens. As shown by Geshkovski et al. (2024a), token representations become increasingly similar as they progress through deeper layers of the network, eventually merging into a single cluster in the limit of infinite training time.

While this clustering behavior shares surface similarities with neural collapse (Papyan et al., 2020), it arises from fundamentally different mechanisms. By modeling self-attention as a system of interacting particles, where each token behaves as a particle influenced by its interactions with other tokens, we gain deeper insights into how Transformers process and organize information. This particle system framework has proven particularly useful for analyzing how and why Transformer architectures converge to their final representations.

In this work, we expand on the work of Geshkovski et al. (2024a) by examining the clustering dynamics in causal attention settings and providing empirical evidence for cluster formation in pretrained transformer models. Our key contributions are threefold: (1) we extend the mathematical framework of interacting particle systems to model causal attention, where tokens can only attend to their predecessors; (2) we empirically demonstrate the emergence of clusters through analysis of token similarity distributions and attention head correlations in pretrained ALBERT models; and (3) we conduct detailed simulations that reveal how different eigenvalue structures and temperature parameters influence cluster formation. Our analysis particularly focuses on understanding the role of attention sinks - tokens that consistently receive high attention weights regardless of semantic content - and how they emerge naturally from the mathematical structure of causal attention.

2 RELATED WORK

2.1 THE TRANSFORMER

A key strength of the Transformer (Vaswani et al., 2017) is its ability to capture relationships between inputs by entangling them in the latent space. In the self-attention mechanism, each object x_i has a query $Q(x_i)$, which it uses to check compatibility with the key $K(x_j)$ of another object x_j through their inner product $\langle Q(x_i), K(x_j)\rangle$. A high inner product indicates a strong match, leading x_i to retrieve x_j's value $V(x_j)$, with the representation u_i formed by summing up values of objects weighted by their compatibility with x_i. When used in isolation, self-attention converges to a rank-1 solution in cubic time (Dong et al., 2023), though skip connections and feedforward networks prevent this collapse, and to study asymptotic clustering behavior with increased depth, we work with ALBERT (Lan et al., 2020), which uses weight sharing across layers to enable training of very deep transformers while maintaining parameter efficiency.

2.2 ATTENTION SINKS

Attention sinks are tokens that receive high attention scores regardless of semantic importance, particularly in initial sequence positions Xiao et al. (2024). These sinks play a crucial role in stabilizing model performance during streaming tasks, as confirmed by studies of LLama-7B showing uniform attention distribution with a disproportionate weighting of the first tokens Agarwal et al. (2024). This phenomenon emerges naturally from the softmax operation in attention computation—since weights must sum to 1, initial tokens become attention sinks during training due to their universal visibility to subsequent tokens in autoregressive models.

2.3 EMERGENCE OF CLUSTERS

Recent theoretical work has established that clustering behavior is fundamentally inherent to transformer architectures, emerging through two distinct mechanisms. Geshkovski et al. (2024b) demonstrates that clustering arises through modeling tokens as particles whose trajectories evolve through network layers, revealing distinct clustering patterns determined by the value matrix V. Their analysis shows three types of clustering: tokens cluster toward vertices of convex polytopes when V is the identity matrix, toward at most three parallel hyperplanes when V has a simple positive leading eigenvalue, and toward hybrid structures combining polytope vertices and linear subspaces for paranormal value matrices.

Building on this, Geshkovski et al. (2024a) proves that clustering emerges naturally in transformers when viewed as interacting particle systems with coupled ODEs. They show that tokens with time-independent weights exhibit clustering toward specific limiting objects, with cluster locations determined by the initial tokens. This mathematically confirms empirical observations from **?** about the emergence of "leader" tokens in sequences. Their analysis reveals that clustering behavior changes based on the spectrum of the value matrix V, and in the one-dimensional case, they prove the self-attention matrix converges to a low-rank Boolean matrix.

2.4 MATH OF TRANSFORMERS

From the work of Geshkovski et al. (2024b), we repeat their formulation of Transformers from the perspective of optimal transport and Ordinary Differential Equations (ODEs). The evolution of tokens through the layers can be modeled as a system of n coupled ODEs describing dynamics of a system of particles $x_1(t), ..., x_n(t)$. One head of the standard attention layer is defined as follows. Given an input sequence represented by the token embeddings $X \in \mathbb{R}^{n\times d}$, where n is the number of tokens and d is the dimension of the embedding space, and the matrices W_Q, W_K, W_V to compute queries, keys, and values, the attention mechanism computes a weighted sum of values based on their relevance to a query in the following form

$$\text{Attn}(X) = \text{softmax}\left(\frac{XW_QW_K^TX^T}{\sqrt{d}}\right)XW_V \tag{1}$$

For notational consistency, we denote a sequence of tokens encoded as particles in d-dimensional embedding space d as $(x_1, ..., x_n)$, corresponding to the columns of X^T. Due to positional encoding,

the order of tokens does not matter, allowing us to view the input as a probability measure of tokens:

$$\{x_1, ..., x_n\} \in \mathbb{R}^d \Longleftrightarrow \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

Additionally, for simplicity we denote $V : W_V^T, Q : W_Q^T, ...$ We also introduce an arbitrary temperature parameter β instead of the fixed scaling factor $1/\sqrt{d}$. One term of attention added to the k-th token can be written explicitly as:

$$\text{attn}(x_1, ..., x_n)_k = \frac{1}{Z_k}\sum_{j=1}^{n} e^{\beta\langle Qx_k, Kx_j\rangle} Vx_j$$

where,

$$Z_k = \sum_{j=1}^{n} e^{\beta\langle Qx_k, Kx_j\rangle} \tag{2}$$

Combining these together, the dynamics of token propagation through the layers can be expressed as:

$$\dot{x}_k(t) = \frac{1}{Z_k(t)} P_{x_k(t)}\Big(\sum_{j=1}^{n} e^{\beta\langle Q(t)x_k(t), K(t)x_j(t)\rangle} V(t)x_j(t)\Big)$$

where $Z_k(t)$ is the time-parameterized version of 2 and the projector P_x defined for any vector v as:

$$P_x(v) := v - \frac{\langle x, v\rangle}{\|x\|^2} x$$

This projection ensures that x_k remains on the sphere by projecting the attention output onto the tangent space at $x_k(t)$. For completeness, we have included a realization of the full Transformer in the equation in Section A.1 of the Appendix.

2.5 Causal Attention

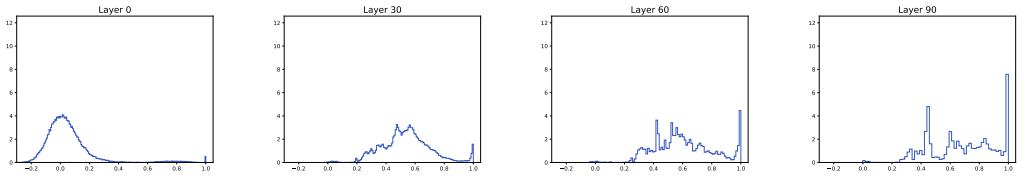
Our focus on causal attention is motivated by its central role in autoregressive language models like ALBERT (Lan et al., 2020), where each token can only attend to previous tokens in the sequence, mirroring the left-to-right nature of text generation. This causality constraint is essential for understanding how these models form meaningful clusters during inference and generation, as each token's representation is built solely from its predecessors. To model this behavior, we borrow from Karagodin et al. (2024) and modify the mean-field dynamics where the k-th token depends on the position of *all tokens* $j \in [n]$, to instead have the dynamics of token k depend only on the position of tokens $j \leq k$:

$$\dot{x}_k(t) = P_{x_k(t)}\left(\frac{1}{Z_k(t)}\sum_{j=1}^{k} e^{\beta\langle Q(t)x_k(t), K(t)x_j(t)\rangle} V(t)x_j(t)\right) \tag{3}$$

The first token is operating fully autonomously from the influence of the other tokens. We can describe its evolution as follows:

$$\dot{x}(t) = P_{x(t)}(Vx(t))$$

Restating the main insight from Karagodin et al. (2024), there are two major forces that drive each token: an internal force determined by the token's own dynamics, and an external force induced by all preceding particles that can be either attractive or repulsive depending on the sign of the top eigenvalue(s) of V. The balance between these forces is mediated by the attention mechanism. Therefore, it is especially useful to analyze the special case where V = Id (eliminating internal forces), where tokens collapse asymptotically to a single point regardless of the choice of Q and K matrices.



(a) Layer 0 (b) Layer 30 (c) Layer 60 (d) Layer 90

Figure 1: Histograms of ALBERT-XL's tokens cosine similarity across different layers. We observe a shift from a normal disquisition to a growing spike at a similarity of 1 (on the x-axis).

3 METHODOLOGY

3.1 EMPIRICAL EVIDENCE OF CLUSTERS

3.1.1 HISTOGRAM OF TOKEN SIMILARITY

We empirically demonstrate the existence of clusters in Transformers by examining the distribution of pairwise cosine similarities between token representations at each layer. By computing histograms of these similarities across layers, we can observe the gradual transformation from an approximately normal distribution to a distinctive spike near similarity=1, providing quantitative evidence for the emergence of token clusters. These distributions reveal how tokens progressively align into tight clusters as they move through the network's layers, with the height and sharpness of the spike indicating the strength of clustering behavior. We conduct this experiment for both a pretrained ALBERT (Lan et al., 2020) and T5 (Raffel et al., 2019) from the Huggingface Transformers library (Wolf et al., 2020).

3.1.2 CORRELATION BETWEEN ATTENTION HEADS

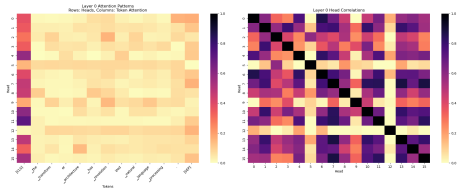
Drawing inspiration from Agarwal et al. (2024), we investigate the relationships between attention heads by examining two key aspects of their operation: the correlation between attention patterns across different heads and their token-specific activation patterns. By computing pairwise correlations between attention distributions of different heads across a large sample of C4 dataset inputs (Raffel et al., 2019), we can identify groups of heads that exhibit similar attention behaviors. Additionally, by visualizing the importance each head assigns to different token positions, we reveal how different heads specialize in attending to specific linguistic patterns or positions in the input sequence.

3.2 SIMULATING ATTENTION DYNAMICS

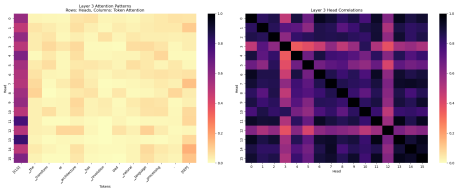
We investigate the geometric properties of attention mechanisms by simulating particle dynamics on the unit sphere. While the dynamics generalize to any dimension, we focus on visualizing particles moving in two or three dimensions. Building from Geshkovski et al. (2024b), we implement a system where each token's state evolves based on attention-weighted interactions with its predecessors from Equation. 3, capturing the causal nature of autoregressive models.

The simulation architecture maintains particles on the sphere through two complementary mechanisms: projection onto the tangent space and explicit renormalization after each integration step. For numerical stability, we use exponential time rescaling and cubic interpolation to ensure smooth trajectory visualization. Our baseline simulations track $n = 64$ particles in $d = 2$ or $d = 3$ dimensions with unit temperature β, integrating for 15 time units with a small step size $dt = 0.1$ to ensure stability.

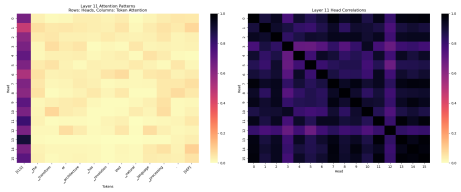
To understand clustering behavior, we investigate several key parameter regimes. We examine how different eigenstructures of the value matrix affect particle interactions by testing the identity matrix (leading to convergence to polytope vertices), matrices with negative eigenvalues (inducing repulsion), and matrices with eigenvalues extracted from pretrained ALBERT attention heads. We also



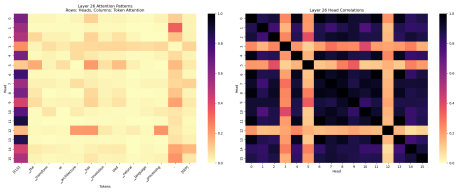
(a) Layer 0



(b) Layer 3



(c) Layer 11



(d) Layer 26

Figure 2: Activations of Multi Head Attention: Left figure shows activation scores for each token for each head in ALBERT. We observe that several heads give similar scores to the sequence. We also observe a strong attention weight given to the first token in the sequence [CLS]. Right figure shows pairwise cross-correlations that generally increase with depth and show the emergence of a show existence of two clusters- Heads [3,5,12] show a strong correlation forming one cluster and the remaining heads form the other

explore temperature variations β, observing that low temperatures produce rapid convergence to a single cluster while higher temperatures allow meta-stable states with multiple clusters to persist.

The simulation framework particularly focuses on causal attention, where each token can only attend to its predecessors in the sequence. This constraint is implemented through careful masking of the attention weights, ensuring that token interactions respect the autoregressive property of language models. Initial conditions are sampled uniformly from the sphere to avoid biasing the dynamics toward particular configurations.

4 Results

4.1 Evidence of Clusters

The histogram analysis in Figure 1 reveals a clear progression of token similarity distributions across ALBERT's layers, transitioning from an initially broad, normal distribution at layer 0 to an increasingly concentrated spike near similarity=1 by layer 90. This evolution provides quantitative evidence for the gradual alignment of token representations into tight clusters as information propagates through the network. Complementing these findings, Figure 2 demonstrates the emergence of structured patterns in attention head behavior, where we observe both the formation of correlated head clusters ([3,5,12] vs others) and the consistent prominence of the [CLS] token across layers. The increasing cross-correlation between attention heads at deeper layers, as shown in the right panels of Figure 2, suggests that the network develops specialized groups of attention heads that work in concert to process different aspects of the input.

4.2 Simulation Attention Dynamics

Our simulation experiments reveal several key insights into the clustering behavior of attention mechanisms. We begin by examining the baseline case of convergence to a single cluster under standard conditions, then explore how this behavior changes with varying temperature and eigenvalue structures.

4.2.1 SINGLE CLUSTER CONVERGENCE

We first investigate the baseline case of convergence under standard conditions where $V = I_d$ and $Q^T K = I_d$ with temperature $\beta = 1.0$. Figure 3 shows the temporal evolution of particles in 3D space, demonstrating how the system naturally converges to a single cluster over time. The sequence of snapshots at $t = 3.0$, $t = 6.0$, and $t = 12.0$ reveals the gradual consolidation of initially dispersed particles into a unified cluster.

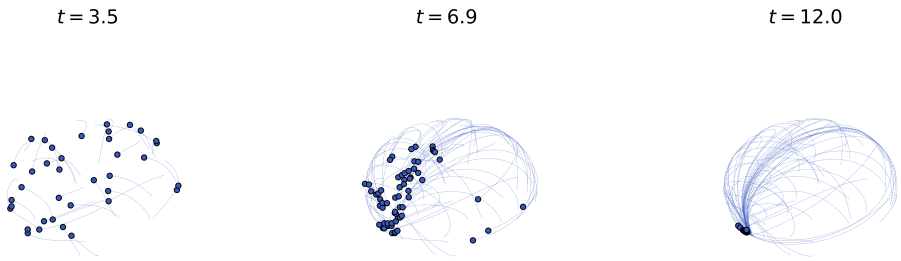


Figure 3: Temporal evolution of particle clustering in 3D space with random Q, K and $V = I_d$ showing convergence to a single cluster when $\beta = 1.0$. The sequence shows snapshots at $t = 3.0$, $t = 6.0$, and $t = 12.0$.

4.2.2 TEMPERATURE-DEPENDENT BEHAVIOR

To understand the role of temperature in cluster formation, we increase the temperature parameter to $\beta \geq \sqrt{d}$, which in our case is $\beta = 9.0 \geq \sqrt{64}$. As shown in Figure 4, this higher temperature leads to dramatically different dynamics, where distinct clusters persist in meta-stable states rather than collapsing to a single point. This behavior suggests that temperature acts as a control parameter for the granularity of cluster formation.

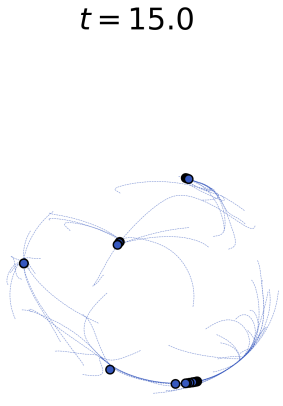
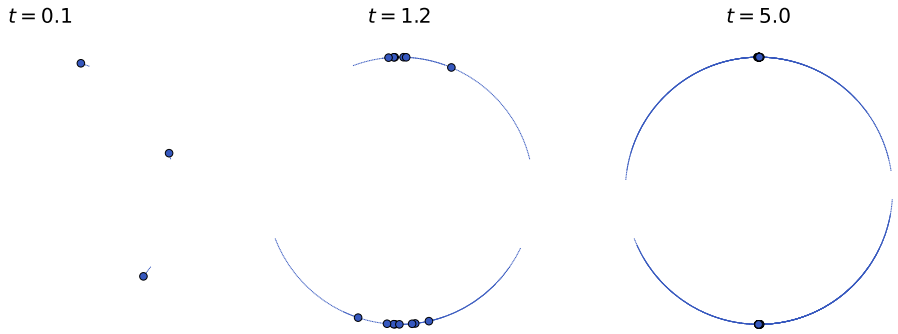


Figure 4: Meta-stable state formation at elevated temperature ($\beta = 9.0$). The higher temperature allows distinct clusters to persist without collapsing to a single point.

4.2.3 REPULSIVE EFFECTS

We next examine how repulsive forces shape cluster formation by studying particle dynamics on a unit circle ($d = 2$) with a negative maximum eigenvalue in the value matrix. Figure 5 tracks this evolution across three time points ($t = 0.1$, $t = 1.2$, and $t = 5.0$), demonstrating how particles organize themselves to maximize separation when subject to repulsive forces. The final configuration shows clusters positioned at opposite sides of the circle, illustrating the system's tendency to maximize inter-cluster distances under repulsive conditions.



(a) $t = 0.1$ (b) $t = 1.2$ (c) $t = 5.0$

Figure 5: Evolution of particle clustering on the unit circle ($d = 2$) with $\beta = 1.0$ and $V_{\lambda_{\max}} = -1$. The repulsion factor drives clusters to maximize their separation across the circle, demonstrating how negative eigenvalues can induce structured separation patterns.

When we extend this analysis to 3D space and increase the temperature to $\beta = 9.0$, the dynamics become notably more complex. Figure 6 shows how the combination of higher dimensionality and temperature leads to richer clustering patterns, where the clusters are separating and becoming more dense, but swirling the point, a geodesic Geshkovski et al. (2024a) called a torus.

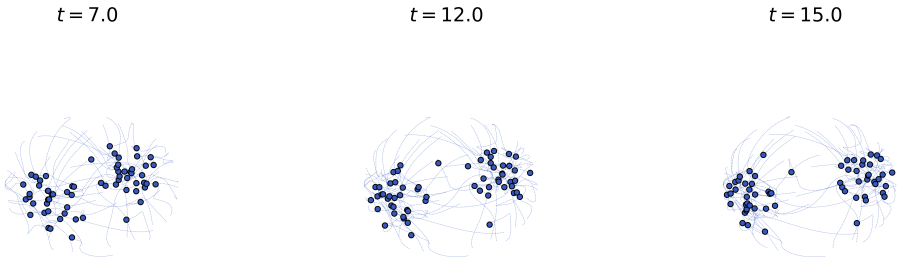


Figure 6: Temporal evolution of particle distributions under negative eigenvalues, showing the formation of structured multi-cluster arrangements.

4.2.4 ALBERT-INSPIRED DYNAMICS

To bridge our theoretical analysis with practical architectures, we simulate particle dynamics using eigenvalues extracted from ALBERT's attention heads. Figure 7 shows the system's evolution under these conditions, where $\beta = 1.0$. The rapid convergence to a meta-stable state by $t = 6.0$, which persists through $t = 15.0$, suggests a natural separation between the initial token and subsequent tokens in the sequence—a finding that aligns with observed behavior in transformer architectures.

To further investigate the unique role of the initial token in causal attention, we track its trajectory by highlighting it in red. Figure 8 follows this token from $t = 0.6$ to $t = 9.0$, revealing its dual

nature: simultaneously attracting subsequent tokens while aligning itself with the eigenvectors of V. This visualization provides direct evidence of how the leading token balances its roles as both an attractor for other tokens and a participant in the broader dynamics governed by the value matrix.

5 DISCUSSION

Our formulation of causal-attention in equation 3 offers insight into why attention sinks emerge naturally in transformer architectures. Note that the first token $x_1(t)$ evolves autonomously, while each subsequent token $x_k(t)$ involves a weighted sum over all previous tokens $(x_1(t), ..., x_k(t))$. The exponential weighting factor $e^{\beta\langle Q(t)x_k(t), K(t)x_j(t)\rangle}$ in equation 3 means that even small alignments between query and key vectors can result in significant attention weights after normalization by $Z_k(t)$. Since the first token is involved in the evolution of every subsequent token's representation, it has more opportunities to develop strong query-key alignments during training. This mathematical structure provides a potential explanation for the emergence of attention sinks - the first tokens naturally become universal attractors in the attention mechanism simply due to their privileged position in the causal structure, rather than any inherent semantic importance. This aligns with the empirical observations of Xiao et al. (2024) and Agarwal et al. (2024) regarding the disproportionate attention weights given to initial sequence tokens.

Future work will explore three key directions: incorporating dynamically optimized MLPs to better reflect real-world transformer behavior, investigating the semantic interpretation of eigenspaces that

Figure 7: Evolution of particle clusters using ALBERT's top-3 eigenvalues with $\beta = 1.0$. The system rapidly converges to a meta-stable state by $t = 6.0$ and maintains this configuration through $t = 15.0$. The configuration suggests segregation between the initial token and subsequent tokens in the sequence.

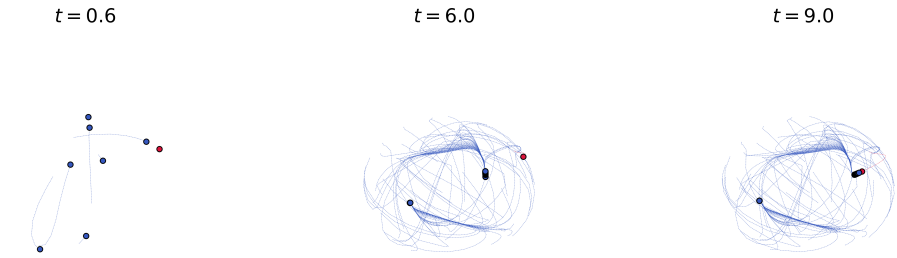


Figure 8: Visualization of the initial token's trajectory (shown in red) from $t = 0.6$ to $t = 9.0$. The leading token exhibits dual behavior: attracting subsequent tokens while simultaneously aligning with the eigenvectors of V, demonstrating the interplay between attention mechanisms and the underlying eigenstructure.

govern token dynamics, and extending our analysis to multiple interacting particle systems as an analog for multi-head attention. These extensions will provide deeper insights into how Transformers organize and process information across multiple attention heads while maintaining mathematical rigor in our particle system framework.

6 CONCLUSION

This work demonstrates that Transformer architectures naturally exhibit clustering behavior through both empirical evidence in pretrained models and theoretical analysis of attention dynamics. Our simulation results reveal that causal attention leads to the emergence of meta-stable clusters influenced by the eigenstructure of value matrices, with the first token playing a crucial role as an attention sink. These findings suggest that the clustering behavior in Transformers is not merely a byproduct of training but rather an inherent feature of the architecture's mathematical structure, particularly in how causal attention shapes token interactions across layers.

REFERENCES

Saurabh Agarwal, Bilge Acun, Basil Hosmer, Mostafa Elhoushi, Yejin Lee, Shivaram Venkataraman, Dimitris Papailiopoulos, and Carole-Jean Wu. Chai: Clustered head attention for efficient llm inference, 2024. URL https://arxiv.org/abs/2403.08058.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023. URL https://arxiv.org/abs/2103.03404.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics, 2024a. URL https://arxiv.org/abs/2305.05465.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, 2024b. URL https://arxiv.org/abs/2312.10794.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking, 2024. URL https://arxiv.org/abs/2411.04990.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020. URL https://arxiv.org/abs/1909.11942.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, September 2020. ISSN 1091-6490. doi: 10.1073/pnas.2015509117. URL http://dx.doi.org/10.1073/pnas.2015509117.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

A MATHEMATICAL GROUNDS

A.1 FORMAL DEFINITION OF TRANSFORMER BLOCK

A transformer block is a parameterized function class $f_\theta : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$. If $x \in \mathbb{R}^{n\times d}$, then $f_\theta(x) = z$, where

$$Q(x_i) = W_Q^T x_i, \quad K(x_i) = W_K^T x_i, \quad V(x_i) = W_V^T x_i, \quad W_Q, W_K, W_V \in \mathbb{R}^{d\times k}$$

$$\alpha_{i,j} = \text{softmax}_j\left(\beta\langle Q(x_i), K(x_j)\rangle\right)$$

$$u_i' = \sum_{h=1}^{H} W_{c,h}^T \sum_{j=1}^{n} \alpha_{i,j} V(x_j), \quad W_{c,h} \in \mathbb{R}^{k\times d}$$

$$u_i = \text{LayerNorm}(x_i + u_i'; \gamma_1, \beta_1), \quad \gamma_1, \beta_1 \in \mathbb{R}^d$$

$$z_i' = W_2^T \text{ReLU}(W_1^T u_i), \quad W_1 \in \mathbb{R}^{d\times m}, \quad W_2 \in \mathbb{R}^{m\times d}$$

$$z_i = \text{LayerNorm}(u_i + z_i'; \gamma_2, \beta_2), \quad \gamma_2, \beta_2 \in \mathbb{R}^d$$