Group 2506
Corte Riccardo
Miotto Alessandro
Rizzi Lorenzo

University of Padua
Physics of Data
Report 2: 24-06-25
Academic Year: 2024-25

Geometry of the attention mechanism

In the previous report, we described the attention mechanism used in GPT-2. Here, we revisit and refine that discussion, focusing on a more compact formulation of the attention scores as a bilinear form.

For a given head h, the attention score matrix $A^{(h)} \in \mathbb{R}^{N\times N}$ is computed using the token buffer $X \in \mathbb{R}^{N\times 768}$, along with the query and key weight matrices $W_Q^{(h)}, W_K^{(h)} \in \mathbb{R}^{768\times 64}$, and their corresponding bias vectors $\boldsymbol{b}_Q^{(h)}, \boldsymbol{b}_K^{(h)} \in \mathbb{R}^{64}$. Each element of the attention score matrix is given by the scaled dot product between the query and key vectors:

$$A_{ij}^{(h)}(\boldsymbol{x}_i, \boldsymbol{x}_j) \propto \boldsymbol{q}_i^{(h)} \cdot \boldsymbol{k}_j^{(h)}$$

where $\boldsymbol{q}_i^{(h)}$, $\boldsymbol{k}_j^{(h)} \in \mathbb{R}^{64}$ are the query and key vectors corresponding to tokens $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively. These are computed as follows:

$$\underbrace{Q^{(h)}}_{N\times 64} = \underbrace{X}_{N\times 768}\underbrace{W_Q^{(h)}}_{768\times 64} + \underbrace{\boldsymbol{1}}_{N\times 1}\underbrace{\boldsymbol{b}_Q^{(h)}}_{1\times 64} \quad \longrightarrow \quad \boldsymbol{q}_i^{(h)} = \boldsymbol{x}_i W_Q^{(h)} + \boldsymbol{b}_Q^{(h)}$$

$$\underbrace{K^{(h)}}_{N\times 64} = \underbrace{X}_{N\times 768}\underbrace{W_K^{(h)}}_{768\times 64} + \underbrace{\boldsymbol{1}}_{N\times 1}\underbrace{\boldsymbol{b}_K^{(h)}}_{1\times 64} \quad \longrightarrow \quad \boldsymbol{k}_i^{(h)} = \boldsymbol{x}_i W_K^{(h)} + \boldsymbol{b}_K^{(h)}$$

Reformulating as a Bilinear Form

We now aim to express the attention score A_{ij} of a specific head h as a bilinear form. This reformulation offers a more concise representation that also integrates the biases into the matrix structure. We define the augmented input vectors $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \in \mathbb{R}^{769}$ by and the block matrix $J \in \mathbb{R}^{769\times 769}$ as:

$$J^{(h)} := \begin{pmatrix} W_Q {W_K}^T & W_Q \boldsymbol{b}_K \\ \boldsymbol{b}_Q W_K^T & \boldsymbol{b}_Q^T \boldsymbol{b}_K \end{pmatrix}, \qquad \tilde{\boldsymbol{x}} := \begin{pmatrix} \boldsymbol{x}_i \\ 1 \end{pmatrix}, \quad \tilde{\boldsymbol{y}} := \begin{pmatrix} \boldsymbol{x}_j \\ 1 \end{pmatrix}.$$

With these definitions, the attention score can now be written as a simpler bilinear form: $B : \mathbb{R}^{769} \times \mathbb{R}^{769} \to \mathbb{R}$ given by:

$$A_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_i) = \tilde{\boldsymbol{x}}^T J \tilde{\boldsymbol{y}} := B(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$$

However, this bilinear form does not define an inner product, as the matrix J is generally neither symmetric nor positive definite. Consequently, it does not induce a norm. This asymmetry arises from the use of different projections into the query and key spaces, which breaks the symmetry required for inner products. To interpret the bilinear form in a way that resembles an inner product, we can perform a singular value decomposition (SVD) of the matrix J:

$$B(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \tilde{\boldsymbol{x}}^T J \tilde{\boldsymbol{y}} = \tilde{\boldsymbol{x}}^T U \Sigma V^T \tilde{\boldsymbol{y}} = (U^T \tilde{\boldsymbol{x}})^T \Sigma (V^T \tilde{\boldsymbol{y}}) := \boldsymbol{u}^T \Sigma \boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle_\Sigma \tag{1}$$

where $U, V \in \mathbb{R}^{n \times m}$ are orthonormal matrices, $\Sigma \in \mathbb{R}^{m \times m}$ is a diagonal matrix of positive singular values, and $m = \text{rank}(J)$.

This expression can be interpreted as a generalized inner product between two transformed vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, each living in different coordinate systems defined by the linear transformation $\pi_U : U^T \boldsymbol{x} \mapsto \boldsymbol{u}$ and $\pi_U : V^T \boldsymbol{y} \mapsto \boldsymbol{v}$, respectively. In this sense, each attention head maps token pairs from the original embedding space into distinct query and key subspaces, and then performs a dot product weighted by Σ.

Distance between Bilinear Forms

We aim to characterize the bilinear form associated with each attention head using a unified mathematical framework. This will enable us to better analyze and compare the differences between the bilinear forms induced by different attention heads.

Let us begin by examining the orthonormal matrices $U, V \in \mathbb{R}^{n \times m}$. Each matrix consists of m linearly independent vectors in $\mathbb{R}^n$, and thus defines a Grassmann manifold $\text{Gr}_m(\mathbb{R}^n)$, which is the space of all m-dimensional linear subset of $\mathbb{R}^n$. The interaction matrix $\Sigma \in \mathbb{R}^{m \times m}$ is diagonal with strictly positive entries. We can thus define a product manifold that compactly encodes the structure of the bilinear form as[1]:

$$\mathcal{M} = \text{Gr}_m(\mathbb{R}^n) \times \text{Gr}_m(\mathbb{R}^n) \times \mathbb{R}^{m \times m} \tag{2}$$

Each attention head h corresponds to a distinct *attention* manifolds $\mathcal{M}_h = (U_h, V_h, \Sigma_h)$. The bilinear form for head h than takes the form:

$$B^{(h)}(\boldsymbol{x}, \boldsymbol{y}) = \langle U_h^T \boldsymbol{x}, V_h^T \boldsymbol{y} \rangle_{\Sigma_h} \tag{3}$$

To compare the bilinear forms across heads, we can define a distance metric between two attention manifolds $\mathcal{M}_i$ and $\mathcal{M}_j$ as follows:

$$d(\mathcal{M}_i, \mathcal{M}_j)^2 = d(U_i, U_j)^2 + d(V_i, V_j)^2 + ||\Sigma_i - \Sigma_j||^2 \tag{4}$$

where $d(U_i, U_j)$ and $d(V_i, V_j)$ denotes a distance on the Grassmann manifold, specifically the geodesic distance between the subspaces spanned by the different U and V. The geodesic distance between Grassmann manifolds is based on the principal angles$\{\theta_1, ..\theta_r\}$ and is defined as:

$$d(U_i, U_j) = \left(\sum_{l=1}^{r} \theta_l^2 \right)^{1/2} \tag{5}$$

This distance lies in the interval $[0, \sqrt{r}\pi/2]$, where 0 indicates that the subspaces are identical, and $\sqrt{r}\pi/2$ corresponds to fully orthogonal subspaces.

By embedding the bilinear forms into this geometric framework, we gain an interpretable way to compare attention heads in terms of the geometry of the subspaces they operate.

[1] *Disclaimer*: My understanding of topology and differential geometry is limited, so this represents a naive attempt to combine and interpret subspaces in a way that allows the introduction of a distance metric between them.

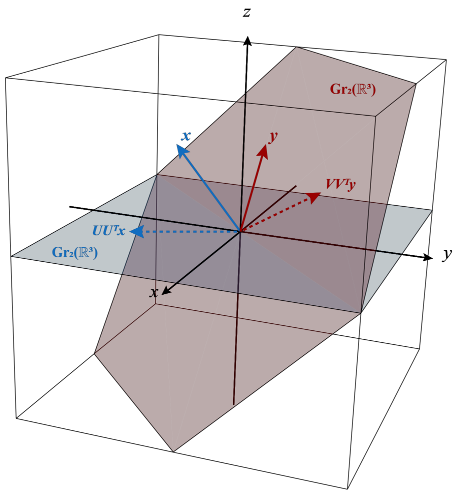


Figure 1: 3D representation of token vectors $\boldsymbol{x}$ (in blue) and $\boldsymbol{y}$ (in red), projected onto two lower-dimensional subspaces spanned by orthonormal bases U and V, respectively. The transformation matrices map the original 3D vectors into a 2D space. For visualization purposes, the projected vectors are re-embedded into the original 3D space. The angle between the planes corresponds to the principal angle between the two subspaces we have mentioned.

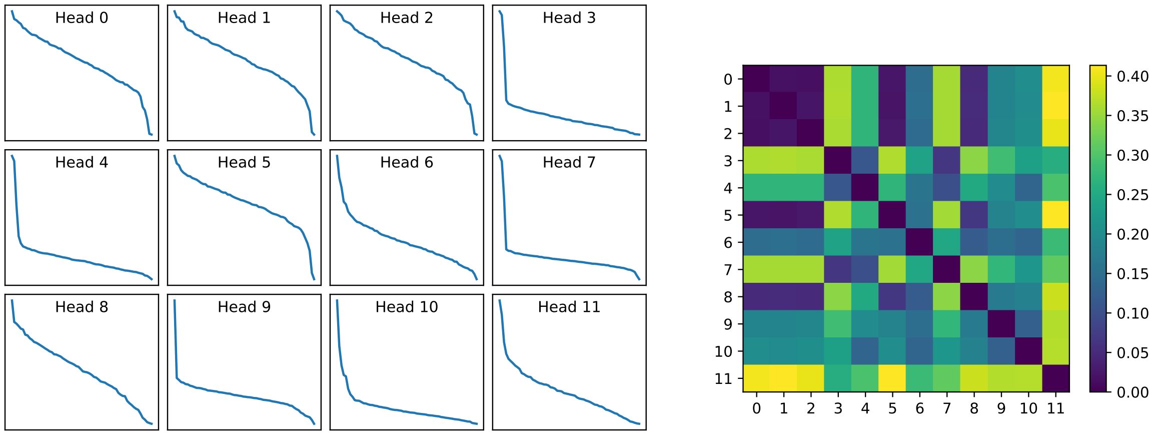
In Figure 1, we illustrate a geometric interpretation of how the attention mechanism operates within each head. Given two tokens $\boldsymbol{x}$ and $\boldsymbol{y}$ in the embedding space $\mathbb{R}^n$, they are transformed into two lower-dimensional subspaces $\mathrm{Gr}_r(\mathbb{R}^n)$ (the query and key subspaces) via linear maps $\pi_U : \boldsymbol{x} \mapsto U^T\boldsymbol{x}$ and $\pi_V : \boldsymbol{y} \mapsto V^T\boldsymbol{y}$. Note that this transformation is not a projection in the strict sense, but rather a submersion from $\mathbb{R}^n$ to $\mathbb{R}^r$.

Once embedded in their respective subspaces, the resulting vectors $U^T\boldsymbol{x}$ and $V^T\boldsymbol{y}$ are combined through an inner product weighted by a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_r)$, which encodes the coupling strength between the query and key. This coupling reflects how strongly the corresponding components interact and contribute to the attention score.

Each attention head h is thus associated with a pair of subspaces $\mathrm{Gr}_r^{(h)}(\mathbb{R}^n) \times \mathrm{Gr}_r^{(h)}(\mathbb{R}^n)$ (for queries and keys), and a distinct coupling matrix $\Sigma^{(h)}$. Geometrically, we can interpret each head as defining a manifold $\mathcal{M}^{(h)}$ formed by the interaction of its query and key subspaces. If two heads attend to similar patterns or features, their associated manifolds will partially overlap. This intuition can be visualized by imagining each head's subspaces as 2D planes in a 3D space: the angles between any two planes $\mathcal{M}_i$ and $\mathcal{M}_j$ quantify their similarity. These are the principal angles, which we use as a metric for comparing attention subspaces.

In addition to subspace orientation, the strength of coupling, encoded in the singular values Σ_i and Σ_j, also plays a crucial role. Therefore, the similarity between attention heads is determined not only by the alignment of their subspaces but also by how their query-key interactions are weighted. In Figure 2 and Figure 3 the normalized singular values for each layer of GPT-2 are shown.

Keys-queries coupling



(a) Singular values $\Sigma^{(h)} = \text{diag}(\sigma_1^{(h)}, ..., \sigma_{64}^{(h)})$ for the attention heads of the first layer of GPT2.

(b) Singular values distance $||\Sigma^{(i)} - \Sigma^{(j)}||$ between the i-th and j-th head.

Figure 2: Singular value distributions for the attention heads in the first layer of GPT-2 (left), and the corresponding pairwise distances based on singular value similarity (right heatmap). Heads with similar singular value profiles—such as heads 1, 2, 3, and 8—exhibit lower distances in the heatmap, indicating similar coupling behavior between their query and key subspaces.

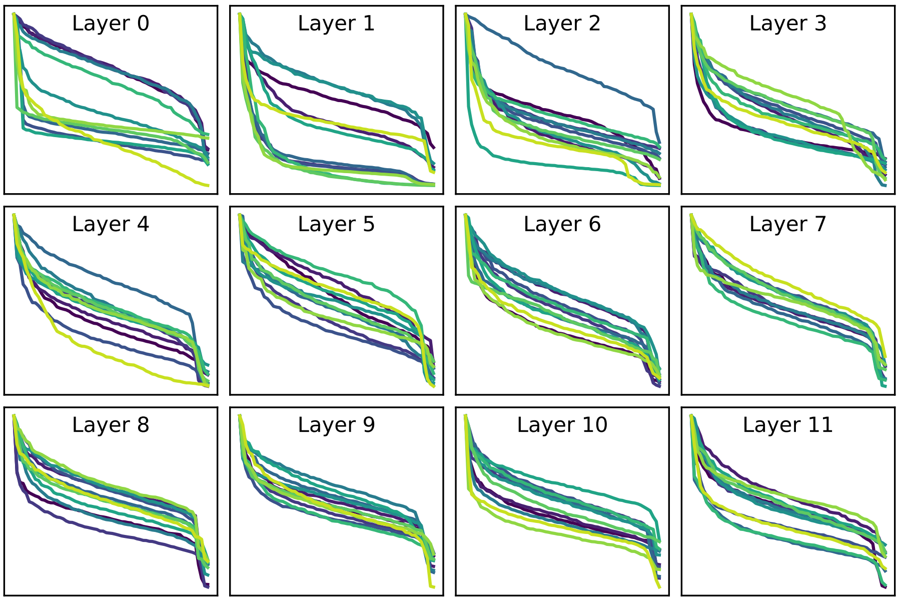


Figure 3: The normalized singular values $\Sigma^{(h,k)} = \text{diag}(\sigma_1^{(h,k)}, ..., \sigma_{64}^{(h,k)})$ for each head h across all layers k of GPT-2. n the first layer, the coupling strengths between keys and queries appear more heterogeneous, suggesting greater variation in how early heads process information. As we move deeper into the transformer, the singular value profiles become more uniform across heads, potentially indicating convergence toward similar query-key interactions. However, interpreting this trend could be misleading, since the attention mechanism depends also on the linear mapping into the key and query subspaces.

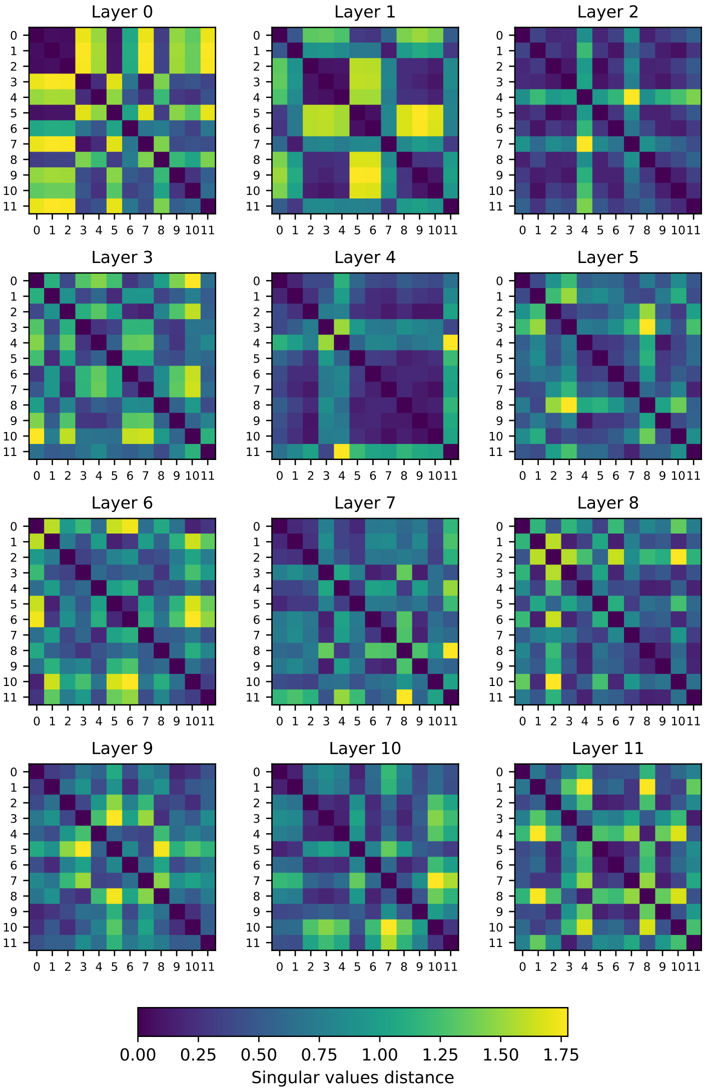


Figure 4: Representation of the coupling between the singular values of each pair of attention heads across all layers of the transformer. This analysis extends the idea shown in Figure 2, but here the pairwise difference is quantified for each layer by computing the norm $||\Sigma_i - \Sigma_j||$, where Σ_i and Σ_j, respectively. Smaller values (blue) indicate similar coupling strengths between the query and key subspaces, while larger values (yellow) reflect divergent head specializations.

Queries U and keys V subspaces

The following heatmap illustrates the similarity between the query and key subspaces across attention heads. Geometrically, we can interpret each layer as decomposing the full embedding space into a set of lower-dimensional subspaces, defined by linear maps π_{U_i} and π_{U_j}. These subspaces, shaped during training, are expected to capture distinct semantic or syntactic features that each head attends to. If two heads span similar subspaces, this suggests they attend to similar linguistic features.

To quantify these relationships, we compute the Grassmann distance (Equation 5) between the row spaces of the query and key projections. The results are visualized in Figure 6 and the distribution is shown in Figure 5.

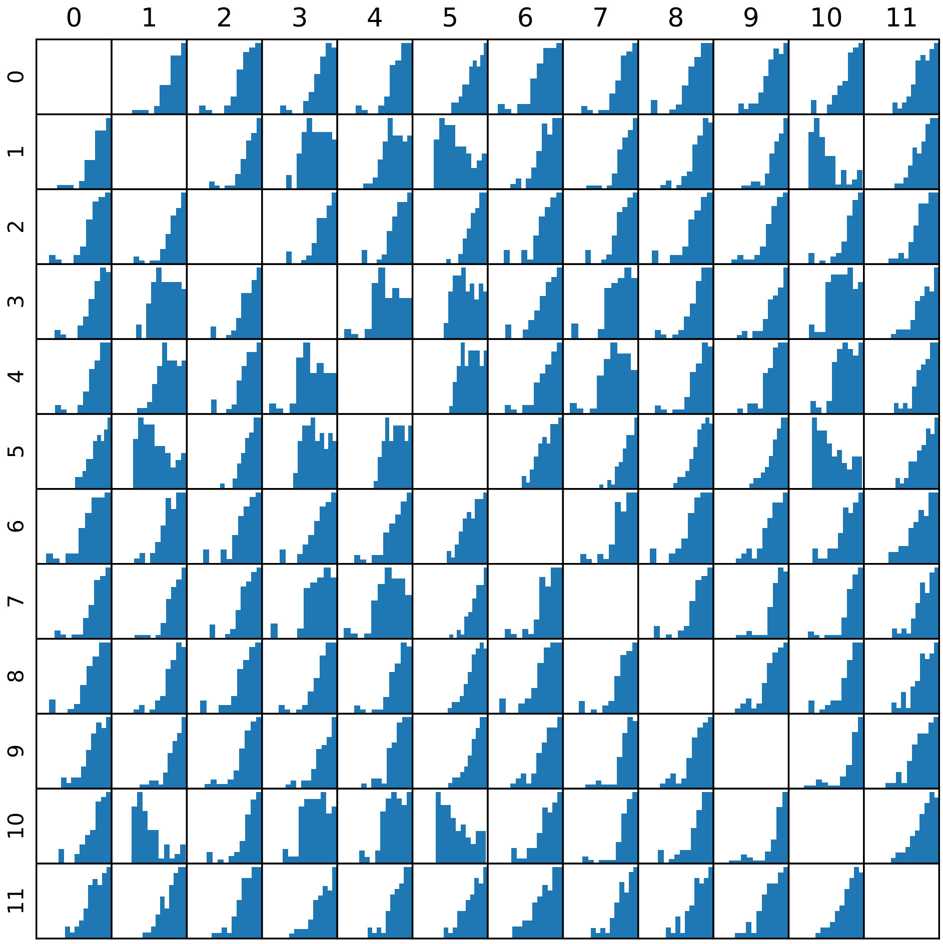
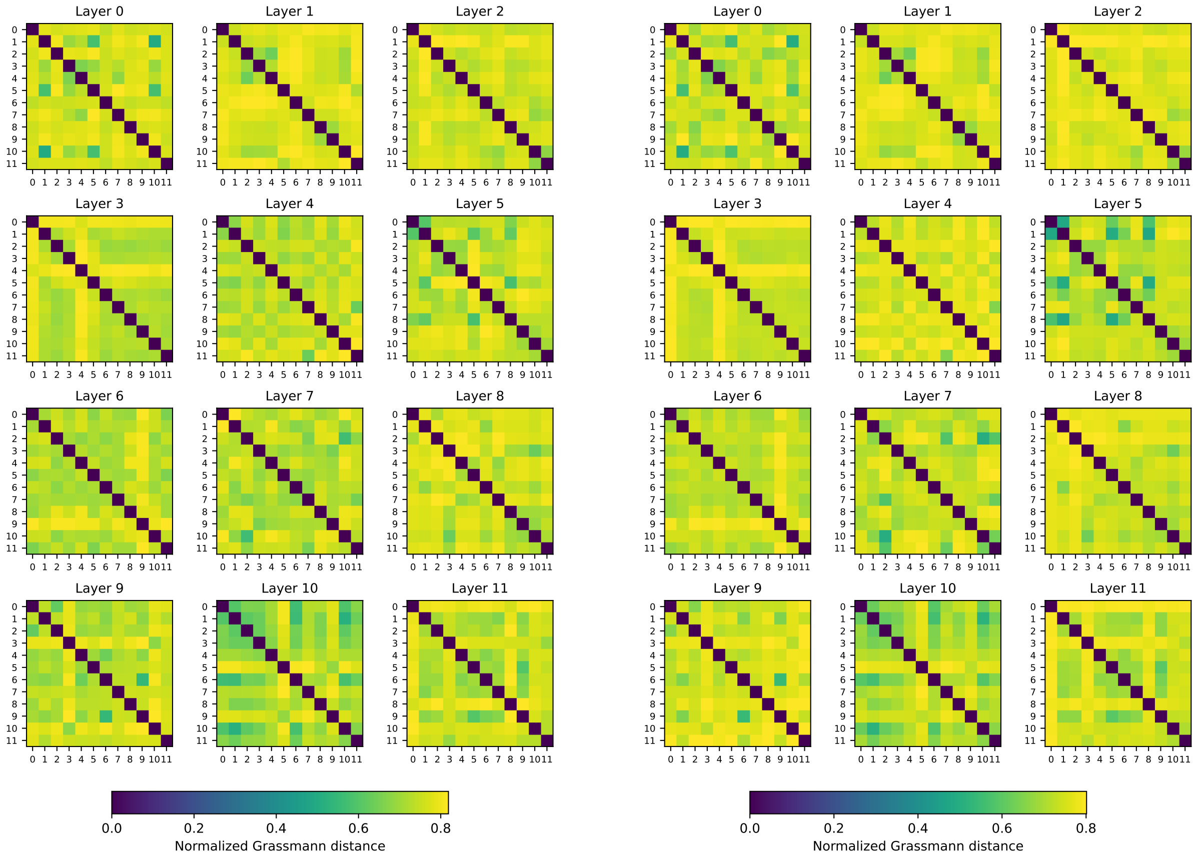


Figure 5: Distribution of the principal angles between query subspaces in Layer 0, defined by the map π_U. As discussed earlier, most of the principal angles are concentrated near $\pi/2$, indicating that the query subspaces of different heads are nearly orthogonal. This aligns with the results shown in Figure 6, where Grassmann distances cluster around 0.8, reinforcing the expected geometric separation between attention heads within the same layer.



(a) Left subspace distances $d(U_i, U_j)$

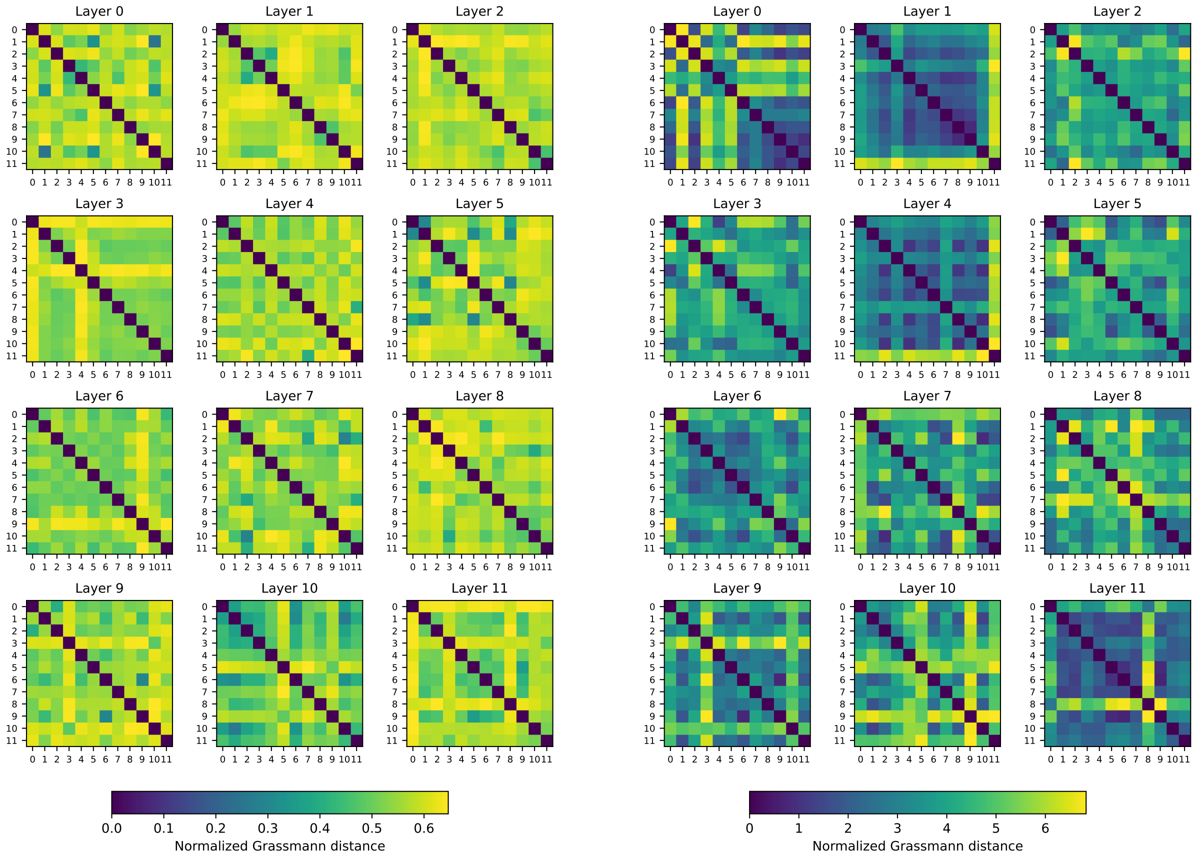
(b) Right subspace distances $d(V_i, V_j)$

Figure 6: Normalized Grassmann distances between the query subspaces (left, induced by the submersion π_U) and key subspaces (right, induced by π_V) for all pairs of attention heads across each layer. A distance of 0 indicates perfect overlap between subspaces, while a distance of 1 corresponds to complete orthogonality. The symmetry observed between the query and key distances suggests a consistent geometric relationship between the two projections. Moreover, the predominance of high distance values indicates that the attention heads are generally well-separated in subspace orientation. This separation is expected, given that each layer divides the 768-dimensional embedding space into twelve distinct 64-dimensional subspaces, and may reflect effective training where heads are not redundant but instead specialize in different aspects of the representation.

Comparison with empirical results

To evaluate the predictive power of this geometric model, we ran an empirical test using 1278 tokens. For each token, we recorded the attention score matrices A_i produced by every head and computed the pairwise differences $||A_i - A_j||$ using the Frobenius norm. These differences were then averaged across all tokens $\frac{1}{N}\sum_k^N ||A_i^k - A_j^k||$. This procedure was repeated for every layer in GPT-2. The empirical results were then compared with the theoretical predictions derived from the subspace distances. The comparison is shown in Figure 7.

In the comparison, we did not include the coupling information provided by the singular values shown in Figure 4. This exclusion is due to two main challenges. First, it is not straightforward to compare the singular values across different decompositions, as their interpretation



(a) Subspace distances $d(U_i, U_j)^2 + d(V_i, V_j)^2$

(b) Empirical heads differences $||A_i - A_j||$

Figure 7: Left: Theoretical distances computed as the sum of Grassmann distances between the query and key subspaces for each pair of attention heads across all layers. Right: Empirical distances obtained by averaging the differences between attention score matrices over 1278 tokens from a set of test inputs. Lower values (blue/green) indicate heads that tend to behave similarly, while higher values (yellow) correspond to heads producing distinct attention patterns. Despite differences in scale, the structural similarity between the theoretical and empirical patterns is, especially in deeper layers (e.g., Layers 9–11), remarkable, given the simplicity of the model.

depends on the alignment of the corresponding subspaces. Second, it is unclear how to meaningfully combine the query and key subspaces U and and V with their respective singular values Σ, since these represent fundamentally different mathematical objects. Although, as shown in Equation 2, we can formally write the attention mechanism using all three components, and define a corresponding distance as in Equation 4, this naive combination does not yield reliable results in the comparison of Figure 7.