LayerNorm Induces Recency Bias in Transformer Decoders

Junu Kim[1,2*] **Xiao Liu**[2] **Zhenghao Lin**[2] **Lei Ji**[2] **Yeyun Gong**[2] **Edward Choi**[1]
[1] KAIST [2] Microsoft Research
{kjune0322,edwardchoi}@kaist.ac.kr
{xiao.liu.msrasia,zhenghaolin,leiji,yegong}@microsoft.com

Abstract

Causal self-attention provides positional information to Transformer decoders. Prior work has shown that stacks of causal self-attention layers alone induce a positional bias in attention scores toward earlier tokens. However, this differs from the bias toward later tokens typically observed in Transformer decoders, known as recency bias. We address this discrepancy by analyzing the interaction between causal self-attention and other architectural components. We show that stacked causal self-attention layers combined with LayerNorm induce recency bias. Furthermore, we examine the effects of residual connections and the distribution of input token embeddings on this bias. Our results provide new theoretical insights into how positional information interacts with architectural components and suggest directions for improving positional encoding strategies.

1 Introduction

In sequence modeling with Transformer decoders (Vaswani et al., 2017), the way positional information is provided to the model is closely tied to performance (Dufter et al., 2022) and its ability to generalize to longer sequence lengths (Zhao et al., 2024). Among the components of a Transformer decoder layer, positional encodings and the causal mask are responsible for supplying positional information (Haviv et al., 2022; Kazemnejad et al., 2023; Chi et al., 2023). While the mechanisms by which positional encodings provide positional information have been extensively studied (Barbero et al., 2025; Su et al., 2024; Press et al., 2022), the corresponding process for the causal mask remains less well understood.

Recent research shows that simply stacking causal self-attention layers can induce an attention bias toward earlier tokens, thereby providing positional information (Wu et al., 2025). However, empirical studies of full Transformer decoder layers yield contrasting results. Specifically, Zuo et al. (2025) show that Transformer decoders exhibit an attention bias toward more recent tokens rather than earlier ones. This phenomenon, known as recency bias, is characteristic of many positional encoding methods (Su et al., 2024; Press et al., 2022; Vaswani et al., 2017). The discrepancy between these findings suggests that additional architectural components, such as LayerNorm (Ba et al., 2016) or residual connections (He et al., 2016), may modulate the positional information induced by the causal mask.

By examining the effects of other architectural components, we show that LayerNorm induces recency bias in Transformer decoders without positional encoding. Formally, stacking causal self-attention layers with LayerNorm induce recency bias, consistent with the observations of Zuo et al. (2025). We further analyze the effects of residual connections and the distribution of input token embeddings on recency bias, both quantitatively and qualitatively. Together, these findings provide theoretical insights into improving positional encoding and length generalization in Transformer decoders.

2 Theoretical Analysis

2.1 Preliminaries

Formally, a single-head pre-LayerNorm (Xiong et al., 2020) Transformer decoder layer defines a function $f^{(l)} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ with $X^{(l)} = f^{(l)}(X^{(l-1)})$, where $X^{(l)}$ denotes the output of the l-th Transformer decoder layer, and n and d correspond to the number of input tokens and the model hidden size, respectively. The superscripts indicate layer indices, and $X^{(0)}$ corresponds to the input token embeddings. We omit the superscripts when clear from context.

First, the normalized input $Y^{(l)}$ is produced by

[*] Work done during an internship at Microsoft Research Asia.

Findings of the Association for Computational Linguistics: ACL 2026, pages 28638–28652

the LayerNorm operation:

$$Y^{(l)} = \text{LayerNorm}(X^{(l-1)}).$$

Then, the query, key, and value matrices Q, K, and V are computed using the learnable weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$:

$$Q = YW_Q, \quad K = YW_K, \quad V = YW_V.$$

The attention score matrix S is computed as:

$$S = \text{Causal}\left(\frac{QK^\top}{\sqrt{d}}\right),$$

where $\text{Causal}(\cdot)$ applies a strictly upper-triangular causal mask (setting masked entries to $-\infty$) to prevent attention to future positions.

Thus, the attention weights A are defined by the row-wise softmax operation:

$$A = \text{Softmax}(S).$$

The attention output $O^{(l)}$ is then computed using a learnable output projection matrix $W_O \in \mathbb{R}^{d\times d}$ together with a residual connection:

$$O^{(l)} = (AV)W_O + X^{(l-1)}.$$

Finally, the hidden state $X^{(l)}$ is computed as:

$$X^{(l)} = \text{FFN}(\text{LayerNorm}(O^{(l)})) + O^{(l)},$$

where FFN denotes a position-wise feed-forward network, typically consisting of two linear layers separated by a nonlinear activation function.

Here, we formally define recency bias as the property that the attention score assigns a higher score to a closer key than to a more distant key for a fixed query.

Definition 1. *The attention score S exhibits recency bias if $S_{ij} > S_{ik}$ for all query indices i and key indices j and k satisfying $i \geq j > k$.*

2.2 LayerNorm

Here, we show that stacked causal self-attention layers with LayerNorm can induce a recency bias, even when the input sequence has no causal dependency, no learnable parameters, and no feed-forward modules. We begin with the case where the input token embeddings follow a normal distribution with zero mean and variance $1/d$, following Reddy (2024) and Wu et al. (2025). We first ignore the residual connection, and then consider cases with residual connections or non-normal input distributions in the following subsections.

Theorem 1. *Let input token embeddings follow $\mathcal{N}(0, \mathbb{I}_d/d)$, and let the architecture be composed of stacked LayerNorm and causal self-attention layers. For hidden sizes $d \gg 1$, the attention score of the second layer $S^{(2)}$ exhibits a recency bias.*

Here, we provide a sketch of the proof; the full version can be found in Appendix A.1.

Proof Sketch Following the assumption, each input token embedding satisfies $x_i^{(0)} \sim \mathcal{N}(0, \mathbb{I}_d/d)$. Under the stated simplifications, a single Layer $f(X)$ acts as:

$$f(X) = \text{Softmax}(\text{Causal}(YY^\top/\sqrt{d}))Y, \quad (1)$$

$$Y = \text{LayerNorm}(X). \quad (2)$$

The operator $\text{Causal}(\cdot)$ applies a strictly upper triangular mask so that a query at position i attends only to keys at position $j \leq i$, and the softmax is applied row-wise:

$$A_{ij}^{(1)} = \begin{cases} \frac{e^{\langle y_i, y_i\rangle/\sqrt{d}}}{e^{\langle y_i, y_i\rangle/\sqrt{d}} + \sum_{k=1}^{i-1} e^{\langle y_i, y_k\rangle/\sqrt{d}}} & (i = j) \\ \frac{e^{\langle y_i, y_j\rangle/\sqrt{d}}}{e^{\langle y_i, y_i\rangle/\sqrt{d}} + \sum_{k=1}^{i-1} e^{\langle y_i, y_k\rangle/\sqrt{d}}} & (i > j) \\ 0 & (i < j) \end{cases}. \quad (3)$$

Since $d \gg 1$, we can apply the law of large numbers. For large d, LayerNorm can be approximated as $\text{LayerNorm}(X) \approx \sqrt{d}X/||X||_2$. Consequently, $\langle y_i^{(1)}, y_i^{(1)}\rangle/\sqrt{d} = \sqrt{d}$. For $i \neq j$, the cross term $\langle y_i^{(1)}, y_j^{(1)}\rangle/\sqrt{d}$ converges in distribution to $\mathcal{N}(0,1)$, which is negligible compared to $\sqrt{d}$. We therefore approximate these terms by zero. Under this approximation,

$$x_i^{(1)} = \sum_{j=1}^{i} A_{ij}^{(1)} y_j^{(1)} = \frac{e^{\sqrt{d}} y_i^{(1)} + \sum_{k=1}^{i-1} y_k^{(1)}}{e^{\sqrt{d}} + i - 1}. \quad (4)$$

Since $x_i^{(1)}$ is a linear combination of zero-mean vectors, the same LayerNorm approximation applies at the next layer. Writing $S_{ij}^{(2)}$ in terms of $x^{(1)}$ and applying the law of large numbers again, for $i > j$ we obtain

$$\begin{aligned} S_{ij}^{(2)} &= \frac{\langle y_i^{(2)}, y_j^{(2)}\rangle}{\sqrt{d}} = \frac{d\langle x_i^{(1)}, x_j^{(1)}\rangle}{\sqrt{d}||x_i^{(1)}|| \cdot ||x_j^{(1)}||} \\ &= \frac{\sqrt{d}(e^{\sqrt{d}} + j - 1)}{\sqrt{e^{2\sqrt{d}} + i - 1}\sqrt{e^{2\sqrt{d}} + j - 1}}. \end{aligned} \quad (5)$$

For fixed i, this expression is strictly increasing in j. For the diagonal case, $S_{ii}^{(2)} = \sqrt{d}$ by construction, and clearly $S_{ii}^{(2)} > S_{i,i-1}^{(2)}$. Therefore, $S_{ij}^{(2)} > S_{ik}^{(2)}$ for all $i \geq j > k$, exhibiting recency bias. □

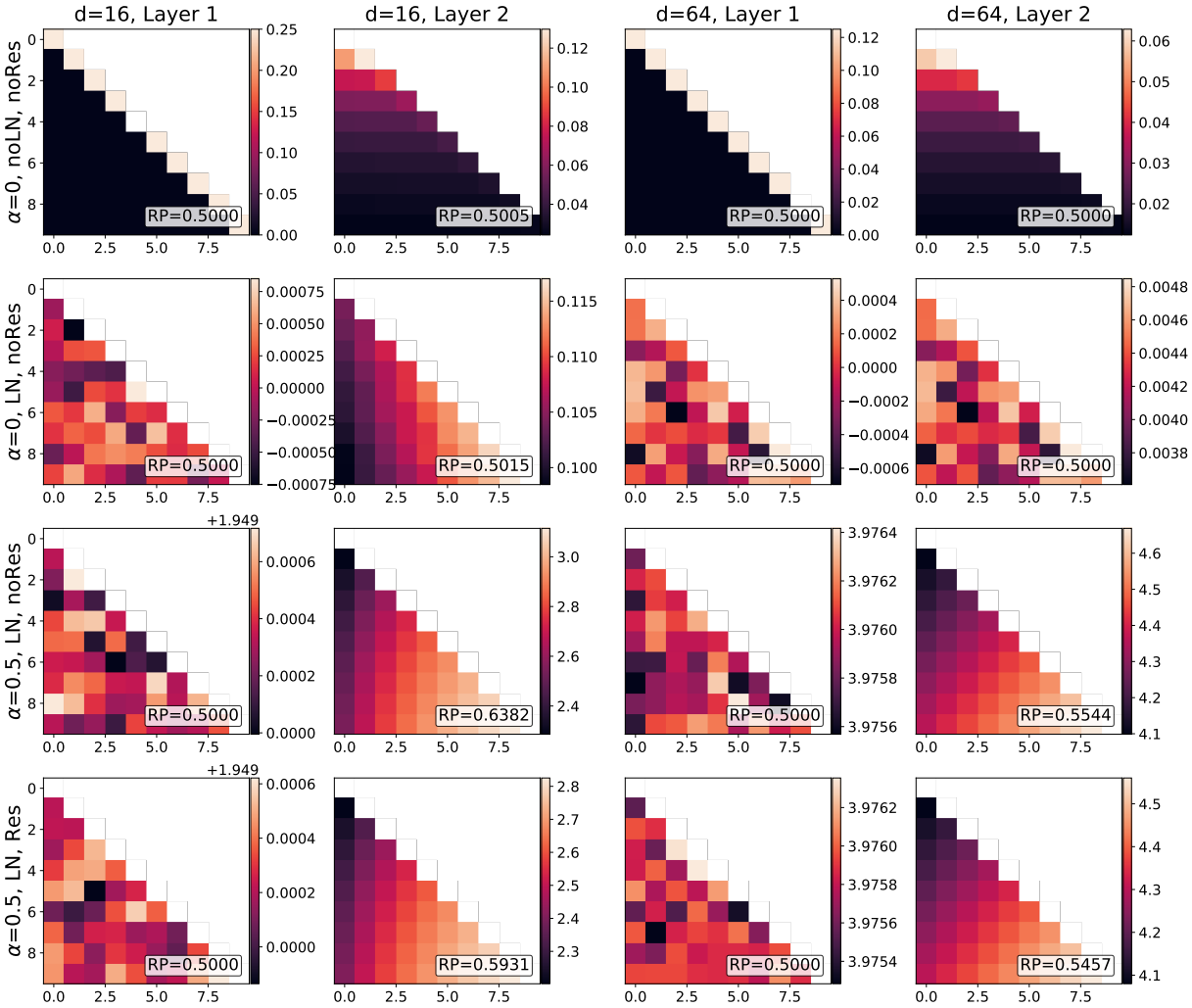

Figure 1: Visualization of the attention scores using a simulation with $d = 16$ (left) and $d = 64$ (right) for layers 1 and 2. LN and Res correspond to LayerNorm and residual connections, respectively. The y-axis represents query indices, and the x-axis represents key indices. To clearly visualize the recency bias, we masked the diagonal elements except those in the first row. Results for layers 3 and 4 are provided in Figure 3, and results for $d = 256$ and $d = 1024$ are provided in Figure 4.

Proposition 1. *Without LayerNorm, $S^{(2)}$ does not exhibit recency bias.*

The proof follows a similar approach to Theorem 1 and can be found in Appendix A.2. Note that this result is consistent with Wu et al. (2025), confirming that simply stacking self-attention layers does not induce recency bias.

To summarize, under the minimal assumptions that the input token embeddings follow a normal distribution and that $d \gg 1$, we show that LayerNorm induces a recency bias at $S^{(2)}$. We empirically show the behavior of later layers in Section 3.

2.3 Residual Connection

In addition, we evaluate the effect of the residual connection on recency bias.

Proposition 2. *Regardless of the existence of the residual connection, $S^{(2)}$ has recency bias.*

The proof can be found in Appendix A.3. We empirically demonstrate its effect on the causal bias in Section 3.

2.4 Distribution of Input Token Embeddings

While we assume that the input token embeddings follow a normal distribution in the previous sections, the input embeddings of pre-trained Transformer decoder models are typically anisotropic; each token embedding has a high cosine similarity with the others (Ethayarajh, 2019). To evaluate the effect of anisotropy in input token embeddings on recency bias, we model anisotropic embeddings $x_i^{(0)}$ with anisotropy level α by adding a shared

vector v to independent Gaussian noise ϵ_i, scaled by a factor $\sqrt{\frac{\alpha}{1-\alpha}}$. Specifically,

$$x_i^{(0)} = \epsilon_i + \sqrt{\frac{\alpha}{1-\alpha}}\, v, \tag{6}$$

where ϵ_i and v are independently drawn from $\mathcal{N}(0, \mathbb{I}_d/d)$.

Proposition 3. *Regardless of the anisotropy of the input token embeddings, $S^{(2)}$ exhibits recency bias.*

The proof can be found in Appendix A.3. We analyze the effect of anisotropic input embeddings in the following section.

3 Empirical Analysis

We further examine how recency bias is induced using a simulation of a Transformer without learnable parameters or positional encodings, as defined in Eq. 2. Specifically, we sample 10 vectors of dimension $d \in \{16, 64\}$ drawn from the distribution defined in Eq. 6. Figure 1 shows the simulated attention scores for layers 1 and 2, with $d = 16$ on the left and $d = 64$ on the right, averaged over 10,000,000 simulations.[1]

To quantitatively assess the presence of recency bias, we introduce the recency probability (RP) metric, inspired by the adjacency probability of Zuo et al. (2025). Given an attention score matrix S_{ij}, the RP is defined as the probability that $S_{ij} > S_{ik}$ for any $i > j > k$, excluding diagonal entries. Under this definition, if the average RP is significantly larger than 0.5, this indicates the presence of a recency bias.

First, consider the case without LayerNorm, corresponding to the first row of Figure 1. As expected, no attention bias is observed in the first layer. In the second layer, the attention scores in each row are uniform across all positions except for the diagonal elements, for both $d = 16$ and $d = 64$. The RP values remain close to 0.5 across all conditions, confirming the absence of recency bias. This behavior is consistent with Proposition 1 and Figure 2 of Wu et al. (2025), which show that stacking causal self-attention layers alone does not induce recency bias, but instead leads to a bias toward earlier tokens.

The second row of Figure 1 corresponds to the case with LayerNorm and $\alpha = 0$. For $d = 64$, the attention scores for the off-diagonal elements are nearly uniform, and no clear recency bias is observed (RP $= 0.5000$). This behavior can be explained by Eq. 5: the $e^{\sqrt{d}}$ term dominates the contribution of j, effectively suppressing the recency bias. However, when $d = 16$, a mild recency bias emerges even with $\alpha = 0$ (RP $= 0.5015$), because the exponential term in the denominator is less dominant at smaller dimensions.

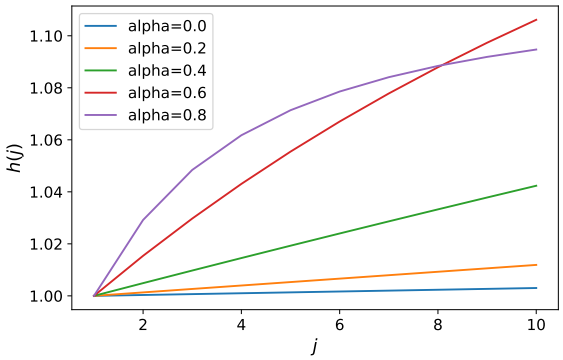

Figure 2: Visualization of $h(j)$ over key index j, for multiple values of α.

The third row of Figure 1 corresponds to the anisotropic input distribution with $\alpha = 0.5$. In the second layer, the attention scores strictly increase as j increases for a fixed i, indicating a clear recency bias for both $d = 16$ (RP $= 0.6382$) and $d = 64$ (RP $= 0.5544$). This behavior can be explained by examining the formulation of $S^{(2)}$. Equation 5 can be decomposed into a term that depends on i and a term that depends on j. Even when $\alpha \neq 0$, a similar decomposition holds, allowing us to write $S^{(2)}_{ij} = g(i)h(j)$. For each row, i is fixed, so $g(i)$ can be treated as a constant, and the variation in the attention scores is governed solely by $h(j)$. The explicit form of $h(j)$ is given in Appendix Eq. 42, and its dependence on α is illustrated in Fig. 2. When $\alpha = 0$, the $e^{\sqrt{d}}$ term dominates, causing $h(j)$ to grow slowly. In contrast, when $\alpha \neq 0$, $h(j)$ increases rapidly, which is consistent with the observed results.

This result is also consistent with Zuo et al. (2025), which show that hidden states associated with nearby query–key pairs exhibit high cosine similarity in Transformer decoders without positional encoding. Since cosine similarity is equivalent to the inner product after ℓ_2 normalization, and ℓ_2 normalization can in turn be approximated by LayerNorm with a scaling factor of $\sqrt{d}$ under mild assumptions, the observed phenomenon can be explained by our theory.

[1]The source code is publicly available at: https://github.com/starmpcc/layernorm_recency_bias.

The last row of Figure 1 corresponds to the case with a residual connection and $\alpha = 0.5$. Compared to row 3, the recency bias observed in row 4 is less pronounced, as reflected by lower RP values ($\mathrm{RP} = 0.5931$ for $d = 16$ and $\mathrm{RP} = 0.5457$ for $d = 64$ in layer 2). Adding a residual connection adds $x_i^{(0)}$ to $x_i^{(1)}$. Since some components are not shared between $x_i^{(0)}$ and $x_j^{(0)}$, the proportion of components shared by $x_i^{(1)}$ and $x_j^{(1)}$ decreases. This reduces the overall scale of the off-diagonal attention scores, making the recency bias less pronounced. A similar analysis for Figure 2 with residual connections (Figure 7), along with attention score visualizations for different values of α (Figures 5 and 6), is provided in Appendix B.

4 Conclusion

In this work, we show that causal self-attention, when combined with LayerNorm, induces a recency bias in attention scores. We further investigate the effects of residual connections and anisotropic input embeddings on this bias. Our theoretical framework naturally extends to modern attention variants used in contemporary LLMs, including Multi-Head Attention (MHA) (Vaswani et al., 2017), Multi-Query Attention (MQA) (Shazeer, 2019), and Grouped-Query Attention (GQA) (Ainslie et al., 2023). In these cases, the $\sqrt{d}$ term in the derivation is replaced by $\sqrt{d_h}$, where d_h is the per-head dimension, and all dimension-dependent terms are correspondingly replaced by d_h, preserving the structure of the derivation and leading to the same qualitative conclusions.

Importantly, although LayerNorm can induce a recency bias, this does not imply relativity. Relativity in positional encoding refers to modeling positional information based on pairwise token distances (Shaw et al., 2018), but the attention scores induced by causal masking and LayerNorm do not satisfy this property. The recency bias in our analysis depends on both the query and key positions rather than solely on their relative distance, meaning that positional interactions between earlier and later tokens may be non-uniform within a single sequence. This non-uniform bias may adversely affect the length generalization capability of Transformer decoders. Therefore, exploring mitigation strategies, such as modifying the causal mask to counteract this uneven bias, appears promising. This suggests that the resulting behavior may differ from that of typical relative positional encodings, including RoPE (Su et al., 2024) and ALiBI (Press et al., 2022). We believe these findings provide valuable insights for the design of future positional encoding methods and for research on length generalization.

Limitations

This study has several limitations. We do not analyze the effects of feed-forward networks, other learnable parameters, or multi-head attention within Transformer decoder layers. Additionally, the relationship between recency bias and overall model performance is not evaluated. Finally, although most modern Transformer decoder-based models use RoPE (Su et al., 2024) for positional encoding, its interaction with the positional information induced by causal self-attention has not been examined.

Acknowledgments

This work was supported by Microsoft Research Asia, the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.RS-2019-II190075, No.RS-2022-II220984, No.RS-2024-00436680), and National Research Foundation of Korea (NRF) grant (RS-2026-25484088), funded by the Korea government (MSIT).

References

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. 2025. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*.

Ta-Chung Chi, Ting-Han Fan, Li-Wei Chen, Alexander Rudnicky, and Peter Ramadge. 2023. Latent positional information is in the self-attention variance of transformer language models without positional embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1183–1193.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2022. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. 2023. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Gautam Reddy. 2024. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

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.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2025. On the emergence of position bias in transformers. In *Forty-second International Conference on Machine Learning*.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International conference on machine learning*, pages 10524–10533. PMLR.

Liang Zhao, Xiachong Feng, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. 2024. Length extrapolation of transformers: A survey from the perspective of positional encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9959–9977.

Chunsheng Zuo, Pavel Guerzhoy, and Michael Guerzhoy. 2025. Position information emerges in causal transformers without positional encodings via similarity of nearby embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9418–9430.

A Proofs

A.1 Proof of Theorem 1

We additionally present the skipped derivation of Equation 5, starting from $S_{ij}^{(2)} = \frac{d\langle x_i^{(1)}, x_j^{(1)} \rangle}{\sqrt{d}||x_i^{(1)}|| \cdot ||x_j^{(1)}||}$.

Proof.

$$\left(\langle x_i^{(1)}, x_j^{(1)} \rangle\right)_{i>j} = \frac{(e^{\sqrt{d}} y_i^{(1)} + \sum_{k=1}^{i-1} y_k^{(1)})(e^{\sqrt{d}} y_j^{(1)} + \sum_{l=1}^{j-1} y_l^{(1)})}{(e^{\sqrt{d}} + i - 1)(e^{\sqrt{d}} + j - 1)} \tag{7}$$

$$= \frac{e^{2\sqrt{d}} \langle y_i^{(1)}, y_j^{(1)} \rangle + e^{\sqrt{d}} \sum_{l=1}^{j-1} \langle y_i^{(1)}, y_l^{(1)} \rangle + e^{\sqrt{d}} \sum_{k=1}^{i-1} \langle y_j^{(1)}, y_k^{(1)} \rangle + \sum_{k=1}^{i-1} \sum_{l=1}^{j-1} \langle y_k^{(1)} y_l^{(1)} \rangle}{(e^{\sqrt{d}} + i - 1)(e^{\sqrt{d}} + j - 1)}. \tag{8}$$

By the law of large numbers, $\langle y_i^{(1)}, y_j^{(1)} \rangle = 0$ for $i \neq j$, and $\langle y_i^{(1)}, y_i^{(1)} \rangle = d$. Thus,

$$\left(\langle x_i^{(1)}, x_j^{(1)} \rangle\right)_{i>j} = \frac{d(e^{\sqrt{d}} + j - 1)}{(e^{\sqrt{d}} + i - 1)(e^{\sqrt{d}} + j - 1)}. \tag{9}$$

For the $||x_i^{(1)}||$ term in the denominator, we again apply the law of large numbers to the case $i = j$ in Eq. 8:

$$||x_i^{(1)}||_2^2 = \langle x_i^{(1)}, x_i^{(1)} \rangle = \frac{d(e^{2\sqrt{d}} + i - 1)}{(e^{\sqrt{d}} + i - 1)^2}. \tag{10}$$

Therefore, for $i > j$,

$$S_{ij}^{(2)} = \frac{\langle y_i^{(2)}, y_j^{(2)} \rangle}{\sqrt{d}} = \frac{d\langle x_i^{(1)}, x_j^{(1)} \rangle}{\sqrt{d}||x_i^{(1)}||_2 \cdot ||x_j^{(1)}||_2} = \frac{\sqrt{d}(e^{\sqrt{d}} + j - 1)}{\sqrt{e^{2\sqrt{d}} + i - 1}\sqrt{e^{2\sqrt{d}} + j - 1}}. \tag{11}$$

Since i and j are positive integers and the order of j in the numerator is larger than that in the denominator, $S_{ij}^{(2)}$ is strictly increasing in j for a fixed i. □

A.2 Proof of Proposition 1

Proof. Without LayerNorm,

$$S_{ij}^{(1)} = \text{Causal}(X^{(0)} X^{(0)\top} / \sqrt{d})_{ij} = \begin{cases} \langle x_i, x_i \rangle / \sqrt{d} & (i = j) \\ \langle x_i, x_j \rangle / \sqrt{d} & (i > j) \\ -\inf & (i < j) \end{cases}. \tag{12}$$

Proceeding in the same manner as in the previous proof, we obtain

$$A_{ij}^{(1)} = \begin{cases} \frac{e^{\langle x_i, x_i \rangle / \sqrt{d}}}{e^{\langle x_i, x_i \rangle / \sqrt{d}} + \sum_{k=1}^{i-1} e^{\langle x_i, x_k \rangle / \sqrt{d}}} & (i = j) \\ \frac{e^{\langle x_i, x_j \rangle / \sqrt{d}}}{e^{\langle x_i, x_i \rangle / \sqrt{d}} + \sum_{k=1}^{i-1} e^{\langle x_i, x_k \rangle / \sqrt{d}}} & (i > j) \\ 0 & (i < j) \end{cases}. \tag{13}$$

Here, we can apply the law of large numbers. Thus,

$$\langle x_i^{(0)}, x_j^{(0)} \rangle = \begin{cases} 1 & (i = j) \\ \approx 0 & (i \neq j). \end{cases} \tag{14}$$

Therefore,

$$A_{ij}^{(1)} = \begin{cases} \frac{e^{1/\sqrt{d}}}{e^{1/\sqrt{d}}+(i-1)} & (i=j) \\ \frac{1}{e^{1/\sqrt{d}}+(i-1)} & (i>j) \\ 0 & (i<j) \end{cases}. \tag{15}$$

Thus,

$$x_i^{(1)} = \sum_{j=1}^{i} A_{ij}^{(1)} x_j^{(1)} = \frac{e^{1/\sqrt{d}} x_i^{(1)} + \sum_{k=1}^{i-1} x_k^{(1)}}{e^{1/\sqrt{d}} + i - 1}, \tag{16}$$

$$\left(S_{ij}^{(2)}\right)_{i>j} = \langle x_i^{(1)}, x_j^{(1)} \rangle = \frac{(e^{1/\sqrt{d}} x_i^{(1)} + \sum_{k=1}^{i-1} x_k^{(1)})(e^{1/\sqrt{d}} x_j^{(1)} + \sum_{l=1}^{j-1} x_l^{(1)})}{(e^{1/\sqrt{d}} + i - 1)(e^{1/\sqrt{d}} + j - 1)} \tag{17}$$

$$= \frac{e^{1/\sqrt{d}} + j - 1}{(e^{1/\sqrt{d}} + i - 1)(e^{1/\sqrt{d}} + j - 1)} \tag{18}$$

$$= \frac{1}{e^{1/\sqrt{d}} + i - 1}. \tag{19}$$

Therefore, the attention score in the second layer without LayerNorm does not exhibit recency bias. □

A.3 Proof of Proposition 2 and 3

Here, we consider both anisotropic input embeddings and residual connections. To incorporate these effects, we rewrite Eq. 2 to include residual networks:

$$f(X) = \mathrm{Softmax}(\mathrm{Causal}(YY^\top)/\sqrt{d})Y + \gamma X, \tag{20}$$

$$Y = \sqrt{d}X/||X||_2, \tag{21}$$

where γ is a constant: $\gamma = 0$ corresponds to the model without residual connections, and $\gamma = 1$ corresponds to the model with residual connections.

Proof. We first compute the expectation of $\langle y_i^{(0)}, y_j^{(0)} \rangle$ when $i \neq j$. For the denominator,

$$\mathbb{E}\left[||x_i^{(1)}||_2^2\right] = \mathbb{E}\left[(\epsilon_i + \sqrt{\frac{\alpha}{1-\alpha}}v)(\epsilon_i + \sqrt{\frac{\alpha}{1-\alpha}}v)\right] \tag{22}$$

$$= \mathbb{E}[\langle \epsilon_i, \epsilon_i \rangle] + 2\sqrt{\frac{\alpha}{1-\alpha}}\mathbb{E}[\langle \epsilon_i, v \rangle] + \frac{\alpha}{1-\alpha}\mathbb{E}[\langle v, v \rangle]. \tag{23}$$

Since $d \gg 1$, the inner product in the second term is negligible compared to those in the first and third terms. Thus, $\mathbb{E}\left[||x_i^{(1)}||_2^2\right] \approx \sqrt{d}/(1-\alpha)$. Therefore,

$$\langle y_i^{(0)}, y_j^{(0)} \rangle = \sqrt{d}\frac{(\epsilon_i + cv)(\epsilon_j + cv)}{\frac{d}{1-\alpha}} = \sqrt{d}\alpha \tag{24}$$

Following the same procedure as in the above proof, we obtain

$$x_i^{(1)} = \sum_{j=1}^{i} A_{ij}^{(1)} y_j^{(1)} + \gamma x_j^{(1)} \tag{25}$$

$$= \frac{e^{\sqrt{d}} y_i^{(1)} + e^{\sqrt{d}\alpha} \sum_{k=1}^{i-1} y_k^{(1)}}{e^{\sqrt{d}} + e^{\sqrt{d}\alpha}(i-1)} + \sqrt{d}\gamma ||x_i^{(1)}||_2 y_i^{(1)}. \tag{26}$$

We denote the denominator by D_i, so that

$$x_i^{(1)} = \frac{(d\gamma D_i + e^{\sqrt{d}})y_i^{(1)} + e^{\sqrt{d}\alpha}\sum_{k=1}^{i-1} y_k^{(1)}}{D_i} \tag{27}$$

$$\begin{aligned}
\left(\langle x_i^{(1)}, x_j^{(1)}\rangle\right)_{i>j} &= \frac{1}{D_i D_j}\Bigg[((d\gamma D_i + e^{\sqrt{d}})y_i^{(1)} + e^{\sqrt{d}\alpha}\sum_{k=1}^{i-1} y_k^{(1)}) \\
&\quad \times ((d\gamma D_j + e^{\sqrt{d}})y_j^{(1)} + e^{\sqrt{d}\alpha}\sum_{l=1}^{j-1} y_l^{(1)})\Bigg] && (28)\\
&= \frac{d}{D_i D_j}\Bigg[\alpha(d\gamma D_i + e^{\sqrt{d}})(d\gamma D_j + e^{\sqrt{d}}) + \alpha e^{\sqrt{d}\alpha}(d\gamma D_i + e^{\sqrt{d}})(j-1) \\
&\quad + \alpha e^{\sqrt{d}\alpha}(d\gamma D_j + e^{\sqrt{d}})(i-2) + e^{\sqrt{d}\alpha}(d\gamma D_j + e^{\sqrt{d}}) + e^{2\sqrt{d}\alpha}(j-1) \\
&\quad + \alpha e^{2\sqrt{d}\alpha}(i-2)(j-1)\Bigg] && (29)\\
&= \frac{d}{D_i D_j}\Bigg[(j-1)(\alpha e^{\sqrt{d}\alpha}(d\gamma D_i + e^{\sqrt{d}}) + e^{2\sqrt{d}\alpha} + \alpha e^{2\sqrt{d}\alpha}(i-2)) \\
&\quad + (d\gamma D_j + e^{\sqrt{d}})(\alpha(d\gamma D_i + e^{\sqrt{d}}) + \alpha e^{\sqrt{d}\alpha}(i-2) + e^{\sqrt{d}\alpha})\Bigg] && (30)\\
&= d\frac{(e^{\sqrt{d}\alpha}(j-1) + d\gamma D_j + e^{\sqrt{d}})(\alpha(d\gamma D_i + e^{\sqrt{d}}) + e^{\sqrt{d}\alpha} + \alpha e^{\sqrt{d}\alpha}(i-2))}{D_i D_j} && (31)\\
&= d\frac{D_j(1+d\gamma)(\alpha(d\gamma D_i + e^{\sqrt{d}}) + e^{\sqrt{d}\alpha} + \alpha e^{\sqrt{d}\alpha}(i-2))}{D_i D_j} && (32)\\
&= d\frac{(1+d\gamma)(\alpha(d\gamma D_i + e^{\sqrt{d}}) + e^{\sqrt{d}\alpha} + \alpha e^{\sqrt{d}\alpha}(i-2))}{D_i} && (33)
\end{aligned}$$

Formally, we aim to show that $S_{ij}^{(2)} > S_{ik}^{(2)}$ for any $i \geq j > k$. Since $S_{ij}^{(2)} = d\frac{\langle x_i^{(1)}, x_j^{(1)}\rangle}{||x_i^{(1)}||\cdot||x_j^{(1)}||}$, and the numerator and $||x_i^{(1)}||_2$ are independent of j, it suffices to show that $||x_j^{(1)}||_2$ is strictly decreasing in j.

$$\begin{aligned}
||x_j^{(1)}||_2^2 &= \frac{((\gamma D_j + e^{\sqrt{d}})y_j^{(1)} + e^{\alpha\sqrt{d}}\sum_{k=1}^{j-1} y_k^{(1)})((\gamma D_j + e^{\sqrt{d}})y_j^{(1)} + e^{\alpha\sqrt{d}}\sum_{k=1}^{j-1} y_k^{(1)})}{D_j D_j} && (34)\\
&= \frac{d}{D_j^2}\Bigg[(\gamma D_j + e^{\sqrt{d}})^2 + 2d\alpha(\gamma D_j + e^{\sqrt{d}})de^{\alpha\sqrt{d}}(j-1) + e^{2\alpha\sqrt{d}}(j-1) \\
&\quad + d\alpha e^{2\alpha\sqrt{d}}(j-2)(j-1)\Bigg] && (35)\\
&= \frac{d}{D_j^2}\Bigg[(\gamma D_j + e^{\sqrt{d}})^2 + 2(\gamma D_j + e^{\sqrt{d}})\alpha(D_j - e^{\sqrt{d}}) + e^{\sqrt{d}\alpha}(D_j - e^{\sqrt{d}}) \\
&\quad + \alpha(D_j - e^{\sqrt{d}} - e^{\sqrt{d}\alpha})(D_j - e^{\sqrt{d}})\Bigg] && (36)\\
&= d\frac{D_j^2(\gamma^2 + 2\alpha\gamma + \alpha) + D_j(1-\alpha)(2\gamma e^{\sqrt{d}} + e^{\alpha\sqrt{d}}) + (1-\alpha)(e^{2\sqrt{d}} - e^{\sqrt{d}+\sqrt{d}\alpha})}{D_j^2} && (37)
\end{aligned}$$

Since $||x_j^{(1)}||_2 \geq 0$, it suffices to show that $||x_{j+1}^{(1)}||_2^2 - ||x_j^{(1)}||_2^2 < 0$:

$$||x_{j+1}^{(1)}||_2^2 - ||x_j^{(1)}||_2^2 \tag{38}$$

$$= d(1-\alpha)(2\gamma e^{\sqrt{d}} + e^{\alpha\sqrt{d}})(\frac{1}{D_{j+1}} - \frac{1}{D_j}) + d(1-\alpha)(e^{2\sqrt{d}} - e^{\sqrt{d}+\alpha\sqrt{d}})(\frac{1}{D_{j+1}^2} - \frac{1}{D_j^2}) \quad (39)$$

$$= d(1-\alpha)(2\gamma e^{\sqrt{d}} + e^{\alpha\sqrt{d}})(\frac{-e^{\alpha\sqrt{d}}}{D_j D_{j+1}}) + d(1-\alpha)(e^{2\sqrt{d}} - e^{\sqrt{d}+\sqrt{d}\alpha})(\frac{-e^{\alpha\sqrt{d}}(2D_j + e^{\alpha\sqrt{d}})}{D_j^2 D_{j+1}^2}) \quad (40)$$

Since $0 \leq \alpha < 1$, both terms are strictly negative. Therefore, $||x_j^{(1)}||_2$ is strictly decreasing in j. Consequently, $S_{ij}^{(2)}$ is strictly increasing in j for fixed i, exhibiting a recency bias regardless of the presence of residual connections or the anisotropy of the input embeddings. □

From the above proof, since $\langle x_i^{(1)}, x_j^{(1)} \rangle$ is independent of j, we have

$$\left(S_{ij}^{(2)}\right)_{i>j} = d\frac{\langle x_i^{(1)}, x_j^{(1)} \rangle}{||x_i^{(1)}|| \cdot ||x_j^{(1)}||} = \left(d\frac{\langle x_i^{(1)}, x_j^{(1)} \rangle}{||x_i^{(1)}||}\right)\left(\frac{1}{||x_j^{(1)}||}\right) = g(i) \cdot h(j). \quad (41)$$

Therefore,

$$h(j) = \frac{D_j}{\sqrt{d}\sqrt{D_j^2(\gamma^2 + 2\alpha\gamma + \alpha) + D_j(1-\alpha)(2\gamma e^{\sqrt{d}} + e^{\alpha\sqrt{d}}) + (1-\alpha)(e^{2\sqrt{d}} - e^{\sqrt{d}+\sqrt{d}\alpha})}}. \quad (42)$$

B Extended Results

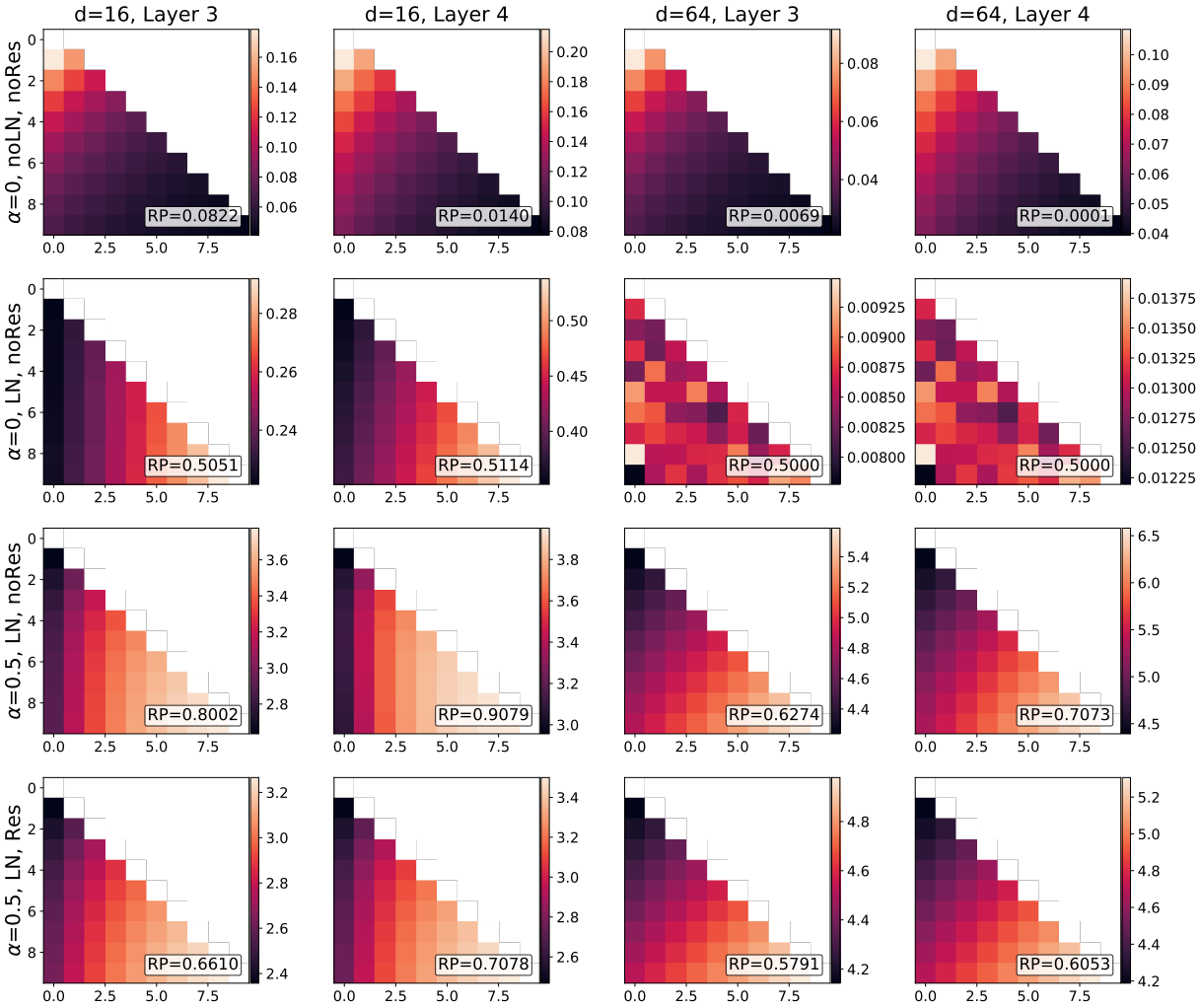

Figure 3: Visualization of the attention scores for layers 3 and 4 with $d = 16$ (left) and $d = 64$ (right). The layout follows Figure 1.

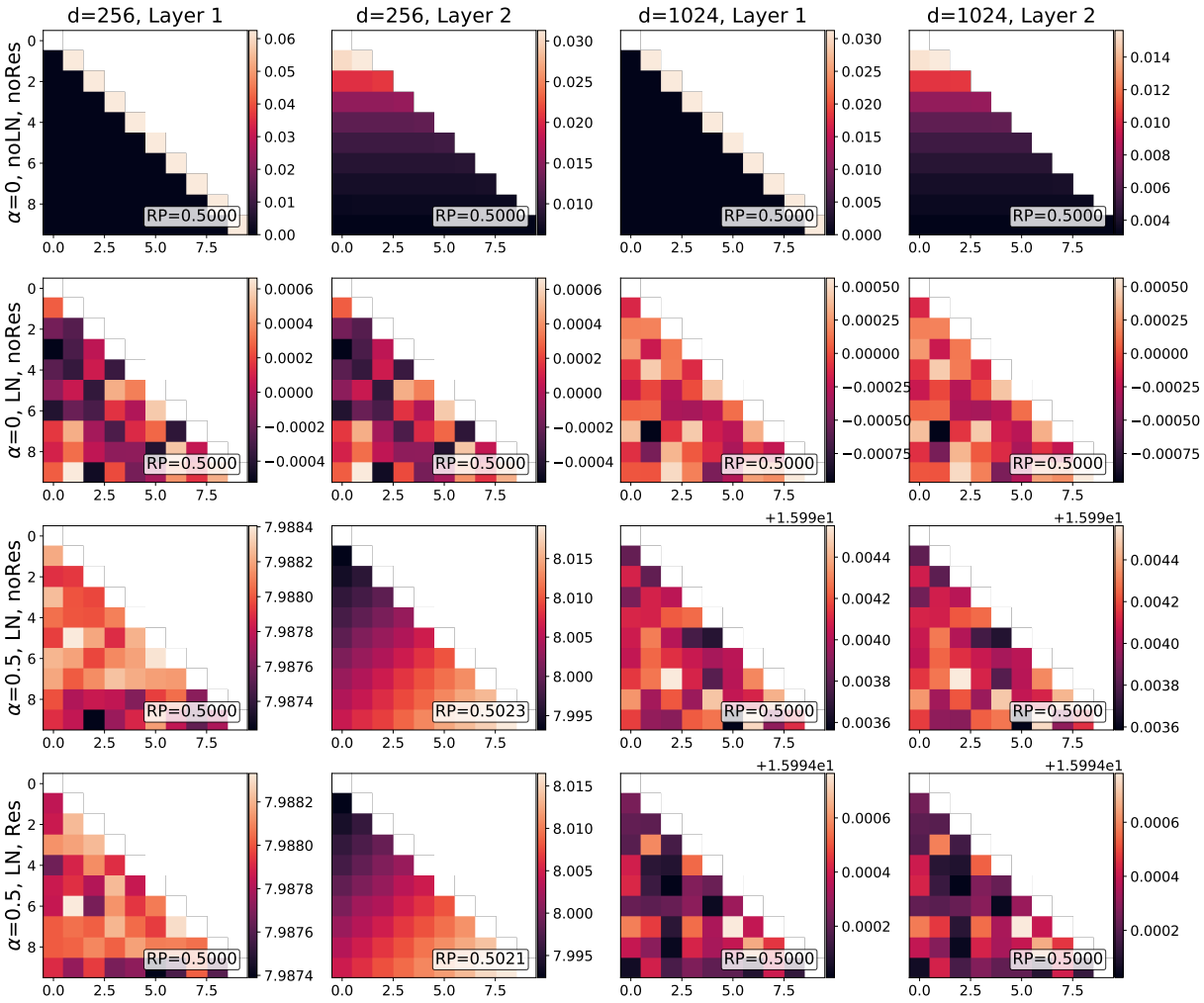

Figure 4: Visualization of the attention scores for layers 1 and 2 with $d = 256$ (left) and $d = 1024$ (right). The layout follows Figure 1.

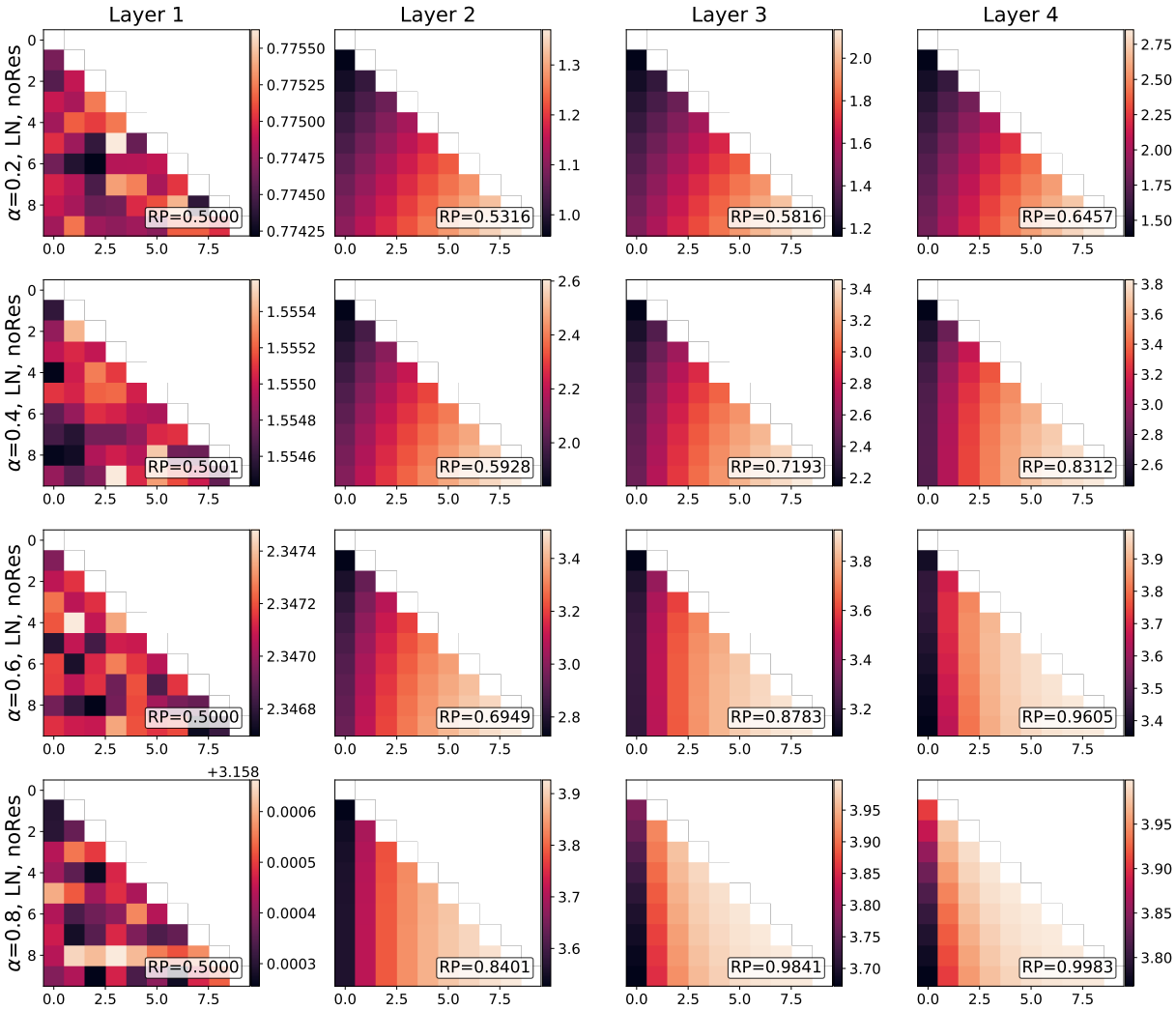

Figure 5: Extended results of Figure 1 with multiple α values and no residual connections.

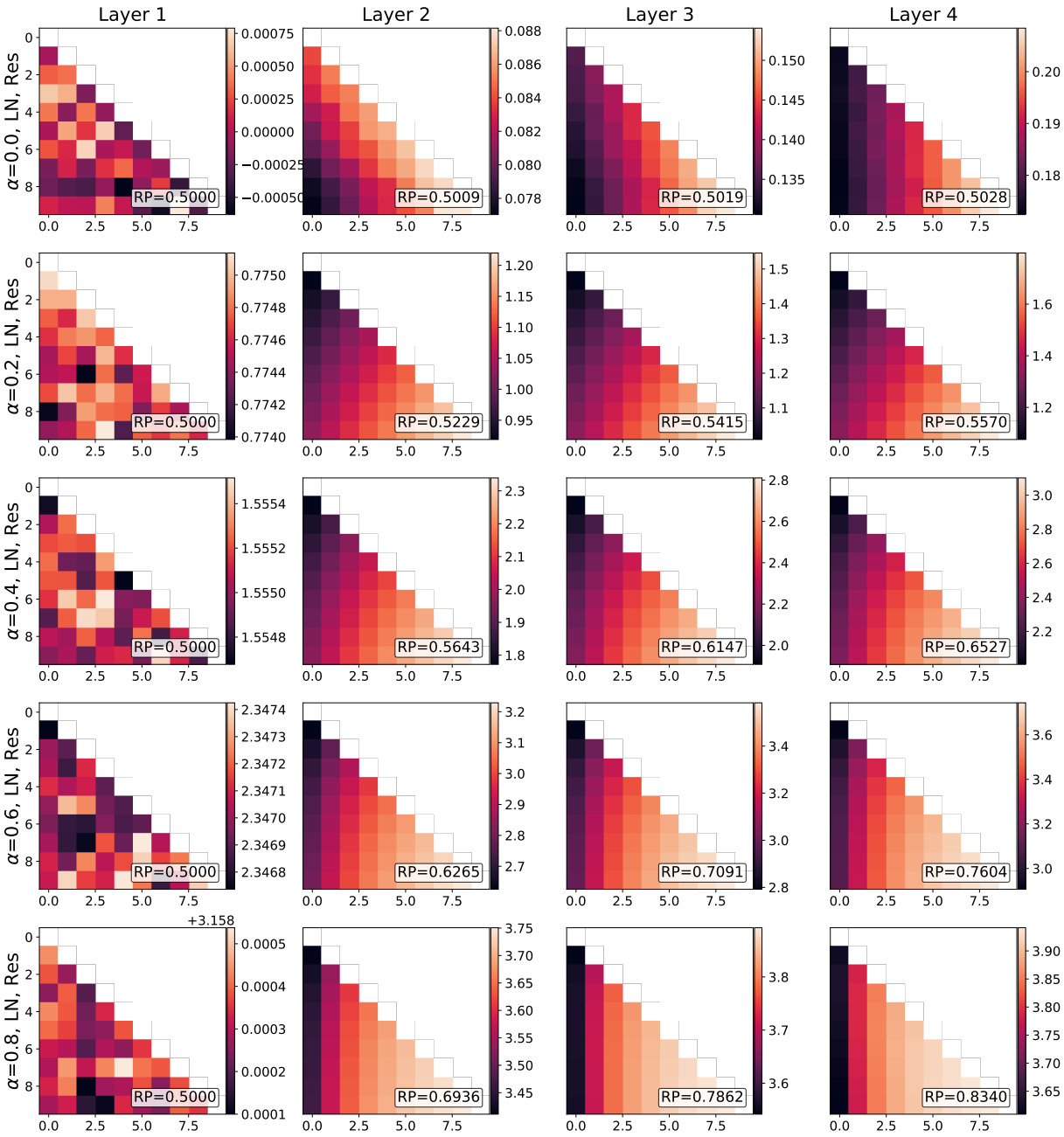

Figure 6: Extended results of Figure 1 with multiple α values and with residual connections.

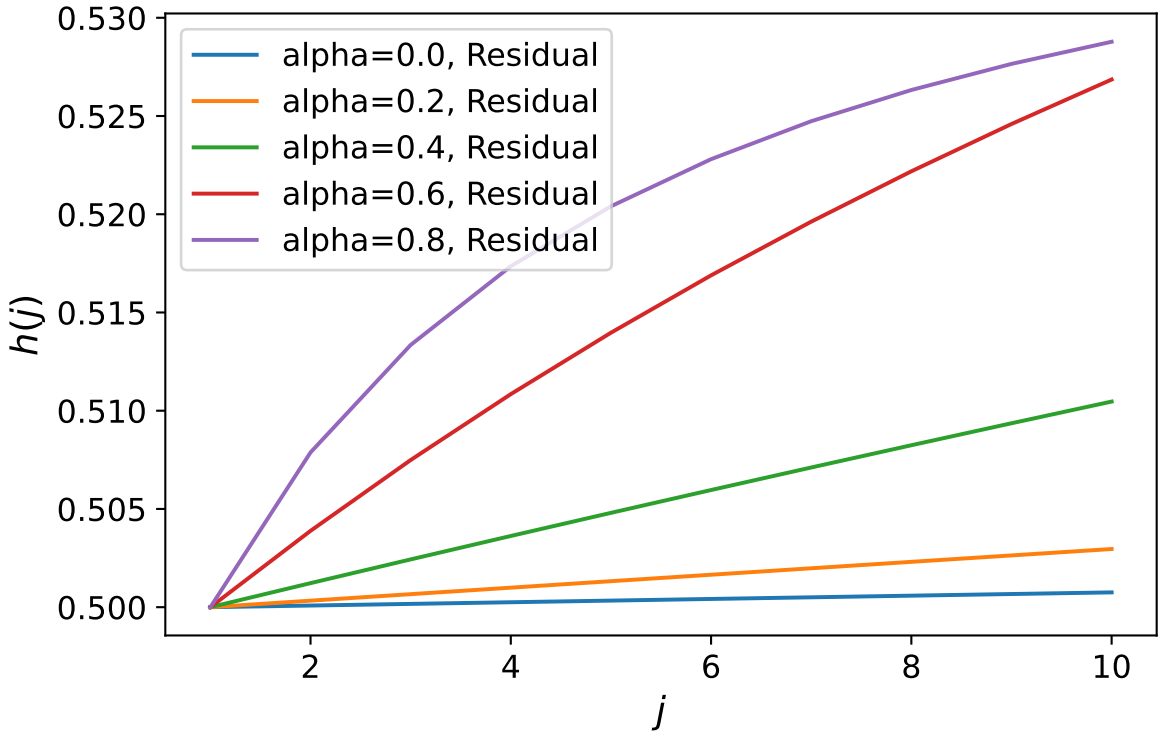

Figure 7: Visualization of $h(j)$ over key index j, for multiple values of α, including residual connections.