Training Instability of Transformers with Softmax and Lipschitz-Kernel Attentions

Anonymous ACL submission

Abstract

Attention-based language models usually rely on the softmax function to convert attention logits into probability vectors. However, this process can lead to *attention entropy collapse*, where the attention concentrates on a single token, causing training instability. In this work, we identify high *variance-entropy sensitivity* of softmax as a root cause of this phenomenon and reproduce it with large language models (LLMs) and a simple Transformer model, demonstrating that *Lipschitz-kernel*-based attention is robust against attention entropy collapse. We demonstrate through controlled and real training settings that Lipschitz-kernel-based and softmax-based attention exhibit differences in sensitivity to *attention logits variance*. We reveal that the high sensitivity of softmax-based attention to the variance contributes to attention entropy collapse. Moreover, we argue that attention entropy collapse leads to training instability because, as attention probabilities become more concentrated, the norm of the attention probability matrix increases, ultimately causing a gradient explosion.

1 Introduction

Attention-based language models convert the attention logits (the query-key dot product) into probability vectors using the softmax function, reflecting each token's relative importance. However, this process can result in excessive focus on a single token, leading to attention entropy collapse (also known as attention sink) (Zhai et al., 2023; He et al., 2024; Xiao et al., 2024; Guo et al., 2024a,b; Yu et al., 2024). Previous studies suggest that multiple factors contribute to this collapse, including large attention logits (Xiao et al., 2024; Wortsman et al., 2024; Dehghani et al., 2023; He et al., 2024), exploding norms of hidden states or activations (Sun et al., 2024), and specific model components such as layer normalization, residual connections, and MLP layers (Gu et al., 2025; Cancedda, 2024).

The core issue of attention entropy collapse in softmax-based attention lies in the exponential nature of the softmax function. The softmax function amplifies differences in attention logits, leading to an increasingly disproportionate focus on a single token as the gap between attention logits grows. This property leads to *attention entropy collapse*, forcing the attention probabilities to collapse into one-hot-like vectors and resulting in training instability (Zhai et al., 2023; Wortsman et al., 2024; He et al., 2024). While several studies have investigated the role of this collapse in training instability, the exact mechanisms through which these instabilities emerge remain unclear.

In this work, we focus on the sensitivity of the softmax function, which amplifies differences among attention logits, causing larger attention logits to dominate the attention probabilities disproportionately. We demonstrate that approximating softmax-based attention with *Lipschitz-kernel* prevents attention entropy collapse and enables more stable training. Specifically, in Figure 1 (Top), based on experiments with open-source LLM, we show that with softmax-based attention, the average attention entropy tends to progressively decrease (the third panel). This collapse leads to an increase in the norm of the attention probability matrix (the fourth panel), ultimately resulting in unstable gradients (the second panel).

Additionally, although prior studies have identified multiple causes of attention entropy collapse, the complexity of LLMs makes it challenging to isolate individual contributing factors. To focus on the attention re-weighting function, we employ a simple and small-scale architecture composed solely of attention layers. As shown in Figure 1 (Bottom), even in this small-scale model, we observe consistent results with those in large-scale experiments. Furthermore, softmax-based attention induces the attention entropy collapse, eventually reducing it to zero, leading to loss divergence.

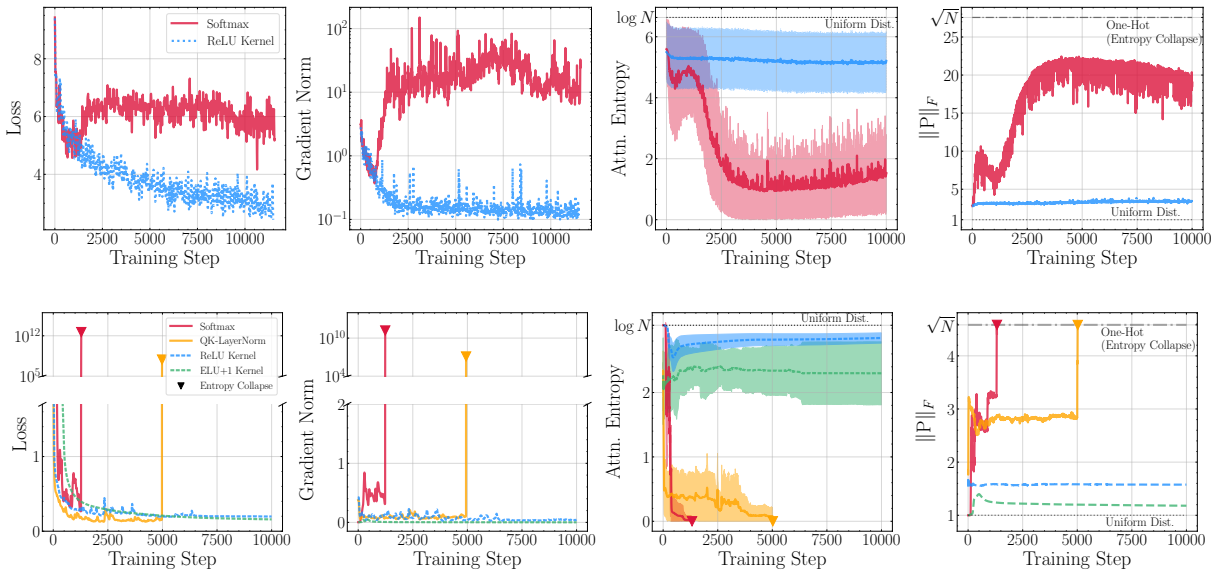

Figure 1: The training behaviors of Llama1-1B (Top, $N = 768$) and a small-scale Transformer model (Bottom, $N = 20$). From left to right, each column shows the training loss (Loss), gradient norm (Gradient Norm), the first layer's average attention entropy with $\pm$ standard deviation of attention entropy (Attn. Entropy), and the average Frobenius norm of the attention probability matrix across all layers ($\|P\|_F$). For the average attention entropy of other layers, see Appendix B. In the third column, as the attention probability becomes uniform, the average attention entropy reaches its maximum ($\log N$, dotted line). In the fourth column, $\|P\|_F$ reaches its maximum ($\sqrt{N}$, dashed-dotted line) when attention entropy collapse (▼) occurs and its minimum (dotted line) under a uniform attention distribution, following Proposition 5.2.

To better understand the distinct behaviors of the two re-weighting functions (softmax and Lipschitz-kernel) in self-attention, we analyze their handling of input bound and variance. Softmax-based attention, with scaling to increase the input bound, amplifies larger attention logits and increases their relative dominance, leading to attention entropy collapse. In contrast, Lipschitz-kernel-based attention applies scaling in a way that affects both the numerator and denominator proportionally, preventing one attention logit from disproportionately dominating the others. Thus, the key factor determining the attention entropy collapse is the level of sensitivity to attention logits variance. To empirically analyze the sensitivity of the two attention mechanisms, we conduct experiments in both controlled and real training settings, increasing variance causes softmax-based attention to exhibit a sharp drop in entropy, whereas Lipschitz-kernel-based attention remains relatively high entropy even with similar variance.

Moreover, as shown in Figure 1 (the second column), the gradient norm explodes around the step where the average attention entropy decreases or approaches zero during training, leading to training instability. This suggests that attention entropy collapse plays a crucial role in training instability, necessitating further analysis. As attention probabilities become increasingly concentrated (attention entropy collapse), the attention probability matrix norm grows rapidly, exploding gradients during backpropagation and causing training instability. Our experiments confirm that softmax-based attention makes this instability more pronounced and more likely to occur, while Lipschitz-kernel-based attention effectively mitigates it by preventing attention entropy collapse.

2 Related Works

Several studies have analyzed the causes and consequences of self-attention excessively focusing on single tokens, a phenomenon called attention entropy collapse or attention sink. One identified issue is that when the query and key weights have large norms, the lower bound of attention entropy becomes tighter, leading to training instability (Zhai et al., 2023). Additionally, as the magnitude of attention logits increases, attention probabilities tend to collapse into one-hot-like vectors, further contributing to training instability (Kedia et al., 2024). This issue can be mitigated through normalization techniques, such as directly normalizing

the attention logits or individually normalizing the query and key (He et al., 2024). Representative methods include QK-LayerNorm (Dehghani et al., 2023), QKNorm (Henry et al., 2020), and Norm-Softmax (Jiang et al., 2023). This phenomenon is often characterized by excessive attention bias toward initial tokens, commonly referred to as an attention sink (Xiao et al., 2024). A few activation units with disproportionately large values concentrate attention probabilities on their corresponding tokens (Sun et al., 2024). Empirical analysis reveals that factors such as QK angles, optimization strategies, data distribution, loss functions, and model architecture also influence this phenomenon (Gu et al., 2025). Moreover, as value norms decrease, residual-state peaks emerge, exacerbating the attention sink problem by causing value-state drains (Guo et al., 2024a).

3 Background

3.1 Softmax-based Attention

Given an input $X \in \mathbb{R}^{N\times D}$, where N denotes the sequence length and D the hidden dimension, we define the three components of a single-head attention mechanism—query $Q \in \mathbb{R}^{N\times D}$, key $K \in \mathbb{R}^{N\times D}$, value $V \in \mathbb{R}^{N\times D}$—by multiplying X by each corresponding weight $W_Q, W_K, W_V \in \mathbb{R}^{D\times D}$. The ith row vector $A_i \in \mathbb{R}^{1\times D}$ of self-attention's output $A \in \mathbb{R}^{N\times D}$ and (i,j)th elements of the attention probability matrix $P \in \mathbb{R}^{N\times N}$ can be defined as follows:

$$A_i = \sum_{j=1}^{N} P_{i,j}V_j \text{ and } P_{i,j} = \frac{\text{sim}(Q_i, K_j)}{\sum_{k=1}^{N}\text{sim}(Q_i, K_j)}, \tag{1}$$

where $\text{sim}(\cdot)$ is a real-valued function that measures the similarity between query and key.

Softmax-based attention uses the exponentiated query-key dot product for the similarity function

$$\text{sim}(Q_i, K_j) = \exp(Q_i K_j^\top)$$

and the corresponding attention probability matrix is

$$P_{i,j} = \frac{\exp(Q_i K_j^\top)}{\sum_{k=1}^{N}\exp(Q_i K_k^\top)}.$$

We refer to $Z = QK^\top \in \mathbb{R}^{N\times N}$ as the attention logits.

3.2 Linear Kernerlized Attention

To mitigate the quadratic complexity of traditional attention mechanisms, several efficient approaches have been proposed, such as sparse pattern (Beltagy et al., 2020; Zaheer et al., 2020), low-rank approximations (Wang et al., 2020; Hu et al., 2022) and kernelized self-attention (Choromanski et al., 2021; Cai et al., 2023). Among these approaches, kernelized self-attention approximates the similarity function using a kernel function $\phi : \mathbb{R}^{1\times D} \to \mathbb{R}^{1\times D}$ as follows:

$$\text{sim}(Q_i, K_j) \approx \phi(Q_i)\phi(K_j)^\top. \tag{2}$$

Instead of directly applying the softmax function, kernelized self-attention reformulates the similarity function with a kernel function ϕ for computational efficiency. Leveraging the associative property of matrix multiplication, it avoids explicit attention probability matrix computation, reducing quadratic to linear complexity as follows:

$$A'_i = \frac{\phi(Q_i)\sum_{j=1}^{N}\phi(K_j)^\top V_j}{\phi(Q_i)\sum_{k=1}^{N}\phi(K_k)^\top} \text{ and } P'_{i,j} = \frac{\phi(Q_i)\phi(K_j)^\top}{\sum_{k=1}^{N}\phi(Q_i)\phi(K_k)^\top}. \tag{3}$$

Based on (3), to compute the output A, we can calculate $\phi(K)^\top V \in \mathbb{R}^{D\times D}$ instead of the query-key dot product, $QK^\top \in \mathbb{R}^{N\times N}$, and reduce the quadratic time complexity to $\mathcal{O}(ND^2) \approx \mathcal{O}(N)$ assuming that the hidden dimension is much smaller than the sequence length.

While most research on kernelized self-attention primarily focuses on selecting an appropriate kernel function ϕ to better approximate the softmax-based attention such as ReLU (with re-weighting) (Qin et al., 2022; Cai et al., 2023; Han et al., 2023), ELU+1 (Katharopoulos et al., 2020), and others (Chen et al., 2021; Arora et al., 2024; Aksenov et al., 2024; Zhang et al., 2024a), our work instead examines kernel function from the perspective of stability.

3.3 Lipschitz-Kernel Function

The softmax function in self-attention lacks Lipschitz continuity because it amplifies small input differences exponentially, leading to unbounded output changes (Dasoulas et al., 2021; Kim et al., 2021). For comparison with the softmax-based attention, we experiment with kernelized attention

with a Lipschitz kernel function, *Lipschitz-kernel-based attention* (see Definition 3.1 below), which is expected to mitigate attention entropy collapse.

Definition 3.1 (Lipschitz Kernel Attention)**.** Kernelized attention in (3) is called *Lipschitz kernel attention* when the kernel function ϕ is Lipschitz, i.e., there is a constant $\alpha > 0$ such that, for any x, x',

$$\|\phi(x) - \phi(x')\| \leq \alpha \|x - x'\|.$$

Specifically, we use ReLU (Qin et al., 2022; Cai et al., 2023; Han et al., 2023) and ELU+1 (Katharopoulos et al., 2020), both simple and widely used Lipschitz kernel functions, which ensure non-negative values with the Lipschitz constant $\alpha = 1$.

3.4 Attention Entropy

The entropy of each row P_i of the attention probability matrix P, also called *attention entropy*, is defined as follows:

$$H(P_i) = -\sum_{j=1}^{N} P_{i,j} \log P_{i,j}. \tag{4}$$

To compute the average attention entropy across all rows, we take the mean of $H(P_i)$ over all N rows:

$$H(P) = \frac{1}{N} \sum_{i=1}^{N} H(P_i). \tag{5}$$

When the attention probabilities in a given row P_i become overly concentrated on a single token, forming a near one-hot distribution, the attention entropy $H(P_i)$ approaches zero. If this occurs for all rows, the average attention entropy also collapses to zero, a phenomenon known as *attention entropy collapse*. This collapse is illustrated in the attention heatmaps in Appendix I.

4 Empirical Analysis of Attention Entropy Collapse and Training Instability

In this section, we present empirical results comparing softmax-based attention and Lipschitz-kernel-based attention, focusing on attention entropy collapse and the resulting training instability. First, in Section 4.1, we report and analyze empirical findings on attention entropy collapse and training instability observed in open-source LLMs, Llama (Touvron et al., 2023) and GPT2 (Radford et al., 2019). Furthermore, in Section 4.2, we conduct controlled experiments on a simple and small architecture composed solely of self-attention layers to isolate the effects of the re-weighting functions, ensuring that the influence of other factors is minimized.

4.1 Analysis on LLM Pre-training

Experimental Setup In this experiment, we pre-train a Llama1-1B model on a subset of the Pile dataset (Gao et al., 2020), consisting of up to 5B tokens. The model is trained with a sequence length of 768 and a batch size of 256. We use AdamW (Loshchilov, 2017) with a learning rate of 1e$-$3, following a cosine scheduling strategy. We train for $10,000$ steps with a weight decay of 0.1 and gradient clipping set to 1. Details on the GPT2-large pre-training setup are provided in Appendix C.

Experimental Result We observe that softmax-based attention (red solid line, Softmax) experiences a gradual decline in the average attention entropy over time, whereas its Lipschitz-kernel-based (blue dashed line, ReLU) approximation maintains a more stable attention entropy, as shown in Figure 1 (Top). As training progresses, this entropy reduction in softmax-based attention is accompanied by an increase in the Frobenius norm of the attention probability matrix. This increasing norm, in turn, leads to exploding gradient norms, further destabilizing training. In contrast, Lipschitz-kernel-based attention sustains relatively higher average attention entropy throughout training while maintaining lower attention probability matrix norms and gradient norms. Moreover, softmax-based attention converges to a higher training loss than Lipschitz-kernel-based attention, and the correspondingly higher validation loss further confirms its inferior generalization performance as detailed in Appendix A. We further conduct experiments on GPT2-large, whose results exhibit similar trends, as detailed in Appendix C.

Causal masking is known to mitigate attention entropy collapse by restricting attention to a limited context, thereby promoting more balanced attention probabilities (Zhai et al., 2023). However, our experimental results indicate that LLMs with softmax-based attention still tend to allocate excessive attention to specific tokens, ultimately leading to entropy collapse and training instability.

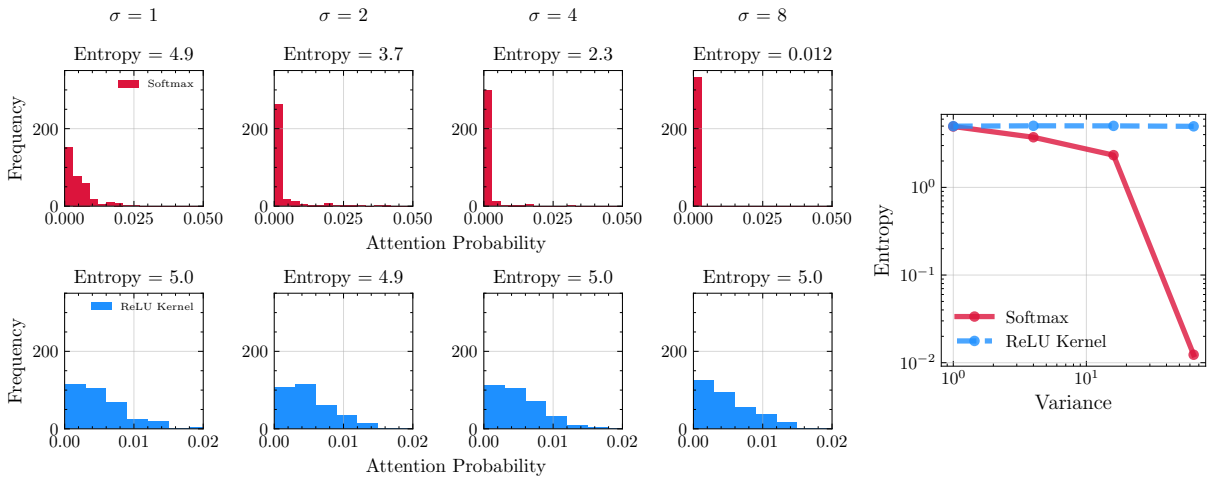

Figure 2: Comparison of the attention probability and attention entropy between softmax-based attention (Top) and Lipschitz-kernel-based attention (Bottom) as the attention logits variance increases. The lines (Rightmost) represent the rate of change (sensitivity) between softmax-based attention (red solid line; Softmax) and Lipschitz-kernel-based attention (blue dashed line; ReLU) as the attention logits variance increases. Here, with $N = 200$, the maximum achievable entropy is $\log N \approx 5.3$.

4.2 Analysis on Simple and Small Transformer

To further clarify the relationship between the Lipschitz continuity of re-weighting functions and attention entropy collapse, we conduct additional experiments in a simplified setting. This collapse is commonly attributed to factors such as model scale, hidden state dimensionality, layer stacking (Sun et al., 2024; He et al., 2024), and MLP layers (Cancedda, 2024). However, to disentangle the role of the re-weighting function from these other influences, we employ a simple and small-scale Transformer model and controlled task settings. Notably, we observe that attention entropy collapse can emerge even solely within attention layers, independent of the other factors, highlighting the fundamental role of the self-attention mechanism itself in driving this effect.

Experimental Setup For this experiment, we employ a simple Transformer architecture composed solely of self-attention layers. The model consists of 5-layers and a 3-dimensional hidden state ($L = 5, D = 3$) and a sequence length of 20 ($N = 20$). Our approach is motivated by findings that Transformers adapt to new tasks from only a few examples without parameter updates, a phenomenon known as in-context learning (Brown et al., 2020), spurring further research, (e.g., Garg et al. 2022; Zhang et al. 2024b; Mahankali et al. 2024; Von Oswald et al. 2023; Ahn et al. 2024). The simple Transformer is trained on an in-context linear regression task, predicting $w^\top x_{n+1}$ from $\{(x_i, y_i)\}_{i=1}^n$ and a query vector x_{n+1}, where (x_i, w) are sampled i.i.d. from $\mathcal{N}(0, I_D)$ and $y_i = w^\top x_i$. Additional implementation details are provided in Appendix D.

Experimental Result In Figure 1 (Bottom), we compare softmax-based attention (solid lines; Softmax, QK-LayerNorm) with Lipschitz-kernel-based attention (dashed lines; ReLU, ELU+1). The results are even more definitive than those observed in the LLMs experiments, as discussed in Section 4.1. Softmax-based attention rapidly collapses to the average attention entropy of zero early in training. At the same step, the gradient norm explodes, causing the loss to spike. Similarly, applying Layer Normalization to both the query and key before the softmax function fails to prevent attention entropy collapse, indicating that this normalization alone is insufficient. In contrast, the Lipschitz-kernel-based attention maintains higher average attention entropy, resulting in more stable training.

5 Why Lipschitz Kernels are Robust to Attention Entropy Collapse and Training Instability

Experimental results indicate that Lipschitz-kernel-based attention is more robust to attention entropy collapse than softmax-based attention, leading to more stable training. In this section, we analyze how the re-weighting function influences attention entropy and examine the causes of attention en-

tropy collapse along with its impact on training instability.

5.1 Scale Sensitivity with Softmax and Lipschitz Kernel

Based on the experiments, attention entropy collapse in self-attention heavily depends on the function used to re-weight the query–key dot product. The main cause is that re-weighting functions either amplify or confine differences between inputs as the input bound increases. To verify the causes and effects of these responses as the input bound increases, we scale the attention logits (query–key dot product). In softmax-based attention, scaling by a constant factor k results in computing $\exp(k \cdot Q_i K_j^\top)$ in both the numerator and denominator, which disproportionately amplifies larger attention logits while suppressing smaller ones due to the exponential growth of the function. Consequently, k-scaling increases the attention logits bound, causing probability mass to concentrate on a single dominant token. For Lipschitz-kernel-based attention, scaling the attention logits by k affects both the numerator and denominator proportionally, ensuring that attention logits remain within a bounded range, preventing attention entropy collapse.

5.2 Entropy Collapse Induced by Variance Sensitivity of Re-weighting Functions

In the previous section, we observe that scaling inputs with softmax-based attention amplifies differences, whereas Lipschitz-kernel-based attention confines these differences within a bounded range. Crucially, each re-weighting function exhibits a different sensitivity to the variance among inputs, and it is this sensitivity that has a major impact on attention entropy collapse. In softmax-based attention, high sensitivity causes the attention probabilities to sharpen excessively as variance increases, resulting in nearly one-hot-like vectors and a higher risk of attention entropy collapse. In contrast, Lipschitz-kernel-based attention exhibits lower sensitivity to input variance, bounding the effects of changes in both the numerator and denominator, thereby preserving balanced attention probabilities even as the *attention logits variance* (defined below) increases.

Definition 5.1 (Attention Logits Variance)**.** The attention logits variance for each row Z_i of the attention logits $Z \in \mathbb{R}^{N \times N}$ is defined as follows:

$$\mathrm{Var}_j[Z_{i,j}] = \frac{1}{N}\sum_{j=1}^{N}\left(Z_{i,j} - \frac{1}{N}\sum_{j'=1}^{N} Z_{i,j'}\right)^2. \tag{6}$$

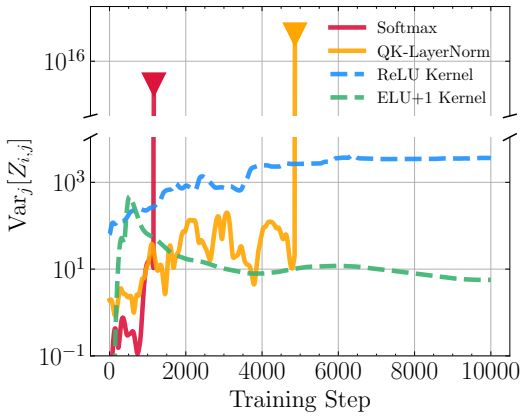

Figure 3: Changes in $\mathrm{Var}_j[Z_{i,j}]$ (attention logits variance, as defined in Definition 5.1) during training, comparing softmax-based attention (solid lines; Softmax, QK-LayerNorm) and Lipschitz-kernel-based attention (dashed lines; ReLU, ELU+1). Attention entropy collapse (▼) occurs as $\mathrm{Var}_j[Z_{i,j}]$ exponentially increases in softmax-based attention. This result is from an intermediate layer, with results from other layers provided in Appendix H.

Entropy Collapse in Controlled Experiment Due to Variance Sensitivity To examine how softmax-based and Lipschitz-kernel-based attention respond to attention logits variance, we control this variance with the unit-norm query and keys sampled from $\mathcal{N}(0, \sigma^2 I)$ at $\sigma = 1, 2, 4, 8$, so that the logit $Z_{i,j} = Q_i K_j^\top \sim \mathcal{N}(0, \sigma^2)$ has a variance of σ^2. Figure 2 presents histograms of the resulting attention weights for a single query (i.e., P_i for Q_i), illustrating how the distribution changes as σ increases. With softmax-based attention, as variance increases, the attention distribution becomes increasingly extreme, concentrating probability mass on a few key vectors and resulting in lower attention entropy. In contrast, Lipschitz-kernel-based attention (ReLU) maintains an attention entropy of around 5.0 as attention logits variance increases, preserving a more evenly distributed attention probability and avoiding entropy collapse. This trend is evident in the rightmost column, which confirms that softmax-based attention is highly sensitive to attention logits variance, exhibiting a steep rate of entropy change as variance increases. Conversely,

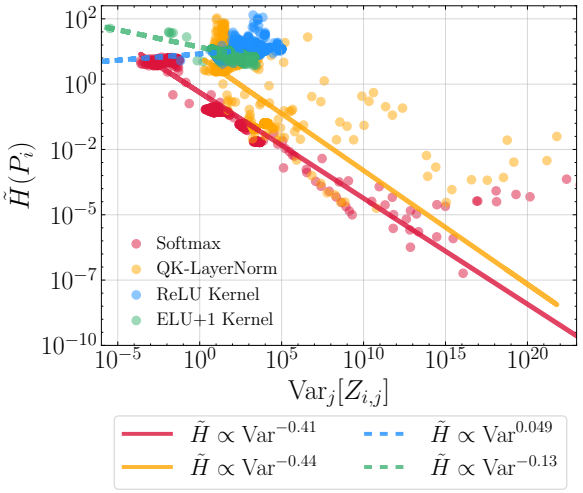

Figure 4: Variance-entropy sensitivity in softmax-based and Lipschitz-kernel-based attention during training. The lines represent the power-law relationship in softmax-based attention (solid lines; Softmax, QK-LayerNorm) and Lipschitz-kernel-based attention (dashed lines; ReLU, ELU+1), as defined in (8).

Lipschitz-kernel-based attention is much less sensitive to attention logits variance, exhibiting an almost flat rate of entropy change. A detailed analysis of the correlation between variance and entropy for a normal distribution is provided in Appendix E.

Entropy Collapse in Training Due to Variance-Entropy Sensitivity Through controlled experiments, we observe that as variance increases, softmax exhibits significantly higher sensitivity than the Lipschitz-kernel, leading to attention entropy collapse. Motivated by this finding, we investigate how attention logits variance changes during training, comparing softmax-based attention with Lipschitz-kernel-based attention. As illustrated in Figure 3, the attention logits variance in softmax-based attention layers grows sharply during training. This spike coincides with the emergence of attention entropy collapse, leading to unstable training.

Building on these observations, we analyze the sensitivity of softmax-based and Lipschitz-kernel-based attention to attention logits variance in relation to attention entropy. Before quantifying this sensitivity, we first define the normalized attention entropy as:

$$\tilde{H}(P_i) = \psi(H(P_i)) = \frac{H(P_i)}{H_{\max} - H(P_i)}, \quad (7)$$

where $H_{\max}$ denotes the maximum attention entropy, which equals $\log N$, and ψ is an increasing function of H. To quantify variance-entropy sensitivity, we assume the following power-law relationship:

$$\tilde{H}(P_i) \propto \mathrm{Var}_j[Z_{i,j}]^{\beta}. \quad (8)$$

Here, β represents the sensitivity at which normalized attention entropy changes in response to attention logits variance.

Figure 4 shows the relationship between normalized attention entropy ($\tilde{H}(P_i)$) and attention logits variance ($\mathrm{Var}_j[Z_{i,j}]$), along with the corresponding power-law exponents (β). For softmax-based attention, β takes on large negative values, meaning that even at the same variance, it results in lower entropy with a steep power-law, indicating high variance-entropy sensitivity. In contrast, Lipschitz-kernel-based attention has β close to zero, resulting in a much flatter power-law and lower variance-entropy sensitivity to attention logits variance.

5.3 Why Attention Entropy Collapse Leads to Training Instability

Attention entropy collapse is associated with unstable gradients, leading to loss spikes and severe training instability. In open-source LLMs training with softmax-based attention, we show that the average attention entropy progressively decreases, while the gradient norm steadily increases (see Figure 1 Top). In contrast, Lipschitz-kernel-based attention maintains higher entropy and stable gradients, preventing training instability. As shown in Figure 1 (Bottom, the second panel), despite being trained with shallow layers composed only of self-attention, the model still experiences gradient explosion, which can even make training entirely infeasible, suggesting a strong correlation between attention entropy collapse and gradient instability.

Entropy-Collapsed Attention Probabilities Explode Gradient The explosion of gradients, along with attention entropy collapse, is closely tied to the Lipschitz constant of self-attention. Specifically, the softmax function is the primary cause, as increases in the input bound or variance result in disproportionately large output changes, leading to an unbounded rate of change and a sharply elevated Lipschitz constant. Previous research has proposed alternative formulations that replace the softmax function in attention mechanisms to address these issues, such as L2 self-attention (Kim et al., 2021) and sigmoid self-attention (Ramapuram et al., 2025), which aim to

enforce a tighter upper bound on the Lipschitz constant.

According to (Dasoulas et al., 2021), the norm of the derivative of the self-attention layers with respect to the input X is upper bounded as follows:

$$\|\mathbf{D}A_X\|_{F,F} \leq \|P\|_F + \sqrt{2}\|X\|_{(2,\infty)} \|\mathbf{D}Z_X\|_{F,(2,\infty)}, \quad (9)$$

where $\|X\|_{(2,\infty)} = \max_j(\sum_i X_{i,j}^2)^{1/2}$ and $\|f\|_{a,b} = \max_{\|x\|_b=1} \|f(x)\|_a$. The attention probability matrix norm $\|P\|_F$ controls the upper bound in (9) and depends on whether the average attention entropy of P is low (one hot) or high (uniform).

Proposition 5.2. *The norm $\|P\|_F$ of the attention probability matrix P lies within the interval $[1, \sqrt{N}]$, attaining the extreme values as follows:*

$$\|P\|_F = \begin{cases} 1 & \text{if each row } P_i \text{ is uniform} \\ \sqrt{N} & \text{if each row } P_i \text{ is one-hot} \end{cases}. \quad (10)$$

On the contrary, the average attention entropy $H(P)$ lies within $[0, \log(N)]$, attaining the extreme values:

$$H(P) = \begin{cases} \log(N) & \text{if each row } P_i \text{ is uniform} \\ 0 & \text{if each row } P_i \text{ is one-hot} \end{cases}. \quad (11)$$

Figure 1 (Rightmost) illustrates how the attention probability matrix norms evolve for softmax-based and Lipschitz-kernel-based attention. At the beginning of training, both models have not yet learned the relevance between tokens in the input sequence. As a result, each row of P is nearly uniform, with a high attention entropy $H(P) \approx \log(N)$ from (11). This uniformity results in stable training dynamics, as indicated by a small Frobenius norm $\|P\|_F \approx 1$ from (10) in Proposition 5.2 and bounded gradients from (9). As training progresses with softmax-based attention, attention probabilities increasingly concentrate on a single token, forming nearly one-hot rows with near-zero attention entropy as described in (11). Consequently, $\|P\|_F$ increases toward $\sqrt{N}$, following (10), leading to larger gradients and increased training instability as indicated in (9). In contrast, Lipschitz-kernel-based attention maintains a significantly lower norm. Furthermore, the positive correlation between the gradient norm and $\|P\|_F$, as indicated by the bound in (9) is empirically validated in Appendix G.

6 Conclusion

In this paper, we identify the critical factor of attention entropy collapse (also known as the attention sink) that occurs during the training of attention-based models. Specifically, through both controlled experiments and real training settings, we demonstrate that softmax-based attention exhibits extremely high sensitivity to variance in attention logits, which serves as a primary factor in attention entropy collapse. In contrast, Lipschitz-kernel-based attention maintains low sensitivity, mitigating this issue. Furthermore, we connect attention entropy collapse to training instability by showing that the increasing norm of the attention probability matrix contributes to the growth of the gradient norm. As a result, these findings suggest that Lipschitz-kernel-based attention is advantageous for designing LLMs, enabling stable training and faster convergence with higher learning rates.

Limitations

Our analysis of attention entropy dynamics does not fully explore their impact on downstream task performance. Comparisons across model families and self-attention variants remain limited, leaving gaps in understanding their differences. The role of optimization choices, including schedules, warm-up strategies, weight decay, and gradient clipping, is not systematically examined. These factors likely influence model behavior and generalization, requiring deeper investigation. Future research should address these limitations to provide a more comprehensive perspective on entropy dynamics in attention-based models.

References

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. 2024. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yaroslav Aksenov, Nikita Balagansky, Sofia Lo Cicero Vaina, Boris Shaposhnikov, Alexey Gorbatovski, and Daniil Gavrilov. 2024. Linear transformers with learnable kernel functions are better in-context models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9584–9597, Bangkok, Thailand. Association for Computational Linguistics.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and

Christopher Ré. 2024. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. 2023. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17302–17313.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808, Bangkok, Thailand. Association for Computational Linguistics.

Yifan Chen, Qi Zeng, Heng Ji, and Yun Yang. 2021. Skyformer: Remodel self-attention with gaussian kernel and nyström method. *Advances in Neural Information Processing Systems*, 34:2122–2135.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

George Dasoulas, Kevin Scaman, and Aladin Virmaux. 2021. Lipschitz normalization for self-attention layers with application to graph neural networks. In *International Conference on Machine Learning*, pages 2456–2466. PMLR.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. 2023. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. 2024a. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. 2024b. Attention score is not all you need for token importance indicator in KV cache reduction: Value also matters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21158–21166, Miami, Florida, USA. Association for Computational Linguistics.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. 2023. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5961–5971.

Bobby He, Lorenzo Noci, Daniele Paliotta, Imanol Schlag, and Thomas Hofmann. 2024. Understanding and minimising outlier features in transformer training. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. 2020. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zixuan Jiang, Jiaqi Gu, and David Z Pan. 2023. Normsoftmax: Normalizing the input of softmax to accelerate and stabilize training. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6. IEEE.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Akhil Kedia, Mohd Abbas Zaidi, Sushil Khyalia, JungHo Jung, Harshith Goka, and Haejun Lee. 2024. Transformers get stable: an end-to-end signal propagation theory for language models. In *Proceedings of*

the 41st International Conference on Machine Learning, ICML'24. JMLR.org.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571. PMLR.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. 2024. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. 2022. cosformer: Rethinking softmax in attention. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russell Webb. 2025. Theory, analysis, and best practices for sigmoid self-attention. In *The Thirteenth International Conference on Learning Representations*.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. In *First Conference on Language Modeling*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie E. Everett, Alexander A. Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. 2024. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan (Celine) Lin. 2024. Unveiling and harnessing hidden attention sinks: enhancing large language models without training through attention calibration. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. 2023. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. 2024a. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2024b. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

A Llama Pre-training Validation Loss

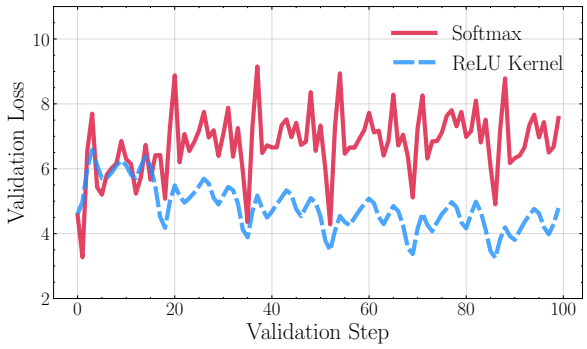

Figure 5: Comparison of validation loss between softmax-based attention (solid line) and Lipschitz-kernel-based attention (dashed line) during Llama1-1B pre-training. Validation loss is evaluated every 100 training steps, showing that Lipschitz-kernel-based attention consistently outperforms softmax-based attention.

B Layer-wise Attention Entropy

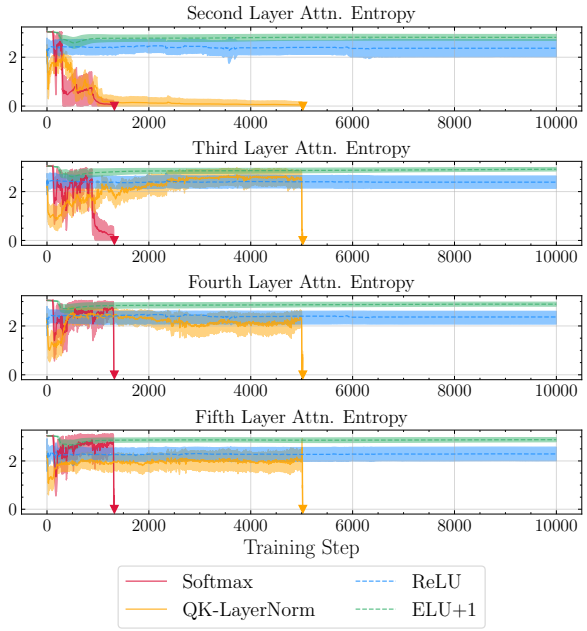

Figure 6: Average attention entropy behavior across layers in softmax-based (solid lines; Softmax, QK-LayerNorm) and Lipschitz-kernel-based attention. Softmax-based attention exhibits the average attention entropy collapse (▼) across all layers, while Lipschitz-kernel-based attention (dashed lines; ReLU, ELU+1) maintains high average attention entropy.

In Figure 1, we show that as the average attention entropy of the first self-attention layer gradually decreases, training instability increases. This result illustrates the dynamics of the average attention entropy across different layers.

C GPT2 Pre-training

We extend our experiments to GPT2-large in addition to the previously conducted Llama1-1B experiments. Figure 7 illustrates that, in softmax-based attention, average attention entropy gradually decreases in the early training steps, eventually approaching zero (the third panel). Almost simultaneously, $\|P\|_F$ increases (the fourth panel), and a sharp increase in gradient magnitude occurs (the second panel), reinforcing the direct relationship between entropy and training stability observed in previous experiments. In contrast, Lipschitz-kernel-based attention preserves higher entropy throughout training, exhibits smaller $\|P\|_F$, and stabilizes gradients.

D Implementation Details

Here are the hyper-parameters we used, and we apply the same ones across all experiments.

Table 1: Hyper-parameters of a Simple Transformer

Hyper-parameter	Value
Optimizer	SGD
Momentum	0.8
Learning rate	0.7
Hidden dimension	3
Sequence length	20
Attention heads	1
Attention layers	5
Training Step	10000

To approximate attention entropy collapse in large models and reproduce it in smaller models, we set a high learning rate of 0.7. As there is no notable difference between the SGD and Adam optimizers, we opt for SGD. The model is configured with a batch size of 4000. Given the small model size, we set it to 5 layers, 1 attention head, a sequence length of 20, and a hidden dimension of 5. To analyze gradient behavior without constraints, gradient clipping is disabled, and training runs for $10,000$ steps.

E Proof of Correlation between Variance and Entropy

If the distribution follows a normal distribution, we can define the probability density function (PDF) of the normal distribution $X \sim N(\mu, \sigma^2)$ for observation x:

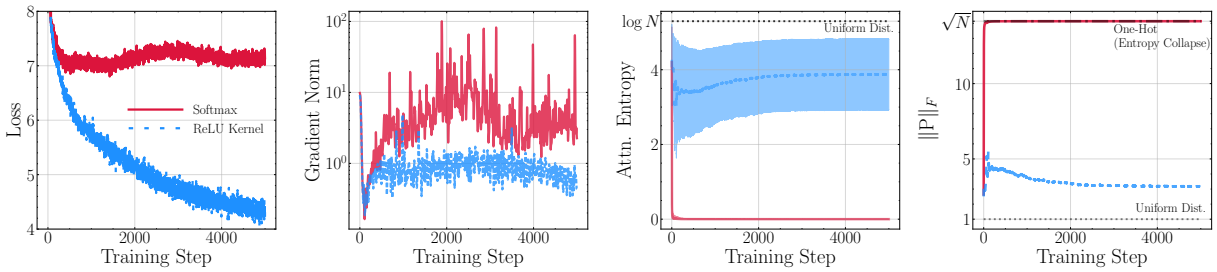

Figure 7: The training behaviors of GPT2-large ($N = 200$) with softmax-based attention (solid line; Softmax) and Lipschitz-kernel-based attention (dashed line; ReLU). From left to right, each panel shows the training loss (Loss), gradient norm (Gradient Norm), the first-layer average attention entropy with $\pm$ standard deviation (Attn. Entropy), and the average Frobenius norm of the attention probability matrix ($\|P\|_F$). In the third panel, as the attention probabilities of Lipschitz-kernel-based attention are nearly uniform, its average attention entropy reaches the maximum value (dotted line; $\log N$), whereas softmax-based attention exhibits an average attention entropy close to 0. In the fourth panel, while the softmax-based attention $\|P\|_F$ reaches its maximum value (dashed-dotted line; $\sqrt{N}$), the Lipschitz-kernel-based attention remains close to its minimum (dotted line) under a uniform attention distribution.

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right) \tag{12}$$

where μ is the mean and σ^2 is the variance of distribution.

Also, we can define the entropy of X as:

$$H(X) = -\int_{-\infty}^{\infty} g(x) \log g(x) dx \tag{13}$$

To compute logarithm of $g(x)$, we can use properties of it:

$$\log g(x) = \log\left(\frac{1}{2\pi\sigma^2}\right) \tag{14}$$

$$+ \log\left(\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \tag{15}$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}. \tag{16}$$

Then, we can calculate $H(X)$ with replacing $\log g(x)$ with (16):

$$H(X) = -\int_{-\infty}^{\infty} g(x) \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right) dx \tag{17}$$

We can separate two terms and the first term can be computed using $-\int_{-\infty}^{\infty} g(x)dx = 1$:

$$-\int_{-\infty}^{\infty} g(x)\left(-\frac{1}{2}\log(2\pi\sigma^2\right) = \frac{1}{2}\log(2\pi\sigma^2) \tag{18}$$

In normal distribution, with $-\int_{-\infty}^{\infty} f(x)(x-\mu)^2 dx = \sigma^2$ we can simplify the second term as:

$$-\frac{1}{2\sigma^2}\int_{-\infty}^{\infty} g(x)(x-\mu)^2 dx = -\frac{1}{2\sigma^2}\sigma^2 = -\frac{1}{2} \tag{19}$$

Therefore, we can define the entropy of normal distribution as:

$$H(X) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \tag{20}$$

F Skewness

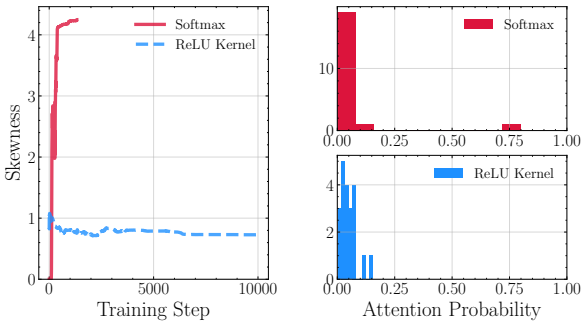

Figure 8: (Left) Skewness of softmax-based attention (solid line; Softmax) and Lipschitz-kernel-based attention (dashed line; ReLU) throughout training. (Right) Distribution of attention probabilities in softmax-based attention (Top) and Lipschitz-kernel-based attention (Bottom) at the point of highest skewness for each method.

After passing through the softmax function, values range between 0 and 1. When the distribution collapses onto a single value, attention entropy collapse occurs. If we consider each row of the attention probabilities as a probability distribution, this represents a highly imbalanced form. Skewness quantifies the degree to which an activation function skews a distribution. If one row vector of the attention probabilities matrix is denoted by $\{p_1, p_2, p_3, ...p_N\}$, with the mean μ and standard deviation σ, the skewness is defined as follows:

$$S = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{a_i - \mu^3}{\sigma} \right). \tag{21}$$

Based on Figure 8 (Left), we observe that the skewness of softmax-based attention rises sharply during training, approaching its maximum value. This indicates that the softmax function tends to learn highly imbalanced distributions, where most tokens attend primarily to a single other token. In Figure 8 (Top-Right), we observe that most values are concentrated around 0 and fall below the mean, demonstrating strong positive skewness. In contrast, Lipschitz-kernel-based attention exhibits relatively lower skewness values. Based on Figure 8 (Bottom-Right), the attention probability distribution is more evenly spread around the mean, indicating low positive skewness.

G Correlation Between Attention Entropy and Probabilities Norm

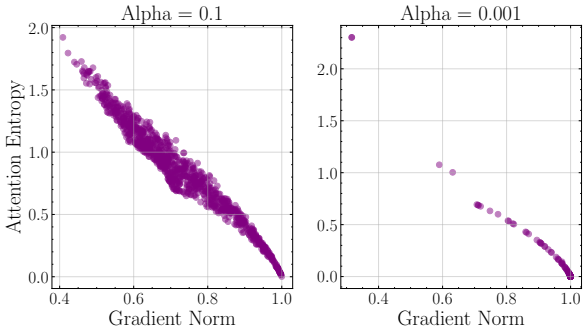

Figure 9: The correlation between the attention entropy and ℓ_2-norm of each row after sampling rows of attention probabilities from a Dirichlet distribution. For this setup, the concentration hyper-parameter α of the Dirichlet distribution is configured as 0.1 and 0.001 during sampling.

To show that as attention entropy decreases, the norm of attention probability matrix increases, we sample attention probability vectors from a Dirichlet distribution, defined as follows:

$$P_i \sim \text{Dirichlet}(\alpha \mathbf{1}) \tag{22}$$

The concentration of the distribution can be controlled using the hyper-parameter $\alpha\mathbf{1}$. When $\alpha\mathbf{1}$ is small, the distribution is concentrated on a single value, resembling attention entropy collapse. In contrast, when $\alpha\mathbf{1}$ is relatively large, the distribution becomes more uniform. Experimental results indicate that when $\alpha\mathbf{1} = 0.001$, attention entropy is significantly lower than at $\alpha\mathbf{1} = 0.1$. Furthermore, it is observed that the attention entropy of P_i and its ℓ_2-norm are inversely related. As attention entropy decreases, $\|P\|_F$ increases, reaching its maximum when attention entropy approaches zero.

H Layer-wise Attention Logits Variance

Based on Figure 10, across all layers, including the intermediate layers shown in Figure 3, we observe that softmax-based attention exhibits a sharp increase in attention logits variance at the step where attention entropy collapse occurs. This variance explosion becomes more pronounced in later layers. In contrast, Lipschitz-kernel-based attention maintains a stable attention logits variance across all layers, demonstrating its robustness to this collapse.

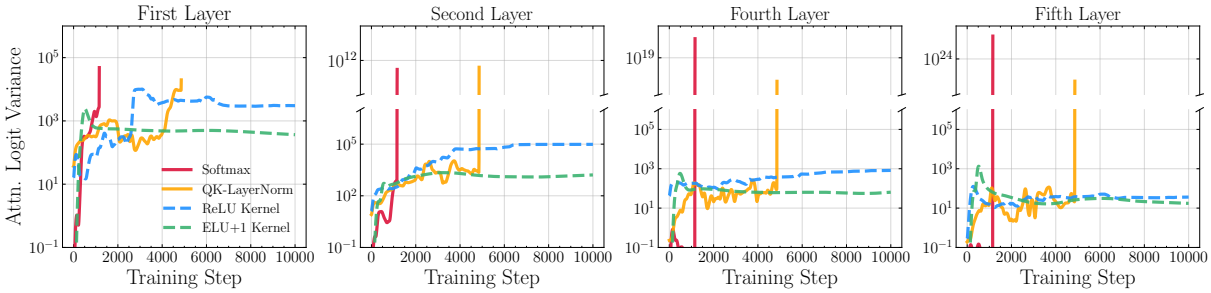

Figure 10: Attention logits variance across layers for softmax-based (solid lines; Softmax, QK-LayerNorm) and Lipschitz-kernel-based attention (dashed lines; ReLU, ELU+1). Softmax-based attention exhibits a sharp increase in variance at the step where attention entropy collapse occurs, with this effect becoming more pronounced in later layers. Lipschitz-kernel-based attention maintains a relatively stable variance across all layers, demonstrating robustness.

I Attention heatmaps

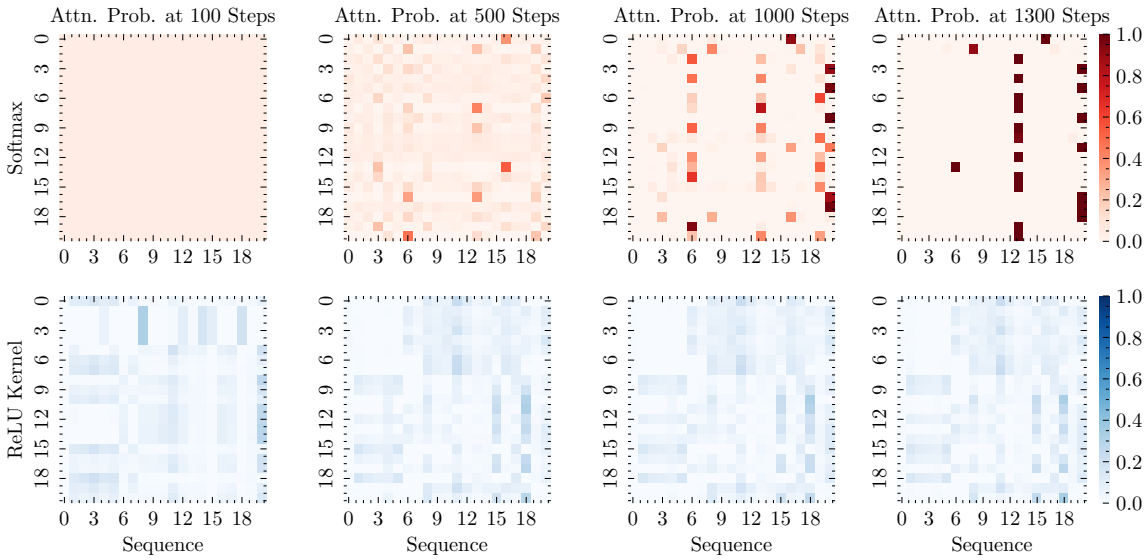

Figure 11: Heatmaps of attention probabilities for softmax-based attention (Top) and Lipschitz-kernel-based attention (Bottom) during training. In softmax-based attention, each row progressively converges to a one-hot-like vector, leading to attention entropy collapse. The attention matrices are from the first layer.