Variance Sensitivity Induces Attention Entropy Collapse in Transformers

Jonghyun Hong, Sungyoon Lee
Department of Computer Science, Hanyang University, Seoul, Republic of Korea
{jody1188, sungyoonlee}@hanyang.ac.kr

Abstract

Attention-based language models commonly rely on the softmax function to convert attention logits into probability distributions. However, this softmax re-weighting can lead to *attention entropy collapse*, in which attention disproportionately concentrates on a single token, ultimately causing training instability. In this work, we identify the high *variance sensitivity* of softmax as a primary cause of this collapse. We show that *entropy-stable* attention methods, which either control or are insensitive to the variance of attention logits, can prevent entropy collapse and enable more stable training. We provide empirical evidence of this effect in both large language models (LLMs) and a small Transformer model composed solely of self-attention and support our findings with theoretical analysis. Moreover, we identify that the concentration of attention probabilities increases the probability matrix norm, leading to the gradient exploding.

1 Introduction

Large language models (LLMs) rely on the attention mechanism, where attention logits (query–key dot products) are converted into probability distributions via the softmax function to capture the relative importance of tokens. 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., 2025, 2024; 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 (He et al., 2016), and MLP layers (Gu et al., 2025; Cancedda, 2024). However, there is still no clear theoretical understanding of why entropy collapse occurs.

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 a disproportionate concentration 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 near one-hot vectors and resulting in training instability.

We compare several attention methods and find that ReLU kernel attention (Choromanski et al., 2021; Qin et al., 2022) and QK-LayerNorm (Gilmer et al., 2023) maintain higher attention entropy and lead to more stable training than softmax-based attention, including Softmax and Window Softmax (Beltagy et al., 2020). Figure 1 illustrates this phenomenon in both open-source LLMs (top) and a simple, attention-only Transformer model (bottom). Specifically, softmax-based attention results in a progressive decrease in attention entropy (third column), which in turn increases the norm of the attention probability matrix (fourth column), leading to unstable gradients and loss spikes (second and first columns, respectively). In contrast, ReLU kernel attention and QK-LayerNorm preserve higher attention entropy and maintain lower norms of the attention matrix and gradients.

To better understand the distinct behaviors of attention methods, we analyze their *insensitivity* and *controllability* with respect to *attention logits variance*. Both theoretical and empirical evidence reveal that, in softmax, entropy decreases with increasing variance. This implies that higher variance results in significantly lower entropy, highlighting its high variance sensitivity. By contrast, our analysis shows that ReLU kernel attention is theoretically *entropy-stable*, as its entropy remains stable even when the variance of the input logits becomes large. We further provide an analysis of

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8372–8390
November 4-9, 2025

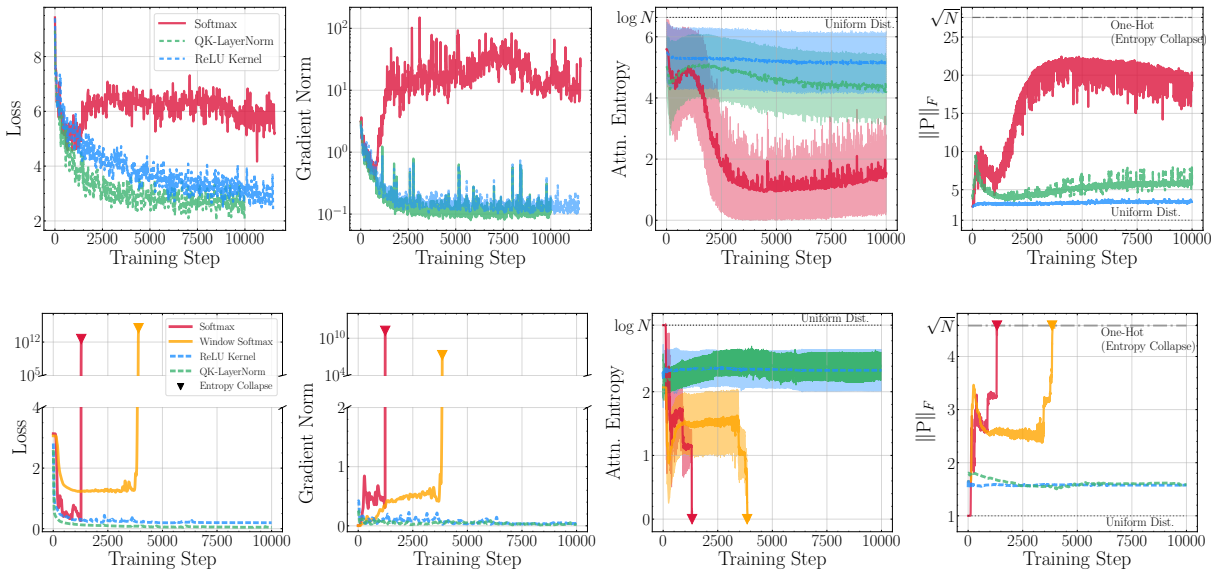

Figure 1: Training behaviors of LLaMA-1B (top, $N = 768$) and a small-scale Transformer model (bottom, $N = 20, W = 8$). From left to right, each column shows the training loss (Loss), gradient norm (Gradient Norm), the attention entropy with $\pm$ standard deviation across all layers (Attn. Entropy), and the average Frobenius norm of the attention probability matrix across all layers ($\|P\|_F$). In the third column, as the attention probability becomes uniform, the 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 5.3.

QK-LayerNorm, introduced to address the issue of large-magnitude attention logits, and show that it effectively controls variance and contributes to preserving attention entropy. However, we also find that, due to the presence of softmax, it remains sensitive to variance, and its behavior strongly depends on the setting of the scaling parameter.

Moreover, we provide a clear and focused analysis of the cause of training instability induced by attention entropy collapse. Several studies have investigated this cause, including softmax saturation and gradient exploding (Dehghani et al., 2023), sharp loss surfaces due to query–key spectral norm blow-up—addressed by the σReparam (Zhai et al., 2023), and outlier activations that disrupt gradient flow (He et al., 2024). However, the exact cause of the instability remains unclear. Our experiments, conducted across both large and small models, reveal a strong correlation between the decrease in attention entropy and spikes in the gradient norm. As shown in Figure 1 (second column), the gradient norm explodes at the point where the attention entropy decreases sharply or approaches zero during training (third column), indicating a direct relationship with instability. As attention probabilities become increasingly concentrated, the norm of the attention probability matrix, $\|P\|_F$, increases rapidly (fourth column), which in turn increases the gradient norm of self-attention output during backpropagation.

To summarize, we make the following contributions:

- We identify the variance sensitivity of the reweighting function as the cause of attention entropy collapse. Empirically, we show that attention methods less sensitive to attention logits variance can prevent this collapse and lead to more stable training, in both small and large models.

- We provide both theoretical and empirical evidence that softmax-based attention is highly sensitive to logit variance, whereas ReLU kernel attention remains *entropy-stable*. Furthermore, QK-LayerNorm offers variance controllability, but retains softmax-induced sensitivity that depends on the scaling parameter.

- We establish that a decrease in attention entropy increases the norm of the attention probability matrix, which increases the gradient norm of the attention output, ultimately leading to exploding gradients.

2 Related Works

Several studies have examined the concentration of probability on a single token, leading to attention entropy collapse. The large spectral norms tighten the lower bound of attention entropy, leading to sharper attention probability distributions, causing training instability (Zhai et al., 2023). As the sequence length grows, a log-scale increase in the top query-key score can cause softmax to disproportionately amplify it, concentrating attention on only a few tokens (Song et al., 2025). Furthermore, as the magnitude of attention logits increases, attention probabilities tend to collapse into near-one-hot vectors, thereby exacerbating training instability (Noci et al., 2023; Kedia et al., 2024). Various normalization methods have been proposed to alleviate the attention entropy collapse. Representative methods include QK-LayerNorm (Dehghani et al., 2023), QKNorm (Henry et al., 2020), Softmax-1 (adding 1 to the denominator) (Kaul et al., 2025; Miller, 2023), NormSoftmax (Jiang et al., 2023), and HybridNorm (Zhuo et al., 2025). This collapse is characterized by an excessive attention bias toward initial tokens (Barbero et al., 2025), commonly referred to as attention sink (Xiao et al., 2024). Large activations in a few units concentrate attention on their associated 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 attention sink by causing value-state drains (Guo et al., 2025). Recent work replaces softmax with unnormalized sigmoid attention to mitigate attention sink and improve training stability (Fu et al., 2025). While prior works focus on attention logit scale, we focus on the *sensitivity to the attention logits variance*.

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 —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 $P \in \mathbb{R}^{N\times N}$ are defined as follows:

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

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

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

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

Window Softmax Attention In window attention, each query at position i attends only to the keys within a fixed window from K_{i-W} to K_{i+W}, where W is the window size. Accordingly, the attention probability in (1) is replaced with:

$$P^{\text{W}}_{i,j} = \frac{\exp(Q_i K_j^\top)}{\sum_{k=i-W}^{i+W}\exp(Q_i K_k^\top)}.$$

Restricting attention to a local window prevents excessive focus on a single token and promotes relatively uniform attention probabilities (Dong et al., 2024; Gu et al., 2025).

3.2 Query-Key Normalization (Gilmer et al., 2023)

To alleviate large attention logits, which can lead to the concentration of attention on a single token, Gilmer et al. (2023) apply Layer Normalization (LN) (Ba et al., 2016) to both Q and K before the dot product, modifying the attention formulation in (1). We define the normalized attention logits of QK-LayerNorm as

$$Z^{\text{LN}}_{i,j} = \text{LN}(Q_i)\text{LN}(K_j)^\top, \quad (2)$$

and compute the attention probability as

$$P^{\text{LN}}_{i,j} = \frac{\exp(Z^{\text{LN}}_{i,j})}{\sum_{k=1}^{N}\exp(Z^{\text{LN}}_{i,k})}. \quad (3)$$

By enhancing training stability (Rybakov et al., 2024), QK-LayerNorm has become a widely used component in many recent LLMs.

3.3 Linear Kernelized Attention

To mitigate the quadratic complexity of traditional attention methods, 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:

$$\text{sim}(Q_i, K_j) \approx \phi(Q_i)\phi(K_j)^\top. \tag{4}$$

Instead of applying softmax directly, kernelized self-attention uses a kernel function ϕ to approximate similarity. By exploiting the associativity of matrix multiplication, it avoids explicit computation of the attention matrix and reduces the quadratic time complexity to linear, as follows:

$$\begin{aligned} A_i^\phi &= \frac{\phi(Q_i)\sum_{j=1}^N \phi(K_j)^\top V_j}{\phi(Q_i)\sum_{k=1}^N \phi(K_k)^\top} \text{ and} \\ P_{i,j}^\phi &= \frac{\phi(Q_i)\phi(K_j)^\top}{\sum_{k=1}^N \phi(Q_i)\phi(K_k)^\top}. \end{aligned} \tag{5}$$

While prior works on kernelized attention mainly focus on choosing kernel functions that better approximate softmax attention such as ReLU (with re-weighting) (Qin et al., 2022; Cai et al., 2023; Han et al., 2023) and ELU+1 (Katharopoulos et al., 2020), our work instead examines kernel function from the perspective of training stability.

In particular, we focus on Lipschitz-continuous kernel functions, which bound the rate of change during the re-weighting from attention logits to probabilities. We use ReLU, ELU+1, and Sigmoid as Lipschitz-continuous kernel functions, which are widely used and ensure non-negative values.

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:

$$H(P_i) = -\sum_{j=1}^N P_{i,j} \log P_{i,j}. \tag{6}$$

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

$$H(P) = \frac{1}{N}\sum_{i=1}^N H(P_i). \tag{7}$$

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 attention entropy also collapses to zero, a phenomenon known as *attention entropy collapse*. This collapse is illustrated in the attention heatmaps in Appendix G.

Although attention sink is a phenomenon similar to attention entropy collapse, it differs in how the statistics are aggregated. The sink metric computes attention column-wise over K, thereby discarding row-wise information (Gu et al., 2025). By contrast, attention entropy is computed for each Q_i over K_j in the row distribution P_i (the dimension along which softmax is applied), and thus it more faithfully captures how the entropy of P_i responds to variance of $Z_{i,j}$ across K_j.

4 Empirical Analysis of Attention Entropy Collapse and Training Instability

In this section, we empirically compare softmax-based and *entropy-stable* attention, focusing on attention entropy collapse leading to 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 GPT-2 (Radford et al., 2019). Furthermore, in Section 4.2, we conduct experiments on a simple regression task using 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. The experimental settings for both experiments are detailed in Appendix C.

4.1 LLM Pre-training

Experimental Results We observe that softmax attention experiences a progressive decrease in attention entropy over time, whereas ReLU kernel attention and QK-LayerNorm maintain a more stable entropy, as shown in Figure 1 (top). As training progresses, this decrease in entropy for softmax attention is accompanied by an increase in the Frobenius norm of the attention probability matrix, which, in turn, increases gradient norms and, ultimately, causes the loss to diverge. In contrast, ReLU kernel attention and QK-LayerNorm maintain relatively higher attention entropy throughout training while keeping the attention probability matrix norms and gradient norms lower. Moreover, softmax attention converges to a higher training loss than these methods. We further conduct experiments on GPT-2 pre-training, which exhibit similar trends, as detailed in Appendix B.

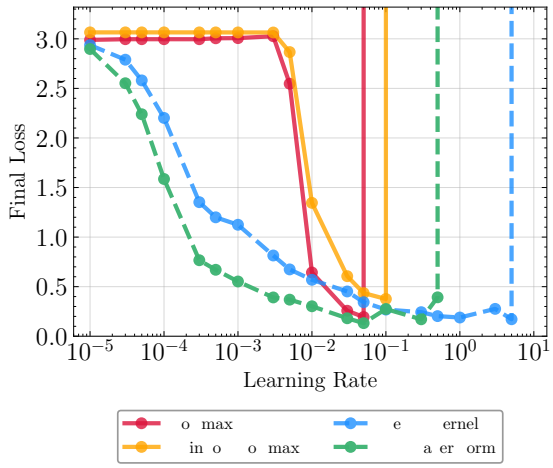

Figure 2: The comparison of training stability with different re-weighting functions is conducted by analyzing the variation in final loss across different learning rates. For each learning rate, the average final loss is computed over five independent runs, comparing softmax-based attention (solid lines; Softmax, Window Softmax) with entropy-stable attention (dashed lines; QK-LayerNorm, ReLU).

Re-Weighting Function	LR Sensitivity
Softmax (Vaswani et al., 2017)	2.30
Window Softmax (Beltagy et al., 2020)	2.20
σReparam (Zhai et al., 2023)	2.18
Sigmoid Kernel	1.97
ELU+1 Kernel (Katharopoulos et al., 2020)	1.95
QK-LayerNorm (Gilmer et al., 2023)	1.14
ReLU Kernel	**1.03**

Table 1: LR sensitivity for various re-weighting functions. It measures the rate of change of final loss with respect to the learning rate. Lower LR sensitivity indicates more stable training.

4.2 Simple and Small Transformer

To further clarify the relationship between the re-weighting functions in attention 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 small Transformer model composed solely of self-attention layers, trained on a simple regression task. Notably, we observe that attention entropy collapse can emerge independently of the other factors, highlighting the fundamental role of the self-attention itself in driving this effect.

Experimental Results The results are even more definitive than those observed in the LLMs experiments, as discussed in Section 4.1. In Figure 1 (bottom), softmax-based attention (solid lines; Softmax, Window Softmax) rapidly collapses to the attention entropy of zero early in training. At the same step, the gradient norm explodes, causing the loss to spike. In contrast, ReLU kernel attention (blue dashed line) and QK-LayerNorm (green dashed line) maintain higher attention entropy, resulting in more stable training. Additional results for other attention variants, including Sigmoid kernel, ELU+1 kernel attention and σReparam, are provided in Appendix A.

4.3 Comparative Analysis of Stability

Across both large and small models, softmax-based attention exhibits attention entropy collapse and training instability, whereas ReLU kernel attention and QK-LayerNorm remain stable. To compare the training stability of these attention methods, we use learning rate sensitivity (LR sensitivity), which quantifies the deviation of the final loss from the optimum across a wide range of learning rates (Wortsman et al., 2024). LR sensitivity is defined as $\mathbb{E}_{\eta\in[a,b]}\left[\min\left(\ell(\mathcal{A}(\eta)),\ell_0\right)-\ell^*\right]$, where $[a,b]$ is the learning rate range, ℓ^* is the loss achieved with the optimal learning rate, ℓ_0 is the loss at initialization, and $\theta=\mathcal{A}(\eta)$ denotes the model parameters obtained after training with learning rate η. We sweep learning rates $\mathrm{lr}\in\{1,\,3,\,5\}\times 10^k$ with $k=-5,-4,\ldots,1$ and $\mathrm{lr}\le 10$, training small-scale models using SGD and reporting results for each re-weighting function as the average over five runs per learning rate.

Experimental Results Figure 2 shows how the final training loss of different attention methods changes over a broad range of learning rates, with a summary in Table 1. ReLU kernel attention (blue dashed lines) achieves the widest stable range and the lowest sensitivity, maintaining low final loss across nearly five orders of magnitude. QK-LayerNorm (green dashed line) also demonstrates strong robustness, with stability comparable to ReLU kernel. In contrast, softmax-based methods (solid lines; Softmax and Window Softmax) remain stable only in a narrow range but exhibit the highest LR sensitivity. Among other attention methods, σReparam remains relatively high, whereas

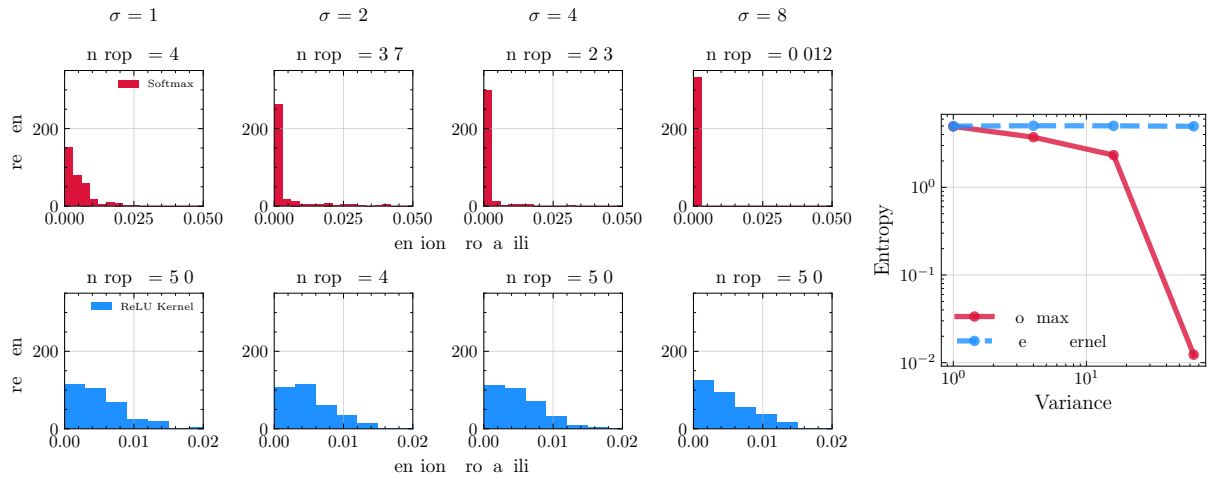

Figure 3: Comparison of attention probabilities and attention entropy between softmax-based attention (top) and ReLU kernel attention (bottom) as the attention logits variance increases. The lines (rightmost) represent the rate of change (variance sensitivity) between softmax-based attention (red solid line; Softmax) and ReLU kernel attention (blue dashed line) as the attention logits variance increases. Here, with $N = 200$, the maximum achievable entropy is $\log N \approx 5.3$.

ELU+1 and Sigmoid are lower (see Appendix A for the corresponding curves).

5 Why Attention Entropy Collapse Emerges and Causes Training Instability

Empirical results show that ReLU kernel attention and QK-LayerNorm avoid attention entropy collapse and enable more stable training than softmax-based attention. This section provides both theoretical insights and experimental analysis to explain the reasons behind this behavior. Furthermore, it demonstrates that the attention entropy collapse increases the gradient norm, leading to training instability.

5.1 Variance Sensitivity Induces Attention Entropy Collapse

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. Softmax-based attention tends to cause entropy collapse because the exponential function excessively amplifies differences in input values as variance increases. As a result, the softmax disproportionately emphasizes larger inputs while suppressing smaller ones. Window attention applies softmax only within local windows rather than across the entire sequence of length N. This local-window restriction prevents any single token from being repeatedly attended to across the entire sequence, which helps limit excessive focus on a single token. However, as demonstrated in previous experiments, attention entropy still tends to decrease or even collapse. Therefore, using re-weighting functions that have low sensitivity and are less affected by logits variance, such as ReLU Kernel, or applying methods like QK-LayerNorm that normalize the variance, can help maintain higher attention entropy and enable stable training.

Theorem 5.1 (Sensitivity of Softmax and ReLU Entropy on Variance)**.** *Let $z \sim \mathcal{N}(0, \sigma^2 I_N)$, $p = softmax(z)$ and $H(p) = -\sum_{i=1}^{N} p_i \log p_i$. Then, for small σ^2,*

$$H(p) = \log N - (N-1)\sigma^2/2N + \mathcal{O}(\sigma^4)$$

and the derivative of $H(p)$ with respect to σ^2 is

$$\frac{\partial H}{\partial \sigma^2} = -\mathbb{E}_z\left[\sum\nolimits_i z_i^2 \cdot p_i\right] < 0.$$

Thus, $H(p)$ is strictly decreasing in σ^2.

By contrast, the entropy of the ReLU kernel attention probability $\tilde{p}$ is given by

$$H(\tilde{p}) = \log N - \mathcal{O}\Big(1/D\Big)$$

and it does not depend on the variance σ^2.

The entropy of the softmax distribution decreases from the maximum value of $\log N$ as σ^2

increases, highlighting softmax's high sensitivity to logits variance and its tendency toward entropy collapse. Notably, the softmax logit scaling by $1/\sqrt{D}$ (used in LLaMA/GPT-2 and omitted in our small model) does not control attention logits variance and therefore neither reduces the variance sensitivity in Theorem 5.1 nor prevents attention entropy collapse. In contrast, the entropy of the ReLU kernel attention distribution remains approximately $\log N$ up to a small correction $\mathcal{O}(1/D)$, and is notably independent of logits variance. The detailed proof is provided in Appendix E.

Variance Controllability with QK-LayerNorm As shown in both Figure 1 and 2, QK-LayerNorm maintains high attention entropy and exhibits stable training. This illustrates how QK-LayerNorm effectively controls the variance of the attention logits in softmax-based attention. Moreover, when the LN scaling parameter γ is bounded, QK-LayerNorm becomes robust to shifts in logits variance, thereby ensuring stable attention behavior during training. Let the inputs be scaled as $Q_i = \sigma_q Q_i$, $K_j = \sigma_k K_j$, with arbitrary scaling factors $\sigma_q, \sigma_k > 0$. Since scaling a vector scales both its norm and variance proportionally, the effect of these scale factors cancels out after LayerNorm is applied, resulting in the normalized attention logits defined in (2) that are invariant to logits variance. Both the attention probability of QK-LayerNorm P_{ij}^{LN} defined in (3) and its entropy depend only on the normalized logits and therefore the attention entropy is invariant to query and key variance, i.e., $\frac{\partial H(P_i)}{\partial \sigma_q^2} = \frac{\partial H(P_i)}{\partial \sigma_k^2} = 0$. However, if the scaling parameters γ_q and γ_k are not bounded, attention entropy may collapse, as detailed in Appendix D.

Controlled Experiment Theoretical analysis demonstrates that the entropy of the softmax function decreases as variance increases, indicating high sensitivity. Unlike softmax, ReLU kernel attention entropy does not depend on the attention logits variance. To provide empirical evidence for the theoretical analysis, we analyze the sensitivity of various re-weighting functions to the *attention logits variance* (defined below).

Definition 5.2 (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 the empirical variance $\text{Var}(\{Z_{i,1}, Z_{i,2}, \cdots, Z_{i,N}\})$.

To examine how softmax-based and entropy-stable 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 3 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 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, ReLU kernel attention maintains an attention entropy of around 5.0 slightly below $\log N$ regardless of the value of the attention logits variance, preserving a more evenly distributed attention probability and avoiding entropy collapse. This trend is evident in Figure 3 (rightmost), confirming that softmax attention is highly sensitive to attention logits variance, with entropy changing steeply as variance increases. In contrast, ReLU kernel attention shows low sensitivity, exhibiting an almost flat rate of entropy change.

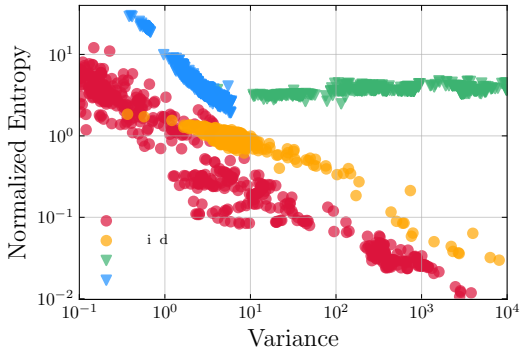

Figure 4: Relationship between attention logits variance and normalized attention entropy defined in (8) during training, across different attention methods. Softmax-based methods (•; Softmax, Window Softmax) and entropy-stable methods (▼; QK-LayerNorm, ReLU kernel) are included for comparison.

Practical Experiment Following controlled experiments, we analyze the relationship between the attention logits variance and entropy of softmax-based and entropy-stable attention methods during training. We define the normalized attention entropy as:

$$\tilde{H}(P_i) = \psi(H(P_i)) = \frac{H(P_i)}{H_{\text{max}} - H(P_i)}, \quad (8)$$

where H_{max} denotes the maximum attention entropy, which equals $\log N$. Note that ψ is increasing in H.

Figure 4 illustrates the relationship between attention logits variance and normalized attention entropy ($\tilde{H}(P_i)$) across different attention methods at each training step. Softmax-based attention exhibits a progressive decrease in entropy as the attention logits variance increases. In contrast, ReLU kernel attention maintains stable attention entropy even as attention logits variance increases, indicating low sensitivity to variance. Even at the same variance level, softmax-based attention produces significantly lower entropy. Notably, QK-LayerNorm shows a trend similar to that of Softmax, but it prevents a sharp drop in entropy by controlling the magnitude of the attention logits variance. On the other hand, Window Softmax exhibits a relatively flatter trend compared to Softmax. Since the variance sensitivity of softmax grows with sequence length N, using a smaller window W slightly reduces the sensitivity to variance, but is not sufficient to mitigate entropy collapse.

5.2 Attention Entropy Collapse Leads to Training Instability

Attention entropy collapse is associated with unstable gradients, leading to loss spikes and training instability. In open-source LLMs pre-training with softmax-based attention, we show that the attention entropy progressively decreases, while the gradient norm steadily increases (see Figure 1 top). In contrast, ReLU kernel attention and QK-LayerNorm maintain 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.

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, because increases in the logits bound or variance result in disproportionately large output changes, leading to an unbounded rate of change and a significantly elevated Lipschitz constant, consistent with prior results that standard dot-product self-attention is not globally Lipschitz. Previous research shows that softmax attention lacks a finite global Lipschitz constant (Kim et al., 2021; Khromov and Singh, 2024). To address this, prior work replaces softmax with alternatives such as L2 self-attention (Kim et al., 2021) and sigmoid self-attention (Ramapuram et al., 2025), which aim to enforce tighter Lipschitz upper bounds.

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:

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

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 attention entropy of P is low (one-hot) or high (uniform).

Proposition 5.3. *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 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 entropy-stable 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 5.3 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, entropy-stable 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 F.

6 Conclusion

In this paper, we identify high variance sensitivity and lack of control in softmax attention as key factors behind attention entropy collapse, as observed even in a model composed solely of self-attention layers. We also provide theoretical and empirical evidence that entropy-stable attention methods, which are either insensitive to or explicitly control attention logits variance, can maintain attention entropy and enable stable training. Furthermore, we link attention entropy collapse to training instability by showing that increased attention matrix norm leads to gradient exploding.

Limitations

Our analysis does not comprehensively cover a wide range of model architectures or self-attention variants, which limits the generality of our findings. Moreover, limited computational resources prevent evaluation of larger-scale models comparable to those used in practice. It remains important to investigate how full attention in encoders and causal attention in decoders differ in their sensitivity to, or ability to control, the attention logits variance in the re-weighting process. Furthermore, additional analysis is needed on training-related factors such as learning rate schedules, warm-up strategies, weight decay, and gradient clipping, which may also influence training stability.

Acknowledgments

This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants (RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University); RS-2023-002206284, Artificial intelligence for prediction of structure-based protein interaction reflecting physicochemical principles) and the National Research Foundation of Korea (NRF) grants (RS-2023-00244896, Implicit bias of optimization algorithms for robust generalization of deep learning; the BK21 FOUR (Fostering Outstanding Universities for Research) project; NRF-2024S1A5C3A02043653, Socio-Technological Solutions for Bridging the AI Divide: A Blockchain and Federated Learning-Based AI Training Data Platform) funded by the Korean government (MSIT). Finally, this paper was written with limited assistance from ChatGPT for language polishing, spell-checking, and clarity of proofs. All outputs were reviewed, edited, and verified by the authors, who take full responsibility for the content.

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.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. 2025. Why do LLMs attend to the first token? In *Second Conference on Language Modeling*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In *Findings of EMNLP*.

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.

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.

Zican Dong, Junyi Li, Xin Men, Wayne Xin Zhao, Bingning Wang, Zhen Tian, Weipeng Chen, and Ji-Rong Wen. 2024. Exploring context window of large language models via decomposed positional vectors. In *Advances in Neural Information Processing Systems*, volume 37, pages 10320–10347. Curran Associates, Inc.

Zichuan Fu, Wentao Song, Yejing Wang, Xian Wu, Yefeng Zheng, Yingying Zhang, Derong Xu, Xuetao Wei, Tong Xu, and Xiangyu Zhao. 2025. Sliding window attention training for efficient large language models. *CoRR*, abs/2502.18845.

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.

Justin Gilmer, Andrea Schioppa, and Jeremy Cohen. 2023. Intriguing properties of transformer training instabilities. *To appear*.

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. 2025. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. In *The Second Conference on Parsimony and Learning (Recent Spotlight Track)*.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. 2024. 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*.

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 (CVPR)*, pages 770–778.

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.

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.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. 2025. From attention to activation: Unraveling the enigmas of large language models. In *The Thirteenth International Conference on Learning Representations*.

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.

Grigory Khromov and Sidak Pal Singh. 2024. Some fundamental aspects about lipschitz continuity of neural networks. In *The Twelfth International Conference on Learning Representations*.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Evan Miller. 2023. Attention is off by one. `https://www.evanmiller.org/attention-is-off-by-one.html`. Blog post.

Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. 2023. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36:54250–54281.

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*.

Oleg Rybakov, Mike Chrzanowski, Peter Dykas, Jinze Xue, and Ben Lanir. 2024. Methods of improving llm training stability. *arXiv preprint arXiv:2410.16682*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Zhao Song, Jing Xiong, and Chiwun Yang. 2025. How sparse attention approximates exact attention?your attention is naturally n^c-sparse. In *Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference*.

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

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*.

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.

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.

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.

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.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Zhijian Zhuo, Yutao Zeng, Ya Wang, Sijun Zhang, Jian Yang, Xiaoqing Li, Xun Zhou, and Jinwen Ma. 2025. Hybridnorm: Towards stable and efficient transformer training via hybrid normalization. *CoRR*, abs/2503.04598.

A Additional Experiments on Attention Variants

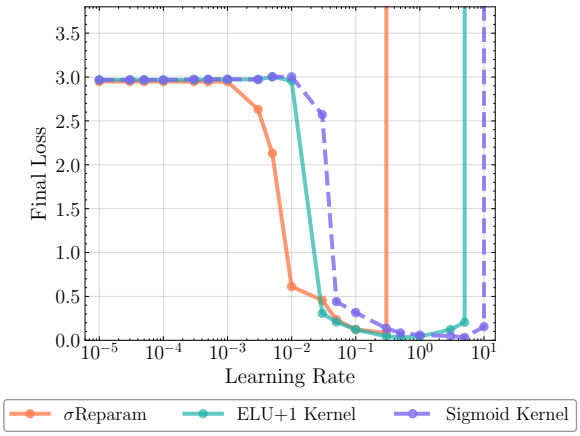

Figure 5: Average final training loss over five independent runs for ELU+1 kernel, Sigmoid kernel, and σReparam method across a range of learning rates.

We additionally experiment with kernelized self-attention with ϕ set to ELU+1 or Sigmoid, as well as σReparam (Zhai et al., 2023), a reparameterization method that normalizes weight matrices by their spectral norm. In this simple and small Transformer setup, which consists solely of self-attention layers, σReparam is applied only to the Q, K, and V projection weights. As shown in Figure 6, ELU+1 and Sigmoid kernel attention maintain stable training with high attention entropy, whereas σReparam undergoes entropy collapse, leading to unstable training. Although σReparam constrains the operator norms and limits the scale of the attention logits, it does not reduce the variance–entropy sensitivity of softmax and thus cannot fully prevent attention entropy collapse. Figure 5 further shows that σReparam has LR sensitivity comparable to softmax attention, which results from its dependence on softmax.

B Analysis on GPT-2 Pretraining

To validate our findings on larger models, we further extend our experiments to GPT-2 Large (774M parameters) with training on WikiText-103 (Merity et al., 2017), beyond the previously reported LLaMA-1B results. Figure 7 illustrates that, in softmax attention, attention entropy gradually decreases in the early training steps, eventually approaching zero (the third panel). Concurrently, $\|P\|_F$ increases (the fourth panel), and the gradient magnitude sharply increases (the second panel), reinforcing the direct relationship between entropy and training stability observed in previous experiments. In contrast, entropy-stable attention mitigates instability, thereby preserving higher entropy, maintaining smaller $\|P\|_F$, and stabilizing gradients.

C Experimental Setups

We specify the hyper-parameters for the large-scale pretraining setup and the small and simple Transformer setup for simple regression.

C.1 LLM-Pretraining Experimental Setup

We pre-train a 1B-parameter LLaMA model on a subset of the Pile dataset ($\approx$ 5B) (Gao et al., 2020), with rotary positional embeddings (RoPE) (Su et al., 2024), a pre-norm structure with RMSNorm (Zhang and Sennrich, 2019), a SwiGLU activation (Shazeer, 2020) in MLP. The model is trained with a sequence length of 768 and a batch size of 256. We use AdamW (Loshchilov and Hutter, 2019) with a learning rate of 1e$-$3, following a cosine scheduling strategy. We train for 10,000 steps ($\approx$ 2B tokens in total) with a weight decay of 0.1 and gradient clipping set to 1.

C.2 Simple and Small Transformer Experimental Setup

For the small-scale regression setup, a simple Transformer architecture composed solely of self-attention layers is employed. The model has 5 layers with a 3-dimensional hidden state ($L = 5, D = 3$) and a sequence length of 20 ($N = 20$). The attention window size is set to 8, yielding the most stable training dynamics and fixed across all experiments. This setup is motivated by findings that Transformers can adapt to new tasks from only a few examples without parameter updates—a phenomenon known as in-context learning (Brown et al., 2020), which has spurred further research (e.g., Garg et al. 2022; Zhang et al. 2024; 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 $y_i = w^\top x_i$. Furthermore, we evaluate a broader set of re-weighting functions, including Sigmoid kernel, ELU+1 kernel attention and σReparam.

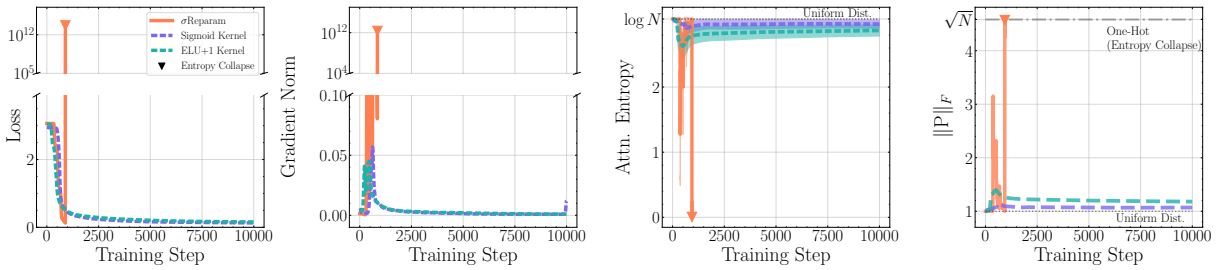

Figure 6: Training behaviors of ELU+1, Sigmoid kernel attention (dotted lines) and σReparam (solid line). The experiments are conducted in a simple and small Transformer, and the figure includes training loss, gradient norm, attention entropy (with $\pm$ standard deviation across all layers), and the average Frobenius norm of the attention probability matrix.

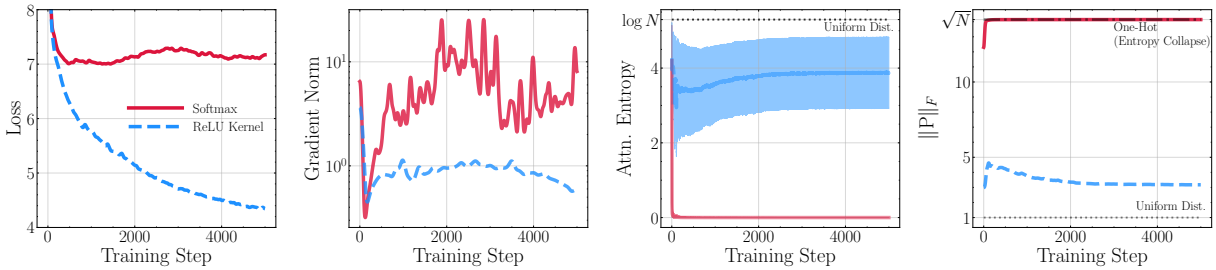

Figure 7: Training behaviors of GPT-2 ($N = 200$) with softmax-based attention (solid line; Softmax) and entropy-stable attention (dashed line; ReLU). From left to right, each panel shows the training loss (Loss), gradient norm (Gradient Norm), the first-layer attention entropy with $\pm$ standard deviation (Attn. Entropy), and the average Frobenius norm of the attention probability matrix ($\|P\|_F$). In the third panel, as the attention probabilities of entropy-stable attention are nearly uniform, its attention entropy reaches the maximum value (dotted line; $\log N$), whereas softmax-based attention exhibits an 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 entropy-stable attention remains close to its minimum (dotted line) under a uniform attention distribution.

D Ablation Study on QK-LayerNorm

Figure 9 presents an ablation study that empirically analyzes the role of the scale parameters γ_q and γ_k in controlling logit variance and preventing attention entropy collapse, comparing four strategies: Gradient Clipping, No Clipping, Fixed $\gamma = 1$, and Weight Clipping. Gradient clipping (top row) does not fully control the norms, leading to significant variation across layers. In layers where $\|\gamma_q\| \cdot \|\gamma_k\|$ becomes large, we observe increased attention logits variance and decreased attention entropy. Without any clipping (second row), the scale parameters grow rapidly and without bound in some layers, resulting in an increase in attention logits variance and diminished attention entropy. Fixing γ_q and γ_k to 1 (third row) enforces a constant attention scale during training, effectively controlling attention logits variance and resulting in higher attention entropy. Weight clipping (bottom row) also constrains the growth of the scale parameters and helps regulate attention behavior, though it exhibits minor variations. These empirical results indicate that QK-LayerNorm can control attention logits variance, thereby improving stability, although this benefit depends critically on the behavior of the scale parameters γ_q and γ_k.

E Proof of Theorem 5.1

E.1 Entropy Approximation for Softmax Version 1

Let $z = (z_1, z_2, ..., z_N) \in \mathbb{R}^N$ be a random vector such that $z_i \sim \mathcal{N}(0, \sigma^2)$ independently. Define the softmax vector $p = \text{softmax}(z)$, where

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{N} \exp z_j}. \tag{12}$$

The entropy of the softmax distribution is given by

$$H(p) = -\sum_{i=1}^{N} p_i \log p_i. \tag{13}$$

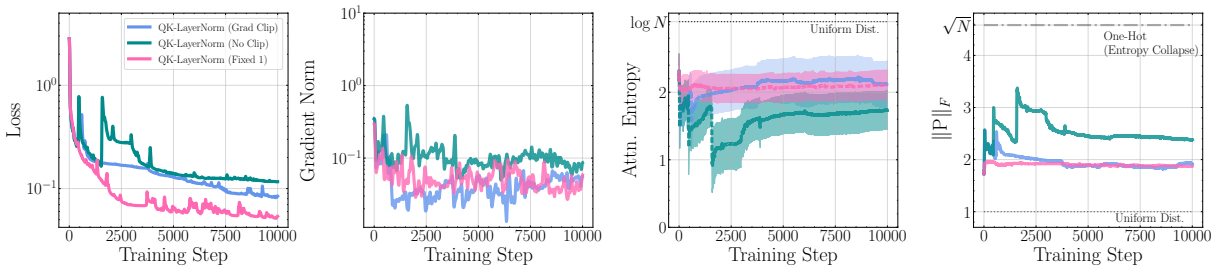

Figure 8: Training behaviors of the scaling parameters γ_q and γ_k are shown under various conditions—including weight clipping, gradient clipping, fixed weights, and no clipping. The experiments are conducted in a simple and small Transformer. From left to right, each column shows the training loss, gradient norm, attention entropy (with $\pm$ standard deviation across all layers), and the average Frobenius norm of the attention probability matrix. Note that the results for weight clipping are shown in Figure 1.

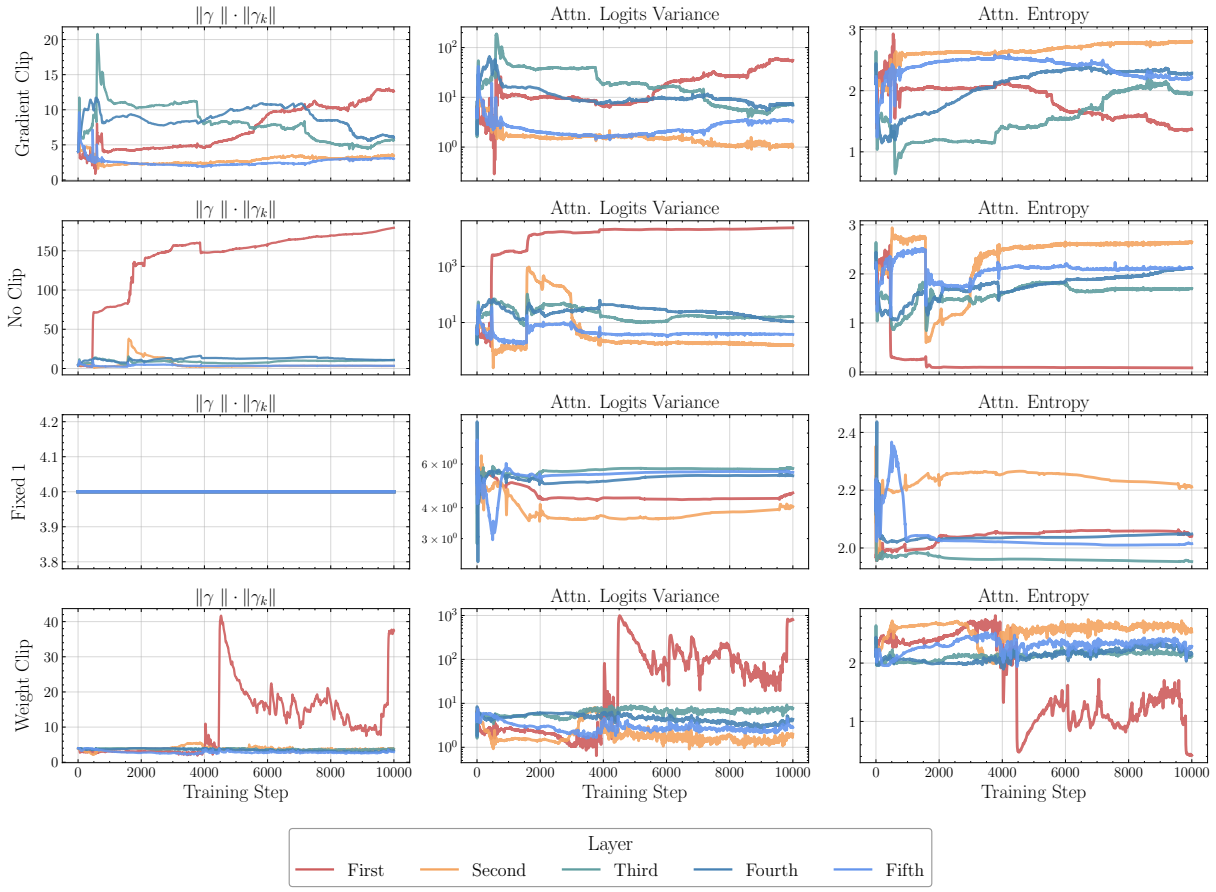

Figure 9: Effects of QK-LayerNorm γ configurations. Each row corresponds to one configuration: Gradient Clip applies gradient clipping to γ; No Clip uses learnable γ without any clipping; Fixed 1 keeps $\gamma_q, \gamma_k = 1$ (non-trainable); and Weight Clip applies value clipping directly to γ. From left to right, for each layer, the parameters γ_q and γ_k with their norm product $\|\gamma_q\| \cdot \|\gamma_k\|$, attention logits variance (Attn. Logits Variance), and attention entropy (Attn. Entropy). Lines with the same color represent the same layer across training steps.

We aim to derive first-order approximation for $H(p)$ in the regime where $\sigma^2 \ll 1$.

When σ^2 is small, the random vector z is concentrated near zero, and hence the softmax output is close to uniform distribution. We can express the softmax probabilities as a perturbation of the uniform vector:

$$p_i = \frac{1}{N} + \zeta_i(z), \tag{14}$$

where the perturbation $\zeta_i(z)$ satisfies $\sum_{i=1}^{N} \zeta_i(z) = 0$, and $\zeta_i(z) = \mathcal{O}(\sigma)$.

Substituting this expansion into the entropy for-

mula yields:

$$H(p) = -\sum_{i=1}^{N} \left(\frac{1}{N} + \zeta_i\right) \log\left(\frac{1}{N} + \zeta_i\right). \quad (15)$$

We perform a Taylor expansion of the logarithm around $\frac{1}{N}$:

$$\log\left(\frac{1}{N} + \zeta_i\right) = \log\left(\frac{1}{N}\right) + N\zeta_i - \frac{N^2}{2}\zeta_i + \mathcal{O}(\zeta_i^3). \quad (16)$$

Therefore, the entropy becomes:

$$\begin{aligned} & H(p) \\ & \approx -\sum_{i=1}^{N} \left(\frac{1}{N} + \zeta_i\right) \left(\log\left(\frac{1}{N}\right) + N\zeta_i - \frac{N^2}{2}\zeta_i^2\right) \\ & = -\log\left(\frac{1}{N}\right) \sum_{i=1}^{N} \left(\frac{1}{N} + \zeta_i\right) \\ & \quad - N \sum_{i=1}^{N} \left(\frac{1}{N} + \zeta_i\right) \zeta_i \\ & \quad + \frac{N^2}{2} \sum_{i=1}^{N} \left(\frac{1}{N} + \zeta_i\right) \zeta_i^2. \end{aligned}$$

Using the fact that $\sum_i \zeta_i = 0$, $\sum_i \frac{1}{N} = 1$, and neglecting higher-order terms, we simplify the expression:

$$\begin{aligned} H(p) &\approx \log N - N \sum_{i=1}^{N} \zeta_i^2 + \frac{N^2}{2} \cdot \frac{1}{N} \sum_{i=1}^{N} \zeta_i^2 \\ &= \log N - \frac{N}{2} \sum_{i=1}^{N} \zeta_i^2. \end{aligned}$$

We now compute the expectation of the perturbation energy:

$$\mathbb{E}_z \left[\sum_{i=1}^{N} \zeta_i^2\right] = \mathbb{E}_z \left[\sum_{i=1}^{N} \left(p_i - \frac{1}{N}\right)^2\right] = \mathrm{Var}(p),$$

which can be approximated by known results for the softmax of a Gaussian:

$$\mathbb{E}_z \left[\sum_{i=1}^{N} \left(p_i - \frac{1}{N}\right)^2\right] \approx \frac{N-1}{N^2}\sigma^2.$$

Substituting this into the entropy expression yields:

$$\begin{aligned} \mathbb{E}_z \left[H(\mathrm{softmax}(z))\right] &\approx \log N - \frac{N}{2} \cdot \frac{N-1}{N^2}\sigma^2 \\ &= \log N - \frac{N-1}{2N}\sigma^2. \end{aligned}$$

E.2 Entropy Approximation for Softmax Version 2

Let $z = (z_1, z_2, ..., z_N) \in \mathbb{R}^N$ be a random vector such that $z_i \sim \mathcal{N}(0, \sigma^2)$ independently. Define the softmax vector $p = \mathrm{softmax}(z)$, where

$$p_i = \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} = \frac{e^{z_i - \bar{z}}}{\sum_{k=1}^{N} e^{z_k - \bar{z}}}, \quad (17)$$

where $\bar{z} = \frac{1}{N}\sum_{k=1}^{N} z_k$ is the empirical mean. We assume the deviations $z_i - \bar{z}$ are small and expand the exponentials using a Taylor expansion up to third order:

$$\begin{aligned} e^{z_k - \bar{z}} &= 1 + \sigma(z_k - \bar{z}) + \frac{1}{2}\sigma^2(z_k - \bar{z})^2 \\ &\quad + \frac{1}{6}\sigma^3(z_k - \bar{z})^3 + \mathcal{O}(\sigma^4). \end{aligned} \quad (18)$$

Then the denominator becomes:

$$\begin{aligned} \sum_{k=1}^{N} e^{z_k - \bar{z}} &= \sum_{k=1}^{N} \Big(1 + \sigma(z_k - \bar{z}) + \tfrac{1}{2}\sigma^2(z_k - \bar{z})^2 \\ &\quad + \tfrac{1}{6}\sigma^3(z_k - \bar{z})^3\Big) + \mathcal{O}(\sigma^4) \end{aligned} \quad (19)$$

By the definition of the mean, $\sum_{k=1}^{n}(z_k - \bar{z}) = 0$. If the data are symmetric with respect to the mean, then $\sum_{k=1}^{n}(z_k - \bar{z})^3 = 0$. Substituting these into (19), we obtain:

$$\begin{aligned} \sum_{k=1}^{N} e^{z_k - \bar{z}} &= N + \frac{1}{2}\,\sigma^2 \sum_{k=1}^{N} (z_k - \bar{z})^2 + \mathcal{O}(\sigma^4) \\ &= N\Big(1 + \tfrac{1}{2}\,\sigma^2 \mathcal{S}_2 + \mathcal{O}(\sigma^4)\Big). \end{aligned} \quad (20)$$

where $\mathcal{S}_2 = \frac{1}{N}\sum_{k=1}^{N}(z_k - \bar{z})^2$. To approximate the softmax, we apply a Taylor expansion to the denominator. This yields:

$$\frac{1}{\sum_k e^{z_k - \bar{z}}} = \frac{1}{N}\left(1 - \frac{1}{2}\sigma^2 \mathcal{S}_2 + \mathcal{O}(\sigma^4)\right). \quad (21)$$

Expanding the numerator similarly:

$$e^{z_i-\bar{z}} = 1 + \sigma(z_i-\bar{z}) + \frac{1}{2}\sigma^2(z_i-\bar{z})^2 \quad (22)$$

$$+ \frac{1}{6}\sigma^3(z_i-\bar{z})^3 + \mathcal{O}(\sigma^4) \quad (23)$$

so the softmax becomes:

$$\begin{aligned} p_i &= \frac{1}{N}\left(1 - \frac{1}{2}\sigma^2\mathcal{S}_2\right)(1 + \sigma(z_i-\bar{z}) \\ &\quad + \frac{1}{2}\sigma 2(z_i-\bar{z})2 + \frac{1}{6}\sigma^3(z_i-\bar{z})^3\Big) + \mathcal{O}(\sigma^4) \\ &= \frac{1}{N}\Big(1 + \sigma(z_i-\bar{z}) \\ &\quad + \sigma^2\left(\frac{1}{2}(z_i-\bar{z})^2 - \frac{1}{2}\mathcal{S}_2\right) \\ &\quad + \sigma^3\left(\frac{1}{6}(z_i-\bar{z})^3 - \frac{1}{2}\mathcal{S}_2(z_i-\bar{z})\right) + \mathcal{O}(\sigma^4)\Big). \end{aligned} \quad (24)$$

The negative log-probability is given by:

$$-\log p_i = -\sigma(z_i-\bar{z}) + \log\sum_k e^{z_k-\bar{z}} \quad (25)$$

$$= -\sigma(z_i-\bar{z}) + \log\left(1 + \frac{1}{2}\sigma^2\mathcal{S}_2 + \mathcal{O}(\sigma^4)\right) \quad (26)$$

$$= \log N - \sigma(z_i-\bar{z}) + \frac{1}{2}\sigma^2\mathcal{S}_2 + \mathcal{O}(\sigma^4). \quad (27)$$

Thus the entropy term is:

$$-p_i\log p_i = \frac{1}{N}\Big[\log N + (\log N - 1)\,\sigma\,(z_i-\bar{z}) \quad (28)$$

$$+ \sigma^2\big(\tfrac{1}{2}\,(z_i-\bar{z})^2 - \tfrac{1}{2}\,\mathcal{S}_2 + \tfrac{1}{2}\,\mathcal{S}_2\log N\big)$$

$$+ \sigma^3\big(\tfrac{1}{6}\,(z_i-\bar{z})^3 - \tfrac{1}{2}\,\mathcal{S}_2\,(z_i-\bar{z})\big) + \mathcal{O}(\sigma^4)\Big]. \quad (29)$$

Summing over i and using $\sum_i(z_i-\bar{z}) = 0$ and $\sum_i(z_i-\bar{z})^2 = N\,\mathcal{S}_2$ then gives

$$\sum_i -p_i\log p_i = \log N - \tfrac{1}{2}\,\sigma^2\,\mathcal{S}_2 + \mathcal{O}(\sigma^4)\,.$$

Summing over i, using $\sum_i(z_i-\bar{z}) = 0$, and $\sum_i(z_i-\bar{z})^2 = N\mathcal{S}_2$, we get:

$$\sum_i -p_i\log p_i = \log N \quad (30)$$

$$+ \sigma^2\left(\tfrac{1}{2}\,\mathcal{S}_2\,\log N - \mathcal{S}_2 + \tfrac{1}{2}\,\mathcal{S}_2\right) \quad (31)$$

$$= \log N - \tfrac{1}{2}\,\sigma^2\,\mathcal{S}_2 + \mathcal{O}(\sigma^4). \quad (32)$$

Taking expectation over z, we obtain:

$$\mathbb{E}_z\left[-\sum_i p_i\log p_i\right] \quad (33)$$

$$= \log N - \frac{1}{2}\sigma^2\mathbb{E}_z[\mathcal{S}_2] + \mathcal{O}(\sigma^4). \quad (34)$$

If we assume the z_i are i.i.d. with unit variance, then:

$$\mathbb{E}_z[\mathcal{S}_2] = \frac{N-1}{N}, \quad (35)$$

and finally:

$$\mathbb{E}_z\Big[-\sum_i p_i\log p_i\Big] \quad (36)$$

$$= \log N - \frac{\sigma^2}{2}\,\frac{N-1}{N} + \mathcal{O}(\sigma^4) \quad (37)$$

$$= \log N - \frac{N-1}{2N}\,\sigma^2 + \mathcal{O}(\sigma^4). \quad (38)$$

E.3 Entropy of Softmax as a Strictly Decreasing Function of Variance

Let $H(\sigma^2)$ denote the expected entropy of the softmax distribution:

$$H(\sigma^2) = \mathbb{E}_z\left[-\sum_{i=1}^{n} p_i(z)\,\log p_i(z)\right].$$

We reparameterize $z = \sqrt{\sigma^2}\,\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I_N)$, and express the softmax distribution as

$$p_i(\varepsilon,\sigma^2) = \frac{\exp(\sqrt{\sigma^2}\,\varepsilon_i)}{\sum_{j=1}^{N}\exp(\sqrt{\sigma^2}\,\varepsilon_j)}.$$

Under this reparameterization, the entropy becomes

$$H(\sigma^2) = \mathbb{E}_\varepsilon\left[\log\left(\sum_j e^{\sqrt{\sigma^2}\,\varepsilon_j}\right) - \sqrt{\sigma^2}\sum_i \varepsilon_i\,p_i(\varepsilon,\sigma^2)\right].$$

Differentiating under the expectation yields

$$\begin{aligned} \frac{\partial H}{\partial\sigma^2} = \;& \mathbb{E}_\varepsilon\Bigg[\frac{1}{2\sqrt{\sigma^2}}\,\frac{\sum_j \varepsilon_j\,e^{\sqrt{\sigma^2}\,\varepsilon_j}}{\sum_k e^{\sqrt{\sigma^2}\,\varepsilon_k}} \\ & - \frac{1}{2\sqrt{\sigma^2}}\sum_i \varepsilon_i\,p_i(\varepsilon,\sigma^2) \\ & - \sqrt{\sigma^2}\sum_i \varepsilon_i^2\,p_i(\varepsilon,\sigma^2) \\ & + \sqrt{\sigma^2}\left(\sum_i \varepsilon_i\,p_i(\varepsilon,\sigma^2)\right)^2\Bigg]. \end{aligned}$$

The first two terms cancel, and substituting back $z = \sqrt{\sigma^2}\,\varepsilon$ gives

$$\frac{\partial H}{\partial \sigma^2} = -\frac{1}{2\sigma^2}\,\mathbb{E}_z\left[\sum_{i=1}^{n} z_i^2\, p_i(z) - \left(\sum_{i=1}^{n} z_i\, p_i(z)\right)^2\right] = -\frac{1}{2\sigma^2}\,\mathbb{E}_z\big[\mathrm{Var}_{p(z)}[z]\big].$$

Because the inner variance is strictly positive almost surely,

$$\frac{\partial H}{\partial \sigma^2} < 0 \quad \text{for all } \sigma^2 > 0.$$

E.4 Entropy Approximation of ReLU kernel Attention

We consider query and key vectors defined as

$$Q_i = \sigma\, g_i, \qquad K_j = \sigma\, h_j,$$

where $g_i, h_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and $\sigma > 0$. We apply the ReLU activation function $\phi(x) = \max(0, x)$), which is positively homogeneous of degree one, i.e., $\phi(\lambda x) = \lambda\phi(x)$ for any $\lambda > 0$. Using this property, we obtain

$$\phi(Q_i) = \sigma\,\phi(g_i), \qquad \phi(K_j) = \sigma\,\phi(h_j).$$

Then we define the unnormalized attention logits as

$$t_{ij} := \phi(g_i)\,\phi(h_j)^\top, \;\; s_{ij} := \phi(Q_i)\,\phi(K_j)^\top = \sigma^2\, t_{ij}.$$

Here, t_{ij} corresponds to the inner product between the vectors g_i and h_j, while s_{ij} is the scaled version of t_{ij} by a factor of σ^2. We then convert these logits into probabilities by applying row-wise normalization:

$$\tilde{p}_{i,j}(\sigma) = \frac{s_{ij}}{\sum_{k=1}^{N} s_{ik}} = \frac{\sigma^2 t_{ij}}{\sigma^2 \sum_{k=1}^{N} t_{ik}} = \tilde{p}_{i,j}(1).$$

Note that the factor σ^2 cancels out, the resulting attention probabilities are invariant to σ. Accordingly, the row-wise entropy is defined as

$$H_i(\sigma) := -\sum_{j=1}^{N} \tilde{p}_{i,j}(\sigma)\,\log \tilde{p}_{i,j}(\sigma),$$

which implies that $H_i(\sigma) = H_i(1)$ for all $\sigma > 0$. For each coordinate $k = 1, \ldots, d$ let $G = g_i^{(k)}$, $H = h_j^{(k)}$, and define

$$X_k Y_k \;=\; \phi\big(g_i^{(k)}\big)\,\phi\big(h_j^{(k)}\big).$$

Each such term contributes to the dot product $t_{i,j}$, and its expectation and variance are given by

$$\mu = \mathbb{E}[X_k Y_k] = \frac{1}{2\pi},$$
$$\tau^2 = \mathrm{Var}(X_k Y_k) = \frac{\pi^2 - 1}{4\pi^2}.$$

By independence and linearity, the mean and variance of $t_{i,j}$ are

$$\mathbb{E}[t_{ij}] = \sum_{k=1}^{D} \mathbb{E}[X_k Y_k] = D\,\mu,$$
$$\mathrm{Var}(t_{ij}) = \sum_{k=1}^{D} \mathrm{Var}(X_k Y_k) = D\,\tau^2.$$

Moreover, since each $X_k Y_k$ has finite variance, central limit theorem applies, giving as $d \to \infty$

$$t_{ij} = \sum_{k=1}^{D} X_k Y_k = D\,\mu + \sqrt{D}\,\tau\,\xi_{ij}, \qquad \xi_{ij} \xrightarrow{D} \mathcal{N}(0,1).$$

Fixing i, define

$$\bar{t}_i = \frac{1}{N}\sum_{j=1}^{N} t_{ij}, \quad \delta_{ij} = \frac{t_{ij} - \bar{t}_i}{\bar{t}_i}, \quad \sum_{j=1}^{N} \delta_{ij} = 0.$$

Since $\bar{t}_i = D\mu + \mathcal{O}_p(\sqrt{D})$, we have $\delta_{ij} = \mathcal{O}_p(D^{-1/2})$. Hence,

$$\tilde{p}_{i,j}(1) = \frac{1}{N}(1 + \delta_{ij}),$$

and a second-order Taylor expansion around the uniform distribution gives

$$H_i(1) = -\sum_{j=1}^{N} \tilde{p}_{ij}(1)\,\log \tilde{p}_{ij}(1) = \log N - \frac{1}{2N}\sum_{j=1}^{N} \delta_{ij}^2 + \mathcal{O}\big(\|\delta_i\|_3^3\big).$$

Finally, since

$$\mathbb{E}[\delta_{ij}^2] = \frac{\tau^2}{D\mu^2} + o(D^{-1}), \quad \mathbb{E}\|\delta_i\|_3^3 = o(D^{-1}),$$

it follows that

$$\mathbb{E}[H_i(1)] = \log N - \mathcal{O}(D^{-1}).$$

F Correlation Between Attention Entropy and Attention Probabilities Norm

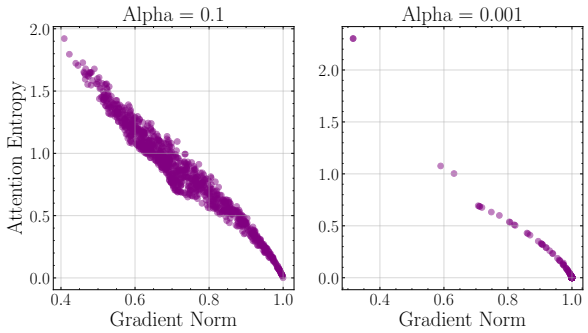

Figure 10: Correlation between the attention entropy and ℓ_2-norm of each row after sampling rows of attention probabilities from a Dirichlet distribution. For this setup, the concentration hyper-parameter α of the Dirichlet distribution is configured as 0.1 and 0.001 during sampling.

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:

$$P_i \sim \text{Dirichlet}(\alpha \mathbf{1}) \tag{39}$$

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, which resembles 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.

G Attention heatmaps

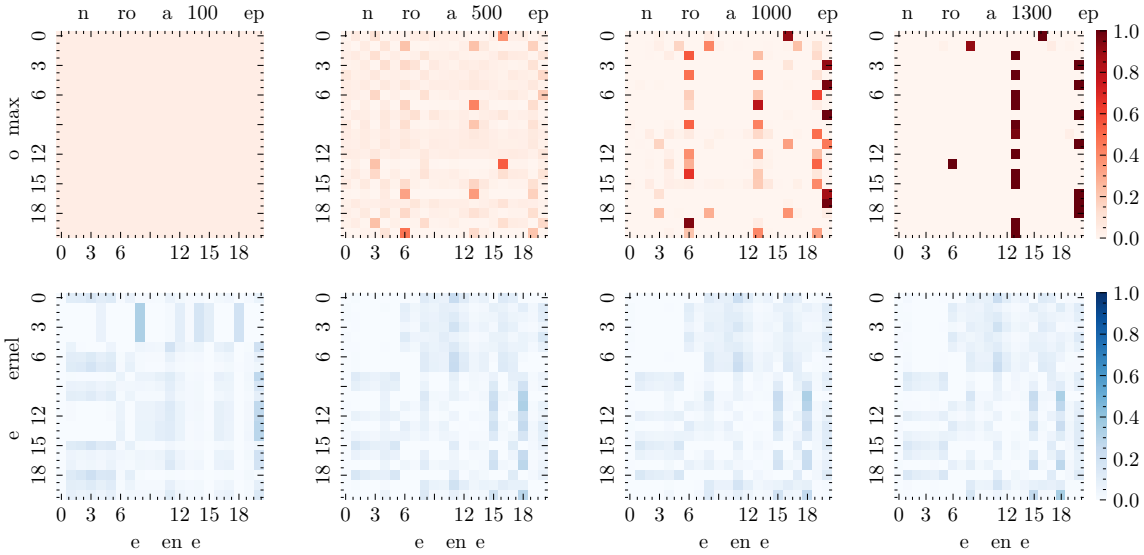

Figure 11: Heatmaps of attention probabilities for softmax-based attention (top) and entropy-stable attention (bottom) during training. In softmax-based attention, each row progressively converges to a one-hot-like vector, leading to attention entropy collapse. The attention matrices are from the first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