PROBING ROTARY POSITION EMBEDDINGS THROUGH FREQUENCY ENTROPY

Yui Oka[1*], **Kentaro Hanafusa**[2*†], **Taku Hasegawa**[1], **Kyosuke Nishida**[1], **Kuniko Saito**[1]
[1]Human Informatics Labs., NTT, Inc.
[2]Ehime University
yui.oka@ntt.com

ABSTRACT

Rotary Position Embeddings (RoPE) are widely used in Transformers to encode positional information in token representations, yet the internal frequency structure of RoPE remains poorly understood. Previous studies have reported conflicting findings on the roles of high- and low-frequency dimensions, offering empirical observations but no unifying explanation. In this paper, we present a systematic framework that bridges these disparate results. We introduce Frequency Entropy (FE), a metric that quantifies the effective utilization of each RoPE frequency dimension, and we provide an analysis of how RoPE's sinusoidal components contribute to model representations on a per-dimension basis. Based on an analysis of the Llama-4 model, which incorporates both RoPE and NoPE layers, we find that the periodicity captured by FE appears in RoPE layers but not in NoPE layers. Furthermore, FE identifies dimensions in which energy concentrates under RoPE. These characteristics are observed across the spectrum rather than being confined to specific dimensions. Moreover, attenuating extreme-entropy dimensions at inference yields downstream accuracy that is statistically indistinguishable from the baseline, with modest perplexity improvements on average, suggesting that such dimensions are often redundant. Overall, FE provides a simple, general diagnostic for RoPE with implications for analysis and design.

1 INTRODUCTION

Position representations in Transformers (Vaswani et al., 2017) are a crucial factor determining their ability to handle long-range dependencies. Among these representations, Rotary Position Embeddings (RoPE) (Su et al., 2023) have become standard in many of the latest large language models, including Llama (Touvron et al., 2023; Grattafiori et al., 2024), Qwen (Qwen et al., 2025; Yang et al., 2025), and Gemma (Gemma Team et al., 2024a), and have contributed to improved performances in long-text processing. However, the design of RoPE was introduced empirically, and the role of each frequency dimension and how they are utilized within the model remain unclear.

In recent years, several analyses at the RoPE dimension level have been reported. For example, Barbero et al. (2025) observed that high-frequency components contribute to positional pattern formation, while low-frequency components contribute to semantic information. They also demonstrated that replacing part of the low-frequency components with NoPE (Kazemnejad et al., 2023) does not significantly affect model performance. On the other hand, Chiang & Yogatama (2025) showed that high-frequency components are scarcely utilized and can be removed without impacting performance. Furthermore, Hong et al. (2024) pointed out that the low-frequency component is essential for modeling long-range dependencies in specific attention heads. Previous analyses of RoPE have largely relied on visual inspection of heatmaps and coarse division into high- versus low-frequency components. The Llama-4 (Meta, 2025) model introduces *iRoPE*, which combines interleaved NoPE layers with frequency scaling to extend context length, though its internal frequency dynamics remain largely unexplored. These examples highlight the conflicting reports regarding the

[*] These authors contributed equally.
[†] Work done during internship at NTT, Inc.

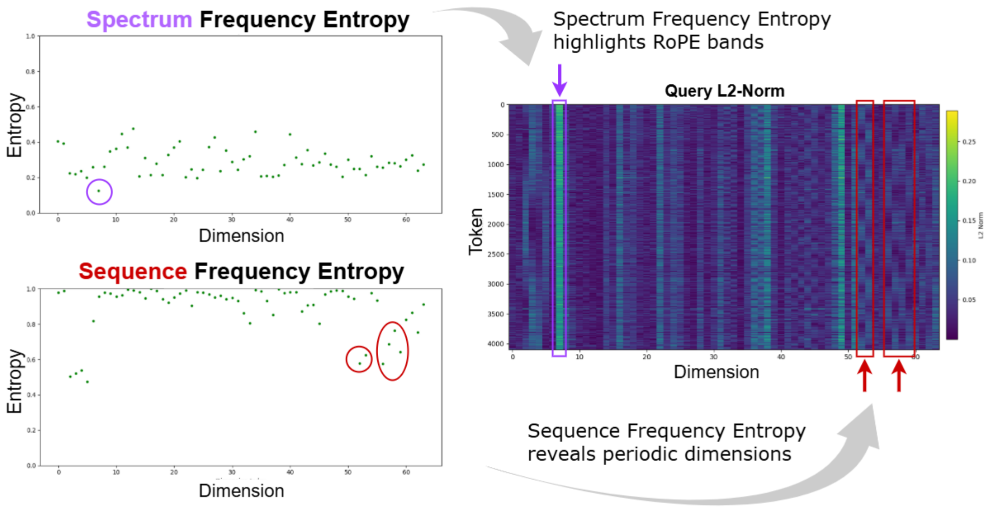


Figure 1: Example of our Frequency Entropy. Left: Frequency Entropy (FE) per rotary pair, where the horizontal axis is the entropy value and the vertical axis is the pair index $j \in \{0, \ldots, d/2 - 1\}$. Right: Query ℓ_2-norms across rotary pairs, where the horizontal axis is sequence length ($L = 4096$) and the vertical axis is the pair index.

role of each RoPE dimension, and analysis remains observation-based, lacking a unified theoretical and empirical understanding.

In this paper, we mathematically formalize the contribution of each frequency dimension and introduce **Frequency Entropy (FE)**, a classical quantitative measurement framework. It comprises two complementary metrics, Spectrum Frequency Entropy (Spectrum FE) and Sequence Frequency Entropy (Sequence FE), and makes RoPE's spectral behavior measurable on a per-dimension basis. Figure 1 illustrates the method. We compute these entropies on a Llama-4 model with RoPE and NoPE layers and find that Spectrum FE quantifies power concentration in the Fourier domain and reveals band-limited rotation pairs with sustained energy, while Sequence FE measures token-wise regularity and detects persistent oscillations of rotation pairs induced by RoPE. Furthermore, these signatures are absent in layers with NoPE.

To probe the functional relevance without fine-tuning, we conduct a targeted attenuation study. For rotation pairs whose FE falls below a threshold, we reduce their contribution on attention during inference by multiplying the corresponding query and key channels by a constant $\alpha < 1$, keeping all other parameters fixed. The experimental results demonstrate that suppressing low-Sequence-FE dimensions leaves perplexity and downstream accuracy unchanged, whereas suppressing low-Spectrum-FE dimensions worsens perplexity. In addition, attenuating high-Spectrum-FE outliers has a negligible effect on the perplexity and the downstream accuracy. These results indicate that persistent oscillations induced by RoPE are largely redundant, while band-limited components carry a task-relevant signal.

Across experiments, we observe that oscillatory and band-limited behaviors occur throughout the spectrum rather than only at specific frequencies or extremes. Consequently, the conventional high- versus low-frequency dichotomy is insufficient. FE provides a spectrum-aware, model-agnostic lens that reconciles prior mixed observations and informs pruning, reweighting, and the design of future positional schemes.

2 Background and Related Work

2.1 Rotary Position Embedding (RoPE)

RoPE (Su et al., 2023) has become the de facto standard positional embedding method in many of the latest large language models, such as Llama (Touvron et al., 2023; Grattafiori et al., 2024) Qwen (Qwen et al., 2025; Yang et al., 2025), and Gemma (Gemma Team et al., 2024a). RoPE introduces position information by applying a rotation to the query and key vectors in the self-attention mechanism. This property allows RoPE to encode relative positional information while preserving compatibility with an absolute token index.

$$A_{m,n} = (R_{n,\theta} q_n)^\top (R_{m,\theta} k_m) = q_n^\top R_{m-n,\theta}\, k_m \tag{1}$$

The rotation matrix $R_{n,\theta} \in \mathbb{R}^{d \times d}$ is the block-diagonal rotation matrix, defined as follows:

$$R_{n,\theta} = \begin{bmatrix} \cos(n\theta_0) & -\sin(n\theta_0) & 0 & 0 & \cdots & 0 & 0 \\ \sin(n\theta_0) & \cos(n\theta_0) & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos(n\theta_1) & -\sin(n\theta_1) & \cdots & 0 & 0 \\ 0 & 0 & \sin(n\theta_1) & \cos(n\theta_1) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos(n\theta_{d/2-1}) & -\sin(n\theta_{d/2-1}) \\ 0 & 0 & 0 & 0 & \cdots & \sin(n\theta_{d/2-1}) & \cos(n\theta_{d/2-1}) \end{bmatrix} \tag{2}$$

where $A \in \mathbb{R}^{L \times L}$, $q_n \in \mathbb{R}^{1 \times d}$ is the n-th query when the number of dimensions is d and n is the absolute position $(0 \leq n \leq L-1)$ when the sequence length is L, and $k_m \in \mathbb{R}^{1 \times d}$ is the m-th key $(0 \leq m \leq L-1)$. Typically, the rotation angles are chosen as the base of RoPE $\theta_j = 10000^{-2j/d}$ ($j = 0, \ldots, \frac{d}{2} - 1$). In practice, the base values θ in RoPE are typically set to be quite large. For example, $\theta = 10{,}000$ is adopted in the Gemma model (Gemma Team et al., 2024a) and Llama-2 (Touvron et al., 2023), $\theta = 500{,}000$ is used in Llama-3 (Xiong et al., 2024), and $\theta = 1{,}000{,}000$ is employed in Qwen-3 (Yang et al., 2025). RoPE provides a complex phase concept of relative offsets for attention across all layers and all heads while remaining parameter-free and hardware-friendly.

2.2 Utilization of information within RoPE

The distribution and utilization of information within RoPE at the dimensional level remain only partially understood. An early perspective at the attention-head level emphasized the low-frequency structure as necessary for long-distance modeling. Hong et al. (2024) identified "positional heads" whose activations align with token distance. Ablating these heads—dominated by lower-frequency (slower-varying) RoPE dimensions—substantially degrades the long-context performance. A second line of empirical evidence dissected pretrained models to associate roles with high and low frequency. (Barbero et al., 2025) reported that high-frequency components drive distinctive off-diagonal "positional" attention patterns, whereas low-frequency components correlate more with semantic content. Further, replacing parts of the low-frequency spectrum with NoPE-like (Kazemnejad et al., 2023) variants leaves performance largely intact or even improved in small-to-mid-scale settings. These results suggest redundancy within portions of the low-frequency subspace. A third line of work has argued almost the converse for the high-frequency end: Chiang & Yogatama (2025) measured the per-dimension utilization and showed that dimensions with larger rotation angles (i.e., higher frequencies) are weakly used. This points to an over-allocation of representational capacity to rapid positional oscillations that downstream layers seldom exploit.

Methodologically, existing analyses have tended to employ broad classifications such as high-frequency/low-frequency, lacking a unified metric that transcends dimensions. Taken together, these findings create a tension: (i) particular heads critically depend on low-frequency structure for long-range dependencies (Hong et al., 2024); (ii) parts of the low-frequency spectrum appear semantically entangled and sometimes dispensable (Barbero et al., 2025); yet (iii) high-frequency dimensions look under-utilized and safely removable (Chiang & Yogatama, 2025).

2.3 iRoPE

Llama-4 (Meta, 2025) introduces *iRoPE*, which augments RoPE with two key changes. First, it utilizes an interleaved layer design where rotary position embeddings are applied only to alternating attention layers, while the others operate without explicit positional signals (NoPE; Kazemnejad et al., 2023), encouraging content-based long-range reasoning. Second, iRoPE scales the RoPE rotation angles by a factor $\alpha < 1$, slowing phase growth and extending the effective positional range well beyond the training context. Despite these advances, the internal behavior of iRoPE remains largely unexplored.

2.4 Frequency Entropy

Frequency entropy (also called *spectral entropy*) quantifies the uncertainty of a signal's frequency distribution by applying Shannon's entropy (Shannon, 1948) to its power spectrum. Let $P(f_i)$ be the power spectral density at the i-th discrete frequency bin, where $i = 1, \ldots, N$ and N is the total number of frequency bins obtained from the discrete Fourier transform. We normalize the spectrum to obtain a probability mass function $p_i = \frac{P(f_i)}{\sum_j P(f_j)}$. The frequency entropy is calculated as

$$H = -\sum_i p_i \log_2 p_i. \tag{3}$$

When desired, H can be normalized by $\log_2 N$ to yield $0 \leq H \leq 1$. A low value of H indicates that the spectral energy is concentrated in a few frequencies and is therefore highly ordered—for example, as in a pure tone—whereas a high value indicates that the energy is spread across many frequencies and is therefore more disordered, as in white noise. Thus, frequency entropy provides a single quantitative measure of the flatness or peakiness of a spectrum. This metric is widely used in signal processing and information theory (Misra et al., 2004; Sucic et al., 2014).

3 Frequency Entropy: A New Lens for Rotary Position Embeddings

After RoPE, each pair is rotated by an angle $n\theta_{d/2-1}$ with $\theta_{d/2-1}$ determined by the RoPE base. We are interested in *how strongly* each rotary pair exhibits narrow-band, RoPE-driven periodicity along the sequence, as opposed to content-driven, broadband variability.

Core idea. For each rotary pair, we construct a 1D block observable along the token axis and measure its frequency entropy (FE), defined as the Shannon entropy of the power spectrum. To quantify the utilization of each dimension, we introduce two new evaluation metrics: **Spectrum Frequency Entropy (SpectrumFE)** and **Sequence Frequency Entropy (SequenceFE)**. SpectrumFE evaluates which frequency components are present at any given moment. SequenceFE evaluates how periodic or irregular the energy fluctuations are in each dimension. Our FE is computed as follows:

1. Split the query into $d/2$ RoPE blocks and compute the ℓ_2-norm of each, treating the resulting length-L vector as a discrete signal (Section 3.1).
2. Compute two variants of power spectrum and applying Shannon entropy (Shannon, 1948) to its power spectrum as shown in Eq. (3) (Sections 3.2 and 3.3).

Low entropy indicates that the time series is dominated by a small number of frequencies, while high entropy indicates spectrally diverse, content-modulated dynamics. This yields a per-pair score that is model-agnostic, scale-free, and comparable across layers, heads, and architectures.

3.1 From RoPE Blocks to a Discrete Frequency Signal

To measure the usage of frequencies, we start by noting any Cauchy-Schwarz equalities following (Barbero et al., 2025). The effect of the j-th frequency component on the activation $A_{n,m}$ is upper bounded by the ℓ_2-norm of the query and key components, i.e., $\left|\langle q_n^{(j)}, k_m^{(j)} \rangle\right| \leq \|q_n^{(j)}\|_2 \, \|k_m^{(j)}\|_2$ $(j = 0, \ldots, \frac{d}{2} - 1)$. For a fixed block j, we collect these vectors across the sequence by

$$\mathbf{s}_j \;:=\; \left[\, \|q_0^{(j)}\|_2,\; \|q_1^{(j)}\|_2,\; \ldots,\; \|q_{L-1}^{(j)}\|_2 \,\right]^\top \;\in\; \mathbb{R}^L. \tag{4}$$

where L is the sequence length. Therefore, we assume that the set of ℓ_2-norms $\|q_n^{(j)}\|_2$ of queries after RoPE constitutes a discrete signal. We calculate the frequency entropy of the $d/2$ discrete signal patterns and measure the frequency utilization rate. In practical terms, measuring the ℓ_2-norm of a query after RoPE is natural: the rotation aligns the representation with the frequency blocks actually used in the logit and preserves norms, so $\|q_n^{(j)}\|_2$ is both position-invariant and directly interpretable. We treat Eq. (4) as a discrete-time signal.

3.2 Spectrum Frequency Entropy

We measure the spectral complexity of this sequence using normalized spectral entropy. For a given signal $\mathbf{s}_j$, first compute the short-time Fourier transform (STFT) to obtain the power spectrum, as

$$S_{k,t} = \left| \sum_{n=0}^{F-1} \mathbf{s}_j[tH+n]\, w[n]\, e^{-i\frac{2\pi}{F}kn} \right|^2 . \tag{5}$$

where $t(t=0,1,2,\dots,T-1)$ is each frame, $w[n]$ is the analysis window, and $S_{k,t}$ is the power spectrum at frequency bin $k(k=0,1,2,\dots,K-1)$. We set the frame length F to 1024, hop length H to 512, and sequence length L to 4096. Therefore, the frequency bin is $K=\frac{F}{2}+1=513$ and number of frames $T=\left\lfloor \frac{L-F}{H} \right\rfloor+1=7$. Second, the power spectrum for each frame is normalized to form a probability distribution, as

$$p_k = \frac{S_k}{\sum_{j=0}^{K-1} S_j}, \quad S_k = \frac{\sum_{t=0}^{T-1} S_{k,t}}{T}. \tag{6}$$

Next, the Shannon entropy H (Shannon, 1948) is calculated as

$$H = -\sum_{k=0}^{K-1} p_k \, \log_2 p_k. \tag{7}$$

To obtain a scale-free measure, we divide by the maximal entropy $H_{\max}=\log_2 K$, yielding the normalized spectral entropy $\tilde{H}=\frac{H}{H_{\max}}$. The normalized spectral entropy $\tilde{H}$ represents the Spectrum Frequency Entropy (SpectrumFE), quantifying the temporal spectral diversity of the query Lℓ_2-norm signal. Finally, since RoPE organizes the embedding into $\frac{d}{2}$ two-dimensional rotary pairs, we compute $\tilde{H}_j$ for each pair index $j \in 0,\dots,\frac{d}{2}-1$. Hence, the entropy is defined along j and yields a length-$\frac{d}{2}$ vector $(\tilde{H}_0,\dots,\tilde{H}_{\frac{d}{2}-1})$, with one value per rotary pair.

3.3 Sequence Frequency Entropy

For a given signal $\mathbf{s}_j$, first compute the discrete Fourier transform (DFT) to obtain the power spectrum as follows:

$$S_k = |\sum_{n=0}^{L-1} \mathbf{s}_j[n]\, e^{-i\frac{2\pi}{L}kn}|^2, \qquad k=0,1,\dots,L-1. \tag{8}$$

Next, discard the DC component ($k=0$) and restrict to the positive frequencies $1 \le k \le \frac{L}{2}-1$. Define the total positive–frequency energy and normalize to obtain a probability distribution over these frequencies:

$$p_k = \frac{S_k}{\sum_{k=1}^{\lfloor L/2 \rfloor -1} S_k}, \qquad k=1,\dots,\left\lfloor \frac{L}{2} \right\rfloor -1. \tag{9}$$

Finally, the Shannon entropy H (Shannon, 1948) is calculated as

$$H = -\sum_{k=1}^{\lfloor L/2 \rfloor -1} p_k \, \log_2 p_k. \tag{10}$$

To obtain a scale-free measure, we divide by the maximal entropy $H_{\max}=\log_2 K$, yielding the normalized spectral entropy $\tilde{H}=\frac{H}{H_{\max}}$. The normalized spectral entropy $\tilde{H}$ represents the Sequence Frequency Entropy (SequenceFE), quantifying the temporal spectral diversity of the query Lℓ_2-norm signal. Finally, since RoPE organizes the embedding into $\frac{d}{2}$ two-dimensional rotary pairs, we compute $\tilde{H}_j$ for each pair index $j \in 0,\dots,\frac{d}{2}-1$. Hence, the entropy is defined along j and yields a length-$\frac{d}{2}$ vector $(\tilde{H}_0,\dots,\tilde{H}_{\frac{d}{2}-1})$, with one value per rotary pair.

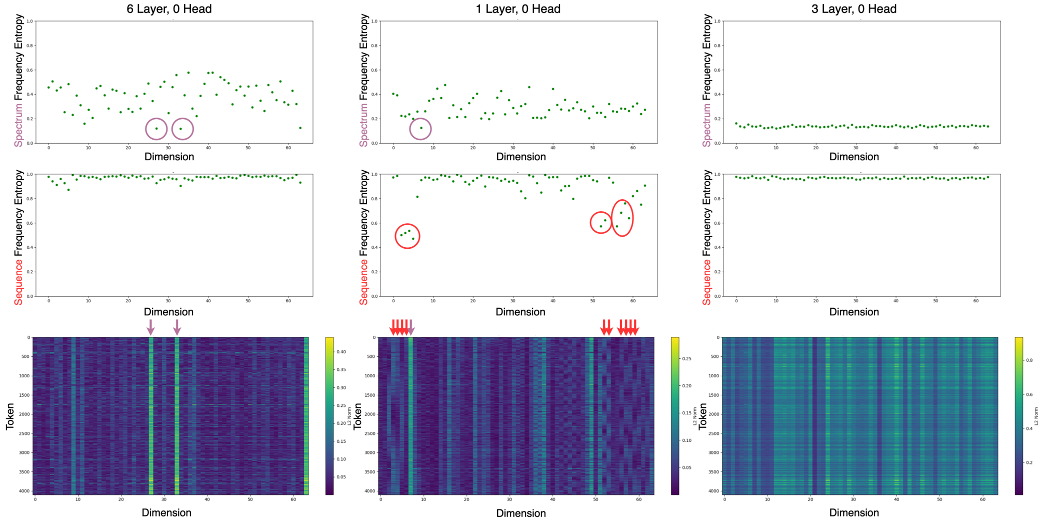


Figure 2: Scatter plots of each FE value in the `Llama-4-Scout-17B-16E-Instruct` model. Columns: layer 6, layer 1, layer 3 (left to right). Rows: SpectrumFE, SequenceFE, and query ℓ_2-norm map. Top/middle: pair index in RoPE j (x) vs. normalized entropy $\tilde{H}_j$ (y). Bottom: j (y) vs. token index n (x); color denotes $\|q_n^{(j)}\|_2$. All results are shown for head 0. The pair index j is rotation patterns. Sequence length $L = 4096$.

4 Analysis via Frequency Entropy

In this section, we investigate the characteristic behavior of RoPE. To isolate their contribution, we compare queries that employ RoPE with those using no positional encoding (NoPE), enabling a direct identification of RoPE-specific effects. For this purpose, we primarily analyze the Llama-4 model with *iRoPE*, which alternates RoPE and NoPE across attention layers. [1]

4.1 Settings

We conduct experiments using the `Llama-4-Scout-17B-16E-Instruct` (Meta, 2025) model with a head dimension of 128 and RoPE with 64 rotation patterns, 48 transformer layers, and 40 attention heads. In the `Llama-4-Scout-17B-16E-Instruct` model, a total of $(64 \times 48 \times 40) = 122{,}880$ frequency-entropy values are computed. For evaluation, we sample text at random from the `wikitext-103-raw-v1` split of the WikiText-103 dataset (Merity et al., 2017) and then concatenate the samples to form sequences of exactly 4096 tokens. Each such sequence is passed to the model, and we extract the attention query vectors during inference. We then compute each frequency entropy from these queries.[2]

4.2 Analysis Results

What characteristics does SpectrumFE capture? In Fig. 2, the top panel shows scatter plots of SpectrumFE per rotary pair, and the bottom panel shows the corresponding query ℓ_2-norm maps. In the query ℓ_2-norm map of layer 6, we observed contiguous stretches of rotary-pair indices with persistently elevated norms. We refer to contiguous ranges of rotary-pair indices that exhibit persistently high query ℓ_2 norms as *frequency bands*. Similar banded patterns were also reported by Barbero et al. (2025) and are especially evident in positional heads. Dimensions with the smallest SpectrumFE align with pronounced band-limited patterns in the ℓ_2-norm maps, indicating that

[1] The analysis of keys is in Appendix A, and the layer-wise analysis is in Appendix B. We also conduct the same analysis on models that apply RoPE in all layers and heads, including Llama-3, Qwen3, and Gemma2. See Appendix C for details.

[2] We additionally provide a context-length analysis in Appendix F a cross-dataset analysis in Appendix F.2, and an ablation study comparing the RoPE-only and NoPE-only models in Appendix G.

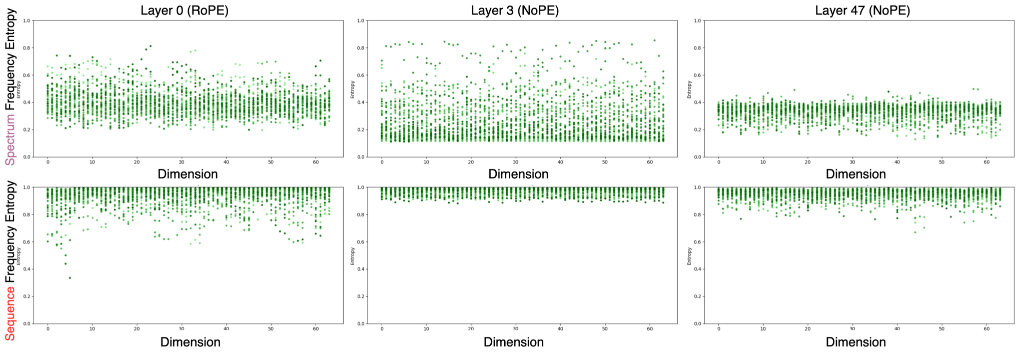


Figure 3: Scatter plots of FE values across all heads in the `Llama-4-Scout-17B-16E-Instruct` model. Columns: layer 0, layer 3, layer 47 (left to right). Layer 0 is the RoPE layer, while the others are NoPE layers. Rows: entropy scatter plots for SpectrumFE and SequenceFE. Scatter plots show pair index in RoPE j (x) vs. normalized entropy $\tilde{H}_j$ (y). We vary the color intensity of the scatter points by head. The pair index j is rotation patterns.

SpectrumFE captures the frequency-band structure. Across RoPE-applied layers, SpectrumFE values predominantly lie in the range 0.2–0.6. The SpectrumFE distribution in the NoPE layer (layer 3) yields a markedly different profile from the RoPE layers. In the NoPE layer (layer 3), multiple frequency bands are observed in the query ℓ_2-norm maps.

What characteristics does SequenceFE capture? In Fig. 2, the middle panel presents scatter plots of SequenceFE per rotary pair. In layer 0 (RoPE layer), pairs with the smallest SequenceFE exhibit clear periodic oscillations in the corresponding query ℓ_2-norm maps, whereas in layer 3 (NoPE layer), no periodic pattern is observed and SequenceFE values rarely fall below 0.8. These observations indicate that SequenceFE is sensitive to periodic structure along the token axis. Across RoPE-applied layers, SequenceFE predominantly ranges from 0.2 to 0.6. In contrast, in the NoPE layer (layer 3), no periodicity emerges and the SequenceFE distribution concentrates near 1.0, yielding a markedly different profile from the RoPE layers (layers 0 and 6).

Effect of the NoPE Layer The bottom row of Fig. 2 shows query ℓ_2-norm heatmaps indicating that RoPE layers and the NoPE layer at layer 3 exhibit markedly different behavior. From SpectrumFE, the shallow NoPE layer (layer 3) exhibits more frequency bands than the RoPE layer. Meanwhile, from SequenceFE, the NoPE layer does not show rotating pairs and therefore does not exhibits clear periodic oscillations.

Entropy Landscapes Across Layers and Heads Fig. 3 shows scatter plots of each FE per rotary pair in all heads. In the SpectrumFE plot, the shallow NoPE layer (layer 3) exhibits more dimensions with frequency-band-like characteristics than the RoPE layer. Furthermore, the SpectrumFE distribution is widely scattered. However, as the layer deepens (final layer 47), the number of dimensions with band-like characteristics decreases, and the SpectrumFE distribution converges within a certain range. In the SequenceFE plot, the shallow RoPE layer (layer 0) indicates that there are several periodic dimensions. However, in the NoPE layer, this periodic dimension disappears for all heads, regardless of whether the layer is shallow or deep. These results suggest that NoPE may attenuate the periodic structure characteristic of RoPE and emphasize frequency bands.

4.3 Why SpectrumFE shows bands and SequenceFE shows periodicity?

In summary, our analysis separates two types of structure in RoPE-driven attention: Spectrum Frequency Entropy reveals band-focused allocation across rotary pairs, and Sequence Frequency Entropy reveals tokenwise periodicity. NoPE may weaken the latter while preserving or highlighting the former. Deep layer may reduce both band sharpness and periodicity.

Why SpectrumFE shows bands? SpectrumFE takes a short-time spectrum of the query-norm for each rotary pair by STFT and measures the Shannon entropy over frequency bins. SpectrumFE measures how narrowly concentrated the local frequency content (via STFT) is. Entropy is low when energy sits in a few bins and high when it is spread out. Therefore, plotting low SpectrumFE across indices (bins) forms a contiguous low-entropy region, i.e., a frequency band. Low SpectrumFE indicates strong frequency-band structure, meaning the model consistently allocates energy to specific rotary pairs.

Why SequenceFE shows periodicity? SequenceFE uses a global Fourier spectrum along the token axis by DFT and measures the entropy over positive frequencies. SequenceFE measures global periodicity of the RoPE-transformed signal (via DFT). Low SequenceFE indicates near–single-tone oscillation driven by RoPE's fixed rotational phase, rather than content. Low entropy means a simple, near-single-tone signal, and high entropy means a complex or noise-like signal. With RoPE active, a rotary pair advances at an almost constant step per token, so the query oscillates at a fixed frequency. If we remove RoPE, the fixed-rate oscillation vanishes, energy spreads, and SequenceFE rises.

Why bands can persist in NoPE for SpectrumFE? Even without rotational oscillation, the model may up-weight some rotary pairs due to content or architectural bias. This uneven allocation still concentrates energy for those indices within short windows, keeping SpectrumFE low over a contiguous set of indices. At the same time, the per-token signal is not strongly periodic, so SequenceFE remains high.

5 Filtering out outlier dimensions

Low SpectrumFE reflects band-limited behavior, while low SequenceFE reflects strong periodicity. The following question arises: *Are frequency bands and periodicity redundant elements for the model, or are they essential components?* In this section, guided by Frequency Entropy, we intervene in RoPE by selectively weighting these rotary pairs to mitigate their contribution.

5.1 Weighted RoPE

If Frequency Entropy is below a certain threshold, we weight the corresponding RoPE dimension to reduce its effect. We call this *Weighted RoPE*. Let $\tilde{H}^{(l,h,j)} \in [0,1]$ denote the Frequency Entropy in the query for layer l, head h, and rotary pair j. Given a threshold $\tau \in (0,1)$ and a RoPE weight $\alpha \in [0, 0.1, ..., 0.9]$, we set a weighted factor as follows:

$$\alpha^{l,h,j} = \begin{cases} \alpha, & \text{if } \tilde{H}^{l,h,j} < \tau, \\ 1, & \text{otherwise.} \end{cases} \tag{11}$$

We modulate the RoPE transformation applied to the j-th rotary pair of the query at token position m. Let $R_{m,\theta}^{(l,h,j)} \in \mathbb{R}^{2\times 2}$ denote the standard RoPE rotation for pair j in layer l and head h. We obtain the weighted query subvector by scaling the usual rotation with $\alpha^{l,h,j}$, which acts as a soft mask that attenuates the contribution of low-entropy pairs while leaving others unchanged, as

$$q_m^{(j)\star} = \alpha^{l,h,j} R_{m,\theta}^{(l,h,j)} q_m^{(j)} \tag{12}$$

Intuitively, for low-entropy pairs, we slow the phase growth and attenuate periodicity; for other pairs, RoPE remains unchanged. We calculate the FE for each key and perform the same processing.

5.2 Perplexity

Settings Four publicly available large language models were used for the evaluation: `Llama-4-Scout-17B-16E` (Meta, 2025), `gemma-2-9b-it` (Gemma Team et al., 2024b), `Qwen3-8B` (Yang et al., 2025), and `Meta-Llama-3-8B` (Grattafiori et al., 2024). We evaluated the inference perplexity on the test set of the wikitext-103 dataset (Merity et al., 2017). For the threshold parameter τ, we adopted different settings according to the entropy metric. For Spectrum Frequency Entropy, we examined two regimes: one where τ was greater than 0.4 and one where τ was less than 0.2. For Sequence Frequency Entropy, we observed that outliers occurred only at low values, so we evaluated the case where τ was less than 0.6. No fine-tuning was performed in any of the experiments.

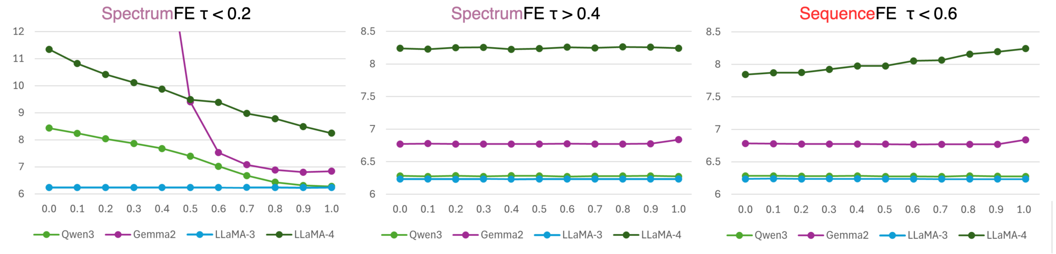


Figure 4: Perplexity as a function of the RoPE weight α under entropy-based gating. The horizontal axis shows the weight α in Weighted RoPE and the vertical axis shows perplexity. We evaluate two settings for Sequence FE: $\tau > 0.4$ and $\tau < 0.2$. For Spectrum FE, we use $\tau < 0.6$ because outliers occur only at low values.

Table 1: Downstream task performance in original RoPE and WeightedRoPE.

Model	HellaSwag		TruthfulQA		MMLU	
	baseline	+WeightedRoPE	baseline	+WeightedRoPE	baseline	+WeightedRoPE
Llama-4 17B	66.67	66.67	97.32	**97.99**	60.05	**60.81**
Llama-3 8B	60.16	60.16	84.85	84.85	34.21	34.21
Qwen3 8B	58.94	58.92	95.31	95.31	57.89	57.89
Gemma-2 9B	61.02	**61.21**	98.83	98.83	72.81	72.81

Results To examine the contribution of outliers, we plotted the perplexity for each weight α to visualize how different weighting values affect model performance. Figure 4 presents the overall perplexity results. When SpectrumFE reduced the dimensions with $\tau < 0.2$, perplexity increased as the weight α decreased. This indicates that dimensions with $\tau < 0.2$ in SpectrumFE contribute to model performance and may be important components for the model—in other words, frequency bands may be important. Furthermore, when SpectrumFE reduced dimensions with $\tau > 0.4$, perplexity decreased slightly as the weight α became smaller. However, the overall performance remained nearly identical, suggesting that dimensions with outlier values of $\tau > 0.4$ may be unnecessary or redundant for the model. Next, when SequenceFE reduced dimensions with $\tau < 0.6$, perplexity decreased slightly as the weight α became smaller. Notably, the decrease in perplexity was larger for the Llama-4 model than for the other models, suggesting that periodicity may be unnecessary or redundant for the model. The Llama-3 model had a smaller impact on perplexity, but the trend was the same as other models.

5.3 Downstream Task

Settings Based on the above experimental results, we hypothesize that the outlier dimensions of SpectrumFE for $\tau > 0.4$ and the periodic dimensions of SequenceFE for $\tau < 0.6$ are redundant. To investigate whether these dimensions affect downstream tasks, we evaluated Weighted RoPE across multiple tasks. We fixed α to 0.1 for Weighted RoPE, applying the weight α to both the outlier dimensions of SpectrumFE at $\tau > 0.4$ and the periodic dimensions of SequenceFE at $\tau < 0.6$. Performance was measured on a diverse set of benchmark tasks including HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), and MMLU (Hendrycks et al., 2021) to assess both reasoning and factual capabilities. [3]

Results Table 1 lists the experimental results. For all tasks, no significant difference was observed between RoPE, which performs no operations, and WeightedRoPE. Only the llama-4 model showed a slight improvement in performance. This suggests that the dimension where SpectrumFE becomes an outlier and the periodic dimension where SequenceFE decreases may not contribute to model performance and could be redundant.

[3] We additionally provide an evaluation of long-context generation on the Needle-in-a-Haystack task in Appendix E.

6 Conclusion

In this work, we introduced frequency entropy (FE), a metric that quantifies the effective utilization of each RoPE frequency dimension, and analyzed how the sinusoidal components of RoPE contribute to the model representation. SpectrumFE can identify the frequency band of RoPE, and reducing the contribution of this band degrades model performance, indicating it is a crucial component. Furthermore, SequenceFE can identify the periodic dimension of RoPE. Reducing the contribution of this dimension does not change model performance and may even improve it for some models. This suggests the periodic dimension may be redundant or unnecessary. Furthermore, these frequency bands and periodic dimensions do not exist in fixed dimensions. This suggests that some inconsistencies in prior work may stem from model-dependent frequency bands whose locations differ across heads and layers, rather than from absolute "low" or "high" frequency effects. Our framework provides a systematic method for interpreting such mixed features individually. Moreover, the fact that low-SequenceFE dimensions can be attenuated with minimal degradation indicates that these components reflect periodic signals induced by RoPE that the model does not functionally use. This finding highlights potential applications to RoPE-aware KV-cache compression and dimension pruning, and it offers a basis for future exploratory work on frequency-targeted architectural variants. We expect FE to function as a practical, model-independent diagnostic tool for positional encoding.

References

Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxZX20qFQ.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GtvuNrk58a.

Ting-Rui Chiang and Dani Yogatama. The rotary position embedding may cause dimension inefficiency in attention heads for long-distance retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13552–13562, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.697/.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context, 2024. URL https://arxiv.org/abs/2402.10171.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024a. URL https://arxiv.org/abs/2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024b. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes

Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher RoPE attention dimension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5877–5888, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.338/.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Drrl2gcjzl.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.229/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Accessed: 2025-09-11.

H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky. Spectral entropy based feature for robust asr. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. I–193, 2004.

Yui Oka, Taku Hasegawa, Kyosuke Nishida, and Kuniko Saito. Wavelet-based positional representation for long context. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OhauMUNW8T.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-4009/.

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Victor Sucic, Nicoletta Saulig, and Boualem Boashash. Analysis of local time-frequency entropy features for nonstationary signal components time supports detection. *Digital Signal Processing*, 34:56–66, 2014. URL https://www.sciencedirect.com/science/article/pii/S1051200414002292.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective longcontext scaling of foundation model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.260.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1472/.

A Additional Experiments on Key

We also compute our spectral entropy (SpectrumFE and SequenceFE) from the keys in `Llama-4-Scout-17B-16E-Instruct` (Meta, 2025) model. The analysis procedure is the same as in Section 4.

Analysis Results Figure 5 shows the scatter plots for each FE and the ℓ_2-norm heatmap for the keys. In RoPE layers, we observe both frequency bands and periodic dimensions in queries and in keys. FE aligns with these observations and shows the same trend for keys. Keys contain a larger number of periodic dimensions than queries (e.g. 5-th Layer). In NoPE layers, we see the same qualitative pattern as in queries: periodic dimensions are rarer, while frequency bands are detected frequently.

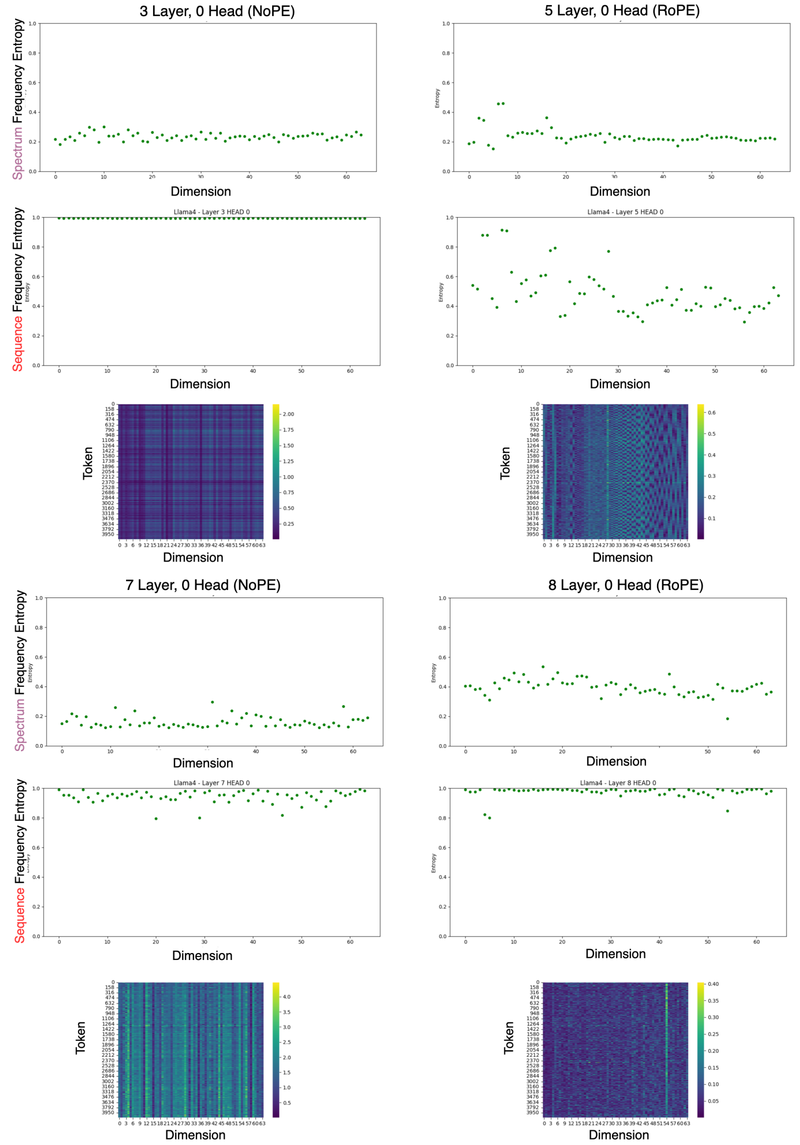


Figure 5: Scatter plots of each FE value in the `Llama-4-Scout-17B-16E-Instruct` model. Columns: layer 6, layer 1, layer 3 (left to right). Rows: SpectrumFE, SequenceFE, and key ℓ_2-norm map. Top/middle: pair index in RoPE j (x) vs. normalized entropy $\tilde{H}_j$ (y). Bottom: j (y) vs. token index n (x); color denotes $\|k_n^{(j)}\|_2$. All results are shown for head 0. The pair index j is rotation patterns. Sequence length $L = 4096$.

B Additional Experiments on All Layers

Figures 6 and 7 show the results of SpectrumFE and SequenceFE across all layers of the `Llama-4-Scout-17B-16E` model.

SpectrumFE in all layers First, we discuss the scatter plots for SpectrumFE in Fig. 6. Comparing the shallow RoPE layer and the NoPE layer, the NoPE layer exhibits a significantly broader distribution, whereas the RoPE layer's distribution is less extensive. However, as the layer depth increases, the NoPE layer's distribution converges. Even in the RoPE layer, the distribution converges, though not as much as in the NoPE layer, suggesting the influence of deepening. On the other hand, low SpectrumFE values are observed regardless of layer depth, indicating that the frequency band is observed in every layer.

SequenceFE in all layers Next, we discuss the scatter plots of SequenceFE in Fig. 7. Comparing the shallow RoPE layer and the NoPE layer, the RoPE layer exhibits a significantly broader distribution, while the NoPE layer's distribution is not as wide. This is exactly opposite to the trend observed in SpectrumFE. However, as the layer deepens, the NoPE layer's distribution widens, indicating that periodic dimensions temporarily emerge in the NoPE layer. Near the final layer, however, the periodic dimensions in the NoPE layer diminish. The RoPE layer maintains a certain number of periodic dimensions even at deeper layers. These periodic dimensions primarily exist near high frequencies. However, since these high-frequency periodic dimensions are not observed in the NoPE layer, NoPE may play a role in mitigating periodic dimensions.

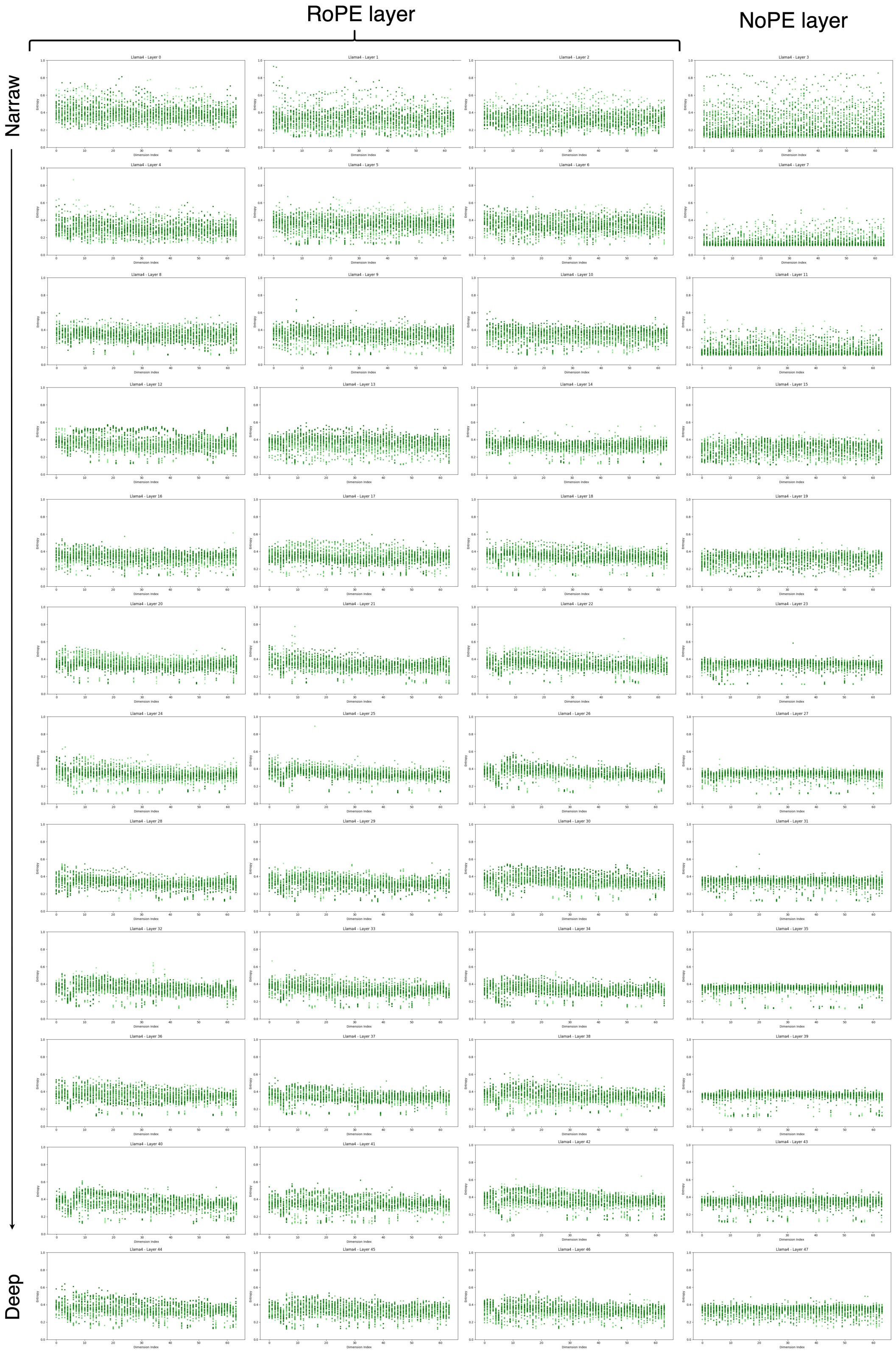


Figure 6: Layer-wise scatter plots of SpectrumFE across all attention heads in `Llama-4-Scout-17B-16E-Instruct`. The figure contains 48 panels arranged in 12 rows × 4 columns, with layer depth increasing left-to-right and then top-to-bottom (layers 0–47). Layers 3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, and 47 use NoPE; all other layers use RoPE.

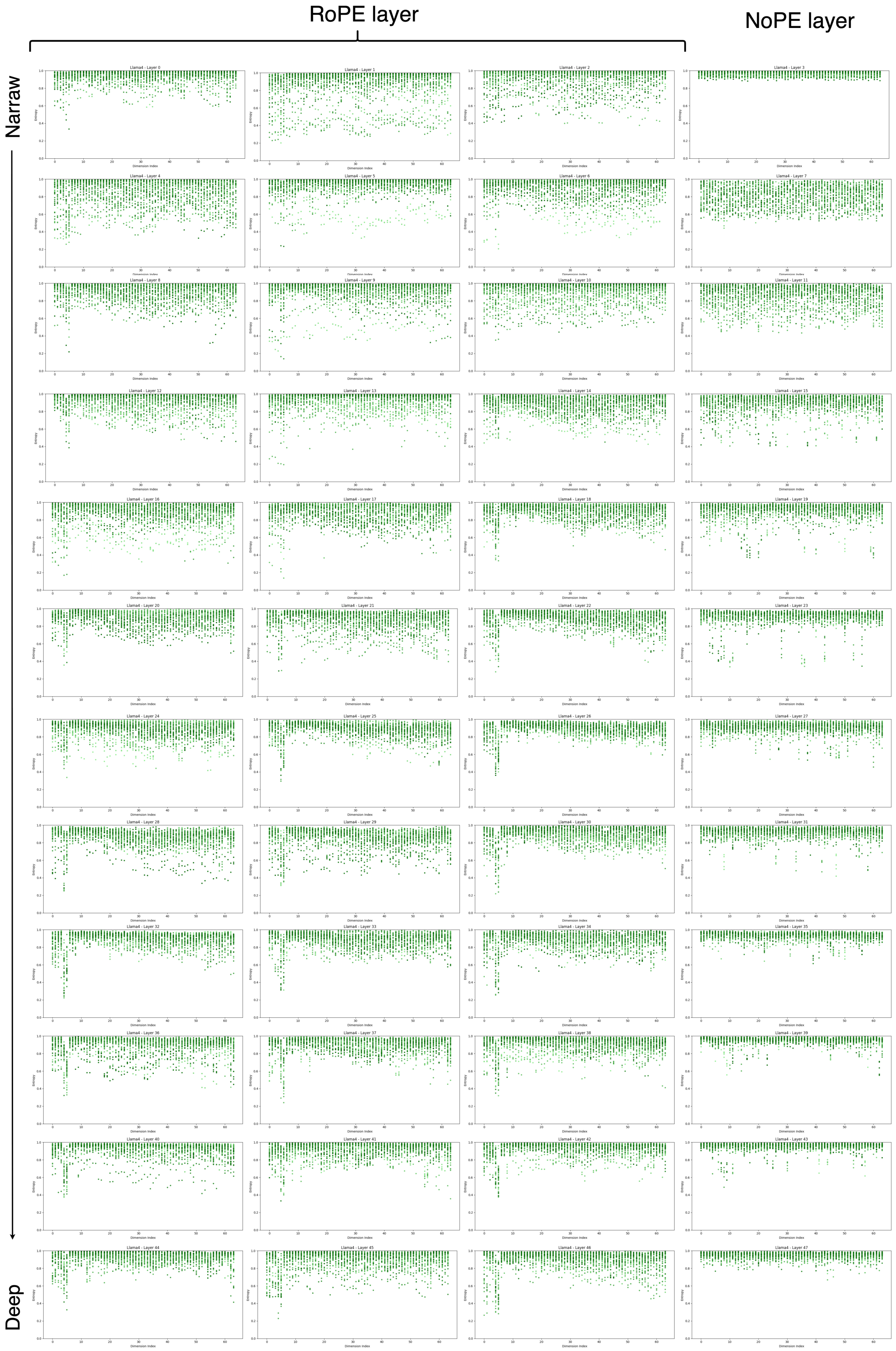


Figure 7: Layer-wise scatter plots of SequenceFE across all attention heads in `Llama-4-Scout-17B-16E-Instruct`. The figure contains 48 panels arranged in 12 rows × 4 columns, with layer depth increasing left-to-right and then top-to-bottom (layers 0–47). Layers 3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, and 47 use NoPE; all other layers use RoPE.

C Additional Experiments on Other Architectures

We performed frequency entropy analysis not only on the `Llama-4-Scout-17B-16E` model but also on `Meta-Llama-3-8B`, `gemma-2-9b-it`, and `Qwen3-8B`. Note that unlike Llama-4, these models do not possess a NoPE layer. All experimental settings are the same as in Section 4.

Llama-3 Figure 8 shows the FE analysis for the Llama-3 model. Only selected salient features are marked. First, consistent with Llama-4 in Section 4, frequency bands are also observed in Llama-3. These bands appear in most heads. In contrast, no periodic dimensions are identified, and SequenceFE remains consistently high overall. The strongest band typically appears between the 39th and 42nd dimensions on average, which differs from the band locations in the RoPE layers of Llama-4.

Gemma-2 Figure 9 shows the FE analysis for the Gemma-2 model. Frequency bands are observed and they appear in most heads. In contrast, no periodic dimensions are identified, and SequenceFE remains consistently high overall. The strongest band typically appears between the 115th and 120th dimensions on average.

Qwen-3 Figure 10 shows the FE analysis for the Qwen-3 model. Frequency bands are observed and they appear in most heads. In contrast, no periodic dimensions are identified, and SequenceFE remains consistently high overall. The strongest band typically appears between the 48th and 50th dimensions on average.

Frequency bands appear across models yet their locations are model dependent. In Llama-3, Gemma-2, and Qwen-3, bands occur in many heads while the peak dimension differs across models, which suggests that band position is governed by the RoPE base, the training length, and architectural factors such as dimensionality and head configuration. In these model, periodic dimensions are ineffective or at most very limited, since SequenceFE remains consistently high and clear periodic components are not observed. Attenuating or pruning these periodic dimensions is likely to cause only a small drop in performance, which is consistent with the downstream task results in Section 5.3. Moreover, since the band is detected in most heads and explicitly reducing the contribution of the band dimensions lowers performance (as shown in Section 5.2), we conclude that a specific frequency range is commonly useful for attention computation.

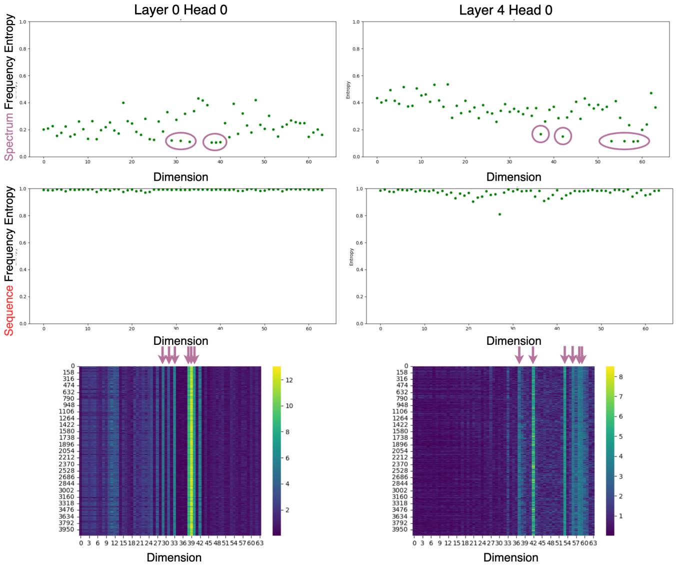


Figure 8: `Meta-Llama-3-8B` model (head 0). Columns: layer 0 and layer 4. Rows: SpectrumFE, SequenceFE, and query ℓ_2-norm map. Top/middle: pair index j (x) vs. normalized entropy $\tilde{H}_j$ (y). Bottom: j (y) vs. token index n (x); color denotes $\|q_n^{(j)}\|_2$. The pair index j is rotation patterns. Sequence length $L = 4096$.

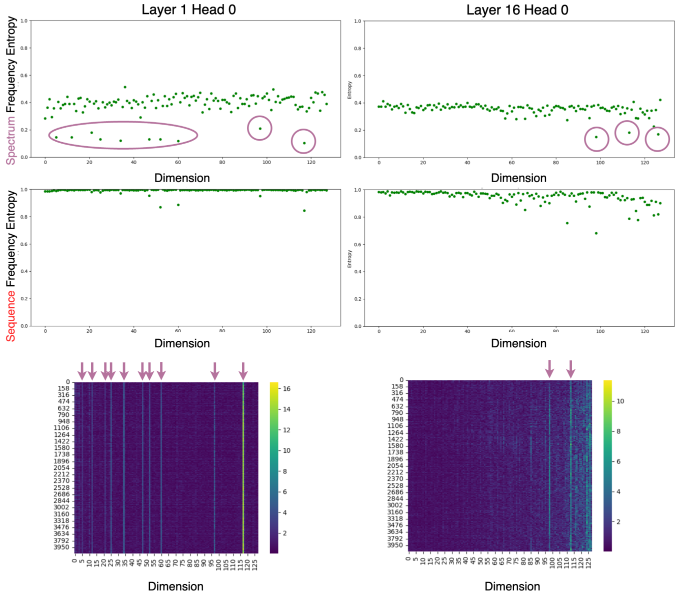


Figure 9: Gemma-2-9b-it model (head 0). Columns: layer 0 and layer 16. Rows: SpectrumFE, SequenceFE, and query ℓ_2-norm map. Top/middle: pair index j (x) vs. normalized entropy $\tilde{H}_j$ (y). Bottom: j (y) vs. token index n (x); color denotes $\|q_n^{(j)}\|_2$. The pair index j is rotation patterns. Sequence length $L = 4096$.

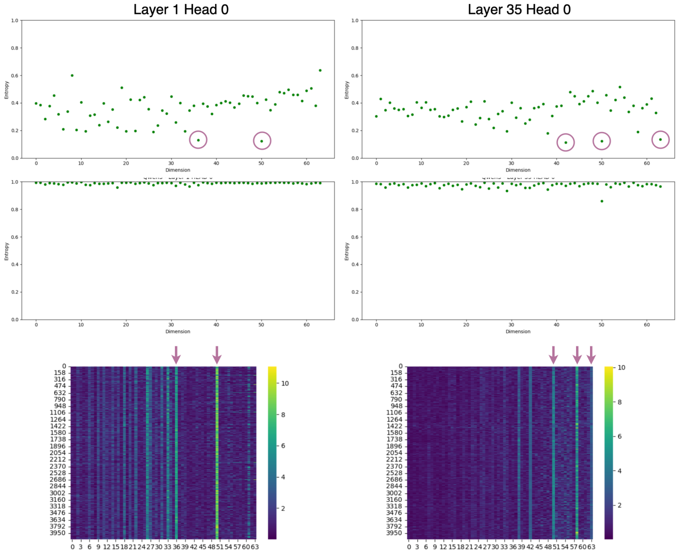


Figure 10: Qwen-3-8B model (head 0). Columns: layer 0 and layer 35. Rows: SpectrumFE, SequenceFE, and query ℓ_2-norm map. Top/middle: pair index j (x) vs. normalized entropy $\tilde{H}_j$ (y). Bottom: j (y) vs. token index n (x); color denotes $\|q_n^{(j)}\|_2$. The pair index j is rotation patterns. Sequence length $L = 4096$.

D Limitation

While our work provides a unified analysis of RoPE's frequency usage, several limitations remain. Our intervention weighted RoPE does not fully disentangle whether performance preservation under SequenceFE-based attenuation reflects true redundancy or compensation from neighboring frequency components. A more granular causal test, such as single-pair perturbations, layer-specific interventions, or head-wise isolated ablations, would be required to conclusively rule out compensatory mechanisms. Finally, although we evaluate four architectures applying RoPE in all layers, our causal claims are limited to the scope of inference-only interventions. Our results should therefore be interpreted as providing first-order causal evidence rather than a complete causal theory of positional encoding. We regard the development of more precise, minimally confounded causal perturbations as an important direction for future work.

E Analyzing Long-Context Behavior Through Weighted RoPE

We evaluated Weighted RoPE using downstream benchmarks such as HellaSwag and MMLU in Section 5.3. These tasks primarily test general knowledge and reasoning, and therefore may not be sensitive to differences in positional encoding. To directly address this concern, we evaluated the impact of Weighted RoPE on long-context understanding in a setting explicitly designed to probe positional robustness. We used the Needle-in-a-Haystack (NIAH) task [4], which is widely used in studies of positional interpolation and long-context evaluation. Given the available GPU memory, we evaluate Llama-4 up to 33,564 tokens and Llama-3 up to 59,615 tokens, which correspond to the maximum feasible context lengths under our inference setup.

E.1 Experimental Setup

We evaluate Llama-3 and Llama-4 under our Weighted RoPE intervention in the NIAH setup. The Weighted RoPE settings are the same as in Section 5.3. We used the implementation of (Fu et al., 2024).

E.2 Results

The results for Llama-3 are shown in Figures 11 and 12, and the results for Llama-4 are shown in Figures 13 and 14.

Consistent with the downstream results reported in Table 1, Llama-3 showed no observable differences between the baseline RoPE and Weighted RoPE across all evaluated context lengths. The retrieval accuracy curves as well as the heatmap structure were effectively identical. This suggests that the RoPE dimensions identified by SpectrumFE as outliers or by SequenceFE as low-periodicity contribute little to Llama-3's long-context retrieval. In other words, the dimensions we suppress or down-scale appear truly redundant for this model, both in short-range downstream tasks and in explicit long-context settings.

For Llama-4, the overall retrieval pattern remained stable, but Weighted RoPE showed slight improvement around the 30k-token context. In the baseline model, this region appears as a cluster of red cells indicating retrieval failure; the Weighted RoPE produces milder failure or partial recovery.

These results provides two insights. (1) Long-context behavior can be affected by localized frequency-band adjustments.Unlike Llama-3, Llama-4 exhibits more sharply defined frequency-band structures in early layers. Weighted RoPE interacts with these bands, and selectively attenuating low-importance rotary pairs can mitigate localized positional instability. (2) Weighted RoPE does not harm long-context performance and may correct fragile regions.Reviewer concerns about potential degradation on long-context–specific tasks are therefore addressed: no regressions were observed, and in Llama-4 certain failure modes were slightly improved.

[4] https://github.com/gkamradt/LLMTest_NeedleInAHaystack

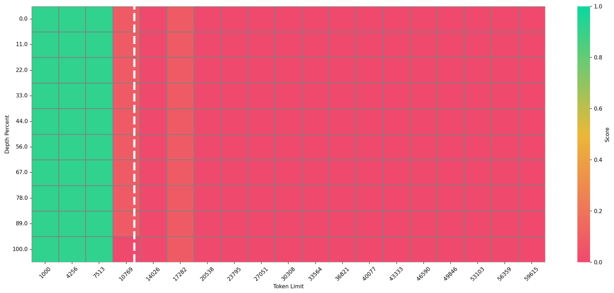


Figure 11: Results for NIAH on `Llama-3-8B`

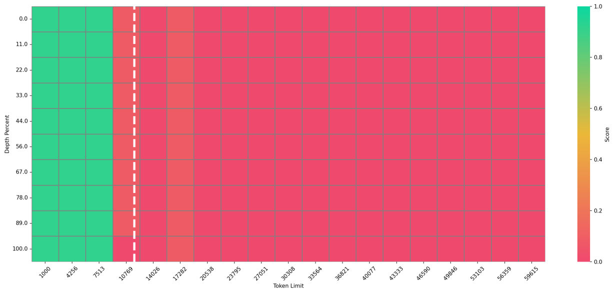


Figure 12: Results for NIAH task on `Llama-3-8B` with Weighted RoPE.

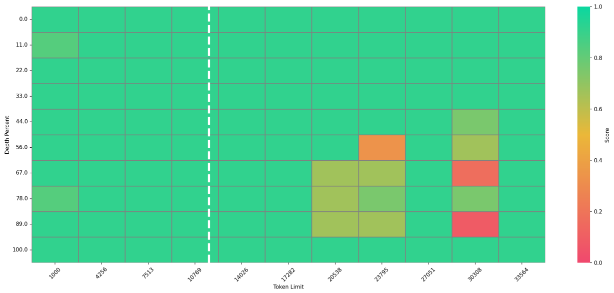


Figure 13: Results for NIAH task on `Llama-4-Scout-17B-16E-Instruct`.

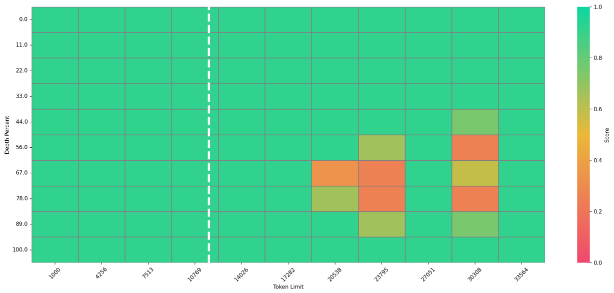


Figure 14: Results for NIAH task on `Llama-4-Scout-17B-16E-Instruct` with Weighted RoPE.

F Additional Experiments Across Datasets and Context Lengths

We extend the analysis via frequency entropy in Section 4 on two additional datasets and across multiple context lengths. In addition to Wikitext-103, we conduct experiments on the C4 dataset (Raffel et al., 2020) [5], which consists of English text from Common Crawl [6], and the PG19 book corpus (Rae et al., 2020). We evaluate models at three context lengths: 2048, 4096, and 8192 tokens.

F.1 Analysis Across Context Lengths

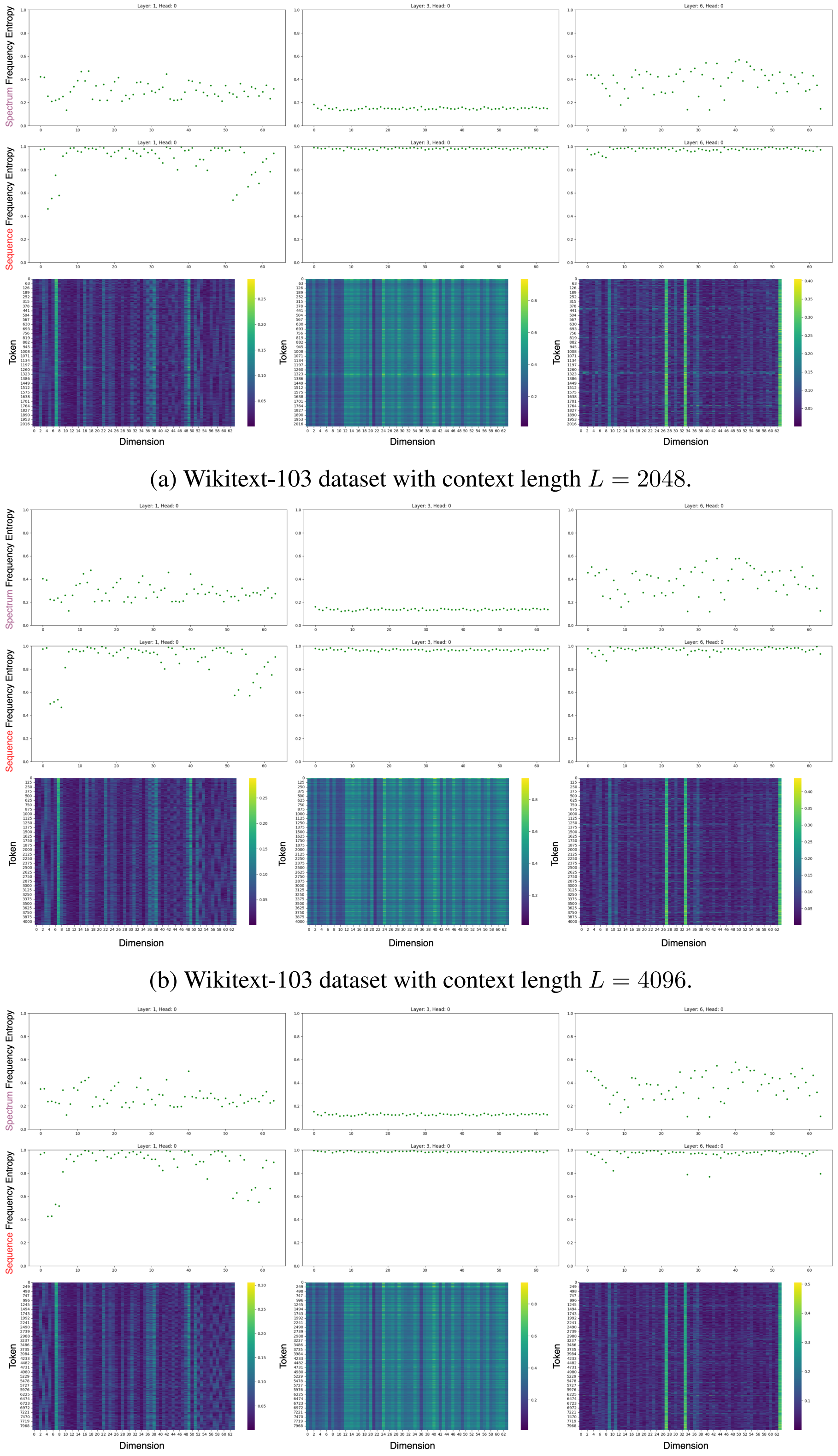
The results for Llama-4 using Wikitext-103 are shown in Figure 15. When we vary the context length, we observe slightly differences in SpectrumFE, SequenceFE, and the 2-norm maps. These changes are expected because the input text itself differs across context-length settings. However, the overall distribution of both SpectrumFE and SequenceFE remains largely unchanged. The frequency bands captured by SpectrumFE persist across all context lengths, and their locations remain stable. Similarly, the periodic patterns captured by SequenceFE also remain present at every context length, with their positions unchanged. As shown in the left 2-norm maps of Figure 15, the visual appearance of the periodic patterns changes as the context length varies, but the underlying dimensions where these patterns occur stay the same. Taken together, these findings indicate that both SpectrumFE and SequenceFE consistently can capture their respective structural properties, and these properties do not shift when the context length is changed.

F.2 Analysis Across Datasets

Figure 16 and Figure 17 show the Llama-4 results on C4 and PG19 respectively, evaluated at context lengths of 2048, 4096, and 8192 tokens. We observe no major differences across context lengths, and changing the dataset similarly leaves the overall distributions of SpectrumFE and SequenceFE largely unchanged. The frequency bands captured by SpectrumFE persist in both the C4 and PG19 datasets, and their locations remain stable. Similarly, the periodic patterns captured by SequenceFE continue to appear in both datasets, with their positions unchanged. Taken together, these findings indicate that both SpectrumFE and SequenceFE consistently can capture their respective structural properties, and these properties do not shift when the input sequence is changed.

[5] We use the 'en' validation split from the processed version provided at the following URL: https://huggingface.co/datasets/allenai/c4

[6] https://commoncrawl.org/

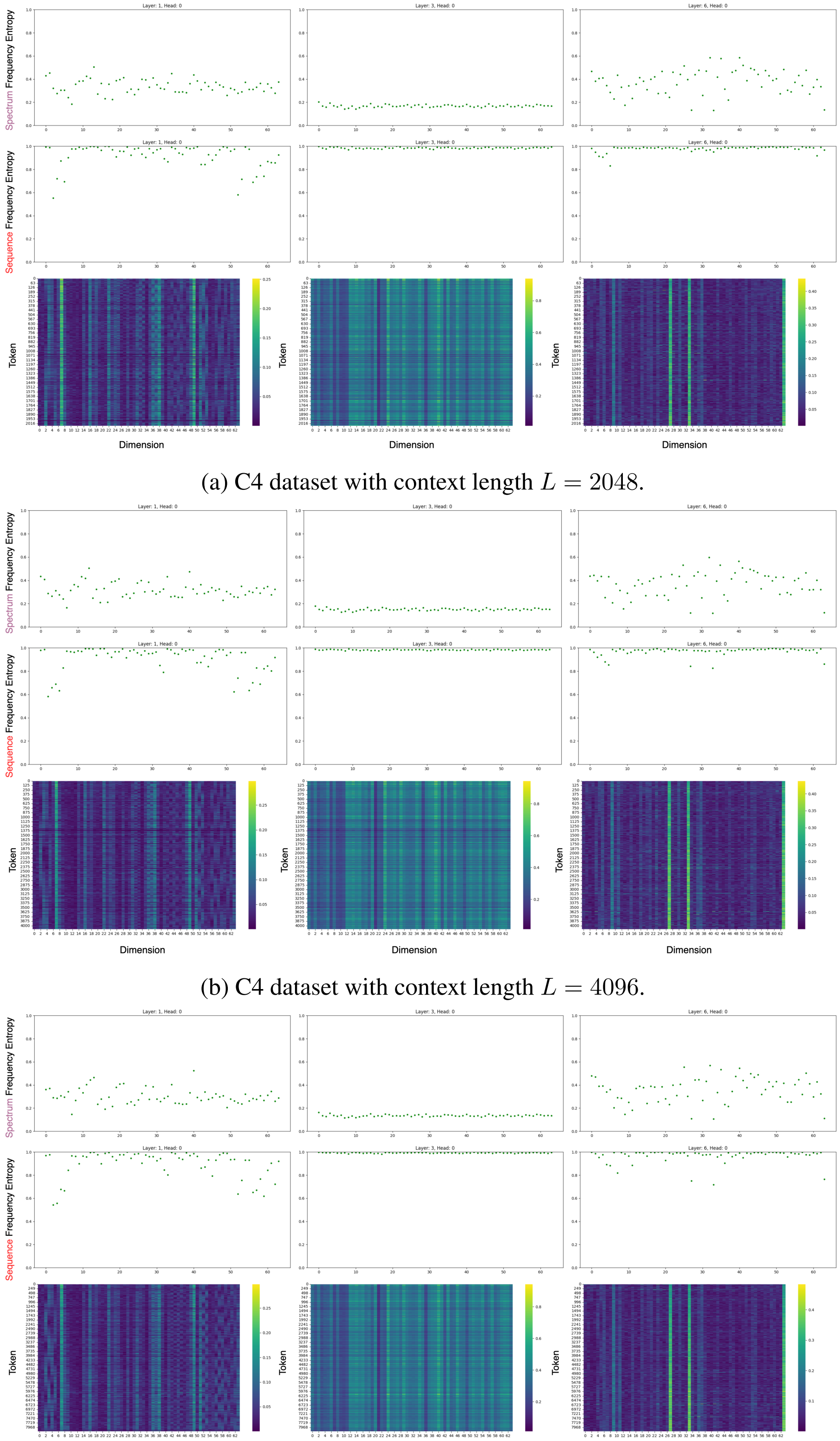


(a) Wikitext-103 dataset with context length $L = 2048$.

(b) Wikitext-103 dataset with context length $L = 4096$.

(c) Wikitext-103 dataset with context length $L = 8192$.

Figure 15: Scatter plots of each FE value in the `Llama-4-Scout-17B-16E-Instruct` model. Columns: layer 6, layer 1, layer 3 (left to right). Rows: SpectrumFE, SequenceFE, and query ℓ_2-norm map. All results are shown for head 0.

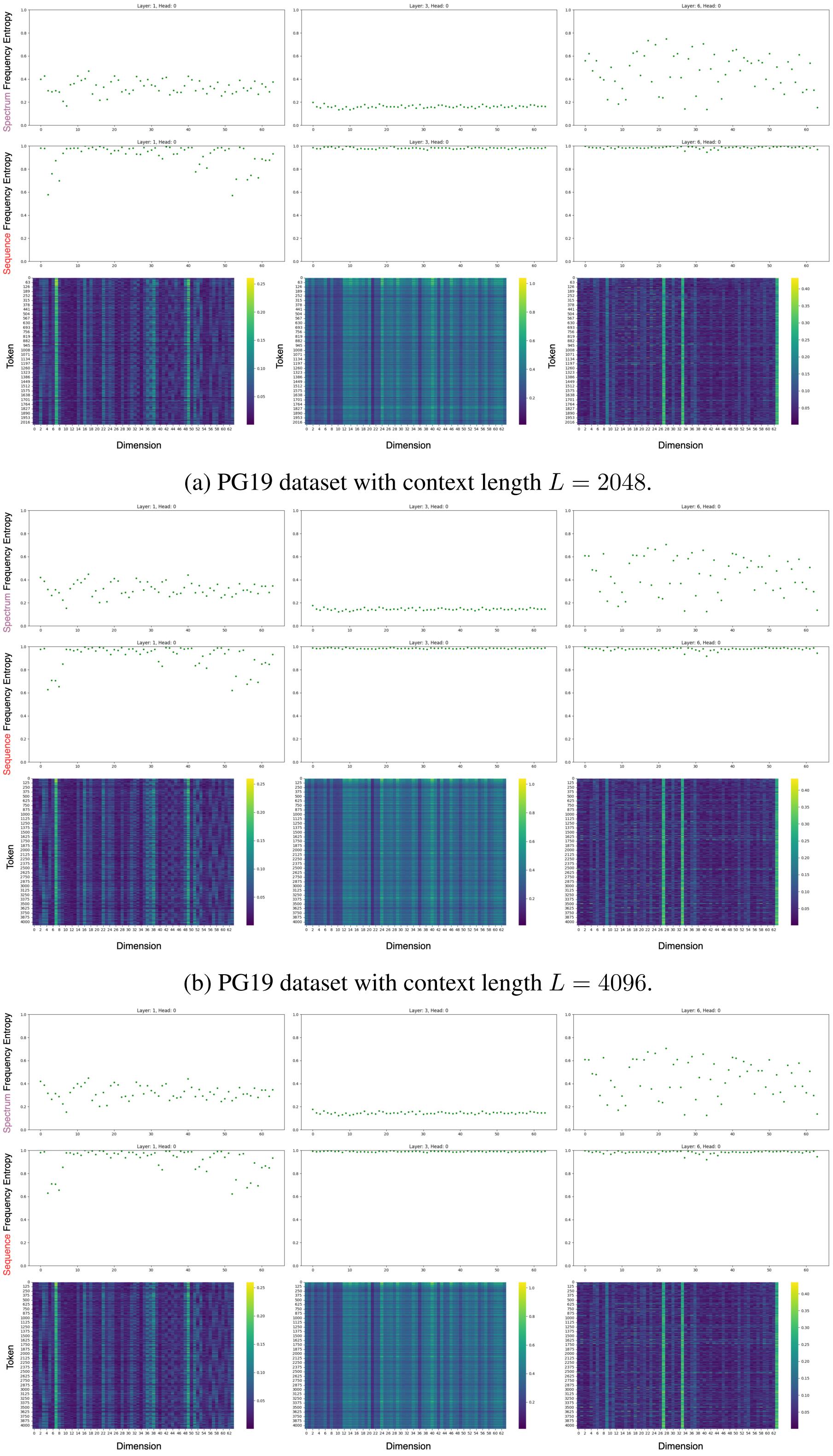


(a) C4 dataset with context length $L = 2048$.

(b) C4 dataset with context length $L = 4096$.

(c) C4 dataset with context length $L = 8192$.

Figure 16: Scatter plots of each FE value in the `Llama-4-Scout-17B-16E-Instruct` model. Columns: layer 6, layer 1, layer 3 (left to right). Rows: SpectrumFE, SequenceFE, and query ℓ_2-norm map. All results are shown for head 0.



(a) PG19 dataset with context length $L = 2048$.

(b) PG19 dataset with context length $L = 4096$.

(c) PG19 dataset with context length $L = 8192$.

Figure 17: Scatter plots of each FE value in the `Llama-4-Scout-17B-16E-Instruct` model. Columns: layer 6, layer 1, layer 3 (left to right). Rows: SpectrumFE, SequenceFE, and query ℓ_2-norm map. All results are shown for head 0.

G Frequency Entropy Comparison Between RoPE-Only and NoPE-Only Models

Llama-4 interleaves RoPE and NoPE layers, causing the effects of the two positional schemes to be mixed within the same network. To disentangle their individual contributions, we pretrain two small models from scratch: one that uses only RoPE and another that uses only NoPE. We then evaluate their layer-wise Frequency Entropy (FE) to examine the independent behavior induced by each encoding.

G.1 Experimental Setup

For pre-training from scratch, we perform a comparative evaluation with a Transformer-based language model (Baevski & Auli, 2019). The dimensionality of the word embedding d_{model} is 1024, the number of heads N is 8, the dimensionality of the heads d is 128, and the number of layers is 16. This implementation used the fairseq (Ott et al., 2019)-based code provided in a previous work(Press et al., 2022), and all hyperparameters were set to the same values as those in the literature(Press et al., 2022). We use the Nesterov's Accelerated Gradient (NAG) optimizer with momentum 0.99, following the Fairseq nag implementation in (Ott et al., 2019). The learning rate schedule is a cosine scheduler: it is initialized at 1e-7, linearly warmed up to 1.0 during the first 16,000 updates, and then decayed with a cosine schedule down to 1e-4. The value 9216 denotes the maximum number of tokens per batch per GPU, with a sequence length of 512 tokens. We train for 286,000 updates, which corresponds to approximately 205 epochs on WikiText-103. We used the WikiText-103 dataset (Merity et al., 2017), which consists of over 103 million tokens of English Wikipedia articles. This setup used in (Press et al., 2022; Oka et al., 2025) [7], where the goal is to analyze structural differences in positional encodings.

G.2 Results

Figure 18 reports the FE measurements for layer 0 of the RoPE-only model, and Figure 19 reports the results for its final layer, layer 15. Figures 20 and 21 provide the corresponding FE results for the NoPE-only model.

First, in the RoPE-only model, the SpectrumFE results in the top row reveal clear frequency bands. The locations of these bands vary across heads and layers, but the band structure itself is consistently present in every head and layer we examined. This behavior matches what we observed in the RoPE layers of Llama-4, as well as in Llama-3, Gemma, and Qwen. In contrast, SequenceFE remains high across all dimensions, indicating little periodic structure. This differs from Llama-4 but aligns with the RoPE-only behavior reported for Llama-3, Gemma, and Qwen in Appendix C.

Next, in the NoPE-only model, the SpectrumFE results show that frequency bands do not appear in all heads. Overall SpectrumFE values are lower than in the RoPE-only model, and the number of dimensions forming identifiable bands is also smaller. Some heads show no band structure at all. For example, the right-side in Figure 18 (0 Layer, 7 Head) displays a pattern closer to noise in its 2-norm map. Moreover, in the final layer, the frequency bands disappear entirely. SequenceFE remains high throughout, showing no periodic patterns. This is expected because NoPE does not introduce any periodic information.

These observations suggest that a NoPE-only model tends to weaken or wash out frequency bands as depth increases. RoPE, in contrast, consistently maintains band structure even in deeper layers. Finally, in architectures like Llama-4 that interleave RoPE and NoPE layers, NoPE may reinforce the band structure introduced by RoPE while also amplifying the periodic components associated with RoPE. This interaction between the two positional schemes may explain the distinct FE patterns observed in mixed-layer settings.

[7] https://github.com/ofirpress/attention_with_linear_biases

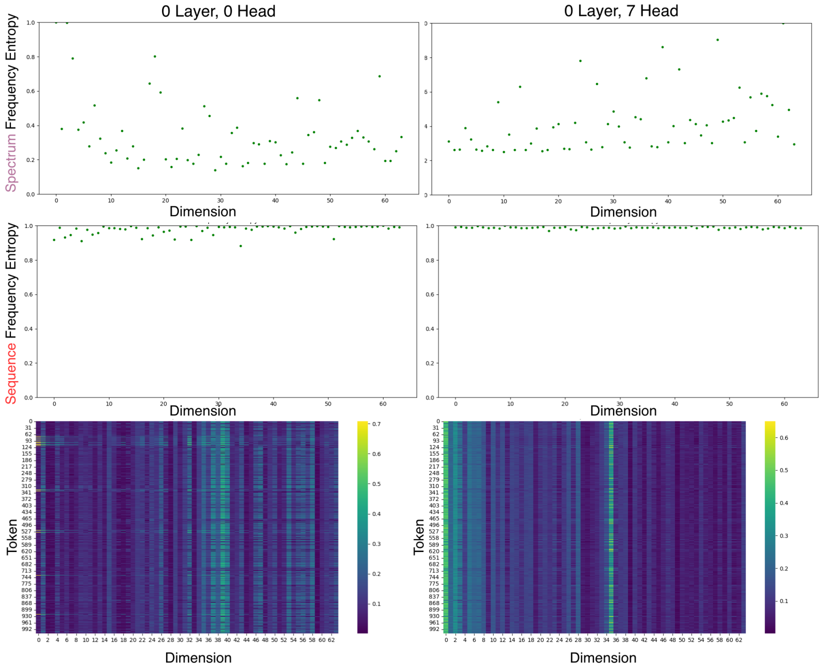


Figure 18: Frequency Entropy (FE) scatter plots for a **RoPE-only** model pretrained from scratch. Shown are the 0th and 7th heads of **layer 0**. Rows depict SpectrumFE, SequenceFE, and the query ℓ_2-norm map.

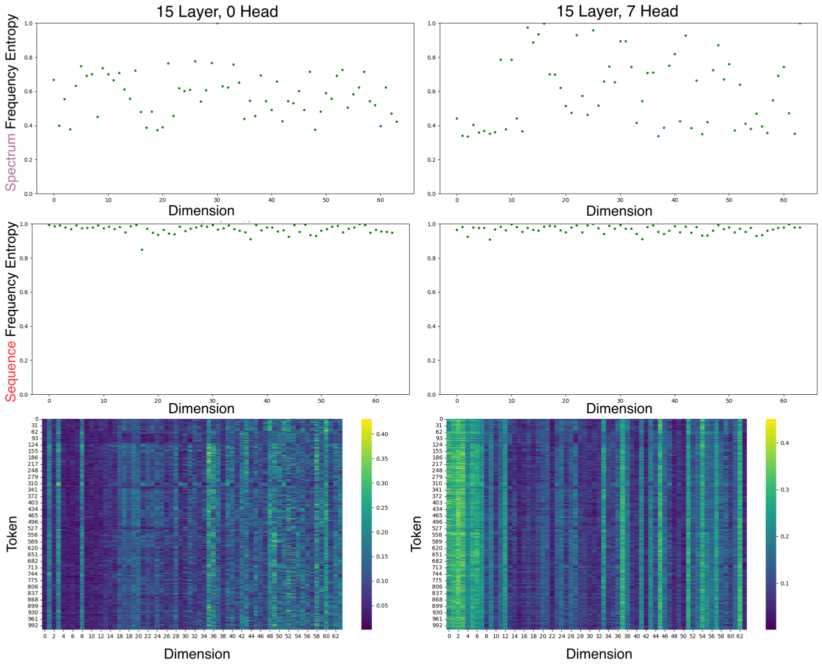


Figure 19: Frequency Entropy (FE) scatter plots for a **RoPE-only** model pretrained from scratch. Shown are the 0th and 7th heads of **last layer**. Rows depict SpectrumFE, SequenceFE, and the query ℓ_2-norm map.

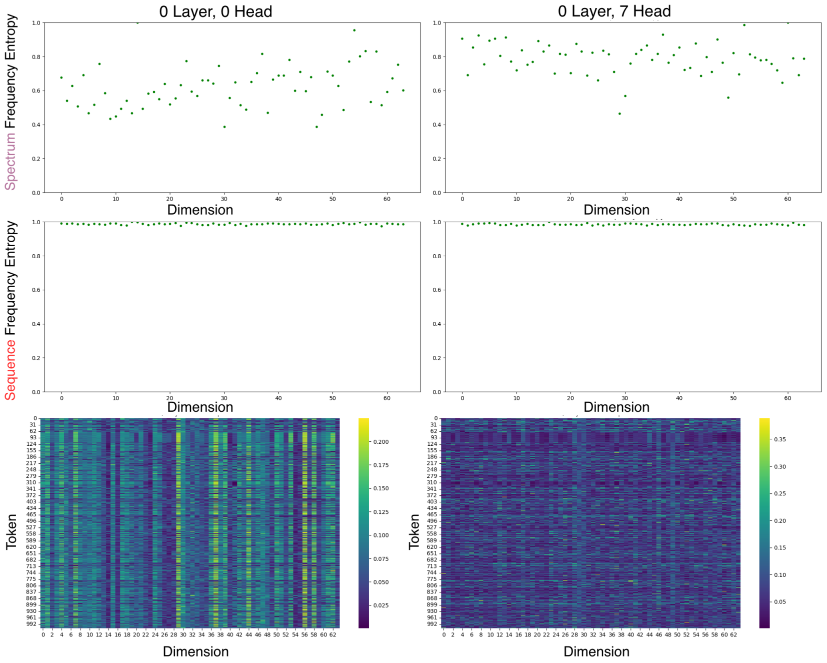


Figure 20: Frequency Entropy (FE) scatter plots for a **NoPE-only** model pretrained from scratch. Shown are the 0th and 7th heads of **layer 0**. Rows depict SpectrumFE, SequenceFE, and the query ℓ_2-norm map.

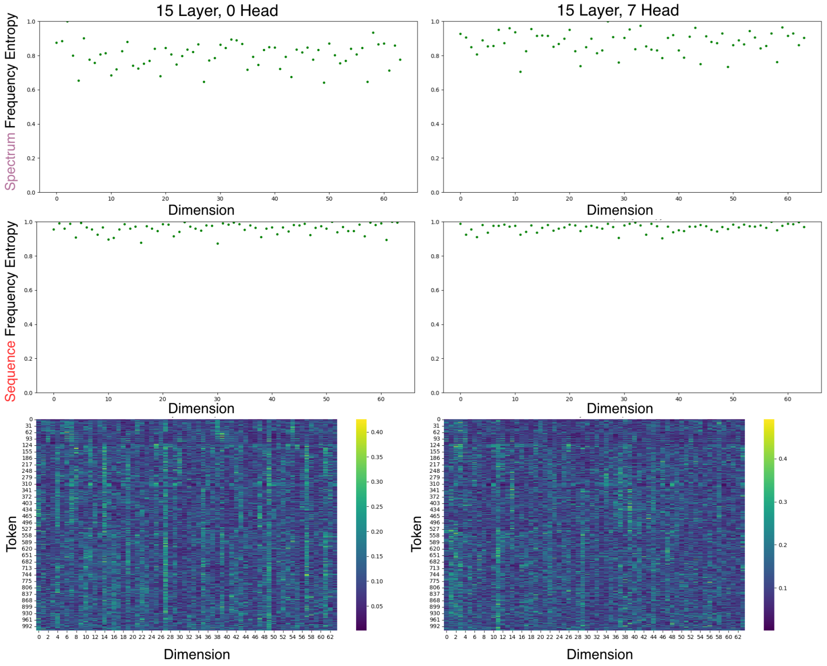


Figure 21: Frequency Entropy (FE) scatter plots for a **NoPE-only** model pretrained from scratch. Shown are the 0th and 7th heads of **last layer**. Rows depict SpectrumFE, SequenceFE, and the query ℓ_2-norm map.