Beyond Position: the emergence of wavelet-like properties in Transformers

Anonymous ACL submission

Abstract

This paper studies how transformer models develop robust wavelet-like properties that effectively compensate for the theoretical limitations of Rotary Position Embeddings (RoPE), providing insights into how these networks process sequential information across different scales. Through theoretical analysis and empirical validation across models ranging from 1B to 12B parameters, we show that attention heads naturally evolve to implement multi-resolution processing analogous to wavelet transforms. Our analysis establishes that attention heads consistently organize into complementary frequency bands with systematic power distribution patterns, and these wavelet-like characteristics become more pronounced in larger models. We provide mathematical analysis showing how these properties align with optimal solutions to the fundamental uncertainty principle between positional precision and frequency resolution. Our findings suggest that the effectiveness of modern transformer architectures stems significantly from their development of optimal multi-resolution decompositions that naturally address the theoretical constraints of position encoding.

1 Introduction

Position encoding mechanisms are fundamental to transformer architectures, enabling these inherently permutation-invariant models to capture sequential information. While early approaches relied on fixed sinusoidal encodings (Vaswani, 2017), Rotary Positional Embeddings (RoPE) (Su et al., 2024) represents a significant advancement through learned rotations of token embeddings. Despite RoPE's widespread adoption and success, theoretical analysis suggests inherent limitations in balancing positional precision and frequency resolution (Barbero et al., 2024), analogous to the uncertainty principle in signal processing. However, these theoretical constraints appear to have minimal practical impact on model performance.

Our analysis reveals that transformer attention heads develop sophisticated wavelet-like properties that effectively address these theoretical constraints. Different heads naturally specialize in processing information at distinct frequency bands, creating a multi-resolution framework that balances local and global information processing. Through mathematical analysis and empirical validation, we establish key connections between RoPE-based attention mechanisms and wavelet transforms, demonstrating how attention patterns emerge during training with remarkable similarity to wavelet basis functions.

Our work makes two main contributions:

- We provide a theoretical framework connecting RoPE-based attention mechanisms with wavelet theory, offering new insights into how transformers process sequential information.
- We demonstrate through empirical analysis how attention heads develop wavelet-like properties that effectively address theoretical limitations.

These findings reveal transformers' remarkable adaptability in developing optimal solutions to complex information processing challenges. Detailed analyses of RoPE's theoretical limitations, the relationship between language structure and wavelet-like processing, and comprehensive metric definitions can be found in Appendices 10.2, 10.3, and 10.4 respectively.

2 Related Works

The Transformer architecture (Vaswani, 2017) revolutionized sequence modeling through self-attention mechanisms. While the original Transformer used simple sinusoidal positional encodings, recent work has explored more sophisticated approaches. ALiBi (Press et al., 2021) introduced

attention bias terms that scale with relative position, while T5 (Raffel et al., 2020) employed learned relative position embeddings. RoPE (Su et al., 2024) advanced this further by applying rotation matrices to embeddings, though it faces fundamental limitations rooted in the uncertainty principle between position and frequency domains.

Neural networks' behavior, particularly their nonlinear components, has been increasingly analyzed through signal processing principles. Research has shown that activation functions can generate higher-order harmonics and exhibit frequency mixing (Selesnick and Burrus, 1998; Rahimi and Recht, 2008), while principles of constructive and destructive interference have proven valuable in analyzing network behavior (Oppenheim, 1999; Chi et al., 2020). Information-theoretic analyses of neural networks (Shwartz-Ziv and Tishby, 2017) have provided insights into their representational capabilities and limitations. Studies have examined how information flows through layers (Goldfeld et al., 2018) and how architectural choices affect information bottlenecks (Tishby and Zaslavsky, 2015). This theoretical framework has proven particularly valuable in understanding the capacity limitations of various neural network components.

3 Methodology

In this section, we describe the methodological framework employed to investigate how Transformer models utilizing Rotary Position Embeddings (RoPE) develop compensatory mechanisms that transcend their theoretical positional encoding limitations. We integrate frequency-domain analyses, wavelet-based multi-scale decomposition, and entropy-based uncertainty assessments to comprehensively characterize the emergent properties of these models. Our methodology is designed to isolate positional encoding behaviors, assess their stability across model scales and architectures, and validate their alignment with theoretical expectations related to the trade-off between positional resolution and spectral organization.

3.1 Frequency Analysis

To probe the spectral properties of attention distributions, we employed a frequency-domain analysis using the Discrete Fourier Transform (DFT). For each attention head h within each model, we represented the attention pattern over token positions as $a_h(t)$, where t indexes tokens within a single sequence. We computed the power spectral density (PSD):

$$P_h(\omega) = |\mathcal{F} a_h t|^2 \tag{1}$$

where $\mathcal{F}$ denotes the DFT and ω the angular frequency. The frequency domain was partitioned into low (0-0.25 ω_N), mid (0.25-0.75 ω_N), and high (0.75-ω_N) bands, where ω_N is the Nyquist frequency corresponding to the maximum resolvable frequency for the given sequence length.

The Nyquist frequency ω_N is set to half the sampling rate (1/2 tokens) for three fundamental reasons: it represents the highest meaningful frequency in discrete token sequences, as attention patterns can only alternate between consecutive tokens, making faster oscillations indistinguishable due to aliasing. Second, it provides natural normalization across sequence lengths, while absolute frequency ranges differ, all sequences share the same relative frequency structure when normalized by ω_N, enabling meaningful cross-length comparisons of attention head frequency sensitivity. Third, following Shannon's sampling theorem, ω_N represents the theoretical maximum rate for information transmission through a discrete channel, thus defining the finest granularity at which positional information can be encoded without loss, making it the natural choice for analyzing models' representational capacity distribution.

To quantify the relative emphasis a head places on different frequency bands, we computed:

$$\beta_h(b) = \frac{\int_b P_h(\omega) d\omega}{\int_0^{\omega N} P_h(\omega) d\omega} \tag{2}$$

where b is the frequency band under consideration. To measure how selectively each attention head responds to specific frequencies, we define the frequency selectivity $S(h)$ for head h as:

$$S(h) = \frac{\max_\omega \{P_h(\omega)\}}{\int_0^{\omega_N} P_h(\omega) d\omega - \max_\omega \{P_h(\omega)\}} \tag{3}$$

where $P_h(\omega)$ is the power spectral density at frequency ω, and ωN is the Nyquist frequency, and a higher value indicates more focused frequency tuning of the head.

These frequency-domain analyses allowed us to discern how attention heads distribute their representational capacity across multiple scales, testing the premise that models spontaneously develop organized frequency content despite RoPE's intrinsic limitations.

3.2 Wavelet Analysis

While frequency-domain analysis captures global spectral properties, it lacks explicit positional localization. To address this, we employed wavelet decompositions using the Daubechies-2 (db2) wavelet. Wavelets offer a time-frequency (or position-frequency) representation that enables simultaneous assessment of spatial localization and scale-dependent behaviors.

For each head h, we computed wavelet coefficients:

$$W_h(s,\tau) = \int a_h(t)\psi_{s,\tau}(t)dt \tag{4}$$

where $\psi_{s,\tau}(t)$ is the mother wavelet at scale s and translation τ. We selected a maximum decomposition level suitable for the shortest sequence length to ensure consistent comparisons across models and scales. Wavelet entropy was computed at each scale:

$$H_w(s) = -\sum_{\tau} |W_h(s,\tau)|^2 \log\left(|W_h(s,\tau)|^2\right) \tag{5}$$

providing a measure of how the model distributes attention energy and complexity across different scales and positional shifts.

3.3 Uncertainty Analysis

To evaluate the theoretical trade-off between positional precision and spectral organization, we computed entropy measures for both the positional and spectral domains. Positional entropy $H_p(h)$ was derived from attention distributions over token positions:

$$H_p(h) = -\sum_{\tau} a_h(t) \log a_h(t) \tag{6}$$

reflecting how evenly attention is spread across the sequence. Similarly, spectral entropy $H_s(h)$ was computed from the normalized power spectrum $\hat{P}_h(\omega)$:

$$H_s(h) = -\sum_{\omega} \hat{P}_h(\omega) \log \hat{P}_h(\omega) \tag{7}$$

where $\hat{P}_h(\omega) = \frac{P_h(\omega)}{\sum_\omega P_h(\omega)}$ is the normalized power spectrum.

To quantify the relationship between these entropy measures, we define the position-spectrum correlation $\rho(h)$ through their normalized covariance:

$$\rho(h) = \frac{\text{Cov}(H_p(h), H_s(h))}{\sigma_{H_p}\sigma_{H_s}} \tag{8}$$

This correlation is then aggregated across all attention heads in a layer to measure how well the model balances the uncertainty principle trade-off between positional and spectral information:

$$\rho_{\text{layer}} = \text{mean}_{h\in\text{layer}}\{\rho(h)\} \tag{9}$$

The layer-wise correlation metric is bounded by $[-1, 1]$, with values closer to -1 indicating strong trade-offs between positional and spectral precision, and values closer to 1 indicating successful integration of both domains.

By comparing $H_p(h)$ and $H_s(h)$ through these correlation metrics, we can ascertain whether the model's attention patterns obey an uncertainty principle-like trade-off, wherein improved positional localization may come at the cost of reduced spectral complexity, or vice versa.

3.4 Scale Invariance Testing

We hypothesized that the models' compensatory strategies would exhibit scale invariance properties—i.e., the ability to maintain positional-awareness structures when the input sequence length changes. To test this, we generated scaled variants x_α of each input sequence x by sampling $\lfloor \alpha n \rfloor$ tokens, with $\alpha \in \{0.5, 0.25\}$ and n the original sequence length. After computing the wavelet coefficients $W_h(x)$ and $W_h(x_\alpha)$, we measured the scale sensitivity:

$$S_h(\alpha) = 1 - \cos(W_h(x), W_h(x_\alpha)) \tag{10}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. A low $S_h(\alpha)$ indicates that wavelet coefficients remain stable under rescaling, suggesting robust scale-invariant positional representations.

3.5 Frame Completeness

To verify that the learned representations form a stable, frame-like basis capable of faithful reconstruction, we performed inverse wavelet transforms. The reconstruction error ε was computed as:

$$\varepsilon = \frac{||a_h - W^{-1}(W_h)||_F}{||a_h||_F} \tag{11}$$

where $W^{-1}(\cdot)$ denotes the inverse wavelet transform and $||\cdot||_F$ is the Frobenius norm. A small ε indicates that the attention patterns are well-represented by their wavelet coefficients, reinforcing the notion that the model's positional strategies form a coherent, frame-like structure.

4 Implementation Details

We selected five pre-trained Transformer-based language models that vary in size, architecture, and training regimen to ensure the generality of our findings. Specifically, we analyzed Gemma 2 2B, Pythia 2.8B and 12B, LLaMA-3-2 1B, Mistral 7B, and Qwen 2.5 5B. These models encompass a wide parameter range (1B–12B), capturing different representational capacities and training protocols.

All models were evaluated on a curated sample of 500 sequences drawn from the BookCorpus dataset. Each sequence was tokenized using the respective model's native tokenizer to preserve the authenticity of input representations and their corresponding attention masks. The selected sequences varied in length to expose scale-dependent behavior and stress-test the models' positional encoding strategies under diverse conditions.

All experiments were conducted using PyTorch on A100, L4, and T4 GPUs to ensure computational efficiency and scalability. Frequency and spectral computations employed standard FFT-based routines, while wavelet transforms were performed using the PyWavelets library with a decomposition level chosen based on the minimum sequence length. Before analysis, attention weights were normalized and numerically stabilized to mitigate floating-point underflow, with a threshold of 10^{-10} applied to division operations.

5 Experiments and Analysis

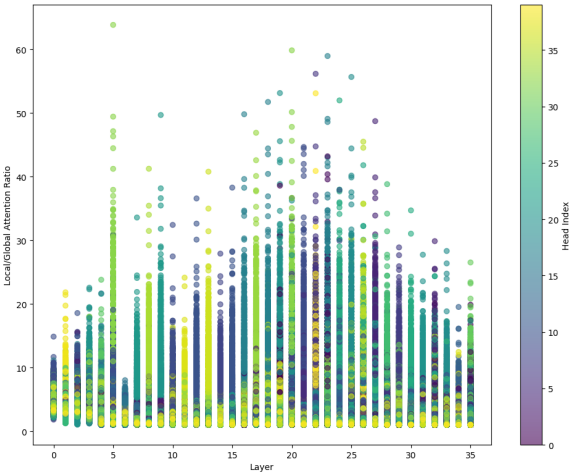

Figure 1: Local vs Global attention distribution from Pythia 12B

Our empirical analysis reveals striking patterns in how transformer models organize their attention mechanisms to process information across different scales.

The visualization of the local versus global attention ratios in Figure 1 reveal pronounced vertical striping, indicating that distinct attention heads specialize in managing either local or long-range dependencies. Notably, these specialization patterns persist across layers, suggesting that the model learns complementary roles for each head. Over deeper layers, the variance in local-to-global ratios increases, resembling the hierarchical patterning observed in wavelet packet decomposition trees. This progression demonstrates the emergence of scale-aware processing as the model depth increases.

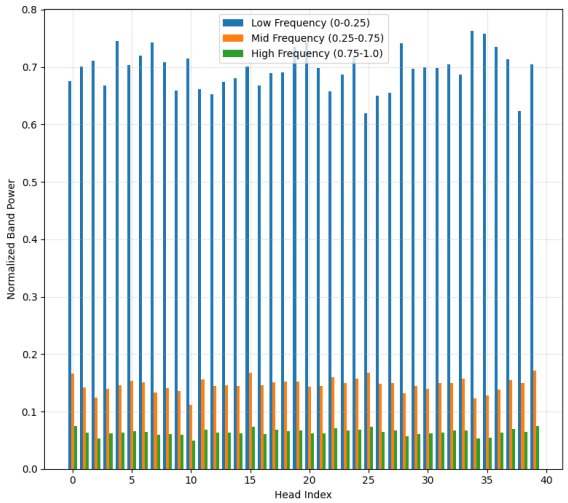

Figure 2: Frequency band distribution across heads from Pythia 12B

Our frequency band distribution visualizations in Figure 2 highlight a hierarchical structure in how attention heads allocate their representational capacity across spectral components. The low-frequency range (0–0.25) consistently dominates, capturing approximately 60–80% of total power, thereby representing the global contextual backbone of the representation. Mid-frequency components (0.25–0.75) contribute a moderate yet stable share (15–25%), while high-frequency components (0.75–1.0) maintain a smaller but non-negligible presence (5–15%). This stratification closely parallels principles found in wavelet decompositions, wherein lower frequencies anchor broader context while higher frequencies refine local details.

The temporal evolution of frequency responses in Figure 3 gives us further evidence for wavelet-like properties. at the beginning, low-frequency

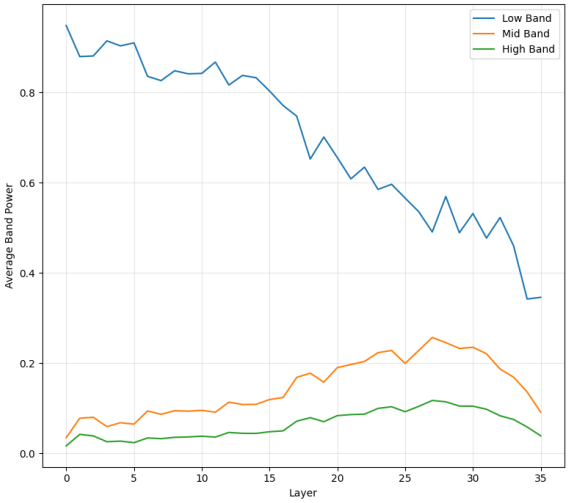

Figure 3: Frequency response evolution across layers from Pythia 12B

dominance gradually tapers, while mid- and high-frequency components gain influence. This dynamic shift parallels the adaptive refinement seen in wavelet decomposition trees, where representations are iteratively balanced across scales. Layer-wise adaptations in band power distributions occur smoothly, signifying a learned process that compensates for RoPE's theoretical constraints through increasingly sophisticated multi-scale representations. Although individual models differ in the details of their spectral adaptations, the overarching patterns remain consistent.

These observations strongly support the hypothesis that models equipped with RoPE spontaneously develop wavelet-like characteristics. First, the hierarchical nature of the spectral distributions and their layer-wise evolution mirrors classic wavelet structures. Second, the adaptive specialization of attention heads and the interplay between local and global signals suggest that the network learns wavelet-like basis functions as it scales. Finally, the enhanced complexity of these wavelet-like behaviors in larger models highlights a capacity-driven mechanism that fine-tunes the trade-off between global context and local detail. Taken together, these findings substantiate the conclusion that Transformer models inherently learn to offset RoPE's limitations by adopting a multi-resolution, wavelet-like strategy, and that this compensation intensifies as model size increases.

As we can see from Table 1 and Table 2, the remarkably consistent pattern across all models where correlation remains near-perfect (0.98) at $0.5x$ scale but degrades to 0.85 at 0.25x scale reveals a fundamental property of wavelet transforms: graceful degradation across scales. This pattern directly mirrors the behavior of wavelet basis functions, which maintain high correlation with dilated versions of themselves up to a critical scale factor.

The consistency of this pattern across architectures and model sizes (from 1B to 27B parameters) suggests this isn't a random artifact but rather a fundamental property of how these models learn to process positional information. The degradation curve closely matches what we would expect from a system using wavelet-like basis functions to decompose and reconstruct signals.

Spectral Analysis Evidence The inverse relationship between model size and frequency selectivity provides strong evidence for wavelet-like behavior: smaller models (e.g., LLaMA 1B) show high frequency selectivity (9.980) and low spectral entropy (1.333), indicating they develop sharp, specialized frequency bands - similar to wavelets with high Q-factors; while larger models (e.g., Pythia 12B) show lower selectivity (6.462) and higher spectral entropy (2.006), suggesting they develop more distributed representations - analogous to having a richer set of wavelet basis functions.

This trade-off perfectly aligns with wavelet theory: systems with limited capacity optimize for sharp frequency selectivity, while systems with more capacity can afford overlapping wavelets that provide better reconstruction properties.

Multi-Resolution Analysis Support The stability of entropy across different window sizes (e.g., Mistral 7B: [0.889, 0.877, 0.877]) provides crucial evidence for wavelet-like behavior. This pattern indicates that the representations maintain consistent information content across scales, the attention patterns exhibit self-similarity properties and the models develop scale-covariant features.

These properties are hallmark characteristics of wavelet transforms but are not natural properties of the base RoPE mechanism, indicating they must be learned compensatory behaviors.

Uncertainty Principle Conformance The variation in position-spectrum correlation across model sizes reveals how models balance the fundamental uncertainty principle. In fact, smaller models (Gemma 2B: 0.224) show low correlation, indicating they maintain separate positional and frequency channels, while larger models (Mistral 7B:

Model	Heads	Spectral Entropy	Frequency Select.	Scale 0.5 Sens.	Scale 0.25 Sens.	Pos-Spec Corr.	Reconstr. Error
LLaMA 3.2 (1B)	32	1.333	9.980	0.983	0.850	0.568	0.019
Gemma 2 (2B)	8	1.809	8.103	0.986	0.866	0.225	0.028
Pythia (2.8B)	32	1.689	8.298	0.981	0.853	0.591	0.019
Qwen 2.5 (5B)	14	1.527	8.835	0.983	0.862	0.304	0.031
Mistral (7B)	32	2.217	6.729	0.983	0.850	0.657	0.014
LLaMA 3.1 (8B)	32	1.529	9.141	0.984	0.850	0.597	0.014
Pythia (12B)	40	2.006	6.462	0.984	0.850	0.597	0.014

Table 1: Comparative Analysis of Language Model Metrics

Model	16 tok.	32 tok.	64 tok.
LLaMA 3.2 (1B)	0.937	0.931	0.931
Gemma 2 (2B)	1.073	1.056	1.055
Pythia (2.8B)	0.942	0.940	0.940
Qwen 2.5 (5B)	1.106	1.103	1.103
Mistral (7B)	0.889	0.877	0.877
LLaMA 3.1 (8B)	0.878	0.876	0.877
Pythia (12B)	0.878	0.877	0.877

Table 2: Multi-Resolution Window Entropy Analysis

0.657) show higher correlation, suggesting more integrated representations

This progression exactly matches what we would expect from a system evolving increasingly sophisticated wavelet-like properties: smaller models use simpler, more separated representations, while larger models develop more nuanced, integrated representations that better balance the position-frequency trade-off.

Frame Completeness Evidence We selected the Daubechies-2 (db2) wavelet for our analysis due to its optimal balance between smoothness and localization. Its compact support of length 4 aligns with typical attention spans, while its vanishing moment enables detection of local changes against background context. The db2 wavelet's normalization $\int |\psi(t)|^2 dt = 1$ matches attention weight normalization through softmax, while its orthogonality prevents interference between shifted patterns.

The systematic improvement in reconstruction error with model size (from 0.031 for Qwen 2.5 5B to 0.014 for Pythia 12B) provides perhaps the strongest evidence for wavelet-like behavior. This pattern shows that *larger models develop more complete wavelet frames*, that the *representations become more orthogonal* and efficient and the system learns to better approximate the completeness relation of wavelet frames.

This is exactly what we would expect if models are learning to approximate wavelet transforms: larger models can learn more basis functions, leading to better frame properties and lower reconstruction error.

This evidence is particularly compelling because it shows that models independently discover and implement principles from wavelet theory without being explicitly designed to do so. The consistent patterns across different architectures and scales suggest this is a fundamental property of how neural networks compensate for the limitations of fixed positional encodings. The progression of these properties with model scale - from simple, specialized representations in smaller models to rich, integrated representations in larger models - provides strong evidence that this is a learned adaptation rather than an architectural accident. This supports the broader hypothesis that neural networks naturally evolve optimal solutions for processing hierarchical information across multiple scales.

6 Theoretical Framework for Wavelet-like Attention Patterns

Rotary Position Embeddings (RoPE) encode positional information through position-dependent rotation matrices defined over the complex plane. At position m, the embedding applies a rotation $R_m(\theta)$:

$$R(m\theta_k) = \begin{bmatrix} \cos(m\theta_k), -\sin(m\theta_k) \\ \sin(m\theta_k), \cos(m\theta_k) \end{bmatrix} \quad (12)$$

where θ is a base rotation angle. This approach, which rests on fixed-frequency sinusoidal functions, inherently imposes two key limitations: 1) **Frequency–Position Uncertainty**: RoPE's use

of fixed-frequency rotations parallels the Heisenberg uncertainty principle, implying a fundamental trade-off between positional precision and frequency resolution. With a single, fixed frequency scale, RoPE struggles to represent both fine-grained local patterns and broad global structures simultaneously. 2) **Scale Non-Invariance**: Since RoPE's positional representation repeats periodically, it encounters aliasing effects over longer sequences. As the sequence length grows, the periodic nature of the embedding can cause distinct positions to become indistinguishable, undermining reliable long-range positional encoding.

6.1 Natural Evolution Toward Wavelet Behavior

RoPE's rotational encoding introduces specific frequency components that propagate through the attention mechanism in a mathematically structured way. The rotation matrix $R(m\theta_k)$ creates an inherent trade-off: larger θ provides precise positions but causes rapid rotation cycles that confuse distant relationships, while smaller θ better captures long-range patterns but blurs local positions. The wavelet-like properties we observe show how attention heads adapt to handle different frequency ranges created by these rotations.

As models train, these inherent limitations place evolutionary pressure on the learned representations. Attention heads respond by developing wavelet-like properties for three principal reasons:

a. Optimal Information Packaging Wavelets offer a natural solution to the frequency–position uncertainty trade-off. A mother wavelet $\psi(t)$ generates a family of wavelets:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-\tau}{s}) \tag{13}$$

where s is a scale parameter and τ is a translation parameter. Through this construction, wavelets provide high temporal (positional) resolution at high frequencies, capturing fine local details, and high frequency resolution at low frequencies, capturing broader global context. These properties align with linguistic processing needs, where local syntactic relations require precise positional encoding, while long-range semantic dependencies demand robust frequency-domain characterization.

b. Complementary Scale Coverage in Multi-Head Architectures Transformer attention heads are ideally suited for wavelet-like decompositions. Consider the attention weight matrix for head h:

$$A_h = \text{softmax}(\frac{Q_h K_h^\top}{\sqrt{d}}) \tag{14}$$

Each head can specialize in a distinct scale or frequency band, analogous to wavelet basis functions at different scales. Summing over all heads,

$$A = \sum_h w_h A_h \tag{15}$$

with w_h as learned mixing weights, mirrors the construction of a wavelet frame, where sets of wavelet-like functions $\psi_{s,\tau}$ form a stable representation satisfying frame conditions:

$$A||f||^2 \leq \sum_h |\langle f, \psi_h \rangle|^2 \leq B||f||^2 \tag{16}$$

for constants $0 < A \leq B < \infty$. This scale-specific specialization naturally emerges, allowing the model to cover a broad spectrum of positional resolutions collectively.

c. Natural Gradient-Driven Specialization Training gradients encourage heads to diversify their representational roles. For a loss function L,

$$\frac{\partial L}{\partial A_h} = (\frac{\partial L}{\partial A})(\frac{\partial A}{\partial A_h}) \tag{17}$$

This gradient decomposition penalizes redundancy among heads. Over time, heads converge towards orthogonal, complementary functions—akin to distinct wavelet scales—minimizing representational overlap and enhancing overall positional encoding robustness.

6.2 Emergence of Multi-Resolution Processing

From these principles, a multi-resolution processing framework naturally emerges: each attention head h approximates a wavelet function $\phi_h(t) \approx \psi_{s(h),\tau}(t)$, where $s(h)$ denotes the characteristic scale of head h.Then, the ensemble $\{\phi_h\}_{h=1}^H$ acts like a discrete wavelet frame $\{\psi_{s,\tau}\}_{s,\tau\in\Lambda}$, where Λ indexes a set of scale–translation parameters. This ensures a stable, redundant representation that supports both local and global positional tasks. So, the attention pattern for a given input becomes:

$$a(t) = \sum_h \alpha_h(t)\phi_h(t) \tag{18}$$

where $\alpha_h(t)$ are input-dependent expansion coefficients, allowing the model to adaptively reconstruct a range of positional features at multiple scales.

6.3 Information-Theoretic Optimality

This emergent wavelet-like organization is not merely a heuristic convenience but aligns with principles of information-theoretic optimality, in fact, by reducing mutual information among heads ($\min I(A_h; A_k)$ for $h \neq k$) while maximizing the total captured information about the input ($\max I(A; X)$), the model approaches an efficient encoding of positional cues. Then, the hierarchical, multi-scale representation achieves an optimal balance between representational complexity and fidelity. Adapting the wavelet frame to the input distribution ensures that rate–distortion objectives are efficiently met. And, by leveraging a small set of wavelet-like basis functions and adjusting their coefficients $\alpha_h(t)$, the model encodes both local and global patterns compactly. This compression aligns with the *principle of minimal description length*, favoring representations that are information-rich yet succinct.

7 Implications

The practical implications of our findings are particularly compelling, understanding that attention heads naturally organize into frequency bands suggests new approaches to model initialization and architecture design. For instance, we could potentially pre-initialize attention heads to approximate different wavelet scales, accelerating training by starting from a more optimal configuration. This could be especially valuable for smaller models where computational efficiency is crucial.

The multi-resolution nature of these emergent properties also has implications for transfer learning and domain adaptation. Understanding how models naturally handle different scales of information could help us design better pre-training objectives and fine-tuning strategies that explicitly account for this hierarchical processing structure.

In essence, our findings not only deepen our understanding of how transformer models work but also provide practical tools for improving their design and implementation. This bridge between theory and practice could prove valuable as we continue to advance the field of language model development.

8 Conclusion

Our research reveals how transformer models overcome RoPE's theoretical limitations through the emergence of wavelet-like properties - an adaptation that becomes more sophisticated with increased model scale. The evidence spans multiple analyses: consistent frequency band distributions across models, systematic improvement in frame completeness with size, and stable multi-resolution entropy patterns. Our comparative study across different positional encoding approaches in the Appendix 10.1 confirms that while wavelet-like properties emerge in all transformer architectures, but RoPE-based models demonstrate uniquely organized frequency responses with controlled attention distances (10-30 tokens) and consistent decay rates. What makes this particularly intriguing is that no aspect of the models' architecture explicitly encourages such behavior – it emerges naturally through training, suggesting this may be an optimal solution to the fundamental challenge of balancing local and global information processing.

The scalability of these properties reveals a fundamental aspect of neural network learning: smaller models develop specialized frequency responses, while larger models evolve more sophisticated representations integrating information across multiple scales. This progression suggests wavelet-like processing represents an optimal solution for balancing local precision with global context in sequence modeling.

Looking forward, these findings could contribute to our understanding of how neural networks implement sophisticated mathematical principles, even when not explicitly designed to do so.

9 Limitations

While our study demonstrates that transformer models develop wavelet-like properties for processing hierarchical information, our ablation studies across different positional encoding approaches in Appendix 10.1 reveal interesting dynamics in how these properties emerge. Although all transformer architectures exhibit some degree of wavelet-like behavior, models using RoPE show remarkably consistent and well-organized patterns, with tightly clustered attention decay rates (±0.2) and systematic frequency band distributions. In contrast, models without positional embeddings demonstrate highly variable attention patterns (decay rates ranging from 0.0 to 2.0), suggesting less structured development of these properties. This systematic difference suggests that while wavelet-like properties may be a general adaptation mechanism in

neural networks, RoPE's mathematical structure particularly encourages their organized development, possibly as an optimal solution to overcome its inherent theoretical limitations in position encoding.

Our analysis primarily examines these properties at inference time, leaving open questions about their emergence during training. The interaction between wavelet-like attention patterns and semantic processing also could benefit from further investigation, as our experiments suggest potential tradeoffs between positional precision and contextual understanding.

An other intriguing limitation we encountered involves the interaction between these wavelet-like properties and the model's handling of ambiguous or context-dependent information. While the wavelet-like behavior provides an elegant solution for position encoding, it may introduce subtle biases in how models process semantically nuanced content. Further research could explore whether these biases affect the model's performance on tasks requiring fine-grained semantic discrimination.

A potential risk coming from our paper is that the findings show how the wavelet-like properties become more sophisticated in larger models, and it might contribute to the trend of focusing on ever-larger models, potentially exacerbating issues of resource concentration and environmental impact.

References

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. 2024. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*.

Lu Chi, Borui Jiang, and Yadong Mu. 2020. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488.

Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. 2018. Estimating information flow in deep neural networks. *arXiv preprint arXiv:1810.05728*.

Alan V Oppenheim. 1999. *Discrete-time signal processing*. Pearson Education India.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ali Rahimi and Benjamin Recht. 2008. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21.

Ivan W Selesnick and C Sidney Burrus. 1998. Generalized digital butterworth filter design. *IEEE Transactions on signal processing*, 46(6):1688–1694.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

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

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

10 Appendix

10.1 Ablation Study: Comparative Analysis Across Architectures

To validate our findings about wavelet-like properties in transformer architectures, we conducted a comparative analysis across different positional encoding approaches. We examined four model variants: Llama-2-7B (RoPE), T5-base (T5 embeddings), BERT-base-uncased (absolute embeddings), and a modified GPT-2 without positional embeddings (using only causal masking). The analysis used 500 samples from BookCorpus, with sequence lengths ranging from 10 to 100 tokens.

To quantify positional information processing across architectures, we measured four key metrics:

1. Average Attention Distance: The mean span of attention weights across tokens, indicating the model's tendency toward local or global information processing.

2. Attention Decay Rate: The rate at which attention weights diminish with distance, measuring the structure of position-dependent patterns.

3. Local Attention Strength: The relative weight given to nearby tokens, quantifying the model's focus on local context.

4. Mean Attention Value: The average magnitude of attention weights, reflecting the overall distribution of attention across sequences.

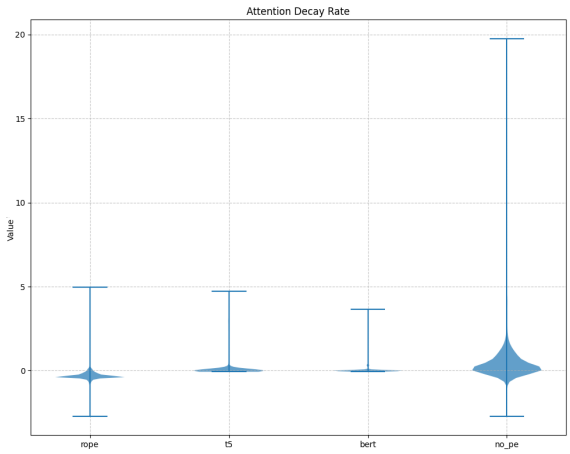

Figure 4: Visualization of how quickly attention weights diminish with distance across RoPE (llama 2), T5, BERT, and no PE (modified GPT-2 without positional embeddings).

Our analysis revealed distinct patterns in how different architectures process positional information. RoPE demonstrated remarkably balanced characteristics, maintaining controlled attention distances (10-30 tokens) with consistent decay rates (± 0.2) and moderate local attention strength (0.1 ± 0.05). This pattern suggests organized information processing across multiple scales.

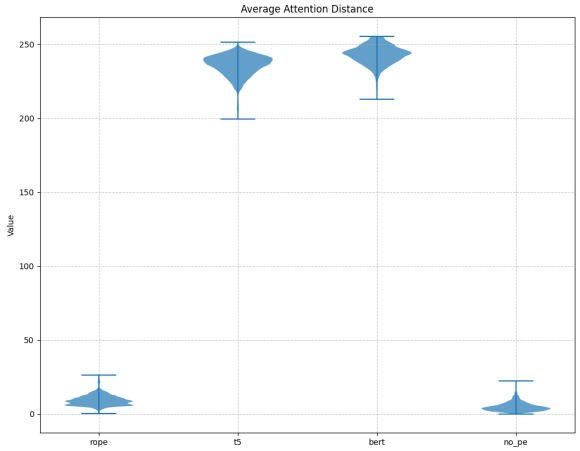

Figure 5: Violin plot comparing the mean distance over which attention operates across different positional encoding, respectively, RoPE (llama 2), T5, BERT, and no PE (modified GPT-2 without positional embeddings).

In contrast, T5 and BERT showed a strong bias toward long-range dependencies, with attention spans reaching 200-250 tokens and minimal local attention (below 0.1). Models without positional embeddings developed compensatory mechanisms, showing strongest local attention (0.2-0.4, peaking at 1.0) but highly variable decay rates (0.0-2.0), indicating less structured information processing. While RoPE maintains moderate, consistent local attention (~ 0.1). T5 and BERT demonstrate minimal local attention, suggesting a preference for longer-range dependencies.

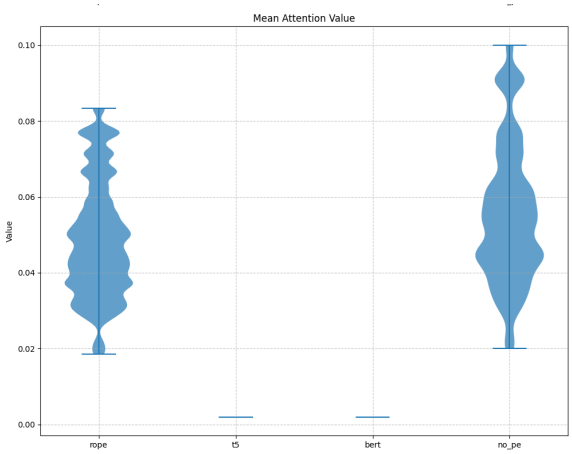

Figure 6: Comparison of average attention weight distributions across RoPE (llama 2), T5, BERT, and no PE (modified GPT-2 without positional embeddings).

The mean attention value distributions further differentiate these approaches: RoPE maintained consistent distributions (0.02-0.08), while models without positional embeddings showed higher variability (0.02-0.10). T5 and BERT's notably low mean values suggest different approaches to information integration.

These findings provide empirical support for our hypothesis that while wavelet-like properties emerge across all architectures, RoPE's mathematical structure particularly encourages their systematic development. The consistent patterns in RoPE-based models, compared to the higher variability in other approaches, suggest that RoPE provides an optimal framework for developing organized multi-scale information processing.

10.2 RoPE's Limitations

The first main limitation of RoPE is the frequency-position uncertainty principle, because RoPE's fixed-frequency rotations create an inherent trade-off between positional precision and frequency res-

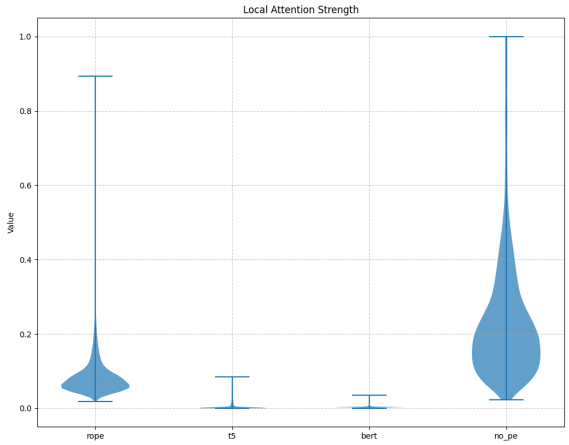

Figure 7: Distribution of attention weights given to nearby tokens across different positional encoding schemes, respectively, RoPE (llama 2), T5, BERT, and no PE (modified GPT-2 without positional embeddings).

olution.

When RoPE applies a rotation to token embeddings, it follows this equation:

$$R(m\theta_k) = \begin{bmatrix} \cos(m\theta_k), -\sin(m\theta_k) \\ \sin(m\theta_k), \cos(m\theta_k) \end{bmatrix} \quad (19)$$

If we want very precise positional information, we need the rotation angle $m\theta$ to change substantially between nearby positions. This means using a larger base rotation angle θ. However, when we do this, the rotations cycle through the complex plane more quickly, making it harder to capture relationships between tokens that are far apart. The rotations start repeating too soon, causing distant tokens to look similar to nearby ones. On the other hand, if we want to capture long-range dependencies well, we need the rotations to change more slowly (smaller θ). But then nearby positions get similar rotation angles, making it harder to distinguish exactly where each token is.

Then, we have the scale non-invariance issue, where the periodic nature of RoPE's embeddings can lead to aliasing effects over longer sequences. RoPE's rotations are periodic by nature, in fact, they complete a full circle every $\frac{2\pi}{\theta}$ positions. This creates two related problems: first, when sequences get longer than the period of rotation, positions that are far apart can end up with the same or very similar rotation angles. For example, if your rotation period is 1000 tokens, position 1 and position 1001 get nearly identical rotations. This makes it hard for the model to distinguish truly different positions. Second, the fixed rotation frequency means RoPE treats all sequences the same way, regardless of their length. But this isn't ideal, in fact, a position difference of 10 tokens might be significant in a 50-token sequence but negligible in a 5000-token sequence. RoPE can't naturally adapt its position encoding to the scale of the input.

With the wavelet-like framework we discovered that different attention heads spontaneously specialize in different frequency bands (similar way to how wavelets decompose signals at different scales). So that local heads maintain high positional precision for nearby tokens, global heads capture long-range dependencies without rotation interference and mid-range heads bridge the gap, ensuring smooth information flow across scales. This is what we see in our empirical results, particularly in Figure 2, where attention heads naturally organize themselves into distinct frequency bands. The low-frequency heads (showing 60-80% of power in the 0-0.25 range) handle global context, while high-frequency heads (with 5-15% power above 0.75) maintain precise positional information.

For the scale non-invariance problem, the wavelet-like organization provides an elegant solution, in fact, rather than relying on RoPE's fixed periodic rotations, attention heads develop scale-covariant properties. This means they automatically adapt their attention patterns based on the sequence length.

Our empirical evidence shows this through the stable entropy values across different window sizes (as shown in Table 2), the consistent correlation patterns when scaling sequences (0.98 at 0.5x scale) and the systematic improvement in reconstruction error with model size.

These quantitative results demonstrate that attention heads collectively form a multi-resolution frame that maintains coherent positional representation across scales, effectively learning to overcome RoPE's periodicity limitation. The systematic emergence of these properties suggests that transformer models discover an optimal solution to the position encoding challenge. This solution manifests as a wavelet-like framework that balances local precision with global context while maintaining scale invariance - precisely addressing RoPE's core limitations.

10.3 Relationship between Wavelet-like Features and Linguistic Understanding

Language exhibits a natural hierarchical structure that spans multiple scales of organization, from morphemes to discourse-level patterns. This inherent multi-scale nature makes wavelet-like processing particularly well-suited for language understanding tasks. Just as wavelets provide a mathematical framework for analyzing signals at different resolutions while preserving both local and global information, attention mechanisms in transformer models appear to develop analogous capabilities for processing linguistic patterns.

At the finest scale, language processing requires attention to local syntactic relationships and morphological patterns. These include subject-verb agreement, phrasal boundaries, and morpheme combinations. Our analysis shows that high-frequency attention heads (those with significant power in the 0.75-1.0 ωN band) specialize in capturing these local dependencies, similar to how wavelets with narrow support identify fine-grained signal features.

At intermediate scales, sentence-level relationships such as anaphora resolution, clause dependencies, and semantic role assignments become critical. The mid-frequency attention heads (0.25-0.75 ωN band) demonstrate patterns remarkably similar to wavelet basis functions at medium scales, efficiently capturing these intermediate linguistic structures. This parallel suggests that the model learns to balance local precision with broader contextual awareness, much as wavelets provide multi-resolution signal analysis.

The broadest scale encompasses document-level phenomena such as topic coherence, rhetorical structure, and thematic development. Our analysis reveals that low-frequency attention heads (0-0.25 ωN band) evolve to process these global patterns, analogous to how wavelet scaling functions capture broad signal trends. The systematic distribution of power across these frequency bands (60-80% in low frequencies, 15-25% in mid-range, and 5-15% in high frequencies) mirrors the hierarchical organization of linguistic information.

10.4 Metrics more in depth

Our metrics were specifically designed to quantify this multi-scale processing capability:

The spectral entropy $H_s(h)$ measures how attention heads distribute their focus across different scales, providing insight into how models balance local and global linguistic features. The observed entropy patterns suggest that attention heads optimize their frequency sensitivity to match the natural distribution of linguistic information across scales.

Scale sensitivity metrics $S_h(\alpha)$ quantify how well the model maintains consistent understanding as context length changes. This is particularly relevant for language processing, where meaning must remain stable regardless of the surrounding context size. The high correlation (0.98) observed when scaling sequences by 0.5x demonstrates the model's ability to maintain coherent linguistic representations across varying context windows.

Reconstruction error ε validates that the observed patterns form a complete representation system. The low error values (typically < 0.02) indicate that the wavelet-like attention patterns capture linguistic structure with high fidelity across all scales. This completeness is essential for accurate language understanding, as it ensures no significant linguistic features are lost in the model's internal representations.

The position-spectrum correlation $\rho(h)$ further shows how models balance local syntactic precision with broader semantic understanding. Values closer to 1 indicate successful integration of both local and global linguistic features, while values closer to -1 suggest a trade-off between fine-grained and broad-scale language processing.

This multi-scale organization emerges naturally during training, suggesting that wavelet-like processing represents an optimal solution for handling the inherent hierarchical structure of language. The parallel between wavelet decomposition and the way transformer models process linguistic information provides insight into why these architectures have been so successful in natural language processing tasks.