LEROPE: LEARNABLE ROPE FREQUENCIES IMPROVE LANGUAGE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary Positional Encodings (RoPE) are currently the most popular positional encodings used in modern language models. RoPE rotates two-dimensional chunks of query and key vectors, operating as a function of their relative positional offset. The position-wise rates of rotation in RoPE typically follow a geometric sequence specified by a fixed base-frequency hyperparameter. Prior work has improved performance by either increasing this parameter to slow rotation or by applying RoPE to only a subset of QK dimensions. In this work we modify RoPE by learning a scalar per frequency, treating frequencies as learnable parameters rather than hyperparameters. We validate **Le**arned **RoPE** by training a ladder of language models from scratch, ranging from 52M to 2.5B parameters. We observe and analyze the emergence of a high-norm, positional LeRoPE band. LeRoPE consistently outperforms RoPE and partial RoPE across all scales, with RoPE requiring 3.4% more compute (FLOPs) to match LeRoPE at the largest scale.

1 INTRODUCTION

Positional encodings are typically included in Transformer-based models (Vaswani et al., 2017) because the self-attention operation on its own does not directly carry positional signal. Rotary Positional Encodings (RoPE) (Su et al., 2021) are used by a number of popular language models (Touvron et al., 2023; Chowdhery et al., 2022; Jiang et al., 2023; Kamath et al., 2025). RoPE splits the query and key vectors into two-dimensional pairs, or *frequency bands*, and rotates each pair by an angle proportional to token position. The effect of these rotations on the query–key inner product depends only on the relative offset between tokens. Each band rotates at its own fixed rate, following a geometric sequence set by a base hyperparameter: high-frequency bands complete many rotations within a context, while the lowest-frequency bands rotate minimally.

The choice of these frequencies matters. Recent work finds that the slowest bands carry semantic rather than positional information (Barbero et al., 2024). This positional invariance allows content matches to hold at any separation, making it beneficial to remove rotation from a subset of bands entirely, or to adjust the base, which sets how many bands rotate appreciably within the context window (Barbero et al., 2024; Men et al., 2024; Oka et al., 2026b).

Motivated by the above, and noting that the RoPE operation is differentiable with respect to its frequency, we propose learning the frequencies. Learned RoPE (LeRoPE) adds one learned scalar per frequency band, 32 parameters in total for each of our models. Training on a ladder of models from 52M to 2.5B parameters, LeRoPE outperforms RoPE at every scale. At the largest scale, RoPE requires 3.4% more compute to match the performance of LeRoPE (Figure 1).

Analyzing models trained with LeRoPE, we observe a consistent frequency profile across scales and random seeds, as well as the emergence of a single band with a large, strongly positional contribution to the attention logits, which we term the "dominant positional band."

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

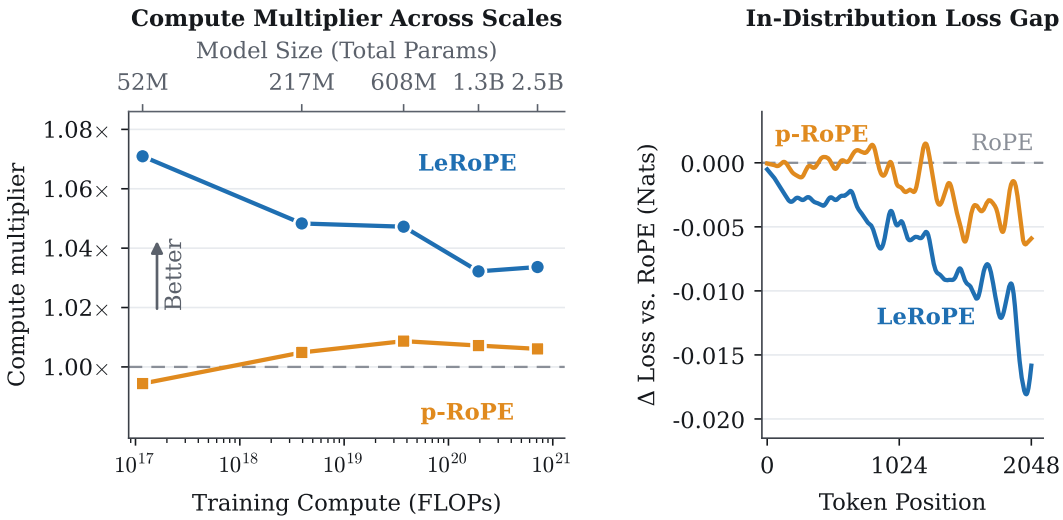

Figure 1: **a**: Compute multipliers vs. RoPE. The compute multiplier is the factor by which the baseline's training compute must increase to match the variant's loss. The multiplier declines from $\sim$1.07 at the smallest scale and plateaus near 1.03–1.04, a small but persistent advantage. This means RoPE needs 3.4% more compute to reach the loss LeRoPE achieves at 2.5B. All runs use a single seed; learning rates are swept per method for models below 1.34B (§4.1). **b**: Average loss gap per token position of p-RoPE and LeRoPE vs. RoPE, in-distribution (smoothed via Gaussian kernel, $\sigma = 30$).

2 Preliminaries

2.1 Positional Embeddings

To break the positional symmetry of the attention mechanism, Vaswani et al. (2017) add fixed sinusoidal positional embeddings directly to the input residual stream. Later models instead incorporated learned absolute positional embeddings (Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019; Zhang et al., 2022; Dosovitskiy et al., 2020).

Shaw et al. (2018) proposed *relative* position representations, which depend only on the relative offset between tokens; a simplified, scalar version of learned relative offsets was adopted by T5 (Raffel et al., 2019). Rotary positional embeddings, proposed by Su et al. (2021), also depend only on relative position, and have become the standard choice for modern language models.

Since RoPE, a number of positional encodings have been proposed. ALiBi (Press et al., 2021) improved out-of-distribution length extrapolation with fixed relative positional biases. CoPE provides the model with the capacity to produce relative positional embeddings from learned, content-dependent features (Golovneva et al., 2024); PaTH also proposes a data-dependent relative transform based on accumulating Householder matrices (Yang et al., 2025b).

Some methods generalize to a broader class of functions, of which RoPE is a member. For instance, LieRE learns skew-symmetric matrices that can rotate along planes other than the standard RoPE plane (Ostmeier et al., 2024). Oka et al. (2026a) generalize RoPE to a broader class of wavelet-function-based positional encodings, finding improved performance from Ricker-based wavelets. We include an expanded related works section in Appendix B.

2.2 RoPE

Su et al. (2021) propose **Ro**tary **P**osition **E**mbedding (**RoPE**), a relative positional embedding method that forms the basis of our study. RoPE rotates key and query vectors according to their position, modifying their inner product according only to their relative offset.

Concretely, consider a single attention head with per-head dimension d_h. The dimensions of this head can be chunked into $d_h/2$ frequency pairs, indexed by $m = 0, \ldots, d_h/2 - 1$. Standard RoPE assigns each pair a frequency from a geometric series:

$$\theta_m = b^{-2m/d_h}, \quad m = 0, \ldots, d_h/2 - 1, \tag{1}$$

where $b > 1$ is the base (typically $b = 10{,}000$). Following the notation of Barbero et al. (2024) we denote $\rho(\theta_m)$ as the matrix form of the rotational transformation of angle θ_m:

$$\rho(\theta_m) = \begin{bmatrix} \cos(\theta_m) & -\sin(\theta_m) \\ \sin(\theta_m) & \cos(\theta_m) \end{bmatrix} \tag{2}$$

RoPE constructs a block-diagonal matrix $\mathbf{R}_s = \oplus_{m=0\ldots d_h/2-1} \rho(\theta^{(m)})^s$ where s is the position of a given token. It then applies $\mathbf{R}_s$ to each key and query vector; that is $\tilde{q}_s = \mathbf{R}_s q_s$ and $\tilde{k}_s = \mathbf{R}_s k_s$.

The rotation-matrix properties $\rho(\theta)\rho(\theta') = \rho(\theta + \theta')$ as well as $\rho(\theta)^\intercal = \rho(-\theta)$, in conjunction with the block-diagonal structure of $\mathbf{R}_s$ gives the relation $\mathbf{R}_s^\intercal \mathbf{R}_t = \mathbf{R}_{t-s}$. The dot product of a RoPE-rotated query-key pair $\tilde{q}_s, \tilde{k}_t$ decomposes into:

$$\tilde{q}_s^\intercal \tilde{k}_t = (\mathbf{R}_s q_s)^\intercal (\mathbf{R}_t k_t) = q_s^\intercal \mathbf{R}_{t-s} k_t = \sum_{m=0\ldots d_h/2-1} \left(q_s^{(m)}\right)^T \rho(\theta_m)^{t-s} k_t^{(m)} \tag{3}$$

where $q_s^{(m)}$ is the 2 dimensional chunk (or band) of the query vector $q_s^{[2m,2m+1]}$ (the same notation holds for the key vectors). Thus, the RoPE operation depends only on the relative offset $t - s$ and not on the absolute positions s and t.

2.3 Frequency Bands and Partial RoPE

Each index m in Eq. equation 1 defines a *frequency band*: a two-dimensional sub-vector rotated at angular frequency θ_m, completing a full period every $2\pi/\theta_m$ tokens. Because the frequencies are geometrically spaced, these periods span several orders of magnitude, and the slowest bands typically do not complete a single period within the training context.

Trained models do not use this spectrum uniformly. Barbero et al. (2024) show that RoPE models concentrate query and key norm in the lowest-frequency bands and argue that these bands, which rotate negligibly over typical offsets, transport semantic rather than positional information. Since even a slow band eventually drifts over long ranges, they propose Partial RoPE (p-RoPE), which removes rotation from the slowest bands entirely:

$$\theta_m = \begin{cases} b^{-2m/d_h} & m < p \cdot d_h/2, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where p is the fraction of bands kept. Setting p=1 recovers RoPE, while p=0 recovers NoPE — omitting positional encoding altogether — which itself improves length generalization (Kazemnejad et al., 2023). Adjusting the base b acts on the same lever continuously: raising it slows every band rather than zeroing the slowest, and is an effective tool for context extension (Liu et al., 2023; Men et al., 2024), adopted at scale in recent open models (Kamath et al., 2025).

Oka et al. (2026b) expand this analysis and identify a contiguous *high-norm band* of RoPE dimensions that emerges early in pre-training, with its location determined jointly by the base b and the training length. Lower frequency bands can be replaced with NoPE at inference at little cost, confirming that the slowest frequencies are only weakly used for position. All of these methods, however, fix the band allocation by hand before training.

3 LEROPE

3.1 EXTENDING ROPE TO LEARNED FREQUENCIES

LeRoPE augments RoPE with one learned scalar α_m per frequency band, rescaling each frequency:

$$\hat{\theta}^{(m)} = e^{\alpha_m}\theta^{(m)}, \qquad \mathbf{R}_s^{\mathrm{LeRoPE}} = \bigoplus_{m=0}^{d_h/2-1} \rho\big(\hat{\theta}^{(m)}\big)^s. \tag{5}$$

We parametrize frequency scalars in logarithmic space, so that steps in α_m make multiplicative changes to $\hat{\theta}^{(m)}$. This yields better conditioning when scaling down frequencies and avoids undesirable gradient dynamics near frequencies of zero. We initialize $\alpha_m = 0$, making our method equivalent to RoPE on initialization. Additionally, we omit weight decay on LeRoPE parameters, and apply gradient clipping to them independently of the rest of the model's parameters.

All layers and heads share the same set of learned frequencies; thus, with d_h as the head dimension, LeRoPE adds $d_h/2$ total parameters to the model. The parametrization also admits finer granularity (e.g., per-layer or per-head), but we find that sharing α_m values across all layers and heads performs consistently best.

3.2 GRADIENT OF A FREQUENCY

Each learned frequency receives its gradient through the per-band contributions of Eq. equation 3. Writing the band-m chunks in polar form and differentiating,

$$\frac{\partial \mathcal{L}}{\partial \hat{\theta}^{(m)}} = -\sum_s \sum_{t<s} (s-t)\, \big\|q_s^{(m)}\big\|\, \big\|k_t^{(m)}\big\|\, \sin\big(\phi_{st}^{(m)} + (s-t)\,\hat{\theta}^{(m)}\big)\, \frac{\partial \mathcal{L}}{\partial z_{st}} \tag{6}$$

where $\phi_{st}^{(m)}$ is the angle from the key to the query chunk and z_{st} the attention logit.

We can express the gradient update as the derivative of a surrogate loss $P(\theta)$. Let α_{st} be the post-softmax attention weight from query q_s to key k_t; let v_t be the value vector at position t and $o_s = \sum_{t<s} \alpha_{st} v_t$ be the attention-aggregated output vector, before projection for simplicity. Further, let $g_s \triangleq \partial \mathcal{L}/\partial o_s$ be the downstream gradient and L the maximum context length.

With these quantities, we can define a spectral function of offset $C_m(\delta)$. The direction of each term is dictated by the relative QK angle $\phi_{st}^{(m)}$, and its scale is the product of QK norms, post-softmax probability, and the downstream gradient term.

$$C_m(\delta) \triangleq \sum_{\substack{s,t \\ s-t=\delta}} \alpha_{st} \|q_s^{(m)}\| \|k_t^{(m)}\| e^{i\phi_{st}^{(m)}} (v_t - o_s)^T g_s = R_m(\delta) e^{i\Phi_m(\delta)}$$

$$P_m(\theta) = \mathrm{Re}\left(\sum_{\delta=1}^{L-1} C_m(\delta) e^{i\delta\theta}\right) = \sum_{\delta=1}^{L-1} R_m(\delta) \cos\left(\Phi_m(\delta) + \delta\theta\right)$$

Lemma 3.1. *Define $P_m(\theta)$ as above: the real component of the discrete-time Fourier Transform of the offset-aggregated gradient signal $C_m(\delta)$ with frequency θ.*

Then $\frac{\partial \mathcal{L}}{\partial \theta^{(m)}} = P_m'\left(\theta^{(m)}\right)$.

Proof. See Lemma C.2. □

This provides one perspective on how $\theta^{(m)}$ is updated. Each offset δ is mapped to a complex phasor $C_m(\delta)$, which aggregates the downstream gradients weighted by QK norms, un-rotated QK relative angles, and the post-softmax attention weights α_{st}.

Table 1: Validation perplexity on C4 across our scaling ladder (lower is better; best per column in bold). LeRoPE improves over the RoPE baseline at every scale; p-RoPE yields a smaller, consistent improvement at all but the smallest scale.

Method	52M	217M	608M	1.34B	2.52B
RoPE	30.2450	18.2377	13.8795	11.5747	10.1720
p-RoPE	30.2695	18.2269	13.8663	11.5665	10.1659
LeRoPE	**29.9479**	**18.1335**	**13.8093**	**11.5384**	**10.1387**

Then, $\theta^{(m)}$ is updated towards the local minimum of the surrogate function $P_m(\theta)$, which is the real component of the discrete-time Fourier Transform of the series, evaluated at frequency θ. The effect of each relative offset on this gradient can be characterized by visualizing the per-offset phasors $C_m(\delta)$: we plot these for each band in Appendix C.2 to illustrate the per-offset geometries; the phasors largely cancel about the origin rather than cohering.

4 Experimental Setup

4.1 Scaling Ladder

To evaluate our proposed method we train a family of decoder-only transformers ranging in size from 52M to 2.5B parameters under a compute-optimal regime. Following previous works (Ferbach et al., 2026; Charles et al., 2025) we co-scale layers and head count keeping the head dimension fixed at 64 so all models have the same number of learned frequency scales. We adopt the Chinchilla token count of $D = 20N$ where N is the total number of model parameters. We use C4 (Raffel et al., 2019) as our pre-training corpus and a sequence length of 2048 with intra-document masking. During evaluation we do not pack documents. An exhaustive overview of our architecture, hyper-parameters and differences from existing works can be found in Appendix A.1.

We compare against RoPE with the standard base $b = 10{,}000$ and p-RoPE with $p = 0.75$ following Barbero et al. (2024). For every model below 1.34B parameters we sweep the learning rate over a geometric grid, $\eta \in \{2^{-i/2} : i \in \mathbb{Z}\}$, until the best value is bracketed by a tested value on either side. We run this sweep independently for each of the three positional-encoding methods, and find that all three share the same optimal learning rate at every scale; comparisons at a given scale therefore use identical hyperparameters. For the two largest models we extrapolate the best grid index from the smaller scales.

4.2 Compute Multipliers

To evaluate the benefit of a proposed architecture change, an important quantity to measure is the amount of compute saved relative to the baseline in order to reach the same performance. We report this through compute multipliers: a compute multiplier of $1.x$ for LeRoPE corresponds to needing $x\%$ more compute to reach the same loss with RoPE.

We measure the training compute of each model as $C = 6ND$ FLOPs, where N is its number of non-embedding parameters. For each method we form a loss–compute curve by piecewise-linear interpolation of the final validation losses of its compute-optimal models in $(\log C, L)$ space and extrapolating beyond the smallest and largest scales with the slope of the nearest segment. Writing L^{RoPE} and L^{var} for the baseline and variant curves, the compute multiplier of a variant trained at compute C^{var} is

$$\mathrm{CM} \;=\; \frac{C^{\mathrm{RoPE}}}{C^{\mathrm{var}}}, \qquad \text{where} \qquad L^{\mathrm{RoPE}}\big(C^{\mathrm{RoPE}}\big) = L^{\mathrm{var}}\big(C^{\mathrm{var}}\big), \tag{7}$$

i.e. the factor by which RoPE's compute must increase to reach the variant's loss, obtained by inverting the interpolated L^{RoPE}.

Our scaling ladder is sparser than that of Ferbach et al. (2026), with adjacent scales differing by a decade or more of compute. Linear interpolation across these gaps overestimates the multiplier for the convex power-law loss curve, so we invert Equation 7 in $\log C$ rather than C to reduce the

Table 2: Validation perplexity and percentage of LeRoPE gain captured by each method on a 217M-parameter model. Fixed LeRoPE uses frozen frequencies learned by LeRoPE on a separate run at the same scale.

Method	Perplexity	% Gain Captured
RoPE	18.2377	0%
p-RoPE	18.2269	10.4%
Fixed LeRoPE	18.1714	63.6%
LeRoPE	**18.1335**	100%

overestimation. We include further discussion on this choice as well as a comparison with the more traditional *efficiency gain* fitted power-law inversion method in Appendix A.2.

5 Results

5.1 Compute Multipliers

We report the compute multipliers for LeRoPE and p-RoPE in Figure 1(a), and the corresponding validation perplexity numbers in Table 1. Across all scales LeRoPE realizes a small but consistent gain, corresponding to RoPE requiring 3.4% more compute to match it at the largest scale. We also include the loss delta relative to RoPE as a function of position in Figure 1 (b). LeRoPE's gain over RoPE grows with context length, suggesting it makes better use of longer context.

To bound run-to-run variability, we repeat the 217M RoPE and LeRoPE runs with three random seeds. The method gap of 0.0066 nats is 3 to 7 times the per-method seed standard deviation; LeRoPE improves on RoPE under every seed, with the worst LeRoPE run outperforming the best RoPE run by 0.0034 nats. The exact numbers as well as additional ablations, including increasing the RoPE base frequency (ABF, Xiong et al. (2023) and per-head experiments can be found in Appendix D. A smaller ladder on code pretraining can be found in Appendix F.

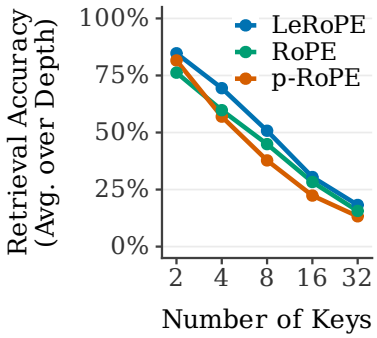

Figure 2: Multi-key Needle-in-a-Haystack accuracy as a function of the number of distractor keys evaluated on sequences in-distribution.

5.2 In-Distribution Needle-in-a-Haystack

Needle in a Haystack (NIAH) is a standard probe of context utilization and retrieval: given a key, the model must retrieve its associated value from a context of filler text and, optionally, distractor key–value pairs. Examples are generated synthetically; we vary the depth of the gold pair (the "needle") within the context, using filler text from the original NIAH repository.[1]

We evaluate the 2.5B models with 1 to 31 distractor pairs across all three methods placing the needle at depths $\{0, 0.1, \ldots, 1.0\} \times L_{\text{train}}$. Figure 2 plots retrieval accuracy against distractor count, averaged over depths: LeRoPE outperforms RoPE and p-RoPE at every shared count, and continues to outperform RoPE at 31 distractor pairs, the hardest setting evaluated. All evaluations here stay within L_{train}, as this section targets in-distribution behavior; we report only the multi-key variant because single-needle retrieval is saturated for all models in-distribution, and we defer length extrapolation to Section 7.

Section Summary: *LeRoPE is 3.4% more compute efficient than RoPE, leading to a final model with better in-distribution performance for a fixed compute budget.*

[1] https://github.com/gkamradt/needle-in-a-haystack

6 Analysis: Learned Frequency Patterns

We now study what LeRoPE learns. We examine the learned frequencies directly and identify a single band that carries most of the positional signal.

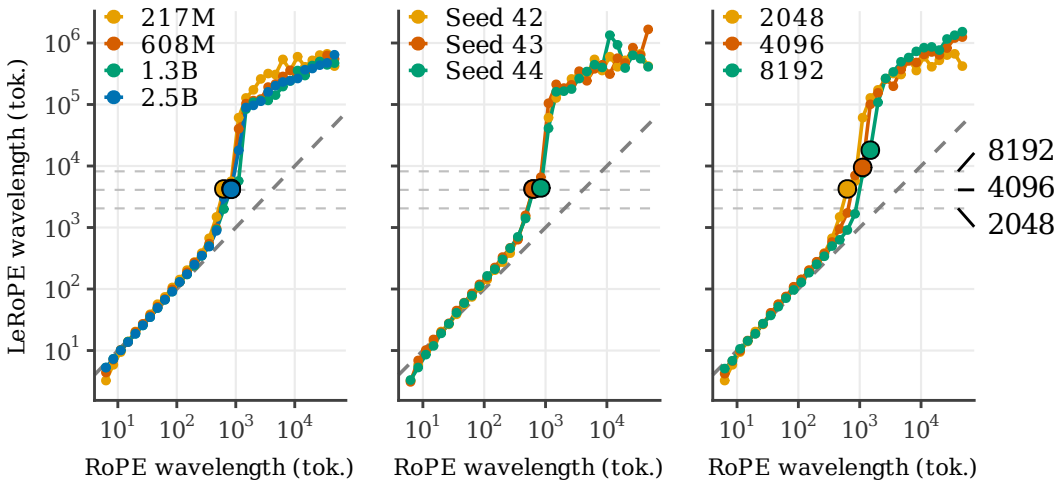

Figure 3: **(a)**: Variation in learned wavelengths across model sizes. **(b)**: Variation of learned wavelengths across 3 random seeds. **(c)**: Variation across different training lengths. Dashed horizontal lines mark shared 2048-, 4096-, and 8192-token context-length references; circled dots mark the dominant positional band; labels appear in the right panel. All plots show LeRoPE learns very similar wavelength profiles regardless of initialization and model size. The center and right plots are for the 217M models.

Learned Frequencies. Figure 3 plots each learned band's wavelength against RoPE's counterpart: low-wavelength bands closely track RoPE ($y = x$), mid-wavelength bands begin to diverge, and high-wavelength bands are pushed to a much larger, near-constant wavelength – i.e. their frequency is driven toward zero. High-frequency (low-wavelength) bands rotate slightly faster, with the effect more pronounced for smaller scales. Mid-frequency bands begin to slow compared to RoPE frequencies, and low frequencies are driven down even closer to zero. This pattern is consistent across all model scales and seeds we test (Figure 3a,b).

The point at which frequencies are suppressed is linked to the sequence lengths seen during training (Figure 3c): models trained on longer contexts diverge from the RoPE frequencies at lower frequencies than their short-context counterparts.

Identifying Dominant Positional Bands. Equation 3 provides a per-band decomposition of the attention logit contributions. By plotting the average contribution per relative distance, we find (Figure 4) that one band dominates in a nearly purely positional pattern for models with LeRoPE.

The dominant band's wavelength tracks the training length across six runs: four model scales at $L_{\text{train}} = 2048$, plus 217M runs at $L_{\text{train}} \in \{4096, 8192\}$. The dominant band's wavelength is $\approx 2.205\, L_{\text{train}}$ (exact wavelengths can be in Table 10). Fixed-frequency RoPE models trained at the same scale do not exhibit a dominant band. Instead, their positional contributions are distributed across several bands at far smaller magnitudes, although, as noted by Oka et al. (2026a), they do contain strongly oscillatory channels.

Learning With Fixed LeRoPE Frequencies. Next, we ask whether the improvements from LeRoPE emerge from jointly training frequencies with the rest of the network, or solely from the discovery of a more optimal set of frequencies. To distinguish these, we train a 217M model with its frequencies frozen to the values learned by an independent LeRoPE run at the same scale (Table 2).

Training with fixed LeRoPE frequencies yields 63% of the improvement of full LeRoPE, suggesting that joint training dynamics and the final frequency set are both partially responsible for performance

gains. Under the same settings, p-RoPE captures 10.4% of the performance gains of LeRoPE. We can interpret p-RoPE as a coarse approximation of final LeRoPE frequencies, preserving faster frequencies while suppressing slower ones.

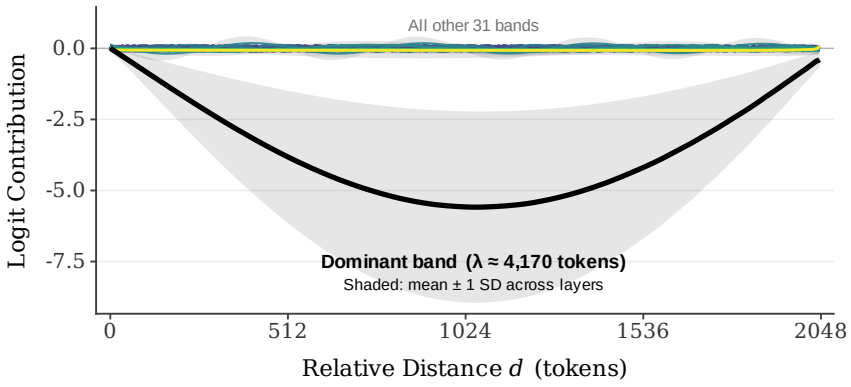

Figure 4: Band-level contributions to the logit score as a function of relative distance $d = s - t$ (in tokens), in a 2.5B LeRoPE model; the shaded band is ± 1 standard deviation across the model's layers. LeRoPE produces a dominant positional band that penalizes logit scores according to their relative position; despite this, LeRoPE outperforms standard RoPE in perplexity and NIAH evaluations at all in-distribution offsets.

Section Summary: *LeRoPE learns the same frequency profile across seeds and scales: frequencies too slow for the training window are suppressed, and a single dominant band emerges at $\lambda \approx 2.2\, L_{train}$. Training with the learned frequencies frozen captures 63.6% of the gain.*

7 Extrapolation

While the main focus of our work is on in-distribution performance, understanding how LeRoPE extrapolates is of practical importance. RoPE-based models degrade sharply when operating on sequence lengths longer than L_{train} with unmodified positions (Press et al., 2021). The standard solution is to rescale frequencies back in-distribution at inference via positional interpolation (PI) (Chen et al., 2023) or NTK-by-parts scaling with attention temperature (Peng et al., 2023).

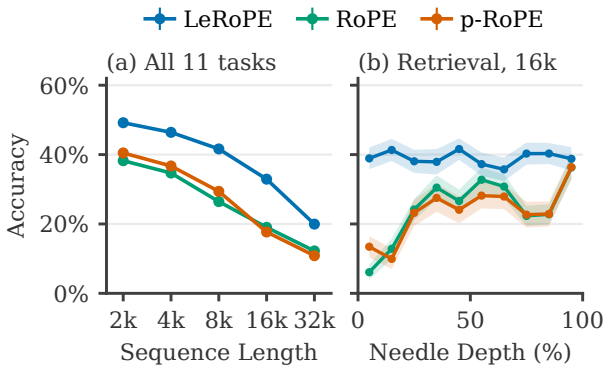

Figure 5: RULER with YaRN for the 2.52B models. (a) Accuracy averaged over all 11 tasks. (b) Retrieval accuracy at 16k over the eight needle-in-a-haystack tasks, by the depth of each prompt's first answer needle (0% is the start of the context). Per-task results are in Table 14.

Naive Extrapolation. With unmodified positions, LeRoPE degrades more sharply than RoPE and p-RoPE (Section E.2, Figure 10). The dominant frequency band's logit contribution follows a negative half-cycle of a sinusoid with period $\approx 2.2\, L_{\text{train}}$, so within the training window it remains negative or near zero. On relative distances beyond those seen during training, the contribution becomes positive. To measure the effect of the dominant positional band alone, we measure the performance of *dominant-band interpolation*, which applies positional interpolation only to the dominant positional band. We find that this prevents the rapid out-of-distribution perplexity explosion, roughly matching the perfor-

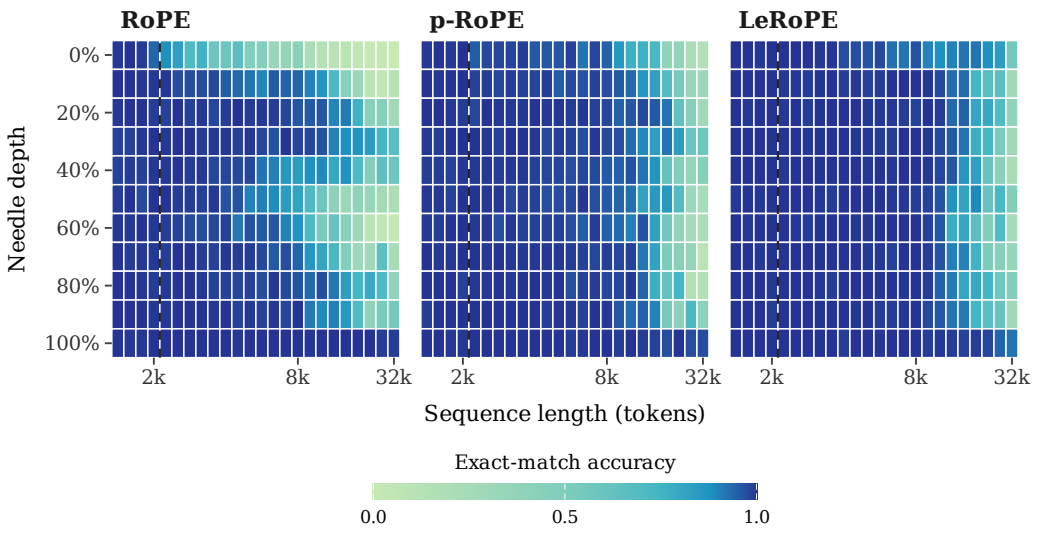

Figure 6: Single-needle NIAH accuracy with NTK-by-parts + YaRN for RoPE, p-RoPE, and LeRoPE on sequences up to 32k. We show that LeRoPE generally outperforms both RoPE and p-RoPE, especially at longer sequence lengths and for keys early in the context.

mance of NTK-by-parts (Table 11). For extrapolation evaluations we filter the validation set of C4 for documents longer than 4096 in length.

Compatibility with YaRN. Standard context-extension methods transfer to LeRoPE directly. LeRoPE with NTK-by-parts scaling and YaRN's attention temperature outperforms the same recipe applied to RoPE and p-RoPE, and also outperforms dominant-band interpolation combined with a temperature. For each method we sweep over YaRN's temperature hyper-parameter to ensure a fair comparison. LeRoPE's in-distribution advantage therefore carries over under the extension tooling practitioners already use; the best configuration for extrapolation overall is LeRoPE with YaRN (Table 11). See Appendix E.2 for a description of NTK-by-parts and YaRN.

Long-context retrieval. We evaluate the 2.52B models on the 11 synthetic tasks of RULER (Hsieh et al., 2024) (excluding QA) at lengths from 2k to 32k and apply YaRN for sequences longer than L_{train}. The complete per-split results can in Table 14 in Appendix E.3. Across all sequence lengths LeRoPE outperforms both RoPE and p-RoPE while maintaining stronger performance over all positions in the context(Figure 5).

To visualize where retrieval differs across each model, we additionally run single-needle NIAH with fixed needles depths ranging from 0% to 100% of the sequence length up to 32k and generate a heatmap (Figure 6). Single-needle retrieval is saturated within L_{train} (§5.2); beyond it, averaged across depths, LeRoPE outperforms both RoPE and p-RoPE at every evaluated length, with the largest gains for needles early in the context.

8 Conclusion

We propose LeRoPE, which makes RoPE's frequencies learnable through one scale per frequency band. Across a scaling ladder from 52M to 2.5B parameters, LeRoPE outperforms RoPE and partial RoPE at every scale, with RoPE requiring 3.4% more compute to match it at the largest scale. Given its consistent performance improvements on in-distribution language modeling, future work may explore the merits of learned RoPE parameters in local layers of local-global models (Kamath et al., 2025).

9 AI USE STATEMENT

We used generative AI tools to assist in the implementation of our training pipelines, drafting and proofing of the text, and verifying the derivations of the gradients of our method. The authors checked the resulting code and text for correctness and take full responsibility for everything presented in this work.

REFERENCES

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Velivckovi'c. Round and round we go! what makes rotary positional encodings useful? *ArXiv*, abs/2410.06205, 2024. URL https://api.semanticscholar.org/CorpusID:273229201.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws, 2023. URL https://arxiv.org/abs/2210.14891.

Zachary Charles, Gabriel Teston, Lucio Dery, Keith Rush, Nova Fallen, Zachary Garrett, Arthur Szlam, and Arthur Douillard. Communication-efficient language model training scales reliably and robustly: Scaling laws for diloco. *ArXiv*, abs/2503.09799, 2025. URL https://api.semanticscholar.org/CorpusID:276961579.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *ArXiv*, abs/2306.15595, 2023. URL https://api.semanticscholar.org/CorpusID:259262376.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022. URL https://api.semanticscholar.org/CorpusID:247951931.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *ArXiv*, abs/2402.13753, 2024. URL https://api.semanticscholar.org/CorpusID:267770308.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Damien Ferbach, Courtney Paquette, Gauthier Gidel, Katie Everett, and Elliot Paquette. Logarithmic-time schedules for scaling language models with momentum. 2026. URL https://api.semanticscholar.org/CorpusID:285303683.

Olga Golovneva, Tianlu Wang, Jason E. Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *ArXiv*, abs/2405.18719, 2024. URL https://api.semanticscholar.org/CorpusID:270094992.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:268536717.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *ArXiv*, abs/2404.06654, 2024. URL https://api.semanticscholar.org/CorpusID:269032933.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Ning Ding, Youbang Sun, Biqing Qi, Yuchen Fan, Xuekai Zhu, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. *ArXiv*, abs/2412.17739, 2024. URL https://api.semanticscholar.org/CorpusID:274992805.

Zhi-Gao Huang and Mu-Sheng Chen. Optimizing the learnable rope theta parameter in transformers. *IEEE Access*, 13:131271–131288, 2025. URL https://api.semanticscholar.org/CorpusID:280291900.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Istvan Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr'as Gyorgy, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Boxi Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, Cj Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Zhiyuan Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, J. Michael Wieting, Jonathan Lai, Jordi Orbay, Joe Fernandez, Joshua Newlan, Junsong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Ping mei Xu, Piotr Stańczyk, Pouya Dehghani Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Ardeshir Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vladimir Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira,

Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab S. Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and L'eonard Hussenot. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025. URL https://api.semanticscholar.org/CorpusID:277313563.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Petros Karypis, Julian J. McAuley, and George Karypis. Extending input contexts of language models through training on segmented sequences. *ArXiv*, abs/2310.14633, 2023. URL https://api.semanticscholar.org/CorpusID:264426090.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *ArXiv*, abs/2305.19466, 2023. URL https://api.semanticscholar.org/CorpusID:258987259.

Haoran Li, Yingjie Qin, Bao-Yuan Ou, Lai Xu, and Ruiwen Xu. Hope: Hybrid of position embedding for long context vision-language models. *Advances in Neural Information Processing Systems 38*, 2025. URL https://api.semanticscholar.org/CorpusID:278911120.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you! 2023.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chen An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. *ArXiv*, abs/2310.05209, 2023. URL https://api.semanticscholar.org/CorpusID:263828829.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. URL https://api.semanticscholar.org/CorpusID:198953378.

Manas Mehta, Fangcong Yin, and Greg Durrett. Randomized yarn improves length generalization for long-context reasoning. *ArXiv*, abs/2606.23687, 2026. URL https://api.semanticscholar.org/CorpusID:289629978.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of rope bounds context length. *ArXiv*, abs/2405.14591, 2024. URL https://api.semanticscholar.org/CorpusID:269983770.

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher. Selective rotary position embedding. *ArXiv*, abs/2511.17388, 2025. URL https://api.semanticscholar.org/CorpusID:283226311.

Yui Oka, Taku Hasegawa, Kyosuke Nishida, and Kuniko Saito. Wavelet-based positional representation for long context. *ArXiv*, abs/2502.02004, 2025. URL https://api.semanticscholar.org/CorpusID:276107428.

Yui Oka, Kentaro Hanafusa, Taku Hasegawa, Kyosuke Nishida, and Kuniko Saito. Probing rotary position embeddings through frequency entropy. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=1JZuEDq62N.

Yui Oka, Itsumi Saito, Kyosuke Nishida, and Kuniko Saito. Frequency bands in roPE: Base frequency and context length shape the interpolation–extrapolation trade-off. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=PR1PPxvG9Q.

Sophie Ostmeier, Brian Axelrod, Maya Varma, Michael E. Moseley, Akshay S. Chaudhari, and Curtis P. Langlotz. Liere: Lie rotational positional encodings. In *International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:276422590.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *ArXiv*, abs/2309.00071, 2023. URL https://api.semanticscholar.org/CorpusID:261493986.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *ArXiv*, abs/2108.12409, 2021. URL https://api.semanticscholar.org/CorpusID:237347130.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019. URL https://api.semanticscholar.org/CorpusID:204838007.

Anian Ruoss, Gr'egoire Del'etang, Tim Genewein, Jordi Grau-Moya, R. Csordás, Mehdi Abbana Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. *ArXiv*, abs/2305.16843, 2023. URL https://api.semanticscholar.org/CorpusID:258947457.

Connor Schenck, Isaac Reid, Mithun George Jacob, Alex Bewley, Joshua Ainslie, David Rendleman, Deepali Jain, Mohit Sharma, Kumar Avinava Dubey, Ayzaan Wahid, Sumeet Singh, Ren'e Wagner, Tianli Ding, Chuyuan Kelly Fu, Arunkumar Byravan, Jake Varley, Alexey Gritsenko, Matthias Minderer, Dmitry Kalashnikov, Jonathan Tompson, Vikas Sindhwani, and Krzysztof Choromanski. Learning the ropes: Better 2d and 3d position encodings with string. *ArXiv*, abs/2502.02562, 2025. URL https://api.semanticscholar.org/CorpusID:276107271.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074/.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, abs/2104.09864, 2021. URL https://api.semanticscholar.org/CorpusID:233307138.

The Microsoft AI Team. Mai-thinking-1: Building a hill-climbing machine. Technical report, Microsoft AI (MAI), 2026. URL https://microsoft.ai/pdf/mai-thinking-1.pdf.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:13756489.

Ali Veisi, Delaram Fartoot, and Hamidreza Amirzadeh. Context-aware rotary position embedding. *ArXiv*, abs/2507.23083, 2025. URL https://api.semanticscholar.org/CorpusID:280400946.

Shaowen Wang, Yuke Zheng, Tansheng Zhu, Shuang Chen, Shao-Hua Liu, Sun-Cong Zheng, and Jian Li. Adarope: Not all attention heads should rotate and scale equally. *ArXiv*, abs/2607.19363, 2026. URL https://api.semanticscholar.org/CorpusID:290442849.

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 long-context scaling of foundation models, 2023. URL https://arxiv.org/abs/2309.16039.

Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr F. Locatelli. Rope to nope and back again: A new hybrid attention strategy. *ArXiv*, abs/2501.18795, 2025a. URL https://api.semanticscholar.org/CorpusID:276079501.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. Path attention: Position encoding via accumulating householder transformations. *ArXiv*, abs/2505.16381, 2025b. URL https://api.semanticscholar.org/CorpusID:278789395.

Hao Yu, Tangyu Jiang, Shuning Jia, Shannan Yan, Shunning Liu, Haolong Qian, Guanghao Li, Shuting Dong, Huaisong Zhang, and Chun Yuan. Comrope: Scalable and robust rotary position embedding parameterized by trainable commuting angle matrices. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4508–4517, 2025. URL https://api.semanticscholar.org/CorpusID:279155450.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022. URL https://api.semanticscholar.org/CorpusID:248496292.

Supplementary Materials
Learnable RoPE Frequencies Improve Language Modeling

A Ladder Notes **16**
- A.1 Architecture . . . 16
- A.2 Compute Multiplier Estimation . . . 16
- A.3 Seed Variability . . . 18

B Extended Related Work **19**

C Gradient Dynamics **20**
- C.1 Definitions and Lemmas . . . 20
- C.2 Phasor Plots . . . 23

D Additional Ablations **26**

E Extrapolation and Dominant Band Wavelengths **28**
- E.1 Dominant Positional Band Wavelength . . . 28
- E.2 Extrapolation Details . . . 28
- E.3 Complete RULER Results . . . 29

F Code Pretraining **30**

A Ladder Notes

A.1 Architecture

We use the Enoki scaling ladder of Charles et al. (2025); Ferbach et al. (2026): the head dimension is fixed at 64 and depth and width are co-scaled, yielding models on the shallower, wider end. The architecture uses RoPE, pre-norm, QK-norm applied *before* RoPE, no biases, and no weight tying. Table 4 lists the model configurations.

We deviate from the Enoki recipe as follows:

- **Data and tokenizer:** C4 (Raffel et al., 2019) with a 32,000-token SentencePiece vocabulary, rather than FineWeb with the 50,304-token GPT-2 tokenizer.
- **Document masking:** We mask attention across document boundaries within packed sequences; prior Enoki work does not state whether it does so.
- **Normalization:** RMSNorm in place of LayerNorm for the pre-norms, QK-norm, and final norm.
- **Learning-rate schedule:** cosine decay to zero rather than to $0.1\times$ peak, with a warmup of 1000 steps rather than 2% of total steps.
- **Optimizer:** AdamW with $\beta_2 = 0.95$ rather than 0.999 ($\beta_1 = 0.9$ in both), and independent weight decay $\lambda = 8/T$, where T is the total number of training steps (with weight decay of $\lambda = \omega/T$, Ferbach et al. report $\omega = 4$ to be robust; both $\omega = 4$ and $\omega = 8$ lie within their swept range).
- **Batch size:** 64 sequences (global batch size of $131{,}072$ tokens) instead of 32, in line with the batch-size findings of Charles et al. (2025) for this ladder.
- **Gradient clipping:** global-norm 1.0 instead of 0.5.

All weights are initialized with fan-in scaling, std $= 1/\sqrt{\text{fan-in}}$; the MLP and attention output projections receive an additional depth-dependent factor, std $= 1/\sqrt{2\, n_{\text{layer}} \cdot \text{fan-in}}$; and token embeddings use std $= 1/\sqrt{d_{\text{embd}}}$, giving each embedding vector unit expected norm. See Table 5 for the complete set of hyper-parameters.

A.2 Compute Multiplier Estimation

Choice of interpolation axis. We use the piecewise-linear baseline of Ferbach et al. (2026), but interpolate in $\log C$ rather than raw C. For compute budgets outside the ladder, we continue the nearest line segment. Loss–compute curves are convex. Linear interpolation in C lies above the true curve between ladder points, so inverting this baseline overestimates the compute budget and the multiplier. With nine scales and adjacent budgets within a factor of $\sim$2–7, the effect is small on their ladder. Across our decade-wide gaps, it is severe in C but smaller in $\log C$, where the power-law curve is far straighter. At the largest scale, LeRoPE's loss is below that of every RoPE model. Estimating the multiplier requires extrapolating the RoPE curve, so the top-scale values are conservative.

Parametric alternative: efficiency gain. The more common estimator fits a scaling law to the baseline ladder (Kaplan et al., 2020; Hoffmann et al., 2022; The Microsoft AI Team, 2026),

$$L = f(C) = AC^{-\alpha} + E, \tag{8}$$

and reports the efficiency gain $\text{EG} = f^{-1}(L')/C'$ for a variant reaching loss L' at cost C'. Fitting the RoPE ladder gives $A = 26.4$, $\alpha = 0.0574$, and $E = 0.641$. The fitted α is within the range reported for fits with an irreducible term by Ferbach et al. (2026, Table 15). Table 6 compares the efficiency gains with the piecewise multipliers of Section 4.2, which differ by at most 0.009 at every scale.

Fit instability in compute multiplier estimates. With five ladder points and three parameters, the irreducible loss E is weakly identified. Every $E \in [0.3, 1.0]$ fits the ladder with RMSE below 0.006

Table 3: Optimal peak learning rates per scale. All three positional-encoding methods share the same optimum at every swept scale; for the two largest models the grid index is extrapolated from the smaller scales, across which the optimum tracks $\eta \propto 1/N_{\text{tot}}$.

Scale	Peak LR η	Determination
52M	1.10×10^{-2}	swept
217M	2.76×10^{-3}	swept
608M	9.77×10^{-4}	swept
1.34B	4.88×10^{-4}	extrapolated
2.52B	3.45×10^{-4}	extrapolated

Table 4: Architectural details of our scaling ladder following Ferbach et al. (2026). Head dimension is fixed to 64 across scale, head count is $\frac{4}{3} n_{\text{layer}}$, width is $d_{\text{embd}} = 64 \times$ heads. We use a vocabulary of size 32,000 and include the non-embedding (N_{ne}) and total (N_{ne}) parameter counts.

Layers	Embd Dim	MLP Hidden	N_{ne}	N_{tot}
6	512	2,048	18.87M	51.64M
12	1,024	4,096	150.99M	216.53M
18	1,536	6,144	509.61M	607.91M
24	2,048	8,192	1.21B	1.34B
30	2,560	10,240	2.36B	2.52B

Table 5: Training Hyperparameters

Parameter	**Value**
Sequence length	2048
Batch size (sequences)	64
Tokens per batch	131,072
Vocabulary size	32,000
Optimizer	AdamW (β_1=0.9, β_2=0.95)
Adam ϵ	1e-8
Warmup window	1000 steps
LR rule, $\gamma(t)$	cosine decay
Final LR	0
Precision/Optimizer state precision	bfloat16/float32
Gradient clipping	1.0 (global norm)
Weight-decay[1], λ	$8/T$ (independent)

[1] T denotes total training steps; weight decay omitted on LeRoPE parameters.

nats (Table 7). The RMSE differences between these fits are comparable to per-run seed variability (Table 9), so run-to-run noise can move the best-fit E across this range. A misspecified fit can have residuals larger than the 0.003–0.010-nat gaps being measured, whereas the piecewise estimator has no residual at measured scales. Ferbach et al. (2026) report the same identifiability issue. Their fitted single-power-law exponent ranges from 0.048 to 0.074 as the assumed saturation level changes. Our ladder is sparser and our effect is smaller, so the problem is more severe.

Sharing the saturation across methods. Following Ferbach et al. (2026), we fit RoPE, p-RoPE, and LeRoPE jointly with a shared E and per-method (A, α). The best fit ($E = 0.643$, joint RMSE 4.7×10^{-4} nats) leaves the efficiency gains unchanged to within 0.001. The identifiability window does not narrow; every shared $E \in [0.3, 1.0]$ still fits within 0.006 nats. Because the three curves are nearly parallel, adding methods provides little information about the floor. The degeneracy comes from the range of scales, not the number of points. The fitted exponents agree to within 0.5% ($\alpha \in [0.0573, 0.0576]$), similarly to Ferbach et al. (2026).

Richer functional forms. Ferbach et al. (2026) fit a double power law $L = a + bC^{-c} + eC^{-f}$ across five optimizer curves with a shared saturation a. It is identifiable with ∼10 scales per curve, but our five-point ladder has no residual degrees of freedom. Additionally, a hinge-type broken power

Table 6: LeRoPE compute multipliers under the two estimators.

	52M	217M	608M	1.34B	2.52B
Piecewise $\log C$ (ours)	1.071	1.048	1.047	1.032	1.034
Efficiency gain (fitted)	1.064	1.047	1.039	1.039	1.031

Table 7: Sensitivity of the fitted efficiency gain to the assumed irreducible loss E. Every $E \in [0.3, 1.0]$ fits the RoPE ladder with RMSE below 0.006 nats, yet the implied per-scale gains vary by up to 0.14 and include impossible values below 1 (bold), despite LeRoPE reaching strictly lower loss at every scale.

			EG per scale				
E	α	RMSE (nats)	52M	217M	608M	1.34B	2.52B
0.0	0.044	$6.4\text{e}{-3}$	1.028	1.118	1.090	1.031	**0.952**
0.3	0.050	$3.8\text{e}{-3}$	1.043	1.089	1.069	1.034	**0.982**
0.5	0.054	$1.8\text{e}{-3}$	1.055	1.065	1.052	1.036	1.008
0.641	0.057	$5.3\text{e}{-4}$	1.064	1.047	1.039	1.039	1.031
0.8	0.062	$2.2\text{e}{-3}$	1.075	1.024	1.022	1.043	1.063
1.0	0.069	$5.5\text{e}{-3}$	1.090	**0.989**	**0.997**	1.051	1.118

law (Caballero et al., 2023) fits our ladder *worse* than the saturating form (RMSE 1.1×10^{-3} vs. 5.3×10^{-4} nats), recovers nearly equal exponents on both sides of a weakly identified break, and exhibits the same efficiency-gain instability, including values below 1. The ladder favors saturation over a break.

We therefore report efficiency gain as corroboration and keep the piecewise estimator primary.

A.3 Seed Variability

Seeds control model initialization and data ordering. To bound their effect, we repeat the 217M runs (as well as RoPE with two additional bases) and the 608M runs under three matched seeds (Table 9). All comparisons are within a seed: the two runs share initialization draws and data order, so their loss fluctuations are strongly correlated and the paired margins vary far less than the marginal deviations ($s_d = 0.0019$ vs. σ up to 0.0020 at 217M; $5s_d = 0.0010$ at 608M). LeRoPE improves on every baseline under every 5 individual seeds at both scales, and per-method variability decreases with scale (σ_{LeRoPE}: $0.0019 \to 0.0004$), supporting the reading that the unseeded larger-scale gaps in Table 1 also exceed noise.

B Extended Related Work

Modified rotary frequencies. Several works offer alterative and sometimes trainable rotational operations for RoPE's. Ostmeier et al. (2024) extend RoPE to Lie groups and replaces RoPE's 2D block diagnal rotation matrix with a learned skew-symmetric matrix. Huang and Chen (2025) learn one multiplier per layer that rescales every band of that layer by the same factor, maintaining a geometric series of frequencies. AdaRoPE (Wang et al., 2026) directly learns log-frequencies per-head per-layer as well as a per-head attention temperature. Other work makes rotations input-dependent, generating frequencies or phases from token content (Veisi et al., 2025; Movahedi et al., 2025) or learn commuting angle matrices (Yu et al., 2025). Outside of language modeling, RoPE-Mixed (Heo et al., 2024) learns per-head, per-layer 2D frequencies for vision. Subsequent work (Schenck et al., 2025) extended this to learnable 2D and 3D generators.

Hand-designed spectra. A second line of work fixes a modified spectrum before training. p-RoPE (Barbero et al., 2024) removes rotation from the slowest bands; FoPE (Hua et al., 2024) models each dimension as a Fourier series and zeroes frequency components whose period exceeds the training length; HoPE (Li et al., 2025) retains only high-frequency components and replaces the rest with position-independent ones. Changing the base moves the whole geometric series: large bases are standard for long-context models (Xiong et al., 2023), the base bounds the usable context length (Men et al., 2024), and matching it to the training length (Oka et al., 2026b) trades interpolation for extrapolation. Hybrid architectures instead interleave RoPE and NoPE layers (Yang et al., 2025a). Oka et al. (Oka et al., 2025) replace RoPE's sinusoids with multi-scale Ricker wavelets. The spectrum LeRoPE learns recovers the element common to many of these designs, suppression of bands too slow for the training window, without choosing the cutoff by hand. It additionally develops a dominant positional band with a period of roughly twice the training length which a FoPE-style cutoff would remove.

Context extension. Positional interpolation (Chen et al., 2023) and YaRN (Peng et al., 2023) rescale frequencies at inference or during short fine-tuning. LongRoPE (Ding et al., 2024) searches for non-uniform per-dimension rescaling factors. Another line of research Ruoss et al. (2023); Karypis et al. (2023); Mehta et al. (2026) proposed efficient fine-tuning methods for length extension by training on short, randomized sequences. These methods adapt to longer sequences after pretraining and can be applied to LeRoPE models to further improve long context performance.

C GRADIENT DYNAMICS

For clarity in this section, we will omit the pre-softmax scaling factor $1/\sqrt{d}$, operate on a single head, and take o_i to be the post-softmax accumulation of values before output projection W_o.

C.1 DEFINITIONS AND LEMMAS

We begin with the following definitions:

$$
\begin{aligned}
g_s &\triangleq \frac{\partial \mathcal{L}}{\partial o_s} \\
S_{st}^{(m)} &\triangleq (s-t)\|q_s^{(m)}\|\|k_t^{(m)}\| \sin\left(\phi_{st}^{(m)} + (s-t)\theta^{(m)}\right) \\
C_m(\delta) &\triangleq \sum_{\substack{s,t \\ s-t=\delta}} \alpha_{st}\|q_s^{(m)}\|\|k_t^{(m)}\| e^{i\phi_{st}^{(m)}} (v_t - o_s)^T g_s \\
P(\theta) &= \text{Re}\left(\sum_{\delta=1}^{L-1} C(\delta) e^{i\delta\theta}\right)
\end{aligned}
$$

Lemma C.1. *Let $g_s = \partial\mathcal{L}/\partial o_s$ be the gradient of the loss with respect to output vector o_s. Let $\partial\mathcal{L}/\partial D_{st}^{(m)}$ be the gradient of the loss with respect to the dot product produced by band m from query position s to key position t.*

Then $\partial\mathcal{L}/\partial D_{st}^{(m)} = \alpha_{st} g_s^T (v_t - o_s)$.

Proof. First, note that the only effect of the value $D_{st}^{(m)}$ is through its effect on the output vector at position s, o_s.

Thus, we have:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial D_{st}^{m}} &= \left(\frac{\partial \mathcal{L}}{\partial o_s}\right)^T \left(\frac{\partial o_s}{\partial D_{st}^{(m)}}\right) \\
&= g_s^T \left(\frac{\partial o_s}{\partial D_{st}^{(m)}}\right)
\end{aligned}
$$

So we only have to find the partial derivative of o_i with respect to the band-s QK product.

The softmax Jacobian gives:

$$
\begin{aligned}
\frac{\partial o_s}{\partial z_{st}} &= \sum_{t'<s} \alpha_{st'}(\mathbb{1}[t=t'] - \alpha_{st}) v_{t'} \\
&= \alpha_{st}(v_t - o_s) \\
\frac{\partial z_{st}}{\partial D_{st}^{(m)}} &= 1 \\
\implies \frac{\partial o_s}{\partial D_{st}^{(m)}} &= \alpha_{st}(v_t - o_s) \\
\implies \frac{\partial \mathcal{L}}{\partial D_{st}^{(m)}} &= \alpha_{st} g_s^T (v_t - o_s)
\end{aligned}
$$

□

Lemma C.2. *Define $P_m(\theta)$ as the real component of the discrete-time Fourier Transform of the signal $C_m(\delta)$ with frequency θ, as above.*

Then $\frac{\partial \mathcal{L}}{\partial \theta^{(m)}} = P'_m(\theta^{(m)})$.

Proof. First, consider the full sum from Equation 6. We may split the sum by relative offsets $\delta \triangleq s-t$, and for notational compactness define $g_{st}^{(m)} \triangleq \partial \mathcal{L}/\partial D_{st}^{(m)}$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta^{(m)}} &= -\sum_{t<s} (s-t)\,\|q_s\|\|k_t\| \sin\left(\phi_{st} + (s-t)\theta^{(m)}\right) \frac{\partial \mathcal{L}}{\partial D_{st}^{(m)}} \\
&= -\sum_{\delta=1}^{L-1} \sum_{\substack{s,t \\ s-t=\delta}} \delta\,\|q_s\|\|k_t\| \sin\left(\phi_{st} + \delta\theta^{(m)}\right) g_{st}^{(m)} \\
&= -\sum_{\delta=1}^{L-1} \sum_{\substack{s,t \\ s-t=\delta}} \delta\,\|q_s\|\|k_t\| \sin\left(\phi_{st} + \delta\theta^{(m)}\right) \alpha_{st}(v_t - o_s)^T g_s
\end{aligned}$$

To split the sine, write it as an exponential and note that all scalar terms outside of the sine are real, allowing us to apply the imaginary operator across the final summed value:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta^{(m)}} &= -\sum_{\delta=1}^{L} \sum_{\substack{s,t \\ s-t=\delta}} \delta \mathrm{Im}\left(e^{i\phi_{st}} e^{i\delta\theta^{(m)}}\right) \|q_s\|\|k_t\|\alpha_{st}(v_t - o_s)^T g_s \\
&= -\mathrm{Im}\left(\sum_{\delta=1}^{L-1} \sum_{\substack{s,t \\ s-t=\delta}} \delta e^{i\phi_{st}} e^{i\delta\theta^{(m)}} \|q_s\|\|k_t\|\alpha_{st}(v_t - o_s)^T g_s\right) \\
&= -\mathrm{Im}\left(\sum_{\delta=1}^{L-1} \delta e^{i\delta\theta^{(m)}} \sum_{\substack{s,t \\ s-t=\delta}} e^{i\phi_{st}} \|q_s\|\|k_t\|\alpha_{st}(v_t - o_s)^T g_s\right) \\
&= -\mathrm{Im}\left(\sum_{\delta=1}^{L-1} \delta C(\delta) e^{i\delta\theta^{(m)}}\right) \\
&= -\mathrm{Im}\left(\sum_{\delta=1}^{L-1} \left(-i\frac{d}{d\theta^{(m)}}\left[C(\delta)e^{i\delta\theta^{(m)}}\right]\right)\right) \\
&= \frac{d}{d\theta^{(m)}} \mathrm{Re}\left(\sum_{\delta=1}^{L-1} C(\delta) e^{i\delta\theta^{(m)}}\right) \\
&= P'_m(\theta^{(m)})
\end{aligned}$$

□

Lemma C.3. *As previously, define $S_{st}^{(m)}$ to be the offset-weighted sine term at RoPE-band m; let v_t be the attention value vector at position s and g_s be the downstream loss gradient $\partial \mathcal{L}/\partial o_s$. Then the gradient with respect to RoPE parameter $\theta^{(m)}$ is as follows:*

$$\frac{\partial \mathcal{L}}{\partial \theta^{(m)}} = -\sum_{s} \mathit{Cov}_{t\sim p_s(t)}\left(v_t^T g_s, S_{st}^{(m)}\right)$$

Proof. As in Lemma C.2, note

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta^{(m)}} &= -\sum_{s=1}^{s_{max}} \sum_{t=0}^{s-1} \alpha_{st}(s-t)\,\|q_s\|\|k_t\| \sin\left(\phi_{st} + (s-t)\theta^{(m)}\right)(v_t - o_s)^T g_s \\ &= -\sum_{s=1}^{s_{max}} \sum_{t=0}^{s-1} \alpha_{st} S_{st}^{(m)} (v_t - o_s)^T g_s \\ &= -\sum_{s=1}^{s_{max}} \mathbb{E}_{t\sim p_s(t)}\left[S_{st}^{(m)} (v_t - o_s)^T g_s\right]\end{aligned}$$

□

Now, note that $\mathbb{E}_{t\sim p_s(t)}[(v_t - o_s)^T g_s] = 0$, as $\mathbb{E}_{t\sim p_s(t)}[v_t] = o_s$, and g_s factors out.

Thus, the expected product of the two terms equals their covariance.

So, as desired, we obtain:

$$\frac{\partial \mathcal{L}}{\partial \theta^{(m)}} = -\sum_{s=1}^{s_{max}} \mathrm{Cov}\left(S_{st}^{(m)}, v_t^T g_s\right)$$

In expectation, with enough samples, in the absence of periodic structure in downstream attention-logit gradients, we would expect this update term to remain centered roughly around zero.

C.2 PHASOR PLOTS

In Figure 8, we plot the per-offset gradient phasors for all bands, for both LeRoPE and RoPE models. The final gradient corresponds to the real component of the aggregate phasor. The dashed line denotes the direction of this aggregate phasor, and the red dot the aggregate phasor, normalized between 0 (full cancellation) and 1 (full coherence).

Gradient (length-weighted) phasors take on similar scales across offsets, as compared to the non-length-weighted objective phasors. This phenomenon is visible in Figure 7.

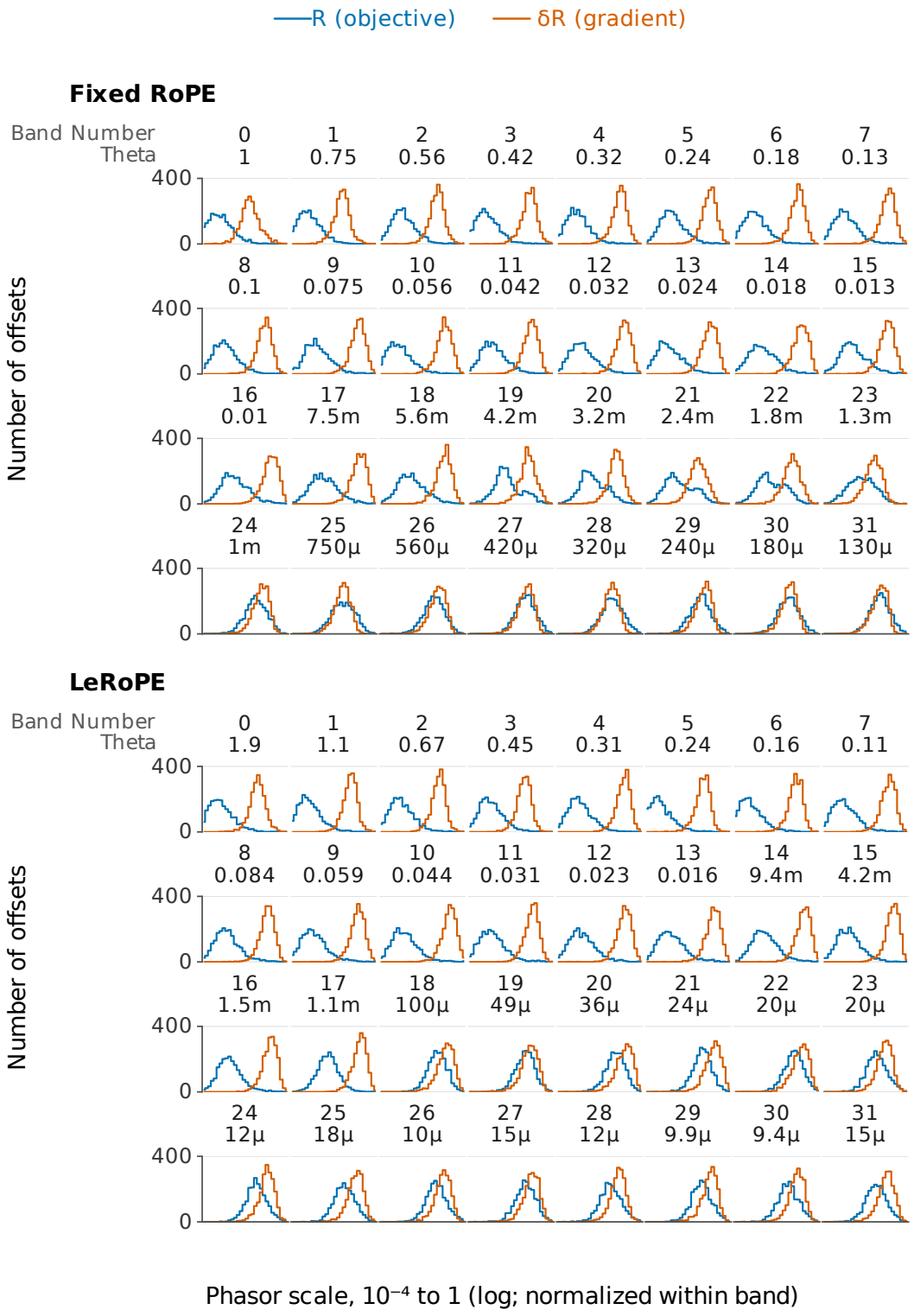

Figure 7: Normalized per-offset phasor-scale distributions for fixed RoPE (top) and LeRoPE (bottom) at $L = 2048$. Panel headers report the band number and operating θ, with m$= 10^{-3}$ and $\mu = 10^{-6}$. Blue curves weight each offset by R_δ (objective), while vermillion curves weight by δR_δ (frequency gradient); each distribution is normalized by its within-band maximum. The gradient-weighted distributions move toward one across bands, indicating more even contributions across offsets.

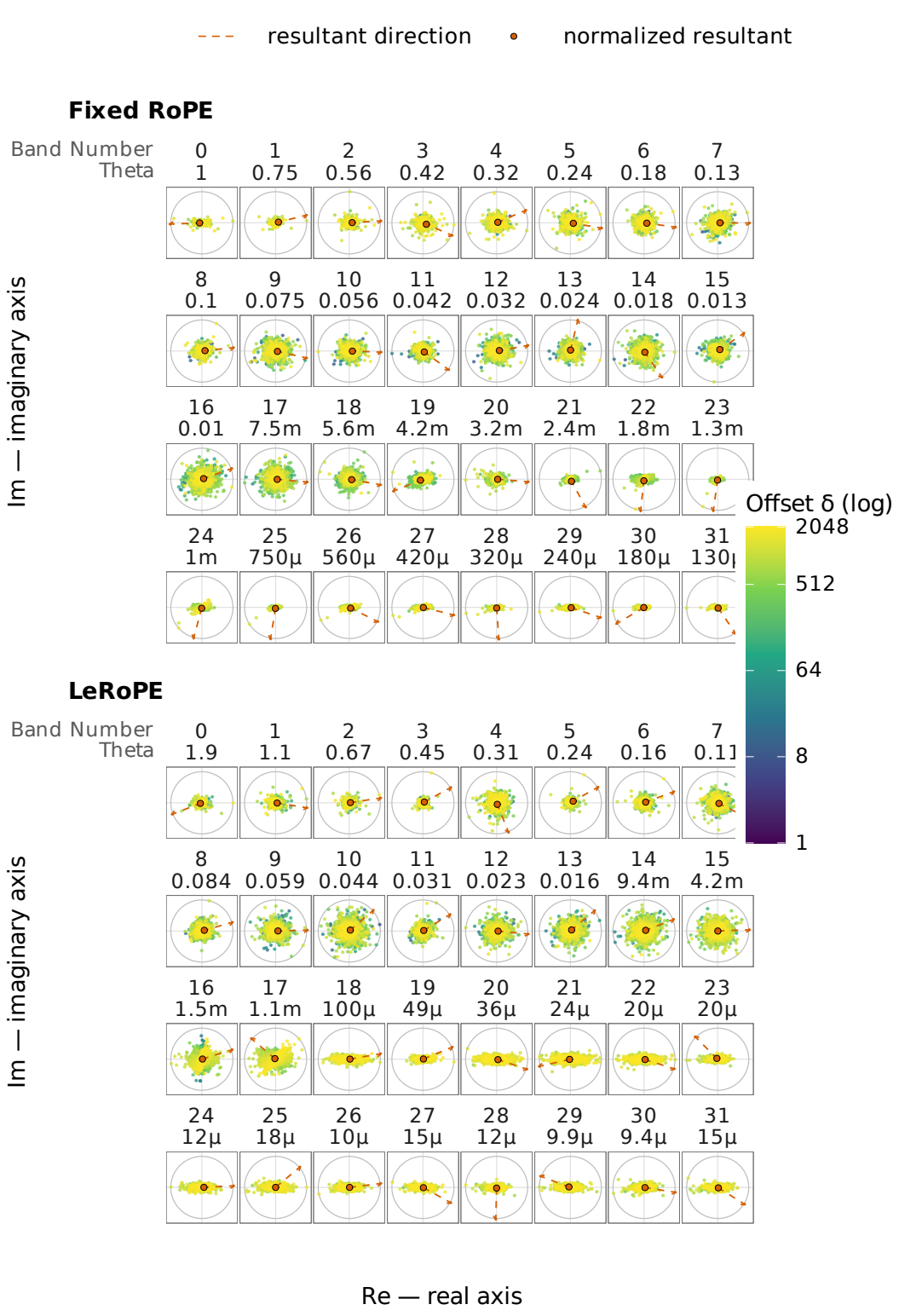

Figure 8: Gradient-weighted per-offset phasors for fixed RoPE (top) and LeRoPE (bottom) at $L = 2048$. Panel headers report the band number and operating θ, with m$= 10^{-3}$ and $\mu = 10^{-6}$. Radius is normalized within band, color encodes log offset δ, the dashed vermillion arrow gives the aggregate direction, and the outlined dot gives the normalized resultant magnitude. The panels show how cancellation and alignment vary across bands.

Table 8: Consolidated ablations at 217M on C4, all under the protocol of §4 and reported on the shared base seed for comparability; Δ is the loss difference to default RoPE (negative is better). Rows marked † have three matched seeds (Table 9). Differences below ~ 0.002 nats are within seed variability, so orderings among the three LeRoPE parametrizations, and between $p \in \{0.5, 0.75\}$, are not meaningful. Fixed LeRoPE trains with frequencies frozen to those of an independent LeRoPE run at the same scale (cf. Table 2).

Variant	Val loss (nats)	Δ vs RoPE
RoPE, $b = 2048^{\dagger}$	2.9107	$+0.0073$
RoPE, $b = 10{,}000$ (default)†	2.9035	—
RoPE, $b = 500{,}000^{\dagger}$	2.8996	-0.0039
p-RoPE, $p = 0.25$	2.9208	$+0.0173$
p-RoPE, $p = 0.5$	2.9029	-0.0006
p-RoPE, $p = 0.75$ (baseline)	2.9029	-0.0006
Shaw-style RPE	2.9068	
LeRoPE (log scalar, shared)†	**2.8978**	$\mathbf{-0.0057}$
linear scalar	2.8978	-0.0057
direct frequencies	2.8980	-0.0055
per layer and head	2.8996	-0.0039
frozen transferred freqs. (Fixed LeRoPE)	2.8999	-0.0036

D Additional Ablations

In this section we include additional ablations that we ran during the process of testing LeRoPE.

Parametrization. LeRoPE learns the per-band scale in log space, $\hat{\theta}_m = e^{\alpha_m}\theta_m$. We compare against a linear scalar, $\hat{\theta}_m = \alpha_m \theta_m$, and against directly learning each frequency, $\hat{\theta}_m = \alpha_m$. Both recover RoPE at initialization.

All three reach the same loss to within 0.0003 nats (Table 8), within seed variability, suggesting the choice is not critical with Adam's per-parameter step normalization. However, we observed very unstable trajectories for the directly learned frequencies and aliasing around 0 for the linear scalars. We retain the log parametrization because it removes these issues.

Finer Learned Granularities. LeRoPE shares one set of scales across all layers and heads. Learning a separate set per layer and head adds parameters but no benefit. It reached 2.8996 nats, 0.0018 behind shared LeRoPE which is within one seed standard deviation (Table 8).

Change of RoPE Base. Prior work finds that $b = 10{,}000$ can be suboptimal for length extrapolation, with gains reported from both raising and lowering the base (Liu et al., 2023). We therefore train 217M RoPE models with two alternative bases (Table 8): $b = L_{\text{train}} = 2048$, following FMRoPE (Oka et al., 2026b), and $b = 500{,}000$, the adjusted base frequency (ABF) adopted for long-context models by Xiong et al. (2023). Matching the base to the training length degrades in-window loss ($+0.0073$ nats), consistent with the interpolation side of the interpolation–extrapolation trade-off described by Oka et al. (2026b). Raising the base to $500{,}000$ improves loss (-0.0039 nats), recovering roughly two-thirds of LeRoPE's 0.0057-nat gain on this seed, though LeRoPE remains best.

Seed and Scale Check for $b = 500{,}000$. Because the larger base is the strongest RoPE variant we tested, we check whether LeRoPE's gains hold across seeds and at a larger scale. At 217M, we repeat all three bases and LeRoPE under two additional matched seeds (Table 9). The ordering b=2048 $>$ b=10,000 $>$ b=500,000 $>$ LeRoPE in validation loss holds under every seed. On average, $b = 2048$ is worse than the default by 0.0069 nats and $b = 500{,}000$ is better by 0.0038 nats, capturing roughly 60% of LeRoPE's gain. At 608M, we train RoPE with $b = 500{,}000$ on the base seed only (Table 9). The larger base again helps but captures only about half (53%) of LeRoPE's gain. Interpolating the $b = 500{,}000$ losses at 217M and 608M in $\log C$ as in Section 4.2 gives LeRoPE a compute multiplier relative to $b = 500{,}000$ of approximately $1.015\times$ at 217M and $1.02\times$ at 608M. The 608M estimate extrapolates beyond the largest $b = 500{,}000$ model and is conservative for the reasons given in Section A.2.

Table 9: C4 validation loss (nats) across three matched random seeds; seed 42 is the run reported in Table 1. $\bar{\Delta}$ is each baseline's mean paired margin behind LeRoPE, which is positive under every individual seed ($s_d = 0.0019$ vs. RoPE and 0.0008 vs. b=500k at 217M; 0.0010 vs. RoPE at 608M). Seed effects are strongly correlated across methods trained on the same seed, so paired margins vary substantially less than the marginal deviations suggest. Per-method seed variability decreases with scale, and at 608M the two methods' ranges do not overlap.

	Seed 42	Seed 43	Seed 44	mean $\pm$ std	$\bar{\Delta}$
217M					
RoPE (b=2048)	2.9107	2.9100	2.9137	2.9115 ± 0.0020	0.0134
RoPE (b=10k)	2.9035	2.9050	2.9053	2.9046 ± 0.0010	0.0066
RoPE (b=500k)	2.8996	2.8997	2.9031	2.9008 ± 0.0020	0.0027
LeRoPE	**2.8978**	**2.8963**	**2.9001**	$\mathbf{2.8981 \pm 0.0019}$	—
608M					
RoPE (b=10k)	2.6304	2.6290	2.6288	2.6294 ± 0.0009	0.0040
RoPE (b=500k)	2.6277	–	–	–	0.0024
LeRoPE	**2.6253**	**2.6250**	**2.6257**	$\mathbf{2.6254 \pm 0.0004}$	—

p-RoPE fraction. Our baseline uses $p = 0.75$, the best-performing value in Barbero et al. (2024). We swept $p \in \{0.25, 0.5, 0.75\}$ at 217M. $p = 0.5$ matches $p = 0.75$ (both within seed variability of RoPE). $p = 0.25$ leaves no band rotating with a period beyond $\sim$50 tokens and degrades substantially ($+0.0173$ nats, $\sim 9\times$ seed variability). Notably, LeRoPE's learned suppression point at $L_{\text{train}} = 2048$ sits near the same boundary (band $\sim$17 of 32, an effective $p \approx 0.5$), while its gain comes from reshaping the bands it retains, which no choice of p can express.

E Extrapolation and Dominant Band Wavelengths

E.1 Dominant Positional Band Wavelength

Table 10: Dominant-band indices and learned wavelengths for different models and training context lengths.

Scale (params)	L	Dominant Band	e^{α}	$\hat{\theta}^{(m)}$	$\lambda_{dominant}$ (tok)	$\hat{\theta}^{(m)}L$
217M	2048	16	0.1481	0.001481	4243.3	3.033
217M	4096	18	0.1178	0.000663	9481.7	2.714
217M	8192	19	0.0823	0.000347	18097.8	2.844
608M	2048	17	0.1945	0.001459	4307.8	2.987
1.34B	2048	17	0.1955	0.001466	4285.6	3.003
2.52B	2048	17	0.2009	0.001507	4170.3	3.086

Leave-one-out band ablation. We zero one learned frequency band at a time and measure the increase in evaluation loss relative to the baseline. Zeroing band 17 increases loss by 0.762 nats. The next-largest loss increase comes from zeroing band 26, which increases loss by only 0.069 nats.

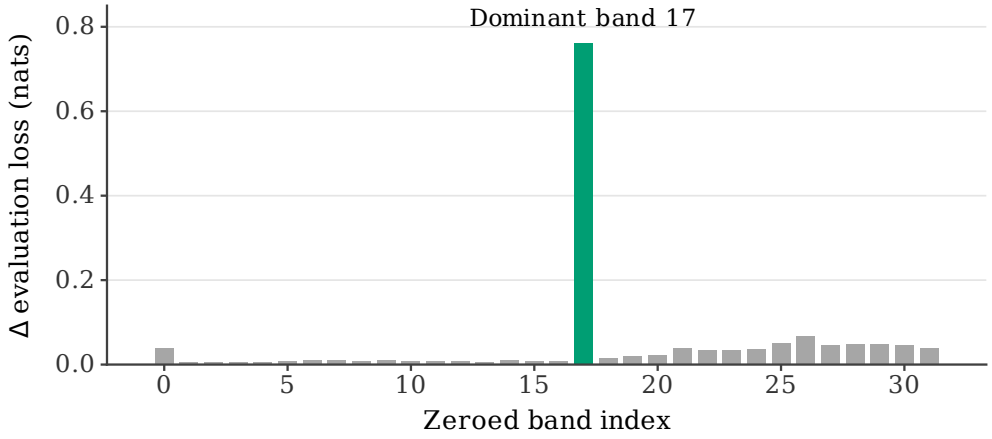

Figure 9: Leave-one-out importance of learned frequency bands. Each bar shows the loss increase from zeroing the indicated band, relative to a baseline loss of 2.309. Zeroing out Band 17 increases evaluation loss by 0.762 nats, far exceeding others.

E.2 Extrapolation Details

NTK-by-parts. NTK-by-parts (Peng et al., 2023) sets each band's interpolation from its number of rotations during training. For band m with frequency φ_m, let $r_m = L_{\text{train}}\varphi_m/2\pi$ be its rotation count during training. In LeRoPE, $\varphi_m = \hat{\theta}_m$. Let $s = \min(1, L_{\text{train}}/L_{\text{doc}})$ be the per-document position interpolation factor. We multiply each frequency by $M_m = s + w_m(1-s)$, where $w_m = \text{clip}\big((r_m - r_{\text{lo}})/(r_{\text{hi}} - r_{\text{lo}}),\, 0,\, 1\big)$, with $r_{\text{lo}} = 1$ and $r_{\text{hi}} = 32$ (Peng et al., 2023). Bands with fewer than r_{lo} rotations are scaled by s, while bands with more than r_{hi} rotations are unchanged. Between these thresholds, the multiplier increases linearly. We apply interpolation per document rather than at a fixed target length and compute rotation counts from each model's own frequencies. LeRoPE's dominant band, slowed to ~ 0.5 rotations per window, is always fully interpolated.

YaRN. YaRN (Peng et al., 2023) adds an attention temperature to NTK-by-parts to keep attention from flattening at longer contexts. With per-document extension factor $s_{\text{ext}} = \max(1, L_{\text{doc}}/L_{\text{train}})$, attention logits are multiplied by $(c \ln s_{\text{ext}} + 1)^2$ before the softmax. We sweep YaRN's temperature coefficient c for each method and use $c = 0.1$, matching Peng et al. (2023). Both mechanisms operate at inference only; no parameters are updated.

Table 11: Average perplexity for 2.52B models under NTK-by-parts + YaRN as well as LeRoPE with only interpolating the dominant band and extrapolating the rest. Split by in-distribution ($< L_{\text{train}}$) and extrapolated ($\geq L_{\text{train}}$) positions over the 0–4095 window. LeRoPE is lowest in every column.

Method	In-dist	Extrap	Full
RoPE	10.138	8.861	10.042
p-RoPE	10.133	8.838	10.035
LeRoPE *(dom. band only)*	10.112	8.965	10.026
LeRoPE	**10.106**	**8.776**	**10.005**

To complement these sequence-level averages, Figure 10 shows per-token loss over the full evaluation range and a zoomed view of the extrapolation region.

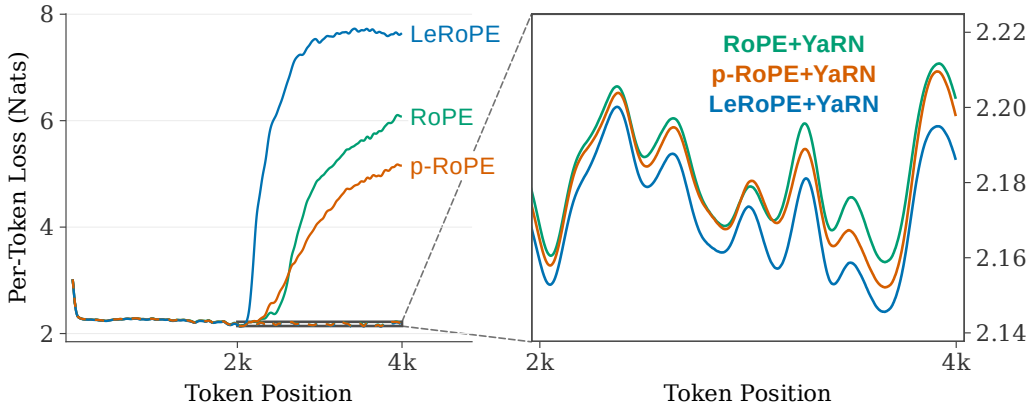

Figure 10: Per-token loss under naive extrapolation (solid) and NTK-by-parts + YaRN (dashed), evaluated up to $2 \times L_{\text{train}}$. Left: Full-range results smoothed with a Gaussian kernel, $\sigma = 20$. Right: YaRN results over the extrapolation region smoothed with $\sigma = 60$.

E.3 Complete RULER Results

The complete per-split, per-task results for the 11 RULER tasks can be found in Table 14. Each length, task pair is evaluated on 500 samples which are fixed across the three models tested. We report both the zero-shot and one-shot results for variable tracking (VT). Traditionally RULER uses one-shot example, but we found that in our case all the models mostly copied the answer from the demonstration. In the zero-shot case RoPE and p-RoPE failed to follow the answer template and would generate noise. LeRoPE here would solve the task but frequently do so by generating multiple variable names. We include both settings for completeness.

F CODE PRETRAINING

We evaluate LeRoPE on the Python split of the StarCoder data (Li et al., 2023) to test whether its gains and frequency patterns in Section 6 are specific to C4 or generalize to code. We pretrain 52M, 217M, and 608M-parameter models with the same setup as Section 4, using a 32,000-token SentencePiece tokenizer trained on code to match vocabulary size and parameter count.

Language modeling. In single-seed runs, LeRoPE reaches lower loss than RoPE at every scale (Table 12). We trained RoPE and LeRoPE at 217M under four seeds, each changing both the initialization and the data order (Table 13). LeRoPE is better under every seed and its worst run is better than RoPE's best. p-RoPE follows a similar trend to C4 with it being only slightly better than RoPE. Again, increasing RoPE's base frequency lead to a stronger performing baseline. Paired seed variability is similar to C4's, though RoPE's per-seed spread is larger (0.0041 versus 0.0010).

Table 12: Validation loss (nats) on StarCoder-Python across model scale under the training recipe of Section 4. Δ is the difference to RoPE with the default base $b = 10{,}000$ (negative is better).

Model	Method	Val loss (nats)	Δ vs RoPE
52M	RoPE	1.6166	—
	LeRoPE	**1.6095**	**−0.0070**
217M	RoPE	1.2333	—
	RoPE, $b = 500{,}000$	1.2281	−0.0052
	LeRoPE	**1.2259**	**−0.0074**
608M	RoPE	1.0531	—
	LeRoPE	**1.0475**	**−0.0056**

Learned frequencies. We analyze 217M-parameter code models trained at $L_{\text{train}} \in \{2048, 4096, 8192\}$ following the procedure in Section 6. On StarCoder, fast bands remain close to their fixed-RoPE frequencies while slow bands learn wavelengths substantially longer than their fixed-RoPE values, as on C4 (Figure 3c). As training length increases, learned frequencies deviate from fixed RoPE at lower frequencies. The dominant wavelength also increases, from $5{,}250$ tokens at $L_{\text{train}} = 2048$ to $16{,}627$ tokens at $L_{\text{train}} = 8192$. At $L_{\text{train}} = 2048$, a single positional band dominates (Figure 11b) with wavelength $\lambda \approx 5{,}250$ tokens. This is larger than those we observed for C4 and believe its partial due to StarCoder having longer average documents than C4 ($\sim$820 vs $\sim$400 tokens).

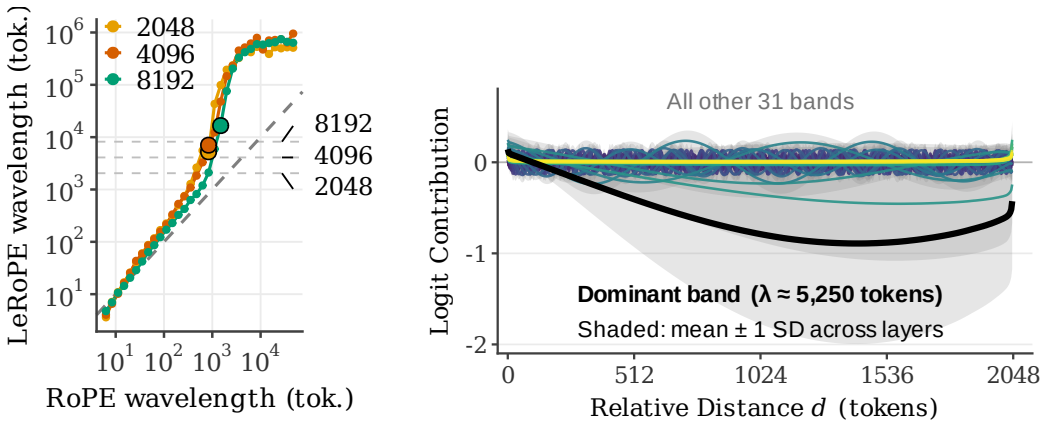

Figure 11: **Left**: Learned versus fixed-RoPE wavelength for each band in 217M LeRoPE models trained on StarCoder-Python at $L_{\text{train}} \in \{2048, 4096, 8192\}$. The diagonal dashed line marks $y = x$, horizontal dashed lines mark the training lengths, and outlined points mark the dominant bands. **Right**: Band-level logit contributions as a function of raw token offset $d = s - t$ for the $L_{\text{train}} = 2048$ model. Shading shows ± 1 standard deviation across layers.

Table 13: Validation loss StarCoder-Python at 217M across four runs. $\bar{\Delta}$ is RoPE's mean paired margin behind LeRoPE over all four runs, which is positive in every run ($s_d = 0.0022$). p-RoPE was trained in run 1 only.

	Run 1	Run 2	Run 3	Run 4	mean ± std (runs 1–3)	$\bar{\Delta}$
RoPE	1.2501	1.2421	1.2471	1.2333	1.2464±0.0041	0.0073
p-RoPE	1.2495	—	—	—	—	—
LeRoPE	**1.2398**	**1.2368**	**1.2407**	**1.2259**	**1.2391±0.0020**	—

Table 14: RULER accuracy per task for the 2.52B models (trained with a 2,048-token context). Bold marks the best of the three models per task and length. Beyond 2,048 tokens every model uses YaRN with scale factor $L/2048$ and temperature coefficient $c = 0.1$.

Task	Model	2K	4K	8K	16K	32K
S-NIAH-1	RoPE	0.984	0.976	0.884	0.814	0.776
	p-RoPE	0.998	0.942	0.682	0.510	0.668
	LeRoPE	**1.000**	**1.000**	**0.998**	**0.996**	**0.988**
S-NIAH-2	RoPE	**1.000**	0.962	0.690	0.364	0.130
	p-RoPE	0.998	0.992	0.852	0.410	0.104
	LeRoPE	**1.000**	**1.000**	**0.994**	**0.782**	**0.188**
S-NIAH-3	RoPE	0.948	0.850	0.578	0.296	0.082
	p-RoPE	**0.994**	0.958	0.808	0.466	0.116
	LeRoPE	0.992	**0.990**	**0.964**	**0.680**	**0.150**
MK-NIAH-1	RoPE	0.562	0.434	0.268	0.192	0.110
	p-RoPE	0.582	0.448	0.320	0.202	0.114
	LeRoPE	**0.684**	**0.646**	**0.456**	**0.250**	**0.194**
MK-NIAH-2	RoPE	0.054	0.018	0.002	0.002	0.000
	p-RoPE	0.062	0.034	0.012	0.004	**0.002**
	LeRoPE	**0.188**	**0.094**	**0.036**	**0.006**	0.000
MK-NIAH-3	RoPE	0.118	0.032	0.006	**0.004**	0.000
	p-RoPE	0.112	0.028	0.008	0.000	0.000
	LeRoPE	**0.126**	**0.060**	**0.022**	0.002	0.000
MV-NIAH	RoPE	0.190	0.175	0.190	0.152	0.062
	p-RoPE	0.281	0.264	0.223	0.153	0.074
	LeRoPE	**0.354**	**0.318**	**0.253**	**0.188**	**0.079**
MQ-NIAH	RoPE	0.217	0.211	0.187	0.133	0.059
	p-RoPE	0.358	0.316	0.278	0.163	0.092
	LeRoPE	**0.521**	**0.478**	**0.367**	**0.220**	**0.113**
Retrieval mean (8)	RoPE	0.509	0.457	0.351	0.245	0.152
	p-RoPE	0.548	0.498	0.398	0.238	0.146
	LeRoPE	**0.608**	**0.573**	**0.511**	**0.390**	**0.214**
VT, one-shot (RULER)	RoPE	0.044	0.244	0.166	0.139	0.185
	p-RoPE	0.041	0.225	0.231	0.134	0.069
	LeRoPE	0.049	0.069	0.043	0.326	0.170
CWE	RoPE	0.005	0.001	0.000	0.001	0.000
	p-RoPE	0.016	0.009	**0.008**	0.002	0.002
	LeRoPE	**0.074**	**0.019**	0.002	**0.004**	**0.003**
FWE	RoPE	**0.114**	**0.139**	**0.097**	**0.120**	**0.109**
	p-RoPE	0.031	0.022	0.015	0.015	0.010
	LeRoPE	0.045	0.046	0.044	0.049	0.065
All 11, one-shot VT	RoPE	0.385	0.368	0.279	0.202	0.138
	p-RoPE	0.407	0.385	0.312	0.187	0.114
	LeRoPE	**0.458**	**0.429**	**0.380**	**0.318**	**0.177**
VT, zero-shot (ablation)	RoPE	0.014	0.017	0.008	0.015	0.016
	p-RoPE	0.019	0.027	0.028	0.021	0.013
	LeRoPE	0.426	0.455	0.443	0.444	0.415
All 11, zero-shot VT	RoPE	0.382	0.347	0.264	0.190	0.122
	p-RoPE	0.405	0.367	0.294	0.177	0.109
	LeRoPE	**0.492**	**0.464**	**0.416**	**0.329**	**0.200**

VT rows are not bolded because neither protocol measures variable tracking faithfully.