BEATS AND ALIASES CAN BREAK ROTARY LENGTH EXTRAPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary position encodings degrade beyond their training length, and a common account blames planes whose period exceeds the training window. We show that a spectrum whose every period is below 13 tokens can still collapse: what is missing is not another frequency but a difference of frequencies. For fixed content a rotary logit is a trigonometric polynomial of the offset, and in that fixed-kernel problem classical super-resolution bounds say that gaps above about $1/L$ let the training offsets determine it stably, while a gap ε with $\varepsilon L \leq 1/2$ can raise the worst-case error near the half-beat offset by a factor $1/(\varepsilon L)$. In a 0.46M-parameter character-level model, twelve pairs detuned by $1/2048$ cycles per token make training on 512 tokens collapse after position 1,024; across a sweep of eight beat periods the onset tracks the beat, and wherever a model has failed, stretching only the gaps at inference removes 52–86% of the damage while an equal shift that keeps them removes at most 9%. Both recur at 3.3M parameters, after ten times longer training and on a second corpus, although in the first two the collapse begins earlier than the beat predicts. In spectra of harmonics of coprime periods, 59% of the strongest out-of-window attention weights fall where the Chinese remainder theorem predicts (base rate 4%), and masking them helps at least three times more than random masking. For RoPE's own geometric spectrum gaps and periods rank the planes alike, our experiments do not separate them, and a beat-specific prediction fails. Spectral arithmetic is an additional diagnostic for rotary extrapolation, computable before training.

1 INTRODUCTION

Rotary position encoding (RoPE) turns positions into rotations of query and key planes (Su et al., 2024) and is the default in open language models (Touvron et al., 2023). Models trained on length L degrade quickly on longer inputs, and a large literature repairs this after training (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024; Tian et al., 2026) or changes the frequencies (Liu et al., 2024; Wang et al., 2024; Barbero et al., 2025; Hua et al., 2025). A common explanation concerns single frequencies: a plane whose period exceeds L never completes a cycle in training, so its angles at longer offsets are out of distribution (Liu et al., 2024; Hua et al., 2025), and several methods respond by clipping or rescaling them. This raises a sharp question: *can a rotary model fail beyond its window when every one of its planes completes dozens of cycles inside it?*

It can. Figure 1 shows two character-level models sharing a spectrum whose every period is below 13 tokens, so the period account sees no plane that could fail. Trained on 64 tokens, the spectrum is one of the two best extrapolators we train at that length; trained on 512 tokens, it tracks a well-behaved control up to position 1,024 and then collapses, while its attention floods back onto tokens about 1,500 positions away. What predicts this is a *difference* of frequencies.

Differences govern the failure for a reason that is classical in signal processing. For fixed content a rotary logit is a trigonometric polynomial of the offset, so how stably L training offsets pin down its continuation depends on the gaps between the frequencies present, not on the frequencies themselves (Theorems 1 and 2, adapted from super-resolution). Two frequencies ε apart beat into a slow clock of period $1/\varepsilon$ that a window of length L sees only along an arc (Proposition 1): the spectrum of Figure 1 has twelve such pairs at $\varepsilon = 1/2048$, and its 512-token model fails right after half of that beat period. Commensurate spectra fail by a second mechanism: harmonics of a few coprime

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.

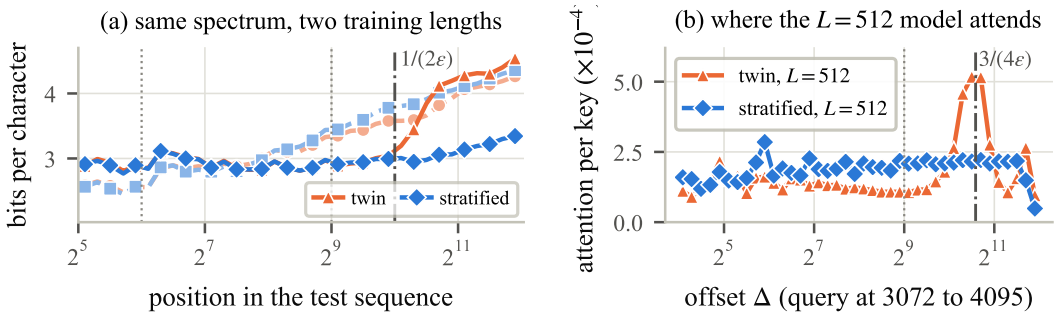


Figure 1: **Every period is short and the model still breaks.** (a) Bits per character by position for a spectrum of twelve frequency pairs detuned by $\varepsilon = 1/2048$ cycles per token ("twin"; every period below 13 tokens) and a stratified spectrum without close pairs. At $L = 64$ (pale, dashed) the beat is unobservable and both degrade alike; at $L = 512$ it is partially resolved and the twin model fails right after half its beat period (dotted: training lengths; dash-dotted: $1/(2\varepsilon)$). (b) Attention per key for queries at 3072–4095, mean over layers and heads. One seed throughout; Figure 2 and Table 1 give the seed spread.

periods make the logit depend on the offset only through residues, so the model re-attends to far offsets that agree with its trained peaks modulo products of those periods (Proposition 2).

Scope. Our theory is about a fixed kernel observed on the training offsets, not about what gradient descent learns; it identifies a hazard, not a guaranteed failure of a trained model, and its conditions are sufficient, not necessary. Our claim is therefore not that periods are the wrong object but that they are an incomplete one: in the spectra we study, gaps can decide where a model fails even when every period is short, and commensurabilities predict where it will attend. For RoPE's own geometric spectrum the two accounts coincide: they rank the planes alike (Table 9), clipping by beats buys nothing over clipping by periods, and a prediction we derived from the beat account for RoPE failed (Section 5.3). All the evidence is character-level models of 0.46M and 3.3M parameters, which use little context beyond a few dozen positions, so what we demonstrate is that the hazard exists and can be intervened on, not that it dominates extrapolation in pretrained models.

Contributions.

- **A finding: short periods do not imply extrapolation.** A spectrum of twelve pairs detuned by $1/2048$ cycles per token, every period below 13 tokens, is one of the best extrapolators at $L = 64$ and no better than RoPE at $L = 512$ (Figure 1). Across a sweep of eight beat periods with 5 seeds each, the position where loss starts to rise moves with the beat and the damage is non-monotone in it (Figure 2); wherever the model fails, stretching only the gaps at inference removes 52–86% of it while an equal shift that keeps them removes at most 9% (Table 1). The collapse and its repair recur in a 3.3M-parameter model, after 40M characters of training and on a second corpus, though in the first two the collapse begins earlier than the beat predicts (Appendix E).
- **A fixed-kernel theory.** Adapting the large sieve and a two-point minimax argument to rotary kernels shows that gaps above about $1/L$ let the window determine the logit stably at every offset, while a gap ε with $\varepsilon L \leq 1/2$ can raise the worst-case error near offset $1/(2\varepsilon)$ by a factor $1/(\varepsilon L)$ over the resolved-pair scale; both hold numerically on fixed kernels (Theorems 1 and 2, Figure 4). Adding one hypothesis about what a trained head reads turns this into predictions for where far attention peaks and where the loss starts to rise, and the Chinese remainder theorem lists the offsets at which commensurate spectra alias (Propositions 1 and 2).
- **Mechanism evidence, and where the account stops.** Heads of residue spectra put 35–100% of their 20 largest out-of-window attention weights on predicted aliases (base rate 4%), and masking those offsets lowers the far loss 3.1–5.4× more than masking random ones, so that attention causally contributes to the excess far loss (Section 5.2); rounding RoPE's wavelengths to integers, as Resonance RoPE does (Wang et al., 2024), creates the same structure. For standard RoPE, however, the two accounts rank the planes alike, our experiments do not separate them, and a beat-specific prediction fails, so what the account buys there is a hazard count and not a better encoding (Section 5.3, Table 9).

2 RELATED WORK

Why rotary encodings fail beyond the training length. Su et al. (2024) motivated RoPE by a long-term decay of the logit bound, and Chen et al. (2023) showed that its attention scores can take very large values just outside the trained offsets. Liu et al. (2024) explained extrapolation through a *critical dimension*: planes whose period exceeds the training length never complete a cycle, so their angles at longer offsets are out of distribution. Xu et al. (2024) bounded the base needed for long-context retrieval; Barbero et al. (2025) and Urrutia et al. (2026) found that high frequencies carry positional patterns while the lowest act as semantic channels, the division of labour our content planes assume; Oka et al. (2026) traced a high-norm band that sets an interpolation–extrapolation trade-off; Wu et al. (2026) argued that models use the frequencies matching the dependencies in the data; Liu (2026) derived Nyquist-type aliasing of single frequencies; Du et al. (2026) proved that RoPE attention loses locality and relevance consistency as the context grows; FoPE (Hua et al., 2025) treats each dimension as a Fourier series and zeroes the undertrained low frequencies; and Wertheimer et al. (2026) trace the failure to a collapse of key and query clusters, repairing it with high-frequency RoPE on a subset of channels. All of this reasons about single frequencies, frequency bands, random phases or cross-frequency mixing in later layers; to our knowledge none uses the separation between frequencies to predict where a spectrum with only short periods fails.

Changing the frequencies. Resonance RoPE rounds wavelengths to integers (Wang et al., 2024), FoPE clips the frequencies that never complete a cycle (Hua et al., 2025), p-RoPE drops the lowest (Barbero et al., 2025), CoPE clips them softly (Li et al., 2026), HoPE (Chen et al., 2025) and RoPE-ID (Wertheimer et al., 2026) keep only high frequencies, and smaller or length-matched bases are used by Liu et al. (2024) and Oka et al. (2026). All select planes by their individual periods; ours selects them by their gaps, the same ordering for the RoPE family but a different threshold, since clipping by period leaves the gap count nonzero while keeping only the highest frequencies empties it (Table 9). Integer wavelengths also make the kernel repeat, which by Proposition 2(ii) aliases trained peaks at predictable offsets (Section 5.3).

Extending the context after training. Position interpolation (Chen et al., 2023), NTK-aware scaling and YaRN (Peng et al., 2024), LongRoPE (Ding et al., 2024), positional skip-wise training (Zhu et al., 2024), dimension-wise index manipulation (Lu et al., 2025) and the mixed-radix view of these rescalings (Tian et al., 2026) change the frequencies or positions of a trained model; they complement our analysis of what the training window can pin down, and the training-free ones are our baselines. Randomized positional encodings (Ruoss et al., 2023) train on subsampled positions, while ALiBi (Press et al., 2022), KERPLE (Chi et al., 2022), FIRE (Li et al., 2024b) and no positional encoding (Kazemnejad et al., 2023) avoid rotation altogether.

Residue codes, super-resolution and the Talbot effect. Residue number systems represent a coordinate by phases with coprime periods; grid cells use them to encode position with exponential range (Fiete et al., 2008; Sreenivasan & Fiete, 2011), as does high-dimensional computing (Kymn et al., 2025), while grid-cell-inspired positional encodings use several spatial scales without the coprime structure (Li et al., 2024a). Fractional Talbot images (Berry & Klein, 1996; Fernández-Pousa, 2017) give such codes in closed form. Super-resolution studies the conditioning of Fourier matrices with close nodes (Candès & Fernandez-Granda, 2014; Moitra, 2015; Batenkov et al., 2020); Theorems 1 and 2 adapt it to rotary attention.

3 ROTARY KERNELS, BEATS AND ALIASES

Rotary kernels. A rotary head with $2M$ query/key dimensions rotates its j-th plane by $\omega_j n$ at position n (Su et al., 2024). We call $\Omega = (\omega_1, \dots, \omega_M) \in [0, \pi]^M$ the *spectrum* of the head and write $f_j = \omega_j / 2\pi$ in cycles per token; a plane with $\omega_j = 0$ carries content only. With the j-th query and key planes as complex numbers $\hat{q}_j, \hat{k}_j$, the logit between a query at position m and a key at n is

$$p(\Delta) := \sum_{j=1}^{M} \mathrm{Re}\big(c_j \, e^{i\omega_j \Delta}\big), \qquad c_j = \hat{q}_j \overline{\hat{k}_j}, \qquad \Delta = m - n. \tag{1}$$

For a fixed pair of token representations, equation 1 makes p a real trigonometric polynomial of the offset, the *kernel* of the pair. With $e(x) = e^{2\pi i x}$ it reads $p(\Delta) = \sum_{\xi\in\Xi} a_\xi e(\xi\Delta)$ over the *node set* $\Xi(\Omega)$ on the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, which contains $\pm f_j$ for every rotating plane and 0 if some plane carries content, with coefficients $a_{\pm f_j} = c_j/2,\ \bar{c}_j/2$ (merged when nodes coincide). A model trained on length L only ever evaluates kernels at the offsets $W_L = \{0, \dots, L-1\}$.

Beats. Two distinct nodes ξ, ξ' *beat* with period $1/\|\xi - \xi'\|$, where $\|x\|$ is the distance from x to the nearest integer. The smallest gap $\eta(\Omega) = \min_{\xi\neq\xi'} \|\xi - \xi'\|$ defines the *slowest beat* $B(\Omega) = 1/\eta(\Omega)$. A beat of gap ε is *resolved* by the window if its period is below $L-1$, the span of the training offsets, *partially resolved* if the window sees between an eighth and all of its cycle ($1/8 \le \varepsilon L \lesssim 1$), and *unobservable* below an eighth (the fractions are conventions). Two kinds of beat involve a single plane: with its own mirror ($\pm f_j$, period $1/\|2f_j\|$) and, when some planes carry content, with the content node (period $1/f_j$); the second is unresolved exactly when the period reaches $L-1$, which up to one token is the critical dimension of Liu et al. (2024). All other beats are between different planes and do not appear in a list of periods.

Spectra studied. *RoPE*: $\omega_j = b^{-(j-1)/M}$ with base b, whose neighbour ratio $r = b^{-1/M}$ makes the low end dense. *Harmonic residue spectra*: for pairwise coprime moduli $q_1 < \cdots < q_K$ the planes rotate by $2\pi a/q_k$, $1 \le a \le \lfloor q_k/2 \rfloor$, and the rest carry content (Appendix C). *Stratified spectra*: one frequency per equal-width cell of a band, jittered by at most a quarter cell. *Twin spectra*: stratified carriers f, each duplicated at $f + \varepsilon$, so ε sets the slowest beat $B = 1/\varepsilon$. Table 4 lists every spectrum with its parameters.

4 Theory: what the training window pins down

Scope. The results below concern one fixed kernel observed on the training offsets: when those L values determine it elsewhere stably, and when a close pair of nodes makes that determination ill-conditioned. They do not model gradient descent, do not bound a language model's loss, and give sufficient, not necessary, conditions; they yield spectral predictions that a trained model is free to violate, and Section 5 tests them directly. The large sieve and the two-point minimax argument are classical (Montgomery & Vaughan, 1974; Moitra, 2015); new here are the rotary kernel as the object and the criterion that follows. Let $R = |\Xi(\Omega)|$ count distinct nodes, let $\|p\|_W$ denote the Euclidean norm of the values of p on the window, and call the Euclidean norm of $(a_\xi)_{\xi\in\Xi}$ the coefficient norm of p. All proofs are in Appendix B; Table 3 collects the notation.

Theorem 1 (Resolved beats: the window determines every offset stably)**.** *If $L - 1 > B(\Omega)$, then for any two kernels p, p' of the spectrum Ω*

$$\sup_{\Delta\in\mathbb{Z}} |p(\Delta) - p'(\Delta)| \ \le\ \sqrt{\frac{R}{L - 1 - B(\Omega)}}\ \|p - p'\|_{W_L}.$$

In particular least squares on noisy window values, $y_\Delta = p(\Delta) + \sigma z_\Delta$ with independent standard Gaussian z_Δ, recovers p at every offset with $\mathbb{E}\sup_\Delta |\hat{p}(\Delta) - p(\Delta)| \le \sigma R/\sqrt{L - 1 - B}$.

In words, once every beat completes a cycle inside the window, two heads that nearly agree on the training offsets nearly agree at all offsets, whatever the test length. A determined kernel need not be benign: it does not decay, so attention dilutes as keys accumulate (Section 5.3). For the stratified spectrum of Section 5 at $L = 512$ ($R = 49$, $B = 92$) the constant is 0.34; for RoPE with $b = 10^4$, $B = 141{,}272$ exceeds every training length, so Theorem 1 never applies to it.

Theorem 2 (Unresolved beats: extrapolation is ill-conditioned)**.** *Suppose two nodes of $\Xi(\Omega)$ are $\varepsilon \le 1/4$ apart and let $\Delta_\star$ be the integer nearest to $1/(2\varepsilon)$. There is a kernel $p_\star$ of Ω with coefficient norm at most $\sqrt{2}$, window energy $\|p_\star\|_{W_L}^2 \le \frac{4\pi^2}{3}\varepsilon^2 L^3$ and $|p_\star(\Delta_\star)| \ge \sqrt{2}$. Consequently, for window observations with noise level σ and every $\Gamma > 0$, there are two kernels p_0, p_1 with coefficient norm at most Γ such that every estimator satisfies*

$$\max_{i\in\{0,1\}} \mathbb{E}_i\big|\hat{p}(\Delta_\star) - p_i(\Delta_\star)\big| \ \ge\ \min\Big\{\frac{\Gamma}{4},\ \frac{\sqrt{6}}{8\pi}\,\frac{\sigma}{\varepsilon L^{3/2}}\Big\}.$$

The bound is worst case over kernels of coefficient norm at most Γ: not every kernel with a close pair is badly determined, only some are. In rates, Theorem 1 bounds the least-squares error on a resolved spectrum by $O(\sigma R/\sqrt{L})$ at every offset, while a single gap ε raises the worst-case error at $\Delta_\star$, which lies beyond the window when $\varepsilon L \le 1/2$, to order $\sigma/(\varepsilon L^{3/2})$: a factor $1/(\varepsilon L)$ above the $\sigma/\sqrt{L}$ scale of a resolved pair, when the coefficient cap is not binding (a cluster of close nodes inflates the error polynomially in the inverse separation, Batenkov et al., 2020).

Both statements hold numerically on fixed kernels, with no network involved: as εL falls below the resolution threshold, the smallest singular value of the window matrix falls in proportion to it, and the least-squares error at $\Delta_\star$ follows the $\sigma/(\varepsilon L^{3/2})$ rate within a factor 1.2–1.4 for $\varepsilon L \le 1/4$ (Figure 4, Appendix B). What a trained model does with a beat it can only partly see is a separate question; to ask it, write the two planes of a beat as a carrier at f with a slowly turning envelope.

Proposition 1 (A partially resolved beat is a wrapping clock)**.** *The joint contribution of nodes f and $f+\varepsilon$ with coefficients c, c' is $\mathrm{Re}\big[e(f\Delta)\,A(\Delta)\big]$ with envelope $A(\Delta) = c + c'e(\varepsilon\Delta)$, and $|A(\Delta)|^2 = |c|^2 + |c'|^2 + 2|cc'|\cos(2\pi\varepsilon\Delta + \phi)$, $\phi = \arg(c'\bar{c})$. Over the training offsets W_L the clock phase $2\pi\varepsilon\Delta + \phi$ sweeps an arc of length $2\pi\varepsilon(L-1)$, and for every integer u*

$$|A(\Delta)| = |A(u/\varepsilon - \Delta - \phi/(\pi\varepsilon))|.$$

Consequently, if $\varepsilon(L-1) < 1$ the clock phase at offsets beyond $L-1$ lies outside the arc swept on W_L, except within $L-1$ after a multiple of $1/\varepsilon$, and offsets near $u/\varepsilon - \phi/(\pi\varepsilon)$ carry the same envelope magnitude as offsets near 0.

From the identity to a hypothesis. Proposition 1 is a statement about the kernel, not about learning. To turn it into a prediction we add one hypothesis that Section 5 tests rather than assumes: *a trained head reads the beat through the envelope magnitude $|A|$*, and *what hurts is an envelope larger than any seen in training*, since larger logits attract attention. Where on the cycle the window sits is set by the learned phase ϕ, which we measure (Appendix D): in our twin models it concentrates near $+\pi/2$ (Section 5.1), so the envelope decreases across the window, returns to its value at $\Delta = 0$ at offset $1/(2\varepsilon)$ and overshoots every trained value on $(1/(2\varepsilon), 1/\varepsilon)$, most at $3/(4\varepsilon)$. If $\varepsilon L \ll 1$ the envelope barely moves on the window and cannot be used; if $\varepsilon L \ge 1$ the whole cycle is observed. Under the hypothesis, damage should be non-monotone in the beat period, failure should begin in the overshoot set $(u/\varepsilon - 1/(2\varepsilon),\, u/\varepsilon]$, whose first interval beyond the window is $(\max\{L, 1/(2\varepsilon)\}, 1/\varepsilon]$, and far attention should peak near $3/(4\varepsilon)$. The hypothesis is what makes these falsifiable: for ϕ near $-\pi/2$ overshoot and undershoot exchange places, and a head that reads the phase of A instead leaves the trained arc right after L, so the onset would predict nothing.

Proposition 2 (Commensurate spectra alias at Chinese-remainder offsets)**.** *For a harmonic residue spectrum with pairwise coprime moduli $q_1 < \cdots < q_K$: (i) $\eta = 1/(q_{K-1}q_K)$, so the spectrum is resolved iff $L - 1 > q_{K-1}q_K$; (ii) every kernel has the additive form $p(\Delta) = p_0 + \sum_k h_k(\Delta \bmod q_k)$ and is periodic with period $H = \prod_k q_k$, which is its least period whenever the fundamental harmonic of every modulus has a nonzero coefficient in that kernel; (iii) for any offset D and set of moduli S, the offsets that agree with D modulo every $q_k \in S$ are $D + \mathbb{Z}\prod_{k\in S} q_k$, and on them $p(\Delta) - p(D)$ depends only on the moduli outside S. In particular two distinct offsets of W_L are S-aliases of each other only if $\prod_{k\in S} q_k \le L - 1$: for a peak $D \in W_L$ the window therefore contains no second offset of its S-alias lattice once the product exceeds $L-1$, while offsets outside the window do lie on it.*

By (ii), if a model learns sharp peaks at a few small offsets $\mathcal{D}$, any offset whose residue modulo q_k lies in $\mathcal{D} \bmod q_k$ for most k inherits most of the peak terms; the window only teaches it to suppress those with modulus products below L, and the others appear at test time, at offsets that can be listed in advance. Coprimality only makes the products least common multiples: for arbitrary integer periods, (ii) and (iii) hold with $H = \mathrm{lcm}_k q_k$ and the S-alias lattice $D + \mathbb{Z}\,\mathrm{lcm}_{k\in S} q_k$, and coprime moduli are the case in which these lattices are sparsest.

Consequence for RoPE. Adjacent RoPE frequencies differ by the factor $r = b^{-1/M}$, so the gap between planes j and $j+1$ is $(1-r)f_j$ and their beat has period $P_j/(1-r)$, where $P_j = 1/f_j$. If only the planes with $P_j \le (1-r)L$ rotate and the others carry content, every beat, including the neighbour beats and the beat of the lowest rotating plane with the content channel, has period at most L, about the span of the window (Appendix B); this *beat rule* is sufficient, not minimal.

The period rule keeps the planes with $P_j \leq L$, whose neighbour beats stay partially resolved for $P_j \in ((1-r)L, L]$; for $b = 10^4$ and $M = 32$, $1 - r = 0.25$, so the beat rule rotates only the planes with period at most $L/4$. Because every RoPE gap is proportional to its frequency, the two rules rank the planes alike and differ only by the factor $1 - r$ in the threshold; beats and periods come apart for spectra that are not geometric. Plain RoPE has no content plane, so its only single-plane beat is the mirror beat, of period $P_j/2$, and the critical dimension of Liu et al. (2024) is the content beat of a spectrum that does have them, such as the clipped models we train.

5 EXPERIMENTS

Setup. We train decoder-only character-level language models on enwik8 (Mahoney, 2011) and, for replication, on 150 PG-19 books (Rae et al., 2020): a base model with 2 layers, width 128 and 2 heads of 32 rotation planes (0.46M parameters) and a scaled model with 4 layers, width 256 and 4 heads (3.3M parameters). Variants differ *only* in the rotation frequencies; architecture, data order, optimiser and budget (4M characters, about 490 optimiser steps at $L = 512$; 12M for the scaled model) are identical (Appendix D). We train with $L \in \{64, 128, 256, 512\}$ and report bits per character (bpc) at every position of validation windows up to $8L$. The *damage* of a model on a range R beyond the window is $\mathrm{bpc}(R) - \mathrm{bpc}([0, L))$, the cost of leaving the window rather than the quality of the model; unless stated otherwise $R = [4L, 8L)$. The spectra (Table 4) are RoPE ($b = 10^4$); *stratified*, 24 frequencies in $[1/17, 1/2]$ cycles per token plus 8 content planes (slowest beat 92); *twins*, 12 stratified carriers each duplicated at $+\varepsilon$; and residue spectra with moduli $(4, 5, 7, 9, 11)$, $(5, 7, 11, 13, 17)$ and $(11, 13, 17, 19)$. We ask where beats (Q1) and aliases (Q2) make models fail, and what follows for RoPE (Q3). Table 5 collects every prediction with its test and outcome, separating consequences of a formal result from hypotheses about the trained model, and including the two that failed.

5.1 Q1: BEATS SET THE FAILURE POINT

Twin spectra isolate beats: across the sweep they differ only in the gap ε, which shortens each upper member's period $1/(f + \varepsilon)$ by at most 3.2% relative to its carrier (0.6% at $\varepsilon = 1/2048$) and leaves every period below 12.6 tokens, so no period criterion separates them; the carriers-only control drops the upper members. Trained on 64 tokens, the twin model with $\varepsilon = 1/2048$ has an unobservable beat ($\varepsilon L = 1/32$) and extrapolates as well as the stratified spectrum (Figure 1a). Trained on 512 tokens the same beat is partially resolved ($\varepsilon L = 1/4$): the model tracks the stratified spectrum up to position $1024 = 1/(2\varepsilon)$ and then degrades to 4.12 bpc on $[2048, 4096)$, no better than RoPE (4.07) and 0.82 bpc above the stratified spectrum (3.30). Its attention shows the wrapping clock: for queries at positions 3072–4095 the attention per key is flat on offsets 256–768 and about five times larger on offsets 1280–1792, around the predicted $3/(4\varepsilon) = 1536$, while the stratified model's profile is flat over the same range (Figure 1b). The learned phase $\phi = \arg(c'\bar{c})$ satisfies $\pi/4 < |\phi| < 3\pi/4$ for 90–98% of layer, head and pair combinations (circular mean 1.51–1.57 rad, 5 seeds).

Figure 2 sweeps the beat period from 384 to 32,768 tokens at $L = 512$. The damage is non-monotone, as predicted: resolved beats ($1/\varepsilon < L - 1$) and unobservable ones ($\varepsilon L < 1/8$) stay at or below the beat-free control, while beats between $2L$ and $4L$ add up to 0.9 bpc on $[4L, 8L)$; a barely visible beat is hardly used (removing the 8,192 beat changes the in-window loss by 0.001 bpc). Take the onset to be the first position at which the excess over the control stays above 0.1 bpc for three consecutive 64-token bins. The envelope overshoots its trained range on $(u/\varepsilon - 1/(2\varepsilon),\ u/\varepsilon]$ for every $u \geq 1$, a set periodic in the beat; the onset of the seed-mean curves lies in one of these intervals for all six partially resolved beats, and in the first one beyond the window for five (for all six at a threshold of 0.05 bpc; Appendix E). The exception is the fastest, $1/\varepsilon = 768$, whose window already covers two thirds of the cycle, so its rise is detectable only one cycle later, at 1,280; for it and for 1,024 the first interval begins at L itself, so they test the prediction less sharply than the other four.

Changing only the gaps repairs the failure. To test causally that the beat, rather than any period, produces the failure, we intervene on the trained twin models at inference (Table 1). When evaluating positions up to sL we move the upper member of every pair from $f + \varepsilon$ to $f + \varepsilon/s$ (*stretch*), with $s = 2, 4, 8$ matched to the three ranges beyond the window as for the training-free baselines of Section 5.3: the carriers are untouched, every period changes by at most 1.5% for the partially

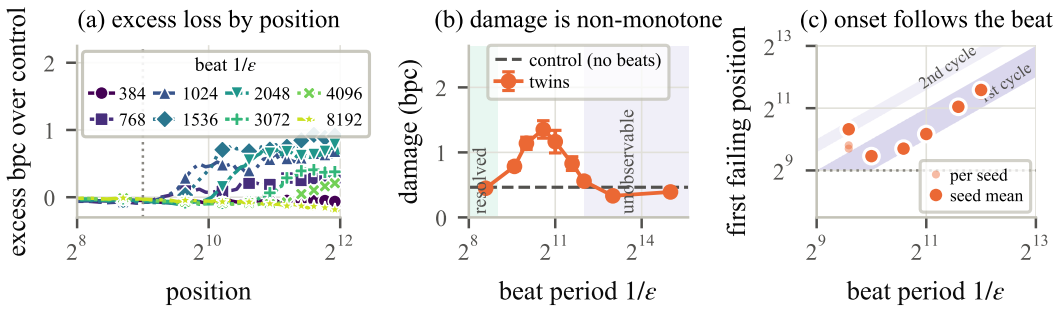


Figure 2: **The failure point follows the beat** (twin spectra, $L = 512$; 5 seeds per beat period, one for the slowest beat, which appears only in (b)). (a) Excess bpc over the carriers-only control by position, seed means, one curve per beat period. (b) Damage on $[4L, 8L)$ against the beat period (error bars: standard deviation over seeds); resolved beats (green) and unobservable ones (violet) are harmless. (c) Onset (first position whose excess exceeds 0.1 bpc for three 64-token bins) for the seed mean (large) and each seed (small), against the intervals where Proposition 1 puts the envelope above its trained range (first and second beat cycle).

Table 1: **Changing only the gaps repairs the failure** (twin models, $L = 512$, interventions at inference, mean over n seeds). Bits per character for the unchanged model (none), the control that shifts both members of every pair by $\varepsilon(1 - 1/s)$ (shift), and the stretch of the upper member from $f + \varepsilon$ to $f + \varepsilon/s$, with $s = 2, 4, 8$ on the three ranges and windows of length sL as in Table 6; $[0, L)$ is the same $s = 1$ evaluation throughout. The carriers-only control (5 seeds) scores 2.99, 3.10, 3.22 and 3.45.

			$[L, 2L)$			$[2L, 4L)$			$[4L, 8L)$		
beat $1/\varepsilon$	n	$[0, L)$	none	shift	stretch	none	shift	stretch	none	shift	stretch
384	5	2.94	3.06	3.09	3.07	3.16	3.17	3.16	3.39	3.42	3.37
768	5	2.92	3.19	3.20	3.02	3.41	3.50	3.08	3.71	3.70	3.29
1,024	5	2.93	3.35	3.33	3.02	3.74	3.77	3.08	4.07	4.08	3.29
1,536	5	2.94	3.20	3.19	3.04	3.99	3.94	3.10	4.30	4.22	3.31
2,048	5	2.95	3.06	3.06	3.05	3.73	3.67	3.11	4.12	4.09	3.33
3,072	5	2.96	3.06	3.06	3.07	3.19	3.18	3.14	3.79	3.72	3.36
4,096	5	2.97	3.07	3.07	3.08	3.12	3.12	3.15	3.53	3.52	3.37
8,192	5	2.98	3.09	3.09	3.10	3.14	3.13	3.17	3.30	3.30	3.38

resolved beats, and the beat phase stays on the arc seen in training. A control moves both members down by $\varepsilon(1 - 1/s)$, a frequency change of the same size that keeps the gaps (*shift*). In every range where a model has failed, that is where its damage exceeds by more than 0.1 bpc that of the beat-free control (itself a mean of 5 seeds), the stretch removes 52–86% of the damage and leaves less damage than that control (0.11, 0.23 and 0.46 bpc, the differences of the four control scores in the caption of Table 1), while the shift removes at most 9%. The range is conditional on those selected ranges: where nothing has failed the stretch is a small net harm, as moving a frequency the model uses correctly should be. In the 3.3M-parameter model (Table 12) and after 40M characters of training (Table 13) the twins fail earlier than the beat predicts, from about 800 and 700, though the stretch still removes 50–88% of the damage there; and on PG-19, in the two failed ranges, it removes 81–94% against at most 16% for the shift (3 seeds, Table 11).

5.2 Q2: Residue spectra alias where the arithmetic says

For the residue spectrum with moduli $(5, 7, 11, 13, 17)$ trained on 512 tokens, the three largest in-window attention weights of every head sit at offsets 0–3, so by Proposition 2(ii) out-of-window attention should land on offsets whose residues modulo most moduli are residues of these peaks. Among the 20 largest out-of-window weights of each head (offsets 512–2047), 35–100% (mean 59% over 12 heads) do so for at least four of the five moduli, against 4% of all offsets and 0–15% for the stratified, twin and RoPE models. The enrichment survives every choice of the statistic's parameters, vanishes for unused moduli (Table 14), and passes a randomization test that randomises

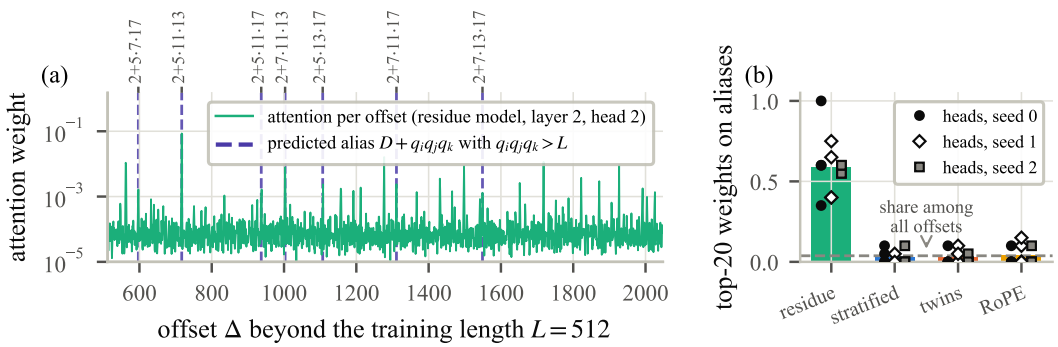


Figure 3: **Aliases at Chinese-remainder offsets** ($L = 512$). (a) Attention weight per offset beyond the training length for the head of the residue model $(5, 7, 11, 13, 17)$ with the sharpest in-window peak ($D = 2$); dashed lines are aliases $D+q_iq_jq_k$ whose modulus product exceeds L, so the window never contained them. (b) Share of each head's 20 largest out-of-window weights on predicted aliases (agreement modulo at least four of the five moduli with the three largest in-window peaks); dots are heads, the dashed line the share among all offsets.

Table 2: **Train short, test long** on enwik8 ($L = 512$): bpc in the window and on $[4L, 8L)$, without and with YaRN's attention temperature; mean $\pm$ standard deviation over n seeds, B the slowest beat. Table 6 gives all ranges and the other baselines.

				$[4L, 8L)$	
spectrum / method	B	n	$[0, L)$	plain	+ temperature
RoPE, $b = 10^4$	141,272	3	2.91 ± 0.01	4.07 ± 0.07	3.91 ± 0.07
+ YaRN	n/a	3	2.90 ± 0.01	3.13 ± 0.06	–
+ period rule ($P \leq L$)	1,413	3	2.93 ± 0.01	3.40 ± 0.02	3.14 ± 0.06
with integer wavelengths	1,409	3	2.93 ± 0.01	3.49 ± 0.07	–
twins, $\varepsilon = 1/2048$	2,048	5	2.95 ± 0.02	4.12 ± 0.16	4.16 ± 0.10
stratified, 24 planes	92	3	2.93 ± 0.01	3.30 ± 0.03	3.09 ± 0.02

the arithmetic while holding each head's attention fixed, at $p \leq 0.0025$ for every residue head (Appendix E). The largest weight, 0.086 at offset $717 = 2 + 5 \cdot 11 \cdot 13$, and the aliases 597 and 1003 have modulus products above L, so the window never contained them (Figure 3). Masking turns the readout into an intervention: forbidding attention to the 397 predicted offsets beyond L lowers the residue models' loss on $[4L, 8L)$ by 0.11–0.22 bpc, against 0.04 for as many random offsets (mean of 5 masks; 3.1–5.4 times less in each of 3 seeds) and 0.07–0.09 for as many of the most-attended other offsets. This last, attention-matched control is the informative one: attention to the predicted set causally contributes to the excess far loss beyond what attending to equally prominent far offsets does. For the stratified model the order reverses (0.03 against 0.08–0.11): aliases are the harmful far offsets only where the arithmetic predicts them.

5.3 Q3: What the account does and does not explain for standard RoPE

In Table 2, RoPE loses 1.16 bpc from the window to $[4L, 8L)$; of the training-free rescalings, YaRN recovers most of it, mainly through its frequency ramp (65% of the damage, against 14% for its temperature alone), while position interpolation without fine-tuning is destructive (Table 7).

Standard geometric RoPE is the main negative control for our account: because its gaps scale with its frequencies, beats and periods rank the planes alike (Section 4) and our experiments do not separate them. Clipping by periods, which here coincides with p-RoPE at $p = 1/2$ (Barbero et al., 2025) and with FoPE's clipping step (Hua et al., 2025), removes more than half of the damage. Clipping by beats also removes the five planes whose neighbour beats are partially resolved; it costs 0.03 bpc in the window for about the same loss beyond it (3.44 against 3.40), and stretching at inference only the planes whose period exceeds L repairs more than also stretching those five, a prediction of the beat account that failed (Appendix E). This is the clearest boundary of the account: on a geometric

spectrum it adds nothing to the period criterion, and the one prediction it makes on its own is wrong. Table 9 shows why, and where the two counts do part company: for RoPE and its rescalings the two orderings coincide, which for a geometric spectrum is an identity rather than a measurement, but clipping by period leaves 28 plane pairs of a Llama-2-sized spectrum beating into $(L - 1, 8L]$ at a period count of zero, and only clipping at $L/8$ empties both.

Integer wavelengths, the premise of Resonance RoPE (Wang et al., 2024), separate the two accounts in the other direction: rounding the wavelengths of period-clipped RoPE moves no period by over half a token, so no plane changes status under the period criterion, while the alias account predicts new structure, its lattices now generated by least common multiples rather than products (the remark after Proposition 2). As predicted before training, the rounded model's heads put 20–25% of their 20 largest out-of-window weights on predicted aliases (1–4% without rounding; base rate 4.2%), masking them helps 2.7, 2.8 and 7.0 times more than random masks in the three seeds, and the loss on $[4L, 8L)$ rises on average from 3.40 to 3.49 (per seed 3.39–3.56, up in two seeds of three and unchanged in the third). The alias causality is weaker here than for the residue spectra: for both rounded and unrounded models the attention-matched mask helps more than the alias mask (Appendix E).

Resolving every beat buys nothing on its own: the residue spectrum, resolved but commensurate, and beat-clipped RoPE are no better than period-clipped RoPE (3.43 and 3.44 against 3.40). The two stratified spectra, resolved and evenly spread, extrapolate best, 0.08–0.10 bpc below clipped RoPE on $[4L, 8L)$, a margin comparable to their seed spread (up to 0.12); but a random spectrum in the same band does as well and both keep every period below $L/4$, so this does not isolate the beats. The attention temperature lowers the far loss of every spectrum except the twins by a similar 0.16–0.26 bpc (Chiang & Cholak, 2022; Nakanishi, 2025); what remains ranks the kernels. With it, the stratified spectra stay within 0.15–0.21 bpc of their in-window loss up to $8L$, on a par with RoPE extended by YaRN (0.23), which needs no retraining, and in the 3.3M-parameter model YaRN is the better of the two (0.23 against 0.31; Appendix E). Designing a spectrum by its gaps therefore buys diagnosis rather than a better encoding.

6 Discussion and conclusion

What changes. When some planes carry content, the critical dimension (Liu et al., 2024) is the beat of a plane with the content channel; clipping by period (Hua et al., 2025; Barbero et al., 2025; Li et al., 2026) removes those beats but keeps the partially resolved neighbour beats of the remaining low planes (Table 9), which at our scale matter little for RoPE, and integer wavelengths (Wang et al., 2024) make the kernel repeat, which in our models also creates aliases that draw attention.

Limitations. Our models are small character-level transformers trained briefly (about 490 optimiser steps): their loss barely improves beyond the first few dozen positions, so they use little long-range context, and ten times longer training does not change this. The damage we measure is therefore the cost far tokens impose on local prediction, not whether far information is still used: a model that ignores distant keys scores well on it, as ALiBi and NoPE do (Table 6). We replicate on a second corpus and a larger model, not on pretrained models, and measure bits per character, not retrieval; differences below about 0.05 bpc lie within our seed spread, and every scale check is one seed. Two predictions failed: for RoPE the beat-specific stretching prediction is wrong, and the onset window, which holds for the briefly trained base models, does not survive scale, the collapse beginning near 800 and 700 in the 3.3M-parameter and 40M-character models. The resolution threshold of Theorem 1 leaves no trace in a sweep over L, and integer wavelengths were tested only in the rounding step. Finally, the theory bounds what any learner can infer from noisy kernel values, not what gradient descent does, and the onset interval assumes a head reads the envelope magnitude, that only an overshoot hurts, and the measured phase near $+\pi/2$.

Conclusion. Whether a rotary encoding extrapolates depends on the arithmetic of its whole spectrum and not only on the periods of its planes: whether every beat is resolved in the training window, and whether the spectrum is commensurate, can be read off the frequency list before training, and in the spectra we study it says where a model can fail and where it will attend. For RoPE the two accounts coincide, so what this buys is a diagnostic rather than a better encoding: it counts a hazard that clipping by period leaves behind, without yet showing what that hazard costs.

AI USE STATEMENT

In this paper, AI helps to aid or polish writing, review the draft, act as a coding agent, take part in discussion and generate plots. AI also helps to create slides for internal meetings; gave feedback on the experimental design, including the choice of the shift control and the attention-matched masking control; implemented the training, evaluation and analysis code; and took part in interpreting the results. We also used it for tasks: drafting and editing the text for readability, producing the figures, searching for and summarising related literature, formatting the references, suggesting the structure of the paper, and preparing slides for internal meetings. It was not used to generate data: all datasets are public. Every table and figure, and most numbers in the text, are produced from saved runs by our own scripts; the remaining numbers were checked against those runs by hand, and the compute total is an estimate. We did not use it for translation, for cleaning or reformatting datasets, or for qualitative or thematic analysis, none of which arise in this work.

All AI-assisted work was reviewed by the authors. The code was inspected and its outputs cross-checked against independent recomputations, which is how the tie handling in one correlation and a stale figure of merit were caught; every reference was checked against its publisher or arXiv record; and the claims were audited against the tables sentence by sentence. The authors take full responsibility for the content of this paper, including the text, the claims, the code and the figures produced with the aid of generative AI.

ETHICS STATEMENT

This work studies positional encodings of small character-level language models trained on public text corpora (enwik8 and public-domain books from PG-19). It involves no human subjects, personal data or deployed systems, and we do not foresee direct negative societal impact beyond that of language modelling in general.

REPRODUCIBILITY STATEMENT

Section 3 defines every spectrum and Table 4 lists their parameters; Appendix D gives the data processing, architectures, optimiser settings, budgets, seeds and evaluation protocol, including the attention readouts and the alias statistic. The main theoretical results state their assumptions in the main text; all proofs are in Appendix B, and the numerical check of the theory in Figure 4 is deterministic given the seed stated in its caption. Every run uses a single 2 GB laptop GPU (4M characters train in about 2.5 minutes for the base model), and the full set of experiments takes about 21 GPU-hours. The training, evaluation and plotting code, the spectrum definitions and the saved per-position losses behind every table and figure accompany this submission as supplementary material; we will also release the dated design log that records the predictions we made before running the corresponding experiments.

REFERENCES

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11091–11104, 2024.

Qingyuan Tian, Wenhong Zhu, Xiaoran Liu, Xiaofeng Wang, and Rui Wang. MrRoPE: Mixed-radix rotary position embedding. In *International Conference on Learning Representations*, 2026.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of RoPE-based extrapolation. In *International Conference on Learning Representations*, 2024.

Suyuchen Wang, Ivan Kobyzev, Peng Lu, Mehdi Rezagholizadeh, and Bang Liu. Resonance RoPE: Improving context length generalization of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 586–598, 2024.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! What makes rotary positional encodings useful? In *International Conference on Learning Representations*, 2025.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24932–24949, 2025.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of RoPE bounds context length. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, and Cristobal Rojas. Decoupling positional and symbolic attention in transformers. In *International Conference on Learning Representations*, 2026.

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 *International Conference on Learning Representations*, 2026.

Xinyi Wu, Siyuan Liu, and Ali Jadbabaie. How data shapes RoPE frequency usage: From positional scale matching to length generalization. *arXiv preprint arXiv:2607.07678*, 2026.

Feilong Liu. Rotary positional embeddings as phase modulation: Theoretical bounds on the RoPE base for long-context transformers. *arXiv preprint arXiv:2602.10959*, 2026.

Yufeng Du, Phillip Harris, Minyang Tian, Eliu A. Huerta, Srikanth Ronanki, Subendhu Rongali, Aram Galstyan, and Hao Peng. RoPE distinguishes neither positions nor tokens in long contexts, provably. *arXiv preprint arXiv:2605.15514*, 2026.

Davis Wertheimer, Aozhong Zhang, Derrick Liu, Penghang Yin, and Naigang Wang. Frayed RoPE and long inputs: A geometric perspective. In *International Conference on Learning Representations*, 2026.

Haoran Li, Sucheng Ren, Alan Yuille, and Feng Wang. CoPE: Clipped RoPE as a scalable free lunch for long context LLMs. *arXiv preprint arXiv:2602.05258*, 2026.

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. HoPE: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23044–23056, 2025.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *International Conference on Learning Representations*, 2024.

Yi Lu, Wanxu Zhao, Xin Zhou, Chenxin An, Chenglong Wang, Shuo Li, Yuming Yang, Jun Zhao, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Effective length extrapolation via dimension-wise positional embeddings manipulation. In *Second Conference on Language Modeling (COLM)*, 2025.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1889–1903, 2023.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Ta-Chung Chi, Ting-Han Fan, Peter J. Ramadge, and Alexander I. Rudnicky. KERPLE: Kernelized relative positional embedding for length extrapolation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *International Conference on Learning Representations*, 2024b.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ila R. Fiete, Yoram Burak, and Ted Brookings. What grid cells convey about rat location. *Journal of Neuroscience*, 28(27):6858–6871, 2008.

Sameet Sreenivasan and Ila Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature Neuroscience*, 14(10):1330–1337, 2011.

Christopher J. Kymn, Denis Kleyko, E. Paxon Frady, Connor Bybee, Pentti Kanerva, Friedrich T. Sommer, and Bruno A. Olshausen. Computing with residue numbers in high-dimensional representation. *Neural Computation*, 37(1):1–37, 2025.

Boyang Li, Yulin Wu, Nuoxian Huang, and Wenjia Zhang. GridPE: A grid cell-inspired unified position embedding for arbitrary-dimensional spaces. *arXiv preprint arXiv:2406.07049*, 2024a.

M. V. Berry and S. Klein. Integer, fractional and fractal Talbot effects. *Journal of Modern Optics*, 43(10):2139–2164, 1996.

Carlos R. Fernández-Pousa. On the structure of quadratic Gauss sums in the Talbot effect. *Journal of the Optical Society of America A*, 34(5):732–742, 2017.

Emmanuel J. Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

Ankur Moitra. Super-resolution, extremal functions and the condition number of Vandermonde matrices. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 821–830, 2015. Section numbering follows the full version, arXiv:1408.1681.

Dmitry Batenkov, Laurent Demanet, Gil Goldman, and Yosef Yomdin. Conditioning of partial nonuniform Fourier matrices with clustered nodes. *SIAM Journal on Matrix Analysis and Applications*, 41(1):199–220, 2020.

H. L. Montgomery and R. C. Vaughan. Hilbert's inequality. *Journal of the London Mathematical Society*, s2-8(1):73–82, 1974.

Matt Mahoney. Large text compression benchmark. `https://mattmahoney.net/dc/textdata.html`, 2011. About the test data, last updated 1 September 2011.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664, 2022.

Ken M. Nakanishi. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

A NOTATION

Table 3: Notation.

symbol	meaning
L, L'	training sequence length; test sequence length
W_L	training offsets $\{0, \dots, L-1\}$
m, n; $\Delta = m - n$	query and key positions; offset
M	number of rotation planes per head (head dimension $2M$)
ω_j, $f_j = \omega_j/2\pi$, $P_j = 1/f_j$	angular frequency, frequency (cycles per token) and period (tokens) of plane j
$\Omega = (\omega_1, \dots, \omega_M)$	spectrum of a head; $\omega_j = 0$ is a content plane
$\hat{q}_j, \hat{k}_j$; $c_j = \hat{q}_j \overline{\hat{k}_j}$	plane j of a query and a key as complex numbers; their product
$p(\Delta)$	logit kernel of a query–key pair, $\sum_j \mathrm{Re}(c_j e^{i\omega_j \Delta})$
$e(x)$; $\|x\|$	$e^{2\pi i x}$; distance from x to the nearest integer
$\Xi(\Omega)$, $R = \lvert\Xi\rvert$, a_ξ	node set ($\pm f_j$ of the rotating planes, and 0 if a plane carries content), its size, kernel coefficient at node ξ
$\eta(\Omega)$, $B(\Omega) = 1/\eta$	smallest gap between nodes; slowest beat (tokens)
ε	a small gap between two nodes; its beat period is $1/\varepsilon$
$A(\Delta)$, ϕ	envelope of a beat and its phase (Proposition 1)
$q_1 < \cdots < q_K$, $H = \prod_k q_k$	moduli of a residue spectrum; its exact period
$h_k(\cdot)$, p_0	per-modulus term and content term of a residue kernel (Proposition 2)
$g_q[t]$, T_q	quadratic Gauss (Talbot) sequence on $\mathbb{Z}_q$; cyclic shift on $\mathbb{Z}_q$
D, $\mathcal{D}$, S	an offset (a peak); the set of in-window peak offsets; a set of moduli
s	scale of a training-free extension or of the beat stretch ($s = 2, 4, 8$ on $[L, 2L)$, $[2L, 4L)$, $[4L, 8L)$)
b, $r = b^{-1/M}$	RoPE base; ratio of neighbouring RoPE frequencies
$\mathbb{T} = \mathbb{R}/\mathbb{Z}$	circle on which the nodes live
σ, z_Δ, Γ	noise level of window observations; independent standard Gaussian noise; bound on the coefficient norm, the Euclidean norm of (a_ξ) (Theorem 2)
$\Delta_\star$, $p_\star$	integer nearest to $1/(2\varepsilon)$; a kernel that is small on the window and large at $\Delta_\star$ (Theorem 2)
$g(\varepsilon)$, λ	window energy of the difference mode of a close pair; ridge parameter (Appendix B)
V_L	Vandermonde matrix $(e(\xi\Delta))_{\Delta \in W_L,\, \xi \in \Xi}$
κ, o	number of in-window peaks and agreement order of the alias statistic (Appendix D)
damage	loss on a range beyond the window minus the in-window loss (bits per character)

B PROOFS

Throughout, $e(x) = e^{2\pi i x}$, $\|x\|$ is the distance from $x \in \mathbb{R}$ to the nearest integer, a kernel of the spectrum Ω is $p(\Delta) = \sum_{\xi \in \Xi} a_\xi e(\xi\Delta)$ with Hermitian coefficients ($a_{-\xi} = \bar{a}_\xi$), $R = |\Xi|$, and $V_L \in \mathbb{C}^{L \times R}$ is the Vandermonde matrix $(V_L)_{\Delta,\xi} = e(\xi\Delta)$, $\Delta = 0, \dots, L-1$, so that the window values of p are $V_L a$.

Lemma 1 (Large sieve in Gram form)**.** *If the nodes are pairwise at circular distance at least η and $L - 1 > 1/\eta$, then for all $a \in \mathbb{C}^R$*

$$(L - 1 - 1/\eta)\,\|a\|_2^2 \;\le\; \|V_L a\|_2^2 \;\le\; (L - 1 + 1/\eta)\,\|a\|_2^2.$$

This is the Montgomery–Vaughan form of the large sieve (Montgomery & Vaughan, 1974); Moitra (2015, Section 2.3) derives both inequalities from the Beurling–Selberg majorant and minorant of the indicator of an interval containing L integers, and translation of the window only multiplies the columns of V_L by unimodular constants. That derivation assumes strict separation, $\|\xi - \xi'\| > \eta$; the non-strict form used here follows by applying it with $\eta' < \eta$ and letting $\eta' \uparrow \eta$, both sides being

continuous in η. For a self-contained argument with a weaker constant: the Gram matrix $G = V_L^* V_L$ has diagonal entries L and off-diagonal entries $|G_{\xi\xi'}| = |\sin(\pi L(\xi - \xi'))/\sin(\pi(\xi - \xi'))| \le 1/(2\|\xi - \xi'\|)$, because $\sin(\pi t) \ge 2t$ on $[0, \frac{1}{2}]$. On each side of ξ the k-th nearest node is at distance at least $k\eta$, so every row has off-diagonal mass at most $2\sum_{k=1}^{\lceil R/2 \rceil}(2k\eta)^{-1} \le (1 + \ln R)/\eta$, and Gershgorin's theorem gives $\lambda_{\min}(G) \ge L - (1 + \ln R)/\eta$. Theorem 1 is stated with the sharp constant, but nothing qualitative depends on it: for the stratified spectrum of Section 5 ($R = 49$, $B = 92$, $L = 512$) the Gershgorin form gives 0.89 in place of 0.34, and both are finite exactly when the slowest beat is resolved.

Proof of Theorem 1. Let a and a' be the coefficient vectors of p and p'. For every $\Delta \in \mathbb{Z}$, $|p(\Delta) - p'(\Delta)| = |\sum_\xi (a_\xi - a'_\xi) e(\xi\Delta)| \le \|a - a'\|_1 \le \sqrt{R}\,\|a - a'\|_2$. By Lemma 1 with $\eta = 1/B$, $\|a - a'\|_2^2 \le \|V_L(a - a')\|_2^2/(L - 1 - B) = \|p - p'\|_{W_L}^2/(L - 1 - B)$. For the estimation statement, least squares $\hat{a} = V_L^+ y$ has $\hat{a} - a = \sigma V_L^+ z$ and $\mathbb{E}\|\hat{a} - a\|_2^2 = \sigma^2 \operatorname{tr}\big((V_L^* V_L)^{-1}\big) \le \sigma^2 R/(L - 1 - B)$; the first part and Jensen's inequality give $\mathbb{E}\sup_\Delta |\hat{p}(\Delta) - p(\Delta)| \le \sqrt{R}\,(\mathbb{E}\|\hat{a} - a\|_2^2)^{1/2} \le \sigma R/\sqrt{L - 1 - B}$. □

Proof of Theorem 2. Let the two nodes be ξ_1, ξ_2 with $\|\xi_1 - \xi_2\| = \varepsilon$ and choose $\theta \in \mathbb{R}$ such that $e(\theta)\big(e(\xi_1\Delta_\star) - e(\xi_2\Delta_\star)\big)$ is a nonnegative real number. Set $p_\star(\Delta) = \operatorname{Re}\big[e(\theta)\big(e(\xi_1\Delta) - e(\xi_2\Delta)\big)\big]$. It is a kernel of Ω (coefficients $\pm e(\pm\theta)/2$ on $\pm\xi_1, \pm\xi_2$), and its coefficient norm is 1 when the four nodes are distinct and at most $\sqrt{2}$ when $\xi_2 = -\xi_1$ or one of them is the content node, in which case coefficients merge. Since $|e(\xi_1\Delta) - e(\xi_2\Delta)| = 2|\sin(\pi\varepsilon\Delta)| \le 2\pi\varepsilon\Delta$, $\|p_\star\|_{W_L}^2 \le 4\pi^2\varepsilon^2 \sum_{\Delta<L} \Delta^2 \le \frac{4\pi^2}{3}\varepsilon^2 L^3$. At $\Delta_\star$, $\pi\varepsilon\Delta_\star \in [\frac{\pi}{2} - \frac{\pi\varepsilon}{2}, \frac{\pi}{2} + \frac{\pi\varepsilon}{2}]$, so $p_\star(\Delta_\star) = 2|\sin(\pi\varepsilon\Delta_\star)| \ge 2\cos(\pi/8) > \sqrt{2}$. For the lower bound, compare $p_0 = 0$ and $p_1 = \alpha p_\star$ with $\alpha = \min\{\Gamma/\sqrt{2},\ \sqrt{3}\,\sigma/(2\pi\varepsilon L^{3/2})\}$, observed as $y_\Delta = p_i(\Delta) + \sigma z_\Delta$ on the window with independent standard Gaussian z_Δ. The Kullback–Leibler divergence between the two observation laws is $\|\alpha p_\star\|_{W_L}^2/(2\sigma^2) \le \frac{1}{2}$, so by Pinsker's inequality their total variation is at most $\frac{1}{2}$. Le Cam's two-point method (Tsybakov, 2009, Section 2.4.2) gives, for any estimator T, $\max_i \mathbb{E}_i|T - p_i(\Delta_\star)| \ge \frac{1}{2}|p_1(\Delta_\star) - p_0(\Delta_\star)|(1 - \frac{1}{2}) \ge \frac{\sqrt{2}}{4}\alpha$, which is the stated bound. □

Proof of Proposition 1. $\operatorname{Re}[c\,e(f\Delta) + c' e((f + \varepsilon)\Delta)] = \operatorname{Re}[e(f\Delta)(c + c' e(\varepsilon\Delta))]$. Expanding $|c + c' e(\varepsilon\Delta)|^2$ gives the cosine formula. The function $\Delta \mapsto \cos(2\pi\varepsilon\Delta + \phi)$ is invariant under $\Delta \mapsto \Delta + u/\varepsilon$ for integer u and under the reflection $\Delta \mapsto -\Delta - \phi/(\pi\varepsilon)$, which proves the identity. On $\Delta \in [0, L)$ the phase $2\pi\varepsilon\Delta + \phi$ covers an arc of length $2\pi\varepsilon L < 2\pi$, and any $\Delta \ge L$ with $\varepsilon\Delta \notin \mathbb{Z} + [0, \varepsilon L)$ has a phase outside it. □

A ridge picture of which beats are used. Consider the two nodes ξ_1, ξ_2 of a pair at distance ε, and suppose for simplicity that the other nodes are resolved and nearly orthogonal to them on the window. The Gram matrix of the pair on the window has eigenvalues $L \pm |\rho|$ with $\rho = \sum_{\Delta<L} e(\varepsilon\Delta)$, so the difference mode has window energy

$$g(\varepsilon) = L - \Big|\frac{\sin(\pi\varepsilon L)}{\sin(\pi\varepsilon)}\Big|, \qquad g(\varepsilon) \approx \tfrac{\pi^2}{6}\,\varepsilon^2 L^3 \ \ (\varepsilon L \ll 1), \qquad g(\varepsilon) \ge L/2 \ \ (\varepsilon L \ge 1),$$

the last because $|\rho| \le 1/\sin(\pi\varepsilon) \le 1/(2\varepsilon)$. With $\lambda = 0$ below (least squares), the coefficient of this mode has standard deviation $\sigma/\sqrt{g}$, and the unit-norm mode has magnitude up to $\sqrt{2}$ at offsets near $L/2 + 1/(2\varepsilon)$, where the two nodes are in antiphase, so the error there is of order $\sqrt{2}\,\sigma/\sqrt{g} \approx (2\sqrt{3}/\pi)\,\sigma/(\varepsilon L^{3/2})$ for $\varepsilon L \ll 1$: least squares attains the rate of Theorem 2 up to a constant. Ridge regression with parameter λ on window values $y_\Delta = p(\Delta) + \sigma z_\Delta$ estimates the coefficient a_- of this mode as $\hat{a}_- = \frac{g}{g+\lambda} a_- + \frac{\sqrt{g}}{g+\lambda}\sigma\zeta$ with a standard Gaussian ζ, so that

$$\mathbb{E}\,|\hat{a}_- - a_-|^2 = \Big(\frac{\lambda}{g + \lambda}\Big)^2 |a_-|^2 + \frac{g}{(g + \lambda)^2}\,\sigma^2.$$

The difference mode has magnitude of order $|a_-|$ at offsets near $L/2 + 1/(2\varepsilon)$, where the two nodes are in antiphase, so this is also the order of the squared extrapolation error there. The variance term vanishes for $g \ll \lambda$ (an unobservable beat is shrunk to zero and not used), peaks at $g = \lambda$

with value $\sigma^2/(4\lambda)$, and decays like σ^2/g once the beat is resolved; the bias term only records that an unobservable beat cannot be used, which costs nothing relative to a spectrum without it. The variance term therefore makes the intermediate gaps, with $\frac{\pi^2}{6}\varepsilon^2 L^3 \approx \lambda$, the hazardous ones, as in the non-monotone damage of Figure 2b. This is a heuristic for trained networks, whose implicit regularisation is not a ridge penalty, not a theorem about them.

Proof of Proposition 2. (i) The nonzero nodes are the fractions a/q_k with $a \not\equiv 0 \pmod{q_k}$, where a ranges over a set of representatives in $(-q_k/2, q_k/2]$, so that the harmonics and their mirrors are both of this form and coincident ones are counted once; for even q_k the harmonic $a = q_k/2$ is its own mirror and contributes the single node $1/2$. Two such fractions with the same modulus differ by at least $1/q_k$ in circular distance unless they are the same node. For $q_i \neq q_j$, $\|a/q_i - b/q_j\| = \|(aq_j - bq_i)/(q_i q_j)\| \geq 1/(q_i q_j)$, since $aq_j - bq_i \not\equiv 0 \pmod{q_i q_j}$ when $q_i \nmid a$ and $\gcd(q_i, q_j) = 1$. Nonzero nodes are at distance at least $1/q_K$ from 0. Hence $\eta \geq 1/(q_{K-1}q_K)$, and Bezout's identity $aq_K - bq_{K-1} = 1$ with $1 \leq a < q_{K-1}$ gives equality. Theorem 1 then yields the resolution statement. (ii) Grouping the planes by modulus, the contribution of modulus q_k is a function of $\Delta \bmod q_k$; since each q_k divides $H = \prod_k q_k$, $p(\Delta + H) = p(\Delta)$. For the least period, note that the nodes of Ξ are distinct, so the coefficients (a_ξ) of a kernel are unique and an integer T is a period of p if and only if $e(\xi T) = 1$ for every node ξ with $a_\xi \neq 0$. If the fundamental harmonic $1/q_k$ of every modulus has a nonzero coefficient, then T must be a multiple of every q_k, hence of H since the moduli are coprime; and H is a period, so it is the least one. (Without that condition a kernel can have a shorter period, for instance when only the second harmonics appear.) (iii) By the Chinese remainder theorem, $\Delta \equiv D \pmod{q_k}$ for all $k \in S$ iff $\Delta \equiv D \pmod{\prod_{k\in S} q_k}$; on these offsets the terms h_k with $k \in S$ coincide with their values at D. Any two distinct offsets of that lattice differ by at least $\prod_{k\in S} q_k$, and two elements of W_L differ by at most $L-1$, so W_L contains two of them only if $\prod_{k\in S} q_k \leq L - 1$. (The restriction to W_L matters: for a reference offset D outside the window, an alias $D - \prod_{k\in S} q_k$ can fall inside it even when the product exceeds L.) □

Proof of the RoPE consequence. With $\omega_j = b^{-(j-1)/M}$ the nonnegative nodes are $f_1 > f_2 > \cdots > f_M$ with $f_{j+1} = r f_j$. The gap between neighbours $j, j+1$ is $(1-r)f_j$, the gap of the lowest node to its mirror is $2f_M$ and to a content node f_M. If only planes with period $P_j = 1/f_j \leq (1-r)L$ rotate, every neighbour gap is at least $(1-r)/P_j \geq 1/L$, and the lowest rotating node is at distance $f_{\min} \geq 1/((1-r)L) > 1/L$ from 0. Non-neighbouring rotating nodes are farther apart than neighbours, a node f and a mirror node $-g$ are $f + g \geq 2f_{\min}$ apart, and the largest RoPE frequency $1/(2\pi) < 1/4$ keeps every gap away from wrap-around; hence every beat has period at most L (in our experiments $B = 335$ for $L = 512$). Conversely a rotating plane with $P_j > (1-r)L$ and a rotating neighbour $j+1$ gives a gap below $1/L$. □

Numerical check of the theorems. Figure 4 evaluates both statements on fixed kernels, with no network involved. A close pair $\{f, f+\varepsilon\}$ sits inside a spectrum of ten well-separated carriers; we sweep εL from $1/32$ to 4, draw 200 random coefficient vectors per point, observe the kernel on the window with Gaussian noise $\sigma = 0.05$, and fit it by least squares.

C Residue spectra and Talbot sequences

For odd q and $\gcd(a, q) = 1$ let $g_{q,a}[t] = q^{-1/2} e(at^2/q)$ on $\mathbb{Z}_q$ and let T_q be the cyclic shift. The quadratic Gauss sum gives $|\widehat{g_{q,a}}[k]| = q^{-1/2}$ at every frequency, equivalently $\langle g_{q,a}, T_q^\tau g_{q,a}\rangle = \mathbf{1}[\tau \equiv 0 \pmod q]$: perfect periodic autocorrelation. For pairwise coprime moduli the absolute code $\Phi(n) = K^{-1/2} \bigoplus_k T_{q_k}^{n \bmod q_k} g_{q_k,a_k}$ therefore satisfies $\langle \Phi(n), \Phi(m)\rangle = \#\{k : q_k \mid n-m\}/K$, the residue-matching kernel in which every harmonic of every modulus carries the same weight; such sequences are the sub-slit weights of fractional Talbot images (Berry & Klein, 1996; Fernández-Pousa, 2017). The harmonic residue spectra of Section 3 are the relative form of the same object: T_q is diagonalised by the discrete Fourier transform with eigenvalues $e(-k/q)$, so $\Phi(n+\Delta) = U^\Delta \Phi(n)$ with U a block of rotations by $2\pi k/q_k$. Two remarks. Any flat-spectrum generator, including a one-hot vector, gives the same Gram matrix, so the chirp is a closed form and not a special ingredient; and what the family buys here is exactness, since the kernel then depends on the offset only through residues, which is what Section 5.2 tests.

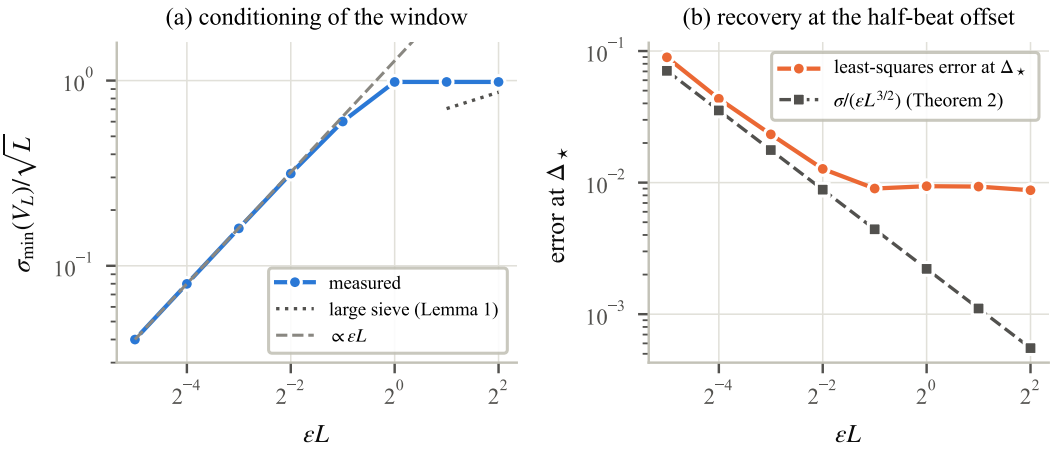


Figure 4: **The theorems in their own regime** (fixed kernels, no network; $L = 512$, $\sigma = 0.05$, 200 draws per point, seed 0). (a) The smallest singular value of the window matrix V_L falls in proportion to εL below the resolution threshold and saturates once the large-sieve bound of Lemma 1 becomes positive. (b) The least-squares error at $\Delta_\star$ follows the $\sigma/(\varepsilon L^{3/2})$ rate of Theorem 2 within a factor 1.2–1.4 for $\varepsilon L \leq 1/4$, and saturates at the resolved-pair floor once the gap is resolved.

D Experimental details

Data. enwik8: the first 10^8 bytes of an English Wikipedia XML dump (Mahoney, 2011), split into the standard 90M/5M/5M bytes for training, validation and test; 205 distinct byte values, vocabulary 256. PG-19 (Rae et al., 2020): the first 150 training books (63.4M bytes) and all 50 validation books (17.7M bytes) of the public release, concatenated with blank lines between books. All numbers in the paper are on the validation splits; the test splits were not used.

Models. Pre-LayerNorm decoder-only transformers without biases, GELU MLPs of width $4d$, tied nothing, no dropout, initialisation $\mathcal{N}(0, 0.02^2)$. Base: $d = 128$, 2 layers, 2 heads (0.46M parameters). Scaled: $d = 256$, 4 layers, 4 heads (3.3M parameters). Both have head dimension 64, i.e. $M = 32$ rotation planes per head, and all heads of a model share the spectrum. Rotations use the half-split convention; angles are formed in float64 and reduced modulo 2π before casting, so positions up to 10^7 keep full phase precision. Evaluation at lengths above the training length uses exact attention computed in query chunks of 512.

Training. AdamW (Loshchilov & Hutter, 2019) with $\beta = (0.9, 0.95)$, weight decay 0.1, gradient clipping at 1.0, 100 warm-up steps and a cosine schedule to 10% of the peak learning rate ($2 \cdot 10^{-3}$ base, 10^{-3} scaled). Batches hold 8,192 characters split into $8192/L$ random windows of length L. Budgets: 4M characters (base; 488 optimiser steps) and 12M characters (scaled; 1,464 steps), identical across spectra; seeds control initialisation and data order.

Evaluation. Per-position negative log-likelihood averaged over random validation windows of length $L' \in \{2048, 4096\}$ with fixed evaluation seed (32 windows, 128 for Figure 1); the same windows are used for every model, so differences below about 0.05 bpc are within the spread of this evaluation and of the seeds; we report bits per character on position ranges. Attention readouts average, per layer and head, the attention weight at each offset over queries at positions 3072–4095 of 6 windows of length 4096. The alias statistic takes the offsets of the $\kappa = 3$ largest in-window weights of a head as its peaks, calls an offset beyond L a predicted alias if, for at least $o = 4$ of the five moduli, its residue is the residue of one of the peaks (possibly of different peaks for different moduli, as the additive form of Proposition 2(ii) allows), and reports the share of predicted aliases among the head's 20 largest weights at offsets L to 2047; the base rate is the share of predicted aliases among all these offsets. With a single peak ($\kappa = 1$) the predicted aliases are the S-aliases of Proposition 2(iii). The masking test of Section 5.2 uses the fixed peak set $\{0, 1, 2, 3\}$ instead of the per-head peaks: it masks every offset in $[L, 8L)$ whose residue modulo at least four of the five

moduli is a residue of $\{0, 1, 2, 3\}$, which is 397 offsets for the moduli $(5, 7, 11, 13, 17)$ and 293 for the integer wavelengths $(6, 8, 11, 15, 20)$ of Section 5.3. The learned envelope phase of a twin pair is $\phi = \arg(c'\bar{c})$ computed from the unrotated query and key planes of the pair; for each layer, head and pair we take the circular mean of ϕ, weighted by $|c||c'|$, over 4,000 random query–key token pairs (query in the second half of the window, key before it) in each of 8 validation windows of length L. Training-free extensions and the interventions of Table 1 evaluate the ranges $[L, 2L)$, $[2L, 4L)$ and $[4L, 8L)$ on windows of length $2L$, $4L$ and $8L$ with the matched scale $s = 2, 4, 8$.

Spectra. Table 4 lists every spectrum with its number of rotating and content planes, its longest period, its slowest beat and, for commensurate spectra, its exact period. No spectrum parameter was tuned on validation data: the bands, jitters and moduli were fixed before training (the 32-plane band so that every period is at most $L/4$ and the planes are evenly spread), and every configuration was trained once per seed.

Table 4: Spectra used in the experiments (32 planes per head): rotating and content planes, longest period $P_{\max}$ and slowest beat B in tokens, and exact period H for commensurate spectra.

spectrum	rot.	content	$P_{\max}$	B	H
RoPE, $b = 10^4$	32	0	47,117	141,272	n/a
RoPE, $b = 500$	32	0	2,587	12,070	n/a
RoPE, $b = 512$ ($b = L$)	32	0	2,647	12,298	n/a
RoPE 10^4, period rule ($P \le 512$)	16	16	471.2	1,413	n/a
RoPE 10^4, $P \le 256$	13	19	198.7	596	n/a
RoPE 10^4, beat rule ($P \le 128$)	11	21	111.7	335	n/a
stratified, 24 in $[1/17, 1/2]$	24	8	14.4	92	n/a
stratified, 32 in $[1/128, 1/2]$	32	0	60.4	111	n/a
stratified, 12 carriers	12	20	12.5	46	n/a
random, 24 in $[1/17, 1/2]$	24	8	16.7	14,265	n/a
twins, 12 carriers, $\varepsilon = 1/2048$	24	8	12.5	2,048	n/a
residue $(4, 5, 7, 9, 11)$	16	16	11.0	99	13,860
residue $(5, 7, 11, 13, 17)$	24	8	17.0	221	85,085
residue $(11, 13, 17, 19)$	28	4	19.0	323	46,189
residue $(5, 9, 11, 13, 15)$ (not coprime)	21	11	15.0	195	6,435

E ADDITIONAL RESULTS

The predictions and their tests. Table 5 lists every prediction the theory of Section 4 makes, the experiment that tests it and its outcome.

The full comparison at $L = 512$. Table 6 gives the comparison at $L = 512$ on all four position ranges, and Table 7 adds position interpolation, which is destructive without fine-tuning, and the same rows evaluated with YaRN's attention temperature.

Short training windows. Table 8 repeats the comparison at $L = 64$, evaluated to $32L$ (one seed for four of the six rows). Every spectrum except the twins degrades faster at this length than at $L = 512$, and the order follows the number of node pairs whose beat is partially resolved at $L = 64$ (period between $L - 1 = 63$ and $8L = 512$ tokens): none for the twins, whose beat of 2,048 tokens is unobservable at this length, 15 for the stratified spectrum, 24 for the residue spectrum, 32 for the random spectrum and 569 for RoPE. The loss on $[16L, 32L)$ grows in the same order.

Resolution thresholds of residue spectra. Proposition 2(i) states that a residue spectrum is resolved once $L - 1$ exceeds the product of its two largest moduli. Figure 5 trains three residue spectra with slowest beats 99, 221 and 323 at $L \in \{64, 128, 256, 512\}$ (one seed each, three for $(5, 7, 11, 13, 17)$ at $L = 512$). The damage on $[2L, 4L)$ falls from 0.39–0.48 bpc at $L = 64$ to 0.21–0.26 bpc at $L = 512$, but not at a sharp threshold: $(4, 5, 7, 9, 11)$ levels off once its slowest beat is resolved, $(11, 13, 17, 19)$ improves most before its beat is resolved, and $(5, 7, 11, 13, 17)$ does

Table 5: **Every prediction, what it rests on, and how it came out.** The middle column separates a consequence of a formal result from a hypothesis about what a trained head does with the kernel: only the first row is a claim about the mathematics alone, and the rows marked *hypothesis* are the ones Section 5 could have falsified. The last two rows did not come out as predicted.

prediction	rests on	test	outcome
An unresolved gap ε costs a factor $1/(\varepsilon L)$ in worst-case accuracy at $\Delta_\star$	Thms 1–2	fixed kernels, no network: a self-consistency check (App. B)	holds within 1.2–1.4×
Resolved and unobservable beats are harmless, partially resolved ones damage	Prop 1 + envelope hypothesis, and the ridge heuristic of App. B	8 beat periods with 5 seeds, plus one slower control (Fig. 2b)	holds; damage non-monotone
Failure begins where the envelope overshoots its trained range	Prop 1 + envelope hypothesis	onsets (Fig. 2c)	6 of 6; 5 of 6 in the first cycle, and 4 of 4 among the beats for which the interval is a sharp test
Far attention peaks near $3/(4\varepsilon)$	Prop 1 + envelope hypothesis	attention by offset (Fig. 1b)	holds; flat for the control
Changing only the gaps repairs the failure, an equal shift does not	beat mechanism, motivated by Thm 2 and Prop 1	stretch vs. shift (Table 1)	52–86% vs. $\leq 9\%$ of the damage
Commensurate spectra reattend at Chinese-remainder offsets	Prop 2	alias share, masking (Sec. 5.2)	59% vs. 4% base rate; 3.1–5.4× random
Rounding wavelengths aliases although no period changes status	Prop 2 and its remark	Resonance-style RoPE (Sec. 5.3)	partly: enrichment and far loss rise, but the attention-matched mask helps more
The onset window also holds at larger scale and after longer training	envelope hypothesis	scaled and 40M-character models (App. E)	**fails**: collapse begins near 800 and 700, before the predicted interval
For RoPE, stretching the beat hazard set repairs more than the period set	beat rule for geometric spectra	stretch on RoPE (App. E)	**fails**: the period set repairs more

not improve after $L = 128$; the stratified reference moves by up to 0.2 bpc between neighbouring lengths. Resolution removes the beats but not the aliases of Proposition 2(iii), which every residue spectrum keeps at any L. With mostly one seed per point, this sweep is not a sharp test, and Theorem 1 bounds how stably the window determines a kernel, not how much loss a model incurs; the twin sweep and the intervention of Section 5.1 are the controlled tests.

Onset measurement. The first failing position in Figure 2c is the first 64-token bin beyond L from which the seed-mean excess over the carriers-only control stays above 0.1 bpc for three bins. We repeated it for thresholds $0.05, 0.1, 0.2$ bpc and runs of two, three and four bins, nine combinations in all. At 0.05 bpc the onsets of all six partially resolved beats lie in the first predicted interval $(\max\{L, 1/(2\varepsilon)\}, 1/\varepsilon]$, at every bin count. At 0.1 bpc five do, and the $1/\varepsilon = 768$ onset is 1,280, inside the second interval $(1{,}152, 1{,}536]$. At 0.2 bpc four do, the 768 onset is 1,536 with two or three bins and 2,240 with four, each inside a later interval, and the 4,096 beat never crosses the threshold. So in all nine combinations every onset that exists falls in an interval where the envelope overshoots its trained range; what the threshold changes is only which excursion is detected first. For the 768 and 1,024 beats the first interval starts at L, so these two points test the prediction less sharply than the other four.

A spectrum that was not built to fail. A random spectrum with 24 frequencies in the band of the stratified one has all periods below 17 tokens, and at $L = 512$ only two of its plane pairs have partially resolved beats (877 and 1,047 tokens), against twelve for the twins. With its first

Table 6: **Train short, test long** on enwik8 (base model, $L = 512$). Bits per character on position ranges of validation windows of length $8L = 4096$; mean $\pm$ population standard deviation over n seeds. Training-free extensions use the scale s matched to each range ($s = 2, 4, 8$) and are evaluated range by range on windows of length sL, so their in-window numbers can differ slightly from those of the unchanged model. B: slowest beat of the spectrum in tokens.

spectrum / method	B	n	$[0, L)$	$[L, 2L)$	$[2L, 4L)$	$[4L, 8L)$
RoPE, $b = 10^4$	141,272	3	2.91 ± 0.01	3.12 ± 0.12	3.70 ± 0.11	4.07 ± 0.07
+ NTK-aware scaling	n/a	3	2.90 ± 0.01	2.99 ± 0.02	3.17 ± 0.10	3.56 ± 0.14
+ YaRN	n/a	3	2.90 ± 0.01	2.96 ± 0.02	3.00 ± 0.03	3.13 ± 0.06
+ YaRN ramp only	n/a	3	2.90 ± 0.01	2.99 ± 0.02	3.08 ± 0.02	3.31 ± 0.02
RoPE, period rule ($P \le L$)	1,413	3	2.93 ± 0.01	2.98 ± 0.01	3.17 ± 0.03	3.40 ± 0.02
with integer wavelengths	1,409	3	2.93 ± 0.01	2.99 ± 0.03	3.21 ± 0.04	3.49 ± 0.07
RoPE, beat rule ($P \le L/4$)	335	3	2.96 ± 0.03	3.02 ± 0.04	3.16 ± 0.03	3.44 ± 0.04
ALiBi	n/a	3	3.15 ± 0.01	3.17 ± 0.01	3.20 ± 0.01	3.16 ± 0.01
NoPE	n/a	3	3.64 ± 0.00	3.79 ± 0.01	3.85 ± 0.01	3.90 ± 0.00
twins, $\varepsilon = 1/2048$	2,048	5	2.95 ± 0.02	3.03 ± 0.01	3.67 ± 0.10	4.12 ± 0.16
residue $(5, 7, 11, 13, 17)$	221	3	2.94 ± 0.01	3.04 ± 0.02	3.19 ± 0.03	3.43 ± 0.04
stratified, 32 planes	111	3	2.91 ± 0.01	2.98 ± 0.03	3.10 ± 0.07	3.32 ± 0.12
stratified, 24 planes	92	3	2.93 ± 0.01	3.01 ± 0.02	3.12 ± 0.03	3.30 ± 0.03

Table 7: **Position interpolation and YaRN's attention temperature** (base model, $L = 512$, enwik8; both use the matched scale s as in Table 6, and the temperature multiplies queries and keys by $0.1 \ln s + 1$). Bits per character, mean $\pm$ population standard deviation over n seeds. The temperature lowers the far loss of every spectrum except the twins by a similar amount, so what remains ranks the kernels.

spectrum	B	n	$[0, L)$	$[L, 2L)$	$[2L, 4L)$	$[4L, 8L)$
RoPE, $b = 10^4$, + PI	n/a	3	2.90 ± 0.01	3.75 ± 0.05	4.34 ± 0.09	4.63 ± 0.10
RoPE, $b = 10^4$, + temperature	141,272	3	2.90 ± 0.01	3.10 ± 0.10	3.55 ± 0.12	3.91 ± 0.07
twins, $\varepsilon = 1/2048$	2,048	3	2.95 ± 0.02	3.04 ± 0.02	3.79 ± 0.07	4.16 ± 0.10
residue $(5, 7, 11, 13, 17)$	221	3	2.94 ± 0.01	3.06 ± 0.02	3.11 ± 0.03	3.24 ± 0.03
RoPE, period rule ($P \le L$)	1,413	3	2.91 ± 0.02	2.99 ± 0.03	3.04 ± 0.04	3.14 ± 0.06
RoPE, beat rule ($P \le L/4$)	335	3	2.96 ± 0.03	3.02 ± 0.04	3.04 ± 0.03	3.23 ± 0.05
stratified, 32 planes	111	3	2.91 ± 0.02	2.99 ± 0.03	3.01 ± 0.05	3.12 ± 0.09
stratified, 24 planes	92	3	2.94 ± 0.01	3.03 ± 0.01	3.02 ± 0.01	3.09 ± 0.02

seed in hand we recorded a prediction for the replication seeds: a small excess damage over the stratified spectrum and no collapse. Over 3 seeds its damage on $[4L, 8L)$ is 0.39 bpc, against 0.37 for the stratified spectrum, and the two spectra reach the same loss there (3.30 against 3.30); no seed collapses. Two partially resolved pairs are not enough for a failure as sharp as the twins', where twelve pairs share one gap.

Do the two counts ever differ on a spectrum in use? The spectra of Section 5 were built so that the two criteria disagree. Table 9 asks the same question of spectra that are used or proposed, at their own training lengths, using nothing but the frequency list. The answer has two halves. On the RoPE family the two criteria *order* the planes identically, and for a purely geometric spectrum this is an identity rather than a measurement: the slowest beat a plane takes part in is its gap to its neighbour, $1/((1 - r)f_j) = P_j/(1 - r)$, a fixed multiple of its own period for every plane but the lowest. Kendall τ between the two per-plane scores is therefore 1 for RoPE at any base and for position interpolation, which divides every frequency by the same factor; the informative rows are YaRN's ramp and the rounded wavelengths, which do perturb the geometric structure, and even there τ stays above 0.996. A gap list is thus not a re-ranking of the planes for these spectra, which is the arithmetic behind the negative result of Section 5.3. The other half is where the counts come apart. Clipping by period, the remedy that FoPE applies, and near which p-RoPE and CoPE cut

Table 8: Training length $L = 64$ (base model, enwik8); bits per character on position ranges up to $32L$, mean±std over n seeds.

spectrum	n	$[0, L)$	$[L, 2L)$	$[2L, 4L)$	$[4L, 8L)$	$[8L, 16L)$	$[16L, 32L)$
RoPE, $b = 10^4$	1	2.61	2.80	3.83	4.32	4.70	4.88
NoPE	1	3.35	3.55	3.77	3.87	4.03	4.07
random, 24 planes	1	2.65	2.70	3.24	3.76	4.26	4.57
residue $(5, 7, 11, 13, 17)$	1	2.66	2.74	3.07	3.43	3.99	4.40
stratified, 24 planes	3	2.67 ± 0.02	2.72 ± 0.02	3.04 ± 0.06	3.31 ± 0.08	3.74 ± 0.11	4.10 ± 0.14
twins, $\varepsilon = 1/2048$	3	2.68 ± 0.02	2.75 ± 0.01	3.05 ± 0.06	3.24 ± 0.08	3.55 ± 0.12	3.86 ± 0.13

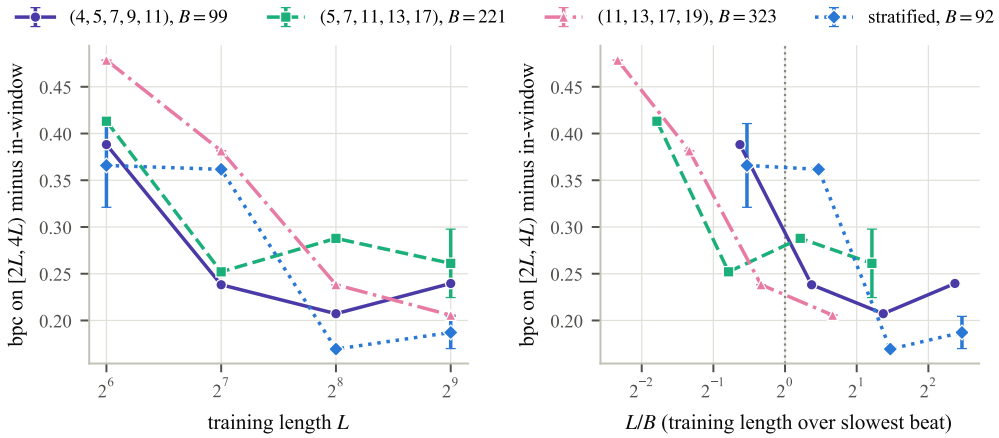


Figure 5: Extrapolation damage (bpc on $[2L, 4L)$ minus in-window) of residue spectra trained at different lengths, against L (left) and against L/B (right), where $B = q_{K-1}q_K$ is the slowest beat; the dotted line marks $L = B$. The stratified spectrum ($B = 92$) is shown for reference. One seed per point, except three for $(5, 7, 11, 13, 17)$ at $L = 512$ and for the stratified spectrum at $L = 64$ and 512 (mean, error bars one standard deviation).

with their own thresholds, drives the period count to zero by construction while 28 plane pairs of the remaining 46 planes still beat into $(L-1, 8L]$, which is 56 node pairs once each plane pair is counted on both sides of the circle; at our own scale the same construction leaves 10 such pairs (Table 10), and it is the spectrum whose rounded-wavelength variant develops the aliases of Section 5.3. Only dropping every plane with a period above $L/8$, which is roughly the beat rule, empties both counts. So on spectra in use the gap criterion is not a better ordering; it is a hazard that survives the standard remedy, and it is visible before training.

Is there one criterion? Periods and gaps are two hazard counts on the same node set, and the period criterion is the special case in which one of the two nodes is the content node at 0. A natural unification is to count every pair of nodes whose difference leaves the arc the window trained, that is every beat period in $(L-1, 8L]$, and to ask whether that count orders the measured damage better than the number of planes with $P > L$. Table 10 does this for the ten spectra we train at $L = 512$. Neither count is sufficient: the twins have no plane with $P > L$ and the largest damage, RoPE has both hazards and the second largest, and five spectra with no hazard of either kind still lose 0.37–0.49 bpc beyond the window, which is the cost of leaving the window that every model pays. Both counts point the same way over the ten spectra, but they differ in how much they can order: the tie-corrected Spearman correlation with the damage is 0.65 ($p = 0.04$) for the gap count and 0.41 ($p = 0.24$) for the period count, and the reason is visible in the table. The period count is nonzero for one of the ten spectra, plain RoPE, so its correlation rests entirely on that one row; drop RoPE and the period count is constant at zero, which makes its correlation undefined, while the gap count is nonzero for five spectra and still orders the remaining nine (0.52, $p = 0.15$). We therefore do not propose the gap count as a replacement criterion, and with ten spectra neither correlation is evidence

Table 9: **Hazard counts for rotary spectra that are used or proposed**, computed from the frequency list alone at each spectrum's own training length L, with no model trained or run: rotating planes, planes whose period exceeds L, node pairs whose beat period lies in $(L-1, 8L]$, and Kendall τ between two per-plane hazard scores, a plane's own period and the slowest beat it takes part in. Every spectrum here has $\tau \geq 0.996$: on the RoPE family the two criteria put the planes in the same order, so they can differ only in where the threshold falls. What they do not agree on is what is left afterwards. Clipping by period, the usual remedy, drives the period count to zero while 56 node pairs still beat into $(L-1, 8L]$; only dropping every plane with a period above $L/8$ removes both.

spectrum	planes	$P > L$	gaps in $(L-1, 8L]$	Kendall τ
RoPE, $b = 10^4$, $L = 4{,}096$	64	18	618	1.000
RoPE, $b = 5 \cdot 10^5$, $L = 8{,}192$	64	29	938	1.000
RoPE, $b = 10^6$, $L = 32{,}768$	64	24	724	1.000
+ position interpolation, $s = 4$	64	28	1192	1.000
+ position interpolation, $s = 8$	64	33	1476	1.000
+ YaRN ramp, $s = 4$	64	27	1020	0.998
+ YaRN ramp, $s = 8$	64	29	984	0.997
+ integer wavelengths	64	18	618	0.997
+ period clipping ($P \leq L$)	46	0	56	1.000
+ high-frequency subset ($P \leq L/8$)	31	0	0	0.999

on its own: what the table shows is that the gap count separates spectra the period count cannot distinguish, because it is the only one of the two that varies among them.

Table 10: Two hazard counts against the measured damage (base model, $L = 512$, seed means): planes whose own period exceeds the window, and node pairs whose beat period lies in $(L-1, 8L]$ (partially resolved gaps; pairs include a plane with its mirror and with the content node, so the period criterion is a special case). Damage is the loss on $[4L, 8L)$ minus the in-window loss, in bits per character. Tie-corrected Spearman correlation with the damage over these 10 spectra: 0.41 ($p = 0.24$) for the period count, which is nonzero for only 1 of them, and 0.65 ($p = 0.04$) for the gap count, which is nonzero for 5.

spectrum	planes with $P > L$	partially resolved gaps	damage
RoPE, $b = 10^4$	16	353	1.16
RoPE, period rule ($P \leq L$)	0	10	0.47
with integer wavelengths	0	10	0.56
RoPE, beat rule ($P \leq L/4$)	0	0	0.48
twins, $\varepsilon = 1/2048$	0	24	1.17
carriers only (12 planes)	0	0	0.46
residue $(5, 7, 11, 13, 17)$	0	0	0.49
random, 24 planes	0	4	0.39
stratified, 24 planes	0	0	0.37
stratified, 32 planes	0	0	0.42

Which RoPE planes need stretching? A failed prediction. The beat rule predicts that the neighbour beats of the RoPE planes with periods in $((1-r)L, L]$ contribute to the failure. We test it on the trained RoPE model by stretching at inference, with the matched scale s and no temperature, either only the planes whose period exceeds L (the period rule's hazard set) or every plane whose period exceeds $(1-r)L = L/4$ (the beat rule's hazard set). The beat account predicts that the second repairs more. It does not: stretching the planes with periods above L gives 2.96, 3.03 and 3.29 bpc on $[L, 2L)$, $[2L, 4L)$ and $[4L, 8L)$, stretching the larger set gives 2.97, 3.08 and 3.37, and YaRN's ramp without its temperature, which stretches both sets in part, gives 2.99, 3.08 and 3.31 (means over 3 seeds; the unchanged model gives 3.12, 3.70 and 4.07). In RoPE a gap cannot be stretched without stretching the periods of the planes that form it, and the planes with periods in $(L/4, L]$ carry patterns the model uses at short offsets, so this test cannot isolate their beats; but its outcome

means that, for RoPE at this scale, the planes with periods above L carry most of the damage and the period rule captures it. The beat mechanism itself is isolated by the twin experiments of Section 5.1, where only the gap changes.

Replication on PG-19 and with a larger model. Tables 11 and 12 repeat the main comparison on PG-19 books and with a 4-layer model of width 256 trained on 12M characters. On PG-19 the pattern of the base model holds: the twins track the carriers-only control up to about position 1,024 and then collapse (4.01 bpc on $[4L, 8L)$ against 3.28 for the control and 2.94 for the stratified spectra, three seeds for the first two); clipping RoPE by periods removes most of its damage (4.17 to 3.09) and clipping by beats adds little (3.08); the residue spectrum sits between RoPE and the stratified spectra; and the evenly spread spectra are best beyond the window and among the best inside it.

With the scaled model (Table 12; 1,464 steps) the picture of the base model holds, with larger effects. RoPE, the best spectrum inside the window (2.08 bpc), degrades most (4.81 on $[4L, 8L)$); YaRN recovers almost all of it (2.33) and the temperature alone very little (4.69). The twins track the stratified spectrum up to about position 800 and then fail (3.69 against 2.65 on $[4L, 8L)$); stretching only their gaps at inference removes 50%, 86% and 77% of the damage on $[L, 2L)$, $[2L, 4L)$ and $[4L, 8L)$, and the equal shift at most 6% (percentages from the unrounded runs behind the three matched rows of Table 12). Their learned phase is again near $\pi/2$ (circular mean 1.42 rad), for which the envelope magnitude regains its value at $\Delta = 0$ near position 1,100, so the scaled twins fail somewhat earlier than the magnitude alone predicts. The residue spectrum loses 0.63 bpc to $[4L, 8L)$, between the stratified spectrum (0.54) and RoPE (2.73), and its heads put 35–100% (mean 57% over 16 heads) of their 20 largest out-of-window weights on predicted aliases (base rate 3.4%; shares of 2.5%, 5.0% and 5.6% for RoPE, the twins and the stratified spectrum). With the temperature the stratified spectrum stays within 0.30 bpc of its in-window loss up to $8L$, while RoPE with YaRN stays within 0.23: at this scale YaRN does better. The scaled models, like the base models, barely improve with position inside the window.

Table 11: Replication on PG-19 ($L = 512$, base model); bits per character on position ranges up to $8L$, mean over the n seeds of each row. The three indented rows repeat the intervention of Table 1 on this corpus and share its matched-window protocol, so the fractions of the damage removed quoted in Section 5.1 follow from them.

spectrum	n	$[0, L)$	$[L, 2L)$	$[2L, 4L)$	$[4L, 8L)$
RoPE, $b = 10^4$	1	2.79	3.20	3.78	4.17
RoPE, period rule ($P \le L$)	1	2.78	2.83	2.97	3.09
RoPE, beat rule ($P \le L/4$)	1	2.81	2.85	2.94	3.08
residue $(5, 7, 11, 13, 17)$	1	2.74	2.81	2.96	3.08
carriers only (12 planes)	3	2.78 ± 0.02	2.85 ± 0.01	3.04 ± 0.06	3.28 ± 0.10
twins, $\varepsilon = 1/2048$	3	2.74 ± 0.02	2.78 ± 0.01	3.58 ± 0.19	4.01 ± 0.24
unchanged, matched windows	3	2.82 ± 0.02	2.82 ± 0.01	3.58 ± 0.18	4.01 ± 0.24
shift (gaps kept)	3	2.82 ± 0.02	2.82 ± 0.01	3.46 ± 0.19	3.94 ± 0.27
stretch (gaps changed)	3	2.82 ± 0.02	2.81 ± 0.02	2.87 ± 0.03	3.04 ± 0.07
stratified, 24 planes	1	2.75	2.76	2.86	2.94
stratified, 32 planes	1	2.73	2.76	2.85	2.94

Ten times longer training. Training the base model on 40M characters (4,882 steps; Table 13, one seed) lowers its in-window loss from about 2.95 to 2.14–2.17 bpc but does not make it fall with position (2.0–2.2 bpc from position 16 to 512), so the longer-trained models still use little long-range context, contrary to what we expected when we started this run. Every spectrum extrapolates worse relative to its in-window loss than after brief training, and the twins far worse: their excess over the carriers-only control starts near position 700 (0.21 bpc on $[704, 768)$, 0.56 on $[768, 832)$, about 2 beyond 1,024) and reaches 1.94 bpc on $[4L, 8L)$, while the stratified spectrum stays best (0.73 above its in-window loss on $[4L, 8L)$, against 1.40 for the control). The model now uses the beat much more inside the window (moving both members onto the carrier costs 0.48 bpc there, against 0.04 after brief training), and changing only the gaps still repairs most of the failure: the stretch removes 87%, 85% and 67% of the damage on $[L, 2L)$, $[2L, 4L)$ and $[4L, 8L)$ and leaves less damage than the control has, while the shift removes at most 4%. The learned phase is again

Table 12: Scaled model (4 layers, width 256, 3.3M parameters, 12M characters, $L = 512$, enwik8, one seed); bits per character on position ranges up to $8L$; YaRN, the temperature and the twin interventions (shift and stretch as in Table 1) use the matched scale s as in Table 6. The three matched twin rows share one protocol, which is the protocol behind the fractions of the damage removed quoted in Section 5.1 and Appendix E (computed from the unrounded runs); the first twin row is the single $8L$ evaluation, comparable with the other spectra.

spectrum	n	$[0, L)$	$[L, 2L)$	$[2L, 4L)$	$[4L, 8L)$
RoPE, $b = 10^4$	1	2.08	3.50	4.45	4.81
+ YaRN	1	2.10	2.17	2.21	2.33
+ temperature only	1	2.10	3.51	4.41	4.69
twins, $\varepsilon = 1/2048$	1	2.12	2.26	3.07	3.69
+ temperature	1	2.13	2.30	3.10	3.58
unchanged, matched windows	1	2.13	2.30	3.13	3.69
shift (gaps kept)	1	2.13	2.29	3.09	3.62
stretch (gaps changed)	1	2.13	2.22	2.27	2.50
residue $(5, 7, 11, 13, 17)$	1	2.12	2.24	2.46	2.75
+ temperature	1	2.15	2.25	2.39	2.57
stratified, 32 planes	1	2.10	2.19	2.35	2.65
+ temperature	1	2.12	2.20	2.27	2.43

near $\pi/2$ (1.65 rad), for which the envelope magnitude regains its value at $\Delta = 0$ near position 970; as in the scaled model, the failure starts earlier than the magnitude alone predicts but after L, consistent with a model that also reads the phase of the envelope, which leaves the trained arc right after L (Proposition 1).

Table 13: Ten times longer training (base model, 40M characters, 4,882 steps, $L = 512$, one seed). The twin rows use the matched-scale windows of Table 1, the other rows windows of length $8L$ as in Table 6.

model	$[0, L)$	$[L, 2L)$	$[2L, 4L)$	$[4L, 8L)$
twins, $\varepsilon = 1/2048$, unchanged	2.17	3.14	5.38	5.52
shift (gaps kept)	2.17	3.10	5.42	5.52
stretch (gaps changed)	2.17	2.30	2.65	3.27
carriers-only control	2.17	2.64	3.08	3.57
stratified, 24 planes	2.14	2.27	2.47	2.87

Exact recurrence of commensurate spectra. To test Proposition 2(ii) directly, we split each validation window into two halves of 256 characters and shift the positions of the second half by a gap G while keeping the text contiguous, so that every query of the second half sees the first half at offsets $\Delta + G$ (Figure 6; 174 gaps between 1 and 200,000, 48 windows). The coprime residue model returns to its no-gap loss up to floating-point precision at $G = 85{,}085$ and 170,170, its period, and the non-coprime model $(5, 9, 11, 13, 15)$ at $G = 6{,}435$, 12,870 and 19,305, multiples of the least common multiple of its moduli; neither returns at the other's period (at least 0.086 and 0.050 bpc above), and the stratified and RoPE models return at no gap beyond the window (at least 0.024 bpc above for every $G \geq 512$). In between, gaps divisible by a product of four moduli cost the coprime model a median of 0.027 bpc, against 0.053 bpc for gaps divisible by no product of two or more moduli, and least at $17{,}017 = 7 \cdot 11 \cdot 13 \cdot 17$ (0.011 bpc): partial agreement buys partial recovery, as the additive form of Proposition 2(ii) predicts, though unevenly across moduli. The exact returns hold by construction for any model with these spectra; what training adds is the partial recovery at partial agreement.

Robustness and specificity of the alias statistic. The statistic of Figure 3b has two free choices, the number κ of in-window peaks and the agreement order o. Table 14 varies both. For the residue model the share of the 20 largest out-of-window weights on predicted aliases exceeds its base rate by a factor of 5 to 55 for every choice and seed, while the stratified, twin and RoPE models stay within

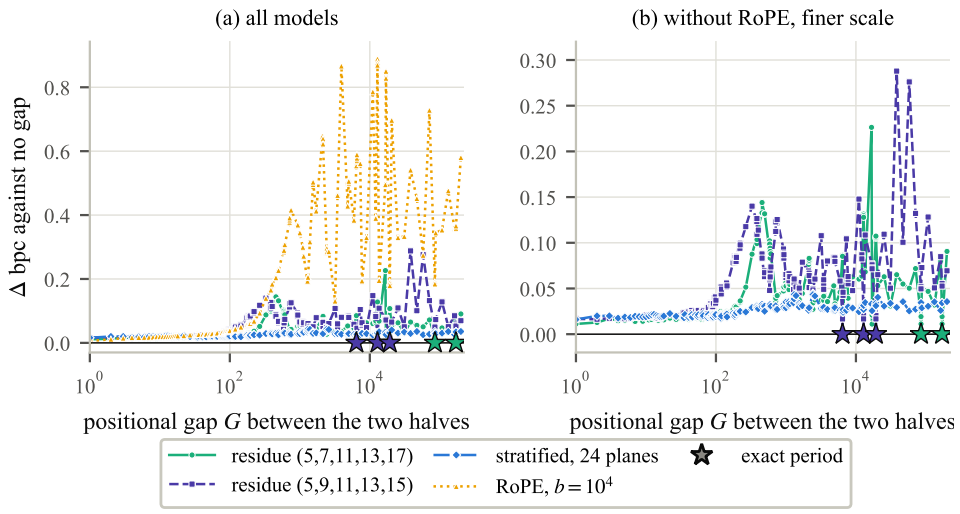


Figure 6: Gap test ($L = 512$). Second-half bits per character relative to no gap, when the positions of the second half of a 512-character window are shifted by G while the text stays contiguous: (a) all four models, (b) the three models other than RoPE on a finer scale. Stars: the loss at the exact periods of the two residue spectra (85,085 and 170,170 for $(5, 7, 11, 13, 17)$; 6,435, 12,870 and 19,305 for $(5, 9, 11, 13, 15)$).

a factor of about three of their base rates (at most 0.26). For significance we use a randomization test rather than a binomial one: the top-20 offsets are order statistics of a correlated attention profile and the alias set is an arithmetic lattice, so the offsets are not independent draws and a binomial p-value would overstate the precision. Instead we hold each head's attention profile fixed, so that the concentration of attention and the identity of the top-20 offsets are common to the null and the observation, and randomise only the arithmetic, replacing the head's own in-window peak set by a random peak set of the same size drawn from the same moduli; the p-value is the share of 2,000 draws whose alias share reaches the observed one, computed with an added draw so that it is never reported as zero. One caveat runs against us and we correct for it. The measured peaks are consecutive offsets, whose residues are distinct modulo every modulus above four, so their alias set is larger than that of a random triple, which usually collides in some modulus; the unrestricted null therefore draws sparser lattices and is mildly anti-conservative. We report both it and a density-matched null that keeps only draws whose alias set is at least as large as the observed one, which can only raise the p-value. Every one of the 12 residue heads over the three seeds gives $p \leq 0.0025$ unrestricted and $p \leq 0.0080$ density-matched, while among the 44 control heads the smallest is 0.0074 and 0.0324 and 41 and 42 respectively exceed 0.05. The statistic is specific to the model's own arithmetic: analysed with the disjoint moduli $(8, 9, 19, 23, 29)$ the residue model shows no enrichment at all, and with the moduli $(5, 9, 11, 13, 15)$, three of which it shares, a weaker one. The enrichment replicates in two further seeds (0.55 and 0.58 in the main setting, against base rates near 0.04), and a residue model with the non-coprime moduli $(5, 9, 11, 13, 15)$ shows it for its own moduli (0.54 against 0.03) but hardly for the coprime ones (0.09 against 0.02).

Masking controls. Table 15 lists every masking run. Besides random offsets, we mask as many of the most-attended offsets beyond L that are not predicted aliases, ranked by the head-averaged attention of Figure 3. For the residue models the alias mask helps more than this attention-matched mask (0.11–0.22 against 0.07–0.09 bpc on $[4L, 8L)$), while for the stratified models it is the other way round (0.03 against 0.08–0.11): the aliases are the harmful far offsets only where the arithmetic predicts them. For both period-clipped RoPE models the attention-matched mask helps most (0.15–0.25 bpc), a sign that much of their far loss comes from attention to distant tokens in general, the dilution that the temperature of Section 5.3 also removes.

Integer wavelengths. Resonance RoPE (Wang et al., 2024) rounds the wavelengths of RoPE to integers. Applied to period-clipped RoPE ($b = 10^4$, planes with period above L carry content),

Table 14: Robustness and specificity of the alias statistic ($L = 512$). Share of each head's 20 largest out-of-window attention weights (offsets 512–2047) on predicted residue aliases, mean over heads, with the share of such offsets among all offsets in parentheses. An offset is a predicted alias if its residues modulo at least o of the moduli are residues of the κ largest in-window peaks (possibly of different peaks for different moduli, as the additive form of Proposition 2(ii) allows); with $\kappa = 1$ these are the S-aliases of Proposition 2(iii). The main setting is $\kappa = 3$, $o \geq 4$, the one used in Figure 3b; significance for it is the randomization test described above, not a binomial one.

model	seed	moduli	$\kappa = 1, o \geq 3$	$\kappa = 3, o \geq 4$	$\kappa = 5, o \geq 4$	$\kappa = 5, o = 5$
twins	0	$(5, 7, 11, 13, 17)$	0.01 (0.012)	0.03 (0.038)	0.17 (0.180)	0.00 (0.027)
stratified, 24 planes	0	$(5, 7, 11, 13, 17)$	0.04 (0.012)	0.04 (0.038)	0.14 (0.171)	0.03 (0.025)
RoPE	0	$(5, 7, 11, 13, 17)$	0.01 (0.012)	0.03 (0.038)	0.20 (0.237)	0.08 (0.037)
residue $(5, 7, 11, 13, 17)$	0	own	0.61 (0.012)	0.64 (0.038)	0.77 (0.104)	0.31 (0.014)
residue $(5, 7, 11, 13, 17)$	0	$(5, 9, 11, 13, 15)$	0.16 (0.031)	0.21 (0.051)	0.36 (0.131)	0.07 (0.017)
residue $(5, 7, 11, 13, 17)$	0	$(8, 9, 19, 23, 29)$	0.00 (0.001)	0.00 (0.007)	0.03 (0.028)	0.00 (0.003)
residue $(5, 7, 11, 13, 17)$	1	own	0.60 (0.012)	0.55 (0.038)	0.84 (0.155)	0.33 (0.021)
residue $(5, 7, 11, 13, 17)$	1	$(5, 9, 11, 13, 15)$	0.16 (0.031)	0.16 (0.054)	0.44 (0.169)	0.08 (0.024)
residue $(5, 7, 11, 13, 17)$	1	$(8, 9, 19, 23, 29)$	0.00 (0.001)	0.00 (0.006)	0.03 (0.034)	0.00 (0.003)
residue $(5, 9, 11, 13, 15)$	0	own	0.50 (0.031)	0.54 (0.029)	0.79 (0.106)	0.24 (0.010)
residue $(5, 9, 11, 13, 15)$	0	$(5, 7, 11, 13, 17)$	0.19 (0.012)	0.09 (0.023)	0.41 (0.114)	0.11 (0.016)
residue $(5, 9, 11, 13, 15)$	0	$(8, 9, 19, 23, 29)$	0.00 (0.001)	0.00 (0.004)	0.01 (0.023)	0.00 (0.002)
stratified, 32 planes	0	$(5, 7, 11, 13, 17)$	0.03 (0.012)	0.05 (0.038)	0.25 (0.214)	0.06 (0.033)
twins	1	$(5, 7, 11, 13, 17)$	0.01 (0.012)	0.04 (0.035)	0.15 (0.160)	0.03 (0.022)
RoPE	1	$(5, 7, 11, 13, 17)$	0.01 (0.012)	0.06 (0.038)	0.26 (0.209)	0.06 (0.032)
stratified, 24 planes	1	$(5, 7, 11, 13, 17)$	0.01 (0.012)	0.01 (0.036)	0.10 (0.150)	0.01 (0.021)
residue $(5, 7, 11, 13, 17)$	2	own	0.66 (0.012)	0.58 (0.033)	0.81 (0.134)	0.26 (0.019)
residue $(5, 7, 11, 13, 17)$	2	$(5, 9, 11, 13, 15)$	0.14 (0.031)	0.12 (0.047)	0.44 (0.164)	0.07 (0.024)
residue $(5, 7, 11, 13, 17)$	2	$(8, 9, 19, 23, 29)$	0.00 (0.001)	0.00 (0.006)	0.03 (0.031)	0.00 (0.003)
stratified, 24 planes	2	$(5, 7, 11, 13, 17)$	0.03 (0.012)	0.03 (0.034)	0.14 (0.157)	0.01 (0.022)
twins	2	$(5, 7, 11, 13, 17)$	0.00 (0.012)	0.01 (0.033)	0.15 (0.177)	0.01 (0.025)
RoPE	2	$(5, 7, 11, 13, 17)$	0.00 (0.012)	0.03 (0.038)	0.16 (0.237)	0.03 (0.037)

rounding gives the periods 6, 8, 11, 15, 20, 26, …, 471, moves none by more than half a token and keeps the slowest beat (1,409 against 1,413 tokens), so the period account predicts no change; but the five fastest planes become commensurate, so that, for instance, $D + 660$ agrees with D modulo 6, 11, 15 and 20, and $D + 1{,}320$ modulo all five. Before training we recorded four predictions (unchanged in-window loss; alias share at least twice that of the unrounded model and of its base rate; alias mask at least twice as helpful as random masks and more helpful than in the unrounded model; far loss not lower); all four held over three seeds each: in-window loss 2.93 for both, alias share 20–25% per seed against 1–4% (base rate 4.2%), alias mask 0.08–0.16 bpc on $[4L, 8L)$ against 0.02–0.03 for the unrounded model and 2.7–7.0 times its random masks, and far loss 3.49 (3.39–3.56) against 3.40. This tests the rounding step only, at our scale and without the interpolation that Resonance RoPE combines it with for the planes above the critical dimension.

Smaller bases and large-model arithmetic. With $b = 500$ or $b = L$ (one seed each), RoPE does worse just beyond the window and slightly better further out (3.26 and 3.23 bpc on $[L, 2L)$, 3.90 and 3.88 on $[4L, 8L)$, against 3.12 and 4.07 for $b = 10^4$). For Llama-3-style RoPE ($b = 5 \cdot 10^5$, 64 planes, $L = 8{,}192$; Grattafiori et al., 2024) the same arithmetic gives 29 planes with periods above L and 8 more (periods 1,594 to 6,695 tokens) whose neighbour beats are partially resolved; we did not test pretrained models.

Table 15: Masking at inference (base models, $L = 512$): bits per character on $[4L, 8L)$ without a mask, with the predicted alias offsets masked, with as many of the most-attended other offsets beyond L masked, and with as many random offsets beyond L masked (mean of 5 masks). Aliases are defined by the moduli $(5, 7, 11, 13, 17)$ for the first two models and by the integer wavelengths $(6, 8, 11, 15, 20)$ for the last two.

model	seed	masked	none	aliases	most attended	random
residue $(5, 7, 11, 13, 17)$	0	397	3.46	3.24	3.37	3.42
residue $(5, 7, 11, 13, 17)$	1	397	3.45	3.32	3.37	3.41
residue $(5, 7, 11, 13, 17)$	2	397	3.37	3.26	3.30	3.34
stratified, 24 planes	0	397	3.27	3.24	3.16	3.24
stratified, 24 planes	1	397	3.33	3.30	3.24	3.30
stratified, 24 planes	2	397	3.30	3.27	3.22	3.27
period-clipped RoPE, integer wavelengths	0	293	3.51	3.43	3.28	3.48
period-clipped RoPE, integer wavelengths	1	293	3.56	3.39	3.34	3.53
period-clipped RoPE, integer wavelengths	2	293	3.39	3.30	3.24	3.36
period-clipped RoPE	0	293	3.43	3.41	3.19	3.41
period-clipped RoPE	1	293	3.38	3.36	3.17	3.35
period-clipped RoPE	2	293	3.39	3.36	3.18	3.35