# HARP: HIERARCHICALLY ANCHORED ROTARY PHASES FOR LONG AND STRUCTURED CONTEXTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Rotary Position Embedding (RoPE) derives one fixed global phase schedule from the raw token index. The raw index must therefore encode local order, long-range distance, and structured layout at once. Recent analyses reveal the overload to long-context retrieval failures, to frequency crowding, and to fragility under the `bfloat16` (BF16) format. We propose Hierarchically Anchored Rotary Phases (HARP), which factors the raw index into a coarse chunk index, a within-chunk offset, and optional ordered structure coordinates. Each coordinate governs its own band of learnable rotary phases, while a fraction of channels stay unrotated as a content bypass. The encoding comes in a general form (HARP-Core) and a separated-band form (HARP-SB). We initialize both exactly on the RoPE submanifold, so training departs from RoPE only when the data rewards the departure. The factorization yields an attention kernel over relative hierarchical coordinates rather than the raw gap alone. The resulting kernel recovers RoPE exactly and strictly enlarges the positional kernel class. Separating the bands further bounds the phase budget at $O(\sqrt{L})$ rather than $O(L)$ in context length $L$. We evaluate HARP through deterministic experiments, synthetic probes, from-scratch pretraining at $160$M and $420$M parameters, a BF16 analysis, and ablations. On structural tasks HARP separates equal-distance token pairs that every purely relative encoding maps to identical logits. Beyond that, the trained model concentrates attention inside the current chunk, carries non-redundant information in all three bands, and HARP-Core raises scope resolution from $4.7\%$ to $42.5\%$. On length and precision at scale, by contrast, the advantage narrows from a clear win at $160$M to parity at $420$M. One reason is that training specializes the global band alone, which puts the bounded phase budget out of reach. HARP therefore encodes hierarchy where RoPE encodes distance alone.

## 1 INTRODUCTION

Transformer-based language models (LMs) are permutation-equivariant, which forces us to inject a position signal before attention can recover token order. Rotary Position Embedding (RoPE) (Su et al., 2024) is the de facto choice in modern decoder-only LMs. The encoding rotates each two-dimensional block of every query and key by an angle proportional to the absolute token index. The attention logit therefore depends only on the relative offset at no parameter cost and with a family of context-extension recipes built on the same rotation (Peng et al., 2024).

The convenience of RoPE comes at the cost of a single overloaded coordinate. One global phase schedule, a geometric sweep of base frequencies over the raw token index (Su et al., 2024), encodes three relations at once. The schedule handles local order within a sentence, long-range distance across a document, and any structured layout the input carries, such as section nesting or syntactic scope (Zhang et al., 2024a). Recent works connect this overload to concrete failures. Frequency bands take on positional and semantic roles that the architecture leaves implicit (Barbero et al., 2025), and the base frequency alone governs how far a model extrapolates (Liu et al., 2024c). Long contexts expose the consequences since models retrieve poorly from the middle of an input (Liu et al., 2024a) and lean on the first tokens as attention sinks (Xiao et al., 2024). Moreover, under the `bfloat16` (BF16) format the accumulated phase makes the relative property fragile at long



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.

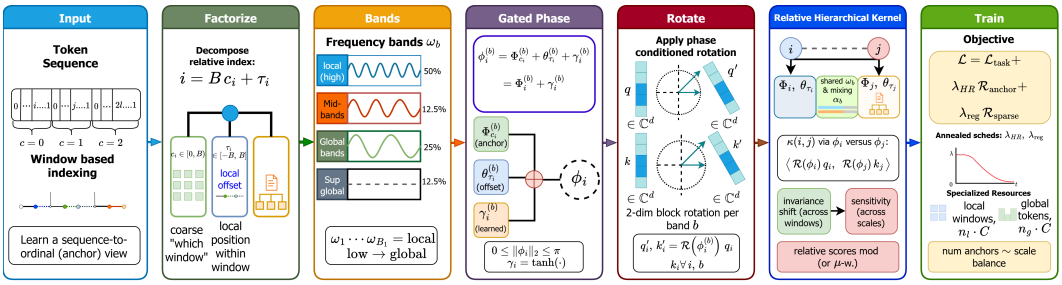



Figure 1: An end-to-end overview of HARP. (1) We split tokens into anchored chunks of size $B$. (2) We factor each index into a chunk index $c_i$, an offset $r_i$, and optional structure coordinates $\boldsymbol{u}_i$. (3) We partition the $M$ rotary blocks into local, global, mixed, and skip bands. (4) Each block accrues the gated phase $\theta_i^{(m)}$ of Equation 8, initialized to reproduce RoPE. (5) We rotate the blocks by $e^{\mathbf{i}\theta_i^{(m)}}$ while skip blocks bypass position. (6) The logit becomes the relative hierarchical kernel of Theorem 1. (7) Training adds warm-anchor and band-sparsity regularizers. Green marks chunk and global, blue offset and local, orange structure, and gray skip.

context (Wang et al., 2025). Rescaling extensions stretch that schedule further (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024), yet a single coordinate still performs every job.

We introduce Hierarchically Anchored Rotary Phases (HARP), which replaces the one monolithic phase wheel with anchored hierarchical coordinates (see Figure 1). HARP factors the raw index into a coarse chunk index, a within-chunk offset, and optional ordered structure coordinates such as section, scope, or nesting depth. Each coordinate governs its own band of learnable rotary phases through per-band gates, so dedicated frequencies handle each relation instead of competing inside one schedule, while a fraction of channels stay unrotated as a content bypass. We report two forms of the encoding. HARP-Core is the general form that exposes the clean theory, whereas HARP-SB is the separated-band form that routes each band into a local-only or global-only specialist and that we describe in our phase-budget analysis later. We initialize both forms exactly on the RoPE submanifold, where the phase collapses to $\omega_m i$ and reproduces RoPE token for token.

The factored geometry amounts to more than a reparameterization. The attention logit depends on differences in chunk index, offset, and structure rather than the raw gap alone (Equation 7), which strictly enlarges the positional kernel class. Separating the bands reshapes the phase budget, because local bands then read only the bounded offset and global bands only the chunk index. The worst-case phase for length $L$ therefore grows as $O(\sqrt{L})$ rather than the $O(L)$ of RoPE and reaches its minimum near $B \approx \sqrt{L}$ (Proposition 2), which cuts kernel-level BF16 drift by 2 to $9.7\times$ once the bands specialize. Our experiments separate structural gains from length and precision gains. On structural tasks the trained HARP kernel separates equal-distance token pairs that every purely relative encoding maps to identical logits, by $0.52$ logits for HARP-SB against $9.9 \times 10^{-6}$ for RoPE. The same kernel holds $0.755$ of its attention mass inside the current chunk against $0.649$ for RoPE, and all three bands carry non-redundant information, since dropping any one of them raises perplexity from $85.4$ to at least $95.6$. HARP-Core raises scope resolution from $4.7\%$ to $42.5\%$, while HARP-SB nearly doubles two-hop retrieval over RoPE and reaches $39.2$ perplexity at 160M against the $41.3$ of RoPE for $112$ extra parameters and an unchanged key-value cache (Appendix C.5). On length and precision at scale the advantage narrows to parity at 420M, since training specializes only the global-target band and the trained per-band phase still grows linearly in the context length $L$.

**The contributions are as follows: 1)** A positional encoding that factors the raw index into anchored hierarchical coordinates with per-band gated phases and a content bypass, in a general and a separated-band form. **2)** HARP induces a relative hierarchical attention kernel (Theorem 1) that strictly enlarges the raw-gap-only kernel class (Proposition 1) and bounds the separated-band phase budget at $O(\sqrt{L})$ (Proposition 2), alongside exact RoPE recovery (Corollary 1) and a finite-precision bound (Proposition 4). **3)** An initialization that starts exactly at RoPE, with warm-anchor and band-sparsity regularizers that let heads specialize only when specialization lowers the training loss. And **4)** a comprehensive evaluation spanning sanity checks, four synthetic tasks, pretraining at 160M and 420M parameters against three rotary and one additive baseline, a BF16 analysis, and ten mechanism analyses.

# 2 METHOD

We develop HARP from the standard RoPE kernel, and the following subsections describe them in detail. We first recast RoPE as a single-coordinate Fourier kernel, then factor the raw index into anchored hierarchical coordinates and derive the attention kernel that those coordinates induce. The separated-band form HARP-SB follows, together with the phase budget that its band structure permits and the training recipe that holds the model at RoPE until the data rewards a departure. Throughout we consider a single attention head and suppress its index.

## 2.1 PRELIMINARIES ON THE ROPE KERNEL

A head has dimension $d_h$, which we split into a rotated portion of dimension $2M$ and an unrotated portion of dimension $d_{\text{skip}}$, so that $d_h = 2M + d_{\text{skip}}$. Standard RoPE sets $d_{\text{skip}} = 0$. The rotated portion pairs into $M$ complex blocks, so the pre-rotation query and key blocks are $q_i^{(m)}, k_j^{(m)} \in \mathbb{C}$ for $m = 1, \ldots, M$, where $i$ and $j$ index the query and key tokens. Standard RoPE multiplies block $m$ at token $i$ by $e^{\mathbf{i}\omega_m i}$, where $\omega_m = \text{base}^{-2m/(2M)}$ is the base frequency of block $m$ and $\text{base} = 10000$. We reserve $d$ for the model dimension, which Table 4 reports. The blockwise contribution to the attention logit is therefore

$$\text{Re}\Big(q_i^{(m)}\,\overline{k_j^{(m)}}\,e^{\mathbf{i}\omega_m(i-j)}\Big), \tag{1}$$

where $\text{Re}(\cdot)$ denotes the real part and $\overline{(\cdot)}$ the complex conjugate. Each block of RoPE therefore reduces to a single complex exponential in the gap $(i - j)$, and the sum over $m$ forms a one-dimensional Fourier family over that one coordinate. This construction explains both the relative-position behavior of RoPE and the pressure that forces one coordinate to encode local order, long-range distance, and structure at once.

## 2.2 ANCHORED HIERARCHICAL COORDINATES

**Coordinates:** We fix a chunk size $B$. For each token $i$, the anchored hierarchical coordinates of token $i$ take the form

$$c_i = \lfloor i/B \rfloor, \quad r_i = i \bmod B, \quad \boldsymbol{u}_i \in \mathbb{Z}^S, \tag{2}$$

where $c_i$ is the coarse chunk index, $r_i \in \{0, \ldots, B-1\}$ is the within-chunk offset, and $\boldsymbol{u}_i$ collects $S$ optional ordered structure coordinates such as section, paragraph, or nesting depth. Equation 2 reconstructs the index exactly as $i = Bc_i + r_i$, so these coordinates do not discard the raw distance, which we recover as

$$i - j = B\,\Delta c_{ij} + \Delta r_{ij}, \tag{3}$$

with $\Delta c_{ij} = c_i - c_j$ and $\Delta r_{ij} = r_i - r_j$. The hierarchy therefore refactors the raw gap into a coarse and a fine part rather than discarding the gap.

**HARP-Core Phase and Rotation:** We call the unrotated part of the head the skip part. For each rotated block $m$, the general HARP-Core phase is

$$\theta_i^{(m)} = \alpha_m\, c_i + \beta_m\, r_i + \boldsymbol{\gamma}_m^\top \boldsymbol{u}_i, \tag{4}$$

where $\alpha_m, \beta_m \in \mathbb{R}$ are the per-block coarse and fine frequencies and $\boldsymbol{\gamma}_m \in \mathbb{R}^S$ are the structure weights. We initialize Equation 4 on the RoPE submanifold with $\alpha_m = B\omega_m$, $\beta_m = \omega_m$, and $\boldsymbol{\gamma}_m = \mathbf{0}$, so that $\theta_i^{(m)} = \omega_m(Bc_i + r_i) = \omega_m i$ reproduces RoPE exactly at step 0. We apply the rotation in complex form,

$$\widetilde{q}_i^{(m)} = e^{\mathbf{i}\theta_i^{(m)}}\, q_i^{(m)}, \qquad \widetilde{k}_j^{(m)} = e^{\mathbf{i}\theta_j^{(m)}}\, k_j^{(m)}, \tag{5}$$

where $\widetilde{q}_i^{(m)}$ and $\widetilde{k}_j^{(m)}$ are the rotated query block at token $i$ and the rotated key block at token $j$. The full attention logit combines the rotated and skip contributions,

$$\ell_{ij} = \frac{1}{\sqrt{d_h}}\Big[\sum_{m=1}^{M} \text{Re}\big(\widetilde{q}_i^{(m)}\overline{\widetilde{k}_j^{(m)}}\big) + \langle q_i^{\text{skip}}, k_j^{\text{skip}}\rangle\Big], \tag{6}$$

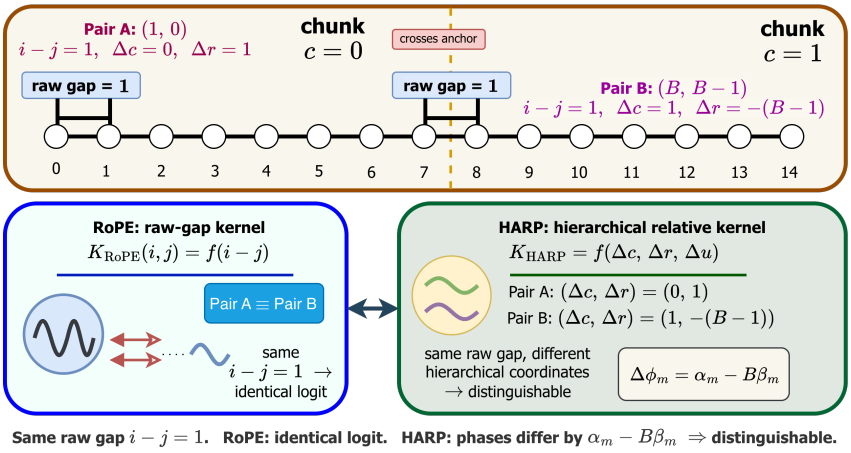



Figure 2: Same-gap discrimination for Proposition 1, where $P_1$ stays inside a chunk and $P_2$ crosses a boundary at the same raw gap.

where $q_i^{\text{skip}}, k_j^{\text{skip}} \in \mathbb{R}^{d_{\text{skip}}}$ are the unrotated channels and $\langle \cdot, \cdot \rangle$ is the real inner product. Equation 6 is the form we actually compute, and expanding the rotated terms yields the central identity behind the whole method.

**Theorem 1** (Relative hierarchical kernel). *For the phase of Equation 4, the attention logit of Equation 6 depends on position only through the relative hierarchical coordinates* $(\Delta c_{ij}, \Delta r_{ij}, \Delta \boldsymbol{u}_{ij})$*, namely*

$$\ell_{ij} = \frac{1}{\sqrt{d_h}} \Big[ \sum_{m=1}^{M} \operatorname{Re}\big( q_i^{(m)} \overline{k_j^{(m)}}\, e^{\mathbf{i}(\alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij} + \boldsymbol{\gamma}_m^\top \Delta \boldsymbol{u}_{ij})} \big) + \langle q_i^{\text{skip}}, k_j^{\text{skip}} \rangle \Big], \tag{7}$$

*where* $\Delta \boldsymbol{u}_{ij} = \boldsymbol{u}_i - \boldsymbol{u}_j$.

**Corollary 1** (RoPE submanifold). *If* $\alpha_m = B\beta_m$ *and* $\boldsymbol{\gamma}_m = \mathbf{0}$ *for every block, then every rotated block of HARP reduces exactly to the RoPE kernel of Equation 1 at base frequency* $\beta_m$*, and a head with* $d_{\text{skip}} = 0$ *reproduces that RoPE model exactly. In the separated-band parameterization of Equation 8 this condition reads* $g_{m,g} = g_{m,l}$*, and the choice* $g_{m,g} = g_{m,l} = 1$ *recovers standard RoPE at its original frequencies* $\omega_m$*.*

Theorem 1 demonstrates that anchoring preserves the relative-position property that makes RoPE effective, and Corollary 1 places RoPE inside the HARP family exactly rather than approximately, which in turn underpins the RoPE-anchored initialization of Section 2.5. Appendix A restates and proves every result of this section.

### 2.3 Geometry and Expressivity

The kernel of Equation 7 forms a Fourier family over the product grid $(\Delta c_{ij}, \Delta r_{ij}, \Delta \boldsymbol{u}_{ij})$ rather than over the single line $(i - j)$. Two consequences follow from this difference, which we state here and prove in Appendix A.

**Proposition 1** (Strict expressivity gain). *Let* $B \geq 2$*. If* $\alpha_m \not\equiv B\beta_m \pmod{2\pi}$ *for some block* $m$*, or* $\boldsymbol{\gamma}_m \neq \mathbf{0}$ *on data where* $\Delta \boldsymbol{u}_{ij}$ *varies independently of* $(i - j)$*, then the block kernel is not a function of the raw gap* $(i - j)$ *alone, because some choice of query and key content assigns different values to two index pairs of equal raw gap. The HARP kernel class therefore strictly contains the raw-gap-only class of RoPE.*

Figure 2 draws the construction behind the proof. The in-chunk pair $(1, 0)$ and the boundary-crossing pair $(B, B - 1)$ share the raw gap $i - j = 1$, so RoPE assigns both pairs the same positional factor, whereas the HARP phases differ by $\alpha_m - B\beta_m$. Section 4 measures the resulting separation on the trained checkpoints themselves.

Exact raw-token shift-equivariance holds only on a measure-zero set of the parameters. The positional kernel stays invariant under every common shift $(i, j) \mapsto (i + t, j + t)$ if and only if $\alpha_m \equiv B\beta_m \pmod{2\pi}$ for all $m$ with no structure terms (Proposition 3), a congruence that the RoPE submanifold of Corollary 1 satisfies with the trivial residue. Breaking that congruence is the intended inductive bias, because HARP may then treat "one token away within a chunk" differently from "one token away across a chunk boundary". Documents, code, and multi-document inputs all carry real boundaries, so exact shift invariance is rarely the property such inputs reward. One important note applies to the content bypass. The skip term of Equation 6 stays phase-free only at the current layer, since earlier layers may still write position into those channels, and the bypass therefore affords a content-only path within a layer rather than a global guarantee.

### 2.4 HARP-SB AND SEPARATED BANDS

HARP-SB is the separated-band specialization of HARP-Core in which each block carries two learnable gates that scale the coarse and fine RoPE frequencies,

$$\theta_i^{(m)} = g_{m,g}\, B\omega_m\, c_i + g_{m,l}\, \omega_m\, r_i + \boldsymbol{\gamma}_m^\top \boldsymbol{u}_i, \tag{8}$$

where $g_{m,g}$ is the global gate and $g_{m,l}$ the local gate. Equation 8 instantiates Equation 4 with $\alpha_m = g_{m,g}B\omega_m$ and $\beta_m = g_{m,l}\omega_m$, so Theorem 1 applies unchanged. The gates start at $g_{m,g} = g_{m,l} = 1$ with $\boldsymbol{\gamma}_m = \mathbf{0}$, which reproduces RoPE at step 0. We partition the $M$ blocks into four groups that differ only in their training-time regularization targets (Section 2.5), namely $50\%$ local-target with $g_{m,g} \to 0$, $25\%$ global-target with $g_{m,l} \to 0$, $12.5\%$ mixed with no specialization pressure, and $12.5\%$ skip with no rotation. We place high-frequency $\omega_m$ in local-target blocks, which resolve fast within-chunk offsets, and low-frequency $\omega_m$ in global-target blocks, which resolve the slower inter-chunk relations.

**Proposition 2** (Separated-band phase budget)**.** *Fix a target length $L = BC$ with $C$ chunks. Suppose local-only bands carry frequencies $|\beta_m| \le \Omega_r$ and global-only bands carry frequencies $|\alpha_m| \le \Omega_c$, with $\Omega_r$ and $\Omega_c$ held fixed independently of $B$ and $L$. Then the separated-band phase budget $\Phi(B)$, the largest absolute phase that any single band must represent, obeys the bound*

$$\Phi(B) \;\le\; f(B) \;:=\; \Omega_r B + \Omega_c \frac{L}{B}, \qquad \min_{B>0} f(B) \;=\; 2\sqrt{\Omega_r \Omega_c L}, \tag{9}$$

*and the bound attains its minimum at $B^\star = \sqrt{(\Omega_c/\Omega_r)L}$. With comparable budgets, $B \approx \sqrt{L}$ yields $O(\sqrt{L})$ phase growth against the $O(L)$ growth of RoPE.*

The hypothesis that $\Omega_c$ does not scale with $B$ describes the specialized configuration rather than the initialization. At the RoPE initialization the global coefficient equals $\alpha_m = B\omega_m$, which grows with $B$ and makes $\Omega_c L/B = \omega_m L$ linear in $L$ once more, exactly as Corollary 1 requires. The $O(\sqrt{L})$ regime therefore demands global bands with $O(1)$ coefficients and local bands that have shed their chunk term, a configuration that training does not reach (Section 4). Our frequency assignment further holds $\Omega_r$ far above $\Omega_c$, which places $B^\star$ below $\sqrt{L}$. A finite-precision bound (Proposition 4) then controls the phase-induced kernel error through the angle perturbation. Perturbing the implemented phases by $\varepsilon_i^{(m)}$ changes the logit by at most $\frac{1}{\sqrt{d_h}} \sum_m |q_i^{(m)}||k_j^{(m)}|(|\varepsilon_i^{(m)}| + |\varepsilon_j^{(m)}|)$. When the BF16 angle error scales with phase magnitude, the $O(\sqrt{L})$ budget of Equation 9 therefore implies a smaller logit drift for the separated-band design than the $O(L)$ budget of RoPE allows. Table 6 in Appendix B.2 contrasts the two variants.

### 2.5 TRAINING AND BAND SPECIALIZATION

We train HARP with the standard LM loss plus three position-specific regularizers,

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{a}}(t)\, \mathcal{R}_{\mathrm{anchor}} + \lambda_{\mathrm{sp}}\, \mathcal{R}_{\mathrm{sp}} + \lambda_\gamma \|\boldsymbol{\gamma}\|^2. \tag{10}$$

The term $\mathcal{R}_{\mathrm{anchor}}$ pulls the gates back to the RoPE submanifold, taking the form $\frac{1}{M}\sum_m[(g_{m,g} - 1)^2 + (g_{m,l} - 1)^2] + \|\boldsymbol{\gamma}\|^2$ for HARP-SB and $\frac{1}{M}\sum_m(\alpha_m - B\beta_m)^2$ for HARP-Core. The term $\mathcal{R}_{\mathrm{sp}}$

drives local-target blocks toward $g_{m,g} \to 0$ and global-target blocks toward $g_{m,l} \to 0$ to encourage specialization, and the last term mildly penalizes the structure weights. The anchor weight $\lambda_{\mathrm{a}}(t)$ follows a warm-anchor schedule that holds its full value for the first $T_a$ steps and then decays linearly to 0 over the next $T_d$ steps, with defaults $\lambda_{\mathrm{a}} = 1.0$, $\lambda_{\mathrm{sp}} = 2{\times}10^{-4}$, $\lambda_{\gamma} = 10^{-5}$, and $T_a, T_d$ at $2\%$ and $8\%$ of total steps. The optimizer therefore leaves the strong RoPE solution only gradually. Together with Corollary 1, this schedule leaves the RoPE solution only where specialization lowers the LM loss by more than the anchor penalty charges for that departure. Algorithm 1 in Appendix B.1 states the complete HARP-SB forward pass, which applies the rotation of Equation 5 in input precision, so that BF16 inputs exercise the bounded-budget arithmetic of Propositions 2 and 4.

## 3 Experimental Setup

**Setup and Protocol:** Each experiment isolates one prediction of Section 2 under an identical recipe. Our comparison of record contrasts HARP-SB with RoPE, since the exact submanifold initialization (Corollary 1) isolates the factorization. We report the unconstrained HARP-Core in every pretraining comparison, and every pretraining run faces two further rotary baselines, partial RoPE (p-RoPE) (Barbero et al., 2025) and the extrapolatable position embedding (xPos) (Sun et al., 2023), together with the additive relative bias of the Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020). Five deterministic tests verify the algebra of Section 2 without training, and four synthetic probes cover raw-gap, boundary, scope, and two-hop retrieval, each trained per task under RoPE, HARP-Core, and HARP-SB. We pretrain 160M- and 420M-parameter decoder-only Transformers from scratch on $70\%$ FineWeb-Edu (Penedo et al., 2024) and $30\%$ SlimPajama (Soboleva et al., 2023) with the Llama-2 tokenizer (Touvron et al., 2023), and evaluate perplexity out to 131072 tokens. Two ablations sweep the chunk size and the band allocation, a comparison between BF16 and single-precision floating point (FP32) tests Propositions 2 and 4, and a qualitative analysis inspects per-example predictions (Appendix C.3). Every run executes on a single NVIDIA RTX A6000 GPU with 48 GB of memory, and Appendix B.2 states the full protocol and the model configurations.

**Research Questions:** Four research questions (RQ) organize Section 4. **RQ1** asks whether anchoring HARP-SB on the RoPE submanifold matches or surpasses RoPE on perplexity. **RQ2** asks whether the model exploits the richer hierarchical kernel by separating same-gap pairs and by improving hierarchical retrieval. **RQ3** asks whether the bands split into local and global roles that carry non-redundant information, and **RQ4** asks whether the $O(\sqrt{L})$ budget improves BF16 stability.

## 4 Results

We run the protocol of Section 3 at seed 42, and at three seeds $(42, 123, 456)$ for the 420M comparison. Unless noted, we report validation perplexity at the 4096-token training context, and Appendix C holds the additional tables that support this section.

### 4.1 Synthetic Mechanism Tasks (RQ2)

Table 1 reports exact-match accuracy on the four probes, each trained per task on 3200 examples. The two hierarchical tasks separate the methods, and a different HARP variant wins each one. On scope, which turns on chunk structure, HARP-Core reaches $42.5\%$ against the $4.7\%$ of RoPE, a $9\times$ gain and the clearest sign that the model exploits the anchored hierarchy. On two-hop lookup, HARP-SB reaches $18.6\%$ against $9.9\%$, with HARP-Core at $4.5\%$. The boundary task saturates near $100\%$ for all encodings and raw-gap stays near $1\%$, as we expect since no encoding exposes an absolute-distance readout. Neither variant wins both hierarchical probes, because HARP-SB reaches only $6.0\%$ on scope and HARP-Core falls below

Table 1: Exact-match accuracy on the four synthetic mechanism tasks.

| Method | raw-gap | bound. | scope | two-hop |
|---|---|---|---|---|
| RoPE | 1.0 | 99.8 | 4.7 | 9.9 |
| HARP-Core | 1.0 | **100.0** | **42.5** | 4.5 |
| HARP-SB | 0.9 | **100.0** | 6.0 | **18.6** |

RoPE on two-hop. The anchored hierarchy therefore helps structured retrieval without either parameterization capturing both kinds of structure at once.

### 4.2 Language-Model Pretraining (RQ1, RQ3)

Figure 3 summarizes the 420M comparison against all four baselines, while Table 7 in Appendix C.1 holds the 160M results. At 160M and an equal 3500-step budget, HARP-SB reaches the best rotary perplexity at **39.2** against the 41.3 of RoPE and the 41.4 of xPos, while HARP-Core at 41.1 matches RoPE. The per-step progression shows what the RoPE-anchored initialization delivers. HARP-SB trails by 5.4 perplexity at step 500 and by 0.6 at step 2500, then crosses ahead at step 3000 and ends 2.1 ahead (Table 7). The anchor weight reaches 0 by step 350. The early deficit therefore measures the cost of learning the gates rather than the pull of the regularizer, and matching RoPE at convergence costs only a small and closing transient. At 420M over three seeds, HARP-Core at $\mathbf{61.3 \pm 3.6}$, HARP-SB at $61.7 \pm 3.2$, and RoPE at $62.1 \pm 3.3$ reach parity with xPos at $61.7 \pm 4.2$, ahead of p-RoPE at $66.5 \pm 0.9$ and T5 at 66.8. The two HARP variants trade places, since HARP-Core wins the 420M mean and five of the nine mechanism metrics (Table 16), whereas HARP-SB wins at 160M and on two-hop retrieval. The separated bands make the phase budget analyzable (Proposition 2), while the unconstrained HARP-Core attains the lower perplexity at 420M.

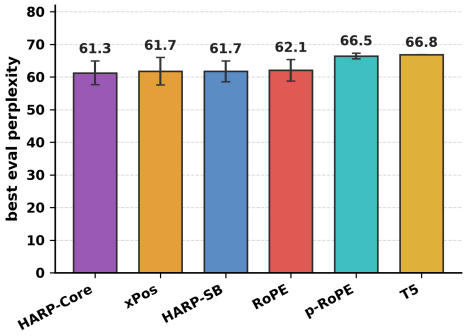



Figure 3: Best validation perplexity at 420M over three seeds.

Table 2: Learned per-band gate means for 420M HARP-SB.

| Band | $g_{m,g}$ | $g_{m,l}$ |
|---|---|---|
| Local-target | 1.03 | 0.99 |
| Mixed | 1.01 | 0.95 |
| Global-target | 1.03 | **0.32** |

The learned gates specialize partially and asymmetrically (**RQ3**, Table 2). Averaged over the 420M HARP-SB blocks, global-target bands drive their local gate to $g_{m,l} = 0.32$ from 1.0 while holding $g_{m,g} \approx 1.03$, which matches the intended global specialist. The local-target bands do not mirror that behavior, since both of their gates stay within 3% of the RoPE initialization, with mixed bands in between. Only one of the two intended specialists therefore emerges at the default $\lambda_{\text{sp}}$. The band-drop test below finds the local band the most load-bearing of the three, which suggests that its gates stay near RoPE because that setting already suits within-chunk structure. HARP-Core, which no gate constrains, departs much further, with per-block deviations $\delta_m = \alpha_m - B\beta_m$ of root-mean-square magnitude 2.51 radians and an anchor penalty of 6.29, even though the block means of $\alpha_m$ and $B\beta_m$ nearly coincide and leave the departures sign-balanced.

### 4.3 Length Generalization (RQ1)

Evaluating the 420M models out to 131072 tokens (Table 11, Appendix C.4), HARP-SB and RoPE coincide at the 4096-token training length at 58.1 and 58.0. Beyond that length HARP-SB degrades faster, reaching 85.4 against 72.7 at 16384, before the two converge at 131072. The factorization therefore does not help raw extrapolation, as we expect of a structure prior whose chunk grid follows the training length, and a deployment that needs length alone is better served by an encoding built for extrapolation than by HARP.

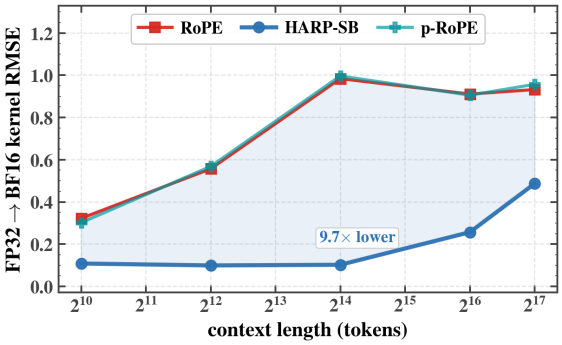



Figure 4: Kernel-level BF16 phase drift as root mean square error (RMSE) against context length.

Table 3: HARP-SB ablations over 500-step runs, sweeping the chunk size $B$ on the left and the band allocation on the right. The sweeps are separate runs, and their shared default therefore differs between the halves.

| Chunk $B$ | Train loss | Perplexity |
|---|---|---|
| 128 | 5.211 | 193.0 |
| 256 | 5.069 | 167.6 |
| 512 | 5.045 | 164.0 |
| 1024 | **5.042** | **163.5** |

| Band split | Train loss | Perplexity |
|---|---|---|
| local-heavy 75/12.5/6.25/6.25 | **5.025** | **161.2** |
| global-heavy 25/50/12.5/12.5 | 5.079 | 169.9 |
| default 50/25/12.5/12.5 | 5.091 | 171.6 |
| all-mixed | 5.108 | 174.3 |

### 4.4 BF16 Numerical Stability (RQ4)

With the fully specialized gates that the design targets, the separated-band partition sharply reduces phase-induced BF16 error. This analysis concerns the rotary encodings alone, since an additive bias accrues no phase. Figure 4 depicts the HARP-SB drift RMSE from FP32 to BF16 staying 2 to 9.7× below RoPE up to 131072 tokens, which matches Propositions 2 and 4, while p-RoPE tracks RoPE. The trained models do not inherit that advantage. Re-running the same measurement with the learned parameters narrows the gap to 1.059 against 1.070 (Appendix C.8), exactly as the gate table predicts, because the local-target bands hold $g_{m,g} \approx 1.03$ rather than approaching 0 and therefore still accrue a phase linear in $L$. At the full-network forward pass HARP-SB falls marginally behind RoPE at every length, by 6% on logit drift and by 11 to 13% on the attention-map Kullback-Leibler (KL) divergence (Appendix C.6). The conditioning gain thus belongs to the specialized configuration rather than to the trained model, and forcing that configuration carries a cost of its own. Overriding the gates at inference raises perplexity at every length, to 59.2, 92.1, and 115.7 against 58.1, 85.4, and 104.1 for the learned gates (Table 15). Bounding the phase budget and minimizing the loss therefore pull in opposite directions, and raising $\lambda_{\text{sp}}$ moves HARP-SB along that trade-off rather than removing the trade-off itself.

### 4.5 Mechanism Analyses (RQ2, RQ3)

Three further results on the 420M corroborate that the model uses the factorization as intended, with details in Appendix C. **(i) Same-gap disambiguation (RQ2):** The trained HARP kernels stay translation-variant at fixed distance, whereas every baseline stays invariant, to round-off for the rotary encodings and exactly for the additive one, which realizes Proposition 1 (Table 8). The separation therefore holds against xPos and the additive T5 bias as well. Two probes agree in direction and differ only in scale, since the kernel probe of Appendix C.2 yields a mean $|\Delta|$ of 0.026 for HARP-SB and 0.032 for HARP-Core against $6 \times 10^{-7}$ for RoPE, while the logit-level sweep yields 0.517 and 0.317 against $9.9 \times 10^{-6}$. **(ii) Band non-redundancy (RQ3):** Dropping any one band type of the trained HARP-SB model degrades perplexity, so all three carry distinct information. **(iii) Chunk locality (RQ2):** HARP-SB concentrates more attention mass inside the current chunk than RoPE, at 0.755 against 0.649, with HARP-Core at 0.833. The post-boundary perplexity penalty agrees in sign at 0.076 against 0.079, and separates the encodings only weakly.

### 4.6 Ablations (RQ3)

Two design-knob sweeps and one band-drop test qualify the defaults (Appendix C.7). Table 3 sweeps chunk size and band allocation. Perplexity falls monotonically in chunk size with no interior optimum, so the sweep never locates $B^\star$ (Proposition 2), and $B = 256$ concedes 2.5% against $B = 1024$. The default band split places third of four, behind local-heavy 75/12.5/6.25/6.25 (161.2 against 171.6). We keep both defaults so every run shares one configuration. Zeroing each band type at 16384 tokens (Figure 5) raises perplexity from 85.4 to 95.6, 98.4, and 117.6 without mixed, global, and local, though per-band phase magnitudes still grow linearly in $L$ (Table 14).

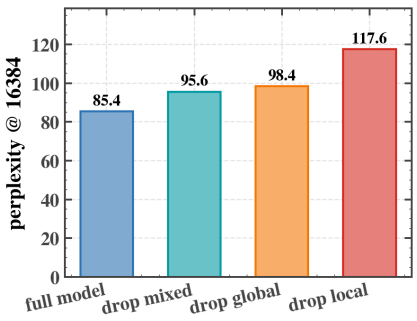



Figure 5: Band-drop ablation for 420M HARP-SB at 16384 tokens.

On the four questions of Section 3, HARP surpasses RoPE at 160M and matches it at 420M (**RQ1**), disambiguates same-gap pairs and improves hierarchical

retrieval (**RQ2**), carries non-redundant roles in all three bands while producing only the global specialist (**RQ3**), and lowers kernel-level BF16 drift only under gates that training does not attain (**RQ4**). The results therefore separate along one axis. Structural findings favor HARP, namely same-gap separation, chunk locality, band non-redundancy, chunk separability, and the better variant on each hierarchical probe. Findings that turn on length or precision at scale favor RoPE or leave the two encodings level, namely perplexity beyond the training window, the network-level BF16 comparison, and the drift measured with trained gates. The factorization therefore produces a positional kernel that encodes hierarchy, without improving the axis that long-context work usually optimizes.

## 5 Related Work

Early Transformers inject position through sinusoidal embeddings (Vaswani et al., 2017), relative schemes bias the logit by a function of the distance (Raffel et al., 2020; Press et al., 2022), and RoPE rotates each query and key block by an angle set by the raw token index (Su et al., 2024). Recent analyses demonstrate that low rotary frequencies carry semantic content while high ones carry position (Barbero et al., 2025), that dropping explicit position stays competitive (Kazemnejad et al., 2023), and that BF16 breaks the relative property at long context (Wang et al., 2025). Most long-context work keeps the raw index and rescales the frequency schedule (Chen et al., 2023; Peng et al., 2024; Xiong et al., 2024; Ding et al., 2024; Zhu et al., 2024), which leaves the single-coordinate structure and the phase budget it forces unchanged.

Two further lines resemble HARP more closely. One generalizes the rotary family itself through learnable commuting or Lie-algebraic generators (Yu et al., 2025; Ostmeier et al., 2025), Fourier or wavelet bases (Hua et al., 2025; Oka et al., 2025), token-dependent rates of rotation (Golovneva et al., 2024; Veisi et al., 2025b), and a split between token and hierarchy levels for code (Zhang et al., 2024a). The other conditions the signal on content, through additive biases read from the attention scores (Zheng et al., 2024a;b), token-specific penalties per head (Veisi et al., 2025a), and head-wise rescaling of position indices (Zhang et al., 2024c). No design in either line combines an anchored coordinate hierarchy, bands separated by role, an unrotated content bypass, and an exact RoPE-submanifold initialization, and an adaptive bias composes with our kernel rather than competing with it (Appendix D). Our baselines cover both families, namely p-RoPE (Barbero et al., 2025) and xPos (Sun et al., 2023) on the rotary side with the T5 relative bias (Raffel et al., 2020) on the additive side.

## 6 Conclusion

In this paper, we introduced anchored hierarchical coordinates in place of the single raw-index coordinate behind RoPE. Specifically, a coarse chunk index, a within-chunk offset, and optional structure coordinates each govern their own band of learnable rotary phases. The resulting attention kernel therefore depends on relative hierarchical coordinates rather than the raw gap alone. Moreover, the same kernel contains RoPE exactly at initialization, and its separated-band phase budget grows as $O(\sqrt{L})$ instead of $O(L)$. Across our experiments, the evidence divides along one axis. On structural tasks HARP separates equal-distance token pairs that every purely relative encoding maps to identical logits. Beyond that, the trained model concentrates attention inside the current chunk, carries non-redundant information in all three rotary bands, and HARP-Core raises scope resolution from $4.7\%$ to $42.5\%$. On length and precision at scale, by contrast, HARP degrades faster than RoPE beyond the training window and records a marginally worse full-network BF16 error. The clear $160$M perplexity win over the rotary baselines therefore narrows to parity at $420$M.

**Limitations:** Three primary limitations bound the experiments we report. First, training produces only one of the two intended band specialists, which leaves the conditioning gain that the phase budget predicts out of reach. Second, the structure channel stays empty in every pretraining run. Third, our largest budget covers $420$M parameters for $3500$ steps, with mechanism probes in place of end-task suites. Appendix D further describes all three in detail.

## REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

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

Egor Bogomolov, Aleksandra Eliseeva, Timur Galimzyanov, Evgeniy Glukhov, Anton Shapkin, Maria Tigina, Yaroslav Golubev, Alexander Kovrigin, Arie van Deursen, Maliheh Izadi, and Timofey Bryksin. Long code arena: a set of benchmarks for long-context code models, 2024. URL https://arxiv.org/abs/2406.11612.

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

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

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. 11091–11105. PMLR, 2024. URL https://proceedings.mlr.press/v235/ding24i.html.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important, 2024. URL https://arxiv.org/abs/2405.18719.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24932–24949. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hua25b.html.

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, pp. 24892–24928, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e85362c02172c0c6567ce593122d31c-Abstract-Conference.html.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012/.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, mar 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org.

Jia Li, Xuyuan Guo, Lei Li, Kechi Zhang, Ge Li, Jia Li, Zhengwei Tao, Fang Liu, Chongyang Tao, Yuqi Zhu, and Zhi Jin. Benchmarking long-context language models on long code understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27309–27327, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1324. URL https://aclanthology.org/2025.acl-long.1324/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Tianyang Liu, Canwen Xu, and Julian McAuley. Repobench: Benchmarking repository-level code auto-completion systems. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=pPjZIOuQuF.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of RoPE-based extrapolation. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=JO7k0SJ5V6.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1gs9JgRZ.

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

Sophie Ostmeier, Brian Axelrod, Maya Varma, Michael Moseley, Akshay S Chaudhari, and Curtis Langlotz. LieRE: Lie rotational positional encodings. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47339–47355. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ostmeier25a.html.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin A. Raffel, Leandro von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems 37*, pp. 30811–30849. Neural Information Processing Systems Foundation, Inc., 2024.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

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. URL https://openreview.net/forum?id=R8sQPpGCv0.

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

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023. URL https://huggingface.co/datasets/cerebras/SlimPajama-627B.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14590–14604, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.816. URL https://aclanthology.org/2023.acl-long.816/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ali Veisi, Hamidreza Amirzadeh, and Amir M. Mansourian. Context-aware biases for length extrapolation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30374–30395, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1545. URL https://aclanthology.org/2025.emnlp-main.1545/.

Ali Veisi, Delaram Fartoot, and Hamidreza Amirzadeh. Context-aware rotary position embedding. *arXiv preprint arXiv:2507.23083*, 2025b. URL https://arxiv.org/abs/2507.23083.

Haonan Wang, Qian Liu, Chao Du, Tongyao Zhu, Cunxiao Du, Kenji Kawaguchi, and Tianyu Pang. When precision meets position: Bfloat16 breaks down rope in long-context training. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=gwXfZ3xkUq.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=293V3bJbmE.

Hao Yu, Tangyu Jiang, Shuning Jia, Shannan Yan, Shunning Liu, Haolong Qian, Guanghao Li, Shuting Dong, and Chun Yuan. Comrope: Scalable and robust rotary position embedding parameterized by trainable commuting angle matrices. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4508–4517, 2025. doi: 10.1109/CVPR52733.2025.00458.

Kechi Zhang, Ge Li, Huangzhao Zhang, and Zhi Jin. HiRoPE: Length extrapolation for code models using hierarchical position. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13615–13627, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.735. URL https://aclanthology.org/2024.acl-long.735/.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. $\infty$Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.814. URL https://aclanthology.org/2024.acl-long.814/.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 60755–60775, 2024c.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, and Yu Li. Dape: Data-adaptive positional encoding for length extrapolation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26659–26700, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f050fa9f0d898e3f265d515f50ae8f9-Abstract-Conference.html.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, and Yu Li. Dape: Data-adaptive positional encoding for length extrapolation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26659–26700, 2024b.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3Z1gxuAQrA.

# Supplementary Material


**A Proofs**
- A.1 Proofs of the Main Results
- A.2 Deterministic Sanity-Check Construction

**B Implementation and Experimental Protocol**
- B.1 Forward-Pass Algorithm
- B.2 Model Configurations and Evaluation Suite

**C Extended Experimental Results**
- C.1 Pretraining Progressions and All Seeds
- C.2 Band Specialization and Same-Gap Disambiguation
- C.3 Qualitative and Error Analysis
- C.4 Length Generalization
- C.5 Compute and Memory Overhead
- C.6 Network-Level BF16 Stability
- C.7 Ablations
- C.8 Mechanism-Analysis Suite

**D Limitations and Future Work**

**E AI Use Statement**

**F Ethics Statement**

**G Reproducibility Statement**


## A PROOFS

This appendix restates and proves the results of Section 2. We use the notation of the main text throughout. We consider one attention head and suppress its index, and we pair the rotated portion into $M$ complex blocks $q_i^{(m)}, k_j^{(m)} \in \mathbb{C}$. The HARP-Core phase of Equation 4 is $\theta_i^{(m)} = \alpha_m c_i + \beta_m r_i + \boldsymbol{\gamma}_m^\top \boldsymbol{u}_i$, and the hierarchical coordinates of Equation 2 are $c_i = \lfloor i/B \rfloor$ and $r_i = i \bmod B$, so that $i = Bc_i + r_i$ with $r_i \in \{0, \dots, B-1\}$. The relative coordinates are $\Delta c_{ij} = c_i - c_j$, $\Delta r_{ij} = r_i - r_j$, and $\Delta \boldsymbol{u}_{ij} = \boldsymbol{u}_i - \boldsymbol{u}_j$. These results are exact algebraic, expressivity, and conditioning facts about the positional kernel, and the protocol of Section 3 carries the empirical claims. Appendix A.2 then verifies each result numerically on random tensors.

### A.1 PROOFS OF THE MAIN RESULTS

**Theorem** (Relative hierarchical kernel, restated from Theorem 1)**.** *For the phase of Equation 4, the attention logit of Equation 6 depends on position only through the relative hierarchical coordinates* $(\Delta c_{ij}, \Delta r_{ij}, \Delta \boldsymbol{u}_{ij})$*, namely*

$$\ell_{ij} = \frac{1}{\sqrt{d_h}} \Big[ \sum_{m=1}^{M} \mathrm{Re}\big( q_i^{(m)} \overline{k_j^{(m)}}\, e^{\mathbf{i}(\alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij} + \boldsymbol{\gamma}_m^\top \Delta \boldsymbol{u}_{ij})} \big) + \langle q_i^{\mathrm{skip}}, k_j^{\mathrm{skip}} \rangle \Big].$$

*Proof.* Fix a rotated block $m$. The rotation of Equation 5 multiplies each block by the unit-modulus scalar $e^{\mathbf{i}\theta_i^{(m)}}$, so $\widetilde{q}_i^{(m)} = e^{\mathbf{i}\theta_i^{(m)}} q_i^{(m)}$ and $\widetilde{k}_j^{(m)} = e^{\mathbf{i}\theta_j^{(m)}} k_j^{(m)}$. The conjugate product expands as

$$\widetilde{q}_i^{(m)} \overline{\widetilde{k}_j^{(m)}} = e^{\mathbf{i}\theta_i^{(m)}} q_i^{(m)} \cdot \overline{e^{\mathbf{i}\theta_j^{(m)}} k_j^{(m)}} = q_i^{(m)} \overline{k_j^{(m)}}\, e^{\mathbf{i}(\theta_i^{(m)} - \theta_j^{(m)})},$$

since $\overline{e^{\mathbf{i}\theta_j^{(m)}}} = e^{-\mathbf{i}\theta_j^{(m)}}$. Substituting the phase of Equation 4 for both tokens, the phase difference separates linearly into the three coordinate channels,

$$\theta_i^{(m)} - \theta_j^{(m)} = \alpha_m (c_i - c_j) + \beta_m (r_i - r_j) + \boldsymbol{\gamma}_m^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) = \alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij} + \boldsymbol{\gamma}_m^\top \Delta \boldsymbol{u}_{ij},$$

which depends on $i, j$ only through the relative coordinates $(\Delta c_{ij}, \Delta r_{ij}, \Delta \boldsymbol{u}_{ij})$. Hence $\widetilde{q}_i^{(m)}\overline{\widetilde{k}_j^{(m)}} = q_i^{(m)}\overline{k_j^{(m)}}\, e^{\mathbf{i}(\alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij} + \boldsymbol{\gamma}_m^\top \Delta \boldsymbol{u}_{ij})}$. Taking real parts and summing over $m$ reproduces the rotated term of Equation 6. The skip inner product $\langle q_i^{\text{skip}}, k_j^{\text{skip}} \rangle$ carries no positional phase and is unchanged by the rotation, completing the identity. For the separated-band variant of Equation 8 the same computation applies with $\alpha_m = g_{m,g} B\omega_m$ and $\beta_m = g_{m,l}\omega_m$. □

**Corollary** (RoPE submanifold, restated from Corollary 1)**.** *If $\alpha_m = B\beta_m$ and $\boldsymbol{\gamma}_m = \mathbf{0}$ for every block, then every rotated block of HARP reduces exactly to the RoPE kernel of Equation 1 at base frequency $\beta_m$, and a head with $d_{\text{skip}} = 0$ reproduces that RoPE model exactly. In the separated-band parameterization the condition reads $g_{m,g} = g_{m,l}$, and $g_{m,g} = g_{m,l} = 1$ recovers standard RoPE at its original frequencies $\omega_m$.*

*Proof.* Assume $\alpha_m = B\beta_m$ and $\boldsymbol{\gamma}_m = \mathbf{0}$. Using the exact index decomposition $i = Bc_i + r_i$ of Equation 2, the phase of Equation 4 collapses to

$$\theta_i^{(m)} = \alpha_m c_i + \beta_m r_i + \mathbf{0} = B\beta_m\, c_i + \beta_m\, r_i = \beta_m(Bc_i + r_i) = \beta_m\, i.$$

Therefore $e^{\mathbf{i}\theta_i^{(m)}} = e^{\mathbf{i}\beta_m i}$, which is the standard RoPE rotation with base frequency $\beta_m$. Substituting into Theorem 1, the block phase difference becomes $\theta_i^{(m)} - \theta_j^{(m)} = \beta_m(i-j)$, so each rotated term reduces to $\mathrm{Re}(q_i^{(m)}\overline{k_j^{(m)}}\, e^{\mathbf{i}\beta_m(i-j)})$, which is exactly the RoPE kernel of Equation 1 at frequency $\beta_m$. For the separated-band form the condition $\alpha_m = B\beta_m$ reads $g_{m,g} B\omega_m = B g_{m,l}\omega_m$, that is $g_{m,g} = g_{m,l}$, and any common gate value $g$ yields $\theta_i^{(m)} = g\omega_m i$, a RoPE kernel at rescaled frequency $g\omega_m$. The particular choice $g_{m,g} = g_{m,l} = 1$ yields $\alpha_m = B\omega_m$ and $\beta_m = \omega_m$, so $\theta_i^{(m)} = \omega_m i$ recovers RoPE at its original base frequency $\omega_m$. With $d_{\text{skip}} = 0$ the equality is exact across the whole head, establishing that the RoPE solution lies on the HARP parameter manifold rather than being merely approximated. □

**Proposition** (Strict expressivity gain, restated from Proposition 1)**.** *Let $B \geq 2$. If $\alpha_m \not\equiv B\beta_m \pmod{2\pi}$ for some block $m$, or $\boldsymbol{\gamma}_m \neq \mathbf{0}$ on data where $\Delta \boldsymbol{u}_{ij}$ varies independently of $(i-j)$, then the block kernel is not a function of the raw gap $(i-j)$ alone, because some choice of query and key content assigns different values to two index pairs of equal raw gap. The HARP kernel class therefore strictly contains the raw-gap-only kernel class of RoPE.*

*Proof.* By Theorem 1, the positional content of block $m$ is carried by the phase factor $\phi_m(i,j) = e^{\mathbf{i}(\alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij} + \boldsymbol{\gamma}_m^\top \Delta \boldsymbol{u}_{ij})}$. Exhibiting two index pairs with the same raw gap $i-j$ and distinct $\phi_m$ therefore suffices, since any kernel that is a function of $(i-j)$ alone must assign equal values to pairs sharing that gap.

Consider the two pairs

$$(i,j) = (1,0) \qquad \text{and} \qquad (i,j) = (B, B-1),$$

both with raw gap $i - j = 1$. With $B \geq 2$ their hierarchical coordinates are

$$(1,0): \quad c_i = 0,\ r_i = 1,\ c_j = 0,\ r_j = 0 \ \Rightarrow\ (\Delta c_{ij}, \Delta r_{ij}) = (0,1),$$

$$(B, B-1): \quad c_i = 1,\ r_i = 0,\ c_j = 0,\ r_j = B-1 \ \Rightarrow\ (\Delta c_{ij}, \Delta r_{ij}) = (1, -(B-1)),$$

where the first pair stays within a chunk and the second crosses the chunk boundary. Taking $\boldsymbol{\gamma}_m = \mathbf{0}$ for this comparison, the two phase arguments are $\alpha_m \cdot 0 + \beta_m \cdot 1 = \beta_m$ and $\alpha_m \cdot 1 + \beta_m \cdot (-(B-1)) = \alpha_m - B\beta_m + \beta_m$, so their difference is

$$\big(\alpha_m - B\beta_m + \beta_m\big) - \beta_m = \alpha_m - B\beta_m.$$

If $\alpha_m \not\equiv B\beta_m \pmod{2\pi}$, then this difference is not an integer multiple of $2\pi$, so the two unit-modulus factors satisfy $\phi_m(1,0) \neq \phi_m(B, B-1)$. A content choice that always separates the two kernels is $q_i^{(m)} = e^{-\mathbf{i}\beta_m}$ and $k_j^{(m)} = 1$, which turns the block kernel into $\mathrm{Re}(e^{\mathbf{i}(\beta_m - \beta_m)}) = 1$ at the first pair and into $\mathrm{Re}(e^{\mathbf{i}(\alpha_m - B\beta_m)}) = \cos(\alpha_m - B\beta_m) < 1$ at the second, where the strict

inequality holds precisely because $\alpha_m - B\beta_m \not\equiv 0 \pmod{2\pi}$. The block kernel therefore takes distinct values at equal raw gap and cannot be any function of $(i-j)$ alone. For the structure channel, if $\boldsymbol{\gamma}_m \neq \mathbf{0}$ and the data admit two pairs with identical $(i-j)$ whose $\Delta \boldsymbol{u}_{ij}$ differs along a direction non-orthogonal to $\boldsymbol{\gamma}_m$, the same argument applies to the term $\boldsymbol{\gamma}_m^\top \Delta \boldsymbol{u}_{ij}$. Since RoPE, by Equation 1, realizes only kernels of the form $g(i-j)$, the strict inclusion follows. The block kernel collapses back to a raw-gap function precisely when $\alpha_m \equiv B\beta_m \pmod{2\pi}$ for all $m$ and $\boldsymbol{\gamma}_m = \mathbf{0}$, a congruence class that contains the RoPE submanifold of Corollary 1. □

Figure 2 in Section 2.3 draws this construction, and Section 4 measures the resulting separation on trained HARP checkpoints.

**Proposition 3** (Exact shift-equivariance condition)**.** *For the fixed-size chunk model with $B \geq 2$ and without structure terms, the positional kernel is invariant under every common token shift $(i,j) \mapsto (i+t, j+t)$, $t \in \mathbb{Z}$, if and only if $\alpha_m \equiv B\beta_m \pmod{2\pi}$ for all $m$. Exactly the RoPE submanifold of Corollary 1 and its $2\pi$ translates satisfy the congruence that this condition states.*

*Proof.* **Sufficiency.** Suppose $\alpha_m = B\beta_m + 2\pi n_m$ for integers $n_m$ and $\boldsymbol{\gamma}_m = \mathbf{0}$. The block phase difference of Theorem 1 becomes

$$\Theta_m = \alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij} = \beta_m (B\Delta c_{ij} + \Delta r_{ij}) + 2\pi n_m \Delta c_{ij} = \beta_m (i-j) + 2\pi n_m \Delta c_{ij}$$

by Equation 3, and $2\pi n_m \Delta c_{ij}$ vanishes from $e^{\mathbf{i}\Theta_m}$ because $\Delta c_{ij}$ is an integer. The kernel therefore depends on $i, j$ only through the gap $i-j$. Since $(i+t)-(j+t) = i-j$ for every shift $t$, the kernel is invariant under all common shifts.

**Necessity.** Suppose the kernel is invariant under all common shifts. The phase difference of block $m$ is, by Theorem 1, $\Theta_m(i,j) = \alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij}$, and invariance for every choice of $q_i^{(m)}, k_j^{(m)}$ forces $\Theta_m(i+t, j+t) \equiv \Theta_m(i,j) \pmod{2\pi}$ for all $t$, because we may zero every block except block $m$ and the map $z \mapsto \mathrm{Re}(z\, e^{\mathbf{i}\Theta})$ on $z \in \mathbb{C}$ determines the unit-modulus factor $e^{\mathbf{i}\Theta}$. Take the in-chunk pair $(i,j) = (1,0)$, with $(\Delta c_{ij}, \Delta r_{ij}) = (0,1)$ and hence $\Theta_m(1,0) = \beta_m$, and shift it by $t = B-1$ to obtain $(i+t, j+t) = (B, B-1)$. This shifted pair crosses a chunk boundary, so the floor operation produces a carry and $(\Delta c_{ij}, \Delta r_{ij}) = (1, -(B-1))$ rather than $(0,1)$, and

$$\Theta_m(B, B-1) = \alpha_m \cdot 1 + \beta_m \cdot (-(B-1)) = \alpha_m - B\beta_m + \beta_m.$$

Shift-invariance requires $\Theta_m(B, B-1) \equiv \Theta_m(1,0) \pmod{2\pi}$, i.e.

$$\alpha_m - B\beta_m + \beta_m \equiv \beta_m \pmod{2\pi} \quad \Longleftrightarrow \quad \alpha_m - B\beta_m \equiv 0 \pmod{2\pi}.$$

As the chosen block was arbitrary, $\alpha_m \equiv B\beta_m \pmod{2\pi}$ for all $m$. The carry across the chunk boundary is precisely what prevents $(c_i, r_i)$ from transforming linearly under shifts away from this submanifold, so any structure term ($\boldsymbol{\gamma}_m \neq \mathbf{0}$ with $\Delta \boldsymbol{u}_{ij}$ not shift-compatible) breaks invariance as well. Exact raw-token shift-equivariance therefore holds exactly on the congruence class $\alpha_m \equiv B\beta_m \pmod{2\pi}$, which contains the RoPE submanifold. Breaking the congruence is the intended inductive bias, which lets the model treat a relation inside one chunk differently from a relation that crosses a chunk boundary. □

**Proposition** (Separated-band phase budget, restated from Proposition 2)**.** *Fix a target length $L = BC$ with $C$ chunks. Suppose local-only bands carry frequencies $|\beta_m| \leq \Omega_r$ and global-only bands carry frequencies $|\alpha_m| \leq \Omega_c$, with $\Omega_r$ and $\Omega_c$ held fixed independently of $B$ and $L$. Then the separated-band phase budget $\Phi(B)$, the largest absolute phase that any single band must represent, obeys the bound*

$$\Phi(B) \;\leq\; f(B) \;:=\; \Omega_r B + \Omega_c \frac{L}{B}, \qquad \min_{B>0} f(B) \;=\; 2\sqrt{\Omega_r \Omega_c L},$$

*the minimum of the bound being attained at $B^\star = \sqrt{(\Omega_c/\Omega_r)L}$. With comparable budgets, $B \approx \sqrt{L}$ therefore yields $O(\sqrt{L})$ phase growth against the $O(L)$ growth that the RoPE schedule itself would incur.*

*Proof.* In the separated-band design of Equation 8 each block is assigned to a single coordinate. A local-only block has phase $\theta_i^{(m)} = \beta_m r_i$ with $|\beta_m| \le \Omega_r$ and $r_i \in \{0, \dots, B-1\}$, so its phase magnitude is bounded by

$$\max_i |\theta_i^{(m)}| \le \Omega_r(B-1) \le \Omega_r B.$$

A global-only block has phase $\theta_i^{(m)} = \alpha_m c_i$ with $|\alpha_m| \le \Omega_c$ and $c_i \in \{0, \dots, C-1\}$, so, using $C = L/B$,

$$\max_i |\theta_i^{(m)}| \le \Omega_c(C-1) \le \Omega_c\, C = \Omega_c \frac{L}{B}.$$

Defining the phase budget $\Phi(B)$ as the largest phase magnitude any single band must represent, the two band maxima yield the upper bound

$$\Phi(B) \le \Omega_r B + \Omega_c \frac{L}{B} =: f(B),$$

where the sum (rather than the maximum) is taken as a conservative budget across the two band types. The function $f(B) = \Omega_r B + \Omega_c L/B$ is a sum of a positive term and its reciprocal scaled by the constant $\Omega_c L$. By the arithmetic-geometric-mean inequality, for any $B > 0$,

$$\Omega_r B + \Omega_c \frac{L}{B} \ge 2\sqrt{\Omega_r B \cdot \Omega_c \frac{L}{B}} = 2\sqrt{\Omega_r \Omega_c L},$$

with equality if and only if the two terms are equal, i.e. $\Omega_r B = \Omega_c L/B$, which solves to

$$B^\star = \sqrt{\frac{\Omega_c}{\Omega_r}}\, L.$$

At the optimum $f(B^\star) = 2\sqrt{\Omega_r \Omega_c L} = \Theta(\sqrt{L})$. When the two band budgets are comparable, $\Omega_c \approx \Omega_r$, the minimizer is $B^\star \approx \sqrt{L}$ and the budget is $O(\sqrt{L})$. By contrast a single RoPE block carries phase $\theta_i^{(m)} = \omega_m i$ with $\max_i |\theta_i^{(m)}| = \omega_m(L-1) = O(L)$ linear in the context length, since one coordinate spans the entire range. The separated-band factorization therefore reduces the worst-case per-band phase growth from $O(L)$ to $O(\sqrt{L})$, which motivates the chunk-size rule $B \approx \sqrt{L}$. □

The hypothesis that $\Omega_c$ is independent of $B$ is what makes the $O(\sqrt{L})$ regime a statement about the specialized configuration rather than about the initialization, and it is worth isolating. At the RoPE initialization of Equation 8 the global coefficient is $\alpha_m = g_{m,g} B \omega_m$ with $g_{m,g} = 1$, so $\Omega_c$ grows linearly in $B$ and the global term of $f(B)$ collapses to $\Omega_c L/B = \omega_m L$, linear in $L$ once more. The collapse reflects no defect of the bound and instead restates Corollary 1, since HARP-SB is RoPE at initialization and must therefore carry the RoPE budget. The bound bites only once training drives $g_{m,g} \to 0$ on local-target blocks, so that those blocks see the bounded offset alone, while the global-target blocks retain $O(1)$ coefficients. Section 4 and Table 14 report that the trained model does not reach that configuration, so the $O(\sqrt{L})$ budget describes a reachable point of the parameter space that our training recipe does not reach.

Figure 6 contrasts the two regimes. The RoPE phase budget grows linearly in the context length $L$, whereas the separated-band budget grows as $O(\sqrt{L})$ and reaches its minimum near $B \approx \sqrt{L}$.

**Proposition 4** (Finite-precision kernel bound)**.** *If the implemented phases are perturbed as* $\hat{\theta}_i^{(m)} = \theta_i^{(m)} + \varepsilon_i^{(m)}$*, then the blockwise kernel and the attention logit satisfy*

$$\left| e^{\mathbf{i}(\hat{\theta}_i^{(m)} - \hat{\theta}_j^{(m)})} - e^{\mathbf{i}(\theta_i^{(m)} - \theta_j^{(m)})} \right| \le |\varepsilon_i^{(m)} - \varepsilon_j^{(m)}|,$$

$$|\hat{\ell_{ij}} - \ell_{ij}| \le \frac{1}{\sqrt{d_h}} \sum_{m=1}^{M} |q_i^{(m)}|\, |k_j^{(m)}| \left( |\varepsilon_i^{(m)}| + |\varepsilon_j^{(m)}| \right).$$

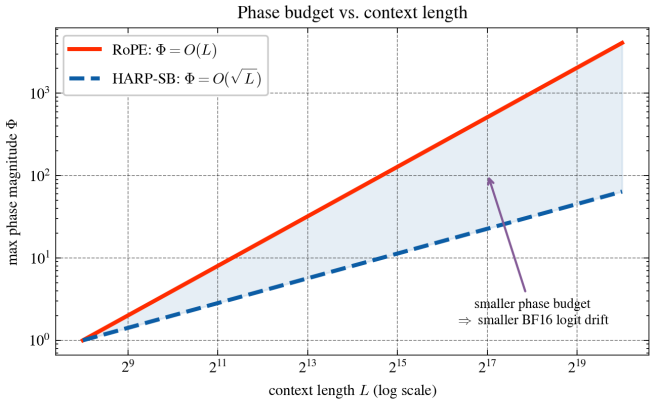



Figure 6: Schematic phase budget for Propositions 2 and 4, showing the two asymptotic regimes rather than any measured values.

*Proof.* For the first inequality, recall the elementary bound, valid for all reals $a, b$,

$$|e^{\mathbf{i}a} - e^{\mathbf{i}b}| = |e^{\mathbf{i}b}|\,|e^{\mathbf{i}(a-b)} - 1| = |e^{\mathbf{i}(a-b)} - 1| = 2|\sin\tfrac{a-b}{2}| \le |a-b|,$$

using $|e^{\mathbf{i}b}| = 1$ and $|\sin x| \le |x|$. Apply this with $a = \hat{\theta}_i^{(m)} - \hat{\theta}_j^{(m)}$ and $b = \theta_i^{(m)} - \theta_j^{(m)}$. Their difference then telescopes as follows,

$$a - b = (\hat{\theta}_i^{(m)} - \theta_i^{(m)}) - (\hat{\theta}_j^{(m)} - \theta_j^{(m)}) = \varepsilon_i^{(m)} - \varepsilon_j^{(m)},$$

so $|e^{\mathbf{i}(\hat{\theta}_i^{(m)} - \hat{\theta}_j^{(m)})} - e^{\mathbf{i}(\theta_i^{(m)} - \theta_j^{(m)})}| \le |\varepsilon_i^{(m)} - \varepsilon_j^{(m)}| \le |\varepsilon_i^{(m)}| + |\varepsilon_j^{(m)}|$ by the triangle inequality.

For the logit bound, write the perturbed and exact block kernels via Theorem 1. The skip term carries no phase and cancels in the difference $\hat{\ell_{ij}} - \ell_{ij}$, leaving only the rotated blocks. Using $|\operatorname{Re}(z) - \operatorname{Re}(w)| \le |z - w|$ and the factorization $z_m = q_i^{(m)}\overline{k_j^{(m)}}\, e^{\mathbf{i}(\theta_i^{(m)} - \theta_j^{(m)})}$ (and $\hat{z}_m$ with the perturbed phase),

$$|\hat{\ell_{ij}} - \ell_{ij}| \le \frac{1}{\sqrt{d_h}} \sum_{m=1}^{M} |\hat{z}_m - z_m| = \frac{1}{\sqrt{d_h}} \sum_{m=1}^{M} |q_i^{(m)}|\,|k_j^{(m)}|\,\left|e^{\mathbf{i}(\hat{\theta}_i^{(m)} - \hat{\theta}_j^{(m)})} - e^{\mathbf{i}(\theta_i^{(m)} - \theta_j^{(m)})}\right|,$$

where the amplitude $|q_i^{(m)}\overline{k_j^{(m)}}| = |q_i^{(m)}|\,|k_j^{(m)}|$ is unaffected by the unit-modulus rotation. Substituting the first inequality and bounding $|\varepsilon_i^{(m)} - \varepsilon_j^{(m)}| \le |\varepsilon_i^{(m)}| + |\varepsilon_j^{(m)}|$ then yields the bound we state above.

The bound controls only the phase-induced component of the error, since it quantifies how angle perturbations propagate through the rotation rather than amplitude rounding of $q^{(m)}, k^{(m)}$ or other numerical error. Under a floating-point model in which the BF16 angle error scales with phase magnitude, $|\varepsilon_i^{(m)}| \lesssim u\,|\theta_i^{(m)}|$ for a machine-dependent constant $u$, the per-block perturbation is controlled by the phase magnitude, so the $O(\sqrt{L})$ budget of Proposition 2 yields a correspondingly smaller worst-case logit drift than the $O(L)$ budget that RoPE carries. □

### A.2 Deterministic Sanity-Check Construction

We verify each result above numerically with a deterministic test that requires no training and runs on random tensors. These tests correspond to the released sanity-check suite that Section 3 references. We describe the constructions below.

**Closed-Form Equality (Theorem 1):** For random $q, k \in \mathbb{C}^M$, random parameters $\alpha_m, \beta_m, \gamma_m$, and random integer positions $i, j$, we compute the logit in two independent ways. The first applies direct rotation, forming $\widetilde{q}_i^{(m)} = e^{\mathbf{i}\theta_i^{(m)}} q_i^{(m)}$ and $\widetilde{k}_j^{(m)} = e^{\mathbf{i}\theta_j^{(m)}} k_j^{(m)}$ and then Equation 6. The second

applies the closed form of Equation 7, namely $\mathrm{Re}(q_i^{(m)}\overline{k_j^{(m)}}\, e^{\mathbf{i}(\alpha_m \Delta c_{ij} + \beta_m \Delta r_{ij} + \boldsymbol{\gamma}_m^\top \Delta \boldsymbol{u}_{ij})})$. Over many trials we require the maximum absolute discrepancy between the two routes to sit at the level of floating-point round-off, which confirms the identity.

**RoPE Special Case (Corollary 1):** With $\alpha_m = B\omega_m$, $\beta_m = \omega_m$, and $\boldsymbol{\gamma}_m = \mathbf{0}$, we compare the HARP rotation of a random vector against the reference RoPE rotation $e^{\mathbf{i}\omega_m i}$ from the standard cosine and sine formula. Agreement to round-off confirms exact containment.

**Same-Gap Counterexample (Proposition 1):** We set the parameters with $\alpha_m \neq B\beta_m$ and compare the logits of the two equal-gap pairs $(1, 0)$ and $(B, B-1)$. A nonzero difference, which stems from the per-block phase offset $\alpha_m - B\beta_m$ of the proof, certifies that the kernel separates pairs that no raw-gap-only kernel can distinguish.

**Shift-Equivariance (Proposition 3):** On the RoPE submanifold, where $\alpha_m = B\beta_m$ and $\boldsymbol{\gamma}_m = \mathbf{0}$, we check that the logits of $(i, j)$ and of the shifted pair $(i + t, j + t)$ coincide to round-off. Off the submanifold the same comparison shows a nonzero discrepancy, which matches the necessity argument across the chunk boundary.

**BF16 Phase Drift (Propositions 2 and 4):** For fixed random $q, k$, we evaluate the attention kernel in FP32 and in BF16 as a function of context length for RoPE and for HARP-SB, and we record the drift between the two precisions. We set the HARP-SB gates to their specialized targets here, with $g_{m,g} = 0$ on local-target blocks and $g_{m,l} = 0$ on global-target blocks, so that no block accrues a phase linear in $L$. The separated-band budget then predicts slower growth of this drift with length for HARP-SB than for RoPE, which matches the phase-induced bound of Proposition 4. Table 16 repeats the same measurement with the gates that training reaches.

All five checks pass to floating-point tolerance, which confirms the algebra of Section 2 on random tensors. The rotated logit matches the closed form to $8.9 \times 10^{-6}$, the RoPE submanifold recovers to $1.2 \times 10^{-4}$, the same-gap pair $(1, 0)$ versus $(B, B-1)$ receives a logit gap of $0.58$ that no raw-gap kernel can produce, a common shift moves HARP-Core by $0.43$ while leaving RoPE invariant, and HARP-SB with fully specialized gates holds its BF16 drift $9.7\times$ below RoPE.

# B Implementation and Experimental Protocol

This appendix records how we implement HARP and how we run every experiment. Appendix B.1 states the HARP-SB forward pass, and Appendix B.2 lists the model configurations, the evaluation suite, the two contributed variants, and the protocol detail that the main text compresses.

## B.1 Forward-Pass Algorithm

Algorithm 1 states the HARP-SB forward pass referenced in Section 2.5. Per token, it forms the hierarchical coordinates of Equation 2, accumulates the gated coarse and fine phases of Equation 8 per block, and applies the complex rotation of Equation 5 to the rotated blocks while passing the skip blocks through unchanged. We compute all phases in the same precision as the inputs.

## B.2 Model Configurations and Evaluation Suite

Table 4 lists the two from-scratch pretraining configurations and Table 5 the controlled evaluation suite, both referenced in Section 3. We report below the detail that the main text compresses for reasons of space.

**Why Not an End-Task Suite:** End-task suites such as RULER, LongBench, HELMET, and $\infty$Bench (Hsieh et al., 2024; Bai et al., 2024; Yen et al., 2025; Zhang et al., 2024b), or RepoBench, LongCodeU, and Long Code Arena (Liu et al., 2024b; Li et al., 2025; Bogomolov et al., 2024), measure aggregate accuracy without isolating which positional mechanism produces a difference. Each experiment we run instead targets one prediction, which is what lets a result implicate the factorization rather than the training recipe. The cost of that choice is external validity, and Appendix D records it.

**Algorithm 1:** HARP-SB phase computation and rotation

**Input:** block of query or key activations $x \in \mathbb{R}^{n \times d}$, positions $\{i\}$, structure coords $\{\boldsymbol{u}_i\}$, chunk size $B$, per-block frequencies $\{\omega_m\}$, gates $\{g_{m,g}, g_{m,l}\}$, weights $\{\boldsymbol{\gamma}_m\}$, and the rotated and skip block index sets.
**Output:** rotated activations $\widetilde{x}$.
**for** *each token* $i$ **do**
  $c_i \leftarrow \lfloor i/B \rfloor, \quad r_i \leftarrow i \bmod B$ `// hierarchical coords, Eq. 2`
  **for** *each rotated block* $m$ **do**
    $\theta_i^{(m)} \leftarrow g_{m,g}\, B\omega_m\, c_i + g_{m,l}\, \omega_m\, r_i + \boldsymbol{\gamma}_m^\top \boldsymbol{u}_i$ `// gated phase, Eq. 8`
    $\widetilde{x}_i^{(m)} \leftarrow e^{\mathbf{i}\theta_i^{(m)}}\, x_i^{(m)}$ `// complex rotation, Eq. 5`
  **foreach** *skip block* **do** $\widetilde{x}_i^{\text{skip}} \leftarrow x_i^{\text{skip}}$ `// content bypass`
**return** $\widetilde{x}$

Table 4: The two from-scratch pretraining configurations, which share the context length, optimizer, precision, and data mixture.

| | 160M | 420M |
|---|---|---|
| Layers | 16 | 28 |
| Model dimension $d$ | 768 | 1024 |
| Attention heads | 12 | 16 |
| Feed-forward width | 3072 | 4096 |
| Context length | 4096 | 4096 |
| Chunk size $B$ | 256 | 256 |
| Backbone learning rate | $3\times10^{-4}$ | $2.5\times10^{-4}$ |
| Positional learning rate | $1.5\times10^{-3}$ | $1.25\times10^{-3}$ |
| Optimizer steps | 3500 | 3500 ($\times$3 seeds) |

**Sanity Checks and Synthetic Tasks:** The five deterministic tests verify the kernel identity (Theorem 1), exact RoPE recovery (Corollary 1), the expressivity gain (Proposition 1), shift-equivariance (Proposition 3), and BF16 phase drift (Propositions 2 and 4). The first four run on the general HARP-Core phase, whose free $\alpha_m, \beta_m, \boldsymbol{\gamma}_m$ exercise the algebra in full, and the fifth runs on HARP-SB with its gates set to the specialized targets, so it measures the configuration the budget describes rather than the one training reaches. On the synthetic side we report exact-match accuracy, with the boundary check and the retrieval tasks targeting **RQ2**. Among the five deterministic tests, the drift check targets **RQ4**.

**Training and Metrics:** Pretraining uses AdamW (Loshchilov & Hutter, 2019), BF16 autocast in the mixed-precision training regime of Micikevicius et al. (2018), separate backbone and positional parameter groups, and the regularizer of Equation 10. The budget of Proposition 2 bounds representable phase rather than extrapolation quality, so we read the length sweep as a description of how each encoding behaves with length. At the network level we measure the attention-logit RMSE between BF16 and FP32 together with the attention-map KL divergence (Kullback & Leibler, 1951). Learned gate means, per-band drop tests, and submanifold deviations $\delta_m = \alpha_m - B\beta_m$ quantify how far heads depart from RoPE and whether the model uses each band. For the qualitative analysis, Appendix C.3 describes per-example records. Two of those records log model predictions rather than per-sequence losses, namely the same-gap sweep, which logs the logit of an interior and a boundary-crossing pair at equal raw distance, and needle retrieval at 16384 tokens, which follows the synthetic needle probes of Hsieh et al. (2024) and logs the gold and predicted token of every item.

## C Extended Experimental Results

We report here the full results behind Section 4, covering in turn the complete pretraining progressions and all seeds, the band specialization and same-gap probes, the qualitative error analysis, the length-generalization and context-extension sweeps, the compute and memory overhead, the network-level BF16 measurement, the ablations, and the mechanism analyses. We

Table 5: The controlled evaluation suite, in which each experiment isolates one prediction of Section 2 or one research question.

| Experiment | Purpose | Tests | RQ |
|---|---|---|---|
| Sanity checks (5 tests) | deterministic verification on random tensors, no training | Thm 1, Cor 1, Props 1–4 | 2, 4 |
| Synthetic tasks | mechanism retrieval over raw-gap, boundary, scope, two-hop | Prop 1 | 2 |
| 160M / 420M pretraining | short-context perplexity and perplexity-vs-length on long docs | Cor 1 | 1, 3 |
| BF16 stability | attention-logit RMSE and attention-map KL, RoPE vs HARP-SB, to 131072 | Props 2, 4 | 4 |
| Qualitative and error analysis | per-pair same-gap logits and needle gold-vs-predicted tokens | Prop 1 | 2 |
| Ablation of chunk size $B$ | $B \in \{128, 256, 512, 1024\}$, test the budget's $B^\star$ prediction | Prop 2 | 3 |
| Ablation of band fractions | default vs all-mixed, local-heavy, and global-heavy allocations | Sec 2.4 | 3 |

Table 6: The two contributed variants, which we both initialize exactly on the RoPE submanifold that Corollary 1 describes.

| | HARP-Core (Eq. 4) | HARP-SB (Eq. 8) |
|---|---|---|
| Per-block parameters | $\alpha_m, \beta_m, \boldsymbol{\gamma}_m$ | $g_{m,g}, g_{m,l}, \boldsymbol{\gamma}_m$ (gates on $B\omega_m, \omega_m$) |
| Initialization | $\alpha_m{=}B\omega_m,\ \beta_m{=}\omega_m,\ \boldsymbol{\gamma}_m{=}\mathbf{0}$ | $g_{m,g}{=}g_{m,l}{=}1,\ \boldsymbol{\gamma}_m{=}\mathbf{0}$ |
| Band structure | no local/global split, 12.5% skip | 50/25/12.5/12.5 local/global/mixed/skip |
| Primary purpose | clean relative-hierarchical theory | $O(\sqrt{L})$ budget, BF16 study, ablations |

measure every perplexity on the held-out mixture, and we take long-context numbers from held-out long documents at the stated length. Inference-based analyses cap the forward pass at 125 to 1000 sequences.

### C.1 Pretraining Progressions and All Seeds

Table 7 lists the 160M validation-perplexity progression for every encoding, and every run trains to 3500 steps. At the shared 3500-step checkpoint the ordering runs HARP-SB $<$ HARP-Core $\approx$ RoPE $\approx$ xPos $<$ T5 $<$ p-RoPE, so HARP-SB leads every other encoding at that budget. The tables that mark values place the best in **bold** and the second best underlined.

Table 7: Validation perplexity by step for 160M controlled pretraining at seed 42, where lower is better.

| Model | 500 | 1000 | 1500 | 2000 | 2500 | 3000 | 3500 |
|---|---|---|---|---|---|---|---|
| HARP-SB | 118.4 | 70.4 | 55.2 | 48.7 | 44.3 | 41.4 | **39.2** |
| HARP-Core | 117.0 | 69.5 | 54.9 | 47.5 | 43.6 | 41.9 | <u>41.1</u> |
| RoPE | 113.0 | 68.3 | 54.9 | 47.5 | 43.7 | 42.1 | 41.3 |
| xPos | 111.5 | 68.1 | 54.8 | 47.7 | 43.8 | 42.2 | 41.4 |
| T5 | 131.9 | 79.5 | 63.1 | 54.6 | 50.1 | 48.1 | 47.2 |
| p-RoPE | 127.4 | 78.7 | 63.4 | 55.4 | 51.0 | 49.1 | 48.1 |

We run the 420M comparison across three seeds $(42, 123, 456)$, and Figure 3 in Section 4 plots the best validation perplexity per method as mean$\pm$std. HARP-Core attains the lowest mean at $61.28 \pm 3.64$, while xPos at $61.73 \pm 4.21$, HARP-SB at $61.74 \pm 3.18$, and RoPE at $62.06 \pm 3.26$ match it inside overlapping standard deviations, and all four lead p-RoPE at $66.45 \pm 0.87$ and T5 at 66.85. The corresponding best validation losses run $4.114 \pm 0.060$, $4.120 \pm 0.069$, $4.122 \pm 0.052$, $4.127 \pm 0.053$, $4.196 \pm 0.013$, and 4.202 in the same order. Per seed, HARP-SB reaches 57.65, 62.18, and 65.39. A consistent seed ordering (seed $42 < 123 < 456$, a $\sim$13% spread) holds for every method, which motivates placing the multi-seed rung at this scale.

### C.2 Band Specialization and Same-Gap Disambiguation

Table 2 lists the learned per-band gates. The global-target band alone collapses its local gate to $0.32$ and becomes a long-range specialist, while local and mixed bands stay near the RoPE initialization. Turning to same-gap translation variance, the trained HARP kernels separate equal-distance pairs that straddle a chunk boundary from interior ones, while every relative baseline is invariant to round-off. We measure this effect with two probes that sweep differently, so their magnitudes are not directly comparable and we report both. The kernel probe over distances $\{8, 64\}$ at sequence length $768$ yields mean $|\Delta| = 0.026$ for HARP-SB and $0.032$ for HARP-Core, against $6.1 \times 10^{-7}$ for RoPE and $5.1 \times 10^{-7}$ for p-RoPE. Table 8 extends the logit-level sweep to xPos and T5, where the rotary xPos stays at round-off and the additive T5 bias scores exactly zero, since a bias defined on the raw distance assigns one value to every pair of equal gap by construction. The logit-level sweep of Table 8 (gaps $\{1, 4, 16, 64\}$, $250$ pairs each) yields $0.517$ and $0.317$ against $9.9 \times 10^{-6}$ and $8.8 \times 10^{-6}$. The separation is stable across gaps in both probes, and only its absolute scale differs.

Table 8: Same-gap translation variance by gap for 420M seed $42$ over $250$ pairs per gap. All rows come from this one sweep, and the additive T5 bias scores exactly zero because it depends on the raw distance alone.

| Method | gap 1 | gap 4 | gap 16 | gap 64 | overall |
|---|---|---|---|---|---|
| HARP-SB | 0.494 | 0.516 | 0.529 | 0.528 | 0.517 |
| HARP-Core | 0.319 | 0.306 | 0.318 | 0.324 | 0.317 |
| RoPE | $9.0 \times 10^{-6}$ | $1.1 \times 10^{-5}$ | $9.3 \times 10^{-6}$ | $1.0 \times 10^{-5}$ | $9.9 \times 10^{-6}$ |
| p-RoPE | $7.8 \times 10^{-6}$ | $1.0 \times 10^{-5}$ | $8.1 \times 10^{-6}$ | $9.1 \times 10^{-6}$ | $8.8 \times 10^{-6}$ |
| xPos | $8.7 \times 10^{-6}$ | $9.8 \times 10^{-6}$ | $7.9 \times 10^{-6}$ | $1.2 \times 10^{-5}$ | $9.5 \times 10^{-6}$ |
| T5 | 0 | 0 | 0 | 0 | 0 |

### C.3 Qualitative and Error Analysis

Two of our per-example records log predictions rather than losses, namely the same-gap logit sweep and needle retrieval at $16384$ tokens. We read both records item by item and report representative examples below.

**Where the Hierarchy Helps:** Table 9 shows six consecutive pairs from the sweep at raw gap $1$, where each row holds an interior pair and a boundary-crossing pair at equal raw distance, which any purely relative encoding must score identically. RoPE does exactly that, with a separation at the floating-point floor for all $250$ pairs, whereas HARP-SB separates $85\%$ of the pairs by more than $0.1$ logit and HARP-Core separates $82\%$, with medians of $0.42$ and $0.28$. The separation stays uneven, ranging from $0.062$ at $j = 264$ to $1.426$ at $j = 288$. The spread tracks query and key content rather than the interior offset, since the phase offset $\alpha_m - B\beta_m$ that separates the two pairs is the same at every interior index. These medians cover gap $1$ alone, whereas the analyses of Section 4.5 in the main text average over gaps $\{1, 4, 16, 64\}$.

**Where Every Encoding Fails:** Needle retrieval at $16384$ tokens stays near zero for all encodings, and the per-item predictions of Table 10 explain the outcome. Every one of the $1000$ items carries the same gold token, identifier $29871$, the SentencePiece (Kudo & Richardson, 2018) whitespace marker that precedes a digit string in the Llama-2 vocabulary (Touvron et al., 2023), so the probe as scored asks the model to emit a space rather than to retrieve the needle value. Each encoding answers with a frequent generic token instead, and HARP-SB in particular answers "a" on $702$ of the $1000$ items, "the" on $126$, and "available" on $91$. This failure belongs to the probe and to the $3500$-step training budget rather than to any encoding, so we read needle retrieval and the $512$-token two-hop trace as a bound on retrieval rather than as a ranking over the encodings that we compare.

### C.4 Length Generalization

Table 11 holds the 420M perplexity-versus-length sweep behind Section 4, computed over $1000$ held-out sequences per length. All four rotary encodings degrade beyond the $4096$ training length. HARP-SB and RoPE coincide at the training length and converge again at $131072$, with RoPE degrading more slowly at the intermediate lengths and p-RoPE strongest among them at the extreme.

Table 9: Per-pair logits at raw gap 1 from the released same-gap sweep, where the query sits at $j+1$ and $|\Delta|$ separates the two pairs.

| Key index $j$ | | HARP-SB logit | | | HARP-Core | RoPE |
|---|---|---|---|---|---|---|
| interior | boundary | interior | boundary | $\lvert\Delta\rvert$ | $\lvert\Delta\rvert$ | $\lvert\Delta\rvert$ |
| 256 | 511 | 0.083 | 0.464 | 0.381 | 0.687 | $1.3 \times 10^{-6}$ |
| 264 | 511 | −0.127 | −0.189 | 0.062 | 0.100 | $4.1 \times 10^{-6}$ |
| 272 | 511 | −0.413 | −0.335 | 0.078 | 0.105 | $8.8 \times 10^{-6}$ |
| 280 | 511 | −2.197 | −1.596 | 0.601 | 0.407 | $5.0 \times 10^{-6}$ |
| 288 | 511 | 1.472 | 0.046 | 1.426 | 0.383 | $9.8 \times 10^{-6}$ |
| 296 | 511 | −0.936 | 0.254 | 1.190 | 0.261 | $7.7 \times 10^{-6}$ |

Table 10: Needle-retrieval predictions at 16384 tokens, drawn from the released per-example files. Depth places the needle in the context, and the gold token is the SentencePiece whitespace marker for every item.

| Depth | Needle value | Gold token | RoPE | HARP-Core | HARP-SB |
|---|---|---|---|---|---|
| 0.1 | 72234 | ▁ (space) | “the” | “a” | “the” |
| 0.3 | 60272 | ▁ (space) | “a” | “a” | “a” |
| 0.5 | 82853 | ▁ (space) | “a” | “a” | “available” |
| 0.7 | 95659 | ▁ (space) | “available” | “available” | “available” |
| 0.9 | 13617 | ▁ (space) | “a” | “a” | “a” |

Table 11: Perplexity versus evaluation context length for 420M models at seed 42, trained at 4096 tokens and measured on held-out long documents rather than the validation mixture of Section 4.2.

| Model | 4096 | 8192 | 16384 | 32768 | 65536 | 131072 |
|---|---|---|---|---|---|---|
| RoPE | <u>58.0</u> | **63.3** | **72.7** | **86.0** | <u>100.9</u> | 117.9 |
| HARP-SB | 58.1 | 70.7 | 85.4 | 95.8 | 104.1 | <u>117.1</u> |
| HARP-Core | **56.9** | <u>70.3</u> | 86.8 | 98.8 | 108.8 | 120.5 |
| p-RoPE | 66.1 | 71.8 | <u>81.0</u> | <u>89.4</u> | **99.7** | **111.5** |

Because HARP is a structure prior rather than an extrapolation method, the standard RoPE rescaling recipes apply unchanged to a HARP checkpoint. Applying them to the frozen 160M RoPE checkpoint at inference, YaRN (Peng et al., 2024) holds perplexity at 50.7 at 131072 where untouched RoPE reaches 155.9, Position Interpolation (Chen et al., 2023) collapses to 166.9, and dynamic neural tangent kernel (NTK) scaling (Peng et al., 2024) lands in between at 81.7. This ordering matches expectation and indicates that long-context deployment of a HARP model stays a separate step, which the existing rescaling recipes already address.

### C.5 Compute and Memory Overhead

HARP changes the phase that each rotary block accrues rather than the shape of the attention computation, so the overhead stays small and easy to account for. One rotary module serves every layer in our implementation, and the module holds $\alpha_m$, $\beta_m$, and, for HARP-SB, the two gates per rotated block. A head carries 32 complex blocks at both scales, of which 28 rotate under the default allocation. HARP-SB therefore adds 112 scalars at four per block and HARP-Core adds 56 at two, a relative increase below $3 \times 10^{-7}$ at 420M. The key-value cache keeps the layout of RoPE, since the unrotated skip channels occupy the same head dimension that a RoPE model rotates, and no extra tensor enters the cache.

Wall-clock cost is where the current implementation pays. Table 12 lists the end-to-end training time of the 420M runs at seed 42, each on one NVIDIA RTX A6000 GPU with 48 GB of memory under the identical configuration of Table 4. HARP-Core costs 6% more than RoPE and HARP-SB costs 31% more, since our phase computation gathers per-band frequencies and gates for every token instead of indexing one cached table of sine and cosine values as the RoPE path does. The gap reflects an unfused research implementation rather than the arithmetic of Equation 8, which performs the same complex multiplication per block that RoPE performs, and a fused kernel that

precomputes the per-band phase table would remove most of it. We report the measurement rather than the projection.

Table 12: End-to-end training time for the 420M seed-42 runs, each on one NVIDIA RTX A6000 GPU with 48 GB of memory under one configuration and one data mixture.

| Encoding | Training time (h) | Relative to RoPE |
|---|---|---|
| p-RoPE | 5.68 | 1.00× |
| RoPE | 5.69 | 1.00× |
| HARP-Core | 6.01 | 1.06× |
| xPos | 6.71 | 1.18× |
| HARP-SB | 7.43 | 1.31× |
| T5 | 24.84 | 4.37× |

### C.6 Network-Level BF16 Stability

Table 13 reports the full-network forward-pass error between FP32 and BF16, as attention-logit RMSE and attention-map KL divergence over 50 batches. Unlike the kernel-level measurement of Figure 4, the two encodings behave comparably here, and the sign even reverses, because HARP-SB comes out marginally worse than RoPE at every length, by 5.9 to 6.1% on logit RMSE and by 11.4 to 12.8% on attention-map KL. At the network level, trained-weight rounding and other numerical effects dominate the phase term, so the kernel-level advantage does not carry into a forward-pass gain at all.

Table 13: Network-level BF16 stability for 420M seed 42, reported as FP32 against BF16 attention-logit RMSE and attention-map KL.

| | RoPE | | HARP-SB | |
|---|---|---|---|---|
| Length | logit RMSE | attn KL | logit RMSE | attn KL |
| 32768 | 0.0236 | 0.0358 | 0.0250 | 0.0404 |
| 65536 | 0.0188 | 0.0229 | 0.0199 | 0.0255 |
| 131072 | 0.0163 | 0.0172 | 0.0173 | 0.0193 |

### C.7 Ablations

**Chunk Size, Band Allocation, and Band-Drop Phase Magnitudes:** Table 3 and Figure 5 in Section 4 hold the chunk-size sweep, the band-allocation sweep, and the band-drop test. We record here the measurement that accompanies the band-drop run. Table 14 lists the largest absolute phase each band represents. Every band grows linearly in $L$, with the local band rising from 976 at $L = 1024$ to $9.74 \times 10^5$ at $L = 1048576$, a factor of 998 across a factor of 1024 in length. A local-only block would be bounded by $\Omega_r B$ and therefore constant in $L$, so the local band is plainly still carrying its chunk term. The measurement restates the gate result of Table 2 in quantitative form. Local-target blocks hold $g_{m,g} \approx 1.03$ instead of approaching 0, and $g_{m,g} B \omega_m c_i$ grows linearly in $L$, which predicts a local-band phase near $1.03\, B\, \omega_m\, L/B$ and matches the measurement to within the spread of $\omega_m$ across blocks. The trained model therefore operates at an $O(L)$ phase budget rather than the $O(\sqrt{L})$ budget of Proposition 2, which is the same shortfall that Section 4 already reports at the kernel level.

Table 14: Largest absolute per-band phase for trained 420M HARP-SB. Every band grows linearly in $L$, so the bounded budget is not realized.

| Band | $L$=1024 | $L$=4096 | $L$=16384 | $L$=1048576 |
|---|---|---|---|---|
| Local | 976 | 3829 | 15240 | $9.74 \times 10^5$ |
| Global | 3.9 | 15.4 | 61.3 | $3.9 \times 10^3$ |
| Mixed | 10.8 | 42.9 | 171.4 | $1.1 \times 10^4$ |

**Gate-Inference Ablation:** Overriding the learned gates at inference isolates what those gates contribute (Table 15). At the training length the learned gates are best (58.1 versus 59.9 for a forced RoPE-submanifold setting), confirming that the model uses the gates. The submanifold setting closes that gap as the evaluation length grows and turns marginally better at 65536 tokens, at 102.5 against 104.1, as we expect of learned structure that the model tunes to in-distribution lengths. Forced full specialization is worse throughout, at 59.2, 92.1, and 115.7. Figure 4 uses exactly that setting to obtain its 2 to 9.7× BF16 advantage, so this row prices the advantage in perplexity and shows the two objectives pulling against one another.

Table 15: Gate-inference ablation for 420M HARP-SB, reported as perplexity versus length under three gate settings.

| Gate setting | 4096 | 16384 | 65536 |
|---|---|---|---|
| learned | **58.1** | **85.4** | 104.1 |
| RoPE-submanifold | 59.9 | 85.6 | **102.5** |
| specialized | 59.2 | 92.1 | 115.7 |

### C.8 Mechanism-Analysis Suite

Ten mechanism analyses run on the 420M seed-42 checkpoints over 125 to 1000 sequences. Eight of them yield a metric comparable across encodings, and Table 16 collects those eight in nine rows, since per-position perplexity (#6) contributes two. The remaining two are internal to HARP, because gate trajectories (#2) and band specialization (#8) have no RoPE counterpart, so Table 2 and Table 14 report them instead. Those two analyses, together with same-gap disambiguation (#3), run deterministically and carry the clearest evidence, so Section 4 reports all three. Among the inference analyses, chunk-boundary attention and per-position perplexity agree in sign, since HARP-SB stays more chunk-local and pays a smaller post-boundary penalty than RoPE.

The content-bypass probe confirms that HARP routes a small nonzero share through the unrotated path, about 0.002 negative log-likelihood (NLL) against 0 for RoPE, which carries no bypass. Chunk separability, which we report as the effect size $\eta^2$ (Cohen, 1988), slightly favors the HARP variants. Two probes do not favor HARP. Length extrapolation (#7) ranks RoPE first, which matches the length results of Section 4, and the 512-token two-hop trace (#5) stays uninformative at this scale, sitting near zero for every encoding, which bounds rather than ranks. Needle retrieval at 16384 tokens, a probe outside the suite, hits the same floor, and Appendix C.3 inspects its predictions directly and traces that floor to the probe itself. Row #4 measures the same kernel-level drift as Figure 4 while substituting the trained positional parameters of the 420M checkpoints for the fully specialized gates that the design targets. The 2 to 9.7× advantage of Figure 4 then narrows to near parity, at 1.059 against 1.070, exactly as Table 2 predicts, because the local-target bands keep $g_{m,g} \approx 1.03$ rather than approaching 0, so those blocks still accrue a phase linear in $L$ and the bounded budget stays out of reach. Row #7 evaluates 125 sequences per length rather than the 1000 sequences of Table 11, so its perplexities run about 3 points lower, and the ranking over the encodings remains unchanged throughout that sweep.

Table 16: Mechanism-analysis suite for 420M seed 42, where **bold** marks the best encoding and underlined the second best.

| # | Metric | HARP-SB | HARP-Core | RoPE |
|---|---|---|---|---|
| 1 | same-chunk attn. mass ↑ | 0.755 | **0.833** | 0.649 |
| 3 | translation variance ↑ | $2.6 \times 10^{-2}$ | $\mathbf{3.2 \times 10^{-2}}$ | $6.1 \times 10^{-7}$ |
| 4 | BF16 phase-drift RMSE at 131072 ↓ | 1.059 | **1.018** | 1.070 |
| 5 | two-hop accuracy (512 tok) ↑ | 0.0 | 0.0 | 0.0 |
| 6 | mean NLL ↓ | 4.055 | **4.035** | 4.053 |
| 6 | post-boundary penalty ↓ | **0.076** | 0.085 | 0.079 |
| 7 | perplexity at 16384 ↓ | 82.5 | 84.2 | **69.9** |
| 9 | bypass ΔNLL (nonzero = uses bypass) | 0.0026 | 0.0020 | 0.0 |
| 10 | chunk separability $\eta^2$ ↑ | 0.026 | **0.031** | 0.025 |

## D Limitations and Future Work

Three boundaries mark what the evidence in this paper covers. The first concerns the phase budget, since Proposition 2 describes a specialized configuration and the default regularization reaches only the global specialist, so the trained models deliver the structural gains of Section 4 rather than the conditioning gains that the budget predicts. The second concerns the structure channel, since every pretraining run sets $S = 0$ and the coordinates $\boldsymbol{u}_i$ therefore carry signal in the synthetic probes rather than on natural markup, a setting that Theorem 1 already covers and that our measurements have yet to reach. The third concerns breadth, since our largest run trains $420$M parameters for $3500$ steps under a controlled rather than a competitive budget, and we report perplexity and mechanism probes in place of end-task suites (Hsieh et al., 2024; Bai et al., 2024; Yen et al., 2025), which secures attribution at the cost of external validity.

Every boundary we identify opens a direction to the work. The gate objective invites the first investigation. An $\ell_1$ or group-lasso penalty, hard gates with a straight-through estimator, or a per-layer annealing schedule could reach the specialized configuration that the quadratic penalty leaves untouched, and the sparsity weight $\lambda_{\mathrm{sp}}$ awaits a sweep. Populating the structure channel deserves the second study. Future runs with $S > 0$ could draw ordered coordinates from nesting depth in structured documents, row and column indices in tables, and scope depth in source code. The implementation raises two further questions worth examining, because a single rotary module currently serves every layer. Per-layer modules would open layer-wise band allocations, and learned or punctuation-aware anchors would replace the fixed chunk grid, both of which merit a dedicated study. Future work could also explore the hierarchical kernel with an adaptive additive bias, apply the rescaling recipes of Appendix C.4 to a HARP checkpoint rather than to the RoPE checkpoint we tested, and run an end-task comparison at a competitive scale.

## E AI Use Statement

We used generative AI tools in this work only as general-purpose assistants, under our own supervision. We used them to polish the writing of text we had already drafted, covering grammar, phrasing, and formatting, and to assist with routine coding tasks such as plotting scripts and refactoring of experiment harnesses. We did not use them for research ideation, for deriving or checking the theoretical results, for generating experimental results or data, or for retrieving and summarizing related work. We reviewed all AI-assisted text and code, and every number we report comes from the released code that we ran ourselves.

## F Ethics Statement

We study positional encodings for transformer language models, and we introduce neither a new dataset nor a human-subjects component. All of our pretraining and evaluation draws on publicly available corpora under their original licenses, as Section 3 describes. Our contribution is architectural, so it inherits the well-known risks of the language models that carry it, which include the reproduction of biases present in the pretraining corpus and the potential for misuse of generated text, and it adds no risk specific to the method. No conflicts of interest, no privacy concerns, and no legal-compliance issues arose over the course of the research we report here.

## G Reproducibility Statement

Sections 2 and 3 specify all architectural definitions, the regularizer schedule, the hyperparameters, the data mixture, and the evaluation protocol, and Appendix B.2 records the protocol detail that the main text compresses. The deterministic sanity checks of Appendix A.2 run on random tensors without training, and Appendix A states the full proof of every theoretical claim. Every number we report comes from a run on a single NVIDIA RTX A6000 GPU with $48$ GB of memory. We release our code, configurations, and run commands upon acceptance of this paper.