BEYOND THE BASE: FREQUENCY ALLOCATION IN ROPE

Anonymous authors
Paper under double-blind review

ABSTRACT

In rotary position embeddings (RoPE), reallocating frequencies within a fixed range can substantially change how effectively a model uses context. Controlled interventions show that performance gains persist even when total log-frequency displacement is also matched. Analysis of complete sine–cosine pairs characterizes positional overlap; reassigning the same frequencies across coordinates changes task performance, demonstrating the role of learned frequency use. Viewing allocation as channel-wise distance scaling motivates TailSpline, a closed-form transition from native frequencies to the extended low-frequency tail. Derived from a discrete tail-connection problem, the construction uses only public RoPE parameters and retains standard rotary computation without weight updates or calibration. At fourfold extension to 32K, TailSpline improves Llama-3-8B's full thirteen-task RULER mean by over 11 percentage points against both YaRN and MrRoPE-Pro. It achieves the highest mean in every main frozen-extension RULER comparison across five model families up to 70B parameters, alongside improvements in language modeling and natural question answering. Allocation also improves native-window language modeling with frozen weights and strengthens extrapolation during training and adaptation. These findings establish internal frequency allocation as a distinct design dimension of RoPE that shapes how models use context within and beyond the native window.

1 INTRODUCTION

Rotary position embeddings (RoPE) encode relative position by rotating each pair of query and key coordinates at its own frequency (Su et al., 2024). A model pretrained with context length L has learned to use these rotations over separations up to L, and training-free extension to sL rewrites the frequencies. Base changes and Position Interpolation preserve the geometric arrangement of frequencies (Xiong et al., 2024; bloc97, 2023; Chen et al., 2023). NTK-aware scaling belongs to this class even though its multipliers vary across channels. YaRN, LongRoPE and MrRoPE change the normalized internal placement (Peng et al., 2024; Ding et al., 2024; Tian et al., 2026). MrRoPE compares profiles within a shared transition band. We identify allocation effects at fixed range, then match total frequency movement and examine learned coordinate use.

We first separate range from internal placement. Ordered from fast to slow, the log-frequencies of any RoPE table can be written as $x_k = a + Rz_k$ with $z_0 = 0$ and $z_{K-1} = 1$. The offset a and span R fix the frequency range; the normalized coordinates z_k, which we call the *allocation*, fix how the frequencies are placed within it. Every geometric table has the same allocation, whatever its base, so changing the base never changes z. At a given extension factor, the tables of NTK-aware scaling and YaRN even share the same range and differ only in allocation. Holding the range and the model weights fixed, we find that allocation alone can decide whether an extended model works: at the same range and attention scaling, OLMo-2-1B scores 0.6% on nine RULER tasks (Hsieh et al., 2024) at 16K with the geometric allocation of NTK-aware scaling and 61% with a coarse-ramp allocation.

Matching total log-frequency displacement still leaves differences in quality, so allocation is not a proxy for how far the frequencies move.

Complete-pair analysis separates shared positional structure from content discrimination: overlapping distance responses can still support distinct content comparisons. An exact kernel criterion separates

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.

changes to the frequency set from coordinate reassignment. These results motivate frozen-model allocations relative to the native table.

TailSpline puts this into practice as a public scaling rule. Like YaRN and MrRoPE, it keeps the fastest channels native and interpolates the slowest by the full factor s; it distributes the resulting extra log-frequency span across the transition band. Its log-stretch increments solve a one-sided tail-connection problem in closed form and shrink along the band, whereas those of MrRoPE's progressive profile, MrRoPE-Pro, grow. TailSpline therefore stretches every transition channel at least as much as MrRoPE-Pro and ends the band much closer to full interpolation (Fig. 1). The rule uses only public RoPE parameters, with no search or calibration.

Without any training, TailSpline achieves the highest mean on RULER's full thirteen-task benchmark (Full-13) in every main-comparison panel (Table 1), across five model families and scales up to 70B. On Llama-3-8B extended fourfold to 32K, it leads both YaRN and MrRoPE-Pro by more than 11 points. TailSpline also has the highest mean on 32K-native models extended to 128K, and RULER gains over MrRoPE-Pro persist at larger factors, including $16\times$ on Llama. Beyond RULER, TailSpline has the lowest extended-context perplexity of the three methods at $s = 4$ and higher means than MrRoPE-Pro in all eight main natural QA and retrieval-augmented comparisons (Table 2).

Our contributions are:

- **Allocation is a design dimension of RoPE separate from its range.** At a fixed range, allocation alone can turn a failing extension into a working one; fixed-range comparisons identify its effect in frozen and trained models, and frozen controls further match total displacement (§5.2, §5.4).
- **Positional overlap and content discrimination are distinct.** Complete-pair geometry separates positional overlap from content discrimination; an exact kernel criterion distinguishes spectral change from coordinate reassignment (§3, §5.2).
- **TailSpline leads YaRN and MrRoPE-Pro without training.** A closed-form rule from public RoPE parameters achieves the highest Full-13 mean in every main-comparison panel across five model families up to 70B (§4.1, §5.1).
- **Allocation improves native context use and learned extrapolation.** A reallocated native table (NCP) lets a frozen model use its earlier context better, and training with Cosh, which moves frequencies toward the fast end, improves extrapolation in pretraining and adaptation (§5.3, §5.4).

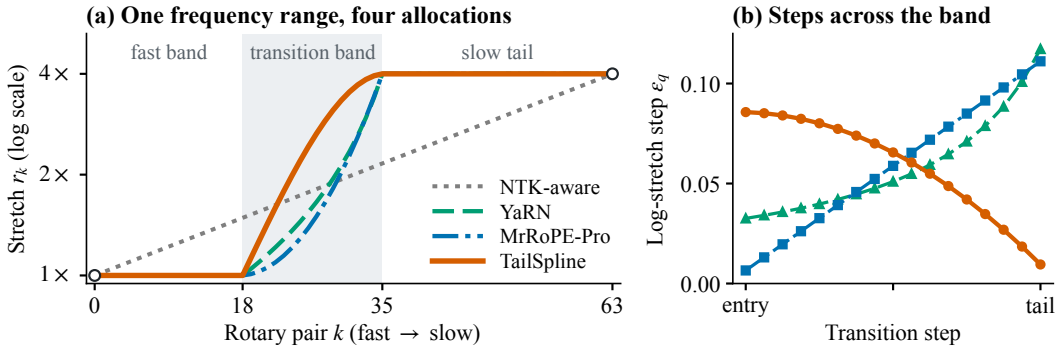


Figure 1: **Same frequency range, different allocations.** Stretch factors $r_k = \omega_k^N/\omega_k'$ of four $s = 4$ extensions of the Llama-3-8B table; a channel with stretch r_k produces at separation $r_k\Delta$ the phase it natively had at Δ. (a) All four tables keep the fastest channel native and divide the slowest by s, so they share one frequency range. NTK-aware base scaling keeps the geometric allocation (straight line on this log scale); YaRN, MrRoPE-Pro and TailSpline keep a fast band native and a slow tail fully interpolated, and differ only inside the shaded transition band. (b) Steps in log-stretch across that band: TailSpline's steps shrink toward the tail, whereas MrRoPE-Pro's and YaRN's grow.

2 FREQUENCY ALLOCATION IN ROPE

A RoPE head with $K = d_{\text{rot}}/2$ rotary pairs uses one frequency per pair; standard RoPE sets $\omega_k = b^{-k/K}$ with base b. Ordering a positive table from fast to slow, $\omega_0 > \cdots > \omega_{K-1}$, its log-frequencies decompose as

$$x_k = -\log \omega_k = a + Rz_k, \quad z_k = \frac{x_k - x_0}{x_{K-1} - x_0}, \quad z_0 = 0,\ z_{K-1} = 1. \tag{1}$$

The offset $a = x_0$ and span $R = x_{K-1} - x_0$ specify the frequency range (also called support); the $K - 2$ interior coordinates z_k form the allocation. Every geometric table has $z_k = k/(K-1)$ irrespective of base: Position Interpolation shifts its range and base scaling widens it, and both preserve the geometric allocation (Chen et al., 2023; bloc97, 2023).

Extension as channel-wise stretch. Native denotes a released model's original table ω^N. An extension to sL, $s > 1$, divides each native frequency by a stretch factor:

$$\omega'_k = \omega^N_k s^{-m_k}, \qquad r_k = \omega^N_k/\omega'_k = s^{m_k}. \tag{2}$$

Channel k reaches at separation $r_k\Delta$ the phase it natively had at Δ: the exponent $m_k = 0$ keeps it native and $m_k = 1$ interpolates it fully. With $m_0 = 0$ and $m_{K-1} = 1$ the endpoints are fixed, so all such tables share one frequency range and the exponents m determine the allocation z (Appendix B). NTK-aware scaling uses $m_k = k/(K-1)$ and keeps z geometric. YaRN-style tables instead keep a *fast band* native and fully interpolate a *slow tail*; with these outer bands fixed, they differ only inside the *transition band* between them (Fig. 1). We apply YaRN's attention gain $g = 1 + 0.1 \ln s$ separately, scaling each pair's query–key logit contribution by g^2. The underlying rotations preserve norms and RoPE's relative-position composition (Su et al., 2024).

Controlled comparisons. At a fixed range, total log-frequency displacement, $D = \sum_k \log(\omega^N_k/\omega'_k) = \log s \sum_k m_k$, is a statistic of the allocation. We use three nested controls. Tables with the same endpoints and gain test whether allocation matters; tables that also share D test whether its distribution matters; and permuting the same frequencies across coordinates, with the weights fixed, tests whether a model depends on which coordinate carries which frequency. Task quality is measured mainly by Full-13, the equal-weight mean of the thirteen official RULER task scores (Hsieh et al., 2024).

3 POSITIONAL OVERLAP AND LEARNED FREQUENCY USE

Why can two tables with the same range perform so differently? Overlapping distance responses can still carry distinct content comparisons. We make this distinction precise with complete-pair geometry, then separate changes to the frequency set from changes to its assignment across learned coordinates.

3.1 AN EXACT RELATION BETWEEN OVERLAP AND EFFECTIVE RANK

Complete rotary pairs give a phase-independent measure of overlap and an exact effective-rank identity. For a fixed content pair, one rotary channel contributes $C_1 \cos(\omega\Delta) + C_2 \sin(\omega\Delta)$ to an attention logit, where Δ is relative token separation. Both coefficients depend on content; the positional object is

$$V_\omega = \text{span}\{\cos(\omega\Delta), \sin(\omega\Delta)\}. \tag{3}$$

To avoid privileging a content phase, let $x_\omega(\Delta) = [\cos(\omega\Delta), \sin(\omega\Delta)]$ and define $S_\omega = \mathbb{E}[x_\omega^\top x_\omega]$, $H_{\omega\nu} = \mathbb{E}[x_\omega^\top x_\nu]$ under a declared separation measure. For nonsingular self-Grams, the singular values of $Q_{\omega\nu} = S_\omega^{-1/2} H_{\omega\nu} S_\nu^{-1/2}$ are the canonical correlations, or cosines of principal angles, of the two spaces (Björck & Golub, 1973). Their mean square,

$$c_{\omega\nu} = \tfrac{1}{2}\|Q_{\omega\nu}\|_F^2 \in [0, 1], \tag{4}$$

measures shared positional directions and is invariant to invertible basis changes within either pair. Whitening normalizes each pair before comparing its directions. Let Γ be the Gram matrix of the

concatenated whitened pairs, with diagonal blocks I_2. The average pairwise overlap $\bar{c}$ gives the exact Rényi-2 effective-rank identity

$$r_2(\Gamma) = \frac{(\operatorname{tr}\Gamma)^2}{\operatorname{tr}(\Gamma^2)} = \frac{2K}{1 + (K-1)\bar{c}}. \tag{5}$$

Here $\operatorname{tr}(\Gamma^2)$ is the frame potential (Benedetto & Fickus, 2003). The effective rank counts distinct positional directions: it equals $2K$ when all pairs are orthogonal and falls as they overlap, concentrating the same $2K$ coordinates into fewer directions.

3.2 Slow frequencies share a limiting positional subspace

Proposition 1 (Shared slow-frequency subspace)**.** *Under uniform measure on* $[0, L]$*, with* $t = \Delta/L$*, as* $x = \omega L \to 0$*,* $\cos(xt) \to 1$ *and* $\sin(xt)/x \to t$*, so* V_ω *converges to* $\operatorname{span}\{1, \Delta\}$*. For* $\epsilon = \max\{|\omega L|, |\nu L|\} \to 0$*, squared chordal separation satisfies* $2 - \|Q_{\omega\nu}\|_F^2 = \mathcal{O}(\epsilon^4)$.

Within a window, a slow channel's cosine is nearly constant and its sine nearly linear in Δ, so distinct slow frequencies approach one shared two-dimensional positional space. On the grid of the paired training runs in §5.4 ($b = 256$, $K = 32$, $L = 256$), the eight slowest pairs span $\omega L \in [1.19, 4]$ and have effective rank $r_2 = 2.11$: their overlap is substantial even beyond the small-phase limit (Appendix A.2 and its Fig. 4).

3.3 Shared positional structure, distinct content uses

Positional overlap need not erase content distinctions. For a block $\mathcal{B}$ of r slow rotary pairs, put $\eta = H \max_{k\in\mathcal{B}} |\omega_k|$. Encoding $2r$ symbols as distinct coordinate unit vectors gives an unscaled matching-score margin of at least $1 - 2\eta$ for all key separations $|d| \le H$ (Appendix A.3). Thus for $\eta < 1/2$ all symbols remain distinguishable, even as their positional functions approach $\operatorname{span}\{1, d\}$ when $\eta \to 0$, consistent with prior observations of positional and semantic frequency use (Barbero et al., 2025).

The same distinction appears in model quality. At Llama 32K, TailSpline's full-pair effective rank is 10.08 and MrRoPE-Pro's 10.21 under uniform separations, yet TailSpline scores 11.7 Full-13 points higher; the analytic equal-displacement control of §5.2 has a slightly higher rank than TailSpline (10.09) and scores 2.1 points lower (Appendices A.3, D.2).

A table change acts through each content coordinate's distance response. For fixed pre-RoPE states, the finite target–competitor log-odds shift sums each stretch-exponent change times its path-averaged signed response. Matching total displacement cancels a response common to all channels, but leaves their signed differences. Equal movement can therefore redistribute attention between target and competitors differently (Appendix A.3).

Frequency changes and coordinate assignment. In the no-alias interval $(0, \pi)$, position-independent invertible Q/K maps preserve the full integer-position rotary kernel exactly if and only if the frequency multisets agree (Appendix A.4), complementing STRING's characterization of relative-translation encodings (Schenck et al., 2025). A changed spectrum therefore cannot be absorbed by such a map, whereas permuting coordinates and frequencies jointly leaves the kernel unchanged. The corresponding fixed-spectrum model interventions are reported in Appendix C.3.

4 DESIGNING ALLOCATIONS

A frozen model associates frequencies with learned content coordinates; a table change alters each coordinate's distance response (Section 3). We therefore use the native table as the reference for two constructions: TailSpline extends the context window, and NCP improves context use within it. For models whose weights will learn, Cosh instead prescribes a new allocation to which they can adapt. All three retain the rotary operator; allocation supplies the design freedom.

4.1 TAILSPLINE: A CLOSED-FORM TRANSITION

Longer inputs combine local dependencies with more widely separated ones (Peng et al., 2024; Wu et al., 2026). Native fast channels preserve their original phase response; fully interpolated slow channels reproduce their native phase at s times the separation. TailSpline connects these regimes through the transition band.

Using the $32/1$-turn rule of YaRN and MrRoPE (Peng et al., 2024; Tian et al., 2026), let l be the last pair above 32 native turns $L\omega_k^N/(2\pi)$ and h the first below one. Put $n = h - l \geq 1$ and $q = \mathrm{clip}(k - l, 0, n)$; the transition accumulates log-stretch increments $\epsilon_q = m_q - m_{q-1}$, with $m_0 = 0$ and $m_n = 1$. For a geometric native grid with gap $c = \log b/K$, each increment widens one log-frequency gap:

$$x'_{l+q} - x'_{l+q-1} = c + \epsilon_q \log s. \tag{6}$$

The increments thus distribute exactly $\log s$ of extra span across the band; in the tail, where $m = 1$, the extra gap returns to zero.

MrRoPE-Pro's cumulative profile $m_q = q(q+1)/[n(n+1)]$ has growing increments, putting its largest extra gaps next to the tail. TailSpline anchors the connection at this constant-stretch tail, leaving the entry free to accumulate log-stretch earlier while the extra gap fades to zero. This design preference gives the one-sided bending cost:

$$\min_{\epsilon:\, \sum_{q=1}^{n} \epsilon_q = 1} \sum_{q=1}^{n-1} (\epsilon_{q+1} - \epsilon_q)^2 + \epsilon_n^2. \tag{7}$$

Theorem 2 (Finite-grid TailSpline)**.** *The unique minimizer of Eq. (7) has positive increments and cumulative profile*

$$\epsilon_q = \frac{3(n+q)(n-q+1)}{n(n+1)(2n+1)}, \qquad m_q^{\mathrm{TS}} = \frac{q(3n^2+3n+1-q^2)}{n(n+1)(2n+1)}. \tag{8}$$

The increments decrease along the band, from $3/(2n+1)$ at the entry to $6/[(n+1)(2n+1)]$ at the tail, so the stretch is front-loaded: $m_q^{\mathrm{TS}} \geq m_q^{\mathrm{Pro}}$ at every transition channel (Fig. 1; derivation in Appendix B.1). The small final step has a direct phase consequence. The last interior pair's residual phase relative to full interpolation ω^N/s satisfies, for $n \geq 2$, $s > 1$ and $\Delta \neq 0$,

$$\delta_M = |\Delta|\, |\omega_{h-1}^M - \omega_{h-1}^N/s|, \qquad \frac{\delta_{\mathrm{TS}}}{\delta_{\mathrm{Pro}}} = \frac{s^{\epsilon_n^{\mathrm{TS}}} - 1}{s^{\epsilon_n^{\mathrm{Pro}}} - 1} \leq \frac{3}{2n+1}. \tag{9}$$

For Llama at $s = 4$, the last interior pair has 92% less residual phase than MrRoPE-Pro relative to full interpolation (ratio 0.080). Appendix B.2 relates phase mismatch to dependency stretch and bounds the finite content response. Section 5 evaluates the resulting table across models and tasks.

Installation. From the public native grid, length and scale, locate the $32/1$-turn band, compute m_k, $\omega'_k = \omega_k^N s^{-m_k}$ and the shared gain $g = 1 + 0.1 \ln s$, and install the table before prefill. Construction costs $O(K)$, needs no calibration or weight updates, and keeps standard rotary computation with one fixed table per checkpoint and target scale.

4.2 NCP: REALLOCATION INSIDE THE NATIVE WINDOW

NCP (native contrastive proximal allocation) slightly lowers interior frequencies while preserving the native endpoints, ordering, weights and unit gain. Each shift reduces a reference loss separating an aligned target from a randomly phased distractor at native distances, with a quadratic penalty on the log-frequency shift. Computed from the public native grid and length, the table is fixed across layers and inputs (Appendix E.1).

4.3 COSH: A TRANSPORT FOR LEARNING

When weights are trained with the table, the allocation can move further from the native one. Cosh moves interior frequencies toward the fast end with a density that balances log-frequency concentration against joint occupancy of the slow tail, where finite-window overlap concentrates (Proposition 1). The unique minimizer of this objective (Appendix B.4) has inverse CDF

$$Q_\tau(u) = 1 - \frac{\operatorname{arcsinh}((1-u)\sinh\tau)}{\tau}, \qquad u_k = \frac{k+1/2}{K}, \tag{10}$$

and recovers the uniform density as $\tau \to 0$. Affine anchoring of $Q_\tau(u_k)$ to chosen endpoints gives a convex allocation; τ sets how far frequencies move (Appendix B.4). Every installation retains the standard rotary operator.

5 EXPERIMENTS

Frozen comparisons install fixed tables in released checkpoints, share prompts and greedy decoding, and report paired 95% bootstrap intervals (Appendix D).

5.1 TAILSPLINE VERSUS YARN AND MRROPE

Our MrRoPE-Pro tables numerically match the released constructor; TailSpline uses the same endpoints, outer bands and gain (Appendix D.6). YaRN retains its public recipe, including Kanana's recommended factor 4.4 (Peng et al., 2024; Kakao, 2025). The main $s = 4$ comparison extends 8K-native Llama-3 to 32K and 32K-native Qwen2.5 and GLM-4 to 128K; TailSpline and MrRoPE-Pro use $s = 2$ for Kanana's 64K.

TailSpline leads across model families and scales. Across five model families from 1B to 70B, YaRN and MrRoPE-Pro exchange ranks while TailSpline has the highest Full-13 mean in every panel (Table 1). In the largest three-method panel, 200 prompts per task on Llama-3-8B at 32K, TailSpline leads MrRoPE-Pro by 11.72 points $[10.32, 13.11]$ and YaRN by 11.16 $[9.78, 12.54]$; on OLMo, its lead over MrRoPE-Pro exceeds 40 points. At 128K it leads both baselines on Qwen2.5-3B and GLM-4-9B, which rotates only half of each head. At the largest scale, Llama-3-70B (NF4), it leads MrRoPE-Pro by 17.72 points; on Kanana, it exceeds the publisher-recommended YaRN recipe by 7.01 points.

Table 1: **TailSpline has the highest Full-13 mean in every panel.** Full-13 (%) of frozen models at their extended length. Within each row, all methods are scored on the same prompts. For OLMo-2-1B, the three-method panel is a subset of the larger TailSpline–MrRoPE-Pro panel. MrRoPE-Pro shares TailSpline's endpoints, outer bands and gain, so TailSpline's gain over it isolates the interior allocation (paired 95% interval, points). Bold marks the highest score; dashes mark comparisons not run. Llama-3-8B at 8K and 16K is shown in Fig. 2; task scores and sample sizes are in Appendix D.

Model	Length	YaRN	MrRoPE-Pro	TailSpline	Gain over MrRoPE-Pro
OLMo-2-1B	4K → 16K	9.81	10.60	**52.24**	+41.64 [35.38, 47.88]
larger panel	4K → 16K	–	9.23	**50.65**	+41.42 [39.99, 42.81]
Qwen2.5-3B	32K → 128K	62.04	59.87	**64.40**	+4.53 [−0.41, 9.49]
Llama-3-8B	8K → 32K	57.11	56.54	**68.27**	+11.72 [10.32, 13.11]
Kanana-1.5-8B[a]	32K → 64K	65.64	70.33	**72.65**	+2.32 [−3.13, 8.04]
GLM-4-9B	32K → 128K	27.85	30.78	**39.65**	+8.87 [4.04, 13.69]
Llama-3-70B (NF4)	8K → 32K	–	60.95	**78.67**	+17.72 [11.65, 23.78]

[a] TailSpline and MrRoPE-Pro use $s = 2$; YaRN uses the publisher-recommended factor 4.4.

One fixed table leads the extension baselines across input lengths. A deployed model serves inputs from its native length up to sL, not only at the target length. The same fixed $s = 4$ Llama table improves Full-13 over MrRoPE-Pro at 8K, 16K and 32K, by 2.76 $[1.29, 4.25]$, 3.39 $[1.53, 5.34]$ and 11.72 points (Fig. 2a), and at 32K the gain covers all four RULER task families (Fig. 2b).

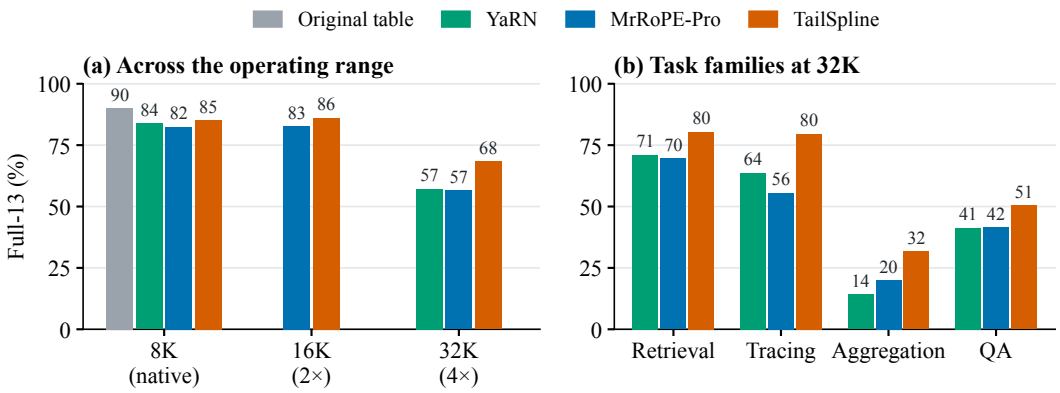


Figure 2: **One fixed TailSpline table outscores the extension baselines at 8K, 16K and 32K.** Full-13 (%) of Llama-3-8B with static $s = 4$ tables; the baselines are MrRoPE-Pro and YaRN, which was evaluated at 8K and 32K. (a) Scores at the native 8K length and at $2\times$ and $4\times$ extension; the gray bar is the original table, which every extension table trails at 8K. (b) The 32K scores by RULER task family: retrieval averages the eight needle tasks, tracing is variable tracking, aggregation averages common- and frequent-word extraction, and QA averages the two QA tasks. Complete task scores are in Appendix D.2.

At the native 8K length, every extension table trails the original one, TailSpline the least, and the shared attention gain accounts for most of the gap (Appendix D.4).

At $16\times$ extension, TailSpline leads MrRoPE-Pro by 20.77 Full-13 points $[15.96, 25.64]$ on Llama-3-8B at 128K (10 prompts per task). At eightfold extension to 256K, TailSpline retains strong single-needle retrieval on Qwen2.5-3B, scoring 93.3% versus 40.0% with MrRoPE-Pro on 15 prompts. Higher-factor book QA and PPL show task-dependent trade-offs (Appendix D.5).

Table 2: **TailSpline has higher means than MrRoPE-Pro in all eight QA/RAG comparisons.** Scores (%). Natural-QA reports five-task LongBench macro F1; 316/631 Llama inputs lie within its native 8K window. Book QA uses official F1 on complete, untruncated InfiniteBench books: pool I, 100K–128K; GLM's disjoint pool II, 70K–95K. LongBench v2 (117 inputs, up to 32K) uses accuracy; HELMET-RAG-derived uses substring exact match. Gains carry paired 95% source-cluster intervals (points). Bold marks the highest mean; dashes mark unevaluated comparisons. Protocols: Appendix D.4.

Model	Benchmark	YaRN	MrRoPE-Pro	TailSpline	Gain over MrRoPE-Pro
OLMo-2-1B	Natural-QA	22.27	21.62	**24.92**	+3.30 [0.39, 6.37]
Qwen2.5-3B	Book QA (I)	18.45	15.90	**19.14**	+3.24 [0.26, 7.51]
Llama-3-8B	Natural-QA	**41.75**	40.88	41.08	+0.20 [−1.48, 1.86]
Llama-3-8B	LongBench v2	–	30.77	**35.04**	+4.27 [−3.39, 11.97]
Llama-3-8B	HELMET-RAG-derived 32K	–	47.27	**48.05**	+0.78 [−3.30, 4.80]
Kanana-1.5-8B†	Book QA	17.85	–	**19.59**	–
GLM-4-9B	Book QA (I)	27.32	25.28	**28.61**	+3.33 [−3.28, 8.98]
GLM-4-9B	Book QA (II)	23.50	22.86	**26.87**	+4.00 [−1.73, 9.68]
Llama-3-70B (NF4)	Natural-QA	–	48.39	**49.28**	+0.89 [−0.60, 2.44]

†TailSpline uses $s = 4$; YaRN uses the publisher-recommended factor 4.4.

The gains extend to natural QA, RAG and language modeling. TailSpline has a higher mean than MrRoPE-Pro in all eight natural QA and retrieval-augmented comparisons, by up to 4.3 points, and exceeds YaRN in five of the six comparisons that include it (Table 2). These tasks test passage and long-document QA (Bai et al., 2024; 2025; Zhang et al., 2024) and retrieval-augmented questions with growing distractor context (Yen et al., 2025). Book QA uses complete books, testing natural text beyond the native window; GLM's gains also hold on a disjoint pool of shorter books. OLMo-2-1B

improves over MrRoPE-Pro on all five Natural-QA tasks. Predictive quality improves as well: at $s = 4$, TailSpline has the lowest extended-context perplexity of the three methods on all four models compared (Appendix Table 10); on OLMo-2-1B at 16K it reaches 8.54, against 12.83 for MrRoPE-Pro and 12.14 for YaRN.

5.2 What the gains identify

TailSpline and MrRoPE-Pro share weights, endpoints, outer bands and gain: Table 1 isolates allocation at fixed range. The following controls address total movement and coordinate assignment; they do not partition that gain into additive shares.

Allocation alone can separate success from failure. At the same range and gain, OLMo-2-1B scores 0.56% on nine held-out RULER tasks at 16K with the geometric allocation of NTK-aware scaling and 61.04% with a coarse ramp (Appendix C.1).

Equal total displacement still permits different quality. With endpoints, outer bands, gain and total displacement matched, different transition profiles span 7.2 Full-13 points on OLMo-2-1B at 16K and 4.8 on Llama at 32K. They include a single step, a plateau and an analytic profile that moves stretch toward the band entry (Appendices B.3, D.2). On the full Llama 32K panel, TailSpline exceeds this analytic control by 2.10 points $[1.11, 3.08]$, while the two stay within half a point at 16K and on OLMo (Appendix D.2). These controls identify allocation effects beyond total displacement; Table 1 evaluates one fixed rule across model families.

Learned coordinate assignment matters beyond the spectrum. Permuting only the interior frequencies of a frozen extension table across coordinates preserves the frequency multiset but lowers Qwen2.5-1.5B's four-task RULER score at 64K from 70% to 0% (Appendix C.3). Together with the kernel criterion in §3.3, this intervention identifies the role of learned coordinate assignment.

5.3 Native reallocation improves earlier-context use

On identical native-length prediction targets, NCP lowers frozen OLMo-2-1B's full-history NLL from 2.915 to 2.902. Comparing full and truncated histories, it gains 0.0079 nat/token more from earlier context than the original table, interval $[0.0055, 0.0103]$: an 8% increase in this history benefit (Appendix E.2). On a 780-prompt development panel, the same fixed table raises OLMo's native 4K Full-13 from 69.37% to 70.78% ($+1.41$ $[0.20, 2.58]$); a separate 130-prompt panel fixed before evaluation gives $+3.26$ $[0.81, 5.92]$, with gains concentrated in variable tracking across reference chains. Appendix E gives the task profiles and a native-window QA comparison.

5.4 Allocation improves learned extrapolation

Paired training isolates an extrapolation gain from interior allocation alone. The 151.9M models train at length 256 with shared architecture, initialization, data order and a 500M-token recipe. The geometric control uses $b = L_{\text{train}}$ (Oka et al., 2026, Sec. 6.1); Cosh ($\tau = 4$) is anchored to the same endpoints, so only the 30 interior z coordinates differ. With the training range retained, Cosh changes mean tail NLL by $+0.026$ at the training length and by -0.281, -0.176 and -0.146 at 512, 1K and 2K tokens, and all three seeds improve at every extended length (Appendix C.2, which separately reports inference-time range retargeting). Swapping runtime tables between the trained models shows that each prefers the table derived from its own training allocation (Appendix Table 4).

At larger scale, a 432M MLA-style model with 16 rotary pairs trains for 500M tokens at 8K across three seeds (DeepSeek-AI, 2024a;b); its Cosh table is not anchored to the geometric endpoints. Cosh improves all five extended lengths in every seed, cutting 16K perplexity by 31%, from 138.8 to 95.6, and the benefit persists under a common wavelength blend (Fig. 3). In matched continuation of a 750M model with fixed endpoints, Cosh nearly halves 16K perplexity, from 45.1 to 24.4, and raises 8K answer-token exact match from 0 to 31 of 40. Relative to matched native-table LoRA, Llama-3-8B LoRA adaptation with Cosh cuts 16K perplexity more than fourfold, from 108.96 to 24.07 (Appendix F, which also reports native-window costs).

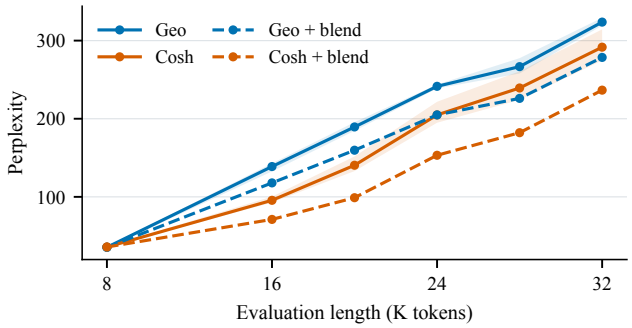


Figure 3: **Cosh lowers extrapolation perplexity in every training seed.** Three-seed mean and range for the 432M MLA model trained at 8K; dashed curves use the common wavelength blend ($s = 4$). Evaluation uses common text chunks.

6 Related Work

Range and allocation. Position Interpolation, NTK-aware scaling and adjusted base frequencies change the frequency range and keep the geometric allocation (Chen et al., 2023; bloc97, 2023; Xiong et al., 2024); analyses of the base relate it to training and attainable context lengths (Liu et al., 2024; Xu et al., 2024; Oka et al., 2026). YaRN and MrRoPE use nonuniform allocation, with MrRoPE comparing transition profiles at shared band boundaries (Peng et al., 2024; Tian et al., 2026). Our matched-displacement and coordinate interventions show that neither total movement nor the frequency multiset alone determines context use.

Search, learning and closed-form construction. LongRoPE and LongRoPE2 search per-dimension scaling factors with model perplexity (Ding et al., 2024; Shang et al., 2025), DPE selects effective distances and key dimensions with task probes (Lu et al., 2025), and LeRoPE learns frequency scales jointly with the weights (Karypis et al., 2026). TailSpline is computed in closed form from public RoPE parameters, without search, calibration or weight updates; Cosh prescribes a table while the weights learn.

Positional overlap and learned use. Prior work studies how models use RoPE frequencies for positional and semantic information (Barbero et al., 2025; Chen et al., 2025; Jin et al., 2025) and characterizes relative-translation encodings (Schenck et al., 2025). Our complete-pair geometry and integer-position criterion separate positional overlap, coordinate assignment and spectral change. Appendix G discusses further related work.

7 Conclusion

RoPE allocation shapes context use beyond frequency range and total movement. Matched fixed-range and equal-displacement comparisons identify its effect, while complete-pair analysis separates positional overlap from content discrimination. Kernel equivalence and coordinate interventions distinguish changing a spectrum from reassigning frequencies to learned coordinates. These findings motivate native-referenced allocation for frozen models.

TailSpline makes this freedom usable through a public, closed-form rule requiring no training. It leads YaRN and MrRoPE-Pro by more than 11 Full-13 points on Llama-3-8B at 32K and has the highest mean in every main-comparison panel across five model families up to 70B. The same Llama $s = 4$ table leads MrRoPE-Pro at 8K, 16K and 32K; the eight main QA/RAG means also exceed MrRoPE-Pro's (Table 2). NCP and Cosh extend the evidence to native context use and learned extrapolation. The preferred allocation depends on both the range and learned frequency use, making it a practical design dimension within and beyond the native window.

ETHICS STATEMENT

This work studies frequency allocation in existing language models using public research corpora and evaluation suites under their stated licenses. It introduces no human-subject study or deployed decision system. The proposed constructions retain the rotary operator and add no trainable parameters; adaptation experiments use the weight updates specified in their protocols.

REPRODUCIBILITY STATEMENT

The appendices define the installed frequency tables, exponent grids, attention amplitudes, matched training and adaptation settings, datasets, and evaluation procedures. The fixed-range experiment reports three paired training seeds; the larger-model and frozen-model tables identify their comparison units and sample counts. The source archive provides the active TeX, vector figures, numerical summaries and CPU checks of the Gram-matrix and allocation identities. It also includes the frozen two- and three-arm data-preparation, table-construction, generation, scoring and paired-report entrypoints with their local dependencies. The runtime guide specifies execution parameters, commands, public data sources and the pinned RULER revision.

The core fixed-support identification uses six runs of $499{,}974{,}144$ tokens each; the 432M comparison uses six 500M-token runs, and the 750M continuation uses two 500M-token runs. The appendices state these protocol-specific budgets alongside the corresponding construction and evaluation settings.

AI USE STATEMENT

Under the authors' direction, generative AI tools assisted with manuscript editing and revision; literature search and organization; research ideation and experiment design, implementation, execution and analysis; and developing and checking mathematical derivations and proofs. The authors set the research direction and made the final methodological and editorial decisions. Verification included checking derivations against their stated assumptions, testing key Gram-matrix and allocation identities with CPU scripts, and checking numerical claims against recorded runs and source results. References were checked against primary sources. The source archive provides the corresponding numerical summaries, checks and implementation entrypoints. The authors take responsibility for the correctness, originality and integrity of the paper and accompanying code.

REFERENCES

Zhangir Azerbayev, Edward Ayers, and Bartosz Piotrowski. Proof-Pile. Dataset, 2022. URL https://huggingface.co/datasets/hoskinson-center/proof-pile.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025. doi: 10.18653/v1/2025.acl-long.183.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

John J. Benedetto and Matthew Fickus. Finite normalized tight frames. *Advances in Computational Mathematics*, 18:357–385, 2003. doi: 10.1023/A:1021323312367.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973. doi: 10.1090/S0025-5718-1973-0348991-3.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, 2022. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

bloc97. NTK-aware scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. Reddit, r/LocalLLaMA, 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. doi: 10.48550/arXiv.2306.15595. URL https://arxiv.org/abs/2306.15595.

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. HoPE: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23044–23056, 2025. doi: 10.18653/v1/2025.acl-long.1123. URL https://aclanthology.org/2025.acl-long.1123/.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. LongLoRA: Efficient fine-tuning of long-context large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6PmJoRfdaK.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a. doi: 10.48550/arXiv.2405.04434. URL https://arxiv.org/abs/2405.04434.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024b. URL https://arxiv.org/abs/2412.19437.

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 *PMLR*, pp. 11091–11104, 2024. URL https://proceedings.mlr.press/v235/ding24i.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

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 *First Conference on Language Modeling*, 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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 24932–24949, 2025. URL https://proceedings.mlr.press/v267/hua25b.html.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. Massive values in self-attention modules are the key to contextual knowledge understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28063–28096. PMLR, 2025. URL https://proceedings.mlr.press/v267/jin25f.html.

Kakao. Kanana-1.5-8B-Instruct-2505 model card. https://huggingface.co/kakaocorp/kanana-1.5-8b-instruct-2505, 2025. Official runtime YaRN configuration.

Petros Karypis, Sean O'Brien, Shreyas Kadekodi, Rui Zhu, and Julian McAuley. LeRoPE: Learnable RoPE frequencies improve language modeling. *arXiv preprint arXiv:2607.10134*, 2026. doi: 10.48550/arXiv.2607.10134. URL https://arxiv.org/abs/2607.10134.

Haoran Li, Yingjie Qin, Baoyuan Ou, Lai Xu, and Ruiwen Xu. HoPE: Hybrid of position embedding for long context vision–language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5430. URL https://papers.nips.cc/paper_files/paper/2025/hash/ee37d51b3c003d89acba2363dde256af-Abstract-Conference.html.

Huayang Li, Tianyu Zhao, Deng Cai, and Richard Sproat. RePo: Language models with context re-positioning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64002. arXiv:2512.14391.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of RoPE-based extrapolation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=JO7k0SJ5V6. arXiv:2310.05209.

Yi Lu, Wanxu Zhao, Xin Zhou, Chenxin An, Chenglong Wang, Shuo Li, Yuming Yang, Jun Zhao, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Effective length extrapolation via dimension-wise positional embeddings manipulation. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Tahpc3iAnO.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Byj72udxe. arXiv:1609.07843.

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher. Selective rotary position embedding. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=AQo1SEElNb.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024. doi: 10.48550/arXiv.2406.17557. URL https://arxiv.org/abs/2406.17557.

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

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

Connor Schenck, Isaac Reid, Mithun George Jacob, Alex Bewley, Joshua Ainslie, David Rendleman, Deepali Jain, Mohit Sharma, Kumar Avinava Dubey, Ayzaan Wahid, Sumeet Singh, René Wagner, Tianli Ding, Chuyuan Fu, Arunkumar Byravan, Jake Varley, Alexey A. Gritsenko, Matthias Minderer, Dmitry Kalashnikov, Jonathan Tompson, Vikas Sindhwani, and Krzysztof Marcin Choromanski. Learning the RoPEs: Better 2D and 3D position encodings with STRING. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53295–53315. PMLR, 2025. URL https://proceedings.mlr.press/v267/schenck25a.html.

Ning Shang, Li Lyna Zhang, Siyuan Wang, Gaokai Zhang, Gilsinia Lopez, Fan Yang, Weizhu Chen, and Mao Yang. LongRoPE2: Near-lossless LLM context window scaling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 54203–54218, 2025. URL https://proceedings.mlr.press/v267/shang25a.html.

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. URL https://doi.org/10.1016/j.neucom.2023.127063.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. doi: 10.48550/arXiv.2501.00656. URL https://arxiv.org/abs/2501.00656.

Qingyuan Tian, Wenhong Zhu, Xiaoran Liu, Xiaofeng Wang, and Rui Wang. MrRoPE: Mixed-radix rotary position embedding. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/69413f87e5a34897cd010ca698097d0a-Abstract-Conference.html.

Unsloth. llama-3-70b-Instruct-bnb-4bit model card. https://huggingface.co/unsloth/llama-3-70b-Instruct-bnb-4bit, 2024.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. VideoRoPE: What makes for good video rotary position embedding? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 66118–66136, 2025. URL https://proceedings.mlr.press/v267/wei25h.html.

Xinyi Wu, Siyuan Liu, and Ali Jadbabaie. How data shapes RoPE frequency usage: From positional scale matching to length generalization. *arXiv preprint arXiv:2607.07678*, 2026. doi: 10.48550/arXiv.2607.07678. URL https://arxiv.org/abs/2607.07678.

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 *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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of RoPE bounds context length. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2773. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9f12dd32d552f3ad9eaa0e9dfec291be-Abstract-Conference.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

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

Z.ai. GLM-4-9B-0414 model card. https://huggingface.co/zai-org/GLM-4-9B-0414, 2025.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024. doi: 10.18653/v1/2024.acl-long.814. URL https://aclanthology.org/2024.acl-long.814/.

Yifan Zhang, Zixiang Chen, Yifeng Liu, Zhen Qin, Huizhuo Yuan, Kangping Xu, Yang Yuan, Quanquan Gu, and Andrew Chi-Chih Yao. Group representational position encoding. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5cb58625f49ddf70fe2d527e9e4bbae5-Paper-Conference.pdf.

SUPPLEMENTARY MATERIAL

The controlled findings in Appendix C establish allocation as a design variable. Appendices A and B connect positional structure and learned frequency use to explicit distance-response designs. The remaining appendices test these designs under frozen extension, native-window use and learning, retaining the comparisons that determine where their benefits hold. Tables and figures abbreviate TailSpline, MrRoPE-Pro (MrPro), YaRN and the equal-displacement control as T, P, Y and C; N or Native denotes a model's original table. The accompanying source archive provides numerical inputs, executable construction and scoring protocols, and CPU verification scripts. Table A1 maps the questions to their controls, findings and detailed evidence; Appendix G discusses further related work.

Table A1: **Questions, controls and findings.**

Question	Control	Finding	Where
Does the range suffice?	Frozen allocations at shared endpoints	No; allocation alone changes quality	§5.2; C.1, D.2
Does total movement suffice?	T–C and additional profiles at matched displacement	No: T–C separates at Llama 32K; step and plateau separate from T on OLMo	§5.2; B.3, D.2
Does coordinate use matter?	Spectrum permutation; weight/table crossing	Yes; weights prefer their own frequency–coordinate association	§5.2, §5.4; C.3
How good is TailSpline in practice?	Frozen retrieval, LM and QA/RAG comparisons across models	Highest T/P/Y Full-13 mean in Table 1; higher mean than P in all eight QA/RAG comparisons	§5.1; D.3–D.6
Can native allocation help?	NCP at fixed native range and weights	More LM benefit from earlier history	§5.3; E
How does learning adapt?	Paired training, continuation, adaptation	Learning uses a changed allocation; preference depends on range	§5.4; C.2, C.4, F

A POSITIONAL GEOMETRY AND LEARNED USE

Complete-pair geometry measures positional overlap, while content responses explain why overlapping directions can still support distinct comparisons.

A.1 COMPLETE-PAIR GEOMETRY

Closed-form cross-Gram. For $\Delta \sim \mathrm{Unif}[0, L]$ and $x_\omega(\Delta) = [\cos(\omega\Delta) \;\; \sin(\omega\Delta)]$, write $\sigma_- = (\omega-\nu)L$, $\sigma_+ = (\omega+\nu)L$, $a(t) = \sin t/t$ and $b_\star(t) = (1-\cos t)/t$. Expanding the products of sinusoids and integrating termwise gives

$$H_{\omega\nu} = \mathbb{E}[x_\omega^\top x_\nu] = \frac{1}{2}\begin{bmatrix} a(\sigma_-) + a(\sigma_+) & b_\star(\sigma_+) - b_\star(\sigma_-) \\ b_\star(\sigma_+) + b_\star(\sigma_-) & a(\sigma_-) - a(\sigma_+) \end{bmatrix}, \tag{11}$$

and $S_\omega = H_{\omega\omega}$. Since $Q_{\omega\nu} = S_\omega^{-1/2} H_{\omega\nu} S_\nu^{-1/2}$ is the whitened cross-Gram of two two-dimensional subspaces, its singular values are the canonical correlations, and $c_{\omega\nu} = \frac{1}{2}\|Q_{\omega\nu}\|_F^2$ measures their overlap independently of the choice of basis.

Basis invariance. Under independent invertible basis changes $x_\omega \mapsto x_\omega A_\omega$ and $x_\nu \mapsto x_\nu A_\nu$, let $U_j = S_j^{1/2} A_j (A_j^\top S_j A_j)^{-1/2}$ for $j \in \{\omega, \nu\}$. Then $U_j^\top U_j = I$, and the transformed whitened cross-Gram is $\widetilde{Q}_{\omega\nu} = U_\omega^\top Q_{\omega\nu} U_\nu$. Left and right orthogonal factors preserve the two canonical correlations and therefore $c_{\omega\nu}$. Thus $c_{\omega\nu}$ is an intrinsic measure of overlap between the two positional subspaces.

Theorem 3 (Complete-pair effective-rank identity)**.** *The block-whitened Gram matrix Γ has Rényi-2 effective rank $r_2(\Gamma) = 2K/[1 + (K-1)\bar{c}]$, as in Eq. (5).*

Proof of Theorem 3. Since $\Gamma_{ii} = I_2$ and $\Gamma_{ij} = Q_{\omega_i\omega_j}$, $\operatorname{tr}\Gamma = 2K$ and $\operatorname{tr}(\Gamma^2) = 2K + \sum_{i\neq j}\|Q_{\omega_i\omega_j}\|_F^2 = 2K[1 + (K-1)\bar{c}]$. Substitution proves Eq. (5). □

Proof of Proposition 1. Put $t = \Delta/L$, $x = \omega L$, and use $\theta = x^2$ as the local parameter. The rescaled pair has the analytic basis

$$u_1(\theta, t) = \cos(\sqrt{\theta}\, t) = 1 - \tfrac{1}{2}\theta t^2 + O(\theta^2), \tag{12}$$

$$u_2(\theta, t) = \frac{\sin(\sqrt{\theta}\, t)}{\sqrt{\theta}} = t - \tfrac{1}{6}\theta t^3 + O(\theta^2). \tag{13}$$

Let $F_\theta : \mathbb{R}^2 \to L_2[0,1]$ have these two columns. Its Gram is nonsingular at $\theta = 0$, where the columns are $1, t$. Hence the orthogonal projector $\Pi_\theta = F_\theta(F_\theta^* F_\theta)^{-1}F_\theta^*$ is analytic near zero, and $V_\omega \to \operatorname{span}\{1, \Delta\}$. With $y = \nu L$ and $\epsilon = \max\{|x|, |y|\}$, the squared chordal distance is

$$2 - \|Q_{\omega\nu}\|_F^2 = \tfrac{1}{2}\|\Pi_{x^2} - \Pi_{y^2}\|_{\mathrm{HS}}^2 = O\big((x^2 - y^2)^2\big) = O(\epsilon^4),$$

where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm. □

A.2 Finite-window overlap and complete-pair comparison

On the paired-training grid of Appendix C.2 ($b = 256$, $K = 32$, $L = 256$), the eight slowest frequencies of $\omega_j = b^{-j/K}$ span $\omega_j L \in [1.1892, 4]$. Evaluating Eq. (11) under uniform separations gives $r_2 = 2.11474$ for their 16 coordinates. The largest two eigenvalues, 7.9681 and 7.5752, account for 97.1% of the trace. This is the geometry in Fig. 4, calculated at the finite frequencies without a small-phase approximation.

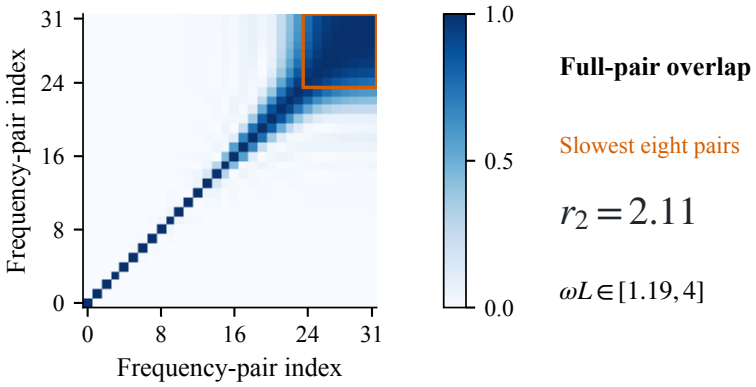


Figure 4: **Finite-window overlap on the paired-training grid.** Complete-pair overlap for $b = 256$, $K = 32$, $L = 256$ under uniform separations. The outlined block marks the eight slowest pairs: their 16 coordinates have normalized positional effective rank $r_2 = 2.11$.

Cosine-only collision can choose the lower-rank table. Take $K = 4$, $L = 8\pi$, and the two descending frequency tables

$$\Omega_A = (1, 5/8, 3/8, 1/4), \qquad \Omega_B = (1, 6.01/8, 4.01/8, 1/4).$$

They share both endpoints. Define $C_{\cos}$ as the mean squared normalized cosine–cosine overlap over distinct pairs. For Ω_A, every cosine is an integer harmonic on $[0, L]$, so $C_{\cos}(A) = 0$ exactly. Mixed harmonic parities nevertheless produce nonzero sine–cosine cross terms. For Ω_B, the small perturbation of an even-harmonic table gives

$$C_{\cos}(A) = 0 < 1.38725 \times 10^{-5} = C_{\cos}(B), \tag{14}$$

$$r_2(A) = 6.29478 < 7.99971 = r_2(B). \tag{15}$$

Thus a strict cosine-only preference can have the opposite full-subspace rank ordering. Both values follow directly from Eq. (11).

A.3 CONTENT COORDINATES AND FINITE ATTENTION RESPONSES

For a block $\mathcal{B}$ of r rotary pairs, write $\mathcal{R}_{\mathcal{B}}(d) = \operatorname{diag}_{k\in\mathcal{B}} R(\omega_k d)$. Each block is orthogonal, so $\mathcal{R}_{\mathcal{B}}(d)$ has rank $2r$ at every distance. This is the rank of an operator on content coordinates; the positional Gram instead measures shared dependence on distance.

Proposition 4 (Content matching in a slow block)**.** *Let $H > 0$ and $\eta = H \max_{k\in\mathcal{B}} |\omega_k|$. For $|d| \le H$,*

$$\left|q_{\mathcal{B}}^{\top}\mathcal{R}_{\mathcal{B}}(d)k_{\mathcal{B}} - q_{\mathcal{B}}^{\top}k_{\mathcal{B}}\right| \le \eta\|q_{\mathcal{B}}\|_2\|k_{\mathcal{B}}\|_2. \tag{16}$$

For $2r$ content symbols represented by distinct coordinate unit vectors, let the query and correct key share a vector and all incorrect keys use other vectors. Their unscaled score margin is at least $1 - 2\eta$ across arbitrary key distances in $[-H, H]$.

Proof. For each pair, $\|R(\omega d) - I\|_2 = 2|\sin(\omega d/2)| \le |\omega d|$. The block-diagonal norm is the maximum pair norm. The bilinear norm bound gives Eq. (16). Unrotated matching scores are one for the correct symbol and zero for an incorrect symbol. Each score changes by at most η, giving the margin bound. In particular, $\eta < 1/2$ preserves the correct ranking. □

Thus the positional subspaces can approach $\operatorname{span}\{1, d\}$ while all $2r$ content symbols remain distinguishable, consistent with prior analyses of frequency use (Barbero et al., 2025).

Task gains with lower positional rank. For the complete Llama native grid $\omega_k^N = 500000^{-k/64}$, native length 8192, $s = 4$ and band $[18, 35]$, Eq. (11) under uniform separations gives TailSpline/MrPro ranks $8.2834/8.7399$ at 16K and $10.0773/10.2149$ at 32K, while Fig. 2 and Table 1 report higher TailSpline task scores. Integer-position Grams, FP32 rounding and causal-separation weights retain the rank ordering. These complete-table comparisons exhibit the distinction between positional diversity and task use.

From distance response to content competition. Fix the pre-RoPE query, keys, valid-key mask and separations d_j for one attention read. Its logits are $\ell_j(\nu) = \ell_j^{\perp} + \operatorname{Re}\sum_k C_{jk}e^{id_j\nu_k}$, with any unrotated contribution $\ell_j^{\perp}$ fixed and the common gain and attention normalization absorbed into C_{jk}. Changing the table gives the finite shift $\eta_j = \operatorname{Re}\sum_k C_{jk}(e^{id_j\nu'_k} - e^{id_j\nu_k})$. For fixed nonempty target and competing groups G, B partitioning the valid keys, let p_G, p_B be their separately normalized original softmax weights. The log-odds $\mathcal{O} = \log\sum_G e^{\ell_j} - \log\sum_B e^{\ell_j}$ obey

$$\mathcal{O}' - \mathcal{O} = \log\mathbb{E}_{p_G}e^{\eta_j} - \log\mathbb{E}_{p_B}e^{\eta_j}. \tag{17}$$

Indeed, factor each original group partition function from its updated sum. If $\min_G \eta_j > \max_B \eta_j$, the target share strictly increases. For $\nu_k(m) = \omega_k^N s^{-m_k}$ at fixed $s > 1$, let $p_{G,m}$ and $p_{B,m}$ denote the group softmax weights at m, and define the signed response $\chi_k = \partial\mathcal{O}/\partial m_k$. For a finite change v supported on I with $\sum_I v_k = 0$, put $\bar{\chi}_k = \int_0^1 \chi_k(m + tv)\,dt$. Then, for any common constant c,

$$\chi_k(m) = \log(s)\nu_k(m)\big(\mathbb{E}_{p_{G,m}} - \mathbb{E}_{p_{B,m}}\big)\left[d_j \operatorname{Im}(C_{jk}e^{id_j\nu_k})\right], \tag{18}$$

$$\mathcal{O}(m + v) - \mathcal{O}(m) = \sum_{k\in I} v_k(\bar{\chi}_k - c). \tag{19}$$

Here $\nu_k = \nu_k(m)$ inside the expectation. The group softmax weights change along the path while the pre-RoPE states stay fixed. Differentiating the logits and group log-partitions gives the first identity; integrating along the path and using the zero-sum constraint gives the second. Thus equal total displacement cancels a common response, leaving signed channel differences.

A.4 INTEGER-POSITION KERNEL EQUIVALENCE

The distinction between positional functions and their content coefficients also separates changing a spectrum from reassigning its coordinates.

Corollary 5 (Integer-position equivalence)**.** *Let Ω, Λ contain K frequencies in $(0, \pi)$ and $\mathcal{R}_\Omega(d) = \operatorname{diag}_k R(\omega_k d)$. Position-independent invertible A, B satisfy $A^{\top}\mathcal{R}_\Omega(d)B = \mathcal{R}_\Lambda(d)$ for $d = 0, 1$ if and only if the frequency multisets agree.*

The statement concerns the full bilinear kernel on the $2K$-dimensional rotary space: equality for all content pairs is matrix equality. At $d = 0$, $A^\top B = I$; at $d = 1$ the two rotation matrices are similar. Their complex eigenvalues are $e^{i\omega_k}, e^{-i\omega_k}$ including multiplicities. On $(0, \pi)$ the positive-angle eigenvalue identifies each frequency uniquely, proving necessity even with repeated frequencies. For sufficiency choose a block permutation P with $\mathcal{R}_\Lambda(d) = P\mathcal{R}_\Omega(d)P^\top$, and set $A = B = P^\top$. This holds for every integer d, including negative distances. Outside $(0, \pi)$, aliases such as ω and $\omega + 2\pi$ are indistinguishable at integer positions.

B Explicit Constructions and Distance Scaling

The finite tail-connection objective determines TailSpline uniquely; its native-distance response then motivates an exact equal-displacement control. The allocation space is $\mathcal{Z}_K = \{0 = z_0 < \cdots < z_{K-1} = 1\}$. For native log-frequency span R_N and stretch exponents with $m_0 = 0$, $m_{K-1} = 1$, the extended allocation is

$$z'_k = \frac{R_N z_k^N + m_k \log s}{R_N + \log s}.$$

Thus allocation and native-relative stretch describe the same table when the native grid and scale are fixed; total displacement is $D = \log s \sum_k m_k$.

B.1 TailSpline and the symmetric construction

Let $D\epsilon = (\epsilon_{q+1} - \epsilon_q)_{q=1}^{n-1}$ and $H = D^\top D + e_n e_n^\top$. A zero value of $\epsilon^\top H \epsilon$ requires constant increments and a zero terminal increment, hence $\epsilon = 0$. Thus H is positive definite. For $n > 1$, the proposed $\epsilon_q = c(n + q)(n - q + 1)$, with $c = 3/[n(n + 1)(2n + 1)]$, satisfies

$$\epsilon_1 - \epsilon_2 = 2c, \quad 2\epsilon_q - \epsilon_{q-1} - \epsilon_{q+1} = 2c, \quad 2\epsilon_n - \epsilon_{n-1} = 2c.$$

Therefore $H\epsilon = 2c\mathbf{1}$, the equality-constrained stationarity equation. The increments are positive and sum to one; summing them yields Eq. (8). The objective minimum is $2c = 6/[n(n + 1)(2n + 1)]$. For $n = 1$, the constraint gives $\epsilon_1 = 1$, agreeing with the formulas. Positivity makes additional nonnegativity constraints inactive.

Adding ϵ_1^2 yields the symmetric objective, with unique solution

$$\epsilon_q^{\mathrm{BM}} = \frac{6q(n - q + 1)}{n(n + 1)(n + 2)}, \qquad m_q^{\mathrm{BM}} = \frac{q(q + 1)(3n + 2 - 2q)}{n(n + 1)(n + 2)}. \tag{20}$$

Both profiles are monotone with $m_0 = 0, m_n = 1$.

What the objective smooths. On the geometric native grid, Eq. (6) makes each change between neighboring log gaps equal to $(\epsilon_{q+1} - \epsilon_q) \log s$. The terminal change is $-\epsilon_n \log s$ because the fully interpolated tail has constant displacement. Thus $(\log s)^2$ times Eq. (7) is exactly the squared log-gap variation within the transition and at its low-frequency junction. The high-frequency junction is deliberately unpenalized.

Write $T_q = m_q^{\mathrm{TS}}$ and $P_q = q(q + 1)/[n(n + 1)]$ for $q = 0, \ldots, n$. Direct subtraction gives

$$T_q - P_q = \frac{q(n - q)(3n + q + 1)}{n(n + 1)(2n + 1)} \geq 0. \tag{21}$$

TailSpline thus assigns at least as much stretch throughout the transition. Its entry increment is $3/(2n + 1)$, compared with MrPro's $2/[n(n + 1)]$; its terminal increment is $6/[(n + 1)(2n + 1)]$, compared with $2/(n + 1)$. For $n > 1$, the larger entry jump accompanies a terminal jump smaller by the factor $3/(2n + 1)$.

B.2 Native phase scale and dependency stretch

The native response at a dependency's original distance supplies the reference for comparing allocations. The turn rule first locates the transition on that scale; the amount of dependency stretch then determines which allocation more closely matches the native phase.

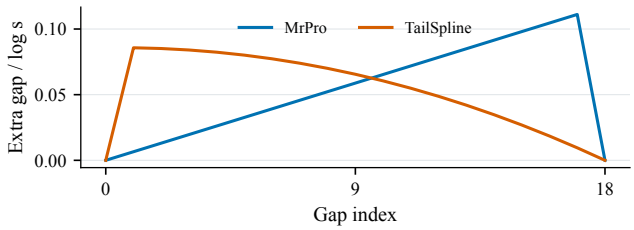


Figure 5: **TailSpline connects the transition to the extended tail.** Extra log-frequency gaps for $n = 17$. Both profiles retain the same outer bands; TailSpline allocates a larger entry gap and smoothly approaches the zero extra gap of the fully interpolated tail.

Turn boundaries across native bases. The $32/1$-turn rule places the transition in a comparable native phase range. For a native geometric grid $\omega_k = \omega_k^N$ with $c = \log b/K$, assume both boundaries lie inside the sampled table without clipping. Let l be the last slot above 32 turns, h the first below one, $n = h - l$ and $u = q/n$. Define $\eta_+ = \log(L\omega_l/(64\pi))$ and $\eta_- = \log(2\pi/(L\omega_h))$. Both belong to $(0, c]$, and

$$nc = \log 32 + \eta_+ + \eta_-, \qquad L\omega_{l+q} = 64\pi\, 32^{-u} \exp\big((1-u)\eta_+ - u\eta_-\big). \tag{22}$$

These identities follow from $\omega_{l+q} = \omega_l e^{-cq}$; in particular, $|\log[L\omega_{l+q}/(64\pi\, 32^{-u})]| \le c$. Together with $d\nu = (d/L)(L\omega)s^{-m}$, this bound relates the finite native grid to the phase at an actual separation d.

Tail-junction phase residual. For one channel, put $\omega = \omega_{l+q}^N$ and $\nu_M = \omega s^{-M_q}$ for $M \in \{T, P\}$. Equation (21) gives $\omega/s \le \nu_T \le \nu_P \le \omega$ for $s > 1$. For $0 \le q < n$ and nonzero distance d, write $\delta_M(q) = d(\nu_M - \omega/s) = (d\omega/s)(s^{1-M_q} - 1)$. The finite-grid formulas give

$$\frac{|\delta_T(q)|}{|\delta_P(q)|} \le \frac{1 - T_q}{1 - P_q} = \frac{(n-q+1)(2n+q+1)}{(2n+1)(n+q+1)}. \tag{23}$$

Indeed, $0 < 1 - T_q \le 1 - P_q$ and $(e^x - 1)/x$ increases for $x > 0$. At $q = n - 1$, the bound is $3/(2n+1)$, recovering Eq. (9).

Dependency stretch and response. Let a reference dependency have separation $d_0 > 0$ and its extended separation be αd_0. The residual phase in channel q is $\delta_M = \omega d_0(\alpha s^{-M_q} - 1)$. For an interior pair with $T_q > P_q$ and $s > 1$, the smaller absolute residual switches from P to T at

$$\alpha_q^* = \frac{2}{s^{-T_q} + s^{-P_q}}. \tag{24}$$

Indeed, with $v_M = s^{-M_q}$, $\delta_T^2 - \delta_P^2 = (\omega d_0)^2 \alpha(v_T - v_P)[\alpha(v_T + v_P) - 2]$. Thus P has smaller residual for $0 < \alpha < \alpha_q^*$, and T for $\alpha > \alpha_q^*$. For Llama $n = 17$, $s = 4$, the maximum interior crossover is $3.670 < 4$ ($q = n$ is the common endpoint). Each dependency has its own stretch α.

To allow a nonzero content phase, write a reference-aligned complex coefficient as $a + ib$, with $a > 0$ and $|b| \le \kappa a$. Its response at residual δ is $f(\delta) = a\cos\delta - b\sin\delta$. If two residuals have the same sign and $0 \le r = |\delta_A| < t = |\delta_B| \le \pi$, then

$$f(\delta_A) - f(\delta_B) \ge 2a \sin\frac{t-r}{2}\left(\sin\frac{t+r}{2} - \kappa\left|\cos\frac{t+r}{2}\right|\right). \tag{25}$$

This follows from the cosine and sine difference identities and $|b| \le \kappa a$; an allowed endpoint value of b attains the bound. A positive bracket guarantees a larger response. At $\kappa = 0$ it recovers the aligned cosine comparison; beyond that case the allowable phase offset is explicit. These are fixed-content responses. Their contribution to attention depends on competing keys, as specified in Appendix A.3. For example, $a = 1$, $b = -1$ and positive residuals $0.1/0.3$ give a response difference of -0.156, illustrating why the phase condition matters.

B.3 AN EXACT EQUAL-DISPLACEMENT CONTROL

At fixed support, the TailSpline–MrPro comparison also changes total displacement D. C–P likewise varies both D and shape, so these contrasts do not partition T–P into additive contributions. To hold D fixed, retain local indices $q = 0, \dots, n$ and write $B_q = m_q^{\mathrm{BM}}$ and $U_q = q/n$. Define the front-loaded profile $F_q = 2U_q - P_q$ and the control

$$C_q = (1 - w_n)U_q + w_n F_q, \qquad w_n = \frac{3n}{2(2n+1)}. \tag{26}$$

Substitution gives $T = (1 - w_n)B + w_n F$, hence

$$T_q - C_q = (1 - w_n)(B_q - U_q) = \frac{q(n-q)(2q-n)}{2n(n+1)(2n+1)}. \tag{27}$$

This difference is antisymmetric under $q \mapsto n-q$, vanishes at both endpoints, and sums to zero. Since U and F have positive increments and $0 < w_n < 1$, C is monotone with $C_0 = 0, C_n = 1$. Extending T and C by the same constant outer bands gives identical endpoints and total log displacement $D = \log s \sum_k m_k$ at a common scale.

The matched displacement also fixes the increment centroid. For any profile with $m_0 = 0$, $m_n = 1$ and $\epsilon_q = m_q - m_{q-1}$, summation by parts gives

$$\sum_{q=1}^{n} q\epsilon_q = n - \sum_{q=1}^{n-1} m_q, \qquad D = Ka' + R' \sum_k z'_k - \sum_k x_k^N. \tag{28}$$

The second identity follows from $x'_k = a' + R' z'_k$ and $D = \sum_k (x'_k - x_k^N)$. Thus, with common outer bands and scale, equal D already entails equal increment centroids. For $n = 1, 2$, $B = U$ and $T = C$; for wider bands the contrast tests the remaining shape difference.

TailSpline puts less stretch in the entry half and more in the tail half. With $v_q = T_q - C_q$ and global slot $k = l + q$, antisymmetry pairs the finite responses in Eq. (19) as

$$\mathcal{O}(T) - \mathcal{O}(C) = \sum_{n/2<q<n} v_q(\bar{\chi}_{l+q} - \bar{\chi}_{l+n-q}), \tag{29}$$

where $\bar{\chi}$ is integrated along $C + t(T - C)$. Since $v_q > 0$ in this sum, a positive weighted tail–entry response difference increases the target attention share. The paired T–C results are in Appendix D.2.

B.4 COSH DENSITY AND FINITE INSTALLATION

Let $S_\rho(t) = \int_t^1 \rho(\phi) d\phi$. For $\alpha > 0$, $\beta \geq 0$, the Cosh transport minimizes

$$\mathcal{C}_{\mathrm{app}}[\rho] = \frac{\alpha}{2} \int_0^1 \rho^2 + \frac{\beta}{2} \int_0^1 S_\rho^2, \qquad \rho \geq 0, \quad \int_0^1 \rho = 1. \tag{30}$$

The first term penalizes concentrated log-frequency density. By Fubini, $\int_0^1 S_\rho(t)^2 dt = \mathbb{E}[\min(\Phi_1, \Phi_2)]$ for independent $\Phi_i \sim \rho$, so the second penalizes joint occupancy toward the slow end. These declared structural preferences define the transport objective. Coercivity and weak lower semicontinuity on the closed convex constraint set give existence; the first term makes the objective strictly convex. For $g(\phi) = \int_0^1 \min(\phi, \psi)\rho(\psi) d\psi$, stationarity is $\alpha\rho + \beta g + \lambda = 0$. Since $g'' = -\rho$, $g'(1) = 0$ and $g'(0) = 1$, this gives $\rho'' = \tau^2 \rho$, $\rho'(0) = -\tau^2$ and $\rho'(1) = 0$, where $\tau^2 = \beta/\alpha$. Thus the unique solution is

$$\rho_\tau(\phi) = \frac{\tau \cosh[\tau(1 - \phi)]}{\sinh \tau}. \tag{31}$$

It is positive and has unit mass, so it satisfies the inequality constraint. At $\beta = 0$ the solution is uniform. Integrating and inverting its CDF yields Eq. (10).

For midpoint quantiles $u_k = (k + 1/2)/K$, anchored installation uses

$$z_k = \frac{Q_\tau(u_k) - Q_\tau(u_0)}{Q_\tau(u_{K-1}) - Q_\tau(u_0)} \leq \frac{k}{K-1}, \qquad \omega_k = \exp(-a - Rz_k). \tag{32}$$

The inequality follows from convexity of Q_τ and its endpoint chord. Unanchored midpoint installation instead uses $\omega_k = b^{-Q_\tau(u_k)}$. Its uniform limit is midpoint Geo; standard Geo uses k/K, differing by the common frequency factor $b^{-1/(2K)}$. Each experimental protocol below specifies its actual grid and support convention. Native denotes the inherited frequency table of a released model. The fixed-support reference strength is $\tau = d_{\text{head}}/\sqrt{L_{\text{train}}}$. The 432M MLA recipe instead uses the empirical effective-dimension convention $128/\sqrt{8192}$, rounded to $\tau = 1.414$; the 750M continuation uses $\tau = 1.5$. Appendix F specifies these training recipes. Appendix C.4 compares neighboring strengths and an RMS-movement-matched exponential transport in the fixed-support setting.

C Controlled Allocation Experiments

Fixed-support interventions isolate interior allocation, while range retargeting and coordinate interventions establish its interaction with the evaluation support and learned frequency use.

C.1 Frozen fixed-support interventions

Frozen OLMo-2-1B and Qwen2.5-1.5B use 64 pairs, their native fast endpoint and the native slow endpoint divided by four. Matched-support arms share gain $1 + 0.1 \ln 4$, inputs, greedy decoder and precision. Uniform z is log-linear between these endpoints, giving the NTK-aware table with $b' = b\, s^{K/(K-1)}$ (bloc97, 2023). The shared gain above is retained. Let X_k contain a weighted sine/cosine pair and X_{-k} the other pairs. Define its residual fraction and normalization by

$$U_k = \frac{\|X_k - X_{-k}X_{-k}^{\dagger}X_k\|_F^2}{\|X_k\|_F^2}, \qquad \bar{U}_k = \frac{U_k - \min_i U_i}{\max_i U_i - \min_i U_i}.$$

The pseudoinverse cutoff is 10^{-10}; a zero normalization span gives $\bar{U}_k = 0$. The frozen controls use native integer separations weighted by $L - d$, grouped in pairs at rounded weighted centroids. The derived arm installs $\omega'_k = \omega_k[1 - 3(1 - \bar{U}_k)^2/4]$ and exact endpoint overwrite. The coarse arm enumerates integer linear-ramp bounds, minimizing MSE to $(1 - \bar{U}_k)^2$ with lexicographic tie-breaking: bounds are $(20, 22)$ for OLMo and $(28, 31)$ for Qwen. These are geometric, label-free controls.

On OLMo's nine held-out tasks (20/task, 16K), uniform/coarse/derived score $0.56/61.04/60.47\%$; Qwen's four development tasks (20/task, 64K) score $57.75/64.00/66.50\%$. The sampled support is identical within each three-arm block. The nine OLMo tasks are held out for confirmation: the tables, amplitude, task split, checkpoint and metric were fixed before this subset was evaluated.

C.2 Fixed-support training and range interaction

The paired 151,898,880-parameter decoder has 12 layers, width 768, 12 heads of dimension 64, MLP width 3072 and vocabulary 50304. Within each seed $(42, 137, 256)$, Geo and anchored Cosh share initialization, token order, optimizer, batch and training length 256. Each arm sees 499,974,144 tokens in 7629 updates of 256 sequences. Seed 42 uses microbatch 64 with accumulation 4; the other seeds use 128×2. Fused AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, clipping 1, peak/minimum LR $6 \times 10^{-4}/6 \times 10^{-5}$ and 762 warmup steps followed by cosine decay.

FineWeb-Edu training and validation use disjoint unshuffled shards `sample/10BT/000_00000.parquet` and `sample/10BT/004_00000.parquet`, revision `87f09149ef47`, with GPT-NeoX-20B tokenizer revision `c292233c833e` (Penedo et al., 2024; Black et al., 2022). Geo uses $b = 256$ and $\omega_k = \exp[-Rk/(K-1)]$, $R = (K-1)\log 256/K$, $K = 32$. Cosh uses anchored midpoint quantiles with $\tau = 4$ and exactly the same endpoints. Only the 30 interior coordinates differ.

Evaluation scores the final 128 target tokens at 32 fixed validation anchors. Seed 20260723 selects one end position from each of 32 equal-width stream bins with a 2048-token guard; the longest windows are disjoint. At anchor e, length H uses tokens $[e - H, e)$ and scores $[e - 128, e)$. The retained-range condition keeps R fixed; the target-matched condition uses $R(H) = (K-1)\log H/K$ while retaining each trained weight set and normalized z. Both use unit gain. Training seed is the replication unit.

Table 3: **Absolute tail NLL at 151.9M under retained and target-matched support.** Geo denotes the matched geometric control and Cosh its anchored $\tau = 4$ counterpart. Absolute scores and paired differences $\Delta = \text{Cosh} - \text{Geo}$ use the final 128 tokens of 32 frozen anchors; $\pm$ denotes sample SD across three training seeds. Negative differences favor Cosh. Target-matched rows move both supports with the evaluation length at inference; Geo is lower there at every extended length in all three seeds.

Length	Geo NLL	Cosh NLL	Δ_{42}	Δ_{137}	Δ_{256}	Mean $\Delta\pm$SD
Retained training support						
256	3.347 ± 0.007	3.373 ± 0.005	$+0.033$	$+0.029$	$+0.017$	$+0.026 \pm 0.008$
512	4.950 ± 0.011	4.669 ± 0.184	-0.478	-0.271	-0.094	-0.281 ± 0.192
1,024	5.769 ± 0.066	5.593 ± 0.032	-0.205	-0.191	-0.132	-0.176 ± 0.039
2,048	6.311 ± 0.073	6.166 ± 0.071	-0.113	-0.181	-0.143	-0.146 ± 0.034
Target-matched support						
256	3.347 ± 0.007	3.373 ± 0.005	$+0.033$	$+0.029$	$+0.017$	$+0.026 \pm 0.008$
512	3.338 ± 0.011	3.399 ± 0.017	$+0.061$	$+0.053$	$+0.067$	$+0.060 \pm 0.007$
1,024	3.579 ± 0.018	3.806 ± 0.118	$+0.182$	$+0.156$	$+0.344$	$+0.227 \pm 0.102$
2,048	4.608 ± 0.029	5.067 ± 0.200	$+0.279$	$+0.399$	$+0.701$	$+0.460 \pm 0.218$

Retargeting lowers absolute NLL for both allocations at every extended length and favors Geo in all three seeds. It moves the support with the evaluation length at inference, so it is a deployment condition rather than evaluation at the retained training support: the fixed-range result identifies z as a training-time variable but does not establish a gain on top of target-matched range transport.

C.3 Learned frequency use and coordinate assignment

For seeds $137, 256$, each Geo- or Cosh-trained checkpoint installs a factor-four table derived from either training allocation. The derivation uses $L_0 = 256$, $\Delta_j = 2j$ for $j = 0, \ldots, 127$, and grouped causal weights $p_j = (2L_0 - 4j - 1)/(L_0(L_0 + 1)/2)$. Applying the residual construction above gives

$$\omega'_k = \omega_k[1 - 3(1 - \bar{U}_k)^2/4].$$

Tables are FP32 with endpoints overwritten to ω_0 and $\omega_{K-1}/4$, and gain $1 + 0.1 \ln 4$. Evaluation uses the same anchors and tail-NLL metric at $512/1024$ tokens. Both seeds favor the table derived from their training allocation. Table 4 reports the 1024-token mean crossing.

Table 4: Two-training-seed mean tail NLL at $1{,}024$ tokens for the 151.9M weights-by-table crossing (lower is better). Each row freezes one trained weight set; each column installs one separately derived runtime table. Compare within rows to isolate runtime-table preference; comparing diagonal cells changes both weights and table.

Frozen weights	Geo-derived table	Anchored-Cosh-derived table
Geo-trained	**3.426**	5.776
Anchored-Cosh-trained	4.455	**3.479**

Spectrum-preserving coordinate reassignment. A block permutation obeys $\mathcal{R}_{\pi\Omega}(d) = P\mathcal{R}_\Omega(d)P^\top$, so jointly permuting Q/K and frequencies preserves the bilinear kernel exactly. With weights fixed, permuting only interior frequency assignments changes model performance. The intervention randomly permutes the interior entries of the final FP32 table while fixing both endpoint slots, so the sorted spectrum is unchanged; the constructor is in the source archive. The log-law table is $\omega'_k = \omega_k^N s^{-(1-\bar{U}_k)^2}$, where $\bar{U}_k$ is the normalized residual fraction above, computed on native causal separations using at most 2048 weighted-centroid bins. The OLMo experiment uses $L_0 = 4096$, $b = 500000$, $K = 64$, $s = 4$, gain $1 + 0.074 \ln 4$, and fixes slots $0, 63$. Its 20-document

PG-19 tail NLL changes from 3.10423 to 6.86493. Qwen2.5-1.5B recomputes the same profile at $L_0 = 32768$, $b = 10^6$; on 20 inputs per task at 64K (single-key, multikey-2, multikey-3, tracking), macro changes from 70 to 0%. These fixed-spectrum interventions isolate coordinate assignment; their table is distinct from TailSpline.

C.4 Multiple shapes at fixed support

The 50.9M decoder has 6 layers, width 512, MLP width 2048 and vocabulary 50304. The 12 configurations cross bases 500K/1M, training lengths $256/1024$ and $4/8/16$ heads (head dimensions $128/64/32$). Each of three seeds $42/137/256$ trains each arm for 8,388,608 tokens, 128 updates, on a shared repeated WikiText-2 raw stream (Merity et al., 2017) with a 1% supervised passkey mix. Execution is float32 on MPS. The preassigned arms are Geo, anchored Cosh at $0.75/1/1.25$ times $d_{\text{head}}/\sqrt{L}$, and an exponential $z_\gamma(t) = (e^{\gamma t} - 1)/(e^\gamma - 1)$ whose RMS node movement matches reference Cosh without task fitting. Every arm shares sampled endpoints. The OOD mean weights ratios $r \in \{2, 4, 8\}$ by $\log_2(r + 1)$; the bootstrap unit is configuration after averaging paired seeds. Training and evaluation use separate WikiText-2 raw train/validation splits, tokenized with GPT-NeoX-20B. Nonempty stripped rows are concatenated without extra special tokens, then independently repeated/truncated to 8,388,608 training and 1,048,576 evaluation tokens. NumPy RandomState(9999) selects four distinct starts at each of $L, 2L, 4L, 8L$, in that order, shared across arms. Full next-token cross-entropy scores $4(rL - 1)$ targets and averages chunk NLLs before exponentiation; configurations, rather than repeated text windows, are the bootstrap units.

Table 5: **Exact-range factorial** (50.9M; 2 bases $\times$ 2 training lengths $\times$ 3 head dimensions $= 12$ configurations, 3 seeds). All arms share sampled extrema and log-span with Geo. Metric: weighted OOD NLL over $2\times/4\times/8\times$, paired within configuration and seed. All arms were pre-specified. CIs are 10,000 configuration-level bootstraps; p is a two-sided exact sign-flip over all 2^{12} assignments, unadjusted across contrasts (App. C).

Contrast	ΔNLL	95% CI	$\Delta<0$ configs	p
Cosh $0.75\times$ − Geo	-0.00912	$[-0.018, -0.001]$	$8/12$	0.082
Cosh rule − Geo	-0.00988	$[-0.021, 0.002]$	$7/12$	0.125
Cosh $1.25\times$ − Geo	$\mathbf{-0.01210}$	$[-0.021, -0.003]$	$\mathbf{10/12}$	0.027
Matched exponential − Geo	-0.01062	$[-0.021, -0.001]$	$9/12$	0.071
The matched exponential reproduces the non-uniform direction:				
Cosh rule − exponential	$+0.00074$	$[-0.005, 0.007]$	$7/12$	0.836

D FROZEN EVALUATION: COMPARISONS AND TASK TRANSFER

Matched frozen comparisons establish allocation gains across lengths and models, then test their transfer to natural tasks and higher extension factors.

D.1 COMPARISON AND SCORING RULES

MrPro is our equation-level implementation of the published quadratic-progressive profile (Tian et al., 2026). T/P share frozen weights, sampled endpoints, outer bands and attention gain; only the interior allocation changes. FP32 frequency tables are installed after loading BF16 weights and remain fixed across layers and input lengths. No table is fitted to model outputs. Evaluation uses batch one, Flash-SDPA and greedy decoding unless specified below. The gain multiplies both cosine and sine and thus contributes g^2 to the rotated Q/K logit.

Llama-3-8B-Instruct (Grattafiori et al., 2024) uses native length 8192, $K = 64$, base 500000 and band $[18, 35]$; OLMo-2-0425-1B-Instruct (Team OLMo et al., 2025) uses 4096, $K = 64$, base 500000 and band $[14, 32]$. Their extension tables use $s = 4$ and gain $1 + 0.1 \ln 4$. Official static YaRN retains its public $32/1$ correction, boundary rounding, blend and amplitude rule, without training; Kanana's recommended recipe is specified separately.

Our RULER protocol pins revision `c3f5e3b4f87f` and takes source records in order, without content padding or depth selection. All arms share prompt tokens, references and generation caps: 128 for NIAH, 30 for tracking, 120 for common words, 50 for frequent words and 32 for QA. Full-13 averages the official scores equally across all thirteen tasks, including empty and capped outputs. Its 95% percentile intervals use $20{,}000$ paired resamples within task and length, retaining fixed task weights. Natural-task intervals resample source-context clusters instead. PPL is the exponential of token-pooled NLL. Direct-baseline PPL contrasts use mean paired document NLL differences and $10{,}000$ document resamples. Scores are never pooled across models or tokenizers. CWE and FWE denote common- and frequent-word extraction, respectively.

Sampling within the completed Llama panel. Several direct-baseline panels use ten prompts per task. To gauge how much a ten-per-task panel varies, we draw ten of the 200 paired Llama 32K inputs per task without replacement and average the thirteen task differences. All $20{,}000$ draws (seed 20260916) favor T over P; their central 95% range is $[5.72, 17.90]$ points around the full-pool difference of 11.72. This measures subsampling variation within the completed pool; paired scores and the CPU reproducer are in the source archive.

D.2 ACROSS-WINDOW GAINS AND THE CONTRIBUTION OF INTERNAL SHAPE

Llama uses $650/650/2600$ matched prompts at $8/16/32$K; OLMo uses 2600 at 16K. These runs use 8192-token prefill chunks. Llama 32K inputs span 28270–32606 tokens, including the fixed QA-1 source supplement needed to reach 200 records. OLMo uses generation seed 20261101 and QA source offset 5600. The tables are fixed before evaluation. Table 6 gives the shorter-window and OLMo scores; Table 8 completes Llama 32K with YaRN and the matched-displacement control.

Llama 32K T–P is $+11.72$ points $[10.32, 13.11]$, with gains in all four task families; the completed three-arm panel gives T–Y $+11.16$ $[9.78, 12.54]$ on the same 2600 prompts. The intermediate 16K gain is $+3.39$ $[1.53, 5.34]$, although tracking and aggregation change by $-2.40/-2.77$ points. OLMo's 16K T–P gain is $+41.42$ $[39.99, 42.81]$, with all thirteen task means increasing.

At native 8K, original RoPE uses its native table and unit gain. T/P/original RoPE score $85.16/82.39/90.03\%$: T–original is -4.87 $[-6.57, -3.15]$ points. The reduction concentrates in FWE, where T/P produce $45/48$ empty answers and original RoPE none. The other twelve tasks have a descriptive mean difference of $+1.23$ points. On 46 matched documents, whole-prefix 8K PPL is original/T/P $5.2759/5.2953/5.2864$; T–original PPL is $+0.0194$ with paired document interval $[0.0066, 0.0349]$.

Equal displacement. C is the analytic profile in Appendix B.3. T/C share outer bands, endpoints, gain and total log-frequency displacement; the deployed FP32 discrepancy in $\sum m_k$ is below 10^{-8}. C evaluates the same $650/2600$ inputs at $16/32$K, reusing completed T/P outputs. At 16K T/C

Table 6: Complete task scores under our RULER protocol (%). N denotes Native; C matches T's total log-frequency displacement. Llama $8/16$K use 50 inputs/task and OLMo 16K uses 200. Llama 32K appears in Table 8.

Task	Llama 8K				Llama 16K			OLMo 16K		
	N	T	P	Y	T	P	C	T	P	C
cwe	97.60	95.40	89.60	94.20	83.80	80.00	83.20	7.55	3.30	6.20
fwe	87.33	9.33	4.00	2.00	25.33	34.67	30.00	51.83	23.83	50.33
multikey_1	100.00	100.00	98.00	98.00	100.00	88.00	98.00	65.00	6.50	65.50
multikey_2	96.00	100.00	94.00	94.00	98.00	86.00	98.00	43.00	0.50	42.50
multikey_3	98.00	98.00	94.00	100.00	92.00	84.00	92.00	4.50	0.00	5.00
multiquery	99.50	96.50	97.00	97.00	98.00	97.00	98.50	73.25	4.13	74.13
multivalue	97.50	97.00	92.50	94.00	98.50	97.50	98.00	77.13	2.25	77.00
single_1	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.50	11.50	93.00
single_2	100.00	100.00	100.00	100.00	100.00	98.00	100.00	100.00	18.00	100.00
single_3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	64.00	0.00	70.00
qa_1	70.00	70.00	72.00	74.00	76.00	70.00	76.00	39.50	31.50	38.50
qa_2	32.00	46.00	44.00	48.00	60.00	50.00	60.00	25.50	18.50	27.50
vt	92.40	94.80	86.00	88.00	87.60	90.00	88.80	9.70	0.00	14.20

score $86.09/86.35\%$, a difference of -0.25 $[-1.41, 0.89]$ points; at 32K they score $68.27/66.17\%$, a difference of $+2.10$ $[1.11, 3.08]$. Both lengths use the same $s = 4$ table. OLMo at 16K uses all 2600 matched rows: T/C score $50.65/51.07\%$, with T–C -0.42 $[-1.43, 0.58]$ points. Its analytic equal-displacement profile is rounded to FP32 with two output-independent one-ulp adjustments, giving absolute error 1.16×10^{-9} in total log-frequency displacement; endpoints, outer bands and gain remain fixed. C–P intervals for all three panels are in Table 7; complete task profiles appear alongside the direct baselines. At Llama 32K, C has slightly higher positional effective rank than T ($10.093/10.077$).

Table 7: **Allocation beyond total displacement.** All tables use $s = 4$. T/C/P: TailSpline/equal-displacement control/MrPro. Full-13 means (%) and paired 95% intervals (pp), on $650/2600/2600$ prompts in row order.

Panel	T / C / P	T–C [95%]	C–P [95%]
Llama 16K	86.09/86.35/82.71	-0.25 [-1.41, 0.89]	+3.64 [1.82, 5.57]
Llama 32K	68.27/66.17/56.54	+2.10 [1.11, 3.08]	+9.63 [8.29, 10.91]
OLMo 16K	50.65/51.07/9.23	-0.42 [-1.43, 0.58]	+41.83 [40.40, 43.29]

Table 8: Complete Full-13 task scores (%): Llama 32K (200/task) and OLMo 16K (10/task). C matches T's total displacement. OLMo's 10/task T/P rows are the first ten records per task of its 200/task panel. The 1600-prompt NIAH endpoints and intervals remain separate in Table 10; Llama NIAH is included in its Full-13 panel.

Task	Llama 200/task				OLMo 10/task		
	T	P	Y	C	T	P	Y
cwe	1.65	0.30	0.25	0.65	7.00	2.00	3.00
fwe	61.83	39.67	28.17	54.83	66.67	13.33	20.00
multikey_1	84.50	62.50	63.00	88.00	70.00	0.00	10.00
multikey_2	47.00	9.00	17.00	36.50	30.00	0.00	0.00
multikey_3	24.00	2.50	1.00	9.50	10.00	0.00	0.00
multiquery	95.25	94.88	96.88	95.25	72.50	2.50	2.50
multivalue	93.13	97.63	97.38	94.88	65.00	0.00	0.00
single_1	100.00	97.00	100.00	99.50	100.00	20.00	40.00
single_2	99.50	99.00	98.50	98.50	100.00	30.00	10.00
single_3	100.00	94.00	94.00	100.00	50.00	0.00	0.00
qa_1	64.00	55.50	54.50	69.50	50.00	30.00	10.00
qa_2	37.00	27.50	28.00	37.00	50.00	40.00	30.00
vt	79.60	55.60	63.70	76.10	8.00	0.00	2.00
Macro	68.27	56.54	57.11	66.17	52.24	10.60	9.81
T–P 95%		[10.32, 13.11]			[35.38, 47.88]		
T–Y 95%		[9.78, 12.54]			[36.17, 48.71]		
T–C 95%		[1.11, 3.08]			–		

Additional profiles at equal displacement. The step and plateau profiles vary the distribution of the same total displacement, retaining T's endpoints, outer bands and gain. Writing $\omega_k = \omega_k^N s^{-m_k}$, the step concentrates the transition into one jump with at most one fractional coordinate, chosen to match $\sum_k m_k$ to T. On OLMo, the plateau sets $m_k = 0.3568$ on coordinates 15–24 and $m_k = 1$ on 25–31; the exact plateau height matches T's sum. The deployed tables' displacement discrepancies are below 3.3×10^{-8}, and all frequencies remain strictly ordered. Table 9 uses the first 50 source-order prompts per task from each large panel, with paired T/C outputs restricted to those same 650 rows. Intervals use 20,000 task-stratified paired resamples (seed 20260925).

Table 9: **Quality varies across profiles with equal total displacement.** Full-13 means (%) and paired 95% profile–T differences (pp), with 50 prompts per task. All profiles share T's frozen weights, endpoints, outer bands, gain and total displacement ($s = 4$). These 650-prompt subsets have their own T/C means; the large-panel means remain in Table 7.

Model / length	Profile	Full-13	Profile–T (95% CI)
OLMo-2-1B / 16K	TailSpline	49.67	–
	C	50.16	$+0.49$ $[-1.47, 2.43]$
	Plateau	46.59	-3.08 $[-5.91, -0.28]$
	Step	53.74	$+4.07$ $[1.50, 6.63]$
Llama-3-8B / 32K	TailSpline	68.24	–
	C	64.90	-3.34 $[-5.43, -1.27]$
	Step	69.70	$+1.46$ $[-0.93, 3.86]$

On the same 650 OLMo prompts, MrRoPE-Pro scores 8.46%, reaggregated from its stored large-panel rows; every equal-displacement profile exceeds it by more than 38 points. On OLMo, plateau–T is -3.08 $[-5.91, -0.28]$ and step–T is $+4.07$ $[1.50, 6.63]$; on Llama, step–T is $+1.46$ $[-0.93, 3.86]$. Step–C differences are $+3.58$ $[1.12, 6.06]$ on OLMo and $+4.79$ $[2.12, 7.46]$ on Llama. Thus the tested equal-displacement profiles span several points on both models, with a resolved step–T advantage only on OLMo.

A separate geometric allocation between T's endpoints scores 0.15% versus T's 49.67% on the same OLMo subset, at shared gain. This comparison changes the outer bands and total displacement as well as interior placement; it complements the independent coarse-ramp comparison in Appendix C.1.

D.3 DIRECT BASELINES, MODEL FAMILIES AND SCALE

The Llama/OLMo NIAH-8 endpoint uses 200 inputs per task (1600 prompts), with its own task-stratified 10,000-resample intervals. These retrieval rows are subsets of the large T/P panels in Appendix D.2. OLMo's three-arm Full-13 comparison instead uses the first 10 inputs per task from its large T/P panel, giving 130 matched T/P/Y prompts. PPL uses 32 Proof-Pile arXiv test documents and 14 PG-19 test-book prefixes (Azerbayev et al., 2022; Rae et al., 2020), with direct prefill. Table 10 reports NIAH scores and PPL.

Llama's retrieval gains concentrate in multikey tasks; its multivalue/multiquery T/P/Y means are $94.19/96.25/97.13\%$. OLMo improves across all three retrieval groups.

Table 10: **Retrieval and language modeling with static $s = 4$ tables.** NIAH uses 200/task for Llama/OLMo and the 10/task Full-13 retrieval subset for Qwen/GLM. PPL uses 46 matched documents for Llama/OLMo and 5 for Qwen/GLM; corpora and tokenizers remain model-specific. NLL contrasts are document means. T/P/Y: TailSpline/MrRoPE-Pro/YaRN.

Model	Length	NIAH ↑ (%)			PPL ↓		
		T	P	Y	T	P	Y
Llama	32K	**80.42**	69.56	70.97	**4.1106**	4.1618	4.1495
OLMo	16K	**65.55**	5.36	6.27	**8.5362**	12.8270	12.1410
Qwen	128K	**79.69**	76.88	76.56	**2.3272**	2.3293	2.3278
GLM	128K	**52.81**	38.44	35.00	**27.3001**	29.9507	30.0744

Paired differences and marginal 95% intervals

Model	Contrast	NIAH (pp)	Interval	ΔNLL	Interval
Llama	T–P	+10.86	[9.06, 12.62]	-0.01238	[-0.01408, -0.01078]
Llama	T–Y	+9.45	[7.78, 11.11]	-0.00941	[-0.01088, -0.00803]
OLMo	T–P	+60.19	[58.28, 62.08]	-0.40724	[-0.43628, -0.38052]
OLMo	T–Y	+59.28	[57.36, 61.16]	-0.35227	[-0.38037, -0.32606]
Qwen	T–P	–	–	-0.00093	[-0.00198, 0.00044]
Qwen	T–Y	–	–	-0.00026	[-0.00111, 0.00110]
GLM	T–P	–	–	-0.09266	[-0.13809, -0.02102]
GLM	T–Y	–	–	-0.09678	[-0.14876, -0.02092]

Qwen2.5-3B-Instruct (Yang et al., 2024) and GLM-4-9B-0414 (Z.ai, 2025) have native reference length 32768, bases $10^6/10^4$ and $64/32$ rotary pairs. T/P use bands $[23, 40]/[17, 30]$. GLM retains its adjacent-pair convention, 64 rotating coordinates per 128-coordinate head, and 32 query/2 key/value heads. At 128K, each model uses 10 inputs/task (130 prompts), including the 5/task subsets. Retrieval uses Qwen and GLM 65536-token prefill chunks. PPL uses five matched InfiniteBench LongBook documents. Table 11 shows the complete task scores and paired intervals.

Llama-3-70B shares the 8B model's RoPE parameters, so its tables, band and gain are identical. The test uses the prequantized Llama-3-70B-Instruct checkpoint (Unsloth, 2024), with NF4 double quantization and BF16 computation. Tokenizer, chat template and special tokens match the 8B evaluation. Direct-prefill 32K evaluation uses the same 10/task subset and five matched PPL documents, without model-specific allocation search. T improves ten task means, ties one and loses two; T/P PPL is $2.4954/2.5075$.

Kanana-1.5-8B-Instruct-2505 has native length 32768, 64 rotary pairs and base 8,000,000 (Kakao, 2025). At 64K, T/P use $s = 2$, band $[20, 35]$ and gain $1 + 0.1 \ln 2$. Publisher-recommended YaRN uses factor 4.4, $64/2$-turn corrections and gain $1 + 0.1 \ln 4.4$. Thus T/P isolates allocation, while T/Y compares complete deployment recipes. Table 12 retains all 130 paired Full-13 prompts.

Table 11: Complete paired Full-13 scores (%, 10/task). These are the direct-baseline panels in Table 1; paired task-stratified 95% intervals appear below each model.

Task	Qwen			GLM			70B NF4	
	T	P	Y	T	P	Y	T	P
cwe	3.00	4.00	6.00	1.00	6.00	2.00	70.00	58.00
fwe	56.67	43.33	50.00	60.00	56.67	60.00	86.67	83.33
multikey_1	100.00	90.00	90.00	60.00	50.00	50.00	80.00	90.00
multikey_2	50.00	40.00	50.00	0.00	0.00	10.00	90.00	20.00
multikey_3	10.00	10.00	0.00	0.00	0.00	0.00	30.00	20.00
multiquery	95.00	92.50	87.50	50.00	30.00	45.00	92.50	87.50
multivalue	82.50	82.50	85.00	62.50	47.50	45.00	97.50	97.50
single_1	100.00	100.00	100.00	80.00	10.00	0.00	100.00	70.00
single_2	100.00	100.00	100.00	70.00	80.00	50.00	100.00	90.00
single_3	100.00	100.00	100.00	100.00	90.00	80.00	100.00	90.00
qa_1	20.00	30.00	30.00	20.00	30.00	20.00	60.00	70.00
qa_2	40.00	30.00	50.00	10.00	0.00	0.00	20.00	10.00
vt	80.00	56.00	58.00	2.00	0.00	0.00	96.00	6.00
Macro	64.40	59.87	62.04	39.65	30.78	27.85	78.67	60.95
T–P 95% (pp)		[-0.41, 9.49]			[4.04, 13.69]		[11.65, 23.78]	
T–Y 95% (pp)		[-2.78, 7.49]			[6.37, 17.32]		–	

Table 12: Kanana at 64K: complete Full-13 scores (%, 10/task) and paired within-task intervals. T/P use $s = 2$ and Y the publisher-recommended factor 4.4.

Task	T	P	Y
single_1	100.00	100.00	100.00
single_2	100.00	100.00	100.00
single_3	100.00	100.00	100.00
multikey_1	80.00	80.00	80.00
multikey_2	60.00	50.00	50.00
multikey_3	70.00	50.00	40.00
multivalue	65.00	57.50	60.00
multiquery	22.50	32.50	25.00
vt	64.00	70.00	82.00
cwe	43.00	51.00	13.00
fwe	90.00	93.33	93.33
qa_1	80.00	70.00	70.00
qa_2	70.00	60.00	40.00
Macro	**72.65**	70.33	65.64
T–P: $+2.32$ pp, 95% interval $[-3.13, 8.04]$			
T–Y: $+7.01$ pp, 95% interval $[0.21, 13.81]$			

D.4 Transfer from retrieval to natural tasks

Natural-QA evaluates whether the synthetic-task gains transfer to passage QA. Its frozen pool has 631 questions from 524 source contexts across HotpotQA, 2WikiMQA, Qasper, NarrativeQA and MultiFieldQA-en (Bai et al., 2024). Membership uses OLMo's > 4096-token threshold and is retained under Llama retokenization: actual inputs span 4110–16319 and 3694–16304 tokens respectively. Llama has $316/315$ questions within/above its native window. Generation caps are 32 for HotpotQA/2Wiki, 128 for Qasper/NarrativeQA and 64 for MultiFieldQA. Llama uses 8192-token prefill chunks; OLMo uses direct prefill. Whole-response normalized token F1 is question-averaged within task, then task-averaged. Source clusters are resampled within task with 20,000 draws, preserving these weights and all outputs.

OLMo improves over P on every task; Llama's task directions are mixed. Llama's within/above-native T–P contrasts are $+0.70/-0.94$ points; OLMo's 4K–8K/8K–16K contrasts are $+6.40/+1.51$. Of these four strata, only OLMo's 4K–8K interval excludes zero. The same question pool on 70B NF4 gives T/P $49.28/48.39\%$; Table 14 reports the paired intervals.

Table 13: Complete Natural-QA631 task scores (F1, %). Each model uses the same question pool across T/P/Y.

Task	Llama-8B			OLMo		
	T	P	Y	T	P	Y
2wikimqa	36.96	35.19	35.61	26.29	21.55	24.72
hotpotqa	52.98	53.09	56.07	34.65	28.63	28.27
multifieldqa_en	51.91	51.33	50.65	35.80	34.49	32.29
narrativeqa	22.82	23.77	26.09	9.85	9.51	11.52
qasper	40.73	41.04	40.33	18.03	13.94	14.56

LongBench v2 adds 117 complete Llama inputs at 8K–32K using the direct-answer multiple-choice adapter, static $s = 4$ tables and matched caps. Correct parsed options score one; incorrect or unparseable answers score zero. T/P answer $41/36$ correctly ($35.04/30.77\%$), difference $+4.27$ $[-3.39, 11.97]$ points under source-context resampling. The single-document subset favors P ($34.43/37.70\%$).

Complete books and source transfer. Qwen and GLM use InfiniteBench English-QA (Zhang et al., 2024) with 35 questions from seven complete books each, five questions/book, no input truncation and a 40-token answer cap. Actual lengths span 100907–127946 for Qwen and 100844–127811 for GLM. All three arms retain their $s = 4$ tables and 65536-token prefill chunks. GLM's second pool has 77 questions from 15 books disjoint from pool I; inputs span 69660–95486 tokens with the same table and decoder. Separate means and source-context intervals test transfer to new sources.

Kanana's 128K comparison uses T with $s = 4$ and YaRN's factor-4.4 recipe. It includes all 118 complete prompts fitting the tokenizer and budget, from 23 books, spanning 69680–130852 tokens with a 40-token output cap. T/Y official F1 is $19.59/17.85\%$; source-context-equal F1 is $19.63/18.04\%$. Kanana's 64K and 128K conditions use separately constructed T tables.

Book QA uses official normalized token F1 against accepted references, retains every answer and resamples whole source books in 20,000 paired draws. Table 14 gives marginal and simultaneous intervals for the natural-QA and three-arm book panels. OLMo Natural-QA and Qwen T–P book-QA have positive marginal intervals, but their simultaneous intervals span zero; the GLM book comparisons also span zero.

Table 14: Natural-QA and book-QA contrasts (pp). Paired source-context bootstrap uses 20,000 draws. The last column gives Bonferroni simultaneous intervals for the two comparisons within each three-arm panel; the 70B panel has one comparison. II identifies GLM's second book pool.

Model	Contrast	Difference	Marginal 95%	Simultaneous 95%
Llama-8B	T–P	+0.20	[-1.48, 1.86]	[-1.74, 2.12]
Llama-8B	T–Y	-0.67	[-2.87, 1.38]	[-3.22, 1.63]
OLMo	T–P	+3.30	[0.39, 6.37]	[-0.02, 6.83]
OLMo	T–Y	+2.65	[0.04, 5.49]	[-0.35, 5.92]
Llama-70B NF4	T–P	+0.89	[-0.60, 2.44]	[-0.60, 2.44]
Qwen	T–P	+3.24	[0.26, 7.51]	[-0.07, 7.89]
Qwen	T–Y	+0.69	[-3.66, 5.21]	[-3.99, 5.86]
GLM	T–P	+3.33	[-3.28, 8.98]	[-4.17, 9.70]
GLM	T–Y	+1.29	[-6.53, 8.61]	[-8.05, 9.29]
GLM II	T–P	+4.00	[-1.73, 9.68]	[-2.57, 10.61]
GLM II	T–Y	+3.37	[-1.58, 8.42]	[-2.38, 9.17]

Matched retrieval augmentation. HELMET-RAG tests the same questions as distractor context grows. Table 15 gives the length-dependent scores and the corresponding evidence-only comparison. The HELMET-RAG-derived panel uses revision `af609c4` of the short-RAG protocol (Yen et al., 2025): seed 20260919 selects 64 distinct questions each from NQ, TriviaQA, PopQA and HotpotQA, balanced across six evidence positions (three for HotpotQA). Matched $50/105/220$-passage pools retain the same evidence and nested distractors at $8/16/32$K. Ranked distractors remain in order;

complete supporting passages occupy the assigned positions, and the largest whole-passage prefix fitting the budget is retained.

Prompts use two disjoint demonstrations and the chat template. Static $s = 4$ T/P decoding stops at EOS, newline or 20 tokens. Substring exact match (SubEM) is task-averaged, with 32K primary and paired source-group intervals. Original RoPE at 8K and the evidence-only arms provide the corresponding context comparisons.

Table 15: HELMET-RAG-derived Llama SubEM (%) on 256 paired questions (Appendix D.4). These rows use the original prompt without the added answer prefix. Prompt-token ranges at $8/16/32$K are 3,966–7,937, 8,363–15,851, and 19,097–31,797; evidence-only uses 518–1,232.

Condition	Window	Native	T	P	T–P 95% (pp)
Long	8K	55.47	63.67	61.72	[-2.41, 6.37]
Long	16K	–	50.39	49.61	[-2.95, 4.67]
Long	32K	–	48.05	47.27	[-3.30, 4.80]
Evidence only	8K	57.42	37.11	41.41	[-9.64, 0.97]

Attention gain and answer format. Setting $g = 1$ leaves each T/P frequency table unchanged. We evaluate the same 650 native-8K Full-13 prompts and 256 evidence-only questions, including all empty outputs (Table 16). The native-length T/P gains are $4.41/6.31$ points, recovering $90.5/82.6\%$ of their respective gaps to Native on this panel. FWE scores rise to 82.00% for both, with $5/6$ empty outputs, versus Native's 87.33% and none. The evidence-only contrasts under the original protocol instead have negative point estimates and intervals spanning zero. Removing gain substantially recovers the RULER loss but leaves this short-QA gap in place. At unit gain, the comparison with Native still changes both support and interior allocation.

With standard T gain $1 + 0.1 \ln 4$ and Native gain one, keeping the same questions, contexts and 20-token, EOS/newline stopping rule, prepending `Answer:` to both assistant continuations gives Native/T evidence-only SubEM of $64.84/62.50\%$: T–Native narrows from -20.31 to -2.34 points $[-5.08, 0.39]$. The corresponding long-8K scores are $62.11/66.41\%$, a gap of $+4.30$ $[0.00, 8.59]$ points versus the original $+8.20$. These prefix contrasts use 20,000 paired source-group resamples within task with fixed task weights. The protocol change affects both gaps.

The Full-13 intervals use 20,000 paired resamples within each task, holding task weights fixed; the original evidence-only gain intervals resample questions.

Table 16: **Gain interventions at fixed frequency tables.** Llama-3-8B, static $s = 4$; standard T/P/Y use $g = 1 + 0.1 \ln 4$. Scores are percentages. Full-13 uses 650 prompts at 8K; evidence-only uses 256 questions under the original protocol without the added answer prefix. FWE counts cover 50 inputs.

Arm	8K Full-13	FWE	Empty	Evidence-only
Native	90.03	87.33	0	57.42
T	85.16	9.33	45	37.11
T ($g = 1$)	89.56	82.00	5	34.77
P	82.39	4.00	48	41.41
P ($g = 1$)	88.70	82.00	6	39.84
Y	83.78	2.00	49	35.55

Unit-gain minus standard-gain contrasts (pp), paired 95% intervals

Arm	8K Full-13	Evidence-only SubEM
T	+4.41 [2.87, 5.93]	-2.34 [-7.81, 3.13]
P	+6.31 [4.69, 7.90]	-1.56 [-7.03, 3.91]

D.5 Behavior at higher extension factors

Llama uses $s = 16$ at 128K and Qwen $s = 8$ at 256K, with T/P sharing weights, public band, amplitude and prompts within each condition. Llama Full-13 and Proof-Pile PPL use direct prefill, as does Qwen's single-NIAH panel.

Table 17: Higher-factor task profiles. Generation scores are percentages; PPL uses its own document set. Intervals are paired 95% intervals.

Model	Endpoint	n	T	P	T–P interval
Llama 128K	Full-13	130	22.05	1.28	[15.96, 25.64] pp
	Proof-Pile PPL	10	2.5870	2.6714	[−0.03505, −0.02918] NLL
	En.QA F1	41	13.21	20.17	[−11.87, −2.07] pp
	En.Dia accuracy	100	12.00	8.00	[−1.00, 10.00] pp
Llama-70B NF4 128K	Proof-Pile PPL	10	2.2449	2.3129	–
Qwen 256K	Single-NIAH	15	93.33	40.00	[33.33, 66.67] pp
	LongBook PPL	5	30.3437	23.9283	[−1.0561, 1.7769] NLL

Llama Full-13 uses ten inputs/task. Qwen uses five inputs for each of single-1/2/3; this three-task endpoint is distinct from NIAH-8. Llama's natural panels retain complete 100K–131K inputs and official InfiniteBench metrics; En.QA has eight source books. Generation intervals resample tasks or source contexts and PPL intervals resample documents. The 70B NF4 PPL comparison uses ten documents, 32768-token prefill chunks and gain $1 + 0.1 \ln 16$; no document interval is supplied. Llama's Full-13 and PPL improve while book-QA F1 decreases; Qwen's single-needle retrieval improves while its five-book PPL is higher.

D.6 Baseline implementation and correspondence with reported results

Frequency tables. Our equation-level MrPro implements the published progressive profile (Tian et al., 2026). Across all 14 model/scale configurations checked, it agrees with the conference-released constructor to within one FP32 ulp (maximum $|\Delta \log \omega| = 1.17 \times 10^{-7}$); bands and gain agree. The Llama band is $[18, 35]$, whose upper endpoint agrees with the reference in their Appendix B.1. Qwen2.5-3B uses $[23, 40]$, also specified there. At Qwen 128K, the checkpoint, $s = 4$, band and evaluation length match their reported setting. Our P/Y means are $59.87/62.04\%$ on 10/task, compared with their reported $53.2/50.1\%$.

Published implementation and checkpoint identity. For the separate diagnostic in Table 18, we directly invoke the released MrPro constructor, rotary forward, generation and scoring functions. Extension arms use $s = 16$, base 500000, an 8K construction reference, shared gain and FlashAttention2/BF16. Our released-code runs use sampling with temperature 0.7, top-$p = 0.9$, a 30-token cap, and EOS/newline stopping. Inputs are retained source-order prompts from our RULER protocol in place of the released HF prompt package: all thirteen tasks, 10 inputs/task (130 total), with paired per-row seeds. Fig. 2 and Table 1 report the separate greedy-decoding comparison with static $s = 4$ Llama tables across $8/16/32$K.

Section 4.1 and Table 2 of Tian et al. (2026) name Llama-3-8B(-Instruct), while its Figure 3(a) is labeled Llama-3.1-8B; the released example also names a Llama-3.1 model path. These are distinct checkpoints: Llama-3 is natively 8K, whereas Llama-3.1 supports 128K with its own RoPE scaling (Grattafiori et al., 2024). Table 18 uses the 8K-native Llama-3 checkpoint. Replacing Llama-3.1's table with an 8K-referenced $s = 16$ construction at $32/128$K is a native-window intervention, which addresses a different question from extending Llama-3.

Observed correspondence. On our 32K panel, released-code Y/P score $56.12/48.15\%$, below the published $92.7/94.3\%$, with their ordering reversed. T scores 55.90%: T–P is $+7.74$ $[1.81, 13.95]$ points and T–Y is -0.22 $[-4.15, 3.86]$. On Llama-3 at 128K, T–P is $+18.08$ $[13.33, 22.88]$.

Table 18: Released-code diagnostic on our retained Full-13 panel (10/task), using the 8K-native Llama-3 checkpoint; empirical means (%). Extension arms use $s = 16$ at both evaluation lengths. Published values are from Tian et al. (2026), Table 2(a), under a different evaluation protocol; dashes denote unmeasured or unreported cells.

Our checkpoint	Arm	Our 32K	Published	Our 128K	Published
Llama-3	Y	56.12	92.7	–	79.9
	P	48.15	94.3	0.51	86.6
	T	55.90	–	18.59	–

E Native-Window Allocation and Context Use

At fixed native support, NCP increases the measured benefit from earlier history. The construction and same-target comparison below connect that language-modeling result to native-context task performance.

E.1 Reference objective and numerical construction

NCP reallocates the interior frequencies within the native support while fixing both sampled endpoints. Let $\phi_k = (L-1)\omega_k$ and $\omega'_k = \omega_k e^{-u_k}$. In Eq. (1), this gives $z'_k = z_k + u_k/R$; zero endpoint shifts preserve a and R. The reference problem is a binary target–distractor comparison with unit amplitudes, aligned target phase and uniformly random distractor phase:

$$r(\theta) = \frac{1}{2\pi}\int_0^{2\pi} \log(1+\exp(\cos\psi - \cos\theta))\, d\psi, \tag{33}$$

$$\mathcal{R}(\phi) = 2\int_0^1 (1-t) r(\phi t) dt. \tag{34}$$

The triangular distance weight is the distribution of absolute separation between two uniform positions on a continuous unit interval. The aligned reference comparison supplies a public design objective. The Transformer task scores below evaluate the resulting fixed construction.

The reference optimization is

$$\min_u \sum_k \left[\mathcal{R}(\phi_k e^{-u_k}) + \frac{9}{8}u_k^2\right], \qquad \omega'_k = \omega_k e^{-u_k}, \tag{35}$$

with fixed $\lambda = 9/4$ (quadratic coefficient $\lambda/2 = 9/8$), subject to

$$u_0 = u_{K-1} = 0, \quad 0 \le u_k \le 2/9, \qquad \log\frac{\omega'_k}{\omega'_{k+1}} \ge \frac{1}{2}\log\frac{\omega_k}{\omega_{k+1}}. \tag{36}$$

The endpoint constraints keep support fixed; the gap condition preserves ordering and at least half each original adjacent log gap. Writing $r(\theta) = a_0 + \sum_{j\ge1} a_j \cos(j\theta)$ gives

$$\mathcal{R}(\phi) = a_0 + \sum_{j\ge1} a_j \operatorname{sinc}^2(j\phi/2), \qquad \operatorname{sinc}(x) = \frac{\sin x}{x}, \quad \operatorname{sinc}(0) = 1. \tag{37}$$

This follows by integrating each cosine against $2(1-t)$. We solve Eq. (35) using the finite approximation $\mathcal{R}_{32}(\phi) = \hat{a}_0 + \sum_{j=1}^{32} \hat{a}_j \operatorname{sinc}^2(j\phi/2)$, with coefficients computed on a 1024-point angular grid. With $q_{32} = d\mathcal{R}_{32}/d\log\phi$, an interior scalar solution satisfies $\lambda u = q_{32}(\phi e^{-u})$. Bisection solves the scalar stationary equations, using zero shift when the native gradient is nonpositive. The two endpoint shifts are reset to zero. The implemented 32-mode scalar objective is strictly convex. For $f(x) = \operatorname{sinc}^2(x/2)$, its log-curvature is $h(x) = 2\cos x - 6\sin x/x + 8(1-\cos x)/x^2$ and $|h(x)| \le 4$: for $0 \le x \le 4$, the integral representation gives $|h(x)| \le x/3 + x^2/6$; for $x \ge 4$, it gives the bound $\sqrt{4 + 36/x^2} + 16/x^2 \le 3.5$. The computed coefficients satisfy $\sum_{j=1}^{32} |\hat{a}_j| < 0.554403$, so scalar curvature is at least $9/4 - 4\sum_{j=1}^{32}|\hat{a}_j| > 0.03238$. Thus the stationary root, or the zero-shift boundary when its derivative is nonnegative, is the unique box minimizer. On the deployed

OLMo grid the log-gap constraints are inactive, so these scalar minimizers also solve the constrained finite-series problem.

The deployed grid has 64 pairs, base 500000 and native length 4096. Its maximum relative frequency reduction is 16.7164% and its minimum log-gap ratio is 0.7681, satisfying the half-gap constraint. The table and unit gain are fixed across layers and inputs; construction reads no checkpoint weights, activations or evaluation outputs. The source archive includes the constructor and an exact FP32 table reproduction check.

E.2 Language modeling and benefit from earlier context

We evaluate 128 nonoverlapping windows from 103 documents in the PG-19 and Proof-Pile test corpora. Each 4097-token window defines 4096 next-token inputs. We score only its final 256 target tokens, once with full history and once with 512 preceding tokens plus the same teacher-forced targets. Shared tokens retain identical relative positions; the checkpoint, targets and score positions are paired. Loss is averaged within document and then equally over documents. Paired document intervals use 20000 draws, seed 20260915.

Writing $B_M = \mathrm{NLL}_{\mathrm{recent},M} - \mathrm{NLL}_{\mathrm{full},M}$, the paired difference $B_{\mathrm{NCP}} - B_{\mathrm{Native}}$ measures the additional benefit from earlier context at identical targets.

Table 19: Same-target NLL (nat/token), document-equal across 103 documents and 128 windows. Context benefit is recent-history minus full-history NLL; positive benefit differences favor NCP.

Context	Native	NCP	NCP–Native	95% interval
Recent history	3.009922	3.005077	-0.004845	[-0.005670, -0.004007]
Full history	2.915269	2.902483	-0.012786	[-0.015168, -0.010440]
Context benefit	0.094653	0.102594	+0.007941	[0.005529, 0.010346]

E.3 Native-context task results

The OLMo native 4K development panel contains 60 source-order inputs per RULER task, 780 total, without content padding. NCP and original RoPE share the checkpoint, unit gain, inputs, BF16 precision, greedy decoder, batch 1 and 4096-token prefill chunks. Their scores are $70.7821/69.3739\%$; the paired task-equal difference is $+1.4081$ points, interval $[0.2009, 2.5833]$.

Table 20: Complete native-window Full-13 scores ($\%$). Development/separate OLMo panels use $60/10$ inputs per task at 4K.

Task	Development (60/task)		Separate (10/task)	
	Native	NCP	Native	NCP
cwe	7.83	4.67	2.00	1.00
fwe	57.78	60.00	53.33	56.67
multikey_1	81.67	78.33	90.00	90.00
multikey_2	91.67	95.00	100.00	100.00
multikey_3	58.33	61.67	80.00	80.00
multiquery	89.17	90.83	85.00	92.50
multivalue	95.42	98.33	97.50	100.00
single_1	100.00	100.00	100.00	100.00
single_2	100.00	100.00	100.00	100.00
single_3	98.33	98.33	100.00	100.00
qa_1	68.33	61.67	70.00	70.00
qa_2	35.00	30.00	50.00	60.00
vt	18.33	41.33	28.00	48.00

A separate OLMo Full-13 panel has ten source-order records per task (130 pairs), fixed before evaluation. This panel reuses the same NCP table and unit gain; its task-equal difference is $+3.2564$ $[0.8077, 5.9231]$ points. The interval uses 20000 paired within-task draws, seed 20260916. Table 20 reports both task panels separately.

Native-QA contains 99 complete inputs within OLMo's native window: 24 2WikiMQA, 5 HotpotQA and 70 Qasper questions. Both arms share prompts, BF16 weights, greedy decoding and task-specific generation caps. Normalized whole-response token F1 is averaged over questions within task, then equally over tasks. Paired source-cluster resampling within task retains question weighting, using 20000 draws and seed 20260917.

The task F1 scores (Native/NCP, %) are $37.96/35.05$ on 2WikiMQA, $45.00/45.00$ on HotpotQA and $35.27/35.32$ on Qasper.

Native/NCP F1 is $39.4116/38.4554\%$, difference -0.9562 $[-4.4410, 1.5666]$ points.

F Learning and Extrapolation with Changed Allocations

Cosh improves long-range prediction in training and adaptation.

F.1 Learning from scratch at a fixed rotary budget

We report allocation at one fixed rotary budget in a 432M-parameter transformer following the MLA architectural pattern (DeepSeek-AI, 2024a), with 24 layers, width 1024, 16 heads, and per-head width 64. Its decoupled rotary subspace has d_{rope}=32 (K=16 frequencies); the remaining d_{nope}=32 dimensions are non-rotary, and the shared KV latent rank is 256. The model trains from scratch on 500M FineWeb-Edu tokens (Penedo et al., 2024) at L_{train}=8192 and base 500K over seeds 42, 43, and 88. The recorded training recipe uses midpoint Geo and unanchored midpoint Cosh at $\tau = 1.414$, an empirical effective-dimension-128 convention (Appendix B.4).

Evaluation corpus and aggregation. Evaluation uses a common held-out FineWeb-Edu stream of approximately five million GPT-2 tokens. At each length (8/16/20/24/28/32K), all methods and seeds share eight text chunks, each scored on all $L - 1$ next-token targets. Per-seed PPL is the exponential of the mean chunk NLL; reported means and sample standard deviations are across the three training seeds.

For the supporting overlay, let $\lambda_k = 2\pi/\omega_k$ and $t_k = \text{clip}((\lambda_k - L_{\text{train}})/(31L_{\text{train}}), 0, 1)$. The MLA wavelength-blend operator sets $\omega'_k = \omega_k/[1 + t_k(s-1)]$: it preserves wavelengths below the training length, applies the full factor s above $32L_{\text{train}}$, and interpolates between them.

At 16K, three-seed mean±SD PPL is 138.8 ± 5.5 for Geo and 95.6 ± 4.1 for Cosh; Cosh is lower at all five extended lengths (16–32K) in each of the three seeds. Mean 8K PPL is $35.4/35.8$. With the common wavelength blend, 16K PPL is 117.9 ± 6.5 versus 71.1 ± 4.1; Cosh plus blend gives the lowest mean PPL at each extended length (Fig. 3).

F.2 Shared-start continuation

Table 21: 750M continued-pretraining check (seed 42; 2K→ 4K; 500M continuation tokens). Passkey retrieval is teacher-forced NLL-gap and is saturated at 8K; AR exact is greedy autoregressive exact match over 40 trials per length, without requiring a following EOS.

Metric	Geo	Cosh
PPL@2K	25.9	26.2
PPL@4K	22.0	22.3
PPL@8K	23.4	**19.6**
PPL@16K	45.1	**24.4**
NLL-gap retrieval @8K	100%	100%
Answer-token exact @8K	0%	**77.5%**

The 750M pair continues a common 2K geometric checkpoint for 500M tokens at 4K, with 3% passkey and 10% downstream mix (seed 42). Cosh uses $\tau = 1.5$, $K = 32$, inclusive nodes $u_k = k/(K-1)$ and $\omega_k = \exp[-(K-1)\log(500000)Q_{1.5}(u_k)/K]$, retaining Geo endpoints. Both arms update all weights. Greedy answer-token exact match generates the expected passkey tokens and closing marker; it does not require a following EOS.

For MLA and 750M continuation, AdamW uses $(0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, clipping 1, no dropout, peak LR 2×10^{-4}, 2% linear warmup and cosine decay to 0.1 of peak. Batch sizes are 6 (MLA) and 14 (continuation: microbatch 7, accumulation 2). Text packing is contiguous without cross-document masking. MLA uses unanchored midpoint Cosh; the continuation fixes sampled endpoints.

F.3 Matched adaptation

The Llama-3-8B-Instruct comparison applies matched rank-64, alpha-128 Q/K/V/O LoRA for 300 steps, seed 42, learning rate 10^{-4}, BF16 and 8192-token sequences. Native-LoRA uses the original frequency grid; Cosh-LoRA uses midpoint Cosh at $\tau = 1.414$. Rows, order, optimizer, schedule, adapter capacity and evaluator match. Adaptation uses 7476 length-filtered LongAlpaca training rows (Chen et al., 2024) and 152 validation rows, with effective batch 8. Evaluation uses 24 frozen temporal packs from arXiv, the Federal Register and Stack Overflow; losses are pooled within domain and averaged equally across domains before exponentiation. All 24 packs favor Cosh at both extended lengths. Table 22 reports PPL across native and extended lengths.

Table 22: **Cosh extrapolation under matched LoRA adaptation.** PPL of Llama-3-8B-Instruct; both arms share rows, order, optimizer, schedule, adapter capacity and evaluator.

Study	Length	Native-table LoRA	Cosh-table LoRA
Llama 8B matched LoRA	8K	6.817	10.068
	16K	108.958	24.068
	32K	991.475	127.911

Matched OLMo routing continuation. OLMo-2-0425-1B-Instruct uses fixed standard-grid Cosh with $\tau = 2$, $b = 500000$ and rank-64, alpha-128 Q/K/V/O LoRA. Native/Cosh share checkpoint, training rows, budgets and evaluation inputs. Starting from each arm's 300-step routing parent, a matched 100-step query-gap continuation uses peak LR 5×10^{-5} and 20 warmup steps; a 32-step answer-plus-EOS stage uses 10^{-5} and 4 warmup steps. Both use seed 20260728, AdamW $(0.9, 0.95)$, zero weight decay, clipping 1, BF16, batch 8 (micro 4, accumulation 2), and cosine LR decay.

Each cycle has two counterfactual-routing updates and one natural-text update. The 1024 training pairs and 128 calibration pairs use different query templates. Examples contain eight key–value blocks; changing the queried value changes the target answer. Answer cross-entropy is combined with a counterfactual margin of 1, weighted by 0.5, on positions where correct and swapped answer tokens differ. The final stage supervises three answer tokens and immediate EOS, with EOS weight one; prompt and padding tokens carry no answer loss. Natural replay uses ordinary next-token loss.

Position schedule. Physical inputs stay at most 4096 tokens. Let q be the semantic query-block start, a the answer start, h the end-exclusive source span and n the active sequence length. Both counterfactual variants use $p_i = i + o\,\mathbf{1}\{i \geq q\}$. Each query-gap update uses four pair offsets: one alternating between 0 and $[1, 4096]$, one in $[4097, 8192]$, and two in $[8193, 12289]$. In the final stage, realized gap targets instead comprise one alternating between contiguous and $[4096, 8191]$, one in $[8192, 12287]$, and two in $[12288, 16079]$. For a noncontiguous target g, the offset is

$$o = \min\{g - (a - h),\ 12289,\ 16384 - \min(n, 4095)\}.$$

The resulting gap must remain in its target band and active position IDs below 16384. Natural replay stays contiguous.

Evaluation uses real contiguous numeric-NIAH prompts, 100 per length, and requires the complete expected answer followed by terminal EOS. Native/Cosh scores are $95/100$ versus $100/100$ at 4K, $18/100$ versus $98/100$ at 8K, and $0/100$ versus $60/100$ at 16K. All outputs terminate with EOS. The Cosh long-gap subsets score $48/50$ at 8K and $31/66$ at 16K. On a separate native-4K RULER panel (13 tasks, 20 fixed prompts per task), the final Native/Cosh adapters score $58.27/26.44\%$.

F.4 WEIGHT ADAPTATION WITH A FIXED TAILSPLINE TABLE

Independent OLMo-2-0425-1B-Instruct and Llama-3-8B-Instruct runs adapt weights while retaining their TailSpline FP32 tables and gains. OLMo uses $s = 4$, Llama $s = 2$, both at 16K; the Llama table therefore differs from the $s = 4$ frozen comparison in Section 5.1. Each run starts from the released model and applies rank-32, alpha-32 LoRA to Q/K/V/O and MLP projections, with zero dropout, for 200 updates with AdamW $(0.9, 0.95)$, zero weight decay, clipping 1, BF16 and seed 20260912. LR rises linearly over 20 updates to 2×10^{-5} and then stays constant. The objective combines long-text next-token loss, long-answer supervision, short-answer replay and KL to Native with weights $1/1/0.25/0.25$. Each update accumulates one 16K PG-19 text example, one intact long-answer example of at most 16K tokens, and one UltraChat replay of at most 4K tokens; each model uses its own tokenized pools. The long-answer mixture has $36/149/15$ examples from LongAlign/LongAlpaca/LongCite. Python shuffles within these pools with seeds $20261610/11/12$ and the source order with 20261600; text and replay each take the first 200 rows after shuffling their respective training pools with seeds 19384 and 5000. KL is $\mathrm{KL}(p_{\mathrm{Native}} \| p_{\mathrm{adapter}})$, averaged over up to 128 evenly spaced answer-prediction positions on the replay sequence. The Native teacher uses original weights, its native grid and unit gain; answer supervision includes terminal EOS. The accompanying source archive retains the pool sizes, selection recipe and exact table values.

Table 23 compares each adapter with its own static table on the same 13×10 RULER prompts. Full-13 weights tasks equally; intervals use 20,000 paired row resamples within task. Both point estimates increase, while both intervals include zero.

Table 23: **Adaptation on fixed TailSpline tables at 16K.** Static and adapted Full-13 scores are percentages; differences and paired 95% intervals are in percentage points.

Model/table	Static	Adapted	Difference	95% interval
OLMo 1B, $s = 4$	52.24	54.99	+2.74	$[-1.87, 7.53]$
Llama 8B, $s = 2$	85.71	87.15	+1.45	$[-2.21, 5.38]$

G FURTHER RELATED WORK

Extension recipes and base selection. YaRN interpolates slow frequencies, keeps fast ones and blends the channels between them (Peng et al., 2024). MrRoPE casts extension methods as radix conversions and introduces uniform and progressive profiles, of which the progressive MrRoPE-Pro performs best (Tian et al., 2026). Base-selection methods keep the geometric allocation: scaling laws and base bounds relate the base to the attainable context length (Liu et al., 2024; Xu et al., 2024), and the choice of base relative to the training or target length shapes the trade-off between interpolation and extrapolation (Oka et al., 2026).

Data-dependent and learned tables. DPE uses task probes and Q/K norms to select effective distances and key dimensions, then remaps relative positions while preserving a local window (Lu et al., 2025). LeRoPE learns frequency log-scales jointly with the weights and analyzes their attention gradients; its transfer experiment trains a new model with a learned table (Karypis et al., 2026). Our frozen constructions read only public parameters, and our Cosh experiments prescribe a table while the weights learn.

Positional structure beyond one static axis. FoPE changes the Fourier components of positional features (Hua et al., 2025), and GRAPE connects group generators, frequencies and norms (Zhang et al., 2026). Selective RoPE and RePo learn input-dependent rotations or positions (Movahedi et al., 2026; Li et al., 2026), whereas our tables are fixed for each model configuration and target scale and shared across inputs. VideoRoPE assigns low frequencies to temporal channels, while multimodal HoPE sets temporal-channel frequencies to zero (Wei et al., 2025; Li et al., 2025); our comparisons vary frequency placement within one positional axis at fixed endpoints.