TRAINING-FREE LONG-CONTEXT EXTENSION BY WARPING ROTARY RELATIVE DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary position embeddings (ROPE) are the default positional mechanism in modern language models, but training-free context extension remains brittle: global position or frequency rescaling can bring distant tokens into the trained range while distorting the local geometry the model learned during pretraining. We introduce WARP (Warped Rotary Positions), a training-free method that keeps the pretrained rotary spectrum fixed and instead applies a monotone nonlinear warp to the relative query–key distance used inside rotary attention. WARP is exactly ROPE at native context length, approximately preserves local distances, and compresses only the far field into the model's trained phase budget. Because the warp is a nonlinear function of the pairwise distance $i - j$, it cannot be represented by a standard per-token position cache, which makes WARP a distinct class of rotary extension from position or frequency rescaling. The evaluated configuration adds a small frequency, temperature, and density calibration and applies the warp through a block-anchored approximation. In ten-seed exact-retrieval stress tests, WARP reaches 50/50 with zero observed variance on Qwen2.5-3B-Instruct and Mistral-7B-v0.3 at both 32k and 64k and on Qwen2.5-0.5B-Instruct at 64k, and reaches 49.20/50 on Llama-2-13B at 16k; Llama-2-13B at 32k remains weak. On official RULER v1 at 64k with Qwen2.5-3B-Instruct, WARP achieves 70.82% macro accuracy, compared with 62.92% for unmodified ROPE, 71.22% for NTK, 73.31% for YaRN, and 73.44% for MrRoPE-Pro. WARP is therefore competitive on broad RULER rather than dominant; its advantage is concentrated in retrieval-heavy behavior, where it is best or tied-best among the strongest methods on several RULER retrieval and QA components. The hardest setting we observe is Llama-2-7B exact retrieval at 32k, where base WARP and every baseline score 0/50; an exploratory variant that adds head routing and a small trained readout adapter recovers up to 31/50. These results suggest that relative-distance warping is a promising, model-adaptable route to long-context extension without finetuning, while exposing architecture-specific failure modes that we report explicitly.

1 INTRODUCTION

Long-context language models are useful only when they can both read distant tokens and preserve their short-range language-modeling behavior. Rotary position embeddings (ROPE) (Su et al., 2024) make this problem especially sharp. A ROPE model is pretrained on a finite range of rotary phases; running it far beyond that range exposes it to query–key phase relations that never occurred during pretraining. The usual training-free remedy is to change the effective position or frequency scale, as in positional interpolation (Chen et al., 2023), YaRN (Peng et al., 2024), LongRoPE (Ding et al., 2024), and MrRoPE-Pro (Tian et al., 2026). These methods are attractive because they require no additional training, but they rely on global or dimension-wise transformations of the rotary coordinates. They can therefore trade one failure mode for another, avoiding far-field extrapolation at the cost of perturbing local distances or changing the model's learned frequency mixture.

This paper frames long-context extension differently. We want the model to see exactly native ROPE behavior when the requested sequence length is native, nearly native geometry for local dependencies, and a compressed but still ordered geometry for distant context. We propose WARP: instead of replacing the rotary spectrum, WARP warps the relative distance $d = i - j$ between each query and

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

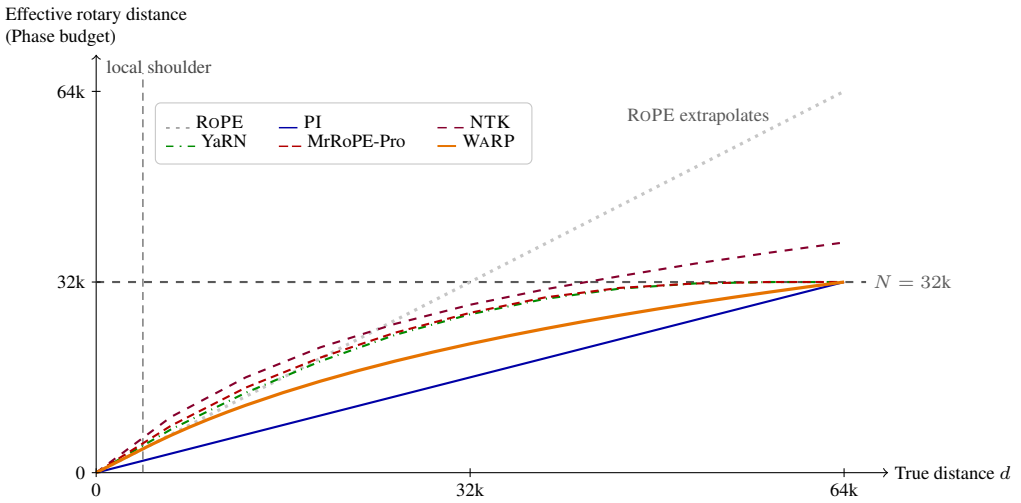


Figure 1: Geometry of training-free rotary extension methods. ROPE extrapolates directly beyond the native phase budget, while PI (blue) compresses all distances uniformly. Frequency-rescaling methods such as NTK (purple), YaRN (green), and MrRoPE-Pro (red) are shown as representative phase-equivalent compression profiles because their effective distance depends on rotary frequency band. WARP (orange) differs by keeping the pretrained frequency spectrum fixed and applying a nonlinear pairwise relative-distance warp: local distances remain near identity, while only the long-distance tail is compressed so the farthest 64k pair lands near the 32k native budget.

key before the rotary phases are evaluated. The pretrained frequencies remain fixed; only the distance supplied to them changes.

The core idea is simple. For short distances we use $f(d) = d$; for long distances we use a monotone logarithmic shoulder that maps the farthest requested distance back into the model's native phase budget. The resulting method is the identity at native length, local near the query, and compressed in the tail. It also has a useful mathematical property: a nonlinear function of $i - j$ cannot be reduced to a per-position map $g(i)$ with cached absolute positions. WARP is thus not another setting of a ROPE cache but a pairwise relative-position transformation. Figure 1 visualizes this geometry.

Our evidence supports a scoped claim. On modern checkpoints that already tolerate long inputs (Qwen2.5 and Mistral), WARP preserves language-modeling quality and reaches the 50/50 exact-retrieval ceiling at 32k–64k, matching the strongest baselines. On official RULER v1 at 64k with Qwen2.5-3B-Instruct, WARP is on par with NTK (70.82% versus 71.22% macro accuracy) and close to the strongest training-free baselines, YaRN (73.31%) and MrRoPE-Pro (73.44%), while matching or leading them on several retrieval and QA components. On the older Llama-2 models the picture is heterogeneous: Llama-2-13B benefits strongly at 16k but is weak at 32k, and Llama-2-7B exposes retrieval failures at 16k and 32k. We treat these failure regimes as central diagnostic results. An exploratory variant with head routing and a small trained readout adapter recovers much of the lost Llama-2-7B retrieval at 32k, but it is not training-free and its robustness across seeds remains an open question.

Our contributions are:

- We introduce WARP, a training-free rotary extension method that preserves the pretrained frequency spectrum and warps pairwise relative distances instead of globally rescaling absolute positions.
- We show that nonlinear distance warps cannot be represented as standard per-token ROPE caches, which separates WARP from frequency- and position-rescaling methods.
- We add model-conditioned frequency and temperature calibration to adapt the same principle across model families and context lengths.

- We evaluate WARP on exact retrieval, perplexity, and official RULER tasks across six models, and report its failure regimes alongside the settings in which it succeeds.

2 BACKGROUND AND RELATED WORK

Rotary position embeddings. ROPE rotates each query/key feature pair by a position-dependent phase (Su et al., 2024). The resulting dot product depends only on the relative offset between query and key, which is one reason ROPE became the default positional scheme in Llama, Qwen, Mistral, and related models. The same relative-position structure also creates the extrapolation problem: when positions exceed the native pretraining window, the model receives phase relations that may not correspond to any learned behavior. This concern belongs to a longer line of relative-position modeling that includes Transformer-XL's segment recurrence and relative positional formulation (Dai et al., 2019) and T5's relative position bias (Raffel et al., 2020).

Relative-position and inference-time extension methods. ALiBi and xPos change positional inductive biases (Press et al., 2022; Sun et al., 2023), while StreamingLLM uses persistent early-token attention sinks for streaming inference (Xiao et al., 2024). ReRoPE and Leaky ReRoPE are especially relevant: they distinguish near and far relative positions, with the leaky variant retaining a compressed far-distance branch (Su, 2023). SelfExtend combines neighbor and grouped attention without finetuning (Jin et al., 2024). Pair-dependent treatment of near and far context is therefore not unique to WARP. Our specific construction uses a logarithmic tail with continuously decreasing slope and a numerically specified endpoint budget. Nonseparability alone establishes neither novelty nor an empirical advantage over these approaches.

Position and frequency rescaling. Position interpolation rescales positions (Chen et al., 2023); YaRN combines frequency-dependent rescaling with attention-temperature adjustment (Peng et al., 2024); LongRoPE searches nonuniform rescaling patterns (Ding et al., 2024); and MrRoPE uses a mixed-radix view of rotary extension (Tian et al., 2026). We distinguish the inference-only transformations evaluated here from published systems that also use continued training. Within any fixed frequency band, constant frequency rescaling produces a phase linear in distance. The ideal logarithmic warp instead produces a nonlinear phase within that band.

Training-based long-context adaptation. Another line of work extends context through continued training or parameter-efficient adaptation. LongLoRA, for example, combines shifted sparse attention during finetuning with LoRA-style adaptation to extend pretrained Llama models to longer contexts (Chen et al., 2024). Such methods can be highly effective but require additional optimization and data. WARP targets the complementary setting in which the model weights are fixed and only the inference-time rotary geometry changes.

Efficient attention implementations. FlashAttention and FlashAttention-2 reduce attention IO and improve GPU parallelism (Dao et al., 2022; Dao, 2024). FlashInfer provides customizable attention templates and JIT compilation for inference (Ye et al., 2025). AttnFuse studies a compositional interface for attention transformations (Dasoju & Zhao, 2026). These systems are implementation context rather than evidence for the novelty or quality of a particular positional schedule. We make no claim that fused rotary transformations themselves are new, and we distinguish kernel-level timings from end-to-end model performance.

Long-context evaluation. Needle-in-a-haystack retrieval is useful but incomplete. RULER (Hsieh et al., 2024) broadens long-context evaluation with multi-needle, multi-value, variable-tracking, aggregation, and QA-style tasks, while LongBench covers retrieval, summarization, question answering, and code (Bai et al., 2024). We therefore use exact-token retrieval as a controlled stress test and RULER as the broader task-level check.

3 METHOD

Let N be the native context length and L the requested context length. For a causal attention pair with query index i and key index $j \leq i$, define $d = i - j$. WARP replaces d by a monotone warped

distance $f_L(d)$:

$$f_L(d) = \begin{cases} d, & L \le N, \\ d, & L > N \text{ and } d < W, \\ W + c\log\left(1 + \frac{d-W}{c}\right), & L > N \text{ and } d \ge W. \end{cases} \tag{1}$$

The shoulder W specifies the identity region of the ideal distance map. For an extended request, the endpoint target is

$$f_L(L-1) = N, \tag{2}$$

using N, rather than $N-1$, as the budget convention. For $W < N < L-1$, the endpoint is strictly increasing in c, ranging from W to $L-1$ as c ranges from zero to infinity, so there is a unique positive solution. The implementation uses 60 geometric-bisection iterations over $c \in [1, 10^7]$; this fixed bracket must contain the solution. When no compression is needed, the implementation uses the identity map. The budget applies to the specified request interval, not to arbitrarily long continuation with a fixed schedule. The logarithmic shoulder is continuously differentiable: both one-sided derivatives at W equal one.

Rotary application. Let ω_m denote the pretrained frequency of rotary pair m. The pure construction uses $\theta_m(i,j) = \omega_m f_L(i-j)$. The evaluated default additionally uses a frequency multiplier r_m:

$$\theta_m(i,j) = \omega_m r_m f_L(i-j), \qquad r_m = s^{-\alpha_m}, \qquad s = \max(1, L/N). \tag{3}$$

The default sets $W = 1$, so it does not preserve an extended local window exactly. Let D be the head dimension, $H = D/2$ the number of rotary pairs, and β the rotary base. For $m = 0, \ldots, H-1$, write $[x]_a^b = \min(b, \max(a, x))$. The default frequency calibration is

$$A(s) = \begin{cases} 0.15 \sin\left(\frac{\pi \log s}{\log 8}\right) \sqrt{\left[\frac{D-64}{64}\right]_0^1 \left(\frac{1}{2} + \frac{1}{2}\left[\log_2\left(\frac{N}{2048}\right)\right]_0^1\right)}, & 1 < s < 8, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

$$k_0 = \left[\left\lceil \frac{D\log(N/(2\pi))}{2\log\beta} \right\rceil\right]_0^H, \tag{5}$$

$$b = \min\left(\frac{1}{2}, \max\left(\frac{1}{H}, \frac{H-k_0}{H}\right)\right), \tag{6}$$

$$\alpha_m = A(s)\left[\frac{m/(H-1) - (1-b)}{b}\right]_0^1, \qquad r_m = s^{-\alpha_m}. \tag{7}$$

The 2π threshold identifies rotary pairs that complete at most one cycle across the native context; calibration is concentrated toward these slow pairs. The band is limited to between one pair and half the spectrum to avoid broadly changing high-frequency, short-range behavior. The head-dimension gate suppresses calibration for 64-dimensional heads, while the native-length factor reduces its strength for short-native models. The sinusoidal envelope makes calibration zero at native length, largest at moderate extension ratios, and zero again by $8\times$ extension, where a fixed positive frequency split performed poorly in development experiments. The peak value 0.15, the $8\times$ cutoff, and the 2048-token reference are empirical design choices, not theoretically derived constants; their sensitivity should be assessed separately.

Because nonzero calibration is restricted to the slow rotary pairs, the fixed-spectrum property holds for the pure construction ($r_m = 1$) but not for calibrated runs with $1 < s < 8$. No language-model weights are updated by this calibration. Native-length identity is a property of the ideal equations, not a claim of bitwise equality between different finite-precision attention kernels.

Temperature and density correction. Writing $q_{i,m}, k_{j,m} \in \mathbb{R}^2$ for the unrotated feature pairs and $R(\theta)$ for planar rotation, the ideal causal logit is

$$\ell_{ij} = \frac{t}{\sqrt{D}} \sum_m q_{i,m}^\top R(-\theta_m(i,j)) k_{j,m} + \lambda \log f_L'(i-j), \qquad j \le i, \tag{8}$$

where t is the logit-temperature multiplier. For the logarithmic tail,

$$f'_L(d) = \begin{cases} 1, & L \le N \text{ or } d < W, \\ c/(c+d-W), & \text{otherwise.} \end{cases} \tag{9}$$

One motivation for the bias is a continuous attention measure uniform in warped coordinates $u = f_L(d)$: replacing du by $f'_L(d)\,dd$ introduces $\log f'_L(d)$ in the log weight. A discrete softmax over the original tokens does not require this choice of measure. We therefore treat λ as an empirical density-bias strength, not a universally necessary change-of-variables correction. The default sets $\lambda = \frac{1}{2}\,\mathrm{clip}((128 - D)/64, 0, 1)$; in particular, the bias is inactive for $D = 128$. Temperature and frequency calibration are also part of the evaluated configuration, not evidence for a distance-only effect.

Block-anchored implementation. The evaluated prefill kernel shares a key-position shift across a block of queries. For block-center anchor a and a key $j < a$, its effective distance away from clipping boundaries is

$$\widehat{f}(i,j) = i - j - \mathrm{round}\big((a-j) - f_L(a-j)\big). \tag{10}$$

Keys at or ahead of the anchor retain their original positions. The default uses integer shifts and unit shift-table granularity; query-block size is selected by the kernel configuration. Without rounding, $\widehat{f}(i,j) = f_L(a-j) + (i-a)$, which generally differs from $f_L(i-j)$. For $0 \le f'_L \le 1$, its distance error is bounded by $|i-a|$, with at most an additional half-token from rounding under these conditions. This does not guarantee small phase or output error. Accordingly, the reported experiments evaluate a block-anchored approximation, not exact pairwise application of Eq. equation 1.

4 Why nonlinear WARP is not a RoPE cache

Many rotary extension methods can be implemented by assigning each token an effective absolute position $g(i)$ and using the difference $g(i) - g(j)$ in the rotary phase. WARP is intentionally more general. Suppose a relative-distance warp f could be represented by absolute positions:

$$g(i) - g(j) = f(i-j) \quad \text{for all } i \ge j \ge 0. \tag{11}$$

Setting $j = 0$ gives $g(i) = g(0) + f(i)$, and substituting back gives

$$f(i) - f(j) = f(i-j). \tag{12}$$

This is Cauchy's additive equation on the integers: with $f(0) = 0$, induction gives $f(d) = f(1)\,d$, so every solution is linear. A genuinely nonlinear warp such as Eq. equation 1 therefore cannot be expressed as an absolute-position cache. WARP is a pairwise relative-position method, not a new interpolation factor.

5 Experimental setup

Models and baselines. We evaluate TinyLlama-1.1B, Llama-2-7B, Llama-2-13B, Qwen2.5-0.5B-Instruct, Qwen2.5-3B-Instruct, and Mistral-7B-v0.3 at context lengths from 8k to 64k, depending on the model. We compare against unmodified RoPE, PI, NTK scaling, YaRN, and MrRoPE-Pro where available.

Evaluation signals. We use three complementary signals. *Exact retrieval* measures whether the model emits a hidden target token from a controlled long-context prompt, reported as hits out of 50 trials. *Perplexity* measures language-modeling quality on long windows. Official RULER v1 measures broader long-context behavior across 13 retrieval, multi-key, multi-value, variable-tracking, extraction, and QA tasks.

Locked protocol. For each headline comparison, all methods share the same model, context length, prompt set, decoding settings, scoring script, and random seeds, and every method-specific option is fixed before scoring. We refer to this setup as the *locked protocol*. RULER results additionally use a padding-safe evaluation in which no sequence is padded within a batch. Seed-variance results use ten seeds; *fresh-seed* results use seeds that were not used while developing the corresponding variant.

Method	Official RULER macro @ 64k	Δ vs. RoPE
PI	2.78%	−60.14
RoPE	62.92%	0.00
NTK	71.22%	+8.30
WARP	70.82%	+7.90
YaRN	73.31%	+10.39
MrRoPE-Pro	**73.44%**	**+10.52**

Table 1: Official RULER v1 13-task macro accuracy for Qwen2.5-3B-Instruct at 64k under the locked, padding-safe protocol. Δ is measured in percentage points against unmodified RoPE; higher is better. WARP improves strongly over RoPE and PI and is on par with NTK, but YaRN and MrRoPE-Pro achieve the highest macro accuracy.

Task	WARP	MrRoPE-Pro	YaRN
NIAH single 1	100.00	100.00	100.00
NIAH single 2	100.00	100.00	100.00
NIAH single 3	99.00	100.00	100.00
NIAH multi-key 1	94.00	96.00	94.00
NIAH multi-key 2	77.00	75.00	75.00
NIAH multi-key 3	32.00	38.00	39.00
NIAH multi-value	85.00	87.25	88.25
NIAH multi-query	94.00	95.00	95.25
Variable tracking	75.40	84.00	88.20
Common-word extraction	1.90	14.10	10.70
Frequent-word extraction	75.33	79.33	77.67
QA 1	48.00	48.00	48.00
QA 2	39.00	38.00	37.00

Table 2: Task-level official RULER v1 accuracy (%) for Qwen2.5-3B-Instruct at 64k under the same locked, padding-safe protocol as Table 1; higher is better. WARP is best or tied-best on several retrieval and QA components but loses macro accuracy on variable tracking, common-word extraction, and the deepest multi-key setting.

6 Results

6.1 Official RULER at 64k

Table 1 reports the main broad-task result. WARP reaches 70.82% macro accuracy on Qwen2.5-3B-Instruct at 64k, a gain of 7.90 points over unmodified RoPE, while PI collapses to 2.78%. WARP is on par with NTK (71.22%) but 2.5–2.6 points below YaRN (73.31%) and MrRoPE-Pro (73.44%). We therefore do not claim broad RULER dominance; WARP is competitive overall and strongest on the retrieval-focused components discussed below.

Table 2 compares WARP with the two macro winners task by task. Among these three methods, WARP is best on NIAH multi-key 2 and QA 2, and ties on NIAH single 1, NIAH single 2, and QA 1. The macro gap comes mainly from variable tracking (75.40 versus 84.00–88.20), common-word extraction (1.90 versus 10.70–14.10), and the deepest multi-key setting (32.00 versus 38.00–39.00).

6.2 Perplexity and efficiency

Table 3 reports long-window perplexity for Qwen2.5-3B-Instruct. WARP does not degrade quality at native length: its 32k perplexity is close to RoPE and slightly better than YaRN and MrRoPE-Pro. At 64k, WARP remains stable but is slightly worse than RoPE, YaRN, and MrRoPE-Pro. Table 4 reports the harder Llama-2-7B setting, whose native context is much shorter. There, WARP matches RoPE at native length, is the best method at 8k–16k, and is comparable to YaRN at 32k, but degrades sharply at 64k. These results support a claim of model-adaptive extension rather than uniform superiority at all lengths.

Method	PPL @ 32k	PPL @ 64k
RoPE	11.383	12.436
PI	21.913	23.163
YaRN	11.626	12.534
MrRoPE-Pro	11.627	12.549
WARP	11.409	12.655

Table 3: Long-window perplexity for Qwen2.5-3B-Instruct at 32k and 64k; lower is better. Only methods with a completed run under the locked protocol are shown. WARP remains close to ROPE, YaRN, and MrRoPE-Pro, so the default preset does not harm language-modeling quality, but it does not achieve the best 64k perplexity.

Method	4k	8k	16k	32k	64k
RoPE	**8.44**	5721.94	7635.42	9239.27	25171.65
PI	225.80	214.75	192.92	224.06	373.11
NTK	15.53	15.94	20.23	71.76	3274.98
YaRN	13.94	15.65	37.78	55.62	131.40
MrRoPE-Pro	12.88	13.17	15.09	**22.55**	**64.51**
WARP	**8.44**	**6.93**	**7.20**	55.50	684.56

Table 4: Llama-2-7B perplexity across evaluation lengths, with every method configured for a fixed 64k (16×) extension; lower is better. WARP matches ROPE at native length and is best at 8k–16k. MrRoPE-Pro is best at 32k–64k; WARP is comparable to YaRN at 32k but degrades at 64k on this short-native model.

The main cost of WARP is latency. In a synthetic H100 attention microbenchmark, WARP prefill is about 3.2× slower than ROPE at 32k (72.6 ms versus 22.8 ms) and about 3.0× slower at 64k (296.7 ms versus 98.5 ms). Peak memory is close to that of YaRN and ROPE in this isolated benchmark, but decode latency is also higher because WARP computes pairwise relative geometry instead of reusing a standard per-token rotary cache. Closing this gap is a target for fused attention kernels.

6.3 EXACT RETRIEVAL ACROSS MODELS

Table 5 reports ten-seed exact-retrieval statistics for WARP, and Figure 2 compares WARP with the baselines across model families and context lengths. On Qwen2.5-3B-Instruct and Mistral-7B-v0.3, WARP reaches the 50/50 ceiling at both 32k and 64k with no observed seed variance. Unmodified ROPE, YaRN, and MrRoPE-Pro reach the same ceiling in these settings, so these endpoints show that WARP does not harm retrieval rather than that it outperforms the baselines. WARP is the best method on Qwen2.5-0.5B-Instruct at 64k (50 versus 47 for MrRoPE-Pro), on Llama-2-7B at 8k (47 versus 39), and on Llama-2-13B at 16k (50 versus 41). On Llama-2-13B at 32k it is the only method above zero, but at 11.60/50 we count this endpoint as a failure rather than a success. WARP is weaker than YaRN and MrRoPE-Pro on TinyLlama at 8k–16k and on Llama-2-7B at 16k (5 versus 35), and every method fails on Llama-2-7B at 32k. We therefore report the failure cases alongside the seed-confirmed successes rather than summarizing the method with a single average.

6.4 THE LLAMA-2-7B 32K CLIFF

The hardest observed setting is Llama-2-7B at 32k, where base WARP and all baselines retrieve 0/50 (Figure 2). Global tuning of the temperature t, the calibration exponents α_m, the budget, shift granularity, and query tiling did not recover this endpoint. We therefore explored targeted adaptive variants: attention-density thresholds, per-head heterogeneous schedules, anchored retrieval routes, and head-preserving adaptive routers. The best variant reaches 31/50 on one fresh seed and 27/50 on another; it includes the trained readout adapter described below. We treat this as a partially repaired cliff and a diagnostic result, not as a solved case or as part of the default method.

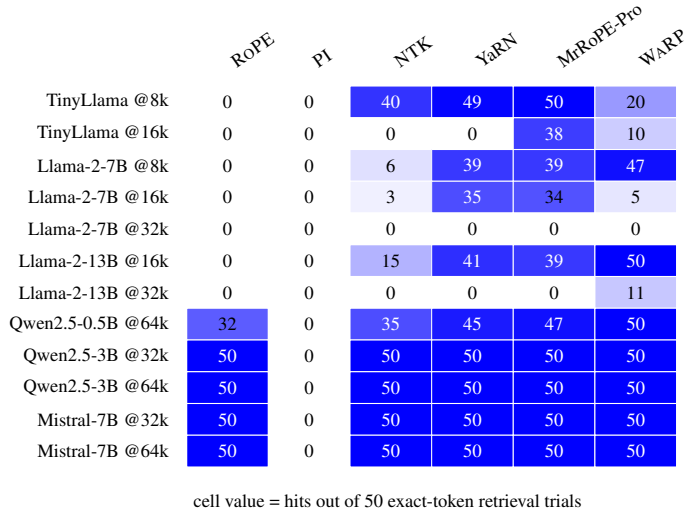


Figure 2: Exact-token retrieval under the original implementation, measured on one fixed seed with 50 trials per model–length–method cell; higher is better. The previously missing PI and NTK cells for Qwen2.5-3B-Instruct and Mistral-7B-v0.3 were evaluated using the same seed and fixed scale. Qwen labels denote Instruct checkpoints. This single-seed matrix is separate from the ten-seed WARP summary: for example, Llama-2-13B at 32k scores 11/50 here, whereas its ten-seed mean is 11.60/50. The matrix predates the fused-kernel boundary corrections and is not a corrected-code comparison.

Model @ length	Mean exact retrieval	Std.
Qwen2.5-3B-Instruct @ 32k	50.00/50	0.00
Qwen2.5-3B-Instruct @ 64k	50.00/50	0.00
Mistral-7B-v0.3 @ 32k	50.00/50	0.00
Mistral-7B-v0.3 @ 64k	50.00/50	0.00
Qwen2.5-0.5B-Instruct @ 64k	50.00/50	0.00
Llama-2-13B @ 16k	49.20/50	0.79
Llama-2-13B @ 32k	11.60/50	2.32

Table 5: Ten-seed exact-token retrieval for WARP under the locked protocol. Each score is the number of successful retrievals out of 50 trials, averaged over ten seeds; the standard deviation is in hits out of 50. Higher is better. The Qwen and Mistral endpoints are seed-stable, whereas Llama-2-13B at 32k remains a failure case.

6.5 TRAINED READOUT EXTENSION FOR THE LLAMA-2-7B CLIFF

The primary WARP method changes inference-time positional geometry without updating model weights. Separately, we explored a trained rank-16 readout adapter for Llama-2-7B. It applies a low-rank residual $h \mapsto h + U(Vh)$ to the final hidden state immediately before the frozen language-model output head; the base model and output head remain frozen. With $V \in \mathbb{R}^{16 \times d_{\text{model}}}$, $U \in \mathbb{R}^{d_{\text{model}} \times 16}$, and $d_{\text{model}} = 4096$, the adapter trains $2 \cdot 16 \cdot d_{\text{model}} = 131{,}072$ parameters. It was optimized with AdamW for 96 updates on synthetic recall examples and language-modeling windows, with recall sampled on 90% of updates. The 31/50 and 27/50 results combine this trained adapter with anchor and head routing; they do not isolate the adapter's contribution. We therefore report them as exploratory, non-training-free extensions, separate from the main WARP comparisons. Both components are model-specific: the adapter's shape depends on the hidden size and the routing on the head layout, so neither is shared across models.

7 DISCUSSION

The experiments support a scoped claim for WARP rather than a claim of universal dominance. Nonlinear relative-distance warping is a distinct training-free extension mechanism: in its pure form it preserves the pretrained rotary spectrum, is exactly native at the original context length, and compresses only long-range distances. Empirically, this design matches the exact-retrieval ceiling on Qwen and Mistral, keeps Qwen2.5-3B-Instruct perplexity stable at long context, and is competitive on broad RULER at 64k. Its RULER profile is uneven: strengths on retrieval and QA are offset by weaker extraction and tracking, which account for the remaining macro gap to YaRN and MrRoPE-Pro.

The main open issue is model heterogeneity. Llama-2-13B and Llama-2-7B do not behave like the newer Qwen and Mistral checkpoints, and Llama-2-7B exposes a sharp retrieval cliff. The gap between the two Llama-2 models is plausibly related to capacity rather than to the warp alone: the 13B model has more layers and attention heads, and therefore potentially more redundant paths for preserving retrieval information under extension. This suggests that long-context extrapolation depends not only on the rotary schedule but also on model-specific attention allocation, head specialization, and pretraining context length. In practice, WARP is most useful when exact retrieval and local quality are the priority, and model-adaptive variants remain an active direction for older short-context checkpoints.

8 LIMITATIONS

First, WARP is weak on Llama-2-7B exact retrieval at 16k, and its repair of the 32k endpoint is seed-sensitive, relies on a trained adapter, and should not be overstated. Second, Llama-2-13B exact retrieval at 32k is weak (11.60/50) under ten-seed evaluation. Third, Llama-2-7B perplexity at 64k degrades sharply, so WARP should not be presented as a universal solution for extreme extension of short-native checkpoints. Fourth, pairwise distance computation makes prefill about $3\times$ slower than ROPE in our current implementation. Fifth, the current implementation requires padding-safe batching; padded batches must be supported before production use. Sixth, the model-conditioned presets are validated empirically; a theory of which architectures prefer which Jacobian and frequency settings remains future work. Seventh, all reported runs use the block-anchored approximation rather than exact pairwise warping.

9 CONCLUSION

We introduced WARP, a training-free rotary extension method that, in its pure form, leaves the pretrained frequency spectrum untouched and instead applies a monotone nonlinear warp to the relative query–key distance. The warp is exactly ROPE at native length, approximately preserves local distances, and compresses only the far field. Because it is nonlinear in $i - j$, it cannot be expressed as a per-token position cache, which places it in a different class from position- and frequency-rescaling methods. Empirically, WARP matches the exact-retrieval ceiling on Qwen and Mistral at 32k–64k, is the best method in several Llama-2 and Qwen2.5-0.5B retrieval settings, keeps Qwen2.5-3B-Instruct perplexity close to ROPE, and is competitive but not dominant on official RULER at 64k, with its advantage concentrated in retrieval and QA components. Its failures are equally informative: older Llama-2 checkpoints expose retrieval cliffs that the base warp cannot remove, and pairwise distance computation currently costs about $3\times$ in prefill latency. Promising next steps include fused kernels for pairwise rotary geometry, a principled theory of model-conditioned presets, and adaptive variants whose Llama-2 repairs are robust across seeds.

Reproducibility statement All reported results use fixed model checkpoints, evaluation presets, random seeds, context lengths, trial counts, and scoring scripts. For each experiment, we record the method configuration, model name, context length, evaluation protocol, and the aggregate outputs used in the corresponding table or figure. The software artifacts of this paper are available at `https://anonymous.4open.science/r/WaRP-E778/README.md`.

AI use statement The authors used AI assistance to aid and polish the writing. The authors reviewed and edited the manuscript and remain responsible for the correctness of the claims, experiments, code, and final submission.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

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.

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.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

Varun Kumar Dasoju and Tian Zhao. AttnFuse: A composable DSL for compiling attentions to fused GPU kernels. *arXiv preprint arXiv:2609.13612*, 2026.

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*, pp. 11091–11104, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. LLM maybe LongLM: Self-Extend LLM context window without tuning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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.

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

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

Jianlin Su. Rectified rotary position embeddings. `https://github.com/bojone/rerope`, 2023. ReRoPE and Leaky ReRoPE.

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

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Albert Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Zihao Ye, Lequn Chen, Ruihang Lai, Wuwei Lin, Yineng Zhang, Stephanie Wang, Tianqi Chen, Baris Kasikci, Vinod Grover, Arvind Krishnamurthy, and Luis Ceze. FlashInfer: Efficient and customizable attention engine for LLM inference serving. In *Proceedings of Machine Learning and Systems*, volume 7, 2025.