# ROTARY ATTENTION IN THE FREQUENCY DOMAIN: VALUE ROTATION AS A DISTANCE FILTER

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

## ABSTRACT

Rotary position embedding (RoPE) encodes position in the attention score but leaves the values unrotated, so what a head reads carries no record of where it came from. We study the variant that also rotates the values and inverse-rotates the head output (QKVO-RoPE), which adds no parameters. We show that this turns each attention head into a filter over distance. Content that a head reads from a single position passes and keeps a phase that records its distance, whereas content that it averages over many positions partly cancels. To test this account, we train matched Transformer pairs from 130M to 1.3B parameters that differ only in the rotation and measure the cancellation in every head. We find that value rotation helps where one position must be read against a diffuse context. Beyond the training length the larger rotated models retrieve needles that their unrotated counterparts miss entirely, and up to 760M the rotated models extrapolate with lower perplexity. We also find that the costs fall where the analysis predicts, in heads that average and in associative recall, whose source lies at a distance set by the query, while the cost at the training length is small. A larger rotary base reduces the cancellation. Our results suggest rotating the values in models that must retrieve from long contexts, preferably with a large rotary base.

## 1 INTRODUCTION

Rotary position embedding (RoPE) is the standard way to encode position in Transformer attention (Su et al., 2024). It rotates each query and key through angles proportional to its position, so that their inner product depends only on their relative position, and it leaves the value vectors unrotated. Position determines which values a head reads, through the attention weights, but not what it reads. A value reaches the head output with no record of how far back it came from.

This work studies what happens when the rotation is also applied to the values and outputs. Each value is rotated by its absolute position and the head output is inverse-rotated at the query position, which leaves the attention score unchanged and adds no parameters. We call this variant QKVO-RoPE and the standard construction QK-RoPE. We analyze what it does and when it helps. DeepSeek-V4 applies it to the rotary dimensions of its shared key-value entries (DeepSeek-AI, 2026), and RoVE reports lower perplexity beyond the training length for GPT-2 models up to 354M (García-Castellanos et al., 2026). Neither asks when the rotation costs. The rotation and its inverse do not cancel. Each value reaches the head output multiplied by a factor $e^{-i\theta\ell}$ that oscillates with its distance $\ell$ from the query, once per rotary frequency $\theta$. What a head reads can depend on how far back it lies. The same oscillation appears in linear recurrences with complex poles, where it is combined with a learned decay (Gu et al., 2022; Lahoti et al., 2026). Value rotation supplies only the oscillation, at the fixed rotary frequencies, and in softmax attention the attention weights take the place of the decay. The value path gains a fixed set of oscillations over distance, not an arbitrary learned dependence on distance. Other positional encodings, including ALiBi (Press et al., 2022) and YaRN (Peng et al., 2024), act only on the attention score, and when rotating the values helps a trained Transformer, and when it hurts, has not been characterized.

We first derive what a rotated head computes (Section 3). Because a head attends to many distances at once, the oscillating factors of the values it reads interfere, and the head output acts as a filter over distance (Figure 1). A head that reads one position, or a fixed distance back, keeps its content. A head that averages over many positions loses part of the content they share, starting with the



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.

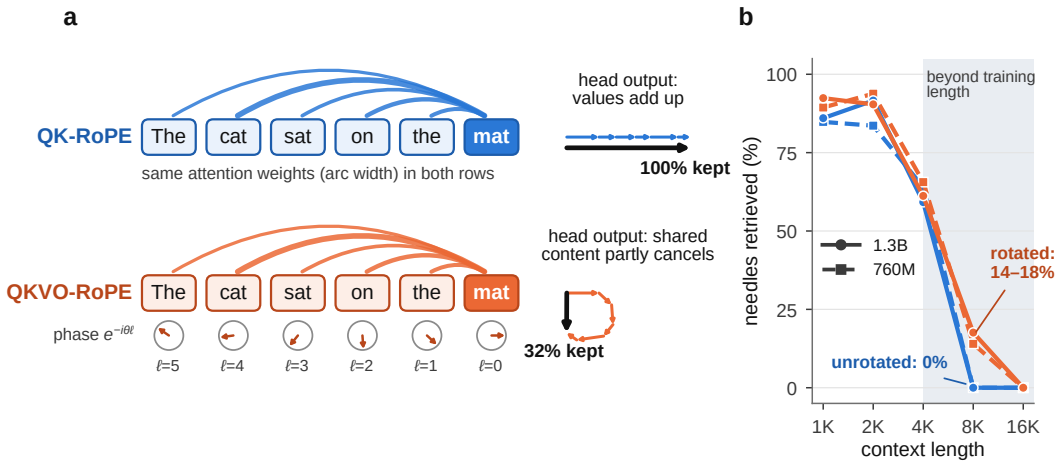



Figure 1: **Value rotation turns each attention head into a filter over distance.** (a) The query "mat" reads the same values with the same attention weights (arc width) in both rows. Under QK-RoPE the weighted values add up unchanged (arrows, drawn head to tail). Under QKVO-RoPE each value arrives multiplied by the phase $e^{-i\theta\ell}$ of its lag $\ell$ (dials, one fast rotary pair), so content shared by the attended positions partly cancels, and here 32% is kept. Section 3.3 derives how much each rotary pair keeps. (b) On trained models, content read from one position survives. At twice the 4K training length the rotated 760M and 1.3B models retrieve 14–18% of needles and their unrotated counterparts none (Section 4.6).

fast-oscillating pairs. Over a uniform window of $W$ positions that content survives in only about a fraction $1 - \ln W / \ln b$ of the pairs at base $b$. The filter should therefore help when a head must pick one position out of a diffuse background, as beyond the training length, and cost when the content a head needs is itself an average or lies at a distance that changes with the query.

We then test these predictions on matched pairs of Transformers at 370M, 760M and 1.3B parameters, together with a 130M pair used for ablations (Section 4). The two members of each pair share data, data order, tokenizer, optimizer, context length and all remaining hyper-parameters, and differ only in whether the value path is rotated. We measure the predicted attenuation head by head on the trained models, train nine of the sixteen combinations of rotating Q, K, V and O at 130M, and include positional-encoding controls and a gated DeltaNet (Yang et al., 2025a) comparison.

Our experiments support most of these predictions. In every model, most heads attend over long distances, so the predicted cancellation occurs in practice, and the rotated models compensate by concentrating their content in the low-frequency rotary pairs. Value rotation improves retrieval of a single position from a long context, because the cancellation suppresses the diffuse background while preserving the content at that position. Beyond the training length, the larger rotated models retrieve needles that their unrotated counterparts fail to retrieve. Conversely, value rotation degrades performance when the required content is an average, lies far back, or lies at a query-dependent distance, as in long-range copying and associative recall. Up to 760M, the rotated models also extrapolate with lower perplexity. At 1.3B this trend reverses, and we attribute the reversal to the attention score rather than to the value path. A larger rotary base reduces the cancellation.

Value rotation is thus neither a uniform improvement nor a uniform cost. The same cancellation that harms averaging heads benefits retrieval heads, and the proportion of each head type can be estimated from a model's attention distances before retraining.

## 2 Related Work

**Oscillation in sequence models.** Diagonal state-space models and linear RNNs place their poles in the complex plane, and the imaginary part gives the state an oscillation that real decay cannot produce (Gu et al., 2022; Gu & Dao, 2024; Orvieto et al., 2023; Dao & Gu, 2024). Value rotation introduces the same oscillation on the value path of attention, at the fixed rotary frequencies and

without the decay (Section 3.2), which is the direct connection to these models. Mamba-3 shows that a complex diagonal SSM equals a real one with data-dependent rotary embeddings on its input and output projections, the key and query of linear attention (Lahoti et al., 2026). RetNet applies a complex exponential decay to queries and keys, on the score path of attention (Sun et al., 2023a).

**Position in the score and in the values.** RoPE (Su et al., 2024) places position in the query-key product, and CRoPE parametrizes the attention projections as complex linear maps over the rotary pairs (Lou et al., 2026), the complex view adopted in Section 3.1. Beyond the training length its low-frequency pairs receive angles never seen in training, and rescaling the frequencies at inference time restores in-range angles (Chen et al., 2023; Peng et al., 2024; Liu et al., 2024), and the base bounds the usable context (Xu et al., 2024; Su, 2024). XPos adds a lag-dependent decay (Sun et al., 2023b), ALiBi replaces the rotation with a distance penalty (Press et al., 2022), and models with no positional encoding learn position from the causal mask (Kazemnejad et al., 2023). Barbero et al. (2025) find that the high-frequency pairs of a trained model detect position while the low-frequency pairs carry content, and Section 3.3 gives the corresponding result for the value path. Su's second rotary encoding rotates values and de-rotates the output (Su, 2025b), which lets latent attention share one latent for keys and values (Su, 2025a) (Appendix I), and DeepSeek-V4 applies the rotation and the inverse output rotation to the last 64 dimensions of its shared key-value entries for this reason (DeepSeek-AI, 2026). RoVE, concurrent with this work, reads the construction as an attentive convolution and reports better in-context learning, lower perplexity beyond the training length and better long-context retrieval for GPT-2 models of 124M and 354M trained at a 1K context (García-Castellanos et al., 2026). We instead derive where the rotation cancels content, measure this in trained heads, and report both its costs and its gains in models up to 1.3B. PaTH replaces the fixed rotation between query and key with accumulated data-dependent Householder transformations and leaves the values untransformed (Yang et al., 2025b).

# 3 METHOD

This section defines QKVO-RoPE, derives what a rotated head computes, and uses that form to predict how much of its values a head retains, both within and beyond the training length.

## 3.1 VALUE-PATH ROTATION

We first define the rotation and the notation used throughout. Consider one attention head with even dimension $d$ and $J = d/2$ rotary frequencies $\theta_j = b^{-2j/d}$, $j = 0, \dots, J-1$, with base $b = 10^4$. The rotation matrix $\mathbf{R}_m \in \mathbb{R}^{d\times d}$ at position $m$ is block diagonal with $2 \times 2$ blocks rotating pair $j$ by angle $m\theta_j$, and pair $j$ of a vector $x = (x_0, \dots, x_{d-1}) \in \mathbb{R}^d$ is the coordinate pair $(x_{2j}, x_{2j+1})$ on which that block acts, which we write as the complex number $x^{(j)} = x_{2j} + i\, x_{2j+1}$, so that $(\mathbf{R}_m x)^{(j)} = e^{im\theta_j} x^{(j)}$. Let $q_m, k_n, v_n \in \mathbb{R}^d$ be the query, key and value at positions $m$ and $n \le m$, and $a_{mn} \ge 0$, $\sum_{n\le m} a_{mn} = 1$, the attention weights. Standard RoPE (QK-RoPE) computes scores from $\mathbf{R}_m q_m$ and $\mathbf{R}_n k_n$ and outputs $o_m = \sum_n a_{mn} v_n$. The value-path variant (QKVO-RoPE) keeps the scores and replaces the output by

$$o_m \;=\; \mathbf{R}_m^\top \sum_{n\le m} a_{mn}\, \mathbf{R}_n v_n. \tag{1}$$

The head output is then multiplied by the output projection $W_o$ as usual. We write $\ell = m - n \ge 0$ for the lag. Figure 1 shows how one head reads its values under the two constructions, and Figure 5 gives the two readings of the same rotation, as poles on the value axis and as a weight applied to each lag.

## 3.2 WHAT A ROTATED HEAD COMPUTES

We now show that a rotated head is ordinary softmax attention in which each attended value is multiplied by a factor that oscillates with its distance from the query. For every pair $j$ and query position $m$,

$$o_m^{(j)} \;=\; \sum_{\ell=0}^{m} a_{m,m-\ell}\; e^{-i\theta_j \ell}\; v_{m-\ell}^{(j)}. \tag{2}$$

Equation 2 holds exactly under softmax attention. The rotation and its inverse do not cancel, and each attended value arrives multiplied by $e^{-i\theta_j \ell}$, whose real part $\cos \theta_j \ell$ weights lag $\ell$ with a sign that alternates with period $2\pi/\theta_j$ (Figure 5(b)). Every later result follows from equation 2, and the derivations are given in Appendix A.

The same factor appears in linear attention with scalar decay $\alpha$, where the construction has a recurrent form (Su, 2025b; Dao & Gu, 2024). The value axis of the state then carries the complex pole $\alpha e^{-i\theta_j}$ instead of the real pole $\alpha$ (Appendix A), whose impulse response $\alpha^\ell e^{-i\theta_j \ell}$ combines a decay with an oscillation. Softmax attention keeps the oscillation and replaces the fixed decay $\alpha^\ell$ by the attention weights, so the value path can weight distance only through the fixed oscillations $\cos \theta_j \ell$, not through an arbitrary learned dependence on distance.

### 3.3 How much of a value is retained

We now derive how much of a value's content a rotated head retains. The answer depends on how the head spreads its attention over lags, and we consider three cases in turn. A head may read one fixed lag, read a lag that depends on the query, or average over many lags.

A head that attends to one fixed lag is unaffected. If $a_{m,m-\ell} = \mathbf{1}[\ell = \ell^\star]$ for all $m$, then $o_m = \mathbf{R}_{-\ell^\star} v_{m-\ell^\star}$, and since $\mathbf{R}_{-\ell^\star}$ is a fixed orthogonal matrix the rotated head with output projection $W_o \mathbf{R}_{\ell^\star}$ computes exactly what the unrotated head computes with $W_o$. The two models have identical function classes on such heads, which covers previous-token heads and any head reading a fixed relative position.

An induction head behaves differently, because it finds its source by content, so its lag $\ell^\star(m)$ varies with the query, and no fixed $W_o$ removes a query-dependent rotation. Content in pair $j$ is then readable only up to a phase $e^{-i\theta_j \ell^\star(m)}$, negligible while $\theta_j \ell^\star \ll 1$ and unreadable by a fixed output projection once $\theta_j \ell^\star$ is of order one.

Most heads instead average content over a range of positions (Section 4.3), and there the rotation removes part of what is read. Consider the case in which the attended values share the same content in pair $j$, so that $v_n^{(j)} = v^{(j)}$ for all $n$ in the support of $a_{m\cdot}$. Then $o_m^{(j)} = \varphi_m(\theta_j)\, v^{(j)}$, where

$$\varphi_m(\theta) = \sum_{\ell \geq 0} a_{m,m-\ell}\, e^{-i\theta\ell} \tag{3}$$

is the characteristic function of the head's lag distribution at query $m$. Its magnitude is at most one, with equality only when the attended lags are congruent modulo $2\pi/\theta_j$. Content that differs across the attended positions is not attenuated in this way. If the values at different positions are independent with zero mean and independent of the attention weights, the rotation changes the phases of their contributions but leaves the expected output energy unchanged, so the attenuation applies to the component of content that the attended positions share. For a uniform window over lags $0, \ldots, W-1$,

$$|\varphi(\theta)| = \frac{|\sin(W\theta/2)|}{W\,|\sin(\theta/2)|}, \tag{4}$$

and the number of pairs with $W\theta_j < 1$, on which content passes almost unattenuated ($|\varphi| > 0.95$), is $J_W = J\,(1 - \ln W/\ln b)$ up to rounding.

Figure 4 plots equation 4 at the 32 frequencies of a 64-dimensional head. For content shared across the window, at $b = 10^4$ a head that averages over 32 tokens keeps 19 of its 32 rotary pairs, one that averages over 2048 tokens keeps 5, and one that averages over a whole 4096-token block keeps 3. The cancelled pairs are the ones that distinguish distances, since content at distances $\pi/\theta_j$ apart arrives in pair $j$ with opposite sign (Appendix H). A head obtains distance selectivity in a given pair only at the cost of cancellation in that same pair. Content that must be retained under averaging is therefore confined to the low-frequency pairs, whose share, $1 - \ln W/\ln b$, falls with the logarithm of the window $W$ and rises with the logarithm of the base $b$. Their number is this share times the number of pairs, so it grows linearly, not faster, with the number of rotary pairs, and the effect is largest for heads that aggregate over long ranges.

**What the filter keeps.** The cancellation that costs an averaging head can help a head that retrieves one item. Suppose a head places weight $a^\star$ on one source at lag $\ell^\star$ and spreads the remaining $1 - a^\star$

over a window of $W$ other positions whose values share a common component, a background that every position of a document carries. Without rotation the background enters the output with weight $1 - a^\star$ on every pair. With rotation it is scaled by the attenuation of the spread part, while the source keeps a phase close to one on the pairs with $\theta_j \ell^\star \ll 1$. On the pairs with $1/W \lesssim \theta_j \ll 1/\ell^\star$ the source passes and the background cancels. Such diffuse attention arises whenever many positions receive similar scores, as over long contexts and beyond the training length. The rotation also leaves in the output a phase that records the lag of what was read, a positional signal that no unrotated value carries and that alone supports copying (Appendix C). Value rotation should therefore help when one position must be read against a diffuse background and cost when the content a head needs is spread over many positions or lies at a distance that changes with the query.

### 3.4 Behavior beyond the training length

We next ask what the rotation does beyond the training length $L$. There the score and the value path both meet angles not seen in training on the $J_L^{\text{low}} = J\,(1 - \ln(L/2\pi)/\ln b)$ slowest pairs, whose angles do not complete a turn within $L$ (Appendix A). On the other pairs the phases wrap, and attention spread over many out-of-range lags is attenuated as in Section 3.3. Beyond the training length the rotation therefore has two opposing effects. The value path adds a second channel with out-of-range phases on the low-frequency pairs, which in a rotated model carry most of the head output that reaches the output projection (71 to 75%, Section 4.3). At the same time the high-frequency pairs cancel the shared, mean component of the content that the out-of-range score collects from far positions, a known failure mode of RoPE extrapolation (Chen et al., 2023; Peng et al., 2024; Liu et al., 2024). Section 4.5 tests which effect is larger by rescaling the score and the value path separately. Frequency rescaling at inference time restores in-range phases for the score and the value path alike, and should remove most of the difference beyond $L$.

### 3.5 Testable predictions

We summarize the analysis as three predictions, each tested in Section 4. First, copying is performed by heads whose lag is set by the prompt, so its cost should grow with the filler length, small while $\theta_j \ell^\star \ll 1$ on most pairs and large at lags in the thousands, where only five of 32 pairs remain readable. Second, within the training length the QKVO-RoPE/QK-RoPE perplexity ratio should rise with position inside a block, and across models the cost should be ordered by the fraction of heads that aggregate over long distances. Third, rotating V without de-rotating O leaves the absolute phase $e^{i\theta_j m}$ of the query in every output, which should give QKV-RoPE an absolute positional code that helps copying, and VO-RoPE, whose score carries no position, should still copy through the lag phase alone (Appendix C).

## 4 Experiments

We test the predictions of Section 3.5 on matched Transformers that differ only in the value rotation.

### 4.1 Experimental setup

**Models.** We compare two positional encodings of the same Transformer. QK-RoPE applies RoPE to queries and keys only, and QKVO-RoPE adds the value rotation and output de-rotation of equation 1, which introduce no learned weights. Every model is a pre-norm Transformer with SwiGLU feed-forward blocks from the `flash-linear-attention` library (Yang & Zhang, 2024), using its standard softmax attention layer (FlashAttention kernels), not its linear-attention layers. The matched pairs are 370M (24 layers, width 1024, 16 heads of dimension 64), 760M (24 layers, width 1536, 16 heads of dimension 96) and 1.3B (24 layers, width 2048, 32 heads of dimension 64), and a fourth pair at 130M (12 layers, width 768, 12 heads of dimension 64) is the ablation scale. Within a pair the two models differ in a single configuration flag, so parameters, data, data order and hyper-parameters are identical.

**Training.** All models are trained from scratch on FineWeb-Edu (Penedo et al., 2024) with `flame` (Zhang & Yang, 2025), with the Mistral tokenizer, AdamW with linear warm-up and cosine decay and a context of 4096 tokens. The peak learning rate, warm-up and weight decay are set per scale

Table 1: Matched-pair results at the 4096-token training length. Bold marks the better member of each pair.

| scale | model | loss | WT103 | PG-19 | LC64 | Copy (%) ↑ | MQAR (%) ↑ | zero-shot ↑ |
|---|---|---|---|---|---|---|---|---|
| 370M | QK-RoPE | **2.496** | **23.55** | 20.46 | **14.02** | 76.77±0.19 | **27.20±1.24** | 45.3 |
| | QKVO-RoPE | 2.501 | 23.84 | **20.35** | 14.10 | **77.23±0.14**† | 20.19±0.89† | **46.1** |
| 760M | QK-RoPE | **2.265** | **18.31** | 17.28 | **11.31** | 78.70±0.05 | 24.19±1.15 | **52.1** |
| | QKVO-RoPE | 2.271 | 18.54 | **17.16** | 11.39 | **79.82±0.12**† | **27.63±0.83**† | 51.5 |
| 1.3B | QK-RoPE | **2.117** | 15.69 | **14.68** | **9.98** | 81.48±0.08 | **28.31±0.73** | **56.8** |
| | QKVO-RoPE | 2.119 | **15.66** | 15.31 | 10.06 | **82.93±0.12**† | 27.93±0.70 | 56.4 |

Copy and MQAR give the mean ± standard deviation over three seeds, averaged over conditions (Table 10). Zero-shot is the six-task average (Table 12). † Difference from QK-RoPE significant after Bonferroni correction (paired tests on identical prompts, Appendix G). Copy uses ordinary subword tokens as the copied content and filler lengths of 0–50 (Appendix J).

and are identical within each pair (Table 9). The pairs are trained on 5.4B (130M), 10.7B (370M), 20.1B (760M) and 30.0B (1.3B) tokens, and Table 9 lists every run with its training loss.

**Controls.** Two further groups of models each answer one question. XPos (Sun et al., 2023b), which is designed for length extrapolation, and a model with no positional encoding serve as controls at 370M, together with a QK-RoPE model trained on the same 86B tokens (Appendix B). Gated DeltaNet (Yang et al., 2025a) with and without rotations asks whether the construction benefits a recurrence that already has its own decay and gate (Appendix E).

**Evaluation.** Perplexity is computed on non-overlapping blocks of a packed token stream from WikiText-103 (Merity et al., 2017), PG-19 (Rae et al., 2020) and the first 2M tokens of LongCrawl64 (Buckman, 2024), for each block and for each 512-token bucket within a block. Three retrieval probes test the three cases of Section 3. **Copying** reads from a distance the prompt fixes and **multi-query associative recall (MQAR)** from a distance the query decides. Both follow Arora et al. (2024), with 1000 trials per condition, three seeds and chance rates near 0.05% for copying and 0.5% for MQAR. **Needle-in-a-haystack** reads a single token from a long context, with 100 trials per cell over five depths and five context lengths. Zero-shot accuracy uses lm-evaluation-harness (Gao et al., 2023) on six tasks. Appendix J defines each evaluation. Differences within a pair are tested with paired tests on identical prompts, Bonferroni-corrected (Appendix G).

### 4.2 Training loss and perplexity

**Value rotation incurs only a small cost at the training length.** We first ask what the rotation costs where both models were trained. Training loss, zero-shot accuracy and training speed are nearly unchanged at every scale (Tables 1, 12 and 14). As the second prediction expects, the small cost that remains is not spread evenly. The rotated model loses ground towards the end of each training block, and it loses most in the model whose heads read farthest (Figure 3(a) and Table 5). Later positions have more context to average over, which is where the filter should cost most.

### 4.3 Head-level attenuation

**Trained heads cancel as the analysis predicts, and the rotated models adapt around it.** We next measure the cancellation directly, by recomputing every head's attention on PG-19 and evaluating its attenuation at each rotary frequency (Appendix J). For a QK-RoPE model this is a counterfactual, the loss its heads would suffer if their values were rotated. In every model most heads average over long distances, so rotating their values would remove a substantial part of what they read, and the rotated models lose at least as much. Training with the rotation does not prevent the cancellation. Instead, the rotated models concentrate most of their head output in the low-frequency rotary pairs, where the analysis predicts that content is preserved under averaging (Figure 7).

### 4.4 Positional recall

**Copying costs appear only far back.** We next ask whether value rotation costs most when the content a head needs lies far away. At short distances the rotation leaves copying almost unchanged from 370M to 1.3B (Table 1). The cost appears where the first prediction places it, when the copied content lies thousands of tokens back and only a few rotary pairs remain readable (Figure 2 and Table 11). There the rotated 370M and 760M models copy less than their baselines, although they match them at shorter distances. At 1.3B the unrotated model, the only one trained with QK-normalization, fails first.

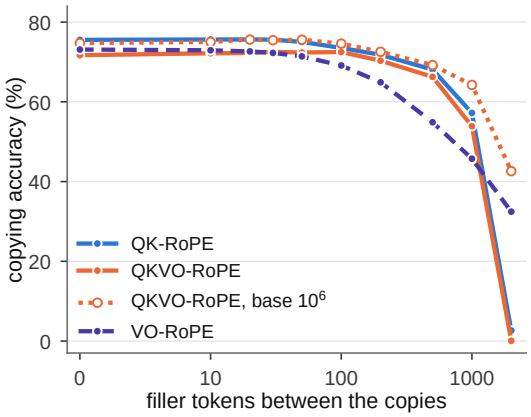



Figure 2: Copying accuracy of the 130M models by the distance between the two copies (ordinary subword tokens, mean over three seeds). Both base-$10^4$ models fail 2000 tokens back, whereas QKVO-RoPE at base $10^6$ and VO-RoPE keep part of the content. Every variant is in Figure 6.

**Associative recall shows the second predicted cost.** An induction head reads from a distance that changes with the query, so no fixed output projection can undo the phase of what it reads. Associative recall accordingly falls at 370M. It rises at 760M, the only pair with larger heads and hence more rotary pairs per head, and is unchanged at 1.3B (Table 17).

**The lag phase alone carries position.** To ask which rotations are needed, we compare nine combinations of rotating Q, K, V and de-rotating O at 130M (Appendix C). QK-RoPE and QKVO-RoPE train best, and three single rotations diverged. As the third prediction expects, VO-RoPE, whose attention score carries no position, still copies nearly as well as QK-RoPE and far better than a model without positional encoding, and it keeps part of the content thousands of tokens back (Figure 2). Contrary to the same prediction, QKV-RoPE copies worse than QKVO-RoPE.

### 4.5 Length extrapolation

**Beyond the training length the rotation helps up to 760M and hurts at 1.3B.** We now test the two opposing effects of Section 3.4. Up to 760M the rotated models degrade more slowly than their baselines on both PG-19 and LongCrawl64, and the gap widens with the context length (Figure 3, Tables 13 and 15). At 1.3B the order reverses.

**The 1.3B reversal lies in the attention score, not in the value path.** To locate it, we rescale the rotary frequencies of the score and of the value path separately (Table 18). Rescaling only the rotated model's score brings it level with the unrotated model at the longest context, whereas rescaling only its value path leaves most of the gap. The rotation leaves the form of the score unchanged, but training with it changes the queries and keys the model learns, and at 1.3B these extrapolate worse. At 760M, rescaling only the value path makes the rotated model worse, so there its out-of-range phases help rather than hurt, as the second effect of Section 3.4 expects. The analysis of Section 3 concerns the value path and does not predict the reversal. The 1.3B pair is also the only one trained with QK-normalization, itself a change to the score, so the reversal may come from how value rotation interacts with QK-normalization rather than from scale. Rescaling both paths together, as NTK-aware rescaling does (Peng et al., 2024), removes most of the difference at every scale (Table 16).

### 4.6 Needle in a haystack

**Rotation helps where one position must be read against a diffuse context.** A needle is a single token read from one position in a long context, the case in which Section 3.3 expects the rotation to help. Beyond the training length the rotated 760M and 1.3B models still retrieve needles, while their unrotated counterparts retrieve none, a difference that remains significant after correction (Table 6). The failure belongs to QK-RoPE, not to every unrotated model, since XPos and a model without positional encoding also retrieve some needles there (Table 3). Within the training length the rotation also helps at short contexts, although the rotated 130M model misses the deepest needles at 2K.

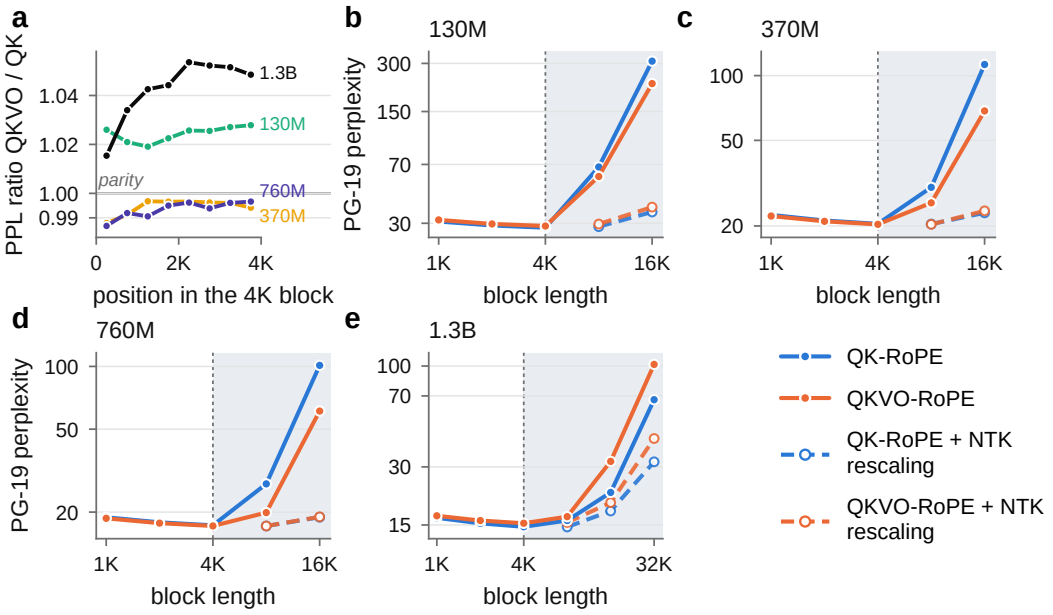



Figure 3: Panels include the 130M ablation pair alongside the three matched pairs. (a) Ratio of QKVO-RoPE to QK-RoPE PG-19 perplexity by position inside the 4096-token block. (b–e) PG-19 perplexity by evaluation block length. The shaded region lies beyond the training length and dashed series apply NTK-aware base rescaling with scale = block length/4096 to both models. LongCrawl64 is in Figure 8.

The 1.3B pair extrapolates worse in perplexity and better in needle retrieval. The two measure different things. A needle is read by one head from one position, and every needle retrieved beyond the training length lies at a distance seen in training. The rotation suppresses what the head collects diffusely from the unseen rest of the context. Perplexity instead sums over all heads and, at 1.3B, is limited by the rotated model's score.

### 4.7 A LARGER ROTARY BASE

**A larger rotary base removes most of the cost.** The analysis predicts that the cost of value rotation falls as the base grows, because more rotary pairs then survive averaging. We trained one further 130M QKVO-RoPE model at base $10^6$ instead of $10^4$, with the recipe of the 130M pair (Table 7). Its heads cancel less, as predicted, and the gain appears where the cost was largest, far back in the context. Where both base-$10^4$ models fail to copy, the larger-base model still recovers much of the content (Figure 2), and it retrieves more needles at long contexts without losing zero-shot accuracy. A larger base also changes the attention score and is known to extend the usable context of RoPE models (Xu et al., 2024; Su, 2024). The head-level measurement concerns the value path alone, but a QK-RoPE model at base $10^6$ is needed to attribute the other gains to it.

## 5 DISCUSSION

**When should the values be rotated?** Our results suggest a simple rule. Value rotation helps heads that read one position against a diffuse background, which matters for retrieval from long contexts and for reading beyond the training length. It costs heads whose content is itself an average, and heads such as induction heads whose source lies at a distance that changes with the query. Averaging heads can be found from a trained model's attention distances before any retraining (Section 4.3).

**How the cost could be reduced.** The analysis points to three ways to keep the positional signal while limiting the cancellation. First, a larger base raises the share of rotary pairs that survive averaging, which a 130M model at base $10^6$ supports pending a QK-RoPE control at the same base (Section 4.7).

Second, the value path could use more frequencies than the score, since the number of surviving pairs grows with the number of rotated pairs. Third, a head could carry both a rotated and an unrotated copy of its values, as partial rotation does in DeepSeek-V4 (DeepSeek-AI, 2026) and in latent attention (Appendix I), so that the model can learn when to read with the phase. The phase should also help tasks that need the position of what was read, such as finding the token just before the most recent occurrence of a query.

## 6 Conclusion

We studied what happens when RoPE is applied not only to the queries and keys of attention but also to its values and outputs. The rotation multiplies every attended value by a factor that oscillates with its distance from the query, the oscillation of a complex pole without its decay, so each head reads its values through a filter over distance. Across four matched pairs from 130M to 1.3B parameters, the filter helps where one token must be retrieved from a long, diffuse context, and beyond the training length the larger rotated models retrieve needles that the unrotated QK-RoPE models miss entirely. It costs where content is averaged or read from a far or query-dependent distance.

Value rotation is a positional signal to be placed where it helps. It should be used where a model must retrieve from long contexts, and its cost falls as more rotary frequencies survive on the value path. A larger base, more frequencies on the value path than on the score, and heads that keep both a rotated and an unrotated copy of their values are natural next steps.

### AI use statement

We did not use generative AI tools to generate research ideas, design the models or experiments, or produce any results. The authors used an AI tool in some parts of the paper to check for grammar mistakes. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including the text, claims, and artifacts produced with the aid of generative AI.

### Reproducibility statement

Derivations of every result in Section 3 are given in Appendix A. Section 4.1 and Table 9 give the training configuration of every model, and Appendix J defines each evaluation exactly, including the tokenization of the perplexity datasets, the construction of the recall and needle prompts, the NTK-aware rescaling rule, the throughput protocol and the head-level measurements. Every result file records the hashes of the model configuration and evaluation script, its arguments and the library versions. The evaluation harness, the analysis scripts that produce every table and figure, and the per-run result files are provided as supplementary material.

## References

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024.

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

Jacob Buckman. LongCrawl64: A long-context natural-language dataset. `https://manifestai.com/articles/longcrawl64/`, 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.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, 2024.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Leo Gao, Jonathan Tow, Baber Abbasi, et al. A framework for few-shot language model evaluation. `https://github.com/EleutherAI/lm-evaluation-harness`, 2023.

Alejandro García-Castellanos, Maurice Weiler, and Erik J. Bekkers. RoVE: Rotary value embeddings attention for relative position-dependent value pathways. *arXiv preprint arXiv:2606.11275*, 2026.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

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

Aakash Lahoti, Kevin Y Li, Berlin Chen, Caitlin Wang, Aviv Bick, J Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, 2026.

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

Beicheng Lou, Zifei Xu, and Vivian W. H. Wong. CRoPE: Efficient parametrization of rotary positional embedding. *arXiv preprint arXiv:2601.02728*, 2026.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, 2023.

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. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 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.

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.

Jianlin Su. The road to transformer upgrade: 18. the principle of choosing the base of RoPE. Blog post, `https://kexue.fm/archives/10122`, 2024.

Jianlin Su. The road to transformer upgrade: 21. what makes MLA good? (Part 2). Blog post, `https://kexue.fm/archives/11111`, 2025a.

Jianlin Su. The road to transformer upgrade: 19. second type of rotary position encoding. Blog post, `https://kexue.fm/archives/10862`, 2025b.

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

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023a.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon 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*, 2023b.

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

Songlin Yang and Yu Zhang. FLA: A Triton-based library for hardware-efficient implementations of linear attention mechanism. `https://github.com/fla-org/flash-linear-attention`, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025a.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. PaTH attention: Position encoding via accumulating Householder transformations. In *Advances in Neural Information Processing Systems*, 2025b.

Yu Zhang and Songlin Yang. Flame: Flash language modeling made easy. `https://github.com/fla-org/flame`, 2025.

# A PROOFS

This appendix derives the results of Section 3 in the order in which they appear there. Each follows from the fact that the blocks of $\mathbf{R}_m$ are rotations, so that $\mathbf{R}_m^\top \mathbf{R}_n = \mathbf{R}_{n-m}$.

**State-space form.** Let attention be unnormalized linear attention with scalar decay, $a_{mn} = \alpha^{m-n}\, \tilde{q}_m^\top \tilde{k}_n$ with $\alpha \in (0,1)$, where $\tilde{q}_m = \mathbf{R}_m q_m$ and $\tilde{k}_n = \mathbf{R}_n k_n$. The QKVO-RoPE output is then $o_m^{(j)} = q_m^\top S_m^{(j)}$, with the state $S_m^{(j)} \in \mathbb{C}^d$ obeying

$$S_m^{(j)} = \alpha\, e^{-i\theta_j}\, \mathbf{R}_{-1} S_{m-1}^{(j)} + k_m\, v_m^{(j)}. \tag{5}$$

The recurrence follows from writing the output as a sum over past positions. The blocks are rotations, so $\mathbf{R}_m^\top \mathbf{R}_n = \mathbf{R}_{n-m}$ and $\tilde{q}_m^\top \tilde{k}_n = q_m^\top \mathbf{R}_{n-m} k_n$. Substituting into equation 1 and taking pair $j$ of the value gives $o_m^{(j)} = \sum_{n \le m} \alpha^{m-n} e^{-i\theta_j (m-n)}\, q_m^\top \mathbf{R}_{n-m} k_n\, v_n^{(j)} = q_m^\top S_m^{(j)}$ with $S_m^{(j)} = \sum_{n \le m} \alpha^{m-n} e^{-i\theta_j (m-n)} \mathbf{R}_{n-m} k_n v_n^{(j)}$. Splitting off the $n = m$ term and writing $\mathbf{R}_{n-m} = \mathbf{R}_{-1}\mathbf{R}_{n-m+1}$ gives equation 5. Here $\mathbf{R}_{-1}$ rotates key pair $j'$ by $-\theta_{j'}$, a real rotation with eigenvalues $e^{\pm i\theta_{j'}}$. Without the value rotation the factor $e^{-i\theta_j(m-n)}$ is absent.

**Lag-phased readout, equation 2.** The readout follows directly from equation 1. Because the blocks are rotations, $\mathbf{R}_m^\top \mathbf{R}_n = \mathbf{R}_{n-m}$ and $(\mathbf{R}_{n-m} v_n)^{(j)} = e^{i(n-m)\theta_j} v_n^{(j)}$, and substituting $\ell = m - n$ into equation 1 gives equation 2.

**Fixed-offset invariance.** The result follows from equation 2 with all attention on the lag $\ell^\star$, because $\mathbf{R}_{-\ell^\star}$ does not depend on $m$ and can therefore be absorbed into the output projection.

**Attenuation by the lag distribution, equation 4.** Shared content is scaled by $\varphi_m(\theta_j)$, obtained by factoring $v^{(j)}$ out of equation 2. The bound $|\varphi_m| \le 1$ is the triangle inequality with $\sum_\ell a = 1$, and equality requires all phases to coincide. For values that are independent with zero mean and independent of the weights, the cross terms vanish in expectation, so $\mathbb{E}|\sum_\ell a_\ell e^{-i\theta_j \ell} v_\ell^{(j)}|^2 = \sum_\ell a_\ell^2\, \mathbb{E}|v_\ell^{(j)}|^2$, the same as without rotation. Equation 4 is the geometric sum. $W\theta_j < 1$ with $\theta_j = b^{-2j/d}$ gives $j > (d/2) \ln W / \ln b$.

**Out-of-range lags.** We show that on the lowest-frequency pairs a lag just beyond the training length produces an angle not seen in training. The angles $\{\theta_j \ell' : \ell' \le L\}$ fill an arc of length $\theta_j L$. If $\theta_j L \ge 2\pi$ the arc is the whole circle and no angle is outside it. If $\theta_j L < 2\pi$ it is not, and $\theta_j \ell$ for $\ell$ just above $L$ lies past its end. Sufficiently large $\ell$ can wrap back into the arc, which is why the statement is restricted to lags near $L$. $\theta_j L < 2\pi$ with $\theta_j = b^{-2j/d}$ gives $j > (d/2)\ln(L/2\pi)/\ln b$.

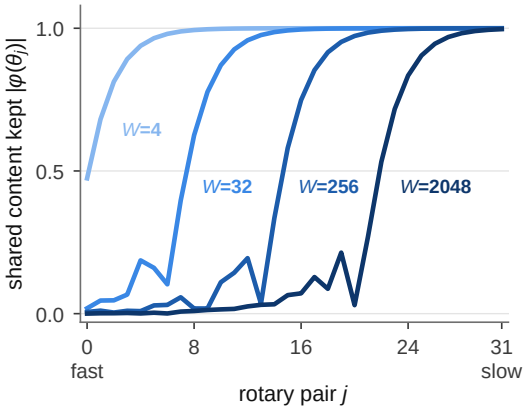



Figure 4: Attenuation $|\varphi(\theta_j)|$ of a uniform lag window of width $W$ at the 32 rotary frequencies of a 64-dimensional head, base $10^4$ (Section 3.3). Shared content is retained only in pairs with $W\theta_j \lesssim 1$.

## B Positional-encoding controls

We compare the matched 370M pair with two positional-encoding controls, XPos and a model with no positional encoding (NoPE), to establish what the two synthetic retrieval tasks measure. They measure different capabilities (Table 2). Associative recall is addressed by content, and the model with no positional encoding has the best associative recall of the five models at 28.02%. Copying needs a positional signal, which a model without positional encoding can learn from the causal mask (Kazemnejad et al., 2023). Trained on 86B tokens it copies 71.9% at short distances and 49.9% with 2000 filler tokens, whereas the 130M model without positional encoding copies 10.7% (Table 4). On copying the rotated 370M model matches its baseline (77.2 against 76.8%), so a variant can score lower on associative recall without losing positional ability. The 7-point MQAR loss of the rotated 370M model (Section 4.4) therefore need not reflect a weaker positional channel. XPos and the model with no positional encoding were trained on eight times as many tokens as the matched pair (86B against 10.7B), so they serve here to characterize the two retrieval tasks, not to rank perplexity.

Table 2: Positional-encoding controls at 370M. XPos and the NoPE run were trained for 86B tokens and the matched pair for 10.7B. QK-RoPE (86B) is the token-matched reference for XPos and NoPE. Copy: ordinary subword tokens, filler lengths 0–50.

| model | tokens (B) | loss | WT103 | PG-19 4K | 8K | 16K | Copy | MQAR | downstream |
|---|---|---|---|---|---|---|---|---|---|
| QK-RoPE | 10.7 | 2.496 | 23.55 | 20.46 | 30.2 | 112.6 | 76.77±0.19 | 27.20±1.24 | 45.3 |
| QKVO-RoPE | 10.7 | 2.501 | 23.84 | 20.35 | 25.6 | 68.5 | 77.23±0.14 | 20.19±0.89 | 46.1 |
| QK-RoPE (86B) | 85.9 | 2.332 | 20.70 | 18.63 | 43.8 | 170.0 | 75.09±0.05 | 22.22±0.96 | 48.7 |
| XPos | 85.9 | 2.330 | 20.46 | 18.90 | 18.7 | 22.4 | 73.81±0.16 | 23.37±1.13 | 48.7 |
| NoPE (86B) | 85.9 | 2.455 | 22.97 | 21.42 | 52.0 | 654.0 | 71.89±0.11 | 28.02±0.62 | 45.9 |

Needle retrieval of the controls is in Table 3. At 8K, twice the training length, XPos retrieves 13.2% and the model with no positional encoding 16.6%, whereas both members of the 370M pair and the QK-RoPE model trained on the same 86B tokens retrieve none. The absence of retrieval at 8K is therefore a property of QK-RoPE, not of every model without value rotation. At 130M, VO-RoPE is the best of the rotation variants at 2K and 4K (89.6 and 61.6%).

## C Rotation ablation

To ask which of the four rotations are needed, we train nine of the sixteen combinations of rotating Q, K, V and de-rotating O at 130M. QK-RoPE and QKVO-RoPE reach the lowest losses of the variants at 130M, and VO-RoPE behaves as Section 3.5 predicts, while QKV-RoPE does not (Table 4). K-RoPE trains to 2.927, because a rotated key is an absolute positional code that an unrotated query can still read.

Table 3: Needle-in-a-haystack accuracy (%) of the controls by context length, averaged over five depths, 100 trials per cell. Models trained at 4K.

| scale | model | 1K | 2K | 4K | 8K | 16K |
|---|---|---|---|---|---|---|
| 370M | QK-RoPE | 82.6 | 75.0 | 53.0 | 0.0 | 0.0 |
| | QKVO-RoPE | 90.2 | 88.4 | 51.4 | 0.0 | 0.0 |
| | XPos (86B) | 82.6 | 78.4 | 54.6 | 13.2 | 0.0 |
| | NoPE (86B) | 82.6 | 66.8 | 40.2 | 16.6 | 0.0 |
| | QK-RoPE (86B) | 70.8 | 69.4 | 47.2 | 0.0 | 0.0 |
| 130M | QK-RoPE | 92.0 | 79.6 | 43.0 | 0.0 | 0.0 |
| | QKVO-RoPE | 97.2 | 62.8 | 38.4 | 0.0 | 0.0 |
| | K-RoPE | 83.6 | 81.6 | 51.6 | 0.0 | 0.0 |
| | QKV-RoPE | 96.0 | 74.4 | 37.2 | 0.0 | 0.0 |
| | VO-RoPE | 92.4 | 89.6 | 61.6 | 0.0 | 0.0 |
| | NoPE (43B) | 83.6 | 59.4 | 19.2 | 3.2 | 0.0 |

Table 4: All combinations of rotating Q, K, V and de-rotating O at 130M on FineWeb-Edu. Copy: ordinary subword tokens, filler lengths 0–50, as in Table 1. Q-, V- and O-RoPE diverged in training (Appendix C).

| model | Q | K | V | $O^{-1}$ | tok (B) | loss | WT103 | PG 4K | PG 16K | Copy | MQAR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NoPE | · | · | · | · | 42.9 | 2.829 | 34.19 | 30.79 | 136.6 | 10.74±0.27 | 10.18±0.17 |
| Q-RoPE | ✓ | · | · | · | 5.4 | 4.567 | 311.88 | 150.83 | 3,126.1 | 0.04±0.00 | 0.08±0.07 |
| K-RoPE | · | ✓ | · | · | 5.4 | 2.927 | 40.03 | 32.61 | 183.6 | 52.42±0.02 | 19.01±0.91 |
| V-RoPE | · | · | ✓ | · | 5.4 | 5.578 | 889.71 | 374.47 | 727.3 | 0.00±0.00 | 0.00±0.00 |
| O-RoPE | · | · | · | ✓ | 5.4 | 4.834 | 474.92 | 215.29 | 3,775.2 | 0.00±0.00 | 0.00±0.00 |
| QKV-RoPE | ✓ | ✓ | ✓ | · | 5.4 | 2.884 | 36.40 | 30.82 | 380.6 | 60.75±0.15 | 18.96±0.72 |
| VO-RoPE | · | · | ✓ | ✓ | 5.4 | 2.881 | 35.94 | 30.68 | 1,410.2 | 72.47±0.19 | 25.78±1.24 |
| QK-RoPE | ✓ | ✓ | · | · | 5.4 | 2.778 | 32.02 | 28.15 | 309.9 | 75.48±0.04 | 26.01±1.43 |
| QKVO-RoPE | ✓ | ✓ | ✓ | ✓ | 5.4 | 2.815 | 33.83 | 28.83 | 224.3 | 72.20±0.13 | 21.34±1.03 |

QKV-RoPE leaves the phase of the query position in every output, which acts as an absolute positional code. It raises the loss by 0.106, against 0.037 for the full construction. Contrary to the prediction of Section 3.5, it does not copy well. It reaches 60.8% at short distances, below QK-RoPE (75.5%), QKVO-RoPE (72.2%) and VO-RoPE (72.5%), and 0.9% with 2000 filler tokens. The absolute phase therefore does not serve as a useful copying signal at this scale. VO-RoPE leaves the score content-only and rotates only the value path. It copies at 72.5%, where NoPE with eight times the tokens copies 10.7%, so a lag-phased readout alone is a usable positional channel. With 2000 filler tokens it still copies 32.4%. Among the variants that train, it degrades most beyond the training length, because that channel is its only positional signal and all of its phases are then out of range.

## D HEAD-LEVEL, NEEDLE AND ROTARY-BASE TABLES

Table 5: Head-level test of Section 3.3. The table describes where each model keeps its value content and how much of it the lag distribution removes. Shared lost: fraction of content shared by all attended positions that the rotation removes ($1 - |\varphi|$ weighted by the value energy in each pair). Meas. lost: fraction of readout energy removed on the actual values. Dist. $> 256$: fraction of heads with mean attention distance above 256. $v$, $o$, $W_o$ low: share of value energy, head-output energy and output-projection weight in the slow half of the pairs.

| scale | model | mean $\lvert\varphi\rvert$ | shared lost | meas. lost | dist. $> 256$ | $v$ low | $o$ low | $W_o$ low |
|---|---|---|---|---|---|---|---|---|
| 370M | QK-RoPE | 0.557 | 0.443 | 0.200 | 0.78 | 0.500 | 0.502 | 0.500 |
| | QKVO-RoPE | 0.532 | 0.431 | 0.287 | 0.83 | 0.538 | 0.744 | 0.538 |
| 760M | QK-RoPE | 0.593 | 0.407 | 0.176 | 0.79 | 0.498 | 0.499 | 0.500 |
| | QKVO-RoPE | 0.546 | 0.430 | 0.255 | 0.89 | 0.530 | 0.715 | 0.554 |
| 1.3B | QK-RoPE | 0.569 | 0.430 | 0.201 | 0.95 | 0.501 | 0.503 | 0.500 |
| | QKVO-RoPE | 0.549 | 0.420 | 0.296 | 0.98 | 0.544 | 0.732 | 0.568 |

Table 6: Needle-in-a-haystack accuracy (%) by context length, averaged over five needle depths. Models trained at 4K. The 130M ablation pair is included for reference.

| scale | model | 1K | 2K | 4K | 8K | 16K |
|---|---|---|---|---|---|---|
| 130M | QK-RoPE | 92.0 | 79.6 | 43.0 | 0.0 | 0.0 |
| | QKVO-RoPE | 97.2 | 62.8 | 38.4 | 0.0 | 0.0 |
| 370M | QK-RoPE | 82.6 | 75.0 | 53.0 | 0.0 | 0.0 |
| | QKVO-RoPE | 90.2 | 88.4 | 51.4 | 0.0 | 0.0 |
| 760M | QK-RoPE | 84.8 | 83.6 | 64.0 | 0.0 | 0.0 |
| | QKVO-RoPE | 89.4 | 93.8 | 65.6 | 14.0 | 0.0 |
| 1.3B | QK-RoPE | 86.0 | 91.6 | 59.2 | 0.0 | 0.0 |
| | QKVO-RoPE | 92.4 | 90.4 | 61.2 | 17.6 | 0.0 |

Table 7: Rotary base $10^6$ against $10^4$ at 130M. Mean $|\varphi|$ and shared lost as in Table 5 (for QK-RoPE, the counterfactual). Copy: ordinary subword tokens, three seeds. "Near" averages fillers of 0–50 tokens, and 1K and 2K are filler lengths. Needle: accuracy at 2K and 4K context. The base-$10^6$ model was trained on one GPU with a different data order.

| model | base | mean $\|\varphi\|$ | shared lost | PG-19 4K | PG-19 16K | copy near | copy 1K | copy 2K | needle 2K | needle 4K | zero-shot |
|---|---|---|---|---|---|---|---|---|---|---|---|
| QK-RoPE | $10^4$ | 0.576 | 0.425 | 28.15 | 309.9 | 75.5 | 57.2 | 2.6 | 79.6 | 43.0 | 40.7 |
| QKVO-RoPE | $10^4$ | 0.557 | 0.406 | 28.83 | 224.3 | 72.2 | 53.9 | 0.1 | 62.8 | 38.4 | 40.7 |
| QKVO-RoPE | $10^6$ | 0.710 | 0.282 | 26.69 | 131.8 | 75.3 | 64.2 | 42.6 | 95.8 | 82.8 | 40.4 |

## E GATED DELTANET

Table 8: Gated DeltaNet at 130M with no rotation, query-key rotation, and query-key-value-output rotation, 5.0B tokens each. Evaluated with the training library (one recall seed). Copy uses the probe with byte-fallback tokens (Appendix J), so its values are not comparable with the copy columns of the other tables.

| model | loss | WT103 | PG-19 | Copy | MQAR | downstream |
|---|---|---|---|---|---|---|
| GDN | 2.757 | 38.45 | 26.78 | 5.06 | 20.97 | 40.6 |
| GDN+QK-RoPE | 2.777 | 39.72 | 29.08 | 1.47 | 18.30 | 40.2 |
| GDN+QKVO-RoPE | 2.789 | 40.17 | 29.19 | 1.82 | 17.50 | 40.3 |

We ask whether value rotation benefits a recurrence that already has its own decay and gate. Gated DeltaNet (Yang et al., 2025a) is a linear recurrence with a data-dependent gate and a delta-rule update. Section 3.2 does not cover it, because its state update is not diagonal. We trained three 130M models for 5.0B tokens each under the protocol of Section 4.1, one without rotation, one with rotated queries and keys and one with all four rotations. The rotation inside the recurrence is implemented in the training library, so Table 8 reports evaluations run with that library rather than with the harness of Appendix J. For the unrotated model the two agree (WikiText-103 38.45 against 38.45, PG-19 26.78 against 26.79, copying 5.06 against 5.15, MQAR 20.97 against 20.96). The unrotated model's perplexity stays flat beyond its training length (PG-19 26.79 at 4K and 25.83 at 16K), as expected for a recurrence without a positional table.

Neither query-key rotation nor full rotation benefits gated DeltaNet at 130M (Table 8). The training loss rises from 2.757 to 2.777 with query-key rotation and to 2.789 with all four rotations, WikiText-103 perplexity rises from 38.45 to 39.72 and 40.17, copying on the byte-token probe falls from 5.06% to 1.47% and 1.82%, and MQAR falls from 20.97% to 18.30% and 17.50%, while zero-shot accuracy is unchanged (40.6, 40.2 and 40.3). A recurrence that already has its own decay and gate therefore does not benefit from fixed-frequency poles at 130M.

## F EXTENDED RESULTS

Tables 9–16 and Figures 6–9 give the per-condition results behind Section 4. Figure 6 extends two of the 130M comparisons beyond the conditions of the main text, to copying with up to 2000 filler tokens and to perplexity at 16K for every rotation variant.

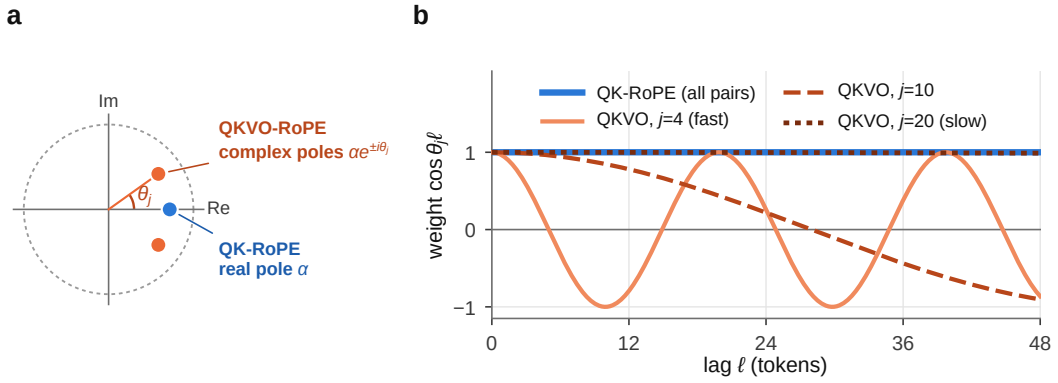



Figure 5: The same rotation admits two readings. (a) Under linear attention with decay $\alpha$ the value path is a recurrence whose real pole becomes a conjugate pair (Section 3.2). (b) Under softmax attention the recurrence is gone, but each pair still weights lag $\ell$ by $\cos\theta_j\ell$ (equation 2), so a high-frequency pair alternates sign within tens of tokens while a low-frequency one passes content unchanged.

At 130M copying declines with distance for every variant, and most sharply between 1000 and 2000 filler tokens (Figure 6(a)). With 100 filler tokens QK-RoPE copies 73.5%, QKVO-RoPE 72.5%, VO-RoPE 69.1% and QKV-RoPE 50.8%. With 1000 the four reach 57.2, 53.9, 45.7 and 31.0%, and with 2000 QK-RoPE, QKVO-RoPE and QKV-RoPE fail (2.7, 0.1 and 0.9%) while VO-RoPE keeps 32.4% and QKVO-RoPE at base $10^6$ keeps 42.6%. The unrotated QK-RoPE model fails as well, so this failure is not specific to value rotation. For the value path, the copying lag is the filler length plus the span, and Section 3.3 shows that content in pair $j$ is readable only while $\theta_j\ell^\star \ll 1$, which at a lag of 2000 leaves five of the 32 pairs at base $10^4$ (Table 19).

Among the variants that rotate both queries and keys, QKVO-RoPE degrades most slowly beyond the training length. At 16K on PG-19 the 130M variants order as NoPE (136.6, trained on eight times the tokens), K-RoPE (183.6), QKVO-RoPE (224.3), QK-RoPE (309.9), QKV-RoPE (380.6) and VO-RoPE (1410). Among the variants that train, QKV-RoPE and VO-RoPE degrade fastest, because every positional signal they carry in the value path is out of range beyond the training length.

Table 9: Matched pairs. Both members of a pair share every setting except the value-path rotation. Peak learning rate, warm-up steps and weight decay (wd) of AdamW, from the training logs. The learning rate decays with a cosine schedule to 3% of its peak (10% at 1.3B).

| scale | model | layers | $d_{\text{model}}$ | $d_{\text{head}}$ | tokens (B) | ctx | peak lr | warm-up | wd | loss |
|---|---|---|---|---|---|---|---|---|---|---|
| 130M | QK-RoPE | 12 | 768 | 64 | 5.4 | 4096 | $10^{-3}$ | 512 | 0.01 | 2.778 |
| 130M | QKVO-RoPE | 12 | 768 | 64 | 5.4 | 4096 | $10^{-3}$ | 512 | 0.01 | 2.815 |
| 370M | QK-RoPE | 24 | 1024 | 64 | 10.7 | 4096 | $10^{-3}$ | 1024 | 0.01 | 2.496 |
| 370M | QKVO-RoPE | 24 | 1024 | 64 | 10.7 | 4096 | $10^{-3}$ | 1024 | 0.01 | 2.501 |
| 760M | QK-RoPE | 24 | 1536 | 96 | 20.1 | 4096 | $2 \times 10^{-3}$ | 2048 | 0.01 | 2.265 |
| 760M | QKVO-RoPE | 24 | 1536 | 96 | 20.1 | 4096 | $2 \times 10^{-3}$ | 2048 | 0.01 | 2.271 |
| 1.3B | QK-RoPE | 24 | 2048 | 64 | 30.0 | 4096 | $8 \times 10^{-4}$ | 572 | 0.1 | 2.117 |
| 1.3B | QKVO-RoPE | 24 | 2048 | 64 | 30.0 | 4096 | $8 \times 10^{-4}$ | 572 | 0.1 | 2.119 |

## G SIGNIFICANCE TESTS AND PATH-SEPARATED RESCALING

Both members of a pair are evaluated on identical prompts, so their per-trial outcomes are paired. Table 17 tests the copying and MQAR averages of Table 1 and the 130M pair. Copying is tested with a Wilcoxon signed-rank test on the per-trial accuracy over 15,000 prompts (five filler lengths, three seeds of 1000, ordinary subword tokens), MQAR with an exact McNemar test over 9,000 prompts, and a Bonferroni correction is applied over the eight comparisons. The per-trial MQAR

Table 10: Copying and MQAR accuracy (%) by condition, mean ± standard deviation over three seeds of 1000 trials each. $k$ is the number of filler tokens, $p$ the number of key-value pairs. Chance is about 0.05% for copying and 0.5% for MQAR. Copying uses ordinary subword tokens (Appendix J), and the long-distance conditions are in Table 11.

| scale | model | $k$=0 | 10 | 20 | 30 | 50 | $p$=4 | 8 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| 130M | QK-RoPE | 75.54±0.44 | 75.63±0.42 | 75.64±0.48 | 75.59±0.16 | 75.03±0.12 | 25.37±1.46 | 27.83±1.87 | 24.83±1.80 |
| | QKVO-RoPE | 71.70±0.52 | 72.17±0.33 | 72.24±0.52 | 72.53±0.13 | 72.36±0.37 | 19.73±1.19 | 22.57±0.95 | 21.73±1.16 |
| 370M | QK-RoPE | 75.44±0.39 | 76.52±0.27 | 77.37±0.10 | 77.19±0.58 | 77.33±0.52 | 29.20±1.70 | 28.60±2.51 | 23.80±0.20 |
| | QKVO-RoPE | 76.51±0.34 | 77.25±0.21 | 77.56±0.16 | 77.32±0.39 | 77.53±0.15 | 19.13±1.53 | 21.60±0.89 | 19.83±0.49 |
| 760M | QK-RoPE | 77.97±0.47 | 78.57±0.49 | 79.13±0.38 | 78.83±0.55 | 78.99±0.26 | 26.37±1.97 | 25.00±1.74 | 21.20±0.26 |
| | QKVO-RoPE | 79.47±0.18 | 79.89±0.27 | 80.05±0.14 | 79.94±0.56 | 79.75±0.17 | 30.10±1.44 | 28.77±0.84 | 24.03±0.83 |
| 1.3B | QK-RoPE | 80.40±0.33 | 81.57±0.15 | 81.99±0.43 | 81.84±0.25 | 81.59±0.06 | 30.33±1.44 | 29.80±0.82 | 24.80±1.25 |
| | QKVO-RoPE | 81.90±0.30 | 82.72±0.23 | 83.51±0.25 | 83.46±0.41 | 83.08±0.07 | 31.53±1.95 | 27.63±1.27 | 24.63±0.70 |

Table 11: Copying accuracy (%) at long distances, ordinary subword tokens, mean ± standard deviation over three seeds of 500 trials each. $k$ is the number of filler tokens between the two copies.

| scale | model | $k$=100 | 200 | 500 | 1000 | 2000 |
|---|---|---|---|---|---|---|
| 130M | QK-RoPE | 73.52±0.56 | 71.78±0.51 | 68.08±0.45 | 57.21±0.53 | 2.65±0.12 |
| | QKVO-RoPE | 72.50±0.13 | 70.31±0.88 | 66.26±0.30 | 53.89±0.20 | 0.07±0.03 |
| 130M | QKVO-RoPE, base $10^6$ | 74.59±0.34 | 72.55±0.56 | 69.21±0.97 | 64.24±0.04 | 42.60±0.32 |
| 370M | QK-RoPE | 76.91±0.06 | 75.35±0.51 | 71.67±0.12 | 68.25±0.31 | 50.25±0.08 |
| | QKVO-RoPE | 77.83±0.25 | 76.32±0.27 | 73.45±0.53 | 68.57±0.03 | 46.42±0.35 |
| 760M | QK-RoPE | 78.89±0.71 | 78.83±0.47 | 76.21±0.69 | 72.26±0.23 | 67.36±0.51 |
| | QKVO-RoPE | 78.09±0.14 | 77.56±0.37 | 77.89±0.47 | 75.27±0.19 | 62.14±0.63 |
| 1.3B | QK-RoPE | 81.10±0.19 | 81.12±0.46 | 78.83±0.18 | 72.26±0.21 | 28.93±0.27 |
| | QKVO-RoPE | 82.97±0.13 | 82.62±0.34 | 81.21±0.22 | 77.02±0.24 | 61.09±0.27 |

outcomes were obtained by re-running the evaluation of Appendix J with per-trial output, which reproduces every MQAR accuracy of Table 10 to two decimals. The copying outcomes come from the ordinary-token evaluation. For needle retrieval, Fisher's exact test on 500 trials per context length gives Bonferroni-corrected $p$ (over the eight 1K and 8K comparisons) of $3.0 \times 10^{-3}$, $4.8 \times 10^{-3}$, 0.30 and $1.2 \times 10^{-2}$ at 1K for 130M, 370M, 760M and 1.3B, and below $10^{-20}$ at 8K for 760M and 1.3B. All tests treat the prompts as the random quantity.

Table 18 rescales the rotary frequencies with the NTK-aware rule of Appendix J in the score only, in the value path only, or in both, for the rotated models, on the first 4M tokens of PG-19. Perplexities differ slightly from Table 13, which uses the full test split.

## H Rotary frequencies and resonance lags

Table 19 lists, for a 64-dimensional head with base $10^4$, each rotary frequency $\theta_j$, the lag $\pi/\theta_j$ at which content arrives with opposite sign in pair $j$, the window width $1/\theta_j$ below which averaging leaves the pair almost unattenuated (Section 3.3), and whether the pair's phases wrap within a 4096-token training length (Section 3.4). The count $J_W$ of Section 3.3 depends on the threshold $|\varphi| > 0.95$: the transition from $W\theta_j = 1$ to $W\theta_j = 2\pi$, over which $|\varphi|$ falls to its first zero, spans a further $J \ln 2\pi / \ln b$ pairs, about six of 32 at $b = 10^4$.

## I Value rotation and latent attention

Value rotation lets multi-head latent attention use one rotated latent for both keys and values. Multi-head latent attention caches one latent $c_n$ per token and derives keys and values from it. Applying RoPE to the latent as a key encodes absolute position into the value path, which is why DeepSeek-V2 (DeepSeek-AI, 2024) keeps a separate small rotated key branch. With the value path rotated and de-rotated, $o_m = \mathbf{R}_m^\top \sum_n a_{mn} \mathbf{R}_n c_n$ and $s_{mn} = (\mathbf{R}_m q_m)^\top (\mathbf{R}_n c_n)$ both depend on relative position only, so one rotated latent can serve both paths (Su, 2025b;a). Su (2025a) reports that sharing keys and values with value rotation on 64 of 256 dimensions slightly improves the loss. By Section 3.3, a fully rotated shared latent would lose its high-frequency pairs in every aggregating head, which

Table 12: Zero-shot accuracy (%). acc_norm for HellaSwag and ARC-challenge, acc otherwise.

| scale | model | LAMBADA | PIQA | HellaSwag | WinoGrande | ARC-e | ARC-c | avg |
|---|---|---|---|---|---|---|---|---|
| 130M | QK-RoPE | 24.5 | 61.3 | 31.4 | 53.3 | 49.2 | 24.3 | 40.7 |
| | QKVO-RoPE | 24.4 | 61.7 | 30.7 | 53.7 | 49.7 | 24.1 | 40.7 |
| 370M | QK-RoPE | 33.6 | 66.3 | 38.5 | 50.5 | 56.2 | 26.6 | 45.3 |
| | QKVO-RoPE | 33.8 | 66.8 | 39.0 | 53.1 | 56.8 | 26.9 | 46.1 |
| 760M | QK-RoPE | 41.8 | 70.0 | 47.6 | 55.3 | 64.7 | 33.0 | 52.1 |
| | QKVO-RoPE | 41.5 | 69.6 | 47.8 | 54.7 | 63.7 | 31.9 | 51.5 |
| 1.3B | QK-RoPE | 48.3 | 72.6 | 55.7 | 57.6 | 69.6 | 37.0 | 56.8 |
| | QKVO-RoPE | 47.2 | 72.2 | 56.1 | 56.6 | 69.0 | 37.3 | 56.4 |

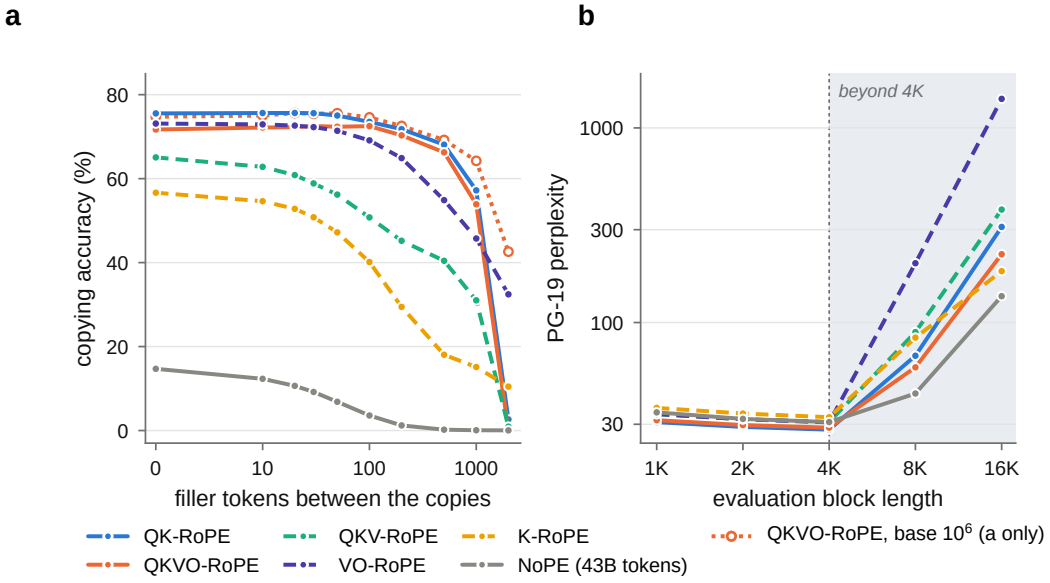



Figure 6: (a) Copying accuracy of the 130M variants with up to 2000 filler tokens (ordinary subword tokens, mean over three seeds), including QKVO-RoPE at base $10^6$. (b) PG-19 perplexity of the same variants by evaluation block length.

rotating only part of the dimensions avoids. DeepSeek-V4 likewise rotates only the last 64 dimensions of its shared key-value entries (DeepSeek-AI, 2026).

## J Evaluation details and provenance

This appendix defines each evaluation in detail to reproduce it and states how the head-level measurements of Section 4.3 are taken.

**Perplexity.** Documents of a split are tokenized with the model tokenizer (which prepends BOS), concatenated in dataset order into one stream, and cut into non-overlapping blocks of the evaluation length, and the remainder is dropped. The loss is the mean next-token cross-entropy over all predicted positions and perplexity is its exponential. For PG-19 the stream is the full test split (100 books, 11.13M tokens). For WikiText-103 it is the test split (327.6K scored tokens at 4096), and for LongCrawl64 the first 48 documents of the training split at a fixed dataset revision, cut at 2M tokens. NTK-aware rescaling multiplies the rotary base by $s^{d/(d-2)}$ with $s = \text{block length}/4096$ in every rotary module, so it applies to the value path as well as the score.

**Recall.** For copying, the prompt is BOS, 20 distinct tokens drawn from ids 1000–2999, $k$ filler tokens drawn with replacement from ids 3000–3999, and the 20 tokens again. Accuracy is the fraction of the 20 repeated positions at which the teacher-forced argmax equals the true token. For MQAR,

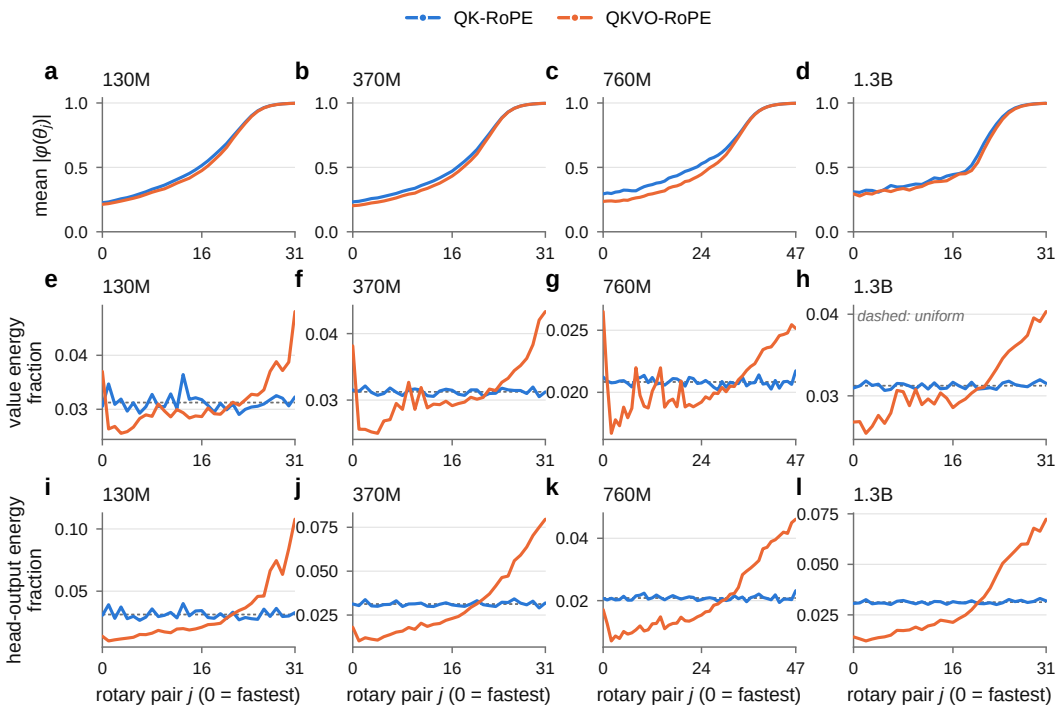



Figure 7: Mean attenuation $|\varphi(\theta_j)|$ over heads (a–d), value energy per rotary pair (e–h) and head-output energy per pair (i–l) for both members of each pair. The dashed line is the uniform fraction $1/J$.

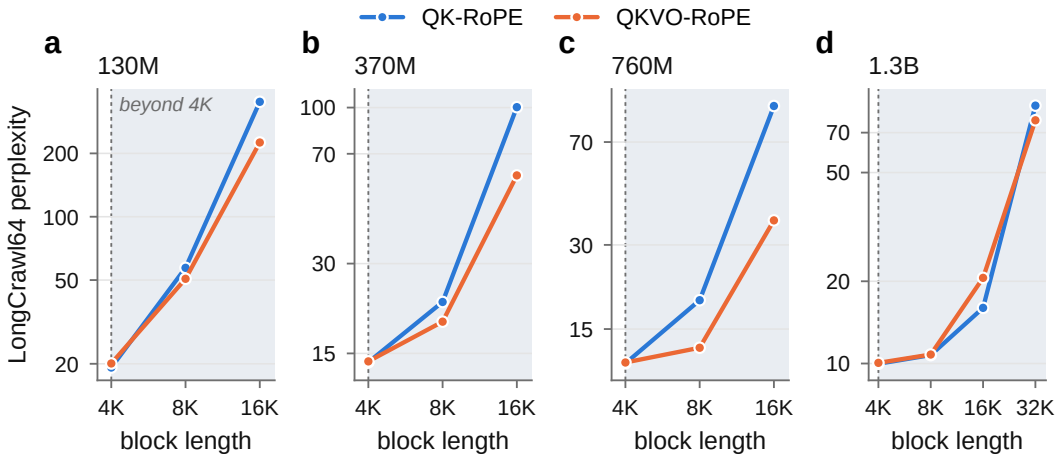



Figure 8: LongCrawl64 perplexity by evaluation block length. The shaded region lies beyond the 4096-token training length.

the prompt is BOS followed by $p$ key-value pairs (keys distinct from ids 400–599, values from ids 600–799) and one query key, and accuracy is whether the argmax at the last position equals the paired value. Each condition uses 1000 trials per seed and seeds 42, 123, 456. The table reports the mean and standard deviation across seeds of the per-seed accuracies. The long-distance conditions ($k = 100$ to 2000) use 500 trials per seed.

Table 13: PG-19 perplexity by evaluation block length and the ratio QKVO-RoPE/QK-RoPE. A ratio below 1 means the value-rotated model degrades less.

| scale | model | 1K | 2K | 4K | 8K | 16K | 32K |
|---|---|---|---|---|---|---|---|
| 130M | QK-RoPE | 30.8 | 29.1 | 28.1 | 67.5 | 309.9 | – |
| | QKVO-RoPE | 31.5 | 29.7 | 28.8 | 58.8 | 224.3 | – |
| | ratio QKVO/QK | 1.02 | 1.02 | 1.02 | 0.87 | 0.72 | – |
| 370M | QK-RoPE | 22.5 | 21.2 | 20.5 | 30.2 | 112.6 | – |
| | QKVO-RoPE | 22.2 | 21.0 | 20.3 | 25.6 | 68.5 | – |
| | ratio QKVO/QK | 0.99 | 0.99 | 0.99 | 0.85 | 0.61 | – |
| 760M | QK-RoPE | 18.9 | 17.9 | 17.3 | 27.3 | 101.2 | – |
| | QKVO-RoPE | 18.7 | 17.7 | 17.2 | 19.9 | 61.1 | – |
| | ratio QKVO/QK | 0.99 | 0.99 | 0.99 | 0.73 | 0.60 | – |
| 1.3B | QK-RoPE | 16.3 | 15.3 | 14.7 | 15.8 | 22.1 | 66.9 |
| | QKVO-RoPE | 16.7 | 15.8 | 15.3 | 16.6 | 32.0 | 101.9 |
| | ratio QKVO/QK | 1.02 | 1.03 | 1.04 | 1.05 | 1.45 | 1.52 |

Table 14: Training-step throughput of the two members of each pair on one RTX 3090 Ti, sequence length 4096, measured alternately in one process.

| scale | QK-RoPE tok/s | QKVO-RoPE tok/s | $\Delta$ (%) | peak memory GiB |
|---|---|---|---|---|
| 130M | 69,104 | 67,755 | -2.0 | 6.31 / 6.59 |
| 370M | 22,625 | 22,286 | -1.5 | 7.58 / 7.96 |
| 760M | 11,964 | 11,804 | -1.3 | 7.25 / 7.53 |
| 1.3B | 7,060 | 6,961 | -1.4 | 11.57 / 11.95 |

**Needle in a haystack.** The haystack is a fixed prose passage repeated, entered at a random offset per trial. The needle is three common three-letter words that are single tokens. It is inserted after a fraction (the depth) of the haystack, and the prompt ends with the first two words. The model is correct if its top-1 prediction is the third word. Each cell uses 100 trials with a seed fixed per cell, so all models see the same trials.

**Downstream.** Zero-shot accuracy uses lm-evaluation-harness with default prompts, and the reported metric is acc_norm for HellaSwag and ARC-challenge and acc otherwise.

**Throughput.** Both members of a pair are loaded in one process on one RTX 3090 Ti in training mode. After two warm-up rounds, five rounds of eight forward-backward steps on random tokens at batch size 4 (130M), 2 (370M) or 1 (760M, 1.3B) and sequence length 4096 are timed alternately, and the median tokens per second is reported.

**Head-level measurements.** For each attention layer a forward hook captures the rotated queries and keys, the unrotated values and the input of the output projection on eight PG-19 blocks of 4096 tokens. Attention is recomputed exactly in single precision. Queries in the first 64 positions of a block are skipped. Rotary pair $j$ occupies dimensions $(j, j + d/2)$ of a head, the half-split layout of the implementation, which is a fixed permutation of the block-diagonal form of Section 3.1.

Table 15: LongCrawl64 perplexity by evaluation block length and the ratio QKVO-RoPE/QK-RoPE.

| scale | model | 4K | 8K | 16K | 32K |
|---|---|---|---|---|---|
| 130M | QK-RoPE | 19.2 | 57.2 | 351.7 | – |
| | QKVO-RoPE | 20.1 | 50.7 | 225.3 | – |
| | ratio QKVO/QK | 1.05 | 0.89 | 0.64 | – |
| 370M | QK-RoPE | 14.0 | 22.3 | 100.3 | – |
| | QKVO-RoPE | 14.1 | 19.2 | 59.2 | – |
| | ratio QKVO/QK | 1.01 | 0.86 | 0.59 | – |
| 760M | QK-RoPE | 11.3 | 19.0 | 94.2 | – |
| | QKVO-RoPE | 11.4 | 12.9 | 36.7 | – |
| | ratio QKVO/QK | 1.01 | 0.68 | 0.39 | – |
| 1.3B | QK-RoPE | 10.0 | 10.7 | 16.0 | 87.9 |
| | QKVO-RoPE | 10.1 | 10.8 | 20.6 | 77.7 |
| | ratio QKVO/QK | 1.01 | 1.00 | 1.29 | 0.88 |

Table 16: PG-19 perplexity beyond the training length without and with NTK-aware base rescaling (scale = block length/4096), applied to every rotary module of both models.

| scale | model | 8K raw / NTK | 16K raw / NTK | 32K raw / NTK |
|---|---|---|---|---|
| 130M | QK-RoPE | 67.5 / 28.6 | 309.9 / 35.3 | – |
| | QKVO-RoPE | 58.8 / 29.7 | 224.3 / 37.9 | – |
| 370M | QK-RoPE | 30.2 / 20.4 | 112.6 / 23.0 | – |
| | QKVO-RoPE | 25.6 / 20.3 | 68.5 / 23.5 | – |
| 760M | QK-RoPE | 27.3 / 17.1 | 101.2 / 18.9 | – |
| | QKVO-RoPE | 19.9 / 17.2 | 61.1 / 19.0 | – |
| 1.3B | QK-RoPE | 15.8 / 14.6 | 22.1 / 17.7 | 66.9 / 31.9 |
| | QKVO-RoPE | 16.6 / 15.5 | 32.0 / 19.6 | 101.9 / 42.1 |

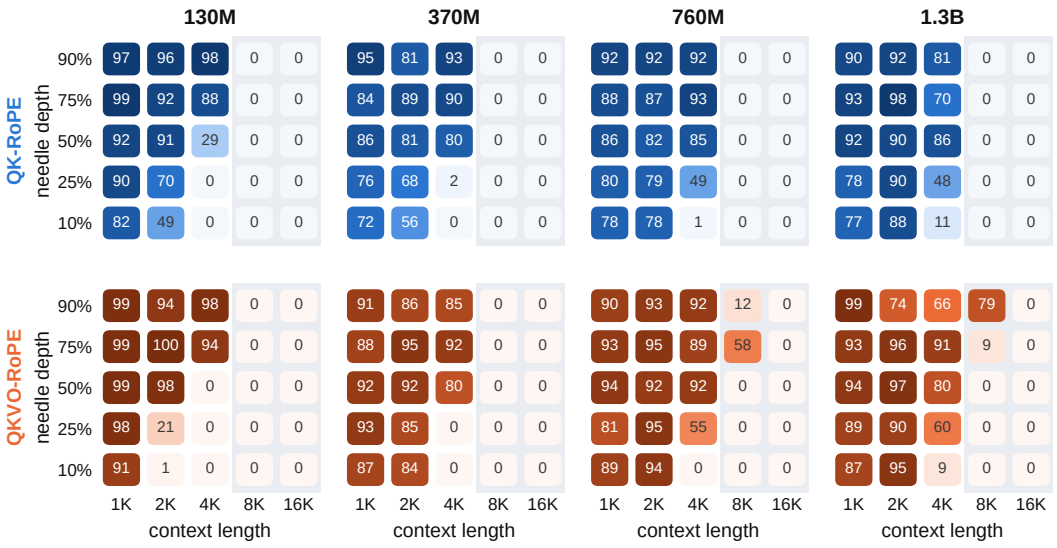



Figure 9: Needle-in-a-haystack accuracy (%) by context length and needle depth, 100 trials per cell. Top: QK-RoPE. Bottom: QKVO-RoPE. Shaded columns lie beyond the training length.

Table 17: Paired tests of QKVO-RoPE against QK-RoPE on identical prompts. Difference in accuracy points with a 95% interval over prompts.

| scale | task | QK-RoPE | QKVO-RoPE | difference | $p$ | $p$ (Bonferroni) |
|---|---|---|---|---|---|---|
| 130M | copying | 75.48 | 72.20 | -3.29 $\pm$ 0.15 | $< 10^{-300}$ | $< 10^{-300}$ |
| 130M | MQAR | 26.01 | 21.34 | -4.67 $\pm$ 0.94 | $4.4 \times 10^{-22}$ | $3.5 \times 10^{-21}$ |
| 370M | copying | 76.77 | 77.23 | +0.46 $\pm$ 0.14 | $1.6 \times 10^{-9}$ | $1.3 \times 10^{-8}$ |
| 370M | MQAR | 27.20 | 20.19 | -7.01 $\pm$ 0.95 | $3.8 \times 10^{-47}$ | $3.0 \times 10^{-46}$ |
| 760M | copying | 78.70 | 79.82 | +1.12 $\pm$ 0.12 | $7.8 \times 10^{-67}$ | $6.2 \times 10^{-66}$ |
| 760M | MQAR | 24.19 | 27.63 | +3.44 $\pm$ 0.94 | $7.3 \times 10^{-13}$ | $5.9 \times 10^{-12}$ |
| 1.3B | copying | 81.48 | 82.93 | +1.46 $\pm$ 0.11 | $2.8 \times 10^{-129}$ | $2.2 \times 10^{-128}$ |
| 1.3B | MQAR | 28.31 | 27.93 | -0.38 $\pm$ 0.90 | 0.42 | 1.00 |

Table 18: PG-19 perplexity (first 4M tokens) beyond the training length with NTK-aware rescaling applied to the score path, the value path or both of the QKVO-RoPE model. Parentheses: ratio to unrescaled QK-RoPE.

| scale | length | QK-RoPE | QKVO-RoPE | score only | value only | both |
|---|---|---|---|---|---|---|
| 130M | 8K | 73.0 | 63.5 (0.87) | 32.9 (0.45) | 63.7 (0.87) | 32.3 (0.44) |
| 130M | 16K | 332.7 | 236.5 (0.71) | 42.3 (0.13) | 183.9 (0.55) | 41.4 (0.12) |
| 370M | 8K | 32.8 | 27.7 (0.84) | 23.0 (0.70) | 28.3 (0.86) | 22.0 (0.67) |
| 370M | 16K | 120.4 | 73.2 (0.61) | 30.7 (0.25) | 67.7 (0.56) | 25.5 (0.21) |
| 760M | 8K | 29.5 | 21.5 (0.73) | 23.0 (0.78) | 35.0 (1.19) | 18.4 (0.62) |
| 760M | 16K | 109.5 | 65.4 (0.60) | 26.8 (0.24) | 139.9 (1.28) | 20.4 (0.19) |
| 1.3B | 8K | 17.0 | 17.8 (1.05) | 17.5 (1.03) | 18.4 (1.08) | 16.6 (0.98) |
| 1.3B | 16K | 23.6 | 34.6 (1.47) | 26.5 (1.12) | 37.0 (1.57) | 21.2 (0.90) |
| 1.3B | 32K | 71.1 | 107.9 (1.52) | 72.2 (1.02) | 100.6 (1.41) | 46.1 (0.65) |

Table 19: Rotary pairs of a 64-dimensional head, base $10^4$.

| $j$ | $\theta_j$ | $\pi/\theta_j$ | $1/\theta_j$ | wraps in 4K | $j$ | $\theta_j$ | $\pi/\theta_j$ | $1/\theta_j$ | wraps in 4K |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.0000 | 3 | 1 | yes | 16 | 0.0100 | 314 | 100 | yes |
| 1 | 0.7499 | 4 | 1 | yes | 17 | 0.0075 | 419 | 133 | yes |
| 2 | 0.5623 | 6 | 2 | yes | 18 | 0.0056 | 559 | 178 | yes |
| 3 | 0.4217 | 7 | 2 | yes | 19 | 0.0042 | 745 | 237 | yes |
| 4 | 0.3162 | 10 | 3 | yes | 20 | 0.0032 | 993 | 316 | yes |
| 5 | 0.2371 | 13 | 4 | yes | 21 | 0.0024 | 1,325 | 422 | yes |
| 6 | 0.1778 | 18 | 6 | yes | 22 | 0.0018 | 1,767 | 562 | yes |
| 7 | 0.1334 | 24 | 7 | yes | 23 | 0.0013 | 2,356 | 750 | no |
| 8 | 0.1000 | 31 | 10 | yes | 24 | 0.0010 | 3,142 | 1,000 | no |
| 9 | 0.0750 | 42 | 13 | yes | 25 | 7.5e-4 | 4,189 | 1,334 | no |
| 10 | 0.0562 | 56 | 18 | yes | 26 | 5.6e-4 | 5,587 | 1,778 | no |
| 11 | 0.0422 | 74 | 24 | yes | 27 | 4.2e-4 | 7,450 | 2,371 | no |
| 12 | 0.0316 | 99 | 32 | yes | 28 | 3.2e-4 | 9,935 | 3,162 | no |
| 13 | 0.0237 | 132 | 42 | yes | 29 | 2.4e-4 | 13,248 | 4,217 | no |
| 14 | 0.0178 | 177 | 56 | yes | 30 | 1.8e-4 | 17,666 | 5,623 | no |
| 15 | 0.0133 | 236 | 75 | yes | 31 | 1.3e-4 | 23,559 | 7,499 | no |