Selective Rotary Position Embedding

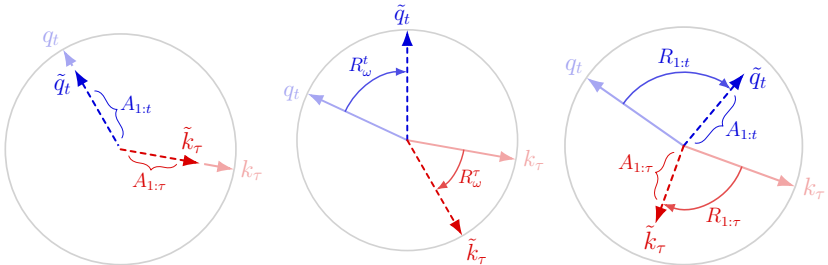
Content-dependent phase for sequence memory

Paper by Sajad Movahedi, Timur Carstensen, Arshia Afzal,
Frank Hutter, Antonio Orvieto, and Volkan Cevher

ICLR 2026 · arXiv:2511.17388v3
Research discussion prepared on 13 October 2026

Learn how far to rotate at each token; retain the backbone's forgetting.

Memory needs control over amplitude and phase



(a) GLA/Mamba2 gating (b) RoPE (c) Selective RoPE

- Forget gates control how much history survives.
- Fixed RoPE advances a clock determined by token index.
- Selective RoPE makes the phase increment depend on content.

Paper, Fig. 1, PDF p. 2. Geometric illustration; general diagonal scaling need not preserve direction.

A fixed state compresses every previous key–value pair

$$S_t = S_{t-1} A_t + v_t k_t^\top, \qquad o_t = S_t q_t$$

$$o_t = \sum_{\tau=1}^{t} v_\tau \left[k_\tau^\top \left(\prod_{\kappa=\tau+1}^{t} A_\kappa \right) q_t \right]$$

- $S_t \in \mathbb{R}^{d_v \times d_k}$; $q_t, k_t \in \mathbb{R}^{d_k}$; $v_t \in \mathbb{R}^{d_v}$.
- A positive diagonal gate rescales channels of the stored state.
- Input-dependent rotations add conditional phase changes and pairwise mixing.

Paper, Eq. (3), PDF p. 3. Product starts at $\tau + 1$, consistent with the recurrence.

The RoPE trick implements complex rotation in real arithmetic

$$R(\omega) = \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix}, \qquad \widetilde{q} = q_1 + iq_2$$

$$\Re\{\overline{\widetilde{k}}\, e^{i\omega}\widetilde{q}\} = k^\top R(\omega) q$$

Move the rotation to query and key

$$k_\tau^\top R^{t-\tau} q_t = (R^\tau k_\tau)^\top (R^t q_t).$$

Two real channels form one complex channel; storage degrees of freedom are unchanged.

Paper, Eqs. (5)–(7), PDF pp. 3–4; App. A.1, pp. 17–18. Consistent conjugation shown.

Learn angle increments and accumulate them along the sequence

$$\omega_t = \mathrm{Conv1D}(W_\omega q)_t, \qquad \psi_{t,j} = \Theta_j \sum_{\kappa=1}^{t} \omega_{\kappa,j}$$

$$C_t = \mathrm{blockdiag}_j\, R(\psi_{t,j}), \qquad q'_t = C_t q_t, \quad k'_t = C_t k_t$$

- W_ω is learned; Θ_j scales the phase of channel pair j.
- The same cumulative rotation acts on the current query and key.
- Feed q'_t, k'_t, v_t to the existing attention backbone.

Paper, Fig. 4 and Sec. 4.1, PDF p. 7. Core path; gates and stabilization add further components.

The RoPE trick also works with cumulative learned phase

Define $R_t = C_{t-1}^\top C_t$ on fixed two-dimensional channel pairs.

$$R_{\tau+1:t} = C_\tau^\top C_t$$

$$k_\tau^\top R_{\tau+1:t} q_t = (C_\tau k_\tau)^\top (C_t q_t)$$

- Constant increments recover ordinary RoPE.
- Learned increments let intervening content change the phase difference.
- Decoding keeps a running phase and short-convolution history.

When a forget gate is present

Preprocessing remains usable. Equivalence to a simple $\Lambda_t R_t$ transition additionally requires compatible, commuting gates.

Paper, Eq. (13), PDF p. 6. Gate condition derived in the backup slides.

Softmax has an RFF representation with rotations

For $\omega \sim \mathcal{N}(0, I)$,

$$e^{q^\top k} = e^{(\|q\|^2+\|k\|^2)/2}\mathbb{E}_\omega\left[\cos\left(\omega^\top(q-k)\right)\right].$$

With equal query norms, each random-feature state has a transition

$$\overline{R}_t = \operatorname{diag}\left(e^{i\Omega^\top(q_t-q_{t-1})}\right).$$

- Finite random features approximate the exponential kernel.
- Softmax also needs its normalization denominator.
- Constant query norm removes an additional real exponential factor.

Paper, Sec. 3.1 and App. A.2, PDF pp. 4, 18–20. Corrected conjugate/phase-difference convention.

Content-dependent phase can implement a parity clock

The RFF increments telescope:

$$\sum_{\kappa=\tau+1}^{t} \Omega^{\top}(q_{\kappa} - q_{\kappa-1}) = \Omega^{\top}(q_t - q_{\tau}).$$

This structural representation alone adds no explicit positional bias.

Teaching example: a content-dependent parity clock

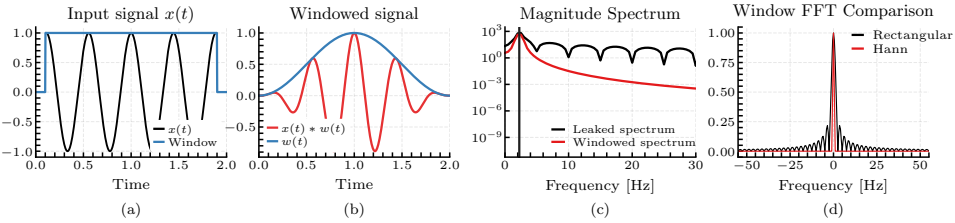
For a binary input b_t, choose $\omega_t = \pi b_t$. Then

$$\cos\psi_t = \cos\left(\pi \sum_{\kappa\le t} b_{\kappa}\right) = (-1)^{\sum_{\kappa\le t} b_{\kappa}}.$$

Input 0 keeps the phase; input 1 flips it.

Path independence: paper, Sec. 3.1, PDF p. 4. Parity construction is a teaching example.

Decay supplies a time window; rotation supplies phase



$$s_n(t) = \int_0^t e^{(-\alpha_n + i\omega_n)(t-u)} k_n v(u)\, du$$

- $\alpha_n > 0$ attenuates old inputs; ω_n creates oscillatory phase.
- Smooth windows suppress spectral sidelobes in the illustrated signal.
- This motivates forgetting; the analysis uses fixed diagonal SSM coefficients.

Paper, Fig. 3 and Sec. 3.2, PDF p. 5; App. A.4, pp. 21–22. Hann and causal exponential windows differ.

Stable optimization requires more than learnable angles

Component	Role reported in the paper
Normalized x_t for angle generation	Replaces query input to reduce instability
Weight normalization or rank 8/16 MLP	Further stabilizes the phase projection
SiLU after convolution	Improves the language-model baseline
Sigmoid phase gate	Controls rotation strength; improves some results
Learnable bias	Helps MAD in combination; limited LM benefit

Norm-preserving rotations do not guarantee easy optimization of their angles.

Paper, Secs. 4.1 and 4.3, PDF pp. 7–9. Fig. 4 omits these full configuration details.

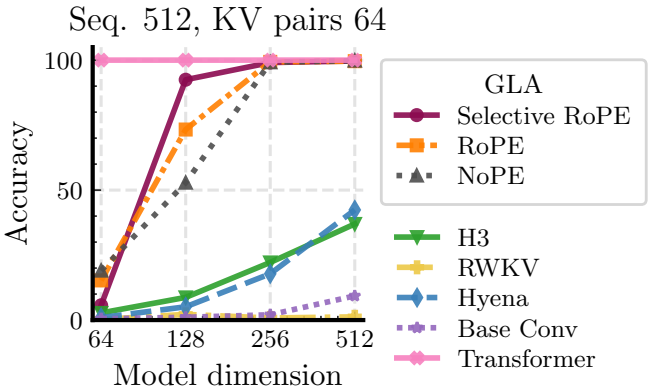
Evaluate recall, state tracking, and language modeling

Experiment	Question	Reported setting
MQAR	Retrieve multiple key–value pairs	Length 512; 64 pairs
MAD	Which memory operations improve?	Six task families
Copying	Generalize exact copying beyond training	Train lengths 2–64
State tracking	Learn input-conditioned state changes	Train length 128
Language modeling	Transfer to pretraining and zero-shot tasks	370M–1.3B models

Within each backbone: NoPE vs. RoPE vs. Selective RoPE.

Paper, Sec. 4, PDF pp. 7–9; App. B, pp. 27–29.

MQAR gains at width 128 do not extend to width 64



Paper, Fig. 6, PDF p. 7. Sequence length 512; 64 key–value pairs.

- At width 128, Selective RoPE leads the GLA variants.
- By width 256, all three approach the ceiling.
- At width 64, Selective RoPE is worse.

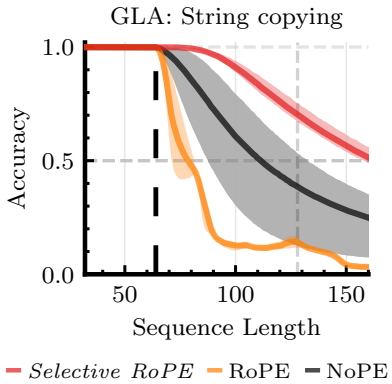
MAD gains depend on the phase gate and bias

GLA variant	IC recall	Memory	Noisy recall	Avg.
NoPE	87.3	38.7	87.6	65.9
RoPE	92.6	**61.4**	91.9	72.5
Selective baseline	94.0	57.1	91.7	72.0
+ phase gate	**96.6**	56.9	94.3	72.3
+ bias	95.0	61.3	91.2	72.7
+ gate and bias	95.9	60.4	**95.0**	**73.2**

- Best configuration: +7.3 percentage points over NoPE; +0.7 over RoPE.
- The selective baseline alone trails RoPE on the average.
- RoPE still leads on Memorize and Selective Copy.

Selected columns from paper, Table 1, PDF p. 7. All scores higher is better.

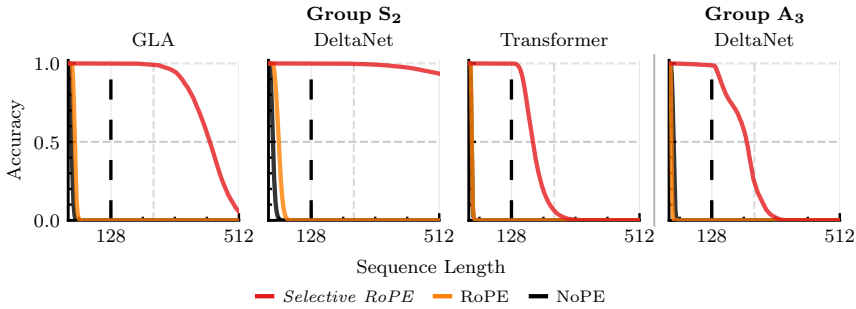
Copying extrapolates better, but accuracy still decays



- Copy the entire input after a copy marker.
- Training lengths: 2–64; dashed line marks the boundary.
- Selective RoPE decays more slowly than both alternatives.
- Near length 160, its accuracy is already roughly 0.5.

Paper, Fig. 5, PDF p. 7; Table 7, p. 29. Configuration lists evaluation to 512; the plotted range is shorter.

State-tracking extrapolation depends on the backbone



- S_2: one-layer models; A_3: two-layer DeltaNet. Train length: 128.
- DeltaNet on S_2 generalizes most strongly toward length 512.
- GLA, Transformer, and A_3 DeltaNet show distinct failure lengths.

Paper, Fig. 7, PDF p. 8; App. B.2, pp. 27–28. Curves use the paper's Accuracy metric.

Language-model comparisons require matching the full setting

- FineWeb; context 4096; Mistral 7B tokenizer with 32,000 tokens.
- GLA/FoX 370M: 10B tokens; GLA 1.3B: 26B tokens.
- Choose learning rates using 4M held-out tokens; evaluate zero-shot.
- Table 3 adds a $\sim$400M Gated DeltaNet sweep with a stable variant.

Read metrics separately

Wiki and LAMBADA perplexity: lower is better.
Task accuracies: higher is better; HellaSwag and ARC-c use normalized accuracy.

Paper, Sec. 4.3 and Table 3, PDF pp. 8–9; Table 5, p. 27. FoX results use output norm OFF.

Small linear models gain accuracy without uniform PPL gains

Backbone	Encoding	Wiki PPL ↓	LMB PPL ↓	Avg. ↑
GLA 370M	NoPE	**25.72**	**15.83**	46.8
	RoPE	26.51	16.32	46.5
	Selective	25.98	16.33	**47.1**
Gated DeltaNet	NoPE	**25.82**	16.27	46.6
∼400M	RoPE	25.92	16.79	46.7
	Selective†	26.13	**15.93**	**47.5**

Against NoPE: +0.3 points for GLA; +0.9 points for Gated DeltaNet.
Wiki perplexity increases in both backbones.

Paper, Table 3, PDF p. 9. All train on 10B tokens. †stable_selective_rope; averages as reported.

FoX and larger GLA expose different metric trade-offs

Backbone	Encoding	Wiki PPL ↓	LMB PPL ↓	Avg. ↑
FoX 370M	NoPE	25.52	17.76	45.3
10B tokens	RoPE	**25.29**	17.71	45.1
	Selective	33.87	**17.29**	**46.1**
GLA 1.3B	NoPE	18.18	8.59	**55.2**
26B tokens	RoPE	18.50	8.88	54.4
	Selective	**17.87**	**8.50**	54.6

FoX: Wiki PPL worsens substantially. GLA 1.3B: PPL improves, average accuracy drops versus NoPE.

Paper, Table 3, PDF p. 9. FoX Avg. values are reported values and do not match the six displayed task means; see backup.

The selective baseline needs additional components to improve

GLA 370M / 10B tokens	Avg-PPL ↓	Avg-Acc ↑
NoPE	**20.78**	46.8
RoPE	21.42	46.5
Selective baseline	21.27	45.8
+ SiLU	21.14	46.7
+ phase gate	21.16	**47.1**

- The phase-gated row improves accuracy by 1.3 points over the baseline.
- NoPE retains the lowest average perplexity.
- Optimization and parameterization are part of the method's benefit.

Paper, Table 2, PDF p. 8. Avg-PPL matches the arithmetic mean of Wiki and LAMBADA PPL.

The RFF temperature theorem needs a correction

Paper's Theorem 1

For unit queries and keys, with $\xi = q^\top k$:

$$\sigma^2 = \frac{1-\xi}{1+\xi}.$$

The proof starts with $q - k$ but later uses terms involving $1 + \xi$.

Reading analysis

The consistent phase-difference kernel has expectation

$$e^{1-\sigma^2(1-\xi)}.$$

As $D \to \infty$, matching e^{ξ} requires $\sigma^2 = 1$ for $\xi < 1$.

The proposed temperature spectrum remains a design heuristic; its stated optimality is not established by this derivation.

Paper, App. A.3, PDF pp. 20–21. Independent derivation and finite-D expression in backup.

Learned phase is the central contribution

1. **Mechanism:** accumulate learned angle increments and rotate queries and keys before attention.
2. **Evidence:** strong changes in recall and state tracking; language-model gains depend on architecture, scale, and metric.
3. **Practical choice:** include normalization, phase gating, and matched NoPE/RoPE baselines in an ablation.

Open empirical questions

Real-language length extrapolation and end-to-end runtime overhead remain unquantified in the reported v3 comparisons.

Paper, Secs. 4–6, PDF pp. 7–10. Interpretation based on the reported results.

References and discussion materials

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher.

Selective Rotary Position Embedding. ICLR 2026. arXiv:2511.17388v3, revised 23 June 2026.

- Paper: arxiv.org/abs/2511.17388v3
- Author code link: github.com/timurcarstensen/selective-rope
- Companion Chinese notes: `reading-notes.zh.md`

Original figures are reproduced from the paper and its source package.
Algebraic corrections and teaching examples are marked as reading analysis.

Backup: the finite-feature error retains a dependence on D

Reading analysis. Let $\xi = q^\top k < 1$, $\|q\| = \|k\| = 1$, and

$$\widehat{K}_D = \frac{e}{D} \sum_{j=1}^{D} \cos\big(\omega_j^\top (q - k)\big), \quad \omega_j \sim \mathcal{N}(0, \sigma^2 I).$$

$$\mathbb{E}[(\widehat{K}_D - e^\xi)^2] = \frac{e^2}{2D} \left(1 - e^{-2\sigma^2(1-\xi)}\right)^2 + \left(e^{1-\sigma^2(1-\xi)} - e^\xi\right)^2.$$

- First term: sampling variance. Second term: squared bias.
- At $\xi = 1/2$, the paper's $\sigma^2 = 1/3$ gives mean 2.301; target 1.649.
- $\sigma = 1$ matches the target mean exactly.

Independent derivation for the consistent real-part RFF estimator; compare App. A.3, PDF pp. 20–21.

Backup: diagonal gates require a change of coordinates

Run the existing gate in the rotated coordinates:

$$\widehat{S}_t = \widehat{S}_{t-1}\Lambda_t + v_t(C_t k_t)^\top, \qquad o_t = \widehat{S}_t C_t q_t.$$

Set $S_t = \widehat{S}_t C_t$. Then

$$S_t = S_{t-1}\underbrace{C_{t-1}^\top \Lambda_t C_t}_{A_t^{\text{eff}}} + v_t k_t^\top.$$

- With a shared scalar gate inside each 2D pair, this reduces to $\Lambda_t R_t$.
- General per-channel gates need not commute with the rotations.
- In either case, $\|A_t^{\text{eff}}\|_2 = \|\Lambda_t\|_2$.

Reading analysis of query–key preprocessing; relates to Sec. 3.3, PDF p. 6.

Backup: FoX averages do not match the displayed scores

FoX 370M / 10B	Reported Avg.	Mean of six shown accuracies
NoPE	45.3	46.18
RoPE	45.1	45.93
Selective RoPE	46.1	46.97

- Six columns: LAMBADA, PIQA, HellaSwag, WinoGrande, ARC-e, ARC-c.
- The text calls Avg. a core multiple-choice macro-average, but the discrepancy is not resolved.
- Ranking is unchanged; cite the aggregation issue with absolute averages.

Paper, Table 3, PDF p. 9. Right column recomputed from displayed values; left column preserved verbatim.

Backup: related mechanisms fit distinct transition structures

Method	Interpretation in this paper
RFF / RFA	Input-dependent unit-modulus phase in a kernel representation
PRF / Performer	Real exponential factor; not necessarily contractive in this representation
CosFormer	A shared frequency $\pi/(2M)$ with sine/cosine state components
ALiBi / FoX	Fixed / input-dependent exponential decay of softmax weights
DeltaNet	Generalized Householder updates mix directions; not always pure rotations

Paper, App. A.5–A.7 and Table 4, PDF pp. 22–27. Interpretations do not equate the complete architectures.

Backup: synthetic protocols use different budgets and seeds

Task	Reported configuration
MAD	66 task settings $\times$ 6 optimizer settings = 396 training runs per model configuration
State tracking	2M train / 0.5M eval sequences; 100 epochs; batch 4096; LR 10^{-3}; five seeds
Copying	Vocabulary 26; LR 5×10^{-5}; seed 42; train lengths 2–64; evaluation configured to 512
MQAR	Eight learning-rate choices; endpoints are described inconsistently in the appendix

Copying confidence bands are shown, but their construction is not specified in the accompanying text.

Paper, App. B.2 and Tables 6–7, PDF pp. 27–29.