Decoupling the "What" and "Where" with Polar Coordinate Positional Embedding

PoPE: a research group discussion

Anand Gopalakrishnan, Róbert Csordás, Jürgen Schmidhuber, Michael C. Mozer

arXiv:2509.10534v3 Published at ICML 2026 (PMLR 306) Official code

Reading date: 28 October 2026

The research question: can attention separate what from where?

A query may need to match a token's content and its relative position independently.

- RoPE gives each two-dimensional content feature an initial phase, then rotates it by position.
- PoPE assigns content to nonnegative magnitudes and position to phases.
- The paper tests this bias with indirect indexing, sequence modeling, and length extrapolation.

Paper, Secs. 1–4, PDF pp. 1–7.

RoPE encodes relative position through paired rotations

For query position t and key position s, a d-dimensional head has $d/2$ two-dimensional components:

$$a_{ts}^{\mathrm{RoPE}} = \sum_{c=1}^{d/2} [R(t\theta_c) q_{tc}]^\top [R(s\theta_c) k_{sc}] = \sum_{c=1}^{d/2} q_{tc}^\top R((s-t)\theta_c) k_{sc}.$$

- $R(\alpha)$ is a planar rotation; θ_c is a channel frequency.
- The position term depends on $s - t$, so shifting both tokens preserves the score pattern.

Paper, Sec. 2, Eq. (1), PDF p. 2.

In polar form, RoPE lets content shift the position peak

$$a_{ts}^{\mathrm{RoPE}} = \sum_{c=1}^{d/2} \mu_{q,tc}\mu_{k,sc} \cos\big((s-t)\theta_c + \phi_{k,sc} - \phi_{q,tc}\big)\,.$$

- $\mu_{q,tc}, \mu_{k,sc}$ are content-dependent magnitudes.
- $\phi_{q,tc}, \phi_{k,sc}$ are also content-dependent phases.
- Their difference changes which offset maximizes a channel's response.

The paper's hypothesis: this coupling can obstruct rules that require separate content and position matches.

Paper, Sec. 2, Eq. (2), PDF p. 2; derivation in App. A.1, p. 12.

Indirect indexing exposes the required operation

Teaching example

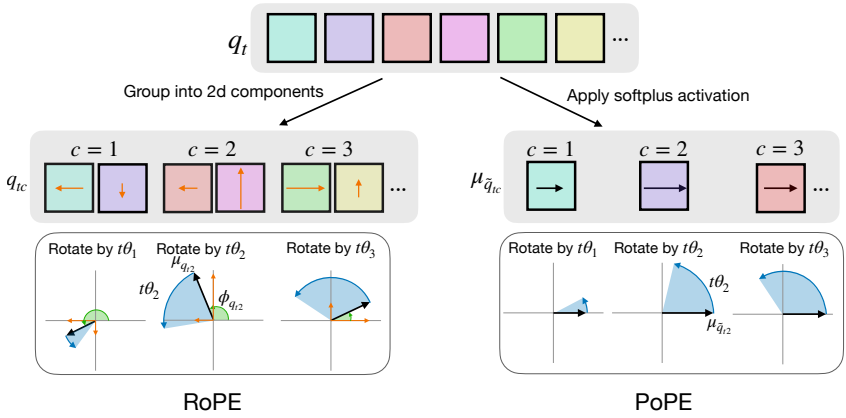
`ABCDE, D, -2` $\rightarrow$ `B`: find the unique anchor `D`, move two positions left, read the target.

1. Match the anchor by content.
2. Shift from that anchor by the requested offset.
3. Predict only the target character.

The actual dataset uses 20–40 distinct letters and offsets from -15 to $+15$.

Paper, Sec. 4, PDF p. 4; data generation in App. B.1, p. 13. The example above is illustrative.

PoPE starts each channel on the real axis



Original paper, Fig. 1, PDF p. 3. Left: RoPE pairs content dimensions. Right: PoPE maps each scalar to a positive magnitude before the position rotation.

PoPE factorizes the attention contribution

$$\tilde{q}_{tc} = \underbrace{\sigma(q_{tc})}_{\text{content magnitude}} e^{it\omega_c}, \qquad \tilde{k}_{sc} = \underbrace{\sigma(k_{sc})}_{\text{content magnitude}} e^{i(s\omega_c+\delta_c)},$$

$$\boxed{a_{ts}^{\mathrm{PoPE}} = \sum_{c=1}^{d} \sigma(q_{tc})\sigma(k_{sc}) \cos((s-t)\omega_c + \delta_c)}$$

- $\sigma(x) = \log(1 + e^x)$ ensures a nonnegative magnitude.
- ω_c is fixed; δ_c is learned per head/channel and shared across positions.
- One complex channel per real feature: d frequencies rather than $d/2$.

Paper, Sec. 3, Eqs. (3)–(6), PDF pp. 2–3.

A fixed phase bias can encode directional offsets

$$g_c(\Delta) = \cos(\Delta\omega_c + \delta_c), \qquad \Delta = s - t.$$

- If $\delta_c = 0$, then $g_c(-\Delta) = g_c(\Delta)$ for an individual channel.
- A learned nonzero δ_c allows different responses to left and right offsets.
- Crucially, δ_c does not vary with the current key/query content.
- The authors constrain $\delta_c \in [-2\pi, 0]$; zero initialization helps length extrapolation.

Paper, Sec. 3, Eq. (6), PDF p. 3. The symmetry statement is a direct deduction from the equation.

Complex attention has a measurable compute cost

$$a_{ts}^{\mathrm{PoPE}} = \sum_{c=1}^{d} \left(x_{q,tc} x_{k,sc} + y_{q,tc} y_{k,sc} \right).$$

- A custom Triton/FlashAttention-style kernel computes the real part without a complex score matrix.
- The appendix estimates extra forward FLOPs as $L/(12D + 2L)$ of one layer's dominant FLOPs.
- At model width $D = 1280$: 5.88% at $L = 1024$; 17.39% at $L = 4096$.
- The current general implementation also increases complex Q/K traffic; in-kernel rotation could reduce it.

Paper, Sec. 3, Eq. (10), PDF p. 4; App. A.2, pp. 12–13. FLOPs are analytical estimates.

Evaluation links each experiment to one claim

Question	Data / setup	Metric / direction
Can it compose content and offset?	Indirect indexing; 3 seeds	Target accuracy ↑
Does it model ordered sequences?	JSB, MAESTRO, human genome	Test NLL ↓
Does it scale on language?	OpenWebText; 124M–774M	Val. perplexity ↓
Does it help downstream?	Six zero-shot tasks	Mean accuracy ↑
Does it extrapolate?	PG-19, up to 10× context	Test perplexity ↓

Paper, Sec. 4, PDF pp. 4–7; App. B, pp. 13–15. Main RoPE/PoPE pairs share model and training settings.

PoPE solves the diagnostic indexing task

Position method	Test accuracy (%, mean $\pm$ SD)
RoPE	11.16 ± 2.45
RoPE, half channels	28.50 ± 6.10
ALiBi, half heads	27.35 ± 10.77
PoPE	$\mathbf{94.82 \pm 2.91}$

PoPE exceeds full RoPE by 83.66 percentage points on this targeted task.

Paper, Table 1, PDF p. 4; additional baselines in App. C, Table 9, p. 15. Three seeds.

Sequence modeling improves in music and genomics

Dataset	RoPE test NLL	PoPE test NLL
JSB Chorales	0.5081	**0.4889**
MAESTRO	1.501	**1.486**
Human Reference Genome	4.217	**4.152**

- Lower is better; these are the reported best test NLL values.
- Music uses up to 2,048 tokens; genome uses up to 1,000 tokens.

Paper, Tables 2–3, PDF pp. 4–5; sequence lengths in App. B.3, p. 15.

OpenWebText gains persist from 124M to 774M

Parameters	RoPE val. perplexity	PoPE val. perplexity
124M	21.55	**21.33**
253M	18.88	**18.55**
774M	15.85	**15.45**

The direction of the perplexity improvement is consistent across the three reported model sizes.

Paper, Table 4, PDF p. 5. OpenWebText pretraining uses 1,024-token contexts.

Ablations: the positive magnitude map matters most

PoPE variant	124M PPL	253M PPL
No softplus	21.57	18.93
ReLU instead of softplus	21.55	18.90
No learned δ	21.42	18.57
Full PoPE	**21.33**	**18.55**

- All values are OpenWebText validation perplexity; lower is better.
- Removing δ costs 0.09 PPL at 124M and 0.02 at 253M.

Paper, Table 5, PDF p. 5.

Downstream mean accuracy improves modestly

Parameters	RoPE mean (%)	PoPE mean (%)	Difference (pp)
124M	45.33	**46.19**	+0.86
253M	48.76	**48.78**	+0.02
774M	51.80	**52.46**	+0.66

- Mean of six zero-shot task accuracies; not every individual task improves.
- At 124M, RoPE remains higher on LAMBADA, HellaSwag, and PIQA.

Paper, Table 6, PDF p. 6. pp = percentage points.

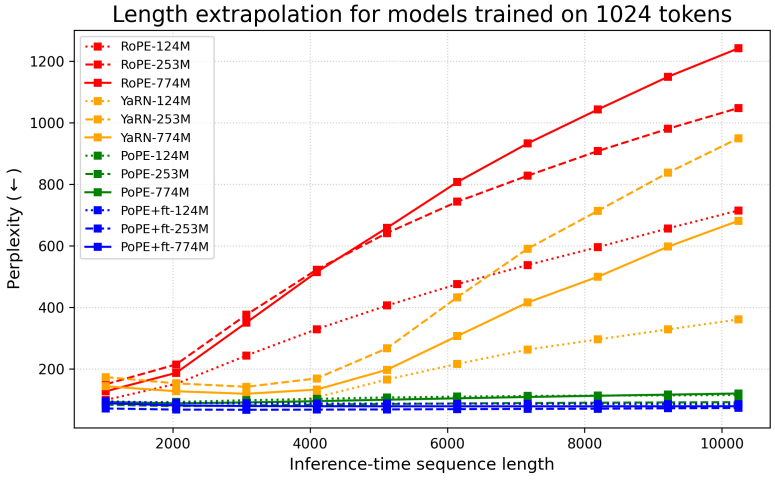
Length extrapolation compares distinct adaptation budgets

- OpenWebText pretraining context: **1,024 tokens**.
- PG-19 test contexts: 1,024 to **10,240 tokens**.
- **RoPE** and **PoPE**: evaluated without long-context fine-tuning.
- **YaRN**: RoPE frequency interpolation plus 500 fine-tuning steps at 4,096 tokens.
- **PoPE+ft**: same 4,096-token, 500-step fine-tuning, without frequency interpolation.

The figure separates the out-of-the-box PoPE result from both fine-tuned alternatives.

Paper, Sec. 4, PDF pp. 5–7, Fig. 2. Fine-tuning uses OpenWebText; evaluation uses PG-19.

PoPE remains flatter out to 10× context length



Original paper, Fig. 2, PDF p. 7. Lower perplexity is better. Green: PoPE zero-shot; blue: PoPE+ft; orange: YaRN; red: RoPE. Line styles denote model size.

Frequency probes support a broader-use interpretation

Model	Observed channel-norm pattern
RoPE	Higher norms in a sparse subset of frequency channels.
PoPE	More distributed use, including higher frequencies except in the first layer.

- Measured across layers for 124M models; the appendix shows a similar 253M pattern.
- PoPE has twice as many frequency rows because each real feature is rotated separately.
- These plots describe usage, while task results establish performance.

Paper, Fig. 3, PDF p. 8; App. D, Figs. 4–5, p. 16.

Takeaways and practical trade-offs

1. **Mechanism:** PoPE fixes position-only phases and puts content in positive magnitudes.
2. **Evidence:** a large indexing gain; lower reported modeling losses; strong PG-19 extrapolation to 10,240 tokens.
3. **Scope:** downstream mean gains can be small (0.02 pp at 253M), with mixed task-level results.
4. **Cost:** complex attention adds arithmetic and memory traffic in the reported implementation.

Paper, Secs. 3–4, PDF pp. 2–8; App. A.2, pp. 12–13.

Paper and code

- Gopalakrishnan, Csordás, Schmidhuber, Mozer. *Decoupling the "What" and "Where" With Polar Coordinate Positional Embedding*. ICML 2026, PMLR 306:36197–36214.

  Official PMLR proceedings record
- Paper version read: arXiv:2509.10534v3 (18 PDF pages).
- Author implementation: agopal42/pope.

Backup: experimental configurations

Dataset	Layers	Context	Training steps
Indirect indexing	8	40	100,000
JSB	6	2,048	3,000
MAESTRO	6	2,048	60,000
Human genome	16	1,000	100,000
OpenWebText	12 / 16 / 36	1,024	100,000

Paper, App. B.2–B.3, Tables 7–8, PDF pp. 14–15. OpenWebText entries correspond to 124M/253M/774M.

Backup: the v3 printed frequency differs from code

- The v3 PDF prints PoPE's frequency after Eq. (4) as $\theta^{(c-1)/d}$.
- The author implementation computes $1/\theta^{(c-1)/d}$ with base $\theta = 10{,}000$.
- The v3 PDF's clamp sentence on p. 3 names θ_c; the implementation clamps δ_c.
- Our mechanism slides denote the fixed frequency by ω_c and discuss the intended bias clamp.

Paper, Sec. 3, PDF pp. 2–3; linked official code.