

Short-Context Gains, Long-Context Drift: Negative Results for Token-Conditioned RoPE

Allan Kazakov

Bahçeşehir University

allan.kazakov@bahcesehir.edu.tr

Abstract

RoPE is a standard positional encoding for modern language models, but its fixed rotations can fail to distinguish positions and tokens in long contexts. Content-aware rotary methods such as CARoPE suggest that adapting rotary geometry to token content can improve language modeling. We report a negative-result study of four token-conditioned RoPE variants for language-model pretraining. DiPE adds token-dependent endpoint phase offsets; Gated DiPE bounds these offsets with learned gates; HSA-RoPE and PPA-RoPE instead modulate accumulated rotary frequency increments at the head and rotary-pair levels. Although these methods produce promising short-context gains, their improvements do not scale reliably. DiPE improves substantially at sequence length 1024 but weakens with larger models and longer contexts, while Gated DiPE fails to recover the lost benefit. HSA-RoPE and PPA-RoPE outperform RoPE at all tested sequence lengths, but their advantage shrinks with longer contexts and they fall substantially behind CARoPE at 8192 tokens.

1 Introduction

Rotary position embeddings (RoPE) encode relative position by rotating query and key vectors with fixed, dimension-specific frequencies (Su et al., 2021). This design is simple and widely used in decoder-only language models (Touvron et al., 2023), but its geometry is independent of token content. Recent work shows that fixed RoPE can fail to distinguish positions and tokens in long contexts (Du et al., 2026), while CARoPE demonstrates that content-aware rotary frequencies can improve language modeling (Veisi et al., 2025). These results motivate a narrower question: when does injecting token content into rotary geometry help, and when does it damage the positional structure that RoPE provides? We study this question

through a negative-result investigation of token-conditioned RoPE variants. Our starting point is Dimension-wise Phase-offset Embeddings (DiPE), which adds token-dependent endpoint phase offsets to standard RoPE. DiPE is initially promising: in small-model training at sequence length 1024, it outperforms both RoPE and CARoPE. However, its gains weaken with larger models and longer contexts. A gated version, which bounds endpoint offsets with learned gates and phase caps, also fails to recover the lost long-context benefit. We then test two accumulated-frequency variants designed to preserve more of RoPE’s interval structure. Head-Scalar Accumulated RoPE (HSA-RoPE) predicts one bounded frequency scale per token and attention head, while Per-Pair Accumulated RoPE (PPA-RoPE) predicts one scale per token, head, and rotary pair. Both are initialized to exactly match RoPE. They outperform RoPE at all tested sequence lengths, but their relative advantage shrinks with context length: they beat CARoPE at 1024 tokens, become weaker on validation at 4096 tokens, and fall substantially behind CARoPE at 8192 tokens. Diagnostics suggest that the degradation is not caused by token conditioning itself, but by the relative phase geometry induced by each parameterization. Endpoint offsets can behave like content-dependent phase noise at long distances. HSA-RoPE learns a persistent slowdown of the rotary clock, producing large accumulated phase drift, while PPA-RoPE can misalign rotary frequency channels despite small average scale shifts. Thus, our results do not imply that content-aware rotary frequencies are inherently flawed: CARoPE remains a strong counterexample. Rather, they show that RoPE initialization and local bounds are not sufficient; successful token-conditioned rotary mechanisms must correct RoPE’s long-context weaknesses while keeping relative phases controlled.

2 Methods

We compare RoPE with four token-conditioned rotary variants under the same Transformer backbone (Vaswani et al., 2017). Let θ_i denote the standard RoPE frequency for rotary pair i , and let x_t be the token embedding at position t . Standard RoPE applies the phase

$$\phi_{t,i} = t\theta_i \quad (1)$$

to each query and key rotary pair. All methods below modify this phase while preserving the same attention, MLP, tokenizer, and training pipeline.

DiPE. Dimension-wise Phase-offset Embeddings add a token-conditioned endpoint offset to RoPE. A linear projection maps each token embedding to angles $\alpha_{t,h,i}$ for each attention head h and rotary pair i :

$$\phi_{t,h,i}^{\text{DiPE}} = t\theta_i + \alpha_{t,h,i}. \quad (2)$$

Thus, the relative phase between positions m and n becomes

$$(m - n)\theta_i + \alpha_{m,h,i} - \alpha_{n,h,i}. \quad (3)$$

This gives attention access to endpoint token content, but the added term is not tied to the interval between the two positions.

Gated DiPE. Gated DiPE uses the same endpoint-offset form, but bounds the learned offset with a phase cap and a learned per-head, per-pair gate:

$$\phi_{t,h,i}^{\text{Gated}} = t\theta_i + g_{h,i} \tilde{\alpha}_{t,h,i}, \quad (4)$$

where $\tilde{\alpha}_{t,h,i}$ is the capped DiPE angle and $g_{h,i}$ is a bounded learned gate. This variant tests whether allowing the model to control the strength of endpoint offsets is sufficient to recover stable long-context behavior.

HSA-RoPE. Head-Scalar Accumulated RoPE removes endpoint offsets and instead predicts one bounded positive frequency scale per token and attention head:

$$s_{t,h} = \exp(c \tanh(W_h x_t)). \quad (5)$$

The local rotary increment is then

$$\Delta\phi_{t,h,i} = \theta_i s_{t,h}, \quad (6)$$

and the final phase is accumulated over positions:

$$\phi_{t,h,i}^{\text{HSA}} = \sum_{u=1}^t \theta_i s_{u,h}. \quad (7)$$

When $s_{t,h} = 1$ for all tokens and heads, HSA-RoPE exactly recovers RoPE. Since the same scale is shared across all rotary pairs within a head, HSA-RoPE preserves the internal RoPE frequency ladder inside each head.

PPA-RoPE. Per-Pair Accumulated RoPE generalizes HSA-RoPE by predicting a separate bounded scale for every token, head, and rotary pair:

$$s_{t,h,i} = \exp(c \tanh(W_{h,i} x_t)). \quad (8)$$

Its phase is

$$\phi_{t,h,i}^{\text{PPA}} = \sum_{u=1}^t \theta_i s_{u,h,i}. \quad (9)$$

PPA-RoPE is more expressive than HSA-RoPE because each rotary frequency can adapt independently. However, this flexibility can also distort the relative alignment between frequency channels.

Initialization. HSA-RoPE and PPA-RoPE are initialized to exactly match RoPE by setting their scale projections to zero, giving $s = 1$ at initialization. DiPE starts close to RoPE through near-zero angle initialization, while Gated DiPE additionally starts with bounded gates. Therefore, performance differences arise from learned departures from the RoPE geometry rather than from different initial positional encodings.

3 Experimental Setup

We evaluate all methods in autoregressive decoder-only language-model pretraining, following the standard next-token prediction setting used by GPT-style models (Radford et al., 2019; Brown et al., 2020). Unless stated otherwise, models are trained on the same packed tokenized corpus and use the same tokenizer, transformer backbone, optimizer settings, training budget, and evaluation protocol. The corpus contains approximately 20B tokens from web, wiki, code, and math sources, with a two-stage mixture schedule: a broad 16B-token stage followed by a 4B-token code/math-upweighted stage. Additional data and training details are given in Appendix A.

Our main experiments use a small decoder-only transformer and vary the positional mechanism while keeping the rest of the architecture fixed. We compare baseline RoPE, CARoPE, HSA-RoPE, and PPA-RoPE at sequence lengths 1024, 4096, and 8192. Because endpoint DiPE and Gated DiPE

are already weaker than the accumulated-frequency variants at 4096 tokens, we reserve the 8192-token runs for HSA-RoPE and PPA-RoPE, since they are the most promising variants. Baseline RoPE and CARoPE use the same architecture and hyperparameters, differing only in the rotary mechanism. For HSA-RoPE and PPA-RoPE, we first run a max-log-scale sweep at sequence length 4096 and select the best validation settings, $c = 0.25$ for HSA-RoPE and $c = 0.1$ for PPA-RoPE. We then keep these values fixed for all sequence-length experiments. The sweep results are reported in Appendix C.

To test whether short-context gains survive model scaling, we additionally run a larger 24-layer model at sequence length 1024, with hidden size 1024, 16 attention heads, and MLP dimension 2816. The vocabulary size is fixed at 50,304. This larger-model study is used only as a scale ablation, not as part of the sequence-length scaling comparison.

We report final training loss, validation loss, and perplexity. For diagnostic analysis, we track learned frequency scales and phase differences relative to RoPE. For PPA-RoPE, we also report wrapped phase differences, which measure angular misalignment modulo 2π . These diagnostics are used to distinguish global rotary-clock drift from per-frequency angular misalignment.

4 Results

Sequence-length scaling. Table 1 shows the main scaling result. These final-checkpoint results reflect stable late-training behavior rather than a noisy endpoint: validation curves show the same ordering near convergence, with no reversal among the main methods (Appendix G). At 1024 tokens, all token-conditioned variants improve over RoPE, with PPA-RoPE giving the best validation loss. However, the relative ordering changes as context length increases. At 4096 tokens, HSA-RoPE and PPA-RoPE still improve over RoPE, but CARoPE becomes best on validation. At 8192 tokens, HSA-RoPE and PPA-RoPE remain slightly better than RoPE, but their gains shrink substantially, while CARoPE improves by a much larger margin. Thus, accumulated-frequency modulation is not automatically long-context robust: it helps over RoPE, but does not preserve its short-context advantage over CARoPE.

Gating endpoint offsets does not fix DiPE. At 4096 tokens, DiPE improves over RoPE, but Gated

Seq. len.	Method	Val loss	Val PPL
1024	RoPE	2.770	15.96
	CARoPE	2.769	15.94
	DiPE	2.739	15.48
	HSA-RoPE	2.734	15.39
	PPA-RoPE	2.724	15.24
4096	RoPE	2.670	14.44
	CARoPE	2.627	13.83
	DiPE	2.644	14.07
	Gated DiPE	2.666	14.38
	HSA-RoPE	2.635	13.95
	PPA-RoPE	2.634	13.93
8192	RoPE	2.638	13.99
	CARoPE	2.594	13.38
	HSA-RoPE	2.627	13.84
	PPA-RoPE	2.628	13.84

Table 1: Validation performance across sequence lengths. HSA-RoPE and PPA-RoPE outperform RoPE at all tested lengths, but their advantage shrinks with context length, while CARoPE becomes strongest at 4096 and 8192 tokens. Full training losses and perplexities are reported in Appendix B with training and validation curves in Appendix G.

DiPE does not: its validation loss is close to RoPE and worse than DiPE, HSA-RoPE, PPA-RoPE, and CARoPE. This suggests that the issue is not simply that endpoint offsets are too strong. Even when the model can bound them with learned gates and phase caps, the endpoint-offset parameterization does not recover robust long-context behavior.

Model scaling weakens the short-context story.

We also test a larger 24-layer model at sequence length 1024. In this setting, DiPE and HSA-RoPE achieve lower training loss than CARoPE, but CARoPE obtains the best validation loss. HSA-RoPE is especially illustrative: it reaches much lower training loss, yet validates slightly worse than CARoPE. This indicates that lower pretraining loss from token-conditioned rotary modulation does not necessarily imply better generalization. We report the full larger-model ablation in Appendix D.

Angular diagnostics. Table 2 summarizes final 8192-token diagnostics for HSA-RoPE and PPA-RoPE. For diagnostics, we compare each learned rotary phase to the corresponding RoPE phase and report the absolute difference. We report both the mean and the 95th percentile (p95), computed over the final evaluation batch, sequence positions, attention heads, and rotary pairs. For PPA-RoPE, we also report the wrapped phase difference, computed as the absolute angular difference modulo 2π .

HSA-RoPE learns a persistent slowdown of the

Method	Scale mean	Scale std.	Phase diff. mean	Phase diff. p95	Wrapped diff. mean	Wrapped diff. p95
HSA-RoPE	0.911	0.154	52.08	293.8	—	—
PPA-RoPE	0.978	0.033	5.75	28.80	0.909	2.789

Table 2: Final angular diagnostics at sequence length 8192. HSA-RoPE did not log wrapped phase metrics, so we report unwrapped phase drift for HSA-RoPE and both unwrapped and wrapped phase differences for PPA-RoPE. Diagnostics for 1024 and 4096 tokens are given in Appendix E.

rotary clock: its mean scale is 0.911, and its accumulated phase drift is large, with mean absolute phase delta 52.08 and p95 293.8. PPA-RoPE has much smaller average scale deviation, but its wrapped phase error is still large: the p95 wrapped phase delta is 2.789 radians. These diagnostics support two different failure modes. HSA-RoPE deforms RoPE mainly through accumulated global drift, whereas PPA-RoPE can misalign individual frequency channels despite smaller average scale changes. The diagnostic results for other sequence lengths are described in Appendix E.

5 Discussion

Our experiments suggest that content-aware rotary modulation is useful but fragile. The strongest short-context results come from mechanisms that allow token content to alter rotary phases or frequencies, but these gains do not reliably transfer to longer contexts or larger models. This matches prior length-extrapolation work showing that positional mechanisms should be evaluated at the target context length rather than only at short or training lengths (Press et al., 2022; Chen et al., 2023; Peng et al., 2023; Ding et al., 2024).

The comparison with CARoPE is important. CARoPE shows that content-aware rotary frequencies can scale well, so our results should not be interpreted as evidence against token-conditioned RoPE in general. Instead, they show that the parameterization of the content-conditioned rotation matters. RoPE initialization ensures that a method starts from a stable geometry, and local bounds limit each token’s immediate effect, but neither prevents harmful cumulative behavior. HSA-RoPE illustrates this through global rotary-clock drift, while PPA-RoPE illustrates it through per-frequency angular misalignment.

These findings also clarify why endpoint phase offsets are difficult to stabilize. DiPE modifies the phase assigned to each endpoint token rather than the frequency path between positions. This can help at short context, where the offset acts as an

additional content signal, but it does not preserve a clean interval structure. Gating and phase caps reduce the magnitude of the perturbation, but they do not change this endpoint-based parameterization.

A practical takeaway is that content-conditioned rotary methods should be evaluated at the target context length, not only at short context. They should also be diagnosed geometrically, since training loss alone can be misleading: in our larger-model ablation, HSA-RoPE obtains the lowest training loss but does not achieve the best validation loss. Future designs may need constraints that directly control accumulated relative phase drift, preserve alignment across rotary frequency channels, or regularize deviations from a well-behaved adaptive frequency path.

6 Conclusion

We presented a negative-result study of token-conditioned RoPE variants for language-model pre-training. DiPE, Gated DiPE, HSA-RoPE, and PPA-RoPE all test plausible ways to make rotary geometry depend on token content, and several improve over RoPE at short or moderate context. However, endpoint offsets do not scale reliably, gating does not fix them, and accumulated-frequency variants lose their advantage over CARoPE at longer context. These failures should not discourage content-aware positional embeddings. Instead, they suggest that improving RoPE remains a promising direction, but successful methods must adapt rotary geometry while controlling the relative phases induced by that adaptation. More broadly, our results support continued work on positional mechanisms that are not only position-dependent, but also token-dependent in a geometrically constrained way.

Code Availability

Code for all models, training scripts, and diagnostics is available at <https://github.com/AllanK24/dipe-hsa-ppa>.

Limitations

Our study is limited to decoder-only language-model pretraining on one corpus and a small number of model scales. We do not run endpoint DiPE or Gated DiPE at 8192 tokens because they are already weaker than the accumulated-frequency variants at 4096 tokens. HSA-RoPE and PPA-RoPE use max-log-scale values selected from a 4096-token sweep; although limited 1024-token checks favored the same settings, we did not retune these bounds separately for every sequence length or model scale. Finally, our diagnostics focus on rotary phase geometry; they explain observed failures but do not fully characterize how these phase changes interact with downstream attention patterns.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. [Extending context window of large language models via positional interpolation](#). *arXiv preprint arXiv:2306.15595*.
- Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. [LongRoPE: Extending LLM context window beyond 2 million tokens](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11091–11104. PMLR.
- Yufeng Du, Phillip Harris, Minyang Tian, Eliu A. Huerta, Srikanth Ronanki, Subendhu Rongali, Aram Galstyan, and Hao Peng. 2026. [RoPE distinguishes neither positions nor tokens in long contexts, provably](#). *arXiv preprint arXiv:2605.15514*.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. [YaRN: Efficient context window extension of large language models](#). *arXiv preprint arXiv:2309.00071*.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2022. [Train short, test long: Attention with linear biases enables input length extrapolation](#). In *International Conference on Learning Representations*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI technical report*.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. [RoFormer: Enhanced transformer with rotary position embedding](#). *arXiv preprint arXiv:2104.09864*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [LLaMA: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Ali Veisi, Delaram Fartoot, and Hamidreza Amirzadeh. 2025. [Context-aware rotary position embedding](#). *arXiv preprint arXiv:2507.23083*.

A Additional Experimental Details

All experiments use decoder-only Transformer language models trained with next-token prediction. Within each comparison, models use the same tokenizer, packed training corpus, optimizer settings, training budget, and evaluation protocol; only the positional mechanism is varied. The vocabulary size is fixed at 50,304. The small-model experiments use a 12-layer Transformer with hidden size 768, 12 attention heads, and MLP dimension 2048. We evaluate this model at sequence lengths 1024, 4096, and 8192. The larger-model ablation uses a 24-layer Transformer with hidden size 1024, 16 attention heads, and MLP dimension 2816, evaluated only at sequence length 1024. The training corpus contains approximately 20B tokens from web, wiki, code, and math sources. Specifically, we use HuggingFaceFW/fineweb_edu_100BT-shuffled, the English subset of HuggingFaceFW/finewiki, HuggingFaceTB/finemath with configuration finemath-4plus, and the Python subset of bigcode/the-stack-dedup. We use a two-stage mixture schedule: a broad 16B-token stage followed by a 4B-token stage with code/math upweighting. Most core experiments are trained with distributed data parallelism on 8 H100 GPUs. We evaluate validation loss on the same held-out

mixture for all methods within each comparison. The number of tokens per optimizer update is fixed at 524,288 across model configurations. As a result, the global batch size varies with sequence length. All runs use learning rate 2×10^{-4} , minimum learning rate 2×10^{-5} , cosine learning-rate decay, 1,000 warmup steps, weight decay 0.1, gradient clipping at 1.0, and random seed 42.

B Full Sequence-Length Results

Table 3 reports the full training and validation results corresponding to Table 1. The main paper reports only validation metrics for compactness.

C Max-Log-Scale Sweeps

HSA-RoPE and PPA-RoPE use a bounded log-scale parameter c that controls the maximum local frequency deformation. We tune this parameter at sequence length 4096 and use the selected values for all HSA-RoPE and PPA-RoPE experiments, including the sequence-length scaling study and the larger-model ablation. Table 4 shows that HSA-RoPE performs best with $c = 0.25$, while PPA-RoPE performs best with the tighter bound $c = 0.1$.

The sweep also shows that PPA-RoPE is more sensitive to the scale bound than HSA-RoPE. Increasing PPA-RoPE from $c = 0.1$ to $c = 0.15$ or $c = 0.25$ substantially worsens validation loss, consistent with the interpretation that per-pair frequency freedom can distort the rotary frequency ladder. We also ran limited checks at sequence length 1024, testing $c = 0.15$ and $c = 0.30$ for HSA-RoPE and $c = 0.05$ and $c = 0.20$ for PPA-RoPE. These checks did not improve over the values selected at 4096, so we use the same settings throughout.

D Larger-Model Scale Ablation

Table 5 reports the 24-layer, sequence-length-1024 ablation. DiPE and HSA-RoPE improve training loss relative to CARoPE, but CARoPE obtains the best validation loss. This supports the main-paper claim that token-conditioned rotary gains in training loss do not necessarily translate to better validation performance.

E Additional Angular Diagnostics

Table 6 reports final angular diagnostics for HSA-RoPE and PPA-RoPE across sequence lengths. For

each method, phase differences are measured relative to the corresponding RoPE phase. Mean and p95 statistics are computed over the final evaluation batch, sequence positions, attention heads, and rotary pairs. HSA-RoPE did not log wrapped phase differences, so only unwrapped phase differences are reported for HSA-RoPE.

The diagnostics show that HSA-RoPE accumulates increasingly large phase drift as sequence length grows: mean absolute phase difference rises from 6.025 at 1024 tokens to 52.08 at 8192 tokens. PPA-RoPE has smaller unwrapped drift, but its wrapped phase difference becomes large by 4096 tokens and remains high at 8192 tokens. This supports the interpretation that HSA-RoPE fails mainly through global rotary-clock drift, while PPA-RoPE fails through per-frequency angular misalignment.

F Parameter Counts

Table 7 reports total parameter counts. All methods keep the Transformer backbone fixed and differ only in the positional mechanism. The additional parameters introduced by the token-conditioned rotary variants are small relative to the full model size. For example, in the small model, HSA-RoPE and CARoPE add only 9.2K parameters over RoPE, while DiPE and PPA-RoPE add about 0.3M parameters.

G Training and Validation Curves

Figure 1 shows training and validation loss curves for the 8192-token setting. The final-checkpoint results in Table 3 are consistent with the late-training trajectories. CARoPE remains consistently below RoPE, HSA-RoPE, and PPA-RoPE, while HSA-RoPE and PPA-RoPE remain close to each other and only modestly better than RoPE near convergence. This supports using final training and validation losses as the main comparison metrics.

Seq. len.	Method	Train loss	Train PPL	Val loss	Val PPL
1024	RoPE	2.685329	14.663021	2.770166	15.961282
1024	CARoPE	2.685213	14.661318	2.768793	15.939384
1024	DiPE	2.651714	14.178326	2.739253	15.475425
1024	HSA-RoPE	2.645509	14.090617	2.733593	15.388072
1024	PPA-RoPE	2.634428	13.935340	2.723719	15.236888
4096	RoPE	3.182270	24.101403	2.669760	14.436501
4096	CARoPE	3.144775	23.214455	2.627078	13.833295
4096	DiPE	3.151549	23.372245	2.643735	14.065646
4096	Gated DiPE	3.184689	24.159778	2.665644	14.377202
4096	HSA-RoPE	3.140824	23.122920	2.635327	13.947866
4096	PPA-RoPE	3.144899	23.217336	2.633718	13.925453
8192	RoPE	2.513150	12.343751	2.638428	13.991198
8192	CARoPE	2.478648	11.925133	2.593681	13.378933
8192	HSA-RoPE	2.501863	12.205205	2.627347	13.837014
8192	PPA-RoPE	2.506641	12.263672	2.627726	13.842258

Table 3: Full sequence-length scaling results. Endpoint DiPE and Gated DiPE were not run at 8192 tokens because they were already weaker than the accumulated-frequency variants at 4096 tokens.

Method	c	Train loss	Train PPL	Val loss	Val PPL
HSA-RoPE	0.10	3.163375	23.650270	2.653924	14.209693
HSA-RoPE	0.15	3.158291	23.530356	2.649613	14.148564
HSA-RoPE	0.25	3.140824	23.122920	2.635327	13.947866
HSA-RoPE	0.30	3.141033	23.127736	2.641458	14.033652
PPA-RoPE	0.10	3.144899	23.217336	2.633718	13.925453
PPA-RoPE	0.15	3.159697	23.563459	2.649472	14.146564
PPA-RoPE	0.25	3.159797	23.565816	2.651016	14.168431

Table 4: Max-log-scale sweep at sequence length 4096. We select $c = 0.25$ for HSA-RoPE and $c = 0.1$ for PPA-RoPE, then keep these values fixed in all reported HSA-RoPE and PPA-RoPE experiments.

Method	Train loss	Train PPL	Val loss	Val PPL
RoPE	2.323361	10.209929	2.442804	11.505257
CARoPE	2.308816	10.062504	2.435992	11.427148
DiPE	2.307163	10.045883	2.439677	11.469333
HSA-RoPE	2.224412	9.248047	2.440316	11.476667

Table 5: Larger-model ablation at sequence length 1024. PPA-RoPE did not finish because TorchInductor compilation errors made the run prohibitively slow.

Seq. len.	Method	Scale mean	Scale std.	Phase mean	Phase p95	Wrapped mean	Wrapped p95
1024	HSA-RoPE	0.9543	0.1769	6.025	34.48	–	–
1024	PPA-RoPE	0.9796	0.0325	0.9018	4.678	0.4288	2.165
4096	HSA-RoPE	0.9294	0.1405	19.87	125.9	–	–
4096	PPA-RoPE	0.9555	0.0636	5.605	32.83	0.9195	2.800
8192	HSA-RoPE	0.9112	0.1538	52.08	293.8	–	–
8192	PPA-RoPE	0.9783	0.0332	5.752	28.80	0.9085	2.789

Table 6: Final angular diagnostics across sequence lengths. Phase columns report absolute phase difference from RoPE. Wrapped columns report absolute angular difference modulo 2π .

Model size	Method	Total params	Extra over RoPE
Small	RoPE	123,587,328	–
	CARoPE	123,596,556	9,228
	DiPE	123,882,240	294,912
	Gated DiPE	123,882,624	295,296
	HSA-RoPE	123,596,556	9,228
	PPA-RoPE	123,882,624	295,296
Large	RoPE	359,842,816	–
	CARoPE	359,859,216	16,400
	DiPE	360,367,104	524,288
	HSA-RoPE	359,859,216	16,400
	PPA-RoPE	360,367,616	524,800

Table 7: Total parameter counts and additional parameters over the RoPE baseline. All methods keep the Transformer backbone fixed.



Figure 1: Training and validation loss curves at sequence length 8192. Training loss is smoothed with a centered rolling window of 31 steps; validation loss is shown for steps after 34K. CARoPE remains consistently below the other methods near convergence, while HSA-RoPE and PPA-RoPE provide only modest improvements over RoPE.