MoPE: Learning One Positional Structure for Short and Long Contexts

Anonymous authors
Paper under double-blind review

Abstract

Positional encoding is central to Transformer language models, yet explicit positional bias is not equally beneficial across attention components. Recent hybrid designs combine rotary position embeddings (RoPE) with no positional encoding (NoPE), but typically prescribe their allocation through fixed rotary fractions or layer schedules. We introduce Mixture of Position Embeddings (MoPE), a simple framework that learns this allocation from data. MoPE jointly optimizes the model backbone and binary gates that select RoPE or NoPE for each layer and rotary frequency. At inference, these gates define a fixed positional structure without input-dependent routing. The resulting gate configurations exhibit distinct layer- and frequency-dependent patterns, revealing heterogeneous positional requirements within Transformers. Experiments on models ranging from 70M to 1B parameters show that MoPE maintains competitive short-context performance while achieving strong long-context retrieval. After 4K pretraining and context extension to 64K, the 410M and 1B models achieve 94.9% and 98.8% NIAH accuracy at 128K, respectively. Combined with YaRN, the 410M model further achieves 99.1% NIAH accuracy at 256K.

1 Introduction

Positional encoding plays an essential role in Transformers (Vaswani et al., 2017), with rotary position embedding (RoPE) (Su et al., 2021) widely adopted for its ability to encode relative positional relationships efficiently. With the growing use of large language models to process long documents, operate over code repositories, and act in multi-turn interactive environments, the ability to integrate information across distant tokens has become increasingly important (Liu et al., 2025; Yao et al., 2025). However, when inference sequences exceed the training context length, the resulting rotary patterns fall outside the training distribution, leading to performance degradation (Liu et al., 2024; Wertheimer et al., 2026). Existing work addresses this issue by rescaling positions, modifying rotary frequencies, or remapping distant positions, including Position Interpolation, NTK-aware scaling, YaRN, and LongRoPE (Chen et al., 2023; bloc97, 2023; Peng et al., 2023; Ding et al., 2024). These methods have substantially improved context-length extension while retaining RoPE as a global architectural prior.

Recent analyses reveal distinct roles for RoPE and attention with no positional encoding, commonly referred to as NoPE (Kazemnejad et al., 2023; Haviv et al., 2022). RoPE makes attention explicitly sensitive to relative displacement, supporting positional patterns and local or recency-biased aggregation, whereas NoPE avoids distance-dependent rotations, allowing semantic similarity to be compared more directly across distant positions (Barbero et al., 2024; Yang et al., 2025). Existing methods exploit these complementary properties in several ways. Partial RoPE applies rotary embeddings to only a fixed fraction of each attention head, while other methods remove selected low-frequency RoPE components and replace them with NoPE (Biderman et al., 2023; Barbero et al., 2024; Chen et al., 2024). Multi-head Latent Attention (MLA)DeepSeek-AI (2024) separates content and RoPE-bearing subspaces, while layer-wise architectures use stronger RoPE for local attention and NoPE or partial RoPE for global attention (DeepSeek-AI, 2024; Yang et al., 2025; Huang et al., 2026). Together, these approaches allocate RoPE and NoPE across different attention subspaces and layers to combine explicit positional modeling with position-independent content retrieval. Nevertheless, existing approaches specify this allocation in advance through manually designed layer schedules, fixed rotary fractions, or preselected low-frequency RoPE components that are replaced

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.

with NoPE. Such fixed designs may not capture how positional requirements vary across layers and frequency components, nor can they reveal the structure that a model would learn from data. This raises a central question: can a Transformer learn where RoPE is needed and thereby discover a positional structure that supports both short- and long-context modeling?

We introduce Mixture of Position Embeddings (MoPE), a trainable routing mechanism that allows the model to discover its positional encoding structure from data. MoPE associates each Transformer layer and rotary frequency with a learnable gate that determines whether the corresponding query and key dimensions use RoPE or remain unrotated (NoPE). By jointly optimizing these gates with the language-model backbone, MoPE replaces a fixed architectural decision with an adaptive layer- and frequency-level routing problem. Once learned, the gates define a fixed positional structure at inference without input-dependent routing. The resulting model can preserve positional components useful for local modeling while weakening components that impede long-context generalization. Jointly learning the positional gates and the model backbone, however, introduces a path-dependent optimization problem: early routing decisions affect the representations learned by the backbone, which can subsequently reinforce those decisions. To maintain exploration during this co-adaptation process, we optimize the gates using Gumbel–Softmax routing (Jang et al., 2017). When the gate probabilities remain uncertain, Gumbel perturbations encourage stochastic exploration of different RoPE–NoPE configurations. As training progresses and the logit margins between RoPE and NoPE increase, the same perturbations become less likely to change the preferred routes, naturally shifting the routing process from exploration toward stable selection. A separate concern is that slowly adapting gates may allow the backbone to become prematurely conditioned on an initial routing configuration. We therefore assign the gate parameters a higher learning rate than the backbone, allowing the routing structure to adapt before the backbone develops a strong dependence on an early configuration.

We evaluate MoPE in two complementary settings: pretraining from scratch with a 4K context window and continued pretraining for context extension to 32K and 64K tokens. Our experiments cover four Pythia model scales up to 1B (Biderman et al., 2023) and compare MoPE with multiple fixed positional encoding baselines. We assess short-context performance on 11 downstream benchmarks and long-context modeling through perplexity and needle-in-a-haystack (NIAH) retrieval (Kamradt, 2023), both within and beyond the training length. When the routing structure learned during 4K pretraining is kept fixed throughout context extension, MoPE preserves short-context performance while achieving 94.9% and 98.8% NIAH accuracy at 128K for the 410M and 1B models, respectively. Combined with YaRN at inference (Peng et al., 2023), the 64K-trained 410M model further supports fourfold context extrapolation to 256K with 99.1% NIAH accuracy. Ablations show that learning an effective routing structure relies on NoPE initialization, stochastic Gumbel exploration, and an appropriate gate learning rate. Moreover, the substantially weaker extrapolation of MoPE-Fixed, which uses the same final routing mask without following the learned routing trajectory, demonstrates that gate–backbone co-adaptation is essential. Finally, the learned gates exhibit distinct layer- and frequency-dependent routing patterns, revealing heterogeneous positional requirements within Transformers. Thus, MoPE serves not only as an effective positional encoding method but also as a tool for positional architecture analysis and discovery.

2 Related Work

Most methods for extending RoPE beyond its training length retain a global rotary structure and modify how it extrapolates. Position Interpolation rescales position indices, NTK-aware scaling and YaRN adjust the frequency spectrum, and LongRoPE searches for non-uniform scaling factors (Chen et al., 2023; bloc97, 2023; Peng et al., 2023; Ding et al., 2024). More recent methods act directly on rotary components: FoPE reformulates them through a Fourier basis and suppresses harmful components, and RoPE-ID applies high-frequency rotations only to selected channels based on a geometric analysis (Hua et al., 2024; Wertheimer et al., 2026). These approaches improve length extrapolation through predefined rescaling or filtering rules.

Another line of work combines RoPE with NoPE to exploit their complementary roles: RoPE provides explicit relative-position information, whereas unrotated representations support content comparison without distance-dependent rotation (Barbero et al., 2024; Yang et al., 2025). Partial RoPE rotates only a fixed fraction of each attention head, and other methods replace or softly clip selected

low-frequency components with NoPE (Biderman et al., 2023; Barbero et al., 2024; Chen et al., 2024; Li et al., 2026). Multi-head Latent Attention separates content and RoPE-bearing query-key subspaces, while layer-wise designs use stronger RoPE for local attention and NoPE or partial RoPE for global attention (DeepSeek-AI, 2024; Yang et al., 2025; Huang et al., 2026). Although these designs demonstrate the benefit of combining position-sensitive and position-independent representations, their channel partitions, frequency selections, or layer schedules are fixed in advance.

Learning-based approaches instead adapt positional transformations from data. DAPE and context-aware RoPE produce input-dependent position signals, while MoICE routes among multiple RoPE-angle experts with a frozen backbone (Zheng et al., 2024; Veisi et al., 2025; Lin et al., 2024). In contrast, MoPE jointly learns the backbone and a binary gate for every layer and rotary frequency, producing a fixed, input-independent structure that can be reused across context lengths.

3 PRELIMINARIES

3.1 FREQUENCY-WISE FORMULATION OF ROPE AND NOPE

RoPE (Su et al., 2021) partitions the query and key of each attention head into two-dimensional subspaces, each associated with a rotary frequency. For clarity, we consider a fixed layer and head and omit their indices. For a head dimension d_h and RoPE base b, the i-th frequency is

$$\theta_i = b^{-2i/d_h}, \qquad i = 0, \ldots, d_h/2 - 1. \tag{1}$$

For $\mathbf{x} \in \{\mathbf{q}, \mathbf{k}\}$, let

$$\mathbf{x}_m^{(i)} = \begin{bmatrix} x_{m,2i} \\ x_{m,2i+1} \end{bmatrix} \tag{2}$$

denote the component associated with frequency θ_i at position m. RoPE transforms this component as

$$\mathbf{x}_m^{\mathrm{R},(i)} = \mathbf{R}(m\theta_i)\mathbf{x}_m^{(i)}, \qquad \mathbf{R}(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \tag{3}$$

In contrast, No Positional Embedding (NoPE) leaves it unchanged:

$$\mathbf{x}_m^{\mathrm{N},(i)} = \mathbf{R}(0)\mathbf{x}_m^{(i)} = \mathbf{x}_m^{(i)}. \tag{4}$$

For a query at position m and a key at position n, RoPE makes the corresponding attention contribution depend explicitly on their relative displacement,

$$\left(\mathbf{q}_m^{\mathrm{R},(i)}\right)^\top \mathbf{k}_n^{\mathrm{R},(i)} = \left(\mathbf{q}_m^{(i)}\right)^\top \mathbf{R}\big((n-m)\theta_i\big)\mathbf{k}_n^{(i)}, \tag{5}$$

while NoPE yields the position-independent contribution $\left(\mathbf{q}_m^{(i)}\right)^\top \mathbf{k}_n^{(i)}$. This unified view makes each frequency subspace a natural unit for selecting between position-sensitive RoPE and position-independent NoPE interactions.

4 METHOD

MoPE formulates positional encoding as discrete routing between RoPE and NoPE. For each layer and rotary frequency, an input-independent binary gate shared across attention heads determines whether the corresponding query-key subspace is rotated. We implement this choice by blending the rotary cosine and sine tensors, which preserves a single-path attention computation. We jointly learn the gates and model backbone using the straight-through Gumbel–Softmax estimator (Jang et al., 2017). At inference, the learned gates define a fixed positional structure.

4.1 MOPE: DISCRETE ROPE–NOPE ROUTING WITH BINARY GATES

MoPE lets each frequency subspace select whether to use RoPE or NoPE. For each layer ℓ and rotary frequency i, MoPE defines one routing unit $u = (\ell, i)$ shared across all attention heads in that layer. Each unit is associated with a binary gate.

$$\mathbf{z}_u = \left[z_u^{\mathrm{R}}, z_u^{\mathrm{N}}\right] \in \{[1,0],[0,1]\}, \qquad z_u^{\mathrm{R}} + z_u^{\mathrm{N}} = 1, \tag{6}$$

where the two entries select the RoPE and NoPE paths, respectively. Denote $g_u = z_u^{\mathrm{R}} \in \{0,1\}$ as the scalar RoPE gate, so that $z_u^{\mathrm{N}} = 1 - g_u$.

Blending RoPE and NoPE. Rather than constructing separate RoPE and NoPE attention branches, we encode the selected transformation directly in the cosine and sine tensors:

$$\begin{aligned} \bar{c}_{m,i} &= (1 - g_u) + g_u \cos(m\theta_i), \\ \bar{s}_{m,i} &= g_u \sin(m\theta_i). \end{aligned} \tag{7}$$

For $\mathbf{x} \in \{\mathbf{q}, \mathbf{k}\}$, the resulting component is

$$\mathbf{x}_m^{\mathrm{M},(i)} = \begin{bmatrix} \bar{c}_{m,i} & -\bar{s}_{m,i} \\ \bar{s}_{m,i} & \bar{c}_{m,i} \end{bmatrix} \mathbf{x}_m^{(i)} = \begin{cases} \mathbf{R}(m\theta_i)\mathbf{x}_m^{(i)}, & g_u = 1, \\ \mathbf{x}_m^{(i)}, & g_u = 0. \end{cases} \tag{8}$$

Thus, the forward pass applies either the complete RoPE rotation or the identity transformation, with no intermediate positional operation. This discrete choice avoids the non-orthogonal, angle-dependent rescaling induced by continuously interpolating between RoPE and NoPE, while yielding an interpretable positional structure. Because routing is applied before attention, MoPE remains compatible with single-path fused attention kernels such as FlashAttention (Dao et al., 2022).

Parameter and computational overhead. For a model with L layers and head dimension d_h, the default layer–frequency routing contains $Ld_h/2$ routing units. Each unit has two logits, giving Ld_h additional trainable parameters in total. The gates do not change the asymptotic complexity of attention. At inference, their fixed binary decisions are folded into the rotary cosine and sine tensors, so no input-dependent router is evaluated.

4.2 Learning Discrete MoPE Gates

The gate in Eq. equation 6 must remain discrete during the forward pass, but a direct one-hot selection is not differentiable. We therefore associate each routing unit u with two trainable logits

$$\boldsymbol{\alpha}_u = [\alpha_u^{\mathrm{R}}, \alpha_u^{\mathrm{N}}], \tag{9}$$

corresponding to the RoPE and NoPE paths, and optimize the gate with the straight-through Gumbel–Softmax estimator.

Straight-through Gumbel–Softmax. We apply the standard Gumbel–Softmax relaxation (Jang et al., 2017) to the routing logits. Let γ_u^c denote independent Gumbel perturbations for the two paths. The relaxed routing probabilities are

$$\widetilde{z}_u^c = \frac{\exp\left((\alpha_u^c + \gamma_u^c)/\tau\right)}{\sum_{c' \in \{\mathrm{R},\mathrm{N}\}} \exp\left((\alpha_u^{c'} + \gamma_u^{c'})/\tau\right)}, \tag{10}$$

where the temperature τ remains fixed during training. We obtain the corresponding one-hot forward decision as

$$\mathbf{z}_u^{\mathrm{hard}} = \mathrm{onehot}\left(\arg\max_c \widetilde{z}_u^c\right). \tag{11}$$

We use the straight-through estimator

$$\mathbf{z}_u = \mathrm{sg}\left(\mathbf{z}_u^{\mathrm{hard}} - \widetilde{\mathbf{z}}_u\right) + \widetilde{\mathbf{z}}_u, \tag{12}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Numerically, $\mathbf{z}_u = \mathbf{z}_u^{\mathrm{hard}}$ in the forward pass, whereas its gradient follows the continuous relaxation $\widetilde{\mathbf{z}}_u$ in the backward pass.

Exploration and gate adaptation. As shown in Eq. equation 10, Gumbel perturbations can change the preferred route when the two routing logits are close. As the logit margin grows relative to these perturbations, they become less likely to alter the routing decision, naturally shifting the process from exploration to exploitation.

The gate and backbone are jointly optimized, making their learning dynamics path-dependent: early routing choices shape the backbone, which can subsequently reinforce those choices. To prevent the gates from adapting too slowly while the backbone commits to an early routing pattern, we set $\eta_{\mathrm{gate}} = \rho\, \eta_{\mathrm{model}}$, with $\rho = 10$ unless otherwise specified.

At inference time, we remove the Gumbel noise and deterministically select $\arg\max_c \alpha_u^c$, yielding a fixed one-hot positional structure with no input-dependent routing. Alternatively, these fixed decisions can be folded into the rotary cosine and sine tensors to eliminate gate computation entirely.

5 EXPERIMENTS

Our experiments investigate whether MoPE provides a robust and general alternative to fixed positional encodings. Specifically, we examine whether MoPE preserves short-context capabilities while improving long-context modeling and length extrapolation, and whether the learned routing exhibits a stable and interpretable structure rather than arising as an optimization artifact. We further conduct ablations to identify the training components required to learn this structure. We also apply MoPE to hybrid-attention models and compare it with layer-wise interleaved positional-encoding baselines in Appendix B.

Experimental Setup. We adopt four model architectures from the Pythia suite, with 70M, 160M, 410M, and 1B parameters (Biderman et al., 2023). Unless otherwise specified, all positional encoding methods use the same model architecture, training data, optimization settings, and token budget to ensure a fair comparison. We organize training into three stages with context lengths of 4K, 32K, and 64K tokens. For the initial 4K pretraining stage, we train all models from scratch on the official 10B-token sample of FineWeb-Edu (Penedo et al., 2024). The 32K and 64K context-extension stages use approximately 1.6B and 1.1B tokens, respectively, drawn from PG-19 (Rae et al., 2020), FineWeb-Edu. We conduct three independent runs for each model and experimental setting and report the mean results. Detailed data compositions and training hyperparameters are provided in Appendix A.

Baselines. We compare MoPE with: (i) Full RoPE, which rotates all rotary dimensions (Su et al., 2021); (ii) Partial RoPE, which rotates a fixed 25% or 50% subset (Biderman et al., 2023); (iii) HoPE, which replaces low-frequency rotary components with identity transformations, and CoPE, which softly clips their rotation angles (Chen et al., 2024; Li et al., 2026); and (iv) FoPE and RoPE-ID, which modify the original RoPE frequency structure to improve length extrapolation (Hua et al., 2024; Wertheimer et al., 2026). RoPE-ID uses temperature scaling by default, for which we additionally report a no-scale variant. All baselines are trained and evaluated under the same experimental protocol; implementation details are provided in Appendix A.2.

Evaluation. We evaluate both short-context capability and long-context generalization. For short-context evaluation, we report validation loss and use the Language Model Evaluation Harness (Gao et al., 2024) to compute word perplexity on WikiText-2 (`wikitext-2-raw-v1`) (Merity et al., 2016) and accuracy on 11 standard downstream benchmarks: LAMBADA (Paperno et al., 2016); PIQA, WinoGrande, and HellaSwag (Bisk et al., 2020; Sakaguchi et al., 2020; Zellers et al., 2019); ARC-Easy, ARC-Challenge, SciQ, BoolQ, OpenBookQA, and CommonsenseQA (Clark et al., 2018; Welbl et al., 2017; Clark et al., 2019; Mihaylov et al., 2018; Talmor et al., 2019); and LogiQA (Liu et al., 2020). We use the unweighted mean accuracy across these tasks as the aggregate short-context score (Short Avg.). For long-context language modeling, we measure perplexity on the PG-19 test set (Rae et al., 2020) and the official Proof-Pile test split (Azerbayev et al., 2022) at multiple sequence lengths both within and beyond the training window. We further evaluate long-range retrieval using needle-in-a-haystack (NIAH) tests (Kamradt, 2023) that vary both context length and needle depth, thereby measuring in-distribution retrieval and length extrapolation. Detailed evaluation-set construction and sample coverage are provided in Appendix A.

5.1 RESULTS OF SHORT-CONTEXT PRETRAINING

We first evaluate models pretrained from scratch with a 4K context window. As shown in Table 1, MoPE achieves the highest Short Avg. at all four model scales, indicating that learning a non-uniform positional structure preserves, and can improve, general short-context capability. Its slightly higher validation loss arises from the Gumbel perturbations used during gate learning: even after the deterministic routing decisions have stabilized, continued routing noise introduces fluctuations into backbone optimization. To isolate this effect, MoPE-Freeze fixes the final hard routing mask once its decisions stop changing and continues training only the backbone under the resulting stationary positional structure. Removing the gate noise consistently reduces validation loss across all model scales.

Table 1: Short-context results for models trained with a 4K context window. MoPE-Freeze fixes the gate parameters once the hard routing decisions stop changing and continues training the backbone. Higher is better for Short Avg.; lower is better for validation loss and WikiText perplexity.

	70M			160M			410M			1B		
Method	Short average ↑	Val. loss ↓	WikiText PPL ↓	Short average ↑	Val. loss ↓	WikiText PPL ↓	Short average ↑	Val. loss ↓	WikiText PPL ↓	Short average ↑	Val. loss ↓	WikiText PPL ↓
Full RoPE	36.90	3.338	55.491	40.77	3.032	37.988	43.05	2.844	30.101	44.06	2.793	28.036
Partial RoPE 50%	37.50	3.327	53.164	40.93	**3.019**	**36.665**	43.60	2.833	**29.017**	44.16	2.794	27.829
Partial RoPE 25%	37.82	3.331	53.516	40.95	3.025	36.708	42.52	2.844	29.495	44.29	2.798	27.938
HoPE	37.56	3.335	54.290	40.60	3.030	37.385	43.66	2.841	29.610	43.86	2.794	27.977
FoPE	37.61	3.352	54.671	40.14	3.036	37.359	43.10	2.843	29.415	44.13	2.801	27.921
RoPE-ID	37.21	3.338	54.085	40.39	3.027	36.952	43.66	2.840	29.644	44.36	2.804	28.341
MoPE	**37.98**	3.328	53.193	**40.97**	3.034	37.233	**44.05**	2.839	29.322	**44.48**	2.802	28.021
MoPE-Freeze	37.96	**3.323**	**52.811**	40.95	3.027	36.980	43.95	**2.832**	29.149	44.20	**2.788**	**27.700**

Table 2: Short-context and NIAH results for 410M and 1B models trained with a 64K context window. OOD Avg. averages NIAH accuracy at 80K, 96K, 112K, and 128K.

Method	410M							1B						
	Average	Val. loss	WikiText PPL	NIAH@4K	NIAH@64K	NIAH@128K	OOD Avg.	Average	Val. loss	WikiText PPL	NIAH@4K	NIAH@64K	NIAH@128K	OOD Avg.
Full RoPE	42.25	2.815	29.841	97.27	54.83	0.6	21.21	43.73	2.770	27.621	**100**	99.7	0	48.86
Partial RoPE 50%	42.61	**2.803**	**28.764**	97.87	93.93	0	36.51	43.68	2.770	27.608	**100**	**100**	0	54.69
Partial RoPE 25%	42.02	2.817	29.410	99.1	95.73	0	23.7	43.96	2.774	27.594	**100**	**100**	0	54.7
CoPE (RoPE init)	42.28	2.815	29.848	98.8	56.33	1.8	22.58	43.79	2.770	27.597	**100**	99.7	0.3	52.87
HoPE	42.9	2.813	29.171	99.7	91.83	53.63	68.33	43.89	2.771	27.499	**100**	**100**	93.03	96.44
FoPE	42.68	2.810	29.226	99.7	88.2	65.43	75.89	43.52	2.775	**27.218**	**100**	99.4	**100**	99.47
RoPE-ID (scaled)	38.42	2.819	37.590	7.57	96.37	85.17	89.54	37.86	2.785	46.327	0	93.97	96.97	96.06
RoPE-ID (no scale)	42.93	2.819	29.485	**100**	95.43	2.13	49.54	44.04	2.783	27.683	**100**	98.5	79.7	87.88
MoPE (Ours)	**43.42**	2.803	29.197	**100**	**100**	**94.87**	**98.49**	**44.15**	**2.765**	27.364	**100**	99.7	98.8	**99.63**

5.2 Results of Long-Context Extension

We next extend the 410M and 1B models to 32K and then 64K through continued pretraining, keeping the routing structures learned during 4K pretraining fixed throughout both stages. This setting directly tests whether a positional structure learned exclusively from short-context pretraining can transfer to substantially longer contexts without relearning the gates. We focus on the final 64K checkpoints and assess whether the transferred structure preserves short-context capability, improves long-context language modeling, and generalizes beyond the training window.

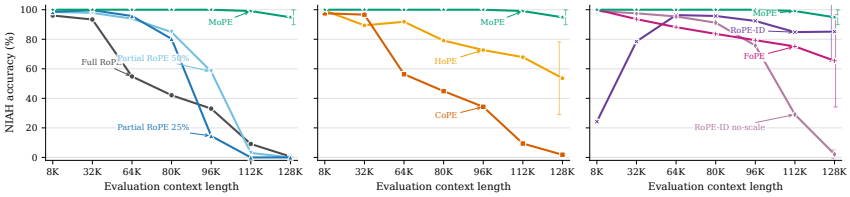


Figure 1: NIAH accuracy across context lengths for 410M models with a 64K context window.

Short-context retention and long-range retrieval. Table 2 shows that MoPE retains the strongest aggregate short-context performance after extension, achieving a Short Avg. of 43.42 for 410M and 44.15 for 1B. It also attains the lowest validation loss for 1B and ties the best result for 410M. The advantage becomes more pronounced beyond the 64K training window. At 128K, MoPE reaches 94.87% and 98.8% NIAH accuracy for 410M and 1B, respectively, while maintaining the highest OOD averages at both scales. The length-wise results in Figure 1 further show that MoPE degrades more gradually than fixed RoPE allocations as the evaluation length increases. RoPE-ID exhibits a clearer trade-off: scaling improves extrapolation but substantially reduces Short Avg., whereas removing scaling restores short-context performance but weakens OOD retrieval; MoPE remains strong on both measures.

Long-context language modeling. Table 3 reports perplexity across evaluation lengths. For the 410M model, MoPE obtains the lowest PG19 perplexity at 128K and the lowest Proof-Pile perplexity at every reported length. For the 1B model, MoPE performs best on PG19 through the 64K training window and on Proof-Pile through 32K, while remaining competitive at longer lengths. FoPE is also strong under 1B length extrapolation, obtaining a PG19 perplexity of 16.072 and 100% NIAH accuracy at 128K, but its Short Avg. is 43.52, compared with 44.15 for MoPE. Thus, whereas FoPE favors extrapolation at the expense of some short-context capability, MoPE provides a stronger overall balance among short-context performance, in-window long-context modeling, and out-of-distribution retrieval.

Table 3: Perplexity on PG19 and Proof-Pile across evaluation lengths for 410M and 1B models trained with a 64K context window.

Method	PG19							Proof-Pile					
	@2K	@4K	@8K	@16K	@32K	@64K	@128K	@2K	@4K	@8K	@16K	@32K	@64K
410M													
Full RoPE	15.274	15.852	16.160	16.048	15.848	15.998	19.921	38.715	33.337	28.628	25.053	23.255	18.558
Partial RoPE 50%	15.237	15.796	16.063	**15.914**	**15.664**	**15.724**	21.867	37.973	32.436	27.632	23.898	21.858	17.119
Partial RoPE 25%	15.495	16.047	16.316	16.157	15.891	15.946	41.457	38.202	32.475	27.478	23.671	21.522	16.932
CoPE clip7	15.277	15.854	16.162	16.050	15.849	15.997	19.440	38.699	33.310	28.607	25.036	23.235	18.550
HoPE	15.289	15.862	16.162	16.040	15.814	15.903	16.542	38.384	32.843	28.082	24.397	22.452	17.738
FoPE	15.335	15.876	16.156	16.008	15.758	15.819	16.600	37.800	32.227	27.371	23.575	21.428	16.689
RoPE-ID (scaled)	24.592	25.817	19.748	17.305	16.146	15.998	16.848	58.469	51.837	34.767	26.505	22.804	17.633
RoPE-ID (no scale)	15.497	16.028	16.307	16.159	15.910	15.996	17.485	39.046	33.523	28.656	24.859	22.805	17.997
MoPE (Ours)	**15.213**	**15.776**	**16.063**	15.920	15.668	**15.724**	**16.367**	**37.463**	**31.998**	**27.172**	**23.376**	**21.261**	**16.677**
1B													
Full RoPE	14.949	15.503	15.789	15.663	15.430	15.510	18.751	35.211	29.919	25.270	21.584	19.631	15.490
Partial RoPE 50%	15.037	15.566	15.829	15.680	15.422	15.472	19.515	34.647	29.421	24.748	21.080	19.110	15.071
Partial RoPE 25%	15.113	15.647	15.904	15.753	15.488	15.536	21.553	35.035	29.653	24.910	21.219	19.128	**15.070**
CoPE clip28	14.946	15.500	15.787	15.662	15.429	15.510	17.965	35.268	29.965	25.312	21.617	19.669	15.525
HoPE	14.931	15.502	15.793	15.669	15.431	15.498	**16.066**	34.988	29.762	25.126	21.511	19.523	15.514
FoPE	15.074	15.623	15.889	15.744	15.479	15.523	16.072	35.165	29.711	24.910	21.152	19.136	15.074
RoPE-ID (scaled)	38.973	43.681	25.518	18.225	16.052	15.732	16.692	68.469	63.041	36.663	25.126	20.546	15.652
RoPE-ID (no scale)	15.214	15.773	16.044	15.899	15.633	15.698	16.389	35.049	29.892	25.330	21.784	19.897	15.680
MoPE (Ours)	**14.927**	**15.473**	**15.749**	**15.607**	**15.346**	**15.390**	16.220	**34.477**	**29.249**	**24.618**	**20.985**	**19.083**	15.177

Table 4: Inference-time extrapolation of 410M models trained at 64K using $4\times$ YaRN.

		PG19 PPL ↓			NIAH ↑					
Method	Short Avg. (%) ↑	64K	128K	256K	4K	16K	32K	64K	128K	256K
Full RoPE	42.72	17.20	17.82	17.70	96.4	97.6	96.1	94.8	80.3	39.1
Partial RoPE 50%	42.44	17.59	18.44	18.35	98.5	97.3	93.9	96.4	96.4	78.5
Partial RoPE 25%	42.01	23.93	30.84	34.21	99.1	98.5	94.8	23.6	2.7	1.8
HoPE	42.63	17.24	17.88	17.77	**100.0**	99.1	98.8	92.1	94.2	75.5
CoPE clip7	42.74	17.20	17.82	17.70	98.5	**100.0**	99.4	98.8	88.8	43.9
FoPE	**42.90**	17.18	17.80	17.63	99.7	97.6	97.3	90.9	84.5	75.2
RoPE-ID	34.90	22.74	21.06	20.57	0.0	0.0	17.9	40.6	61.8	44.5
RoPE-ID no-scale	41.52	21.20	22.24	25.87	77.9	83.3	75.8	52.4	34.2	0.3
MoPE (Ours)	42.77	**16.68**	**17.14**	**16.83**	**100.0**	**100.0**	**100.0**	**99.7**	**99.7**	**99.1**

5.3 Inference-Time Extrapolation with YaRN

We show that MoPE can work effectively with YaRN for inference-time extrapolation. We apply the same $4\times$ YaRN configuration (Peng et al., 2023) to all 64K-trained 410M models without additional training. As shown in Table 4, MoPE remains stable as the inference context grows from 64K to 256K: its NIAH accuracy changes only from 99.7% to 99.1%, while its PG19 perplexity changes from 16.68 to 16.83. At 256K, the best non-MoPE results are 78.5% NIAH accuracy and a PG19 perplexity of 17.63, respectively, compared with 99.1% and 16.83 for MoPE.

5.4 Analysis of Learned MoPE Routing

The routing structures learned by MoPE exhibit common, interpretable patterns across model scales along both network depth and rotary frequency (Figure 2; the 1B results are shown in Appendix 4).

Depth-wise structure. Across model sizes, shallow layers strongly favor RoPE, later-middle layers increasingly select NoPE, and the final layers partially recover RoPE. A plausible interpretation is that shallow layers rely on explicit relative positions to construct local patterns, whereas increasingly contextualized representations in deeper layers support more position-independent, long-range interactions. The recovery near the output suggests that explicit positional information again becomes useful when producing the final token representations.

Frequency-wise structure. MoPE also learns a characteristic non-monotonic structure across rotary dimensions. Under 4K pretraining with the standard RoPE base $b = 10{,}000$, the 70M–410M models, which use $d_{\text{head}} = 64$, exhibit a localized NoPE notch at frequency-pair indices 23–25. This notch begins near the first rotary component whose phase coverage falls below one full cycle

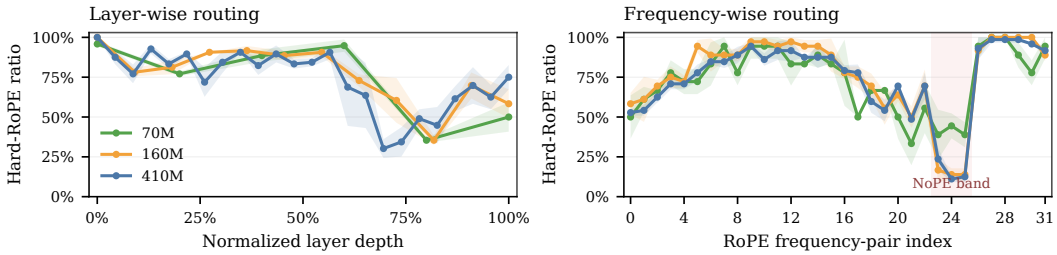


Figure 2: Learned hard-routing structure after 4K pretraining. **Left:** RoPE usage is high in shallow layers, decreases in later layers, and partially recovers near the output. **Right:** frequency-pair indices 23–25 form a localized NoPE notch followed by renewed RoPE selection at lower frequencies. Curves and bands show the mean and standard deviation over three runs.

Table 5: Gate initialization and gate–backbone co-adaptation. The NoPE-initialized model is standard MoPE and the shared baseline.

Ablation	Configuration	Val. loss ↓	Final Hard NoPE (%)	NIAH OOD Avg. ↑
Initialization	RoPE	2.8083	15.41	87.96
	Neutral	**2.7999**	21.35	88.26
	NoPE (MoPE)	2.8034	26.39	**98.49**
Co-adaptation	L0 RoPE@0	2.7970	31.08	80.54
	L0 RoPE@500	2.7968	30.34	86.44
	L0 RoPE@1000	2.8003	31.12	93.20
	MoPE-Fixed	**2.7956**	26.39	72.58

within the training window. RoPE is selected again at the subsequent ultra-low frequencies, showing that the learned structure is not a simple low-frequency cutoff. The 1B model, with $d_{\text{head}} = 256$, exhibits an analogous notch around index 91, where the corresponding one-cycle boundary occurs. Thus, the learned routing tracks the phase variation observable within the training context rather than fixed frequency indices. Accordingly, for the 32K and 64K extension stages, we choose the RoPE base separately for each head dimension to approximately preserve this one-cycle boundary (Appendix A). Additional experiments show that MoPE remains effective under alternative RoPE-base schedules (Appendix F).

5.5 Ablation Studies

We examine how gate initialization, gate–backbone co-adaptation, the gate learning-rate multiplier, and stochastic routing affect the learned structure. Unless otherwise specified, all ablations use the 410M model and report means over three runs.

5.5.1 Effect of Gate Initialization

We compare NoPE-biased, neutral, and RoPE-biased initialization while keeping all other settings fixed. As shown in Table 5, NoPE initialization achieves a NIAH OOD Avg. of 98.49, more than 10 points above the other initializations. We therefore use NoPE initialization by default. However, because initialization also changes the optimization trajectory, this gain cannot be explained by the final NoPE ratio alone.

5.5.2 Gate–Backbone Co-adaptation

We next test whether the gain from NoPE initialization arises from the final routing structure or from its optimization trajectory. In the standard 410M runs, the first layer starts from NoPE but eventually routes entirely to RoPE. We shorten this initial NoPE phase by fixing the first layer to RoPE at steps 0, 500, or 1,000, while standard MoPE leaves it learnable throughout. NIAH OOD Avg. increases from 80.54 to 86.44, 93.20, and 98.49 as fixation is delayed. This trend is not explained by the final NoPE ratio: the early-fix variants retain 30–31% NoPE routes, compared with 26.39% for MoPE, yet extrapolate substantially worse. Moreover, MoPE-Fixed trains a fresh backbone with MoPE's

Table 6: Gate optimization dynamics for 410M models trained at 4K. Hard NoPE is the percentage of hard routes assigned to NoPE; Agreement is the percentage of decisions matching the final hard mask.

		Step 500		Step 1,000		Final
Router (MLR)	Val. loss ↓	Hard NoPE	Agreement	Hard NoPE	Agreement	Hard NoPE
Gumbel (1)	5.2958	100.00	100.00	100.00	100.00	100.00 ± 0.00
Gumbel (5)	2.8454	59.51	63.48	32.23	88.80	33.66 ± 5.62
Gumbel (10)	2.8391	22.66	74.13	17.01	87.41	26.39 ± 0.85
Gumbel (20)	**2.8302**	3.19	80.47	10.16	88.22	21.16 ± 4.51
STE (10)	2.8401	98.11	86.91	93.03	89.00	85.81 ± 0.92

final routing mask fixed from the start but achieves only 72.58 NIAH OOD Avg. Notably, it attains a lower validation loss than MoPE, showing that improved short-context fitting does not necessarily translate into better extrapolation. Thus, the final structure alone does not explain the long-context gain; the gates and backbone co-adapt along the routing trajectory.

5.5.3 Gate Learning-Rate Multiplier

The gate learning-rate multiplier (MLR) controls how quickly the gates adapt relative to the backbone. As shown in Table 6, MLR=1 remains at the all-NoPE initialization and incurs a validation loss of 5.2958, whereas MLR values of 5, 10, and 20 allow the gates to escape this collapsed solution. By step 1,000, these non-collapsed settings agree with more than 87% of their respective final routing masks. However, MLR=20 overshoots toward RoPE, reaching only 3.19% Hard NoPE at step 500 before recovering to 21.16% at the end. Its final NoPE ratio is also less stable across runs than that of MLR=10 ($\pm 4.51\%$ versus $\pm 0.85\%$), suggesting that overly rapid updates can cause premature commitment to a local routing optimum. MLR=5 is also less reproducible, with a final NoPE standard deviation of 5.62%. We therefore use MLR=10, which balances rapid adaptation with a stable final structure.

5.5.4 Stochastic versus Deterministic Routing

We replace the default hard Gumbel-Softmax router with a deterministic straight-through estimator (STE) (Bengio et al., 2013). Despite achieving a similar validation loss, STE remains close to the NoPE-biased initialization, with 85.81% final Hard NoPE compared with 26.39% for Gumbel routing. Deterministic routing therefore stabilizes prematurely, whereas Gumbel noise enables exploration of substantially different routing structures.

6 Conclusion

We introduced MoPE, which replaces a globally prescribed positional encoding with a discrete layer- and frequency-wise routing structure learned jointly with the language-model backbone. Across model scales and context-extension stages, MoPE preserves short-context capability while improving long-context language modeling and retrieval beyond the training length. Its gains remain complementary to inference-time YaRN scaling and extend to a hybrid full-attention and sliding-window architecture. Beyond performance, MoPE consistently discovers structured routing patterns across depth and rotary frequency, including a localized NoPE notch aligned with the phase variation observable within the training context. Together, these results suggest that the allocation of positional information is itself a learnable architectural choice and that a structure learned at short context can remain effective as the context window grows. Our evaluation is currently limited to models up to 1B parameters, and the long-context experiments primarily measure perplexity and synthetic retrieval; the hybrid-attention study is further restricted to 4K training. Establishing whether the same routing principles persist in substantially larger models and translate to complex long-context tasks therefore remains an important direction. More broadly, the reproducible structures identified by MoPE may serve not only as learned positional encodings, but also as empirical guidance for designing simpler positional mechanisms across model architectures and training regimes.

AI USE STATEMENT

In this work, we used generative AI tools for paper polishing, finding related works and research excution (implementing experiments and result analysis). We drafted the experiments part with AI tools. We have not used generative AI tools for generated research ideas, and generating synthetic datasets, proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We have reviewed the draft with 2 authors and the code implementation and results are carefully examined. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Zhangir Azerbayev, Edward Ayers, and Bartosz Piotrowski. Proof-Pile: A pretraining dataset of mathematical text. Hugging Face dataset, test split, 2022. URL https://huggingface.co/datasets/hoskinson-center/proof-pile/tree/main/test.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Velivckovi'c. Round and round we go! what makes rotary positional encodings useful? *ArXiv*, abs/2410.06205, 2024. URL https://api.semanticscholar.org/CorpusID:273229201.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *ArXiv*, abs/2304.01373, 2023. URL https://api.semanticscholar.org/CorpusID:257921893.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

bloc97. NTK-aware scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. Reddit post in r/LocalLLaMA, 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *ArXiv*, abs/2306.15595, 2023. URL https://api.semanticscholar.org/CorpusID:259262376.

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. Hope: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. *ArXiv*, abs/2410.21216, 2024. URL https://api.semanticscholar.org/CorpusID:273653948.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1189. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *ArXiv*, abs/2402.13753, 2024. URL https://api.semanticscholar.org/CorpusID:267770308.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, July 2024. URL https://zenodo.org/records/12608602.

Yoav Gelberg, Koshi Eguchi, Takuya Akiba, and Edoardo Cetin. Extending the context of pretrained llms by dropping their positional embeddings. *ArXiv*, abs/2512.12167, 2025. doi: 10.48550/arXiv.2512.12167. URL https://arxiv.org/abs/2512.12167.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.99. URL https://aclanthology.org/2022.findings-emnlp.99/.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Ning Ding, Youbang Sun, Biqing Qi, Yuchen Fan, Xuekai Zhu, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. *ArXiv*, abs/2412.17739, 2024. URL https://api.semanticscholar.org/CorpusID:274992805.

Ailin Huang, Ang Li, Aobo Kong, Bin Wang, Binxing Jiao, Bo Dong, Bojun Wang, Boyu Chen, Brian Li, Buyun Ma, Chang Su, Changxing Miao, Changyi Wan, Chao Lou, Chen Hu, Chen Xu, Chenfeng Yu, Chengting Feng, Cheng-Yuan Yao, Chunrui Han, Dandan Ma, Da Shi, Da-Xin Jiang, Dehua Ma, Deshan Sun, Di Qi, Enle Liu, Fa-Jun Zhang, Fan-Qi Wan, Guanzhe Huang, Gulin Yan, Guo-Liang Cao, Guopeng Li, Han Cheng, Hang-Yu Guo, Han-Shan Zhang, Hao Nie, Haonan Jia, Haoran Lv, Hebin Zhou, He Lv, Heng Wang, Heung-Yeung Shum, Hongbo Huang, Hong Peng, Hongyu Zhou, Hongyuan Wang, Hou-Tong Chen, Huan Zhu, Huimin Wu, Hui Min Guo, Jia Wang, Jian Zhou, Jianjian Sun, Jiao Jiao Wu, Jiaran Zhang, Jiashu Lv, Jiashuo Liu, Jiaying Fu, Jia-Yu Liu, Jie Cheng, Jie Luo, Jie Yang, Jie Zhou, Jie Hou, Jing Bai, Jing-Cheng Hu, Jin Xie, Jingwei Wu, Jingyang Zhang, Jishi Zhou, Junfeng Liu, Junzhe Lin, Ka Man Lo, Kaiqu Liang, Kaibo Liu, Kaijun Tan, Kaiwen Yan, Kai yang Li, Kang An, Kangheng Lin, Lei Yang, Liang Lv, Liang Zhao, Liangyu Chen, Lieyu Shi, Liguo Tan, Lin Lin, Lina Chen, Luck Ma, Mengqiang Ren, Michael Li, Ming Li, Mingliang Li, Mingming Zhang, Mingrui Chen, Mitt Huang, Na Wang, Peng Liu, Qiu-Ying Han, Qian Zhao, Qinglin He, Qi Guang Du, Qiu-Ping Wu, Quan Sun, Rong Yang, Ruihang Miao, Rui Han, Ruosi Wan, Ruyan Guo, Shanyi Wang, Shaoliang Pang, Shaowen Yang, Shuchuan Fan, Shijie Shang, Shi Qiang Yang, Shiwei Li, Shuangshuang Tian, Siqi Liu, Siye Wu, Siyu Chen, Song Yuan, Tiancheng Cao, Tianchi Yue, Tianhao Cheng, Tianning Li, Ting-Dan Luo, Wang You, Wei Ji, Wei Yuan, Wei Zhang, Wei-Bo Wu, Wei-Bing Xie, Wenzhe Sun, Wen-Yu Deng, Wen Xiao Zheng, Wuxun Xie, Xiangfeng Wang, Xian-Chang Kong, Xiangyu Liu, Xiangyu Zhang, Xiaobo Yang, Xiao-Jun Liu, Xiao-Peng Yuan, Xiao fei Jiao, Xiao-Yong Ren, Xiaoyun Zhang, Xin Li, Xin Liu, Xin Wu, Xing Chen, Xingping Yang, Xinran Wang, Xu Zhao, Xuan He, Xu Feng, Xuedan Cai, Xuqiang Zhou, Yanbo Yu, Yang Li, Yang Xu, Yanlin Lai, Yanming Xu, Yaoyu Wang, Ye-Qing Shen, Yibo Zhu, Yichen Lv, Yicheng Cao, Yifeng Gong, Yi-Jing Yang, Yikun Yang, Yin Zhao, Yin Zhao, Yinmin Zhang, Yitong Zhang, Yixuan Zhang, Yiyang Chen, Yong-Chi Zhao, Yongshen Long, Yongyao Wang, Yadong Guan, Yu Zhou, Yuang Peng, Yu-Lai Ding, Yu-Jin Fan, Yuanzhen Yang, Yu-Lan Luo, Yudi Zhao, Yue Peng, Yueqiang

Lin, Yufan Lu, Yulin Zhao, Yunzhou Ju, Yurong Zhang, Yusheng Li, Yuxiang Yang, Yu-Yang Chen, Yu-Zhuo Cai, Zejia Weng, Zetao Hong, Zexi Li, Zhe Xie, Zheng Ge, Zheng Gong, Zheng-Wen Zeng, Zhenyi Lu, Zhewei Huang, Zhichao Chang, Zhiguo Huang, Zhi-Hang Hu, Zi-Kang Yang, Zili Wang, Zi-Qi Ren, Zixin Zhang, and Zixuan Wang. Step 3.5 flash: Open frontier-level intelligence with 11b active parameters. 2026. URL https://api.semanticscholar.org/CorpusID:285470688.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax, 2017. URL https://arxiv.org/abs/1611.01144.

Gregory Kamradt. Needle in a haystack: Pressure testing LLMs. GitHub repository, 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

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, pp. 24892–24928. Curran Associates, Inc., 2023. doi: 10.52202/075280-1082. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4e85362c02172c0c6567ce593122d31c-Paper-Conference.pdf.

Haoran Li, Sucheng Ren, Alan L. Yuille, and Feng Wang. Cope: Clipped rope as a scalable free lunch for long context llms. *ArXiv*, abs/2602.05258, 2026. URL https://api.semanticscholar.org/CorpusID:285303675.

Hongzhan Lin, Ang Lv, Yuhan Chen, Chen Zhu, Yang Song, Hengshu Zhu, and Rui Yan. Mixture of in-context experts enhance llms' long context awareness. *ArXiv*, abs/2406.19598, 2024. doi: 10.48550/arXiv.2406.19598. URL https://arxiv.org/abs/2406.19598.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, Zhejian Zhou, Ruijie Zhu, Junlan Feng, Yang Gao, Shizhu He, Zhoujun Li, Tianyu Liu, Fanyu Meng, Wenbo Su, Yingshui Tan, Zili Wang, Jian Yang, Wei Ye, Bo Zheng, Wangchunshu Zhou, Wenhao Huang, Sujian Li, and Zhaoxiang Zhang. A comprehensive survey on long context language modeling, 2025. URL https://arxiv.org/abs/2503.17407.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/501. URL https://doi.org/10.24963/ijcai.2020/501. Main track.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of RoPE-based extrapolation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=J07k0SJ5V6.

Yi Lu, Wanxu Zhao, Xin Zhou, Chenxin An, Chenglong Wang, Shuo Li, Yuming Yang, Jun Zhao, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Effective length extrapolation via dimension-wise positional embeddings manipulation. *ArXiv*, abs/2504.18857, 2025. doi: 10.48550/arXiv.2504.18857. URL https://arxiv.org/abs/2504.18857.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Yui Oka, Itsumi Saito, Kyosuke Nishida, and Kuniko Saito. Frequency bands in rope: Base frequency and context length shape the interpolation–extrapolation trade-off. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 94947–94975, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/993dbaec0418a6449090f1debbcb8844-Paper-Conference.pdf.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144/.

Guilherme Penedo, Hynek Kydlicek, 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*, volume 37, 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *ArXiv*, abs/2309.00071, 2023. URL https://api.semanticscholar.org/CorpusID:261493986.

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. URL https://arxiv.org/abs/1911.05507.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, abs/2104.09864, 2021. URL https://api.semanticscholar.org/CorpusID:233307138.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ali Veisi, Delaram Fartoot, and Hamidreza Amirzadeh. Context-aware rotary position embedding. *ArXiv*, abs/2507.23083, 2025. URL https://api.semanticscholar.org/CorpusID:280400946.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Davis Wertheimer, Ao-Zhong Zhang, Derrick Liu, Peng-Hang Yin, and Naigang Wang. Frayed rope and long inputs: A geometric perspective. *ArXiv*, abs/2603.18017, 2026. URL https://api.semanticscholar.org/CorpusID:286669770.

Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr F. Locatelli. Rope to nope and back again: A new hybrid attention strategy. *ArXiv*, abs/2501.18795, 2025. URL https://api.semanticscholar.org/CorpusID:276079501.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, pp. 9965–10017, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b126cc38b8638e07bef37e7b2bb72bf-Paper-Conference.pdf.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael K. Ng, Xin Jiang, Zhenguo Li, and Yu Li. Dape: Data-adaptive positional encoding for length extrapolation. *Advances in Neural Information Processing Systems 37*, 2024. URL https://api.semanticscholar.org/CorpusID:269982134.

Table 7: Training configurations across model sizes and context lengths. GBS denotes the global batch size in sequences, and the data size is rounded to the nearest 0.1B tokens.

Model	Context	Stage	Learning Rate	GBS	Tokens/Step	RoPE θ	Data Size
70M	4K	Base training	1.0×10^{-3}	128	0.52M	1×10^4	10.0B
160M	4K	Base training	7.0×10^{-4}	128	0.52M	1×10^4	10.0B
410M	4K	Base training	5.0×10^{-4}	128	0.52M	1×10^4	10.0B
1B	4K	Base training	3.0×10^{-4}	128	0.52M	1×10^4	10.0B
410M	32K	4K $\rightarrow$ 32K extension	2.5×10^{-4}	16	0.52M	2×10^5	1.6B
1B	32K	4K $\rightarrow$ 32K extension	1.5×10^{-4}	16	0.52M	1.8×10^5	1.6B
410M	64K	32K $\rightarrow$ 64K extension	1.0×10^{-4}	8	0.52M	6×10^5	1.1B
1B	64K	32K $\rightarrow$ 64K extension	1.0×10^{-4}	16	1.05M	5×10^5	1.1B

A Experiment setup details

A.1 Data Composition

We use two subsets of FineWeb-Edu (Penedo et al., 2024): `sample-10BT` and `sample-100BT`, which contain approximately 10B and 100B GPT-2 tokens, respectively.

4K data. The 4K corpus consists of approximately 10B tokens from the FineWeb-Edu `sample-10BT` subset.

32K data. The 32K corpus contains approximately 1.6B tokens. It comprises 1.0B tokens from PG19 documents of 32K–128K tokens, 0.2B replay tokens from FineWeb-Edu `sample-10BT`, and 0.3B tokens from 8K–32K documents selected from FineWeb-Edu `sample-100BT`. We additionally include approximately 0.1B tokens of NIAH data constructed using `sample-10BT` documents as background text.

64K data. The 64K corpus contains approximately 1.1B tokens. It comprises 0.6B tokens from PG19 documents of 64K–128K tokens, 0.2B replay tokens from FineWeb-Edu `sample-10BT`, and 0.2B tokens from 8K–32K documents selected from FineWeb-Edu `sample-100BT`. The two FineWeb-Edu subsets are disjoint from their counterparts in the 32K corpus. We further include approximately 0.1B tokens of NIAH data constructed using `sample-10BT` documents as background text.

Validation data. For each context setting, we hold out a validation split with an approximately constant token budget of 8.4M tokens. Specifically, the 4K setting uses 2,048 validation sequences, the 32K setting uses 256 sequences, and the 64K setting uses 128 sequences. Each validation sequence has the same length as the corresponding training context, yielding 8,388,608 validation tokens at every stage. These held-out sequences are excluded from optimization. The corpus sizes reported above refer to the complete tokenized datasets and therefore include the held-out validation data.

A.2 Baseline Implementations

For the HoPE baseline (Chen et al., 2024), we determine the cutoff using the critical angular frequency $2\pi/L$, where L is the training context length. This yields a cutoff of 9 for the 70M–410M models with head dimension 64 and 37 for the 1B model with head dimension 256. CoPE (Li et al., 2026) is applied only during the 32K and 64K context-extension stages. Following the original method, we set its clipping range to 75% of the low-frequency dimensions identified by its frequency criterion, resulting in clip sizes of 7 and 28 for models with head dimensions 64 and 256, respectively. RoPE-ID (Wertheimer et al., 2026) uses temperature scaling by default; we additionally evaluate a no-scale variant to isolate the effect of this scaling on short-context performance. Unless otherwise stated, all baselines use the same model architecture, training data, optimization settings, and token budget as MoPE.

Table 8: Number of eligible test documents at each perplexity evaluation length. The initial pools contain 95 PG-19 books and 100 Proof-Pile documents, each with at least 8K tokens. Documents shorter than the target evaluation length are excluded.

Evaluation Length	PG-19	Proof-Pile
2K	95	100
4K	95	100
8K	95	100
16K	93	73
32K	78	28
64K	54	4
128K	20	0
256K	6	0

A.3 Evaluation

Short-context evaluation. We compute WikiText word perplexity and all 11 downstream benchmark accuracies using the corresponding task implementations in the Language Model Evaluation Harness (Gao et al., 2024). The harness's `wikitext` task evaluates the `wikitext-2-raw-v1` test split and reports word perplexity (Merity et al., 2016). We report the unweighted mean accuracy over the 11 downstream tasks as Short Avg.

Long-context perplexity data. We evaluate long-context language modeling on PG-19 Rae et al. (2020) and Proof-Pile Azerbayev et al. (2022). PG-19 is a book-level language modeling benchmark consisting of full-length Project Gutenberg books published before 1919. Its test split contains 100 books. We tokenize each book in full without padding or truncation and discard books containing fewer than 8,192 tokens. Five books do not satisfy this minimum-length requirement, leaving a fixed evaluation pool of 95 PG-19 books.

Proof-Pile is a mathematical-text corpus containing both informal and formal mathematics. We use its official test split, tokenize each document without padding or truncation, and retain documents containing at least 8,192 tokens. We then select up to 100 eligible Proof-Pile documents for evaluation.

For each evaluation length L, we include only documents containing at least L tokens and truncate each eligible document to its first L tokens. Since both evaluation pools are initially filtered using an 8K-token threshold, all retained documents are available for the 2K, 4K, and 8K evaluations. At longer lengths, the sample count decreases according to the number of documents that satisfy the target-length requirement. Table 8 reports the exact sample coverage.

NIAH evaluation set. We construct a single-needle Needle-in-a-Haystack (NIAH) evaluation set (Kamradt, 2023) using English passages from the Paul Graham Essays as background text. Each example contains a synthetic needle associating a city with a randomly generated seven-digit "magic number." The needle is inserted at one of 11 relative depths: 0%, 10%, 21%, 31%, 42%, 52%, 63%, 73%, 84%, 94%, or 100%. The model is then asked to retrieve the corresponding number using a completion-style prompt. We generate 10 examples for each context-length–depth combination, resulting in 440 examples in total. The Pythia tokenizer (Biderman et al., 2023) is used to control the target context length.

B Extension to Hybrid Attention Models

We further examine whether MoPE extends beyond full-attention models to hybrid architectures. We use 410M models in which full-attention and sliding-window attention (SWA) layers (Beltagy et al., 2020) are interleaved at a 1:3 ratio, with a window size of 512 tokens. All models in this study are trained with a 4K context. For fixed baselines, we use X/Y to denote the positional encoding in full-attention/SWA layers; P25 and P50 denote Partial RoPE applied to 25% and 50% of the head dimensions. We denote our learned-routing method by *MoPE*, and its retrained fixed-structure counterpart by *MoPE-Fixed*.

Table 9: Comparison of 410M hybrid-attention models trained at 4K, with a 3:1 ratio of full-attention to SWA layers. NIAH-in and NIAH-out average accuracy over 1K–4K and 5K–8K, respectively. Each PPL cell lists 2K/4K/8K/16K results. Values are three-run means without standard deviation. Best and second-best results at each evaluation length are bolded and underlined.

Method	Short avg. ↑	Wiki PPL ↓	NIAH-in ↑	NIAH-out ↑	PG19 PPL ↓ (2K/4K/8K/16K)	Proof-Pile PPL ↓ (2K/4K/8K/16K)
RoPE/RoPE	43.74	30.18	62.28	24.84	33.86/31.03/30.86/36.07	59.37/56.09/58.82/83.42
NoPE/RoPE	43.60	29.86	80.82	<u>56.60</u>	33.82/30.80/**29.17/28.30**	58.17/52.68/49.14/**47.37**
P25/P25	<u>43.89</u>	29.69	<u>91.67</u>	43.86	33.83/30.87/29.86/30.59	56.63/51.20/49.80/51.69
P50/P50	43.78	29.50	89.40	51.14	34.08/31.15/30.72/31.77	<u>56.33</u>/**50.88**/49.28/52.57
P25/RoPE	43.82	29.48	80.15	32.27	33.61/30.75/29.77/30.88	57.01/51.62/50.32/51.62
P50/RoPE	43.59	<u>29.46</u>	85.60	40.60	<u>33.38</u>/<u>30.30</u>/29.78/30.87	57.34/52.11/50.60/53.14
MoPE	43.69	29.66	83.47	47.11	33.67/30.65/<u>29.26</u>/<u>29.54</u>	57.19/51.87/<u>49.10</u>/<u>50.32</u>
MoPE-Fixed	**44.04**	**29.21**	**93.64**	**57.43**	**33.15/30.16**/29.31/30.52	**56.14**/<u>50.89</u>/**48.98**/51.96

Transferability of the learned structure. To test whether MoPE discovers a useful architecture rather than merely a transient optimization state, we extract its final discrete layer–frequency routing structure and retrain a fresh model from scratch with this structure fixed. MoPE-Fixed achieves the best overall result in this 4K-trained setting, including the highest 11-task Short Avg. (44.04), the lowest WikiText perplexity (29.21), and the highest mean NIAH scores within and beyond the training length (93.64 and 57.43) (Table 9). Thus, fixing the learned structure and retraining from scratch preserves strong performance, indicating that MoPE identifies an effective structure rather than only benefiting from its joint optimization trajectory.

Extrapolation and scope. MoPE remains competitive under length extrapolation, ranking second at 16K on both PG19 (29.54) and Proof-Pile (50.32), behind NoPE/RoPE (Table 9). In contrast, RoPE/RoPE degrades to a Proof-Pile perplexity of 83.42 at 16K. This study is limited to models trained with a 4K context, so the present evidence should be viewed as an initial validation of MoPE for hybrid architectures. We leave training and evaluation at substantially longer context lengths to future work.

Routing in the hybrid attention model. We separately analyze the learned gates of full-attention and sliding-window attention (SWA) layers in the hybrid 410M model as shown in Figure 3. Averaged over three runs, full-attention layers retain RoPE for 65.3% of their frequency dimensions, compared with only 57.1% in SWA layers. This difference is particularly pronounced at the low-frequency indices 23–31, where the RoPE ratio decreases from 43.2% in full-attention layers to 14.2% in SWA layers. This behavior is consistent with their different receptive fields: within the 512-token local window of SWA, long-wavelength rotary components undergo limited phase variation and may provide less useful relative-position information. In contrast, full-attention layers can observe their variation over the complete 4K context and therefore preserve more low-frequency RoPE components. These results suggest that MoPE adapts its routing not only to depth and frequency, but also to the effective receptive field of each attention layer.

C Additional Analysis of Learned MoPE Routing

C.1 Cross-Scale Routing Statistics

We quantify the phase coverage of the i-th rotary component within a context of length L as

$$C_i(L) = \frac{L\theta_i}{2\pi}, \tag{13}$$

where θ_i is its angular frequency. For the 70M, 160M, and 410M models, which share $d_{\text{head}} = 64$, approximately three quarters of all routing decisions retain RoPE. The first layer selects RoPE for nearly every frequency pair, whereas later layers contain substantially more NoPE routes. Under 4K pretraining, index 23 is the first component with less than one complete cycle of phase coverage ($C_{22}(4096) > 1$ and $C_{23}(4096) < 1$), aligning with the start of the learned NoPE notch at indices 23–25.

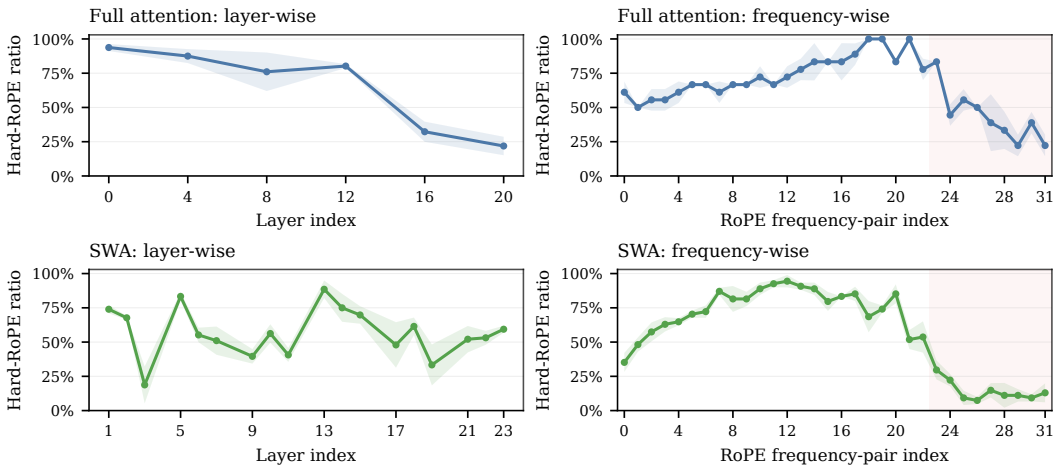


Figure 3: **Learned MoPE routing in the hybrid attention model.** Depth-wise and frequency-wise hard-RoPE ratios are reported separately for full-attention and SWA layers. SWA layers select substantially more NoPE components at low frequencies, while full-attention layers retain more low-frequency RoPE. Curves and shaded regions denote the mean and standard deviation over three runs.

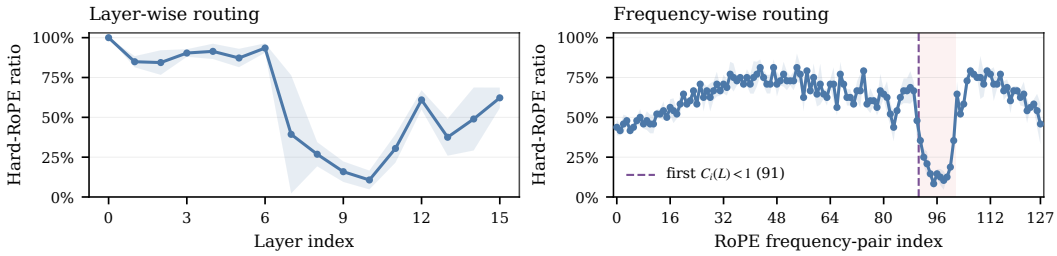
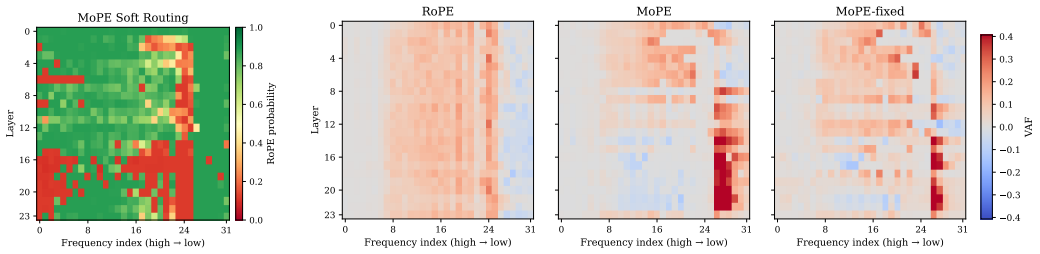


Figure 4: Learned MoPE routing in the 1B full-attention model at a 4K context length. The left panel shows the hard-RoPE selection ratio across layers, while the right panel shows the ratio across RoPE frequency pairs. Curves and shaded regions denote the mean and standard deviation over three runs. RoPE dominates the early layers but is substantially reduced in the middle layers before partially recovering. Frequency-wise routing exhibits a pronounced NoPE band at indices 91–101, aligned with the first sub-cycle component satisfying $C_i(L) < 1$ (dashed line). These consistent, structured patterns indicate that MoPE learns scale-dependent routing rather than collapsing to a uniform positional-encoding choice.

The 1B model provides a cross-head-dimension comparison. With $d_{\text{head}} = 256$, it retains RoPE for 60.3% of all routing decisions. Its first seven layers predominantly select RoPE, its middle layers strongly favor NoPE, and its final layers partially recover RoPE. The first component with $C_i(4096) < 1$ shifts to index 91: indices 91–101 retain only 18.9% RoPE, whereas indices 104–119 recover to 71.6%. These statistics support the same localized, non-monotonic NoPE notch observed in the smaller models.

C.2 How MoPE Reshapes Q–K Interactions

We measure how strongly each frequency aligns with the complete 128K Q–K attention-logit profile using variance accounted for (VAF), defined as $\text{VAF}_f = 1 - \|a - a_f\|_2^2 / \|a\|_2^2$, where a_f and a denote the frequency-specific and complete logit profiles. Full RoPE concentrates positive VAF at indices 24–25. MoPE routes most components in indices 23–25 to NoPE, reducing their mean VAF from 0.134 to 0.019, while shifting the dominant alignment to the lower-frequency RoPE components at indices 26–31, whose mean VAF increases from -0.015 to 0.150. This broad low-frequency band



(a) (b)

Figure 5: **Learned MoPE structure and VAF analysis.** (a) Learned RoPE probability across layers and frequency pairs, where red and green denote NoPE- and RoPE-dominant routing, respectively. (b) Layer-wise single-frequency VAF at 128K for Full RoPE, MoPE, and MoPE-fixed. MoPE shifts the dominant alignment from indices 24–25 to a broader low-frequency band at indices 26–31, whereas MoPE-fixed remains concentrated near index 26.

varies more slowly with relative distance, providing a mechanism for MoPE's improved extrapolation beyond the 64K training length.

MoPE-fixed isolates whether this behavior follows from the final mask alone. Although it uses exactly the same routing pattern, its VAF over indices 26–31 reaches only 0.074 and is concentrated mainly at index 26. In contrast, MoPE distributes high VAF across indices 26–29 in the middle and later layers. The broader organization therefore arises from co-adaptation between the learned gate and the LLM backbone, rather than from the discrete routing mask alone.

D MoPE GATES

Why discrete routing? A continuous interpolation between RoPE and NoPE does not generally preserve the geometry of either transformation. For a gate $g \in [0, 1]$ and rotation angle ϕ, the interpolated operator is

$$\mathbf{A}(g, \phi) = (1 - g)\mathbf{I} + g\mathbf{R}(\phi). \tag{14}$$

Its Gram matrix satisfies

$$\mathbf{A}(g, \phi)^\top \mathbf{A}(g, \phi) = [1 - 2g(1 - g)(1 - \cos \phi)]\, \mathbf{I}. \tag{15}$$

For $0 < g < 1$, this operator is generally non-orthogonal and introduces an angle-dependent rescaling in addition to positional rotation. By using a binary gate in the forward pass, MoPE preserves either the exact RoPE rotation or the identity transformation of NoPE, while also producing an interpretable discrete positional structure.

E GATE DYNAMICS DURING TRAINING.

The learned routing structure emerges early and becomes stable well before convergence. Figure 6 shows a representative 410M training run. After approximately the first tenth of training, more than 90% of the hard-routing decisions already agree with the final mask. Subsequent training primarily refines a small number of remaining decisions, and no hard-mask changes are observed during the final tenth of training. This behavior indicates that MoPE rapidly establishes its principal routing structure and does not rely on unstable gate oscillations near convergence.

F EFFECT OF THE ROPE BASE FREQUENCY.

We evaluate MoPE under two RoPE base-frequency schedules on 410M models. The large-θ setting uses $\theta = 10^6$ for 32K training, a common choice for long-context models, and $\theta = 5 \times 10^6$ for its 64K continuation. We additionally consider a frequency-aligned schedule with $\theta = 2 \times 10^5$ at 32K

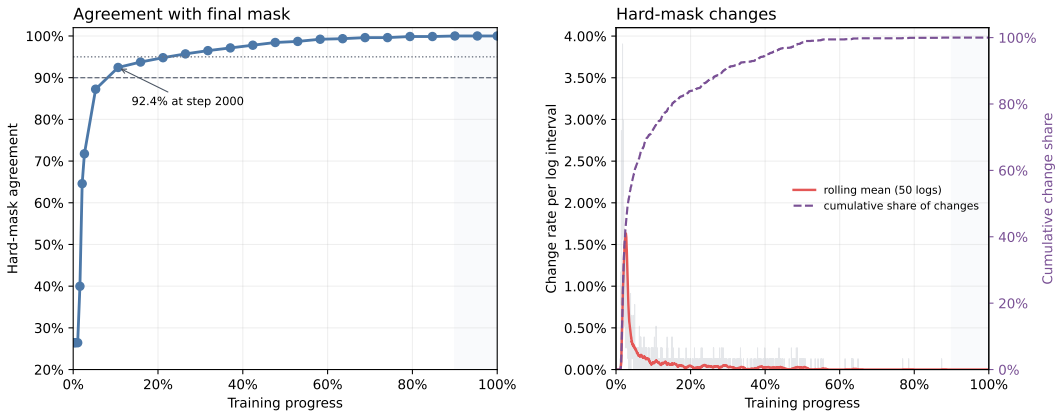


Figure 6: Gate dynamics during training.

and $\theta = 6 \times 10^5$ at 64K. For a head dimension $d_h = 64$, the j-th RoPE frequency is $\omega_j = \theta^{-2j/d_h}$, and the frequency index that completes one rotation over the training context satisfies

$$j^\star = \frac{d_h}{2} \frac{\log(L/2\pi)}{\log \theta}. \tag{16}$$

The frequency-aligned choices give $j^\star \approx 22.4$ at 32K and $j^\star \approx 22.3$ at 64K, thereby preserving approximately the same spectral boundary across training lengths. This schedule is intended as a controlled spectral comparison rather than as the θ that maximizes absolute extrapolation performance.

Table 10 reports both retrieval beyond the training length and short-context performance. A larger θ improves absolute extrapolation performance, and MoPE remains the best method: it obtains Magic OOD averages of 100.00 and 99.47 after 32K and 64K training, respectively. In contrast, the smaller, frequency-aligned θ constitutes a more difficult extrapolation setting. It degrades most methods, but MoPE deteriorates less severely than most baselines, increasing its margin over the strongest non-MoPE method from 2.73 to 6.36 points at 32K and from 15.61 to 22.60 points at 64K. Consequently, the $2 \times 10^5/6 \times 10^5$ schedule demonstrates MoPE's robustness to a challenging spectral setting rather than providing an absolute gain for MoPE itself. We use this frequency-aligned schedule to compare training lengths under a consistent spectral boundary, while the large-θ results show that MoPE's advantage is not tied to this choice.

The long-context gains do not come at the expense of short-context capability. MoPE achieves the highest numerical Short Avg. in all four settings, exceeding the strongest baseline by 0.39–0.69 points. Its Short Avg. also changes by at most 0.08 points between the two θ schedules at either training length, suggesting that θ primarily affects length extrapolation rather than short-context performance.

F.1 Detailed evaluation results

Table 10: Effect of the RoPE base frequency on long- and short-context performance. Magic OOD Avg. averages the evaluation lengths beyond the training context: 48K and 64K for 32K-trained models, and 80K, 96K, 112K, and 128K for 64K-trained models. Short Avg. averages 11 short-context benchmarks. CoPE uses clip size 7 with RoPE initialization, whereas HoPE uses clip size 9. Results are averaged over three runs, except for CoPE under 64K training with $\theta = 5 \times 10^6$, which has two runs, and the 64K HoPE results, which have one run. Higher is better. Bold denotes the best result, and underlining denotes the strongest non-MoPE baseline.

	32K training				64K training			
	$\theta = 10^6$		$\theta = 2 \times 10^5$		$\theta = 5 \times 10^6$		$\theta = 6 \times 10^5$	
Method	Magic OOD	Short Avg.	Magic OOD	Short Avg.	Magic OOD	Short Avg.	Magic OOD	Short Avg.
Full RoPE	59.68	42.34	11.37	42.43	75.83	42.46	21.21	42.25
Partial RoPE 25%	97.27	41.98	0.15	42.02	83.86	41.90	23.70	42.02
Partial RoPE 50%	90.60	42.76	2.12	42.76	82.87	42.85	36.51	42.61
FoPE	82.28	42.78	77.12	42.78	81.74	42.76	75.89	42.68
CoPE (clip7)	60.00	42.28	12.88	42.45	74.54	42.16	22.58	42.28
HoPE	85.60	42.79	66.20	42.90	51.60	42.58	37.50	42.73
MoPE	**100.00**	**43.37**	**83.48**	**43.29**	**99.47**	**43.47**	**98.49**	**43.42**

Table 11: Short-task results for 70M models trained at 4K context.

Benchmark	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	HoPE	FoPE	RoPE-ID	MoPE	MoPE freeze
ARC-C	22.84 ± 0.687	23.18 ± 0.713	23.07 ± 0.341	**23.61 ± 0.423**	22.7 ± 0.294	23.58 ± 0.132	22.95 ± 0.450	22.81 ± 0.663
ARC-E	42.34 ± 0.729	42.2 ± 0.973	42.38 ± 0.112	42.1 ± 0.251	42.48 ± 0.506	42.24 ± 0.756	42.68 ± 0.223$^+$	**42.79 ± 0.089**$^+$
BoolQ	45.43 ± 5.148	52.76 ± 5.771	**54.25 ± 3.721**	50.83 ± 3.816	52.54 ± 1.518	48.79 ± 5.815	53.78 ± 1.604	52.45 ± 2.326
CSQA	19.66 ± 0.294	19.36 ± 0.610	20.2 ± 0.601	19.24 ± 0.294	19.74 ± 0.144	19.11 ± 1.556	20.39 ± 0.456$^+$	**20.61 ± 1.055**$^+$
HellaSwag	28.54 ± 0.169	28.46 ± 0.098	28.47 ± 0.132	28.56 ± 0.098	28.49 ± 0.104	28.59 ± 0.154	28.73 ± 0.185$^+$	**28.8 ± 0.190**$^+$
LAMBADA	17.38 ± 0.425	17.24 ± 0.192	17.32 ± 0.822	18.2 ± 0.497	17.68 ± 0.462	17.65 ± 0.600	**19.75 ± 0.278**$^+$	19.41 ± 0.254$^+$
LogiQA	27.8 ± 0.615	**28.57 ± 1.009**	26.88 ± 0.310	27.19 ± 1.623	28.32 ± 0.583	28.21 ± 0.358	26.57 ± 0.815	26.73 ± 1.712
OBQA	27.07 ± 1.617	27.4 ± 1.039	27.33 ± 0.643	**27.8 ± 0.529**	27.33 ± 0.577	26.47 ± 0.987	27.4 ± 0.917	27.53 ± 0.462
PIQA	59.25 ± 0.572	58.78 ± 0.486	59.05 ± 0.327	59.32 ± 0.542	**59.47 ± 0.144**	59.25 ± 0.641	58.94 ± 0.454	59.07 ± 0.835
SciQ	63.77 ± 0.473	63.57 ± 1.795	65.43 ± 0.551	**65.83 ± 0.833**	63.33 ± 1.909	65.03 ± 0.681	65.57 ± 0.924	65.77 ± 0.907
WinoGrande	**51.8 ± 0.641**	50.96 ± 1.739	51.67 ± 0.795	50.52 ± 0.410	51.62 ± 0.710	50.38 ± 0.251	51.04 ± 0.753	51.54 ± 1.061
Average	36.9 ± 0.535	37.5 ± 0.151	37.82 ± 0.285	37.56 ± 0.481	37.61 ± 0.296	37.21 ± 0.613	**37.98 ± 0.133**$^+$	37.96 ± 0.196$^+$

Table 12: Short-task results for 160M models trained at 4K context.

Benchmark	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	HoPE	FoPE	RoPE-ID	MoPE	MoPE freeze
ARC-C	**25.11 ± 0.872**	24.6 ± 0.819	25.03 ± 0.604	24.63 ± 0.501	24.63 ± 0.552	24.74 ± 1.294	24.92 ± 0.532	24.63 ± 0.565
ARC-E	49.3 ± 0.460	48.32 ± 1.130	49.69 ± 0.525	48.68 ± 0.355	49.19 ± 1.452	48.95 ± 0.634	**50.11 ± 0.615**$^+$	49.7 ± 0.552$^+$
BoolQ	57.99 ± 4.241	**58.55 ± 2.737**	54.86 ± 1.134	56.31 ± 3.856	50.5 ± 4.255	53.55 ± 9.257	57.06 ± 2.853	57.06 ± 3.012
CSQA	19.44 ± 0.806	19.85 ± 0.842	20.17 ± 1.690	18.65 ± 0.548	19.66 ± 0.803	18.84 ± 0.215	**20.47 ± 0.427**$^+$	20.06 ± 0.498
HellaSwag	31.81 ± 0.135	**31.91 ± 0.260**	31.83 ± 0.215	31.55 ± 0.117	31.75 ± 0.127	31.81 ± 0.090	31.79 ± 0.015	31.88 ± 0.065
LAMBADA	24.97 ± 0.320	24.36 ± 0.036	24.59 ± 0.486	24.47 ± 0.506	24.31 ± 0.538	24.66 ± 0.299	**25.78 ± 0.529**$^+$	25.58 ± 0.197$^+$
LogiQA	26.52 ± 1.779	25.81 ± 0.765	27.09 ± 1.079	**27.09 ± 0.846**	26.22 ± 0.847	25.65 ± 0.669	26.21 ± 0.179	26.17 ± 0.724
OBQA	**29.53 ± 1.007**	29.4 ± 1.510	29.33 ± 1.973	28.93 ± 1.553	28.27 ± 1.665	29.47 ± 0.611	27.93 ± 0.833	28.13 ± 0.702
PIQA	61.24 ± 0.136	61.7 ± 0.330	61.22 ± 0.385	61.39 ± 0.082	61.32 ± 0.608	**61.81 ± 0.534**	61.16 ± 0.663	61.26 ± 0.667
SciQ	70.67 ± 0.777	74.83 ± 0.321	75.3 ± 0.500	73.47 ± 0.874	**75.4 ± 0.917**	73.67 ± 1.704	74.53 ± 2.203	75.23 ± 2.316
WinoGrande	**51.93 ± 0.968**	50.88 ± 1.506	51.3 ± 0.876	51.46 ± 0.139	50.28 ± 1.633	51.17 ± 0.318	50.72 ± 1.624	50.75 ± 1.757
Average	40.77 ± 0.229	40.93 ± 0.352	40.95 ± 0.618	40.6 ± 0.506	40.14 ± 0.154	40.39 ± 0.735	**40.97 ± 0.436**$^+$	40.95 ± 0.527$^+$

Table 13: Short-task results for 410M models trained at 4K context.

Benchmark	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	HoPE	FoPE	RoPE-ID	MoPE	MoPE freeze
ARC-C	**27.25 ± 0.386**	26.99 ± 0.793	26.22 ± 0.393	27.13 ± 0.595	26.94 ± 0.890	27.11 ± 0.597	26.59 ± 0.176	27.13 ± 1.120
ARC-E	55.24 ± 0.431	54.91 ± 0.827	54.69 ± 0.106	54.94 ± 0.423	55.81 ± 0.382	55.61 ± 0.343	**56.07 ± 0.590**$^+$	55.92 ± 0.925$^+$
BoolQ	54.18 ± 5.695	58.42 ± 1.153	47.97 ± 1.017	58.04 ± 2.232	54.29 ± 1.822	58.65 ± 2.009	**60.56 ± 1.812**$^+$	60.55 ± 1.140$^+$
CSQA	19.98 ± 0.732	19.25 ± 0.754	19.9 ± 1.065	20.75 ± 0.948	19.28 ± 1.116	20.15 ± 0.376	**20.77 ± 0.122**$^+$	20.42 ± 0.822
HellaSwag	35.69 ± 0.119	35.79 ± 0.111	35.61 ± 0.104	35.84 ± 0.405	35.57 ± 0.195	35.83 ± 0.252	36.03 ± 0.070$^+$	**36.09 ± 0.084**$^+$
LAMBADA	29.61 ± 0.265	30.36 ± 0.110	29.57 ± 0.482	29.96 ± 0.363	29.67 ± 0.572	29.97 ± 0.187	**30.94 ± 0.454**$^+$	29.94 ± 0.186
LogiQA	27.04 ± 1.200	26.27 ± 1.625	26.32 ± 1.079	**27.5 ± 0.958**	25.6 ± 0.087	25.45 ± 0.468	26.52 ± 1.079	26.01 ± 2.438
OBQA	30.27 ± 1.901	30.4 ± 0.400	30.47 ± 1.206	31.27 ± 1.815	30.8 ± 0.917	30.47 ± 0.987	**31.53 ± 1.514**$^+$	**31.53 ± 1.172**$^+$
PIQA	64.05 ± 0.416	63.86 ± 0.698	63.84 ± 0.346	**64.38 ± 0.270**	63.46 ± 0.549	64.33 ± 0.355	63.69 ± 0.396	63.75 ± 0.354
SciQ	80.4 ± 1.400	82.1 ± 0.458	**82.77 ± 0.153**	79.8 ± 1.308	81.33 ± 0.153	82.03 ± 1.106	80.13 ± 1.966	80.47 ± 2.219
WinoGrande	49.83 ± 0.846	51.25 ± 1.461	50.38 ± 1.156	50.7 ± 0.244	51.27 ± 1.194	50.64 ± 0.680	**51.69 ± 0.824**$^+$	51.67 ± 0.397$^+$
Average	43.05 ± 0.260	43.6 ± 0.017	42.52 ± 0.256	43.66 ± 0.250	43.1 ± 0.323	43.66 ± 0.190	**44.05 ± 0.422**$^+$	43.95 ± 0.430$^+$

Table 14: Short-task results for 1B models trained at 4K context.

Benchmark	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	HoPE	FoPE	RoPE-ID	MoPE	MoPE freeze
ARC-C	28.3 ± 1.357	28.78 ± 0.556	27.99 ± 1.066	27.7 ± 0.980	28.76 ± 0.758	**28.78 ± 0.519**	28.38 ± 0.248	27.99 ± 0.372
ARC-E	55.99 ± 0.906	55.93 ± 0.761	55.89 ± 0.154	56.3 ± 0.889	55.96 ± 0.239	56.2 ± 0.321	56.69 ± 0.584^{+}	**57.03 ± 1.095**$^{+}$
BoolQ	58.27 ± 2.789	57.53 ± 4.984	59.61 ± 2.386	58.6 ± 3.196	58.24 ± 3.837	**60.14 ± 0.706**	58.12 ± 1.531	57.87 ± 1.781
CSQA	20.75 ± 1.247	20.72 ± 1.420	20.28 ± 0.743	20.12 ± 0.616	20.83 ± 0.046	**21.02 ± 1.680**	20.88 ± 2.012	20.64 ± 1.772
HellaSwag	36.89 ± 0.249	36.67 ± 0.270	36.66 ± 0.356	36.77 ± 0.159	36.47 ± 0.251	36.45 ± 0.176	**36.97 ± 0.293**$^{+}$	36.86 ± 0.338
LAMBADA	30.74 ± 0.388	30.96 ± 0.702	30.86 ± 1.123	30.66 ± 0.207	30.78 ± 0.723	**31.19 ± 0.059**	31.08 ± 0.930	30.49 ± 0.291
LogiQA	26.27 ± 0.403	**26.99 ± 0.874**	26.37 ± 0.235	24.37 ± 1.042	25.29 ± 0.388	25.55 ± 1.747	26.32 ± 0.937	26.52 ± 1.833
OBQA	32.13 ± 1.102	**33.67 ± 0.987**	31.93 ± 1.747	32.53 ± 0.416	31.87 ± 0.757	32.2 ± 0.721	33.27 ± 0.416	32.13 ± 0.462
PIQA	64.64 ± 0.385	**65 ± 0.382**	64.83 ± 0.311	64.82 ± 0.367	64.56 ± 0.086	64.58 ± 0.724	64.56 ± 0.331	64.65 ± 0.445
SciQ	80.3 ± 1.082	80.1 ± 0.100	81.1 ± 0.529	80.17 ± 0.833	81.63 ± 0.681	79.83 ± 1.343	**82.23 ± 0.751**$^{+}$	81.37 ± 0.551
WinoGrande	50.41 ± 1.071	49.41 ± 0.777	51.67 ± 0.659	50.36 ± 0.752	50.96 ± 1.081	**52.01 ± 1.838**	50.73 ± 1.377	50.62 ± 1.690
Average	44.06 ± 0.600	44.16 ± 0.273	44.29 ± 0.282	43.86 ± 0.520	44.13 ± 0.330	44.36 ± 0.321	**44.48 ± 0.272**$^{+}$	44.2 ± 0.255

Table 15: Short-task results for 410M models trained at 64K context.

Benchmark	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	CoPE clip7 (RoPE init)	HoPE	FoPE	RoPE-ID	RoPE-ID no-scale	MoPE freeze
ARC-C	26.48 ± 0.272	26.11 ± 0.392	25.51 ± 0.372	26.25 ± 0.556	**26.65 ± 0.810**	26.56 ± 0.856	23.15 ± 0.910	26.19 ± 0.309	26.05 ± 1.100
ARC-E	52.67 ± 0.384	52.97 ± 1.429	52.36 ± 1.322	52.41 ± 0.343	52.45 ± 0.615	53.42 ± 0.564	43.76 ± 0.608	53.84 ± 0.461	**53.89 ± 0.613**$^{+}$
BoolQ	53.66 ± 6.179	56.47 ± 2.618	51.22 ± 0.455	53.57 ± 6.252	57.17 ± 3.760	57.02 ± 3.279	59.88 ± 0.301	58.77 ± 1.903	**60.52 ± 1.413**$^{+}$
CSQA	18.48 ± 0.734	18.87 ± 0.734	19.08 ± 0.410	18.84 ± 0.619	**20.42 ± 0.789**	18.65 ± 0.548	20.04 ± 0.905	20.15 ± 1.107	20.23 ± 0.780
HellaSwag	34.69 ± 0.306	34.61 ± 0.372	34.24 ± 0.180	34.64 ± 0.251	34.63 ± 0.182	34.35 ± 0.306	30.95 ± 0.225	34.81 ± 0.266	**34.89 ± 0.155**$^{+}$
LAMBADA	30.36 ± 0.220	29.92 ± 1.134	30.05 ± 0.382	30.34 ± 0.110	30.56 ± 1.853	30.33 ± 0.707	13.5 ± 0.323	**30.84 ± 0.462**	30.53 ± 0.687
LogiQA	25.66 ± 0.266	**26.98 ± 1.699**	25.86 ± 1.616	26.52 ± 0.387	26.83 ± 0.537	25.55 ± 0.936	25.4 ± 2.361	25.5 ± 1.230	26.42 ± 1.513
OBQA	29.93 ± 0.902	29.27 ± 1.102	30.13 ± 2.318	30 ± 0.529	**30.27 ± 0.462**	29.87 ± 0.702	27.47 ± 1.332	29.87 ± 0.643	**30.27 ± 0.945**$^{+}$
PIQA	62.89 ± 0.663	62.89 ± 0.382	63.26 ± 0.401	63.15 ± 0.457	63.16 ± 0.474	62.11 ± 0.440	60.26 ± 0.839	**63.28 ± 0.239**	63.02 ± 0.257
SciQ	79 ± 0.300	79.97 ± 2.329	**80.93 ± 0.850**	78.97 ± 0.115	78.07 ± 1.050	80.33 ± 0.945	66.67 ± 2.065	78.87 ± 1.405	80 ± 1.868
WinoGrande	50.93 ± 0.483	50.65 ± 0.939	49.51 ± 1.568	50.41 ± 0.449	51.62 ± 0.802	51.33 ± 0.848	51.51 ± 0.201	50.17 ± 1.641	**51.88 ± 0.641**$^{+}$
Average	42.25 ± 0.557	42.61 ± 0.662	42.02 ± 0.482	42.28 ± 0.564	42.9 ± 0.360	42.68 ± 0.474	38.42 ± 0.346	42.93 ± 0.291	**43.42 ± 0.352**$^{+}$

Table 16: Short-task results for 1B models trained at 64K context.

Benchmark	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	CoPE clip28 (RoPE init)	HoPE	FoPE	RoPE-ID	RoPE-ID no-scale	MoPE gn0.75 freeze
ARC-C	28.07 ± 0.729	**28.16 ± 0.150**	27.53 ± 0.301	28.01 ± 0.708	27.9 ± 0.596	26.94 ± 0.354	21.84 ± 0.783	27.87 ± 0.386	27.27 ± 0.543
ARC-E	54.03 ± 0.950	54.48 ± 0.441	53.31 ± 0.713	54.1 ± 0.890	54.61 ± 1.133	54.11 ± 0.409	43.46 ± 0.384	**55.05 ± 0.696**	54.91 ± 0.758
BoolQ	57.58 ± 1.148	59.42 ± 3.003	60.05 ± 1.965	57.05 ± 1.703	**60.09 ± 1.781**	55.96 ± 6.008	56.1 ± 3.429	59.53 ± 1.714	59.57 ± 0.829
CSQA	21.08 ± 0.092	20.01 ± 0.810	19.63 ± 1.198	21.59 ± 0.046	20.75 ± 1.273	20.86 ± 0.329	**22 ± 1.214**	20.5 ± 0.617	20.88 ± 0.294
HellaSwag	**36.12 ± 0.130**	35.72 ± 0.235	35.83 ± 0.187	36.11 ± 0.079	35.83 ± 0.142	35.65 ± 0.173	30.14 ± 0.099	35.59 ± 0.217	35.98 ± 0.323
LAMBADA	30.97 ± 0.162	31.3 ± 1.014	31.85 ± 1.010	31.08 ± 0.250	31.86 ± 0.495	31.77 ± 0.829	14.64 ± 0.465	**33.2 ± 0.601**	31.54 ± 0.988
LogiQA	26.52 ± 1.108	24.99 ± 1.450	26.37 ± 0.173	**26.94 ± 1.570**	24.88 ± 1.312	26.01 ± 1.233	25.75 ± 0.621	25.75 ± 1.976	26.37 ± 0.991
OBQA	31.93 ± 1.793	**32.2 ± 1.058**	30.87 ± 0.902	32.07 ± 2.082	30.53 ± 1.447	30.6 ± 1.058	28.53 ± 0.702	31.4 ± 0.200	32.13 ± 1.804
PIQA	64.15 ± 0.450	64.42 ± 0.500	64.47 ± 0.425	64.15 ± 0.285	64.38 ± 0.635	64.29 ± 0.452	59.45 ± 0.570	**64.51 ± 0.548**	63.84 ± 0.219
SciQ	80.93 ± 0.987	80.57 ± 1.266	81.27 ± 0.950	80.73 ± 0.981	80.13 ± 0.473	81.97 ± 1.290	64.6 ± 0.854	79.6 ± 0.889	**82.3 ± 0.200**$^{+}$
WinoGrande	49.7 ± 1.938	49.23 ± 0.989	**52.43 ± 0.356**	49.88 ± 1.537	51.75 ± 0.239	50.62 ± 1.623	49.94 ± 1.109	51.41 ± 1.048	50.91 ± 1.123
Average	43.73 ± 0.335	43.68 ± 0.301	43.96 ± 0.332	43.79 ± 0.319	43.89 ± 0.318	43.52 ± 0.473	37.86 ± 0.185	44.04 ± 0.036	**44.15 ± 0.310**$^{+}$

Table 17: Magic-number NIAH accuracy (%) for 1B models trained at 64K context.

Method	@8K	@10K	@12K	@14K	@16K	@24K	@32K	@48K	@64K	@80K	@96K	@112K	@128K
Full RoPE	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	99.7 ± 0.520	97.6 ± 1.873	85.43 ± 11.064	12.4 ± 17.586	0 ± 0.000
Partial RoPE 50%	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	18.77 ± 13.329	0 ± 0.000
Partial RoPE 25%	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	18.8 ± 25.098	0 ± 0.000
CoPE clip28 (RoPE init)	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	99.4 ± 0.520	99.7 ± 0.520	96.67 ± 3.691	85.73 ± 9.985	28.8 ± 20.000	0.3 ± 0.520
HoPE	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	98.5 ± 2.598	97.27 ± 4.734	96.97 ± 5.254	93.03 ± 10.546
FoPE	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	99.7 ± 0.520	**100 ± 0.000**	99.4 ± 1.039	99.4 ± 0.520	99.4 ± 1.039	99.1 ± 0.900	**100 ± 0.000**
RoPE-ID	3.63 ± 3.301	13.67 ± 5.532	25.77 ± 6.374	45.17 ± 13.692	49.1 ± 10.708	75.17 ± 9.113	88.17 ± 8.736	96.97 ± 2.793	93.97 ± 7.550	97.87 ± 2.950	94.87 ± 5.173	94.53 ± 4.179	96.97 ± 3.695
RoPE-ID no-scale	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	**100 ± 0.000**	99.1 ± 1.559	98.8 ± 2.078	99.1 ± 1.559	98.2 ± 2.381	98.5 ± 1.375	96.07 ± 2.136	92.13 ± 4.990	83.63 ± 11.173	79.7 ± 13.275
MoPE gn0.75 freeze	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	99.7 ± 0.520	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	99.7 ± 0.520	**100 ± 0.000**$^{+}$	99.7 ± 0.520	**100 ± 0.000**$^{+}$	98.8 ± 1.039

Table 18: Magic-number NIAH accuracy (%) for 410M models trained at 64K context.

Method	@8K	@10K	@12K	@14K	@16K	@24K	@32K	@48K	@64K	@80K	@96K	@112K	@128K
Full RoPE	96.07 ± 6.813	98.17 ± 3.175	96.07 ± 6.813	97.57 ± 4.215	95.17 ± 8.372	96.97 ± 5.254	93.37 ± 6.813	66.97 ± 8.563	54.83 ± 6.050	42.1 ± 6.972	33.03 ± 5.008	9.1 ± 1.800	0.6 ± 0.520
Partial RoPE 50%	98.5 ± 1.873	98.8 ± 1.039	98.8 ± 2.078	97.9 ± 2.265	97.57 ± 2.804	99.1 ± 1.559	97.87 ± 2.950	96.97 ± 3.803	93.93 ± 3.695	84.83 ± 12.138	58.17 ± 14.636	3.03 ± 2.950	0 ± 0.000
Partial RoPE 25%	98.5 ± 2.598	**100 ± 0.000**	99.4 ± 1.039	99.7 ± 0.520	**100 ± 0.000**	99.7 ± 0.520	99.7 ± 0.520	99.7 ± 0.520	95.73 ± 3.803	80.27 ± 14.392	14.53 ± 6.293	0 ± 0.000	0 ± 0.000
CoPE clip7 (RoPE init)	97.57 ± 4.215	98.8 ± 1.375	99.4 ± 1.039	98.5 ± 2.598	97.57 ± 4.215	98.17 ± 3.175	96.67 ± 2.804	69.4 ± 6.421	56.33 ± 4.734	44.87 ± 7.271	34.27 ± 2.136	9.4 ± 1.873	1.8 ± 1.559
HoPE	99.4 ± 1.039	98.8 ± 2.078	98.17 ± 3.175	98.5 ± 2.598	97.27 ± 4.734	95.17 ± 6.872	89.4 ± 17.586	88.5 ± 14.470	91.83 ± 11.884	79.1 ± 27.008	72.7 ± 26.302	67.9 ± 34.812	53.63 ± 24.601
FoPE	**100 ± 0.000**	98.8 ± 1.375	99.7 ± 0.520	97.27 ± 4.734	98.5 ± 2.598	95.77 ± 7.332	93.63 ± 9.511	91.83 ± 14.145	88.2 ± 17.414	83.63 ± 25.230	79.37 ± 30.283	75.13 ± 29.650	65.43 ± 31.261
RoPE-ID	24.23 ± 9.072	38.5 ± 10.912	36.07 ± 10.538	42.7 ± 7.465	46.97 ± 10.661	61.5 ± 14.832	78.5 ± 5.196	93.93 ± 6.369	96.37 ± 3.287	95.77 ± 7.332	92.4 ± 8.266	84.83 ± 12.200	85.17 ± 20.281
RoPE-ID no-scale	99.1 ± 0.900	**100 ± 0.000**	99.7 ± 0.520	99.1 ± 0.900	99.4 ± 1.039	98.5 ± 2.598	97.57 ± 2.804	98.2 ± 2.381	95.43 ± 3.980	91.2 ± 9.115	75.77 ± 12.036	29.07 ± 13.411	2.13 ± 2.950
MoPE freeze	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**100 ± 0.000**$^{+}$	**99.1 ± 0.900**$^{+}$	**94.87 ± 5.014**$^{+}$

Table 19: Short-task results for 410M models trained at 64K context and extrapolated with YaRN. Values are accuracy (%, mean ± sample standard deviation over three runs). Average is the unweighted mean over all 11 tasks, including LAMBADA. Normalized accuracy is used when available; otherwise accuracy is used.

Benchmark	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	CoPE clip7	HoPE	FoPE	RoPE-ID	RoPE-ID no-scale	MoPE (Ours)
ARC-C	27.70 ± 0.859	27.25 ± 1.114	26.05 ± 0.300	27.47 ± 1.210	27.50 ± 0.806	**27.93 ± 0.355**	22.81 ± 1.080	25.71 ± 0.130	26.82 ± 0.982
ARC-E	52.55 ± 0.450	**53.40 ± 0.413**	52.88 ± 0.466	52.50 ± 0.499	52.92 ± 0.697	53.23 ± 0.502	39.51 ± 1.142	50.34 ± 0.367	53.04 ± 0.779
BoolQ	53.77 ± 6.301	51.70 ± 3.497	48.02 ± 3.359	54.07 ± 6.127	56.77 ± 3.494	54.74 ± 3.013	47.44 ± 4.529	**60.22 ± 1.395**	58.75 ± 2.711
CSQA	20.17 ± 0.288	19.14 ± 1.251	18.67 ± 0.573	20.28 ± 0.581	20.42 ± 1.112	**20.48 ± 0.591**	20.26 ± 0.898	19.63 ± 0.949	19.36 ± 1.112
HellaSwag	**35.43 ± 0.211**	35.30 ± 0.210	34.83 ± 0.322	35.38 ± 0.214	35.04 ± 0.254	35.09 ± 0.346	28.05 ± 0.144	33.04 ± 0.333	35.10 ± 0.160
LAMBADA	32.22 ± 0.784	31.52 ± 1.428	**32.71 ± 1.630**	31.75 ± 0.639	30.43 ± 1.782	29.24 ± 1.218	9.28 ± 0.432	23.99 ± 0.695	26.88 ± 0.858
LogiQA	25.19 ± 0.670	25.24 ± 1.427	24.32 ± 0.846	25.55 ± 1.279	24.63 ± 0.582	25.40 ± 1.515	24.37 ± 1.153	**25.86 ± 1.877**	25.35 ± 2.009
OBQA	29.33 ± 0.987	30.60 ± 0.400	29.80 ± 1.200	29.87 ± 0.231	29.60 ± 1.800	**30.87 ± 0.462**	26.40 ± 1.114	29.00 ± 0.346	30.60 ± 0.872
PIQA	63.38 ± 0.249	63.08 ± 1.205	**63.87 ± 0.392**	63.22 ± 0.326	63.28 ± 0.425	63.04 ± 0.408	58.22 ± 0.930	62.80 ± 0.545	63.69 ± 0.638
SciQ	79.30 ± 0.721	78.53 ± 1.401	80.23 ± 0.808	79.13 ± 0.808	78.17 ± 1.815	**80.43 ± 0.808**	56.33 ± 1.756	75.83 ± 1.890	78.80 ± 1.311
WinoGrande	50.91 ± 0.418	51.09 ± 1.155	50.75 ± 1.026	50.96 ± 0.277	50.22 ± 1.527	51.41 ± 0.537	51.28 ± 1.739	50.30 ± 0.733	**52.07 ± 0.462**
Average	42.72 ± 0.455	42.44 ± 0.564	42.01 ± 0.038	42.74 ± 0.514	42.63 ± 0.368	**42.90 ± 0.197**	34.90 ± 0.557	41.52 ± 0.147	42.77 ± 0.484

Table 20: PG19 perplexity for the same 410M YaRN models (mean ± sample standard deviation over three runs; lower is better). Lengths are binary: 64K, 128K, and 256K correspond to 65,536, 131,072, and 262,144 tokens. The evaluations contain 54, 20, and 6 eligible texts, respectively.

Length	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	CoPE clip7	HoPE	FoPE	RoPE-ID	RoPE-ID no-scale	MoPE (Ours)
64K	17.20 ± 0.033	17.59 ± 0.404	23.93 ± 0.519	17.20 ± 0.028	17.24 ± 0.074	17.18 ± 0.080	22.74 ± 1.346	21.20 ± 1.097	**16.68 ± 0.109**
128K	17.82 ± 0.039	18.44 ± 0.442	30.84 ± 0.141	17.82 ± 0.029	17.88 ± 0.090	17.80 ± 0.069	21.06 ± 0.990	22.24 ± 1.195	**17.14 ± 0.113**
256K	17.70 ± 0.028	18.35 ± 0.452	34.21 ± 0.921	17.70 ± 0.018	17.77 ± 0.080	17.63 ± 0.064	20.57 ± 0.719	25.87 ± 2.908	**16.83 ± 0.117**

Table 21: Single-needle magic-number retrieval accuracy for the same 410M YaRN models (%, mean ± sample standard deviation over three runs; higher is better).

Length	Full RoPE	Partial RoPE 50%	Partial RoPE 25%	CoPE clip7	HoPE	FoPE	RoPE-ID	RoPE-ID no-scale	MoPE (Ours)
4K	96.37 ± 3.147	98.48 ± 1.390	99.09 ± 1.576	98.48 ± 2.627	**100.00 ± 0.000**	99.70 ± 0.525	0.00 ± 0.000	77.88 ± 26.544	**100.00 ± 0.000**
8K	96.67 ± 5.774	96.67 ± 5.774	98.18 ± 1.576	**100.00 ± 0.000**	**100.00 ± 0.000**	99.39 ± 1.051	0.00 ± 0.000	83.94 ± 3.191	**100.00 ± 0.000**
16K	97.58 ± 4.197	97.27 ± 2.408	98.48 ± 1.051	**100.00 ± 0.000**	99.09 ± 1.576	97.58 ± 4.197	0.00 ± 0.000	83.33 ± 9.719	**100.00 ± 0.000**
32K	96.06 ± 5.326	93.94 ± 10.496	94.85 ± 6.702	99.39 ± 1.051	98.79 ± 2.102	97.27 ± 4.723	17.88 ± 10.055	75.76 ± 6.945	**100.00 ± 0.000**
48K	95.15 ± 8.400	99.09 ± 1.576	46.06 ± 19.840	98.79 ± 1.390	95.76 ± 7.350	95.76 ± 7.350	34.24 ± 5.555	65.46 ± 9.619	**99.70 ± 0.525**
64K	94.85 ± 7.406	96.36 ± 4.810	23.64 ± 11.818	98.79 ± 1.390	92.12 ± 13.649	90.91 ± 13.455	40.61 ± 8.639	52.43 ± 10.055	**99.70 ± 0.525**
96K	89.70 ± 8.248	89.39 ± 14.438	8.79 ± 9.719	95.15 ± 2.285	87.58 ± 19.962	92.12 ± 12.869	53.03 ± 12.273	44.85 ± 12.736	**99.39 ± 1.051**
128K	80.31 ± 12.410	96.36 ± 4.723	2.73 ± 2.402	88.79 ± 5.555	94.24 ± 9.971	84.54 ± 25.212	61.82 ± 12.729	34.24 ± 19.609	**99.70 ± 0.525**
160K	66.67 ± 12.605	84.85 ± 11.435	2.43 ± 1.390	76.67 ± 5.007	88.18 ± 20.467	86.06 ± 24.145	62.12 ± 8.446	21.51 ± 11.583	**99.09 ± 0.910**
192K	52.43 ± 9.105	85.76 ± 4.575	1.52 ± 1.051	60.30 ± 1.051	86.67 ± 21.537	80.30 ± 34.116	58.18 ± 6.860	6.06 ± 6.945	**99.09 ± 1.576**
224K	46.97 ± 5.916	83.33 ± 14.093	2.12 ± 1.390	48.79 ± 1.051	73.03 ± 42.833	80.30 ± 32.552	54.24 ± 19.504	1.52 ± 2.627	**99.09 ± 0.910**
256K	39.09 ± 8.671	78.48 ± 17.636	1.82 ± 2.408	43.94 ± 5.555	75.45 ± 35.426	75.15 ± 37.662	44.55 ± 16.544	0.30 ± 0.525	**99.09 ± 0.910**