LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens

Yiran Ding, **Li Lyna Zhang**, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, Mao Yang

Microsoft Research

https://github.com/microsoft/LongRoPE

Background

❖ **Context window: How far the LLM can see**

- Pre-trained LLMs have a limited context window
- GPT-4/LLaMA2: 4k tokens
- ~10 pages in a book, ~14 seconds of a video

❖ **Larger context window -> Greater Capabilities**

- LLM with a 2 million context window can:
- Read 7 Harry Potter Books at one shot
- Watch a 2-hour movie
- Listen to a 20-hour audio

Preliminary and Key Challenges

RoPE Interpolation and then fine-tuning, can effectively extend LL context window

RoPE:

$[cos(n\theta_0), sin(n\theta_0), cos(n\theta_1), \cdots, cos(n\theta_{d/2-1}), sin(n\theta_{d/2-1})]$

Rescaled-RoPE (NTK, PI, YaRN):

$$\left[cos\left(\frac{n}{\lambda(\beta)^0}\right), sin\left(\frac{n}{\lambda(\beta)^0}\right), cos\left(\frac{n}{\lambda(\beta)^1}\right), \cdots, sin\left(\frac{n}{\lambda(\beta)^{d/2-1}}\right)\right]$$

Where $\beta = \theta^{2/d}$, θ is 10000

Challenges in further extending LLM context window:

❖ **Non-uniformities in RoPE embedding. Current RoPE-based extension do not fully consider the subtle non-uniformities**

Method	λ
PI	Extension ratio, $\lambda = s$
NTK	$\lambda = s^i$
YaRN	Divide RoPE dims into 3 groups, perform PI, NTK and direct extrapolation

❖ **Fine-tuning is extremely expensive and long text data is scarce**

❖ **Performance drop on the original short context**

Methodology

Step1: Non-uniform RoPE Interpolation and Extrapolation

- evolution search for RoPE rescaling factors

$$\underset{\mathbf{x}\in\mathbf{X};\,|\mathbf{x}|\geq L'}{\arg\min}\ \mathcal{L}\left(\text{LLM}(\text{RoPE}, \mathbf{X})\right), \text{where}$$

$$\underset{\substack{i=0,\cdots,\frac{d}{2}-1;\\ n\in[0,|\mathbf{x}|);}}{\text{RoPE}(n)} = \left[\cdots, cos\left(\mathbb{I}(\hat{\lambda}_i, \hat{n}) \times \frac{n}{\beta^i}\right), sin\left(\mathbb{I}(\hat{\lambda}_i, \hat{n}) \times \frac{n}{\beta^i}\right), \cdots\right]$$

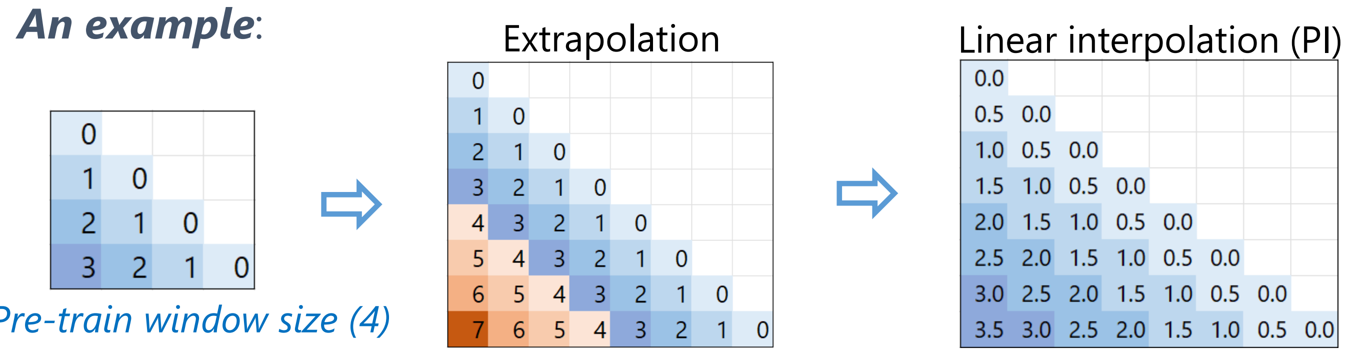
$$\text{where}\ \mathbb{I}(\hat{\lambda}_i, \hat{n}) = \begin{cases} 1 & n < \hat{n} \\ \frac{1}{\lambda_i} & n \geq \hat{n} \end{cases}$$

Algorithm 1 The search algorithm

Input: target LLM, input samples $\mathbf{X}$, population size P, mutation size N_1, crossover size N_2, max iterations $\mathcal{T}$, mutate probability p

```
1: Top-k=ϕ;
2: P0=Initialize_population_with_optimization (P, X, p);
3: for i=1 to T do
4:    Compute_perplexity (LLM, P_{i-1}, X);
5:    Top-k = Update_Topk (Top-k, P_{i-1});
6:    P_mutation=Mutation_with_mono_constraint (Top-k, N1, p);
7:    P_crossover=Crossover_with_mono_constraint (Top-k, N2);
8:    P_i = P_mutation ∪ P_crossover ∪ Top-k;
9: end for
10: Return the individual with lowest perplexity in Top-k;
```

An example:

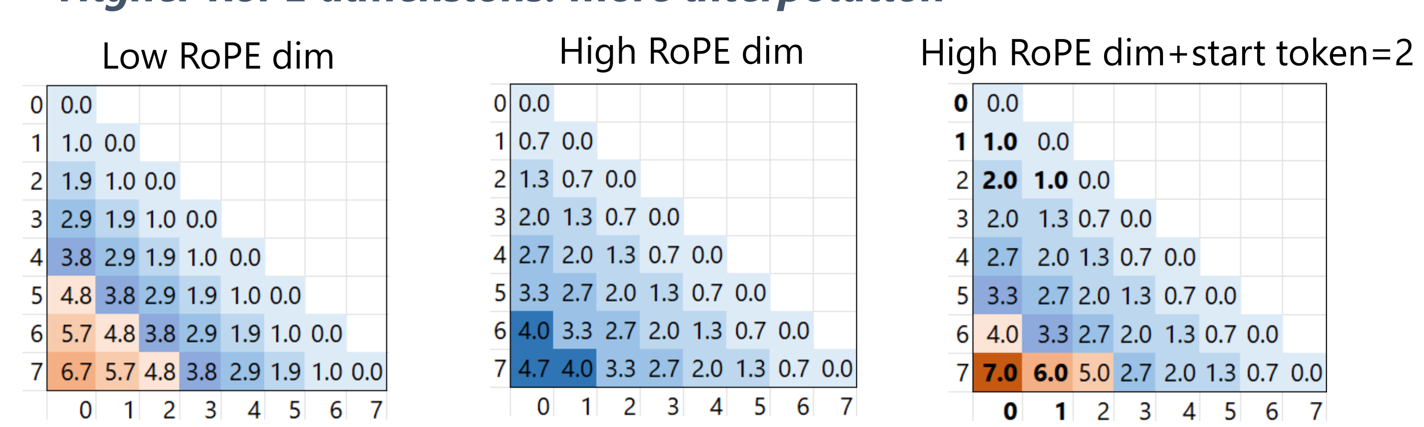


Pre-train window size (4) *Inference sequence length (8)* *Rescale factor=2*

Our searched Non-Uniform RoPE:

✓ ***Lower RoPE dimensions and initial token positions: less interpolation***

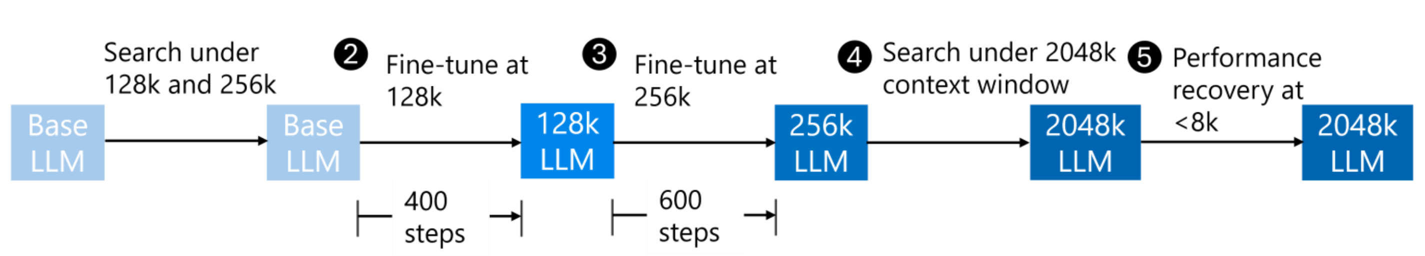
✓ ***Higher RoPE dimensions: more interpolation***



Rescale factor=1.05 *Rescale factor=1.5* *Rescale factor=1.5, initial 2 tokens with no interpolation*

Step2: Progressive Extension to 2 Million Context Window

- 1k fine-tuning steps at 256k text lengths
- Non-uniform positional interpolation allows 8x extension without fine-tuning



Step3: Short Performance Recovery

- Attention becomes dispersed as it's spread thinly across vast positions
- Readjust RoPE on shorter context lengths, less interpolation
- Increase the attention entropy via introducing a temperature t

$$Attention = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \rightarrow softmax\left(\frac{QK^T}{t\sqrt{d}}\right)V$$

Experiments

❖ **Long sequence language modeling**

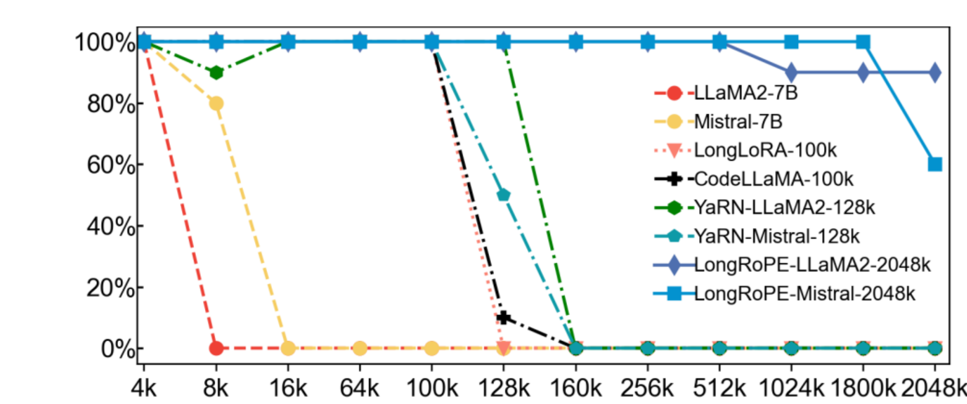
Table 5. Proof-pile perplexity of models with various positional interpolation methods. ft: the context window size used in fine-tuning. Even with a context window 16× longer than current long-context models, our models also outperform them within 256k context length.

Base LLM	Model Name	Context Window	Extension Method	4096	8192	32768	65536	98304	131072	262144
LLaMA2-7B	LLaMA2-7B	4k	-	**3.58**	$>10^4$	$>10^4$	$>10^4$	$>10^4$	$>10^4$	$>10^4$
	Together	32k	PI	3.69	3.50	2.64	$>10^2$	$>10^3$	$>10^4$	$>10^4$
	LongLoRA	100k	PI	3.83	3.62	2.68	2.44	2.33	9.89	$>10^3$
	Code LLaMA	100k	NTK	3.95	3.71	2.74	2.55	2.54	2.71	49.33
	YaRN (s=16)	64k	YaRN	3.69	3.51	2.65	2.42	$>10^1$	$>10^1$	$>10^4$
	YaRN (s=32)	128k	YaRN	3.75	3.56	2.70	2.45	2.36	2.37	99.64
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	3.67	**3.49**	**2.60**	**2.36**	**2.27**	**2.26**	**1.88**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	3.69	3.64	**2.63**	**2.38**	**2.28**	**2.26**	**1.87**
Mistral-7B	Mistral v0.1	8k	-	**3.09**	2.96	$>10^2$	$>10^3$	$>10^3$	$>10^3$	$>10^4$
	YaRN (s=8)	64k	YaRN	3.18	3.04	2.37	2.20	10.39	57.4	$>10^4$
	YaRN (s=16)	128k	YaRN	3.21	3.08	2.41	2.24	2.18	2.19	4.91
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	**3.09**	**2.95**	**2.31**	**2.12**	**2.06**	**2.06**	**1.77**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	3.10	**2.96**	**2.30**	**2.12**	**2.06**	**2.06**	**1.77**

Table 6. Perplexity evaluation on Books3 dataset. Without additional fine-tuning, our LongRoPE-2048k models, with a training context window size of 128k and 256k, effectively scale to an extremely long context size of 2048k. 1k=1024 tokens.

Base LLM	Model Name	Context Window	Extension Method	8k	16k	32k	64k	128k	256k	512k	1024k	2048k
LLaMA2-7B	LongLoRA	100k	PI	6.99	6.80	6.66	6.59	20.57	246.45	$>10^3$	$>10^4$	$>10^4$
	Code LLaMA	100k	NTK	7.68	7.49	7.38	7.88	9.80	98.30	$>10^3$	$>10^4$	$>10^4$
	YaRN (s=16)	64k	YaRN	**6.33**	**6.20**	**6.11**	**6.06**	$>10^4$	$>10^4$	$>10^4$	$>10^4$	$>10^4$
	YaRN (s=32)	128k	YaRN	6.38	6.25	6.16	6.11	**6.12**	$> 10^4$	$>10^4$	$>10^4$	$>10^4$
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	6.53	6.35	6.24	6.18	6.17	**6.17**	**6.36**	**6.83**	**7.80**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	6.79	6.66	6.31	6.27	6.21	**6.17**	**6.17**	**6.35**	**7.08**
Mistral-7B	Mistral v0.1	8k	-	**6.32**	66.61	$>10^2$	$>10^3$	$>10^3$	$>10^3$	-	-	-
	YaRN (s=8)	64k	YaRN	6.59	6.48	6.42	6.45	104.15	727.20	$> 10^3$	$> 10^4$	$> 10^4$
	YaRN (s=16)	128k	YaRN	6.70	6.63	6.65	6.72	6.85	99.90	$> 10^3$	$> 10^4$	$> 10^4$
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	6.42	**6.25**	**6.14**	**6.18**	**6.31**	**6.51**	**6.93**	**7.51**	**9.48**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	6.44	**6.28**	**6.19**	**6.19**	**6.35**	**6.61**	**7.40**	**7.75**	**11.25**

❖ **Long context retrieve**



❖ **Short performance at original context window**

Table 8. Comparison of long-context LLMs with original LLaMA2 and Mistral on the Hugging Face Open LLM benchmark.

Model	Context Window	ARC-c	HellaSwag	MMLU	TruthfulQA
(a) LLaMA2-7B with extended context window					
Original LLaMA2-7B	4k	53.1	78.6	46.6	39.0
Together	32k	47.6	76.1	43.3	39.2
Code LLaMA	100k	42.4	64.8	40.1	37.1
YaRN (s=16)	64k	52.4	**78.7**	42.4	38.2
YaRN (s=32)	128k	52.2	78.5	41.8	37.4
LongRoPE-2048k (ft=128k)	2048k	**53.3**	77.6	**45.2**	**39.6**
LongRoPE-2048k (ft=256k)	2048k	**54.1**	77.8	**44.4**	38.9
(b) Mistral-7B with extended context window					
Original Mistral-7B	8k	60.6	83.2	63.6	42.6
MistralLite (Amazon, 2023)	16k	59.2	81.6	50.4	38.3
YaRN (s=8)	64k	59.3	81.3	**61.3**	42.5
YaRN (s=16)	128k	59.0	80.5	60.5	42.5
LongRoPE-2048k (ft=128k)	2048k	59.0	**81.7**	60.9	**43.9**
LongRoPE-2048k (ft=256k)	2048k	**59.8**	81.4	60.9	**44.1**

❖ **Ablation study on the non-uniformities**

Table 11. Ablation study on the two forms of non-uniformities.

Methods	LLaMA2-7B PG19 Perplexity 16k	LLaMA2-7B PG19 Perplexity 32k	LLaMA2-7B (ft=256k) Books3 Perplexity 2048k
Linear interpolation (PI)	14.88	136.30	20.17
RoPE dim (Ours)	7.28	13.00	7.08
RoPE dim+Start tokens (Ours)	**7.22**	**11.51**	**7.08**

LongRoPE in Phi3-128k series

❖ **More challenging long-context benchmarks**

Ruler

Models	Context Window	4k	8k	16k	32k	64k	128k	Avg
Gemini-1.5-pro	1M	96.7	95.8	96	95.9	95.9	94.4	95.8
GPT-4-1106-preview	128k	96.6	96.3	95.2	93.2	87	81.2	91.6
Command-R-plus (104B)	128k	95.6	95.2	94.2	92.0	84.3	63.1	87.4
GradientAI/LLaMA3 (70B)	1M	95.2	93.4	93.4	89.4	82.6	72	87.7
Phi3-mini-128k (3.8B)	**128k**	**92.3**	**91.2**	**90.8**	**87.7**	**79.8**	**65.3**	**84.5**
Mixtral-8x22B	64k	95.6	94.9	93.4	90.9	84.7	31.7	81.9
LVM (7B)	1M	82.3	78.4	73.7	69.1	68.1	65.0	72.8
FILM-7B	32k	92.8	88.2	88.1	86.9	70.1	27.1	75.5
ChatGLM (6B)	128k	87.8	83.4	78.6	69.9	56.0	42.0	69.6
LongChat (7B)	32k	84.7	79.9	70.8	59.3	0	0	49.1

Long context code understanding (RepoQA)

		Python	cpp	java	typescript	rust	avg
GPT-4o-2024-05-13	128k	95	80	85	96	97	90.6
Gemini-1.5-pro-latest	1M	91	81	91	94	96	90.6
claude-3-opus-20240229	200k	93	83	88	95	94	90.6
phi3-mini-128k-instruct	**128k**	**86**	**64**	**73**	**94**	**71**	**77.6**
GPT-4-turbo-2024-04-09	128k	84	79	75	89	55	76.4
Mixtral-8x22B-Instruct-v0.1	64k	60	67	74	83	55	67.8

❖ **More short tasks**

	Phi3-mini-128k-instruct	Mistral-7B	Gemma 7B	LLaMA3-Instruct-8B	Mixtral 8x7B
MMLU	68.1	61.7	63.6	66.5	**68.4**
GSM8K	**83.6**	46.4	59.8	77.4	64.7
MedQA	55.3	49.6	50	60.5	**62.2**
AGIEval	36.9	35.1	42.1	42	**45.2**
BBH-Hard	**71.5**	57.3	59.6	51.5	69.7
HumanEval	57.9	28	34.1	60.4	37.8

❖ **Multi-modality long context support**

	Phi3-vision-128k-instruct	LLaVA-1.6-vicuna-7B	QWEN-VL-Chat	LLaMA3-LLaVA-Next-8B	Claude-3-Haiku	Gemini 1.0 Pro V	GPT-4V-Turbo
MMMU	40.4	34.2	39	36.4	40.7	42	**55.5**
MMBench	80.5	76.3	75.8	79.4	62.4	80	**86.1**
ScienceQA	**90.8**	70.6	67.2	73.7	72	79.7	75.7
MathVista	44.5	31.5	29.4	34.8	33.2	35.0	**47.5**
InterGPS	38.1	20.5	22.3	24.6	32.1	28.6	**41.0**
ChartQA	**81.4**	55.0	50.9	65.8	59.3	58.0	62.3

Conclusion

❖ We present LongRoPE, a method that remarkably extends pretrained LLMs context window beyond 2 million tokens, while maintaining capabilities within original short context window

❖ LongRoPE exploits two forms of non-uniformities in RoPE using an efficient evolution search

❖ LongRoPE introduces a progressive extension strategy to reach a 2048k context window without direct fine-tuning at 2M text length

❖ Extensive experiments and the application on Phi3-128k-series demonstrate the effectiveness of LongRoPE