LLM Maybe LongLM: SelfExtend LLM Context Window Without Tuning

Hongye Jin [1 *] Xiaotian Han [1 *] Jingfeng Yang [2] Zhimeng Jiang [1] Zirui Liu [3] Chia-Yuan Chang [1]
Huiyuan Chen [4] Xia Hu [3]

ICML 2024 Spotlight

Presenter: Jiyun Bae, Chanyoung Gwak, Junhee Cho

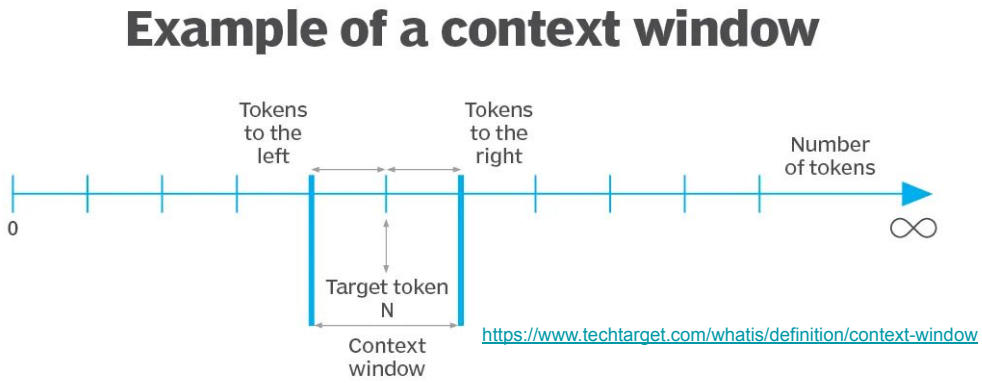
Contents

01 Background & Preliminary

What is context window?

“The context window refers to the maximum amount of text that an language model can process at one time.”

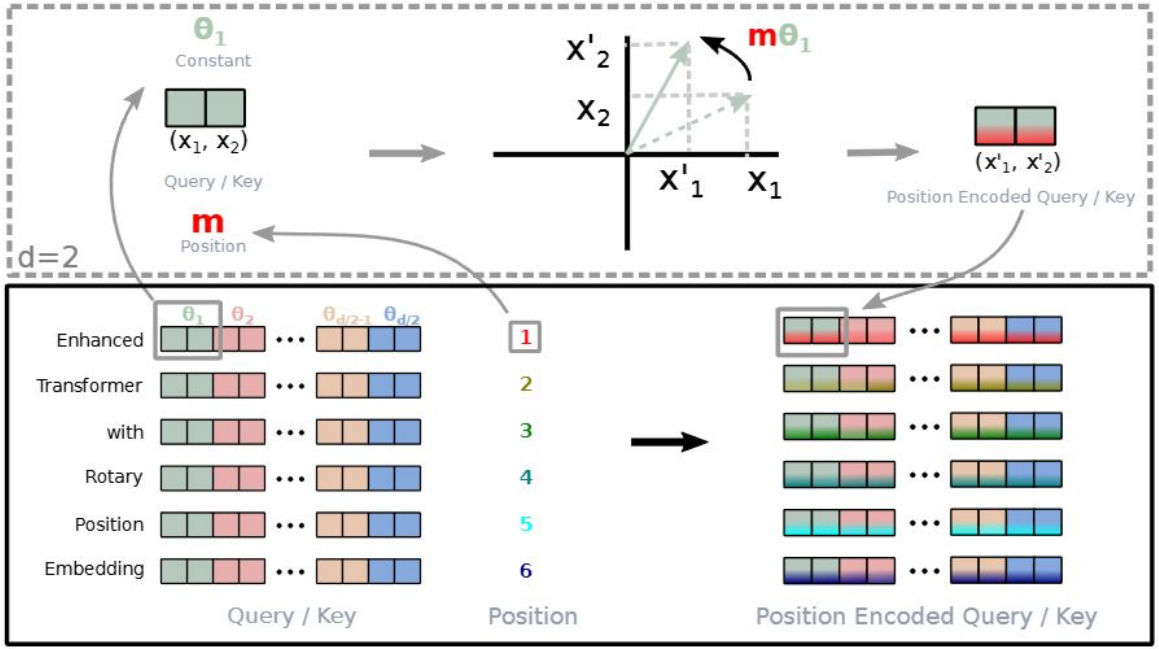
Tokenization @ https://huggingface.co/spaces/Xenova/the-tokenizer-playground, model: gpt-4



If context window is 4, the model processes only, "The", " context", " window", “ refers”, distorting subsequent tokens as they exceed the window.

Then is a larger context window size always better?

Rotary Position Embedding (RoPE)



Su, Jianlin, et al. "Roformer: Enhanced transformer with rotary position embedding." *Neurocomputing* 568 (2024): 127063.

- The position information is encoded through the rotation of vectors by $m\theta$.

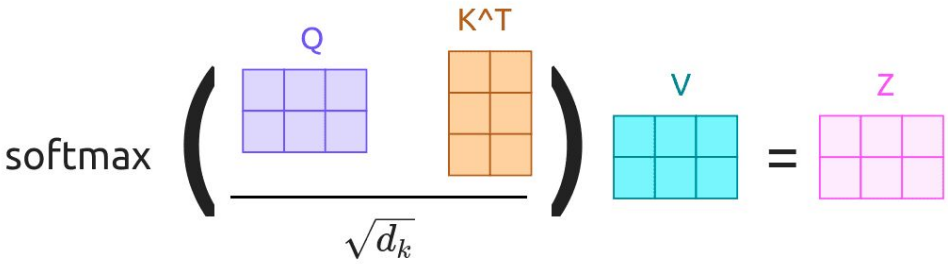
$$\mathbf{q_m} = f_q\left(\mathbf{x_m}, m\right) = W_q \mathbf{x_m} e^{im\theta}$$

$$\mathbf{k_n} = f_k\left(\mathbf{x_n}, n\right) = W_k \mathbf{x_n} e^{in\theta}$$

- The position-encoded query/key vectors are applied during the attention calculation.

Self attention integrated with RoPE

- Self-attention computes relationships with other tokens within the window.



https://jalammar.github.io/illustrated-transformer/

$$Attention\,(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

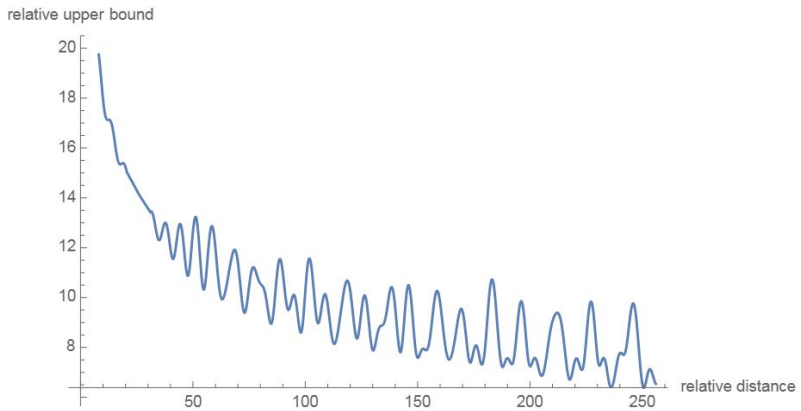
Merits of RoPE

- Computational efficient

$$\boldsymbol{R}^d_{\Theta,m}\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_{d-1} \\ x_d \end{pmatrix} \otimes \begin{pmatrix} \cos m\theta_1 \\ \cos m\theta_1 \\ \cos m\theta_2 \\ \cos m\theta_2 \\ \vdots \\ \cos m\theta_{d/2} \\ \cos m\theta_{d/2} \end{pmatrix} + \begin{pmatrix} -x_2 \\ x_1 \\ -x_4 \\ x_3 \\ \vdots \\ -x_d \\ x_{d-1} \end{pmatrix} \otimes \begin{pmatrix} \sin m\theta_1 \\ \sin m\theta_1 \\ \sin m\theta_2 \\ \sin m\theta_2 \\ \vdots \\ \sin m\theta_{d/2} \\ \sin m\theta_{d/2} \end{pmatrix}$$

$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

- Long-term decay

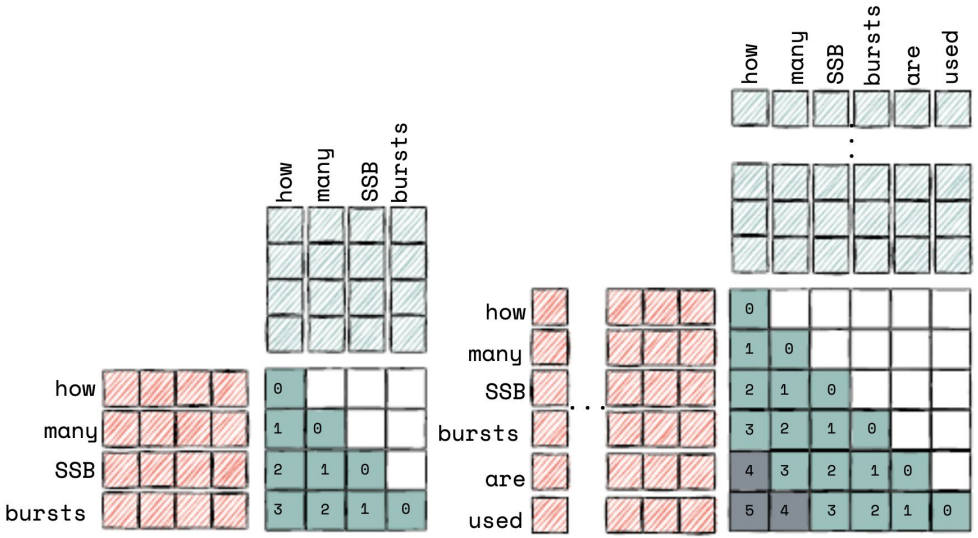


Su, Jianlin, et al. "Roformer: Enhanced transformer with rotary position embedding." *Neurocomputing* 568 (2024): 127063.

Positional O.O.D. issue

- The Positional Out-of-Distribution (O.O.D.) issue

  If sequences exceeding the pre-trained context window, self-attention and RoPE become distorted due to novel relative distances.

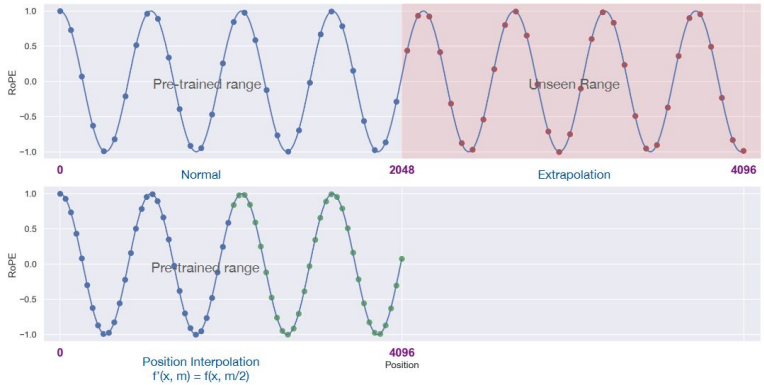


Alabbasi, Nouf, et al., *arXiv preprint arXiv:2411.02617* (2024).

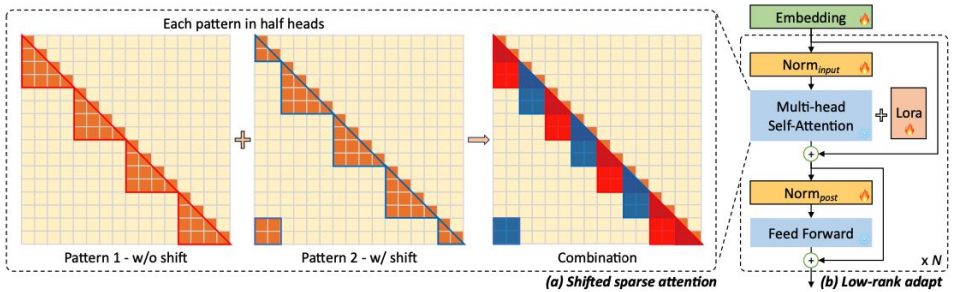
Fine-tuning methods

Fine-tuning is the process of further training a pre-trained LLM to extend the window size.

Position Interpolation (PI)



LongLoRA



- Overfitting: Fine-tuning with limited data leads to performance degradation and inefficiency in processing shorter sequences.

Then is a larger context window size always better?

Not Always Better

- **Compute & memory**

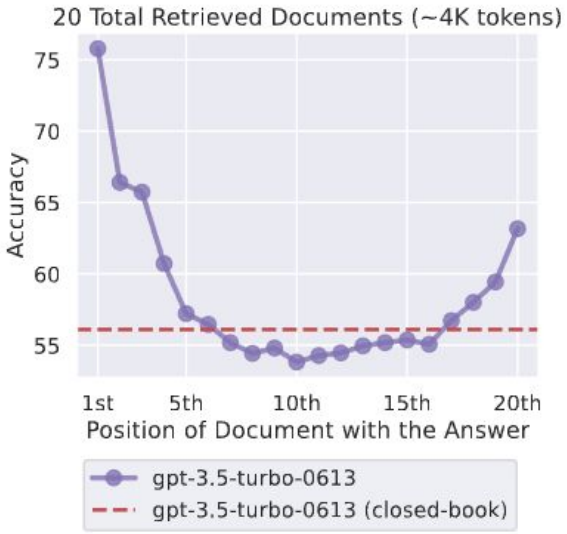
  Increase quadratic complexity($O(n^2)$).

- **Data scarcity**

  Truly long sequences are rare in pre-training corpora.

- **Lost-in-the-Middle**

  Models often fail to recall information placed in the middle of the prompt.



Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." Transactions of the Association for Computational Linguistics 12 (2024): 157-173.

02 Method

Method Overview

- LLM fails when encountered with positional out-of-distribution.

Hello	,	World	!	How	are	you	?
0	1	2	3	4	5	6	7

PE used in the pre-training

Hello	,	World	!	How	are	you	?	I	*'m	fine	,	thank	you	and	you
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

PE within the context window size

PE Out-of-distribution

Method Overview

- We can avoid avoid out-of-distribution by applying floor division to the positional embedding.

Hello	,	World	!	How	are	you	?
0	1	2	3	4	5	6	7

PE used in the pre-training

Hello	,	World	!	How	are	you	?	I	*'m	fine	,	thank	you	and	you
0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7

New PE with floor division

$$P_g = P \;//\; G_s,$$

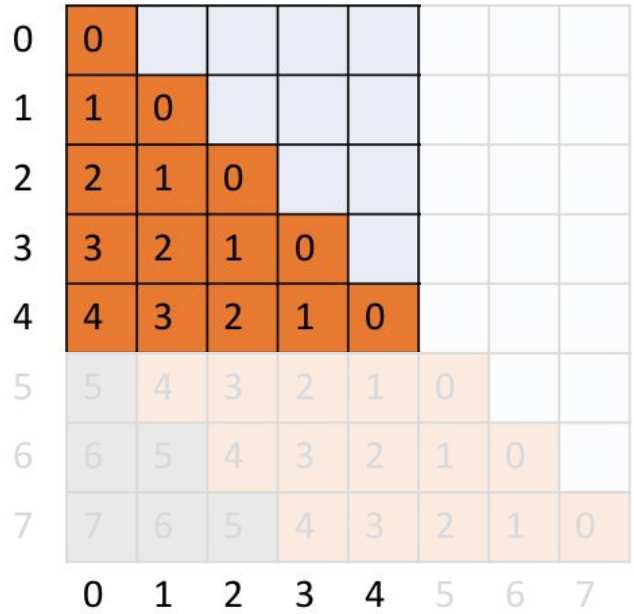
New PE — original PE — group size

Method Overview

- With RoPE, distance between query and key is encoded into the attention map.

Mapping to the attention matrix:

$$\langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n)\rangle_{\mathbb{R}} = g(\mathbf{x}_m, \mathbf{x}_n, m-n)$$



Pretraining context window size L=5

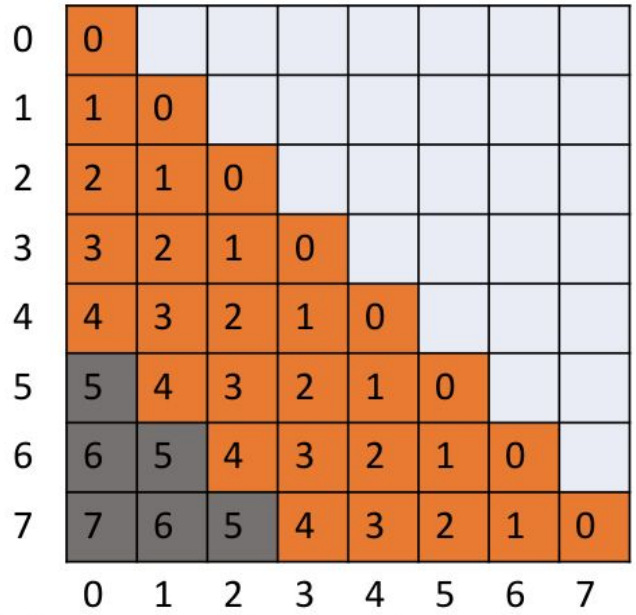
In-distribution Relative PE: [0, 4]

Method Overview

- With RoPE, distance between query and key is encoded into the attention map.

Mapping to the attention matrix:

$$\langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n)\rangle_{\mathbb{R}} = g(\mathbf{x}_m, \mathbf{x}_n, m-n)$$

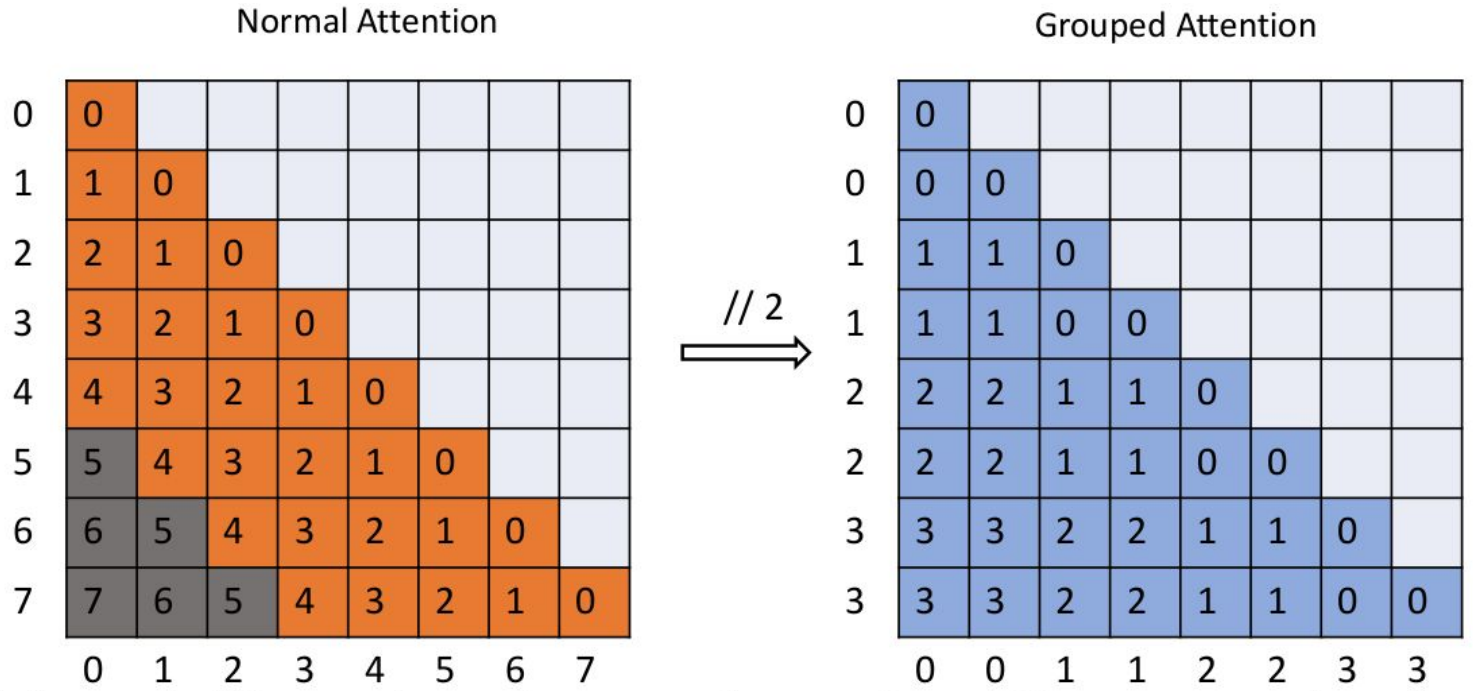


Pretraining context window size L=5

In-distribution Relative PE: [0, 4]
Out-of-distribution Relative PE: [5, 7]

Method Overview

- Grouped attention extends the context window by floor division.

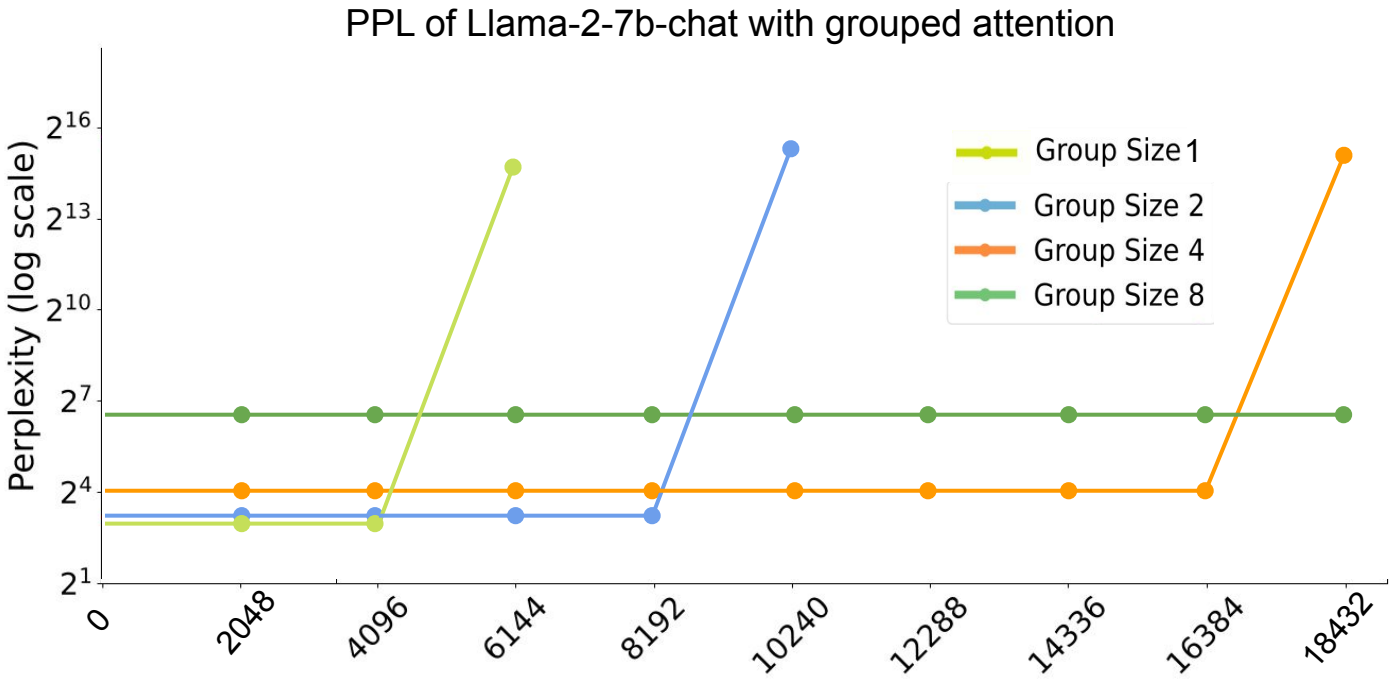


Method Overview

Tradeoffs in using Grouped Attention

(+) Extends the context window to avoid PPL explosion

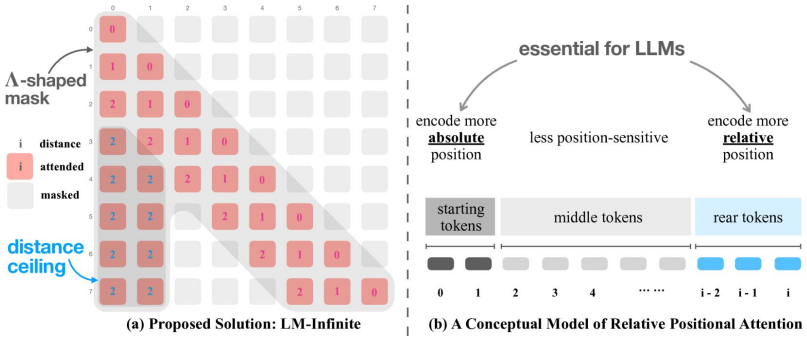
(–) Introduces some drop in performance with the group size



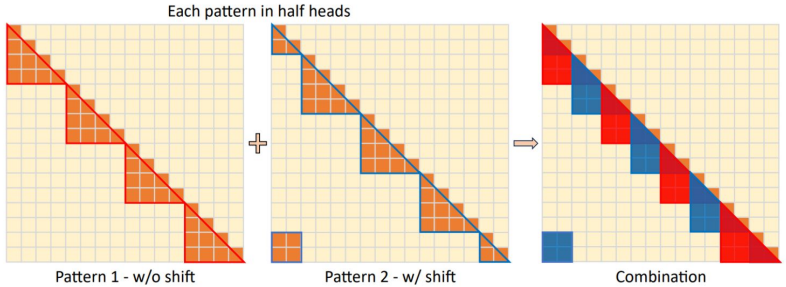
Method Overview

Relative positional encoding near the target token matters

- In the process of generating next tokens, **the immediate neighbors of a target token play a crucial role**.
- it is important to maintain the standard attention mechanism for tokens in close proximity to the target token.



Position sensitivity proposed in LM-Infinite

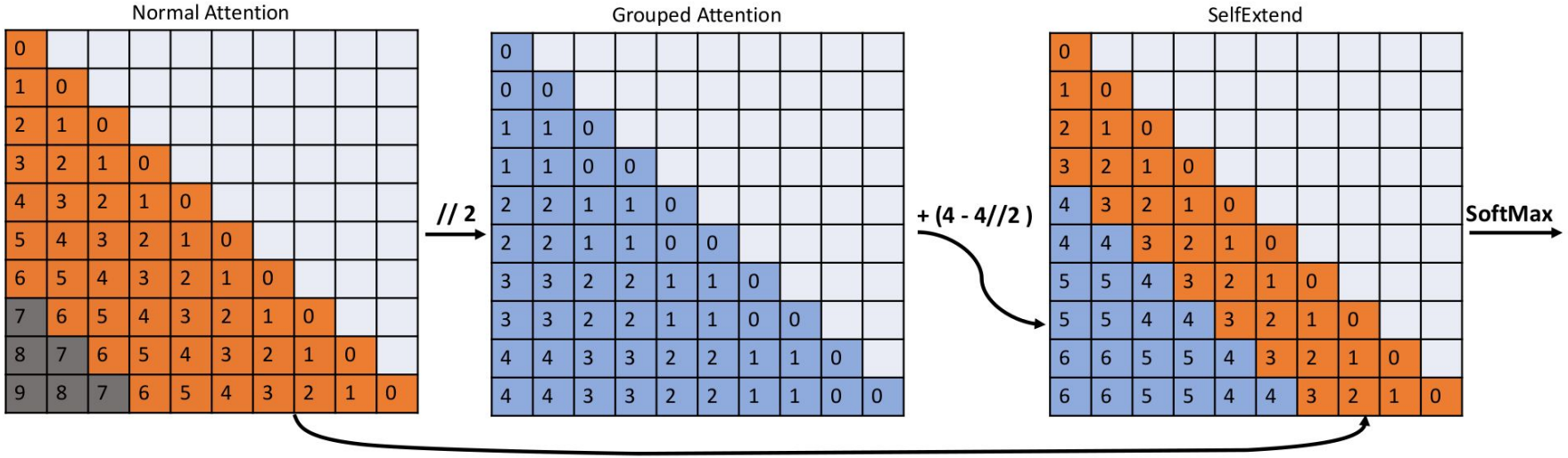


Shifted Sparse Attention proposed in LongLoRA

Han et al., “LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models”, NAACL 2024 Outstanding Paper
Chen et al, “LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models”, ICLR 2024 Oral

Method Overview

Q. How do we restore degraded language modeling ability caused by groped attention?
A. By re-introducing normal attention in the neighboring area.



Pretraining context window size $L = 7$

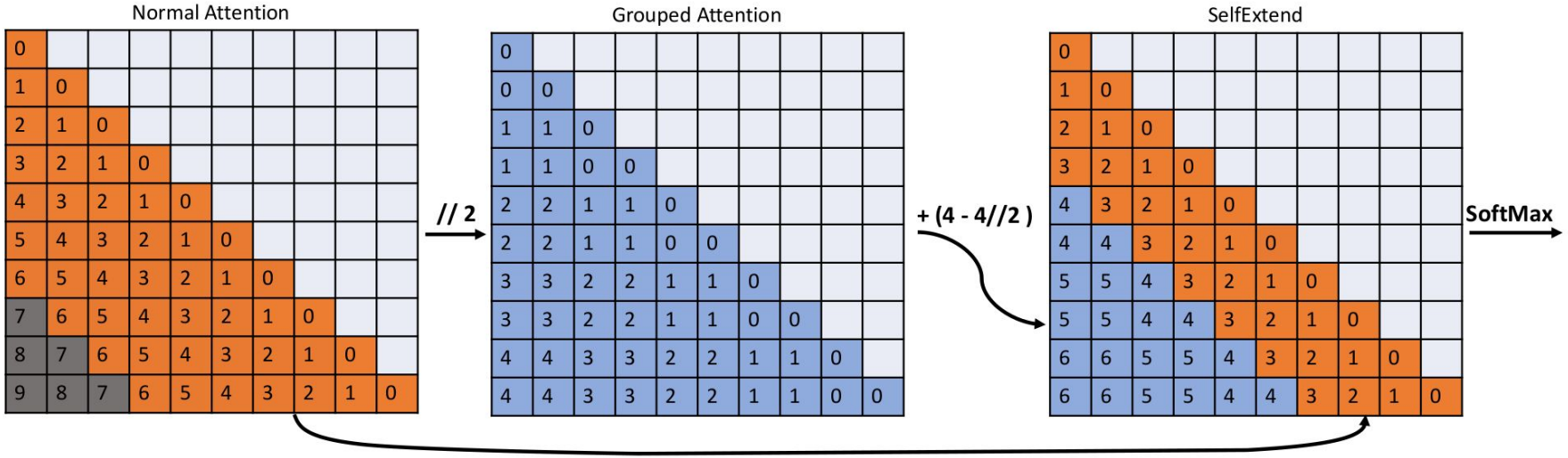
Neighbor window $w_n = 4$

Group size $G_s = 2$

Shift: $w_n - w_n//G_s$

Method Overview

Q. How do we restore degraded language modeling ability caused by groped attention?
A. By re-introducing normal attention in the neighboring area.



Normal attention:

$$\langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n)\rangle_{\mathbb{R}} = g(\mathbf{x}_m, \mathbf{x}_n, m-n)$$

Grouped attention:

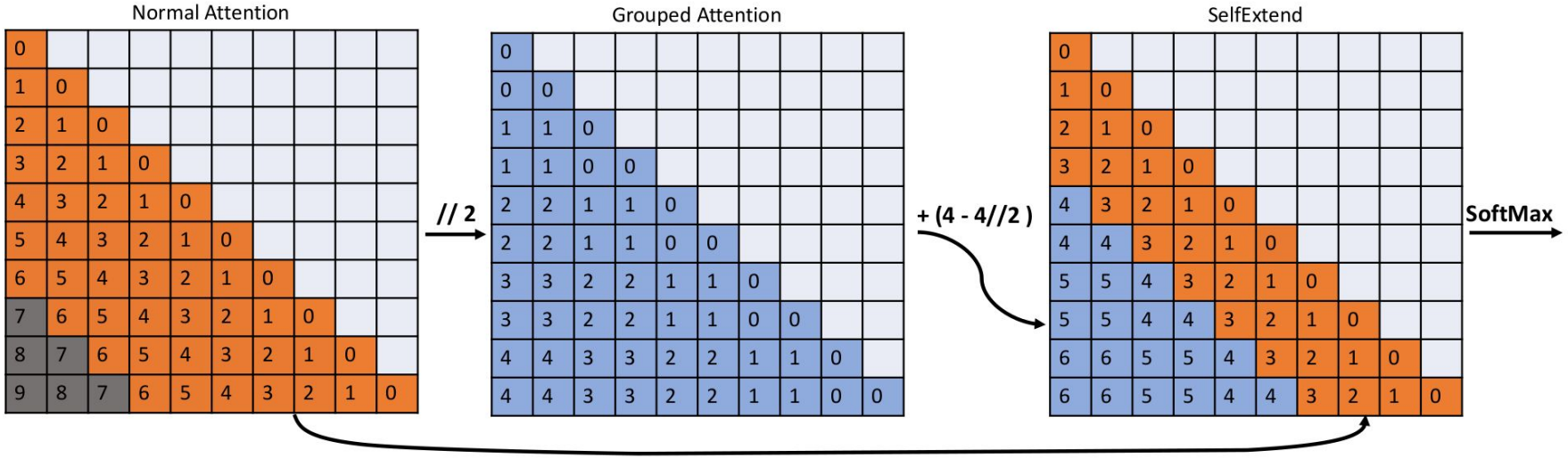
$$\langle f_q(\mathbf{x}_m, m//G_s), f_k(\mathbf{x}_n, n//G_s)\rangle_{\mathbb{R}} = g(\mathbf{x}_m, \mathbf{x}_n, m//G_s - n//G_s)$$

SelfExtend:

$$\begin{cases} g(\mathbf{x}_m, \mathbf{x}_n, m-n) & \text{if } m-n \le w_n \\ g(\mathbf{x}_m, \mathbf{x}_n, m//G_s - n//G_s + w_n - w_n//G_s) & \text{otherwise.} \end{cases}$$

Method Overview

Q. How do we restore degraded language modeling ability caused by groped attention?
A. By re-introducing normal attention in the neighboring area.



Normal attention:

maximum context window size: L

Grouped attention:

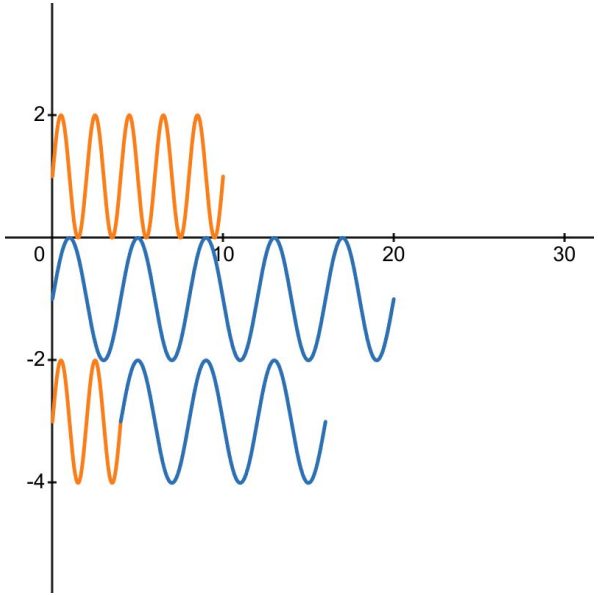
maximum context window size: $L \cdot G_s$

SelfExtend:

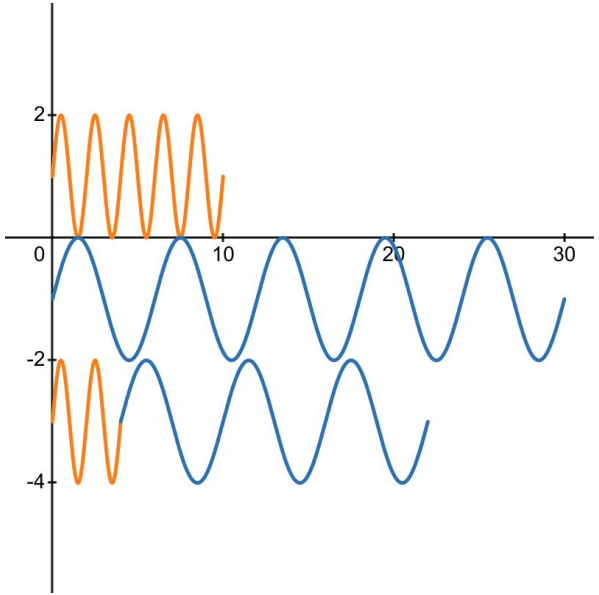
maximum context window size: $(L - w_n) \cdot G_s + w_n$

Method Overview

- SelfExend stretches positional embedding of the attention, while the neighboring region remains unchanged.



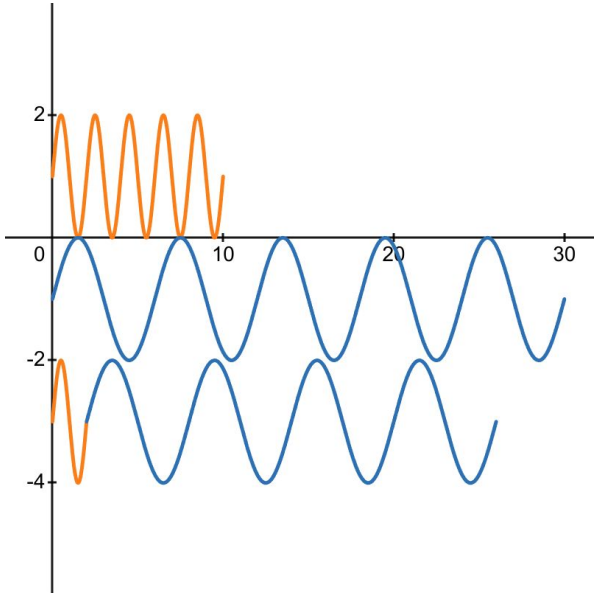
Pretraining context window size $L = 10$
Neighbor window $w_n = 4$
Group size $G_s = 2$



Pretraining context window size $L = 10$
Neighbor window $w_n = 4$
Group size $G_s = 3$

Method Overview

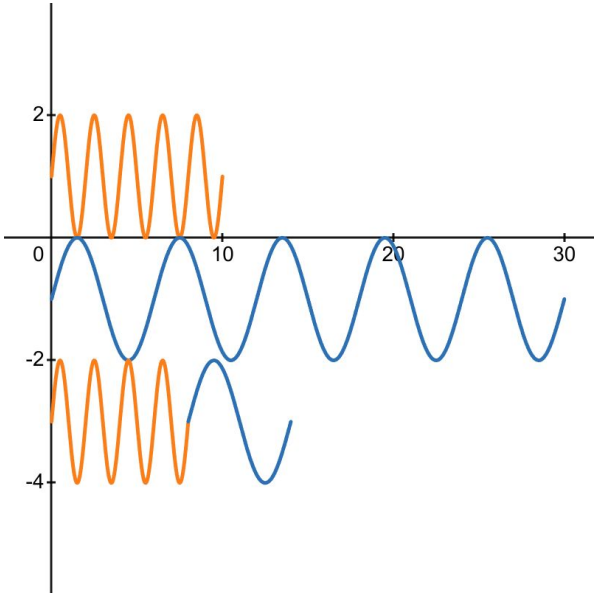
- SelfExend stretches positional embedding of the attention, while the neighboring region remains unchanged.



Pretraining context window size $L = 10$

Neighbor window $w_n = 2$

Group size $G_s = 3$



Pretraining context window size $L = 10$

Neighbor window $w_n = 8$

Group size $G_s = 3$

03 Experiment & Ablation

Experiment #1:
Language Modeling

- The most fundamental and the least requirement for LLMs

Setup.

- Model: Llama-2 and its families, Phi-2, Mistral, SOLAR
- Metric: Perplexity (PPL)
- Dataset: PG19

Experiment #1:
Language Modeling

Observation.

- SelfExtend: maintaining a ↓ PPL out of the pretraining context window

Model Name	Evaluation Context Window Size 4096	6144	8192	10240	12288	14336	16384
Llama-2-7b-chat	9.181	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
SelfExtend-Llama-2-7b-chat	8.885	8.828	9.220	8.956	9.217	9.413	9.274
Mistral-7b-instruct-0.1 w/ SWA	9.295	9.197	9.532	9.242	9.198	9.278	9.294
Mistral-7b-instruct-0.1 w/o SWA	9.295	9.205	10.20	55.35	$> 10^3$	$> 10^3$	$> 10^3$
SelfExtend-Mistral-7b-instruct-0.1	9.272	9.103	9.369	9.070	8.956	9.022	9.128

Experiment #1:
Language Modeling

Observation.

- SelfExtend: maintaining a ↓ PPL out of the pretraining context window

Model Name	Evaluation Context Window Size 4096	6144	8192	10240	12288	14336	16384
Llama-2-7b-chat	9.181	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
SelfExtend-Llama-2-7b-chat	8.885	8.828	9.220	8.956	9.217	9.413	9.274
Mistral-7b-instruct-0.1 w/ SWA	9.295	9.197	9.532	9.242	9.198	9.278	9.294
Mistral-7b-instruct-0.1 w/o SWA	9.295	9.205	10.20	55.35	$> 10^3$	$> 10^3$	$> 10^3$
SelfExtend-Mistral-7b-instruct-0.1	9.272	9.103	9.369	9.070	8.956	9.022	9.128

Experiment #1:

PPL as a Metric for Long Context Capabilities

Infinite.

Since RoPE uses bounded functions(cos, sin) to encode relative position differences, it can handle infinitely long contexts

- Evaluating on PG19, Llama-2-7b-chat w/ ‘infinite’ achieves ↓ PPL scores

Table 6. Perplexity on the PG19 dataset: For ‘Infinite’, we set three different local window sizes: 1024, 2048, and 4096. We have also included the results from Table 1 for comparison.

Model	Evaluation Context Window Size						
Name	**4096**	**6144**	**8192**	**10240**	**12288**	**14336**	**16384**
Llama-2-7b-chat	9.181	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
SelfExtend-Llama-2-7b-chat	8.885	8.828	9.220	8.956	9.217	9.413	9.274
1024-‘Infinite’–Llama-2-7b-chat	9.556	9.393	9.728	9.266	9.400	9.369	9.142
2048-‘Infinite’–Llama-2-7b-chat	9.288	9.045	9.478	8.993	9.128	9.105	8.872
4096-‘Infinite’–Llama-2-7b-chat	9.181	9.045	9.506	8.993	9.165	9.105	8.856
Mistral-7b-instruct-0.1 w/ SWA	9.295	9.197	9.532	9.242	9.198	9.278	9.294
Mistral-7b-instruct-0.1 w/o SWA	9.295	9.205	10.20	55.35	$> 10^3$	$> 10^3$	$> 10^3$
SelfExtend-Mistral-7b-instruct-0.1	9.272	9.103	9.369	9.070	8.956	9.022	9.128

Experiment #1:

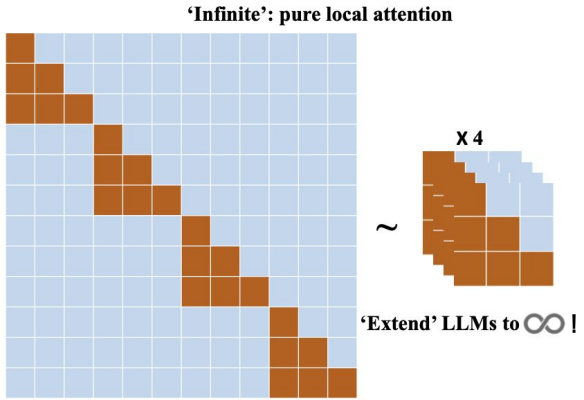
PPL as a Metric for Long Context Capabilities

- ‘Infinite’ divides long sequences into short segments

→ failing to fully address long-context reasoning

- ↓ PPL: accurately predicting the majority of tokens

→ not critical for understanding long contexts or answering questions



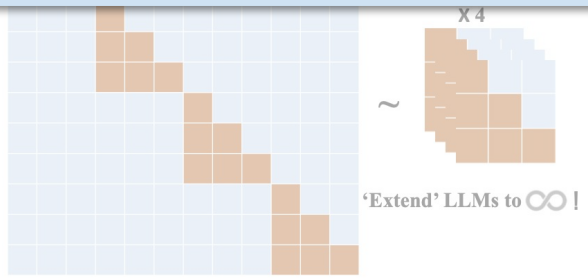
Experiment #1:

PPL as a Metric for Long Context Capabilities

- ‘Infinite’ divides long sequences into short segments

- While ↓ PPL is essential for a good model,
  it does Not guarantee strong long-context understanding

- Proper evaluation of long-context ability requires real tasks beyond PPL



Experiment #2:

Synthetic Long Context Task

Question.

Does low PPL guarantee good performance?

Setup.

- Model: Mistral (sliding window size: 4096)
- Task: Passkey Retrieval Task

Example:

Prompt: There is an important info hidden inside a lot of irrelevant text. Find it and memorize it. I will quiz you about the important information there ..... back again. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.The pass key is 60151. Remember it. 60151 is the pass key. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.The grass is green. The sky ..... What is the passkey?

Ground Truth: 60151

Experiment #2:

Synthetic Long Context Task

Observation.

- w/o fine-tuning, SelfExtend obtains 100% passkey retrieval accuracy
- Mistral w/ SWA can only access information within its sliding window

Conclusion.

↓ PPL score does not necessarily indicate proficiency in handling long context

Experiment #3:
Real-World Long Context Task

Motivation.
The previous two experiments fail to fully capture the long-context capabilities of LLMs

- Passkey retrieval task is overly straightforward

Setup.

- Model: Llama-2 and its families, Phi-2, Mistral, SOLAR
- Benchmarks: LongBench, L-Eval

Experiment #3:
Real-World Long Context Task

LongBench.

- Evaluates LLMs' long-context understanding (up to 16k tokens)
- Covers QA, summarization, and multi-hop reasoning tasks

L-Eval.

- Evaluates LLMs on retrieval, reasoning, and generation with long contexts (up to 32k tokens)
- Includes human-verified benchmarks across multiple domains

LongBench evaluates diverse long-context tasks, while L-Eval verifies accuracy and reliability

Experiment #3:

Real-World Long Context Task

LongBench. (Base models)

- Llama-2-7B. SelfExtend mostly outperforms baseline model
- Mistral-7B. SelfExtend improves its long context ability over the base model
- SoLAR-10.7B, Phi-2. SelfExtend obtain substantial performance improvements

		Single-Document QA			Multi-Document QA			Summarization			Few-shot Learning			Synthetic		Code	
	LLMs[a]	NarrativeQA	Qasper	MultiField-en	HotpotQA	2WikiMQA	Musique	GovReport	QMSum	MultiNews	TREC	TriviaQA	SAMSum	PassageCount	PassageRe	Lcc	RepoBench-P
SelfExtend	Llama-2-7B-chat-4k*	18.7	19.2	**36.8**	25.4	32.8	9.4	27.3	20.8	25.8	61.5	77.8	40.7	2.1	**9.8**	52.4	43.8
	SE-Llama-2-7B-chat-16k+	**21.69**	25.02	35.21	34.34	30.24	14.13	27.32	**21.35**	25.78	**69.50**	**81.99**	40.96	**5.66**	5.83	**60.60**	**54.33**
	SE-Llama-2-7B-chat-25k+	21.37	**26.68**	34.63	**35.47**	**30.46**	**15.51**	**27.51**	21.30	**25.87**	68.50	78.79	**41.29**	3.90	3.50	59.69	53.83
	Mistral-7B-ins-0.1-16k w/ SWA+	19.40	34.53	37.06	42.29	32.49	14.87	27.38	22.75	26.82	65.00	**87.77**	42.34	**1.41**	28.50	**57.28**	53.44
	Mistral-7B-ins-0.1-8k w/o SWA+	20.46	35.36	39.39	34.81	29.91	11.21	24.70	21.67	26.67	68.00	86.66	41.28	0.18	24.00	56.94	**55.85**
	SE-Mistral-7B-ins-0.1-16k+[b]	**23.56**	**39.33**	**49.50**	**45.28**	**34.92**	**23.14**	**30.71**	**24.87**	**26.83**	**69.50**	86.47	**44.28**	1.18	**29.50**	55.32	53.44
	Phi-2-2k+	4.46	7.01	19.98	**9.43**	8.55	**4.62**	25.64	14.32	**24.03**	50.50	74.55	**1.71**	**2.83**	**4.17**	**58.96**	54.14
	SE-Phi-2-8k+	**12.04**	**12.10**	**20.15**	8.22	**9.68**	3.89	**27.90**	**14.58**	22.13	**61.00**	**82.82**	1.40	2.37	2.83	57.87	**56.42**
	SOLAR-10.7B-ins-4k+	16.50	24.06	46.76	44.03	**36.05**	22.76	**31.39**	19.81	**26.36**	70.00	87.91	42.49	**4.5**	26.5	41.04	54.36
	SE-SOLAR-10.7B-ins-16k+	**22.63**	**32.49**	**47.88**	**46.19**	34.32	**27.88**	30.75	**22.10**	25.62	**74.50**	**89.04**	**42.79**	4.0	**28.0**	**53.73**	**56.47**

Experiment #3:

Real-World Long Context Task

LongBench. (Base models)

- Llama-2-7B. SelfExtend mostly outperforms baseline model
- Mistral-7B. SelfExtend improves its long context ability over the base model
- SoLAR-10.7B, Phi-2. SelfExtend obtain substantial performance improvements

Lcc. reliance on local codes, shorter dataset lengths → Higher base model performance

		Single-Document QA			Multi-Document QA			Summarization			Few-shot Learning			Synthetic		Code	
	LLMs[a]	NarrativeQA	Qasper	MultiField-en	HotpotQA	2WikiMQA	Musique	GovReport	QMSum	MultiNews	TREC	TriviaQA	SAMSum	PassageCount	PassageRe	Lcc	RepoBench-P
SelfExtend	Llama-2-7B-chat-4k*	18.7	19.2	**36.8**	25.4	32.8	9.4	27.3	20.8	25.8	61.5	77.8	40.7	2.1	**9.8**	52.4	43.8
	SE-Llama-2-7B-chat-16k+	**21.69**	25.02	35.21	34.34	30.24	14.13	27.32	**21.35**	25.78	**69.50**	**81.99**	40.96	**5.66**	5.83	**60.60**	**54.33**
	SE-Llama-2-7B-chat-25k+	21.37	**26.68**	34.63	**35.47**	**30.46**	**15.51**	**27.51**	21.30	**25.87**	68.50	78.79	**41.29**	3.90	3.50	59.69	53.83
	Mistral-7B-ins-0.1-16k w/ SWA+	19.40	34.53	37.06	42.29	32.49	14.87	27.38	22.75	26.82	65.00	**87.77**	42.34	**1.41**	28.50	**57.28**	53.44
	Mistral-7B-ins-0.1-8k w/o SWA+	20.46	35.36	39.39	34.81	29.91	11.21	24.70	21.67	26.67	68.00	86.66	41.28	0.18	24.00	56.94	**55.85**
	SE-Mistral-7B-ins-0.1-16k+[b]	**23.56**	**39.33**	**49.50**	**45.28**	**34.92**	**23.14**	**30.71**	**24.87**	**26.83**	**69.50**	86.47	**44.28**	1.18	**29.50**	55.32	53.44
	Phi-2-2k+	4.46	7.01	19.98	**9.43**	8.55	**4.62**	25.64	14.32	**24.03**	50.50	74.55	**1.71**	**2.83**	**4.17**	**58.96**	54.14
	SE-Phi-2-8k+	**12.04**	**12.10**	**20.15**	8.22	**9.68**	3.89	**27.90**	**14.58**	22.13	**61.00**	**82.82**	1.40	2.37	2.83	57.87	**56.42**
	SOLAR-10.7B-ins-4k+	16.50	24.06	46.76	44.03	**36.05**	22.76	**31.39**	19.81	**26.36**	70.00	87.91	42.49	**4.5**	26.5	41.04	54.36
	SE-SOLAR-10.7B-ins-16k+	**22.63**	**32.49**	**47.88**	**46.19**	34.32	**27.88**	30.75	**22.10**	25.62	**74.50**	**89.04**	**42.79**	4.0	**28.0**	**53.73**	**56.47**

Experiment #3:

Real-World Long Context Task

LongBench. (Fine-Tuned Models)

- Llama-2-7B. mostly outperforms models
  - 25k variant < 16k due to the trade-off between larger context and positional precision
- Mistral-7B. The fine-tuned variant MistralLite achieves the best performance on most datasets (many of these datasets were includes in MistralLite's fine-tuning data)

		Single-Document QA			Multi-Document QA			Summarization			Few-shot Learning			Synthetic		Code	
	LLMs[a]	NarrativeQA	Qasper	MultiField-en	HotpotQA	2WikiMQA	Musique	GovReport	QMSum	MultiNews	TREC	TriviaQA	SAMSum	PassageCount	PassageRe	Lcc	RepoBench-P
Other Methods	LongChat1.5-7B-32k*	16.9	27.7	41.4	31.5	20.6	9.7	30.8	22.7	26.4	63.5	82.3	34.2	1.0	30.5	53.0	55.3
	together/llama-2-7b-32k+	15.65	10.49	33.43	12.36	12.53	6.19	29.28	17.18	22.12	71.0	87.79	43.78	1.0	23.0	63.79	61.77
	CLEX-7B-16k*	18.05	23.68	44.62	28.44	19.53	9.15	32.52	22.9	25.55	68	84.92	42.82	0	11.5	59.01	56.87
	CodeLLaMA-7B-16k*	22.93	30.69	43.37	33.05	27.93	14.2	28.43	24.18	26.84	70	84.97	43.43	2	13.5	64.35	55.87
	SE-Llama-2-7B-chat-16k+	21.69	25.02	35.21	34.34	30.24	14.13	27.32	21.35	25.78	69.50	81.99	40.96	5.66	5.83	60.60	54.33
	SE-Llama-2-7B-chat-25k+	21.37	26.68	34.63	35.47	30.46	15.51	27.51	21.30	25.87	68.50	78.79	41.29	3.90	3.50	59.69	53.83
	Vicuna1.5-7B-16k*	19.4	26.1	38.5	25.3	20.8	9.8	27.9	22.8	27.2	71.5	86.2	40.8	6.5	4.5	51.0	43.5
	SE-Vicuna1.5-7B-16k+	21.88	35.16	42.00	31.14	22.51	13.33	28.47	22.24	26.70	69.50	86.31	40.54	3.56	7.50	60.16	44.07
	SE-Vicuna1.5-7B-25k+	22.46	34.42	42.58	30.95	24.33	12.72	27.75	22.26	27.21	72.00	84.02	40.38	3.01	7.00	58.86	43.86
	MistralLite-16k+	32.12	47.02	44.95	58.5	47.24	31.32	33.22	26.8	24.58	71.5	90.63	37.36	3	54.5	66.27	65.29
	SE-Mistral-7B-ins-0.1-16k+	23.85	37.75	46.93	45.35	34.54	23.28	30.45	23.58	26.94	69.50	85.72	43.88	0.59	28.50	54.92	53.44

Experiment #3:

Real-World Long Context Task

L-Eval.

- SelfExtend achieves superior performance on nearly all datasets

Model	Tokens	Coursera	GSM	QuALITY	TOEFL	CodeU	SFiction	Avg.
Llama2-7b-chat	4k	29.21	19.00	37.62	51.67	1.11	60.15	33.12
Longchat1.5-7b-32k	32k	32.99	18.00	37.62	39.77	**3.33**	57.02	31.45
Llama2-7b-NTK	16k	32.71	19.00	33.16	52.78	0.00	**64.84**	33.74
SE-Llama2-7B-chat+	16k	**35.76**	**25.00**	**41.09**	**55.39**	1.11	57.81	**36.02**
Vicuna1.5-7b-16k	16k	**38.66**	19.00	39.60	**55.39**	**5.55**	60.15	36.39
SE-Vicuna1.5-7B+	16k	37.21	**21.00**	**41.58**	**55.39**	3.33	**63.28**	**36.96**
Llama2-13b-chat	4k	35.75	39.00	**42.57**	60.96	1.11	54.68	39.01
Llama2-13b-NTK	16k	36.48	11.00	35.64	54.64	1.11	**63.28**	33.69
Llama2-13b-NTK(Dyn)	16k	30.08	**43.00**	41.58	64.31	1.11	35.15	35.87
SE-Llama2-13B-chat+	16k	**38.95**	42.00	41.09	**66.17**	1.11	**63.28**	**42.10**
Mistral-7b-ins-0.1 w/ SWA+	16k	44.77	44.00	46.53	60.59	2.22	**64.06**	43.70
Mistral-7b-ins-0.1 w/o SWA+	8k	43.60	49.00	45.05	60.59	**4.44**	60.94	43.94
MistralLite+	16k	29.23	32.00	46.04	17.47	3.33	14.06	23.69
SE-Mistral-7b-ins-0.1+	16k	**45.20**	**51.00**	**48.02**	**64.68**	3.33	59.38	**45.27**
Phi-2+	2k	38.37	64.00	**42.08**	55.76	3.33	**52.34**	42.64
SE-Phi-2+	8k	**42.44**	**65.00**	41.08	**62.83**	**4.44**	**52.34**	**44.69**
SOLAR-10.7b-Instruct-v1.0+	4k	48.84	**72.00**	59.90	77.32	**4.44**	69.53	55.34
SE-SOLAR-10.7b-v1.0+	16k	**50.44**	**72.00**	**70.30**	**79.18**	**4.44**	**73.44**	**58.30**

Conclusion.

SelfExtend achieves comparable or better performance, compared to methods that requires further fine-tuning

Experiment #4:

Short Context Task

Motivation.
Previous fine-tuning based methods usually undergo performance degradation on short-context tasks.

Setup.

- Model: Llama-2 and its families, Mistral, Phi-2
  - Llama-2: group size 16, neighbor window 1024
  - Mistral: group size 6, neighbor window 1024
  - Phi-2: group size 12, neighbor window 512

- Tasks:
  - 25-shot ARC Challenge
  - 10-shot HellaSwag
  - 5-shot MMLU
  - 0-shot TruthfulQA
  - 5-shot GSM8K

Experiment #4:

Short Context Task

Observation.

SelfExtend can maintain the performance of the short-context tasks

Conclusion.

SelfExtend operates only at inference, allowing LLMs to retain their original behavior for short contexts without any fine-tuning

Size	Name	ARC-c	Hellaswag	MMLU	TruthfulQA	GSM8k
7B	Llama-2	**52.99**	**78.66**	46.58	**38.97**	**14.94**
7B	SE-Llama 2	**52.99**	78.65	**46.68**	**38.97**	14.71
7B	Llama-2-chat	**52.73**	**78.49**	**48.20**	45.32	18.73
7B	SE-Llama-2-chat-16k	**52.73**	**78.49**	48.09	**45.33**	**18.88**
7B	Mistral-instruct-v0.1	54.35	**75.72**	55.57	**55.89**	30.93
7B	SE-Mistral-instruct-v0.1	**54.44**	75.71	**55.59**	**55.89**	**31.39**
2.7B	Phi-2	**61.17**	75.13	58.20	**44.54**	55.11
2.7B	SE-Phi-2	61.00	**75.20**	**58.29**	**44.54**	**55.42**

Ablation.

Trade-offs: Group Size and Neighbor Window

Group size.

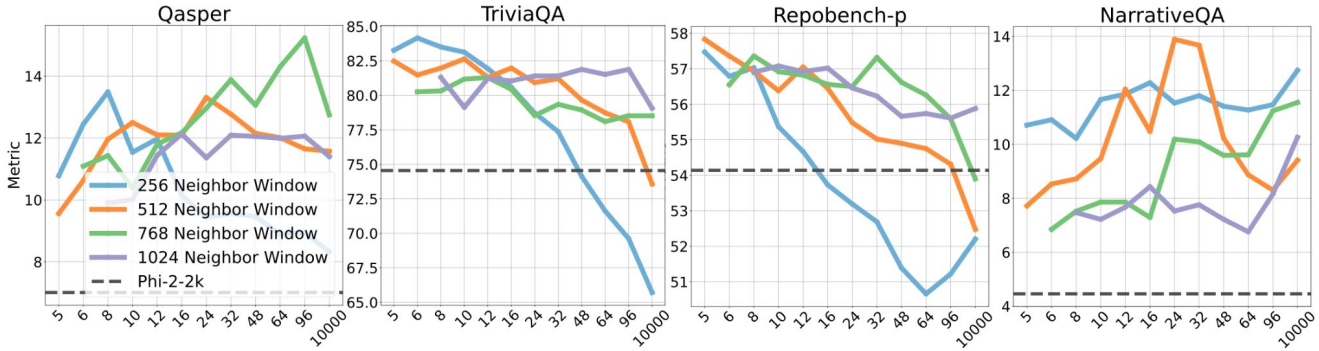
↑ ~ position information becomes more coarse. performance ↓.

↓ ~ utilizing larger position embeddings which are less trained compared to smaller one

Neighbor window size.

↑ ~ ↑ precise information about neighbor tokens

: ↑ group size is required ~ becoming coarse



04 Conclusion & Limitations

Conclusion.

- LLMs have the inherent ability to handle long sequences
- SelfExtend: mapping unseen relative positions into those seen during pretraining
- Without any tuning or further training,
  SelfExtend can effectively improve LLMs' long context performance

Limitation.

- ↑ computation cost with naive implementations
  : performing extra attention across all query-key pairs
  - But, with optimizations like blocked kernels (e.g., Flash Attention) → linear
  - For long input sequences, the marginal cost is small enough to be ignored

- The current observation may not be applicable to all models
  : achieving optimal performance may still require moderate hyperparameter tuning

Thank You !