LONG CONTEXT LANGUAGE MODELS

Allen Huo, Caroline Xu, Evan Zhang

OVERVIEW

- LLMs primarily perform tasks via prompting, and task specification and data is mostly processed as textual input ("context")
- These input can be thousands and thousands of tokens long (lengthy documents, documents from search engines, conversation history)
- This raises three key questions: Do they use long context effectively? Can we extend context efficiently? How can retrieval improve long context ability?
- Lost in the Middle: Long Context LLM performance depends on document
- LLM Maybe LongLM: Modifying attention can extend the context window without finetuning
- Retrieval meets Long Context LLM: Retrieval + Long Context gives the best results

Lost in the Middle: How Language Models Use Long Contexts

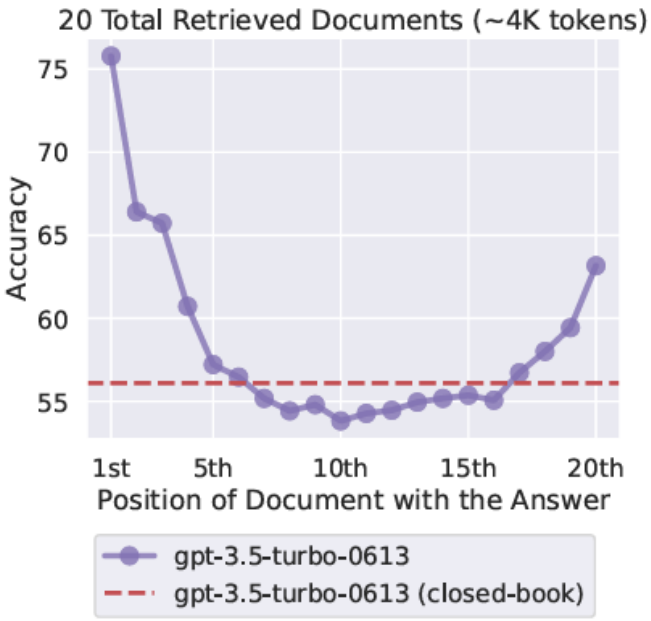
Nelson F. Liu[1*] **Kevin Lin**[2] **John Hewitt**[1] **Ashwin Paranjape**[3]
Michele Bevilacqua[3] **Fabio Petroni**[3] **Percy Liang**[1]
[1]Stanford University [2]University of California, Berkeley [3]Samaya AI
nfliu@cs.stanford.edu

MOTIVATION

- When input contexts contain thousands of tokens (long documents or model augmented with external information), LMs need to operate over very long sequences
- Current solution: Transformers
  - Memory and compute increases quadratically in sequence length
  - LMs are trained on small context windows (512-2048 tokens)
  - Recent developments made in hardware/algorithms have enabled larger context windows, but it's unclear on how these LMs specifically use their input contexts on downstream tasks

BACKGROUND

- Models (alongside humans) have primacy and recency bias
- Current research indicates: providing LM with more context may help with performance in downstream task, but also increases the amount of content the model reasons over, potentially decreasing accuracy
- **To claim a language model can robustly use information within long input contexts, it is necessary to show that its performance is minimally affected by the position of the relevant information in the input context**
- It's currently unclear how extended-context LMs make use of their input context



EXPERIMENT – MULTI-DOCUMENT Q/A

- How do LMs use their input context?
- Data – NaturalQuestions-Open (queries from Google and answers from Wikipedia)
  - 2655 Queries – document with the answer and k-1 distractor documents from a retrieval system related to the query, but don't contain any of the answers from the dataset

Input Context

```
Write a high-quality answer for the given question using only the provided search
results (some of which might be irrelevant).

Document [1](Title: Asian Americans in science and technology) Prize in physics for
discovery of the subatomic particle J/ψ. Subrahmanyan Chandrasekhar shared...
Document [2](Title: List of Nobel laureates in Physics) The first Nobel Prize in
Physics was awarded in 1901 to Wilhelm Conrad Röntgen, of Germany, who received...
Document [3](Title: Scientist) and pursued through a unique method, was essentially
in place. Ramón y Cajal won the Nobel Prize in 1906 for his remarkable...

Question: who got the first nobel prize in physics
Answer:
```

Desired Answer

```
Wilhelm Conrad Röntgen
```

EXPERIMENT – MULTI-DOCUMENT Q/A

- Changing order of correct document in input context
- Changing input context length

Input Context

```
Write a high-quality answer for the given question
using only the provided search results (some of
which might be irrelevant).

Document [1](Title: List of Nobel laureates in
Physics) ...
Document [2](Title: Asian Americans in science and
technology) ...
Document [3](Title: Scientist) ...

Question: who got the first nobel prize in physics
Answer:
```

Desired Answer

```
Wilhelm Conrad Röntgen
```

Input Context

```
Write a high-quality answer for the given question
using only the provided search results (some of
which might be irrelevant).

Document [1](Title: Asian Americans in science and
technology) ...
Document [2](Title: List of Nobel laureates in
Physics) ...
Document [3](Title: Scientist) ...
Document [4](Title: Norwegian Americans) ...
Document [5](Title: Maria Goeppert Mayer) ...

Question: who got the first nobel prize in physics
Answer:
```

Desired Answer

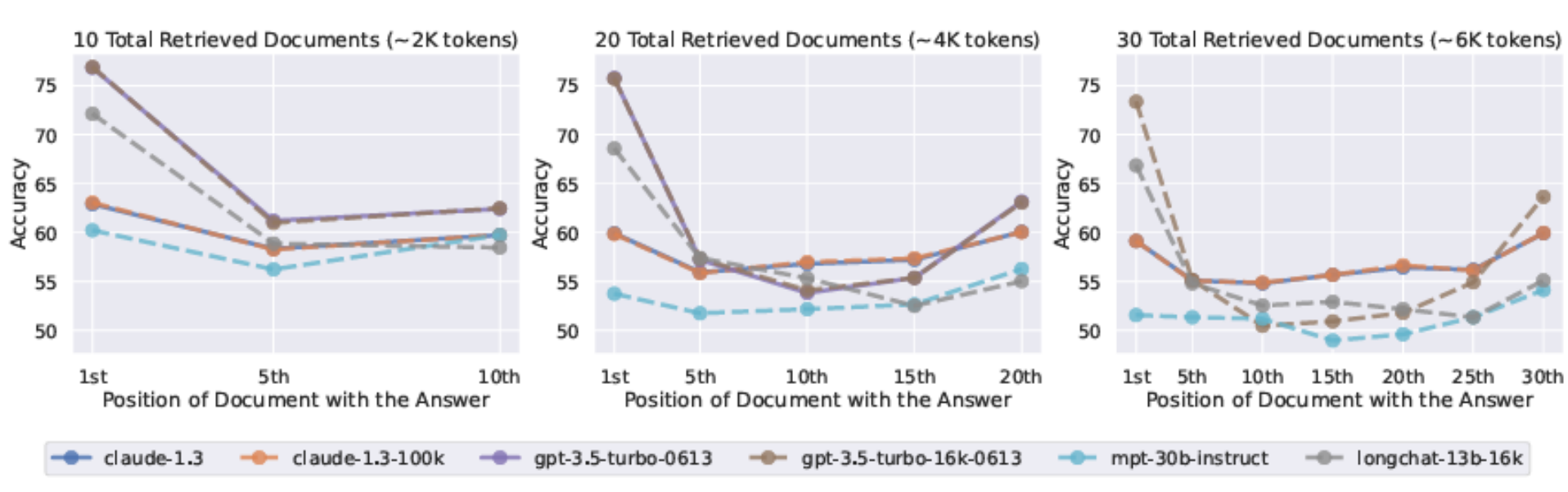
```
Wilhelm Conrad Röntgen
```

EXPERIMENT – MULTI-DOCUMENT Q/A

- Analyzed with open and closed LMs
  - MPT-30B-Instruct: 8192 tokens
  - LongChat-13B-16K
  - GPT-3.5-Turbo: 4K tokens
  - GPT-3.5-Turbo-16K
  - Claude-1.3: 8K Tokens
  - Claude-1.3-100K

Model	Closed-Book	Oracle
LongChat-13B (16K)	35.0%	83.4%
MPT-30B-Instruct	31.5%	81.9%
GPT-3.5-Turbo	56.1%	88.3%
GPT-3.5-Turbo (16K)	56.0%	88.6%
Claude-1.3	48.3%	76.1%
Claude-1.3 (100K)	48.2%	76.4%

RESULTS



- Accuracy is highest when relevant information occurs at the beginning/end of its input context
  - U-shaped curve indicates higher accuracy when using information at the beginning (primacy bias) and very end (recency bias) of the input context
- Extended-context models are not necessarily better at using input context
  - When input context fits in the context window of both models, performance is nearly identical

EXPERIMENT – KEY/VALUE RETRIEVAL

- How well can LMs retrieve from input context?
- Inputs – String-serialized JSON object with k key-value pairs where the keys and values are randomly generated 128 bit UUIDs; Key within the JSON object
- Goal: return the value associated with the specified key
  - Removes as much natural language semantics as possible to remove confounding variables
- Edit input context length by adding or removing random keys, thus changing the number of distractor key-value pairs

Input Context

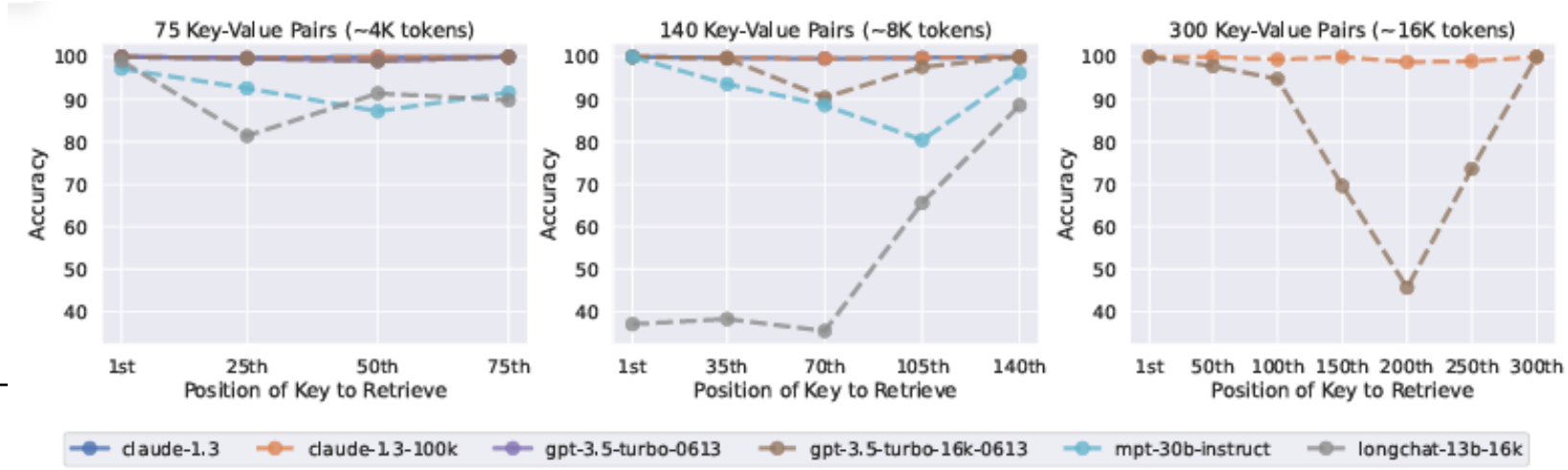
```
Extract the value corresponding to the specified key in the JSON object below.

JSON data:
{"2a8d601d-1d69-4e64-9f90-8ad825a74195": "bb3ba2a5-7de8-434b-a86e-a88bb9fa7289",
 "a54e2eed-e625-4570-9f74-3624e77d6684": "d1ff29be-4e2a-4208-a182-0cea716be3d4",
 "9f4a92b9-5f69-4725-ba1e-403f08dea695": "703a7ce5-f17f-4e6d-b895-5836ba5ec71c",
 "52a9c80c-da51-4fc9-bf70-4a4901bc2ac3": "b2f8ea3d-4b1b-49e0-a141-b9823991ebeb",
 "f4eb1c53-af0a-4dc4-a3a5-c2d50851a178": "d733b0d2-6af3-44e1-8592-e5637fdb76fb"}

Key: "9f4a92b9-5f69-4725-ba1e-403f08dea695"
Corresponding value:
```

Desired Output

```
703a7ce5-f17f-4e6d-b895-5836ba5ec71c
```

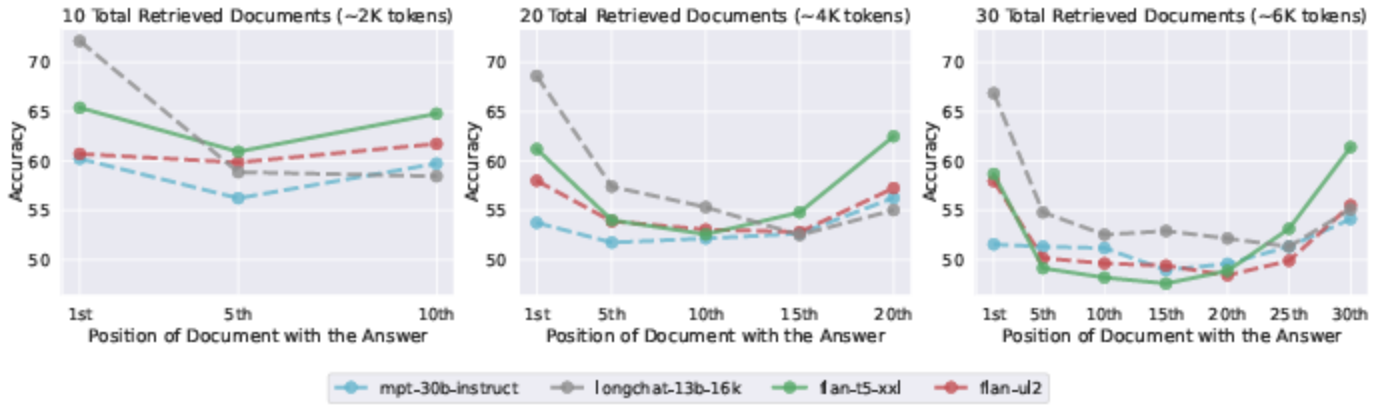
RESULTS

- Experiment with input contexts containing 75, 140, 300 key-value pairs (500 examples each), with the same models from the previous experiment
- Claude-1.3 and its counterpart do nearly perfect on all examples
- GPT-3.5-Turbo and its counterpart, and MPT-30B-Instruct have lowest performance when accessing key-value pairs in the middle of input context
    - When relevant information is placed at the start of the input context, LongChat-13B-16K tends to generate code to retrieve the key, instead of outputting the value directly



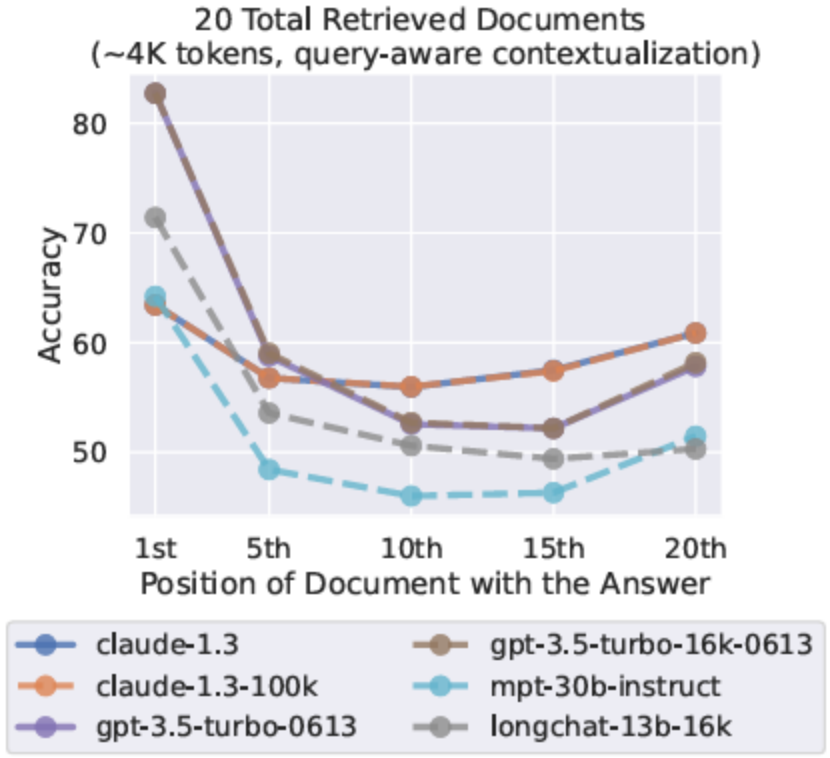
MODEL ARCHITECTURE

- Decoder-only (open models) vs. Encoder-decoder LMs
- Flan-UL2 and Flan-T5-XXL (encoder-decoder; 2048 tokens context window) - use relative positional embeddings, so can extrapolate beyond maximum context legnths (sequences of up to 8K tokens)
    - Longer input contexts (than initially trained on) results in a greater performance degradation when relevant information is in the middle of the input context



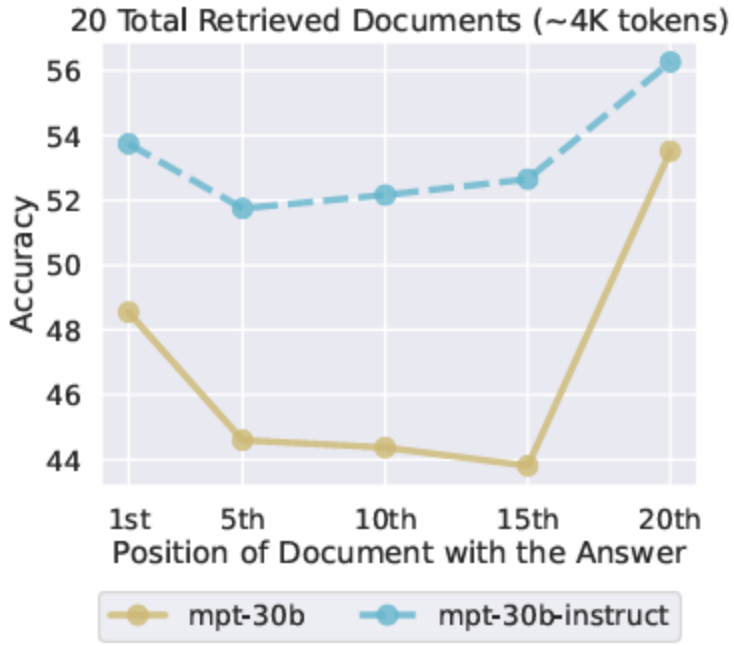
QUERY-AWARE CONTEXTUALIZATION

- Previous experiments place query after the processed data, so decoder-only models cannot attend to query tokens when contextualizing documents
  - Can we improve decoder-only models by placing query before and after the data?
- Query-aware contextualization significantly improves key-value retrieval (near perfect performance, without context., worst case is 45%), but performance did not improve much for multi-document Q/A



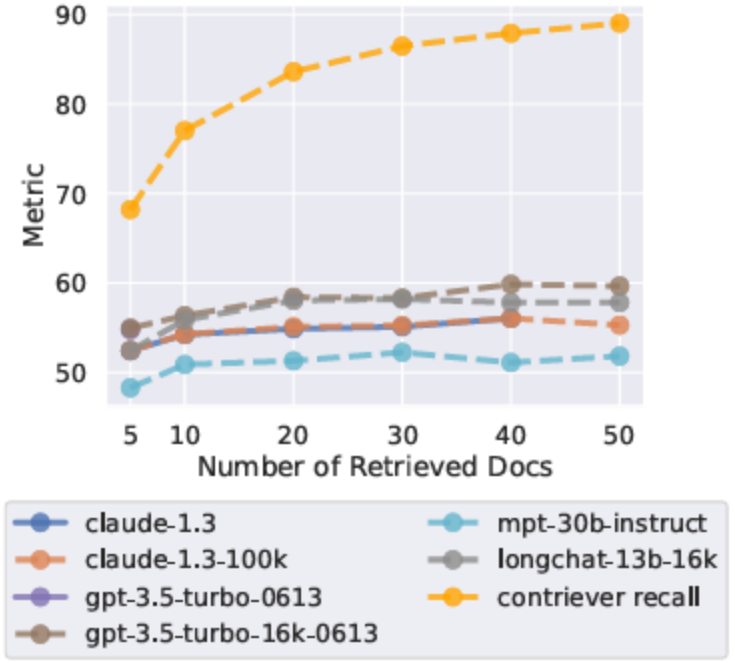
INSTRUCTION FINE-TUNING

- Models are all instruction finetuned – task specification is commonly placed at the beginning of input context -> Potentially more weight from LMs onto the start of the input context
- Performance trends are relatively similar between finetuned vs. not finetuned model
- Instruction finetuning slightly reduces worst case performance disparity from best to worst case from 10% to 4%



TRADEOFF WITH LONGER INPUT CONTEXT

- Providing LM with more information may help it perform downstream task, but also increases the amount of content the model must reason over, potentially decreasing accuracy
- Test on retriever recall and reader accuracy, with retrieval system that takes in an input query and returns k documents from Wikipedia with highest relevance score
- Reader model performance saturates long before retriever performance



CONTRIBUTION

- Language model performance degrades significantly when changing the position of relevant information -> Models struggle to robustly access information in long input contexts
  - More context does not necessarily mean better performance
- Performance is lowest when models use information in the middle of long input context
  - Not all parts of the context are usable memory
- Investigation of the role of model architecture, query-aware contextualization, and instruction fine-tuning

LIMITATIONS

- Only focused on two specific tasks that don't cover all real-world uses: Multi-document Q/A and synthetic key-value retrieval
- Main finding was that this U-shaped performance (primacy/recency bias) exists, but doesn't explain why (in terms of model architecture) this occurs
  - With closed models, this would be hard to determine
- Encoder-decoder models are more robust only within their training context length – this paper does not provide a solution

FUTURE DIRECTION

- Experiment on a more diverse set of tasks
- Precise knowledge access on long contexts (difficult to get away with quadratic sentence length complexity)
- Effective reranking of retrieved documents or ranked list truncation for improving how language-model-based readers use retrieved context

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]

OVERVIEW & MOTIVATION

- LLMs cannot generalize to long contexts/have trouble processing long input sequences during inference
  - Once the length of the input texts exceeds the pretraining context window, LLM performance suffers
- Current approach: fine-tuning
- LLMs have the inherent ability to handle long contexts

POSITION ENCODINGS

- Absolute position embeddings: embeds the absolute position i into position vector
- Relative positional encodings: uses relative distance information between tokens, usually applied in attention layers
- RoPE (Rotary Position Embedding):
  - Stores positional information into the query and key vectors (q and k)
  - Finding the inner product of q and k gives relative position information

$$\mathbf{q}_m = f_q(\mathbf{x}_m, m) \in \mathbb{R}^{|L|},\ \mathbf{k}_n = f_k(\mathbf{x}_n, n) \in \mathbb{R}^{|L|}, \quad (1)$$

$$f_q(\mathbf{x}_m, m) = W_q \mathbf{x}_m e^{im\theta}$$

$$\langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n)\rangle_{\mathbb{R}} = \mathbf{Re}(\langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n)\rangle_{\mathbb{C}})$$
$$= \mathbf{Re}(\mathbf{x}_m^* W_q^* W_k \mathbf{x}_n e^{i\theta(m-n)}) = g(\mathbf{x}_m, \mathbf{x}_n, m-n), \quad (2)$$

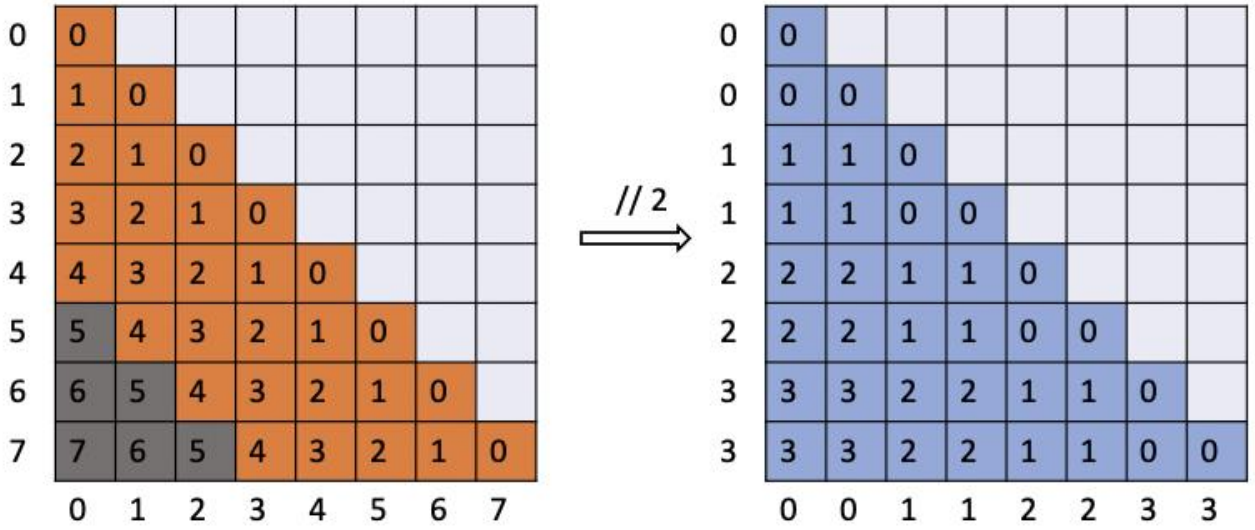
THE PROBLEM: O.O.D ISSUES

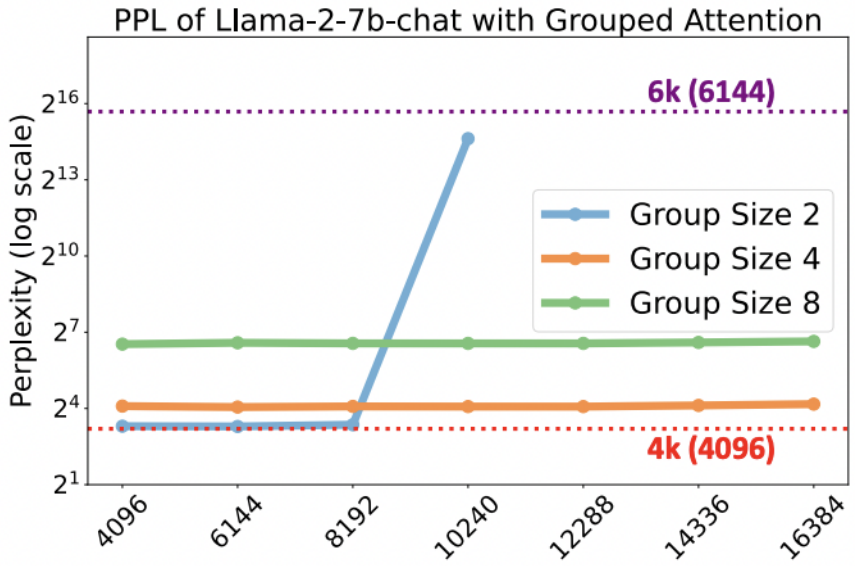
- Out of Distribution (O.O.D)
- Long contexts: LLMs are exposed to new relative distances not seen to them before.
- Attention distribution is different vs. pretraining
- **Potential solution:** Floor operation

GROUPED ATTENTION

- Using Floor: maps original large position to smaller discrete set of values, avoiding OOD
- Needs a set group size G_s

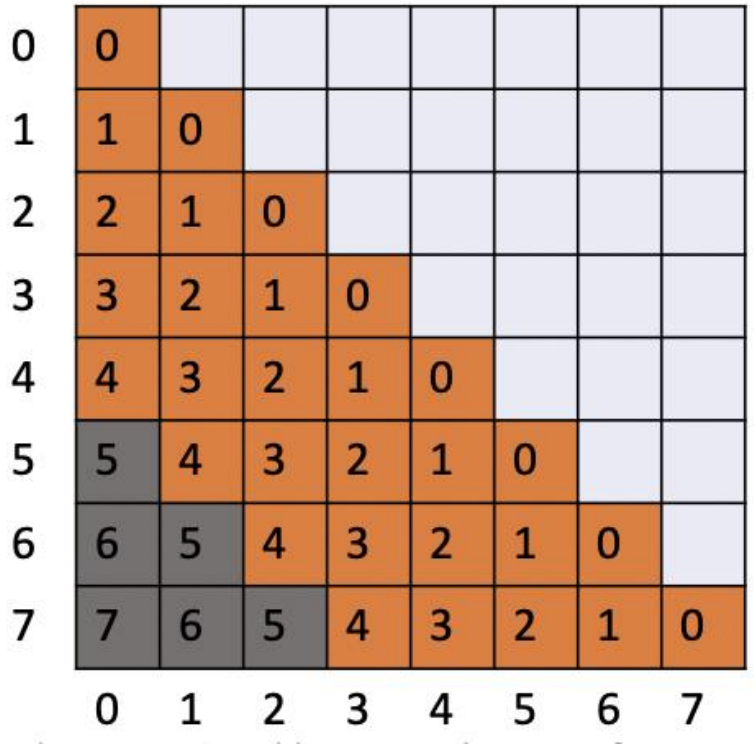
$$P_g = P \;//\; G_s, \tag{3}$$





NORMAL ATTENTION

- When generating next tokens, neighbors of the target are still important
- Standard attention mechanism must be preserved
- Needs a window size w_n

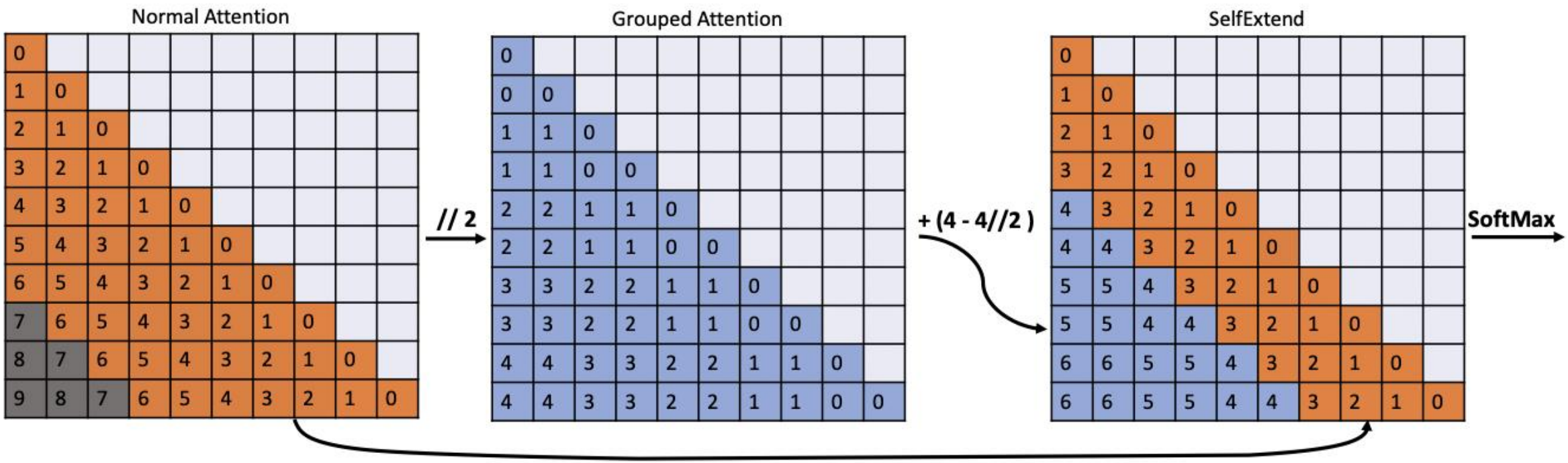


SELF-EXTEND

- **Self-Extend:** enhances LLM's long context abilities *without* fine-tuning
- Intuition: exact word position becomes less important in long texts compared to overall meaning and relative order of information.
- Combines grouped and normal attention
  - Normal Attention for neighbor tokens
  - Grouped Attention for non-neighbor tokens (far apart)
- Only modifies attention during inference: no additional fine tuning
- Ideal max length of the extended context window:

$$(L - w_n) * G_s + w_n.$$

$$w_n - w_n//G_s$$


Normal Attention
Grouped Attention
SelfExtend
// 2
+ (4 - 4//2)
SoftMax

EXPERIMENT

- Evaluated SelfExtend with Llama-2, Phi-2, Mistral, and SOLAR
- Language modeling Tasks
- Synthetic long context tasks
- Real-world long context tasks
- Short-context tasks
- Tradeoffs with Group Size, Neighbor Window Size

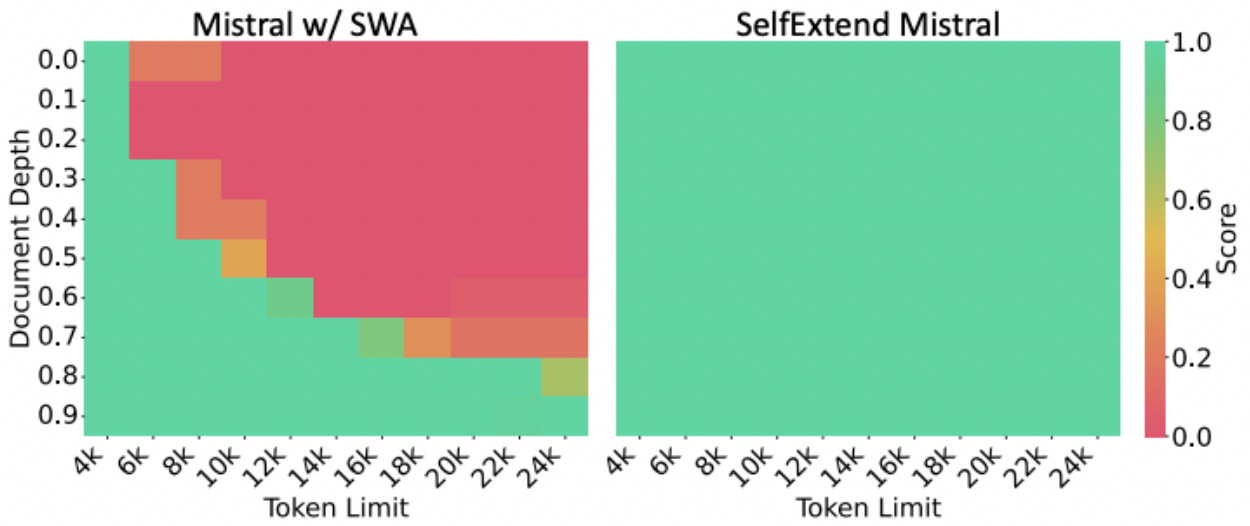
RESULTS – LANGUAGE MODELING

- PG19 Dataset
  - Consists of lengthy books
- SelfExtend maintains a low PPL out of the pretraining context window
  - Mistral w/ Sliding Window Attention also maintains a low PPL, even without SelfExtend

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
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

RESULTS – SYNTHETIC LONG CONTEXT TASKS

- Synthetic long context task: passkey retrieval
  - retrieving a simple passkey in a long sequence of text
  - placed at various document depths
  - Tested across different context lengths
- SelfExtend obtains a 100% passkey retrieval across all tested depths and context lengths



RESULTS – REAL WORLD LONG CONTEXT TASKS

- Synthetic tasks like passkey retrieval aren't enough to fully assess long-context capabilities
- Two benchmarks: LongBench and L-Eval
  - LongBench: 21 datasets across 6 task categories in both English and Chinese, average length of 6,711 words (English) and 13,386 characters (Chinese)
  - L-Eval: evaluation suite with: 20 sub-tasks, 508 long documents, over 2,000 human-labeled query-response pairs of diverse question styles, domains, and input length (3k - 200k tokens)
- Llama-2: increased contexts to 4k, 8k, 16k
- Mistral: increased contexts to 16k
- SelfExtends achieves comparable or better performance compared to methods that require finetuning

RESULTS – REAL WORLD LONG CONTEXT TASKS

	LLMs[a]	Single-Document QA			Multi-Document QA			Summarization			Few-shot Learning			Synthetic		Code	
		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**
	Llama-3-8B-ins-8k+	21.71	44.24	44.54	46.82	36.42	21.49	30.03	22.67	**27.79**	74.5	90.23	**42.53**	NA	67.00	**57.00**	51.22
	SE-Llama-3-8B-ins-16k+	**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**
	SE-Llama-3-8B-ins-32k 10/96+	21.50	43.96	50.26	48.18	28.18	25.58	34.88	23.83	26.96	75.50	88.26	42.01	4.12	88.0	36.58	37.73
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
	Gradient-Llama-3-8B-Inst-262k(32k)+	21.71	44.24	44.54	46.82	36.42	21.49	30.03	22.67	**27.79**	74.5	90.23	**42.53**	NA	67.00	**57.00**	51.22
	Gradient-Llama-3-8B-Inst-1M(32k)+	**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**
	SE-Llama-3-8B-ins-32k 10/96+	**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**

Model	Tokens	Coursera	GSM	QuALITY	TOEFL	CodeU	SFiction	Avg.
Claude1.3-100k	100k	60.03	88.00	73.76	83.64	17.77	72.65	65.97
GPT-4-32k	32k	75.58	96.00	82.17	84.38	25.55	74.99	73.11
Turbo-16k-0613	16k	63.51	84.00	61.38	78.43	12.22	64.84	60.73
Chatglm2-6b-8k	2k	43.75	13.00	40.59	53.90	2.22	54.68	34.69
XGen-7b-8k (2k-4k-8k)	2k	26.59	3.00	35.15	44.23	1.11	48.43	26.41
Chatglm2-6b-8k	8k	42.15	18.00	44.05	54.64	2.22	54.68	35.95
Chatglm2-6b-32k	32k	47.81	27.00	45.04	55.01	2.22	57.02	39.01
XGen-7b-8k	8k	29.06	16.00	33.66	42.37	3.33	41.40	27.63
MPT-7b-65k	8k	25.23	8.00	25.24	17.84	0.00	39.06	19.22
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**

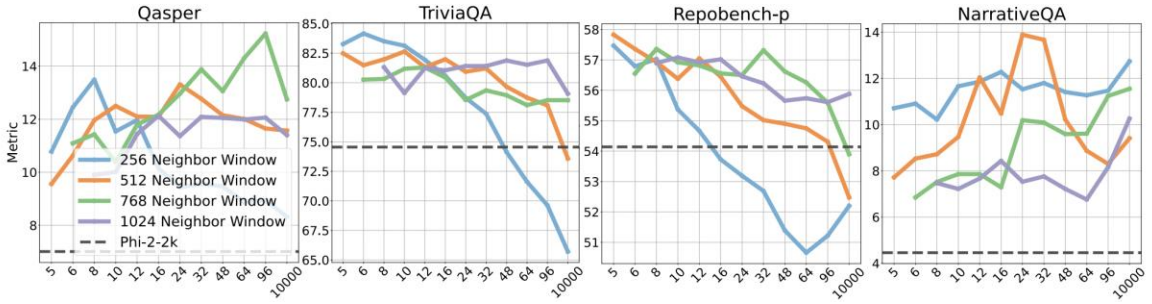
RESULTS – SHORT CONTEXT TASKS

- Finetuned methods can lead to degraded performance on short-context tasks
- Evaluate performance on five public short context tasks (Hugging Face Open LLM benchmark)
- SelfExtend maintains performance of short-context tasks
- Can be readily adopted, as it is automatically "disabled" when it comes to short-text sequences

Model	Tokens	Coursera	GSM	QuALITY	TOEFL	CodeU	SFiction	Avg.
Claude1.3-100k	100k	60.03	88.00	73.76	83.64	17.77	72.65	65.97
GPT-4-32k	32k	75.58	96.00	82.17	84.38	25.55	74.99	73.11
Turbo-16k-0613	16k	63.51	84.00	61.38	78.43	12.22	64.84	60.73
Chatglm2-6b-8k	2k	43.75	13.00	40.59	53.90	2.22	54.68	34.69
XGen-7b-8k (2k-4k-8k)	2k	26.59	3.00	35.15	44.23	1.11	48.43	26.41
Chatglm2-6b-8k	8k	42.15	18.00	44.05	54.64	2.22	54.68	35.95
Chatglm2-6b-32k	32k	47.81	27.00	45.04	55.01	2.22	57.02	39.01
XGen-7b-8k	8k	29.06	16.00	33.66	42.37	3.33	41.40	27.63
MPT-7b-65k	8k	25.23	8.00	25.24	17.84	0.00	39.06	19.22
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**

TRADEOFFS & ADDITIONAL FINDINGS

- Tradeoffs with group size & neighbor window size
  - When group size becomes too large, position information becomes less precise
  - Small group sizes requires the use of larger position embeddings
  - larger window size means more accurate info about neighbor tokens, BUT requires larger group size
- Varying context window length
  - Longer context window performs better, but peaks
- Varying Passkey length



Context Length	2k (vanilla)	4k	6k	8k
Document QA				
NarrativeQA	4.46	6.49 (+45.52%)	8.98 (+101.35%)	12.04 (+169.96%)
Qasper	7.01	11.16 (+59.20%)	12.84 (+83.17%)	12.10 (+72.61%)
Summarization				
Gov_report	25.46	27.91 (+9.62%)	28.14 (+10.53%)	27.51 (+8.05%)
Qmsum	14.32	14.88 (+3.91%)	16.72 (+16.76%)	18.58 (+29.75%)
Few-shot Learning				
Trec	50.5	60.0 (+18.81%)	62.5 (+23.76%)	60.0 (+18.81%)
Triviaqa	74.55	84.88 (+13.86%)	82.64 (+10.85%)	81.31 (+9.07%)
Coding				
Repobench-p	54.14	56.18 (+3.77%)	56.76 (+4.84%)	57.05 (+5.37%)
Lcc	58.96	59.06 (+0.17%)	58.88 (-0.14%)	59.42 (+0.78%)

CONCLUSION AND LIMITATIONS

- Main Concepts:
  - The degrading performance caused by long contexts is due to O.O.D issues in relative position embeddings (the long distance between tokens in context exceeds training window)
  - Grouped attention: using a floor operation to map these unknown distances to ones encountered during training
  - Combining grouped attention with normal attention at inference time allows LLMs to better handle long context without additional finetuning
- Limitations:
  - Extra computation costs (extra attention across all query-key pairs, processing entire sequences)
  - Performance degrades with large group size, so no indefinite context

DISCUSSION

- LLMs have an inherent ability to understand long contexts, even when not trained on them
- Introduces SelfExtend, a method that enhances LLM's capability to process long contexts without fine-tuning or further training.
  - SelfExtend surpasses many fine-tuning based models
- Future Work
  - More sophisticated methods of remapping

RETRIEVAL MEETS LONG CONTEXT LARGE LANGUAGE MODELS

Peng Xu[†], Wei Ping[†], Xianchao Wu, Lawrence McAfee
Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina
Mohammad Shoeybi, Bryan Catanzaro
NVIDIA
[†]{pengx, wping}@nvidia.com

MOTIVATION

- Which is better for model performance on downstream tasks, long context windows or retrieval-augmentation?
- Is there a way to combine both methods?

EXPERIMENTS

Template for LLM to follow:
"System: {System}\n\nUser: {Question}\n\nAssistant: {Answer}
"System: {System}\n\n{Context}\n\nUser: {Question}\n\nAssistant: {Answer}"

- LLMs:
  - GPT-43B and Llama2-70B
- Context Window Extension:
  - Positional Interpolation to extend GPT-43B from 4k -> 16K, Llama2-7B from 4k -> 32k
- Retriever:
  - Dragon, Contriever, OpenAI Embeddings
- Instruction Tuning:
  - Soda dataset, Open Assistant dataset, Dolly dataset, and a proprietary conversational dataset

DATASETS

	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
# of samples	200	1,726	2,000	2,000	200	200	150
avg doc length	14,140	4,912	84,770	6,592	16,198	13,319	7,185
avg top-5 chunks	2,066	2,071	2,549	2,172	2,352	2,322	2,385
avg top-10 chunks	4,137	3,716	5,125	4,018	4,644	4,554	4,305
avg top-20 chunks	8,160	4,658	10,251	5,890	9,133	8,635	6,570

- QMSum (QM)
  - Query-focused summarization dataset
- Qasper (QASP)
  - Question answering (QA) dataset
- NarrativeQA (NQA)
  - Long-context QA dataset
- QuALITY (QLTY)
  - Long-document multiple-choice QA dataset
- HotpotQA (HQA)
  - Wikipedia-based multi-hop QA dataset
- MuSiQue (MSQ)
  - Challenging multi-hop QA dataset
- MultiFieldQA-en (MFQA)
  - Long-context QA dataset

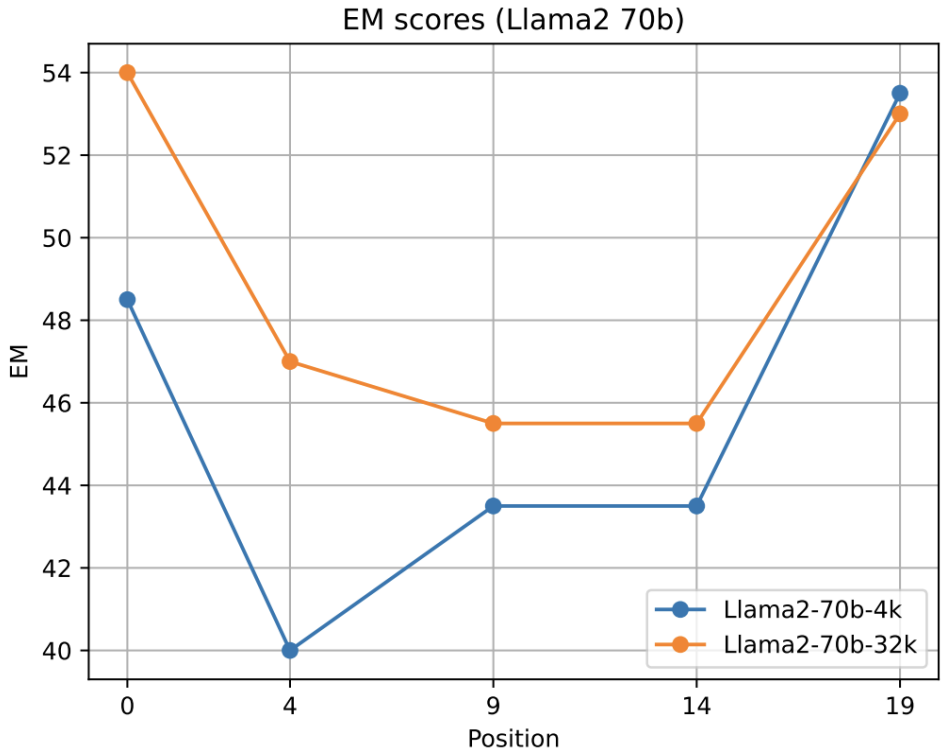
RESULTS

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
Llama2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
Llama2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
Llama2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**
Llama2-7B	4k	22.65	14.25	22.07	14.38	40.90	8.66	23.13	35.20
+ ret	4k	**26.04**	16.45	22.97	18.18	43.25	14.68	26.62	40.10
Llama2-7B	32k	**28.20**	16.09	23.66	19.07	44.50	15.74	31.63	46.71
+ ret	32k	27.63	17.11	23.25	19.12	43.70	15.67	29.55	45.03

- Retrieval helped a lot with 4k models
  - Llama2-70B-4K: 31.61 to 36.02
  - GPT-43B-4K: 26.44 to 29.32
- HotpotQA (HQA) especially favors long sequence models with sequence length 4k -> 16k
  - Llama2-70B: 34.64 to 43.97
  - GPT-43B: 28.91 to 37.48

LOST IN THE MIDDLE PHENOMENON

- It is quite interesting that the retrieval-augmented long context LLM (e.g., 16K and 32K) can obtain better results than retrieval-augmented 4K context LLM, even with the same top 5 chunks of evidence as input.



CONFLICTING RESULTS

- LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding (Bai et al., 2023)

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
Llama2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
Llama2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
Llama2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**
Llama2-7B	4k	22.65	14.25	22.07	14.38	40.90	8.66	23.13	35.20
+ ret	4k	**26.04**	16.45	22.97	18.18	43.25	14.68	26.62	40.10
Llama2-7B	32k	**28.20**	16.09	23.66	19.07	44.50	15.74	31.63	46.71
+ ret	32k	27.63	17.11	23.25	19.12	43.70	15.67	29.55	45.03

COMPARING TO OPENAI MODELS

- Models:
  - Llama2-70B-32k, Davinci003, GPT-3.5-turbo (4k), GPT-3.5-turbo-16k
- Scores with * were prepared by the organizers of the benchmarks
  - Avg-7 is average across all 7 datasets, Avg-4 is average across the ZeroSCROLLS benchmark

Model	Avg-7	Avg-4*	QM*	QASP*	NQA*	QLTY*	MSQ	HQA	MFQA
Davinci003 (175B)	39.2	40.8*	16.9*	52.7*	24.6*	69.0*	22.1	41.2	47.8
GPT-3.5-turbo (4k)	38.4	39.2*	15.6*	49.3*	25.1*	66.6*	21.2	40.9	49.2
+ret							24.4	49.5	49.5
GPT-3.5-turbo-16k	42.8	42.4	17.6	50.5	28.8	72.6	26.9	51.6	52.3
+ret							30.4	46.6	52.8
Llama2-70B-32k	40.9	42.4	15.6	45.9	28.4	79.6	19.1	49.5	48.4
Llama2-70B-32k-ret	**43.6**	**43.0**	18.5	46.3	31.5	75.6	26.7	53.9	52.9

DIFFERENT RETRIEVERS

- Retrievers, regardless of specific model, can improve performance on both short and long contexts.

Seq len	Setting	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
4k	baseline (w/o ret)	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
	Dragon	35.73	18.14	29.20	23.39	70.30	20.09	41.54	47.45
	Contriever	**36.02**	17.41	28.74	23.41	70.15	21.39	42.06	48.96
	OpenAI-embedding	35.79	17.76	28.85	23.57	70.70	19.92	41.76	47.99
32k	baseline (w/o ret)	37.36	15.37	31.88	23.59	73.80	19.07	49.49	48.35
	Dragon	**39.60**	18.34	31.27	24.53	69.55	26.72	53.89	52.91
	Contriever	38.85	17.60	31.56	23.88	69.00	26.61	49.65	53.66
	OpenAI-embedding	39.34	18.24	32.07	24.36	69.45	24.90	51.64	54.75

MORE RETRIEVED CHUNKS

- Best results for top 5 or top 10 chunks.

Seq len	Setting	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
4k	base	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
	top-5	**35.73**	18.14	29.20	23.39	70.30	20.09	41.54	47.45
	top-10	34.62	16.54	28.67	24.38	68.70	19.00	42.18	42.84
	top-20	34.61	16.52	28.67	24.38	68.70	19.00	42.18	42.84
16k	base	36.78	16.72	30.92	22.32	76.10	18.78	43.97	48.63
	top-5	37.23	18.70	29.54	23.12	70.90	23.28	44.81	50.24
	top-10	**38.31**	18.41	30.20	25.53	73.60	22.78	47.72	49.91
	top-20	36.61	17.26	29.60	25.81	72.30	22.69	41.36	47.23
32k	base	37.36	15.37	31.88	23.59	73.80	19.07	49.49	48.35
	top-5	**39.60**	18.34	31.27	24.53	69.55	26.72	53.89	52.91
	top-10	38.98	17.71	30.34	25.94	70.45	22.80	55.73	49.88
	top-20	38.38	16.36	30.42	24.42	69.60	24.51	54.67	48.65

RETRIEVAL FOR FEW-SHOT TASKS

- Uses two more datasets: Trec and SAMSum
- Llama2-70B-32k-ret outperforms its non-retrieval Llama2-70B-32k baseline as well as GPT-3.5-turbo-16k by a large margin.
  - Confirms benefits of using retrieval together with long context models.

Model	Trec	SAMSum
GPT-3.5-turbo-16k	68	41.7
Llama2-70B	73	46.5
Llama2-70B-ret	76	47.3

CONCLUSION

- Retrieval largely boost performance of both 4K short context LLM and 16K/32K long context LLMs
- 4K context LLMs with simple retrieval-augmentation can perform comparable to 16K long context LLMs while being more efficient at inference
- After context window extension and retrieval-augmentation, best model Llama2-70B-32k-ret can outperform GPT-3.5-turbo-16k and Davinci003

LIMITATIONS & FUTURE DIRECTIONS

- Task Scope
- Noise and Retrieval Set-Up
- Long-Context Extension Method

- Develop advanced methods for existing pretrained large language models
- Further extending context window to 64k
- How to mitigate "lost-in-middle"

FINAL SUMMARY

- Long context language models can be used to process thousands of tokens at a time, but exhibits primacy and recency bias, where target tokens are lost when placed in the middle of the context.
- Lost in the Middle: Long Context LLMs perform better when the answer is in the beginning or end of context, which is explained by primacy and recency bias.
- LLM Maybe LongLM: Mapping relative positional encodings allows the LLM to generalize better to longer context without having to retrain on a larger context window.
- Retrieval meets Long Context LLM: Retrieval, paired with short or long context models, outperforms long context on its own.