slido

Please download and install the
Slido app on all computers you use

Audience Q&A

ⓘ Start presenting to display the audience questions on this slide.

Announcement

HW6 mini-project should be well underway.

If you have not started, you are behind.

Final time announced, check the official calendar.

Final's scope is every lecture after Midterm.

Long Context Language Models

Large Language Models: Methods and Applications

Daphne Ippolito and Chenyan Xiong

Learning Objectives

Learn the scenarios where long-context is explored

Learn the technologies that pretrain long-context models

Understand the benefits and limitations of current long-context models

Outline

Motivation

Probing Long Context Ability

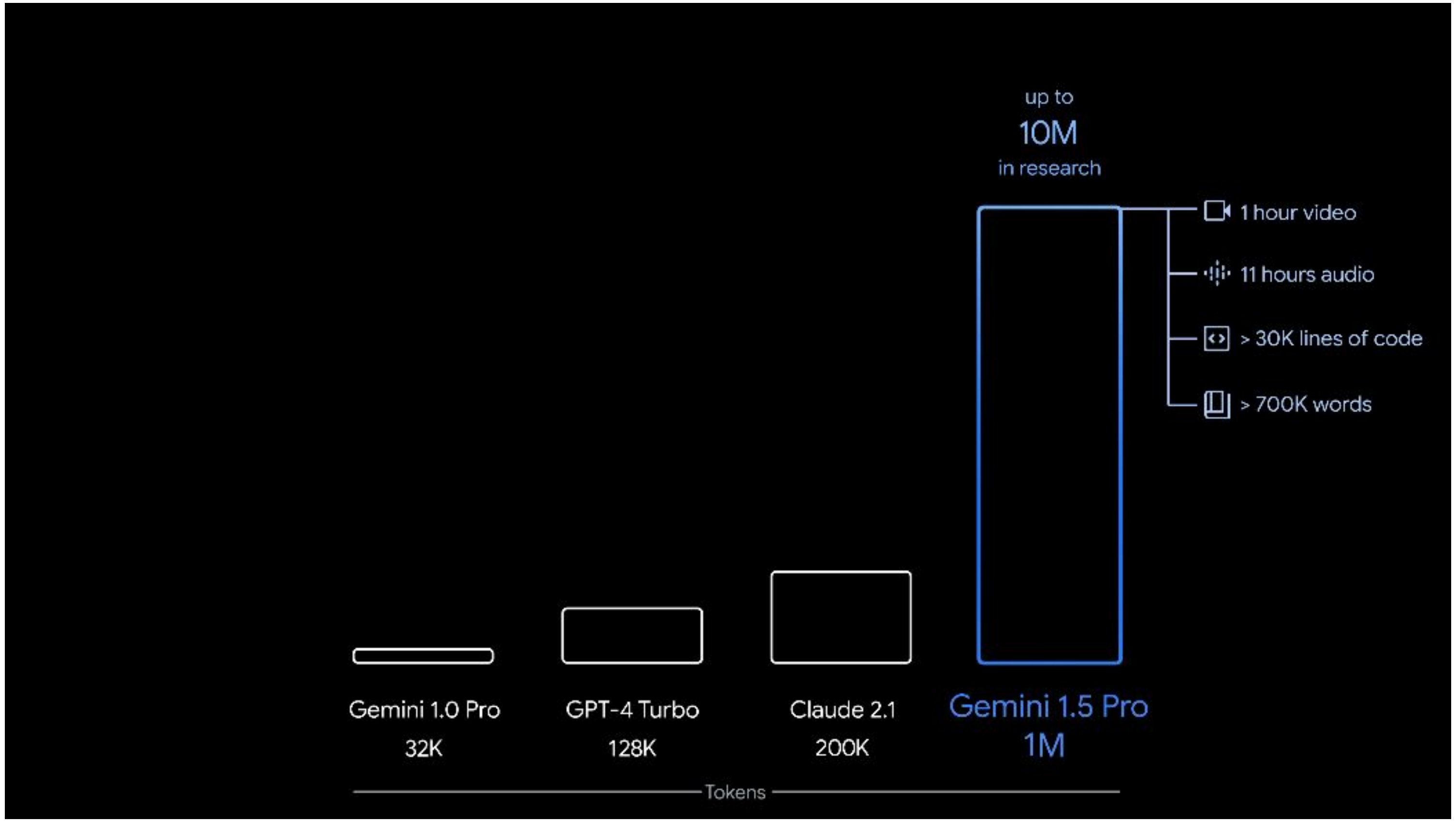
Evaluation on Real Scenarios

Adapting LLMs to Long Context Tasks

Efficient Serving

Long-Context Ability of LLMs

One of the main "competing" metric of industry LLMs in the past year [1]



[1] https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

slido

Please download and install the Slido app on all computers you use

Why we need long context in LLMs?

ⓘ Start presenting to display the poll results on this slide.

Why Long-context?

Many scenarios naturally needs long inputs. 4K token is not enough

- Chatbot: long conversation history
- RAG: lots of retrieved documents
- Code: large repository
- Fancy Prompts: can be very long

[4] Hsieh et al. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models?

Why Long-context?

Many scenarios naturally needs long inputs. 4K token is not enough

- Chatbot: long conversation history
- RAG: lots of retrieved documents
- Code: large repository
- Fancy Prompts: can be very long

Ideally, a lot of imagination towards AGI

- Short term memory
- Long term reasoning across multiple text pieces
- Global understanding
- Bring the AGI power of LLMs to all the above

Long-context Demo of Gemini

https://www.youtube.com/watch?v=LHKL_210CcU&t=107s&ab_channel=Google

Outline

Motivation

Probing Long Context Ability

Evaluation on Real Scenarios

Adapting LLMs to Long Context Tasks

Efficient Serving

How LLMs Use Long Context?

Multi-document QA Task: Answer the question from one relevant document places in the context

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:

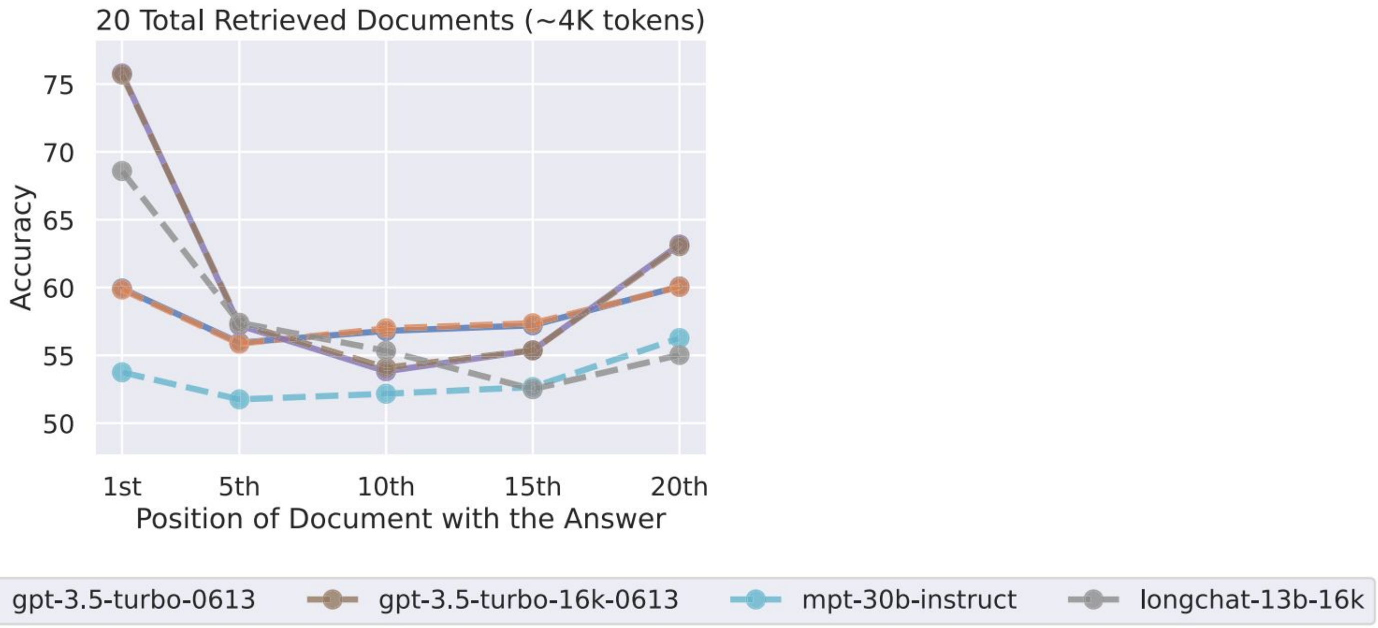
Desired Answer

Wilhelm Conrad Röntgen

Probing QA Ability with Multiple Document Contexts [2]

[2] Liu et al. 2024. Lost in the Middle: How Language Models Use Long Contexts.

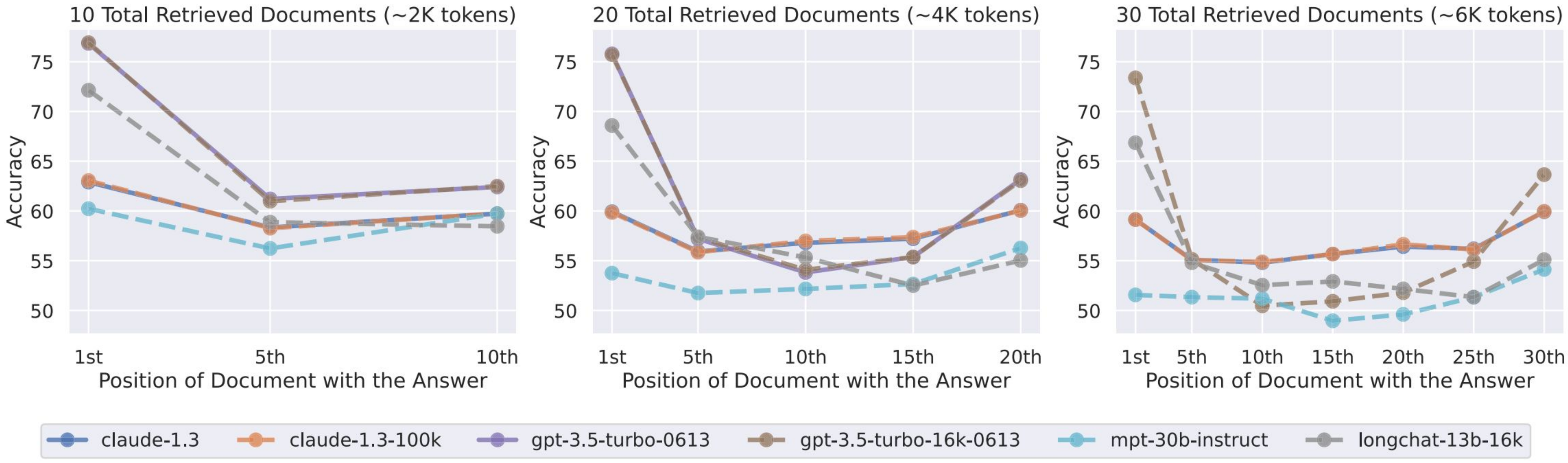
How LLMs Use Long Context?



QA Accuracy with Relevant Docs at Different Positions in Context [2]

[2] Liu et al. 2024. Lost in the Middle: How Language Models Use Long Contexts.

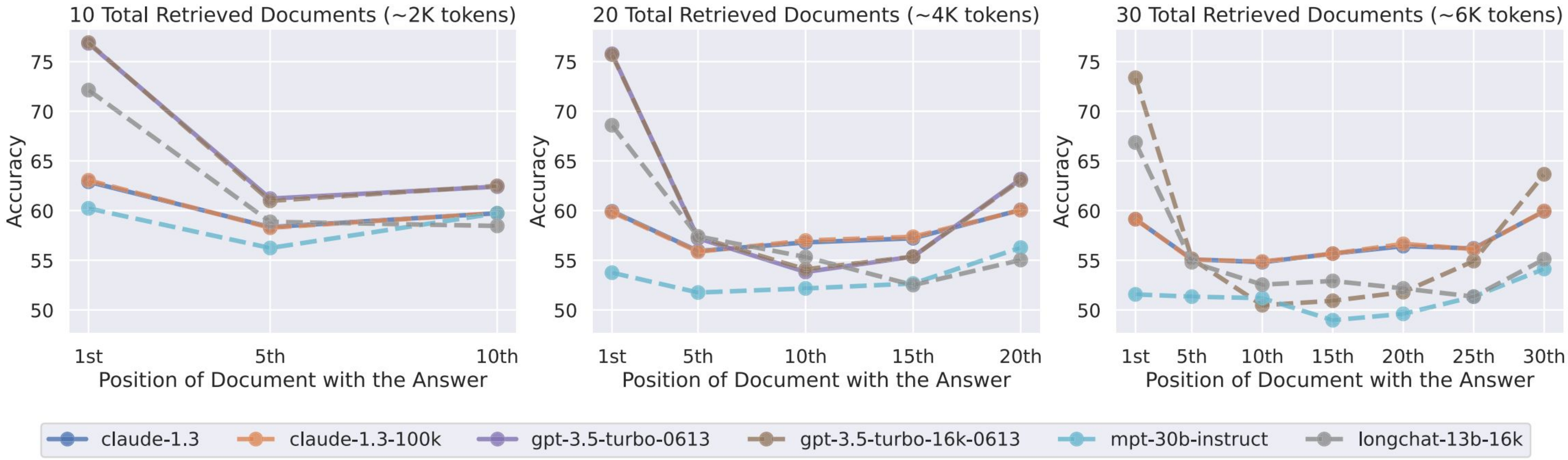
How LLMs Use Long Context?



QA Accuracy with Relevant Docs at Different Positions in Context [2]

[2] Liu et al. 2024. Lost in the Middle: How Language Models Use Long Contexts.

How LLMs Use Long Context?



QA Accuracy with Relevant Docs at Different Positions in Context [2]

More irrelevant contexts distract LLMs

Lost-in-the-middle: worst at finding relevant information in the middle

[2] Liu et al. 2024. Lost in the Middle: How Language Models Use Long Contexts.

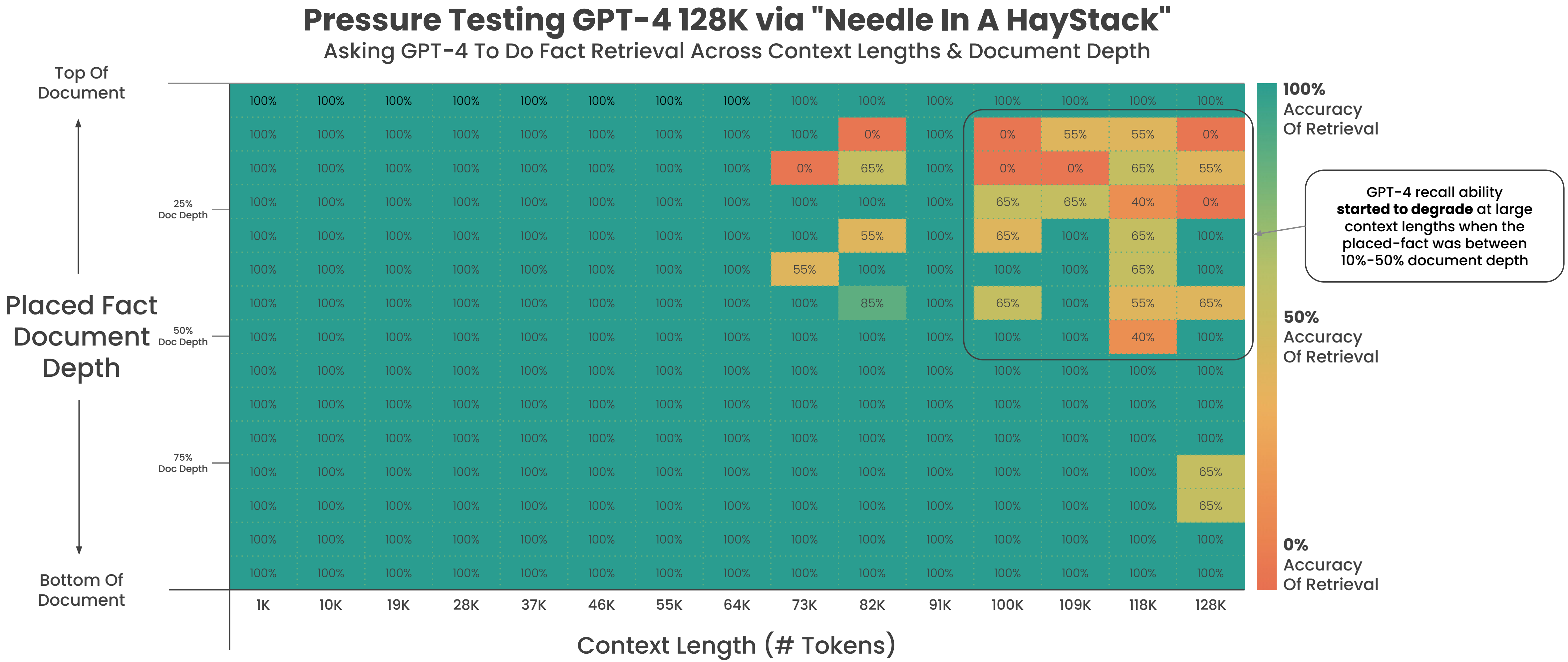
Needle in the Hack Test

Placing a “needle” into the context, and test if LLM can retrieve it.

- Needle: a random fact that unlikely to be part of the LLM’s parametric knowledge
  - E.g. "The 5 best things to do in San Francisco are: 1) Go to Dolores Park. 2) Eat at Tony's Pizza Napoletana. 3) Visit Alcatraz. 4) Hike up Twin Peaks. 5) Bike across the Golden Gate Bridge"
- Context: other unrelated documents
- Test: if the LLM can extract the answer perfectly
  - E.g. for question “What are the 5 best things to do in San Franscisco?”

[3] https://github.com/gkamradt/LLMTest_NeedleInAHaystack

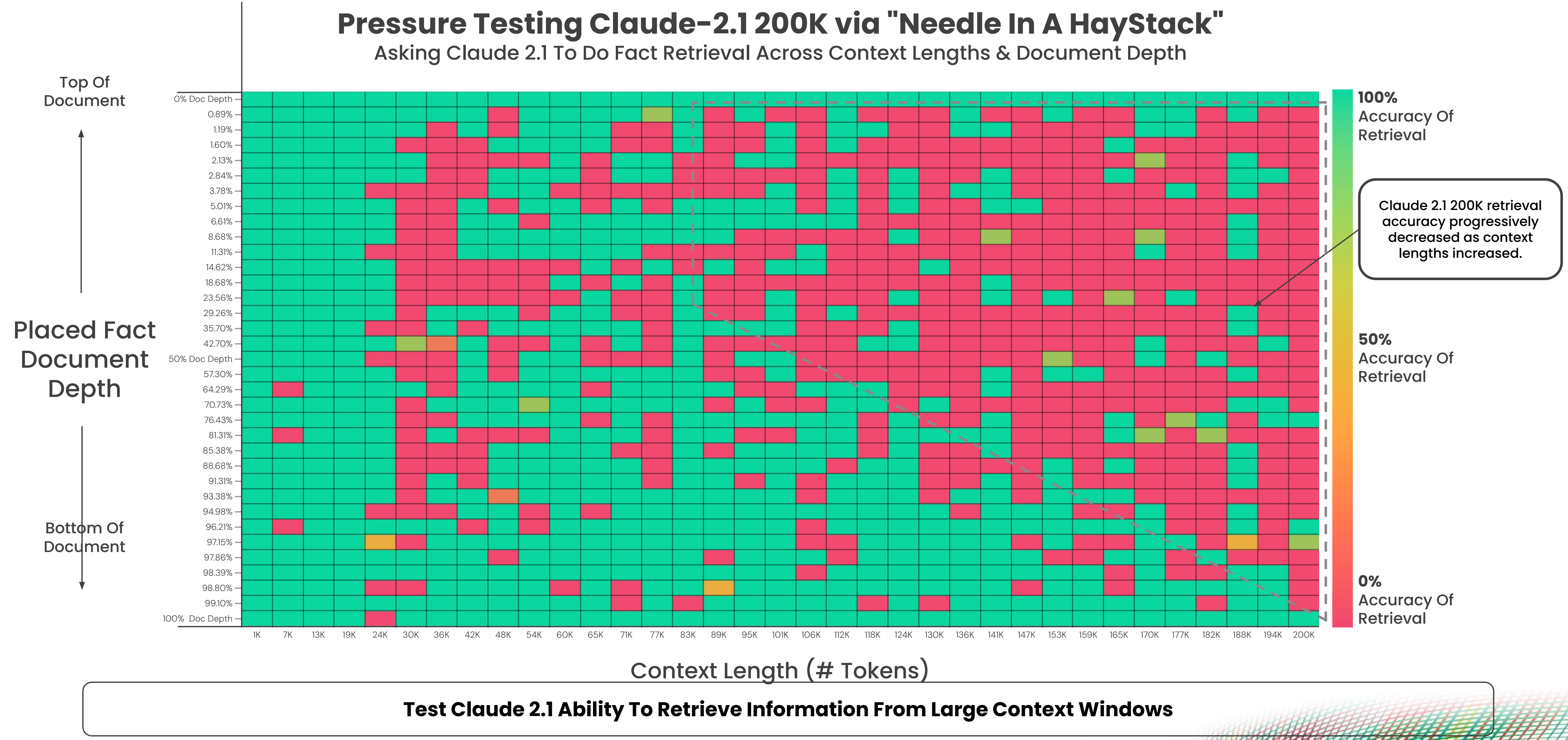
Needle in the Hack Test



Goal: Test GPT-4 Ability To Retrieve Information From Large Context Windows

[3] https://github.com/gkamradt/LLMTest_NeedleInAHaystack

Needle in the Hack Test



Needle in the Hack Test

Lots of varieties of synthetic tasks [4]:

Task	Configuration	Example
Single NIAH (S-NIAH)	type_key = word type_value = number type_haystack = essay size_haystack ∝ context length	(essays) ...... One of the special magic numbers for long-context is: 12345. ...... What is the special magic number for long-context mentioned in the provided text? Answer: 12345

Needle in the Hack Test

Lots of varieties of synthetic tasks [4]:

Task	Configuration	Example
Single NIAH (S-NIAH)	type_key = word type_value = number type_haystack = essay size_haystack ∝ context length	(essays) ...... One of the special magic numbers for long-context is: 12345. ...... What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-keys NIAH (MK-NIAH)	num_keys = 2 type_key = word type_value = number type_haystack = essay size_haystack ∝ context length	(essays) ...... One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. ...... What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-values NIAH (MV-NIAH)	num_values = 2 type_key = word type_value = number type_haystack = essay size_haystack ∝ context length	(essays) ...... One of the special magic numbers for long-context is: 12345. One of the special magic numbers for long-context is: 54321. ...... What are all the special magic numbers for long-context mentioned in the provided text? Answer: 12345 54321
Multi-queries NIAH (MQ-NIAH)	num_queries = 2 type_key = word type_value = number type_haystack = essay size_haystack ∝ context length	(essays) ...... One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. ...... What are all the special magic numbers for long-context and large-model mentioned in the provided text? Answer: 12345 54321

[4] Hsieh et al. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models?

Needle in the Hack Test

Lots of varieties of synthetic tasks [4]:

Task	Configuration	Example
Variable Tracking (VT)	num_chains = 2 num_hops = 2 size_noises ∝ context length	(noises) ...... VAR X1 = 12345 ...... VAR Y1 = 54321 ...... VAR X2 = X1 ...... VAR Y2 = Y1 ...... VAR X3 = X2 ...... VAR Y3 = Y2 ...... Find all variables that are assigned the value 12345. Answer: X1 X2 X3
Common Words Extraction (CWE)	freq_cw = 2, freq_ucw = 1 num_cw = 10 num_ucw ∝ context length	aaa bbb ccc aaa ddd eee ccc fff ggg hhh iii iii ...... What are the 10 most common words in the above list? Answer: aaa ccc iii ......
Frequent Words Extraction (FWE)	$\alpha = 2$ num_word ∝ context length	aaa bbb ccc aaa ddd eee ccc fff ggg aaa hhh aaa ccc iii iii ...... What are the 3 most frequently appeared words in the above coded text? Answer: aaa ccc iii
Question Answering (QA)	dataset = SQuAD num_document ∝ context length	Document 1: ...... aaa ...... Document 2: ...... bbb ...... Document 3: ...... ccc ...... Question: question Answer: bbb

Needle in the Hack Test

What is the extractive ability of LLMs on long inputs?

Models	Claimed Length	Effective Length	4K	8K	16K	32K	64K	128K	Avg.
Llama2 (7B)	4K	-	85.6						
Gemini-1.5-Pro	1M	>128K	96.7	95.8	96.0	95.9	95.9	94.4	95.8
GPT-4	128K	64K	96.6	96.3	95.2	93.2	87.0	81.2	91.6
Llama3.1 (70B)	128K	64K	96.5	95.8	95.4	94.8	88.4	66.6	89.6
Qwen2 (72B)	128K	32K	96.9	96.1	94.9	94.1	79.8	53.7	85.9
Command-R-plus (104B)	128K	32K	95.6	95.2	94.2	92.0	84.3	63.1	87.4
GLM4 (9B)	1M	64K	94.7	92.8	92.1	89.9	86.7	83.1	89.9
Llama3.1 (8B)	128K	32K	95.5	93.8	91.6	87.4	84.7	77.0	88.3
GradientAI/Llama3 (70B)	1M	16K	95.1	94.4	90.8	85.4	80.9	72.1	86.5
Mixtral-8x22B (39B/141B)	64K	32K	95.6	94.9	93.4	90.9	84.7	31.7	81.9
Yi (34B)	200K	32K	93.3	92.2	91.3	87.5	83.2	77.3	87.5
Phi3-medium (14B)	128K	32K	93.3	93.2	91.1	86.8	78.6	46.1	81.5
Mistral-v0.2 (7B)	32K	16K	93.6	91.2	87.2	75.4	49.0	13.8	68.4
LWM (7B)	1M	<4K	82.3	78.4	73.7	69.1	68.1	65.0	72.8
DBRX (36B/132B)	32K	8K	95.1	93.8	83.6	63.1	2.4	0.0	56.3
Together (7B)	32K	4K	88.2	81.1	69.4	63.0	0.0	0.0	50.3
LongChat (7B)	32K	<4K	84.7	79.9	70.8	59.3	0.0	0.0	49.1
LongAlpaca (13B)	32K	<4K	60.6	57.0	56.6	43.6	0.0	0.0	36.3

[4] Hsieh et al. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models?

Why Long-context?

Many scenarios naturally needs long inputs. 4K token is not enough

- Chatbot: long conversation history
- RAG: lots of retrieved documents
- Code: large repository
- Fancy Prompts: can be very long

Ideally, a lot of imagination towards AGI

- Short term memory
- Long term reasoning across multiple text pieces
- Global understanding
- Bring the AGI power of LLMs to all the above

[4] Hsieh et al. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models?

slido

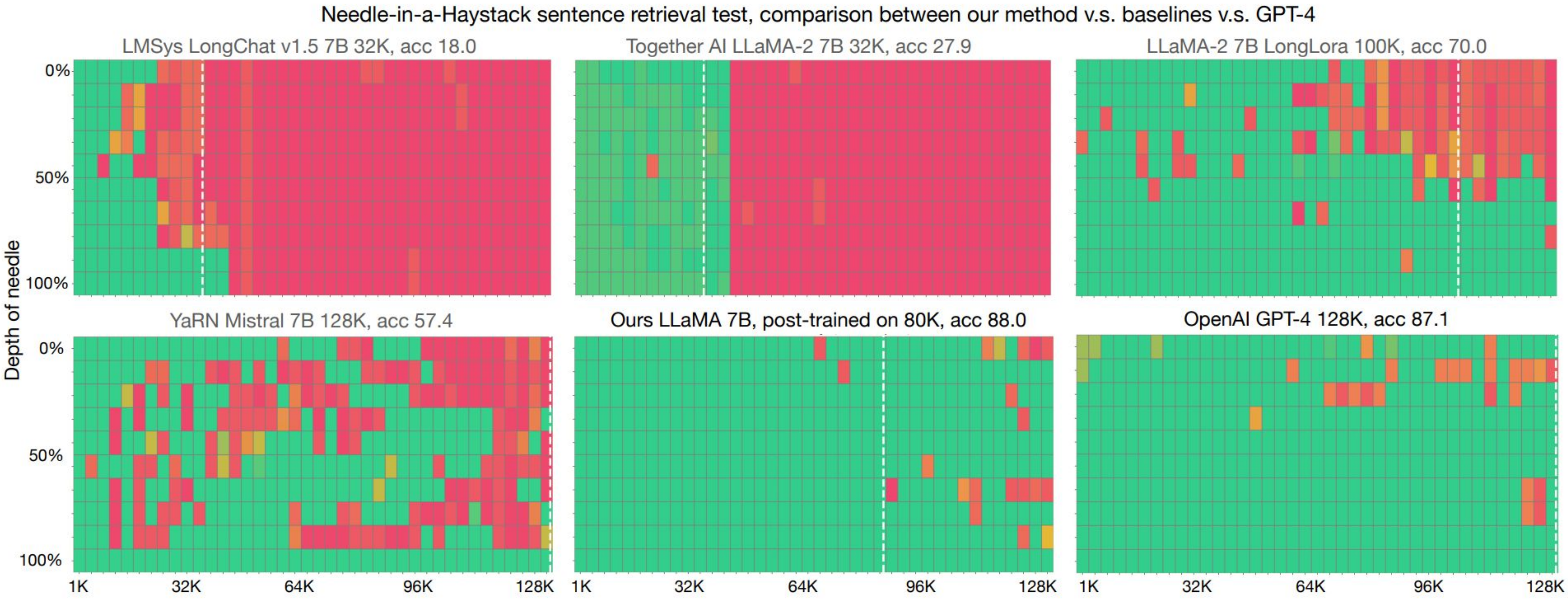
Please download and install the
Slido app on all computers you use

Can Needle In the Hack evaluation reflect LLM's long context ability

ⓘ Start presenting to display the poll results on this slide.

Needle in the Hack Test

Post training on longer documents yield almost perfect extraction [5]



[5] Fu et al. 2024. Data Engineering for Scaling Language Models to 128K Context.

Needle in the Hack Test

Easy to achieve 100 Needle in the Hach score (NIAH)

		HELMET		
Models	NIAH	Recall	RAG	Re-rank
Fu et al. (2024)	100	95.8	52.1	23.1
Llama-3.1-8B	100	99.4	56.3	37.0
Llama-3.1-70B	100	100	62.1	49.2

Outline

Motivation

Probing Long Context Ability

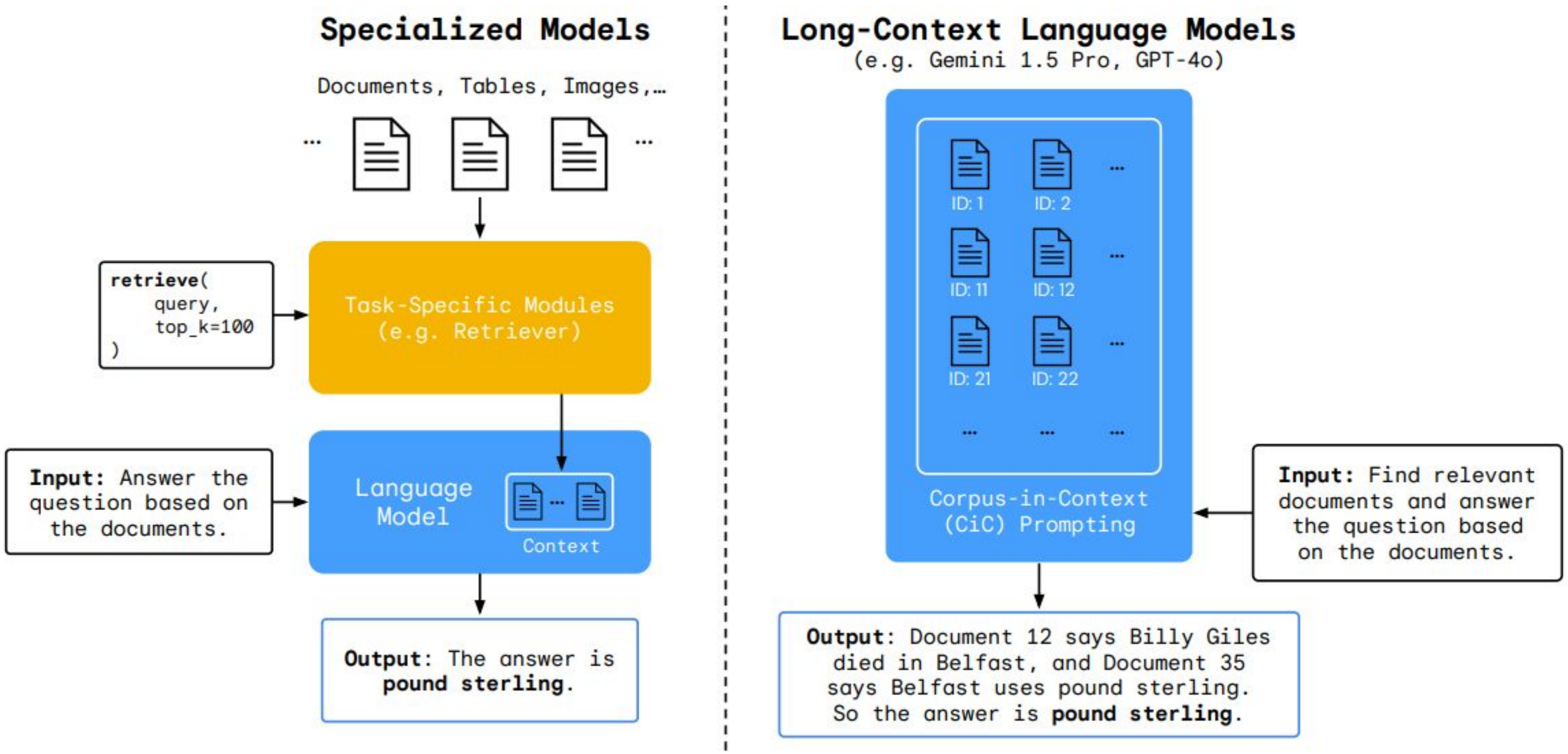
Evaluation on Real Scenarios

Adapting LLMs to Long Context Tasks

Efficient Serving

Real Task Performances

Can Long-context LLM replace task-specific models, like retriever?

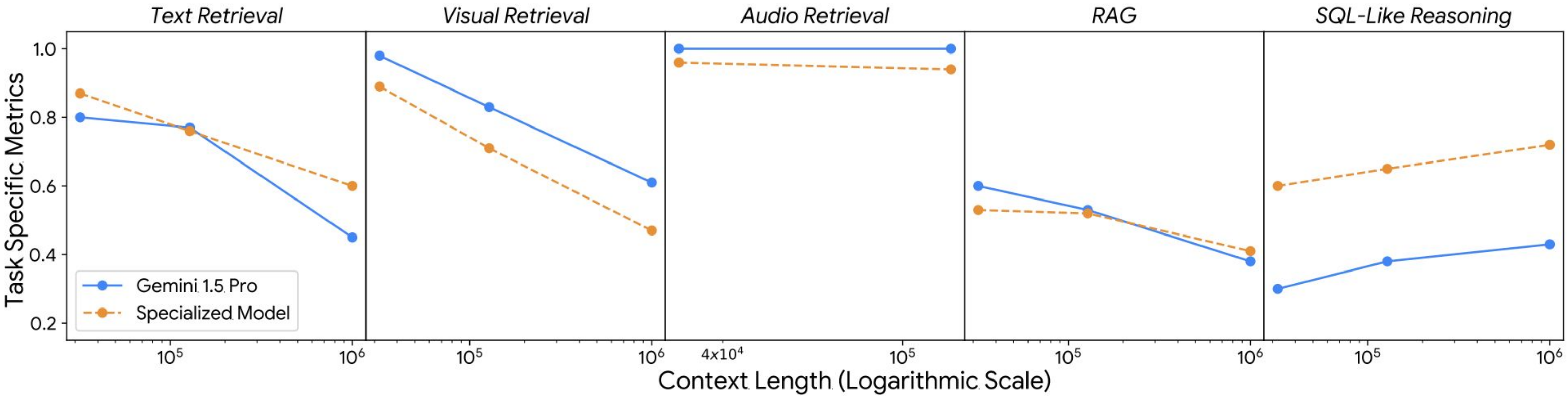


Specialized Models versus Long-Context LLMs [6]

[6] Lee et al. 2024. Can Long-Context Language Models Subsume Retrieval, RAG, SQL, and More?

Real Task Performances

Can Long-context LLM replace task-specific models, like retriever?



Specialized Models versus Long-Context LLMs [6]

[6] Lee et al. 2024. Can Long-Context Language Models Subsume Retrieval, RAG, SQL, and More?

Real Task Performances

Can Long-context LLM replace task-specific models, like retriever?

- Similar ish performances
- Pros: Convenience
- Cons: Cost

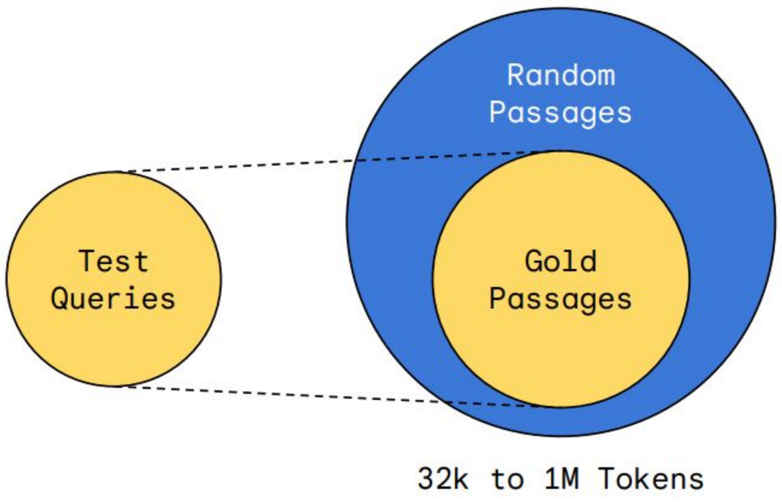
[6] Lee et al. 2024. Can Long-Context Language Models Subsume Retrieval, RAG, SQL, and More?

Real Task Performances

Can Long-context LLM replace task-specific models, like retriever?

- Similar ish performances
- Pros: Convenience
- Cons: Cost

Note, tested on distractor settings, much simpler than real retrieval



Also: 1M tokens are merely 500 documents

slido

Please download and install the
Slido app on all computers you use

Audience Q&A

ⓘ Start presenting to display the audience questions on this slide.

Outline

Motivation

Probing Long Context Ability

Evaluation on Real Scenarios

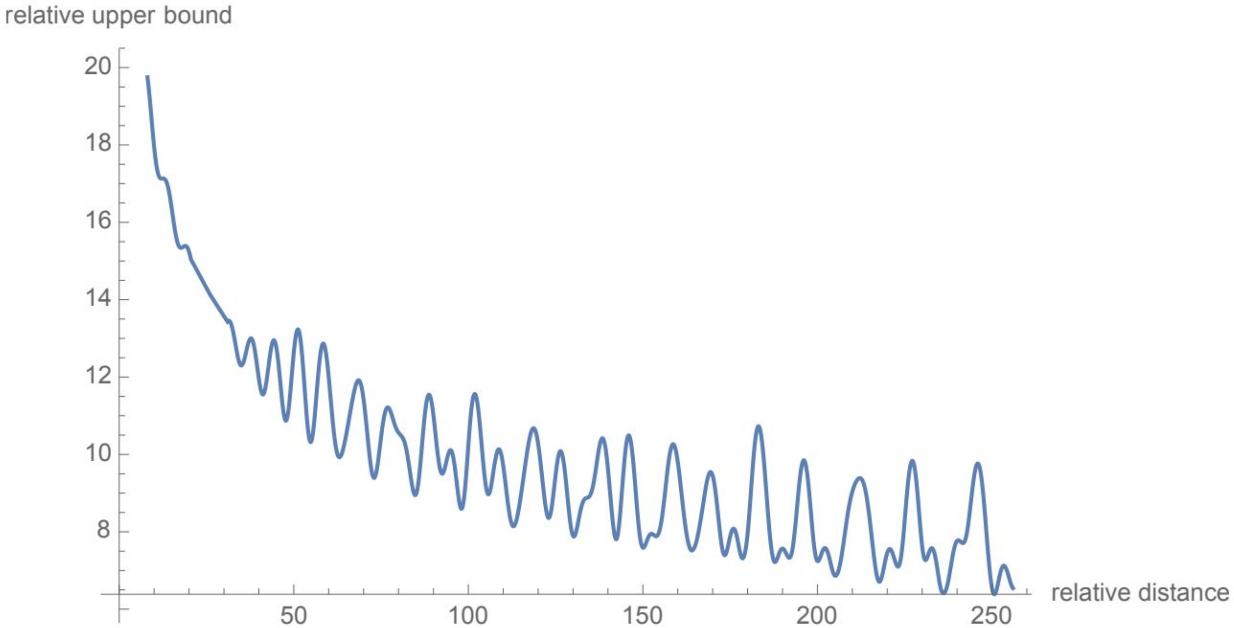
Adapting LLMs to Long Context Tasks

Efficient Serving

What are the gaps?

Positional encodings only capture shorter context

- Absolute and relative: never learned long context positions
- RoPE: strong decay over distance



Long-term Decay of RoPE [7]

[7] Sun et al. 2023. RoFormer: Enhanced transformer with Rotary Position Embedding

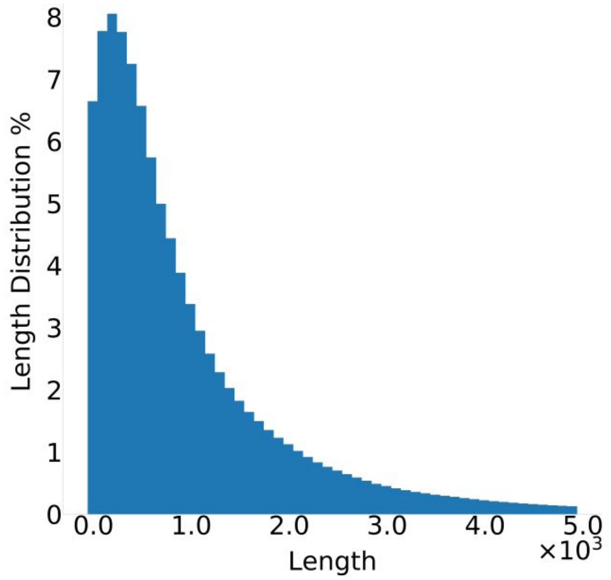
What are the gaps?

Format: Positional encodings only capture shorter context

- Absolute and relative: never learned long context positions
- RoPE: strong decay over distance

Distribution Shift:

- Pretraining data are mainly “short” documents
- Empirically no attentions across document boundary



Distribution of Web Page Length in Tokens [8]

[8] Overwijk et al. 2022. Clueweb22: 10 billion web documents with visual and semantic information.

What are the gaps?

Format: Positional encodings only capture shorter context

- Absolute and relative: never learned long context positions
- RoPE: strong decay over distance

Distribution Shift:

- Pretraining data are mainly “short” documents
- Empirically no attentions across document boundary

Source of Intelligence: how do LLMs learn long-term dependency or global reasoning?

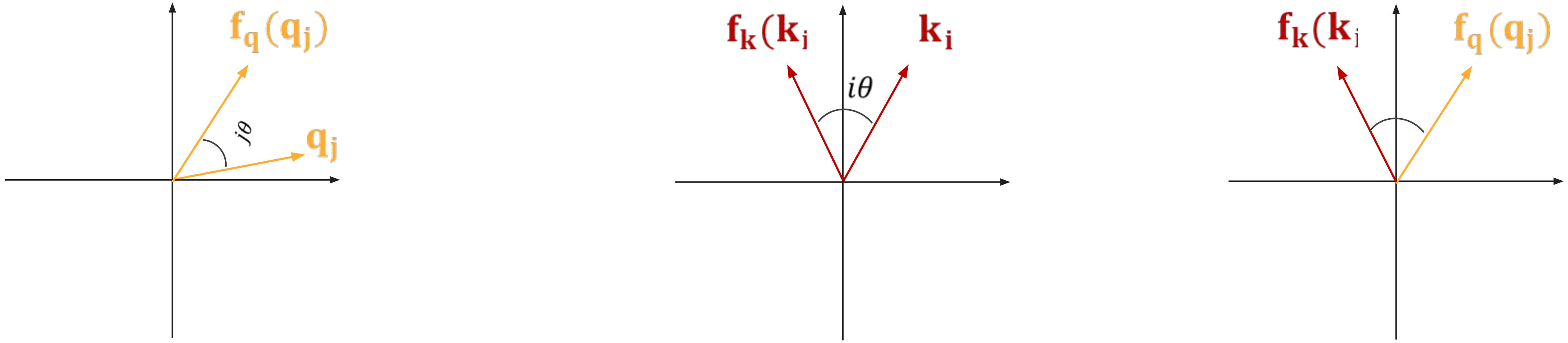
- Pretrained on next token prediction task solely
- How many next token prediction require information 128K tokens away?

[8] Overwijk et al. 2022. Clueweb22: 10 billion web documents with visual and semantic information.

Position Encoding: Rotational Position Embedding

Incorporate the vector rotation in the attention mechanism (2d space) [7]:

$$f_q(q_j) = \begin{pmatrix} \cos j\theta & -\sin j\theta \\ \sin j\theta & \cos j\theta \end{pmatrix} \begin{pmatrix} q_j^1 \\ q_j^2 \end{pmatrix} \qquad f_k(k_i) = \begin{pmatrix} \cos i\theta & -\sin i\theta \\ \sin i\theta & \cos i\theta \end{pmatrix} \begin{pmatrix} k_i^1 \\ k_i^2 \end{pmatrix} \qquad \theta_k = (1/b^{2(i-1)/d})$$



Rotate q by $j\theta$

Rotate k by $i\theta$

Attention only depends on $i - j$

Attention score by the dot prod of rotated vectors:

$$\text{Attention score}(\mathbf{q}_j, \mathbf{k}_i) = f_q(q_j) \cdot f_k(k_i) = \begin{pmatrix} q_j^1 \\ q_j^2 \end{pmatrix}^T \begin{pmatrix} \cos j\theta & -\sin j\theta \\ \sin j\theta & \cos j\theta \end{pmatrix}^T \begin{pmatrix} \cos i\theta & -\sin i\theta \\ \sin i\theta & \cos i\theta \end{pmatrix} \begin{pmatrix} k_i^1 \\ k_i^2 \end{pmatrix}$$

[7] Sun et al. 2023. RoFormer: Enhanced transformer with Rotary Position Embedding

Adapting Positional Encoding

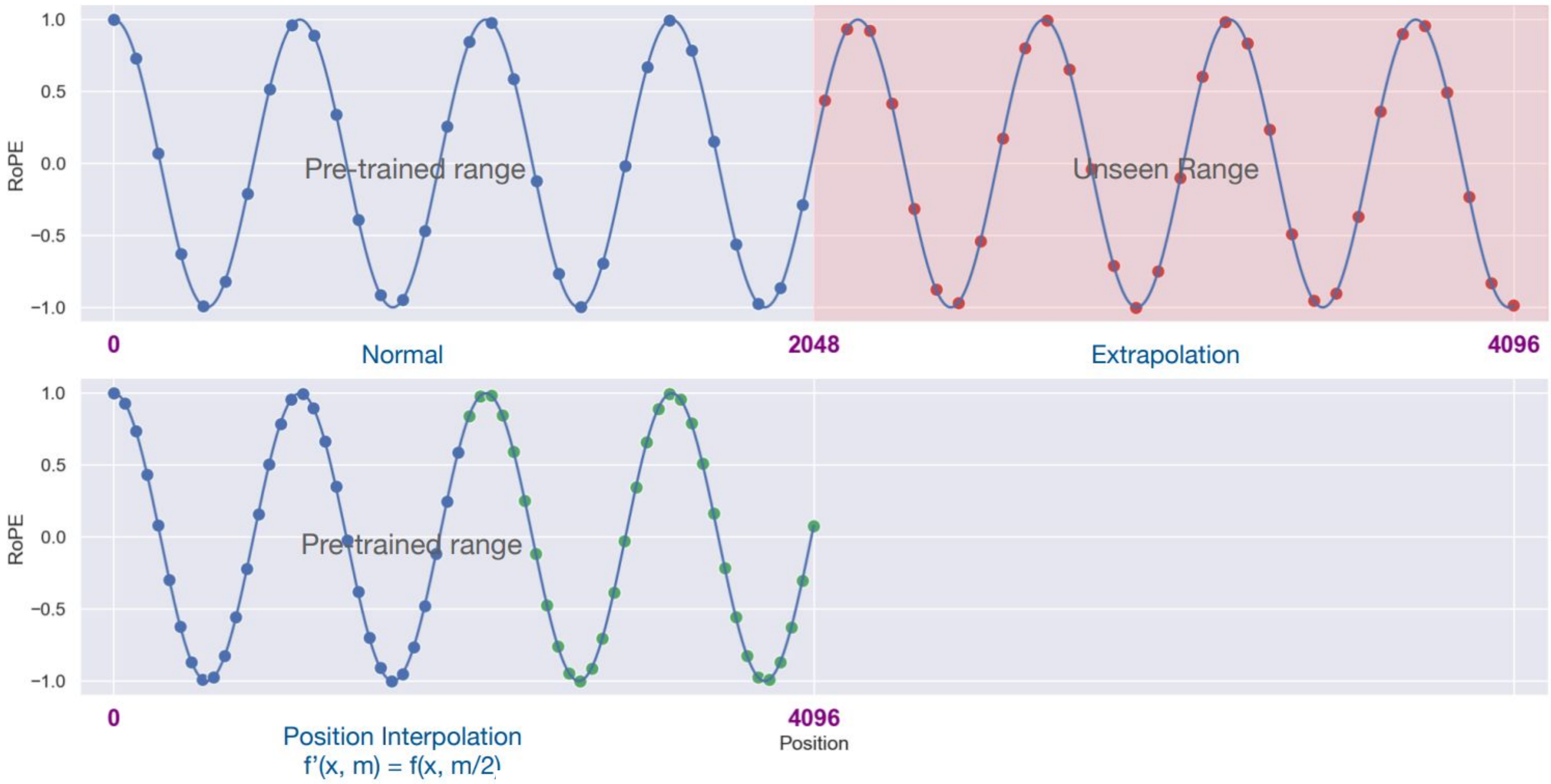
Direct application leads to unseen range (also decayed long term dependency)

Adaptation of RoPE Encoding [9]

[9] Chen et al. 2023. extending context window of large language models via positional interpolation

Adapting Positional Encoding

Interpolate the position into a smaller range (increasing RoPE frequency)



Adaptation of RoPE Encoding [9]

[9] Chen et al. 2023. extending context window of large language models via positional interpolation

Adapting Positional Encoding

Interpolate the position into a smaller range (increasing RoPE frequency b) $\theta_k = (1/b^{2(i-1)/d})$

RoPE Base ($\times 10^6$)	Long-Context							Short-Context
	Recall	RAG	Re-rank	ICL	QA	Summ.	Avg.	Avg.
0.5	25.8	37.0	4.4	73.8	17.5	16.3	29.1	65.0
4.0	81.3	47.8	18.2	76.5	31.8	36.3	48.7	65.3
8.0	96.0	54.9	29.4	73.9	35.7	37.9	**54.6**	**65.5**

Performance with Different RoPE frequency [10]

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

What are the gaps?

Format: Positional encodings only capture shorter context

- Absolute and relative: never learned long context positions
- RoPE: strong decay over distance

Distribution Shift:

- Pretraining data are mainly “short” documents
- Empirically no attentions across document boundary

Source of Intelligence: how do LLMs learn long-term dependency or global reasoning?

- Pretrained on next token prediction task solely
- How many next token prediction require information 128K tokens away?

[8] Overwijk et al. 2022. Clueweb22: 10 billion web documents with visual and semantic information.

Learning Long Context in Pretraining

Increase the fraction of longer text sequence in pretraining data

- Let model see more long sequences in pretraining
- Hope it naturally learns long context
    - LLMs do learn a lot of things from next token prediction

Where to get long pretraining sequences?

Source of Pretraining Sequences

Concatenating documents together?

Attention	Long-Context							Short-Context
	Recall	RAG	Re-rank	ICL	QA	Summ.	Avg.	Avg.
No doc masks	97.4	53.6	20.4	76.6	37.2	36.3	53.6	64.9
Document masks	96.0	54.9	29.4	73.9	35.7	37.9	**54.6**	**65.5**

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

Source of Pretraining Sequences

Concatenating documents together?

Attention	Long-Context							Short-Context
	Recall	RAG	Re-rank	ICL	QA	Summ.	Avg.	Avg.
No doc masks	97.4	53.6	20.4	76.6	37.2	36.3	53.6	64.9
Document masks	96.0	54.9	29.4	73.9	35.7	37.9	**54.6**	**65.5**

Attention cross document boundary hurts significantly [10]
Can be mitigated but still often underperforms

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

Source of Pretraining Sequences

Create synthetic data for long context tasks

Simple dictionary key-value retrieval (with an answer template)

```
Do a task using the list of dictionaries below.

Dictionary [1] {122: 765, 4548: 1475, 4818: 4782}
Dictionary [2] {526: 290, 9205: 9318, 9278: 1565}
...
Dictionary [32] {2931: 8364, 196: 1464, 812: 5363}
...
Dictionary [85] {344: 1579, 116: 617, 330: 411}

Above is a list of dictionaries such that each key and value is an integer. Report the
value of key 2931 and the dictionary it is in. Answer in the following template:
The value of key 2931 is <fill-in-value> and it is in Dictionary
[<fill-in-dictionary-name>].
```

```
Desired answer: The value of key 2931 is 8364 and it is in Dictionary [32].
```

Synthetic Key-value Retrieval Task to Fine-tune Long-context LLMs [11]

[11] Xiong et al. 2024. From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in LLMs by Finetuning on Synthetic Data

Source of Pretraining Sequences

Create synthetic data for long context tasks

Simple dictionary key-value retrieval (with an answer template)

```
Do a task using the list of dictionaries below.

Dictionary [1] {122: 765, 4548: 1475, 4818: 4782}
Dictionary [2] {526: 290, 9205: 9318, 9278: 1565}
...
Dictionary [32] {2931: 8364, 196: 1464, 812: 5363}
...
Dictionary [85] {344: 1579, 116: 617, 330: 411}

Above is a list of dictionaries such that each key and value is an integer. Report the
value of key 2931 and the dictionary it is in. Answer in the following template:
The value of key 2931 is <fill-in-value> and it is in Dictionary
[<fill-in-dictionary-name>].
```

```
Desired answer: The value of key 2931 is 8364 and it is in Dictionary [32].
```

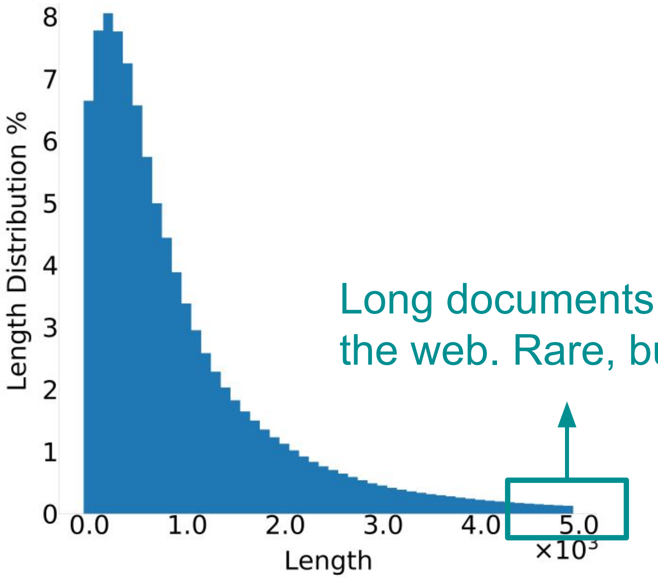
Synthetic Key-value Retrieval Task to Fine-tune Long-context LLMs [11]

Hard to believe this leads to general long-context ability.
Also may lead to model collapse (next lecture).

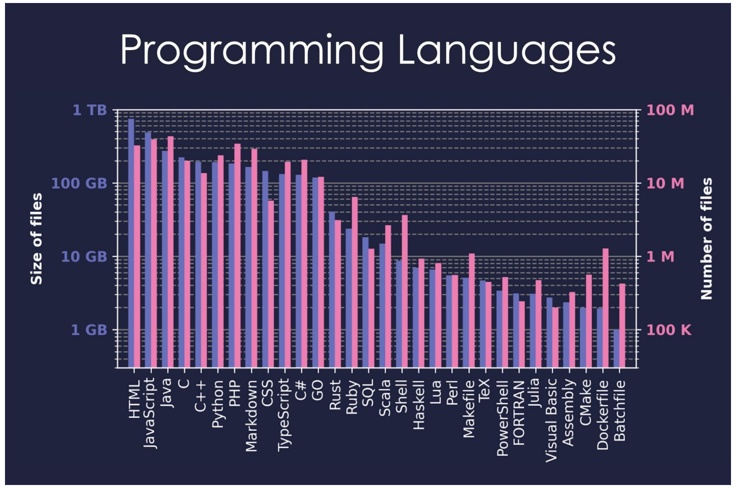
[11] Xiong et al. 2024. From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in LLMs by Finetuning on Synthetic Data

Source of Pretraining Sequences

Up sample long documents exist in organic data



Web Pages [8]



Code Repos [12]

Google books

Google Scholar

Other Long Documents

[12] Kocetkov et al. 2022. The Stack: 3 TB of permissively licensed source code.

Mixing Long Pretraining Data in Pretraining

Continue pretrain LLaMA on a mix of long texts [10]

Data	#Long tokens
Code Repos	**98.8B**
SP/Books	**33.2B**
SP/CC	15.3B
SP/Arxiv	5.2B
SP/GitHub	2.8B
SP/Wiki	0.1B
SP/StackEx	<0.1B
SP/C4	<0.1B

Long Data Mixture [10]

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

Mixing Long Pretraining Data in Pretraining

Continue pretrain LLaMA on a mix of long texts [10]

Data	#Long tokens
Code Repos	**98.8B**
SP/Books	**33.2B**
SP/CC	15.3B
SP/Arxiv	5.2B
SP/GitHub	2.8B
SP/Wiki	0.1B
SP/StackEx	<0.1B
SP/C4	<0.1B

Long Data Mixture [10]

Long Data (60%)	Long-Context						
	Recall	RAG	Re-rank	ICL	QA	Summ.	Avg.
CommonCrawl	84.1	53.3	28.1	67.5	35.2	37.0	50.9
Books	94.9	53.9	**30.7**	72.2	33.2	37.7	53.8
Code Repos	**99.2**	53.8	29.0	61.2	34.7	36.2	52.3
Books/Repos 1:1	96.0	**54.9**	29.4	**73.9**	**35.7**	**37.9**	**54.6**

Performance when continue pretrained on long data mixture [10]

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

Mixing Long Pretraining Data in Pretraining

Continue pretrain LLaMA on a mix of long texts [10]

Data	#Long tokens
Code Repos	**98.8B**
SP/Books	**33.2B**
SP/CC	15.3B
SP/Arxiv	5.2B
SP/GitHub	2.8B
SP/Wiki	0.1B
SP/StackEx	<0.1B
SP/C4	<0.1B

Long Data Mixture [10]

Long Data (60%)	Long-Context							Short-Context
	Recall	RAG	Re-rank	ICL	QA	Summ.	Avg.	Avg.
CommonCrawl	84.1	53.3	28.1	67.5	35.2	37.0	50.9	66.5
Books	94.9	53.9	**30.7**	72.2	33.2	37.7	53.8	65.5
Code Repos	**99.2**	53.8	29.0	61.2	34.7	36.2	52.3	65.9
Books/Repos 1:1	96.0	**54.9**	29.4	**73.9**	**35.7**	**37.9**	**54.6**	65.5

Performance when continue pretrained on long data mixture [10]

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

Mixing Long Pretraining Data in Pretraining

Continue pretrain LLaMA on a mix of long texts [10]

Data	#Long tokens
Code Repos	**98.8B**
SP/Books	**33.2B**
SP/CC	15.3B
SP/Arxiv	5.2B
SP/GitHub	2.8B
SP/Wiki	0.1B
SP/StackEx	<0.1B
SP/C4	<0.1B

Long Data Mixture [10]

Long Data (60%)	Long-Context							Short-Context
	Recall	RAG	Re-rank	ICL	QA	Summ.	Avg.	Avg.
CommonCrawl	84.1	53.3	28.1	67.5	35.2	37.0	50.9	66.5
Books	94.9	53.9	**30.7**	72.2	33.2	37.7	53.8	65.5
Code Repos	**99.2**	53.8	29.0	61.2	34.7	36.2	52.3	65.9
Books/Repos 1:1	96.0	**54.9**	29.4	**73.9**	**35.7**	**37.9**	**54.6**	65.5

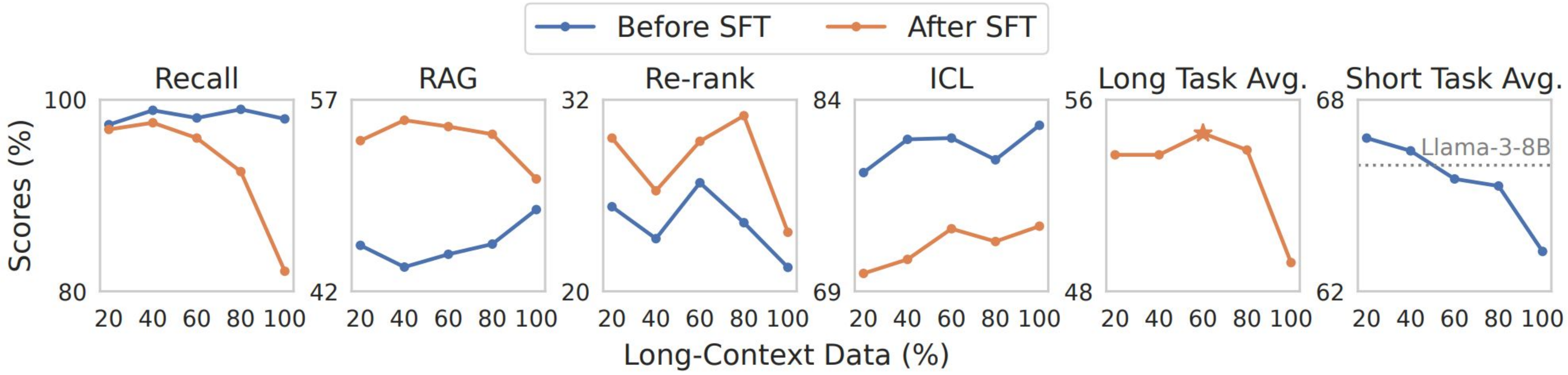
Performance when continue pretrained on long data mixture [10]

Improved on long-context tasks, but hurt on standard LLM tasks

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

Mixing Long Pretraining Data in Pretraining

Mixing back the standard pretraining data



Performance when continue pretrained with different fraction of long data [10]

Long-context Training Receipt

Start from an open-source LLM, then continue pretrain

Manual?
Data Mixing

Continued Long-context Training		
Data	30% code repos, 30% books, 3% textbooks, 37% ShortMix	
	ShortMix:	27% FineWeb-Edu, 27% FineWeb, 11% Tulu-v2, 11% StackExchange, 8% Wikipedia, 8% OpenWebMath, 8% ArXiv
Length Curriculum	Stage 1 (64K):	Code repos, books, and textbooks at length 64K
	Stage 2 (512K):	Code repos: 50% at length 512K, 50% at length 64K Books: 17% at length 512K, 83% at length 64K Textbooks at length 512K
Steps	Stage 1: 20B tokens (2.2K H100 hours), Stage 2: 20B tokens (12.2K H100 hours)	
Model	Initialization:	Llama-3-8B-Instruct (original RoPE base freq. 5×10^5)
	RoPE:	Stage 1: 8×10^6, Stage 2: 1.28×10^8
	Attention:	Full attention with cross-document attention masking
Optim.	AdamW (weight decay = 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.95$)	
	LR:	$1e-5$ with 10% warmup and cosine decay to $1e-6$, each stage
	Batch size:	4M tokens for stage 1, 8M tokens for stage 2

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

Long-context Training Receipt

Start from an open-source LLM, then continue pretrain

Continued Long-context Training			
Data	30% code repos, 30% books, 3% textbooks, 37% ShortMix		
	ShortMix:	27% FineWeb-Edu, 27% FineWeb, 11% Tulu-v2, 11% StackExchange, 8% Wikipedia, 8% OpenWebMath, 8% ArXiv	
Length Curriculum	Stage 1 (64K):	Code repos, books, and textbooks at length 64K	
	Stage 2 (512K):	Code repos: 50% at length 512K, 50% at length 64K Books: 17% at length 512K, 83% at length 64K Textbooks at length 512K	
Steps	Stage 1: 20B tokens (2.2K H100 hours), Stage 2: 20B tokens (12.2K H100 hours)		
Model	Initialization:	Llama-3-8B-Instruct (original RoPE base freq. 5×10^5)	
	RoPE:	Stage 1: 8×10^6, Stage 2: 1.28×10^8	
	Attention:	Full attention with cross-document attention masking	
Optim.	AdamW (weight decay = 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.95$)		
	LR:	$1e-5$ with 10% warmup and cosine decay to $1e-6$, each stage	
	Batch size:	4M tokens for stage 1, 8M tokens for stage 2	

Manual?
Data Mixing

Curriculum leaning to grow the length

Long-context Training Receipt

Start from an open-source LLM, then continue pretrain

Manual? Data Mixing

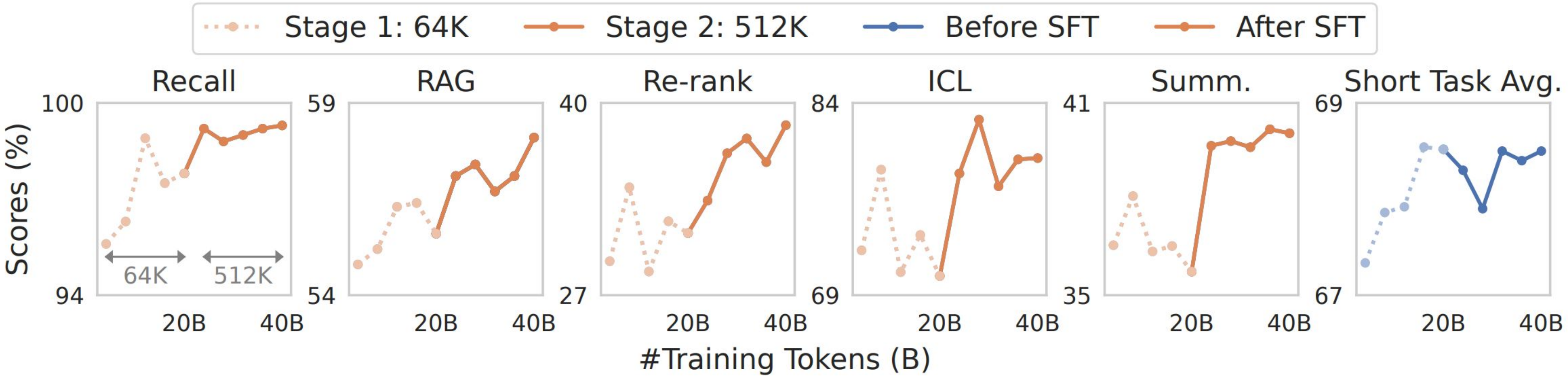
Curriculum leaning to grow the length

Standard Continue Pretraining

Continued Long-context Training		
Data	30% code repos, 30% books, 3% textbooks, 37% ShortMix	
	ShortMix:	27% FineWeb-Edu, 27% FineWeb, 11% Tulu-v2, 11% StackExchange, 8% Wikipedia, 8% OpenWebMath, 8% ArXiv
Length Curriculum	Stage 1 (64K):	Code repos, books, and textbooks at length 64K
	Stage 2 (512K):	Code repos: 50% at length 512K, 50% at length 64K Books: 17% at length 512K, 83% at length 64K Textbooks at length 512K
Steps	Stage 1: 20B tokens (2.2K H100 hours), Stage 2: 20B tokens (12.2K H100 hours)	
Model	Initialization: RoPE: Attention:	Llama-3-8B-Instruct (original RoPE base freq. 5×10^5) Stage 1: 8×10^6, Stage 2: 1.28×10^8 Full attention with cross-document attention masking
Optim.	AdamW (weight decay = 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.95$)	
	LR: Batch size:	$1e-5$ with 10% warmup and cosine decay to $1e-6$, each stage 4M tokens for stage 1, 8M tokens for stage 2

Long-context Training Performance

Improved long-context ability with maintained short task performance



Performance with Long-context Continue Pretraining [10]

[10] Gao et al. 2024. How to Train Long-Context Language Models (Effectively)

slido

Please download and install the Slido app on all computers you use

Why only need to do long-context training in continuous pretraining but not pretraining from scratch?

ⓘ Start presenting to display the poll results on this slide.

Why Long-context?

Many scenarios naturally needs long inputs. 4K token is not enough

- Chatbot: long conversation history
- RAG: lots of retrieved documents
- Code: large repository
- Fancy Prompts: can be very long

Ideally, a lot of imagination towards AGI

- Short term memory
- Long term reasoning across multiple text pieces
- Global understanding
- Bring the AGI power of LLMs to all the above

[4] Hsieh et al. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models?

How are Long-Context Ability Evaluated

Category	Dataset	Metrics	Description
Retrieval-augmented generation	Natural Questions	SubEM	Factoid question answering
	TriviaQA	SubEM	Trivia question answering
	PopQA	SubEM	Long-tail entity question answering
	HotpotQA	SubEM	Multi-hop question answering
Passage re-ranking	MS MARCO	NDCG@10	Rerank passage for a query
Generation with citations	ALCE ASQA	Recall, Cite	Answer ambiguous questions with citations
	ALCE Qampari	Recall, Cite	Answer factoid questions with citations
Long-document QA	NarrativeQA	Model-based	Book and movie script QA
	∞BENCH QA	ROUGE F1	Novel QA with entity replacement
	∞BENCH MC	Accuracy	Novel multiple-choice QA with entity replacement
Summarization	∞BENCH Sum	Model-based	Novel summarization with entity replacement
	Multi-LexSum	Model-based	Summarizing multiple legal documents
Many-shot in-context learning	TREC Coarse	Accuracy	Question type classification, 6 labels
	TREC Fine	Accuracy	Question type classification, 50 labels
	NLU	Accuracy	Task intent classification, 68 labels
	BANKING77	Accuracy	Banking intent classification, 77 labels
	CLINC150	Accuracy	Intent classification, 151 labels
Synthetic recall	JSON KV	SubEM	Retrieve a key in JSON dictionary
	RULER MK Needle	SubEM	Retrieve the needle (a number) within noisy needles
	RULER MK UUID	SubEM	Retrieve the needle (a UUID) within noisy needles
	RULER MV	SubEM	Retrieve multiple values for one needle (key)

Long-Context Evaluation Tasks [13]

[13] Yen et al. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly

How are Long-Context Ability Evaluated

Category	Dataset	Metrics	Description
Retrieval-augmented generation	Natural Questions	SubEM	Factoid question answering
	TriviaQA	SubEM	Trivia question answering
	PopQA	SubEM	Long-tail entity question answering
	HotpotQA	SubEM	Multi-hop question answering
Passage re-ranking	MS MARCO	NDCG@10	Rerank passage for a query
Generation with citations	ALCE ASQA	Recall, Cite	Answer ambiguous questions with citations
	ALCE Qampari	Recall, Cite	Answer factoid questions with citations
Long-document QA	NarrativeQA	Model-based	Book and movie script QA
	∞BENCH QA	ROUGE F1	Novel QA with entity replacement
	∞BENCH MC	Accuracy	Novel multiple-choice QA with entity replacement
Summarization	∞BENCH Sum	Model-based	Novel summarization with entity replacement
	Multi-LexSum	Model-based	Summarizing multiple legal documents
Many-shot in-context learning	TREC Coarse	Accuracy	Question type classification, 6 labels
	TREC Fine	Accuracy	Question type classification, 50 labels
	NLU	Accuracy	Task intent classification, 68 labels
	BANKING77	Accuracy	Banking intent classification, 77 labels
	CLINC150	Accuracy	Intent classification, 151 labels
Synthetic recall	JSON KV	SubEM	Retrieve a key in JSON dictionary
	RULER MK Needle	SubEM	Retrieve the needle (a number) within noisy needles
	RULER MK UUID	SubEM	Retrieve the needle (a UUID) within noisy needles
	RULER MV	SubEM	Retrieve multiple values for one needle (key)

Long-Context Evaluation Tasks [13]

How many of them are unique to long-context ability?

[13] Yen et al. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly

How are Long-Context Ability Evaluated

Category	Dataset	Metrics	Description
Retrieval-augmented generation	Natural Questions	SubEM	Factoid question answering
	TriviaQA	SubEM	Trivia question answering
	PopQA	SubEM	Long-tail entity question answering
	HotpotQA	SubEM	Multi-hop question answering
Passage re-ranking	MS MARCO	NDCG@10	Rerank passage for a query
Generation with citations	ALCE ASQA	Recall, Cite	Answer ambiguous questions with citations
	ALCE Qampari	Recall, Cite	Answer factoid questions with citations
Long-document QA	NarrativeQA	Model-based	Book and movie script QA
	∞BENCH QA	ROUGE F1	Novel QA with entity replacement
	∞BENCH MC	Accuracy	Novel multiple-choice QA with entity replacement
Summarization	∞BENCH Sum	Model-based	Novel summarization with entity replacement
	Multi-LexSum	Model-based	Summarizing multiple legal documents
Many-shot in-context learning	TREC Coarse	Accuracy	Question type classification, 6 labels
	TREC Fine	Accuracy	Question type classification, 50 labels
	NLU	Accuracy	Task intent classification, 68 labels
	BANKING77	Accuracy	Banking intent classification, 77 labels
	CLINC150	Accuracy	Intent classification, 151 labels
Synthetic recall	JSON KV	SubEM	Retrieve a key in JSON dictionary
	RULER MK Needle	SubEM	Retrieve the needle (a number) within noisy needles
	RULER MK UUID	SubEM	Retrieve the needle (a UUID) within noisy needles
	RULER MV	SubEM	Retrieve multiple values for one needle (key)

Long-Context Evaluation Tasks [13]

How many of them are unique to long-context ability?

How many cannot be solved by divide-and-conquer?

[13] Yen et al. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly

Performance of Current LLMs on Real Tasks

Model	Recall 8k	Recall 16k	Recall 32k	Recall 64k	Recall 128k	RAG 8k	RAG 16k	RAG 32k	RAG 64k	RAG 128k
GPT-4	99.5	93.5	93.1	88.6	72.8	75.3	73.6	70.9	68.1	65.0
GPT-4o-05	94.7	93.4	91.2	87.9	81.6	74.1	73.1	71.8	71.1	71.0
GPT-4o-08	99.8	99.4	97.9	97.0	97.0	73.4	73.8	72.4	71.1	70.8
GPT-4o-mini	100.0	99.8	99.1	92.0	83.6	72.6	71.0	69.6	68.3	66.7
Claude-3.5-sonnet	99.9	97.2	96.2	95.2	93.3	60.4	52.8	51.1	39.8	41.1
Gemini-1.5-Flash	93.5	93.6	93.2	92.5	87.8	71.6	69.9	69.6	68.6	67.6
Gemini-1.5-Pro	81.3	83.6	86.9	87.1	84.1	73.0	72.9	71.6	71.9	70.9
Llama-3.1-8B	99.4	99.6	97.2	98.3	91.1	69.1	67.9	64.8	64.6	59.0
Llama-3.1-70B	99.9	99.8	98.0	87.4	84.4	73.0	72.2	71.5	70.3	55.8
Mistral-Nemo	93.6	83.3	52.3	21.5	12.1	68.4	63.6	56.9	47.6	39.9
MegaBeam-Mistral	93.9	90.0	81.6	83.6	76.0	62.6	62.6	61.8	57.4	55.2
Phi-3-mini-128k	90.3	84.9	81.1	80.1	42.3	61.2	60.6	57.9	55.7	46.0
Phi-3-small-128k	91.0	89.3	73.5	66.7	59.0	66.5	65.8	62.5	61.3	58.1
Phi-3-med-128k	76.1	70.6	62.5	51.8	14.4	65.3	64.5	62.7	56.9	45.2
Phi-3.5-mini	95.8	90.7	83.1	77.2	40.7	59.8	57.9	55.6	51.0	41.4
Jamba-1.5-Mini	87.3	85.6	85.0	79.7	76.8	66.2	65.0	64.0	63.4	56.6
ProLong	96.6	95.8	93.0	92.9	85.5	68.8	69.3	66.6	66.5	64.8

[13] Yen et al. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly

Performance of Current LLMs on Real Tasks

	Recall					RAG					Cite					Re-rank				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	99.5	93.5	93.1	88.6	72.8	75.3	73.6	70.9	68.1	65.0	43.8	45.2	28.8	3.6	3.1	76.4	72.3	63.9	37.8	16.8
GPT-4o-05	94.7	93.4	91.2	87.9	81.6	74.1	73.1	71.8	71.1	71.0	43.7	44.2	44.1	44.1	40.6	74.4	74.3	67.2	56.9	46.8
GPT-4o-08	99.8	99.4	97.9	97.0	97.0	73.4	73.8	72.4	71.1	70.8	45.8	47.1	46.4	45.7	45.3	75.6	73.1	67.4	59.5	47.9
GPT-4o-mini	100.0	99.8	99.1	92.0	83.6	72.6	71.0	69.6	68.3	66.7	36.1	33.7	31.3	28.0	24.5	68.9	65.2	56.4	40.5	30.5
Claude-3.5-sonnet	99.9	97.2	96.2	95.2	93.3	60.4	52.8	51.1	39.8	41.1	36.7	32.9	30.5	26.4	12.5	76.3	46.1	36.0	14.5	9.1
Gemini-1.5-Flash	93.5	93.6	93.2	92.5	87.8	71.6	69.9	69.6	68.6	67.6	48.4	46.6	43.0	36.7	29.0	75.1	73.9	68.9	59.3	50.7
Gemini-1.5-Pro	81.3	83.6	86.9	87.1	84.1	73.0	72.9	71.6	71.9	70.9	47.1	43.0	44.7	45.1	42.5	75.8	73.2	71.7	65.9	58.6
Llama-3.1-8B	99.4	99.6	97.2	98.3	91.1	69.1	67.9	64.8	64.6	59.0	35.4	26.9	12.6	12.8	3.4	58.7	45.9	42.0	31.9	15.0
Llama-3.1-70B	99.9	99.8	98.0	87.4	84.4	73.0	72.2	71.5	70.3	55.8	44.5	42.1	39.5	30.9	7.6	73.3	69.7	58.4	40.0	19.4
Mistral-Nemo	93.6	83.3	52.3	21.5	12.1	68.4	63.6	56.9	47.6	39.9	33.7	8.6	3.7	1.3	0.5	56.8	46.0	13.1	0.0	0.0
MegaBeam-Mistral	93.9	90.0	81.6	83.6	76.0	62.6	62.6	61.8	57.4	55.2	22.3	13.8	9.7	4.5	4.0	49.9	36.2	34.2	21.7	15.9
Phi-3-mini-128k	90.3	84.9	81.1	80.1	42.3	61.2	60.6	57.9	55.7	46.0	22.8	16.9	9.3	2.7	0.8	44.1	28.7	25.6	16.6	5.8
Phi-3-small-128k	91.0	89.3	73.5	66.7	59.0	66.5	65.8	62.5	61.3	58.1	18.9	15.9	8.9	4.6	2.9	38.3	32.1	28.1	17.2	6.5
Phi-3-med-128k	76.1	70.6	62.5	51.8	14.4	65.3	64.5	62.7	56.9	45.2	39.1	27.1	10.2	5.8	3.3	43.2	33.3	25.5	11.9	5.8
Phi-3.5-mini	95.8	90.7	83.1	77.2	40.7	59.8	57.9	55.6	51.0	41.4	22.1	17.2	7.1	2.0	2.5	42.4	29.6	23.2	18.0	9.1
Jamba-1.5-Mini	87.3	85.6	85.0	79.7	76.8	66.2	65.0	64.0	63.4	56.6	15.4	10.0	5.7	3.1	2.5	53.5	43.0	35.6	23.2	14.6
ProLong	96.6	95.8	93.0	92.9	85.5	68.8	69.3	66.6	66.5	64.8	33.7	24.0	11.0	2.3	1.2	53.8	43.9	39.3	33.3	25.0

Performance of Current LLMs on Real Tasks

	LongQA					Summ				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	47.8	48.1	49.5	47.6	45.7	28.0	29.6	33.6	36.2	35.8
GPT-4o-05	45.2	46.9	51.8	56.0	61.4	29.5	34.9	40.7	42.1	45.1
GPT-4o-08	40.4	45.7	50.1	53.3	57.4	29.0	35.7	40.9	41.8	43.6
GPT-4o-mini	37.8	39.6	46.0	51.0	52.2	28.3	33.8	36.7	39.2	40.8
Claude-3.5-sonnet	29.8	25.8	32.5	18.2	19.9	27.4	33.0	40.1	37.2	40.7
Gemini-1.5-Flash	33.1	40.2	44.9	51.9	57.9	25.2	30.0	33.2	36.2	39.9
Gemini-1.5-Pro	33.9	40.2	48.9	53.4	61.7	30.1	32.6	39.3	45.5	45.0
Llama-3.1-8B	24.6	32.3	39.4	43.3	45.6	23.2	25.7	29.2	30.2	31.5
Llama-3.1-70B	31.5	38.9	45.8	55.4	58.4	27.7	31.7	35.5	35.8	35.7
Mistral-Nemo	32.2	32.0	26.8	24.6	24.2	26.0	23.2	25.3	21.5	20.1
MegaBeam-Mistral	23.1	30.6	34.0	36.3	34.7	21.6	24.3	27.9	30.4	29.4
Phi-3-mini-128k	24.4	31.7	31.6	34.1	27.7	20.8	24.2	26.9	28.1	28.8
Phi-3-small-128k	22.8	29.0	33.4	40.3	36.4	18.1	20.7	25.4	25.3	26.6
Phi-3-med-128k	22.6	21.5	20.4	20.4	27.1	23.1	24.5	25.4	31.4	28.6
Phi-3.5-mini	25.2	28.3	29.9	31.2	27.8	21.2	23.9	24.6	29.4	27.9
Jamba-1.5-Mini	34.8	41.1	46.2	52.2	52.7	18.0	19.0	19.1	19.6	19.3
ProLong	27.2	34.1	36.2	42.3	43.5	21.1	25.9	26.1	27.2	29.0

[13] Yen et al. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly

Performance of Current LLMs on Real Tasks

	LongQA					Summ					ICL				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	47.8	48.1	49.5	47.6	45.7	28.0	29.6	33.6	36.2	35.8	78.6	81.6	70.2	55.0	40.1
GPT-4o-05	45.2	46.9	51.8	56.0	61.4	29.5	34.9	40.7	42.1	45.1	77.8	79.2	76.2	65.4	47.7
GPT-4o-08	40.4	45.7	50.1	53.3	57.4	29.0	35.7	40.9	41.8	43.6	83.2	85.6	88.8	86.6	84.6
GPT-4o-mini	37.8	39.6	46.0	51.0	52.2	28.3	33.8	36.7	39.2	40.8	73.2	77.8	80.0	79.6	80.0
Claude-3.5-sonnet	29.8	25.8	32.5	18.2	19.9	27.4	33.0	40.1	37.2	40.7	86.0	87.8	88.4	89.6	59.8
Gemini-1.5-Flash	33.1	40.2	44.9	51.9	57.9	25.2	30.0	33.2	36.2	39.9	70.4	67.0	53.6	39.5	21.9
Gemini-1.5-Pro	33.9	40.2	48.9	53.4	61.7	30.1	32.6	39.3	45.5	45.0	77.3	79.6	81.0	79.5	76.3
Llama-3.1-8B	24.6	32.3	39.4	43.3	45.6	23.2	25.7	29.2	30.2	31.5	69.8	74.6	77.0	78.6	83.4
Llama-3.1-70B	31.5	38.9	45.8	55.4	58.4	27.7	31.7	35.5	35.8	35.7	71.6	74.2	77.0	79.4	83.6
Mistral-Nemo	32.2	32.0	26.8	24.6	24.2	26.0	23.2	25.3	21.5	20.1	66.8	75.4	80.0	81.6	84.4
MegaBeam-Mistral	23.1	30.6	34.0	36.3	34.7	21.6	24.3	27.9	30.4	29.4	72.0	77.2	78.4	82.8	85.0
Phi-3-mini-128k	24.4	31.7	31.6	34.1	27.7	20.8	24.2	26.9	28.1	28.8	61.6	71.4	73.2	77.0	80.0
Phi-3-small-128k	22.8	29.0	33.4	40.3	36.4	18.1	20.7	25.4	25.3	26.6	67.6	73.2	79.2	82.6	84.2
Phi-3-med-128k	22.6	21.5	20.4	20.4	27.1	23.1	24.5	25.4	31.4	28.6	58.8	61.2	70.6	72.4	72.0
Phi-3.5-mini	25.2	28.3	29.9	31.2	27.8	21.2	23.9	24.6	29.4	27.9	61.2	69.0	74.2	77.8	78.4
Jamba-1.5-Mini	34.8	41.1	46.2	52.2	52.7	18.0	19.0	19.1	19.6	19.3	77.6	82.0	85.6	88.4	91.2
ProLong	27.2	34.1	36.2	42.3	43.5	21.1	25.9	26.1	27.2	29.0	66.4	72.2	78.0	81.4	84.0

[13] Yen et al. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly

Performance of Current LLMs on Real Tasks

	LongQA					Summ					ICL					Avg.				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	47.8	48.1	49.5	47.6	45.7	28.0	29.6	33.6	36.2	35.8	78.6	81.6	70.2	55.0	40.1	64.2	63.4	58.6	48.1	39.9
GPT-4o-05	45.2	46.9	51.8	56.0	61.4	29.5	34.9	40.7	42.1	45.1	77.8	79.2	76.2	65.4	47.7	62.8	63.7	63.3	60.5	56.3
GPT-4o-08	40.4	45.7	50.1	53.3	57.4	29.0	35.7	40.9	41.8	43.6	83.2	85.6	88.8	86.6	84.6	63.9	65.8	66.3	65.0	63.8
GPT-4o-mini	37.8	39.6	46.0	51.0	52.2	28.3	33.8	36.7	39.2	40.8	73.2	77.8	80.0	79.6	80.0	59.5	60.1	59.9	57.0	54.1
Claude-3.5-sonnet	29.8	25.8	32.5	18.2	19.9	27.4	33.0	40.1	37.2	40.7	86.0	87.8	88.4	89.6	59.8	59.5	53.7	53.5	45.8	39.5
Gemini-1.5-Flash	33.1	40.2	44.9	51.9	57.9	25.2	30.0	33.2	36.2	39.9	70.4	67.0	53.6	39.5	21.9	59.6	60.2	58.1	55.0	50.7
Gemini-1.5-Pro	33.9	40.2	48.9	53.4	61.7	30.1	32.6	39.3	45.5	45.0	77.3	79.6	81.0	79.5	76.3	59.8	60.7	63.5	64.1	62.7
Llama-3.1-8B	24.6	32.3	39.4	43.3	45.6	23.2	25.7	29.2	30.2	31.5	69.8	74.6	77.0	78.6	83.4	54.3	53.3	51.7	51.4	47.0
Llama-3.1-70B	31.5	38.9	45.8	55.4	58.4	27.7	31.7	35.5	35.8	35.7	71.6	74.2	77.0	79.4	83.6	60.2	61.2	60.8	57.0	49.3
Mistral-Nemo	32.2	32.0	26.8	24.6	24.2	26.0	23.2	25.3	21.5	20.1	66.8	75.4	80.0	81.6	84.4	53.9	47.4	36.9	28.3	25.9
MegaBeam-Mistral	23.1	30.6	34.0	36.3	34.7	21.6	24.3	27.9	30.4	29.4	72.0	77.2	78.4	82.8	85.0	49.3	47.8	46.8	45.2	42.9
Phi-3-mini-128k	24.4	31.7	31.6	34.1	27.7	20.8	24.2	26.9	28.1	28.8	61.6	71.4	73.2	77.0	80.0	46.4	45.5	43.7	42.0	33.1
Phi-3-small-128k	22.8	29.0	33.4	40.3	36.4	18.1	20.7	25.4	25.3	26.6	67.6	73.2	79.2	82.6	84.2	46.2	46.6	44.4	42.6	39.1
Phi-3-med-128k	22.6	21.5	20.4	20.4	27.1	23.1	24.5	25.4	31.4	28.6	58.8	61.2	70.6	72.4	72.0	46.9	43.2	39.6	35.8	28.1
Phi-3.5-mini	25.2	28.3	29.9	31.2	27.8	21.2	23.9	24.6	29.4	27.9	61.2	69.0	74.2	77.8	78.4	46.8	45.2	42.5	41.0	32.5
Jamba-1.5-Mini	34.8	41.1	46.2	52.2	52.7	18.0	19.0	19.1	19.6	19.3	77.6	82.0	85.6	88.4	91.2	50.4	49.4	48.7	47.1	44.8
ProLong	27.2	34.1	36.2	42.3	43.5	21.1	25.9	26.1	27.2	29.0	66.4	72.2	78.0	81.4	84.0	52.5	52.2	50.0	49.4	47.6

[13] Yen et al. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly

Remarks

Solution: Mixing in longer organic data in a dedicated continuous pretraining

Same capability as in short-text, nothing more.

Many scenarios not as effective as divide-and-conquer solutions, but very convenient (though costly)

Remarks

Solution: Mixing in longer organic data in a dedicated continuous pretraining

Same capability as in short-text, nothing more.

Many scenarios not as effective as divide-and-conquer solutions, but very convenient (though costly)

Many questions remain unanswered:

- What scenarios require true long-context ability?
- What is true long-context ability?
- How can be obtain such long-context ability?
- Will scaling up go to lead us there?

slido

Please download and install the
Slido app on all computers you use

Audience Q&A

ⓘ Start presenting to display the audience questions on this slide.

Outline

Motivation

Probing Long Context Ability

Evaluation on Real Scenarios

Adapting LLMs to Long Context Tasks

Efficient Serving

Serving Extremely Long Contexts

Main bottleneck: Attention mechanism

- 1 million context length == 1 TB GPU memory!
- Very realistic length in specific scenarios
  - DNA sequences
  - Autonomous Driving

[14] Liu et al. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context

Serving Extremely Long Contexts

Main bottleneck: Attention mechanism

- 1 million context length == 1 TB GPU memory!
- Very realistic length in specific scenarios
  - DNA sequences
  - Autonomous Driving

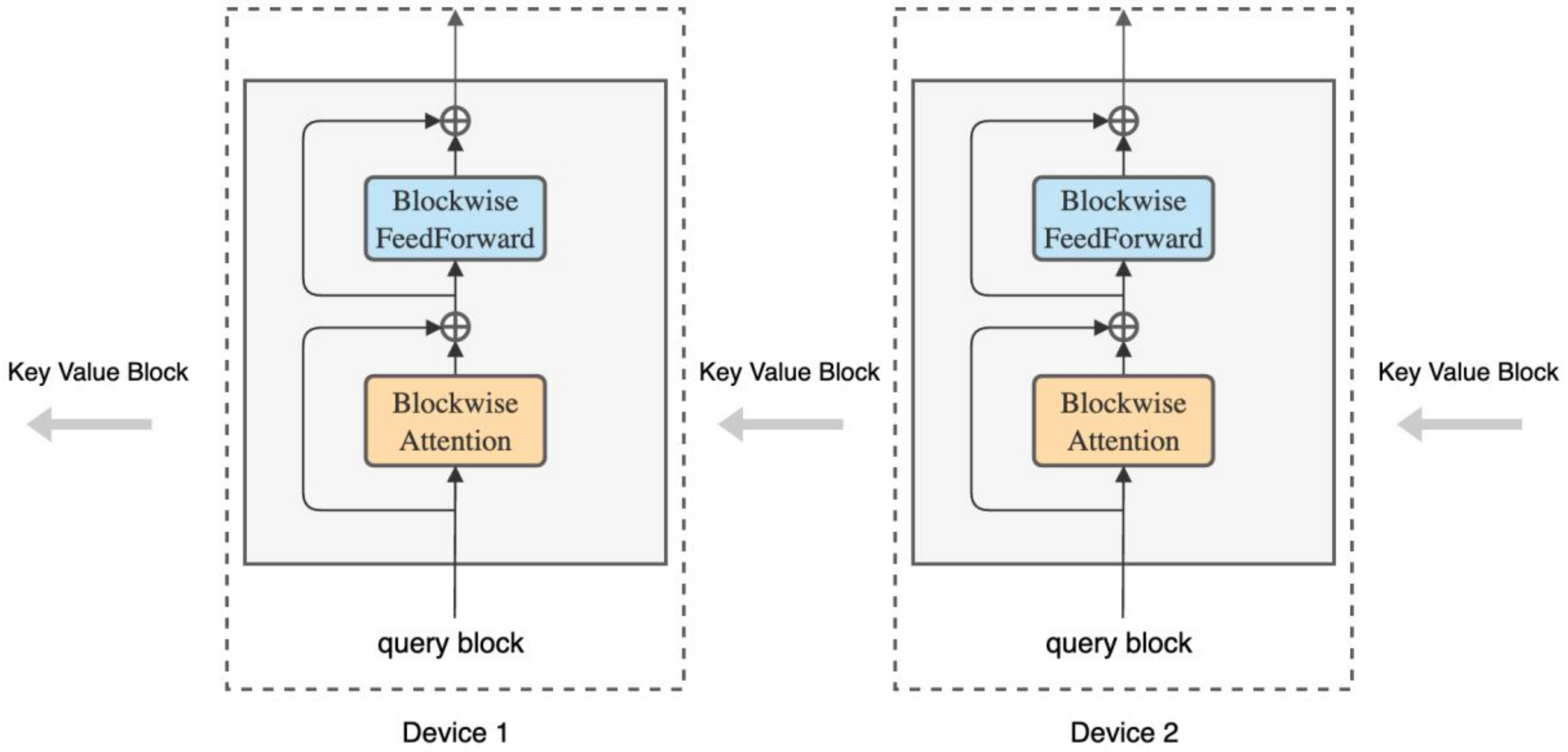
Many ways to approximate long-context attention

- Sparsity
- Recurrent
- Attention Sink

[14] Liu et al. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context

Efficient Serving: Ring Attention

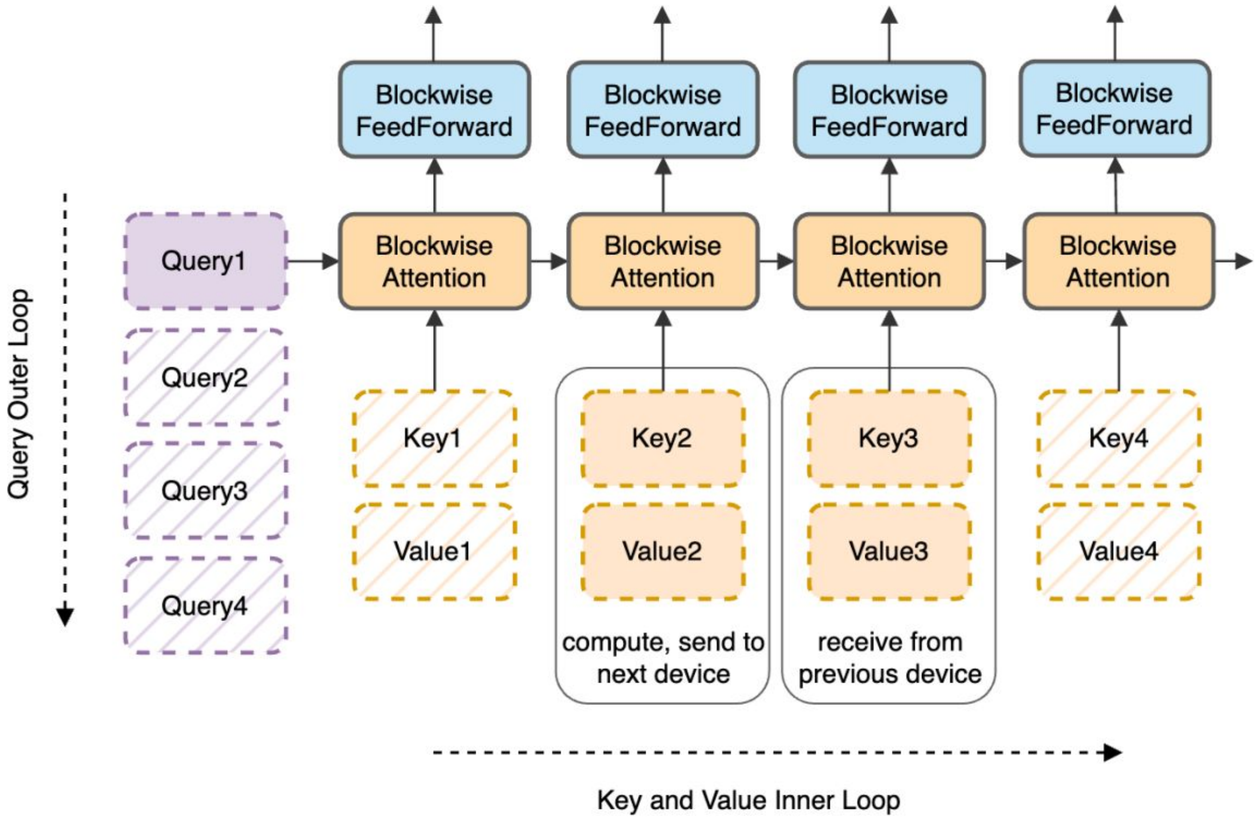
Compute Attention Block-Wise



[14] Liu et al. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context

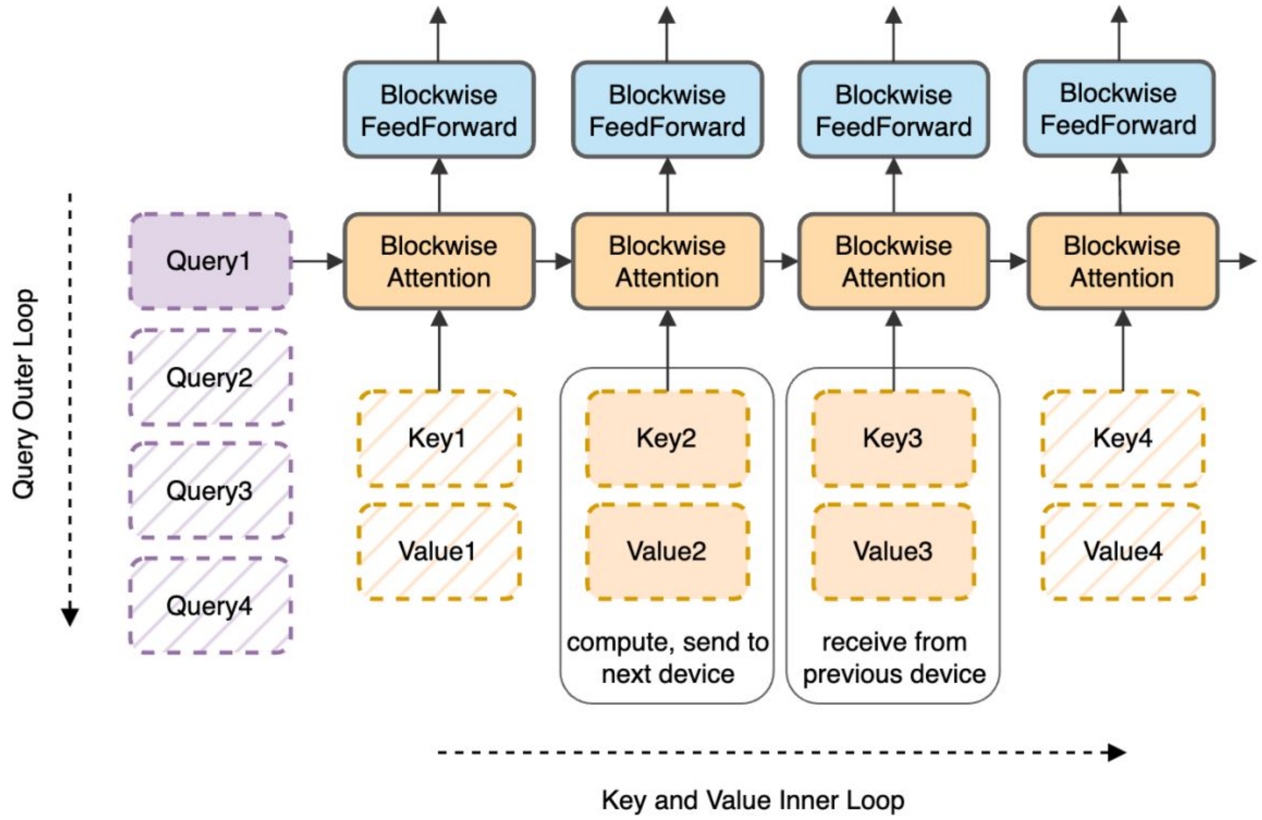
Efficient Serving: Ring Attention

Form a communication ring to pass KV blocks around



Efficient Serving: Ring Attention

Form a communication ring to pass KV blocks around

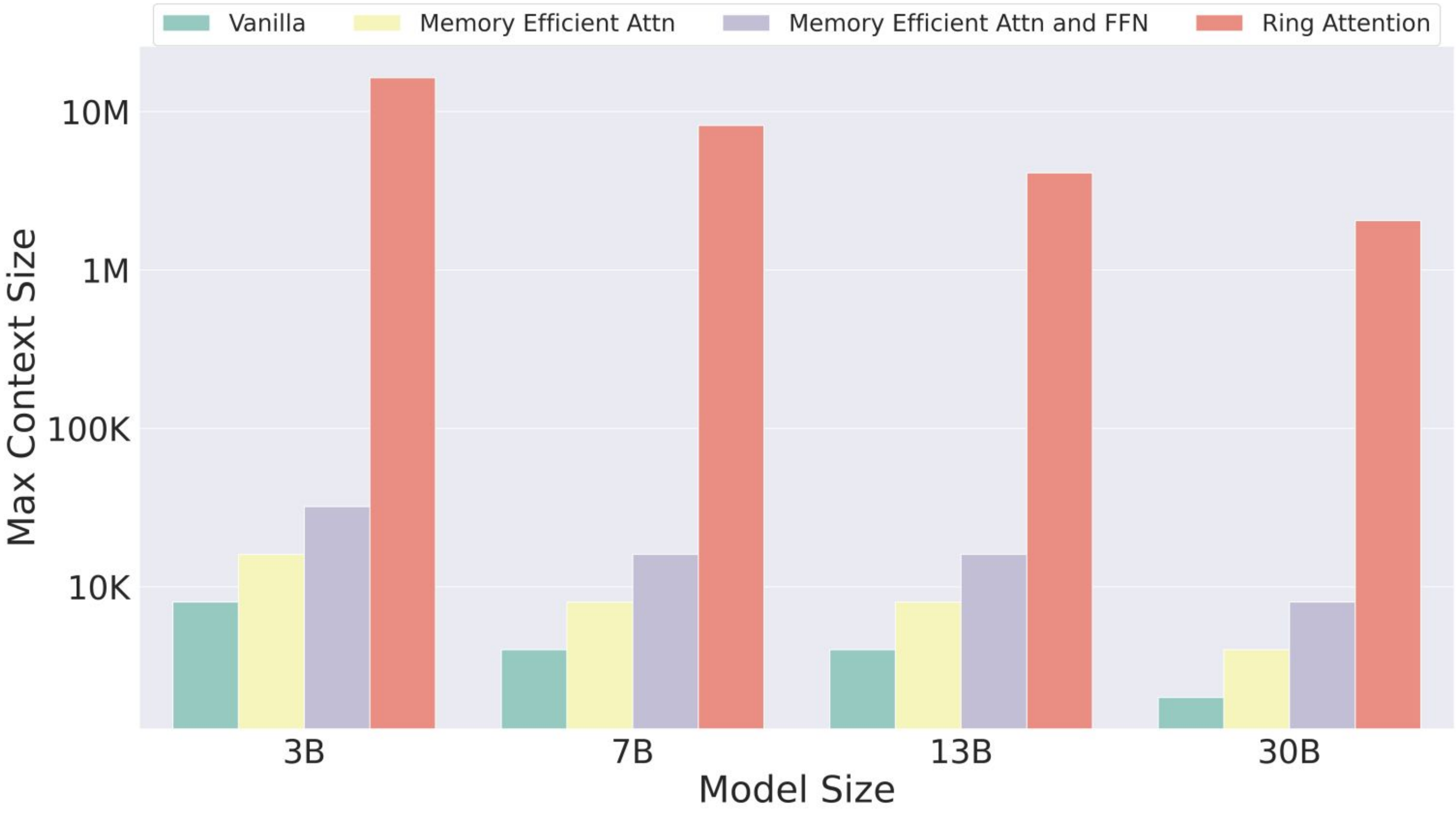


No specific order required as attention is a set wise operation
If communication < compute, then no extra latency

[14] Liu et al. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context

Efficient Serving: Ring Attention

Long-context parallelism for inference efficiency



Max Context-Length in Large Scale Pretraining [14]

[14] Liu et al. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context

slido

Please download and install the
Slido app on all computers you use

Audience Q&A

ⓘ Start presenting to display the audience questions on this slide.