Long Context Modeling in LLM Era

— Advances and Challenges

Speaker: Juntao Li & Zecheng Tang

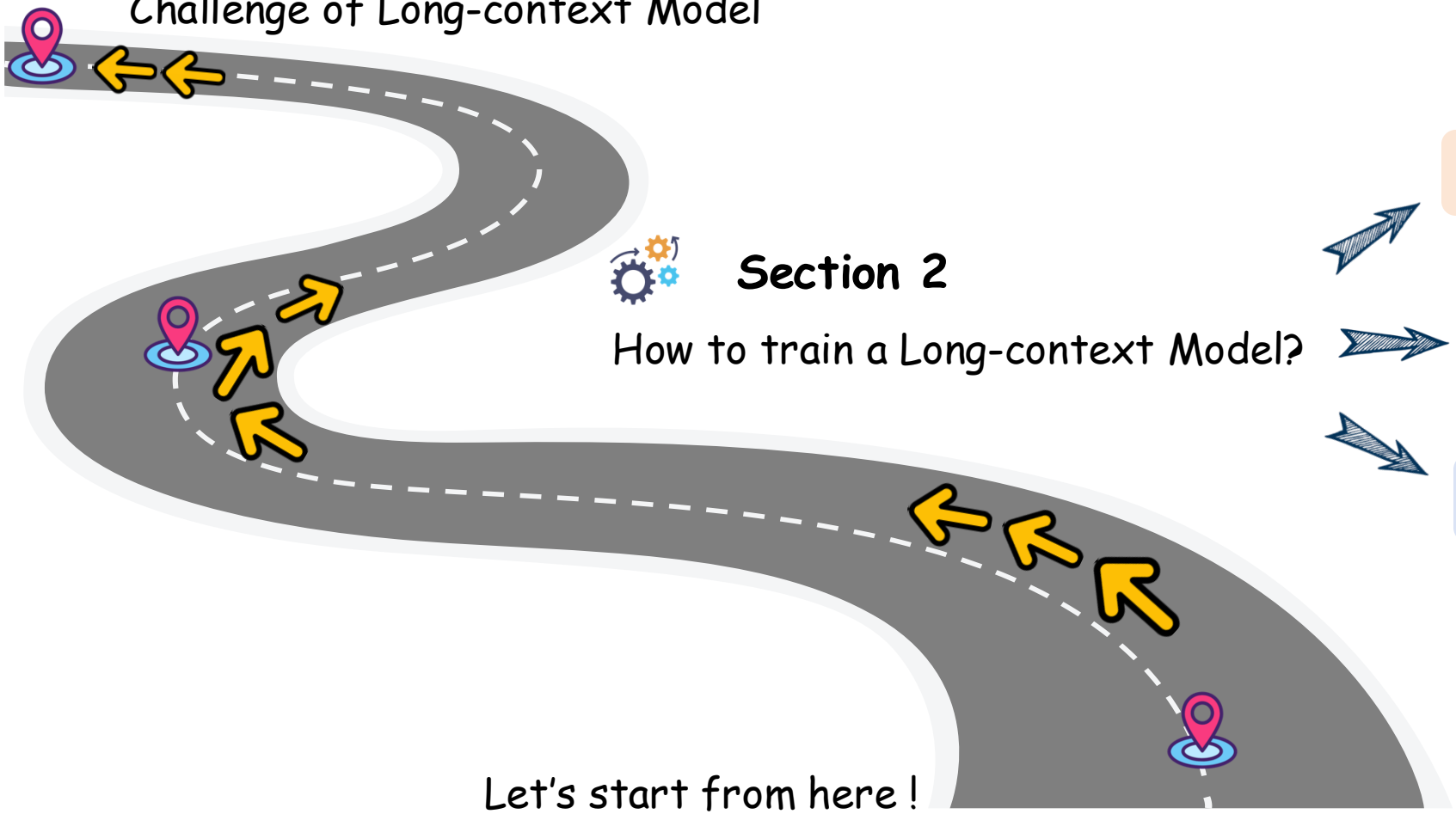
OpenNLG group @ Soochow University

Slides Link

Tutorial Road Map

Section 3:

Challenge of Long-context Model



Section 2

How to train a Long-context Model?

- 2.1 Modeling
- 2.2 Data
- 2.3 Evaluation

Section 1

What is Long-context Model?

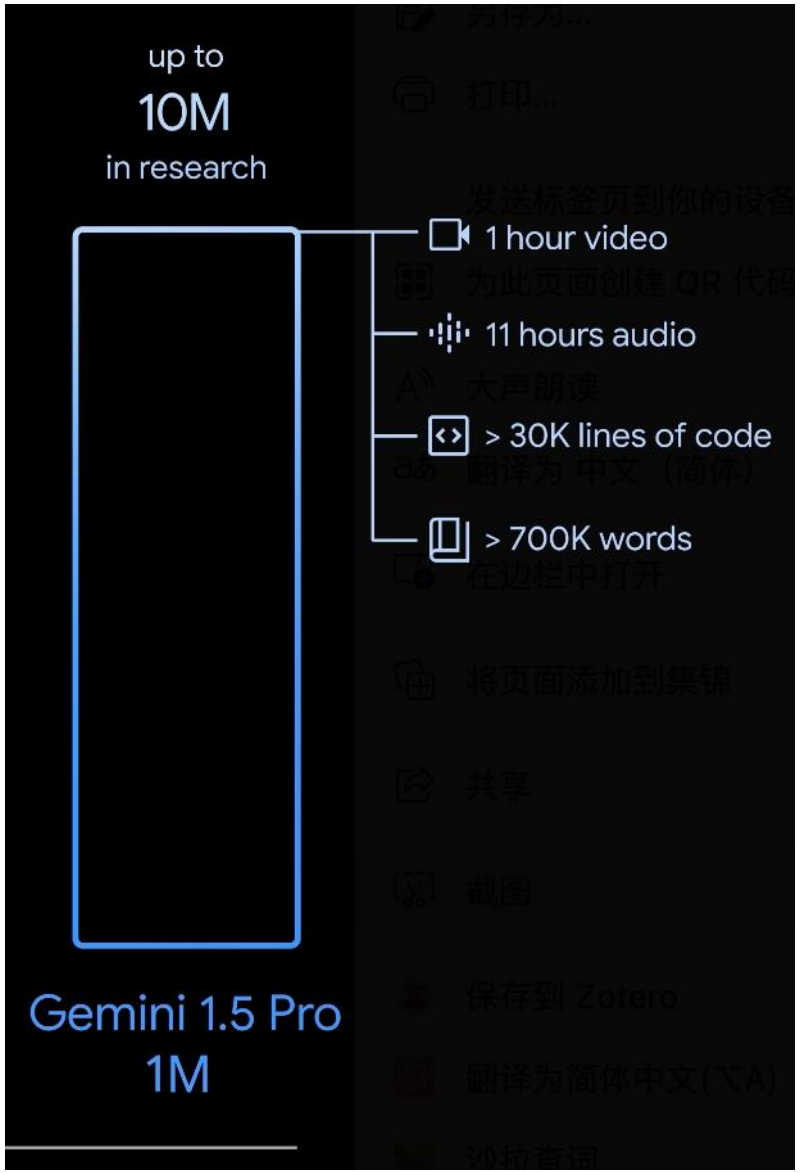
Let's start from here !

Section 1

What is Long Context Model?

The Era of Long-context Models

	Close-source	Open-source
2022 Nov	ChatGPT-3.5 *16K*	
2023 Apr		ChatGLM2 *32K*
2023 July	Claude 2 *100K*	MPT *65K*
2023 Sep		Qwen1.5 *32K* Baichuan2 *192K*
2023 Nov	GPT4 Turbo *128K*	Yi-6B *200K*
2023 Dec	Gemini 1.0 Pro* *32K*	Mistral 7B-0.2 *32K*
2024 Mar	Claude3 *200K*	Deepseek-V2 *128K*
2024 May	Gemini 1.5 Pro *10M*	
2024 July	GLM-4-long *1M*	Mistral Nemo *128K* Meta-Llama 3.1 *128K*



** we report the model announcement time*

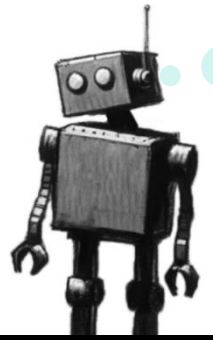
Source: Google Blog

Long-context Magic

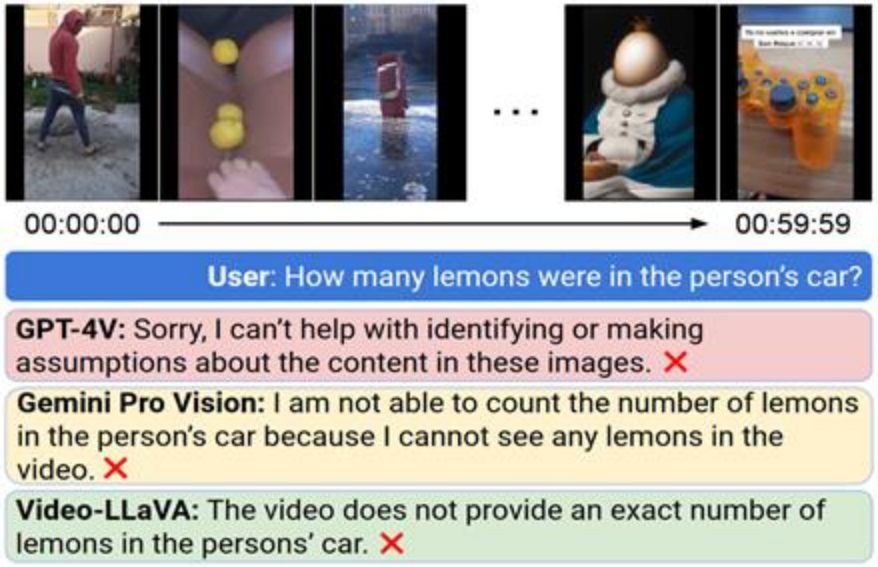
Considering some important scenarios in our daily life:

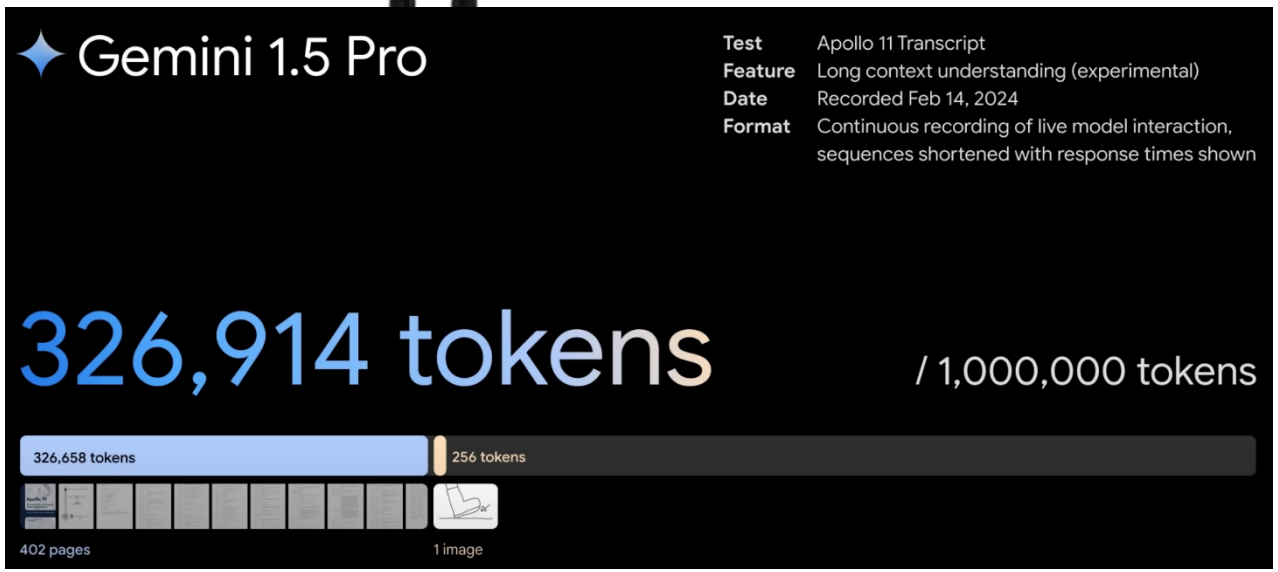
- Book and document analysis
- Web content reading
- Code bases writing
- High-res images
- Audio recordings and Videos
- …

Long



"See" More
"Memory" More
"Think" More
"Speak" More
…





What is Long-context Model ?

"A long context model, in the realm of natural language processing, refers to a type of language model that is capable of processing and understanding **extensive sequences of text**, far beyond the **typical context window size** that standard large language models (LLMs) can handle."

relative and vague concept

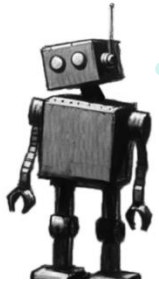
Nikolay Savinov

Research Scientist

Google DeepMind

Google Blog, Gemini Team

"10 million tokens at once is already close to the thermal limit of our Tensor Processing Units — we don't know where the limit is yet, and the model might be capable of even more as the hardware continues to improve" from Nikolay Savinov.



What's the boundary of Long-context Model?

"Sky's the limit"

Section 2

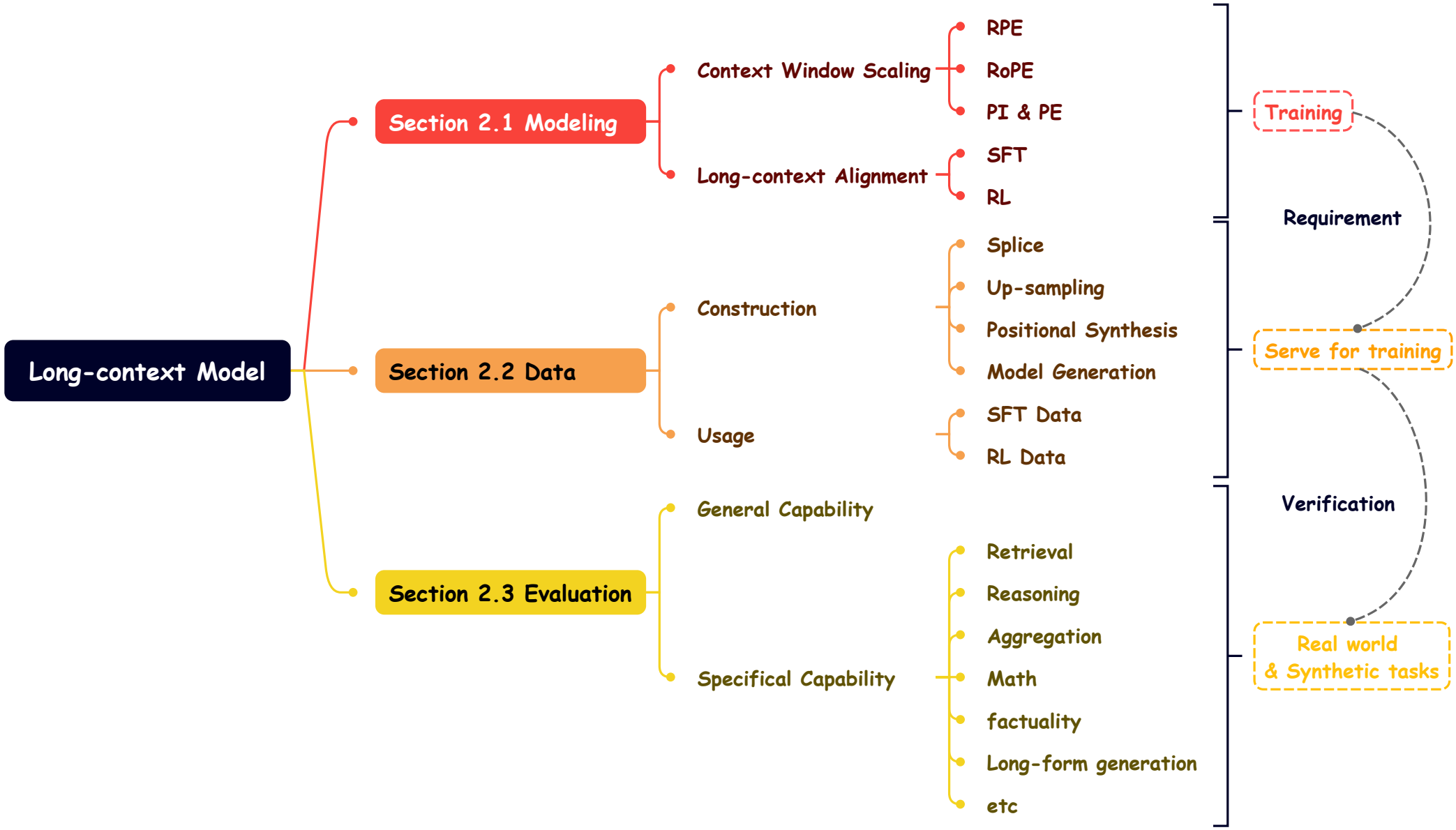
How to train a Long-context Model?

Section 2.1 Modeling

Section 2.2 Data

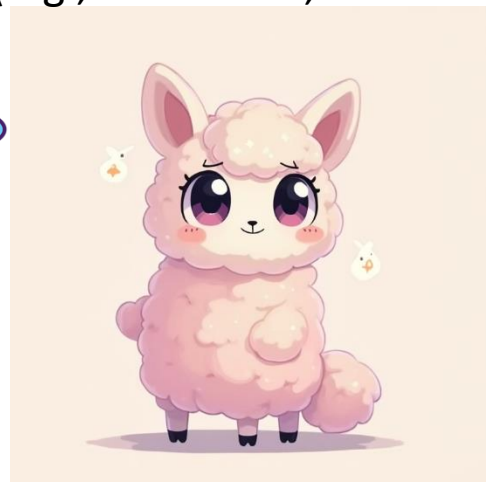
Section 2.3 Evaluation

Overall Structure of Sec.2

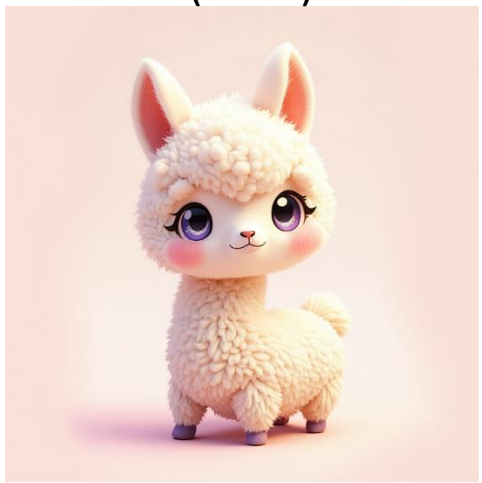


Sec. 2.1 Modeling

Open-source short-context model
(e.g., Llama2-4K, Llama3-8K)



Model with long context window
(>32K)



Powerful long-context model

Context Window Scaling

- Relative Positional Embedding (RPE)
- Rotary Positional Embedding (RoPE)
  - Position Interpolation (PI)
  - Position Extrapolation (PE)

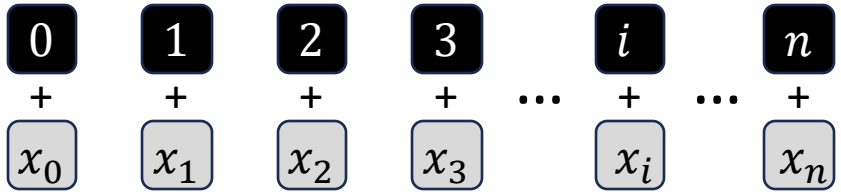
Long-context Alignment

- Supervised Fine-Tuning (SFT)
- Reinforcement Learning (RL)

Context Window Scaling • RPE

Absolute Position Embedding (APE):

$$A_{i,j} = \left(E_{x_i} + P_i\right)^{T} W_q^{\mathrm{T}} W_k \left(E_{x_j} + P_j\right)^{T}$$

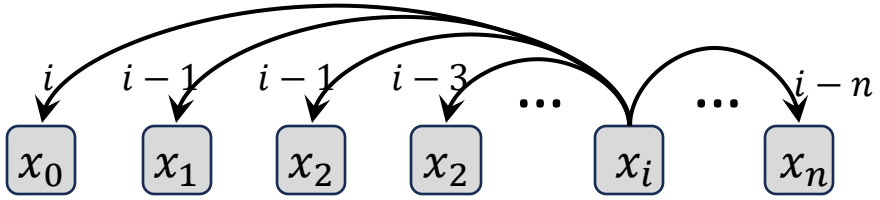


<0,0>	<0,1>	<0,2>	<0,3>	<0,4>	<0,5>	<0,6>	<0,7>	<0,8>	<0,9>
<1,0>	<1,1>	<1,2>	<1,3>	<1,4>	<1,5>	<1,6>	<1,7>	<1,8>	<1,9>
<2,0>	<2,1>	<2,2>	<2,3>	<2,4>	<2,5>	<2,6>	<2,7>	<2,8>	<2,9>
<3,0>	<3,1>	<3,2>	<3,3>	<3,4>	<3,5>	<3,6>	<3,7>	<3,8>	<3,9>
<4,0>	<4,1>	<4,2>	<4,3>	<4,4>	<4,5>	<4,6>	<4,7>	<4,8>	<4,9>
<5,0>	<5,1>	<5,2>	<5,3>	<5,4>	<5,5>	<5,6>	<5,7>	<5,8>	<5,9>
<6,0>	<6,1>	<6,2>	<6,3>	<6,4>	<6,5>	<6,6>	<6,7>	<6,8>	<6,9>
<7,0>	<7,1>	<7,2>	<7,3>	<7,4>	<7,5>	<7,6>	<7,7>	<7,8>	<7,9>
<8,0>	<8,1>	<8,2>	<8,3>	<8,4>	<8,5>	<8,6>	<8,7>	<8,8>	<8,9>

- Transformer-based models rely on PE to identify the position of each token;
- RPE focuses more on the relative position relationship

Relative Position Embedding (RPE):

$$A_{i,j} = Att(E_{x_i}, E_{x_j}, i-j)$$

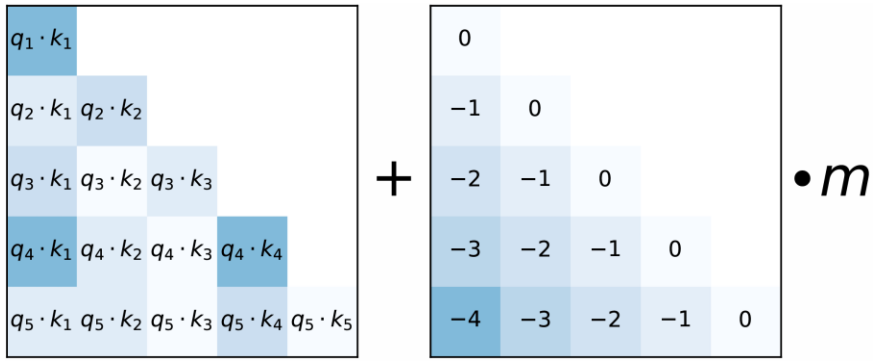


0	1	2	3	4	5	6	7	8	9	10
-1	0	1	2	3	4	5	6	7	8	9
-2	-1	0	1	2	3	4	5	6	7	8
-3	-2	-1	0	1	2	3	4	5	6	7
-4	-3	-2	-1	0	1	2	3	4	5	6
-5	-4	-3	-2	-1	0	1	2	3	4	5
-6	-5	-4	-3	-2	-1	0	1	2	3	4
-7	-6	-5	-4	-3	-2	-1	0	1	2	3
-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0

Context Window Scaling • RPE

Attention with Linear Biases Enables Input Length Extrapolation (ALiBi) (ICLR 2022)

ALiBi Function: $A_{i,j} = E_{x_i}^T W_q^T W_k E_{x_j} - \lambda|i-j|$



0	1	2	3	4	5	6	7	8	9	10	11
1	0	1	2	3	4	5	6	7	8	9	10
2	1	0	1	2	3	4	5	6	7	8	9
3	2	1	0	1	2	3	4	5	6	7	8
4	3	2	1	0	1	2	3	4	5	6	7
5	4	3	2	1	0	1	2	3	4	5	6
6	5	4	3	2	1	0	1	2	3	4	5
7	6	5	4	3	2	1	0	1	2	3	4
8	7	6	5	4	3	2	1	0	1	2	3
9	8	7	6	5	4	3	2	1	0	1	2
10	9	8	7	6	5	4	3	2	1	0	1
11	10	9	8	7	6	5	4	3	2	1	0

Pros

- Simple yet effective (MPT, 2023), scaling to 65K

Cons

- Single direction: unable to identify relative left or right position;
- Relative positional weight severe attenuation as the sequence length increases.

Context Window Scaling • RPE

Bucket RPE (Oct 19), First from to T5 (JMLR, 2020)

Core Function: $A_{i,j} = E_{x_i}^T W_q^{\mathrm{T}} W_k E_{x_j} + \beta_{i,j}$

$$b_{i,j} = \begin{cases} i-j, & if \ \ |i-j| < n_b/4 \\ \dfrac{i-j}{|i-j|} \cdot \min\left(\dfrac{n_b}{2} - 1, \dfrac{n_b}{4} + \left\lfloor \dfrac{\log\left(\frac{(i-j)}{n_b/4}\right)}{\log\left(\frac{\max d}{n_b/4}\right)} \cdot \dfrac{n_b}{4} \right\rfloor \right), & else \end{cases}$$

max distance, also the starting distance of the farthest bucket

the number of buckets

$i-j$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$f(i-j)$	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9
$i-j$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	...
$f(i-j)$	10	10	10	10	10	10	10	11	11	11	11	11	11	11	11	...

``The closer, the more accurate; The farther, the blurrier''

0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9	10	10	10	10
-1	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9	10	10	10
-2	-1	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9	10	10
-3	-2	-1	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9	10
-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9
-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9
-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	8	8	8	9	9
-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	8	8	8	9
-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	8	8	8
-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	8	8
-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	8
-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
-9	-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
-9	-9	-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5
-9	-9	-9	-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4
-10	-9	-9	-9	-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3
-10	-10	-9	-9	-9	-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
-10	-10	-10	-9	-9	-9	-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0	1
-10	-10	-10	-10	-9	-9	-9	-9	-8	-8	-8	-8	-7	-6	-5	-4	-3	-2	-1	0

Context Window Scaling • RoPE

$$< f_q(x_m, m), f_k(x_n, n) > = g(x_m, x_n, m - n)$$

General term of RoPE

(Relative position)

How to directly applied to the attention?

Utilize the properties of exponential and complex number operations

Definition of complex number operation:

Let q and k be complex numbers. Their inner product, denoted as $\langle q, k \rangle$, is defined as the real part of the product of q and the complex conjugate of k:

$$\langle q, k \rangle := \mathrm{R}[qk^*],$$

- R[.] denotes the real part
- k^* is the complex conjugate of k

Context Window Scaling • RoPE

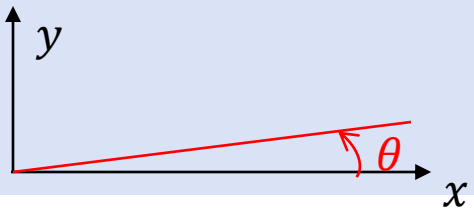
$$< f_q(x_m, m), f_k(x_n, n) > = g(x_m, x_n, m - n)$$ Relative position

Is there another way to integrate RPE into Attention?

Utilize the properties of exponential and complex number operations

Definition of Rotation Matrix

- Complex number $z = a + ib$ can be represented in matrix form as: $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$
- Rotation matrix is given by $\begin{pmatrix} \cos\theta & -sin\theta \\ sin\theta & \cos\theta \end{pmatrix}$
- Geometric interpretation of complex number multiplication: rotate the vector counterclockwise by an angle θ



Definition of complex number

Let q and k be complex numbers. Their inner pro-duct, denoted as $\langle q, k \rangle$, is defined as the real part of the product of q and the complex conjugate of k:

$$\langle q, k \rangle := \mathrm{R}[qk^*]$$

- R[.] denotes the real part
- k^* is the complex conjugate of k

Context Window Scaling • RoPE

Definition of complex number

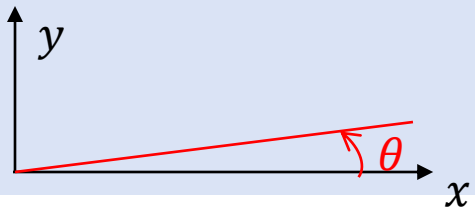
Let q and k be complex numbers. Their inner pro-duct, denoted as $\langle q, k\rangle$, is defined as the real part of the product of q and the complex conjugate of k:

$$\langle q, k\rangle := \mathrm{R}[qk^*]$$

- R[.] denotes the real part
- k^* is the complex conjugate of k

Definition of Rotation Matrix

- Complex number $z = a + ib$ can be represented in matrix form as: $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$
- Rotation matrix is given by $\begin{pmatrix} \cos\theta & -sin\theta \\ sin\theta & \cos\theta \end{pmatrix}$
- Geometric interpretation of complex number multiplication: rotate the vector counterclockwise by an angle θ



Take the rotation matrix as positional embedding

$$\begin{aligned} A_{m,n} = q_m^T \mathrm{k_n} &= \left(P_m W_q x_m\right)^{\mathrm{T}} \left(P_n W_k x_n\right) \\ &= \left(W_q x_m\right)^{\mathrm{T}} P_m^{\mathrm{T}} P_n (W_k x_n) \\ &= \left(W_q x_m\right)^{\mathrm{T}} P_{m-n}^{\mathrm{T}} (W_k x_n) \end{aligned}$$

- P_m, P_n denote the position of i-th and j-th token
- $q_m, \mathrm{k_n}$ denote the *i-th* query, *j-th* key;
- E_m, E_n denote the *i-th, j-th* token.

Let's express f_q and f_k in complex number form

$$q_m = f_q(x_m, m) = \left(W_q x_m\right) e^{im\theta} \quad (1)$$

$$k_n = f_k(x_n, n) = (W_k x_n) e^{in\theta} \quad (2)$$

$$g(x_m, x_n, m-n) = R[\left(W_q x_m\right)(W_k x_n)^* e^{i(m-n)\theta}]$$

Context Window Scaling • RoPE

Position P_m: Two-dimensional case

The complex number is similar to the 2-dim rotation matrix below.

P_m denotes the matrix representing the position m:

$$P_m = \begin{pmatrix} \mathrm{R}[\mathrm{e}^{im\theta}] & \mathrm{I}[\mathrm{e}^{-im\theta}] \\ \mathrm{I}[\mathrm{e}^{im\theta}] & \mathrm{R}[\mathrm{e}^{im\theta}] \end{pmatrix} = \begin{pmatrix} \cos \mathrm{m}\theta & -\sin \mathrm{m}\theta \\ \sin \mathrm{m}\theta & \cos \mathrm{m}\theta \end{pmatrix}$$

I[.] denotes the imaginary part

Take the rotation matrix as positional embedding

$$\begin{aligned} A_{m,n} = q_m^T \mathrm{k_n} &= \left(P_m W_q x_m\right)^{\mathrm{T}} \left(P_n W_k x_n\right) \\ &= \left(W_q x_m\right)^{\mathrm{T}} P_m^{\mathrm{T}} P_n (W_k x_n) \\ &= \left(W_q x_m\right)^{\mathrm{T}} P_{m-n}^{\mathrm{T}} (W_k x_n) \end{aligned}$$

- q_m, $\mathrm{k_n}$ denote the *i-th* query, *j-th* key;
- E_m, E_n denote the *i-th, j-th* token.

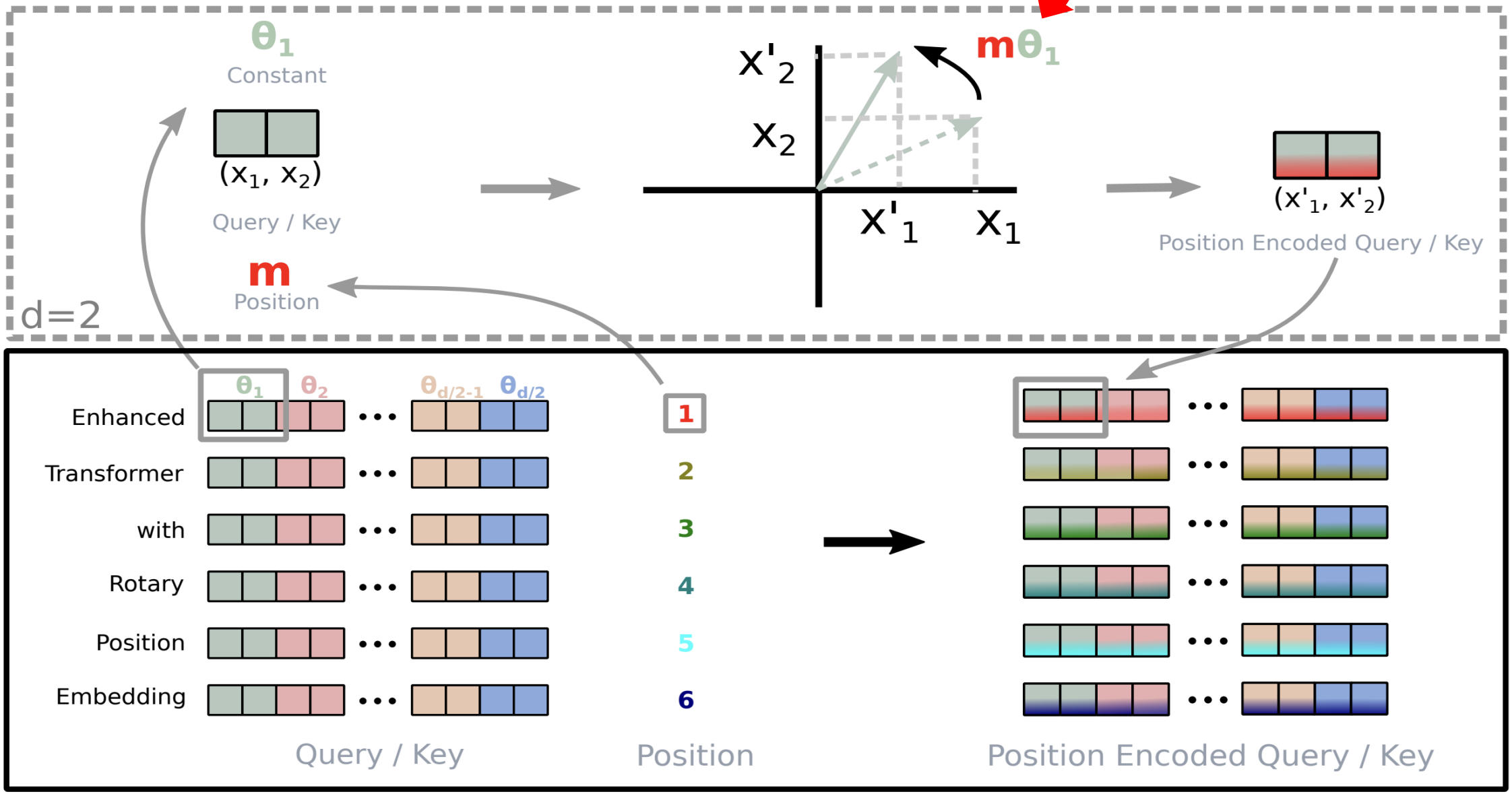
A d-dim form can be as follows

$$P_m = \begin{pmatrix} \cos m\theta_0 & -\sin m\theta_0 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_0 & \cos m\theta_0 & 0 & 0 & & 0 & 0 \\ 0 & 0 & \cos m\theta_1 & -\sin m\theta_1 & & 0 & 0 \\ 0 & 0 & \sin m\theta_1 & \cos m\theta_1 & & 0 & 0 \\ & & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & 0 & & \cos m\theta_{d/2-1} & -\sin m\theta_{d/2-1} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2-1} & \cos m\theta_{d/2-1} \end{pmatrix}$$

$$\theta_i = 1000^{-2(i-1)/d}, i \in [1,2,\ldots,d/2]$$

Context Window Scaling • RoPE

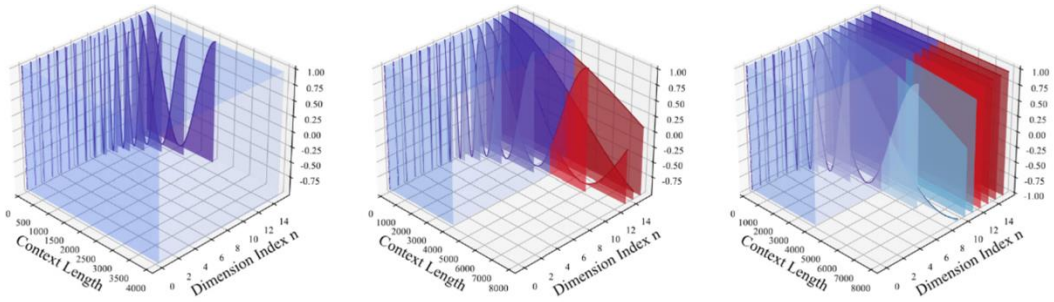
Rotate the components of the token vector by an angle in pairs



Context Window Scaling • RoPE

Implementation

Here is a brief code of RoPE implemented by pytorch



(a) RoPE base=500 (b) RoPE base=10000 (c) RoPE base=1000000

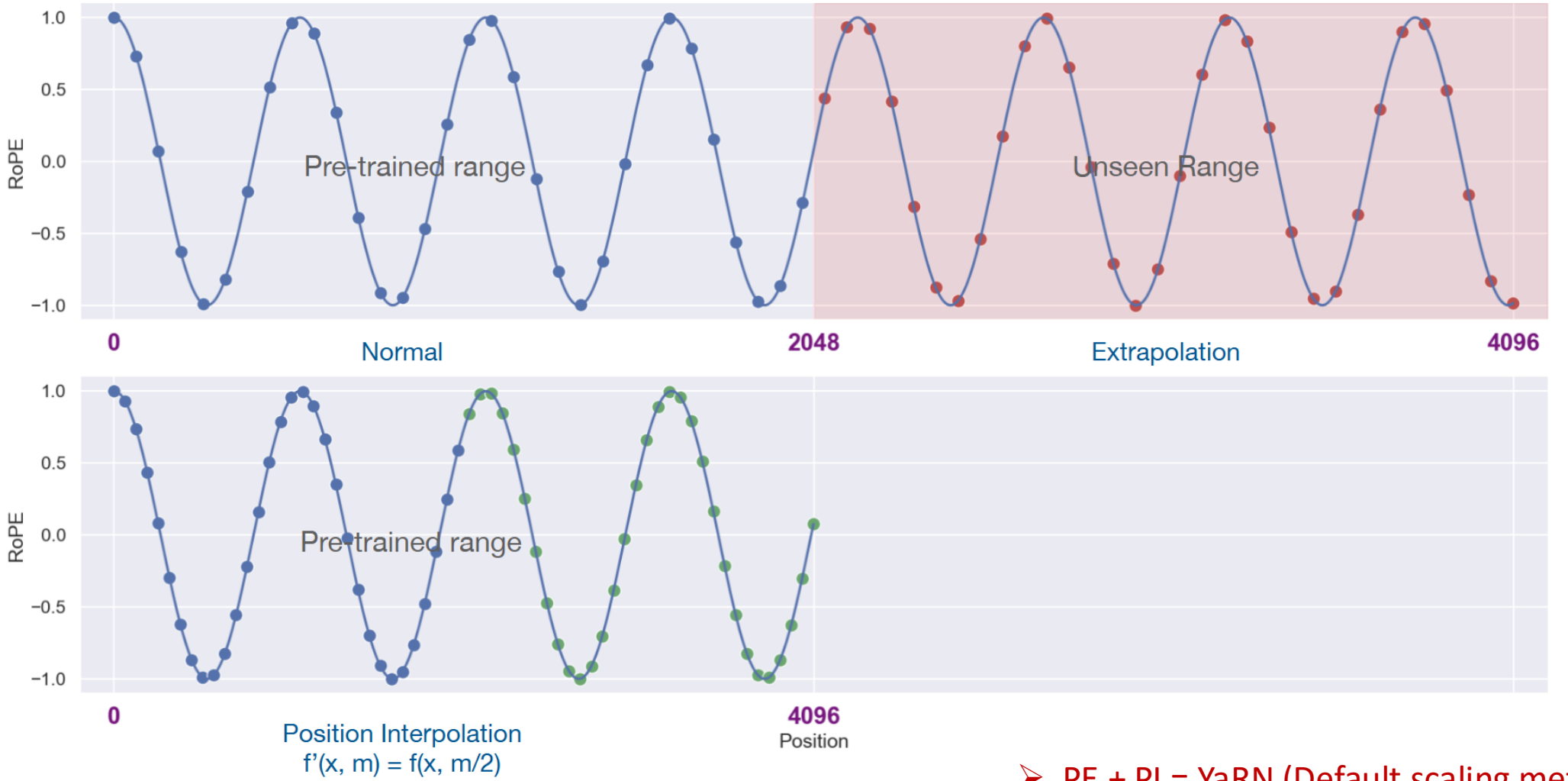
- Large θ means flatter wavelengths, allowing the model to accommodate longer context length
- Llama-2 θ: 10K, Llama-3 θ: 500K

```python
def _init_weight(weight: nn.Parameter):
    #return the sinusodial value
    n_pos, d = weight.shape
    position_enc = np.array(
        [[pos / np.power(10000, 2 * (j // 2) / d) for j in range(d)]\
         for pos in range(n_pos)])# \theta_i ,i=[0,0,1,1,...,d//2-1,d//2-1]
    weight.requires_grad = False
    sentinel = d // 2 if d % 2 == 0 else (d // 2) + 1
    weight[:, 0:sentinel] = torch.FloatTensor(np.sin(position_enc[:, 0::2]))
    weight[:, sentinel:] = torch.FloatTensor(np.cos(position_enc[:, 1::2]))
    weight.detach_()
    return weight

def apply_rotary(x,past_key_values_length=0):
    # x.shape [batch, num_heads, seq_len, d]
    #n_pos:position nums # d: d_model//num_heads
    pos_embedding=torch.randn(n_pos,d)
    weight=_init_weight(pos_embedding)
    sinusoidal_pos=weight[
        past_key_values_length: past_key_values_length+x.size(-2)
    ][None,None,:,:]

    sin, cos = sinusoidal_pos.chunk(2,dim=-1)
    x1, x2 = x[..., 0::2], x[..., 1::2]
    # [cos_nθ, -sin_nθ] [x1]
    # [sin_nθ,  cos_nθ] [x2]
    # => [x1 * cos_nθ - x2 * sin_nθ, x1 * sin_nθ + x2 * cos_nθ]
    return torch.cat([x1 * cos - x2 * sin, x1 * sin + x2 * cos], dim=-1)
```

https://github.com/JunnYu/RoFormer_pytorch/blob/roformer_v2/src/roformer/modeling_roformer.py#L156

Context Window Scaling • PE & PI

Extending Context Window of Large Language Models via Positional Interpolation (Meta, 2023)



Position Extrapolation (PE)

Position Interpolation (PI)

Hugging Face

➢ PE + PI = YaRN (Default scaling method of HF)

YaRN: Efficient Context Window Extension of Large Language Models (ICLR 2024)

Context Window Scaling • RoPE

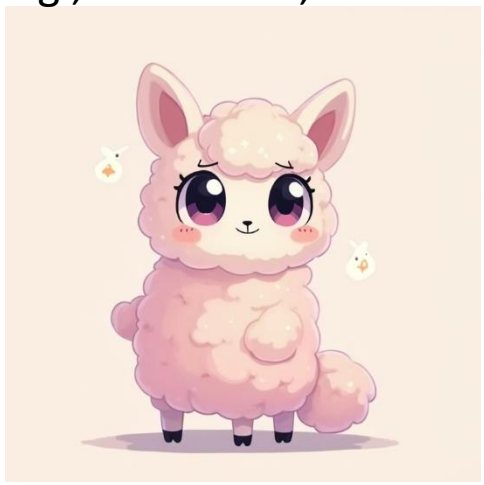
Summarization

Why RoPE becomes mainstream method in Long-context Models?

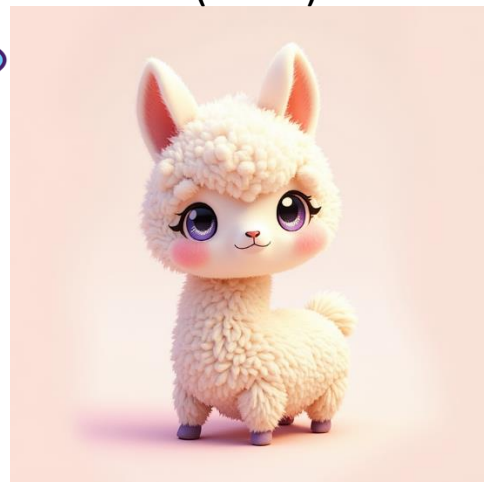
- Unified APE and RPE
- Flexibility in sequence length: decaying inter-token dependency
- Compatibility with *Efficient Attention Mechanisms*
- Simplicity and ease of Implementation
- …

Long-context Alignment

Open-source short-context model
(e.g., Llama2-4K, Llama3-8K)



Model with long context window
(>32K)



Powerful long-context model

Context Window Scaling

- ✓ Relative Positional Embedding (RPE)
- ✓ Rotary Positional Embedding (RoPE)
  - ✓ Position Interpolation (PI)
  - ✓ Position Extrapolation (PE)

Long-context Alignment

- ➢ Supervised Fine-Tuning (SFT)
- ➢ Reinforcement Learning (RL)

Long-context Alignment

Data

With data and algorithm, we can conduct long-context alignment …

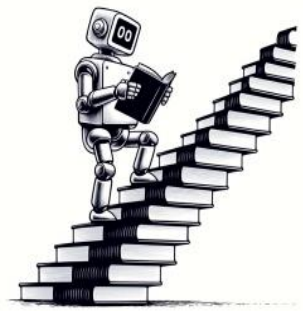
Algorithm

Supervised Fine-Tuning (SFT)

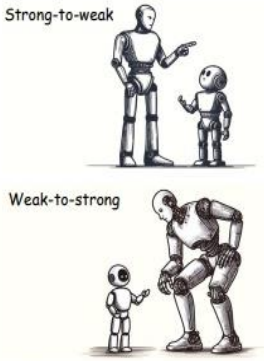
SFT the pre-trained model with task-specific datasets

Reinforcement Learning (RL)

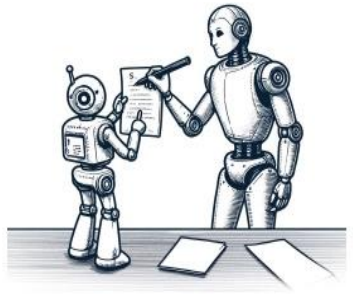
Utilize signal from reward model to update the model



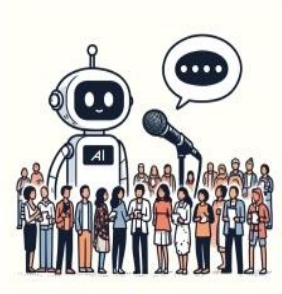
(a) Aligning through inductive bias



(b) Aligning through behavior imitation



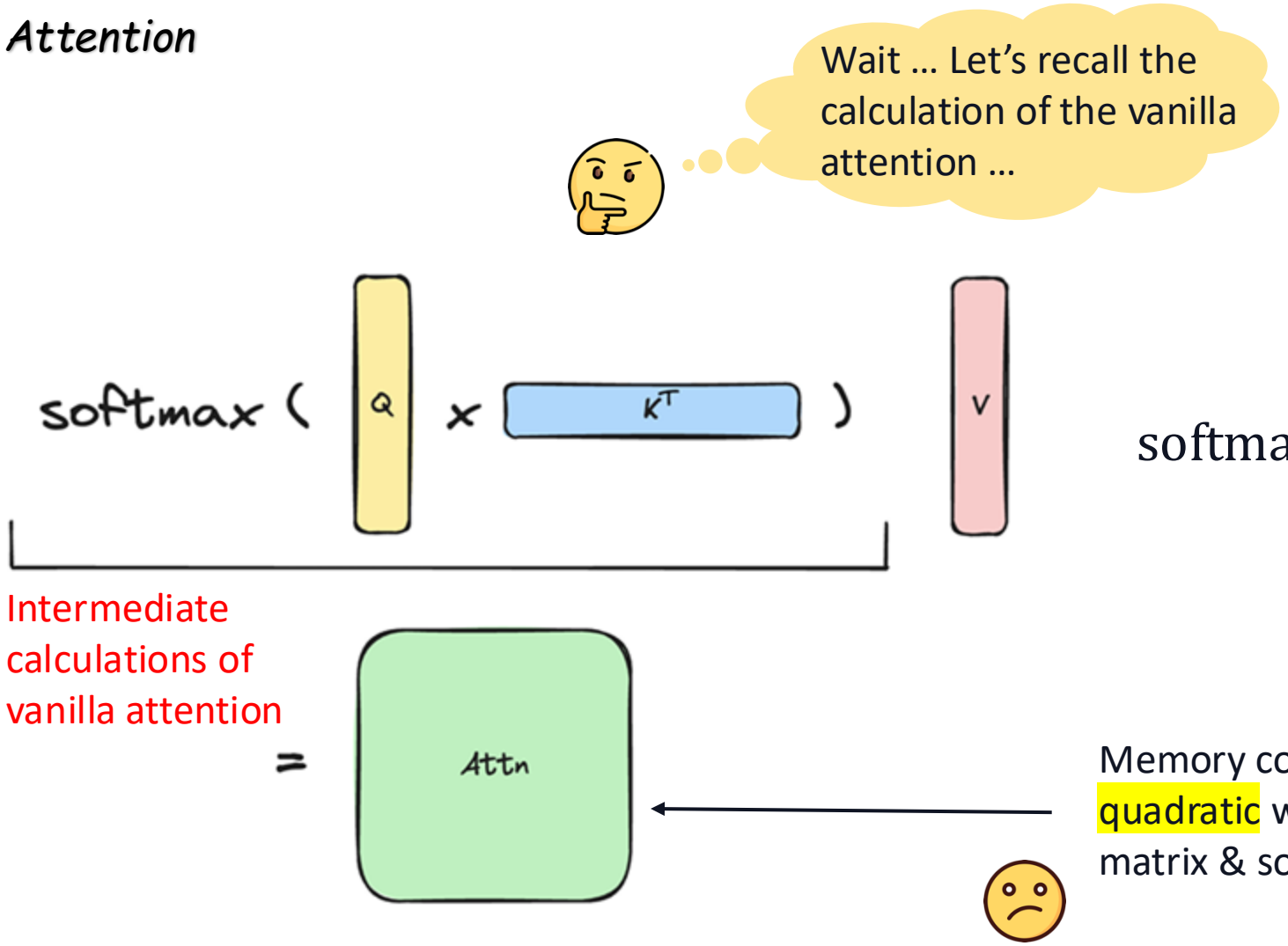
(c) Aligning through model feedback



(d) Aligning through environment feedback

Long-context Alignment

Vanilla Attention



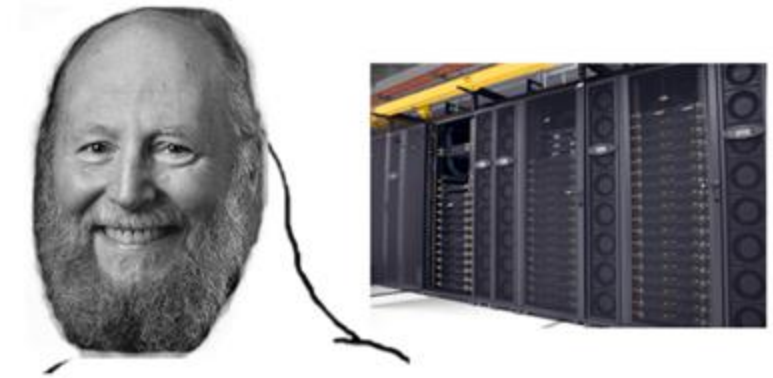
$$\text{softmax}(QK^T)V$$

Memory complexity of naive attention is quadratic with sequence length (score matrix & softmax output).

Long-context Alignment

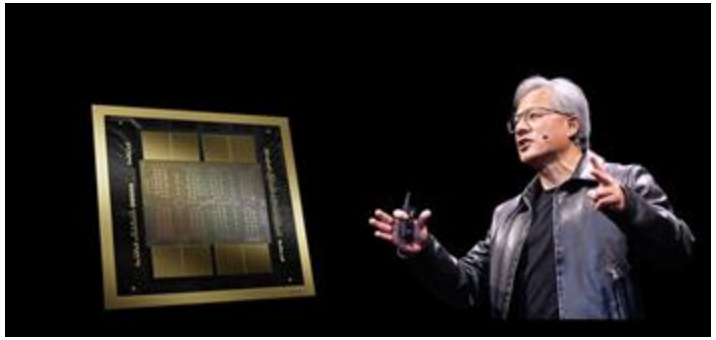
Challenge: We Run Out of Memory

``with a batch size of 1, processing 100 million tokens requires over 1000 GB of memory for a modest model with a hidden size of 1024'' —— Ring Attention, 2023, Hao Liu et al.



Why ?

- Input has to be materialized in GPU;
- Memory scales linearly with Flash-Attention
  - Need to store input QKV (Prefilling Memory)
  - Output (Prediction Memory)
  - LSE (logits or similarity estimations, Intermediate calculations)
  - Dout (gradient of output) for backward



Memory of current high-end GPUs

- NVIDIA H200: 141 GB
- AMD MI300X: 192 GB
- NVIDIA GB200 (Blackwell): 288 GB (available late 2024)
- *NVIDIA H100/A100/A800*: 80GB*

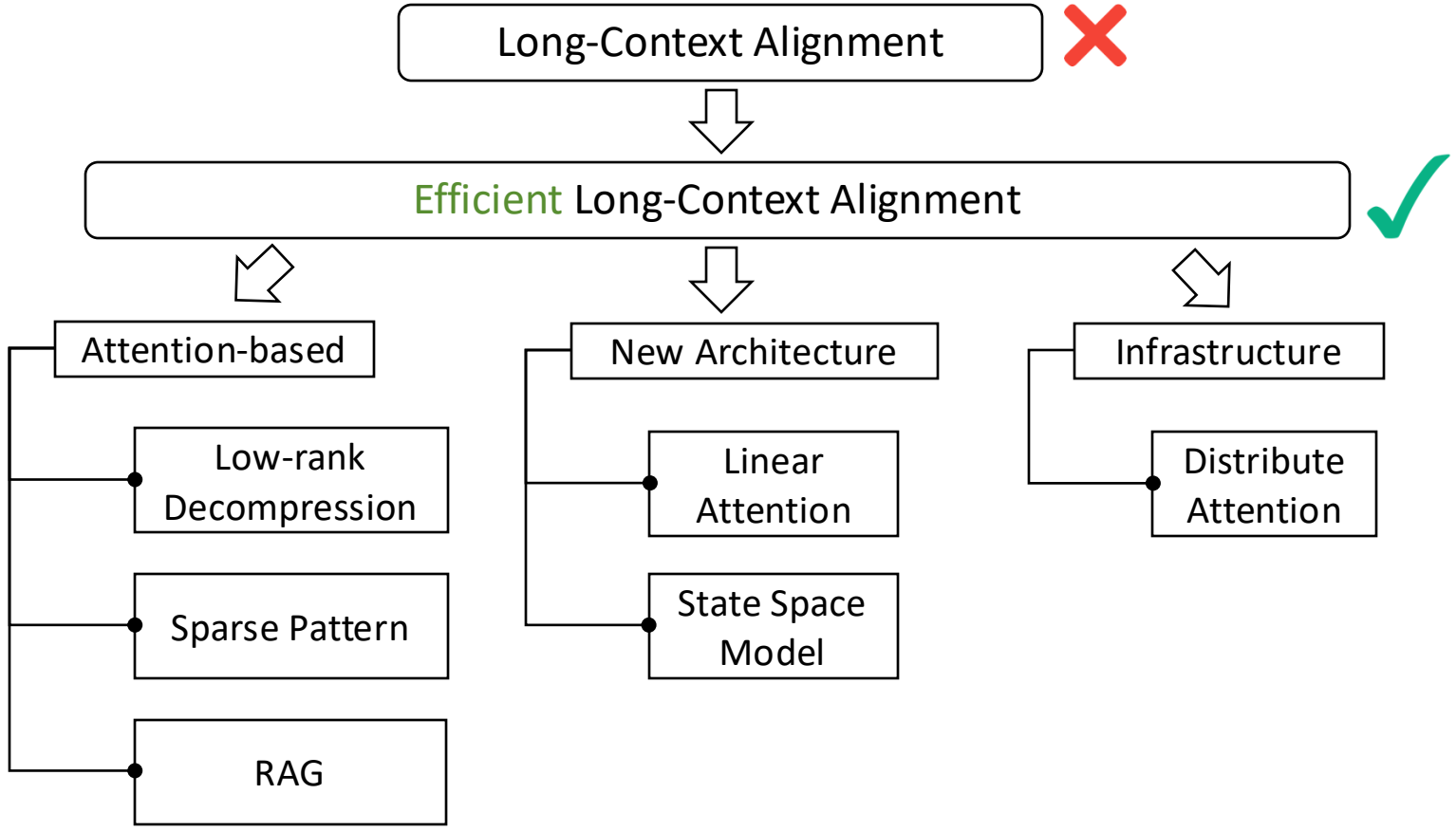
Long-context Alignment

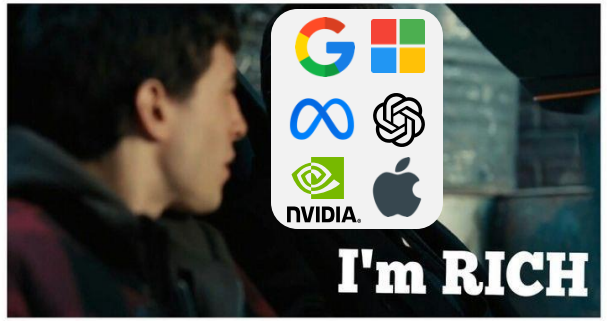
Challenge: We Run Out of Memory

Algorithm ?

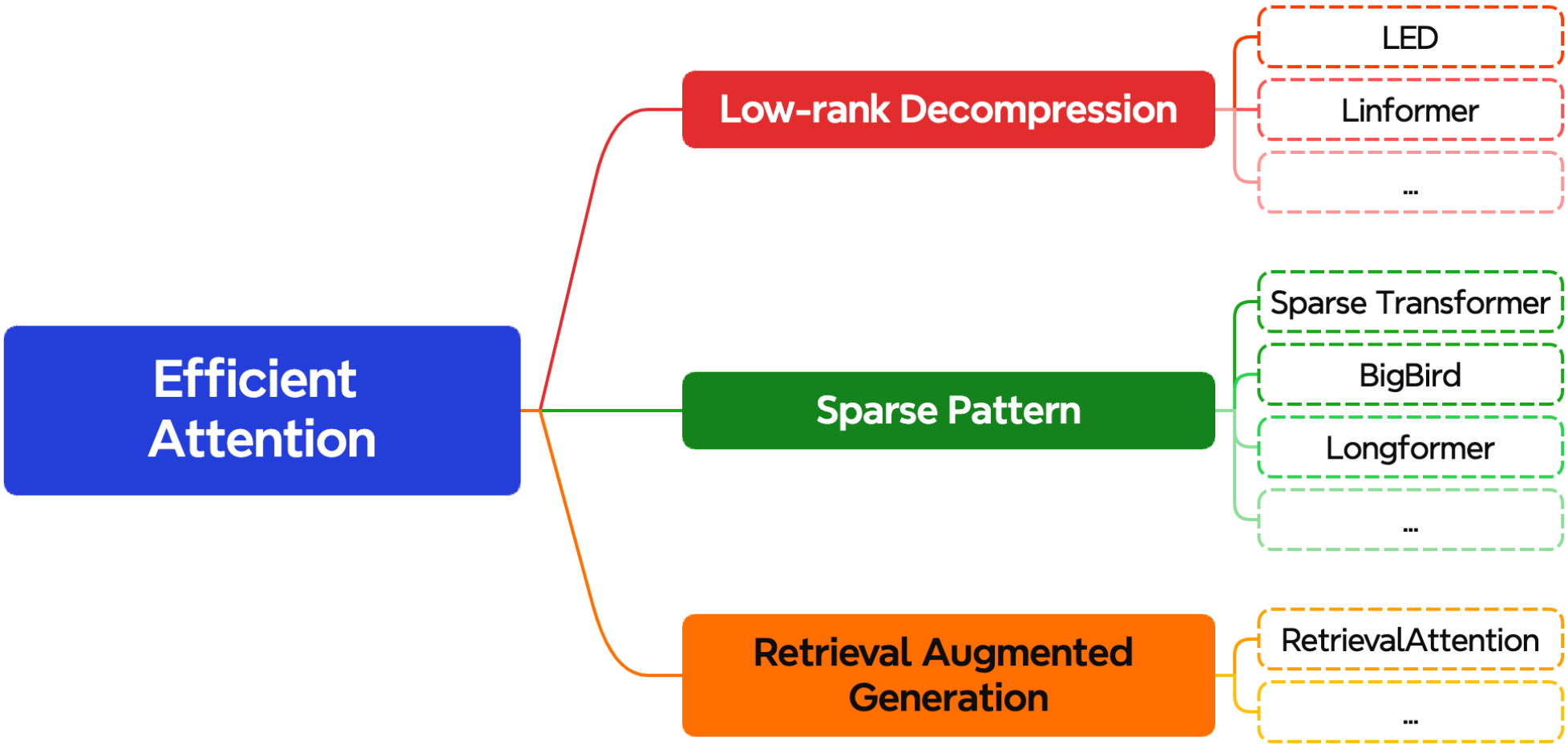
Implementation !

For academic research, before designing ``superpower'' training objective and strategy, thinking about ``how to implement it''!

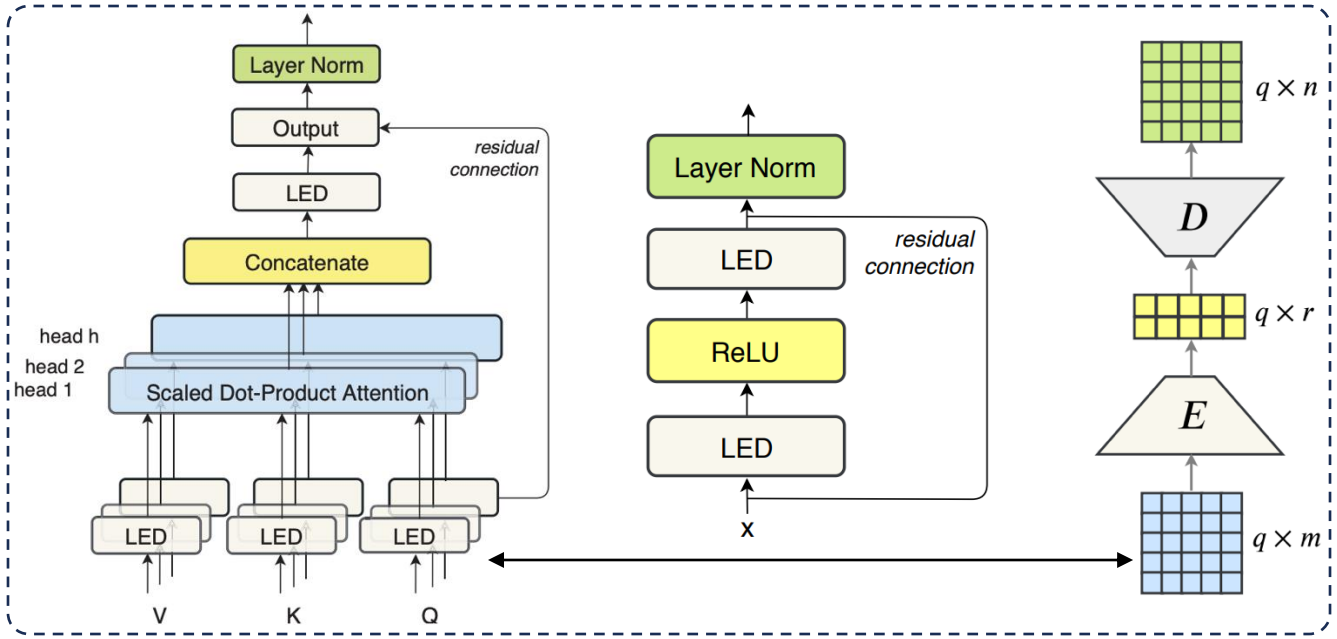




Efficient Attention-based Models

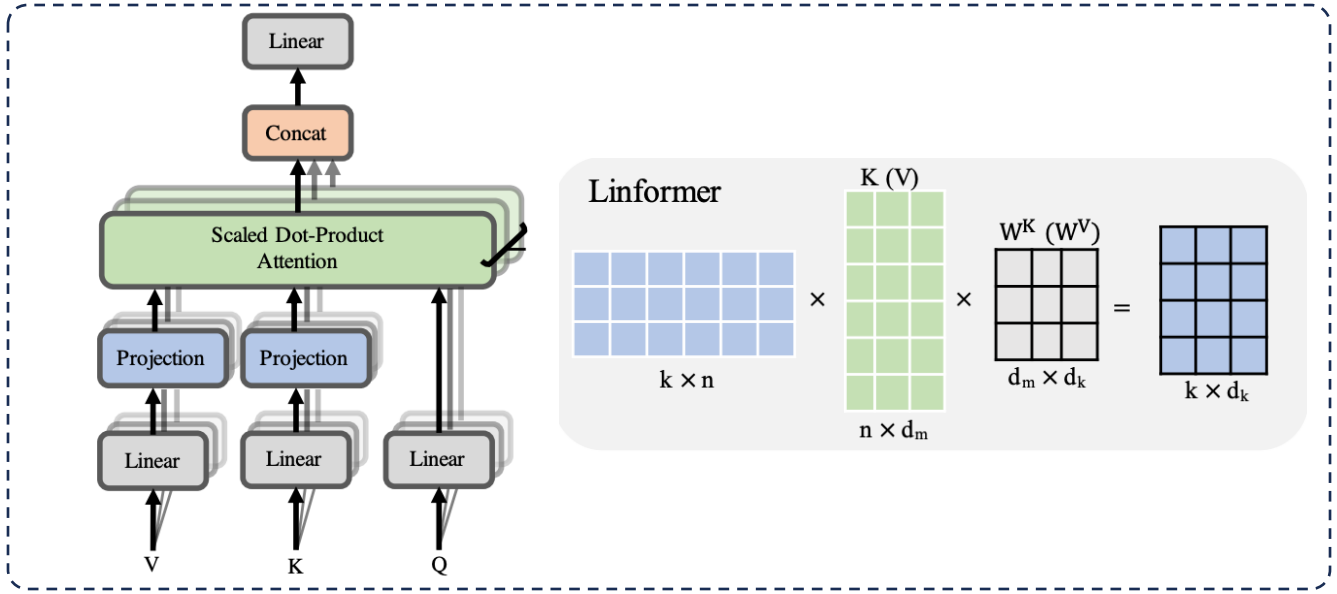


Low-rank decompression



LED: Lightweight and efficient end-to-end speech recognition using low-rank transformer (ICASSP 2022)

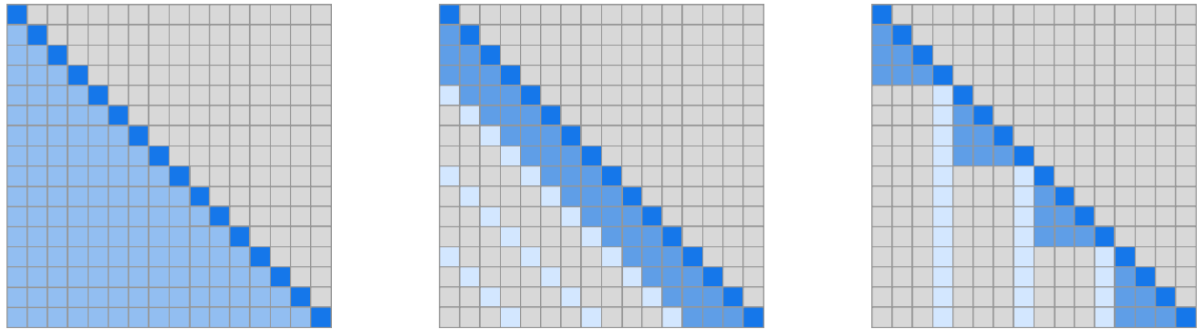
- Utilize VAE model to reduce the sequence length
- Reduce the matrix size of Q, K, V for approximation of linear parameter efficiency.



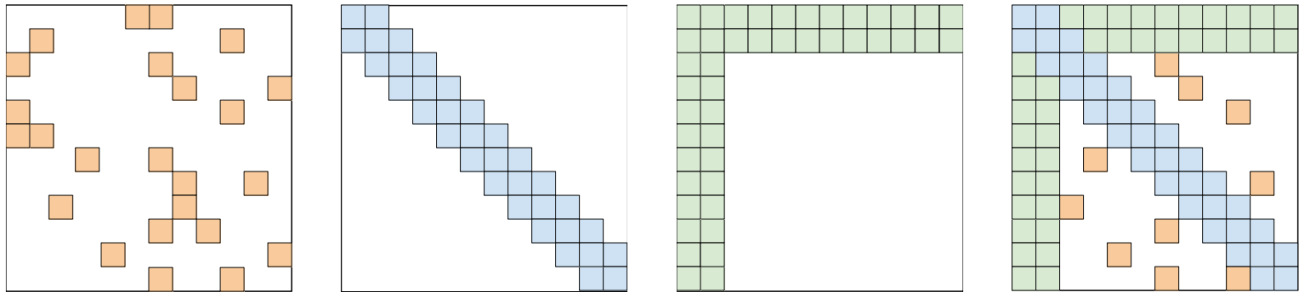
Linformer: Self-attention with linear complexity (Meta, 2020)

- Using a low-rank approximation of the K, V attention matrix
- "Automatically" compressing the context length ≈ dropping "irrelevant" tokens

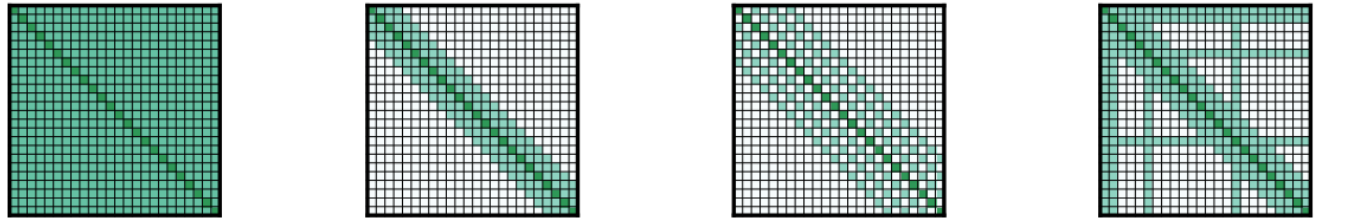
Sparse Attention



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Sparse Transformer (OpenAI, 2019)

- Limit the attention within a local window.

BigBird (NeurIPS 2020)

- Attend to a random subset of previous tokens as well as several globally accessible tokens.

Longformer (AllenAI, 2020)

- Introduce dilated sliding window patterns to increase attention's receptive field and manually picks the window sizes for each layer.

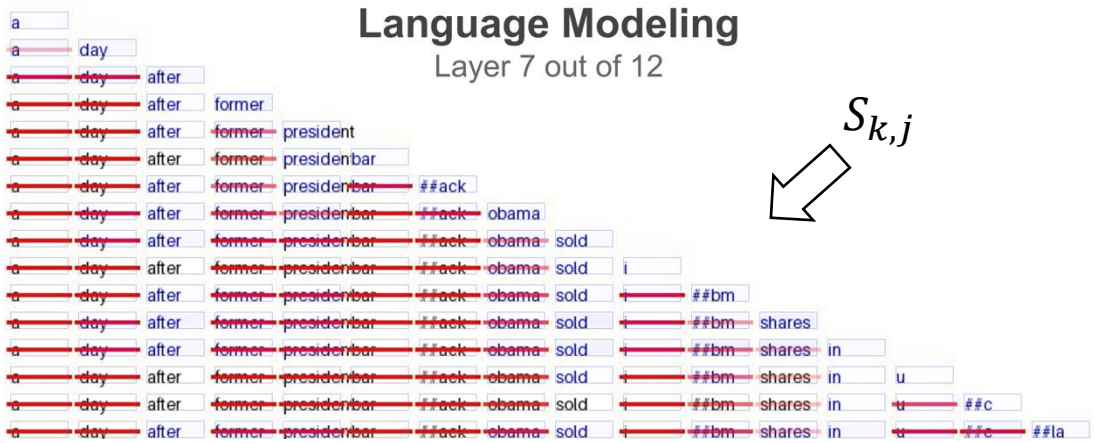
Sparse Attention

"a token wanting to mask another is after absorbing its contents" —— speculative decoding, ICML 2023

Selective Masking Strategy

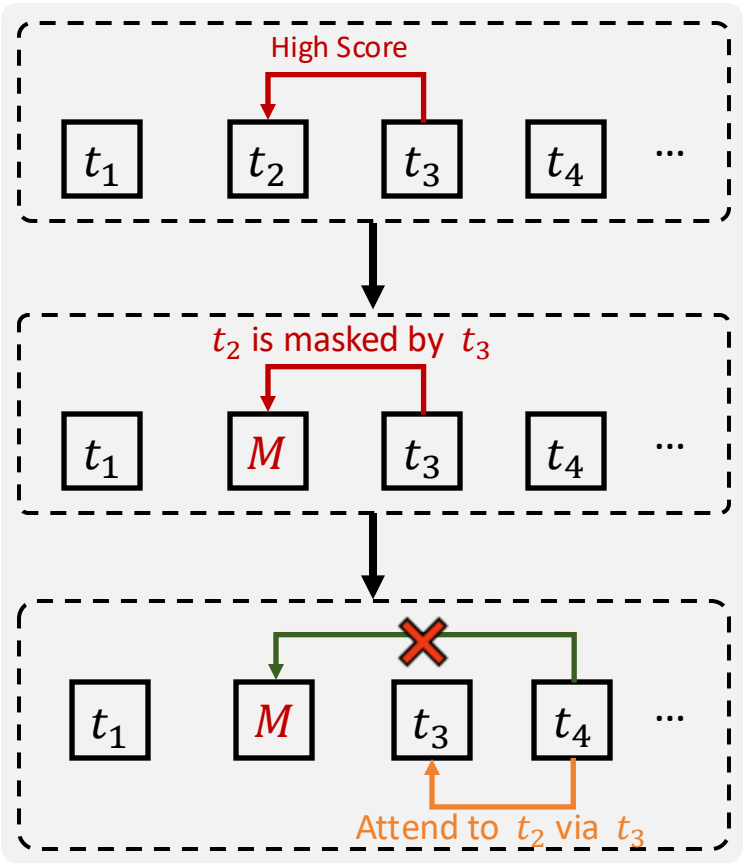
$$\text{SelectiveAttention}(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} - F\right)V,$$

where $F_{i,j} = \sum_{k \le i-1} S_{k,j}$

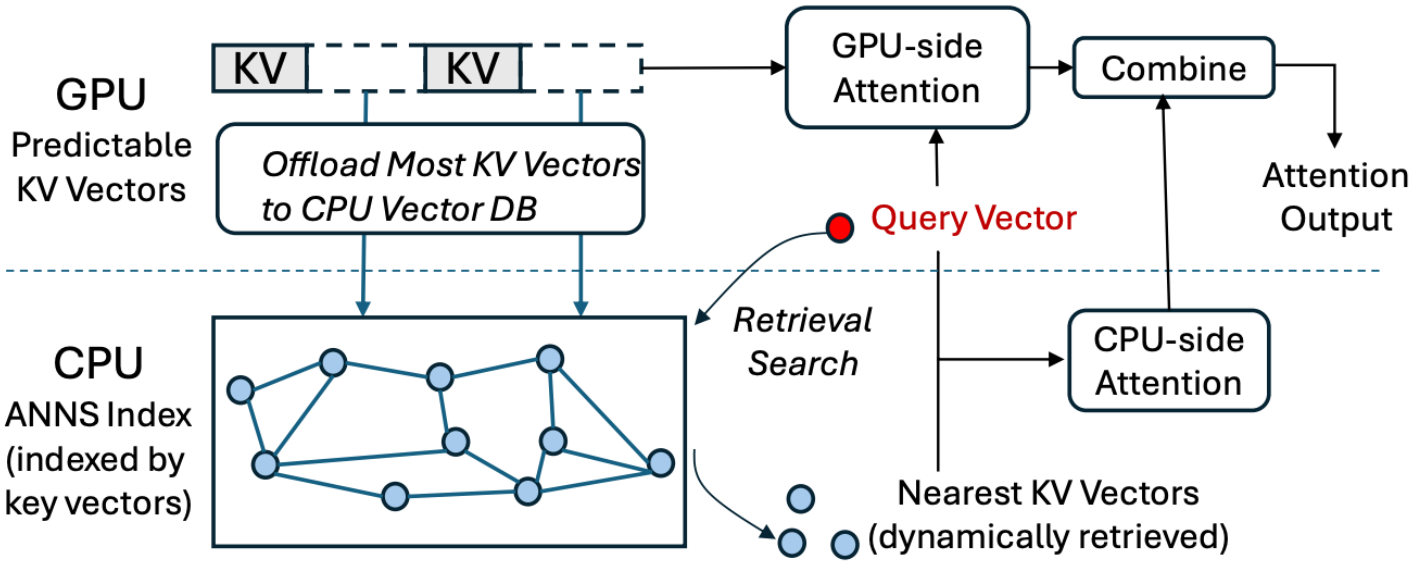


- A token is masked by selective attention, it will not contribute meaningfully to any future attention operations

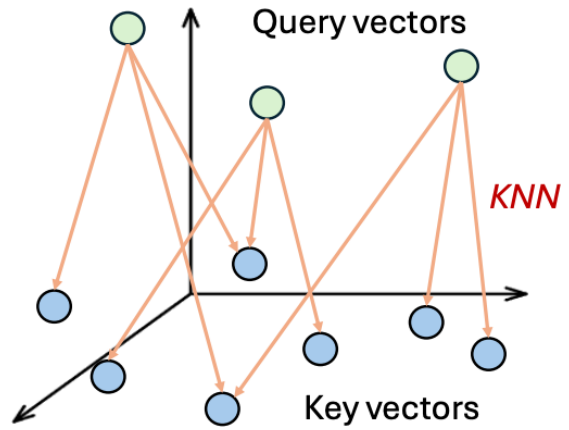
Selective Attention (Google, 2024)



RAG



(a) Overall design of RetrievalAttention.



(b) Key building procedure of OOD-aware index.

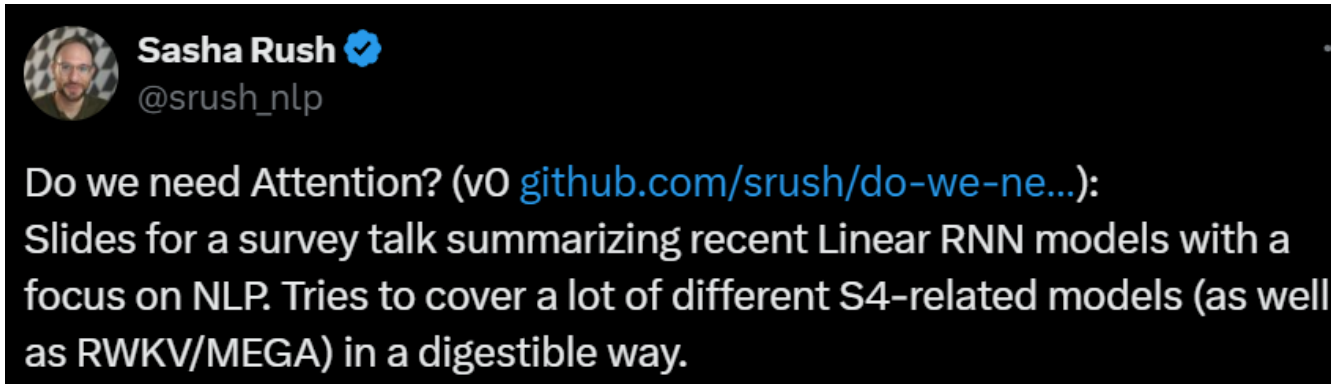
- RetrievalAttention
  - Compute attention on CPU and GPU separately → Reduce excessive commutation overhead over PCIe
  - Link each query to its exact top-k nearest key vectors using KNN → Mitigate the distribution gap of query and key vectors

Liu D, Chen M, Lu B, et al. RetrievalAttention: Accelerating Long-Context LLM Inference via Vector Retrieval[J]. arXiv preprint arXiv:2409.10516, 2024.

New Model Architectures

Is Attention the only way?

Can we use other efficient architectures? e.g., RNN

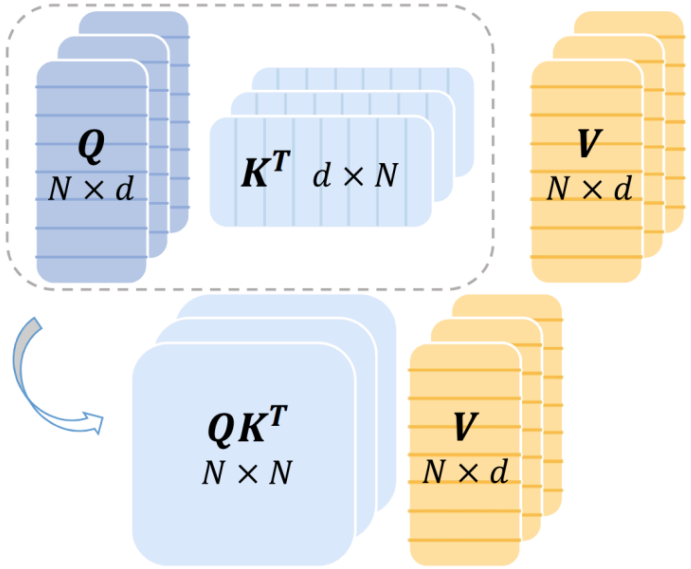


Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$

New Model Architectures · Linear Attention

$$\text{Vanilla Attention}(Q, K, V) = \text{Softmax}(QK^T)\,V$$

- In self-attention, $Q,\ K,\ V \in R^{N\times d}$, $N >> d$
- The step QK^T yields an $N \times N$ matrix, leading to $O(N^2)$ complexity.



Softmax Attention $\mathcal{O}(N^2 d)$

Challenge: The $\text{Softmax}(\cdot)$ operation needs to be computed over full rows of the score matrix QK^T, outputs depend on the sum in the denominator.

$$s\left(x_i\right) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

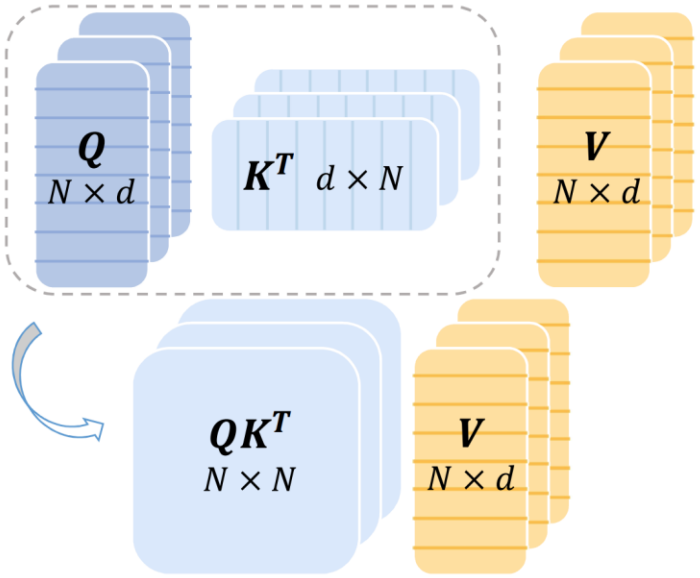
New Model Architectures · Linear Attention

$$\text{Vanilla Attention}(Q, K, V) = \text{Softmax}(QK^T)\,V$$

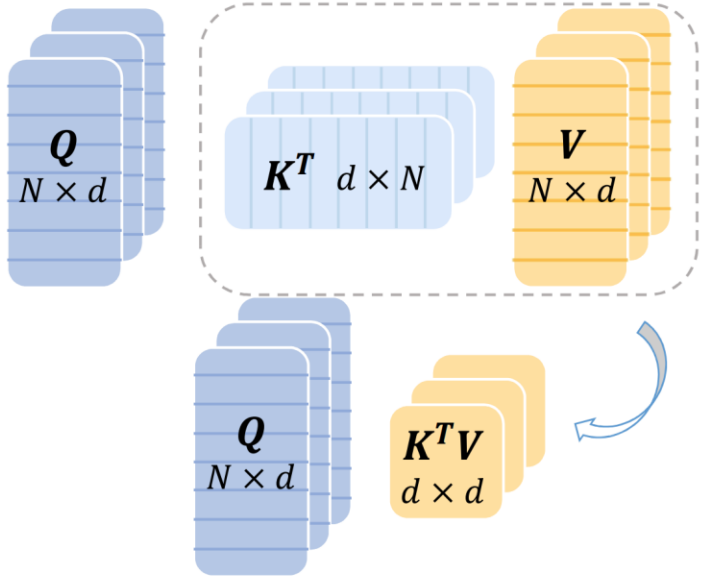
- In self-attention, Q, K, $V \in R^{N\times d}$, $N >> d$
- The step QK^T yields an $N \times N$ matrix, leading to $O(N^2)$ complexity.

If there is no $\text{Softmax}(\cdot)$ function:

- We can calculate $K^T V$ first to obtain a matrix of $d \times d$
- Since d<<n, the complexity is only $O(n)$



Softmax Attention $\mathcal{O}(N^2 d)$



Linear Attention $\mathcal{O}(Nd^2)$

New Architecture · Linear Attention

Vanilla Attention

$$Q, K, V = XW_Q, XW_K, XW_V$$
$$O = \text{softmax}\big((QK^T) \odot M\big)V,$$

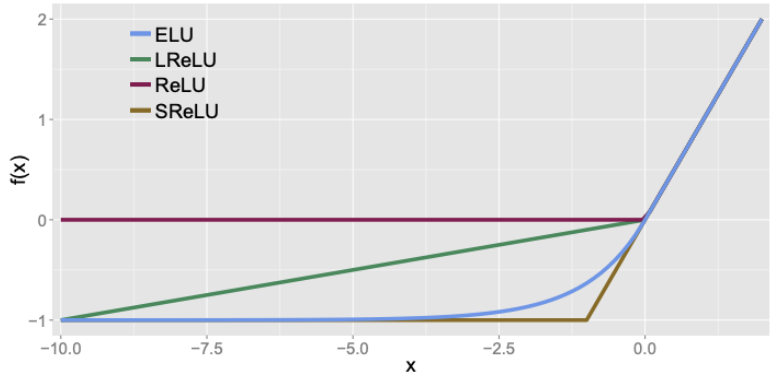
where $W_Q, W_K, W_V \in R^{d \times d}$ are learnable matrices and M is the mask matrices.

$$q_t, k_t, v_t = x_t W_Q, x_t W_K, x_t W_V$$
$$o_t = \frac{\sum_{i=1}^{t} \exp(q_t k_i^T) v_i}{\sum_{i=1}^{t} \exp(q_t k_i^T)}$$

which calculates the query (q_t), key (k_t), and value(v_t) vectors given the current token's representation $x_t \in R^{1 \times d}$.

We can replace $\exp(q_t k_i^T)$ with a kernel function $g(q_t, k_i)$ with an associated feature map ϕ (i.e., $g(q_t, k_i) = <\phi(q_t), \phi(k_i)>$)

$$o_t = \frac{\sum_{i=1}^{t} \phi(q_t)\phi(k_i)^T v_i}{\sum_{i=1}^{t} \phi(q_t)\phi(k_i)^T} = \frac{\phi(q_t) \sum_{i=1}^{t} \phi(k_i)^T v_i}{\phi(q_t) \sum_{i=1}^{t} \phi(k_i)^T}$$



e.g., $\phi(x) = \text{elu}(x) + 1$

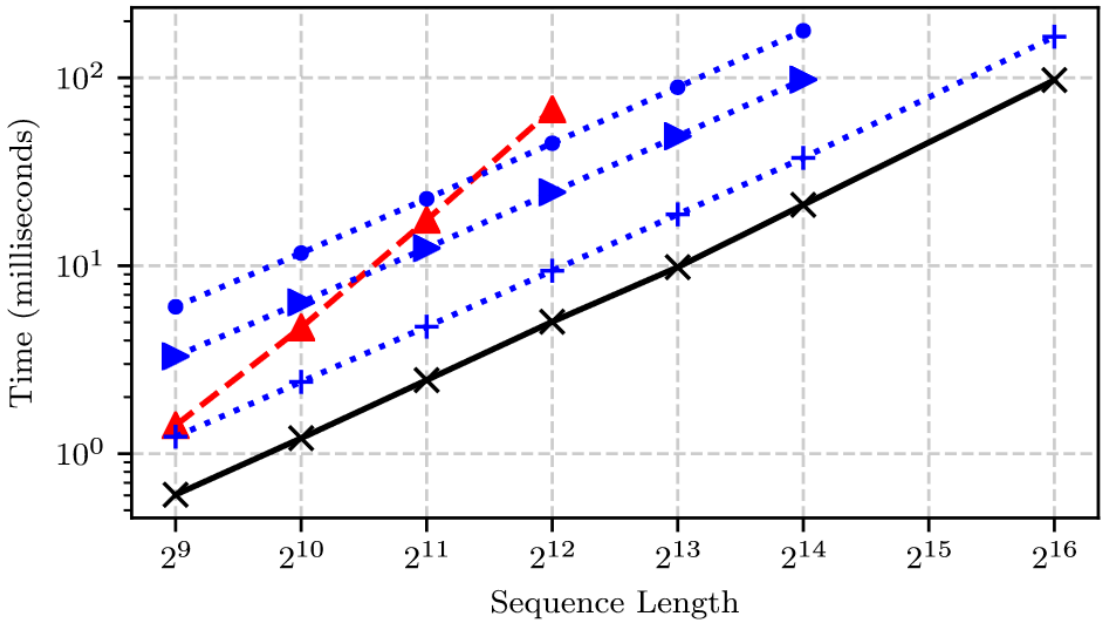
We can also rewrite the function of Linear Attention into RNN format:

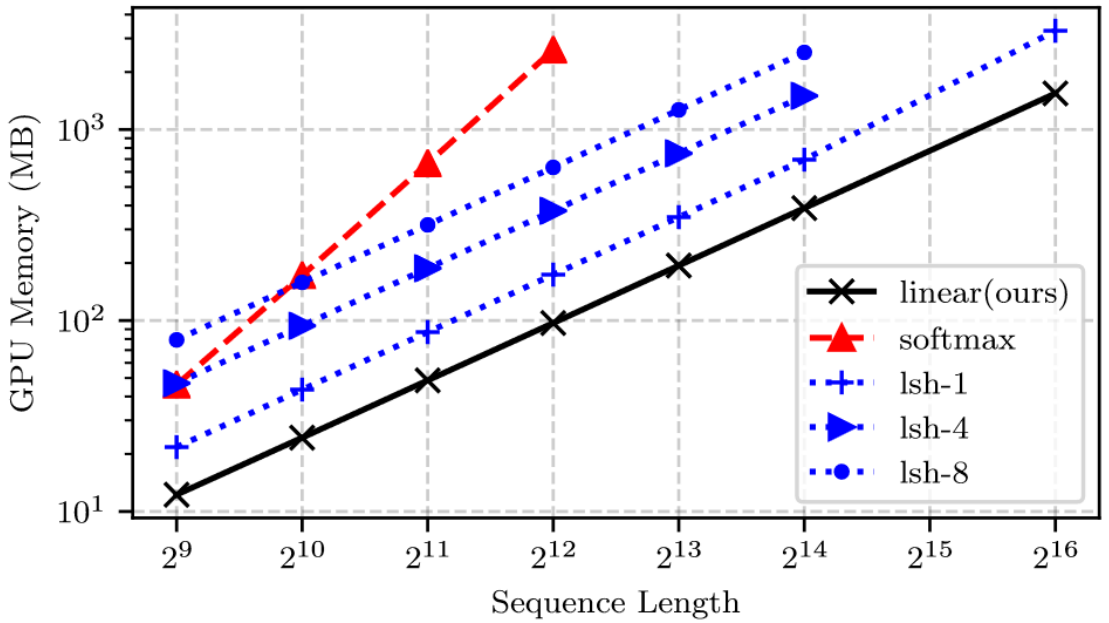
$$S_t = \sum_{i=1}^{t} \phi(k_i)^T v_i \ (S_t \in R^{d \times d}),$$
$$z_t = \sum_{i=1}^{t} \phi(k_i)^T \ (z_t \in R^{d \times 1})$$

$$\Rightarrow$$

$$S_t = S_{t-1} + \phi(k_t)^T v_t,$$
$$z_t = z_{t-1} + \phi(k_t)^T,$$
$$o_t = \frac{\phi(q_t) S_t}{\phi(q_t) z_t}$$

New Architecture · Linear Attention

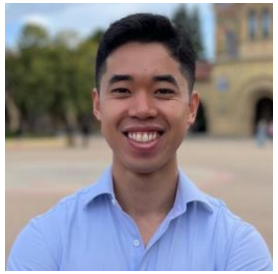




Linear attention scale linearly with the sequence length unlike softmax which scales with the square of the sequence length both in memory and computation

New Architecture • State Space Models

Albert Gu
Assistant Professor @ CMU



Tri Dao
Assistant Professor @ Princeton

“ Quadratic attention has been indispensable for information-dense modalities such as language …

…

Announcing Mamba: a new SSM arch. that has linear-time scaling, **ultra long context**, and most importantly outperforms Transformers everywhere we've tried ”

New Architecture · State Space Models

Transformer
(NeurIPS,2017)

H3 (ICLR,2023)

Hyena (ICML,2023)

Monarch (NeurIPS,2023)

Mamba (COLM,2024)

State Space Model (SSM)

$$h'_t = Ah(t) + Bx(t)$$
$$y_t = Ch(t) + Dx(t)$$

Structed State Space Model (S4)

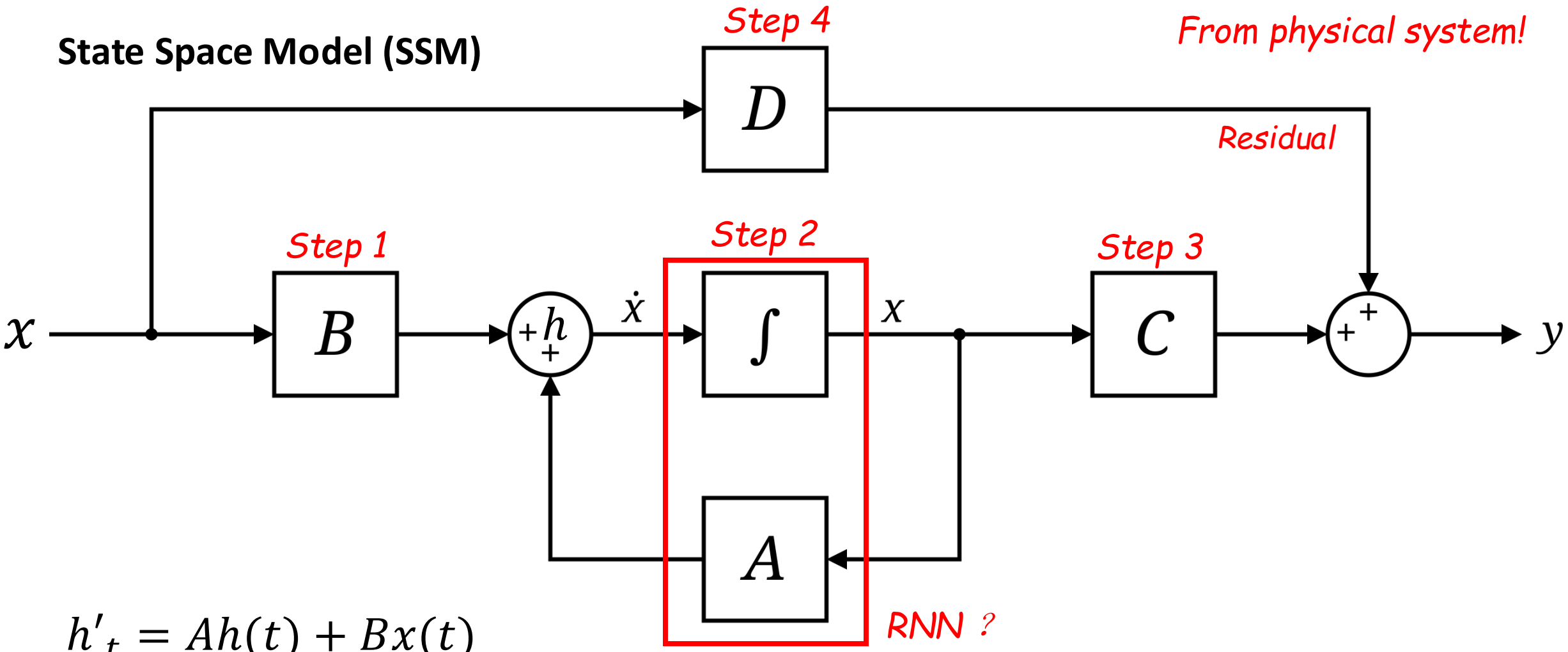
$$h_t = \bar{A}h_{t-1} + \bar{B}x_t$$
$$y_t = Ch_t$$

Selective State Space Model (Mamba)

$$h_t = s_{\bar{A}}(x_t)h_{t-1} + s_{\bar{B}}(x_t)x_t$$
$$y_t = s_{\bar{C}}(x_t)h_t$$

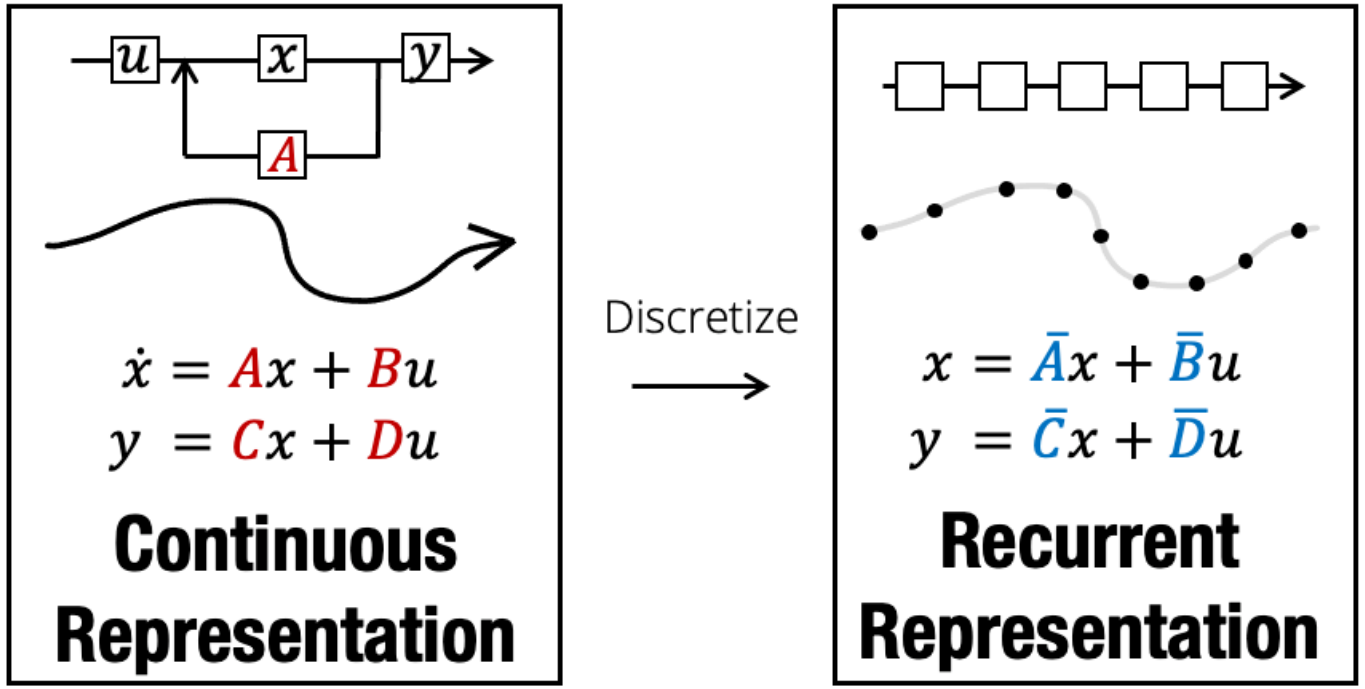
New Architecture · State Space Models

State Space Model (SSM)



$$h'_t = Ah(t) + Bx(t)$$
$$y_t = Ch(t) + Dx(t)$$

Structed State Space Model (S4)



Natural language is discrete

Selective Structed State Space Model (Mamba)

- In S4, discrete parameters $\bar{A}$, $\bar{B}$ and $\bar{C}$ are constant, which we called Linear Time Invariance (LTI)
- Model pays equal attention to the output of each position.

S4: $h_t = \bar{A}h_{t-1} + \bar{B}x_t$
$y_t = Ch_t$

Mamba: $h_t = s_{\bar{A}}(x_t)h_{t-1} + s_{\bar{B}}(x_t)x_t$
$y_t = s_{\bar{C}}(x_t)h_t$

Algorithm 1 SSM (S4)

Input: x : (B, L, D)
Output: y : (B, L, D)
1: $\boldsymbol{A}$: (D, N) ← Parameter
▷ Represents structured $N \times N$ matrix
2: $\boldsymbol{B}$: (D, N) ← Parameter
3: $\boldsymbol{C}$: (D, N) ← Parameter
4: Δ : (D) ← τ_Δ(Parameter)
5: $\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}$: (D, N) ← discretize($\Delta, \boldsymbol{A}, \boldsymbol{B}$)
6: y ← SSM($\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C}$)($x$)
▷ Time-invariant: recurrence or convolution
7: **return** y

Algorithm 2 SSM + Selection (S6)

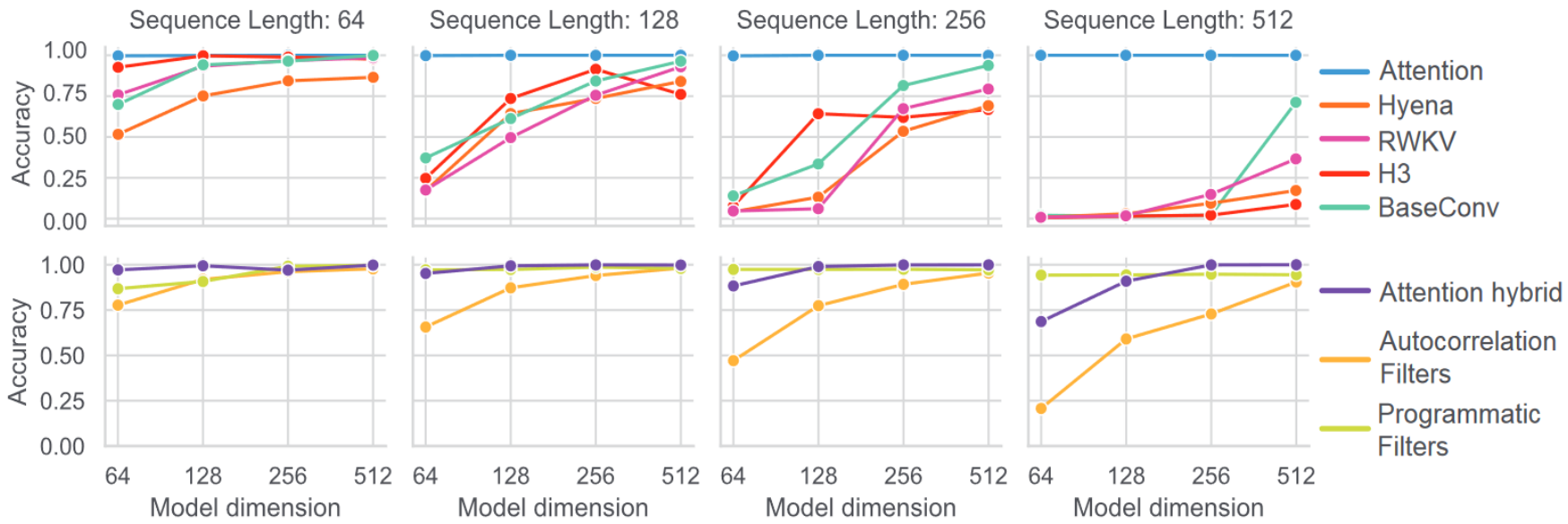
Input: x : (B, L, D)
Output: y : (B, L, D)
1: $\boldsymbol{A}$: (D, N) ← Parameter
▷ Represents structured $N \times N$ matrix
2: $\boldsymbol{B}$: (B, L, N) ← $s_B(x)$
3: $\boldsymbol{C}$: (B, L, N) ← $s_C(x)$
4: Δ : (B, L, D) ← τ_Δ(Parameter+$s_\Delta(x)$)
5: $\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}$: (B, L, D, N) ← discretize($\Delta, \boldsymbol{A}, \boldsymbol{B}$)
6: y ← SSM($\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C}$)($x$)
▷ Time-varying: recurrence (*scan*) only
7: **return** y

Selective Structed State Space Model (Mamba)

Model	Token.	Pile ppl ↓	LAMBADA ppl ↓	LAMBADA acc ↑	HellaSwag acc ↑	PIQA acc ↑	Arc-E acc ↑	Arc-C acc ↑	WinoGrande acc ↑	Average acc ↑
Hybrid H3-130M	GPT2	—	89.48	25.77	31.7	64.2	44.4	24.2	50.6	40.1
Pythia-160M	NeoX	29.64	38.10	33.0	30.2	61.4	43.2	24.1	**51.9**	40.6
Mamba-130M	NeoX	**10.56**	**16.07**	**44.3**	**35.3**	**64.5**	**48.0**	**24.3**	**51.9**	**44.7**
Hybrid H3-360M	GPT2	—	12.58	48.0	41.5	68.1	51.4	24.7	54.1	48.0
Pythia-410M	NeoX	9.95	10.84	51.4	40.6	66.9	52.1	24.6	53.8	48.2
Mamba-370M	NeoX	**8.28**	**8.14**	**55.6**	**46.5**	**69.5**	**55.1**	**28.0**	**55.3**	**50.0**
Pythia-1B	NeoX	7.82	7.92	56.1	47.2	70.7	57.0	27.1	53.5	51.9
Mamba-790M	NeoX	**7.33**	**6.02**	**62.7**	**55.1**	**72.1**	**61.2**	**29.5**	**56.1**	**57.1**
GPT-Neo 1.3B	GPT2	—	7.50	57.2	48.9	71.1	56.2	25.9	54.9	52.4
Hybrid H3-1.3B	GPT2	—	11.25	49.6	52.6	71.3	59.2	28.1	56.9	53.0
OPT-1.3B	OPT	—	6.64	58.0	53.7	72.4	56.7	29.6	59.5	55.0
Pythia-1.4B	NeoX	7.51	6.08	61.7	52.1	71.0	60.5	28.5	57.2	55.2
RWKV-1.5B	NeoX	7.70	7.04	56.4	52.5	72.4	60.5	29.4	54.6	54.3
Mamba-1.4B	NeoX	**6.80**	**5.04**	**64.9**	**59.1**	**74.2**	**65.5**	**32.8**	**61.5**	**59.7**
GPT-Neo 2.7B	GPT2	—	5.63	62.2	55.8	72.1	61.1	30.2	57.6	56.5
Hybrid H3-2.7B	GPT2	—	7.92	55.7	59.7	73.3	65.6	32.3	61.4	58.0
OPT-2.7B	OPT	—	5.12	63.6	60.6	74.8	60.8	31.3	61.0	58.7
Pythia-2.8B	NeoX	6.73	5.04	64.7	59.3	74.0	64.1	32.9	59.7	59.1
RWKV-3B	NeoX	7.00	5.24	63.9	59.6	73.7	67.8	33.1	59.6	59.6
Mamba-2.8B	NeoX	**6.22**	**4.23**	**69.2**	**66.1**	**75.2**	**69.7**	**36.3**	**63.5**	**63.3**
GPT-J-6B	GPT2	–	4.10	68.3	66.3	75.4	67.0	36.6	64.1	63.0
OPT-6.7B	OPT	–	4.25	67.7	67.2	76.3	65.6	34.9	65.5	62.9
Pythia-6.9B	NeoX	6.51	4.45	67.1	64.0	75.2	67.3	35.5	61.3	61.7
RWKV-7.4B	NeoX	6.31	4.38	67.2	65.5	76.1	67.8	37.5	61.0	62.5

For each model size, Mamba is best-in-class on every single evaluation result and matches baselines at twice the model size.

Selective Structed State Space Model (Mamba)



RNN-based models still lag behind Attention-based models on **Recall task**

Simran Arora
PhD student
@ Stanford

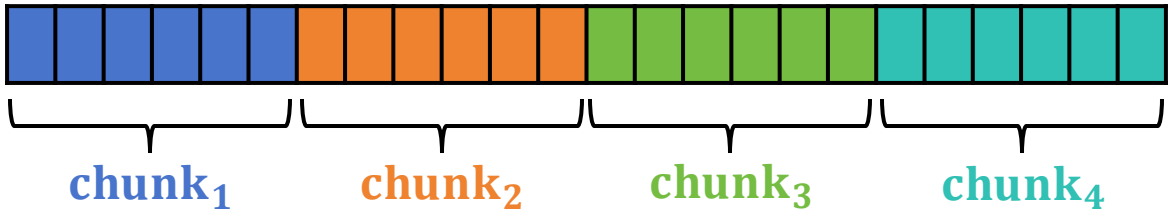
Hazy Research Blog, Stanford

"We find that there is a consistent perplexity gap between the SoTA attention-free models and Transformers."

"We found that all architectures obeyed a fundamental tradeoff: the less memory the model consumed during inference, the worse it did on associative recall. In attention, the state is commonly referred to as the KV-cache, and it grows with the length of the sequence."

Long-context Alignment • Infrastructure

The quadratic complexity comes from the dot product between Q (L) and K(L), what if we calculate it chunk by chunk?



After obtaining attention results for each chunk, how we get the final attention weights for the entire sequence?

Vanilla Attention

Vanilla Attention: $\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$

Let $w_{xy} = \frac{q_x k_y^T}{\sqrt{d}}$, then $\text{Attn}(q_x,K,V) = \frac{\sum_{y=1}^{L} e^{w_{xy}} v_y}{\sum_{y=1}^{L} e^{w_{xy}}}$, where L is the sequence length.

Attention for each chunk

Split K,V into N chunks: $K = \{k_1, k_2, \cdots, k_N\}, V = \{v_1, v_2, \cdots, v_N\}$, where B is the chunk size.

Attention for each chunk can be written as:

$$\text{Attn}(q_x, K_i, V_i) = \frac{\sum_{y=i\times B}^{(i+1)\times B} e^{w_{xy}} v_y}{\sum_{y=i\times B}^{(i+1)\times B} e^{w_{xy}}}$$

Long-context Alignment · Chunk Attention

Attention on entire sequence = Merge all chunks

Let $\mathrm{Attn}(q_x, K_i, V_i) = \frac{\sum_{y=i\times B}^{(i+1)\times B} e^{w_{xy}} v_y}{\sum_{y=i\times B}^{(i+1)\times B} e^{w_{xy}}} = \frac{\sum_{j=1}^{n} A_j}{\sum_{j=1}^{n} B_j}$, we have $B_{1\cdots n} = \sum_{i=1}^{n} B_i$

$$\mathrm{attn}_{12} = \frac{A_{12}}{B_{12}} = \frac{A_1+A_2}{B_1+B_2} = \frac{A_1}{B_1} * \frac{B_1}{B_1+B_2} + \frac{A_2}{B_2} * \frac{B_2}{B_1+B_2} = \mathrm{attn}_1 \frac{B_1}{B_{12}} + \mathrm{attn}_2 \frac{B_2}{B_{12}}$$

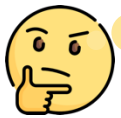
$$\mathrm{attn}_{123} = \mathrm{attn}_{12} \frac{B_{12}}{B_{123}} + \mathrm{attn}_3 \frac{B_3}{B_{123}}$$

…

$$\text{(full attention) } \mathrm{attn}_{1\cdots n} = \mathrm{attn}_{1\cdots n-1} \frac{B_{1\cdots n-1}}{B_{1\cdots n}} + \mathrm{attn}_n \frac{B_n}{B_{1\cdots n}}$$

Therefore, we only need to record $\mathrm{attn_i}$ and its corresponding $\mathrm{B_i}$, and we can calculate the Full Attention through an iterative approach.

We can compute attention in distribute setting



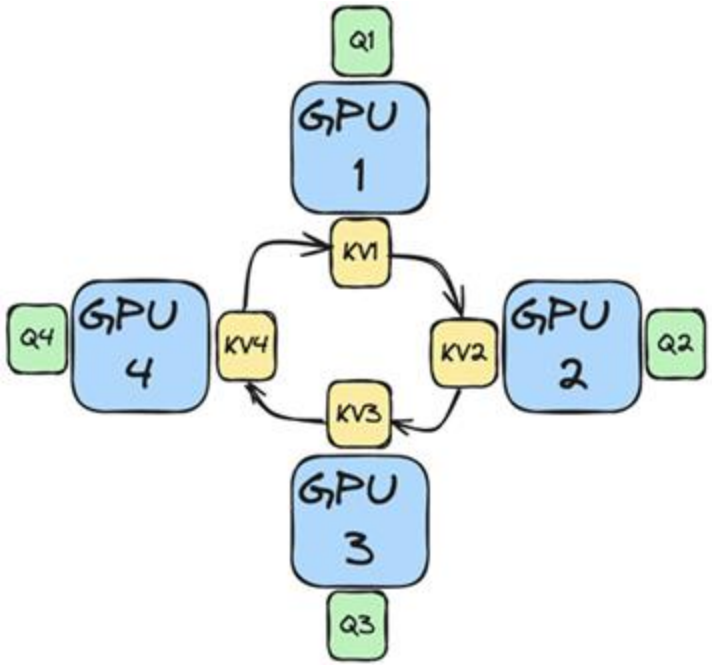
If I have more than 1 GPU, can I leverage the iterative strategy above to accelerate the computation process?

Long-context Alignment • Ring Attention

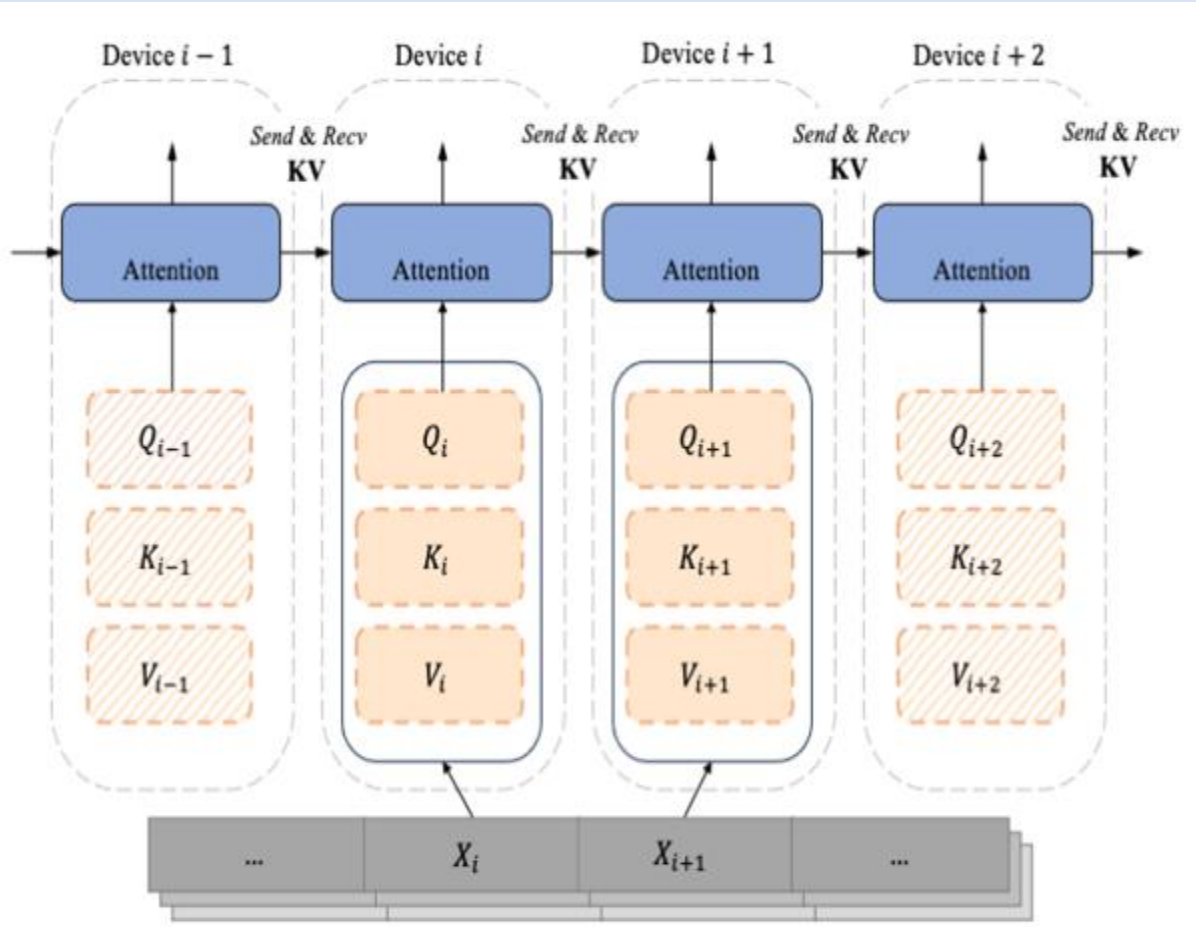
Main Concept of Ring Attention

Hosts Form a Conceptual Ring to exchange KV segments; one pass completes when every node has seen all parts of the KV

- Arbitrary Order of Block Computations
- Distributing QKV Sequence Across Hosts

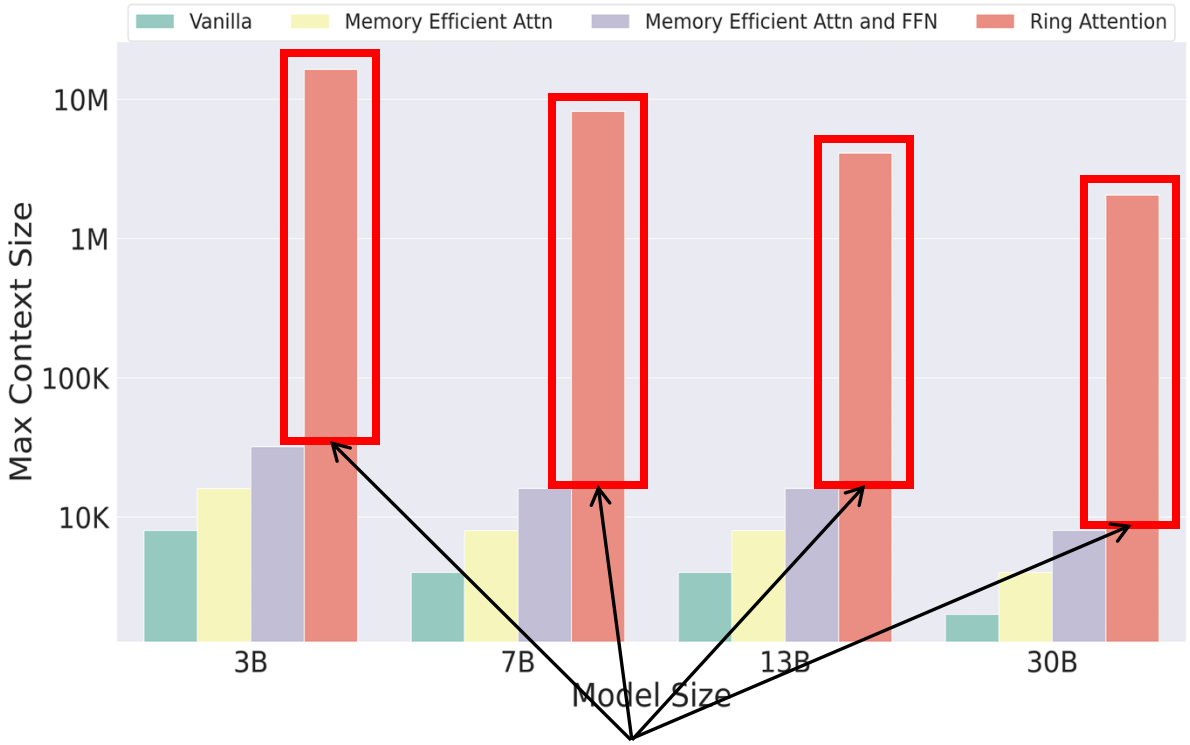


Sequence Parallel



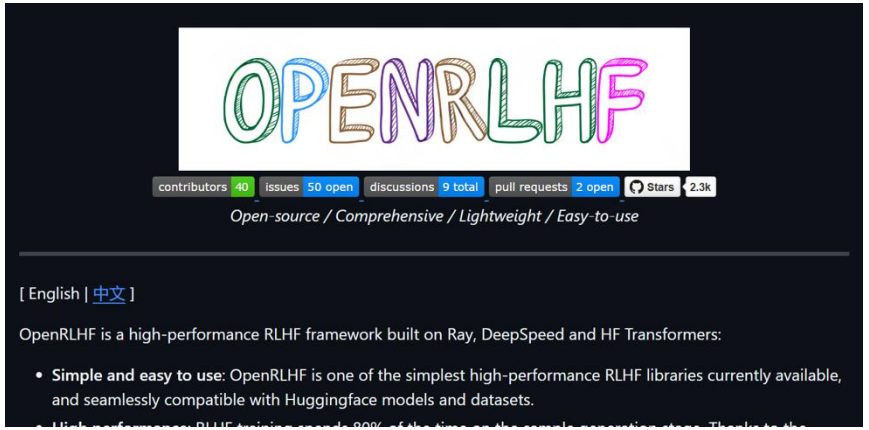
Liu H, Zaharia M, Abbeel P. Ring attention with blockwise transformers for near-infinite context[J]. arXiv preprint arXiv:2310.01889, 2023.

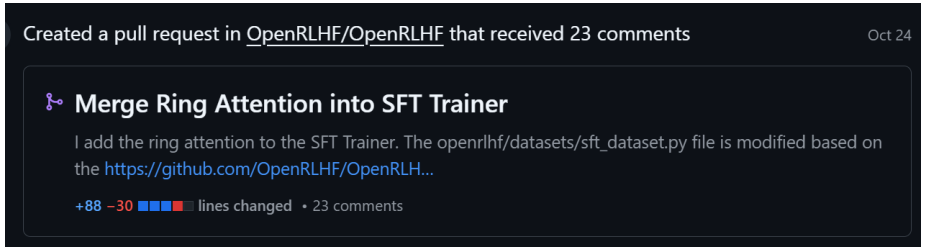
Long-context Alignment • Ring Attention



Ring attention significantly reduces GPU memory compared to other efficient methods.

Two open-source repos for Ring Attention

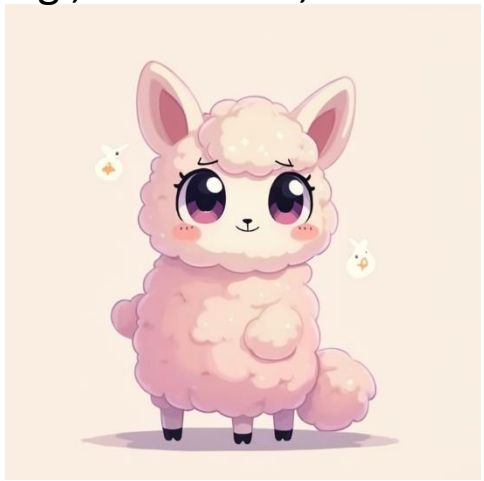




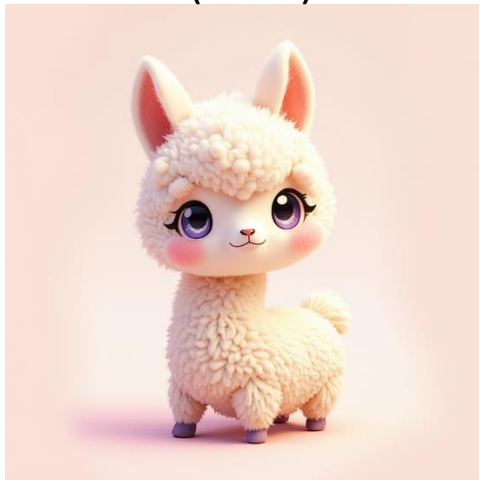
If encounter any problem when utilizing SFT Ring_Attn in OpenRLHF, plz start an issue 😊

Long-context Window Alignment

Open-source short-context model
(e.g., Llama2-4K, Llama3-8K)



Model with long context window
(>32K)



Powerful long-context model

Context Window Scaling

- ✓ Relative Positional Embedding (RPE)
- ✓ Rotary Positional Embedding (RoPE)
  - ✓ Position Interpolation (PI)
  - ✓ Position Extrapolation (PE)

Long-context Alignment

- ✓ Supervised Fine-Tuning (SFT)
- ✓ Reinforcement Learning (RL)

Sec. 2.2 Data

Data Resource

- *General Pretraining Data*
  - Code
  - Book
  - Web
  - *etc*
- *Synthetic Data*

Construction

- *Splice*
- *Up-sampling & Mixture*
- *Positional Synthesis*
- *Model Generation*

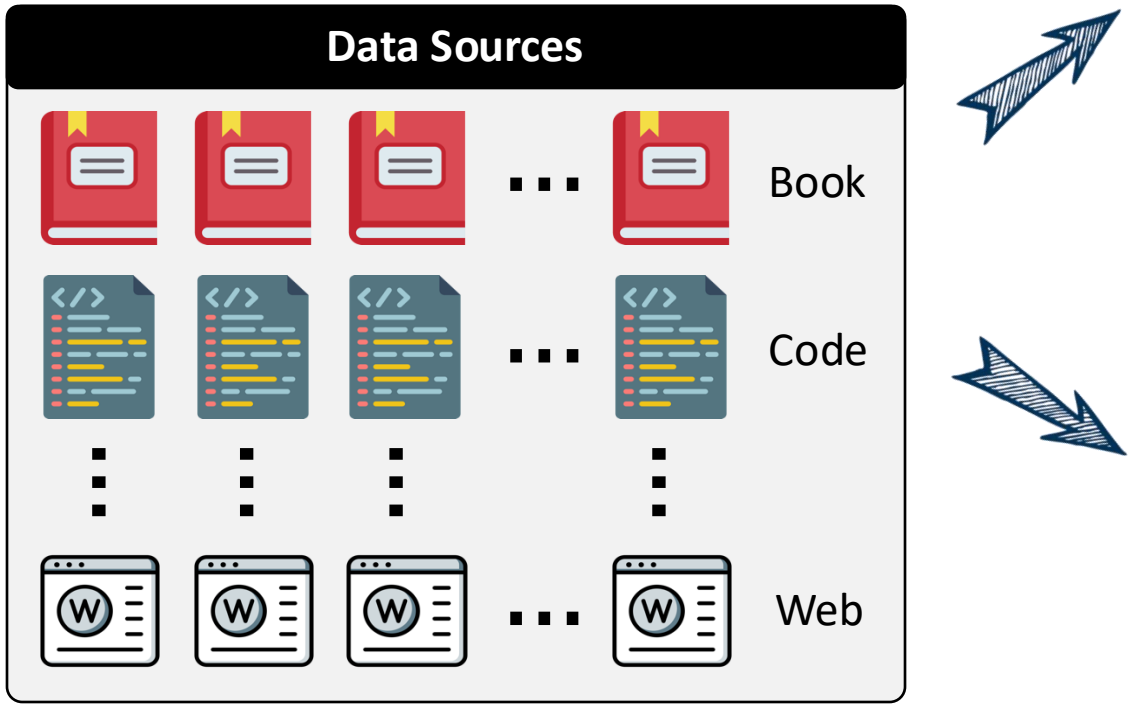
Usage

- *Supervised Fine-tuning*
- *Reinforcement Learning*

Long-context Data Construction · Splice & Up-sampling

- **Splicing:** Chunking and randomly sampling short-context data from different data sources and then concatenating them

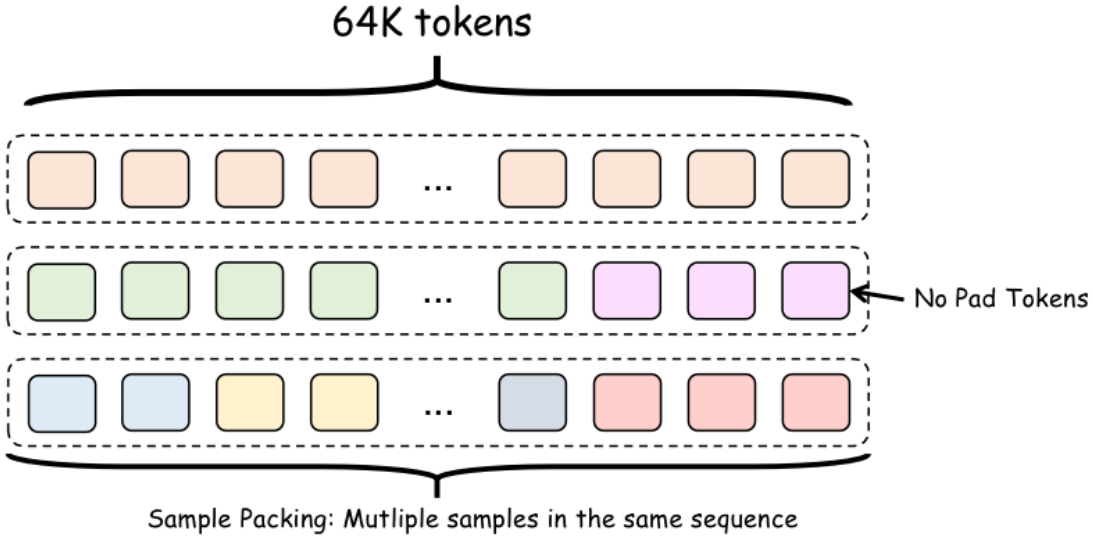
- **Up-sampling**: Directly search for text that naturally has long context length.



4K - 128K Context Length

Original	1.560	1.650	0.786	1.075	1.313	1.852	0.447
v.s. Per-source	– .010	– .010	– .006	– .011	– .044	– .014	+ .002
v.s. Global	– .010	– .006	– .001	– .016	– .040	– .018	– .007
v.s. Code↑	– .008	– .002	– .003	– .007	– .042	– .010	– .029
v.s. Book↑	– .010	– .006	+ .001	– .007	– .037	– 0.30	+ .000
v.s. Arxiv↑	– .008	– .002	+ .002	– .036	– .039	– .010	– .004

Data engineering for scaling language models to 128k context (ICML, 2024)

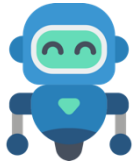


Effective long-context scaling of foundation models (NAACL, 2024)

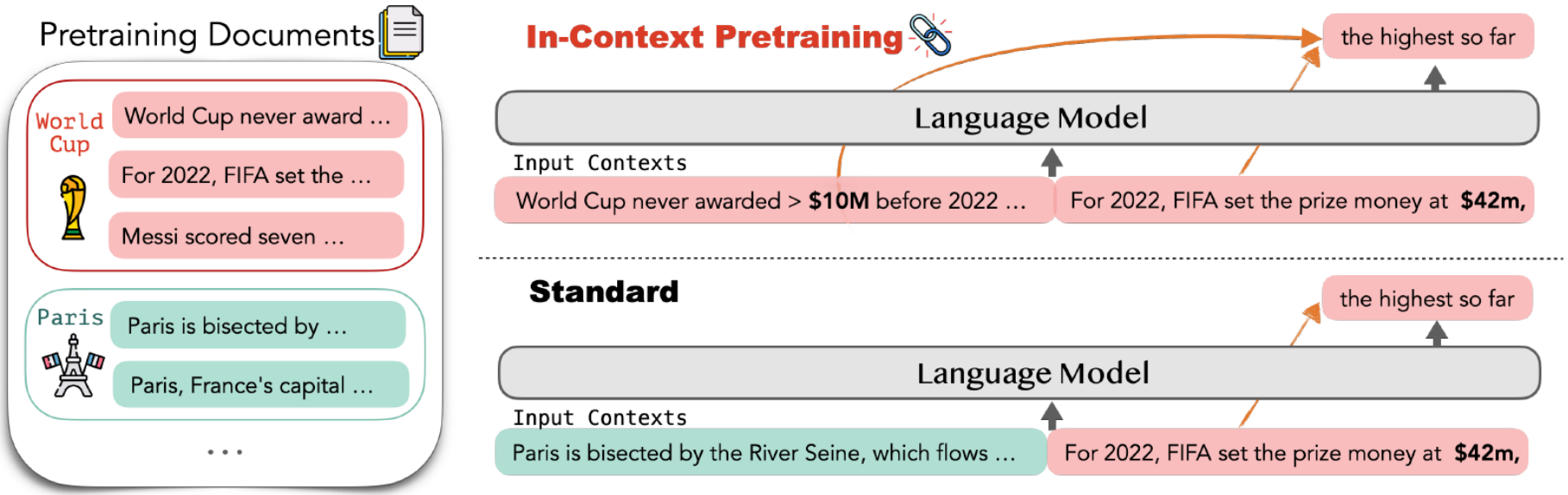
Long-context Data Construction · Splice

How to select the short-context data and how to concatenating?

Maybe similarity-based method is ok…

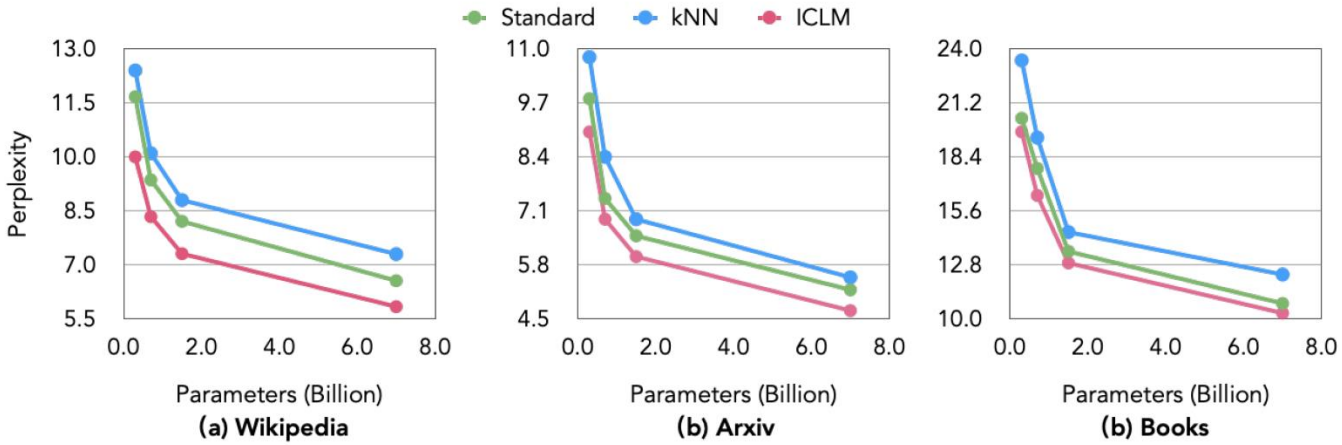


ICLM (In-context Pretraining: Language Modeling beyond Document Boundaries, ICLR 2024)



Different from the standard pretraining strategy that place randomly shuffled documents in the input context, ICLM places related documents in the same context, making models learn to reason across prior documents

Long-context Data Construction • Splice



ICLM outperforms the standard (trained with randomly spliced data) and kNN-based (trained with kNN-sorted data) LLM.

When constructing data with **Splice** method, try to consider

- Context dependency
- Segment quality is balanced
- Whether the segment will interfere with the final answer
- *etc.*

Low quality —— often serve for context window scaling

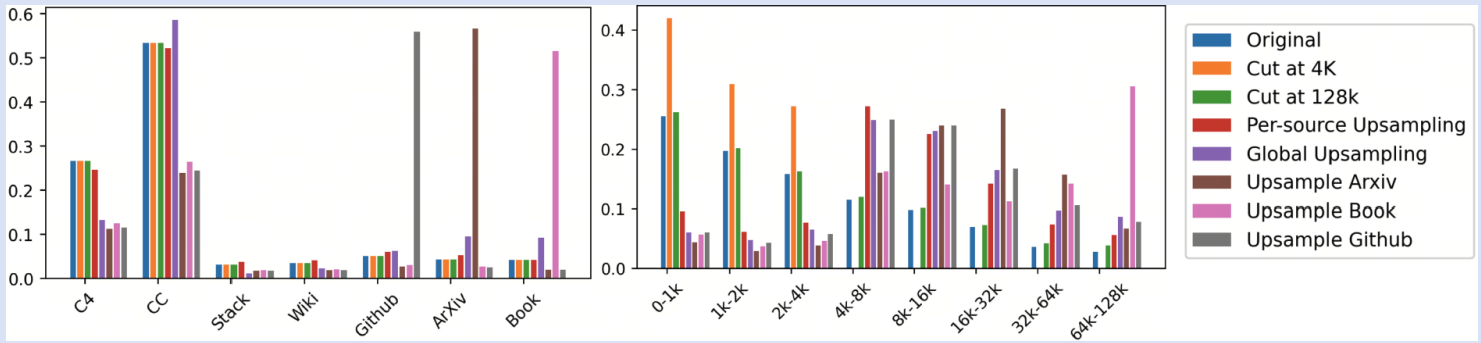
Long-context Data Construction • Up-sampling

One way to improve the data quality:
Utilize Data with Natural Long Context.

	Ctx.	Needle.	MMLU
Non-LLaMA Models			
GPT-4-Turbo	128K	**87.1**	**86.4**
GPT-3.5-Turbo	16K	-	67.3
YaRN Mistral 7B	128K	57.4	59.4
LLaMA-2 7B Based Models			
Together LLaMA-2 7B	32K	27.9	**44.8**
LongChat v1.5 7B	32K	18.0	42.3
LongLoRA 7B	100K	70.0	37.9
Ours LLaMA-2 7B	80K	**88.0**	43.3
LLaMA-2 13B Based Models			
LongLoRA 13B	64K	54.1	50.1
Ours LLaMA-2 13B	64K	**90.0**	**52.4**

c) Up-sampling long-context data from pre-training corpus can maintain the model performance on short-context testing set.

Data Engineering for Scaling Language Models to 128K Context (ICML, 2024)



a) There exists long-context data (book and code) in the pre-training corpus

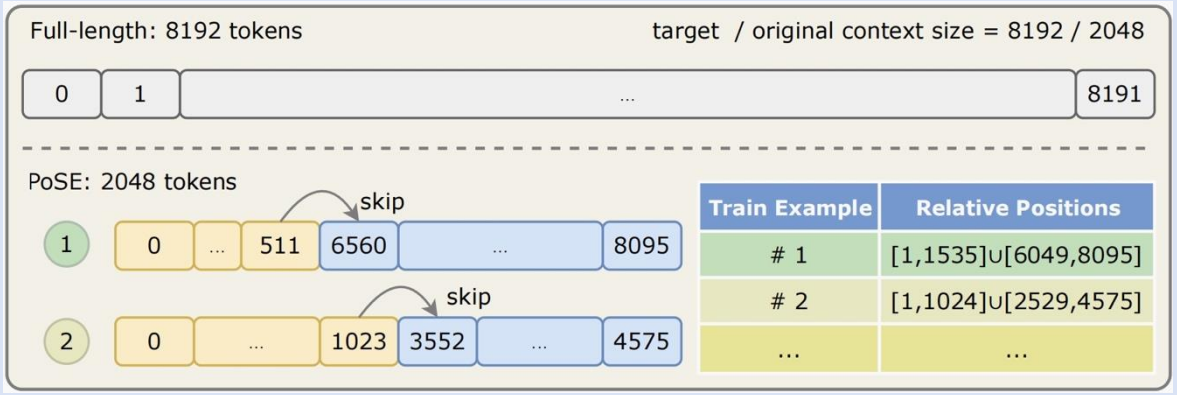
	C4	CC	Stack	Arxiv	Wiki	Book	Github
0 - 4K Context Length							
Original	2.038	1.760	1.519	1.660	1.424	2.085	0.907
v.s. Per-source	+ .002	+ .008	– .001	– .008	– .040	– .065	– .008
v.s. Global	+ .008	+ .010	+ .015	– .020	– .020	– .140	+ .015
v.s. Code↑	+ .010	+ .016	+ .010	+ .006	– .026	+ .030	– .023
v.s. Book↑	+ .010	+ .016	+ .021	+ .000	– .010	– .175	+ .029
v.s. Arxiv↑	+ .006	+ .016	+ .013	– .060	– .030	+ .040	+ .025
4K - 128K Context Length							
Original	1.560	1.650	0.786	1.075	1.313	1.852	0.447
v.s. Per-source	– .010	– .010	– .006	– .011	– .044	– .014	+ .002
v.s. Global	– .010	– .006	– .001	– .016	– .040	– .018	– .007
v.s. Code↑	– .008	– .002	– .003	– .007	– .042	– .010	– .029
v.s. Book↑	– .010	– .006	+ .001	– .007	– .037	– 0.30	+ .000
v.s. Arxiv↑	– .008	– .002	+ .002	– .036	– .039	– .010	– .004

Train on 80K length data, test on different length data

b) Data distribution matte-rs; sampling data from a single domain can impair model's performance on other domains, especially in short-context test setting.

Long-context Data Construction • Positional Index Synthesis

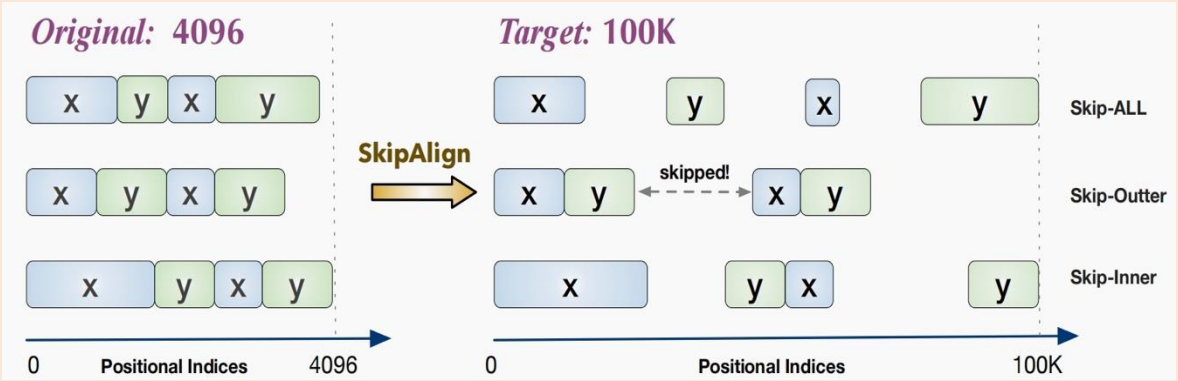
Instead of increasing the actual input sequence length, we can construct "long'' context data by manipulating positional indices.



Positional Skip-wisE (PoSE), ICLR 2024

Method	Context size	GovReport					Proof-pile				
	Train / Target	2k	4k	8k	16k	32k	2k	4k	8k	16k	32k
Original	- / -	4.74	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	2.83	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
Full-length	16k / 16k	4.87	4.70	4.61	4.59	-	2.93	2.71	2.58	2.53	-
RandPos	2k / 16k	11.63	11.17	11.54	15.16	-	7.26	6.83	6.76	7.73	-
	2k / 32k	93.43	95.85	91.79	93.22	97.57	60.74	63.54	60.56	63.15	66.47
PoSE (Ours)	2k / 16k	4.84	4.68	4.60	4.60	-	2.95	2.74	2.61	2.60	-
	2k / 32k	4.91	4.76	4.68	4.64	4.66	3.01	2.78	2.66	2.60	2.59

PoSE: language modelling results (PPL on long-context tasks)



Step-Skipping Alignment (SkipAlign), Arxiv 2024

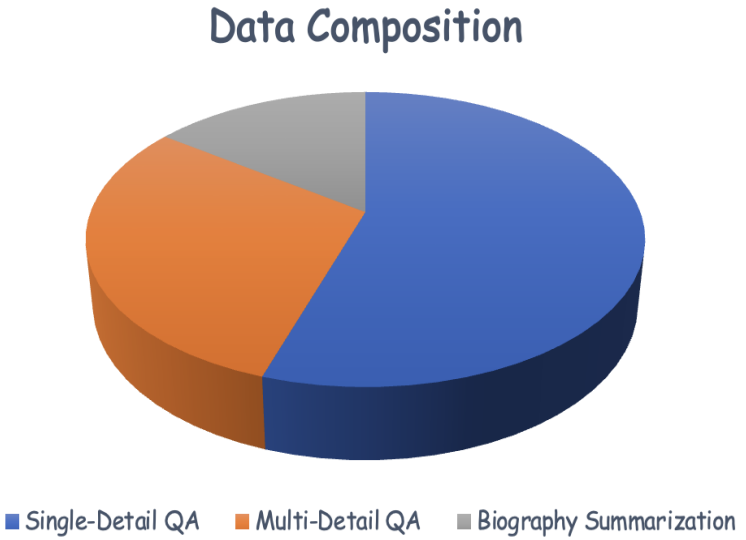
Model	Avg.	S-Doc QA	M-Doc QA	Summ	Few-shot	Code
GPT-3.5-Turbo-16k	44.6	39.7	38.7	26.5	67.0	54.2
LLAMA-2-7B Based Models						
LLAMA-2-7B-chat-4k	35.2	24.9	22.5	25.0	60.0	48.1
SEext-LLAMA-2-7B-chat-16k	38.7	27.3	26.2	24.8	64.2	57.5
LongChat1.5-7B-32k	36.9	28.7	20.6	26.6	60.0	54.2
LLAMA-2-7B-NTK32k	31.7	16.2	7.3	15.4	66.7	**63.4**
+ Normal-SFT	41.5	31.3	32.7	26.0	65.3	57.4
+ PackedSFT-16k	42.6	31.6	32.8	26.2	67.9	60.5
+ PackedSFT-32k	41.6	30.0	32.2	26.2	67.3	58.0
+ PackedSFT-50k	43.6	36.0	**37.0**	**27.7**	63.8	58.5
+ SkipAlign	**44.1**	**38.6**	33.8	26.1	**67.6**	59.6

SkipAlign: real-world tasks (LongBench)

Long-context Data Construction · Model Generation

Extending Llama-3's Context Ten-Fold Overnight (Arxiv 2024)

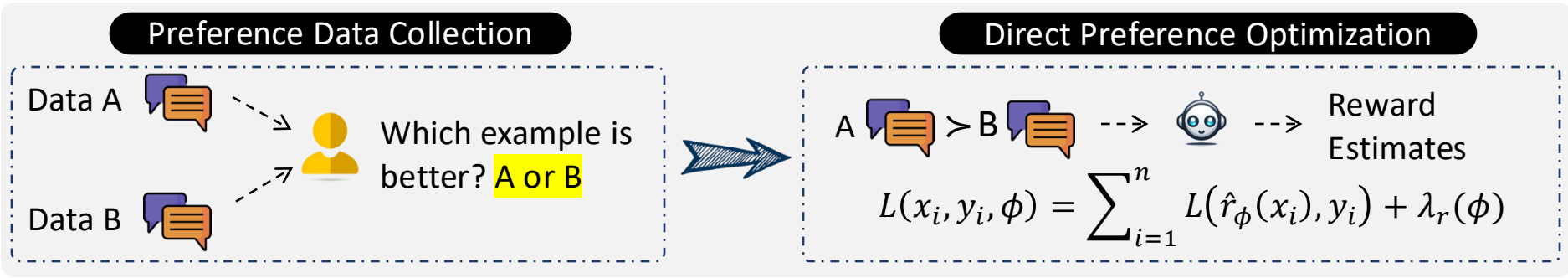
Single-Detail QA	Prompt GPT-4 to answer question of one specific detail in the long context
Multi-Detail QA	Prompt GPT-4 to aggregate and reason over multiple details in the long context
Biography Summarization	Prompt GPT-4 to write a biography for each main character in a given book.



Model	Single-Doc	Multi-Doc	Summ.	Few-Shot	Synthetic	Code	Avg
Llama-3-8B-Instruct	37.33	36.04	26.83	**69.56**	37.75	53.24	43.20
Llama-3-8B-Instruct-262K	37.29	31.20	26.18	67.25	44.25	**62.71**	43.73
Llama-3-8B-Instruct-80K-QLoRA	**43.57**	**43.07**	**28.93**	69.15	**48.50**	51.95	**47.19**

Using synthetic data constructed from GPT-4 can greatly improve the long-context model performance.

Long-context Alignment • Dataset for RL

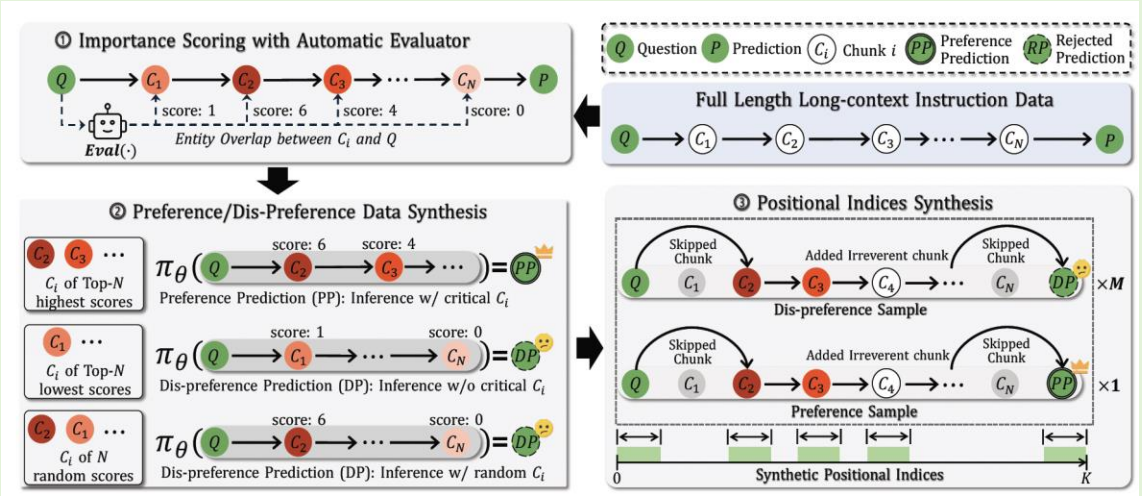


Create long-context preference data is challenge:

- Lack of ground truth label
  - Model annotation: costly
  - Human annotation: extremely hard
- Hard to distinguish between good & negative samples
  - Prompts lead to universally good or poor responses
  - Lack of evaluation model

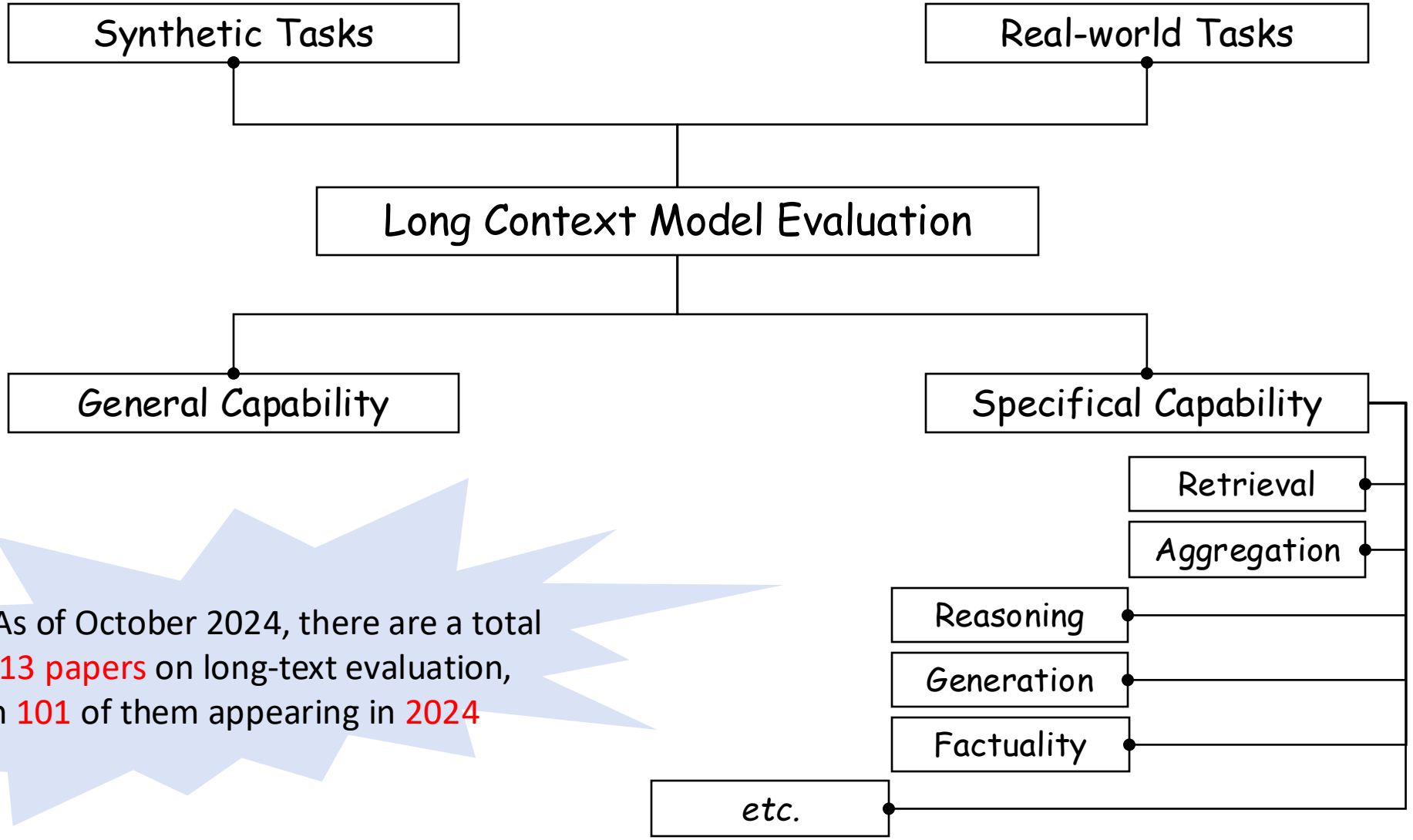
- Lack of ground truth label
  - ✓ Create labels from short-context data
  - ✓ Expanding the short-text with context-scaling methods, such as using relevant context, positional indices synthesis.
- Hard to distinguish between good & negative samples
  - ✓ Each to check with short-context data

LOGO —— Long cOntext aliGnment via efficient preference Optimization (Arxiv, 2024)



TL;DR Utilize critical segments within the context to guide short-context model to generate synthetic data

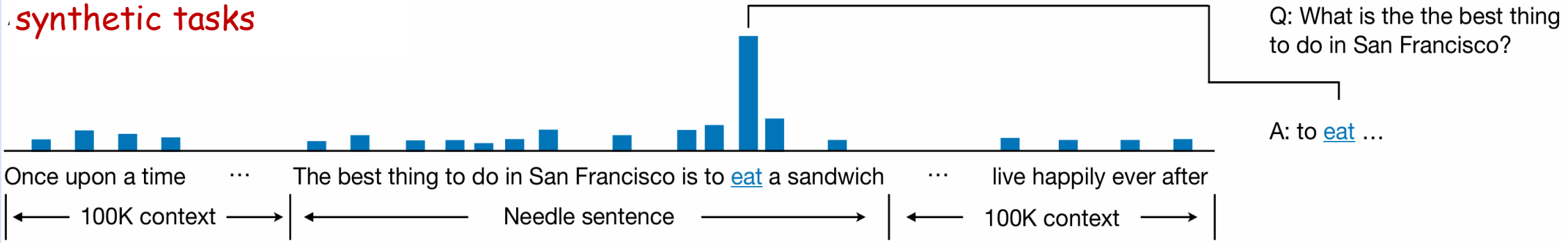
Long-context Model Evaluation



As of October 2024, there are a total of 113 papers on long-text evaluation, with 101 of them appearing in 2024

Long-context Model Evaluation

synthetic tasks



Pros of synthetic tasks:

- More controllable
- Mitigating the impact of LCMs' intrinsic Knowledge

General format of synthetic tasks:

- Long irreverent context
- Key information within the context

real-world tasks

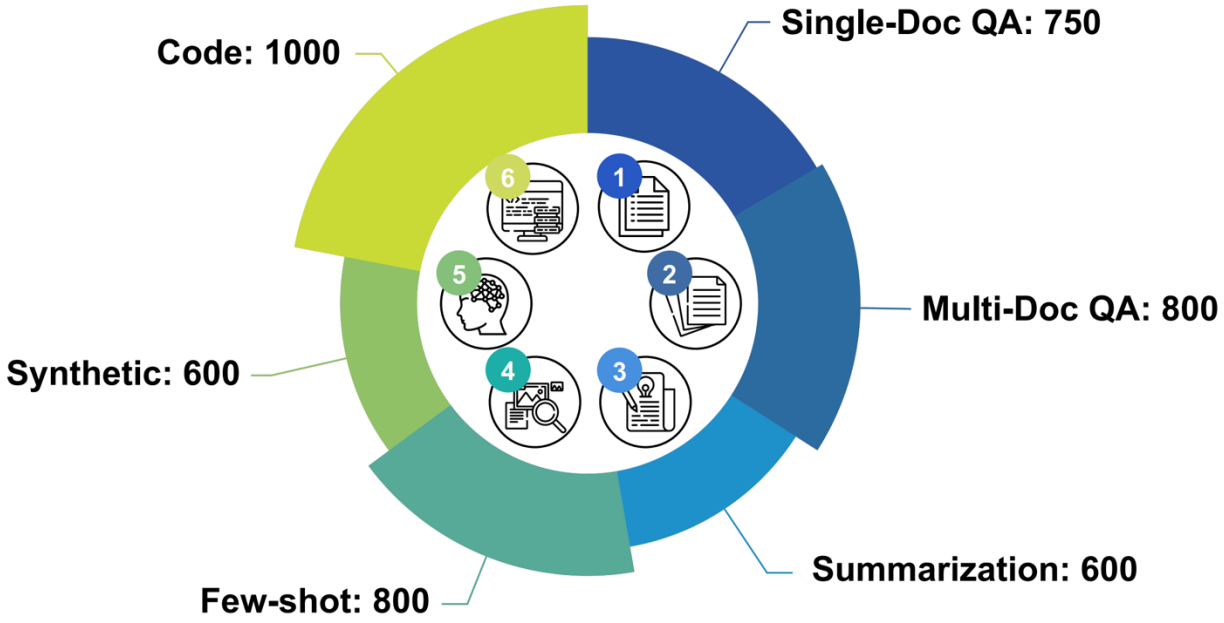
Closed - Ended Tasks						
TOEFL	Multiple choice	English test	3,907	4,171	269	15
GSM(16-shot)†	Solving math problems	In-context examples	5,557	5,638	100	100
QuALITY †	Multiple choice	Gutenberg	7,169	8,560	202	15
Coursera*	Multiple choice	Advanced courses	9,075	17,185	172	15
TopicRet†	Retriving topics	Conversation	12,506	15,916	150	50
SFcition*	True or False Questions	Scientific fictions	16,381	26,918	64	7
CodeU*	Deducing program outputs	Python Codebase	31,575	36,509	90	90

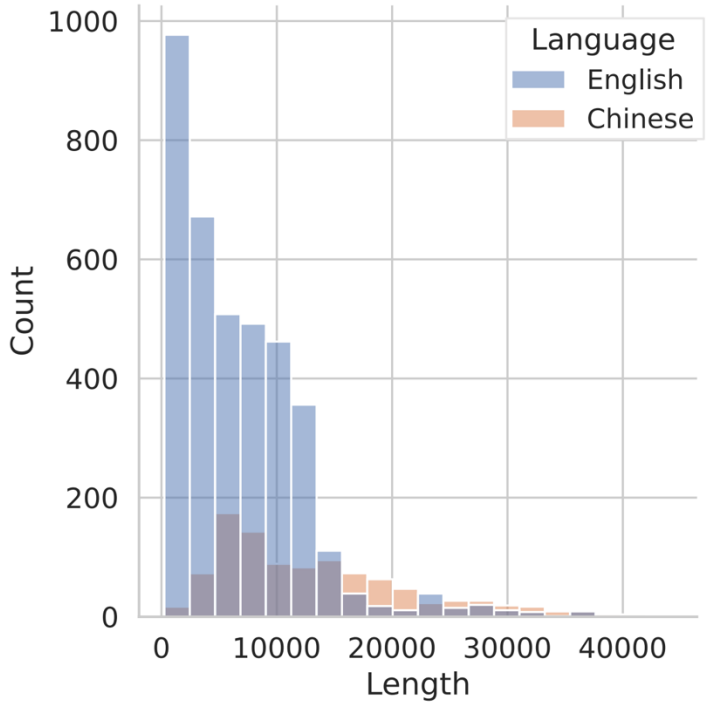
L-Eval: Instituting Standardized Evaluation for Long Context Language Models (ACL 2024)

- Requiring the model to respond while ``resisting'' interference information

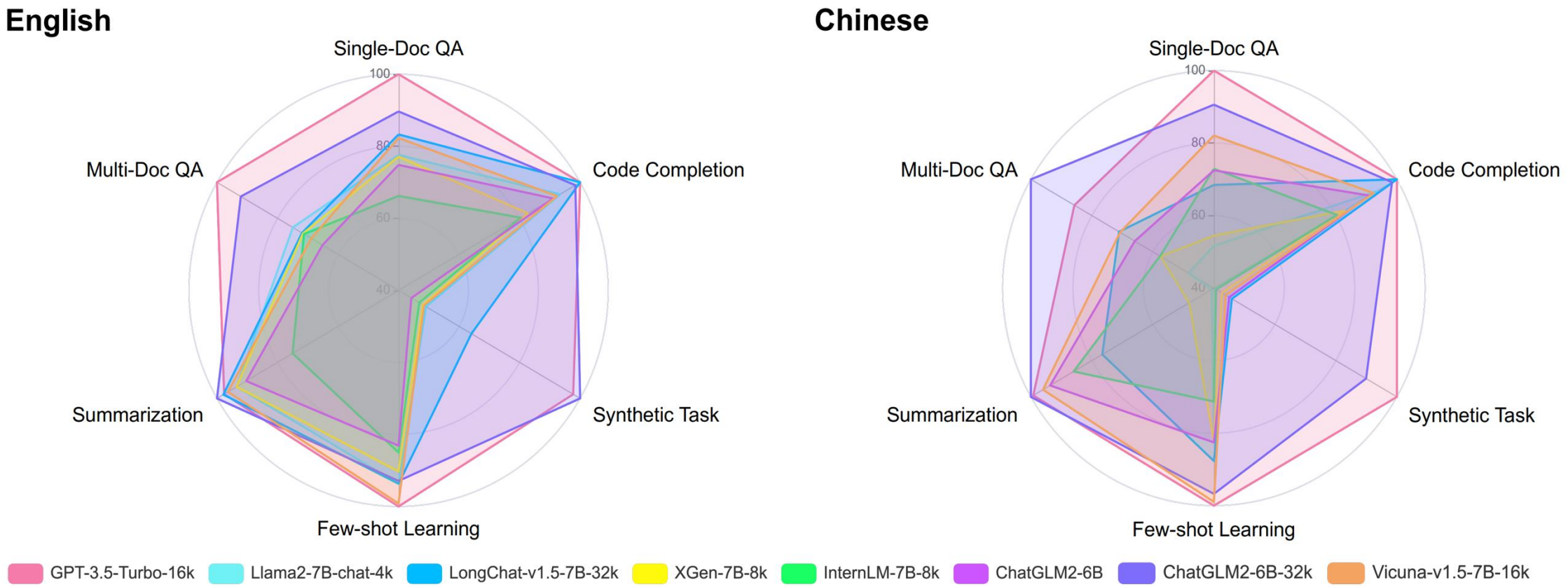
Long-context Model Evaluation · General Capability

LongBench *(ACL 2024)* is the first bilingual, multi-task benchmark for long-context understanding in long-context understanding (both real-world & synthetic tasks)





Long-context Model Evaluation · General Capability



- Gaps between open-source models and close-source models (GPT-3.5-Turbo)
- Models benefit from scaled positional embedding and continued training on longer context, as ChatGLM2-6B-32K and LongChat-v1.5-7B-32K obtain relative improvements of 62% and 19%, respectively

Evaluation • General Capability

RULER: What's the Real Context Size of Your Long-Context Language Models? (COLM 2024)

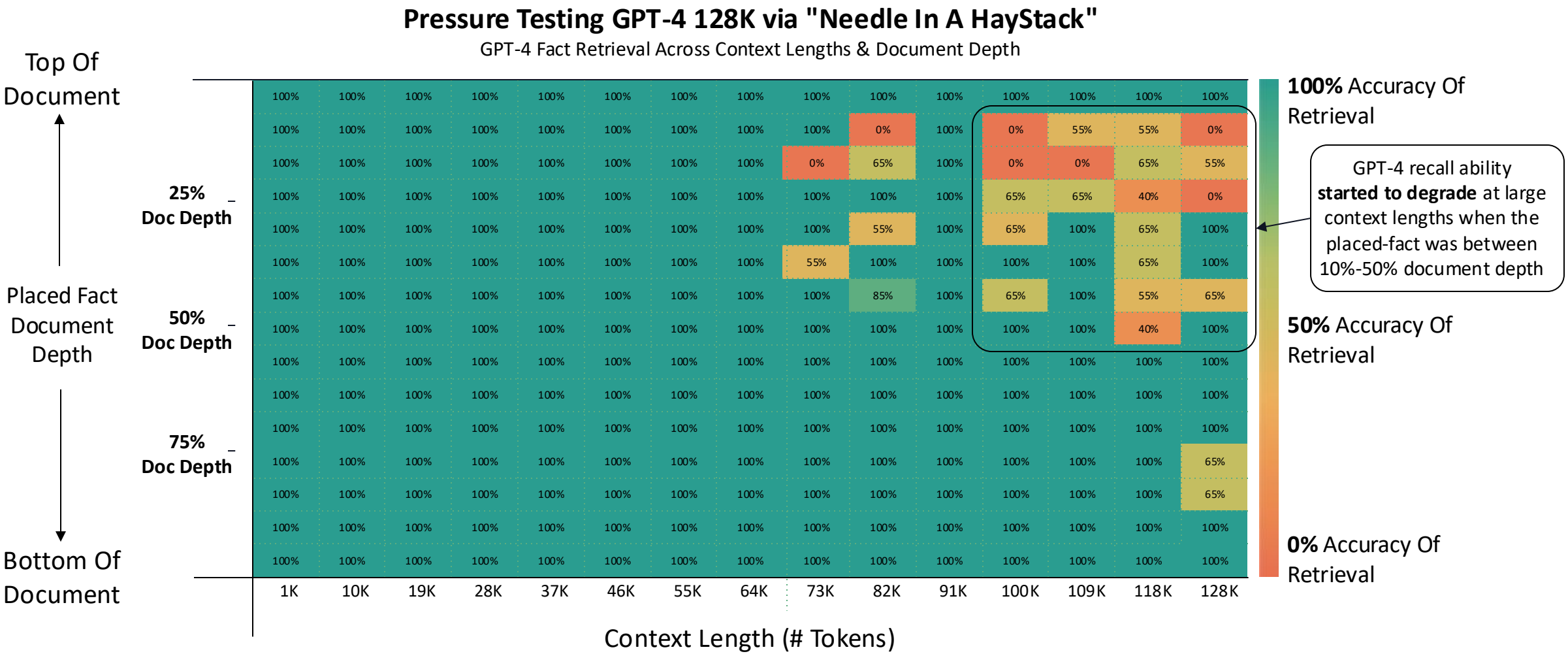
Task	Configuration	Example
Single NIAH (S-NIAH)	type_key = word type_value = number type_haystack = essay size_haystack ∝ context length	(essays) ..... One of the special magic numbers for long-context is: 12345. ..... What is the special magic number for long-context mentioned in the provided text?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

Evaluation · General Capability

Long-context Model Arena

Models	Claimed Length	Effective Length	4K	8K	16K	32K	64K	128K	Avg.	wAvg. (inc)	wAvg. (dec)
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	95.5 (1st)	96.1 (1st)
GPT-4	128K	64K	96.6	96.3	95.2	93.2	87.0	81.2	91.6	89.0 (2nd)	94.1 (2nd)
Llama3.1 (70B)	128K	64K	96.5	95.8	95.4	94.8	88.4	66.6	89.6	85.5 (4th)	93.7 (3rd)
Qwen2 (72B)	128K	32K	96.9	96.1	94.9	94.1	79.8	53.7	85.9	79.6 (9th)	92.3 (4th)
Command-R-plus (104B)	128K	32K	95.6	95.2	94.2	92.0	84.3	63.1	87.4	82.7 (7th)	92.1 (5th)
GLM4 (9B)	1M	64K	94.7	92.8	92.1	89.9	86.7	83.1	89.9	88.0 (3rd)	91.7 (6th)
Llama3.1 (8B)	128K	32K	95.5	93.8	91.6	87.4	84.7	77.0	88.3	85.4 (5th)	91.3 (7th)
GradientAI/Llama3 (70B)	1M	16K	95.1	94.4	90.8	85.4	80.9	72.1	86.5	82.6 (8th)	90.3 (8th)
Mixtral-8x22B (39B/141B)	64K	32K	95.6	94.9	93.4	90.9	84.7	31.7	81.9	73.5 (11th)	90.3 (9th)
Yi (34B)	200K	32K	93.3	92.2	91.3	87.5	83.2	77.3	87.5	84.8 (6th)	90.1 (10th)
Phi3-medium (14B)	128K	32K	93.3	93.2	91.1	86.8	78.6	46.1	81.5	74.8 (10th)	88.3 (11th)
Mistral-v0.2 (7B)	32K	16K	93.6	91.2	87.2	75.4	49.0	13.8	68.4	55.6 (13th)	81.2 (12th)
LWM (7B)	1M	<4K	82.3	78.4	73.7	69.1	68.1	65.0	72.8	69.9 (12th)	75.7 (13th)
DBRX (36B/132B)	32K	8K	95.1	93.8	83.6	63.1	2.4	0.0	56.3	38.0 (14th)	74.7 (14th)
Together (7B)	32K	4K	88.2	81.1	69.4	63.0	0.0	0.0	50.3	33.8 (15th)	66.7 (15th)
LongChat (7B)	32K	<4K	84.7	79.9	70.8	59.3	0.0	0.0	49.1	33.1 (16th)	65.2 (16th)
LongAlpaca (13B)	32K	<4K	60.6	57.0	56.6	43.6	0.0	0.0	36.3	24.7 (17th)	47.9 (17th)

Evaluation • Retrieval Capability



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

A fact was placed within a document. GPT-4 (1106-preview) was then asked to retrieve it. The output was evaluated for accuracy.
This test was run at 15 different document depths (top > bottom) and 15 different context lengths (1K >128K tokens).
2x tests were run for larger contexts for a larger sample size.

Evaluation • Reasoning Capability

Counting-Stars: A Simple, Efficient, and Reasonable Strategy for Evaluating Long-Context Large Language Models (Arxiv 2024)

Long-Context Multi-evidence Acquisition *English Version*	November 2005In the next few years, venture capital funds will find themselves squeezed from four directions. They're already stuck with a seller's market, because of the huge amounts they raised at the end of the Bubble and still haven't invested. This by itself is not the end of the world. In fact, it's just a more extreme version of the norm in the VC business: too much money chasing too few deals.Unfortunately, those few deals now want less and less money, because it's getting so cheap to start a startup ... ***The little penguin counted {number1} ★*** ... Moore's law, which makes hardware geometrically closer to free; the Web, which makes promotion free if you're good; and better languages, which make development a lot cheaper.When we started our startup in 1995, the first three were our biggest expenses. We had to pay $5000 for the Netscape Commerce Server, the only software that then supported secure http connections ... ***The little penguin counted {number2} ★*** ... people throw away computers more powerful than our first server ... . ***On this moonlit and misty night, the little penguin is looking up at the sky and concentrating on counting ★. Please help the little penguin collect the number of ★, for example: "little_penguin": [x, x, x,...]. The summation is not required, and the numbers in [x, x, x,...] represent the counted number of ★ by the little penguin. Only output the results in JSON format without any explanation.***
Long-Context Multi-evidence Reasoning *English Version*	November 2005In the next few years, venture capital funds will find themselves squeezed from four directions. They're already stuck with a seller's market, because of the huge amounts they raised at the end of the Bubble and still haven't invested. This by itself is not the end of the world. In fact, it's just a more extreme version of the norm in the VC business: too much money chasing too few deals.Unfortunately, those few deals now want less and less money, because it's getting so cheap to start a startup ... ***The little penguin counted {wrong number1} ★, but found that a mistake had been made, so the counting was done again, and this time {number1} ★ was counted correctly.*** ... Moore's law, which makes hardware geometrically closer to free; the Web, which makes promotion free if you're good; and better languages, which make development a lot cheaper.When we started our startup in 1995, the first three were our biggest expenses. We had to pay $5000 for the Netscape Commerce Server, the only software that then supported secure http connections ... ***The little penguin counted {wrong number2} ★, but found that a mistake had been made, so the counting was done again, and this time {number2} ★ was counted correctly.*** ... people throw away computers more powerful than our first server .... ***On this moonlit and misty night, the little penguin is looking up at the sky and concentrating on counting ★. Please help the little penguin collect the correct number of ★, for example: "little_penguin": [x, x, x,...]. The summation is not required, and the numbers in [x, x, x,...] represent the correctly counted number of ★ by the little penguin. Only output the results in JSON format without any explanation.***

Evaluation · Reasoning Capability

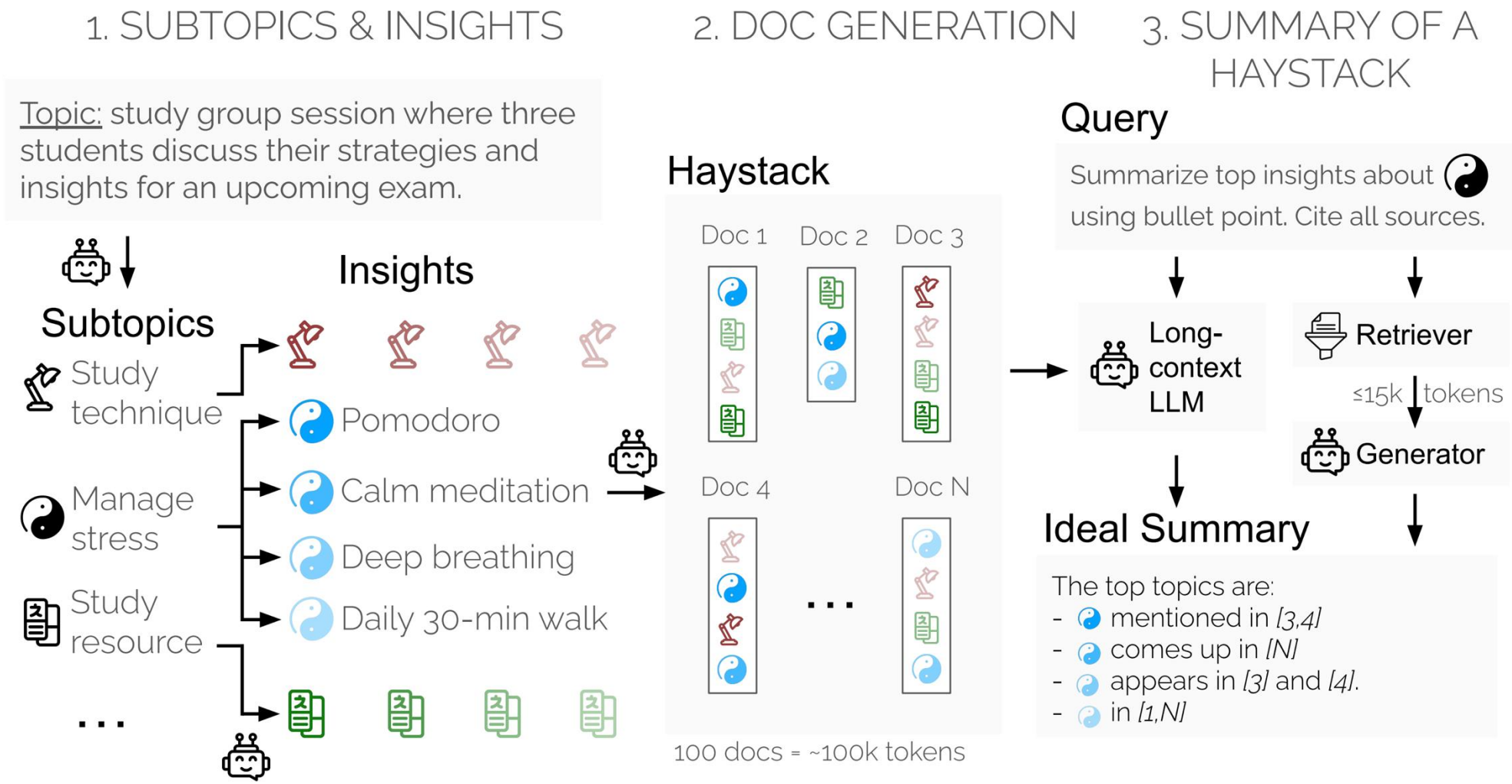
Counting-Stars: A Simple, Efficient, and Reasonable Strategy for Evaluating Long-Context Large Language Models (Arxiv 2024)

Models	GPT-4 Turbo		Gemini 1.5 Pro	Claude3			GLM-4	Moonshot-v1
	1106	*0125*		Opus	Sonnet	Haiku		
Multi-evidence Acquisition (ZH)	0.697	0.663	0.775	0.807	0.788	0.698	0.682	0.606
Multi-evidence Acquisition (EN)	0.718	0.662	0.833	0.705	-	-	0.389	0.559
Multi-evidence Reasoning (ZH)	0.473	0.386	0.575	0.488	-	-	0.475	0.344
Multi-evidence Reasoning (EN)	0.651	0.610	0.371	0.374	-	-	0.179	0.460
Average Score	0.635_2	0.580_4	0.639_1	0.594_3	-	-	0.431_6	0.492_5

"Although cutting-edge long-context LLMs have achieved nearly perfect performance on the needle-in-a-haystack task, they still perform poorly on the Counting-Stars test, which indicates that the needle-in-a-haystack is too simple to truly show the capabilities of LLMs in processing long texts"

Evaluation • Aggregation Capability

Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems (Arxiv 2024)



Evaluation · Aggregation Capability

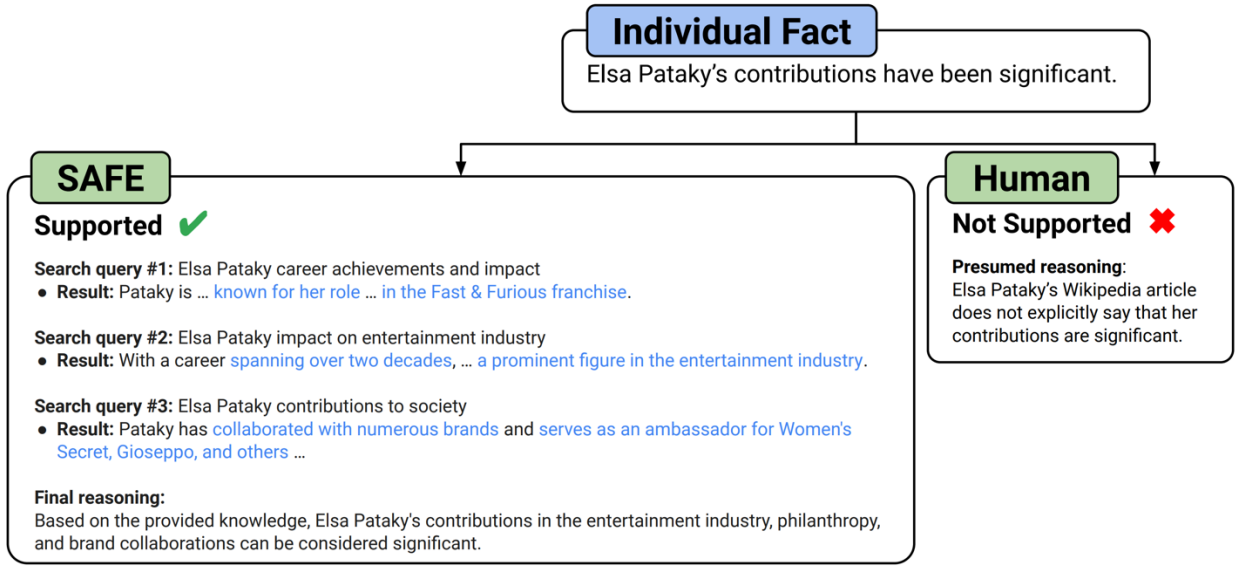
Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems (Arxiv 2024)

	Coverage Score (↑)							Citation Score (↑)							Joint Score (↑)							
Summarizer	Rand	Vect	LongE	KWs	RR3	Orac	Full	Rand	Vect	LongE	KWs	RR3	Orac	Full	Rand	Vect	LongE	Kws	RR3	Orac	Full	$\#W_b$
GPT3.5	36.2	45.8	46.0	48.4	51.9	56.2	–	9.3	15.2	15.0	15.9	16.8	23.0	–	3.6	7.3	7.2	7.9	9.0	13.2	–	28.2
Claude 3 Haiku	49.9	64.9	62.3	63.4	66.6	72.1	62.3	13.4	25.1	25.5	26.5	28.8	35.6	14.1	7.1	17.4	17.2	17.7	20.1	26.8	9.2	31.9
GPT4-turbo	49.4	61.0	56.7	61.2	61.8	67.1	57.9	17.9	28.6	28.1	31.1	31.8	41.4	5.5	9.6	18.7	16.9	20.1	20.6	28.9	3.2	37.9
Command-r	47.0	54.8	53.5	56.0	55.2	60.4	50.3	17.7	34.6	34.9	37.5	40.4	53.8	30.9	8.9	19.6	19.6	21.9	23.6	33.9	16.2	33.1
Gemini-1.5-flash	49.7	58.1	58.9	61.8	62.6	65.1	59.4	17.4	31.9	31.8	34.2	43.6	51.7	32.8	9.2	19.4	20.0	22.0	28.7	34.9	21.0	31.6
Command-r +	44.2	56.4	53.1	56.2	58.9	61.0	44.5	20.4	41.7	41.7	43.1	46.8	60.2	19.9	9.6	24.7	24.0	25.7	29.3	38.3	9.7	25.5
Claude 3 Sonnet	55.8	70.6	69.7	72.1	73.1	77.7	73.6	18.0	34.9	36.6	37.3	41.1	51.7	23.5	11.0	26.1	27.2	28.5	31.4	41.2	18.3	33.5
Claude 3 Opus	56.5	72.4	69.6	72.5	76.5	81.4	76.2	17.7	34.3	35.8	37.3	39.4	50.7	22.3	11.1	26.5	26.7	28.6	31.9	42.5	18.0	29.3
GPT-4o	54.0	67.1	67.8	66.6	70.4	76.6	66.1	21.9	38.4	38.0	38.6	41.3	54.6	16.2	12.6	27.3	27.6	27.3	30.8	43.4	11.4	36.5
Gemini-1.5-pro	53.0	63.5	64.9	63.6	68.4	67.6	70.0	21.9	43.1	44.5	46.6	49.7	64.1	51.0	12.3	28.6	31.0	30.8	36.0	44.6	37.8	30.2
Human Perf.	–	–	–	–	–	74.5	–	–	–	–	–	–	73.9	–	–	–	–	–	–	56.1	–	29.7

Insights: Long-context models can be on par with or even better than the RAG method.

Evaluation • Factuality Capability

Long-form factuality in large language models (Arxiv, 2024)



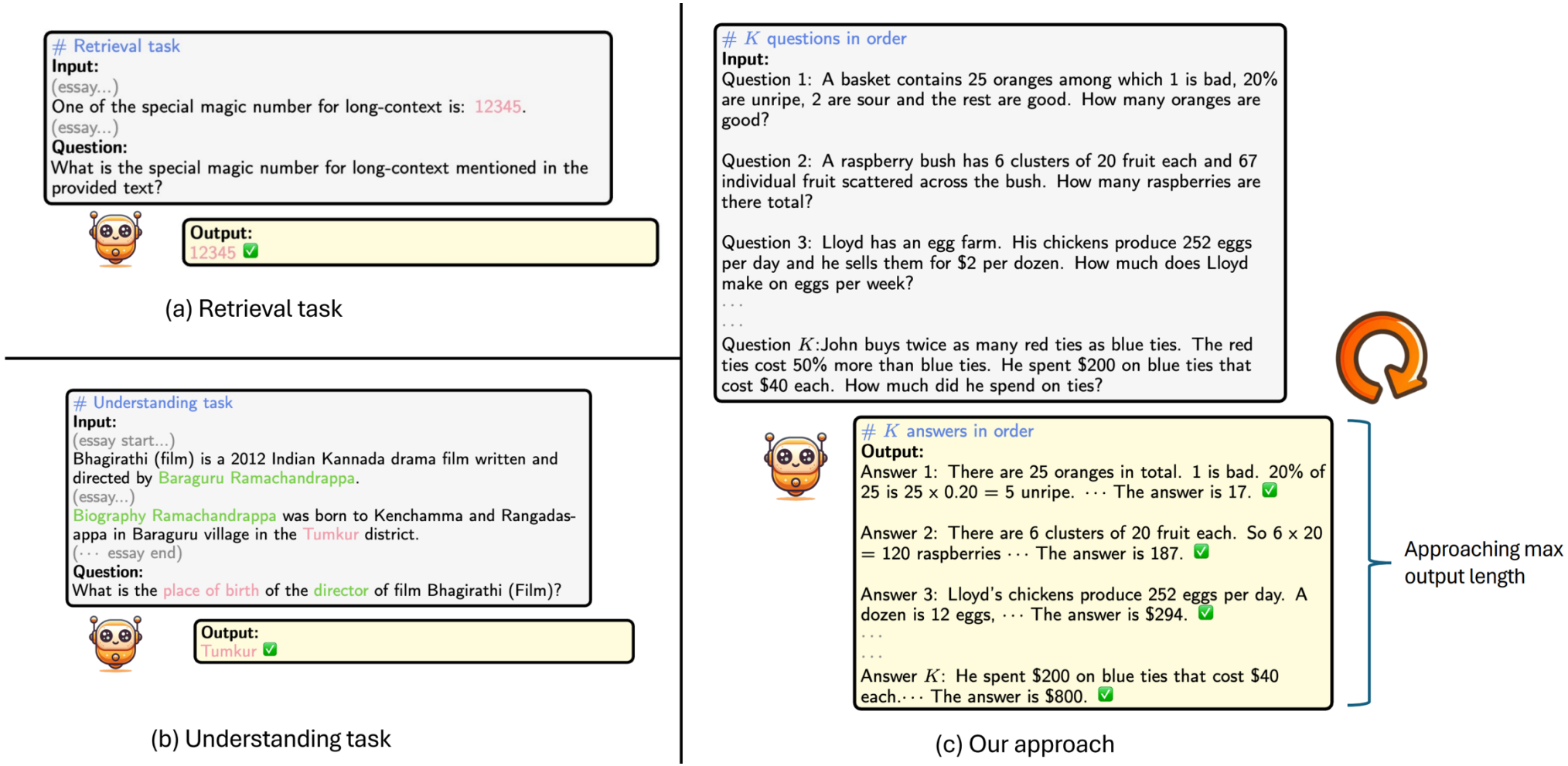
Model	Raw metrics			Aggregated metrics				
	S	NS	I	Prec	R_{64}	R_{178}	$F_1@64$	$F_1@178$
Gemini-Ultra	83.4	13.1	7.6	86.2	98.3	46.9	91.7	60.3
Gemini-Pro	66.6	12.1	5.5	82.0	88.5	37.4	83.7	50.4
GPT-4-Turbo	93.6	8.3	6.1	91.7	99.0	52.6	95.0	66.4
GPT-4	59.1	7.0	2.4	89.4	88.3	33.2	88.0	48.0
GPT-3.5-Turbo	46.9	4.5	1.2	90.8	72.4	26.4	79.6	40.5
Claude-3-Opus	63.8	8.1	2.8	88.5	91.4	35.9	89.3	50.6
Claude-3-Sonnet	65.4	8.4	2.6	88.5	91.4	36.7	89.4	51.4
Claude-3-Haiku	39.8	2.8	1.3	92.8	62.0	22.4	73.5	35.8
Claude-2.1	38.3	6.5	1.9	84.8	59.8	21.5	67.9	33.7
Claude-2.0	38.5	6.3	1.3	85.5	60.0	21.6	68.7	34.0
Claude-Instant	43.9	8.1	1.3	84.4	68.4	24.6	73.8	37.6
PaLM-2-L-IT-RLHF	72.9	8.7	4.3	89.1	94.0	41.0	91.0	55.3
PaLM-2-L-IT	13.2	1.8	0.1	88.8	20.6	7.4	31.1	13.2

After a set number of steps, the model performs reasoning to determine whether the fact is supported by the search results.

Larger models achieve better long-form factuality

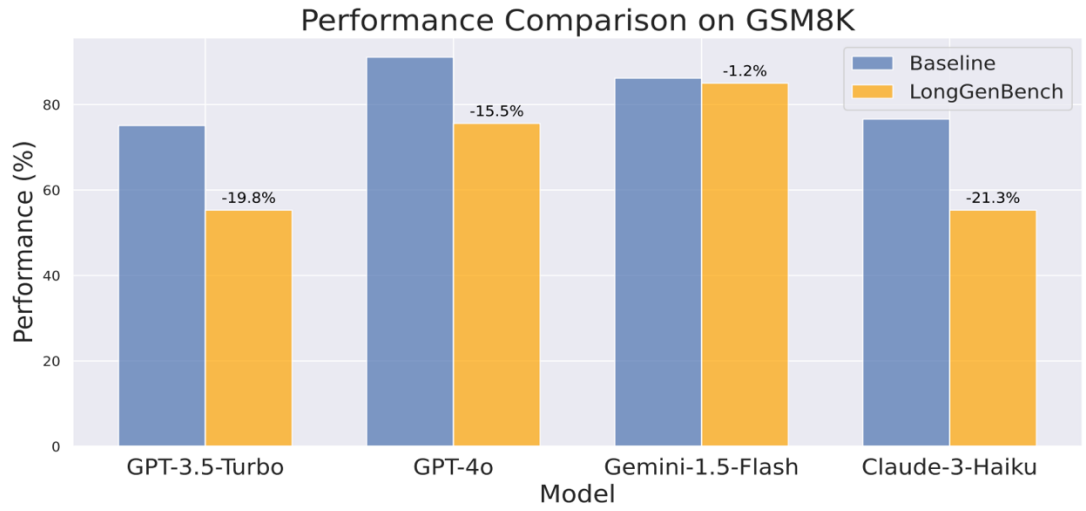
Evaluation • Long-Form Generation Capability

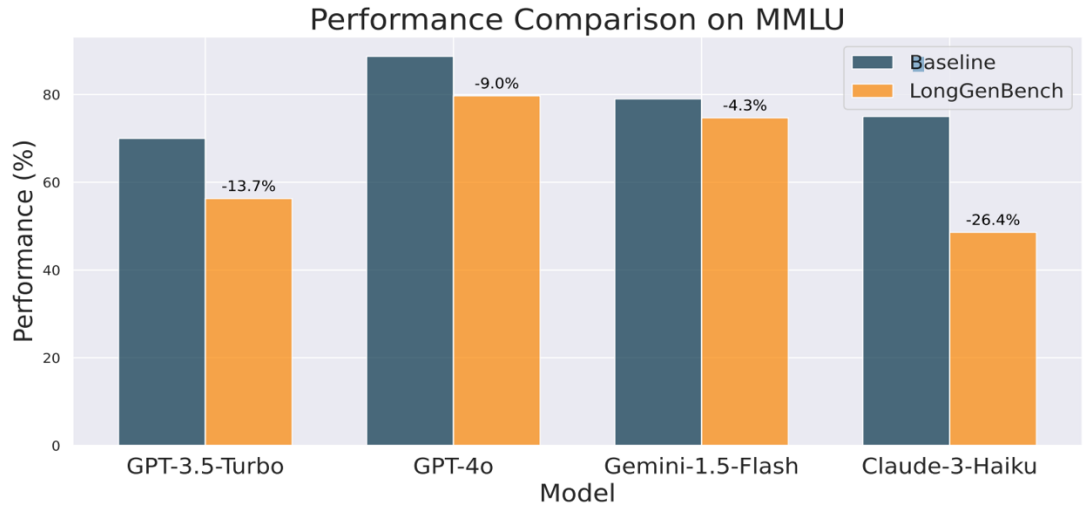
LONGGENBENCH: Long-context Generation Benchmark (EMNLP 2024)



LongGenBench requires LLMs to sequentially understand and respond to each question in a single response.

Evaluation • Long-Form Generation Capability





- Mainstream LLMs exhibit performance degradation when tasked with long-context generation.

Model	GSM8K (%) Baseline↑	LongGenBench↑	Delta Δ
LLaMA-3-8B-Instruct	79.6	32.5	-47.1∇
LLaMA-3-70B-Instruct	93.0	83.2	-9.8∇
Qwen2-7B-Instruct	82.3	63.9	-18.4∇
Qwen2-57B-A14B-Instruct	79.6	71.2	-8.4∇
Qwen2-72B-Instruct	91.1	85.7	-5.4∇
ChatGLM4-9B-Chat	79.6	68.8	-10.8∇
DeepSeek-v2-Chat	92.2	86.5	-5.7∇

(a) Performance on GSM8K dataset

Model	MMLU (%) Baseline↑	LongGenBench↑	Delta Δ
LLaMA-3-8B-Instruct	68.4	50.4	-18.0∇
LLaMA-3-70B-Instruct	82.0	71.2	-10.8∇
Qwen2-7B-Instruct	70.5	59.4	-11.1∇
Qwen2-57B-A14B-Instruct	75.4	66.7	-8.7∇
Qwen2-72B-Instruct	82.3	75.8	-6.5∇
ChatGLM4-9B-Chat	72.4	63.0	-9.4∇
DeepSeek-v2-Chat	77.8	72.0	-5.8∇

(b) Performance on MMLU dataset

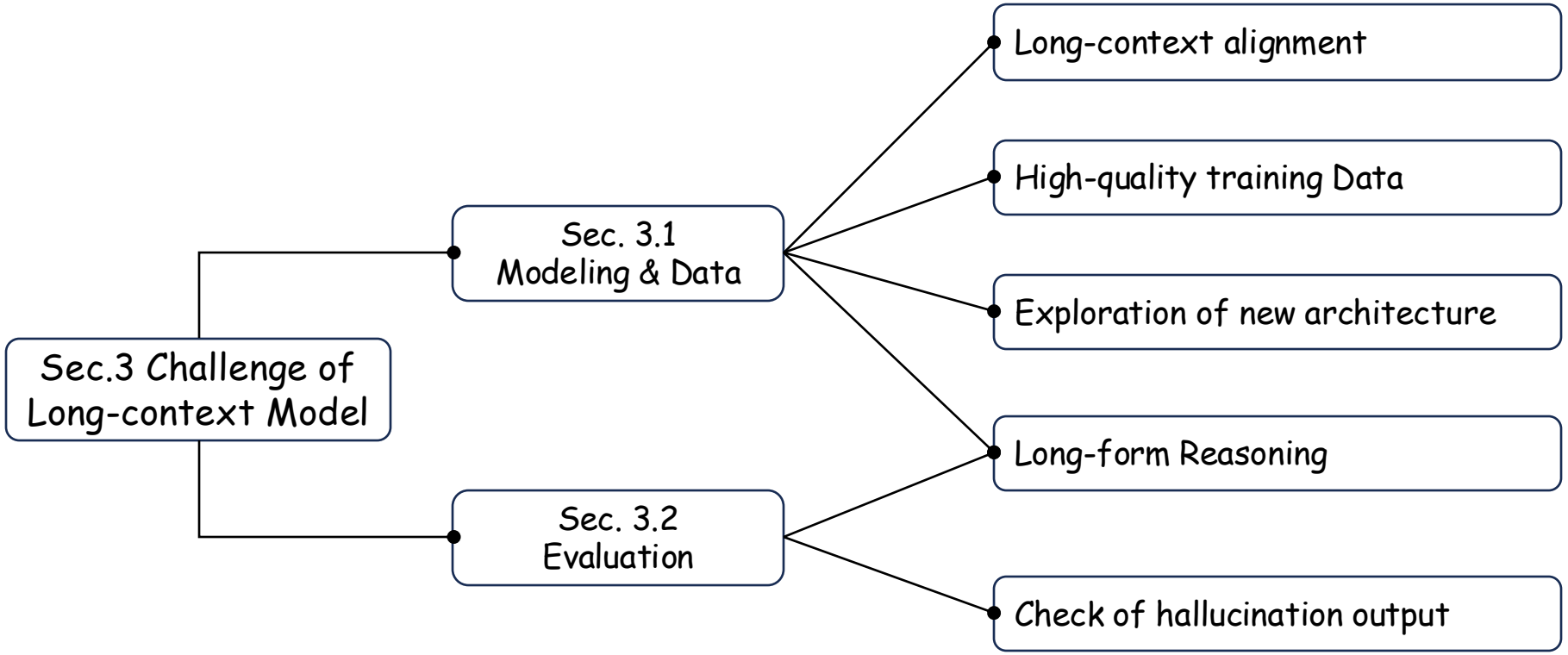
Model	Long Input + Short Output	Long Input + Long Output	Performance Drop
GPT-3.5-Turbo	74.3	55.3	-19.0
Gemini-1.5-Flash	86.1	85.0	-1.1

- Primary challenge in LongGenBench lies in
- the generation of long outputs rather than comprehension of long inputs.

Section 3

Challenge for Current Long-context Model

Challenge of Long-context Model



Long-context Alignment

***Challenge 1**: Lack of effective long-context alignment method.*

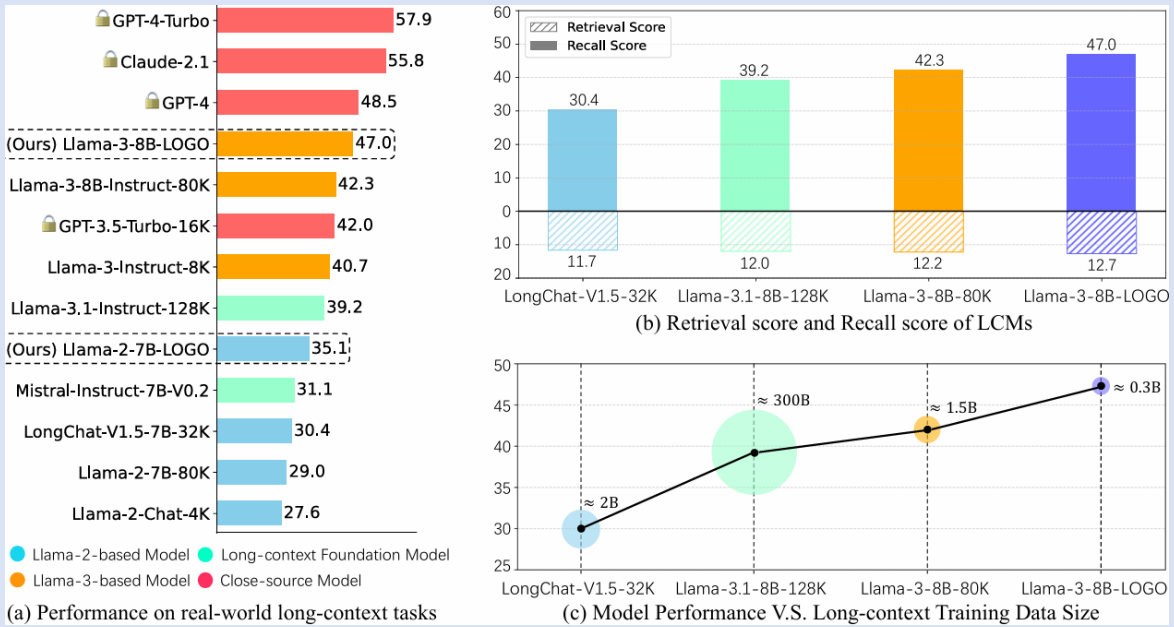
- Lack of high-quality long-context training data;
- Long-context training is costly and challenging, lacking a comprehensive training framework specifically designed for long-context models, such as ring attention.

Efficient Long-context preference optimization

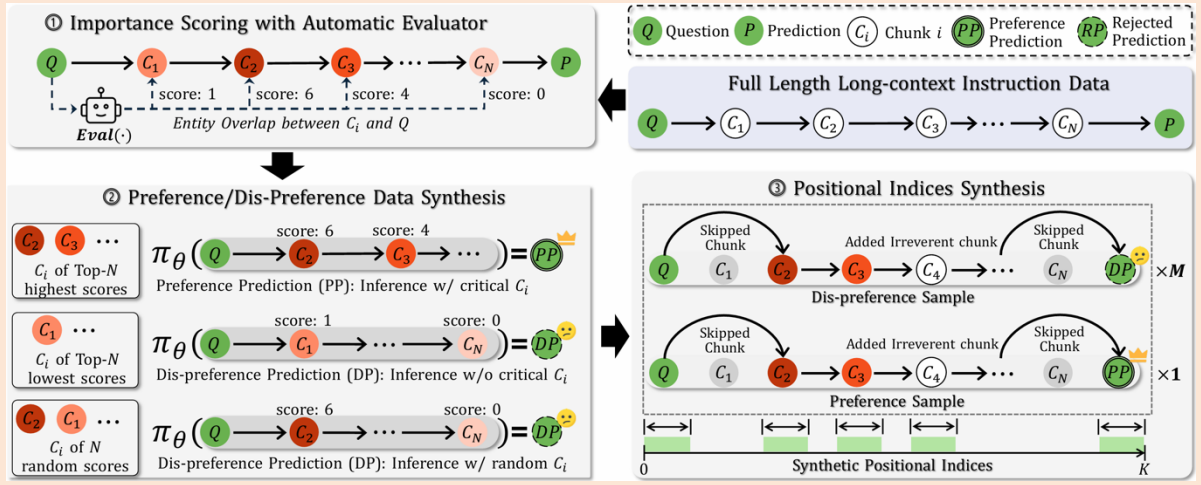
$$\mathcal{L}_{\text{LOGO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l^{(1\cdots M)})}\left[\log\sigma\left(\frac{\beta}{|y_w|}\log\pi_\theta(y_w|x) - \frac{\beta}{M|y_l|}\sum_{j=1}^{M}\log\pi_\theta(y_l^{(j)}|x) - \gamma\right)\right]$$

$$\mathcal{L}^*_{\text{LOGO}}(\pi_\theta) = \mathcal{L}_{\text{LOGO}}(\pi_\theta) + \lambda\mathbb{E}_{(x,y_w)}\log\pi_\theta(y_w|x))$$

LOGO -- Long cOntext aliGnment via efficient preference Optimization (Arxiv 2024)



Data construction



- Constructing long-context preference data from short-context data
  - Leveraging critical segments to build preference data
  - Leveraging irreverent segments to build dis-preference data
- Utilizing Synthetic Positional Indices to scale the context length

Long-context Alignment

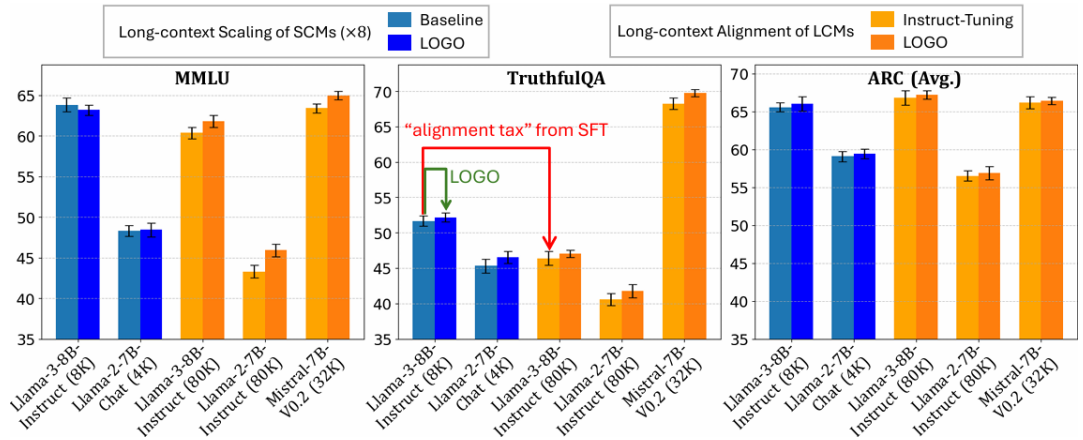
LOGO -- Long cOntext aliGnment via efficient preference Optimization (Arxiv 2024)

https://github.com/ZetangForward/LCM_Stack

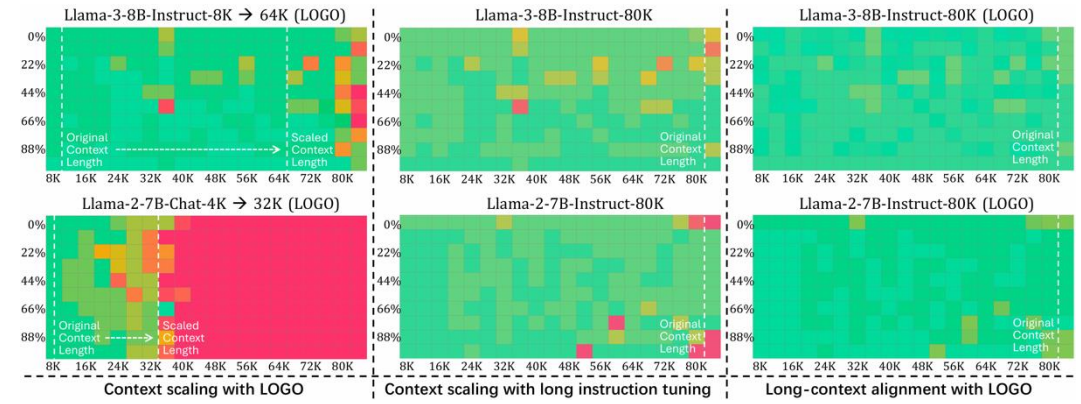
With 0.3B tokens, LOGO significantly improves the model performance on long-context real-world tasks

Models	S-Doc QA	M-Doc QA	Summ	Few-shot	Synthetic	Avg.
GPT-3.5-Turbo-16K	39.8	38.7	26.5	67.1	37.8	42.0
LongChat-v1.5-7B-32k	28.7	20.6	26.7	60.0	15.8	30.4
LLama-3.1-8B-Instruct-128K	23.9	15.8	28.9	69.8	57.5	39.2
Results on SCMs (scaling ×8 context window)						
Llama-3-8B-Instruct-8K	39.3	36.2	24.8	63.5	39.9	40.7
+ YaRN-64K†	38.0	36.6	27.4	61.7	40.9	40.9
+ RandPOS-64K	32.5	30.5	26.5	61.3	33.4	36.8
+ LOGO-64K	**39.8**	**36.7**	**28.8**	**65.4**	**49.0**	**43.9**
Llama-2-7B-Chat-4K	24.9	22.6	24.7	60.0	5.9	27.6
+ LOGO-32K	**26.7**	**23.3**	**26.3**	**63.1**	**11.1**	**30.1**
Results on LCMs (long-context alignment)						
Llama-3-8B-Instruct-80K	43.0	39.8	22.2	64.3	46.3	42.3
+ Instruct Tuning (Full)	38.8	35.0	24.6	65.9	44.5	41.8
+ Instruct Tuning (Partial)	39.3	36.2	26.8	63.5	48.0	42.8
+ LOGO-80K	**44.0**	**41.2**	**28.1**	**68.6**	**53.0**	**47.0**
Llama-2-7B-Instruct-80K	26.9	23.8	21.3	65.0	7.9	29.0
+ LOGO-80K	**33.6**	**28.0**	**29.4**	**65.1**	**24.5**	**36.1**
Mistral-Instruct-7B-V0.2-32K	31.7	30.6	16.7	58.4	17.9	31.1
+ LOGO-32K	**38.3**	**37.6**	**26.1**	**67.0**	**31.5**	**40.1**

LOGO can maintain the model performance on short-context tasks, while SFT methods lead to the model forgetting intrinsic knowledge



LOGO can also scale the model context window size



Exploration of New Architecture

Hugging Face tiiuae/falcon-mamba-7b

model name	IFEval	BBH	MATH LvL5	GPQA	MUSR	MMLU-PRO	Average
Pure SSM models							
FalconMamba-7B	33.36	19.88	3.63	8.05	10.86	14.47	**15.04**
TRI-ML/mamba-7b-rw*	22.46	6.71	0.45	1.12	5.51	1.69	6.25
Hybrid SSM-attention models							
recurrentgemma-9b	30.76	14.80	4.83	4.70	6.60	17.88	13.20
Zyphra/Zamba-7B-v1*	24.06	21.12	3.32	3.03	7.74	16.02	12.55
Transformer models							
Falcon2-11B	32.61	21.94	2.34	2.80	7.53	15.44	13.78
Meta-Llama-3-8B	14.55	24.50	3.25	7.38	6.24	24.55	13.41
Meta-Llama-3.1-8B	12.70	25.29	4.61	6.15	8.98	24.95	13.78
Mistral-7B-v0.1	23.86	22.02	2.49	5.59	10.68	22.36	14.50
Mistral-Nemo-Base-2407 (12B)	16.83	29.37	4.98	5.82	6.52	27.46	15.08
gemma-7B	26.59	21.12	6.42	4.92	10.98	21.64	**15.28**
RWKV models							
RWKV-v6-Finch-7B*	27.65	9.04	1.11	2.81	2.25	5.85	8.12
RWKV-v6-Finch-14B*	29.81	12.89	1.13	5.01	3.16	11.3	10.55

Challenge 2: Insufficient exploration of new model architectures

- Training difficulties, including frequent numerical overflow (nan loss) and loss explosion problems;
- Significant capability loss due to computational trade-off

SoTA RNN-based model in benchmark

Transformer-based models perform stable

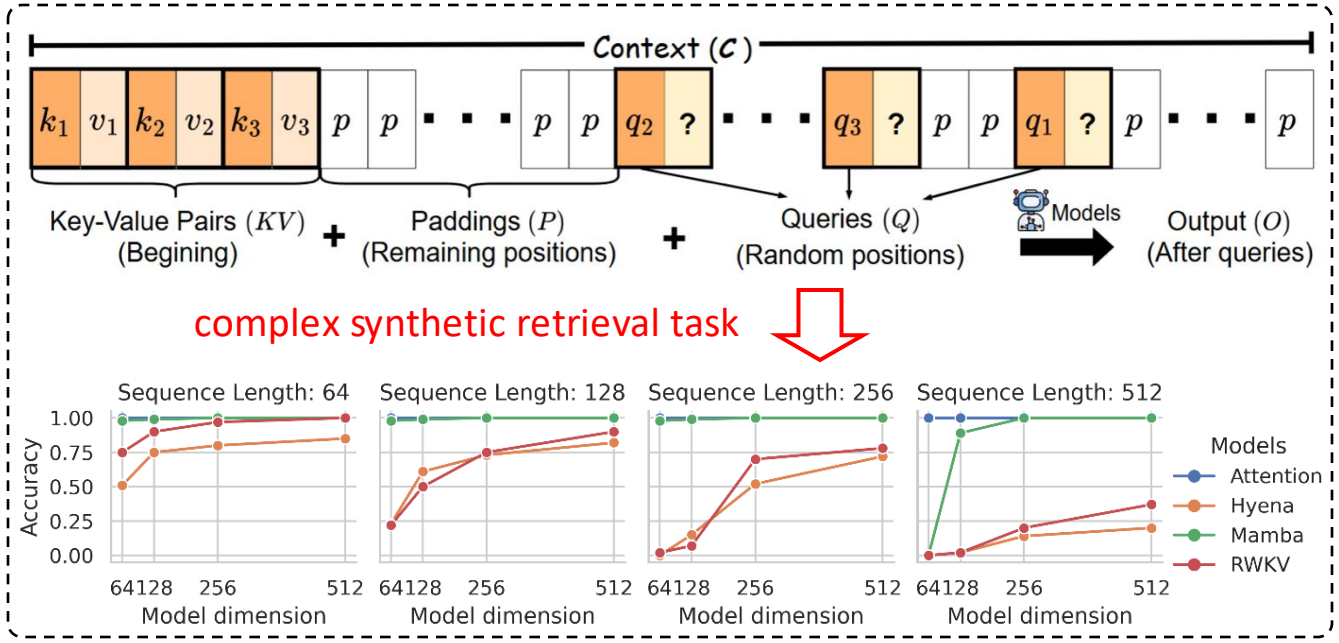
SoTA Transformer-based model in benchmark

Exploration of New Architecture

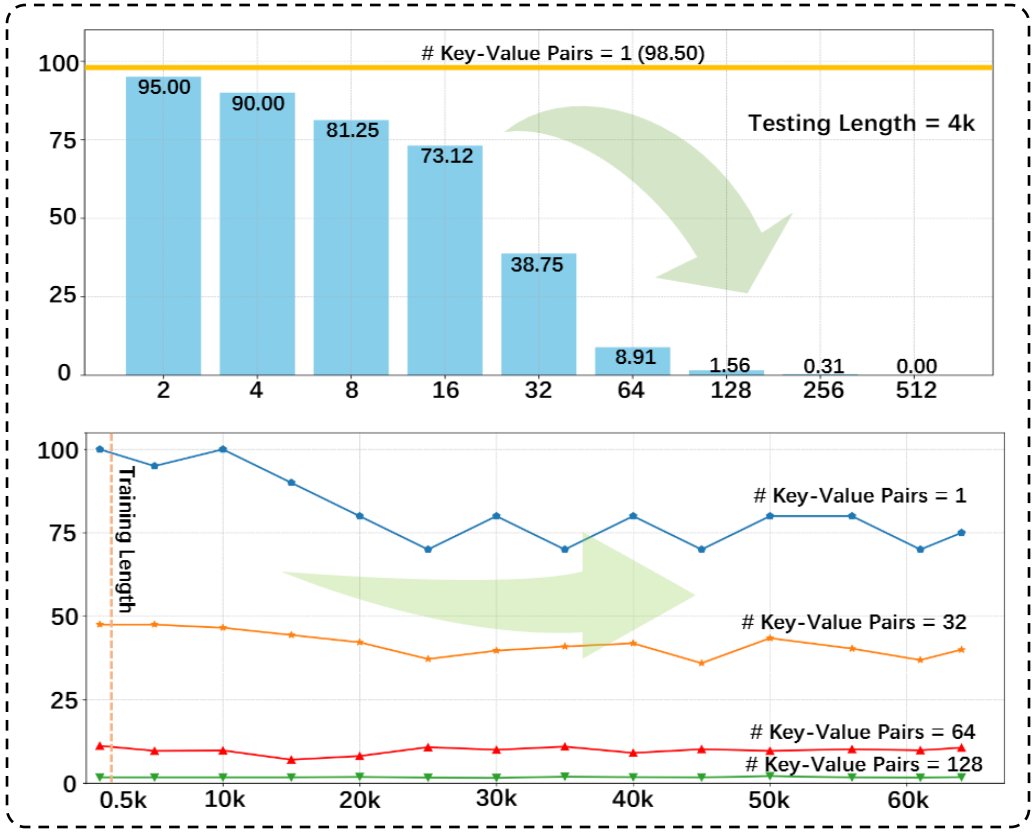
Revealing and Mitigating the Local Pattern Shortcuts of Mamba (Arxiv 2024)

https://github.com/ZetangForward/Global_Mamba

Mamba outperforms other RNN-based models at the same model dimension with a fixed recurrent state size.



Mamba can generalize on Sequence Length, but it fails to generalize on Information Density.

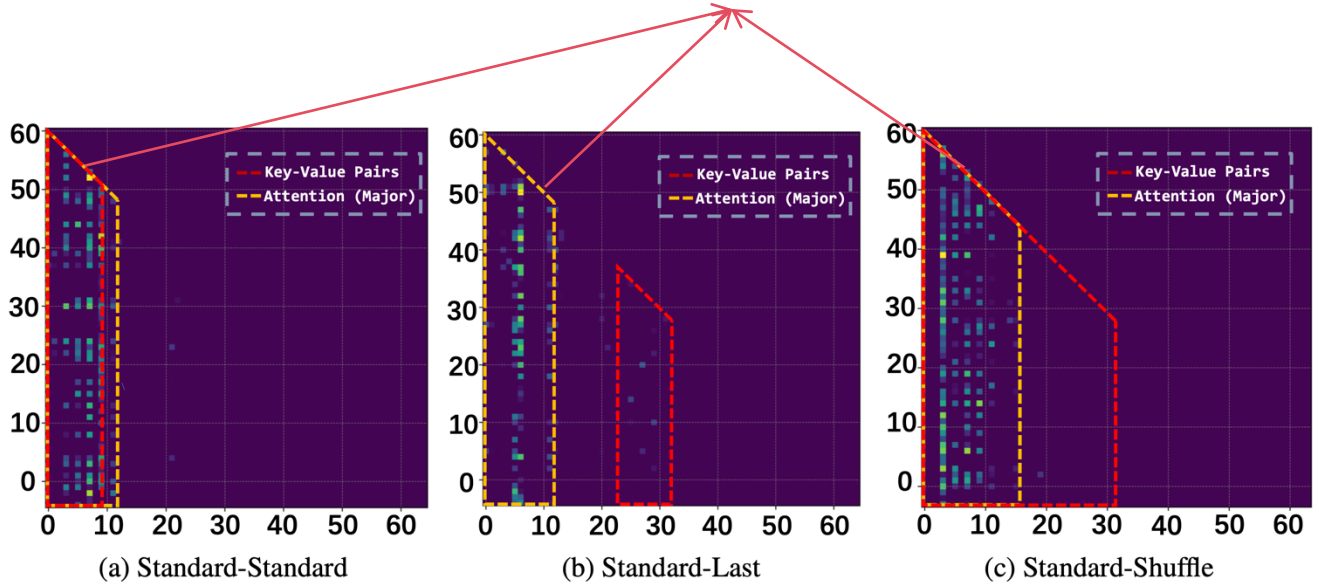


Exploration of New Architecture

Revealing and Mitigating the Local Pattern Shortcuts of Mamba (Arxiv 2024)

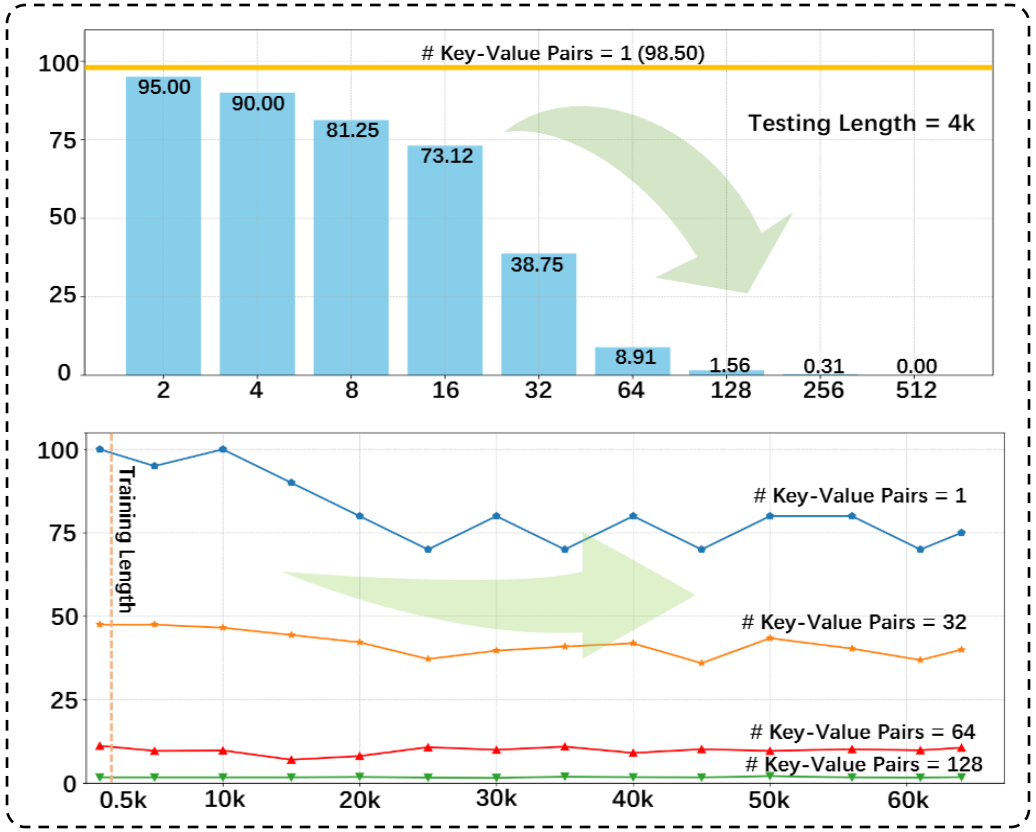
https://github.com/ZetangForward/Global_Mamba

Attention map of Mamba on various task settings, where attention is concentrated at the beginning (shortcuts).



(a) Standard-Standard (b) Standard-Last (c) Standard-Shuffle

Mamba can generalize on Sequence Length, but it fails to generalize on Information Density.



Exploration of New Architecture

Revealing and Mitigating the Local Pattern Shortcuts of Mamba (Arxiv 2024)

https://github.com/ZetangForward/Global_Mamba

Simply add a global (long) gate to the transformation matrix can break the shortcuts !

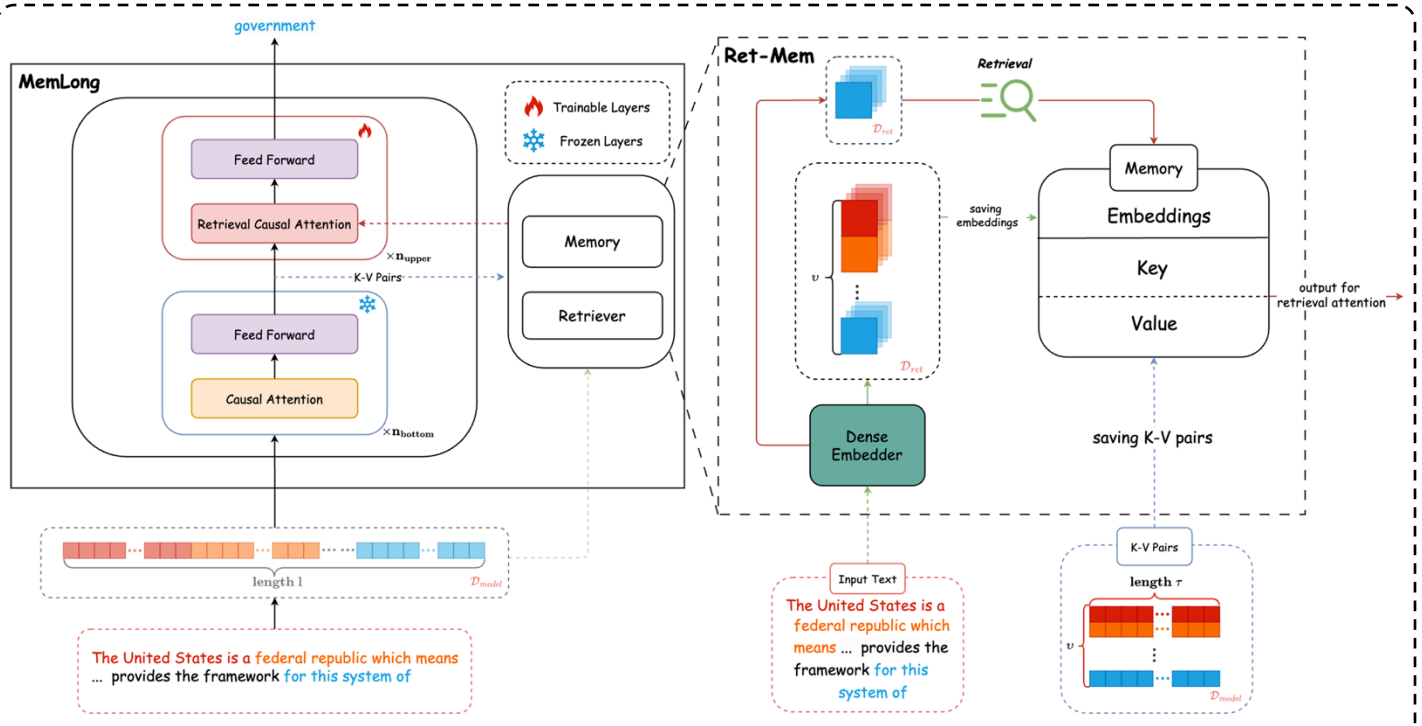
$$\Delta_t = \mathbf{W}_2 \cdot \sigma\left(\mathbf{W}_1 \cdot \mathrm{Conv}_{\mathrm{short}}(\mathbf{X}_t)\right) \odot \sigma\left(\mathrm{Conv}_{\mathrm{long}}(\mathbf{X}_t)\right)$$

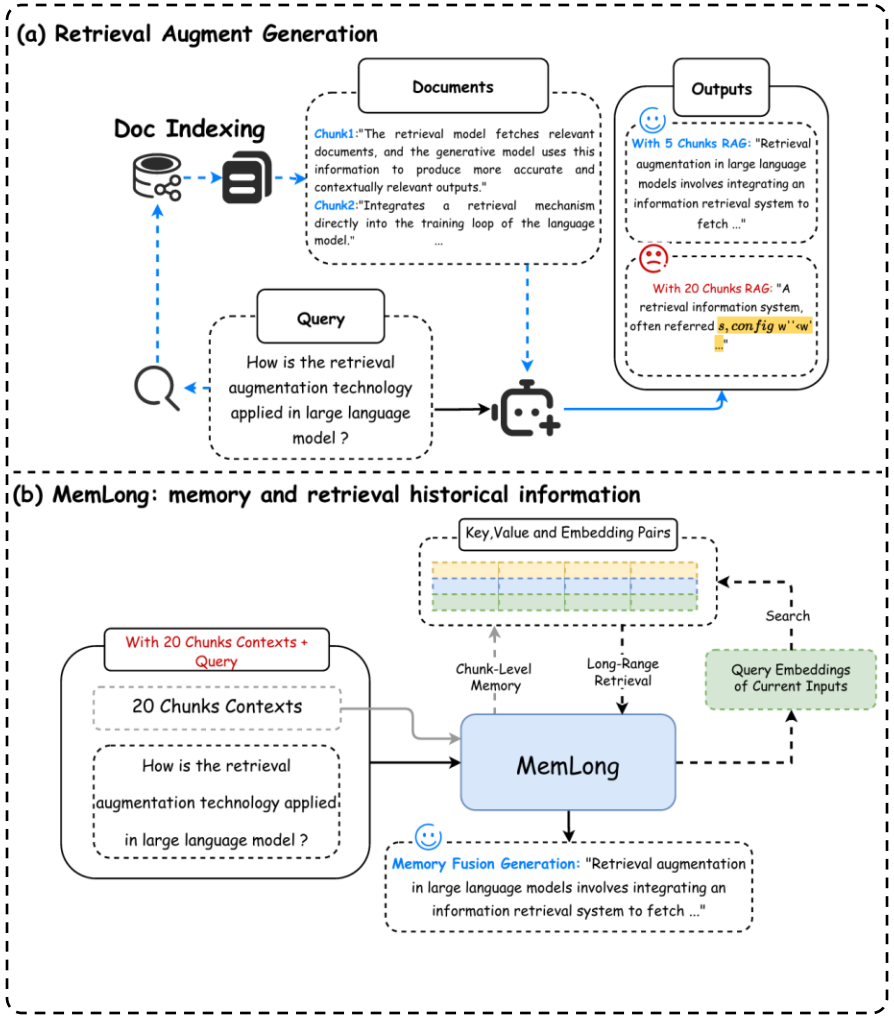
Models	Scale	Shuffle	Std-Last	Std-Shuffle	K2V2	K2V2-Robustness	K4V8-Shuffle
Pythia (Biderman et al., 2023)	*133m*	99.82	93.75	94.31	99.99	99.99	99.99
Hyena (Poli et al., 2023)	*153m*	✗	✗	✗	77.62	65.92	22.51
RWKV (Peng et al., 2023)	*153m*	✗	✗	✗	85.99	72.62	6.57
Mamba (Gu and Dao, 2023)	*129m*	80.98	15.44	22.37	99.98	66.01	✗
w/ 2×*State Size*	*130m*	88.57	40.22	31.88	99.84	78.90	✗
w/ 4×*State Size*	*134m*	96.92	35.89	32.88	99.84	57.11	✗
w/ *Global Selection*	*133m*	90.45	41.97	35.73	99.06	81.46	80.54

Exploration of New Architecture

MemLong: Memory-Augmented Retrieval for Long Text Modeling (Arxiv 2024)

https://github.com/Bui1dMySea/MemLong





- Decouple the computation and storage of the Key-Value (KV) cache.
  - Memory serves for storing the KV cache
  - CPU serves for retrieving requires memory (KV cache)
  - GPU serves for computation
- Maintain the memory using a First-In-First-Out (FIFO) queue.

Exploration of New Architecture

MemLong: Memory-Augmented Retrieval for Long Text Modeling (Arxiv 2024)

https://github.com/Bui1dMySea/MemLong

Model	PG19				Proof-pile				BookCorpus				Wikitext-103			
	1k	2k	4k	16k	1k	2k	4k	16k	1k	2k	4k	16k	1k	2k	4k	16k
7B Model																
LLaMA-2-7B	10.82	10.06	8.92	-	3.24	3.40	2.72	-	8.73	7.91	6.99	-	10.82	6.49	5.66	-
LongLoRA-7B-32k	9.76	9.71	10.37	7.62	3.68	3.35	3.23	2.60	14.99	12.66	11.66	6.93	7.99	7.83	8.39	5.47
YARN-128k-7b	7.22	7.47	7.17	-	3.03	3.29	2.98	-	7.02	7.54	7.06	-	5.71	6.11	5.71	-
3B Model																
OpenLLaMA-3B	11.60	9.77	$> 10^3$	-	**2.96**	**2.70**	$> 10^3$	-	**8.97**	8.77	$> 10^3$	-	10.57	8.08	$> 10^3$	-
LongLLaMA-3B*	10.59	10.02	$> 10^3$	-	3.55	3.15	$> 10^3$	-	10.70	9.83	$> 10^3$	-	8.88	8.07	$> 10^3$	-
LongLLaMA-3B†	10.59	10.25	9.87	-	3.55	3.22	**2.94**	-	10.14	9.62	9.57	-	10.69	8.33	7.84	-
Phi3-128k	11.31	9.90	**9.66**	- / *9.65*	4.25	3.11	2.77	- / *3.08*	11.01	**9.22**	**8.98**	- / *9.27*	**7.54**	**7.22**	7.01	- / *7.20*
MemLong-3B*	10.66	10.09	$> 10^3$	-	3.58	3.18	$> 10^3$	-	10.37	9.55	$> 10^3$	-	8.72	7.93	$> 10^3$	-
w/ 4K Memory	10.54	9.95	9.89	**9.64**	3.53	3.16	3.15	2.99	10.18	9.50	9.57	9.61	8.53	7.92	7.87	7.99
w/ 32K Memory	**10.53**	**9.85**	**9.83**	9.73	3.51	3.15	3.11	**2.99**	9.64	9.56	**9.51**	**9.54**	8.02	7.58	**6.89**	**7.09**

Plug-and-play method

Performance on Language Modeling tasks, where all the experiments are conduct on 3090 GPU (24GB memory size)

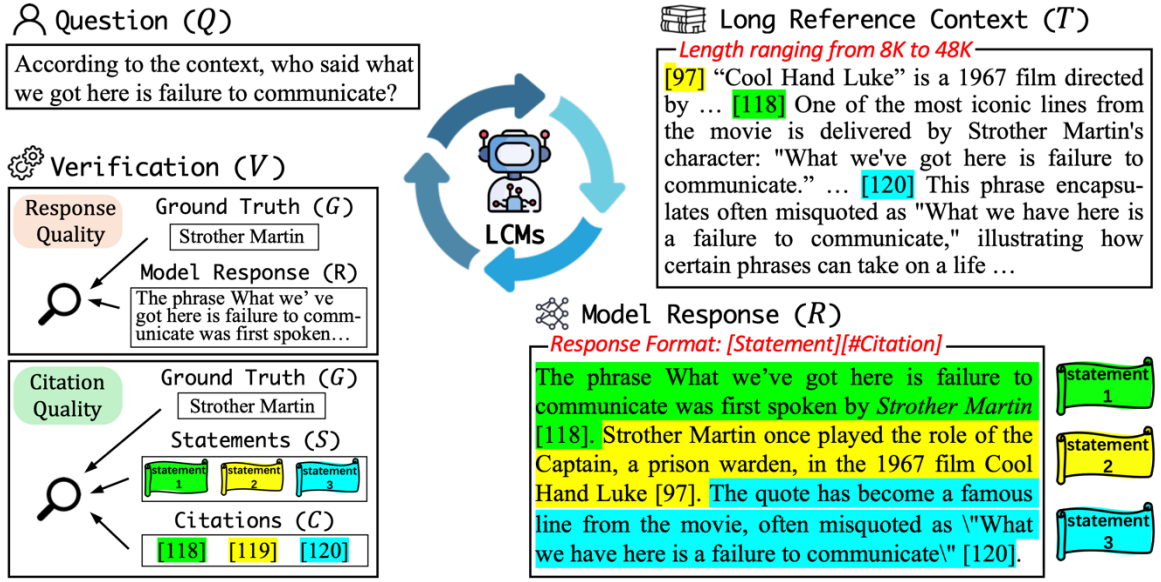
Evaluation of Long-context Model

L-CiteEval: Do Long-Context Models Truly Leverage Context for Responding? (Arxiv 2024)

https://github.com/ZetangForward/L-CITEEVAL

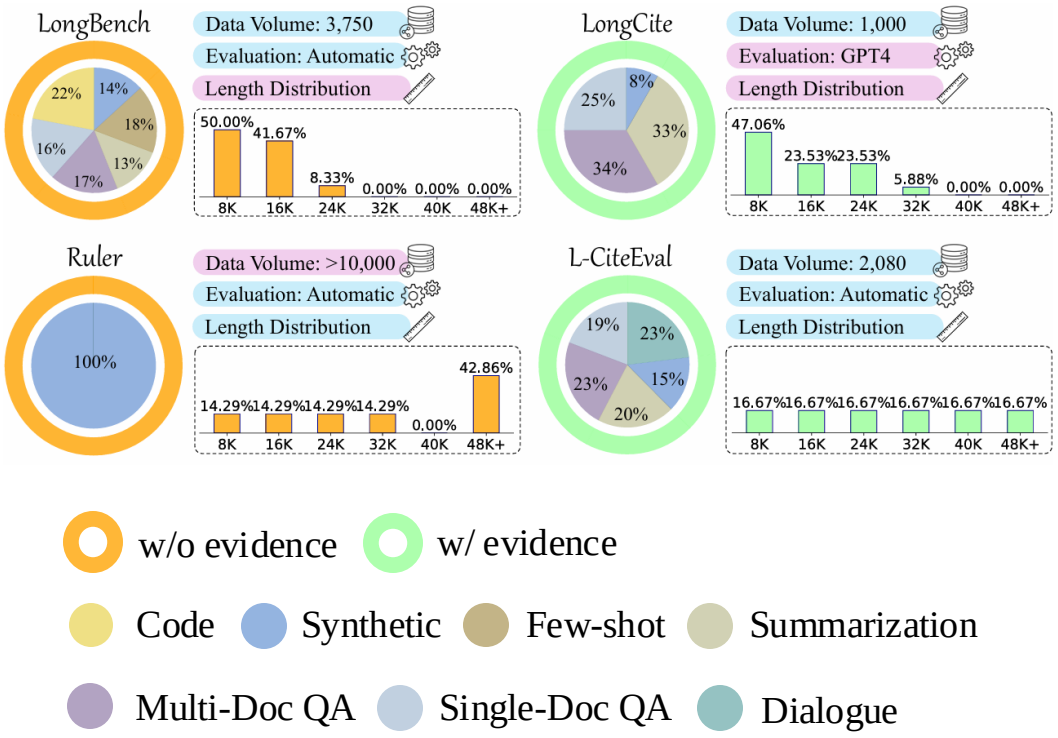
Challenge 3: Lack of interpretability and faithfulness for long-context model evaluation

In addition to focusing on the performance of the Long-context model, it is also essential to pay attention to whether the long-text model responds based on the context (faithfulness).



Task format and pipeline of L-CiteEval benchmark

Comparison among different long-context benchmarks



Evaluation of Long-context Model

L-CiteEval: Do Long-Context Models Truly Leverage Context for Responding? (Arxiv 2024)

https://github.com/ZetangForward/L-CITEEVAL

Models	Single-Doc QA				Dialogue Understanding				Needle in a Haystack			
	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N
Closed-source LCMs												
GPT-4o	32.05	**38.12**	33.48	2.02	53.90	**64.25**	**56.76**	2.17	**76.25**	**76.67**	**76.39**	1.12
Claude-3.5-sonnet	**38.70**	37.79	**37.43**	3.54	**54.45**	50.48	51.45	2.83	65.00	68.33	65.97	1.04
o1-mini	29.83	35.33	31.66	3.38	45.54	50.74	47.21	2.63	25.42	28.33	26.25	1.58
Open-source LCMs												
Qwen2.5-3b-Ins	7.13	5.83	6.00	1.75	9.53	9.71	8.41	2.33	12.08	12.50	12.22	1.12
Phi-3.5-mini-Ins	21.06	20.46	19.14	2.86	20.39	24.27	20.57	2.27	11.11	12.50	11.53	1.20
Llama-3.1-8B-Ins	22.68	24.73	22.64	2.59	51.86	**57.58**	53.50	2.08	34.31	35.83	34.72	0.99
Glm-4-9B-chat	**29.00**	**28.66**	**28.05**	2.21	**54.54**	55.62	**53.58**	1.78	46.53	**50.83**	**47.78**	1.23
Mistral-Nemo-Ins	4.34	3.68	3.76	0.68	23.91	24.33	23.50	1.35	11.11	12.50	11.53	1.18
Qwen2-57B-A14B-Ins	4.90	3.43	3.82	1.27	22.63	22.54	21.61	1.80	15.28	15.83	15.42	1.17
Llama-3.1-70B-Ins	25.89	26.89	26.11	1.23	51.71	56.20	53.19	1.76	**46.67**	46.67	46.67	0.82
ChatQA-2-70B	21.75	22.54	21.92	1.12	47.67	51.25	48.77	1.29	38.33	38.33	38.33	0.95

Models	Multi-Doc QA				Summarization				Counting Stars			
	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N
Closed-source LCMs												
GPT-4o	57.48	**58.50**	56.10	1.71	34.37	54.28	41.60	22.86	**83.37**	**81.18**	**81.71**	4.54
Claude-3.5-sonnet	**66.85**	55.62	**58.58**	2.44	**36.70**	**55.03**	**43.45**	17.70	73.01	75.83	73.15	4.81
o1-mini	49.95	49.60	48.58	1.78	20.23	33.61	24.83	19.58	34.06	46.46	38.45	6.73
Open-source LCMs												
Qwen2.5-3b-Ins	13.17	8.04	9.37	1.96	7.72	12.15	9.09	9.52	3.82	1.48	2.01	1.66
Phi-3.5-mini-Ins	11.89	10.25	10.53	1.71	10.90	10.94	9.60	8.23	4.19	3.67	4.09	3.48
Llama-3.1-8B-Ins	43.41	42.15	41.64	1.62	19.57	23.03	20.83	18.31	16.87	18.26	19.18	4.19
Glm-4-9B-chat	47.91	44.75	45.09	1.64	**29.16**	**37.29**	**31.92**	11.38	18.15	15.69	16.21	4.52
Mistral-Nemo-Ins	17.61	15.45	15.85	0.70	11.21	14.85	12.40	5.45	3.09	2.92	3.26	2.32
Qwen2-57B-A14B-Ins	17.30	12.07	13.61	1.06	4.01	3.37	3.19	3.81	4.37	4.37	4.24	4.24
Llama-3.1-70B-Ins	**49.64**	**54.02**	**50.74**	1.42	25.50	31.99	27.91	11.78	**66.85**	**61.74**	**63.73**	4.37
ChatQA-2-70B	47.20	49.51	47.92	1.10	19.57	23.60	20.89	11.81	14.02	11.22	13.22	3.49

Citation quality of LCMs on L-CiteEval

Models	Single-Doc QA		Multi-Doc QA		Summ.	Dialogue		Synthetic	
	Prec.	Rec.	Prec.	Rec.	Rouge-L	Prec.	Rec.	Rouge-1†	Acc‡
Closed-source LCMs									
GPT-4o	**11.78**	70.37	**10.34**	**87.38**	20.15	**9.81**	**65.35**	**89.24**	**91.88**
Claude-3.5-sonnet	5.96	**71.96**	4.30	80.77	**22.06**	3.71	57.80	88.33	69.65
o1-mini	10.30	66.44	7.36	64.25	19.22	7.02	54.27	54.98	57.29
Open-source LCMs									
Qwen2.5-3b-Ins	8.91	60.28	3.82	52.41	22.39	4.58	40.77	84.49	26.81
Phi-3.5-mini-Ins	8.62	62.34	7.82	64.54	19.48	11.39	52.77	73.83	61.32
Llama-3.1-8B-Ins	10.11	**68.13**	7.66	68.84	20.90	11.07	58.84	85.11	33.75
Glm-4-9B-chat	11.22	67.25	7.88	**77.97**	21.42	7.69	51.25	90.81	58.82
Mistral-Nemo-Ins	10.53	59.71	8.78	67.70	20.83	9.27	49.26	87.88	18.06
Qwen2-57B-A14B-Ins	12.93	61.71	15.25	57.53	**22.95**	14.32	52.23	**91.30**	63.61
Llama-3.1-70B-Ins	15.23	67.08	12.50	76.40	22.29	19.62	**62.91**	88.18	**89.03**
ChatQA-2-70B	**43.25**	61.20	**34.95**	55.64	22.06	**26.57**	58.34	70.14	78.68

Generation quality of LCMs on L-CiteEval

Performance of Closed-Source Models:

- GPT-4o and Claude-3.5-sonnet excel in generation quality and citation quality, particularly in citation accuracy and recall.

Performance of Open-Source Models:

- Open-source models are comparable to closed-source models in generation quality but lag behind in citation quality.
- Larger open-source models, such as Llama-3.1-70B-Instruct, perform nearly as well as closed-source models in some tasks.

Conclusion

Long Context Modeling in LLM Era — Advances and Challenges

Introduction

01

- Long-context Model
  - Progress
  - Application
  - Definition

Long-context Training & Evaluation

02

- Modeling
  - Positional Embedding
  - Long-context Alignment
    - Efficient Training
    - New Architecture
    - Infrastructure
- Data
  - Resources
  - Construction
    - Splice / Up-sampling
    - Positional Synthesis
    - Model Generation
  - Usage
    - SFT
    - RL
- Evaluation
  - General Capability
  - Specifical Capability

Challenges

03

- Modeling and Data
  - Long-context training data
  - Efficient training method
  - New architectures
- Evaluation
  - Faithfulness

Q & A