Long Context LLMs: Challenges & Solutions

Large Language Models: Introduction and Recent Advances

ELL881 · AIL821

Gaurav Pandey
Research Scientist, IBM Research

Agenda

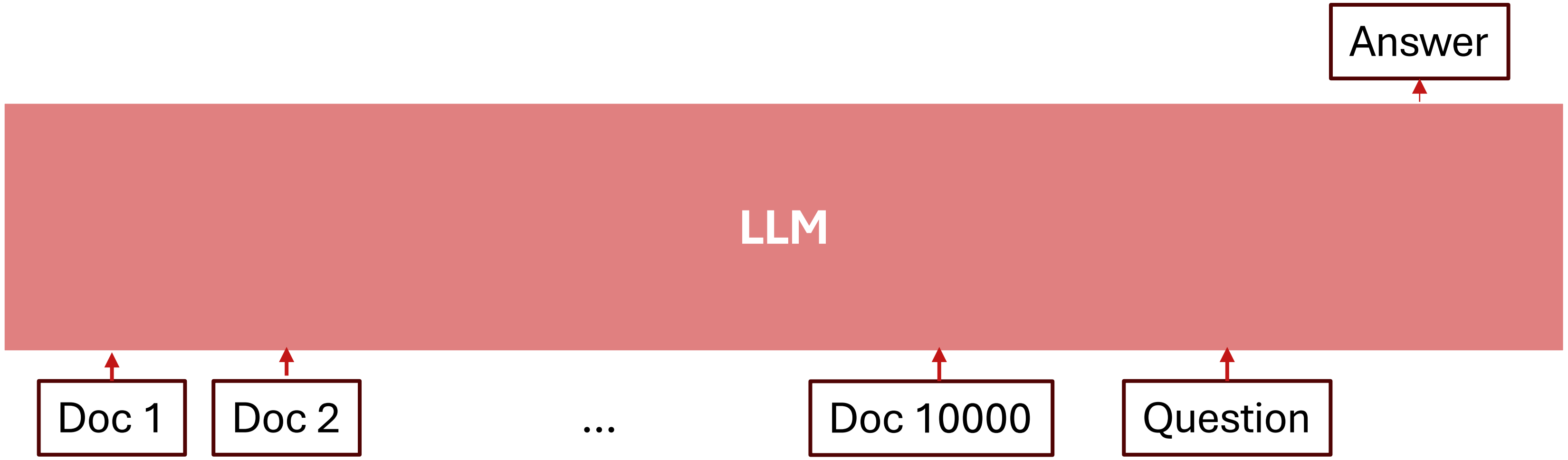
- Long Contexts & Challenges
- Key Papers and Their Contributions
  - LongNet
  - ALiBi
  - Positional Interpolation
  - Lost in the Middle
- Discussion and Future Directions

Agenda

- **Long Contexts & Challenges**
- Key Papers and Their Contributions
  - LongNet
  - ALiBi
  - Positional Interpolation
  - Lost in the Middle
- Discussion and Future Directions

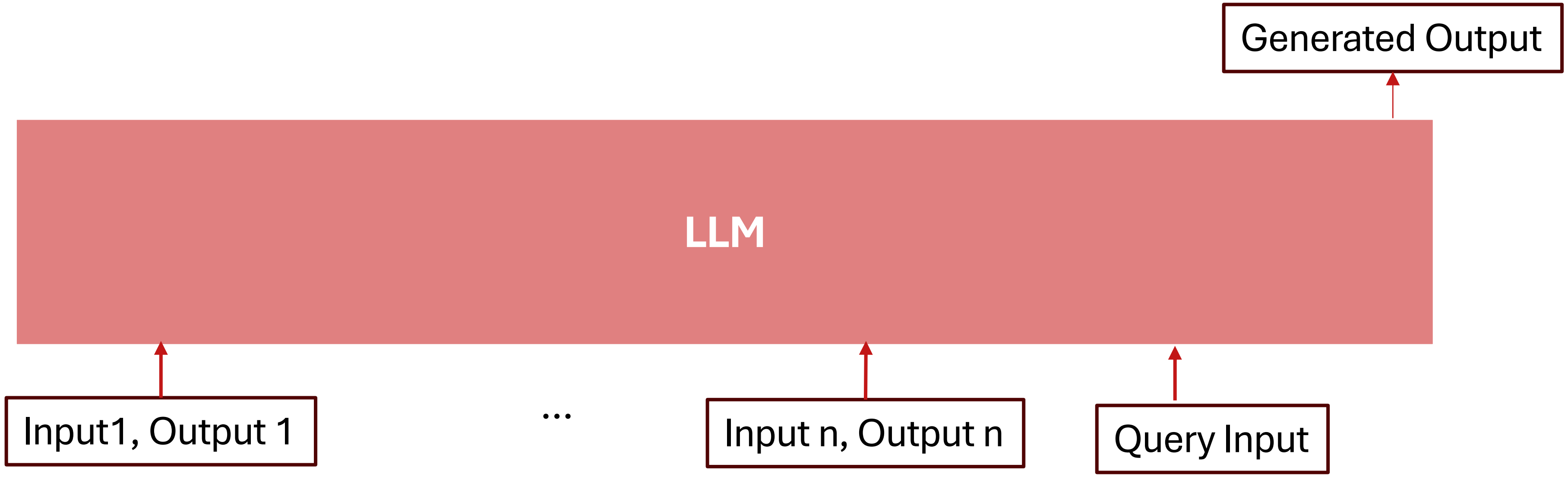
Why is Long Context desirable?

- Allows summarization & generation of entire books while maintaining continuity.
- Allows inference over long videos & entire codebase.
- Retrieval Augmented generation
  - Doesn’t require the retriever to be accurate



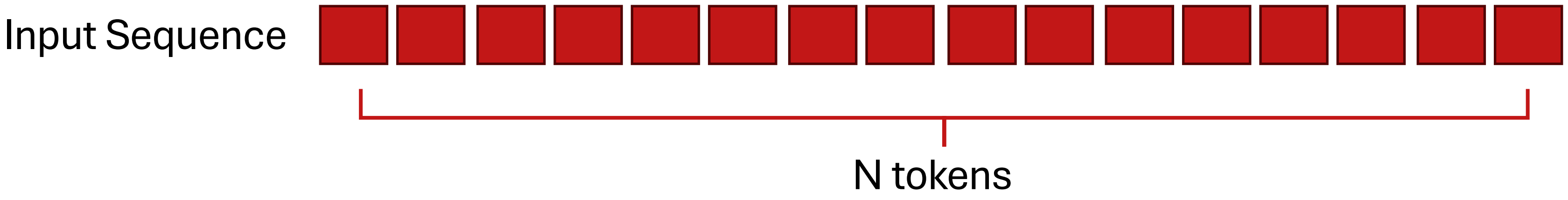
Why is Long Context desirable?

- Can replace task-specific finetuning with in-context learning.



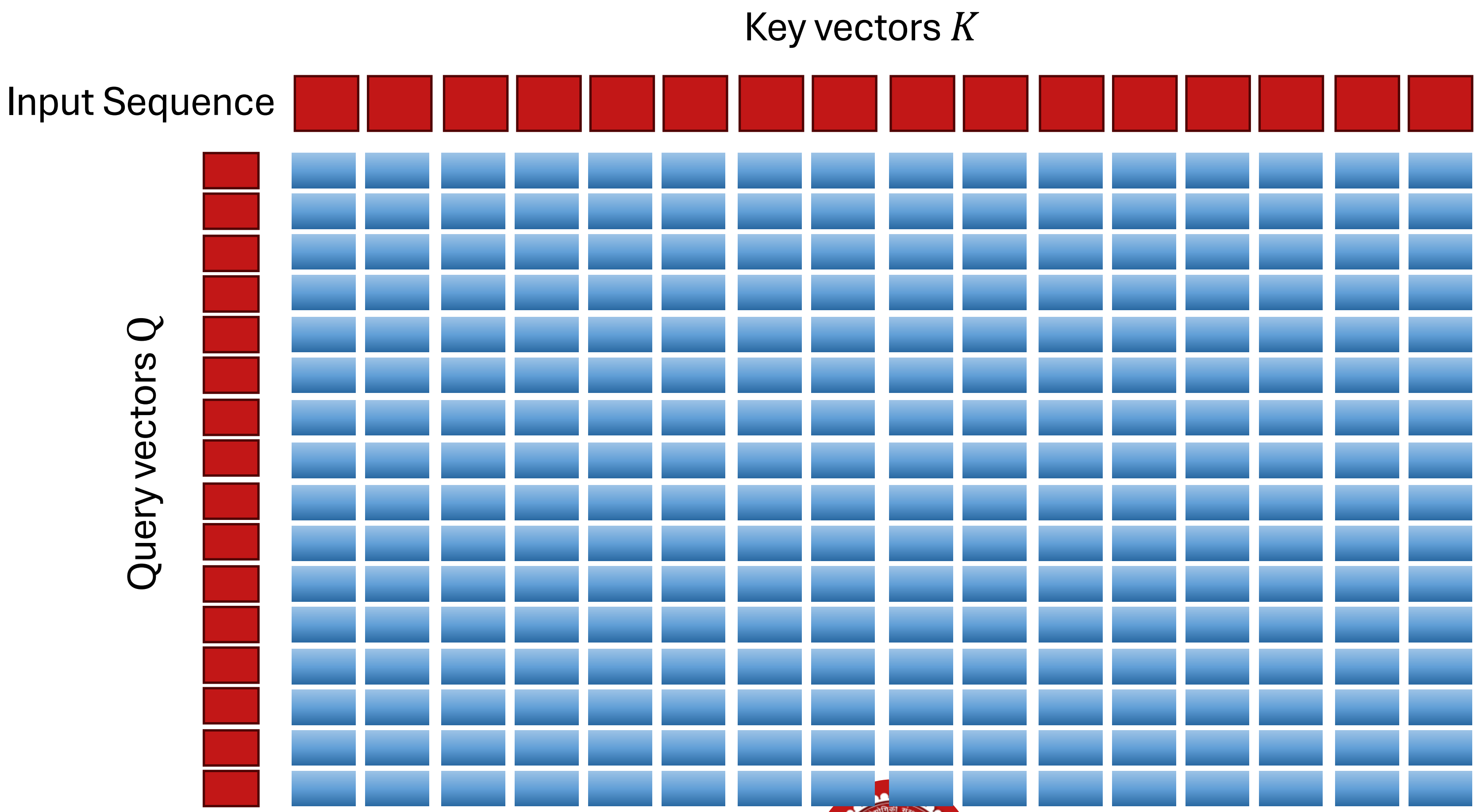
Challenges

- Computational complexity of self-attention



Challenges

- Computational complexity of self-attention



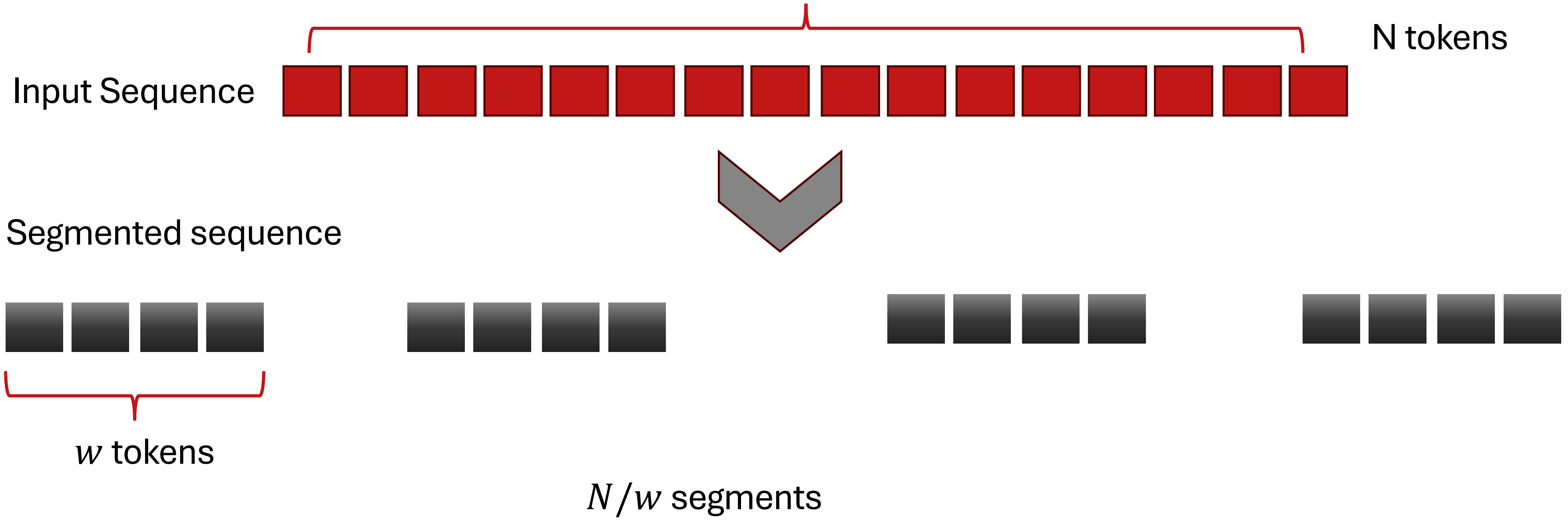
$O(N^2)$ computations

Agenda

- Long Contexts & Challenges
- Key Papers and Their Contributions
  - **LongNet**
  - ALiBi
  - Positional Interpolation
  - Lost in the Middle
- Discussion and Future Directions

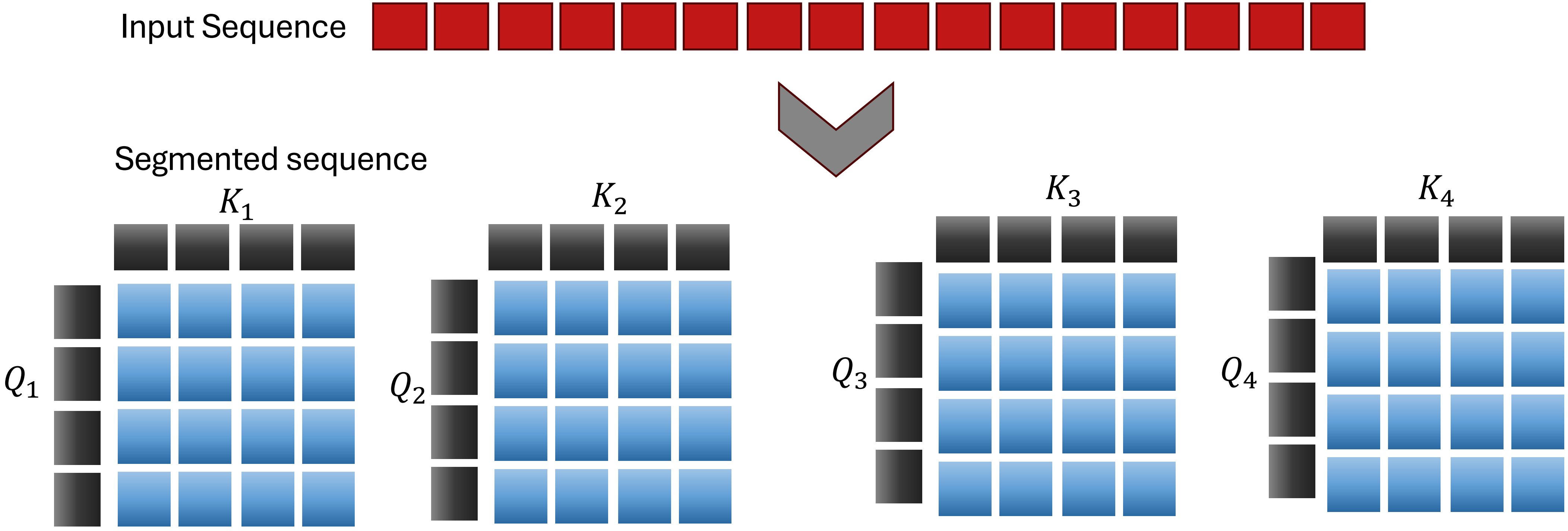
LongNet: Scaling Transformers to 1,000,000,000 Tokens

LongNet – Input Segmentation



Adapted from https://www.youtube.com/watch?v=VMu0goeii3g

LongNet – Segment Attention

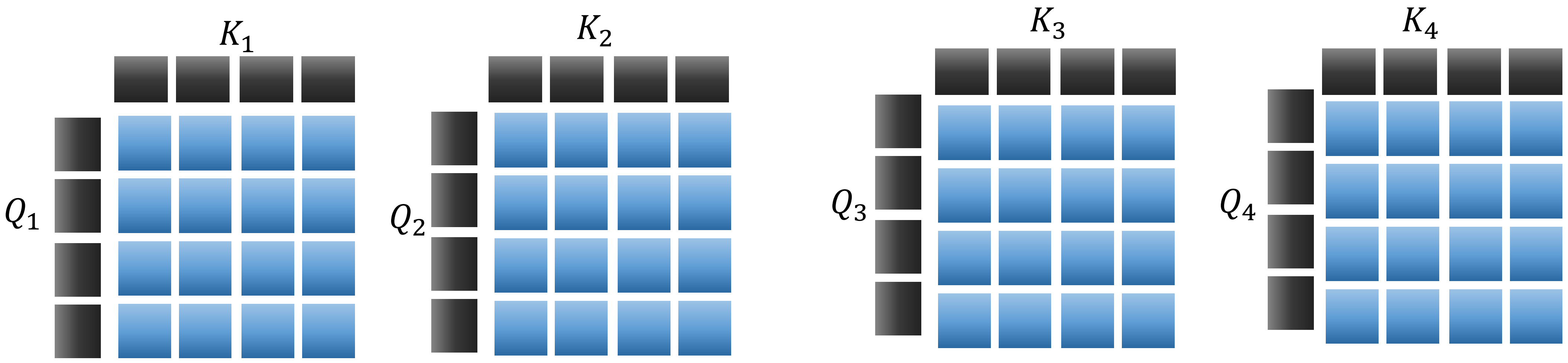


$O\left(\frac{N}{w}w^2\right) = O(Nw)$ computations

w is the length of each segment

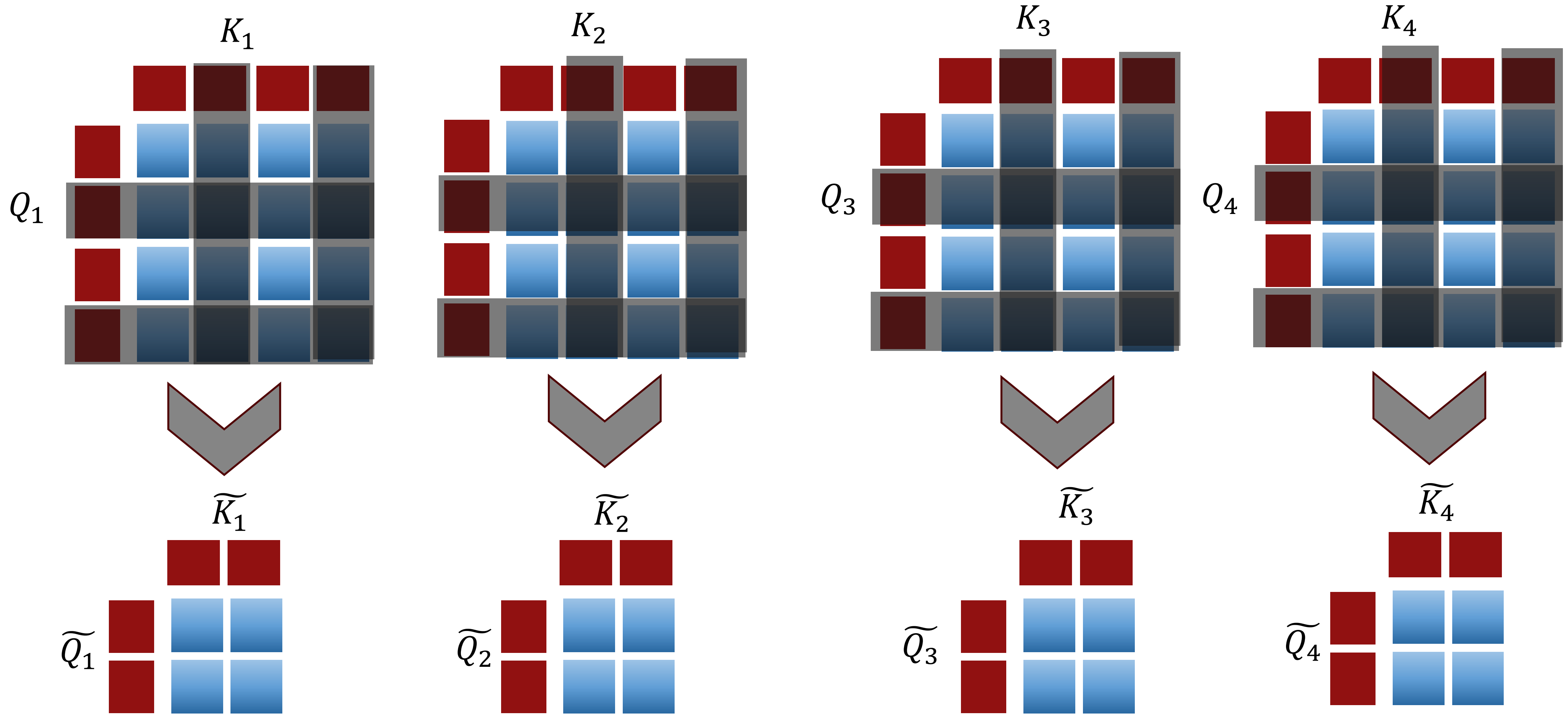
Adapted from https://www.youtube.com/watch?v=VMu0goeii3g

LongNet sparsification



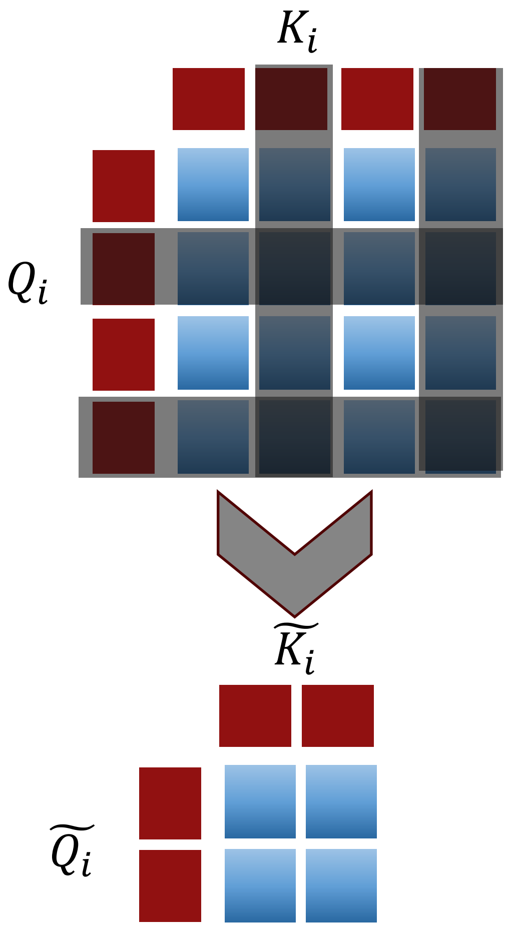
Adapted from https://www.youtube.com/watch?v=VMu0goeii3g

LongNet sparsification



Sparsification rate $r = 2$

LongNet sparsification



Sparsification rate $r = 2$

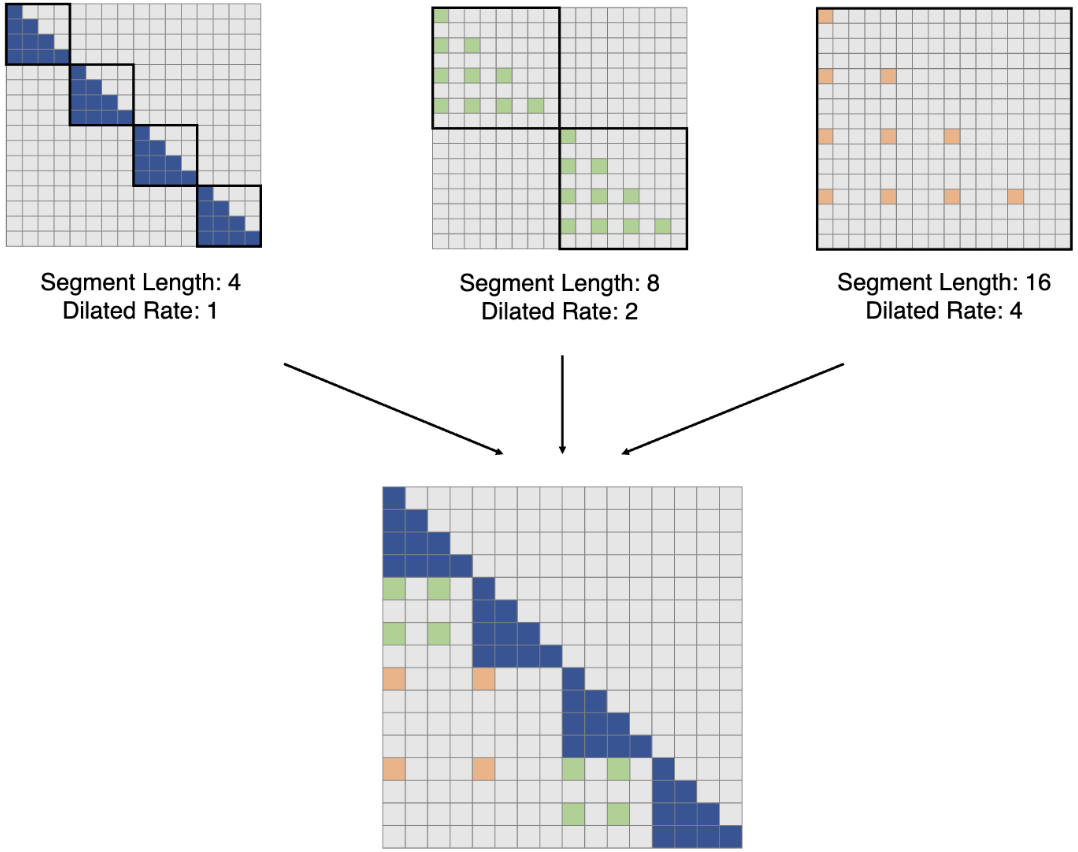
$$\widetilde{O}_i = \mathrm{softmax}(\widetilde{Q}_i \widetilde{K}_i^T)\widetilde{V}_i$$

$$\hat{O}_i = \{\widetilde{O}_{i,j} | j \bmod r = 0; 0 | j \bmod r \neq 0\}$$

$$O = [\hat{O}_0, \hat{O}_1, \ldots, \hat{O}_{\frac{N}{w}-1}]$$

https://arxiv.org/abs/2307.02486

LongNet – segment/sparsification mixture



A mixture of segment sizes and dilation rates are used.

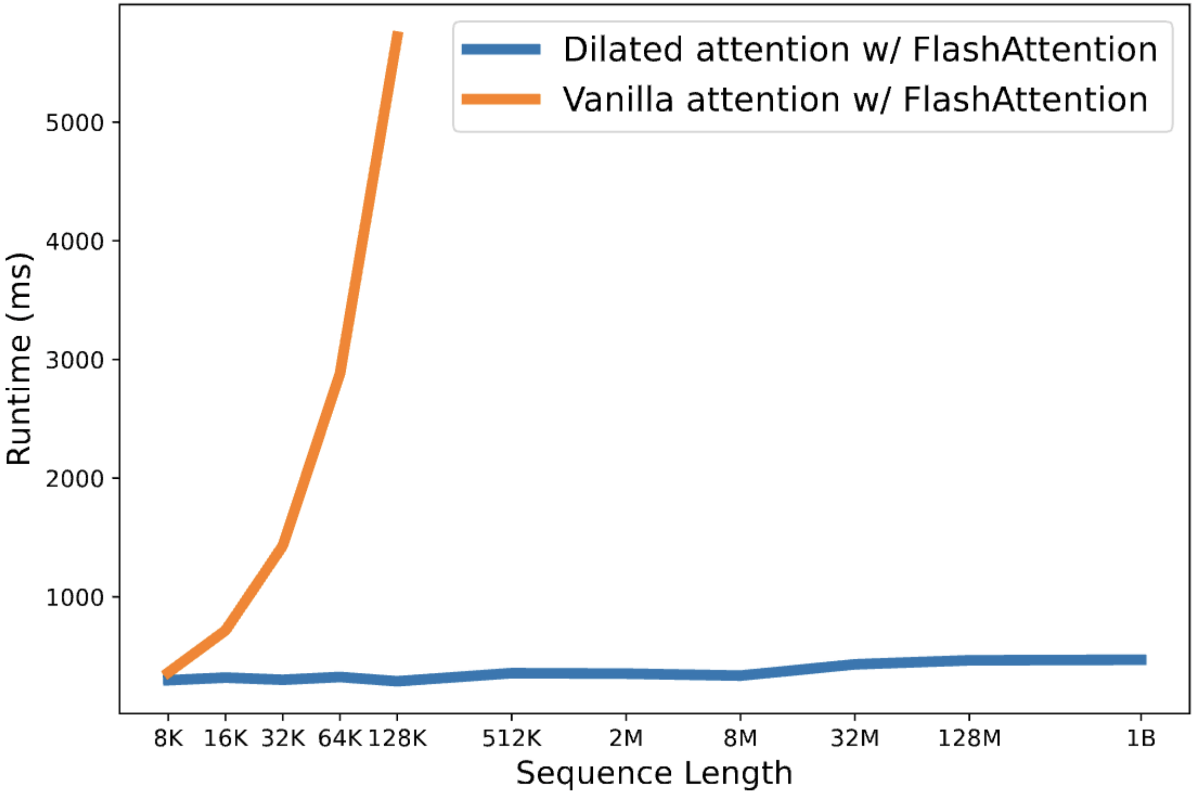
$$O = \sum_{i=1}^{k} \alpha_i O|_{r_i, w_i}$$

$$\alpha_i = \frac{s_i}{\sum_j s_j}$$

s_i is the denominator of softmax for $O|_{r_i, w_i}$

https://arxiv.org/abs/2307.02486

LongNet - runtime



- Time taken for vanilla attention grows quadratically.
- For dilated attention, (w_i, r_i) can be chosen to keep it constant or linear albeit at the cost of performance.

  - Still requires training with data with long-context length
    - Expensive compute
  - Can we:
    - Train with short contexts
    - Test with long contexts

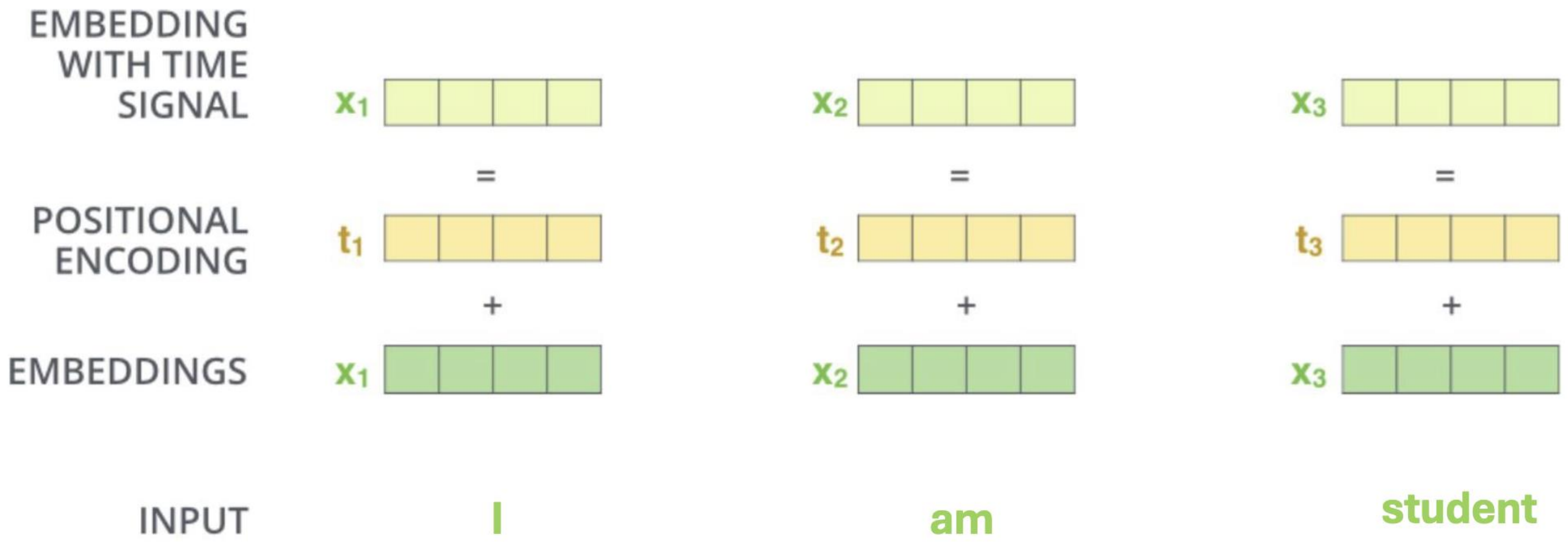
https://arxiv.org/abs/2307.02486

Agenda

- Long Contexts & Challenges
- Key Papers and Their Contributions
  - LongNet
  - **ALiBi**
  - Positional Interpolation
  - Lost in the Middle
- Discussion and Future Directions

Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation

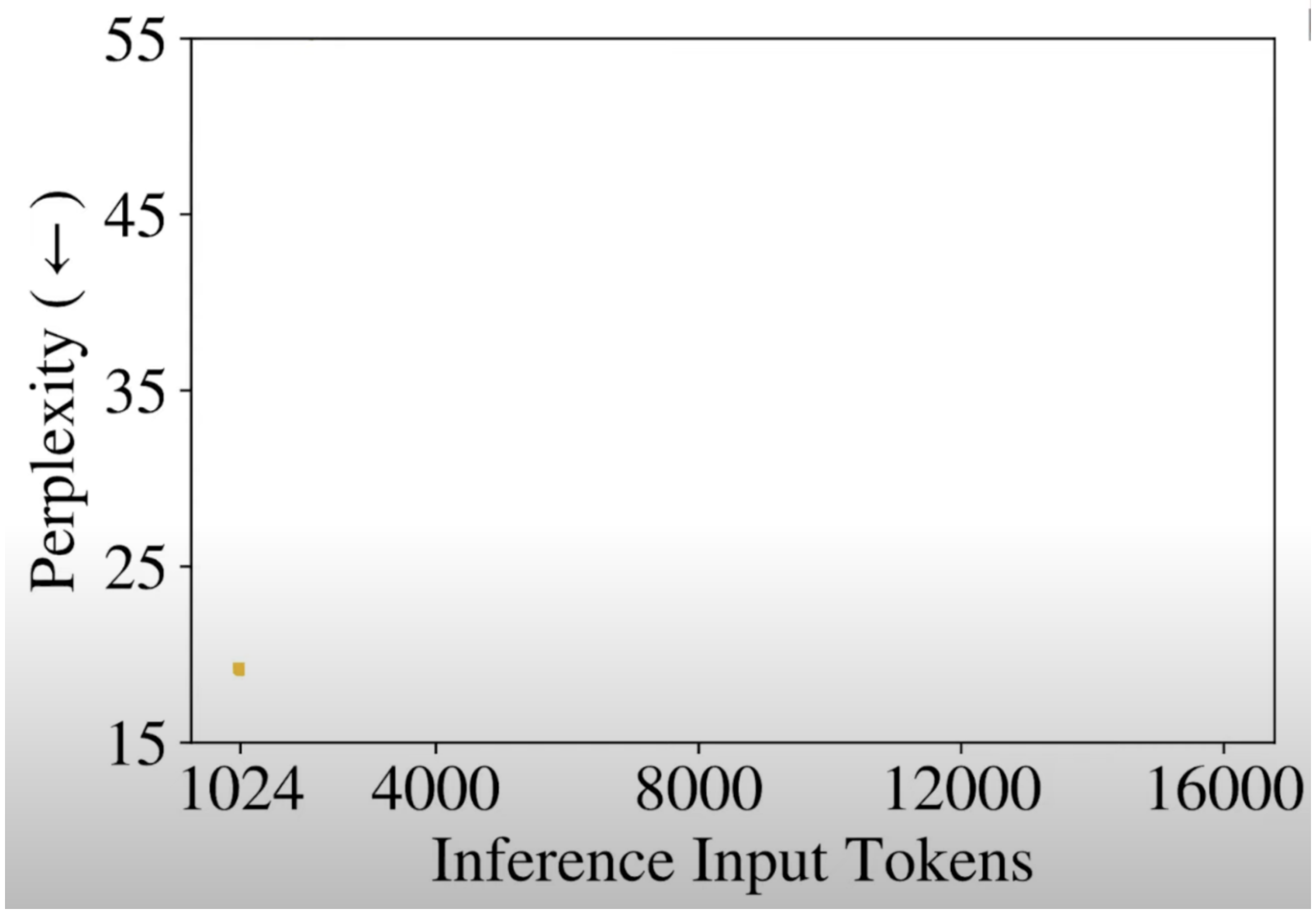
Recap: Position Embeddings



Question: What if the model was trained for max 1024 tokens? Can I still use t_{2048}?

From: Introduction to Transformer (Part II) of this course

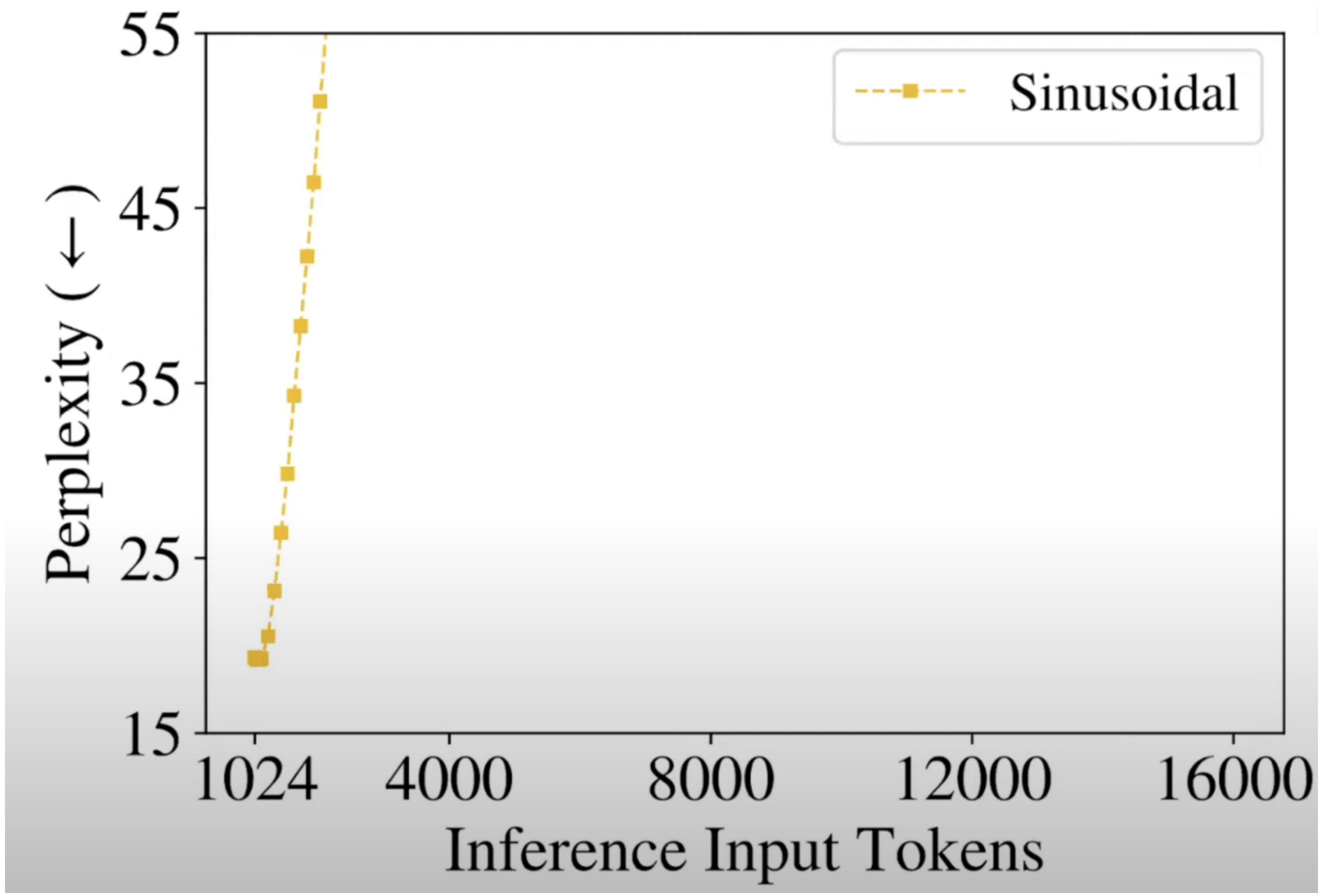
Extrapolating sinusoidal embeddings



Language model trained with
L=1024 context length
on WikiText103
247M parameters

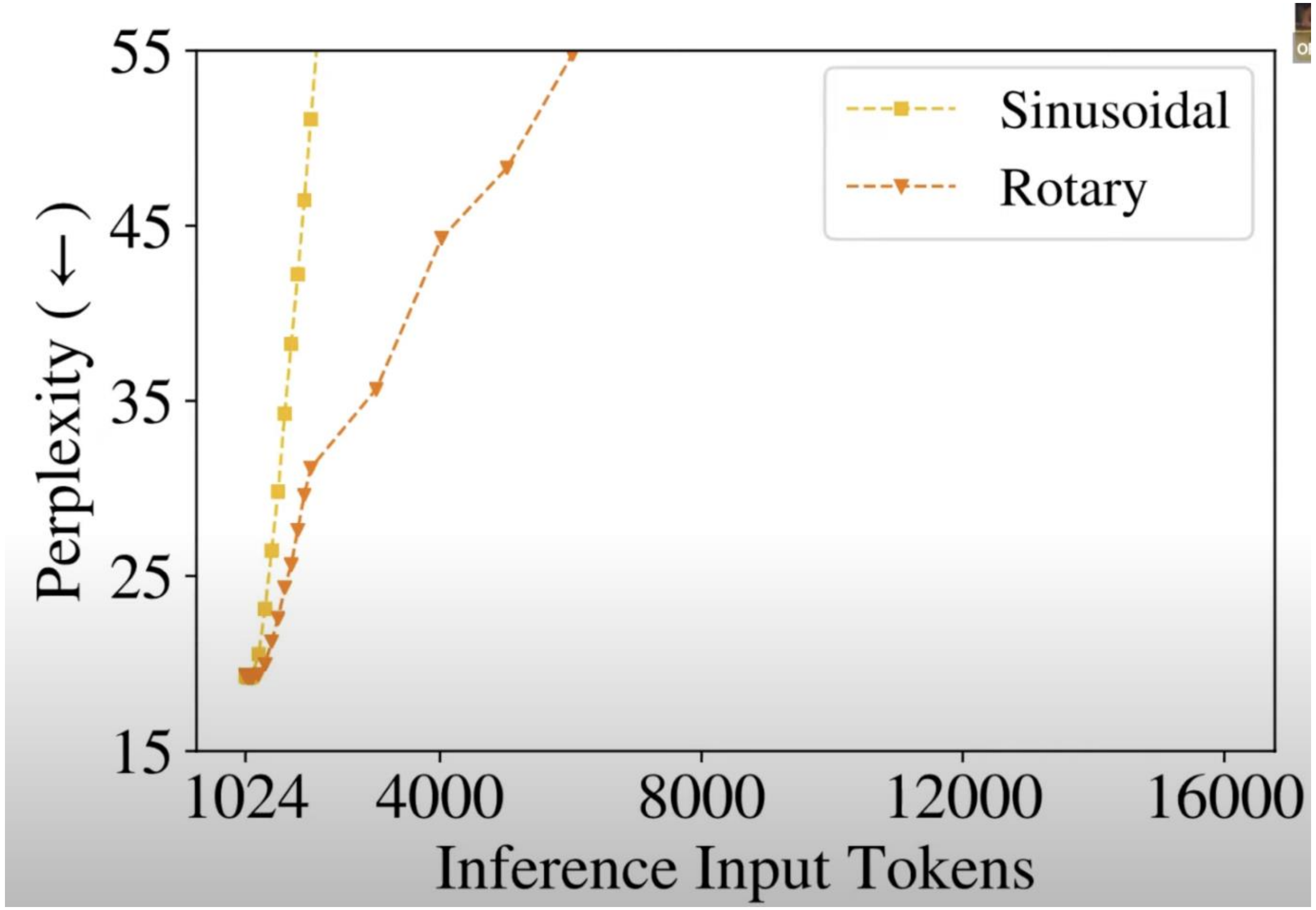
https://www.youtube.com/watch?v=Pp61ShI9VGc

Extrapolating sinusoidal embeddings



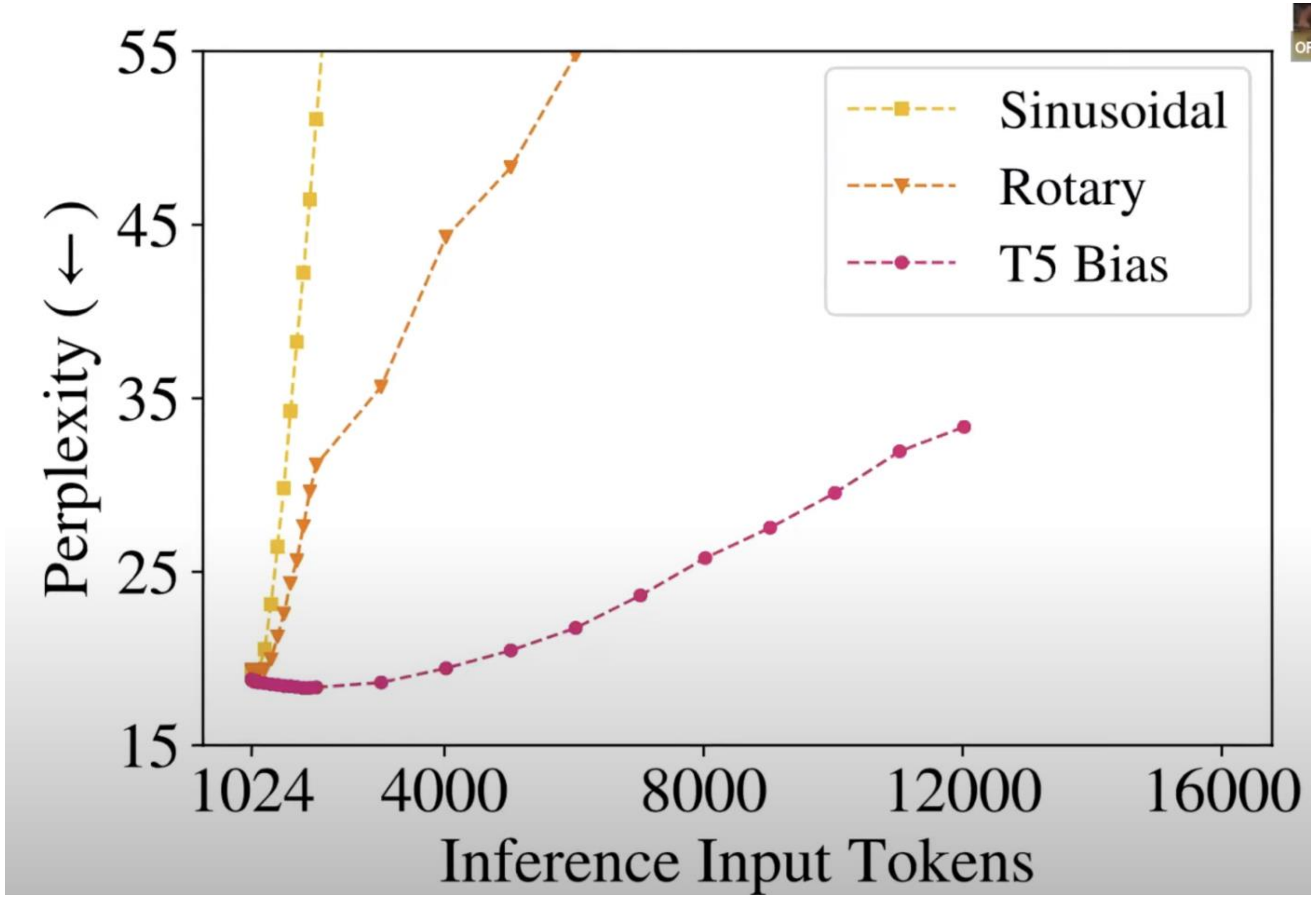
https://www.youtube.com/watch?v=Pp61ShI9VGc

Extrapolating rotary embeddings



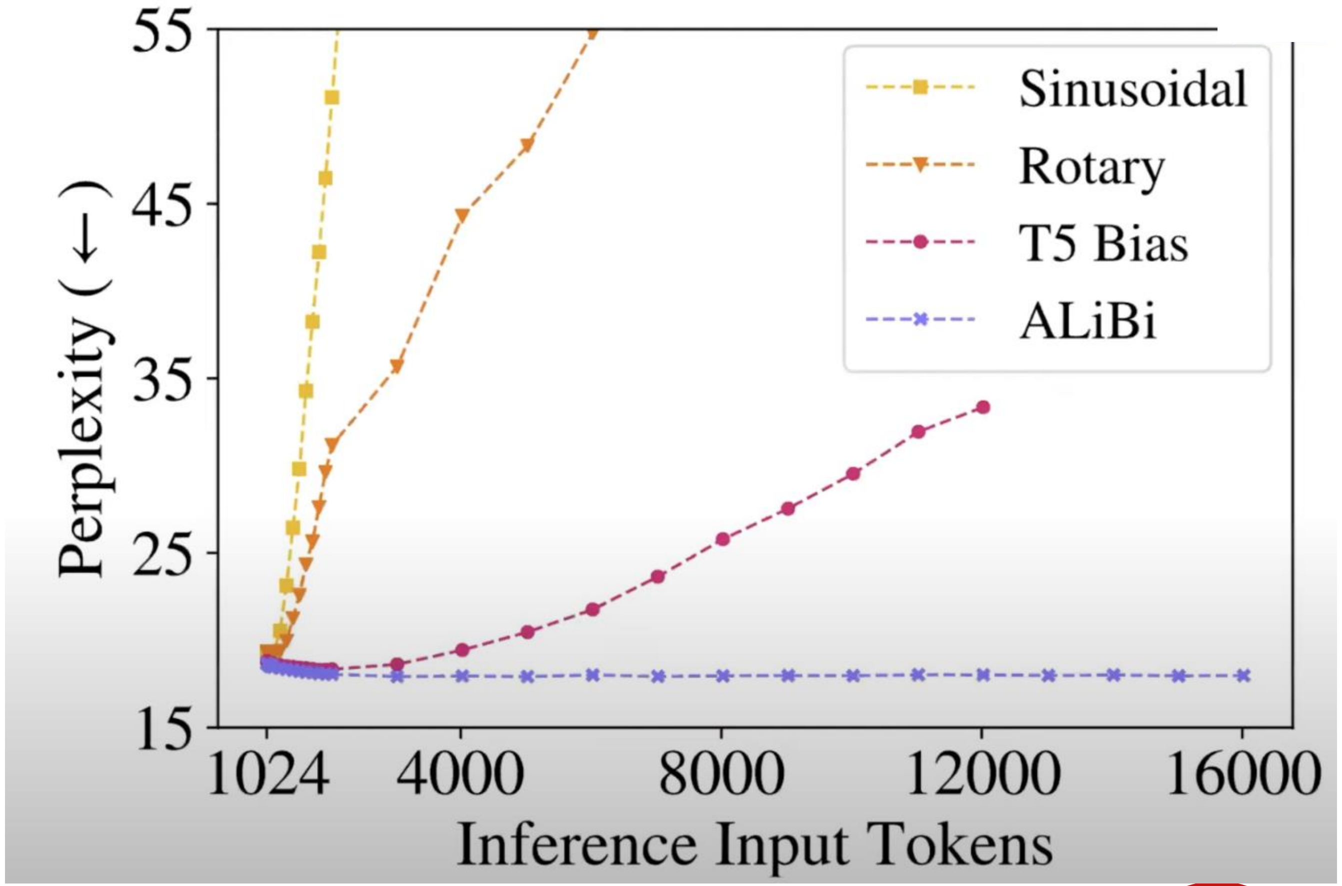
https://www.youtube.com/watch?v=Pp61ShI9VGc

Extrapolating position embeddings



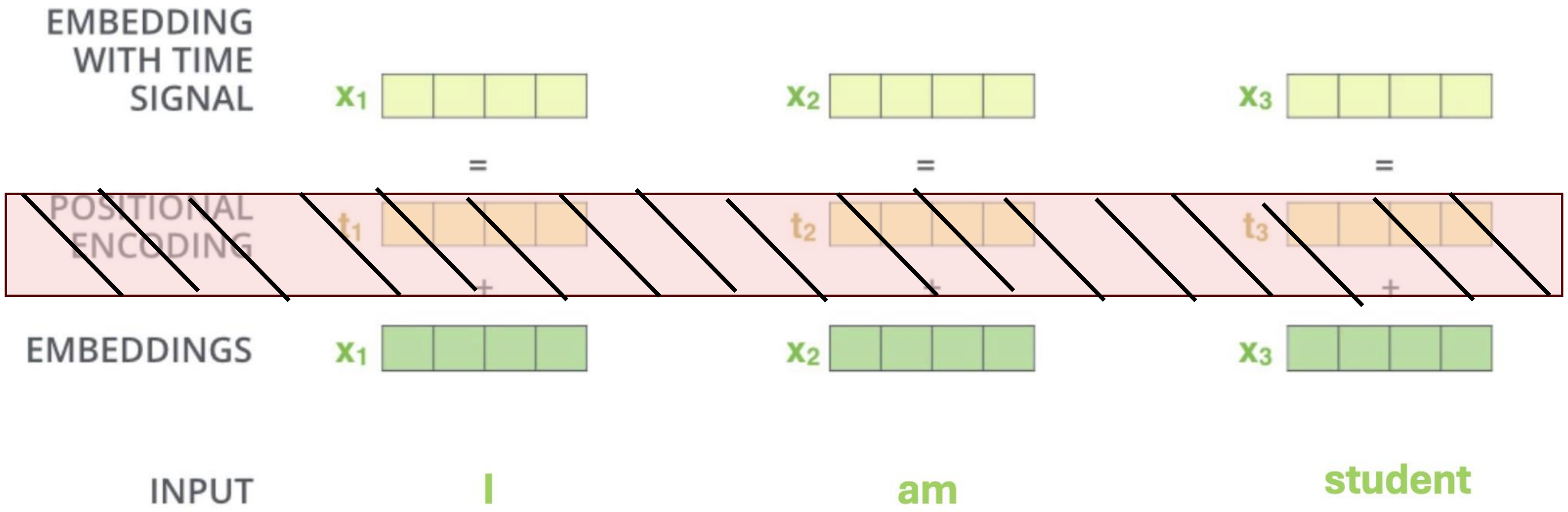
https://www.youtube.com/watch?v=Pp61ShI9VGc

Extrapolating position embeddings



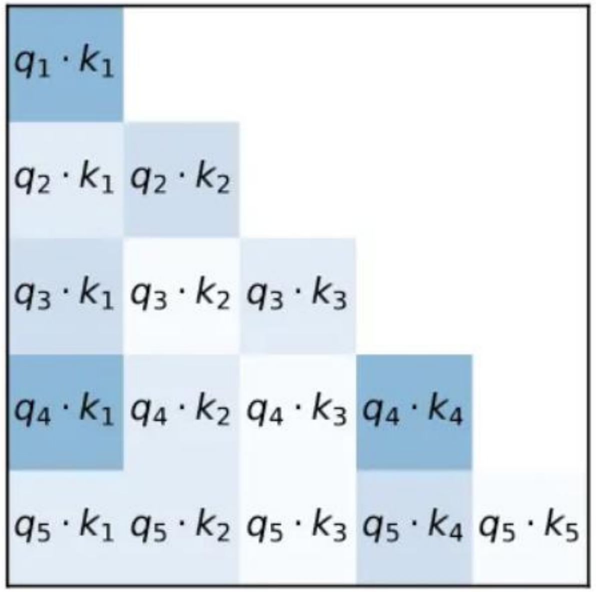
https://www.youtube.com/watch?v=Pp61ShI9VGc

Remove position embeddings altogether



Can I remove position embeddings altogether and add position information only in the attention matrix?

Standard attention

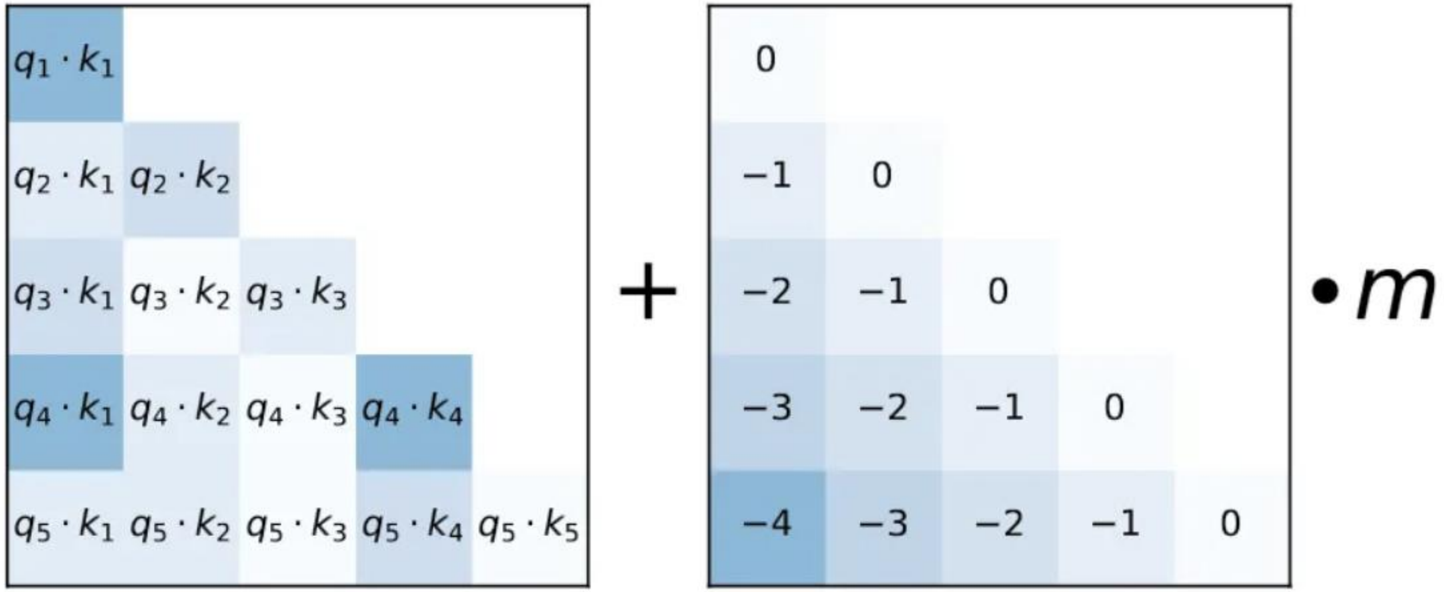


$$\mathrm{softmax}(\mathbf{q}_i \mathbf{K}^\top)$$

Let us add position information only in the attention matrix?

https://www.youtube.com/watch?v=Pp61ShI9VGc

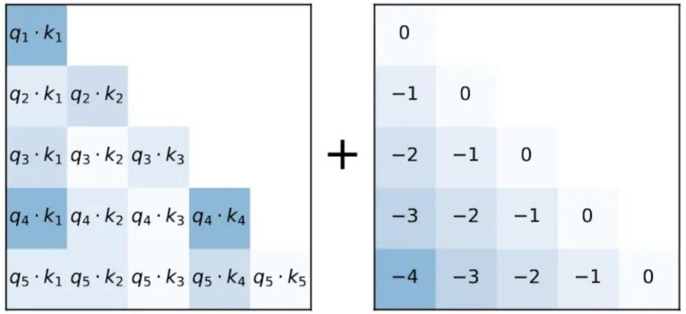
Attention with linear bias (ALiBi)



$$\mathrm{softmax}(\mathbf{q}_i \mathbf{K}^\top + m \cdot [-(i-1), \ldots, -2, -1, 0])$$

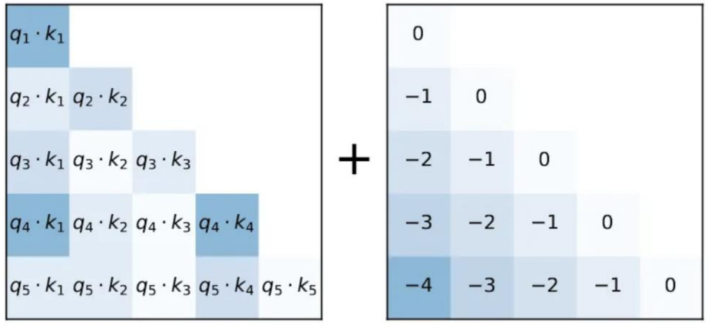
https://www.youtube.com/watch?v=Pp61ShI9VGc

ALiBi with multiple heads



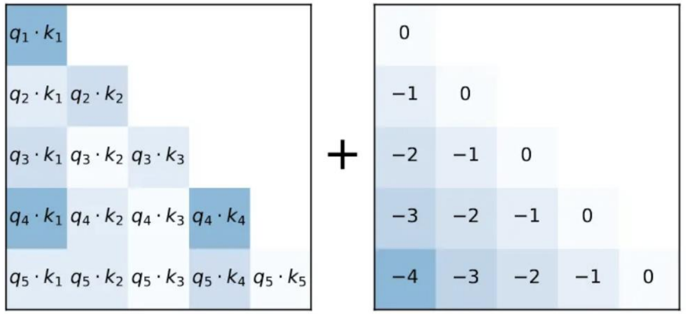
$\bullet\, m \qquad m = \frac{1}{2^{\frac{8}{n}}}$

Head 0



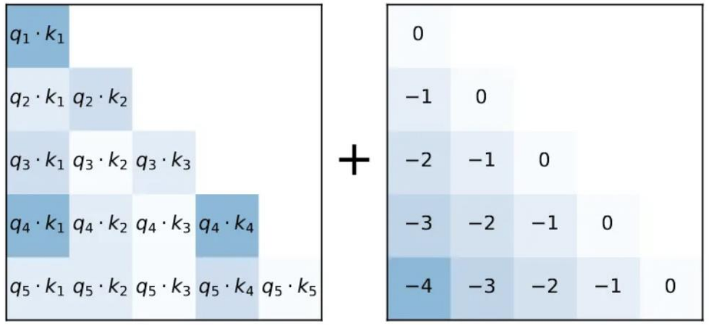
$\bullet\, m \qquad m = \frac{1}{2^{\frac{16}{n}}}$

Head 1



$\bullet\, m \qquad m = \frac{1}{2^{\frac{8*(n-1)}{n}}}$

Head n-1



$\bullet\, m \qquad m = \frac{1}{2^{8}}$

Head n

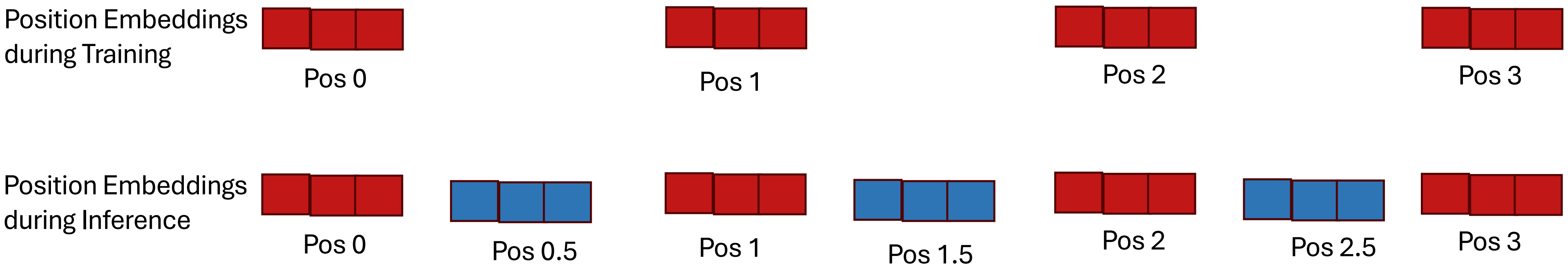
Agenda

- Long Contexts & Challenges
- Key Papers and Their Contributions
  - LongNet
  - ALiBi
  - **Positional Interpolation**
  - Lost in the Middle
- Discussion and Future Directions

Extending Context Window of Large Language Models via Positional Interpolation

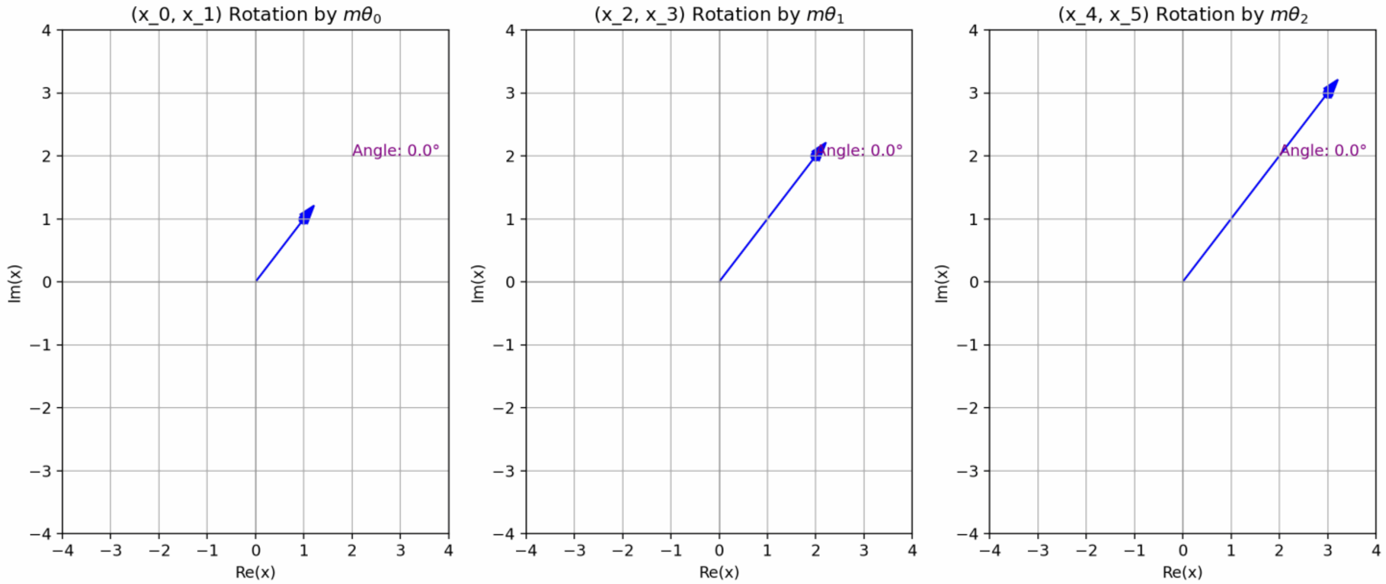
From ALiBi to position interpolation

- ALiBi - Some attention heads have narrow vision
  - Bias can get very large for long contexts
  - Model effectively learns to ignore distant tokens
- Back to Position Embeddings
  - Can we interpolate between the position embeddings used during training?



Recap: Rotary Position Embeddings (RoPE)

- Given $x = (x_0, \ldots, x_{d-1}) \in R^d$ at position m
  - $f(x, m) = \left[(x_0 + ix_1)e^{im\theta_0}, (x_2 + ix_2)e^{im\theta_1}, \ldots, (x_{d-1} + ix_d)e^{im\theta_{\frac{d}{2}-1}}\right]^T$



Extending Context Window of Large Language Models via Positional Interpolation

Attention scores for RoPE

- q is the query vector, k is the key vector
- Apply RoPE on both query & key vector

$$f(q,m) = \left[(q_0 + iq_1)e^{im\theta_0}, (q_2 + iq_2)e^{im\theta_1}, \ldots, (q_{d-1} + iq_d)e^{im\theta_{\frac{d}{2}-1}}\right]^T$$

Extending Context Window of Large Language Models via Positional Interpolation

Attention scores for RoPE

- q is the query vector, k is the key vector
- Apply RoPE on both query & key vector

$$f(q,m) = \left[(q_0 + iq_1)e^{im\theta_0}, (q_2 + iq_2)e^{im\theta_1}, \dots, (q_{d-1} + iq_d)e^{im\theta_{\frac{d}{2}-1}}\right]^T$$

$$f(k,n) = \left[(k_0 + ik_1)e^{in\theta_0}, (k_2 + ik_2)e^{in\theta_1}, \dots, (k_{d-1} + ik_d)e^{in\theta_{\frac{d}{2}-1}}\right]^T$$

Extending Context Window of Large Language Models via Positional Interpolation

Attention scores for RoPE

$$a(m,n) = Re < f(q,m), f(k,n) >$$

$$= Re\left[\sum_{j=0}^{\frac{d}{2}-1} \underbrace{(q_{2j} + iq_{2j+1})(k_{2j} - ik_{2j+1})}_{h_j} e^{i(m-n)\theta_j}\right]$$

Intuitively, it is a function of

- The inner product between $(q_{2j} + iq_{2j+1})$ & $(k_{2j} + ik_{2j+1})$
- The angle between position embeddings $(m-n)\theta_j$

Extending Context Window of Large Language Models via Positional Interpolation

Attention scores for RoPE

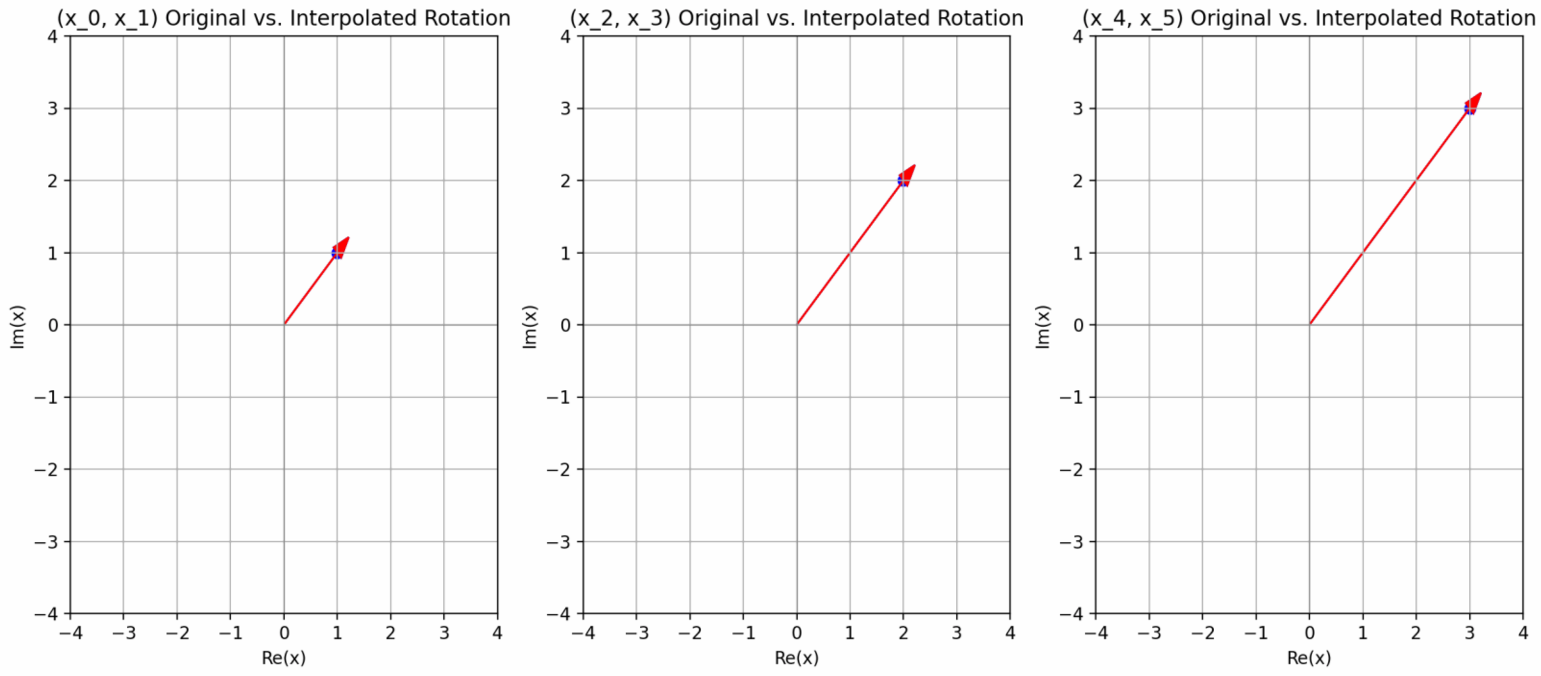
$$a(m,n) = Re < f(q,m), f(k,n) >$$

$$= Re\left[\sum_{j=0}^{\frac{d}{2}-1} (q_{2j} + iq_{2j+1})(k_{2j} - ik_{2j+1})e^{i(m-n)\theta_j}\right]$$

$$= \underbrace{Re\left[\sum_{j=0}^{\frac{d}{2}-1} h_j e^{i\left((m-n)\theta_j\right)}\right]}_{\text{Is bounded by } \frac{d}{2}\max_j |h_j|}$$

Extending Context Window of Large Language Models via Positional Interpolation

Position Interpolation



$\frac{L}{L'} = 2$ in this figure

This is called linear interpolation

$$f'(x, m) = f\left(x, \frac{mL}{L'}\right)$$

L is the max length used during pretraining
L' is the desired max-length

Interpolation bound

- The original LLM has been trained with integer values of $m - n$
- During interpolation, this can be fractional.
- How do the attention scores for these new fractional values look like?

Interpolation bound

- The original LLM has been trained with integer values of $m - n$
- During interpolation, this can be fractional.
- How do the attention scores for these new fractional values look like?

In practice, further finetuning is desirable on small fraction of examples

Theorem 2.1 (Interpolation bound). *For attention score* $a(s) = \mathrm{Re}\left[\sum_{j=0}^{d/2-1} h_j e^{\mathrm{i} s\theta_j}\right]$, *where* $\theta_j = c^{-2j/d}$, *its interpolation value* $a(s)$ *for* $s \in [s_1, s_2]$ *is bounded as follows:*

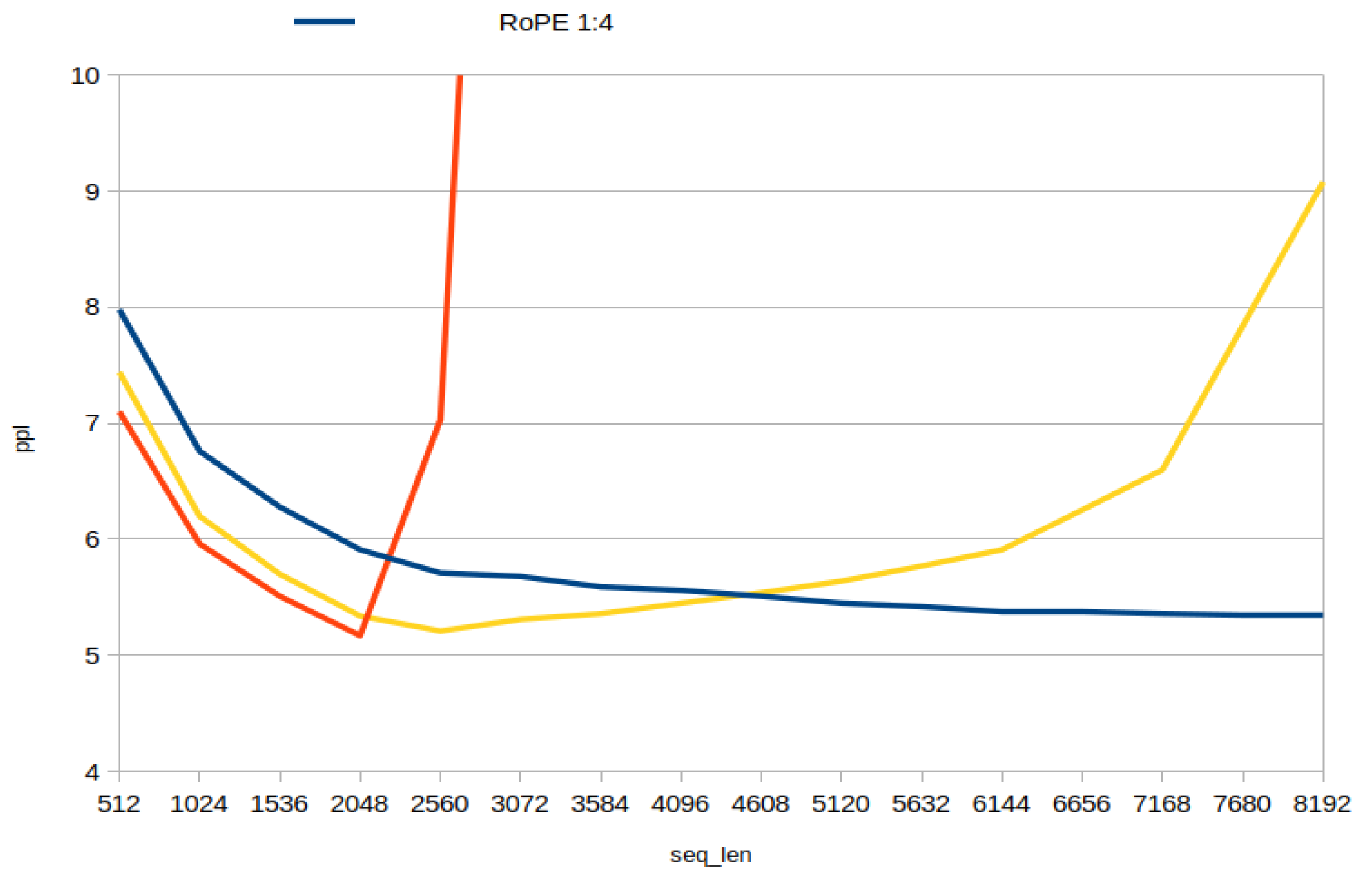
$$|a(s) - a_{\text{linear}}(s)| \leq d \left(\max_j |h_j|\right) \frac{(s - s_1)(s_2 - s)}{8 \ln c} \tag{5}$$

where $a_{\text{linear}}(s)$ *is the linear interpolation of two grid point* $a(s_1)$ *and* $a(s_2)$ *that are known to behave well, enforced by LLM pre-training:*

$$a_{\text{linear}}(s) := (1 - \lambda(s))a(s_1) + \lambda(s)a(s_2), \qquad \lambda(s) := \frac{s - s_1}{s_2 - s_1} \tag{6}$$

Extending Context Window of Large Language Models via Positional Interpolation

Position interpolation works



- Llama-13B trained on 2048 tokens
- Further finetuned on $\frac{L}{L'} = 0.25$

More complex interpolation schemes have been proposed:
- NTK inspired
- YaRN

From: https://github.com/ggerganov/llama.cpp/discussions/1965#discussioncomment-6256563

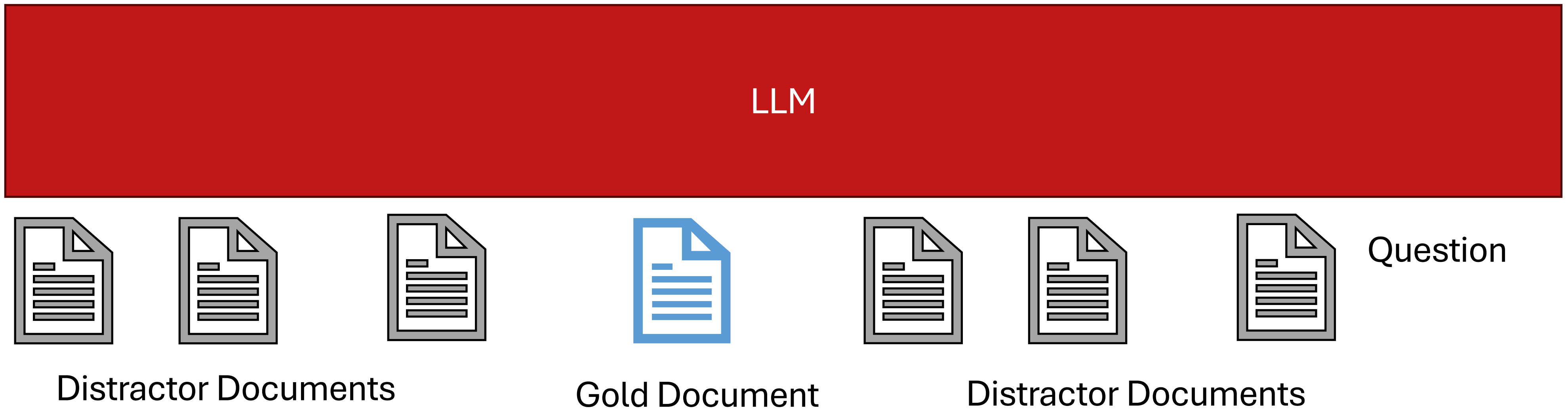
Agenda

- Long Contexts & Challenges
- Key Papers and Their Contributions
  - LongNet
  - ALiBi
  - Positional Interpolation
  - **Lost in the Middle**
- Discussion and Future Directions

How well do these models use long-context?

Experimental setting

- Vary the number of distractor documents.
- Vary the position of the gold document



An example

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...

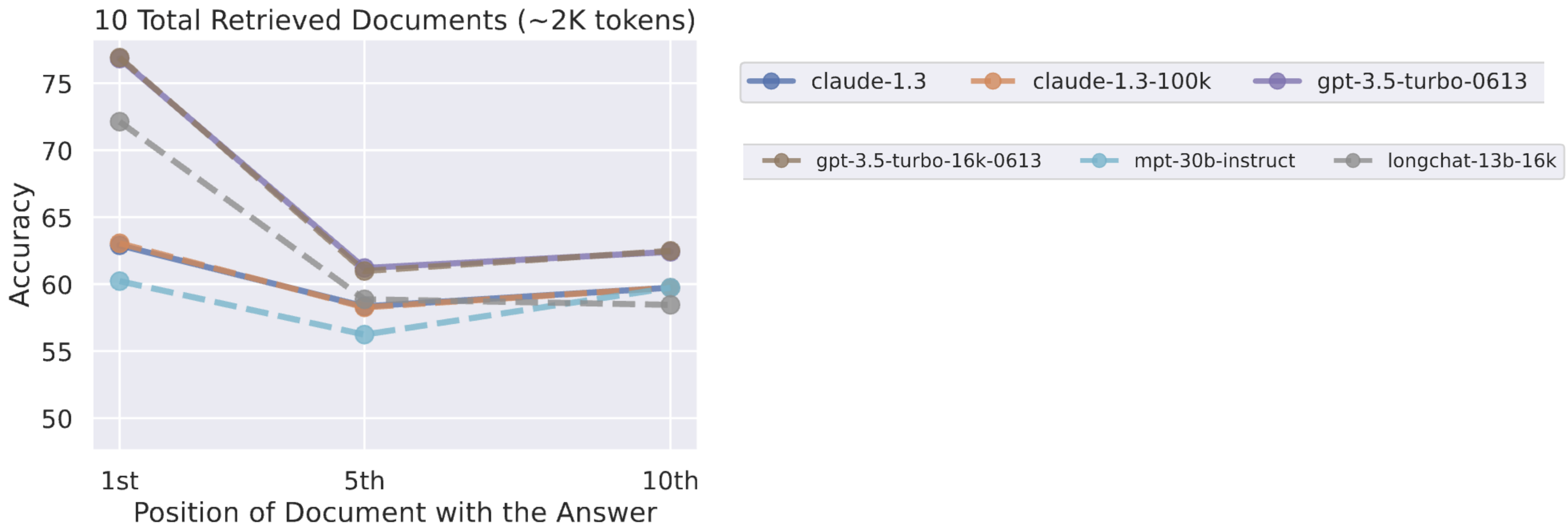
```
Document [3](Title: Scientist) and pursued through a unique method, was essentially
in place. Ramón y Cajal won the Nobel Prize in 1906 for his remarkable...

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

Desired Answer

```
Wilhelm Conrad Röntgen
```

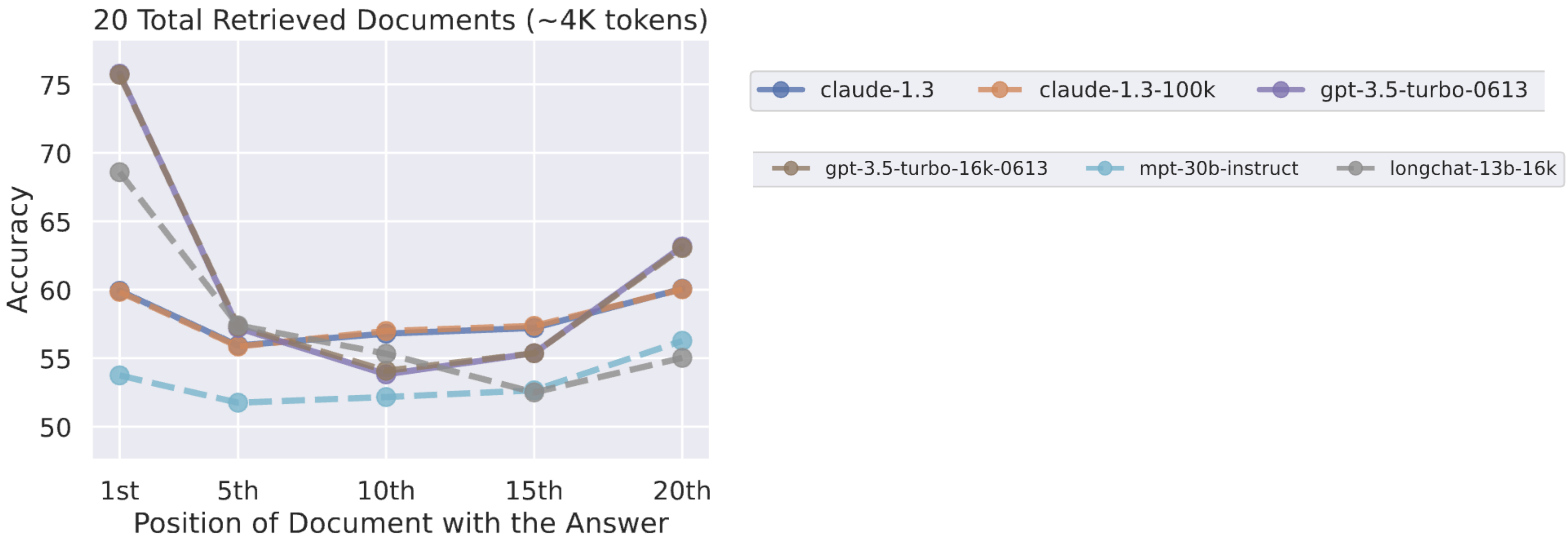
Lost in the Middle: How Language Models Use Long Contexts

Effect of changing the position of gold document



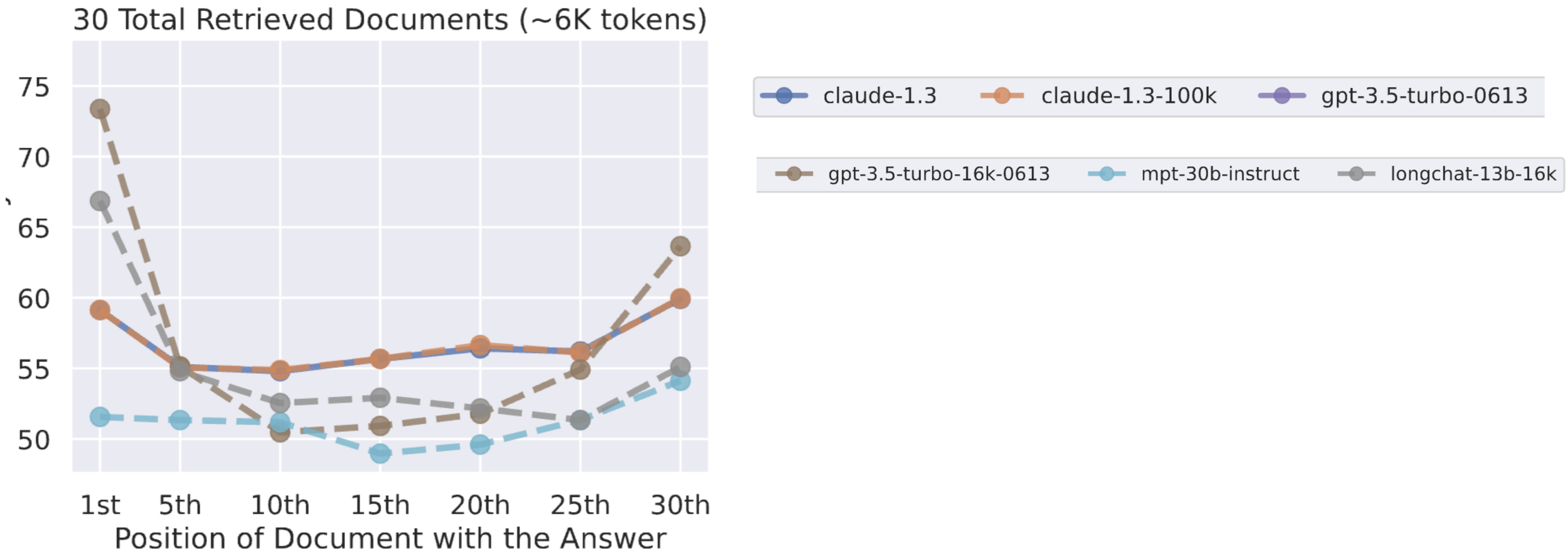
Lost in the Middle: How Language Models Use Long Contexts

Effect of changing the position of gold document



Lost in the Middle: How Language Models Use Long Contexts

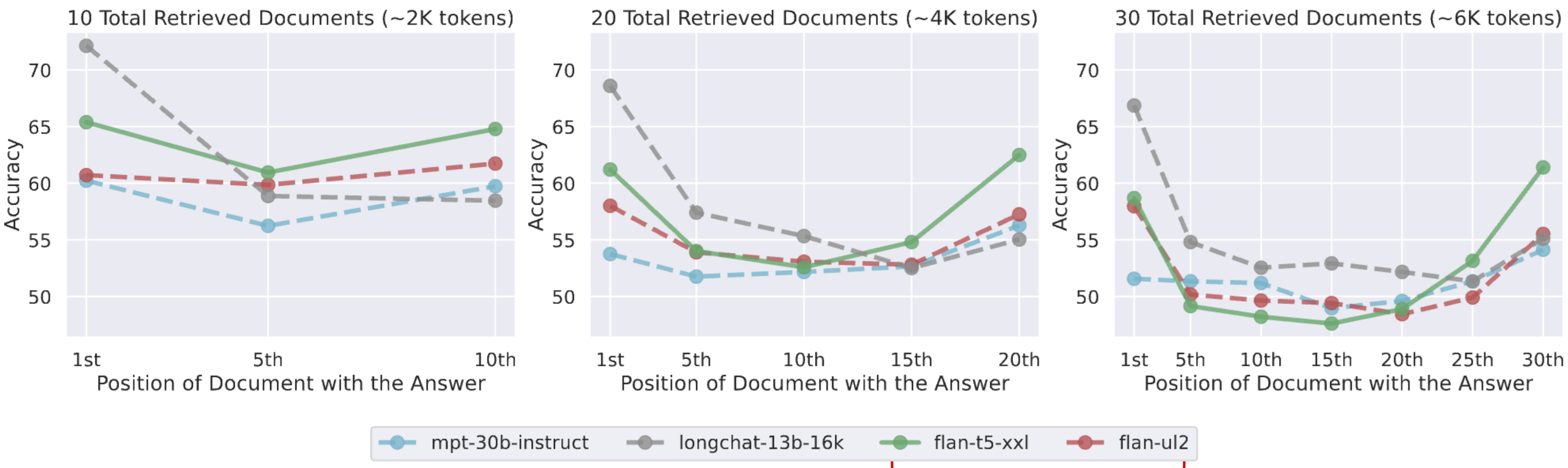
Effect of changing the position of gold document



Lost in the Middle: How Language Models Use Long Contexts

Effect of model architecture

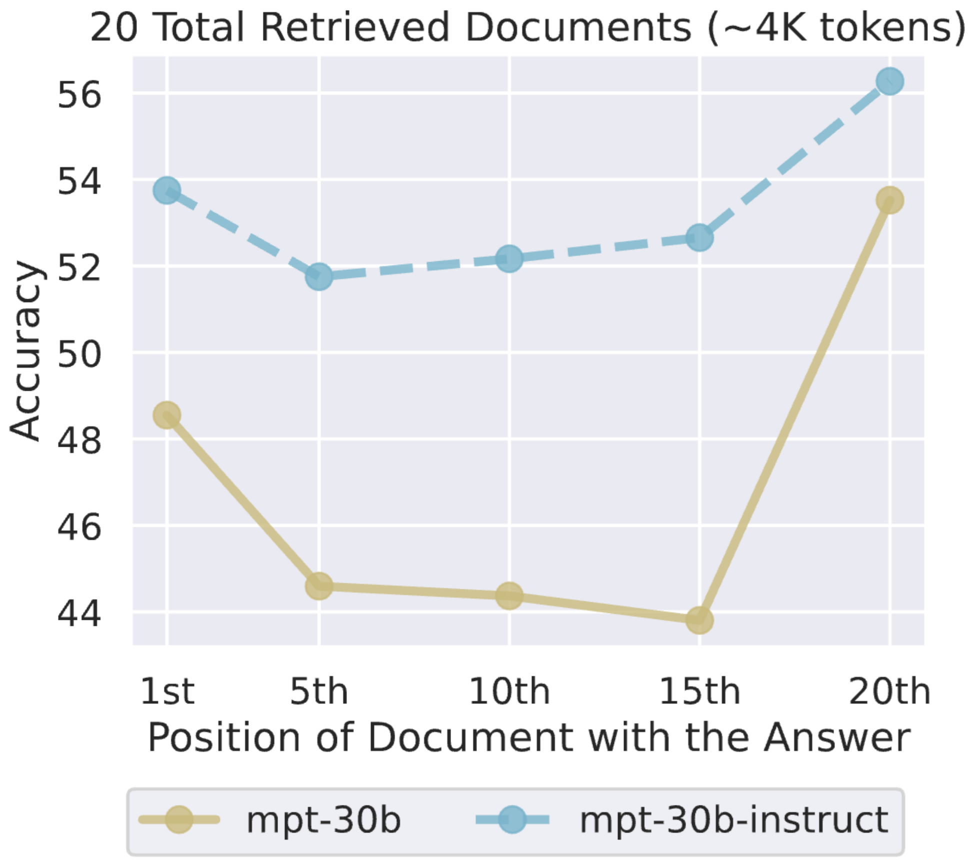
- Do bidirectional models fair better?



Encoder-decoder models

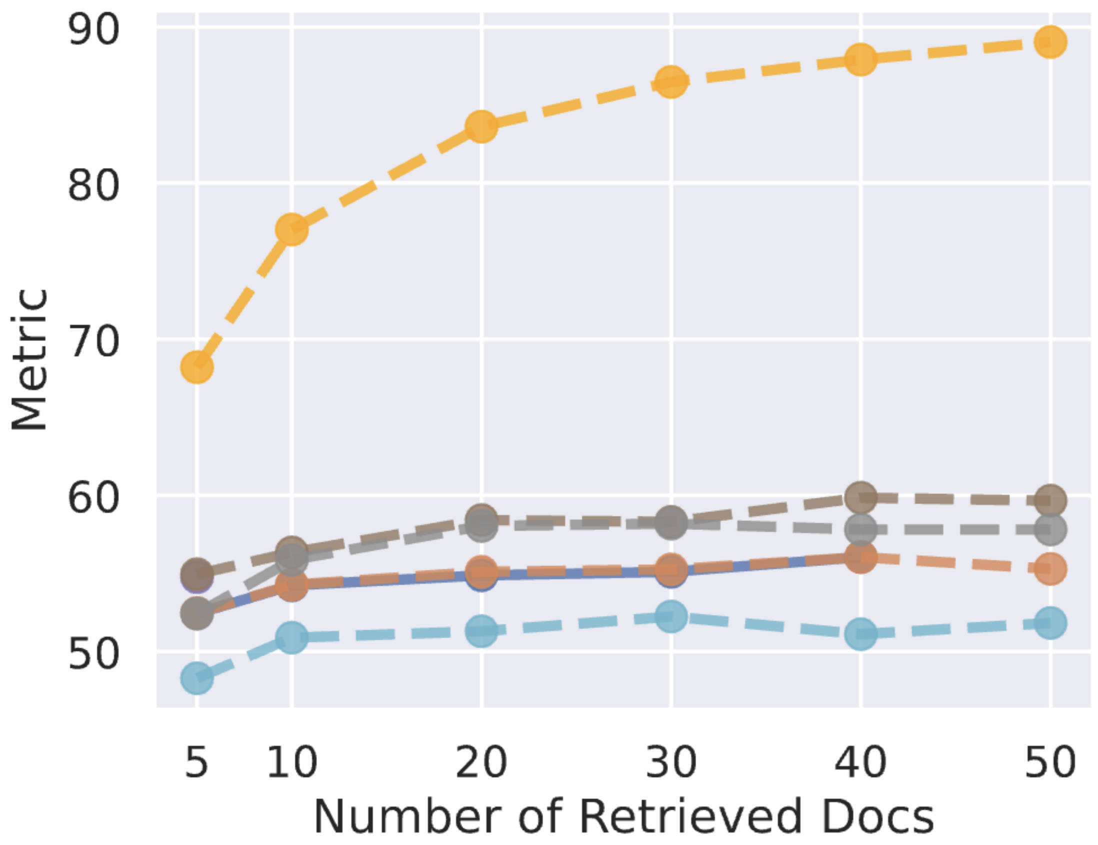
Lost in the Middle: How Language Models Use Long Contexts

Effect of instruction fine-tuning



Lost in the Middle: How Language Models Use Long Contexts

Retriever recall vs model performance



Retrieving more documents doesn't lead to improved performance

Lost in the Middle: How Language Models Use Long Contexts

Discussion

- Techniques like **LongNet**, **ALiBi**, and **Positional Interpolation** aim to enable LLMs to process longer sequences efficiently.
  - **LongNet** – Dilated attention to reduce computational complexity
  - **ALiBi**, and **Positional Interpolation** – Improve how models handle position information
- As noted in **"Lost in the Middle"**, models may still overlook important information in lengthy inputs.
- Ensuring model's performance when extending beyond trained context lengths remains an active research area.