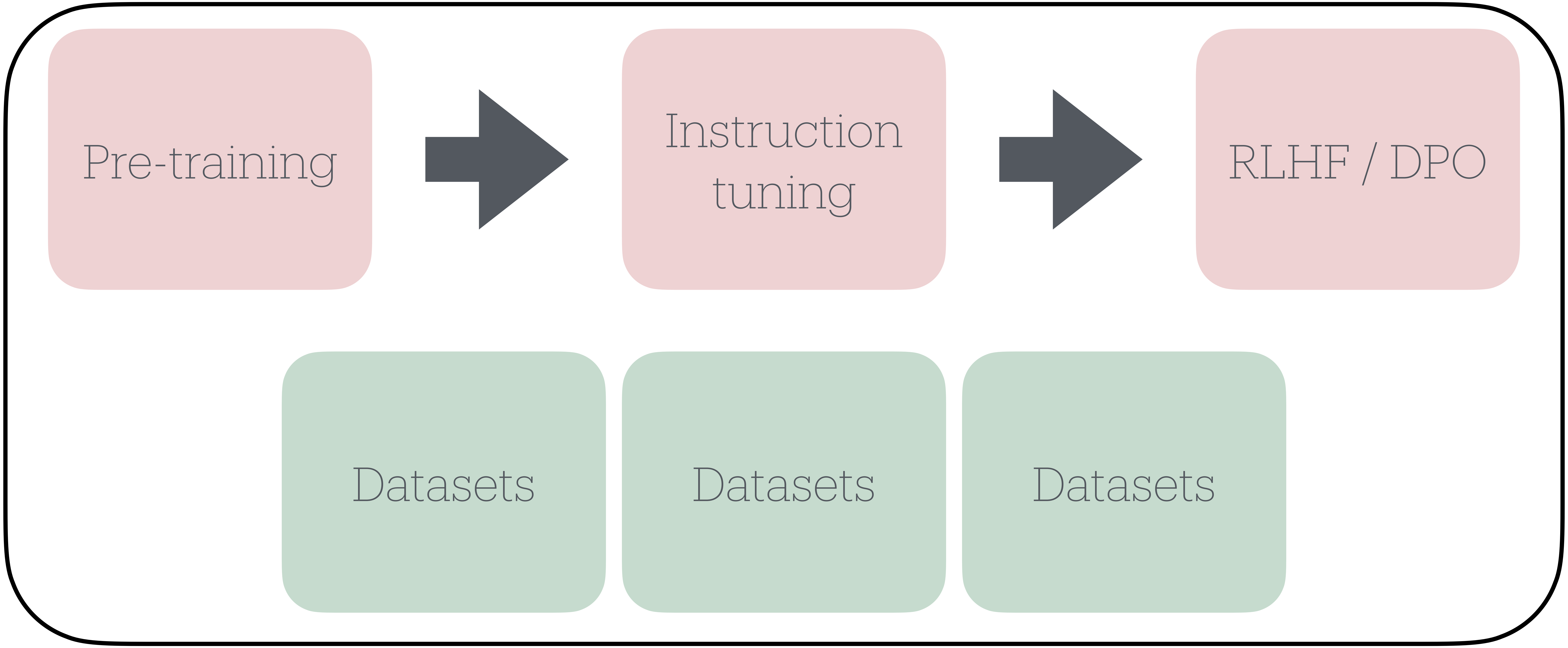
Long Context

Philipp Krähenbühl, UT Austin

Full Picture

Basic LLM



Pre-training

Instruction tuning

RLHF / DPO

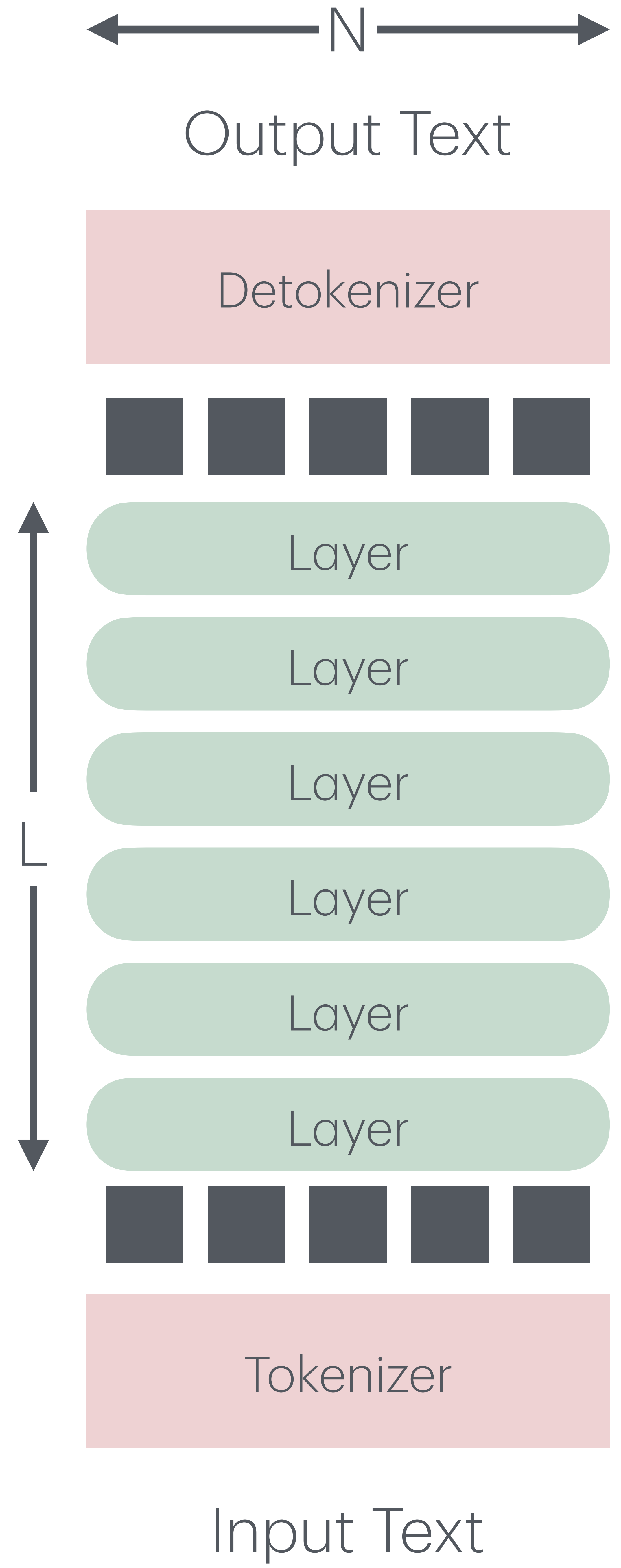
Datasets

Datasets

Datasets

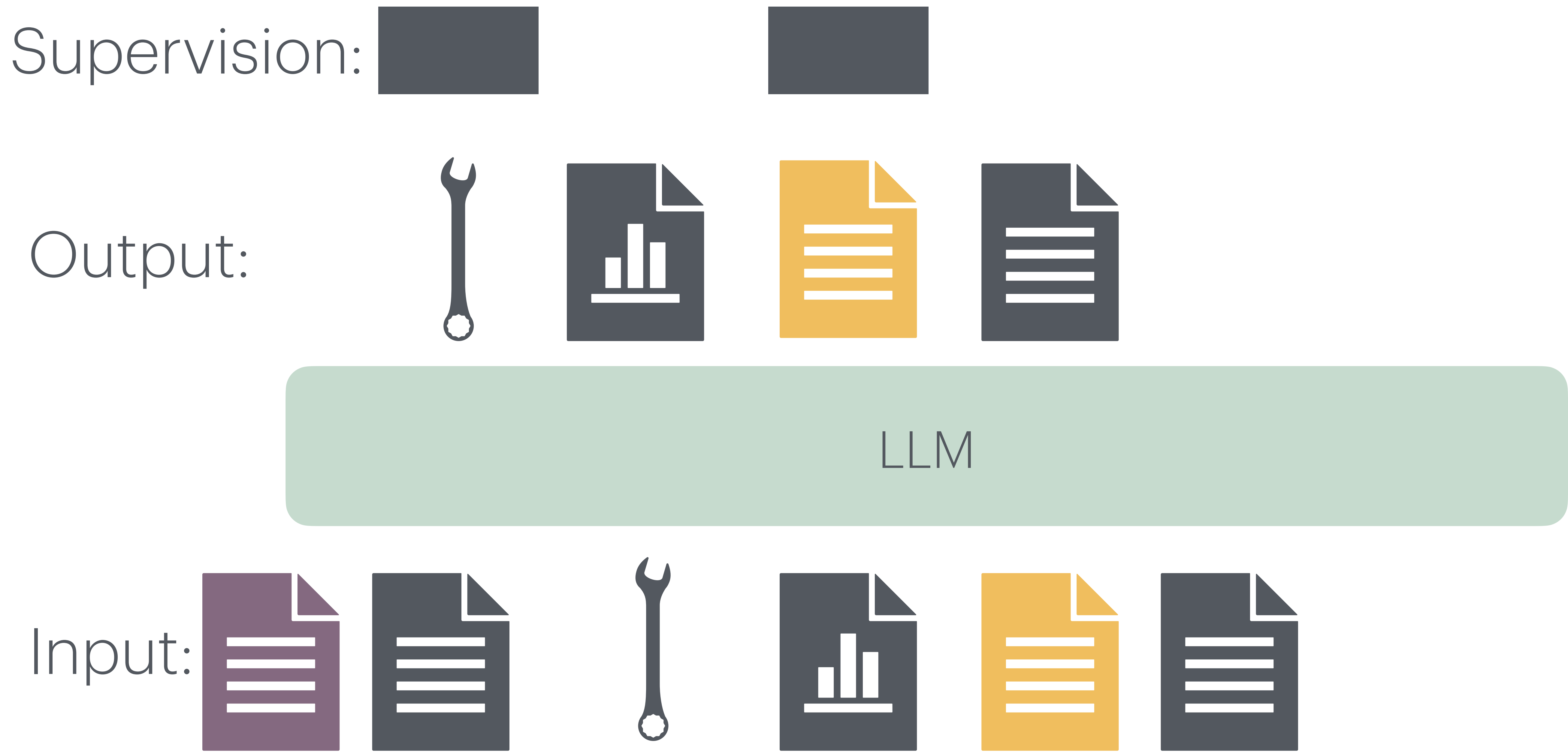
Training and Generation

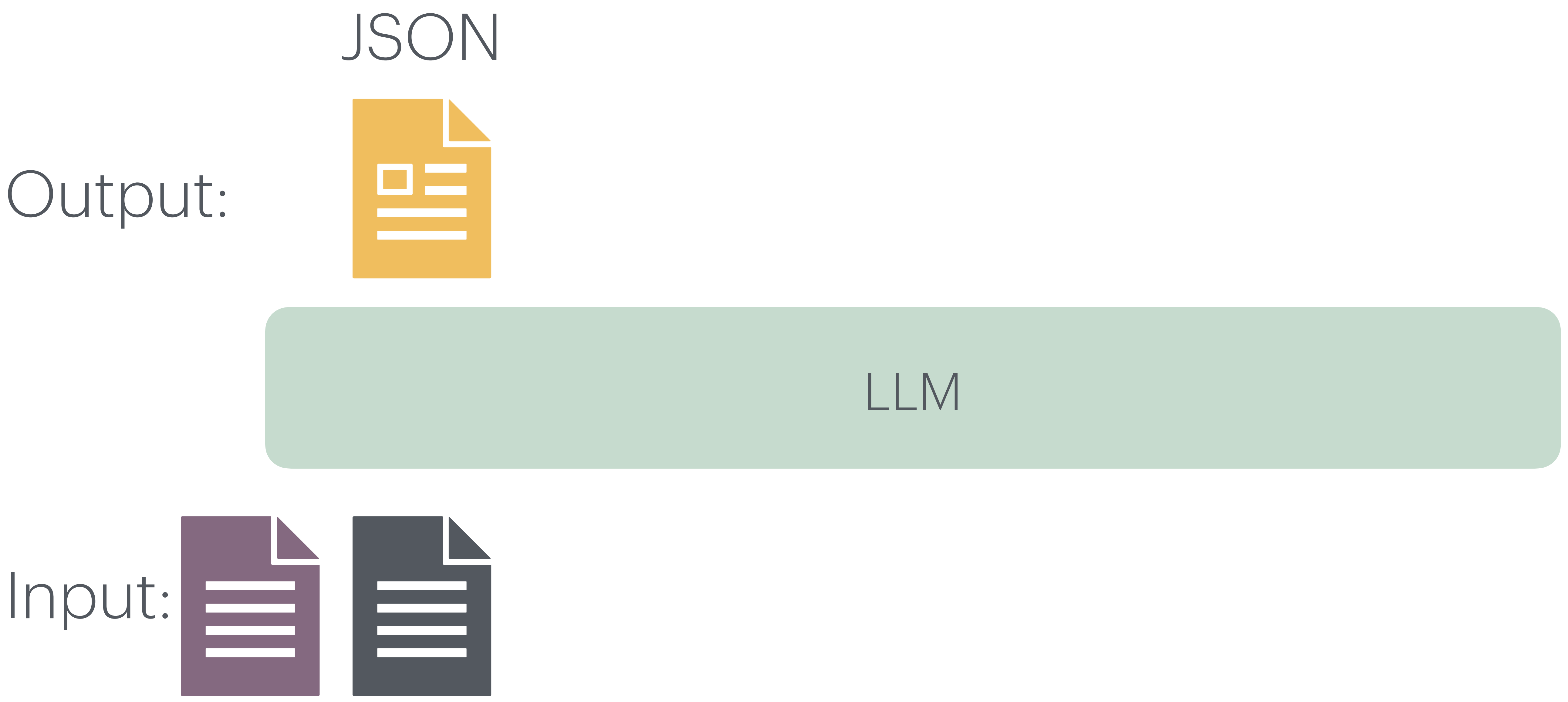
	Training	Training - Checkpointi	Generation	Paged Attention	Speculative decoding
Peak Memory	O(NL)	$O(NL^{\frac{1}{2}})$	O(N)	O(NL)	O(NL)
Runtime	$O(N^2L)$	$O(2\ N^2L)$	$O(N^3L)$	$O(N^2L)$	$O(N^2L)$
# forward	1	1	N	N	N / α



Tools and Structured outputs

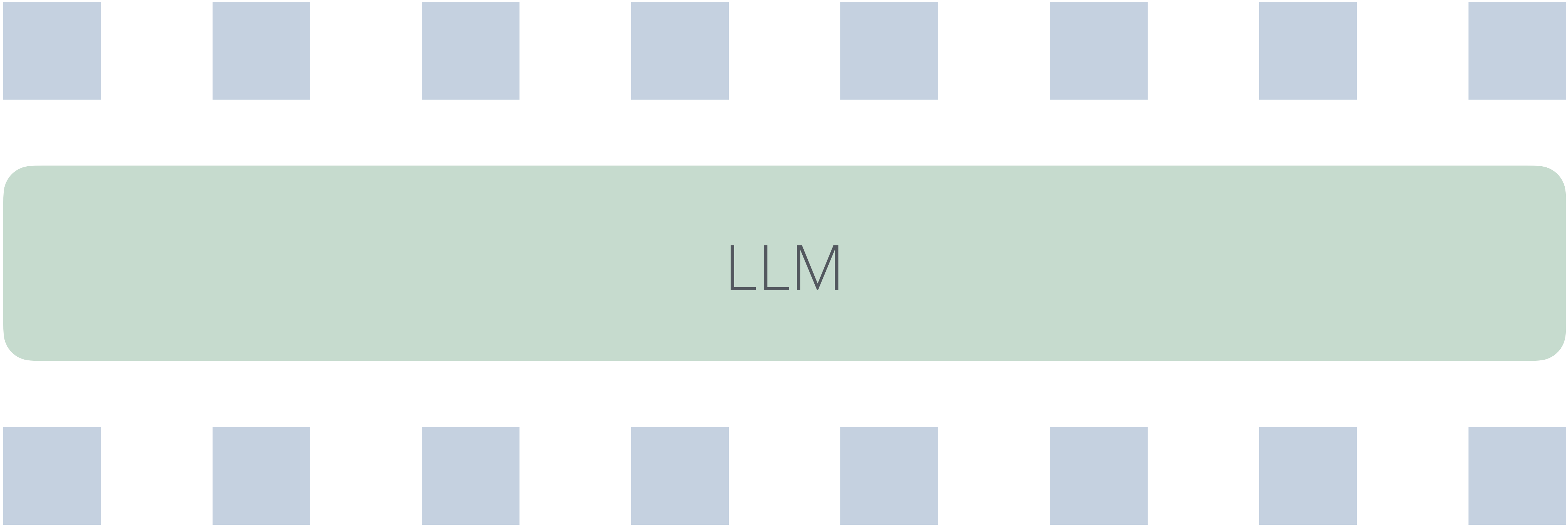
- Tools
  - Special tags, Special chat-template
- Structured output
  - Option 1.1: Write a robust parser (in python)
    - Let LLM know that you failed to parse
  - Option 1.2: Constrain output
  - Option 2: Use a tool, arguments = json fields





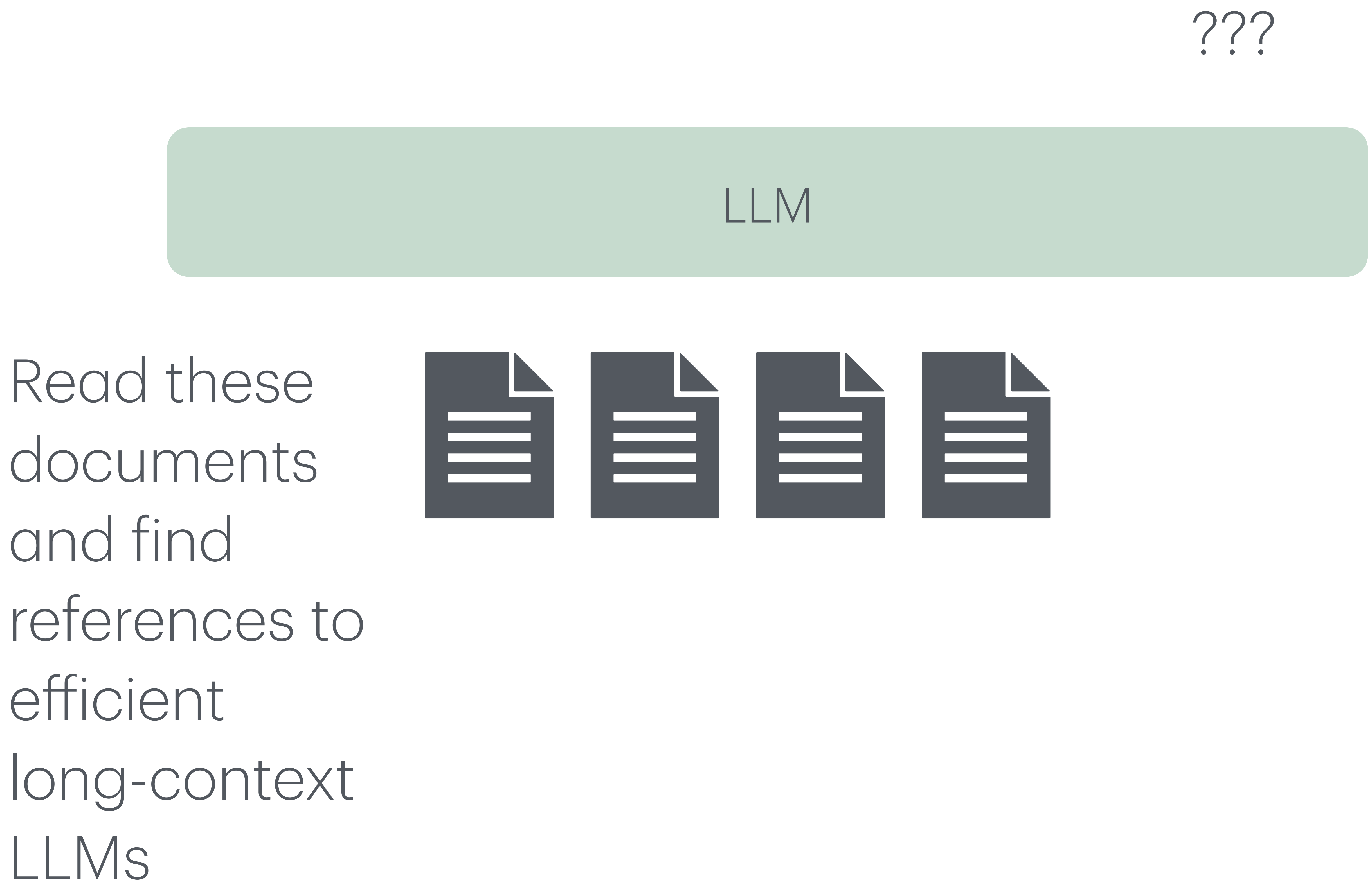
Long Context

- Current model are **pre-trained** on **2-8k** token sequences



Long Context

What happens if we feed ten’s of thousands of tokens into an LLM?



???

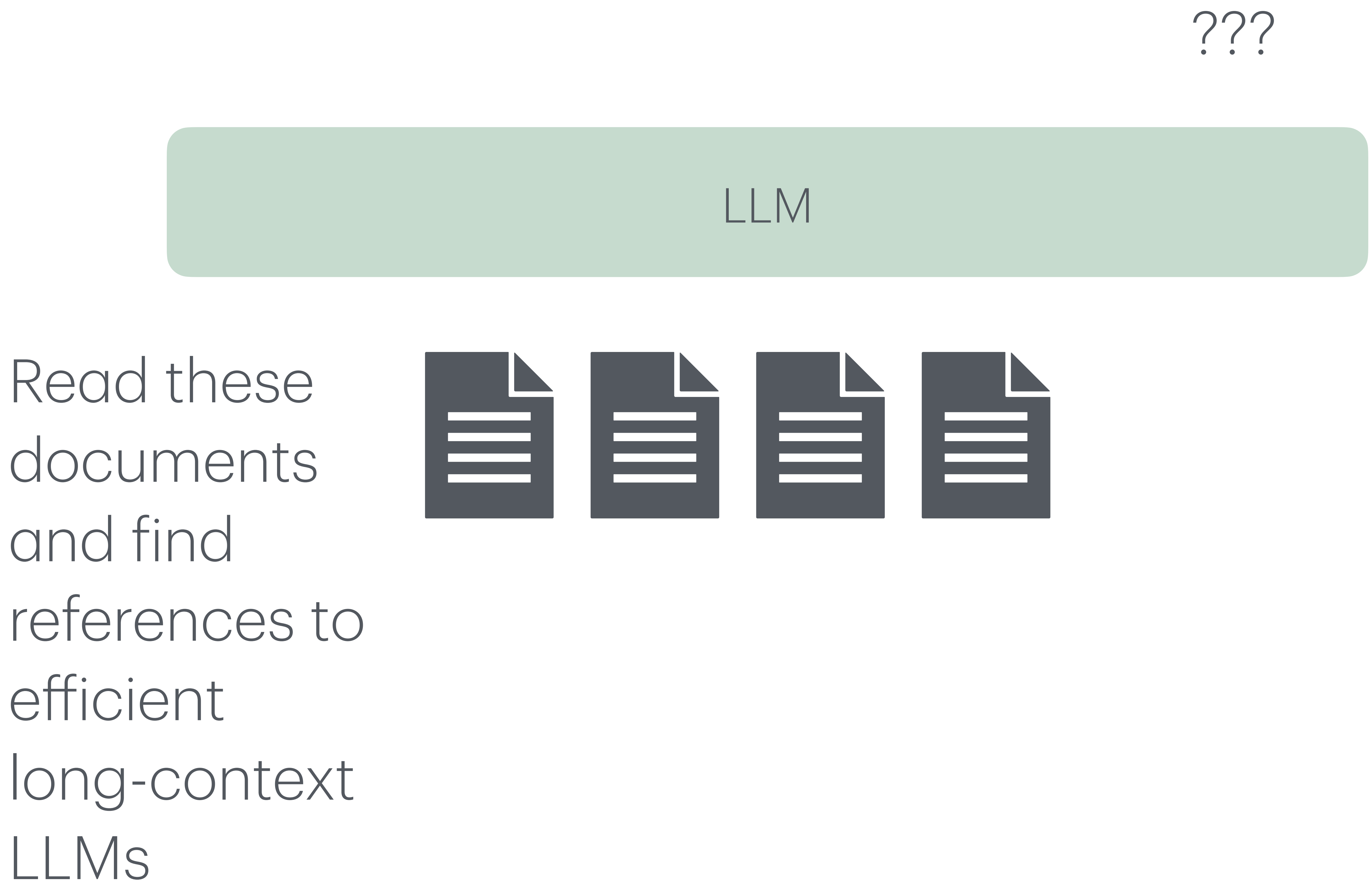
LLM

Read these documents and find references to efficient long-context LLMs

Long Context

What happens if we feed ten’s of thousands of tokens into an LLM?

1. OOM (Out Of Memory)

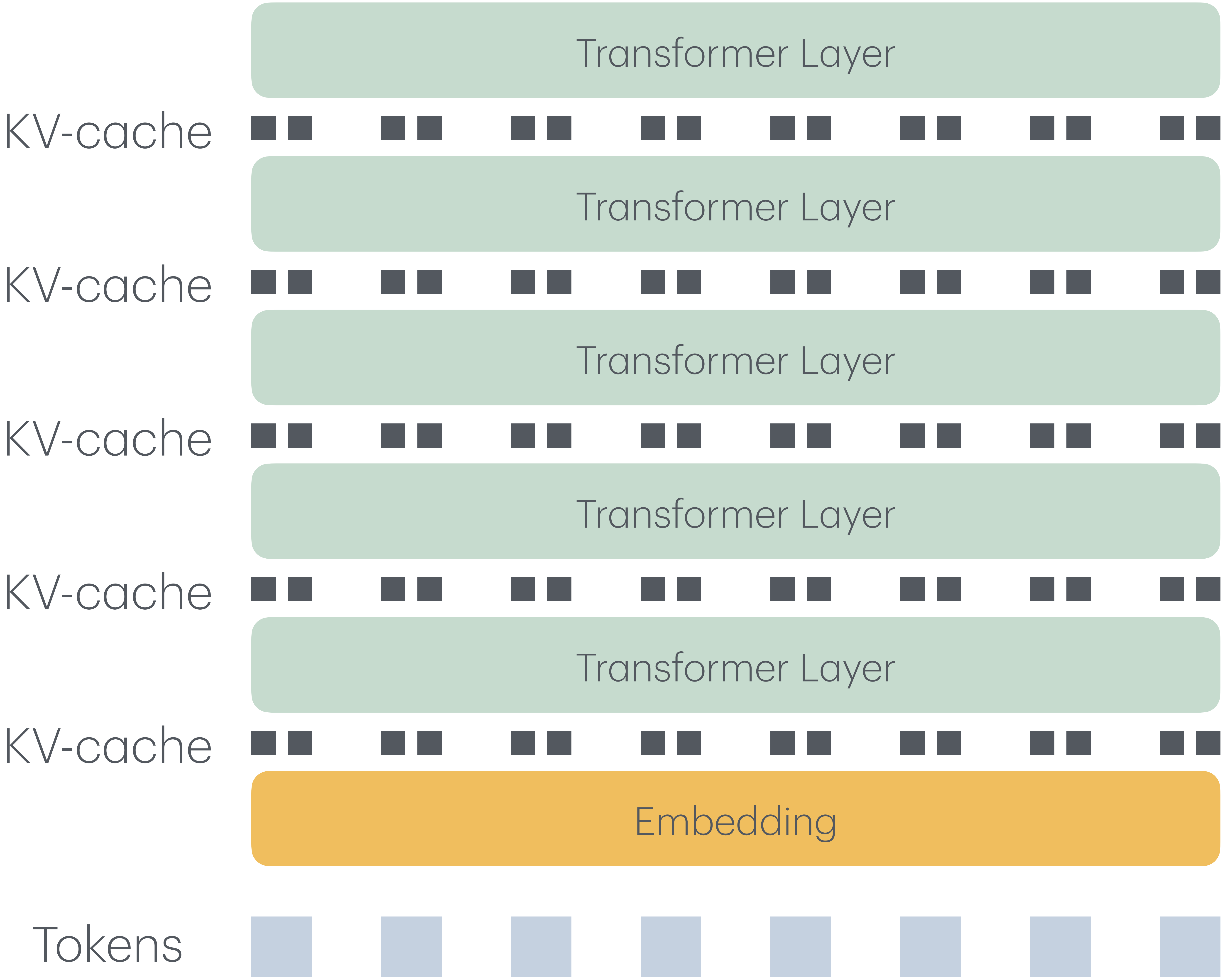


Read these documents and find references to efficient long-context LLMs

Long Context

What happens if we feed ten's of thousands of tokens into an LLM?

1. OOM (Out Of Memory)

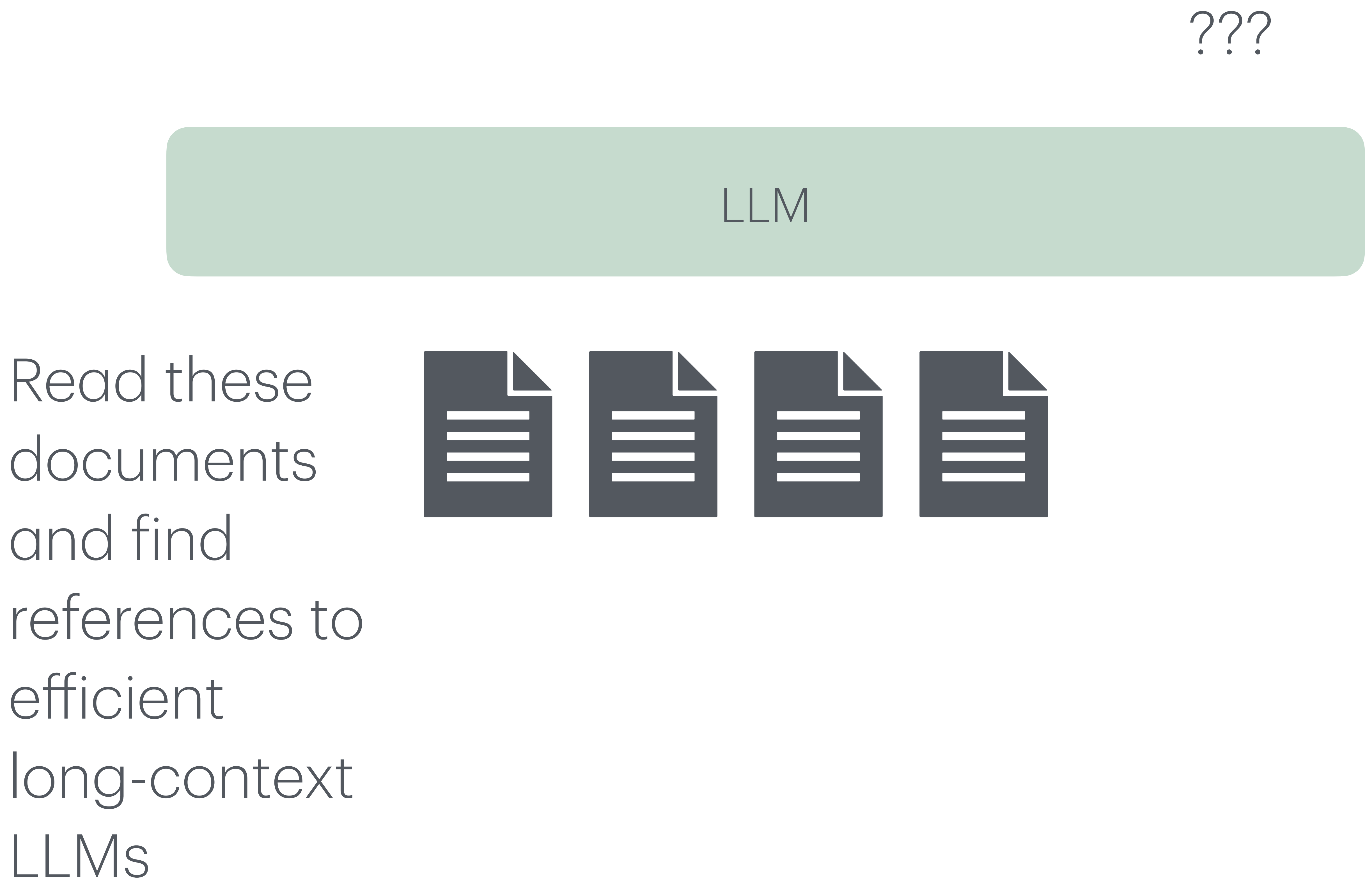


Long Context

What happens if we feed ten's of thousands of tokens into an LLM?

1. OOM (Out Of Memory)

2. Model will be very slow



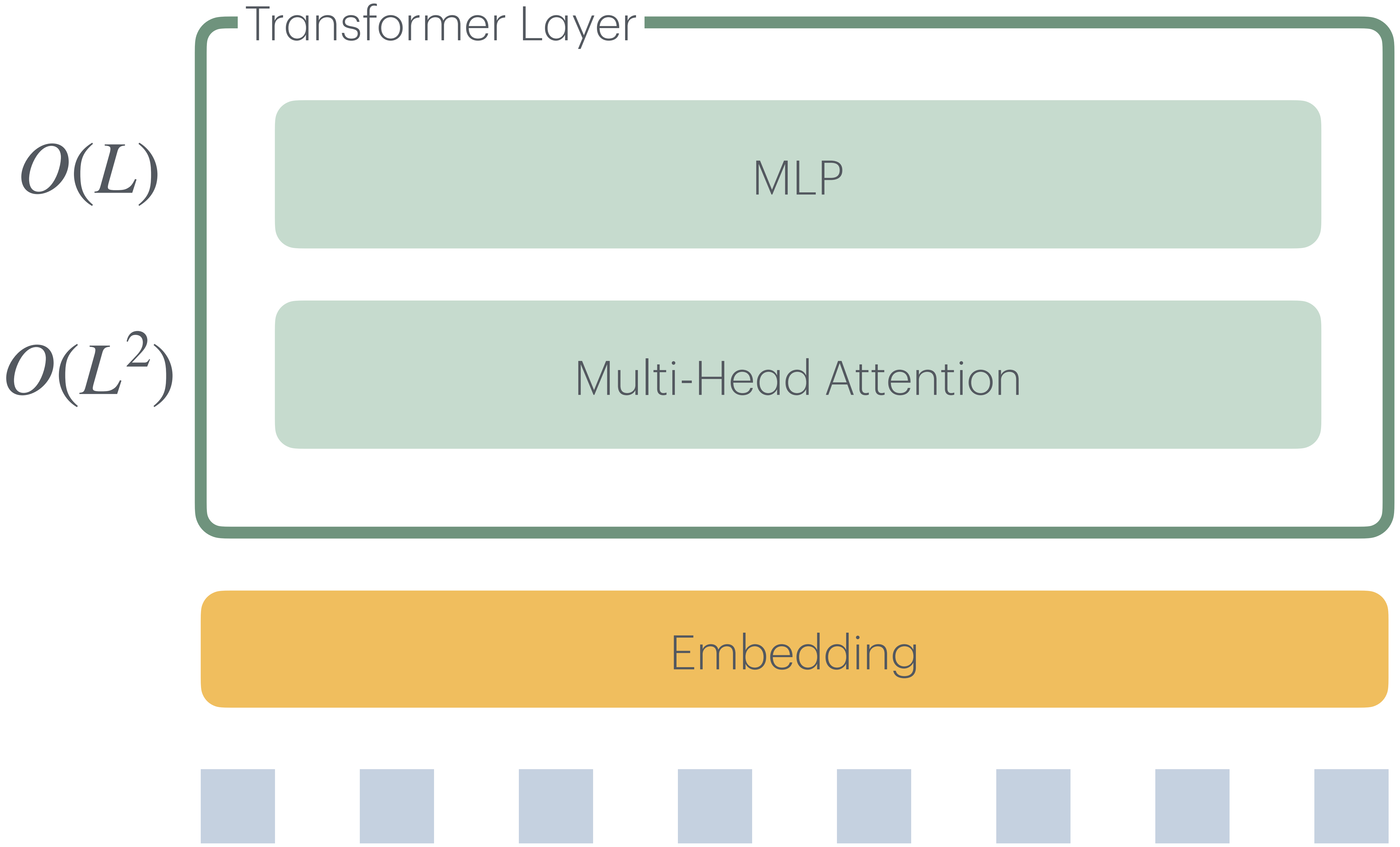
Read these documents and find references to efficient long-context LLMs

Long Context

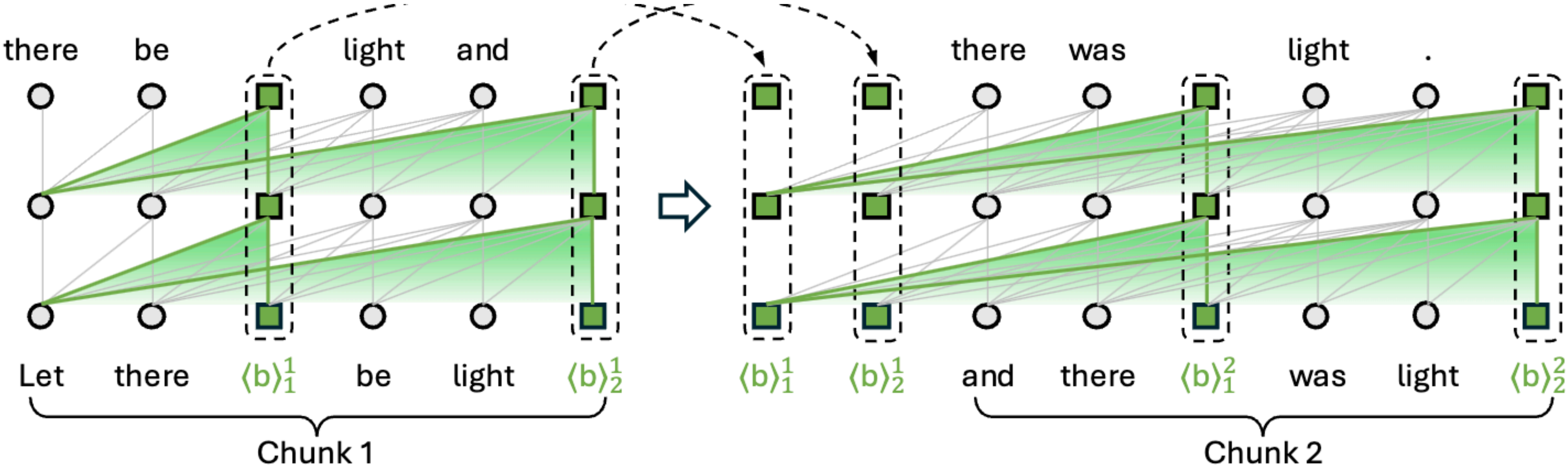
What happens if we feed ten's of thousands of tokens into an LLM?

1. OOM (Out Of Memory)

2. Model will be very slow

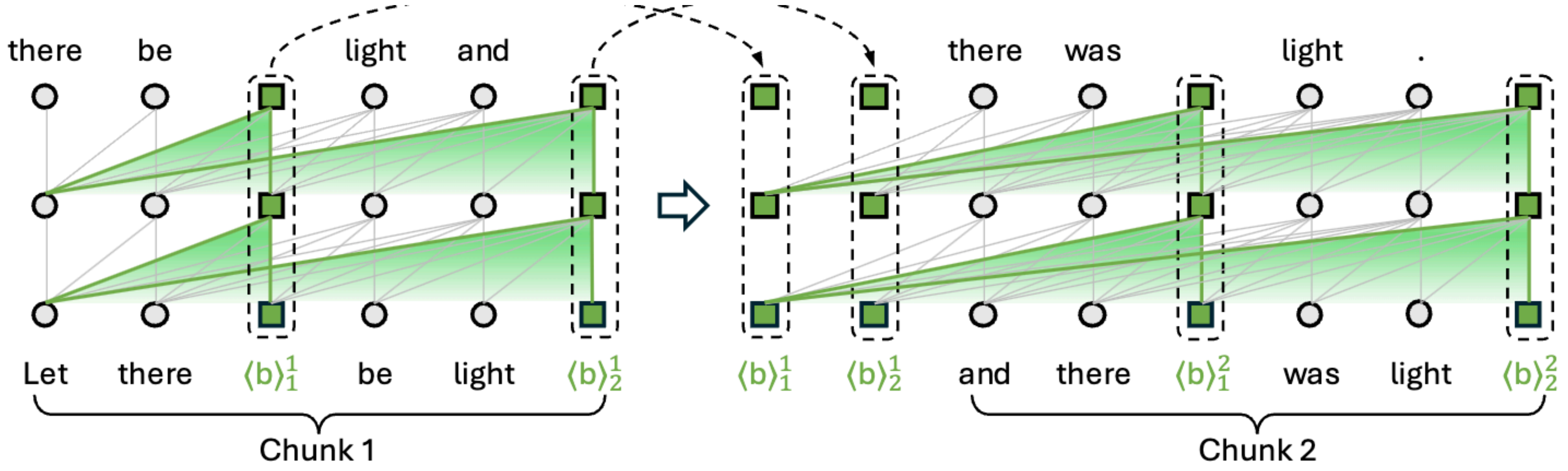


Activation Beacon

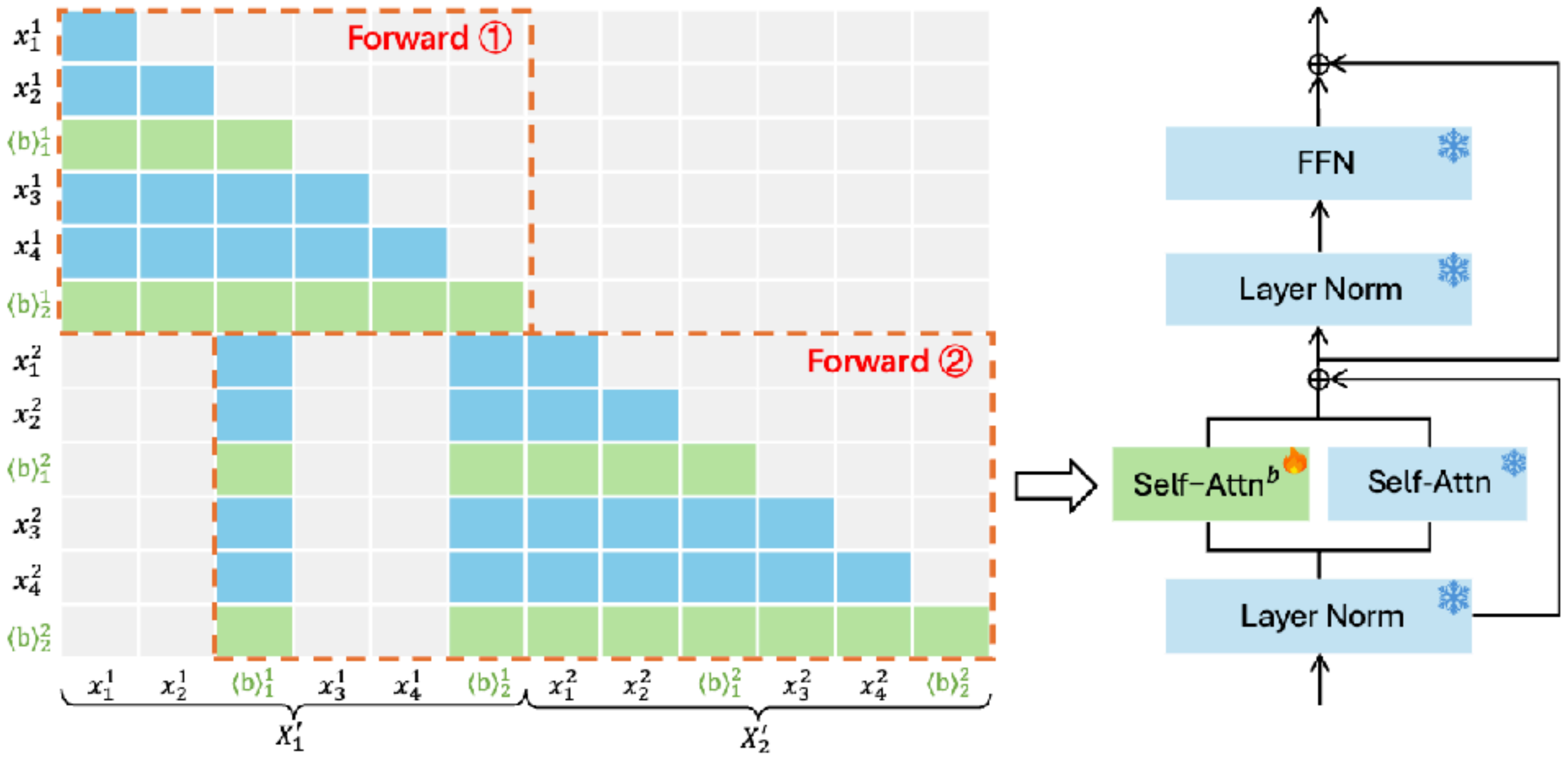


Long Context Compression with Activation Beacon, Zhang etal 2024

Activation Beacon



- Start from pre-trained model
- Partition sequence into chunks of 1024
- Pick k “beacons” per chunk
- Chunk n only sees beacons of chunks 1…n-1
- Fine-tune



Long Context Compression with Activation Beacon, Zhang etal 2024

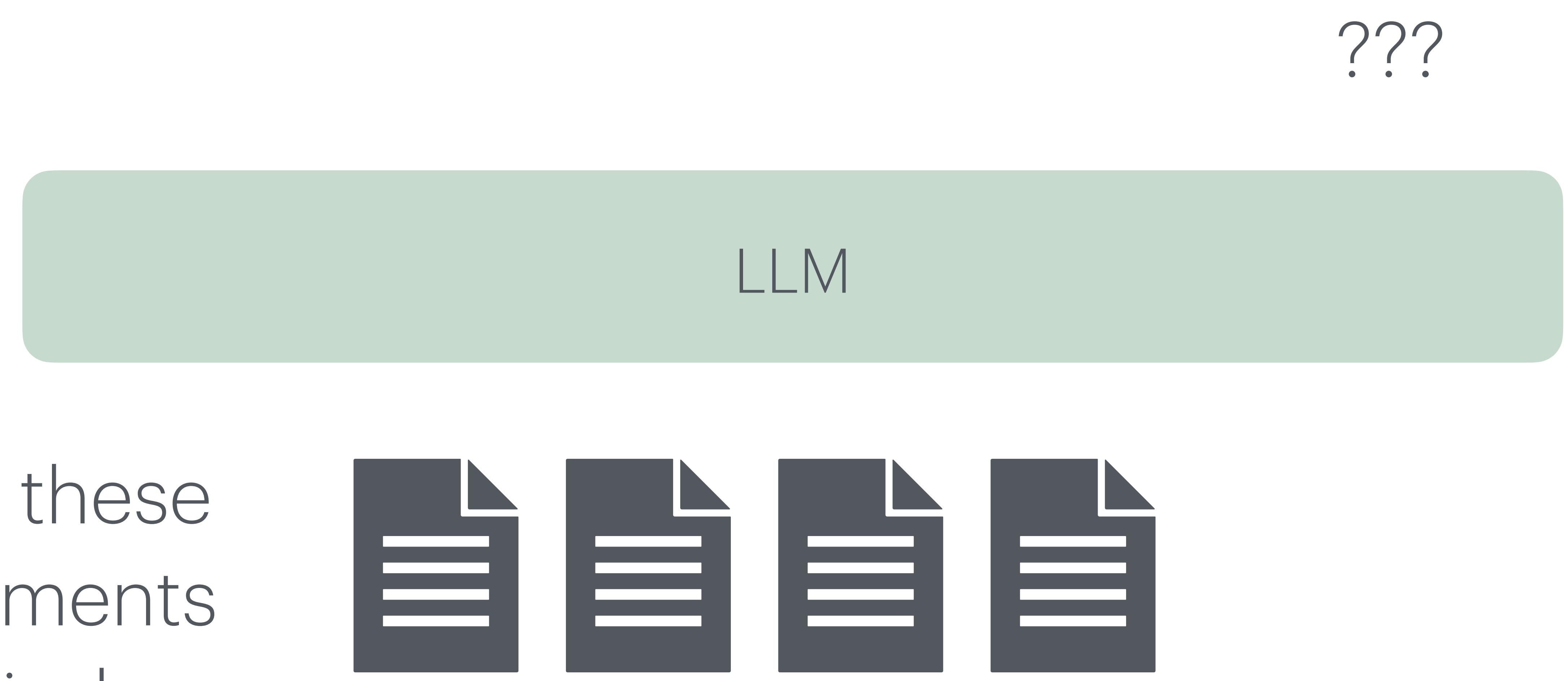
Long Context

What happens if we feed ten's of thousands of tokens into an LLM?

1. ~~OOM (Out Of Memory)~~
2. ~~Model will be very slow~~

Activation Beacons and friends

Read these documents and find references to efficient long-context LLMs

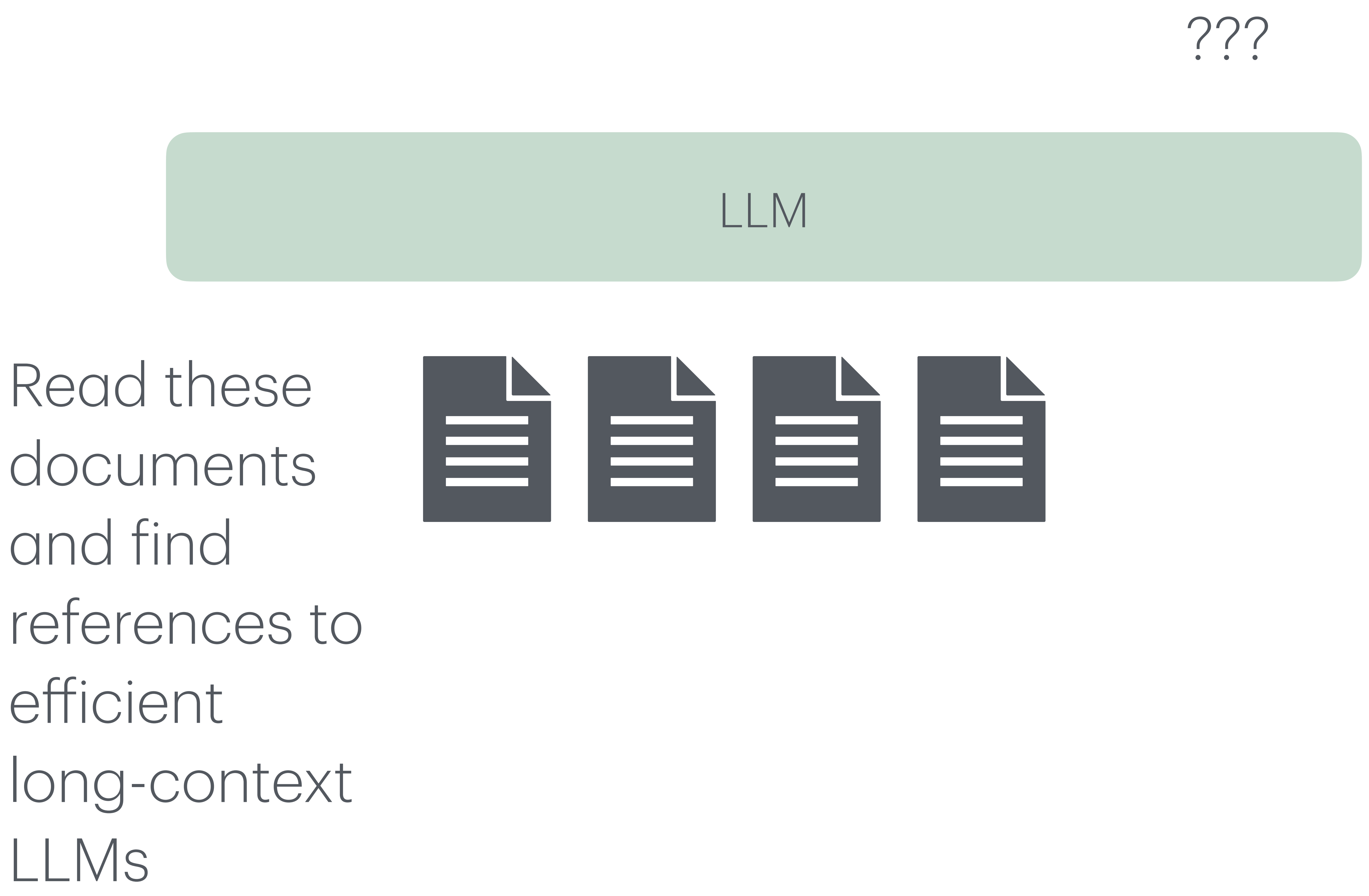


Long Context

What happens if we feed ten's of thousands of tokens into an LLM?

1. ~~OOM (Out Of Memory)~~
2. ~~Model will be very slow~~
3. Model will produce garbage outputs

Activation Beacons and friends



Read these documents and find references to efficient long-context LLMs

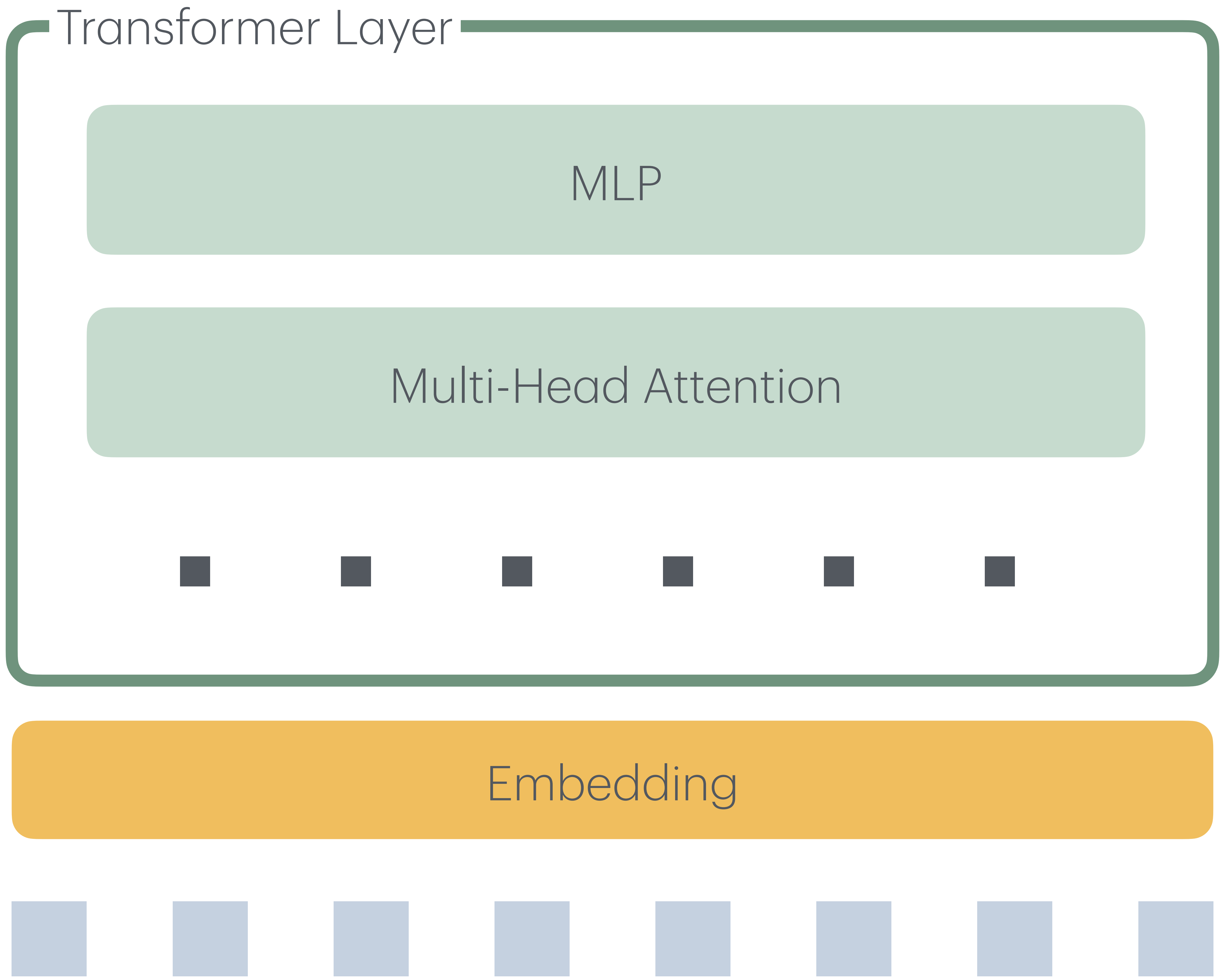
Long Context

What happens if we feed ten’s of thousands of tokens into an LLM?

1. ~~OOM (Out Of Memory)~~
2. ~~Model will be very slow~~
3. Model will produce garbage outputs

Activation Beacons and friends

Positional embedding



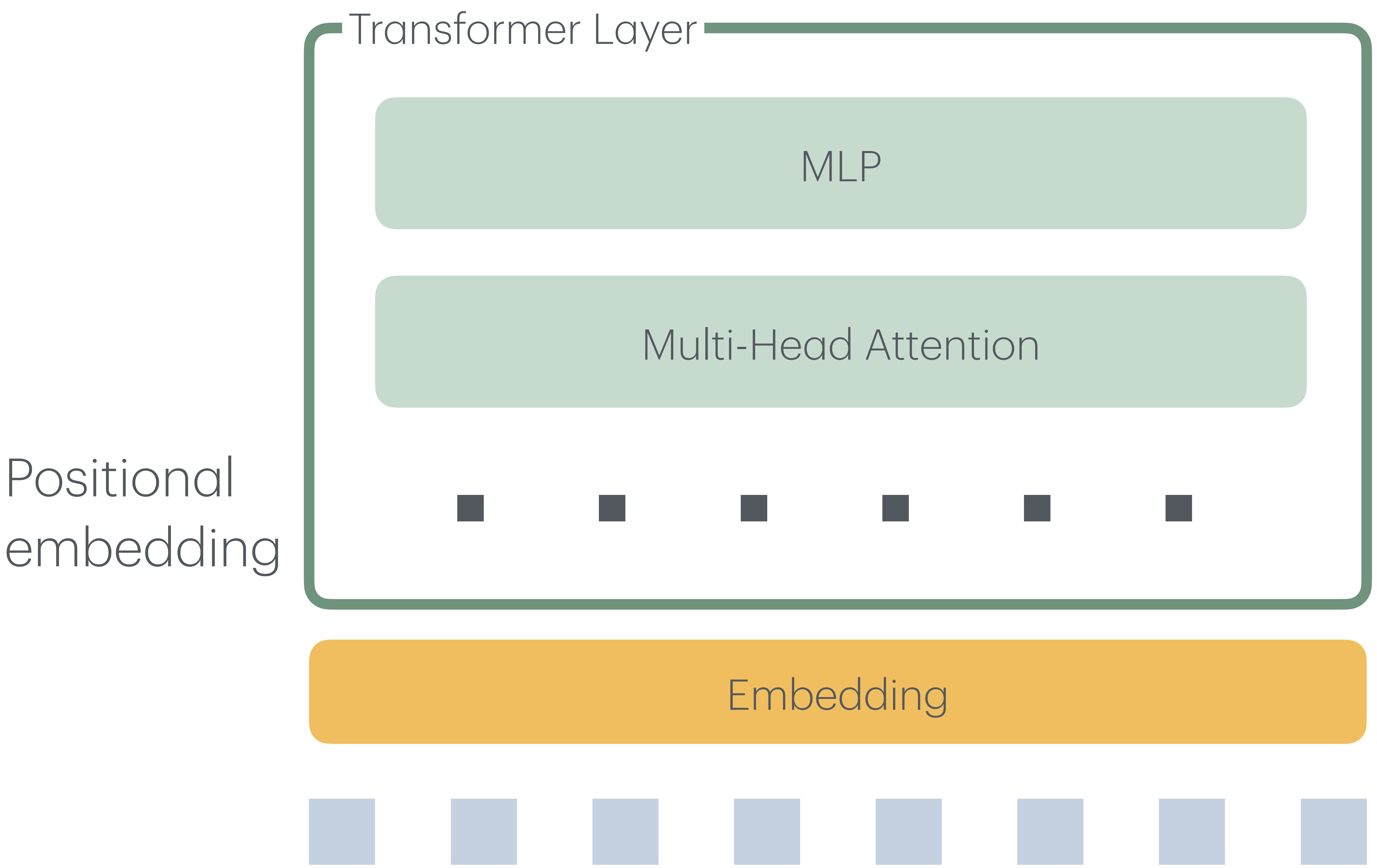
Positional Embedding

- Rotary Embeddings

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_{\{q,k\}} \boldsymbol{x}_m$$

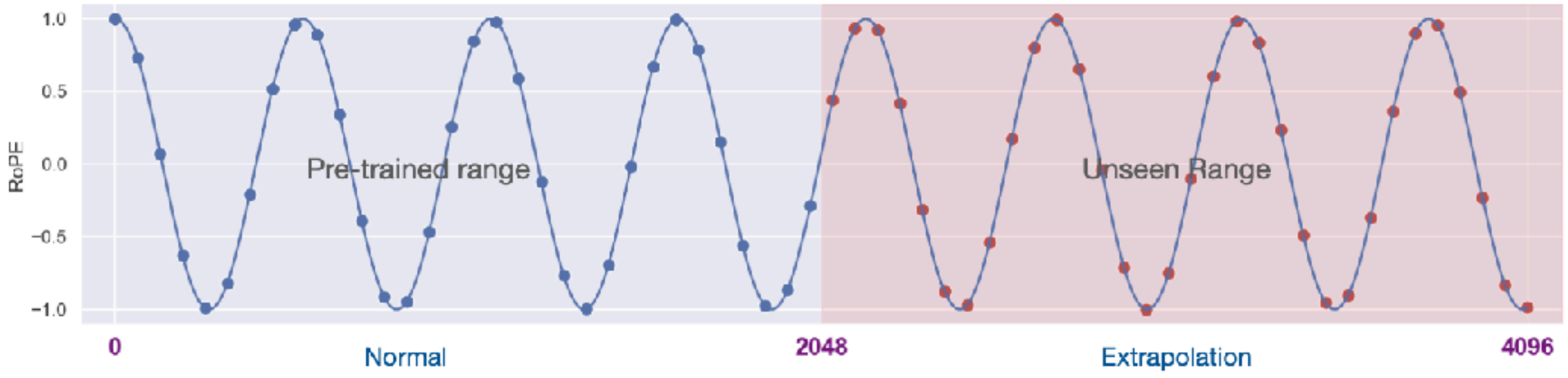
$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots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

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = (\boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_q \boldsymbol{x}_m)^\intercal (\boldsymbol{R}^d_{\Theta,n} \boldsymbol{W}_k \boldsymbol{x}_n) = \boldsymbol{x}^\intercal \boldsymbol{W}_q R^d_{\Theta,n-m} \boldsymbol{W}_k \boldsymbol{x}_n$$



RoFormer: Enhanced Transformer with Rotary Position Embedding, Su etal 2021

Positional Embedding



- Rotary Embeddings
- Fixed context length during training
  - Longer context for inference

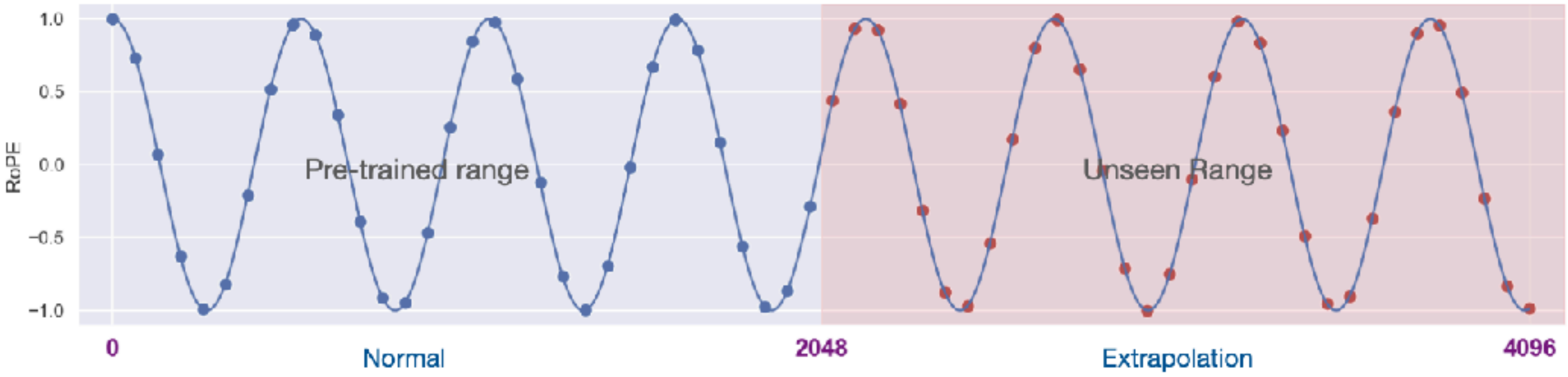
$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_{\{q,k\}} \boldsymbol{x}_m$$

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

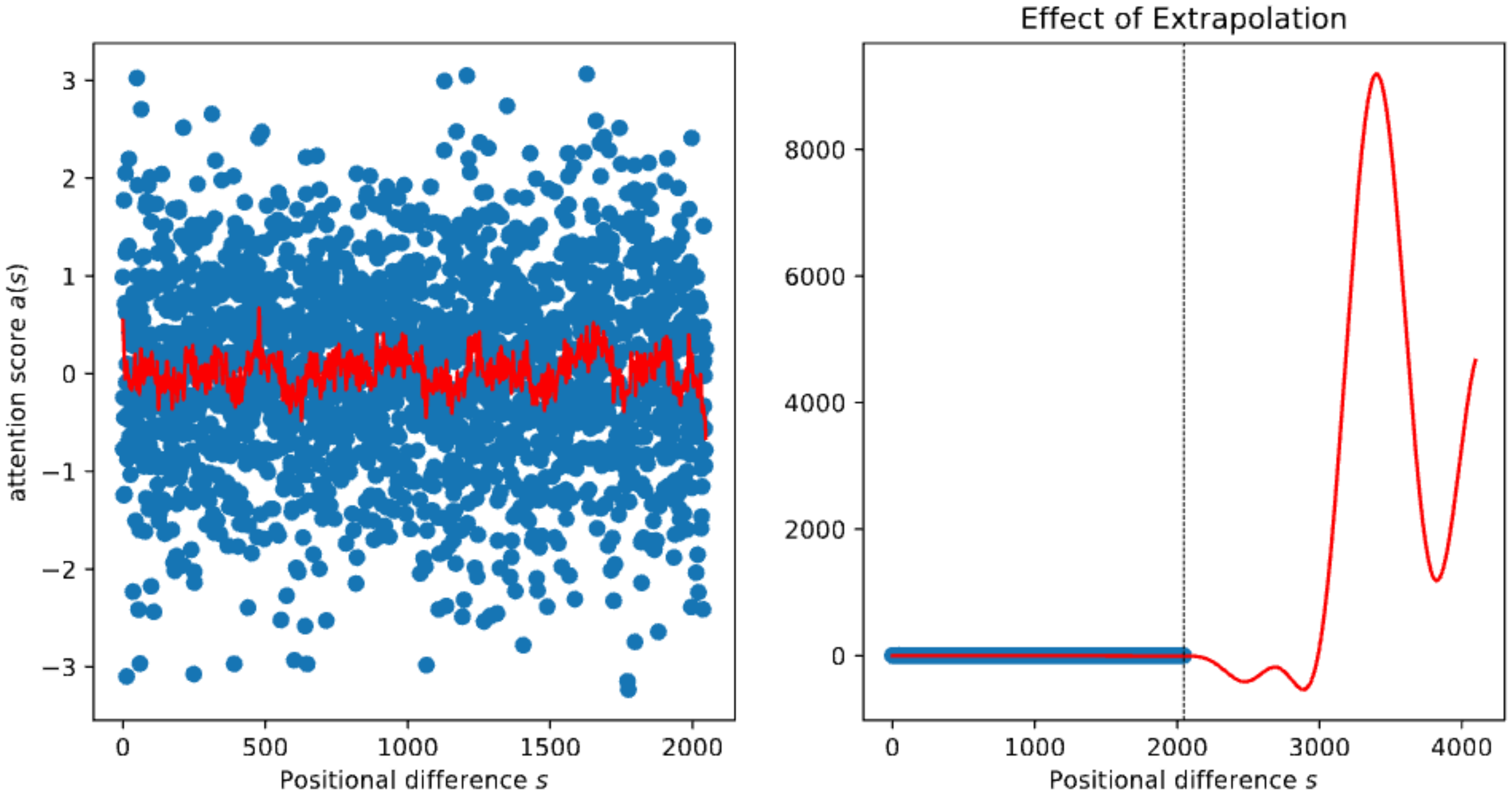
$$\boldsymbol{q}_m^\top \boldsymbol{k}_n = (\boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_q \boldsymbol{x}_m)^\top (\boldsymbol{R}^d_{\Theta,n} \boldsymbol{W}_k \boldsymbol{x}_n) = \boldsymbol{x}^\top \boldsymbol{W}_q R^d_{\Theta,n-m} \boldsymbol{W}_k \boldsymbol{x}_n$$

RoFormer: Enhanced Transformer with Rotary Position Embedding, Su etal 2021
Extending Context Window of Large Language Models via Positional Interpolation, Chen etal 2023

Positional Embedding

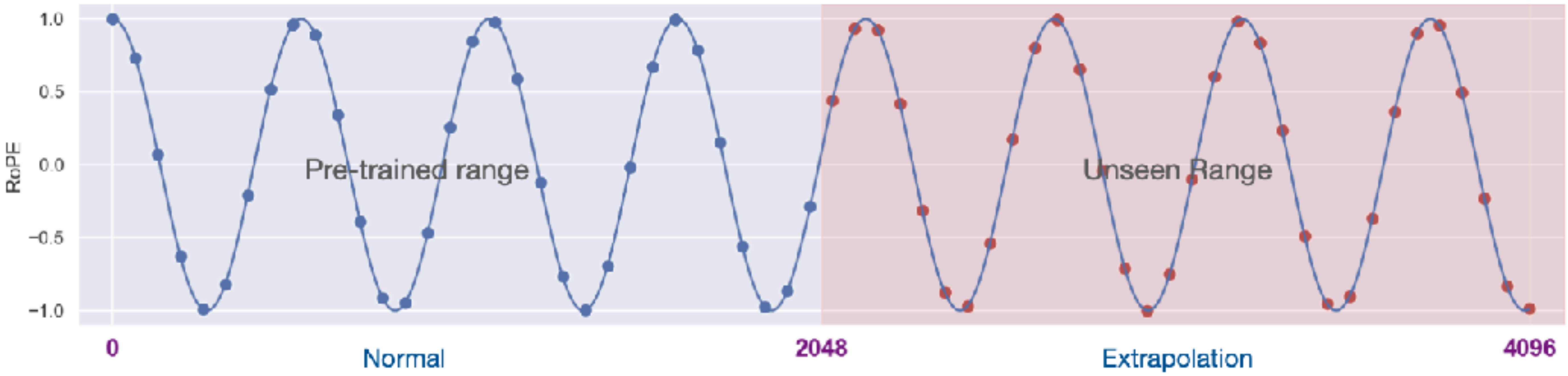


- Rotary Embeddings
  - Do not extrapolate well

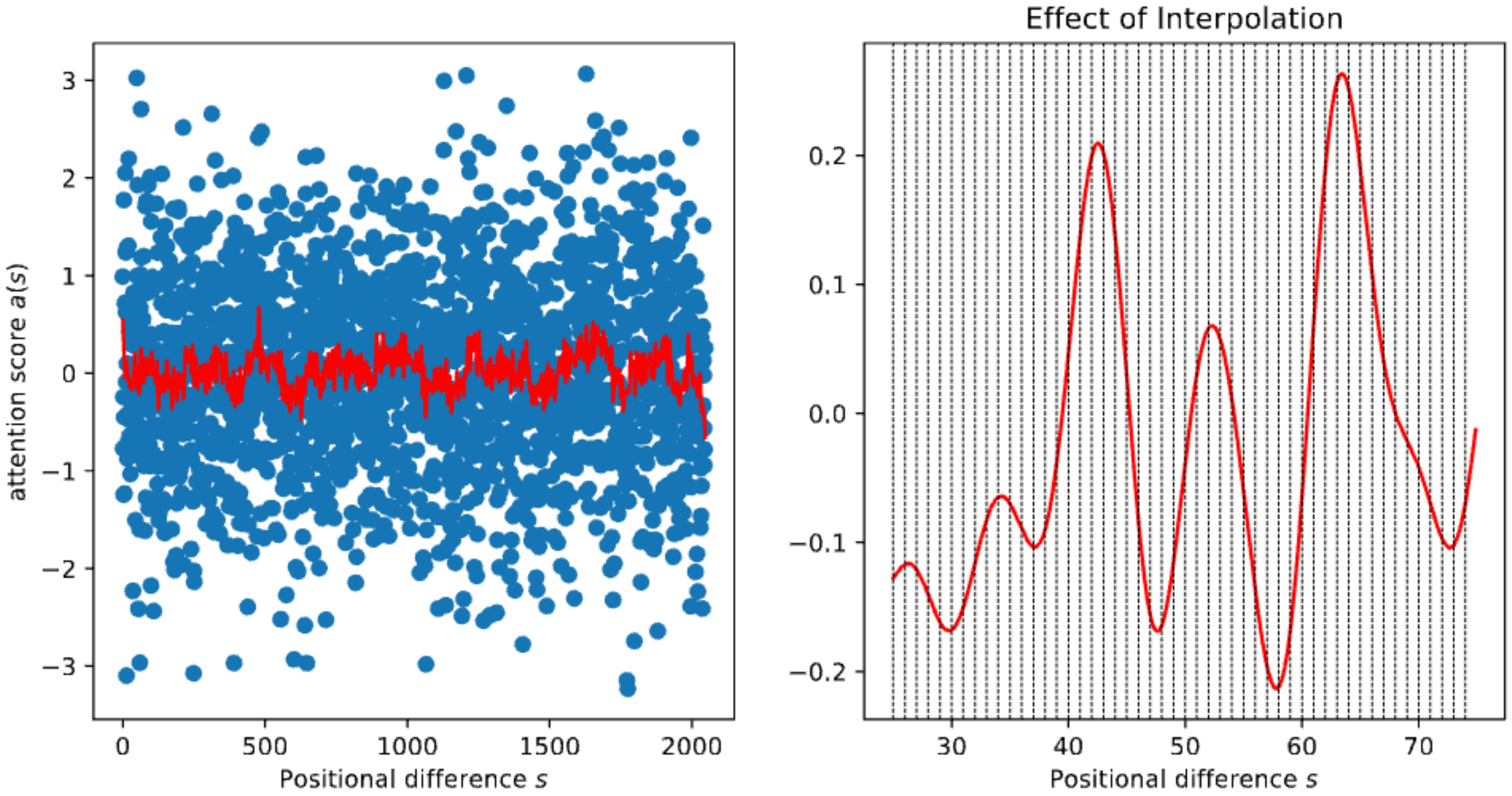


RoFormer: Enhanced Transformer with Rotary Position Embedding, Su etal 2021
Extending Context Window of Large Language Models via Positional Interpolation, Chen etal 2023

Positional Embedding



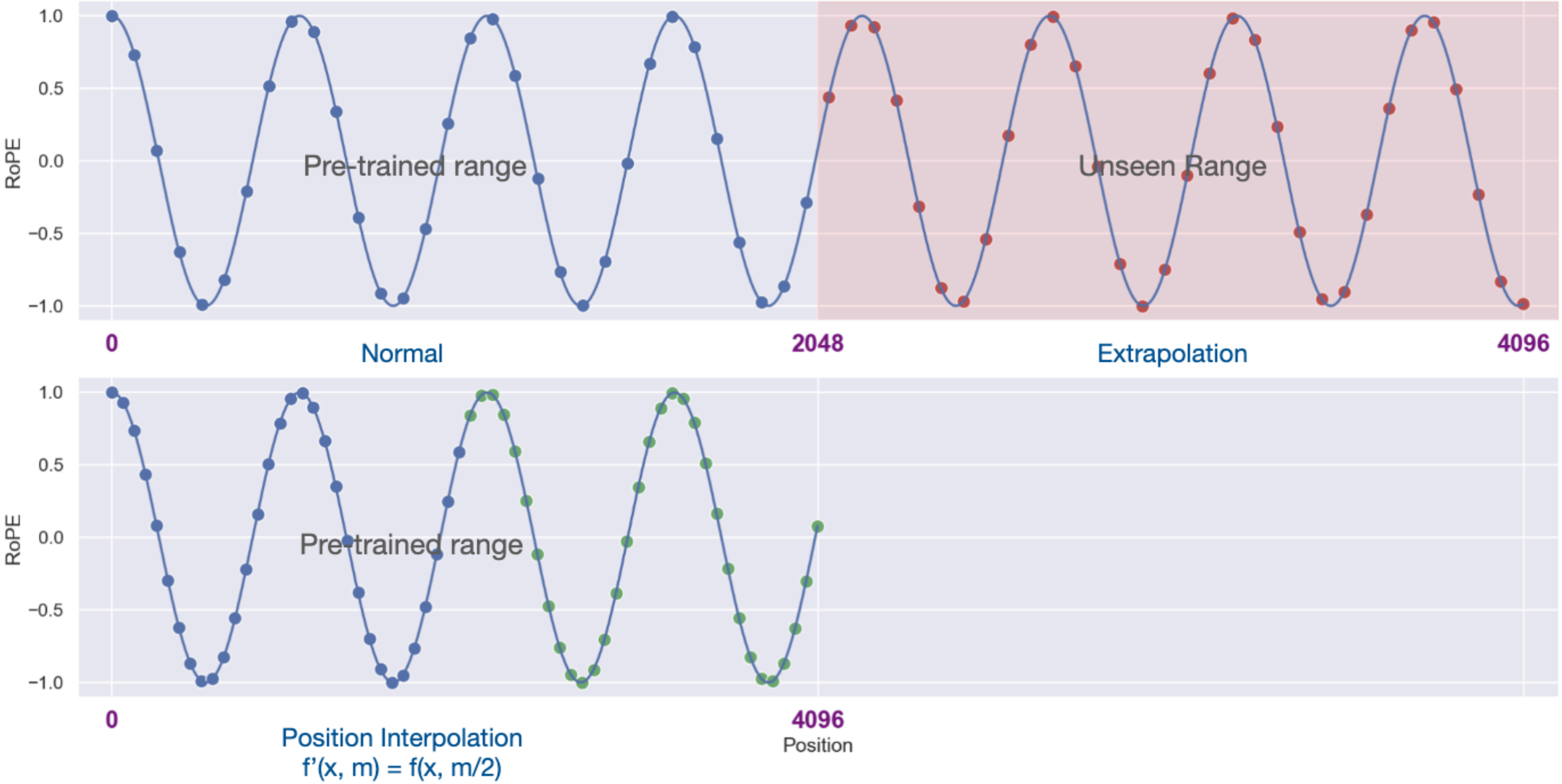
- Rotary Embeddings
  - Do not extrapolate well
  - But they interpolate



RoFormer: Enhanced Transformer with Rotary Position Embedding, Su etal 2021

Extending Context Window of Large Language Models via Positional Interpolation, Chen etal 2023

RoPE Scaling

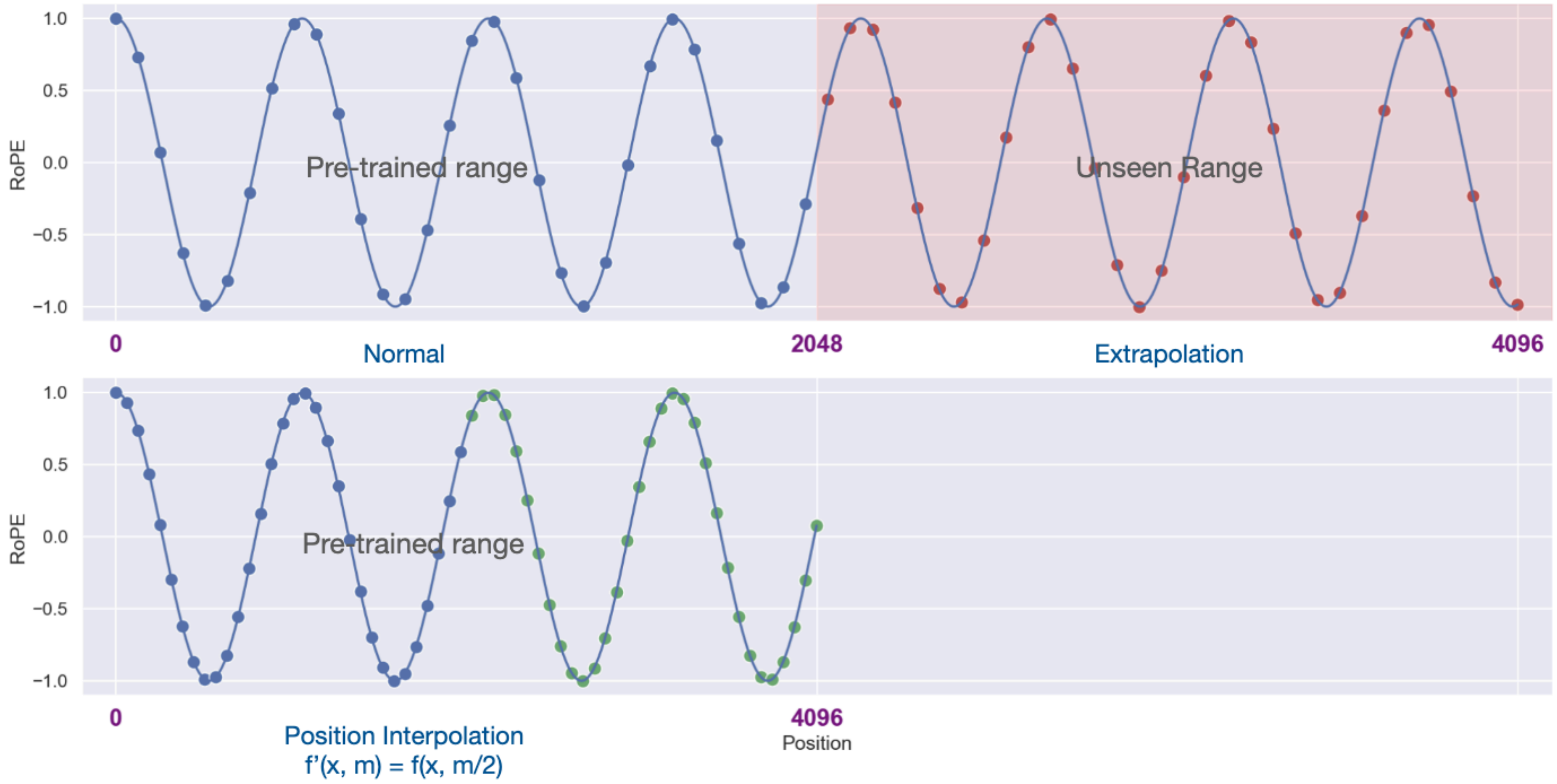


RoFormer: Enhanced Transformer with Rotary Position Embedding, Su etal 2021

Extending Context Window of Large Language Models via Positional Interpolation, Chen etal 2023

RoPE Scaling

- Extrapolation
  - Make token stream **longer**
  - Does not generalize
- RoPE Scaling
  - Make token stream **denser**
  - Model generalizes
- Widely used



RoFormer: Enhanced Transformer with Rotary Position Embedding, Su etal 2021
Extending Context Window of Large Language Models via Positional Interpolation, Chen etal 2023

Long Context

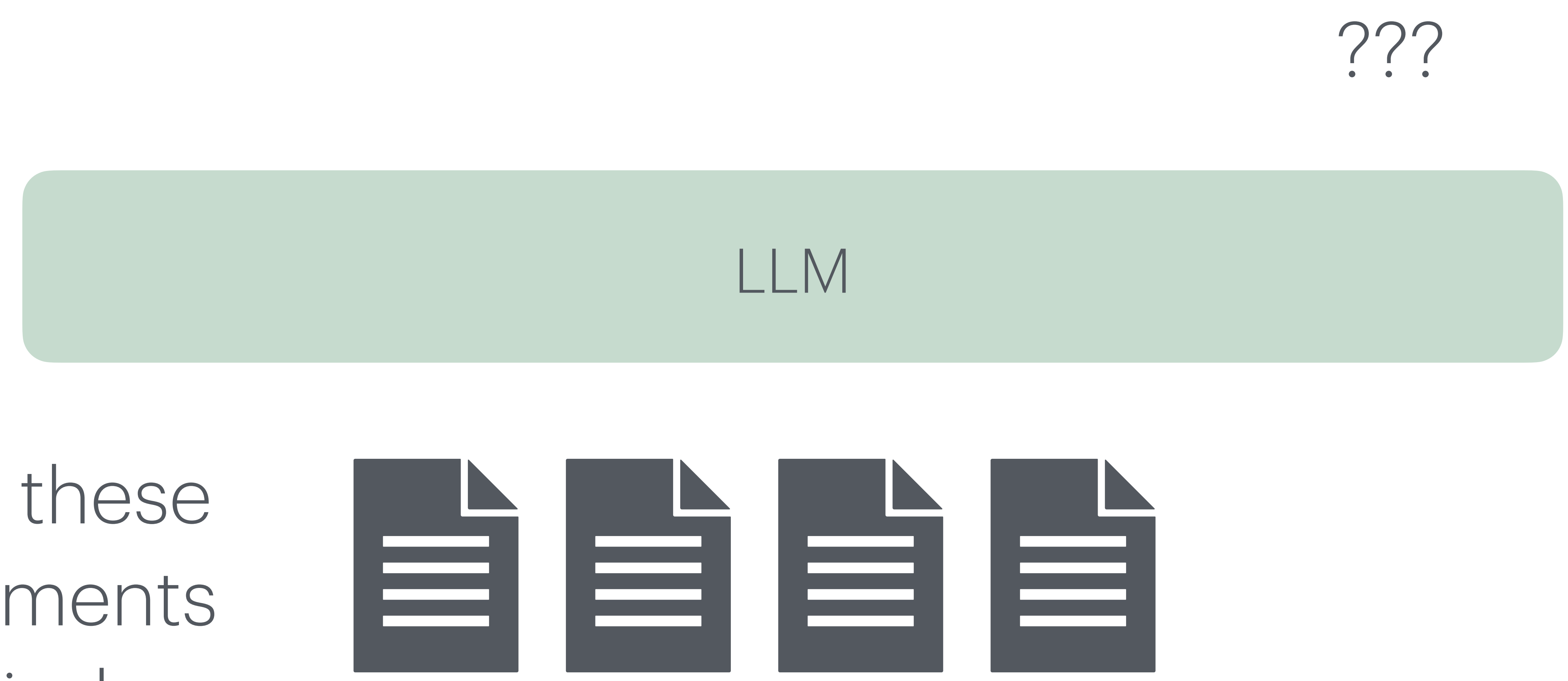
What happens if we feed ten’s of thousands of tokens into an LLM?

1. ~~OOM (Out Of Memory)~~
2. ~~Model will be very slow~~
3. ~~Model will produce garbage outputs~~

Activation Beacons and friends

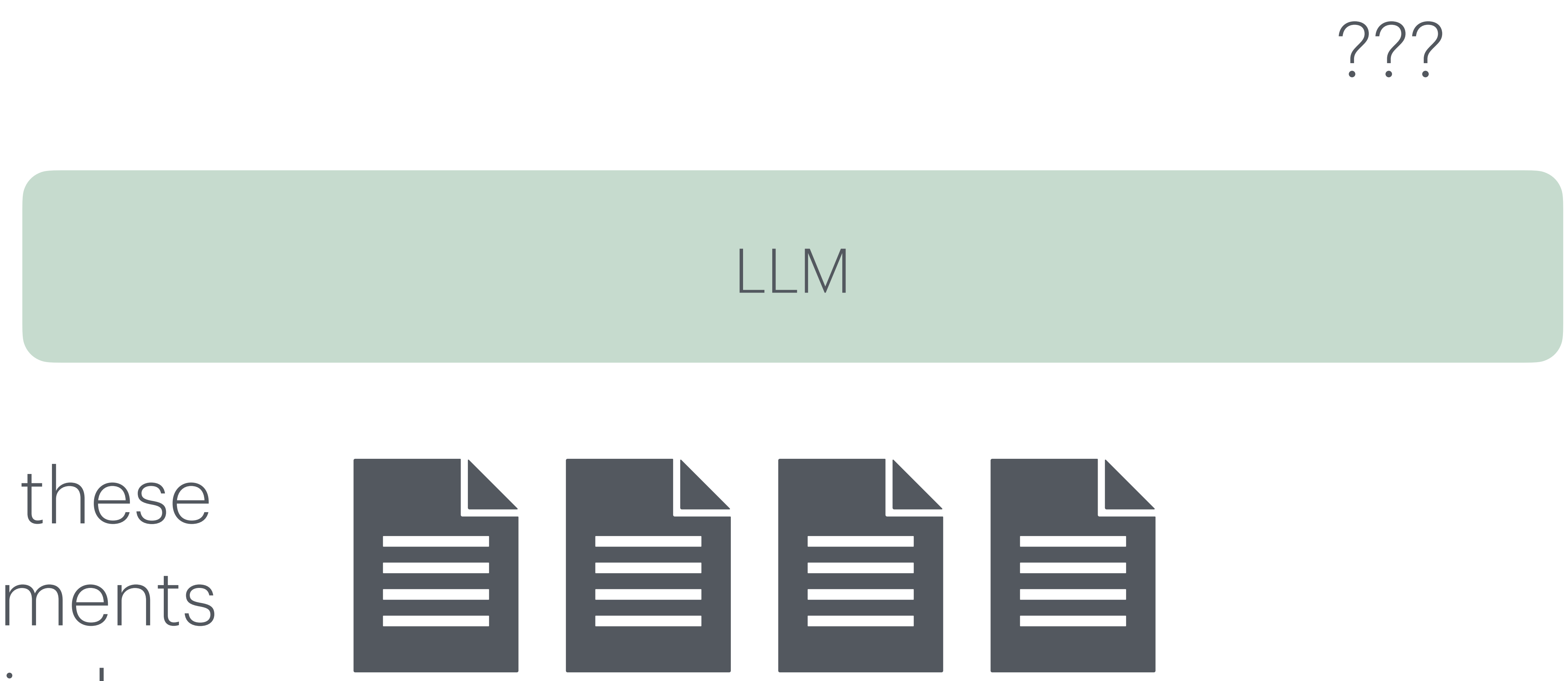
RoPE scaling

Read these documents and find references to efficient long-context LLMs



Long Context

- Current model are **pre-trained** on **2-8k** token sequences
- Late stage pre-training **8k-128k**
  - RoPE Scaling
- Fine-tuned on variable length sequences



Read these documents and find references to efficient long-context LLMs

References

- [1] Long Context Compression with Activation Beacon, Zhang et al. 2024 (link)
- [2] RoFormer: Enhanced Transformer with Rotary Position Embedding, Su etal 2021 (link)
- [3] Extending Context Window of Large Language Models via Positional Interpolation, Chen et al 2023 (link)