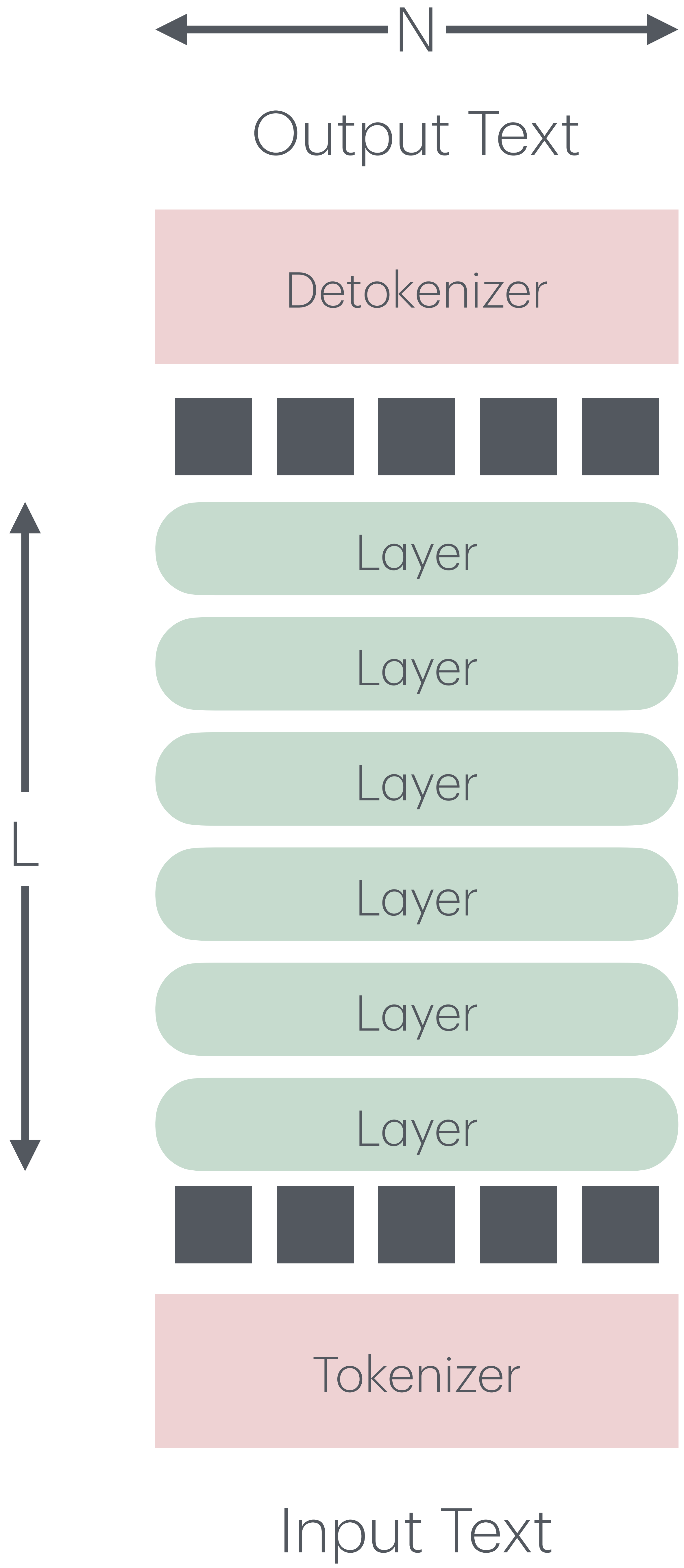
Page Attention

Philipp Krähenbühl, UT Austin

Training and Generation

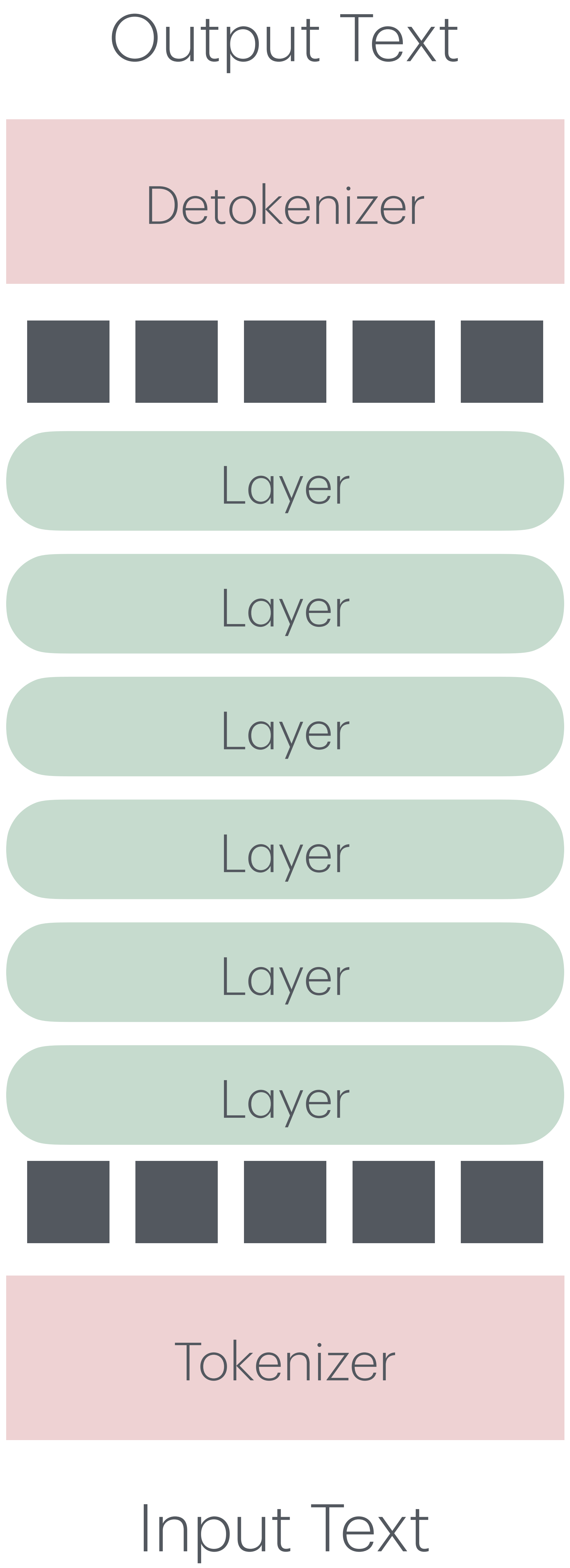
Vanilla Generation

	Training	Training - Checkpointing	Generation
Peak Memory	O(NL)	O(NL½)	O(N)
Runtime	$O(N^2L)$	$O(2\ N^2L)$	$O(N^3L)$
# forward calls	1	1	N



Generation

- Step 1: Tokenize
- Step 2: $N \times$ Forward
- Step 3: Detokenize



Output Text

Detokenizer

Layer

Layer

Layer

Layer

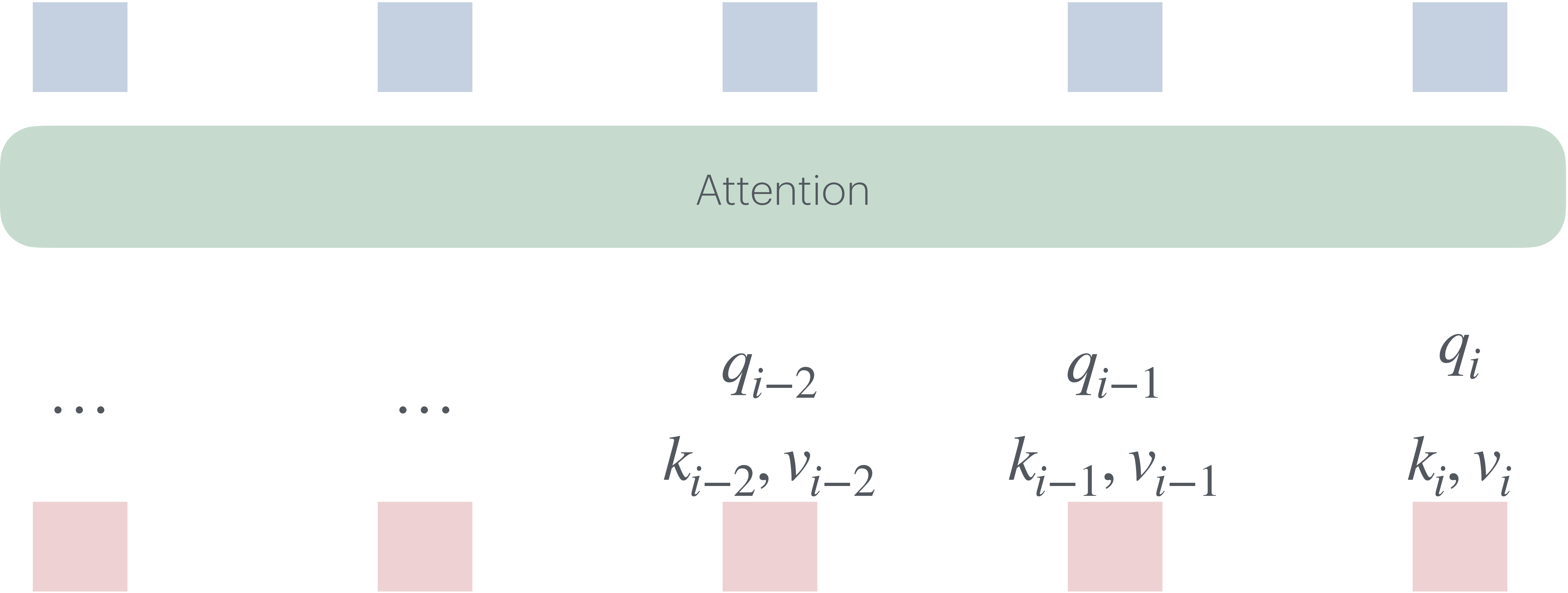
Layer

Layer

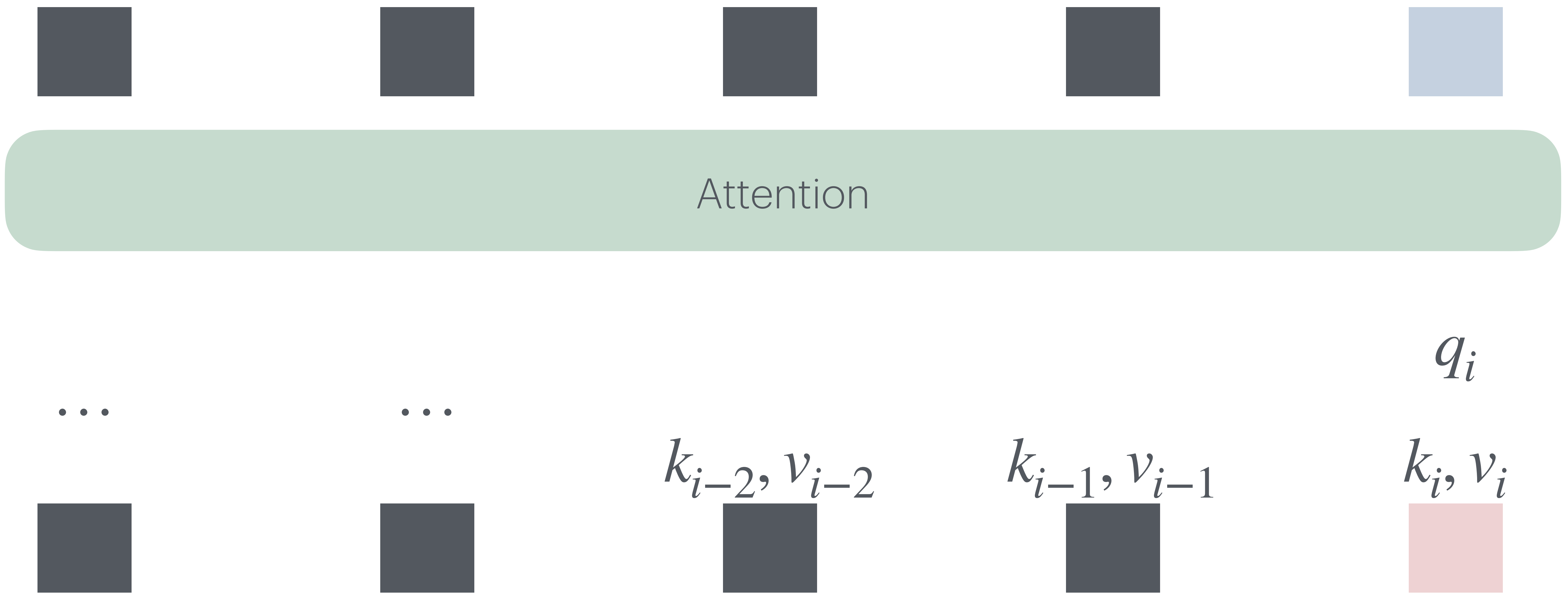
Tokenizer

Input Text

Vanilla Attention

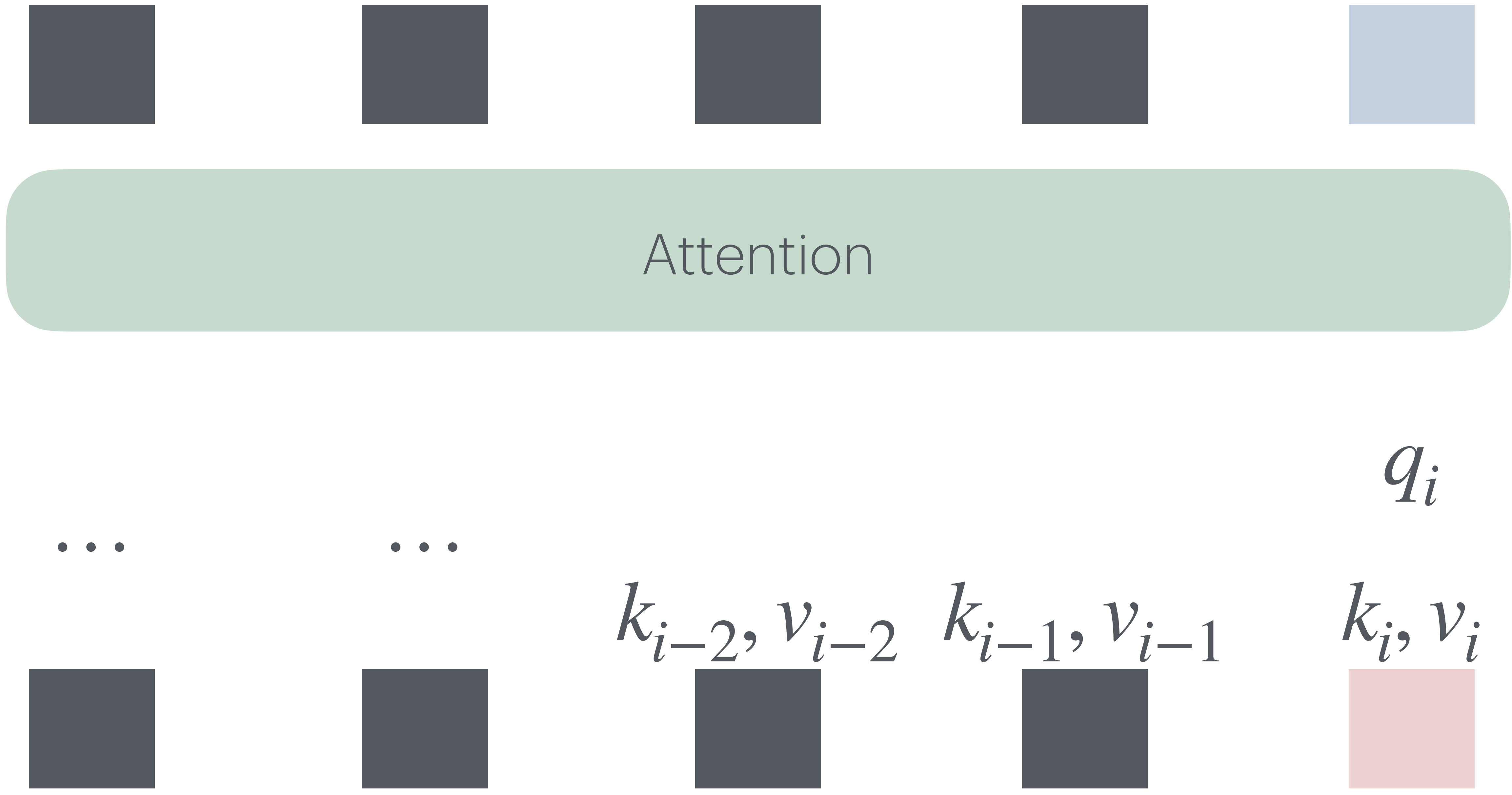


Paged Attention



Efficient Memory Management for Large Language Model Serving with PagedAttention, Kwon etal 2023

Paged Attention - Analysis



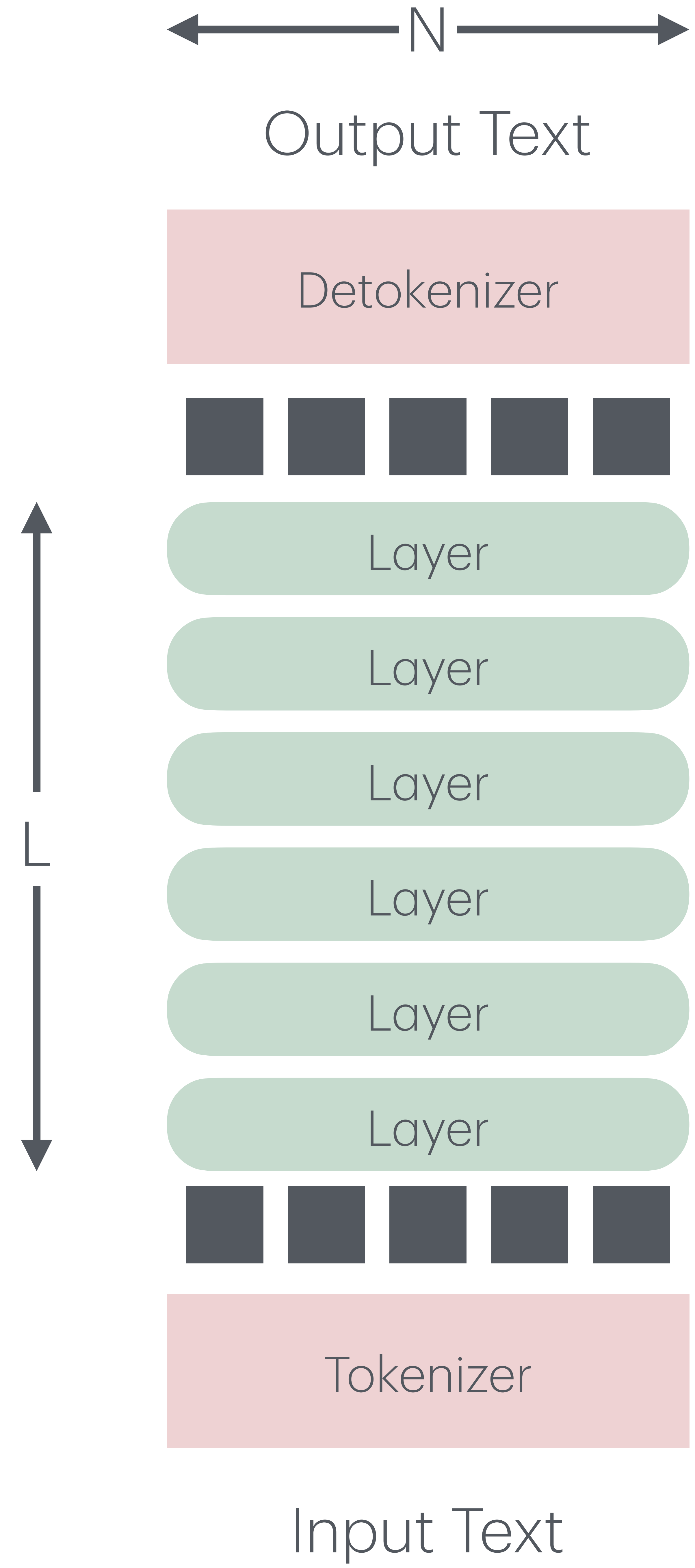
- Cache keys and value
  - $O(NL)$ memory
- Less computation
  - $O(NL)$ per output token

Efficient Memory Management for Large Language Model Serving with PagedAttention, Kwon etal 2023

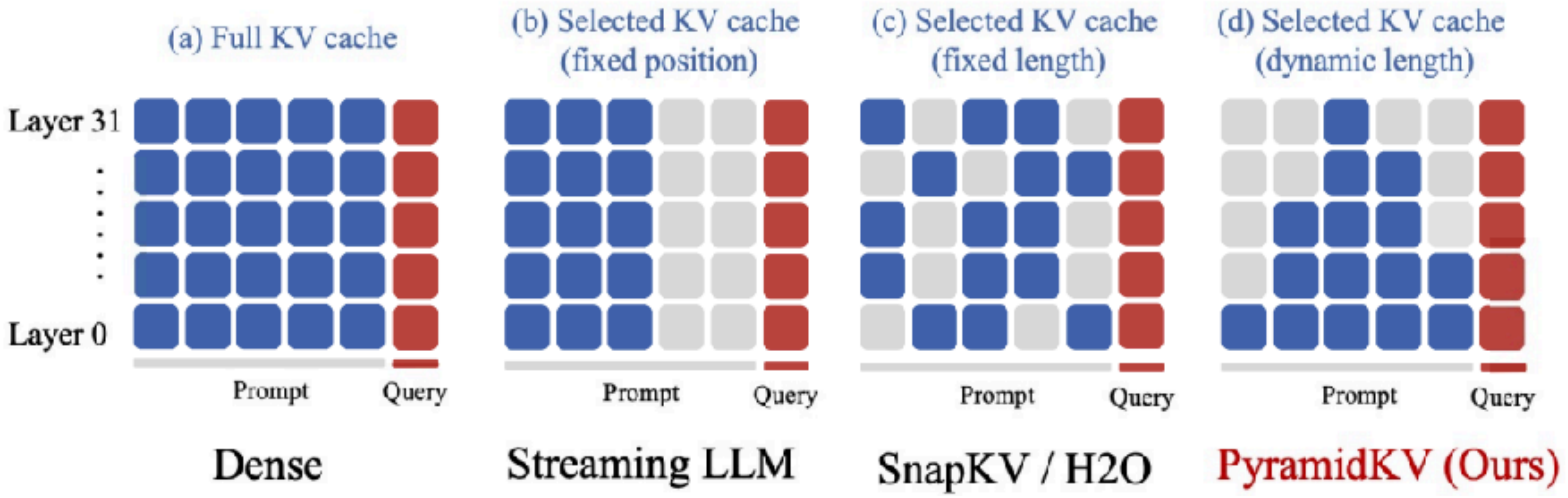
Training and Generation

Paged Attention

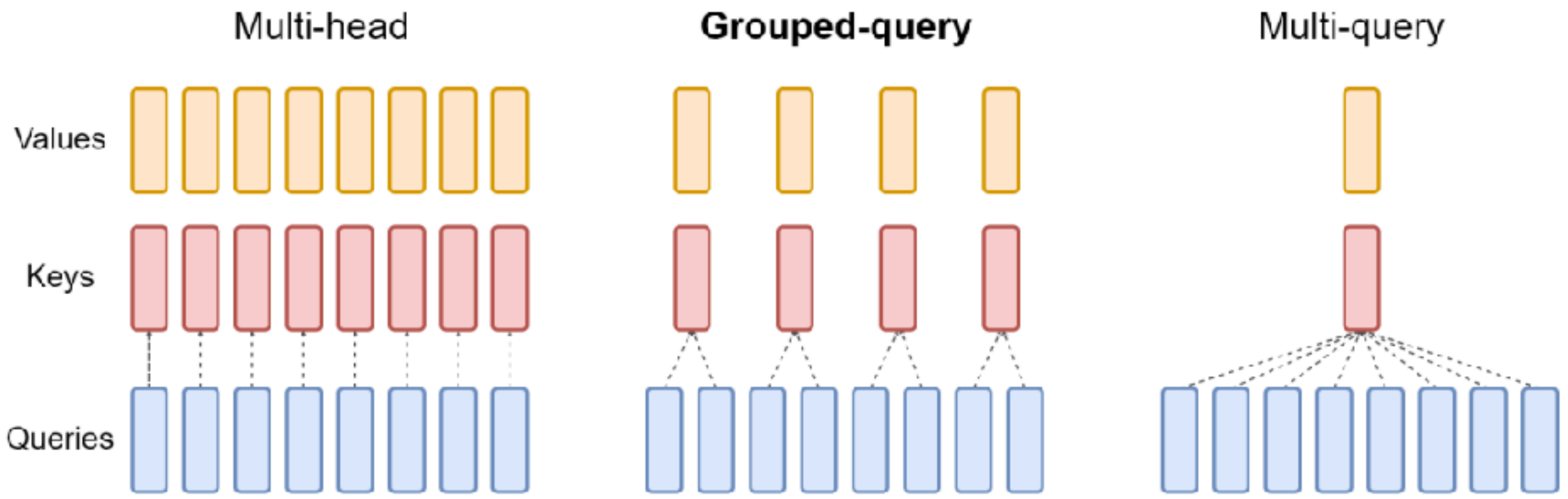
	Training	Training - Checkpointing	Generation	Paged Attention
Peak Memory	O(NL)	O(NL½)	O(N)	O(NL)
Runtime	$O(N^2L)$	$O(2\ N^2L)$	$O(N^3L)$	$O(N^2L)$
# forward calls	1	1	N	N



Open Problem


(a) Full KV cache
(b) Selected KV cache (fixed position)
(c) Selected KV cache (fixed length)
(d) Selected KV cache (dynamic length)
Layer 31
Layer 0
Prompt
Query
Dense
Streaming LLM
SnapKV / H2O
PyramidKV (Ours)

- A more efficient KV-Cache
  - Group Query Attention
  - Pruning
  - Low-rank representations?
- Connection to state-space models



PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling, Cai etal 2024
GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints, Ainslie etal 2023

References

- [1] Efficient Memory Management for Large Language Model Serving with PagedAttention, Kwon et al 2023. (link)

- [2] PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling, Cai et al 2024. (link)

- [3] GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints, Ainslie et al 2023. (link)