Keyformer: KV Cache reduction through key tokens selection for Efficient Generative Inference

Muhammad Adnan, Akhil Arunkumar, Gaurav Jain, Prashant J. Nair,
Ilya Soloveychik, Purushotham Kamath

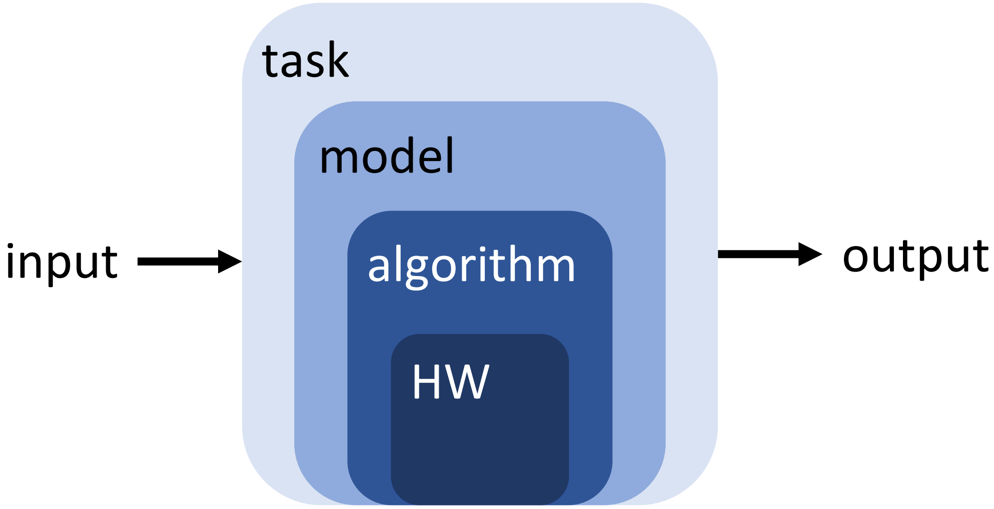
MLSys, 2024

THE UNIVERSITY OF BRITISH COLUMBIA

d-Matrix

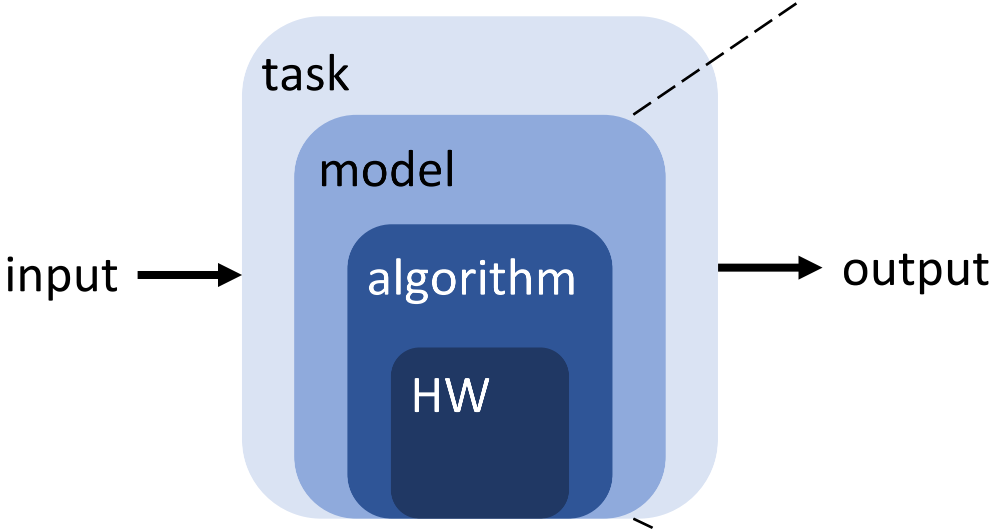
Background

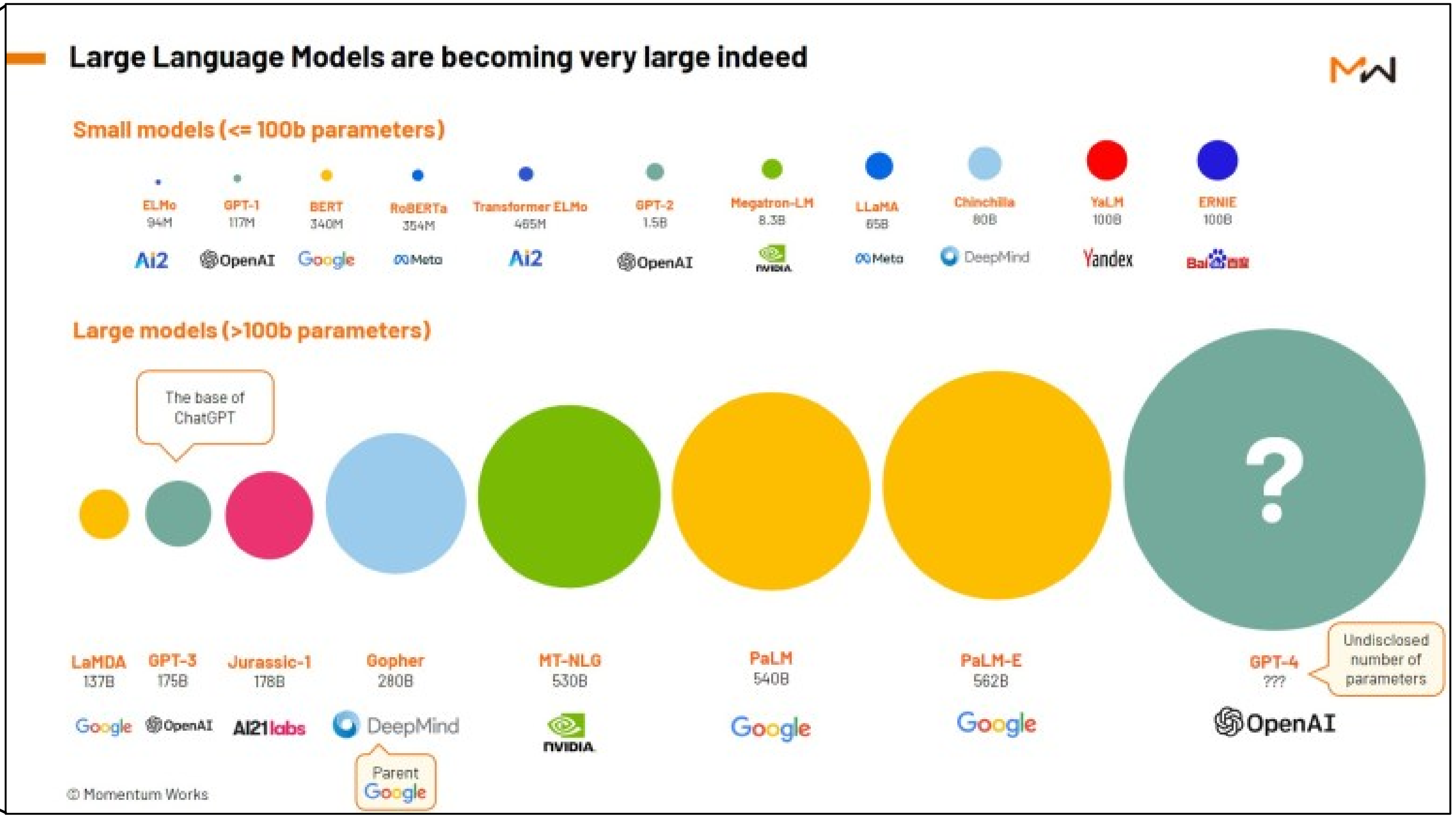
Large Language Models (LLMs) abstraction



Background

Models are getting bigger

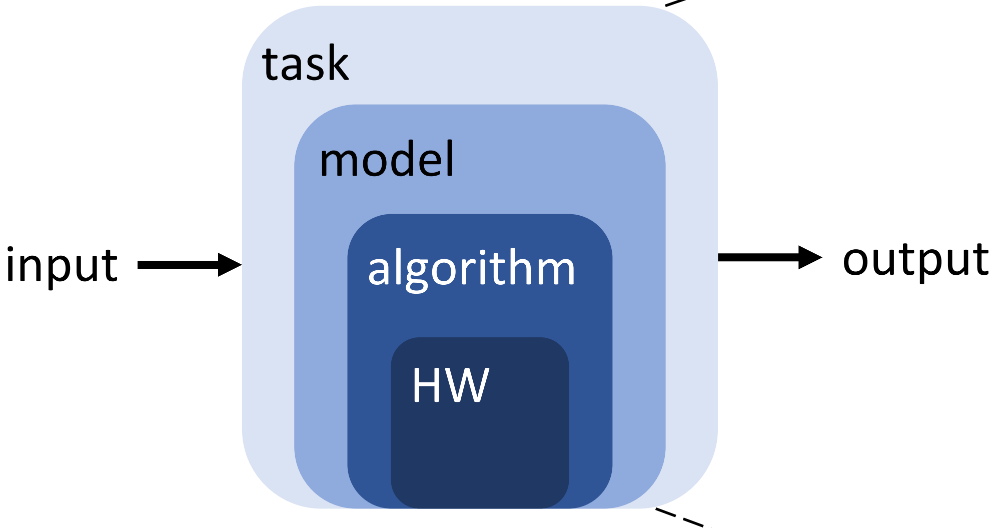


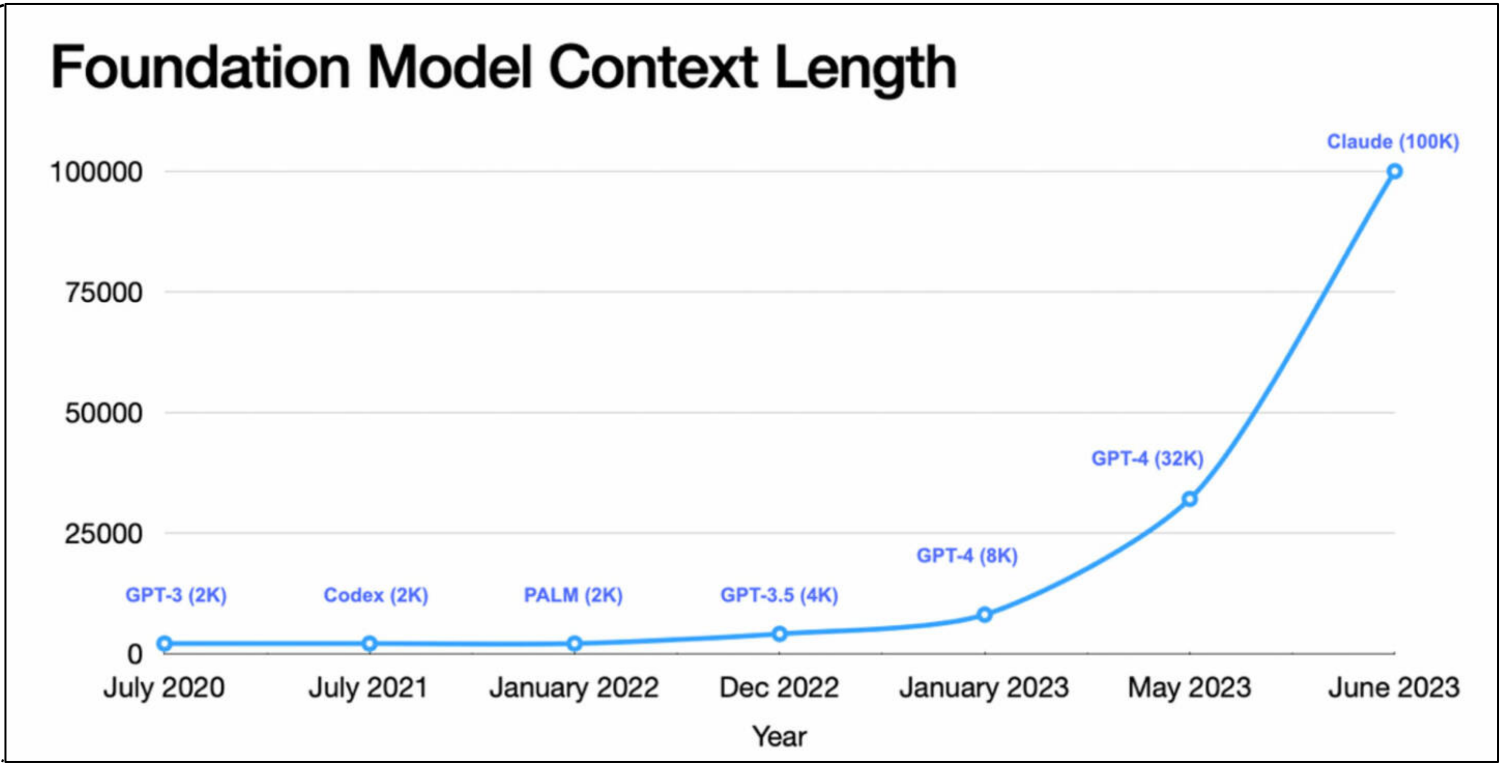


The Low Down Blog by Momentum Works. 2023

Background

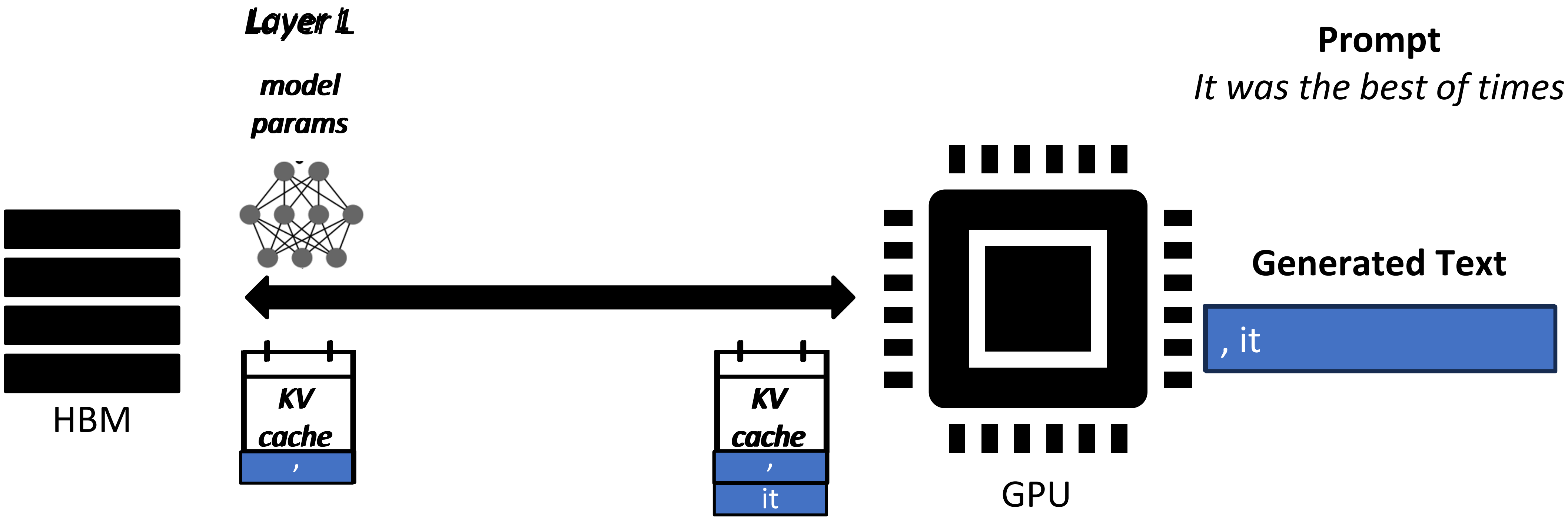
Tasks require longer context length





Hazy Research Blog. 2023

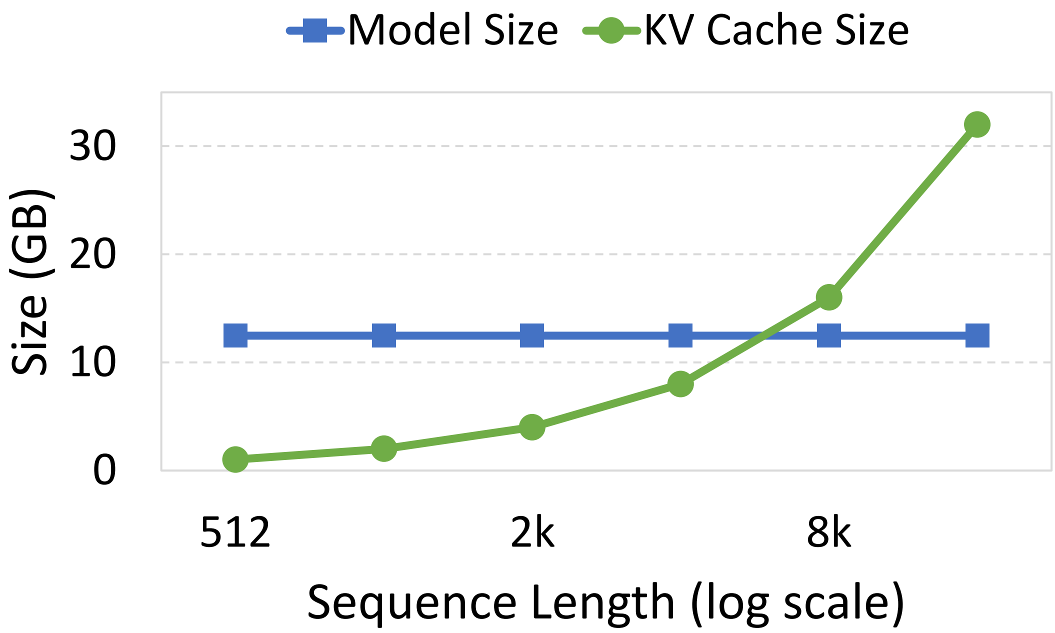
Off-chip Data Movement



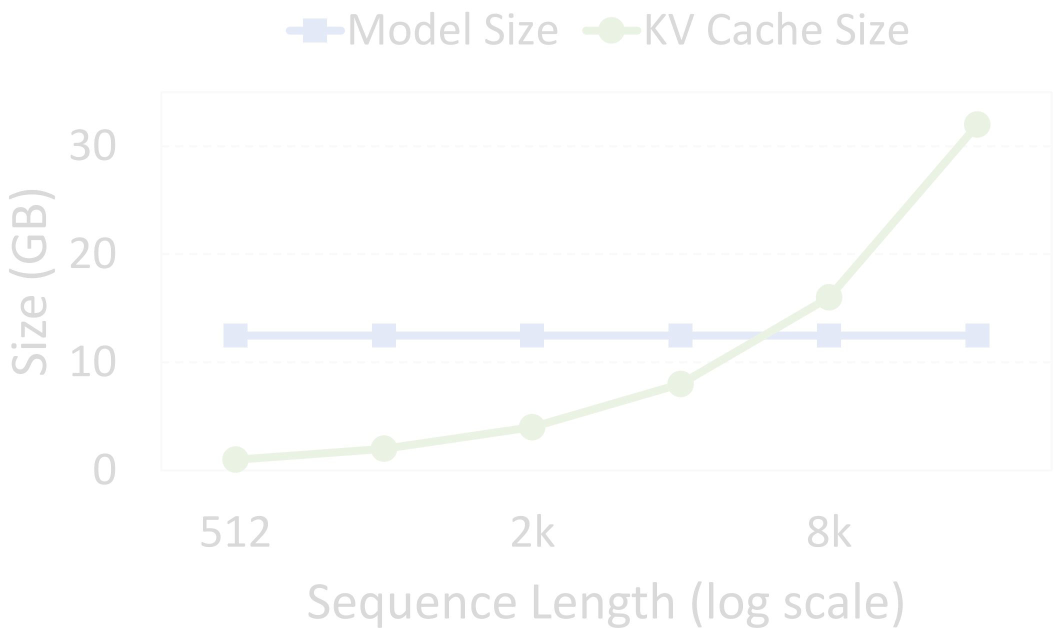
Prompt

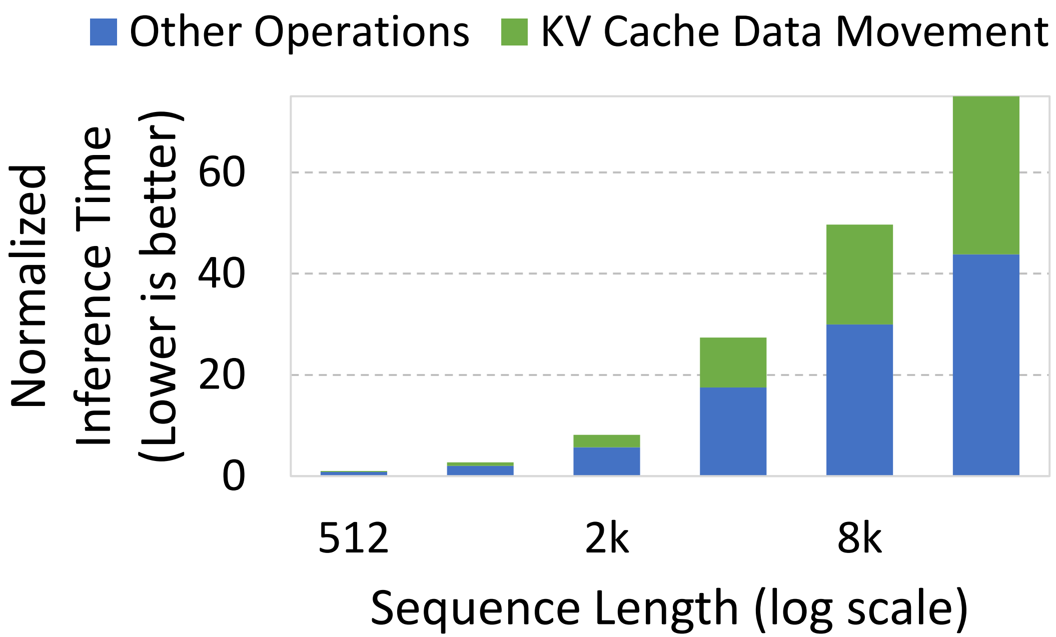
It was the best of times

Problem: Capacity, Bandwidth

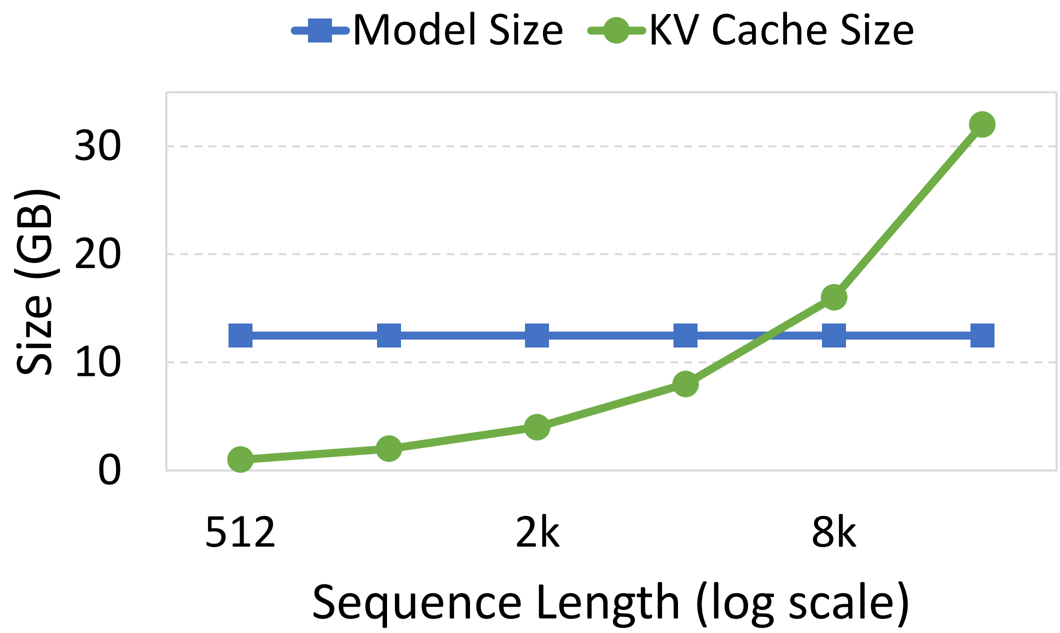


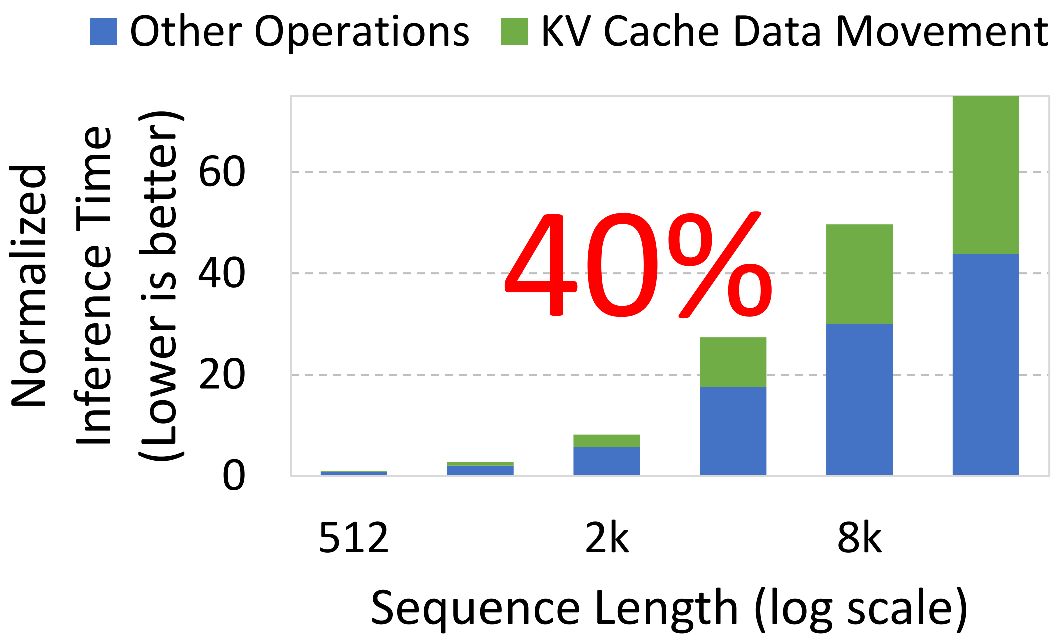
Problem: Inference Time





Problem: Capacity, Bandwidth & Time





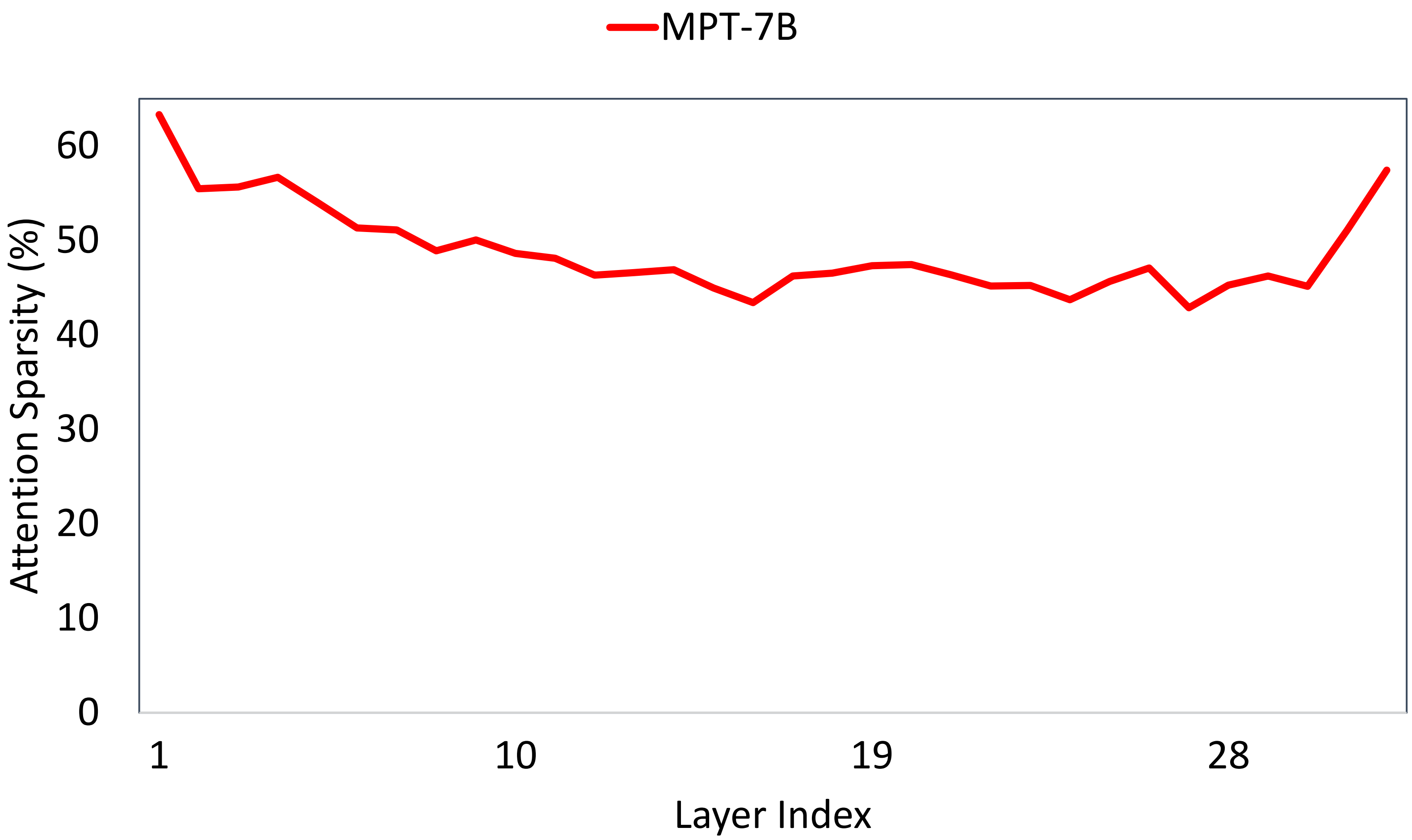
Profiling with MPT-7B model using varying Sequence Length (50% context + 50% text generation) using NVIDIA A100 (80GB) GPU and synthetic data.

Goals

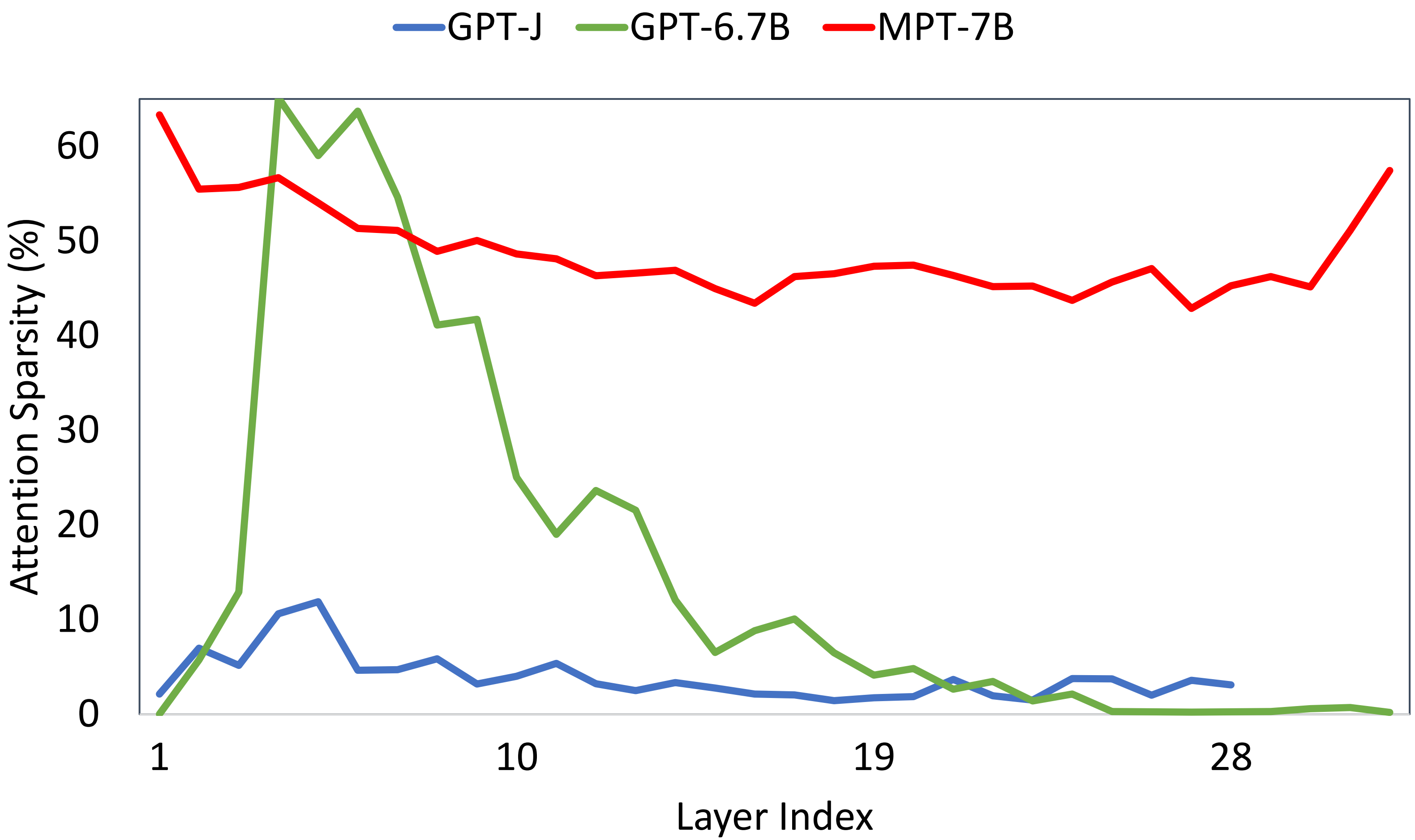
KV cache reduction without accuracy drop

KV cache reduction on the fly without any retraining or finetuning

Key Insight – I: Sparsity

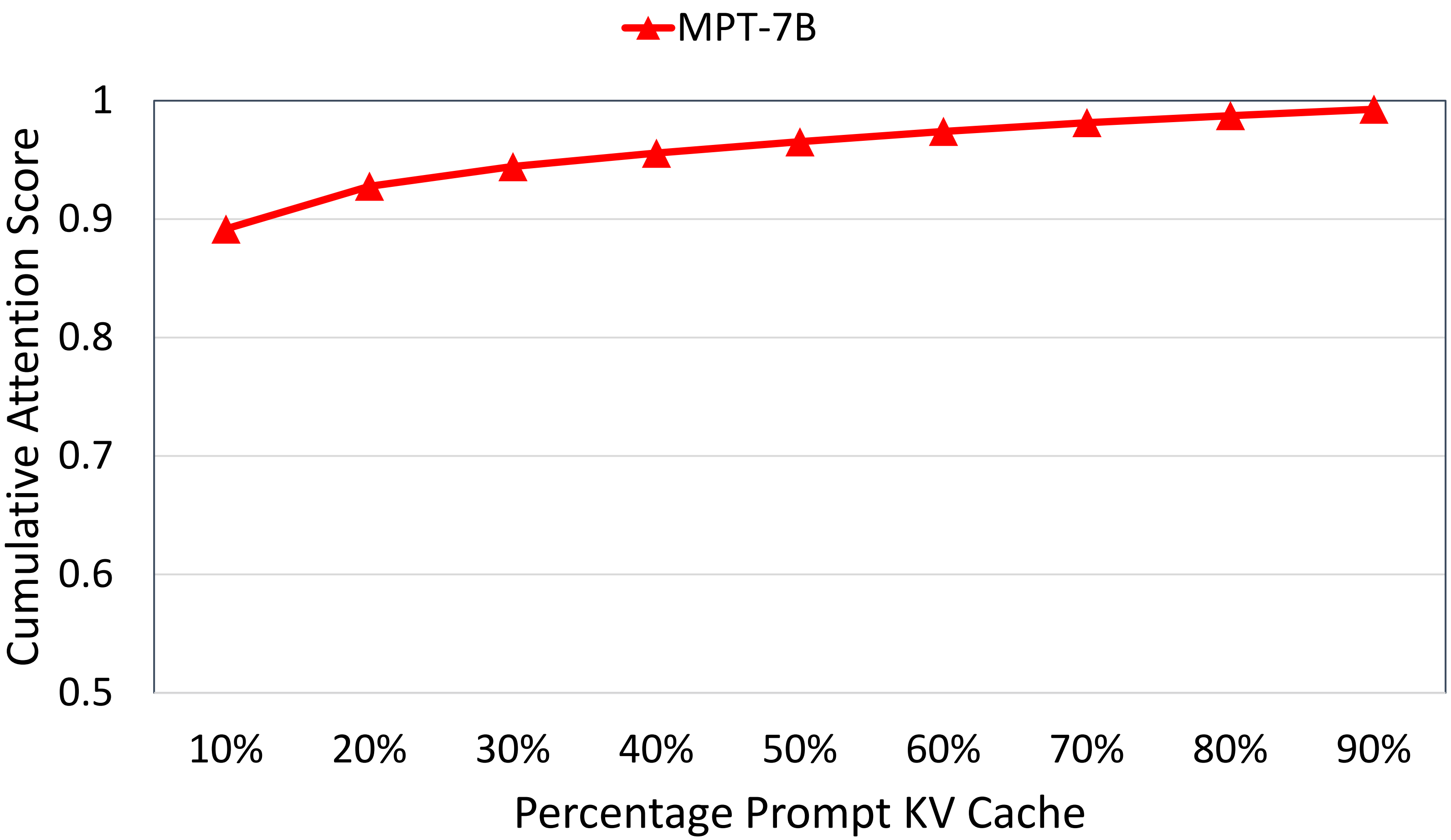


Key Insight – I: Sparsity

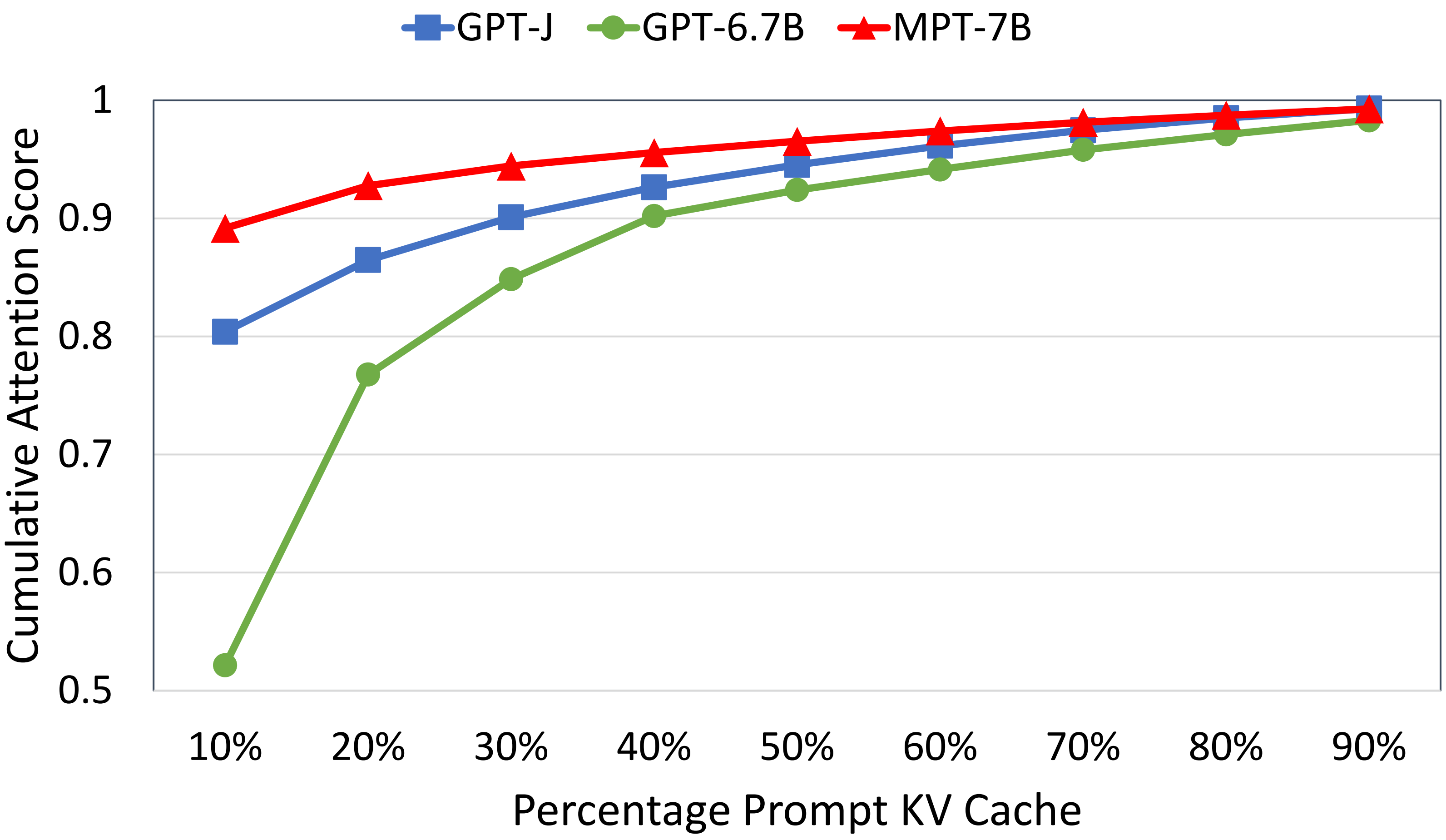


Average Attention Sparsity using CNN/DailyMail dataset for Summarization Task.

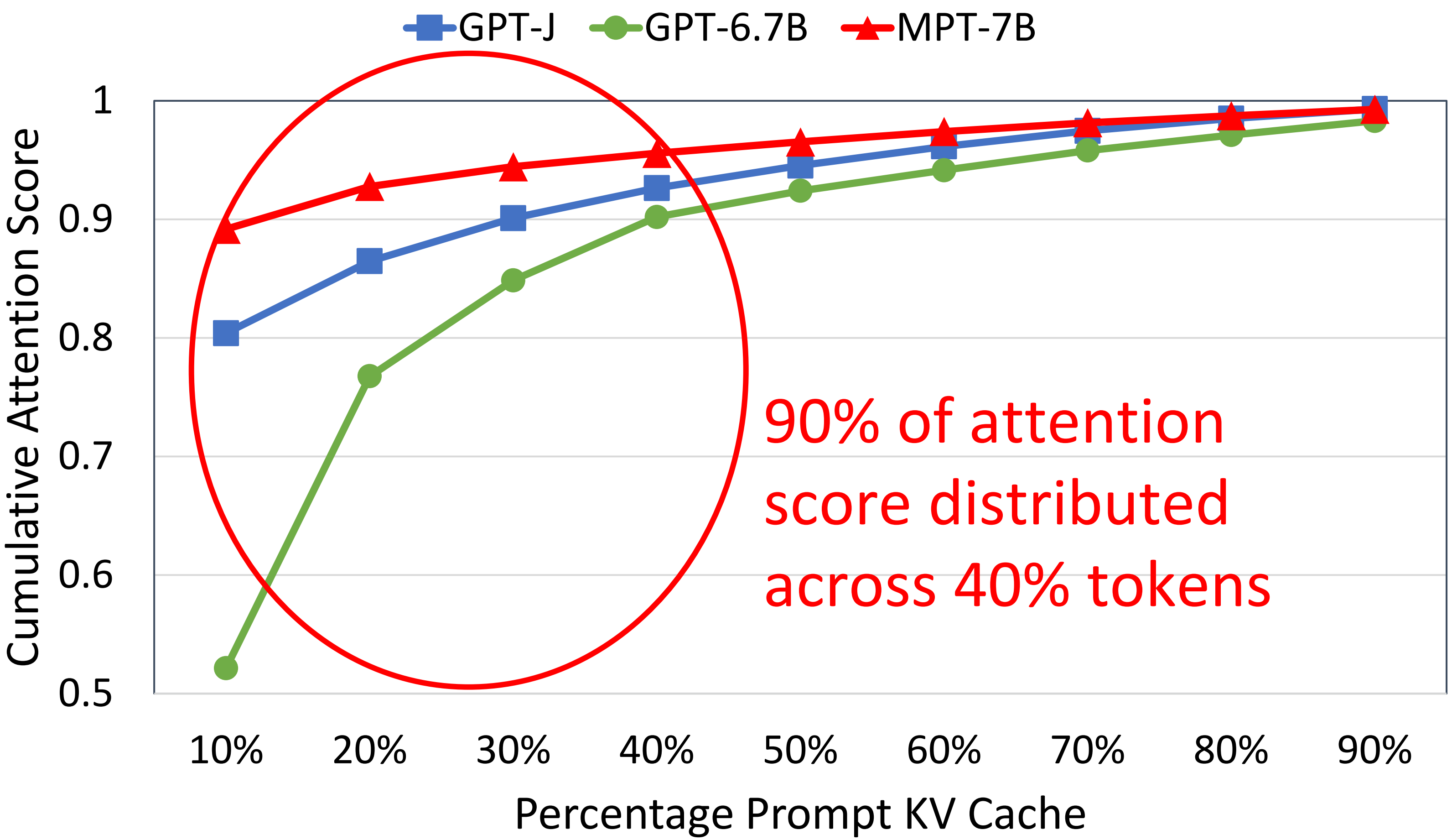
Key Insight – II: Attention Distribution



Key Insight – II: Attention Distribution



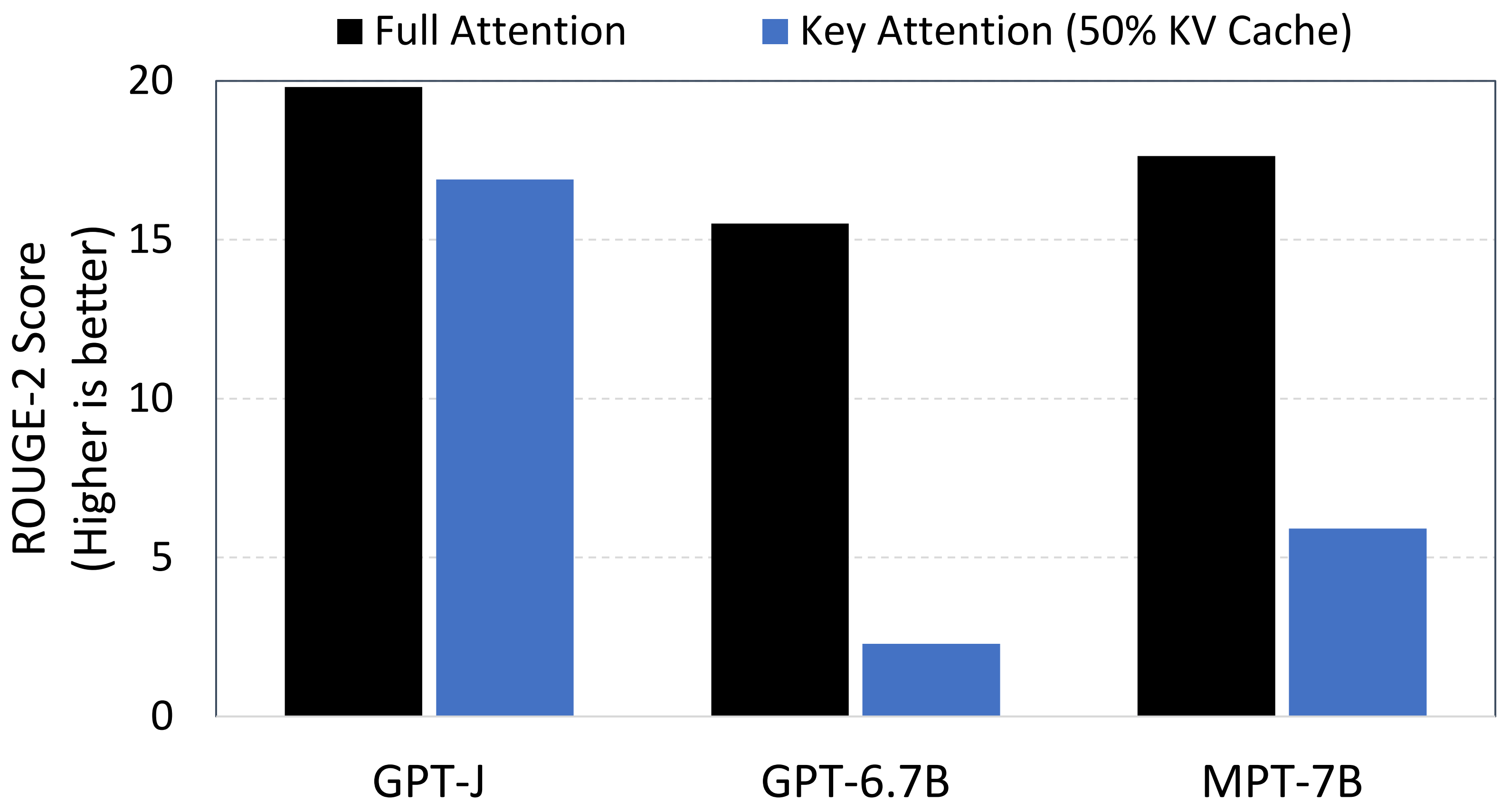
Key Insight – II: Attention Distribution



90% of attention score distributed across 40% tokens

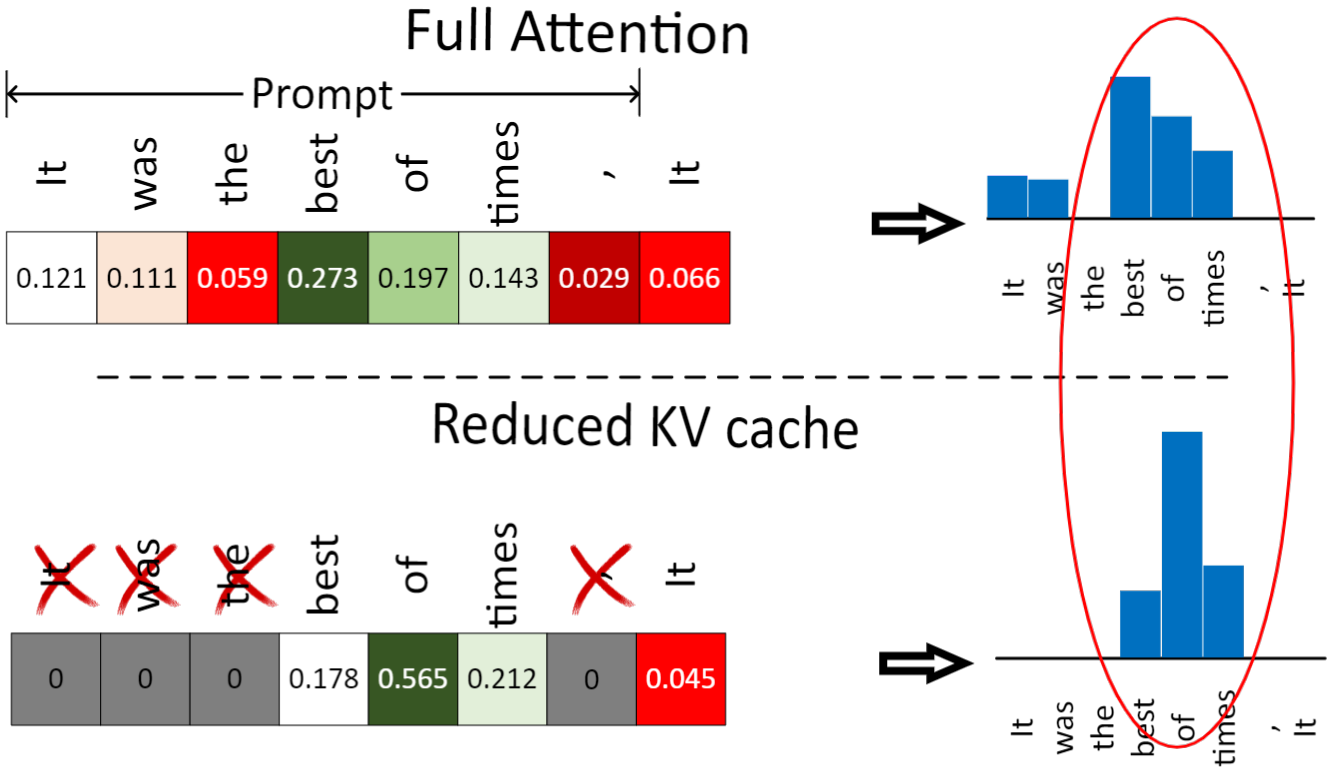
Attention Score Distribution using CNN/DailyMail dataset for Summarization Task.

Limitation – Key Tokens



Accuracy Drop with 50% KV Cache for Summarization Task with CNN/DailyMail dataset.

Key Insight – III: Change in Score Distribution



Full Attention

It	was	the	best	of	times	,	It
0.121	0.111	0.059	0.273	0.197	0.143	0.029	0.066

Reduced KV cache

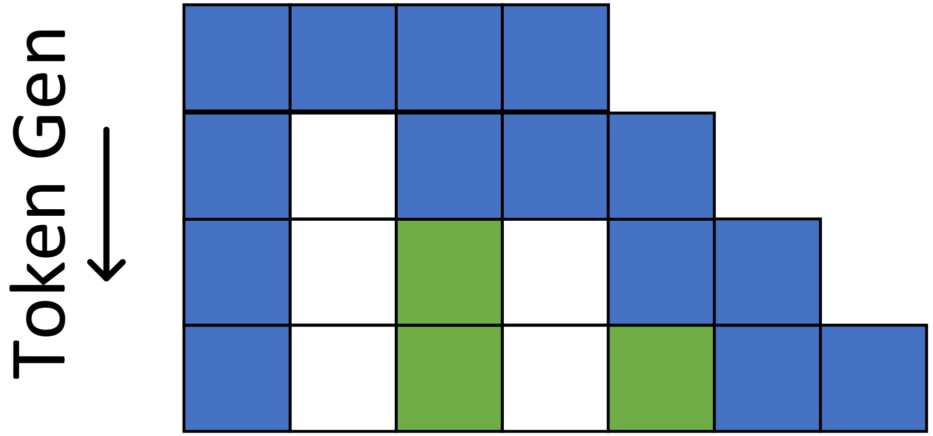
~~It~~	~~was~~	~~the~~	best	of	times	~~,~~	It
0	0	0	0.178	0.565	0.212	0	0.045

MPT-7B model layer 1 and Head 1 stats.

Challenges

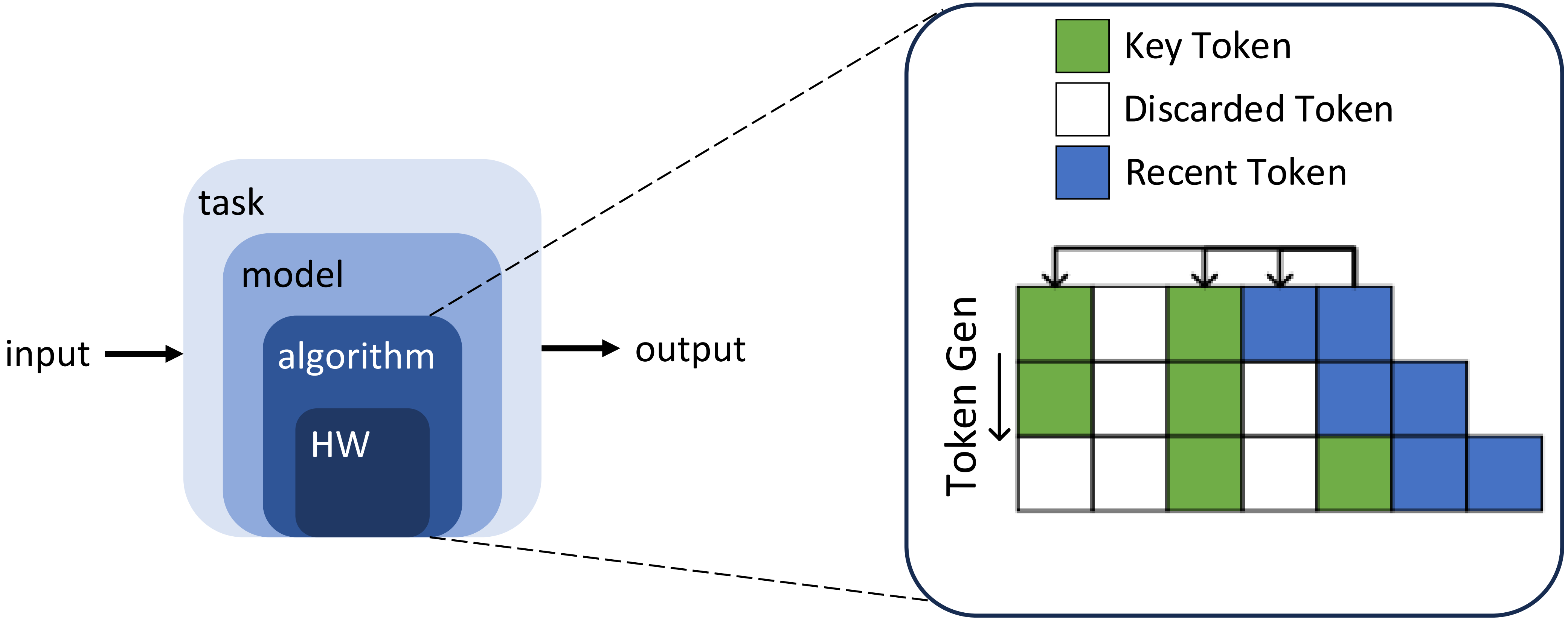
🟩 Key Token ⬜ Discarded Token

Key Tokens Identification



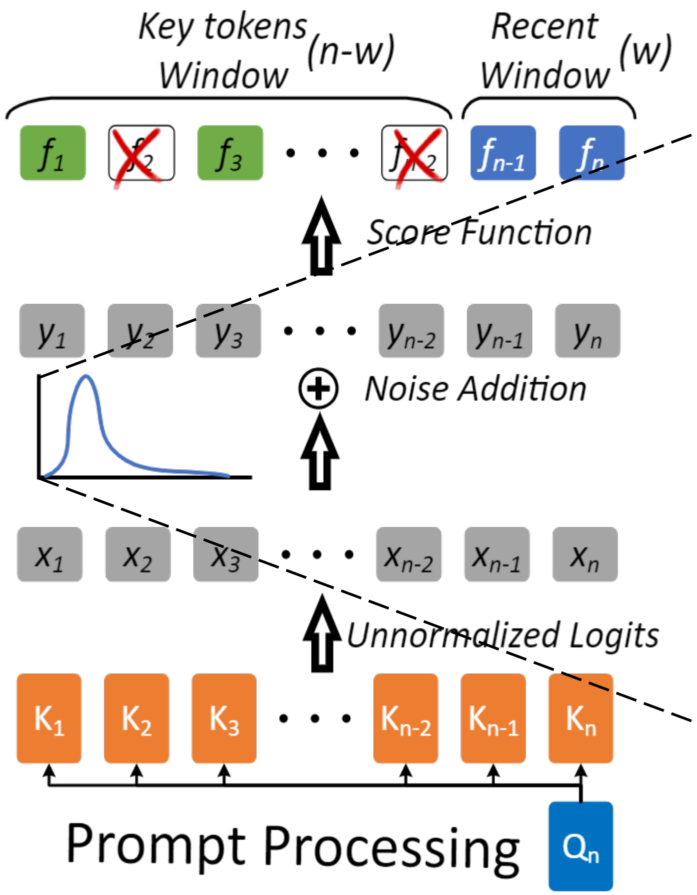
Discarded Tokens Utilization

Keyformer – Regularization based Key Token



Keyformer - KV Cache Reduction

🟩 *High Score* □ *Low Score* 🟦 *Recent* ✗ *Discarded Tokens*



Key tokens Window (n-w)

Recent Window (w)

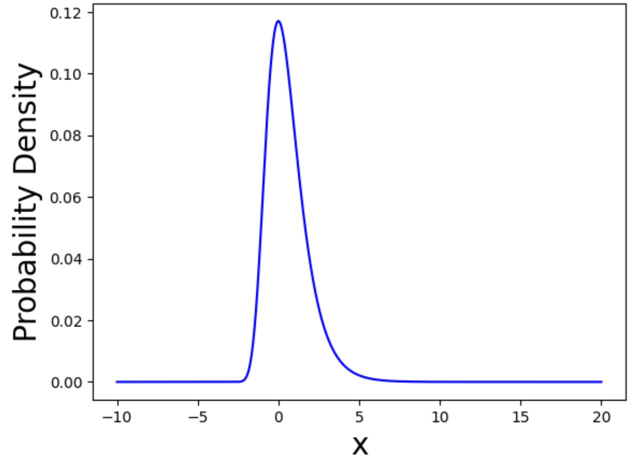
Score Function

Noise Addition

Unnormalized Logits

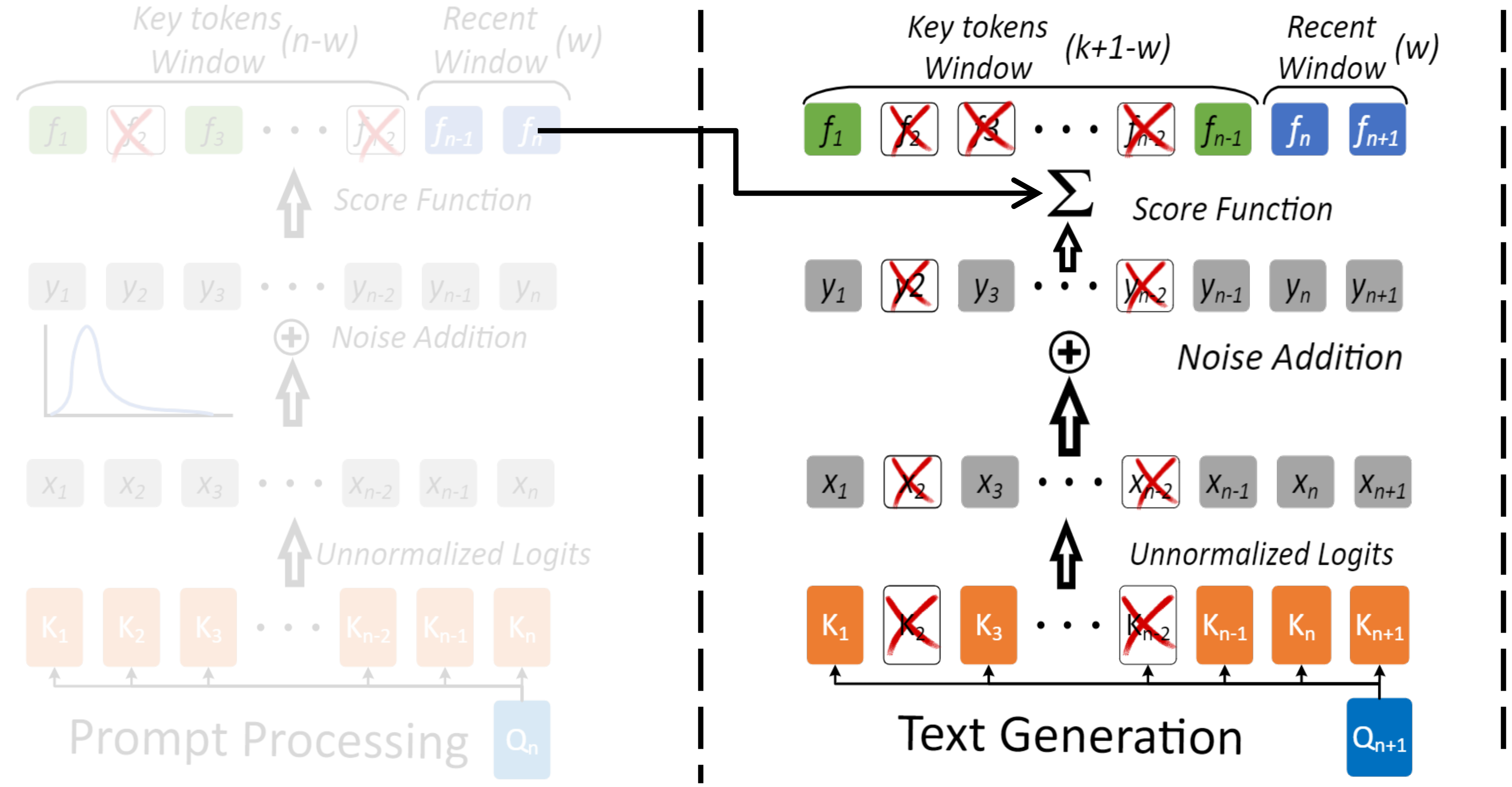
Prompt Processing

- Bias towards Initial tokens
- Long tail Distribution

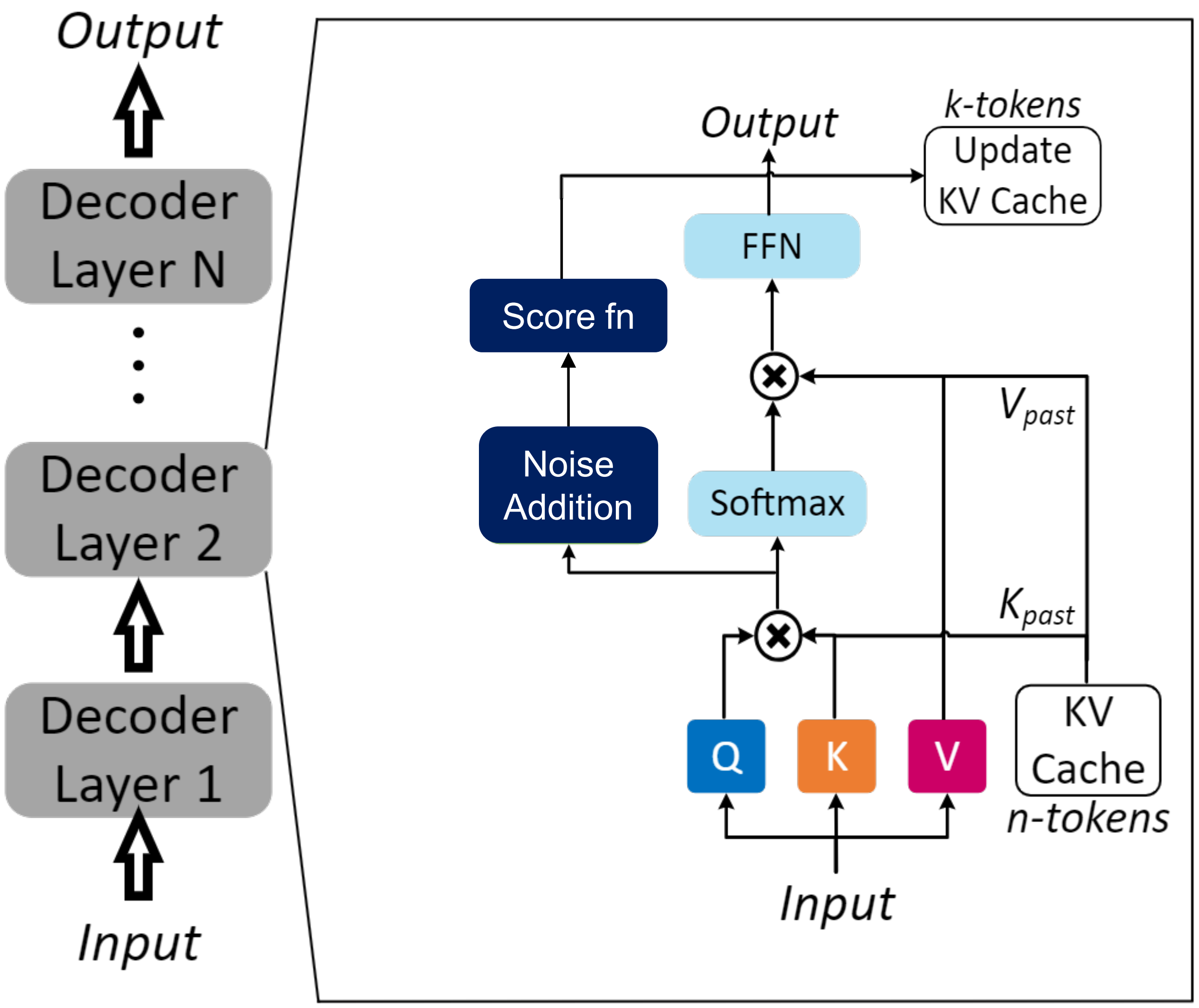


Keyformer - KV Cache Reduction

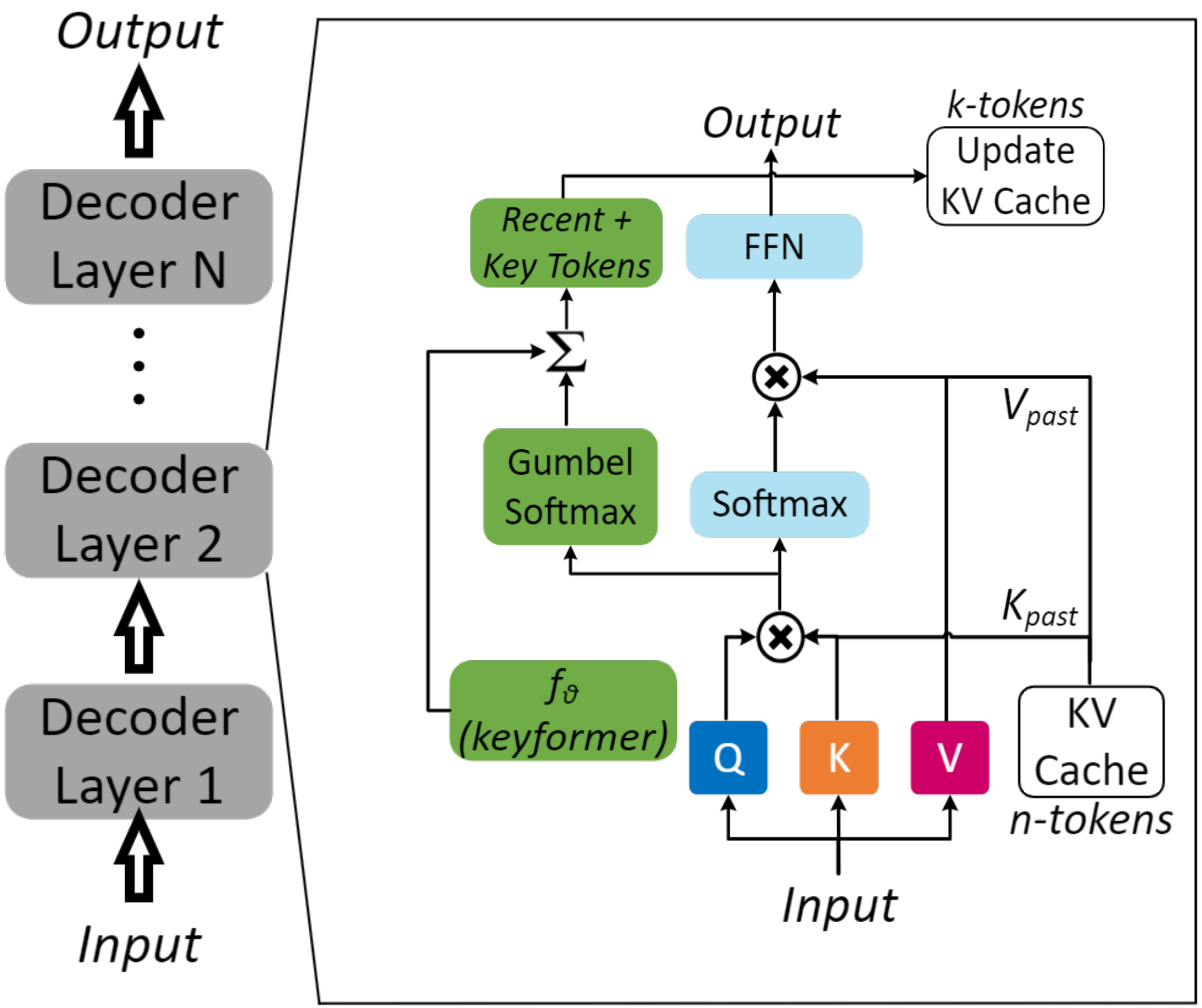
🟩 *High Score* ⬜ *Low Score* 🟦 *Recent* ❌ *Discarded Tokens*


Key tokens Window (n-w)
Recent Window (w)
f_1 f_2 f_3 . . . f_{n-2} f_{n-1} f_n
Score Function
y_1 y_2 y_3 . . . y_{n-2} y_{n-1} y_n
Noise Addition
x_1 x_2 x_3 . . . x_{n-2} x_{n-1} x_n
Unnormalized Logits
K_1 K_2 K_3 . . . K_{n-2} K_{n-1} K_n
Prompt Processing
Q_n
Key tokens Window (k+1-w)
Recent Window (w)
f_1 f_2 f_3 . . . f_{n-2} f_{n-1} f_n f_{n+1}
Σ
Score Function
y_1 y_2 y_3 . . . y_{n-2} y_{n-1} y_n y_{n+1}
Noise Addition
x_1 x_2 x_3 . . . x_{n-2} x_{n-1} x_n x_{n+1}
Unnormalized Logits
K_1 K_2 K_3 . . . K_{n-2} K_{n-1} K_n K_{n+1}
Text Generation
Q_{n+1}

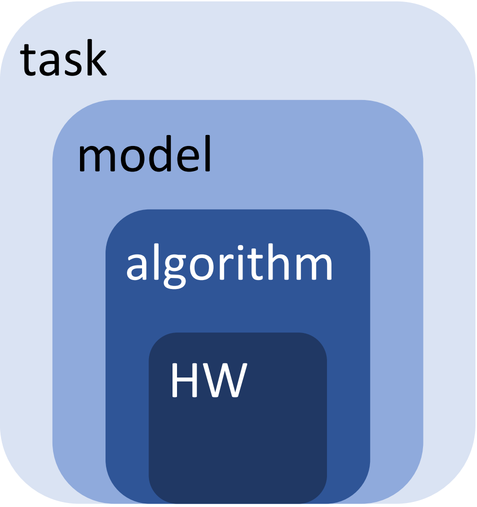
Keyformer - Decoder



Keyformer - Decoder



Evaluation



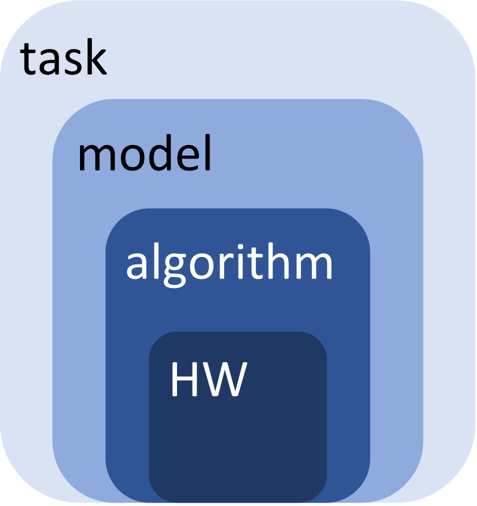
Summarization
Conversation,
COPA (Commonsense Reasoning)
OpenBookQA (Commonsense Reasoning)
Winogrande (Language Understanding)
PIQA (Commonsense Reasoning)

GPT-J (finetuned – Summarization) - *RoPE*
MPT- chat (finetuned – Dialogue) - *ALiBi*
Cerebras GPT (pretrained) - *Absolute*
MPT (pretrained) - *ALiBi*

Full Attention
Window Attention
H_2O

NVIDIA A100 (80GB)

Evaluation



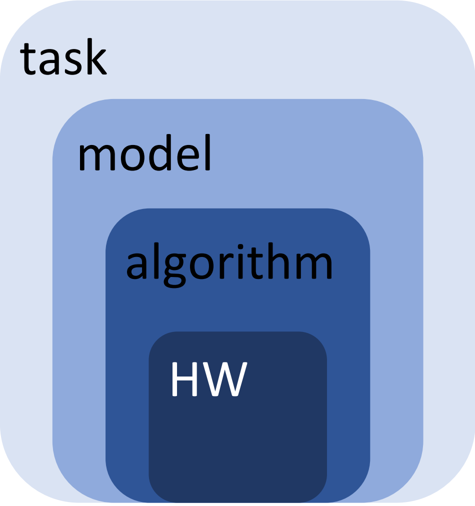
Summarization
Conversation,
COPA (Commonsense Reasoning)
OpenBookQA (Commonsense Reasoning)
Winogrande (Language Understanding)
PIQA (Commonsense Reasoning)

GPT-J (finetuned – Summarization) - *RoPE*
MPT- chat (finetuned – Dialogue) - *ALiBi*
Cerebras GPT (pretrained) - *Absolute*
MPT (pretrained) - *ALiBi*

Full Attention
Window Attention
H_2O

NVIDIA A100 (80GB)

Evaluation



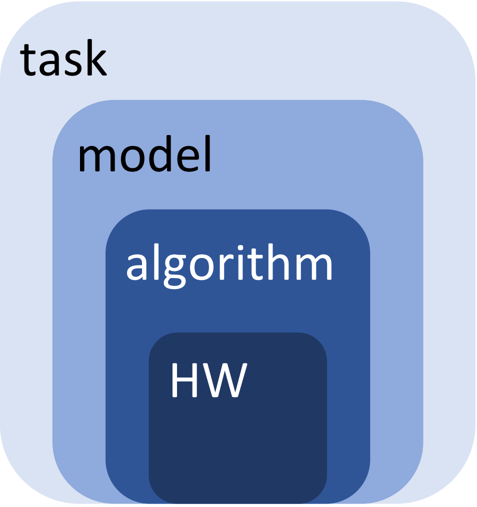
Summarization
Conversation,
COPA (Commonsense Reasoning)
OpenBookQA (Commonsense Reasoning)
Winogrande (Language Understanding)
PIQA (Commonsense Reasoning)

GPT-J (finetuned – Summarization) - *RoPE*
MPT- chat (finetuned – Dialogue) - *ALiBi*
Cerebras GPT (pretrained) - *Absolute*
MPT (pretrained) - *ALiBi*

Full Attention
Window Attention
H_2O

NVIDIA A100 (80GB)

Evaluation



task

model

algorithm

HW

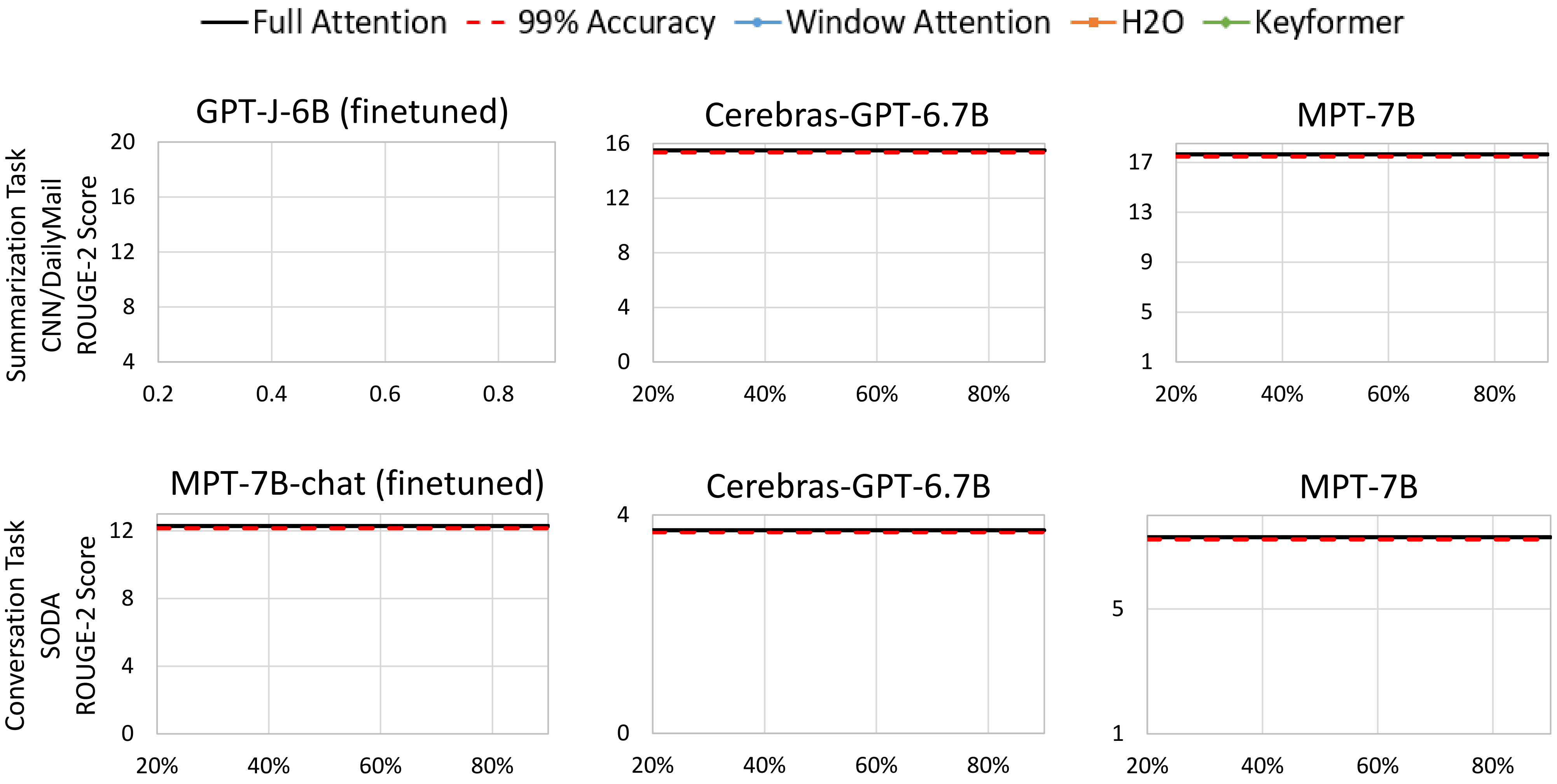
Summarization
Conversation,
COPA (Commonsense Reasoning)
OpenBookQA (Commonsense Reasoning)
Winogrande (Language Understanding)
PIQA (Commonsense Reasoning)

GPT-J (finetuned – Summarization) - *RoPE*
MPT- chat (finetuned – Dialogue) - *ALiBi*
Cerebras GPT (pretrained) - *Absolute*
MPT (pretrained) - *ALiBi*

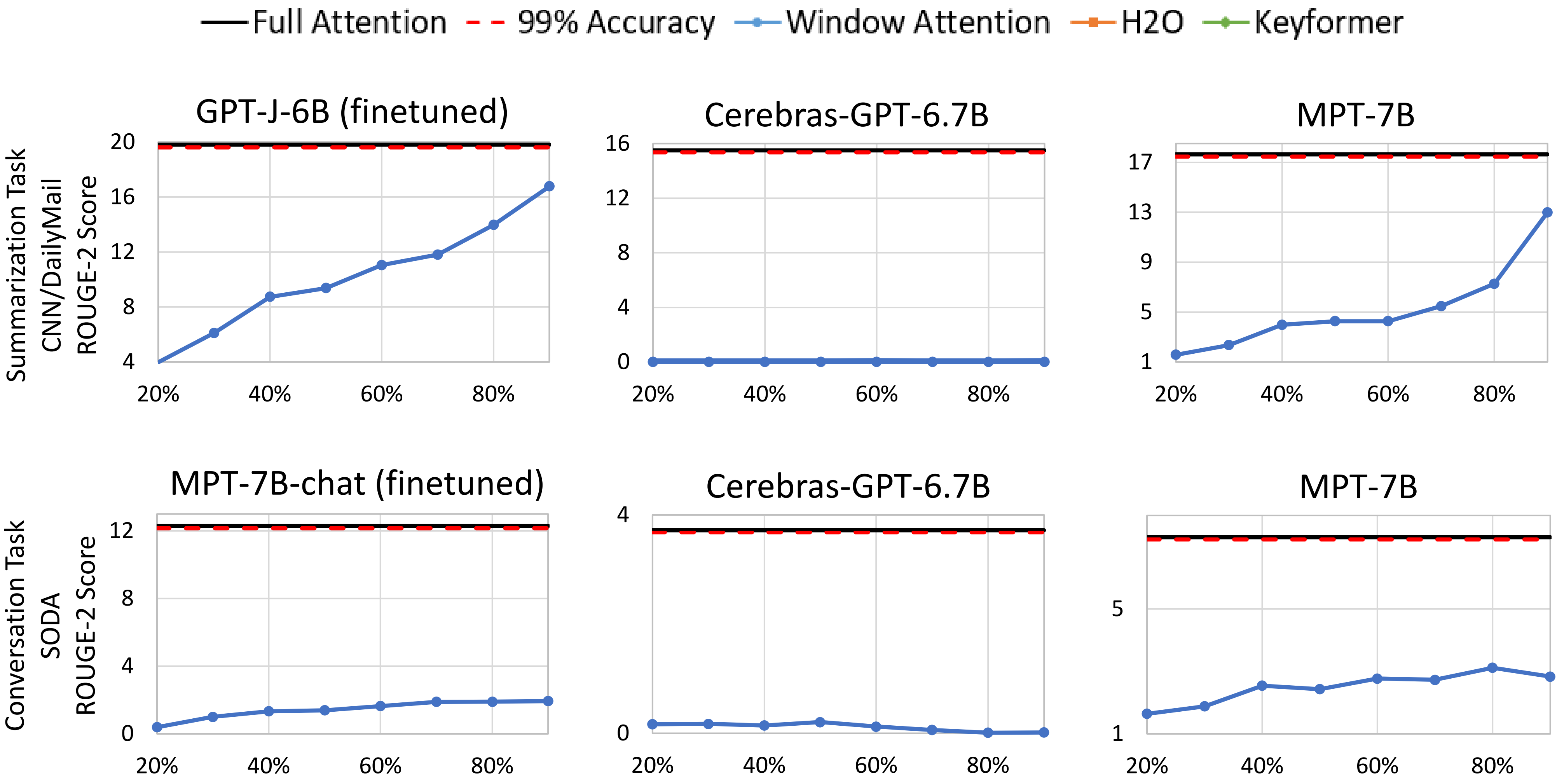
Full Attention
Window Attention
H_2O

NVIDIA A100 (80GB)

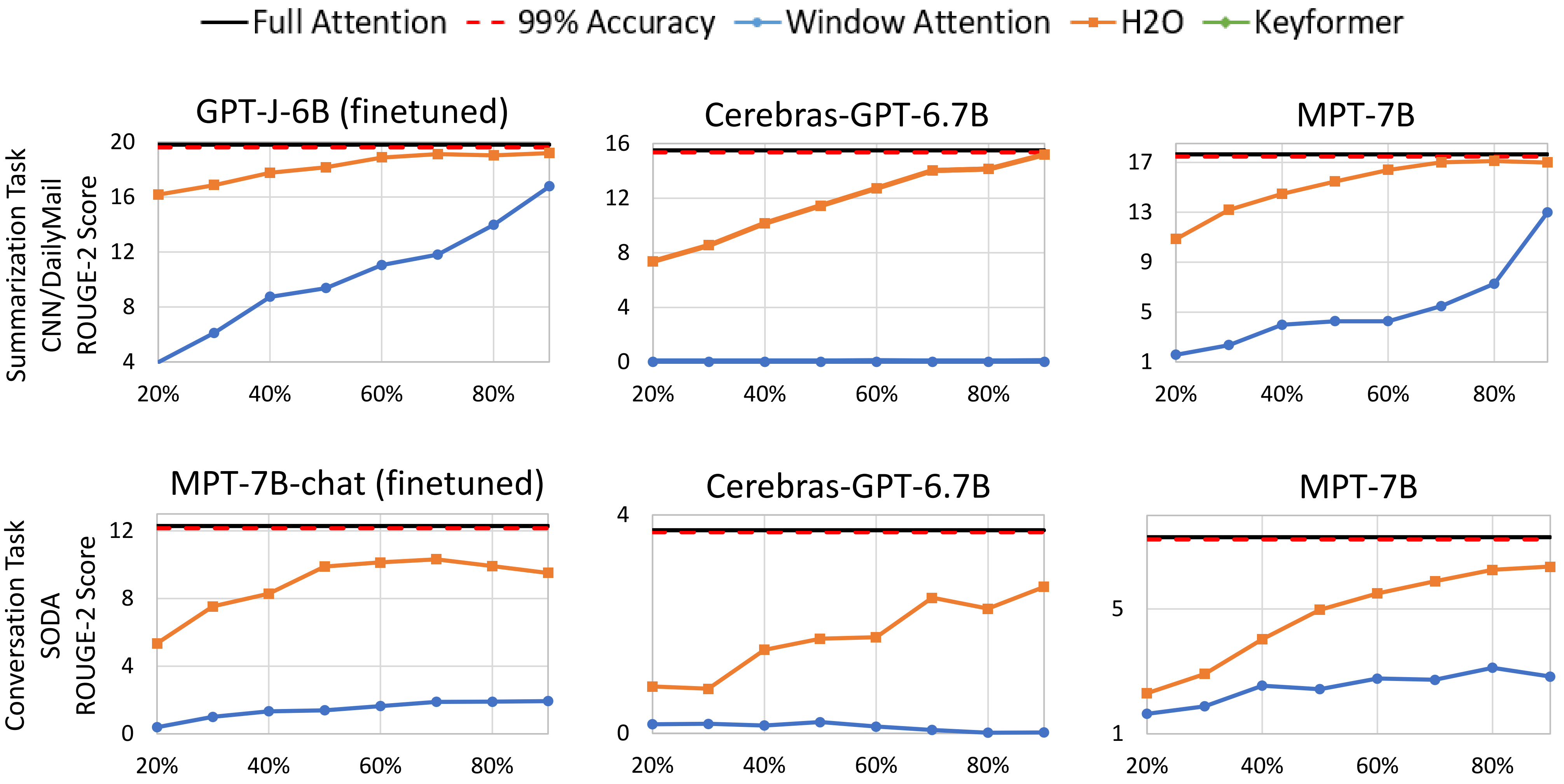
Accuracy Tradeoff



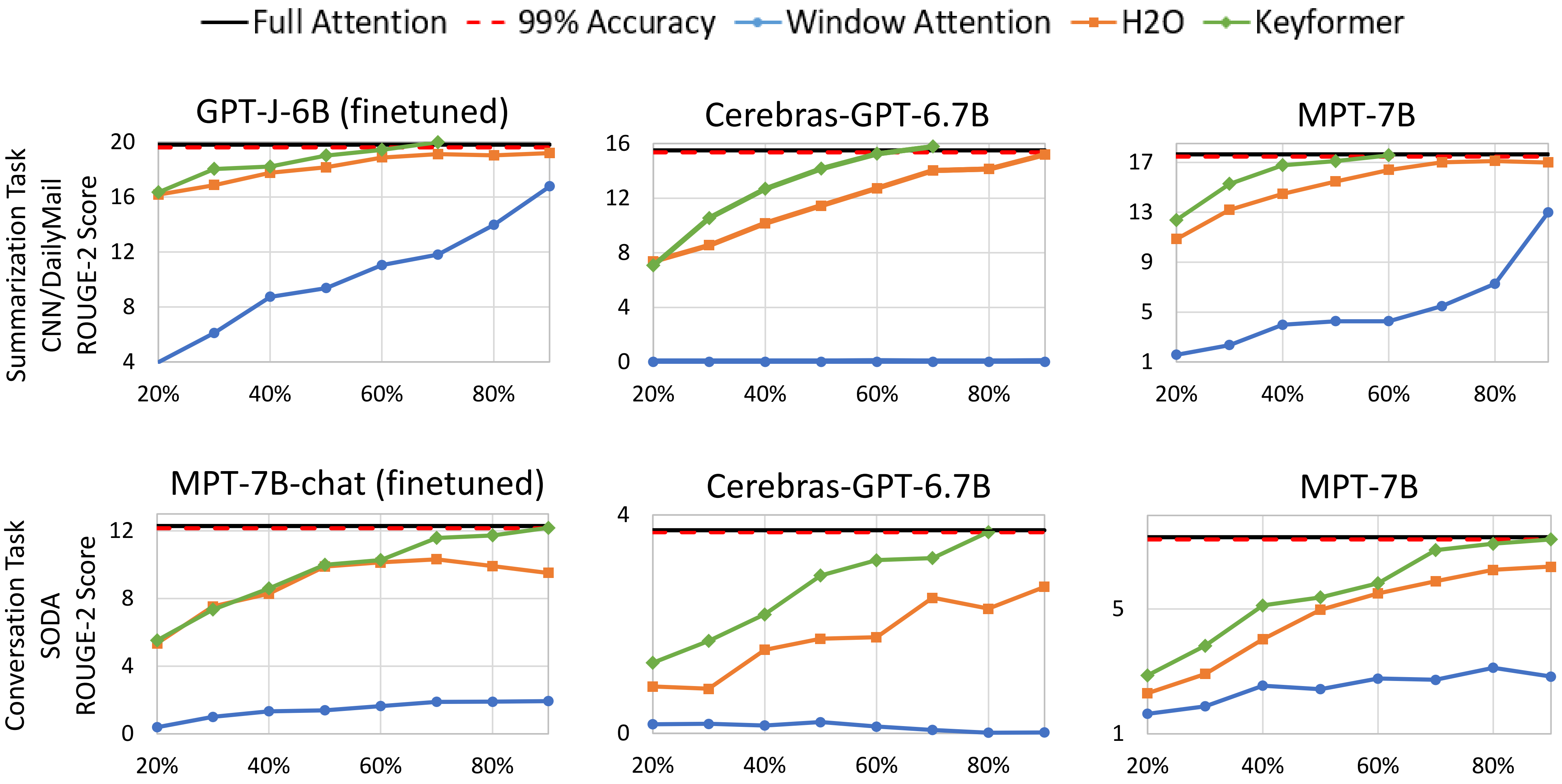
Accuracy Tradeoff



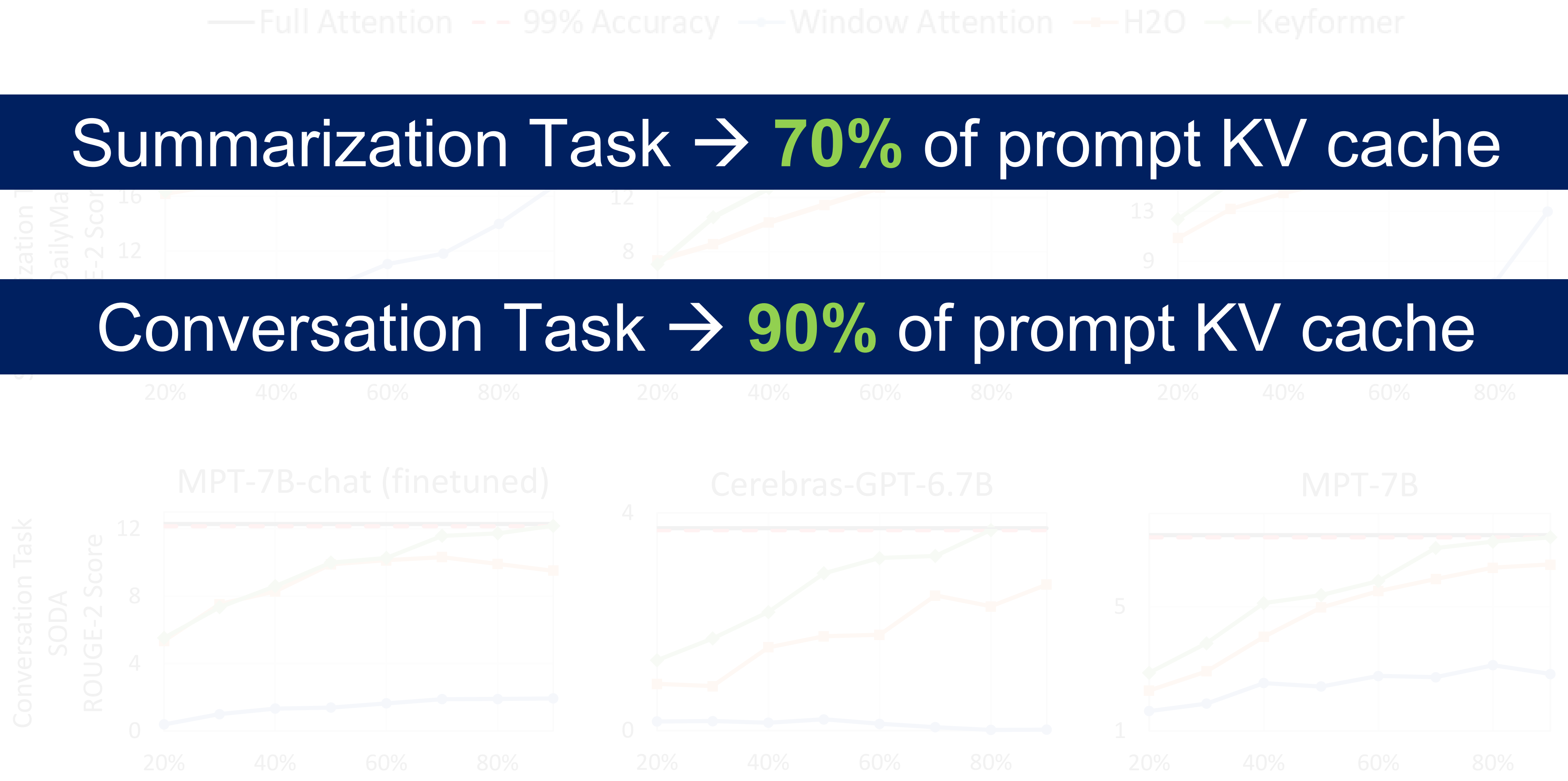
Accuracy Tradeoff



Accuracy Tradeoff



Accuracy Tradeoff



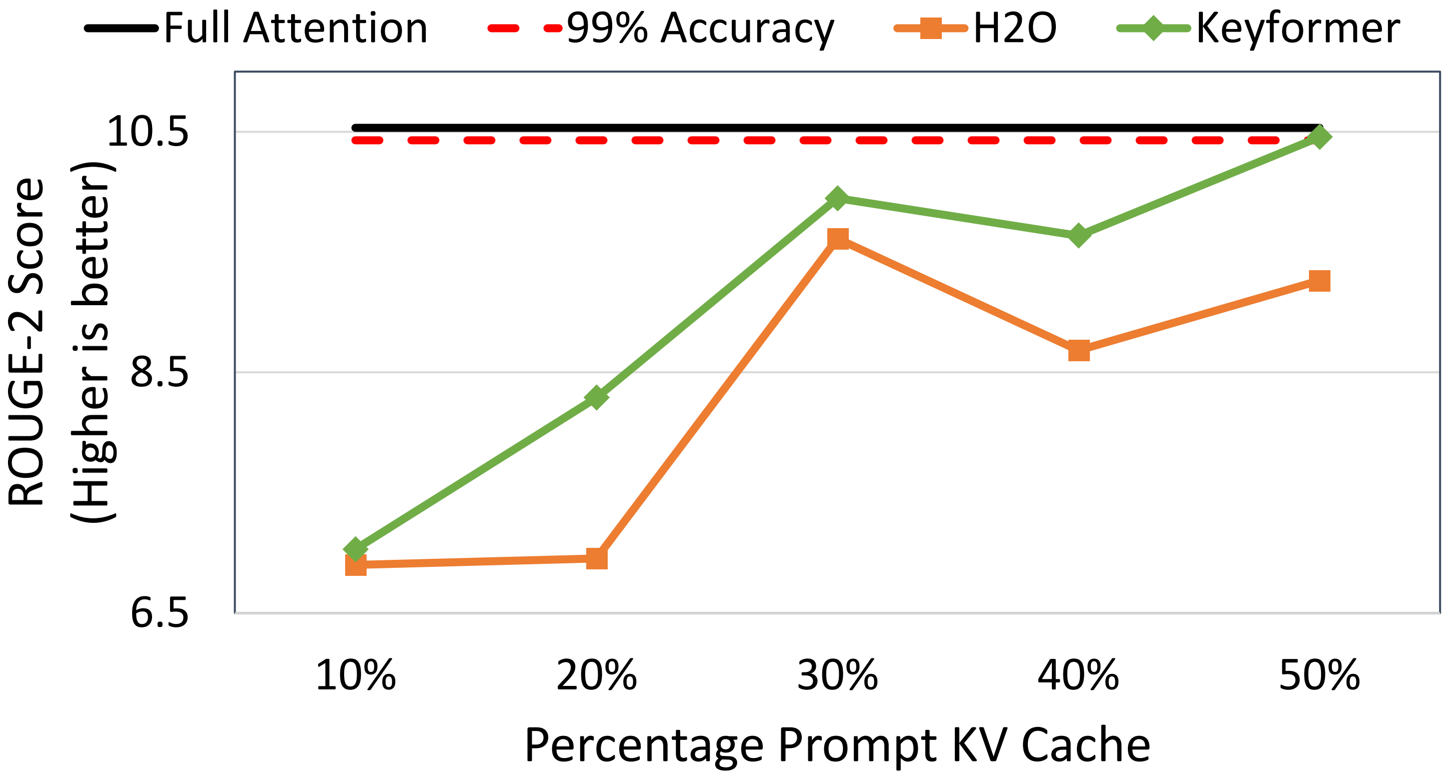
Summarization Task → **70%** of prompt KV cache

Conversation Task → **90%** of prompt KV cache

Long Context

Model : MPT-7B-storywriter (65k seq len)

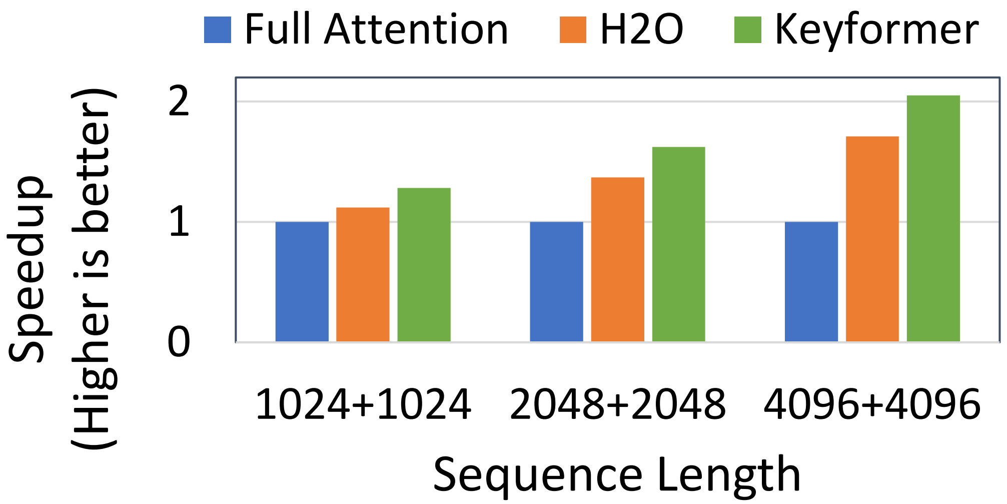
Dataset : Gov Reports (Mean Context: 9k)



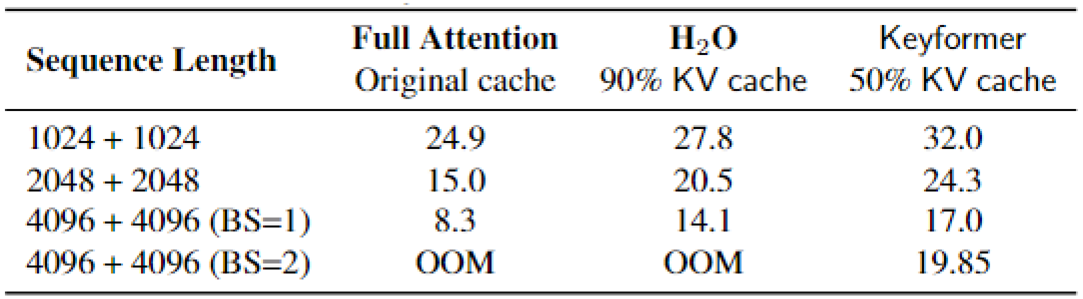
Long Context

Model : MPT-7B-storywriter (65k seq len)

Dataset : Gov Reports (Mean Context: 9k)



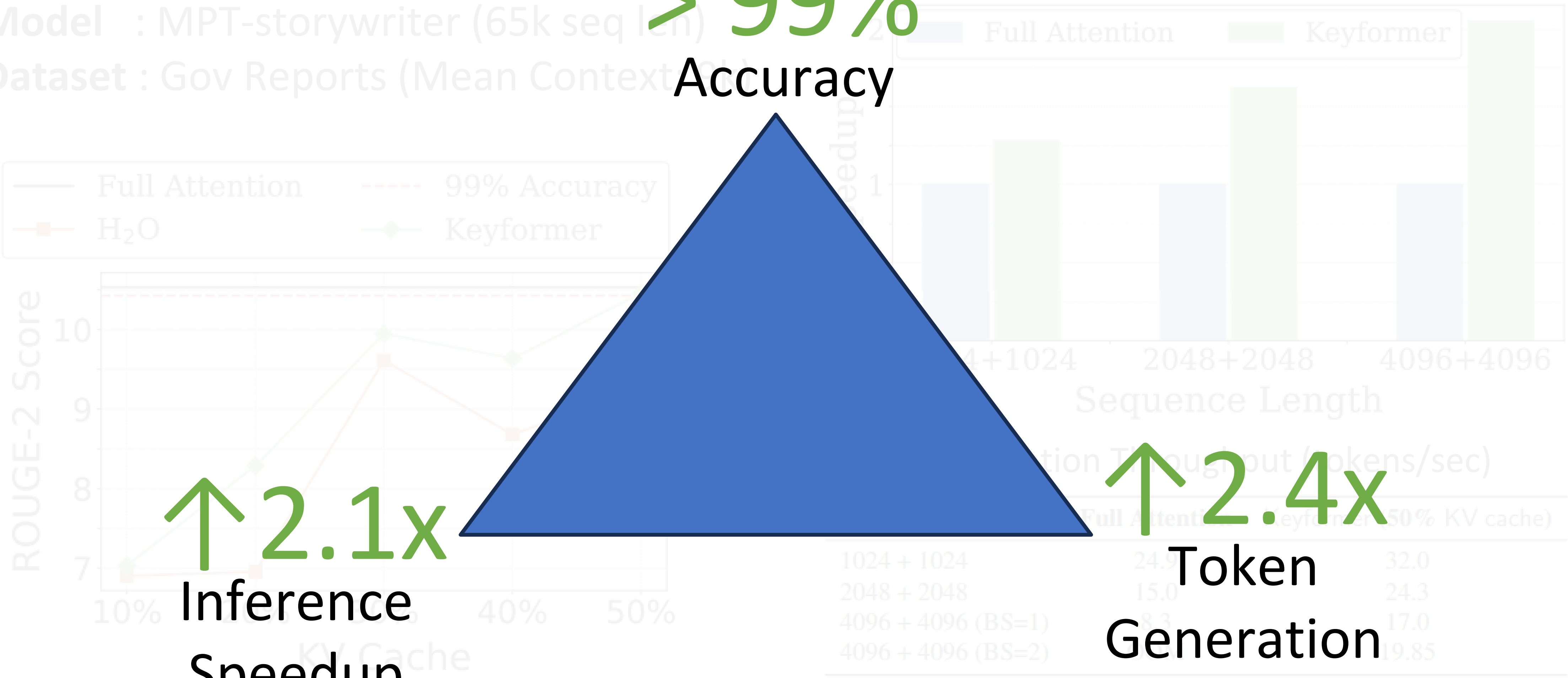
Generation Throughput (tokens/sec)



Sequence Length	Full Attention Original cache	H_2O 90% KV cache	Keyformer 50% KV cache
1024 + 1024	24.9	27.8	32.0
2048 + 2048	15.0	20.5	24.3
4096 + 4096 (BS=1)	8.3	14.1	17.0
4096 + 4096 (BS=2)	OOM	OOM	19.85

Long Context

Model : MPT-storywriter (65k seq len)
Dataset : Gov Reports (Mean Context 8k)



> 99%
Accuracy

↑2.1x
Inference Speedup

↑2.4x
Token Generation Throughput

Generation Throughput (tokens/sec)

	Full Attention	Keyformer (50% KV cache)
1024 + 1024	24.9	32.0
2048 + 2048	15.0	24.3
4096 + 4096 (BS=1)	8.3	17.0
4096 + 4096 (BS=2)	[illegible]	19.85

Conclusion

- LLMs → Inherent sparsity within attention
- Sparsity → KV cache reduction
- Token Discarding → Effect on attention score distribution
- Keyformer → Regularization based Key token Identification

> 99%
Accuracy

↑2.1x
Inference Speedup

↑2.4x
Token Generation Throughput

Questions