ICLR

FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference

Xunhao Lai [1], Jianqiao Lu [2]

Yao Luo [3], Yiyuan Ma [3], Xun Zhou [3]

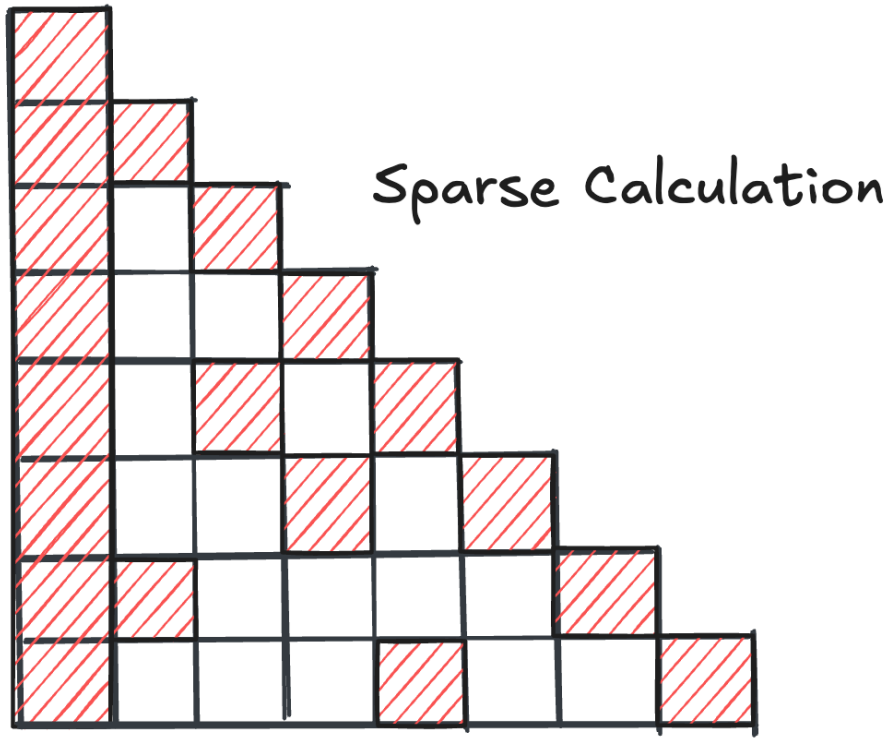
[1] Peking University [2] The University of Hong Kong [3] ByteDance Inc

北京大學
PEKING UNIVERSITY

香 港 大 學
THE UNIVERSITY OF HONG KONG

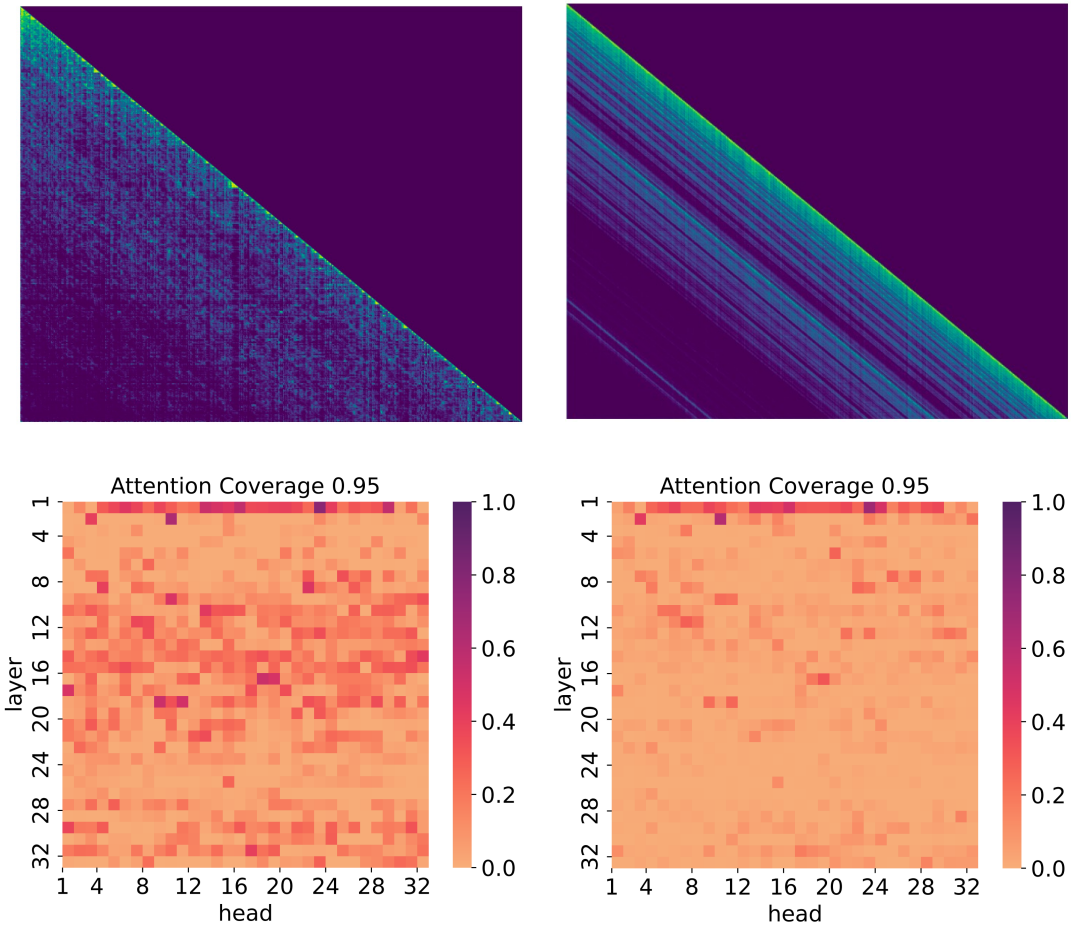
Introduction

- Long-sequence inference in LLMs is computationally intensive.
- The attention pre-filling phase scales quadratically with sequence length.
- Softmax attention scores are usually sparse.



Understanding Attention Sparsity

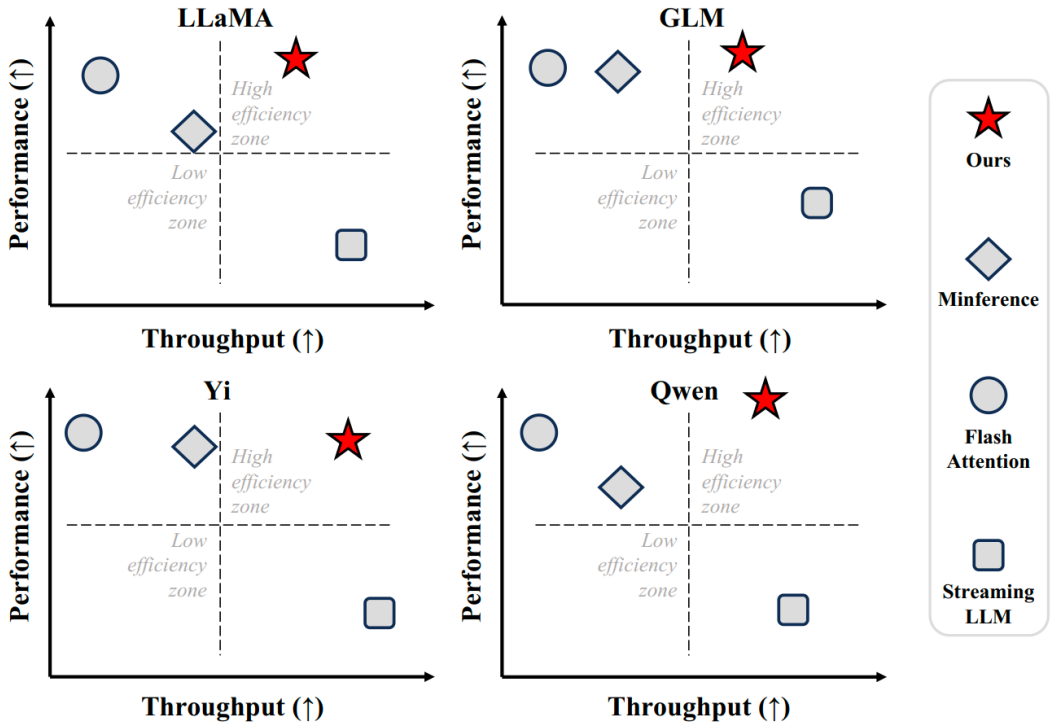
- Attention Patterns Vary
  - Diverse, unstructured patterns.
  - Structured patterns (e.g., vertical-slash [1]).
- Sparse Ratio Vary
  - Input length & task type
  - Different layers and heads



[1] Jiang, Huiqiang, et al. "Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention." Advances in Neural Information Processing Systems 37 (2024): 52481-52515.

Introducing FlexPrefill

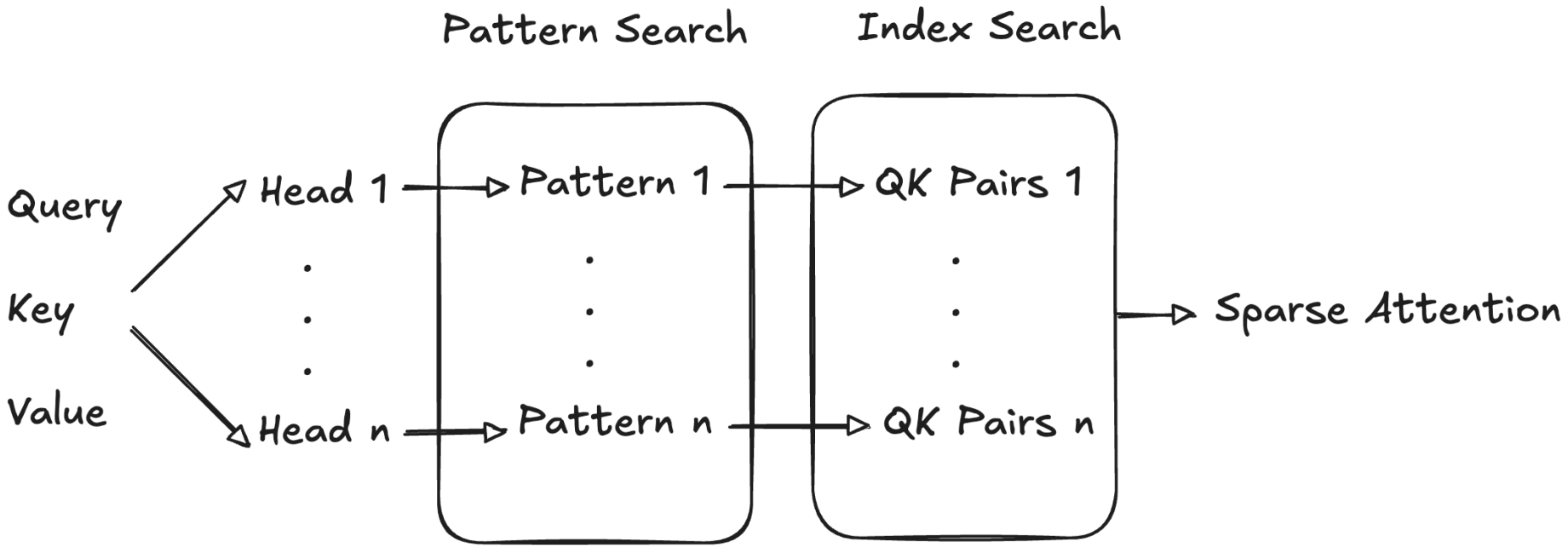
- What is FlexPrefill ?
  - A flexible, context-aware sparse attention mechanism for faster long-context pre-filling.
- Key Features:
  - Dynamic sparsity
  - Plug-and-play



Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025

Algorithm - Overview

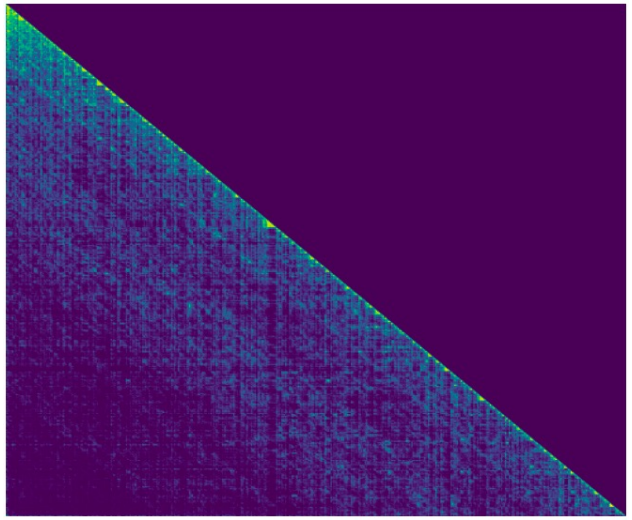
1. Determine sparse pattern.
2. Select critical query-key pairs based on the pattern.
3. Compute sparse attention only on selected QK pairs.



Algorithm – 1. Sparse Pattern Determination

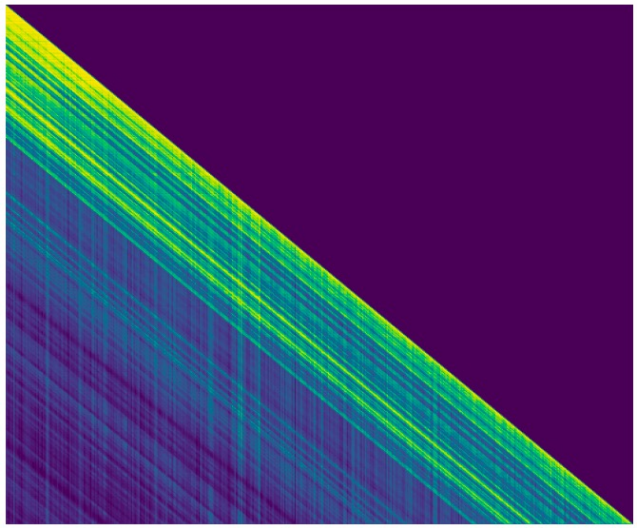
- **Query-Aware Pattern**:

  Attention is concentrated on various positions and does not follow a specific structure.



- **Structured-Sparse (Vertical-Slash) Pattern**:

  Attention is concentrated along vertical and slash like structures, which is consistent across queries.



Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025

Algorithm – 1. Sparse Pattern Determination

How to handle these two types of patterns ?

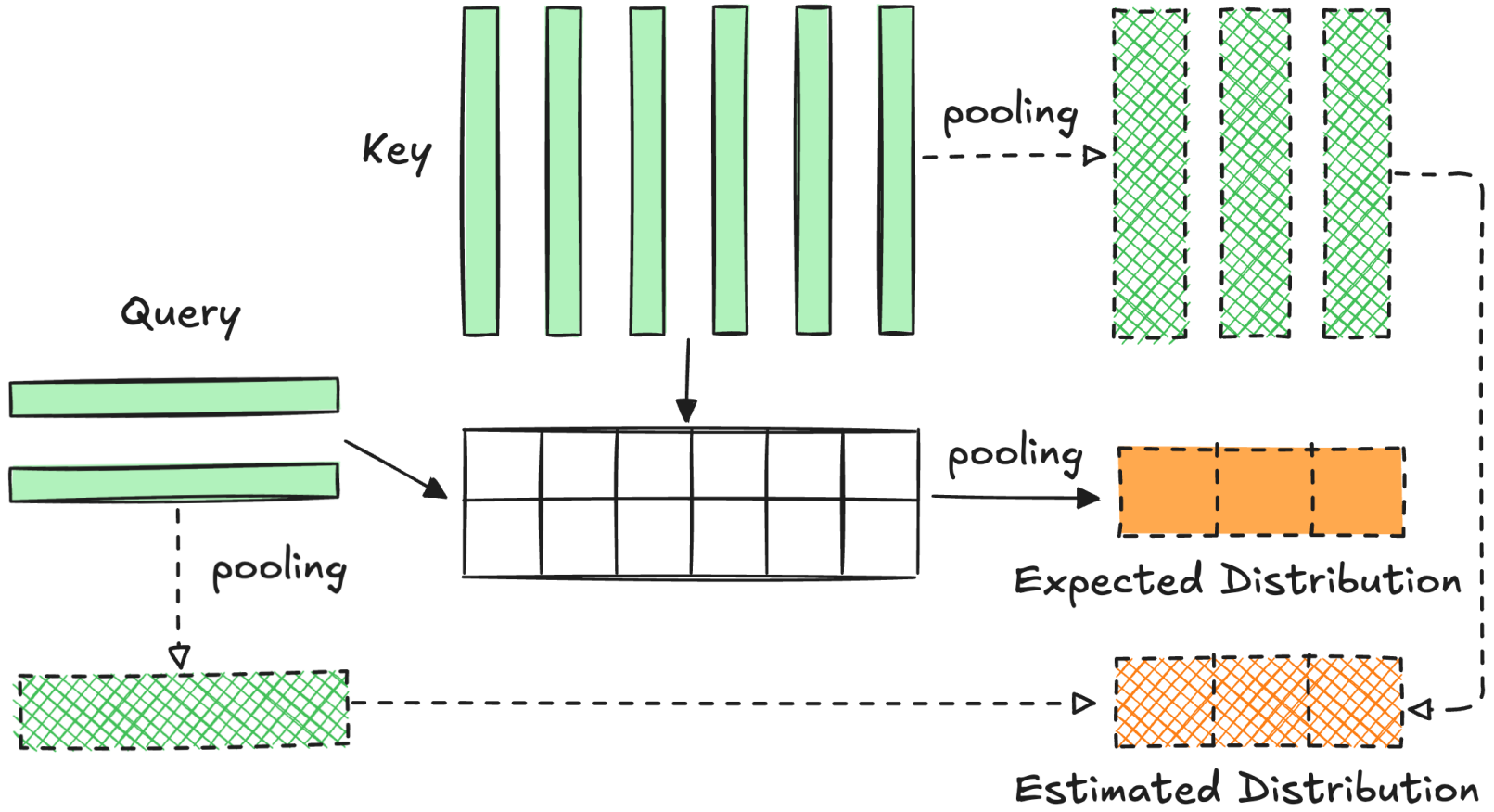
- Use block-wise attention score approximation to identify important key-value blocks for each query block.
- If the approximation is inaccurate, apply a predefined vertical-slash pattern.

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025

Algorithm – 1. Sparse Pattern Determination

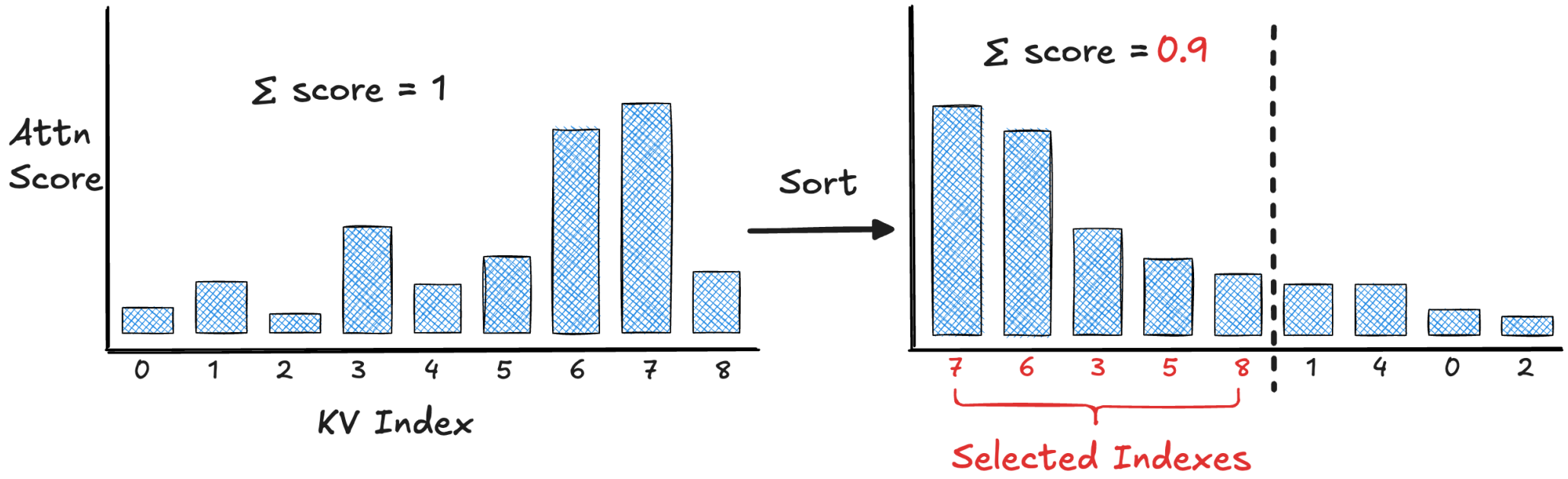
Compare true and estimated attention distributions using JS divergence.

If $D_{JS} < \tau$, use Query-Aware pattern, else fallback to Vertical-Slash pattern.



Algorithm – 2. Sparse Index Selection

- Define a threshold γ for cumulative attention scores.
- Select tokens with the highest scores until the threshold is reached.

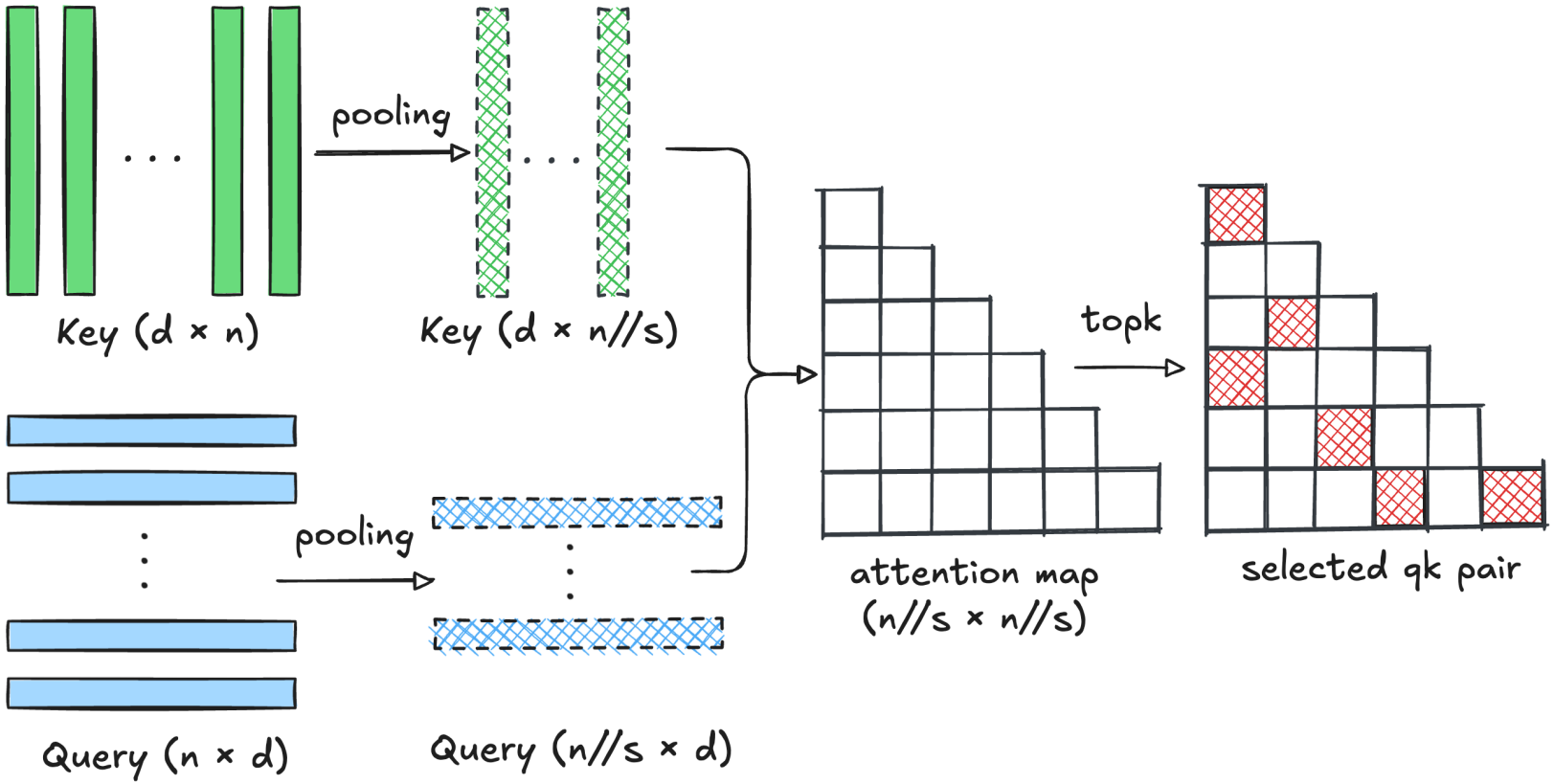


Algorithm – 2. Sparse Index Selection

Query-Aware Pattern:

1. Compute a pooled attention map.

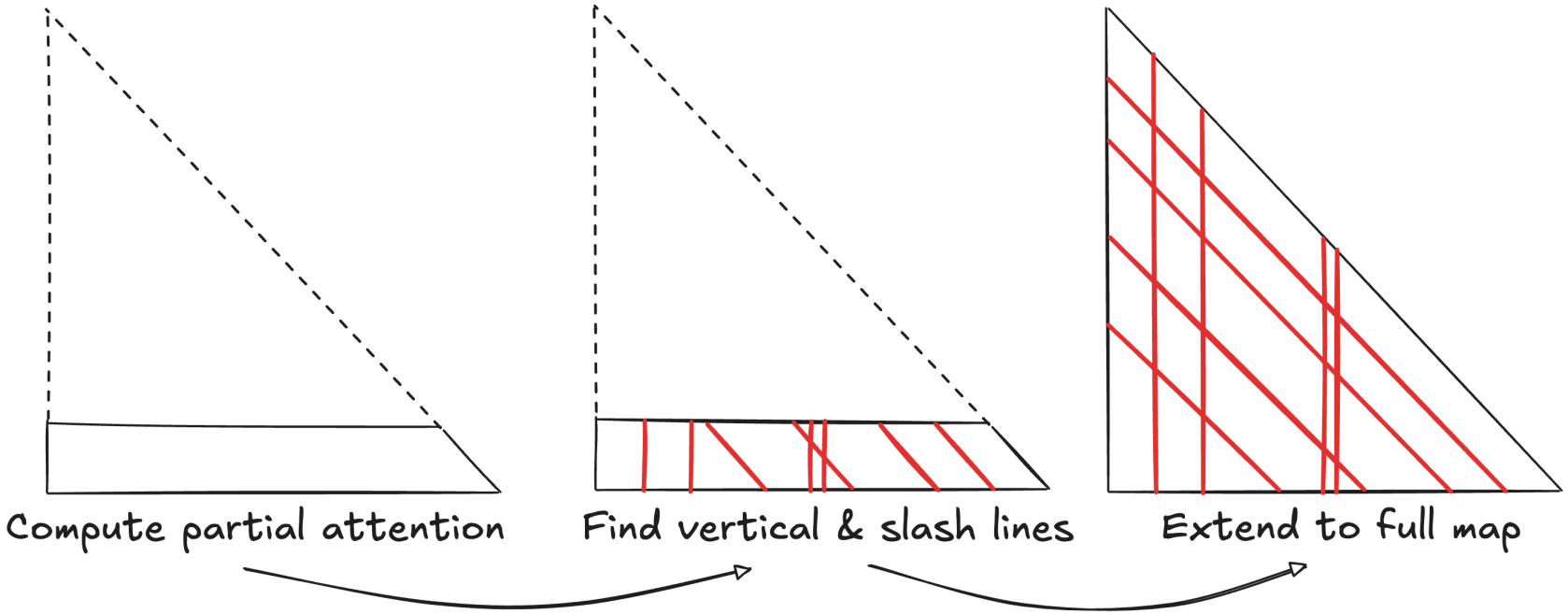
2. Select important KV blocks for each query block.



Algorithm – 2. Sparse Index Selection

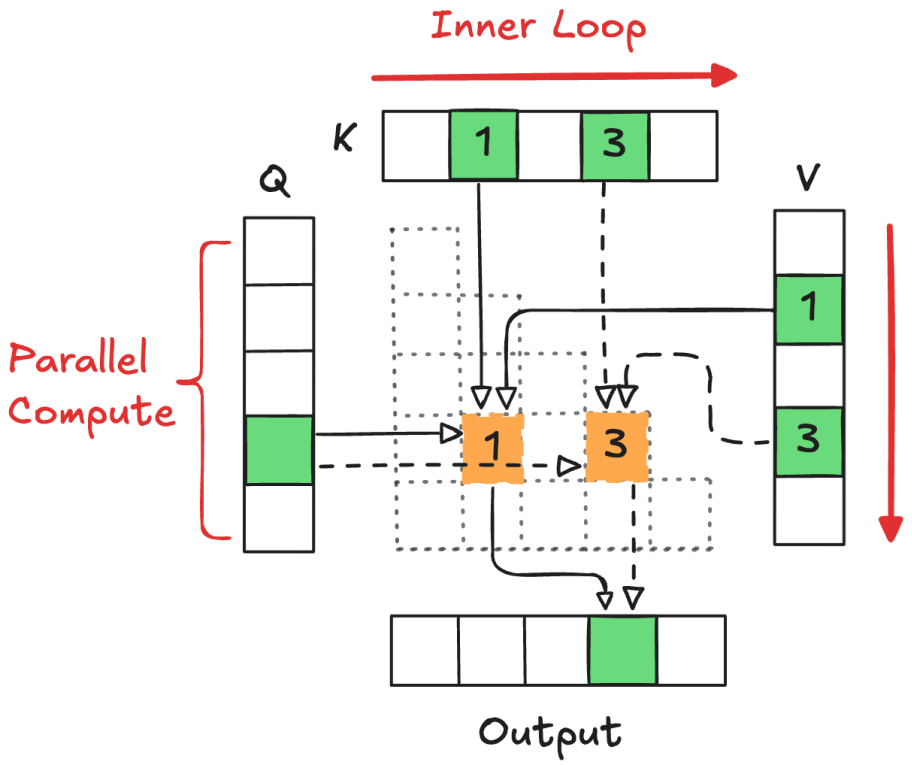
Vertical-Slash Pattern:

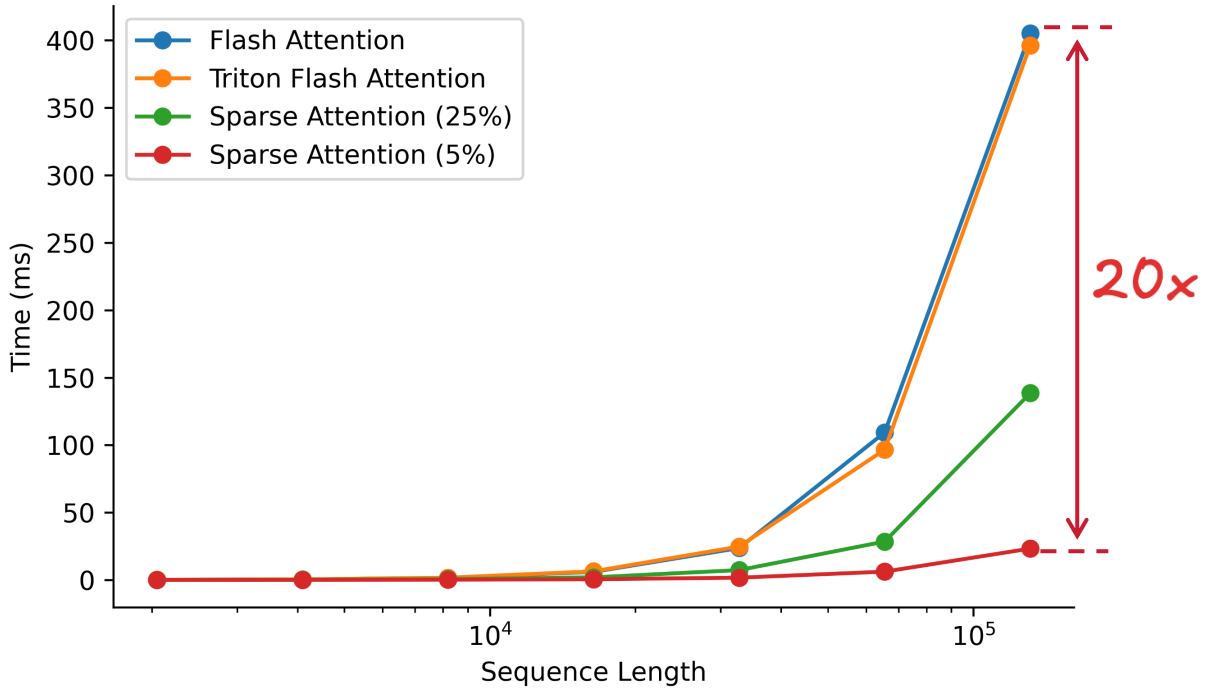
1. Use a subset of query vectors to identify important vertical and slash lines.
2. Extend to the full attention map to get QK pairs.



Algorithm – 3. Sparse Attention Calculation

- Based on FlashAttention [1], we only calculate attention over selected KV blocks
- Implement using Triton [2] and achieve expected acceleration ratio





[1] Dao, Tri. "Flashattention-2: Faster attention with better parallelism and work partitioning." *arXiv preprint arXiv:2307.08691* (2023).
[2] https://github.com/triton-lang/triton

Experiments - Performance

Results on RULER:

- FlexPrefill consistently preserves models' performance across multiple context lengths.

Models	Methods	4k	8k	16k	32k	64k	128k	Avg
LLaMA	Full-attn	95.67	93.75	93.03	87.26	84.37	78.13	88.70
	Streaming LLM	95.43	**93.99**	74.76	48.56	26.20	30.77	61.62
	MInference	**95.67**	**93.99**	93.27	86.54	**84.86**	58.17	85.42
	Ours	95.43	93.51	**94.71**	**89.42**	82.93	**79.09**	**89.18**
GLM	Full-attn	93.75	93.03	89.66	90.63	85.34	81.97	89.06
	Streaming LLM	**93.75**	92.79	76.92	56.97	40.14	34.86	65.91
	MInference	**93.75**	**93.03**	**90.38**	89.90	85.10	82.93	89.18
	Ours	93.51	91.83	89.90	**91.35**	**86.06**	**83.41**	**89.34**
Yi	Full-attn	92.79	86.06	85.34	76.93	69.47	66.35	79.49
	Streaming LLM	93.03	83.41	65.38	51.68	39.18	34.61	61.22
	MInference	92.79	85.58	82.69	**73.32**	63.94	**57.69**	76.00
	Ours	**93.27**	**86.78**	**83.89**	72.36	**64.66**	56.97	**76.32**
Qwen	Full-attn	89.90	88.70	80.77	79.33	56.49	17.79	68.83
	Streaming LLM	90.14	88.94	57.93	43.03	25.48	12.26	52.96
	MInference	89.90	88.70	79.33	78.61	49.04	10.82	66.07
	Ours	**90.39**	**89.91**	**83.17**	**81.25**	**59.14**	**20.67**	**70.75**

Experiments - Performance

Results on InfiniteBench:

- FlexPrefill outperforms baselines across multiple models and tasks.

Models	Methods	En.Sum	En.QA	En.MC	En.Dia	Zh.QA	Code.Debug	Math.Find	Retr.PassKey	Retr.Number	Retr.KV	Avg
LLaMA	Full-attn	31.91	25.92	69.43	21.50	31.95	16.75	24.29	99.15	99.66	60.00	48.06
	Streaming LLM	30.15	10.15	41.05	8.50	22.38	8.63	17.71	2.71	5.93	0.00	14.72
	Minference	31.04	22.00	63.76	14.50	28.70	5.33	27.43	56.78	77.12	14.00	34.06
	Ours	**31.82**	**24.82**	**69.43**	**19.50**	**35.46**	**16.75**	**31.14**	**98.64**	**99.83**	**44.00**	**47.14**
GLM	Full-attn	28.60	20.70	42.36	35.00	15.69	32.99	26.29	99.15	100.00	19.00	41.98
	Streaming LLM	28.15	11.58	31.00	18.50	13.77	26.65	20.29	15.08	100.00	6.60	27.16
	Minference	27.27	19.42	**44.98**	**29.50**	**15.83**	**36.29**	**23.71**	**100.00**	100.00	48.60	44.56
	Ours	**28.90**	**20.27**	42.36	**29.50**	15.55	33.50	**23.71**	99.15	100.00	**55.40**	**44.83**
Yi	Full-attn	8.32	11.47	65.50	1.50	17.49	20.81	23.14	100.00	99.66	29.00	37.69
	Streaming LLM	6.01	11.21	42.79	**3.50**	17.15	19.54	22.86	10.17	36.10	3.60	17.29
	Minference	6.06	10.27	62.45	2.00	**17.74**	**24.37**	**24.86**	**100.00**	**97.63**	**29.00**	**37.44**
	Ours	**6.62**	**11.89**	**65.07**	2.50	16.87	19.54	24.00	99.83	91.53	24.00	36.18
Qwen	Full-attn	4.65	5.57	34.50	9.00	11.27	24.87	24.57	95.42	75.42	0.00	29.53
	Streaming LLM	**18.54**	6.43	**39.30**	12.50	10.94	**24.11**	28.00	30.85	61.69	0.00	23.24
	Minference	7.45	3.94	14.85	4.50	10.17	13.20	30.00	83.90	58.47	0.00	22.65
	Ours	14.27	**6.55**	34.06	**13.50**	**11.51**	20.81	**30.86**	**96.95**	**72.20**	0.00	**30.07**

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025

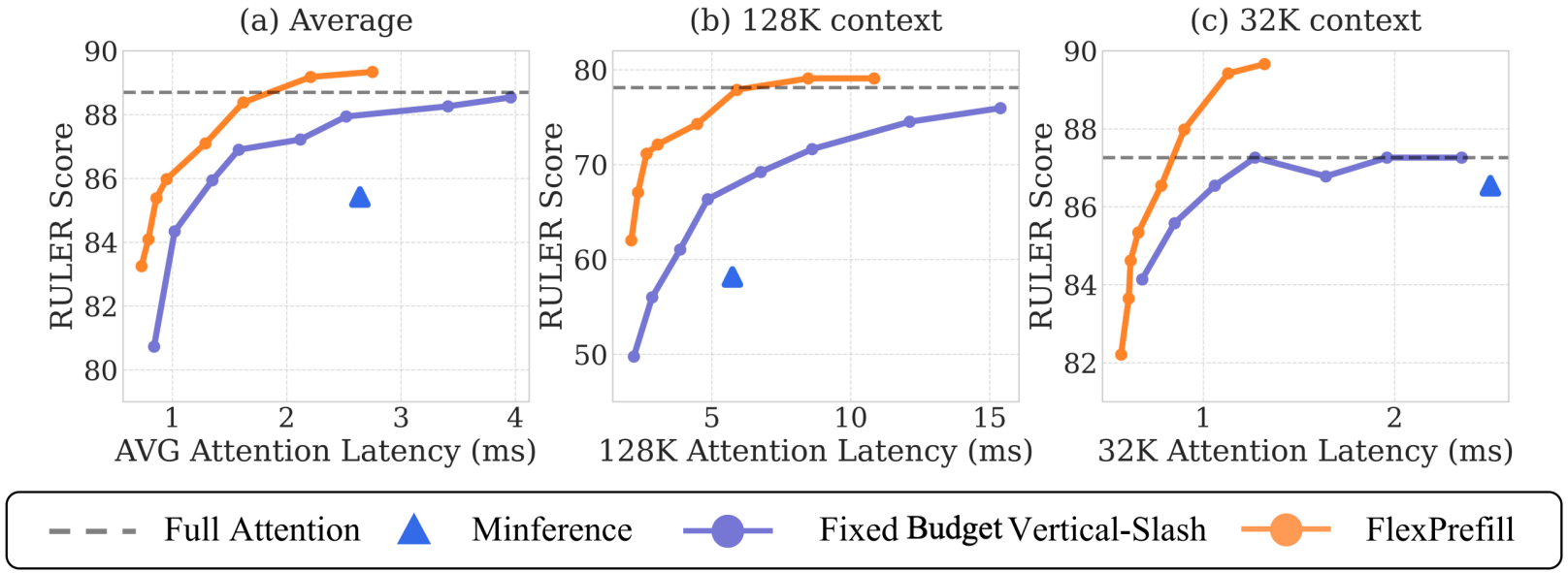
Experiments - Latency

- Tune the threshold γ to balance speed and performance.
- Achieves high acceleration ratios for pre-filling while maintaining performance.

γ	4k	8k	16k	32k	64k	128k	Avg	128k speedup
1 (Full-Attn)	95.67	93.75	93.03	87.26	84.37	78.13	88.70	-
0.97	95.43	93.99	94.47	89.66	83.41	79.09	89.34	1.89x
0.95	95.43	93.51	94.71	89.42	82.93	79.09	89.18	2.43x
0.9	95.43	93.27	94.23	87.98	81.49	77.89	88.38	3.49x
0.85	95.91	93.51	92.31	86.54	80.05	74.28	87.10	4.60x
0.8	96.39	92.79	92.31	87.74	80.05	71.39	86.78	5.20x
0.75	95.19	93.51	92.31	86.06	78.13	70.43	85.94	6.75x
0.7	95.67	93.03	91.35	84.62	76.44	71.15	85.38	7.75
0.65	95.91	92.55	91.35	83.65	74.04	67.07	84.09	8.79x
0.6	95.67	93.27	91.83	82.21	74.52	62.02	83.25	9.81x

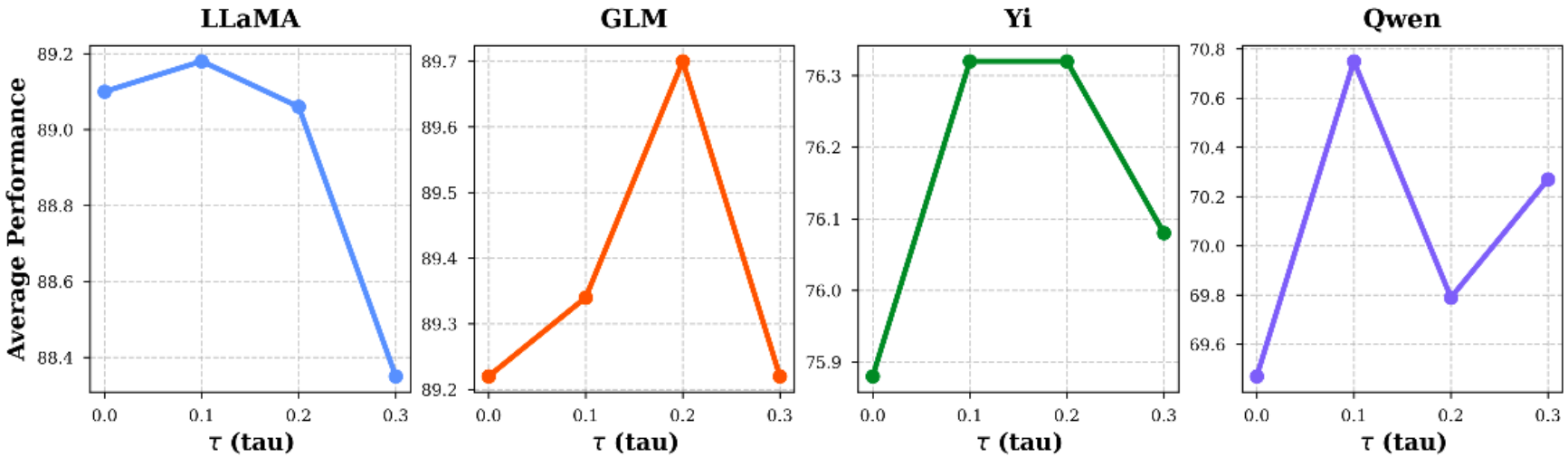
Ablation Study - Dynamic Allocation

- Dynamic allocation improves performance and balances inference speed with effectiveness.



Ablation Study - Impact of τ:

- Setting an appropriate τ enhances performance.
- Overly large τ will misclassify some heads as Query-Aware pattern, leading to degraded performance.



Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025

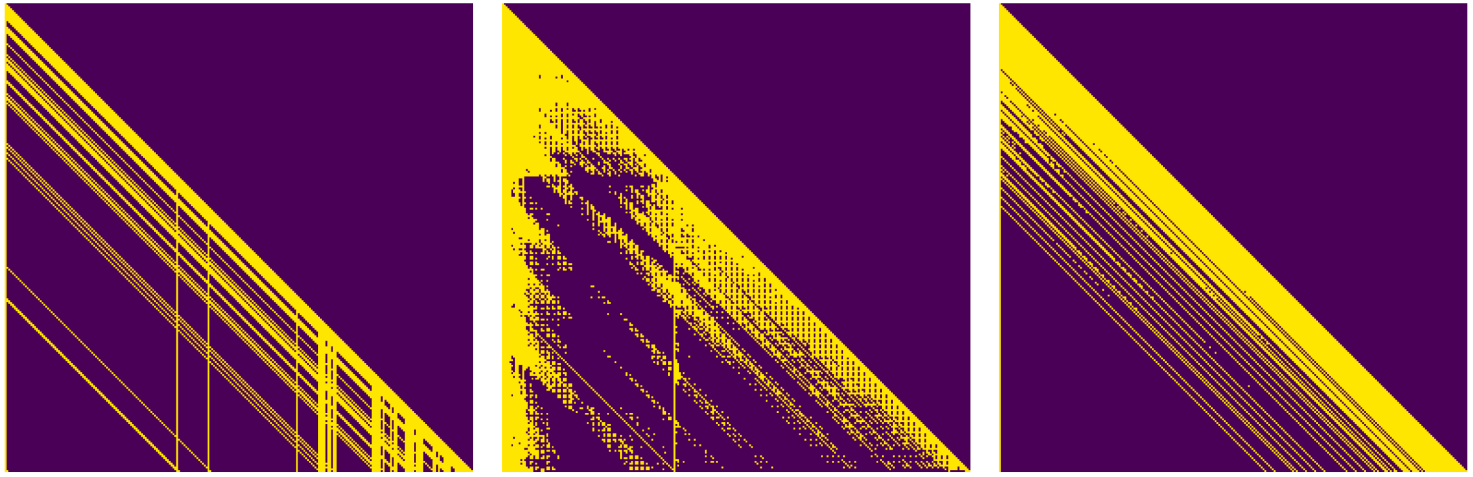
Visualization Insights

- Sparse Masks:

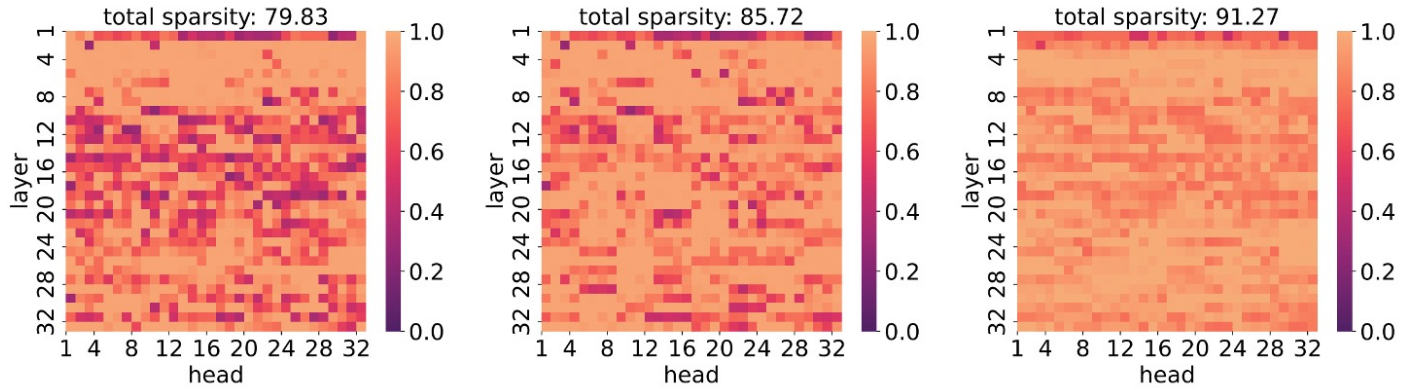
Query-Aware pattern capture scattered blocks, some resembling vertical-slash lines.

- Sparsity Variations:

Different tasks and context lengths yield distinct sparsity ratios.



(a) *Vertical-Slash* pattern (b) *Query-Aware* pattern (c) *Query-Aware* pattern



(a) 64k multi needle (b) 64k single needle (c) 256k multi needle

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025

Open Source

Code available at https://github.com/bytedance/FlexPrefill

Accelerate your long-context LLMs with just a few lines of code:

```
from transformers import AutoModelForCausalLM, AutoTokenizer
+ from flex_prefill import patch_model

tokenizer = AutoTokenizer.from_pretrained("meta-llama/Llama-3.1-8B-Instruct")
model = AutoModelForCausalLM.from_pretrained(
    "meta-llama/Llama-3.1-8B-Instruct",
    torch_dtype=torch.bfloat16,
    _attn_implementation="flash_attention_2",
).cuda()

+ flex_prefill_config =  {"block_size": 128, "flex_prefill_gamma": 0.9, "flex_prefill_tau": 0.1}
+ patch_model(model, "flex_prefill", flex_prefill_config)

input_ids = tokenizer(prompt, return_tensors="pt", return_attention_mask=False).input_ids.cuda()
output_ids = model.generate(input_ids, max_new_tokens=64)
output = tokenizer.decode(output_ids[0], skip_special_tokens=True)
```

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025

Thanks!

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, Xun Zhou. "FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference." ICLR 2025