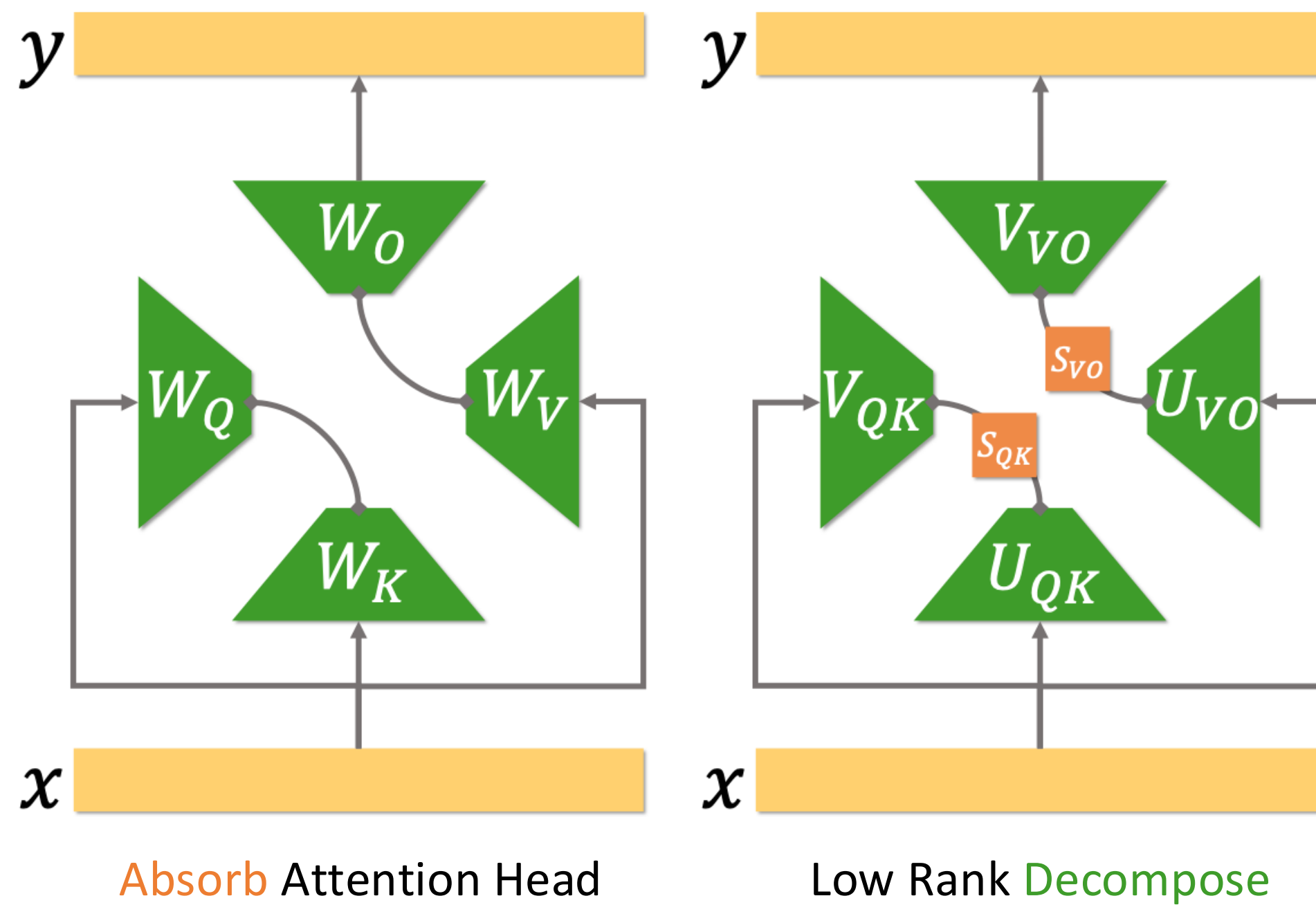




Cross-Layer Orthogonal Vectors



Attention:

$$\text{attn}(Q_h, K_h) = \text{softmax} \left(\frac{Q_h K_h^\top}{\sqrt{d}} \right)$$

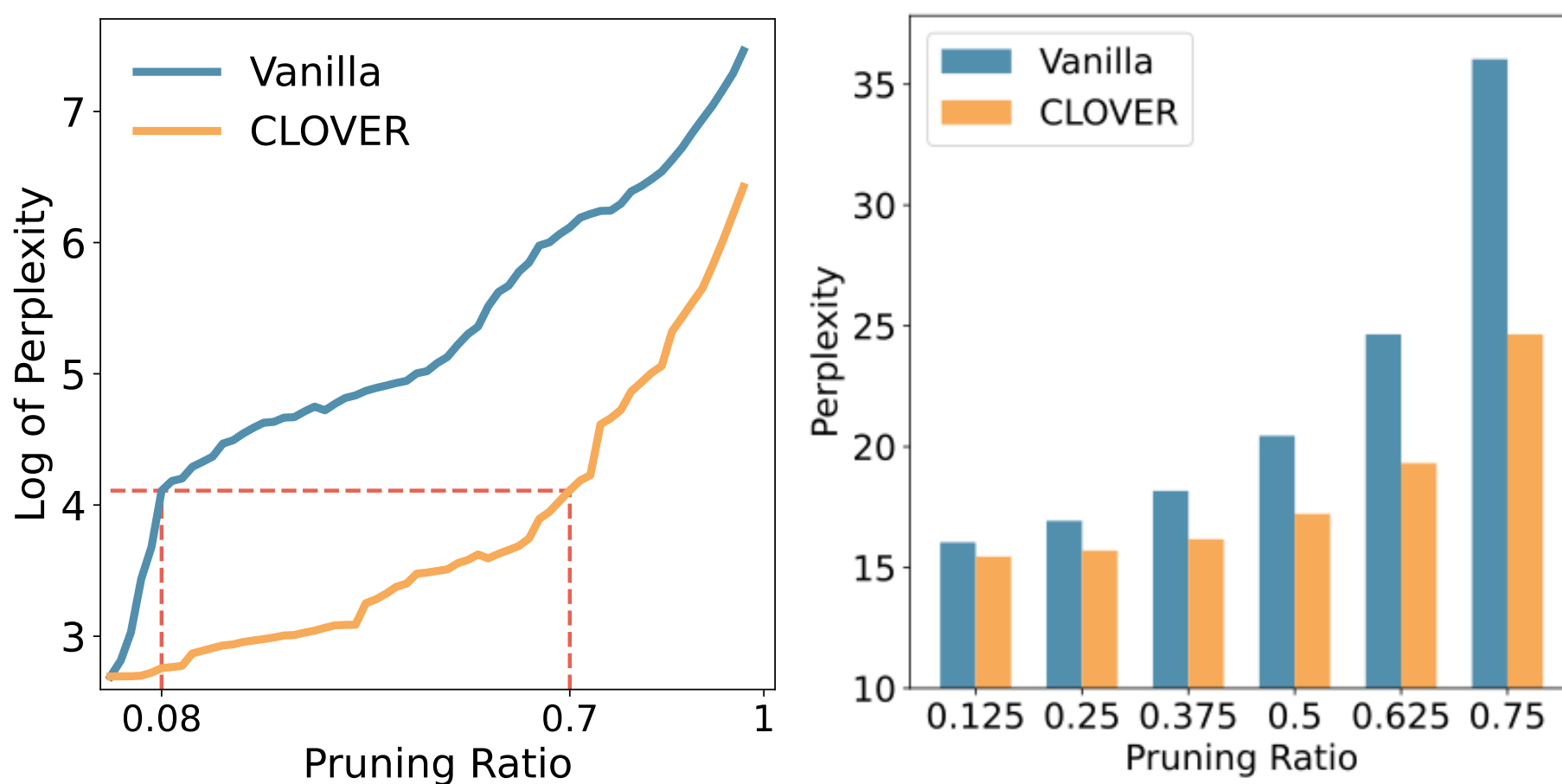
$$= \text{softmax} \left(\frac{XW_Q^{[:,h,:]} (XW_K^{[:,h,:]})^\top}{\sqrt{d}} \right).$$

Absorb and Decompose :

$$W_Q^{[:,h,:]} (W_K^{[:,h,:]})^\top = W_{QK}^h = U_{QK}^h S_{QK}^h V_{QK}^h.$$

Guide Pruning !

Large Ratio Pruning



CLOVER Pruning

Mr. Birkett Foster's landscapes smile at one much in the same way that Mr. Carker used to flash his teeth. And, and Mr. John Collier gives his sitter a cheerful slap on the back before he says, like a shampooer in a Turkish bath, next man.

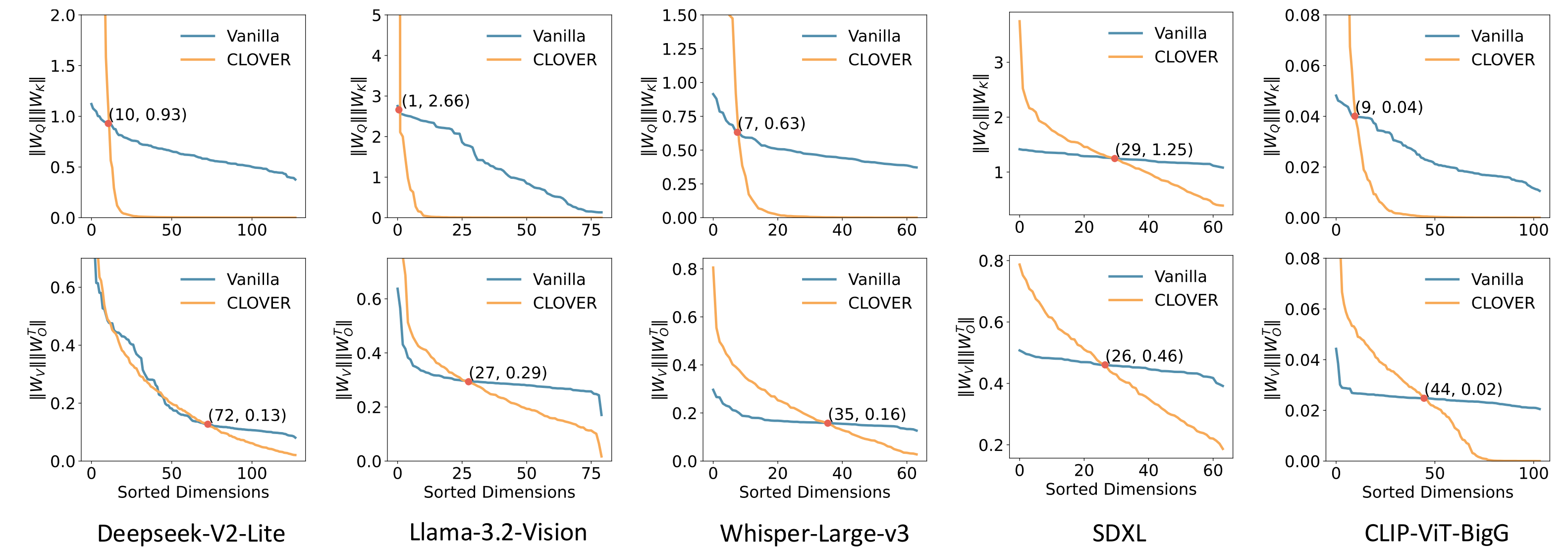
Vanilla Pruning

.....
.....
.....
.....

Full-Rank Finetuning

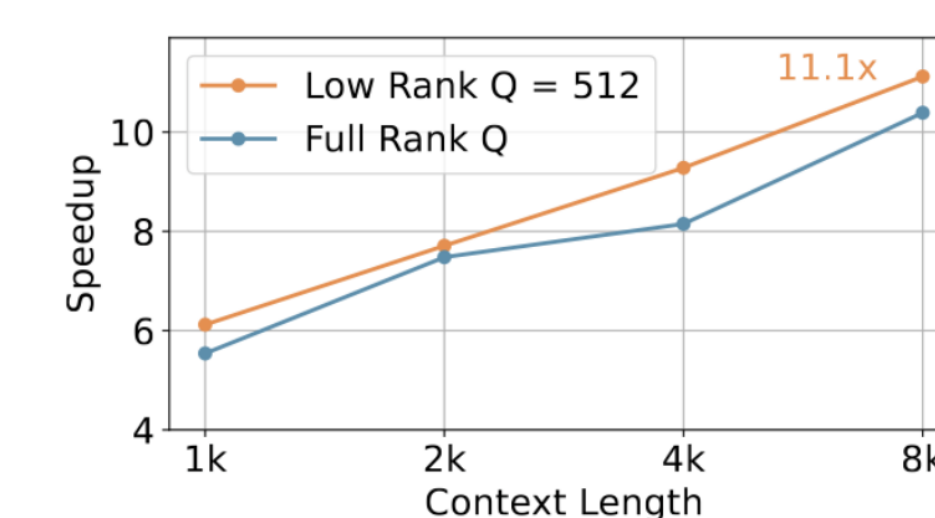
Model	Method	Params	BoolQ	PIQA	SIQA	Hella Swag	Wino Grande	ARC-e	ARC-c	OBQA	Avg.
LLaMA2-7B	LoRA	0.83%	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
	DoRA	0.84%	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
	HiRA	0.83%	71.2	83.4	79.5	88.1	84.0	86.7	73.8	84.6	81.4
	PiSSA	0.83%	75.0	87.0	81.6	95.0	86.5	88.5	75.9	86.4	84.5
	CLOVER	0.83%	75.0	86.4	82.0	95.1	87.5	89.6	76.6	89.4	85.2

Information Aggregation

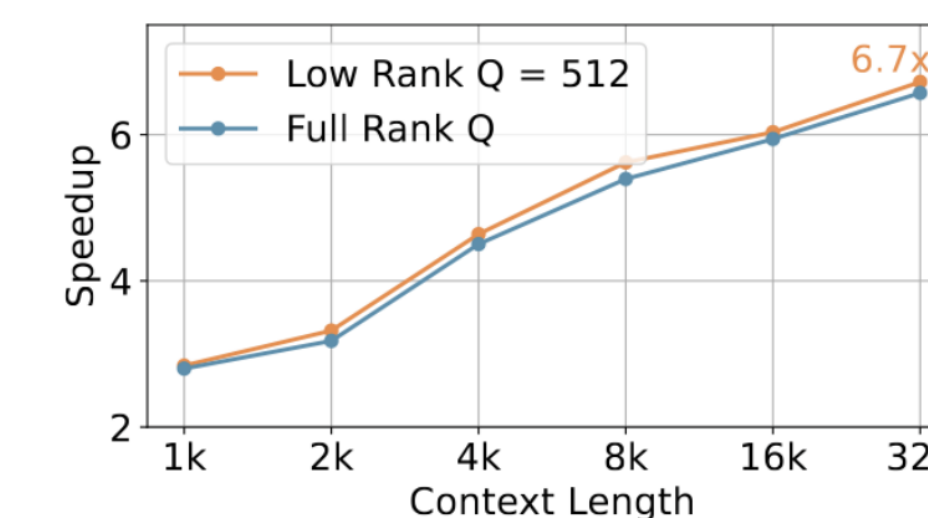


Combined with Other Methods

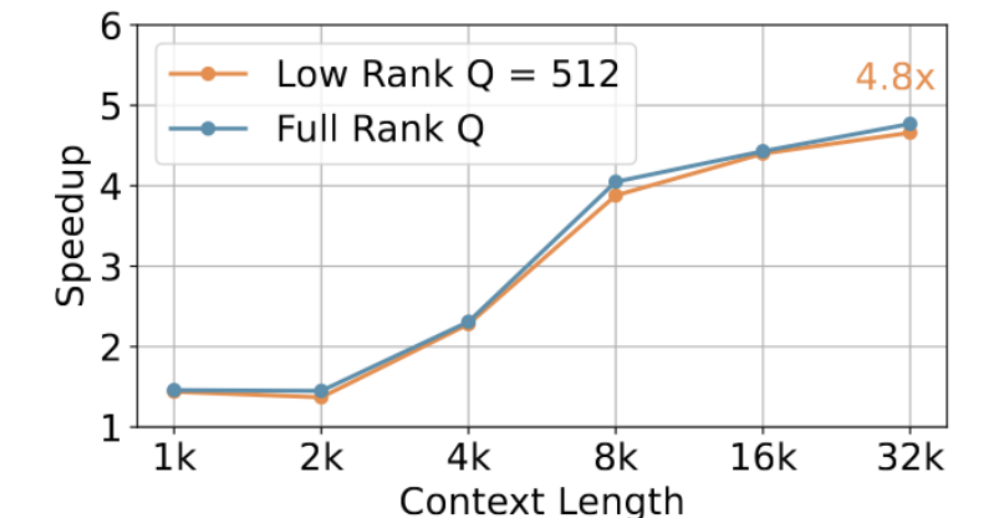
Model	Base	Additional	Avg.	MMLU	ARC	PIQA	HS	OBQA	WG
DS V2 Lite	--	--	61.54	43.29	60.39	79.92	74.51	45.40	65.75
• SliceGPT	-6.25%	--	58.92	40.52	56.94	78.84	70.91	43.20	63.14
	-12.50%	--	58.00	38.70	56.27	78.18	68.73	44.00	62.12
• SliceGPT + CLOVER	-6.25%	-50%	58.17	39.21	55.67	78.56	69.35	43.80	62.43
	-12.50%	-50%	56.32	37.89	55.54	77.58	66.65	39.60	60.69
LLaMA-2-7B	--	--	59.85	41.43	59.24	78.40	73.29	41.80	64.96
• TransMLA	-68.75%	--	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	-92.97%	--	59.19	40.41	58.68	77.53	70.39	45.00	63.14
• TransMLA + CLOVER	-68.75%	-50%	59.40	40.91	58.97	78.35	71.32	43.40	63.46
	-92.97%	-50%	59.13	40.69	60.03	77.09	69.65	45.20	62.12



(a) 165.2 TFLOPS — 24GB



(b) 312 TFLOPS — 40GB



(c) 320 TFLOPS — 64GB

Summary of Contributions

- By decomposing the absorbed head, we orthogonalize the attention head without adding extra transformation matrices.
- This orthogonalization reduces linear redundancy, is compatible with previous pruning method, and allows for higher pruning ratios.
- CLOVER enables full-rank updates, surpassing SOTA methods.

