

QJL: 1-Bit KV cache Quantization



Amir Zandieh

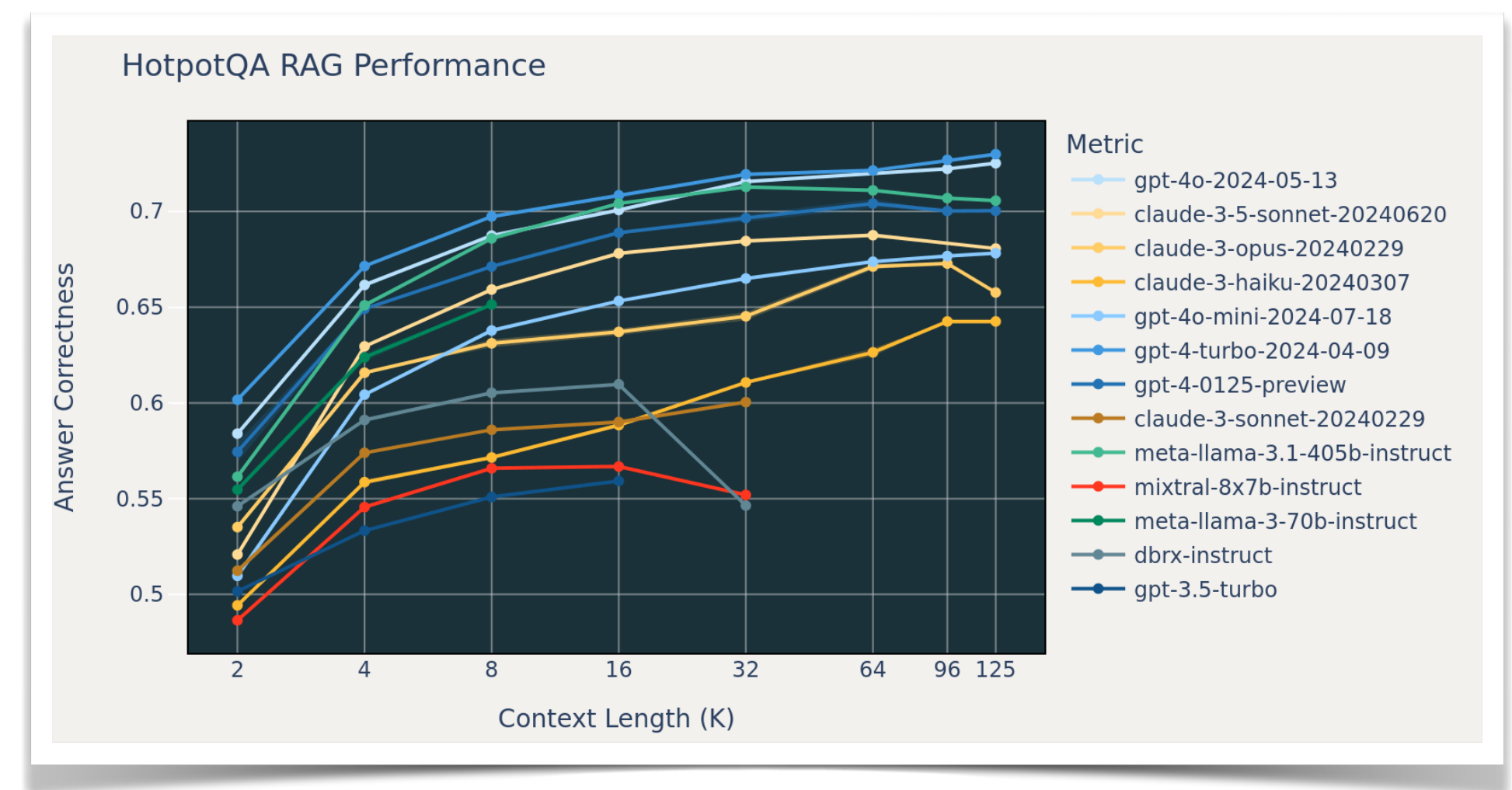
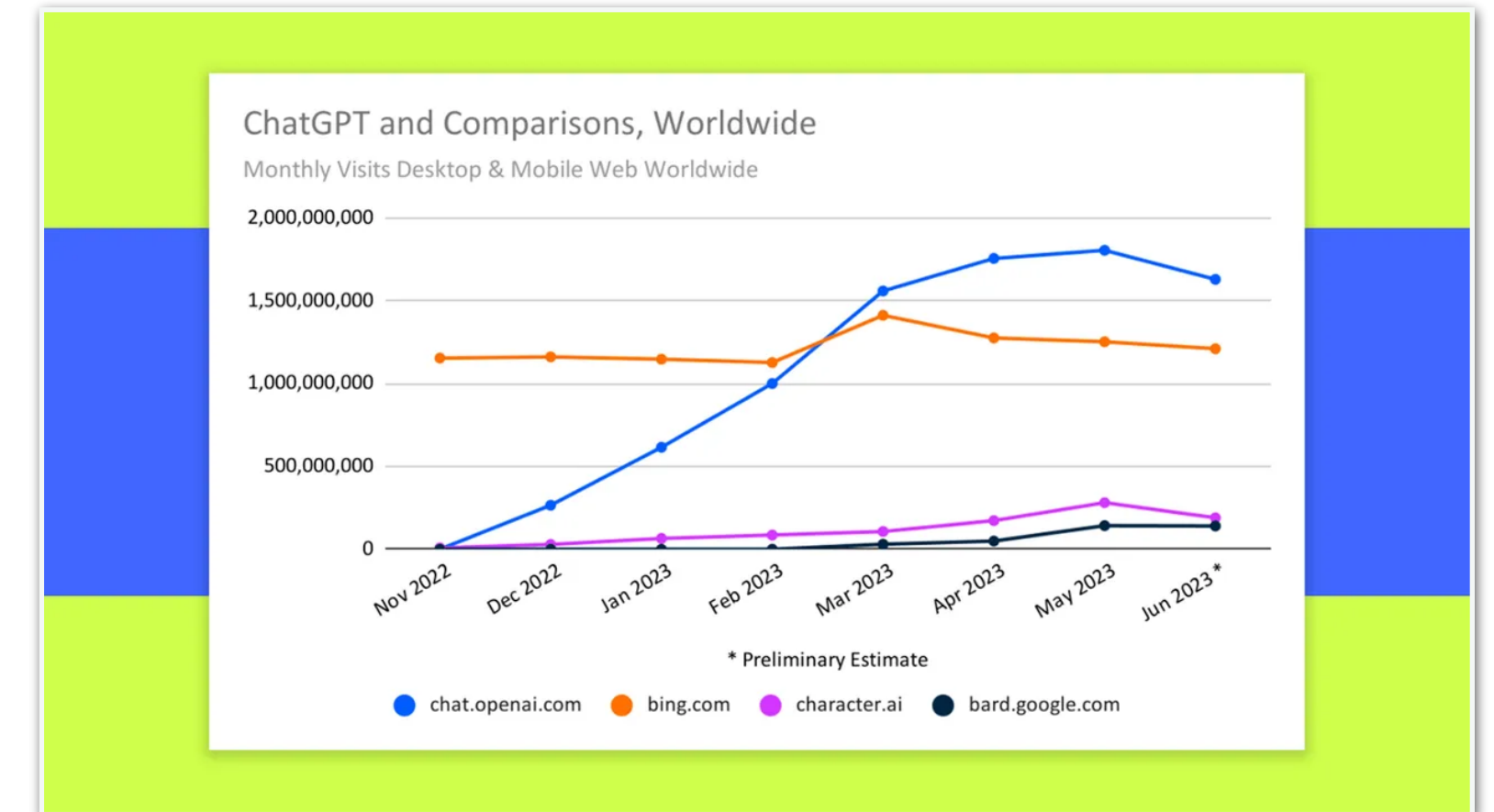
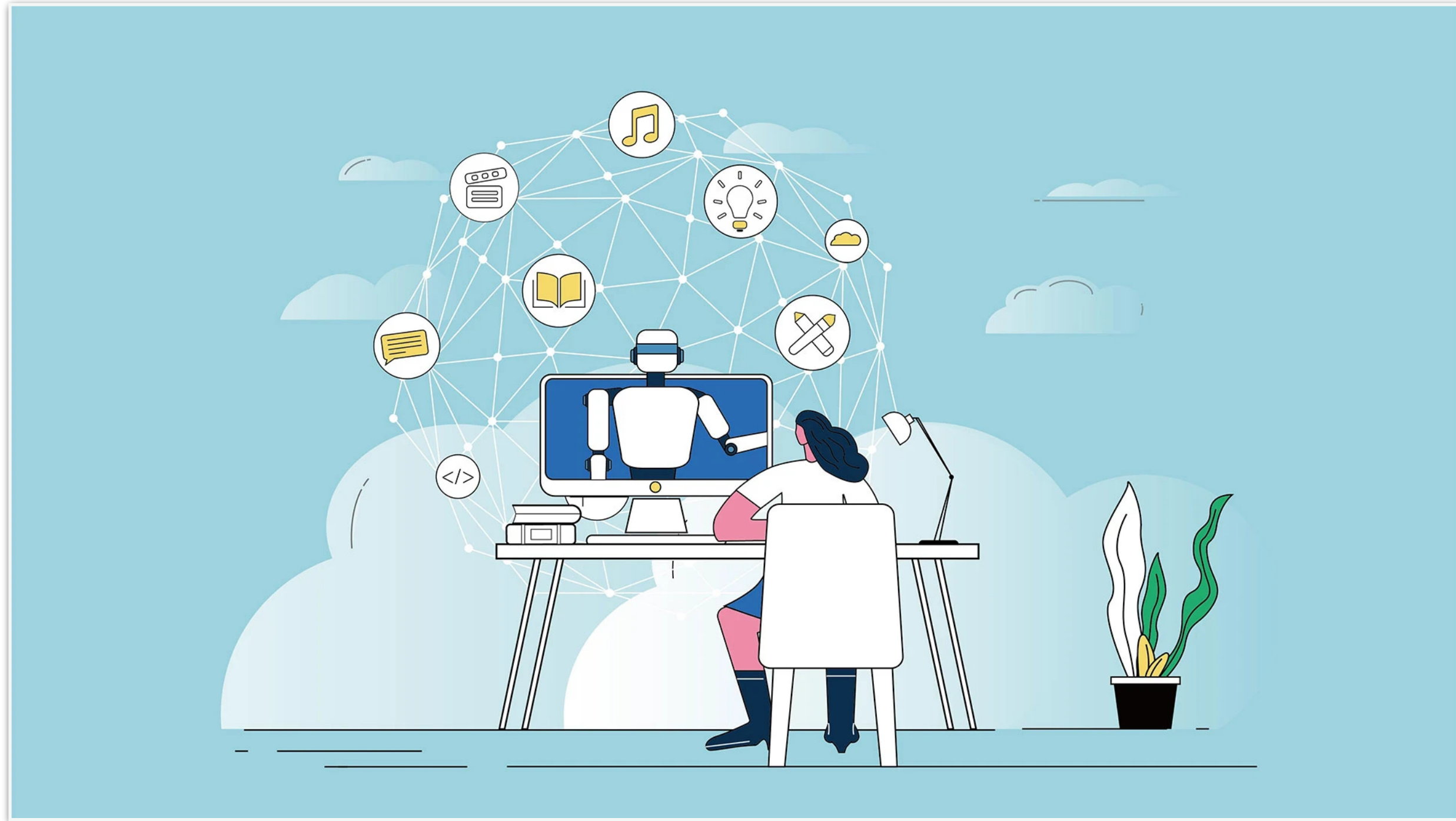
Google



Insu Han

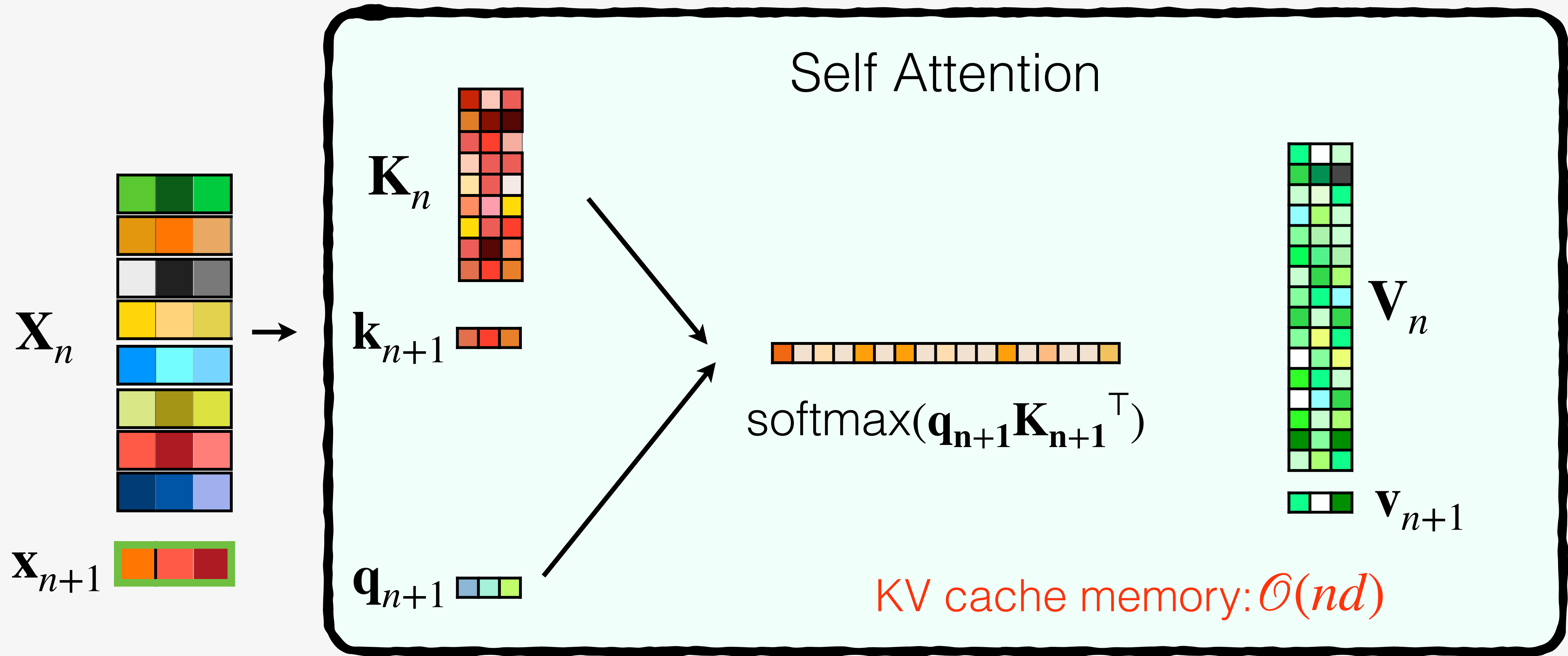
KAIST

AI Revolution

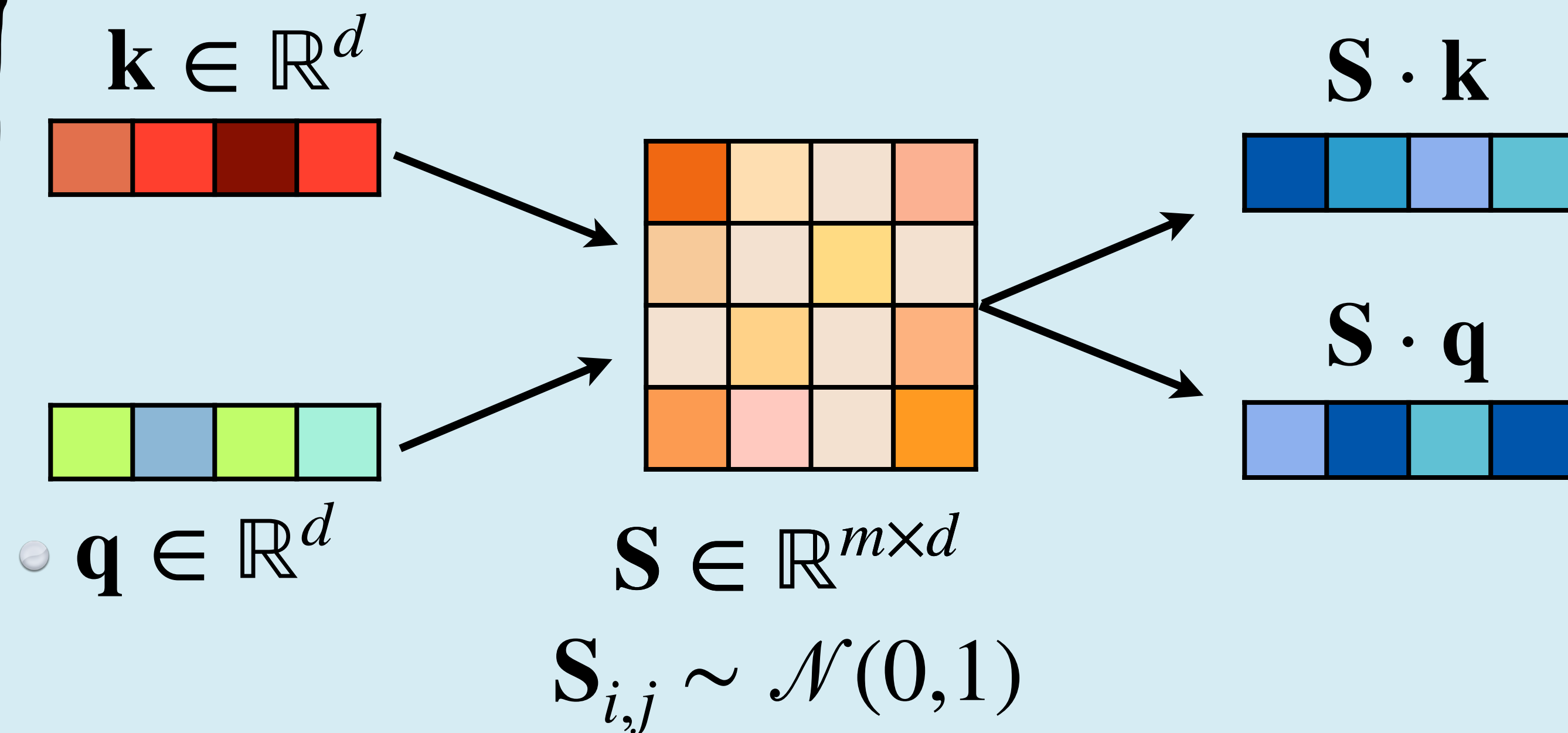


Generation: From Context to Output

Transformer



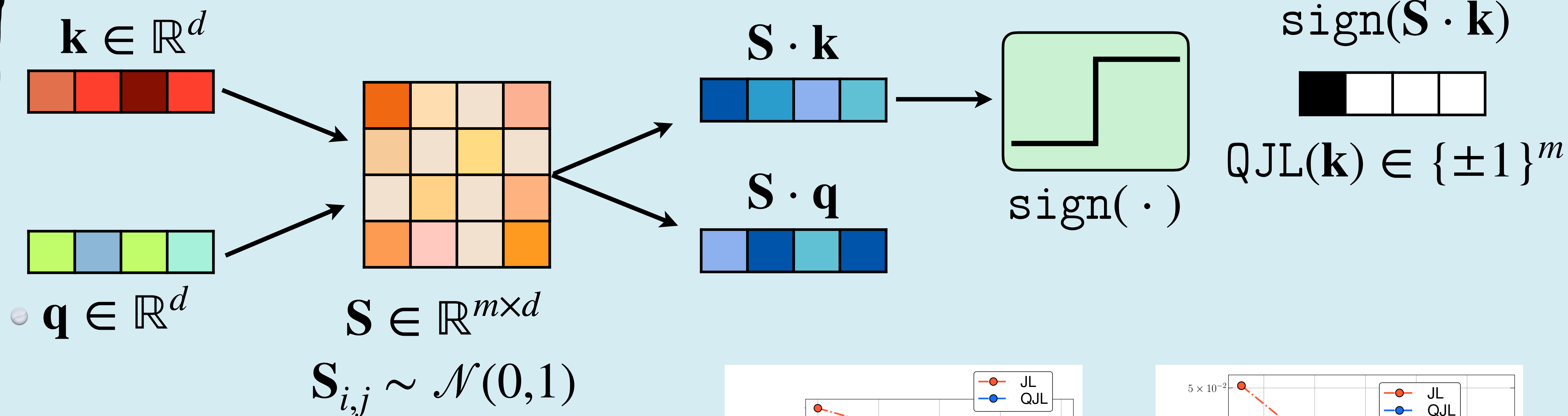
JL Transform



JL Lemma: For $m = \frac{1}{\epsilon^2} \log(\frac{1}{\delta})$ w.p $1 - \delta$

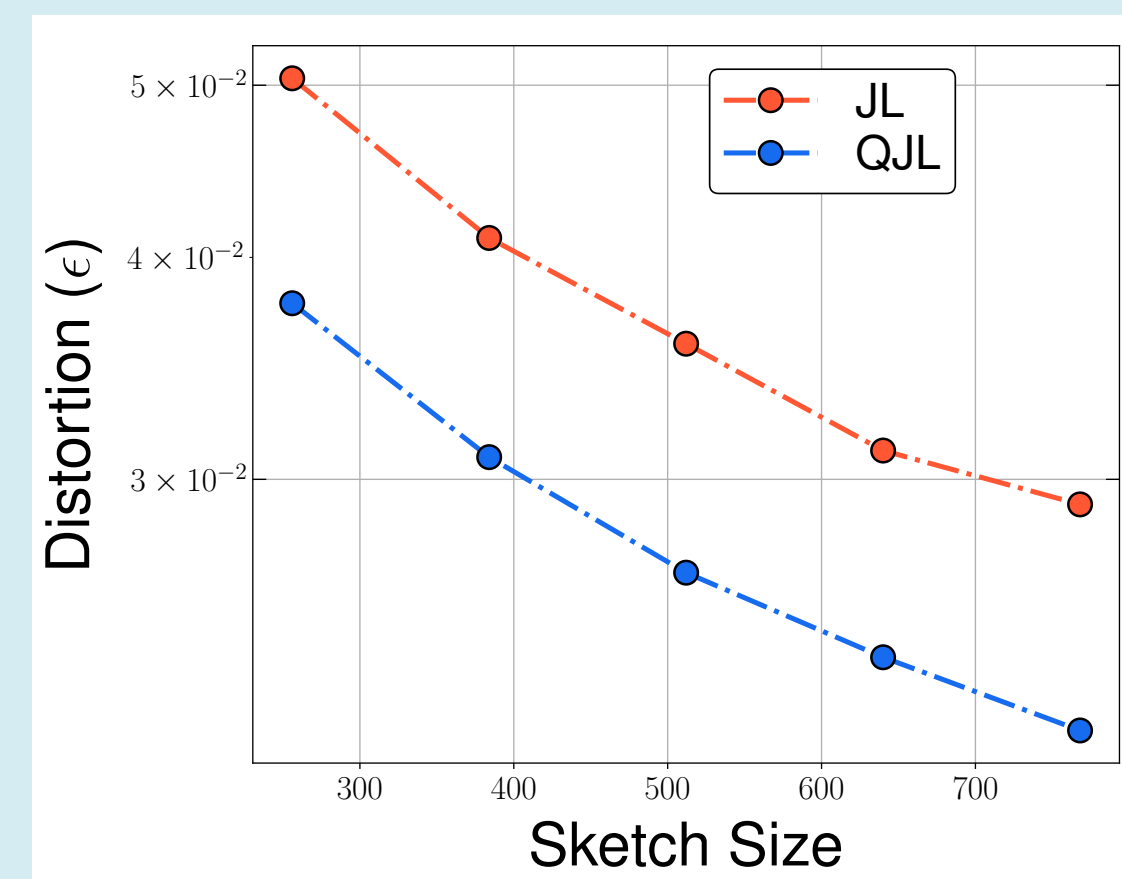
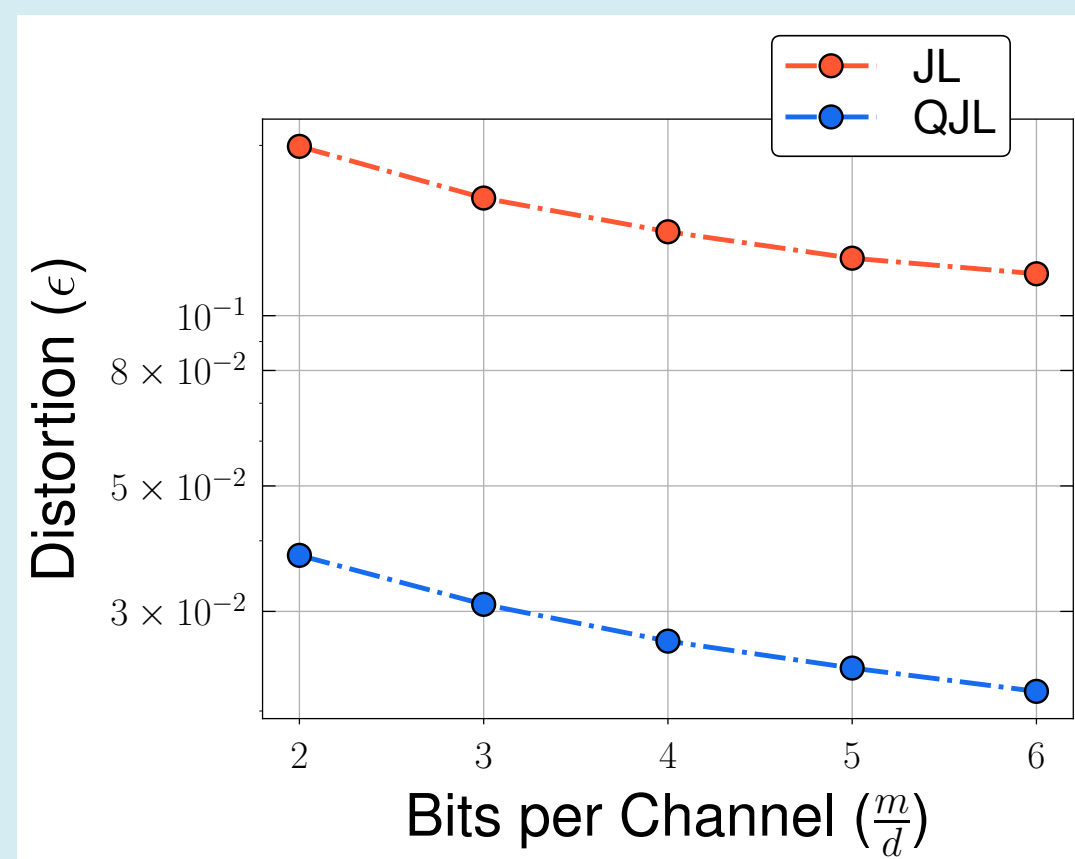
$$\langle \mathbf{S}\mathbf{q}, \mathbf{S}\mathbf{k} \rangle \approx_{\epsilon} \langle \mathbf{q}, \mathbf{k} \rangle$$

QJL Quantize



QJL Lemma: For $m = \frac{1}{\epsilon^2} \log(\frac{1}{\delta})$ w.p $1 - \delta$

$$\langle \mathbf{S}\mathbf{q}, \text{sign}(\mathbf{S}\mathbf{k}) \rangle \approx_{\epsilon} \langle \mathbf{q}, \mathbf{k} \rangle$$



Evaluation

Methods	Bits	Datasets from LongBench (Bai et al. 2023)					
		NarrativeQA	Qasper	MultiQA-en	MultifQA-zh	HotpotQA	2WikiMultiQA
FP16 (Longchat-7b-v1.5-32k)	16	20.79	29.42	42.83	34.33	33.05	24.14
KIVI (Liu et al. 2024b)	3	20.96	29.01	40.93	34.75	32.79	23.01
QJL (ours)	3	20.67	28.48	40.94	29.71	35.62	23.60
KVQuant (Hooper et al. 2024)	4.3	20.14	28.77	44.22	34.44	34.06	23.05
QJL (ours)	4.3	20.72	30.02	41.18	31.73	34.22	22.63
KIVI (Liu et al. 2024b)	5	20.49	28.90	43.24	34.66	33.07	24.86
QJL (ours)	5	21.09	29.11	41.58	31.86	35.65	24.61

Methods	Bits	Datasets from LongBench (Bai et al. 2023)								Avg.
		Qmsum	Qasper	MultiNews	TREC	TriviaQA	SAMSum	LCC	RepoBench-P	
BP16 (Llama-3.1-8B-Instruct)	16	25.22	44.98	26.99	73.5	91.79	43.87	62.99	56.39	53.47
QJL (ours)	3	23.68	42.76	26.82	72.5	88.55	43.43	63.98	56.09	52.48
QJL (ours)	4	23.40	43.09	27.02	72.5	89.69	43.30	64.75	56.79	52.82

