QJL: 1-Bit KV cache Quantization

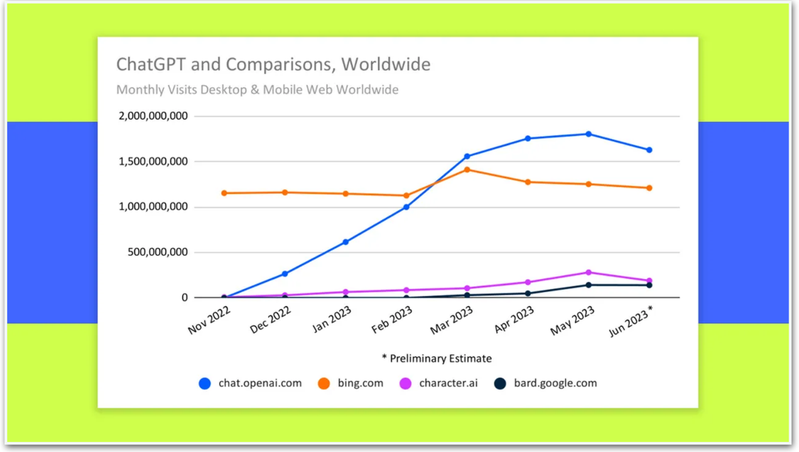
Amir Zandieh

Google

Insu Han

KAIST

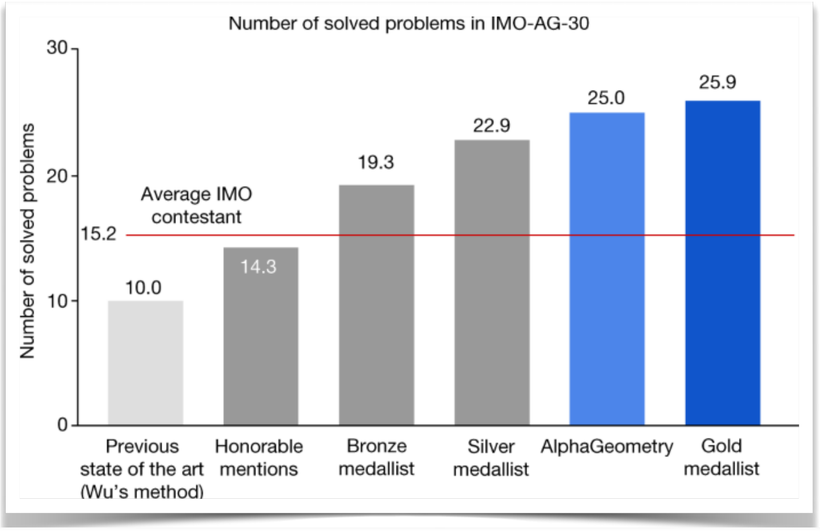
AI Revolution

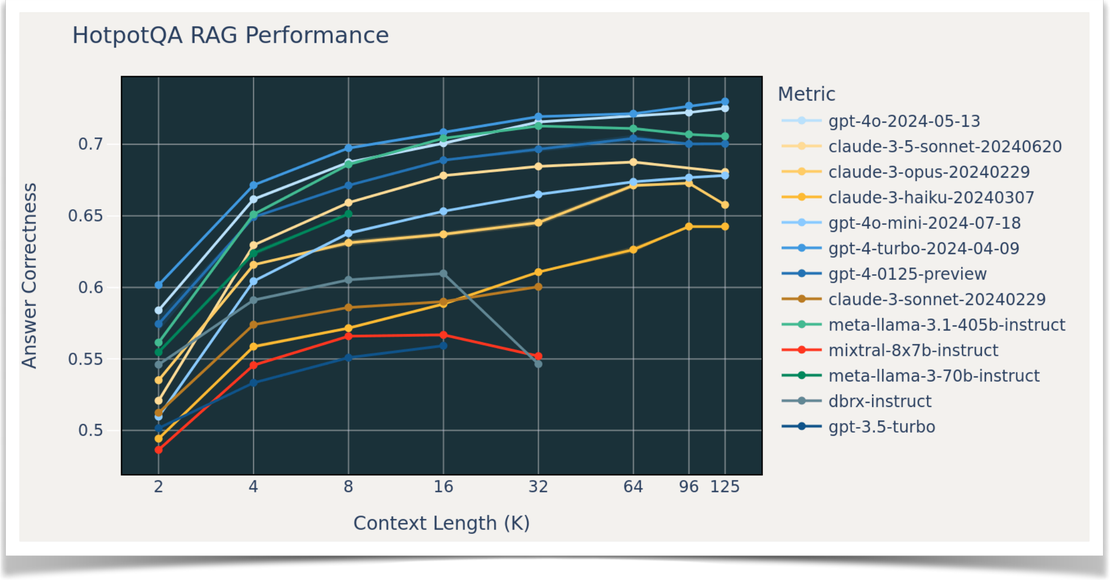


Gemini

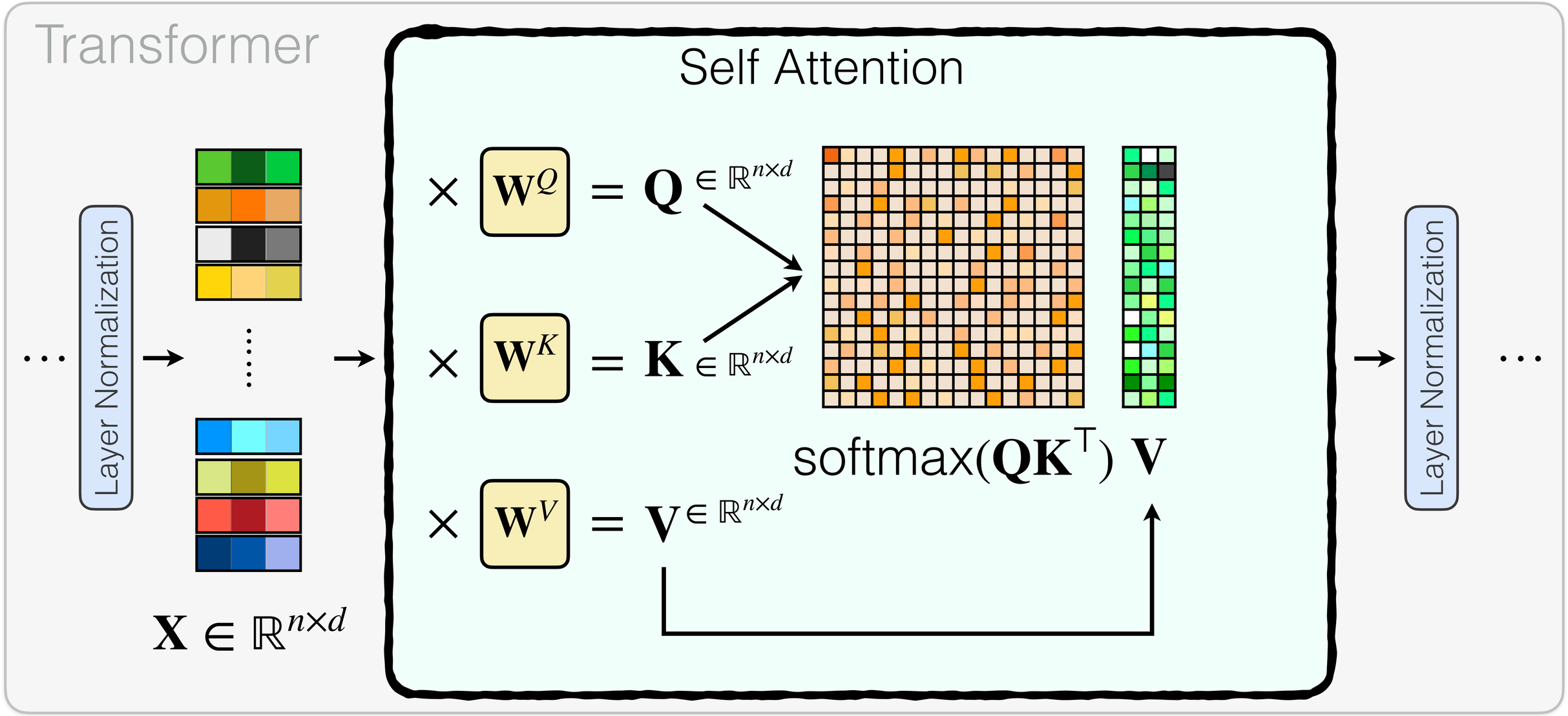
Claude

Long Context LLM

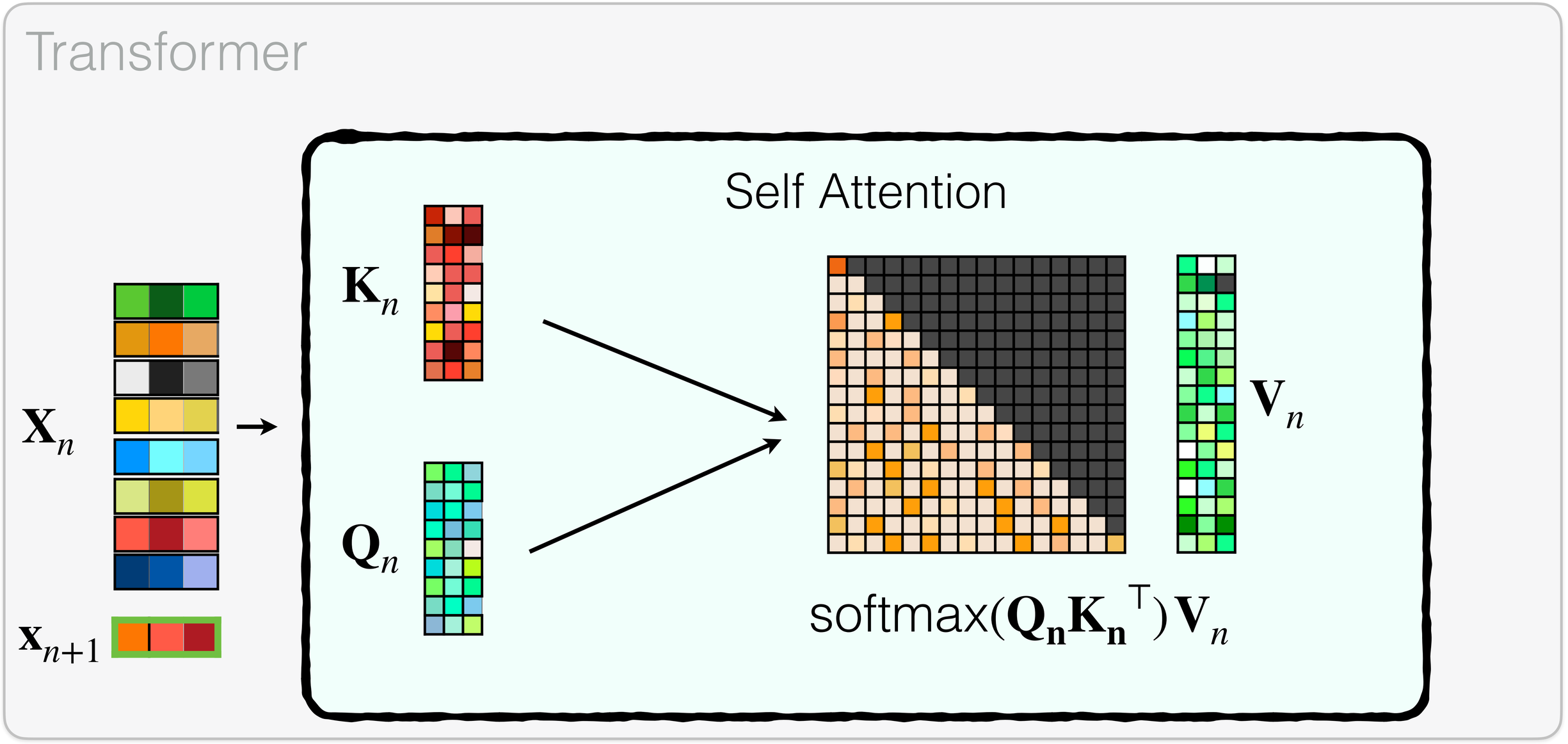




Challenges with Long Contexts

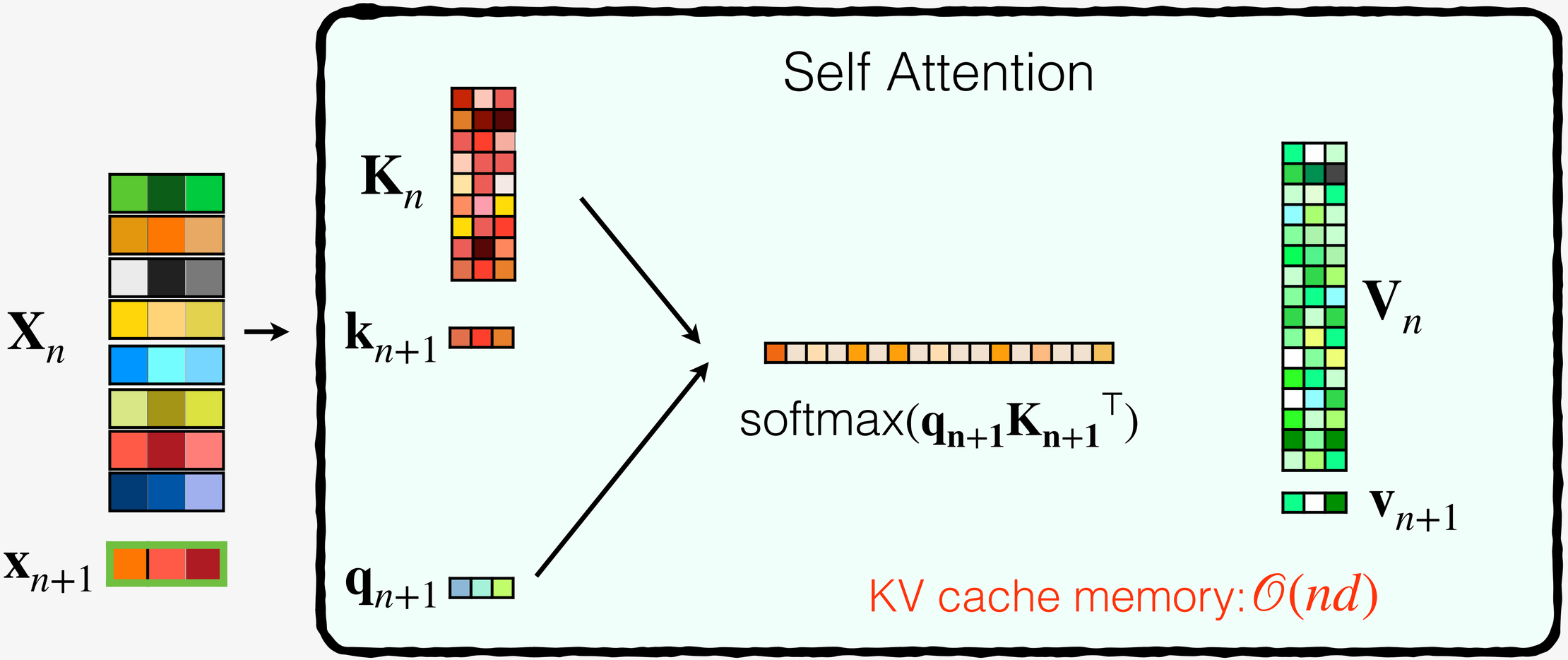


Prefill: Context Prepration

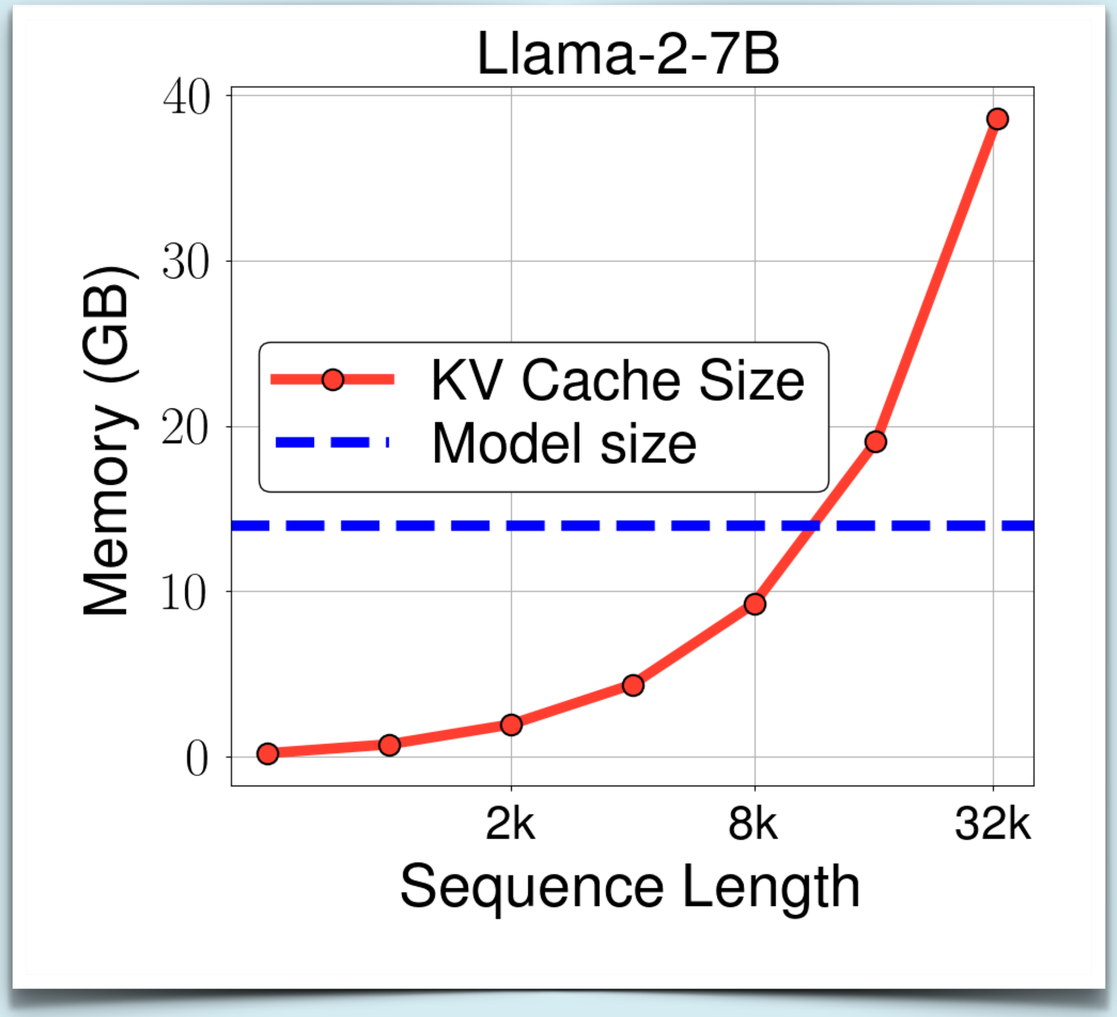


Generation: From Context to Output

Transformer

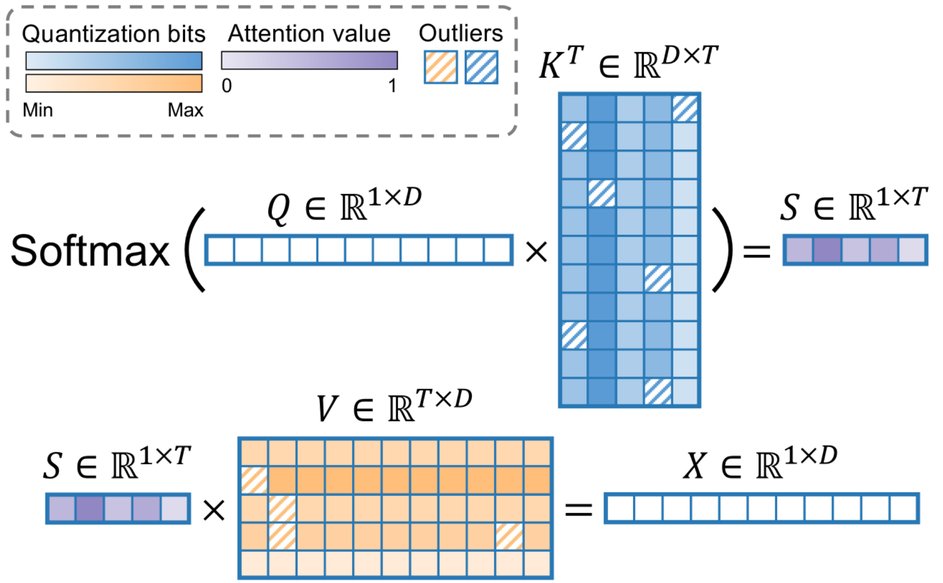


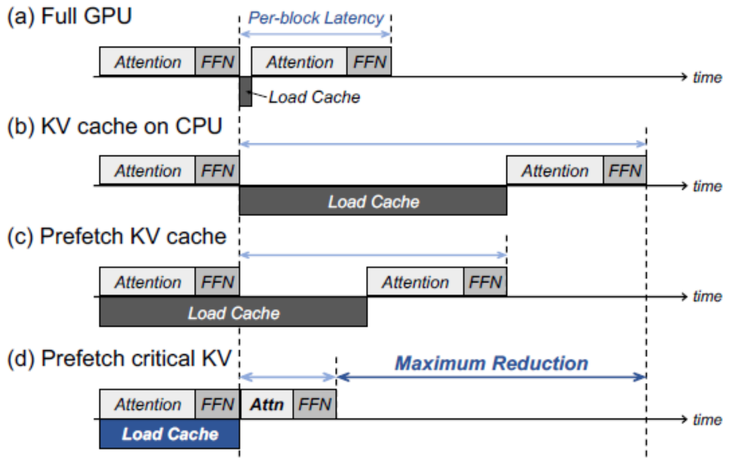
Generation: From Context to Output

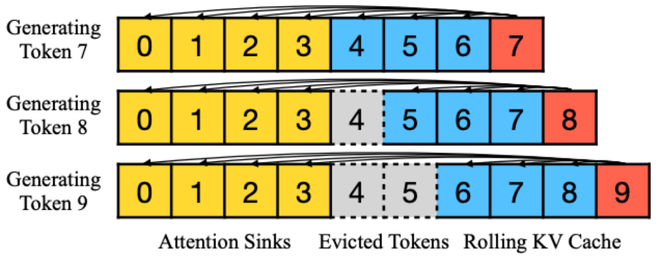


Previous Works

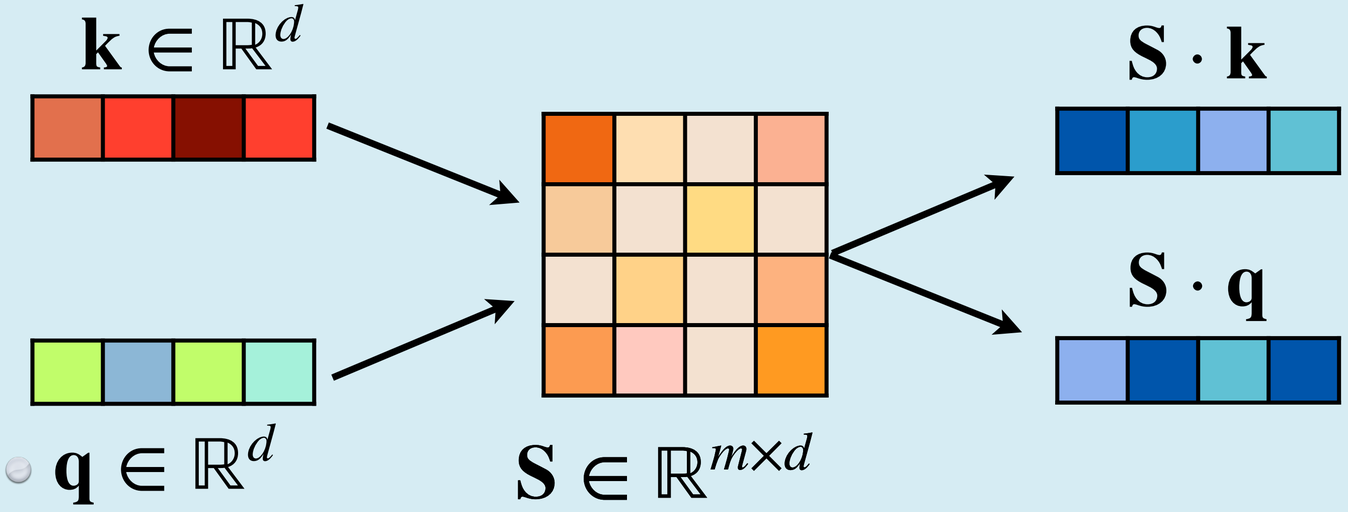
- **Quantization**
  - GEAR (Kang et al. 24)
  - KIVI (Liu et al., 24')
  - KVQuant (Hooper et al. 24)
  - Lexico (Compressed Sensing)
- **Eviction**
  - H_2O (Zhang et al., 23')
  - Sink (Xian et al., 24')
  - SubGen (Zandieh, 24')
- **Offloading**
  - InstInfer (Pan et al., 24')
  - InfiniGen (Lee et al., 24')
  - ShadowKV (Sun et al., 21')



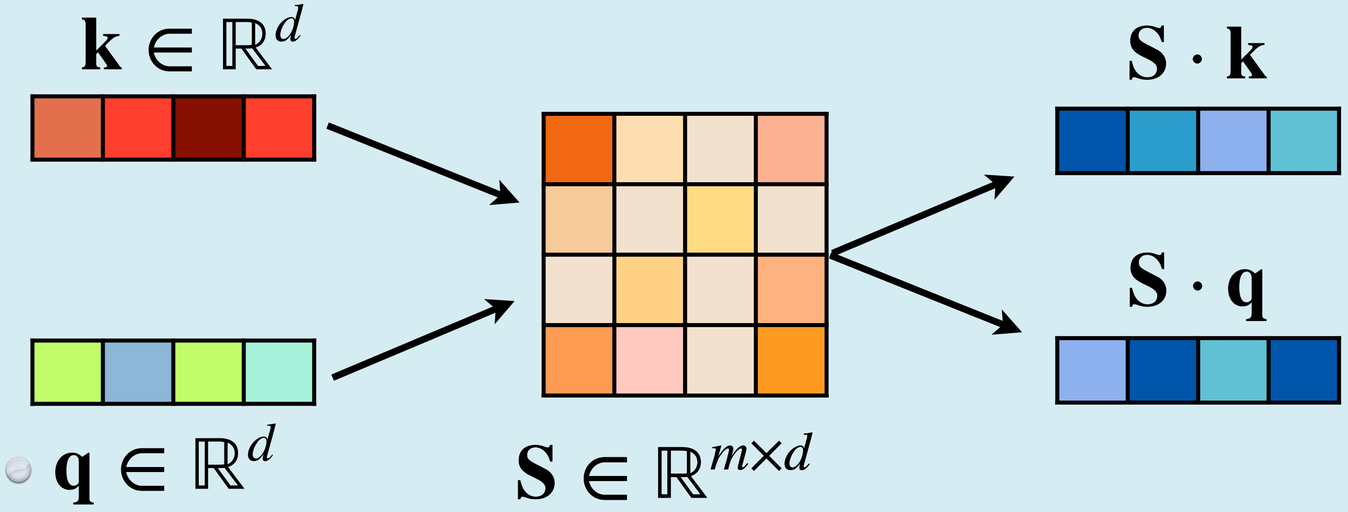




JL Transform



JL Transform

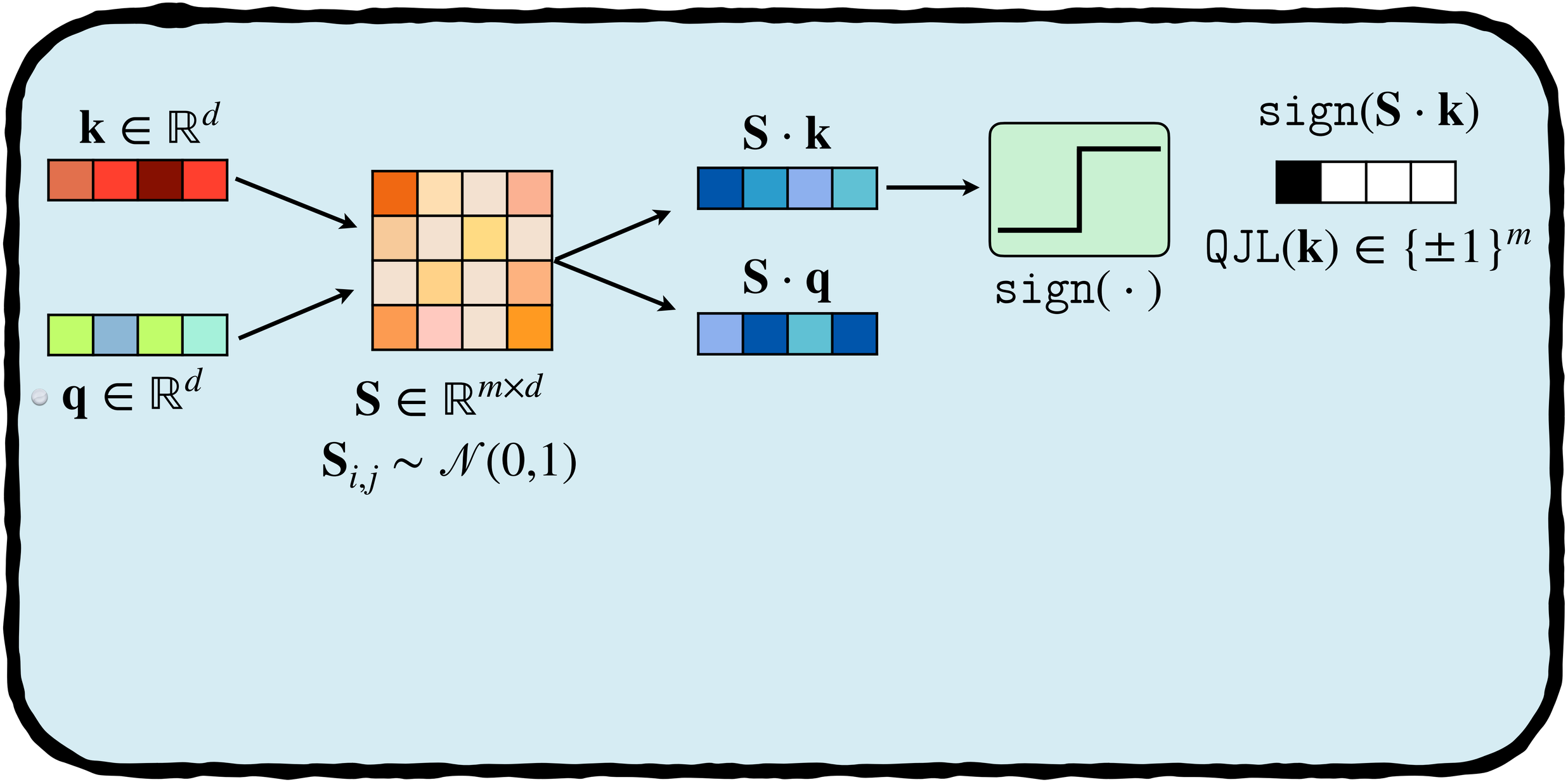


$$\mathbf{S}_{i,j} \sim \mathcal{N}(0,1)$$

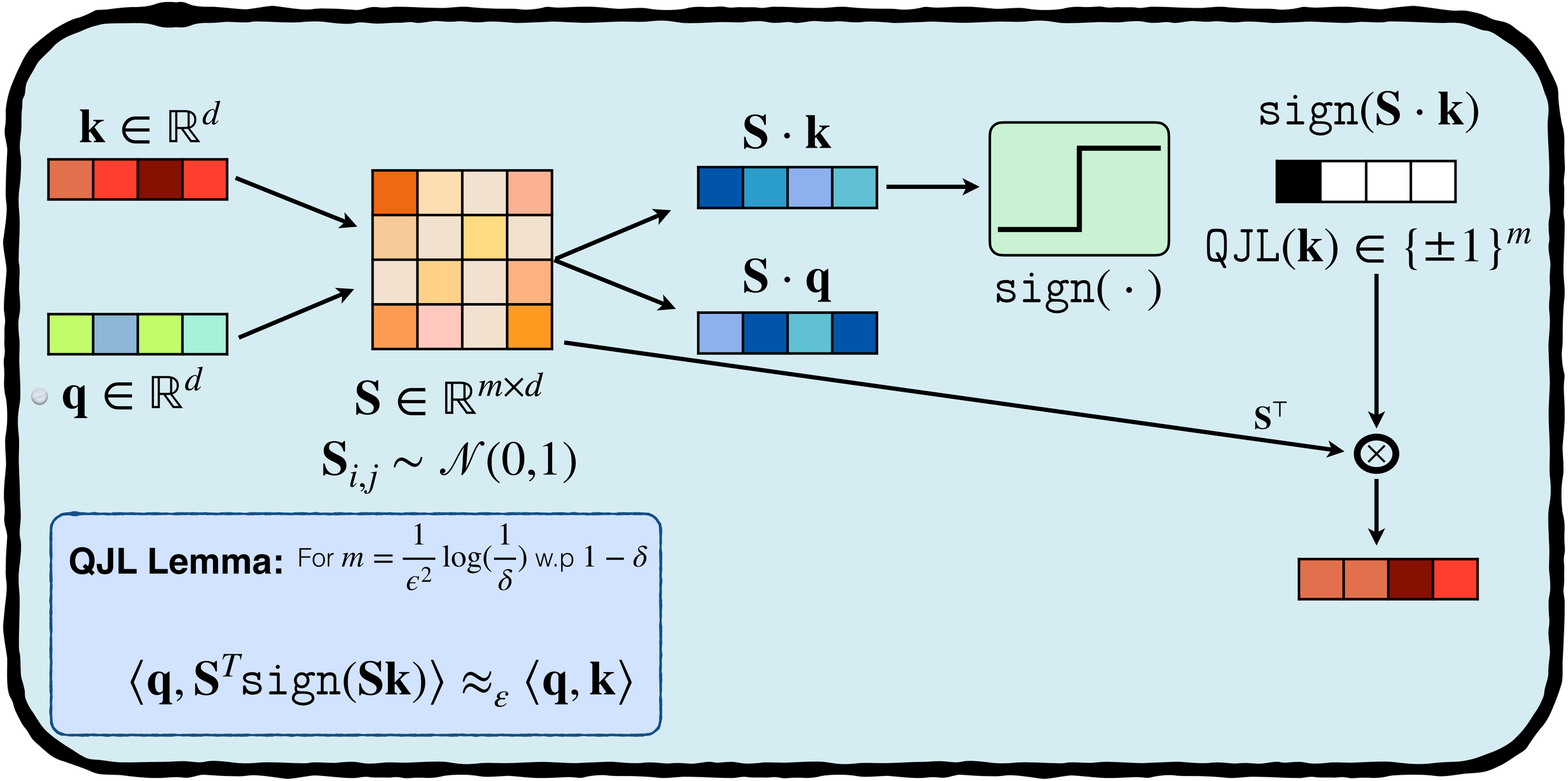
JL Lemma: For $m = \frac{1}{\epsilon^2}\log(\frac{1}{\delta})$ w.p $1-\delta$

$$\langle \mathbf{Sq}, \mathbf{Sk} \rangle \approx_{\varepsilon} \langle \mathbf{q}, \mathbf{k} \rangle$$

QJL


$\mathbf{k} \in \mathbb{R}^d$
$\mathbf{q} \in \mathbb{R}^d$
$\mathbf{S} \in \mathbb{R}^{m \times d}$
$\mathbf{S}_{i,j} \sim \mathcal{N}(0,1)$
$\mathbf{S} \cdot \mathbf{k}$
$\mathbf{S} \cdot \mathbf{q}$
$\texttt{sign}(\cdot)$
$\texttt{sign}(\mathbf{S} \cdot \mathbf{k})$
$\texttt{QJL}(\mathbf{k}) \in \{\pm 1\}^m$

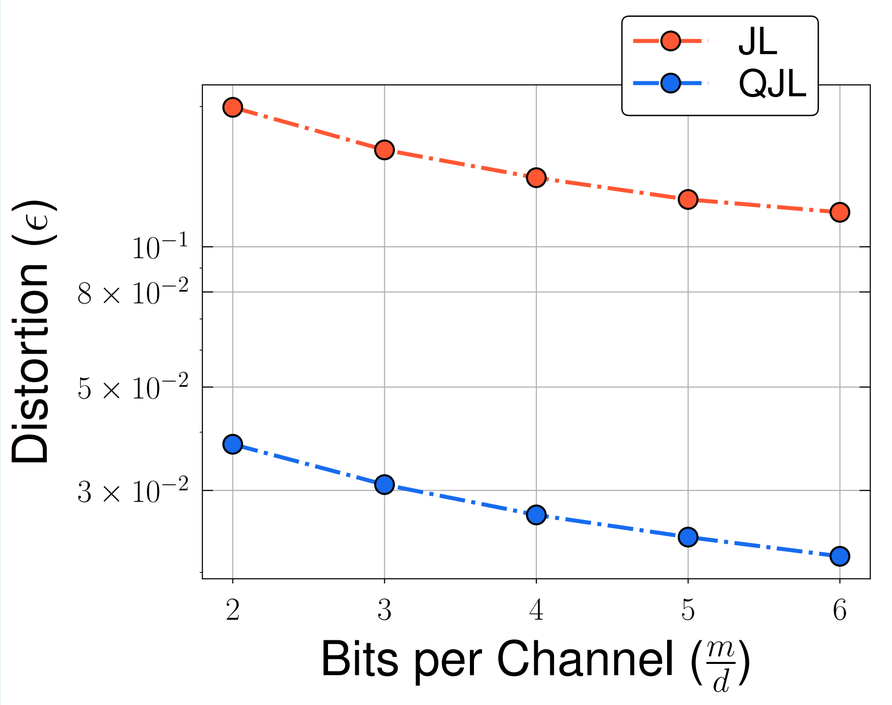
QJL

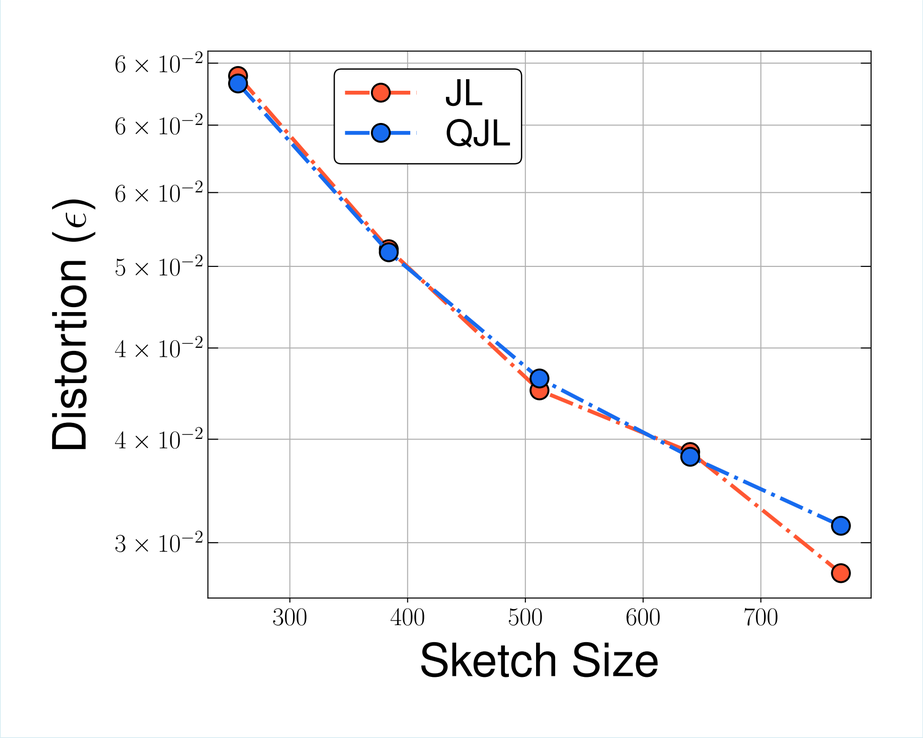


QJL Lemma: For $m = \frac{1}{\epsilon^2} \log(\frac{1}{\delta})$ w.p $1 - \delta$

$$\langle \mathbf{q}, \mathbf{S}^T \mathtt{sign}(\mathbf{S}\mathbf{k}) \rangle \approx_\varepsilon \langle \mathbf{q}, \mathbf{k} \rangle$$

QJL Quantize





QJL Proof of Unbiasedness

$$\texttt{Prod}_{\texttt{QJL}}(q,k) = \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \|k\|_2 \cdot s_i^\top q \cdot \texttt{sign}(s_i^\top k)$$

$$q^{\perp k} := q - \frac{\langle q,k\rangle}{\|k\|_2^2} \cdot k$$

QJL Proof of Unbiasedness

$$\texttt{Prod}_{\texttt{QJL}}(q,k) = \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \|k\|_2 \cdot s_i^\top q \cdot \texttt{sign}(s_i^\top k)$$

$$q^{\perp k} := q - \frac{\langle q,k\rangle}{\|k\|_2^2} \cdot k$$

$$= \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \frac{\langle q,k\rangle}{\|k\|_2} \cdot s_i^\top k \cdot \texttt{sign}(s_i^\top k) + \|k\|_2 \cdot s_i^\top q^{\perp k} \cdot \texttt{sign}(s_i^\top k)$$

QJL Proof of Unbiasedness

$$\texttt{Prod}_{\texttt{QJL}}(q,k) = \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \|k\|_2 \cdot s_i^\top q \cdot \texttt{sign}(s_i^\top k) \qquad q^{\perp k} := q - \frac{\langle q,k\rangle}{\|k\|_2^2} \cdot k$$

$$= \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \frac{\langle q,k\rangle}{\|k\|_2} \cdot s_i^\top k \cdot \texttt{sign}(s_i^\top k) + \|k\|_2 \cdot s_i^\top q^{\perp k} \cdot \texttt{sign}(s_i^\top k)$$

$$= \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \frac{\langle q,k\rangle}{\|k\|_2} \cdot |s_i^\top k| + \|k\|_2 \cdot s_i^\top q^{\perp k} \cdot \texttt{sign}(s_i^\top k)$$

QJL Proof of Unbiasedness

$$\texttt{Prod}_{\texttt{QJL}}(q,k) = \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \|k\|_2 \cdot s_i^\top q \cdot \texttt{sign}(s_i^\top k) \qquad q^{\perp k} := q - \frac{\langle q,k\rangle}{\|k\|_2^2} \cdot k$$

$$= \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \frac{\langle q,k\rangle}{\|k\|_2} \cdot s_i^\top k \cdot \texttt{sign}(s_i^\top k) + \|k\|_2 \cdot s_i^\top q^{\perp k} \cdot \texttt{sign}(s_i^\top k)$$

$$= \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \frac{\langle q,k\rangle}{\|k\|_2} \cdot |s_i^\top k| + \|k\|_2 \cdot s_i^\top q^{\perp k} \cdot \texttt{sign}(s_i^\top k)$$

$$\mathbb{E}[\texttt{Prod}_{\texttt{QJL}}(q,k)] = \sqrt{\pi/2}\left(\frac{\langle q,k\rangle}{\|k\|_2} \cdot \mathbb{E}\left[|s_1^\top k|\right] + \|k\|_2 \cdot \mathbb{E}\left[s_1^\top q^{\perp k} \cdot \texttt{sign}(s_1^\top k)\right]\right)$$

QJL Proof of Unbiasedness

$$\texttt{Prod}_{\texttt{QJL}}(q,k) = \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \|k\|_2 \cdot s_i^\top q \cdot \texttt{sign}(s_i^\top k) \qquad q^{\perp k} := q - \frac{\langle q,k\rangle}{\|k\|_2^2} \cdot k$$

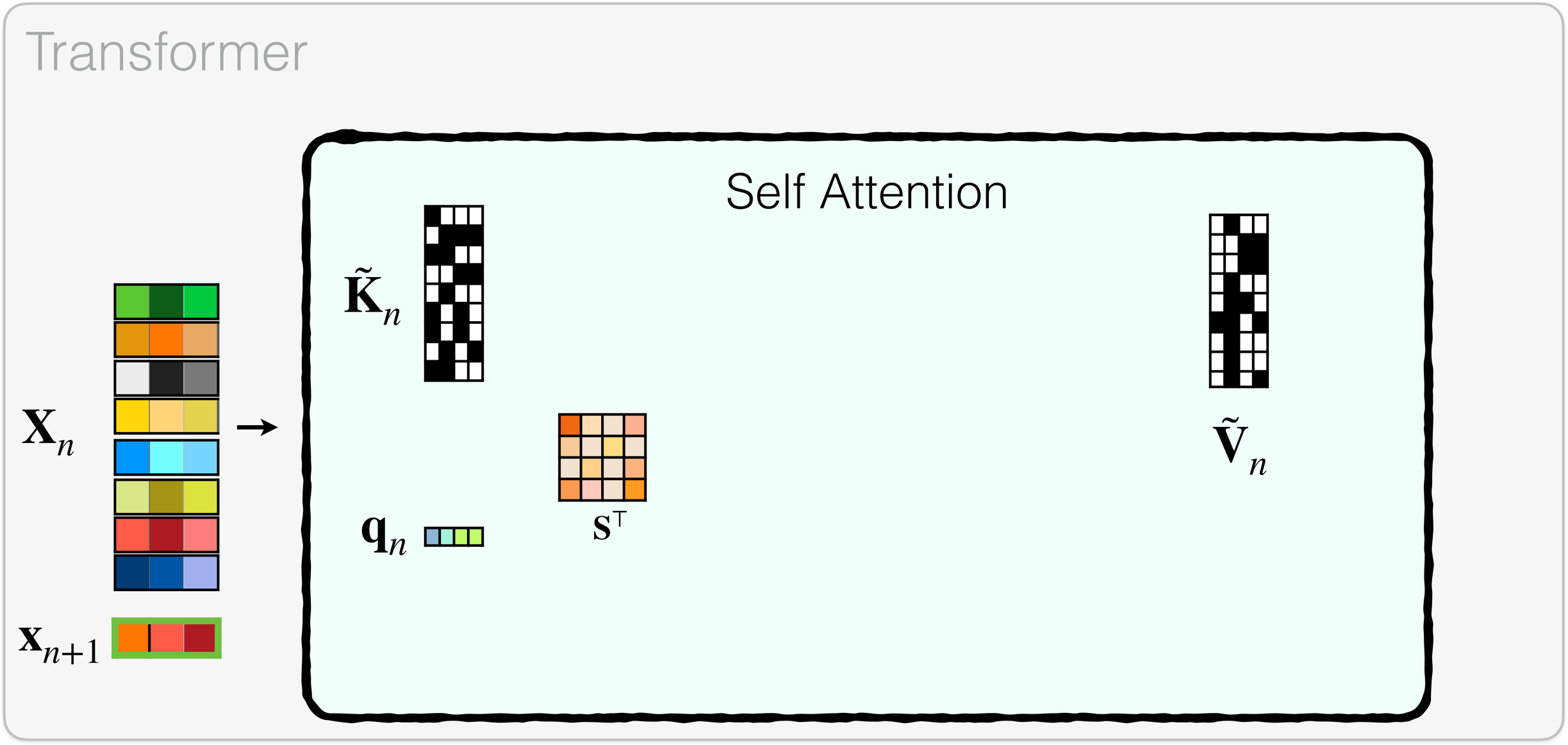
$$= \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \frac{\langle q,k\rangle}{\|k\|_2} \cdot s_i^\top k \cdot \texttt{sign}(s_i^\top k) + \|k\|_2 \cdot s_i^\top q^{\perp k} \cdot \texttt{sign}(s_i^\top k)$$

$$= \frac{\sqrt{\pi/2}}{m} \sum_{i\in[m]} \frac{\langle q,k\rangle}{\|k\|_2} \cdot |s_i^\top k| + \|k\|_2 \cdot s_i^\top q^{\perp k} \cdot \texttt{sign}(s_i^\top k)$$

$$\mathbb{E}[\texttt{Prod}_{\texttt{QJL}}(q,k)] = \sqrt{\pi/2}\left(\frac{\langle q,k\rangle}{\|k\|_2} \cdot \mathbb{E}\left[|s_1^\top k|\right] + \|k\|_2 \cdot \mathbb{E}\left[s_1^\top q^{\perp k} \cdot \texttt{sign}(s_1^\top k)\right]\right)$$

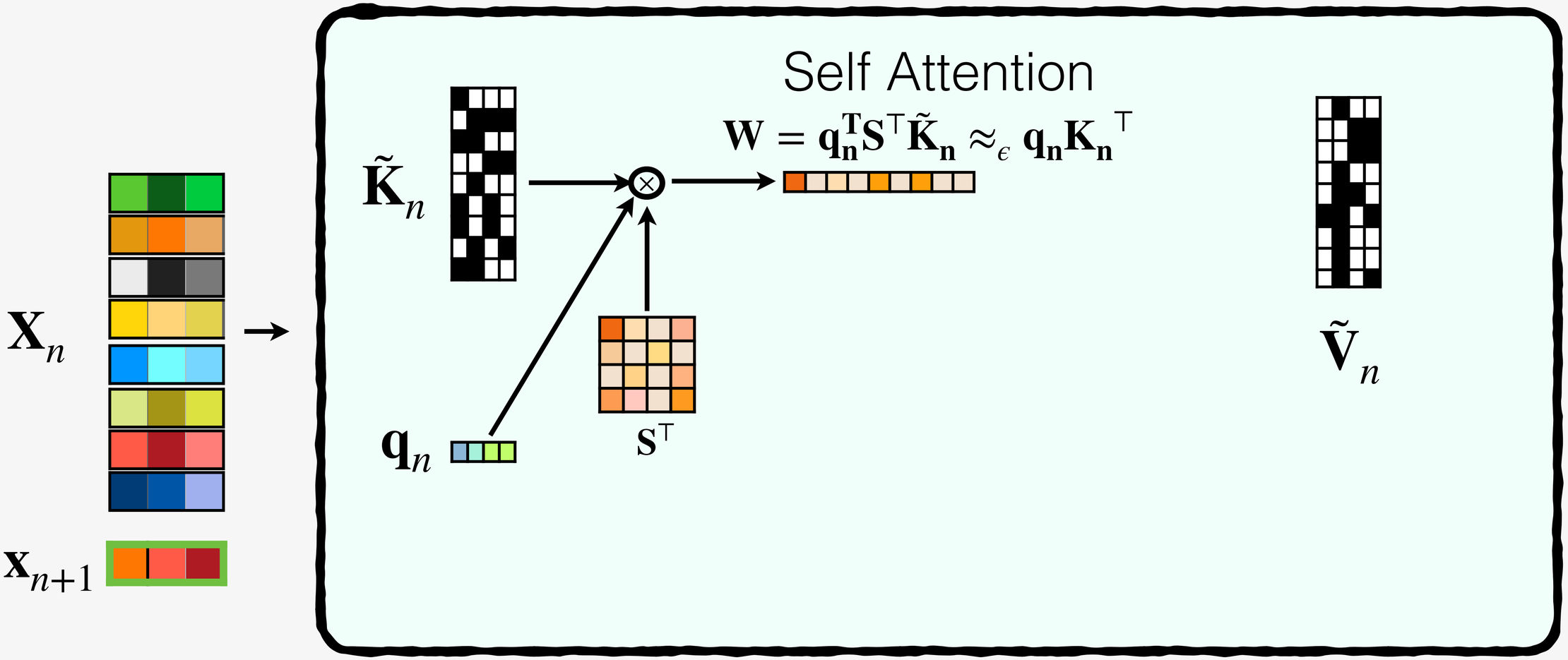
$$= \sqrt{\pi/2}\left(\frac{\langle q,k\rangle}{\|k\|_2} \cdot \mathbb{E}_{x\sim N(0,\|k\|)}\left[|x|\right]\right)$$

QJL: KV cache Quantization



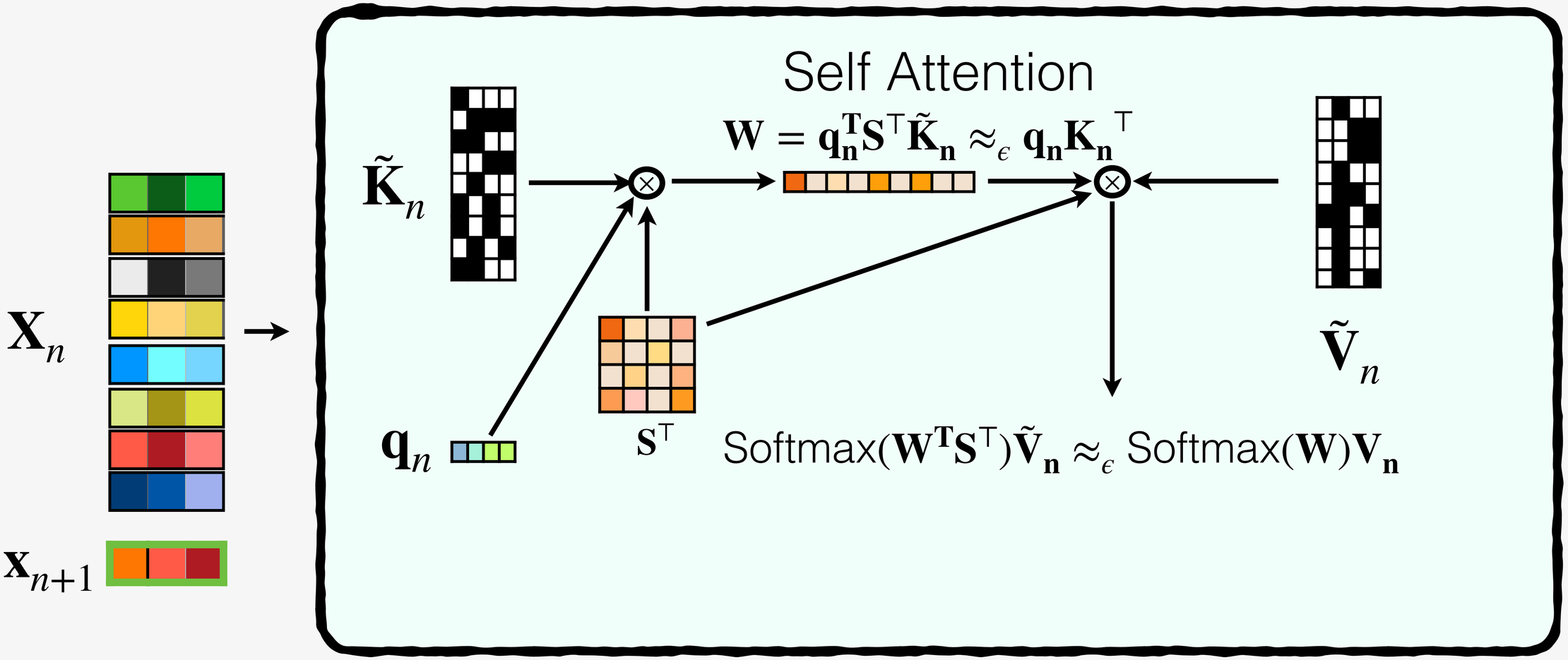
QJL: KV cache Quantization

Transformer



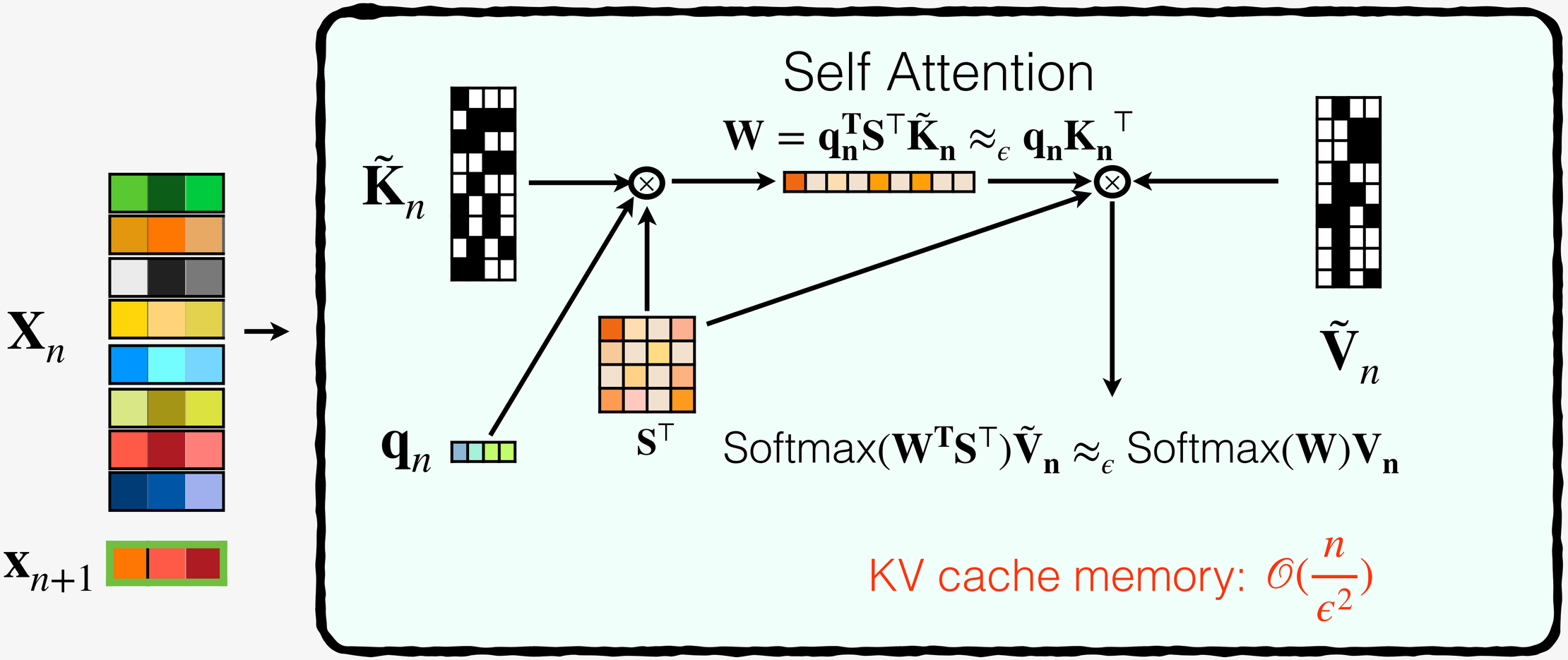
QJL: KV cache Quantization

Transformer



QJL: KV cache Quantization

Transformer



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

Evaluation

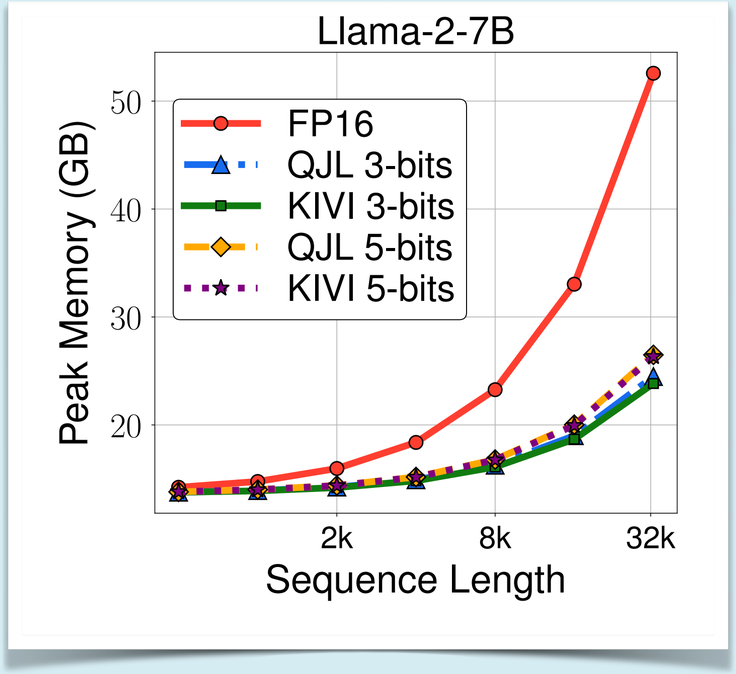
Method	KV Size	Qasper	QMSum	MultiNews	TREC	TriviaQA	SAMSum	LCC	RepoBench-P	Average
					Llama-3.1-8B-Instruct					
Full Cache	100%	22.54	24.57	27.44	72.5	91.65	43.47	63.15	56.76	50.26
KIVI-4	33.2%	**22.83**	23.72	**27.95**	71.0	90.39	**44.25**	**62.93**	55.48	49.78
Lexico $_{s=24}$	30.6%	21.68	**24.25**	27.20	**72.5**	**91.58**	42.93	62.92	**56.51**	**49.95**
KIVI-2	21.1%	13.77	22.72	**27.35**	71.0	90.85	**43.53**	62.03	53.00	48.03
Lexico $_{s=16}$	21.4%	**15.45**	**23.13**	25.78	**72.5**	**92.25**	42.02	**63.01**	**55.58**	**48.71**

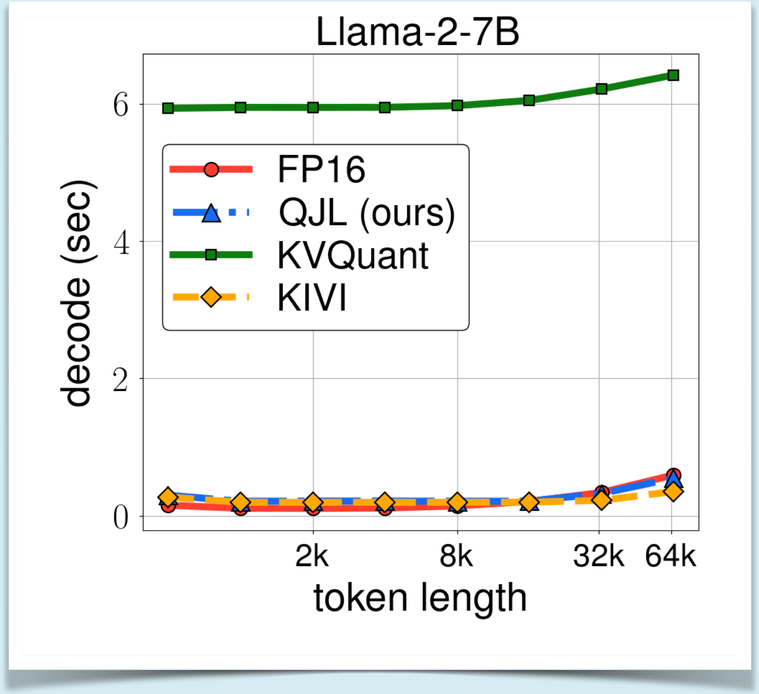
Method	KV Size	Qasper	QMSum	MultiNews	TREC	TriviaQA	SAMSum	LCC	RepoBench-P	Average
					Llama-3.1-8B-Instruct					
Full Cache	100%	44.98	25.22	26.99	73.5	91.79	43.87	62.99	56.39	53.47
QJL (ours)	25.0%	43.09	23.40	27.02	72.5	89.69	43.30	64.75	56.79	52.82
QJL (ours)	18.75%	42.76	23.68	26.82	72.5	88.55	43.43	63.98	56.09	52.48

Evaluation

Models	Methods	Bits	Datasets from LM-eval (Gao et al. 2023)				
			Lambada-OpenAI	HellaSwag	PIQA	MathQA	MMLU
Llama-2-7B	FP16 (baseline)	16	73.90	57.18	78.07	28.11	41.85
	KIVI (Liu et al. 2024b)	3	73.88	57.13	78.07	28.11	41.81
	QJL (ours)	3	73.88	57.14	78.07	28.17	41.78
Llama-3-8B	BF16 (baseline)	16	75.59	60.17	79.65	40.64	62.09
	QJL (ours)	3	75.61	60.13	79.87	40.60	62.12

Evaluation


Llama-2-7B
Peak Memory (GB)
50
40
30
20
FP16
QJL 3-bits
KIVI 3-bits
QJL 5-bits
KIVI 5-bits
2k
8k
32k
Sequence Length


Llama-2-7B
decode (sec)
6
4
2
0
FP16
QJL (ours)
KVQuant
KIVI
2k
8k
32k
64k
token length