SpinQuant:

LLM quantization with learned rotations

(ICLR 2025, Meta,Zechun Liu)

Taekhyun Park

Graduate School of Data Science, PNU, Busan

pthpark1@pusan.ac.kr

부산대학교

PUSAN NATIONAL UNIVERSITY

Motivation

- The key to effective Quantization is “how to control **outliers**”. (Smoothquant, GPTQ, Bitnet f4.8)
- Rotating the weight or activation matrix $(X \cdot R)$ can reduce outliers in the network

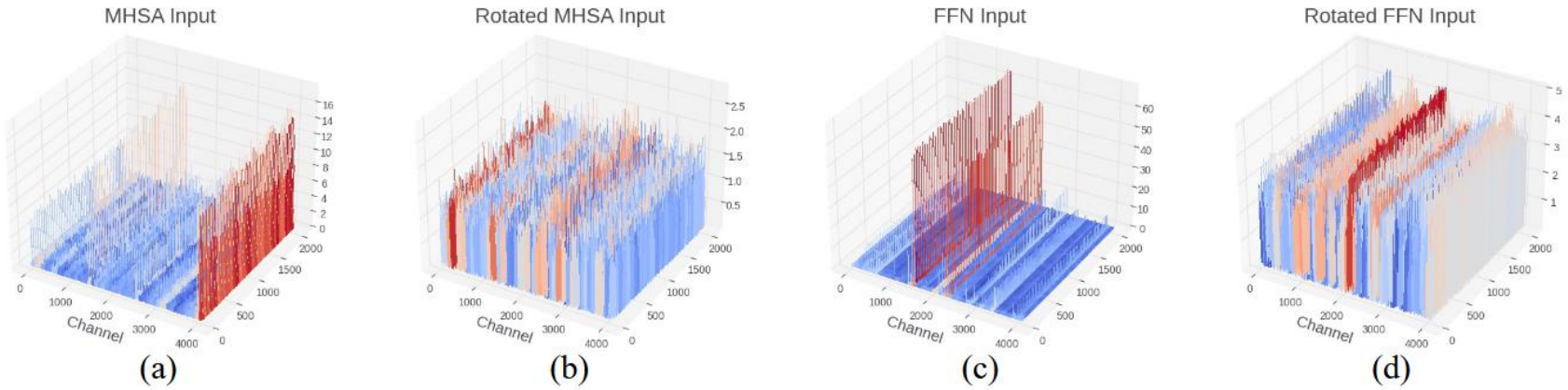


Figure 2: Activation distribution in LLaMA-2 7B model before and after rotation. Outliers exist in particular channels before rotation. Since channel-wise quantization is not supported in most hardware, outlier removal using rotation enables accurate token-wise or tensor-wise quantization.

Motivation

- Rotating the weight or activation matrix $(X \cdot R)$ can reduce outliers in the network

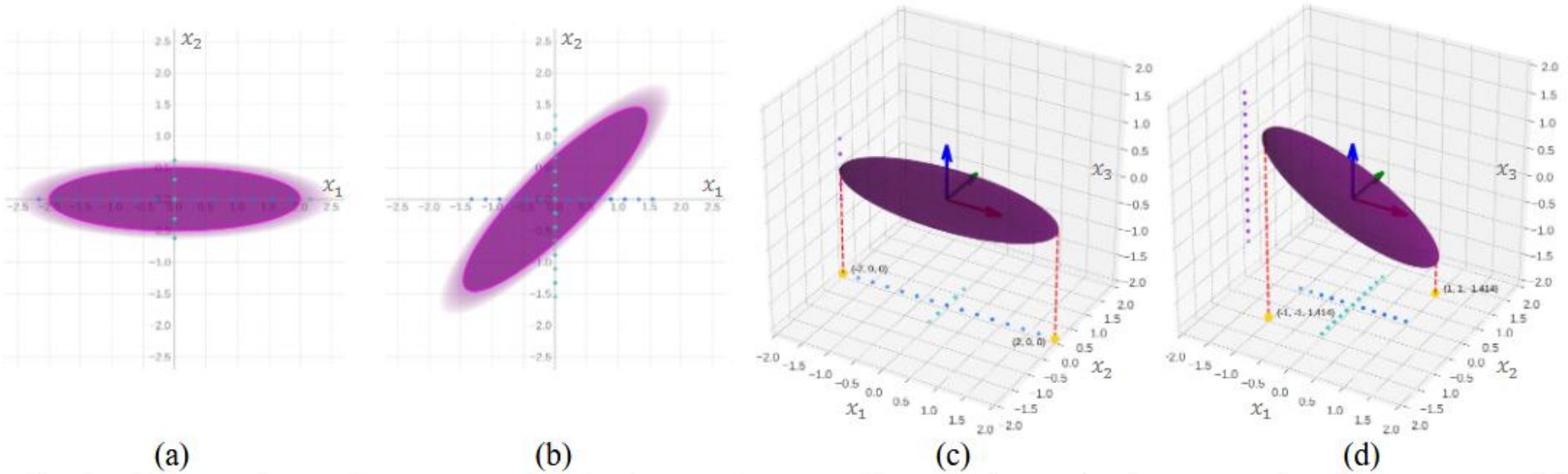


Figure 6: An illustration of how rotation helps reduce outliers and maximize quantization range utilization.

Intuition the rotation helps reduce outliers

- Outliers only exist in certain basis
- Rotate a matrix is changing the basis
- Random rotation matrix statistically makes the projection of the tensor to each axis have more similar magnitudes

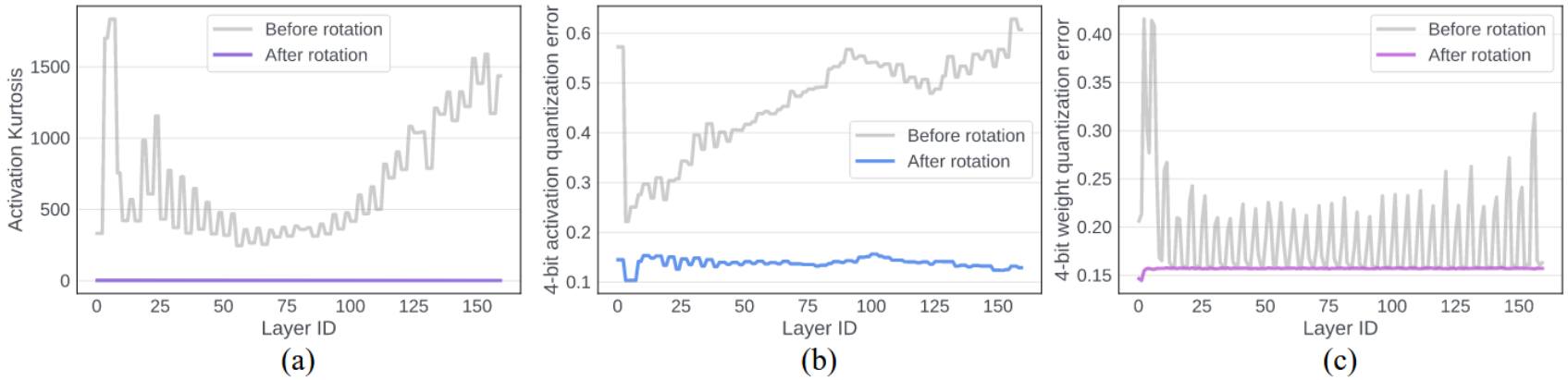


Figure 3: Outlier measurement and quantization error across input activation and weights in the five layers that take inputs from the residual (Q/K/V/Up/Gate-projection) of each block in the LLaMA-2 7B model. (a) After rotation, *kurtosis* of activation distributions is significantly reduced to approximately three across all layers. Quantization error is reduced after rotation in both (b) activations and (c) weights.

Apply rotation matrix in LLM

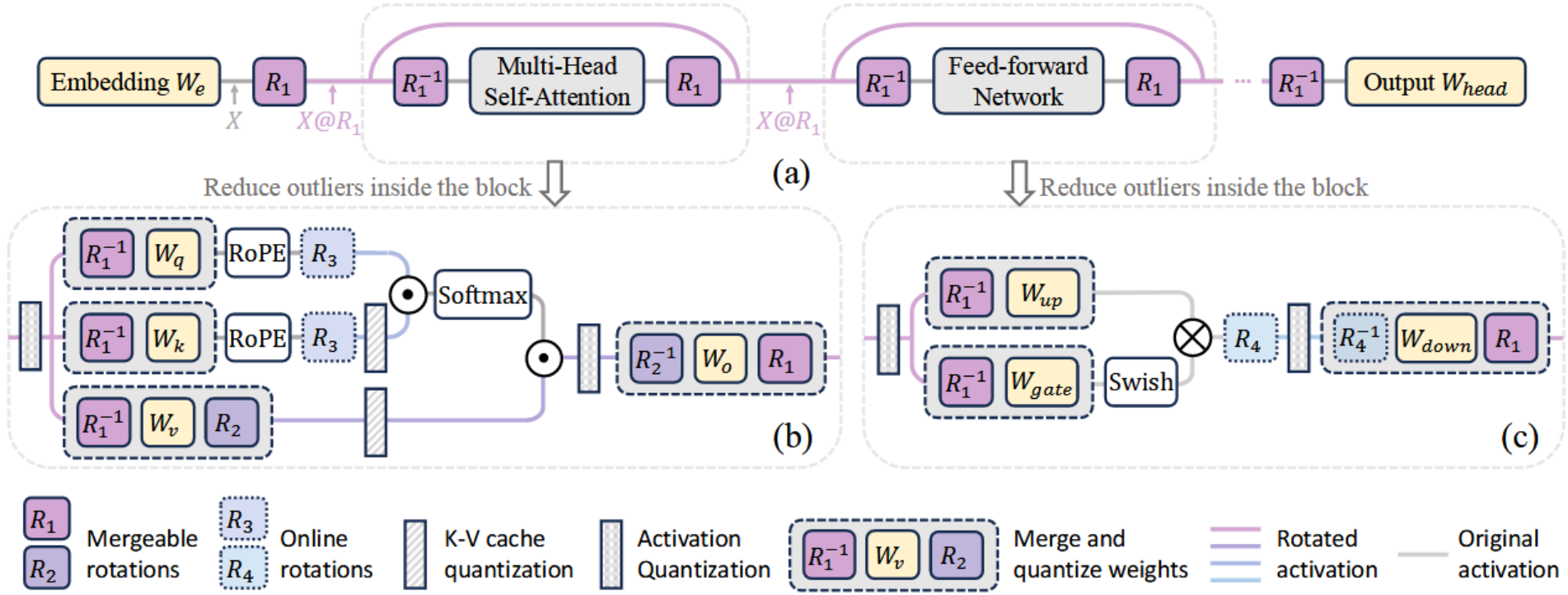
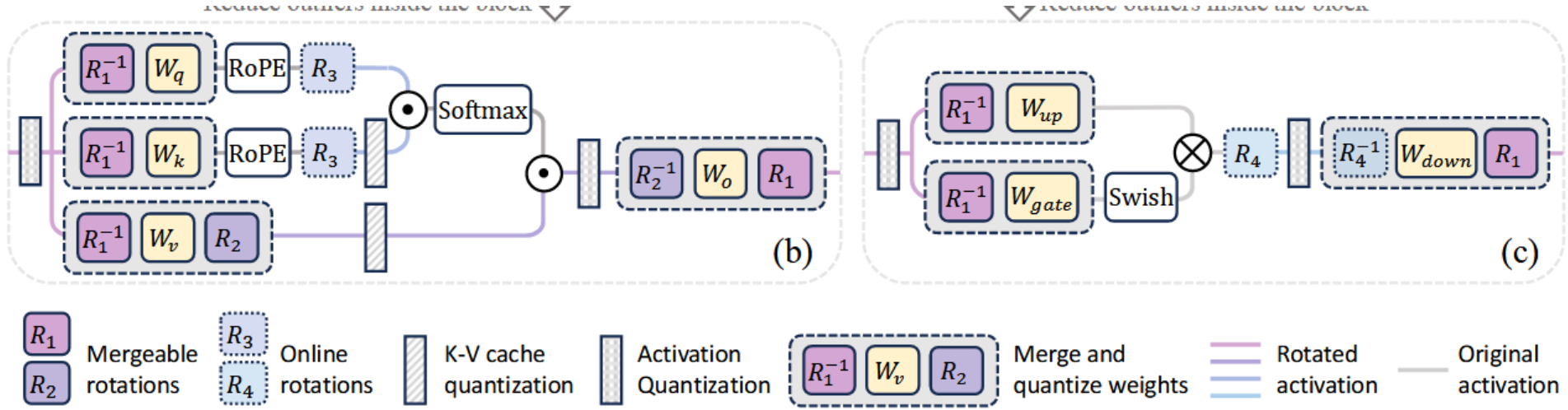


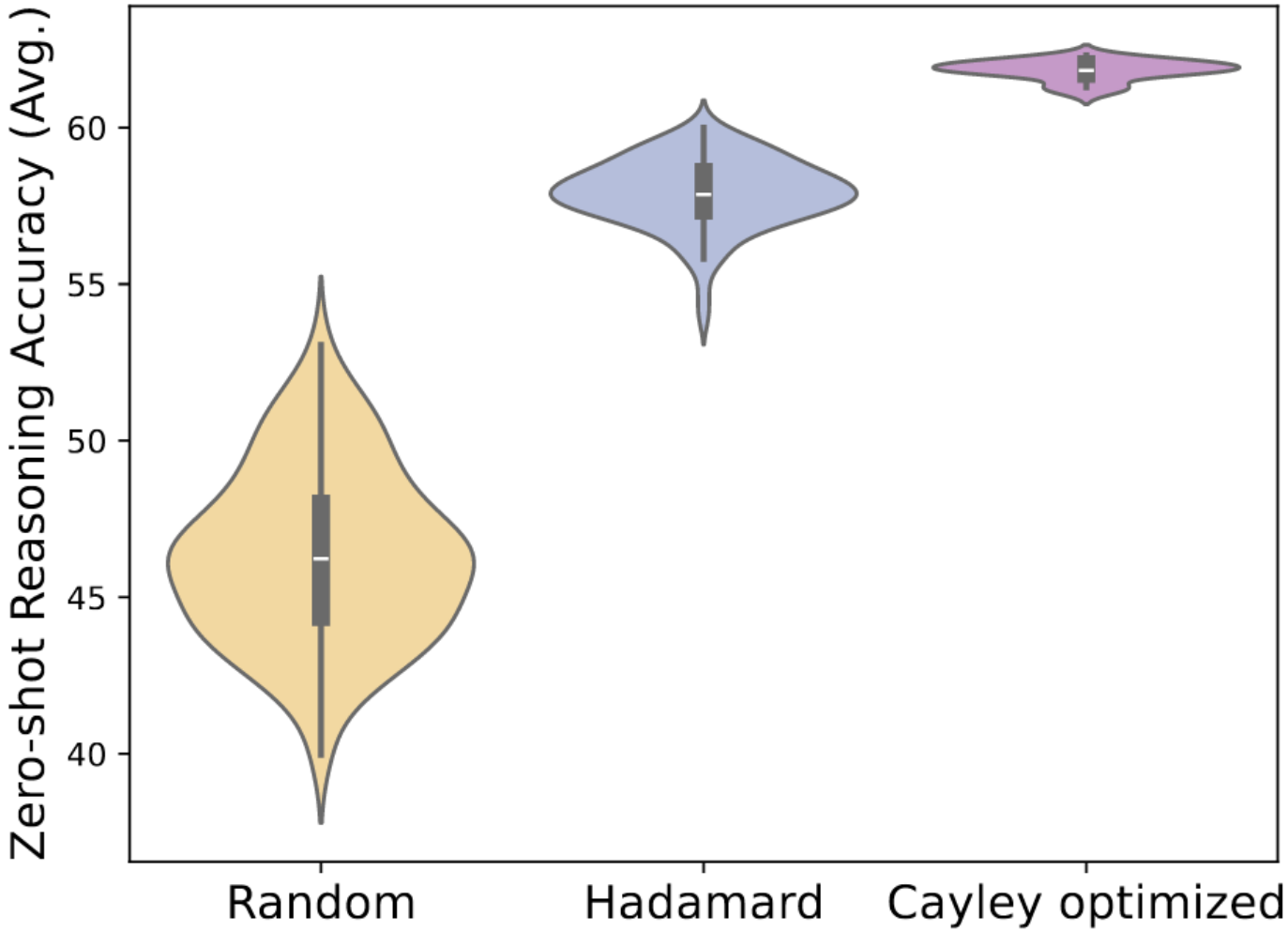
Figure 1: **Overall diagram of rotation.** (a) The residual stream can be rotated in the transformer network, resulting in numerically equivalent floating point networks before and after rotation. The rotated activations exhibit fewer outliers and are easier to quantize. (b) & (c) The rotation matrix can be integrated with the corresponding weight matrices and we further define R_2, R_3, and R_4 for reducing outliers inside the block.

Apply rotation matrix in LLM

- $SpinQuant_{no\ had}$
  - Use R_1, R_2 only. Merges rotation matrices into pre-trained weights without altering the network architecture. Suitable for low-bit weight quantization (W4A16, W4A8)
- $SpinQuant_{had}$
  - R_1, R_2 are mergable, R_3, R_4 use online Hadamard rotation, which is fast to compute.
  - Suitable for low-bit activation / KV cache quantization (W4A4,W4A4KV4).



Rotation matrix type



Rotation matrix type

- Hadamard matrix (H)
  - A Hadamard matrix H is a special type of rotation matrix, where the entries of the matrix are solely $\pm\sqrt{n}$. Given a Hadamard matrix H, we can generate 2^n different random Hadamard matrices by multiplying with S, a diagonal matrix with elements s_i randomly chosen from $\{-1,1\}$.
- Hadamard rotation can be implemented with fast Hadamard transform and introduce negligible overhead. ($n \log n$)

$$H_1 = [1],$$

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

$$HH^\mathsf{T} = nI_n$$

$$\frac{1}{\sqrt{n}} H^\mathsf{T} = \sqrt{n} H^{-1}$$

$$\det(H) = \pm n^{n/2},$$

Rotation matrix type

- $Cayley$-optimized rotation
  - Learn rotation matrices (R_1, R_2) to minimize quantization loss and improve accuracy using Cayley SGD
  - $\mathcal{M}$ is Stiefel manifold i.e., the set off all orthogonal matrices.

$$\arg\min_{R \in \mathcal{M}} \mathcal{L}_Q(R_1, R_2 \mid W, X)$$

$$Y = \hat{G} - \hat{G}^\top, \qquad \hat{G} := GR^\top - \frac{1}{2} RR^\top GR^\top \qquad G := \nabla_R \mathcal{L}_Q$$

Make skew-symmetric matrix Y — Projection G — Calculate loss G

$$R' = \Delta R(Y)R := \left(I - \frac{\alpha}{2} Y\right)^{-1} \left(I + \frac{\alpha}{2} Y\right) R$$

Updating Rotation matrix R

Experimental results

Table 1: Comparison of the perplexity score on WikiText2 and averaged accuracy on eight Zero-shot Common Sense Reasoning tasks. Results for SmoothQuant (Xiao et al., 2022), LLM-QAT (Liu et al., 2023c), GPTQ (Frantar et al., 2022) were obtained using their publicly released codebase. While OmniQuant (Shao et al., 2023), AWQ (Lin et al., 2023), and QuIP# (Tseng et al., 2024) results were quoted from their papers. Full results are in the Appendix.

#Bits (W-A-KV)	Method	LLaMA-2 7B 0-shot[8] Avg.(↑)	LLaMA-2 7B Wiki (↓)	LLaMA-2 13B 0-shot[8] Avg.(↑)	LLaMA-2 13B Wiki (↓)	LLaMA-2 70B 0-shot[8] Avg.(↑)	LLaMA-2 70B Wiki (↓)	LLaMA-3.2 1B 0-shot[8] Avg.(↑)	LLaMA-3.2 1B Wiki (↓)	LLaMA-3.2 3B 0-shot[8] Avg.(↑)	LLaMA-3.2 3B Wiki (↓)	LLaMA-3 8B 0-shot[8] Avg.(↑)	LLaMA-3 8B Wiki (↓)	Mistral-7B 0-shot[8] Avg.(↑)	Mistral-7B Wiki (↓)
16-16-16	FloatingPoint	66.9	5.5	68.3	5.0	72.9	3.3	56.9	13.4	63.9	10.7	69.6	6.1	71.0	5.4
4-8-16	RTN	62.4	7.9	57.3	6.7	68.6	5.0	55.4	20.7	58.6	29.0	65.5	8.2	59.3	6.8
	SmoothQuant	58.9	7.5	63.6	6.1	70.6	4.1	47.1	1e2	55.6	3e2	61.0	10.7	–	–
	LLM-QAT	64.8	11.4	67.5	14.5	–	–	53.2	21.0	60.8	41.1	67.2	7.7	–	–
	AWQ (w4)	–	6.2	–	5.1	–	–	–	–	–	–	–	–	–	–
	OmniQuant (w4)	–	5.7	–	5.0	–	3.5	–	–	–	–	–	–	–	–
	QuIP# (w4)	–	**5.6**	–	5.0	–	**3.4**	–	–	–	–	–	–	–	–
	GPTQ	64.9	20.2	65.2	5.9	71.7	4.3	55.0	17.3	58.7	25.2	64.5	7.2	51.7	8.6
	SpinQuant$_{no\ had}$	**65.7**	5.8	**68.2**	5.1	72.1	3.7	56.0	15.3	61.4	11.6	**68.6**	6.7	68.8	5.7
	SpinQuant$_{had}$	**65.7**	5.7	68.1	**5.0**	**72.7**	3.5	**56.5**	**14.4**	**63.2**	**11.5**	68.4	**6.5**	**69.9**	**5.5**
4-8-8	RTN	62.5	7.9	57.6	6.7	68.4	5.0	55.7	20.7	58.4	28.8	65.3	8.2	58.9	6.7
	SmoothQuant	58.8	7.5	63.4	6.1	70.5	4.1	47.1	1e2	55.5	3e2	60.9	10.7	–	–
	LLM-QAT	64.6	11.4	67.5	14.2	–	–	53.1	21.0	60.5	39.3	66.9	7.6	–	–
	GPTQ	64.8	20.2	65.3	5.9	71.6	4.3	54.8	17.3	58.7	24.1	64.6	7.2	51.7	8.6
	SpinQuant$_{no\ had}$	**65.8**	5.8	68.1	**5.1**	72.2	3.7	55.7	15.3	61.8	11.7	68.6	6.7	69.4	5.7
	SpinQuant$_{had}$	**65.8**	**5.7**	**68.2**	**5.1**	**72.7**	**3.5**	**55.8**	**14.3**	**63.2**	**11.2**	**68.8**	**6.5**	**70.2**	**5.5**
4-4-16	RTN	35.6	2e3	35.3	7e3	35.1	2e5	41.2	1e2	42.1	7e2	43.9	2e2	41.4	4e2
	SmoothQuant	41.8	3e2	44.9	34.5	57.7	57.1	37.9	2e3	43.6	4e2	40.3	9e2	–	–
	LLM-QAT	47.8	12.9	34.3	4e3	–	–	42.0	62.1	46.9	37.6	44.9	42.9	–	–
	GPTQ	36.8	9e3	35.2	5e3	35.5	2e6	41.6	1e2	43.4	3e2	40.6	2e2	40.4	3e2
	SpinQuant$_{no\ had}$	57.0	9.2	61.8	7.2	61.0	7.3	44.8	48.4	52.9	22.4	51.9	18.6	52.7	13.4
	SpinQuant$_{had}$	**64.1**	**5.9**	**67.2**	**5.2**	**71.0**	**3.8**	**53.5**	**15.3**	**61.0**	**11.1**	**65.8**	**7.1**	**68.4**	**5.7**
4-4-4	RTN	37.1	2e3	35.5	7e3	35.0	2e5	40.6	2e2	41.2	8e2	43.1	3e2	41.4	4e2
	SmoothQuant	39.0	7e2	40.5	56.6	55.9	10.5	36.5	2e3	40.0	6e2	38.7	2e3	–	–
	LLM-QAT	44.9	14.9	35.0	4e3	–	–	41.5	76.2	45.9	42.0	43.2	52.5	–	–
	GPTQ	36.8	9e3	35.2	5e3	35.6	1e6	41.6	1e2	41.1	4e2	40.5	2e2	41.3	2e2
	SpinQuant$_{no\ had}$	56.0	9.2	60.7	7.1	62.0	7.4	45.3	47.7	52.9	22.4	52.6	18.6	52.4	13.7
	SpinQuant$_{had}$	**64.0**	**5.9**	**66.9**	**5.3**	**71.2**	**3.8**	**53.4**	**15.9**	**60.5**	**11.4**	**65.5**	**7.3**	**68.6**	**5.8**

Ablation learned rotation vs random rotation

Table 2: Compared to Hadamard rotation, SpinQuant learned rotation consistently outperform by a significant margin. Results are averaged accuracy on eight Zero-shot CommonSense Reasoning tasks.

	LLaMA-3.2 3B		LLaMA-3 8B		Mistral-7B	
	4-4-16	4-4-4	4-4-16	4-4-4	4-4-16	4-4-4
Random Hadamard $R_{\{1,2\}}$	49.8	49.6	49.5	50.0	51.4	51.5
SpinQuant$_{no\ had}$ $R_{\{1,2\}}$	**52.9** (↑3.1)	**52.9** (↑3.3)	**51.9** (↑2.4)	**52.6** (↑2.5)	**52.7** (↑1.3)	**52.4** (↑0.9)
Random Hadamard $R_{\{1,2,3,4\}}$	59.0	58.4	64.2	63.9	52.7	52.4
SpinQuant$_{had}$ $R_{\{1,2,3,4\}}$	**61.0** (↑2.1)	**60.5** (↑2.2)	**65.8** (↑1.6)	**65.5** (↑1.6)	**68.4** (↑15.7)	**68.6** (↑16.2)

Ablation SpinQuant with another methods

Table 4: Floating-point(FP) rotation vs Hadamard rotation on a LLaMA-2 7B model.

#Bits (W-A-KV)	Task	No *Cayley* + RTN FP	No *Cayley* + RTN Hadamard	*Cayley* + RTN FP init.	*Cayley* + RTN Hadamard init.
4-16-16	0-shot[8] Avg.(↑)	62.5 ±0.8	62.4 ±1.0	64.9 ±0.4	64.6 ±0.3
	Wiki(↓)	6.7 ±0.12	6.9 ±0.45	5.5 ±0.01	5.5 ±0.01
4-4-16	0-shot[8] Avg.(↑)	49.4 ±2.8	59.0 ±1.0	61.6 ±0.4	61.8 ±0.4
	Wiki(↓)	15.9 ±4.04	8.2 ±0.73	6.2 ±0.06	6.1 ±0.03
4-4-4	0-shot[8] Avg.(↑)	48.3 ±2.7	58.7 ±1.0	61.5 ±0.8	61.5 ±0.3
	Wiki(↓)	18.2 ±4.35	8.2 ±0.36	6.3 ±0.08	6.2 ±0.03

Table 5: Comparison with QuaRot (Ashkboos et al., 2023b).

	LLaMA-3 8B (FP: 69.6, 6.1) 4-4-16		**LLaMA-3 8B** (FP: 69.6, 6.1) 4-4-4		**LLaMA-3 70B** (FP: 74.5, 2.8) 4-4-16		**LLaMA-3 70B** (FP: 74.5, 2.8) 4-4-4	
	0-shot[8] Avg.(↑)	Wiki (↓)	0-shot[8] Avg.(↑)	Wiki (↓)	0-shot[8] Avg.(↑)	Wiki (↓)	0-shot[8] Avg.(↑)	Wiki (↓)
QuaRot+RTN	59.5	10.4	58.6	10.9	41.5	91.2	41.3	92.4
SpinQuant$_{had}$+RTN	**64.6**	**7.7**	**64.1**	**7.8**	**70.1**	**4.1**	**70.1**	**4.1**
QuaRot+GPTQ	63.8	7.9	63.3	8.0	65.4	20.4	65.1	20.2
SpinQuant$_{had}$+GPTQ	**65.8**	**7.1**	**65.5**	**7.3**	**69.5**	**5.5**	**69.3**	**5.5**