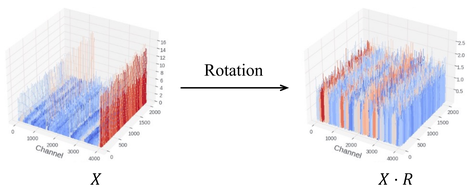
SpinQuant: LLM Quantization with Learned Rotations

Zechun Liu*, Changsheng Zhao*, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, Tijmen Blankevoort

① Observation

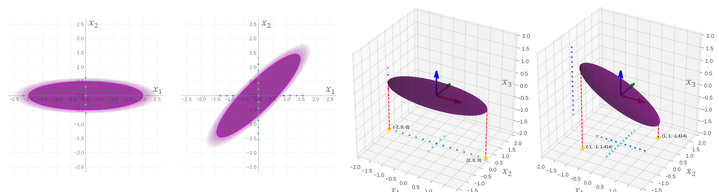
LLM contain a lot of outliers Outliers disappear after rotation



Outliers will dominate quantization range and reduce the precision for the majority of normal valued weights and harm quantization accuracy.

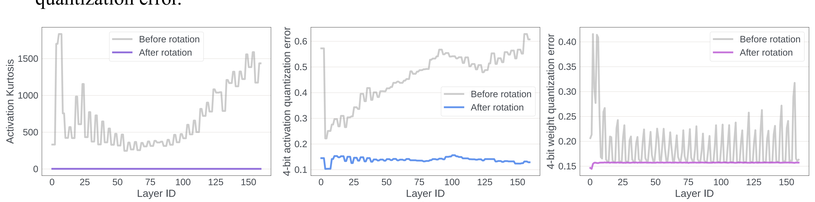
② Intuitive understanding

Intuitively, how does rotation help reduce outliers and maximize quantization range utilization?



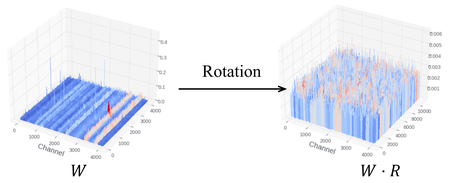
③ Statistically effective

Statistically, rotation significantly reduces the skewness of the activation distribution and lowers quantization error.



④ Two bird one stone

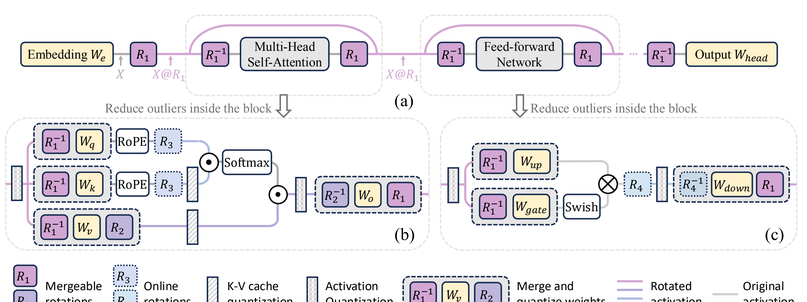
Both weights and activations become easier to quantize



Numerically Identical

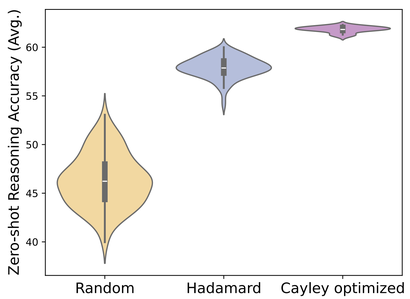
$$A \cdot W^T = A \cdot R \cdot R^T \cdot W^T = \underbrace{(A \cdot R)}_{\text{Easy to quantize}} \cdot (W \cdot R)^T$$

⑤ Create an rotation-invariant transformer



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.

⑥ Not all rotations are equally effective



⑦ Optimize rotation matrix to make it even better

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

$\mathcal{M}$: Stiefel manifold

- $\mathcal{L}_Q(\cdot)$: the quantization network task loss. It is a function of {R_1, R_2}, given the fixed pretrained weights W.
- We employ the *Cayley SGD* method which is an efficient optimization algorithm on the *Stiefel* manifold.
- Only need to run 100 iterations on 800 Wiki examples.

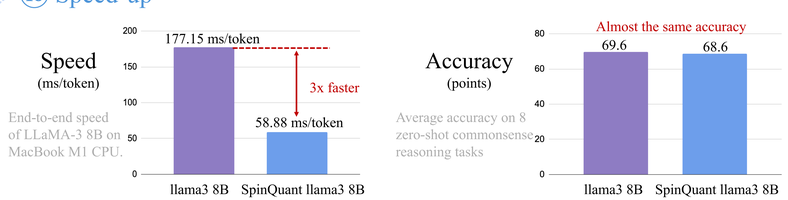
⑧ Experimental results

#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**

⑨ Comparison to random rotation

	LLaMA-3.2 3B 4-4-16	LLaMA-3.2 3B 4-4-4	LLaMA-3 8B 4-4-16	LLaMA-3 8B 4-4-4	Mistral-7B 4-4-16	Mistral-7B 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**$_{(\uparrow 3.1)}$	**52.9**$_{(\uparrow 3.3)}$	**51.9**$_{(\uparrow 2.4)}$	**52.6**$_{(\uparrow 2.5)}$	**52.7**$_{(\uparrow 1.3)}$	**52.4**$_{(\uparrow 0.9)}$
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**$_{(\uparrow 2.1)}$	**60.5**$_{(\uparrow 2.2)}$	**65.8**$_{(\uparrow 1.6)}$	**65.5**$_{(\uparrow 1.6)}$	**68.4**$_{(\uparrow 15.7)}$	**68.6**$_{(\uparrow 16.2)}$

⑩ Speed-up

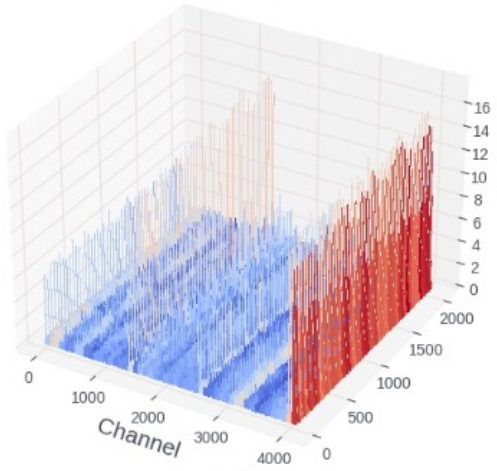


⑪ What's next?

ParetoQ: Scaling Laws in Extremely Low-bit LLM Quantization (Arxiv)

SpinQuant

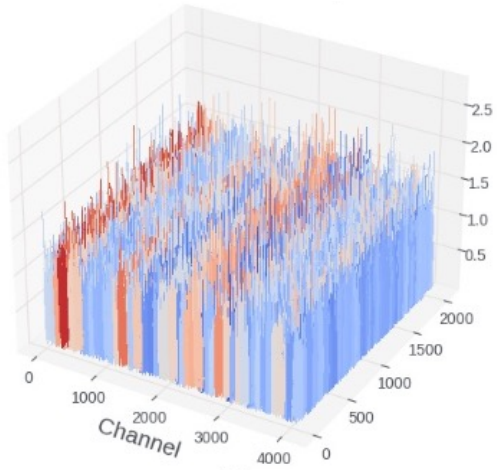
LLM contain a lot of outliers



X

Rotation

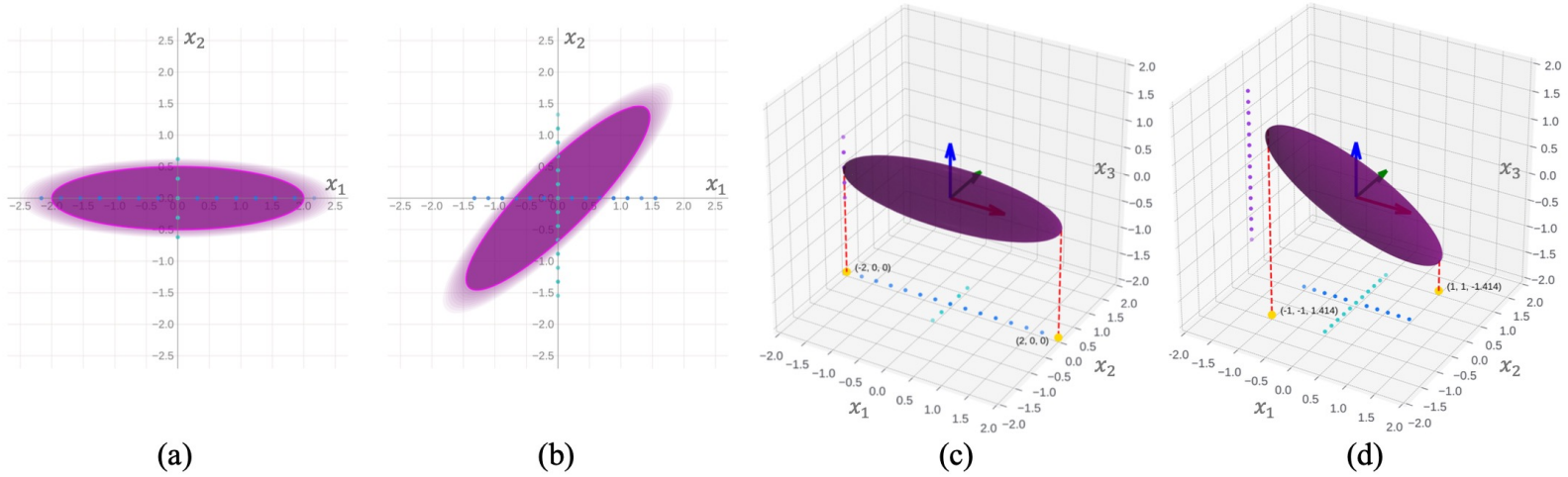
Outliers disappear after rotation



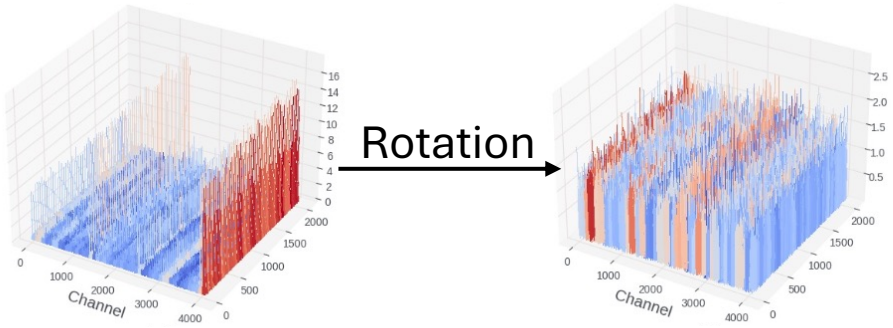
$X \cdot R$

SpinQuant

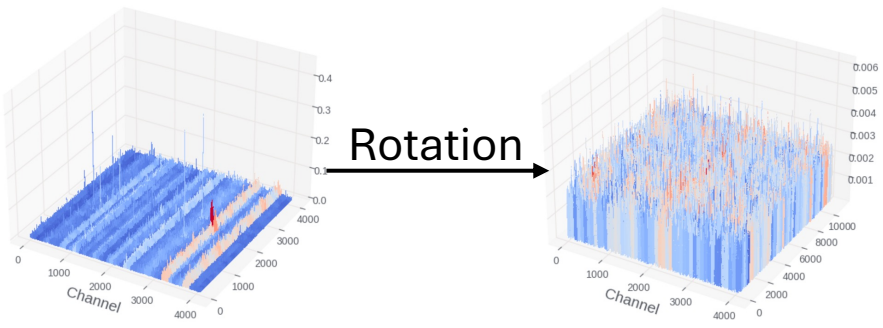
- How does rotation eliminate outliers?



SpinQuant



A $A \cdot R$



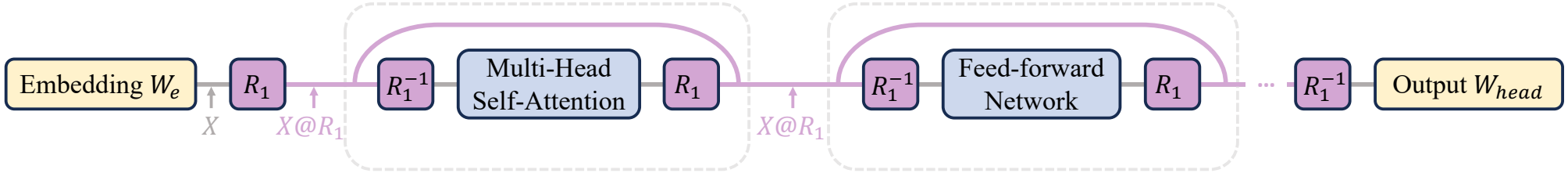
W $W \cdot R$

Numerically Identical

$$A \cdot W^T = A \cdot R \cdot R^T \cdot W^T = \underbrace{(A \cdot R)}_{\text{Easy to quantize}} \cdot \underbrace{(W \cdot R)^T}_{\text{Easy to quantize}}$$

SpinQuant

Network-level rotation invariance in the transformer



SpinQuant

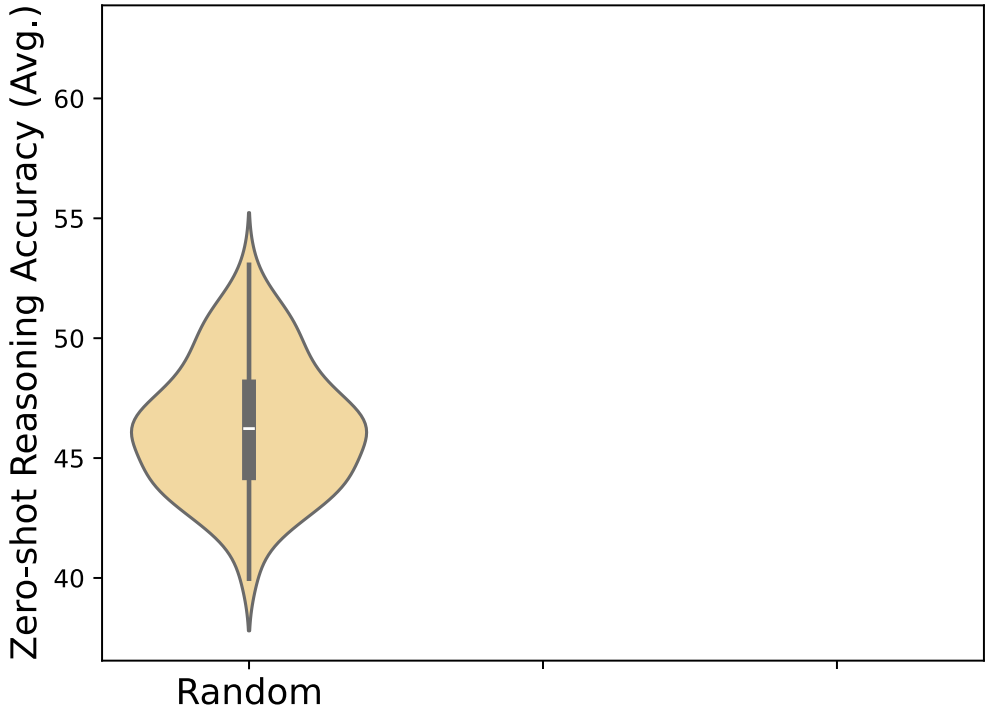
Meta

Which rotation to choose?

- Random rotation
- Hadamard Rotation (H),
  A special type of rotation matrix, where the entries of the matrix are solely $\pm\sqrt{n}$.
- Or learnable?

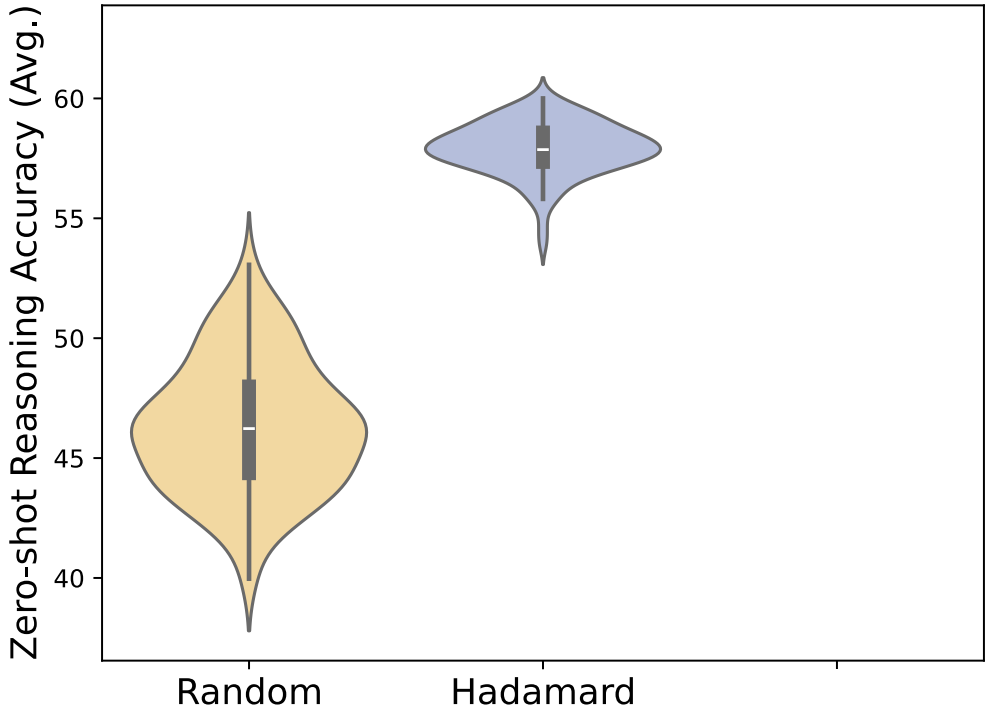
SpinQuant

Random rotation introduce high variance



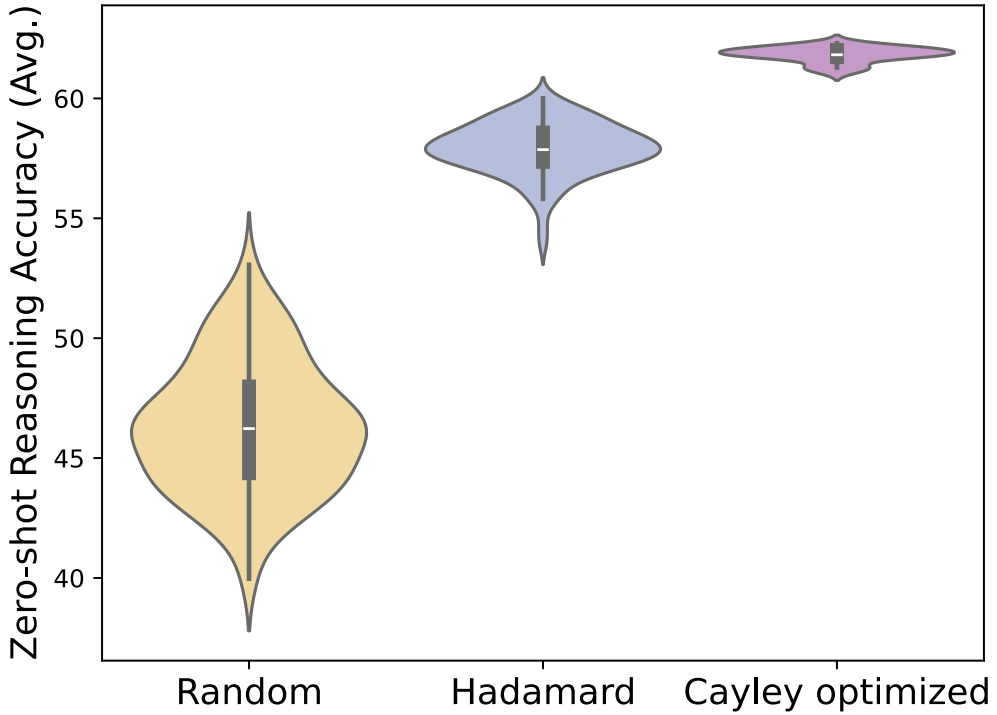
SpinQuant

Random Hadamard rotation is slightly better



SpinQuant

Learned rotation achieves highest accuracy with smallest variance.



SpinQuant

- Optimize the rotations with respect to the final loss of a quantized network

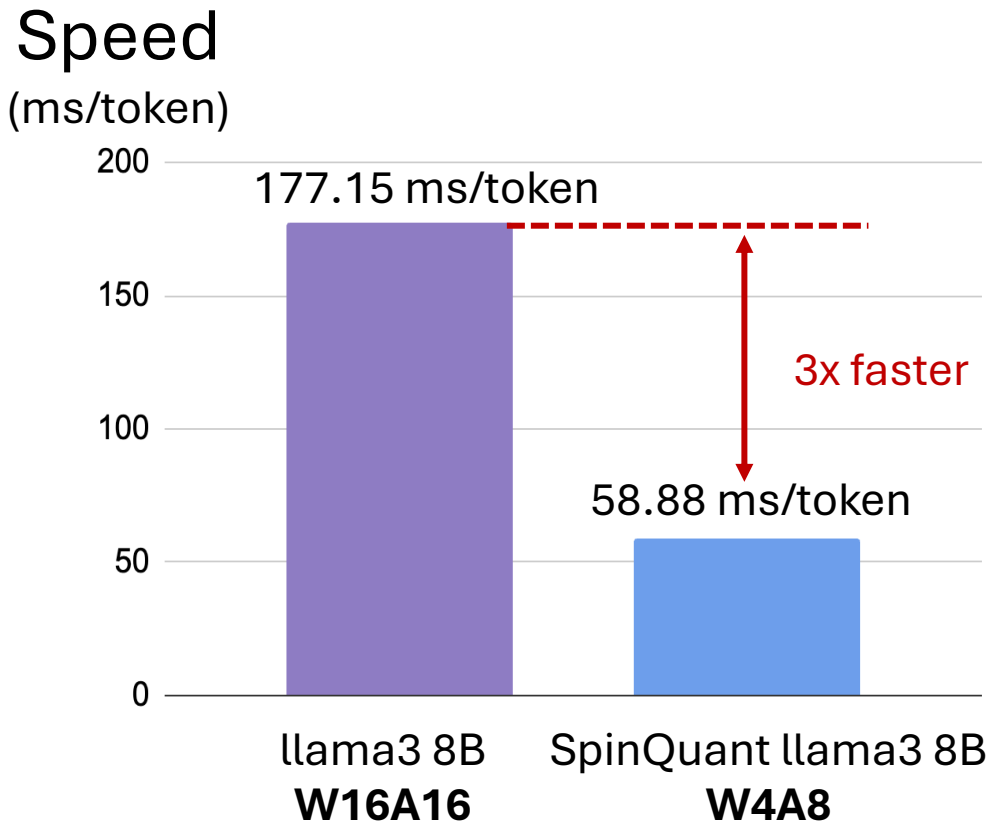
$$\arg\min_{R \in \mathcal{M}} \mathcal{L}_Q(R_1, R_2 \mid W, X)$$

SpinQuant

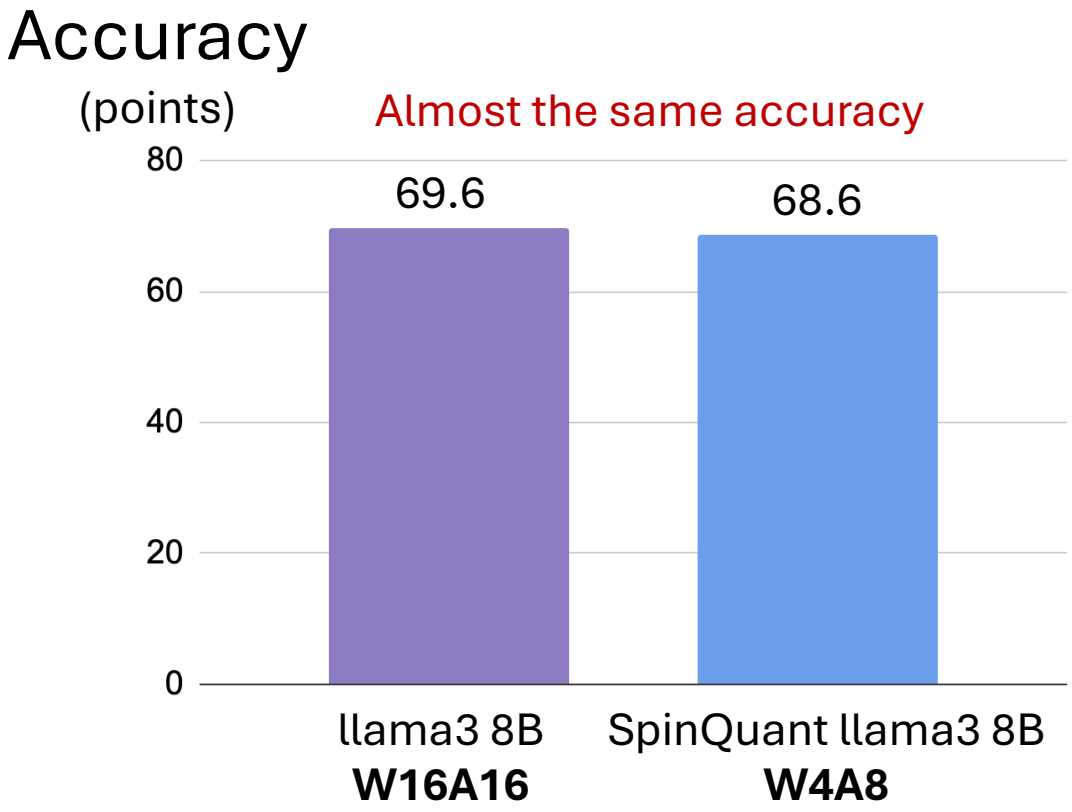
#Bits (W-A-KV)	Method	LLaMA-2 7B 0-shot[8] Avg.(↑)	LLaMA-2 7B Wiki (↓)	LLaMA-2 13B 0-shot[8] Avg.(↑)	LLaMA-2 13B Wiki (↓)	LLaMA-2 70B 0-shot[8] Avg.(↑)	LLaMA-2 70B Wiki (↓)	LLaMA-3.2 1B 0-shot[8] Avg.(↑)	LLaMA-3.2 1B Wiki (↓)	LLaMA-3.2 3B 0-shot[8] Avg.(↑)	LLaMA-3.2 3B Wiki (↓)	LLaMA-3 8B 0-shot[8] Avg.(↑)	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**

SpinQuant: LLM quantization with learned rotations (ICLR 2025)

SpinQuant



End-to-end speed of LLaMA-3 8B on MacBook M1 CPU.



Average accuracy on 8 zero-shot commonsense reasoning tasks

SpinQuant: LLM quantization with learned rotations (ICLR 2025)