OSTQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformation for Better Distribution Fitting

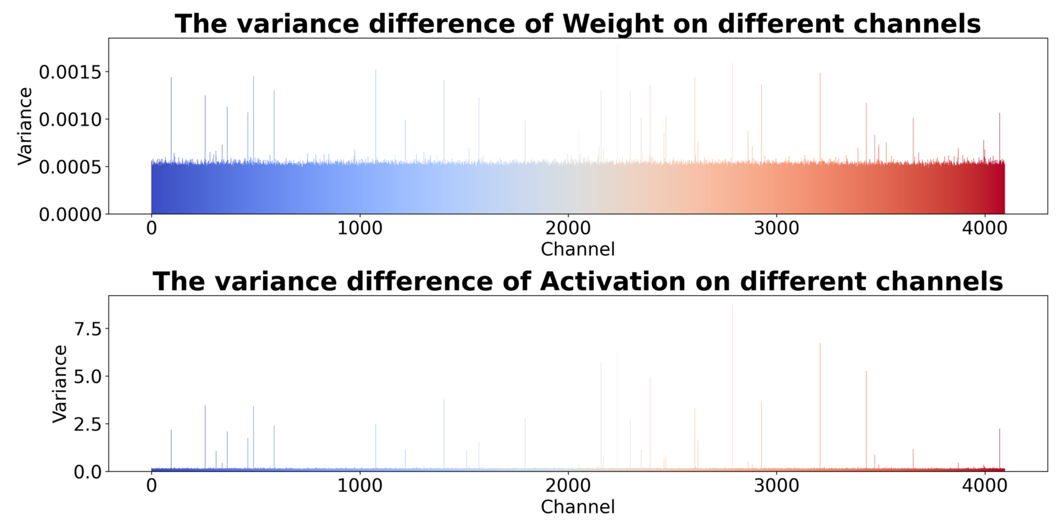
Xing Hu[1*], Yuan Cheng[12*], Dawei Yang[1*]✉, Zukang Xu[1], Sifan Zhou[3], Jiangyong Yu[1], Chen Xu[1], Zhe Jiang[3], Zhihang Yuan[1]✉

1. Houmo AI 2. Nanjing University 3. Southeast University

Paper Code Email

Challenges

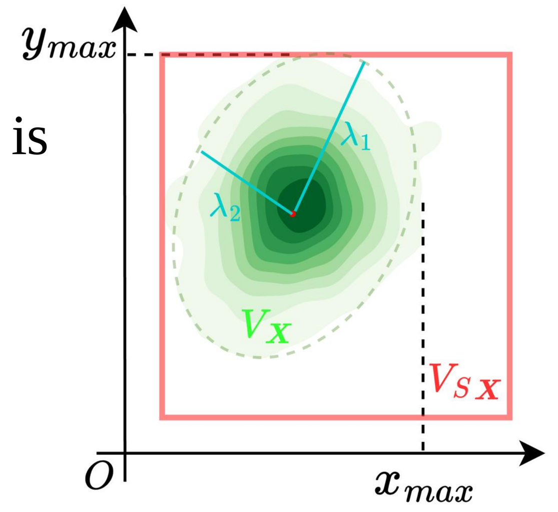
Uneven & **heavy-tailed** distributions in LLMs expands the quantization range, thereby making the quantization for LLMs challenging.



Previous approaches **remain heuristic** and do not optimize the distribution across the entire quantization space.

QSUR

The Quantization Space Utilization Rate(**QSUR**) of X is defined as the ratio of the quantization space V_{S_X} to the hypervolume occupied by $\boldsymbol{X}$:



$$\mathrm{QSUR}_X = \frac{V_X}{V_{S_X}} \qquad (1)$$

For $X \in N(\mu, \Sigma)$, We get:

$$\mathrm{QSUR}_X = \frac{\frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot \left(\chi_d^2(\alpha)\right)^{d/2} \cdot \sqrt{det(\Lambda)}}{2^d\left(max\left(\sqrt{\chi_d^2(\alpha)\cdot\lambda_1}\cdot \boldsymbol{q}_1\right)\right)^d} = \frac{\frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot \sqrt{\prod_{i=1}^{d}\lambda_i}}{2^d\left(max\left(\sqrt{\lambda_1}\cdot \boldsymbol{q}_1\right)\right)^d}$$

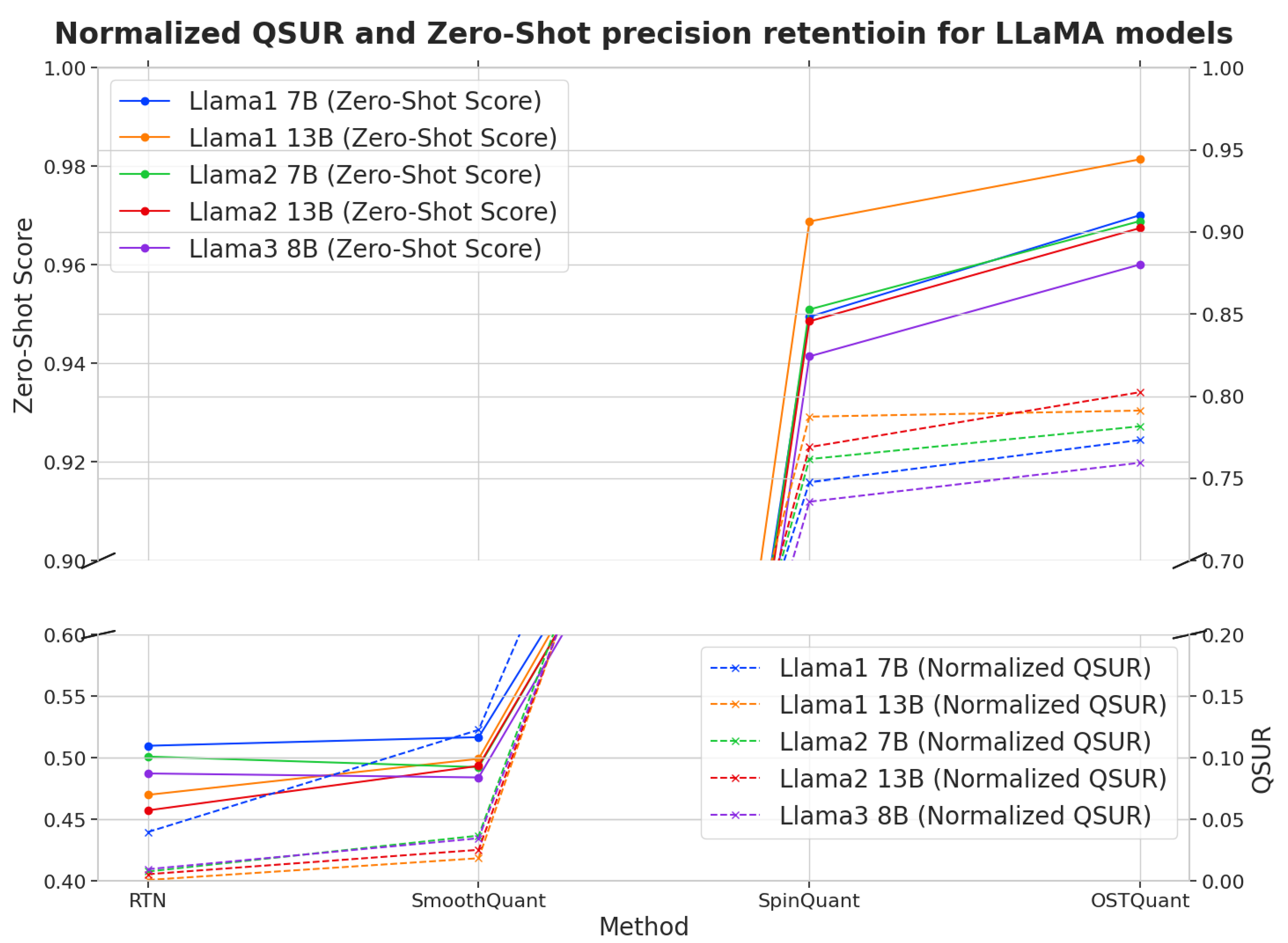
From equation above, We find that:

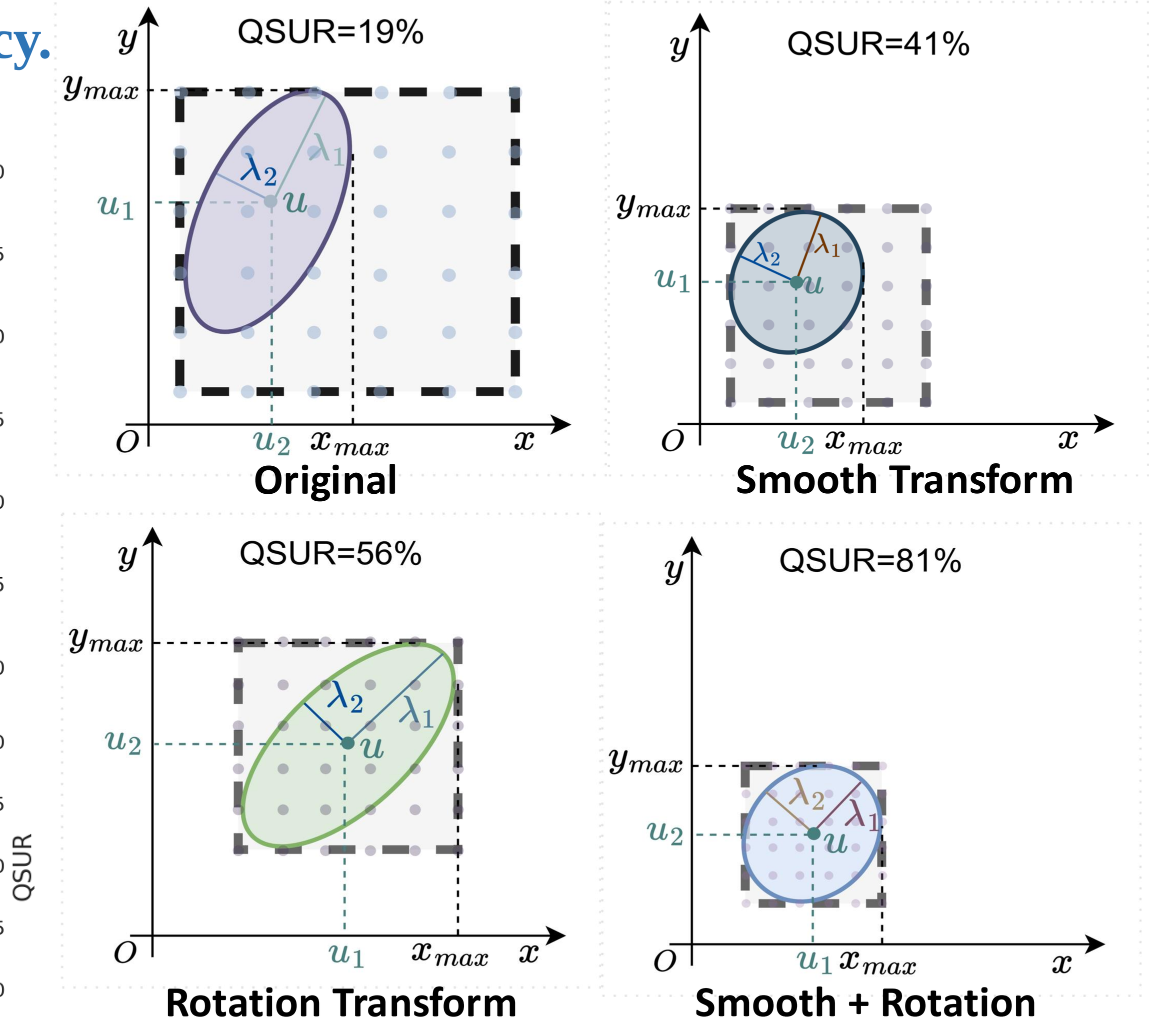
1) QSUR is proportional to the product of the ratios of each eigenvalue λ_i to the largest eigenvalue λ_1

2) The maximum component of the eigenvector q_1 is inversely proportional to QSUR.

Motivation

➤QSUR exhibits a positive correlation with accuracy.

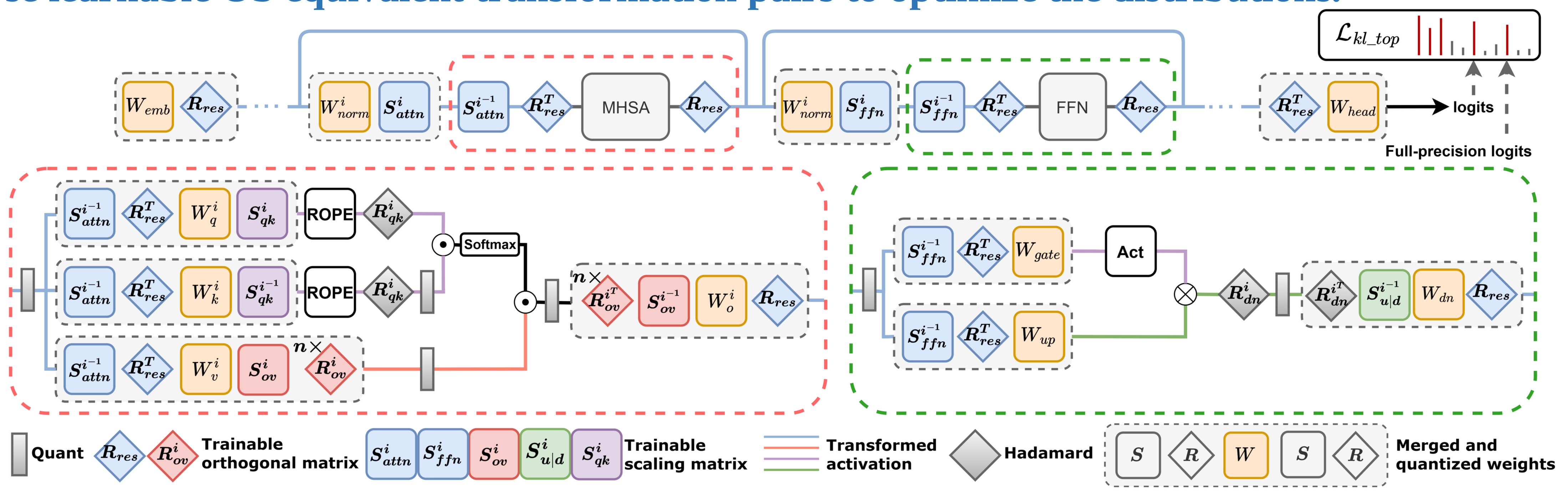




➤Orthogonal & Scaling enhance QSUR.

Method

➤Use learnable OS equivalent transformation pairs to optimize the distributions.



Equivalent Transformation Pair $\boldsymbol{T} = \boldsymbol{A}\boldsymbol{Q}^{\top}$**, forward like:**

$$y = Q(x\boldsymbol{W}_1\boldsymbol{O}\boldsymbol{\Lambda})Q(\boldsymbol{\Lambda}^{-1}\boldsymbol{O}^T\boldsymbol{W}_2)$$

Optimization objective:

$$\arg\min_{\boldsymbol{A}_i,\boldsymbol{O}_i} \mathcal{L}(\hat{y}, y; \boldsymbol{A}_i, \boldsymbol{O}_i, \theta)$$

Initialization:

$$\boldsymbol{O} = \boldsymbol{E}_{\boldsymbol{W}}\boldsymbol{H} \qquad \mathbf{A} = \boldsymbol{I}$$

➤ Use KL-Top loss function to focus optimization on model's main predictions.

$$idxs = \mathrm{topk}(\boldsymbol{z})$$

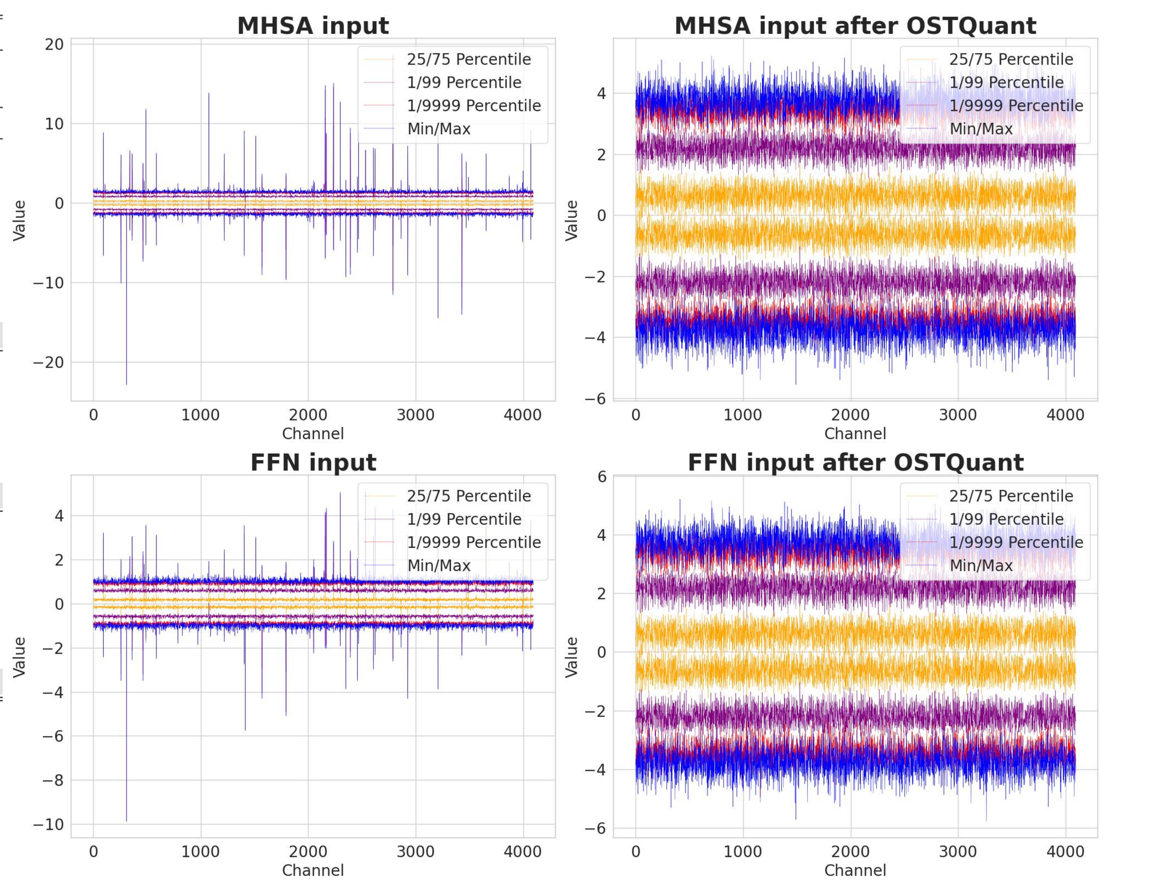
$$\mathcal{L} = \sum_{i \in idxs} \boldsymbol{z}[i] \log\left(\frac{\boldsymbol{z}[i]}{\hat{\boldsymbol{z}}[i]}\right)$$

Experiments

➤Performance across different tasks and models.

#Bits W-A-KV	Method	LLaMA-3 8B 0-shot[9] Avg.(↑)	LLaMA-3 8B Wiki (↓)	LLaMA-3 70B 0-shot[9] Avg.(↑)	LLaMA-3 70B Wiki (↓)	LLaMA-2 7B 0-shot[9] Avg.(↑)	LLaMA-2 7B Wiki (↓)	LLaMA-2 13B 0-shot[9] Avg.(↑)	LLaMA-2 13B Wiki (↓)	LLaMA-2 70B 0-shot[9] Avg.(↑)	LLaMA-2 70B Wiki (↓)	LLaMA 7B 0-shot[9] Avg.(↑)	LLaMA 7B Wiki (↓)	LLaMA 13B 0-shot[9] Avg.(↑)	LLaMA 13B Wiki (↓)	LLaMA 30B 0-shot[9] Avg.(↑)	LLaMA 30B Wiki (↓)
16-16-16	FloatingPoint	68.09	6.14	73.81	2.86	65.21	5.47	67.61	4.88	71.59	3.32	64.48	5.68	66.67	5.09	70.00	4.10
4-16-16	RTN	63.70	8.13	31.15	1e5	61.27	7.02	60.24	6.39	69.62	3.87	62.67	7.94	63.45	8.60	65.69	6.13
	SmoothQuant	62.79	8.12	67.94	6.70	58.88	8.03	62.03	5.86	65.93	5.50	62.24	7.46	62.69	18.75	65.69	5.80
	GPTQ	61.03	7.43	31.45	9e3	60.86	9.84	64.71	5.79	70.96	3.94	60.15	7.93	64.36	6.58	66.95	5.26
	Omniquant	65.66	7.19	-	-	63.19	5.74	66.38	5.02	71.04	3.47	63.42	5.86	66.22	5.21	69.07	4.25
	AWQ	67.03	7.36	68.92	5.92	63.89	5.83	66.25	5.07	70.88	4.03	63.30	5.97	65.58	5.28	69.44	4.28
	QuaRot	67.27	6.53	72.93	3.53	64.30	5.62	66.95	5.00	71.21	3.41	63.40	5.83	65.91	5.20	69.73	4.27
	SpinQuant	66.54	**6.49**	72.90	3.49	63.59	**5.58**	67.14	5.00	71.12	3.43	63.94	**5.76**	66.32	**5.16**	69.62	4.21
	OSTQuant	**67.80**	6.53	**73.69**	**3.19**	**64.37**	5.64	**67.31**	**4.94**	**71.48**	**3.41**	**64.13**	5.81	**66.62**	5.21	**69.84**	**4.19**
4-4-16	RTN	33.42	6e2	31.21	8e3	32.44	nan	30.86	8e3	30.90	7e4	32.51	7e3	31.63	3e4	31.57	2e3
	SmoothQuant	33.04	1e3	34.67	2e2	32.13	nan	34.26	1e3	35.86	3e2	34.42	3e2	33.29	6e2	34.64	1e3
	GPTQ	32.98	5e2	31.47	4e4	32.72	nan	30.11	4e3	30.86	nan	32.12	1e3	31.51	3e3	30.88	2e3
	QuaRot	61.69	8.02	65.56	6.35	61.87	6.05	65.13	5.35	69.96	3.78	61.76	6.22	64.46	5.50	68.14	4.57
	SpinQuant	64.11	7.28	66.99	6.10	57.37	6.78	63.23	5.24	70.58	3.68	61.82	6.08	64.59	**5.36**	68.08	4.53
	OSTQuant	**65.14**	**7.24**	**72.21**	**3.97**	**63.90**	**5.60**	**66.24**	**5.14**	**70.92**	**3.57**	**62.72**	**6.04**	**65.80**	5.40	**68.52**	**4.43**
4-4-4	RTN	33.18	7e2	30.82	8e3	32.67	nan	30.93	7e3	31.73	7e4	32.87	1e4	31.33	3e4	31.64	2e3
	SmoothQuant	32.96	1e3	33.76	3e2	32.12	nan	33.36	1e3	35.54	3e2	33.32	3e2	33.28	5e2	34.65	1e3
	GPTQ	33.71	6e2	31.20	4e4	33.52	nan	27.85	5e3	31.09	nan	31.80	2e3	30.63	3e3	31.07	2e3
	Omniquant	32.33	4e2	-	-	48.40	14.26	50.35	12.30	-	-	48.46	11.26	45.63	10.87	45.04	12.35
	QuaRot	61.38	8.18	65.33	6.6	61.48	6.11	65.16	5.39	70.30	3.80	61.22	6.26	64.59	5.53	68.08	4.60
	SpinQuant	64.10	7.35	66.31	6.24	62.01	5.96	64.13	5.74	70.57	3.61	61.32	6.12	64.95	**5.39**	68.14	4.55
	OSTQuant	**65.37**	**7.29**	**71.69**	**4.01**	**63.18**	**5.91**	**65.41**	**5.25**	**70.84**	**3.59**	**62.55**	**6.07**	**65.43**	5.40	**68.20**	**4.42**

➤Distribution pre&after OSTQuant



➤Speedup and Memory saving.

Model	Decoder Speed (tokens/sec) FP	Quantized	Speed up	Memory Use (GB) FP	Quantized	Memory Saving
LLaMA-2-7B	47.32	89.4	1.89x	13.94	4.32	3.23x
LLaMA-3-8B	38.33	77.71	2.03x	15.83	5.88	2.69x
LLaMA-2-13B	23.7	55.35	2.34x	23.7	8.5	2.79x
LLaMA-30B	OOM	30.49	-	OOM	18.19	-
LLaMA-3-70B	OOM	14.68	-	OOM	38.41	-

Model Size	Prefill Speedup (Seqlen) 256	512	1024	2048	4096	8192	Memory Saving Factor (Seqlen) 256	512	1024	2048	4096	8192
7B	2.24x	2.27x	2.23x	2.14x	2.11x	2.02x	3.48x	3.34x	3.12x	2.86x	2.57x	2.34x
8B	2.42x	2.52x	2.52x	2.43x	2.36x	2.23x	3.48x	3.36x	3.12x	2.77x	2.38x	2.00x
13B	2.62x	2.68x	2.63x	2.52x	2.83x	2.32x	3.64x	3.51x	3.30x	3.02x	2.70x	2.43x
30B	3.18x	3.01x	2.98x	3.40x	2.84x	2.68x	3.70x	3.59x	3.42x	3.15x	2.83x	2.53x

➤Ablation Study.

Ablation on loss function

Model	Loss Type	Wiki PPL	Arc-Easy Score	Arc-Challenge Score
LLaMA-2-7B	Origin	**5.38**	69.87	42.41
	KL-Top	5.94	**72.69**	**44.62**
LLaMA-2 13B	Origin	**5.12**	75.09	46.08
	KL-Top	5.25	**75.29**	**47.10**
LLaMA-3 8B	Origin	**6.80**	76.68	49.26
	KL-Top	7.29	**76.73**	**49.32**

Setting	Metric	k=5	k=50	k=100	k=500	k=1000	k=5000	k=10000
W3 Only	Zero-Shot[9] Score	61.87	61.88	61.75	62.18	**62.30**	61.25	61.21
	Wiki PPL	6.06	6.116	6.13	6.07	**6.06**	6.06	6.12
W4A4KV4	Zero-Shot[9] Score	62.4	62.13	62.38	62.34	**63.18**	62.44	62.11
	Wiki PPL	5.99	5.96	5.95	5.96	5.96	**5.93**	5.94

Ablation on methods of initialization

Model	Quant Setting	Method	Zero-Shot[9]	Wiki PPL
LLaMA-2-7B	Full-Precision	-	65.21	5.47
	W4A16KV16	Hadamard	63.32	5.62
	W4A16KV16	WOMI	63.45	5.59
	W4A4KV4	Hadamard	61.47	6.11
	W4A4KV4	WOMI	61.52	6.09
LLaMA-3-8B	Full-Precision	-	68.09	6.14
	W4A16KV16	Hadamard	67.27	6.53
	W4A16KV16	WOMI	67.41	6.48
	W4A4KV4	Hadamard	61.38	8.18
	W4A4KV4	WOMI	61.40	8.17

Ablation on different transformation matrices

Metric	Baseline	$+\boldsymbol{R}_{res}$	$+\boldsymbol{S}_{res}$	$+\boldsymbol{R}_{dn}$	$+\boldsymbol{S}_{u\|d}$	$+\boldsymbol{R}_{qk}$	$+\boldsymbol{S}_{qk}$	$+\boldsymbol{R}_{ov}$	$+\boldsymbol{S}_{ov}$
Wiki PPL	nan	9.70	9.46	6.16	6.00	5.92	5.92	5.94	**5.91**
Zero-shot[9]	33.51	54.33	53.74	61.75	61.79	62.35	62.56	63.11	**63.18**

Conclusion

➤Introduce QSUR as an effective metric and support it to guide optimization and method design.

➤OSTQuant: a fast and effective PTQ method helps quantization by optimizing distributions.

➤SOTA: shows strong performance, maintaining high accuracy even at extremely low bitwidths.