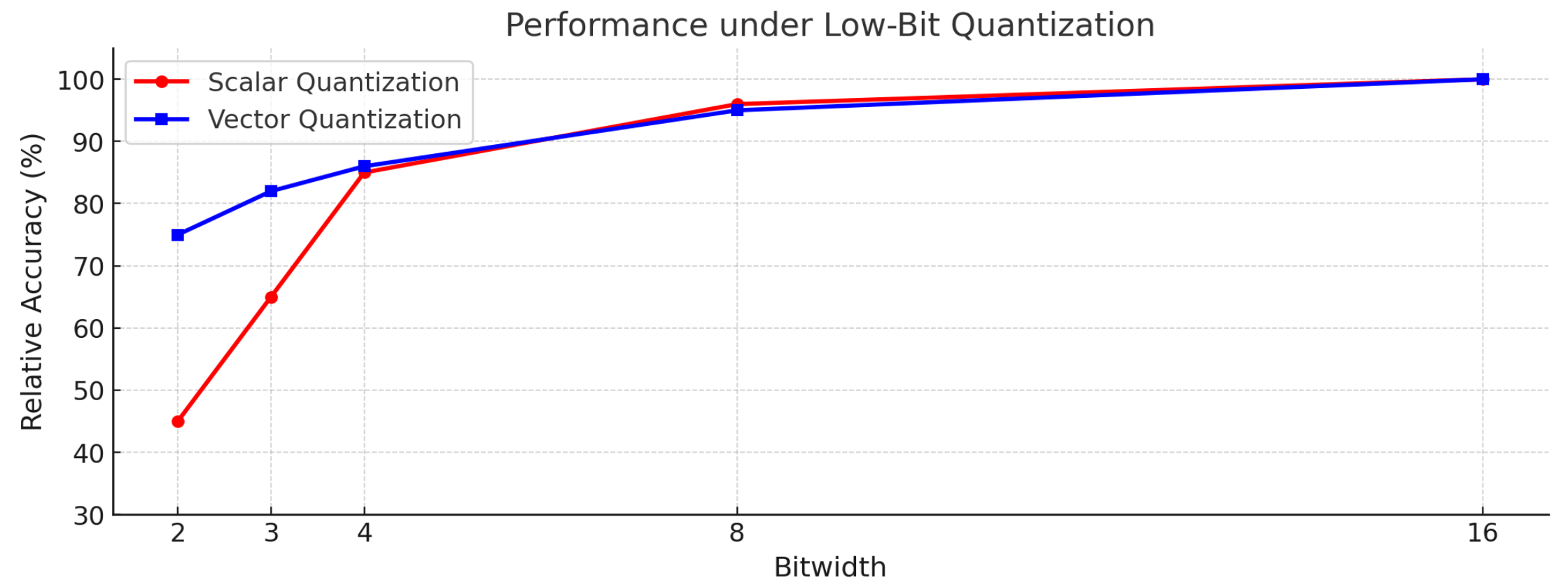
后摩智能
HOUMO.AI

RSAVQ: Riemannian Sensitivity-Aware VectorQuantization for Large Language Models

Zukang Xu*, Xing Hu*, Qiang Wu, Dawei Yang ✉

Motivation

Scalar Quantization struggles at low bitwidths



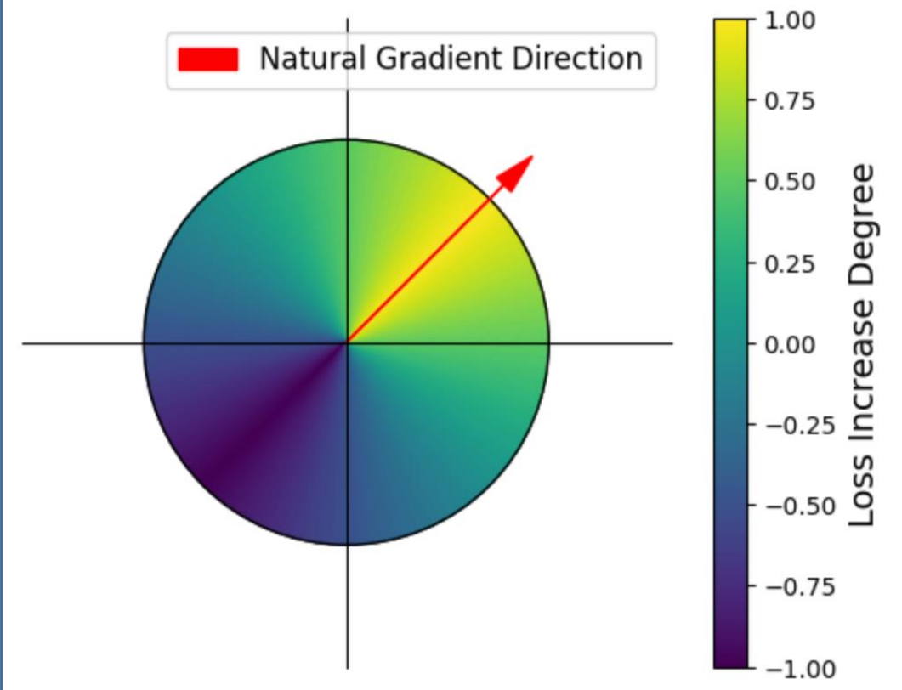
Vector Quantization shows strong potential under ultra-low bits.

However, existing explorations of VQ for LLMs remain limited:

- ➢Limited study in ultra-low-bit regimes
- ➢Ignoring error direction & channel sensitivity.
- ➢Inefficient for real-world deployment.

Challenge

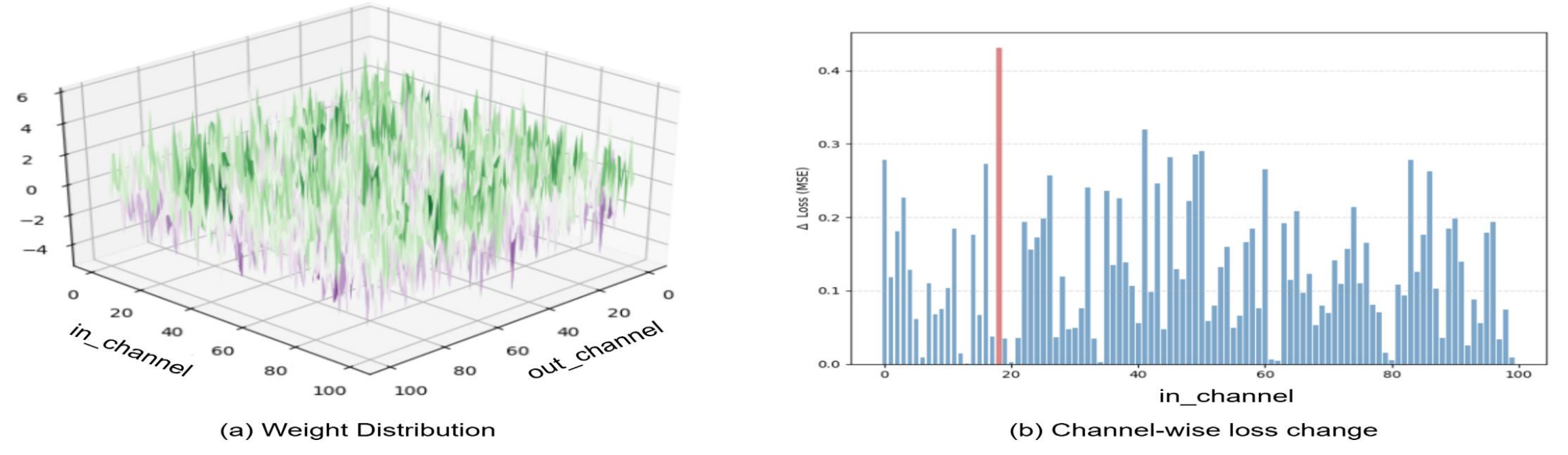
1. Error Direction Sensitivity Ignored



Most quantization methods assume isotropic error (Euclidean view), ignoring that errors along natural gradient directions have very different impacts.

2. Channel Sensitivity Overlooked

Current VQ and uniform quantization treat all weight channels equally, missing the large variance in channel-wise impact on loss.



(a) Weight Distribution

(b) Channel-wise loss change

Method(Part I): EDSG

Error Direction Sensitivity Guidance

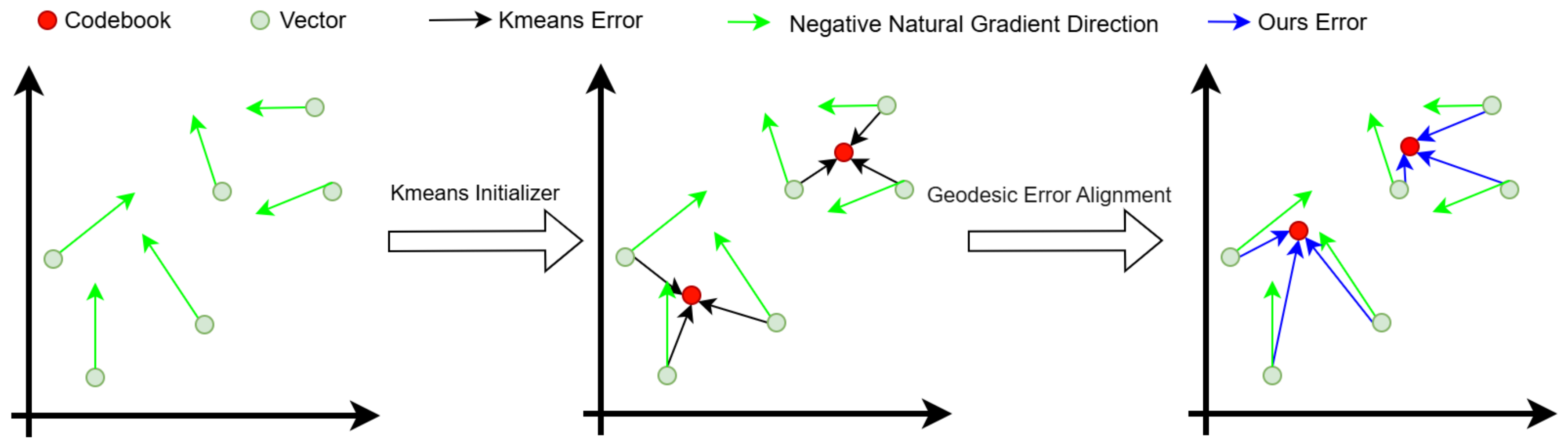
➢Quantization error inevitably occurs, defined as

$$E = W - \mathcal{C}(W)$$

➢We introduce an alignment loss to geometrically constrain quantization errors:

$$\mathcal{L}_{\text{project}} = \|E + \lambda * \tilde{\nabla}\mathcal{L}\|_{\mathbf{F}}^2$$

➢ Clustering process of error projection along negative natural gradient direction



Method(Part II): WCSG

Weight Channel Sensitivity Guidance

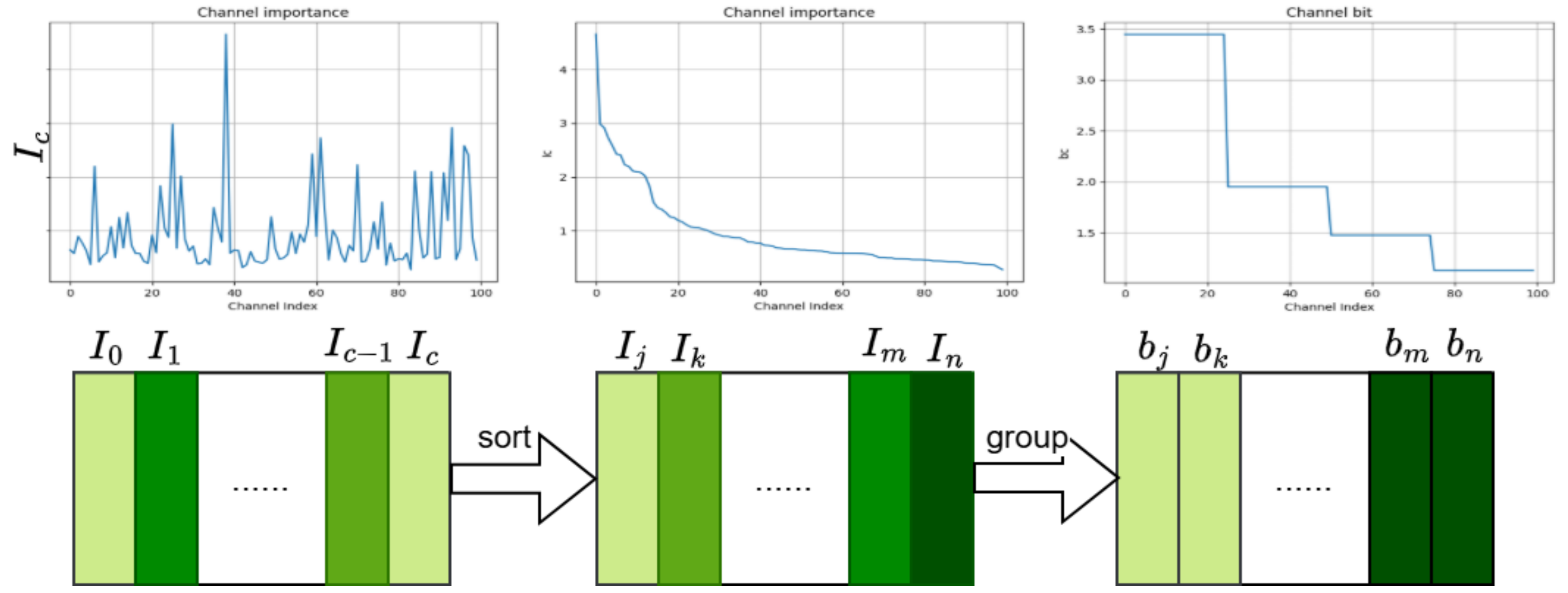
➢Channel sensitivity measured via Fisher geometry

$$I_c = \frac{1}{2}| - \tilde{\nabla}\mathcal{L}_c|_W^2 = \frac{1}{2}(\tilde{\nabla}\mathcal{L}_c)^\top \mathbf{F}_c \tilde{\nabla}\mathcal{L}_c.$$

➢Optimal Bit Allocation under Sensitivity-Aware Constraint

$$b_c = Round\left(B_{\max} \cdot \frac{\log_2 I_c}{\sum_{c=1}^{C} \log_2 I_c}\right) \cdot b_g = \text{Round}\left(\frac{1}{|G_g|}\sum_{c\in G_g} b_c\right)$$

➢ Channel sensitivity-driven channel grouping and bit assignment



Experiments

Performance Comparison of Quantization Methods on LLaMA under Low-Bit Settings

Methods	LLaMA-2 7B			LLaMA-2 13B			LLaMA-2 70B		
	Bits	W2↓	0-shot Avg↑	Bits	W2↓	0-shot Avg↑	Bits	W2↓	0-shot Avg↑
FP16	16	5.12	64.7	16	4.57	67.82	16	3.12	70.21
GPTQ	2	50.75	39.16	2	43.84	43.72	2	–	59.18
GPTVQ	2.25	6.71	56.14	2.25	5.72	61.56	2.25	4.25	68.55
DB-LLM	2.01	7.23	55.12	2.01	6.19	59.41	2.01	4.64	65.83
AQLM	2.29	6.29	58.57	2.18	5.41	61.58	2.07	3.94	68.75
VPTQ	2.02	6.13	58.13	2.02	5.32	62.37	2.07	3.93	68.61
QuIP#	2	6.19	58.22	2	5.35	61.96	2	3.91	68.94
RSAVQ	2	**5.97**	**58.66**	2	**5.29**	**62.84**	2	**3.55**	**69.05**
GPTQ	3	8.06	53.1	3	5.85	59.61	3	4.4	65.41
GPTVQ	3.125	5.44	62.69	3.125	4.8	59.63	3.125	–	–
AQLM	3.04	5.46	60.88	3.03	4.82	63.49	3.01	3.36	69.86
VPTQ	3.02	5.43	61.72	3.03	4.79	64.21	3.01	3.34	69.58
QuIP#	3	5.41	–	3	4.78	–	3	3.35	–
RSAVQ	3.01	**5.26**	**62.7**	3.01	**4.74**	**66.12**	3.01	**3.25**	**70.42**

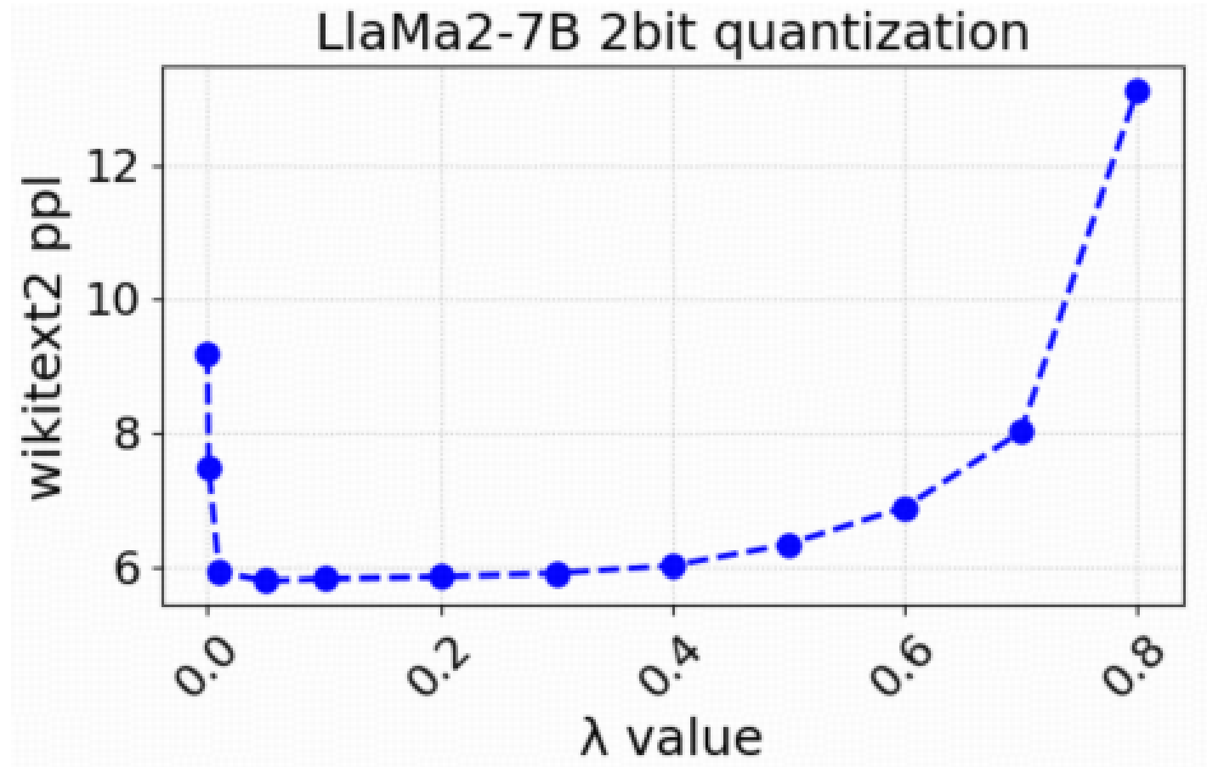
Methods	LLaMA-3 8B			LLaMA-3 70B		
	Bits	W2↓	0-shot Avg↑	Bits	W2↓	0-shot Avg↑
FP16	16	6.14	68.66	16	2.9	75.32
GPTQ	2	210	36.16	2	11.9	45.42
QuIP	2	85.1	36.81	2	13	48.66
QuIP#	2	9.11	–	2	5.6	–
VPTQ	2.08	9.29	60.22	2.07	5.66	70.74
RSAVQ	2	8.79	61.72	2	5.6	71.3
GPTQ	3	8.2	61.7	3	5.2	70.58
QuIP	3	7.5	63.72	3	4.7	72.56
QuIP#	3	6.77	–	3	3.8	–
VPTQ	3.03	6.97	66.66	3.01	3.81	73.68
RSAVQ	3.01	6.34	66.38	3.01	3.69	74.26

Ablation Experiment For RSAVQ

Bits	Methods	LLaMA-2 7B						
		W2↓	AC	AE	HE	QA	WI	Acc Avg↑
FP16		5.12	43.3	76.3	57.1	78.1	68.7	64.70
2bit	Kmeans	9.20	28.9	62.5	43.3	71.5	63.3	53.90
	+EDSG	7.29 (-1.91)	31.5	66.0	46.6	73.3	63.6	56.10 (+2.20)
	+WCSG	5.81 (-3.39)	37.2	64.4	50.7	75.4	65.7	58.69 (+4.79)
3bit	Kmeans	7.25	35.0	68.6	47.7	73.4	64.6	57.86
	+EDSG	5.63 (-1.62)	40.1	72.8	53.9	76.6	66.2	61.77 (+3.91)
	+WCSG	5.26 (-1.99)	41.0	73.0	54.7	76.7	68.2	62.70 (+4.84)

On LLaMA-2 7B, ablation results show that EDSG and WCSG improve 2-bit accuracy from **53.90 → 56.10 → 58.69.** Similar gains at 3-bit confirm both modules effectively reduce quantization error under ultra-low-bit settings.

Ablation Experiment For λ in EDSG



Ablation experiments show that RSAVQ achieves the best trade-off between error reduction and accuracy when **λ is tuned within \[0.01, 0.1]**, a range that remains robust across models under 2-bit quantization.

Conclusion

➢RSAVQ introduces a geometry-driven vector quantization framework that leverages error direction and channel sensitivity guidance to tackle extreme low-bit quantization in LLMs.

➢It bridges information geometry with quantization by modeling parameter space as a Riemannian manifold, enabling principled error control and adaptive bit allocation.

➢Experiments on LLaMA show state-of-the-art 2-bit performance, highlighting RSAVQ's potential for efficient LLM deployment in resource-constrained environments.