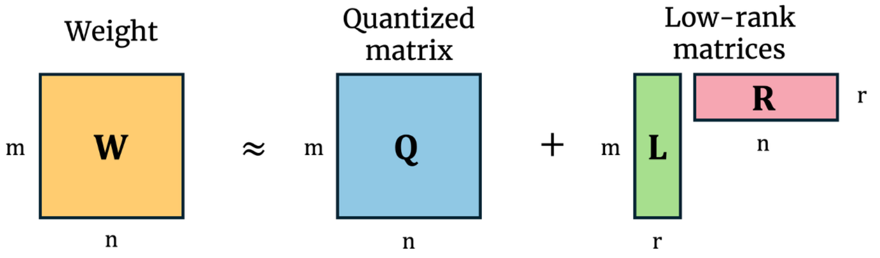
Preserve then Quantize: Dominant-Subspace Guided Low-Rank Reconstruction

Yoonjun Cho*, Dongjae Jeon*, Soeun Kim, Albert No

Preliminary

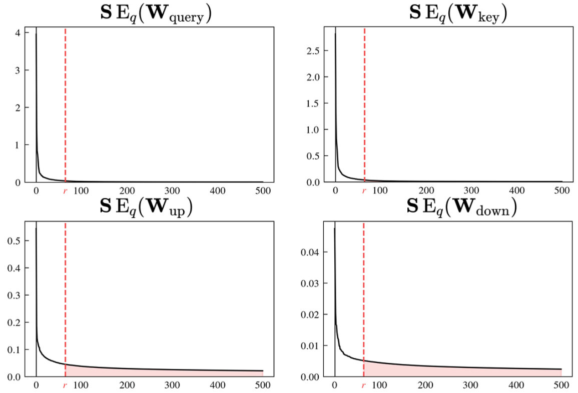
- **Quantization Error Reconstruction (QER):**
  - Approximates the quantization error $(\mathbf{W} - \mathbf{Q})$ with a low-rank term **LR**.
  - Activation statistics can be incorporated via a scaling matrix **S**.
  - **LR** is computed by SVD of scaled error $\mathbf{S}\,\mathrm{E}_q(\mathbf{W}) := \mathbf{S}(\mathbf{W} - \mathbf{Q})$

- **Quantized Parameter-Efficient Fine-Tuning (QPEFT):**
  - Only the low-rank **LR** is updated for downstream tasks, keeping **Q** fixed.



Is Quantization Error Sufficiently Low-Rank?

- **Problem:**
  - The scaled error $\mathbf{S}\,\mathrm{E}_q(\mathbf{W})$ is often not low-rank.
  - **LR** captures only a small portion of the error, resulting suboptimality



Capture Low-rank First, then Quantize Residual

- **Structured Residual Reconstruction (SRR):**
  - Dominant directions are preserved explicitly.
  - Quantization error remains small (only the low-energy tail is quantized).

$$\mathbf{W} = \mathbf{U}_h \mathbf{\Sigma}_h \mathbf{V}_h^\top + \mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{V}_\ell^\top$$

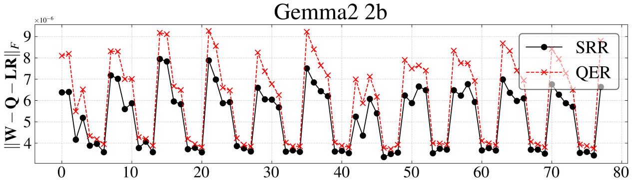
top low-rank
bypasses
quantization

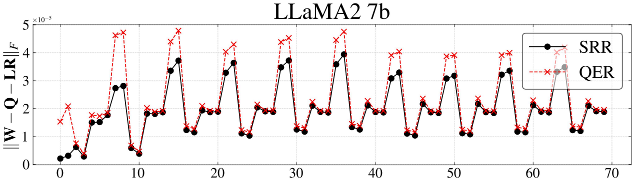
$$\mathbf{Q} := \mathcal{Q}\left(\mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{V}_\ell^\top\right)$$

$$\mathbf{LR} \leftarrow \mathbf{W} - \mathbf{Q} = \underbrace{\mathbf{U}_h \mathbf{\Sigma}_h \mathbf{V}_h^\top}_{\text{(1) preserved top-ranks}} + \underbrace{\mathrm{E}_q\left(\mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{V}_\ell^\top\right)}_{\text{(2) yields small error}}$$

SRR vs. QER: When It Works and When It Fails

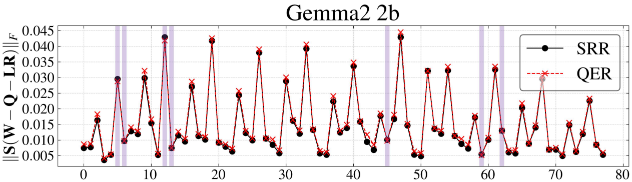
- **SRR** outperforms **QER** when activation-statistics are not used ($\mathbf{S} = \mathbf{I}$).

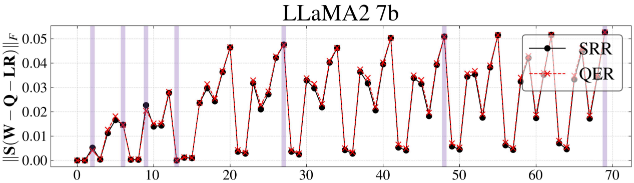




SRR vs. QER: When It Works and When It Fails

- Performance drops when activation-statistics are applied ($\mathbf{S} \neq \mathbf{I}$).





Mismatch with Activation-Statistics

- With activation statistics, quantization targets the component mapped back to **W**-space, not the low-energy tail of **SW**.

  - $\mathbf{S} = \mathbf{I}$ (SRR outperforms)

$$\mathbf{W} - \mathbf{Q} = \underbrace{\mathbf{U}_h \boldsymbol{\Sigma}_h \mathbf{V}_h^\top}_{\text{(1) preserved top-ranks}} + \underbrace{\mathrm{E}_q\left(\mathbf{U}_\ell \boldsymbol{\Sigma}_\ell \mathbf{V}_\ell^\top\right)}_{\text{(2) yields small error}}$$

  - $\mathbf{S} \neq \mathbf{I}$ (SRR fails)

$$\mathbf{S}(\mathbf{W} - \mathbf{Q}) = \underbrace{\mathbf{U}_h \boldsymbol{\Sigma}_h \mathbf{V}_h^\top}_{\text{(1) preserved top-ranks}} + \mathbf{S}\,\mathrm{E}_q\big(\underbrace{\mathbf{S}^{-1} \mathbf{U}_\ell \boldsymbol{\Sigma}_\ell \mathbf{V}_\ell^\top}_{\text{(2) Unknown}}\big)$$

Adaptive Strategy: Select Aligned Directions

- Only dominant directions in **SW** that are also important in **W** are preserved.

$$\mathbf{SW} = \sum_{i=1}^{n} \sigma_i u_i v_i^\top,$$

$$\text{score}_i = \sigma_i \|\mathbf{S}^{-1} u_i\|_2.$$

How each direction in **SW** contributes to **W**

$$\mathcal{H} := [r] \cap \text{Top-}r(\text{score}_i) \quad \mathcal{L} := [n] \setminus \mathcal{H}$$

$$\mathbf{SW} = \underbrace{\mathbf{U}_\mathcal{H} \boldsymbol{\Sigma}_\mathcal{H} \mathbf{V}_\mathcal{H}^\top}_{\text{(1) preserved top-ranks}} + \underbrace{\mathbf{U}_\mathcal{L} \boldsymbol{\Sigma}_\mathcal{L} \mathbf{V}_\mathcal{L}^\top}_{\text{(2) only quantized}}$$

Final decomposition

Adaptive SRR Wins QER

- **SRR** with the adaptive strategy outperforms **QER** in over **90%** of cases under optimal activation-statistics (QERA-exact[1]).

Method	Gemma-2 2B		LLaMA-2 7B	
	Win-rate (↑)	PPL (↓)	Win-rate (↑)	PPL (↓)
QERA-exact	-	19.36	-	10.68
w/ SRR (Naive)	76.37%	19.07	83.93%	10.61
w/ SRR (Adaptive)	**89.56%**	**18.65**	**95.98%**	**10.53**

(a) **Win-rate**: fraction of layers with lower reconstruction loss than QER (3-bit, $r = 64$).

[1] Zhang, Cheng, et al, “Qera: an analytical framework for quantization error reconstruction.”, *ICLR*, 2025.

PTQ results

- **SRR** outperforms under various scaling matrices **S**.

		Method	TinyLlama 1.1B		Gemma-2 2B		LLaMA-2 7B		LLaMA-2 13B		LLaMA-3.1 8B	
			$r=32$	$r=64$	$r=32$	$r=64$	$r=32$	$r=64$	$r=32$	$r=64$	$r=32$	$r=64$
		BF16	13.98		13.08		8.71		7.68		7.55	
Quantization Bits	**3.25**	*w-only*	32.82		41.13		13.33		10.25		18.96	
		ZeroQuant-V2 (Yao et al. 2024)	**28.31**	25.90	36.27	33.09	**13.18**	**12.99**	**10.04**	**10.03**	**20.09**	19.28
		w/ SRR	31.93	**25.18**	**26.77**	**24.71**	15.36	13.30	11.43	10.97	20.95	**18.44**
		LQER (Zhang et al. 2024a)	21.95	20.63	22.99	21.37	14.51	15.14	9.18	9.13	12.39	11.90
		w/ SRR	**21.10**	**19.86**	**22.61**	**21.02**	**11.24**	**11.05**	**9.12**	**9.00**	**12.27**	**11.76**
		QERA-approx (Zhang et al. 2025)	21.68	20.52	23.31	21.83	11.15	10.99	9.11	9.04	12.51	11.72
		w/ SRR	**20.83**	**19.54**	**22.02**	**19.98**	**10.92**	**10.75**	**9.05**	**8.95**	**11.99**	**11.45**
		QERA-exact (Zhang et al. 2025)	20.10	19.59	20.10	19.36	10.84	10.68	9.04	8.97	11.37	11.00
		w/ SRR	**19.61**	**18.70**	**19.55**	**18.65**	**10.76**	**10.53**	**9.01**	**8.90**	**11.20**	**10.74**

(b) Perplexity (↓) on WikiText2 with 3-bit MXINT quantizer under two low-rank settings ($r = 32, 64$).

PTQ in Iterative setting

- **SRR** shows consistent gains across iterations.

	Method	TinyLlama 1.1B			Gemma-2 2B			LLaMA-2 7B			LLaMA-3.1 8B		
		$i=1$	$i=5$	$i=10$	$i=1$	$i=5$	$i=10$	$i=1$	$i=5$	$i=10$	$i=1$	$i=5$	$i=10$
$r=32$	QERA-exact	20.10	19.41	19.15	19.59	18.89	18.83	10.84	10.69	10.63	11.37	11.04	10.97
	w/ SRR	**19.61**	**18.88**	**18.56**	**19.55**	**18.60**	**18.35**	**10.76**	**10.63**	**10.54**	**11.20**	**10.90**	**10.84**
$r=64$	QERA-exact	19.23	18.22	17.93	19.36	17.96	17.73	10.68	10.48	10.44	11.00	10.60	10.51
	w/ SRR	**18.70**	**17.77**	**17.58**	**18.65**	**17.33**	**17.00**	**10.53**	**10.37**	**10.30**	**10.76**	**10.39**	**10.28**

(c) Perplexity (↓) on WikiText2 with 3-bit MXINT after $i = 1$, 5, and 10 reconstruction steps. **SRR** vs. **QER** at ranks $r = 32$ and 64; best values in **bold**.

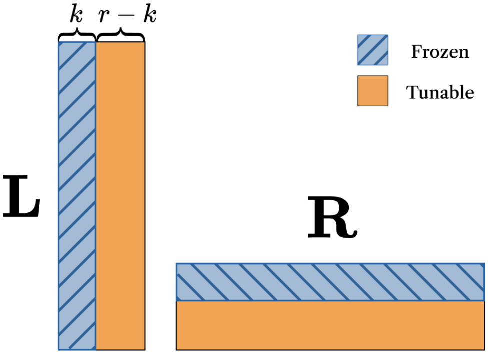
Applying SRR to QPEFT

- Fixed dominant directions; only residual subspace is updated.

$$\mathcal{H} := [r] \cap \text{Top-}r(\text{score}_i)$$

$$k = |\mathcal{H}|$$

Top-k directions dominate in both **SW** and **W**, but tuning them often degrades performance.



QPEFT results

- **SRR** outperforms baselines on GLUE across various bit-widths.

	Method	Rank	MNLI Acc.	QNLI Acc.	RTE Acc.	SST Acc.	MRPC Acc.	CoLA Matt.	QQP Acc.	STSB P/S Corr.	Avg.
16	Full FT	–	87.62	93.03	76.53	95.18	89.95	61.79	91.55	90.28/90.05	85.73
	LoRA (Hu et al., 2022)	8	87.59	92.68	72.76	95.07	89.76	61.08	90.95	90.09/89.84	84.92
4.25	QLoRA (Dettmers et al., 2023)	8	86.91	92.29	66.06	94.15	86.76	56.24	90.45	88.95/88.82	82.72
	LoftQ (Li et al., 2023)		**87.13**	91.63	64.26	93.46	87.75	59.07	90.46	88.95/88.84	82.83
	QERA (Zhang et al., 2025)		87.07	92.20	64.98	94.15	87.99	58.55	90.45	89.86/89.68	83.14
	LQ-LoRA (Guo et al., 2024)		85.89	90.96	54.15	92.32	82.35	42.60	88.67	85.89/85.73	77.84
	SRR		87.09	**92.64**	**72.20**	**94.84**	**88.48**	**60.58**	**90.48**	**90.06/89.77**	**84.53**
3.25	QLoRA (Dettmers et al., 2023)	8	86.14	90.76	54.87	90.83	78.92	10.83	89.91	86.77/86.28	73.60
	LoftQ (Li et al., 2023)		86.38	90.24	57.04	91.63	81.13	14.52	89.27	86.55/86.24	74.58
	QERA (Zhang et al., 2025)		**86.49**	89.46	57.40	91.74	84.56	28.98	89.26	**87.90/87.61**	76.95
	LQ-LoRA (Guo et al., 2024)		84.70	88.74	54.51	91.63	74.75	24.37	87.61	85.16/85.31	73.95
	SRR		86.06	**91.87**	**59.93**	**93.46**	**87.50**	**50.11**	**90.01**	87.97/87.50	**80.84**
2.50	QLoRA (Dettmers et al., 2023)	64	78.58	85.34	50.98	89.22	68.63	0	88.08	66.14/66.35	65.88
	LoftQ (Li et al., 2023)		81.30	86.63	50.37	91.06	71.08	0	88.48	82.63/82.85	68.96
	QERA (Zhang et al., 2025)		84.24	88.61	54.25	90.83	81.37	21.93	89.48	83.61/83.51	74.28
	LQ-LoRA (Guo et al., 2024)		83.33	87.26	52.71	89.79	71.83	0	88.32	78.45/79.39	69.02
	SRR		**85.64**	**90.96**	**59.57**	**92.89**	**85.78**	**38.22**	**90.24**	**87.43/87.13**	**78.82**

QPEFT results (Cont'd)

Quantization Bits		Method	Rank	LLaMA-2 7B (Δ_{acc})
	16	LoRA (Hu et al., 2022)	64	35.41
	4.25	QLoRA (Dettmers et al., 2023)	64	32.21
		LoftQ (Li et al., 2023)		28.35
		QERA (Zhang et al., 2025)		32.13
		LQ-LoRA (Guo et al., 2024)		29.82
		SRR		**32.87**
	2.50	QLoRA (Dettmers et al., 2023)	64	14.03
		LoftQ (Li et al., 2023)		15.69
		QERA (Zhang et al., 2025)		18.76
		LQ-LoRA (Guo et al., 2024)		16.67
		SRR		**18.95**

(d) GSM8K results for LLaMA-2 7B fine-tuned with PEFT under 4-/2-bit MXINT (block size 16/32, rank 64). LoftQ and LQ-LoRA use 5 iterations. Best accuracy in **bold**