

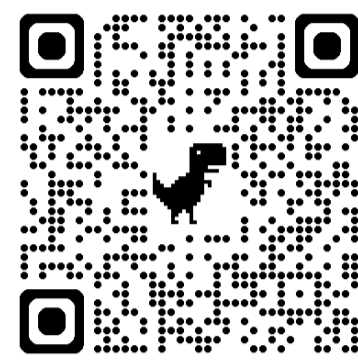
TTQ: Activation-Aware Test-Time Quantization to Accelerate LLM Inference On The Fly

Toshiaki Koike-Akino, Jing Liu, Ye Wang

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA.

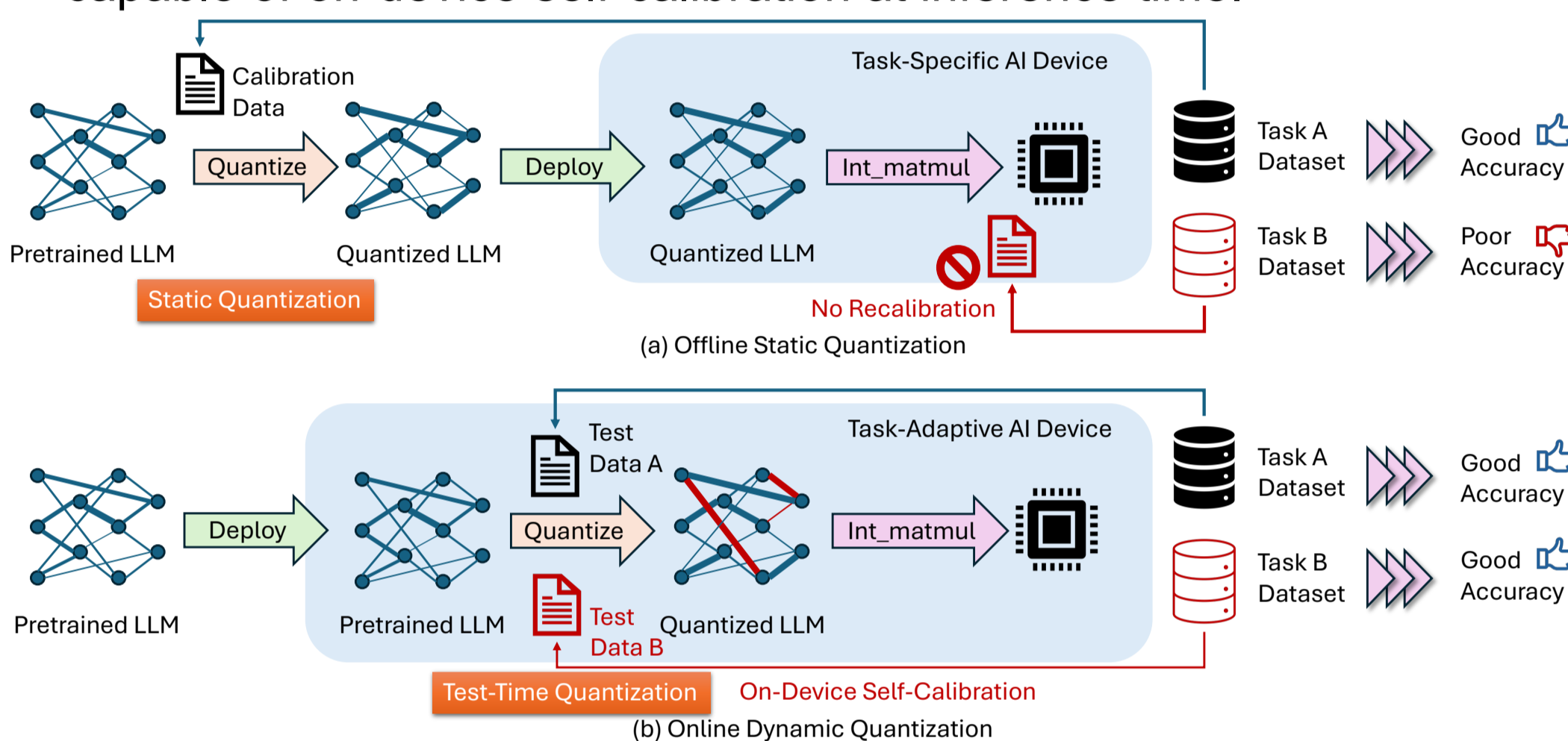
Highlights

- We propose **TTQ** to accelerate LLMs at inference time.
- We introduce activation-aware quantization to compress LLMs on the fly, with negligible overhead.
- We integrate low-rank decomposition into TTQ.
- We solve domain shift issue of baseline quantization.
- We demonstrate the benefit of TTQ for several LLM/VLM/VLA benchmarks.



Test-Time Quantization

- Offline static quantization (e.g., **AWQ**) requires offline calibration data, incurs domain shift, and cannot be recalibrated after deployment.
- TTQ is online dynamic quantization, with zero offline calibration, capable of on-device self-calibration at inference time.



Activation-Aware Instant Quantization

Online Activation-Aware Quantization: Given weights $W \in \mathbb{R}^{d' \times d}$ and input activations $X \in \mathbb{R}^{d' \times T}$, we want to minimize quantization error:

$$\mathcal{L} \triangleq \mathbb{E}_X [\| (W - \hat{W})X \|^2] = \| (W - \hat{W})C^{1/2} \|^2, \quad (1)$$

where $\hat{W} \in \mathbb{R}^{d' \times d}$ is quantized weights, and $C \triangleq \mathbb{E}_X [XX^T]$ is correlation. We use diagonal preconditioner depending on online test data X :

$$D = \text{diag}[(\|X_{:,i}\|_p^2 + \lambda)^\alpha].$$

to employ scaled quantization:

$$\hat{W} = \mathcal{Q}[WD^{1/2}]D^{-1/2},$$

where $\mathcal{Q}[\cdot]$ is groupwise quantization-dequantization.

Vanishing Overhead. The additional cost over offline AWQ:

$$\rho = \frac{\mathcal{O}(dT + 3d'd)}{\mathcal{O}(d'dT)} = \mathcal{O}\left(\frac{1}{d'} + \frac{1}{T}\right) \xrightarrow{d', T \gg 1} 0. \quad (2)$$

Thus, TTQ introduces *negligible extra complexity* at inference time.

Low-Rank Factors: We may integrate TTQ with low-rank decomposition:

$$\hat{W} = \mathcal{Q}[(W - BA)D^{1/2}]D^{-1/2} + BA, \quad (3)$$

where $B \in \mathbb{R}^{d' \times r}$ and $A \in \mathbb{R}^{r \times d}$ are rank- r factors. Low-rank projection has also vanishing complexity:

$$\rho' = \frac{\mathcal{O}[r(d' + d)T]}{\mathcal{O}[d'dT]} = \mathcal{O}\left(\frac{r}{d} + \frac{r}{d'}\right) \xrightarrow{r \ll d, d'} 0. \quad (4)$$

Runtime Analysis

- TTQ improves throughput up to **5-fold**, even with low-rank factors.

Table: Runtime Speed (k tokens/sec) of Query Projection Module for **Qwen3** Models with 4-bit AWQ and TTQ, on NVIDIA RTX4090 GPU.

Qwen3	0.6B	1.7B	4B	8B	14B	32B
FP16	116.82	77.04	72.45	58.44	46.78	10.76
AWQ (awq_gemm)	113.84	90.42	80.19	62.94	51.32	50.45
AWQ (marlin_gemm)	120.62	115.20	112.61	101.00	80.45	72.34
TTQ ($r=0$)	108.00	104.11	103.45	92.23	74.72	67.40
TTQ ($r=16$)	77.88	73.47	72.55	66.62	58.65	53.17

LLM Experiments

- Benefit in LLM benchmark for **OPT**, **Qwen3**, and **Gemma3**.

Table: Calibration length impact at 3-bit quantization with $g = 32$ for **OPT-350M** model.

Calib Tokens T	TTQ		AWQ (C4 Calib)						
	$0_{(r=0)}$	$0_{(r=16)}$	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}
WT2 Perplexity (\downarrow)	24.93	24.17	25.73	25.56	25.62	25.55	25.42	25.07	25.02

Table: Groupsize impact on **WT2** perplexity at 3-bit quantization for **Qwen3-1.7B** model.

Groupsize g	8	16	32	64	128	256	512
RTN	33.39	66.61	120.89	114.97	1450.28	15503.10	949478.44
AWQ (WT2 Calib)	17.64	19.00	20.35	20.69	24.96	31.41	36.14
TTQ ($r=16$)	17.54	17.81	18.29	19.46	22.33	24.81	31.43

Table: Perplexity (\downarrow) of quantized OPT/Qwen3/Gemma3 models. It shows macro average across **WT2/PTB/C4** datasets. Groupsize is $g = 32$ for all cases. Calibration token length is $T = 2^{17}$ for AWQ. **Bold** and underline denote the best and second best, respectively. Asterisk "*" indicates competitive performance to the un-compressed LLM.

Quantization q	2 bits	3 bits	4 bits	5 bits	Quantization q	2 bits	3 bits	4 bits	5 bits
OPT-125M: PPL 31.1					Qwen3-1.7B: PPL 24.2				
RTN	5058.5	56.3	33.5	31.8	RTN	1.4e6	162.8	30.6	26.1
AWQ (WT2 Calib)	381.7	37.4	32.3	31.4	AWQ (WT2 Calib)	1864.7	30.0	24.5	24.4
AWQ (PTB Calib)	375.3	37.3	32.2	31.3	AWQ (PTB Calib)	2309.6	29.8	24.7	24.5
AWQ (C4 Calib)	451.7	37.7	32.6	31.3	AWQ (C4 Calib)	2364.7	28.2	24.9	24.4
TTQ ($r=0$)	257.4	36.6	31.9	31.2	TTQ ($r=0$)	522.6	27.3	24.4	24.3
TTQ ($r=16$)	141.7	35.8	31.8	*31.1	TTQ ($r=16$)	264.6	26.4	24.3	*24.1
OPT-1.3B: PPL 17.0					Gemma3-1B: PPL 91.1				
RTN	11514.4	27.2	18.1	17.2	RTN	8.6e5	209.4	111.1	96.6
AWQ (WT2 Calib)	32.4	18.0	17.3	*17.0	AWQ (WT2 Calib)	4734.7	134.7	91.9	95.9
AWQ (PTB Calib)	32.6	18.1	17.3	*17.0	AWQ (PTB Calib)	9326.6	138.9	99.2	95.5
AWQ (C4 Calib)	31.7	18.0	17.2	*17.0	AWQ (C4 Calib)	5486.9	150.8	93.5	93.6
TTQ ($r=0$)	32.2	18.2	17.2	*17.0	TTQ ($r=0$)	1928.5	127.3	*89.9	*90.2
TTQ ($r=16$)	27.2	17.9	17.1	*17.0	TTQ ($r=16$)	1804.9	114.5	91.7	*90.3

VLM Experiments: Visual Reasoning QA

- Benefit for **Qwen3-VL** VLM in **TextVQA** visual reasoning tasks.

Table: Accuracy (\uparrow) of Qwen3-VL models on TextVQA benchmark ($g = 32$).

Quantization q	2 bits	3 bits	4 bits	5 bits
Qwen3-VL-4B: Acc 81.44%				
RTN	0.03%	74.79%	80.77%	*81.47%
AWQ (COCO-Cap Calib)	0.13%	78.68%	80.49%	81.20%
AWQ (OK-VQA Calib)	0.18%	77.89%	80.58%	80.75%
AWQ (ChartQA Calib)	0.42%	77.37%	80.67%	*81.47%
AWQ (TextVQA Calib)	0.25%	77.49%	80.28%	81.01%
TTQ ($r=0$)	1.51%	79.93%	81.29%	*81.48%
TTQ ($r=16$)	7.47%	79.45%	81.22%	*81.49%
Qwen3-VL-8B: Acc 81.72%				
RTN	0.17%	78.51%	80.63%	*81.73%
AWQ (COCO-Cap Calib)	41.39%	80.37%	81.52%	*81.79%
AWQ (OK-VQA Calib)	31.62%	79.51%	81.01%	81.55%
AWQ (ChartQA Calib)	41.17%	78.65%	81.23%	81.39%
AWQ (TextVQA Calib)	39.60%	78.43%	81.25%	81.69%
TTQ ($r=0$)	42.56%	80.77%	81.57%	*81.81%
TTQ ($r=16$)	48.01%	81.04%	81.71%	*81.85%

VLA Experiments: Robot Manipulation Task

- $\pi_{0.5}$ VLA models for **LIBERO** robot manipulation tasks ($g = 2$).

Benchmark	Libero Spatial	Libero Object	Libero Goal	Libero Long	Avg
BF16	97.5%	100.0%	97.0%	96.5%	97.75%
RTN	57.0%	65.0%	27.5%	2.0%	37.88%
AWQ (Spatial Calib)	90.5%	100.0%	85.0%	82.0%	89.34%
AWQ (Object Calib)	91.5%	98.5%	92.0%	78.0%	90.00%
AWQ (Goal Calib)	92.5%	100.0%	93.5%	84.5%	92.63%
AWQ (10 Calib)	94.0%	99.5%	92.5%	76.5%	90.63%
TTQ ($r=0$)	93.0%	99.5%	93.5%	87.0%	93.25%
TTQ ($r=16$)	94.5%	100.0%	93.5%	87.5%	93.88%