Table A: Multi-scale (grouping) performance for LLaMA 7B and 13B on Wikitext2 where g indicates the group size. The perplexity consistently decreases as we take on more learnable parameters. Our table provides the results of a 3-bit group-wise OPTQ experiment conducted on LLaMA 7B and 13B, which can be compared with PEQA. Please note that there is a typographical error in Table 8 where the order of LLaMA 13B and 7B was mistakenly switched. The PEQA results remain as shown in Table 8, and a 3-bit group-wise experiment was additionally conducted for LoRA+OPTQ.

Model	Method	W Bits	Channel-Wise	$g256$	$g128$	$g64$
LLaMA 7B	LoRA+OPTQ	3	17.79	10.15	12.92	10.12
	PEQA	3	**6.19**	**5.96**	**5.91**	**5.89**
LLaMA 13B	LoRA+OPTQ	3	7.53	5.76	5.61	5.48
	PEQA	3	**5.54**	**5.40**	**5.37**	**5.34**

Table B: Comparison of fine-tuning zero-points only, quantization scales only (PEQA), and both quantization scales and zero-points for LLaMA 7B and LLaMA 13B on Wikitext2 with only weights being quantized into 4-bit. The lower PPL, the better.

Model	Zero-points only	Scales only (PEQA)	Both zero-points and scales
LLaMA 7B	11.56	5.84	5.86
LLaMA 13B	9.83	5.30	5.34

Table C: Comparison of Quantization-Aware PEFT with full fine-tuning, PEFT, PEFT+PTQ, PTQ+PEFT using LLaMA 65B on the DRAM usage and training time during fine-tuning, the DRAM storage for deployment, the inference acceleration, and task-switching efficiency. The DRAM usage estimation for PEFT is based on LoRA.

	Fine-Tuning		Deployment		
Method	DRAM	Tuning Speed	DRAM	Inference Speed	Task-Switching
Full Fine-Tuning	457GB	Slow	131GB	Slow	Slow
PEFT	131GB	Fast	131GB	Slow	Fast
PEFT+PTQ	131GB	Fast	33GB	Fast	Slow
PTQ+PEFT	33GB	Fast	33GB	Slow	Fast
Quantization-Aware PEFT (Ours)	**33GB**	**Fast**	**33GB**	**Fast**	**Fast**

Table D: Peak memory usage of PEQA and LoRA for fine-tuning LLaMA 7B on Wikitext2 on a single NVIDIA A100 80GB GPU without gradient accumulation.

Batch Size	LoRA	PEQA
2	59GB	**43GB**
4	OOM	**80GB**