Cherry on Top: Parameter Heterogeneity and Quantization in Large Language Models

Wanyun Cui and Qianle Wang

Shanghai University of Finance and Economics

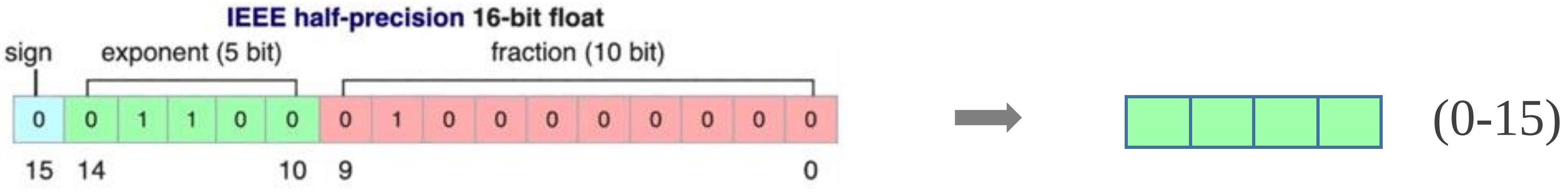
Table of Contents

The (unexpected) high robustness of LLMs to quantization

Cherry on Top: Parameter Heterogeneity and Quantization in Large Language Models

What is quantization?

- **Reduction of 32/16-bit models to low-bit counterparts**



FP16 INT4

- **Despite the extreme low bits in quantized models, what causes the high robustness of LLMs to quantization?**

Full precision

4-bit precision

FP16	3.32
RTN	3.46(+0.14)
GPTQ	3.42(+0.10)
AWQ	3.41**(+0.09)**

Perplexity of Different Quantization Methods on LLaMA-2-70B

No significant performance degradation

Parameter Heterogeneity Phenomenon

Cherry on Top: Parameter Heterogeneity and Quantization in Large Language Models

Heterogeneity can explain

➢ ***Heterogeneity* is not an isolated case!**

The significant variation in the influence of quantization on different parameters

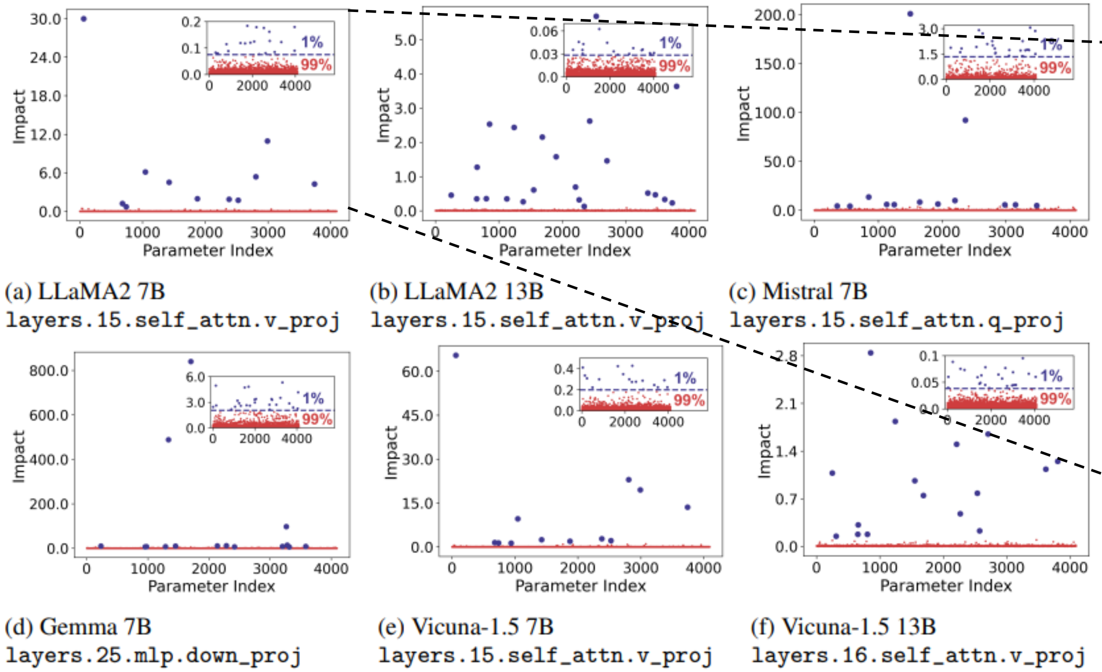
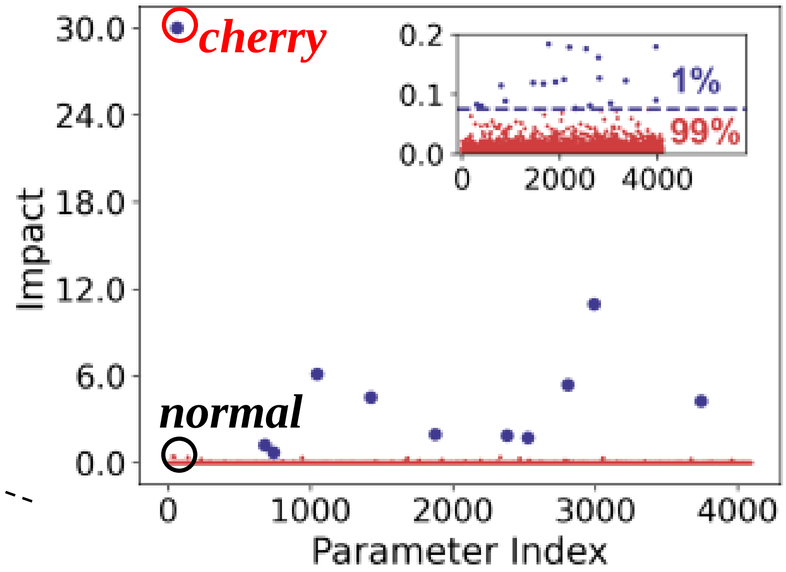


Figure 1: Scatter plot of parameter impacts in different LLMs. We randomly sampled 4096 parameters from the corresponding parameter matrix. Each point represents the impact of an individual parameter. Insets show the zoomed-in y-axis. The heterogeneity is found across different model scales (1a,1b), different model families (1c, 1d), and both base models and chat models (1e, 1f).



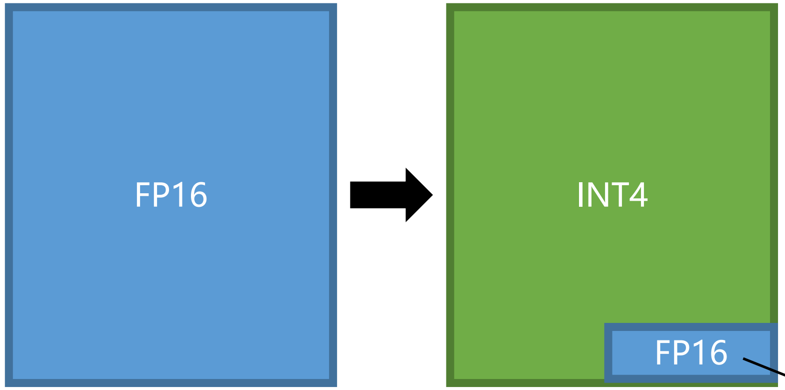
(a) LLaMA2 7B
`layers.15.self_attn.v_proj`

Heterogeneity can explain

- **Why LLMs are tolerant of quantization?**

  For the vast majority (> 99%) of parameters, the effect of their quantization to the model are minimal and can thus be alleviated or ignored

- **Why is mixed-precision quantization effective?**



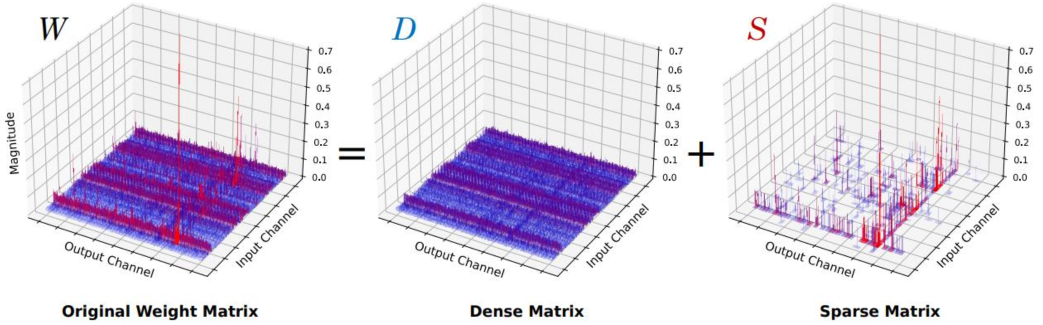
Mixed-precision quantization

< 1%

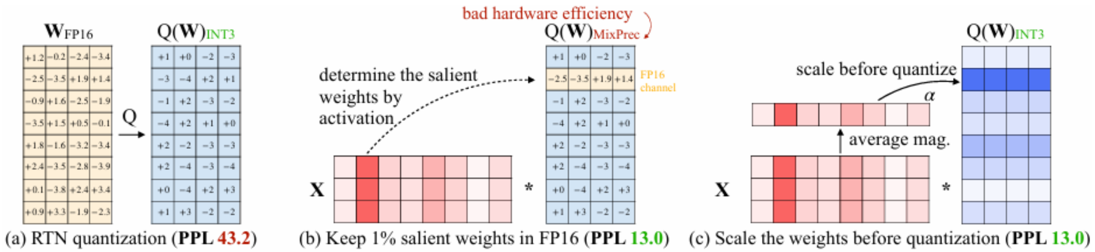
By preserving a small proportion of parameters with high precision, the quantization performance can be effectively improved

What is the best ***Cherry*** parameter selection strategy?

➢ **Weight-based strategy**



➢ **Activation-based strategy**



➢ **Impact-based strategy**

$$\mathbf{Impact}(w_i) = \mathbf{L}(w_i + \Delta w) - \mathbf{L}(w_i)$$

➢ **Impact > Activation > Weight?**

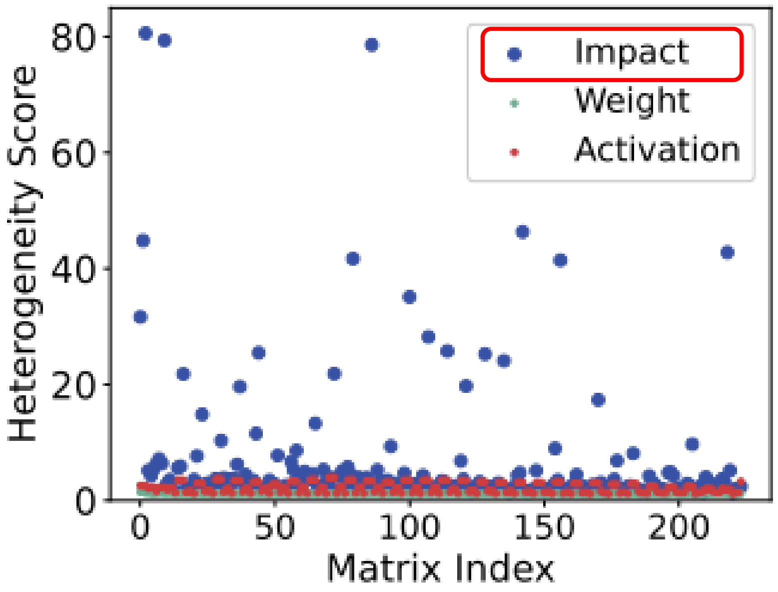
Effective strategy should exhibit high heterogeneity that differentiates the influence of cherry parameters and normal parameters of the model!

[1] Kim, Sehoon, et al. *Forty-first International Conference on Machine Learning*
[2] Lin, Ji, et al. *Proceedings of Machine Learning and Systems* 6 (2024): 87-100

What is the best *Cherry* parameter selection strategy?

- **Based on *heterogeneity score*, what is the best strategy?**

$$\textbf{Heterogeneity Score}(f) = \frac{\textbf{Mean}(f(w_i)_{\text{top1\%}})}{\textbf{Max}(f(w_i)_{\text{bottom1\%}})}$$



(a) LLaMA-2 7B

The ***Impact*** metric better distinguishes between the normal and cherry parameters, thus the best Cherry parameter selection strategy!

How to quantize parameters according to parameter heterogeneity?

- ***CherryQ*: End-to-End Mixed Precision Quantization**
  - ***Cherry*** parameters are updated using standard gradient descent
  - Normal parameters employ the Straight-Through Estimator (STE) trick for low-precision gradient descent

Algorithm 1 Cherry Parameter and Quantization-aware Training (CherryQ)

Require: Model parameters $\mathbf{W}$, quantization function $Quant(\cdot)$, threshold τ, learning rate η
Ensure: Quantized model parameters

1: $\mathbf{C} \leftarrow \{w_i \in \mathbf{W} \mid H_{ii} > \tau\}$ ▷ Identify cherry parameters
2: $\mathbf{N} \leftarrow \mathbf{W} \setminus \mathbf{C}$ ▷ Identify normal parameters
3: **for** each training batch x **do**
4: $\quad L \leftarrow \text{model}(x; \mathbf{C} \cup Quant(\mathbf{N}))$ ▷ Compute loss
5: $\quad \mathbf{C} \leftarrow \mathbf{C} - \eta \frac{\partial L}{\partial \mathbf{C}}$ ▷ Standard gradient descent
6: $\quad \mathbf{N} \leftarrow \mathbf{N} - \eta \cdot \text{STE}(\frac{\partial L}{\partial \mathbf{N}})$ ▷ Gradient descent with gradient approximation by STE
7: **end for**
8: **return** $\mathbf{C} \cup Quant(\mathbf{N})$

[1] Liu, Zechun, et al. *arXiv preprint arXiv:2305.17888* (2023).

Experiments

Cherry on Top: Parameter Heterogeneity and Quantization in Large Language Models

3/4-bit quantization experiment

➢ Perplexity results

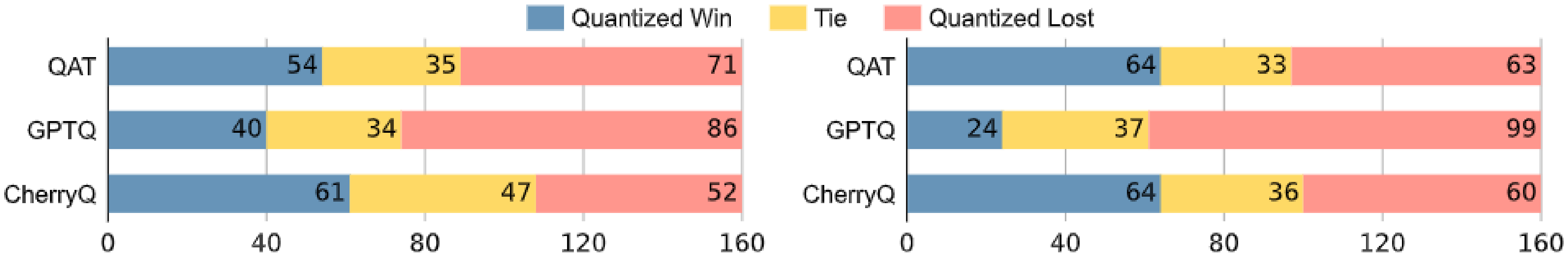
Method	Avg. bit	7B-3bit-g128		Avg. bit	7B-3bit-g64		Avg. bit	13B-3bit-g128		Avg. bit	13B-3bit-g64	
		c4	wiki2		c4	wiki2		c4	wiki2		c4	wiki2
FP16	16	6.97	5.47	16	6.97	5.47	16	6.47	4.88	16	6.47	4.88
QAT	3.13	9.25	6.90	3.25	8.74	7.13	3.13	7.19	5.63	3.25	7.02	5.48
GPTQ	3.15	8.28	6.74	3.30	8.20	6.62	3.15	7.24	5.63	3.30	7.10	5.56
AWQ	3.15	7.84	6.24	-	-	-	3.15	6.94	5.32	-	-	-
OmniQuant	3.15	7.75	6.03	-	-	-	3.15	6.98	5.28	-	-	-
SqueezeLLM	-	-	-	3.24	7.51	5.96	-	-	-	3.24	6.82	5.23
CherryQ	3.17	**7.39**	**5.93**	3.30	**7.34**	**5.87**	3.17	**6.80**	**5.26**	3.29	**6.76**	**5.21**

Perplexity (↓) of 3-bit quantization on LLaMA2 models. gX means the group size is X

CherryQ consistently outperforms all other approaches across both model sizes (7B and 13B) and grouping sizes (64 and 128), achieving the lowest perplexity on both the C4 and WikiText-2 datasets

3/4-bit quantization experiment

- **Effect of Chat LLM Quantization**



Comparison of 3-bit quantized models to FP16 Vicuna-1.5
(Left) Comparisons to Vicuna1.5-7B (Right) Comparisons to Vicuna-1.5-13B
CherryQ even shows competitive quality compared to the 16-bit counterpart

2-bit quantization experiment

➢ Perplexity results

Method	LLaMA2-7B-2bit		LLaMA2-13B-2bit	
	c4	wiki2	c4	wiki2
FP16	6.97	5.47	6.47	4.88
GPTQ-g64	19.40	20.85	12.48	22.44
AWQ-g64	$> 10^5$	$> 10^5$	$> 10^4$	$> 10^5$
OmniQuant-g64	12.72	9.62	10.05	7.56
CherryQ-g64	**9.08**	**7.84**	**8.02**	**6.72**
FP16	6.97	5.47	6.47	4.88
GPTQ-g128	33.70	36.77	20.97	28.14
AWQ-g128	$> 10^5$	$> 10^5$	$> 10^4$	$> 10^5$
OmniQuant-g128	15.02	11.06	11.05	8.26
CherryQ-g128	**9.55**	**8.34**	**8.40**	**7.20**

Perplexity (↓) of 2-bit quantization on LLaMA2 models

Compared to existing methods such as GPTQ, AWQ, and OmniQuant, our proposed CherryQ method demonstrates superior performance across all metrics

Thanks!

Cherry on Top: Parameter Heterogeneity and Quantization in Large Language Models