

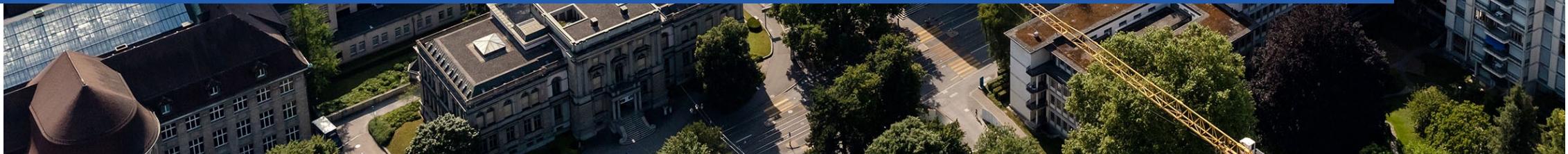


Accurate LoRA-Finetuning Quantization of LLMs via Information Retention

Haotong Qin¹, Xudong Ma², Xingyu Zheng², Xiaoyang Li³, Yang Zhang³, Shouda Liu³, Jie Luo², Xianglong Liu², Michele Magno¹

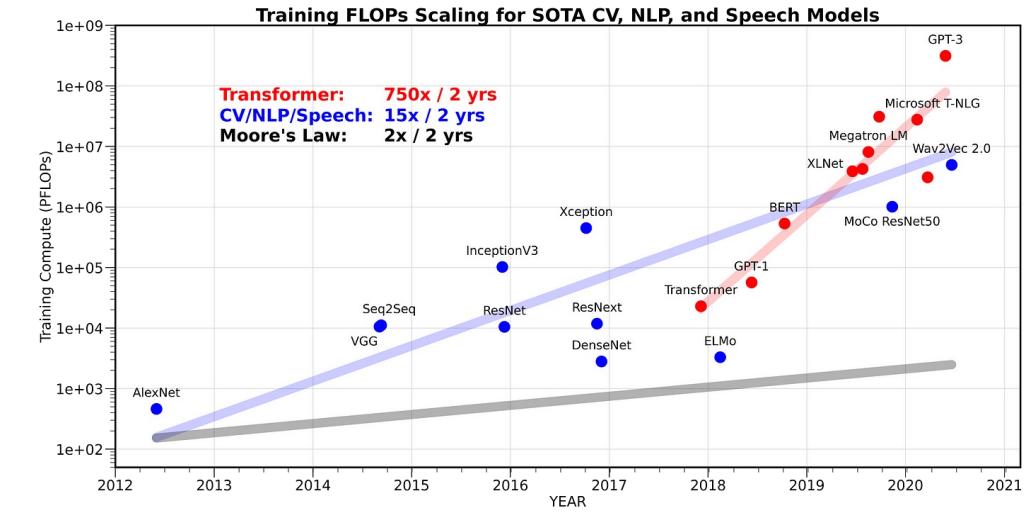
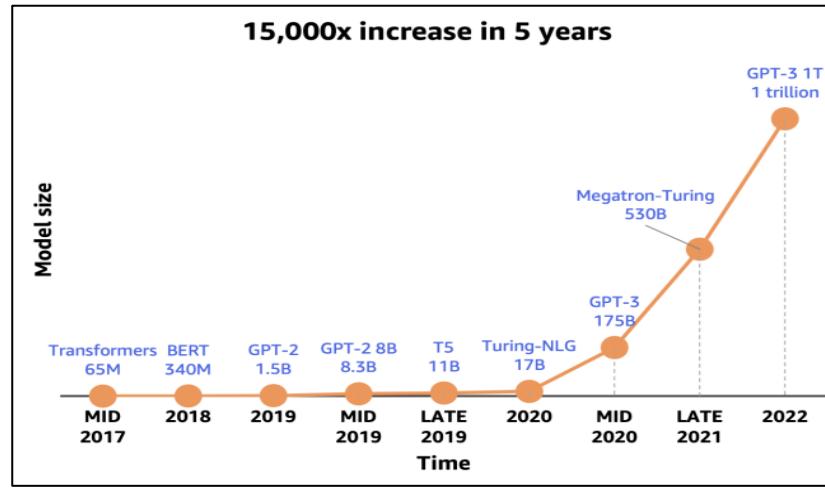
¹ETH Zurich, ²Beihang University, ³ByteDance Inc.

Dr. Haotong Qin
Postdoctoral Researcher
Center of Project-Based Learning, D-ITET



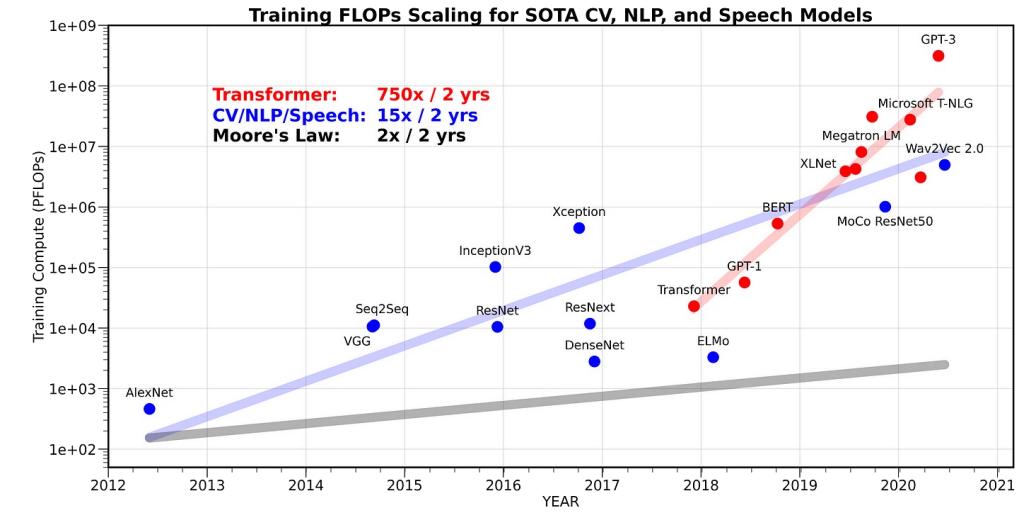
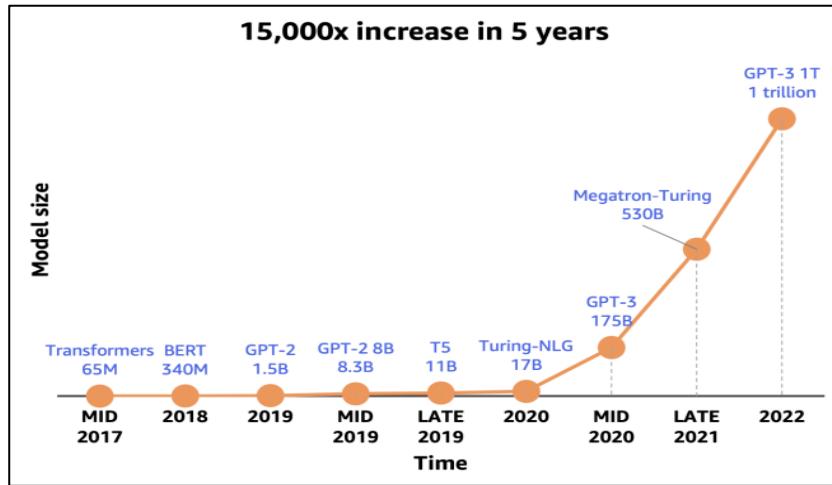
Background: LLMs Suffer Significant Cost

Massive Training Data
Huge Model Scale

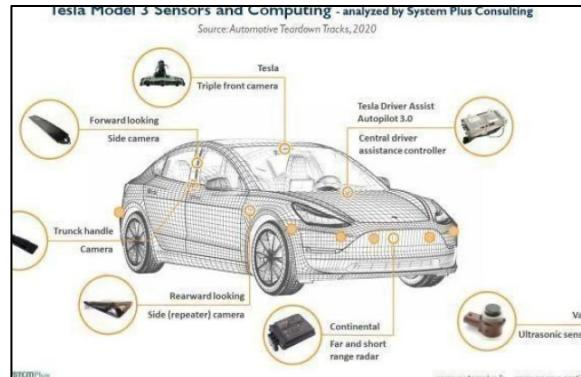
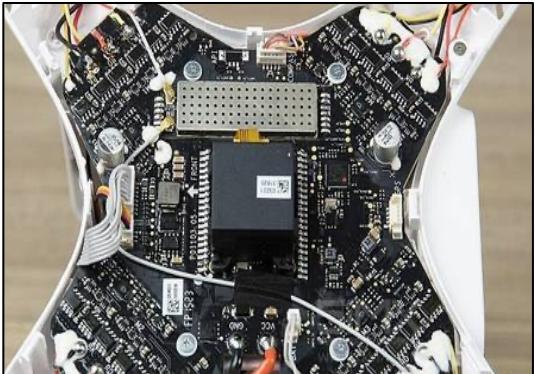


Background: LLMs Suffer Significant Cost

Massive Training Data
Huge Model Scale

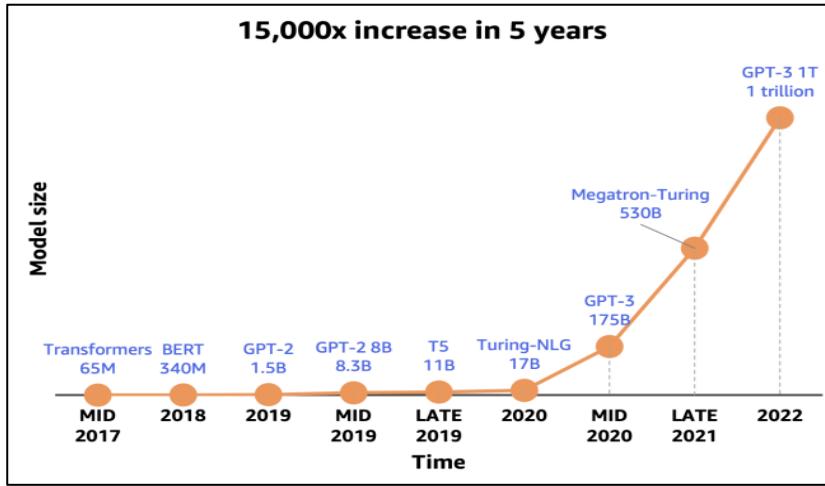


Diverse Scenarios
Constrained Resources

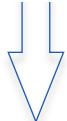


Background: LLMs Suffer Significant Cost

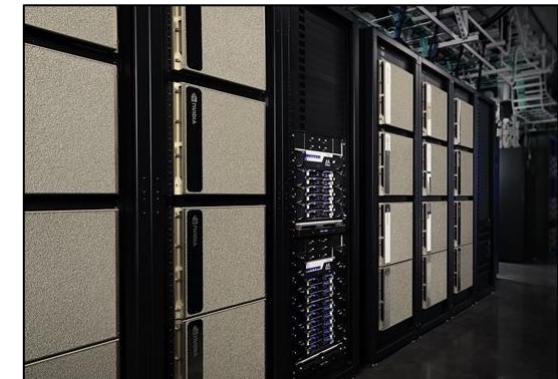
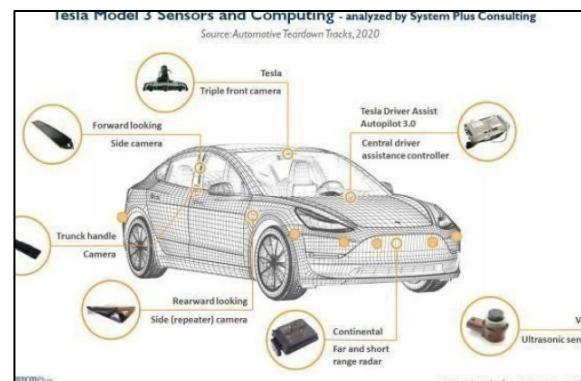
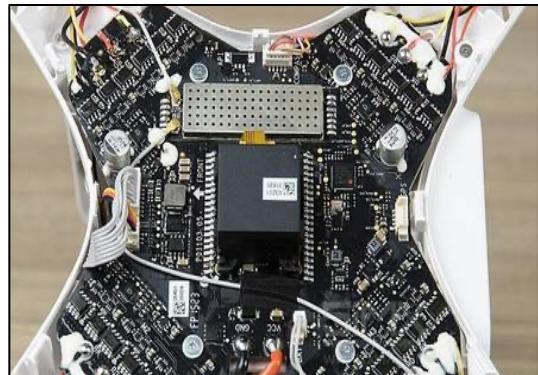
Massive Training Data
Huge Model Scale



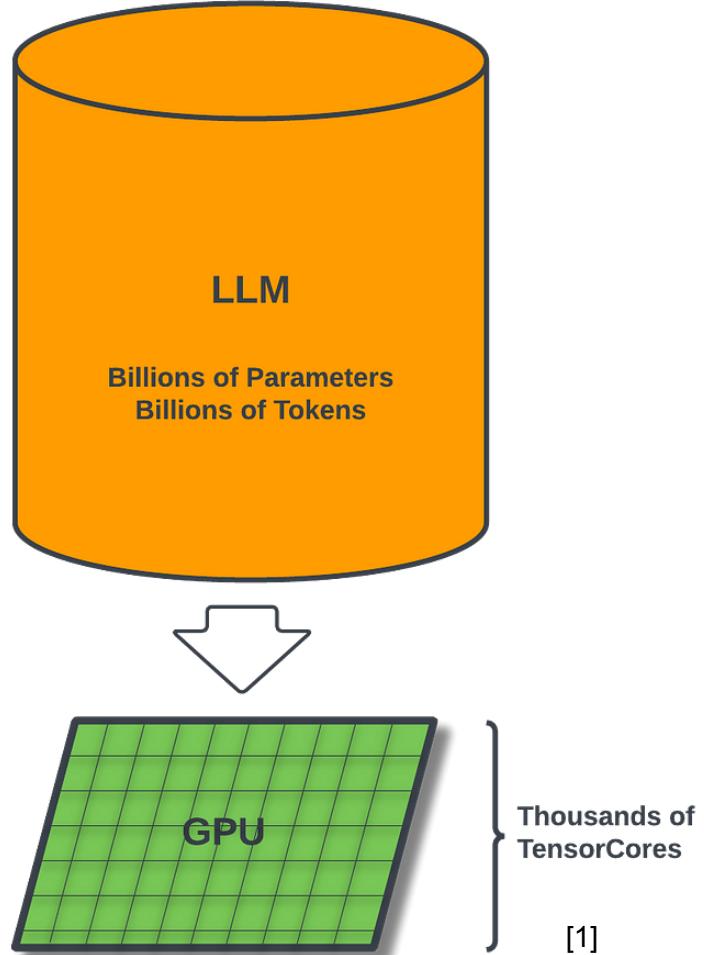
**Significant Conflicts
Exists!**



Diverse Scenarios
Constrained Resources



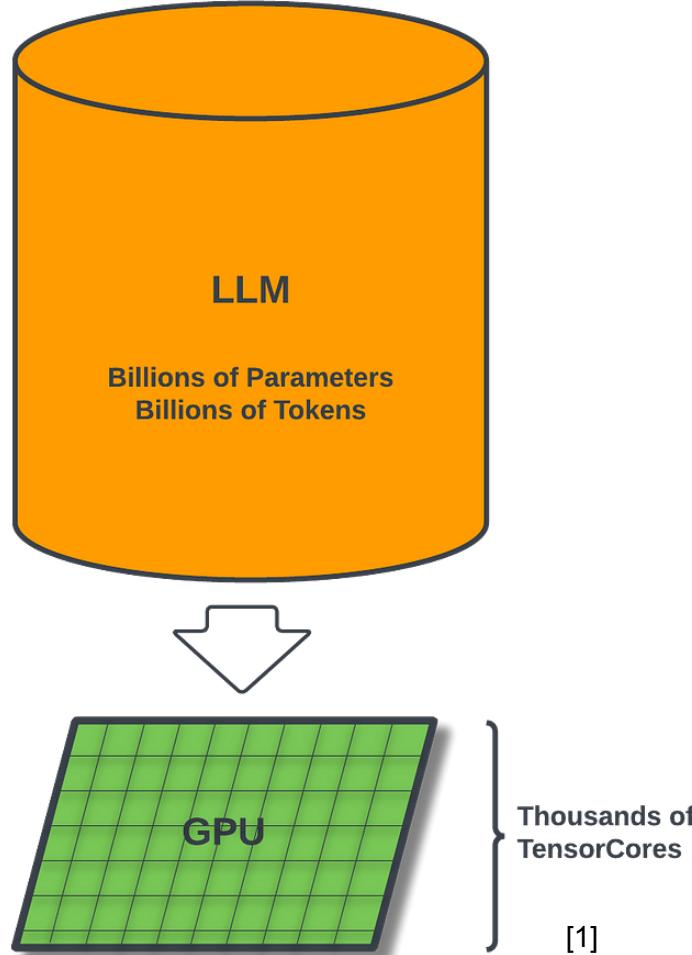
Background: LLMs Suffer Significant Cost



[1] Performance bottlenecks in deploying LLMs—a primer for ML researchers. 2023

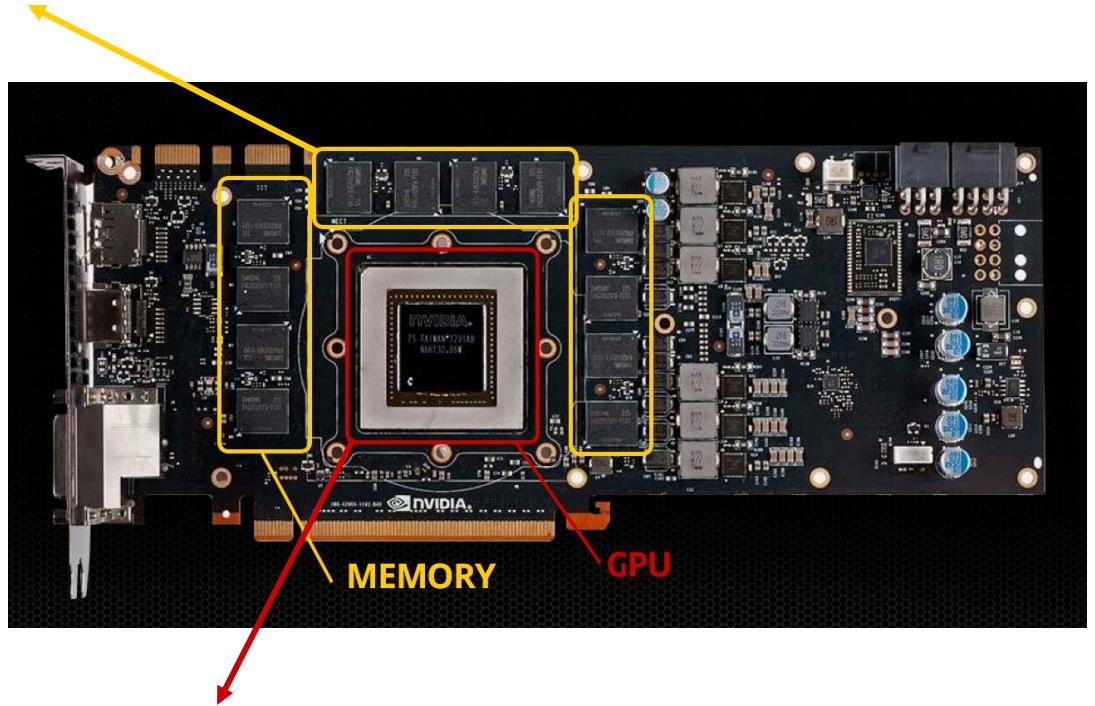
[2] LLaMA: Open and Efficient Foundation Language Models. 2023

Background: LLMs Suffer Significant Cost



Memory Cost: **196.52 GB** (FP16/BF16, w/o datasets) vs. 141GB (Nvidia H200)

LLaMA-13B
hidden_size: 5120
model_max_length: 4096
intermediate_size: 13696
num_attention_heads: 40
num_hidden_layers: 40
batch_size: 1



Training Time: **135,168 GPU Hours** (NVIDIA A100-80G) [2]

[1] Performance bottlenecks in deploying LLMs—a primer for ML researchers. 2023

[2] LLaMA: Open and Efficient Foundation Language Models. 2023

Preliminaries: LoRA-Finetuning Quantization

LLMs: NormalFloat (NF) Quantization

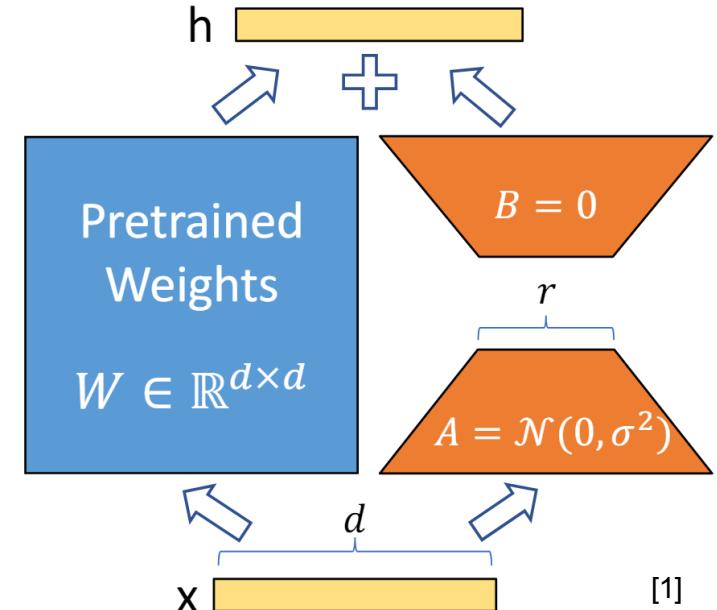
$$\hat{\mathbf{w}}^{\text{NF}k} = \text{NF}k \left(\frac{\mathbf{w}}{s} \right) = \text{NF}k \left(\frac{\mathbf{w}}{\text{absmax}(\mathbf{w})} \right), \quad (1)$$

$$q_i = \frac{1}{2} \left(Q \left(\frac{i}{2^k + 1} \right) + Q \left(\frac{i + 1}{2^k + 1} \right) \right), \quad (2)$$

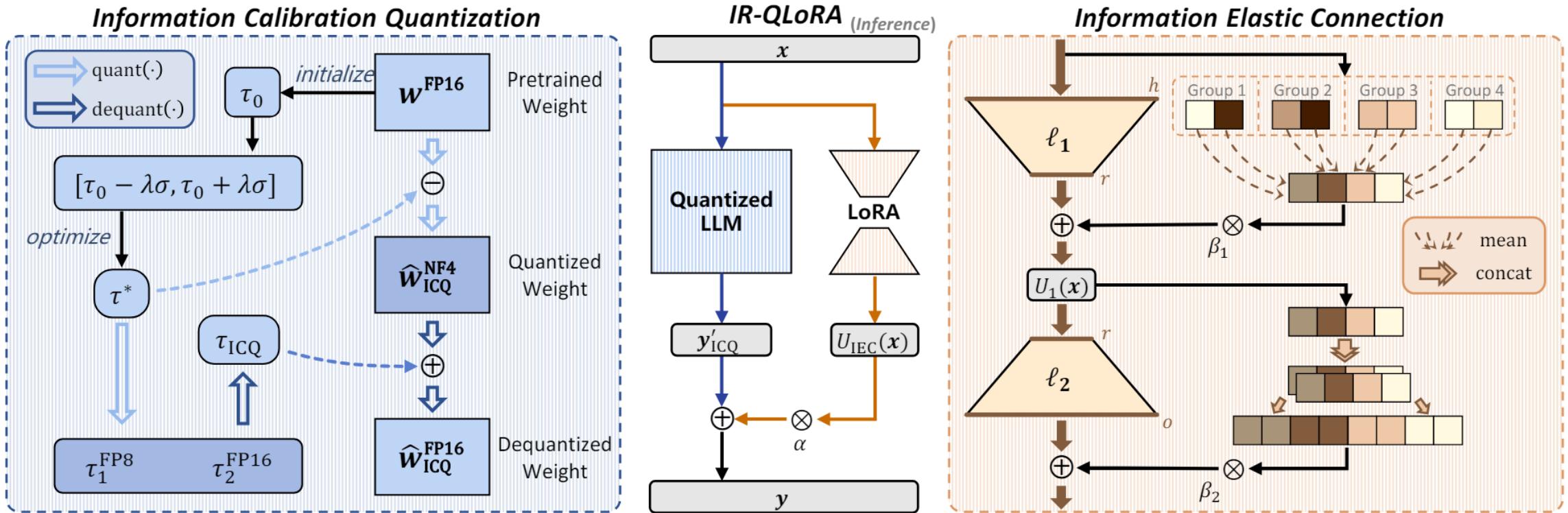
$$\mathbf{y}' = \mathbf{x} \hat{\mathbf{w}}^{\text{FP16}} = \mathbf{x} \left(\hat{\mathbf{w}}^{\text{NF}k} \text{dequant}(s_1^{\text{FP8}}, s_2^{\text{FP16}}) \right), \quad (3)$$

LoRA: Low Rank Adapters

$$\mathbf{y} = \mathbf{y}' + \alpha \mathbf{x} \ell_1 \ell_2, \quad (4)$$



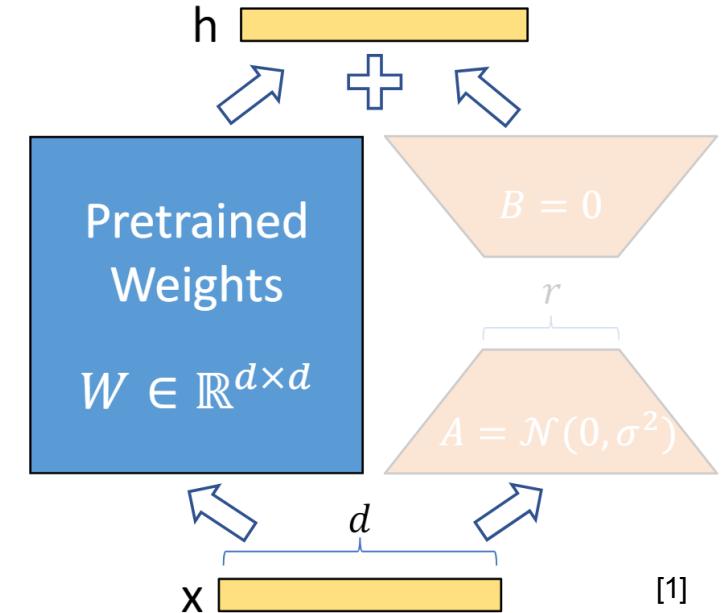
IR-QLoRA: Overview



IR-QLoRA: Information Calibration Quantization

Degeneration of Quantized LLMs

$$\mathcal{I}(\hat{\mathbf{w}}^{\text{FP16}}; \mathbf{w}) = \mathcal{H}(\hat{\mathbf{w}}^{\text{FP16}}) - \underbrace{\mathcal{H}(\hat{\mathbf{w}}^{\text{FP16}} | \mathbf{w})}_{0}, \quad (5)$$

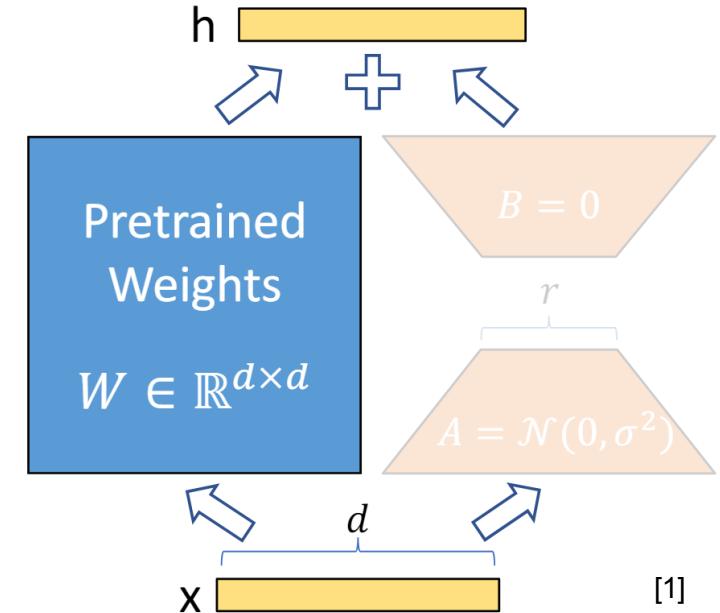


IR-QLoRA: Information Calibration Quantization

Degeneration of Quantized LLMs



$$\mathcal{I}(\hat{\mathbf{w}}^{\text{FP16}}; \mathbf{w}) = \mathcal{H}(\hat{\mathbf{w}}^{\text{FP16}}) - \underbrace{\mathcal{H}(\hat{\mathbf{w}}^{\text{FP16}} | \mathbf{w})}_{0}, \quad (5)$$



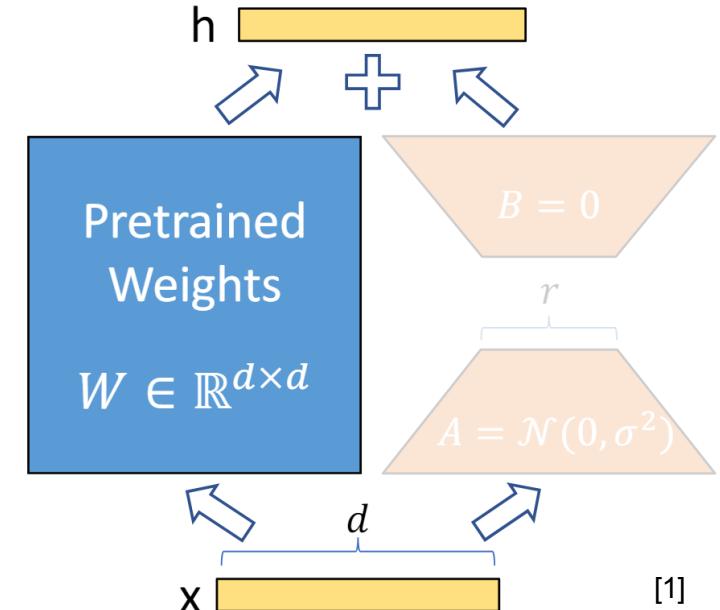
IR-QLoRA: Information Calibration Quantization

Degeneration of Quantized LLMs

$$\mathcal{I}(\hat{\mathbf{w}}^{\text{FP16}}; \mathbf{w}) = \mathcal{H}(\hat{\mathbf{w}}^{\text{FP16}}) - \underbrace{\mathcal{H}(\hat{\mathbf{w}}^{\text{FP16}} | \mathbf{w})}_{0}, \quad (5)$$

$$\underset{s, s_1^{\text{FP8}}, s_2^{\text{FP16}}}{\operatorname{argmax}} \mathcal{H}(\hat{\mathbf{w}}^{\text{FP16}}; s, s_1^{\text{FP8}}, s_2^{\text{FP16}}). \quad (6)$$

$$\mathcal{H}(\hat{\mathbf{w}}^{\text{NF}k}; s) = - \sum_{i=1}^{2^k-1} P(q_i) \log_2 P(q_i), \quad (7)$$



Reduced bit-width of parameters constrains the representation capabilities of LLMs

IR-QLoRA: Information Calibration Quantization

Information Calibration Quantization (ICQ) for Representation Recovery


$$\hat{\mathbf{w}}^{\text{NF}k} = \text{NF}k \left(\frac{\mathbf{w}}{s} \right) = \text{NF}k \left(\frac{\mathbf{w}}{\text{absmax}(\mathbf{w})} \right), \quad (1)$$

$$\hat{\mathbf{w}}^{\text{NF}k} = \text{NF}k \left(\frac{\mathbf{w} - \tau}{s} \right). \quad (8)$$

IR-QLoRA: Information Calibration Quantization

Information Calibration Quantization (ICQ) for Representation Recovery


$$\hat{\mathbf{w}}^{\text{NF}k} = \text{NF}k \left(\frac{\mathbf{w}}{s} \right) = \text{NF}k \left(\frac{\mathbf{w}}{\text{absmax}(\mathbf{w})} \right), \quad (1)$$


$$\hat{\mathbf{w}}^{\text{NF}k} = \text{NF}k \left(\frac{\mathbf{w} - \tau}{s} \right). \quad (8)$$


$$\underset{s}{\operatorname{argmax}} \mathcal{H}(\hat{\mathbf{w}}^{\text{NF}k}; s)$$


$$\underset{\tau, s}{\operatorname{argmax}} \mathcal{H} \left(\hat{\mathbf{w}}^{\text{NF}k}; \tau, s \right)$$

IR-QLoRA: Information Calibration Quantization

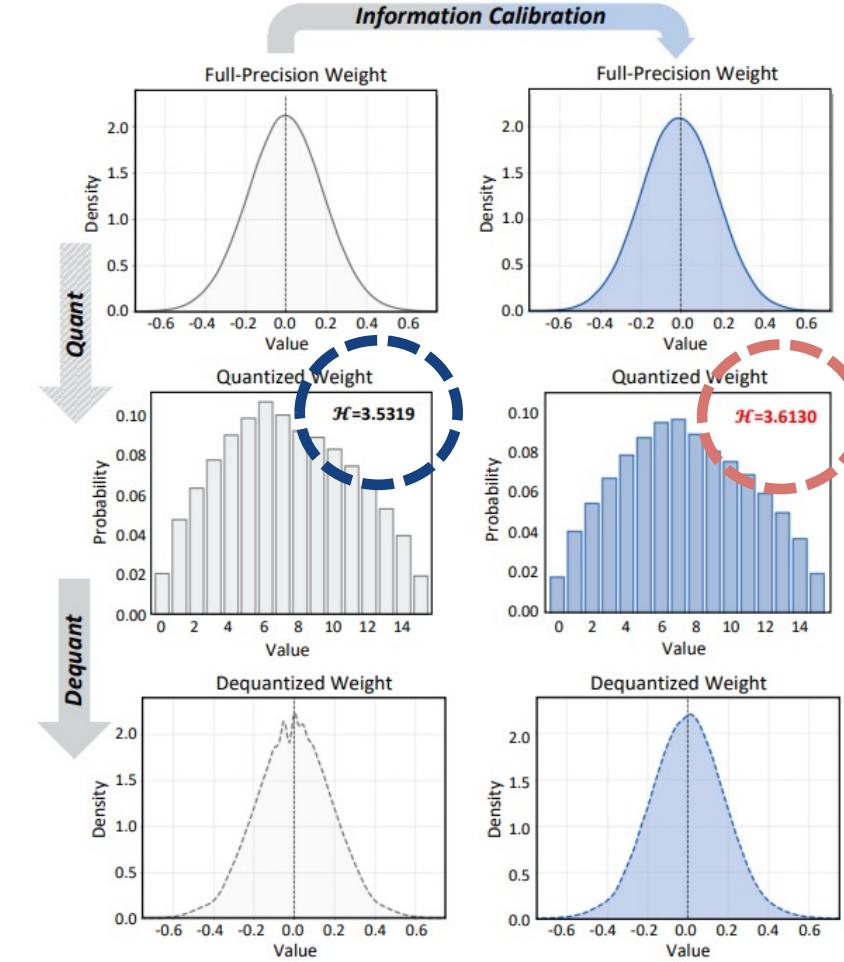
Information Calibration Quantization (ICQ) for Representation Recovery

$$\hat{w}^{\text{NF}k} = \text{NF}k \left(\frac{w}{s} \right) = \text{NF}k \left(\frac{w}{\text{absmax}(w)} \right), \quad (1)$$

$$\hat{w}^{\text{NF}k} = \text{NF}k \left(\frac{w - \tau}{s} \right). \quad (8)$$

$$\underset{s}{\operatorname{argmax}} \mathcal{H}(\hat{w}^{\text{NF}k}; s)$$

$$\underset{\tau, s}{\operatorname{argmax}} \mathcal{H} \left(\hat{w}^{\text{NF}k}; \tau, s \right)$$



IR-QLoRA: Information Calibration Quantization

Information Calibration Quantization (ICQ) for Representation Recovery

1. Solve the τ

- Initialization

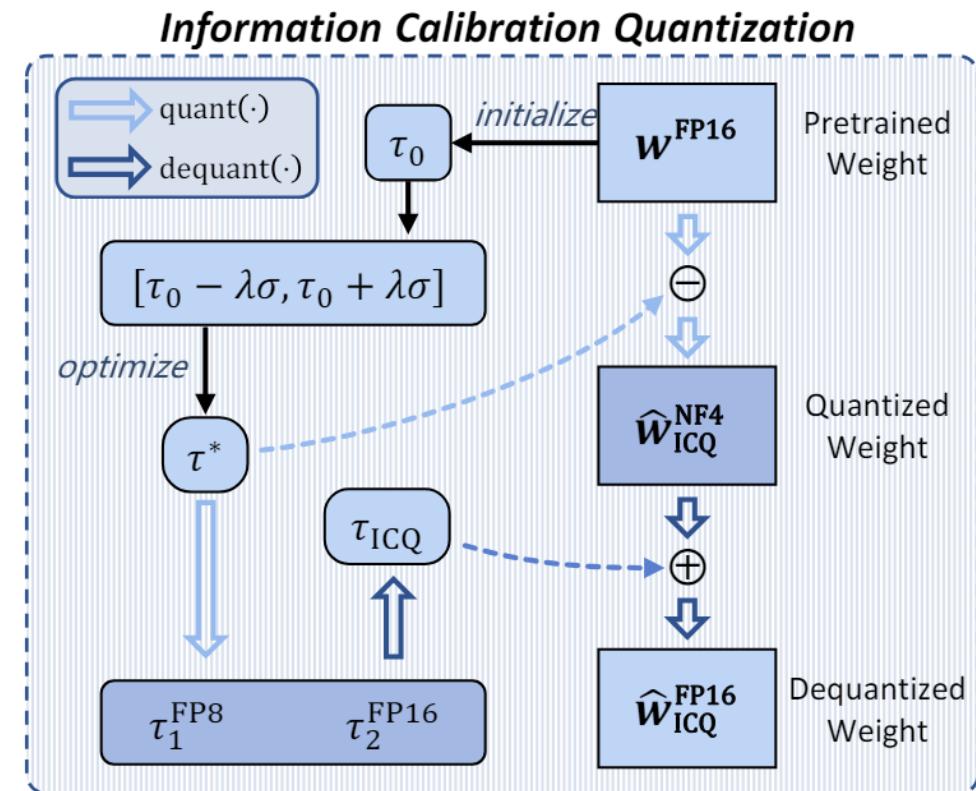
$$\tau_0 = \text{quantile}_{\frac{1}{2}}(\mathbf{w})$$

- Optimization

$$\tau^* \leftarrow \text{search in } [\tau_0 - \lambda\sigma, \tau_0 + \lambda\sigma]$$

2. Solve the s

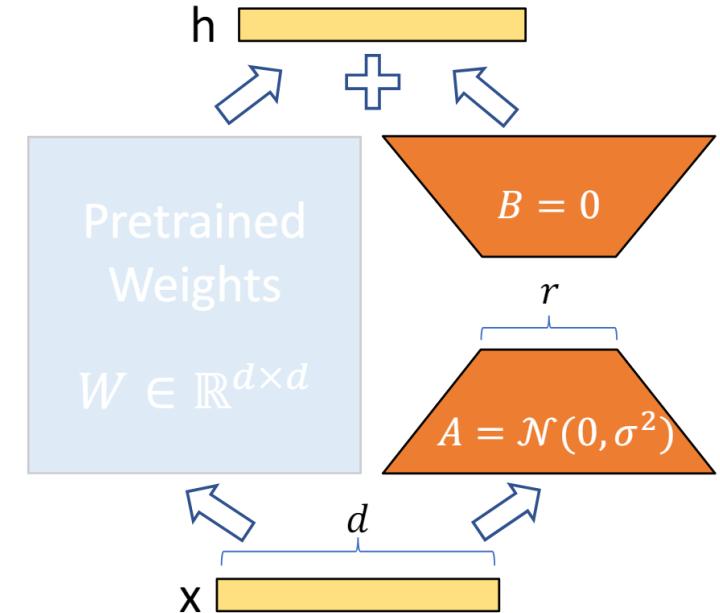
$$s \leftarrow \text{absmax}(\mathbf{w} - \tau^*)$$



IR-QLoRA: Information Elastic Connection

Limitation of Finetunable LoRA

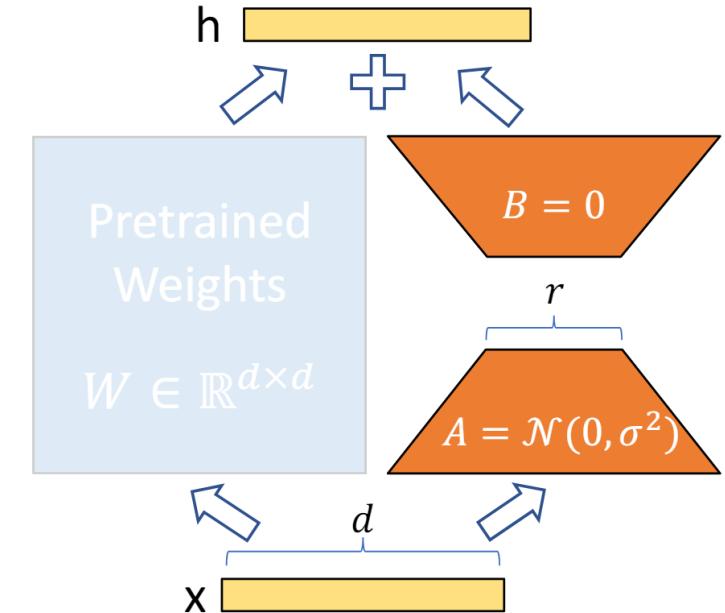
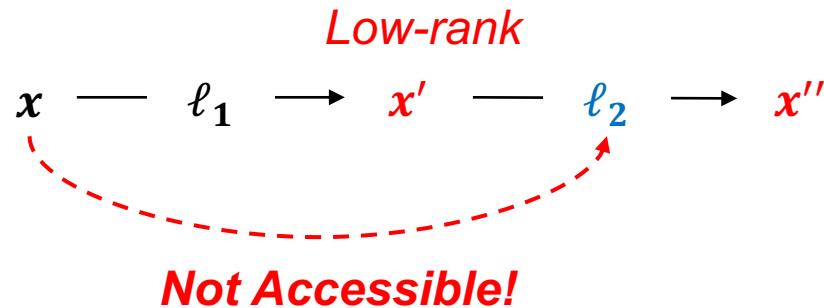
$$\mathbf{y} = \mathbf{y}' + \alpha \mathbf{x} \ell_1 \ell_2, \quad (4)$$



IR-QLoRA: Information Elastic Connection

Limitation of Finetunable LoRA

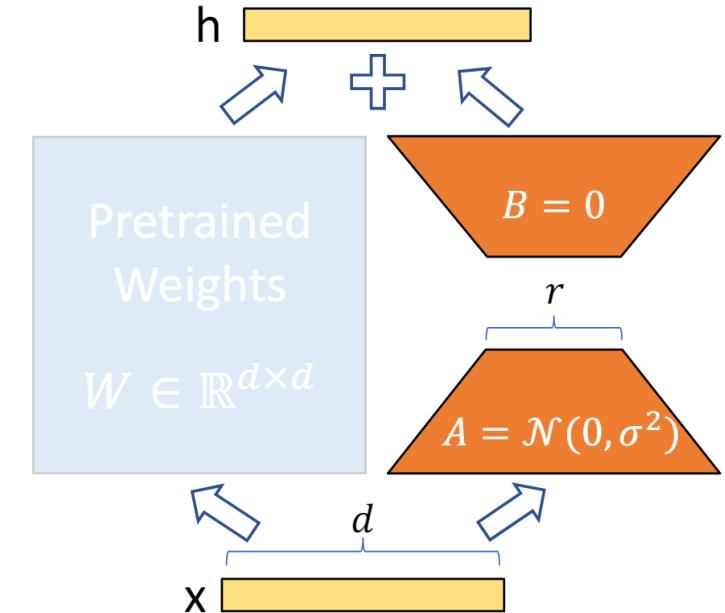
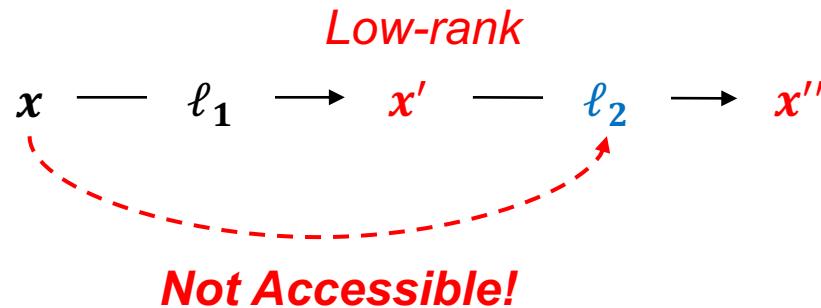
$$y = y' + \alpha x \ell_1 \ell_2, \quad (4)$$



IR-QLoRA: Information Elastic Connection

Limitation of Finetunable LoRA

$$y = y' + \alpha x \ell_1 \ell_2, \quad (4)$$



Low-rank transformations distort the information of representation extracted by LLMs

IR-QLoRA: Information Elastic Connection

Information Elastic Connection (IEC) for Information Enhancement

$$U_1(\boldsymbol{x}) = \boldsymbol{x} \boldsymbol{\ell}_1 + \beta_1 \prod_{r=1}^{\frac{h}{\gcd(h,r)}} \left(\frac{\gcd(h,r)}{h} \sum_{i=1}^{\gcd(h,r)} \boldsymbol{x}^{\left[(i-1)\frac{h}{\gcd(h,r)} : i\frac{h}{\gcd(h,r)} - 1 \right]} \right)$$

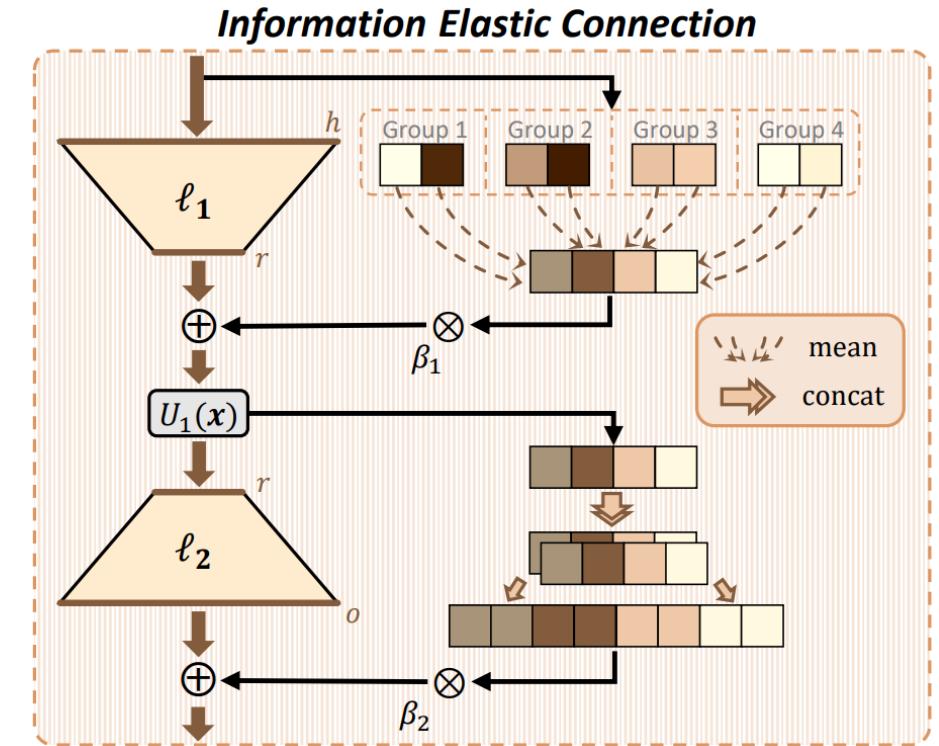
$$U_2(\boldsymbol{x}') = \boldsymbol{x}' \boldsymbol{\ell}_2 + \beta_2 \prod_{r=1}^{\frac{o}{\gcd(o,r)}} \left(\frac{\gcd(o,r)}{r} \sum_{i=1}^{\gcd(o,r)} \boldsymbol{x}'^{\left[(i-1)\frac{r}{\gcd(o,r)} : i\frac{r}{\gcd(o,r)} - 1 \right]} \right)$$

IR-QLoRA: Information Elastic Connection

Information Elastic Connection (IEC) for Information Enhancement

$$U_1(\mathbf{x}) = \mathbf{x} \ell_1 + \beta_1 \prod_{i=1}^{\frac{r}{\gcd(h,r)}} \left(\frac{\gcd(h,r)}{h} \sum_{i=1}^{\gcd(h,r)} \mathbf{x}^{[(i-1)\frac{h}{\gcd(h,r)} : i\frac{h}{\gcd(h,r)} - 1]} \right)$$

$$U_2(\mathbf{x}') = \mathbf{x}' \ell_2 + \beta_2 \prod_{i=1}^{\frac{o}{\gcd(o,r)}} \left(\frac{\gcd(o,r)}{r} \sum_{i=1}^{\gcd(o,r)} \mathbf{x}'^{[(i-1)\frac{r}{\gcd(o,r)} : i\frac{r}{\gcd(o,r)} - 1]} \right)$$



IR-QLoRA: Information Elastic Connection

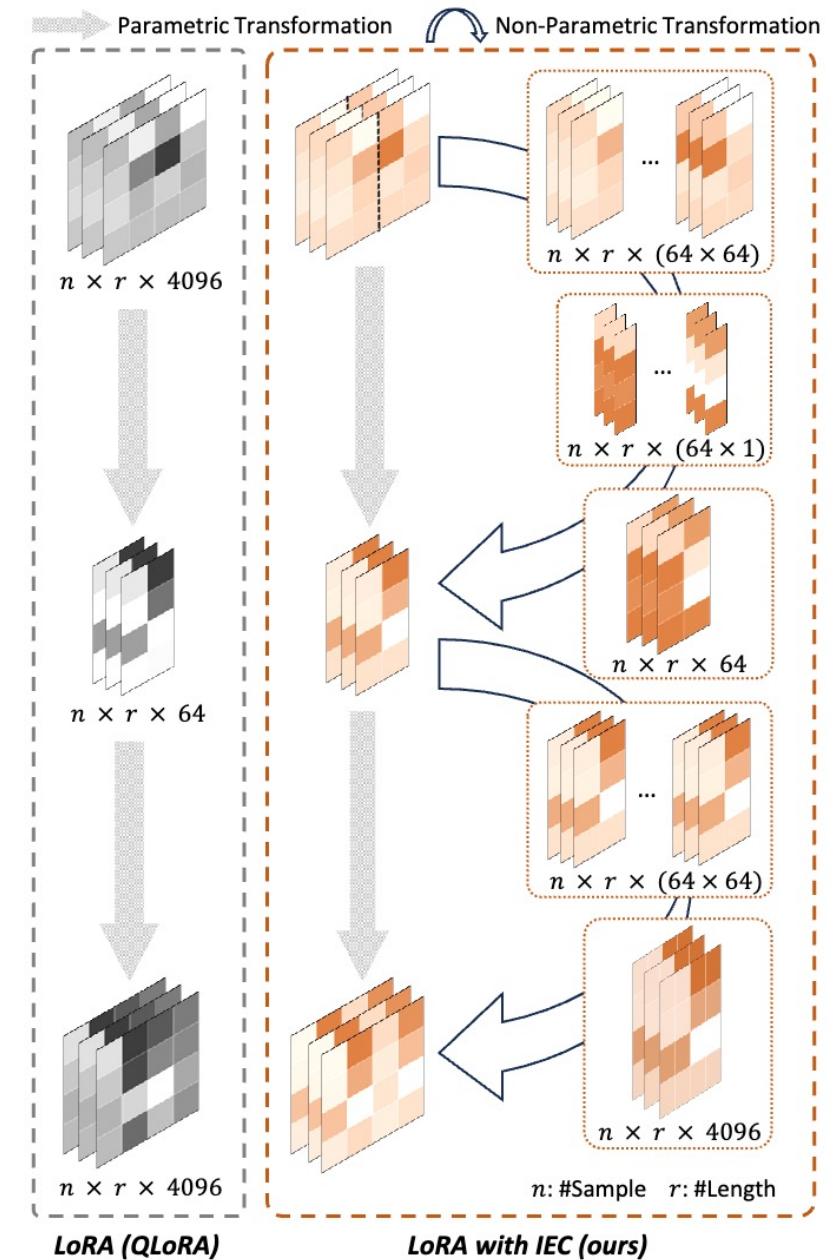
Information Elastic Connection (IEC) for Information Enhancement

$$U_1(\mathbf{x}) = \mathbf{x} \ell_1 + \beta_1 \prod_{i=1}^{\frac{r}{\gcd(h,r)}} \left(\frac{\gcd(h,r)}{h} \sum_{i=1}^{\gcd(h,r)} \mathbf{x}^{[(i-1)\frac{h}{\gcd(h,r)} : i\frac{h}{\gcd(h,r)} - 1]} \right)$$

$$U_2(\mathbf{x}') = \mathbf{x}' \ell_2 + \beta_2 \prod_{i=1}^{\frac{o}{\gcd(o,r)}} \left(\frac{\gcd(o,r)}{r} \sum_{i=1}^{\gcd(o,r)} \mathbf{x}'^{[(i-1)\frac{r}{\gcd(o,r)} : i\frac{r}{\gcd(o,r)} - 1]} \right)$$

$$U_1(\mathbf{x}) = \mathbf{x} \ell_1 + \beta_1 \frac{r}{h} \sum_{i=1}^r \mathbf{x}^{[(i-1)\frac{h}{r} : i\frac{h}{r} - 1]}$$

$$U_2(\mathbf{x}') = \mathbf{x}' \ell_2 + \beta_2 \prod_{i=1}^{\frac{o}{r}} \mathbf{x}'$$



IR-QLoRA: Experiments

Quantitative Results: Ablation Study

Table 4: Accuracy (%) ablation on MMLU

Method	#Bit	MMLU				
		Hums.	STEM	Social	Other	Avg.
LLaMA-7B	16	33.3	29.8	37.8	38.0	34.6
Vanilla	4	36.1	31.9	42.0	44.5	38.4
ICQ	4	37.9	33.6	43.9	46.7	40.3
IEC (U_1)	4	37.9	31.9	43.4	44.8	39.4
IEC (U_2)	4	38.0	32.3	43.6	45.1	39.7
IEC	4	38.3	33.0	44.5	45.7	40.2
IR-QLoRA (ours)	4	38.6	34.6	45.2	45.5	40.8

Accuracy

Table 6: Efficiency ablation on the different sizes of LLaMA

Method	#Bit	#Params(GB)		Time(h)	
		7B	13B	7B	13B
LLaMA	16	12.55	24.24	-	-
Vanilla	4	2.34	4.13	15.33	26.18
ICQ	4	2.39	4.22	15.40	26.26
IEC	4	2.34	4.13	15.33	26.18
IR-QLoRA	4	2.39	4.22	15.40	26.26

Efficiency

IR-QLoRA: Experiments

Quantitative Results: Main Results

Method	#Bit	MMLU				
		Hums.	STEM	Social	Other	Avg.
LLaMA-7B	16	33.3	29.8	37.8	38.0	34.6
PEQA	4	34.9	28.9	37.5	40.1	34.8
NormalFloat	4	33.1	30.6	38.8	38.8	35.1
QLoRA w/ GPTQ	4	33.8	31.3	37.4	42.2	36.0
QLoRA	4	36.1	31.9	42.0	44.5	38.4
QA-LoRA	4	36.6	32.4	44.8	44.9	39.4
IR-QLoRA (ours)	4	38.6	34.6	45.2	45.5	40.8
LLaMA-13B	16	40.6	36.7	48.9	48.0	43.3
NormalFloat	4	43.0	34.5	51.8	51.4	45.0
PEQA	4	43.0	37.7	53.6	49.0	45.0
QLoRA	4	45.4	37.4	55.7	54.3	48.0
QLoRA w/ GPTQ	4	48.4	38.3	54.9	55.2	49.2
QA-LoRA	4	48.4	38.3	54.9	55.2	49.2
IR-QLoRA (ours)	4	47.2	39.0	56.5	55.0	49.3
LLaMA-30B	16	56.2	45.9	67.1	63.9	58.2
NormalFloat	4	55.3	44.7	66.2	63.3	57.3
QLoRA	4	55.4	46.0	66.4	63.6	57.7
QLoRA w/ GPTQ	4	55.8	46.4	67.0	64.0	58.1
QA-LoRA	4	55.8	46.4	67.0	64.0	58.1
IR-QLoRA (ours)	4	56.7	46.7	66.5	63.2	58.2
LLaMA-65B	16	61.4	51.9	73.6	67.6	63.4
QA-LoRA	4	60.8	50.5	72.5	66.7	62.5
NormalFloat	4	60.7	52.3	72.6	67.3	63.0
QLoRA w/ GPTQ	4	60.4	52.5	73.0	67.2	63.0
QLoRA	4	60.3	52.7	72.9	67.4	63.1
IR-QLoRA (ours)	4	60.1	50.1	74.4	68.7	63.1

Alpaca

Method	#Bit	MMLU				
		Hums.	STEM	Social	Other	Avg.
LLaMA-7B	16	33.3	29.8	37.8	38.0	34.6
NormalFloat	4	33.1	30.6	38.8	38.8	35.1
QLoRA w/ GPTQ	4	33.8	31.3	37.4	42.2	36.0
QLoRA	4	41.4	35.0	49.8	52.0	44.3
QA-LoRA	4	43.9	38.0	54.3	53.0	47.0
IR-QLoRA (ours)	4	44.2	39.3	54.5	52.9	47.4
LLaMA-13B	16	40.6	36.7	48.9	48.0	43.3
NormalFloat	4	43.0	34.5	51.8	51.4	45.0
QLoRA w/ GPTQ	4	48.4	38.3	54.9	55.2	49.2
QLoRA	4	49.9	40.1	60.2	57.9	51.9
QA-LoRA	4	50.0	41.5	60.5	58.4	52.4
IR-QLoRA (ours)	4	49.2	41.2	62.1	59.2	52.6
LLaMA-30B	16	56.2	45.9	67.1	63.9	58.2
NormalFloat	4	55.3	44.7	66.2	63.3	57.3
QLoRA w/ GPTQ	4	55.8	46.4	67.0	64.0	58.1
QLoRA	4	57.2	48.6	69.8	65.2	60.0
QA-LoRA	4	57.9	48.8	71.0	65.5	60.6
IR-QLoRA (ours)	4	58.1	49.4	70.7	65.8	60.8
LLaMA-65B	16	61.4	51.9	73.6	67.6	63.4
NormalFloat	4	60.7	52.3	72.6	67.3	63.0
QLoRA w/ GPTQ	4	60.4	52.5	73.0	67.2	63.0
QLoRA	4	59.8	52.9	75.0	69.6	63.9
QA-LoRA	4	57.6	51.1	73.9	67.4	62.1
IR-QLoRA (ours)	4	61.6	52.0	75.6	68.9	64.3

Flanv2

IR-QLoRA: Experiments

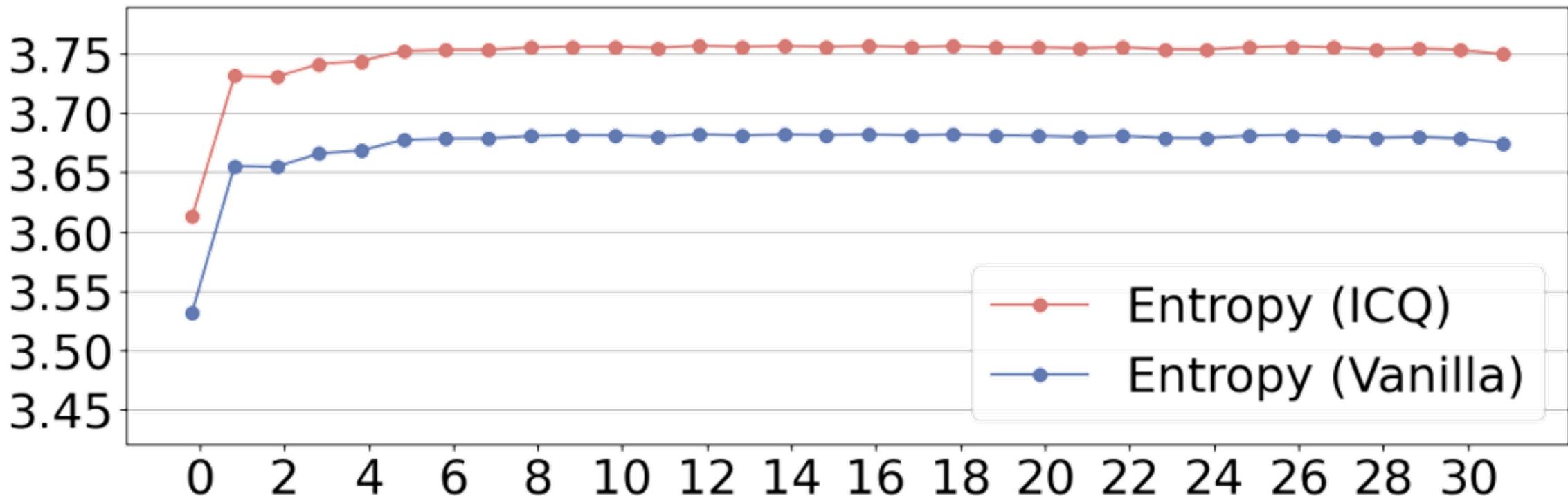
Quantitative Results: More Evaluation Benchmark

Table 6: Accuracy (%) comparison on the Commonsense QA datasets

Method	#Bit	CommonsenseQA							Avg.
		HellaSwag	PIQA	WinoGrande	ARC-e	ARC-c	BoolQ	OBQA	
LLaMA-7B	16	56.3	78.2	67.1	67.3	38.2	72.9	28.4	58.3
NormalFloat	4	56.7	78.7	70.6	75.7	41.6	74.7	33.2	61.6
QLoRA w/ GPTQ	4	57.4	77.6	66.2	70.9	41.8	73.5	31.2	59.8
QLoRA	4	61.8	78.1	68.4	75.8	43.6	73.7	32.8	62.0
QA-LoRA	4	58.6	78.0	66.9	71.2	43.9	79.9	34.0	61.8
IR-QLoRA (ours)	4	54.7	78.8	72.6	76.6	45.1	80.6	37.2	63.7

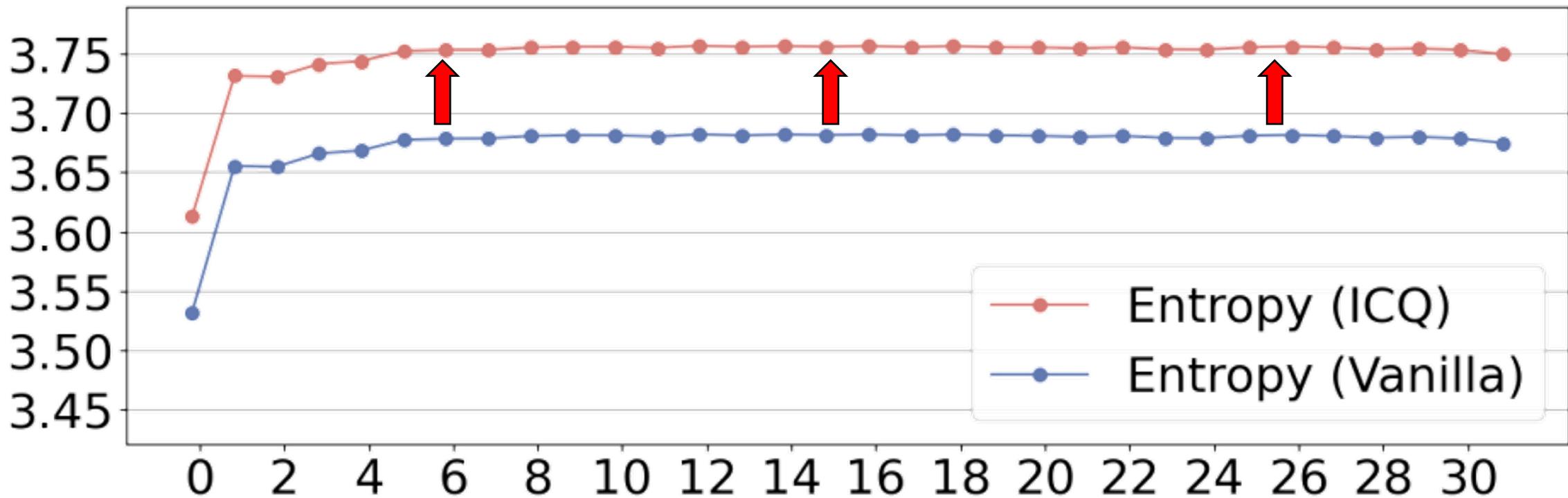
IR-QLoRA: Experiments

Analysis: Information Entropy



IR-QLoRA: Experiments

Analysis: Information Entropy



IR-QLoRA: Experiments

Analysis: Real Cases

[1] **Input:** Alice has 15 apples. She buys 10 more from the market and then gives 7 apples to her friend. How many apples does Alice have now?

Answer: Alice have 18 apples now.

Model	Output
QLoRA-7B	22
IR-QLoRA-7B (ours)	Alice now has 18 apples.
QLoRA-13B	Alice had 15 apples. Alice buys 10 more apples. Alice has $15 + 10 = 25$ apples. Alice gives 7 apples to her friend. Alice has $25 - 7 = 18$ apples.
IR-QLoRA-13B (ours)	Alice started with 15 apples and then bought 10 apples from the market, so her total apples inventory is now 25 apples ($15 + 10 = 25$). Then she gave 7 apples to her friend, so her total apples inventory is now 18 apples ($25 - 7 = 18$). Alice now has 18 apples.

IR-QLoRA: Experiments

Analysis: Real Cases

[3] Input: Imagine you are a historian living in ancient Rome, describe a day in your life.	
Model	Output
QLoRA-7B	I wake up in my bed and eat breakfast. I put on my clothes and walk to school. I take my lunch and sit in my desk. We have a history lesson and a math lesson. I walk home and play with my friends.
IR-QLoRA-7B (ours)	My day begins with waking up early to tend to my animals. I then head to the forum to listen to the debates between my fellow Romans. Afterwards, I might head to the library to research some more, before I finally return home to my family in the evening.
QLoRA-13B	You wake up in your apartment in the center of the city. You roll out of your bed and stretch your arms and legs. You yawn and open your eyes. You sit up and put your feet on the floor. You get up and walk to the bathroom. You wash your face, your hands and your body. You then brush your teeth. You return to your bedroom and get dressed in your clothes. You then put on your shoes and walk out of your apartment.
IR-QLoRA-13B (ours)	As a historian living in ancient Rome, I would wake up in the morning when my slave would bring me coffee. Then I would go to the library and read ancient texts to see how Rome developed over time. Later, I would visit a slave market where I would buy my slaves. Then, I went to bed.



Thank You! (Q&A)

Dr. Haotong Qin
Postdoctoral Researcher
Center of Project-Based Learning, D-ITET

