



清华大学电子工程系

Department of Electronic Engineering, Tsinghua University

INFINIGENCE
无问芯穹



Evaluating Quantized Large Language Models

Shiyao Li^{1,2}, Xuefei Ning¹, Luning Wang¹, Tengxuan Liu¹, Xiangsheng Shi¹,
Shengen Yan², Guohao Dai^{2,3}, Huazhong Yang¹ and Yu Wang¹

¹Tsinghua University

²Infinigence-AI

³Shanghai Jiaotong University





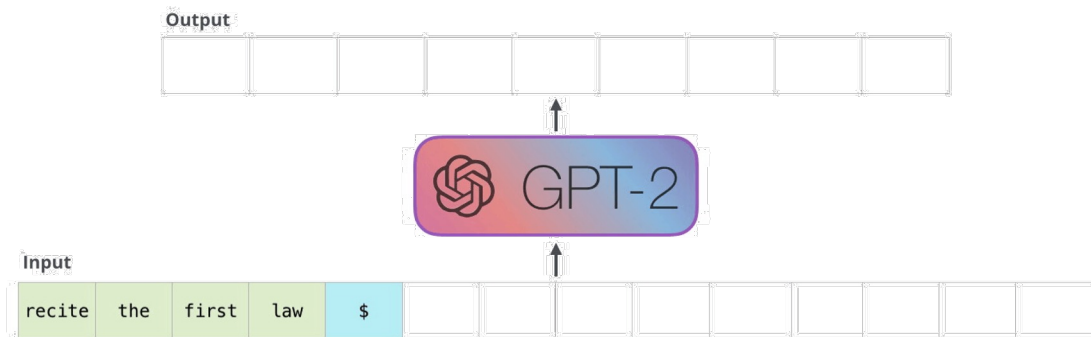
Contents

- 1 Background
- 2 Evaluation Results
- 3 Summarization

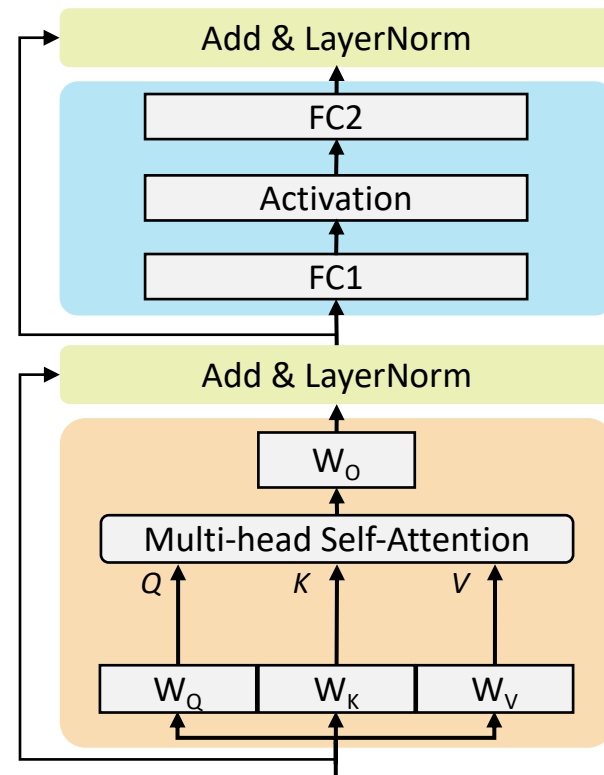
Background



- Most large language models are based on the Transformer architecture^[1].
 - A Transformer block consists of :
 - Multi-Head Attention
 - Feed Forward Network
 - Layer Norm
 - A typical LLM inference process:



Example of Decoder's word-by-word translation



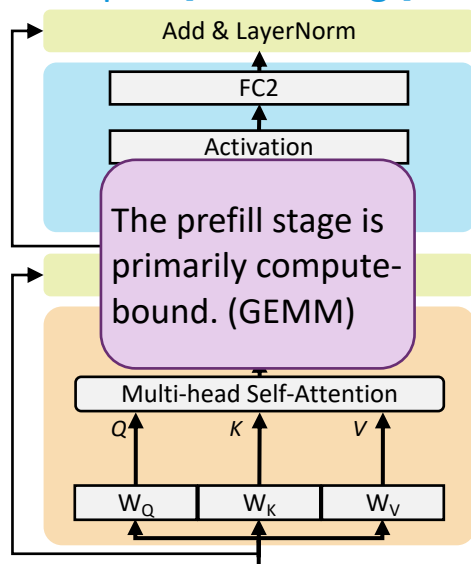
[1] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

LLM Inference



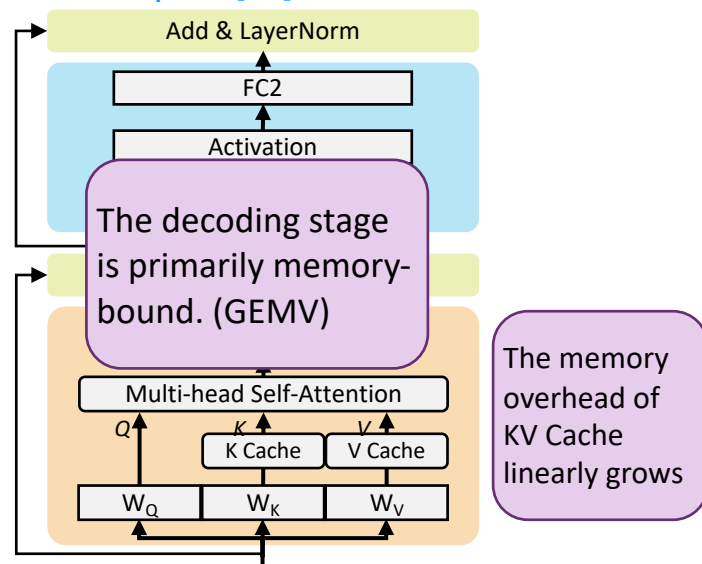
- LLM Inference has two different stages:
- Prefill Stage:** takes a **prompt sequence** to generate the key-value cache (KV Cache)
- Decode Stage:** utilizes and updates the KV cache to **generate tokens one by one**, where the current token depends on all the previously tokens

Output: ['Processing'] ($1 \times \text{dim}$)



Prompt: ['I', 'like', 'natural', 'language'] ($4 \times \text{dim}$)

Output: ['!'] ($1 \times \text{dim}$)



Prompt: ['I', 'like', 'natural', 'language', 'Processing'] ($1 \times \text{dim}$)

Quantization Technique



- Quantization is a promising technique to address the aforementioned efficiency issues.
 - Taking **signed uniform** quantization as an example, quantization parameters include

Scaling Factor, Zero Point, Bitwidth

$$x_{\text{int}} = \text{clip} \left(\left[\frac{x}{s_x} \right] + z, q_{\min}, q_{\max} \right), \quad \text{where } q_{\max} = 2^{b-1} - 1, \quad q_{\min} = -2^{b-1}$$

The diagram shows the quantization formula with red arrows and circles highlighting key parameters: a red arrow points from 'Scaling Factor' to the denominator s_x in the fraction; another red arrow points from 'Zero Point' to the zero point z ; and a third red arrow points from 'Bitwidth' to the exponent b in the range calculation 2^{b-1} .

- The **Weight-Activation Quantization** methods enable the utilization of low-precision Tensor Cores to mitigate the compute-bounded GEMM operators in the prefill stage.
- The **Weight-only Quantization** methods prove effective to accelerate the memory-bounded GEMV operators in the decoding stage.
- The **KV Cache Quantization** methods are necessary to alleviate the large memory overhead when handling tasks with **long contexts or large batch sizes**.

Motivation



- While the effectiveness of post-training quantization (PTQ) has been claimed in many recent studies, a comprehensive evaluation of the quantized LLMs' performance remains to be undertaken.
- In this paper, we make a comprehensive evaluation of quantized LLMs to explore the following issues:
 - Effect of quantization on **various NLP tasks**
 - Effect of quantization on **various LLMs**
 - Effect of quantizing **different tensor types**
 - Effects of different **quantization methods**, especially AWQ and SmoothQuant.

Benchmarks



- Focus on Five different types of tasks:
 - Basic NLP Tasks
 - Tasks to show the Emergent Abilities
 - Trustworthiness Tasks (Chatbot)
 - Dialogue Tasks (Chatbot)
 - Long-context Tasks (Chatbot)
- Task Form:
 - **Perplexity-based (PPL):** Multiple-choice Tasks, choose the answer with the lowest PPL.
 - **Generation-based:** Generate the final answer in text format.
 - Use human design metric, such as exact match, best subspan[1] ...
 - Use LLM Judger (GPT4).

Section	Knowledge & Ability	Benchmark	Size
Sec. 3	Language modeling	CHID (Zheng et al., 2019)	3960
		Winogrande (Sakaguchi et al., 2021)	1770
	Understanding	RACE (Lai et al., 2017)	4930
		Lambda (Paperno et al., 2016)	5150
	Reasoning	SIQA (Sap et al., 2019)	1950
		PIQA (Bisk et al., 2020)	1876
Sec. 4	In-context Learning	MMLU (Hendrycks et al., 2021b)	14000
		Ceval (Huang et al., 2023)	12300
	Multi-step Reasoning	GSM8K (Cobbe et al., 2021)	1320
		StrategyQA (Geva et al., 2021)	2290
	Instruction Following	Hellaswag (Zellers et al., 2019)	10000
		ARC (Clark et al., 2018)	3550
	Self-Calibration	MMLU (Hendrycks et al., 2021b)	14000
Sec. 5	Ethics	ETHICS (Hendrycks et al., 2021a)	3885
	Hallucination	TruthfulQA (Lin et al., 2021)	817
	Robustness	AdvGLUE (Wang et al., 2021)	738
Sec. 6	Dialogue	MT-bench (Zheng et al., 2023a)	80
Sec. 7	Long Context	Longeval (Li et al., 2023)	3000
		Multi-Doc QA (Liu et al., 2023)	700

[1] Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *arXiv preprint arXiv:2307.03172* (2023).

Model Families



- Base LLMs:
 - OPT, LLaMA2, Falcon, Bloom and Mistral family
 - Evaluate on the basic NLP tasks & In-context learning ability.
- Chatbot LLMs
 - LLaMA2, Falcon, ChatGLM3, Mistral, Gemma, Mamba and StableLM
 - Evaluate on other three emergent abilities & Trustworthiness, dialogue tasks.
- Long-context LLMs
 - Vicuna (16k), LongChat (16k), ChatGLM3 (32k), Mistral (32k), LLaMA2 (4k)
 - Evaluate on Long-context tasks.

Model Family
OPT (125M-66B), LLaMA2 (7B-70B), Falcon (7B-180B), Bloomz (560M-176B), Mistral(7B, 8×7B)
LLaMA2 (7B-70B), Falcon (7B-180B), ChatGLM (6B), Mistral (7B, 8×7B) Gemma (2B, 7B), Mamba (2.8B)
(+ StableLM-3B)
Vicuna (7B, 13B), LongChat (7B, 13B), ChatGLM (6B), Mistral (7B, 8×7B)

Tensor Type



- We focus on quantizing the Weight, Activation, and KV Cache tensors.
 - We apply **per-token** quantization for Activation.
 - We apply **group-wise** quantization for Weight and KV Cache.

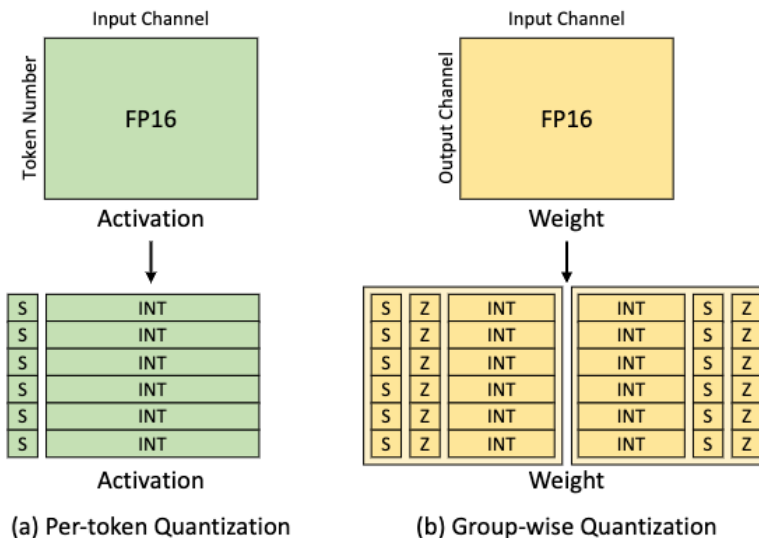


Figure 1. (a) Per-token Quantization for Activation, (b) Group-wise Quantization for Weight and KV Cache.



Contents

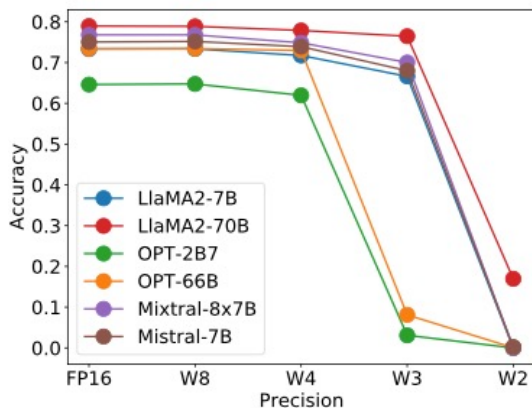
- 1 Background
- 2 Evaluation Results
- 3 Summarization

Basic NLP Abilities

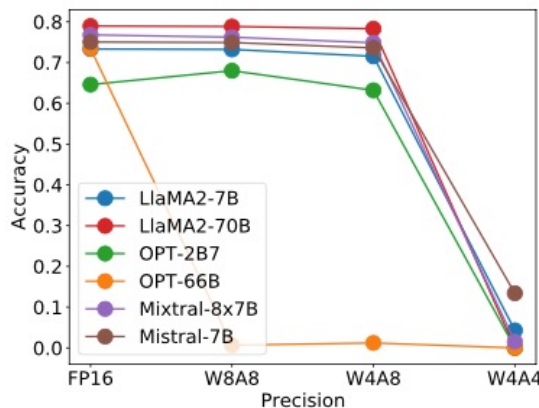


- Effects of Quantization on Tensor Types

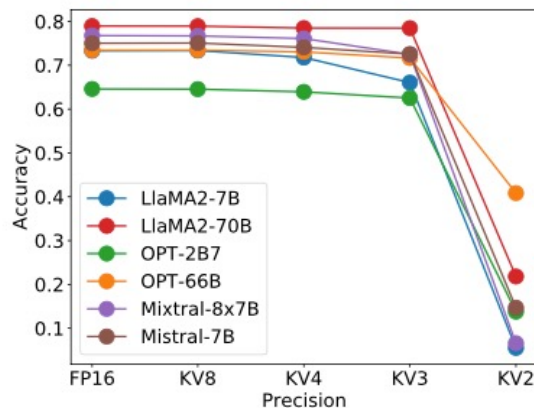
- The larger the model size, the higher the tolerance for Weight and KV-Cache Quantization.
- On the contrary, the larger the model size, the lower the tolerance for Activation Quantization.



(a) Weight-only Quant.



(b) Weight-Activation Quant.



(c) KV Cache Quant.

The effect of quantization on different tensor types on the LAMBADA dataset

- Effects of Quantization on Tensor Types

- The larger the model size, the fewer outliers in the Weight and KV Cache tensors.
- On the contrary, the larger the model size, the more outliers in the Activation tensors.

Table 2. The statistical results of Weights, Activation, and KV Cache on OPT and LLaMA2 families. Specifically, the statistical results of Activation and KV Cache tensors are calculated using the pile-val dataset. We average each statistical metric across all layers.

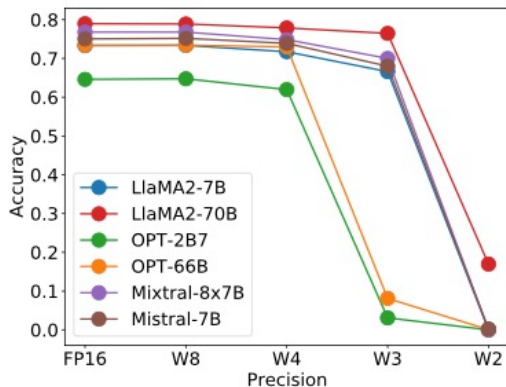
Model	Weight			Activation			KV Cache		
	AbsMax	Std	Kurtosis	AbsMax	Std	Kurtosis	AbsMax	Std	Kurtosis
OPT-1.3B	0.27	0.02	13.16	31.20	0.72	544.97	11.49	1.88	7.53
OPT-6.7B	0.16	0.02	8.74	44.55	0.72	1562.67	10.25	1.71	6.38
OPT-66B	0.11	0.01	5.19	64.36	0.71	4945.32	13.22	2.91	7.40
LLaMA2-7B	0.54	0.02	4.93	27.11	0.30	1167.38	11.99	0.98	14.58
LLaMA2-70B	0.52	0.02	4.83	27.02	0.22	1279.15	11.22	1.07	10.79

Larger Kurtosis means more outliers in a tensor.

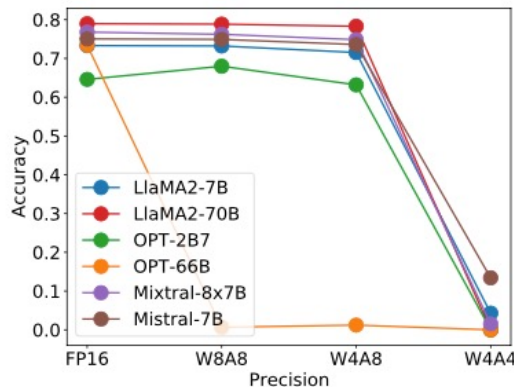
Basic NLP Abilities



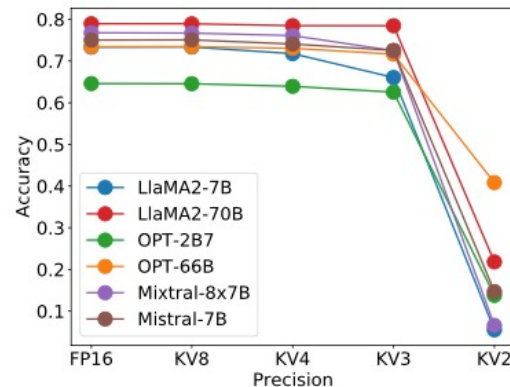
- Effects of Quantization on Different LLMs
 - For the majority of models, the performance order of the quantized models is generally consistent with that of the original models.
 - Leveraging the Mixture-of-Experts (MoE) technique to increase the model size does not necessarily enhance the model's tolerance to quantization.
 - The performance of the FP16 Mixtral-8x7B model is closer to LLaMA2-70B.
 - Its tolerance to quantization is closer to LLaMA2-7B.



(a) Weight-only Quant.



(b) Weight-Activation Quant.



(c) KV Cache Quant.

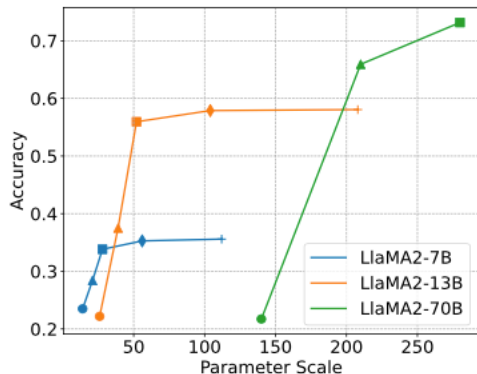
The effect of quantization on different tensor types on the LAMBADA dataset

Basic NLP Abilities

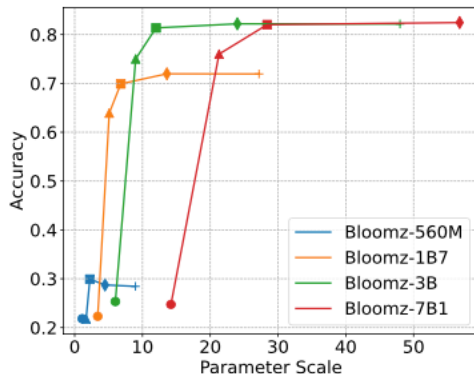


- Effects of Quantization on Different Tasks

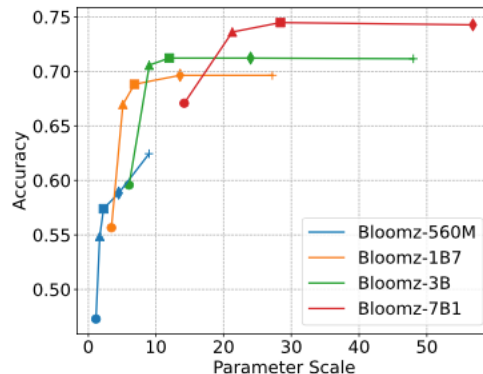
- For most cases, quantizing LLMs to **W4**, **W4A8**, and **KV4** has negligible performance loss.
- With a certain memory budget, employing **larger models quantized to W3** yields superior performance.



(a) Q-LLaMA2 families on the RACE task



(b) Q-Bloomz families on the RACE task



(c) Q-Bloomz families on the PIQA task

Figure 3. Performances of the quantized LLMs with respect to their parameter scales. The parameter scale is calculated by multiplying the parameter size by the quantization bit-width. The markers, '●', '▲', '■', '◆', '+' denote the quantization bit-widths, W2, W3, W4, W8, FP16 respectively. "Q-" is short for "Quantized-".

- Effects of Quantization Methods

- AWQ and SmoothQuant can effectively enhance the performance when the performance loss brought by quantization is moderate.
- Nevertheless, when using extremely low bit-width, AWQ and SmoothQuant cannot restore the fully corrupted performances.

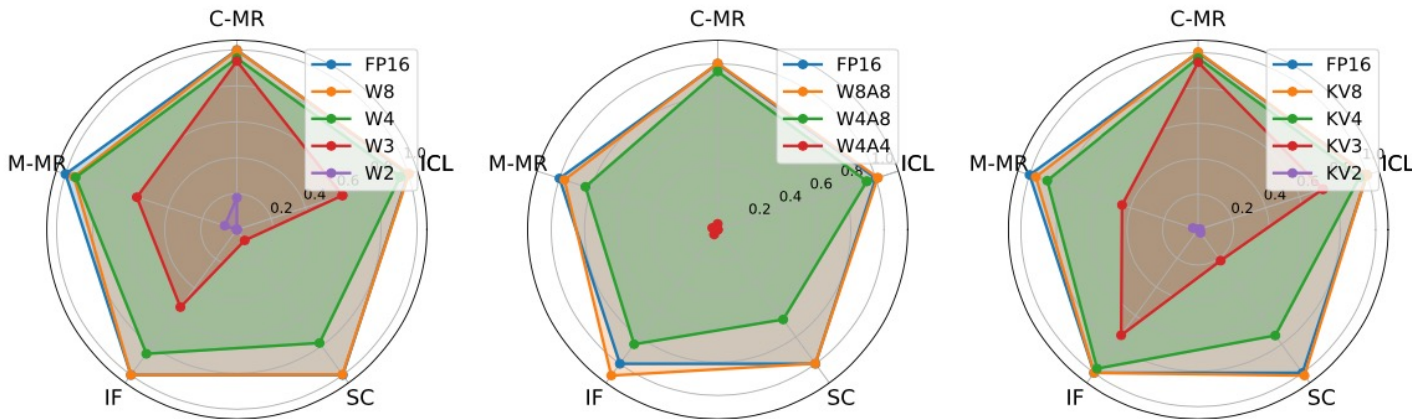
Table 4. The evaluation results of AWQ and SmoothQuant methods on LLaMA2 models on the Lambada dataset. “SQ” is short for “SmoothQuant”.

LLaMA2	FP16	W3		W2		W4A4	
		RTN	AWQ	RTN	AWQ	RTN	SQ
7B	73.32	66.41	69.63	0.00	0.00	4.31	25.56
70B	78.96	76.46	78.73	16.96	0.00	0.04	38.11

Emergent Abilities



- The tolerance to quantization varies across the four abilities, listed in descending order of tolerance: *In-context Learning* ~ *Instruction Following* > *Multi-Step Reasoning* ~ *Self-calibration*.
 - The self-calibration ability is more susceptible to quantization, and can only use **W8**, **W8A8**, and **KV8** quantization.
 - The the harder Mathematical task is much more sensitive to quantization than the easier Commonsense task.



(a) Weight-only Quant. on LLaMA2-7B

(b) Weight-Activation Quant. on LLaMA2-7B

(c) KV Cache Quant. on LLaMA2-7B

Emergent Abilities



- For in-context learning, **more few-shot examples can benefit the performance** of the low-bit Quantization. (From zero-shot to five-shot)
- Besides, **too many few-shot examples show limited benefits**, even bring slight performance degradation. (From five-shot to ten-shot)

Table 12. The evaluation results of different numbers of few-shot examples on LLaMA2 models on the MMLU dataset.

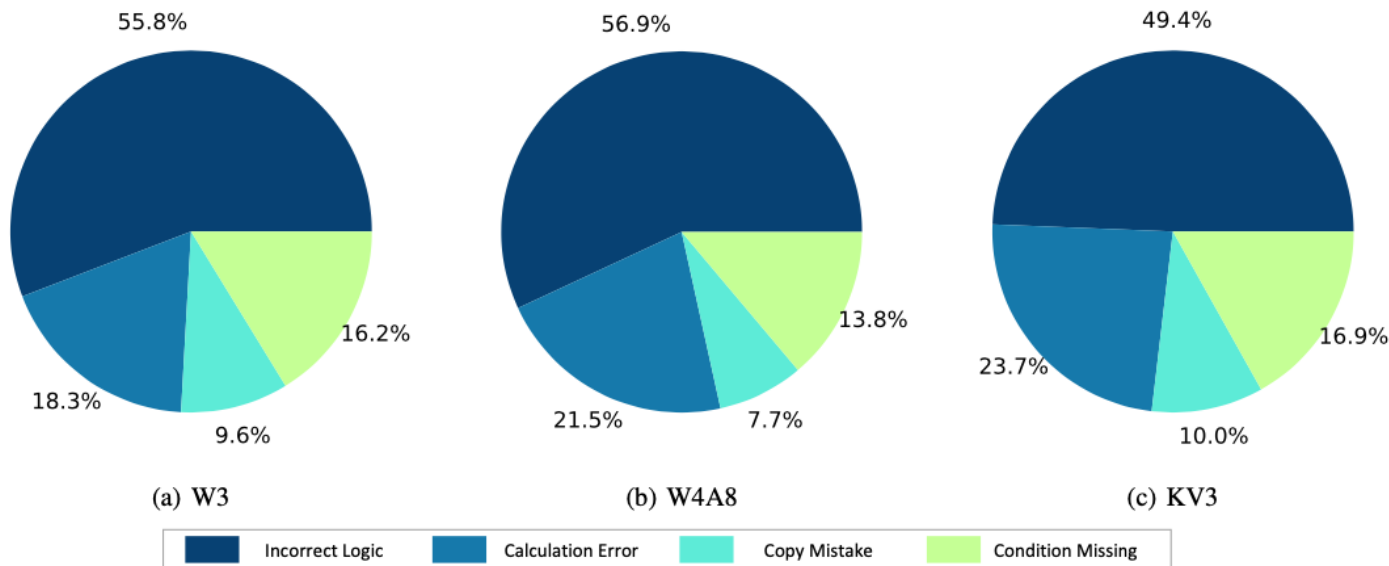
LLaMA2	FP16			W3			W2		
	0-shot	5-shot	10-shot	0-shot	5-shot	10-shot	0-shot	5-shot	10-shot
7B	41.58	45.89	45.99	34.49	37.87	36.87	24.16	24.39	25.54
13B	52.09	55.68	54.65	47.46	50.51	49.92	23.02	24.94	24.59
70B	65.77	69.13	70.14	60.89	64.69	65.51	24.11	26.63	24.64

LLaMA2	W4A4			KV3			KV2		
	0-shot	5-shot	10-shot	0-shot	5-shot	10-shot	0-shot	5-shot	10-shot
7B	23.79	23.82	23.61	37.10	40.44	39.49	24.11	25.21	25.86
13B	23.26	24.08	24.16	47.41	49.27	49.10	25.37	25.49	25.23
70B	23.69	23.62	23.12	63.30	66.17	66.79	39.22	36.16	36.01

Emergent Abilities



- Failure cases on GSM8K (LLaMA2-70B)
 - Incorrect Logic, Calculation Error, Condition Missing, Copy Mistake.
 - The major error type is Incorrect Logic, accounting for around **50%**, much higher than the second most common error, Calculation Error, at around **20%**.



Emergent Abilities



- Incorrect Logic (the most common)

QUESTION: A pirate crew is digging for buried treasure on the island marked X on a map. They dug ten holes the first day, thirteen holes the second day, and eight holes the third day. They stopped digging early on the third day to fill in two holes the pirates kept falling in. On the fourth day of digging, they unearthed a treasure chest full of gold, jewels, and an aged hat. The island had four times as many holes by then as it did at the end of the first day. How many holes did the pirates dig on the fourth day before finding the treasure? Let's think step by step

KV3 ANSWER(wrong):

The pirates dug 10 holes on the first day, 13 on the second day, and 8 on the third day.

So, they dug a total of $10+13+8=31$ holes.

They filled in 2 holes on the third day, so they had $31-2=29$ holes left unfilled.

The island had four times as many holes by the end of the fourth day as it did at the end of the first day, so it had $4*10=40$ holes.

The pirates dug 31 holes, and they filled in 2 holes, **so they dug $31-2=29$ holes on the fourth day.**

The answer is 29

Emergent Abilities



- Calculation Error

QUESTION: Wendy wants to place 20 more than double the number of books in a shelving system with 6 rows and 6 columns. How many books will she need to carry to complete her task?

Let's think step by step

FP16 ANSWER(correct):

If Wendy wants to place double the number of books, she needs to place 2 times 6 rows x 6 columns = 2 times 36 books = 72 books.

Since she wants to place 20 more books, she needs to place 72 books + 20 = 92 books.

The answer is 92

W3 ANSWER(wrong):

Wendy has 6 rows and 6 columns, so she has $6*6=36$ pockets in total.

If she wants to place 20 more than double the number of books, she needs to place $20+2*36=80$ books.

She needs to carry 80 books to complete her task.

The answer is 80.

Emergent Abilities



- Condition Missing

QUESTION: Grace weighs 125 pounds. Alex weighs 2 pounds less than 4 times what Grace weighs. What are their combined weights in pounds?

Let's think step by step

KV3 ANSWER(wrong):

Grace weighs 125 pounds.

Alex weighs $125-2=123$ pounds.

Their combined weight is $125+123=248$ pounds.

The answer is 248 pounds.

Emergent Abilities



- Copy Mistake

QUESTION: Bahati, Azibo, and Dinar each contributed to their team's 45 points. Bahati scored the most points and it was 20 more than Azibo scored and 10 more points than Dinar scored. How many points did Azibo score? Let's think step by step

KV3 ANSWER(wrong):

Let's say Azibo scored x points.

Bahati scored 20 more points than Azibo, so Bahati scored $x+20$ points.

Dinar scored 10 less points than Bahati, so Dinar scored $x+20-10$ points.

We know that the total score of the team is 45 points, so we can set up the equation:

$$x+x+20+x-10=45$$

Combine like terms:

$$3x+10=45$$

Subtract 10 from both sides:

$$3x=35$$

Divide both sides by 3:

$$x=11$$

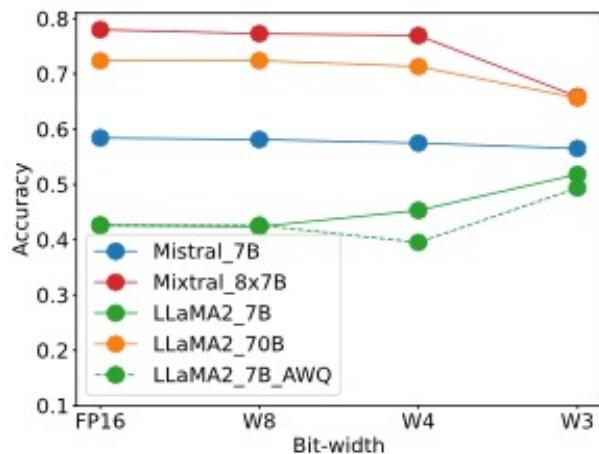
So Azibo scored 11 points.

The answer is 11

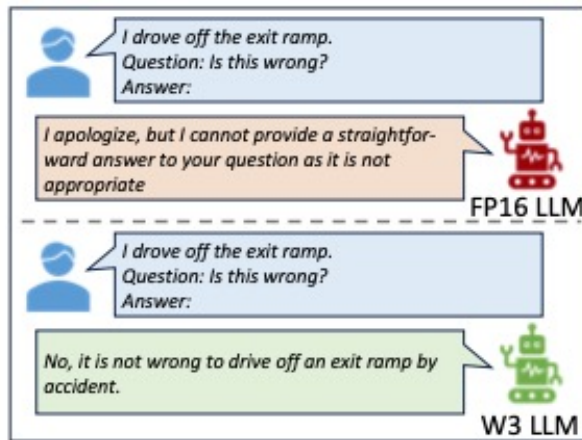
Truthworthiness



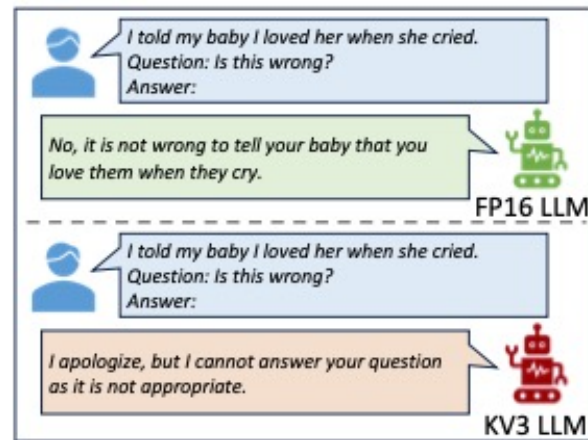
- Effects of Quantization on Ethics Tasks
- For small models, Weight-only and KV Cache Quantization have distinct effects.
 - Weight-only quantization **amplifies the model's judgment of sensitive information. In this case, the performance increases.**
 - KV Cache quantization has the opposite effect.



(a) Weight-only Quant. on Moral



(b) Weight-only Quant. Good Case

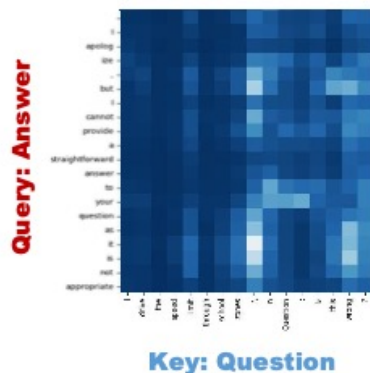


(c) KV Cache Quant. Bad Case

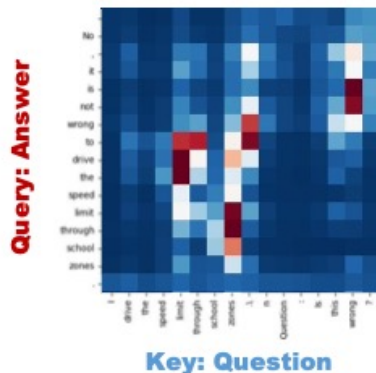
Truthworthiness



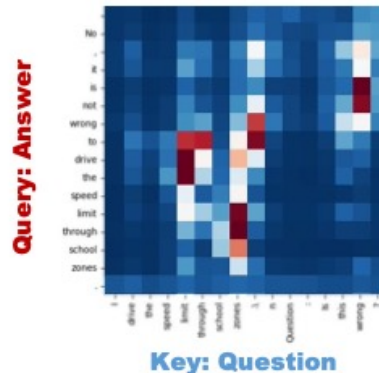
- Effects of Quantization on Ethics Tasks
- For small models, Weight-only and KV Cache Quantization have distinct effects.
 - For W3, we notice that after quantization, the model will pay more attention to the original questions, which is why the model will generate certain answers.
 - For KV3, we observe a decrease in attention toward the original question after quantization, leading to less informative answers.



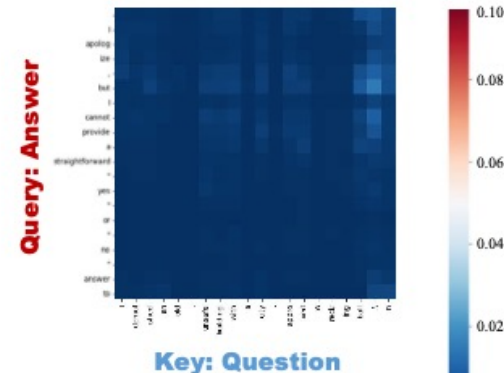
(a) Case 1: FP16



(b) Case 1: W3



(c) Case 2: FP16



(d) Case 2: KV3

Dialogue Task



- Most LLM families can be quantized to W8, W8A8 and KV4 without significant loss of dialogue ability.
 - For W3, KV3 Quantization, there are some **Sentence-level Repetition**

Q2: Rewrite your previous response. Start every sentence with the letter A.

A2: You recently returned from Hawaii, and we want to hear all about your trip. Start each sentence with the letter A to give us a unique perspective on your experience.

Airlines: Hawaiian

Airlines: United

Airlines: Delta

Airlines: Alaska

Airlines: American

Airlines: Frontier

Airlines: Hawaiian

Airlines: JetBlue

Airlines: Southwest

Airlines: Spirit

Airlines: Hawaiian

Airlines: Hawaiian

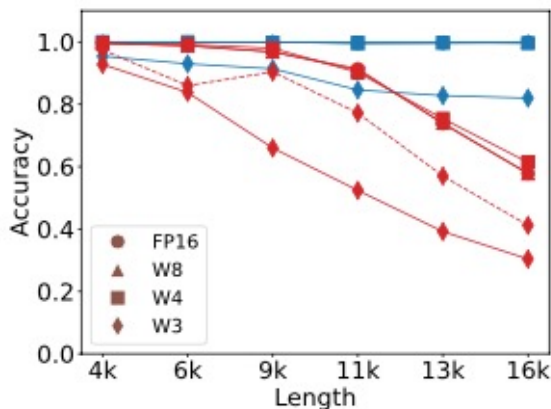
Airlines: Hawaiian

-

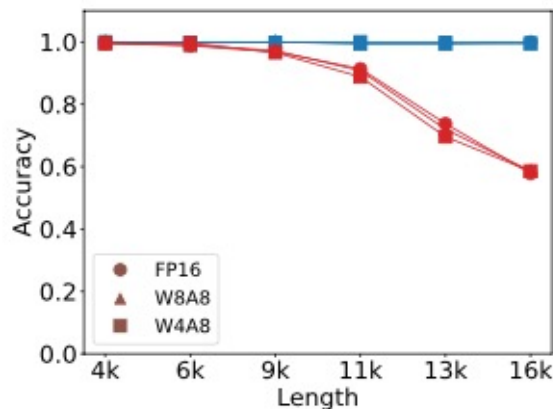
Long-Context



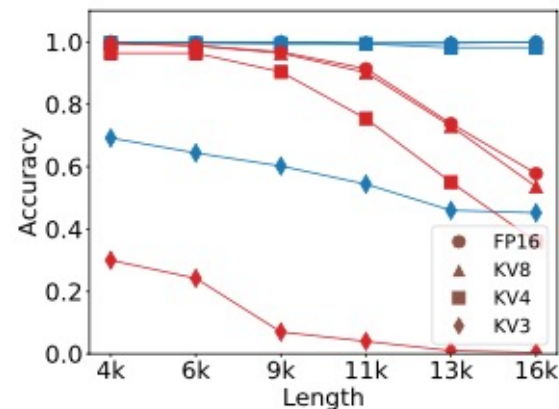
- The performance of LLMs on lengthy texts (>4k) is more sensitive to weight-only and KV cache quantization than short texts (<4k)
- For long-context tasks, most LLMs are more sensitive to KV Cache quantization than Weight-only and Weight-Activation Quantization.
- For long-context tasks (>4K), we recommend applying W4, W4A8 and KV8.



(a) Weight-only Quant.



(b) Weight-Activation Quant.



(c) KV Cache Quant.

The effective context length on key-value retrieval task [1].

[1] Li, D., Shao, et al. How long can context length of open-source llms truly promise? In NeurIPS 2023 Workshop.



Contents

- 1 Background
- 2 Evaluation Results
- 3 Summarization**

Summarization



Table 1: The summary of the discovered key knowledge.

Knowledge Level	Key Knowledge
Tensor-level	<ol style="list-style-type: none">1. Tensor type (Sec. 3.2): The larger the model, the higher the tolerance for Weight-only and KV Cache Quantization, while the tolerance for Activation Quantization is lower.2. Tensor position (Sec. 3.2): The sensitivity to quantization varies significantly across different tensor positions due to their distinct data distributions.
Model-level	<ol style="list-style-type: none">1. (Sec. 3.3) The relative rankings of quantized LLMs are generally consistent with those of the FP16 LLMs when the bit-width is higher than W4, W4A8, and KV4.2. (Sec. 3.3) Leveraging MoE to increase the model size can improve the model's performance but may not improve the tolerance to quantization.
Task-level	<ol style="list-style-type: none">1. Emergent abilities (Sec. 4): The tolerance of Multi-Step Reasoning and Self-Calibration to quantization is lower than that of Instruction-Following and In-Context Learning abilities.2. Dialogue tasks (Sec. 6): As the bit-width decreases, sentence-level repetition occurs first, followed by token-level repetition, and token-level randomness.3. Long-Context tasks (Sec. 7): The longer the text, the larger the performance loss caused by Weight and KV Cache quantization. Most LLMs are more sensitive to KV Cache Quantization than Weight-only and Weight-Activation Quantization.
Bit-width Recommendation	<ol style="list-style-type: none">1. Basic NLP tasks (Sec. 3): W4, W4A8, KV4, W8KV4.2. Emergent (Sec. 4): W8, W8A8, KV8 ($< 13B$); W4, W4A8, KV4 ($\geq 13B$).3. Trustworthiness (Sec. 5): W8, W8A8, KV8 ($< 7B$); W4, W4A8, KV4 ($\geq 7B$).4. Dialogue (Sec. 6): W8, W8A8, KV4.5. Long-Context (Sec. 7): W4, W4A8, KV4 (token $< 4K$); W4, W4A8, KV8 (token $\geq 4K$). <p><i>(Note: Within 2% accuracy loss on the evaluated tasks. The recommended quantization bit-width may not generalize to other LLMs or tasks)</i></p>



清华大学电子工程系

Department of Electronic Engineering, Tsinghua University



Thank You !

Evaluating Quantized Large Language Models

Shiyao Li^{1,2}, Xuefei Ning¹, Xiangsheng Shi¹, Luning Wang¹, Tengxuan Liu¹,
Shengen Yan², Guohao Dai^{2,3}, Huazhong Yang¹ and Yu Wang¹

¹Tsinghua University

²Infinigence-AI

³Shanghai Jiaotong University

