NEURAL INFORMATION PROCESSING SYSTEMS

OneBit: Towards Extremely Low-bit Large Language Models

Yuzhuang Xu Xu Han Zonghan Yang Shuo Wang
Qingfu Zhu Zhiyuan Liu Weidong Liu Wanxiang Che

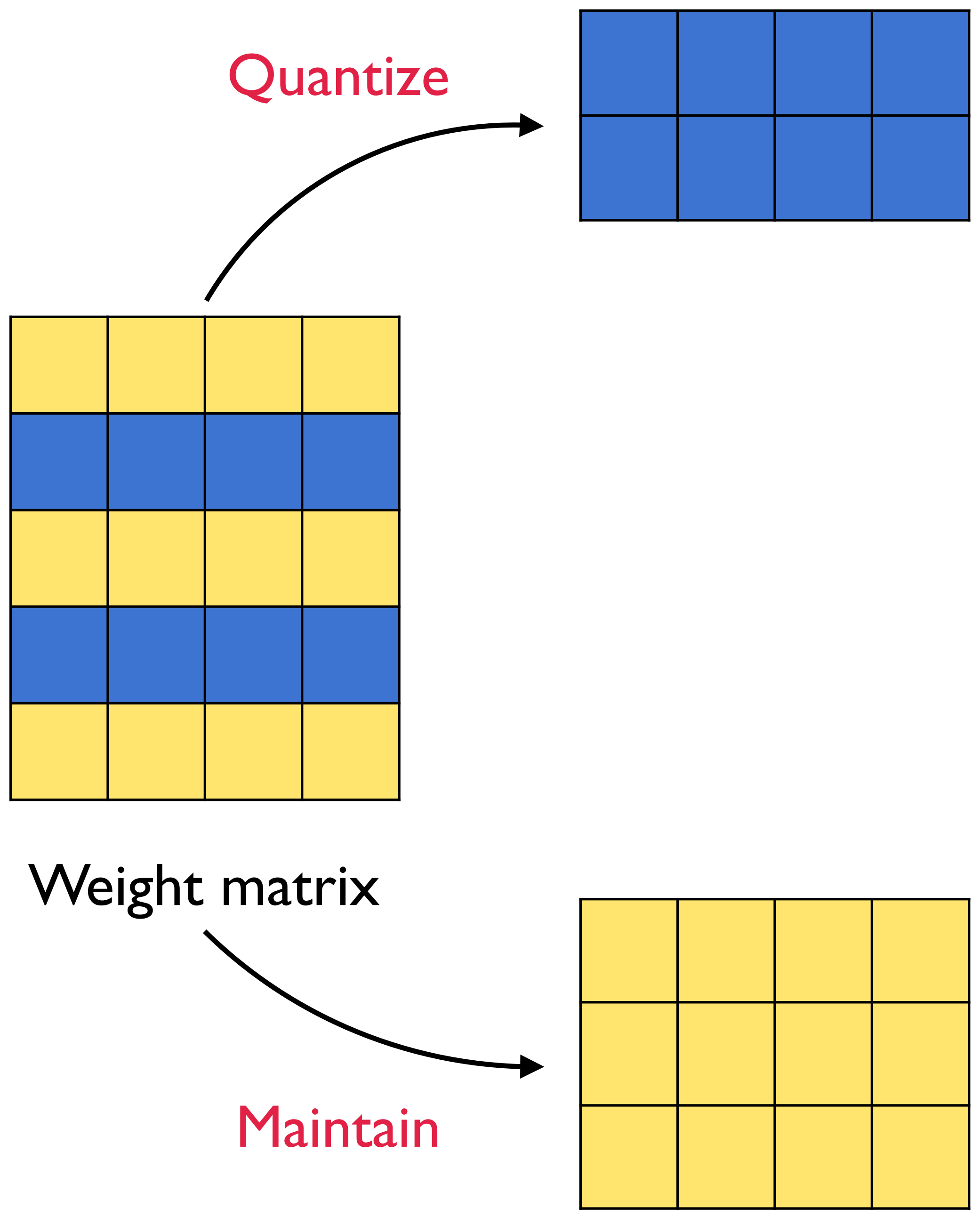
Model Quantization

- Model quantization is one of the main methods of model compression
- Convert weights from high precision to low-bit representation
- Model performance is almost unchanged or the loss is tolerable
- Its core problem lies in how to convert weights into appropriate low-precision representations

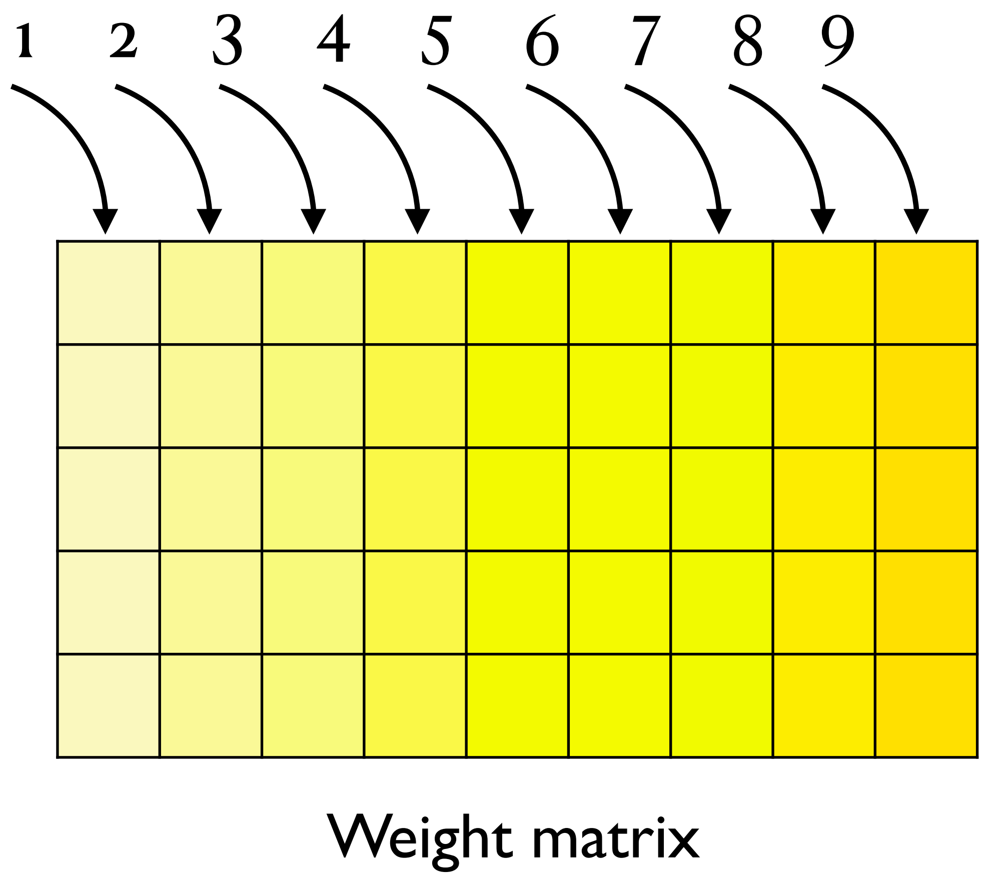
FP16

INT4

Weight Representation Methods

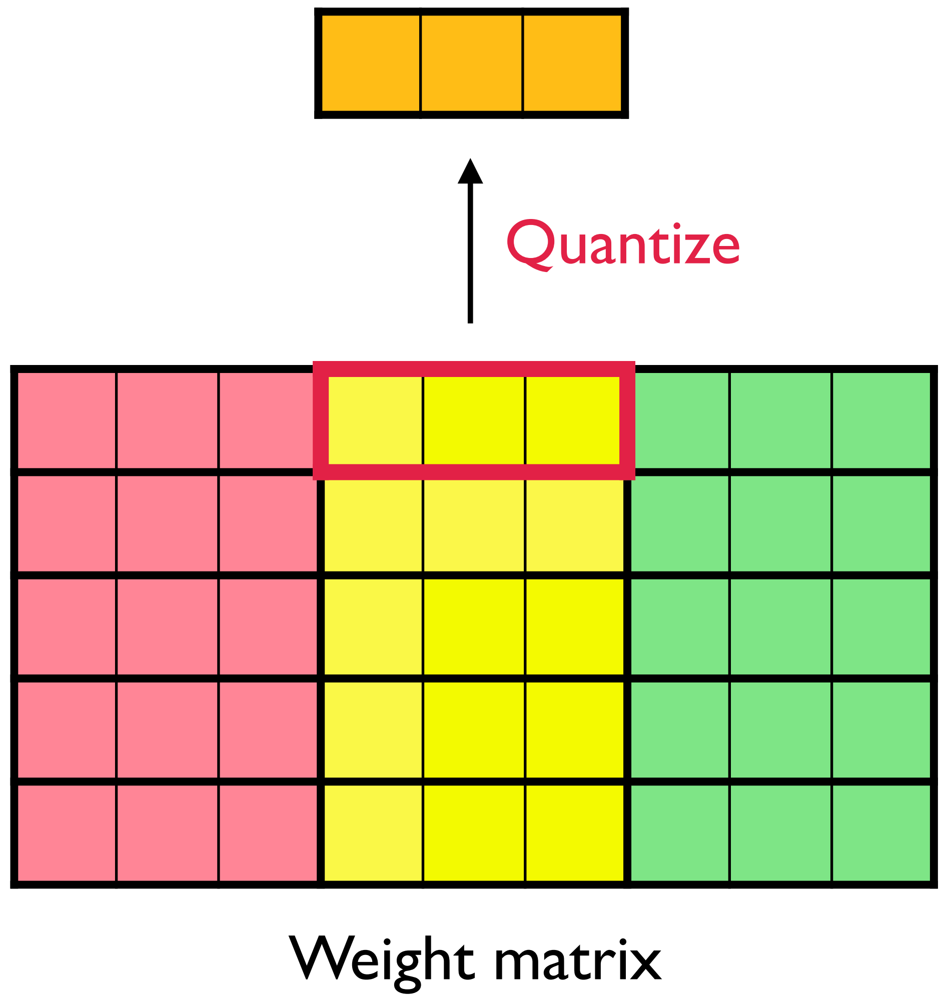


Mixed precision[1]



Quantize the matrix iteratively

Compensate for the loss simultaneously[2]



Vector quantization[3]

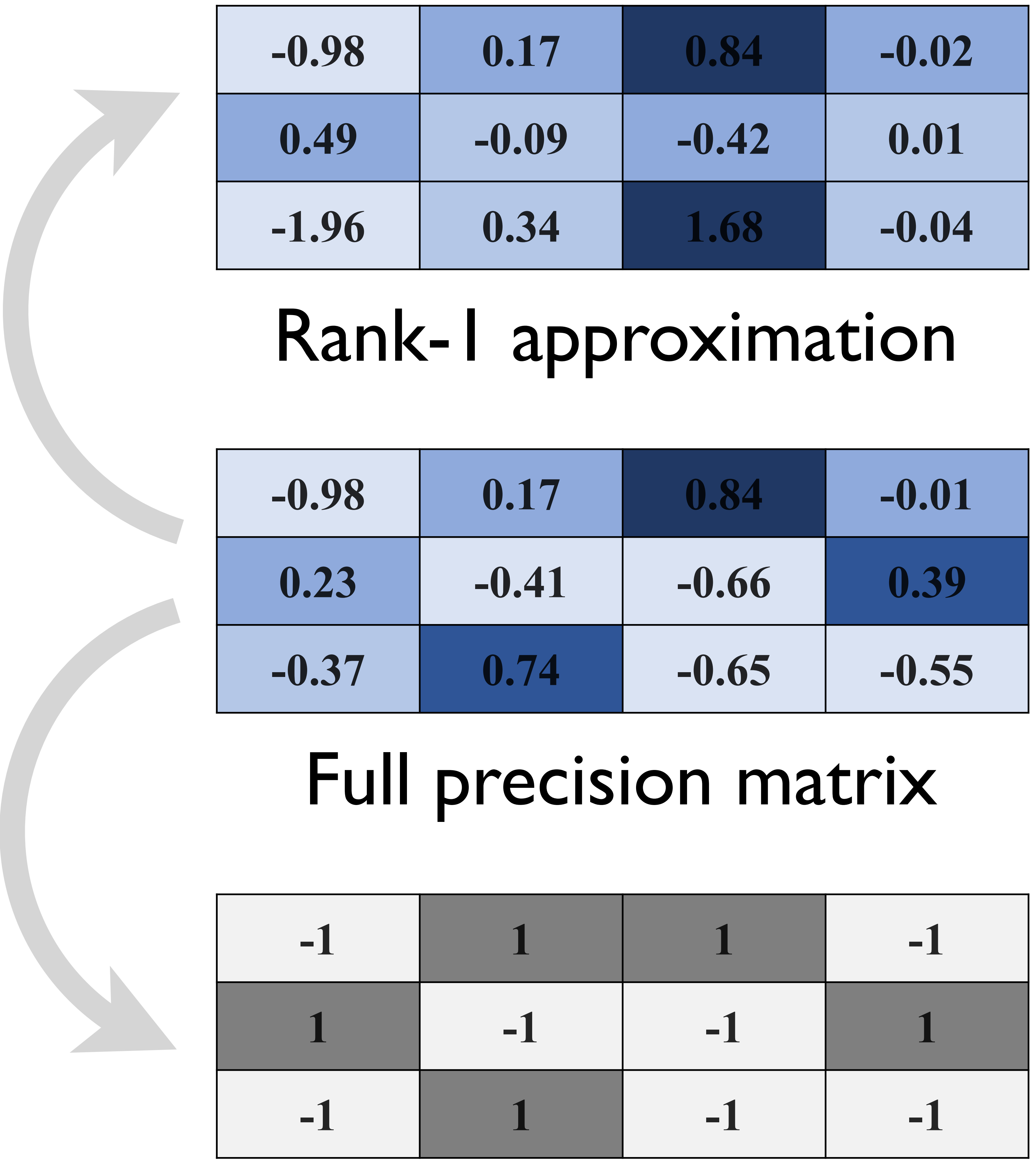
[1] LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. Tim Dettmers, et al. 2022.

[2] GPTQ: Accurate Post-training Quantization for Generative Pre-trained Transformers. Elias Frantar, et al. 2023.

[3] Extreme Compression of Large Language Models via Additive Quantization. Vage Egiazarian, et al. 2024.

Rank & Floating Precision

- Our motivation is that both floating point precision and rank are very important



-0.98	0.17	0.84	-0.02
0.49	-0.09	-0.42	0.01
-1.96	0.34	1.68	-0.04

Rank-1 approximation

- high floating precision
- low rank

-0.98	0.17	0.84	-0.01
0.23	-0.41	-0.66	0.39
-0.37	0.74	-0.65	-0.55

Full precision matrix

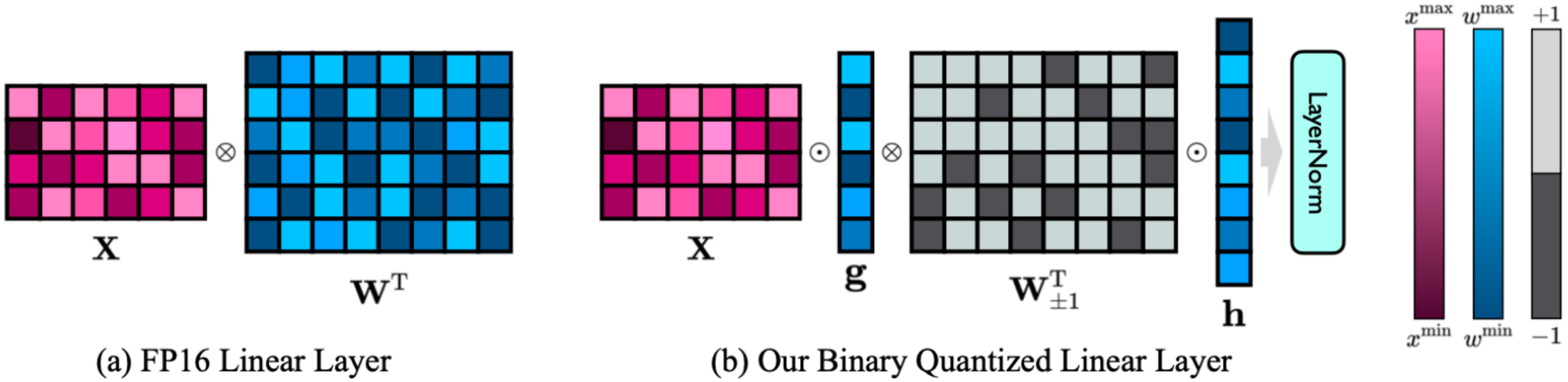
- high floating precision
- high rank

-1	1	1	-1
1	-1	-1	1
-1	1	-1	-1

1-bit symbol matrix

- low floating precision
- high rank

1-bit Linear Architecture

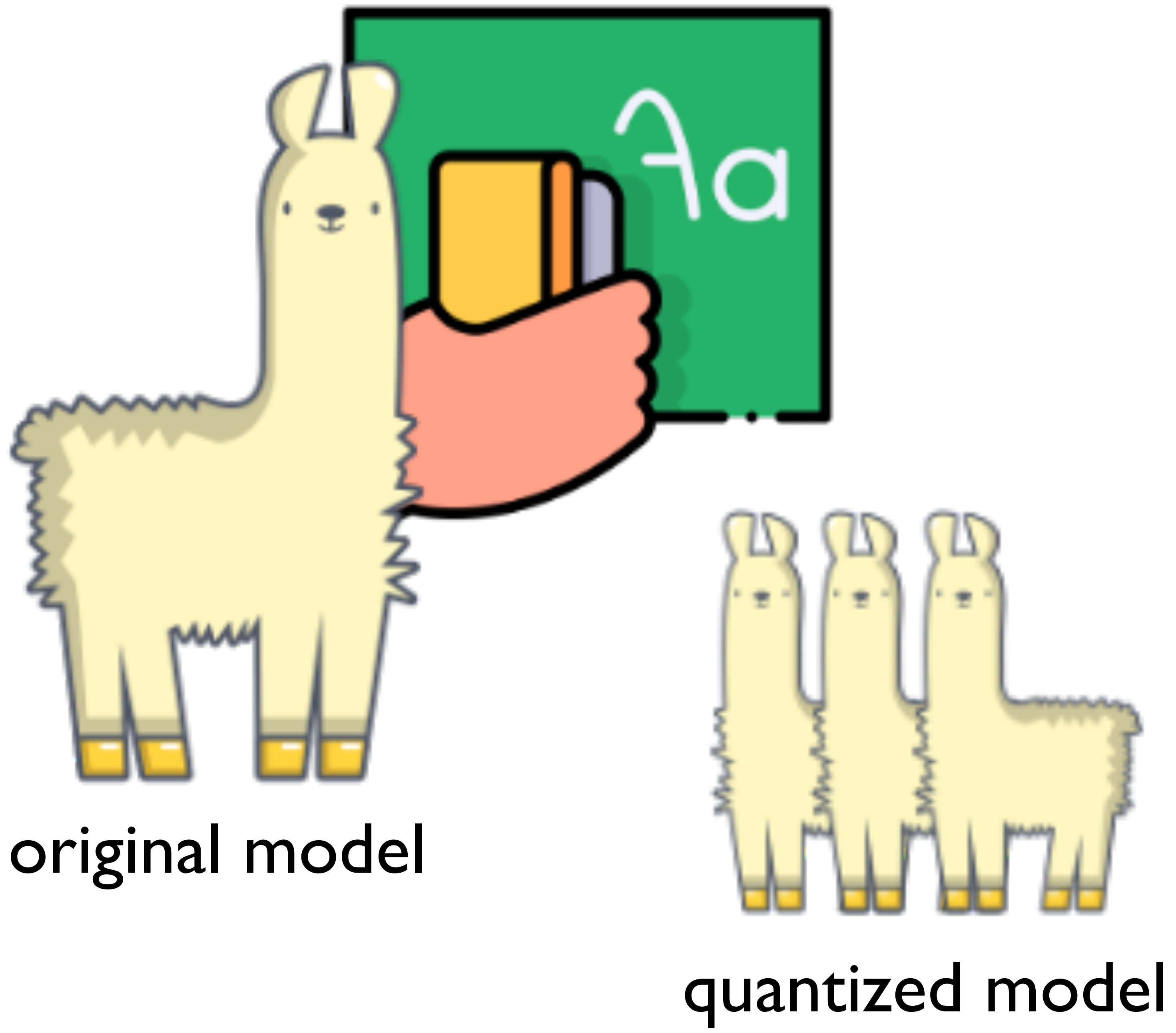


(a) FP16 Linear Layer

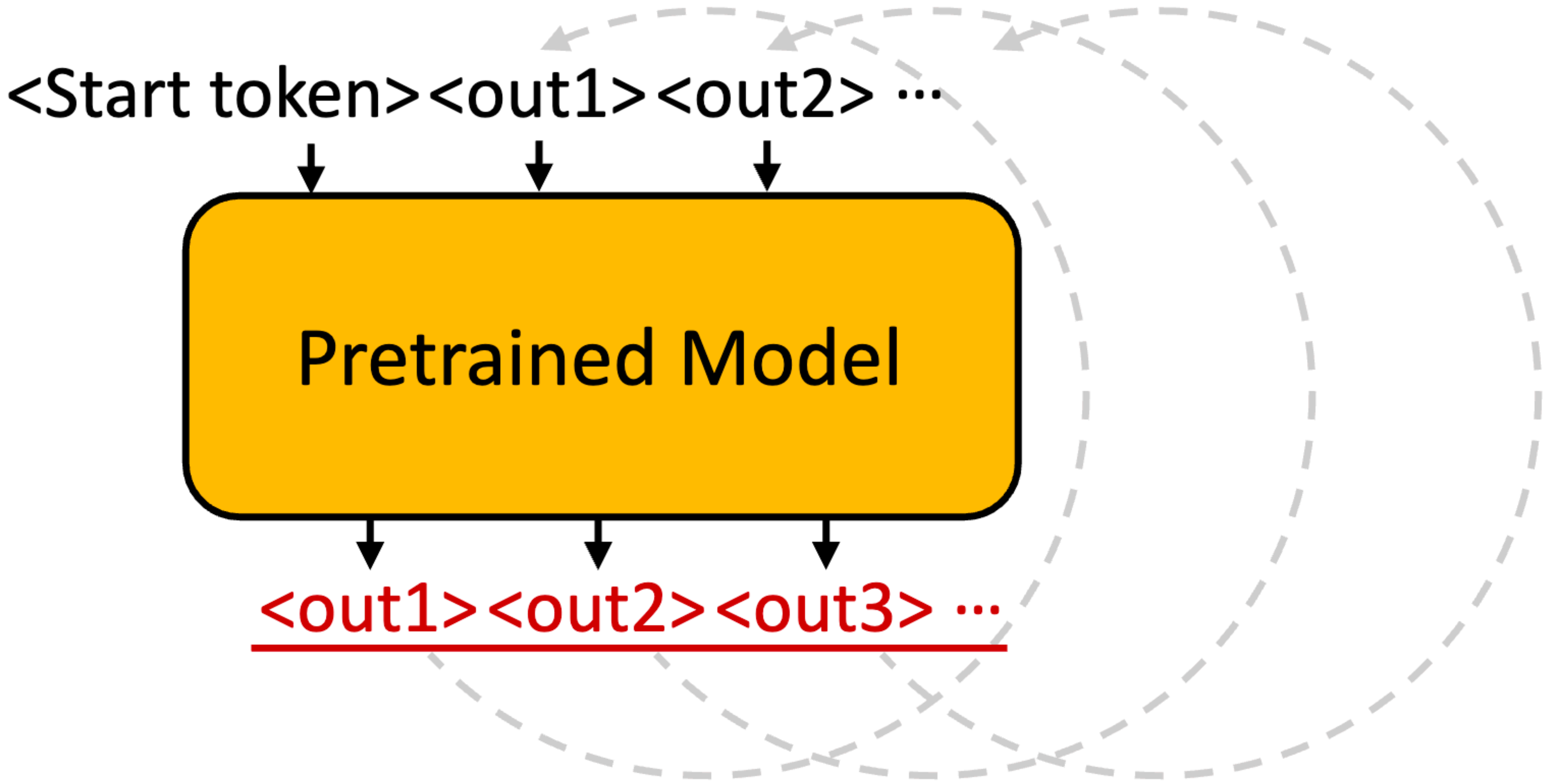
(b) Our Binary Quantized Linear Layer

Figure 2: The main idea of our method OneBit. The left is the original FP16 `Linear` Layer, in which both the activation $\mathbf{X}$ and the weight matrix $\mathbf{W}$ are in FP16 format. The right is our proposed architecture. Only value vectors $\mathbf{g}$ and $\mathbf{h}$ are in FP16 format and the weight matrix consists of ± 1 instead.

Knowledge Transfer



Teacher guide student



<Start token><out1><out2> …

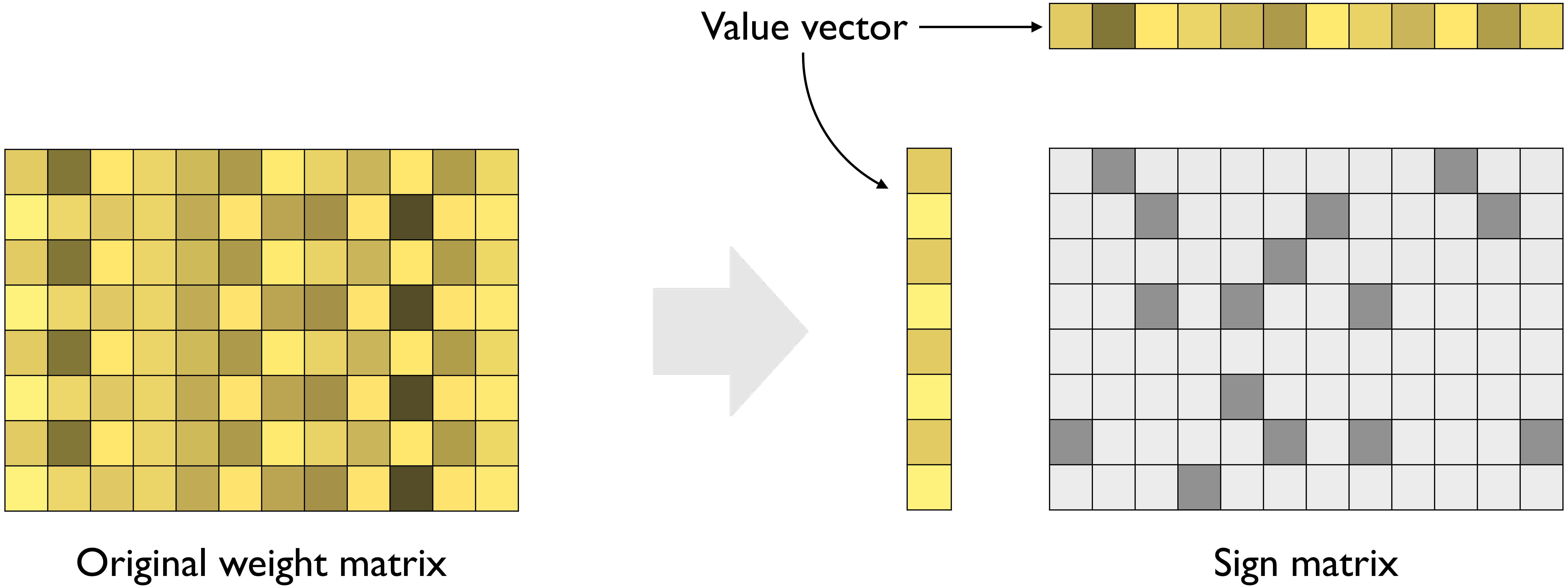
Pretrained Model

<out1><out2><out3> …

Data Generation[1]

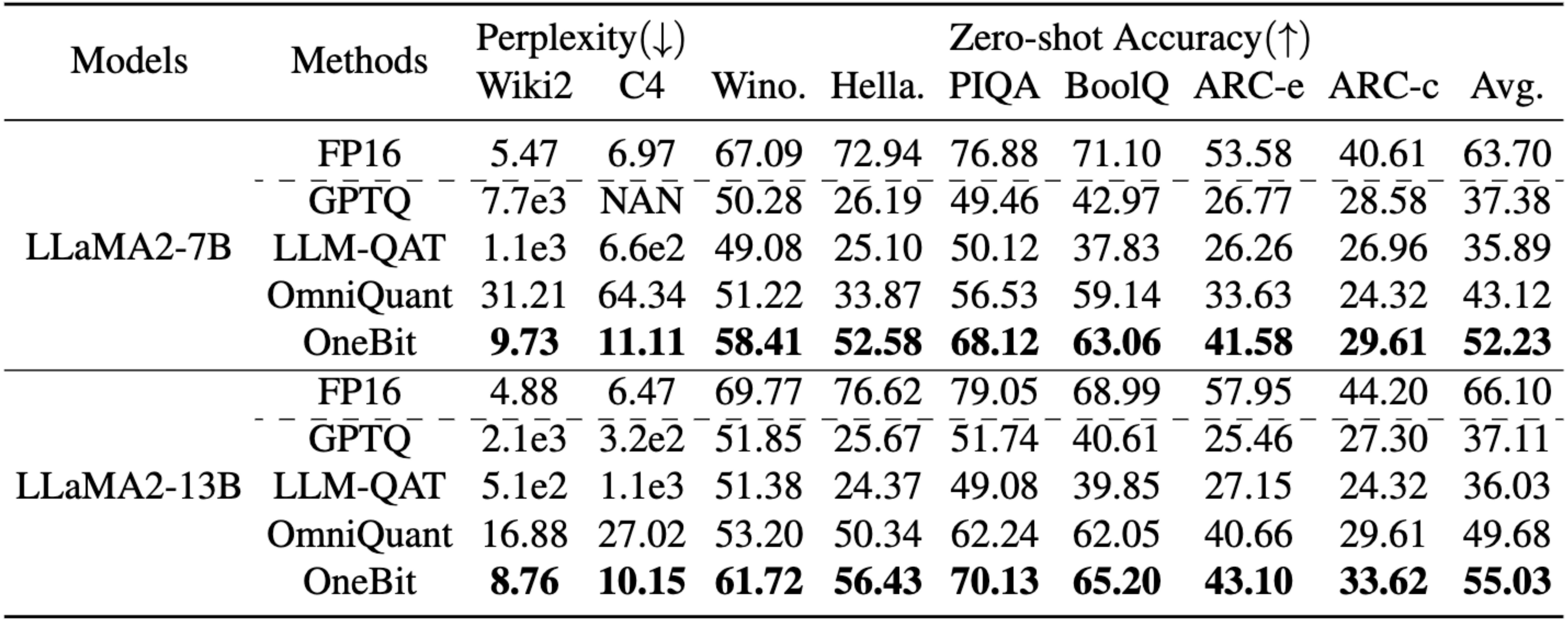
[1] LLM-QAT: Data-Free Quantization Aware Training for Large Language Models. Zechun Liu, et al. 2023.

Weight Initialization



Sign-Value-Independent-Decomposition

Capability Evaluation



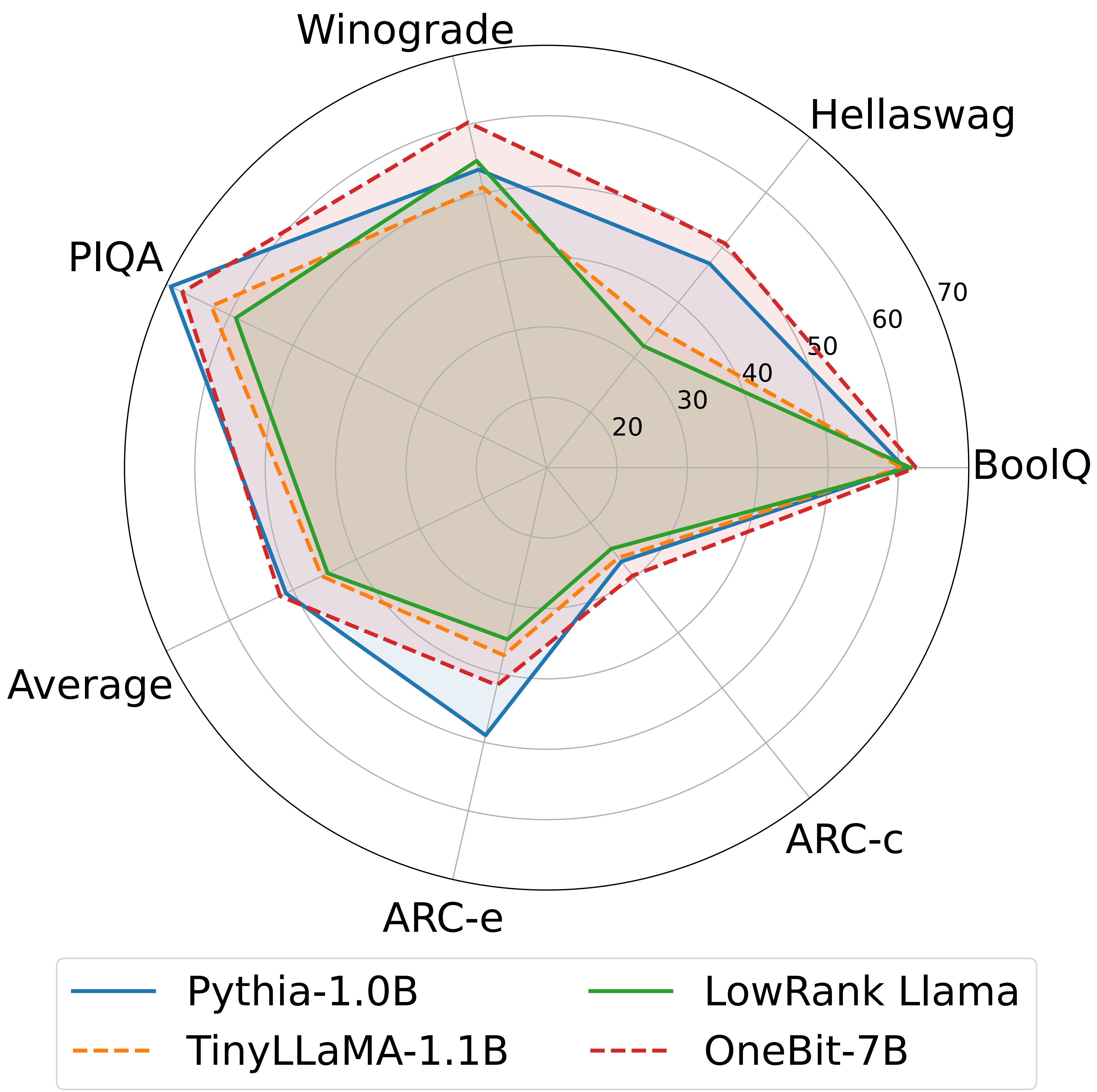
Models	Methods	Perplexity(↓)		Zero-shot Accuracy(↑)						
		Wiki2	C4	Wino.	Hella.	PIQA	BoolQ	ARC-e	ARC-c	Avg.
LLaMA2-7B	FP16	5.47	6.97	67.09	72.94	76.88	71.10	53.58	40.61	63.70
	GPTQ	7.7e3	NAN	50.28	26.19	49.46	42.97	26.77	28.58	37.38
	LLM-QAT	1.1e3	6.6e2	49.08	25.10	50.12	37.83	26.26	26.96	35.89
	OmniQuant	31.21	64.34	51.22	33.87	56.53	59.14	33.63	24.32	43.12
	OneBit	**9.73**	**11.11**	**58.41**	**52.58**	**68.12**	**63.06**	**41.58**	**29.61**	**52.23**
LLaMA2-13B	FP16	4.88	6.47	69.77	76.62	79.05	68.99	57.95	44.20	66.10
	GPTQ	2.1e3	3.2e2	51.85	25.67	51.74	40.61	25.46	27.30	37.11
	LLM-QAT	5.1e2	1.1e3	51.38	24.37	49.08	39.85	27.15	24.32	36.03
	OmniQuant	16.88	27.02	53.20	50.34	62.24	62.05	40.66	29.61	49.68
	OneBit	**8.76**	**10.15**	**61.72**	**56.43**	**70.13**	**65.20**	**43.10**	**33.62**	**55.03**

Zip Ratio

90.4%

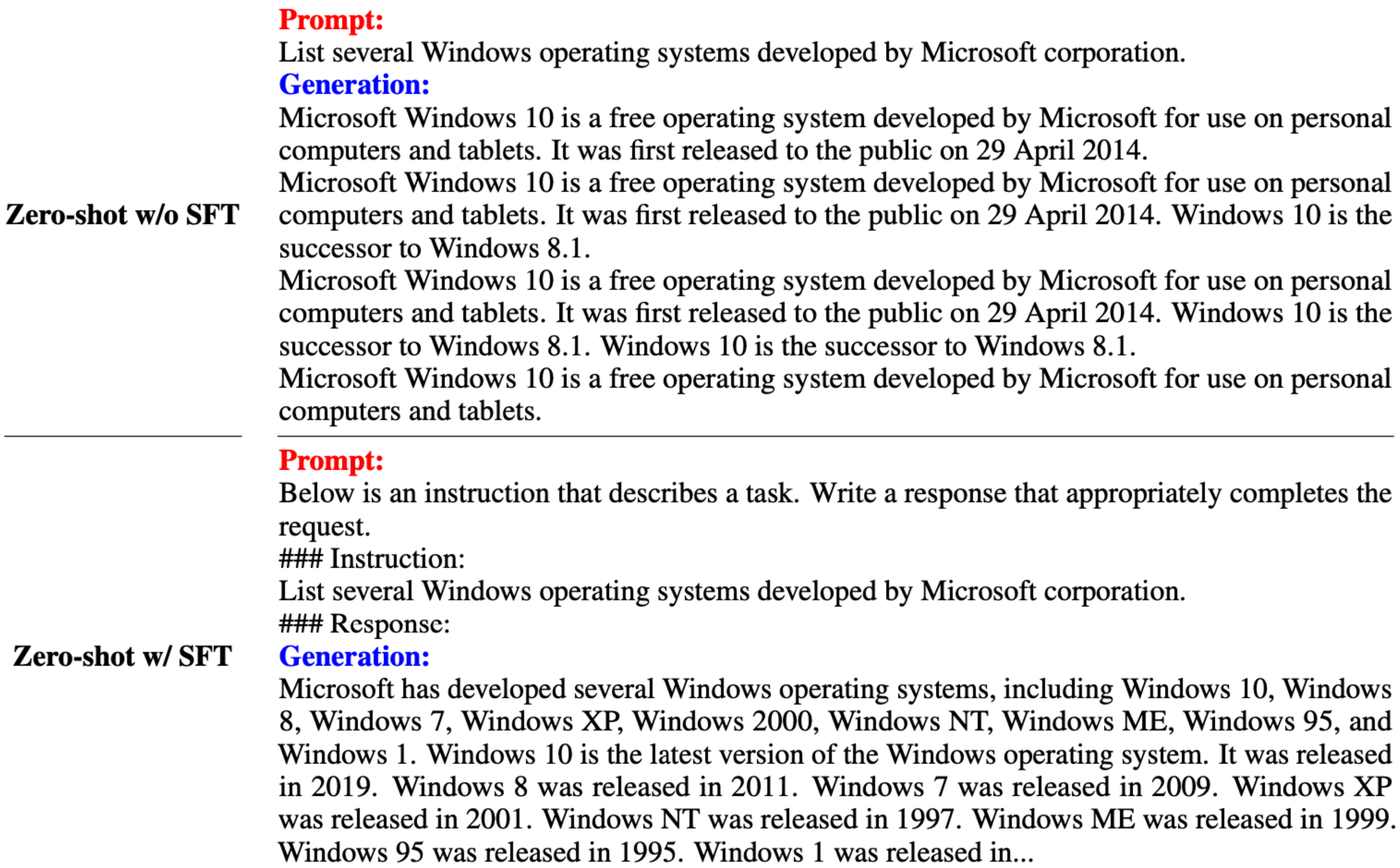
91.5%

Problem Solving Ability



- Similar to but stronger than other baseline models
- The ability is significantly reduced on some datasets

Instruction Following Ability



Zero-shot w/o SFT	**Prompt:** List several Windows operating systems developed by Microsoft corporation. **Generation:** Microsoft Windows 10 is a free operating system developed by Microsoft for use on personal computers and tablets. It was first released to the public on 29 April 2014. Microsoft Windows 10 is a free operating system developed by Microsoft for use on personal computers and tablets. It was first released to the public on 29 April 2014. Windows 10 is the successor to Windows 8.1. Microsoft Windows 10 is a free operating system developed by Microsoft for use on personal computers and tablets. It was first released to the public on 29 April 2014. Windows 10 is the successor to Windows 8.1. Windows 10 is the successor to Windows 8.1. Microsoft Windows 10 is a free operating system developed by Microsoft for use on personal computers and tablets.
Zero-shot w/ SFT	**Prompt:** Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: List several Windows operating systems developed by Microsoft corporation. ### Response: **Generation:** Microsoft has developed several Windows operating systems, including Windows 10, Windows 8, Windows 7, Windows XP, Windows 2000, Windows NT, Windows ME, Windows 95, and Windows 1. Windows 10 is the latest version of the Windows operating system. It was released in 2019. Windows 8 was released in 2011. Windows 7 was released in 2009. Windows XP was released in 2001. Windows NT was released in 1997. Windows ME was released in 1999. Windows 95 was released in 1995. Windows 1 was released in...

Efficiency at Inference-time

$$\begin{bmatrix}-0.2 & 1.5 & 0.3\end{bmatrix} \otimes \begin{bmatrix}0.4\\-0.1\\0.9\end{bmatrix} \qquad -0.2 \times 0.4 + 1.5 \times (-0.1) + 0.3 \times 0.9$$

$$\begin{bmatrix}-0.2 & 1.5 & 0.3\end{bmatrix} \otimes \begin{bmatrix}+1\\-1\\+1\end{bmatrix} \qquad \mathrm{Pos}(-0.2) + \mathrm{Neg}(1.5) + \mathrm{Pos}(0.3)$$

Pos(·) indicates the instruction to set the sign of the floating point number to positive;

Neg(·) indicates the instruction to set the sign of the floating point number to negative.

Thanks!

Paper

Code

HomePage