

Table 6: Complete omparison of the perplexity score on WikiText2 and averaged accuracy on Zero-shot Common Sense Reasoning tasks on **LLaMA-2**. We reported the mean and standard deviation across six trails for SpinQuant as well as our reproduced results of GPTQ and QuaRot.

Model	#Bits W-A-KV	Method	ARC-e (↑)	ARC-c (↑)	BoolQ (↑)	PIQA (↑)	SIQA (↑)	HellaS. (↑)	OBQA (↑)	WinoG. (↑)	Avg. (↑)	Wiki2 (↓)
7B	16-16-16	Full Precision	75.0	50.8	77.3	78.9	48.5	76.0	59.3	69.5	66.9	5.5
		RTN	71.3	46.0	73.5	76.9	47.2	72.5	54.2	66.9	63.6	7.2
	4-16-16	SmoothQuant	66.2	42.5	67.4	75.8	44.1	67.2	44.6	64.6	59.1	7.5
		LLM-QAT	73.3	48.6	73.2	78.2	48.8	73.6	55.0	68.4	64.9	5.9
		OmniQuant	67.8	37.9	—	77.1	—	—	—	67	—	5.7
		AQLM	68.9	40.3	—	77.7	—	—	—	67.3	—	—
		QuIP#	69.1	40.5	—	78.4	—	—	—	67.6	—	—
		GPTQ	72.6 ±0.3	46.8 ±0.5	73.9 ±0.6	78.2 ±0.4	46.8 ±0.6	73.8 ±0.2	55.5 ±1.1	68.6 ±0.7	64.5 ±0.3	11.3 ±0.97
		QuaRot*	69.5 ±1.9	45.3 ±1.4	72.8 ±1.2	77.0 ±0.8	46.4 ±0.8	69.8 ±2.1	52.5 ±2.0	66.3 ±1.5	62.4 ±1.0	6.9 ±0.45
		QuaRot	74.2 ±0.4	50.0 ±0.3	75.3 ±0.8	78.2 ±0.3	48.3 ±0.4	74.7 ±0.2	57.2 ±1.4	68.1 ±1.5	65.8 ±0.3	5.6 ±0.01
		SpinQuant*	72.2 ±0.9	48.6 ±0.5	73.4 ±2.2	78.2 ±0.3	46.9 ±0.4	74.2 ±0.3	55.5 ±1.2	67.9 ±0.2	64.6 ±0.3	5.5 ±0.01
		SpinQuant	73.8 ±0.4	49.4 ±0.5	76.0 ±1.2	78.4 ±0.3	47.9 ±0.4	74.9 ±0.3	57.8 ±0.8	69.1 ±0.3	65.9 ±0.3	5.6 ±0.01
	4-4-16	RTN	26.6	22.1	44.3	50.9	38.9	26.2	26.6	49.4	35.6	2,167.2
		SmoothQuant	37.8	27.1	51.9	59.4	40.2	34.3	31.6	52.4	41.8	254.5
		LLM-QAT	46.2	32.4	61.8	62.0	41.3	47.6	36.1	54.7	47.8	12.9
		GPTQ	27.6 ±1.0	24.9 ±0.8	47.4 ±2.7	50.7 ±0.7	38.6 ±0.4	26.9 ±0.2	28.3 ±1.9	49.9 ±1.2	36.8 ±0.4	8,949.0
		QuaRot*	65.3 ±2.1	41.8 ±1.8	69.6 ±1.0	74.2 ±1.0	44.9 ±0.6	64.9 ±2.3	48.6 ±1.8	62.4 ±0.7	59.0 ±1.0	8.2 ±0.73
		QuaRot	71.8 ±1.2	46.8 ±1.3	73.4 ±0.8	76.7 ±0.3	47.1 ±0.5	72.9 ±0.2	52.6 ±1.4	67.0 ±1.0	63.5 ±0.3	6.1 ±0.01
		SpinQuant*	67.7 ±0.5	44.8 ±0.9	71.4 ±1.9	76.6 ±0.7	45.8 ±0.7	71.3 ±0.5	51.5 ±2.5	65.2 ±1.1	61.8 ±0.4	6.1 ±0.03
		SpinQuant	72.1 ±0.9	47.5 ±1.4	74.4 ±1.1	77.0 ±0.4	47.3 ±0.5	73.2 ±0.3	54.4 ±1.9	66.9 ±0.8	64.1 ±0.4	5.9 ±0.00
	4-4-4	RTN	27.1	24.4	44.8	51.4	39.4	26.7	33.0	50.0	37.1	2,382.5
		SmoothQuant	31.4	24.8	51.4	54.1	39.4	29.1	31.9	50.0	39.0	698.7
		LLM-QAT	42.0	27.7	59.5	58.9	41.0	43.1	33.5	53.3	44.9	14.9
		GPTQ	27.6 ±1.1	23.6 ±0.8	47.8 ±1.0	51.0 ±1.6	38.7 ±0.5	27.0 ±0.4	28.5 ±2.8	50.3 ±0.9	36.8 ±0.6	9,253.1
		QuaRot*	65.3 ±1.6	40.9 ±1.6	69.3 ±0.5	74.4 ±1.4	45.0 ±1.1	64.7 ±2.1	48.4 ±3.1	61.4 ±1.9	58.7 ±1.0	8.2 ±0.36
		QuaRot	70.1 ±0.8	46.1 ±1.2	72.0 ±0.4	76.8 ±0.5	46.8 ±0.6	71.8 ±0.5	52.4 ±1.0	64.5 ±0.6	62.5 ±0.3	6.4 ±0.01
		SpinQuant*	68.1 ±0.9	44.4 ±1.1	71.4 ±0.7	75.7 ±0.7	45.7 ±0.5	71.0 ±0.7	51.4 ±1.2	64.4 ±0.7	61.5 ±0.3	6.2 ±0.03
		SpinQuant	72.6 ±0.9	47.5 ±0.5	73.9 ±1.2	77.0 ±0.3	47.2 ±0.6	73.0 ±0.4	54.1 ±2.4	66.9 ±0.5	64.0 ±0.3	5.9 ±0.01
13B	16-16-16	Full Precision	75.3	51.4	79.8	80.4	50.5	79.8	56.8	72.5	68.3	5.0
		RTN	63.7	40.3	69.5	74.0	46.5	60.4	47.0	61.4	57.9	6.4
	4-16-16	SmoothQuant	72.0	45.6	71.4	78.4	46.8	72.9	51.0	68.4	63.3	6.1
		OmniQuant	70.2	43.1	—	78.4	—	—	—	67.8	—	—
		QuIP	73.3	44.9	—	79	—	—	—	69.7	—	—
		AQLM	72.2	43.9	—	78.6	—	—	—	70.4	—	—
		QuIP#	73.9	45.5	—	78.9	—	—	—	69.9	—	—
		GPTQ	73.2 ±1.4	48.4 ±1.0	76.9 ±0.6	78.2 ±0.3	48.5 ±0.7	71.2 ±2.5	53.1 ±1.0	68.3 ±1.4	64.7 ±0.9	5.6 ±0.01
		QuaRot*	75.3 ±1.3	51.2 ±1.4	78.6 ±2.4	79.5 ±0.6	49.1 ±0.5	76.6 ±0.5	55.3 ±1.2	71.2 ±0.9	67.1 ±0.5	5.5 ±0.04
		QuaRot	76.3 ±0.6	52.5 ±1.1	80.7 ±0.5	80.4 ±0.2	50.3 ±0.5	78.8 ±0.2	55.6 ±0.7	72.0 ±0.6	68.3 ±0.2	5.0 ±0.01
		SpinQuant*	76.3 ±0.8	51.0 ±1.4	77.8 ±1.7	80.0 ±0.4	49.3 ±0.8	78.8 ±0.2	55.1 ±1.2	71.0 ±0.5	67.4 ±0.6	4.9 ±0.01
		SpinQuant	77.0 ±0.5	51.9 ±0.5	80.6 ±0.6	80.4 ±0.2	50.0 ±0.4	78.9 ±0.2	56.6 ±0.7	72.7 ±0.6	68.5 ±0.1	5.0 ±0.00
	4-4-16	RTN	26.0	26.0	40.6	49.7	38.7	26.0	25.4	49.9	35.3	7,216.7
		SmoothQuant	45.2	27.1	55.4	62.5	40.5	44.3	33.4	50.8	44.9	34.5
		GPTQ	26.6 ±0.5	24.7 ±1.3	37.9 ±0.2	49.3 ±0.6	39.2 ±0.4	26.2 ±0.3	27.7 ±1.6	50.3 ±1.3	35.3 ±0.5	5,245.3
		QuaRot*	72.5 ±1.3	48.6 ±1.2	76.1 ±2.1	77.9 ±0.4	47.9 ±0.5	73.8 ±0.3	52.6 ±1.6	68.7 ±1.0	64.8 ±0.6	6.1 ±0.06
		QuaRot	74.4 ±1.1	49.3 ±1.3	78.8 ±1.0	79.3 ±0.6	49.0 ±0.6	76.9 ±0.3	54.7 ±1.6	71.1 ±0.9	66.7 ±0.3	5.4 ±0.01
		SpinQuant*	73.7 ±1.1	49.5 ±1.7	77.1 ±1.9	78.4 ±0.7	48.4 ±1.2	76.1 ±0.5	54.4 ±0.9	69.1 ±0.7	65.8 ±0.5	5.4 ±0.01
		SpinQuant	75.9 ±0.8	50.8 ±0.8	78.1 ±0.8	79.5 ±0.1	49.4 ±0.5	77.5 ±0.2	55.2 ±1.3	70.8 ±0.9	67.2 ±0.3	5.2 ±0.01
	4-4-4	RTN	26.1	24.3	40.3	48.7	39.6	25.8	29.2	49.6	35.4	7,428.8
		SmoothQuant	36.9	24.8	49.4	57.2	39.6	33.3	31.2	51.7	40.5	56.6
		GPTQ	26.6 ±0.5	24.1 ±1.4	37.9 ±0.2	48.8 ±0.8	38.9 ±0.5	26.1 ±0.2	29.3 ±2.0	50.1 ±1.4	35.2 ±0.3	5,237.1
		QuaRot*	72.4 ±1.7	47.9 ±1.2	75.1 ±1.7	77.9 ±0.6	47.4 ±0.5	73.4 ±0.4	53.5 ±1.6	67.8 ±0.8	64.4 ±0.6	6.2 ±0.07
		QuaRot	74.0 ±0.7	48.8 ±1.0	78.7 ±0.7	78.8 ±0.6	48.7 ±0.2	76.4 ±0.2	53.6 ±1.6	70.7 ±0.4	66.2 ±0.4	5.4 ±0.01
		SpinQuant*	73.8 ±1.4	48.8 ±1.1	75.4 ±3.6	78.3 ±0.5	48.3 ±1.1	76.1 ±0.3	53.4 ±1.2	69.9 ±0.6	65.5 ±0.5	5.4 ±0.01
		SpinQuant	75.7 ±1.0	50.5 ±1.0	79.3 ±1.0	79.5 ±0.2	49.1 ±0.4	77.1 ±0.1	53.8 ±1.1	69.9 ±0.5	66.9 ±0.1	5.3 ±0.00
70B	16-16-16	Full Precision	80.2	60.5	85.1	82.8	50.8	84.3	59.0	80.6	72.9	3.3
		RTN	77.7	54.6	82.7	81.5	47.7	78.4	56.2	75.2	69.2	4.6
	4-16-16	SmoothQuant	79.7	56.7	81.3	81.4	50.2	81.4	54.8	76.4	70.2	4.1
		OMNIQ	77.9	49.8	—	80.7	—	—	—	75.8	—	—
		QuIP	74.3	47	—	80.3	—	—	—	76	—	—
		AQLM	78.1	51	—	81.4	—	—	—	76.9	—	—
		QuIP#	78.1	50.6	—	81.4	—	—	—	77.1	—	—
		GPTQ	80.1 ±0.2	58.6 ±0.6	83.6 ±0.7	82.4 ±0.3	50.8 ±0.3	82.9 ±0.1	58.1 ±0.6	78.8 ±0.4	71.9 ±0.2	3.9 ±0.02
		QuaRot*	79.5 ±0.7	58.6 ±1.0	84.3 ±0.5	82.3 ±0.4	49.6 ±0.7	82.4 ±0.3	59.5 ±0.9	78.1 ±0.6	71.8 ±0.4	3.7 ±0.01
		QuaRot	79.4 ±0.7	59.4 ±1.0	84.7 ±0.4	82.5 ±0.5	50.3 ±0.4	83.4 ±0.2	58.7 ±0.3	79.3 ±0.2	72.2 ±0.1	3.5 ±0.00
		SpinQuant*	79.8 ±0.6	59.0 ±1.0	84.0 ±0.7	82.3 ±0.4	50.3 ±0.6	83.7 ±0.2	59.6 ±1.8	78.5 ±0.6	72.2 ±0.2	3.5 ±0.00
		SpinQuant	79.7 ±0.6	59.8 ±0.5	84.9 ±0.2	82.5 ±0.3	50.4 ±0.2	83.6 ±0.3	59.9 ±0.4	79.6 ±0.5	72.6 ±0.2	3.5 ±0.00
	4-4-16	RTN	26.0	23.2	43.5	48.9	37.0	26.0	25.6	50.5	35.1	2e5
		SmoothQuant	9.5	71.7	29.0	66.6	73.1	45.1	67.4	39.4	64.6	57.1
		GPTQ	25.3 ±0.5	25.8 ±0.6	45.7 ±1.1	50.1 ±0.3	36.4 ±0.6	25.8 ±0.4	24.6 ±2.6	50.0 ±0.8	35.5 ±0.4	2e6
		QuaRot*	77.9 ±0.8	55.8 ±0.4	81.5 ±0.9	80.6 ±0.2	48.5 ±0.7	80.3 ±0.4	57.1 ±1.1	75.9 ±1.1	69.7 ±0.5	4.2 ±0.01
		QuaRot	78.1 ±0.6	56.1 ±0.5	83.0 ±0.5	81.0 ±0.4	49.7 ±0.7	81.9 ±0.2	57.1 ±1.6	76.3 ±0.4	70.4 ±0.3	3.9 ±0.01
		SpinQuant*	79.2 ±0.8	57.0 ±1.1	81.6 ±1.6	81.8 ±0.3	50.5 ±0.7	82.6 ±0.2	60.3 ±0.8	76.0 ±0.7	71.1 ±0.5	3.9 ±0.01
		SpinQuant	78.4 ±0.4	57.0 ±1.2	82.7 ±0.5	81.4 ±0.3	50.2 ±0.3	83.0 ±0.2	58.5 ±1.4	77.0 ±1.0	71.0 ±0.3	3.8 ±0.01
	4-4-4	RTN	25.5	24.5	43.2	50.2	36.7	26.6	24.2	49.3	35.0	2e5
		SmoothQuant	68.1	31.9	65.8	72.0	43.5	64.2	38.2	63.1	55.9	10.5
		GPTQ	26.1 ±1.0	25.2 ±1.4	45.7 ±1.7	49.5 ±1.2	36.8 ±0.5	26.0 ±0.3	25.4 ±2.4	50.2 ±1.1	35.6 ±0.4	1e6
		QuaRot*	77.8 ±0.6	55.1 ±0.8	80.8 ±0.5	80.3 ±0.8	48.7 ±0.1	80.2 ±0.4	57.8 ±1.0	75.6 ±1.0	69.5 ±0.4	4.2 ±0.01
		QuaRot	78.4 ±1.0	56.8 ±0.6	82.3 ±0.6	81.0 ±0.5	49.3 ±0.3	81.8 ±0.2	57.3 ±1.3	75.7 ±0.7	70.3 ±0.4	3.9 ±0.01
		SpinQuant*	78.7 ±0.8	56.9 ±0.3	81.2 ±0.9	81.1 ±0.5	49.6 ±0.7	82.5 ±0.3	58.5 ±1.7	75.9 ±1.2	70.5 ±0.4	3.9 ±0.01
		SpinQuant	78.3 ±0.3	57.6 ±0.8	82.1 ±0.9	81.7 ±0.4	50.1 ±0.4	82.9 ±0.2	59.8 ±1.6	77.3 ±0.7	71.2 ±0.4	3.8 ±0.00