Table 5: Complete omparison of the perplexity score on WikiText2 and averaged accuracy on Zero-shot Common Sense Reasoning tasks on **LLaMA-3**. 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 (↓)
8B	16-16-16	Full Precision	77.6	57.7	83.3	80.7	48.7	79.6	55.8	73.7	69.6	6.1
	4-16-16	RTN	74.7	49.0	73.0	77.0	47.6	76.6	53.4	71.4	65.4	7.8
		SmoothQuant	67.6	41.3	72.6	74.4	46.7	70.6	48.0	67.0	61.0	10.7
		LLM-QAT	77.1	53.0	82.4	79.0	48.1	76.6	54.4	71.3	67.7	7.1
		GPTQ	73.5 ±1.9	50.2 ±1.0	79.7 ±1.5	77.9 ±0.7	48.8 ±0.3	76.4 ±0.2	52.6 ±0.9	72.6 ±1.0	66.5 ±0.6	7.2 ±0.02
		QuaRot*	73.8 ±1.4	51.4 ±0.9	80.0 ±1.7	77.7 ±1.1	47.8 ±0.8	75.9 ±0.4	52.1 ±0.8	71.1 ±0.7	66.2 ±0.6	7.5 ±0.02
		QuaRot	77.0 ±0.8	55.2 ±0.8	82.2 ±0.6	79.5 ±0.4	48.6 ±0.5	78.3 ±0.3	54.4 ±0.9	72.1 ±0.8	68.4 ±0.2	6.4 ±0.01
		`SpinQuant*`	76.5 ±1.1	54.8 ±1.5	79.8 ±2.1	79.0 ±0.8	47.6 ±1.0	78.0 ±0.3	53.3 ±1.1	71.6 ±0.8	67.6 ±0.6	6.5 ±0.01
		`SpinQuant`	77.6 ±0.7	55.5 ±0.9	81.4 ±1.3	79.4 ±0.2	48.2 ±0.4	78.4 ±0.2	55.1 ±1.3	72.4 ±1.0	68.5 ±0.2	6.4 ±0.01
	4-4-16	RTN	31.8	27.6	47.2	53.8	39.7	30.8	28.2	48.9	38.5	923.9
		SmoothQuant	36.3	26.3	50.6	54.1	40.3	31.4	30.6	52.9	40.3	867.5
		LLM-QAT	44.1	29.7	58.0	61.5	42.1	39.9	33.0	51.3	44.9	42.9
		GPTQ	31.4 ±0.9	24.7 ±1.4	42.5 ±1.3	52.7 ±1.0	39.1 ±0.9	27.8 ±0.3	27.3 ±2.5	50.7 ±1.1	37.0 ±0.7	955.9
		QuaRot*	66.0 ±1.2	42.5 ±1.0	70.5 ±2.0	72.5 ±0.6	45.4 ±0.9	68.6 ±0.9	46.7 ±2.1	63.5 ±1.7	59.5 ±0.6	10.4 ±0.26
		QuaRot	72.4 ±1.1	48.0 ±1.1	75.8 ±1.4	75.9 ±0.6	47.1 ±0.8	73.7 ±0.6	51.0 ±1.8	66.7 ±1.2	63.8 ±0.5	7.9 ±0.04
		`SpinQuant*`	74.1 ±1.6	49.7 ±1.7	75.8 ±3.2	77.0 ±0.6	46.4 ±0.9	74.7 ±0.4	52.0 ±1.9	67.1 ±1.0	64.6 ±0.8	7.7 ±0.05
		`SpinQuant`	75.0 ±1.0	50.9 ±1.2	78.9 ±0.6	77.5 ±0.7	47.2 ±0.6	75.9 ±0.4	52.9 ±1.6	68.5 ±1.0	65.8 ±0.2	7.1 ±0.02
	4-4-4	RTN	31.9	26.1	46.2	52.3	39.9	29.9	28.6	51.0	38.2	1,118.5
		SmoothQuant	33.5	25.1	49.6	53.1	40.3	28.8	29.6	49.6	38.7	1,530.5
		LLM-QAT	40.5	26.6	52.7	59.9	42.3	37.5	33.6	52.7	43.2	52.5
		GPTQ	31.0 ±0.9	24.9 ±1.0	41.9 ±1.1	52.8 ±0.6	38.4 ±0.3	27.9 ±0.4	28.2 ±2.0	51.4 ±1.1	37.1 ±0.3	1,071.7
		QuaRot*	65.9 ±3.0	41.3 ±2.2	69.5 ±2.3	71.9 ±0.9	44.8 ±1.1	67.2 ±1.6	46.5 ±1.8	61.9 ±1.5	58.6 ±0.8	10.9 ±0.26
		QuaRot	71.6 ±0.9	48.0 ±1.2	74.9 ±1.8	75.1 ±0.5	46.8 ±0.9	73.1 ±0.7	50.4 ±1.0	66.1 ±1.4	63.3 ±0.3	8.0 ±0.05
		`SpinQuant*`	72.6 ±1.4	49.5 ±2.3	74.8 ±6.3	76.6 ±0.8	46.4 ±0.4	74.3 ±1.0	50.6 ±3.1	67.9 ±1.0	64.1 ±1.7	7.8 ±0.05
		`SpinQuant`	74.4 ±1.3	50.4 ±1.7	77.7 ±1.6	76.9 ±0.6	47.2 ±0.5	75.5 ±0.2	52.0 ±1.1	67.2 ±1.4	65.2 ±0.6	7.3 ±0.02
70B	16-16-16	Full Precision	80.6	64.5	87.4	83.7	51.7	85.3	62.0	80.5	74.5	2.8
	4-16-16	RTN	27.3	27.2	37.8	51.0	39.1	25.6	26.2	49.8	35.5	1e5
		SmoothQuant	76.7	45.5	80.6	81.2	48.7	81.1	46.2	75.5	66.9	12.0
		GPTQ	28.2 ± 1.3	24.7 ± 1.6	37.9 ± 0.1	50.7 ± 0.8	39.0 ± 0.6	26.8 ± 1.7	27.7 ± 3.6	50.5 ± 0.6	35.7 ± 0.6	1e5
		QuaRot*	65.9 ± 1.4	44.3 ± 2.7	67.0 ± 4.9	75.2 ± 2.0	44.1 ± 2.1	59.8 ± 8.6	42.5 ± 3.8	58.5 ± 2.9	57.2 ± 2.7	41.6 ± 16.76
		QuaRot	74.4 ± 0.7	58.6 ± 2.6	86.4 ± 0.5	83.8 ± 0.4	51.9 ± 0.4	83.7 ± 0.1	47.7 ± 2.2	76.1 ± 0.6	70.3 ± 0.7	7.9 ± 0.21
		`SpinQuant*`	78.5 ± 2.2	57.9 ± 1.5	84.5 ± 1.0	82.3 ± 0.6	50.3 ± 0.6	82.6 ± 0.4	57.4 ± 5.9	77.5 ± 1.9	71.4 ± 1.5	3.9 ± 0.47
		`SpinQuant`	78.0 ± 2.7	59.1 ± 0.9	85.2 ± 1.1	82.8 ± 1.0	50.6 ± 0.6	83.5 ± 0.2	54.8 ± 6.6	78.5 ± 1.6	71.6 ± 1.1	4.8 ± 1.97
	4-4-16	RTN	27.6	27.0	38.2	50.1	38.5	26.0	28.4	49.1	35.6	1e5
		SmoothQuant	59.5	35.7	62.4	70.3	44.3	61.7	44.2	63.9	55.3	18.0
		GPTQ	27.0 ± 0.4	26.1 ± 1.5	39.1 ± 1.2	50.4 ± 0.4	38.9 ± 0.4	25.7 ± 0.2	25.7 ± 1.7	49.6 ± 0.9	35.3 ± 0.2	1e5
		QuaRot*	40.8 ± 4.8	26.2 ± 2.8	52.4 ± 3.8	58.4 ± 3.5	40.0 ± 0.9	33.8 ± 3.9	30.0 ± 3.6	50.2 ± 1.7	41.5 ± 2.5	91.2 ± 24.05
		QuaRot	72.4 ± 1.5	52.2 ± 1.6	78.5 ± 2.4	78.9 ± 0.8	49.0 ± 1.1	78.5 ± 0.9	45.2 ± 1.9	68.2 ± 3.0	65.4 ± 1.3	20.4 ± 3.23
		`SpinQuant*`	77.2 ± 0.9	55.9 ± 1.1	81.7 ± 1.7	80.9 ± 0.6	49.0 ± 0.5	80.9 ± 0.4	58.7 ± 1.9	76.2 ± 0.8	70.1 ± 0.4	4.1 ± 0.02
		`SpinQuant`	76.7 ± 1.9	55.6 ± 2.1	82.3 ± 1.3	80.6 ± 1.1	49.8 ± 0.9	81.0 ± 1.6	55.4 ± 6.3	74.5 ± 3.9	69.5 ± 2.1	5.5 ± 2.56
	4-4-4	RTN	27.0	24.1	38.5	50.4	38.8	25.8	25.2	51.8	35.2	1e5
		SmoothQuant	55.0	34.9	62.2	66.8	43.1	59.4	39.8	58.0	52.4	22.1
		GPTQ	27.1 ± 0.3	24.8 ± 1.7	38.8 ± 0.7	50.8 ± 0.6	39.0 ± 0.4	25.5 ± 0.2	24.8 ± 2.9	49.8 ± 0.6	35.1 ± 0.2	1e5
		QuaRot*	40.5 ± 4.9	25.7 ± 3.7	50.9 ± 4.5	57.7 ± 3.3	39.9 ± 0.8	33.9 ± 3.9	31.0 ± 2.6	51.0 ± 1.6	41.3 ± 2.6	92.4 ± 24.18
		QuaRot	72.3 ± 1.9	51.6 ± 1.0	77.5 ± 2.0	78.9 ± 1.1	49.0 ± 1.0	78.2 ± 0.9	45.5 ± 1.7	67.8 ± 2.6	65.1 ± 1.1	20.2 ± 3.12
		`SpinQuant*`	77.3 ± 1.0	56.0 ± 1.1	81.8 ± 1.7	80.8 ± 0.4	49.3 ± 0.5	80.9 ± 0.3	58.7 ± 0.7	76.4 ± 0.3	70.1 ± 0.5	4.1 ± 0.01
		`SpinQuant`	76.8 ± 2.1	55.8 ± 2.6	82.2 ± 1.7	81.0 ± 0.8	49.5 ± 1.0	81.2 ± 1.3	53.8 ± 6.3	74.2 ± 3.8	69.3 ± 2.2	5.5 ± 2.59

A Appendix / supplemental material

A.1 Complete results of main result table

In Tables 5 and 6, we show the complete results of Table 1. We compare the accuracy on eight zero-shot commonsense reasoning tasks including ARC-easy, ARC-challenge [9], BoolQ [8], PIQA [6], SIQA [34], HellaSwag [45], OBQA [28], and WinoGrande [33]. We compare our results with previous works including SmoothQuant[43], LLM-QAT[25], GPTQ [14], OmniQuant [35], AQLM [12], ATOM [47], AWQ [23], QuIP [7], QuIP# [41], and QuaRot [5].

A.2 *Cayley* optimization choice

In Table 7, we evaluate the impact of varying the number of samples and iterations used in *Cayley* optimization. Given the relatively small number of trainable parameters in the rotation matrix compared to the original weight parameters, and considering it as a constraint optimization, we only need a minimal amount of calibration data and iterations to enhance the rotation for improved quantization. The findings indicate that rotation optimization is resilient to modifications in the number of samples. Even though we used 800 samples in our experiments, reducing this to 128 samples does not lead to a significant change in the perplexity. Furthermore, we examined the optimal number of iterations and found that the wiki perplexity ceases to decrease and stabilizes at 100 iterations. Consequently, we chose to use 100 iterations in all our experiments.