

A FULL RESULTS

Table 6: Comparison of the WikiText-2 perplexity (\downarrow) results for LLaMA and Mistral. The 4-4-4, 4-4-8, 4-4-16, 4-8-4, 4-8-8, 4-8-16 represent W4A4KV4, W4A4KV8, W4A4KV16, W4A4KV4, W4A4KV8 and W4A4KV16, respectively. We show the failed GPTQ experiments using NaN and the perplexity results >100 by Inf. Results for RTN, GPTQ and QuaRot (Ashkboos et al., 2024b) are obtained using QuaRot publicly released codebase. Results for SpinQuant (Liu et al., 2024) are obtained using their publicly released codebase.

Method	LLaMA2-7B			LLaMA2-13B			LLaMA3-8B			Mistral-7B		
Baseline	5.47			4.88			6.13			5.25		
	4-4-4	4-4-8	4-4-16	4-4-4	4-4-8	4-4-16	4-4-4	4-4-8	4-4-16	4-4-4	4-4-8	4-4-16
RTN	NaN	NaN	NaN	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
+QuaRot	9.04	8.70	8.69	6.31	6.22	6.23	11.06	10.5	10.47	6.26	6.19	6.19
+SpinQuant	6.20	6.15	6.17	5.51	5.47	5.40	8.00	7.80	7.82	5.60	5.58	5.58
+DuaRot	5.88	5.85	5.86	5.27	5.25	5.19	7.49	7.40	7.41	5.48	5.42	5.43
GPTQ	NaN	NaN	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
+QuaRot	6.27	6.19	6.20	5.51	5.46	5.47	8.20	8.03	8.02	5.75	5.71	5.71
+SpinQuant	5.94	5.88	5.91	5.25	5.21	5.21	7.34	7.24	7.25	5.62	5.58	5.57
+DuaRot	5.79	5.75	5.74	5.13	5.11	5.12	7.22	7.17	7.13	5.55	5.50	5.50
	4-8-4	4-8-8	4-8-16	4-8-4	4-8-8	4-8-16	4-8-4	4-8-8	4-8-16	4-8-4	4-8-8	4-8-16
RTN	7.92	7.37	7.37	5.79	5.47	5.47	19.38	10.89	10.88	12.70	11.21	10.65
+QuaRot	6.89	6.75	6.75	5.51	5.46	5.46	7.81	7.64	7.64	5.70	5.65	5.65
+SpinQuant	5.56	5.51	5.52	4.97	4.94	4.93	6.79	6.71	6.70	5.28	5.24	5.25
+DuaRot	5.51	5.49	5.48	4.94	4.91	4.91	6.76	6.67	6.67	5.24	5.22	5.24
GPTQ	7.13	6.55	6.60	5.40	5.21	5.21	Inf	Inf	Inf	6.17	6.01	6.01
+QuaRot	5.66	5.61	5.61	5.04	5.00	5.01	6.62	6.52	6.51	5.37	5.34	5.34
+SpinQuant	5.65	5.61	5.62	5.04	5.00	5.00	6.63	6.54	6.55	5.37	5.34	5.34
+DuaRot	5.65	5.61	5.61	5.03	5.00	5.00	6.62	6.53	6.51	5.37	5.34	5.34

Table 7: Zero-shot accuracy of LLaMA2-7B with RTN and GPTQ on PIQA, WinoGrande (WG), HellaSwag (HS), ARC-challenge (ARC-c), ARC-easy (ARC-e), OBQA, BoolQ and SIQA.

W-A-KV	Method	PIQA	WG	HS	ARC-c	ARC-e	OBQA	BoolQ	SIQA	Avg.
16-16-16	Baseline	79.11	69.22	76.00	46.33	74.54	44.20	76.97	46.11	64.06
4-4-4	RTN	50.60	51.14	26.08	26.96	27.02	25.80	42.20	33.93	35.47
	+QuaRot	70.29	60.77	63.87	36.86	59.55	33.40	66.33	40.02	53.89
	+SpinQuant	75.52	63.85	70.91	39.76	65.07	40.60	72.08	43.91	58.96
	+DuaRot	74.97	65.67	71.43	41.30	69.87	39.60	72.54	42.94	59.79
4-4-8	RTN	50.65	50.51	25.84	24.66	26.81	25.00	42.11	34.08	34.96
	+QuaRot	69.97	61.09	65.30	35.67	60.61	35.60	69.72	40.58	54.82
	+SpinQuant	75.79	63.30	70.92	38.57	61.74	39.40	71.68	41.66	57.88
	+DuaRot	76.50	65.27	72.36	41.47	66.84	39.80	71.04	42.63	59.49
4-4-16	RTN	50.22	48.46	25.79	24.74	25.67	27.40	41.87	34.03	34.77
	+QuaRot	70.02	60.06	64.80	35.92	59.22	33.00	68.50	40.07	53.95
	+SpinQuant	75.84	63.06	70.90	39.85	64.39	39.60	70.98	41.04	58.21
	+DuaRot	75.79	65.35	71.55	41.38	66.79	37.60	74.34	43.35	59.52
4-8-4	RTN	74.59	60.93	67.87	39.33	66.67	35.60	66.21	42.58	56.72
	+QuaRot	76.01	64.96	70.77	40.70	67.38	37.20	73.33	42.37	59.09
	+SpinQuant	77.31	65.98	74.25	42.32	70.45	42.20	73.61	42.53	61.08
	+DuaRot	76.71	66.54	74.29	42.83	70.83	42.60	73.09	43.65	61.32
4-8-8	RTN	75.57	65.27	70.72	43.09	67.93	40.60	67.92	43.14	59.28
	+QuaRot	75.79	65.90	71.34	41.04	68.48	37.80	73.39	42.22	59.50
	+SpinQuant	77.53	66.61	74.11	42.66	70.08	42.60	74.13	42.53	61.28
	+DuaRot	77.91	66.77	74.05	44.71	72.35	43.00	74.10	42.17	61.88
4-8-16	RTN	75.95	64.72	70.58	42.75	68.31	42.00	67.52	42.68	59.31
	+QuaRot	76.33	65.51	71.34	40.70	68.39	37.60	73.46	42.32	59.46
	+SpinQuant	77.58	67.25	74.44	43.00	69.49	41.60	74.89	42.32	61.32
	+DuaRot	77.86	67.01	74.75	45.22	72.22	42.80	74.71	44.58	62.39
4-4-4	GPTQ	49.08	48.30	25.77	25.77	27.95	25.40	49.48	33.32	35.63
	+QuaRot	77.15	65.82	72.76	41.47	69.44	37.80	71.87	43.35	59.96
	+SpinQuant	76.66	65.98	72.78	42.06	70.92	38.40	73.82	44.73	60.67
	+DuaRot	76.50	67.09	72.69	42.15	71.51	42.00	72.60	44.68	61.15
4-4-8	GPTQ	50.71	49.25	26.75	26.54	27.57	28.00	47.22	33.88	36.24
	+QuaRot	76.39	65.67	72.88	41.81	69.61	39.60	73.18	43.55	60.34
	+SpinQuant	77.09	67.48	73.39	43.00	69.87	40.40	75.20	43.81	61.28
	+DuaRot	76.22	65.11	72.90	43.60	71.09	41.80	75.08	44.47	61.28
4-4-16	GPTQ	50.27	48.15	26.26	26.96	27.23	25.60	47.68	34.90	35.88
	+QuaRot	77.31	65.43	72.95	41.55	70.03	39.00	73.46	43.91	60.46
	+SpinQuant	75.24	66.14	72.82	40.44	68.77	39.20	74.53	43.76	60.11
	+DuaRot	76.82	66.30	72.77	42.66	72.39	41.20	72.17	45.29	61.20
4-8-4	GPTQ	76.82	63.61	71.54	40.53	68.14	43.00	70.80	42.99	59.68
	+QuaRot	78.56	68.67	75.08	43.77	73.11	43.20	75.87	45.29	62.94
	+SpinQuant	78.07	69.22	74.66	45.65	73.57	43.80	75.02	45.14	63.14
	+DuaRot	78.78	68.51	74.69	43.09	72.77	43.20	76.02	45.29	62.79
4-8-8	GPTQ	77.91	66.06	73.40	42.15	70.66	42.20	72.51	43.71	61.08
	+QuaRot	78.94	69.46	75.09	43.52	73.53	43.00	76.21	45.14	63.11
	+SpinQuant	77.31	67.32	74.92	42.92	72.35	43.00	74.62	44.83	62.16
	+DuaRot	78.18	68.75	74.50	44.11	73.23	43.00	76.64	45.80	63.03
4-8-16	GPTQ	77.20	66.85	73.33	42.32	69.95	42.40	73.43	43.76	61.15
	+QuaRot	78.84	68.98	75.12	43.94	73.53	42.80	76.18	45.19	63.07
	+SpinQuant	77.86	68.03	74.85	43.43	72.81	43.20	74.16	44.78	62.39
	+DuaRot	77.97	68.51	74.68	43.94	73.15	43.20	74.22	45.04	62.59

Table 8: Zero-shot accuracy of LLaMA2-13B with RTN and GPTQ on PIQA, WinoGrande (WG), HellaSwag (HS), ARC-challenge (ARC-c), ARC-easy (ARC-e), OBQA, BoolQ and SIQA.

W-A-KV	Method	PIQA	WG	HS	ARC-c	ARC-e	OBQA	BoolQ	SIQA	Avg.
16-16-16	Baseline	80.52	72.22	79.38	48.98	77.53	45.20	80.09	47.34	66.41
4-4-4	RTN	47.99	50.36	26.55	27.99	26.35	25.00	39.33	35.11	34.84
	+QuaRot	76.66	66.30	72.19	43.77	70.08	39.60	73.70	42.32	60.58
	+SpinQuant	77.97	65.82	75.39	45.82	74.41	42.80	75.78	45.39	62.92
	+DuaRot	78.13	69.06	76.55	46.33	73.99	41.80	72.08	45.60	62.94
4-4-8	RTN	49.13	51.22	26.35	25.68	26.39	25.20	38.35	35.36	34.71
	+QuaRot	77.75	65.67	72.71	44.37	71.25	39.60	74.77	41.45	60.95
	+SpinQuant	78.18	67.32	75.41	46.08	74.33	42.80	77.77	47.49	63.67
	+DuaRot	78.56	68.90	76.82	45.82	72.85	43.80	78.01	45.60	63.80
4-4-16	RTN	50.92	49.88	26.01	29.01	26.05	25.40	38.90	34.60	35.10
	+QuaRot	76.55	66.22	72.86	44.37	70.33	40.60	74.62	41.91	60.93
	+SpinQuant	78.13	69.14	76.08	46.33	75.08	43.40	77.65	46.06	63.98
	+DuaRot	78.35	67.72	76.52	45.56	73.53	43.20	75.14	44.68	63.09
4-8-4	RTN	77.86	67.01	75.64	45.90	74.20	40.80	76.64	45.75	62.97
	+QuaRot	79.11	69.53	75.84	46.50	74.28	42.20	78.35	44.47	63.79
	+SpinQuant	79.98	70.88	77.95	48.72	76.77	45.40	79.30	46.78	65.72
	+DuaRot	79.27	70.96	78.50	47.01	75.08	45.20	78.04	44.63	64.84
4-8-8	RTN	78.67	71.35	77.24	48.55	75.76	43.60	79.20	46.72	65.14
	+QuaRot	78.89	68.75	75.99	47.27	74.92	43.40	79.36	44.63	64.15
	+SpinQuant	80.20	70.80	78.37	49.57	77.02	44.80	79.11	46.32	65.77
	+DuaRot	80.03	71.03	78.71	47.78	76.52	44.40	77.49	44.78	65.09
4-8-16	RTN	79.00	71.03	77.37	48.21	75.29	43.60	79.05	46.62	65.02
	+QuaRot	78.89	68.90	76.00	46.84	74.92	43.40	78.87	44.68	64.06
	+SpinQuant	80.03	69.06	78.45	49.32	77.23	44.60	79.05	46.52	65.53
	+DuaRot	79.11	69.61	78.93	47.61	76.05	43.00	78.26	45.04	64.70
4-4-4	GPTQ	49.95	49.72	25.86	27.39	26.81	23.40	40.06	34.95	34.77
	+QuaRot	77.75	69.93	75.82	45.65	73.65	43.20	76.73	45.19	63.49
	+SpinQuant	78.78	70.24	76.63	47.35	75.93	44.40	75.81	46.11	64.41
	+DuaRot	79.22	70.96	77.35	46.08	76.01	43.40	78.29	45.70	64.63
4-4-8	GPTQ	48.37	51.30	25.85	26.79	26.85	23.60	39.85	34.54	34.64
	+QuaRot	77.53	69.61	76.13	46.59	74.62	44.60	76.91	45.60	63.95
	+SpinQuant	78.73	69.69	77.48	48.04	76.52	45.60	77.61	45.96	64.95
	+DuaRot	79.43	69.77	77.40	47.01	75.13	42.60	77.95	45.75	64.38
4-4-16	GPTQ	47.39	47.99	26.08	25.68	26.94	24.40	39.39	33.88	33.97
	+QuaRot	78.51	69.77	75.55	45.65	74.49	42.00	77.09	45.19	63.53
	+SpinQuant	78.45	68.59	77.26	47.44	75.38	44.40	77.77	47.29	64.57
	+DuaRot	80.03	68.27	76.67	47.70	75.08	42.20	78.01	46.83	64.35
4-8-4	GPTQ	77.09	70.32	75.63	45.82	73.65	42.60	77.09	46.06	63.53
	+QuaRot	79.11	71.74	78.34	46.76	76.77	44.80	79.20	46.32	65.38
	+SpinQuant	80.09	70.96	78.15	47.35	76.94	44.60	79.60	46.62	65.54
	+DuaRot	79.76	71.51	78.14	50.17	77.65	45.00	79.51	46.98	66.09
4-8-8	GPTQ	79.11	70.24	76.71	46.33	75.38	43.80	76.97	45.70	64.28
	+QuaRot	79.82	71.43	78.42	47.87	76.81	44.40	79.17	46.72	65.58
	+SpinQuant	80.03	71.35	78.33	48.46	76.39	43.40	79.08	46.37	65.43
	+DuaRot	79.71	73.01	78.28	49.74	77.44	46.00	79.60	46.93	66.34
4-8-16	GPTQ	79.16	70.48	76.60	46.93	75.63	43.20	76.67	45.75	64.30
	+QuaRot	79.87	71.35	78.44	47.95	76.81	44.20	79.20	46.52	65.54
	+SpinQuant	80.36	71.51	78.51	49.15	77.40	44.20	79.54	46.72	65.92
	+DuaRot	79.65	70.96	78.49	49.23	76.26	44.20	78.69	47.34	65.60

Table 9: Zero-shot accuracy of LLaMA3-8B with RTN and GPTQ on PIQA, WinoGrande (WG), HellaSwag (HS), ARC-challenge (ARC-c), ARC-easy (ARC-e), OBQA, BoolQ and SIQA.

W-A-KV	Method	PIQA	WG	HS	ARC-c	ARC-e	OBQA	BoolQ	SIQA	Avg.
16-16-16	Baseline	80.79	72.53	79.15	53.41	77.74	45.00	81.62	47.08	67.17
4-4-4	RTN	50.71	48.86	27.06	24.66	28.45	27.60	47.22	34.60	36.15
	+QuaRot	70.24	63.69	68.70	38.05	60.65	35.20	69.14	41.30	55.87
	+SpinQuant	75.90	66.30	74.16	44.54	70.58	42.40	72.60	44.78	61.41
	+DuaRot	78.40	66.54	76.36	48.55	74.45	43.00	75.38	45.39	63.51
4-4-8	RTN	52.23	51.46	27.79	24.83	27.48	25.60	47.16	33.06	36.20
	+QuaRot	71.60	63.85	69.38	36.69	59.55	37.60	68.96	42.43	56.26
	+SpinQuant	77.04	68.98	75.00	44.03	70.12	39.60	75.11	43.96	61.73
	+DuaRot	78.84	70.09	76.70	48.98	77.06	42.40	77.46	46.57	64.76
4-4-16	RTN	51.90	48.78	28.06	22.70	28.79	23.80	48.20	34.08	35.79
	+QuaRot	72.96	63.22	69.36	37.20	60.86	35.40	71.10	42.32	56.55
	+SpinQuant	77.75	68.59	74.92	43.86	72.05	41.20	74.92	45.09	62.30
	+DuaRot	77.69	67.72	75.85	47.78	73.99	42.60	77.34	45.60	63.57
4-8-4	RTN	60.39	55.72	52.24	30.38	48.27	31.00	62.29	40.63	47.62
	+QuaRot	77.53	71.11	75.67	43.77	72.22	40.60	79.33	44.47	63.09
	+SpinQuant	78.78	73.01	77.79	49.83	72.52	41.60	75.84	44.83	64.28
	+DuaRot	79.49	71.74	77.87	51.71	78.07	44.00	78.23	44.98	65.76
4-8-8	RTN	74.97	71.19	72.45	42.58	67.34	40.40	74.37	43.86	60.90
	+QuaRot	77.97	72.06	76.16	45.48	72.31	41.40	79.82	45.34	63.82
	+SpinQuant	79.54	71.59	78.10	48.55	72.98	42.80	75.66	45.39	64.33
	+DuaRot	79.49	73.48	78.24	50.26	78.87	44.60	80.03	46.21	66.40
4-8-16	RTN	75.35	70.17	72.30	42.41	68.06	41.00	74.83	43.76	60.99
	+QuaRot	77.91	71.74	76.13	45.56	72.47	41.80	79.51	45.04	63.77
	+SpinQuant	79.82	72.93	78.01	49.49	75.46	42.40	78.78	46.01	65.36
	+DuaRot	79.60	73.01	77.89	51.62	75.42	43.20	76.94	46.42	65.51
4-4-4	GPTQ	52.34	48.62	25.73	22.87	27.40	30.20	48.44	33.42	36.13
	+QuaRot	74.65	68.11	72.82	42.24	68.48	40.80	72.17	44.37	60.46
	+SpinQuant	77.58	68.43	75.23	47.78	75.21	42.40	76.24	44.78	63.46
	+DuaRot	78.13	69.77	76.38	47.78	75.17	43.20	78.13	44.63	64.15
4-4-8	GPTQ	50.76	48.54	26.75	25.85	24.62	26.80	48.47	33.32	35.64
	+QuaRot	75.90	65.75	73.39	44.88	70.29	40.20	71.53	43.09	60.63
	+SpinQuant	78.35	69.77	76.06	47.44	75.46	43.00	77.92	44.73	64.09
	+DuaRot	78.29	68.98	75.55	49.06	76.18	43.40	78.17	45.34	64.37
4-4-16	GPTQ	49.08	50.51	26.68	26.88	26.14	26.40	46.79	33.21	35.71
	+QuaRot	76.22	69.14	73.84	43.86	70.58	40.40	71.62	42.73	61.05
	+SpinQuant	78.94	68.43	74.95	47.53	75.13	41.80	77.58	45.04	63.68
	+DuaRot	78.18	69.22	76.34	47.61	75.84	42.40	78.01	46.06	64.21
4-8-4	GPTQ	54.35	52.01	38.32	24.06	36.11	27.20	53.03	34.49	39.95
	+QuaRot	78.94	73.72	77.22	50.85	77.48	43.80	79.08	46.57	65.96
	+SpinQuant	79.82	73.64	77.38	51.11	77.65	43.80	80.31	45.80	66.19
	+DuaRot	80.58	71.90	77.38	50.60	75.08	44.00	78.32	45.55	65.43
4-8-8	GPTQ	53.97	55.72	45.51	21.42	31.82	32.20	57.31	32.19	41.27
	+QuaRot	79.49	73.56	77.54	50.68	77.44	44.40	79.69	45.96	66.10
	+SpinQuant	79.98	73.48	78.01	51.62	77.78	45.40	80.18	47.19	66.71
	+DuaRot	80.36	72.77	78.09	50.60	75.76	44.20	80.43	46.06	66.03
4-8-16	GPTQ	55.50	56.12	47.11	21.59	31.86	33.20	56.91	33.83	42.02
	+QuaRot	79.60	73.64	77.73	50.77	77.57	44.60	79.88	46.21	66.25
	+SpinQuant	79.38	73.32	77.81	52.13	78.32	44.00	80.64	47.34	66.62
	+DuaRot	79.60	73.48	78.18	52.30	78.20	43.80	81.65	47.08	66.79

Table 10: Zero-shot accuracy of Misral-7B with RTN and GPTQ on PIQA, WinoGrande (WG), HellaSwag (HS), ARC-challenge (ARC-c), ARC-easy (ARC-e), OBQA, BoolQ and SIQA.

W-A-KV	Method	PIQA	WG	HS	ARC-c	ARC-e	OBQA	BoolQ	SIQA	Avg.
16-16-16	Baseline	81.88	73.95	80.98	54.95	80.18	44.40	83.49	46.83	68.33
4-4-4	RTN	53.21	49.88	26.96	25.51	26.89	23.60	39.88	34.54	35.06
	+QuaRot	79.33	67.96	75.79	46.76	74.12	40.60	78.26	43.76	63.32
	+SpinQuant	79.87	70.48	78.48	48.81	76.52	42.00	82.35	44.37	65.36
	+DuaRot	81.56	70.48	78.56	50.51	78.07	41.00	80.76	43.30	65.53
4-4-8	RTN	49.67	48.62	27.47	27.22	27.48	25.60	39.76	33.32	34.89
	+QuaRot	77.69	68.82	75.87	46.67	75.46	39.80	78.35	43.65	63.29
	+SpinQuant	81.23	70.88	78.43	50.00	77.69	40.60	79.88	45.14	65.48
	+DuaRot	81.12	72.30	79.01	52.39	79.67	45.20	81.41	44.98	67.01
4-4-16	RTN	52.45	50.59	27.57	28.07	27.86	24.20	39.42	34.34	35.56
	+QuaRot	79.00	68.11	75.82	46.76	73.95	42.40	78.53	44.06	63.58
	+SpinQuant	80.41	69.14	77.74	48.04	77.10	41.40	81.90	44.98	65.09
	+DuaRot	82.15	71.67	79.35	49.15	76.43	43.20	82.48	44.47	66.11
4-8-4	RTN	78.89	66.61	76.60	48.63	73.86	40.40	70.43	44.88	62.54
	+QuaRot	81.12	70.96	78.77	50.94	77.78	43.40	81.25	45.09	66.16
	+SpinQuant	81.28	72.61	79.92	52.30	79.08	43.00	82.42	46.01	67.08
	+DuaRot	81.18	72.06	79.80	54.27	79.76	44.00	81.96	45.24	67.28
4-8-8	RTN	79.92	68.90	77.36	48.63	75.21	43.80	72.02	44.68	63.82
	+QuaRot	81.28	72.14	78.92	50.77	78.37	44.80	81.87	44.88	66.63
	+SpinQuant	82.10	72.61	80.10	52.22	79.21	44.00	82.66	45.60	67.31
	+DuaRot	81.72	72.85	79.83	54.18	80.05	44.80	82.42	45.34	67.65
4-8-16	RTN	80.09	69.14	77.56	48.89	74.75	42.00	71.62	45.45	63.69
	+QuaRot	81.28	72.14	78.99	51.11	78.37	44.60	82.11	44.83	66.68
	+SpinQuant	81.66	72.30	79.80	52.05	78.58	44.40	82.42	46.42	67.20
	+DuaRot	81.94	73.01	80.20	51.71	79.12	43.40	81.68	45.50	67.07
4-4-4	GPTQ	53.32	48.78	27.38	26.45	31.06	27.60	42.78	34.14	36.44
	+QuaRot	79.92	68.59	78.13	50.43	76.35	41.20	80.37	45.91	65.11
	+SpinQuant	79.60	70.40	78.55	50.68	77.61	43.00	81.62	46.47	65.99
	+DuaRot	79.98	71.43	78.95	50.26	78.41	43.00	80.67	45.39	66.01
4-4-8	GPTQ	53.32	48.38	27.45	25.51	30.60	26.60	44.13	36.28	36.53
	+QuaRot	79.05	70.24	78.07	50.09	76.52	41.40	80.21	44.17	64.97
	+SpinQuant	79.71	71.51	78.67	51.37	78.41	43.00	81.31	44.98	66.12
	+DuaRot	80.58	72.85	78.69	51.45	78.03	43.80	81.28	46.11	66.60
4-4-16	GPTQ	53.97	48.54	27.55	24.74	31.73	23.40	45.47	33.73	36.14
	+QuaRot	79.82	69.38	78.13	49.15	76.56	41.80	80.64	44.78	65.03
	+SpinQuant	79.87	71.67	78.83	51.19	79.00	44.60	79.97	45.96	66.39
	+DuaRot	80.30	73.16	79.02	51.28	78.87	44.00	81.96	45.85	66.81
4-8-4	GPTQ	80.20	69.46	78.42	50.09	76.26	40.40	77.49	45.96	64.79
	+QuaRot	81.23	73.40	79.93	53.41	79.71	42.40	82.78	45.96	67.35
	+SpinQuant	81.61	72.93	80.14	51.79	79.76	44.20	83.06	46.16	67.46
	+DuaRot	81.72	74.35	80.15	53.50	80.18	43.20	82.87	46.42	67.80
4-8-8	GPTQ	80.63	72.69	79.15	50.60	78.03	42.20	78.17	46.32	65.97
	+QuaRot	81.45	73.95	80.05	52.82	80.22	42.60	83.03	46.37	67.56
	+SpinQuant	81.72	73.64	80.51	53.16	80.35	45.20	82.72	46.16	67.93
	+DuaRot	80.96	74.43	80.10	52.73	80.26	43.80	82.57	45.60	67.56
4-8-16	GPTQ	80.63	71.27	79.34	51.02	77.74	41.80	78.23	46.37	65.80
	+QuaRot	81.45	73.64	80.01	52.82	80.09	43.00	83.21	46.26	67.56
	+SpinQuant	81.72	73.32	80.42	53.24	79.84	44.60	82.84	45.91	67.74
	+DuaRot	81.23	72.85	80.22	52.99	79.76	44.60	83.73	46.47	67.73