#Bits	Method	PPL ↓			Accuracy (%) ↑								
		WikiText2	C4	Avg.	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	WinoG.	Avg.
Full Prec.	-	17.56	22.17	19.86	53.04	39.51	68.34	23.00	35.14	60.00	61.36	25.94	45.79
w2a16g128	RTN	2.27e+07	3.06e+07	2.66e+07	51.38	34.19	**52.61**	**18.80**	25.84	**47.74**	24.66	**21.33**	**34.57**
	GPTQ	1.30e+04	1.04e+04	1.17e+04	52.57	33.16	50.98	16.80	**25.94**	45.26	**27.86**	20.82	34.17
	AWQ	**1.02e+04**	**8.18e+03**	**9.18e+03**	48.93	**34.44**	51.03	15.60	25.62	38.84	26.30	20.48	32.65
w3a16g128	RTN	91.65	96.75	94.20	48.38	36.59	60.88	19.00	30.21	49.54	46.84	21.93	39.17
	GPTQ	**32.89**	**40.29**	**36.59**	51.93	37.15	61.53	**20.80**	**31.39**	**58.56**	**51.89**	22.53	**41.97**
	AWQ	54.20	55.94	55.07	50.36	**37.41**	**62.40**	17.00	31.28	52.23	51.56	**24.49**	40.84
w4a16g128	RTN	22.54	28.04	25.29	53.67	38.54	66.38	**23.00**	**34.18**	**62.05**	58.04	**25.85**	**45.21**
	GPTQ	**20.03**	**25.01**	**22.52**	52.01	**39.71**	65.56	22.00	33.99	56.36	**58.84**	24.74	44.15
	AWQ	21.42	26.19	23.81	52.25	38.02	**66.76**	22.80	34.07	58.96	58.54	25.77	44.65
w4a4	RTN	2.60e+03	2.22e+03	2.41e+03	50.91	33.73	52.61	17.40	26.40	43.73	30.77	18.77	34.29
	SmoothQuant	331.70	441.95	386.82	52.09	33.32	53.70	**18.20**	27.38	**44.46**	37.42	20.65	35.90
	OS+	**263.76**	**389.67**	**326.71**	52.49	**35.62**	55.39	14.60	**27.56**	43.46	41.46	**20.73**	**36.41**
	QuaRot	472.15	567.85	520.00	49.17	34.34	**56.37**	14.60	27.08	41.01	**43.01**	**20.73**	35.79
w6a6	RTN	22.84	27.45	25.14	49.41	38.28	65.07	20.00	33.02	58.23	56.73	**25.68**	43.30
	SmoothQuant	20.37	25.12	22.74	53.91	38.13	64.64	**22.80**	32.52	59.02	59.22	25.00	44.41
	OS+	**19.67**	**25.00**	**22.33**	51.54	**39.71**	**66.81**	21.20	32.88	**59.85**	**60.19**	24.32	44.56
	QuaRot	20.26	25.02	22.64	52.25	39.05	66.32	22.40	**33.06**	57.77	60.14	**25.68**	**44.58**
w8a8	RTN	17.75	22.45	20.10	52.57	39.05	**68.01**	21.80	35.07	**60.37**	61.45	25.09	45.43
	SmoothQuant	17.68	22.35	20.01	52.64	**39.66**	67.74	21.80	**35.14**	60.15	61.49	25.43	45.51
	OS+	**17.67**	**22.32**	**19.99**	53.51	39.00	67.79	**23.00**	**35.14**	60.09	**61.66**	**25.85**	**45.76**
	QuaRot	17.77	22.42	20.10	52.33	39.15	**68.01**	22.80	**35.14**	60.34	61.15	25.34	45.53

Table 1: Quantization Results for SmolLM-135M model. Activation clipping and online rotation within QuaRot are canceled for a fair comparison. “HellaS.” and “WinoG.” represent HellaSwag and WinoGrande, respectively. We mark the best results in **bold**.

#Bits	**Method**	**PPL** ↓			**Accuracy** (%) ↑								
		WikiText2	C4	Avg.	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	WinoG.	Avg.
Full Prec.	-	13.10	17.68	15.39	58.25	41.25	71.33	25.20	41.63	55.20	69.82	33.28	49.49
w2a16g128	RTN	2.98e+06	2.60e+06	2.79e+06	51.22	32.91	51.74	**16.40**	25.72	**47.95**	25.21	**20.90**	**34.01**
	GPTQ	**797.15**	**812.25**	**804.70**	48.62	**34.95**	50.22	16.00	26.24	39.30	27.69	18.69	32.71
	AWQ	3.12e+03	2.67e+03	2.90e+03	48.93	34.08	**52.50**	15.20	**26.82**	42.11	**30.68**	19.97	33.79
w3a16g128	RTN	32.13	39.52	35.83	53.35	36.80	**67.30**	**22.20**	36.23	**62.02**	57.87	29.44	**45.65**
	GPTQ	**21.14**	**26.85**	**24.00**	52.64	37.77	65.56	19.40	36.47	51.96	57.95	27.99	43.72
	AWQ	23.24	28.91	26.08	53.75	**38.28**	66.76	21.00	**37.86**	53.91	**61.41**	**29.86**	45.35
w4a16g128	RTN	15.11	20.20	17.65	56.20	40.53	**70.46**	**24.20**	40.39	54.37	65.87	32.00	48.00
	GPTQ	**14.80**	**19.72**	**17.26**	55.72	39.36	69.91	23.80	39.75	**54.43**	66.20	31.14	47.54
	AWQ	15.17	20.08	17.63	57.06	**40.94**	69.26	23.00	**41.00**	51.74	**68.27**	**32.85**	**48.02**
w4a4	RTN	645.64	613.99	629.82	51.14	33.88	54.52	13.80	26.51	43.61	33.96	19.37	34.60
	SmoothQuant	123.40	233.90	178.65	48.70	35.36	**59.47**	**17.20**	30.35	45.17	44.87	**24.66**	**38.22**
	OS+	**80.14**	**122.98**	**101.56**	49.96	**35.41**	58.43	13.20	**30.46**	**47.06**	**48.70**	21.67	38.11
	QuaRot	157.89	158.13	158.01	49.41	34.44	57.73	15.80	28.28	39.08	40.57	21.16	35.81
w6a6	RTN	15.32	21.15	18.24	55.17	40.23	69.15	23.00	39.44	48.78	66.46	30.89	46.64
	SmoothQuant	14.26	19.17	16.72	53.20	40.99	69.53	**26.80**	40.84	53.98	67.85	32.08	**48.16**
	OS+	**14.15**	**19.01**	**16.58**	54.14	**41.40**	**69.75**	23.00	**40.86**	53.88	67.34	**32.42**	47.85
	QuaRot	14.36	19.24	16.80	54.30	40.84	69.64	24.40	40.41	**55.05**	**68.35**	32.00	48.12
w8a8	RTN	13.31	17.97	15.64	56.04	40.58	70.67	25.80	41.64	55.20	**70.24**	33.79	49.24
	SmoothQuant	13.27	17.90	15.58	56.75	**41.30**	70.95	25.80	41.67	**55.96**	70.03	33.53	49.50
	OS+	**13.24**	**17.85**	**15.55**	55.96	41.10	**71.16**	**26.20**	41.67	55.84	70.16	**34.04**	**49.52**
	QuaRot	13.26	17.90	15.58	56.75	40.89	**71.16**	25.20	**41.73**	53.82	69.87	33.87	49.16

Table 2: Quantization Results for SmolLM-350M model.

#Bits	**Method**	**PPL** ↓			**Accuracy** (%) ↑								
		WikiText2	C4	Avg.	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	WinoG.	Avg.
Full Prec.	-	9.58	13.92	11.75	60.93	43.65	75.79	30.00	49.55	65.93	76.47	43.43	55.72
w2a16g128	RTN	1.40e+07	1.06e+07	1.23e+07	49.64	33.42	53.10	**17.20**	25.85	44.50	25.42	22.61	33.97
	GPTQ	465.98	319.93	392.95	51.70	34.60	51.25	15.60	27.03	51.38	30.68	19.28	35.19
	AWQ	**91.93**	**122.20**	**107.06**	49.64	**34.65**	**60.72**	16.40	**31.11**	**56.36**	**50.38**	**23.38**	**40.33**
w3a16g128	RTN	17.57	23.43	20.50	56.99	41.20	72.36	**28.60**	**45.72**	61.47	70.20	**39.93**	52.06
	GPTQ	**12.10**	16.85	14.47	58.56	40.89	73.01	27.80	45.21	61.56	71.09	37.37	51.94
	AWQ	12.11	**16.68**	**14.40**	57.70	**41.81**	**73.34**	28.20	45.22	**63.91**	**72.81**	39.76	**52.84**
w4a16g128	RTN	10.56	15.13	12.85	60.30	**44.52**	75.08	**31.20**	**49.12**	63.00	**76.05**	**43.52**	**55.35**
	GPTQ	**10.05**	14.45	12.25	60.54	43.76	74.97	29.40	48.43	65.29	75.67	42.41	55.06
	AWQ	10.05	**14.43**	**12.24**	60.77	43.50	**75.79**	29.60	48.56	**65.57**	75.97	42.92	55.34
w4a4	RTN	1.34e+07	8.32e+07	4.83e+07	50.59	33.06	50.98	14.80	24.50	48.87	29.38	22.18	34.30
	SmoothQuant	285.34	222.59	253.96	51.62	34.24	54.46	15.60	29.47	55.78	42.68	23.29	38.39
	OS+	403.41	882.42	642.91	47.99	36.03	55.77	17.40	29.64	54.04	47.60	25.00	39.18
	QuaRot	**37.41**	**49.55**	**43.48**	50.20	**37.15**	**60.07**	**17.80**	**34.05**	**58.90**	**52.10**	**26.45**	**42.09**
w6a6	RTN	11.71	16.65	14.18	56.20	41.97	73.29	28.60	46.47	63.73	72.81	38.40	52.68
	SmoothQuant	10.71	15.54	13.12	59.35	42.27	**74.43**	**30.40**	47.87	64.46	74.37	39.85	54.12
	OS+	10.51	15.13	12.82	58.96	**42.43**	73.99	29.20	48.25	64.83	73.78	40.44	53.98
	QuaRot	**10.35**	**14.99**	**12.67**	58.09	**42.43**	73.83	29.60	**48.65**	**65.14**	**74.66**	**40.70**	**54.14**
w8a8	RTN	9.73	14.21	11.97	59.67	**43.86**	**76.01**	**30.60**	49.40	66.02	76.35	42.92	55.60
	SmoothQuant	9.65	14.04	11.84	61.33	43.50	75.63	30.40	49.37	65.81	76.60	43.00	55.71
	OS+	9.64	14.01	11.83	60.46	43.65	75.63	30.00	**49.43**	**66.36**	**76.73**	**44.03**	**55.79**
	QuaRot	**9.64**	**14.01**	**11.82**	59.91	43.30	75.79	30.20	49.33	**66.36**	76.64	43.43	55.62

Table 3: Quantization Results for SmolLM-1.7B model.

#Bits	Method	PPL ↓			Accuracy (%) ↑								
		WikiText2	C4	Avg.	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	WinoG.	Avg.
Full Prec.	-	8.60	13.74	11.17	60.62	45.04	74.48	23.20	50.09	68.23	70.37	36.26	53.54
w2a16g128	RTN	7.86e+03	1.61e+04	1.20e+04	50.91	34.54	53.10	13.80	25.90	40.70	26.39	22.44	33.47
	GPTQ	**71.23**	**101.64**	**86.44**	48.86	36.03	57.24	16.20	29.00	**43.12**	33.88	19.45	35.47
	AWQ	100.70	197.93	149.31	52.64	**38.18**	**60.83**	**16.60**	**31.88**	42.60	**44.78**	**22.95**	**38.81**
w3a16g128	RTN	11.00	17.70	14.35	60.77	41.50	72.42	19.60	46.76	63.79	63.93	33.28	50.26
	GPTQ	10.34	16.44	13.39	60.62	42.99	71.60	21.80	46.40	60.64	65.11	**35.24**	50.55
	AWQ	**10.01**	**16.23**	**13.12**	59.67	**44.52**	**72.63**	**22.40**	**47.07**	**65.38**	**66.84**	34.30	**51.60**
w4a16g128	RTN	8.98	14.35	11.67	59.43	44.37	73.07	23.20	**49.42**	67.13	**69.40**	36.35	52.80
	GPTQ	8.89	**14.23**	**11.56**	60.46	44.06	**73.39**	22.80	49.00	69.24	68.64	36.09	52.96
	AWQ	**8.87**	14.25	11.56	61.17	**45.19**	73.29	**23.40**	49.36	**71.01**	69.32	**36.60**	**53.67**
w4a4	RTN	35.70	50.17	42.93	52.17	39.05	64.09	16.60	36.29	57.34	51.47	25.51	42.81
	SmoothQuant	19.75	30.51	25.13	52.33	40.48	65.45	19.00	**40.68**	60.40	**55.18**	28.07	**45.20**
	OS+	21.72	33.72	27.72	51.22	**40.79**	**65.67**	**20.20**	40.59	59.82	53.79	**28.67**	45.09
	QuaRot	**19.18**	**30.01**	**24.60**	52.01	35.31	59.30	18.00	28.77	**63.39**	41.25	26.54	40.57
w6a6	RTN	9.09	14.44	11.77	61.01	**44.58**	**74.16**	22.40	49.33	69.30	69.11	**36.60**	**53.31**
	SmoothQuant	9.03	14.39	11.71	60.06	44.11	73.18	**23.40**	49.25	69.24	**69.70**	36.09	53.13
	OS+	9.05	**14.38**	11.72	59.83	44.52	73.88	**23.40**	**49.48**	68.93	68.94	36.01	53.12
	QuaRot	**9.01**	14.41	**11.71**	58.80	36.85	65.29	20.20	31.16	**69.48**	47.35	29.35	44.81
w8a8	RTN	8.65	13.80	11.23	62.04	44.17	**74.48**	**23.80**	49.86	68.10	70.08	**36.52**	**53.63**
	SmoothQuant	8.64	13.79	11.21	59.91	44.11	74.32	22.20	49.90	68.32	**70.29**	35.67	53.09
	OS+	**8.63**	**13.78**	**11.21**	59.91	**44.63**	74.16	23.00	**49.92**	68.17	70.08	35.67	53.19
	QuaRot	8.64	13.79	11.22	60.38	37.15	64.96	22.40	31.53	**68.99**	48.06	29.61	45.38

Table 4: Quantization Results for MiniCPM-1B model.

#Bits	Method	PPL ↓			Accuracy (%) ↑								
		WikiText2	C4	Avg.	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	WinoG.	Avg.
Full Prec.	-	8.16	13.00	10.58	63.14	47.24	76.22	28.60	52.88	73.58	74.66	42.58	57.36
w2a16g128	RTN	612.79	880.31	746.55	49.01	35.52	56.64	15.80	28.51	58.93	31.86	20.14	37.05
	GPTQ	29.60	45.30	37.45	47.75	36.44	60.88	15.20	32.90	55.87	38.85	21.67	38.69
	AWQ	**24.28**	**36.25**	**30.26**	55.09	**40.07**	**66.10**	**16.80**	**39.54**	**63.70**	**55.89**	**29.35**	**45.82**
w3a16g128	RTN	9.79	15.54	12.66	60.22	44.58	**74.48**	25.80	50.42	71.83	70.12	**40.78**	54.78
	GPTQ	9.56	15.29	12.43	61.33	43.91	73.50	25.40	50.23	73.12	69.74	37.97	54.40
	AWQ	**9.18**	**14.68**	**11.93**	60.85	**46.21**	74.05	**27.20**	**51.07**	**73.36**	**71.76**	40.27	**55.60**
w4a16g128	RTN	8.40	13.43	10.92	64.96	47.34	**76.22**	**28.80**	52.77	73.70	74.45	**42.24**	**57.56**
	GPTQ	8.50	13.59	11.04	61.88	**47.39**	75.30	27.40	52.65	**75.14**	73.40	41.89	56.88
	AWQ	**8.32**	**13.39**	**10.85**	61.33	46.88	75.73	**28.80**	**52.80**	74.65	**74.96**	41.72	57.11
w4a4	RTN	33.64	52.72	43.18	53.35	38.64	64.85	18.00	37.21	62.60	52.23	26.71	44.20
	SmoothQuant	17.20	**28.01**	**22.61**	53.99	**41.91**	68.39	**23.20**	42.71	**63.88**	**60.02**	**33.28**	**48.42**
	OS+	**17.15**	28.22	22.68	53.75	41.15	**68.50**	20.40	**43.25**	63.82	59.22	31.91	47.75
	QuaRot	19.87	31.97	25.92	53.51	35.98	61.04	16.80	27.36	61.50	40.91	25.09	40.27
w6a6	RTN	8.46	13.58	11.02	63.14	**45.96**	75.03	27.60	**52.21**	73.21	73.78	40.61	56.44
	SmoothQuant	**8.43**	**13.49**	**10.96**	62.59	45.24	**75.63**	27.80	52.03	72.75	73.99	41.21	56.41
	OS+	8.45	13.51	10.98	61.33	45.55	74.65	**28.00**	52.01	**74.07**	**74.16**	**41.81**	**56.45**
	QuaRot	8.48	13.55	11.01	61.56	38.33	65.18	20.60	28.49	72.23	52.36	31.91	46.33
w8a8	RTN	**8.13**	13.04	**10.59**	63.77	46.57	**76.39**	29.40	**52.97**	73.94	**74.66**	42.24	**57.49**
	SmoothQuant	8.17	13.04	10.60	63.06	46.93	76.33	29.20	52.80	73.73	74.62	**42.32**	57.37
	OS+	8.18	**13.04**	10.61	63.06	**47.19**	76.17	**29.60**	52.80	**74.01**	74.28	41.89	57.38
	QuaRot	8.18	13.04	10.61	62.75	38.43	66.05	22.40	28.57	73.58	53.07	31.83	47.09

Table 5: Quantization Results for MiniCPM-2B model.

#Bits	**Method**	**PPL ↓**			**Accuracy (%) ↑**								
		WikiText2	C4	Avg.	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	WinoG.	Avg.
Full Prec.	-	13.58	18.97	16.27	57.70	43.04	69.48	21.40	38.35	61.04	54.76	25.51	46.41
w2a16g128	RTN	2.09e+05	1.97e+05	2.03e+05	51.30	34.03	53.37	14.40	25.48	44.86	25.00	**22.70**	33.89
	GPTQ	**1.34e+03**	**1.39e+03**	**1.37e+03**	50.04	**34.49**	**53.70**	13.80	25.95	44.28	27.86	20.90	33.88
	AWQ	9.73e+03	8.82e+03	9.27e+03	48.54	33.06	53.37	**15.00**	**26.30**	**46.36**	**28.75**	20.14	**33.94**
w3a16g128	RTN	32.82	45.18	39.00	52.64	37.51	62.62	18.80	33.34	45.08	46.04	23.38	39.93
	GPTQ	**19.62**	**28.06**	**23.84**	52.96	37.10	**66.43**	**19.00**	34.71	**59.60**	**51.98**	**24.49**	**43.28**
	AWQ	22.72	30.28	26.50	52.96	**39.00**	66.16	18.00	**35.18**	57.34	49.28	23.72	42.70
w4a16g128	RTN	15.75	21.90	18.83	54.54	40.69	67.68	**21.20**	37.42	**62.32**	51.01	23.98	44.86
	GPTQ	**14.86**	**20.80**	**17.83**	55.49	41.04	67.90	21.00	37.47	59.17	**56.86**	24.57	**45.44**
	AWQ	14.90	20.86	17.88	56.99	**41.20**	**68.44**	19.20	**37.50**	59.45	52.90	**24.83**	45.06
w4a4	RTN	1.09e+03	1.01e+03	1.05e+03	48.86	34.60	52.45	13.20	26.23	41.71	27.86	19.62	33.07
	SmoothQuant	172.65	232.83	202.74	49.72	34.49	54.73	12.80	27.93	45.66	32.58	21.33	34.90
	OS+	261.88	271.76	266.82	52.09	33.93	56.75	15.20	28.44	46.02	33.84	21.33	35.95
	QuaRot	**57.48**	**78.85**	**68.16**	51.54	**35.21**	**59.63**	**15.80**	**30.25**	**48.20**	**38.97**	**21.42**	**37.63**
w6a6	RTN	15.79	21.99	18.89	53.99	40.33	67.08	21.00	37.14	50.46	53.66	26.37	43.75
	SmoothQuant	15.29	21.25	18.27	55.09	41.04	67.19	**21.40**	37.63	54.34	53.37	**26.54**	44.58
	OS+	15.32	21.22	18.27	54.78	**42.07**	**68.44**	20.40	**37.92**	53.33	54.76	25.85	44.69
	QuaRot	**14.93**	**20.82**	**17.87**	55.17	41.56	67.63	**21.40**	37.62	**57.40**	**55.72**	25.43	**45.24**
w8a8	RTN	13.85	19.37	16.61	56.12	42.22	69.37	**21.80**	**38.32**	58.93	54.59	25.17	45.81
	SmoothQuant	13.72	19.20	16.46	56.99	42.37	**69.80**	21.00	38.29	59.97	54.71	25.60	46.09
	OS+	**13.70**	**19.16**	**16.43**	58.33	**42.73**	**69.80**	21.20	38.29	59.79	**55.51**	**25.85**	**46.44**
	QuaRot	13.70	19.17	16.44	55.88	42.48	69.64	**21.80**	38.26	**60.61**	55.22	25.26	46.14

Table 6: Quantization Results for Qwen2-0.5B model.

#Bits	**Method**	**PPL ↓**			**Accuracy (%) ↑**								
		WikiText2	C4	Avg.	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	WinoG.	Avg.
Full Prec.	-	9.84	14.36	12.10	64.72	46.11	75.57	26.80	48.31	71.96	65.87	33.45	54.10
w2a16g128	RTN	3.41e+04	2.45e+04	2.93e+04	50.20	32.75	50.60	**15.00**	25.84	**46.12**	25.29	20.48	33.28
	GPTQ	482.42	462.96	472.69	52.49	34.08	54.30	14.60	26.58	44.40	28.96	**20.82**	34.53
	AWQ	**326.04**	**398.14**	**362.09**	51.38	**34.95**	**55.98**	14.80	**28.12**	42.72	**34.72**	20.31	**35.37**
w3a16g128	RTN	15.24	21.27	18.26	61.72	43.19	70.95	23.00	43.65	68.01	60.14	**32.00**	50.33
	GPTQ	**12.39**	**18.54**	**15.47**	61.25	43.24	**71.98**	**25.20**	**44.39**	**68.44**	**62.50**	30.89	**50.99**
	AWQ	13.47	19.40	16.43	62.04	**43.45**	71.22	23.20	44.11	65.96	59.55	28.07	49.70
w4a16g128	RTN	10.59	15.29	12.94	64.01	44.73	74.59	**26.40**	47.21	**72.39**	62.46	31.48	52.91
	GPTQ	**10.28**	**15.02**	**12.65**	66.14	45.29	74.54	**26.40**	**47.68**	71.07	65.07	**32.68**	53.61
	AWQ	10.41	15.16	12.79	66.69	**46.57**	**75.24**	26.00	47.22	70.55	**65.40**	31.83	**53.69**
w4a4	RTN	275.87	265.84	270.85	50.99	34.54	55.77	13.60	28.63	44.68	31.48	20.56	35.03
	SmoothQuant	85.82	105.29	95.56	48.93	35.16	**59.52**	16.60	32.30	45.60	37.29	**23.98**	37.42
	OS+	98.76	115.03	106.89	50.67	**37.10**	56.96	13.00	31.65	46.79	36.41	21.42	36.75
	QuaRot	**42.19**	**56.01**	**49.10**	52.17	35.82	58.65	**17.80**	**34.72**	**50.86**	**38.38**	21.50	**38.74**
w6a6	RTN	11.02	15.83	13.42	63.93	43.91	72.80	25.80	46.91	63.64	62.88	31.83	51.46
	SmoothQuant	10.94	15.74	13.34	63.30	44.83	73.12	25.80	**47.41**	65.57	63.80	32.42	52.03
	OS+	**10.84**	**15.59**	**13.22**	63.77	45.14	73.18	**27.40**	47.27	62.35	62.25	32.25	51.70
	QuaRot	10.86	15.61	13.24	64.17	**46.37**	**74.21**	26.60	47.21	**67.52**	**65.28**	**34.13**	**53.19**
w8a8	RTN	9.96	14.43	12.19	64.88	46.37	75.30	26.80	**48.13**	72.26	65.87	33.11	54.09
	SmoothQuant	9.97	14.41	12.19	65.67	**47.13**	**75.35**	**27.40**	47.98	72.20	**67.34**	**33.19**	**54.53**
	OS+	9.93	14.31	12.12	65.82	46.88	**75.35**	26.40	**48.13**	**72.42**	65.53	**33.19**	54.22
	QuaRot	**9.89**	**14.31**	**12.10**	65.59	46.06	75.03	26.60	48.09	71.65	65.87	33.02	53.99

Table 7: Quantization Results for Qwen2-1.5B model.