Appendix

A. Experiment on more LLMs

Table 2: The performance of the LLaMA2-7B model on five zero-shot benchmarks by reporting their average accuracy and one language modeling task using perplexity.

Method	#Bit	PIQA	Hella.	Wino.	Arc-e	OpenBookQA	Avg. (↑)	Wiki (↓)
LLaMA2-7B	16.0	76.88	72.95	67.25	53.49	41.00	62.31	8.79
RTN	4.0	77.04	72.57	66.54	55.39	40.00	62.31	9.21
	3.0	75.52	71.10	67.01	52.78	40.20	61.32	11.21
	2.0	51.31	26.27	49.49	27.40	26.80	36.25	2e6
AWQ [8]	4.0	76.71	72.58	66.69	53.28	41.20	62.09	9.04
	3.0	76.66	70.66	65.43	52.65	40.60	61.20	10.30
	2.0	49.73	26.14	49.80	26.52	32.40	36.92	1e7
SqueezeLLM [7]	4.0	76.71	72.23	68.27	53.41	40.00	62.12	9.50
	3.0	76.50	69.95	65.90	51.47	40.40	60.84	10.56
LLM-MQ (Ours)	4.0	76.82	72.42	67.96	53.79	41.40	62.48	8.96
	3.8	77.09	72.35	67.01	53.41	41.20	62.21	9.04
	3.6	76.61	71.85	66.85	52.74	39.80	61.57	9.23
	3.4	75.95	71.55	67.25	51.56	41.00	61.46	9.34
	3.2	76.39	71.28	64.80	52.57	39.40	60.89	9.52
	3.0	76.28	71.03	65.98	51.47	40.00	60.95	9.65
	2.8	76.01	70.23	66.22	51.30	39.60	60.67	10.17
	2.6	75.03	69.53	64.96	50.13	40.40	60.01	10.96
	2.4	74.21	67.67	64.56	48.36	38.20	58.60	12.01
	2.2	74.48	65.51	63.85	47.43	37.60	57.77	13.37
	2.0	72.31	61.14	59.91	45.88	38.80	55.61	15.93

Table 3: The performance of the OPT-6.7B model on five zero-shot benchmarks by reporting their average accuracy and one language modeling task using perplexity.

Method	#Bit	PIQA	Hella.	Wino.	Arc-e	OpenBookQA	Avg. (↑)	Wiki (↓)
OPT-6.7B	16.0	76.44	67.17	65.19	60.10	37.20	61.22	12.29
RTN	4.0	76.33	65.69	64.09	58.59	37.60	60.46	13.02
	3.0	72.42	58.41	59.91	52.40	34.60	55.55	43.15
	2.0	49.46	26.08	47.43	24.92	27.60	35.10	2e5
AWQ (8)	4.0	76.39	66.84	64.56	60.10	36.60	60.90	12.44
	3.0	75.79	65.45	64.80	57.00	38.00	60.21	12.99
	2.0	71.06	56.72	59.91	52.74	35.00	55.09	18.77
SqueezeLLM (7)	4.0	76.12	66.55	64.48	60.27	37.20	60.92	12.45
	3.0	75.68	63.96	64.88	58.67	35.40	59.72	13.17
LLM-MQ (Ours)	4.0	76.39	67.16	65.11	60.19	37.80	61.33	12.41
	3.8	76.39	66.91	65.11	60.10	38.20	61.34	12.46
	3.6	76.39	66.32	64.40	59.60	38.20	60.98	12.97
	3.4	75.84	65.77	64.88	58.67	39.00	60.83	13.17
	3.2	75.73	65.34	65.43	58.92	36.60	60.40	13.48
	3.0	76.12	65.93	63.77	59.81	38.20	60.77	12.84
	2.8	75.84	65.71	63.30	59.60	38.60	60.61	13.32
	2.6	74.97	63.13	64.40	57.66	36.40	59.31	14.94
	2.4	75.52	64.14	64.96	58.71	36.40	59.95	14.64
	2.2	74.43	62.70	63.38	57.74	36.40	58.93	15.70
	2.0	74.27	60.87	61.88	55.93	35.20	57.63	17.09

Table 4: The performance of the OPT-13B model on five zero-shot benchmarks by reporting their average accuracy and one language modeling task using perplexity.

Method	#Bit	PIQA	Hella.	Wino.	Arc-e	OpenBookQA	Avg. (↑)	Wiki (↓)
OPT-13B	16.0	76.88	69.81	65.04	61.78	39.00	62.50	11.49
RTN	4.0	76.22	68.21	64.64	62.46	37.80	61.87	11.88
	3.0	70.51	45.57	58.17	50.17	33.20	51.52	45.36
	2.0	49.46	26.20	49.49	25.25	27.60	35.60	1e6
AWQ (8)	4.0	76.01	69.66	65.43	61.74	39.00	62.37	11.60
	3.0	76.55	68.25	64.56	59.97	36.60	61.19	12.03
	2.0	71.98	57.07	60.54	50.17	35.20	54.99	16.06
SqueezeLLM (7)	4.0	76.61	68.91	64.88	62.29	39.00	62.34	11.62
	3.0	75.30	66.36	64.72	59.01	38.40	60.80	13.37
LLM-MQ (Ours)	4.0	76.66	69.37	65.43	61.36	38.00	62.16	11.59
	3.8	76.77	69.39	65.27	60.94	39.00	62.27	11.62
	3.6	76.22	68.98	64.48	61.74	39.60	62.20	11.82
	3.4	76.71	68.86	64.88	62.04	39.20	62.34	11.92
	3.2	75.95	67.24	67.09	60.27	37.80	61.67	12.70
	3.0	76.22	68.43	64.64	62.08	39.40	62.15	11.99
	2.8	74.92	67.13	64.80	61.45	38.00	61.26	12.88
	2.6	75.35	66.88	64.56	60.02	36.60	60.68	13.39
	2.4	74.70	65.32	64.17	59.09	36.20	59.90	14.45
	2.2	74.59	62.59	65.35	57.58	35.00	59.02	15.79
	2.0	73.23	60.63	64.17	54.55	34.20	57.36	17.24