

## A More Results

In this section, we report more results in our LLMCBench. The results for the first compression performance track are shown in Table 7. We also observe that quantization approaches have higher overall metrics on Vicuna, which is similar to the observation in our main paper.

For track 2 the generalization ability track, we report INT8 quantization results in Table 8. For track 3 the training consumption track, we report full results in Table 4. We also show the results of training consumption on Vicuna-7B in Table 9. For track 4 the inference consumption track, we report more results in Table 10. For track 5 the hardware acceleration track, we report full results in Table 5, and the results on Vicuna-7B are shown in Table 11. For track 6 the trustworthiness track, we report full results in Table 6 and the results on Vicuna-7B in Table 12.

## B Evaluation Model Selection

We choose LLaMA2 and LLaMA3 as the baseline models for evaluation in tracks 1, 3, 4, 5, 6. The reason is that the mainstream LLMs include: LLaMA [26], LLaMA2 [27], LLaMA3, Vicuna [35], OPT [34], and ChatGLM [4]. Among these models, LLaMA2, LLaMA3, Vicuna, and ChatGLM are recently proposed and their performances are promising. The structure of LLaMA2, Vicuna, and ChatGLM are similar, and most compression approaches are tested on LLaMA-like models in their own experiments. Therefore, we choose the most representative one LLaMA2 as one of our baseline model. As LLaMA3 has recently been proposed and achieves state-of-the-art performance, we also choose LLaMA3 as the baseline model in our evaluation.

Table 4: Full results of Track 3: Training consumption in our LLMCBench.

Method	Model	Sparsity/#Bits	Training time	GPU memory	OM <sub>train</sub>
Sparsity					
LLM-Pruner	LLaMA2-7B	50%	0.5min (sparsity) + 121min (retrain)	33.79G	1.77
	LLaMA3-8B	50%	0.5min (sparsity) + 183min (retrain)	37.61G	
Wanda	LLaMA2-7B	50%	4min	26.96G	43.78
	LLaMA3-8B	50%	10min	31.58G	
Wanda	LLaMA2-7B	2:4	4min	26.96G	42.14
	LLaMA3-8B	2:4	9min	31.58G	
SparseGPT	LLaMA2-7B	50%	16min	25.80G	11.47
	LLaMA3-8B	50%	33min	31.35G	
SparseGPT	LLaMA2-7B	2:4	17min	25.80G	11.03
	LLaMA3-8B	2:4	33min	31.35G	
Quantization					
GPTQ	LLaMA2-7B	INT8	17min	26.36G	13.28
	LLaMA3-8B	INT8	19min	40.31G	
SmoothQuant	LLaMA2-7B	INT8	7min	13.51G	39.45
	LLaMA3-8B	INT8	15min	15.96G	
AWQ	LLaMA2-7B	INT8	12min	11.70G	19.73
	LLaMA3-8B	INT8	10min	20.10G	
OmniQuant	LLaMA2-7B	INT8	325min	29.45G	1.08
	LLaMA3-8B	INT8	307min	30.61G	

Table 5: Full results of Track 5: Hardware Acceleration in our LLMCBench.

Method	Model	Sparsity/#Bits	Tokens/s			OM <sub>hard</sub>
			TensorRT-LLM	vLLM	MLC-LLM	
Sparsity						
Dense	LLaMA2-7B	0	95.62	85.86	107.45	100
	LLaMA3-8B	0	79.65	77.42	91.83	
Structured sparsity	LLaMA2-7B	50%	145.28	129.61	155.29	146.97
	LLaMA3-8B	50%	116.03	115.31	128.01	
Structured 2:4 sparsity	LLaMA2-7B	2:4	120.51	85.94	107.44	106.56
	LLaMA3-8B	2:4	88.07	78.21	90.91	
Unstructured sparsity	LLaMA2-7B	50%	96.15	85.85	107.29	100.11
	LLaMA3-8B	50%	79.76	77.86	91.43	
Quantization						
Full-Precision	LLaMA2-7B	FP16	95.62	85.86	107.45	100
	LLaMA3-8B	FP16	79.65	77.42	91.83	
Quantization	LLaMA2-7B	INT8	150.42	115.48	154.27	146.25
	LLaMA3-8B	INT8	126.12	112.32	125.80	
Quantization	LLaMA2-7B	INT4	182.29	152.46	186.19	180.74
	LLaMA3-8B	INT4	153.39	145.97	146.55	

Table 6: Full results of Track 6: Model Trustworthiness in our LLMCBench.

Method	Model	Sparsity/#Bits	Robustness	Truthfulness	OM <sub>trust</sub>
Sparsity					
Dense	LLaMA2-7B	0	39.01	44.77	100
	LLaMA3-8B	0	46.52	56.21	
LLM-Pruner	LLaMA2-7B	50%	39.72	43.79	94.09
	LLaMA3-8B	50%	46.05	42.37	
Wanda	LLaMA2-7B	50%	36.71	39.27	87.32
	LLaMA3-8B	50%	41.61	43.36	
Wanda	LLaMA2-7B	2:4	36.63	41.95	90.95
	LLaMA3-8B	2:4	44.10	45.34	
SparseGPT	LLaMA2-7B	50%	36.67	41.53	88.88
	LLaMA3-8B	50%	41.43	44.35	
SparseGPT	LLaMA2-7B	2:4	37.92	38.70	88.41
	LLaMA3-8B	2:4	42.72	43.22	
Quantization					
Full-Precision	LLaMA2-7B	FP16	39.01	44.77	100
	LLaMA3-8B	FP16	46.52	56.21	
GPTQ	LLaMA2-7B	INT8	38.75	42.66	97.22
	LLaMA3-8B	INT8	46.51	52.94	
SmoothQuant	LLaMA2-7B	INT8	40.18	45.62	98.35
	LLaMA3-8B	INT8	45.49	50.71	
AWQ	LLaMA2-7B	INT8	38.84	42.80	97.51
	LLaMA3-8B	INT8	46.46	53.36	
OmniQuant	LLaMA2-7B	INT8	39.02	45.20	99.55
	LLaMA3-8B	INT8	46.72	54.38	

Table 7: More results on compression performance track evaluated on Vicuna-7B. H.S. means HellaSwag.

Method	Sparsity /#Bits	Knowledge ability			Inference ability					
		MMLU	ARC-c	ARC-e	H.S.	PIQA	Wino	QNLI	MNLI	Wiki↓
Sparsity										
Dense	0	48.80	45.90	71.21	73.81	77.97	69.53	57.17	53.55	6.33
LLM-Pruner	50%	24.01	29.35	48.65	47.32	66.16	54.85	49.59	34.84	23.98
Wanda	50%	40.98	42.75	68.43	69.64	74.65	67.80	54.42	54.11	7.95
	2:4	25.99	35.49	60.44	57.80	71.44	62.43	50.61	37.03	13.44
SparseGPT	50%	41.87	41.98	66.62	69.79	75.68	67.01	55.36	53.91	7.94
	2:4	34.46	37.20	62.46	61.24	72.85	66.22	51.23	47.32	11.99
Quantization										
Full Prec.	FP16	48.80	45.90	71.21	73.81	77.97	69.53	57.17	53.55	6.33
GPTQ	INT8	48.70	45.90	71.34	73.70	78.02	69.69	56.65	53.75	6.33
SmoothQuant	INT8	47.11	44.45	70.29	72.23	75.41	67.64	51.75	42.71	6.61
AWQ	INT8	41.97	45.65	71.13	73.81	78.13	69.61	58.00	53.75	6.29
OmniQuant	INT8	48.71	44.88	71.25	73.69	77.80	69.22	56.49	53.16	6.35

Table 8: Generalization ability performance of different 8-bit quantization methods.

Model	Dense	GPTQ	SmoothQuant	AWQ	OmniQuant
LLaMA-7B	5.68	5.68	5.72	5.68	5.69
LLaMA-13B	5.09	5.09	5.12	5.09	5.09
LLaMA-30B	4.10	4.10	4.20	4.10	4.12
LLaMA-65B	3.53	3.53	3.66	3.53	3.53
LLaMA2-7B	5.12	5.12	5.51	5.12	5.13
LLaMA2-13B	4.57	4.57	4.92	4.57	4.59
LLaMA2-70B	3.12	3.12	3.18	3.12	3.38
LLaMA3-8B	5.54	5.54	5.64	6.14	2349.16
LLaMA3-70B	2.59	2.59	2.97	2.59	22202.36
Vicuna-7B	6.33	6.34	6.83	6.34	6.35
Vicuna-13B	5.57	5.57	6.12	5.57	5.59
OPT-1.3B	14.62	15.76	14.79	14.61	14.89
OPT-2.7B	12.47	12.48	12.51	12.47	12.48
OPT-6.7B	10.86	10.86	10.87	10.86	10.86
OPT-13B	10.13	10.13	10.15	10.13	11.55
OPT-30B	9.56	9.56	9.59	9.56	12.55
ChatGLM2-6B	105.58	106.02	640.33	106.08	957.51
ChatGLM3-6B	6.21	6.22	6.89	6.22	6.37

Table 9: Training consumption of different LLM compression methods evaluated on Vicuna-7B.

Method	Sparsity/#Bits	Training time	GPU memory
<b>Sparsity</b>			
LLM-Pruner	50%	0.5min (sparsity) + 120min (retrain)	33.79G
Wanda	50%	4min	26.96G
	2:4	4min	26.96G
SparseGPT	50%	16min	25.80G
	2:4	15min	25.80G
<b>Quantization</b>			
GPTQ	INT8	16min	26.36G
SmoothQuant	INT8	6min	13.51G
AWQ	INT8	10min	11.70G
OmniQuant	INT8	332min	29.45G

Table 10: Inference consumption of different compression methods tested on Vicuna-7B.

Method	Sparsity/#Bits	GPU Memory	Model Size	#MACs
<b>Sparsity</b>				
Dense	0	22.96G	12.55G	0.85T
LLM-Pruner	50%	13.31G	6.75G	0.51T
Wanda	50%	22.96G	12.55G	0.43T
	2:4	22.96G	12.55G	0.43T
SparseGPT	50%	22.96G	12.55G	0.43T
	2:4	22.96G	12.55G	0.43T
<b>Quantization</b>				
Full-Precision	FP16	22.96G	12.55G	0.85T
GPTQ	INT8	15.16G	6.67G	0.23T
SmoothQuant	INT8	23.62G	12.55G	0.23T
AWQ	INT8	15.15G	6.71G	0.23T
OmniQuant	INT8	15.13G	6.53G	0.23T

Table 11: Hardware acceleration of different LLM compression methods on Vicuna-7B.

Method	Sparsity/#Bits	Tokens/s		
		TensorRT-LLM	vLLM	MLC-LLM
Sparsity				
Dense	0	95.62	85.93	107.45
Structured sparsity	50%	131.75	130.89	155.19
Structured 2:4 sparsity	2:4	120.08	85.72	107.33
Unstructured sparsity	50%	95.98	85.35	107.37
Quantization				
Full-Precision	FP16	95.62	85.86	107.45
Quantization	INT8	151.10	122.28	155.16
Quantization	INT4	185.09	167.84	186.78

Table 12: Trustworthiness of different LLM compression methods on Vicuna-7B.

Method	Sparsity/#Bits	Robustness	Truthfulness
<b>Sparsity</b>			
Dense	0	50.46	55.51
LLM-Pruner	50%	51.62	44.21
Wanda	50%	44.96	52.54
	2:4	41.35	45.62
SparseGPT	50%	46.32	53.25
	2:4	42.40	45.48
<b>Quantization</b>			
Full-Precision	FP16	50.46	55.51
GPTQ	INT8	50.44	52.81
SmoothQuant	INT8	51.37	53.11
AWQ	INT8	50.48	52.80
OmniQuant	INT8	50.46	52.80