LLMCBench: Benchmarking Large Language Model Compression for Efficient Deployment

Ge Yang, Changyi He, Jinyang Guo*, Jianyu Wu, Yifu Ding, Aishan Liu, Haotong Qin, Pengliang Ji, Xianglong Liu

Background and motivation

Existing LLM compression works are still far away from practical usage due to two main challenges:

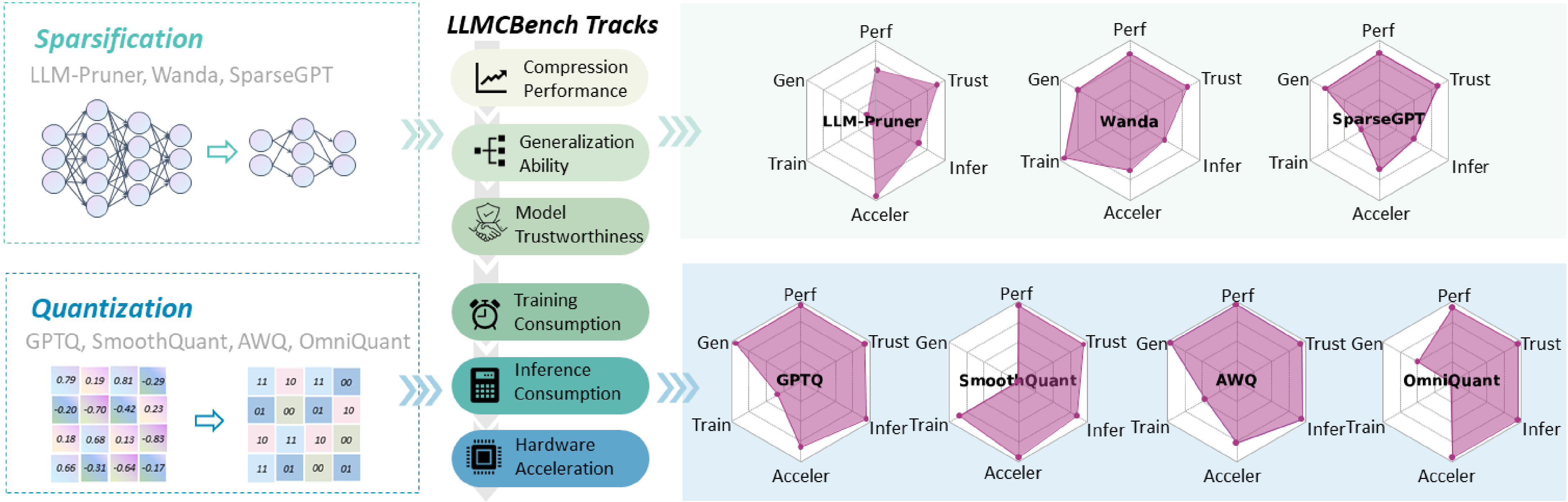
- *Challenge 1: Performance evaluation scope is limited.*

  - Current LLM compression researches often use different types of LLMs for evaluation, which cannot form a fair comparison between different methods.
  - The base model performance is different in current works. (*e.g.*, LLaMA-7B--PPL: 12.62 in LLM-Pruner, 5.68 in OmniQuant)

- *Challenge-2: Efficiency evaluation metric remains theoretical.*

  - Current methods lack a comprehensive evaluation of broader efficiency metrics in actual deployment scenarios. (*e.g.*, practical acceleration, GPU memory reduction)

Overview of our LLMCBench



LLMCBench: Tracks and Metrics

Track 1: Compression Performance

- Knowledge ability: whether the LLM knows the world
- Inference ability: whether the LLM can reason based on its knowledge

$$\mathrm{OM}_{\mathrm{perf}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\frac{A^{c}_{\mathrm{ability}_i}}{A_{\mathrm{ability}_i}}\right)^2}$$

Track 2: Generalization Ability

An effective LLM compression algorithm should be effective for various model types and sizes.

- Model type: LLaMA2, LLaMA3, OPT, ChatGLM2, ChatGLM3, Vicuna
- Model size: 7B, 13B, 30B, 70B, etc.

$$\mathrm{OM}_{\mathrm{gen}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\frac{A^{c}_{\mathrm{mod}_i}}{A_{\mathrm{mod}_i}}\right)^2}$$

LLMCBench: Tracks and Metrics

Track 3: Training Consumption

An effective LLM compression algorithm should require small resources to finish the compression process.

- Training time
- GPU memory

$$\mathrm{OM}_{\mathrm{train}} = \sqrt{\frac{1}{2}\left(\mathbb{E}\left(\frac{T_{\mathrm{train}}^{\max}}{T_{\mathrm{train}}}\right)^2 + \mathbb{E}\left(\frac{M_{\mathrm{train}}^{\max}}{M_{\mathrm{train}}}\right)^2\right)}$$

Track 4: Inference Consumption

- MACs
- GPU memory
- Model size

$$\mathrm{OM}_{\mathrm{inf}} = \sqrt{\frac{1}{3}\left(\mathbb{E}\left(\frac{M_{\mathrm{inf}}}{M_{\mathrm{inf}}^{c}}\right)^2 + \mathbb{E}\left(\frac{S_{\mathrm{inf}}}{S_{\mathrm{inf}}^{c}}\right)^2 + \mathbb{E}\left(\frac{F_{\mathrm{inf}}}{F_{\mathrm{inf}}^{c}}\right)^2\right)}$$

LLMCBench: Tracks and Metrics

Track 5: Hardware Acceleration

Existing LLM compression approaches seldom extensively compare this important aspect, making the acceleration performance remain theoretical.

- TensorRT-LLM
- vLLM
- MLC-LLM

$$\mathrm{OM}_{\mathrm{hard}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\frac{V^{c}_{\mathrm{lib}_i}}{V_{\mathrm{lib}_i}}\right)^{2}}$$

Track 6: Trustworthiness

Since the compressed LLMs need to be deployed in real-world scenarios, their trustworthiness is also a key aspect to avoid negative social impacts.

- Robustness
- Truthfulness

$$\mathrm{OM}_{\mathrm{trust}} = \sqrt{\frac{1}{2}\left(\mathbb{E}\left(\frac{A^{c}_{\mathrm{rob}}}{A_{\mathrm{rob}}}\right)^{2} + \mathbb{E}\left(\frac{A^{c}_{\mathrm{tru}}}{A_{\mathrm{tru}}}\right)^{2}\right)}$$

Evaluation and analysis

Track 1: Compression Performance

- Quantization offers better overall performance.
- Sparsity is better for inference ability, while quantization is better for knowledge ability.

Method	Model	Sparsity /#Bits	Knowledge ability MMLU	ARC-c	ARC-e	Inference ability H.S.	PIQA	Wino	QNLI	MNLI	Wiki↓	OM_{ka}	OM_{ia}	OM_{perf}
Sparsity														
Dense	LMA2	0	40.52	46.33	74.58	75.98	79.11	69.06	50.53	44.31	5.12	100	100	100
	LMA3	0	61.38	53.50	77.74	79.12	80.69	73.24	50.86	63.48	5.54			
LLM-Pruner	LMA2	50%	24.15	27.47	46.52	47.76	68.44	54.14	49.45	34.33	20.66	60.51	75.85	68.61
	LMA3	50%	29.90	32.17	55.09	55.93	69.70	62.51	50.60	40.71	14.22			
Wanda	LMA2	50%	29.67	42.75	69.07	70.78	76.66	68.90	50.67	35.28	6.46	83.25	90.19	86.79
	LMA3	50%	40.59	44.97	68.18	68.23	76.01	70.17	50.60	54.57	8.61			
Wanda	LMA2	2:4	23.63	32.25	58.46	55.11	71.71	62.43	50.64	35.12	6.51	62.53	78.78	71.12
	LMA3	2:4	27.57	28.84	50.04	47.86	66.10	59.83	50.60	32.44	19.98			
SparseGPT	LMA2	50%	34.62	42.24	67.89	71.04	76.44	69.69	50.62	35.16	6.51	85.10	91.29	88.25
	LMA3	50%	48.33	42.15	65.70	71.66	76.71	70.32	50.60	54.96	7.55			
SparseGPT	LMA2	2:4	25.76	33.62	60.23	58.68	72.36	66.14	50.61	36.05	10.28	67.53	81.29	74.73
	LMA3	2:4	28.27	33.87	57.15	56.02	68.28	63.69	50.60	42.50	10.96			
Quantization														
Full Prec.	LMA2	FP16	40.52	46.33	74.58	75.98	79.11	69.06	50.53	44.31	5.12	100	100	100
	LMA3	FP16	61.38	53.50	77.74	79.12	80.69	73.24	50.86	63.48	5.54			
GPTQ	LMA2	INT8	40.77	46.25	74.33	76.00	79.11	68.90	50.62	39.53	6.88	99.97	97.17	98.58
	LMA3	INT8	61.36	53.41	77.69	79.06	80.63	72.85	50.77	63.44	5.54			
SmoothQuant	LMA2	INT8	39.02	44.28	73.36	74.41	78.18	66.93	50.22	38.53	5.53	97.50	96.55	97.03
	LMA3	INT8	58.30	51.96	79.67	78.13	79.54	72.61	51.40	62.90	6.28			
AWQ	LMA2	INT8	40.90	46.16	74.41	75.98	79.05	69.22	50.64	38.86	5.12	99.89	98.89	99.39
	LMA3	INT8	61.22	53.22	77.57	79.15	80.59	72.45	50.46	63.43	5.54			
OmniQuant	LMA2	INT8	40.32	45.65	74.75	75.94	79.00	69.22	50.55	43.59	5.12	99.21	99.63	99.42
	LMA3	INT8	61.19	52.13	77.61	79.23	80.52	72.61	50.73	62.56	5.55			

Track 2: Generalization Ability

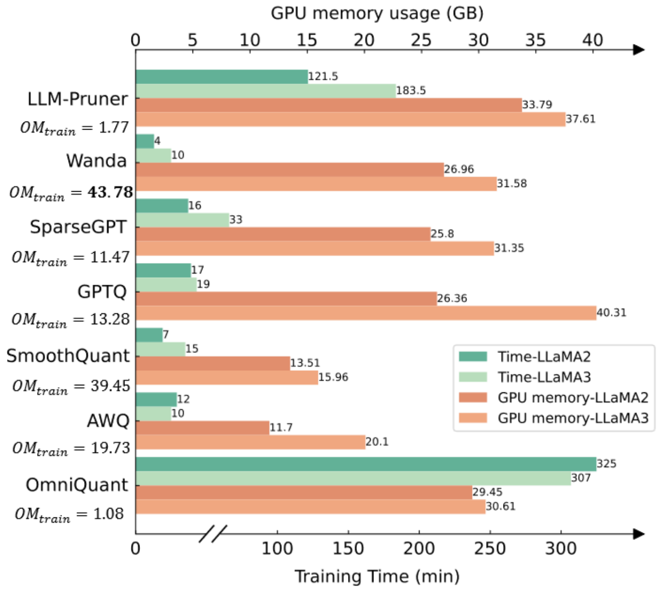
- Weight-only quantization methods have good generalization ability under lower bit.

Model	Dense	LLM-Pruner	Wanda	SparseGPT	GPTQ	SmoothQuant	AWQ	Omniquant
LLaMA-7B	5.68	19.20	7.09	6.73	6.61	380.77	5.78	11.26
LLaMA-13B	5.09	14.15	6.03	5.85	5.20	552.8	5.19	10.86
LLaMA-30B	4.10	9.86	5.18	5.07	4.25	1057.91	4.20	10.63
LLaMA-65B	3.53	8.34	4.55	4.37	3.76	890.32	3.61	9.17
LLaMA2-7B	5.12	18.43	6.46	6.51	5.25	1887.53	5.23	14.26
LLaMA2-13B	4.57	14.10	5.47	5.34	4.66	403.44	4.65	12.29
LLaMA2-70B	3.12	6.34	3.91	3.81	3.31	1306.59	3.21	9604.32
LLaMA3-8B	5.54	15.35	8.61	7.55	5.75	799.70	6.14	12735.95
LLaMA3-70B	2.59	8.40	5.01	4.92	4.71	274.00	3.06	37026.54
Vicuna-7B	6.33	19.11	7.95	7.90	6.50	2636.98	6.51	87.39
Vicuna-13B	5.57	15.99	6.63	6.44	5.66	494.89	5.65	60.22
OPT-1.3B	14.62	124.01	18.41	17.55	16.41	1412.51	14.92	98.6
OPT-2.7B	12.47	163.81	14.22	13.46	12.81	8749.80	12.70	360.26
OPT-6.7B	10.86	119.49	11.98	11.60	11.05	21492.23	10.96	12.24
OPT-13B	10.13	113.89	11.93	11.15	10.22	13176.12	10.29	11.65
OPT-30B	9.56	76.00	10.03	9.77	9.59	12765.02	9.61	10.31
ChatGLM2-6B	105.58	43499.38	3916.7	2534.85	122.97	5887.32	128.58	3624.92
ChatGLM3-6B	6.21	301.05	20.58	33.86	6.34	1175.5	6.4	494.41
OM_{gen}	100	28.89	76.41	79.06	93.80	0.82	96.13	48.51

Evaluation and analysis

Track 3: Training Consumption

◆ Learning is the bottleneck.



Track 4: Inference Consumption

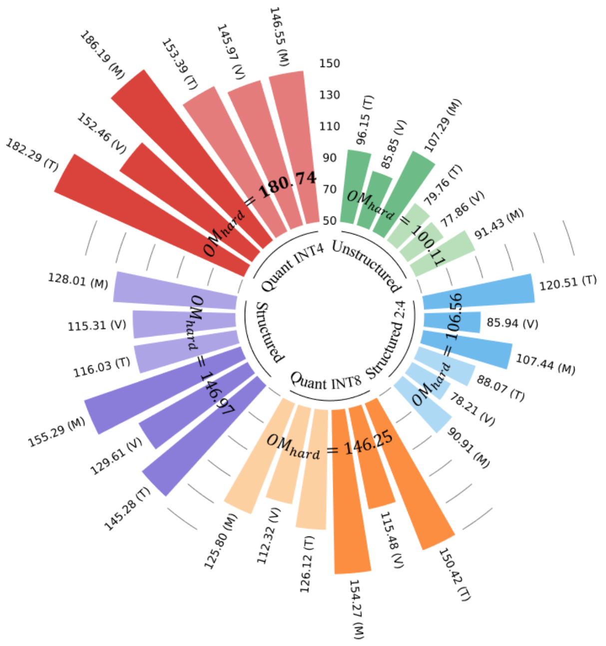
◆ Quantization generally has less inference consumption.

Method	Model	Sparsity/#Bits	GPU Memory	Model Size	#MACs	OM_{inf}
			Sparsity			
Dense	LMA2 LMA3	0 0	22.96G 25.35G	12.55G 14.96G	0.85T 0.97T	100
LLM-Pruner	LMA2 LMA3	50% 50%	13.50G 18.50G	6.75G 9.97G	0.51T 0.62T	161.86
Wanda	LMA2 LMA3	50% 50%	22.96G 25.35G	12.55G 14.96G	0.43T 0.57T	134.76
Wanda	LMA2 LMA3	2:4 2:4	22.96G 25.35G	12.55G 14.96G	0.43T 0.57T	134.76
SparseGPT	LMA2 LMA3	50% 50%	22.96G 25.35G	12.55G 14.96G	0.43T 0.57T	134.76
SparseGPT	LMA2 LMA3	2:4 2:4	22.96G 25.35G	12.55G 14.96G	0.43T 0.57T	134.76
			Quantization			
Full-Precision	LMA2 LMA3	FP16 FP16	22.96G 25.35G	12.55G 14.96G	0.85T 0.97T	100
GPTQ	LMA2 LMA3	INT8 INT8	15.16G 17.03G	6.67G 8.62G	0.23T 0.29T	245.91
SmoothQuant	LMA2 LMA3	INT8 INT8	23.62G 25.02G	12.55G 14.96G	0.23T 0.29T	220.58
AWQ	LMA2 LMA3	INT8 INT8	15.15G 17.72G	6.71G 8.66G	0.23T 0.29T	245.11
OmniQuant	LMA2 LMA3	INT8 INT8	15.13G 17.19G	6.53G 8.61G	0.23T 0.29T	246.34

Evaluation and analysis

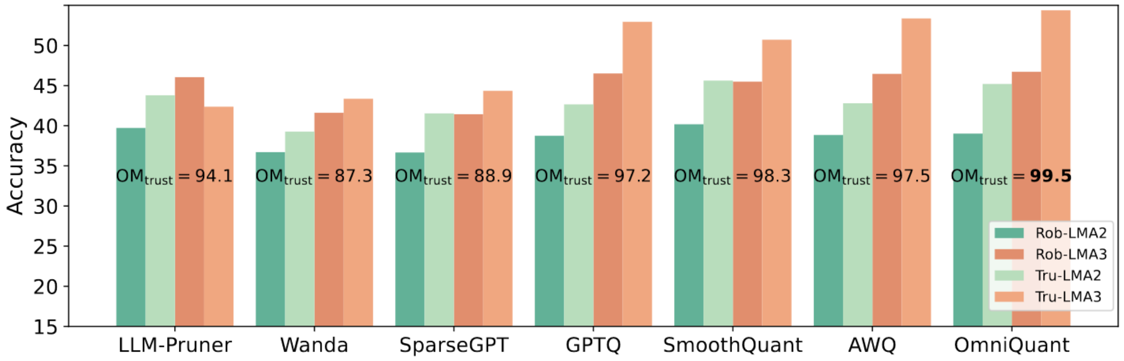
Track 5: Hardware Acceleration

- INT4 quantization has the best acceleration performance.
- Structured sparsity ≈ INT8 quantization.
- Structured 2:4 sparsity is not well-supported.



Track 6: Trustworthiness

- Quantization brings better trustworthiness.
- Better compression performance ≠ better trustworthiness.



Conclusion

- Quantization is preferable for LLM compression due to improved performance and hardware compatibility.
- Weight-activation quantization is better in terms of inference efficiency (inference consumption and hardware acceleration).
- Sparsity generally has better training efficiency. However, its hardware/library support is not well constructed in the current stage.

Resources

- GitHub: https://github.com/AboveParadise/LLMCBench
- Contact: jinyangguo@buaa.edu.cn