“Give Me BF16 or Give Me Death”? Accuracy-Performance Trade-Offs in LLM Quantization

Eldar Kurtić[1,2], Alexandre Marques[1], Shubhra Pandit[1], Mark Kurtz[1], Dan Alistarh[1,2]

[1]Red Hat AI, [2]Institute of Science and Technology Austria

Correspondence: ekurtic@redhat.com, dalistar@redhat.com

ACL 2025

HUMANE Lab, 석사과정 최종현

랩 세미나

2025.09.17

Backgrounds

- Quantization is used to accelerate LLM inference
- Trade-offs between accuracy and performance remain unclear
- This creates uncertainty due to lack of systematic benchmarks
- This paper conducts a comprehensive analysis to provide clear, data-driven guidelines

Research question

"What are the practical accuracy-performance trade-offs for popular quantization formats?"

Quantization formats

W8A8-FP

W8A8-INT

W4A16-INT

Quantization methods

- W8A8-FP: RTN
- W8A8-INT: GPTQ+SmoothQuant
- W4A16-INT: GPTQ

Experiments setup

- Model: Llama 3.1 (8B, 70B, 405B)
- Quantization formats: W8A8-FP / W8A8-INT / W4A16-INT
- Evaluations
  - Academic benchmarks
  - Real-world benchmarks
  - Text similarity analysis

Experiments setup

1. Academic benchmarks - Open LLM Leaderboard V1, V2

- V1 (GSM, MMLU, ARC-C, Winogrande, HellaSwag, TruthfulQA)
- V2 (MMLU-Pro, GPQA, BBH, MuSR, MATH Level 5, IFEval)

Experiments setup

2. Real-world benchmakrs for practical scenarios
    - Instruction following
    - Long-context
    - Code generation
- Arena-Hard-Auto-v0.1
- HumanEval / HumanEval+
- RULER

Experiments setup

3. Text similarity analysis

- ROUGE
- BERTScore
- Semantic Textual Similarity

Results

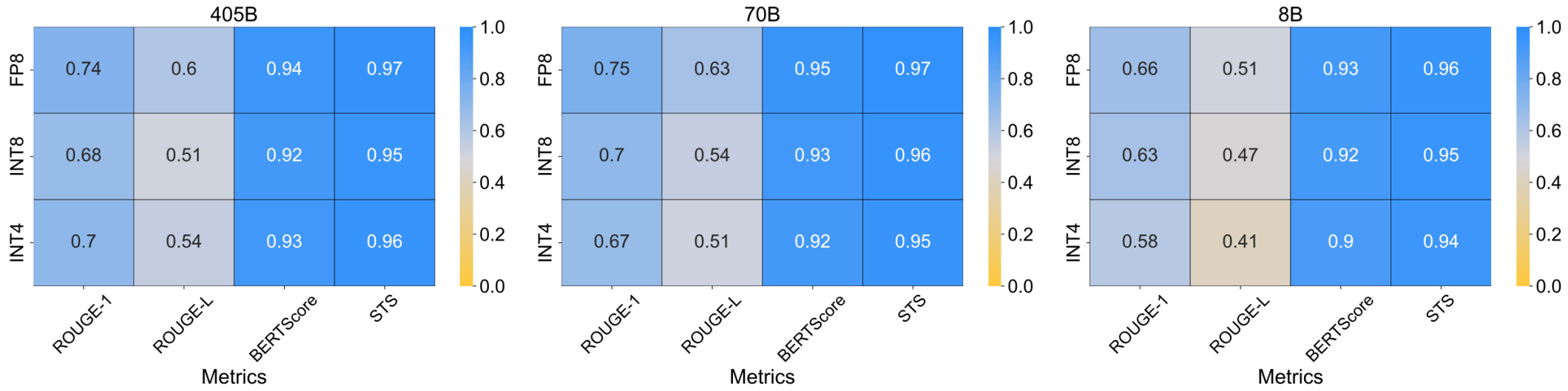
		Recovery %	Average Score	MMLU 5-shot	MMLU CoT 0-shot	ARC-C 0-shot	GSM8k CoT 8-shot	HellaSwag 10-shot	Winogrande 5-shot	TruthfulQA 0-shot
8B	BF16	100.00	74.06	68.3	72.8	81.4	82.8	80.5	78.1	54.5
	W8A8-FP	99.31	73.55	68.0	71.6	81.2	82.0	80.0	77.7	54.3
	W8A8-INT	100.31	74.29	67.8	72.2	81.7	84.8	80.3	78.5	54.7
	W4A16-INT	98.72	73.11	66.9	71.1	80.2	82.9	79.9	78.0	52.8
70B	BF16	100.00	84.40	83.8	86.0	93.3	94.9	86.8	85.3	60.7
	W8A8-FP	99.72	84.16	83.8	85.5	93.5	94.5	86.6	84.6	60.6
	W8A8-INT	99.87	84.29	83.7	85.8	93.1	94.2	86.7	85.1	61.4
	W4A16-INT	99.53	84.00	83.6	85.6	92.8	94.4	86.3	85.5	59.8
405B	BF16	100.00	86.79	87.4	88.1	95.0	96.0	88.5	87.2	65.3
	W8A8-FP	100.12	86.89	87.5	88.1	95.0	95.8	88.5	88.0	65.3
	W8A8-INT	99.32	86.20	87.1	87.7	94.4	95.5	88.2	86.1	64.4
	W4A16-INT	99.98	86.78	87.2	87.7	95.3	96.3	88.3	87.4	65.3

- All formats show high recovery rates
- W8A8-FP formats have near perfect re covery rates

			Academic Benchmarks (Open LLM Leaderboard V2)							Real-World Benchmarks			
		Recovery %	Average Score	IFEval 0-shot	BBH 3-shot	Math lvl 5 4-shot	GPQA 0-shot	MuSR 0-shot	MMLU-Pro 5-shot	Arena-Hard Win-Rate	HumanEval pass@1	HumanEval+ pass@1	RULER Score
8B	BF16	100.0	27.6	77.8	30.1	15.7	3.7	7.6	30.8	25.8	67.3	60.7	82.8
	W8A8-FP	101.2	27.9	77.2	29.6	16.5	5.7	7.5	31.2	26.8	67.3	61.3	82.8
	W8A8-INT	101.5	28.0	77.9	30.9	15.5	5.4	7.6	30.9	27.2	67.1	60.0	82.8
	W4A16-INT	96.1	26.5	76.3	28.9	14.8	4.1	6.3	28.8	24.0	67.1	59.1	81.1
70B	BF16	100.0	41.7	86.4	55.8	26.1	15.4	18.1	48.1	57.0	79.7	74.8	83.3
	W8A8-FP	100.0	41.7	87.6	54.9	28.0	14.6	17.2	47.7	57.7	80.0	75.0	83.0
	W8A8-INT	97.3	40.5	86.6	55.2	23.9	13.6	16.8	47.1	57.0	78.7	74.0	82.5
	W4A16-INT	97.4	40.6	85.7	55.0	24.4	13.8	17.2	47.2	56.3	80.5	74.2	82.2
405B	BF16	100.0	48.7	87.7	67.0	38.9	19.5	19.5	59.7	67.4	86.8	80.1	-
	W8A8-FP	99.9	48.7	86.8	67.1	38.8	18.9	20.8	59.4	66.9	87.0	81.0	-
	W8A8-INT	98.3	47.9	86.9	66.7	35.8	20.4	19.2	58.4	64.6	86.9	80.4	-
	W4A16-INT	98.9	48.2	88.0	67.5	37.6	17.5	19.4	59.3	66.5	85.1	78.9	-

- All models have recovery rates of at lea st 96%
- Smaller models have higher variance in GPQA, MuSR

Results



- Large quantized models (70B and 405B) are close with BF16 counterparts
- 8B models exhibit slightly higher variability – though they still maintain strong semantic fidelity
- Quantized models generate high-quality outputs across all sizes and schemes

Sync and async

- Synchronous: single query is processed at a time
- Asynchronous: multiple query is processed at a time (e.g., vLLM)

	Input tokens	Output tokens
Code completion	256	1024
Instruction following	256	128
Summarization	4096	512
Multi-turn chat	512	256
RAG	1024	128
Docstring generation	768	128
Code fixing	1024	1024

Sync and async

Size	GPU	#	Format	CR	Code Completion		Docstring Generation		Code Fixing		RAG		Instruction Following		Multi-Turn Chat		Summarization	
					Lat.	Q/$	Lat.	Q/$	Lat.	Q/$	Lat.	Q/$	Lat.	Q/$	Lat.	Q/$	Lat.	Q/$
8B	A6000	1	BF16	–	24.5	183	3.2	1,395	25.0	180	3.3	1,374	3.1	1,445	6.2	723	13.4	335
		1	INT8	1.54	15.9	284	2.1	2,157	16.3	276	2.1	2,139	2.0	2,249	4.0	1,120	8.9	506
		1	INT4	**2.39**	9.7	**462**	1.4	**3,290**	10.1	**445**	1.4	**3,136**	1.3	**3,543**	2.5	**1,787**	6.1	**736**
70B	A6000	4	BF16	–	61.7	18	6.6	170	62.6	18	8.1	138	8.0	141	15.8	71	32.6	35
		2	INT8	1.94	63.4	35	7.1	317	63.8	35	8.4	267	8.0	280	16.2	139	34.0	66
		2	INT4	**2.96**	39.2	**57**	5.0	**453**	40.4	**56**	5.8	**390**	5.1	**440**	10.2	**221**	23.5	**96**
	A100	2	BF16	–	50.7	20	2.9	343	51.2	20	6.8	148	6.4	156	12.9	78	27.3	37
		1	INT8	1.81	54.3	37	4.0	500	54.8	37	7.2	279	6.9	291	13.8	146	29.3	69
		1	INT4	**2.67**	35.0	**57**	2.8	**718**	35.8	**56**	5.2	**390**	4.6	**439**	9.2	**220**	21.0	**96**
	H100	2	BF16	–	31.3	18	4.0	139	31.5	18	4.1	138	4.0	142	7.9	71	16.4	34
		1	FP8	1.84	32.8	33	4.3	256	33.1	33	4.3	254	4.2	262	8.3	132	17.4	63
		1	INT4	**2.11**	28.6	**38**	3.8	**289**	28.2	**39**	3.8	**287**	3.7	**299**	7.1	**153**	15.3	**72**
405B	A100	16	BF16	–	81.9	2	10.8	12	81.2	2	11.2	11	10.6	12	20.9	6	44.1	3
		8	INT8	3.27	50.1	5	6.6	38	50.5	5	6.8	37	6.4	39	12.8	20	26.9	9
		4	INT4	**6.38**	48.9	**10**	7.0	**71**	49.5	**10**	7.3	**68**	6.4	**79**	12.7	**39**	29.4	**17**
	H100	16	BF16	–	50.6	1	6.5	12	50.3	1	6.6	11	6.4	12	13.0	6	26.5	3
		8	FP8	3.17	31.7	5	4.2	36	31.9	5	4.2	36	4.1	37	8.0	19	16.7	9
		4	INT4	**5.15**	37.5	**8**	5.0	**58**	37.8	**8**	5.1	**57**	4.8	**60**	9.2	**32**	20.4	**14**

†**CR**: Cost Reduction factor compared to BF16 baseline. Higher is better.
Lat.: Latency in seconds (lower is better). Q/$: Queries per USD (higher is better).

- Focus on latency and queries per USD
- INT4 shows lowest latency
- Synchronous task are memory-bound (how fast can it move from memory)
- INT4 is more efficient at lower latencies – ideal for applications requiring rapid response times
- INT4 is highly effective for synchronous deployment

Sync and async

Size	HW	Format	Speedup	Code Compl.		Doc. Gen.		Code Fixing		RAG		Inst. Following		Multi-Turn Chat		Summarization	
				QPS	Q/$	QPS	Q/$	QPS	Q/$	QPS	Q/$	QPS	Q/$	QPS	Q/$	QPS	Q/$
8B	1×A6000	BF16	–	1.5	6.8k	5.6	25.1k	1.1	4.8k	4.4	19.9k	11.8	53.0k	5.3	24.0k	0.7	3.2k
		INT8	**1.38**	2.2	9.8k	**7.7**	**34.6k**	**1.4**	**6.4k**	**6.1**	**27.6k**	**16.5**	**74.5k**	**7.2**	**32.3k**	**1.0**	**4.4k**
		INT4	1.08	**2.2**	**9.8k**	5.3	24.0k	1.3	6.0k	4.1	18.6k	11.2	50.5k	5.4	24.3k	0.7	3.1k
70B	4×A6000	BF16	–	0.4	0.4k	1.4	1.6k	0.3	0.3k	1.4	1.6k	3.3	3.8k	1.5	1.7k	0.2	0.3k
		INT8	1.91	0.7	0.8k	**3.9**	**4.4k**	0.5	0.6k	**2.8**	**3.1k**	**6.9**	**7.7k**	2.2	2.5k	**0.3**	**0.4k**
		INT4	**1.92**	**1.2**	**1.4k**	2.7	3.1k	**0.7**	**0.8k**	1.9	2.1k	5.2	5.9k	**2.6**	**3.0k**	0.3	0.3k
	4×A100	BF16	–	1.4	0.7k	6.9	3.5k	1.0	0.5k	3.3	1.6k	8.7	4.4k	4.3	2.2k	0.7	0.4k
		INT8	**1.87**	**2.4**	**1.2k**	15.9	8.0k	**1.8**	**0.9k**	**6.1**	**3.1k**	**16.5**	**8.3k**	**8.0**	**4.0k**	**1.2**	**0.6k**
		INT4	1.64	2.3	1.2k	**22.8**	**11.5k**	1.4	0.7k	4.3	2.2k	11.9	6.0k	5.8	2.9k	0.8	0.4k
	4×H100	BF16	–	3.5	1.0k	10.0	2.9k	2.6	0.7k	8.0	2.3k	20.3	5.9k	9.9	2.9k	1.7	0.5k
		FP8	**1.77**	**6.9**	**2.0k**	**17.8**	**5.2k**	**4.0**	**1.2k**	**14.3**	**4.2k**	**38.3**	**11.1k**	**18.4**	**5.4k**	**2.6**	**0.8k**
		INT4	1.55	5.9	1.7k	16.4	4.8k	3.1	0.9k	13.0	3.8k	35.8	10.4k	16.1	4.7k	2.2	0.6k
405B	16×A100	BF16	–	0.8	59	2.5	187	0.3	20	2.1	156	4.6	347	2.1	158	0.3	22
		INT8	**2.53**	1.3	98	**4.8**	**358**	1.1	79	**3.8**	**282**	**10.1**	**760**	**4.9**	**366**	**0.8**	**63**
		INT4	2.21	**1.9**	**144**	3.6	271	**1.2**	**93**	2.8	211	8.2	616	4.0	304	0.6	43
	16×H100	BF16	–	0.7	52	6.1	456	0.6	44	4.8	363	8.5	638	5.3	398	0.6	46
		FP8	3.04	**4.4**	**329**	9.6	725	**2.7**	**200**	7.6	571	20.7	1561	10.4	780	**1.7**	**125**
		INT4	**3.09**	4.0	304	**11.1**	**833**	2.5	192	**8.7**	**652**	**24.7**	**1856**	**11.6**	**872**	1.6	122

QPS: Queries per second (higher is better). Q/$: Queries per USD (higher is better).
Numbers denoted with *k* represent thousands (e.g., 20.3k = 20,300).

- Most INT8 and FP8 models show highes t query per second throughput
- Async task uses batching to process qu eries, thus compute-bound
- INT8 and FP8 models' weight and activa tions are both 8-bit (fast compute time)
- INT8 and FP8 is better suited for batch processing

Conclusion

- Experiments on FP8, INT8, INT4
- FP8 is nearly lossless, INT8 is effective, INT4 is competitive
- Quantized models produce similar text
- Optimal format depends on the use case (latency vs throughput)