FMBench Results

Generated: 2025-12-09 17:25

Scenario: ARC Challenge

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Llama-3.2-3B	0.23	0.017	2.0	0.90	400.3	15.0	10035	57.7	6903
Llama-3.2-1B-quantized.w8a8	0.62	0.018	2.0	0.50	592.2	16.5	10200	58.0	3351
Qwen2.5-1.5B	7.27	0.024	128.0	1.00	5060.1	16.9	9855	47.3	3722
Llama-3.2-1B	0.11	0.021	2.0	0.30	164.9	25.3	10442	42.0	3112
Qwen3-8B-quantized.w4a16	-	-	-	-	33.5	0.0	6734	0.0	7508
Qwen2.5-1.5B-quantized.w8a8	58.29	0.017	128.0	1.00	42744.0	15.7	10039	59.9	4371
Qwen2.5-7B-quantized.w8a8	-	-	-	-	3029.0	16.1	13852	84.2	7672
Qwen2.5-7B-Instruct	293.70	0.242	115.7	1.00	148884.0	14.9	13821	67.7	6166
Qwen3-8B	456.52	0.245	128.0	1.00	214239.7	15.5	14014	60.1	6228
falcon-7b-instruct	14.10	0.187	9.6	0.60	7034.3	15.8	12531	93.4	7123

Table 1: Results for ARC Challenge

Scenario: ARC Easy

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-8B-quantized.w4a16	-	-	-	-	38.5	0.0	6794	0.0	7505
Qwen2.5-1.5B	7.21	0.018	128.0	1.00	4930.4	16.5	9881	47.1	3720
Qwen3-8B	451.81	0.229	128.0	1.00	214915.5	16.0	13997	59.9	6237
Llama-3.2-1B-quantized.w8a8	0.60	0.036	2.0	0.60	570.4	17.6	10478	57.1	3351
Qwen2.5-7B-quantized.w8a8	-	-	-	-	2714.6	15.9	14209	78.6	7625
Qwen2.5-1.5B-quantized.w8a8	58.59	0.019	128.0	1.00	43527.2	16.0	9963	59.6	4372
Llama-3.2-1B	0.09	0.017	2.0	0.40	127.2	22.2	10437	34.5	3107
Qwen2.5-7B-Instruct	349.20	0.252	127.1	1.00	163531.8	14.9	13720	67.8	6170
Llama-3.2-3B	0.17	0.069	3.9	0.90	309.8	18.9	10030	54.3	6912
falcon-7b-instruct	11.26	0.229	7.8	0.60	5481.1	13.9	12353	91.5	7124

Table 2: Results for ARC Easy

Scenario: CountBenchQA

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
llava-1.5-7b-hf	0.045	0.60	5128.5	13.4	14373	86.6	6931
SmolVLM-256M-Instruct	0.014	0.60	1513.5	18.9	7790	82.7	1692

Table 3: Results for CountBenchQA

Scenario: GTSRB

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
llava-1.5-7b-hf	0.046	0.90	18876.3	13.7	14498	91.4	6928
SmolVLM-256M-Instruct	0.017	0.90	6982.8	17.3	7724	35.1	1683

Table 4: Results for GTSRB

Scenario: HaGRID

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.014	0.10	5449.4	16.2	7357	31.8	1525
llava-1.5-7b-hf	0.062	0.70	16454.0	15.1	14575	90.8	6910

Table 5: Results for HaGRID

Scenario: Idle Baseline

Model	Energy (J)	Idle Power (W)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
granite-timeseries-patchtst	334.7	33.5	9.8	5623	9.0	525
Qwen2.5-7B-Instruct	87.1	8.7	15.1	11630	1.8	5742
Qwen2.5-1.5B-quantized.w8a8	368.2	36.8	10.7	10006	1.0	2846
SmolVLM-256M-Instruct	433.4	43.3	12.5	6775	0.0	1034
Llama-3.2-1B-quantized.w8a8	250.1	25.0	18.9	10003	8.6	2661
arima	81.5	8.2	12.0	5991	1.0	433
Qwen2.5-1.5B	375.9	37.6	14.2	9883	1.0	3716
falcon-7b-instruct	405.4	40.5	9.8	12591	2.0	7109
llava-1.5-7b-hf	440.5	44.1	12.7	8276	0.0	6437
Qwen2.5-7B-quantized.w8a8	232.4	23.2	11.3	10476	0.0	6083
Llama-3.2-1B	443.9	44.4	17.5	10057	5.2	3102
Qwen3-8B	73.6	7.4	10.3	12205	3.0	5814
Llama-3.2-3B	114.7	11.5	12.9	10277	1.6	6880
Qwen3-8B-quantized.w4a16	217.5	21.7	7.4	7079	0.0	7514
chronos-t5-small	298.3	29.8	12.7	5880	0.0	679

Table 6: Results for Idle Baseline

Scenario: M3 Monthly Forecasting

Model	sMAPE (%)	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
granite-timeseries-patchtst	200.00	33.8	0.0	5863	0.0	530
arima	35.37	23.1	10.6	6235	1.0	435
chronos-t5-small	34.75	519.9	19.2	6426	32.5	704

Table 7: Results for M3 Monthly Forecasting

Scenario: VQAv2

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
llava-1.5-7b-hf	0.062	0.00	4185.6	17.5	13889	85.9	6955
SmolVLM-256M-Instruct	0.015	0.70	2106.1	17.5	7831	43.5	1515

Table 8: Results for VQAv2

Scenario: classification

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
falcon-7b-instruct	36.24	0.194	24.5	0.10	17111.8	13.3	12912	93.2	7122
Qwen3-8B	447.31	0.235	128.0	0.60	210400.0	15.2	13982	61.3	6240
Llama-3.2-3B	5.18	0.028	91.4	0.20	5201.3	15.5	10143	79.8	6926
Qwen2.5-1.5B	0.87	0.017	19.5	0.50	897.0	24.2	9916	49.8	3739
Llama-3.2-1B	4.56	0.019	116.8	0.40	3223.9	17.8	10691	53.9	3122
Qwen3-8B-quantized.w4a16	-	-	-	-	42.3	0.0	6697	0.0	7498
Llama-3.2-1B-quantized.w8a8	38.79	0.024	128.0	0.50	29488.0	17.4	10319	67.2	3461
Qwen2.5-7B-Instruct	246.83	0.216	105.0	0.70	100532.3	15.4	15202	89.3	6001
Qwen2.5-1.5B-quantized.w8a8	1.33	0.019	13.2	0.40	4935.9	17.5	10077	60.2	4357
Qwen2.5-7B-quantized.w8a8	-	-	-	-	3384.0	14.0	13912	86.9	7683

Table 9: Results for classification

Scenario: docvqa

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
llava-1.5-7b-hf	0.043	0.20	11443.0	13.3	14766	91.6	7005
SmolVLM-256M-Instruct	0.013	0.60	1885.5	16.8	7714	68.0	1647

Table 10: Results for docvqa

Scenario: ner

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Llama-3.2-3B	3.70	0.025	80.0	0.60	4353.5	15.7	10122	79.8	6916
Qwen3-8B	457.57	0.267	128.0	1.00	216912.0	15.8	14035	60.1	6226
falcon-7b-instruct	92.68	0.206	60.4	0.70	40746.1	14.6	13099	94.0	7122
Qwen3-8B-quantized.w4a16	-	-	-	-	42.4	0.0	6766	0.0	7505
Llama-3.2-1B	2.26	0.024	71.8	0.35	1930.9	18.0	10713	52.8	3108
Llama-3.2-1B-quantized.w8a8	6.55	0.019	26.4	0.20	6245.6	14.9	10348	66.6	3453
Qwen2.5-7B-Instruct	340.35	0.192	126.2	0.80	161042.7	15.3	13840	68.1	6168
Qwen2.5-1.5B	1.10	0.019	24.8	0.70	993.2	17.4	9923	49.5	3734
Qwen2.5-1.5B-quantized.w8a8	5.40	0.017	18.2	0.60	6563.5	15.2	10066	59.3	4360
Qwen2.5-7B-quantized.w8a8	-	-	-	-	2641.2	16.7	13574	86.7	7629

Table 11: Results for ner

Scenario: perplexity_c4

Model	Perplexity	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen2.5-7B-Instruct	21.93	3263.1	15.8	14566	68.7	6139
Qwen3-8B	29.99	3806.6	17.8	13838	63.2	6326
Llama-3.2-1B-quantized.w8a8	23.84	762.4	14.6	10285	68.7	4531
Qwen2.5-1.5B-quantized.w8a8	27.20	812.0	16.4	9991	78.3	5148
Qwen2.5-1.5B	24.96	282.8	18.7	9806	92.0	4871
falcon-7b-instruct	20.20	2404.8	14.3	13309	86.0	7157
Llama-3.2-1B	23.78	305.9	16.4	10440	64.7	4332
Qwen3-8B-quantized.w4a16	-	42.4	0.0	6760	0.0	7498
Llama-3.2-3B	19.73	736.0	18.2	10262	79.1	7783
Qwen2.5-7B-quantized.w8a8	-	4042.4	18.9	13085	91.9	7829

Table 12: Results for perplexity_c4

Scenario: perplexity_wikitext2

Model	Perplexity	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen2.5-1.5B	700.37	12.0	0.0	9560	1.0	3705
Qwen2.5-1.5B-quantized.w8a8	2109.38	212.6	17.7	9917	49.3	3901
falcon-7b-instruct	132.72	805.6	14.9	12381	71.7	7128
Llama-3.2-3B	248.85	130.0	13.1	9796	51.0	6987
Qwen2.5-7B-Instruct	1029.40	1696.0	18.9	14805	30.0	5984
Llama-3.2-1B	495.51	44.1	0.0	10219	0.0	3090
Qwen3-8B	174.36	1682.3	19.6	15078	29.8	5916
Llama-3.2-1B-quantized.w8a8	517.40	141.7	12.9	10135	75.0	3100
Qwen2.5-7B-quantized.w8a8	-	1187.7	18.4	12956	75.6	7702
Qwen3-8B-quantized.w4a16	-	42.7	0.0	6744	0.0	7498

Table 13: Results for perplexity_wikitext2

Scenario: sentiment

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen2.5-7B-Instruct	214.65	0.223	108.0	1.00	106645.8	18.0	15246	88.7	6013
Llama-3.2-3B	1.64	0.042	27.2	0.40	2177.0	20.9	10181	88.1	7014
Qwen2.5-1.5B-quantized.w8a8	1.05	0.029	2.0	0.80	1042.1	14.7	10045	70.0	4245
Llama-3.2-1B-quantized.w8a8	38.15	0.030	128.0	1.00	28682.6	15.2	10217	67.1	3456
Qwen2.5-1.5B	0.25	0.016	2.0	0.80	249.0	18.2	9809	54.0	3739
Llama-3.2-1B	3.15	0.029	78.0	0.70	2302.0	17.9	10676	58.5	3224
Qwen3-8B	293.81	0.235	128.0	0.80	142446.1	15.1	15609	87.3	6118
falcon-7b-instruct	32.78	0.185	17.4	0.30	13201.9	13.9	13001	94.1	7122
Qwen2.5-7B-quantized.w8a8	-	-	-	-	3698.7	16.5	13920	89.4	7718
Qwen3-8B-quantized.w4a16	-	-	-	-	42.6	0.0	6755	0.0	7505

Table 14: Results for sentiment

Scenario: summarization

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Llama-3.2-1B-quantized.w8a8	38.33	0.021	128.0	0.00	28785.1	15.9	10511	68.1	3546
Qwen2.5-1.5B	7.42	0.017	126.6	0.00	5794.1	15.2	10083	53.5	3921
Qwen3-8B	283.46	0.246	128.0	0.00	139126.2	16.4	15409	88.1	6510
Qwen2.5-1.5B-quantized.w8a8	55.90	0.017	122.8	0.00	41840.4	15.7	10163	60.9	4425
Qwen2.5-7B-quantized.w8a8	-	-	-	-	8292.2	13.8	13487	96.7	7888
Llama-3.2-1B	4.73	0.018	128.0	0.00	4228.2	15.6	10592	61.8	3568
Qwen3-8B-quantized.w4a16	-	-	-	-	42.8	0.0	6799	0.0	7505
Qwen2.5-7B-Instruct	250.98	0.235	125.5	0.00	123580.6	16.5	15354	89.4	6219
falcon-7b-instruct	133.90	0.190	102.9	0.00	71183.9	14.6	13698	94.3	7132
Llama-3.2-3B	7.75	0.033	114.8	0.00	7366.7	19.0	10380	86.6	7331

Table 15: Results for summarization

Scenario: translation

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-8B	425.61	0.268	128.0	0.00	197221.2	16.0	14298	63.8	6216
Llama-3.2-3B	7.10	0.022	123.5	0.10	6239.2	19.5	10039	80.0	6898
Qwen3-8B-quantized.w4a16	-	-	-	-	42.8	0.0	6812	0.0	7505
Qwen2.5-7B-quantized.w8a8	-	-	-	-	2567.9	15.1	13654	84.2	7633
Qwen2.5-1.5B	1.24	0.037	19.3	0.00	888.6	16.0	9938	44.2	3723
Qwen2.5-1.5B-quantized.w8a8	9.64	0.018	19.2	0.00	7027.9	14.9	10104	59.5	4352
Llama-3.2-1B	3.93	0.021	118.8	0.00	3148.9	16.9	10510	53.7	3106
Llama-3.2-1B-quantized.w8a8	38.07	0.033	128.0	0.00	28353.2	15.7	10214	66.7	3452
falcon-7b-instruct	26.88	0.175	18.4	0.10	12341.6	13.6	12781	92.5	7123
Qwen2.5-7B-Instruct	199.26	0.226	102.8	0.10	100209.7	14.4	15347	89.6	6064

Table 16: Results for translation