FMBench Results

CPU: AMD64 Family 25 Model 33 Stepping 2, AuthenticAMD
GPU: NVIDIA GeForce RTX 3070 (GPU 0)
Generated: 2025-12-09 16:09:02

Benchmark Summary

Evaluated Models

Model Name
Llama-3.2-1B
Llama-3.2-1B-quantized.w8a8
Qwen2.5-1.5B
Qwen2.5-1.5B-quantized.w8a8
Qwen2.5-7B-Instruct
Qwen2.5-7B-quantized.w8a8
Qwen3-0.6B
Qwen3-4B
Qwen3-4B-quantized.w4a16
SmolVLM-256M-Instruct
arima
chronos-t5-small
granite-timeseries-patchtst
llava-1.5-7b-hf

Scenario Configurations

Scenario	Flags
ARC Challenge	num samples: 10
ARC Easy	num samples: 10
CountBenchQA	num samples: 10
FEV-Bench	
GIFT-EVAL	
GTSRB	num samples: 10
HaGRID	num samples: 10
Idle Baseline	idle duration: 10
M3 Monthly Forecasting	num samples: 10
VQAv2	num samples: 10
classification	num samples: 10
docvqa	num samples: 10
ner	num samples: 10
perplexity_c4	num samples: 10
perplexity_wikitext2	num samples: 10
sentiment	num samples: 10
summarization	num samples: 10
	use expensive metrics: True
translation	num samples: 10
	use expensive metrics: True

LLM Scenarios

ARC Challenge

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	11.00	0.015	128.0	1.00	5884.0	11.3	10148	13.4	2577
Llama-3.2-1B	0.10	0.013	2.0	0.60	101.5	10.7	9773	0.0	3475
Llama-3.2-1B-quantized.w8a8	0.37	0.012	2.0	0.50	317.0	11.6	9952	37.0	3638
Qwen2.5-1.5B	8.71	0.012	128.0	1.00	5012.8	11.1	10139	21.5	4110
Qwen2.5-1.5B-quantized.w8a8	45.92	0.012	128.0	1.00	37980.2	11.2	10034	30.7	4763
Qwen3-4B	31.67	0.037	128.0	1.00	23930.1	10.3	11635	44.8	7288
Qwen3-4B-quantized.w4a16	92.92	0.016	128.0	1.00	89902.4	7.5	12726	93.0	8167
Qwen2.5-7B-Instruct	130.13	0.142	128.0	1.00	89817.7	9.4	19499	54.1	5770
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1117.1	8.6	16218	77.5	7404

Table 1: Results for ARC Challenge

ARC Easy

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	10.56	0.015	128.0	1.00	5845.7	11.6	10146	13.3	2573
Llama-3.2-1B	0.10	0.012	2.0	0.60	105.1	11.2	9759	1.0	3473
Llama-3.2-1B-quantized.w8a8	0.47	0.012	2.0	0.60	411.3	10.6	10005	42.0	3679
Qwen2.5-1.5B	8.60	0.012	128.0	1.00	5082.0	11.1	10137	21.5	4110
Qwen2.5-1.5B-quantized.w8a8	40.86	0.012	128.0	1.00	36836.8	11.4	10078	31.8	4763
Qwen3-4B	28.84	0.041	121.9	1.00	22642.7	10.3	11647	45.8	7278
Qwen3-4B-quantized.w4a16	92.68	0.016	128.0	1.00	89363.1	7.7	12733	93.0	8167
Qwen2.5-7B-Instruct	128.87	0.129	125.3	1.00	88312.2	9.4	19512	53.6	5792
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1098.7	8.1	16167	72.1	7287

Table 2: Results for ARC Easy

classification

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	10.10	0.016	128.0	0.90	5770.7	11.6	10153	14.4	2594
Llama-3.2-1B	3.66	0.013	116.8	0.60	2869.4	12.3	10204	31.3	3502
Llama-3.2-1B-quantized.w8a8	27.78	0.012	128.0	0.40	24779.3	10.7	10152	34.7	3843
Qwen2.5-1.5B	0.84	0.019	19.3	0.40	731.6	11.8	10095	20.2	4110
Qwen2.5-1.5B-quantized.w8a8	2.82	0.012	13.1	0.30	3813.8	11.7	10026	33.8	4730
Qwen3-4B	31.55	0.036	128.0	0.50	24039.7	10.4	11662	45.8	7300
Qwen3-4B-quantized.w4a16	90.79	0.016	128.0	0.60	88801.0	7.6	12743	93.1	8167
Qwen2.5-7B-Instruct	129.70	0.141	118.5	0.70	83484.5	9.5	19510	54.0	5801
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1377.0	8.3	16464	81.4	7615

Table 3: Results for classification

ner

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	10.40	0.016	128.0	0.80	5616.6	11.7	10186	14.2	2573
Llama-3.2-1B	2.34	0.012	64.5	0.30	1510.7	11.6	10174	29.1	3488
Llama-3.2-1B-quantized.w8a8	4.42	0.016	26.4	0.20	4903.5	11.2	10163	38.5	3828
Qwen2.5-1.5B	0.93	0.012	24.8	0.70	812.4	12.8	10137	27.3	4108
Qwen2.5-1.5B-quantized.w8a8	6.86	0.012	25.7	0.70	7518.8	11.3	10069	31.6	4748
Qwen3-4B	36.23	0.037	128.0	0.90	26020.5	10.0	11665	40.4	7283
Qwen3-4B-quantized.w4a16	93.18	0.016	128.0	0.90	90104.0	7.6	12744	92.4	8167
Qwen2.5-7B-Instruct	121.13	0.129	121.7	0.80	85599.6	9.4	19374	53.5	5800
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1027.6	8.5	16257	71.2	7627

Table 4: Results for ner

perplexity_c4

Model	Perplexity	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	60.38	102.7	8.7	9637	0.0	3575
Llama-3.2-1B	23.78	49.2	0.0	9377	0.0	3460
Llama-3.2-1B-quantized.w8a8	23.84	215.8	9.8	9864	65.5	4702
Qwen2.5-1.5B	24.97	116.4	8.4	9767	32.0	5179
Qwen2.5-1.5B-quantized.w8a8	27.32	318.2	11.7	9807	44.7	5225
Qwen3-4B	42.56	383.5	15.3	11963	46.0	7078
Qwen3-4B-quantized.w4a16	45.13	3552.7	7.9	13181	98.9	8061
Qwen2.5-7B-Instruct	21.92	968.9	8.7	19076	60.0	5755
Qwen2.5-7B-quantized.w8a8	-	1142.0	8.8	16180	81.6	7667

Table 5: Results for perplexity_c4

perplexity_wikitext2

Model	Perplexity	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	1437.28	48.9	0.0	9301	0.0	2528
Llama-3.2-1B	494.60	49.2	0.0	9340	0.0	3460
Llama-3.2-1B-quantized.w8a8	608.76	112.1	9.0	9719	1.0	3478
Qwen2.5-1.5B	698.49	45.9	0.0	9432	0.0	4090
Qwen2.5-1.5B-quantized.w8a8	2080.65	173.6	8.7	9728	22.5	4291
Qwen3-4B	373.75	181.5	8.4	11705	25.5	6748
Qwen3-4B-quantized.w4a16	293.20	1219.1	9.4	11498	91.8	7878
Qwen2.5-7B-Instruct	1032.62	379.9	10.0	19048	46.0	5619
Qwen2.5-7B-quantized.w8a8	-	475.0	9.6	15451	62.2	7316

Table 6: Results for perplexity_wikitext2

sentiment

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	9.19	0.015	128.0	1.00	5337.6	12.2	10141	17.4	2609
Llama-3.2-1B	2.94	0.013	102.5	0.90	2649.9	12.0	10179	31.9	3592
Llama-3.2-1B-quantized.w8a8	25.68	0.013	128.0	1.00	24597.6	11.0	10168	37.1	3839
Qwen2.5-1.5B	0.11	0.012	2.0	0.80	103.4	10.7	9789	41.0	4113
Qwen2.5-1.5B-quantized.w8a8	2.90	0.012	6.1	0.80	1869.5	11.4	10028	34.0	4696
Qwen3-4B	36.32	0.038	128.0	0.90	27034.2	10.0	12171	40.5	6740
Qwen3-4B-quantized.w4a16	91.50	0.018	128.0	0.80	91292.5	7.7	12799	93.4	8165
Qwen2.5-7B-Instruct	106.96	0.129	115.3	1.00	80811.9	9.4	19358	55.3	5790
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1384.3	8.4	16266	86.5	7620

Table 7: Results for sentiment

summarization

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	10.08	0.015	128.0	0.00	6798.3	10.6	10377	18.0	3255
Llama-3.2-1B	4.84	0.012	128.0	0.00	4736.4	9.8	10362	30.4	4576
Llama-3.2-1B-quantized.w8a8	27.91	0.012	128.0	0.00	25792.4	10.7	10426	36.8	4509
Qwen2.5-1.5B	7.73	0.012	126.6	0.00	6017.7	10.1	10377	24.4	5061
Qwen2.5-1.5B-quantized.w8a8	43.69	0.012	128.0	0.00	39242.6	10.9	10308	31.3	5202
Qwen3-4B	37.80	0.036	128.0	0.00	29410.8	9.8	12138	42.3	6807
Qwen3-4B-quantized.w4a16	109.26	0.016	128.0	0.00	111528.8	7.3	13665	96.5	8166
Qwen2.5-7B-Instruct	108.86	0.130	125.2	0.00	80361.1	9.4	19549	52.8	6673
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1942.3	8.2	16534	87.0	7779

Table 8: Results for summarization

translation

Model	Latency (s)	TTFT (s)	Tok/Out	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	10.18	0.016	128.0	0.10	6620.4	11.3	12750	15.8	2653
Llama-3.2-1B	4.48	0.014	128.0	0.00	4188.7	11.4	12830	31.6	3679
Llama-3.2-1B-quantized.w8a8	27.15	0.012	128.0	0.00	24523.6	11.2	12749	36.7	3756
Qwen2.5-1.5B	1.47	0.012	19.3	0.00	1707.3	10.3	12835	27.0	4570
Qwen2.5-1.5B-quantized.w8a8	6.88	0.012	19.1	0.00	6215.9	11.4	12730	34.8	4609
Qwen3-4B	38.52	0.036	128.0	0.10	27557.6	9.7	14481	39.0	6740
Qwen3-4B-quantized.w4a16	93.14	0.015	128.0	0.10	91119.3	7.6	15095	92.3	8163
Qwen2.5-7B-Instruct	107.29	0.131	95.8	0.10	69319.9	9.1	21952	52.8	5829
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1025.8	8.5	16329	63.4	7639

Table 9: Results for translation

VLM Scenarios

CountBenchQA

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.012	0.60	617.7	9.3	9795	38.9	1554
llava-1.5-7b-hf	0.032	0.60	2197.4	9.0	18084	53.3	6913

Table 10: Results for CountBenchQA

GTSRB

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.011	0.90	7276.0	11.5	9856	15.3	1645
llava-1.5-7b-hf	0.033	0.90	11669.3	10.2	18517	53.8	6913

Table 11: Results for GTSRB

HaGRID

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.014	0.10	6594.0	12.2	9934	13.5	1487
llava-1.5-7b-hf	0.034	0.70	10860.8	8.4	18114	47.7	6895

Table 12: Results for HaGRID

VQAv2

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.011	0.70	1666.5	11.7	10005	15.0	1470
llava-1.5-7b-hf	0.033	0.00	1707.1	9.3	17909	53.3	6891

Table 13: Results for VQAv2

docvqa

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.016	0.60	959.0	9.9	9837	48.4	1593
llava-1.5-7b-hf	0.035	0.20	7149.4	9.4	18509	54.1	6852

Table 14: Results for docvqa

Time-Series Scenarios

FEV-Bench

Model	sMAPE (%)	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
granite-timeseries-patchtst	11.24	46.3	0.0	8967	0.0	474
chronos-t5-small	200.00	45.8	0.0	8981	0.0	622
arima	200.00	45.9	0.0	8870	0.0	293

Table 15: Results for FEV-Bench

GIFT-EVAL

Model	sMAPE (%)	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
granite-timeseries-patchtst	-	45.9	0.0	9148	0.0	474
chronos-t5-small	200.00	45.6	0.0	9165	0.0	622
arima	200.00	45.8	0.0	9053	0.0	293

Table 16: Results for GIFT-EVAL

M3 Monthly Forecasting

Model	sMAPE (%)	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
granite-timeseries-patchtst	200.00	45.9	0.0	8953	0.0	474
chronos-t5-small	200.00	45.7	0.0	8959	0.0	622
arima	200.00	46.0	0.0	8858	0.0	293

Table 17: Results for M3 Monthly Forecasting

Baseline Scenarios

Idle Baseline

Model	Energy (J)	Idle Power (W)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	487.2	48.7	3.6	9504	14.0	2561
Llama-3.2-1B	492.0	49.2	0.3	9316	2.0	3460
Llama-3.2-1B-quantized.w8a8	496.1	49.6	0.3	9342	2.0	3022
Qwen2.5-1.5B	475.0	47.5	0.5	9379	2.0	4090
Qwen2.5-1.5B-quantized.w8a8	459.5	46.0	0.2	9392	2.0	3224
Qwen3-4B	473.2	47.3	0.2	10636	0.0	7174
Qwen3-4B-quantized.w4a16	461.3	46.1	0.3	9412	0.0	3770
Qwen2.5-7B-Instruct	477.9	47.8	0.3	18583	0.0	5488
Qwen2.5-7B-quantized.w8a8	462.1	46.2	0.3	11614	0.0	5766
SmolVLM-256M-Instruct	465.5	46.6	0.3	8989	2.0	980
llava-1.5-7b-hf	460.0	46.0	0.4	10111	0.0	6346
granite-timeseries-patchtst	465.5	46.5	0.6	8931	7.0	474
chronos-t5-small	460.2	46.0	0.2	8942	1.0	622
arima	460.6	46.1	0.3	8850	0.0	293

Table 18: Results for Idle Baseline