HoliBench Results

CPU: AMD64 Family 25 Model 33 Stepping 2, AuthenticAMD
GPU: NVIDIA GeForce RTX 3070 (GPU 0)
Generated: 2025-12-13 17:36:22

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
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	10.86	0.016	128.0	1.00	5929.8	11.2	8999	12.9	2244
Llama-3.2-1B	0.09	0.016	2.0	0.60	108.9	9.1	8586	24.0	3142
Llama-3.2-1B-quantized.w8a8	0.49	0.016	2.0	0.50	400.8	10.9	8867	32.0	3346
Qwen2.5-1.5B	8.11	0.016	128.0	1.00	4861.4	11.1	8754	22.5	3776
Qwen2.5-1.5B-quantized.w8a8	45.55	0.016	128.0	1.00	37156.0	11.3	8878	30.7	4429
Qwen3-4B	30.30	0.031	128.0	1.00	23936.6	9.8	10545	47.3	7109
Qwen3-4B-quantized.w4a16	97.32	0.016	128.0	1.00	94518.1	7.3	11820	94.3	8107
Qwen2.5-7B-Instruct	110.78	0.047	122.6	1.00	82734.5	9.1	18031	58.7	6121
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1124.6	8.1	15097	79.3	7493

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	10.16	0.016	128.0	1.00	5548.6	11.3	9001	14.8	2240
Llama-3.2-1B	0.07	0.000	2.0	0.50	102.8	11.4	8564	2.0	3140
Llama-3.2-1B-quantized.w8a8	0.47	0.016	2.0	0.60	384.4	10.8	8851	28.0	3346
Qwen2.5-1.5B	7.75	0.016	128.0	1.00	4819.4	11.1	8743	22.7	3776
Qwen2.5-1.5B-quantized.w8a8	44.76	0.000	128.0	1.00	37250.3	11.0	8875	29.7	4429
Qwen3-4B	30.15	0.031	128.0	1.00	23625.6	9.7	10609	47.6	7104
Qwen3-4B-quantized.w4a16	91.44	0.016	124.6	1.00	90721.1	7.6	11892	93.8	8138
Qwen2.5-7B-Instruct	119.21	0.032	124.4	1.00	83789.8	9.2	18276	58.7	6074
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1178.0	7.8	15081	80.1	7337

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	10.91	0.016	128.0	1.00	6018.8	11.2	9002	13.6	2260
Llama-3.2-1B	4.09	0.016	112.8	0.30	2871.9	11.3	8982	28.0	3169
Llama-3.2-1B-quantized.w8a8	27.17	0.016	128.0	0.40	24527.4	10.7	8995	35.5	3510
Qwen2.5-1.5B	0.83	0.016	19.3	0.40	743.5	12.0	8704	20.6	3776
Qwen2.5-1.5B-quantized.w8a8	2.63	0.000	13.1	0.30	3640.8	11.7	8890	34.6	4396
Qwen3-4B	30.29	0.047	128.0	0.50	23419.3	9.9	10582	47.8	7113
Qwen3-4B-quantized.w4a16	94.55	0.016	128.0	0.50	92233.3	7.3	11670	94.0	8143
Qwen2.5-7B-Instruct	119.72	0.047	119.3	0.70	81611.2	9.4	17942	58.3	6132
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1371.3	8.1	15314	82.2	7667

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	10.60	0.016	128.0	0.80	6078.9	11.0	8989	13.1	2240
Llama-3.2-1B	0.70	0.016	14.4	0.10	319.8	12.2	8841	38.0	3149
Llama-3.2-1B-quantized.w8a8	4.19	0.016	26.4	0.20	4564.0	11.5	8994	41.1	3493
Qwen2.5-1.5B	1.05	0.016	24.8	0.70	919.4	12.7	8720	22.9	3774
Qwen2.5-1.5B-quantized.w8a8	7.35	0.000	25.7	0.70	7707.7	10.4	8926	28.2	4416
Qwen3-4B	30.19	0.047	128.0	1.00	22900.4	9.7	10558	47.3	7096
Qwen3-4B-quantized.w4a16	97.26	0.016	128.0	0.90	93813.8	7.5	11764	93.7	8137
Qwen2.5-7B-Instruct	97.33	0.047	101.7	0.80	70766.5	9.1	17953	57.1	6113
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1116.1	8.2	15209	65.6	7606

Table 4: Results for ner

perplexity_c4

Model	Perplexity	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
Qwen3-0.6B	60.38	112.0	8.0	8482	55.0	3242
Llama-3.2-1B	23.78	49.1	0.0	8177	0.0	3127
Llama-3.2-1B-quantized.w8a8	23.84	208.3	9.8	8709	27.0	4133
Qwen2.5-1.5B	24.97	102.8	8.0	8393	95.0	4845
Qwen2.5-1.5B-quantized.w8a8	27.32	292.0	10.6	8661	44.7	4801
Qwen3-4B	42.56	389.1	8.7	10520	86.5	7043
Qwen3-4B-quantized.w4a16	45.13	3829.0	8.2	12246	96.6	8051
Qwen2.5-7B-Instruct	21.92	946.4	9.1	17639	59.2	6098
Qwen2.5-7B-quantized.w8a8	-	1212.2	8.3	15179	78.8	7615

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	49.1	0.0	8129	0.0	2195
Llama-3.2-1B	494.60	49.1	0.0	8135	0.0	3127
Llama-3.2-1B-quantized.w8a8	608.76	108.3	8.4	8560	32.0	3145
Qwen2.5-1.5B	698.49	45.6	0.0	8062	0.0	3756
Qwen2.5-1.5B-quantized.w8a8	2080.65	97.0	8.2	8468	3.0	3660
Qwen3-4B	373.75	97.3	8.1	10352	5.0	6665
Qwen3-4B-quantized.w4a16	293.20	1348.3	9.0	10547	91.8	7860
Qwen2.5-7B-Instruct	1032.62	415.3	9.6	17571	54.6	5996
Qwen2.5-7B-quantized.w8a8	-	459.9	8.6	14050	68.6	6991

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	10.79	0.016	128.0	1.00	5916.4	11.2	9007	15.4	2276
Llama-3.2-1B	2.42	0.016	67.5	0.70	1766.5	11.3	8968	33.2	3257
Llama-3.2-1B-quantized.w8a8	27.05	0.016	128.0	1.00	24571.7	10.7	8999	35.7	3506
Qwen2.5-1.5B	0.11	0.016	2.0	0.80	111.8	10.4	8419	25.0	3779
Qwen2.5-1.5B-quantized.w8a8	3.01	0.000	6.1	0.80	1856.5	11.0	8904	30.9	4366
Qwen3-4B	29.97	0.031	128.0	0.90	23207.8	9.9	10947	48.9	6726
Qwen3-4B-quantized.w4a16	94.67	0.016	128.0	0.80	95056.0	7.6	11891	93.9	8135
Qwen2.5-7B-Instruct	112.58	0.047	121.7	1.00	84874.6	9.2	17908	58.0	6152
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1209.1	8.3	14970	75.2	7585

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	10.81	0.016	128.0	0.00	7394.7	9.9	9226	17.2	2930
Llama-3.2-1B	4.49	0.016	116.0	0.00	4304.5	9.6	9213	29.3	4247
Llama-3.2-1B-quantized.w8a8	25.97	0.016	128.0	0.00	25215.2	10.5	9060	36.6	4166
Qwen2.5-1.5B	7.98	0.016	126.6	0.00	6061.0	9.8	8998	23.0	4717
Qwen2.5-1.5B-quantized.w8a8	42.99	0.000	128.0	0.00	38561.1	11.4	9308	32.9	5000
Qwen3-4B	31.60	0.031	128.0	0.00	25296.6	9.5	10937	50.6	6793
Qwen3-4B-quantized.w4a16	115.10	0.016	128.0	0.00	117586.1	7.2	12620	96.9	8112
Qwen2.5-7B-Instruct	106.93	0.047	126.1	0.00	80629.9	9.1	18024	55.2	7015
Qwen2.5-7B-quantized.w8a8	-	-	-	-	2166.6	7.7	15423	90.2	7692

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	8.81	0.016	128.0	0.10	6214.4	11.6	11444	18.1	2348
Llama-3.2-1B	4.95	0.016	123.3	0.00	4687.5	9.9	11623	26.5	3432
Llama-3.2-1B-quantized.w8a8	28.05	0.016	128.0	0.00	25084.6	10.8	11345	35.7	3441
Qwen2.5-1.5B	1.67	0.000	19.3	0.00	1790.1	9.8	11367	27.6	4195
Qwen2.5-1.5B-quantized.w8a8	6.90	0.016	19.1	0.00	6280.0	12.1	11727	34.4	4444
Qwen3-4B	30.27	0.031	128.0	0.10	24183.4	9.6	13341	48.0	6732
Qwen3-4B-quantized.w4a16	98.31	0.016	128.0	0.10	96329.1	7.4	14298	93.1	8143
Qwen2.5-7B-Instruct	84.83	0.031	87.5	0.10	58824.6	9.0	20240	57.4	6144
Qwen2.5-7B-quantized.w8a8	-	-	-	-	1105.8	8.1	15036	68.6	7368

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	0.016	0.60	520.3	8.6	8251	32.2	1816
llava-1.5-7b-hf	0.031	0.60	2073.4	8.6	16555	57.1	7172

Table 10: Results for CountBenchQA

GTSRB

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.000	0.90	6975.3	11.4	8310	14.4	1928
llava-1.5-7b-hf	0.031	0.90	10983.0	9.0	17056	57.4	7194

Table 11: Results for GTSRB

HaGRID

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.016	0.10	6744.8	11.4	8419	12.9	1770
llava-1.5-7b-hf	0.031	0.70	10134.7	9.3	16582	51.0	7174

Table 12: Results for HaGRID

VQAv2

Model	TTFT (s)	Accuracy	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
SmolVLM-256M-Instruct	0.016	0.70	1650.6	11.6	8479	22.1	1753
llava-1.5-7b-hf	0.031	0.00	1621.3	9.4	16390	56.1	7154

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	945.0	10.3	8346	40.4	1876
llava-1.5-7b-hf	0.047	0.20	6843.1	8.9	17040	55.3	7146

Table 14: Results for docvqa

Time-Series Scenarios

M3 Monthly Forecasting

Model	sMAPE (%)	Energy (J)	CPU Util (%)	RAM (MB)	GPU Util (%)	VRAM (MB)
granite-timeseries-patchtst	200.00	45.5	0.0	7517	0.0	757
chronos-t5-small	200.00	45.2	0.0	7487	0.0	905
arima	200.00	45.5	0.0	7376	0.0	576

Table 15: 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	487.4	48.7	0.7	8160	0.0	2200
Llama-3.2-1B	489.5	48.9	0.2	8107	2.5	3127
Llama-3.2-1B-quantized.w8a8	492.8	49.3	0.3	8196	3.0	2689
Qwen2.5-1.5B	466.4	46.6	0.3	8005	2.0	3756
Qwen2.5-1.5B-quantized.w8a8	455.8	45.6	0.4	8053	0.0	2890
Qwen3-4B	492.1	49.2	0.4	9914	0.0	6996
Qwen3-4B-quantized.w4a16	451.1	45.1	0.2	8576	0.0	3755
Qwen2.5-7B-Instruct	486.4	48.6	0.4	17667	0.0	5754
Qwen2.5-7B-quantized.w8a8	458.1	45.8	0.3	9970	0.0	6103
SmolVLM-256M-Instruct	461.2	46.1	0.2	7443	2.0	1263
llava-1.5-7b-hf	454.2	45.4	0.4	8582	3.0	6629
granite-timeseries-patchtst	461.5	46.1	0.6	7497	0.0	757
chronos-t5-small	454.3	45.4	0.4	7586	0.0	905
arima	455.1	45.5	0.2	7370	0.0	576

Table 16: Results for Idle Baseline