# FMBench Results

**Jetson**: NVIDIA Jetson
Generated: 2025-12-10 00:07:35

## 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 |
| SmolVLM-256M-Instruct |
| arima |
| chronos-t5-small |
| granite-timeseries-patchtst |

### 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 |
| translation | num samples: 10 |

# LLM Scenarios

## ARC Challenge

| Model | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|---|---|---|---|
| Llama-3.2-1B | 2.14 | 3.001 | 2.0 | 0.40 | 417.4 | - | 7762 |
| Llama-3.2-1B-quantized.w8a8 | 3.48 | 0.779 | 2.0 | 0.50 | 549.7 | - | 4239 |
| Qwen2.5-1.5B | 103.37 | 0.872 | 128.0 | 1.00 | 16872.9 | 0.0 | 8757 |
| Qwen2.5-1.5B-quantized.w8a8 | 220.05 | 0.871 | 128.0 | 1.00 | 33375.1 | - | 3995 |

Table 1: Results for ARC Challenge

## ARC Easy

| Model | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|---|---|---|---|---|---|---|
| Llama-3.2-1B | 1.70 | 0.818 | 2.0 | 0.50 | 307.7 | 7810 |
| Llama-3.2-1B-quantized.w8a8 | 3.26 | 0.739 | 2.0 | 0.60 | 502.5 | 4178 |
| Qwen2.5-1.5B | 103.68 | 0.896 | 128.0 | 1.00 | 16827.4 | 9399 |
| Qwen2.5-1.5B-quantized.w8a8 | 224.82 | 0.891 | 128.0 | 1.00 | 33566.7 | 3976 |

Table 2: Results for ARC Easy

## classification

| Model | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|---|---|---|---|
| Llama-3.2-1B | 62.69 | 1.848 | 116.9 | 0.40 | 12214.7 | 0.0 | 7850 |
| Llama-3.2-1B-quantized.w8a8 | 153.16 | 0.801 | 128.0 | 0.50 | 25036.0 | - | 4240 |
| Qwen2.5-1.5B | 12.24 | 0.743 | 19.3 | 0.40 | 2796.2 | - | 8419 |
| Qwen2.5-1.5B-quantized.w8a8 | 6.21 | 0.800 | 13.4 | 0.20 | 3785.5 | - | 3993 |

Table 3: Results for classification

## ner

| Model | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|---|---|---|---|---|---|---|
| Llama-3.2-1B | 13.36 | 0.772 | 30.0 | 0.40 | 3226.2 | 7904 |
| Llama-3.2-1B-quantized.w8a8 | 27.75 | 0.767 | 26.4 | 0.20 | 5259.2 | 4311 |
| Qwen2.5-1.5B | 15.80 | 0.849 | 24.8 | 0.70 | 3280.7 | 8392 |
| Qwen2.5-1.5B-quantized.w8a8 | 23.66 | 0.856 | 18.9 | 0.70 | 5095.7 | 4003 |

Table 4: Results for ner

## perplexity_c4

| Model | Perplexity | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|---|
| Llama-3.2-1B | 23.79 | 841.1 | 0.0 | 8262 |
| Llama-3.2-1B-quantized.w8a8 | 23.88 | 972.1 | 0.8 | 4816 |
| Qwen2.5-1.5B | 24.97 | 943.1 | - | 8841 |
| Qwen2.5-1.5B-quantized.w8a8 | 27.11 | 1205.5 | - | 4427 |

Table 5: Results for perplexity_c4

## perplexity_wikitext2

| Model | Perplexity | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|---|
| Llama-3.2-1B | 495.42 | 135.7 | - | 7863 |
| Llama-3.2-1B-quantized.w8a8 | 551.70 | 219.2 | 0.1 | 4275 |
| Qwen2.5-1.5B | 689.63 | 179.4 | - | 8334 |
| Qwen2.5-1.5B-quantized.w8a8 | 1742.01 | 274.6 | 1.2 | 4044 |

Table 6: Results for perplexity_wikitext2

## sentiment

| Model | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|---|---|---|---|
| Llama-3.2-1B | 39.62 | 0.897 | 67.0 | 0.60 | 7512.9 | 0.0 | 7920 |
| Llama-3.2-1B-quantized.w8a8 | 152.41 | 0.788 | 128.0 | 1.00 | 25412.5 | - | 4399 |
| Qwen2.5-1.5B | 4.13 | 0.801 | 2.0 | 0.80 | 858.8 | 0.4 | 8418 |
| Qwen2.5-1.5B-quantized.w8a8 | 6.86 | 0.804 | 2.0 | 0.80 | 1211.2 | 0.2 | 4005 |

Table 7: Results for sentiment

## summarization

| Model | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|---|---|---|---|
| Llama-3.2-1B | 73.95 | 2.693 | 118.1 | 0.00 | 14086.0 | 0.0 | 8374 |
| Llama-3.2-1B-quantized.w8a8 | 158.91 | 0.766 | 128.0 | 0.00 | 26244.5 | - | 4699 |
| Qwen2.5-1.5B | 109.88 | 0.797 | 126.6 | 0.00 | 18373.4 | - | 9130 |
| Qwen2.5-1.5B-quantized.w8a8 | 209.75 | 0.811 | 124.7 | 0.00 | 33367.9 | - | 3733 |

Table 8: Results for summarization

## translation

| Model | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|---|---|---|---|
| Llama-3.2-1B | 61.49 | 0.863 | 117.9 | 0.10 | 12205.9 | 0.0 | 8133 |
| Llama-3.2-1B-quantized.w8a8 | 150.21 | 0.770 | 128.0 | 0.00 | 24829.1 | - | 4471 |
| Qwen2.5-1.5B | 17.66 | 0.775 | 19.1 | 0.00 | 2593.5 | - | 8886 |
| Qwen2.5-1.5B-quantized.w8a8 | 38.62 | 0.865 | 19.5 | 0.00 | 5207.5 | - | 3476 |

Table 9: Results for translation

# VLM Scenarios

## CountBenchQA

| Model |
|---|
| SmolVLM-256M-Instruct |

Table 10: Results for CountBenchQA

## GTSRB

| Model |
|---|
| SmolVLM-256M-Instruct |

Table 11: Results for GTSRB

## HaGRID

| Model |
|---|
| SmolVLM-256M-Instruct |

Table 12: Results for HaGRID

## VQAv2

| Model |
|---|
| SmolVLM-256M-Instruct |

Table 13: Results for VQAv2

## docvqa

| Model | Energy (J) | VRAM (MB) |
|---|---|---|
| SmolVLM-256M-Instruct | 314.7 | 4076 |

Table 14: Results for docvqa

# Time-Series Scenarios

## FEV-Bench

| Model | sMAPE (%) |
|---|---|
| granite-timeseries-patchtst | 11.23 |
| chronos-t5-small | 200.00 |
| arima | 200.00 |

Table 15: Results for FEV-Bench

## GIFT-EVAL

| Model | sMAPE (%) | Energy (J) | VRAM (MB) |
|---|---|---|---|
| granite-timeseries-patchtst | 19.84 | 20.8 | 2580 |
| chronos-t5-small | 200.00 | - | - |
| arima | 200.00 | - | - |

Table 16: Results for GIFT-EVAL

## M3 Monthly Forecasting

| Model | sMAPE (%) |
|---|---|
| granite-timeseries-patchtst | 200.00 |
| chronos-t5-small | 200.00 |
| arima | 200.00 |

Table 17: Results for M3 Monthly Forecasting

# Baseline Scenarios

## Idle Baseline

| Model | Energy (J) | GPU Util (%) | VRAM (MB) |
|---|---|---|---|
| Llama-3.2-1B | 54.7 | - | 7910 |
| Llama-3.2-1B-quantized.w8a8 | 54.7 | - | 3982 |
| Qwen2.5-1.5B | 60.7 | 1.3 | 9694 |
| Qwen2.5-1.5B-quantized.w8a8 | 56.1 | 4.1 | 4067 |
| SmolVLM-256M-Instruct | 50.6 | 0.3 | 3347 |
| granite-timeseries-patchtst | 52.2 | - | 2325 |
| chronos-t5-small | 52.2 | - | 2413 |
| arima | 51.5 | - | 2316 |

Table 18: Results for Idle Baseline