

# FM Bench Results

**Raspberry Pi:** Raspberry Pi  
Generated: 2025-12-10 05:06:37

## Benchmark Summary

### Evaluated Models

| Model Name                  |
|-----------------------------|
| Llama-3.2-1B                |
| Llama-3.2-1B-quantized.w8a8 |
| Qwen3-0.6B                  |

### Scenario Configurations

| Scenario             | Flags                       |
|----------------------|-----------------------------|
| ARC Challenge        | num samples: 10             |
| ARC Easy             | num samples: 10             |
| Idle Baseline        | idle duration: 10           |
| classification       | 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) | VRAM (MB) |
|-----------------------------|-------------|----------|---------|----------|------------|-----------|
| Qwen3-0.6B                  | 202.99      | 0.001    | 128.0   | 1.00     | 3637.8     | 3583      |
| Llama-3.2-1B                | 10.66       | 0.020    | 2.0     | 0.30     | 232.1      | 5722      |
| Llama-3.2-1B-quantized.w8a8 | 25.50       | 0.004    | 2.0     | 0.60     | 504.0      | 6432      |

Table 1: Results for ARC Challenge

### ARC Easy

| Model                       | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|-----------------------------|-------------|----------|---------|----------|------------|-----------|
| Qwen3-0.6B                  | 198.44      | 0.025    | 128.0   | 1.00     | 3706.9     | 3585      |
| Llama-3.2-1B                | 8.31        | 0.039    | 2.0     | 0.40     | 176.1      | 5741      |
| Llama-3.2-1B-quantized.w8a8 | 23.73       | 0.009    | 2.0     | 0.50     | 459.6      | 6575      |

Table 2: Results for ARC Easy

### classification

| Model                       | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|-----------------------------|-------------|----------|---------|----------|------------|-----------|
| Qwen3-0.6B                  | 203.52      | 0.001    | 128.0   | 0.90     | 3700.7     | 3600      |
| Llama-3.2-1B                | 420.90      | 0.029    | 104.1   | 0.30     | 7419.9     | 5785      |
| Llama-3.2-1B-quantized.w8a8 | 1415.87     | 0.002    | 128.0   | 0.40     | 26847.1    | 6535      |

Table 3: Results for classification

### ner

| Model                       | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|-----------------------------|-------------|----------|---------|----------|------------|-----------|
| Qwen3-0.6B                  | 198.99      | 0.001    | 128.0   | 0.80     | 3613.4     | 3574      |
| Llama-3.2-1B                | 160.26      | 0.036    | 26.4    | 0.10     | 1805.2     | 5772      |
| Llama-3.2-1B-quantized.w8a8 | 252.15      | 0.004    | 26.4    | 0.20     | 5583.8     | 6616      |

Table 4: Results for ner

### perplexity\_c4

| Model                       | Perplexity | Energy (J) | VRAM (MB) |
|-----------------------------|------------|------------|-----------|
| Qwen3-0.6B                  | 60.38      | 409.1      | 3674      |
| Llama-3.2-1B                | 23.78      | 892.9      | 5830      |
| Llama-3.2-1B-quantized.w8a8 | 23.89      | 953.7      | 6202      |

Table 5: Results for perplexity\_c4

## perplexity\_wikitext2

| Model                       | Perplexity | Energy (J) | VRAM (MB) |
|-----------------------------|------------|------------|-----------|
| Qwen3-0.6B                  | 1448.93    | 58.8       | 3529      |
| Llama-3.2-1B                | 495.39     | 105.7      | 5793      |
| Llama-3.2-1B-quantized.w8a8 | 454.20     | 182.5      | 5923      |

Table 6: Results for perplexity\_wikitext2

## sentiment

| Model                       | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|-----------------------------|-------------|----------|---------|----------|------------|-----------|
| Qwen3-0.6B                  | 209.96      | 0.002    | 128.0   | 1.00     | 3942.2     | 3725      |
| Llama-3.2-1B                | 441.66      | 0.020    | 91.7    | 0.70     | 6957.9     | 5788      |
| Llama-3.2-1B-quantized.w8a8 | 1390.01     | 0.012    | 128.0   | 1.00     | 27051.6    | 6573      |

Table 7: Results for sentiment

## summarization

| Model                       | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|-----------------------------|-------------|----------|---------|----------|------------|-----------|
| Qwen3-0.6B                  | 223.63      | 0.002    | 128.0   | 0.00     | 4475.8     | 3795      |
| Llama-3.2-1B                | 504.82      | 0.025    | 128.0   | 0.00     | 9825.0     | 5802      |
| Llama-3.2-1B-quantized.w8a8 | 1464.52     | 0.002    | 128.0   | 0.00     | 28145.5    | 6725      |

Table 8: Results for summarization

## translation

| Model      | Latency (s) | TTFT (s) | Tok/Out | Accuracy | Energy (J) | VRAM (MB) |
|------------|-------------|----------|---------|----------|------------|-----------|
| Qwen3-0.6B | 201.76      | 0.002    | 128.0   | 0.10     | 3944.9     | 6652      |

Table 9: Results for translation

## Baseline Scenarios

### Idle Baseline

| Model                       | Energy (J) | Idle Power (W) | VRAM (MB) |
|-----------------------------|------------|----------------|-----------|
| Qwen3-0.6B                  | 9.2        | 0.9            | 3734      |
| Llama-3.2-1B                | 10.3       | 1.0            | 5594      |
| Llama-3.2-1B-quantized.w8a8 | 11.1       | 1.1            | 2292      |

Table 10: Results for Idle Baseline