Architectural Evaluation of On-Device LLM Inference

Scaling, Energy Efficiency, and Thermal Stability on Apple Silicon M3 (16GB Unified Memory)

Shaun Jerome

February 2026

Abstract

This work presents a structured empirical evaluation of transformer-based large language model (LLM) inference on consumer-grade Apple Silicon M3 (16GB unified memory). Rather than reporting raw performance alone, this study investigates scaling behavior across 1B, 3B, and 7B parameter models, power consumption characteristics, energy cost per token, unified memory utilization, and sustained thermal stability under extended workloads.

Log–log regression of throughput versus parameter count yields an empirical scaling exponent $\alpha \approx 0.725$, indicating sub-linear inverse scaling relative to ideal compute-bound behavior. Sustained 7B inference maintained stable throughput under nominal thermal pressure on a fan-less system, demonstrating effective dynamic frequency management under transformer workloads.

1 Introduction

Large language models are commonly benchmarked on server-class GPUs with discrete memory and active cooling. However, increasing demand for on-device AI systems requires understanding how transformer inference behaves under unified-memory architectures and constrained thermal envelopes.

Apple Silicon introduces several architectural characteristics:

- Unified CPU–GPU memory pool
- Integrated GPU compute
- Fan-less thermal envelope (MacBook Air)
- Aggressive dynamic frequency scaling

This study evaluates inference scaling behavior under these constraints.

2 Experimental Setup

2.1 Hardware Configuration

- Device: MacBook Air M3 (Fan-less)

- CPU: 8-core (4 Performance + 4 Efficiency)
- GPU: 8-core integrated
- Memory: 16GB unified memory
- OS: macOS 15.2

2.2 Software Configuration

- Framework: Apple MLX
- Quantization: 4-bit
- Power Sampling: macOS `powermetrics`

2.3 Models Evaluated

Model	Parameters	Quantization
Llama 3.2	1B	4-bit
Qwen 2.5	3B	4-bit
Qwen 2.5	7B	4-bit

Table 1: Models evaluated in this study

3 Methodology

For each model:

- Multiple inference iterations executed
- Throughput (tokens/sec) recorded
- CPU + GPU combined power sampled
- Memory usage and swap monitored

Energy per token computed as:

$$E_{token} = \frac{Power}{Throughput} \tag{1}$$

A sustained 30-minute inference test was conducted for the 7B model.

4 Results

4.1 Throughput Scaling

Average throughput observed:

- 1B: ~ 37.95 tokens/sec
- 3B: ~ 20.6 tokens/sec
- 7B: ~ 9.1 tokens/sec

Scaling behavior modeled as:

$$Throughput \propto N^{-\alpha} \tag{2}$$

Log–log regression yields:

$$\alpha \approx 0.725 \tag{3}$$

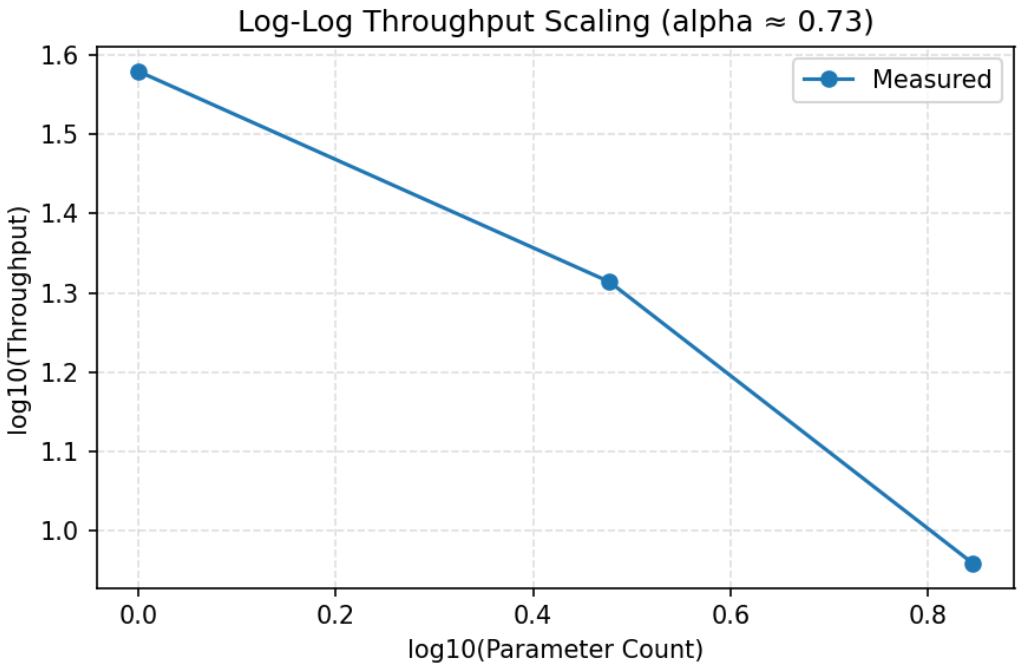

Figure 1: Log–log throughput scaling across 1B, 3B, and 7B models

4.2 Throughput Comparison

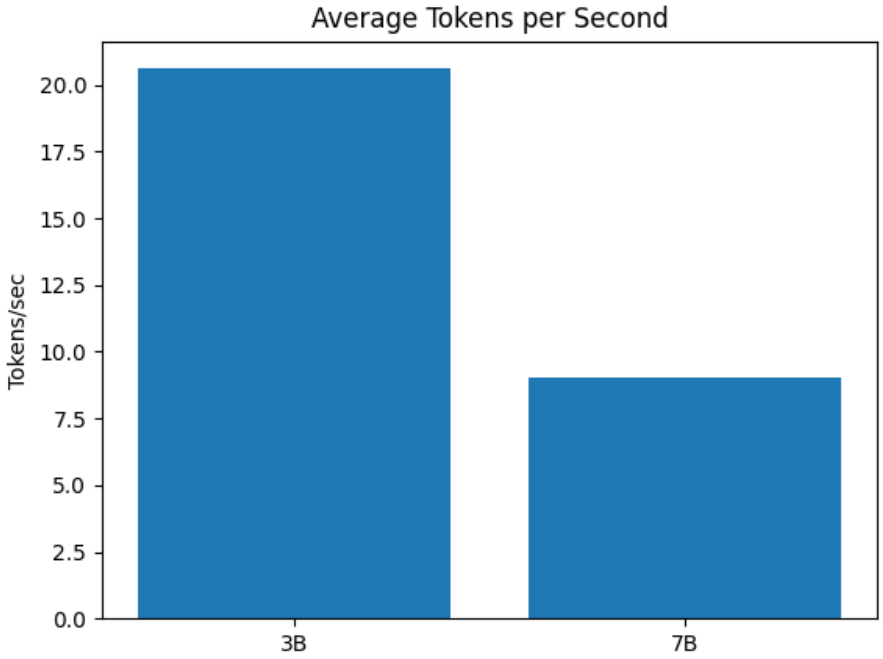

Figure 2: Average throughput comparison

4.3 Power Behavior

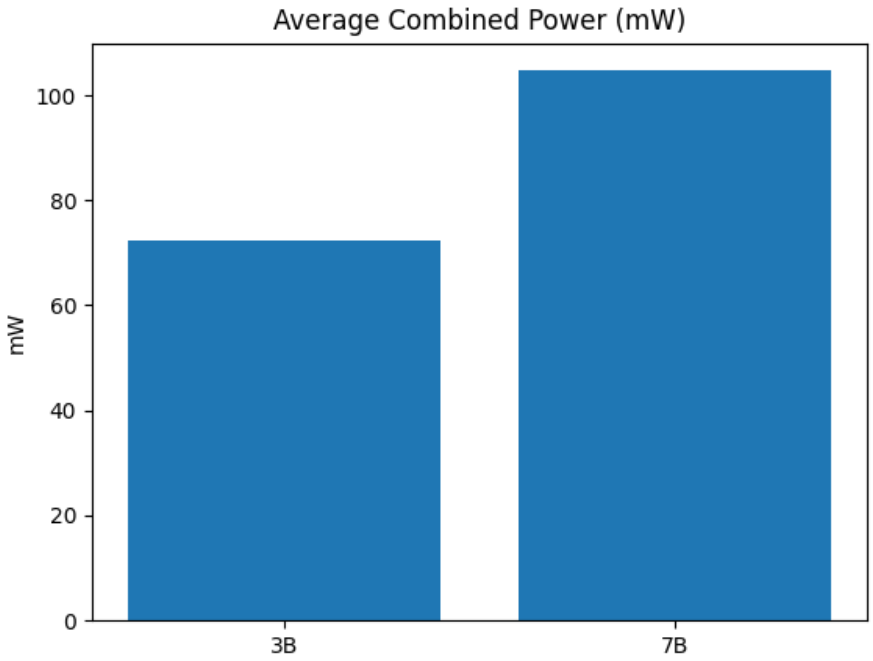

Figure 3: Average combined power consumption

4.4 Energy Efficiency

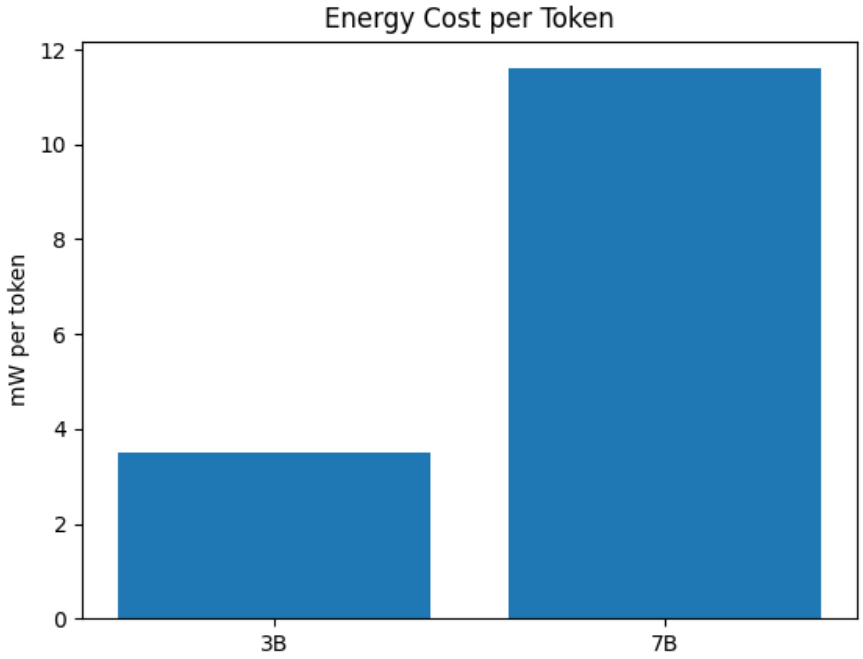

Figure 4: Energy cost per token

Energy per token increases super-linearly with model size, indicating diminishing efficiency at larger scales.

4.5 Throughput vs Power Regime

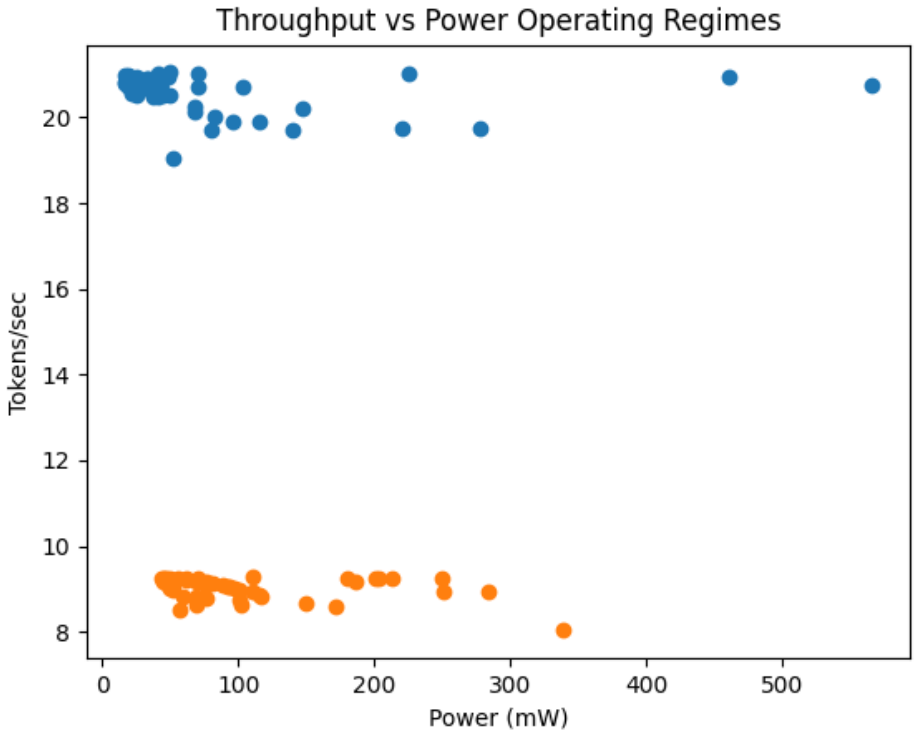

Figure 5: Operating regimes: throughput vs power

4.6 Sustained Thermal Stability

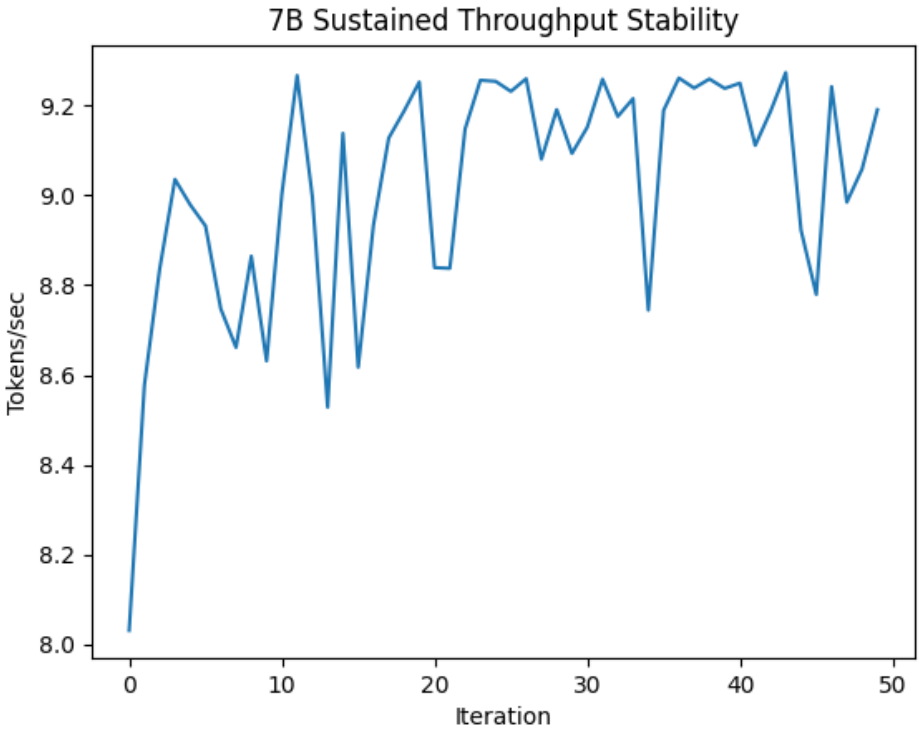

Figure 6: 7B sustained throughput over time

Thermal pressure remained nominal throughout testing. No observable throttling occurred.

4.7 Unified Memory Scaling

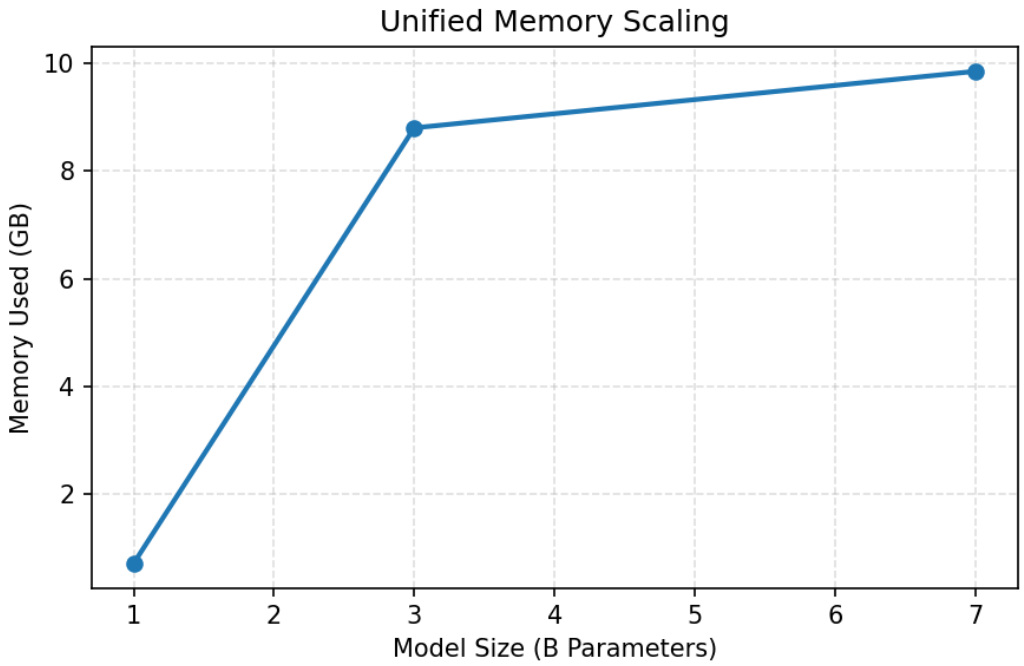

Figure 7: Unified memory utilization scaling

Memory usage remained within safe limits (approx. 9–10GB for 7B), with negligible swap.

5 Discussion

Ideal compute-bound scaling implies $\alpha = 1$:

$$Throughput \propto \frac{1}{N} \tag{4}$$

Observed $\alpha < 1$ suggests hybrid compute–memory regime. Possible contributors:

- Unified memory arbitration
- Memory bandwidth effects
- Cache hierarchy reuse
- Kernel scheduling overhead

Moderate-scale models ($\leq$ 3B) provide favorable throughput-to-energy balance.

6 Limitations

- Single hardware platform

- Single quantization level
- Limited parameter range
- No cross-platform comparison

7 Conclusion

This study demonstrates that consumer unified-memory architectures can sustain moderate-scale transformer inference without thermal throttling. Throughput scales sub-linearly ($\alpha \approx 0.725$), energy cost increases super-linearly, and unified memory remains stable under sustained 7B workloads.

These findings highlight architectural tradeoffs and hardware–software co-design considerations for on-device AI workloads.