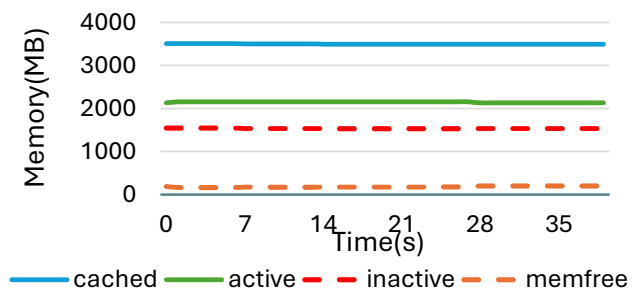
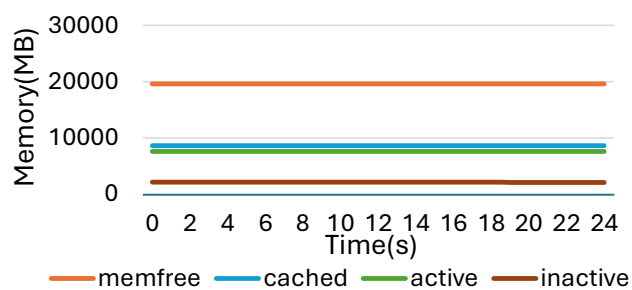




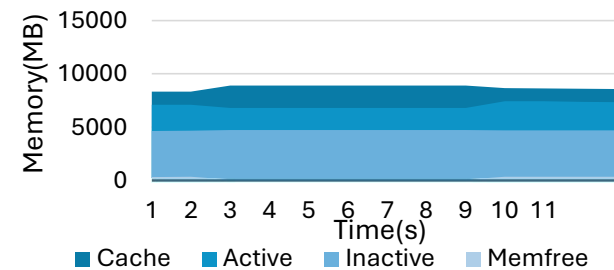
(a) TinyLlama on Raspberry Pi-5



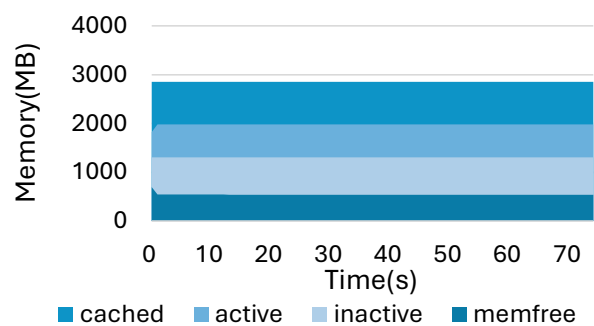
(b) TinyLlama on Jetson Orin



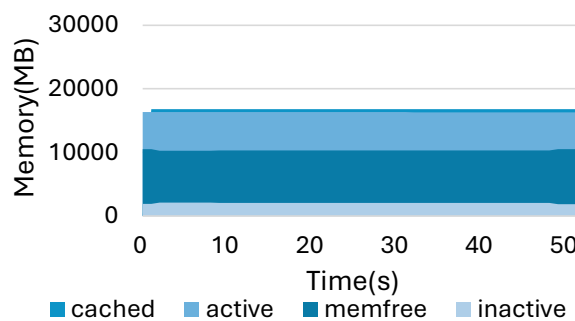
(c) TinyLlama on Mac Mini



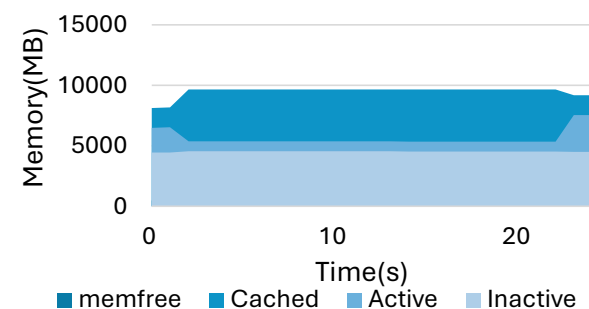
(d) Phi-3 on Raspberry Pi-5



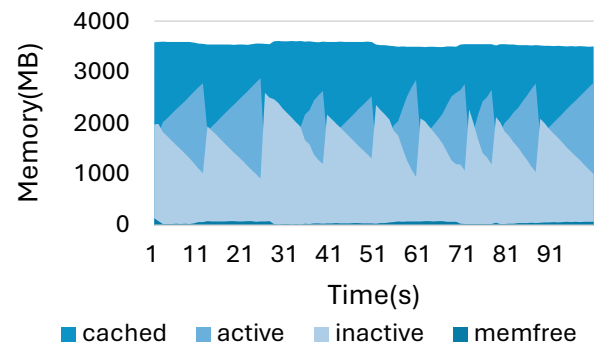
(e) Phi-3 on Jetson Orin



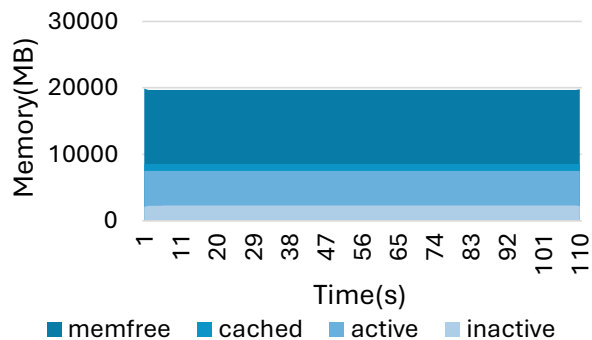
(f) Phi-3 on Mac Mini



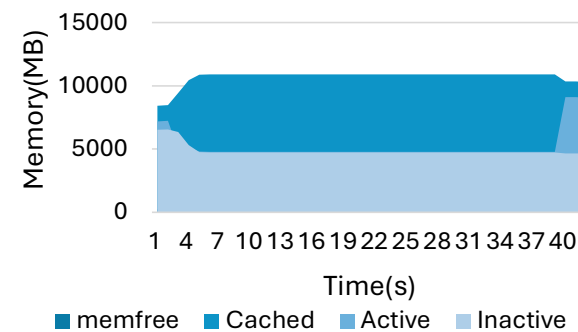
(g) LLama-3 on Raspberry Pi-5



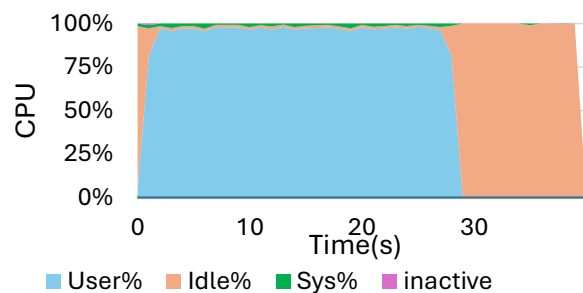
(h) LLama-3 on Jetson Orin



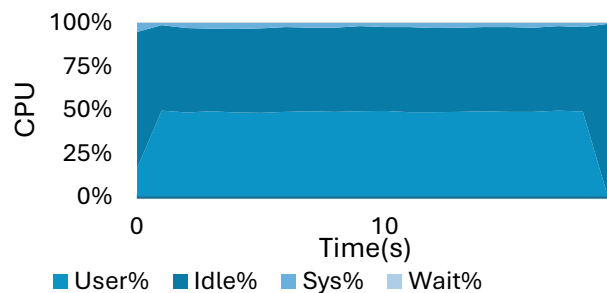
(i) LLama-3 on Mac Mini



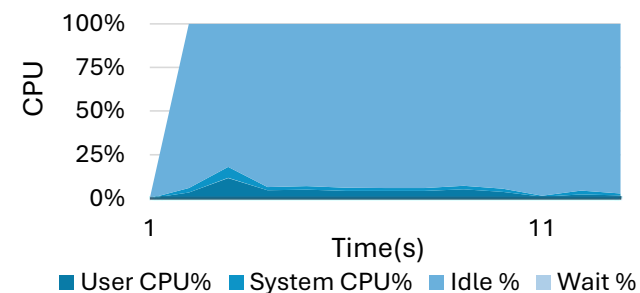
(a) TinyLlama on Raspberry Pi-5



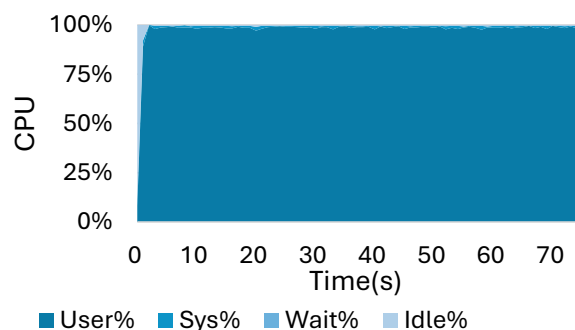
(b) TinyLlama on Jetson Orin



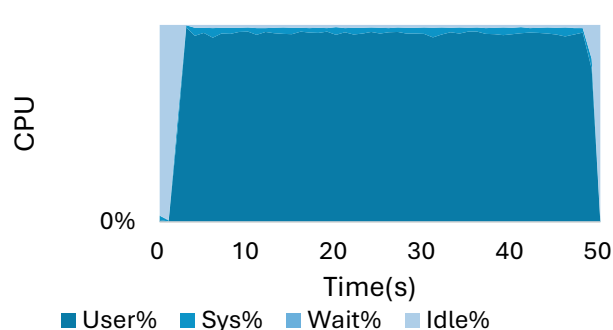
(c) TinyLlama on Mac Mini



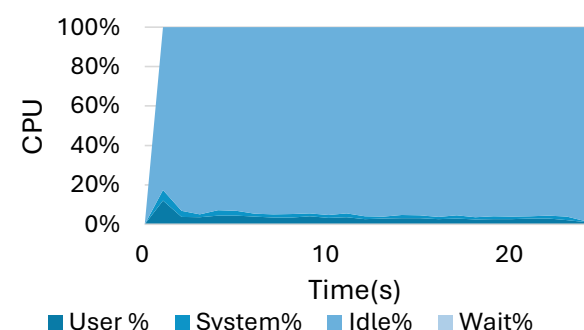
(d) Phi-3 on Raspberry Pi-5



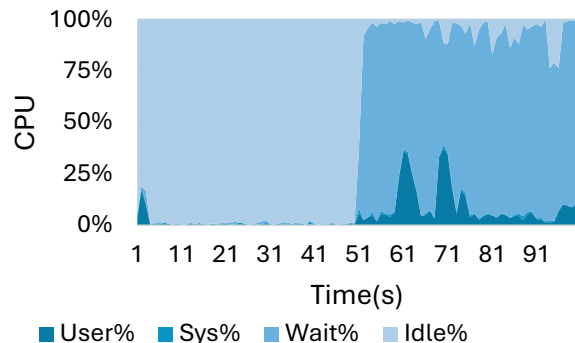
(e) Phi-3 on Jetson Orin



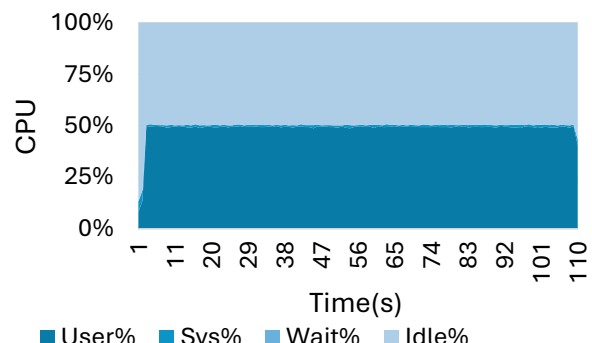
(f) Phi-3 on Mac Mini



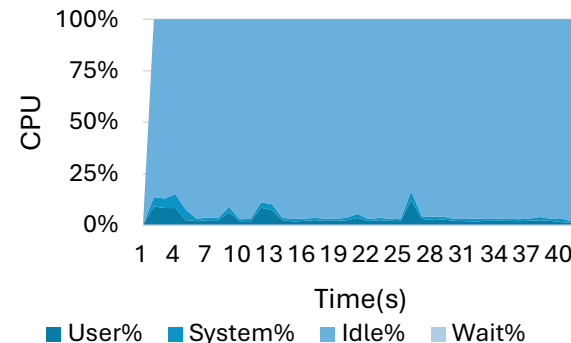
(g) LLama-3 on Raspberry Pi-5



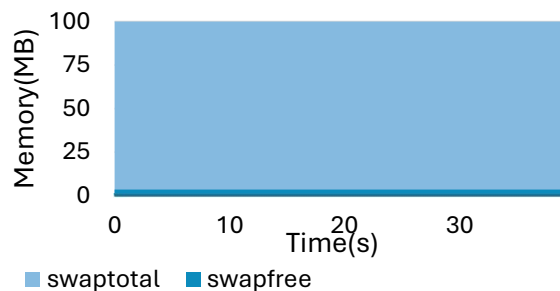
(h) LLama-3 on Jetson Orin



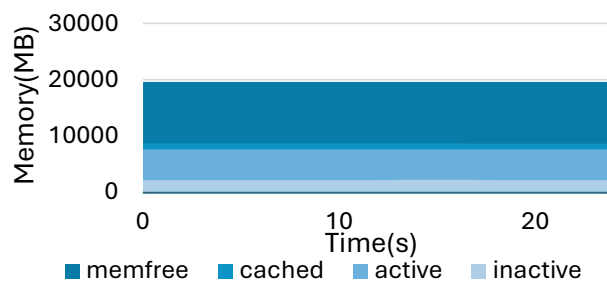
(i) LLama-3 on Jetson Orin



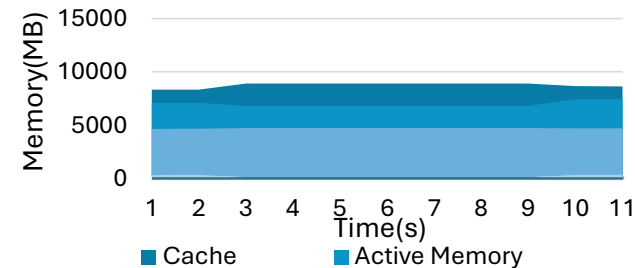
(a) TinyLlama on Raspberry Pi-5



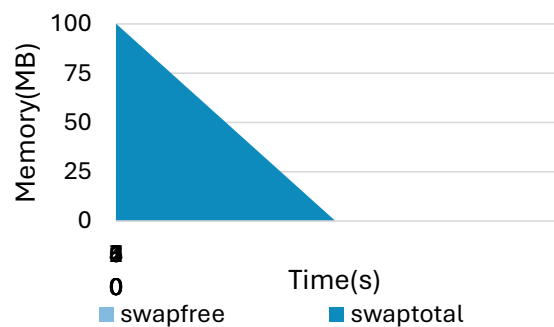
(b) TinyLlama on Jetson Orin



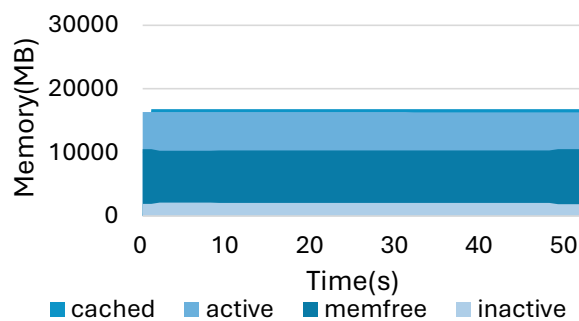
(c) TinyLlama on Mac Mini



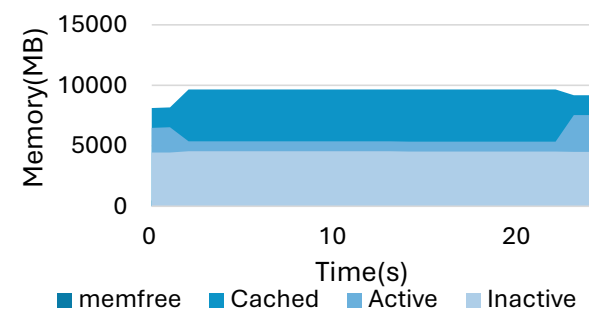
(d) Phi-3 on Raspberry Pi-5



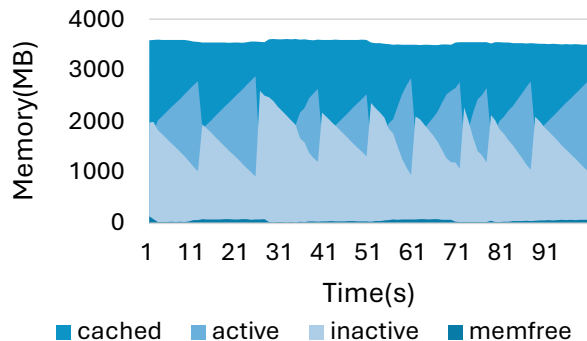
(e) Phi-3 on Jetson Orin



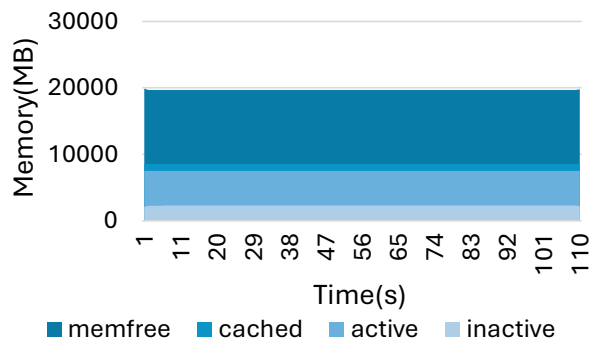
(f) Phi-3 on Mac Mini



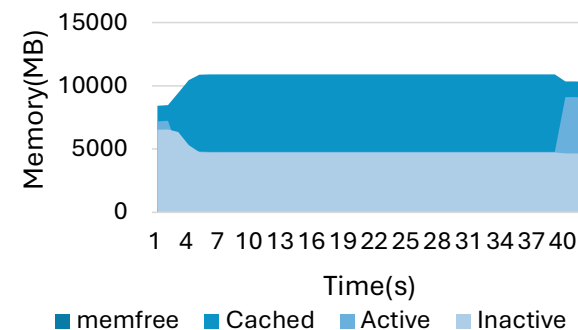
(g) LLama-3 on Raspberry Pi-5

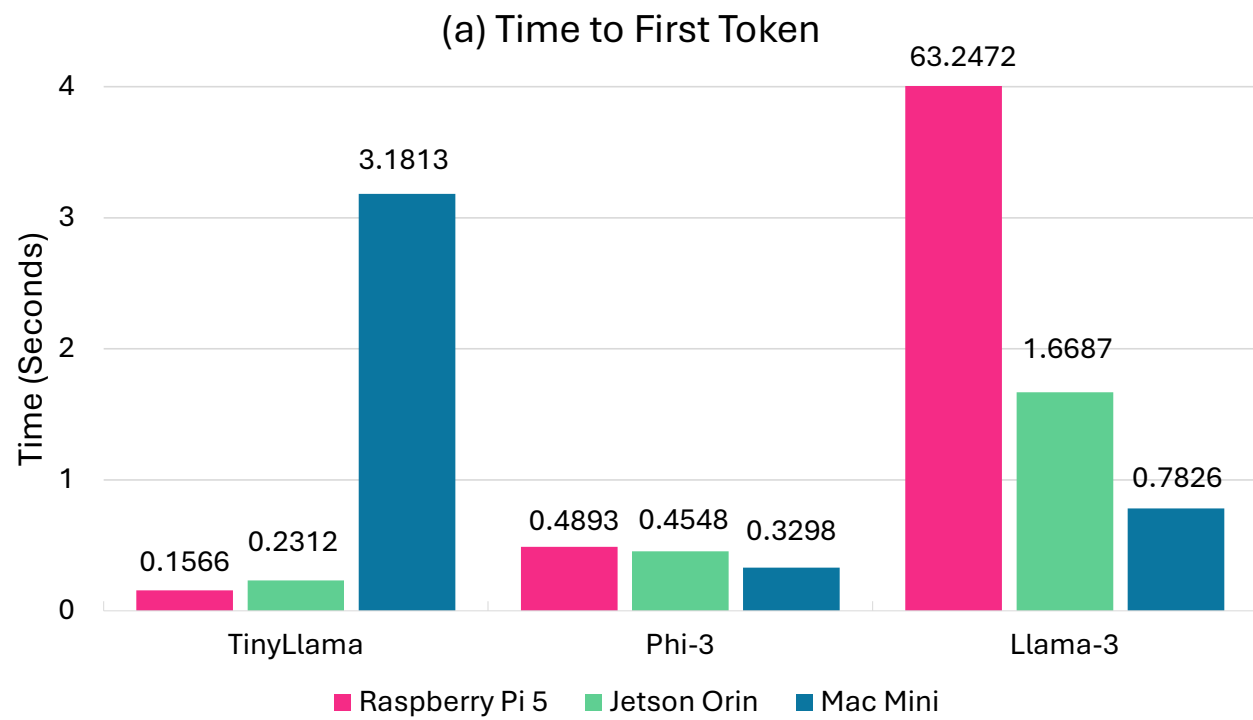


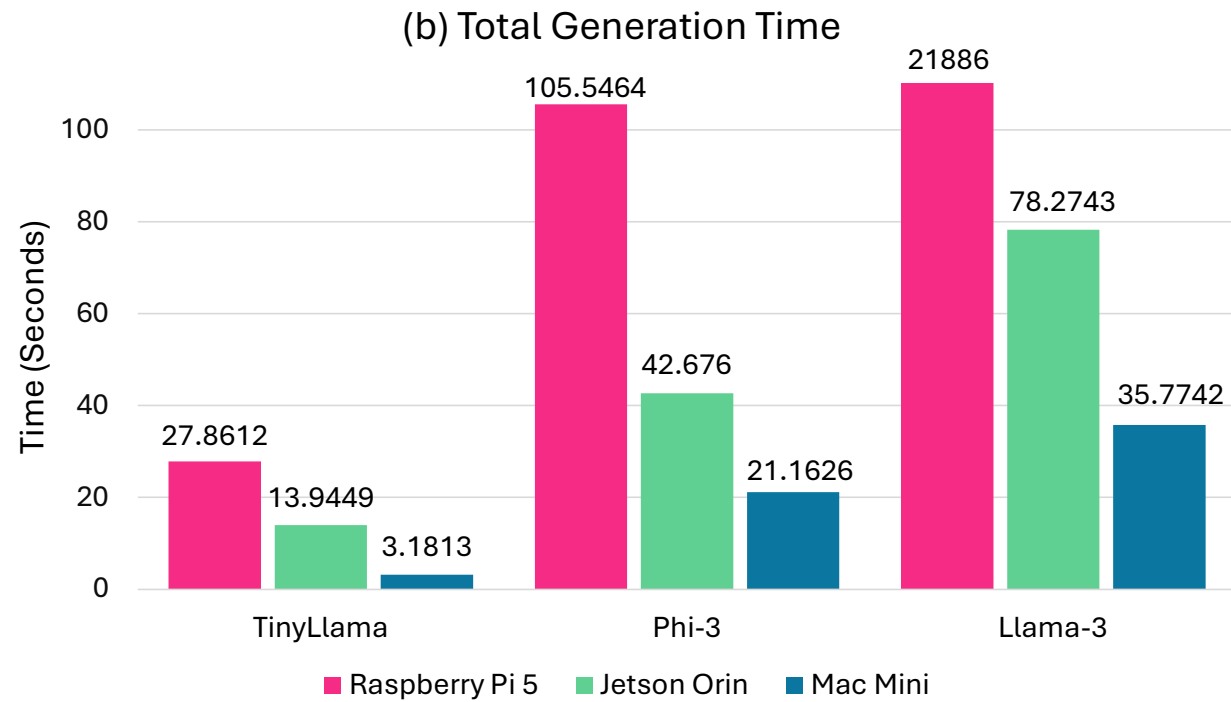
(h) LLama-3 on Jetson Orin



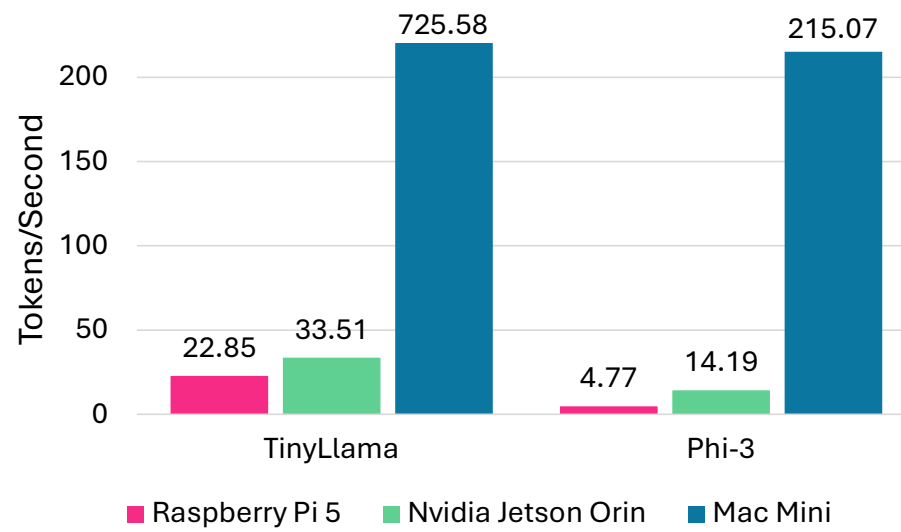
(i) LLama-3 on Mac Mini





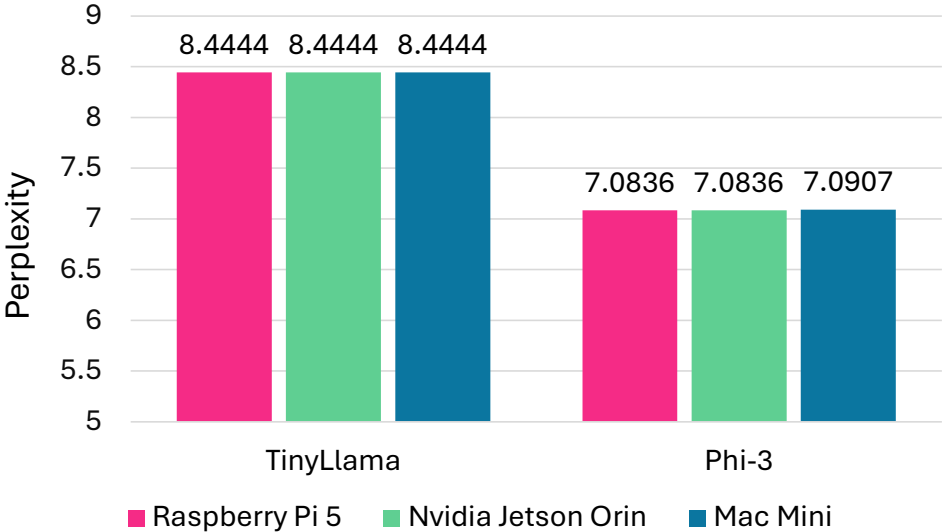


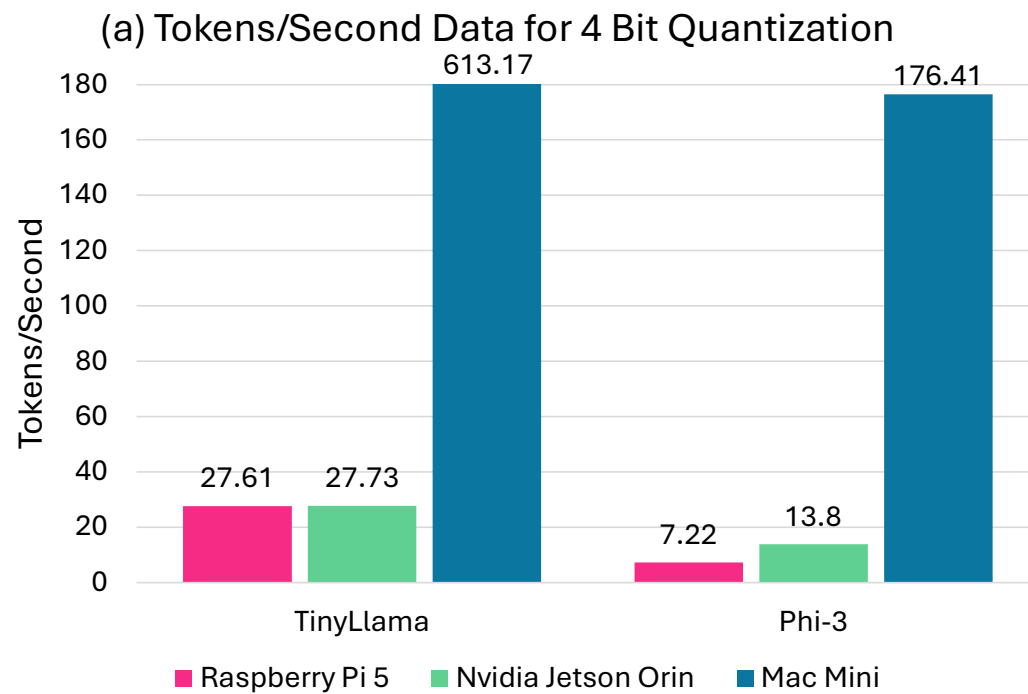
(a) Tokens/Second Data for F16 Quantization



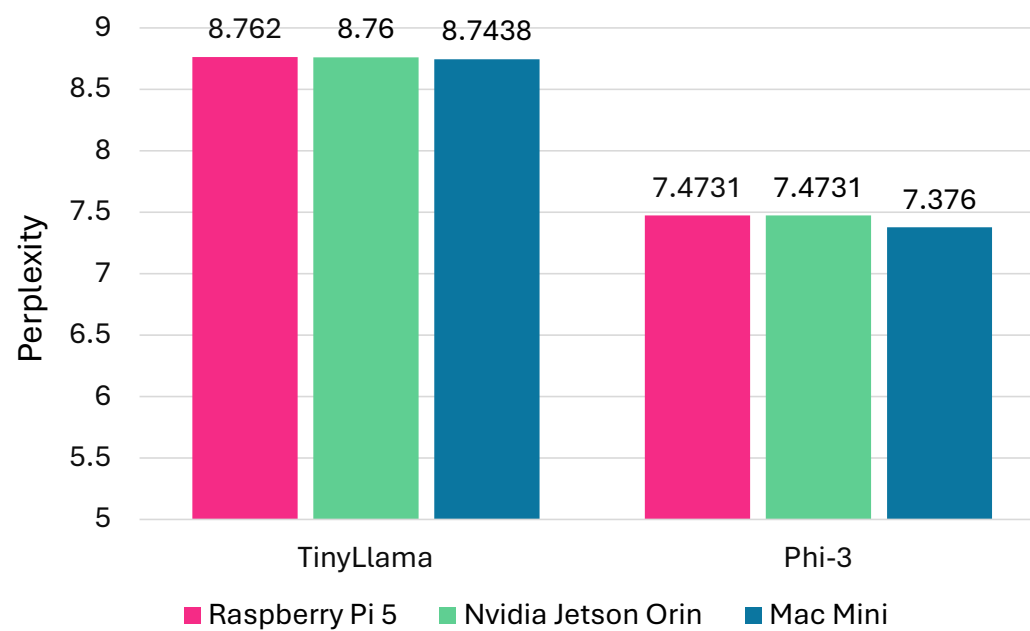


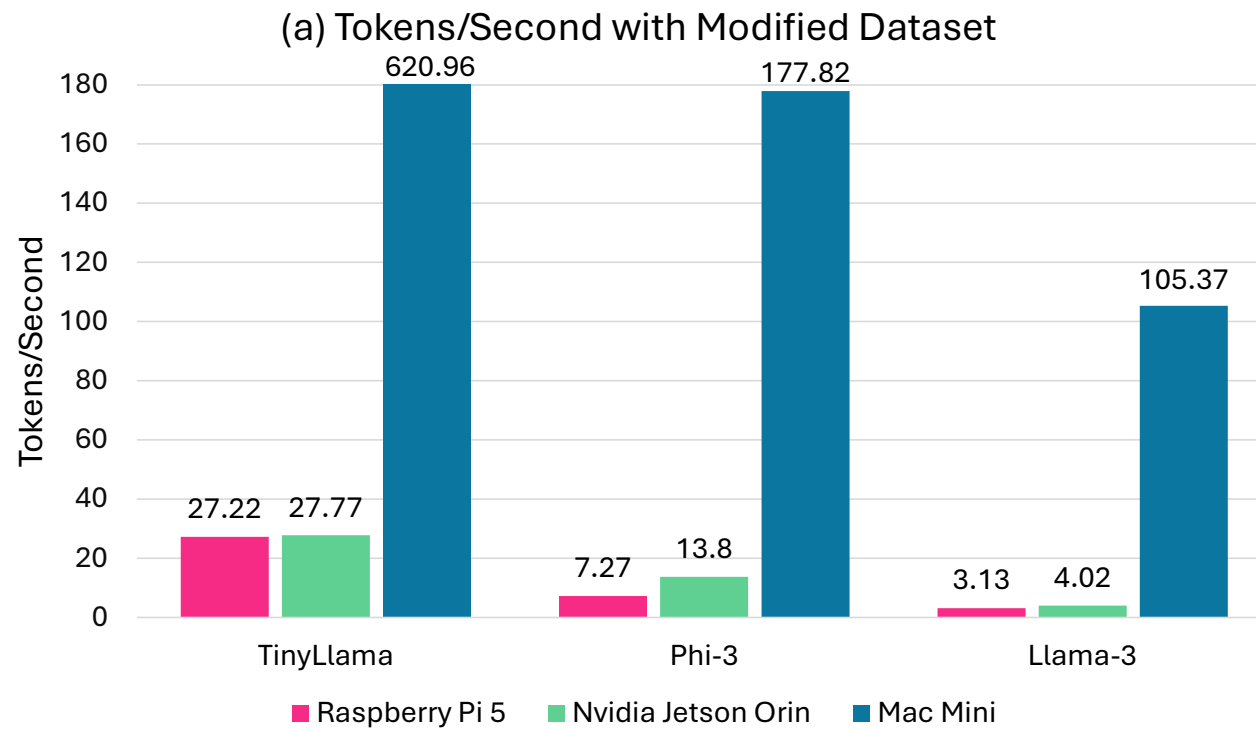
(b) Perplexity Score for F16 Quantization





(b) Perplexity Score for 4 Bit Quantization





(b) Perplexity Score with Modified Dataset

