(b) Perplexity Score for F16 Quantization 9 8.4444 8.4444 8.4444 8.5 8 7.5 7.0836 7.0836 7.0907 Perplexity 6.5 6 5.5 5 TinyLlama Phi-3 ■ Raspberry Pi 5
■ Nvidia Jetson Orin Mac Mini