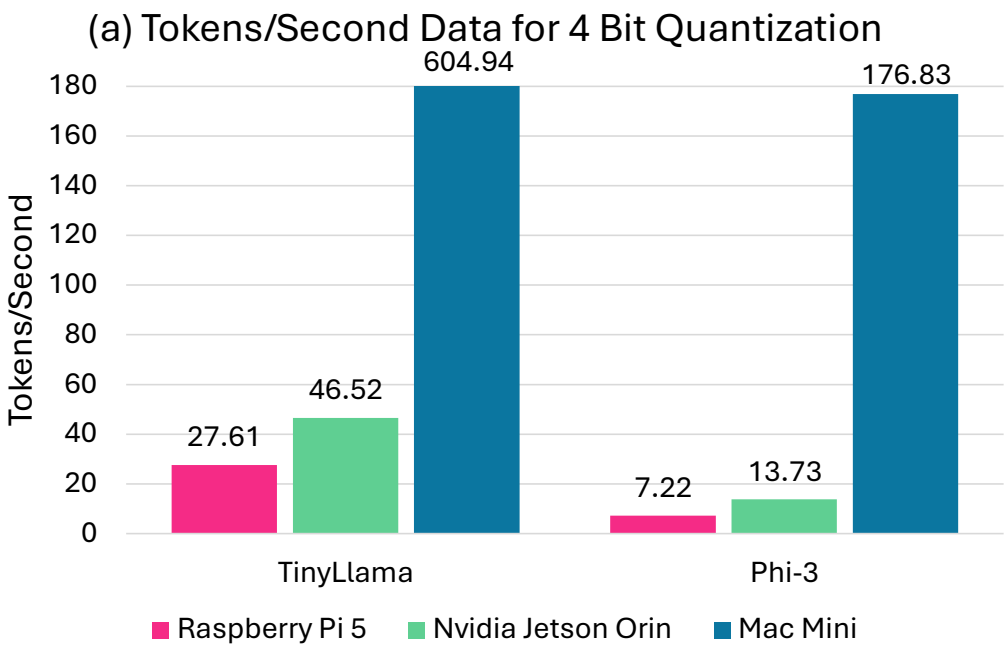
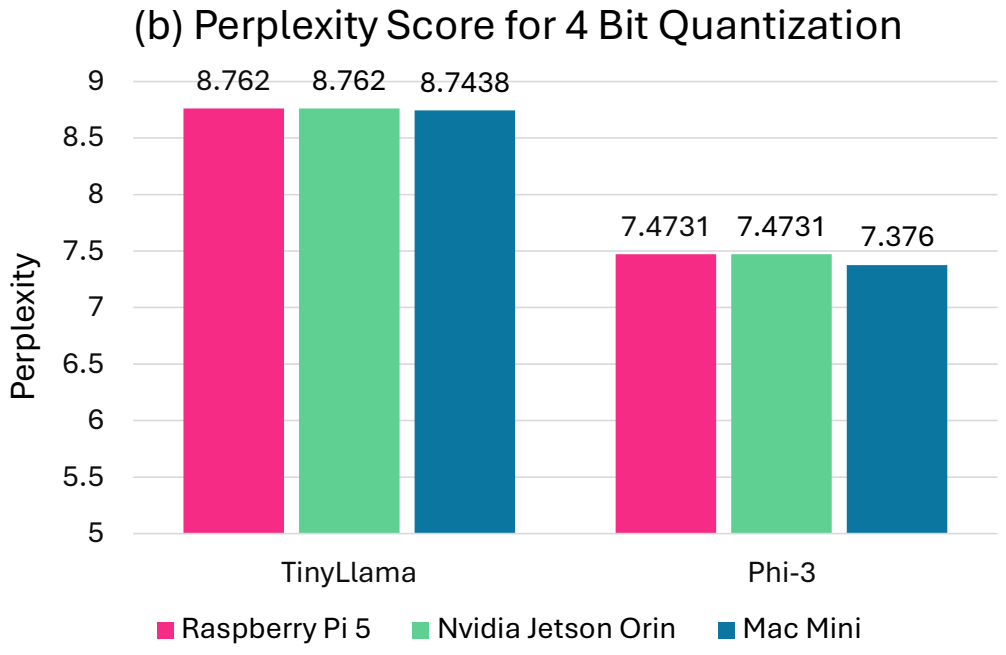

(a) Tokens/Second Data for 4 Bit Quantization
Tokens/Second
180
160
140
120
100
80
60
40
20
0
27.61
46.52
604.94
7.22
13.73
176.83
TinyLlama
Phi-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini


(b) Perplexity Score for 4 Bit Quantization
Perplexity
9
8.5
8
7.5
7
6.5
6
5.5
5
8.762
8.762
8.7438
7.4731
7.4731
7.376
TinyLlama
Phi-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini