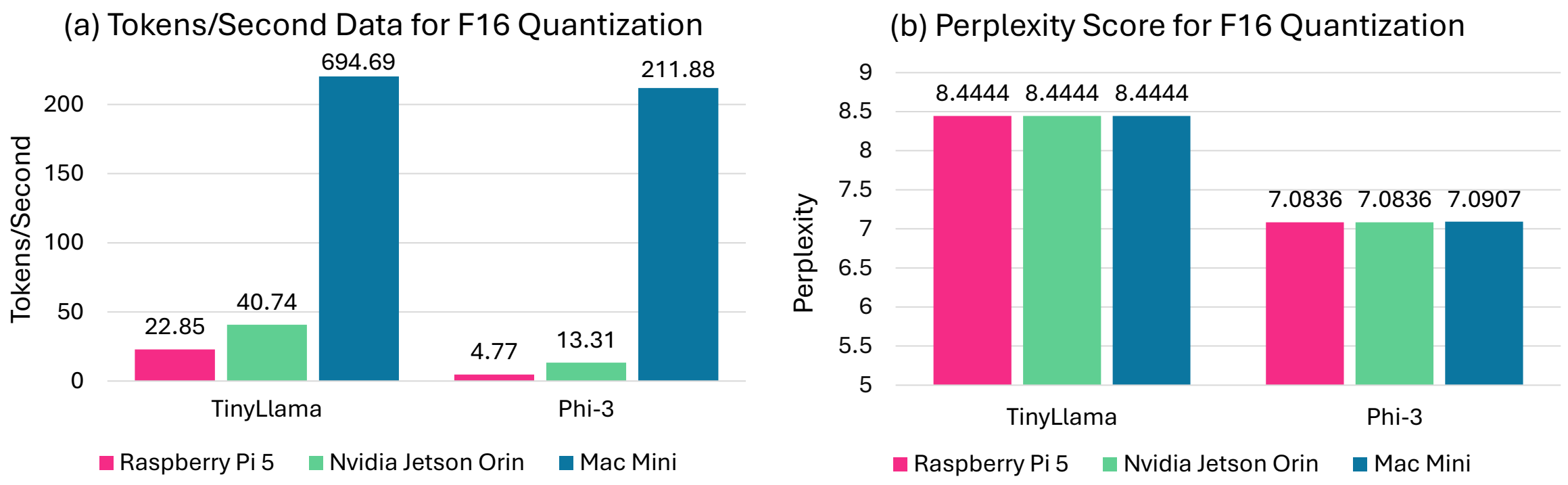
(a) Tokens/Second Data for F16 Quantization
Tokens/Second
200
150
100
50
0
694.69
211.88
40.74
22.85
13.31
4.77
TinyLlama
Phi-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini
(b) Perplexity Score for F16 Quantization
Perplexity
9
8.5
8
7.5
7
6.5
6
5.5
5
8.4444
8.4444
8.4444
7.0836
7.0836
7.0907
TinyLlama
Phi-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini