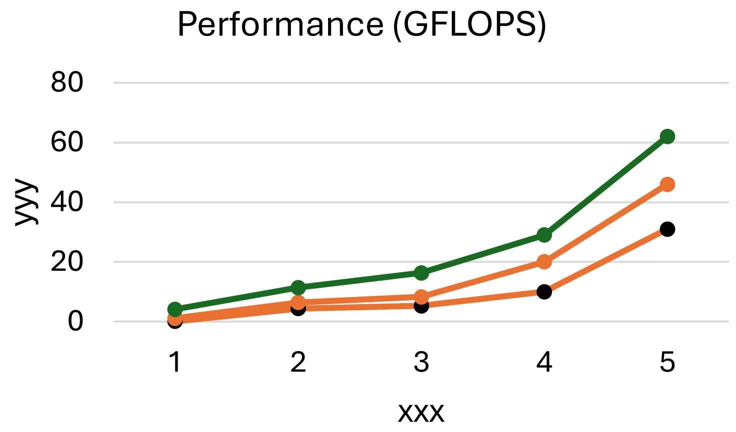
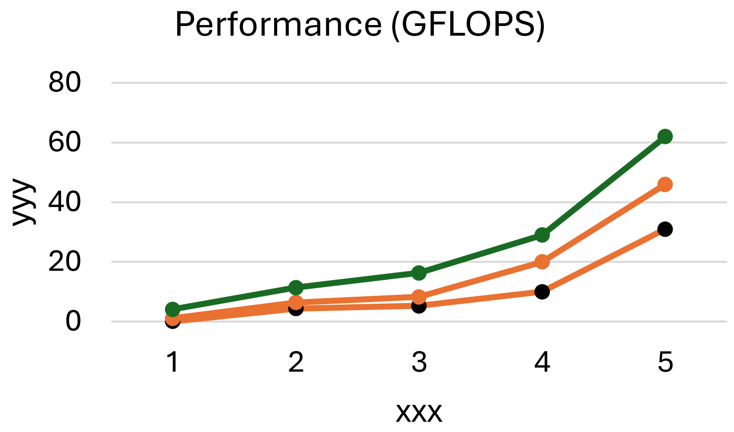
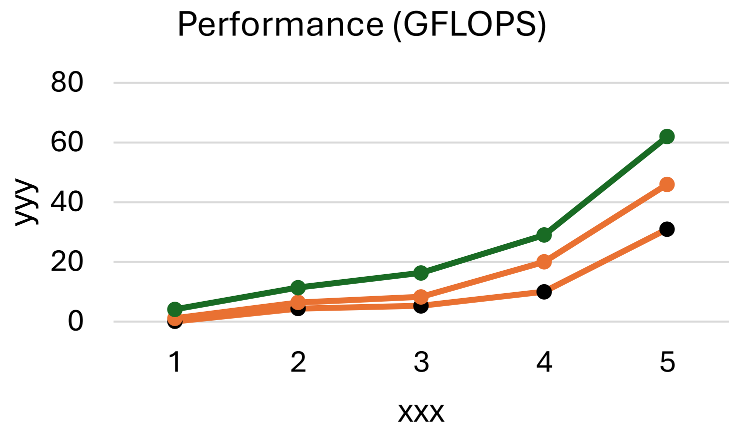
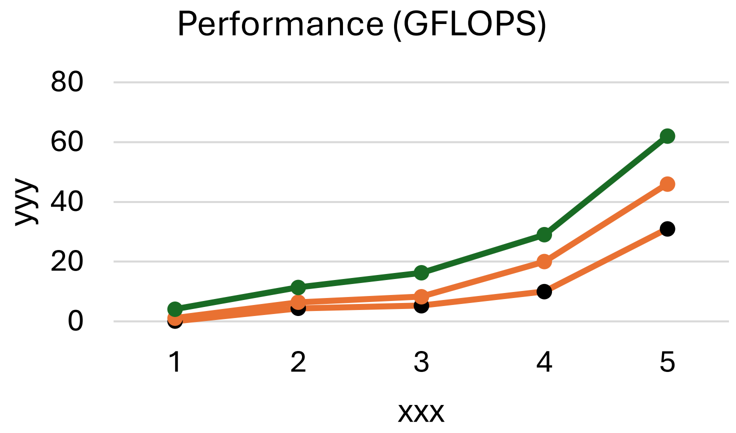
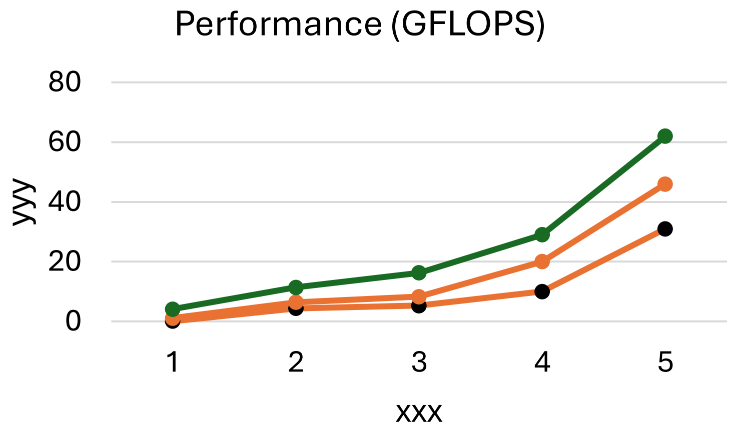
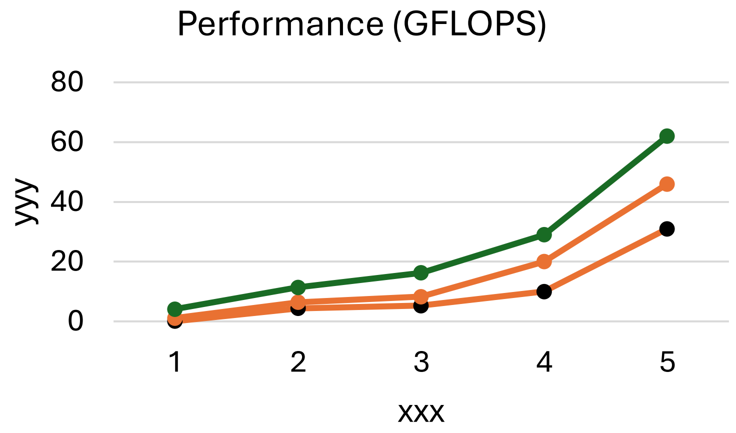

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

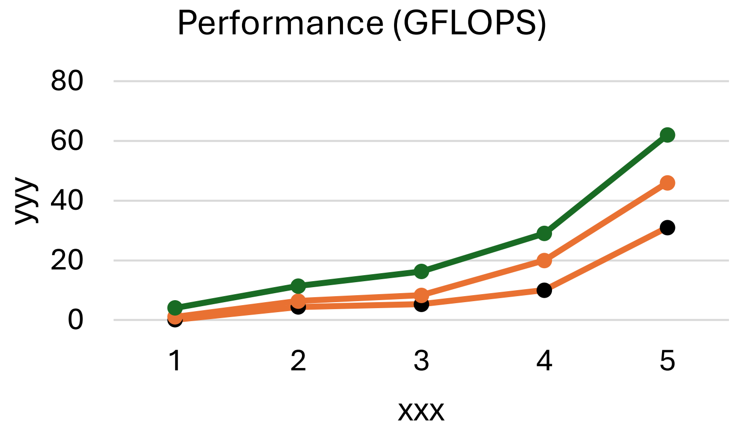

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

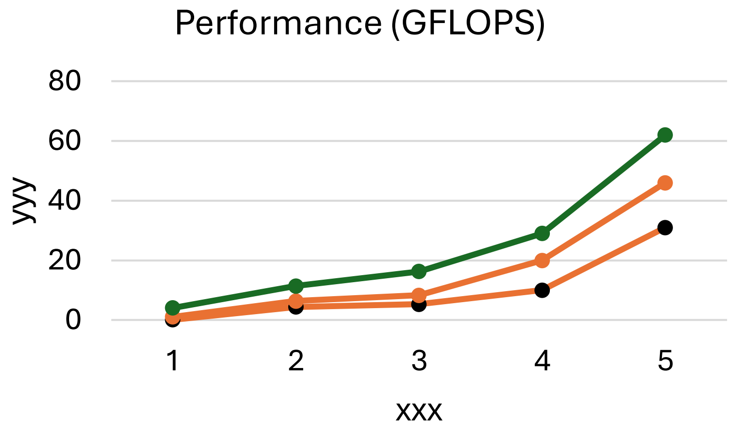

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

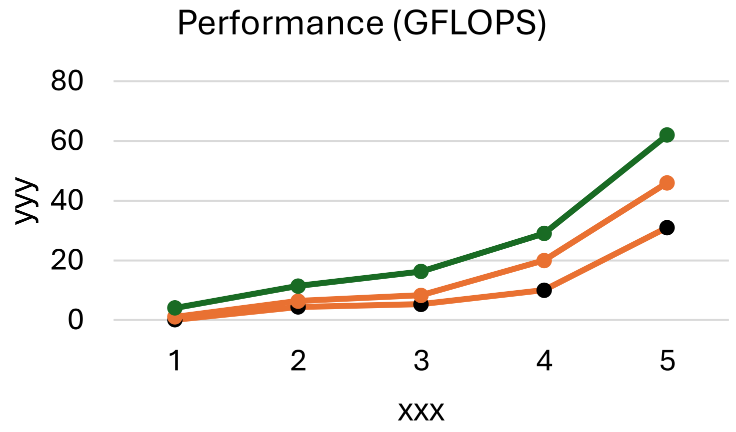

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

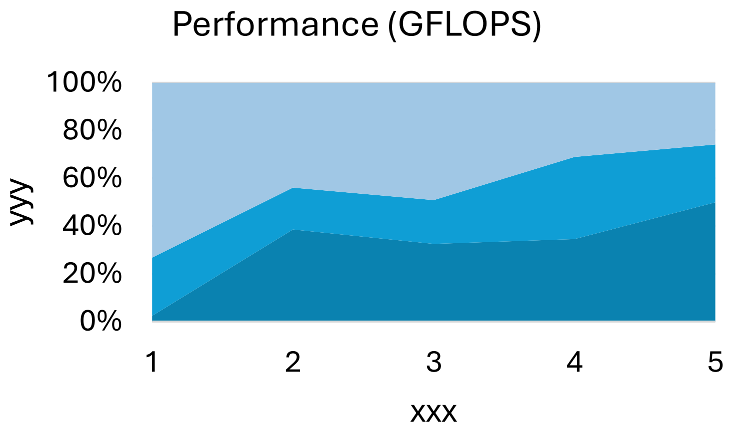

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

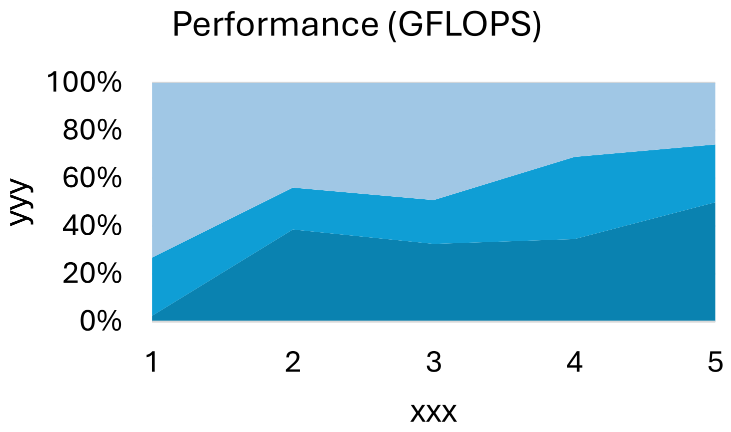

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

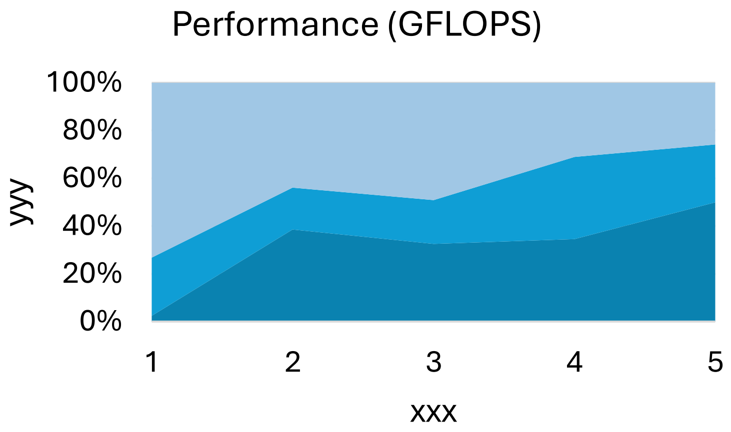

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

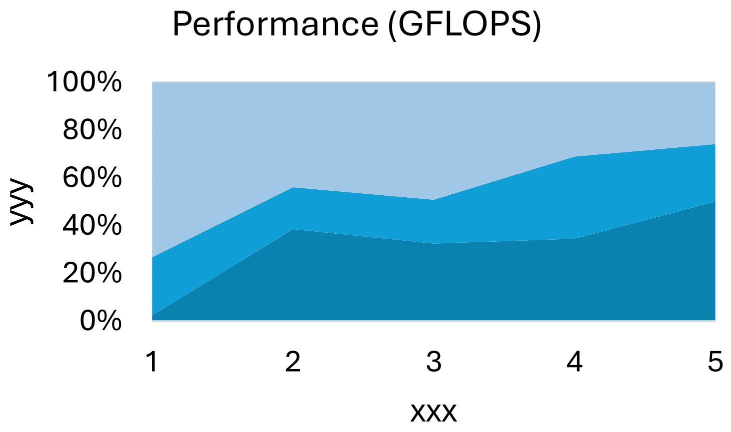

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

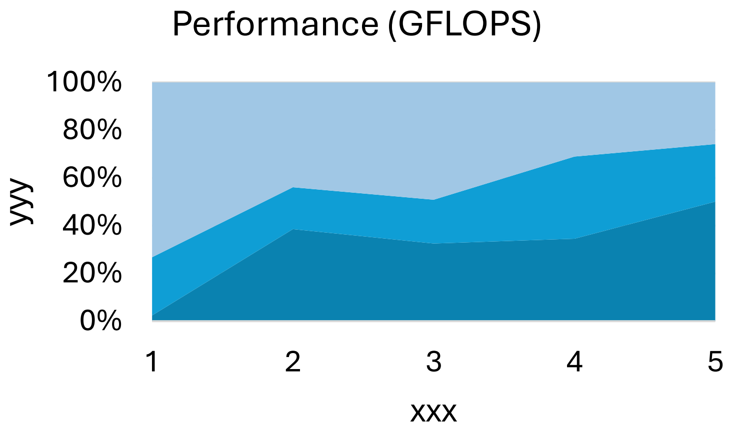

Performance (GFLOPS)
80
60
40
20
0
yyy
1
2
3
4
5
xxx

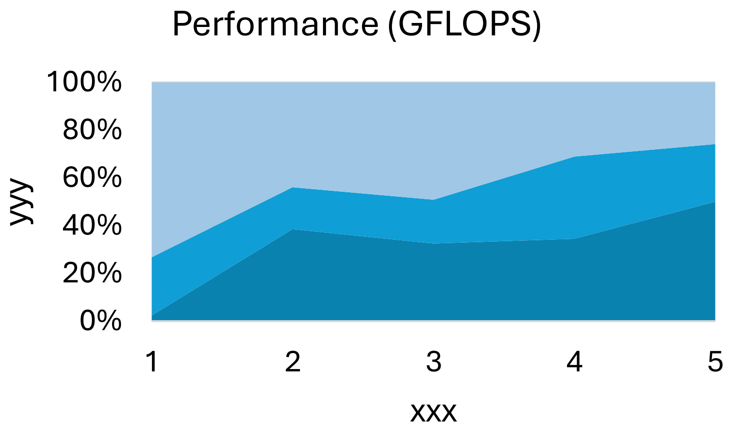

Performance (GFLOPS)
100%
80%
60%
40%
20%
0%
yyy
1
2
3
4
5
xxx

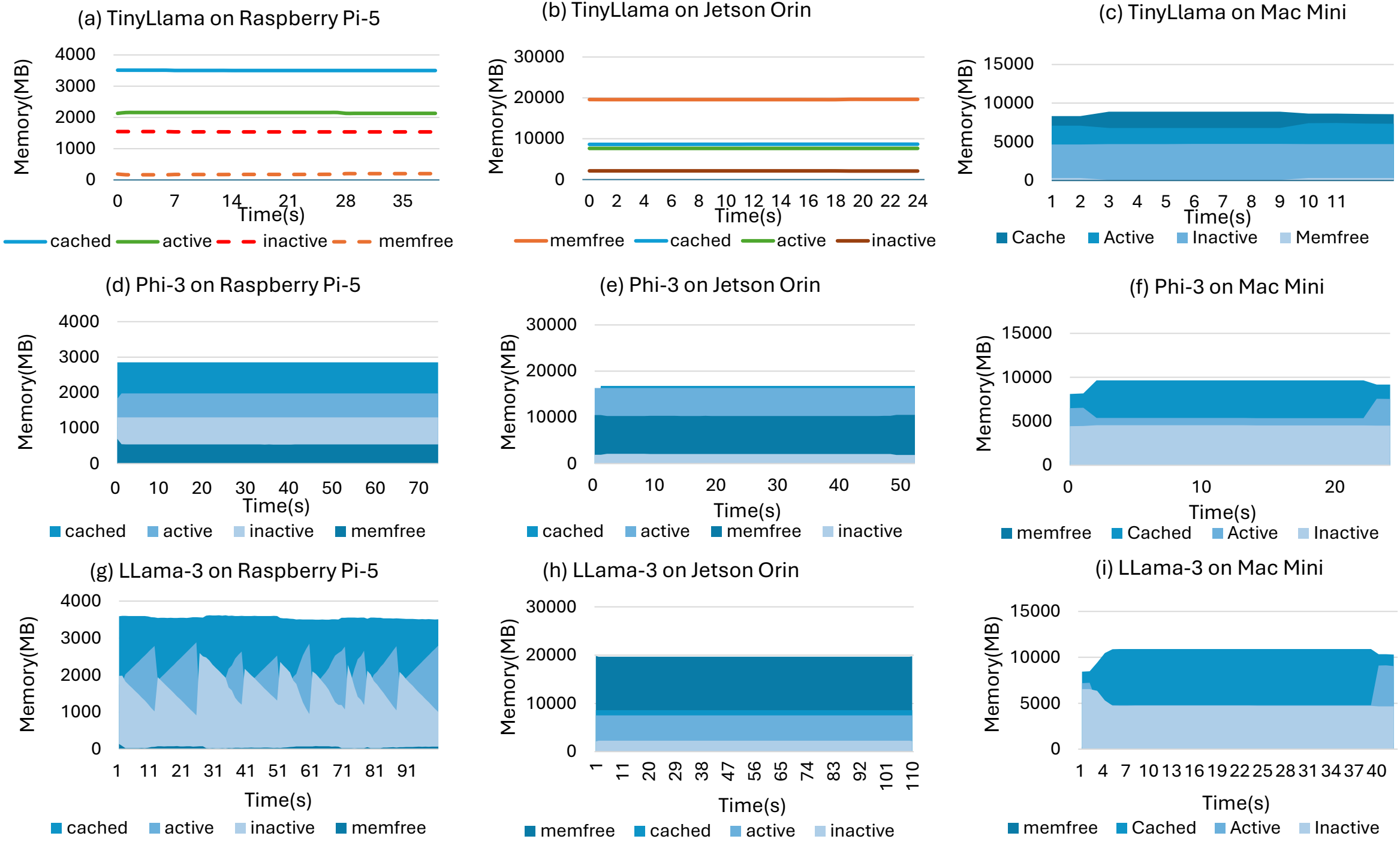

Performance (GFLOPS)
100%
80%
60%
40%
20%
0%
yyy
1
2
3
4
5
xxx

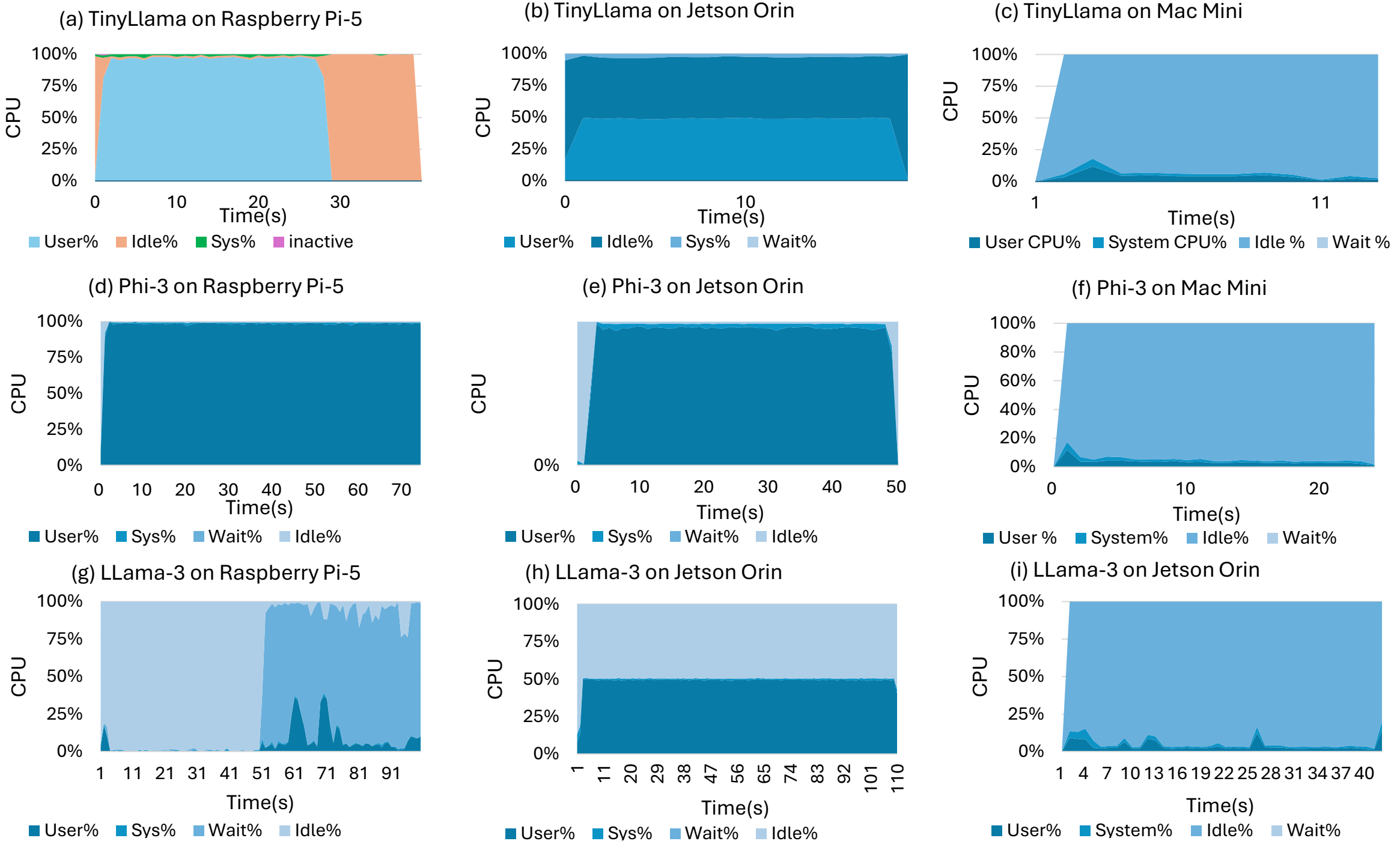

Performance (GFLOPS)
100%
80%
60%
40%
20%
0%
yyy
1
2
3
4
5
xxx

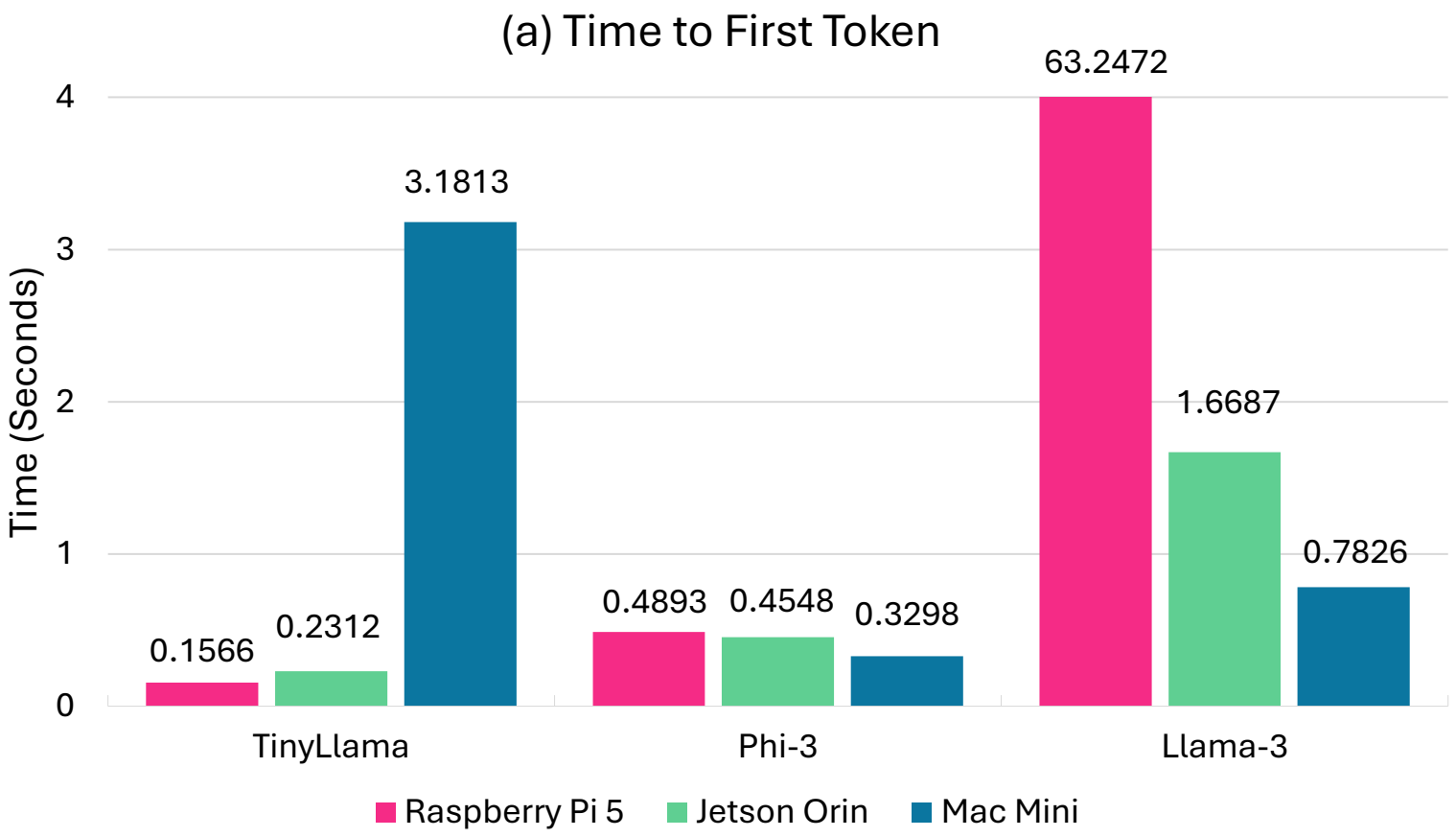

Performance (GFLOPS)
100%
80%
60%
40%
20%
0%
yyy
1
2
3
4
5
xxx

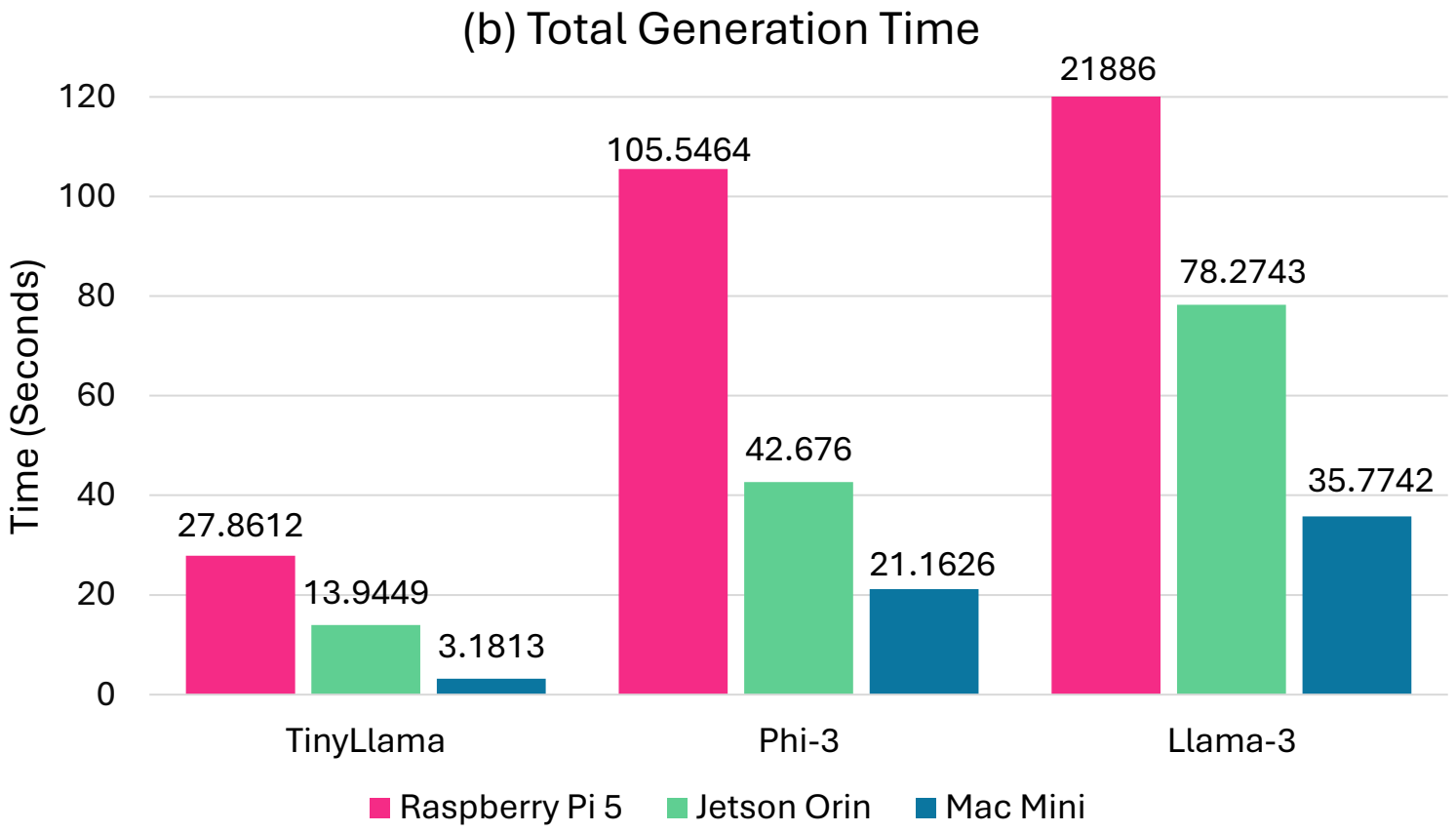

Performance (GFLOPS)
100%
80%
60%
40%
20%
0%
yyy
1
2
3
4
5
xxx

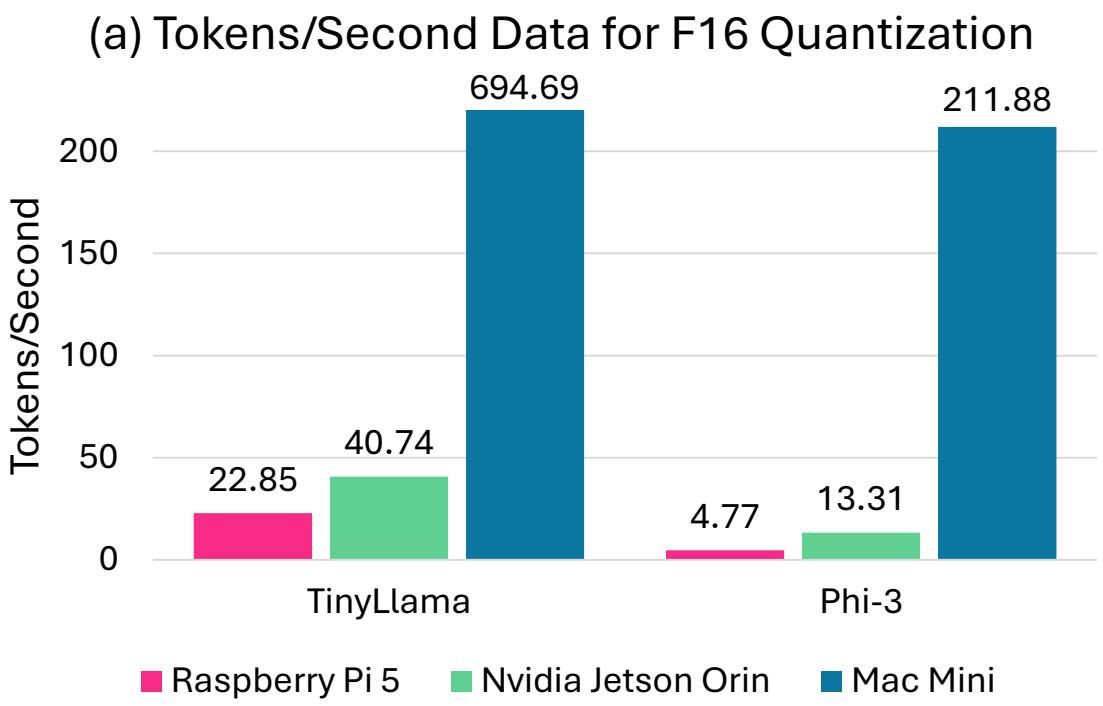

Performance (GFLOPS)
100%
80%
60%
40%
20%
0%
yyy
1
2
3
4
5
xxx

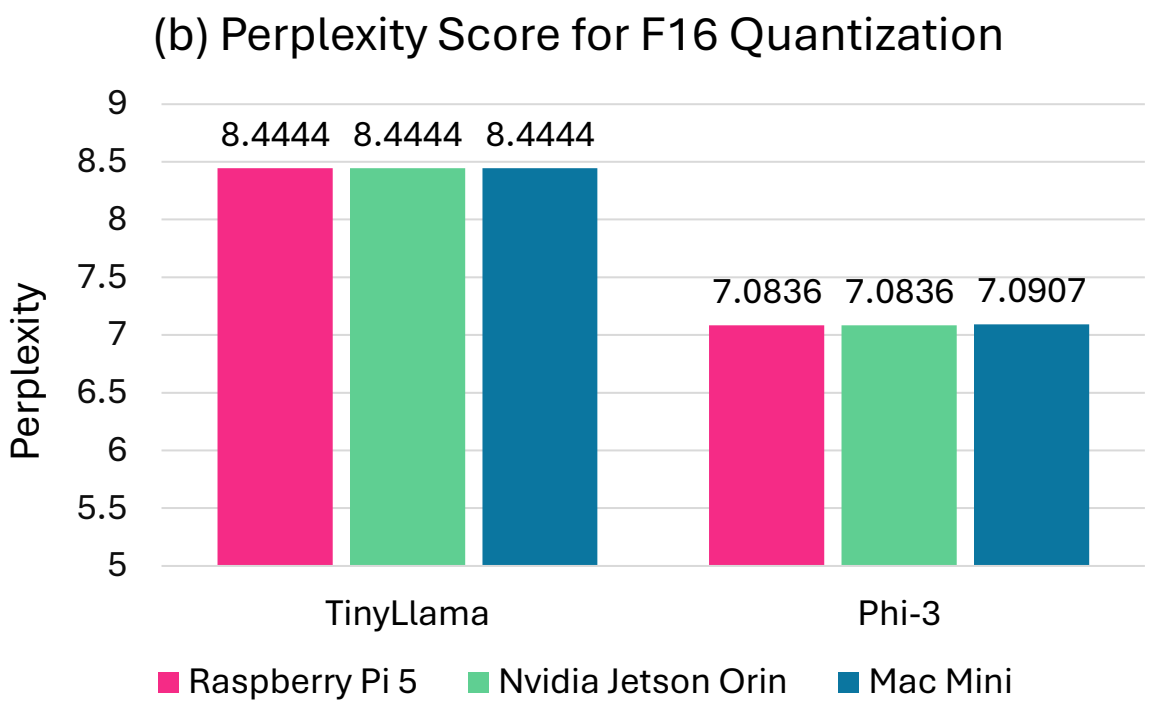

(a) TinyLlama on Raspberry Pi-5
Memory(MB)
4000
3000
2000
1000
0
0 7 14 21 28 35
Time(s)
cached active inactive memfree
(b) TinyLlama on Jetson Orin
Memory(MB)
30000
20000
10000
0
0 2 4 6 8 10 12 14 16 18 20 22 24
Time(s)
memfree cached active inactive
(c) TinyLlama on Mac Mini
Memory(MB)
15000
10000
5000
0
1 2 3 4 5 6 7 8 9 10 11
Time(s)
Cache Active Inactive Memfree
(d) Phi-3 on Raspberry Pi-5
Memory(MB)
4000
3000
2000
1000
0
0 10 20 30 40 50 60 70
Time(s)
cached active inactive memfree
(e) Phi-3 on Jetson Orin
Memory(MB)
30000
20000
10000
0
0 10 20 30 40 50
Time(s)
cached active memfree inactive
(f) Phi-3 on Mac Mini
Memory(MB)
15000
10000
5000
0
0 10 20
Time(s)
memfree Cached Active Inactive
(g) LLama-3 on Raspberry Pi-5
Memory(MB)
4000
3000
2000
1000
0
1 11 21 31 41 51 61 71 81 91
Time(s)
cached active inactive memfree
(h) LLama-3 on Jetson Orin
Memory(MB)
30000
20000
10000
0
1 11 20 29 38 47 56 65 74 83 92 101 110
Time(s)
memfree cached active inactive
(i) LLama-3 on Mac Mini
Memory(MB)
15000
10000
5000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40
Time(s)
memfree Cached Active Inactive

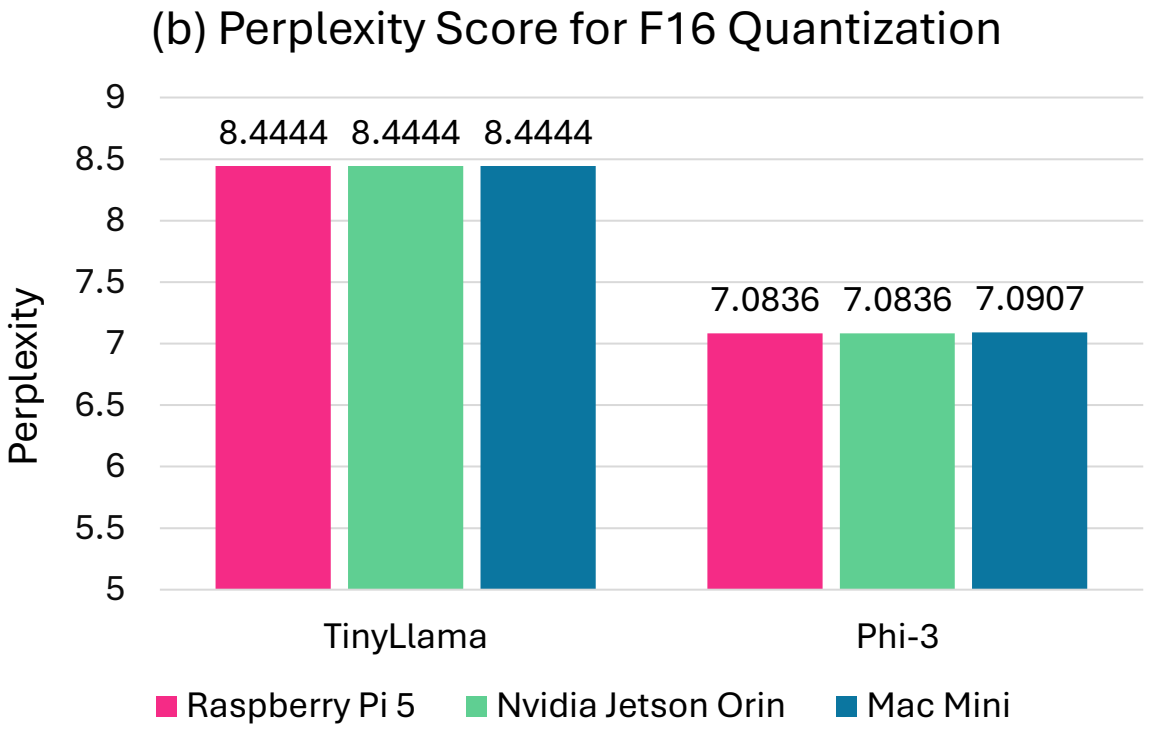

(a) TinyLlama on Raspberry Pi-5
CPU
100%
75%
50%
25%
0%
0
10
20
30
Time(s)
User%
Idle%
Sys%
inactive
(b) TinyLlama on Jetson Orin
CPU
100%
75%
50%
25%
0%
0
10
Time(s)
User%
Idle%
Sys%
Wait%
(c) TinyLlama on Mac Mini
CPU
100%
75%
50%
25%
0%
1
11
Time(s)
User CPU%
System CPU%
Idle %
Wait %
(d) Phi-3 on Raspberry Pi-5
CPU
100%
75%
50%
25%
0%
0
10
20
30
40
50
60
70
Time(s)
User%
Sys%
Wait%
Idle%
(e) Phi-3 on Jetson Orin
CPU
0%
0
10
20
30
40
50
Time(s)
User%
Sys%
Wait%
Idle%
(f) Phi-3 on Mac Mini
CPU
100%
80%
60%
40%
20%
0%
0
10
20
Time(s)
User %
System%
Idle%
Wait%
(g) LLama-3 on Raspberry Pi-5
CPU
100%
75%
50%
25%
0%
1
11
21
31
41
51
61
71
81
91
Time(s)
User%
Sys%
Wait%
Idle%
(h) LLama-3 on Jetson Orin
CPU
100%
75%
50%
25%
0%
1
11
20
29
38
47
56
65
74
83
92
101
110
Time(s)
User%
Sys%
Wait%
Idle%
(i) LLama-3 on Jetson Orin
CPU
100%
75%
50%
25%
0%
1
4
7
10
13
16
19
22
25
28
31
34
37
40
Time(s)
User%
System%
Idle%
Wait%

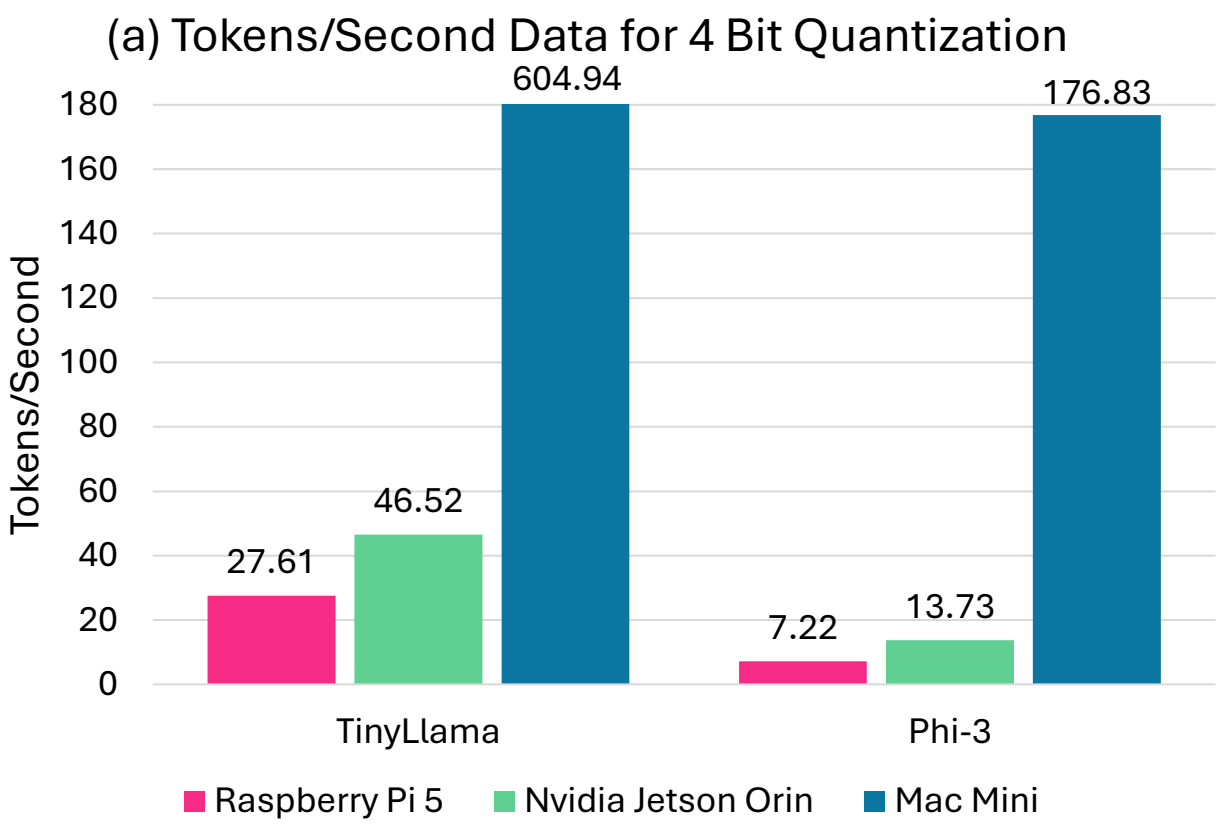

(a) Time to First Token
Time (Seconds)
4
3
2
1
0
0.1566
0.2312
3.1813
0.4893
0.4548
0.3298
63.2472
1.6687
0.7826
TinyLlama
Phi-3
Llama-3
Raspberry Pi 5
Jetson Orin
Mac Mini

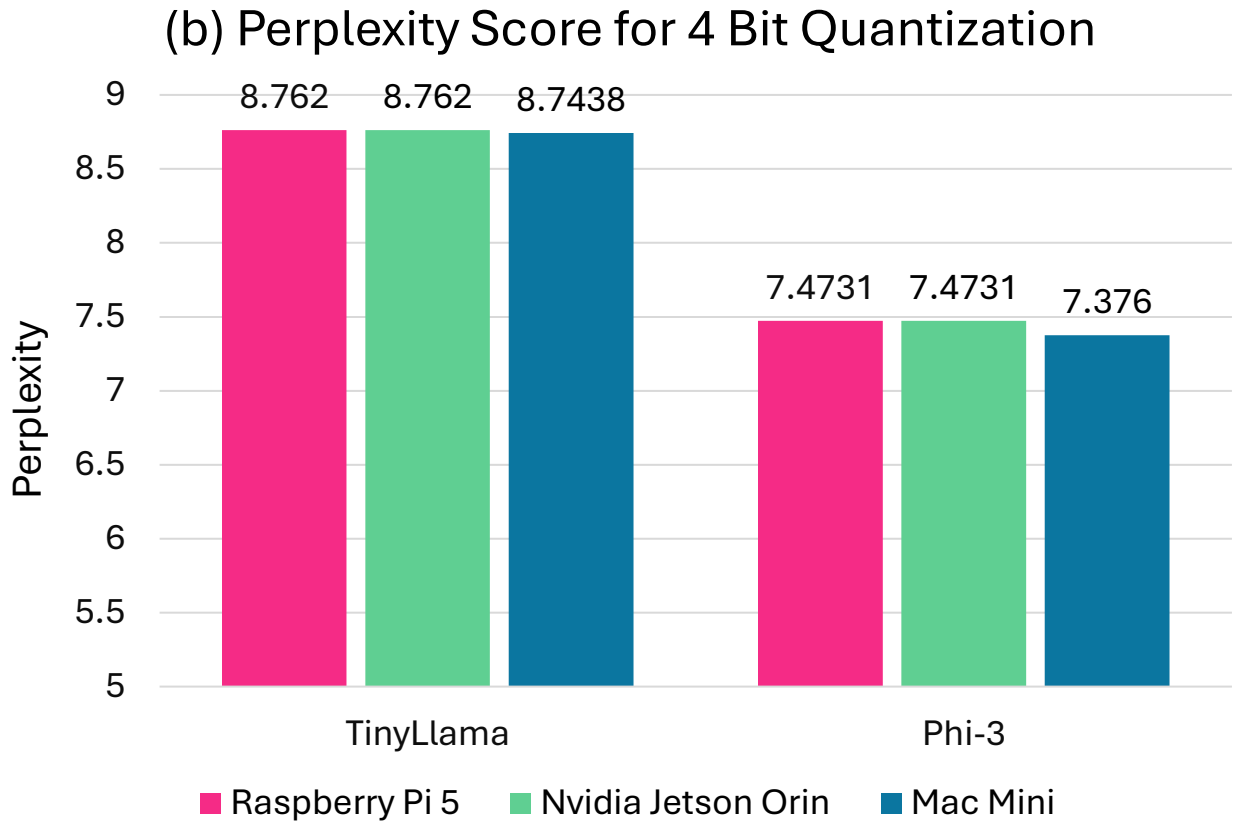

(b) Total Generation Time
Time (Seconds)
120
100
80
60
40
20
0
27.8612
13.9449
3.1813
105.5464
42.676
21.1626
21886
78.2743
35.7742
TinyLlama
Phi-3
Llama-3
Raspberry Pi 5
Jetson Orin
Mac Mini

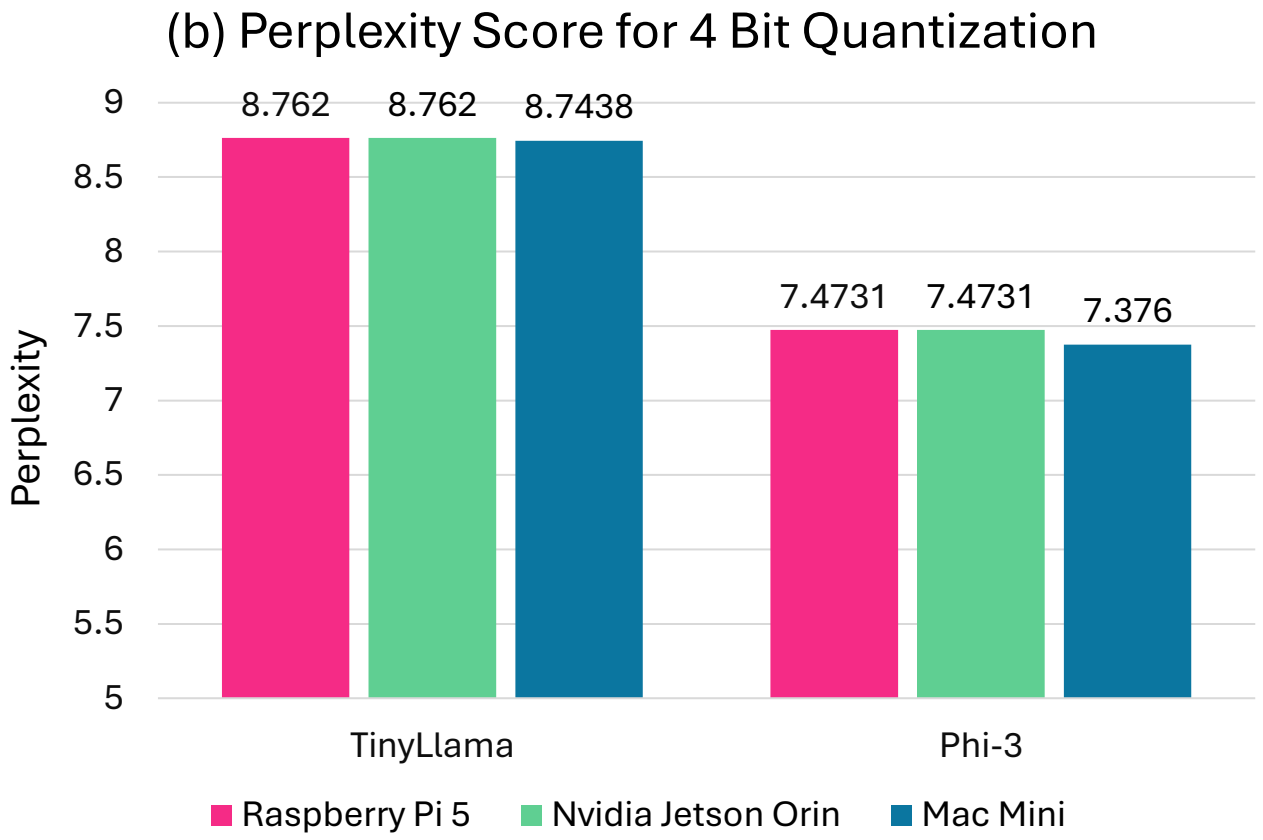

(a) Tokens/Second Data for F16 Quantization
Tokens/Second
200
150
100
50
0
694.69
211.88
40.74
22.85
4.77
13.31
TinyLlama
Phi-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini

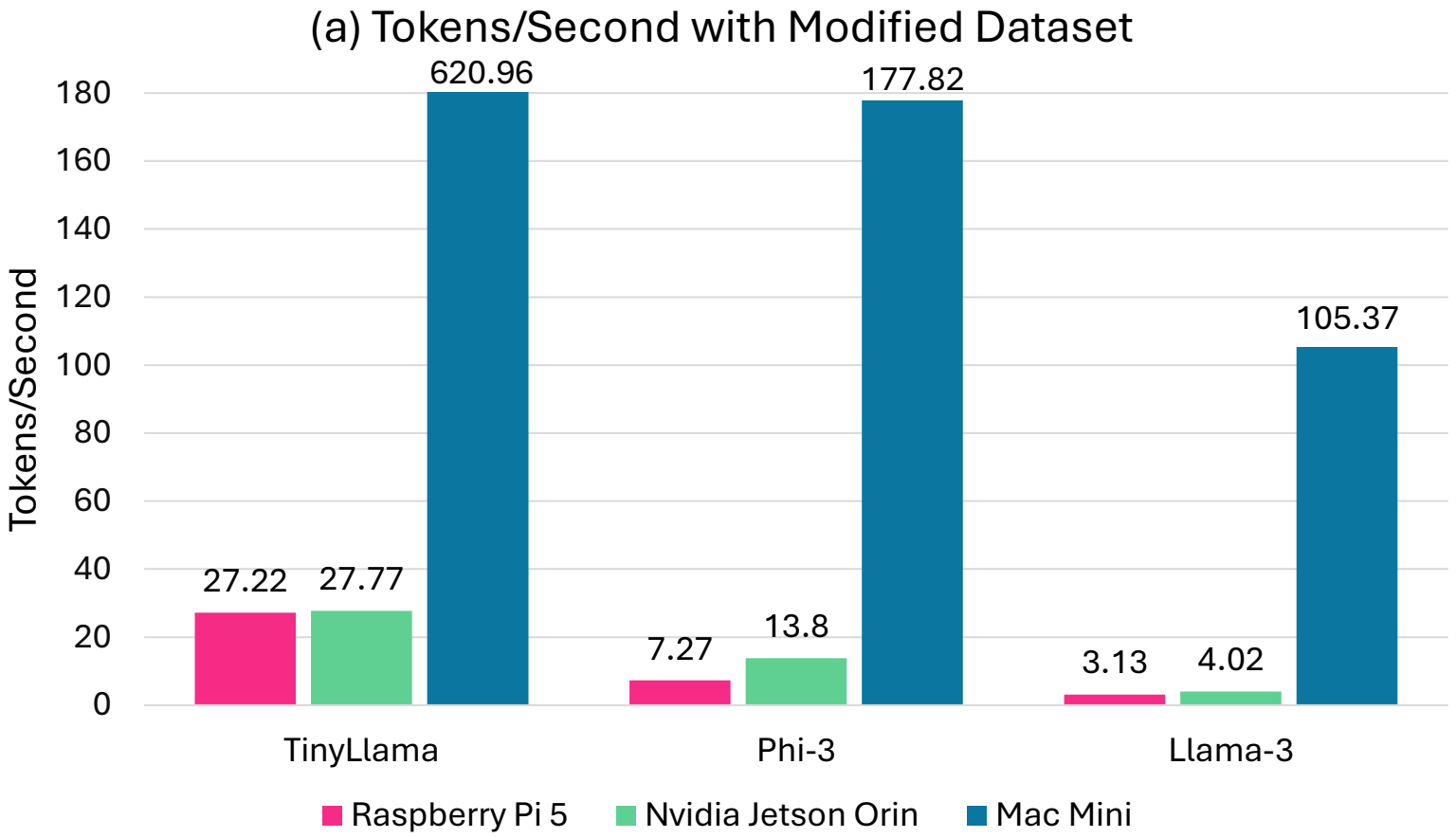

(b) Perplexity Score for F16 Quantization
Perplexity
9
8.5
8
7.5
7
6.5
6
5.5
5
8.4444
8.4444
8.4444
7.0836
7.0836
7.0907
TinyLlama
Phi-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini

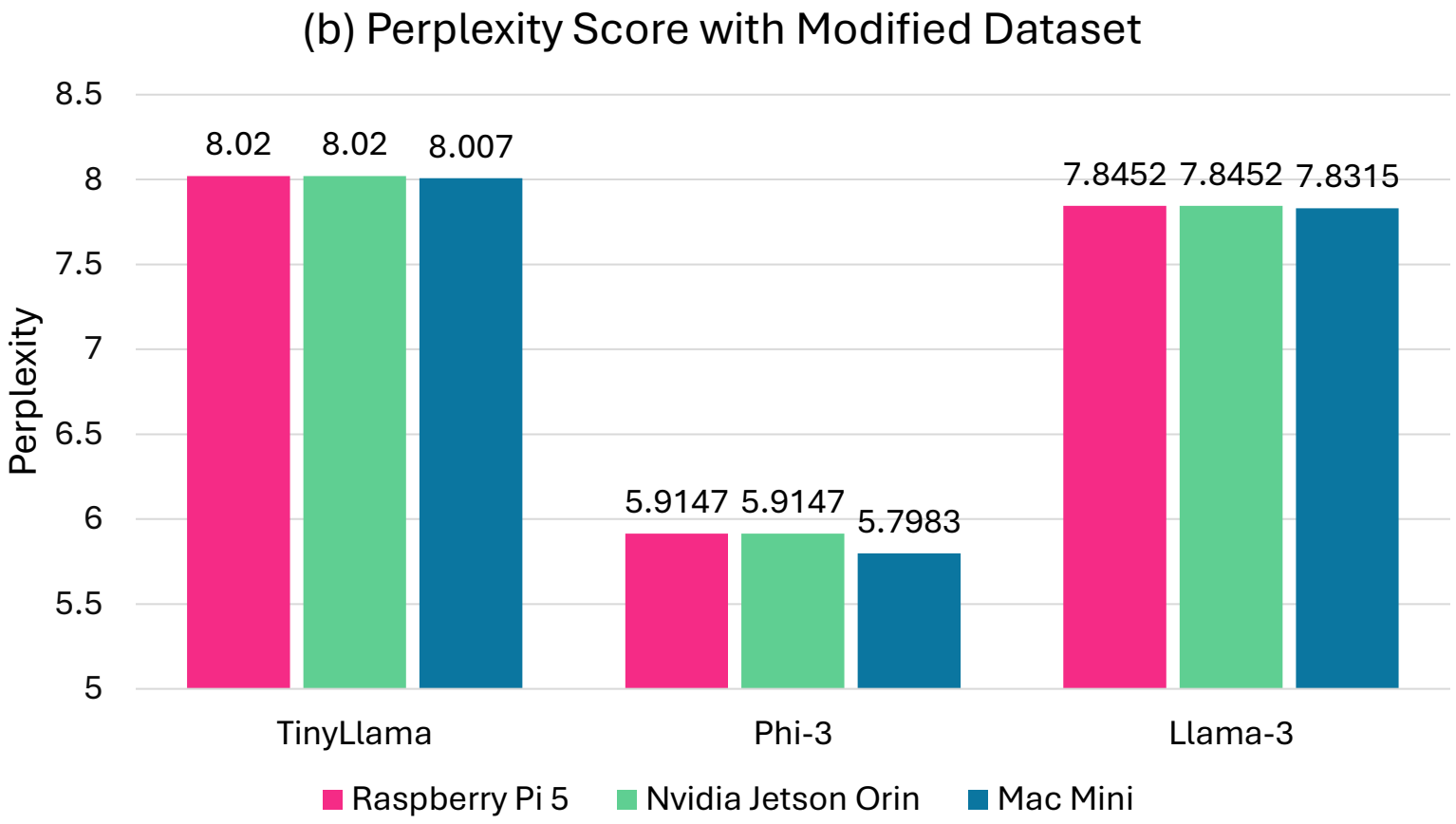

(b) Perplexity Score for F16 Quantization
Perplexity
9
8.5
8
7.5
7
6.5
6
5.5
5
8.4444
8.4444
8.4444
7.0836
7.0836
7.0907
TinyLlama
Phi-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini


(a) Tokens/Second Data for 4 Bit Quantization
Tokens/Second
180
160
140
120
100
80
60
40
20
0
604.94
176.83
46.52
27.61
13.73
7.22
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


(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


(a) Tokens/Second with Modified Dataset
Tokens/Second
180
160
140
120
100
80
60
40
20
0
620.96
177.82
105.37
27.22
27.77
7.27
13.8
3.13
4.02
TinyLlama
Phi-3
Llama-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini


(b) Perplexity Score with Modified Dataset
Perplexity
8.5
8
7.5
7
6.5
6
5.5
5
8.02
8.02
8.007
5.9147
5.9147
5.7983
7.8452
7.8452
7.8315
TinyLlama
Phi-3
Llama-3
Raspberry Pi 5
Nvidia Jetson Orin
Mac Mini