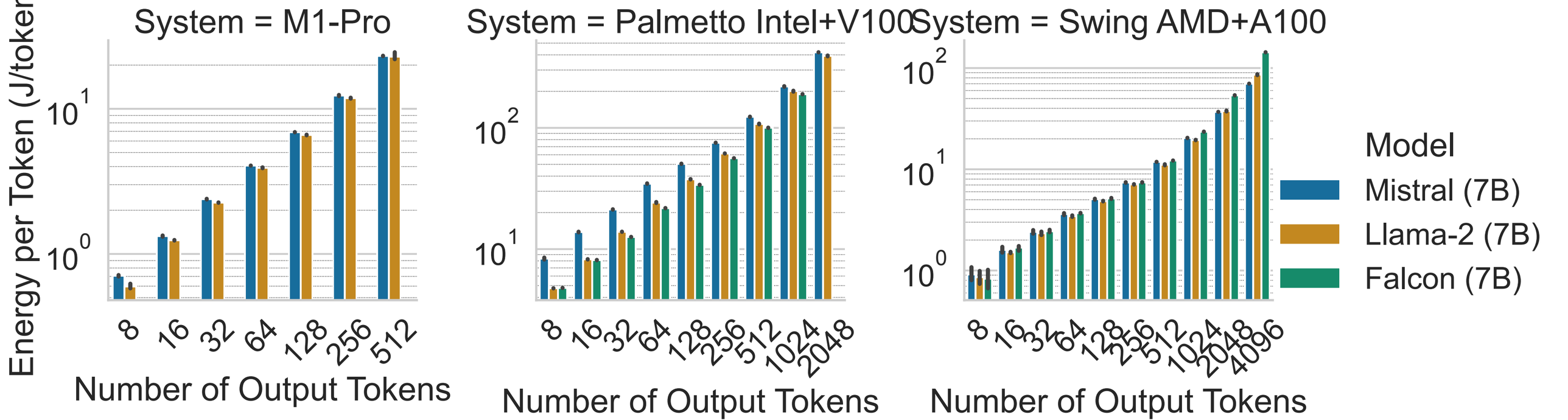
System = M1-Pro
System = Palmetto Intel+V100
System = Swing AMD+A100
Energy per Token (J/token)
10^0
10^1
10^2
8
16
32
64
128
256
512
1024
2048
4096
Number of Output Tokens
Model
Mistral (7B)
Llama-2 (7B)
Falcon (7B)