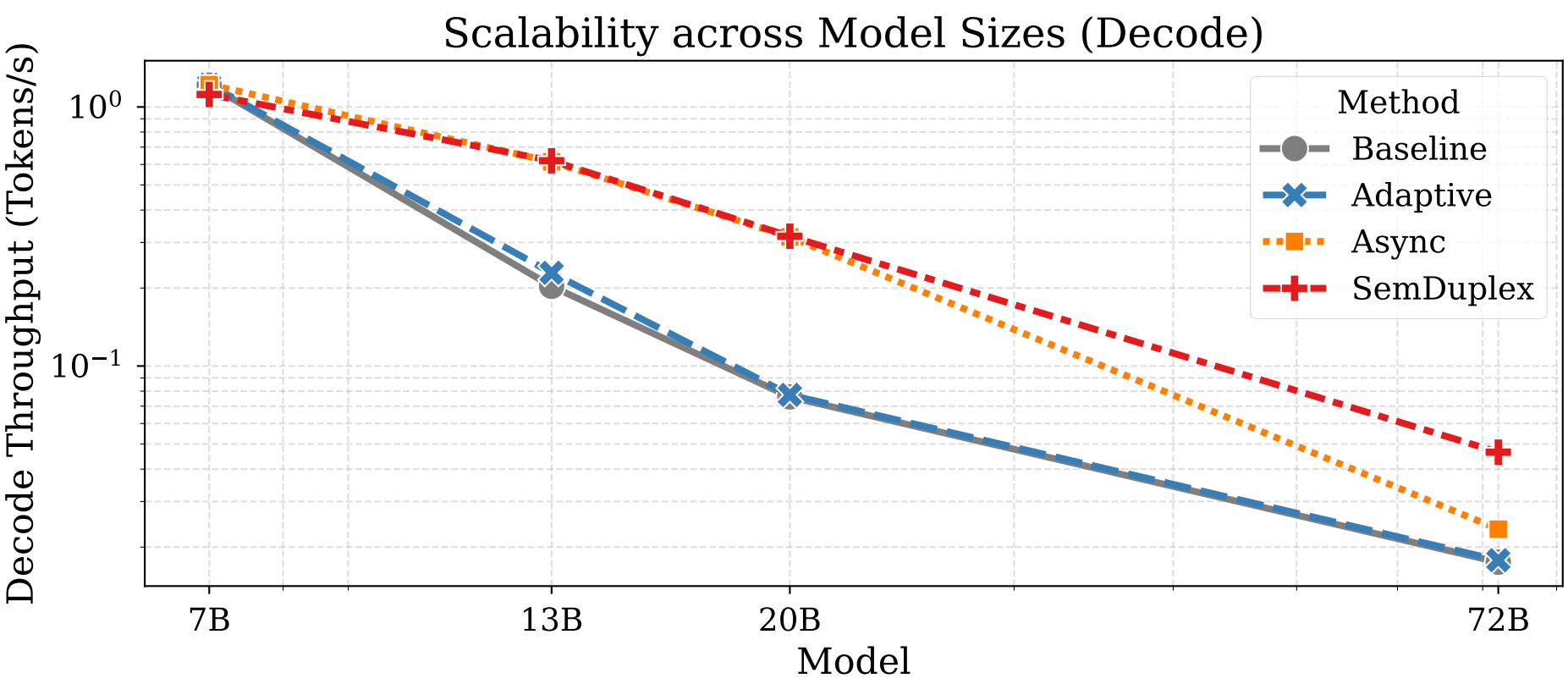
Scalability across Model Sizes (Decode)
Decode Throughput (Tokens/s)
10^0
10^{-1}
7B
13B
20B
72B
Model
Method
Baseline
Adaptive
Async
SemDuplex