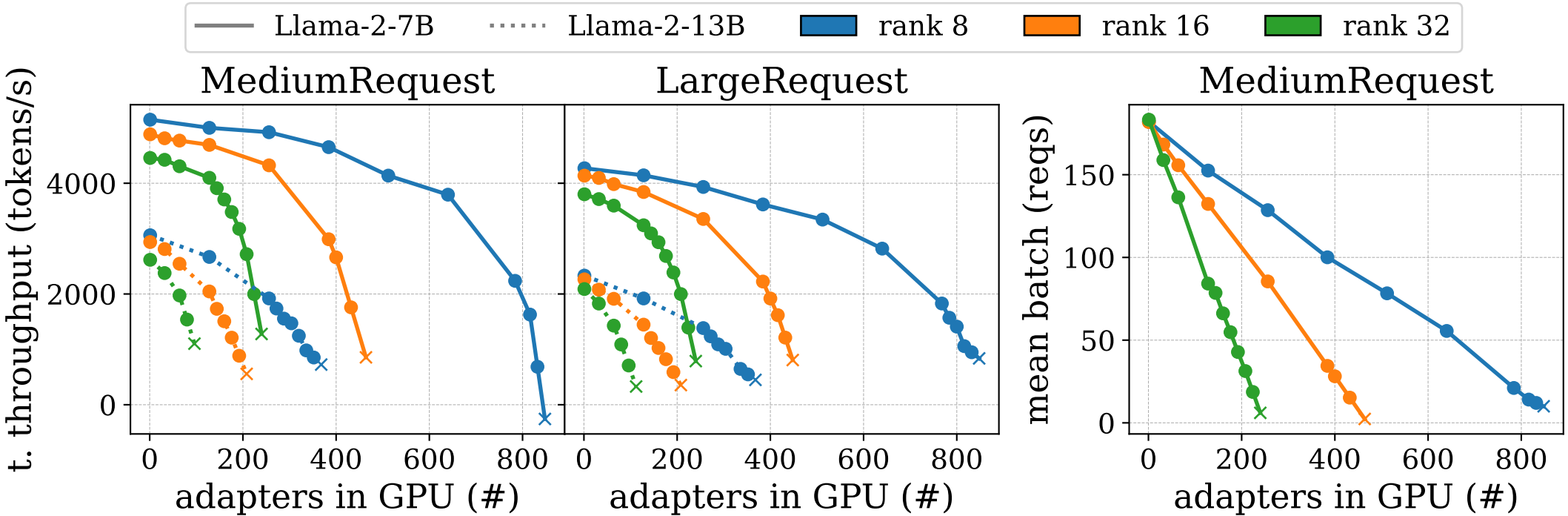
Llama-2-7B
Llama-2-13B
rank 8
rank 16
rank 32
MediumRequest
LargeRequest
MediumRequest
t. throughput (tokens/s)
mean batch (reqs)
adapters in GPU (#)
adapters in GPU (#)
adapters in GPU (#)
0
2000
4000
0
50
100
150
0
200
400
600
800