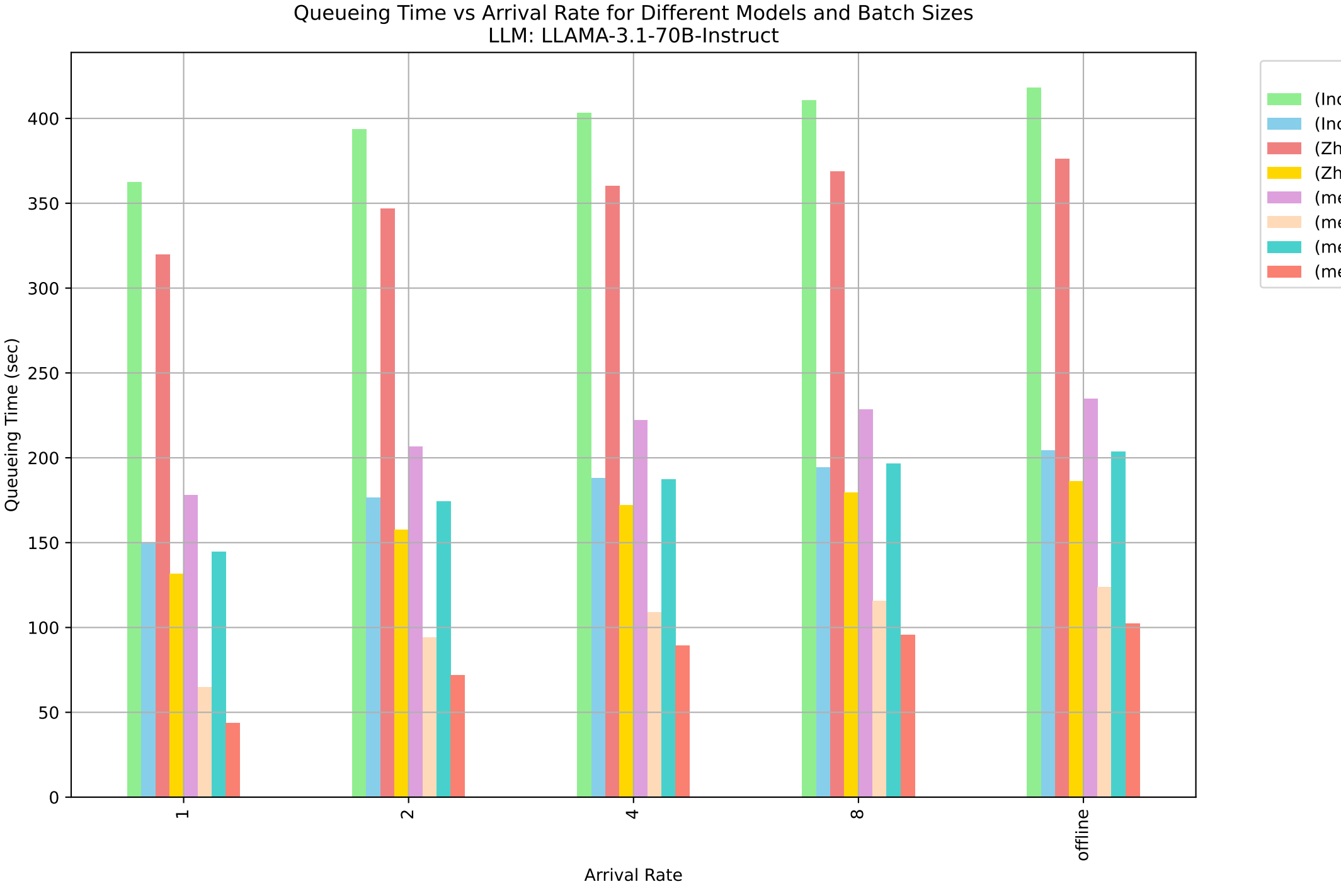
Queueing Time vs Arrival Rate for Different Models and Batch Sizes
LLM: LLAMA-3.1-70B-Instruct
Queueing Time (sec)
0
50
100
150
200
250
300
350
400
1
2
4
8
offline
Arrival Rate
(Inc
(Inc
(Zh
(Zh
(me
(me
(me
(me