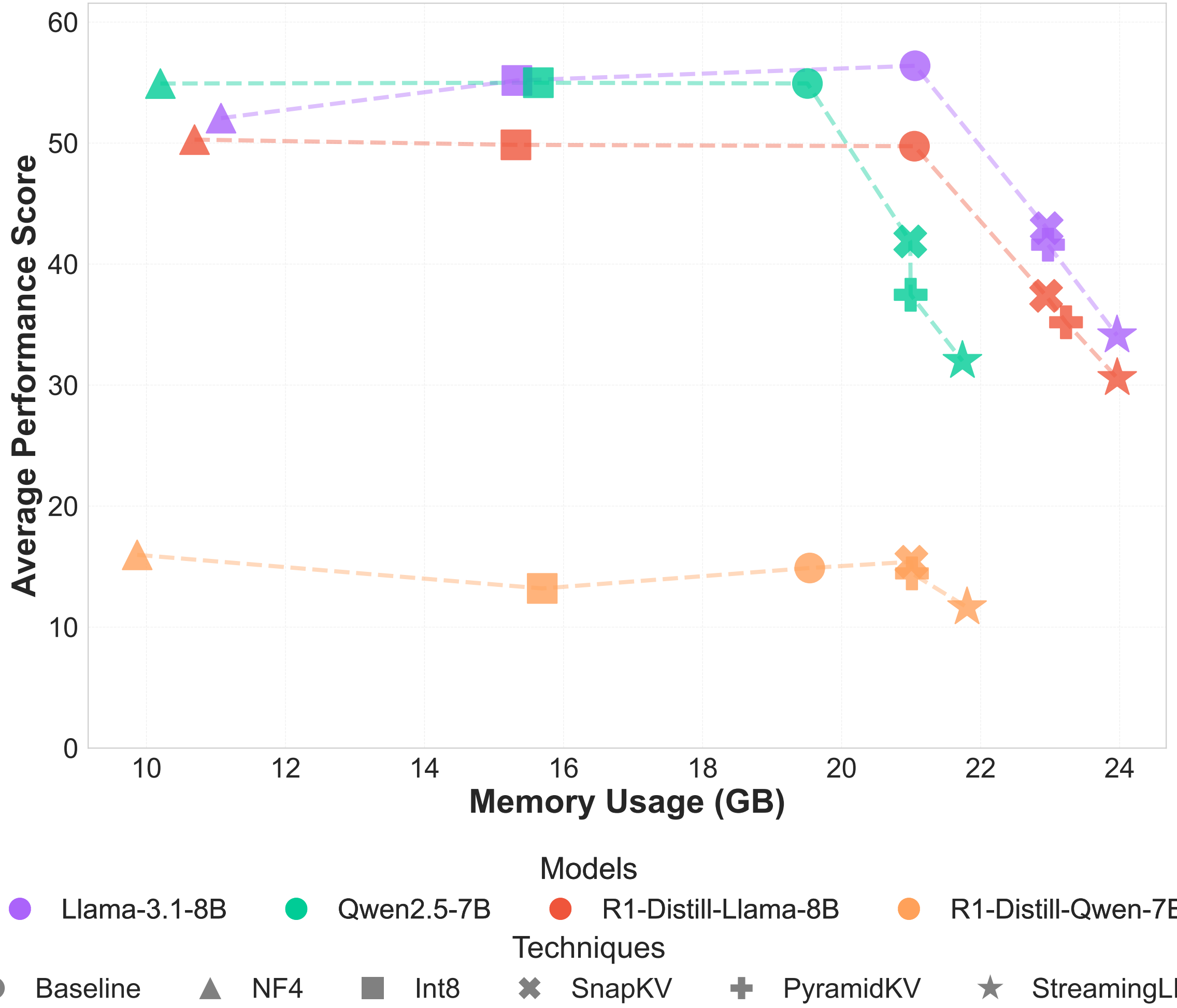
Average Performance Score
60
50
40
30
20
10
0
10
12
14
16
18
20
22
24
Memory Usage (GB)
Models
Llama-3.1-8B
Qwen2.5-7B
R1-Distill-Llama-8B
R1-Distill-Qwen-7B
Techniques
Baseline
NF4
Int8
SnapKV
PyramidKV
StreamingLLM