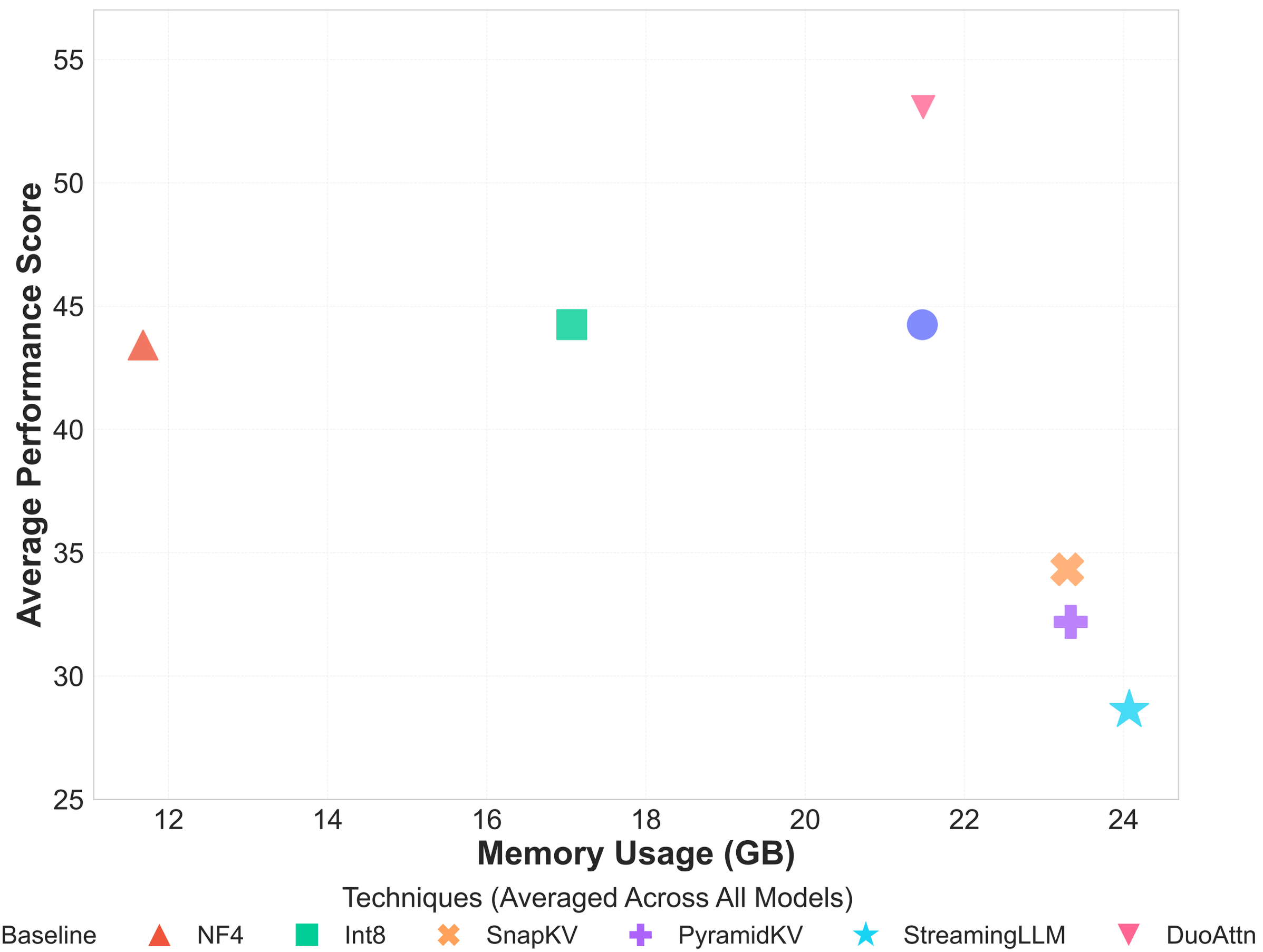
Average Performance Score
55
50
45
40
35
30
25
12
14
16
18
20
22
24
Memory Usage (GB)
Techniques (Averaged Across All Models)
Baseline
NF4
Int8
SnapKV
PyramidKV
StreamingLLM
DuoAttn