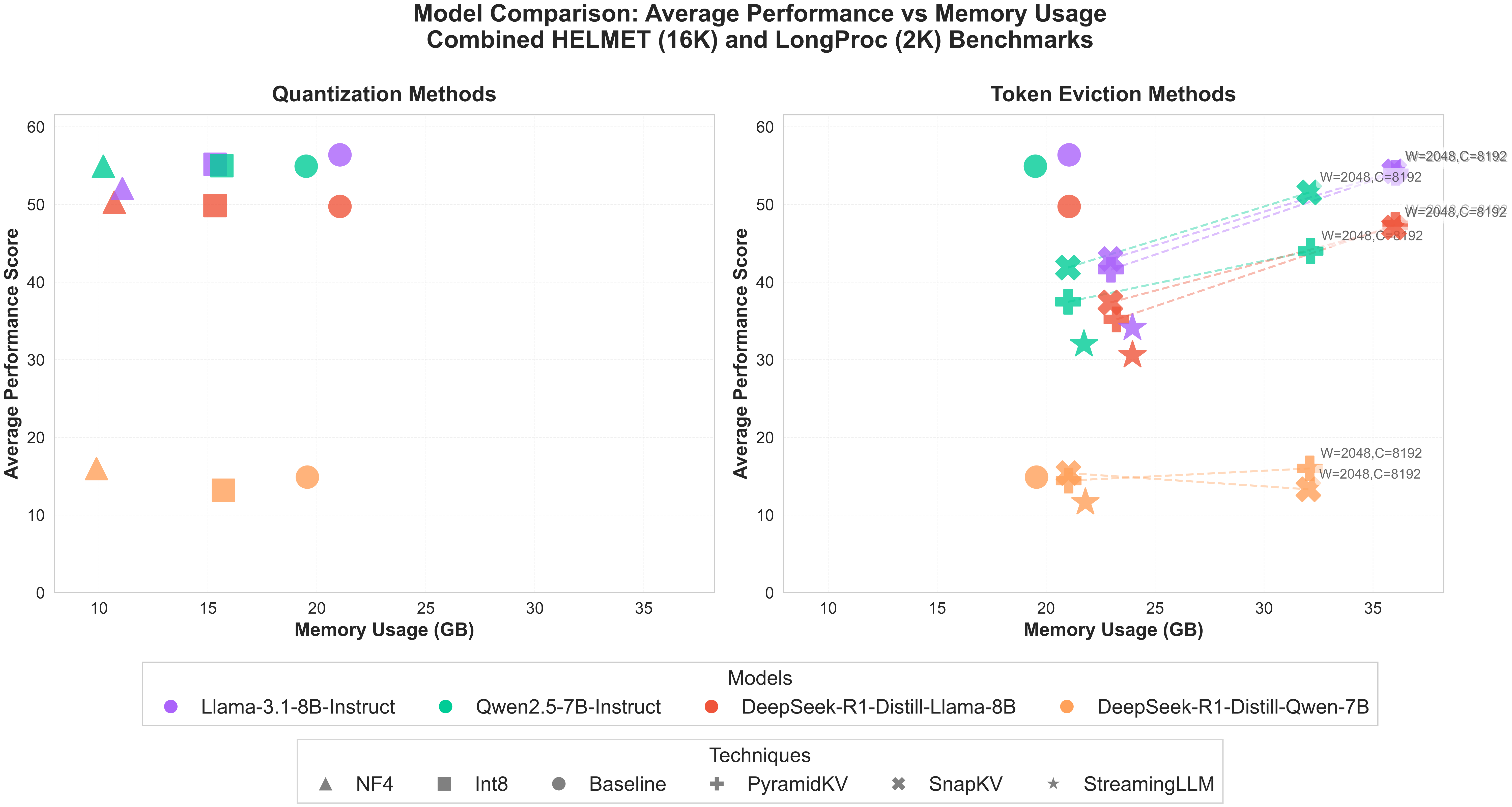
Model Comparison: Average Performance vs Memory Usage
Combined HELMET (16K) and LongProc (2K) Benchmarks
Quantization Methods
Token Eviction Methods
Average Performance Score
Memory Usage (GB)
0
10
20
30
40
50
60
10
15
20
25
30
35
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
Models
● Llama-3.1-8B-Instruct
● Qwen2.5-7B-Instruct
● DeepSeek-R1-Distill-Llama-8B
● DeepSeek-R1-Distill-Qwen-7B
Techniques
▲ NF4
■ Int8
● Baseline
✚ PyramidKV
✖ SnapKV
★ StreamingLLM