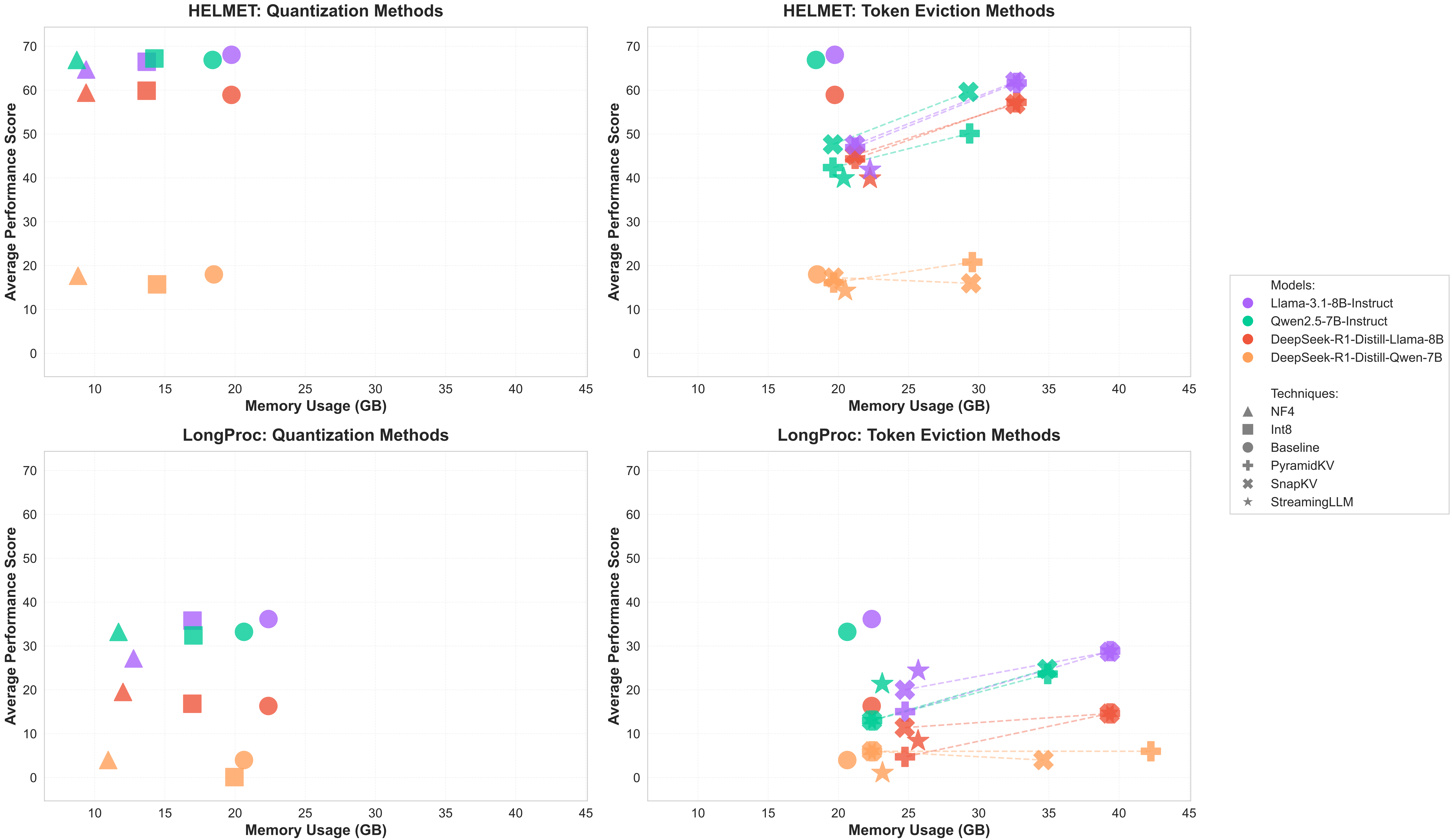
Model Comparison: Average Performance vs Memory Usage
HELMET (16K) and LongProc (2K) Benchmarks


HELMET: Quantization Methods
HELMET: Token Eviction Methods
LongProc: Quantization Methods
LongProc: Token Eviction Methods
Average Performance Score
Memory Usage (GB)
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