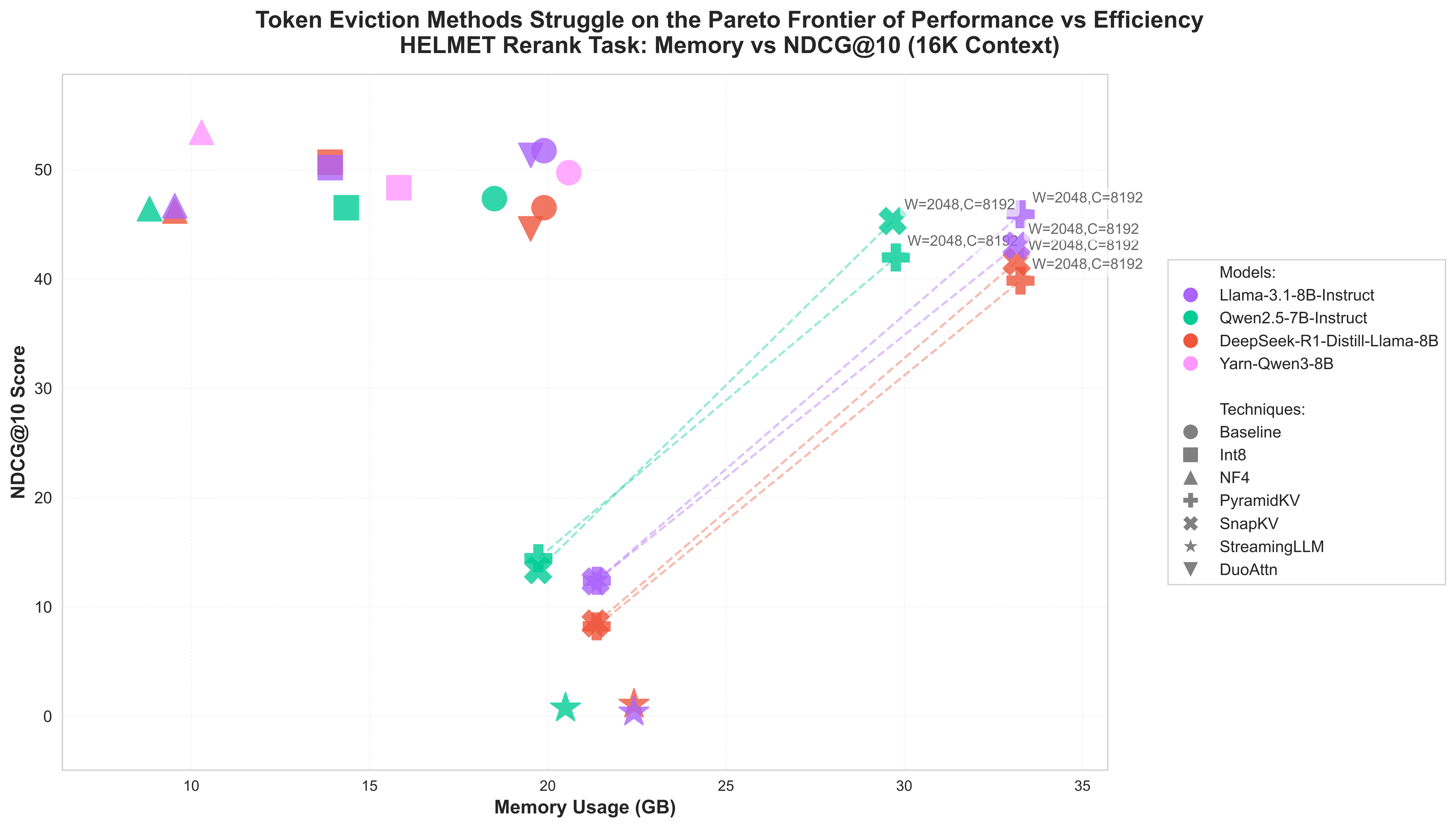
Token Eviction Methods Struggle on the Pareto Frontier of Performance vs Efficiency
HELMET Rerank Task: Memory vs NDCG@10 (16K Context)
NDCG@10 Score
Memory Usage (GB)
0
10
20
30
40
50
10
15
20
25
30
35
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
W=2048,C=8192
Models:
Llama-3.1-8B-Instruct
Qwen2.5-7B-Instruct
DeepSeek-R1-Distill-Llama-8B
Yarn-Qwen3-8B
Techniques:
Baseline
Int8
NF4
PyramidKV
SnapKV
StreamingLLM
DuoAttn