Task-wise Performance Delta: Efficient Techniques vs Baseline Averaged Across Models and Context Lengths

Y-axis: **Δ Performance = Efficient Technique Performance − Baseline Performance**

X-axis: **Tasks from HELMET (16K/32K) and LongProc (500/2K)**

Panels (top to bottom):
- DuoAttention (retrieval head sparsity = 50%)
- SnapKV (cache size setting of w=256, c=2048, k=7, maxpool)
- SnapKV (cache size setting of w=2048, c=8192, k=7, maxpool)
- PyramidKV (cache size setting of w=256, c=2048, k=7, avgpool)
- PyramidKV (cache size setting of w=2048, c=8192, k=7, avgpool)
- StreamingLLM (n_local = 4092, n_init = 4)
- NF4
- Int8

X-axis task labels: niah, recall_jsonkv, rag_hotpotqa, rag_nq, cite_str_em, cite_rec, cite_prec, rerank, icl_clinc, icl_banking, multi_lexsum, pseudo_to_code, html_to_tsv, travel_planning