

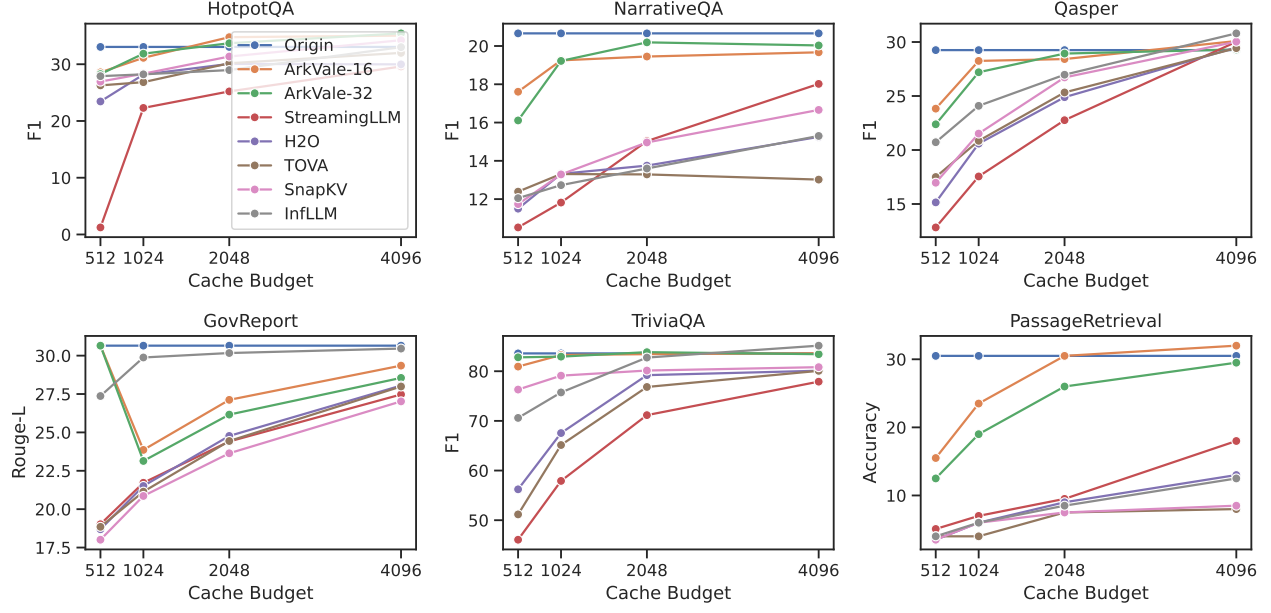


Figure 1: Evaluation of LongChat-7B-v1.5-32k on LongBench with different methods and cache-budgets.

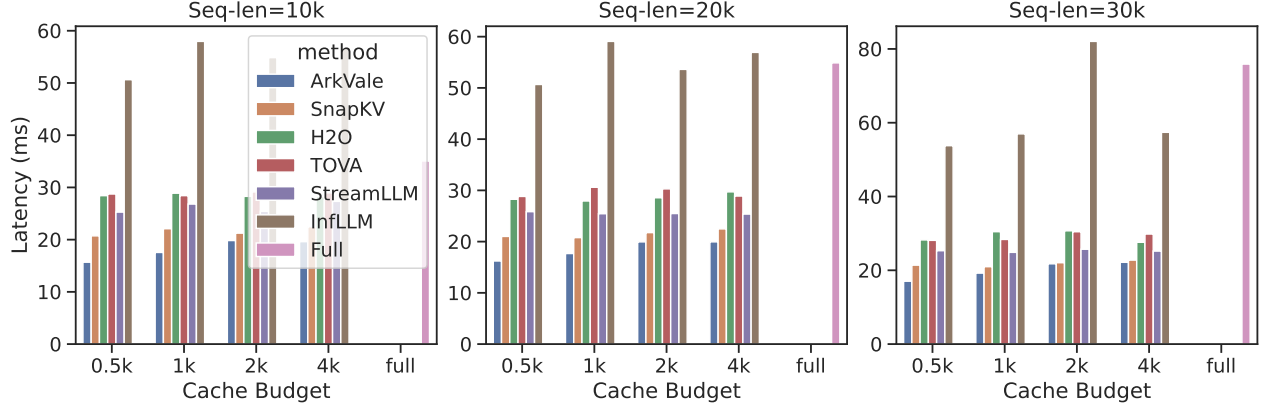
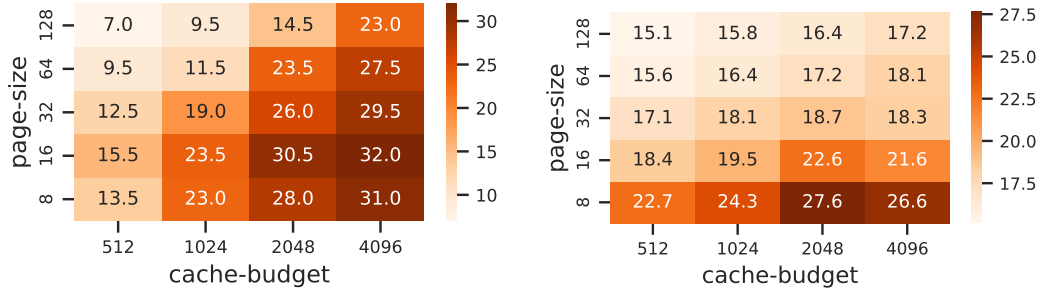


Figure 2: Per-step decoding latency of LongChat-7B-v1.5-32k with different methods, sequence-lengths, and cache-budgets.



(a) Accuracy

(b) Average per-step decoding latency (ms)

Figure 3: Evaluation of ARKVALE adapted to LongChat-7b-v1.5-32k on Passage-Retrieval with different cache-budgets and page-sizes