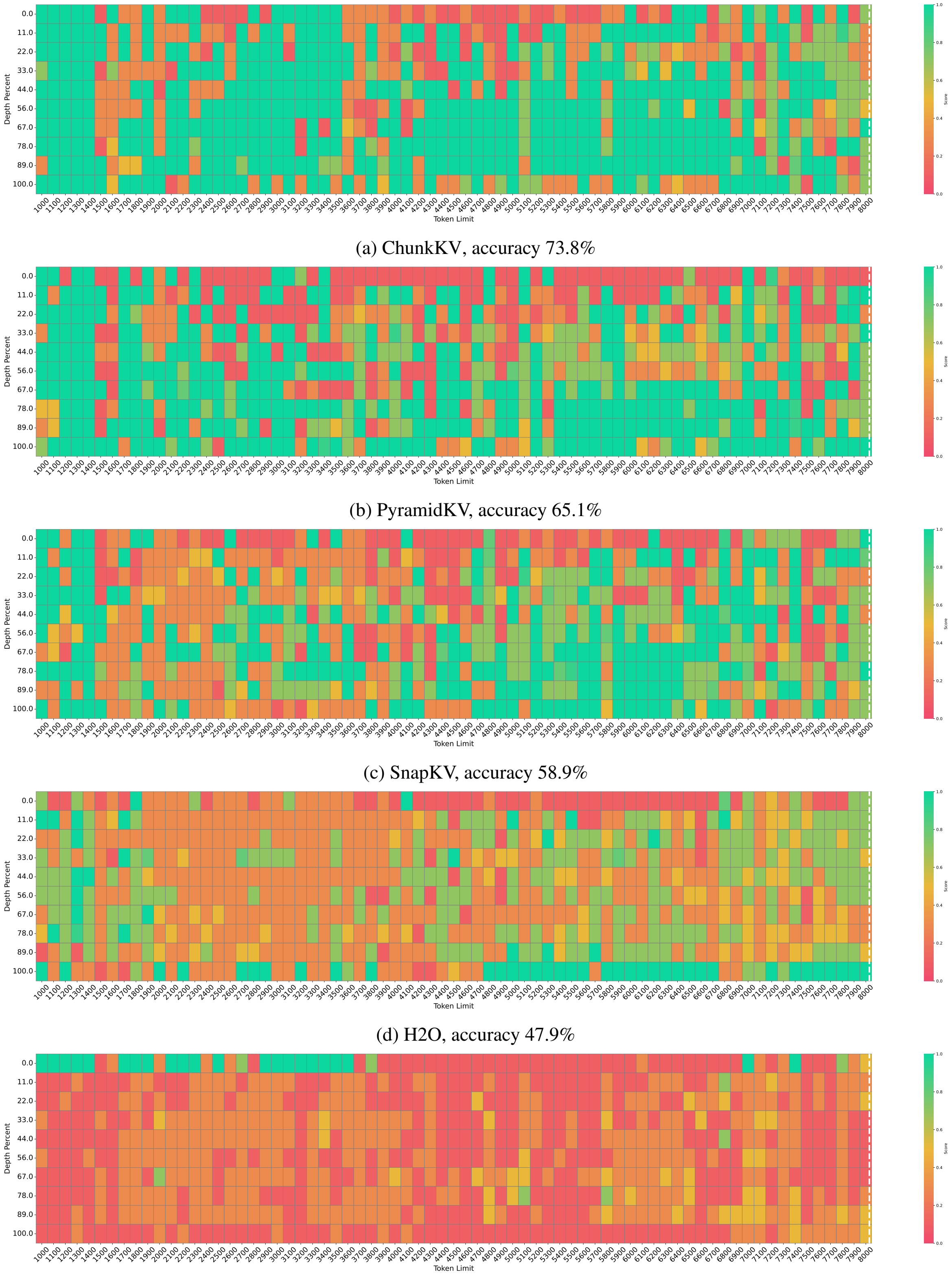

(a) ChunkKV, accuracy 73.8%

(b) PyramidKV, accuracy 65.1%

(c) SnapKV, accuracy 58.9%

(d) H2O, accuracy 47.9%

(e) StreamingLLM, accuracy 23.7%

Figure 17: NIAH benchmark for LLaMA-3-8B-Instruct with KV cache size=128 under 8k context length