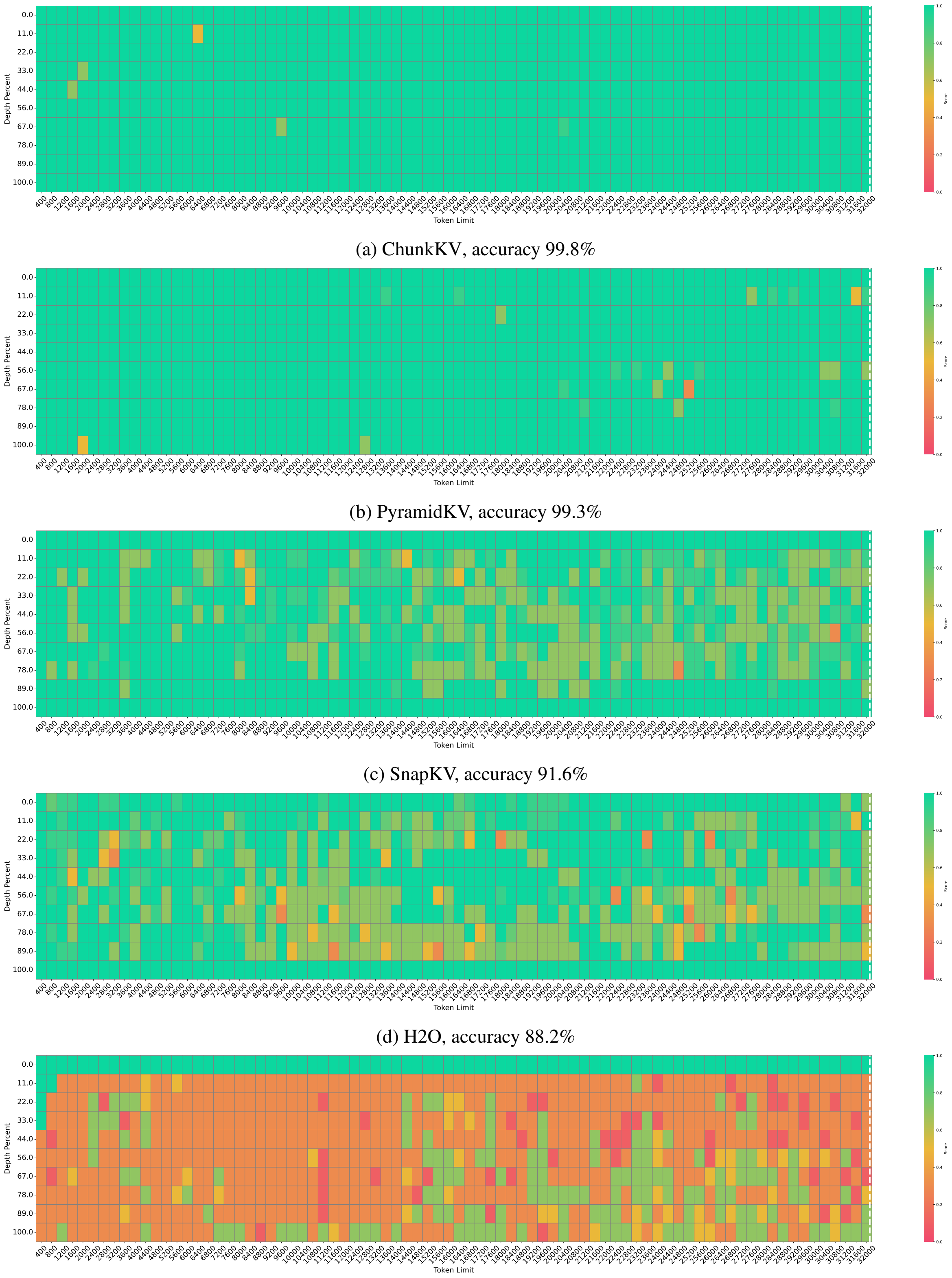

(a) ChunkKV, accuracy 99.8%

(b) PyramidKV, accuracy 99.3%

(c) SnapKV, accuracy 91.6%

(d) H2O, accuracy 88.2%

(e) StreamingLLM, accuracy 44.3%

Figure 18: NIAH benchmark for Mistral-7B-Instruct with KV cache size=128 under 32k context length