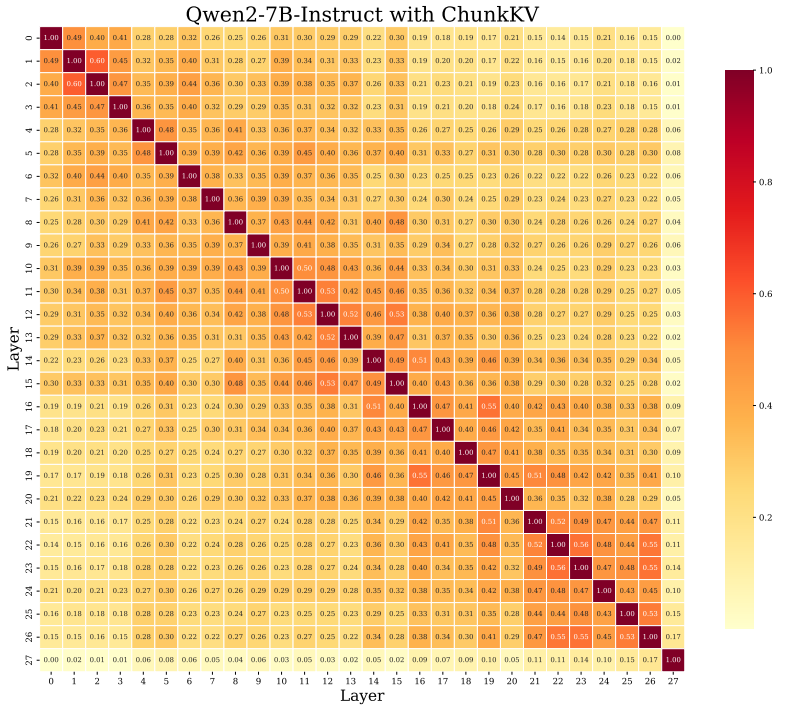

Figure 15: Layer-wise similarity heatmaps of the preserved KV cache indices by ChunkKV on Qwen2-7B-Instruct

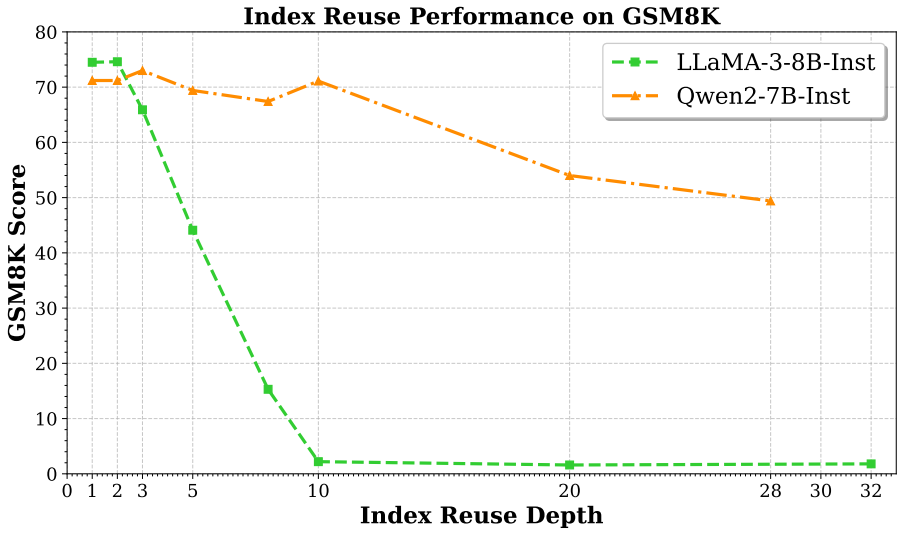

Figure 16: GSM8K Performance Comparison with different index reuse layers

Table 13 shows the performance of ChunkKV with different numbers of index reuse layers in GSM8K. The number of index reuse layers is set from 1 to the number of layers in the model, where a index reuse layer of 1 corresponds to the normal ChunkKV without index reuse, and 28/32 is the maximum number of layers for LLaMA-3-8B-Instruct and Qwen2-7B-Instruct. The significant performance drop of LLaMA-3-8B-Instruct raises another question: whether the KV cache compression method is more sensitive to the model's mathematical reasoning ability.