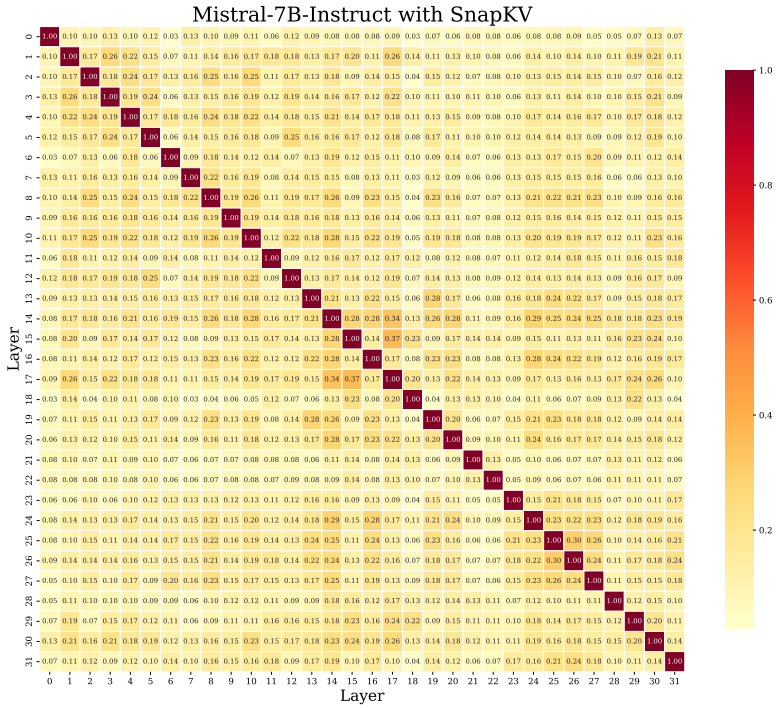

Figure 11: Layer-wise similarity heatmaps of the preserved KV cache indices by SnapKV on Mistral-7B-Instruct

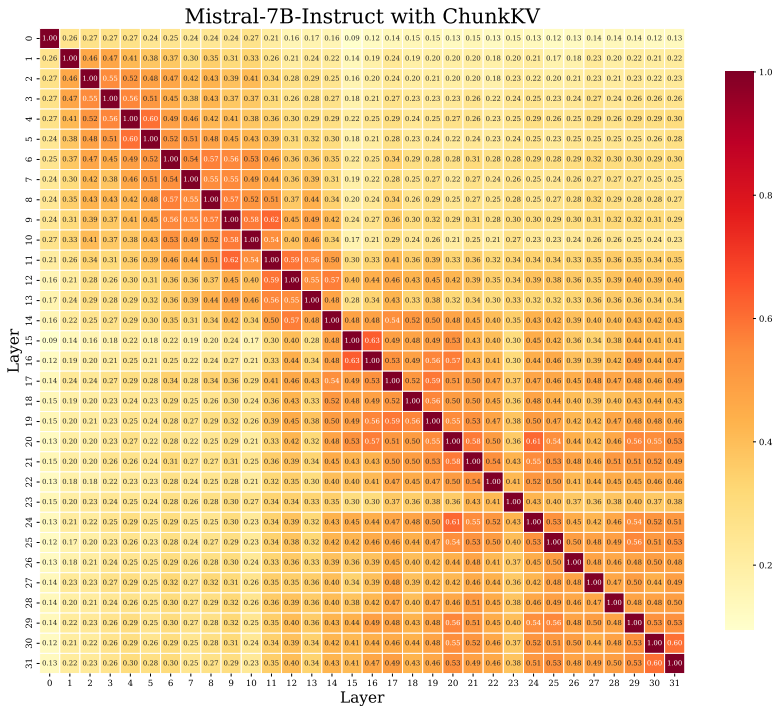

Figure 12: Layer-wise similarity heatmaps of the preserved KV cache indices by ChunkKV on Mistral-7B-Instruct

B.1.3. Index Reuse Performance

Figure 16 illustrates the performance of ChunkKV with varying index reuse layers on the GSM8K benchmark. The experiment reveals that math problems are more sensitive to index reuse layers compared to LongBench. Both LLaMA3-8B-