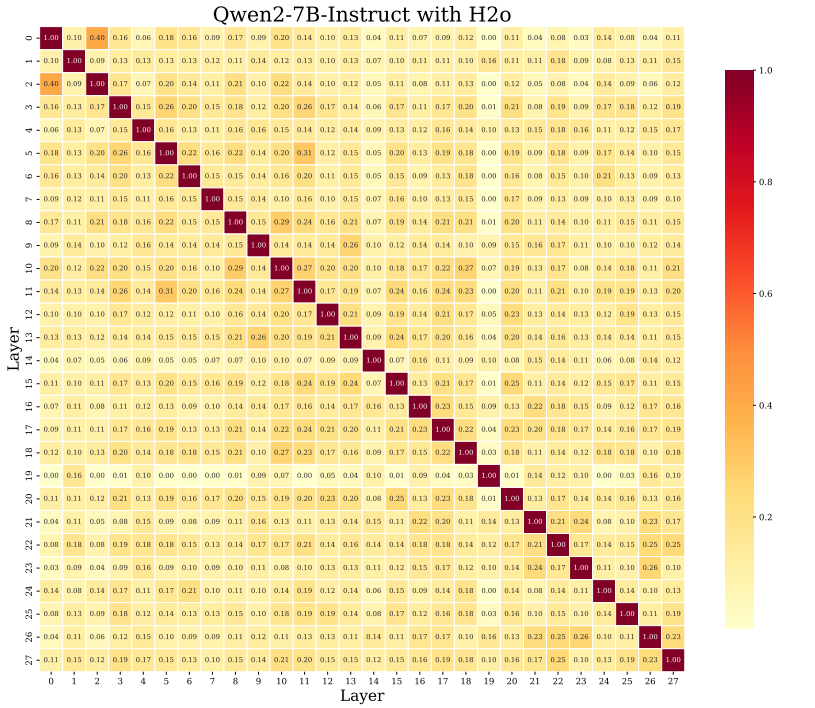

Figure 13: Layer-wise similarity heatmaps of the preserved KV cache indices by H2O on Qwen2-7B-Instruct

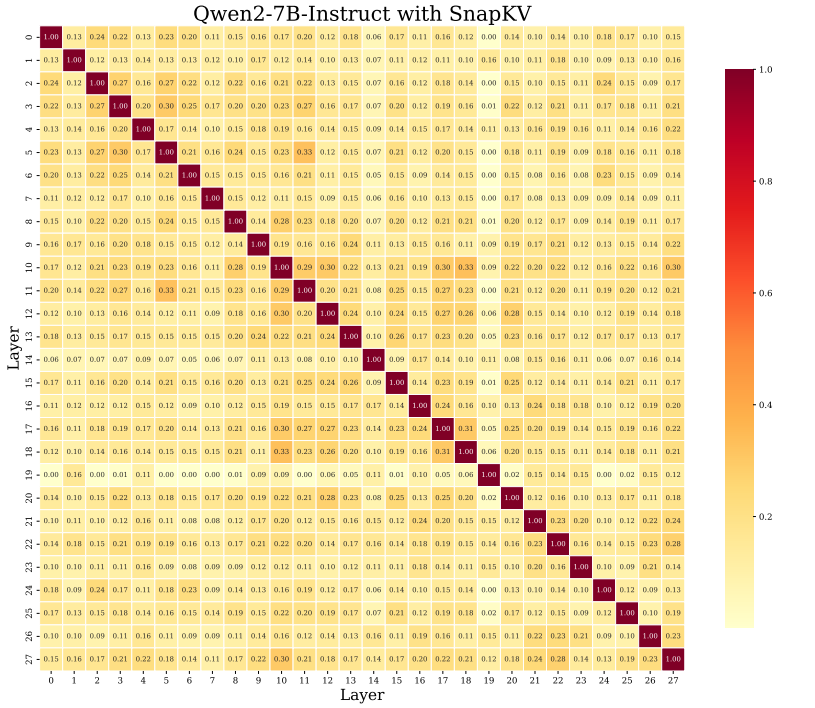

Figure 14: Layer-wise similarity heatmaps of the preserved KV cache indices by SnapKV on Qwen2-7B-Instruct

Instruct and Qwen2-7B-Instruct exhibit significant performance degradation, with LLaMA3-8B-Instruct experiencing a steeper decline after two layers of index reuse than Qwen2-7B-Instruct. This suggests that the Qwen2-7B-Instruct model may be more robust to index reuse.