the log file. For multi-head attention, only the indices of the first head are saved. PyramidKV, which has varying preserved index sizes across different layers, is not applicable for this experiment. Then we calculate the Jaccard similarity of the preserved indices of adjacent layers for different models. Table 12 shows the Jaccard similarity of the preserved indices of adjacent layers for different models.

Table 12: Retained KV Cache Indices Similarity of Adjacent Layers for Different Models.

Method	**H2O**	**SnapKV**	**ChunkKV**
LLaMA-3-8B-Instruct	25.31%	27.95%	**57.74%**
Qwen2-7B-Instruct	14.91%	16.50%	**44.26%**
Mistral-7B-Instruct	15.15%	15.78%	**52.16%**

Figures 7-9 (LLaMA-3-8B-Instruct), 10-12 (Mistral-7B-Instruct), and 13-15 (Qwen2-7B-Instruct) display the heatmaps of layer-wise indices similarity of the preserved KV cache indices by H2O, SnapKV and ChunkKV on different models. The pattern of the layer-wise indices similarity heatmap is consistent across different models, aligning with our findings in Section 3.3.