Table 1: Performance of different KV eviction policies with Llama-3.1-8B-Instruct on LongBench, where the **bold** results indicate the highest value in each row excluding FullKV column. The size of KV cache is set to 512.

Dataset	FullKV	StreamingLLM	SnapKV	IntentKV
2WikiMQA	50.2	44.2	47.9	**49.8**
SAMSum	43.2	41.4	**42.1**	42.0
QMSum	25.4	21.0	23.3	**24.0**
TriviaQA	92.3	88.8	91.2	**92.0**
HotpotQA	54.3	46.3	52.7	**54.7**
RepoBench-P	52.0	49.0	49.7	**50.3**
MultiNews	26.9	23.0	24.0	**24.1**
TREC	73.0	54.5	58.5	**65.0**
Qasper	47.2	24.8	34.4	**41.6**
PassageCount	6.6	**7.5**	7.3	**7.5**
LCC	62.0	59.1	60.9	**61.8**
MuSiQue	32.4	24.6	29.8	**31.6**
MultiFieldQA-en	55.3	33.8	49.7	**53.9**
NarrativeQA	31.1	26.5	**29.7**	29.5
PassageRetrieval-en	100.0	96.5	**99.5**	**99.5**
GovReport	33.9	22.0	23.9	**24.5**
Avg.	49.1	41.4	45.3	**47.0**

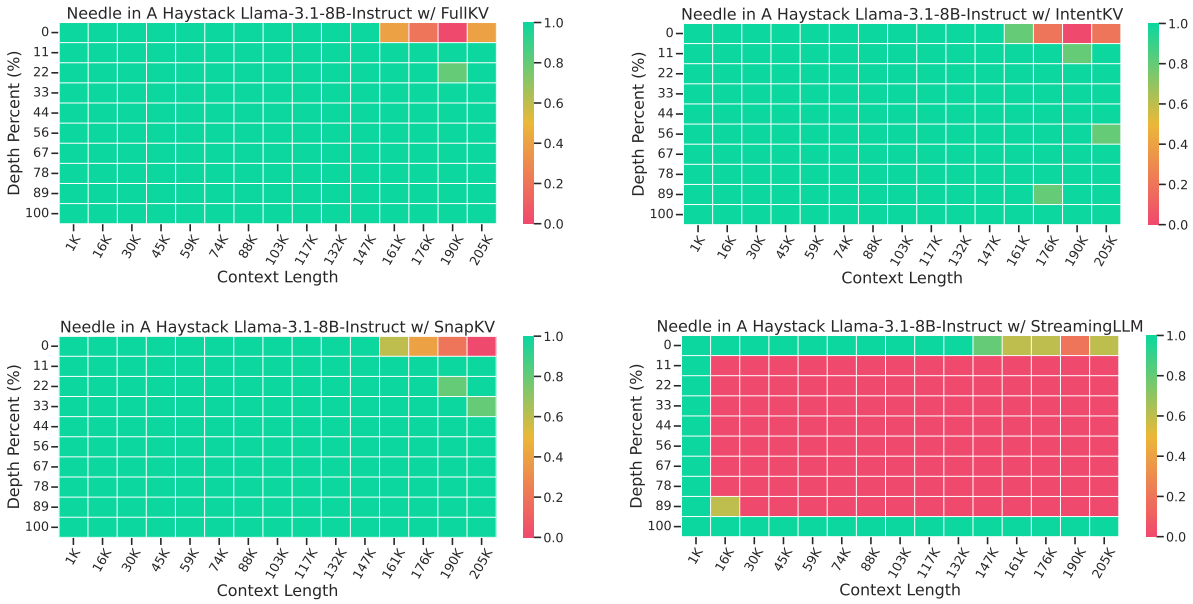

Figure 1: Performance of different KV eviction policies with Llama-3.1-8B-Instruct on NIAH test, where the size of KV cache is set to 2K.