B.3. Needle-In-A-Haystack

Figure 17 and 18 visualizes the performance of ChunkKV on the NIAH benchmark for LLaMA-3-8B-Instruct and Mistral-7B-Instruct with a KV cache size of 128 under 8k and 32k context length. The performance of ChunkKV is consistently better as the context length increases.