B.4. Chunk Size

Table 15 shows the performance of ChunkKV with different chunk size on the LongBench benchmark.

Table 15: LongBench Performance Comparison with different chunk sizes

Model	Chunk Size					Full KV
	3	5	10	20	30	
LLaMA-3-8B-Instruct	40.49	40.47	**40.51**	40.05	39.57	41.46
Mistral-7B-Instruct	46.45	46.51	**46.71**	46.42	45.98	48.08
Qwen2-7B-Instruct	40.38	40.33	40.66	**40.88**	40.73	40.71

Table 16 shows the performance of ChunkKV with different chunk size on the GSM8K benchmark. Figure 19 shows that the ChunkKV with different chunk sizes on GSM8K displays the same curve pattern as LongBench. The CoT prompt length for GSM8K is only 1K tokens, so the optimal chunk size range is smaller.

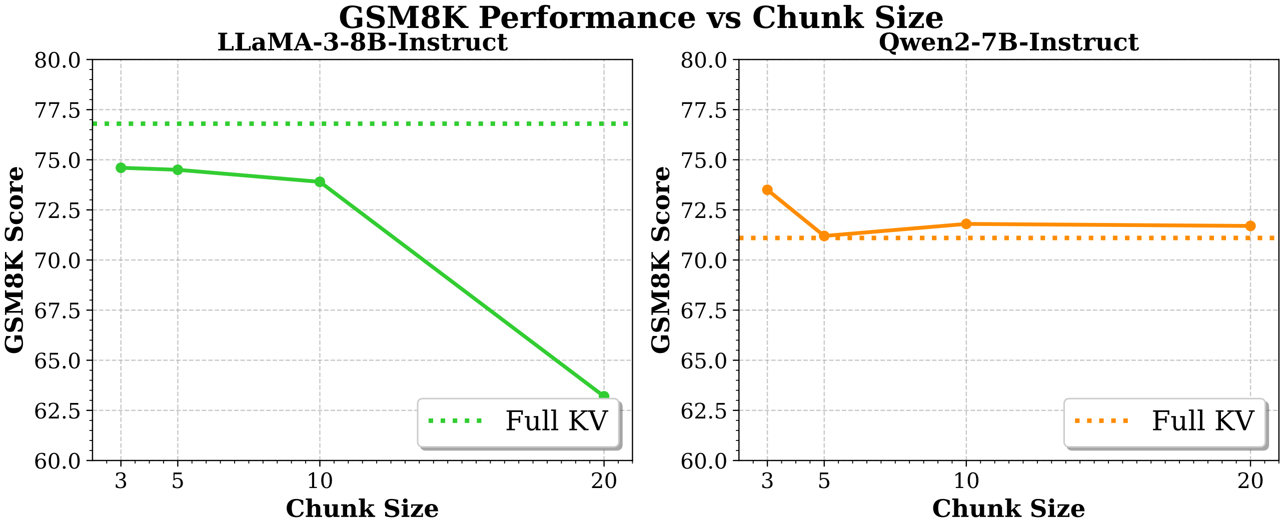

Figure 19: GSM8K Performance Comparison with different chunk size

Table 16: GSM8K Performance Comparison with different chunk sizes

Model	Chunk Size				Full KV
	3	5	10	20	
LLaMA-3-8B-Instruct	**74.6**	74.5	73.9	63.2	76.8
Qwen2-7B-Instruct	**73.5**	71.2	71.8	71.7	71.1

B.5. Multi-Lingual

Table 17 is the Chinese support model Qwen2-7B-Instruct evaluated on the LongBench Chinese subtask, where ChunkKV achieves better performance than other compression methods and the full KV cache performance. Both the English and Chinese results indicate that ChunkKV is a promising approach for maintaining crucial information in the KV cache.