Table 1: Comparison of SnapKV (compress to 1024 tokens) with various observation window and pooling kernel sizes on LongBench. Here w=32 k=7 is the configuration used in the paper, and w=32 k=1 refers to SnapKV without pooling, where we focus on tasks that do not involve information retrieval.

		Single-Doc. QA		Multi-Doc. QA		Summarization		Few-shot Learning			Synthetic		Code	
	LLMs *	Qasper	MF-en	HotpotQA	2WikiMQA	GovReport	MultiNews	TREC	TriviaQA	SAMSum	PCount	PRe	Lcc	RB-P
Mistral-7B	w=32 k=7	**29.51**	**49.25**	40.94	25.7	25.89	26.11	**69.5**	86.48	**42.06**	2.98	88.56	55.65	**51.87**
	w=16 k=7	27.14	48.9	41.02	**27.06**	28.2	26.13	67.0	86.84	40.9	4.51	**91.56**	60.55	50.25
	w=64 k=7	27.28	48.99	40.95	26.95	26.41	**26.18**	67.0	86.84	40.85	4.44	**91.56**	**60.79**	50.25
	w=32 k=5	26.79	48.7	40.07	26.74	29.65	24.55	64.29	86.73	40.21	**4.74**	90.49	57.06	48.57
	w=32 k=9	27.18	49.19	**41.39**	26.55	26.58	24.61	65.33	**86.87**	39.74	4.51	**91.56**	60.56	50.25
	w=32 k=1	-	-	-	-	**33.23**	26.04	67.33	86.84	40.9	4.51	**91.56**	60.66	50.25

Table 2: Performance comparison between original Llama-3-8B-Instruct-Gradient-1048k (1M) model with and without SnapKV (compress to 1024 tokens) on LongBench

	Single-Doc. QA	Multi-Doc. QA	Summarization	Few-shot Learning	Synthetic	Code
LLMs *	Qasper	HotpotQA	GovReport	TREC	PCount	Lcc
All KV	20.05	**38.11**	**34.71**	**67.0**	**5.0**	37.42
SnapKV: 1024	**20.17**	38.1	34.65	**67.0**	**5.0**	**37.48**

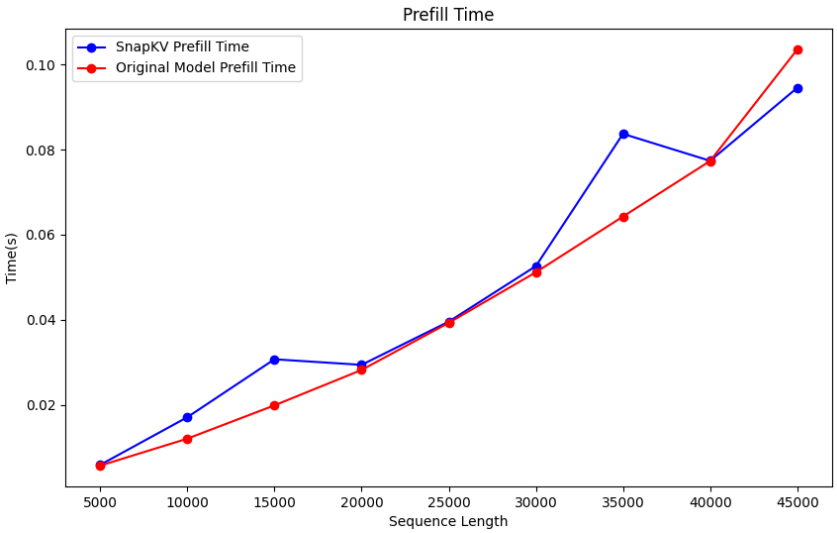

Figure 1: The prefilling time comparison between Mistral model with and without SnapKV on an H100.

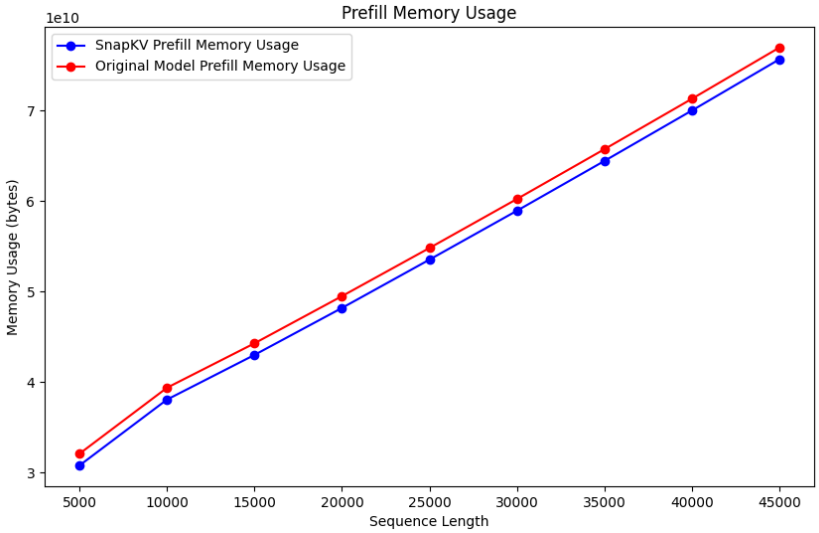

Figure 2: The maximum memory allocated comparison between Mistral model with and without SnapKV on an H100.

Table 3: Sequence Length Statistics for LongBench Benchmark

Dataset	Min	Max	Avg
Qasper	1443	14722	4620
HotpotQA	111	12480	6658
2WikimQA	881	23442	7141
GovReport	111	12480	6658
MultiNews	374	27973	6000
MF-en	505	10337	4559
TREC	746	13034	5475
TriviaQA	804	15960	6685
SAMSum	936	12403	6170
PCount	1407	14537	6117
PRe	2358	10607	6115
Lcc	386	14106	4283
RB-P	785	18864	6067