Table 13: Reusing Indexing Performance Comparison on GSM8K

Model	Number of Index Reuse Layers							
	1	2	3	5	8	10	20	28/32
LLaMA-3-8B-Instruct	74.5	74.6	65.9	44.1	15.3	2.20	1.60	1.80
Qwen2-7B-Instruct	71.2	71.2	73.0	69.4	67.4	71.1	54.0	49.4

B.2. LongBench

The Table 14 shows the average performance of KV cache compression methods in the LongBench English subtask categories. The ChunkKV achieves the best performance on the overall average, and the Multi-Document QA category, which supports that chunk method is more effective for semantic preservation.

Table 14: Comprehensive performance comparison of KV cache compression methods across LongBench English subtasks. Results are shown for various models and tasks, highlighting the effectiveness of different compression techniques.

Method	Single-Document QA			Multi-Document QA			Summarization			Few-shot Learning			Synthetic		Code		Avg. ↑
	NrtvQA	Qasper	MF-en	HotpotQA	2WikiMQA	Musique	GovReport	QMSum	MultiNews	TREC	TriviaQA	SAMSum	PCount	PRe	Lcc	RB-P	
Avg len	18,409	3,619	4,559	9,151	4,887	11,214	8,734	10,614	2,113	5,177	8,209	6,258	11,141	9,289	1,235	4,206	
LlaMa-3-8B-Instruct, KV Size = Full																	
FullKV	25.70	29.75	41.12	45.55	35.87	22.35	25.63	23.03	26.21	73.00	90.56	41.88	4.67	69.25	58.05	50.77	41.46
LlaMa-3-8B-Instruct, KV Size Compression Ratio = 10%																	
StreamingLLM	20.62	13.09	22.10	36.31	28.01	15.61	21.47	21.05	19.39	62.00	84.18	40.27	4.62	69.10	58.84	55.26	35.74
H2O	24.80	17.32	31.80	40.84	33.28	18.90	22.29	22.29	21.82	40.00	90.51	40.55	5.79	**69.50**	58.04	55.26	37.06
SnapKV	25.08	22.02	**37.95**	43.36	35.08	20.29	22.94	22.64	21.37	71.00	90.47	40.15	5.66	69.25	58.69	56.50	40.15
PyramidKV	**25.58**	20.77	35.85	**43.80**	33.03	**21.45**	**23.68**	22.26	**21.85**	71.50	90.47	**41.66**	5.84	69.25	58.52	55.91	40.08
ChunkKV	24.89	**22.96**	37.64	43.27	**36.45**	20.65	22.80	**22.97**	20.82	**71.50**	**90.52**	40.83	**5.93**	69.00	**60.49**	**57.48**	**40.51**
LlaMa-3-8B-Instruct, KV Size Compression Ratio = 20%																	
StreamingLLM	23.35	18.97	32.94	42.39	29.37	18.76	**25.78**	21.92	**25.16**	71.00	88.85	40.82	5.04	69.00	56.46	51.12	38.80
H2O	25.60	21.88	35.36	42.06	32.68	19.72	23.54	22.77	22.72	45.50	**90.57**	**41.67**	**5.51**	69.25	54.97	50.95	37.79
SnapKV	25.50	25.95	38.43	44.12	35.38	20.49	24.85	23.36	23.51	**72.50**	90.52	40.91	5.23	69.25	56.74	51.75	40.53
PyramidKV	25.36	26.88	37.99	44.21	**35.65**	**21.43**	25.52	**23.43**	23.47	72.00	90.56	41.45	5.26	**69.50**	56.55	50.93	40.63
ChunkKV	**26.13**	**28.43**	**38.59**	**44.46**	34.13	21.06	24.72	23.11	22.91	71.50	90.56	41.51	5.09	69.00	**58.17**	**52.51**	**40.74**
LlaMa-3-8B-Instruct, KV Size Compression Ratio = 30%																	
StreamingLLM	24.49	22.53	35.30	**44.33**	32.81	19.00	**27.12**	22.19	**25.93**	72.50	89.84	41.75	**5.41**	69.00	60.40	55.13	40.48
H2O	25.87	23.03	37.06	43.71	33.68	20.93	24.56	23.14	23.58	50.50	**90.77**	41.96	4.91	69.25	59.38	55.39	39.23
SnapKV	25.15	28.75	**39.28**	43.57	36.16	21.58	25.56	**23.19**	24.30	**73.00**	90.52	41.70	4.96	69.25	60.27	55.74	41.43
PyramidKV	25.42	27.91	38.81	44.15	**36.28**	**21.72**	26.50	23.10	24.28	72.00	90.56	41.87	4.67	**69.50**	60.09	55.19	41.37
ChunkKV	**25.88**	**29.58**	38.99	43.94	34.16	21.70	26.50	23.15	23.95	72.00	90.56	**42.47**	5.34	69.25	**61.68**	**56.35**	**41.59**
Mistral-7B-Instruct-v0.3, KV Size = Full																	
FullKV	29.07	41.58	52.88	49.37	39.01	28.58	34.93	25.68	27.74	76.00	88.59	47.59	6.00	98.50	61.41	62.39	48.08
Mistral-7B-Instruct-v0.3, KV Size Compression Ratio = 10%																	
StreamingLLM	25.15	25.47	30.08	44.39	32.49	19.40	24.11	20.85	19.55	65.00	88.21	44.83	4.50	79.50	59.48	58.82	40.11
H2O	29.35	33.39	50.39	49.58	36.76	27.42	25.16	24.75	22.12	42.00	89.00	**47.04**	5.50	98.50	57.58	59.24	43.61
SnapKV	28.54	**36.88**	53.42	50.15	38.17	27.99	26.67	**25.21**	**22.33**	72.00	89.36	45.44	**5.50**	**99.00**	59.79	**61.63**	46.38
PyramidKV	29.40	35.39	52.96	49.93	38.67	28.63	**27.59**	24.99	22.77	74.00	**90.02**	46.07	4.00	98.50	58.54	60.88	46.39
ChunkKV	**29.75**	36.82	**53.99**	**50.33**	**38.72**	**29.01**	27.03	24.76	21.42	**76.00**	88.73	46.49	5.00	98.00	**59.98**	61.47	**46.71**
Qwen2-7B-Instruct, KV Size = Full																	
FullKV	25.11	42.64	44.29	14.25	13.22	9.08	36.38	23.43	26.53	77.00	89.99	44.88	6.75	75.92	60.17	61.84	40.71
Qwen2-7B-Instruct, KV Size Compression Ratio = 10%																	
StreamingLLM	25.15	45.42	41.46	13.66	11.95	8.72	32.79	21.49	**26.24**	77.50	89.15	44.54	7.50	50.50	60.03	60.91	38.56
H2O	26.17	44.33	42.54	12.81	12.46	9.15	**33.24**	22.69	25.94	76.50	89.44	44.32	8.00	76.00	**61.28**	**62.39**	40.45
SnapKV	26.84	**45.96**	**45.79**	14.27	**13.35**	9.91	32.62	22.70	25.83	77.00	89.19	44.71	7.50	71.50	60.35	61.37	40.55
PyramidKV	**27.51**	44.45	43.59	13.35	13.13	9.12	32.28	22.60	25.45	77.00	**89.44**	44.53	7.00	73.50	60.91	61.24	40.31
ChunkKV	26.48	44.19	45.04	**15.94**	12.60	**10.52**	32.38	**22.87**	25.91	**77.50**	89.22	**44.78**	**8.50**	**76.50**	60.64	61.32	**40.88**