Table 9: Effect of retrieval depth on performance across conversation lengths (100K–10M) and memory abilities. Results are shown for different numbers of retrieved documents ($K \in \{5, 10, 15, 20\}$).

Length	Memory Ability	K=5	K=10	K=15	K=20
100K	Abstention	0.475	0.500	0.625	0.625
	Contradiction Resolution	0.037	0.025	0.025	0.031
	Event Ordering	0.216	0.191	0.218	0.210
	Information Extraction	0.502	0.450	0.412	0.391
	Instruction Following	0.312	0.362	0.475	0.462
	Knowledge Update	0.337	0.375	0.350	0.300
	Multi-Hop Reasoning	0.307	0.322	0.321	0.309
	Preference Following	0.550	0.591	0.562	0.575
	Summarization	0.231	0.231	0.218	0.213
	Temporal Reasoning	0.112	0.162	0.137	0.137
	Average	0.308	0.321	**0.334**	0.325
500K	Abstention	0.600	0.514	0.614	0.642
	Contradiction Resolution	0.014	0.021	0.071	0.071
	Event Ordering	0.246	0.229	0.238	0.247
	Information Extraction	0.508	0.531	0.503	0.507
	Instruction Following	0.375	0.341	0.390	0.373
	Knowledge Update	0.257	0.307	0.326	0.326
	Multi-Hop Reasoning	0.206	0.188	0.234	0.213
	Preference Following	0.557	0.597	0.628	0.607
	Summarization	0.323	0.354	0.375	0.376
	Temporal Reasoning	0.178	0.128	0.121	0.135
	Average	0.326	0.321	**0.350**	0.350
1M	Abstention	0.500	0.521	0.600	0.585
	Contradiction Resolution	0.021	0.021	0.057	0.053
	Event Ordering	0.200	0.224	0.240	0.242
	Information Extraction	0.366	0.398	0.377	0.391
	Instruction Following	0.419	0.476	0.439	0.446
	Knowledge Update	0.357	0.350	0.400	0.407
	Multi-Hop Reasoning	0.209	0.189	0.209	0.190
	Preference Following	0.551	0.596	0.535	0.514
	Summarization	0.316	0.317	0.325	0.351
	Temporal Reasoning	0.154	0.154	0.119	0.199
	Average	0.309	0.325	**0.330**	0.330
10M	Abstention	0.550	0.600	0.650	0.600
	Contradiction Resolution	0.012	0.012	0.025	0.025
	Event Ordering	0.197	0.210	0.213	0.236
	Information Extraction	0.350	0.150	0.300	0.300
	Instruction Following	0.350	0.150	0.450	0.400
	Knowledge Update	0.275	0.200	0.300	0.300
	Multi-Hop Reasoning	0.125	0.100	0.125	0.150
	Preference Following	0.308	0.175	0.275	0.275
	Summarization	0.220	0.089	0.196	0.164
	Temporal Reasoning	0.000	0.025	0.000	0.000
	Average	0.238	0.171	**0.253**	0.245

In a complementary experiment, we examined the impact of retriever choice. Our base architecture employs a dense retriever, which we compare against the sparse Splade-V2 retriever (Formal et al., 2022). As shown in Figure 5 in Appendix C.2, Splade yields performance gains of 2.01% at 100K tokens and 0.8% at 1M, but leads to slight degradations of 0.003% at 500K and 0.71% at 10M. On average, the sparse retriever provides a modest improvement across conversation lengths. The complete results comparing the dense retriever with the SPLADE retriever are provided in Table 10.