C Detailed Experiments

C.1 Ablation Study

In this section, we present the complete results of our ablation experiments. We evaluate the contribution of individual components in our proposed module as shown in table 8.

Table 8: Ablation study showing the impact of removing key memory components (retrieval, scratchpad, working memory, and noise filtering) on performance across various conversation lengths (100K–10M).

Length	Memory Ability	Base	w/o Retrieval from Index	w/o Scratchpad	w/o Working Memory	w/o Noise Filtering
100K	Abstention	0.475	0.725	0.600	0.575	0.700
	Contradiction Resolution	0.037	0.043	0.012	0.043	0.018
	Event Ordering	0.216	0.190	0.194	0.220	0.200
	Information Extraction	0.502	0.329	0.510	0.451	0.485
	Instruction Following	0.312	0.375	0.287	0.387	0.312
	Knowledge Update	0.337	0.237	0.350	0.362	0.312
	Multi-Hop Reasoning	0.307	0.201	0.248	0.303	0.181
	Preference Following	0.550	0.675	0.533	0.579	0.491
	Summarization	0.231	0.266	0.143	0.223	0.103
	Temporal Reasoning	0.112	0.075	0.125	0.125	0.087
	Average	0.308	0.311	0.300	**0.327**	0.289
500K	Abstention	0.600	0.571	0.585	0.657	0.585
	Contradiction Resolution	0.014	0.007	0.014	0.017	0.014
	Event Ordering	0.246	0.222	0.266	0.262	0.229
	Information Extraction	0.508	0.254	0.466	0.485	0.464
	Instruction Following	0.375	0.307	0.316	0.334	0.286
	Knowledge Update	0.257	0.192	0.285	0.235	0.314
	Multi-Hop Reasoning	0.206	0.104	0.227	0.192	0.247
	Preference Following	0.557	0.553	0.450	0.547	0.465
	Summarization	0.323	0.312	0.225	0.353	0.203
	Temporal Reasoning	0.178	0.042	0.116	0.114	0.130
	Average	**0.326**	0.256	0.295	0.320	0.294
1M	Abstention	0.500	0.664	0.600	0.557	0.507
	Contradiction Resolution	0.021	0.021	0.035	0.042	0.032
	Event Ordering	0.200	0.215	0.221	0.227	0.199
	Information Extraction	0.366	0.246	0.391	0.397	0.366
	Instruction Following	0.419	0.427	0.335	0.384	0.351
	Knowledge Update	0.357	0.185	0.321	0.400	0.285
	Multi-Hop Reasoning	0.209	0.129	0.227	0.221	0.169
	Preference Following	0.551	0.602	0.536	0.597	0.540
	Summarization	0.316	0.310	0.169	0.330	0.128
	Temporal Reasoning	0.154	0.050	0.111	0.121	0.111
	Average	0.309	0.285	0.295	**0.328**	0.269
10M	Abstention	0.550	0.800	0.650	0.650	0.600
	Contradiction Resolution	0.012	0.000	0.012	0.000	0.000
	Event Ordering	0.197	0.199	0.199	0.209	0.181
	Information Extraction	0.350	0.000	0.200	0.150	0.200
	Instruction Following	0.350	0.175	0.175	0.175	0.050
	Knowledge Update	0.275	0.050	0.300	0.150	0.225
	Multi-Hop Reasoning	0.125	0.000	0.125	0.125	0.075
	Preference Following	0.308	0.191	0.241	0.200	0.175
	Summarization	0.220	0.119	0.068	0.0083	0.050
	Temporal Reasoning	0.000	0.000	0.050	0.075	0.000
	Average	**0.238**	0.153	0.202	0.181	0.155

C.2 Retrieval Budget

We investigate the impact of the retrieval budget through two sets of experiments: (i) varying the retrieval depth by setting the number of retrieved documents $K \in \{5, 10, 15, 20\}$, and (ii) comparing a dense retriever against a sparse retriever (SPLADE).

The full results examining the effect of different retrieval depths (number of retrieved documents) are presented in Table 9.