

Table 1: Comparison of different LLMs and methods across conversation lengths and memory abilities using the created benchmark. Methods with the best performance per evaluation are bolded.

Length	Memory Ability	Qwen 2.5			Llama Maverick			Gemini 2 Flash			GPT-4.1-nano		
		Vanilla	RAG	Ours	Vanilla	RAG	Ours	Vanilla	RAG	Ours	Vanilla	RAG	Ours
100K	Abstention	0.300	0.650	0.475	0.200	0.800	0.600	0.800	0.800	0.675	0.475	0.800	0.575
	Contradiction Resolution	0.031	0.025	0.037	0.025	0.031	0.031	0.006	0.050	0.018	0.012	0.018	0.031
	Event Ordering	0.192	0.201	0.205	0.190	0.162	0.166	0.181	0.191	0.166	0.181	0.169	0.177
	Information Extraction	0.425	0.338	0.479	0.510	0.392	0.518	0.333	0.341	0.464	0.273	0.362	0.538
	Instruction Following	0.400	0.375	0.362	0.412	0.375	0.412	0.275	0.287	0.362	0.425	0.350	0.400
	Knowledge Update	0.437	0.275	0.362	0.300	0.350	0.450	0.125	0.325	0.300	0.275	0.375	0.375
	Multi-Hop Reasoning	0.222	0.203	0.281	0.152	0.225	0.353	0.200	0.148	0.225	0.178	0.263	0.365
	Preference Following	0.554	0.379	0.566	0.450	0.512	0.625	0.300	0.416	0.462	0.437	0.550	0.625
	Summarization	0.128	0.074	0.232	0.065	0.111	0.238	0.018	0.093	0.139	0.028	0.083	0.202
	Temporal Reasoning	0.112	0.112	0.100	0.275	0.187	0.187	0.150	0.125	0.112	0.125	0.162	
500K	Average	0.280	0.269	0.311	0.240	0.323	0.358	0.242	0.280	0.294	0.239	0.309	0.345
	Abstention	0.314	0.728	0.571	0.185	0.785	0.628	0.714	0.800	0.685	0.557	0.828	0.600
	Contradiction Resolution	0.053	0.017	0.017	0.035	0.028	0.042	0.010	0.021	0.021	0.017	0.025	0.035
	Event Ordering	0.185	0.221	0.244	0.209	0.186	0.197	0.215	0.189	0.200	0.188	0.180	0.204
	Information Extraction	0.166	0.400	0.506	0.608	0.402	0.535	0.469	0.343	0.478	0.142	0.382	0.491
	Instruction Following	0.304	0.350	0.295	0.403	0.447	0.390	0.133	0.334	0.280	0.244	0.286	0.342
	Knowledge Update	0.111	0.226	0.278	0.276	0.338	0.264	0.171	0.180	0.223	0.107	0.288	0.240
	Multi-Hop Reasoning	0.125	0.187	0.214	0.219	0.313	0.350	0.198	0.135	0.157	0.070	0.233	0.266
	Preference Following	0.567	0.477	0.571	0.560	0.525	0.623	0.379	0.427	0.532	0.450	0.577	0.684
	Summarization	0.137	0.187	0.344	0.266	0.197	0.373	0.136	0.165	0.250	0.109	0.184	0.334
1M	Temporal Reasoning	0.035	0.114	0.121	0.064	0.078	0.190	0.150	0.078	0.092	0.057	0.161	0.154
	Average	0.200	0.291	0.316	0.283	0.330	0.359	0.257	0.267	0.292	0.194	0.314	0.335
	Abstention	0.342	0.650	0.500	0.221	0.742	0.435	0.642	0.750	0.735	0.492	0.778	0.678
	Contradiction Resolution	0.035	0.035	0.021	0.046	0.028	0.042	0.010	0.028	0.007	0.050	0.028	0.021
	Event Ordering	0.183	0.195	0.200	0.214	0.179	0.193	0.190	0.198	0.185	0.191	0.179	0.211
	Information Extraction	0.138	0.407	0.366	0.489	0.431	0.474	0.374	0.380	0.341	0.153	0.399	0.410
	Instruction Following	0.383	0.300	0.419	0.440	0.338	0.433	0.120	0.290	0.380	0.226	0.271	0.394
	Knowledge Update	0.064	0.378	0.357	0.164	0.342	0.414	0.107	0.278	0.264	0.150	0.342	0.392
	Multi-Hop Reasoning	0.102	0.163	0.209	0.174	0.245	0.270	0.083	0.134	0.147	0.091	0.293	0.278
	Preference Following	0.486	0.491	0.551	0.535	0.514	0.610	0.273	0.470	0.472	0.435	0.513	0.576
10M	Summarization	0.122	0.157	0.316	0.207	0.145	0.315	0.091	0.125	0.224	0.060	0.152	0.290
	Temporal Reasoning	0.073	0.078	0.154	0.097	0.107	0.176	0.104	0.057	0.085	0.061	0.064	0.107
	Average	0.193	0.285	0.309	0.259	0.307	0.336	0.199	0.271	0.284	0.191	0.302	0.336
	Abstention	0.250	0.600	0.550	0.050	0.700	0.450	0.750	0.650	0.650	0.450	0.650	0.400
	Contradiction Resolution	0.050	0.000	0.012	0.025	0.000	0.000	0.025	0.000	0.000	0.012	0.025	
	Event Ordering	0.180	0.221	0.197	0.190	0.220	0.176	0.220	0.266	0.193	0.215	0.201	0.173
	Information Extraction	0.100	0.350	0.350	0.075	0.375	0.300	0.075	0.275	0.150	0.050	0.300	0.350
	Instruction Following	0.175	0.200	0.350	0.250	0.350	0.500	0.025	0.125	0.250	0.075	0.175	0.250
	Knowledge Update	0.100	0.300	0.275	0.100	0.375	0.325	0.050	0.325	0.200	0.050	0.325	0.300
	Multi-Hop Reasoning	0.125	0.050	0.125	0.000	0.075	0.125	0.000	0.125	0.125	0.012	0.091	0.135
10M	Preference Following	0.241	0.291	0.308	0.291	0.316	0.483	0.075	0.300	0.150	0.175	0.366	0.425
	Summarization	0.114	0.106	0.220	0.065	0.053	0.277	0.000	0.045	0.136	0.020	0.063	0.179
	Temporal Reasoning	0.000	0.000	0.000	0.000	0.025	0.025	0.025	0.025	0.075	0.050	0.000	0.025
	Average	0.133	0.211	0.238	0.104	0.249	0.266	0.122	0.216	0.192	0.109	0.218	0.226

for other LLMs we adopt their default maximum output length. For experiments involving both the RAG baseline and our proposed method, we employ FAISS as the vector database (Douze et al., 2024). For dense retrieval, we use the embedding model *BAAI/bge-small-en-v1.5* (Xiao et al., 2023).

4.2 EMPIRICAL FINDINGS

Main Results: Across all four conversation lengths (100K–10M tokens), our method consistently outperforms both long-context LLMs and RAG baselines (Table 1). At shorter contexts (100K), we observe strong gains, such as +49.1% for Llama-4-Maverick and +44.3% for GPT-4.1-nano over long-context baselines, showing that structured memory helps even when full history can be processed. The benefits grow with context length: at 1M tokens, improvements reach +75.9% for GPT-4.1-nano and +60.1% for Qwen2.5-32B. At 10M tokens—where no baseline natively supports the full context—our method achieves dramatic improvements, including +155.7% for Llama-4-Maverick and +107.3% for GPT-4.1-nano. The only exception is Gemini-2.0-flash at 10M, where our method surpasses the long-context baseline (+57.3%) but slightly trails RAG, likely due to model-specific retrieval behavior. Overall, these findings underscore the scalability and robustness of our framework across diverse architectures and extreme context lengths.

When evaluated across the ten memory abilities, our method shows the largest relative gains in summarization (+160.6%), multi-hop reasoning (+27.2%), and preference following (+76.5%). Strong improvements are also observed in information extraction (+56.7%), instruction following (+39.5%), and temporal reasoning (+56.3%). These results highlight that our method is particularly effective for tasks requiring long-range recall and integration of dispersed information. In contrast, all methods—including ours—perform strongest in abstention and weakest in contradiction resolution, indicating that contradiction detection remains a challenging open problem.