



Figure 5: Performance comparison between dense retrieval and sparse retrieval (SPLADE) in LIGHT.

Table 10: Comparison of dense and sparse retrieval strategies across conversation lengths (100K–10M) and ten memory abilities. The table reports performance when using the default dense retriever versus a sparse retriever (SPLADE).

Length	Memory Ability	Base (Dense retriever)	Sparse retriever (SPLADE)
100K	Abstention	0.475	0.525
	Contradiction Resolution	0.037	0.43
	Event Ordering	0.216	0.181
	Information Extraction	0.502	0.596
	Instruction Following	0.312	0.400
	Knowledge Update	0.337	0.350
	Multi-Hop Reasoning	0.307	0.267
	Preference Following	0.550	0.562
	Summarization	0.231	0.230
	Temporal Reasoning	0.112	0.125
	Average	0.308	0.328
500K	Abstention	0.600	0.557
	Contradiction Resolution	0.014	0.025
	Event Ordering	0.246	0.226
	Information Extraction	0.508	0.559
	Instruction Following	0.375	0.345
	Knowledge Update	0.257	0.307
	Multi-Hop Reasoning	0.206	0.212
	Preference Following	0.557	0.565
	Summarization	0.323	0.330
	Temporal Reasoning	0.178	0.107
	Average	0.326	0.323
1M	Abstention	0.500	0.564
	Contradiction Resolution	0.021	0.028
	Event Ordering	0.200	0.196
	Information Extraction	0.366	0.392
	Instruction Following	0.419	0.401
	Knowledge Update	0.357	0.371
	Multi-Hop Reasoning	0.209	0.193
	Preference Following	0.551	0.595
	Summarization	0.316	0.300
	Temporal Reasoning	0.154	0.133
	Average	0.309	0.317
10M	Abstention	0.550	0.700
	Contradiction Resolution	0.012	0.000
	Event Ordering	0.197	0.202
	Information Extraction	0.350	0.350
	Instruction Following	0.350	0.250
	Knowledge Update	0.275	0.375
	Multi-Hop Reasoning	0.125	0.125
	Preference Following	0.308	0.200
	Summarization	0.220	0.090
	Temporal Reasoning	0.000	0.025
	Average	0.238	0.231