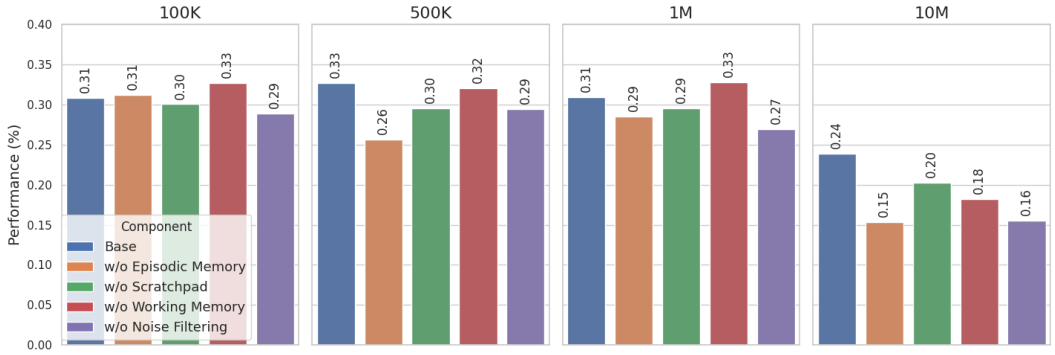


Figure 3: Ablation study of the effect of different components in LIGHT.

**Ablation:** We conduct an ablation to assess the role of each component—*episodic memory*, *scratchpad*, *working memory*, and *noise filtering*—across conversation lengths (Figure 3). At 100K, retrieval slightly hurts performance (+0.28% when removed), since the scratchpad alone suffices and extra retrieval introduces noise, while removing scratchpad or noise filtering reduces performance (–0.08%, –1.89%). Working memory also degrades results here (–1.89%), consistent with the low proportion of probing questions targeting recent turns (Table 7). At 500K, removing any component reduces performance, confirming their utility at this scale. At 1M, retrieval, scratchpad, and noise filtering remain beneficial, but removing working memory slightly improves performance, again reflecting its limited usefulness when few questions depend on the most recent turns. By 10M, all components are essential, with removals leading to large drops (–8.5% for retrieval, –3.7% for scratchpad, –5.7% for working memory, –8.3% for noise filtering). Overall, the ablations show that each module contributes increasingly as context length grows, and the full architecture consistently achieves the best performance. Detailed results across all memory abilities are provided in Table 8.

**Effect of Retrieval Budget:** We examine the effect of retrieval budget ($K$), testing 5, 10, 15, and 20 documents (Figure 4). Performance consistently improves when increasing $K$ from 5 to 15, with the best results at $K$=15 (+8.5%, +7.3%, +6.6%, and +6.1% at 100K, 500K, 1M, and 10M). Increasing further to $K$=20 slightly degrades performance, likely due to noisy context. Results at $K$=10 are mixed—helpful at 100K and 1M but harmful at 500K and 10M—indicating additional documents sometimes add noisy information. Full results across memory abilities are shown in Table 9. We also conducted complementary experiments analyzing the effect of retriever choice, where we did not observe a considerable difference between sparse and dense retrieval. The full results and discussion are provided in Appendix C.2.

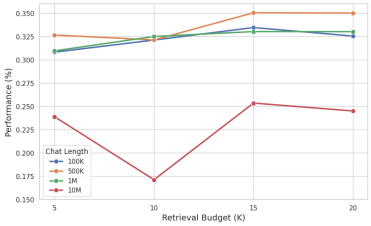


Figure 4: Effect of varying retrieval budget ($K$) on the performance.

**Case Study** A case study demonstrating the usefulness of the scratchpad is provided in Appendix F.

**Human Evaluation:** We conducted a human evaluation to assess the quality of the generated conversations. Three dimensions were considered: *Coherence and Flow*, *Realism*, and *Complexity and Depth*, each rated on a 5-point Likert scale (1 = lowest, 5 = highest). The average scores across all conversations were 4.53, 4.57, and 4.64, respectively, indicating consistently high quality. The evaluation rubric and detailed scores are provided in Appendix B.2.

## 5 Related Work

The detailed related work is provided in Appendix A; here we present a concise summary.

Context windows of LLMs have expanded dramatically, from early limits of 512–2K tokens (GPT-2/3; (Radford et al., 2019; Brown et al., 2020)) to 128K–1M (Claude-3, GPT-4-Turbo, Gemini 2.0;