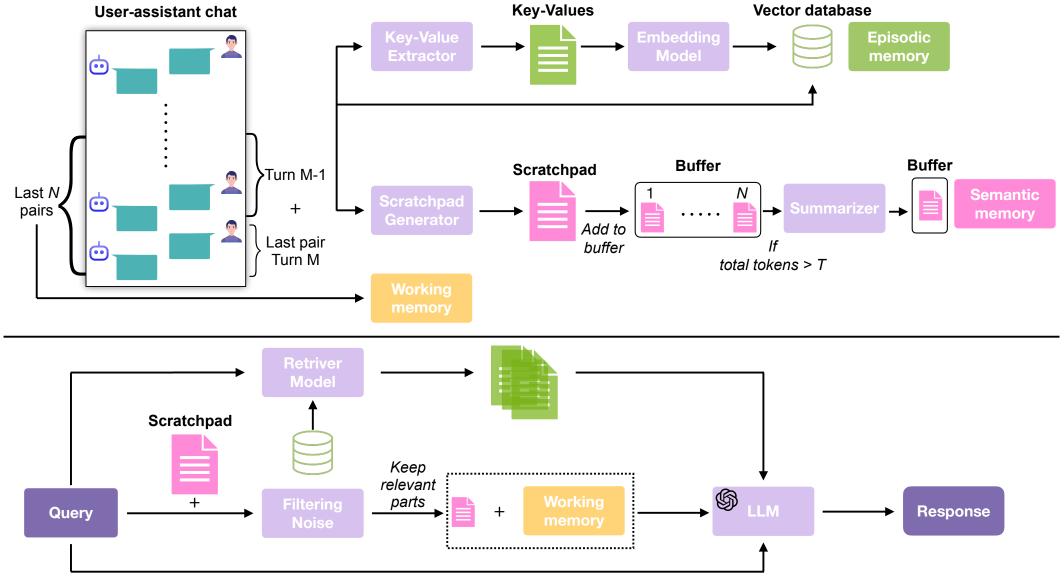

Figure 2: Overview of the LIGHT framework.

Appendix G), which assigns 0 (unsatisfied), 0.5 (partially satisfied), or 1 (fully satisfied). Scores are averaged across nuggets to produce ability-level metrics. This nugget-based procedure applies to nine memory abilities; the exception is event ordering, where quality depends on both recall and correct sequence. We evaluate event ordering using the Kendall tau-b coefficient (Kendall, 1945), which considers both order and presence. To apply this metric, an LLM equivalence detector (using the prompt in Listing 21 in Appendix G) aligns events in system responses with nuggets, outputting `yes` if two snippets denote the same event/topic and `no` otherwise. Kendall tau-b is then computed over the aligned sequences, capturing both recall and ordering fidelity. Examples of nugget construction for each memory ability are provided in Appendix D.

3 LIGHT: Improving Memory Capabilities of LLMs

Inspired by research in human cognitive science (Sridhar et al., 2023; Binder & Desai, 2011), humans employ two primary mechanisms for remembering and using knowledge: *episodic memory*, the ability to recall specific personal experiences along with their context, and *working memory*, the capacity to retain and manipulate information about recent events over short periods. In addition, maintaining notes on a *scratchpad* provides an external record that supports long-term recall and later retrieval. Since answering questions in long-context conversations similarly requires integrating past experiences and accumulated knowledge, we introduce a method that emulates these strategies by combining episodic recall, short-term working memory, and an external scratch-pad mechanism.

Overview: An overview of our method is shown in Figure 2. Given a question x about a conversation $\mathcal{T} = \{t_i\}_{i=1}^{|\mathcal{T}|}$, where $|\mathcal{T}|$ is the total number of turns, the framework first queries a retrieval model R to obtain k relevant segments from $\mathcal{T}$, simulating recall from episodic memory: $E = R(x, k, \mathcal{T})$. Next, the most recent z dialogue pairs of the conversation are selected to form the working memory, $W = \{t_{|\mathcal{T}|-i}\}_{i=0}^{z}$. In parallel, a pre-constructed scratchpad $S_{|\mathcal{T}|}$ contains up to m salient notes. A filtering function f retains only the items pertinent to x, yielding $S_x = f(S_{|\mathcal{T}|}, x)$. Finally, the LLM π generates the answer by conditioning on the question and these three memory components, $y = \pi(x, E, W, S_x)$ using the prompt shown in Listing 44 in Appendix G. The remainder of this section details the construction and logic of each component in this pipeline.