For 10M-token conversations, a single plan cannot adequately capture the scope and continuity required at this scale. To address this, we construct ten distinct yet interlocking conversation plans that together produce a coherent long-term narrative. While the process begins with a main seed that defines the global topic and theme of the conversation, a single seed is insufficient for producing ten plans. Instead, we generate ten distinct conversation seeds—one for each plan—so that the narrative can unfold across multiple stages. The procedure for deriving these seeds—and the plans that follow—differs depending on the strategy. We propose two strategies for constructing them:

- **Sequential Expansion:** The conversation seed is used as the first seed in the sequence. The remaining seeds are generated to represent successive stages of the user's life, extending the storyline chronologically. For instance, if the main seed concerns an international trip, the first plan covers the trip itself, the second covers the period after returning (e.g., job search), and subsequent seeds correspond to later milestones. We generate these seeds using the prompt shown in Listing 28, which conditions on the main seed, user profile, and timeline to produce a sequence of temporally aligned seeds. Each conversation plan is then generated sequentially, with every plan conditioned on its predecessor to maintain continuity, as specified in line 12 of Algorithm 1 in Appendix B.3.5. The plans are generated using the prompt shown in Listing 30, yielding a temporally ordered series of interconnected narrative arcs. To maintain realism, the user's core relationships (e.g., parents, children, partner) remain fixed across plans, while new acquaintances are gradually introduced.

- **Hierarchical Decomposition:** Instead of extending the seed chronologically, the main seed is decomposed into ten sub-seeds, each corresponding to a distinct topical or temporal slice of the overall storyline. Together, these seeds span the full narrative. For example, if the main seed concerns an international trip, the first three seeds may cover preparation steps (e.g., reservations, document gathering), the next five capture events during the trip, and the final two represent post-trip activities (e.g., reflections, recounting experiences). Like in Sequential Expansion, the user's core relationships (e.g., parents, children, partner) remain fixed across plans, while new acquaintances are gradually introduced. We generate these ten sub-seeds using the prompt shown in Listing 29, which takes the main seed, user profile, and timeline, and outputs ten derived seeds.

Each plan is assigned explicit topical and temporal boundaries to prevent redundancy or thematic overlap, ensuring that sub-themes unfold in the correct stage of the narrative. These boundaries are encoded in the conversation seed itself. For coherence, summaries of all prior plans are provided to the LLM when generating a new plan, allowing contextual references to past events. Moreover, when generating each plan, future seeds are also supplied, encoding their own topical and temporal boundaries. This design allows earlier plans to anticipate upcoming events with consistent references (e.g., booking tickets for the correct travel dates before the trip actually occurs). This strategy is implemented in line 20 of Algorithm 1 in Appendix B.3.5. Conversation plans are generated using the prompt shown in Listing 31, which takes as input the main seed, the current sub-seed, the number of sub-plans, the narrative set, the user profile, core and newly introduced relationships, the preceding and subsequent sub-seeds, the previous plan, the summary of all previous plans, the index of the current sub-seed, and a binary indicator specifying whether the plan is the first in the sequence (in which case the introduction of the user is included). The output is a fully specified conversation plan.

After the conversation plan is constructed, it is expanded into user-turn questions and subsequently assistant responses, yielding complete dialogues that can be used to evaluate memory abilities. However, in its initial form, the plan may not include sufficient information to evaluate three critical memory abilities: *contradiction resolution*, *knowledge update*, and *instruction following*. To address this, after the initial plan generation, we pass the plan to GPT-4.1 to generate high-quality plans and augment each sub-plan with additional bullet points specifically designed to enable evaluation of these abilities. Importantly, this augmentation is performed in a second stage rather than during the initial plan generation, since incorporating such information directly in a single-pass generation leads to lower quality and less reliable coverage of these abilities. The augmentation is implemented using the prompt shown in Listing 27, which takes an existing conversation plan as input and outputs a revised version where each sub-plan includes three additional bullet points targeting these abilities.