Given the conversation seed, this model produces narrative elements capturing the evolving storyline, forming the backbone of a coherent conversation.

Conversation plans consist of N *sub-plans*, each representing a distinct stage in the conversation. Each sub-plan contains M *bullet points*, defined by a *narrative*, a descriptive statement of its role in the storyline, and a *time anchor*. For conversations of 128K, 500K, and 1M tokens, a single plan is generated (line 4 in Algorithm 1, Appendix B.3.5) by conditioning the LLM on the conversation seed, profile, relationship graph, timeline, and specified counts of sub-plans, bullet points, and narratives (prompt in Listing 24, Appendix G). The number of sub-plans varies with domain and target length to meet the token requirement; e.g., coding domains generally require fewer turns than broader domains. For 10M-token conversations, one plan cannot capture the scope, so we create ten interlocking plans forming a coherent longer narrative. The process begins with a global seed defining the overall topic and theme, but a single seed is insufficient; instead, we derive ten distinct seeds—one per plan—so the narrative can evolve across stages. We propose two strategies:

- **Sequential Expansion:** The global seed defines the initial point in the conversation's chronology. Subsequent seeds represent successive events (e.g., a trip, job search, later milestones). Using the prompt in Listing 28 (Appendix G), each new seed is generated from the main seed, profile, and timeline. Plans are then produced sequentially (line 12 in Algorithm 1, Appendix B.3.5), with each plan conditioned on its predecessor to maintain continuity. Core relationships (e.g., parents) remain fixed, while new acquaintances are gradually introduced to reflect the evolving context.
- **Hierarchical Decomposition:** The main seed is decomposed into ten sub-seeds, each representing a distinct topical and temporal segment. Together, these sub-seeds span the full storyline (e.g., an international trip: first three for preparation, next five for trip events, final two for reflections). Similar to sequential expansion, the user's core relationships remain constant, while new acquaintances are introduced to reflect the evolving context. These ten sub-seeds are generated using the prompt in Listing 29 (Appendix G), conditioned on the main seed, profile, and timeline.

Each conversation plan is assigned explicit topical and temporal boundaries—encoded in the seed—to avoid redundancy and ensure sub-themes appear in the right narrative stage. For coherence, the LLM conditions on summaries of prior plans and future seeds when producing a new plan, allowing anticipation of upcoming events (e.g., reserving tickets for travel dates). This procedure is implemented in line 20 of Algorithm 1 (Appendix B.3.5). Plans are generated using the prompt in Listing 31 (Appendix G), conditioned on the main seed, current sub-seed, number of sub-plans, narrative set, user profile, core and new relationships, preceding and subsequent sub-seeds, previous plan, a summary of earlier plans, current sub-seed index, and a binary flag for the first plan (triggering user introduction). Since initial plans may not sufficiently test three key memory abilities—*contradiction resolution*, *information update*, and *instruction following*—we apply a two-stage augmentation: first generate the base plan, then use GPT-4.1 (Listing 27) to augment each sub-plan with three targeted bullet points. Performing augmentation separately improves coverage and fidelity. The refinement follows the prompt in Listing 27 (Appendix G), which takes plan as input and outputs the revised version. The detailed process for plan generation is reported in Appendix B.3.2.

2.2.2 User Utterance Generation

Once conversation plans are constructed, user utterances are synthesized from the sub-plans. Each sub-plan contains M bullet points, which are divided sequentially into K contiguous batches of equal size. Batching narrows the LLM's focus, reducing repetition and low-quality outputs that can occur when conditioning on the entire sub-plan. For each batch, the LLM generates I user questions (line 6 in Algorithm 2 in Appendix B.3.5) using the prompt in Listing 32 (Appendix G), conditioned on the conversation seed, the current batch, preceding batches, and context from earlier sub-plans. Each generated user question constitutes a user turn in the dialogue, ensuring coherence and continuity across extended conversations. Values of K and I are manually specified based on domain and target conversation length to meet the token budget, with configurations reported in Table 6 (Appendix B). This provides fine-grained control over user interaction density, preventing undergeneration or redundancy. To balance quality and cost, question generation uses the open-source LLaMA-3.3 70B model (AI, 2024), which produces high-quality outputs efficiently as the backbone LLM. The details of this procedure for user utterance generation are provided in Appendix B.3.3.