Domain	Chat Titles
Asking Recommendation	Finding the Perfect Smartphone for Photography and Gaming • Choosing a Lightweight Laptop for Work, Travel, and Entertainment • Selecting a Must-Read Fiction Series for Winter Evenings • Finding the Best Streaming Movies for a Family Weekend • Choosing Comfortable and Stylish Sneakers for Daily Wear
Legal & Administrative	Filing for a Marriage-Based Green Card in the United States • Creating a Legally Valid Will and Estate Plan • Applying for a Patent to Protect a New Invention
Philosophical & Ethical Discussion	Deciding Whether to Use AI to Automate Hiring in My Company • Considering Whether to Believe in and Live by the Idea of Free Will

B.3.2 CONVERSATION PLAN GENERATION

A *conversation plan* serves as the central scaffold of each conversation, providing a coherent storyline that evolves chronologically. The process of constructing conversation plans is anchored by a *seed* that specifies the *domain* of the dialogue (e.g., sports, finance, programming, mathematics), a *title* representing the high-level topic, and a *theme* that provides a more detailed instantiation of the title. The seed also includes a set of *subtopics*, which enumerate finer-grained subtopics and details to ensure topical diversity. However, a title, theme, and subtopics alone are insufficient to support detailed and information-rich conversations. To enrich the narrative, we introduce *narratives set* that define the evolving aspects of a conversation (e.g., career progression, goals, relationships). Each narrative is paired with descriptive details that specify its scope and trajectory.

In addition to the seed and narrative set, each conversation incorporates a *user profile*, a *relationship graph*, and an explicit *timeline*. The user profile includes attributes such as name, age, gender, location, profession, and personality traits. To avoid redundancy, personality traits are grounded in the Myers–Briggs Type Indicator (MBTI). Specifically, we randomly select six MBTI types, provide their descriptions, and instruct an LLM to synthesize a composite trait profile, enabling the creation of 8,008 unique user profiles. Relationship graphs are then constructed, linking the main user to family members (parents, partner, children), friends, and acquaintances, subject to constraints (e.g., plausible age gaps) to preserve realism. The timeline specifies the temporal span of the conversation, defining the range between its beginning and end.

In order to generate titles and themes of the chats, target domains are first specified by human. Given these domains, GPT-4.1 (OpenAI, 2025a) is prompted using the prompt shown in Listing 22 in Appendix G, to produce candidate titles, themes, and subtopics. These candidates are refined by human to ensure topical diversity by removing the similar chat titles and selecting diverse chat titles. Finally, for each conversation, we generate 15–20 narratives using open-source LLaMA-3.3 70B (AI, 2024) with the prompt shown in Listing 23 to save cost. In this prompt, given the conversation seed as input, the LLM produces narratives that capture evolving aspects of the storyline, providing the backbone for constructing coherent conversation plans.

Conversation plans are structured as a sequence of N *sub-plans*, where each sub-plan corresponds to a distinct stage of the conversation. Each sub-plan contains a fixed number of M *bullet-points*, and each bullet-point is defined by a *narrative* and a descriptive statement specifying how that narrative unfolds in the storyline. To maintain temporal coherence, each sub-plan also includes a *time anchor* specifying a concrete date or period.

For conversations of sizes 128K, 500K, and 1M tokens, a single conversation plan is generated, as shown in line 4 of Algorithm 1 in Appendix B.3.5. The plan is produced by conditioning the LLM on the conversation seed, user profile, relationship graph, timeline, the number of sub-plans, the number of bullet points within each sub-plan and narrative set, using the prompt shown in Listing 24 in Appendix G. The number of sub-plans is not fixed but varies with both the domain and the target conversation length, in order to adhere to the length budget. For instance, domains such as coding typically require fewer dialogue turns to reach the same token budget compared to more general domains.