

LifeBench: A Benchmark for Long-Horizon Multi-Source Memory

Zihao Cheng^{1,2}, Weixin Wang^{1,2}, Yu Zhao³, Ziyang Ren³, Jiaxuan Chen^{1,2},
Ruiyang Xu³, Shuai Huang³, Yang Chen³, Guowei Li³, Mengshi Wang³, Yi Xie³, Ren Zhu³,
Zeren Jiang³, Keda Lu³, Yihong Li³, Xiaoliang Wang¹, Liwei Liu³, Cam-Tu Nguyen^{1,2}

¹State Key Laboratory for Novel Software Technology, Nanjing University

²School of Artificial Intelligence, Nanjing University

³Huawei Technologies Co., Ltd.

Nanjing, Jiangsu, China

zihao_cheng@smail.nju.edu.cn

Abstract

Long-term memory is fundamental for personalized agents capable of accumulating knowledge, reasoning over user experiences, and adapting across time. However, existing memory benchmarks primarily target declarative memory, specifically semantic and episodic types, where all information is explicitly presented in dialogues. In contrast, real-world actions are also governed by non-declarative memory, including habitual and procedural types, and need to be inferred from diverse digital traces.

To bridge this gap, we introduce LifeBench, which features densely connected, long-horizon event simulation. It pushes AI agents beyond simple recall, requiring the integration of declarative and non-declarative memory reasoning across diverse and temporally extended contexts. Building such a benchmark presents two key challenges: ensuring data quality and scalability. We maintain data quality by employing real-world priors, including anonymized social surveys, map APIs, and holiday-integrated calendars, thus enforcing fidelity, diversity and behavioral rationality within the dataset. Towards scalability, we draw inspiration from cognitive science and structure events according to their partonomic hierarchy; enabling efficient parallel generation while maintaining global coherence. Performance results show that top-tier, state-of-the-art memory systems reach just 55.2% accuracy, highlighting the inherent difficulty of long-horizon retrieval and multi-source integration within our proposed benchmark. The dataset and data synthesis code are available at <https://github.com/1754955896/LifeBench>.

1 Introduction

Psychology and cognitive science [22] have long suggested that memory is not a single entity but a collection of distinct systems that process different kinds of information and operate under different principles. Declarative memory—comprising semantic memory for factual knowledge and episodic memory for personal experiences—supports the ability to recall events and contextual details. In contrast, non-declarative memory facilitates the gradual establishment of skills, habitual behaviors, preference and emotion-conditioned actions. Together, these systems enable humans to both recall the past and adapt behavior based on accumulated patterns, jointly driving daily decision-making and activities.

Recent studies in AI agent have primarily emphasized declarative memory, focusing on retrieving and reasoning over explicitly stored facts or episodic records. This paradigm is exemplified by benchmarks such as LOCOMO [13], and LongMemEval [28]—and memory-augmented systems, such as HippoRAG [9], MemOS [12].

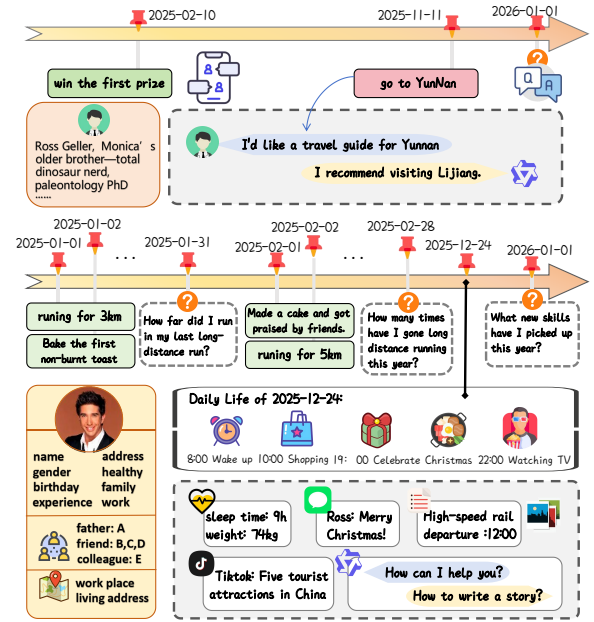


Figure 1: LifeBench captures rich human activities, which are driven by not only declarative memory (semantic or episodic) but also by non-declarative one (habits, skills, etc.). The user actions and habits need to be inferred from fragmented and diverse digital traces (lower part).

While these benchmarks have enhanced the capabilities of memory agents, they offer limited mechanisms for evaluating non-declarative reasoning, which is vital for personalized intelligence. Such knowledge enables systems to align task scheduling with user habits, suggest behavioral adjustments for healthier lifestyles, or optimize device operations for energy efficiency. Furthermore, at scale, anonymized commuting patterns can provide critical insights for urban design. Developing systems that reason over both memory types is therefore essential for building long-term AI agents that understand not just what users know, but the fundamental rhythms of how they live.

Capturing human memory is challenging, as it requires dense, long-term behavioral data that reflect skill acquisition, evolving routines, and causal events. Real-world datasets are often limited by privacy, ethical constraints, and sparsity. To address this, recent

Table 1: LifeBench vs existing memory benchmarks: Ctx depth is defined as the average number of tokens in user history. Information with (*) is estimated and reported in [28]. Core memory abilities include Information Extraction (IE), Multi-hop reasoning (MR), Temporal and Knowledge Updating (TKU), Nondeclarative Memory Reasoning (ND), and Unanswerable (UA).

Benchmark	Domain	Artifacts	Events	#Questions	Ctx Depth	Core Memory Abilities				
						IE	MR	TKU	ND	UA
MSC [30]	Open-Domain	Chats	-	-	1K	✗	✗	✗	✗	✗
PerLTQA [3]	Personal	Chats	-	8593	1M (*)	✓	✗	✗	✗	✗
LOCOMO [13]	Personal	Chats	-	7512	10K	✓	✓	✗	✗	✓
LongMemEval [28]	Chats	Chats	-	500	115K, 1.5M	✓	✓	✓	✗	✓
Mem-Pal [7]	Personal	Chats, Logs	Sparse	-	-	✓	✓	✗	✗	✗
LifeBench (ours)	Personal	Chats, Apps, Health Records	Dense	2003	3.66M	✓	✓	✓	✓	✓

work uses synthetic data generation with statistical models [34], GANs [29], and VAEs [26], and Large Language Models (LLM) [10, 33]. However, these approaches usually focus on isolated behavioral dimensions (e.g. mobility), without capturing the full spectrum of experiences driven by multi-source memory.

This study addresses these limitations by introducing a simulation environment that generates long-horizon, densely connected user experiences governed by human-inspired memory systems. Unlike existing benchmarks that rely solely on explicit dialogue, our framework reflects real-world conditions where memory must be inferred from fragmented signals such as chats, calendar entries, notes, SMS, and health records (Figure 1). Consequently, LifeBench provides a vital resource for developing and evaluating memory agents with tasks that mirror real-world complexity while eliminating the privacy concerns associated with real-user data.

Building such a benchmark presents two fundamental challenges: ensuring data quality and maintaining scalability. To ensure data quality, we incorporate real-world priors such as anonymized social surveys, map APIs, and holiday-integrated calendars. These priors ground the synthesized activities in behavioral rationality and diversity, ensuring that simulated events remain consistent and plausible. To achieve scalability, we organize events within a partonomic hierarchy that enables parallelized simulation while maintaining global coherence. This parallelization strategy significantly accelerates the pipeline, reducing the total simulation time from 58 hours to 8 hours per user-year. Performance results show that state-of-the-art systems achieve only 55.2% accuracy, highlighting the inherent challenge of LifeBench.

Our work makes four key contributions: (1) **Human Cognition-Inspired Synthesis**: we introduce a long-horizon synthesis system that generates year-scale behavioral trajectories grounded in cognitive science. Specifically, the framework models multiple human memory systems and organizes events according to a partonomic hierarchy to maintain consistency across various temporal scales; (2) **Dataset and Benchmark**: we construct a multi-source dataset spanning a full year, encompassing personas, daily events, phone artifacts, health records, and monthly summaries to support the modeling of human-like memory. To systematically evaluate reasoning capabilities, we design 2,003 questions across four main categories including information extraction, multi-source reasoning, temporal evolution, non-declarative memory reasoning; (3) **Empirical**

Evaluation: we perform an evaluation of state-of-the-art memory systems such as MemU [14], Hindsight [11], and MemOS [12]. By analyzing their performance and error patterns on our benchmark, we uncover key challenges inherent in long-term memory reasoning and multi-source integration; (4) **Reproducibility Package**: we release the complete dataset, synthesis framework, and documentation under the Apache License 2.0 to facilitate transparent and responsible research. We invite the community to utilize these toolkits for large-scale data generation and further investigation into personalized intelligence.

2 Related Work

Recent years have witnessed growing interest in modeling long-term user history for personalized conversational agents. The MSC dataset [30] focuses on open-domain chats across multiple sessions. PerLTQA [3] augments historical chats with social relationships, whereas Mem-Pal [7] includes both historical dialogues and application logs. LOCOMO [13] and LongMemEval [28] benchmark a range of memory reasoning tasks across multiple chat sessions. These datasets have laid the groundwork for a variety of memory systems, including A-Mem [31], Zep [20], Mem0 [2], MemInsights [21], HippoRAG [9], MemoryBank [35], MIRIX [27], and Hindsight [11].

These benchmarks, however, do not model densely connected events and activities, focusing mainly on semantic and episodic memory, and often assume that user information is explicitly and fully expressed through chats [3, 13, 28, 30]. In reality, human actions are governed by multiple memory systems and are observable only in fragmented form, distributed across various phone applications. Although Mem-Pal [7] incorporates log data, the application logs are simplified into plain text with explicit event descriptions. Moreover, user events in Mem-Pal are relatively sparse (around 2.5 logs or chats per day). As a result, Mem-Pal fails to capture the real challenge of extracting core memories from irrelevant data.

A detailed comparison between LifeBench and related datasets is provided in Table 1. Our benchmark is uniquely characterized by dense event information occurring at a rate of around 14 events per day. These events are partially observable across diverse digital artifacts (24 artifacts per day), which results in a significantly greater context depth exceeding twice the length of the trajectories in LongMemEval. Furthermore, LifeBench is the only framework that includes a specific task for non-declarative memory reasoning.

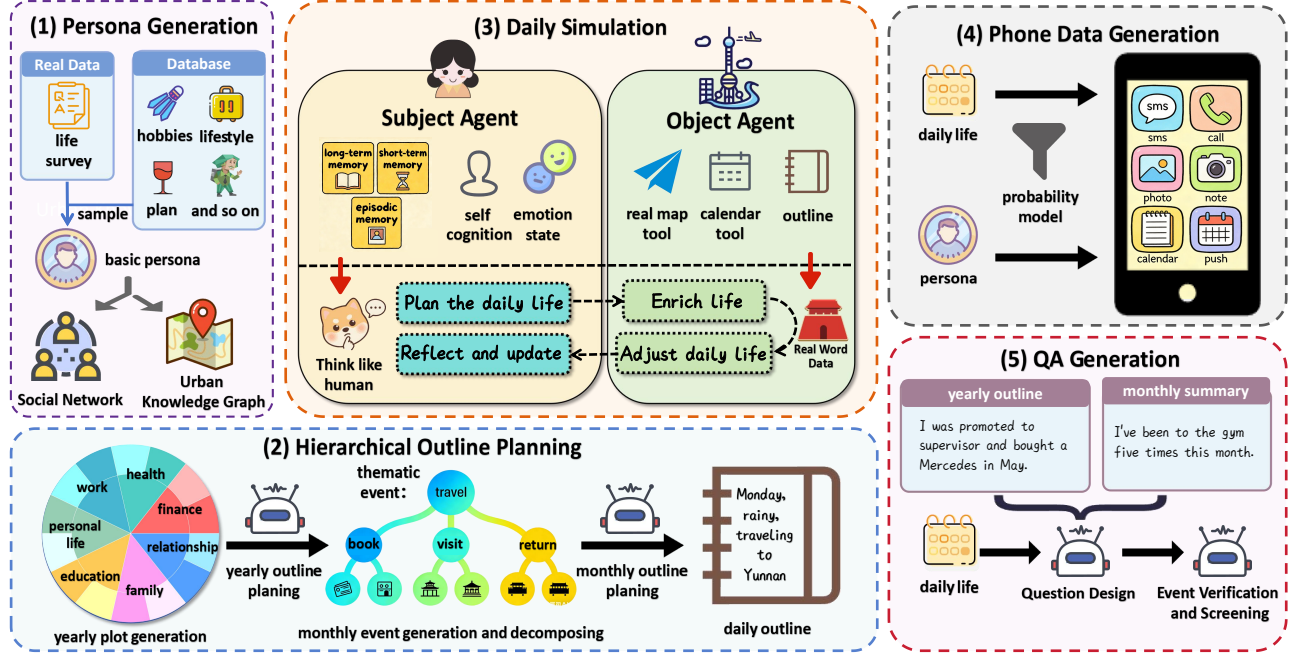


Figure 2: The LifeBench data synthesis pipeline architecture and key algorithmic components.

3 LifeBench: Memory Simulation Framework

3.1 Design Principles

Human Cognition Inspired Synthesis: our framework is built on two principles: First, we explicitly model how multiple human memory systems [22], including declarative memory (semantic and episodic) and non-declarative memory (habits, skills, emotional conditioning) influence human daily activities. Second, we represent events through a partonomic hierarchy [32] that captures their nested structure and ensures consistency across temporal scales.

Rationality and Diversity: We integrate diverse persona-conditioned generation and incorporate multi-source real-world priors, such as maps and calendars, to ensure behavioral diversity and contextual realism in the synthesized data.

Phone Artifacts: Human events and activities can only be observed through their digital traces, such as phone artifacts including calls, messages, calendar entries, agent chats, photos, notes, push notifications, and health records. This highlights a practical challenge that memory agents must address.

3.2 Synthesis Pipeline Architecture

Figure 2 illustrates our overall framework for simulating long-horizon, densely connected user activities. The framework comprises five LLM-based modules operating in a pipeline: the **persona synthesis module** (Section 3.2.1), the **hierarchical outline planning module** (Section 3.2.2), the **daily activity simulation module** (Section 3.2.3), and the **phone data generation module** (Section 3.2.4), and **QA generation** (Section 3.2.5). We use DeepSeek-R1¹ as the base LLM in our synthesis framework.

3.2.1 Persona Synthesis. Each user is modeled with a rich persona capturing demographics (e.g., age, gender, occupation, education, income), lifestyle (e.g., hobbies, habits), personality (e.g., traits, values, emotion), social relationships (e.g., family, friends, colleagues, and interaction patterns), and long-term goals (e.g., career, health, milestones). These personas are generated in three-steps: basic information generation, social networks construction, and real urban location assignments. The specific attribute fields and structures can be found in Appendix H.

Basic information generation. Base personas are constructed by sampling attributes from various sources, which are then refined and expanded through LLM rewriting and generation. First, background and lifestyles are sampled from anonymized survey data², which include attributes such as ethnic background, marital status, occupation, and health conditions. Second, additional attributes such as habits, plans, characters, and skills are sampled from the PersonaHub [4] dataset. Third, LLMs generate sensitive attributes (fiction names, addresses, etc.), and adjust sampled fields to maintain consistency. This process grounds each persona in realistic demographic and behavioral priors while ensuring that none of the generated personas correspond to real individuals.

Social network. Leveraging base personas, LLMs infer and generate social circles for each individual, and define the specific members and their relational networks within each circle. The resulting overall social network comprises approximately 20–30 associated individuals, each equipped with a detailed role profile.

¹<https://www.deepseek.com/>

²<https://www.cnsda.org/index.php>

Personalized urban landscape. We further construct a personalized urban landscape for each individual to ground their activities in real-world geography. Using the map API³ and LLMs, we map each persona's geographical information real-world locations within their respective city. Based on these mappings, we identify plausible anchor locations that serve as the primary sites of each individual's daily activities.

3.2.2 Hierarchical Outline Planning. We construct a personal event partonomic tree to support the modeling of user activities. Higher-level nodes represent long-span thematic events, capturing a participant's intentions and plans, whereas leaf nodes represent atomic events, defined as those shorter than one day and closely related to daily activities. We first describe the procedure for generating the event tree, and then explain how daily activities are subsequently derived from it (Section 3.2.3).

Plot Generation. To ensure coherence, we first establish a plot outline, a global story that serves as the narrative of an individual. Using a predefined database of approximately 50 candidate plots of seven categories (work, health, finance, relationship, family, education, and personal life), the LLM selects up to 10 plots tailored to the user's persona. To allow for controlled synthesis, our simulation framework supports customization via priority labels, enabling human-assigned priors to weight the selection of specific plots. The selected plots are ultimately integrated through temporal sequencing and optimization to produce a seamless, one-year chronological outline. The plot schema can be found in Appendix H.

Thematic Event Generation and Decomposition. Guided by the plot outline, the LLM generates thematic events month-by-month in an autoregressive manner. Each event includes attributes such as location, participants, duration, start and end dates. We assign thematic events to a month if their start dates fall within the given month window, though events may naturally span across multiple months. To ensure narrative and temporal consistency, the prompt includes the global plot outline, a summary of the preceding month's events, and relevant holiday data. Finally, we directly ask the LLM to decompose each thematic event into nested subevents until reaching atomic events, which are defined as those lasting less than one day, thereby constructing the event partonomic tree.

Atomic Event Optimization. Atomic events generated in the preceding phase are all mapped to their corresponding thematic events. Nevertheless, certain atomic events are inherently short-term in nature, and all events are susceptible to contextual factors including emotional states, self-cognition such as hobbies, and habits (**non-declarative memory**). To capture these dynamics and refine atomic events, the LLM analyzes and adjusts those within each two-week time window based on contextual factors derived from the user's persona and history. New atomic events can then be introduced to reflect users' daily routines such as regular physical exercise, while existing ones can be modified to better align with contextual conditions like weather conditions and personal moods.

3.2.3 Dual-Agent Daily Activity Simulation. Daily activities can be viewed as the smallest units of events, typically lacking explicit goals or plans. Each atomic event may consist of one or more daily

activities, and each activity may or may not have digital traces on the user phone (referred to as phone artifacts). To generate these daily activities, we adopt a dual-agent simulation framework: a subjective agent simulates human-like reasoning to produce plausible activities, while an objective agent verifies and grounds these activities in real-world data to ensure fidelity and consistency.

Subjective Agent: The Subjective Agent is an LLM that simulates human behavior to generate daily activities conditioned on atomic events, user profiles, and daily contextual factors (**non-declarative memory**). The LLM is equipped with three types of memory: *long-term memory*, derived from summaries of the preceding month and the current month; *short-term memory*, which stores detailed records of activities from the last seven days; and *episodic memory*, comprising the three most relevant daily activities retrieved from the accumulated storage of past activities (here, the day's atomic events are used as the query for retrieval).

Objective Agent: The Objective Agent enforces physical and temporal constraints, validating the feasibility of daily plans by checking for time conflicts, travel times via map APIs, venue availability, and consistency with established facts, while also introducing supplementary events and external environmental feedback. By doing so, the Objective Agent ensures the rationality and authenticity of the simulated activity data.

Overall Procedure: The overall daily simulation follows a day-by-day autoregressive manner, whereas the simulation for each day proceeds iteratively as follows: the Subjective Agent first proposes a daily plan based on scheduled atomic events, emotional state, habitual cognition; the Objective Agent then validates the plan, provides feedback, generates sparse time slots for supplementary events, and assigns real location and date information; finally, the Subjective Agent updates and optimizes the plan according to this feedback, refining the associated attribute information.

3.2.4 Phone data generation. Phone data artifacts serve as the raw observations for inferring underlying user activities, events, and habits. The main types of phone data are as follows:

- **Contacts:** Generated based on the persona's social network.
- **SMS, Calls, Push Notifications, Photos, Notes, Calendar, Agent-chats:** The LLM analyzes daily activities and estimates the likelihood of each type of mobile operation. Sampling is then applied based on the estimated likelihood, and the LLM generates the corresponding data. To simulate realistic interpersonal conversations, the LLM also generates topic-specific SMS content unrelated to personal events—such as discussions of third-party incidents, social topics, personal thoughts, or online content.
- **Noise Data:** Event-irrelevant mobile content, including random conversations, advertisements, web browsing records, and device notifications—some meaningless, others containing user information—is synthesized to simulate realistic background activity.

The complete phone data schema is provided in Appendix H. Due to the presence of noise and the fact that not all daily activities are captured in phone data, this introduces a unique challenge for benchmarking memory systems.

³<https://lbs.amap.com/>

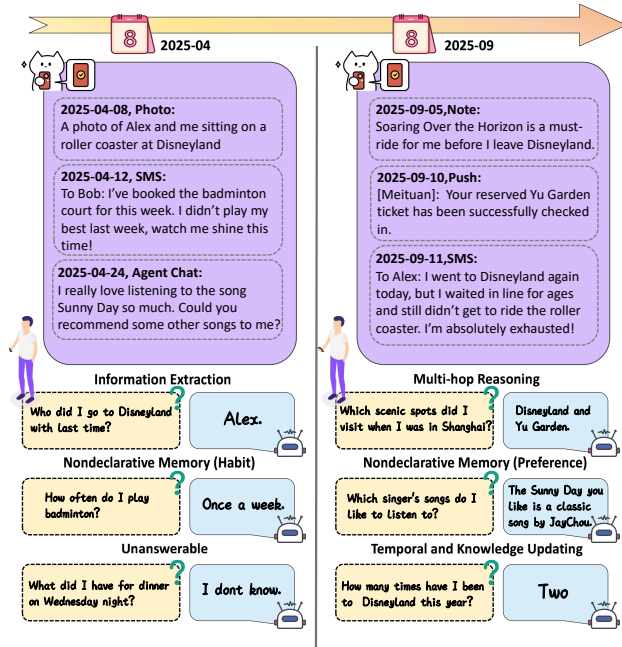


Figure 3: Samples of LifeBench question types.

Health and Sports Records. Daily health data include step counts, travel distances, and workout records (gym, running, cycling, swimming) with associated duration, intensity, and calorie consumption; sleep duration annotated with categorical tags (restful, interrupted); daily activity summaries; and physical recovery indicators. All health metrics are generated conditioned on daily activities, the daily contextual factors (persona’s habits, and fluctuations in daily emotion and stress levels). The full Health and Sports Records data schema is provided in Appendix H.

3.2.5 Question Answering Generation. An in-progress question generation mechanism produces queries dynamically during daily activity simulation rather than post hoc, enabling exact grounding and evidence assignment from corresponding phone data. The question design aims to address several distinct challenges that have been largely overlooked in existing memory benchmarks. Specifically, it targets temporal evolution testing, where time-dependent queries assess whether models can integrate new information as data accumulates; causal reasoning, which requires linking temporally distant events to infer underlying causes; and ambiguous concept handling, which tests the ability to manage updates and distinctions in attributes, names, and semantic nuances. We further include nondeclarative memory reasoning questions that probe emotions, habits, and preferences, requiring models to infer patterns from long-range daily activities.

Our benchmark includes five question categories with sample questions are shown in Figure 3. Each question category is designed to evaluate a distinct aspect of memory:

- **Information Extraction (IE):** Tests direct fact retrieval from individual memory records, such as locating the time of a specific gym session.

- **Multi-hop Reasoning (MR):** Requires aggregating and reasoning over multiple memory points, for example counting how many times the user met colleagues from the marketing team in Q1.
- **Temporal and Knowledge Updating (TKU):** Assesses temporal reasoning and the ability to handle evolving information, such as updating the total number of swimming sessions as time progresses.
- **Nondeclarative Memory Reasoning (ND):** Evaluates system understanding of user habits, skills, emotion and preferences, e.g., identifying the user’s usual gym routine.
- **Unanswerable (UA):** Contains questions that cannot be inferred from available data or are factually false, requiring the model to respond that the query is unanswerable (e.g., “How many bottles of milk did I drink on May 8th?”).

Each category includes two question formats. Open-ended QA prompts open-ended responses to evaluate completeness, coherence, and the appropriate level of abstraction. Multiple-choice QA provides four candidate answers plus a “not in memory” option to test discriminative ability and reduce hallucination. Paired questions targeting the same memory content allow direct comparison between recall and recognition performance.

Phone Data Supplementation: After formulating the questions, we analyze the related phone data to assess whether it provides sufficient information to answer the questions. If gaps are found, additional phone data are generated to fill in the missing details, ensuring that all questions—except unanswerable ones—can be answered based on the phone data.

Data Filtering and Manual Checking: An initial screening is performed using an LLM to remove questions that are unreasonable or whose answers can be directly inferred from the question itself. The remaining questions are then manually reviewed, with unreasonable items removed and answers and options rewritten or optimized as needed. This process produces the final QA dataset.

3.3 Scalability

A major challenge in LLM-based simulations of long-horizon, causally linked events is the high computational cost. A naive implementation based on sequential generation takes 58 hours to simulate one year of activities for a single user, with 3-4 hours for hierarchical outline planing, 40 hours for daily activity generation, and 14 hours for phone data generation. To overcome this, we accelerate the simulation based on the following key insights.

Parallel thematic event decomposition. In the previous section, thematic events are decomposed independently, but the degree of parallelism is limited by the number of thematic events. To accelerate this process, we design a parallelization method that maintains an event tree and recursively decomposes each “splittable” leaf node into subevents. This approach increases the degree of parallelism to the number of leaf nodes in the tree. A detailed description is provided in Algorithm 1 in Appendix B. We ultimately achieved an acceleration of event decomposition from approximately 2 hours to around 30 minutes through 24-thread parallelization.

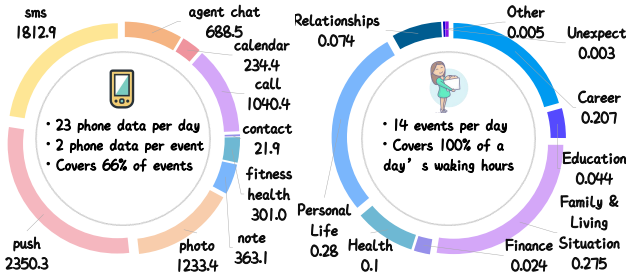


Figure 4: Statistics of Events and Phone Data.

Time-sliced daily simulation. The primary bottleneck lies in simulating daily activities. Using a day-by-day autoregressive approach, generating data for a single individual takes approximately 40 hours. To accelerate this process, we assume that activities separated by two weeks can be treated independently when conditioned on their corresponding atomic events. This assumption allows activities from different two-week periods to be generated in parallel, while those within the same period remain sequential. Since explicit activity memory is limited in this parallel setting, we supplement the long-term, short-term, and episodic memories with the corresponding atomic events, which were previously generated during the event decomposition stage. We ultimately completed the processing of one-year daily simulation in approximately 2 hours via 24-thread parallelization.

Parallel phone and health data generation. Mobile phone data and health records depend solely on daily activities and contextual factors, which were previously generated during the atomic event generation and daily activity simulation steps; therefore, data for each day can be generated independently in parallel. This process takes approximately 1-2 hours, keeping the total simulation time for one person under 8 hours.

3.4 Data Statistics and Quality Assessment

3.4.1 Data Statistics. Using the data simulation framework, we synthesize one year long of user activities for 10 users. Each user has a total of 5,149 events, averaging roughly 14 events per day, with a diverse distribution illustrated in Figure 4. Additionally, each user has 8046 phone data and health records, with their distribution shown on the left of Figure 4. Full data statistics are provided in Table 5 in Appendix.

Question Statistics. A total of 2,003 questions cover five categories, with the distribution shown in Table 2. Each question is associated with a timestamp to evaluate the system's ability to capture memory evolution, and the temporal distribution of the questions is illustrated in Figure 5.

3.4.2 Data Quality Evaluation. We employ automated and manual evaluations to assess the consistency, authenticity, and diversity of the synthesized event data.

Manual Evaluation. We evaluated the *rationality* of daily activity and phone data for five individuals over one month (about 400–500 activities per user). For daily activities, we checked for temporal

Table 2: Statistics on QA pairs of LifeBench

Type	Subtype	Ques. (Percent)
Category	IE	718 (35.85%)
	MR	597 (29.81%)
	TKU	229 (11.43%)
	ND	429 (21.42%)
	UA	30 (1.50%)
Format	Multiple Choice	517 (25.81%)
	Short Answer	1486 (74.19%)
Total		2003 (100.00%)

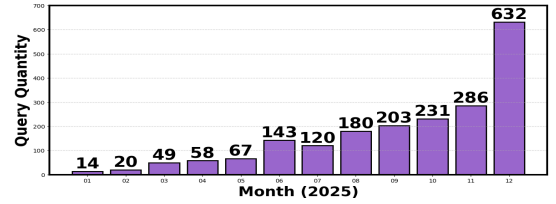


Figure 5: Distribution of Query Time.

coherence, content, location, and participant consistency, marking any unreasonable entries as invalid. For phone data, each data entry was matched to the corresponding activities and assessed for consistency and rationality, with invalid entries recorded. The proportions of valid data entries were then calculated.

For diversity evaluation, we exploited a 10-point scale based on five criteria: (1) inter-individual variability, (2) intra-individual variability across days, (3) non-stereotypical representativeness: a balanced mix of individual-specific and common events that avoid rigid character stereotypes, (4) fine-grained metric diversity: fluctuation in fine-grained metrics such as sleep or diet, and (5) plausibility: reflecting the complexity and unpredictability of real life while remaining coherent. Each criterion contributes up to 2 points, totaling 10 points. The detailed information for manual evaluation rubrics is given in Appendix D.1.

Automatic Evaluation. We used LLM and the AMap API to automatically evaluate the generated event data along four dimensions: *relation consistency* (alignment with predefined social relations), *location authenticity* (verification of places against real-world maps), *trip authenticity* (plausibility of travel durations and routes), and *event diversity* (balance and variety of event types measured by Normalized Shannon Entropy and Simpson Diversity). Higher scores across all dimensions indicate greater realism and diversity in the generated data. Detailed definitions and computation procedures are provided in Appendix D.

3.4.3 Quality Assessment Results. Figure 6 presents the manual evaluation results. Overall, the synthesized data demonstrates high rationality (0.962–1.0) across both daily activity and phone data. Minor irrationalities mainly stem from repeated events or step jumps due to hallucinations. No major logical inconsistencies are observed,

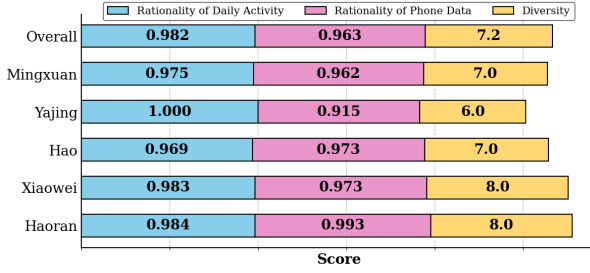


Figure 6: Manual Evaluation on Synthetic Data.

Table 3: Automatic Data Quality Assessment Results

Category	Metric	Value
Consistency	Relation consistency	0.83
Authenticity	Location authenticity	0.83
	Trip authenticity	0.75
Diversity	Normalized Shannon Entropy	0.75
	Simpson Diversity Index	0.78

and most content remains coherent and authentic. In terms of diversity, the data performs well in both inter- and intra-individual variability but shows room for improvement in fine-grained metric diversity and the non-stereotypical representativeness.

Table 3 presents the automated assessment of synthetic data quality. Relational consistency and location authenticity meet the expected levels, reflecting the effectiveness of our design that integrates human-curated personas with real-world priors. Trip authenticity is slightly lower, likely due to temporal constraints in the real-world datasets, which limit access to transportation conditions for specific simulated periods. Regarding diversity, both Normalized Shannon Entropy (0.75) and Simpson Diversity Index (0.78) show moderately high diversity, i.e. the generated events are diverse with only mild concentration in certain categories.

4 Evaluating Memory Systems on LifeBench

4.1 Memory Systems

To illustrate the difficulty of LifeBench, we conduct empirical evaluations with a number of state-of-the-art memory systems that have recently achieved top performance on existing benchmarks, including LoCoMo [13] and LongMemEval [28]. These systems represent the current frontier of long-term memory modeling in LLMs, covering diverse architectures and retrieval mechanisms. By testing them under our LifeBench, we aim to examine their drawbacks in more complex, temporally extended scenarios.

MemU [13] is a memory framework for 24/7 proactive agents, using a dual-agent setup where a primary agent executes tasks while a memory agent (MemU Bot) monitors interactions, summarizes memories, and predicts user intentions.

Hindsight [11] structures memory agent into four networks: world facts, personal experiences, synthesized observations, and evolving opinions. This design separates raw data from inferred

knowledge. It implements three core operations (retain, recall, reflect) through TEMPR (Temporal Entity Memory Priming Retrieval) for *memory storage and retrieval*, and CARA (Coherent Adaptive Reasoning Agents) for *preference-conditioned reflection*. The system uses multi-strategy retrieval combining semantic, graph, temporal, and maintains an opinion network with updated confidence scores.

MemOS [12] serves as a memory operating system that *integrates multi-source LLM memories* such as plaintext, activation (KV-cache), and LLM parametric memories within a unified “MemCube” abstraction. It manages the entire memory lifecycle with metadata recording source history, version control, and access policies. It adopts a three-tier architecture comprising working memory, long-term storage, and a cold archive, with dynamic memory migration across tiers. An intent-aware scheduler proactively preloads or fuses relevant memories based on anticipated task demands.

4.2 Experimental Setup

We evaluate the memory systems using a unified base LLM, standardized input formats, and a consistent evaluation protocol.

Base LLM and Data Processing. GPT-5.1-Mini [17] is used for both memory operations and answer generation, while text-embedding-3-small [18] handles text embeddings. As for input data, we convert our dataset to the standard format of LoCoMo test configurations [11–13], which are supported by all the memory systems. Since existing benchmarks are primarily designed for conversational text and cannot directly process structured mobile phone data (e.g., contacts, calls, SMS, calendar entries, agent chats, photos, notes, push notifications, and health data), we create a summarized info field for structured types. This field condenses contextual information into textual form, enabling easier processing by current memory systems. Future memory systems may choose to leverage either the summarized or raw structured data.

Evaluation Protocol. We use an LLM-as-judge approach, employing GPT-5.1-Mini as the judge LLM. The standard LoCoMo evaluation prompt [13] is used to determine if a model correctly answers each question. The model score is calculated as the proportion of its correctly answered questions.

4.3 Experimental Results

Figure 7 shows overall accuracy across all question types as well as by category. The top-performing model, MemOS, achieves only 55.22% overall accuracy, with Hindsight ranking second at 40.99%. Notably, Hindsight scores around 90% on existing memory benchmarks (LoCoMo and LongMemEval) [11], highlighting that current benchmarks are no longer sufficient for testing state-of-the-art memory systems due to their simplified assumptions. In contrast, LifeBench presents a much-needed challenge to drive further progress in memory system research.

Although MemOS is the top-performing model on average, its performance varies across QA categories, as shown in Figure 7. Specifically, MemOS outperforms Hindsight on IE, MR, and TKU, but falls behind on Non-Declarative Memory Reasoning (ND) and Unanswerable (UA) questions. Its strong results on factual and temporal questions (IE, MR, TKU) is likely due to the benefit of integrating multiple source LLM memories (plaintext, KV cache,

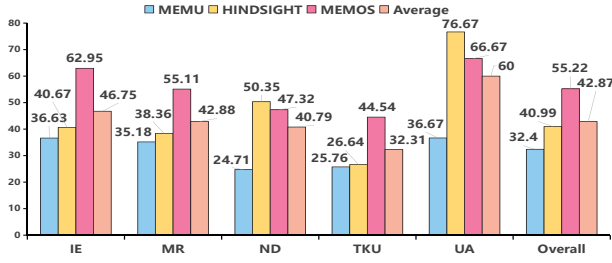


Figure 7: Performance of Memory Systems on LifeBench.

LLM model parameters) within the unified MemCube. In addition, MemOS’s success on TKU questions demonstrates the effectiveness of its memory version management in handling evolving memories.

Despite ranking second overall, Hindsight excels on questions about preferences (a part of ND type), thanks to its ability to model preference-conditioned reflection and opinion networks. Additionally, its higher performance on UA reflects its tendency to refrain from answering ambiguous queries. Conversely, MemOS attempts to answer most questions, which lowers its UA scores but allows it to perform better on other categories.

Among the three systems, MemU performs significantly worse than Hindsight and MemOS, despite achieving better results over Hindsight on LoCoMo [12]. This is mainly because MemU often produces brief memory summaries that fail to capture the complexity and nuances of long-horizon daily activities in LifeBench.

4.4 Error Analysis

This section provides a qualitative analysis to draw insights into the memory reasoning abilities of MemOS and Hindsight.

The first case study (Question 1) is presented in Table 4, where MemOS provides the correct answer while Hindsight does not. Analysis of the retrieved memories reveals that Hindsight’s outputs omit the event of making fabric kitten pendants, focusing instead on general content related to cats and neighbors. The retrieval strength of MemOS originates from its design to standardize multi-source (photos, sms) in a standardized memory units (the MemCube). Further examination of their reasoning processes shows that Hindsight overlooked temporal cues such as “right after” and key relational terms like “parent-child,” resulting in an incorrect response based on an unrelated cat event from a different month. MemOS, however, successfully incorporated these contextual and temporal constraints to produce the correct reasoning outcome.

The second case study (Question 2) is presented in Table 4, where Hindsight produces the correct answer while MemOS does not. Analysis of the retrieved results shows that both systems identified the event of taking care of the cat for the first time. However, Hindsight’s memory units include richer descriptions of the user’s psychological states and behaviors, whereas MemOS focuses primarily on factual event records. During response generation, Hindsight tends to analyze the user’s emotions and reasoning to produce more detailed and comprehensive answers, while MemOS favors concise factual statements. More case studies are given in Appendix E.2 that reveal common errors of MemOS and Hindsight include

Table 4: Case Studies of Memos vs Hindsight

Question 1: After the neighbor came to express their gratitude, what cat-themed parent-child activity did I do right after?

A: Went to the neighbor’s house with my daughter...

B: Told the whole family about the experience ... the cat.

C: Watched cat videos and made fabric kitten pendants.

D: Discussed the possibility of keeping a pet in the future.

MemOS: C; **Hindsight:** A, **Groundtruth:** C

Question 2: How was Ma Xiulan’s mood or state when she took care of her neighbor’s cat independently for the first time?

MemOS: Happy and pleasantly surprised

Hindsight: She was initially a little nervous, scared and cautious, but after finishing the first feeding and litter box cleaning, she felt at ease and built up a sense of security and confidence. Over the following days, the cat gradually relaxed ...

Groundtruth: She felt a little nervous and her movements were unskilled, but she successfully completed the task.

failure to retrieve correct evidence, incomplete evidence retrieval, hallucinated content, incorrect reasoning, and memory omission.

Implications for Future Memory System Design: A system that addresses multiple facets of human memory is essential. Unfortunately, matching, linking, and aligning memories across different data sources and memory types remains a significant challenge. In addition, temporal reasoning and knowledge updating remain major bottlenecks because current systems rely mainly on similarity-based retrieval and lack mechanisms for temporal constraints or reasoning. Developing specialized temporal indexing and time-aware query processing is therefore necessary to accurately track the evolution of memory. Furthermore, a large scale of synthetic data can facilitate memory training [12]. The implementation of such systems must prioritize ethical and responsible use, as discussed further in Appendix G.

5 Conclusion

In this paper, we introduce LifeBench, a synthetic data framework capable of generating long-term, densely connected personal events and activities. Inspired by cognitive science, our framework is built on two principles: 1) modeling multiple human memory systems, including declarative and non-declarative memory, in shaping human actions, and 2) representing events through a partonomic hierarchy that captures their decomposable structure. We also design a parallelization solution that reduces the synthesis time for year-long dense activities for a single individual from 58 hours to 8 hours.

Based on this framework, we develop a year-long dataset covering 10 users, including detailed personas and activities with underlying causal events. Evaluations of state-of-the-art memory systems reveal the challenges posed by LifeBench, highlighting its value as a resource for advancing AI memory systems. These benchmarks and datasets can also support future applications in personalized AI assistants, digital health coaching, behavioral research. We release the complete dataset, generation code, and documentation to support reproducible and responsible research.

References

- [1] Vannevar Bush et al. 1945. As we may think. *The atlantic monthly* 176, 1 (1945), 101–108.
- [2] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshray Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv:2504.19413* (2025).
- [3] Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjuan Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. PerLTQA: A Personal Long-Term Memory Dataset for Memory Classification, Retrieval, and Fusion in Question Answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*. 152–164.
- [4] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv:2406.20094* (2024).
- [5] Jim Gemmell, Gordon Bell, and Roger Lueder. 2006. MyLifeBits: a personal database for everything. *Commun. ACM* 49, 1 (2006), 88–95.
- [6] Ao Guo, Jianhua Ma, I Kevin, and Kai Wang. 2018. From user models to the cyber-i model: approaches, progresses and issues. In *Proceedings of 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress*. 33–40.
- [7] Zhaopei Huang, Qifeng Dai, Guozheng Wu, Xiaopeng Wu, Kehan Chen, Chuan Yu, Xubin Li, Tiezheng Ge, Wenxuan Wang, and Qin Jin. 2026. Mem-PAL: Towards Memory-based Personalized Dialogue Assistants for Long-term User-Agent Interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [8] Zihan Huang, Tong Li, and Yong Li. 2025. AppGen: Mobility-aware App Usage Behavior Generation for Mobile Users. *IEEE Transactions on Mobile Computing* (2025).
- [9] Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems* 37 (2024), 59532–59569.
- [10] Chenlu Ju, Jiaxin Liu, Shobhit Sinha, Hao Xue, and Flora Salim. 2025. Trajllm: A modular llm-enhanced agent-based framework for realistic human trajectory simulation. In *Companion Proceedings of the ACM on Web Conference 2025*. 2847–2850.
- [11] Chris Latimer, Nicol   Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. 2025. Hindsight is 20/20: Building Agent Memory that Retains, Recalls, and Reflects. *arXiv:2512.12818* (2025).
- [12] Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, et al. 2025. Memos: A memory os for ai system. *arXiv:2507.03724* (2025).
- [13] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *The 62nd Annual Meeting of the Association for Computational Linguistics*. 13851–13870.
- [14] MemU [n. d.]. <https://github.com/NevaMind-AI/memU>.
- [15] Cataldo Musto, Giovanni Semeraro, Cosimo Lovascio, Marco De Gemmis, and Pasquale Lops. 2018. A framework for holistic user modeling merging heterogeneous digital footprints. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 97–101.
- [16] Ilan Naiman, N Benjamin Erichson, Pu Ren, Michael W Mahoney, and Omri Azenkot. 2023. Generative modeling of regular and irregular time series data via koopman VAEs. *arXiv:2310.02619* (2023).
- [17] OpenAI. 2025. GPT-5.1 mini. <https://openai.com>.
- [18] OpenAI. 2025. text-embedding-3-small. <https://openai.com>.
- [19] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.
- [20] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: a temporal knowledge graph architecture for agent memory. *arXiv:2501.13956* (2025).
- [21] Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. MemInsight: Autonomous Memory Augmentation for LLM Agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 33136–33152.
- [22] Larry R Squire. 2004. Memory systems of the brain: a brief history and current perspective. *Neurobiology of learning and memory* 82, 3 (2004), 171–177.
- [23] Wang-Chiew Tan, Jane Dwivedi-Yu, Yuliang Li, Lambert Mathias, Marzieh Saeidi, Jing Nathan Yan, and Alon Y. Halevy. 2023. TimelineQA: A Benchmark for Question Answering over Timelines. In *Findings of the Association for Computational Linguistics: ACL 2023*. 77–91.
- [24] The Lifelog Search Challenge 2018-2026. <https://lifelogsearch.org/lsc/>.
- [25] Quang-Linh Tran, Binh Nguyen, Gareth JF Jones, and Cathal Gurrin. 2025. Open-LifelogQA: An Open-Ended Multi-Modal Lifelog Question-Answering Dataset. *arXiv:2508.03583* (2025).
- [26] Zhiqiang Wan, Yazhou Zhang, and Haibo He. 2017. Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 IEEE symposium series on computational intelligence (SSCI)*. 1–7.
- [27] Yu Wang and Xi Chen. 2025. Mirix: Multi-agent memory system for llm-based agents. *arXiv:2507.07957* (2025).
- [28] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The Thirteenth International Conference on Learning Representations*.
- [29] Gang Xiong, Zhishuai Li, Meihua Zhao, Yu Zhang, Qinghai Miao, Yisheng Lv, and Fei-Yue Wang. 2023. TrajSGAN: A semantic-guiding adversarial network for urban trajectory generation. *IEEE Transactions on Computational Social Systems* 11, 2 (2023), 1733–1743.
- [30] Jing Xu, Arthur Szlam, and Jason Weston. [n. d.]. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). 5180–5197.
- [31] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems* (2025).
- [32] Jeffrey M Zacks and Barbara Tversky. 2001. Event structure in perception and conception. *Psychological bulletin* 127, 1 (2001), 3.
- [33] Jiarui Zhang. 2024. Guided Profile Generation Improves Personalization with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*. 4005–4016.
- [34] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)* 42, 4 (2017), 1–41.
- [35] Wanjuan Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19724–19731.

A Related Work on Behavior Modeling

Beyond memory benchmarking, our research is closely situated within the broader field of human behavior modeling. The idea of organizing personal information traces back to the Memex system [1], and has since been advanced through projects such as MyLifeBits [5], Musto [15], and Cyber-I [6], among other notable lifelogging research [24, 25]. MyLifeBits [5] utilized simple SQL-based architectures and keyword indexing to manage personal information. Musto [15] proposed an integrated user modeling framework that fuses heterogeneous digital footprints from social networks and personal devices, while Cyber-I [6] extended this idea by modeling affective, behavioral, and cognitive dimensions. The Lifelog Search Challenge (LSC) [24, 25] further explored records of personal events generated from digital sources such as photos or shopping data. Despite these advances, real-world lifelog data remain limited in scale and availability due to privacy concerns, ethical limitations, and prohibitive costs.

Synthetic data generation has emerged as a promising approach to address these challenges for modeling human behaviors. Early work relied on statistical methods [34] and generative models such as GANs [29] and VAEs [26] [16]. More recently, large language models (LLMs) have enabled the creation of rich textual and structured personal data [10, 33]. However, these studies mainly focus on synthesizing single-aspect data, such as user profiles [33], urban mobility traces [10, 29], or simple app usage trajectories [8], typically lacking user-centric full memory modeling. Recent works, including TimelineQA [23] and Generative Agents [19], partially address this issue but struggle with data fidelity. These approaches either do not ground on realistic temporal and location information [19] or rely on rigid template-based generation [23], failing to reflect the complexity of actual human experience.

B Scalability

Parallel thematic event decomposition is provided in Algorithm 1.

Algorithm 1 Parallel Event Decomposition (1-Day Threshold)

Require: *EventTree* with all thematic events as leaves, Time threshold $\tau = 1$ day

Ensure: Expanded *EventTree* with partonomic hierarchy

```

1: Function DFSParallelEventDecomp(EventTree,  $\tau$ )
2: dfs_stack  $\leftarrow$  Empty Stack
3: Assign all leaf nodes with splittable = True
4: for each leaf node in EventTree where node.splittable = True
   do
5:   Push(dfs_stack, node)
6: end for
7: while dfs_stack is not Empty do
8:   current_node  $\leftarrow$  Pop(dfs_stack)
9:   Assign current_node to an available worker
10:  The following is done in parallel for all popped nodes in current batch
11:  subevents  $\leftarrow$  LLMdecomp(current_node,  $\tau$ );
12:  for each sube  $\in$  subevents do
13:    if sube.duration <  $\tau$  then
14:      sube.splittable = False
15:    else
16:      sube.splittable = True
17:      Push(dfs_stack, sube)
18:    end if
19:  end for
20:  current_node.subevents = subevents
21: end while
22: RETURN EventTree

```

C Dataset Statistics

The statistical data of the dataset can be found in Table 5.

D Detailed Evaluation Metrics

D.1 Manual Evaluation

Scoring Criteria for Rationality Evaluation:

- (1) For each piece of daily activity data, we assessed the rationality of its temporal coherence, content, location, and participant. If any of these dimensions was deemed unreasonable, the data was marked as invalid, and the proportion of invalid data was calculated at last.
- (2) For each piece of phone data, we traced it back to the corresponding daily activity data in the matching time period, then evaluated the consistency between phone operations and daily activities as well as the rationality of phone operation content. If either dimension was unreasonable, the data was labeled as invalid, with the final proportion of invalid data statistically tabulated.

Scoring Criteria for Diversity Evaluation:

- (1) There are significant differences between the daily activities of the subject and those of other subjects, which can reflect the individual's unique characteristics (2 points).

Table 5: Dataset statistics

Metric	Value
Total users	10
Duration per user	1 year
Total events	51491
Events per user (mean)	5149
Events per user (median)	5147
Persona tokens (mean)	17730.8
Relationship (mean)	24.4
location (mean)	15
Contacts per user (mean)	21.9
Calls per user (mean)	1040.4
SMS messages per user (mean)	1812.9
Calendar events per user (mean)	234.4
Agent chat conversations per user	688.5
Photos per user (mean)	1233.4
Notes per user (mean)	363.1
Push notifications per user (mean)	2350.3
Fitness health records per user	301.0
Monthly summaries per user	12
Summary length (words, per user)	1737.7
Context depth (tokens) per user	3.66M
Total dataset size	332 MB
Generation time per user	8 hours

- (2) There are significant differences in the activities of the same subject on different dates, and the activities are not homogeneous (2 points).
- (3) On the basis of reflecting individual characteristics, the data includes a certain number of randomly diverse events and population common events, rather than only events that follow the stereotypes of character portraits (2 points).
- (4) On the basis of criteria (1) and (2), the fine-grained metrics of the subject (e.g., sleep duration, body weight, and dietary intake) also exhibit obvious fluctuations (2 points).
- (5) Verified by manual analysis, the data can reflect the complexity, unpredictability and rich diversity of real life, and the overall content is coherent and without sense of incongruity (2 points).

Total Score: 10 points (2 points are assigned to each of the 5 evaluation criteria, 0 points for completely unsatisfactory, 1 point for relatively satisfactory, and 2 points for highly satisfactory.).

D.2 Automatic Evaluation Metrics

We implement a comprehensive and fully automated data quality evaluation pipeline (Figure 8) to assess the **consistency**, **authenticity**, and **diversity** of the generated event data. Overall, the proposed framework consists of two complementary components:

- *Intra-source consistency evaluation*, which verifies whether generated events remain coherent with the original structured source data, such as persona profiles and predefined

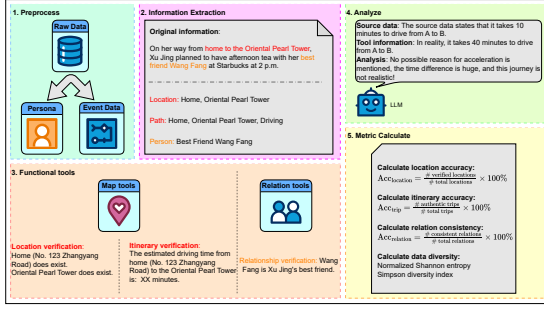


Figure 8: Automated data quality evaluation pipeline for consistency, authenticity, and diversity assessment.

social relations. This component primarily relies on lookup-table construction and LLM-assisted information extraction to determine whether referenced entities and relations can be consistently traced back to the source data.

- *Extra-source authenticity evaluation*, which validates generated content against real-world knowledge beyond the source data. In particular, we leverage real-world geographic and transportation information provided by the map API to assess the plausibility of locations and travel trajectories mentioned in the generated events.

Based on the above two components, we evaluate the quality of generated events from four complementary aspects: *relation consistency*, *location authenticity*, *trip authenticity*, and *event diversity*.

Relation Consistency. To evaluate character and relationship consistency, we construct a relation lookup dictionary from persona profiles, which specifies the expected social network for each character. We then apply LLM-based information extraction to identify character mentions and relational references in generated event narratives.

If a character appears in an event but cannot be matched to any entry in the predefined relation dictionary, the mention is marked as inconsistent. Relation consistency accuracy is computed as:

$$\text{Acc}_{\text{person}} = \frac{\# \text{ consistent relation mentions}}{\# \text{ total relation mentions}} \times 100\%$$

This metric quantifies the degree to which generated events adhere to the original interpersonal structure.

Location Authenticity. For location authenticity, place names are automatically extracted from event descriptions and queried against the AMap API. Locations that can be successfully verified by the map service are labeled as authentic, while unverifiable or ambiguous locations are considered invalid.

Location authenticity accuracy is calculated as:

$$\text{Acc}_{\text{location}} = \frac{\# \text{ verified locations}}{\# \text{ total extracted locations}} \times 100\%$$

This evaluation ensures that generated locations correspond to real-world geographic entities.

Trip Authenticity. For events involving travel or movement, we further evaluate trip-level authenticity. Given extracted origin-destination pairs and temporal information from event narratives,

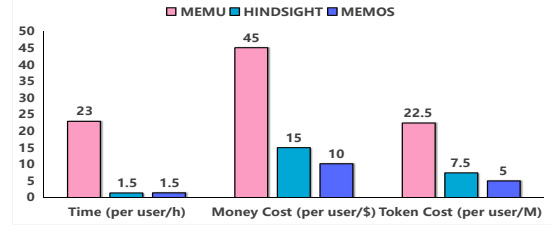


Figure 9: Time and cost of each system.

we query the AMap API to obtain reference travel durations and routes.

A synthesized trip is considered authentic if its temporal deviation falls within a predefined tolerance threshold ($\pm 20\%$ relative deviation or ± 30 minutes absolute deviation) compared to the API-provided reference. Trips exceeding this threshold are still considered authentic if the original data provides a plausible explanation for the temporal discrepancy (e.g., traffic congestion). Trip authenticity accuracy is defined as:

$$\text{Acc}_{\text{trip}} = \frac{\# \text{ authentic trips}}{\# \text{ total evaluated trips}} \times 100\%$$

This metric assesses whether generated travel descriptions are consistent with realistic transportation constraints.

Diversity Evaluation. To quantify event diversity, we analyze the distribution of event types using two complementary metrics.

Normalized Shannon Entropy measures the uncertainty and coverage of event types:

$$H = - \sum_{i=1}^K p_i \log p_i, \quad H_{\text{norm}} = \frac{H}{\log K}$$

where K denotes the total number of event categories and p_i is the proportion of events belonging to category i .

Simpson Diversity Index captures the evenness of the distribution. We first compute the concentration measure:

$$D = \sum_{i=1}^K p_i^2$$

and define the Simpson diversity index as:

$$\text{Simpson} = 1 - D = 1 - \sum_{i=1}^K p_i^2$$

Higher values of both metrics indicate more balanced and diverse event distributions, whereas lower values suggest concentration in a small number of event types.

E Additional Results

E.1 Time Complexity and Cost Analysis

We conducted a coarse-grained analysis and estimation of the time and cost required by each system to process a single user, as summarized in Figure 9. It is evident that MEMU incurs substantially higher time overhead and monetary cost than Hindsight and MemOS. This difference primarily stems from MEMU's reliance on frequent LLM invocations throughout its pipeline. Beyond text embedding,

MEMU invokes the LLM at nearly every processing stage, with a single memory operation requiring multiple LLM calls. In contrast, Hindsight and MemOS support batch processing and avoid repeated LLM invocations during memory handling, resulting in significantly lower computational overhead and cost.

E.2 Extension of Error Analysis

E.2.1 Failure to retrieve correct evidence.

- *Question:* After Ye Mingxuan started his weight loss plan in May, did he perform his first aerobic exercise from the plan before or after his first strength training session?
- *Generated Answer from MemU:* Uncertain: There is a strength training record from May (around May 17), but there is no clear first aerobic record from May for comparison.
- *Standard Answer:* After.
- *Analysis:* The system's memory clearly recorded that Ye Mingxuan first performed weight training and then aerobic training, but the system did not realize that weight training was strength training, which is why it could not retrieve the specific content.

E.2.2 Incomplete evidence retrieval.

- *Question:* What specific tasks did Feng Haoran complete in February in preparation for his promotion performance review report?
- *Generated Answer from MemU:* February work: Organize annual learning materials; refine the "Teamwork and Leadership" section; draft a six-month learning plan; send the initial draft of the performance review to General Manager Qian and set a schedule for optimization.
- *Standard Answer:* He conceived the report framework, organized team collaboration materials, compiled personal continuous learning documentation, wrote the section on team contributions, refined the future learning plan, and integrated the findings to complete the first draft of the report.
- *Analysis:* The answer is incomplete. The system's memory records that a report framework was conceived, team collaboration materials were organized, and personal continuous learning documents were compiled, but these were not included in the answer.

E.2.3 Hallucinated content.

- *Question:* What are the specific dates and locations for Yu Xiaowen's high school reunion (initiated by Chen Siyu) on May 30, 2025?
- *Generated Answer from MemU:* May 30th entry: Scheduled to return to Luoyang on June 1st, 2025 for a classmate reunion; another reunion on June 18th, 2025 at Dian Dou De Restaurant (Chen Siyu will be present).
- *Standard Answer:* Unable to answer.
- *Analysis:* The data only showed that a gathering took place but did not specify when the next gathering would take place, leading to over-inference in the system's response.

E.2.4 Incorrect reasoning.

- *Question:* Where did Feng Haoran mainly work in January?

- *Generated Answer from MemU:* Primarily based in the company office in Changsha; during holidays or business trips, work remotely or via video from hotels or homes.
- *Standard Answer:* Huatai Securities Co., Ltd. has its office located at No. 88, Furong Middle Road, Kaifu District, Changsha.
- *Analysis:* The system ignored the crucial limiting information that Feng Haoran worked at the company throughout January, and only in later months did he work in more locations.

E.2.5 Memory omission.

- *Question:* During the Spring Festival, while helping his mother prepare ingredients, Feng Haoran had an in-depth discussion with his father about which report?
- *Generated Answer from MemU:* The provided materials did not specify which report it was, and the information was insufficient to determine the source.
- *Standard Answer:* Central Bank Monetary Policy Report.
- *Analysis:* The system considered the Central Bank Monetary Policy Report information unimportant and did not fully incorporate it into its memory, merely using "report" as a substitute. This resulted in a lack of memorized information when answering the question, leading to an inability to respond correctly.

E.2.6 Hallucination in Memory Unit Construction.

- *Question:* In July, when Ma Xiulan asked her neighbor for advice on aphid control, what two homemade remedies did the neighbor recommend?
- *Generated Answer from MemOS:* Tobacco water and chili-garlic water
- *Standard Answer:* Tobacco water and laundry detergent water.
- *Analysis:* We analyzed the corresponding memory units of Memos and found that it did capture the note data recording tobacco water and laundry detergent water. However, it autonomously added the non-existent hallucination of chili-garlic water—which was not present in the original data—during the process of generating summaries for memory units, leading to an error.

E.2.7 Retrieval without Temporal Constraints.

- *Question:* When handling the annual fixed expenses due in August, what specific arrangements did I make on the day before the actual payment?
- *Generated Answer from Hindsight:* There is no explicit record in the memory of the specific arrangements you made on "the day before the actual payment" (i.e., August 10, 2025).
- *Standard Answer:* I withdrew 600 yuan in cash from the family emergency cash box, retrieved the house property certificate and household registration booklet from the chest of drawers, put them into a kraft paper document bag, and placed the bag in a prominent position on the TV cabinet in the living room.
- *Analysis:* We found that the payment event was successfully retrieved; however, the retrieval purely based on similarity introduced irrelevant financial payment-related information

for the entire year that was unrelated to the query target, and failed to retrieve events from the day before the payment based on time constraints, resulting in the inability to answer the query.

F Impact and Use Cases

F.1 Implications for Memory System Design

Our benchmark results reveal several insights for designing effective memory-augmented personal agents:

Comprehensive Memory Structure Modeling Developing a comprehensive system that addresses the multiple dimensions of human memory—such as declarative and non-declarative memory—is vital for realizing truly personalized AI agents. While frameworks like Hindsight and MemOS have addressed to some extent, a system that holistically addresses all aspects of LifeBench has yet to be developed.

Temporal Reasoning and Knowledge Updating Are a Major Bottleneck. Current systems rely primarily on similarity-based retrieval and lack the temporal reasoning necessary to handle time-sensitive data. This limitation introduces significant noise and can lead to the omission of critical information. To address this, it is recommended to implement specialized temporal indexing and time-aware query processing.

Multi-source integration remains challenging. Systems that performed exceptionally well on previous benchmarks consisting solely of chat-based data exhibit a significant performance decline on ours. This indicates that integrating and analyzing multi-type data to infer personal life patterns poses considerable challenges, and future systems need to optimize the organization and reasoning of multi-source information.

F.2 Enabling Research Directions

The LifeBench dataset and benchmark enable systematic investigation of numerous research questions:

Memory consolidation and forgetting. How should systems decide what to remember, summarize, or forget? Our dataset presents challenges of a long time span and a large volume of long texts.

Personalized memory organization. Do optimal memory structures vary by persona (e.g., detail-oriented vs. big-picture individuals)? Our diverse personas enable persona-conditioned memory research.

Proactive memory retrieval. Beyond reactive QA, how can systems proactively surface relevant memories? Our rich event context supports evaluation of memory-triggered interventions.

Privacy-preserving memory. What information can be safely retained vs. summarized for privacy? Our synthetic data allows experimentation without real privacy risks.

F.3 Application Scenarios

Personal AI assistants. Agents that remember user preferences, track commitments, and provide personalized recommendations based on long-term behavioral patterns.

Digital health coaching. Systems that monitor health trends, correlate lifestyle factors with outcomes, and provide personalized interventions grounded in individual history.

Behavioral research. Controlled experimentation on human behavior modeling, habit formation, and lifestyle dynamics without privacy-sensitive real data.

Memory-augmented LLM development. Standardized benchmark for comparing memory architectures, retrieval strategies, and summarization approaches.

G Limitations and Ethical Considerations

G.1 Limitations

Synthetic Realism Boundaries. Despite our efforts to ensure coherence and diversity, synthetic data cannot fully capture the complexity, unpredictability, and rich contextual nuances of real human lives. Edge cases, rare events, and genuine emotional depth remain challenging to synthesize convincingly.

Coverage Gaps. Our dataset focuses on individuals in Chinese adult population with mobile phones. It does not represent all demographics, cultures, or lifestyles. Users with disabilities, non-standard living situations, or marginalized communities may be underrepresented.

Persona and Event Assumptions. Persona generation relies on anonymized survey data and social priors that may encode societal biases. Event synthesis uses LLM-based generation, which inherits biases from training data. We attempt to mitigate this through diversity controls and manual review, but systematic biases may remain.

Health Data Disclaimers. Health and sports records are *illustrative and non-clinical*. They should not be used for medical research, diagnosis, or treatment recommendations. The data represent simplified tracking signals, not comprehensive medical records.

Evaluation Limitations. Open-ended QA scoring relies on LLM judges, which may introduce subjectivity and variance. Scoring points are designed to be objective, but edge case interpretation may vary. Multiple-choice questions are more objective but limited in capturing nuanced reasoning.

G.2 Ethical Considerations

Privacy and Consent. Although our dataset is entirely synthetic, it represents realistic personal scenarios that could inadvertently resemble real individuals. We take precautions: (1) all personas are composites from anonymized aggregates, not copies of real people; (2) sensitive attributes are removed or transformed; (3) generation is seeded randomly to prevent reidentification.

Misuse Risks. Realistic personal data, even synthetic, could be misused for:

- Training surveillance or manipulation systems
- Generating deceptive personas for social engineering
- Creating plausible fake identities for fraud

We mitigate these risks by: (1) releasing data under a license prohibiting harmful use; (2) providing clear documentation of synthetic origins; (3) encouraging responsible use for research and beneficial applications only. However, we cannot fully prevent misuse. We rely on community norms and encourage dataset users to consider ethical implications of their work.

Health-Related Cautions. Health data in our dataset are *not medical-grade*. They represent simplified activity tracking for

lifestyle context. Any system trained on this data should *not* be deployed for clinical decision-making without extensive validation on real medical data and regulatory approval.

Informed Use. We commit to transparency:

- Comprehensive documentation of generation process, assumptions, and limitations
- Clear labeling as synthetic data in all releases
- Ongoing engagement with the research community to address emerging ethical concerns

G.3 Responsible Use Recommendations

The code repository for this study is released under the Apache License 2.0. We recommend that researchers and practitioners using LifeBench:

- (1) Clearly disclose use of synthetic data in publications and deployments
- (2) Validate findings on real data before drawing strong conclusions
- (3) Consider demographic and cultural limitations when generalizing results
- (4) Implement additional safeguards (bias audits, privacy protections) when building systems on this data
- (5) Adhere to ethical guidelines for AI research and deployment in their respective domains

H Dataset Schema Details

H.1 Persona Schema

```
{
  "name": "Yu_Xiaowei",
  "birth": "1986-03-14",
  "age": 35,
  "nationality": "Han",
  "home_address": {
    "province": "Hong_Kong_Special_Administrative_Region",
    "city": "Hong_Kong",
    "district": "Kowloon_City_District",
    "street_name": "Prince_Edward_Road_West",
    "street_number": "No._328"
  },
  "birth_place": {
    "province": "Hong_Kong_Special_Administrative_Region",
    "city": "Hong_Kong",
    "district": "Kowloon_City_District"
  },
  "gender": "Female",
  "education": "General_Senior_High_School",
  "job": "Insurance_Agent",
  "occupation": "AIA_Hong_Kong",
  "workplace": {
    "province": "Hong_Kong_Special_Administrative_Region",
    "city": "Hong_Kong",
    "district": "Wan_Chai_District",
```

```
    "street_name": "Gloucester_Road",
    "street_number": "AIA_Central_No._138"
  },
  "belief": "No_religious_belief",
  "salary": 300000.0,
  "body": {
    "height": 155,
    "weight": 55.0,
    "BMI": 22.89
  },
  "family": "First_marriage_with_spouse,_has_2_children_(one_son_and_one_daughter)",
  "personality": {
    "mbti": "ESFJ",
    "traits": [
      "Career-oriented",
      "Traditional",
      "Spiritual_satisfaction-oriented"
    ]
  },
  "hobbies": [
    "Yoga_and_fitness",
    "Marathon",
    "Gathering_with_friends",
    "Shopping",
    "Listening_to_music_at_home",
    "Doing_hand_embroidery,_focusing_on_traditional_Cantonese_embroidery_craft",
    "Keeping_pets_(cats)"
  ],
  "favorite_foods": [
    "Hong_Kong-style_milk_tea",
    "Pineapple_bun_with_butter",
    "Wonton_noodles",
    "Roasted_meat_platter"
  ],
  "memory_date": [
    "2010-05-20:_Wedding_Anniversary",
    "2015-08-15:_Eldest_son's_birthday"
  ],
  "aim": [
    "Promote_to_Regional_Sales_Manager_within_three_years",
    "Take_the_family_on_an_in-depth_trip_to_Japan_within_five_years",
    "Insist_on_completing_at_least_two_half_marathons_every_year"
  ],
}
```

"healthy_desc": "The_individual_is_in_good_overall_health_with_no_chronic_diseases._She_regards_exercise_as_part_of_her_lifestyle_and_regularly_takes_yoga_and_marathon_training,_but_her_exercise_mode_tends_to_be_goal-oriented_training_and_competition_participation_rather_than_daily_routine_activities._She_has_an_irregular_physical_examination_habit,_seeking_medical_treatment_about_once_a_year._She_has_good_personal_care_habits,_focusing_on_diet_and_intermittent_high-intensity_exercise_to_maintain_health.",

"lifestyle_desc": "Her_living_habits_are_centered_on_social_interaction_and_family,_with_frequent_leisure_activities._She_mainly_obtains_information_through_the_Internet_and_surfs_the_web_every_day._In_terms_of_media_use,_she_listens_to_music_at_home_every_day,_watches_TV_and_does_handcrafts_several_times_a_month._She_is_very_active_in_social_activities,_gathering_with_friends,_shopping_and_meeting_with_relatives_who_do_not_live_together_several_times_a_week,_but_rarely_participates_in_activities_such_as_watching_movies,_reading_books_or_visiting_cultural_events._She_is_away_from_home_for_11_to_20_nights_a_year_due_to_vacation_or_visiting_relatives,_showing_a_certain_frequency_of_travel._Her_lifestyle_is_pragmatic_and_focuses_on_interpersonal_relationships,_enjoying_interaction_and_relaxation_in_her_familiar_social_circle.",

"economic_desc": "The_family's_economic_situation_is_stable._Her_personal_annual_income_is_about_300,000_Hong_Kong_dollars,_her_spouse's_annual_income_is_about_100,000_Hong_Kong_dollars,_and_the_total_family_annual_income_is_about_400,000_Hong_Kong_dollars,_with_income_exceeding_expenditure._She_owns_a_property_and_a_family_car._In_terms_of_consumption_habits,_she_occasionally_uses_overdraft_methods_such_as_credit_cards._At_present,_the_family_has_no_financial_investment_activities_such_as_stocks,_funds_or_bonds,_with_a_conservative_financial_management_style,_focusing_more_on_the_safety_and_liquidity_of_assets,_and_investment_activities_may_be_concentrated_on_owner-occupied_real_estate.",

"work_desc": "She_is_an_insurance_agent_employed_by_a_private_enterprise,_working_about_40_hours_a_week_with_regular_day_shifts,_and_the_one-way_commute_time_is_about_60_minutes._Her_work_content_is_mainly_sales_and_customer_relationship_maintenance,_with_few_cases_of_heavy_physical_or_mental_labor._She_is_very_satisfied_with_her_current_job_and_holds_multiple_part-time_jobs_at_the_same_time,_which_may_involve_the_agency_of_different_insurance_products_or_customer_development,_showing_a_strong_career_drive_and_professional_ability.",

"experience_desc": "Born_in_Kowloon_City_District_of_Hong_Kong,_she_has_lived_there_all_the_time_and_has_never_moved_her_household_registration._After_graduating_from_senior_high_school_in_2005,_her_first_job_was_a_packaging_worker_(unskilled_worker)_in_a_factory,_employed_by_others._She_studied_insurance_knowledge_through_night_school,_obtained_the_insurance_agent_qualification_in_2010,_joined_AIA_Hong_Kong,_and_transformed_into_the_insurance_sales_industry._Relying_on_excellent_communication_skills_and_diligence,_she_gradually_built_her_own_customer_network,_established_a_family_and_career_in_Hong_Kong,_and_raised_two_children._Her_life_experience_reflects_a_career_leap_from_grassroots_labor_to_professional_sales,_as_well_as_a_life_trajectory_of_rooting_in_the_local_area_and_managing_a_family.",

"description": "Yu_Xiaowei_is_a_35-year-old_female_insurance_agent_in_Hong_Kong,_working_at_AIA_Hong_Kong._Born_and_raised_in_Hong_Kong,_she_has_a_senior_high_school_education_background._As_an_ESFJ_(Consul)_personality,_she_is_enthusiastic,_responsible_and_values_interpersonal_relationships,_with_values_inclined_to_be_career-oriented,_traditional_and_spiritual_satisfaction-oriented._She_is_married_with_two_children,_the_family's_annual_income_is_about_400,000_Hong_Kong_dollars,_with_a_stable_economic_situation,_owning_a_property_and_a_car,_and_a_conservative_investment_style._Yu_Xiaowei_is_in_good_health_with_no_chronic_diseases,_integrating_exercise_into_her_life_and_loving_yoga_and_challenging_marathons._Her_lifestyle_is_highly_social,_gathering_with_friends_and_relatives_and_shopping_frequently_every_week,_while_using_the_Internet_to_obtain_information_and_enjoy_music_every_day._At_work,_she_is_a_diligent_and_satisfied_insurance_salesperson_with_a_long_commute_time_but_good_at_time_management,_and_holds_multiple_part-time_jobs_to_expand_her_career._In_life,_she_often_organizes_family_gatherings_and_plans_to_participate_in_marathon_events_regularly._In_the_future,_she_hopes_to_be_promoted_to_regional_manager,_take_the_family_to_travel_to_Japan,_and_continue_her_running_hobby.",

"relation": [

```
[
  {
    "name": "Chen_Meiling",
    "relation": "Mother",
    "social_circle": "Family_Circle",
    "gender": "Female",
    "age": 62,
    "birth_date": "1962-08-22",
    "home_address": {
      "province": "Hong_Kong_Special_Administrative_Region",
      "city": "Hong_Kong",
      "district": "Kowloon_City_District",
      "street_name": "Prince_Edward_Road_West",
      "street_number": "No._328"
    },
    "birth_place": {
      "province": "Hong_Kong_Special_Administrative_Region",
      "city": "Hong_Kong",
      "district": "Kowloon_City_District"
    },
    "personality": "ISFJ",
    "economic_level": "Comfortable_Living",
```

```
"occupation": "Retired_Primary_School_Teacher",
"organization": "Kowloon_Tong_Christian_Primary_School_(Retired)",
"nickname": "Ma_Ma",
"relation_description": "Mother_of_Yu_Xiaowei,_who_has_been_living_with_her_daughter's_family_all_the_time_and_helps_take_care_of_her_grandchildren._The_mother-daughter_relationship_is_close,_and_Yu_Xiaowei_chats_with_her_mother_every_day_after_work,_sharing_work_experiences_and_interesting_stories_about_the_children._The_family_dinner_on_weekends_is_usually_cooked_by_the_mother,_who_prepares_traditional_Cantonese_cuisine._The_mother_is_an_important_emotional_support_for_Yu_Xiaowei,_and_is_always_the_first_person_to_talk_to_and_ask_for_help_when_she_is_under_work_pressure_or_needs_parenting_advice."
```

```
},
{
  "name": "Yu_Guoqiang",
  "relation": "Father",
  "social_circle": "Family_Circle",
  "gender": "Male",
  "age": 65,
  "birth_date": "1959-12-05",
  "home_address": {
    "province": "Hong_Kong_Special_Administrative_Region",
    "city": "Hong_Kong",
    "district": "Kowloon_City_District",
    "street_name": "Prince_Edward_Road_West",
    "street_number": "No._328"
  },
  "birth_place": {
    "province": "Hong_Kong_Special_Administrative_Region",
    "city": "Hong_Kong",
    "district": "Kowloon_City_District"
  },
  "personality": "ISTJ",
  "economic_level": "Comfortable_Living",
  "occupation": "Retired_Bus_Driver",
  "organization": "Kowloon_Motor_Bus_Co.,_Ltd._(Retired)",
  "nickname": "Lao_Dou",
```



```

    "relation_description": "Father_of_Yu_Xiaowei, who lives with his wife after retirement and likes to play chess and read newspapers in the park on weekdays. The father-daughter relationship is traditional and implicit, with communication mostly centered on family affairs and grandchildren's education. Yu_Xiaowei regularly gives her father pocket money and goes out for morning tea together on important days such as Father's Day. The father occasionally expresses worry about Yu_Xiaowei's marathon hobby, but mostly supports her silently and sometimes waits for her at the finish line."
  },
  .....
]
]
}

[
{
  "name": "Kowloon_City_Plaza_Shopping_Mall",
  "location": "114.192980,22.328095",
  "formatted_address": "No. 128 Carpenter Road, Kowloon_City_District, Hong_Kong_Special_Administrative_Region",
  "city": [],
  "district": "Kowloon_City_District",
  "streetName": "Carpenter_Road",
  "streetNumber": "No. 128",
  "description": "Yu_Xiaowei's permanent residence, located in Kowloon_City_District_of_Hong_Kong, where she lives with her family on a daily basis."
},
{
  "name": "Wan_Chai_Immigration_Tower",
  "location": "114.177960,22.277046",
  "formatted_address": "No. 7 Gloucester Road, Wan_Chai_District, Hong_Kong_Special_Administrative_Region",
  "city": [],
  "district": "Wan_Chai_District",
  "streetName": "Gloucester_Road",
  "streetNumber": "No. 7",
  "description": "Yu_Xiaowei's workplace, located in Wan_Chai_District_of_Hong_Kong, her daily office as an insurance agent."
},
  .....
]

```

H.2 Event Schema

```

[
  {
    "event_id": "4398",
    "name": "On-site_Inspection_and_Appreciation_Dinner_Restaurant_Selection",
    "date": [
      "2025-11-04_20:40:00_to_2025-11-04_21:40:00"
    ],
    "type": "Career",
    "description": "Yu_Xiaowei informed her family that she needed to go out to inspect restaurants for the appreciation dinner. She took the MTR from home to Wan_Chai, and conducted on-site inspections of two mid-to-high-end Cantonese restaurants around Lockhart_Road, focusing on the private room environment, privacy and menu. She initially selected an elegantly decorated one and made a reservation for a food tasting the next noon.",
    "participant": [
      {
        "name": "Yu_Xiaowei",
        "relation": "Herself"
      }
    ],
    "location": "Lockhart_Road_Area, Wan_Chai_District, Hong_Kong"
  },
  {
    "event_id": "4399",
    "name": "Overtime_Processing_of_Documents_for_a_Client's_Urgent_Request",
    "date": [
      "2025-11-04_21:40:00_to_2025-11-04_22:30:00"
    ],
    "type": "Unexpected_Events",
    "description": "After returning home, Yu_Xiaowei received an urgent WeChat message from an important client, who needed to adjust the beneficiary clause of an insurance policy. She immediately turned on her laptop, connected to the company system, carefully checked the terms, communicated and confirmed repeatedly with the client, and prepared the revised documents. Her planned post-exercise relaxation time was occupied.",
    "participant": [
      {
        "name": "Yu_Xiaowei",
        "relation": "Herself"
      },
      {
        "name": "Important_Client",

```

```

    "relation": "Client"
  }
],
"location": "Study_at_Home,_Kowloon_City_Plaza
,_No._128_Carpenter_Road,_Kowloon_City_
District,_Hong_Kong_Special_Administrative
_Region"
},
.....
]

```

H.3 Phone Data Schema

- **Contact:** User-associated contacts (family, friends, colleagues) with attributes of name, phone number, relationship and email address.
- **Call:** Call records including caller/callee information, timestamp, duration and call type (incoming/outgoing/missed), with filtering based on contact relationships and event triggers.
- **SMS:** Short message records capturing real-time text communication interactions.
- **Calendar:** Scheduled events, appointments and reminders that are synced with daily life activities.
- **Agent Chat:** Conversation records with AI assistants, covering scenario-based interactions such as scheduling, information seeking and casual conversation.
- **Photo:** Image records with timestamps and textual descriptions of visual content.
- **Note:** Personal content including memos, to-do items and reflective notes.
- **Push:** App push notifications from social media platforms, news channels and various service applications.

```

{
  "event_id": "0",
  "date": "2025-01-01",
  "type": "agent_chat",
  "conversation": {
    "turn_1": {
      "user": {
        "action": "topic_query",
        "content": "Hello!_This_afternoon,_my_
          whole_family_played_Monopoly_
          together._I_took_the_chance_to_
          explain_some_simple_financial_
          concepts_like_'buying_land'_and_'
          collecting_rent'_to_my_8-year-old_
          Zixuan_and_5yearold_Yuxin._They_had_
          a_great_time_and_listened_carefully._
          This_got_me_thinking_besides_board_
          games,_are_there_any_other_fun_and_
          effective_ways_to_cultivate_
          financial_literacy_suitable_for_
          their_age_group?"
      },
      "assistant": {
        "action": "need_inference",

```

```

        "content": "It_sounds_like_you_are_a_
          mother_who_attaches_great_importance_
          to_family_education,_hoping_to_
          integrate_financial_knowledge_into_
          life_in_a_way_that_children_love,_
          right?_Are_you_looking_for_some_
          specific_and_operable_financial_
          education_ideas_or_activity_
          suggestions_targeted_at_your_8-year-
          old_and_5-year-old_children?"
      }
    },
    "turn_2": {
      "user": {
        "action": "need_confirmation",
        "content": "Yes,_exactly!_I_work_as_an_
          insurance_agent_myself_and_have_some_
          understanding_of_financial_planning_
          ,_but_I_want_to_know_how_to_teach_
          children_in_a_language_and_way_they_
          can_understand._For_example,_how_to_
          help_them_understand_the_meaning_of_
          'savings',_or_how_to_distinguish_
          between_'wants'_and_'needs'?_It_
          would_be_best_if_it_can_be_combined_
          with_our_daily_life_in_Hong_Kong."
      },
      "assistant": {
        "action": "solution_proposal",
        "content": "Understood._Combining_your_
          children's_ages_and_daily_life_
          scenarios_in_Hong_Kong,_I_have_a_few_
          ideas._First,_you_can_start_with_
          pocket_money_management._Give_Zixuan_
          a_three-compartment_savings_jar,_
          labeled_'Save',_'Spend',_and_'Share'_
          respectively,_and_let_him_allocate_
          the_money_himself._For_Yuxin,_you_
          can_use_a_more_intuitive_transparent_
          jar_so_she_can_watch_the_coins_pile_
          up_and_establish_a_preliminary_
          concept_of_'saving_money'."
      }
    },
    "turn_3": {
      "user": {
        "action": "solution_feedback",
        "content": "This_savings_jar_idea_is_
          great!_It's_clearly_categorized_and_
          simple_to_operate._What_about_the_
          concept_of_'wants_vs._needs'?_My_own_
          investment_style_is_relatively_
          conservative,_and_I_also_hope_to_
          subtly_pass_on_a_prudent_financial_
          concept_to_them."
      },
      "assistant": {

```

```

    "action": "solution_discussion",
    "content": "You've thought this through very thoroughly. For 'wants vs. needs', you can play the 'supermarket shopping game'. Before going to the supermarket, make a list with the children and distinguish between what is 'needed' (such as milk, bread) and what is 'wanted' (such as new toys, snacks). Within the budget, allow them to use the money in their 'Spend' compartment to buy one 'wanted' item, this will let them personally experience choice and trade off. Your conservative and prudent style can be passed on perfectly by emphasizing the importance of the 'Save' section in the savings jar, for example, tell them that saving money is for more important goals, like the in-depth family trip to Japan you mentioned earlier."
  },
  "turn_4": {
    "user": {
      "action": "information_request",
      "content": "Connecting family goals (like travel) with savings is such a great angle! It will give them more motivation. Besides these daily practices, are there any picture books or books about money or financial management suitable for parent-child reading that you can recommend? Preferably in Chinese."
    },
    "assistant": {
      "action": "information_provision",
      "content": "Yes, there are. For 5-year-old Yuxin, you can read the 'Little Rabbit Learns to Spend Money' series of picture books, which explain the concepts of 'spending, saving, and earning money' through lovely stories. For 8-year-old Zixuan, you can try 'Money Dog' (Juvenile Edition), it's highly story-driven and can inspire children to think about the relationship between dreams, goals, and actions. These books should be available in libraries or bookstores in Hong Kong."
    }
  },
  "turn_5": {
    "user": {
      "action": "solution_feedback",
      "content": "Thank you so much! From the specific savings jar method and life games to the picture book recommendations, everything is very practical and perfectly fits the personalities and ages of my two children. I plan to try this right away, starting with the compartmentalized savings jar and the supermarket game. Playing Monopoly this afternoon was a great start, and now I feel more directionally clear."
    },
    "assistant": {
      "action": "talk",
      "content": "You're welcome, Ms. Yu. Financial literacy education is originally a pleasant parent-child journey integrated into life. You've already taken a wonderful first step. Remember, the most important thing is the interaction and guidance in the process, not teaching all concepts at once. Enjoy the time exploring and growing together with Zixuan and Yuxin. If you have new ideas or questions during subsequent practice, feel free to come and chat anytime."
    }
  },
  "phone_id": 0
},
{
  "type": "calendar",
  "event_id": "4284",
  "title": "Collect Physical Examination Report at Hong Kong Sanatorium and Hospital",
  "description": "Collect the comprehensive physical examination report and have an in-person consultation with Dr. Zhou Wenbin, Time: 2025-10-28 14:50:00, Location: Department of Family Medicine, Hong Kong Sanatorium and Hospital, Shan Tsuen Road, Wan Chai District, Hong Kong, Source: Hospital appointment",
  "start_time": "2025-10-28 14:50:00",
  "end_time": "2025-10-28 15:40:00",
  "datetime": "2025-10-27 20:15:00",

```

```

    "summarized_info": "Yu_Xiaowei_set_a_schedule_to_collect_the_physical_examination_report_at_Hong_Kong_Sanatorium_and_Hospital,_Time:_2025-10-28_14:50:00,_Location:_Department_of_Family_Medicine,_Hong_Kong_Sanatorium_and_Hospital,_Shan_Tsuen_Road,_Wan_Chai_District,_Hong_Kong,_Source:_Hospital_appointment",
    "phone_id": 0
}

{
  "type": "call",
  "event_id": "1",
  "phoneNumber": "85291234573",
  "contactName": "Lin_Jiaxin",
  "datetime": "2025-01-01_08:35:00",
  "datetime_end": "2025-01-01_08:42:30",
  "direction": 1,
  "call_result": "Connected",
  "phone_id": 0
}

{
  "name": "Chen_Meiling",
  "relation": "Mother",
  "gender": "Female",
  "nickname": "Ma_Ma",
  "phoneNumber": "85291234567",
  "personalEmail": "chenmeiling_hk@163.com",
  "workEmail": "chenmeiling@kltxdxx.com",
  "idNumber": "810101196208224028",
  "phone_id": "1"
}

{
  "type": "note",
  "event_id": "4409",
  "title": "Key_Points_of_Son's_Parent-Teacher_Day_Interview",

  "content": "I._Teacher's_Feedback:_1._Made_significant_progress_in_Math;_2._Needs_to_improve_Chinese_handwriting;_II._Follow-up_Actions:_1._Arrange_extra_calligraphy_practice_sheets;_2._Keep_an_eye_on_weekend_parent-child_activity_information",
  "datetime": "2025-11-05_16:35:00",

  "summarized_info": "Yu_Xiaowei_recorded_the_key_points_of_her_son's_Parent-Teacher_Day_interview,_including:_I._Teacher's_Feedback:_1._Made_significant_progress_in_Math;_2._Needs_to_improve_Chinese_handwriting;_II._Follow-up_Actions:_1._Arrange_extra_calligraphy_practice_sheets;_2._Keep_an_eye_on_weekend_parent-child_activity_information",
  "phone_id": 2
}

{
  "event_id": "3",
  "type": "photo",
  "caption": "Yu_Xiaowei_took_a_selfie_with_her_colleague_Su_Lishan_in_a_cha_chaan_teng_in_Wan_Chai._Pineapple_buns_with_butter_and_Hong_Kong-style_milk_tea_are_placed_in_front_of_them,_with_a_bustling_restaurant_environment_in_the_background.",
  "title": "IMG_20231001_121045",
  "datetime": "2023-10-01_12:10:45",
  "location": {
    "province": "Hong_Kong_Special_Administrative_Region",
    "city": "Hong_Kong",
    "district": "Wan_Chai_District",
    "streetName": "Lockhart_Road",
    "streetNumber": "No._XX",
    "poi": "Kam_Fung_Cha_Chaan_Teng"
  },
  "faceRecognition": [
    "Yu_Xiaowei",
    "Su_Lishan"
  ],
  "imageTag": [
    "Colleague_Gathering",
    "Selfie",
    "Pineapple_Bun_with_Butter",
    "Hong_Kong-style_Milk_Tea",
    "Cha_Chaan_Teng",
    "Lunch_Break"
  ],
  "ocrText": "None",
  "shoot_mode": "Portrait",
  "image_size": "3024x4032",
  "summarized_info": "Yu_Xiaowei_and_her_colleague_Su_Lishan_took_a_selfie_during_their_lunch_break_gathering._The_main_content_includes_Kam_Fung_Cha_Chaan_Teng_in_Wan_Chai,_pineapple_buns_with_butter_and_Hong_Kong-style_milk_tea_in_front_of_them,_and_the_bustling_restaurant_environment.",
  "phone_id": 4
}

```



```
{
  "type": "push",
  "event_id": "5",
  "title": "Alipay: Investment Earnings Credited",
  "content": "[Alipay] Your Yu'E Bao earnings of HK$15.23 for yesterday have been credited. Please check in the app.",
  "datetime": "2023-10-01 07:05:00",
  "source": "Alipay",
  "push_status": "Unread",
  "jump_path": "Alipay_to_My_Account_to_Yu'E_Bao_to_Earnings_Details",
  "summarized_info": "I received a notification from Alipay that HK$15.23 of yesterday's Yu'E Bao earnings has been credited to my account.",
  "phone_id": 0
}

{
  "type": "sms",
  "event_id": "24",
  "message_content": "Xin Xin, are you free this weekend? I've been so tired lately and want to ask you to relax at that Japanese-style spa in Tsim Sha Tsui. I'll check out what packages they have available.",
  "contactName": "Lin Jiaxin",
  "phoneNumber": "85291234573",
  "datetime": "2025-01-02 13:25:00",
  "message_type": "Sent",
  "phone_id": 7
}
```

H.4 Health Data Schema

```
{
  "Date": "2025-01-01",
  "City": "Hong Kong",
  "Daily_Activities": {
    "Steps": "8560_steps",
    "Distance": "6.2_km",
    "Calories_Burned": "420_kcal",
    "Exercise_Duration": "45_minutes",
    "Active_Hours": "8_hours"
  },
  "Running": {
    "Sport_Type": "Indoor_Running",
    "Sport_Time": "2025/01/01/07:30:00-2025/01/01/07:45:00",
    "Weather": "Sunny",
    "Distance": "2.5_km",
    "Average_Heart_Rate": "135_bpm",
    "Average_Cadence": "165_spm",
    "Elevation_Gain": "0_m",
```

```
    "Elevation_Loss": "0_m",
    "Average_Pace": "06:00_min/km",
    "Best_Pace": "05:45_min/km",
    "Total_Steps": "3125_steps",
    "Calories_Burned": "150_kcal"
  },
  "Cycling": {
    "Sport_Type": "Outdoor_Cycling",
    "Sport_Time": "2025/01/01/00:00:00-2025/01/01/00:00:00",
    "Weather": "Sunny",
    "Distance": "0_km",
    "Average_Speed": "0_km/h",
    "Average_Heart_Rate": "0_bpm",
    "Average_Cadence": "0_rpm",
    "Average_Power": "0_W",
    "Best_Speed": "0_km/h",
    "Max_Cadence": "0_rpm",
    "Calories_Burned": "0_kcal"
  },
  "Walking": {
    "Sport_Type": "Outdoor_Walking",
    "Sport_Time": "2025/01/01/10:50:00-2025/01/01/11:15:00",
    "Weather": "Sunny",
    "Distance": "1.2_km",
    "Average_Heart_Rate": "95_bpm",
    "Average_Cadence": "110_spm",
    "Total_Steps": "1600_steps",
    "Average_Pace": "12:30_min/km",
    "Best_Pace": "11:00_min/km",
    "Calories_Burned": "65_kcal"
  },
  "Sleep": {
    "Bedtime": "22:45:00",
    "Wake-up_Time": "08:30:00",
    "Total_Sleep_Duration": "585_minutes",
    "Light_Sleep_Duration": "280_minutes",
    "Deep_Sleep_Duration": "180_minutes",
    "REM_Sleep_Duration": "125_minutes",
    "Awake_Duration": "25_minutes",
    "Number_of_Awakenings": "3_times",
    "Deep_Sleep_Continuity_Score": "85_points",
    "Sleep_Score": "88_points",
    "Nap_Duration": "0_minutes"
  },
  "Heart_Rate_Statistics": {
    "Average_Heart_Rate": "78_bpm",
    "Average_Resting_Heart_Rate": "62_bpm",
    "Heart_Rate_Variability": "45_ms"
  },
  "Body_Temperature_Statistics": {
    "Average_Body_Temperature": "36.5"
  },
  "Blood_Glucose_Statistics": {
```

```

    "Average_Blood_Glucose_Level": "5.2_mmol/L"
  },
  "Body_Weight": {
    "Weight": "55.0_kg"
  },
  "Stress": {
    "Stress_Score": "25_points"
  },
  "Diet_Record": {
    "Calories_Consumed": "1850_kcal"
  },
  "User_Interaction_Events": [
    {
      "Time": "2025/01/01_07:30",
      "Description": "Started_indoor_running_
        training_with_a_smart_watch"
    },
    {
      "Time": "2025/01/01_10:50",
      "Description": "Started_recording_walking_
        activity_for_grocery_shopping"
    },
    {
      "Time": "2025/01/01_21:30",
      "Description": "Checked_the_summary_of_
        today's_steps_and_activity_calorie_
        burn"
    },
    {
      "Time": "2025/01/01_22:15",
      "Description": "Set_tomorrow's_fitness_
        goal:_30_minutes_of_morning_yoga"
    }
  ],

  "summarized_info": "Today_is_New_Year's_Day_
    2025. _Yu_Xiaowei_woke_up_at_08:30_and_went_
    _to_bed_at_23:00, _maintaining_a_body_
    weight_of_55.0_kg. _As_a_marathon_
    enthusiast, _she_kept_up_her_habit_of_a_2.5_
    _km_morning_indoor_run. _Her_day_was_mainly_
    _filled_with_family_activities, _including_
    collaborating_with_her_husband_to_refresh_
    the_living_room_decor, _grocery_shopping, _
    and_preparing_family_afternoon_tea_and_
    dinnerlight_housework_in_total. _She_
    accumulated_8560_steps, _covering_6.2_km_
    and_burning_420_kcal. _Her_sleep_quality_
    was_good_with_a_sleep_score_of_88, _and_her_
    _stress_level_was_low.",

  "phone_id": 0
}

```

H.5 Monthly Summary Schema

The monthly summary provides a retrospective analysis of the user's health metrics, lifestyle patterns, and significant life events including both routine and anomalous. As a rich repository of personal data, it captures a comprehensive digital trajectory and serves as a reference (groundtruth) for uncovering latent activities hidden within mobile artifacts. While this data remains untapped by the evaluated memory systems, it represents a critical frontier for future research.

Your monthly summary is as follows:

Monthly Overview

- **Basic Information:** Yu Xiaowei, Female, 35 years old, Occupation: Insurance Salesperson (Financial Insurance Industry).
- **Total Exercise Volume:**
 - Running: 11 sessions / Total distance approximately 104 km / Total duration approximately 660 minutes (estimated).
 - Yoga: 4 sessions / Total duration 240 minutes.
 - Interval Run Training: 2 sessions / Total duration approximately 120 minutes (estimated).
 - Monthly Total Exercise: 17 sessions / Total duration approximately 1020 minutes (17 hours). Exercise performance steadily improved from easy runs at the beginning of the month to systematic training including tempo runs, long run workouts and interval runs by the end of the month.
- **Health Status Summary:**
 - **Health Score Change:** Based on data inference, status was stable at the beginning of the month (estimated score 75/100), fluctuated slightly in the middle of the month due to increased training load and work pressure (estimated score 70/100), and rebounded overall at the end of the month through adjustment and recovery with high-quality training completed (estimated score 78/100). The monthly average score is estimated to be 74/100.
 - **Core Tags:** "Launch of systematic marathon preparation", "Stress management under work-family balance", "Increased cross-training in accordance with doctor's advice".
 - **Overall Health Rating: Good.** Exercise performance improved steadily, family and social activities were active, but certain dual pressure from work and training was endured in the middle of the month, requiring attention to recovery.
- **Key Events:**
 - **Sports Milestone:** January 12: Co-determined and successfully registered for the 2025 Hong Kong Green Half Marathon (April) with running friend Shen Peiwen, officially launching the season.
 - **Health Indicator Alerts:** January 14: First experienced mild soreness in the right calf muscle after an 8km tempo run, marking the entry into the intensity enhancement phase of training.
 - **Major Life Changes:** January 28-31: Entered the Spring Festival Holiday, with life rhythm switching from work mode to family gathering and leisure mode.
- **Goal Completion Status:**
 - **Monthly Goals:**
 - (1) **Total running distance greater than 80km** (lay the foundation for half marathon preparation).
 - (2) **Increase low-impact exercise** (in accordance with the doctor's advice from the physical examination on January 10).
 - (3) **Complete systematic learning of Unit 1 of Mandarin.**
 - **Completion Degree:**
 - (1) Completed approximately 104km of running, **130% overachieved**.
 - (2) Yoga was added 4 times (unplanned), cross-training goal **achieved**.
 - (3) Learning, practice and testing of Unit 1 of Mandarin were all completed, **achieved**.
 - **Unfinished Reasons:** All sports and learning goals set for this month were **fully achieved**.
- **Core Changes from Last Month:**

- **Exercise Volume:** This month's running sessions (11 vs 7 last month) increased by 57%, and total distance (104km vs an estimated 45km last month) increased significantly by 131%. Regular yoga (4 sessions) and interval run training (2 sessions) were newly added, making the exercise structure more diverse.
- **Health Indicators:** **Weight:** 55.0kg at the beginning of the month, 55.2kg at the end of the month, with an overall fluctuation range of 0.9kg (54.7kg - 55.8kg). Weight slightly rebounded at the end of the month due to Spring Festival gatherings, but remained stable in the range of 54.8-55.2kg during the training period. **Sleep Duration:** May have greater volatility than last month due to overtime work, social activities and holidays.
- **Life Rhythm: Work and Rest:** Stability of bedtime decreased due to overtime work (e.g., January 2, 22), social activities (e.g., January 11, 27) and holidays. **Business Trips/Travel:** No long-distance business trips this month. **Screen Time:** Average daily usage may have increased due to online learning, work and Spring Festival social activities.
- **Core Focus Areas:** This month's high-frequency concerns are **Balancing marathon training load and physical responses (especially calf soreness)**, **Ensuring Mandarin learning progress with fragmented time**, and **Maintaining exercise habits amid intensive work and family affairs**.

Monthly Review 2.1 Sports Review

- **Exercise Overview:**
 - A total of 17 structured exercise sessions were conducted throughout the month. Exercise type proportion: Running (64.7%), Yoga (23.5%), Interval Run Training (11.8%).
 - Exercise intensity distribution (estimated): Zone 2 moderate intensity (easy runs, long run workouts) accounted for approximately 70%, with a cumulative duration of about 714 minutes; Zone 3 high intensity (tempo runs, interval runs) accounted for approximately 30%, with a cumulative duration of about 306 minutes.
- **Sports Statistics and Best Performances:**
 - **Running:**
 - * Total Distance: Approximately 104 km.
 - * Average Pace: Based on training data, estimated to be approximately 6'20"/km.
 - * Single Longest Distance: January 19: 15km long run workout.
 - * Fastest Pace: Reached Zone 3 high intensity interval in tempo runs on January 14, 16 and interval runs on January 23, 30, with the fastest pace in interval run segments (estimated to be within 5'40"/km).
 - * Average/Maximum Heart Rate: Data not directly provided, but based on training descriptions, heart rate reached the upper limit of the aerobic endurance zone and above during tempo runs and interval runs.
 - **Yoga:**
 - * Total Duration: 240 minutes (4 sessions, 60 minutes each).
 - * Single Longest Duration: 60 minutes.
 - * Core Effect: Significantly relieved calf muscle tension after running (January 15), and balanced running load as a low-impact cross-training.
 - **Interval Run Training:**
 - * Total Sessions: 2 sessions (January 23, 30).
 - * Core Content: Conducted high-intensity interval runs under the guidance of a coach, focusing on improving speed and running efficiency.
- **Exercise Habit Preferences:**
 - **Weekly Frequency:** High-frequency exercise days are **Saturday and Sunday** (3-4 exercise records each), followed by **Tuesday and Thursday** (morning running days).
 - **Daily Time Periods:** High-frequency exercise periods in a day are **6:00-7:30 in the early morning** (morning running, accounting for approximately 65%), followed by **19:00-20:30 in the evening** (yoga/-training sessions, accounting for approximately 35%).
 - **Exercise Locations:** Mainly running in the **neighborhood near home** (11 sessions), taking yoga classes at the **community center** (4

sessions), and taking interval run classes at the **designated training ground** (2 sessions).

- **Exercise Consistency:** The longest consecutive exercise days are **3 days** (e.g., January 14-16). There was **1 obvious exercise gap** this month, from January 20-22 (3 days in total), due to "mild cold and physical discomfort" and "overtime work to handle customer urgent needs".

2.2 Health Review

- **Sleep:**
 - Average Daily Sleep Duration (estimated): Approximately 6.5 hours on workdays (wake up 6:30-7:00, go to bed 22:30-23:30), approximately 7.5 hours on weekends (wake up 7:00-8:30, go to bed around 23:00).
 - Sleep Stage Distribution: Data not provided, but inferred from descriptions of "deep sleep feeling" (e.g., after SPA, yoga) and stress periods (after overtime work), the proportion of deep sleep may reach more than 25% on relaxation days and less than 20% on high-pressure days.
 - Bedtime/Wake-up Stability: On workdays, the fluctuation range of bedtime is approximately 60 minutes (22:30-23:30), and the fluctuation range of wake-up time is approximately 30 minutes (6:00-7:00); greater fluctuation on weekends. The latest bedtime record is 23:50 (January 18, 27).
 - Average Daily Sleep Interruptions: Data not provided, but based on the family situation with young children, it is estimated that there may be 1-2 short interruptions at night.
- **Physiological Indicators:**
 - **Weight:** 55.0kg at the beginning of the month, 55.2kg at the end of the month. Weekly averages: 55.1kg in Week 1, 55.1kg in Week 2, 55.2kg in Week 3, 55.5kg in Week 4 (affected by Spring Festival). Maximum fluctuation +0.9kg (reached 55.4kg after family feast on January 12, reached the monthly peak of 55.8kg on January 29 during Spring Festival). Overall trend: Steady and slightly declining during the training period (early to mid-month), significantly rising during holidays, and quickly rebounding after holidays.
 - **Body Fat Percentage:** Data not provided, but combined with weight changes and increased exercise volume, it is estimated that there may be a decrease of approximately 0.5% at the end of the month compared to the beginning (increased muscle mass offsets part of the fat).
 - **Resting Heart Rate:** Data not provided. Inferred based on exercise performance and stress events: Approximately 58-62 beats per minute on normal days; may rise to 65-68 beats per minute on the day after high-intensity training or after overtime and staying up late (e.g., January 22).
 - **HRV:** Data not provided. Trend inference: At a relatively high level after high-quality sleep and yoga practice (e.g., January 15); at a relatively low level during periods of continuous work pressure and insufficient sleep (e.g., January 20-22).
- **Mental Health:**
 - **Stress:** Estimated average daily stress score 5/10. Weekly averages: 4/10 in Weeks 1 and 2, rose to 6/10 in Week 3 due to urgent work tasks and performance pressure, and dropped to 3/10 in Week 4 during the Spring Festival Holiday. High-pressure periods: Afternoon of January 20 (customer complaint), January 22 (urgent overtime), January 24 (brief family quarrel caused by performance pressure).
 - **Mood:** Mainstream mood tags are **Fulfilling, Satisfied, Expectant**. Mood fluctuation periods: Felt "emotional fluctuation" due to sudden work problems on January 20; felt "slight emotional ups and downs" due to pressure on January 24; both were alleviated through social interaction (talking to best friends), exercise (yoga) and family communication.
 - **Health Score:** As mentioned in the overview, the estimated average daily score is 74. The lowest estimated score is 70 (high-pressure period in Week 3), and the highest estimated score is 80 (after completing the long run on January 19 and after the happy family day on January 26).

2.3 Life Rhythm Review

- **Work and Rest Habits:**

- Average Daily Bedtime: Approximately 23:05 on workdays, approximately 23:15 on weekends.
- Average Daily Wake-up Time: Approximately 6:40 on workdays, approximately 8:00 on weekends.
- Workday vs Weekend Difference: On average, weekends are **10 minutes later to bed** and **80 minutes later to wake up** than workdays, with a total sleep duration of approximately 1 hour more.
- Irregular Schedule Days: There were **8 days** this month with bedtime later than 23:30, due to overtime work, social gatherings, Spring Festival activities, etc.
- **Commuting/Activities:**
 - Average Daily Non-Exercise Steps: Data not provided, but based on the occupation of an insurance agent (part of the time out to meet customers) and an active lifestyle, it is estimated to be approximately 8,000 steps per day. Steps may be higher during the Spring Festival Holiday.
 - Commuting Method/Duration: Mainly take the MTR, approximately 60 minutes one way, with an average daily commuting duration of approximately 2 hours. Commuting time is the core period for Mandarin learning.
 - Business Trips/Travel: No work business trips this month.
- **Daily Life:**
 - Screen Time: Estimated average daily total duration is approximately 6-8 hours, of which approximately 4-5 hours are work-related (customer communication, product learning), and approximately 2-3 hours are entertainment and social interaction (WeChat, videos). The highest daily duration may exceed 10 hours (e.g., overtime day on January 22).
 - Eating Habits: High regularity of three meals a day, approximately 25 days of regular meals. The diet structure is biased towards traditional Hong Kong style (high carbohydrates, moderate protein), with high-calorie gatherings on weekends and festivals (e.g., January 4, 12, 28). Fewer takeout meals (estimated < 5 times), mainly home-cooked meals. Data on average daily water intake is not provided, but based on health awareness, basic intake can be guaranteed.

Monthly Trend Analysis 3.1 Sports Trend

- **Weekly Exercise Volume Changes:**
 - **Week 1 (January 1-5):** Low total exercise volume (only 1 5km run), in the holiday recovery and plan formulation period.
 - **Week 2 (January 6-12):** Significant increase in exercise volume, 3 running sessions (19km total), 1 yoga session. Month-on-month growth of approximately 300%. Reason: Restored the morning running habit, launched systematic training, and completed the marathon registration.
 - **Week 3 (January 13-19):** Peak exercise volume, 4 running sessions (41km total, including 15km long run), 2 yoga sessions. Running distance increased by 116% month-on-month. Reason: Officially entered the marathon preparation cycle, with both training intensity and volume increasing.
 - **Week 4 (January 20-26) and Spring Festival Week (January 27-31):** Exercise volume first decreased then increased. There was a gap at the beginning of Week 4 due to cold and work, and training resumed in the second half of Week 4 and Spring Festival Week, completing 4 running sessions (39km total), 1 yoga session, and 2 interval run sessions. The overall exercise rhythm was affected by the holiday but remained consistent.
- **Training Cycle Peaks and Valleys:**
 - **Peak Week:** Week 3 (January 13-19), characterized by the largest running volume (41km), completion of the first long run workout (15km), and exercise performance and confidence reaching a high point.
 - **Valley Period:** Early Week 4 (January 20-22), with 3 consecutive days of no exercise due to physical discomfort (cold) and work pressure, which is an obvious monthly valley.
- **Exercise Performance Changes:**
 - **Running Pace/Heart Rate Trend:** From easy runs (mainly Zone 2) in Weeks 1-2, to adding tempo runs and long runs (mixed Zone 2-Zone 3) in Week 3, to adding interval runs (Zone 3 peak) in Week 4

and the end of the month. On the basis of stable aerobic capacity, high pace capacity received targeted stimulation. Data shows that with the progress of training, the subjective sense of fatigue in completing the same distance (e.g., 10km) is reduced.

3.2 Health Trend

• Sleep Trend:

- **Duration and Quality:** Sleep was relatively regular and sufficient in Weeks 1-2. In Week 3, sleep duration was compressed and quality may have decreased due to work pressure (January 20) and urgent overtime (January 22). During the Spring Festival Holiday in Week 4, sleep duration was sufficient but work and rest regularity was disrupted (going to bed late and waking up late). The fluctuation of deep sleep proportion is negatively correlated with stress events and exercise recovery needs.
- **Sleep Continuity:** Throughout the month, sleep was affected by work, social activities and family (children), with few nights of complete uninterrupted sleep, but no severe insomnia was reported.

• Physiological Health Trend:

- **Weight Curve:** Showing a "stepwise" fluctuation. Weight was stable and slightly declining during training weeks (Weeks 2, 3) (54.8-55.2kg). It immediately increased by 0.3-0.6kg after every social/family feast (Weekends of Weeks 2, 4), and quickly rebounded in subsequent training days. Indicating active metabolism but sensitivity to diet.
- **Resting Heart Rate and HRV Correlation:** Inferred that HRV was higher and resting heart rate was lower on days such as January 15 (after yoga) and January 19 (sufficient rest after long run). HRV was lower and resting heart rate was higher on days such as January 22 (after overtime and staying up late) and January 24 (after emotional stress).

• Mental Health Trend:

- **Stress and Mood Curve:** Stress score is highly synchronized with work intensity. Week 3 was an obvious high-pressure period, accompanied by brief emotional fluctuations. Exercise (especially social running with friends and yoga) and in-depth social interaction (chatting with best friends) are effective stress release valves. After each in-depth social interaction or high-quality training, emotional and stress scores improved significantly.

• Injury Impact:

- January 14: Experienced mild soreness in the right calf muscle. Direct impact: Prompted targeted stretching in the yoga class on January 15, and more attention to warm-up in subsequent training. Did not cause training interruption, but marked the body's entry into the intensity adaptation phase.

3.3 Life Rhythm Trend

- **Work and Rest and Mental State:** Data shows that on the day after sufficient sleep (>7 hours) and stable bedtime (before 23:00) (e.g., January 7, 21), Yu Xiaowei's morning running pace was more stable, and the subjective sense of fatigue in completing tempo runs or long runs was reduced. On the contrary, after a single day of excessive screen time (>8 hours) or handling urgent work in the evening, bedtime was delayed by an average of 40 minutes on the same day, and heart rate was higher and recovery was slower during exercise the next day.
- **Work/Life Rhythm Changes:**
 - **Work Intensity Changes:** Week 3 (especially January 20-22) had the highest work intensity, involving handling customer complaints and urgent case overtime, directly compressing exercise and sleep time.
 - **Life Changes:** Week 4 entered the Spring Festival Holiday, with life focus completely shifting to family reunion and leisure. Exercise shifted from "training mode" to "maintenance mode", with increased sleep duration but decreased regularity.

Personalized Insights

- **Sleep, Exercise Performance:** Data shows that on the day after Yu Xiaowei had sufficient sleep and a stable bedtime (before 23:00), her morning running pace was more stable, and her subjective sense of fatigue in completing tempo runs or long runs was reduced. For example, she had a good rest the night before the long run on January 19 and

successfully completed the 15km workout. On the contrary, after a single day of excessive work screen time (greater than 8 hours) or handling urgent work in the evening, her bedtime was delayed by an average of 40 minutes on the same day, and she had a higher heart rate and slower recovery during exercise the next day.

- **Business Trips/Stress , Health:** Although there were no business trips this month, high-pressure work events (such as customer complaints and urgent overtime) had a significant impact on health indicators. Data shows that during the concentrated pressure period of January 20-22, the estimated resting heart rate increased by 5-7 beats per minute, sleep duration decreased by approximately 1 hour, and directly triggered a brief emotional fluctuation on January 24. She buffered this effectively through yoga and social interaction with best friends.
- **Exercise Type , Physical State:** Comparison shows that during periods of adhering to greater than 2 yoga sessions per week (e.g., Week 3), discomfort such as calf muscle soreness was relieved faster, and the subjective feeling of physical flexibility was improved. When the weekly running volume increased to more than 30km (e.g., Week 3), weight showed a downward trend (from 55.5kg to 55.1kg), but sufficient nutrition and sleep need to be matched to prevent fatigue accumulation. The calf soreness on January 14 occurred just after a sharp increase in running volume (8km including tempo run), reminding that volume increase should be cautious.
- **Behavioral Habits , Overall Health:**
 - **Work and Rest Stability:** More than 20 days this month, she went to bed before 23:30, which accumulated good recovery ability for her to cope with the end-of-month work wrap-up and Spring Festival Holiday, and the monthly health score was maintained in the good range.
 - **Social Quality:** High-quality in-depth social interaction (e.g., in-depth communication with best friends on January 11 and 19) contributed greatly to her mental health. Data shows that after such activities, her positive emotional state can last for 2-3 days, and indirectly improve patience in exercise and work.
 - **Family Support:** In-depth family interaction and support on January 12 and 26 were the core buffer zones for coping with work pressure and maintaining exercise motivation. The short-term weight gain caused by family gatherings was quickly adjusted through active metabolism and subsequent training.

Abnormal Event Summary

- **Sports Injury or Discomfort:**
 - Occurrence Time: After morning running on Tuesday, January 14.
 - Type: Mild soreness and tightness in the right calf muscle.
 - Impact Scope: Did not cause training interruption, but aroused her attention to training intensity and recovery. Prompted targeted stretching in the yoga class on January 15, and more attention to warm-up in subsequent training.
 - Recovery Status: Basically relieved by the morning running on January 16. The entire event was successfully managed as a physical signal and did not develop into an injury.
- **Exercise Gap Event:**
 - Gap Period: January 20 (Monday) to January 22 (Wednesday), 3 days in total.
 - Causes: Voluntarily adjusted on January 20 due to "mild cold and sore throat"; worked overtime late on January 22 due to "customer urgent needs".
 - Recovery Performance after Gap: Immediately resumed morning running and attended the interval run class on January 23, with good performance ("greatly encouraged"), indicating that the 3-day gap was mainly for physical recovery and transactional adjustment, and did not cause significant physical fitness decline.
- **Health Indicator Abnormal Change:**
 - Abnormal Indicator: Weight.
 - Abnormal Time and Magnitude: After family feast on Saturday, January 12, weight increased from 55.1kg to 55.4kg (+0.3kg); rose to the monthly peak of 55.8kg on January 29 (Spring Festival) (+0.4kg from the previous day).

- Abnormal Causes: High-calorie, high-carbohydrate family/festival gatherings.
- Subsequent Changes: Both quickly fell back to the basic range of 55.2-55.3kg within 1-3 days after restoring normal diet and exercise. Indicating that her body has rapid adjustment ability.

- **Life Rhythm Abnormal Event:**

- Abnormal Period: Wednesday, January 22.
- Abnormal Type: Worked until 23:50 due to handling customer urgent needs, resulting in severe late bedtime.
- Impact on Health and Exercise: Directly led to no exercise on the same day, severely insufficient sleep duration (estimated < 6 hours), and relied on coffee to refresh mental state the next day (January 23), but boosted mental state through morning running.
- Recovery Measures and Effects: Went to bed before 22:30 on the evening of January 23 for sleep compensation, and recovered on January 24.

H.6 QA Schema

```
{
  "question": "During the preparation for the daughter's graduation ceremony in June, on which day did the whole family (including the maternal grandmother, husband and son) participate in the final rehearsal or practice activity before the graduation ceremony?",
  "options": [
    {
      "option": "A",
      "content": "2025-06-22"
    },
    {
      "option": "B",
      "content": "2025-06-25"
    },
    {
      "option": "C",
      "content": "2025-06-28"
    },
    {
      "option": "D",
      "content": "2025-06-29"
    }
  ],
  "answer": "C",
  "required_events_id": [
    "2580"
  ],
  "required_events_date": [
    "2025-06-28"
  ],
  "ask_time": "2025-07",
  "question_type": "temporal",
  "evidence": [
    {
      "type": "agent_chat",
      "id": 409
    }
  ],
}
```

```

    {
      "type": "sms",
      "id": 1078
    }
  ]
}

{
  "question": "What_is_the_basis_for_Yu_Xiaowei's_proposal_of_the_budget_adjustment_plan_at_the_family_video_conference_on_June_15,_2025?",
  "answer": "It_is_based_on_the_data_of_Chen_Meiling's_medical,_medication_and_caregiver_expenses_in_the_past_three_months,_as_well_as_the_key_points_of_the_optimized_car_insurance_premium_case_she_compiled.These_data_helped_her_present_the_family's_financial_situation_and_serve_as_specific_cases_and_starting_points_for_discussing_financial_optimization.",
  "score_points": [
    {
      "description": "Based_on_the_data_of_the_mother's_medical,_medication_and_caregiver_expenses_in_the_past_three_months",
      "score": 5
    },
    {
      "description": "Referring_to_the_key_points_of_the_optimized_car_insurance_premium_case_as_specific_cases_and_starting_points_for_discussing_family_financial_optimization",
      "score": 5
    }
  ],
  "required_events_id": [
    "2396",
    "2398",
    "2402"
  ],
  "required_events_date": [
    "2025-06-15"
  ],
  "ask_time": "2025-07",
  "question_type": "mutihop_reasoning",
  "evidence": [
    {
      "type": "push",
      "id": 1182
    },
    {
      "type": "calendar",
      "id": "130"
    }
  ],

```

```

    {
      "type": "note",
      "id": "185"
    },
    {
      "type": "photo",
      "id": 877
    },
    {
      "type": "push",
      "id": 1183
    },
    {
      "type": "sms",
      "id": 983
    },
    {
      "type": "calendar",
      "id": "156"
    },
    {
      "type": "call",
      "id": 675
    },
    {
      "type": "note",
      "id": "238"
    }
  ]
}

```

H.7 Plot Schema

```

{
  "topic": "The_Road_to_Promotion:_The_Leap_from_Individual_Performance_to_Team_Management",

```


"detailed_description": "This topic focuses on Xiaowei's journey to achieving her core career goal being promoted to Regional Sales Manager. The timeline starts in April 2025, where she accumulates key capital for promotion through the success of a landmark performance project. In the third quarter of 2025, under sustained performance pressure and with her contributions to team collaboration, she successfully passed the assessment and achieved a career leap within three years. After the promotion, she faced the challenge of transitioning from an individual contributor to a manager, taking on the role of a mentor for new team members for the first time and experiencing a difficult period of adaptation and integration. Relying on her inherent sense of responsibility and pragmatic spirit, she ultimately guided the new employees to success, gained initial recognition from the team, and capped off this phase with the honor of attending the Annual Excellent Sales Summit.",

"monthly_description": [

{
 "month": "April",
 "content": "Yu Xiaowei collaborated closely with her colleague Su Lishan. Leveraging Su Lishan's connections in Shenzhen and her own solid professional follow-up, she successfully designed a cross-border group insurance plan for a start-up tech company operating in both Shenzhen and Hong Kong, securing a major annual order. This project stands as an excellent demonstration of her cross-regional collaboration capabilities and proficiency in handling complex business.",

"impact": "This large policy significantly boosted Yu Xiaowei's quarterly performance and became a highlight in her promotion application materials. Meanwhile, the successful collaboration with Su Lishan strengthened her partnerships within the team and accumulated positive experience for her in managing colleague relationships in her subsequent managerial role.",

"core_events": [
 "Collaborated with colleague Su Lishan to secure a large group insurance policy"

],

{

"month": "May-August",

"content": "After securing the major order, Yu Xiaowei continued to maintain her high-end client network and deliver consistent performance. At the same time, the comprehensive wealth planning project for high-net-worth clients, which she had been collaborating on with Li Guodong, a bank wealth management manager, for nearly a year, entered the final critical stage. She devoted substantial effort to in-depth communication with clients and coordinated service integration between AIA and the bank, demonstrating professional planning and resource integration capabilities that go beyond pure sales. Her supervisor Liang Zhenhua noticed her all-round performance and began to consciously give her more opportunities to showcase herself in team meetings, while privately encouraging her to prepare for promotion.",

"impact": "Her consistent high-quality performance and the ongoing major cooperation project earned her recognition and strong recommendation from her supervisor, a key prerequisite for passing the promotion assessment. This period was a sprint phase for her career goals, bringing increased work pressure but a strong sense of purpose.",

"core_events": []

},

{

"month": "September",

"content": "Early in the month, the important client asset planning project in collaboration with Li Guodong was successfully signed, a family wealth inheritance plan integrating insurance, trust and banking services with a substantial value. In the middle of the month, relying on this project and many other outstanding achievements, a stable client network, and the strong recommendation of her supervisor Liang Zhenhua, Yu Xiaowei officially passed the company assessment and was promoted to Regional Sales Manager, responsible for managing a small sales team.",

```

"impact": "The success of the project was a key milestone for her promotion, greatly enhancing her professional reputation and internal influence. The realization of the promotion is an important step in her career planning, bringing a higher professional status, income expectations and brand-new managerial responsibilities, marking her role transition from an individual contributor to a manager.",
"core_events": [
  "Completed the important client asset planning project in collaboration with Li Guodong",
  "Was promoted to Regional Sales Manager"
],
{
  "month": "October - November",
  "content": "After the promotion, Yu Xiaowei faced her first managerial challenge: the company assigned a fresh graduate to her for her to act as a mentor. Initially, she tried to impart all her years of accumulated sales skills, but found that the new employee struggled to absorb the knowledge quickly and lacked initiative in practical work. This took up a great deal of her time originally dedicated to serving high-end clients, leading to a slowdown in her short-term individual performance growth and causing her anxiety. With the guidance of her husband Zhang Weiming and suggestions from her supervisor Liang Zhenhua, she adjusted her approach, shifting from meticulous hands-on guidance to setting clear goals and frameworks, encouraging independent exploration by the new employee and conducting regular reviews. Utilizing the enthusiastic and responsible traits of her ESFJ personality, she not only provided work guidance but also cared for the new employee in daily life, helping them integrate into the team.",

```

```

"impact": "The mentorship experience was a painful growing pain of transition, testing her patience and leadership style. Through adjustments, she gradually found a balance between management and personal business. Successfully developing the new employee also initially proved her leadership potential, enhanced team cohesion, and boosted her confidence in her new managerial position.",
"core_events": [
  "Took on the role of a mentor for new team members for the first time"
],
{
  "month": "December",
  "content": "Based on her annual target-accomplishing performance, Yu Xiaowei was invited to attend the company's Annual Excellent Sales Summit held in Hong Kong. She attended the summit with her supervisor Liang Zhenhua and her colleague and close friend Su Lishan. At the summit, she participated in high-end training and briefly shared a case in the sharing session where she led the new employee to collaboratively complete a small insurance policy.",
  "impact": "Attending the summit is the company's recognition of her work achievements over the past year, greatly enhancing her professional sense of honor and sense of belonging to the company. The new connections established at the summit may bring referral opportunities for the team's business in the future. It also draws a perfect close to the theme of her 2025 career promotion.",
  "core_events": [
    "Attended the company's Annual Excellent Sales Summit (Hong Kong Session)"
  ]
}
}

```