MOBILEMEM

2026-08-20
ChinaRxiv

Abstract

As the capabilities of foundation models continue to improve, artificial intelligence is evolving from generative intelligence toward agents capable of long-term user understanding, continual learning, and autonomous evolution. Memory will become an important foundational capability for next-generation artificial intelligence, following perception, generation, and action, while edge devices serve as key carriers of long-term personal memory. Centered on edge-side memory, this white paper systematically analyzes core scientific issues such as long-term individual simulation, memory organization, and continual cognitive evolution, and proposes a personal memory layer for terminal operating systems and a knowledge-driven memory synthesis method. On this basis, it constructs MobileMem, a long-term, multimodal, and evolvable edge-side memory benchmark, promoting the transition of edge-side memory from "storing information" to "forming cognition." The white paper further summarizes a three-stage development roadmap of edge-cloud collaboration, edge-side autonomy, and trusted collaboration, and looks ahead from personal memory toward long-term cognitive infrastructure across devices and spaces, providing a new technical framework and industrial perspective for the development of next-generation personalized intelligence and intelligent terminals.

Full Text

oppo | OpenKG.CN
Chinese Open Knowledge Graph

MOBILEMEM

A New Era of On-Device Intelligent Memory

ChinaXiv:202608.00183v1

This version posted 2026-08-20.

Contents

Chapter 1 Introduction

1.1 Four Paradigm Leaps in the Development of Artificial Intelligence

The development of human civilization is, in essence, a history of the continuous evolution of information-processing capabilities. From language, writing, and printing to computers and the Internet, every technological revolution has continuously improved the efficiency of information acquisition, knowledge dissemination, and social collaboration. Artificial intelligence further endows machines with the ability to understand, generate, and use knowledge, and drives the continuous evolution of intelligent-system capabilities. From perception and generation to action, and then to the emerging capability of memory, artificial intelligence is undergoing a series of key capability leaps, gradually evolving from an information-processing tool into an intelligent agent with the capacities for continuous learning, long-term cognition, and autonomous growth.

The first leap is **perceptual intelligence**. Represented by the maturation of convolutional neural networks and deep learning, this stage was the first in which artificial-intelligence systems acquired visual recognition and speech-perception capabilities surpassing those of humans. They were able to extract effective information from unstructured data such as images and audio, achieving a breakthrough from "weak perception" to "strong understanding."

The second leap is **generative intelligence**. Represented by large language models and multimodal generative models, artificial intelligence has moved from perception toward creation, becoming capable of generating high-quality text, images, and audiovisual content. The hallmark achievements of this stage are the models' outstanding performance in question answering, content creation, and complex reasoning, marking the evolution of artificial intelligence from "recognizing the world" to "participating in creation."

The third leap is **action intelligence**. The capability boundaries of large models have expanded from "answering questions"to "executing tasks."Agent systems formulate execution plans, invoke external tools, and continuously interact with the environment, achieving a transition from "passive response" to "active action." Artificial intelligence is no longer merely an information-processing tool, but an executing subject capable of independently completing complex tasks.

The fourth leap is now taking place, namely **memory intelligence**. Although today's most advanced artificial-intelligence models perform excellently in generation and execution, their essence remains stateless. This is reflected in the fact that they can provide accurate answers in each conversation, yet cannot continuously accumulate over time, update, and truly remember a user' s preferences, habits, relationship networks, and life context. Each interaction starts almost from zero, and users cannot obtain personalized services derived from past experience. In view of this, building long-term memory that can continuously evolve is the key to enabling agents to achieve continuous learning and long-term adaptation. Long-term memory enables agents to continually accumulate experience, understand users, and support long-term planning, decision-making, and execution, pushing artificial intelligence from one-off question answering toward agents that truly possess long-term cognitive capabilities.

As foundation-model capabilities tend toward homogenization, what is truly difficult to replicate is no longer the model itself, but the individual cognitive assets formed through long-term interaction. Therefore, memory is becoming a new type of artificial-intelligence infrastructure after foundation models. Future competition

The landscape will therefore undergo a fundamental change: the technical barriers will shift away from general indicators such as model parameter scale and context-window length, and toward the ability to construct memory for individual users efficiently, with organizational quality, timely updating, and privacy and security safeguards.

1.2 Basic Models Are Not the End Point

Although the most advanced basic models in the current field of artificial intelligence perform outstandingly in knowledge question answering, content generation, and complex reasoning, their essence remains stateless. Each interaction is treated as an independent event. The model can provide accurate answers in a single round of dialogue, but it cannot precipitate the information obtained in that interaction so as to understand the next question from the same user. This limitation is mainly manifested in three aspects.

First, there is no long-term experience. The knowledge of basic models comes from the static corpus injected during the training stage, rather than from long-term memory formed through continuous interaction with a specific user. They know the general laws of the world, but they do not know what the user in front of them said yesterday, what they did last week, or where they went last year. Because they lack cross-time accumulation of experience, almost every conversation is like a first encounter, and every service requires re-establishing context; they cannot provide truly continuous, personalized intelligent services.

Second, there is no continuous growth. The core capabilities of a basic model are basically fixed after training is completed. Although it can use context to complete the current task, it will not continuously optimize itself through long-term service to the same user. As the user' s occupation, interests, goals, and social relationships constantly change, the model cannot actively absorb these changes, update its cognition, and adjust its behavior. It is therefore difficult for it to develop long-term adaptive capabilities, let alone grow into a true long-range intelligent agent.

Third, there is no individual cognition. What a basic model understands is the commonality of human beings, not the individuality of this particular person. It knows how to answer general travel suggestions, but it does not know that the user prefers window seats, dislikes seafood, and is accustomed to getting up early. This lack of individual cognition means that all personalized services can only remain at the shallow level of preferences actively set by the user, and cannot touch true deep understanding.

What basic models provide is general intelligence shared across users, whereas what a memory system forms is long-term cognition oriented toward the continuous evolution of an individual. The former enables an intelligent agent to understand the current task; the latter enables an intelligent agent to understand the user and history behind the task. Together, the two constitute a complete capability system for artificial intelligence to move from one-off interaction toward long-term service.

1.3 Memory Is the Foundational Capability of the Next Generation of Artificial Intelligence

If basic models solve knowledge coverage in the spatial dimension, then memory systems solve knowledge accumulation in the temporal dimension. Basic models can use existing knowledge to complete current tasks, whereas memory systems can continuously form, update, and evolve cognition about users through long-term interaction, enabling intelligent agents to truly possess the capabilities of cross-time learning and continuous adaptation. Artificial intelligence' s move toward long-term intelligence is, in essence, the establishment of continuously evolving cognitive capabilities in the temporal dimension.

The focus of competition in future artificial intelligence will also gradually shift from model capability to memory capability. Current competition mainly revolves around model parameter scale, context windows, and benchmark-test performance. These indicators reflect the level of general intelligence; whereas long-term intelligent services

The core lies in whether the agent can continuously understand the same user and, through long-term interaction, form stable, accurate, and evolvable individual cognition.

As the capabilities of foundation models rapidly converge toward homogenization, the models themselves are gradually becoming general-purpose infrastructure. The performance gap between open-source and closed-source models continues to narrow, and the marginal returns brought by parameter scale and context length continue to decline. When high-performance models become commonplace, what is truly difficult to replicate is no longer the model itself, but the user memory formed by the agent over the course of long-term service, as well as the sustained cognitive capabilities established from it.

Therefore, an important direction for the next stage of artificial intelligence competition will involve not only model scale and reasoning ability themselves, but also whether long-term memory can be constructed, organized, updated, and utilized more accurately, efficiently, and safely. Memory will become the core foundational capability supporting the continuous evolution of artificial intelligence and personalized services, following perception, generation, and action.

1.4 What Is On-Device Memory

On-device memory refers to a memory system that continuously accumulates, organizes, updates, and invokes an individual' s long-term experience on the user' s terminal device. Its core objective is not simply to store historical data, but, under the premise of ensuring as much as possible that user data remains locally controllable, to provide the agent with long-term cognitive capabilities related to a specific user.

Where to deploy the memory system is not merely an engineering choice; it requires comprehensive consideration of the accompanying nature, continuity, privacy, and on-device resource constraints of long-term memory. Long-term memory needs to accumulate and evolve continuously with the user' s daily use, and it must also control the scope of personal data circulation. Therefore, terminal devices such as mobile phones, computers, wearable devices, cars, home devices, and embodied robots are important locations for carrying personal memory. Based on this characteristic, a complete on-device memory system must simultaneously attend to systematicity, continuity, long-termness, privacy, and efficiency: it must systematically organize user cognition, continuously accumulate and evolve long-term experience, and, under the premise of ensuring that data remains locally controllable, provide the agent with stable and reliable long-term cognitive capabilities at lower token overhead, storage pressure, latency, and power consumption.

Systematicity. Memory is not a pileup of isolated fragments, but an organic whole. The user' s preferences, habits, relationship networks, and life contexts intertwine with one another, forming a dynamically evolving semantic network. The system must possess structured organization and association capabilities, so that new information can be absorbed by existing cognitive frameworks while also reshaping old structures in reverse, forming a spiraling upward cognitive cycle.

Continuity. The value of memory lies in sustained accumulation along the time dimension. The user' s preferences, habits, relationship networks, and life contexts can form meaningful cognition only over spans of months or even years.

This requires the memory system to remain always online and always available, and not be reset because of network interruptions or service switching.

Long-term nature. Truly valuable cognition does not come from a single interaction, but is gradually formed through long-term accumulation. A memory system not only needs to preserve information over the long term, but, more importantly, must continuously update, integrate, abstract, and evolve it, so that the cognitive structure is constantly improved over time rather than merely piling up historical data.

Privacy. Personal memory naturally contains highly sensitive information, including the user's social relationships, health status, financial situation, location trajectories, and consumption habits. Once such information leaves the terminal device and enters the cloud, no matter how complete the encryption measures are, the relationship of control over the data and the exposure surface are fundamentally changed.

Efficiency. On-device memory must meet the requirements of efficient operation under resource-constrained conditions. Long-term memory continuously accumulates a large amount of interaction content; if all historical records are directly input into the foundation model, it will cause unacceptable token overhead, latency, power consumption, and storage pressure. Therefore, an on-device memory system needs to complete necessary information filtering, compression, summarization, indexing, and hierarchical management locally, activating only a small amount of high-value memory relevant to the task, so as to avoid linear expansion of memory scale with the number of interaction rounds. Efficiency determines whether on-device memory can move from short-term demonstrations toward long-term stable operation.

Taken together, at present only on-device devices can simultaneously satisfy the above five requirements. Terminal devices are always with users and can continuously perceive behaviors and interactions in the real world; data remaining locally helps strengthen users' access, management, and deletion control over memory; even if the network is interrupted, memory can still continue to accumulate and will not be interrupted by service migration. Therefore, on-device is not a deployment preference, but an important carrier of memory intelligence.

After satisfying the above basic conditions, a truly usable memory system also needs to possess two key capabilities: robustness, and forgetting and activation.

Robustness. The real world is full of differences; different cultures, languages, lifestyles, and social relationships all affect the forms in which memory is expressed. The memory system needs to maintain stable organization, retrieval, and reasoning capabilities, so that it can continuously provide accurate, consistent, and traceable memory services across different user groups and application scenarios. Robustness is not an additional capability, but a basic requirement for long-term operation.

Forgetting and activation. The efficiency of memory does not come from unlimited storage, but from reasonable management. The system needs to actively forget redundant, outdated, or conflicting information to prevent cognition from continuously swelling; at the same time, it should be able to actively activate relevant memories according to the current context, providing the most valuable information at the appropriate moment. Forgetting keeps memory lightweight, while activation keeps memory fresh.

As shown in Figure 1-1, terminal devices accompany users throughout the day and naturally carry behavioral data that is closer to real usage scenarios; keeping data locally as much as possible helps reduce privacy leakage and cross-platform transfer risks at the source; without relying on the network, memory can continuously accumulate, update, and evolve along a continuous timeline. Precisely because of this, the on-device memory intelligence can gradually form long-term cognition of users, achieve the ability of continuous evolution in which "the more it is used, the better it understands the user; the more it is used, the closer it becomes to the user," and

Figure 1-1 Examples of multiple scenarios for mobile memory

Ultimately, it supports truly long-range agents. This capability to move from one-off answers toward continuous growth will become a fundamental feature distinguishing the next generation of intelligent assistants from previous products.

1.5 The Significance of On-Device Memory

On-device memory is a key foundational capability for artificial intelligence to move from general capabilities toward individual intelligence. Foundation models endow artificial intelligence with broad knowledge, reasoning, and generation capabilities, giving it the general intelligence basis for handling complex tasks; on-device memory further combines these general capabilities with the long-term experiences of specific users, enabling intelligent systems to form an understanding of individuals through continuous interaction. The core of on-device memory is not simply to store historical information, but to continuously accumulate, organize, and update experience and knowledge related to the individual on the user side, so that the intelligent system can continually optimize its own behavior based on long-term interaction and provide services that better fit user needs.

First, on-device memory will profoundly change the way artificial intelligence serves individuals. Traditional intelligent systems mainly rely on current input to respond; different interactions lack continuity, making it difficult to truly understand users' long-term behavioral patterns and preferences. By contrast, an intelligent system with long-term memory can integrate the user' s past behaviors, habits, preferences, and historical experiences to make more accurate judgments about current needs. For example, in daily-life scenarios, the system can not only answer the user' s current question, but also incorporate long-accumulated experience to understand the user' s usual decision-making style and personalized needs. As a result, artificial intelligence services will gradually shift from instant responses based on single interactions to an intelligent service model oriented toward long-term companionship and continuous optimization for the individual.

Furthermore, edge-side memory provides an important foundation for intelligent agents to achieve continuous learning and long-term adaptation. An intelligent agent capable of autonomously completing complex tasks not only needs planning and execution capabilities, but also must be able to accumulate experience from past actions and apply it in future tasks. Which strategies have proved effective, which attempts have led to failure, what needs users have previously expressed, and what constraints exist in particular environments—all these experiences need to be continuously preserved, summarized, and invoked. For embodied intelligence and robots, this capability is especially critical. Digital agents mainly face tasks in information spaces, whereas robots must operate for long periods in real, open, and dynamically changing physical environments. Household robots need to gradually understand family members and living environments; service robots need to adapt to the long-term needs of different users; and industrial robots also need to accumulate operational experience related to specific equipment and production processes. Therefore, edge-side memory will become an important support for driving embodied intelligence from merely possessing action-execution capability toward possessing environmental adaptation and experiential accumulation capability.

At the same time, large-scale application of edge-side memory also brings new security and governance challenges. As memory content becomes increasingly rich, intelligent systems can provide more precise services, but they may also involve large amounts of personal privacy and sensitive information. Therefore, future memory systems must not only focus on the effective storage and utilization of information, but also address key issues such as memory access permissions, content management, dynamic deletion, and user autonomous control. Keeping personal memories on the user side as much as possible helps enhance users' control over their data and reduce the risks caused by the circulation of sensitive information across different platforms. Thus, edge-side memory is not only an important direction for improving artificial intelligence capabilities, but also an important foundation for building trustworthy, controllable, human-centered personal AI systems.

From a longer-term perspective, the significance of edge-side memory lies in introducing a temporal dimension of continuous evolution into artificial intelligence. Foundation models give artificial intelligence a broad understanding of world knowledge, while memory enables it to accumulate long-term experience related to specific individuals and specific environments. The former answers how artificial intelligence understands the world; the latter determines how artificial intelligence understands a particular person and how it continuously adapts and changes through long-term interaction. Combining the two will promote the development of artificial intelligence from a general model that relies on pre-trained knowledge into a personalized intelligent system capable of continuous learning, proactive adaptation, and co-growth with users. Future competition in artificial intelligence will lie not only in model scale and reasoning capability, but also in whether systems can understand users over the long term, accumulate experience, and form distinctive individual intelligence. Edge-side memory is therefore expected to become a key bridge connecting foundation models and personal intelligence, driving artificial intelligence into a new stage centered on long-term companionship, continuous adaptation, and personalized evolution.

Chapter 2 Scientific Questions

The construction of on-device memory is, in essence, not about developing a new information-storage mechanism, but about exploring how artificial intelligence can form human-like sustained cognitive abilities over long time scales. Current foundation models mainly rely on large-scale pretraining to acquire static world knowledge; their capability boundaries are determined by training data and model parameters, making it difficult for them to continuously accumulate experience, form individual understanding, and achieve autonomous growth through long-term interaction with real users.

The core advantage of human intelligence lies not only in possessing rich knowledge, but also in continuously recording events, summarizing experience, forming cognition, and continually adjusting one' s own behavior as the environment changes over many years of life experience. By comparison, existing artificial intelligence still lacks the ability to model an individual' s long-term experience; it cannot simulate the behavioral evolution process of real users over a scale of years, nor can it transform fragmented, multimodal, dynamically changing personal experiences into a stable and continuously updatable cognitive system.

Therefore, on-device memory needs to break through the existing static-knowledge paradigm of artificial intelligence and establish a new cognitive mechanism oriented toward long-term individual intelligence. The following discussion introduces three key scientific questions: long-horizon individual simulation, on-device memory organization, and continuous cognitive evolution.

2.1 Long-Horizon Individual Simulation Mechanism

Human understanding of the world and of the self comes from long-term life experience, rather than from the accumulation of isolated events. A person' s interests, habits, values, and behavioral patterns often require years or even longer periods of observation and interaction to gradually take shape. However, current artificial intelligence mainly acquires capabilities through offline data training and lacks a mechanism for simulating the long-term experiences and behavioral evolution of real individuals.

For on-device intelligence, it is necessary to explore how to build an individual-experience simulation environment spanning months, years, or even longer time scales, so that artificial intelligence can undergo a continuous life process similar to that of real users. In this process, the system must not only generate reasonable user behavior trajectories, but also model changes in user goals, shifts in interests, environmental influences, and long-term interactive relationships between humans and intelligent agents.

The core challenge of long-horizon individual simulation is that real human behavior is highly complex and uncertain; there are significant differences between short-term behavior and long-term characteristics, and accidental events, environmental changes, and personal growth all affect future behavior. Therefore, it is necessary to establish a simulation mechanism capable of reflecting the laws of long-term individual change, enabling artificial intelligence to learn from long-term experience how to understand and predict individuals.

Therefore, the fundamental scientific question that needs to be answered is:

How can we break through the existing short-duration interaction paradigm of artificial intelligence and construct an individual-experience simulation mechanism oriented toward a multi-year scale, so that an intelligent agent can form a deep understanding of user behavior, preferences, and environmental changes through long-term interaction?

2.2 On-Device Memory Organization Mechanism

Long-term individual intelligence requires not only experience, but also the effective organization of experience. In fact, the data generated by users on the device side are highly heterogeneous, including multimodal information such as text communication, images, videos, location trajectories, application behaviors, sensory information, and environmental feedback. Such information is not merely a simple data record, but an important component of an individual' s life experience. At present, memory mechanisms in artificial intelligence mainly revolve around textual context or information retrieval, making it difficult to handle long-term, multimodal, dynamically changing personal experiences. How to identify key events from massive fragmented data, establish temporal associations, causal relationships, and semantic links among events, and form a structured and callable personal memory system is a core challenge for realizing on-device intelligence.

The essence of on-device multimodal memory organization is to transform raw experiences into cognitive representations that are understandable, inferable, and transferable. This process involves key mechanisms such as memory selection, information compression, semantic abstraction, relationship modeling, and cross-modal fusion. At the same time, the system needs to distinguish short-term states from long-term features, identify incidental events and stable patterns, and avoid cognitive confusion caused by the accumulation of ineffective memories.

Therefore, the fundamental scientific question that needs to be answered is:

How can key experiences be extracted from long-term, multimodal, dynamically changing personal experiences, and how can an on-device memory organization mechanism be established that supports understanding, reasoning, and decision-making, so as to achieve the leap from data records to the formation of individual knowledge?

2.3 Continuous Cognitive Evolution Mechanism

Memory formation is not the endpoint of individual intelligence, but the foundation for continuous growth. Human cognition is not fixed; rather, it is continuously adjusted through experience, feedback, and reflection. Future on-device intelligence must likewise possess the ability to evolve over the long term, enabling it to continuously update its self-knowledge as users, environments, and tasks change. The core of continuous cognitive evolution is to study how artificial intelligence can realize a learning and adaptation process similar to that of living systems under limited resource constraints. An intelligent agent needs to autonomously decide which experiences are worth retaining, which knowledge needs to be updated, and which experiences should be forgotten, and to continuously optimize its own behavioral patterns through long-term feedback.

However, on-device environments are characterized by limited computational resources, continuously growing data, and constantly changing user needs. How to avoid knowledge conflicts, cognitive drift, and efficiency degradation caused by long-term accumulation, while ensuring that the memory update process is safe, controllable, and trustworthy, is an important challenge faced by continuous cognitive evolution.

Therefore, the fundamental scientific question that needs to be answered is:

How can the static-model paradigm of artificial intelligence be broken through, so that intelligent agents can autonomously update, reorganize, and optimize their cognitive structures during long-term interaction, thereby realizing individual-oriented continuous cognitive evolution?

Overall, the core objective of on-device memory research is not to make artificial intelligence simply remember more information, but to explore the fundamental laws by which artificial intelligence moves from static knowledge intelligence toward dynamic experience intelligence. Long-range individual simulation addresses the question of how intelligence acquires long-term experiences; on-device multimodal memory organization addresses the question of how experiences are transformed into cognition; and continuous cognitive evolution addresses how cognition can continuously ···

the problem of sustained growth. Together, the three promote artificial intelligence from a general model that depends on pre-trained knowledge toward a

Figure 3-1 Personal Memory Layer Architecture Diagram

Figure 1: Figure 3-1 Personal Memory Layer Architecture Diagram

new generation of individual intelligent systems capable of understanding individuals, adapting to environments, and developing continuously.

Important future breakthroughs in artificial intelligence will lie not only in possessing larger model scales and stronger reasoning capabilities, but also in whether, like humans, it can form distinctive cognition through long-term experience. Edge-side memory may therefore become a key theoretical foundation linking foundation models with individual intelligence, driving artificial intelligence to evolve from a one-off interaction tool into an intelligent entity capable of long-term companionship, continuous learning, and shared growth.

Chapter 3 Personal Memory Layer

Figure 3-1 Personal Memory Layer Architecture Diagram

This chapter points out, from the perspective of scientific problems, that end-side memory needs to solve the questions of how long-term experience is acquired, how it is organized, and how it continuously evolves. It focuses on discussing how the personal memory layer receives multi-source behavioral data in terminal systems and organizes it into long-term memory that is searchable, updatable, and inferable. In other words, the personal memory layer is the system implementation carrier for the end-side multimodal memory organization mechanism and the continuous cognitive evolution mechanism.

3.1 Definition

The personal memory layer is the core middleware in a terminal operating system responsible for the management of user cognitive data. It undertakes system-level functions such as behavioral perception, memory organization, and intelligent retrieval. As shown in Figure 3-1, in the system architecture of an end-side intelligent agent, the personal memory layer is clearly defined as foundational middleware at the operating-system level, serving as a key bridge between user behavioral data and the agent' s reasoning and decision-making. Its core functions are to continuously collect behavioral signals from users' active interactions and from passive system observations on the terminal, perform local encoding, organization, updating, and persistent storage of information, and, when the intelligent agent initiates a request, provide precise and efficient memory retrieval services with contextual awareness. This design makes the personal memory layer the foundation on which terminal intelligent agents achieve continuous cognition and personalized services.

The proposal of the personal memory layer marks the migration of memory capability from application functions to operating-system-level infrastructure. At present, storage management in mainstream operating systems centers on file systems and databases, serving the data-persistence needs of applications; it is a passive

…, static, warehouse-style management. By contrast, the personal memory layer takes the agent's cognition of the user as the organizing thread, emphasizing the association, abstraction, and evolution of information; it is a proactive, dynamic form of cognitive management. This transformation echoes the deep evolution of operating-system design philosophy. At present, mainstream operating systems in the industry are undergoing a paradigm shift from interaction tools to agents; system-level AI runtimes, heterogeneous computing architectures, and device-cloud collaborative model matrices have become the core capabilities for building 端-side intelligence. Under this trend, the personal memory layer, as a key subsystem of the agent operating system, provides the upper-layer agent with a cognitive infrastructure that supports continuous accumulation and dynamic evolution. Just as an operating system manages a computer's hardware and software resources, the personal memory layer manages cognitive resources about the user, by organizing fragmented behavioral data into a coherent cognitive network, thereby transforming static factual records into dynamic knowledge assets.

From the perspective of system boundaries, the inputs to the personal memory layer are heterogeneous, continuously generated behavioral streams. These behaviors may be actively initiated by the user, such as conversations with an intelligent assistant and operation records in various applications; they may also be obtained through passive system observation, such as changes in screen content, the presentation of notification-bar information, and the collection of sensor data. In a real mobile environment, users and intelligent assistants not only interact with each other, but also connect extensively with third-party applications. This gives behavioral streams a naturally high degree of heterogeneity, encompassing natural-language instructions, application operation events, interface-state snapshots, and application-specific interaction records.

The outputs of the personal memory layer are structured memory contents that can be directly invoked by the agent. These contents need to support multiple access methods, including implicit retrieval based on semantic similarity, explicit queries based on entity relationships, and filtering and ranking based on time ranges. The output forms may include factual statements, event summaries, preference inference, process guidance, or relationship descriptions, with the specific form depending on the needs of downstream tasks.

There is an essential difference between the personal memory layer and the storage systems on which traditional software systems rely. Traditional storage systems take files or records as their units, emphasizing persistent preservation and accurate restoration of data; they are a passive, static, warehouse-style management system. The personal memory layer, however, takes the agent's cognition of the user as its organizing thread, emphasizes the association, abstraction, and evolution of information, and is a proactive, dynamic form of cognitive management. The latter must not only answer factual questions such as "what happened," but also support judgmental questions such as "what kind of person is the user" and "what might the user need."

3.2 Key Challenges

In real mobile environments, the implementation and deployment of the personal memory layer face a series of challenges that have not yet been systematically resolved. These challenges run through the entire chain of data access, memory representation, cognitive reasoning, and system deployment.

Unified access and organization of heterogeneous data. User behavior is dispersed across a large number of applications, including communication, social networking, audio and video, travel, shopping, and health. The data cover multiple modalities such as text, images, video, speech, location trajectories, and sensor information, and are characterized by heterogeneous formats, semantic fragmentation, and temporal discreteness, forming a large number of information silos. How to achieve unified access, semantic alignment, and temporal organization of cross-application and cross-modal data in the 端-side environment is the foremost challenge in constructing the personal memory layer.

Unified representation and continuous evolution of long-term memory. Personal memory requires a unified representation of different types of information, such as facts, events, tasks, and user preferences, while also supporting continuous accumulation, dynamic updating, and reasonable forgetting. As user behavior changes continuously, the system must both avoid catastrophic forgetting and eliminate outdated information while controlling storage scale, so as to achieve long-term, stable evolution. In addition, different types of memory have different update frequencies and storage requirements, creating an inherent tension between unified representation and efficient management.

From information storage to users' cognitive understanding. The value of a memory system lies not only in recording and retrieval, but also in forming a sustained understanding of the user. An intelligent agent needs to abstract user intentions, behavioral patterns, and long-term preferences from discrete events, and to establish cross-time and cross-task associative reasoning capabilities, thereby supporting long-term planning, decision-making, and personalized services.

System optimization under end-side resource constraints. The above capabilities all need to be deployed on mobile terminals, where they are constrained by token consumption, storage, latency, power consumption, computing power, and memory. Traditional memory systems process all interaction data without filtering, causing resource consumption to grow linearly with the number of rounds; this is dramatically amplified on the end side. It is therefore necessary to design lightweight storage, hierarchical memory management, incremental updating, and efficient reasoning mechanisms, so as to achieve long-term

Figure 4-1 Example of a knowledge-driven memory synthesis method

Figure 2: Figure 4-1 Example of a knowledge-driven memory synthesis method

stable operation under resource-limited conditions.

The above challenges cut across the four levels of data acquisition, memory management, cognitive understanding, and system deployment, and they are mutually coupled and mutually constraining. Data organization affects memory representation; the mode of representation determines the capacity for continuous evolution; and cognitive capability is in turn constrained by end-side resources. Therefore, the construction of the personal memory layer requires coordinated design of data, memory, cognition, and computing resources at the system-architecture level, rather than reliance on a single algorithm or local optimization.

Paradigmatic limitations of existing memory methods. Current research on end-side memory still largely follows an externalized memory paradigm of information storage and on-demand retrieval: that is, the interaction content generated by an intelligent agent is treated as external data independent of the model' s cognitive process, stored by means of summarization, vectorization, knowledge graphs, or retrieval augmentation, and re-injected into the model when required by a task. In essence, this paradigm regards memory as an auxiliary storage space outside the model' s capabilities. It mainly addresses the problems of how to remember and how to find, but it struggles to answer deeper questions such as what to remember, how to organize it, when to update it, and how to form sustained cognition from experience. Especially in long-term, continuous, real-world mobile environments, user experiences are not mutually independent static pieces of information, but continuous processes that accumulate, associate, revise, and evolve over time. Simple memory appending and retrieval easily lead to information redundancy, semantic fragmentation, and conflicts between old and new knowledge, making it difficult for memory to form stable user cognition. Therefore, the personal memory layer must break through the traditional retrieval-centered paradigm, shifting from adding a memory store to the model toward establishing a continuous cognitive layer for the intelligent agent; it must elevate memory from a passive storage mechanism into an active system that connects experience acquisition, memory organization, cognitive formation, and continuous evolution, enabling the intelligent agent to continuously form, revise, and update its understanding of the user through long-term experience.

Chapter 4 Knowledge-Driven Memory Synthesis

Figure 4-1 Example of a knowledge-driven memory synthesis method

To build an evaluable end-side memory system in real mobile environments, one faces a fundamental data challenge: although the behavioral data generated by users in daily life are abundant, they are naturally fragmented, heterogeneous in source and structure, and highly private. They are difficult to use directly for system development and evaluation, and they also cannot be simply stitched together into long-term trajectories with temporal continuity. To address this problem, we propose a knowledge-driven memory synthesis method that constructs simulated personal experiences with long time spans, interpersonal relationship constraints, and multimodal interaction features, and on this basis forms the end-side memory evaluation benchmark MobileMem. This method mainly responds to the problem of long-range individual simulation mechanisms in Chapter 2, while also providing reproducible evaluation data for the organization, retrieval, updating, and reasoning capabilities of the personal memory layer described in Chapter 3.

As shown in Figure 4-1, the knowledge-driven memory synthesis method takes personal knowledge-graph construction as its central anchor and proceeds step by step according to the workflow of prior-driven user modeling, knowledge-graph construction, long-range trajectory generation, multimodal memory synthesis, and question construction. It transforms users' static images, historical behaviors, and interaction information into long-term memory with **temporal continuity, semantic consistency, and state evolvability**.

interactive trajectories, providing a controllable and reproducible data foundation for the systematic evaluation of terminal memory systems, and promoting terminal memory research from an exploratory stage characterized by nonuniform data construction methods and inconsistent evaluation standards toward a standardized stage featuring unified criteria, comparability, reproducibility, and iterability.

Knowledge-driven memory synthesis methods are not only used for constructing data for terminal-memory benchmark tests; more importantly, they provide a technical pathway by which terminal memory systems can move from data accumulation toward sustained understanding and long-term cognitive evolution. This method centers on three core capabilities: long-horizon temporal-trajectory construction, cross-modal knowledge alignment, and continuous evolution of user states. It promotes the evolution of terminal memory from static information storage toward a dynamic, structured, and inferable cognitive system.

First, long-horizon temporal-trajectory construction. In response to the problem that, during real users' long-term use, information is scattered across different times, different applications, and different interaction carriers, events, sessions, and trajectories are taken as the basic organizational units. Multi-source information such as users' multi-turn textual dialogues, application operation records, system behaviors, and screenshots is uniformly modeled and temporally aligned, forming long-term interaction trajectories with clear temporal structure, event associations, and contextual continuity. By organizing discrete interactive behaviors into continuous event chains, the system can comprehensively characterize the occurrence process of user behavior and its contextual evolution, thereby providing a structured foundation for the storage, updating, and retrieval of long-term memory.

Second, cross-modal knowledge alignment. For complex scenarios in terminal memory where multimodal information coexists and semantics are mutually related, a unified knowledge-organization framework is built through a personal knowledge graph, mapping information from different modalities into a unified event and relation structure, and realizing semantic alignment and associative modeling among visual, linguistic, and behavioral information. On this basis, joint understanding and inference across modalities are supported, enabling information scattered across different modalities to complement and verify one another, thereby improving the overall ability of terminal memory to understand complex user experiences.

Third, continuous evolution of user states and personalized adaptation. For dynamic scenarios in which user preferences, behavior patterns, social relationships, and event states continuously change during long-term interaction, user memory is modeled as a structured knowledge system that is continuously updated and evolves over time. By continuously absorbing new interaction information, dynamically updating event states, user preferences, and relationship structures, and revising, integrating, and, when necessary, forgetting historical information, the user model can evolve continuously along with the real usage process. This enables dynamic characterization of users' long-term states and personalized adaptation, promoting the evolution of terminal memory from a statically constructed knowledge base into an individual cognitive system that continuously learns, updates, and adapts.

4.1 Overall Approach

The construction of a terminal memory system starts from personal information modeling, abstracting users' fragmented behavioral data into structured knowledge representations that can be understood and inferred by an intelligent agent. On the basis of personal information modeling, the knowledge-driven memory synthesis method transforms user images into long-term interaction trajectories with temporal continuity and semantic consistency. As shown in Figure 4-2, its basic process uses the user 先

Figure 4-2 Schematic diagram of the construction method of MobileMem (this figure is a schematic for the multimodal Omni version)

Starting from prior knowledge, the user' s personal knowledge graph is constructed by organizing the user' s static profile, social relationships, and important milestones into a user-centered personal knowledge graph with the user as the central node and frequently contacted people as peripheral nodes. This serves as the semantic anchor for all subsequent generation, ensuring that the generated content remains consistent in terms of interpersonal relationships, spatiotemporal background, and life logic. Based on this knowledge graph, the system generates an annual event sequence anchored to important personal and

Figure 4-2 schematic: MobileMem construction method. The diagram shows a flow from memory pattern/image agent to personal graph/Wenshengtu model/reference photo, then events and dialogue. It includes a personal knowledge graph with roles such as client, colleague, and landlord; basic information (name: xxx; birthday: 199x-01-01; identity: intermediate engineer; interests: technology, investment; ⋯); notes and screen recordings; and memory-related dialogues.

Figure 3: Figure 4-2 schematic: MobileMem construction method. The diagram shows a flow from memory pattern/image agent to personal graph/Wenshengtu model/reference photo, then events and dialogue. It includes a personal knowledge graph with roles such as client, colleague, and landlord; basic information (name: xxx; birthday: 199x-01-01; identity: intermediate engineer; interests: technology, investment; ⋯); notes and screen recordings; and memory-related dialogues.

cultural dates. Each event is further decomposed into temporally ordered sub-event steps, with annotations of the possible types of mobile applications and visual materials involved. The core simulation process occurs in the dialogue synthesis stage. Each sub-event step is decomposed into fine-grained memory points, representing key information fragments that need to be retained during the user–AI interaction process. The memory points further generate dialogues, and all dialogue fragments are concatenated in chronological order to form a complete mobile interaction trajectory, providing a controllable and reproducible data basis for evaluating on-device memory systems.

4.2 MobileMem Construction

To evaluate the core capabilities of on-device memory systems in long-term mobile interactions, we construct an on-device memory benchmark evaluation system covering both text and multimodal scenarios. The overall framework is based on user prior knowledge; through knowledge-driven synthesis of long-range experience, it constructs user interaction trajectories with temporal continuity and semantic consistency, and further automatically generates multi-level evaluation questions from these trajectories. The entire process consists of four modules: construction of user prior knowledge, synthesis of long-range interaction trajectories, synthesis of question–answer pairs, and quality control. Among them, the knowledge-driven memory synthesis framework is the core of the entire data construction process. It is used to organize dispersed user knowledge and existing interactions into continuously evolving long-term experiences, and on this basis to form data for memory evaluation.

Prior-driven user modeling. Directly relying on large language models to generate complete user data can easily lead to problems such as insufficient user diversity, limited personalization, and lack of constraints on long-term behavior. Therefore, we adopt prior-driven user modeling, providing a basis for subsequent long-term traject-

It provides stable identity, state, and behavioral constraints for trajectory synthesis. Depending on differences in data sources and modality scope, the text benchmark MobileMem and the multimodal extended version MobileMem-Omni adopt different prior-knowledge construction strategies. For data involving real users, we perform the necessary anonymization and privacy desensitization.

Long-horizon trajectory generation. Given user prior knowledge, our goal is to generate user interaction trajectories that can span relatively long time ranges, evolve continuously, and remain internally consistent. Directly generating a complete trajectory in one pass can easily lead to a lack of constraints among events and makes it difficult to ensure consistency over a long timeline. Therefore, we decompose long-horizon trajectory construction into a hierarchical synthesis process from coarse to fine, and use existing user knowledge as persistent constraints.

We propose a knowledge-driven memory synthesis framework. Its core idea is to use user prior knowledge and existing interactions as knowledge anchors, and, under the joint guidance of user profiles and temporal constraints, organize scattered knowledge into a unified long-term interaction flow, while continuously generating and revising subsequent experiences as the user' s history gradually unfolds. Thus, a trajectory is not a static concatenation, but instead gradually expands and evolves under the joint action of existing knowledge and newly generated experiences. Specifically, the framework adopts a combination of top-down knowledge guidance and bottom-up experience evolution. First, starting from the user' s initial profile and time span, the complete temporal span is divided into life events of different granularities and recursively refined into specific interaction sessions; in this process, existing knowledge anchors are mapped to corresponding events and used as inviolable contextual constraints. Subsequently, concrete human–computer interactions are generated at leaf nodes, and the newly generated experiences further exert a reverse influence on future events that have not yet unfolded, thereby dynamically adjusting event structure, temporal relations, and user state. This process forms a continuous feedback loop of planning, anchoring, synthesis, and revision, enabling the generated trajectory to simultaneously satisfy knowledge constraints, temporal consistency, and experience evolution.

MobileMem adopts different concrete implementations under the above unified framework. MobileMem mainly retains knowledge anchors and textual interactions, and fully uses hierarchical event planning and an experience feedback mechanism; MobileMem-Omni, by contrast, performs simplified event unfolding according to a predefined annual event sequence, and further incorporates multimodal constraints such as participants, mobile applications, and visual content. Because the Omni trajectories are relatively large in scale and their complete trajectory length can exceed the million-token level, explicit low-level feedback revision is omitted under the premise that preceding and subsequent events remain historically and contextually consistent, in order to reduce the cost of data synthesis.

Evaluation benchmark generation. After obtaining long-horizon interaction trajectories and multimodal memories, we further automatically construct evaluation question–answer pairs from the trajectories and their corresponding user knowledge. Because the trajectories themselves have a hierarchical structure, we adopt a bottom-up question synthesis strategy to fully use memory information at different temporal scales and different granularities. Specifically, starting from a predefined set of question types, at leaf nodes we generate basic question–answer pairs based on specific user attributes or interaction sessions; at internal nodes, questions generated by child nodes are used as construction units and further combined to form questions that require information integration across attributes, events, or sessions. Unused child-node questions and newly generated complex questions are passed upward together, enabling the questions to expand progressively along with the hierarchical structure,

from single-point fact retrieval to cross-session multi-hop reasoning. During synthesis, the set of question types can also be dynamically expanded according to newly emerging reasoning patterns.

Data quality control. To ensure the reliability of the synthesized data, we establish a multi-level quality-control process covering trajectories, visual content, and evaluation questions. For long-range interaction trajectories, because existing automatic metrics are difficult to use to reliably measure global consistency across long time spans, we combine structured constraints with manual sampling review for quality control. During the automatic synthesis process, temporal constraints, event-level constraints, consistency of knowledge anchors, and contextual verification are used to reduce obvious conflicts; trajectory segments are then manually sampled for inspection, with particular attention to temporal relations among events, changes in user states, and whether clear contradictions exist between earlier and later interactions. A small amount of natural noise, such as numerical deviations that may occur in data resembling real users, is retained so long as it does not affect overall semantic consistency.

For multimodal content, we further conduct dedicated image quality control. First, multimodal large language models are used to check the semantic consistency between images and their corresponding visual descriptions, and to filter out images that are clearly mismatched, unsafe, or of low visual quality. For images involving users or persons in the personal knowledge graph, a face-recognition model is further used to compare the generated results with the corresponding reference persons; images below the threshold are regenerated, and samples that still fail to meet the requirements after multiple rounds of generation are discarded. After automatic filtering, manual sampling is used to examine the consistency between the visual content and the interaction trajectory in which it appears.

For evaluation question–answer pairs, large language models are first used to verify whether the questions can be fully answered by evidence in the source memory, and to detect questions with high similarity, insufficient information, or incorrect answers. Low-quality and redundant samples are rewritten or removed, and finally manual sampling review is used to further verify the correctness of the questions and the sufficiency of the evidence. For some multi-hop questions generated through automatic composition, multiple relatively independent single-hop questions may have been combined. Although such questions have certain limitations in naturalness, their answers still have clear evidential support and can effectively test a memory system' s ability to retrieve across multiple information fragments and integrate evidence.

Through the above process, we ultimately construct an end-side memory benchmark that covers text and multimodal mobile interactions, accommodates long-term temporal consistency and multi-level memory capabilities, and provides a basis for unified and reproducible evaluation of different memory systems in real mobile environments.

Chapter 5 Benchmark Evaluation and Ecosystem Applications

The construction of the Mobile Memory benchmark and the establishment of an evaluation system provide a standardized data foundation and assessment tools for research on on-device memory, enabling system capabilities to move from qualitative demonstration toward measurability, comparability, and sustainable improvement. On this basis, the Mobile Memory benchmark has been deeply integrated into the productization practice of end-side artificial intelligence, producing substantive impact at multiple levels, including the industrial ecosystem, standardization, on-device intelligence, user value, and concrete application scenarios.

5.1 Dataset Details

Table 5-1 presents the data statistics of MobileMem-Omni (the multimodal version). This dataset includes balanced English and Chinese users, with each user associated with a large number of events spanning a relatively long time period. In terms of interaction statistics, each user in the benchmark has a relatively long context length, each user includes multiple sessions, and each session contains multiple rounds of dialogue. A large number of images are distributed across the sessions, reflecting the multimodal characteristics of real-world mobile usage. The distribution of question types covers seven reasoning categories, enabling a comprehensive evaluation of different memory capabilities.

Table 5-1 Statistical Information of MobileMem-Omni

Statistical Category	Metric	Value
User statistics	Total number of users	16

Statistical Category	Metric	Value
User statistics	English-interaction users	8
User statistics	Chinese-interaction users	8
User statistics	Total number of events	1,589
Interaction statistics	Average context length (tokens/user)	1.72 million
Interaction statistics	Average number of sessions (per user)	202.6
Interaction statistics	Average dialogue rounds (per session)	48.2
Interaction statistics	Total dialogue rounds	155,670
Interaction statistics	Average number of images (per session)	5.88
Interaction statistics	Total number of images	19,060
Question-type distribution	Total number of questions	7,415
Question-type distribution	Single-hop retrieval	986
Question-type distribution	Multi-hop reasoning	1,135
Question-type distribution	Knowledge updating	1,010
Question-type distribution	Temporal reasoning	773
Question-type distribution	Implicit preference inference	1,226
Question-type distribution	Refusal to answer	1,208
Question-type distribution	Visual reasoning	1,077

As shown in Table 5-2, compared with representative dialogue-memory benchmarks (including Mem-Gallery, HaluMem, PersonaMem, LoCoMo, and LongMemEval), MobileMem-Omni exhibits several distinctive characteristics. Unlike most existing benchmarks oriented toward general scenarios, MobileMem-Omni is designed specifically for mobile screenshots and supports multimodal input. Among the compared benchmarks, MobileMem-Omni adopts bilingual Chinese-English and role-based language settings, whereas the other listed benchmarks are mainly in English. In addition, MobileMem-Omni uses synthesized personalized images as image sources, while other multimodal benchmarks use open-source images. In terms of role construction, MobileMem-Omni includes both real persons and roles generated by large models, whereas all other benchmarks rely only on roles generated by large models. These unique features make MobileMem-Omni a dialogue-memory evaluation framework with greater authenticity and privacy awareness in mobile environments.

Table 5-2 Comparison of MobileMem-Omni with Other Memory Datasets

	MobileMem-Omni	Mem-Gallery	HaluMem	PersonaMem	LOCOMO	LongMemEval
Dialogue time span	One year	Several months	Several years	Several years	Several months	Several years
Average single-session length	8.5k words	–	8.3k words	6k words	477 words	3k words
Maximum context length	2.10 million words	–	1.00 million words	1.00 million words	9k words	1.50 million words
Multimodal support	Yes	Yes	No	No	Yes	No
Target platform	Mobile	General	General	General	General	General
Average number of images per session	5.9	4.2	–	–	3.4	–
Image source	Synthetic private images	Open-source images	–	–	Open-source images	–
Language	English + Chinese	English	English	English	English	English
Role source	Real users + large-model generation	Large-model generation	Large-model generation	Large-model generation	Large-model generation	Large-model generation
Number of questions	7,415	1,711	3,467	Approx. 6,000	7,512	500

5.2 Industrial Ecosystem Applications

MobileMem is built on OPPO' s native mobile memory ecosystem, covering multiple long-term memory sources such as calendar, photo albums, notes, documents, to-dos, voice notes, Xiaobu Memory, and video notes. Together, these heterogeneous memory sources record temporal information, visual content, textual records, browsing behavior, multimedia consumption, and personal behavioral trajectories generated during users' long-term use of mobile devices, providing an authentic and continuous user-memory foundation for next-generation mobile AI with long-term memory capabilities.

(1) Health Management and Medication Reminders

> An elderly user routinely records blood pressure and blood glucose, and has also saved previous medical examination reports and doctors' instructions. Over the past few months, after each follow-up visit, the user has recorded the next appointment time in the calendar and casually noted the doctor' s instructions in notes.

When organizing medications, they browsed the contraindication information for a commonly used drug in the browser and saved lab reports and prescription photos through the album. One early morning, the user felt a little dizzy and opened the mobile assistant to ask: "What drugs was I allergic to before? Help me check whether any of the medicines I have taken recently include them." The assistant only needs to match the allergy history in the health records with the current medication list, and can immediately provide a clear risk reminder.

This scenario demonstrates the application value of long-term personal memory in health management, and at the same time reflects MobileMem' s comprehensive evaluation capability for cross-source memory association, multimodal information fusion, long-term factual memory, temporal reasoning, and complex memory retrieval.

(2) Itinerary planning for travel abroad

> A foreign tourist plans to travel to Chengdu. In advance, they translated many travel notes and guides, and casually bookmarked several of them; the more they read, the more they were attracted by the scenery along the western Sichuan loop. On the night before departure, the user said to the assistant: "Based on everything I' ve bookmarked, what' s a good two-day route around Chengdu?" The assistant does not simply recommend Kuanzhai Alley and Jinli in a generic way, but instead, based on the contents of those bookmarked guides and in combination with the transportation methods the user searched for and their accommodation preferences, integrates a specific two-day itinerary. On the first day, depart from downtown Chengdu, pass through Dujiangyan and arrive at Yingxiu; on the second day, cross the Balang Mountain pass to see Siguniang Mountain, and return in the evening.

This scenario shows that long-term memory can not only support information retrieval, but can also support cross-source knowledge integration, user-preference modeling, multi-hop reasoning, and long-term personalized planning. MobileMem provides a real, continuous, and reproducible evaluation environment for such capabilities.

> **(3) Personal watchlist and drama-following records**
>
> A drama enthusiast is used to writing detailed reviews in notes every time, together with several memorable screenshots, and has long recorded viewing files for hundreds of film and television works. Today, after just finishing a new series, they feel that this work resonates with some of the works they have recorded before, but they cannot remember the specific details. The user says to the assistant: "Based on my previous drama-following records, organize all content related to this one and write a comparative long review." The assistant does not need to search online; instead, it browses through years of accumulated fragments of impressions in the notes, filters out works with similar styles or themes, compares them one by one from several angles, and generates a review with viewpoints, materials, and traces of personal memory.
>
> This scenario embodies the importance of long-term memory for interest modeling, cross-time knowledge association, and personalized generation, and also reflects MobileMem' s evaluation value in long-term knowledge organization, multimodal retrieval, and generative reasoning.
>
> **(4) Work review and mid-year summary**
>
> An office worker is preparing to write a mid-year summary. Over the past half year, they attended countless meetings, revised plans again and again, and read many technical articles and industry reports, but they always feel that what they learned is scattered and unclear. The user says to the assistant: "Help me review my work over the past half year.

In combination with the previous technical sharing session, Assistant then asks, "What am I mainly trying to fill in, and where am I getting stuck at work?" The assistant integrates the timeline of all meetings and project milestones in the calendar over the past six months, the records of successive presentations and draft revisions in documents, the fragmented thoughts in notes, and the themes of technical articles browsed in screen memory, distilling a clear main thread of learning. At the same time, by comparing work tasks with actual outputs, it identifies the links in which the user' s feedback has been stalled, and generates a well-structured midyear review summary.

This scenario demonstrates the application potential of long-term memory in knowledge management, continuous learning, and analysis of personal growth.

It also reflects MobileMem' s systematic evaluation value for core capabilities such as long-term temporal reasoning, cross-source information fusion, complex memory organization, and personalized summary generation.

Future outlook. The current version of the end-side memory benchmark focuses on long-term characteristics, and naturally supports ecosystem expansion toward broader personalized artificial intelligence. Future versions will incorporate memory sources from wearable devices, tablet computers, personal computers, smart home devices, and in-vehicle systems, thereby enabling unified assessment of full-ecosystem, lifelong memory. At the same time, we plan to expand multilingual coverage, increase the diversity of user groups, and introduce more challenging reasoning tasks involving long-term planning, collaborative memory, continuous personalization, and cross-modal knowledge integration. We hope that the mobile memory benchmark will develop into a comprehensive benchmark connecting academic research with industrial deployment, while promoting the development of trustworthy, personalized, and continuously evolving AI assistants.

Chapter 6 Roadmap for On-Device Memory

Long-term memory is an important foundational capability for artificial intelligence to move toward sustained cognition and autonomous evolution, and it is also a key infrastructure for the next generation of intelligent terminals. Over the next decade, the development of on-device memory will undergo a three-stage evolution: from device-cloud collaboration, to on-device autonomous intelligence, and then to trustworthy collaborative intelligence, driving artificial intelligence from single-device intelligence toward a sustained cognitive system that spans terminals, space, and time.

6.1 Device-Cloud Collaboration: Building the Infrastructure for On-Device Memory

For some time to come, device-cloud collaboration will remain the main form of on-device memory development. The device side is responsible for the continuous accumulation, organization, and invocation of personal long-term memory, while the cloud side undertakes foundation-model training, world-knowledge updating, and complex reasoning computation; together, the two form a unified long-term cognitive architecture.

Future device-cloud collaboration will evolve from traditional data synchronization into cognitive collaboration. The device side will continuously precipitate individual long-term experience and form user-oriented cognitive assets; the cloud side will continuously evolve general knowledge models, providing knowledge enhancement, capability upgrades, and model optimization for the device side. Through incremental model updates, incremental knowledge synchronization, and the exchange of memory summaries, capability sharing can be achieved as far as possible, reducing the sharing of raw personal data and continuously improving the capabilities of intelligent systems while ensuring that personal data remain locally controllable.

Future device-cloud collaboration will gradually form a new artificial-intelligence infrastructure characterized by "the evolution of public knowledge in the cloud and the precipitation of individual cognition on the device," providing a unified foundation for large-scale on-device intelligence.

6.2 On-Device Autonomy: Moving Toward Low-Power, Continuously Evolving Intelligence

With the continuous development of on-device computing power, storage media, and dedicated AI chips, long-term memory is expected to further reduce its dependence on cloud-side continuous computing resources and enter a stage of autonomous continuous evolution.

Future on-device memory systems will not only be able to continuously record user behavior, but will also autonomously complete memory formation, knowledge abstraction, experience compression, long-term consolidation, dynamic forgetting, and cognitive updating, continuously forming a more complete and more accurate individual cognitive model throughout the device life cycle.

This stage requires breaking through the key bottlenecks of long-term operation under resource constraints, establishing a lightweight continuous-learning system for on-device environments, and achieving ultra-low-power operation, long-term online service, adaptive resource scheduling, and cross-version continuous evolution, so that memory systems can continue to grow along with the long-term operation of devices rather than relying on periodic centralized training and model reconstruction.

Future on-device memory will evolve from "continuous recording" to "continuous growth,"becoming an important support for the long-term autonomous evolution of intelligent terminals.

6.3 Trustworthy Collaboration: Building a Distributed On-Device Memory Network

As personal digital spaces continue to expand, many types of terminals—mobile phones, tablets, PCs, smart glasses, wearable devices, smart vehicles, home robots, and others—will jointly support users' long-term digital lives. In the future, long-term memory will break through the boundaries of individual devices and gradually form a distributed end-side memory network oriented toward the full personal space.

Different terminals will form local memories based on their own perceptual capabilities, and will realize long-term cross-device cognitive sharing through trusted collaboration mechanisms. During collaboration, the exchange of raw private data should be avoided as much as possible; instead, cross-device knowledge sharing should be completed through privacy-preserving methods such as memory summaries, cognitive representations, and knowledge parameters, enabling each terminal to continuously acquire new cognitive capabilities while ensuring that personal data always remains stored locally.

Furthermore, multiple individual terminals will also explore the establishment of trusted memory collaboration mechanisms. Under the premise of user authorization and privacy protection, knowledge sharing and experiential collaboration among multiple entities—such as family members, research teams, and enterprise organizations—can be realized, exploring the possibility of extending from individual intelligence to collaborative intelligence.

In the future, a new generation of distributed cognitive networks will be formed that are individual-centered, device-autonomously collaborative, and privacy-trustworthy in sharing, providing a secure and reliable long-term cognitive infrastructure for an intelligent society.

6.4 Outlook: Toward Long-Term Cognitive Infrastructure

The development of end-side memory is not only an enhancement of terminal AI capabilities, but also represents an important evolution of the foundational architecture of artificial intelligence. In the future, long-term memory will gradually evolve from an application function into operating-system-level infrastructure; from the capability of a single device into cross-terminal collaborative capability; and from static data management into a mechanism for continuous cognitive evolution, ultimately forming a long-term cognitive network covering individuals, devices, and groups.

The new theories, new architectures, new standards, and new ecosystems formed around end-side memory will become an important direction for the future development of artificial intelligence. Taking the lead in breaking through key technologies for long-term end-side memory, building autonomously controllable memory infrastructure, and establishing an open industrial ecosystem will be expected to occupy the strategic high ground in the development of next-generation artificial intelligence and to promote the transformation of AI from generative intelligence toward continuous cognitive intelligence.

Acknowledgments

We hope that this white paper will provide a valuable reference for industry developers, corporate decision-makers, policymakers, and all colleagues concerned with the development of mobile intelligent agents. We sincerely thank all the institutions, organizations, and individuals who participated in the preparation of this white paper.

Collaborating organizations: OPPO Guangdong Mobile Communications Co., Ltd.; Chinese Open Knowledge Graph (OpenKG);

Core authors: Xue Yida, Deng Huanle, Ru Xiangyuan;

Data applications: Chen Yijun, Xu Buqiang, Mao Mingjun, Liu Xinhao, Zhang Chi, Zhong Ruobin, Fang Jizhan, Xu Haoming, Jiang Chen, Li Haibo, Chen Qiqi, Zhang Nandi, Chen Ming;

Legal compliance: Jia Fan, Lu Da, Ye Zilong, Wen Xiaohua, Lei Yunxia, Li Zhentong, Qu Yaohao;

Technical guidance: Jiang Xianchen, Wang Lizhong, Wang Jianfeng, Zeng Li, Zhang Xiaofan, Ju Minxin, Zhang Ningyu, Zhang Wen, Chen Yongrui, Sun Zequn, Deng Shumin, Wu Tianxing, Wang Meng, Chen Xiang, Bi Yong, Hu Wei, Wang Haofen, Qi Guilin, Chen Huajun;

Although every expert on our editorial committee has worked diligently to make the work as thorough and polished as possible, omissions and shortcomings in the white paper are nevertheless unavoidable. We sincerely welcome criticism and guidance from experts and colleagues in the industry.

Note: Figure translations are in progress. See original paper for figures.