

oppo



MOBILEMEM

开启端侧智能体记忆

新时代

目录

第 1 章 引言.....	3
1.1 人工智能发展的四次范式跃迁.....	3
1.2 基础模型不是终点.....	2
1.3 记忆是下一代人工智能的基础能力.....	2
1.4 什么是端侧记忆.....	3
1.5 端侧记忆的意义.....	5
第 2 章 科学问题.....	7
2.1 长程个体模拟机制.....	7
2.2 端侧记忆组织机制.....	7
2.3 持续认知演化机制.....	8
第 3 章 个人记忆层.....	10
3.1 定义.....	10
3.2 关键挑战.....	11
第 4 章 知识驱动的记忆合成.....	13
4.1 总体思路.....	14
4.2 MOBILEMEM 构建.....	15
第 5 章 基准评测与生态应用.....	18
5.1 数据集详情.....	18
5.2 产业生态应用.....	19
第 6 章 端侧记忆路线图.....	22
6.1 端云协同：构建端侧记忆基础设施.....	22
6.2 端侧自主：迈向低功耗持续演化智能.....	22
6.3 可信协同：构建分布式端侧记忆网络.....	22
6.4 展望：迈向长期认知基础设施.....	23
致谢.....	24

第 1 章 引言

1.1 人工智能发展的四次范式跃迁

人类文明的发展，本质上是信息处理能力持续演进的历史。从语言、文字、印刷术到计算机与互联网，每一次技术革命都不断提升信息获取、知识传播和社会协作的效率。人工智能则进一步赋予机器理解、生成和利用知识的能力，并推动智能系统能力不断演化。从感知、生成到行动，再到正在兴起的记忆，人工智能正经历一系列关键能力跃迁，由信息处理工具逐步演进为具备持续学习、长期认知和自主成长能力的智能体。

第一次跃迁是**感知智能**。这一阶段以卷积神经网络和深度学习的成熟为代表，人工智能系统首次具备了超越人类的视觉识别与语音感知能力，能够从图像、声音等非结构化数据中提取有效信息，实现了从“弱感知”到“强理解”的突破。

第二次跃迁是**生成智能**。以大语言模型和多模态生成模型为代表，人工智能从感知走向创造，能够生成高质量的文本、图像和音视频内容。这一阶段的标志性成就是模型在问答、内容创作与复杂推理上的卓越表现，标志着人工智能从“识别世界”进化为“参与创造”。

第三次跃迁是**行动智能**。大模型的能力边界从“回答问题”扩展到“执行任务”。智能体系统制定执行计划、调用外部工具与环境持续交互，实现了从“被动响应”到“主动行动”的跨越。人工智能不再是单纯的信息处理工具，而是能独立完成复杂任务的执行主体。

第四次跃迁正在发生，即**记忆智能**。当前最先进的人工智能模型虽然在生成与执行上表现出色，但其本质仍是无状态的。这体现在它每一次对话中都能够给出精准回答，却无法在时间维度上持续积累、更新并真正记住一个用户的偏好、习惯、关系网络与生活脉络。每一次交互都近乎从零开始，用户得不到源于过往经验的个性化服务。鉴于此，构建持续演化的长期记忆，是实现智能体持续学习和长期适应的关键。长期记忆使智能体能够不断积累经验、理解用户，并支撑长期规划、决策与执行，推动人工智能从一次性问答迈向真正具备长期认知能力的智能体。

随着基础模型能力趋于同质化，真正难以复制的不再是模型，而是在长期交互过程中形成的个体认知资产。因此，记忆正在成为继基础模型之后的新型人工智能基础设施。未来的竞争

格局将因此发生根本变化，技术壁垒将从模型参数规模、上下文窗口长度等通用指标，转向对个体用户记忆的构建效率、组织质量、更新时效与隐私安全保障能力。

1.2 基础模型不是终点

当前人工智能领域最先进的基础模型虽然在知识问答、内容生成与复杂推理上表现出色，但其本质仍是无状态的。每一次交互都被视为独立事件，模型能够在单轮对话中给出精准回答，却无法将本次交互中获取的信息沉淀下来，用于理解同一用户的下一次提问。这一局限主要体现在三个方面。

第一，没有长期经验。基础模型的知识来自训练阶段注入的静态语料，而非与特定用户持续互动形成的长期记忆。它知道世界的普遍规律，却不知道眼前这个用户昨天说过什么、上周做过什么、去年去过哪里。由于缺乏跨时间的经验积累，每一次对话几乎都是初见，每一次服务都需要重新建立上下文，无法提供真正连续、个性化的智能服务。

第二，没有持续成长。基础模型的主体能力在训练完成后基本固定，虽然能够借助上下文完成当前任务，却不会因为长期服务同一用户而不断优化自身。随着用户职业、兴趣、目标和社交关系不断变化，模型无法主动吸收这些变化、更新认知并调整行为，因此难以形成长期适应能力，更无法成长为真正的长程智能体。

第三，没有个体认知。基础模型理解的是人类的共性，而非这个人的个性。它知道如何回答一般性的旅行建议，却不知道用户偏好靠窗座位、厌恶海鲜、习惯早起。这种个体认知的缺失，使所有个性化服务都只能停留在用户主动设定的浅层偏好之上，无法触及真正的深度理解。

基础模型提供的是跨用户共享的通用智能，记忆系统形成的是面向个体持续演化的长期认知。前者让智能体能够理解当前任务，后者让智能体能够理解任务背后的用户与历史。二者共同构成人工智能从一次性交互走向长期服务的完整能力体系。

1.3 记忆是下一代人工智能的基础能力

如果说基础模型解决的是空间维度上的知识覆盖，那么记忆系统解决的则是时间维度上的知识积累。基础模型能够利用已有知识完成当前任务，而记忆系统能够在长期交互中持续形成、更新和演化对用户的认知，使智能体真正具备跨时间学习与持续适应的能力。人工智能迈向长期智能，本质上是在时间维度建立持续演化的认知能力。

未来人工智能的竞争焦点也将由模型能力逐步转向记忆能力。当前的竞争主要围绕模型参数规模、上下文窗口和基准测试性能展开，这些指标反映的是通用智能水平；而长期智能服务

的核心，则在于智能体能否持续理解同一用户，在长期交互中形成稳定、准确且可演化的个体认知。

随着基础模型能力快速趋于同质化，模型本身正逐渐成为通用基础设施。开源与闭源模型之间的性能差距不断缩小，参数规模和上下文长度带来的边际收益持续下降。当高性能模型成为普遍能力后，真正难以复制的不再是模型本身，而是智能体在长期服务过程中形成的用户记忆，以及由此建立的持续认知能力。

因此，下阶段人工智能竞争的一个重要方向，将不只是模型规模和推理能力本身，还包括能否更精准、更高效、更安全地构建、组织、更新和利用长期记忆。记忆将成为继感知、生成和行动之后，支撑人工智能持续演化和个性化服务的核心基础能力。

1.4 什么是端侧记忆

端侧记忆是指在用户终端设备上持续积累、组织、更新和调用个人长期经验的记忆系统。其核心目标不是简单保存历史数据，而是在尽可能保证用户数据本地可控的前提下，为智能体提供与具体用户相关的长期认知能力。

将记忆系统部署在何处，并非单纯的工程选型，而需要综合考虑长期记忆的伴随性、连续性、隐私性以及端侧资源约束。长期记忆需要随用户日常使用持续积累和演化，也需要控制个人数据的流转范围。因此，手机、电脑、可穿戴设备、汽车、家庭设备和具身机器人等终端设备，是承载个人记忆的重要位置。基于这一特性，一个完整的端侧记忆系统需要同时关注系统性、连续性、长期性、隐私性和效率性：系统化组织用户认知，持续积累与演化长期经验，在保障数据本地可控的前提下，以较低的词元（Token）开销、存储压力、延迟和功耗，为智能体提供稳定、可信的长期认知能力。

系统性。记忆不是孤立片段的堆砌，而是一个有机整体。用户的偏好、习惯、关系网络与生活脉络彼此交织，构成一张动态演化的意义网络。系统必须具备结构化组织与关联能力，使新信息能够被已有认知框架吸纳，同时反向重塑旧有结构，形成螺旋上升的认知循环。

连续性。记忆的价值在于时间维度上的持续积累。用户的偏好、习惯、关系网络与生活脉络，只有在数月甚至数年的跨度上才能形成有意义的认知。这要求记忆系统始终在线、始终可用，不能因为网络中断或服务切换而重置。

长期性。真正有价值的认知并非来自一次交互，而是在长期积累中逐渐形成。记忆系统不仅需要长期保存信息，更需要持续更新、整合、抽象和演化，使认知结构随着时间不断完善，而不是简单堆积历史数据。

隐私性。个人记忆天然包含高度敏感的信息，它包含了用户的社交关系、健康状态、财务状况、位置轨迹与消费习惯。这些信息一旦脱离终端设备进入云端，无论加密措施多么完善，都从根本上改变了数据的控制权关系与暴露面。

效率性。端侧记忆必须满足资源受限条件下的高效运行要求。长期记忆会持续积累大量交互内容，如果直接将全部历史记录输入基础模型，将带来不可接受的词元开销、延迟、功耗和存储压力。因此，端侧记忆系统需要在本地完成必要的信息筛选、压缩、摘要、索引和分层管理，只在任务相关时激活少量高价值记忆，避免记忆规模随交互轮次线性膨胀。效率性决定了端侧记忆能否从短期演示走向长期稳定运行。

综合来看，目前只有端侧设备能够同时满足上述五项要求。终端设备始终伴随用户，能够持续感知真实世界中的行为与交互；数据留存在本地，有助于增强用户对记忆的访问、管理和删除控制；即使网络中断，记忆仍然能够持续积累，不会因为服务迁移而中断。因此，端侧不是一种部署偏好，而是记忆智能的重要载体。

在满足上述基础条件之后，一个真正可用的记忆系统还需要具备两项关键能力：一是鲁棒性，二是遗忘与激活能力。

鲁棒性。真实世界充满差异，不同文化、语言、生活方式和社会关系都会影响记忆的表达形式。记忆系统需要保持稳定的组织、检索和推理能力，使其能够在不同用户群体和应用场景下持续提供准确、一致、可追溯的记忆服务。鲁棒性不是额外能力，而是长期运行的基本要求。

遗忘与激活。记忆的效率并不来自无限存储，而来自合理管理。系统需要主动遗忘冗余、过时或冲突的信息，避免认知不断膨胀；同时能够依据当前情境主动激活相关记忆，在恰当的时机提供最有价值的信息。遗忘让记忆保持轻盈，激活让记忆保持鲜活。

如图 1-1 所示，终端设备伴随用户全天候活动，天然承载着更接近真实使用场景的行为数据；数据尽可能保留在本地，有助于从源头降低隐私泄露和跨平台流转风险；无需依赖网络，记忆能够在连续的时间轴上不断积累、更新和演化。正因如此，端侧记忆智能体能够逐渐形成对用户的长期认知，实现“越用越懂用户，越用越贴近用户”的持续进化能力，并



图 1-1 移动记忆的多场景示例

最终支撑真正的长程智能体。这种从一次性回答走向持续成长的能力，将成为下一代智能助手区别于前代产品的根本特征。

1.5 端侧记忆的意义

端侧记忆是人工智能从通用能力走向个体智能的关键基础能力。基础模型赋予人工智能广泛的知识、推理与生成能力，使其具备处理复杂任务的通用智能基础；而端侧记忆则进一步将这些通用能力与具体用户的长期经历相结合，使智能系统能够在持续交互过程中形成对个体的理解。端侧记忆的核心并非简单存储历史信息，而是在用户侧持续积累、组织和更新与个体相关的经验知识，使智能系统能够根据长期交互不断优化自身行为，并提供更加贴合用户需求的服务。

首先，端侧记忆将深刻改变人工智能服务个人的方式。传统智能系统主要依赖当前输入进行响应，不同交互之间缺乏连续性，难以真正理解用户长期形成的行为模式和偏好。而具备长期记忆的智能系统能够综合用户过去的行为、习惯、偏好以及历史经历，对当前需求进行更加准确的判断。例如，在日常生活场景中，系统不仅能够回答用户当前的问题，还能够结合长期积累的经验理解用户通常的决策方式和个性需求。由此，人工智能服务将逐渐从基于单次交互的即时响应，转变为面向个体长期陪伴和持续优化的智能服务模式。

进一步而言，端侧记忆为智能体实现持续学习和长期适应提供了重要基础。一个能够自主完成复杂任务的智能体，不仅需要具备规划和执行能力，还需要能够从过去的行动中积累经验，并在未来任务中加以利用。哪些策略曾经有效，哪些尝试导致失败，用户曾经提出过哪些需求，以及特定环境中存在怎样的约束，这些经验都需要被持续保存、总结和调用。对于具身智能和机器人而言，这种能力尤为关键。数字智能体主要面对信息空间中的任务，而机器人则需要长期运行于真实、开放且动态变化的物理环境中。家庭机器人需要逐渐理解家庭成员和生活环境，服务机器人需要适应不同用户的长期需求，工业机器人也需要积累特定设备和生产流程中的操作经验。因此，端侧记忆将成为推动具身智能从具备动作执行能力走向具备环境适应和经验积累能力的重要支撑。

与此同时，端侧记忆的大规模应用也带来了新的安全与治理挑战。随着记忆内容不断丰富，智能系统能够提供更加精准的服务，但同时也可能涉及大量个人隐私和敏感信息。因此，未来的记忆系统不仅需要关注信息的有效存储和利用，更需要解决记忆访问权限、内容管理、动态删除以及用户自主控制等关键问题。将个人记忆尽可能保留在用户侧，有助于提升用户对数据的掌控能力，减少敏感信息在不同平台之间流转带来的风险。由此，端侧记忆不仅是人工智能能力提升的重要方向，也是构建可信、可控、以人为中心的个人 AI 系统的重要基础。

从更长远的视角来看，端侧记忆的意义在于为人工智能引入了持续演化的时间维度。基础模型使人工智能具备了对世界知识的广泛理解，而记忆使其能够积累与特定个体、特定环境相关的长期经验。前者回答人工智能如何理解世界，后者决定人工智能如何理解一个具体的人以及如何在长期互动中不断适应变化。二者结合，将推动人工智能从依赖预训练知识的通用模型，进一步发展为能够持续学习、主动适应并与用户共同成长的个体化智能系统。未来的人工智能竞争，不仅在于模型规模和推理能力，更在于系统是否能够长期理解用户、积累经验并形成独特的个体智能。端侧记忆因此有望成为连接基础模型与个人智能之间的关键桥梁，推动人工智能进入以长期陪伴、持续适应和个性化进化为核心的新阶段。

第 2 章 科学问题

端侧记忆的构建，本质上不是开发一种新的信息存储机制，而是探索人工智能如何在长期时间尺度上形成类似人类的持续认知能力。当前基础模型主要依赖大规模预训练获得静态世界知识，其能力边界由训练数据和模型参数决定，难以在与真实用户长期交互过程中持续积累经验、形成个体理解并实现自主成长。

人类智能的核心优势并不只是拥有丰富知识，而是在多年生活经历中不断记录事件、总结经验、形成认知，并随着环境变化持续调整自身行为。相比之下，现有人工智能仍然缺乏对个体长期经历的建模能力，无法模拟真实用户多年尺度上的行为演化过程，也无法将碎片化、多模态、动态变化的个人经历转化为稳定且可持续更新的认知体系。

因此，端侧记忆需要突破现有人工智能静态知识范式，建立面向长期个体智能的新型认知机制。下面将围绕**长程个体模拟**、**端侧记忆组织**、**持续认知演化**这三个关键科学问题展开介绍。

2.1 长程个体模拟机制

人类对世界和自身的理解来源于长期生活经历，而非孤立事件的积累。一个人的兴趣、习惯、价值观和行为模式，往往需要经过数年甚至更长时间的观察和交互才能逐渐形成。然而，当前人工智能主要通过离线数据训练获得能力，缺乏模拟真实个体长期经历和行为演化的机制。面向端侧智能，需要探索如何构建跨越月份、年份甚至更长时间尺度的个体经历模拟环境，使人工智能能够经历类似真实用户的连续生活过程。在这一过程中，系统不仅需要生成合理的用户行为轨迹，还需要建模用户目标变化、兴趣迁移、环境影响以及人与智能体之间长期互动关系。

长程个体模拟的核心挑战在于：真实人的行为具有高度复杂性和不确定性，短期行为与长期特征之间存在显著差异，偶然事件、环境变化和个人成长都会影响未来行为。因此，需要建立能够反映个体长期变化规律的模拟机制，使人工智能能够从长期经历中学习如何理解和预测个体。

因此，需要回答的基础科学问题是：

如何突破现有人工智能短时交互范式，构建面向多年尺度的个体经历模拟机制，使智能体能够通过长期交互形成对用户行为、偏好和环境变化的深层理解？

2.2 端侧记忆组织机制

长期个体智能不仅需要经历，更需要有效组织经历。真实用户在端侧产生的数据具有高度异构性，包括文本交流、图像、视频、位置轨迹、应用行为、传感信息以及环境反馈等多模态信息。这些信息并非简单的数据记录，而是构成个人生活经验的重要组成部分。当前人工智能中的记忆机制主要围绕文本上下文或信息检索展开，难以处理长期、多模态、动态变化的个人经历。如何从海量碎片化数据中识别关键事件，建立事件之间的时间关联、因果关系和语义联系，并形成结构化、可调用的个人记忆体系，是实现端侧智能的核心挑战。

端侧多模态记忆组织的本质，是将原始经历转化为可理解、可推理、可迁移的认知表示。这一过程涉及记忆选择、信息压缩、语义抽象、关系建模以及跨模态融合等关键机制。同时，系统需要区分短期状态与长期特征，识别偶然事件与稳定规律，避免无效记忆积累导致认知混乱。

因此，需要回答的基础科学问题是：

如何从长期、多模态、动态变化的个人经历中提取关键经验，建立能够支持理解、推理和决策的端侧记忆组织机制，实现从数据记录到个体知识形成的跨越？

2.3 持续认知演化机制

记忆形成并不是个体智能的终点，而是持续成长的基础。人的认知并非固定不变，而是在不断经历、反馈和反思中持续调整。未来端侧智能同样需要具备长期演化能力，使其能够随着用户变化、环境变化和任务变化不断更新自身认知。持续认知演化的核心，是研究人工智能如何在有限资源约束下实现类似生命系统的学习与适应过程。智能体需要自主决定哪些经历值得保留，哪些知识需要更新，哪些经验应该遗忘，并通过长期反馈不断优化自身行为模式。

然而，端侧环境具有计算资源有限、数据持续增长、用户需求不断变化等特点。如何避免知识冲突、认知漂移和长期积累导致的效率下降，同时保证记忆更新过程的安全、可控和可信，是持续认知演化面临的重要挑战。

因此，需要回答的基础科学问题是：

如何突破人工智能静态模型范式，使智能体能够在长期交互过程中自主更新、重组和优化认知结构，实现面向个体的持续认知演化？

总体而言，端侧记忆研究的核心目标不是让人工智能简单地记住更多信息，而是探索人工智能从静态知识智能走向动态经验智能的基本规律。长程个体模拟解决智能如何获得长期经历的问题，端侧多模态记忆组织解决经历如何转化为认知的问题，持续认知演化解决认知如何不

断成长的问题。三者共同推动人工智能从依赖预训练知识的通用模型，迈向能够理解个体、适应环境并持续发展的新一代个体智能系统。

未来人工智能的重要突破，不仅在于拥有更大的模型规模和更强的推理能力，更在于是否能够像人类一样通过长期经历形成独特认知。端侧记忆因此可能成为连接基础模型与个体智能的关键理论基础，推动人工智能从一次性交互工具演进为能够长期陪伴、持续学习和共同成长的智能体。

第3章 个人记忆层

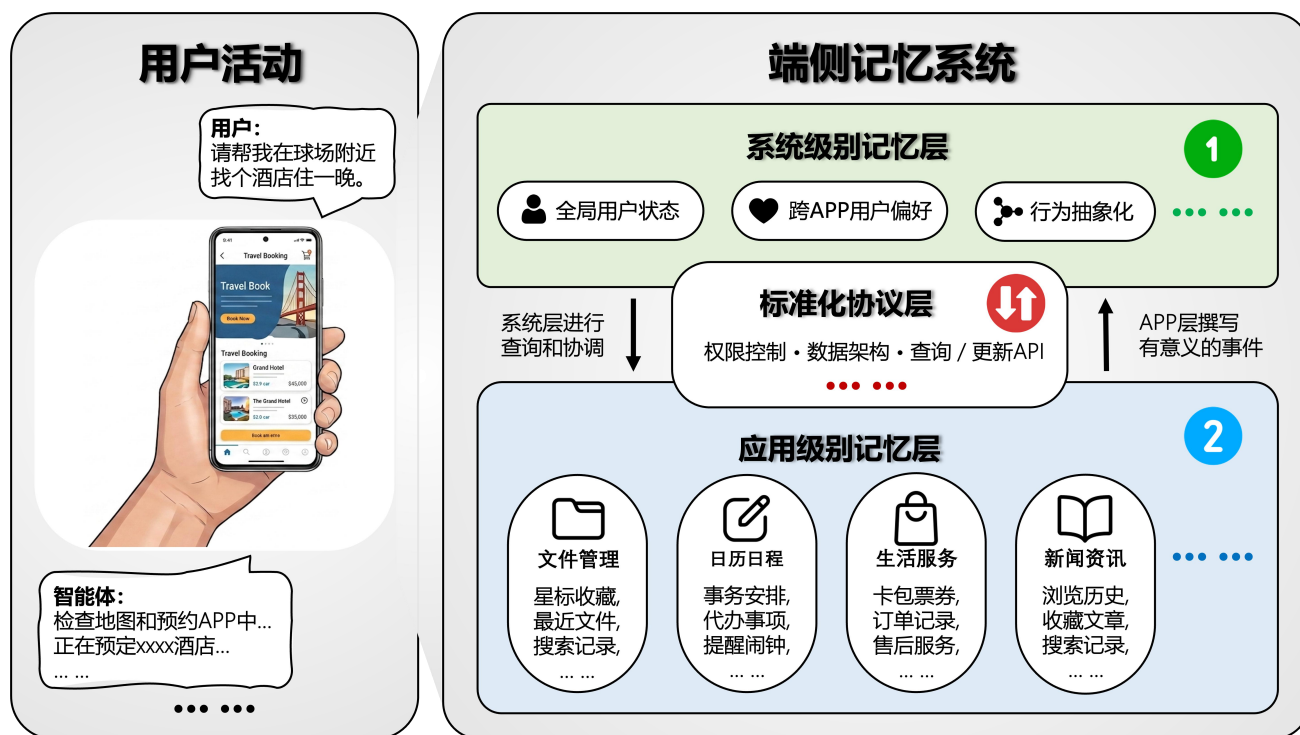


图 3-1 个人记忆层架构图

本章从科学问题角度指出，端侧记忆需要解决长期经历如何获得、如何组织以及如何持续演化的问题，重点讨论个人记忆层如何在终端系统中承接多源行为数据，并将其组织为可检索、可更新、可推理的长期记忆。换言之，个人记忆层是端侧多模态记忆组织机制和持续认知演化机制的系统实现载体。

3.1 定义

个人记忆层是终端操作系统中负责用户认知数据管理的核心中间件，承担行为感知、记忆组织与智能检索的系统级职能。如图 3-1 所示，在端侧智能体的系统架构中，个人记忆层被明确定义为操作系统层级的基础中间件，承担着用户行为数据与智能体推理决策之间的关键桥梁角色。其核心功能为：持续采集来自用户主动交互与系统被动观测的行为信号，在终端本地执行信息的编码、组织、更新与持久化存储，并在智能体发起请求时，提供精准、高效且具有上下文感知能力的记忆检索服务。这一设计使个人记忆层成为终端智能体实现持续认知与个性化服务的基础。

个人记忆层的提出，标志着记忆能力从应用功能向操作系统级基础设施的跃迁。当前主流操作系统的存储管理以文件系统和数据库为核心，服务于应用的数据持久化需求，是一种被动

的、静态的仓库式管理。而个人记忆层以智能体对用户的认知为组织线索，强调信息的关联、抽象与演化，是一种主动的、动态的认知式管理。这一转变呼应了操作系统设计理念的深层演进。当前，业界主流操作系统正经历从交互工具向智能体的范式转变，系统级人工智能运行时、异构计算架构与端云协同模型矩阵成为构建端侧智能的核心能力。在这一趋势下，个人记忆层作为智能体操作系统的关键子系统，为上层智能体提供持续积累、动态演化的认知基础设施。正如操作系统管理计算机的硬件与软件资源，个人记忆层管理的是关于用户的认知资源，通过将碎片化的行为数据编织为连贯的认知网络，从而把静态的事实记录转化为动态的知识资产。

从系统边界来看，个人记忆层的输入是异构的、持续产生的行为流。这些行为可能由用户主动发起，如与智能助手的对话、在各类应用中的操作记录；也可能由系统被动观测获得，如屏幕内容的变化、通知栏信息的呈现、传感器数据的采集。在真实移动环境中，用户和智能助手不仅彼此交互，还广泛与第三方应用发生联系，这使得行为流天然具有高度异构性，涵盖自然语言指令、应用操作事件、界面状态快照以及各类应用特定的交互记录。

个人记忆层的输出则是结构化的、可被智能体直接调用的记忆内容。这些内容需要支持多种访问方式，包括基于语义相似度的隐式检索、基于实体关系的显式查询，以及基于时间范围的过滤与排序。输出的形式可以是事实陈述、事件摘要、偏好推断、流程指导或关系描述，具体形态取决于下游任务的需求。

个人记忆层与传统软件系统所依赖的存储体系存在本质区别。传统存储系统以文件或记录为单位，强调数据的持久保存与精确还原，是一种被动的、静态的仓库式管理。而个人记忆层以智能体对用户的认知为组织线索，强调信息的关联、抽象与演化，是一种主动的、动态的认知式管理。后者不仅需要回答“发生过什么”的事实性问题，更需要支撑“用户是什么样的人”、“用户可能会需要什么”这样的判断性问题。

3.2 关键挑战

个人记忆层在真实移动环境中的落地部署面临一系列尚未被系统性解决的挑战，这些挑战贯穿数据接入、记忆表示、认知推理与系统部署全链路。

异构数据的统一接入与组织。用户行为分散于通讯、社交、影音、出行、购物、健康等大量应用，数据涵盖文本、图像、视频、语音、位置轨迹及传感器信息等多种模态，具有格式异构、语义割裂和时间离散等特点，形成大量信息孤岛。如何在端侧环境下实现跨应用、跨模态数据的统一接入、语义对齐与时序组织，是个人记忆层构建的首要挑战。

长期记忆的统一表示与持续演化。个人记忆需要统一表示事实、事件、任务和用户偏好等不同类型的信息，同时支持持续积累、动态更新和合理遗忘。随着用户行为不断变化，系统既要避免灾难性遗忘，又要淘汰过时信息并控制存储规模，实现长期稳定演化。此外，不同类型记忆具有不同更新频率和存储需求，统一表示与高效管理之间存在天然矛盾。

从信息存储到用户认知理解。记忆系统的价值不仅在于记录和检索，更在于形成对用户的持续理解。智能体需要从离散事件中抽象用户意图、行为模式和长期偏好，建立跨时间、跨任务的关联推理能力，为长期规划、决策与个性化服务提供支撑。

端侧资源约束下的系统优化。上述能力均需部署于移动终端，受限于词元消耗、存储、延时、功耗、算力和内存等。传统的记忆系统不加过滤地处理全部交互数据，导致资源消耗随轮次线性增长，在端侧被急剧放大，需要设计轻量化存储、分层记忆管理、增量更新和高效推理机制，在资源受限条件下实现长期稳定运行。

上述挑战贯穿数据获取、记忆管理、认知理解和系统部署四个层面，并相互耦合、彼此制约。数据组织影响记忆表示，表示方式决定持续演化能力，而认知能力又受限于端侧资源约束。因此，个人记忆层的构建需要从系统架构层面对数据、记忆、认知和计算资源进行协同设计，而非依赖单一算法或局部优化。

现有记忆方法的范式局限。当前端侧记忆研究总体仍沿袭信息存储与按需检索的外挂式记忆范式，即将智能体产生的交互内容视为独立于模型认知过程之外的外部数据，通过摘要、向量化、知识图谱或检索增强等方式进行保存，并在任务需要时重新注入模型。这一范式本质上将记忆视为模型能力之外的附加存储空间，主要解决如何记住和如何找到的问题，却难以回答记住什么、如何组织、何时更新以及如何由经历形成持续认知等更深层次问题。尤其在长期、连续的真实移动环境中，用户经历并非相互独立的静态信息，而是随时间不断累积、关联、修正和演化的连续过程；简单的记忆追加与检索容易造成信息冗余、语义碎片化以及新旧知识冲突，使记忆难以形成稳定的用户认知。因此，个人记忆层需要突破以检索为中心的传统范式，从为模型增加一个记忆库转向为智能体建立持续认知层，将记忆从被动存储机制提升为连接经历获取、记忆组织、认知形成与持续演化的主动系统，使智能体能够在长期经历中不断形成、修正和更新对用户的理解。

第 4 章 知识驱动的记忆合成

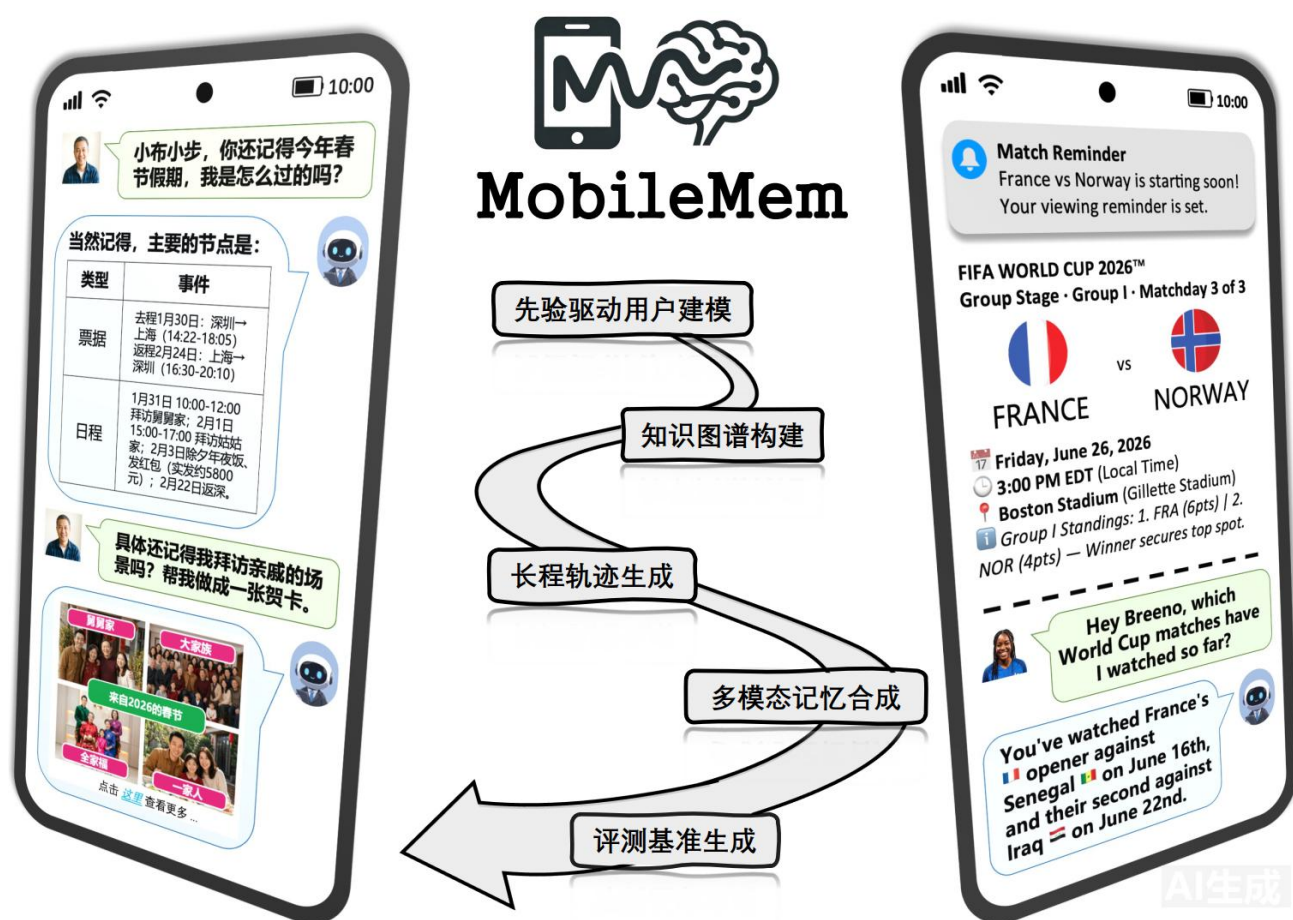


图 4-1 知识驱动的记忆合成方法示例

在真实移动环境中构建可评测的端侧记忆系统，面临一个根本性的数据难题：用户日常产生的行为数据虽然丰富，却天然具有碎片化、多源异构与高度私密的特征，既难以直接用于系统开发与评测，也无法简单拼接为具有时间连贯性的长期轨迹。为解决这一问题，我们提出知识驱动的记忆合成方法，构建具有长期时间跨度、人物关系约束和多模态交互特征的模拟个人经验，并在此基础上形成端侧记忆评测基准 MobileMem。该方法主要回应第 2 章中的长程个体模拟机制问题，同时为第 3 章所述个人记忆层的组织、检索、更新和推理能力提供可复现的评测数据。

如图 4-1 所示，知识驱动的记忆合成方法以个人知识图谱构建为核心锚点，按照先验驱动用户建模、知识图谱构建、长程轨迹生成、多模态记忆合成、问题构造的流程逐步展开，将用户静态画像、历史行为与交互信息转化为具有时间连贯性、语义一致性和状态可演化性的长期

交互轨迹，为端侧记忆系统的系统化评估提供可控且可复现的数据基础，推动端侧记忆研究从数据构造方式不统一、评价标准不一致的探索阶段，逐步走向统一标准、可比较、可复现、可迭代的规范化发展阶段。

知识驱动的记忆合成方法不仅用于端侧记忆基准测试的数据构造，更重要的是为端侧记忆系统提供从数据积累走向持续理解与长期认知演化的技术路径。该方法围绕长程时序轨迹构建、跨模态知识对齐、用户状态持续演化三类核心能力展开，推动端侧记忆由静态信息存储向动态、结构化、可推理的认知系统演进。

第一，长程时序轨迹构建。面向真实用户长期使用过程中信息分散在不同时间、不同应用和不同交互载体的问题，以事件、会话、轨迹为基本组织单元，对用户多轮交互中的文本对话、应用操作记录、系统行为以及屏幕截图等多源信息进行统一建模与时序对齐，形成具有清晰时间结构、事件关联和上下文连续性的长程交互轨迹。通过将离散交互行为组织为连续事件链条，使系统能够完整刻画用户行为的发生过程及其上下文演变，为长期记忆的存储、更新与检索提供结构化基础。

第二，跨模态知识对齐。面向端侧记忆中多模态信息并存且语义相互关联的复杂场景，通过个人知识图谱构建统一的知识组织框架，将不同模态信息映射到统一的事件与关系结构中，实现视觉、语言与行为信息之间的语义对齐与关联建模。在此基础上支持跨模态信息的联合理解与推理，使分散在不同模态中的信息能够相互补充与相互验证，从而提升端侧记忆对复杂用户经历的整体理解能力。

第三，用户状态持续演化与个性化适应。面向长期交互过程中用户偏好、行为模式、社交关系及事件状态持续变化的动态场景，将用户记忆建模为随时间不断更新与演化的结构化知识体系。通过持续吸收新的交互信息，动态更新事件状态、用户偏好及关系结构，并对历史信息进行修正、融合与必要的遗忘，使用户模型能够随真实使用过程持续演化，从而实现对用户长期状态的动态刻画与个性化适应，推动端侧记忆由一次性构建的静态知识库向持续学习、持续更新、持续适应的个体认知系统演进。

4.1 总体思路

端侧记忆系统的构建以个人信息建模为起点，将用户碎片化的行为数据抽象为可被智能体理解与推理的结构化认知表示。在个人信息建模的基础上，知识驱动的记忆合成方法将用户画像转化为具有时间连贯性与语义一致性的长期交互轨迹。如图 4-2，其基本流程以用户先

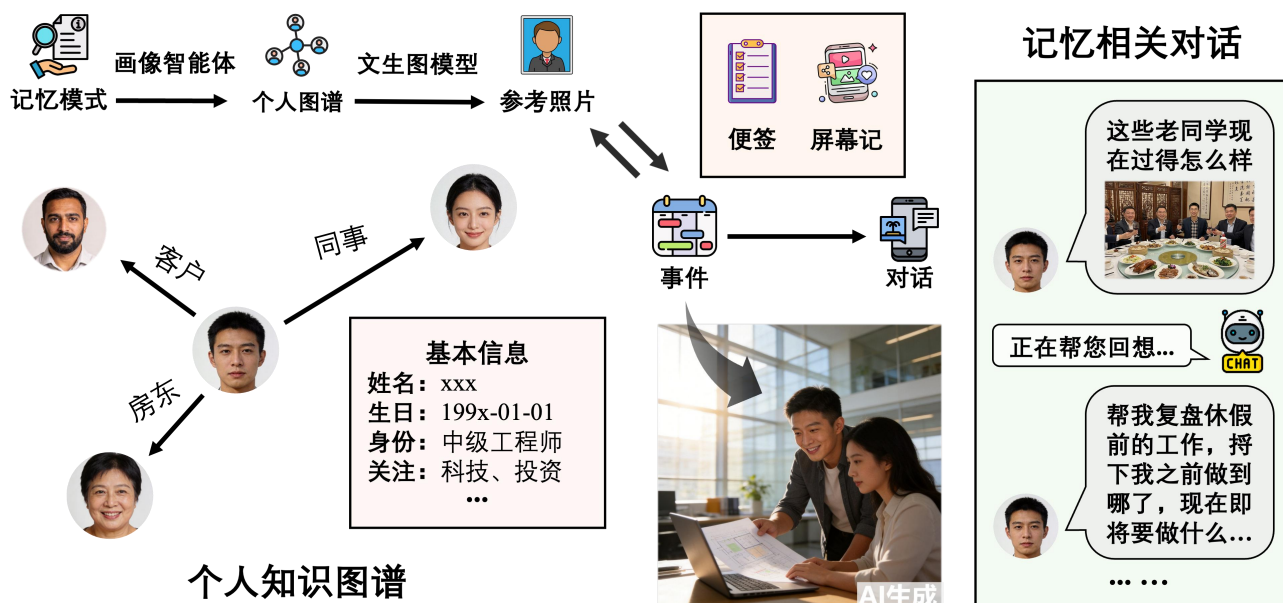


图 4-2 MobileMem 的构建方式示意图（此图为多模态 Omni 版本的示意图）

验知识构建个人知识图谱为起点，将用户的静态画像、社会关系与重要里程碑组织为以用户为中心节点、以频繁联系人作为外围节点的个人知识图谱，作为后续所有生成的语义锚点，确保生成内容在人物关系、时空背景与生活逻辑上保持一致。基于这一知识图谱，系统生成锚定于个人与文化重要日期的年度事件序列，每个事件进一步分解为时间有序的子事件步骤，标注可能涉及的移动应用与视觉素材类型。核心的模拟过程发生在对话合成阶段。每个子事件步骤被分解为细粒度的记忆点，代表用户与人工智能交互过程中需要保留的关键信息片段。记忆点进一步生成对话，所有对话片段按时间顺序拼接，形成完整的移动交互轨迹，为端侧记忆系统的评估提供可控、可复现的数据基础。

4.2 MobileMem 构建

为系统评估端侧记忆系统在长期移动交互中的核心能力，我们构建了覆盖文本与多模态场景的端侧记忆基准评测体系。整体框架以用户先验知识为基础，通过知识驱动的长程经验合成构建具有时间连贯性和语义一致性的用户交互轨迹，并进一步从轨迹中自动生成多层次评测问题。整个流程由四个模块组成：用户先验知识构建、长程交互轨迹合成、问答对合成与质量控制。其中，知识驱动的记忆合成框架是整个数据构建过程的核心，用于将分散的用户知识与已有交互组织为持续演化的长期经验，并在此基础上形成用于记忆评测的数据。

先验驱动用户建模。直接依赖大语言模型生成完整用户数据，容易产生用户差异不足、个性化程度有限以及长期行为缺乏约束等问题。因此，我们以先验驱动用户建模，为后续长期轨

迹合成提供稳定的身份、状态与行为约束。根据数据来源和模态范围的不同，文本基准 MobileMem 与多模态扩展版本 MobileMem-Omni 采用不同的先验知识构建策略。对于涉及真实用户的数据，我们均进行必要的匿名化与隐私脱敏处理。

长程轨迹生成。给定用户先验知识，我们的目标是生成能够跨越较长时间范围、持续演化且内部一致的用户交互轨迹。直接一次性生成完整轨迹容易造成事件之间缺乏约束，并难以保证长期时间线的一致性。因此，我们将长程轨迹构建分解为由粗到细的层次化合成过程，并将已有用户知识作为持续约束。

我们提出知识驱动的记忆合成框架。其核心思想是将用户先验知识与已有交互作为知识锚点，在用户画像和时间约束的共同引导下，将分散的知识组织为统一的长期交互流，并随着用户经历逐步展开而持续生成和修正后续经验。由此，轨迹并非静态拼接，而是在已有知识与新生成经验的共同作用下逐步扩展和演化。具体而言，该框架采用自顶向下的知识引导与自底向上的经验演化相结合的思路。首先，从用户初始画像和时间范围出发，将完整时间跨度划分为不同粒度的生活事件，并递归细化为具体交互会话；在此过程中，将已有知识锚点映射至相应事件，并将其作为不可违背的上下文约束。随后，在叶节点生成具体的人机交互，新生成的经验进一步反向影响尚未展开的未来事件，从而动态调整事件结构、时间关系和用户状态。该过程形成规划、锚定、合成、修正的持续反馈回路，使生成轨迹能够同时满足知识约束、时间一致性和经验演化。

MobileMem 在上述统一框架下采用不同的具体实现。MobileMem 主要保留知识锚点和文本交互，并完整使用层次化事件规划与经验反馈机制；MobileMem-Omni 则根据预定义的年度事件序列进行简化的事件展开，并进一步加入参与者、移动应用和视觉内容等多模态约束。由于 Omni 轨迹规模较大，其完整轨迹长度可超过百万级词元，因此在保证前后事件通过历史上下文保持一致的前提下，省略显式的底层反馈修正，以降低数据合成成本。

评测基准生成。在获得长程交互轨迹和多模态记忆后，我们进一步从轨迹及其对应的用户知识中自动构建评测问答对。由于轨迹本身具有层次化结构，我们采用自底向上的问题合成策略，以充分利用不同时间尺度和不同粒度的记忆信息。具体而言，从预定义的问题类型集合出发，在叶节点处基于具体用户属性或交互会话生成基础问答对；在内部节点处，则以子节点生成的问题作为构建单元，进一步组合形成需要跨属性、跨事件或跨会话进行信息整合的问题。未被使用的子节点问题与新生成的复杂问题一同向上传递，使问题能够随着层次结构逐步扩展，

从单点事实检索发展到跨会话多跳推理。在合成过程中，问题类型集合还可以根据新出现的推理模式动态扩展。

数据质量控制。为保证合成数据的可靠性，我们建立覆盖轨迹、视觉内容和评测问题的多层次质量控制流程。对于长程交互轨迹，由于现有自动指标难以可靠衡量跨长时间跨度的全局一致性，我们结合结构化约束与人工抽样审核进行质量控制。自动合成过程中，通过时间约束、事件层级约束、知识锚点一致性以及上下文验证减少明显冲突；随后对轨迹片段进行人工抽样检查，重点关注事件之间的时间关系、用户状态变化以及前后交互是否存在明显矛盾。对于少量类似真实用户数据中可能出现的数值偏差等自然噪声，在不影响整体语义一致性的情况下予以保留。

对于多模态内容，我们进一步进行专门的图像质量控制。首先利用多模态大语言模型检查图像与对应视觉描述之间的语义一致性，并过滤明显不匹配、不安全或视觉质量较低的图像。对于涉及用户或个人知识图谱中人物的图像，进一步利用人脸识别模型将生成结果与对应参考人物进行相似度比对，低于阈值的图像重新生成；经过多轮生成仍无法满足要求的样本直接丢弃。自动过滤后，再通过人工抽样检查视觉内容与其所在交互轨迹之间的一致性。

对于评测问答对，首先利用大语言模型验证问题是否能够由源记忆中的证据充分回答，并检测高相似度、信息不足或答案错误的问题。低质量和冗余样本进行重写或删除，最终通过人工抽样审核进一步验证问题的正确性与证据充分性。对于部分由自动组合产生的多跳问题，可能存在将多个相对独立的单跳问题进行组合的现象。尽管这类问题在自然性上存在一定局限，但其答案仍具有明确的证据支撑，并能够有效测试记忆系统跨多个信息片段检索和整合证据的能力。

通过上述流程，我们最终构建了覆盖文本与多模态移动交互、兼顾长期时序一致性与多层次记忆能力的端侧记忆基准，为不同记忆系统在真实移动环境下的统一、可复现评估提供基础。

第5章 基准评测与生态应用

移动记忆基准的构建与评测体系的建立，为端侧记忆研究提供了标准化的数据基础与评估工具，使系统能力从定性展示走向可度量、可比较、可持续提升。在此基础上，移动记忆基准已深度融入终端人工智能的产品化实践，在产业生态、标准化、端侧智能、用户价值与具体应用场景等层面产生了实质性影响。

5.1 数据集详情

表 5-1 展示了 MobileMem-Omni（多模态版本）的数据统计。该数据集包含均衡的英语和中文用户，每位用户关联大量跨越较长时间跨度的事件。在交互统计方面，该基准每位用户具有较长的上下文长度，每位用户包含多个会话，每个会话包含多轮对话。大量图像分布在各个会话中，反映了真实世界移动使用的多模态特性。问题类型分布在七个推理类别中，能够全面评估不同的记忆能力。

表 5-1 MobileMem-Omni 的统计信息

统计类别	指标	数值
用户统计	用户总数	16
	英文交互用户	8
	中文交互用户	8
	事件总数	1,589
交互统计	平均上下文长度（词元/用户）	172 万
	平均会话数（每用户）	202.6
	平均对话轮次（每会话）	48.2
	对话轮次总数	155,670
	平均图像数（每会话）	5.88
	图像总数	19,060
问题类型分布	问题总数	7,415
	单跳检索	986
	多跳推理	1,135
	知识更新	1,010
	时序推理	773
	隐式偏好推断	1,226
	拒答	1,208
	视觉推理	1,077

如表 5-2 所示，与代表性的对话记忆基准（包括 Mem-Gallery、HaluMem、PersonaMem、LoCoMo 和 LongMemEval）相比，MobileMem-Omni 表现出若干显著特征。与大多数面向通用场景的现有基准不同，MobileMem-Omni 专门针对移动截图设计，并支持多模态输入。在对比的基准中，MobileMem-Omni 采用中英文双语和基于角色的语言设置，而其他列举基准主要以英语为主。此外，MobileMem-Omni 使用合成的个性化图像作为图像来源，而其他多模态基准则使用开源图像。在角色构建方面，MobileMem-Omni 同时包含真实人物和大模型生成的角色，而所有其他基准仅依赖大模型生成的角色。这些独特特征使 MobileMem-Omni 成为移动环境中更具真实性和隐私意识的对话记忆评估框架。

表 5-2 MobileMem-Omni 与其他记忆数据集的对比

	MobileMem-Omni	Mem-Gallery	HaluMem	PersonaMem	LOCOMO	LongMemEval
对话时间跨度	一年	数月	数年	数年	数月	数年
平均单会话长度	8.5k 词元	—	8.3k 词元	6k 词元	477 词元	3k 词元
最大上下文长度	210 万 词元	—	100 万 词元	100 万 词元	9k 词元	150 万 词元
多模态支持	是	是	否	否	是	否
目标平台	移动端	通用	通用	通用	通用	通用
平均单会话图数	5.9	4.2	—	—	3.4	—
图像来源	合成私有图像	开源图像	—	—	开源图像	—
语言	英文+中文	英文	英文	英文	英文	英文
角色来源	真实用户+大模型生成	大模型生成	大模型生成	大模型生成	大模型生成	大模型生成
问题数量	7,415	1,711	3,467	约 6,000	7,512	500

5.2 产业生态应用

MobileMem 构建于 OPPO 原生移动端记忆生态之上，覆盖日历、相册、便签、文档、待办、语音记、小布记忆以及视频记等多种长期记忆来源。这些异构记忆来源共同记录了用户长期使用移动设备过程中产生的时间信息、视觉内容、文本记录、浏览行为、多媒体消费以及个人行为轨迹，为下一代具备长期记忆能力的移动 AI 提供了真实且连续的用户记忆基础。

(1) 健康管理 与 用药提醒

一位年长的用户日常会记录血压和血糖，也保存过过往的体检报告和医生叮嘱。过去几个月里，该用户每次复诊后都会在日历中记下下次预约时间，在笔记中随手记下医生调

整用药时的嘱咐，在浏览器搜索过某种常用药物的禁忌说明，并通过相册保存了化验单和处方照片。某天清晨，用户感觉有些头晕，打开手机问助手：“我之前对什么药物过敏来着？帮我看看最近吃的药里有没有。”助手只需匹配健康记录中的过敏史和当前用药清单，立刻给出明确的风险提醒。

该场景体现了长期个人记忆在健康管理中的应用价值，同时反映了 MobileMem 对跨来源记忆关联、多模态信息融合、长期事实记忆、时间推理以及复杂记忆检索能力的综合评测能力。

（2）异国旅途的行程规划

一位外国游客计划来成都旅行，提前翻了不少游记和攻略，零零散散收藏了好几篇，越看越对川西环线上的一路山水心动。出发前天晚上，用户对助手说：“Based on everything I've bookmarked, what's a good two-day route around Chengdu?” 助手不是泛泛地推荐宽窄巷子和锦里，而是基于那几篇收藏过的攻略内容，结合用户搜索过的交通方式和住宿偏好，整合出一份两日的具体行程。第一天从成都市区出发，经都江堰到映秀，第二天翻过巴朗山垭口看四姑娘山，傍晚返回。

该场景说明长期记忆不仅能够支持信息检索，更能够支撑跨来源知识整合、用户偏好建模、多跳推理以及长期个性化规划。MobileMem 为此类能力提供了真实、连续且可复现的评测环境。

（3）私人影单与追剧记录

一位追剧党习惯每次都在笔记中详细撰写剧评，并配上几张印象深刻的截图，长期记录了上百部影视作品的观感档案。今天刚看完一部新剧，觉得这部和曾经记录过的某部有所共鸣，但具体细节记不清了。用户对助手说：“根据我之前的追剧记录，把和这部有关的内容都整理出来，写篇对比长评。”助手不需要联网搜索，而是翻阅笔记中积攒多年的观感碎片，把风格或主题相近的作品筛选出来，换几个角度逐项比对，生成一篇有观点、有素材、有个人记忆痕迹的剧评。

该场景体现了长期记忆对于兴趣建模、跨时间知识关联以及个性化生成的重要意义，也反映了 MobileMem 在长期知识组织、多模态检索和生成式推理方面的评测价值。

（4）工作复盘与年中总结

一位工作党准备写年中总结，过去半年里开会无数、方案改了又改，技术文章和行业资讯也读得不少，但总觉得学得杂、理不清。用户对助手说：“帮我复盘这半年的工作。

结合上次技术分享会，捋一下我主要在补什么，工作上又卡在哪里。”助手整合日历中这半年所有会议和项目节点的时间线、文档中历次汇报和方案稿的修改记录、笔记中碎片化的思考，以及屏幕记忆里浏览过的技术文章主题，提炼出一条清晰的学习主线，同时对比工作任务和实际产出，定位出用户反复卡壳的环节，生成一份结构清晰的年中复盘摘要。

该场景展示了长期记忆在知识管理、持续学习以及个人成长分析中的应用潜力，也体现了 MobileMem 对长期时间推理、跨来源信息融合、复杂记忆组织以及个性化总结生成等核心能力的系统评测价值。

未来展望。当前版本的端侧记忆基准聚焦于长期性，天然支持迈向更广泛的个性化人工智能生态扩展。未来版本将纳入来自可穿戴设备、平板电脑、个人电脑、智能家居设备与车载系统等记忆来源，实现全生态终身记忆的统一评估。同时计划扩展多语言覆盖范围，增加用户群体的多样性，并引入涉及长期规划、协作记忆、持续个性化与跨模态知识整合等更具挑战性的推理任务。我们期望移动记忆基准发展成为一个连接学术研究与产业部署的综合基准，同时促进可信赖、个性化且持续演化的人工智能助手的发展。

第 6 章 端侧记忆路线图

长期记忆是人工智能迈向持续认知、自主进化的重要基础能力，也是下一代智能终端的重要基础设施。未来十年，端侧记忆的发展将经历由端云协同向端侧自主智能再向可信协同智能的三阶段演进，推动人工智能从单设备智能走向跨终端、跨空间、跨时间的持续认知体系。

6.1 端云协同：构建端侧记忆基础设施

未来一段时期，端云协同仍将是端侧记忆发展的主要形态。端侧负责个人长期记忆的持续积累、组织与调用，云端承担基础模型训练、世界知识更新和复杂推理计算，两者形成统一的长期认知架构。

未来端云协同将由传统的数据同步演进为认知协同。端侧持续沉淀个体长期经验，形成面向用户的认知资产；云端持续演化通用知识模型，为端侧提供知识增强、能力升级和模型优化。通过模型增量更新、知识增量同步和记忆摘要交换，尽可能实现能力共享，减少原始个人数据共享，在保障个人数据本地可控的前提下不断提升智能体能力。

未来端云协同将逐步形成“公共知识云端演化、个体认知端侧沉淀”的新型人工智能基础架构，为大规模端侧智能提供统一底座。

6.2 端侧自主：迈向低功耗持续演化智能

随着端侧算力、存储介质和专用 AI 芯片不断发展，长期记忆有望逐步降低对云端持续计算资源的依赖，进入自主持续演化阶段。

未来端侧记忆系统不仅能够持续记录用户行为，更能够自主完成记忆形成、知识抽象、经验压缩、长期巩固、动态遗忘和认知更新，在设备生命周期内不断形成更加完整、更加准确的个体认知模型。

这一阶段需要突破资源受限条件下长期运行的关键瓶颈，建立面向端侧环境的轻量化持续学习体系，实现超低功耗运行、长期在线服务、自适应资源调度以及跨版本持续演化，使记忆系统能够伴随设备长期运行而持续成长，而非依赖周期性的集中训练与模型重构。

未来端侧记忆将由“持续记录”发展为“持续成长”，成为智能终端长期自主进化的重要支撑。

6.3 可信协同：构建分布式端侧记忆网络

随着个人数字空间不断扩展，手机、平板、PC、智能眼镜、可穿戴设备、智能汽车、家庭机器人等多类终端将共同承载用户长期数字生活。未来长期记忆将突破单设备边界，逐步形成面向个人全空间的分布式端侧记忆网络。

不同终端基于自身感知能力形成局部记忆，通过可信协同机制实现跨设备长期认知共享。协同过程中应尽量避免交换原始隐私数据，而是通过记忆摘要、认知表示、知识参数等隐私保护方式完成跨设备知识共享，使各终端能够持续获得新的认知能力，同时保持个人数据始终留存在本地。

进一步地，多个个体终端之间也将探索建立可信记忆协同机制，在用户授权和隐私保护前提下，实现家庭成员、科研团队、企业组织等多主体之间的知识共享与经验协同，探索从个体智能向协同智能扩展的可能。

未来形成以个人为中心、设备自主协同、隐私可信共享的新一代分布式认知网络，为智能社会提供安全可信的长期认知基础设施。

6.4 展望：迈向长期认知基础设施

端侧记忆的发展不仅是终端人工智能能力的提升，更代表人工智能基础架构的重要演进。未来，长期记忆将由应用功能逐步演化为操作系统级基础设施，由单设备能力扩展为跨终端协同能力，由静态数据管理升级为持续认知演化机制，最终形成覆盖个人、设备与群体的长期认知网络。

围绕端侧记忆形成的新理论、新架构、新标准和新生态，将成为未来人工智能的重要发展方向。率先突破端侧长期记忆关键技术、构建自主可控记忆基础设施并建立开放产业生态，将有望占据下一代人工智能发展的战略制高点，推动人工智能由生成智能迈向持续认知智能。

致谢

我们期望本白皮书能够为行业开发者、企业决策者、政策制定者以及所有关注移动智能体发展的同仁们提供一份有价值的参考。在此衷心感谢参加本白皮书编写的各单位、组织及个人。

合作单位：OPPO 广东移动通信有限公司，中文开放知识图谱（OpenKG）；

核心作者：薛逸达，邓歆乐，茹湘原；

数据应用：陈毅俊，徐步强，毛铭骏，刘忻洁，张弛，钟若彬，方继展，徐浩铭，姜晨，李海博，陈琪琦，张南迪，陈明；

法务合规：贾凡，路达，叶子龙，温晓桦，雷云霞，李镇通，翟钰皓；

技术指导：姜昱辰，王立中，王剑锋，曾理，张晓帆，屠敏欣，张宁豫，张文，陈永锐，孙泽群，邓淑敏，吴天星，王萌，陈想，毕祯，胡伟，王昊奋，漆桂林，陈华钧；

虽然我们编委的每一位专家都努力地尽力把工作做到尽善尽美，但白皮书仍不可避免有各种纰漏，恳请行业内专家和同仁不吝赐教。