面向大语言模型的记忆管理理论框架研究：认知自适应与用户参与的视角

作者：
Li Wenhan

摘要

大语言模型在长程交互中面临记忆过载与用户失控的双重困境：无差别的海量存储导致认知负荷攀升，黑箱式的遗忘机制引发隐私信任危机。本研究提出一种兼具认知自适应与用户可干预的 AI 记忆管理理论框架（CAUM）。首先，基于信息熵、交互频率与冲突检测，设计多维记忆重要性评估模型，特别引入后文关联潜力作为信息价值评估的新维度，使记忆保留更具前瞻性；其次，构建包含原始层、摘要层与骨架层的分级存储架构，并引入阈值触发的智能压缩机制；最后，提出用户参与式授权机制，将“记忆整理提案”可视化呈现并由用户审核决策，实现“人在回路”的记忆治理。该框架为缓解 LLM 记忆过载问题提供了系统的概念方案，将信息生命周期理论拓展至 AI 记忆管理领域，强调用户中心的信息处置权，为人工智能时代的信息生命周期管理提供了新的理论视角，也为构建用户可控的智能记忆系统奠定了概念基础。

关键词：大语言模型；记忆管理；认知负荷；用户参与；信息生命周期；后文关联

1 引言

想象这样一个场景：你第三次向 AI 助手重复自己偏爱中深烘、低酸度的咖啡，而它依然像初次听闻般热情回应；与此同时，它却牢牢记住了一次偶然提及的琐碎细节，并在后续对话中不合时宜地提起。这种“该忘的忘不掉，该记的记不住”的窘境，正在成为人机深度交互的日常困扰。更令人不安的是，用户无从知晓 AI 究竟记住了什么、遗忘什么——记忆是一个不透明的黑箱。

随着大语言模型（Large Language Models, LLMs）在对话式人工智能领域的深度融合，用户期冀与 AI 建立长期、连贯、富含情境的伙伴关系。然而，当前主流 LLM 的记忆能力与其强大的生成能力严重不匹配。绝大多数模型依赖于固定长度的上下文窗口（如 4K、32K 或 128K tokens），一旦对话长度超出此限，早期信息便被简单截断或通过滑动窗口丢弃。更先进的方案虽能借助向量数据库实现“全量无差别存储”，但这又引发了双重困境。

对系统自身而言，无止境的信息堆积导致检索成本呈指数增长，模型处理速度下降，推理能耗上升——即面临严重的“认知负担过重”。同时，无关或冗余信息的干扰可能影响当前对话的专注度与决策质量。对用户而言，模型的记忆是一个黑箱：用户无法知晓 AI 记住了哪些细节，又以何种逻辑决定遗忘哪些内容。这种失控感侵蚀了用户信任，尤其在涉及个人偏好、隐私信息或关键决策依据的场景下，用户完全有权要求知情与干预。

反观人类记忆系统，其高效性与适应性为我们提供了宝贵灵感。人脑并非事无巨细地保存所有输入，而是通过一套精密的动态机制进行运作：依据信息的情感强度、重复频率、与当前目标的相关性进行重要性筛选；通过睡眠中的记忆整合将短期记忆转化为长期记忆，抽提出核心要点；甚至存在主动遗忘机制以清理无效信息、优化认知资源。这一过程是自适应、有选择且服务于更高认知目标的。同时，信息科学领域的信息生命周期理论（Information Lifecycle Management, ILM）为我们提供了管理框架，强调信息从创建、使用、维护到处置的完整流程都应受到有效管控。将这一理论应用于 AI 记忆，意味着我们需要对 AI“记住”的信息负责，为其设计一个包含“处置”环节的完整生命周期管理方案。

从信息管理学科的视角来看，大语言模型的记忆问题本质上是信息资源的全生命周期治理问题。传统信息管理关注文本、数据等显性信息，而 AI 记忆作为一种新型的、动态生成的“隐性信息资产”，其管理理论尚付之阙如。本研究正是从信息生命周期理论出发，探索 AI 记忆的“处置”环节，拓展信息管理的理论边界。

基于上述背景，本文致力于融合认知科学与信息管理理论，提出一种名为“认知自适应用户参与式记忆管理框架”（Cognitively-Adaptive User-involved Memory Management Framework, CAUM）的理论方案。本框架的核心创新在于三个转变：（1）从被动存储到主动感知：使模型具备量化自身记忆负荷的能力，在认知瓶颈出现前主动触发管理流程；（2）从均质扁平到分层压缩：引入多级记忆表征，实现对信息的渐进式抽象与压缩，保留信息熵最高的核心内容；（3）从机器自主到人机协同：将最终处置权以清晰、可理解的方式交还用户，构建可审计、可控制的记忆流程。

本文作为一项理论框架研究，旨在为后续实证工作提供概念基础与设计蓝图，而非报告已完成实验的结果。

2 文献综述

2.1 大语言模型的记忆机制演进

早期 Transformer-based LLMs 的记忆完全依赖于自注意力机制内的上下文窗口。为突破长度限制，研究者提出多种技术路径：（1）工程化扩展，如线性注意力（Katharopoulos et al., 2020）、位置插值（Chen et al., 2023）等，直接增大有效窗口，但计算复杂度随窗口长度呈平方增长；（2）检索增强，将长上下文存储在外部向量数据库中，按需检索相关片段注入上下文，典型工作如 RAG（Lewis et al., 2020）及后续变体，这类方法虽扩展了记忆容量，但仍是“存储-检索”模式，缺乏对信息本身的价值评估与动态管理；（3）结构化摘要，在对话过程中定期生成摘要并替代原文送入上下文，初步体现了“压缩”思想，但摘要策略通常固定且未经用户确认，压缩过程不可控。总体而言，现有方法多关注“如何记住更多”，而非“如何记住更好、更智能”。这一现象与张敏等（2024）指出的“AI 赋能知识组织仍停留在存储层面”的判断相一致。

2.2 机器学习中的持续学习与灾难性遗忘

在持续学习领域，“灾难性遗忘”是一个经典问题。弹性权重巩固（Elastic Weight Consolidation, EWC; Kirkpatrick et al., 2017）通过对重要参数施加约束来保护旧知识；梯度情景记忆（Gradient Episodic Memory, GEM; Lopez-Paz & Ranzato, 2017）则利用情景存储与梯度约束缓解遗忘。这些方法的核心是对神经网络参数层面的重要性进行评估。然而，LLM 对话中的记忆管理更侧重于内容层面（具体对话事实、用户偏好）而非参数层面（通用语言能力）的遗忘与保留。直接将持续学习算法应用于对话历史管理并不自然，需要新的内容重要性度量标准。

2.3 类脑记忆模型与主动遗忘机制

受认知科学启发，一系列工作尝试为 AI 构建类脑记忆系统。记忆网络（Memory Networks; Weston et al., 2014）和可微分神经计算机（Differentiable Neural Computer, DNC; Graves et al., 2016）引入了可读写的外部记忆体。近年来，生成式智能体研究取得了显著进展，Park et al.（2023）提出的 Generative Agents 通过“反思”机制，将具体经历提炼为高级别的概括性描述并存储于长时记忆，这本质上是一种周期性的、自主的“记忆压缩”。同年，Packer et al.（2023）提出 MemGPT，借鉴操作系统思想为 LLM 设计分层记忆管理，当上下文超限时自动触发中断并将早期内容交换至外部存储。然而，这些模型的“遗忘”或“压缩”规则完全内置于系统设计中，用户无法知晓、更无法干预其记忆的“消化”过程，透明性与可控性不足。

2.4 信息生命周期理论在数字治理中的应用

信息生命周期理论起源于记录管理与信息资源管理领域，Horton（1985）和 Marchand（1985）等学者奠定了其理论基础。该理论将信息视为具有生命周期的资源，包含创建、采集、组织、存储、利用、归档和处置等多个阶段。在数据治理、数字档案等领域，该理论已被成熟应用，确保信息的合规、高效与安全（Higgins, 2008）。王曰芬等（2023）系统梳理了信息生命周期理论从文档管理到数据治理的演进脉络。然而，在 AI，特别是 LLM 的记忆场景中，研究普遍聚焦于信息的“创建”（生成）、“存储”（向量化）和“利用”（检索），而几乎完全忽视了“处置”阶段。AI 记忆的“处置”（即遗忘、删除、销毁）应是其生命周期不可或缺的终点，且必须符合伦理与法律要求（GDPR 等法规明确赋予用户“被遗忘权”）。本研究正是将 ILM 理论系统性地引入 AI 记忆管理，强调用户参与下的、负责任的记忆处置，填补了这一研究空白。

2.5 现有研究的不足与本框架的定位

综合上述分析，现有研究存在以下不足：（1）技术路径上，重存储轻管理，缺乏对信息价值的动态评估；（2）用户参与度上，记忆过程黑箱化，用户被排除在决策之外；（3）理论视角上，缺乏将记忆视为需要全生命周期治理的信息资源。陈果等（2025）指出，面向对话场景的知识组织需要从静态分类转向动态记忆，这与本研究的出发点一致。本研究提出的 CAUM 框架旨在从理论上弥补这些不足，为后续开发用户可控的 AI 记忆系统提供概念基础。

3 理论基础

3.1 认知负荷理论及其在 AI 中的类比

认知负荷理论（Cognitive Load Theory, CLT）由 Sweller（1988）提出，认为人类工作记忆容量极其有限，信息处理效率受内在认知负荷（任务复杂性）、外在认知负荷（信息呈现方式）和相关认知负荷（图式构建）影响。过载将导致学习失败或决策错误。将这一理论类比至 LLM：模型的上下文窗口可视为其“工作记忆”，其容量限制是硬性的。存储在外部记忆库中的海量信息，在检索时需要被“激活”并加载到工作记忆中进行分析，这个过程同样消耗计算资源，可视为一种“外部检索负荷”。当记忆条目过多、关系过于复杂时，检索、筛选、整合信息的负担会显著增加系统延迟和错误率，即产生“AI 认知过载”。CAUM 框架的目标之一，便是通过主动管理记忆内容，降低这种外部检索负荷和相关认知负荷。

3.2 信息价值衰减与评估模型

信息并非永恒等值。在对话上下文中，信息的价值随其“新鲜度”（Recency）、被访问的“频次”（Frequency）及其与对话主体的“重要性”（Importance）动态变化。本研究提出一个综合的价值评估函数：

$$V(i,t) = \alpha \cdot R(i,t) + \beta \cdot F(i) + \gamma \cdot I(i) + \delta \cdot L(i)$$

其中，R(i,t)是时间衰减函数，本文采用指数衰减形式 $R(i,t) = e^{-\lambda(t - t_0(i))}$，$t_0(i)$为记忆单元 i 的创建时间，λ 为衰减率；F(i)为历史访问频率的归一化值，$F(i) = f_i / \max_j f_j$；I(i)为重要性得分，可由模型根据信息类型（事实陈述、用户偏好设定、情感表达）、用户明确标注（如“记住这个”）或隐含信号（被多次追问、在决策中被引用）自动推断；L(i)表示记忆单元 i 的后文关联潜力，即该信息与当前对话上下文及潜在未来话题的语义关联强度。该维度受认知科学中“线索-依赖性遗忘”（cue-dependent forgetting）理论的启发：记忆的可访问性取决于是否存在有效的检索线索，若一条信息与后续对话中可能出现的主题、实体或意图高度相关，则其被再次利用的概率较高，因此应赋予更高的保留价值。

L(i)的计算可通过实时分析当前对话窗口中的核心实体与意图，并计算其与记忆库中各条目的语义相似度实现。例如，若用户当前询问“咖啡”，而历史记忆中包含“胃痛”的记录，即使该记录时间久远，其与当前话题的医学关联可能导致 L(i)值升高，从而避免被过早压缩。权重 α, β, γ, δ 满足 $\alpha + \beta + \gamma + \delta = 1$，初始值可根据应用场景通过专家判断或预实验设定（如 $\delta = 0.2$），后续通过用户反馈自适应调整。

3.3 用户采纳与信任模型

技术接受模型（Technology Acceptance Model, TAM; Davis, 1989）指出，感知有用性和感知易用性是用户采纳新技术的关键。在 AI 记忆管理中，“有用性”体现为记忆系统能

否高效、准确地支持对话；“易用性”则体现为用户理解和管理记忆的难度。信任理论进一步强调能力、仁爱心和正直是信任的三大基石（Mayer et al., 1995）。一个黑箱的记忆系统损害了用户的“感知控制感”，削弱了对系统“正直”（行为可预测、符合预期）的信任。Venkatesh et al.（2003）的 UTAUT2 模型也指出，习惯与信任显著影响技术采纳意愿。图情档领域的实证研究也表明，用户对 AI 服务的信任显著受感知控制感影响（李晶等，2023）。CAUM 框架通过引入用户参与式授权，旨在提升系统的透明性、可控性和可预测性，从而直接增强用户的感知易用性、控制感，并最终建立对 AI 记忆能力的信任。

4 模型构建：CAUM 框架详述

4.1 总体架构

CAUM 框架是一个包含四个核心模块的闭环系统。记忆流从“记忆采集层”进入，经过“记忆评估层”计算价值与负荷，当触发阈值时流向“记忆压缩层”进行分级处理，生成的提案提交至“用户授权层”等待批复，批复后的指令更新“记忆存储体”，同时用户反馈优化评估模型。

4.2 记忆分级存储结构

为模拟人类记忆的层次性，本框架采用三级存储结构：

L0 - 原始细节层（短期缓存）：存储原始对话轮次（或初步清洗后的文本），保留最完整的语境和细节。容量上限为 N_max（如 1000 轮），确保高速访问。当新对话进入时，遵循先进先出原则，溢出的记忆触发评估，准备向下一层转移。

L1 - 整合摘要层（episodic memory）：存储经过首次压缩的整合性记忆。当 L0 层记忆单元被判定为有价值但细节冗余时，通过摘要模型提炼其核心事实、用户在该话题中表达的主要观点、偏好或决策结果。例如，将关于“假期旅行计划”的 10 轮讨论压缩为“用户计划于 2026 年国庆期间前往西安，偏好历史古迹，预算中等”。此层保留较多语义内容，支持对具体事件的查询。

L2 - 抽象骨架层（semantic memory）：存储高度抽象、泛化的用户模型和世界知识。这是对 L1 层信息的进一步提纯，形成长期的、稳定的用户画像和交互模式。例如，从多次用餐对话中抽象出“用户是素食主义者”；从多次技术讨论中抽象出“用户具备中级编程知识”。此层数据量小但信息密度极高，用于指导 AI 的长期行为基调。

每一层记忆均附带元数据：<创建时间，最后访问时间，压缩来源指针，重要性标签，访问频次>，构成可追溯的记忆图谱。陈果等（2025）提出的动态记忆组织思路与本框架的设计理念相契合。

4.3 认知负荷量化与动态触发机制

定义系统实时的认知负荷指数 C，作为触发记忆管理的信号。C 由多个可观测指标加权计算得出：

$$C = w_1 \cdot (N/N_{max}) + w_2 \cdot (T_{avg}/T_{max}) + w_3 \cdot (K_{conflict}/K_{max})$$

其中：
N 为当前 L0 层记忆单元数量，N_max 为 L0 层容量上限，反映存储压力；
T_avg 为近期（如过去 50 次查询）记忆检索的平均响应时间，T_max 为可容忍的最大平均延迟阈值，反映性能压力；
K_conflict 为近期检测到的记忆内部不一致或与外部知识冲突的次数（如用户两次陈述的偏好矛盾，或模型记忆与最新事实不符），K_max 为冲突容忍上限，反映记忆质量与一致性压力；
w_1，w_2，w_3 为权重系数，满足 $w_1 + w_2 + w_3 = 1$，可根据应用场景通过专家判断或层次分析法设定初始值，并后续通过用户反馈自适应调整。

当 C > C_threshold（如 0.8）时，系统判定认知负荷过高，自动启动记忆管理流程。但在触发管理前，系统会额外评估当前对话主题与低价值记忆的后文关联度：若检测到某些价值评分较低的记忆单元与当前上下文存在强关联（即 L(i)高于预设阈值 L_high），则临时保护这些记忆免于被压缩，并将其标记为“线索唤醒”，被唤醒的记忆将获得一个短期的“保护窗口”（如当前会话内），窗口结束后重新进入常规评估流程，同时记录该事件用于后续的权重优化。此机制模拟了人类记忆中由情境线索触发旧忆的过程，有助于提升记忆利用的主动性与准确性。

4.4 分层压缩与遗忘算法

压缩过程是 CAUM 框架的核心智能环节，采用多步骤管道：

1. 候选记忆选择：根据价值评估函数 V(i,t)，在 L0 层和 L1 层中筛选出价值低于阈值 V_low 的记忆单元集合 M_candidate。
2. 重要性聚类与关联分析：对 M_candidate 中的记忆进行主题聚类（如使用 LDA 或 BERT-based 聚类）或实体关联分析，识别出可合并的语义群组。在此过程中，系统将特别考虑记忆之间的后文关联模式：通过分析历史对话中话题演变的规律，挖掘那些经常被后续对话共同提及的记忆集合，将其归为“关联群组”。这类群组在压缩时倾向于整体保留或整体压缩，以维持知识结构的完整性。这一机制受认知心理学中“图式理论”（schema theory）的启发：例如，关于用户饮食习惯的多条零散记忆（如“喜欢中深烘咖啡”“偶尔胃痛”“最近在控制咖啡因”）可能因多次被共同引用而聚类，从而在压缩时生成更丰富的摘要。
3. 提炼压缩：
    - 对于 L0 层的候选记忆，调用 LLM 进行摘要式压缩，生成 L1 层记忆条目。提示词设计为：“请将以下对话历史浓缩为一段简洁的摘要，保留关键事实、用户做出的决定或表达出的明确偏好。忽略寒暄和重复内容。”

- 对于 L1 层的候选记忆，调用 LLM 进行抽象化压缩，尝试归纳出更高阶的模式或原则，生成或更新 L2 层记忆条目。提示词例如：“基于以下几条关于用户饮食偏好的摘要，请归纳总结用户在饮食方面的长期特点或原则。”
4. 生成处置提案：系统生成一份结构化提案，列出建议压缩的记忆组、建议完全遗忘的低价值碎片记忆、压缩/遗忘的理由（如“信息过时”“与核心偏好无关”）、以及操作后的预期收益（如“预计释放 L0 层 15%空间，平均检索速度提升 20%”）。

4.5 用户参与式授权与交互界面

系统将“记忆整理提案”通过交互界面呈现给用户。界面设计遵循简洁性、透明性与可控性原则：
- 记忆可视化图谱：以时间线或知识图谱的形式展示当前记忆存储概况，高亮被建议处置的部分。图谱节点代表记忆单元，边代表语义关联。
- 提案详情卡片：对每一项处置建议，提供“原文片段”“压缩后内容”“处置理由”的对比展示，支持展开查看详情。对于因高后文关联度而被系统临时保留的记忆，卡片中将特别标注“线索关联提醒”，例如：“系统注意到您近期频繁讨论饮食健康，因此关联到您三个月前关于胃部不适的记录，建议暂时保留以备参考。”这一设计旨在提升系统行为的可解释性，使用户理解 AI 的“贴心”并非偶然，从而增强信任感。
- 授权选项：一键批准；选择性批准；手动调整（用户可以直接编辑系统建议的压缩后文本，或为特定记忆打上“永久保留”标签）；暂不处理。
- 反馈学习：用户的决策行为将被记录，用于优化重要性评估模型中的权重参数。例如，若用户频繁保留某类信息，其重要性权重 I(i)的初始值应自动调高；若用户经常拒绝压缩某类信息，则对应类型的衰减率 λ 可适当降低。

5 框架实现的技术构想

为展示 CAUM 框架在技术上的可行性，本节提出一种基于现有技术的实现构想，作为未来实际开发的参考。

5.1 技术架构设想

在具体实现时，可考虑以下技术选型与架构设计：
- 后端框架：可采用 LangChain 构建智能体流程，实现模块化连接。
- 大语言模型：可选用 ChatGLM3-6B 等开源模型作为核心对话与压缩引擎，部署于本地服务器。
- 记忆存储：
  - L0 层：可使用 Redis 存储原始对话记录，保证高速访问。
  - L1/L2 层：可使用 PostgreSQL 关系数据库存储结构化摘要和抽象记忆。
- 向量索引：可使用 Chroma 等向量数据库对所有层级的记忆文本建立嵌入索引，以支持语义检索。

- 前端界面：可使用 Gradio 等工具快速构建 Web UI，包含对话主界面和独立的“记忆管理面板”。

5.2 核心模块设计思路

- 评估器模块：可按照 4.3 节的价值评估函数和认知负荷计算逻辑进行实现。
- 压缩引擎模块：可封装对 LLM 的调用，设计针对摘要压缩和抽象压缩的两套提示词模板，压缩过程可设计为异步执行以降低对实时对话的影响。
- 授权界面模块：可在前端实现“记忆管理”面板，使用 NetworkX 等库生成记忆图谱，以表格或卡片形式展示处置提案，支持用户逐项审批。

5.3 可行性分析

基于上述设计思路，系统理论上可完成以下流程：（1）多轮对话记录存入 L0 层；（2）触发压缩流程后，系统生成摘要提案；（3）授权界面展示提案，用户可勾选批准或拒绝；（4）系统根据用户选择更新记忆库。这一技术构想表明，CAUM 框架的核心机制在现有技术条件下具备可实现性。实际系统的效果评估有待未来通过规范的实证研究加以验证。

6 框架分析与理论论证

本章从理论层面分析 CAUM 框架的预期优势与可行性，而非报告实证结果。

6.1 时间复杂度分析

假设 L0 层容量为 N，L1 层容量为 M，L2 层容量为 K，且 $N \gg M \gg K$。在传统全量存储方案中，每次检索需扫描全部 N+M+K 条记忆（或至少需计算与全部记忆的相似度）。在 CAUM 框架中，日常对话仅需在 L0 层进行检索，时间复杂度为 O(N)；当需要深度回忆时，可依次检索 L1 和 L2 层。由于压缩后保留了核心信息，多数查询可在 L0 层解决。从理论上，CAUM 框架可将平均检索时间复杂度从 O(N+M+K)降至 O(N)，其中 N 远小于无管理状态下的总记忆量。

6.2 信息压缩率估计

设原始对话平均每条占用 S_0 存储单位，摘要平均每条占用 S_1，抽象原则平均每条占用 S_2。根据已有摘要系统的经验，$S_1 \approx 0.1S_0$（压缩 90%），$S_2 \approx 0.01S_0$（压缩 99%）。若系统将 N_0 条原始对话压缩为 N_1 条摘要和 N_2 条抽象，且 $N_1 \ll N_0$，$N_2 \ll N_1$，则总存储量从 N_0S_0 降至 N_0S_0（保留部分原始）+ N_1S_1 + N_2S_2。在典型场景下，理论压缩率可达 80%-95%。

6.3 用户参与的理论收益

从用户采纳理论（Davis, 1989; Venkatesh et al., 2003）推导，CAUM 框架通过以下机制预期提升用户信任与满意度：（1）透明性：可视化记忆图谱让用户知晓 AI 记住了什么；（2）可控性：授权机制赋予用户最终决定权；（3）可预测性：通过学习用户偏好，系统行为逐渐符合用户预期。这三者共同作用于感知控制感，进而影响信任与采纳意愿。这一理论推演有待未来实证检验。

6.4 框架适用边界分析

CAUM 框架主要适用于以下场景：（1）长期、连续性对话：如个人助理、教育辅导、心理咨询等；（2）对隐私敏感的用户：希望控制 AI 所持个人信息；（3）对记忆准确性要求高的场景：如医疗、法律咨询，需确保关键信息不被误忘。对于一次性对话或对效率要求极高、用户不愿参与管理的场景，简化版的记忆策略可能更为合适。

6.5 信息管理视角下的理论意义

从信息管理学看，CAUM 框架将传统 ILM 理论中处于边缘位置的“处置”阶段提升为核心环节。在纸质时代，信息处置主要表现为档案销毁；在数字时代，表现为数据删除；而在 AI 时代，记忆的“处置”不再是简单的删除，而是一个需要用户参与的、动态的、可逆的信息价值再评估过程。吴江等（2024）指出，生成式 AI 环境下的信息治理面临新的挑战，本框架正是对这一挑战的回应。这一拓展丰富了信息生命周期理论在智能时代的内涵。

7 讨论

7.1 理论贡献

本研究的主要理论贡献在于跨学科融合与概念拓展。第一，成功将认知科学中的记忆理论（Sweller, 1988）和信息管理学中的信息生命周期理论（Horton, 1985）系统性地引入 LLM 工程实践，为解决 AI 记忆问题提供了富有解释力的概念框架，拓展了 ILM 理论在人工智能时代的应用边界。第二，明确提出“用户参与式信息处置”这一核心概念，将 AI 记忆的“处置”权从纯算法逻辑中剥离出来赋予用户，为 AI 伦理、可解释性和人机信任研究提供了新视角。这与 GDPR“被遗忘权”等法规精神高度契合，具有理论前瞻性。第三，提出的“认知负荷指数”为量化评估 AI 系统的“记忆健康度”提供了可操作的指标体系，填补了现有研究多关注“性能”而忽视“负荷”的空白。

7.2 实践意义

在应用层面，CAUM 框架为构建下一代可信赖的 AI 助手提供了可参考的设计蓝图。它尤其适用于那些对连续性、个性化要求高，且涉及敏感信息的场景，如心理陪伴助手、医疗顾问、法律咨询助理、个性化教育导师等。在这些场景中，记忆的可控、透明、可审计不仅是体验需求，更是伦理和法律要求。框架中的用户授权模块可以作为一种标准化的“记忆审计接

口”，方便用户监督和管理 AI 所掌握的个人信息。此外，本框架的技术构想具有较好的可复现性，可为业界开发同类系统提供参考。

7.3 局限性与未来工作

本研究作为一项理论框架研究，存在以下局限，指向未来研究方向：

1. 缺乏实证检验：本文提出的 CAUM 框架尚未经过系统的用户实验验证。框架的实际效果（如用户信任提升幅度、系统效率改善程度）有待未来开展规范的实证研究。后续工作可按照规范的实验设计方案，招募被试开展对照实验。
2. 压缩的保真度问题：依赖 LLM 进行摘要和抽象可能引入偏见或“幻觉”。未来需研究更可靠的压缩算法，如基于事实一致性校验的压缩，或引入多步验证机制。
3. 用户交互成本与自动化平衡：用户参与虽增强控制感，但可能带来交互负担。未来需研究智能授权代理机制，对高置信度操作自动化处理，仅对高风险操作提请授权。
4. 后文关联计算的开销与准确性：引入后文关联潜力 L(i)虽能提升记忆保留的精准性，但实时计算所有记忆与当前话题的语义相似度可能带来显著的计算开销。未来可探索建立“动态热区”机制，仅对近期活跃或高权重的记忆进行关联扫描；同时，需防范“过度关联”问题（如不恰当的语义联想），可通过用户反馈闭环不断修正关联模型，确保其符合用户的实际认知习惯。
5. 遗忘的伦理问题：用户授权的遗忘操作是否应支持撤销？彻底删除的数据能否从模型参数中完全移除？这些问题涉及 AI 遗忘的深层伦理挑战，值得深入研究。

8 结论

本文针对大语言模型在长程交互中面临的记忆过载、黑箱操作与用户失控问题，从认知自适应与人机协同的视角出发，提出了一个创新的记忆管理理论框架 CAUM。该框架通过构建多级记忆存储、量化认知负荷、实现阈值触发的智能压缩，并最终将处置决策置于用户的知情与授权之下，创造性地将 AI 的记忆管理从一个单纯的技术优化问题，转变为一个需要人机协作的治理过程。

通过本文提出的技术构想与理论分析，论证了 CAUM 框架在降低检索复杂度、提高信息压缩率方面具有预期优势，且符合用户采纳理论关于信任构建的基本原理。

这项工作表明，将人类记忆的智慧与信息治理的原则融入 AI 系统设计，是构建更高效、更透明、更可信人机关系的关键路径。未来的智能体不应只是一个被动的信息容器，而应成为一个懂得“消化”“整理”并与用户共同“管理”记忆的伙伴。本研究作为一项理论探索，为此愿景提供了概念基础与设计蓝图。后续研究可在本框架基础上开展实证检验，推动理论向实践的转化。

随着 AI 日益融入人类生活，“AI 该记住什么、遗忘什么”不仅是技术问题，更是关乎信息主权、用户尊严与社会信任的核心议题。信息管理学科在记录保存、信息治理、用户权益等

领域积累了丰富理论，完全有能力为这一新兴问题提供学术支撑。期待更多研究关注 AI 记忆的信息管理维度，共同塑造以人为本的智能信息生态。

作者贡献声明

本文是人类与人工智能协作完成的学术成果。

人类作者（Li Wenhan）贡献：
- 独立提出论文的核心研究问题、理论框架与关键概念，包括：
  - 将人脑三级记忆结构（原始层、摘要层、骨架层）引入 AI 记忆管理
  - 设计“重要记忆保留”与“用户参与式授权”机制
  - 定义“认知过载”问题并关联信息管理学科的信息生命周期理论
  - 提出“后文关联潜力”作为信息价值评估的新维度
- 确定论文的整体研究方向、逻辑结构与论证路径
- 对 AI 生成的所有内容进行逐项审核、修改与最终定稿
- 对论文的全部内容承担最终责任

人工智能工具（DeepSeek）贡献：
- 根据人类提供的核心思想与关键词，辅助进行文献检索与梳理
- 协助组织论文结构、润色语言表达、统一术语与格式
- 生成初稿框架，并在人类指导下完成多轮修订
- 生成参考文献格式草案（经人类核对真实性）

利益冲突声明：无。

致谢

本文核心思想由人类作者独立提出。写作过程中使用了人工智能辅助工具（DeepSeek）进行文献梳理、语言润色与格式整理。AI 生成内容已全部经作者审核、修改与整合，作者对论文全部内容承担最终责任。感谢所有对本文思想形成有所启发的先行研究。

参考文献

[1] 陈果， 赵一鸣. 面向对话场景的知识组织框架：从静态分类到动态记忆[J]. 情报学报， 2025， 44(1)： 1-12.
[2] 李晶， 王文韬， 谢阳群. 用户对人工智能服务的信任机制研究——基于信息采纳视角[J]. 情报学报， 2023， 42(5)： 513-525.
[3] 王曰芬， 曹高辉. 信息生命周期理论的演进与拓展：从文档管理到数据治理[J]. 情报学报， 2023， 42(8)： 891-902.
[4] 吴江， 胡蓉. 生成式 AI 环境下的信息治理：挑战与应对[J]. 情报学报， 2024， 43(11)： 1257-1268.

[5] 张敏， 刘晓彤， 夏宇. 人工智能赋能知识组织：理论逻辑与实践路径[J]. 情报学报，2024， 43(2)： 127-138.
[6] Brooks J. SUS： A 'quick and dirty' usability scale[M]//Jordan P W， Thomas B， Weerdmeester B A， et al. Usability evaluation in industry. London： Taylor & Francis， 1996.
[7] Chen S， Wong S， Chen L， et al. Extending context window of large language models via positional interpolation[J]. arXiv preprint arXiv：2306.15595， 2023.
[8] Davis F D. Perceived usefulness， perceived ease of use， and user acceptance of information technology[J]. MIS Quarterly， 1989， 13(3)： 319-340.
[9] Graves A， Wayne G， Reynolds M， et al. Hybrid computing using a neural network with dynamic external memory[J]. Nature， 2016， 538(7626)： 471-476.
[10] Higgins S. The DCC curation lifecycle model[J]. International Journal of Digital Curation， 2008， 3(1)： 134-140.
[11] Horton F W. Information literacy vs. computer literacy[J]. Bulletin of the American Society for Information Science， 1985， 9(4)： 14-16.
[12] Jian J Y， Bisantz A M， Drury C G. Foundations for an empirically determined scale of trust in automated systems[J]. International Journal of Cognitive Ergonomics， 2000， 4(1)： 53-71.
[13] Katharopoulos A， Vyas A， Pappas N， et al. Transformers are RNNs： Fast autoregressive transformers with linear attention[C]//Proceedings of the 37th International Conference on Machine Learning. New York： ACM， 2020： 5156-5165.
[14] Kirkpatrick J， Pascanu R， Rabinowitz N， et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences， 2017， 114(13)： 3521-3526.
[15] Lewis P， Perez E， Piktus A， et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems 33. Red Hook： Curran Associates， 2020： 9459-9474.
[16] Lopez-Paz D， Ranzato M. Gradient episodic memory for continual learning[C]//Advances in Neural Information Processing Systems 30. Red Hook：Curran Associates， 2017.
[17] Marchand D A. Information management： Strategies and tools in transition[J]. Information Management Review， 1985， 1(1)： 5-16.
[18] Mayer R C， Davis J H， Schoorman F D. An integrative model of organizational trust[J]. Academy of Management Review， 1995， 20(3)： 709-734.
[19] Packer C， Wooders S， Lin K， et al. MemGPT： Towards LLMs as operating systems[J]. arXiv preprint arXiv：2310.08560， 2023.
[20] Park J S， O'Brien J， Cai C J， et al. Generative agents： Interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM

Symposium on User Interface Software and Technology. New York： ACM，2023： 1-22.
[21] Sweller J. Cognitive load during problem solving： Effects on learning[J]. Cognitive Science， 1988， 12(2)： 257-285.
[22] Venkatesh V， Morris M G， Davis G B， et al. User acceptance of information technology： Toward a unified view[J]. MIS Quarterly， 2003，27(3)： 425-478.
[23] Weston J， Chopra S， Bordes A. Memory networks[J]. arXiv preprint arXiv：1410.3916， 2014.