

基于图形的智能体记忆： 分类法、技术与应用

Chang Yang[†], Chuang Zhou[†], Yilin Xiao[†], Su Dong, Luyao Zhuang, Yujing Zhang, Zhu Wang, Zijin Hong, Zheng Yuan, Zhishang Xiang, Shengyuan Chen[‡], Huachi Zhou[‡], Qinggang Zhang[‡], Ninghao Liu, Jinsong Su, Xinrun Wang, Yi Chang, Xiao Huang

摘要—记忆在基于大语言模型 (LLM) 的智能体执行长时程复杂任务 (例如多轮对话、游戏博弈、科学发现) 中成为核心模块, 其中记忆能够实现知识积累、迭代推理和自我演化。在多种范式中, 图结构因其固有的建模关系依赖、组织层次化信息以及支持高效检索的能力而脱颖而出, 成为智能体记忆的有力结构。本综述从基于图形的视角对智能体记忆进行了全面回顾。首先, 我们提出了智能体记忆的分类体系, 包括短期记忆与长期记忆、知识记忆与经验记忆、非结构化记忆与结构化记忆, 并从图结构实现的角度进行阐述。其次, 根据智能体记忆的生命周期, 系统分析了基于图的智能体记忆中的关键技术, 涵盖记忆提取 (将数据转换为内容)、存储 (高效组织数据)、检索 (从记忆中获取相关内容以支持推理) 以及演化 (更新记忆中的内容)。第三, 我们总结了支持自演化智能体记忆开发与评估的开源库和基准测试。同时, 我们也探讨了多样化的应用场景。最后, 我们指出了当前面临的关键挑战与未来研究方向。本综述旨在为推动更高效、更可靠的基于图的智能体记忆系统的发展提供可操作的洞察。所有相关资源, 包括研究论文、开源数据及项目, 均已整理并汇总于 <https://github.com/DEEP-PolyU/Awesome-GraphMemory> 供社区使用。

[†]Equal contribution.

[‡]Corresponding authors: Qinggang Zhang, Huachi Zhou, Shengyuan Chen.

Chang Yang, Chuang Zhou, Yilin Xiao, Su Dong, Luyao Zhuang, Yujing Zhang, Zhu Wang, Zijin Hong, Zheng Yuan, Shengyuan Chen, Huachi Zhou, Qinggang Zhang, Ninghao Liu, and Xiao Huang are with the The Hong Kong Polytechnic University, Hong Kong SAR, China (e-mail: {chang.yang, chuang-qzj.zhou, yilin.xiao, su.dong, luyao.zhuang, yu-jing.zhang, julia.zhu.wang, zijin.hong, yzheng.yuan, huachi.zhou}@connect.polyu.hk, {sheng-yuan.chen, qinggang.zhang, ninghao-prof.liu, xiao.huang}@polyu.edu.hk).

Jinsong Su and Zhishang Xiang are with the School of Information, Xiamen University, China (e-mail: xiangzhishang@stu.xmu.edu.cn, jssu@xmu.edu.cn).

Xinrun Wang is with the School of Computing and Information Systems, Singapore Management University, Singapore (e-mail: xr-wang@smu.edu.sg).

Yi Chang is with the School of Artificial Intelligence, Jilin University, Changchun, China (e-mail: yichang@jlu.edu.cn).

Index Terms—智能体, 多智能体系统, 智能体记忆, 知识图谱, 自演化, 基于图形的记忆

I. 引言

近几年见证了基于大模型 (LLM) 的智能体的快速发展, 这些智能体在多个领域中复杂且长周期的任务上表现出色, 涵盖了从软件工程 [1]、数学推理 [2] 到多智能体任务 [3] 和开放世界探索 [4] 等。大模型固有的语言理解、生成和推理能力使得基于大模型的智能体能够自主感知环境并做出决策, 从而减少人工干预, 并重塑智能系统的范式 [5]。

尽管取得了显著进展, 基于大模型的智能体仍然受到大模型内在局限性的制约。(i) 知识截止: 大模型在静态数据集上进行预训练, 具有固定的时序边界, 导致知识截止问题, 使其无法融入实时信息 (如当前金融数据) 或超出其预训练语料库的领域特定知识。这一限制削弱了其适应动态环境和开放性场景的能力。(ii) 工具使用能力不足: 尽管工具使用是基于大模型智能体的核心能力 [6], [7], 但现有大模型在高效学习和应用新工具方面表现出有限的容量, 这显著制约了智能体的性能。(iii) 性能饱和: 由于无法积累任务相关的洞察并利用历史经验来优化长期交互中的决策策略, 基于大模型的智能体在迭代性、长时程任务中表现出持续失败。因此, 智能体可能反复犯下类似错误, 而未展现出纠正错误的学习行为以成功完成任务。

为应对这些挑战, 记忆 [8] 已成为推动大模型智能体实现四大目标的关键组件: i) **个性化与特定化**。[9]: 记忆使智能体能够捕捉用户偏好、交互历史以及任务相关的上下文信息, 从而提供定制化响应, 例如在软件工程中记住工作流习惯, 或在对话场景中识别沟通风格。记忆将通用知识与具体上下文相连接, 既存储普遍事

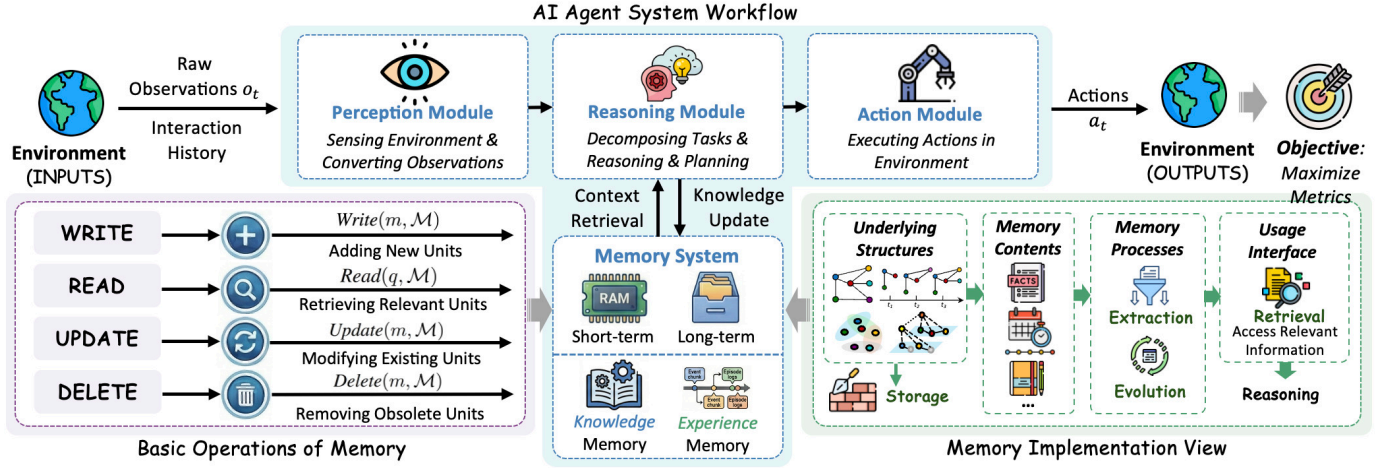


图 1: 一张展示人工智能智能体系统工作流及记忆系统详细实现的示意图。

实，也保留特定历史，使响应基于个性化、上下文感知的信息 [10]。ii) **超越上下文窗口的长期推理**。尽管大模型在有限的上下文窗口内运行，依赖静态参数化知识，但记忆系统提供了无界的外部存储，支持持续学习与适应。记忆使智能体能够在更长的时间跨度上保留信息，积累部署后的经验（包括成功与失败），并动态优化策略，而无需重新训练模型。iii) **自我改进** [11]：通过累积经验知识、推理模式和反馈，智能体记忆支持适应性与性能的迭代提升，从而实现基于大模型的智能体在不更新参数的情况下对任务的自我改进。iv) **幻觉缓解** [12]：将输出建立在结构化、可验证的记忆内容之上，减少了对可能不可靠的参数化知识的依赖。本质上，记忆将无状态的反应式模型转变为有状态的自适应实体，具备关系构建、基于轨迹的学习能力，以及日益复杂的个性化行为。

传统智能体记忆的实现主要采用线性的、非结构化的或简单的键值存储范式，例如固定长度的 token 序列、向量数据库和基于日志的缓冲区 [13], [14]。尽管这些框架能够实现基本的信息存储与检索，但智能体记忆需要更复杂的功能，例如关系建模、层次化组织以及因果依赖。基于图形的智能体记忆 [15], [16] 已成为 2025–2026 年研究的前沿方向，从被动的“日志”式事实记录转变为一种结构化的拓扑模型，用以保留信息随时间关联的方式。与传统的线性或非结构化记忆不同，基于图形的记忆由于其固有的实体关系建模能力、层次语义捕捉能力以及灵活的遍历与推理支持，能够自然地编码记忆元素之间的关系依赖。即便是普通的记忆，也可被视为具有平凡关系的退化图，这使得基于图形的智能体记忆成为智能体记忆设计的一种通用且灵活的框架。近

期，针对大语言模型智能体的基于图形的记忆架构研究迅速兴起，包括知识图谱 (KG)、时序图、超图、层次化树/图以及混合图 [17], [18]，这些方法在多种场景中展现出显著成效，例如层次化任务规划、多会话对话理解以及神经符号推理。

因此，我们提出了一项全面的综述，整合了基于图形的智能体记忆领域的最新进展，对其核心技术进行分类，综合其应用，并识别出开放性挑战。我们的贡献有四个方面：

- 我们提出了一种智能体记忆的分类体系，包括短期记忆与长期记忆、知识记忆与经验记忆、非结构化记忆与结构化记忆，并从基于图形的记忆实现视角进行阐述（第 III 节）。
- 我们系统地分析了关键的内存管理技术，涵盖内存提取（第 IV 节）、内存存储（第 V 节）、内存检索（第 VI 节）和内存演化（第 VII 节）。
- 我们总结了开源库和基准（第 VIII 节），这些资源支持在多样化应用场景（第 IX 节）中自演化基于图形的智能体记忆的开发与评估。
- 我们识别出关键挑战，并概述了未来的研究方向，以推动高效且可靠的基于图形的智能体记忆系统的发展（第 X 章）。

本综述旨在全面概述基于图形的智能体记忆，为研究人员提供有价值的见解以推动记忆设计的发展，并帮助实践者为特定应用选择合适的结构与技术。

II. 初步研究

Definition II.1 (AI Agents). 一个智能体是一种基于大语言模型的系统，由四个核心模块组成：

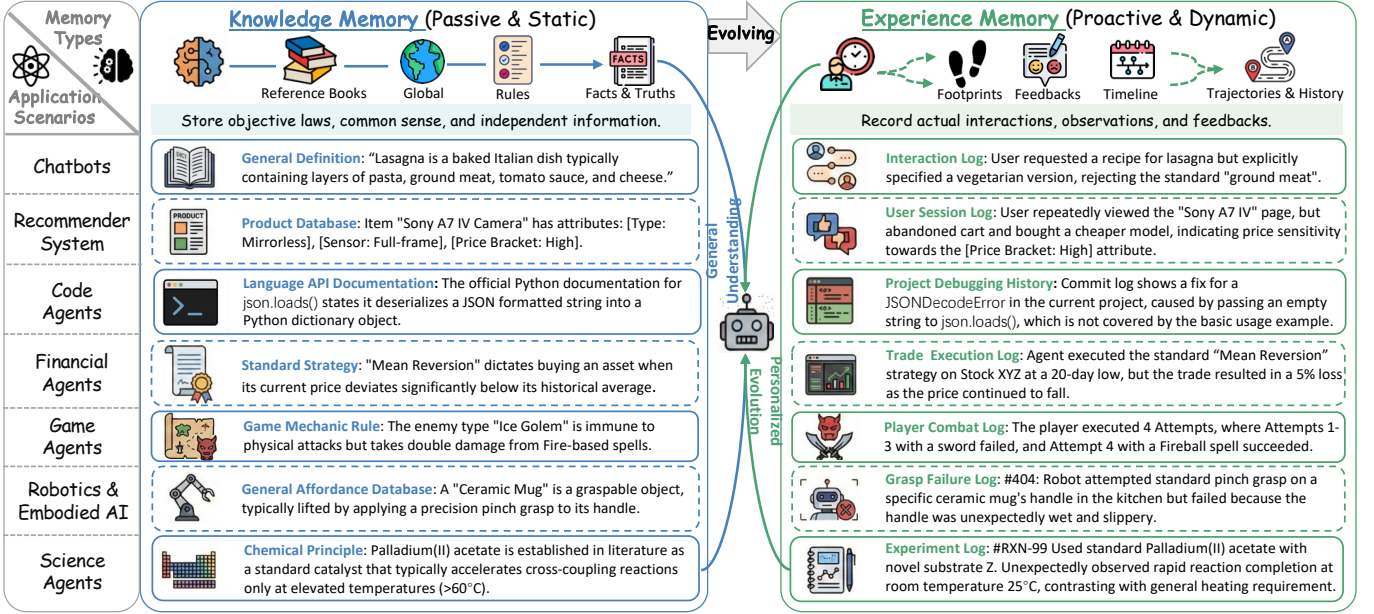


图 2: 两种类型的智能体记忆, 即 Knowledge Memory 和 Experience Memory, 以及它们在不同智能体场景中的应用。静态知识记忆与动态经验记忆之间的协同作用, 使智能体既能理解世界规则, 又能适应个性化交互。

- **感知模块:** 感知环境并将外部观测转换为内部表示。
- **推理模块:** 将复杂任务分解, 推理依赖关系, 与记忆交互, 并制定执行策略。
- **记忆系统:** 包括用于即时推理的短期记忆和用于经验保留以支持推理的长期记忆。
- **动作模块:** 在环境中执行动作。

人工智能智能体的目标是最大化期望的评估指标, 例如准确率、成功率和奖励。

在本节中, 我们介绍人工智能智能体的基础知识。人工智能智能体是一种基于大模型的系统, 能够利用大模型的推理、记忆和决策能力, 自主完成复杂任务。其形式化定义如定义 II.1 所示。人工智能智能体的典型执行范式是感知-推理-动作循环: 智能体通过感知模块从环境中接收输入, 然后利用大模型对输入信息、大模型的内部知识以及存储在记忆系统中的内容进行推理, 并最终输出动作以作用于环境。在此范式中, 记忆系统在智能体的推理与动作过程中起着重要作用。记忆系统的基本操作定义如定义 II.2 所示。

Definition II.2 (Basic Operations of Memory). 智能体记忆的基本操作定义了操纵记忆 \mathcal{M} 的基本动作。这些原子操作包括:

- **写入:** $Write(m, \mathcal{M}) \rightarrow \mathcal{M}'$, 将一个新的记忆单元 m 添加到记忆仓库 \mathcal{M} 中。

- **读取:** $Read(q, \mathcal{M}) \rightarrow \mathcal{M}_q$, 根据查询 q 检索相关的记忆单元 $\mathcal{M}_q \subseteq \mathcal{M}$ 。
- **更新:** $Update(m, \mathcal{M}) \rightarrow \mathcal{M}'$, 根据新信息修改现有的记忆单元。
- **删除:** $Delete(m, \mathcal{M}) \rightarrow \mathcal{M}'$, 从 \mathcal{M} 中移除过时或不相关的记忆单元。

这些操作共同实现了对智能体记忆系统的动态管理与演化。

尽管基本操作描述了对内存的原子动作, 但理解这些操作随时间如何协调对于智能体记忆系统的完整功能至关重要。因此, 我们通过生命周期的概念进一步形式化了内存的时序动态, 该概念描述了信息如何在不同处理阶段之间流动。

Definition II.3 (Lifecycle of Agent Memory). 智能体记忆的生命周期描述了智能体内部信息处理的连续性时间流, 定义了原始数据如何转化为存储的知识, 以及知识如何随时间演变。生命周期 \mathcal{L} 定义为一个包含四个不同阶段的循环过程:

- **记忆提取:** 将原始的非结构化观测 o_t 转变为结构化的记忆单元 m 。
- **记忆存储:** 将提取的单元通过适当的索引和结构化组织, 放入记忆结构 \mathcal{M} 中的组织与放置。
- **记忆检索:** 在响应查询 q 时, 检索相关存储信息的机制 $\mathcal{M}_{rel} \subset \mathcal{M}$ 。

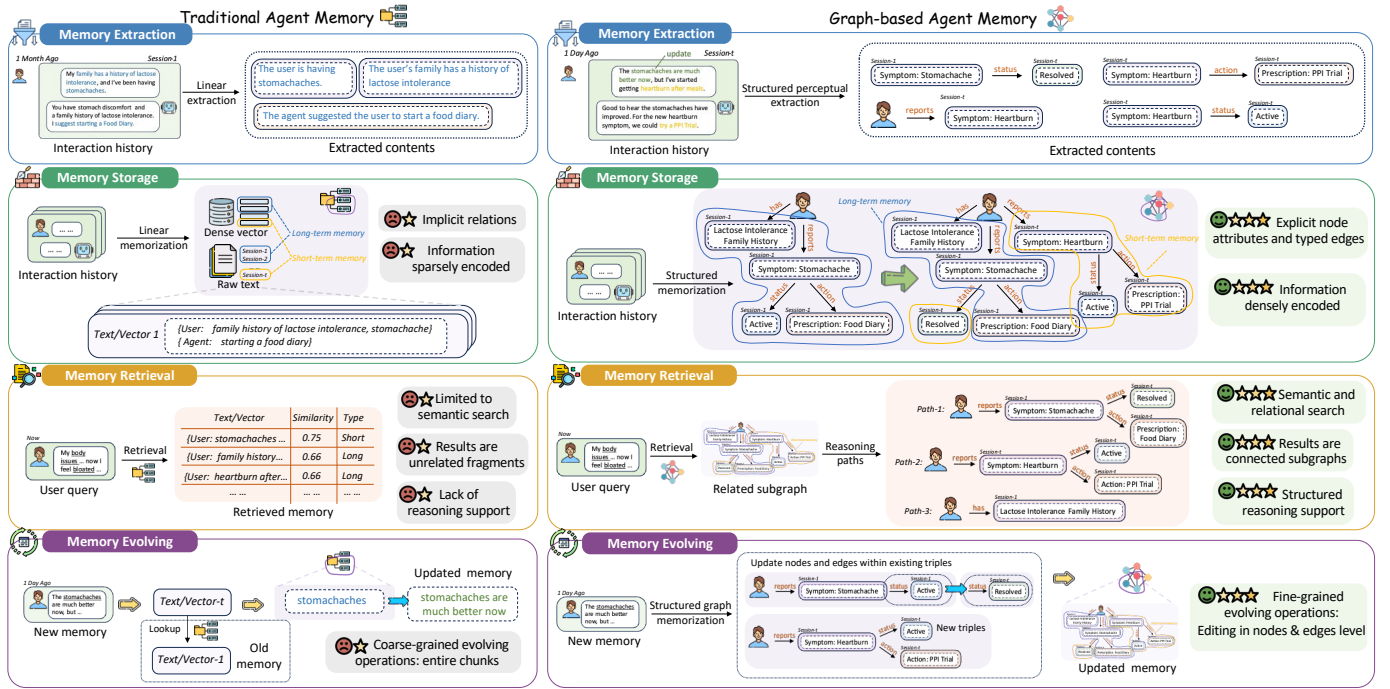


图 3: 传统智能体记忆与基于图形的智能体记忆的比较

- **记忆演化**: 后处理阶段, 其中记忆被优化, 包括内部自演化机制 (例如, 整合、抽象) 和外部自探索过程 (例如, 新的环境反馈)。

这一生命周期确保记忆保持最新、相关且结构优化, 以支持智能体的推理。

III. 智能体记忆的分类

在本节中, 我们从不同角度提出了智能体记忆的综合分类体系。具体而言, 记忆根据多个维度进行分类, 包括时间范围、功能角色和表示结构。随后, 我们引入基于图形的记忆作为智能体记忆的统一视角, 并从实现的角度对本节进行总结。

A. 长期记忆与短期记忆

记忆的一种最简单的分类是时间维度, 即信息保持的持续时间:

- **短期记忆**: 以快速访问和有限容量维持近期、立即相关的信息。这包括当前对话上下文、活跃的推理迹和临时状态变量。短期记忆具有易失性, 通常在即时任务完成后被丢弃, 但重要元素可能会被整合到长期存储中。
- **长期记忆**: 在会话之间存储持久信息, 包括累积的知识、历史交互、学成的模式和用户偏好。长期记忆支持回合之间的连续性, 实现迁移学习, 并为长时间尺度上的个性化和自适应行为提供基础。

B. 记忆的认知结构

我们介绍了记忆的认知结构, 如 [8]:

- **语义记忆**: 存储通用的、去情境化的世界知识和事实信息 (例如, “巴黎是法国的首都”)。在图结构中, 这构成了稳定的本体, 为智能体的推理提供常识和领域特定的事实基础。
- **程序性记忆**: 编码技能、常规和不可变规则。它表示“如何做”的知识, 例如标准操作流程或游戏规则。在应用中, 这使得智能体在标准条件下能够自动执行复杂任务。
- **联想记忆**: 在知识库中不同概念或数据点之间建立潜在的关联。通过连接相关的信息 (例如, 症状与诊断、产品与类别), 它促进了创造性推理和类比思维, 充当知识图谱的连接纽带。
- **工作记忆**: 作为智能体的“思维草稿板”, 用于存储即时体验。它临时保存当前对话轮次、中间推理步骤以及瞬时变量。尽管短暂, 却是所有经验记忆的入口, 具有快速访问的特点, 并对下一步动作产生直接影响。
- **情景记忆**: 记录过去会话的时间顺序。它将短暂的工作记忆变换为持久且可查询的自传式历史。例如, 它会记录客户曾请求过素食选项或订单被取消。这使得智能体能够召回“发生了什么以及何时发生的”, 提供时间上的定位。

- **情感记忆**：捕捉从交互中得出的情感基调或情绪。通过记录用户反馈或挫败程度，智能体可以在后续对话中调整其共情能力和表达风格。这一层为原始交互日志增添了定性维度。

C. 知识与经验记忆

为了具体定义智能体记忆，我们可以将其与人类认知系统进行类比。人类记忆是一个涉及信息的编码、存储和检索的结构化过程，通常被划分为不同类型。类似地，智能体记忆可以通过图 2 中所示的两个主要且互补的类别来理解：

知识记忆（被动且静态）：这代表智能体被动且静态的客观、全局和可验证信息存储库。它充当内部参考图书馆或教科书，包含正则事实、普遍规则、既定流程以及关于世界的普遍真理。这种记忆通常预先加载、更新缓慢且与上下文无关。其目的是为推理和行动提供一个稳定可靠的基石。例如，它存储概念的事实定义、游戏的不可变规则或标准科学原理。在应用中，购物智能体的产品数据库或机器人的一般可用性模型构成了知识记忆。通常，这种记忆是被动的，作为推理的可靠、事实性基础。

经验记忆（主动且动态）：这是智能体的个人日志簿，主动记录其具体的交互、观察以及动作结果。它包括用户对话历史、执行日志、试错轨迹和反馈。例如，它会记录某位用户曾要求一份素食千层面变体，基于均值回归策略的一次交易实际导致了损失，或标准抓取程序在湿杯子上失败。此记忆具有动态性、个性化特征，并构成从实践中学习并适应特定情境的基础。

D. 非结构性记忆与结构性记忆

传统的智能体记忆系统通常采用简单的存储范式，包括：

- **线性或缓冲区式记忆**，例如固定长度的 token 窗口或对话历史，虽然能保持近期交互，但存在信息丢失和缺乏关系上下文的问题。
- **基于向量的记忆**，将经验编码为存储在向量数据库中的稠密嵌入，支持语义相似度搜索，但在结构化推理和层次关系方面存在困难。
- **键值或日志型记忆**，通过顺序日志或属性-值对记录事件，支持直接查找，但在复杂查询或动态更新方面能力有限。

这些传统范式的一个共同特征是将记忆视为顺序的、平坦的或隐式结构化的存储。虽然在处理某些模式

时有效，但它们往往无法显式表示和高效推理知识片段之间的复杂关系网络，而这种能力对于复杂的规划、因果理解以及叙事连贯性至关重要。更重要的是，尽管这些方法能够实现基本的召回和短期上下文管理，但在需要层次化知识组织、时间追踪以及长期自适应学习的场景中表现出明显的局限性。这些约束在长时程任务、多会话交互以及知识动态演化的领域中尤为明显。

E. 基于图形的记忆：一种统一且通用的视角

基于图形的智能体记忆作为一种强大的通用化和增强型传统记忆框架，展现出显著优势。基于图形的智能体记忆的核心思想是将记忆内容建模为一种动态的、结构化的**记忆图**。在此范式中，记忆单元（如事件、实体、概念、观测）被抽象为**结点**，它们之间的语义、时间、因果或逻辑关系则被抽象为**边**。这种显式的结构表示将记忆从扁平的条目列表或隐状态向量，转变为丰富且相互关联的知识网络。

通过将记忆元素表示为结点，其关系表示为边，图结构天然支持：**①** 显式关系建模，使智能体能够推理记忆项之间的因果依赖和语义关联；**②** 层次化组织，从细粒度的事实三元组到一般的主题簇或子图；**③** 时序与动态结构，其中具备时间感知的边可以捕捉事件序列、状态转移和知识演化；**④** 高效的结构化检索，支持遍历、子图提取以及超越单纯语义相似度的多跳关系查询。

值得注意的是，传统的记忆形式可以被视为图记忆范式中的退化或简化情况。例如，线性缓冲区对应于图中的一条链，而向量记忆可被解释为具有相似度加权边的全连接图。因此，基于图形的记忆不仅取代了现有设计，还提供了一个统一且可扩展的框架。

本质上，基于图形的智能体记忆将记忆从被动的、扁平的“日志”提升为针对智能体生活经验量身定制的主动的、结构化的“知识图谱”。它不仅记录“发生了什么”，更重要的是建模“这些事物是如何相互关联的”。这为联想推理、长时依赖建模以及实现可解释的智能体行为提供了强大的基础。图 3 总结了传统智能体记忆与基于图形的智能体记忆之间的关键区别。总而言之，基于图形的智能体记忆将记忆重新概念化为一种主动的、结构化的经验模型。通过将关系视为第一类公民，它为复杂推理、长期一致性以及复杂自主智能体所需的自适应行为提供了强大基础。基于图形的记忆架构，如知识图谱和超图，在需要多会话一致性、个性化适应、复杂任务规划以及幻觉减少的应用中已展现出卓越性能。后

续章节将深入探讨此类记忆系统的技术实现，涵盖构建、检索、更新以及领域特定应用。

F. 实施视角下的另一种看法

在本文中，我们使用生命周期来介绍记忆，然而，从实现的角度来看，我们也可以对记忆提供另一种视角，这有助于研究人员或工程师理解本文。具体而言，该框架始于潜在结构，这些结构涵盖了基础的数据表示形式，包括基于图形的结构、嵌入以及时间序列，它们构成了记忆存储的基础（第 V 节）。存储的表示形式表现为各种记忆内容，包括对话历史、时间记录和结构化知识库。这些内容由记忆过程中的提取（第 IV 节）与演化（第 VII 节）进行管理，在此过程中，相关信息根据上下文和使用模式被筛选和优化。最后，使用接口实现了相关资讯的检索（第 VI 节），并通过访问经过整理的记忆来支持推理任务。这一实现视角从实现层面系统性地展示了语言模型如何维持、处理并利用记忆以支持复杂推理。

IV. 记忆提取：数据的变换

记忆提取过程始于原始数据的收集与预处理。这些格式各异的原始输入构成了智能体内部记忆构建的基础材料。从这一视角出发，记忆提取可从两个互补的维度来理解：记忆的来源以及提取后的记忆最终编码的形式。根据前一节所定义的内容，智能体记忆大致可分为知识记忆和经验记忆，二者主要区别在于其来源以及在系统中的作用。

经验记忆从智能体自身的交互历史中提取，反映了随时间积累的、与具体任务相关的情境化经验。其主要来源包括与用户的多轮对话、任务执行过程中产生的动作和观测序列，以及显式或隐式的反馈信号。**知识记忆**相比之下，来自独立于智能体个人经验的来源，旨在表示客观且普遍有效的信息。典型来源包括经过筛选的知识库、领域特定的数据库、正式文档、教学文本及其他权威语料库。此外，相当一部分知识记忆是通过在大规模文本数据上的预训练隐式获得的，其中事实性和程序性信息被嵌入模型参数中。来自这些来源的信息通常具有稳定性、上下文无关性，并具有广泛的适用性。因此，知识记忆的提取侧重于将正则的事实、规则和程序提炼为持久的表示，为跨任务和跨领域的推理提供可靠的基石。

潜在信息以多种不同格式呈现。这些数据源可能表现为非结构化或半结构化文本，如文档、对话转录和执

行日志。另一些则为非文本形式，包括图像、感官观测结果，或在交互过程中生成的结构化记录。数据源模式的多样性意味着原始信息无法直接作为记忆存储。相反，必须以反映数据源特征和所构建记忆类型的方式进行处理与抽象。因此，不同类型的数据源自然需要采用不同的提取策略。以下是针对不同类型数据源所使用的主要技术。

a) 从文本来源中提取：从非结构化文本数据（如对话日志或文档）中提取记忆，重点在于识别和构建语义信息 [19]。关键方法包括：(1) 结构化信息提取：使用命名实体识别和关系抽取模型，或提示大型语言模型，直接以（实体-关系-实体）三元组的形式提取实体、属性及其关系 [20], [21]。(2) 语义嵌入编码：利用 Sentence-BERT 等模型将句子或段落编码为稠密向量表示，将语义内容转换为适用于基于相似度检索的嵌入表示 [22]。(3) 摘要生成：应用抽取式或抽象式摘要模型（包括大型语言模型），将长篇文本或对话历史浓缩为简洁且信息丰富的记忆片段 [23]。

b) 从连续轨迹中提取：对于时间序列交互数据，如动作-观测序列，提取的目标是捕捉时间结构：(1) 事件分割与时间标记：将连续轨迹分割为离散且有意义的事件或回合，并为每个事件精确标注时间戳 [24]。(2) 动态状态快照：超越静态快照，该方法捕捉关键状态的演化过程。它涉及在关键时刻周期性地捕获并存储智能体或环境状态的紧凑表示，例如嵌入或特征向量 [25]。(3) 模式挖掘：应用离线序列挖掘或聚类算法，发现频繁出现的连续子序列或策略模式，并将其抽象为过程记忆模板。

c) 多模态数据提取：对于视觉或音频等感官数据，特征提取弥合了原始信号与语义意义之间的鸿沟：(1) 描述生成：使用视觉-语言模型 [26] 或音频字幕模型 [27] 生成视觉场景或听觉事件的文本描述，随后将这些描述作为文本进行处理。(2) 交互内容：该过程侧重于检测对象或从智能体的动作与环境响应（如原始像素或信号）之间提炼信息。(3) 联合多模态嵌入：利用定制化模型将不同模态的数据编码到统一的向量空间中，生成一个联合嵌入，作为多模态体验的紧凑表示 [28]。

如图5所示，提取过程将原始数据变换为结构化且语义丰富的表示。这种变换通常发生在三个抽象层次上：从原始的数据流，到中间提取出的实体和关系，最终转化为服务于智能体内部特定认知与操作目的的组织化功能记忆类型，相关内容将在第 III 节中介绍。

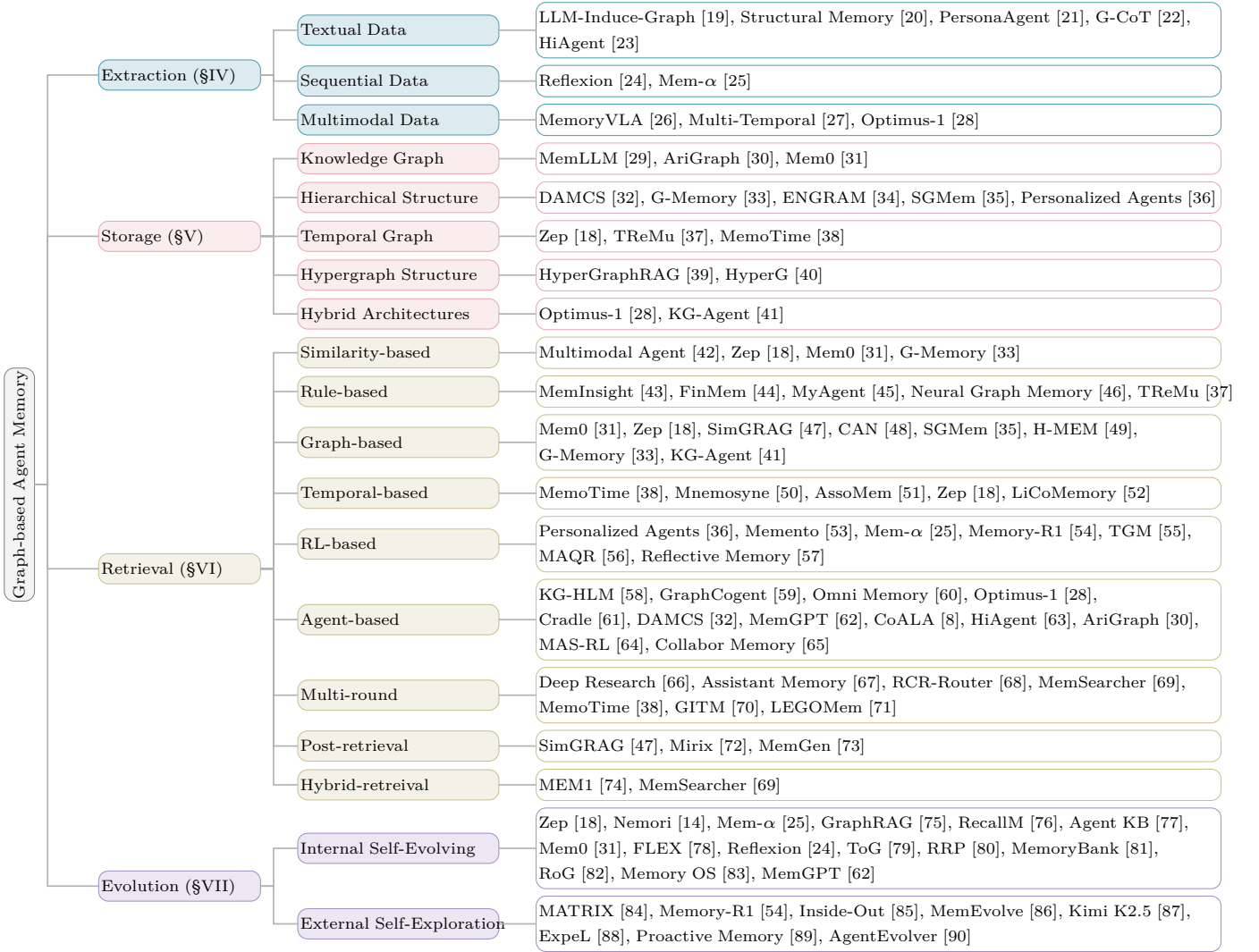


图 4: 基于图形的大型语言模型智能体记忆管理综合分类法。

V. 记忆存储：整理心智

上述提取阶段生成了一组语义丰富的成果，包括识别出的实体与关系、稠密语义嵌入、简洁摘要、带时间戳的事件分段以及多模态标题，这些共同构成了下游记忆架构的操作基础。因此，构建智能体记忆的核心挑战在于，将这些异构的提取成果转化为能够保留相关语义的同时支持高效检索和可靠更新的存储格式。与静态知识库不同，智能体记忆还必须适应动态性、个性化和经验锚定，这些因素共同决定了提取信息应如何编码与维护。

选择特定的图结构记忆机制实质上是在不同设计目标之间做出明确的权衡。准确率和显式的多跳推理倾向于使用关系图；压缩和概念抽象倾向于使用层次化或树状摘要；时间保真度则推动使用时间知识图谱和时间索引的回合；而跨模态泛化或模糊召回通常更青睐向

量存储或混合系统。考虑到这些权衡，本节首先聚焦于知识图谱作为正则的关系基础，阐述三元组的生成、集成与维护方式，随后综述了层次化、时间性、超图以及混合架构，它们在这一设计谱系中补充了不同的方面。图 6 概括了下文讨论的构建范式。

A. 知识图谱结构

知识图谱 (Knowledge Graph, KG) 作为一种结构化的记忆范式，专门用于存储和推理事实性知识 [29]。它将信息表示为一个由相互连接的三元组组成的网络，每个三元组的形式为 (头实体, 关系, 尾实体)。这种关系结构使其成为特定类型智能体记忆的强大基础。

a) 知识图谱建模：为智能体构建知识图谱涉及从非结构化交互中持续提取并整合结构化的三元组。其主要方法依赖于大模型作为强大的、开放词表的

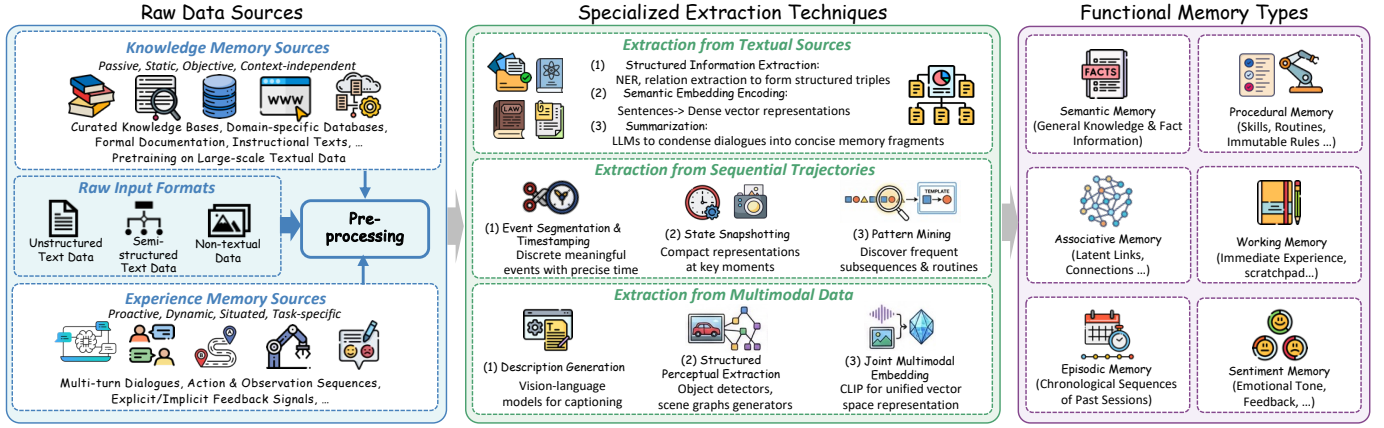


图 5: 智能体记忆提取概览。该图展示了从多种数据资源中构建智能体记忆的统一流水线。来自经验记忆和知识记忆的原始输入，通过专门的提取技术被变换为结构化且紧凑的表示。这些提取出的单元随后被组织成不同功能的记忆类型，使智能体能够支持推理和下游任务。

解析引擎。例如，在 AriGraph 世界模型 [30] 中，语言模型会解析环境中的每条文本观测，识别相关对象，并以三元组形式提取它们之间的关系。类似地，Mem0 [31] 等系统在其提取阶段也采用语言模型，将对话消息转换为实体-关系三元组。该方法优于传统的命名实体识别 (NER) 和关系分类流水线，因为大模型能够对新型细粒度实体类型进行泛化，并基于上下文理解挖掘隐含关系。

提取出的三元组随后进入更新阶段，将其整合到持久化的图存储中。该阶段并非简单操作，涉及冲突检测 (例如处理矛盾的事实)、关系剪枝和模式演化等操作。Mem0 通过专门的更新机制显式建模这一过程，将新记忆与现有的相似记忆进行对比评估。这种基于大模型提取、再经推理整合的两步流程，构成了从智能体原始感知构建动态、经验驱动的知识图谱的核心流水线。

b) 相应的存储类型：知识图谱 (KG) 的三元组结构使其特别适合实现长期静态记忆。这包括一般世界事实 (例如, *(Paris, Capital_Of, France)*) 和不可变的领域特定概念。显式的关联格式允许高效存储，并支持复杂且多跳的查询，这是向量检索器难以实现的。此外，通过为三元组添加时间元数据或将其与情景事件关联，其中情景顶点与来自同一观察的三元组相连，知识图谱也可以支持情景记忆。

B. 分层内存结构

层次结构是智能体记忆系统中组织知识最常见且最直观的范式 [32], [33], [36]。通过将信息排列成多级树形结构，它提供了一种合理的层级架构，将大量经验压

缩为可管理的模式。树的本质是有向非循环图 (DAG)，能够显式地建模父-子和包含关系，从而能够表示从广泛类别到具体实例，以及从高层次目标到细粒度执行步骤的概念。这一固有特性使得系统能够高效地进行自顶向下查询抽象主题，同时实现自下而上汇总以保持大规模记忆存储的一致性。诸如 MemTree 等系统采用此方法，通过动态地将新信息路由至层次结构中，在现有结点对相似内容进行聚类，同时为新信息创建新的分支，并递归更新所有祖先结点的语义摘要，以反映整合后的知识。

层次化记忆的构建主要依赖于两个过程：**语义聚类**用于组织，以及**递归摘要**用于抽象 [34]。为了保持层级结构作为压缩知识模式的价值，每个父结点动态地综合其子树内所有信息的简洁语言摘要。除了上述静态事实知识外，这种结构还表现为会话或序列记忆。在此，单个句子或对话回合主要通过其在特定会话内的时序顺序或对话流相连接。这形成了一条时间线或叙事链，对于建模单次交互回合中的对话连贯性及用户意图演化至关重要。SGMEM [35] 系统为长期会话智能体提出了一种句级树结构，利用话语之间的顺序和指代关系来构建连贯的交互。

C. 时间图结构

现实世界的交互本质上是动态的，其中事实仅在特定的时间窗口内有效。时间知识图谱 (TKGs) 将标准三元组扩展为四元组 (s, r, o, t) ，但简单的时间戳往往不足以处理复杂的智能体工作流。近期的架构引入了更

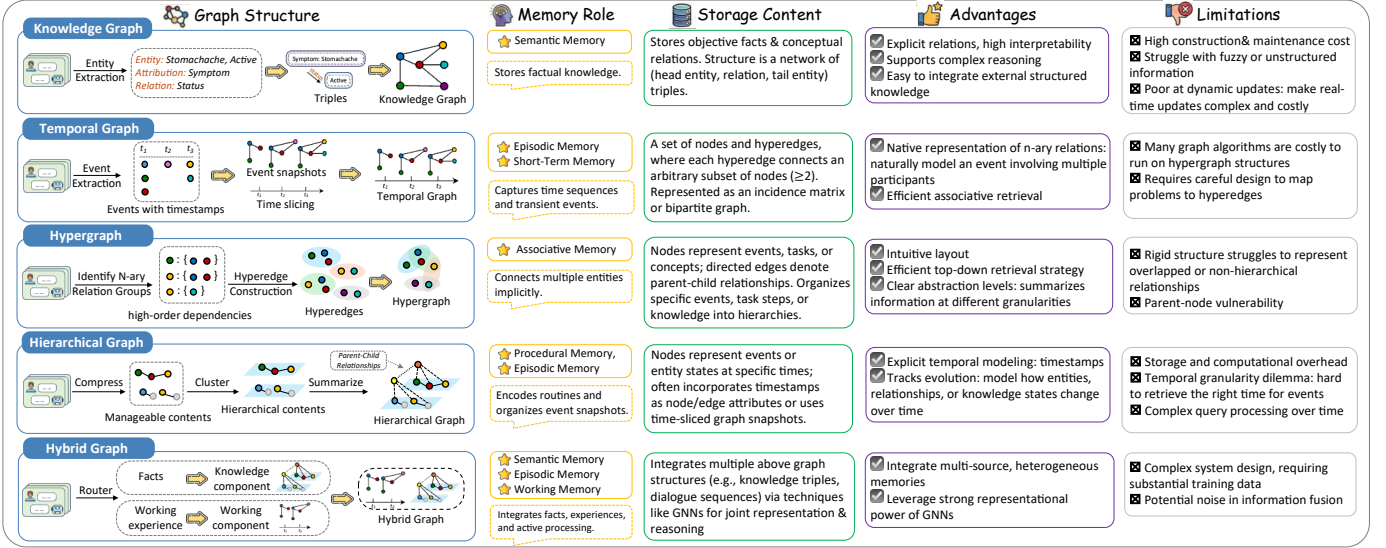


图 6: 智能体记忆系统中图构建范式的综合分类: 方法论、相应的记忆功能, 以及优缺点的对比分析。

细粒度的时间建模, 以应对有效性、分歧以及推理单调性等问题。

a) 双时相建模: 智能体记忆中的一个关键区别在于事件发生的时间 (有效时间) 与记录事件的时间 (事务时间)。**Graphiti** [18] 实现了一个双时间模型, 用于追踪两条不同的时间线。这使得系统能够处理不断演进的对话, 其中在 t_1 引入的事实可能事后描述 t_2 发生的事件。通过显式跟踪创建时间 ($t'_{created}$) 和过期时间 ($t'_{expired}$), 并结合有效性时间区间, 系统可以通过时间无效化来解决矛盾, 而非直接覆盖, 从而保持状态变化历史的忠实性。

b) 区分提及时间与事件时间: 在多层对话中, 相对时间表达 (例如 “上周五”) 常常引发分歧。**TReMu** [37] 采用一种时间感知的记忆机制, 将提及时间 (会话时间戳) 与推断出的事件时间解耦。**TReMu** 将记忆结构化为 “时间线摘要”, 其中事件按其推断出的绝对时间步进行分组和索引。该结构支持神经符号推理, 智能体生成 Python 代码对日期进行精确计算, 然后再检索对应的时间线结点。

c) 分层时间约束: **MemoTime** [38] 被提出以防止推理链中的逻辑幻觉 (例如, 检索发生在原因之前的效应)。该框架将时序知识图谱的推理过程组织成层次化的 “时间树”。与扁平化检索不同, 这种结构强制实现时间单调性, 确保任何检索到的推理路径 $e_1 \rightarrow e_2 \rightarrow e_3$ 严格遵守时间约束 ($t_1 \leq t_2 \leq t_3$)。这种面向操作符的设计使智能体能够有效剔除语义相关但时间上无效的证据。

D. 超图结构

虽然二元图 (连接两个实体) 效率较高, 但在表示复杂、多实体交互时会损失信息 (例如, 三种药物共同作用导致副作用)。超图通过使用可连接任意数量结点的超边来解决这一问题, 从而保留了 n -元关系的完整性。

a) N 元关系中的信息完整性: 超图记忆的主要动机在于防止将复杂事实分解为二元边所引起的稀疏性和语义碎片化问题。**HyperGraphRAG** [39] 表明, 超图结构表示在信息论上比二元等价形式更加全面。通过将自然语言知识片段及其所有相关实体视为单一超边 $e_i = (e_i^{text}, \{v_1, \dots, v_n\})$, 系统实现了 “双重检索”, 能够同时检索相关实体以及连接它们的超边, 从而获取完整的事实以用于生成。

b) 表格数据中的结构依赖: 超图在关系本质上为组群结构的数据中特别有效。**HyperG** [40] 使用超图对表格知识进行建模。它为行、列以及整个表格构建不同的超边, 以捕捉高阶依赖关系, 例如列内的语义一致性以及标题与单元之间的层次关系。为了增强推理能力, **HyperG** 采用了一种提示注意力超图学习 (PHL) 模块, 该模块根据具体查询动态地在结点与超边之间传播注意力, 从而有效模拟人类对相关数据子结构的关注。

E. 混合图架构

图结构在准确率和多跳推理方面表现优异, 但可能在向量检索的广度或非结构化缓冲区的灵活性方面有

所欠缺。混合架构通过将图与其他数据结构结合，以平衡这些权衡，通常将“静态知识”与“动态经验”分离开来。

a) 知识-经验解耦：一种主流的设计模式是将世界规则与智能体轨迹分离。*Optimus-1* [28] 提出了一种面向智能体的 **混合多模态记忆**。该模型结合了 分层有向知识图谱 (*HDKG*)，用于存储静态、结构化的游戏机制（以有向非循环图的形式建模制作配方），以及抽象多模态经验池 (*AMEP*)。AMEP 作为动态向量存储，保留了多模态的成功与失败轨迹。这种分离使得智能体能够在基于图形的知识基础上进行规划，同时通过从池中检索增强的经验来优化执行。

b) 外部图与内部工作记忆：另一种混合方法聚焦于外部大规模图与内部轻量级状态之间的交互。**KG-Agent** 框架 [41] 将外部知识图谱与内部知识记忆相集成。与纯图形化智能体不同，KG-Agent 维护一个结构化的临时记事本，迭代更新推理历史、工具定义以及从外部知识图谱中检索到的中间观察结果。这种混合设计使智能体能够利用灵活且不断演化的内部上下文来导航大规模静态图，实现了符号存储与神经推理之间的桥梁。

总而言之，不同类型的记忆通常需要不同的图构建方法，不存在适用于所有场景的单一范式。知识记忆通常具有稳定性、结构性和上下文无关性，非常适合强调关系或包含结构的图，例如知识图谱。这类图能够捕捉事实实体之间的显式关系，并支持对全局有效信息的高效推理。相比之下，经验记忆具有动态性、个性化和上下文依赖性，通常更受益于强调时间序列、交互轨迹或用户-动作网络的图。此类图能够捕捉智能体交互或偏好随时间演变的模式。因此，图类型的选取与底层记忆的特性密切相关。稀疏性、时间动态性、模态多样性以及增量更新需求等因素都会影响结点与边的定义方式。此外，有时需要采用混合或多层次的图来整合知识记忆与经验记忆。总体而言，图的构建需仔细考虑记忆类型的特点。

VI. 记忆检索：回忆过去

这些构建选择直接决定了内存使用情况。一旦内存结构确定，系统就必须决定如何访问这些内存以支持下游推理。这促使了存储流水线的下一阶段——检索阶段的到来。检索阶段通过定义可执行的操作符来操作图内存。基础检索操作符可以归纳为三种范式。这些范式与两种内存角色——知识和经验——相互作用。

语义检索包括基于相似度的算子，作用于提取的文本或多模态块及其嵌入。它支持模糊匹配和基本概念对齐，常被用作知识与经验记忆的候选生成器。

结构化检索包括基于规则的算子、时间算子和基于图形的算子。这些算子对结构化实体（如知识图谱、层次结构、时间图和超图）施加显式约束。这使得证据选择具有可验证性和可解释性：每次检索决策都可以通过潜在的图结构或规则集进行追溯。结构化检索在知识记忆中尤为关键。对于知识记忆，正确性和一致性是首要关注点。

基于策略的检索包括强化学习和 Agentic 检索算子，将检索视为序列决策过程。系统选择查询哪种记忆类型，并决定应用哪些算子。它分配计算资源并决定何时停止检索。基于策略的检索对于经验记忆尤为重要，因为经验记忆具有动态性、个性化和时效性特征。

在实际应用中，系统通常会组合使用这些基本操作符，例如语义锚定 → 结构化扩展 → 策略控制的停止与剪枝。

A. 检索技术的分类

检索过程可分解为三个基本操作：(i) 查询预处理，(ii) 候选检索，以及 (iii) 剪枝。它们通常以简单的流水线形式执行，从记忆图中提取少量相关证据。在本节中，我们介绍一组基础检索算子，这些算子可以灵活组合，用于实现后两项操作。结合预处理，这些算子构成一个端到端的检索流水线。除了这些算子外，我们还描述了检索增强策略，这些是叠加在基础算子之上的辅助方法，旨在提升检索质量。

1) 基于相似度的算子：基于相似度的检索是一种粗粒度的检索算子，它将用户查询编码为向量，然后在嵌入空间中检索与查询最相似的前- k 个记忆条目 [42]。

对于知识记忆，基于相似度的算子主要支持确切抽象概念或实体的召回。对于经验记忆，该算子作为特定记忆单元的召回机制，例如与当前查询匹配的回合、摘要或智能体状态。在实际应用中，系统通常将基于相似度的检索与经验记忆和知识记忆相结合 [31]。对于经验记忆，系统使用摘要图，其中检索器首先匹配高层摘要，然后深入到支持性的原始语料块。对于知识记忆，系统使用知识图谱，其中检索出的实体或三元组作为后续图谱扩展的锚点。

在简单的应用中，直接相似度搜索可以有效 [18]。它也可以起到辅助作用，例如，可以使用查询-查询 [33]

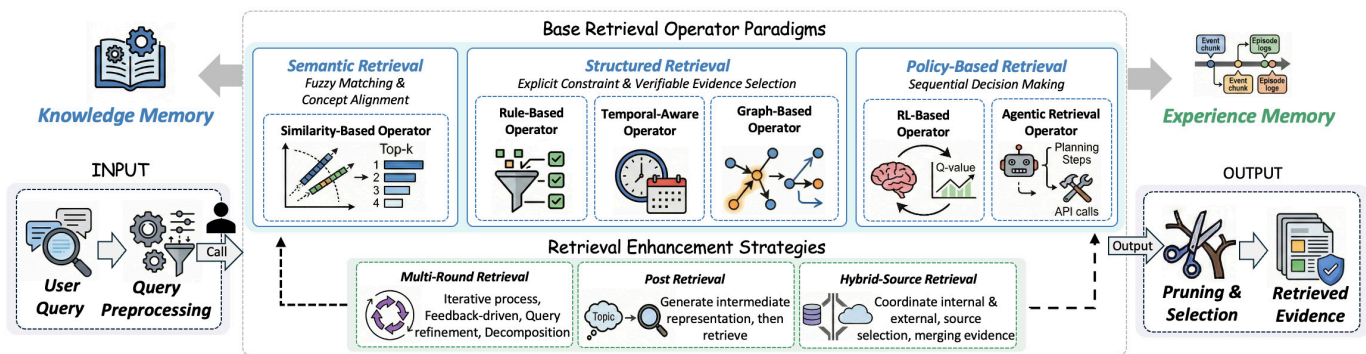


图 7: 检索流水线架构，整合基础操作符与增强策略。左侧：用户查询在检索前经过预处理。上方：六个基础检索操作符被组织成三种范式：语义、结构化和基于策略的检索，它们与知识和经验记忆类型进行交互。下方：检索增强策略层叠加在这些操作符之上：多轮检索、后检索以及混合源检索，将内部记忆与外部资源进行协调。右侧：最终的剪枝与选择生成经过排序的检索证据，供下游推理使用这些操作符与策略。

相似度来过滤无关的记忆。然而，这种方法存在明显的局限性：

- **相似度并不保证相关性。**词汇上或语义上相似的文本可能无法匹配任务所需的特定记忆。
- **多跳推理能力差。**复杂查询的答案通常依赖于原始查询中不存在的实体。单纯的相似度匹配无法弥补这一差距。
- **时间意识常常被忽略。**在动态情境中进行记忆检索需要理解时间与情境的相关性，而不仅仅是语义上的接近性。
- **可扩展性引入了噪声。**随着内存增大，语义相似的片段数量也随之增加。这导致了冗余或不相关的检索，从而降低了准确率。

这些挑战促使人们需要更复杂的检索方法，而不仅仅是相似度搜索。实际系统应利用与记忆组织结构相一致的结构化检索。此外，检索应由策略驱动：智能体根据潜在的记忆类型选择并组合操作符。

2) 基于规则的运算符：基于规则的图记忆运算符使用符号规则和可执行过滤器来判断相关记忆，而非单纯依赖语义相似度。

对于知识记忆，基于规则的算子发挥双重作用。它可以作为第一阶段选择器，对下游检索的候选对象进行预处理，也可以作为检索后的验证器，通过施加硬约束来剔除检索到的记忆。对于经验记忆，基于规则的检索主要支持对动态回合（如对话历史和执行轨迹）的时间范围限定，因为经验具有噪声且持续演化 [45]。在实际应用中，对于知识记忆，基于规则的算子可将候选对象限制在兼容的实体或关系类型内，优先选择高置信度三

元组，并排除被预定义规则标记为冲突的三元组。对于经验记忆，它们利用常见的规则，如时间窗口过滤和任务阶段约束，例如执行 [43]。

一种常见变体使用手工设计的关联启发式规则，灵感源自赫布型更新 [46]：经常共同被检索的记忆之间其连接会被加强，而最近写入或反复引用的项目会随着时间推移被赋予更高的权重。这类轻量级更新规则可通过检索相互关联的记忆块而非孤立项目来提升整体一致性 [44]。

基于规则的检索也经常应用确定性符号过滤来满足显式的查询约束。在结构化后端中，例如 SQL 数据库，这些约束可以编译为可执行的查询或程序，例如查询和连接表的 SQL 或 Python 代码，从而提高准确率并使检索流水线可审计 [37]。

3) 基于时间的算子：基于时间的算子对于处理依赖事件发生时间的查询、追踪随时间变化的事实以及保留对话中交互的顺序至关重要 [38], [50]。

对于知识记忆，普遍真理无论何时学成都保持有效，因此时间操作符主要作为过滤规则，而非核心检索机制。对于记录互动与事件的经验记忆，其相关性随时间变化，时间操作符便成为主要手段。用户的兴趣会演变，近期的工具故障比旧的更值得关注，系统必须区分上周有效的方案与去年有效的方案。因此，时间操作符按事件发生的新近程度对回合进行排序，随时间推移降低过时偏好的权重，并检索形成连贯叙事的事件链，而非来自不同时期的孤立记录。

在实际应用中，Zep [18] 为每个事实维护显式的时序窗口，标记其生效时间和过期时间。这使

得系统能够在查询时判断某个事实是否仍然适用。LiCoMemory [52] 在排序过程中应用衰减函数,降低较旧事实的权重,同时保留高度相关的信息。系统还会解析用户的查询以推断意图的时间范围,例如从“上个月我讨论了什么?”中提取“上个月”,然后将检索限制在该时间窗口内 [51]。这些方法提升了那些本身依赖时间信息的问题的召回率。

4) 基于图形的算子: 基于图形的检索算子遍历经验或知识图谱记忆中的显式关系链接。它使用查询条件驱动的遍历方式,从锚点结点扩展到同一图谱或其他图谱中与任务相关的子图。

对于知识记忆,图通过将事实编码为关联的单元来显式支持推理。检索可以施加结构化的关系约束,并返回诸如短路径或诱导子图等解释性结构。它支持组合查询,能够跨图连接多个事实 [41]。对于经验记忆,图通过保留回合之间的连通性来支持推理。它使智能体能够重建连贯的状态-动作-奖励链,并将奖励归因于特定条件。它能够检索与后续动作紧密相关的邻近动作,这些动作对主动适应至关重要 [35]。

在实际应用中,知识记忆与经验记忆共享一个共同的工作场景: 首先,识别关键实体作为锚点; 其次,通过邻域遍历扩展候选节点,例如在 n 跳以内; 第三,对结点或路径进行得分并剪枝,以获得紧凑的证据。在知识记忆中,锚点通常是实体或概念,遍历通常在知识图谱上受关系约束,以支持类逻辑检索。在经验记忆中,锚点通常是来自对话片段和执行日志的情境描述符,如工具输出或环境状态,遍历沿着事件和时间链接展开,以恢复情景上下文 [47]。

a) 层内遍历: 例如,Mem0 [31] 采用以实体为中心的方法,在识别出的实体周围扩展关系。Zep [18] 通过广度优先的邻域扩展来增强检索,以收集限定跳数范围内的候选结点或边,之后可根据相关性进行重新排序。H-MEM [49] 在层次化记忆树上使用基于索引的路由,逐层检索相关内容。一些方法进一步采用基于图神经网络的模型,以获取跨层表示用于检索和打分 [48]。

b) 层间遍历: 在分层图记忆中,遍历也通过层间边在抽象层次间进行,例如摘要结点与对话片段之间的边,实现自下而上抽象。G-Memory [33] 执行双向遍历,从包含详细交互日志的交互图中合成泛化策略,构建洞察图。LiCoMemory [52] 使用显式的超链接,从抽象摘要遍历到包含支持证据的精确对话片段。可训练图记忆 [55] 在查询、转移路径和元认知层之间聚合跨层

路径强度,以推导出用于检索的相关性得分。

5) 基于强化学习的操编辑: 强化学习 (RL) 正越来越多地用作训练算子,通过优化由潜在记忆定义的下流任务奖励来学习自适应检索策略。

两种记忆类型具有不同的用途: 知识记忆主要为该操作符提供稳定的支撑和约束,而经验记忆则支持上下文敏感的适应性。上下文敏感的适应性是指根据智能体当前的上下文(如当前任务状态),通过召回并排序先前观察到的状态-动作-奖励回合来调整检索动作。它们具有相同的目标: 检索始终有用的证据,同时降低语义相似但系统性误导的证据候选者的权重 [25]。这使得鲁棒性超越了固定的相似度启发式方法。在实践中,基于强化学习的检索操作符可应用于任何结构化记忆系统,只要检索动作定义明确、存在反映检索效用的奖励信号,并且策略能够训练以平衡性能提升与检索成本之间的关系 [56]。

一种常见的实现方式是将基于嵌入的候选生成与一个学成的动作-价值函数 $Q(s,a)$ 相结合,其中状态 s 概括了当前查询,可选地还包括智能体状态。动作 a 对应于检索决策,例如选择记忆项、选择图导航步骤或发出工具调用 [36], [53]

除了单步选择之外,一些研究将检索建模为一个序列决策过程,并使用策略优化训练专用的记忆智能体,例如 PPO/GRPO [55]。在典型的流水线中,记忆智能体决定要检索的内容,在某些设计中还决定写入内存的内容。随后,它将检索到的记忆 M_{ret} 传递给一个可能固定的回答智能体,以生成最终响应。奖励基于任务特定的回答质量指标计算,例如问答任务中的 EM/F1,用于更新检索策略以最大化性能 [54]。

由于在大模型上进行端到端的在线强化学习可能代价高昂,一种替代方案是通过回顾性标注实现无强化学习的策略学习。在这种方法中,一个专家大模型评估哪些检索到的记忆真正有用,生成监督信号以训练一个轻量级的检索器,例如一个 MLP 排序器。另一种方式是当候选集较大时,使用轻量级启发式评分函数来指导选择 [54], [64]。当大规模在线强化学习不可行时,这提供了一种在自适应性训练成本之间的实用权衡 [57]。

6) 基于智能体的操作员: 基于智能体的算子将检索视为一个开放的规划-反馈环,其中智能体能够突破内部记忆图谱的边界 [30], [58]。与基于强化学习的方法的核心区别在于,该方法能够调用外部工具和 API 来补充内部记忆,然后将经过验证的结果存储回记忆中以

供未来使用 [65]。这使得检索能够超越从固定知识库中选择的范畴。

对于知识记忆，智能体通过自规划遍历知识图谱以收集支持性事实 [59]。图遍历使智能体能够发现结构化记忆中的推理链。当内部图覆盖不完整时，智能体可利用内部知识来验证信息的正确性。对于经验记忆，智能体不遍历图，而是通过追踪回合间的动作迹来恢复情景上下文 [8], [63]。与导航结构化关系不同，智能体从带时间戳的日志中重构事件以做出决策 [47]。

动作空间包括选择相应的存储器和索引，并在知识图谱索引（用于规则或事实）与时间索引或超图索引（用于过去的回合）之间进行选择 [61]。智能体通过选择下一个要访问的结点或边，或跨抽象层移动来导航所选的记忆结构。智能体维护一个有限的工作记忆，记录当前问题、部分计划、每条证据来自哪个记忆源、先前访问过的结点以及到目前为止收集到的证据。这使得闭环导航决策成为可能 [28], [60]。

除了预定义的动作空间外，智能体还可以主动编写 API 调用（例如 SQL 查询或搜索请求），按需检索信息。这使得检索决策与智能体的规划过程保持一致 [32], [62]。在复杂系统中，责任可以分配给多个智能体，例如候选提案与排序，以提高鲁棒性。

B. 检索增强策略

除了基础检索算子外，近年来的图记忆系统越来越重视检索增强策略。这些方法本身并不定义新的算子，而是通过在基础算子周围增加额外步骤来提升检索效果，而非一次性查找。这些策略在前述两种记忆角色中均能发挥作用。知识记忆需要强调可靠性和冲突处理的增强策略，因为稳定的知识和规则必须在多次检索中保持一致。经验记忆则需要优先考虑时间连贯性、个性化以及迭代证据收集的增强策略，因为它记录了智能体与用户实际经历的情况。

在本综述中，我们将它们分为三类：i) **多轮检索**，通过重复检索来增加搜索深度和覆盖范围；ii) **后检索**，通过先生成中间表示（例如主题或意图描述），使查询更清晰，然后再进行检索；iii) **混合源检索**，通过从内部记忆和外部源同时检索，并结合结果来提高答案的完整性。

1) 多轮检索：多轮检索将记忆访问视为一个迭代过程，而非对记忆的单次查询 [66]。每一轮都基于原始查询和之前检索到的记忆生成下一个检索查询，检索更

多记忆，然后聚合并评估累积证据是否足够 [68], [69]。如果足够，则终止；否则，触发下一轮。更广泛地说，多轮检索可以通过策略实现，这些策略决定是否重新查询、如何优化查询以及何时停止。这种环路使检索过程显式地依赖反馈，而非静态 [66], [67]。

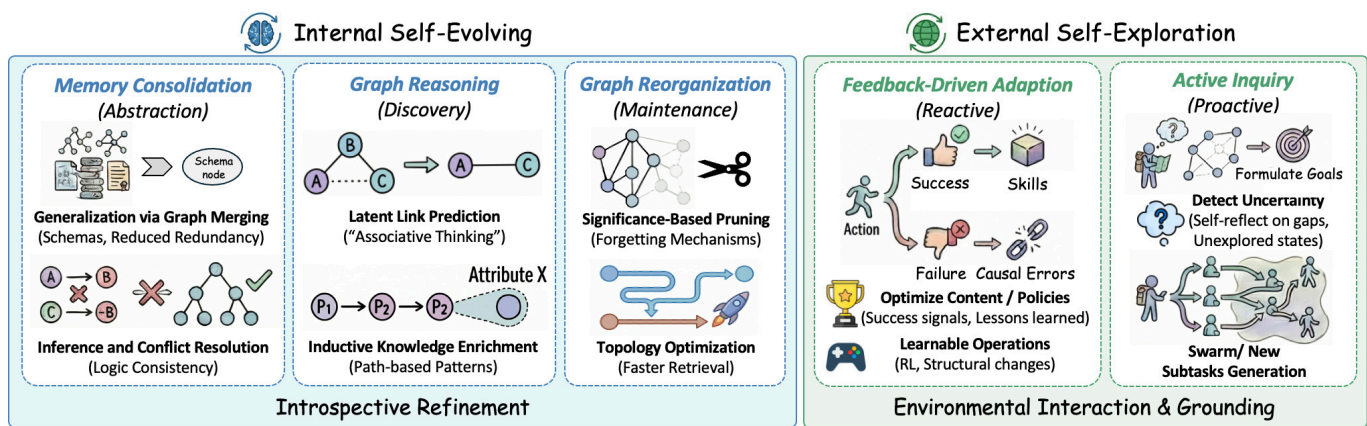
此外，一些系统将复杂查询分解为子查询 [38]，并在子查询层面执行检索，可选地重写每个子查询以使查询更具表现力。这种细粒度的分解可以减少查询的语义漂移，并实现有针对性的证据收集 [15]。

对于知识记忆，通常会使用额外的回合来收集支持性事实或规则，并验证检索项之间的一致性。对于经验记忆，通常会使用额外的回合来减少上下文缺失，例如最近的用户请求、最近的失败情况或最新的环境变化。在实际应用中，这也改变了停止规则：以知识为导向的环在证据一致且得到充分支持时停止，而以经验为导向的环则在收集到足够多最近且相关的回合后停止 [70], [71]。

2) 检索后：尽管大多数记忆增强系统在生成前检索记忆，后检索则遵循“先生成再检索”的模式。系统首先生成一个中间表示，例如主题、意图描述符、假设的实体和关系，或草稿结构，然后基于该中间表示进行检索。例如，模型可能在回答之前生成一个高层主题，并利用该主题来检索记忆，或者将查询转换为一个想象中的子图，并选择使图距离最小化的候选子图 [47]。这种设计的动机源于交互场景中的一种常见失败模式：用户查询通常模糊或表述不清，即使基于大语言模型的查询重写也可能无法可靠地将此类查询映射到正确的记忆上。通过引入主题生成步骤，后检索对查询表面表述的敏感性降低 [72]。

除了显式的符号查询之外，还存在生成式检索的变体，其中模型利用其当前的潜在推理状态来生成一系列潜在记忆 token 表示。然后，它通过在该潜在空间中的嵌入相似度来检索记忆。模型可以使用答案质量作为学习目标进行训练，从而鼓励那些能够揭示更有帮助记忆的潜在记忆 token 表示 [73]。

对于知识记忆，中间表示通常被调整为更接近正则形式，以便系统能够匹配规则和事实并加以验证。对于经验记忆，中间表示通常包含用户未明确说出的缺失上下文，例如用户目标、约束条件或已尝试过的内容。这在系统必须从大量相似日志中恢复正确回合时非常有用。



3) 混合源检索: 图记忆检索在混合源设置中受到越来越多关注, 其中系统需协调内部记忆与外部资源。在此情景下, 外部知识被视为一种额外的可检索资源, 与记忆存储并列 [69]。例如, 本地文档索引、返回标题和摘要及网址的在线搜索 API, 以及可通过智能体接口访问的任务环境。一个关键挑战是源选择问题。系统必须决定何时依赖内部记忆, 何时进行外部检索, 并如何融合证据或解决不同来源间的冲突。

混合源检索自然支持知识记忆与经验记忆。外部源在提供频繁变化的信息或超出系统已有认知范围的内容方面表现优异, 这更接近于知识记忆。内部经验主要提供个性化和局部的细节, 例如用户的先前购买历史, 或过去尝试中哪些工具失败了。当两者存在分歧时, 合并规则应取决于所检索的内容。若系统检索的是事实, 则应优先考虑可通过多个独立来源验证并追溯至权威原点的证据, 因为过时或不可靠的外部数据比可信的内部知识记忆更差。若系统检索的是个人经验, 则应优先采用与正确用户及正确时间范围相匹配的内部记录, 因为外部数据无法捕捉这种个体性 [74]。

VII. 记忆演化: 随时间学习

当智能体系统在长时间内与动态环境交互时, 其记忆不应保持静态, 而必须不断演化以融入新信息、解决不一致之处, 并适应变化的上下文。基于图形的记忆结构因其对关系连接、时间依赖性和有效性的显式建模, 特别适合演化, 能够通过结点/边/子图操作实现直接更新。受认知科学的启发——人类记忆通过突触可塑性等机制进行巩固和适应——基于图形的智能体记忆通过内部自我反思与外部探索来实现演化。这种双重方法解

决了新旧记忆之间的冲突 (例如过时的事实与近期观察结果), 同时增强了长期一致性与适应性。如图 8 所示, 我们将记忆演化分为两种互补范式: i) **内部自演化**, 专注于内在图操作以在无外部输入的情况下维持一致性; ii) **外部自探索**, 通过与环境的主动交互自动实现记忆的锚定与优化。这些机制将被动的知识存储库转变为积极的学习系统。

A. 内部自我演化

内部自演化将智能体的记忆图视为一个闭环系统, 具备自我反思与优化的能力。这一过程类似于人类在睡眠或休息时的记忆巩固, 大脑在此期间整理近期经历, 抽象出通用规则, 并遗忘琐碎细节, 而无需依赖新的外部感官输入。在基于图形的智能体记忆背景下, 这涉及重新组织图结构以提升检索效率、逻辑一致性和泛化能力。与传统仅简单追加日志的存储方式不同, 内部演化对图结构施加结构性变换, 通常聚焦于三个方面: **记忆巩固**、**图推理**和 **图重构**。

1) 记忆巩固: 内部演化的核心是从原始经验数据中抽象出高层次知识。智能体积累了大量的经验记忆 (具体轨迹或交互日志)。内部演化机制 [14], [18], [25] 分析这些回合, 以推导出泛化的知识。

- **通过图合并实现泛化:** 当智能体观察到多个相似的独立事件时, 它可以将这些子图实例合并为一个泛化的模式结点 [75]。这减少了存储冗余, 并创建了“技能”或“事实”的正则表示 [76], [77]。更先进的方法如 Mem0 [31] 和 FLEX [78] 采用基于大模型的语义门控机制, 其中模型评估新轨迹的信息增益

- **推理与冲突消解**: 通过内部推理（例如逻辑推理或图遍历），智能体可以检测结点之间的隐式矛盾。例如，若结点 $A \text{ implies } B$ ，但结点 $C \text{ implies } \neg B$ ，系统将触发自我修正过程，基于得分或时间新近性来解决冲突，更新图结构以保持逻辑一致性 [24], [79], [80]。

2) 图推理: 除了整合现有结点外，内部演化还涉及发现潜在关系以解决记忆稀疏性问题。智能体主动扫描其记忆图，识别缺失的边，或推断出未被明确观察到但可从现有事实中推导出的新关系。这一过程将记忆从一组稀疏孤立的轨迹转变为一个高度连接的知识网络。

- **潜在链接预测**: 利用大模型的语义能力，智能体可以预测不相连子图之间的潜在关系。例如，如果记忆中包含 $(A \xrightarrow{\text{cause}} B)$ 和 $(B \xrightarrow{\text{cause}} C)$ ，智能体可以自主推断并插入一条传递边 $(A \xrightarrow{\text{leads_to}} C)$ [79], [82]。该机制模拟了人类认知中的“联想思维”，使智能体能够在无需外部新输入的情况下，将时间上相距较远的事件“串联起来”。

- **归纳知识增强**: 智能体可以采用基于路径的推理来推导新的属性或事实。通过遍历图结构，系统识别出模式，并以推断出的“协变量”或高阶关系 [75] 来丰富图结构。这有效缓解了初始记忆构建的不完整性，确保未来检索能够访问那些从未明确写入日志但具有逻辑导数关系的信息。

3) 图重组: 记忆的无限积累会导致检索延迟和噪声。内部自我演化的机制会采用类似生物系统的“遗忘”机制，以维持图结构的健康与质量。

- **基于重要性的剪枝**: 智能体采用算法（例如 PageRank 变体或衰减函数）来评估记忆结点的效用。那些很少被访问或对近期决策贡献较低的结点将被剪枝或压缩 [62], [81], [83]。
- **拓扑最优化**: 这涉及重新调整边的结构，以缩短频繁关联概念之间的路径长度 [76], [91]。通过增加边的权重或在相关结点之间创建捷径，智能体对图进行优化，以实现未来更快的检索，将其经验整合为更高效的结构。

B. 外部自我探索

尽管内部演化基于内在一致性对记忆进行优化，但无法验证知识在现实世界中的有效性。**外部自我探索**通过环境交互来实现记忆演化的基础，弥合了这一差距。与被动记录所有事件不同，有效的探索利用环境反馈（例

如成功信号、错误）来区分信号与噪声，并主动寻求缺失信息。我们将其大致分为反馈驱动的适应（反应式）和主动探究（主动式）。

第一种范式，**反馈驱动的适应性**，侧重于根据过去动作的结果来优化记忆内容和管理策略。在开放式的环境中，原始的交互日志通常具有噪声；因此，智能体利用任务执行结果作为监督信号，提炼出可操作的知识。像 ExpeL [88] 和 Matrix [84] 这类方法会根据成功或失败进行差异化处理：成功的轨迹被凝练为可重用的技能，而失败的轨迹则通过对比分析来识别因果错误，显式地编码“学成的经验”，以促进高价值行为。这种演化逻辑不仅延伸至记忆内容本身，也扩展到记忆管理策略。与依赖固定规则不同，Memory-R1 [54] 和 Inside Out [85] 等系统将记忆操作（如添加或删除）视为强化学习框架中的可学习动作。通过接收下游任务准确率的奖励，智能体能够自主学习最优的图结构维护策略。类似地，MemEvolve [86] 将这一反馈驱动原则应用于架构层面，动态调整存储结构和索引机制，使系统能够适应部署环境复杂性的变化。

然而，仅依赖反应式适应存在覆盖偏差问题，导致智能体的知识始终局限于其被分配的特定任务。为克服这一局限，第二种范式——**主动探究**，将智能体从被动学习器转变为积极探索者。先进的框架使智能体能够检测当前图中的不确定性，并自主制定目标以解决该问题。例如，ProMem [89] 使智能体能够对缺失的结点或模糊的转移关系进行批判性自我反思，进而生成具体的查询以填补这些知识空白。AgentEvolver [90] 将此能力扩展至任务空间，通过自主生成新的子任务，强制智能体导航至未探索的状态，预先在记忆中填充多样化的经验。为进一步加速这一过程，KIMI K2.5 [87] 引入了可扩展的群体机制，生成并行的子智能体以探索问题空间的不同分支。这使得中心记忆能够快速吸收多样化的视角和边缘案例，将系统从被动的经验容器转变为知识的主动构建者。

尽管在反馈驱动的适应和主动探究方面取得了进展，但显式利用图结构进行探索的研究仍较为不足。本文提出一些在自我探索中使用图结构的建议。首先，拓扑引导的探索可优先关注连接稀疏的簇，或连接孤立的子图，从而系统性地提升知识覆盖范围与连通性。其次，多粒度策略可在追求高层次关系模式的同时，通过有针对性的交互填充具体实例，实现发现过程的协调。这些方法使记忆图从被动存储库转变为动态构建并不断优

化自身知识边界的主动架构。

VIII. 开源库与基准

A. 开源库

在附录 B 的表 IV 中, 我们对十一类具有代表性的开源内存库在关键功能维度上进行了系统性比较。

多个库提供了基于图形的记忆表示与利用, 包括 Cognee [92]、Mem0 [31]、OpenMemory、MemMachine、Memory 以及 Graphiti。基于图形的结构天然与结构化检索相结合, 支持多跳或关系查询, 这对于实体、事件或概念的推理至关重要。在功能覆盖方面, OpenMemory 和 Mem0 [31] 作为最全面的图记忆工具, 支持记忆构建、由交互驱动的更新、生命周期管理、时间感知以及图管理。基于图形的记忆工具可直接集成到智能体架构中, 动态利用结构化知识, 支持长期检索、时间推理以及多步任务执行。Cognee [92] 提供可查询的图嵌入, Mem0 [31] 与 OpenMemory 支持会话感知的记忆更新, Graphiti 则支持用于多步规划的时间图推理。对于非图记忆系统, 记忆构建主要以交互或会话为驱动, 例如 LangMem、LightMem [93] 以及 O-Mem [60], 具备增量更新机制并支持检索。尽管它们的基本检索功能基于嵌入相似度。虽然轻量级工具如 Memori 和 MemMachine 专注于模块化记忆管理, 强调易集成性, 并通过 API 支持智能体条件化。

B. 数据集和基准

与标准的自然语言处理任务不同, 评估智能体记忆时应考虑信息在整个扩展交互过程以及不断变化的系统环境中的分布情况。该领域的有效基准更注重智能体在有限上下文窗口和计算成本限制下重用观测数据的能力。接下来, 我们重点介绍专为增强记忆的智能体设计的关键基准。这些选择如表 I 所总结, 基于统一的评估标准 (包括模态、环境真实性以及记忆类型) 进行对比, 以提供一个全面的基准概览。

a) 情景分类法: 我们通过基于情景的分类体系对现有基准进行归类, 以反映增强记忆智能体在多样化应用环境中的使用情况。该分类基于三个关键方面: 从多轮对话到长时程任务的交互模式, 包括基于网络或具身化以及工具辅助的运行界面, 以及信息复用的时间跨度。依照此结构, 我们识别出七个代表性情景: 交互、个性化、网络、长上下文、持续学习、环境以及工具/生成。

1) 交互: 多轮跨会话对话记忆: 交互情景中的基准测试关注智能体在多轮及跨会话对话中保持连续性的能力。在这些情景中, 对话早期引入的相关信息必须在后续准确回忆并应用。此类基准包括 LoCoMo [94]、LongMemEval [95]、MemoryAgentBench [96]、MEM-TRACK [97]、MADial-Bench [98]、MemSim [99]、ChMapData [100]、MSC [101]、MMRC [102]、MemBench [103]、StoryBench [104]、DialSim [105] 和 RealMem [106] 等数据集。这些基准优先考虑长程上下文、不同会话间的一致性, 以及在进行对话时对先前提及的用户提供的事实检索能力。

评估通常聚焦于长历史记忆召回、信息的上下文复用以生成连贯回应, 以及保持一致性以避免自相矛盾。这些能力对于助手型智能体至关重要, 因为记忆失效会直接导致用户体验下降。常用指标包括任务级别的准确率、以召回率 @k 为代表的检索导向度量, 以及对话层面的一致性比率。此外, 一些基准测试会追踪多轮任务中的成功率, 以检验召回信息是否被有效整合到回应中。一个显著的局限是, 许多交互基准在处理冲突事实时缺乏对记忆更新的显式监督, 因此智能体如何覆盖或遗忘过时信息的机制, 远不如其召回能力得到系统性评估。

2) 个性化: 用户画像、偏好设置与记忆更新: 个性化基准测试考察智能体管理持久性用户中心事实和配置属性的能力, 如 PersonaMem [107]、PerLTQA [108]、MemoryBank [81]、MPR [109]、PrefEval [110] 和 LOCCO [111] 所示。这些基准测试评估智能体是否能够构建稳定的用户模型, 并随时间整合新的用户信息。对于现实世界中的助手而言, 一个主要挑战是避免人格漂移, 即智能体忘记或违背先前建立的偏好。然而, 这些任务当前的一个局限性在于它们频繁依赖于明确的监督信号以确定应存储的内容。相比之下, 实际部署需要具备选择性写入和隐私敏感的保留能力, 而这两方面在现有研究中仍缺乏充分关注。

3) 网页: 长时浏览与多步骤在线任务: 网页基准测试在扩展的动作轨迹中评估智能体的记忆能力, 要求智能体在多个步骤中跟踪环境状态和中间结果。像 WebShop [114] 和 WebArena [115] 这样的数据集聚焦于电子商务和功能型网站的交互。WebChoreArena [112] 针对复杂的浏览流程, 而 MT-Mind2Web [113] 强调对话式导航, MMInA [116] 则评估多跳网页交互。这些任务对经验记忆提出了较高要求, 因为智能体通常需要缓存

表 I: 按场景分组的智能体记忆基准。

Name	Scenario	Modality	Feature	Environment	Memory type	Link(Feb 2026)
LoCoMo [94]	Interaction	Text+Image	Long conversational memory	real	Factual	• Website
LongMemEval [95]	Interaction	Text	Long-term interactive memory	simulated	Factual	• GitHub
MemoryAgentBench [96]	Interaction	Text	Multi-turn interactions	simulated	Factual + Experiential	• GitHub
MEMTRACK [97]	Interaction	Text+Code+Logs	Long-term interactive memory	simulated	Factual + Experiential	• Website
MADial-Bench [98]	Interaction	Text	Memory-augmented dialogue generation	simulated	Factual	• GitHub
MemSim [99]	Interaction	Text	Bayesian memory simulation	simulated	Factual + Experiential	• GitHub
ChMapData [100]	Interaction	Text	Memory-aware proactive dialogue	simulated	Factual	• GitHub
MSC [101]	Interaction	Text	Multi-session chat	simulated	Factual	• Website
MMRC [102]	Interaction	Text+Image	Multi-modal real-world conversation	simulated	Factual	• GitHub
MemBench [103]	Interaction	Text	Interactive scenarios	simulated	Factual + Experiential	• GitHub
StoryBench [104]	Interaction	Text	Interactive fiction memory	mixed	Factual + Experiential	• Website
DialSim [105]	Interaction	Text	Multi-dialogue understanding	real	Factual + Experiential	• Website
RealMem [106]	Interaction	Text	Project-oriented long-term memory interaction	simulated	Factual + Experiential	• GitHub
PersonaMem [107]	Personalization	Text	Dynamic user profiling	simulated	Factual	• GitHub
PerLTQA [108]	Personalization	Text	Social personalized interactions	simulated	Factual	• Website
MemoryBank [81]	Personalization	Text	User memory updating	simulated	Factual	• GitHub
MPR [109]	Personalization	Text	User personalization	simulated	Factual	• GitHub
PrefEval [110]	Personalization	Text	Personal preferences	simulated	Factual	• Website
LOCCO [111]	Personalization	Text	Chronological conversations	simulated	Factual	• GitHub
WebChoreArena [112]	Web	Text+Image	Tedious web browsing	real	Factual + Experiential	• GitHub
MT-Mind2Web [113]	Web	Text	Conversational web navigation	real	Factual + Experiential	• GitHub
WebShop [114]	Web	Text+Image	E-commerce web interaction	simulated	Experiential	• GitHub
WebArena [115]	Web	Text+Image	Web interaction	real	Experiential	• GitHub
MMInA [116]	Web	Text+Image	Multihop web agent	real	Factual + Experiential	• Website
NQ [117]	LongContext	Text	Natural question answering	simulated	Factual	• Website
TriviaQA [118]	LongContext	Text	Large-scale question answering	simulated	Factual	• Website
PopQA [119]	LongContext	Text	Adaptive retrieval augmentation	simulated	Factual	• GitHub
HotpotQA [120]	LongContext	Text	Explainable multi-hop QA	simulated	Factual	• Website
2wikimultihopQA [121]	LongContext	Text	Multi-hop QA	simulated	Factual	• GitHub
Musique [122]	LongContext	Text	Multi-hop QA	simulated	Factual	• GitHub
LongBench [123]	LongContext	Text	Long-context understanding	mixed	Factual	• GitHub
LongBench v2 [124]	LongContext	Text	Long-context multitasks	mixed	Factual	• GitHub
RULER [125]	LongContext	Text	Long-context retrieval	simulated	Factual	• GitHub
BABILong [126]	LongContext	Text	Long-context reasoning	simulated	Factual	• GitHub
MM-Needle [127]	LongContext	Text+Image	Multimodal needle retrieval	simulated	Factual	• Website
HaluMem [128]	LongContext	Text	Memory hallucination eval	simulated	Factual	• GitHub
MemoryBench [129]	Continual	Text	Continual learning	simulated	Factual + Experiential	• GitHub
LifelongAgentBench [130]	Continual	Text	Lifelong learning	simulated	Factual + Experiential	• Website
StreamBench [131]	Continual	Text	Continuous online learning	simulated	Factual + Experiential	• Website
Evo-Memory [132]	Continual	Text	Test-time learning	simulated	Factual + Experiential	• Website
Ego4D [133]	Environments	Video+Audio	Egocentric episodic memory	real	Experiential	• Website
EgoLife [134]	Environments	Video+Audio	Long-context life QA	real	Experiential	• Website
ALFWorld [135]	Environments	Text	Household tasks	simulated	Factual + Experiential	• Website
BabyAI [136]	Environments	Text	Language navigation	simulated	Experiential	• Website
ScienceWorld [137]	Environments	Text	Multi-step science experiments	simulated	Factual + Experiential	• GitHub
AgentGym [138]	Environments	Text	Multiple environments	mixed	Experiential	• Website
AgentBoard [139]	Environments	Text	Multi-round interaction	mixed	Experiential	• GitHub
SWE-Bench [140]	Tool/Gen	Text+Code	Code repair	real	Experiential	• Website
GAIA [141]	Tool/Gen	Text	Deep research tasks	real	Experiential	• Website
xBench-DS [142]	Tool/Gen	Text+Image	Deep-search evaluation	real	Experiential	• Website
ToolBench [143]	Tool/Gen	Text→API	API tool use	real	Experiential	• Website
GenAI-Bench [144]	Tool/Gen	Text+Image	Visual generation eval	real	Experiential	• Website

页面状态以避免重复动作。同时，它们也凸显了资源高效记忆的必要性，因为过多的工具调用可能导致高昂的操作成本。一个普遍的挑战是，这些基准测试的成功有时可通过简单的启发式方法实现，因此要隔离记忆的具体影响，需采用受控情景，例如限制记忆容量。

4) 长文档理解与检索：长上下文基准测试衡量智能体在高容量输入和检索密集型情景下的表现。已建立的问答套件，包括 NQ [117]、TriviaQA [118]、PopQA [119]、HotpotQA [120]、2wikimultihopQA [121] 以及 Musique [122]，用于测试证据聚合与多步推理能力。更近期的框架如 LongBench [123] 和 LongBench v2 [124] 提供多任务评估，而 RULER [125]、BABI-Long [126]、MM-Needle [127] 以及 HaluMem [128] 则专注于针堆中找针式的检索和幻觉评估。尽管这些基准测试对于建模证据获取至关重要，但它们并不总是智能体记忆能力的完美指标。许多任务仍为单轮交互，不求智能体主动向持久化内存存储写入信息，可能导致长上下文处理与专用记忆机制混淆。

5) 持续性：终身学习与测试时适应：持续性基准评估智能体在不经历史灾难性遗忘的情况下是否能够随时间不断改进，通常在流式或顺序任务分布下进行。MemoryBench [129]、LifelongAgentBench [130]、StreamBench [131] 和 Evo-Memory [132] 等框架捕捉了在线更新和测试时适应的要素。此类基准严格意义上代表了终身记忆，要求模型在获取新知识的同时保持对早期任务的熟练度。尽管其重要性显著，但该领域缺乏标准化报告，因为遗忘和迁移收益的度量差异较大。此外，往往难以判断性能提升是源于参数更新还是对过往日志的检索。

6) 环境：具身化与互动世界：基于环境的基准测试在模拟或物理交互情景中评估智能体，要求智能体在部分可观测条件下从观测中提炼记忆。Ego4D [133] 和 EgoLife [134] 专注于第一人称情景记忆与多模态生活记录。ALFWorld [135] 和 BabyAI [136] 强调指令遵循与导航能力，而 ScienceWorld [137] 则测试多步骤实验能力。更广泛的评测套件如 AgentGym [138] 和 AgentBoard [139] 采用以规划为核心的分析方法，提供多轮评估。这些基准主要测试智能体在不同环境变化下的经验记忆与鲁棒性。然而，由于性能通常与特定环境技能密切相关，若要宣称记忆相关优势，必须仔细控制规划与工具使用等变量。

7) 工具/生成：工具使用与 workflow 执行：工具/生成基准测试评估在涉及外部 API 和迭代推理的工作流中的内存能力。ToolBench [143] 专注于 API 调用，而 SWE-Bench [140] 通过迭代调试聚焦于软件工程。GAIA [141]、xBench-DS [142] 和 GenAI-Bench [144] 测量复杂的科研与生成行为。这些任务强调过程记忆或保留中间假设和失败尝试的能力。它们也突出了操作性问题，例如迹的可追溯性和重试带来的财务成本。一个显著的障碍是评估复杂性，因为成功取决于环境稳定性和专用评分脚本的设计。

a) 评估景观的构建：总体而言，这些基准通过在交互式或长周期任务中测试智能体，提供了对智能体能力的全面视角。尽管记忆并非始终是唯一指标，但这些环境中的成功取决于智能体存储相关数据并利用过往经验做出当前决策的能力。未来的研究可通过消融实验来衡量记忆的具体影响，并重点关注智能体在动态情景中处理变化信息的方式。通过使用清晰的效率指标并确保环境可复现，这些基准有助于更好地理解实际中记忆行为的表现。

b) 战略基准选择：选择合适的基准取决于所研究的记忆能力。交互和个人化数据集最适合测试对话的持续性。网络和环境数据集更适用于评估智能体处理长序列动作和经验数据的能力。长上下文任务仍然是检查大输入中事实检索的标准。对于研究随时间学习的智能体，持续性基准可衡量信息保持的效果，而工具/生成任务则用于评估在执行复杂技术步骤时的记忆表现。

IX. 应用

基于图形的智能体记忆在对话聊天机器人、具身机器人以及科学智能体等领域具有广泛应用。通过解决长期知识保留、个性化交互、多步推理和自我演化等挑战，记忆能够提升大语言模型智能体在众多应用领域中的有效性和可靠性。本节将系统性地讨论当前及未来可能的应用。

A. 会话智能体

会话智能体，如 Claude¹ 以及 ChatGPT²，是基于大语言模型系统的最常见应用场景之一。智能体在多轮对话中面临长期保持连贯且个性化的对话内容的挑战，需要复杂的记忆系统来有效更新其知识和用户偏好。

¹<https://claude.ai/>

²<https://chatgpt.com/>

早期研究集中于记忆系统的可控性与稳定性。Memory Sandbox [145] 和 LD-Agent [146] 项目通过强调记忆系统的透明性以及事件记忆与个性之间的区别,为提升记忆系统的可信度奠定了基础。当代关于记忆的研究则转向利用结构最优化方法解决多会话对话系统中的上下文碎片化问题。SeCom [147] 通过提取对话系统中的主题来优化对话结构,而 SGMem [35] 则依赖语义图谱连接碎片化的对话会话。这些研究将记忆系统从简单的线性结构升级为知识图谱,实现了更精确的记忆召回。

除了记忆存储之外,高级基于记忆的对话智能体还需要具备动态推理和时间演化能力。RMM [57] 通过基于反思的自我修正机制扩展了记忆系统,而 TReMu [37] 则利用时间知识图谱来捕捉复杂的基于时间的关系。然而,近期的 ENGRAM [34] 项目对记忆系统复杂性日益增加的趋势提出了挑战,该研究显示,一种具有基本类型记忆结构(包括情景记忆、语义记忆和程序记忆)的高效记忆系统,能够达到与复杂记忆系统相当的性能。

B. 代码智能体

代码生成 [148] 和代码仿真 [149] 的软件工程流程对智能体的内存组件提出了独特挑战,因为它们需要遵循严格的结构要求和软件编程的逻辑流程。在初始阶段,任务混淆问题通过建模常规的人类工作流以及角色与职责的定义,在 MetaGPT [150] 和 ChatDev [151] 中得到了解决。SWE-agent [1] 进一步通过引入智能体-计算机接口,增强了这一方法,使智能体能够对源代码控制系统执行操作。然而,随着任务变得更加复杂,线性方法已不再足够。TALM [152] 提出了一种新范式,摒弃了传统的线性工作流,转向动态的树状架构。通过采用分而治之的方法并充分利用智能体的长期记忆,TALM 展示了层级结构的必要性,这种结构能够处理代码生成中涉及的非线性依赖关系。

除了任务的编排之外,智能体还应导航软件中复杂的资讯空间,这类似于依赖关系与逻辑构成的知识图谱。尽管 Reflexion [153] 通过一种口头反馈环引入了基本的自我修正形式,但近期研究更关注结构化上下文。在这方面,RepoAudit [154] 尝试解决与代码仓库相关的问题。引入了一个审计智能体,能够自主探索代码库。这类似于在文件依赖关系上的图遍历算法,从而确保了分析的准确性。MemGovern [155] 和 Multi-Agent RL Debugging [156] 是进一步的扩展,将从 GitHub 网

站获取的外部信息以及反馈环以可检索的知识库形式进行组织。这清晰地表明了与软件智能体相关的记忆正在演化,正从简单的记忆转向更结构化的记忆,从而支持对拓扑结构的基本推理。

C. 推荐系统

智能体记忆被应用于推荐系统,以解决推荐智能体难以处理长期且动态的用户历史记录的问题 [157]。实际上,推荐智能体常常截断用户历史,导致短期噪声(例如意外的交互行为)覆盖了稳定的长期偏好。同时,通过微调智能体参数来追踪不断变化的用户偏好成本较高 [158]。外部记忆是一种更高效的解决方案,它通过减少重新训练和 token 开销来缓解该问题。

当前基于智能体的推荐系统采用了三层记忆维护方法,从粗粒度检索开始,最终实现细粒度推理。在第一层历史作为记忆阶段,MAP [159] 和 AgentCF++ [160] 均关注可扩展性。对于后者,提出了一种双层方法以实现噪声过滤与社交情境化。在第二层结构化键值记忆阶段,交互被组织为语义结构。MemoCRS [161] 和 Agent4Rec [162] 使用实体-键对来关联评分。最后,在第三层记忆即认知阶段,通过反思与规划使记忆变得主动。在此层次上,CRAVE [163] 和 AgentCF [164] 总结偏好规则。RecMind [165] 采用战略性的多步规划。为了使这些动态系统具体化并扎根于现实,KGLA [166] 引入知识图谱以注入具体的元数据。MR.Rec [167] 则利用强化学习实现自适应记忆检索与逻辑推理。

D. 金融智能体

金融市场因其信息衰减模式、来自不同来源的异构数据流,以及需要在模式与最新市场状况之间取得平衡,被认为对智能体记忆系统具有挑战性 [168]。金融智能体需要具备高优先级的近期性记忆能力以及模式保持能力。此外,金融决策还要求可解释性和风险管理,而记忆系统在反事实推理、最优化和适应性方面至关重要。

在个体认知仿真方面已开展初步尝试,其中 FinMem [44] 通过受人类交易者行为启发的分层记忆结构,解决认知瓶颈问题。为应对个体人工智能智能体固有的偏见,TradingGPT [169] 提出一种扩展方案,引入智能体间辩论机制,以集体智能实现去偏。迈向专业能力阶段,FinCon [170] 为分析师设立了管理层级结构,赋予智能体更强的记忆能力,以维护动作历史、盈

亏序列以及投资信念的变化记录,支持风险管控策略。最后,FinAgent [171] 通过将记忆扩展至多模态感知,弥合信息鸿沟,使智能体能够处理 K 线图和工具增强的数据,实现对市场的全面分析。

当前记忆增强型金融智能体的实现仍相对有限。未来基于图形的记忆系统在金融领域具有变革性潜力,例如,层次化图可通过对跨资产相关系数的建模来提升多资产组合管理能力,时序图则可改善风险管理和尾部风险分析。

E. 游戏智能体

游戏环境因其动态性、复杂规则和长期目标,对智能体的记忆系统构成了重大挑战——尤其是在需要多步推理和探索式学习的开放世界中。记忆系统必须存储经验性知识和世界知识(包括成功与失败),并支持高效实时访问,以促进技能获取。

早期关于游戏智能体的研究聚焦于通过逐步增强的记忆与感知机制来掌握像 Minecraft 这样的开放环境。Ghost in the Minecraft (GITM) [70] 采用基于字典的记忆系统,通过文本交互捕获空间布局和制作知识。Voyager [4] 进一步推进这一范式,引入终身学习,利用不断增长的可执行代码库作为过程性记忆,实现技能的获取与复用。Jarvis-1 [172] 进一步融合了多模态记忆,将文本推理与视觉感知对齐,实现了在 Minecraft 中掌握超过 200 项任务。最近,Optimus-1 [28] 通过将经验组织成分层有向知识图谱,并结合抽象化经验池,解决了长时程规划问题。

最近,范式已转向能够通过统一接口在多种环境中运行的通用智能体。Cradle [61] 通过将交互建立在截图和键盘鼠标操作的基础上,打破了游戏专用 API 的界限,构建了一个类人的统一接口。这使得智能体能够运行多款商业游戏,并需要情景记忆来保留跨交互上下文以及累积的游戏经验。沿着这一方向,SIMA [173] 和 SIMA 2 [174] 聚焦于可指令控制的智能体,能够在多个三维环境中执行任意自然语言命令。最新版本进一步扩展了这一范式,引入了更高级别的推理与自我改进能力,从被动遵循指令转向主动技能获取。这些进展对记忆系统提出了更高的要求,包括长期的情景记忆,以及在不同环境中积累和复用策略的机制。未来的研究应关注具有表现力的记忆架构,例如用于潜在技能层次发现的动态与时间图结构构建,以及具备时间感知能力的因果推理。

F. 机器人学与具身智能体

具身智能体必须持续在动态且部分可观测的物理世界中进行决策,执行跨越多个交互回合的长时程操作任务。因此,嵌入物理或虚拟环境中的智能体面临特殊挑战,需要先进的记忆机制来应对。

HELPER [175] 和 MAP-VLA [176] 均利用记忆来连接高层语言指令与可执行动作,但其方式不同:前者采用键值记忆,将自然语言指令直接映射到机器人代码;后者则使用可复用的记忆库,以检索并适配特定的操作策略。进一步地,STRAP [177] 通过引入一种灵活的记忆机制,利用动态时间规整匹配大规模多样化数据集中长度可变的动作子序列,从而提升轨迹检索效果。相比之下,TrackVLA++ [178] 采用双记忆架构,旨在提升感知稳定性并解决长期记忆维护问题。

综上所述,具身智能体所使用的记忆系统仍保持扁平化或弱结构化,难以实现复杂的关联推理和层级抽象。未来的工作可聚焦于基于图形的记忆架构,以表征物体之间的空间关系、任务的层级结构以及动作与效果之间的因果关系。

G. 医疗健康智能体

医疗智能体需要先进的记忆系统以支持高风险医疗决策、长期患者照护以及基于证据的诊断。记忆系统帮助医疗智能体在多次就诊过程中保留患者的病史,整合医学领域的最新知识,进行鉴别诊断,并提供个性化的患者照护,同时确保安全性和可解释性。具体而言,AgentClinic [179] 和 AgentMental [180] 专注于通过记忆模块模拟医生与患者之间的互动,以记录和追踪随时间变化的诊断信息;而 AgentMental [180] 则采用动态树状结构记忆,用于组织和管理医学知识及对话数据。医疗智能体中的记忆系统反映了真实临床环境中长期病史和诊断准确性的关键作用,使智能体能够保留先前信息,从而提出全面的建议。相比之下,AgentHospital [181] 强调一个全流程的虚拟医院环境,智能体通过在成功与失败案例中的大规模交互不断演化,依托更丰富的数据整合和更全面的系统功能得到支持。

这些研究共同强调了在医疗智能体中采用结构化且具备交互感知能力的记忆的重要性。基于医学本体(如 UMLS [182])的图结构记忆能够实现多跳临床推理、显式建模药物相互作用,以及对疾病进展和治疗过程的时间追踪。这类结构化表示为更稳健的医疗智能体提供

了有前景的基础,有助于提升其上下文推理能力、诊断规划能力以及长期决策支持能力 [183]。

H. 科学智能体

在科学发现中,智能体正从被动的数据分析逐渐转变为能够高效搜索广阔空间并调用特定工具的主动实验伙伴。智能体记忆使得理论与实验在复杂的科研工作流中实现迭代整合,其作用如同一个动态的工作空间,而非静态的信息检索系统。ChatNT [184]、ChemCrow [185] 和 El Agente [186] 代表了特定领域的科学研究支持智能体。它们均致力于提升大语言模型在专业研究任务中的能力,但在研究范围上各有侧重,例如生物序列推理、化学分析以及量子化学仿真。这些系统主要作为智能工具,在科研过程的局部阶段辅助研究人员,而非建模整个研究生命周期。

此外,一些复杂的智能体被构建为具有复杂环境的系统级科学智能体。在生物研究领域,Biomni [187] 超越了任务层面的辅助,支持复杂的生物医学研究工作流,使智能体能够自动识别、组合并执行多步骤的实验流水线。类似地,CRESt [188] 面向材料科学领域,实现了计算推理与物理实验之间的闭环,使智能体能够迭代生成假设并通过机器人合成进行验证。总体而言,VirtualLab [189] 在组织层面运行,通过协作智能体团队模拟人类研究机构,标志着从单个科学工具向集体科学智能的转变,是全面的智能体代表。

总体而言,现有科学智能体的发展历程反映了从局部科学辅助向整体化、自主化科学系统转变的趋势,这需要有效的记忆模块。随着智能体能力的扩展,记忆需求也相应变得更加复杂,要求对异构领域知识进行结构化整合、组织与管理,而基于图形的记忆结构为此提供了清晰且高效解决方案。

X. 局限性和未来方向

尽管基于图形的智能体记忆取得了显著进展,但仍存在若干基本挑战,这些挑战为该领域的进一步发展提供了关键机遇。

记忆图的质量。记忆图的质量从根本上限制了基于图形的智能体记忆系统的性能、可靠性和适应性 [190], [191]。与传统记忆系统不同,后者质量通常主要与下游任务中的事实准确率相关,而基于图形的记忆系统应引入多维度的质量标准,包括结构、语义、时间和操作等方面,每一项都直接塑造智能体的能力 [18], [103]。然

而,目前缺乏专门用于显式评估记忆图内在质量的度量指标 [103], [192]。

可扩展性与效率。随着智能体在长时间交互中积累经验,记忆系统面临计算瓶颈,图操作表现出二次方或更差的复杂度 [15]。未来研究应探索专为图结构设计的记忆压缩技术 [193]、避免完全重新计算的增量更新算法 [194],以及以牺牲部分准确率换取显著效率提升的近似检索方法 [195]。通过专用图处理单元 [196] 和分布式架构 [197] 实现硬件加速,能够管理数百万个结点的同时保持快速访问。

隐私保护与安全。个人助理应用在实现有意义的个性化的同时,需要对敏感信息进行强有力的保护 [9]。基于图形的记忆结构引入了独特的漏洞,其中关系模式可能通过推理攻击无意中暴露私人数据。关键的研究方向包括:(1) 开发针对图记忆系统的差分隐私机制 [198]; (2) 联邦架构,支持设备端处理以最小化数据暴露;以及 (3) 安全多方计算协议,使智能体能够在不损害个体隐私的前提下受益于集体经验。除了隐私泄露之外,记忆系统还面临来自对抗攻击的新兴威胁。类似于针对大模型的提示注入和数据污染攻击 [199], [200],攻击者可以操纵记忆内容以破坏智能体行为或注入恶意知识。内存内容验证、异常检测以及鲁棒审计协议等防御机制对于确保记忆完整性至关重要。

动态模式学习与知识迁移。当前的图模式通常具有领域特定性,可复用性有限,对于新应用需要大量重新工程 [11]。未来的系统应追求动态模式学习,使智能体能够从原始经验中自动识别相关的实体类型和关系模式。元学习方法可实现对新领域的快速适应 [201],结合通用图本体 [202] 和领域无关的抽象机制,以促进任务间有效的知识迁移。

可解释性与可信性。为了使智能体能够在高风险领域中部署,其记忆系统必须既易于人类理解,又在操作上透明。基于图形的记忆架构在可解释性方面具有独特优势:其明确的关系结构自然契合人类的心理模型,使用户能够检查并理解智能体如何组织和利用信息 [79], [82], [191]。关键研究方向包括开发记忆溯源追踪系统,创建交互式可视化界面,使用户能够从多个层次探索记忆图 [145]。通过确保对智能体记忆过程的人类监督与理解,这些系统能够促进适当的信任校准,帮助用户识别潜在偏差、验证关键信息,并保持对智能体行为的控制 [180], [203]。

理论基础。建立严格的数学框架对于推动该领域的

发展仍然至关重要。优先研究方向包括能够提供完备性和一致性保证的形式化模型、确立构建与检索操作理论边界复杂度分析，以及记忆增强型智能体的缩放法则 [204]。与人类认知架构的对比分析有助于识别基本差距，并为与生物记忆系统相一致的架构改进提供机会 [205]。

多智能体系统中的记忆协调。在多智能体或智能体集群的情景中，记忆不再是一个孤立的组件，而是一种共享资源，直接影响任务完成情况和协调效率。无效的记忆共享或不一致的记忆更新可能导致冲突决策。设计记忆同步机制、角色感知的记忆访问以及可扩展的协调方法仍然是一个开放性挑战，尤其是在通信约束条件下。

XI. 结论

随着基于大语言模型的智能体朝着日益自主和通用的方向演进，记忆已成为关键组成部分。基于图形的记忆架构标志着从简单存储机制向结构化、关系型表示的范式转变，能够实现复杂的推理、个性化以及持续学习。本文综述从基于图形的角度全面回顾了智能体记忆。首先，提出一个智能体记忆分类体系，包括短期与长期记忆、知识与经验记忆、非结构化与结构化记忆，并重点关注基于图形的记忆实现。其次，按生命周期系统分析关键的基于图形的智能体记忆技术，涵盖提取、存储、检索与演化。第三，总结开源库、数据集与基准测试，以及自演化智能体记忆的多样化应用场景。最后，指出当前面临的挑战与未来研究方向。我们希望本综述能为推动智能体记忆系统前沿研究的科研人员，以及致力于构建更强大、可靠、可信人工智能智能体的实践者提供有价值的参考。

参考文献

- [1] J. Yang, C. E. Jimenez, A. Wettig, K. Lieret, S. Yao, K. Narasimhan, and O. Press, “SWE-agent: agent-computer interfaces enable automated software engineering,” in *NeurIPS*, 2024.
- [2] Y. Lin, S. Tang, B. Lyu, J. Wu, H. Lin, K. Yang, J. Li, M. Xia, D. Chen, S. Arora *et al.*, “Goedel-prover: A frontier model for open-source automated theorem proving,” *arXiv preprint arXiv:2502.07640*, 2025.
- [3] Meta, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu *et al.*, “Human-level play in the game of diplomacy by combining language models with strategic reasoning,” *Science*, 2022.
- [4] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, “Voyager: An open-ended embodied agent with large language models,” *Transactions on Machine Learning Research*, 2024.
- [5] Y. Su, D. Yang, S. Yao, and T. Yu, “Language agents: Foundations, prospects, and risks,” in *EMNLP: Tutorial Abstracts*, 2024.
- [6] Z. Wang, Z. Cheng, H. Zhu, D. Fried, and G. Neubig, “What are tools anyway? a survey from the language model perspective,” in *COLM*, 2024.
- [7] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, dahai li, Z. Liu, and M. Sun, “ToolLLM: Facilitating large language models to master 16000+ real-world APIs,” in *ICLR*, 2024.
- [8] T. Sumers, S. Yao, K. R. Narasimhan, and T. L. Griffiths, “Cognitive architectures for language agents,” *Transactions on Machine Learning Research*, 2024.
- [9] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun *et al.*, “Personal LLM agents: Insights and survey about the capability, efficiency and security,” *arXiv preprint arXiv:2401.05459*, 2024.
- [10] C. Yang, X. Wang, Q. Zhang, Q. Jiang, and X. Huang, “Efficient integration of external knowledge to llm-based world models via retrieval-augmented generation and reinforcement learning,” in *Findings, EMNLP*, 2025, pp. 9484–9501.
- [11] Y. Shang, Y. Li, K. Zhao, L. Ma, J. Liu, F. Xu, and Y. Li, “Agentsquare: Automatic LLM agent search in modular design space,” in *ICLR*, 2025.
- [12] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, “Cognitive mirage: A review of hallucinations in large language models,” *arXiv preprint arXiv:2309.06794*, 2023.
- [13] H. Yu, T. Chen, J. Feng, J. Chen, W. Dai, Q. Yu, Y.-Q. Zhang, W.-Y. Ma, J. Liu, M. Wang *et al.*, “Memagent: Reshaping long-context llm with multi-conv rl-based memory agent,” *arXiv preprint arXiv:2507.02259*, 2025.
- [14] J. Nan, W. Ma, W. Wu, and Y. Chen, “Nemori: Self-organizing agent memory inspired by cognitive science,” *arXiv preprint arXiv:2508.03341*, 2025.
- [15] R. Zeng, J. Fang, S. Liu, and Z. Meng, “On the structural memory of llm agents,” *arXiv preprint arXiv:2412.15266*, 2024.
- [16] Z. Jia, J. Li, Y. Kang, Y. Wang, T. Wu, Q. Wang, X. Wang, S. Zhang, J. Shen, Q. Li, S. Qi, Y. Liang, D. He, Z. Zheng, and S.-C. Zhu, “The AI hippocampus: How far are we from human memory?” *Transactions on Machine Learning Research*, 2025.
- [17] H. Sun and S. Zeng, “Hierarchical memory for high-efficiency long-term reasoning in llm agents,” *arXiv preprint arXiv:2507.22925*, 2025.
- [18] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef, “Zep: a temporal knowledge graph architecture for agent memory,” *arXiv preprint arXiv:2501.13956*, 2025.
- [19] R. B. Yousuf, A. Khatri, S. Xu, M. Sharma, and N. Ramakrishnan, “Can an llm induce a graph? investigating memory drift and context length,” *arXiv preprint arXiv:2510.03611*, 2025.
- [20] R. Zeng, J. Fang, S. Liu, and Z. Meng, “On the structural memory of llm agents,” *arXiv preprint arXiv:2412.15266*, 2024.
- [21] S. Liang, Y. Zhang, and Y. Guo, “Personaagent with graphrag: Community-aware knowledge graphs for personalized llm,” *arXiv preprint arXiv:2511.17467*, 2025.

- [22] C. Huan, Z. Meng, Y. Liu, Z. Yang, Y. Zhu, Y. Yun, S. Li, R. Gu, X. Wu, H. Zhang *et al.*, “Scaling graph chain-of-thought reasoning: A multi-agent framework with efficient llm serving,” *arXiv preprint arXiv:2511.01633*, 2025.
- [23] M. Hu, T. Chen, Q. Chen, Y. Mu, W. Shao, and P. Luo, “Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.
- [24] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: language agents with verbal reinforcement learning,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.
- [25] Y. Wang, R. Takanobu, Z. Liang, Y. Mao, Y. Hu, J. McAuley, and X. Wu, “Mem- $\{\alpha\}$: Learning memory construction via reinforcement learning,” *arXiv preprint arXiv:2509.25911*, 2025.
- [26] H. Shi, B. Xie, Y. Liu, L. Sun, F. Liu, T. Wang, E. Zhou, H. Fan, X. Zhang, and G. Huang, “Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation,” *arXiv preprint arXiv:2508.19236*, 2025.
- [27] J. H. Yeo, M. Kim, and Y. M. Ro, “Multi-temporal lip-audio memory for visual speech recognition,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [28] Z. Li, Y. Xie, R. Shao, G. Chen, D. Jiang, and L. Nie, “Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks,” *Advances in neural information processing systems*, 2024.
- [29] A. Modarressi, A. Köksal, A. Imani, M. Fayyaz, and H. Schütze, “Memllm: Finetuning llms to use an explicit read-write memory,” *arXiv preprint arXiv:2404.11672*, 2024.
- [30] P. Anokhin, N. Semenov, A. Sorokin, D. Evseev, A. Kravchenko, M. Burtsev, and E. Burnaev, “Arigraph: Learning knowledge graph world models with episodic memory for llm agents,” *arXiv preprint arXiv:2407.04363*, 2024.
- [31] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav, “Mem0: Building production-ready ai agents with scalable long-term memory,” *arXiv preprint arXiv:2504.19413*, 2025.
- [32] H. Yang, J. Chen, M. Siew, T. Llorido-Botran, and C. Joe-Wong, “Llm-powered decentralized generative agents with adaptive hierarchical knowledge graph for cooperative planning,” *arXiv preprint arXiv:2502.05453*, 2025.
- [33] G. Zhang, M. Fu, G. Wan, M. Yu, K. Wang, and S. Yan, “G-memory: Tracing hierarchical memory for multi-agent systems,” *arXiv preprint arXiv:2506.07398*, 2025.
- [34] D. Patel and S. Patel, “Engram: Effective, lightweight memory orchestration for conversational agents,” *arXiv preprint arXiv:2511.12960*, 2025.
- [35] Y. Wu, Y. Zhang, S. Liang, and Y. Liu, “Sgmemo: Sentence graph memory for long-term conversational agents,” *arXiv preprint arXiv:2509.21212*, 2025.
- [36] Z. Wang, Z. Li, Z. Jiang, D. Tu, and W. Shi, “Crafting personalized agents through retrieval-augmented generation on editable memory graphs,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.
- [37] Y. Ge, S. Romeo, J. Cai, R. Shu, Y. Benajiba, M. Sunkara, and Y. Zhang, “Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues,” in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.
- [38] X. Tan, X. Wang, X. Xu, X. Yuan, L. Zhu, and W. Zhang, “Memotime: Memory-augmented temporal knowledge graph enhanced large language model reasoning,” *arXiv preprint arXiv:2510.13614*, 2025.
- [39] H. Luo, G. Chen, Y. Zheng, X. Wu, Y. Guo, Q. Lin, Y. Feng, Z. Kuang, M. Song, Y. Zhu *et al.*, “Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation,” *arXiv preprint arXiv:2503.21322*, 2025.
- [40] S. Huang, H. Li, Y. Gu, X. Hu, Q. Li, and G. Xu, “Hyperg: Hypergraph-enhanced llms for structured knowledge,” in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.
- [41] J. Jiang, K. Zhou, W. X. Zhao, Y. Song, C. Zhu, H. Zhu, and J.-R. Wen, “Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.
- [42] L. Long, Y. He, W. Ye, Y. Pan, Y. Lin, H. Li, J. Zhao, and W. Li, “Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory,” *arXiv preprint arXiv:2508.09736*, 2025.
- [43] R. Salama, J. Cai, M. Yuan, A. Currey, M. Sunkara, Y. Zhang, and Y. Benajiba, “Meminsight: Autonomous memory augmentation for llm agents,” *arXiv preprint arXiv:2503.21760*, 2025.
- [44] Y. Yu, H. Li, Z. Chen, Y. Jiang, Y. Li, J. W. Suchow, D. Zhang, and K. Khashanah, “Finmem: A performance-enhanced llm trading agent with layered memory and character design,” *IEEE Transactions on Big Data*, 2025.
- [45] Y. Hou, H. Tamoto, and H. Miyashita, “my agent understands me better”: Integrating dynamic human-like memory recall and consolidation in llm-based agents,” in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024.
- [46] M. Fisher, “Neural graph memory: A structured approach to long-term memory in multimodal agents,” 2025.
- [47] Y. Cai, Z. Guo, Y. Pei, W. Bian, and W. Zheng, “Simrag: Leveraging similar subgraphs for knowledge graphs driven retrieval-augmented generation,” in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.
- [48] X. Wu, Y. Shen, C. Shan, K. Song, S. Wang, B. Zhang, J. Feng, H. Cheng, W. Chen, Y. Xiong *et al.*, “Can graph learning improve planning in llm-based agents?” *Advances in Neural Information Processing Systems*, 2024.
- [49] H. Sun and S. Zeng, “Hierarchical memory for high-efficiency long-term reasoning in llm agents,” *arXiv preprint arXiv:2507.22925*, 2025.
- [50] A. Jonelagadda, C. Hahn, H. Zheng, and S. Penachio, “Mnemosyne: An unsupervised, human-inspired long-term memory architecture for edge-based llms,” *arXiv preprint arXiv:2510.08601*, 2025.
- [51] K. Zhang, X. Zhang, E. Ahmed, H. Jiang, C. Kumar, K. Sun, Z. Lin, S. Sharma, S. Oraby, A. Colak *et al.*, “Assomem: Scalable memory qa with multi-signal associative retrieval,” *arXiv preprint arXiv:2510.10397*, 2025.

- [52] Z. Huang, Z. Tian, Q. Guo, F. Zhang, Y. Zhou, D. Jiang, and X. Zhou, "Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning," *arXiv preprint arXiv:2511.01448*, 2025.
- [53] H. Zhou, Y. Chen, S. Guo, X. Yan, K. H. Lee, Z. Wang, K. Y. Lee, G. Zhang, K. Shao, L. Yang *et al.*, "Memento: Fine-tuning llm agents without fine-tuning llms," *arXiv preprint arXiv:2508.16153*, 2025.
- [54] S. Yan, X. Yang, Z. Huang, E. Nie, Z. Ding, Z. Li, X. Ma, K. Kersting, J. Z. Pan, H. Schütze *et al.*, "Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning," *arXiv preprint arXiv:2508.19828*, 2025.
- [55] S. Xia, Z. Xu, J. Chai, W. Fan, Y. Song, X. Wang, G. Yin, W. Lin, H. Zhang, and J. Wang, "From experience to strategy: Empowering llm agents with trainable graph memory," *arXiv preprint arXiv:2511.07800*, 2025.
- [56] M. Xu, G. Liang, K. Chen, W. Wang, X. Zhou, M. Yang, T. Zhao, and M. Zhang, "Memory-augmented query reconstruction for llm-based knowledge graph reasoning," *arXiv preprint arXiv:2503.05193*, 2025.
- [57] Z. Tan, J. Yan, I.-H. Hsu, R. Han, Z. Wang, L. Le, Y. Song, Y. Chen, H. Palangi, G. Lee *et al.*, "In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.
- [58] T. Kim, V. François-Lavet, and M. Cochez, "Leveraging knowledge graph-based human-like memory systems to solve partially observable markov decision processes," *arXiv preprint arXiv:2408.05861*, 2024.
- [59] R. Wang, S. Liang, Q. Chen, Y. Huang, M. Li, Y. Ma, D. Zhang, K. Qin, and M.-F. Leung, "Graphcogent: Mitigating llms' working memory constraints via multi-agent collaboration in complex graph understanding," *arXiv preprint arXiv:2508.12379*, 2025.
- [60] P. Wang, M. Tian, J. Li, Y. Liang, Y. Wang, Q. Chen, T. Wang, Z. Lu, J. Ma, Y. E. Jiang *et al.*, "Omni memory system for personalized, long horizon, self-evolving agents," *arXiv preprint arXiv:2511.13593*, 2025.
- [61] W. Tan, W. Zhang, X. Xu, H. Xia, Z. Ding, B. Li, B. Zhou, J. Yue, J. Jiang, Y. Li, R. An, M. Qin, C. Zong, L. Zheng, Y. Wu, X. Chai, Y. Bi, T. Xie, P. Gu, X. Li, C. Zhang, L. Tian, C. Wang, X. Wang, B. F. Karlsson, B. An, S. YAN, and Z. Lu, "Cradle: Empowering foundation agents towards general computer control," in *ICML*, 2025.
- [62] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "Memgpt: Towards llms as operating systems," *arXiv preprint arXiv:2310.08560*, 2023.
- [63] M. Hu, T. Chen, Q. Chen, Y. Mu, W. Shao, and P. Luo, "Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.
- [64] Z. Jia, J. Li, X. Qu, and J. Wang, "Enhancing multi-agent systems via reinforcement learning with llm-based planner and graph-based policy," *arXiv preprint arXiv:2503.10049*, 2025.
- [65] A. Rezazadeh, Z. Li, A. Lou, Y. Zhao, W. Wei, and Y. Bao, "Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control," *arXiv preprint arXiv:2505.18279*, 2025.
- [66] B. Yan, C. Li, H. Qian, S. Lu, and Z. Liu, "General agentic memory via deep research," *arXiv preprint arXiv:2511.18423*, 2025.
- [67] Y. Zhang, W. Yuan, and Z. Jiang, "Bridging intuitive associations and deliberate recall: Empowering llm personal assistant with graph-structured long-term memory," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.
- [68] J. Liu, Z. Kong, C. Yang, F. Yang, T. Li, P. Dong, J. Nanjeyye, H. Tang, G. Yuan, W. Niu *et al.*, "Rcr-router: Efficient role-aware context routing for multi-agent llm systems with structured memory," *arXiv preprint arXiv:2508.04903*, 2025.
- [69] Q. Yuan, J. Lou, Z. Li, J. Chen, Y. Lu, H. Lin, L. Sun, D. Zhang, and X. Han, "Memsearcher: Training llms to reason, search and manage memory via end-to-end reinforcement learning," *arXiv preprint arXiv:2511.02805*, 2025.
- [70] X. Zhu, Y. Chen, H. Tian, C. Tao, W. Su, C. Yang, G. Huang, B. Li, L. Lu, X. Wang *et al.*, "Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory," *arXiv preprint arXiv:2305.17144*, 2023.
- [71] D. Han, C. Couturier, D. M. Diaz, X. Zhang, V. Rühle, and S. Rajmohan, "Legomem: Modular procedural memory for multi-agent llm systems for workflow automation," *arXiv preprint arXiv:2510.04851*, 2025.
- [72] Y. Wang and X. Chen, "Mirix: Multi-agent memory system for llm-based agents," *arXiv preprint arXiv:2507.07957*, 2025.
- [73] G. Zhang, M. Fu, and S. Yan, "Memgen: Weaving generative latent memory for self-evolving agents," *arXiv preprint arXiv:2509.24704*, 2025.
- [74] Z. Zhou, A. Qu, Z. Wu, S. Kim, A. Prakash, D. Rus, J. Zhao, B. K. H. Low, and P. P. Liang, "Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents," *arXiv preprint arXiv:2506.15841*, 2025.
- [75] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.
- [76] B. Kynoch, H. Latapie, and D. van der Sluis, "Recallm: An adaptable memory mechanism with temporal understanding for large language models," *arXiv preprint arXiv:2307.02738*, 2023.
- [77] X. Tang, T. Qin, T. Peng, Z. Zhou, D. Shao, T. Du, X. Wei, P. Xia, F. Wu, H. Zhu *et al.*, "Agent kb: Leveraging cross-domain experience for agentic problem solving," *arXiv preprint arXiv:2507.06229*, 2025.
- [78] Z. Cai, X. Guo, Y. Pei, J. Feng, J. Su, J. Chen, Y.-Q. Zhang, W.-Y. Ma, M. Wang, and H. Zhou, "Flex: Continuous agent evolution via forward learning from experience," *arXiv preprint arXiv:2511.06449*, 2025.
- [79] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. Ni, H.-Y. Shum, and J. Guo, "Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph," in *The Twelfth International Conference on Learning Representations*, 2024.

- [80] Y. Xiao, C. Zhou, Q. Zhang, B. Li, Q. Li, and X. Huang, "Reliable reasoning path: Distilling effective guidance for llm reasoning with knowledge graphs," *IEEE Transactions on Knowledge and Data Engineering*, 2026.
- [81] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "Memorybank: Enhancing large language models with long-term memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [82] L. LUO, Y.-F. Li, G. Haffari, and S. Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," in *The Twelfth International Conference on Learning Representations*, 2024.
- [83] J. Kang, M. Ji, Z. Zhao, and T. Bai, "Memory os of ai agent," *arXiv preprint arXiv:2506.06326*, 2025.
- [84] Z. Xu, R. Zhou, Y. Yin, H. Gao, M. Tomizuka, and J. Li, "Matrix: multi-agent trajectory generation with diverse contexts," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- [85] Z. Gekhman, E. B. David, H. Orgad, E. Ofek, Y. Belinkov, I. Szpektor, J. Herzig, and R. Reichart, "Inside-out: Hidden factual knowledge in llms," *arXiv preprint arXiv:2503.15299*, 2025.
- [86] G. Zhang, H. Ren, C. Zhan, Z. Zhou, J. Wang, H. Zhu, W. Zhou, and S. Yan, "Memevolve: Meta-evolution of agent memory systems," *arXiv preprint arXiv:2512.18746*, 2025.
- [87] Moonshot AI, "Kimi k2.5," 2026. [Online]. Available: <https://www.kimi.com/blog/kimi-k2-5.html>
- [88] A. Zhao, D. Huang, Q. Xu, M. Lin, Y.-J. Liu, and G. Huang, "Expel: Llm agents are experiential learners," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [89] C. Yang, Z. Sun, W. Wei, and W. Hu, "Beyond static summarization: Proactive memory extraction for llm agents," *arXiv preprint arXiv:2601.04463*, 2026.
- [90] Y. Zhai, S. Tao, C. Chen, A. Zou, Z. Chen, Q. Fu, S. Mai, L. Yu, J. Deng, Z. Cao *et al.*, "Agentevolver: Towards efficient self-evolving agent system," *arXiv preprint arXiv:2511.10395*, 2025.
- [91] B. J. Gutiérrez, Y. Shu, W. Qi, S. Zhou, and Y. Su, "From RAG to memory: Non-parametric continual learning for large language models," in *Forty-second International Conference on Machine Learning*, 2025.
- [92] V. Markovic, L. Obradovic, L. Hajdu, and J. Pavlovic, "Optimizing the interface between knowledge graphs and llms for complex reasoning," *arXiv preprint arXiv:2505.24478*, 2025.
- [93] J. Fang, X. Deng, H. Xu, Z. Jiang, Y. Tang, Z. Xu, S. Deng, Y. Yao, M. Wang, S. Qiao *et al.*, "Lightmem: Lightweight and efficient memory-augmented generation," *arXiv preprint arXiv:2510.18866*, 2025.
- [94] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating very long-term conversational memory of LLM agents," in *ACL*, 2024.
- [95] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu, "LongMemEval: Benchmarking chat assistants on long-term interactive memory," in *ICLR*, 2025.
- [96] Y. Hu, Y. Wang, and J. McAuley, "Evaluating memory in llm agents via incremental multi-turn interactions," *arXiv preprint arXiv:2507.05257*, 2025.
- [97] D. Deshpande, V. Gangal, H. Mehta, A. Kannappan, R. Qian, and P. Wang, "Memtrack: Evaluating long-term memory and state tracking in multi-platform dynamic agent environments," *arXiv preprint arXiv:2510.01353*, 2025.
- [98] J. He, L. Zhu, R. Wang, X. Wang, G. Haffari, and J. Zhang, "Madial-bench: Towards real-world evaluation of memory-augmented dialogue generation," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.
- [99] Z. Zhang, Q. Dai, L. Chen, Z. Jiang, R. Li, J. Zhu, X. Chen, Y. Xie, Z. Dong, and J.-R. Wen, "Memsim: A bayesian simulator for evaluating memory of llm-based personal assistants," *arXiv preprint arXiv:2409.20163*, 2024.
- [100] B. Wu, W. Wang, L. Lihaoran, Y. Deng, Y. Li, J. Yu, and B. Wang, "Interpersonal memory matters: A new task for proactive dialogue utilizing conversational history," in *Proceedings of the 29th Conference on Computational Natural Language Learning*, 2025.
- [101] J. Xu, A. Szlam, and J. Weston, "Beyond goldfish memory: Long-term open-domain conversation," in *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 2022.
- [102] H. Xue, F. Tang, M. Hu, Y. Liu, Q. Huang, Y. Li, C. Liu, Z. Xu, C. Zhang, C.-M. Feng *et al.*, "Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation," *arXiv preprint arXiv:2502.11903*, 2025.
- [103] H. Tan, Z. Zhang, C. Ma, X. Chen, Q. Dai, and Z. Dong, "Mem-bench: Towards more comprehensive evaluation on the memory of llm-based agents," *arXiv preprint arXiv:2506.21605*, 2025.
- [104] L. Wan and W. Ma, "Storybench: A dynamic benchmark for evaluating long-term memory with multi turns," *arXiv preprint arXiv:2506.13356*, 2025.
- [105] J. Kim, W. Chay, H. Hwang, D. Kyung, H. Chung, E. Cho, Y. Jo, and E. Choi, "Dialsim: A real-time simulator for evaluating long-term multi-party dialogue understanding of conversation systems," *arXiv preprint arXiv:2406.13144*, 2024.
- [106] H. Bian, Z. Yao, S. Hu, Z. Xu, S. Zhang, Y. Guo, Z. Yang, X. Han, H. Wang, and R. Chen, "Realmem: Benchmarking llms in real-world memory-driven interaction," *arXiv preprint arXiv:2601.06966*, 2026.
- [107] B. Jiang, Z. Hao, Y.-M. Cho, B. Li, Y. Yuan, S. Chen, L. Ungar, C. J. Taylor, and D. Roth, "Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale," *arXiv preprint arXiv:2504.14225*, 2025.
- [108] Y. Du, H. Wang, Z. Zhao, B. Liang, B. Wang, W. Zhong, Z. Wang, and K.-F. Wong, "Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering," *arXiv preprint arXiv:2402.16288*, 2024.
- [109] Z. Zhang, Y. Zhang, H. Tan, R. Li, and X. Chen, "Explicit vs implicit memory: Exploring multi-hop complex reasoning over personalized information," *arXiv preprint arXiv:2508.13250*, 2025.
- [110] S. Zhao, M. Hong, Y. Liu, D. Hazarika, and K. Lin, "Do LLMs recognize your preferences? evaluating personalized preference following in LLMs," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [111] Z. Jia, Q. Liu, H. Li, Y. Chen, and J. Liu, "Evaluating the long-term memory of large language models," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

- [112] A. Miyai, Z. Zhao, K. Egashira, A. Sato, T. Sunada, S. Onohara, H. Yamanishi, M. Toyooka, K. Nishina, R. Maeda *et al.*, “Webchorearena: Evaluating web browsing agents on realistic tedious web tasks,” *arXiv preprint arXiv:2506.01952*, 2025.
- [113] Y. Deng, X. Zhang, W. Zhang, Y. Yuan, S. K. Ng, and T.-S. Chua, “On the multi-turn instruction following for conversational web agents,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.
- [114] S. Yao, H. Chen, J. Yang, and K. Narasimhan, “Webshop: Towards scalable real-world web interaction with grounded language agents,” *Advances in Neural Information Processing Systems*, 2022.
- [115] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried *et al.*, “Webarena: A realistic web environment for building autonomous agents,” in *12th International Conference on Learning Representations, ICLR 2024*, 2024.
- [116] S. Tian, Z. Zhang, L.-Y. Chen, and Z. Liu, “Mmina: Benchmarking multihop multimodal internet agents,” in *Findings of the Association for Computational Linguistics*, 2025.
- [117] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee *et al.*, “Natural questions: a benchmark for question answering research,” *Transactions of the Association for Computational Linguistics*, 2019.
- [118] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, “Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017.
- [119] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, “When not to trust language models: Investigating effectiveness of parametric and non-parametric memories,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.
- [120] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018.
- [121] X. Ho, A.-K. D. Nguyen, S. Sugawara, and A. Aizawa, “Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.
- [122] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, “Musique: Multihop questions via single-hop question composition,” *Transactions of the Association for Computational Linguistics*, 2022.
- [123] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou *et al.*, “Longbench: A bilingual, multitask benchmark for long context understanding,” in *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2024.
- [124] Y. Bai, S. Tu, J. Zhang, H. Peng, X. Wang, X. Lv, S. Cao, J. Xu, L. Hou, Y. Dong *et al.*, “Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.
- [125] C.-P. Hsieh, S. Sun, S. Krizan, S. Acharya, D. Rekes, F. Jia, and B. Ginsburg, “RULER: What’s the real context size of your long-context language models?” in *First Conference on Language Modeling*, 2024.
- [126] Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, and M. Burtsev, “Babilong: Testing the limits of llms with long context reasoning-in-a-haystack,” *Advances in Neural Information Processing Systems*, 2024.
- [127] H. Wang, H. Shi, S. Tan, W. Qin, W. Wang, T. Zhang, A. Nambi, T. Ganu, and H. Wang, “Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models,” in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.
- [128] D. Chen, S. Niu, K. Li, P. Liu, X. Zheng, B. Tang, X. Li, F. Xiong, and Z. Li, “Halumem: Evaluating hallucinations in memory systems of agents,” *arXiv preprint arXiv:2511.03506*, 2025.
- [129] Q. Ai, Y. Tang, C. Wang, J. Long, W. Su, and Y. Liu, “Memorybench: A benchmark for memory and continual learning in llm systems,” *arXiv preprint arXiv:2510.17281*, 2025.
- [130] J. Zheng, X. Cai, Q. Li, D. Zhang, Z. Li, Y. Zhang, L. Song, and Q. Ma, “Lifelongagentbench: Evaluating llm agents as lifelong learners,” *arXiv preprint arXiv:2505.11942*, 2025.
- [131] C.-K. Wu, Z. R. Tam, C.-Y. Lin, Y.-N. Chen, and H.-y. Lee, “Streambench: Towards benchmarking continuous improvement of language agents,” *Advances in Neural Information Processing Systems*, 2024.
- [132] T. Wei, N. Sachdeva, B. Coleman, Z. He, Y. Bei, X. Ning, M. Ai, Y. Li, J. He, E. H. Chi *et al.*, “Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory,” *arXiv preprint arXiv:2511.20857*, 2025.
- [133] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, “Ego4d: Around the world in 3,000 hours of egocentric video,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [134] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang *et al.*, “Egolife: Towards egocentric life assistant,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.
- [135] M. Shridhar, X. Yuan, M.-A. Cote, Y. Bisk, A. Trischler, and M. Hausknecht, “{ALFW}orld: Aligning text and embodied environments for interactive learning,” in *International Conference on Learning Representations*, 2021.
- [136] M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, and Y. Bengio, “BabyAI: First steps towards grounded language learning with a human in the loop,” in *International Conference on Learning Representations*, 2019.
- [137] R. Wang, P. Jansen, M.-A. Côté, and P. Ammanabrolu, “Science-world: Is your agent smarter than a 5th grader?” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.
- [138] Z. Xi, Y. Ding, W. Chen, B. Hong, H. Guo, J. Wang, X. Guo, D. Yang, C. Liao, W. He *et al.*, “Agentgym: Evaluating and training large language model-based agents across diverse environments,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

- [139] M. Chang, J. Zhang, Z. Zhu, C. Yang, Y. Yang, Y. Jin, Z. Lan, L. Kong, and J. He, “Agentboard: An analytical evaluation board of multi-turn llm agents,” *Advances in neural information processing systems*, 2024.
- [140] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. R. Narasimhan, “SWE-bench: Can language models resolve real-world github issues?” in *The Twelfth International Conference on Learning Representations*, 2024.
- [141] G. Mialon, C. Fourrier, T. Wolf, Y. LeCun, and T. Scialom, “Gaia: a benchmark for general ai assistants,” in *The Twelfth International Conference on Learning Representations*, 2023.
- [142] K. Chen, Y. Ren, Y. Liu, X. Hu, H. Tian, T. Xie, F. Liu, H. Zhang, H. Liu, Y. Gong *et al.*, “xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations,” *arXiv preprint arXiv:2506.13651*, 2025.
- [143] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, dahai li, Z. Liu, and M. Sun, “ToolLLM: Facilitating large language models to master 16000+ real-world APIs,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [144] B. Li, Z. Lin, D. Pathak, J. Li, Y. Fei, K. Wu, T. Ling, X. Xia, P. Zhang, G. Neubig *et al.*, “Genai-bench: Evaluating and improving compositional text-to-visual generation,” *arXiv preprint arXiv:2406.13743*, 2024.
- [145] Z. Huang, S. Gutierrez, H. Kamana, and S. MacNeil, “Memory sandbox: Transparent and interactive memory management for conversational agents,” in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.
- [146] H. Li, C. Yang, A. Zhang, Y. Deng, X. Wang, and T.-S. Chua, “Hello again! llm-powered personalized agent for long-term dialogue,” in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.
- [147] Z. Pan, Q. Wu, H. Jiang, X. Luo, H. Cheng, D. Li, Y. Yang, C.-Y. Lin, H. V. Zhao, L. Qiu, and J. Gao, “Secom: On memory construction and retrieval for personalized conversational agents,” in *ICLR*, 2025.
- [148] Y. Dong, X. Jiang, J. Qian, T. Wang, K. Zhang, Z. Jin, and G. Li, “A survey on code generation with llm-based agents,” *arXiv preprint arXiv:2508.00083*, 2025.
- [149] M. A. Islam, M. E. Ali, and M. R. Parvez, “Codesim: Multi-agent code generation and problem solving through simulation-driven planning and debugging,” *arXiv preprint arXiv:2502.05664*, 2025.
- [150] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin *et al.*, “Metagpt: Meta programming for a multi-agent collaborative framework,” in *The Twelfth International Conference on Learning Representations*, 2023.
- [151] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong *et al.*, “Chatdev: Communicative agents for software development,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.
- [152] M.-T. Shen and Y.-J. Joung, “Talm: Dynamic tree-structured multi-agent framework with long-term memory for scalable code generation,” *arXiv preprint arXiv:2510.23010*, 2025.
- [153] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” *Advances in Neural Information Processing Systems*, 2023.
- [154] J. Guo, C. Wang, X. Xu, Z. Su, and X. Zhang, “Repoaudit: An autonomous llm-agent for repository-level code auditing,” *arXiv preprint arXiv:2501.18160*, 2025.
- [155] Q. Wang, Z. Cheng, S. Zhang, F. Liu, R. Xu, H. Lian, K. Wang, X. Yu, J. Yin, S. Hu *et al.*, “Memgovern: Enhancing code agents through learning from governed human experiences,” *arXiv preprint arXiv:2601.06789*, 2026.
- [156] A. Krishnamoorthy, K. Ivatury, and B. Ahmadnia, “Multi-agent reinforcement learning for interactive code debugging with human feedback and memory,” in *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, 2025.
- [157] S. Cai, J. Zhang, K. Bao, C. Gao, Q. Wang, F. Feng, and X. He, “Agentic feedback loop modeling improves recommendation and user simulation,” in *Proceedings of the 48th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2025.
- [158] F. Liu, X. Lin, H. Yu, M. Wu, J. Wang, Q. Zhang, Z. Zhao, Y. Xia, Y. Zhang, W. Li *et al.*, “Recoworld: Building simulated environments for agentic recommender systems,” *arXiv preprint arXiv:2509.10397*, 2025.
- [159] J. Chen, “Memory assisted llm for personalized recommendation system,” *arXiv preprint arXiv:2505.03824*, 2025.
- [160] J. Liu, S. Gu, D. Li, G. Zhang, M. Han, H. Gu, P. Zhang, T. Lu, L. Shang, and N. Gu, “Agentcf++: Memory-enhanced llm-based agents for popularity-aware cross-domain recommendations,” in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.
- [161] Y. Xi, W. Liu, J. Lin, B. Chen, R. Tang, W. Zhang, and Y. Yu, “Memocrs: Memory-enhanced sequential conversational recommender systems with large language models,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024.
- [162] A. Zhang, Y. Chen, L. Sheng, X. Wang, and T.-S. Chua, “On generative agents in recommendation,” in *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, 2024.
- [163] Y. Zhu, H. Steck, D. Liang, Y. He, N. Kallus, and J. Li, “Llm-based conversational recommendation agents with collaborative verbalized experience,” *Proceedings of the Proc. of EMNLP Findings*, 2025.
- [164] J. Zhang, Y. Hou, R. Xie, W. Sun, J. McAuley, W. X. Zhao, L. Lin, and J.-R. Wen, “Agentcf: Collaborative learning with autonomous language agents for recommender systems,” in *Proceedings of the ACM Web Conference 2024*, 2024.
- [165] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, Y. Lu, X. Huang, and Y. Yang, “Recmind: Large language model powered agent for recommendation,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024.
- [166] T. Guo, C. Liu, H. Wang, V. Mannam, F. Wang, X. Chen, X. Zhang, and C. K. Reddy, “Knowledge graph enhanced language agents for recommendation,” *arXiv preprint arXiv:2410.19627*, 2024.

- [167] J. Huang, X. Zou, L. Xia, and Q. Li, “Mr. rec: Synergizing memory and reasoning for personalized recommendation assistant with llms,” *arXiv preprint arXiv:2510.14629*, 2025.
- [168] H. Li, Y. Cao, Y. Yu, S. R. Javaji, Z. Deng, Y. He, Y. Jiang, Z. Zhu, K. Subbalakshmi, J. Huang *et al.*, “Investorbench: A benchmark for financial decision-making tasks with llm-based agent,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.
- [169] Y. Li, Y. Yu, H. Li, Z. Chen, and K. Khashanah, “Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance,” *arXiv preprint arXiv:2309.03736*, 2023.
- [170] Y. Yu, Z. Yao, H. Li, Z. Deng, Y. Jiang, Y. Cao, Z. Chen, J. Suchow, Z. Cui, R. Liu *et al.*, “Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making,” *Advances in Neural Information Processing Systems*, 2024.
- [171] W. Zhang, L. Zhao, H. Xia, S. Sun, J. Sun, M. Qin, X. Li, Y. Zhao, Y. Zhao, X. Cai *et al.*, “A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist,” in *Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining*, 2024.
- [172] Z. Wang, S. Cai, A. Liu, Y. Jin, J. Hou, B. Zhang, H. Lin, Z. He, Z. Zheng, Y. Yang *et al.*, “Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [173] M. A. Raad, A. Ahuja, C. Barros, F. Besse, A. Bolt, A. Bolton, B. Brownfield, G. Buttimore, M. Cant, S. Chakera *et al.*, “Scaling instructable agents across many simulated worlds,” *arXiv preprint arXiv:2404.10179*, 2024.
- [174] A. Bolton, A. Lerchner, A. Cordell, A. Moufarek, A. Bolt, A. Lampinen, A. Mitenkova, A. O. Hallingstad, B. Vujatovic, B. Li *et al.*, “Sima 2: A generalist embodied agent for virtual worlds,” *arXiv preprint arXiv:2512.04797*, 2025.
- [175] G. Sarch, Y. Wu, M. Tarr, and K. Fragkiadaki, “Open-ended instructable embodied agents with memory-augmented large language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.
- [176] R. Li, W. Guo, Z. Wu, C. Wang, H. Deng, Z. Weng, Y.-P. Tan, and Z. Wang, “Map-VLA: Memory-augmented prompting for vision-language-action model in robotic manipulation,” *arXiv preprint arXiv:2511.09516*, 2025.
- [177] M. Memmel, J. Berg, B. Chen, A. Gupta, and J. Francis, “STRAP: Robot sub-trajectory retrieval for augmented policy learning,” in *ICLR*, 2025.
- [178] J. Liu, Y. Qi, J. Zhang, M. Li, S. Wang, K. Wu, H. Ye, H. Zhang, Z. Chen, F. Zhong *et al.*, “TrackVLA++: Unleashing reasoning and memory capabilities in VLA models for embodied visual tracking,” *arXiv preprint arXiv:2510.07134*, 2025.
- [179] S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, and M. Moor, “Agentclinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments,” *arXiv preprint arXiv:2405.07960*, 2024.
- [180] J. Hu, A. Wang, Q. Xie, H. Ma, Z. Li, and D. Guo, “Agentmental: An interactive multi-agent framework for explainable and adaptive mental health assessment,” *arXiv preprint arXiv:2508.11567*, 2025.
- [181] J. Li, Y. Lai, W. Li, J. Ren, M. Zhang, X. Kang, S. Wang, P. Li, Y.-Q. Zhang, W. Ma *et al.*, “Agent hospital: A simulacrum of hospital with evolvable medical agents,” *arXiv preprint arXiv:2405.02957*, 2024.
- [182] O. Bodenreider, “The unified medical language system (umls): integrating biomedical terminology,” *Nucleic acids research*, 2004.
- [183] M. Moritz, E. Topol, and P. Rajpurkar, “Coordinated ai agents for advancing healthcare,” *Nature Biomedical Engineering*, 2025.
- [184] B. P. de Almeida, G. Richard, H. Dalla-Torre, C. Blum, L. Hexemer, P. Pandey, S. Laurent, C. Rajesh, M. Lopez, A. Laterre *et al.*, “A multimodal conversational agent for dna, rna and protein tasks,” *Nature Machine Intelligence*, 2025.
- [185] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, “Augmenting large language models with chemistry tools,” *Nature Machine Intelligence*, 2024.
- [186] Y. Zou, A. H. Cheng, A. Aldossary, J. Bai, S. X. Leong, J. A. Campos-Gonzalez-Angulo, C. Choi, C. T. Ser, G. Tom, A. Wang *et al.*, “El agente: An autonomous agent for quantum chemistry,” *Matter*, 2025.
- [187] K. Huang, S. Zhang, H. Wang, Y. Qu, Y. Lu, Y. Roohani, R. Li, L. Qiu, G. Li, J. Zhang *et al.*, “Biomni: A general-purpose biomedical ai agent,” *biorxiv*, 2025.
- [188] Z. Zhang, Z. Ren, C.-W. Hsu, W. Chen, Z.-W. Hong, C.-F. Lee, A. Penn, H. Xu, D. J. Zheng, S. Miao *et al.*, “A multimodal robotic platform for multi-element electrocatalyst discovery,” *Nature*, 2025.
- [189] K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, and J. Zou, “The virtual lab of AI agents designs new SARS-CoV-2nanobodies,” *Nature*, 2025.
- [190] Z. Xiong, Y. Lin, W. Xie, P. He, J. Tang, H. Lakkaraju, and Z. Xiang, “How memory management impacts llm agents: An empirical study of experience-following behavior,” *arXiv preprint arXiv:2505.16067*, 2025.
- [191] Y. Bei, W. Zhang, S. Wang, W. Chen, S. Zhou, H. Chen, Y. Li, J. Bu, S. Pan, Y. Yu *et al.*, “Graphs meet ai agents: Taxonomy, progress, and future opportunities,” *arXiv preprint arXiv:2506.18019*, 2025.
- [192] Y. Du, W. Huang, D. Zheng, Z. Wang, S. Montella, M. Lapata, K.-F. Wong, and J. Z. Pan, “Rethinking memory in ai: Taxonomy, operations, topics, and future directions,” *arXiv preprint arXiv:2505.00675*, 2025.
- [193] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, “Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters,” *Internet Mathematics*, 2009.
- [194] W. Fan, C. Hu, and C. Tian, “Incremental graph computations: Doable and undoable,” in *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017.
- [195] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [196] J. Ahn, S. Hong, S. Yoo, O. Mutlu, and K. Choi, “A scalable processing-in-memory accelerator for parallel graph processing,” in *Proceedings of the 42nd annual international symposium on computer architecture*, 2015.

- [197] Y. Shao, H. Li, X. Gu, H. Yin, Y. Li, X. Miao, W. Zhang, B. Cui, and L. Chen, “Distributed graph neural network training: A survey,” *ACM Computing Surveys*, 2024.
- [198] T. T. Mueller, D. Usynin, J. C. Paetzold, D. Rueckert, and G. Kaissis, “Sok: Differential privacy on graph-structured data,” *arXiv preprint arXiv:2203.09205*, 2022.
- [199] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, “Universal and transferable adversarial attacks on aligned language models,” *arXiv preprint arXiv:2307.15043*, 2023.
- [200] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr, “Poisoning web-scale training datasets is practical,” in *2024 IEEE Symposium on Security and Privacy (SP)*, 2024.
- [201] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *ICML*, 2017.
- [202] A. Hogan, E. Blomqvist, M. Cochez, C. d’ Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, “Knowledge graphs,” *ACM Computing Surveys*, 2021.
- [203] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, “Relational inductive biases, deep learning, and graph networks,” *arXiv preprint arXiv:1806.01261*, 2018.
- [204] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [205] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, “Neuroscience-inspired artificial intelligence,” *Neuron*, 2017.

附录 A

初步研究

本节提供了理解基于图形的智能体记忆系统所需的基础概念和形式化定义。由于图作为基本结构，我们首先介绍图论基础，然后引入基于大语言模型的智能体架构，最后形式化记忆组件。在给出正式定义之前，常用符号列于表 II 中。

A. 图论基础

a) 图的定义：我们将图定义为 $G = (V, E, X)$ ，其中 V 表示结点集合， $E \subseteq V \times V$ 表示编码结点之间成对关系的边集合， X 表示与结点相关联的特征。

图结构由邻接矩阵 $A \in \mathbb{R}^{|V| \times |V|}$ 表示，其中每个元素 A_{ij} 描述了结点 v_i 与 v_j 之间的关系。若 $A_{ij} \in \{0, 1\}$ ，该图为无权图，表示边的有无。若 $A_{ij} \in \mathbb{R}$ ，该图为加权图，相应边的权重由 w_{ij} 表示。

结点特征 X 可能由连续向量或非结构化文本组成。在文本属性图中， X 对应于与结点相关联的文本描述，边关系通常是二元且无向的，即 $e_{ij} = e_{ji}$ 。在知识图谱中，结点特征可表示为文本或向量，而边编码语义关系，通常为有向，即 $e_{ij} \neq e_{ji}$ 。知识图谱广泛使用边标签来表示关系类型（例如，“located_in”，“has_property”）。当存在边权重时，它们进一步量化关系的强度或置信度。

1) 不同的图表：

a) 知识图谱：一个**知识图谱**是统一图表示 $G = (V, E, X)$ 的一种实例，其中结点对应实体，边编码了带有类型的语义关系。它通常被形式化为 $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ ，其中 \mathcal{E} 表示实体， \mathcal{R} 表示关系类型， $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ 是关系三元组的集合。每个三元组 (h, r, t) 对应一条从 $v_i = h$ 到 $v_j = t$ 的有向边 e_{ij} ，关系语义附加在边上。知识图谱通常是多关系的有向图，结点特征 X 以文本描述、学成嵌入或两者兼具的形式表示。

b) 时序图：一个**时序图**通过将边与来自时间域 T 的时间相关联的信息相结合，扩展了静态图 $G = (V, E, X)$ 。每条边 $e_{ij} \in E$ 都与一个时间戳或时间段 $\tau(e_{ij}) \subseteq T$ 相关联，使得邻接结构 A_{ij} 能够随时间变化。这种形式化支持动态的邻域定义以及对演化交互、事件序列和状态转移的时序推理，同时保持相同的结点特征表示 X 。

c) 超图：一个**超图**通过允许每条边连接多于两个结点，推广了 $G = (V, E, X)$ 的成对型边结构。形式上，边被定义为顶点的子集，即 $E \subseteq 2^V$ ，从而能够表示无法简化为二元交互的高阶和多向关系。结点特征 X 仍与单个顶点相关联，而邻域关系则由共享的超边成员身份而非成对邻接所诱导。

d) 图的变体：几种常用的图结构变体可被视为对统一表示 $G = (V, E, X)$ 的任务驱动实例，通过在结点和边定义上施加特定约束而获得。**二值图**将边值限制为 $A_{ij} \in \{0, 1\}$ ，仅建模关系的存在性。**文本属性图**为结点关联文本特征，支持语言感知的表示学习。**分块图**进一步将文本分解为更细粒度的单元作为结点，边编码上下文或语义关系。**层次图**引入结构约束以表示多层级关系。这些变体共享相同的潜在图结构形式，仅在 V 、 E 与 X 的具体实例化方式上有所不同，以满足实际建模需求。

这些图结构变体为检索、表示学习与分类提供了灵活的抽象。在智能体图记忆系统中，它们支持对异构信息的高效组织与推理，同时保持统一的结构基础。

2) 图论算法：

a) 图嵌入：图嵌入旨在将结点编码到连续向量空间中，同时保留图的结构或语义信息。给定一个图 $G = (V, E, X)$ ，节点嵌入被定义为一种映射

$$f: V \rightarrow \mathbb{R}^d,$$

其中每个结点 $v_i \in V$ 与一个 d 维表示 $\mathbf{h}_i = f(v_i)$ 相关联。

b) 基于拓扑的嵌入：一种典型的传统方法是 *Node2Vec*，它通过在有偏随机游走生成的结点序列上优化 Skip-gram 目标来学习节点嵌入。形式上，通过最大化 $\log \Pr(v_j | v_i)$ 来学习嵌入 \mathbf{h}_i ：

$$\Pr(v_j | v_i) = \frac{\exp(\mathbf{h}_i^\top \mathbf{h}_j)}{\sum_{v_k \in V} \exp(\mathbf{h}_i^\top \mathbf{h}_k)},$$

其中 $\mathcal{N}_l(v_i)$ 表示由长度为 l 的有偏随机游走诱导的 v_i 的上下文结点。该公式通过结点共现模式捕捉图结构。

c) 基于图神经网络的嵌入：图神经网络通过迭代的邻域聚合计算节点嵌入。在第 k 层，结点 v_i 的嵌入更新为

$$\mathbf{h}_i^{(k+1)} = \sigma(\mathbf{W}^{(k)} \cdot \text{AGG}^{(k)}(\{\mathbf{h}_j^{(k)} | v_j \in \mathcal{N}(v_i)\})),$$

其中 $\text{AGG}(\cdot)$ 为置换不变的聚合函数， $\mathbf{W}^{(k)}$ 为可学习的参数矩阵， $\sigma(\cdot)$ 表示非线性激活。

表 II: 图论基础符号

Symbol	Formal Definition	Meaning
G	$G = (V, E, X)$	Graph represented by node set, edge set, and node features
V	$V = \{v_1, \dots, v_{ V }\}$	Set of nodes (vertices)
E	$E \subseteq V \times V$	Set of edges encoding pairwise relations between nodes
X	$X \in \mathcal{X}$	Node feature set, consisting of vectors or unstructured texts
v_i, v_j	$v_i, v_j \in V$	The i -th and j -th nodes
e_{ij}	$e_{ij} = (v_i, v_j)$	Edge from node v_i to node v_j
$\mathcal{N}(v_i)$	$\mathcal{N}(v_i) = \{v_j \mid e_{ij} \in E\}$	Neighborhood set of node v_i
$d(v_i)$	$d(v_i) = \mathcal{N}(v_i) $	Degree of node v_i
A	$A \in \mathbb{R}^{ V \times V }$	Adjacency matrix representing graph structure
A_{ij}	$A_{ij} \in \{0, 1\}$ or \mathbb{R}	Adjacency matrix entry encoding the relation between v_i and v_j
w_{ij}	$w_{ij} \in \mathbb{R}$	Weight associated with edge e_{ij} (if applicable)
\mathbf{h}_i	$\mathbf{h}_i \in \mathbb{R}^d$	d -dimensional embedding of node v_i

基于语言模型的嵌入。 对于一个具有文本属性的结点 v_i ，其关联的 token 序列为 X_i ，基于 Transformer 的语言模型将其输入编码为

$$\mathbf{H}_i = \text{Transformer}(X_i),$$

其中 $\mathbf{H}_i \in \mathbb{R}^{L \times d}$ 表示最后一层的 token 表示。节点嵌入定义为

$$\mathbf{h}_i = \mathbf{H}_{i, [\text{CLS}]},$$

即，特殊分类标记的表示，遵循 BERT 类模型的标准做法。

d) 图遍历：图遍历方法定义了探索图拓扑的系统性或随机性过程 $G = (V, E)$ ，其目标是发现可达的结点、结构路径或支持检索、推理和表示学习的任务相关子图。

- **广度优先搜索 (BFS)** 按照从源结点 v_i 出发的最短路径距离的递增顺序探索图，迭代地访问相继的邻域 $N_1(v_i), N_2(v_i), \dots$ 中的结点，常用于无权图中的局部邻域扩展和最短路径发现。
- **深度优先搜索 (DFS)** 通过递归地尽可能深入地遍历相邻结点，然后回溯，从而实现路径的高效枚举、连通性分析和环路检测。
- **随机游走** 定义了一种随机遍历过程，其中根据转移概率生成结点序列 (v_0, v_1, \dots, v_k)

$$p(v_{j+1} = v \mid v_j) = \frac{A_{v_j v}}{\sum_{u \in \mathcal{N}(v_j)} A_{v_j u}},$$

并构成了经典节点嵌入方法的理论基础。

- **最短路径遍历** 寻找一条路径 $P = \langle v_0 = s, \dots, v_k = t \rangle$ ，以最小化累积边权重

$$\sum_{i=0}^{k-1} w_{v_i v_{i+1}},$$

其中 w_{ij} 表示与边 e_{ij} 相关联的权重。

- **子图提取** 识别以 v_i 为中心的局部诱导子图，通常定义为 k -跳邻域

$$N_k(v_i) = \{v_j \in V \mid d(v_i, v_j) \leq k\},$$

广泛用于局部检索和下游基于图形的建模。

B. 基于大型语言模型的智能体

表 III 总结了智能体系统和基于图形的记忆公式中使用的关键符号及其语义解释。

1) 智能体系统：一个基于语言模型的智能体是一种决策系统，它以大型语言模型作为其核心推理组件，与环境交互并完成给定任务。形式上，一个智能体 \mathcal{A} 在状态空间为 \mathcal{S} 且任务规范为 Q 的环境中运行。

在每个时间步 t ，智能体选择一个动作 a_t ，该动作通过转移过程 Ψ 驱动环境的演化，基于部分观测 o_t 而非对潜在状态的完整访问。智能体的行为由参数化语言模型所诱导的策略决定，该策略整合了观测、交互历史 h_t ，以及可选的外部记忆或工具，以支持推理和决策，而具体的操作流程则由后续引入的智能体交互环指定。

a) 智能体交互环：具体而言，在时间步 t 期间，智能体通过结构化的感知-检索-推理-动作更新环与环境交互：

- **感知。** 智能体接收一个观测

$$o_t = O(s_t, h_t, Q),$$

这编码了在任务规范 Q 下关于环境状态 s_t 的部分信息。

- **检索。** 给定当前观测和交互历史，智能体查询其外部记忆：

$$c_t = \text{Retrieve}(\mathcal{M}, o_t, h_t),$$

Symbol	Description
\mathcal{A}	LLM-based agent
\mathcal{S}	Environment state space
s_t	Environment state at time step t
Q	Task specification
t	Discrete interaction time step
a_t	Action selected by the agent at time t
o_t	Partial observation received by the agent at time t
r_t	Feedback or reward signal at time t
h_t	Agent interaction history up to time t
Ψ	Environment transition process
$O(\cdot)$	Observation function
\mathcal{M}_θ	Parameterized LLM policy with parameters θ
\mathcal{M}	Agent memory module
c_t	Retrieved memory context at time t
q_t	Memory query derived from task or observation
\mathcal{M}_G	Graph-based memory module
G_t	Memory graph at time t
V_t	Set of memory nodes at time t
E_t	Set of edges (relations) at time t
X_t	Node and edge attributes of the memory graph
v_i	A memory node
e_{ij}	A directed or undirected relation between v_i and v_j
\mathcal{D}	Static corpus/experience set for memory construction
Δ_t	Online memory update signal (o_t, a_t, r_t)

表 III: 智能体交互环与基于图形的记忆框架中使用的符号。

其中 c_t 表示检索到的上下文信息。

- **推理。** LLM 主干网络整合了观测信息、检索到的记忆以及历史信息，以进行推理和决策：

$$a_t \sim \mathcal{M}_\theta(o_t, c_t, h_t).$$

- **动作。** 选定的动作 a_t 在环境中执行，引发状态转移

$$s_{t+1} \sim \Psi(s_{t+1} \mid s_t, a_t).$$

- **更新。** 智能体将新的经验与反馈融入记忆中：

$$\mathcal{M} \leftarrow \text{Update}(\mathcal{M}, o_t, a_t, r_t).$$

记忆模块 \mathcal{M} 存储一组经验元组和知识表征，并支持检索和更新操作。其内部结构在此阶段未作具体说明，后续章节中可实例化为键-值数据库、情景缓冲区或结构化图记忆。

2) 从记忆中构建提示：在基于大语言模型的智能体系统中，记忆主要通过提示词条件化影响智能体的行

为。在每个时间步 t ，智能体构建一个复合提示词，该提示词整合了系统级指令、检索到的记忆内容以及当前任务上下文。形式上，提示词可抽象为

$$\text{Prompt}_t = \text{System}(I) \oplus \text{Memory}(c_t) \oplus \text{Task}(o_t),$$

其中， I 指定智能体的角色和操作约束， c_t 表示检索到的记忆上下文， o_t 代表当前的观测或查询。

内存上下文 c_t 通过使用与任务相关的查询 q_t 向内存模块进行查询获得，通常是通过基于相似度的检索：

$$c_t = \text{TopK}_{m_i \in \mathcal{M}}(\text{sim}(q_t, m_i)),$$

其中 m_i 表示个体记忆单元， $\text{sim}(\cdot, \cdot)$ 是在其表示上定义的相似度函数。该公式提供了一个统一的接口，使存储的经验和知识能够融入智能体的推理过程。

C. 基于图形的内存

在基于大语言模型的智能体系统中，记忆作为存储、组织和检索过往经验与知识的持久化机制。当记忆以结构化形式实现时，可自然地表示为图结构，从而支持关系建模、高效检索以及随时间的增量更新。

1) 记忆的图格式：我们将基于图形的智能体记忆形式化为一个动态属性图

$$\mathcal{M}_G \triangleq G_t = (V_t, E_t, X_t),$$

其中， V_t 表示内存结点的集合， $E_t \subseteq V_t \times V_t$ 表示结点间的关系集合， X_t 表示与结点和边相关的属性。该表述与之前介绍的基本图结构一致，同时允许针对特定任务进行实例化。

具体而言，每个结点 $v_i \in V_t$ 对应一个记忆单元，例如实体、事件、概念或文本片段。每条边 $e_{ij} \in E_t$ 表示记忆单元之间的关系，可能编码语义、时间或因果依赖。结点属性 $X_t(v_i)$ 通常包括文本内容或向量嵌入，而边属性可能包括关系类型、置信得分或时间信息。

2) 图记忆机制：

a) 图记忆构建：记忆构建指的是从非结构化或半结构化信息中独立于在线智能体动作，初始形成图结构记忆的过程。形式上，给定一个语料库或经验集 \mathcal{D} ，图构建定义了一个映射

$$\text{Construct} : \mathcal{D} \rightarrow G_0 = (V_0, E_0, X_0),$$

其中结点 V_0 为提取的存储单元，边 E_0 编码它们之间的关系。

构建通常包括 (i) 结点提取, 即识别实体、事件、概念或文本片段作为结点, 以及 (ii) 关系抽取, 即识别语义、时间或结构关系以形成边。生成的图可表示为三元组 (v_i, e_{ij}, v_j) , 并作为初始状态。

b) 图记忆检索: 记忆检索对应于查询图以识别相关的一组结点或子图:

$$\text{Retrieve} : (q, G_t) \rightarrow \mathcal{S}_t \subseteq V_t,$$

其中, 查询 q 可以是文本型、结构型或基于嵌入的。检索可通过结点表示的语义相似度、从与查询相关的结点进行图遍历, 或两者的结合来实现。

c) 图记忆更新: 记忆更新描述了智能体与环境交互所驱动的现有记忆图的在线演化。在时间步 t , 更新信号

$$\Delta_t \triangleq (o_t, a_t, r_t)$$

由智能体的观测、动作和反馈推导而来。给定当前的记忆图 G_t , 更新操作定义为

$$\text{Update} : (G_t, \Delta_t) \rightarrow G_{t+1}.$$

更新过程通过添加或修改结点和边, 调整其属性或修订关系, 将新观测到的信息整合到图中, 以反映新获取的证据。与从静态数据构建初始图的记忆构造不同, 记忆更新采用增量方式, 在智能体交互过程中实现对记忆结构的持续适应。

3) 图记忆质量: 为了系统评估智能体系统中的基于图形的记忆机制, 有必要同时评估记忆检索的质量、图结构表述所引起的结构性质, 以及它们对下游任务性能的影响。我们从三个互补的维度总结了具有代表性的评估标准。

a) 检索效果: 检索质量衡量记忆图在响应查询时揭示相关信息的能力。常用指标包括 Precision@K 和 Recall@K, 它们量化了检索出的前 K 个结点中的相关性, 以及平均倒数秩 (MRR), 用于捕捉首个相关记忆的排名位置。

b) 图结构质量: 图质量度量评估所构建的记忆图是否对存储的知识提供了连贯且忠实的表示。典型标准包括: 连贯性, 反映结构一致性和连通性; 完整性, 衡量关键信息的覆盖程度; 冗余性, 捕捉不必要的重复内容; 以及时间一致性, 评估时间关系是否被正确保留。

c) 任务级效用: 任务性能指标用于评估图记忆在智能体决策中的功能有效性, 包括任务成功率、交互效率以及对未见任务或领域的泛化能力。

附录 B

开源库

在表 IV 中, 我们对十一款具有代表性的开源内存库在关键功能维度上进行了系统性比较。列项涵盖了内存系统的重要方面, 包括许可证类型、构建模式 (基于外部知识或交互驱动)、对基于图形的内存的支持、检索机制、生命周期管理、时间建模、推理能力、条件控制、个性化、层次结构以及与智能体框架的集成。这一结构化概览使得智能体内存设计中的内存设计选择能够进行细致入微的对比。

³<https://docs.cognee.ai/>

⁴<https://langchain-ai.github.io/langmem/>

⁵<https://docs.mem0.ai/open-source/overview>

⁶<https://github.com/zjunlp/LightMem>

⁷<https://github.com/OPPO-PersonalAI/O-Mem>

⁸<https://github.com/CaviraOSS/OpenMemory>

⁹<https://github.com/GibsonAI/Memori>

¹⁰<https://github.com/MemMachine/MemMachine>

¹¹<https://github.com/kingjulio8238/Memary>

¹²<https://github.com/getzep/graphiti>

¹³<https://github.com/memvid/memvid>

表 IV: 基于图形的内存系统开源库比较

ID	Library	License		External Construct.	Interaction Construct.	Graph Memory	Retrieval	Lifecycle	Temporality	Reasoning	Conditioning	Personalization	Hierarchy	Agent Integration
		Apache2.0	MIT											
1	Cognee ³	✓		✓		✓	✓							✓
2	LangMem ⁴		✓		✓		✓	✓			✓	✓		✓
3	Mem0 ⁵	✓			✓	✓	✓	✓			✓	✓	✓	✓
4	LightMem ⁶		✓		✓		✓	✓		✓		✓	✓	✓
5	O-Mem ⁷	✓			✓		✓	✓			✓	✓	✓	✓
6	OpenMemory ⁸	✓		✓	✓	✓	✓	✓	✓		✓		✓	✓
7	Memori ⁹	✓			✓		✓			✓				
8	MemMachine ¹⁰	✓				✓				✓				
9	Memary ¹¹		✓	✓	✓	✓	✓				✓	✓		✓
10	Graphiti ¹²	✓		✓	✓	✓	✓			✓				
11	Memvid ¹³	✓		✓	✓		✓	✓	✓					