arXiv:2601.06490v1 [cs.MA] 10 Jan 2026

Bi-Mem：通过归纳反思智能体双向构建个性化大语言模型的分层记忆

Wenyu Mao[1]*、Haosong Tan[1]* Shuchang Liu[2] Haoyang Liu[1],Yifan Xu[1], HuaxiangJi[1], Xiang Wang[1]† 、[1] 中国科学技术大学[2]快手科技

摘要

从用户的长期对话中构建记忆能够克服大语言模型的上下文局限，并实现个性化交互。近期研究聚焦于分层记忆，通过聚类和聚合历史对话来建模用户的多粒度行为模式。然而，对话噪声和记忆幻觉可能在聚类过程中被放大，导致局部聚合记忆与用户的全局用户画像不一致。为缓解这一问题，我们提出了 Bi-Mem，这是一种通过双向构建来确保分层记忆保真度的智能体框架。具体而言，我们部署了一个归纳智能体来形成分层记忆：它从原始对话中提取事实信息，形成事实级记忆；通过图聚类将这些记忆聚合为主题场景（即 局部场景级记忆）；并将用户画像推断为全局人物级记忆。同时，我们设计了一个反思智能体，利用从人物设定级记忆中得到的全局约束，对局部场景级记忆进行校准，从而强化全局-局部对齐。为实现连贯的记忆召回，我们提出了一种关联检索机制：在初始分层搜索之外，扩散激活过程使事实能够唤起情境场景，而场景层级匹配则检索出具有显著性的支持性事实信息。实证评估表明，在长期个性化对话任务中，Bi-Mem 在问答性能方面取得了显著提升。



图 1：在朴素的层级记忆系统中，局部聚合记忆与用户的全局用户画像不一致，导致大语言模型生成违背用户画像的回答。

1 引言

大语言模型（LLM）的个性化Chen 等人，2024；Zhang 等人，2025b）旨在满足用户的个性化需求，例如回忆共同经历，以及根据历史对话生成符合偏好的建议。为支持长期对话任务中的个性化交互，由于大语言模型上下文窗口有限，构建记忆（Chhikara 等人，2025年；Li 等，2025年；Xu等，2025年）已成为一种重要范式。早期研究侧重于从历史对话中提取关键实体或对话摘要，将其作为事实性记忆（Packer 等人，2023年；Chhikara 等人，2025年）。然而，这类碎片化的事实记忆难以捕捉事实间关系或用户较高层次的行为模式（Rezazadeh 等人，2025年；Pan 等人，2025年）。因此，近期研究探索通过聚类并汇总这些孤立事实，形成不同粒度的层次化记忆结构（例如，多层图（Li 等，2025年；Gutierrez 等人，2024）或树结构（Rezazadeh 等人，2025年））。通过支持由粗到细检索（Zhao 等人，2025年），分层系统能够提供更全面的上下文知识，从而实现更连贯、更个性化的交互。

尽管分层记忆取得了成功，这种聚类与聚合仍面临两个关键局限：

*共同贡献
†通讯作者：xiangwang@ustc.edu.cn。

噪声放大和幻觉累积（Huang 等，2025年；Zhang 和 Zhang，2025年）。首先，对话噪声——例如无关闲聊和偶发交互——往往会在局部簇中被放大，导致聚合后的局部记忆偏向琐碎内容。其次，事实层面幻觉（Chen 等，2025年）和提取错误会在层级整合过程中传播并累积。这些“级联错误”共同导致局部聚合记忆与用户的全局用户画像不一致（Wang 等，2025b）。如图1所示，设想一位用户经常为家常饭菜寻找清淡口味的食谱，却会为了与喜爱辣味食物的朋友参加一次社交聚餐而寻找辛辣菜肴。朴素的聚类方法往往会放大这种短暂的社交情境，进而将这些情境聚合成一条局部行为规则：“外出就餐时偏好辛辣菜肴。”相比之下，从更广泛的对话中得出的全局用户画像是“讨厌辣味食物，口味清淡”。因此，聚合后的局部记忆会偏离用户的全局用户画像，导致大语言模型在家庭外出就餐场景中生成违背用户画像的建议。

为解决这些问题，我们提出了 Bi-Mem，这是一种通过双向构建而非单向聚合来确保分层记忆保真度的智能体框架（Li 等人，2025年）。具体而言，我们采用归纳智能体来管理自底向上的层级结构构建：从原始对话中提取原子事实层级记忆，通过图聚类将其聚合为主题场景层级记忆（Pattnaik 等人，2024），并进一步将其提炼为五维人物设定级记忆（Zhong 等人，2024），作为全局约束。与此同时，我们设计了反思智能体来执行自顶向下的校准，将全局约束施加于局部场景记忆，以实现全局-局部对齐。如图 1所示，通过注入“我喜欢清淡口味”这一全局约束来校准局部场景记忆，可以解决全局与局部之间的矛盾，从而使大语言模型能够生成符合偏好的建议。为连接不同粒度的记忆，我们引入了联想检索机制。在初始的分层检索之外，该机制还采用扩散激活过程（Zhang 等人，2025a）：检索到的事实会唤起相关的主题场景，而场景层级匹配则会检索出显著支持事实。实证结果表明，Bi-Mem 显著提升了个性化长期对话任务中的问答性能。

总结如下对比ibuti ons o f thi s wor k

•我们将全局人格与局部聚合记忆失调识别为分层记忆面临的一项普遍挑战，其成因是噪声和幻觉。

• 我们设计了 Bi-Mem，这是一种智能体框架，通过双向构建（即归纳过程和反思过程）确保分层记忆保真度。

• 我们提出一种联想检索机制，以促进连贯的记忆召回，从而加强对细粒度事实信息与上下文场景的整合。

• 在复杂的个性化长期对话任务上开展的大量实证评估，证明了 Bi-Mem 的有效性。

2 相关工作

个性化大语言模型。 个性化大语言模型旨在基于用户与大语言模型的交互历史，使模型回复适应特定用户的偏好与个性（Chen 等，2024；Zhang 等，2025b）。与仅提供通用知识不同（Kim 等，2023），它们能够满足用户的特定需求，例如回忆共同经历或生成符合偏好的建议，从而提升用户的参与度与满意度。为处理长期交互历史，现有方法要么基于对话数据微调用户专属模块（Tan 等，2024b，a），要么采用记忆架构（Wang 等，2025b；Zhong 等，2024），在不修改模型参数的情况下存储和管理用户专属对话。考虑到微调所需的计算资源以及大语言模型的上下文窗口限制，我们专注于为用户构建长期对话历史记忆，以实现此类个性化。

个性化大语言模型的记忆。 个性化大语言模型的记忆（Zhang 等人，2024；Hu 等人，2025年）是一种灵活的组件，用于存储和利用用户的历史对话信息，以弥补大语言模型上下文窗口的局限。它主要由两部分构成：记忆构建（Xu 等人，2025年）（从用户与大语言模型的交互中提取关键信息，并将其整理成结构化形式）和记忆检索（Wang 等人，2025a）（通过语义相似度检索相关记忆，以支持大语言模型的响应生成）。主流实现方式

采用基于 token 的存储方式。例如，Mem0（Chhikara 等人，2025年）通过支持添加/更新/删除/无操作等操作，从对话信息中构建记忆，并通过向量相似度检索相关记忆。MemoryBank（Zhong 等人，2024）通过整合类似人类记忆的存储方式（对话、事件摘要、用户画像）和稠密检索，为个性化大语言模型提供支持。

分层记忆。个性化大语言模型的基于令牌的记忆有两种核心结构：扁平记忆（Packer 等人，2023）将信息存储为彼此独立的单元（例如，原始交互和摘要会话），不建立显式连接。分层记忆（Hu 等人，2025年）采用具有不同粒度和关联关系的多层架构，并通过跨层连接支持复杂推理。例如，CAM（Li 等人，2025年）采用增量重叠聚类算法构建分层记忆，并通过剪枝-增长策略进行检索。HippoRAG（Gutierrez 等人，2024）采用无模式的分层知识图谱（KG）进行记忆构建，并利用个性化 PageRank实现跨记忆推理。

3 初步介绍

3.1 任务定义

设 $\mathcal{C} = \{c_1, ..., c_N\}$ 表示用户与大语言模型代理之间的长期对话历史，其中每个 $c_i = (q_i, r_i)$ 表示第 i 次交互（用户请求 q_i，大语言模型响应 r_i）。记忆代理的任务由记忆构建和检索任务组成（Pan 等人，2025年）。

• **记忆构建：** 通过记忆构建函数 $\mathcal{M}$ 对对话历史进行编码 $\mathcal{C}$，构建结构化记忆库：

$$\mathcal{M} = f_{\text{cons}}(\mathcal{C}) \tag{1}$$

其中，f_{cons} 将 $\mathcal{C}$ 构建为离散记忆单元 $\boldsymbol{m} \in \mathcal{M}$。

• **记忆检索：** 给定目标用户查询 q^*和记忆库 $\mathcal{M}$，通过检索函数提取与 q 相关的 K 记忆单元 f_R：

$$\{\boldsymbol{m}_k \in \mathcal{M}\}_{k=1}^{K} \leftarrow f_R(q^*, \mathcal{M}, K) \tag{2}$$

• **响应生成：** 大语言模型个性化（Wang 等人，2025b）的目标是生成最优的个性化响应 r^*，以满足当前查询 q^*，基于检索到的 N 记忆单元：

$$r^* = f_{\text{LLM}}\left(q^*, \{\boldsymbol{m}_k\}_{k=1}^{K}\right) \tag{3}$$

其中，r^* 与用户画像相符，或能准确召回相关上下文细节。

4 方法

本节将详细介绍我们提出的 Bi-Mem，这是一种采用双向构建分层记忆并进行联想检索的智能体框架，如图 2 所示。我们首先在4.1节定义所设计的分层记忆的三个层级。然后，在4.2节介绍双向记忆构建，包括归纳-反思过程；随后在4.3节详细说明联想检索机制。详细算法见附录A。

4.1 分层记忆的定义

Bi-Mem 中的分层记忆 $\mathcal{M} = (\mathcal{F}, \mathcal{S}, \mathcal{P})$ 由三个粒度层次构成：

• **事实级记忆 ($\mathcal{F}$)** 是记忆的基本单元，每次对话交互都映射为一个事实单元。形式上，$\mathcal{F} = \{f_1, f_2, ..., f_N\}$，其中每个单元 f_i 定义为元组 $(i, \text{con}_i, t_i, e_i)$：$i$ 是事实的索引，con_i 是交互的摘要内容，t_i 是时间戳，e_i 是连接到相关事实或交互索引的连接边。

• **场景级记忆($\mathcal{S}$)**将相关事实聚类，形成主题场景（例如，“与喜欢吃辣的朋友外出就餐”）。形式上，$\mathcal{S} = \{s_1, s_2, ..., s_J\}$，其中每个场景 s_j 是事实级记忆 $\mathcal{F}_j \subseteq \mathcal{F}$:$s_j = \frac{1}{|\mathcal{F}_j|} \sum_{f \in \mathcal{F}_j} f$的一个子集的聚合。

• **人物设定级记忆 ($\mathcal{P}$)**代表从所有汇总场景$\mathcal{S}$中提炼出的用户核心概况，充当确保记忆一致性的全局约束。形式化地，$\mathcal{P} = [p_1, p_2, p_3, p_4, p_5]$ 是一个五维人格，其中 p_1（基本信息）涵盖姓名、年龄、职业和所在地，p_2（兴趣）涵盖爱好、喜欢和不喜欢的事物，p_3（人格）涵盖人格特征和情绪模式，p_4（价值观）涵盖核心价值观、信念和动机，p_5（社会关系）涵盖重要的社会关系。

4.2 双向记忆构建

双向构建整合了自底向上的归纳过程（用于记忆形成）和自顶向下的反思过程（用于自顶向下的校准），实现了全局-局部对齐。

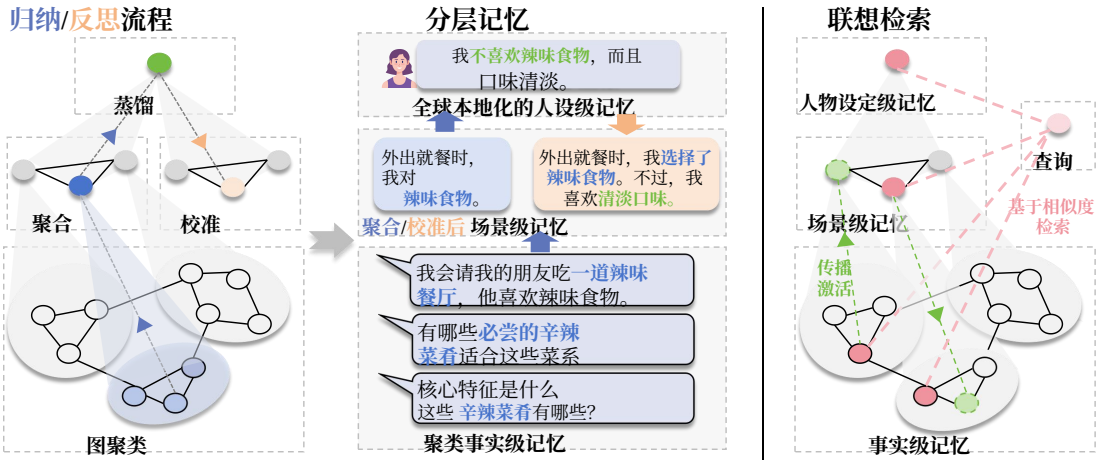

图2：我们提出的 Bi-Mem 框架，包括分层记忆的双向构建（归纳过程和反思过程）以及联想检索。

4.2.1 归纳过程（自底向上）

归纳智能体通过三个连续步骤，从原始对话中构建分层记忆：

事实级记忆提取：为了为对话中的每次交互构建事实单元 $f_i = (i,\ \text{con}_i,\ t_i,\ e_i)$，我们首先通过大语言模型应用摘要算子 $\mathcal{E}$，提取经过总结的交互内容 con_i，定义如下：

$$\text{con}_i = \arg\max_{\text{con}\in\text{Con}} P_{\mathcal{E}}(\text{con}|c_i), \tag{4}$$

其中，$i \in [1, 2, \ldots, N]$，Con 是由大语言模型生成的摘要候选集。索引i 和时间戳 t_i 直接取自交互元数据。通过计算当前内容 con_i 与其他事实单元中的 con_l 之间的语义相似度，确定连接边 e_i。形式上，$l \in e_i$ 若：

$$\text{sim}(\phi(\text{con}_i), \phi(\text{con}_l)) > \tau, \tag{5}$$

其中，sim(·) 是通过预训练嵌入模型 ϕ 得到的嵌入之间的余弦相似度，而 τ 是预定义的阈值。

场景级记忆聚合：为了提供更多上下文知识并捕捉用户的行为模式，我们将零散的原子事实转化为高层次的主题场景。我们将事实记忆建模为无向图 $G = (\mathcal{F},\ E)$，其中，若满足$f_i,\ f_l) \in E$ 或满足 $l \in e_i$，则建立一条边 $i \in e_l$。随后，我们将标签传播算法（LPA）（UganderandBackstrom，2013年）应用于 G，进行图聚类。这一过程利用局部结构共识（Lanci-chinetti and Fortunato，2012）将事实 集 $\mathcal{F}$划分为 J 主题簇 $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_J\}$。具体而言，对于每个事实簇 $\mathcal{F}_j \subseteq \mathcal{F}$，我们使用大语言模型应用聚合算子 $\mathcal{A}$，通过最大化生成概率来导出相应的场景s_j：

$$s_j = \arg\max_{s\in\mathcal{S}_{can}} P_{\mathcal{A}}(s \mid f_k \in \bar{\mathcal{F}}_j), \tag{6}$$

其中，$\mathcal{S}_{can}$ 是候选生成场景的集合。因此，我们可以构建主题场景级记忆 $\mathcal{S} = \{s_1, s_2, \ldots, s_J\}$。

人格级记忆蒸馏：为了获得记忆的全局约束，我们从全部场景级记忆中蒸馏出五维人格 $\mathcal{P}$，详见4.1。具体而言，我们定义一个蒸馏算子 $\mathcal{D}$，利用大语言模型将集合 $\mathcal{S}$ 投射到各个人格维度上：

$$p_d = \arg\max_{p\in\mathcal{P}_{cand}} P_{\mathcal{D}}(p \mid \mathcal{S}, \text{per}_d), \tag{7}$$

其中，$d \in 1$、.....、5，per_d 是针对特定维度的指令，用于引导大语言模型关注 $\mathcal{S}$ 中的相关场景。

4.2.2 反思过程（自顶向下）

为确保全局与局部对齐，并减少归纳过程中的级联错误，我们实现了一个反思智能体来执行自顶向下校准，确保每个局部场景 $s_j \in \mathcal{S}$ 都以稳定的人格约束 $\mathcal{P}$ 为依据。

具体而言，反思智能体首先评估局部场景 s_j 是否与相关的全局用户画像相矛盾，或未能反映该画像

（位于 $\mathcal{P}$ 内）。随后，它应用校准算子 $\mathcal{R}$，生成补偿条件 Δs_j，明确地将局部场景 s_j 与全局用户画像联系起来：

$$\Delta s_j = \arg\max_{\delta \in \Delta_{cand}} P_{\mathcal{R}}(\delta \mid s_j, \mathcal{P}), \quad (8)$$

其中，Δ_{cand} 表示候选补偿条件的空间。校准后的场景形式化定义为：$s'_j = s_j \oplus \Delta s_j$，其中，如果局部场景已与全局用户画像一致，则 Δs_j 为空。此反思过程将全局约束注入局部场景，从而形成包含场景$\mathcal{S}' = \{s'_1, s'_2, \ldots, s'_J\}$的自洽分层记忆。由于事实级记忆由详细的对话事实构成，我们不对其进行校准以使其与现实保持一致。

4.3 联想记忆检索

为实现跨记忆粒度的连贯回忆，我们提出了一种联想检索机制。该机制在分层记忆的初始搜索后利用激活传播$\mathcal{M} = (\mathcal{F}, \mathcal{S}', \mathcal{P})$，兼顾细粒度细节与高层语境。

初始分层搜索： 给定一个查询 q^*，我们首先使用预训练嵌入模型 $\phi(\cdot)$将查询和所有记忆映射到潜在空间中。初始相关性得分 a_0 通过计算查询与每个记忆单元之间的余弦相似度得出：

$$a_0(x) = \text{sim}(\phi(q^*), \phi(x)), \quad (9)$$

其中，$x \in \mathcal{F} \cup \mathcal{S}' \cup \mathcal{P}$ 表示分层记忆中的记忆单元，sim(·)表示余弦相似度。选择得分最高的前k x个 a_0 作为检索记忆集 M_{ret}。

联想扩散激活：为捕捉细粒度事实与主题场景之间的结构依赖关系，我们实现了一种双向联想机制：检索到的事实会激活其父场景，而检索到的场景则会触发对其最具代表性事实的召回。具体而言，从初始检索记忆集 $\mathcal{M}_{ret}$ 出发，检索集通过两条联想路径扩展。首先，对于每个事实 $f \in \mathcal{M}_{ret}$，其唯一的父场景 s 会自动与之关联，并加入扩展后的集合。其次，对于每个场景 $s_j \in \mathcal{M}_{ret}$，该机制会将激活传播至其对应的事实簇 $\mathcal{F}_j$。我们选取与场景的相关性得分 a_1最高的前m 个事实：

$$a_1(x) = \text{sim}(\phi(s_j), \phi(x)), \quad \forall x \in \bar{\mathcal{F}}_j. \quad (10)$$

对于每个用户画像 $p \in \mathcal{M}_{ret}$，我们将其视为全局锚点，不与其他记忆关联。最终检索记忆集 $\mathcal{M}_{ret}$更新为初始检索结果与相关记忆的并集。

响应生成： 最终检索记忆集 $\mathcal{M}_{ret}$涵盖了从原子事实到全局角色特征的多粒度证据，并将这些证据整合到统一的提示上下文中。形式上，给定当前查询 q^* 以及 $\mathcal{M}_{ret}$中的检索信息，大语言模型通过最大化条件概率生成个性化响应 r^*：

$$r^* = \arg\max P_{\text{LLM}}(r \mid q^*, \mathcal{M}_{ret}). \quad (11)$$

5 实验

本节将通过回答以下问题，开展广泛实验，以评估我们提出的框架在长期个性化对话任务中的有效性：

- RQ1：我们的框架 Bi-Mem 与领先的记忆基线方法相比，表现如何?
- RQ2：事实、场景和角色各层级以及反思校准分别对构建分层记忆有何贡献?
- RQ3：关联式检索机制在多大程度上提升了问答准确率?
- RQ4：该框架对初始检索规模这一超参数 k有多敏感?
- 研究问题5：与现有记忆框架相比，Bi-Mem 的效率如何?

5.1 实验设置

数据集。 我们采用 LoCoMo（Maharana 等人，2024）开展实验。LoCoMo 是个性化长期对话任务中广泛使用的数据集，包含 50 段对话，每段平均有 305 轮次、20 个会话和 9,000 个词元。与现有对话数据集相比（Jang 等人，2023；Xu 等人，2022），它更适用于构建大语言模型的记忆。LoCoMo 通过严谨的人类与大语言模型协同创作流程构建：具有独特人设的基于大语言模型的智能体生成初始对话，人工标注员再对其进行完善，以确保长期一致性。对于

模型	方法	LoCoMo									
		单跳		多跳		时间		开放域		平均值	
		$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
GPT-4o-mini	LONGCONTEXT	32.69	20.85	22.87	16.84	12.47	8.86	7.87	5.89	25.08	16.65
	RAG	52.45	47.94	27.50	20.13	46.07	40.35	23.23	17.94	44.67	39.33
	SECOM	16.67	13.96	15.06	11.93	14.58	12.20	17.29	14.76	15.97	13.25
	A-MEM	44.65	37.06	27.02	20.09	45.85	36.67	12.14	12.00	39.61	32.27
	MEM0	47.65	38.72	38.72	27.13	48.93	40.51	28.64	21.58	45.08	35.88
	LightMem	41.79	37.83	29.78	24.80	43.71	39.72	16.89	13.92	38.41	34.32
	CAM	49.94	46.68	36.63	25.37	18.85	12.90	22.71	18.93	39.25	33.92
	Bi-Mem	**53.68**	**47.99**	**39.17**	**31.81**	**54.44**	**41.56**	**30.99**	**26.90**	**49.74**	**42.33**
Qwen2.5-14b-Instruct	LONGCONTEXT	21.00	13.17	18.09	13.21	10.54	5.82	7.31	4.71	17.41	11.10
	RAG	47.87	42.79	26.38	19.54	30.78	25.97	14.16	10.52	38.20	32.95
	SECOM	22.26	18.71	19.15	14.26	19.58	16.30	15.75	13.01	20.71	17.02
	A-MEM	33.75	30.04	22.09	15.28	27.19	22.05	13.49	10.74	28.95	24.43
	MEM0	42.58	35.15	31.73	24.82	28.96	26.24	15.03	11.28	35.99	29.88
	LIGHTMEM	34.92	31.22	25.45	19.61	32.03	27.70	15.81	11.81	31.37	27.12
	CAM	32.69	23.15	24.78	18.98	5.21	2.71	15.25	12.45	24.37	17.42
	Bi-Mem	**48.07**	**42.95**	**32.42**	**25.53**	**44.21**	**29.29**	**18.38**	**16.11**	**42.51**	**35.19**

表1：不同方法在 LoCoMo 基准上的总体表现。最高分以粗体标出，表示具有统计显著性的提升（$p < 0.05$）；次高分则加下划线。

基于历史对话（构建的记忆）进行进一步问答，LoCoMo 包含 7,512 个问答对，涵盖多种类型：源自单个会话的单跳问题；需要综合跨会话信息的多跳问题；用于评估对时间相关线索把握程度的时间推理问题；将对话语境与外部事实相结合的开放域知识问题；以及旨在测试模型识别不可回答查询能力的对抗性问题。请注意，我们的工作未采用对抗性问题类别，因为它与记忆构建和检索的研究范围不符。

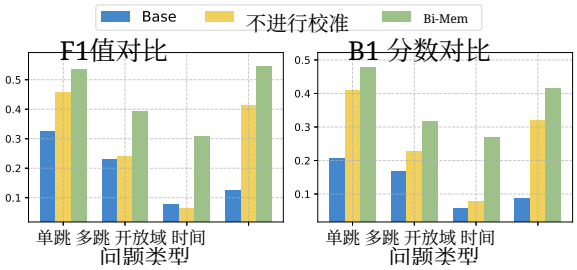

图3：层次化记忆构建中反思校准的消融研究。“基线”表示使用长上下文大语言模型骨干网络且不使用记忆进行答案生成；“无校准”表示在反思过程中未经校准的单向分层记忆构建。

基线方法 我们全面评估了 Bi-Mem 相较于多种领先基线方法的表现，包括：LongContext（不使用记忆的 LLM 骨干模型）、RAG（Lewis 等人，2020年）、Mem0（Chhikara 等人，2025年）、LightMem（Fang 等人，2025年）、A-MEM（Xu 等人，2025年）、SeCom（Pan 等人，2025年），以及 CAM（Li 等人，2025年）。这些基线方法的详细介绍见附录B.1。

实现细节 为确保公平比较，我们为 Bi-Mem 和所有基线使用相同的大语言模型骨干网络和嵌入模型。行。具体而言，我们采用 GPT-4o-mini（Hurst 等人，2024）和 Qwen2.5-14B-Instruct（Bai 等人，2025年）（128K词元上下文窗口）来评估 Bi-Mem 的有效性和泛化能力。对于嵌入模型，我们使用 all-MiniLM-L6-v2（记为 $\phi(\cdot)$，该模型同时支持图边构建（用于事实级记忆）和检索任务。事实单元之间边形成的预设阈值 τ（公式（5））设为 0.2。关于检索器实现，我们采用混合检索方法，将基于余弦相似度的稠密检索器与 BM25 检索器相结合，加权因子为 0.5。初始检索的 k 设为

方法	单跳		多跳		开放域		时间		平均值	
	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
仅事实	49.99	44.66	27.11	20.46	9.43	7.80	51.28	36.60	43.49	36.19
事实+场景	50.12	44.66	32.01	23.56	25.75	22.08	52.35	37.30	45.72	37.81
事实+角色	49.26	43.99	27.91	20.60	9.65	9.21	51.84	37.09	43.38	36.04
Bi-Mem	**53.68**	**47.99**	**39.17**	**31.81**	**30.99**	**26.90**	**54.44**	**41.56**	**49.74**	**42.33**

表2：用于评估分层记忆中事实-场景-人设层级各自贡献的消融研究。“仅事实”表示仅在事实层级构建的记忆；“事实+场景”和“事实+角色”分别表示结合事实层级与场景层级或角色层级信息的分层记忆。

方法	单跳		多跳		开放域		时间		平均值	
	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
自顶向下	13.87	11.85	12.92	9.51	14.47	13.43	16.63	11.58	14.32	11.46
自底向上	42.59	37.50	29.01	20.95	21.26	17.03	47.05	33.26	39.68	32.27
分层	51.26	45.76	32.28	24.58	20.17	17.46	51.42	36.16	45.84	38.06
+场景 _to_Fact	52.97	47.31	34.02	26.81	22.72	19.38	53.48	38.05	47.68	39.83
+事实_to_场景	52.40	46.75	34.16	26.96	25.14	20.51	51.62	36.44	47.16	39.29
Bi-Mem	**53.68**	**47.99**	**39.17**	**31.81**	**30.99**	**26.90**	**54.44**	**41.56**	**49.74**	**42.33**

表3：不同检索策略的消融研究。“自顶向下”和“自底向上”分别表示从顶层或底层检索相关记忆，然后逐层关联。“分层”指同时从全部三个层级检索 top-k 个相关记忆。“+ 场景_到_事实”和“+ 事实_到_场景”则在分层检索的基础上，分别从场景扩展到其子事实，或从事实扩展到其父场景。

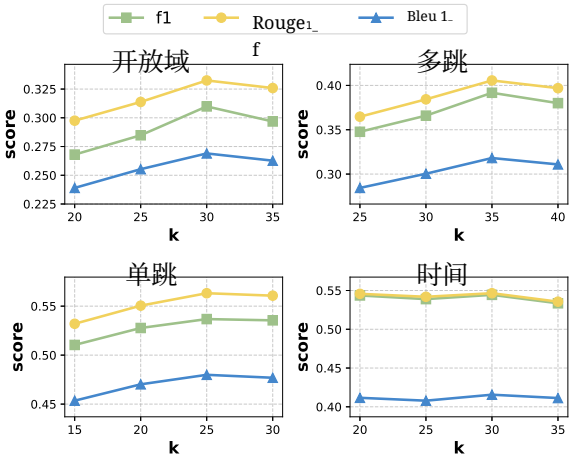

图4：Bi-Mem 对超参数 k 的敏感性分析，该参数表示层级搜索阶段初始检索的记忆单元数量。

公式（9）中的搜索选取 $[15, 20, 30, 35, 40]$，而公式（m）中的联想扩散 10）范围为 $[1, 2, 3]$。最优超参数设置见附录 B.2。为量化记忆的有效性，我们通过比较各方法的问答准确率来进行评估，采用两个指标 F_1（用于平衡答案精确率和召回率）以及 BLEU-1（记为 B_1，用于衡量词语重叠率生成答案与真实答案之间的重叠）。

5.2 主要结果（RQ1）

为回答 Q1，我们在 LoCoMo 基准测试的四类问题上，将 Bi-Mem 与领先的基线方法进行比较。每项实验均使用不同的随机种子至少进行三次，并在表1中报告平均结果。我们可以看到，Bi-Mem 在 GPT-4o-mini 和 Qwen-2.5-14B-Instruct 两种骨干模型上始终优于所有基线方法，这验证了我们提出的双向记忆构建和面向大语言模型个性化的联想检索方法的有效性与泛化能力。值得注意的是，在我们的设置中，检索增强生成（Lewis 等人，2020年）作为扁平事实记忆，仅在单跳问题上表现更优。这表明，纯事实记忆难以处理需要捕捉事实间关系或识别用户高层行为模式的复杂问题，凸显了多粒度记忆结构的必要性。

5.3 记忆构建消融实验（RQ 2)

为回答 Q2，我们开展实验，以验证每种粒度化记忆在

分层结构中的必要性。如表2所示，纯事实记忆构建在除单跳以外的所有问题类型中表现最差，凸显了高层记忆的重要性。得益于角色级校准提供的全局约束，“事实+场景”取得了相近的表现。此外，Bi-Mem（整合全部三个层级）表现最佳，证实了每种粒度层级的必要性。为进一步评估双向构建的有效性，我们针对反思过程校准开展了消融研究。如图3所示，移除反思过程（无校准）会导致所有问题类型的表现显著下降。此外，我们在附录C.2 中给出一个案例，直观地验证反思校准的重要性。

5.4 联想检索消融实验（RQ3）

为评估我们的联想检索机制的有效性，我们将其与表 3 中展示的多种检索策略进行比较。结果表明，单向策略在连接多粒度记忆方面存在明显局限。“自顶向下”策略可能无法检索到问题所涉及的精确细节，而“自底向上”策略则常常受到语义碎片化的影响。“分层”基线同时从所有层级检索排名靠前的k 个候选项，虽然提升了性能，但仍受限于各个检索到的记忆单元之间相互独立。"+ 事实_到_场景" 和 "+场景_到_事实" 所带来的显著性能提升，体现了初始分层搜索后进行扩散激活的强大效果。最终，完整的 Bi-Mem 检索配置取得了最佳结果，证明双向关联能够有效弥合抽象人物画像与具体对话事实之间的差距。

5.5 敏感性分析（RQ4）

为回答 RQ4，我们分析 Bi-Mem 对超参数的敏感性 k （层级搜索阶段最初检索到的记忆单元数量）。 k 在 $[15, 20, 25, 30, 35, 40]$范围内变化。如图4所示，所有任务的答案准确率最初都随着 k 的增加而提升——这表明，较大的初始检索预算能够使我们的联想检索捕获更全面的记忆集合。达到峰值后，性能随即下降，因为检索到过多记忆会引入额外噪声，对相关信息

指标	A-MEM	CAM	Bi-Mem
记忆构建（秒）	3327	443	3448
检索-回答（秒）	7.30	8.86	4.54
总计（秒）	4437	1790	4138
答案质量（F1值）	39.65	39.25	49.73

表4：成本效益分析：我们测量了单个用户的记忆构建时间以及每个问题的平均检索与回答时间，总耗时包括记忆构建和随后回答152个问题的时间。

的提取产生负面影响。这表明，最优的 k 能够有效弥合零散事实与主题场景之间的差距。关于其他超参数的更多敏感性分析（例如， m 公式10中的联想扩散参数）见附录C.1。

5.6 效率（研究问题5）

为评估效率，我们将 Bi-Mem 的记忆构建和问答时间开销与基于记忆的基线方法进行比较（A-Mem（Xu 等人，2025年）、CAM（Li 等人，2025年））。如表4所示，Bi-Mem 的问答效率更高（每个问题4.54秒），显著快于 A-Mem（7.30秒）和 CAM（8.86秒）。这种提速源于我们的联想检索机制——预先建立的层次链接能够比扁平检索或复杂的重新排序更快地识别多粒度证据。尽管其双向归纳—反思过程使构建时间（3,448秒）长于 CAM（443秒），但与 A-Mem 的效率相当（3,327秒）。值得注意的是，这种权衡因准确率的大幅提升而合理，证实了双向分层记忆构建的价值。

6 结论

本文提出 Bi-Mem，这是一种通过双向构建和联想检索机制来确保分层记忆保真度的框架。Bi-Mem 不同于通过单向聚合构建分层记忆的方式，而是采用归纳智能体进行自底向上构建，并采用反思智能体进行自顶向下校准，从而缓解局部记忆与用户整体人设之间的不匹配。为确保连贯回忆，联想检索机制通过扩散激活连接不同粒度的记忆。实验结果表明，Bi-Mem 显著提升了长期个性化对话任务中的问答准确率。

局限性

尽管 Bi-Mem 显著改进了个性化大语言模型的记忆系统，但仍存在若干局限：

1. 对模型推理能力的敏感性：双向记忆构建的保真度本质上取决于底层大语言模型的指令遵循和推理能力。大语言模型指令的差异可能会影响归纳和反思过程的有效性，这表明需要采用更稳健的提示策略。

2. 静态与动态角色建模：尽管我们的框架通过稳定的角色锚点确保一致性，但其主要针对人设相对稳定的用户而设计。未来的工作可以扩展校准机制，以更好地捕捉用户偏好的时间演变和动态角色变化。

3. 通过强化学习进行优化的潜力：当前框架侧重于记忆构建与检索的结构设计。尽管这种设计有效，但仍有机会通过强化学习（RL）或偏好优化进一步改进特定智能体的行为，例如反思智能体的校准策略，从而更好地适应多样化的用户交互风格。

伦理声明。

本研究旨在提升层次化记忆系统的保真度，将记忆构建过程表述为局部场景与全局用户画像之间的双向对齐。生成式人工智能用于辅助编程和写作。我们预计本研究不会对社会造成任何直接、即时或负面影响。

可复现性声明。

本研究中的所有结果均可复现。我们已在附录中讨论最优超参数以及设备和软件环境的详细信息B.2。

参考文献

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wen- bin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。2025年。Qwen2. 5-vl 技术报告。*arXiv* 预印本 *arXiv:2502.13923*。

Ding Chen、Simin Niu、Kehang Li、Peng Liu、Xiangping Zheng、Bo Tang、Xinchi Li、Feiyu Xiong 和 Zhiyu Li。2025年。Halumem：评估智能体记忆系统中的幻觉。*CoRR*，abs/2511.03506。

Jin Chen、Zheng Liu、Xu Huang、Chenwang Wu、Qi Liu、Gangwei Jiang、Yuanhao Pu、Yuxuan Lei、Xiaolong Chen、Xingmei Wang、Kai Zheng、Defu Lian 和 Enhong Chen。2024。当大语言模型遇上个性化：挑战与机遇展望。万维网（*WWW*），27(4):42。Prateek Chhikara、Dev Khant、Saket Aryan、Taranjeet Singh 和 Deshraj Yadav。2025年。Mem0：构建具备可扩展长期记忆的生产级 AI 智能体。*CoRR*，abs/2504.19413。Jizhan Fang、Xinle Deng、Haoming Xu、Ziyan Jiang、Yuqi Tang、Ziwen Xu、Shumin Deng、Yunzhi Yao、Mengru Wang、Shuofei Qiao、Huajun Chen 和 Ningyu Zhang。2025年。Lightmem：轻量高效的记忆增强生成。*CoRR*，abs/2510.18866。

Bernal Jimenez Gutierrez、Yiheng Shu、Yu Gu、Michihiro Yasunaga 和 Yu Su。2024。Hipporag：受神经生物学启发的面向大语言模型的长期记忆。收录于*NeurIPS*。

YuyangHu、Shichun Liu、YanweiYue、Guibin Zhang、Boyang Liu、Fangyi Zhu、Jiahang Lin、Honglin Guo、Shihan Dou、Zhiheng Xi，等 1 位作者。2025年。AI智能体时代的记忆。*arXiv* 预印本 *arXiv:2512.13564*。

Lei Huang、Weijiang Yu、Weitao Ma、Weihong Zhong、Zhangyin Feng、Haotian Wang、Qianglong Chen、WeihuaPeng、Xiaocheng Feng、Bing Qin 和 Ting Liu。2025年。大型语言模型中的幻觉综述：原理、分类体系、挑战与开放问题。*ACM* 信息系统汇刊，43(2):42:1–42:55。

Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、AidanClark、AJOstrow、Akila Welihinda、Alan Hayes、Alec Radford，等 1 位作者。2024。GPT-4o 系统卡片。*arXiv* 预印本 *arXiv:2410.21276*。

Jihyoung Jang、Minseong Boo 和 Hyounghun Kim。2023。Conversation chronicles：探索多会话对话中多样的时间与关系动态。载于*EMNLP*，第 13584–13606 页。计算语言学协会。

Jiho Kim、Yeonsu Kwon、Yohan Jo 和 Edward Choi。2023。KG-GPT：一种利用大语言模型对知识图谱进行推理的通用框架。载于*EMNLP(Findings)*，第 9410–9421 页。计算语言学协会。

Andrea Lancichinetti 和 Santo Fortunato。2012。复杂网络中的共识聚类。*Scientificreports*，2(1):336。

Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel、Sebastian Riedel 和 Douwe Kiela。2020年。面向知识密集型NLP任务的检索增强生成。在*NeurIPS*。

Rui Li、Zeyu Zhang、Xiaohe Bo、Zihang Tian、Xu Chen、Quanyu Dai、Zhenhua Dong 和 Ruiming Tang。2025年。CAM：基于大语言模型的阅读理解中智能体记忆的建构主义视角。在第三十九届神经信息处理系统年会。

Adyasha Maharana、Dong-Ho Lee、Sergey Tulyakov、Mohit Bansal、Francesco Barbieri 和 Yuwei Fang。2024年。评估大语言模型智能体的超长期对话记忆。发表于 *ACL*（*1*），第 13851–13870 页。计算语言学协会。

Charles Packer、Vivian Fang、Shishir G. Patil、Kevin Lin、Sarah Wooders 和 Joseph E. Gonzalez。2023年。Memgpt：迈向将大语言模型用作操作系统。*CoRR*，abs/2310.08560。

Zhuoshi Pan、Qianhui Wu、Huiqiang Jiang、Xufang Luo、Hao Cheng、Dongsheng Li、Yuqing Yang、Chin-Yew Lin、H. Vicky Zhao、Lili Qiu 和 Jianfeng Gao。2025年。SECOM：面向个性化对话智能体的记忆构建与检索。发表于第十三届国际学习表征会议。

Anup Pattnaik、Cijo George、Rishabh Kumar Tripathi、Sasanka Vutla 和 Jithendra Vepa。2024。通过大语言模型引导的多视角聚类表示改进层次文本聚类。载于*EMNLP*（工业赛道），第 719–727 页。计算语言学协会。

Alireza Rezazadeh、Zichao Li、Wei Wei 和 Yujia Bao。2025。从孤立的对话到层次化模式：面向大语言模型的动态树记忆表示。载于*ICLR*。OpenReview.net。

Zhaoxuan Tan、Zheyuan Liu 和 Meng Jiang。2024a。个性化组件：通过协作打造高效的个性化大语言模型。载于*EMNLP*，第6459–6475页。计算语言学协会。

Zhaoxuan Tan、Qingkai Zeng、Yijun Tian、Zheyuan Liu、Bing Yin 和 Meng Jiang。2024b。通过个性化参数高效微调，让大语言模型惠及大众。载于*EMNLP*，第 6476–6491 页。计算语言学协会。

Johan Ugander 和 Lars Backstrom。2013年。用于划分大规模图的平衡标签传播。载于*WSDM*，第 507–516 页。ACM。

Chengbing Wang、Yang Zhang、Fengbin Zhu、Jizhi Zhang、Tianhao Shi 和 Fuli Feng。2025a。利用记忆检索增强基于大语言模型的生成式推荐。载于*WWW*（伴随卷），第 1346–1350 页。ACM。

Zehong Wang、Junlin Wu、Zhaoxuan Tan、Bolian Li、Xianrui Zhong、Zheli Liu 和 Qingkai Zeng。2025b。《从个人到集体：论局部记忆与全局记忆在大语言模型个性化中的作用》。*CoRR*，abs/2509.23767。

Jing Xu、Arthur Szlam 和 Jason Weston。2022。《超越金鱼记忆：开放域长时对话》。发表于*ACL (1)*，第 5180–5197 页。计算语言学协会。

WujiangXu、Zujie Liang、KaiMei、HangGao、Juntao Tan 和 Yongfeng Zhang。2025年。A-MEM：面向大语言模型智能体的智能体记忆。载于第三十九届神经信息处理系统年会。

BY Yan、Chaofan Li、Hongjin Qian、Shuqi Lu 和 Zheng Liu。2025年。通过深度研究实现通用智能体记忆。*arXiv* 预印本 *arXiv:2511.18423*。

Wan Zhang 和 Jing Zhang。2025。检索增强大语言模型的幻觉缓解：综述。*Mathematics*，13(5):856。

Yujie Zhang、Weikang Yuan 和 Zhuoren Jiang。2025a。连接直觉联想与审慎回忆：通过图结构长期记忆赋能大语言模型个人助理。发表于 *ACL*（*Findings*），第 17533–17547 页。计算语言学协会。

Zeyu Zhang、Xiaohe Bo、Chen Ma、Rui Li、Xu Chen、Quanyu Dai、Jieming Zhu、Zhenhua Dong 和 Ji-Rong Wen。2024。基于大语言模型的智能体记忆机制综述。*arXiv* 预印本 *arXiv:2404.13501*。

Zhehao Zhang、Ryan A. Rossi、Branislav Kveton、Yijia Shao、Diyi Yang、Hamed Zamani、Franck Dernoncourt、Joe Barrow、Tong Yu、Sungchul Kim、Ruiyi Zhang、Jiuxiang Gu、Tyler Derr、Hongjie Chen、Junda Wu、Xiang Chen、Zichao Wang、Subrata Mitra、Nedim Lipka 等另外两位作者。2025b。大语言模型个性化：综述。*Trans. Mach.Learn. Res.*，2025年。

Xinping Zhao、Yan Zhong、Zetian Sun、Xinshuo Hu、Zhenyu Liu、Dongfang Li、Baotian Hu 和 Min Zhang。2025年。Funnelrag：一种用于检索增强生成的由粗到细渐进式检索范式。载于 *NAACL*（*Findings*），第 3029–3046 页。计算语言学协会。

Wanjun Zhong、Lianghong Guo、Qiqi Gao、He Ye 和 Yanlin Wang。2024。《Memorybank：通过长期记忆增强大语言模型》。载于*AAAI*，第 19724–19731 页。AAAI出版社。

算法 1 双向分层记忆构建 输入：对话历史 $\mathcal{C}$，相似度阈值 τ输出：校准后的分层记忆 $\mathcal{M} = (\mathcal{F}, \mathcal{S}', \mathcal{P})$ 1：// 阶段 1：归纳过程（自底向上）2： $\mathcal{F} \leftarrow$ 事实提取 $(\mathcal{C})$，通过算子 $\mathcal{E}$3： $G \leftarrow$ 构建事实图$(\mathcal{F}, \tau)$ 4： $\{\bar{\mathcal{F}}_j\}_{j=1}^{J} \leftarrow \text{LPA}(G)$ ▷基于图的聚类 5： $\mathcal{S} \leftarrow$ 场景聚合$(\{\mathcal{F}_j\})$，通过算子 $\mathcal{T}$6： $\mathcal{P} \leftarrow$ 提炼人物画像$(\mathcal{S})$，通过算子 $\mathcal{D}$ ▷五维人物画像 7：// 阶段 2：反思过程（自顶向下） 8： $\mathcal{S}' \leftarrow \emptyset$9：对每个场景 $s_j \in \mathcal{S}$ 执行 10：如果 s_j 与 $\mathcal{P}$ 不一致，则 11：$\Delta s_j \leftarrow$ 校准$(s_j, \mathcal{P})$，通过算子 $\mathcal{R}$12： $s'_j \leftarrow (s_j, \Delta s_j)$ ▷ 基于条件的校准 13：否则 14： $s'_j \leftarrow (s_j, \emptyset)$ 15：结束判断 16： $\mathcal{S}' \leftarrow \mathcal{S}' \cup \{s'_j\}$17：结束循环 18：返回 $\mathcal{M} = \langle \mathcal{F}, \mathcal{S}', \mathcal{P} \rangle$

算法 2联想检索机制

输入： 分层记忆 $\mathcal{M} = \langle \mathcal{F}、\mathcal{S}、\mathcal{P} \rangle$，查询 q^*，参数 k， m
输出： 最终检索记忆集 $\mathcal{M}_{ret}$
1: **// 阶段 1：初始分层搜索**
2: 为所有对象计算 $a_0(x) = \text{sim}(\phi(q^*), \phi(x))$ $x \in \mathcal{F} \cup \mathcal{S} \cup \mathcal{P}$
3: $\mathcal{M}_{ret} \leftarrow$前$k(\{x \mid a_0(x)\})$ ▷ 选择初始跨层种子
4: **// 阶段 2：联想扩散**
5: **遍历**每个元素 $u \in \mathcal{M}_{ret}$ **do**
6: **如果** $u \in \mathcal{F}$ **则**
7: $s \leftarrow$ 查找父节点(u) ▷ 自底向上：关联父场景
8: $\mathcal{M}_{ret} \leftarrow \mathcal{M}_{ret} \cup \{s\}$
9: **否则如果** $u \in \mathcal{S}$ **则**
10: $\mathcal{F}_u \leftarrow$事实簇(u) ▷ 自顶向下：关联组成事实
11: $\mathcal{F}_{top_m} \leftarrow$ 自顶向下-m($\in \bar{\mathcal{F}}_u \mid$ 相似度$(u, f)\}$) $\{f$
12: $\mathcal{M}_{ret} \leftarrow \mathcal{M}_{ret} \cup \mathcal{F}_{top_m}$
13: 结束条件判断
14: **结束循环**
15: **返回**$\mathcal{M}_{ret}$

A 算法

本文列出了 Bi-Mem 的双向分层记忆构建流程算法，见算法 1；联想检索机制见算法 2。

B 详细实验设置

B.1 基线方法详情

这里，我们详细介绍基线方法。
• LongContext：该方法直接使用大语言模型（不采用记忆机制）执行 QA 任务，并将历史对话和查询整合到提示词中。
• 检索增强生成（Lewis 等人，2020年），对话被拆分为每块 2,048 个词元，并按照 GAM（Yan 等人，2025年）的设置，使用检索到的排名前 5 的文本块执行问答任务。
• A-MEM（Xu 等人，2025年）是一种智能体框架，具备结构化记忆、动态链接和语义匹配驱动的演化机制。
• LightMem（Fang 等人，2025年）是一个轻量级系统，采用三阶段记忆（感官/短期/长期）。
• Mem0（Chhikara 等人，2025年）通过添加/更新/删除/无操作操作提取记忆，并通过向量相似度进行检索。
• SeCom（Pan 等人，2025年）从主题连贯的片段构建去噪记忆，并在片段级别进行检索。
• CAM（Li 等人，2025年）是一种分层记忆框架，采用增量式重叠聚类算法进行构建，并采用剪枝与生长策略进行检索。

B.2 更多实现细节

为实现第 5.4 节中的“自顶向下”和“自底向上”消融变体，我们参照主流配置，分别为事实、场景和角色层级将选定的相关记忆单元数量设为 1、15 和 25。为每位用户构建分层记忆所消耗的词元总数约为 200,000–300,000，而为每位用户回答后续问题所消耗的词元数则为 100,000 至 20,000。嵌入模型“all-MiniLM-L6-v2”部署在 A800 上，既支持公式（5）中的边构建，也支持公式（9）、（10）中的检索过程。此外，超参数的详细设置列于表 5。

	单跳	多跳	时序开放域	
k	35	25	30	25
m	3	3	3	3
τ	0.2	0.2	0.2	0.2
α	0.5	0.5	0.5	0.5
L_w	128k	128k	128k	128k

表 5：不同问题类型的超参数设置。α 表示检索记忆时稠密检索器与 BM25检索器之间的加权因子。L_w 表示大语言模型骨干网络的上下文窗口。

为便于复现，数据和代码可在 https://github.com/tohsaka-sb/Bi-Mem获取。

C 更多实验结果

C.1 更多敏感性分析

为了验证 Bi-Mem 对超参数 m（表示联想扩散激活期间从检索场景中激活的事实数量）的敏感性，我们在表 6 中展示了实验结果。我们观察到，性能在$m = 3$时达到峰值，平均值 F_1 为 49.74%。而 $m = 1$ 和 $m = 5$ 都会导致性能略微下降。具体而言，$m = 1$（较小的值）无法为复杂任务提供充足的事实记忆，而将 m 增至 5，则可能引入冗余噪声，分散大语言模型对用户核心意图的注意力。这表明，平衡的联想范围能够有效检索出最相关的记忆，从而准确回答问题。

C.2 案例研究

为生动说明分层记忆双向构建的重要性，我们展示一个具有代表性的案例研究，重点关注归纳式记忆构建和反思校准。如图 5 （原始对话）和图 6 （事实级记忆内容）所示，从原始对话中提取的事实信息可能包含噪声（例如，“照顾好自己并不总是那么容易”）和幻觉（例如，“我正在寻求帮助”）。这类不准确之处可能会扭曲聚合后的场景级记忆（图 7），导致其与用户的全局人格级记忆不一致——在全局人格级记忆中，Caroline 被定义为富有同理心，并有勇气拥抱真实自我（图 8）。通过反思校准，我们将人物设定作为全局约束注入场景级记忆，最终生成校准后的场景级记忆（图 9），其中融合了内在的在人格特质（例如，同理心和拥抱真实自我的勇气），并确保场景级记忆与核心人格严格一致。

D 提示模板

下面给出为双向分层记忆构建的各个对应步骤设计的提示。

事实级记忆提取 $\mathcal{E}$

为以下交互内容 c_i 生成结构化事实 con_i，请执行以下操作：1. 识别最显著的关键词（重点关注名词、动词和关键概念）2. 提取核心主题和上下文要素 3. 创建相关的类别标签 请将回复格式设为 JSON 对象：{ - keywords: [“关键词1”、“关键词2”.....],- context: “用一句话概述交互内容”，- tags: [“标签1”、“标签2”.....]}

场景级记忆聚合 $\mathcal{A}$

你是一名专门进行事实理解的场景综合器。任务：将一组相关事实记忆总结为连贯的“场景记忆”。输入事实记忆：facts_content 说明：1. 找出这些事实共同的核心主题。2. 生成一段描述性摘要，呈现对话事实的发展过程。3. 提取关键实体和主题。请将回复格式化为 JSON 对象：{ - scene_memory: “全面的场景摘要”, - keywords: [“关键词1”, “关键词2”, ...],- tags: [“标签1”, “标签2”, ...]}

方法	单跳		多跳		开放域		时间		平均值	
	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
m=1	52.79	46.95	34.56	27.83	25.22	23.01	51.12	**44.77**	47.12	41.25
m=5	53.58	47.82	37.52	29.98	29.91	26.03	53.24	41.03	49.06	41.73
Bi-Mem$(m = 3)$	**53.68**	**47.99**	**39.17**	**31.81**	**30.99**	**26.90**	**54.44**	41.56	**49.74**	**42.33**

表 6：超参数 m 的敏感性实验。该超参数表示在进行联想扩散激活时，Scene_to_Fact 中扩展事实的数量。

人格级记忆蒸馏 $\mathcal{D}$

你是一名专门从事心理与行为分析的人物设定合成师。任务：根据提供的场景记忆，创建一份全面的用户画像。输入场景：所有_场景_内容。说明：1. 深入分析这些场景，寻找行为、情绪和选择方面的模式。2. 针对以下每个维度撰写一段详细的分析（5 至 10 句话）。不要写得过于简略。3. 使用场景中的具体示例来支持分析。请将回复格式化为 JSON 对象：{ - basic_info：“详细背景.....”， - 兴趣：“爱好的完整清单，以及他们如何参与其中.....”， - personality：“深入的个性分析.....”， - values：“ 核心信念和动机 .....”， - relationships：“详细的社交互动模式.....”

}

反思校准。

你是一名场景记忆校准器。你的目标是使给定场景与用户画像相契合。人物设定级记忆：user_persona 场景级记忆：current_scene 指令：1. 阅读人物设定级记忆，了解用户的主要兴趣、价值观和特质。2. 检查当前场景级记忆。它是否遗漏了场景中可能存在的、与用户画像相关的具体联系？3. 如果是，请添加一个补偿条件，附加到原场景之后。此补充内容应明确说明场景如何与人物设定相契合（例如：“这与她对.....的兴趣相符。”）。4. 重要：不要改写现有摘要。只生成要添加的文本。5. 如果当前摘要已经很完善，则为“添加条件”返回空字符串。请按以下格式返回 JSON 对象：{ - needs_calibration: true/false, - added condition: “要作为条件添加的文本（若无则为空字符串）”, - reason: 决定是否进行校准的原因。” }

原始对话：

1. 我完全同意，Melanie。照顾好自己非常重要——即使这并不总是容易。很高兴你把自我照护放在优先位置。2. 我分享了自己的经历、曾经遇到的困难，以及出柜后成长了多少。3. 我曾饱受心理健康问题困扰，而我获得的支持真的很有帮助。4. 最近，我一直在考虑从事心理咨询和心理健康相关工作。我想帮助那些经历过和我相同事情的人。

图5：一组原始对话中的案例研究。

事实级记忆内容：

1. 照顾好自己并不总是容易的。（噪声）2. 我分享了自己的经历、曾经的挣扎，以及走出来的艰难过程。3. 我曾面临心理健康方面的困扰，目前正在寻求帮助。（幻觉）4. 我正在考虑从事心理咨询／心理健康相关职业，帮助有类似经历的人。

图6：对话聚类中提取的事实级记忆内容案例研究，其中存在噪声和幻觉。

汇总后的场景级记忆：

由于自我照护方面的挑战和心理健康困扰，她经常分享自己的经历，在这一过程中寻求支持，并希望从事心理健康／心理咨询领域的职业，帮助有类似经历的人。

图7：关于从事实簇汇总而来的场景级记忆的案例研究，该记忆与全局用户画像不一致。

相关人物设定级记忆：

Caroline以同理心、创造力和韧性为特质，并勇于接纳真实自我。她以温暖亲切的举止、对家人坚定不移的忠诚以及极强的支持他人之心而著称。

图8：与该场景相关的全局人物级记忆案例研究。

校准后的场景级记忆：

由于面临自我照护方面的挑战和心理健康困扰，她经常分享自己的个人历程，在此过程中寻求支持，并希望从事心理健康／心理咨询领域的工作，帮助有类似经历的人。**事实上，她的本性体现为富有同理心、乐于支持他人，并有勇气接纳真实自我。**

Fig图9：关于注入人物设定约束的校准后的场景级记忆的案例研究 t.