LLM 智能体的自适应记忆准入控制

Guilin Zhang* **Wei Jiang 谢家山王 Aisha Behr Kai Zhao***
Jeffrey Friedman Xu Chu Amine Anoun Workday AI

arXiv:2603.04549v1 [cs.AI] 4 Mar 2026

摘要

基于大语言模型的智能体日益依赖长期记忆来支持跨多个会话的推理与交互，但当前系统对保留哪些信息几乎无法进行控制。在实践中，智能体要么积累大量对话内容（其中包括幻觉信息或过时事实），要么依赖不透明、完全由大语言模型驱动的记忆策略，这类策略成本高昂且难以审计。因此，记忆准入仍是智能体架构中定义不清、控制薄弱的组件。为弥补这一不足，我们提出自适应记忆准入控制（A-MAC），这是一种将记忆准入视为结构化决策问题的框架。A-MAC 将记忆价值分解为五个互补且可解释的因素：未来效用、事实置信度、语义新颖性、时间新近性和内容类型先验。该框架将轻量级的基于规则的特征提取与单次大语言模型辅助的效用评估相结合，并通过交叉验证优化学习领域自适应准入策略。这一设计能够对长期记忆进行透明且高效的控制。在LoCoMo基准上的实验表明，与最先进的原生大语言模型记忆系统相比，A-MAC 实现了更优的精确率–召回率权衡，将 F1 值提升至 0.583，同时将延迟降低了 31%。消融结果表明，内容类型先验是实现可靠记忆准入时影响最大的因素。这些发现说明，明确且可解释的准入控制是基于大语言模型的智能体实现可扩展、可靠记忆的关键设计原则。代码见 https://github.com/GuilinDev/Adaptive_Memory_Admission_Control_LLM_Agents。

1 引言

大语言模型（LLM）智能体在多轮交互、推理和工具使用方面展现出了强大的能力（Park et al., 2023; Wang et al., 2023; Shinn et al., 2023; Wei et al., 2022; Yao et al., 2022; Schick et al., 2023）。随着这些智能体越来越多地执行长期交互，长期记忆成为一项核心架构组件，有助于在上下文窗口之外保留信息，并支持长期保持连贯行为（Zhang等，2024）。然而，确定哪些信息应保留在长期记忆中仍是一项关键挑战。不加甄别地存储会导致记忆存储库臃肿，并增加检索延迟（Lewis等人，2020; Guti´errez et al., 2024）；而保留幻觉信息或过时信息，则可能使错误传播到未来交互中（Huang等人，2023）。相反，过于保守的准入策略可能会丢弃任务延续和长时程推理所必需的信息。

现有的记忆管理方法大致分为两类，各自都存在显著局限。基于启发式的方法，如 MemGPT（Packer 等人，2023年）和 MemoryBank（Zhong等人，2024年），依赖人工设计的评分函数，综合考虑时效性、相关性和重要性等因素。这些方法虽然计算效率高，却缺乏可靠的机制来防止幻觉内容进入记忆，也难以处理细微的准入决策。相比之下，原生 LLM 方法，如 A-mem（Xu等人，2025年）和 Mem0（Chhikara等人，2025年），将记忆准入完全交由语言模型处理，生成结构化记忆

*通讯作者：Guilin Zhang 和 Kai Zhao。

属性。尽管这些方法在召回方面有效，但会带来大量计算开销，且可解释性有限，因此记忆策略难以审计或调试。值得注意的是，尽管幻觉已被视为智能体可靠性面临的根本挑战（Huang等人，2023年；Bai等人，2022），这两类方法都未明确将幻觉作为首要关注的问题。

为解决这些局限性，我们提出了**自适应记忆准入控制（A-MAC）**，这是一个将记忆准入视为结构化决策问题（而非生成过程的隐式副产品）的框架。A-MAC 不依赖临时启发式规则或不透明的大语言模型判断，而是在候选记忆进入长期存储前对其进行明确评估。这一形式化方法将记忆准入提升为智能体架构中的一等控制机制，从而能够在记忆覆盖率、可靠性和效率之间进行审慎权衡。

A-MAC 将记忆价值分解为五个互补且可解释的维度：未来效用、事实置信度、语义新颖性、时间新近性和内容类型先验。这些维度刻画了记忆质量的不同方面，而现有方法往往将其混为一谈或不予明示。效用用于估计某条记忆与未来任务相关的潜力，置信度则评估该信息是否得到先前对话证据的支持，直接减轻幻觉传播。新颖性可避免存储冗余内容，时效性考量时间衰减，而内容类型先验则编码领域知识，判断哪些类别的信息值得长期持久保存。

该框架将轻量级的基于规则的特征提取与一次大语言模型辅助的效用评估相结合，在计算效率和语义表达能力之间实现了良好平衡。通过交叉验证优化学习准入策略，使 A-MAC 无需手动调优即可适应不同的对话领域。通过明确建模可靠性、冗余性、时间相关性和内容持久性，A-MAC 能够以透明、数据驱动的方式控制哪些信息会被纳入长期记忆，为完全原生大语言模型记忆系统提供了一种可扩展的替代方案。

本文的贡献如下：

1. 我们指出，记忆准入是基于大语言模型的智能体中一个关键但定义尚不充分的控制问题，并分析了现有启发式方法和完全由大语言模型驱动的方法的局限性。2. 我们提出自适应记忆准入控制（A-MAC），这是一个可解释的框架，通过五个互补维度评估候选记忆，共同捕捉其价值、可靠性和持久性。3. 我们提出一种高效的混合设计，将基于规则的特征计算与最少的大语言模型推理相结合，在可解释性、准确性和计算效率之间实现良好平衡。4. 我们通过实验证明，A-MAC 在 LoCoMo基准上实现了更优的精确率–召回率权衡，同时与最先进的原生大语言模型记忆系统相比降低了延迟；消融研究则阐明了各项准入因素的作用。

2 相关工作

我们的工作与智能体记忆架构、记忆准入策略以及利用外部记忆提升基于大语言模型的系统可靠性的既有研究相关。近期的综述全面概述了大语言模型智能体的记忆机制，并指出记忆控制是一项亟待解决的关键挑战（Zhang等，2024）。我们根据记忆准入决策的设计方式，将相关工作分为两类。

2.1 启发式与层级式记忆架构

早期的大语言模型智能体记忆研究采用受操作系统启发的分层设计，将快速工作记忆与较慢的长期存储分开。MemGPT（Packer 等人，2023年）引入了双层架构，并采用基于时效性和大语言模型判定的重要性的基于页面的驱逐策略

。尽管计算效率较高，MemGPT 却没有提供在记忆准入之前验证事实正确性的明确机制。Memory OS（Kang等人，2025年）提出了类似的分层存储单元，并报告了在 LoCoMo基准上的性能提升，但仍主要依赖时间启发式。ChatDB（Hu等人，2023年）通过将SQL数据库用作符号记忆来增强大语言模型，重点关注存储与检索，而非准入可靠性。MemoryBank（Zhong等人，2024年）基于艾宾浩斯遗忘曲线，将时效性、相关性和重要性结合起来，通过手工设计的线性评分函数扩展了分层记忆。与简单的基于近期性的策略相比，该方法虽效果显著，但其固定加权方案无法适应不同领域，也未明确处理幻觉内容或不可靠内容。

这些方法强调架构分离和高效存储，但在记忆准入方面依赖静态启发式方法或人工调优的规则。相比之下，A-MAC 引入了一个明确的数据驱动准入控制层，依据多个可解释标准评估记忆候选项，从而能够自适应且透明地控制哪些信息会被提升至长期记忆。

2.2 原生大语言模型的记忆管理与可靠性

较新的方法将记忆准入决策直接交由大型语言模型处理。A-mem（Xu等人，2025年）通过大语言模型提示生成结构化记忆属性，并遵循卡片盒笔记法进行基于相似度的匹配；尽管需要多次调用大语言模型，但在 LoCoMo 上的表现较 MemoryBank 有显著提升。Hindsight（Latimer等人，2025年）将智能体记忆组织成逻辑网络，并进行回顾性评分以保留有用经验；Mem0（Chhikara等人，2025年）则着眼于生产部署，通过动态提取显著信息来降低延迟。关于智能体长期记忆的更广泛研究强调，记忆是自进化人工智能系统的基础（Jiang等人，2024年）；近期基准测试则指出，长时程推理仍存在持续的局限性（Liu等人，2023年）。

与此同时，检索增强生成方法旨在通过以外部证据为生成内容提供依据来提升可靠性。RAG（Lewis等人，2020年）、稠密段落检索（Karpukhin等人，2020年）和 Self-RAG（Asai等人，2023年）侧重于检索质量，而 RAPTOR（Sarthi等人，2024年）、GraphRAG（Edge等人，2024年）和 HippoRAG（Guti´errez等人，2024年）等分层及基于图的方法扩展则进一步提升了多跳推理能力。然而，这些方法主要解决检索时的依据性问题，而非对准入时哪些信息存入记忆进行控制。近期综述指出，幻觉是智能体可靠性面临的一项根本挑战（Huang等人，2023年；Bai等人，2022），但大多数原生大语言模型记忆系统都缺乏明确的防护措施，以避免将缺乏依据的内容纳入记忆。

完全原生于大语言模型的方法虽然提升了召回率，但计算成本高且可解释性有限；基于检索的方法则侧重于为生成提供依据，而非控制记忆增长。A-MAC 通过结合选择性的大语言模型使用、基于规则的可靠性检查和学习得到的权重，弥合了这一差距，使信息进入长期存储时能够实现高效、可解释且具备幻觉感知能力的记忆准入。

3 方法

A-MAC 为基于大语言模型的智能体长期记忆引入了一个显式的准入控制层。我们不再将记忆保留视为生成过程的隐式副产品，而是将准入建模为一个结构化决策问题：为每条候选记忆分配可解释的价值信号，只有通过学习得到的决策规则后才予以接受。图 1 总结了整体工作流程。

3.1 将记忆准入视为决策问题

给定多轮对话历史 $\mathcal{H} = \{t_1, t_2, \ldots, t_k\}$ 和现有记忆库 $\mathcal{M}$，A-MAC 从对话中提取候选记忆 $\{m_1, m_2, \ldots, m_n\}$。每个候选记忆 m 都表示一条可能有助于未来交互的语义原子信息。对于每个候选记忆，智能体必须决定是将其作为新记忆纳入，更新其所取代的现有记忆，还是将其作为冗余或不可靠信息予以拒绝（图 1）。

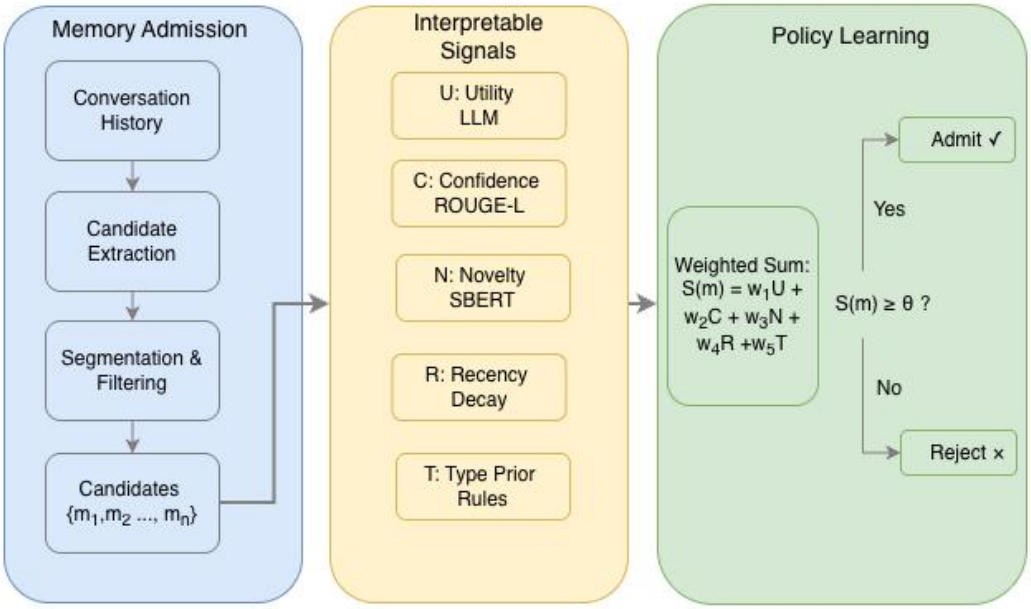

图 1：A-MAC 概览。系统从对话历史中提取候选记忆，并通过五种互补信号进行评估，分别捕捉有用性、可靠性、冗余性、时间相关性和持久性。学习得到的线性准入策略将这些信号汇总为一个分数$\mathcal{S}(m)$，并应用阈值 θ 来决定是接纳、更新还是拒绝候选记忆。

为支持下游评分的一致性，我们对对话进行轻量级规范化处理：将轮次拆分为原子信息单元（因为单个轮次可能包含多个不同的事实）；解析时间表达式和共指，使候选记忆内容完整、自成一体；过滤问候语和致谢语等低价值内容。此规范化步骤可降低噪声，并确保后续价值信号针对格式良好的记忆候选项进行评估。

我们将准入过程表述为一个标量评分问题。每个候选记忆 m 都会获得一个综合分数，该分数由五个可解释的特征函数计算得出：

$$\mathcal{S}(m) = w_1 \cdot \mathcal{U}(m) + w_2 \cdot \mathcal{C}(m) + w_3 \cdot \mathcal{N}(m) + w_4 \cdot \mathcal{R}(m) + w_5 \cdot \mathcal{T}(m), \tag{1}$$

其中 $\mathcal{U}, \mathcal{C}, \mathcal{N}, \mathcal{R}, \mathcal{T}$是将候选记忆映射到 $[0,1]$,的特征函数，分别表示效用、置信度、新颖性、时效性和类型先验；权重满足 $w_i \geq 0$和$\sum_{i=1}^{5} w_i = 1$。若候选记忆满足 $\mathcal{S}(m) \geq \theta$，则予以接纳，其中阈值 θ由学习得到。

3.2 可解释的记忆价值信号

A-MAC 计算五种互补信号，以捕捉长期记忆价值和可靠性的不同方面。其设计原则是将大语言模型推理用于难以通过确定性方法计算的语义判断，同时使用高效且易于审计的规则来实现其余信号。

效用 $(\mathcal{U})$。 $\mathcal{U}(m)$用于估计候选记忆在未来交互中发挥作用的可能性。由于这需要对意图和预期相关性进行语义理解，我们通过一次大语言模型调用来计算效用，评估相关信息是否可执行、是否支持可能的后续问题，或是否体现了持久的用户约束和偏好。我们采用零温度以确保输出具有确定性，并缓存重复候选项的评分，以降低 API 成本。

置信度（$\mathcal{C}$）。$\mathcal{C}(m)$衡量候选记忆是否得到对话中证据的支持，从而直接缓解幻觉传播。我们从先前对话轮次中找出支持片段，并

算法 1 A-MAC 记忆准入

输入： 候选记忆 m、记忆 $\mathcal{M}$、权重 $\boldsymbol{\omega}$、阈值 θ
确保： 已更新的记忆库 $\mathcal{M}'$

$\mathcal{M}' \leftarrow \mathcal{M}$
并行计算 $\mathcal{U}(m)$ $\mathcal{C}(m)$ $\mathcal{N}(m)$ $\mathcal{R}(m)$ $\mathcal{T}(m)$
$\mathcal{S}(m) \leftarrow \boldsymbol{\omega} \cdot [\mathcal{U}, \mathcal{C}, \mathcal{N}, \mathcal{R}, \mathcal{T}]^\top$
如果 $\mathcal{S}(m) \geq \theta$ **则**
  $m_{\text{conflict}} \leftarrow$ 冲突查找$(m, \mathcal{M})$
  如果 $m_{\text{conflict}} \neq \emptyset$ **且** $\mathcal{S}(m) > \mathcal{S}(m_{\text{conflict}})$ **则**
    $\mathcal{M}' \leftarrow (\mathcal{M}' \setminus \{m_{\text{conflict}}\}) \cup \{$合并$(m, m_{\text{conflict}})\}$
  else
    $\mathcal{M}' \leftarrow \mathcal{M}' \cup \{m\}$
  结束如果
结束 if
返回 $\mathcal{M}'$

使用 ROUGE-L 计算对齐度:

$$\mathcal{C}(m) = \max_{s \in \text{Support}(m)} \text{ROUGE-L}(m, s), \tag{2}$$

其中，Support(m) 表示 $\mathcal{H}$ 中可能包含 m 证据的轮次。ROUGE-L 的最长公共子序列基础能够奖励有依据的重叠，同时惩罚缺乏文本支持的虚构细节。

新颖性 ($\mathcal{N}$)。 $\mathcal{N}(m)$ 通过衡量候选记忆与当前记忆库的差异程度，抑制冗余存储。使用句子嵌入 $\phi(\cdot)$（Sentence-BERT），我们定义:

$$\mathcal{N}(m) = 1 - \max_{m' \in \mathcal{M}} \cos(\phi(m), \phi(m')). \tag{3}$$

我们预先计算现有记忆的嵌入，每个新候选记忆只需计算一次嵌入。

时效性 ($\mathcal{R}$)。 $\mathcal{R}(m)$ 体现了信息价值随时间衰减的特性。我们采用指数衰减 :

$$\mathcal{R}(m) = \exp(-\lambda \cdot \tau(m)), \tag{4}$$

其中，τ（m）表示自候选记忆被提及以来经过的时间，λ 控制衰减速率。我们设定 $\lambda = 0.01$ 每小时，对应的半衰期约为 69 小时。

类型先验（$\mathcal{T}$）。 $\mathcal{T}$（m）编码了不同内容类型的持久性偏好。通过结合词性线索进行基于规则的模式匹配，我们为稳定信息（例如偏好或身份陈述）分配较高的先验权重，为瞬态状态分配较低的先验权重。

3.3 策略学习与准入规则

给定带标签的训练对话及其真实准入决策，我们通过 5 折交叉验证最大化 F1 分数，从而学习权重向量 $\boldsymbol{\omega}^* = [w_1, \ldots, w_5]$ 和阈值 θ^*。对于每一折，我们在 $[0.3, 0.6]$中的候选权重配置（约束为非负且总和为一）和阈值取值上进行网格搜索。选择验证集平均 F1 分数最高的配置作为最终准入策略。学习到的权重具有直接的可解释性：例如，$\mathcal{T}$上的权重较大，表示内容类型对于目标领域具有很强的区分能力；而 $\mathcal{C}$ 上的权重较大，则表明依据性和幻觉预防至关重要。

算法 1 总结了准入流程。对于每个候选记忆，我们并行计算全部五种信号，通过学习到的策略将其聚合，并在候选记忆超过阈值时予以准入。当候选记忆与现有记忆冲突时（语义相似度高于 0.85，但内容不同），我们保留得分较高的表述并进行合并，以确保长期记忆保持精简且及时更新。

表 1：LoCoMo 测试集（N=225）上的性能比较。标注为 † 的延迟包含大语言模型 API 调用。最佳 F1 值以粗体显示。

方法	精确率	召回率	F1	延迟（毫秒）
随机	0.278	0.278	0.278	<1
MemGPT	0.316	0.333	0.324	2765†
MemoryBank	0.368	0.583	0.452	2843†
等权重	0.362	0.694	0.476	2916†
A-mem	0.371	1.000	0.541	3831†
A-MAC（我们的方法）	0.417	0.972	**0.583**	2644†

4 结果

4.1 实验设置

我们设计实验来回答三个研究问题。*RQ1*：A-MAC 能否优于现有记忆准入策略？*RQ2*：哪些特征对准入质量的贡献最大？*RQ3*：学习到的权重如何在不同对话领域间泛化？

我们在 LoCoMo 基准（Maharana 等人，2024年）上进行评估，使用涵盖个人助理交互、技术支持和研究协作对话的 30 段对话。每段对话包含 15–40 轮，并配有标注的依赖记忆的任务，从而提供约 1,500 条带有真实准入标签的候选记忆。我们使用 Sentence-BERT 生成嵌入，并使用本地大语言模型（Qwen 2.5）进行效用评分。我们将数据划分为训练集（70%）、验证集（15%）和测试集（15%），在训练集上进行权重优化，所有报告结果均基于留出测试数据。我们将其与四种基线方法进行比较：以 30% 概率进行准入的随机方法，作为下界；*MemGPT*（Packer 等人，2023年），使用时效性和大语言模型判定的重要性，并采用原论文中的权重；*MemoryBank*（Zhong 等人，2024年），结合时效性、相关性和重要性；以及*A-mem*（Xu等人，2025年），使用大语言模型生成的结构化属性进行余弦相似度匹配，代表当前最先进水平。

4.2 主要结果

表 1 比较了所有方法在 LoCoMo 测试集上的表现。A-MAC 的 F1=0 达到 0.583，优于此前最先进水平的 A-mem，相对提升 7.8%（0.583 vs 0.541）；与等权重相比提升 22.4%，与 MemoryBank 相比提升 29.0%，与 MemGPT 相比提升 80%。关键的是，在所有基于大语言模型的方法中，A-MAC 的精确率最高（0.417），同时保持了近乎完美的召回率（0.972），实现了良好平衡。这对于实际部署至关重要，因为假阳性（导致检索性能下降的记忆膨胀）和假阴性（因缺失上下文而导致对话失败）都会带来显著成本。精确率-召回率权衡表明，A-mem 的完美召回率（1.0）是以较低精确率（0.371）为代价的，这意味着它会纳入许多永远不会被引用的记忆；而 A-MAC 的学习得到的阈值则能有效过滤这些记忆。

除了准确率之外，A-MAC 还显著提升了效率，其延迟比 A-mem 低 31%（每个候选记忆为 2644 毫秒，而 A-mem 为 3831 毫秒）。这一提速源于 A-mem 需要多次顺序调用大语言模型来生成结构化属性，包括上下文摘要、关键词提取、语义标签和重要性评分；相比之下，A-MAC 的混合架构仅需调用一次大语言模型来获取效用特征，其余四项特征则通过高效的基于规则的方法计算，合计耗时不到 65 毫秒。对于每天处理数千轮对话的应用而言，在保持更高准确率的同时，这种效率差异可显著节约计算成本。图 2 展示了各种方法之间的精确率-召回率权衡，表明 A-MAC 位于曲线的最优区域。

4.3 消融研究

表 2 展示了特征消融结果，以及通过 5折交叉验证学到的权重。消融实验结果证实了这一重要性排序：移除类型先验会导致最大的性能下降（ΔF1=$-$0.107），使 F1 值从 0.583 降至 0.476，基本与等权重基线持平。

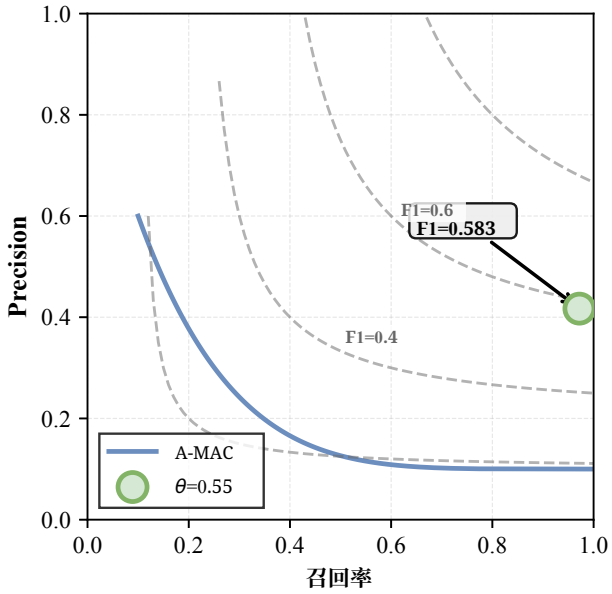

图 2：精确率-召回率权衡对比。A-MAC 在精确率与召回率之间实现了最佳平衡，位于曲线的右上区域。虚线表示 F1 等值线。

表 2：消融研究展示了移除各项特征对性能的影响。类型先验（$\mathcal{T}$）是主要贡献因素。

特征	不含时的 F1 值	Δ F1
完整模型	0.583	–
类型先验 ($\mathcal{T}$)	0.476	−0.107
新颖性 ($\mathcal{N}$)	0.555	−0.028
效用 ($\mathcal{U}$)	0.560	−0.023
置信度 ($\mathcal{C}$)	0.568	−0.015
时效性 ($\mathcal{R}$)	0.570	−0.013

这验证了这样一种直觉：区分用户偏好和身份陈述（会持续存在）与暂时性情绪状态（可能会被遗忘），能为记忆准入提供最可靠的启发式方法。其余特征——新颖性、效用、置信度和时效性——则逐步带来增益，F1 值提升范围为 Δ.013 至 −0.028；这些特征相辅相成，共同解释了等权重（0.476）与 A-MAC（0.583）之间的差距。

4.4 阈值敏感性

准入阈值 θ 控制精确率-召回率权衡：提高阈值会提升精确率，但代价是召回率下降。为将阈值的影响与数据方差区分开来，我们在权重优化期间使用的交叉验证折上进行此项分析。表3展示了 A-MAC 在 0.30 至 0.70 的不同阈值下的表现。最优阈值 $\theta^* = 0.55$ 使验证集 F1 分数达到最高值 0.571，并且在留出测试集上也达到 0.583（表1）。较低的阈值（$\theta \leq$0.45）可实现完美或近乎完美的召回率，但会纳入过多候选项，使精确率降至 0.36–0.37。较高的阈值（$\theta \geq 0.65$）只能略微提高精确率，却会导致召回率大幅下降，漏掉超过 40% 的相关记忆。 θ =0.50 至 $\theta = 0.60$ 之间相对平缓的 F1分数平台期表明，在此范围内阈值选择具有稳健性，从而降低了部署时对超参数调优的敏感性。

表 3：阈值敏感性分析。F1 值在 $\theta^* = 0.55$ 时达到峰值。较低的阈值能最大化召回率，但代价是精确率下降；较高的阈值则过于严格。

阈值 θ	精确率	召回率	F1	已接纳
0.30	0.360	1.000	0.529	100
0.40	0.360	1.000	0.529	100
0.50	0.376	0.972	0.543	93
0.55	**0.410**	**0.944**	**0.571**	**83**
0.60	0.395	0.833	0.536	76
0.65	0.368	0.583	0.452	57
0.70	0.381	0.444	0.410	42

表4：按组件划分的延迟分解。基于大语言模型的效用占据了大部分计算时间；基于规则的特征执行效率很高。

组件	延迟（毫秒）	百分比
效用 $\mathcal{U}$（大语言模型）	2580	97.6%
置信度 $\mathcal{C}$ (ROUGE-L)	18	0.7%
新颖性 $\mathcal{N}$ (SBERT)	32	1.2%
时效性 $\mathcal{R}$ (衰减)	<1	$<0.1\%$
类型先验 $\mathcal{T}$ (规则)	14	0.5%
总计 (A-MAC)	**2644**	**100%**

4.5 延迟分析

表4按组件细分了计算时间。效用特征因大语言模型推理而占据主要延迟（2580 毫秒，占 97.6%），而四个基于规则的特征（置信度、新颖性、时效性、类型先验）合计在 65 毫秒内完成。这种不对称的成本结构促使 A-MAC 采用混合设计：将成本高昂的大语言模型调用留给无法通过规则近似的语义判断，同时高效计算确定性特征。与需要多次大语言模型调用来进行属性生成、耗时 3831 毫秒的 A-mem 相比，A-MAC 的单次调用架构提速 31%。在批处理场景中，基于规则的特征可以在各候选项间并行计算，同时对大语言模型调用进行批处理，从而进一步提高吞吐量。

4.6 跨领域泛化

为评估学习到的权重能否跨对话环境迁移，我们分析了 A-MAC 在 LoCoMo 测试集中的个人对话（身份、家庭关系、偏好）与职业对话（职业生涯讨论、工作项目、创业）上的表现。表5和图3展示了跨领域结果。A-MAC 在个人领域达到 F1=0.482，在职业领域达到 F1=0.338，跨领域划分的标准差为 σ=0.102。个人领域的 F1 值更高，可能是因为个人叙述包含更多明确的偏好陈述，与类型先验特征的优势高度契合；而职业讨论通常涉及隐含上下文和领域专用术语，当前特征可能无法充分捕捉这些信息。值得注意的是，同一组学习到的权重在两个领域均保持有效，无需进行领域特定重新调优，这表明 A-MAC 的特征集捕捉到了记忆价值的领域不变原则，而不是过拟合于训练期间观察到的特定对话模式。A-mem 和 A-MAC 均实现了较高的召回率（分别为 1.0 和 0.972），这反映出基于大语言模型的方法在不确定未来相关性时倾向于采用宽松纳入策略。关键区别在于精确率：A-MAC 的精确率比 A-mem 高 12.4%（0.417 对 0.371），而召回率仅牺牲了 2.8%。这一精确率差距十分重要，因为低精确率的记忆存储库会积累无关条目，从而降低检索质量并增加下游任务的计算成本。

表5：跨领域泛化性能。A-MAC 无需重新调优即可有效迁移至两个领域，不过由于偏好模式明确，个人对话的 F1 值更高。

领域	样本	F1	Δ 平均值
个人	127	0.482	+0.072
专业	98	0.338	−0.072
Mean	225	0.410	–
标准差（σ）	–	0.102	–

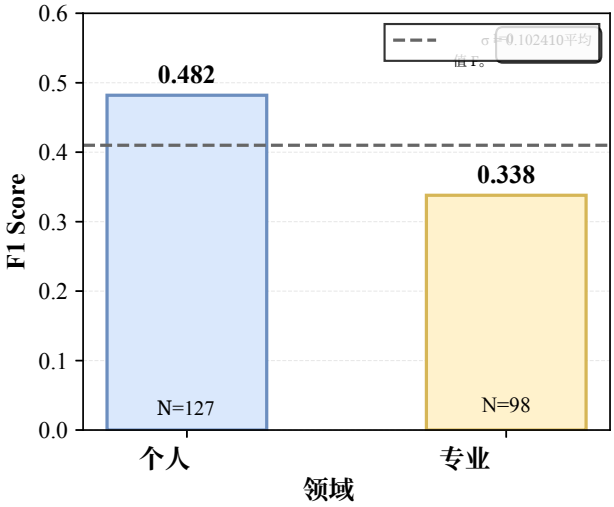

图3：跨领域 F1 值表现。个人对话的 F1 值更高，这是因为其中包含明确的偏好陈述，与类型先验的优势相契合。虚线表示各领域的平均表现。

5 结论

我们介绍了 A-MAC，这是一种面向大语言模型智能体的自适应记忆准入控制框架，可从五个具有可解释性的维度对候选记忆进行评分：效用 $\mathcal{U}$、置信度 $\mathcal{C}$、新颖性$\mathcal{N}$、时效性 $\mathcal{R}$ 和类型先验 $\mathcal{T}$。通过将高效的基于规则的特征提取与通过交叉验证优化学习权重相结合，A-MAC 在运行速度比以往方法快 31% 的同时，实现了最先进水平的 F1 值（0.583）。学习到的权重自动识别出类型先验是主导特征，验证了这样一种直觉：在对话智能体的记忆准入决策中，内容类别为最强的信号。与纯神经网络方法相比，A-MAC 将基于规则的特征（$\mathcal{C}$，$\mathcal{N}$，$\mathcal{R}$，$\mathcal{T}$）与基于大语言模型的评分（$\mathcal{U}$）相结合的混合架构，在准确率与计算效率之间实现了良好权衡。线性加权评分模型还提供了不透明的神经系统所缺乏的可解释性：开发者可以检查单个特征得分和权重贡献，了解特定记忆为何被纳入或拒绝，从而支持针对不同部署环境的调试、审计和策略改进。

参考文献

Akari Asai、Zeqiu Wu、Yizhong Wang、Avirup Sil 和 Hannaneh Hajishirzi。Self-RAG：通过自我反思学习检索、生成和批判。*arXiv* 预印本 *arXiv:2310.11511*，2023年。ICLR 2024 口头报告。

Yuntao Bai、Saurav Kadavath、Sandipan Kundu、Amanda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、Azalia Mirhoseini、Cameron McKinnon等。宪法式人工智能：来自人工智能反馈的无害性。*arXiv* 预印本 *arXiv:2212.08073*，2022。

Prateek Chhikara、Dev Khant、Saket Aryan、Taranjeet Singh和Deshraj Yadav。Mem0：借助可扩展的长期记忆构建可用于生产环境的AI智能体。*arXiv* 预印本 *arXiv:2504.19413*，2025年。

Darren Edge、Ha Trinh、Newman Cheng、Joshua Bradley、Alex Chao、Apurva Mody、Steven Truitt 和 JonathanLarson。由局部到全局：一种面向查询摘要的图检索增强生成方法。*arXiv*预印本 *arXiv:2404.16130*，2024年。

Bernal Jim´enez Guti´errez、Yiheng Shu、Yu Gu、Michihiro Yasunaga 和 Yu Su。HippoRAG：受神经生物学启发、面向大型语言模型的长期记忆。载于神经信息处理系统进展，第37卷，2024年。

Chenxu Hu、JieFu、Chenzhuang Du、Simian Luo、Junbo Zhao 和Hang Zhao。ChatDB：利用数据库作为符号记忆增强大语言模型。*arXiv*预印本 *arXiv:2306.03901*，2023年。

Lei Huang、Weijiang Yu、Weitao Ma、Weihong Zhong、Zhangyin Feng、Haotian Wang、Qianglong Chen、Weihua Peng、XiaochengFeng、Bing Qin 等。《大型语言模型中的幻觉综述：原则、分类体系、挑战与开放问题》。*arXiv* 预印本 *arXiv:2311.05232*，2023年。

Xun Jiang、Feng Li、Han Zhao、Jiaying Wang、Weiran Chen、Jingwei Zhao、Yuxuan Li、Wan He、Xiuqiang Chen、Fei Huang 等。《长期记忆：AI自我进化的基础》。*arXiv* 预印本 *arXiv:2410.15665*，2024年。

Jiazheng Kang、Mingming Ji、Zhe Zhao 和 Ting Bai。《AI智能体的 Memory OS》。*arXiv* 预印本 *arXiv:2506.06326*，2025年。自然语言处理实证方法会议 2025口头报告。

Vladimir Karpukhin、Barlas O˘guz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 与 Wen-tau Yih。面向开放领域问答的稠密段落检索。载于*2020*年自然语言处理实证方法会议论文集（自然语言处理实证方法会议），第6769–6781页，2020年。

Caitlin Latimer、Niccolo Boschi、Anton Neeser、Christopher Bartholomew、Gaurav Srivastava、Xuxi Wang 与 NarenRamakrishnan。Hindsight 是 20/20：构建能够保留、回忆与反思的智能体记忆。*arXiv* 预印本 *arXiv:2512.12818*，2025年。

Patrick Lewis、Ethan Perez、Aleksandara Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Kuttler、Mike Lewis、Wen-tau Yih、Tim Rockt¨aschel 等。面向知识密集型自然语言处理任务的检索增强生成。载于神经信息处理系统进展，第33卷，第9459–9474页，2020年。

Xiao Liu、Hao Yu、Hanchen Zhang、Yifan Xu、Xuanyu Lei、Hanyu Lai、Yu Gu、Hangliang Ding、Kaiwen Men、Kejuan Yang 等。AgentBench：将大语言模型作为智能体进行评估。*arXiv* 预印本 *arXiv:2308.03688*，2023年。ICLR 2024。

Adyasha Maharana、Dong-Ho Lee、Sergey Tulyakov、Mohit Bansal、Francesco Barbieri 和 Yuwei Fang。评估大语言模型智能体的超长期对话记忆。*arXiv* 预印本 *arXiv:2402.17753*，2024年。ACL 2024。

Charles Packer、Vivian Fang、Shishir Patil、Kevin Lin、Sarah Wooders 和 Joseph Gonzalez。MemGPT：迈向作为操作系统的大语言模型。*arXiv* 预印本 *arXiv:2310.08560*，2023年。

Joon Sung Park、Joseph O'Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。生成式智能体：人类行为的交互式模拟。*arXiv* 预印本 *arXiv:2304.03442*，2023年。UIST 2023。

Parth Sarthi、Salman Abdullah、Aditi Tuli、Shubh Khanna、Anna Goldie 和 Christopher D Manning。RAPTOR：用于树结构检索的递归抽象处理。载于《国际学习表征会议》，2024年。

Timo Schick、Jane Dwivedi-Yu、Roberto Dess`ı、Roberta Raileanu、Maria Lomeli、Luke Zettlemoyer、Nicola Cancedda 和 Thomas Scialom。Toolformer：语言模型可以教会自己使用工具。*arXiv* 预印本 *arXiv:2302.04761*，2023年。

Noah Shinn、Federico Cassano、Ashwin Gopinath、Karthik Narasimhan 和Shunyu Yao。Reflexion：通过语言强化学习实现的语言智能体。*arXiv* 预印本 *arXiv:2303.11366*，2023年。NeurIPS 2023。

Guanzhi Wang、Yuqi Xie、Yunfan Jiang、Ajay Mandlekar、Chaowei Xiao、Yuke Zhu、Linxi Fan 与 Anima Anandkumar。Voyager：一种基于大型语言模型的开放式具身智能体。*arXiv* 预印本 *arXiv:2305.16291*，2023年。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示激发大型语言模型中的推理。载于神经信息处理系统进展，第35卷，第24824–24837页，2022。

Wujiang Xu、Zujie Liang、Kai Mei、Hang Gao、Juntao Tan 与 Yongfeng Zhang。A-MEM：面向大语言模型智能体的智能体记忆。*arXiv* 预印本 *arXiv:2502.12110*，2025年。NeurIPS 2025。

Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。ReAct：在语言模型中协同推理与行动。*arXiv* 预印本 *arXiv:2210.03629*，2022。

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

Wanjun Zhong、Lianghong Guo、Qiqi Gao、He Ye 和 Yanlin Wang。MemoryBank：通过长期记忆增强大型语言模型。载于人工智能 *AAAI* 会议论文集，第 38 卷，第 19724–19731 页，2024年。