arXiv:2602.07092v1 [cs.MA] 6 Feb 2026

LEMON AGENT 技术报告

Haipeng Jiang[1], **Kailong Ren**[1], **Zimo Yin**[1], **Zhetao Sun**[1], **Xin Gan**[1], **Guangyi Lv**[1], **Ming He**[1, †], **Peng Wang**[1, †], **Congli Yin**[1], **Hong Pan**[1], **Changwen Zhang**[1], **Shan Tong**[1], **Zhengyu Xu**[1], **Zeping Chen**[1], **Yubin Huangfu**[1], **Yanzhi Xu**[1],**Xing Su**[1],**Qin Feng**[1],**Dong An**[1],**Jianping Fan**[1]

[1] 联想 CTO 组织 AI 实验室

摘要

近期先进的大语言模型驱动的智能体系统在处理复杂的长周期任务方面展现出卓越能力。然而，它们在资源效率、上下文管理和多模态感知方面仍存在固有局限。基于这些观察，**LemonAgent** 应运而生。这是一个基于全新提出的 AgentCortex 框架构建的多智能体编排器-工作器系统，该框架通过自适应任务执行机制对经典的规划器-执行器-记忆范式进行了形式化。我们的系统集成了分层自适应调度机制，在整体编排器层和工作器层均可运行。该机制能够根据任务复杂度动态调整计算强度，使编排器能够分配一个或多个工作器来并行执行子任务，同时工作器还可以通过并发调用工具进一步提高运行效率。凭借这种双层架构，系统在全局任务协调与局部任务执行之间实现协同平衡，从而优化复杂场景中的资源利用率和任务处理效率。为减少并行步骤中的上下文冗余并提高信息密度，我们采用了三层渐进式上下文管理策略。为更充分地利用历史信息，我们提出了一种自进化记忆系统，可从所有历史经验中提取多维有效信息，以辅助完成类似任务。此外，我们还提供了增强版 MCP 工具集。在权威基准测试上的实证评估表明，我们的 Lemon Agent 在 GAIA 上实现了最先进的 91.36% 总体准确率，并以 77+ 的分数位居 xbench-DeepSearch 排行榜榜首。我们的实验结果凸显了自适应资源协调与持续经验学习之间的协同作用，如何有效弥合智能体的理论潜力与现实世界部署的严格要求之间的差距。

项目：https://github.com/Open-LemonAgent/LemonAgent

1 引言

现代人工智能智能体已在各种任务和应用领域展现出卓越的能力（Su et al., 2025；Zhang et al., 2025a），能够在高度非结构化环境中掌握复杂推理和自适应决策（Team, 2025；Zhang et al., 2025b；Yu et al., 2025）。这些突破使人工智能自主智能体成为下一代应用的基础技术，适用于精准医疗、科学发现和可持续基础设施管理等关键领域。自主人工智能智能体不仅能大幅提升企业运营效率，还能让公众日常使用先进人工智能能力来解决现实世界问题（Maslej et al., 2025；at JDCHO, 2025）。

尽管自主人工智能智能体取得了长足进展，但仍面临至少三个关键问题：(1) 面对复杂程度各异的任务，采用静态资源分配策略，即使是简单任务也过度部署高性能大型模型，导致运营成本高得令人难以承受（Du 等，2025；Luo 等，2025）。这削弱了其在资源受限的现实场景中的环境适应性。(2) 低效的记忆演化策略依赖二元任务成功/失败结果来进行知识更新，因此在缺乏明确真实标注的复杂现实任务中，可能无法捕捉部分成功或格式偏差导致的失败所蕴含的宝贵见解（Fang 等，2025；Gao 等，2025）。由于无法从早期经验（即任务执行轨迹）中充分学习，这会妨碍其实现自主进化。(3) 感知不足

†通讯作者。邮箱：heming01@foxmail.com，wangpeng31@lenovo.com。

在处理高维多模态输入时尤为明显。这限制了现实场景应用中人机交互的丰富性和自然度（Nwankwo & Rueckert，2024年）。

基于这些观察，我们提出了**Lemon Agent**，这是一个构建于 AgentCortex 框架之上的多智能体系统，旨在实现高效、可靠的任务执行。Lemon Agent 无缝整合了五项核心结构创新：(1)*AgentCortex* 框架，通过实现规划器-执行器-记忆范式，确保工业级可靠性和学术通用性，同时支持面向服务的部署；(2) 分层自适应调度机制，在宏观编排和微观执行两个层面运作，根据任务复杂度动态调节计算强度和工具并行化；(3) 三层渐进式上下文管理策略，结合工具内截断、自适应摘要和追溯压缩，缓解长时程轨迹中的信息衰减；(4) 自进化语义记忆（*SES Memory*）模块，无论任务最终结果如何，均能从所有执行轨迹中提取可迁移技能片段，从而持续优化能力；(5)增强型 *MCP* 工具集，通过自适应图像理解工具、街景智能体、多源搜索工具、强大的文件读取工具等专用模块，增强环境交互。这些组件协同作用，使 Lemon Agent 能够将自适应资源管理、持续知识演化与高保真环境感知相结合，从而取得卓越表现。

实验结果表明，Lemon Agent 在多个不同的基准测试中均表现出色。截至2026年2月6日，我们的系统在 GAIA 基准上的总体准确率达到 91.36%，并以 77+的准确率位居 xbench-DeepSearch 基准测试榜首（Mialon 等，2024；Chen 等，2025）。这些结果验证了我们管理庞大搜索空间并解决复杂多步推理问题的能力。除学术指标外，AgentCortex 框架在联想超级智能体中的部署也展现了其工业级可靠性。这一商业实现的交易额达数亿元，并获评 2025 CCF 企业数字化优秀案例。

我们的主要贡献总结如下：

• 我们推出 Lemon Agent，这是一个基于 AgentCortex 基础设施构建的统一多智能体编排系统，将规划器-执行器-记忆范式标准化。我们的系统整合了分层自适应调度机制、三层上下文管理策略以及 SES Memory 协议，以持续完善技能；同时还配备了增强型工具套件，涵盖自适应视觉理解、智能搜索和精准地理空间导航。
• 我们通过在包括 GAIA 和 xbench DeepSearch 在内的权威基准测试上开展广泛的实证研究，验证了系统的卓越性能与架构通用性；在这些基准测试中，Lemon Agent 实现了业界领先的结果。为鼓励社区研究并促进透明验证，我们提供完整的实现作为开源项目，以支持自主智能体技术的发展。

2 LEMON AGENT

我们介绍了 Lemon Agent，这是一种模块化多智能体系统，旨在复杂动态环境中实现稳健且高效的任务执行。图 1 展示了系统的全面概览。我们的系统通过分层自适应调度机制协调专门的子工作器集群，在宏观编排与微观执行层面调节计算强度。为在延长轨迹中保持推理连贯性，我们引入了三层渐进式上下文管理策略和自进化语义记忆系统，以从多样化的执行轨迹中持续开展经验学习。这种集成方法还借助增强型工具套件得到进一步强化，其中包括自适应视觉优化、多源搜索和精准地理空间导航。相关细节将在以下小节中介绍。

2.1 AgentCortex 框架

我们使用 AgentCortex 实现了该系统。AgentCortex 是一种面向智能体的技术框架和设计范式，旨在弥合学术研究与现实世界产品部署之间的差距。

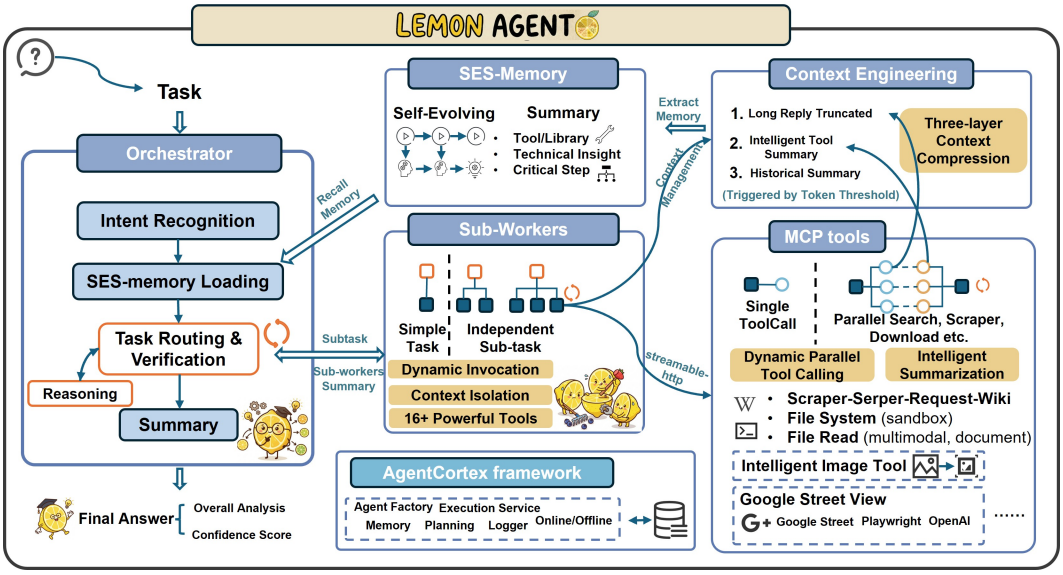


图 1：Lemon Agent 系统概览。我们的编排系统集成了多智能体协作、子工作者集群、自进化记忆和工具模块。

该基础设施通过提供以下技术特性，为构建自主系统奠定了稳健且模块化的基础：

• **模块化架构**。 该框架将智能体拆解为定义明确的组件，包括意图理解、任务分解、规划、工具执行、知识检索、记忆管理和任务总结。这些组件通过统一且高度抽象的接口连接，从而实现低耦合、高灵活性和可持续的系统演进。

• **技术融合**。 AgentCortex 的一项关键优势在于能够将算法创新与严谨的工程实践融为一体。研究人员可以快速将最先进的算法或自主开发的算法集成到完整的智能体流水线中进行实验，同时保留直接迈向生产级实现的路径。

• **生产级基础设施**。 该框架提供内置的企业级能力，包括高性能微服务引擎、完善的日志记录与监控基础设施，以及稳健的数据库访问机制。这确保智能体系统能够在各种真实场景中可靠地大规模部署。

AgentCortex 的实际成效在于，它能够显著缩短从实验性原型到可投入生产的智能体的转化路径。通过将研究工作流与工业工程实践相结合，该框架能够使复杂的智能体行为无缝过渡到稳定的商业运营中。这些成果验证了 AgentCortex 先进的架构设计，以及其在复杂多样的任务环境中的工业级可靠性。

2.2 分层自适应调度

为优化资源分配和执行效率，Lemon Agent 采用一种分层自适应调度机制，在宏观编排和微观执行两个层面发挥作用。这种双层自适应机制使系统能够根据任务本身的复杂性和结构要求，动态调整计算强度。通过平衡并行探索的广度与顺序推理的精确性，该框架能够在各种工作负载下实现高性能，同时保持可持续的资源占用。调度逻辑分为以下两个不同的自适应维度：

2.2.1 宏观层面动态编排器-工作器

主代理根据全局目标在结构上的独立性，判断是否需要多专家协作。对于简单任务，系统倾向于仅调用一个子工作器，以精简配置，尽可能减少代理间的通信开销和推理延迟。然而，如果任务包含多个相互独立的方面或子目标，主代理便会启用由专门子工作器组成的协作团队。这种宏观层面的调整确保系统只会在任务复杂度足以证明额外计算开销合理时，才扩展专家资源。

2.2.2 微　　观层面动态工具并行　　化

在每个子工作器内部，系统实施细粒度的执行策略，可在一到五个实例之间动态调整并行工具调用的数量。并行化程度由子任务的具体功能需求决定。在新闻检索等信息密集型场景中，子工作器可以执行多个并发搜索引擎查询，随后进行并行化的网络爬取，以加速数据采集。同样，对于涉及图像识别的多模态任务，智能体会同时触发多张图像的下载与并发视觉分析，以确保快速完成证据综合。相反，当遇到复杂推理链或步骤间依赖度较高的任务时，调度机制会优先采用顺序的单工具执行模式。在逻辑密集型场景中采用单次调用的这一设置，可以保持推理轨迹的完整性。

2.3 分层上下文管理策略

Lemon Agent 采用三级渐进式上下文压缩机制。该策略旨在降低信息衰减和上下文窗口溢出的风险，这些问题在涉及大规模数据检索的多步轨迹中十分常见。分层管理策略包含以下三个功能层：

• **工具内截断与元数据记录**。 第一层在单个工具层级实施，通过严格的字符数限制来发挥作用。当单个工具的原始输出超过预设阈值时，系统会先进行截断，避免一次检索操作耗尽上下文预算。在此过程中，代理会记录相应的元数据。这些元数据充当结构锚点，使系统能够追踪被截断内容的来源和性质，以便在后续层级进行更精准的处理。

• **轮次内自适应摘要**。 第二层侧重于单轮交互中的累积输出。当所有工具响应及其相关子查询的总长度达到特定启发式阈值时，就会触发此机制。在这种情况下，系统会对该轮的完整执行轨迹进行智能综合。如果总长度显著超过阈值且包含先前被截断的数据，摘要引擎会优先精确重构这些截断片段。这样可确保代理在不被未经整理的原始数据淹没的情况下，保留对该轮的高层语义理解。

• **跨轮次追溯压缩**。 第三层通过追溯压缩协议管理长期历史。当累积的历史上下文达到容量上限，且系统识别出已有被截断工具记录时，便会启动这一流程。该机制会回溯交互历史，定位这些被截断的节点，并进行二次摘要。随后，系统会将原始结果就地替换为压缩后的版本。这种追溯更新在保留多步推理所必需的关键逻辑联系的同时，显著减少了整体上下文占用。为防止重复摘要导致信息丢失，系统会在完成跨轮次更新后立即重置截断记录表。

2.4 自进化语义记忆

我们首先提出一种名为 SES-Memory 的自演化语义记忆，用于从执行轨迹中提取可迁移技能。不同于受二元

成功/失败评估限制的传统记忆系统，SES-Memory 以过程为中心的技能提取方式模仿人类经验学习，从而实现持续的能力优化。具体而言，传统方法只从成功的任务中提取完整解决方案。SES-Memory 则从所有智能体执行轨迹中提取高价值技能片段，不论最终任务成功还是失败。这模仿了人类专家从所有过往工作中同等学习的方式，因为即使是失败的任务，也可能在设计或实现方面蕴含宝贵见解。例如，数据分析任务可能因最终结论错误而失败，但其中间步骤仍可能包含有价值的信息。SES-Memory 能够提取并概括这些局部亮点，将其转化为独立技能。

通过反复尝试，智能体能够从历史执行中自动获取丰富的先验知识。SES-Memory 不仅限于历史摘要，还涵盖从经验中提炼出的多种可复用知识，包括以往使用过的代码片段、工具和库、技术洞见以及关键决策步骤。智能体还可以重新组合过去不同执行中的要素，以替代且更高效的方式解决新任务。
此外，为确保记忆质量和可扩展性，SES-Memory 引入了两阶段控制机制。在检索出最相关的 top-k 条记忆后，会应用基于相似度的阈值过滤器，防止低价值或含噪记忆进入上下文。此外，当一次查询检索到多条相似度始终较高的记忆时，系统不会存储新生成的记忆，从而避免冗余条目和记忆不受控制地增长。

2.5 工具增强

为应对用户交互需求日益增长的复杂性和多样性，Lemon Agent 系统性地扩展其工具集，以赋予智能体现实世界问题解决能力。通过为智能体配备更完善的感知、分析和检索工具，我们的系统能够有效地获取、处理并综合来自不同模态和来源的信息。这不仅拓宽了 Lemon Agent 可处理任务的范围，也大幅提升了其在现实世界环境中的适应性和可靠性。

2.5.1 智能图像工具

当前用于图像理解的视觉语言模型（VLM）存在固有的分辨率限制。一方面，超过特定阈值的输入图像通常会在特征提取和理解之前进行缩放，这会导致细粒度细节大量丢失。另一方面，这些模型对图像中非主要区域的感知能力往往有限。这两项缺陷共同导致视觉语言模型作为视觉感知工具时，无法准确分析小范围局部区域中的目标信息。值得注意的是，现代视觉语言模型的空间定位和定位能力持续提升。利用这一优势，我们设计了一种智能图像工具，以增强对小范围局部区域的理解。工作流程首先由视觉语言模型判断是否需要进行局部分析。如果认为无需进一步细化，模型便直接输出结果。反之，如果需要查看局部区域，视觉语言模型会为目标区域生成对角相对坐标，记为 $r = [x_{min}, y_{min}, x_{max}, y_{max}]$，其中坐标值归一化至 0 到 1 之间。随后，根据原始图像尺寸（W， H），将这些相对值转换为绝对像素坐标 $P = [X_{min}, Y_{min}, X_{max}, Y_{max}]$，转换方式如下：

$$X_i = x_i \cdot W, \quad Y_i = y_i \cdot H$$

随后，系统根据这些绝对坐标裁剪原始图像，并将生成的高分辨率局部片段重新输入视觉语言模型，以获得最终的精细化结果。

2.5.2 街景代理

为弥合高层子任务与谷歌地图低层空间导航之间的差距，我们引入了街景智能体，这是一个专门用于处理三维环境中复杂地理空间任务的模块。浏览器操作代理常常受困于一般网页浏览的高延迟和冗余信息，而街景智能体则充当专用接口，将抽象的导航目标转化为精确的 API驱动的操作。借助这一机制，主代理能够专注于高层策略，而街景智能体则负责处理

空间感知和基于坐标的移动的复杂细节。我们的街景智能体在技术上的优势体现在以下三大核心方面：

• **全景感知**。 空间推理的一项主要挑战是，通过标准界面探索环境需要过多的交互轮次。为了解决这一问题，我们的智能体采用了全景信息合成机制。它不再依次调整视角，而是通过 Google 地图 API 获取地理空间元数据，以编程方式将四张正交视角图像拼接成一幅全景表示。我们的智能体在单轮交互中向大语言模型提供全面的环境快照。

• **精确的基于坐标的导航**。 与网页浏览器中的基于启发式的导航不同，后者往往依赖不精确的方向点击；我们实现了坐标驱动的空间控制。每次移动时，智能体都会根据所需位移明确计算目标纬度和经度。通过直接调用 API 获取这些精确坐标处的图像，它能够实现厘米级定位。这种精细控制使智能体能够灵活执行微移动，从而细致地搜索局部区域。

• **以目标为中心的定位**。 我们还集成了新提出的图像工具，以支持主动视觉定位和空间重新定位。当在全景表示中检测到候选目标时，智能体会调用自适应裁剪机制，生成场景的高分辨率局部片段。通过从细化过程中计算出的绝对坐标推导出精确方位，智能体将这些像素级空间数据点转换为方向导航向量。借助这一集成，智能体能够自主移动至最合适的观察位置，在那里目标最清晰可见且遮挡最少。细化工具与导航界面的协同作用，确保收集到成功执行任务所需的高保真视觉证据。

2.5.3 多源搜索工具套件

为实现高保真信息锚定，Lemon Agent 使用了多元化的搜索套件，其中包括 Google 搜索、维基百科历史修订版本和专业网络爬取能力。该工具套件的核心是稳健的三级回退协议：优先使用 Jina.ai Reader API 进行结构化提取；对于广泛查询，则回退至 Serper MCP 服务；最后，使用原生 HTTP 请求作为备用方案。这种分层设计辅以指数退避重试算法，以及一个智能输入清理层；该清理层使用正则表达式，自动修正格式错误或欠佳的查询结构。除了原始信息检索，该系统还会动态过滤冗余信息，并根据具体错误类型向智能体提供细致的诊断反馈，从而提升信噪比。通过将失败转化为后续规划周期中可执行的建议，这一闭环搜索解决方案确保 Lemon Agent 即使在与不可预测的外部数据源交互时，也能保持稳定、连贯的推理轨迹。

2.6 任务执行流程

算法 1 展示了 Lemon Agent 的运行工作流：通过一系列结构化步骤，整合上述架构组件。Lemon Agent 首先初始化 AgentCortex 框架，包括智能体模块和执行服务。系统执行意图识别，以确定全局目标，并从 SES Memory 中检索技术洞见。随后，系统根据检索到的洞见加载特定技能和功能库，然后进入推理循环。

在循环中，Lemon Agent 通过宏观编排执行任务路由和验证，以确定协作模式。系统部署动态子代理来处理相互独立的子任务，或分配一个子工作器进行线性执行。执行引擎通过可流式传输的 HTTP 协议调用 MCP 工具，并在单个工具调用和动态并行工具调用之间进行选择。系统会触发专用模块，例如智能图像工具或街景智能体，以处理多模态数据和地理空间数据。

系统会监控累计令牌数，并在达到预设阈值时应用三层上下文压缩策略。此过程包括工具内截断、轮次内摘要以及跨轮次追溯更新，以维持架构稳定性。

算法 1 Lemon 智能体执行流水线

要求： 用户任务 T
确保： 最终答案 A，置信度分数为 S_{conf}
使用 AgentCortex.Framework：实例化核心服务（智能体、MCP 工具、执行）
SES 记忆加载：检索与 T相关的前-k 个条目 $\mathcal{M}_{\text{SES}}$
可选推理：如果已启用，对 T进行推理，以更好地理解任务意图
主循环（任务路由与动态子代理）：
    初始化智能体集合 $\mathcal{A} \leftarrow \emptyset$
    当任务 T 尚未完全解决时：
        任务路由：将 T路由至适当的子代理 $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$
        对于每个子代理 $a_i \in \mathcal{A}$**并行执行**：
            为子代理 a_i初始化上下文 C_i
            当子智能体 a_i 未终止时：
                独立 ReAct：让 a_i 针对其分配的子任务进行 ReAct
                动态并行工具调用：
                    为 a_i选择工具 $\mathcal{T}_i = \{t_1, t_2, \ldots, t_m\}$
                    并行执行所有工具（位于 $\mathcal{T}_i$ 中），收集结果 $\mathcal{R}_i$
                上下文压缩：
                    1. **长回复已截断**：
                        对于每个结果 $r_j \in \mathcal{R}_i$：
                            如果 $|r_j| > \tau_{\text{max}}$：使用标记截断 r_j
                    2. **智能工具摘要**：
                        如果有多个工具且 $\sum_{r_j \in \mathcal{R}_i} |r_j| > \tau_{\text{multi}}$：
                            在将 $\mathcal{R}_i$ 提供给 a_i之前，进行智能摘要
                    3. **历史摘要**：
                        如果 $|C_i| > \tau_{\text{context}}$：
                            通过摘要压缩历史上下文 C_i
                使用处理结果更新上下文 C_i
            验证：子智能体 a_i 决定是否验证其结果
            子代理摘要：子智能体 a_i 对其结果进行摘要，以减少主代理的上下文长度
        整合子代理 a_i 输出
结果聚合：汇总所有子代理的答案，生成最终答案 A，并附上置信度分数 S_{conf}
SES 记忆写入：将执行上下文、任务 T、答案 A 和分数 S_{conf} 写入 SES 记忆
返回 A , S_{conf}

Lemon Agent 会在任务完成时，结合全面的整体分析和置信度分数生成最终答案。系统通过从执行轨迹中提取高价值技能片段，并将这些片段存回 SES Memory，以持续积累经验，从而结束本轮周期。

3 实验

3.1 实验设置

我们在 GAIA 和 xbench-DeepSearch 基准测试上评估 Lemon Agent 的通用适应能力。**GAIA** 是一个知名基准测试，旨在评估通用人工智能助手处理实际问题的能力，这些问题需要运用多项核心技能（Mialon 等，2024）。对人类而言，这些问题在概念上仍然简单，但对当今最先进的人工智能系统来说却极具挑战。**xbench-DeepSearch**是 xbench 的 AGI-Aligned 系列的一部分，评估在搜索和信息检索场景中使用工具的能力（Chen 等，2025）。它聚焦于整合规划、多步搜索和逻辑综合，以解决专家级查询。

在实验中，我们采用 GPT-5 作为核心语言模型骨干，并设置了特定参数。对于主代理，我们将推理级别设为高，并将最大

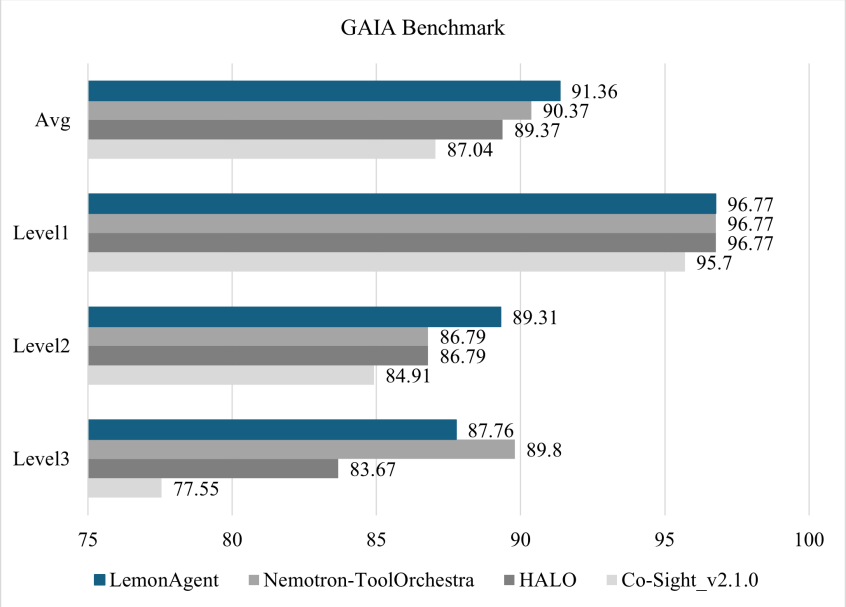


图2：Lemon Agent 在 GAIA 基准的全部三个难度级别中与其他表现出色的智能体的性能对比。

交互轮数设为 10。对于子工作器，我们同样将推理级别设为高，同时允许执行序列最长达到 20 轮。摘要模块采用中等推理级别，以平衡计算效率与输出质量。我们将模型温度设为 1.0，以确保输出多样性，并将最大输出词元数限制为 128,000。我们的经验策略采用 pass@3 方法，即 Lemon Agent 会对每项任务尝试 3 次。评估流程支持多进程并行执行，默认同时运行 3 个进程，以确保可扩展性。

3.2 基准测试结果

截至2026年2月6日，Lemon Agent 在 GAIA 基准的三个级别中总体准确率达到91.36%。其各级别成绩分别为：第1级任务96.77%，第2级任务89.31%，第3级任务87.76%。与其他智能体相比，Lemon Agent 在第1级和第2级均取得最先进的性能。这些成绩如图2所示，展现了其在复杂扩展任务序列中的多步推理能力取得了显著进步。

此外，截至2026年2月6日，Lemon Agent 在 xbench-DeepSearch 排行榜上位居榜首。表 1 总结了其与其他领先系统的性能对比。其总体准确率达到 77+，超过了包括 ChatGPT 5 Pro 和 SuperGrok Expert 在内的领先模型。这一结果表明，Lemon Agent 在处理庞大搜索空间以及从多源信息流中整合高保真证据方面具有卓越能力。

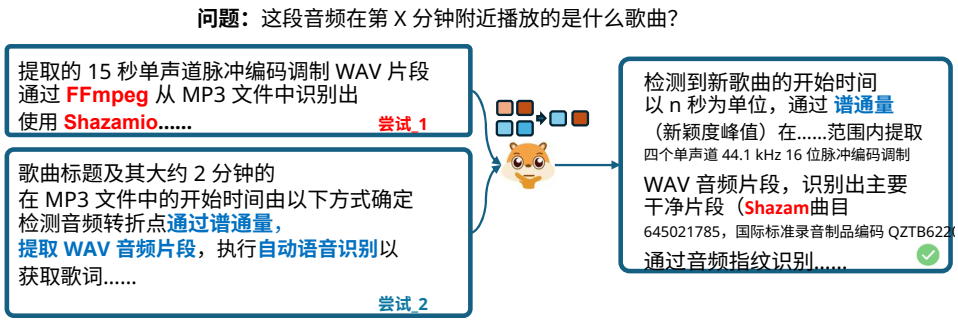
表1：Lemon Agent 在 xbench-DeepSearch 排行榜上的表现。

产品	组织	准确率	评估日期
Minimax Agent	Minimax	35+	2025年-11-04
阶跃星辰研究	阶跃星辰	35+	2025-08-28
Flowith	Flowith	35+	2025-08-30
SuperGrok Expert	xAI	40+	2025-08-28
ChatGPT-5-Pro	OpenAI	75+	2025-08-28
Lemon Agent	**联想首席技术官组织的 LR AILab**	**77+**	**2026年2月4日**

3.3 SES-MEMORY 结果

SES-Memory 模块会分析人工智能智能体如何完成任务，并从这些经验中提取可复用技能。它将特定代码片段转化为可应用于类似问题的通用模式。例如，详细的音频识别步骤会变成一个可复用模板，如“使用 Shazam 识别音频”。常见的工作流序列，例如搜索网络、抓取内容以及识别 PDF 文件中的文本，也会被识别为可重复的模式。我们的方法还会捕捉决策逻辑，形成诸如“任务需要 X 时采用方法 A，否则采用方法 B”这样的规则。这些提取出的技能会以结构化的自然语言描述形式存储，既便于人类理解，又能保留可执行的技术知识。

如图3所示，以识别音乐为例，看看这一过程是如何运作的。第一次尝试时，智能体使用 ffmpeg 提取音乐片段，并使用 Shazamio 进行识别。第二次尝试时，它采用了不同的方法，使用谱通量提取音频，并通过自动语音识别（ASR）进行识别。第三次尝试时，智能体利用 SES-Memory 中存储的前两次尝试所得经验，将谱通量提取与 Shazam库识别相结合，成功完成了任务。值得注意的是，SES-Memory 无需加载完整的代码库。相反，它会在其记忆结构中以简单提示和指令的形式存储实用的编程知识。这种方法使系统能够在不必管理庞大代码仓库的情况下，提升处理复杂任务的性能。



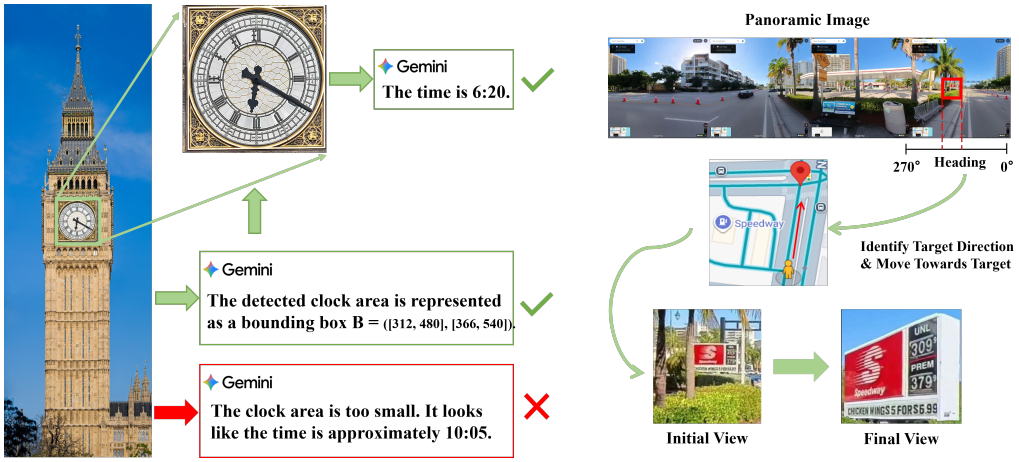
Fig图3：结合多次历史尝试中的实用技能完成任务的案例 via SES-Memory 模块。 s

3.4 工具增强结果

我们的智能图像工具在细粒度视觉感知方面展现出显著的实证改进。我们使用 GAIA 基准中一项具有代表性的 Level 2 任务来展示其有效性。该任务要求计算两个时钟之间的时间差。当我们使用标准视觉语言模型（如 Gemini 3 Pro）直接处理原始图像时，在目标物体仅占图像总面积极小比例的场景中，结果经常不准确。这一失败源于激进的分辨率压缩，导致模型无法区分时针和分针，或完全误判其方向。相反，我们的自适应工具会动态地分离并裁剪钟面，将其转化为高分辨率局部片段。这一细化使智能体能够提取精确的视觉证据，并绕过全局图像缩放的限制，从而得出正确结论。我们在图 4a 中展示了这一具体的局部细化流程及其对推理准确度的影响。

此外，我们通过 GAIA 基准中的一项复杂地理空间推理任务，展示了街景智能体的性能。图 4b 展示了一项 Level 2 任务，该任务要求识别加油站标牌上的燃油价格信息。我们的街景智能体首先检索目标位置的全景图像，由此开始整个流程。为提高识别精度，智能体先将全景图裁剪为可能存在目标的候选区域，并进行初步识别，以确认目标是否存在。在本示例中，一旦检测到加油站标志，智能体便根据裁剪区域在全景图像中的相对坐标计算所需的移动角度。随后，它沿道路导航，并精确地将自身定位到标志正前方最清晰的视角。最后，智能体调用智能图像工具，对检索到的标志进行局部细化。通过将标志裁剪为高分辨率片段，如

从初始视图到最终视图的转变所示，智能体准确识别出标牌上所描绘的具体油价。该示例成功展示了我们的空间导航模块与自适应视觉优化工具之间的协同作用，同时证实了街景智能体在动态三维环境中的通用性。



(a) 智能图像工具　　(b) 街景智能体

Figu图 4：展示智能图像工具和街景智能体性能的 GAIA 案例体

4 局限性与未来工作

尽管 Lemon Agent 在多个基准测试中取得了最先进的成果，但我们也认识到一些限制，这些限制将指导我们未来的研究。目前，工具交互通常需要在智能体与外部环境之间传输大量数据，从而增加通信开销和延迟。为解决这一局限，我们正在探索集成沙箱执行环境，使智能体能够通过直接函数调用来调用模块。这一设计无需反复上传和下载图像或结构化数据等大型文件，从而显著缩短上下文长度。我们还计划探索更复杂的多智能体协作拓扑，以处理更加零散且相互正交的子目标。通过在沙箱中本地执行工具操作，并提升 SES Memory 模块的精确度，我们希望进一步提高整个系统的效率和可扩展性。

5 结论

Lemon Agent 通过将多智能体协作与实际工程考量相结合，推动了工业级规模的先进部署。我们提出了一种模块化系统，可实现经济高效、可持续的部署，同时确保在各种运行环境中保持稳健的性能。我们提出的系统在统一的规划器-执行器-记忆范式中，成功协调了分层调度、渐进式上下文管理和自演进语义记忆。在 GAIA 和 xbench DeepSearch 基准测试上的实验结果，验证了 Lemon Agent 卓越的推理能力以及管理海量搜索空间的能力。此外，在联想超级智能体中的成功部署，证明了我们方案具备工业级可靠性和商业价值。我们将完整实现作为开源项目提供，以促进社区协作，并支持自主智能体研究的发展。这项工作朝着构建均衡的人工智能系统迈出了重要一步，这类系统既重视高保真能力，也注重长期可持续性。

参考文献

JDCHO 智能体团队。Joyagent-jdgenie，2025年。网址：https://github.com/jd-opensource/joyagent-jdgenie。

Kaiyuan Chen、Yixin Ren、Yang Liu、Xiaobo Hu、Haotong Tian、Tianbao Xie、Fangfu Liu、Haoye Zhang、Hongzhang Liu、Yuan Gong、Chen Sun、Han Hou、Hui Yang、James Pan、Jianan Lou、Jiayi Mao、Jizheng Liu、Jinpeng Li、Kangyi Liu、Kenkun Liu、Rui Wang、Run Li、Tong Niu、Wenlong Zhang、Wenqi Yan、Xuanzheng Wang、Yuchen Zhang、Yi-Hsin Hung、Yuan Jiang、Zexuan Liu、Zihan Yin、Zijian Ma 和 Zhi wen Mo。xbench：通过职业对齐的现实世界评估追踪智能体生产力的规模化增长。*CoRR*，abs/2506.13651，2025年。doi：10.48550/ARXIV。2506.13651。URL https://doi.org/10.48550/arXiv.2506.13651。

Shangheng Du、Jiabao Zhao、Jinxin Shi、Zhentao Xie、Xin Jiang、Yanhong Bai 和 Liang He。《基于大语言模型的智能体优化综述》。*CoRR*，abs/2503.12434，2025年。doi: 10.48550/ARXIV.2503.12434。URL https://doi.org/10.48550/arXiv.2503. 12434。

Jinyuan Fang、Yanwen Peng、Xi Zhang、Yingxu Wang、Xinhao Yi、Guibin Zhang、Yi Xu、Bin Wu、Siwei Liu、Zihao Li、Zhaochun Ren、Nikos Aletras、Xi Wang、Han Zhou 和 Zaiqiao Meng。《自我进化人工智能智能体综合综述：连接基础模型与终身智能体系统的新范式》。*CoRR*，abs/2508.07407，2025年。doi: 10.48550/ARXIV.2508.07407。URL https://doi.org/10.48550/arXiv.2508.07407。

Huan-ang Gao、Jiayi Geng、Wenyue Hua、Mengkang Hu、Xinzhe Juan、Hongzhang Liu、Shilong Liu、Jiahao Qiu、Xuan Qi、Yiran Wu、Hongru Wang、Han Xiao、Yuhang Zhou、Shaokun Zhang、Jiayi Zhang、Jinyu Xiang、Yixiong Fang、Qiwen Zhao、Dongrui Liu、Qihan Ren、Cheng Qian、Zhenhailong Wang、Minda Hu、Huazheng Wang、Qingyun Wu、Heng Ji 和 Mengdi Wang。《自我进化智能体综述：迈向人工超级智能之路》。*CoRR*，abs/2507.21046，2025年。doi：10.48550/ARXIV.2507.21046。URL：https://doi.org/10.48550/arXiv. 2507.21046。

Junyu Luo、Weizhi Zhang、Ye Yuan、Yusheng Zhao、Junwei Yang、Yiyang Gu、Bohan Wu、Binqi Chen、Ziyue Qiao、Qingqing Long、Rongcheng Tu、Xiao Luo、Wei Ju、Zhiping Xiao、Yifan Wang、Meng Xiao、Chenwu Liu、Jingyang Yuan、Shichang Zhang、Yiqiao Jin、Fan Zhang、Xian Wu、Hanqing Zhao、Dacheng Tao、Philip S. Yu 和 Ming Zhang。《大语言模型智能体：方法论、应用和挑战综述》。*CoRR*，abs/2503.21460，2025年。doi：10.48550/ ARXIV.2503.21460。URL：https://doi.org/10.48550/arXiv.2503.21460。

Nestor Maslej、Loredana Fattorini、C. Raymond Perrault、Yolanda Gil、Vanessa Parli、Njenga Kariuki、Emily Capstick、Anka Reuel、Erik Brynjolfsson、John Etchemendy、Katrina Ligett、Terah Lyons、James Manyika、Juan Carlos Niebles、Yoav Shoham、Russell Wald、Tobi Walsh、Armin Hamrah、Lapo Santarlasci、Julia Betts Lotufo、Alexandra Rome、Andrew Shi 和 Sukrut Oak。2025年人工智能指数报告。*CoRR*，abs/2504.07139，2025年。doi：10.48550/ARXIV.2504.07139。URL：https://doi.org/10.48550/arXiv.2504.07139。

Grégoire Mialon、Clémentine Fourrier、Thomas Wolf、Yann LeCun 和 Thomas Scialom。GAIA：通用人工智能助手的基准测试。发表于第十二届国际学习表征会议，*ICLR 2024*，维也纳，奥地利，*2024* 年*5*月*7*日至*11*日。OpenReview.net，2024年。URL：https://openreview.net/forum?id=fibxvahvs3。

Linus Nwankwo 和 Elmar Rueckert。使用预训练语言和视觉基础模型进行多模态人类与自主智能体交互。*CoRR*，abs/2403.12273，2024年。doi：10.48550/ ARXIV.2403.12273。URL：https://doi.org/10.48550/arXiv.2403.12273。

Hongjin Su、Shizhe Diao、Ximing Lu、Mingjie Liu、Jiacheng Xu、Xin Dong、Yonggan Fu、Peter Belcak、Hanrong Ye、Hongxu Yin、Yi Dong、Evelina Bakhturina、Tao Yu、Yejin Choi、Jan Kautz 和 Pavlo Molchanov。Toolorchestra：通过高效的模型与工具编排提升智能，2025年。URL：https://arxiv.org/abs/2511.21689。

MiroMind AI 团队。Miroflow：高性能开源研究智能体框架。https: // github.com/MiroMindAI/MiroFlow，2025年。

Chengyue Yu、Siyuan Lu、Chenyi Zhuang、Dong Wang、Qintong Wu、Zongyue Li、Runsheng Gan、Chunfeng Wang、Siqi Hou、Gaochi Huang、Wenlong Yan、Lifeng Hong、Aohui Xue、Yanfeng Wang、Jinjie Gu、David Tsai 与 Tao Lin。Aworld：编排智能体人工智能的训练配方，2025年。URL https://arxiv.org/abs/2508.20404。

Hongwei Zhang、Ji Lu、Shiqing Jiang、Chenxiang Zhu、Li Xie、Chen Zhong、Haoran Chen、Yurui Zhu、Yongsheng Du、Yanqin Gao、Lingjun Huang、Baoli Wang、Fang Tan 和 Peng Zou。Co-sight：通过冲突感知元验证以及基于结构化事实的可信推理，增强基于大语言模型的智能体。*arXiv* 预印本 *arXiv:2510.21557*，2025a。网址：https://arxiv.org/abs/ 2510.21557。

Wentao Zhang、Liang Zeng、Yuzhen Xiao、Yongcong Li、Ce Cui、Yilei Zhao、Rui Hu、Yang Liu、Yahui Zhou 和 Bo An。AgentOrchestra：用于通用任务求解的分层多智能体框架，2025b。网址：https://arxiv.org/abs/2506.12508。