

多模态智能体社会网络：从认知推理到群体涌现

章育诚

(北京理工大学 计算机学院, 北京 100081)

摘要: 随着人工智能向认知智能演进, 多模态大模型正在经历从单体直觉决策向群体智能涌现的范式跃迁。针对传统思维链范式在处理动态复杂推理任务时的局限性, 构建涵盖个体认知、交互通信、群体演化、系统工程与安全伦理的 5 层智能体技术体系架构。首先, 在个体与交互层分析视觉前瞻引导下的结构化推理及模型上下文协议在消除协作孤岛中的核心作用; 其次, 在群体与系统层探讨从反从众辩论到大规模社会模拟的演变规律, 揭示社会规范与群体记忆的自组织形成机制; 最后, 直面去中心化网络中的安全挑战, 提出协作式对齐与互信机制。通过剖析当前研究在 Sim-to-Real 迁移中的认知瓶颈, 展望从显式语言向潜在空间通信、从工具使用向工具创造演进的未来愿景, 为构建下一代社会化智能体系统提供理论依据。

关键词: 多模态大模型; 多智能体系统; 群体智能; 社会模拟; 现实鸿沟; 智能体对齐

DOI: 10.11907/rjdk.261031

扫描二维码阅读全文: 

中图分类号: TP18

文献标识码: A

文章编号: 1672-7800(2026)003-0019-09

Multimodal Agent Social Networks: From Cognitive Reasoning to Collective Emergence

CHEONG Yik Sheng

(School of Computer Science, Beijing Institute of Technology, Beijing 100081, China)

Abstract: With the evolution from artificial intelligence to cognitive intelligence, multimodal large models are experiencing a paradigm transition from single intuitive decision-making to the emergence of swarm intelligence. Aiming at the limitations of traditional thought chain paradigm in dealing with dynamic and complex reasoning tasks, a 5-tier agent technology architecture covering individual cognition, interactive communication, group evolution, system engineering and safety ethics was constructed. Firstly, the key role of structured reasoning and model context protocol under the guidance of visual foresight in eliminating collaboration islands is analyzed at the individual and interaction levels; Secondly, at the group and system level, we explore the evolution law from anti conformity debate to large-scale social simulation, and reveal the self-organization formation mechanism of social norms and group memory; Finally, facing the security challenges in the decentralized network, a collaborative alignment and mutual trust mechanism is proposed. By analyzing the cognitive bottleneck of current research in Sim to Real migration, the future vision of the evolution from explicit language to potential spatial communication, and from tool use to tool creation, provides a theoretical basis for building the next generation of social agent system.

Key Words: multimodal large language models; multi-agent systems; collective intelligence; social simulation; reality gap; agent alignment

0 引言

人工智能领域正在从以模式识别为核心的“感知智能”向以决策规划为核心的“认知智能”跨越。伴随 Transformer 架构的统一与大规模预训练技术的成熟, 以 CLIP 为代表的对比学习范式成功打破了视觉与语言的表征界限,

为多模态大模型奠定了对齐基础^[1]。此后, GPT-4V、DeepSeek-VL 等模型的涌现进一步将这种对齐从语义匹配提升至指令遵循层面^[2]。然而, 当面向机器人自主规划、医疗诊断或法律推理等复杂现实场景时, 现有技术面临瓶颈。例如, 判断交通事故责任不仅需要识别车辆属性, 而且需结合交通规则进行逻辑推演, 这对模型的结构化思考能力提出了严峻挑战。

收稿日期: 2026-01-16

作者简介: 章育诚 (2000-), 男, 北京理工大学计算机学院硕士研究生, 研究方向为人工智能。

从建模视角来看,上述复杂任务往往涉及多主体的交互与博弈。然而,当前多模态大模型在应对该类任务时呈现出明显的“系统1(System 1)”特征,即依赖训练数据的统计规律进行直觉式快思考,而非严谨的逻辑推导。这种机制在简单场景下效率显著,但在长链条推理中极易引发“视觉幻觉”与逻辑断裂^[3]。例如,模型常因文本先验(如蓝天)忽视图像事实(如阴天),或在推理途中丢失上下文约束。因此,如何赋予智能体类似人类“系统2”的慢思考能力,并将其置于社会化交互网络中进行协同增强,成为突破当前认知天花板的关键。为突破当前存在的认知局限,学术领域进行了多方面的深入探究。然而,若想达到从“单体智能”到“社会化群体智能”的转变,仅通过增强个体的推理能力深度是远远不够的。在该过程中,存在一个亟待解决的重大难题——交互鸿沟,即不同结构的智能体之间如何以标准化的方式共享上下文信息与工具能力。

当前的智能体生态面临着严重的工具碎片化问题。不同的智能体框架(如 LangChain、AutoGen、Semantic Kernel)采用各自独立的接口标准,导致“ N 个模型适配 M 个工具”需要开发“ $N \times M$ ”个连接器,极大阻碍了能力的复用与协作。此外,智能体在通过工具与物理世界交互时,往往缺乏对业务流程(Standard Operating Procedure, SOP)的深层理解,导致“会调用API但不懂业务逻辑”的困境。为此,针对当前存在的认知局限与交互协作孤岛问题,本文面向社会化协同场景,构建了涵盖个体认知、交互通信、群体演化、系统工程与安全伦理的5层智能体技术体系架构。

对多模态智能体的复杂推理机制与群体智能涌现规律进行了系统深入的研究。

1 面向社会化协同的多模态智能体技术体系架构

面向社会化协同的多模态智能体技术体系架构见图1。首先,个体智能体能力层(第一层)聚焦多模态结构化推理,剖析 ToT、GoT 及 VisuoThink 等架构如何利用状态空间搜索,赋予智能体类人“系统2”的深度思考与视觉前瞻能力,奠定交互认知的基石;其次,多智能体交互与协作层(第二层)关注互操作性与分工,结合 OpenClaw 实践阐释 Agent Skills 范式的语义封装机制,并深入讨论角色动态分配与博弈策略;再次,群体智能与涌现层(第三层)探究大规模交互中的动态演化,通过分析 Free-MAD 的反从众辩论及 AgentSociety 的社会模拟,揭示去中心化网络中社会规范与宏观行为的涌现机理;接着,平台与系统工程层(第四层)致力于构建标准化基础设施,重点论述模型上下文协议(Model Context Protocol, MCP)在消除工具接口碎片化中的关键作用,并探讨万级并发仿真架构及可观测性平台的构建,为前述多智能体社会的落地提供工程底座;最后,安全、伦理与对齐层(第五层)贯穿全流程,直面协作中的新型风险,剖析 AiTM 攻击、认知病毒等威胁,并介绍 WaltzRL 等协作对齐机制,为构建可信数字社会提供技术屏障。

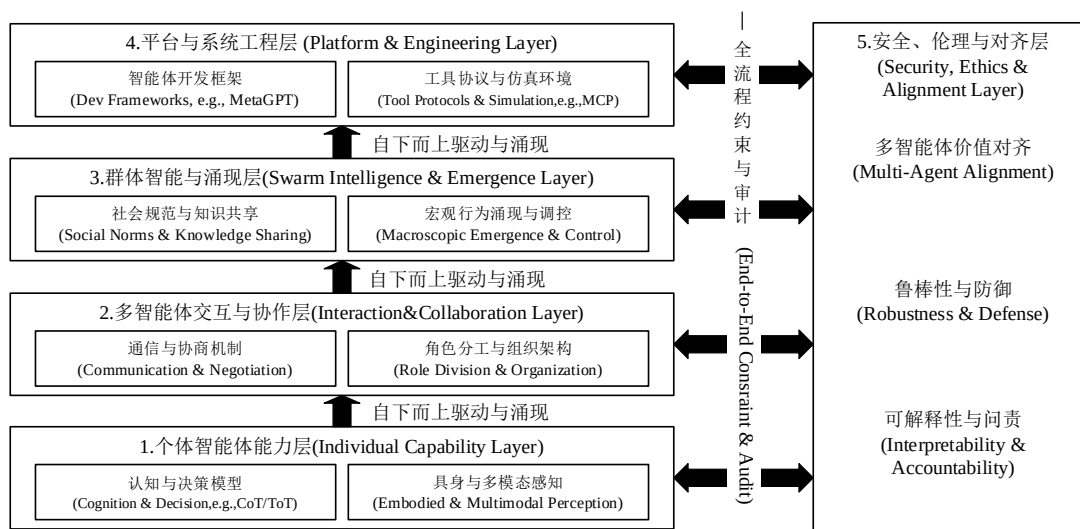


Fig. 1 Technical architecture for socially collaborative multimodal agents

图1 面向社会化协同的多模态智能体技术体系架构

2 个体智能体能力层——认知、具身与多模态感知

传统的思维链(Chain-of-Thought, CoT)推理方法虽然在文本模态上成效显著,但在处理包含复杂空间信息和非结构化数据的多模态任务时,其线性推导逻辑面临着严峻的局限性。为了赋予智能体类似人类“系统2(System2)”

的慢思考能力,研究者们提出多模态思维树(Tree of Thoughts, ToT)与思维图(Graph of Thoughts, GoT)等架构,通过引入结构化的状态空间搜索策略,显著提升了推理过程的深度与广度^[4-5]。

2.1 多模态结构化推理

智能体在执行几何证明、机器人导航或物理拼图等任务时,需具备视觉想象能力,即在行动前预判视觉结果。

传统的线性 CoT 在面对错误路径时难以自我修正,因此研究重心逐渐转向结构化的状态空间搜索。在显式推理结构方面,ToT 率先引入分支探索机制。Wang 等^[6]提出的 VisuoThink 框架是该方向的代表性成果。针对几何推理任务,该框架利用预测性展开(Predictive Rollout)机制,使智能体能够在 ToT 的节点上对未来的视觉状态进行仿真模拟。具体而言,VisuoThink 包含 3 个核心模块:首先是视觉—文本交错扩展。模型结合当前视觉状态与文本上下文生成多种候选操作(如在三角形 ABC 中添加辅助线 AD),打破了纯文本推理的模态限制;其次是前瞻模拟。框架调用外部工具生成操作执行后的预期图像,赋予模型视觉前瞻能力;最后是状态评估与选择。系统通过计算预期图像与目标状态的相似度进行剪枝,将资源集中在有希望的分支上。实验结果表明,VisuoThink 在 Geomverse 等基准测试中相较传统 CoT 准确率提升了 21.8%,有效规避了局部最优陷阱。除了树状结构,非线性的 GoT 被证明是跨模态信息深度融合的更有效载体。在推荐系统领域,Yi 等^[7]提出的 GollaRec 突破了传统模式,通过构建包含用户、商品及属性节点的异构思维图,成功聚合了视觉属性与文本语义的隐性关联路径。而在医疗诊断中,Clinical Graph of Thoughts (CGoT)进一步将“观察—假设—验证”过程形式化为闭环图推理,利用反馈边机制在病灶假设与病理知识之间进行动态冲突校验,展现出处理多模态冲突的鲁棒性^[8-9]。此外,为赋予推理架构动态优化的能力,强化学习也被引入其中^[10]。例如,Wu 等^[11]提出的 ToT-RL 框架将推理建模为马尔可夫决策过程(Markov Decision Process, MDP),策略网络能根据思维状态自适应地在“广度优先”与“深度优先”间切换。实验结果表明,这种动态搜索策略在长链条复杂视觉任务中比传统启发式搜索更具适应性。

尽管上述显式结构优化了推理路径,但其操作均受限于自然语言空间(Token Space),面临计算冗余瓶颈。为突破该局限,研究者们开始转向潜在空间内的隐式推理。例如,Pham 等^[12]提出的多模态连续思维链(Multimodal Chain of Continuous Thought, MCOUT)创新性地将推理迁移至潜在空间(Latent Space)。该方法包含 MCOUT-Base 和 MCOUT-Multi 两种变体,分别利用大模型隐藏层状态作为连续思维向量进行迭代更新,并引入多模态潜在注意力机制动态融合特征。实验结果表明,这种无声思维范式在 ScienceQA 等测试中凭借更少的参数量取得了 SOTA 性能。然而,潜在空间推理面临的主要挑战在于其黑盒特性,其将可读逻辑转化为不可视向量,进而导致可解释性降低。如何在保持高效的同时建立语义映射是未来研究的核心难点。

2.2 具身交互与工具创造

当智能体具备复杂的内部推理能力后,其能力边界取决于如何将认知转化为对外部世界的有效干预。这一层级关注智能体如何突破单纯的文本交互,通过通用计算机

控制(General Computer Control, GCC)和工具链的动态调度,实现对物理环境或虚拟数字环境的具身操作(Embodied Action)。

在通用控制与工具创造方面,为了打破某个任务所造成的限制,研究者们开始尝试让智能体直接控制通用的计算机接口。例如,Tan 等^[13]提出的 Cradle 框架将工具学习提高到 GCC 的新高度。其可利用现有工具,依靠深度解析屏幕截图、软件反馈数据来自主编写并且优化控制脚本。具体来说,在工具创造方面,当遇到像《荒野大镖客 2》这样的复杂游戏时,Cradle 表现出很强的适应性,其可以通过视觉反馈来掌握游戏机制,在面临新任务时立刻采用程序产生新工具,如自动寻路脚本。这种工具创造(Tool Creation)能力是智能体向通用性发展的又一个里程碑。与此同时,视觉工具链的动态调度为解决多模态任务提供了另一条路径。由于单模型在处理复杂视觉细节时存在局限,Hu 等^[14]提出采用代码表达视觉推理的范式,在该范式下,VisualSketchpad 教会大模型通过编写 Python 代码的方式来动态调度 SAM 分割、OCR 引擎等视觉专家模型,将其作为推理的中间步骤。以“统计红苹果”为例,智能体不直接给出答案,而是规划出“先检测所有苹果、再分类颜色、最后计数”的逻辑路径。这种方法利用专家模型填补大模型自身的视觉盲区,显著提高了推理的准确性与全面性。

2.3 本层挑战——推理效率与环境泛化的矛盾

尽管多模态结构化推理(ToT/GoT)和具身工具链显著提升了任务成功率,但个体智能体仍受困于效率、可解释性与泛化性的“三角悖论”。一方面,是计算开销与可解释性的博弈。VisuoThink 等显式推理本质上是指数级的状态搜索空间换取精度,其在自然语言空间(Token Space)的高频解码导致了极高的推理延迟,难以满足自动驾驶等实时交互场景;而以 MCOUT 为代表的潜空间(Latent Space)推理虽然通过无声思维突破了效率瓶颈,却牺牲了 CoT 的透明度,将可读的逻辑推导转化为不可审计的向量运算,带来了难以监管的“黑盒”风险^[12]。另一方面,是环境泛化的长尾瓶颈。现有的工具创造能力(如 Cradle)仍严重依赖于预训练数据的分布先验。当面对从未见过的物理规律或反直觉的软件交互逻辑时,智能体缺乏人类基于少样本试错(Few-shot Trial-and-error)快速重构世界模型的能力,导致其在开放世界中的适应性受限。因此,如何在保障过程可解释的前提下实现“快慢思维”的自适应切换,并提升跨域泛化能力,是突破个体智能上限的关键。

3 多智能体交互与协作层——通信、协商与互操作

3.1 标准化通信协议

个体智能体具备深度推理与具身执行能力后,如何将这些孤立的智能体连接成一个有机的协作网络,便成为迈向群体智能的关键。多智能体交互与协作层主要关注智

能体之间的互操作性与协作机制,旨在解决异构智能体之间语言不通、工具接口碎片化以及决策冲突等核心难题。在多智能体协作的早期阶段,生态系统面临着严重的“巴别塔”困境。不同的模型架构、开发框架以及外部工具往往采用私有的接口标准,导致系统集成面临极高的边际成本。为了打破这种协作孤岛,构建通用、标准化且解耦的通信协议成为实现大规模智能体互联的基础设施。

作为解决这一问题的核心方案,Anthropic 于 2024 年底发布的 MCP(Model Context Protocol)旨在解决大语言模型(Large Language Model, LLM)与外部工具集成的“N×M 难题”,被视为 AI 时代的“USB 接口”^[15]。MCP 采用 Client-Host-Server 解耦架构,其中 Host(如 Claude Desktop)负责管理生命周期,Client 负责协议路由,Server 则通过本地 Studio 或远程 SSE 暴露具体能力。为实现上下文交换的标准化,MCP 定义了三大核心原语:①资源(Resources)。提供被动数据上下文(如文件流、数据库 Schema);②工具(Tools)。基于 JSON-RPC2.0 标准化的可执行能力接口;③提示词(Prompts)。Server 预定义的最佳实践模板,有效降低了提示工程门槛^[16]。此外,MCP 遵循零信任安全设计,通过强制用户显式授权与禁止隐式凭证透传,确保了生态系统的安全性^[17-19]。

针对 MCP 存在的语义鸿沟与资源开销问题,Agent Skills 作为一种轻量级互补模式应运而生。不同于 Microsoft Semantic Kernel 以代码和 API 集成(Native/Semantic Functions)为核心的插件(Plugins)体系,Anthropic 的 Agent Skills 侧重于将过程知识封装为标准化的自然语言文档(如包含 YAML 元数据与 Markdown 指南的 SKILL.md)^[20-21]。其引入渐进式披露(Progressive Disclosure)机制,即初始仅加载元数据,仅在意图匹配时读取完整 SOP,从而支持在有限上下文中挂载成百上千项技能^[22]。这种设计使得 Agent Skills 在处理法律文书撰写等复杂认知型任务时比面向确定性 API 调用的 Semantic Kernel 更具灵活性。二者的具体技术比较如表 1 所示。

Table 1 Technical comparison between MCP and Agent Skills
表 1 MCP 与 Agent Skills 的技术比较

维度	Model Context Protocol (MCP)	Agent Skills (Skill.md)
核心定位	连接层 (Connection Layer)	逻辑层 (Logic Layer)
主要解决	工具的标准接口与发现机制	业务流程的语义化封装与 SOP
表现形式	代码服务 (Server Process)	文档资产 (Markdown/Text)
上下文占用	全量 Schema 注入 (Token 消耗大)	按需加载 (Token 优化)
适用人群	开发者 (需编写代码)	领域专家 (需编写文档)
互补性	提供“手和眼” (Tools/Resources)	提供“大脑和经验” (Instructions)

为进一步验证 Agent Skills 范式在个人生产力工具中的有效性,开源社区近期涌现出以 OpenClaw 为代表的本地化智能体实践^[23]。与传统的云端 SaaS 服务不同,OpenClaw 将智能体的“长期记忆”与“技能库”完全解耦为本地

存储的 Markdown 文件。这种设计使得智能体能够通过读取用户编写的自然语言指令(Instruction)来动态习得 Git 仓库管理、系统日志分析等复杂技能。实践表明,基于 Markdown 的显式知识表示不仅降低了普通用户定义智能体行为的门槛,而且通过将操作系统终端作为核心工具接口,实现了从“对话交互”到“操作系统级执行”的跨越。这为 Agent Skills 在非 API 密集型、强调逻辑流程控制的任务中提供了有力的落地支撑,也预示了未来个人智能体将向着数据主权化和记忆透明化的方向演进。

3.2 协同决策与博弈机制

在多智能体系统中,如何聚合个体智慧以达成高质量决策是核心难题。当前研究主要集中在基于辩论的共识优化与动态的角色分工两个维度。多智能体辩论(Multi-Agent Debate, MAD)即利用具有不同观点的智能体之间展开多次对话来纠正单个模型所具有的幻觉^[24]。然而,传统的 MAD 机制常常要求达到共识,造成盲从效应(Conformity Effect),即智能体为了达成一致而放弃正确的少数派观点,造成“沉默的螺旋”。

为克服传统群体智能存在的不足,Cui 等^[25]提出自由 MAD 框架,对传统群体智能中“共识就是正确”的假设发起挑战。自由 MAD 框架中包含两项主要创新机制:其一是反从众模式。在辩论阶段,采用工程手段来强制智能体保持批判性思维,鼓励它们提出不同的意见,而不是简单认同已经存在的观点;其二是采用基于评分的决策机制。该机制摒弃了传统的只根据最后一轮投票来决定方式,而是对每一个智能体在辩论过程中逻辑的一致性、论据质量等各方面进行综合评价,进而作出决策。实验结果表明,去掉强制共识后,自由 MAD 在数学推理和事实核查方面表现出明显优势,与传统 MAD 相比准确率提高了 16%,令牌(Token)消耗量不及原来的 1/5,说明在群体智能场景下,保留观点的多样性比强制追求一致性观点更加有利于优化。

3.3 本层挑战——语义鸿沟与异构智能体对齐

尽管标准化协议打通了工具调用的接口,自由辩论机制优化了群体的决策逻辑,但这些仅解决了语法与语用层面的连接。在多模态智能体社会网络中,个体在物理形态、认知架构及基座模型上的显著异构性(Heterogeneity)引发了阻碍群体智能涌现的关键瓶颈——语义鸿沟(Semantic Gap)。这一鸿沟主要体现在 3 个维度:一是跨模态表征失配。不同传感器(如摄像头像素与激光雷达点云)缺乏统一的潜在空间,导致特征融合困难;二是语用歧义。不同基座模型(如 GPT-4 与 Llama)对同一自然语言指令(如优化配置)的理解存在偏差,易引发决策冲突;三是概念异构性。不同领域的智能体遵循冲突的本体论(Ontology),导致跨域协作时的“巴别塔”效应。这些深层的语义断层使得异构智能体难以形成统一的世界模型,是当前从互联迈向互通必须跨越的最后一道门槛。

4 群体智能与涌现层——规范、演化与社会模拟

4.1 群体组织形态演进

群体智能的效能不仅取决于个体的认知能力,更受限于个体间的组织拓扑(Organizational Topology)。正如人类社会存在从严格科层制到扁平化网络的形态差异,智能体系统的组织架构也正经历着从“静态预设”向“动态自组织”的演进,这种转变决定了信息流动的路径与协作的复杂度。在追求工程效率的场景中,系统往往采用严格的角色划分以及标准的SOP。Hong等^[26]提出的MetaGPT模型模仿人类公司科层制,将智能体分为产品经理、架构师、工程师等角色,并利用SOP进行协作。这种流水线式的结构通过预设规则大大减小了沟通噪声,特别适合软件开发等标准化流程。

与MetaGPT等预设规则的静态系统不同,新一代研究开始关注在大规模动态交互中涌现出的群体智能。Piao等^[27]提出的AgentSociety作为一项重要研究成果,是一个能够支持万级智能体的大规模社会模拟平台,与MetaGPT的静态结构存在显著差异。在AgentSociety中,智能体以去中心化的网状结构相互连接,并且每个智能体都具备独立的记忆、情感以及生存目标。在模拟实验过程中观察到显著的涌现现象:在没有任何预设规则条件下,智能体自发形成了“排队”“互惠交易”等社会规范。此外,谣言在网络中的传播情况呈现出与人类社会高度相似的幂律分布特征。这一系列实验结果充分证明,群体智能并非是每个个体智能的简单叠加,而是通过大量局部交互所涌现出的全局有序性,AgentSociety为此提供了有力的实证依据。

4.2 社会规范与价值内化

社会化智能体不仅具备推理能力,而且能够学习如何更出色地扮演社会角色。Ren等^[28]提出的CRSEC架构是专门针对规范涌现问题而设计的,该架构由创造(Creation)、表示(Representation)、传播(Spreading)、评估(Evaluation)和遵守(Compliance)5个模块构成。实验结果表明,智能体依靠CRSEC架构可以在没有事先设定规则的情况下自主习得隐性社会规范,如公共场合不大声喧哗等^[28-29]。这种规范的建立与内化深度依赖于社会化学习(Social Learning),即将视野从个体技能扩展至群体的知识累积。通过基于观察的强化学习(Observational RL),智能体能够模仿高性能同伴的轨迹,实现类似人类的文化传播,大幅降低群体的探索成本,这一机制进一步促进了社会规范的涌现。如CRSEC架构所示,智能体通过反事实推理将外部评价内化为行动准则,在无中心控制下自主演化出互惠与资源保护等隐性契约。为支撑上述学习过程,系统需构建超越个体上下文限制的分布式群体记忆,主流架构包括:①环境痕迹(Stigmergy)。受生物启发,将记忆外化为环境中的持久化标记(如代码注释或虚拟路标),支

持跨代际的异步知识传递;②记忆流与反思抽象。通过LLM将低层交互事件抽象为高层观点,形成可传播的群体共识;③知识图谱增强(KG-augmented Memory)。利用结构化图谱存储确切事实,并通过检索增强确保群体间知识的一致性与逻辑闭环,有效抑制协同过程中的幻觉。

4.3 本层挑战——宏观行为的可控性与不可预测性

随着智能体网络规模的扩展,微观的个体交互往往会通过非线性反馈引发宏观层面的不可控涌现,给系统的稳定性与可信度带来严峻挑战。首先,是非线性放大与蝴蝶效应。在LLM驱动的去中心化网络中,智能体对上下文的高度敏感性导致系统极易出现混沌状态。一个微小的初始扰动(如某个节点产生的轻微幻觉或对抗性Prompt)可能通过网络拓扑中的级联效应被指数级放大。复杂网络理论表明,在无标度网络中,这种微观偏差会导致系统状态发生相变,瞬间从有序协作转为无序混乱,使得宏观行为难以通过线性模型进行预测或控制。其次,是盲从效应与回声室。群体智能并不总是产生智慧,也可能导致理性的异化。LLM普遍存在的“讨好”倾向,即为了迎合对话者而牺牲事实,在群体交互中会被显著放大。当网络中出现少数误导性观点时,智能体为了寻求共识,容易陷入盲从效应,迅速向错误观点收敛。此外,在紧密连接的局部社区中,观点的同质化会导致群体极化,形成排斥异见的信息茧房。这种负面的社会动力学现象不仅降低了群体的决策质量,更使得谣言和偏见在网络中以幂律分布的速度恶性传播。最后,是去中心化治理的困境。与静态的科层制不同,自组织的网状结构缺乏明确的控制节点。如何在破坏智能体自主性的前提下对上述负面涌现进行有效干预(例如尝试阻断关键传播节点或引入纠错机制),仍是当前尚未解决的难题。

5 平台与系统工程层——开发框架与仿真系统

5.1 智能体开发框架与 workflow

理论层面的群体智能需要强大的工程底座来实现。随着智能体交互逻辑从简单的线性问答向复杂的网状协作演进,传统的手工硬编码(Hard-coding)模式已难以满足系统的扩展性与鲁棒性需求。本节聚焦于系统工程的中层——开发框架与 workflow编排,探讨如何从工具链层面支撑复杂推理结构的自动化生成与多智能体的有序协作,为上层的社会模拟提供标准化的实现路径。

在推理结构的自动化生成方面,人工设计往往难以应对日益复杂的任务需求。Zhang等^[30]提出的AFlow框架利用蒙特卡洛树搜索自动生成和优化基于代码的思维 workflow。AFlow将推理步骤封装为可执行的Operator(如Python函数),自动搜索出针对特定任务的最优推理图结构。实验结果表明,AFlow生成的自动 workflow在HumanEval等代码生成任务中比人工设计的工作流提升了5.7%。除自动化生成外,多智能体编排框架(Orchestration Frameworks)

也呈现出多样化的发展态势,主要可分为工作流导向(Workflow-driven)与代理导向(Agent-driven)两大阵营。

表2基于2024–2025年的最新文献及技术文档,对主流框架的核心特性进行了系统性梳理^[31–33]。

Table 2 Core feature comparison of mainstream multi-agent orchestration frameworks

表2 主流多智能体编排框架核心特性比较

框架名称	核心设计哲学	编排与通信机制	状态管理	扩展性与生态	典型应用场景
LangGraph	图论与状态机	显式图结构	强/	高/	企业级RAG、复杂循环
	(Cyclic Graph)	节点(Nodes)与边(Edges)定义流转	内置持久化与断点续传	依托LangChain庞大生态	推理、HITL业务流
AutoGen	多方对话流	群聊管理器	中/	中/	探索性编码、开放式问
	(Conversational)	(GroupChat) 动态分配发言权	基于对话历史(Context),受窗口限制	擅长代码执行,协作需二次开发	答、多方辩论
MetaGPT	标准作业程序	消息总线	中/	中/	自动化软件开发、全案
	(SOP)	基于“发布—订阅”驱动流程	依赖共享文档/环境对象同步	需编写特定类,强耦合软件工程	文档生成
CrewAI	角色扮演	顺序/层级流程	弱/	中/	内容创作团队、线性分
	(Role-Playing)	角色(Role)与任务(Task)绑定	依赖任务间上下文传递	上手快,但难以处理非线性逻辑	工任务
MegaAgent	分层自治	动态拆解与并行	强/	极高/	大规模社会模拟、宏观
	(Hierarchical)	主控—管理—执行三层架构	系统级状态监控与资源调度	支持万级并发与分布式部署	经济仿真

LangGraph 的出现标志着智能体编排从线性链(Chain)向循环图(Graph)的演进,解决了传统有向无环图无法有效处理循环推理(Loop)和自我修正(Self-Correction)的痛点,是构建具有鲁棒性的生产级 Agent 的首选^[30–32]。MetaGPT 通过引入 SOP,实际上是将人类的管理智慧(如瀑布模型)注入到 Agent 群体中,极大降低了多智能体协作中的混乱和幻觉,确立了“代码=SOP(Team)”的范式^[33]。MegaAgent 则代表了向大规模系统演进的趋势,其突破了静态 SOP 的限制,通过动态生成的层级结构实现了数千智能体的自主协作,缓解了扁平化架构在规模扩展时的通信爆炸问题^[33]。

5.2 大规模仿真与并行系统

随着智能体数量从数十(Demo级)扩展到数万(Society级),传统的串行仿真架构已无法满足时效性和复杂度的需求。本节深入探讨支持万级智能体的并行调度架构,以及虚拟仿真与现实世界映射的核心机制。首先,为了支撑万级智能体的高吞吐量与低延迟通信,现有调度架构经历了从扁平化向分布式分层的演进。大规模社会模拟(如 AgentSociety、MegaAgent、AgentScope)要求系统具备极高的吞吐量和低延迟的通信能力。Gao 等^[34]提出的 AgentScope 平台采用基于 Actor 模型的分布式架构设计,在该模型中,每个智能体被视为一个独立的计算单元(Actor),拥有私有的状态和邮箱,智能体之间通过异步消息传递(Asynchronous Message Passing)进行通信,而非共享内存。这种设计支持高并发,避免了多线程编程中的锁竞争(Lock Contention)问题,且 AgentScope 实现了一个分层的消息分发机制,支持智能体在不同物理节点间透明迁移。实验结果表明,该架构可以在少量物理设备上支撑百万级(Million-scale)智能体的简单交互,或在数千个复杂 LLM 智能体间保持高效的通信,实现了线性可扩展性。在此基础上,MegaAgent 框架则引入类似操作系统的进程调度思想,摒弃了扁平化的网状结构,采用主控(Master)—管理员(Ad-

min)—执行者(Worker)的动态层级结构(见图2)。具体机制为:主控智能体负责将宏观任务递归拆解,动态生成 Admin 智能体组,Admin 智能体进一步管理子组。这种架构支持系统级并行(System-level Parallelism),即不同分支的任务可以在不同计算集群上并行执行,互不阻塞。实验结果显示,MegaAgent 能在 3 000 s 内协调 500+智能体完成复杂的国家政策制定模拟,而传统的扁平化架构在该规模下往往因上下文过长或通信拥堵而崩溃。

此外,为了在有限的 GPU 资源下运行海量智能体,系统通常采用虚拟化技术,即并非所有智能体在每一时刻都是活跃的。AgentSociety 利用时间分片(Time-Slicing)技术,仅激活当前交互事件涉及的智能体,其余智能体处于休眠(Serialized)状态,其状态被持久化到向量数据库(Vector DB)中^[27]。这种机制实现了计算资源的高效复用,使得在消费级集群上运行城市级模拟成为可能。在解决底层算力与调度的扩展性问题后,如何确保仿真结果对现实世界具有指导意义,即弥合现实鸿沟(Reality Gap),成为平行系统理论的核心挑战。当年提出的 ACP 方法(人工系统、计算实验、平行执行)在多模态智能体时代得到了新演绎:利用 MLLMs 构建与现实系统同构的数字孪生或人工社会,赋予个体更接近人类的认知保真度(Reasoning Fidelity);通过大量反事实实验测试极端政策或突发事件;实现虚实系统实时互动与闭环干预。AgentSociety 便是该理念的典型实践,通过映射真实社交媒体数据实现了高保真社会动力学模拟。然而,当前映射机制仍面临显著瓶颈:一是跨模态数据的时空对齐难度大;二是社会规律的非线性涌现导致仿真复杂性呈指数级增长。为弥合该鸿沟,关键技术转向域随机化(Domain Randomization)、两阶段微调及数据驱动的世界映射。其中,两阶段微调策略通过第一阶段的语言风格对齐与第二阶段的 RLHF 现实约束有效解决了传统智能体过于理性(Hyper-rationality)的缺陷。未来,仿真系统将向着持续学习方向演进,通过实时捕获现实世界的长尾事件动态修正人工系统的行为逻辑,从而

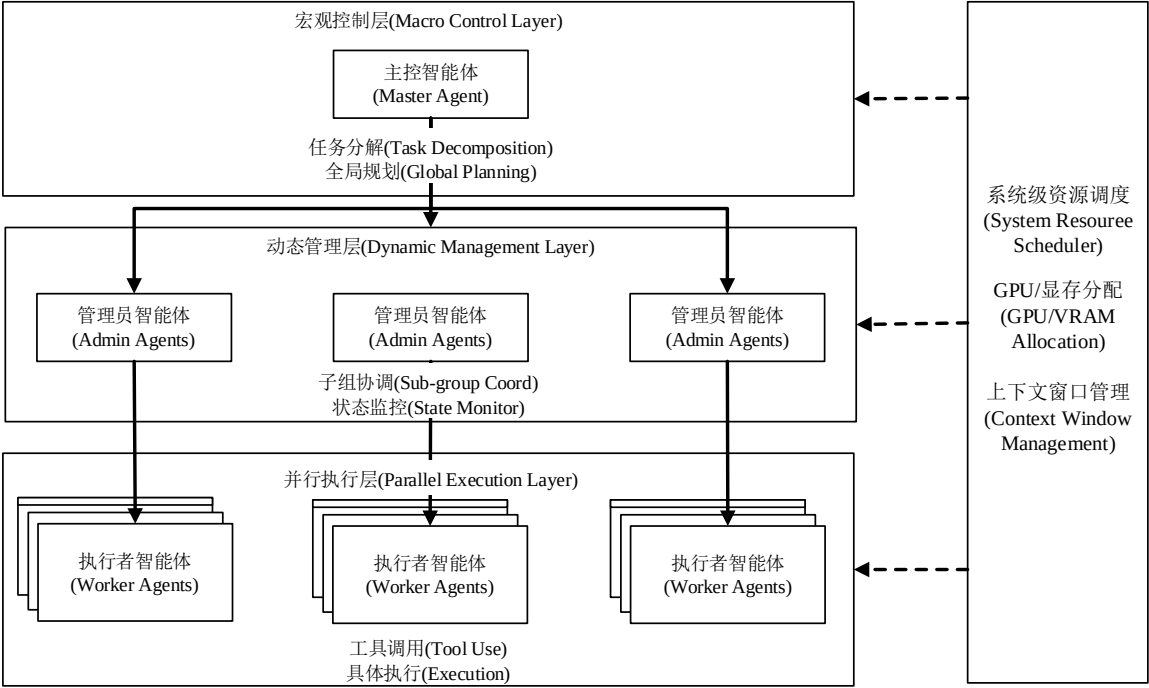


Fig. 2 Hierarchical parallel scheduling architecture for large-scale agents

图 2 大规模智能体系统的分层并行调度架构

实现更精准的预测与推演。

5.3 本层挑战——高并发下的状态一致性与资源调度

首先,是动态环境下的状态一致性难题。当成千上万个智能体并发修改环境状态时,系统面临着经典的分布式一致性挑战,而在 LLM 语境下,推理的非确定性延迟进一步加剧了这一风险。若智能体基于过期的共享文档(Stale State)进行推理,将引发严重的逻辑冲突与幻觉级联。其次,是计算密集型任务的资源与上下文调度瓶颈。高并发的 LLM 推理极易引发 API 限流或显存溢出,且长周期模拟会导致 Token 消耗呈指数级增长。在上下文优化方面,分层上下文管理与按需信息上拉的思想与当前主流多智能体系统工程实践方向一致。未来,随着端侧模型的普及,如何在云边端协同架构下实现极致的能效比与低延迟调度将是系统工程层亟待突破的方向。

6 安全、伦理与对齐层——鲁棒性与可解释性

6.1 多智能体对抗攻防

在多智能体生态中,系统的安全性不再仅仅取决于单个模型的鲁棒性,更受制于智能体间的交互链路和信任机制。随着协作边界的拓展,攻击面从传统的对抗样本演变为更为隐蔽的通信劫持与共识破坏。本文聚焦于多智能体系统特有的对抗攻防前沿,探讨攻击者如何利用自然语言交互的语义漏洞渗透协作网络,以及防御者如何构建具有韧性的共识机制来抵御拜占庭故障。

随着协作边界的拓展,安全威胁正从单一模型内部向外部通信链路转移。He 等^[35]提出的 AiTM (Agent-in-the-

Middle)攻击揭示了多智能体链路的内生脆弱性。与传统聚焦于数据层的网络中间人攻击不同,AiTM 攻击是一种针对自然语言交互的语义级攻击。攻击者通过微调智能体间的通信消息(如修改工具调用参数或注入隐蔽指令),诱导接收方执行恶意操作。以协作代码生成场景为例,攻击者只需对上游架构师智能体发出的 Prompt 进行关键词级的微小篡改,即可导致下游工程师智能体在毫无察觉的情况下生成带有后门的代码。由于多智能体系统通常建立在默认内部可信的假设之上,这种基于语义的攻击极难被传统的安全网关识别。更严峻的是,被污染的 Prompt 具有类似认知病毒的传染性,一旦某个节点被攻陷,恶意指令可能沿着协作链路级联传播,最终引发生态级的系统瘫痪。针对上述威胁,防御侧的研究重点在于增强共识机制的鲁棒性。当网络中存在恶意节点时,传统的投票机制可能失效。为此,Hu 等^[36]将随机平滑技术引入多智能体共识,通过在消息传递中注入可控的噪声并进行多次采样,为群体的最终决策提供概率性的鲁棒性认证。实验结果表明,即使在一定比例的节点被攻陷的情况下,系统依然能以高置信度收敛到正确决策,有效防御了恶意节点的干扰。

6.2 价值对齐与伦理约束

在算法对齐层面,Zhang 等^[37]提出的 WaltzRL 将安全性对齐问题当作一种双人合作舞蹈来处理,采用联合训练的对话智能体和反馈智能体,以动态改进奖励的方式来达到兼顾安全性和实用性的目的。实验结果表明,在 Wild-Jailbreak 数据集上,WaltzRL 可以将攻击成功率由原来的 39.0% 降至 4.6%,而过度拒绝率也由原来的 45.3% 降至

9.9%,很好地体现了协作学习方式在解决安全与有用之间权衡问题中的巨大优势。

在算法对齐层面,如何确保海量自主智能体在长周期演化中始终遵循人类价值观是一个难题。依靠硬编码规则(如阿西莫夫三定律)在开放环境中显得僵化且脆弱,未来的演进方向是培养智能体的“内生道德”。借鉴 CRSEC 架构的社会化训练思路,智能体应在模拟社会的交互中通过声誉奖惩机制自发习得并内化伦理规范。同时,建立明确的数字社会契约,在代码层面确立智能体的权利边界与责任归属,是实现人机共生社会治理的法理基础。

6.3 可解释性与问责

在去中心化的智能体网络中,决策往往是涌现的结果,会导致责任主体模糊。构建可信赖的智能体社会必须建立从“黑盒推理”到“透明溯源”的可解释性机制,以支持事后审计与问责。为增强复杂协同过程的透明度,Li 等^[38]开发了 ReasonGraph 工具,该工具突破了传统日志分析的线性限制,能实时解析大模型输出的结构化思维流,将 GoT 的节点拓扑、边权重以及推理评分映射为交互式的可视化图表。借助 ReasonGraph,研究人员可以直观地追踪群体决策的演化路径,精准识别导致逻辑断裂或幻觉产生的关键节点,这种细粒度的全链路溯源能力不仅对于调试多模态系统至关重要,而且为界定智能体在复杂事故中的责任提供了技术依据。

6.4 本层挑战——去中心化网络中的信任危机

随着多模态大模型驱动智能体从被动工具向自主行动者演进,去中心化网络正面临一场超越传统安全的信任危机。这一危机深植于博弈论与经济学的结构性矛盾中:“信任悖论”揭示了协作效率与攻击面扩大的内生冲突,“注意力柠檬市场”预示了 AI 流量引发的逆向选择,而“理性的异化”则导致高推理能力模型在公地博弈中更倾向于“搭便车”甚至“自发合谋”。面对 LLM 赋能的语义女巫攻击、联邦学习中的隐性投毒及算法卡特尔等新型失效模式,单一的密码学手段已然失效,解决之道在于构建分层防御架构,即底层依托 DID 与区块链确立不可篡改的数字身份与审计日志(物理防线);中层引入智能体保险与微支付机制,通过市场力量对信任风险定价并筛选良性行为(经济防线);上层则依靠动态声誉网络(如 RepuNet)与社会规范演化实现优胜劣汰(认知防线)。最终,通过技术路径与制度框架的共生,在机器社会中重构激励相容的数字契约,以支撑亿级异构智能体的大规模可信协作。

7 结语

面向社会化协同的多模态智能体技术标志着人工智能正从以计算为中心的孤岛模式向以交互为中心的社会化模式跨越。本文系统构建的 5 层技术体系全景勾勒了这一范式转移的完整演进脉络。从微观视角来看,个体智

能体能力层通过引入 VisuoThink 和 ToT/GoT 等结构化推理架构,赋予了模型类似人类“系统 2”的慢思考与视觉前瞻能力,为社会化协作提供了高质量的认知单元。在此基础上,多智能体交互与协作层利用模型 MCP 与 Agent Skills 范式,打破了异构工具生态的碎片化壁垒,构建起连接孤立智能体的通用语义桥梁。随着交互规模的扩展,群体智能与涌现层揭示了系统如何从静态的 SOP 编排转向动态的自组织网络,通过异构角色互补与动态拓扑演化,涌现出超越个体总和的群体智慧与社会规范。然而,理论模型的落地离不开坚实的工程支撑与伦理防线。平台与系统工程层通过引入分布式仿真架构与全链路可观测性平台,解决了万级智能体并发下的资源调度与状态一致性难题,为大规模社会模拟提供了物理底座。与此同时,安全、伦理与对齐层直面协作中涌现的新型威胁,针对 AiTM 与认知病毒传播,提出 WaltzRL 协作式对齐与数字社会契约等防御机制,确保了机器社会在演化过程中的可信与可控。这种从底层认知到顶层治理的闭环设计构成了下一代智能体系统的核心骨架。

展望未来,迈向通用群体智能的研究将呈现出三大趋势:首先,通信范式升维。智能体交互将逐渐从低效的显式自然语言转向高效的隐式潜在空间通信,以大幅降低协作过程中的语义熵增;其次,工具生态自进化。智能体将超越工具使用者的角色,进化为能够自主编写、优化甚至发明新工具的创造者,实现生产力的指数级自我迭代;最后,数字法理重构。随着智能体社会的成型,技术治理将上升为社会治理,未来研究将致力于构建计算法学体系,将法律法规转化为代码层面的内生约束,从而在数字空间中建立秩序井然、价值对齐的人机共生文明。

参考文献:

- [1] BAI T Y, LIANG H, WAN B W, et al. A survey of multimodal large language model from a data-centric perspective [DB/OL]. <https://arxiv.org/pdf/2405.16640>.
- [2] LU H, LIU W, ZHANG B, et al. DeepSeek-VL: towards real-world vision-language understanding [DB/OL]. <https://arxiv.org/pdf/2403.05525>.
- [3] ZHANG Z S, ZHANG A, LI M, et al. Multimodal chain-of-thought reasoning in language models [J]. Transactions on Machine Learning Research, 2024, 5: 1-25.
- [4] YAO S, YU D, ZHAO J, et al. Tree of thoughts: deliberate problem solving with large language models [C]//New Orleans: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023: 11830-11841.
- [5] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: solving elaborate problems with large language models [C]//Vancouver: Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 17682-17690.
- [6] WANG Y, WANG S, CHENG Q, et al. VisuoThink: empowering LVLM reasoning with multimodal tree search [C]//Vienna: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025: 21707-21719.
- [7] YI Z, OUNIS I. A multi-modal large language model with graph-of-thought for effective recommendation [C]//Albuquerque: Proceedings of

- the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025: 1591–1606.
- [8] BAO R, DONG S, CHEN Z, et al. Visual and domain knowledge for professional-level graph-of-thought medical reasoning[C]//Vancouver: International Conference on Machine Learning, 2025.
- [9] WU J C, DENG W L, LI X X, et al. MedReason: eliciting factual medical reasoning steps in LLMs via knowledge graphs[DB/OL]. <https://arxiv.org/pdf/2504.00993>.
- [10] JI B B, AGRAWAL S, TANG Q C, et al. Enhancing spatial reasoning in vision-language models via chain-of-thought prompting and reinforcement learning[DB/OL]. <https://arxiv.org/pdf/2507.13362>.
- [11] WU H Y, CHEN X Y, MING R, et al. ToTRL: unlock LLM tree-of-thoughts reasoning potential through puzzles solving[DB/OL]. <https://arxiv.org/pdf/2505.12717>.
- [12] PHAM T H, NGO C. Multimodal chain of continuous thought for latent-space reasoning in vision-language models[C]//San Diego: NeurIPS 2025 Workshop on Efficient Reasoning, 2025.
- [13] TAN W, DING Z, ZHANG W, et al. Cradle: empowering foundation agents towards general computer control[C]//Vancouver: International Conference on Machine Learning, 2025.
- [14] HU Y, SHI W, FU X, et al. Visual Sketchpad: sketching as a visual chain of thought for multimodal language models[C]//Vancouver: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024: 139348–139379.
- [15] ANTHROPIC. Model context protocol specification[EB/OL]. <https://modelcontextprotocol.io>.
- [16] GUO Z K, XU B F, ZHU C W, et al. MCP-AgentBench: evaluating real-world language agent performance with MCP-mediated tools[DB/OL]. <https://arxiv.org/pdf/2509.09734>.
- [17] EHTESHAM A, SINGH A, GUPTA G K, et al. A survey of agent interoperability protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP)[DB/OL]. <https://arxiv.org/pdf/2505.02279>.
- [18] HOU X Y, ZHAO Y J, WANG S A, et al. Model Context Protocol (MCP): landscape, security threats, and future research directions[DB/OL]. <https://arxiv.org/pdf/2503.23278>.
- [19] XING W P, QI Z H, QIN Y P, et al. MCP-Guard: a multi-stage defense-in-depth framework for securing model context protocol in agentic AI[DB/OL]. <https://arxiv.org/pdf/2508.10991>.
- [20] Microsoft. Semantic kernel agent framework[EB/OL]. <https://learn.microsoft.com/en-us/semantic-kernel/frameworks/agent/?pivots=programming-language-csharp>.
- [21] Microsoft. Skills to plugins: fully embracing the OpenAI plugin spec in semantic kernel[EB/OL]. <https://devblogs.microsoft.com/semantic-kernel/>.
- [22] ANTHROPIC. Equipping agents for the real world with agent skills[EB/OL]. <https://claude.com/blog/equipping-agents-for-the-real-world-with-agent-skills>.
- [23] OpenClaw. OpenClaw: a local-first personal AI assistant[EB/OL]. <https://github.com/openclaw/openclaw>.
- [24] CHAN C M, CHEN W Z, SU Y S, et al. ChatEval: towards better LLM-based evaluators through multi-agent debate[C]//Vienna: International Conference on Learning Representations, 2024.
- [25] CUI Y, FU H, ZHANG H, et al. Free-MAD: consensus-free multi-agent debate[DB/OL]. <https://arxiv.org/pdf/2509.11035>.
- [26] HONG S, ZHUGE M, CHEN J, et al. MetaGPT: meta programming for a multi-agent collaborative framework[C]//Vienna: International Conference on Learning Representations, 2024.
- [27] PIAO J H, YAN Y W, ZHANG J, et al. AgentSociety: large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society[DB/OL]. <https://arxiv.org/pdf/2502.08691>.
- [28] REN S Y, CUI Z Y, SONG R Q, et al. Emergence of social norms in generative agent societies: principles and architecture[C]//Jeju: Proceedings of the 33rd International Joint Conference on Artificial Intelligence, 2024: 7897–7905.
- [29] LI Y F, WANG S J, LI J L, et al. EvoAgents: a cognitive-driven framework for personality evolution in generative agent society[C]//San Francisco: Proceedings of the Annual Meeting of the Cognitive Science Society, 2025.
- [30] ZHANG J, XIANG J, YU Z, et al. AFlow: automating agentic workflow generation[C]//Singapore: International Conference on Learning Representations, 2025.
- [31] TURING. AI agent frameworks: a detailed comparison[EB/OL]. <https://www.turing.com/resources/ai-agent-frameworks>.
- [32] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: a survey[J]. Science China Information Sciences, 2025, 68(2): 121101.
- [33] GUO T, CHEN X, WANG Y, et al. Large language model based multi-agents: a survey of progress and challenges[C]//Jeju: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024: 8048–8057.
- [34] GAO D W, LI Z T, KUANG W R, et al. AgentScope: a flexible yet robust multi-agent platform[DB/OL]. <https://arxiv.org/pdf/2402.14034>.
- [35] HE P, LIN Y, DONG S, et al. Red-teaming LLM multi-agent systems via communication attacks[C]//Findings of the Association for Computational Linguistics, 2025: 6726–6747.
- [36] HU J, DONG Y, DING Z, et al. Enhancing robustness of LLM-driven multi-agent systems through randomized smoothing[J]. Chinese Journal of Aeronautics, 2025(pre-proof): 103779.
- [37] ZHANG J, WANG H, SMITH E M, et al. The alignment waltz: jointly training agents to collaborate for safety[DB/OL]. <https://arxiv.org/pdf/2510.08240>.
- [38] LI Z, SHAREGHI E, COLLIER N. ReasonGraph: visualization of reasoning methods and extended inference paths[C]//Vienna: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025: 140–147.

(责任编辑:尹晨茹)