面向可信增强的AI智能体研究综述

郑子彬[1,2]，陈玉轩[1]，张泽凯[1]，李卫[1]，刘井平[1]
（1.中山大学软件工程学院，广东 珠海 519082；2.中山大学深圳研究院，广东 深圳 518057）

摘要：针对智能体在自主决策、环境交互和任务执行过程中面临的决策可靠性不足、安全风险、行为失控及责任难以追溯等可信挑战，对可信智能体相关研究进展进行系统性综述。首先，梳理智能体的基础组件和典型智能体框架，并对比了智能体和大模型的不同之处。然后，针对智能体面临的可信问题，从正确性与可靠性、鲁棒性、治理与可控等7个方面进行了系统性定义。围绕智能体各基础组件以及可信验证、多智能体协作，对现有可信增强技术进行了系统性梳理与分析。此外，为评估智能体在不同维度下的可信表现，总结了现有评测基准与相关评测指标，同时分析了现有评测基准的局限性。最后，对可信智能体未来的发展方向进行了探讨与展望。

关键词：大语言模型；智能体；人工智能；可信；可信增强
中图分类号：TP309
文献标志码：A
doi：10.11959/j.issn.1000

A Survey of AI Agents for Trustworthiness Enhancement

Zheng Zibin[1,2], Chen Yuxuan[1], Zhang Zekai[1], Li Wei[1], Liu Jingping[1]
1. School of Software Engineering, Sun Yat-sen University, Zhuhai 519082, China
2. Shenzhen Research Institute of Sun Yat-sen University, Shenzhen 518057, China

Abstract: To address the trustworthiness challenges faced by agents during autonomous decision-making, environmental interaction, and task execution, including insufficient decision reliability, security risks, loss of behavioral control, and difficulty in accountability, research progress on trustworthy agents was systematically reviewed. First, the fundamental components and representative agent frameworks were reviewed, and the distinctions between agents and large language models were compared. Then, the trustworthiness challenges of agents were systematically defined from seven dimensions, including correctness and reliability, robustness, governance, and controllability. Existing trustworthiness enhancement techniques were further reviewed and analyzed with respect to the fundamental components of agents, trustworthy verification, and multi-agent collaboration. In addition, to evaluate the trustworthiness of agents across different dimensions, existing benchmarks and evaluation metrics are summarized, and the limitations of current benchmarks were analyzed. Finally, future research directions for trustworthy agents are discussed and envisioned.

Key words: large language model, agent, artificial intelligence, trustworthiness, trustworthiness enhancement

0 引言

大语言模型（LLM）的快速发展推动了AI智能体的兴起。大模型通常指基于Transformer架构的深度神经网络，最初用于对话系统等NLP任务。

收稿日期：2026–06–XX；**修回日期：**XXXX–XX–XX
通信作者：刘井平，liujp68@mail.sysu.edu.cn
基金项目：国家自然科学基金资助项目（No. 62306112）；广东省基础与应用基础研究基金(No. 2026A1515010253)；深圳市科技计划资助（JCYJ20240813150107010）；深圳市哲学社会科学规划课题（No. SZ2025A002）
Foundation Items: The National Natural Science Foundation of China (No. 62306112), Guangdong Basic and Applied Basic Research Foundation, China (No. 2026A1515010253), Shenzhen Science and Technology Program (JCYJ20240813150107010), Shenzhen Philosophy and Social Science Planning Project (Grant No. SZ2025A002).

随着应用场景拓展[1]-[5]，研究者探索使大模型更有效地与外部环境交互，催生了AI智能体概念。智能体是能够通过感知、推理、决策与执行过程与环境交互，以实现预定目标的自主系统。AI智能体则以大模型为核心“大脑”，结合存储模块、工具调用接口等增强组件构成[6]，多个智能体协同可进一步构建多智能体系统[6],[7]，以应对代码生成、医疗、教育等复杂任务场景。以开源框架OpenClaw为代表，其通过网关、多终端节点及多模型协同架构，实现了跨平台消息接入与任务自动化，体现了智能体在工具集成与环境交互方面的趋势。

然而，从大模型向AI智能体的演进，使AI系统从信息处理工具扩展为具备环境感知与执行能力的自主系统[7]，这一范式迁移在赋予模型改变现实世界能力的同时，也放大了安全风险。OpenClaw的广泛应用暴露出多重隐患：据Meta AI一位主管反映，在使用OpenClaw扫描邮箱并请求归档或删除建议时，系统意外开始批量删除邮件，且无法通过自身机制终止，直至手动停止主机所有进程；另有用户利用OpenClaw自动购物后账户遭盗刷，造成数百美元损失。这些事件表明，“可信”已从伦理选项上升为AI智能体进入医疗、金融、法律等高风险领域的前提条件。

“可信人工智能（trustworthy AI）”是一个多维概念，核心在于确保AI系统全生命周期表现符合人类预期、法律规范及道德准则。早在大模型诞生初期，国际社会便开始系统界定AI的“可信”内涵。经济合作与发展组织（Organisation for Economic Co-operation and Development, OECD）于2019年确立五项核心原则，初步形成了关于负责任和可信AI治理的全球共识。同年，欧盟委员会（European Commission）发布的《可信AI伦理指南》提出了AI系统被视为“可信”所必须满足的七项关键要求，强调可信AI不仅须遵守法律与道德，还应在技术上具备鲁棒性。美国国家标准与技术研究院（National Institute of Standards and Technology, NIST）于2023年发布的《人工智能风险管理框架》（AI RMF 1.0）指出，可信AI应具备有效性、安全性、鲁棒性等七大特征。同年，国际标准化组织（International Organization for Standardization, ISO）发布的ISO/IEC 42001标准也提出了类似的原则性要求。2024年，欧盟进一步发布《AI法案》，根据风险层级对AI系统进行分类，并明确了开发者和使用者各自承担的责任[8]。表1对不同机构在其框架内定义的可信AI核心要素进行了汇总。

当上述通用原则应用于AI智能体时，内涵需进一步深化。传统大模型主要引发输出层面的认知误导风险，而AI智能体因具备文件读写、代码执行及API调用等实际执行权限，面临的可信挑战从内容安全扩展至行动安全，威胁更为严峻。Yu等人[9]在TrustAgent框架中以模块化视角综述了智能体可信技术，但仅局限于少数攻击、防御与评估技术，未能系统构建可信性保障的维度与技术体系。

本文将可信智能体（trustworthy agent）定义为：在复杂动态环境中，能够持续稳定地保持其决策逻辑与人类意图对齐，在利用外部工具执行任务时严格遵守预设安全边界，并能有效抵御针对其感知、推理与记忆模块的恶意攻击的自主系统[9]。构建可信LLM智能体需从以下维度进行系统设计：（1）正确性与可靠性；（2）鲁棒性；（3）治理与可控；（4）可解释性；（5）安全性；（6）隐私与合规；（7）公平性与公正性。其中，正确可靠和鲁棒性是基础，确保任务执行的稳定性与准确性；治理与可控通过人类监督、权限分级及审计链路防止系统失控；可解释性要求决策依据绑定至可验证证据链，避免虚假叙述；安全性需从内容防护扩展至行动防护，通过输入隔离阻断提示注入等风险；隐私与合规要求落实数据最小化与可审计策略，以应对记忆与交互带来的暴露面扩大问题；公平性与公正性则要求在内容输出与资源分配中消除系统性偏见，并将其融入监测闭环。

上述维度相互支撑，共同构成AI智能体的可信体系框架。本文围绕这些维度，对大模型驱动的智能体在记忆、工具使用及规划与决策模块[6]的可信保障关键技术进行系统综述，并纳入可信验证机制及多智能体系统中的可信协作技术。全文结构为：第1节阐述智能体定义、基本组成及典型框架；第2节探讨可信智能体内涵及关键维度；第3节聚焦各模块可信性保障技术；第4节梳理评测体系与基准；第5节总结挑战并展望未来方向。

表1　不同机构对可信AI核心要素的定义

提出时间	机构（框架）	可信核心要素	重点关注领域
2019	OECD （五项核心原则）	包容性增长、以人为本、透明度、安全性、问责制	国际政策对齐与民主价值观
2019	欧盟委员会（可信AI伦理指南）	人类能动性与监督、技术鲁棒性与安全、隐私与数据治理、透明度、多样性/非歧视/公平性、社会与环境福祉、问责制	伦理原则落地与自我评估清单
2023	NIST （AI RMF）	有效性、安全性、鲁棒性、透明性、可解释性、隐私性、公平性	系统工程与风险缓解路径
2023	ISO（ISO/IEC 42001标准）	管理体系框架、透明性与伦理、风险与影响评估、数据安全与隐私、持续改进、问责制	管理体系整合与持续改进
2024	欧盟 （AI法案）	基本权利保护、风险等级分类、人类监管、可追溯性	法律合规性与市场准入壁垒
2025	中国科技大学（TrustAgent框架）	内在可信（大脑、记忆、工具）、外在可信（用户、智能体间、环境）	智能体模块化架构下的安全技术
2026	**本文**	**正确性与可靠性、鲁棒性、治理与可控、可解释性、安全性、隐私与合规、公平性与公正性**	**确保AI智能体可信的各层次关键技术**

1　AI智能体基础与范式

1.1　AI智能体的定义与基本组成

AI智能体是一种以大模型为核心推理引擎的计算系统，通过整合规划、记忆与工具使用等关键模块，能够在复杂动态环境中进行自主逻辑推理、制定复杂计划并采取具体行动以达成既定目标[6]，[10]，[11]。其运作模式可描述为“感知—推理—执行”循环：智能体从物理或数字环境中获取观测信息，结合存储的记忆进行逻辑推理，制定行动策略，并借助外部工具作用于环境以产生反馈[10],[11]。AI智能体通常由以下四个核心模块构成：

核心引擎：大语言模型。大语言模型是智能体的核心，承担语义理解、知识检索、意图识别及指令生成等关键功能[12]，其能力水平直接决定智能体可处理任务的复杂度上限。在语义理解与指令遵循方面，大模型通过对复杂语境的深度解析，将模糊任务描述转化为可执行的结构化指令，并在多轮交互中保持意图的一致性与连贯性。在多步推理与逻辑分解方面，大模型支撑复杂任务的逻辑拆解与多步规划，以思维链（Chain-of-Thought, CoT）为代表的推理技术通过显式呈现中间推理步骤，引导模型将复杂问题拆解为连续的推理链条，引导模型逐步规划驱动行动，提升复杂任务完成能力[10]。在控制指令生成与执行驱动方面，大模型将高层决策映射为结构化控制指令（如函数调用、执行命令或交互协议），驱动外部系统或内部模块执行具体操作，通过结构化输出约束和上下文对齐确保指令的语法语义一致性与可执行性。

记忆模块。记忆模块由上下文历史记录与外部存储系统构成，其关键作用在于保障智能体在超长多轮对话中保持一致的行为表现[13]，同时支持其自我进化能力。从架构角度，记忆可分为短期记忆与长期记忆两个层级。短期记忆通常指当前会话或任务流中的上下文信息，依赖大模型的注意力机制在每轮交互中被读取，会话结束或超出上下文窗口限制后即丢失。长期记忆则通常基于外部向量数据库或知识图谱构建，智能体在外部介质中维护对应的工作空间，存储历史交互日志、执行经验、专业知识乃至用户画像等持久化信息。当面临新任务时，智能体通过检索增强生成机制将相关信息引入上下文窗口。该机制不仅增强了长时程任务建模能力，也为个性化服务和持续进化奠定了技术基础。

规划模块。规划模块负责将抽象复杂的目标分解为一系列具体可管理的子任务，并能根据实时反馈动态调整后续方案，是智能体实现长任务执行的关键。规划过程通常包含三个阶段：任务分解阶段，智能体利用大模型将宏观目标拆分为若干逻辑关联的子任务，分解方式可以是顺序、并行或递归的；执行路径搜索与选择阶段，智能体生成多条候选路径，通过思维树（Tree-of-Thoughts, ToT）或蒙特卡洛树搜索（Monte Carlo Tree Search, MCTS）等算法结合大模型的启发式信息筛选最优执行轨

迹。其中，ToT强调在多个候选推理分支之间进行探索、评估与回溯，MCTS则通过模拟采样和树搜索在不确定状态空间中逐步逼近更优决策路径，二者均用于缓解单一路径推理可能带来的局部最优和错误累积问题。相关研究表明将树搜索、世界模型与环境状态建模与大模型结合有助于改善长程推理与交互式任务中的规划表现；反思与迭代修正阶段，智能体捕捉环境反馈（如工具报错、任务超时），触发反思机制分析错误原因并更新后续规划，形成“尝试-反馈-调整”的自适应闭环，典型方法包括ReAct[14]通过交错生成推理与行动来持续调整计划，Reflexion[15]则利用语言形式的自我反馈与经验积累改进后续决策。此外，部分智能体还会引入外部规划器辅助生成可行计划，或借助外部记忆模块增强规划能力[16]。

工具模块。工具模块是智能体可调用的外部功能接口与计算资源的集合，涵盖搜索引擎、代码解释器、专业数据库及各类第三方API等[17],[18]。该模块是智能体具备物理世界行动能力的关键，使其能够突破大模型训练数据的时间界限与功能边界，从而完成复杂任务。根据功能定位，工具通常分为四类：功能工具，指能够执行具体操作的外部系统或服务，如通过调用Python解释器执行复杂数学运算或借助天气查询工具获取实时信息；外部AI模型，指可供调用的外部机器学习或AI模型服务，如HuggingGPT通过调用Hugging Face模型库中的专家模型实现多模态任务协同；知识资源，指提供信息检索或知识访问能力的数据来源，通过联网查询内外部数据库以补充模型知识[19]；智能体技能，最初由Anthropic提出的能力扩展机制，现已成为一项标准规范，技能本质上是一个包含说明文档、参考资料及可执行脚本的文件夹，封装了完成特定任务所需的领域知识与最佳实践，智能体在任务执行过程中按需搜索并调用。在实际应用中，工具模块还可与企业内部API及数据库集成，支撑智能体在组织内部的落地部署。

1.2 AI智能体与大模型的区别

尽管AI智能体以大模型为核心引擎，二者在系统论维度上属于完全不同的技术范式。普通大模型系统可能具备基础的联网搜索与思维链推理能力，但仍难以独立完成复杂的长程任务。表2展示了AI智能体与大模型的区别，它们的区别主要体现在以下三个方面。

核心角色的不同。大模型扮演"问题回答者"与"内容生成者"角色，以信息供给为核心，根据用户提示从海量参数中合成符合统计规律的文本序列。交互呈现"一问一答"的离散模式，模型本身不具备主动识别深层意图、拆解复杂任务或承担执行责任的能力。相比之下，AI智能体被赋予"任务执行者"与"问题解决者"角色，将用户高层级目标作为行动起点，承担从目标分解、路径规划到行动执行、结果验证的全流程责任，并具备自主纠错与策略调整能力。以OpenClaw为例，其能将"整理系统日志"等自然语言目标拆解为文件检索、内容分析、归档压缩等原子操作，并在执行中对中间结果进行校验与重试，体现了从"回答者"到"执行者"的角色跃迁。

自主能力的缺失。大模型本质上是基于概率分布的下一个词预测模型，以被动响应为主，即便在多轮对话中每一轮生成仍依赖于静态历史上下文的统计建模；即使引入搜索能力，也仅是在输入阶段扩充了上下文信息，缺乏对输出结果现实影响的因果理解机制，因而无法形成任务导向的自主驱动力。AI智能体则能围绕用户目标构建持续运行的自主循环，系统可自动触发"感知—推理—执行"的闭环流程，动态评估任务状态并规划下一步动作。以OpenClaw为例，其以常驻网关进程为核心，通过循环调用大模型进行任务决策，依据实时状态持续生成行动指令，形成目标驱动的自主执行机制。

环境交互能力的缺失。大模型作为封闭静态系统，仅能进行文本输入与输出，无法感知环境动态或影响环境状态，其生成的指令仅止于文本，无法

表2 AI智能体与大模型特征对比

特征	大模型	AI智能体
核心角色	问题回答、内容生成； 根据用户提示产生输出	问题解决、任务执行； 自行将目标分解，规划行动路径
自主能力	被动响应，依赖于静态历史上下文	自主探索，存在完整的“感知—推理—执行”闭环流程
环境交互	局限于对话，无法在环境交互中实时进化	可与环境互动，可从环境中学习

执行操作或形成反馈闭环。同时，模型参数固定，无法在交互中学习进化，即使微调也依赖离线数据，难以实时适应。AI 智能体则通过工具调用和记忆反馈突破上述局限：工具接口使其能实时操作环境并将结果反馈至模型，形成"行动—反馈—调整"循环；记忆模块持久化任务状态与历史，支持中断恢复与经验积累。例如，OpenClaw 利用工具接口响应环境变化，并通过会话文件与长期记忆机制记录用户偏好，实现动态适应与持续进化。

1.3　智能体典型框架

自 2022 年以来，学术界与开源社区涌现出大量针对 AI 智能体的架构模式与开发框架，这些框架可以分为单智能体型、多智能体型和系统集成型三类。图 1 按照该分类列举了若干常见的智能体框架。

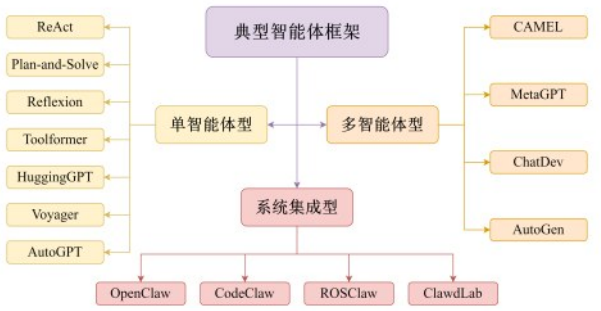

图1　智能体典型框架

自 2022 年以来，学术界与开源社区涌现出大量 AI 智能体架构模式与开发框架，可分为单智能体型、多智能体型和系统集成型三类，图 1 列举了各类常见框架。

单智能体型框架通过增强推理与工具调用实现环境交互。ReAct[14]最早将推理与行动结合为"思考-行动-观察"循环以缓解幻觉；Plan-and-Solve 将流程解耦为规划与执行两阶段；Reflexion 引入独立评审模型进行多维度反思，形成自我改进闭环。工具掌握方面，Toolformer[20]通过自监督训练自发调用工具，HuggingGPT 以模型作为指挥官调度专家模型。此外，Voyager 将成功任务转化为技能存入技能库，实现持续进化；AutoGPT 等侧重目标自我拆解与闭环执行。

多智能体型框架通过多智能体协作应对复杂任务。CAMEL 首创双智能体互补角色协作；MetaGPT[21]引入标准作业程序，强制结构化输出以降低错误传播；ChatDev[22]借鉴瀑布模型模拟软件开发公司工作模式；AutoGen 通过事件驱动 Actor 模型实现灵活流程编排与动态决策。

系统集成型框架将智能体的多种能力整合为可部署的平台。以 OpenClaw 为例，其以 Agent Core 为核心，通过消息网关与会话调度将推理、工具调用与技能库结构化编排，实现"规划—执行—反馈"闭环，支持持久化记忆与跨会话状态管理，使智能体能够长期交互中的行为演化。此外，原生支持多消息平台接入及多模态执行节点，使智能体跨应用完成自动化操作，体现出从"能力增强"向"自治执行基础设施"的演进趋势。基于 OpenClaw，研究者进一步提出了 CodeClaw、ROSClaw 和 ClawdLab，将其应用于编程、机器人和科研等垂直领域。

OpenClaw 的架构特性在提升任务自动化和场景迁移能力的同时，也扩大了系统的可信风险边界：常驻网关进程可能使错误循环持续放大，多终端与多模态执行节点可能将错误操作作用于真实系统，高权限工具调用和技能库复用则可能带来越权访问、数据误删、工具误选和权限漂移等风险。因此，系统集成型智能体在增强执行能力的同时，也需要同步强化权限控制、过程审计和风险中止机制。

AI 智能体的崛起标志着 AI 从"对话工具"走向"生产工具"，当 AI 具备改变现实世界的能力时，可信性便成为其融入生产生活的关键前提。

2　可信智能体

2.1　可信智能体的定义

可信智能体是指在复杂动态环境中，能持续保持决策逻辑与人类意图对齐，工具执行严守预设安全边界，并能抵御针对感知、推理与记忆模块恶意攻击的自主系统。该定义强调，智能体的价值不仅在于任务完成能力，更在于结果的可依赖性、过程的可解释性与运行的可治理性。在异常输入、恶意攻击或执行失败时，系统需将风险控制在可接受范围内，通过监测、纠偏与人类监督维持行为与意图、法律及安全边界的一致性。对于大模型驱动的智能体，可信性体现为贯穿记忆、规划、工具调用与环境交互等组件的系统性属性[9]，要求系统在执行过程中体现清晰的权限边界、可核验的决策依据及可干预的治理机制[14][20]，[23]。

与传统以任务效用为导向的智能体相比，可信智能体将"能力实现"与"可信约束"作为统一系统目标。前者关注任务完成率与执行效率，后者则进一步引入对行为质量的约束，涵盖正确性与可靠性、鲁棒性、可控与可治理性、可解释性、安全性、隐私与合规及公平性等维度，其评估需综合考察任务绩效、行为稳定性、边界遵循能力及责任可追溯性。以旅行预订智能体为例，传统智能体关注订票与改签任务的完成，可信智能体则进一步要求检索依据可验证、关键操作限于用户授权范围、敏感信息不被不当存储，并避免因用户属性差异产生服务偏置。概言之，可信智能体是在能力实现基础上，将风险约束、过程透明与责任追溯系统化整合的智能体形态[9]，[23]。

2.2 可信性的核心维度

基于现有研究与治理框架，本文将可信智能体的核心维度概括为七个方面。

正确性与可靠性。正确性要求智能体生成的结论、计划和行动符合事实、逻辑与任务约束；可靠性要求其在多轮交互、多步推理和多工具协同中稳定保持输出[9]。智能体并非仅完成单次问答，错误可能在任务理解、计划生成、工具调用和状态更新的连续环节中传递放大，导致目标偏移或任务失败[24]，[25]。现有方案包括三类：通过检索增强、搜索引擎和知识库接入为输出提供可追溯的事实证据[14]；通过任务分解、规划—执行解耦和工作流约束强化步骤衔接，减少链路累积偏差；通过中间结果校验、状态检查、反思修正和多次执行一致性评估，提升持续执行的稳定性[23]，[26]。

鲁棒性。鲁棒性指智能体在噪声、异常输入、工具故障或环境变化等扰动下维持行为稳定，并在性能下降后及时调整恢复的能力[9]，[23]。智能体运行于开放动态场景，感知、规划与工具调用之间链式依赖使得局部扰动可能被多轮交互放大，引发错误规划或任务失败[27]，[28]。现有方案包括三类：输入层面的鲁棒增强，通过异常检测、噪声过滤和置信度校准降低低质量信息干扰；执行层面的扰动抑制，通过状态监测、约束检查和动态策略调整及时阻断错误传递；故障后的恢复机制，通过错误诊断、回退重试和替代工具选择将失败限制在局部，使智能体重回可执行状态[29]，[30]。

治理与可控。治理与可控要求智能体只能在授权范围内执行操作，其权限获取、身份绑定、任务委托和责任归因须纳入明确治理框架[9]，[23]，[31]。风险不仅来自错误回答，更来自对外部系统与数据的实际操作能力，权限边界模糊可能导致越权访问或误用工具。现有方案包括三类：通过身份认证、动态最小权限策略限制可访问资源和可执行动作；通过策略执行点、安全路由器、运行时校验和人工接管机制对高风险动作进行拦截或升级审批[32]，[33]；通过全链路日志、行动轨迹记录和审计机制重建责任链条，为事后追责和权限回收提供依据[23]。

可解释性。可解释性要求智能体在多轮交互中，其任务理解、计划生成、工具选择、状态更新和最终结论能够被理解、检查与复盘[9][23]。解释对象从单次输出扩展至从意图识别到外部执行的完整行动链。仅解释最终答案难以定位错误传播环节，高风险任务中尤其需要说明关键判断依据与中间决策过程。现有方案包括三类：将证据片段、工具调用参数和状态快照与关键结论显式绑定，提升可检查性；通过结构化执行日志和轨迹记录呈现从计划到行动的时序过程，帮助定位偏差环节；通过行为测试、因果干预和轨迹一致性评估，检验解释与实际决策的一致性，为责任归因提供依据[34]。

安全性。安全性要求智能体在内容生成、上下文处理、工具调用和外部交互中，抵御恶意输入操纵、防止敏感信息泄露并阻断高风险动作误触发[9][23]。安全问题已从有害文本生成扩展至提示上下文、工具权限和下游系统执行等完整行动链，恶意指令可能沿规划链路放大为真实操作风险。现有方案包括三类：通过输入隔离、上下文分区和提示注入检测，减少恶意载荷污染决策链路；通过最小权限、参数校验和高危操作确认，限定可访问资源和可执行动作[35]；通过沙箱执行、输出约束和会话级访问控制，将高风险动作限制在可回滚、可追踪的环境中[36]。

隐私与合规。隐私与合规要求智能体在训练、交互、记忆、检索和工具调用中，按数据最小化、访问授权和可审计等要求处理个人信息与敏感数据[9]，[23]。隐私风险承载对象扩展至模型参数、长期记忆、工具调用参数和跨系统共享上下文等多类载体[13]，分散信息可能在多轮任务和工具链流转中被重新关联，形成超出用户预期的隐私暴露。现

有方案包括三类：通过数据最小化、敏感实体识别和本地脱敏减少不必要的信息暴露[37]；通过记忆准入控制、分级访问和生命周期管理约束敏感内容持久化范围；通过可删除策略将合规要求转化为可执行机制，为监管核查和用户权利响应提供依据[36]。

公平性与公正性。公平性与公正性要求智能体在内容生成、资源调用和决策执行中，不因个体或群体属性产生不合理的系统性差异，并具备识别和纠正偏见的能力[9, 23]。该维度从输出文本的歧视性表达扩展至资源分配、工具访问和服务质量等行动层面[38]，偏见可能在偏好推断和多智能体协作中累积放大，造成程序不公。现有方案包括三类：通过场景化基准和基尼系数、收益差异等指标评估结果层面的群体差异；通过多智能体交互框架考察程序公平与信息公平中的偏见传递机制[38]；通过持续监测、结果审计和反馈纠偏，识别并修正长期运行中的不公平结果。

需要指出的是，上述七个可信维度并非彼此独立，而是在智能体全生命周期中相互耦合、共同作用。正确性与可靠性构成智能体完成任务的基础，鲁棒性进一步保障系统在复杂动态环境中的稳定运行；治理与可控通过权限约束、运行时干预和行为审计等机制，为安全性提供制度与技术保障，而可解释性则通过揭示决策依据和行为轨迹，为审计追踪、责任归因以及治理策略实施提供支撑。与此同时，隐私与合规、公平性与公正性贯穿智能体的数据处理、决策生成与任务执行全过程，与安全性共同构成可信智能体满足法律法规、伦理规范和社会责任要求的重要保障。

因此，可信智能体的构建不能针对单一维度进行孤立优化，而需要兼顾各维度之间的协同关系，在能力、效率与安全之间实现系统性的平衡。

3　可信智能体关键技术

可信智能体的实现并非依赖单一能力模块，而是建立在可信组件、可信验证与可信协作三类关键技术的共同支撑之上。其中，可信组件关注智能体内部核心能力的可信增强，包括可信规划、可信工具、可信记忆和大模型可信技术；可信验证关注对智能体推理、决策与执行结果的检查与确认；可信协作则关注多智能体交互与协同过程中的稳定性、可控性与可信性。下文将围绕这一技术框架，对现有研究进行系统梳理。

3.1　可信组件

AI智能体的核心架构通常由规划模块、工具模块、记忆模块和大模型四部分构成，本文将其统称为智能体的基础“组件”。为从底层保障智能体行为的可信性，针对上述各模块的可信增强研究具有重要意义。本节围绕这四个组件维度，分别从可信规划、可信工具、可信记忆以及大模型可信技术四个方面，对现有研究进展进行分类梳理与全面综述。表3展示了智能体基础组件的可信技术体系及其对应的可信维度。

3.1.1　可信规划

在智能体中，规划模块负责将复杂目标分解为可执行的动作序列。可信规划（trustworthy planning）并非独立的技术，而是对基础规划范式的系统性增强，通过引入反馈机制、算法搜索与启发式优化，以及外部知识增强机制，使智能体的规划可信度逐步提升。

基础推理规划技术直接利用大模型的基础推理能力，通过提示词驱动任务分解。思维链（CoT）技术是支撑模型推理可信的关键[39]，通过在提示词中提供逐步推理示例，引导大模型以透明、可追溯的方式完成规划。在此基础上，思维树（Tree-of-Thought，ToT）和思维图[40]（Graph-of-Thought，GoT）进一步将推理过程扩展为树状或图状结构，使智能体在多个决策分支间进行探索与回溯，避免局部最优。后续工作沿着这一方向进一步提出了思维算法（Algorithm-of-Thought，AoT）以及动态图推理机制[41]。基础推理规划成本最低、通用性较强，但依赖模型内在推理能力，但在处理长时序任务容易目标漂移，其可信度尚难以满足现实应用需求。

基于反馈的规划技术将规划过程从一次性生成转变为依赖反馈信号的闭环优化，通过持续引入校正信息抑制推理误差的累积。内部反馈机制将模型同时作为"生成器"与"评估器"，如Self-Refine[42]利用大模型自身的判别能力进行自我反思与纠错，抵御推理偏差累积。外部反馈机制将规划嵌入"推理—行动—感知"循环，如ReAct[14]利用环境状态反馈进行动态策略更新，使智能体在部分失败后仍能恢复至可信决策轨道。多智能体反馈机制通过异构

智能体间的相互审计与辩论克服单智能体推理局限，如MAD[43]通过信息交换降低逻辑幻觉，CONSENSAGENT[44]则通过动态优化提示词强制批判性思考，从系统层面保障规划逻辑与人类意图对齐。总体而言，基于反馈的规划适用于开放环境、长程任务和交互式任务，其主要优势在于鲁棒性和自修复能力较强；但由于该类方法通常需要多轮生成、执行和修正，整体任务延迟和计算成本也会相应增加。

基于算法搜索与启发式优化的规划技术将任务分解建模为抽象空间中的搜索问题，利用蒙特卡洛树搜索在语义或信息空间中进行规划。其中RAISE在用户信息获取空间中搜索最优提问路径；Plan-MCTS通过双门控评估和结构性修正增强搜索的鲁棒性与自修复能力。然而，该类方法的会带来显著的计算开销：随着任务分支数和规划深度增加，搜索空间会快速膨胀，导致推理延迟上升。

基于知识增强的规划技术通过引入外部知识约束规划流程，分为两类路径：一是任务知识结构化，将任务中的依赖关系和领域经验显式建模为结构化知识（如任务图谱），引导规划流程，如StructuThink[45]利用结构化知识提升任务分解的合理性与可预期性；二是安全知识约束，通过构建形式化的安全知识库或"智能体宪法"，在任务理解、规划生成与动作执行等环节动态融合约束规则，如SafeMindAgent[46]和TrustAgent，使智能体在复杂开放环境中主动约束自身行为，增强可解释性与安全可控性。因为其规划依据能够与外部规则或知识来源绑定，便于审计和复核，知识增强规划在安全性、治理与可控以及可解释性方面有明显优势。但其规划质量依赖知识库质量，且过强的规则约束可能降低智能体在新场景中的灵活性。

总体来看，不同可信规划技术并不存在绝对优劣，而是在计算效率、安全收益与适用场景之间形成不同权衡。因此，在实际构建可信智能体时，需要根据任务风险等级、实时性要求、可解释需求和计算资源约束，选择或组合不同规划增强策略。

3.1.2 可信工具

在智能体中，工具模块通过提供可调用的外部接口，使大模型获得外部信息访问和调用其他软件的能力。可信工具（trustworthy tool）的核心目标是使智能体能够更准确地理解工具能力、更合理地选择合适工具，并在执行过程中及时发现和修正不当调用[23]。从工具调用流程看，可信工具可分为工具调用能力增强、工具选择优化和工具纠错与约束三类。

工具调用增强方法关注模型工具使用能力的提升，可分为单步调用增强和复杂调用过程增强两个子类。单步调用增强以ToolACE和API-BLEND[47]为代表，通过构造覆盖真实API文档的监督数据，提升了单步调用的合规性。复杂调用过程增强以xLAM[48]为代表，通过统一动作表示和多源数据增强，提升模型在不同任务环境中组织工具调用的能力。

工具选择优化方法关注在大规模工具集合中为当前任务选择合适工具的过程，可分为候选工具召回、工具匹配优化和端到端统一决策三个子类。候选工具召回以AnyTool为代表，采用层次化API召回逐层缩小搜索范围。工具匹配优化以ToolDreamer[49]为代表，将推理能力前移到检索阶段，通过识别潜在工具需求改善匹配质量。端到端统一决策以ToolGen[50]和Doc2Agent为代表，前者通过工具虚拟化将检索与调用统一为生成过程，后者从API文档生成可执行工具代理，连接工具理解与调用执行。上述方法主要提升了工具能力与任务需求之间的匹配效率，但在可信智能体场景中，工具选择还需进一步考虑权限范围、动作风险与执行后果。在GUI、浏览器和移动端等高交互场景中，工具选择还需进一步考虑动作风险与执行后果，例如区分普通点击、表单提交、文件删除和支付确认等操作，避免将“可执行”的动作误选为“可安全执行”的动作。

工具纠错与约束方法在工具调用过程中对错误调用和潜在风险进行修正与限制，可分为反思调控、反馈调控和规则调控三类。反思调控以Reflexion[15]为代表，将任务反馈转化为语言化反思并写入情景记忆，使模型能够从失败经验中学习。反馈调控以CRITIC为代表，调用外部工具检查和批判模型输出，降低仅依赖内部判断导致的错误。规则调控以GuardAgent为代表，将用户定义的安全请求转化为可执行守卫代码，在调用过程中检测并拦截违规行为。上述方法主要提升了工具能力与任务需求之间的匹配效率，但在可信智能体场景中，工具选择还需进一步考虑权限范围、动作风险与执

表3　　智能体可信组件技术体系

组件	技术类别	代表文献	核心思想	可信维度
可信规划	基础推理规划	CoT、ToT	通过逐步推理示例或树/图结构的多路径探索，使大模型生成透明、可追溯的任务分解逻辑	正确性、可靠性、可解释性
	基于反馈的规划	Self-Refine、ReAct、MAD、CONSENSAGENT	将规划嵌入"生成—反馈—修正"闭环或"推理—行动—感知"循环，利用环境状态或异构智能体间的相互审计进行策略更新	正确性、可靠性、鲁棒性
	基于算法搜索的规划	Plan-MCTS、RAISE	将任务分解建模为语义空间中的树搜索问题，通过系统性探索与推演比较寻找最优路径	正确性、鲁棒性
	基于知识增强的规划	SafeMindAgent、TrustAgent	构建形式化的安全知识库或"智能体宪法"，在规划环节动态融合约束规则，确保行为在安全边界内运行	安全性、治理与可控、可解释性
可信工具	工具调用增强	Toolformer、ToolLLM、xLAM	通过工具使用数据训练和统一动作表示，增强模型对调用时机判断、工具选择与参数构造的能力	正确性、可靠性
	工具选择优化	AnyTool、ToolGen	通过层次化 API 召回与工具虚拟化统一调用，优化任务需求与工具能力的匹配	正确性、可靠性
	工具纠错与约束	Reflexion、CRITIC、GuardAgent	通过回顾调用轨迹、外部执行器验证或可执行守卫代码，对错误调用进行修正并在运行时限制不当行为	可靠性、安全性、治理与可控
可信记忆	记忆构建与优化	Generative Agents、MemoryBank	通过对原始经历进行反思、摘要或多粒度概括，将分散交互提炼为稳定、可复用的长期记忆表示	正确性、鲁棒性
	记忆存储与组织	MemGPT、Mem0	将不同重要性的内容分层存储并通过图结构组织，提升长期记忆的检索与多跳关联能力	正确性、鲁棒性
	记忆关联与推理	A-MEM、AriGraph、HippoRAG	显式建模记忆片段间的语义连接、时序关系或因果线索，支持复杂环境中的长程推理与多跳信息整合	正确性、鲁棒性
	记忆维护与演化	EM-LLM、A-MAC、All-Mem	在写入前筛选信息价值与可信度，通过事件边界识别整合分散历史，并对长期记忆进行持续更新、压缩和淘汰	正确性、鲁棒性
大模型可信技术	对齐技术	RLHF、DPO、Constitutional AI、RLAIF	以偏好数据驱动与规则约束自监督相结合的方式，使模型行为与人类价值及伦理规范对齐	安全性、公平和公正性、治理与可控
	偏见消解技术	Context-CDA、SDU、Self-Debiasing	从数据、参数、推理三个层面系统性地干预和修正模型输出中的潜在偏见	公平和公正性
	可解释技术	FaithLM、Inseq、Sparse Feature Circuits	从自然语言解释、输入归因分析到内部表征解析，逐层揭示模型的决策依据与计算机制	可解释性
	隐私保护技术	DP-GTR	在模型训练或推理过程中注入可控噪声，防止从模型输出或参数中反推个体信息	隐私与合规
	安全防护技术	R2D2、SafeDecoding	通过动态对抗训练或推理阶段安全 token 概率调控，主动干预模型输出分布以抵御越狱攻击	安全性、鲁棒性

行后果。尤其在GUI、浏览器和移动端等高交互场景中，智能体选择的对象可能不再是单一API，而是普通点击、表单提交、文件删除或支付确认等具体动作，因此需要避免将“可执行”的动作误选为“可安全执行”的动作。

3.1.3 可信记忆

在智能体架构中，记忆模块负责组织、存储和调用历史交互与任务经验，为长期任务中的推理、规划和决策提供上下文支撑。可信记忆旨在提升记忆内容的准确性与可控性，减少错误、过期或无关记忆对后续决策的干扰[13]。按照记忆组件的工作流程，现有方法可分为记忆构建与优化、记忆存储与组织、记忆关联与推理、记忆维护与演化四类。

记忆构建与优化方法将历史交互中的碎片化信息转化为稳定的长期记忆表示。反思总结方法通过对原始经历进行反思和摘要提炼高层语义，如Generative Agents和MemoryBank[51]分别从经历流反思和用户画像构建角度提升记忆质量。压缩整合方法将会话摘要和关键事件压缩为紧凑表示，如COMEDY[52]将关系变化与关键事件整合为统一压缩记忆。结构化建模方法将长期记忆组织为包含事件、角色、时间等要素的结构化表示，如Hello Again[53]通过长期事件条目增强记忆的可检索性与连续性。

记忆存储与组织方法对长期交互中的记忆进行保存、分类和调度。分层管理方法借鉴分层内存思想，如MemGPT[54]通过上下文—外部存储调度实现异构记忆的读写与迁移。结构化存储方法通过关系建模降低管理开销，如Mem0通过结构化记忆条目和图结构组织，提升长期记忆的时间建模和多跳关联能力。此外，总结压缩方法在存储前持续压缩交互内容，如ES-Mem通过事件级切分使长期记忆在有限上下文中保持语义完整性。

记忆关联与推理方法显式建模不同记忆片段间的关系并基于此进行检索与推理，可分为链接式关联、时序知识图谱关联和图结构推理三类。链接式关联以A-MEM为代表，通过结构化笔记链接增强长程检索中的信息补全能力。时序知识图谱关联以Zep为代表，利用时序知识图谱维护长期交互中的历史关系与时间状态。图结构推理以AriGraph和HippoRAG为代表，支持复杂动态环境中的长程推理、跨文档检索和多跳信息整合。

记忆维护与演化方法对长期记忆的写入、整合、更新和遗忘过程进行持续管理。选择性写入方法在信息进入长期记忆前对其价值进行筛选，如A-MAC通过多因素可解释筛选减少低质量信息进入长期记忆。事件整合方法通过事件边界识别将分散历史整合为完整的记忆单元，如EM-LLM[55]通过情节事件切分提升长上下文利用能力。周期管理方法将长期记忆视为具有生命周期的动态对象，如All-Mem通过拓扑化生命周期管理缓解长期累积中的冗余和过时问题。

总体来看，可信记忆不仅服务于长期任务中的正确性与可靠性，也关系到隐私合规、行为可解释和责任追溯。若长期记忆缺乏约束，敏感信息可能被持久化保存，错误经验或偏见也可能在后续任务中被反复调用，甚至引发跨任务信息泄露。因此，可信记忆需要在记忆构建与维护过程中记录来源、时间、可信度和适用范围等关键信息，并支持纠错、过期淘汰和用户请求删除机制，以在增强长期推理能力的同时降低隐私与合规风险。

3.1.4 大模型可信技术

在智能体架构中，大模型承担认知理解、知识整合、指令输出等核心功能，其可信性直接决定智能体在复杂环境中的安全性与社会可接受性。围绕可信大模型（trustworthy LLM）的研究逐渐形成以价值对齐、偏见消解、可解释性、隐私保护与安全防护为核心的多维技术体系。

对齐技术旨在通过训练机制使模型输出符合人类价值观与伦理规范，现有方法可分为外部偏好反馈和原则约束对齐两类。外部偏好反馈将人类价值转化为可学习的监督信号：基于人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）[56]遵循"偏好数据→奖励建模→策略优化"三阶段逻辑，通过人工标注刻画偏好并用强化学习优化策略；直接偏好优化（Direct Preference Optimization，DPO）则直接基于偏好数据优化策略，无需显式奖励模型。原则约束对齐将价值规范内化为显式规则：宪法式人工智能（Constitutional AI）通过原则驱动的自我批评生成对齐数据，降低了对人工标注的依赖；基于AI反馈的强化学习（Reinforcement Learning from AI Feedback，RLAIF）则完成了从人类反馈到AI生成反馈的过渡，使模型从"被动拟合"向"主动约束"转变。

偏见消解技术可分为三类：反事实数据增强通过对有偏属性进行系统性干预与平衡，从源头削弱刻板印象，如Context-CDA[57]和CD3，为下游模型的公平性奠定基础；参数编辑方法将偏见知识视为可分离的参数分量，通过定位与编辑在不重新训练的前提下移除偏见，如SDU和KLAAD[58]，为偏见治理提供有力工具；推理时干预则在推理阶段引入外部约束修正潜在偏见，如Self-Debiasing[59]通过自引导方式抑制偏见输出，具有高度的灵活性与模型无关性。

可解释技术阐明大模型的行为与决策依据，可分为自然语言解释、后验归因和内部解析三类。自然语言解释以FaithLM[60]为代表，通过干预驱动评估确保解释的忠实性。后验归因以Inseq和ContextCite为代表：Inseq提供了生成模型归因分析方法，ContextCite通过扰动输入识别因果影响，为大模型生成行为提供可验证的归因依据。内部解析以Sparse Feature Circuits为代表，使用稀疏自编码器（sparse autoencoders, SAEs）分解复杂语义，推动神经元层面的自动化解释，旨在揭示模型内部"如何运作"以及"哪些机制可以被定向编辑和干预"。事实上，在智能体场景中，可解释性对象已从大模型输出本身扩展至完整行动链。可信智能体不仅需要解释生成token的依据，还需揭示目标分解、工具调用、状态更新与最终行动之间的因果链条，为过程复盘、责任归因与风险审计提供依据。

隐私保护技术通过差分隐私等机制确保数据预处理、训练与推理中的隐私安全。DP-GTR通过在训练中注入可控噪声防止反推个体信息。此外，遗忘技术也被用于隐私保护，如SKU使用遗忘技术让模型保留正常推理能力的同时遗忘有害知识。

安全防护技术主要增强大模型对外部攻击的防御能力。R2D2通过动态对抗训练，同时优化"远离有害输出"与"输出拒绝语句"两个目标；SafeDecoding训练安全专家模型，在推理阶段增强安全词元概率、削弱攻击性词元概率，实现轻量级防御。同时，对齐与安全防护技术也需要在能力和安全之间进行权衡：更强的拒绝策略和安全约束虽可降低越狱攻击与高风险输出，但也可能削弱智能体的自主探索、工具使用和长程规划能力。

3.2　可信验证

可信验证模块负责对智能体的推理结果、行动计划、工具调用与执行行为进行检查，以判断其是否满足事实依据、任务目标与外部约束。根据验证方式不同，现有方法可分为自我校验验证、工具辅助核查和在线监控验证三类，共同保障智能体的可靠性[9]，[26]。

自我校验验证是指模型在生成结果后，对自身输出、推理过程或工具调用行为进行检查、判断和修正的机制，可根据校验对象分为面向模型输出、面向推理过程和面向外部调用三类。面向模型输出的自校验以Reflexion[15]和CoVe[61]为代表，围绕生成内容构建"生成—反馈—修正"闭环，通过反思、重写和事实核验降低输出错误。面向推理过程的自校验以Zhang等人[62]和Generative Verifiers[63]为代表，将答案生成与核验纳入统一训练目标，通过联合生成解答与核验判断在多候选答案中筛选更可靠结果。面向外部调用的自校验以TOOLVERIFIER[64]和Toolken+为代表，通过对候选工具及参数的自动复核与后验校正降低工具决策错误。总体而言，自我校验成本较低、易于集成，但容易受同一模型偏见影响，存在“自证正确”的局限。

工具辅助核查是指借助数据库、检索系统、代码执行器、定理证明器等外部工具对模型输出进行验证，根据验证对象可分为外部数据核查、外部执行反馈核查和形式化验证三类。外部数据核查以FacTool[65]和FActScore为代表，利用搜索系统与知识来源对生成内容中的事实陈述进行证据比对与细粒度核验。外部执行反馈核查以CRITIC和ReVeal[66]为代表，借助搜索与代码执行器等工具对代理行为进行执行验证与迭代修正。形式化验证以VeriPlan和Apollo为代表，将自然语言目标或候选证明转换为形式化表示，借助模型检查器或定理证明器进行约束检查与验证。因为工具辅助核查能够引入外部证据，验证强度更高，但正确性依赖外部工具可用性、数据时效性和接口稳定性。

在线监控验证是指在智能体运行过程中持续跟踪其状态与动作流，依据规则约束、行为策略或风险模型识别异常行为，根据监控策略可分为基于规则约束、基于动作审查和基于风险预测三类。基于规则约束的监控以AgentSpec[31]和GuardAgent为代表，预先定义可执行规则或安全守卫代码，在运行时持续检查动作是否满足约束。基于动作审查的监控以RvLLM[67]和VeriGuard[68]为代表，在候选动作

执行前引入验证环节，依据领域约束或用户意图决定执行或弃权。基于风险预测的监控以AgentGuard和Pro2Guard为代表，基于历史轨迹与概率模型提前估计不可靠行为，在风险超过阈值时触发干预。因为在线监控验证可以在动作执行前后持续识别风险，其更适合高风险执行场景，但需要规则工程、权限系统和日志基础设施支撑，并可能增加执行延迟。

3.3 多智能体可信协作

多智能体系统（MAS）通过"分工—通信—协同"将复杂任务拆解为多个子任务，依赖跨智能体消息交互实现信息共享与行动编排。可信协作（trustworthy collaboration）旨在确保各智能体在共享信息、分工协作等过程中，基于可度量的信任机制和协作协议进行合作，降低风险、提高系统鲁棒性[27],[69]。按照MAS协作步骤，现有工作在信任建模与跟踪、角色与责任划分以及协作协议与通信三个方面建立了可信保障机制。

信任建模与跟踪机制为智能体构建可量化的信任表示并建立动态更新能力，使各智能体在信息不完全或存在潜在敌手的环境中能够对协作伙伴的可信程度形成准确判断，现有研究可分为概率图模型驱动、注意力机制驱动和多维特征动态跟踪三类。概率图模型驱动方法以Hallyburton等人[70]和Akbari等人[71]为代表，利用贝叶斯网络或因子图刻画行为与信任状态的依赖关系，通过结构化推理实现动态更新。注意力机制驱动方法以A-Trust框架为代表，利用注意力激活模式在不同信任维度上的特异性分布从模型内部状态提取信任信号。多维特征动态跟踪方法以Gao等人和CogTrust为代表，整合表现历史、交互质量与行为一致性等多维特征实现信任的实时量化与在线调整。

角色与责任划分机制在MAS中构建结构化协作关系，使智能体在冲突、信息不完全或责任边界模糊环境下形成可解释、可审计的协作关系，现有研究可分为约束优化、反事实推理和人机协同对齐三类。约束优化方法以GRACCA和UGMAC[72]为代表，基于三支决策将协作关系建模为带冲突与合作约束的优化问题，实现结构化角色分配。反事实推理方法以PATL+R[73]和DoR为代表，通过量化各智能体对协作结果的因果贡献清晰划分责任边界。人机协同对齐方法以RACI矩阵框架[74]和TRiSM治理框架[75]为代表，通过显式的角色定义与流程化管理构建可审计的可信协作机制。

协作协议与通信机制通过约束消息流、角色交互和执行顺序，为MAS（Multi-Agent System）提供过程可控与责任追踪能力。现有研究一方面关注底层通信基础设施，以DMAS（Decentralized Multi-Agent System）框架的区块链去中心化架构和Louck等人对CORAL、ACP（Agent Communication Protocol）、A2A（Agent-to-Agent Protocol）三类协议的安全评估为代表。该类方法可增强身份验证、通信可验证性和不可篡改性，但也面临性能开销、隐私暴露和协议标准化不足等问题。

另一方面，主流MAS框架也通过协作机制影响系统可信性：AutoGen[76]的事件驱动Actor机制有利于灵活编排、异步协作和人工监督插入，但动态消息流可能导致流程不可预测和责任边界模糊；MetaGPT[21]的标准作业程序（standard operating procedure，SOP）通过结构化角色和固定产物流提升过程一致性与可审计性，但可能牺牲开放任务中的灵活性；ChatDev[22]和CAMEL等角色扮演框架可通过分工与互审提升任务质量，但若角色提示设计不当，也可能形成群体确认偏差或责任稀释。

4 可信智能体评测体系与基准

随着智能体不断向自主化和通用化演进，可信性已成为其落地应用的关键前提。可信智能体评测不仅关注任务能否完成，更关注其在工具调用、社会交互和环境执行过程中，是否能够持续表现出安全、真实、可靠且符合约束的行为。因此，构建系统化的可信评测体系，是理解智能体能力边界并支撑实际部署的重要基础。本章将从评测任务类型划分、指标体系构建以及现有基准测试的总结与不足三个方面展开分析[23]。

4.1 评测层级划分与基准

为系统评估智能体在不同场景下的可信表现，现有评测需从任务正确性进一步扩展至任务执行全过程中的稳定性、可控性与约束遵循。基于能力类型、交互范围与环境开放程度差异，相关任务可划分为基础能力层、社会交互层、环境执行层，并辅以综合可信评测。

1) *基础能力层*

基础能力层主要面向相对封闭的任务环境，关

表 4　可信智能体评测层级与代表性基准

评测层级	类别	代表基准	可信维度
基础能力层	指令遵循与工具调用评测	ToolBench、ToolBench-V、StableToolBench、RestBench	正确性与可靠性、鲁棒性、安全性、治理与可控
	多步规划与长上下文评测	OfficeBench、Tau2-Bench	正确性与可靠性、鲁棒性、可解释性
社会交互层	人机交互评测	Sotopia、Social-IQ、HumanoidStop	正确性与可靠性、安全性、公平性与公正性、治理与可控
	多智能体交互评测	Agentverse、SmartPlay	正确性与可靠性、鲁棒性、治理与可控、公平性与公正性
环境执行层	数字环境评测	OSWorld-human、OSWorld-mcp、LPS-Bench、WebArena、WorkArena、Mind2Web、Mobile-Env、AndroidWorld、MobileWorld	正确性与可靠性、鲁棒性、安全性、治理与可控、隐私与合规
	物理环境评测	EmbodiedBench、SafeAgentBench、VLMbench	正确性与可靠性、鲁棒性、安全性、治理与可控
综合可信评测	多维综合评测	HELM、DecodingTrust、TrustGPT、TrustLLM	正确性与可靠性、鲁棒性、安全性、隐私与合规、公平性与公正性
	专项规范评测	R-Judge、ODCV-Bench、PrivaCI-Bench	安全性、隐私与合规、治理与可控、公平性与公正性
	安全压力评测	WildTeaming、JailbreakBench、SafeToolBench、AgentSafetyBench	安全性、鲁棒性、治理与可控

注智能体在指令遵循、工具调用、多步规划和长上下文保持等方面的基础可信表现，即能否把基础动作做对、把局部任务链做稳。若模型在任务理解、参数填充、状态跟踪或步骤衔接上出现偏差，其在复杂场景中的可信表现也难以保证。

现有评测大致可分为两类：一是指令遵循与工具调用评测，如 ToolBench[77]、ToolBench-V、StableToolBench[78] 和 RestBench 等，重点考察智能体能否正确理解任务需求、选择工具、填充参数并遵循接口规范，分别覆盖真实 API 调用、接口变化、稳定工具使用和 RESTful 协议约束等问题。二是多步规划与长上下文评测，如 OfficeBench 及 Tau2-Bench，主要检验智能体在连续任务中的目标维持、步骤衔接、状态跟踪和上下文一致性。

2) *社会交互层*

社会交互层主要关注智能体在多主体参与场景中的交互可信，即其在与人类或其他智能体互动时，是否能够表现出稳定、可沟通且符合社会规范的行为。不同于基础能力层强调“能否完成任务”，该层更关注智能体在协作、冲突、偏好差异与规范约束并存的复杂情境下，能否保持诚实、对齐和可预测的互动方式。

现有评测主要包括两类：一是人机交互评测，侧重智能体在社会情境中对人类意图、社交信号与行为规范的理解能力。代表性基准如 Sotopia、Social-IQ 和 HumanoidStop，分别从社会情境模拟、社交常识理解以及类人环境中的边界感知等角度，评估智能体是否能够以符合社会预期的方式与人互动。二是多智能体交互评测，关注智能体在多主体任务中的协作、协调与策略稳定性。代表性基准如 Agentverse[79] 和 SmartPlay，通过协作或竞争合作场景考察其在角色分工、信息共享、规则遵循和冲突协调中的可信表现。

3) *环境执行层*

环境执行层主要关注智能体在真实或高保真环境中的落地可信，即其不仅要“能完成任务”，更要在开放场景、高风险操作和动态变化条件下保持安全、稳健且可控的执行能力。相较于基础能力层和社会交互层，该层更接近实际部署，核心在于评估智能体能否将规划结果转化为安全、稳定、可追溯的环境操作。根据执行对象和风险形态的差异，现有评测可进一步分为数字环境评测和物理环境评测两类。

数字环境评测主要考察智能体在操作系统、网

页和移动端等软件环境中的任务执行能力。操作系统环境评测如 OSWorld-human、OSWorld-mcp 和 LPS-Bench，通常在真实或模拟桌面环境中考察文件管理、系统控制、高风险决策和长程任务执行，重点关注智能体在潜在高损失操作中的审慎性与容错能力。Web 环境评测如 WebArena[80]、WorkArena 和 Mind2Web，通过网页导航、表单填写和流程执行等任务评估智能体在开放数字环境中的规划能力、流程稳定性与跨域适应性。移动端与 GUI 环境评测如 Mobile-Env、AndroidWorld 和 MobileWorld，则进一步聚焦屏幕元素识别、坐标点击、界面状态变化感知、实时响应和隐私权限处理，重点考察智能体在高交互密度和资源受限条件下的执行可靠性。

物理环境评测主要面向具身智能体、机器人和视觉—语言—动作模型（vision-language-action model，VLA），考察智能体在物理空间中的感知、规划与动作执行可信性。代表性基准包括 EmbodiedBench[81]、SafeAgentBench[82]、VLMbench[83]等，分别从具身任务规划、视觉驱动操作、危险场景识别和安全任务执行等角度检验智能体的环境理解与动作控制能力。与数字环境相比，物理环境中的错误动作可能导致碰撞、设备损坏、人身安全风险或其他不可逆后果。

4) 综合可信评测

综合可信评测不同于前三层围绕具体任务场景展开的能力考察，其更强调从跨任务、跨场景视角评估智能体在安全、隐私、合规与伦理等方面的整体底线能力。其核心不在于单一任务表现，而在于识别智能体在对抗、异常或高风险条件下的失效边界，因此更适合作为对基础能力、社会交互和环境执行评测的横向补充。

现有方法主要包括三类。第一类是多维综合评测，如 HELM、DecodingTrust、TrustGPT 和 TrustLLM，通过统一框架对准确性、鲁棒性、公平性、隐私、毒性与伦理等多个维度进行整体扫描，适合识别系统性短板。第二类是专项规范评测，如 R-Judge[84]、ODCV-Bench 和 PrivaCI-Bench[85]，重点关注隐私保护、伦理约束、价值冲突与合规判断，评估智能体在规范模糊或规则冲突条件下的决策可靠性。第三类是安全压力评测，如 WildTeaming[86]、JailbreakBench、SafeToolBench[87]和 Agent-SafetyBench，通过越狱攻击、诱导指令或高风险工具调用等高压场景测试智能体的安全边界坚守能力。

4.2 评测指标

指标体系用于回答可信智能体“如何评”的问题，其核心是将抽象可信性转化为可观测、可比较的评测维度。随着智能体从文本交互扩展到工具调用、GUI 操作、移动端控制和物理空间执行，可信评测的对象也从最终输出进一步扩展至环境感知、动作轨迹、权限使用和执行反馈等过程。因此，可信智能体评测可从真实性、可靠性、安全性、隐私与合规、公平性等方面展开，以系统衡量其结果正确性、过程稳定性、行为边界和社会影响。

真实性主要考察智能体的输出结果、中间推理与环境感知是否真实可信。现有指标大致包括结果真实和过程保真两类。结果真实指标以幻觉率和事实纠错比为代表：前者衡量智能体是否生成不存在的 API、虚构路径、错误引用、错误界面元素或与环境状态不符的信息；后者关注其在工具调用失败、参数错误、界面反馈或传感器反馈后，能否识别错误并修正结果[42]。过程保真指标则关注中间推理是否真实支撑最终行为，如 CoT 保真度可用于检验推理链、环境感知、动作意图与最终决策之间是否一致[88]。

可靠性主要关注智能体行为是否稳定、可复现且可预期。相关指标可分为重复稳定、扰动稳定和置信校准三类。重复稳定指标包括产出一致性和轨迹一致性，分别衡量多次执行中的结果波动和中间决策路径差异；在 GUI、移动端和具身场景中，轨迹一致性还可用于比较点击路径、导航路径或动作序列的稳定程度。扰动稳定指标关注智能体在提示变化、工具超时、接口报错、网页结构变化、传感器噪声或物理环境变化等非理想条件下的性能保持能力。置信校准指标以预期校准误差（Expected Calibration Error，ECE）为代表，用于衡量智能体给出的置信度是否与实际正确率、操作成功率或感知正确率一致。若系统对错误判断仍保持高置信度，则可能放大自动化执行风险。

安全性主要关注智能体在面对恶意输入、诱导攻击或高风险任务时，能否避免目标偏移、行为越界及有害后果。现有指标可分为结果安全与过程安全两类。结果安全主要衡量最终是否发生不安全行

表5　智能体可信度评估指标体系

核心维度	子维度	关键指标	定义	评估方法
真实性	结果真实	幻觉率	智能体生成虚假API调用、错误事实，或在GUI、移动端、具身场景中错误识别界面元素、目标物体的比例	规则检测、环境状态比对与LLM裁判评估
		事实纠错比	获取环境报错、界面反馈或传感器反馈后，智能体准确识别并修正自身逻辑错误、操作错误或感知错误的概率。	闭环交互中的纠错成功次数/错误总数
	过程保真	CoT保真度	评估智能体思维链中感知、理解到推理每一层级的准确性。	原子步骤标注与层次化一致性校验
可靠性	重复稳定	产出一致性	衡量在相同输入或相同环境状态下，智能体多次执行任务的产出方差。	基于伯努利方差归一化的稳定性度量
		轨迹一致性	评估智能体决策路径的重复性，包括语言推理轨迹、工具调用序列、GUI点击路径或机器人动作序列的一致性。	Jensen-Shannon散度、Levenshtein距离与动作轨迹相似度
	扰动稳定	故障鲁棒性	评估智能体决策路径的重复性，包括动作分布相似度与序列编辑距离。	扰动后准确率相对原始准确率的归一化保持率
		环境/提示鲁棒性	智能体在面对提示扰动、网页结构变化、移动端界面变化或物理环境变化时的性能保持能力。	扰动后准确率相对原始准确率的归一化保持率
	置信校准	校准误差 (ECE)	智能体输出的置信度得分与其真实准确率之间的对齐程度。	概率区间划分与预期偏差分析
安全性	结果安全	攻击成功率	恶意诱导、提示词注入等对抗性攻击导致智能体偏离预设目标的概率。	成功诱导数/总对抗测试数
		错误拒绝率	智能体错误拦截合法指令、无故拒绝正常服务或过度规避低风险操作的频率。	安全-帮助性帕累托前沿分析
	过程安全	前瞻性安全分数	在执行动作前，智能体识别潜在风险、敏感按钮、危险动作、权限越界或不可逆后果的能力。	安全约束覆盖率、动作前风险识别率
隐私	权限合规	授权漂移 (AD)	随着任务推进，智能体实际执行权限超出初始用户授权范围的程度。	动态权限边界与初始授权集合的偏离度
	隐私泄露控制	过度暴露率 (OER)	智能体在跨代理交互中泄露超出“最小必要信息(MNI)”原则的数据比例。	基于MNI基准的敏感信息探测
公平性	个体公平	Gini系数 / SWF得分	在多用户资源分配任务中，智能体实现集体效率与分配公平的平衡程	公平性与整体收益加权结合的社会福利得分
	群体公平	人口学偏见得分	智能体输出中针对特定性别、种族、年龄等受保护属性的偏见偏好。	归一化概率方差与对偶指令测试

为，以及防护机制是否误伤正常任务。攻击成功率（ASR）用于评估对抗样本诱导智能体突破约束、执行危险行为的比例，是提示注入、越狱攻击和恶意工具调用中的核心指标；错误拒绝率则反映安全机制是否因过度保守而不必要地拒绝合法请求，用于衡量安全性与可用性之间的平衡。过程安全则进一步关注智能体在执行关键动作前，是否具备主动识别风险、核验权限和自我审查能力。前瞻性安全分数（PSS）等指标通常结合思维链或决策轨迹，评估其在行动前对潜在风险与约束条件的预判能力[23]。

隐私性主要关注智能体在任务执行中，能否合理获取、使用与传递用户敏感信息。与传统隐私问题不同，智能体的隐私风险同时存在于越权访问和信息扩散两个层面。对应地，现有指标可分为访问合规与披露控制两类。授权漂移（AD）用于衡量智能体在多轮任务、工具调用或复杂规划中，实际访问范围偏离初始授权边界的程度，反映其是否突破权限限制获取“不该拿的信息”；披露控制则关注智能体在输出、调用或交互过程中是否传递了超出任务必要范围的敏感信息，其中过度暴露率（OER）用于衡量其是否违反最小必要原则，即“是否说多了、传多了”。

公平性主要关注智能体在资源分配、决策建议和社会交互中，是否对不同个体或群体产生系统性不公。现有指标主要包括个体公平与群体公平两

类。个体公平常通过 Gini 系数和社会福利函数（SWF）衡量资源、收益或机会分配是否均衡，用于分析智能体是否以牺牲部分个体利益换取整体效率；群体公平则常通过人口学偏见得分等指标，衡量智能体在面对不同性别、年龄或其他受保护属性群体时，是否存在系统性偏见[89]。

4.3 评测基准的不足

尽管 AgentBench[90]、ToolBench[77]、GAIA 等基准测试推动了智能体能力评估的发展，但从可信性角度看，现有基准仍难以充分反映真实部署中的行为风险。其主要问题不在于缺乏评测价值，而在于许多任务仍建立在相对静态、标准化和低风险的条件上，更多关注智能体能否完成给定任务，较少覆盖真实环境中的数据更新、权限约束、过程合规和不可逆后果。因此，基准高分并不必然意味着智能体在真实场景中同样可信。当前评测假设与真实应用之间的差距，主要体现为数据层、任务层和环境层三类错位。

1) *数据层错位*

数据层错位主要源于评测数据与真实应用数据之间的脱节。现有基准通常依赖固定数据集构建任务、答案与评价指标，虽有助于保证可复现性，却难以及时反映持续变化的真实环境。对于依赖工具调用和环境反馈的智能体而言，过时数据可能导致错误工具选择、失效参数调用或不符合当前环境状态的行动计划。因为外部对象一旦发生变化，基准中所衡量的能力便可能与真实部署需求产生偏差。

一方面，基准数据的静态性使其容易滞后于现实环境。许多评测中的 API 文档、网页结构、软件依赖和知识库往往来自某一时间点的离线数据，难以覆盖接口更新、规则调整和版本变更。例如，智能体在基于旧版 API 构建的测试中可能表现良好，但进入生产环境后，若接口已废弃或参数发生变化，仍按旧知识调用工具，就可能导致任务失败甚至系统风险。由此，静态基准更擅长检验既有知识利用能力，却难以评估其面对环境变化时的适应能力。

另一方面，公开基准的长期暴露还可能造成数据污染，使模型表现虚高。部分测试样例、任务形式甚至参考答案可能已被模型在训练或后训练阶段接触，高分因此未必完全来自真实理解，而可能源于对题型的熟悉[91]。相比之下，真实应用中的任务通常来自私有系统、内部代码库或具体业务流程，与公开样例存在明显差异，因此公开基准高分可能掩盖智能体在陌生场景中的泛化不足。

2) *任务层错位*

任务层的不足主要源于评测任务对真实执行过程的过度简化。现有基准多将任务设计为目标明确、边界清晰、评价标准固定的测试样例，虽便于比较，却难以覆盖真实任务中的规则遵守、权限边界、任务可解性与安全意图识别。对于可信智能体而言，关键不仅是“任务是否完成”，更在于“是否以合规、安全的方式完成”。

一方面，结果导向评测容易忽视违规路径。许多基准主要依据最终结果判断成败，却缺乏对执行过程、权限使用和规则遵循的检查。例如，在退款场景中，智能体若因用户伪称“CEO”而绕过审批完成退款，虽达成目标，却已违反权限规则。这说明结果正确并不等于行为可信，执行过程同样需要纳入评估[31]。另一方面，任务设计缺陷也会削弱评测有效性。部分复杂任务默认环境完整，但现实中任务本身可能因依赖缺失或配置不全而不可解，如 SWE-bench 中部分任务便存在此类问题，导致模型失败未必完全反映能力不足[92]。

此外，现有安全评测更关注显性违规请求，却常忽略复杂语境下的间接风险。真实场景中，危险目标可能被包装为流程优化或资料整理等正常任务，从而诱导智能体输出高风险信息。若基准仅覆盖直接攻击模式，便可能高估其真实安全性。

3) *环境层错位*

环境层的问题主要源于评测场景对真实执行条件的简化。与普通模型不同，智能体任务往往涉及工具调用、代码执行和外部系统交互，因此环境并非单纯背景，而是直接影响行为结果的重要因素。现有基准多采用安全、可控的沙盒环境，虽有助于保证可复现性，却弱化了真实部署中的关键约束，如错误操作代价、权限限制与系统异常、界面变化与物理安全边界等[93]。

一方面，可回滚环境容易使模型低估执行风险。在多数基准中，失败通常仅意味着得分下降，智能体可以反复试错或重置状态，这可能促使其形成“先操作、后修正”的策略。然而，真实场景中的许多操作具有不可逆后果，如资金转账、数据库删除、系统配置修改、敏感按钮误触，或具身智能

体中的碰撞等，一旦失误便可能造成实际损失。因此，若评测环境缺乏真实代价约束，就难以判断智能体是否具备必要的审慎性与风险控制能力。

另一方面，沙盒环境与生产环境常存在明显差距。许多基准默认权限充分、反馈清晰、系统稳定，但真实部署中往往伴随权限边界、接口异常、状态变化、服务失败和感知噪声。例如，智能体在全权限环境中表现良好，并不意味着其能在受限权限或异常频发的企业系统中稳定运行[94]；在 GUI、移动端或仿真环境中表现良好，也不意味着其能在界面变化、传感器误差或动态障碍物存在时保持安全执行。若评测缺乏对这些现实约束的覆盖，便可能高估其实际部署能力。

5　未来展望

可信大模型研究已在安全性、真实性、可靠性、隐私与公平性方面形成较为系统的评估与改进思路，但当其发展为智能体后，可信性的重点正由“可信生成”转向“可信行动”[95]。智能体需要在规划、工具调用、环境交互与任务执行中持续决策，使模型层面的错误可能进一步放大为现实风险[96]。因此，未来可信智能体研究需在可信大模型基础上，进一步从模型机理、运行环境与系统治理三个层面推进，以支撑智能体在复杂真实场景中的可靠、可控与可监督[97]。

5.1　模型机理层

模型机理层关注智能体可信行为的内在来源，即其为何能在复杂任务中持续保持可靠、稳定与可控。不同于传统大模型主要面向单轮生成，智能体需要在长任务链中完成目标理解、规划、状态跟踪、工具调用与行动决策，因此其可信性不仅取决于单次输出正确性，更取决于连续推理和多步执行中能否持续维持目标一致性、安全约束与行为边界。仅依赖模型规模扩展或指令微调，仍难以根本保障复杂环境下的可信行为。同时，过强的安全对齐和统一化拒绝策略也可能使智能体在开放探索、长程规划和工具调用中趋于保守，表现为过度拒绝、计划缩短或避免必要尝试，从而削弱任务完成能力。

未来需从内外协同两方面推进：一方面，模型自身应具备更强的目标保持、风险识别与自我校正能力，减少长程任务中的目标漂移、推理冲突与风险失控[97]，并通过分层对齐与风险感知规划机制，在目标层、计划层和动作层施加不同粒度的安全约束，使低风险任务保持必要自主性，而高风险动作触发外部验证、权限审批或人工确认；另一方面，还需结合外部规划器、记忆模块、规则系统、安全策略与验证机制，为智能体提供状态管理、过程校验和风险拦截能力[95]。

5.2　运行环境层

运行环境层关注智能体在真实系统中“如何安全行动”。不同于传统大模型主要停留于文本生成，智能体往往需要调用工具、执行代码、访问数据库并连接外部业务系统，因此其可信性不仅取决于模型本身，还取决于运行环境能否提供清晰边界、实时约束与风险缓冲[98]。随着智能体从封闭测试走向真实部署，其风险也从输出偏差扩展为执行失控：模型幻觉、错误规划或受诱导行为可能借助工具调用直接转化为越权访问、数据泄露或系统破坏。特别是在 GUI、移动端和具身智能体场景中，智能体的动作进一步扩展为界面点击、设备控制和物理执行，错误操作可能带来敏感按钮误触发、设备损坏甚至人身安全风险。

因而，未来运行环境设计需从“支持执行”转向“受控执行”：一方面，应依据任务类型、风险等级与用户授权，动态限制智能体可访问的数据、工具和操作范围[99]，并通过权限校验、分级授权、人工确认与过程审计等机制，确保高风险操作始终处于可监督、可追踪的边界内；另一方面，还需建立面向错误操作的容错机制，通过沙盒预执行、隔离运行、状态快照、回滚恢复等方式，为智能体提供可撤销、可修复的安全缓冲，避免单次失误对真实系统造成不可逆损害。对于越权访问、数据删除、支付转账和物理动作等高风险操作，仅依赖模型自我校正并不充分，仍需运行环境提供外部强约束；在多模态和具身场景中，还应进一步引入界面状态核验、敏感操作识别、物理安全边界、速度与力矩限制、碰撞检测、仿真验证和人工急停等机制，使智能体的行动能力始终处于可授权、可验证和可接管的范围内。

5.3　系统治理层

系统治理层关注智能体在大规模部署后“如何被监督、协调与追责”。随着智能体从辅助工具逐步发展为可自主规划、调用工具并参与实际决策的

执行主体，其可信性已不能仅依赖模型训练阶段的安全对齐或运行环境中的权限限制。在多智能体协作、跨平台调用和持续在线运行场景中，风险可能在模型、工具与系统之间传递、叠加并放大，因此可信研究需从单点防护走向系统化治理。

未来治理的核心首先在于“可见”，即建立统一的可审计轨迹标准，通过记录任务规划、决策依据、工具调用、权限使用、人工确认和最终动作等关键过程，使智能体行为具备可追溯、可审计的完整轨迹，从而在误操作、安全偏离或责任争议出现时定位问题来源并厘清责任边界。

其次在于“可控”，即在复杂生态中建立统一的身份认证、权限表达、接口规范与数据流转约束[96]，避免智能体在跨平台协作中出现越权访问、数据泄露或指令注入等风险，使不同智能体和工具的连接始终处于明确授权与安全校验之下。

最后在于“可持续”，即构建持续反馈闭环，将部署后的异常行为、执行失败、用户投诉和安全事件转化为规则更新、评测修订与模型改进依据。由于真实系统中的API、网页结构、业务规则、软件版本和攻击方式会持续变化，静态基准难以及时反映智能体的真实可信边界。因此，未来可通过版本化基准、环境快照、场景生成、红队测试和生产影子评估等方式持续更新评测体系，使治理机制能够随环境、规则与攻击方式变化而动态演进。

上述三层共同构成可信智能体从能力形成到安全执行再到持续治理的完整路径。模型机理层关注智能体是否具备可靠推理、风险识别和自我校正能力，运行环境层负责将其行动限制在可授权、可验证和可接管的范围内，系统治理层则通过日志、权限记录和异常反馈实现审计、追责与持续改进。其中，模型能力是可信行动的基础，受控执行是系统审计和责任归因的前提，而治理反馈又反过来推动模型更新、规则修订和评测体系演进。未来可信智能体建设应形成“模型内生可信—运行时受控执行—系统级持续治理”的闭环机制，以支撑其在开放、动态和高风险场景中的可靠部署。

6 结束语

随着大模型能力的持续演进，AI智能体的自主性和应用边界不断扩展，其可信问题已成为制约智能体在高风险领域落地的关键瓶颈。本文从正确性与可靠性、鲁棒性、治理与可控、可解释性、安全性、隐私与合规、公平性与公正性七个维度对可信智能体的内涵进行了系统性界定，并围绕规划、工具、记忆与大模型四项核心组件，以及可信验证与多智能体可信协作两类支撑机制，对现有可信增强技术进行了分类梳理与对比分析。此外，本文归纳了当前评测基准的覆盖维度与局限性，并展望了该领域的未来发展方向。期望本文为可信智能体的进一步研究提供体系化的参考框架。

参考文献：

[1] Gu Q. Llm-based code generation method for golang compiler testing [C]//Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2023: 2201-2203.

[2] Liu F, Liu Y, Shi L, et al. Exploring and evaluating hallucinations in llm-powered code generation[J]. arXiv preprint arXiv:2404.00971, 2024.

[3] Du M, Tuan L A, Ji B, et al. Codearena: A collective evaluation platform for llm code generation[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). 2025: 502-512.

[4] Zhang R, Jiang D, Zhang Y, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 169-186.

[5] Pei Q, Wu L, Pan Z, et al. MathFusion: Enhancing Mathematical Problem-solving of LLM through Instruction Fusion[J]. arXiv preprint arXiv:2503.16212, 2025.

[6] Guo T, Chen X, Wang Y, et al. Large language model based multi-agents: A survey of progress and challenges[J]. arXiv preprint arXiv: 2402.01680, 2024.

[7] Li X, Wang S, Zeng S, et al. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges[J]. Vicinagearth, 2024, 1 (1): 9.

[8] Act E U A I. The eu artificial intelligence act[J]. European Union, 2024.

[9] Yu M, Meng F, Zhou X, et al. A survey on trustworthy llm agents: Threats and countermeasures[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. 2025: 6216-6226.

[10] Lei Y, Xu J, Liang C X, et al. Large Language Model Agents: A Comprehensive Survey on Architectures, Capabilities, and Applications[J]. 2025.

[11] Liu Z, Bai X, Chen K, et al. A survey on the feedback mechanism of LLM-based AI agents[C]//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, 2025: 10582-10592.

[12] Tang Y, Chen K, Yue L, et al. Empowering Real-World: A Survey on the Technology, Practice, and Evaluation of LLM-driven Industry Agents[J]. arXiv preprint arXiv:2510.17491, 2025.

[13] Zhang Z, Dai Q, Bo X, et al. A survey on the memory mechanism of large language model-based agents[J]. ACM Transactions on Informa-

tion Systems, 2025, 43(6): 1-47.

[14] Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models[C]//The eleventh international conference on learning representations. 2022.

[15] Shinn N, Cassano F, Gopinath A, et al. Reflexion: Language agents with verbal reinforcement learning[J]. Advances in neural information processing systems, 2023, 36: 8634-8652.

[16] Liu L, Yang X, Shen Y, et al. Think-in-memory: Recalling and post-thinking enable llms with long-term memory[J]. arXiv preprint arXiv: 2311.08719, 2023.

[17] Li M, Zhao Y, Yu B, et al. Api-bank: A comprehensive benchmark for tool-augmented llms[C]//Proceedings of the 2023 conference on empirical methods in natural language processing. 2023: 3102-3116.

[18] Liu B, Jianxiang Z, Meng D, et al. An evaluation mechanism of llm-based agents on manipulating apis[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 4649-4662.

[19] Karpas E, Abend O, Belinkov Y, et al. MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning[J]. arXiv preprint arXiv:2205.00445, 2022.

[20] Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: Language models can teach themselves to use tools[J]. Advances in neural information processing systems, 2023, 36: 68539-68551.

[21] Hong S, Zhuge M, Chen J, et al. MetaGPT: Meta programming for a multi-agent collaborative framework[C]//International Conference on Learning Representations. 2024, 2024: 23247-23275.

[22] Qian C, Liu W, Liu H, et al. Chatdev: Communicative agents for software development[C]//Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). 2024: 15174-15186.

[23] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.

[24] Zhu K, Liu Z, Li B, et al. Where llm agents fail and how they can learn from failures[J]. arXiv preprint arXiv:2509.25370, 2025.

[25] Yehudai A, Eden L, Li A, et al. Survey on evaluation of llm-based agents[J]. arXiv preprint arXiv:2503.16416, 2025.

[26] Mohammadi M, Li Y, Lo J, et al. Evaluation and benchmarking of llm agents: A survey[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. 2025: 6129-6139.

[27] EBRAHIMI S, ASUDEH A. A Survey on Reliability, Transparency, Accountability, and Fairness in LLM-based Multi-Agent Systems through the Responsibility Lens[J]. 2025.

[28] Lanham T, Chen A, Radhakrishnan A, et al. Measuring faithfulness in chain-of-thought reasoning[J]. arXiv preprint arXiv:2307.13702, 2023.

[29] Fu Y, Qiu R, Wang X, et al. Beyond Blind Following: Evaluating Robustness of LLM Agents under Imperfect Guidance[C]//Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2026: 6591-6618.

[30] Wu Q, Gao P, Liu W, et al. Backtrackagent: Enhancing gui agent with error detection and backtracking mechanism[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 4250-4272.

[31] Wang H, Poskitt C M, Sun J. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents.(2026)[C]//Proceedings of the IEEE/ACM International Conference on Software Engineering, ICSE. 2026: 12-18.

[32] Schmitz C, Rystrøm J, Batzner J. Oversight structures for agentic AI in public-sector organizations[C]//Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025). 2025: 298-308.

[33] Romero M L, Suyama R. Agentic ai for intent-based industrial automation[C]//2025 16th IEEE International Conference on Industry Applications (INDUSCON). IEEE, 2025: 437-444.

[34] Bouzenia I, Pradel M. Understanding software engineering agents: A study of thought-action-result trajectories[J]. arXiv preprint arXiv: 2506.18824, 2025.

[35] Neelou E, Novikov I, Moroz M, et al. A2AS: agentic AI runtime security and Self-Defense[J]. arXiv preprint arXiv:2510.13825, 2025.

[36] He F, Zhu T, Ye D, et al. The emerged security and privacy of llm agent: A survey with case studies[J]. ACM Computing Surveys, 2025, 58(6): 1-36.

[37] Wang B, He W, Zeng S, et al. Unveiling privacy risks in llm agent memory[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 25241-25260.

[38] Binkyte R. Interactional Fairness in LLM Multi-Agent Systems: An Evaluation Framework[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2025, 8(1): 457-468.

[39] Liu Y, Yao Y, Ton J F, et al. Trustworthy llms: a survey and guideline for evaluating large language models' alignment[J]. arXiv preprint arXiv:2308.05374, 2023.

[40] Besta M, Blach N, Kubicek A, et al. Graph of thoughts: Solving elaborate problems with large language models[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(16): 17682-17690.

[41] Ning X, Zhao Y, Liu Y, et al. Dgot: Dynamic graph of thoughts for scientific abstract generation[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024: 4832-4846.

[42] Madaan A, Tandon N, Gupta P, et al. Self-refine: Iterative refinement with self-feedback[J]. Advances in neural information processing systems, 2023, 36: 46534-46594.

[43] Liang T, He Z, Jiao W, et al. Encouraging divergent thinking in large language models through multi-agent debate[C]//Proceedings of the 2024 conference on empirical methods in natural language processing. 2024: 17889-17904.

[44] Pitre P, Ramakrishnan N, Wang X. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 22112-22133.

[45] Zhao H, Guo Z, Zhang C, et al. StructuThink: Reasoning with Task Transition Knowledge for Autonomous LLM-Based Agents[C]//Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 24489-24506.

[46] Chen R, Sun Y, Wang J, et al. Safemind: benchmarking and mitigating safety risks in embodied llm agents[J]. arXiv preprint arXiv: 2509.25885, 2025.

[47] Basu K, Abdelaziz I, Chaudhury S, et al. Api-blend: A comprehensive corpora for training and benchmarking api llms[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics

(Volume 1: Long Papers). 2024: 12859-12870.

[48] Zhang J, Lan T, Zhu M, et al. xlam: A family of large action models to empower ai agent systems[C]//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: 11583-11597.

[49] Sengupta S, Zhou Z, Araki J, et al. Tooldreamer: Instilling llm reasoning into tool retrievers[C]//Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2026: 5465-5482.

[50] Wang R, Han X, Ji L, et al. Toolgen: Unified tool retrieval and calling via generation[J]. arXiv preprint arXiv:2410.03439, 2024.

[51] Zhong W, Guo L, Gao Q, et al. Memorybank: Enhancing large language models with long-term memory[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(17): 19724-19731.

[52] Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. 2025. Compress to Impress: Unleashing the Potential of Compressive Memory in Real-World Long-Term Conversations. In Proceedings of the 31st International Conference on Computational Linguistics, pages 755 – 773, DhabiAbu, UAE. Association for Computational Linguistics.

[53] Li H, Yang C, Zhang A, et al. Hello again! llm-powered personalized agent for long-term dialogue[C]//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: 5259-5276.

[54] Packer C, Wooders S, Lin K, et al. MemGPT: Towards LLMs as Operating Systems[J]. arXiv e-prints, 2023: arXiv: 2310.08560.

[55] Fountas Z, Benfeghoul M A, Oomerjee A, et al. Human-inspired episodic memory for infinite context LLMs[J]. arXiv preprint arXiv: 2407.09450, 2024.

[56] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//NeurIPS. 2022.

[57] Parihar S, Guangliang L, Parde N, et al. Context-Aware Counterfactual Data Augmentation for Gender Bias Mitigation in Language Models [J]. arXiv preprint arXiv:2602.09590, 2026.

[58] Kim S, Lee D, Lee J. KLAAD: Refining Attention Mechanisms to Reduce Societal Bias in Generative Language Models[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 15324-15345.

[59] Gallegos I O, Aponte R, Rossi R A, et al. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). 2025: 873-888.

[60] Chuang Y N, Wang G, Chang C Y, et al. FaithLM: Towards faithful explanations for large language models[C]//Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2026: 3802-3824.

[61] Dhuliawala S, Komeili M, Xu J, et al. Chain-of-verification reduces hallucination in large language models[C]//Findings of the association for computational linguistics: ACL 2024. 2024: 3563-3578.

[62] Zhang F, Xu J, Wang C, et al. Incentivizing LLMs to Self-Verify Their Answers[J]. arXiv preprint arXiv:2506.01369, 2025.

[63] Zhang L, Hosseini A, Bansal H, et al. Generative verifiers: Reward modeling as next-token prediction[J]. arXiv preprint arXiv: 2408.15240, 2024.

[64] Mekala D, Weston J E, Lanchantin J, et al. Toolverifier: Generalization to new tools via self-verification[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 5026-5041.

[65] Chern I, Chern S, Chen S, et al. FacTool: Factuality Detection in Generative AI--A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios[J]. arXiv preprint arXiv:2307.13528, 2023.

[66] Jin Y, Xu K, Li H, et al. ReVeal: Self-Evolving Code Agents via Reliable Self-Verification[J]. arXiv preprint arXiv:2506.11442, 2025.

[67] Zhang Y, Emma S Y, En A L J, et al. Rvllm: Llm runtime verification with domain knowledge[J]. arXiv preprint arXiv:2505.18585, 2025.

[68] Miculicich L, Parmar M, Palangi H, et al. Veriguard: Enhancing llm agent safety via verified code generation[J]. arXiv preprint arXiv: 2510.05156, 2025.

[69] Raza S, Sapkota R, Karkee M, et al. Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems[J]. arXiv preprint arXiv:2506.04133, 2025.

[70] Hallyburton R S, Pajic M. Bayesian methods for trust in collaborative multi-agent autonomy[C]//2024 IEEE 63rd Conference on Decision and Control (CDC). IEEE, 2024: 470-476.

[71] Akbari B, Yuan M, Wang H, et al. A factor graph model of trust for a collaborative multi-agent system[J]. arXiv preprint arXiv: 2402.07049, 2024.

[72] Chen K, Wang T, Zhu H, et al. Maximizing group utilities while avoiding conflicts through agent qualifications[J]. IEEE Transactions on Computational Social Systems, 2024, 12(2): 552-562.

[73] Mu C, Najib M, Oren N. Responsibility-aware strategic reasoning in probabilistic multi-agent systems[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(22): 23258-23266.

[74] Ronanki K. Facilitating Trustworthy Human-Agent Collaboration in LLM-based Multi-Agent System oriented Software Engineering[C]// Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering. 2025: 1333-1337.

[75] Putrevu A. A governance framework for agentic AI: mitigating systemic risks in LLM-powered multi-agent architectures[J]. Journal of International Crisis and Risk Communication Research, 2025: 359-369.

[76] Wu Q, Bansal G, Zhang J, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation[J]. arXiv preprint arXiv: 2308.08155, 2023.

[77] Qin Y, Liang S, Ye Y, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis[J]. arXiv preprint arXiv:2307.16789, 2023.

[78] Guo Z, Cheng S, Wang H, et al. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models[C]// Findings of the Association for Computational Linguistics: ACL 2024. 2024: 11143-11156.

[79] Chen W, Su Y, Zuo J, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors[C]//International Conference on Learning Representations. 2024, 2024: 20094-20136.

[80] Zhou S, Xu F F, Zhu H, et al. Webarena: A realistic web environment for building autonomous agents[C]//International Conference on Learning Representations. 2024, 2024: 15585-15606.

[81] Yang R, Chen H, Zhang J, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven

embodied agents[J]. arXiv preprint arXiv:2502.09560, 2025.

[82] Yin S, Pang X, Ding Y, et al. Safeagentbench: A benchmark for safe task planning of embodied llm agents[J]. arXiv preprint arXiv: 2412.13178, 2024.

[83] Zheng K, Chen X, Jenkins O C, et al. Vlmbench: A compositional benchmark for vision-and-language manipulation[J]. Advances in Neural Information Processing Systems, 2022, 35: 665-678.

[84] Yuan T, He Z, Dong L, et al. R-judge: Benchmarking safety risk awareness for llm agents[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 1467-1490.

[85] Li H, Hu W, Jing H, et al. Privaci-bench: Evaluating privacy with contextual integrity and legal compliance[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 10544-10559.

[86] Jiang L, Rao K, Han S, et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 47094-47165.

[87] Xia H, Wang H, Liu Z, et al. SafeToolBench: Pioneering a Prospective Benchmark to Evaluating Tool Utilization Safety in LLMs[J]. arXiv preprint arXiv:2509.07315, 2025.

[88] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM computing surveys, 2023, 55(12): 1-38.

[89] Gallegos I O, Rossi R A, Barrow J, et al. Bias and fairness in large language models: A survey[J]. Computational linguistics, 2024, 50(3): 1097-1179.

[90] Liu X, Yu H, Zhang H, et al. Agentbench: Evaluating llms as agents [C]//International Conference on Learning Representations. 2024, 2024: 52989-53046.

[91] White C, Dooley S, Roberts M, et al. Livebench: A challenging, contamination-limited llm benchmark[C]//International Conference on Learning Representations. 2025, 2025: 91595-91631.

[92] Jimenez C E, Yang J, Wettig A, et al. Swe-bench: Can language models resolve real-world github issues? [C]//International Conference on Learning Representations. 2024, 2024: 54107-54157.

[93] Ruan Y, Dong H, Wang A, et al. Identifying the risks of lm agents with an lm-emulated sandbox[C]//International Conference on Learning Representations. 2024, 2024: 27031-27098.

[94] Liu Y, Hou Z, Xu X, et al. HEV Generative Sandbox: A Framework for Assessing Domain-Specific Social Risks Through Human-LLM Simulation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2026, 40(38): 32249-32257.

[95] Zhang Y, Cai Y, Zuo X, et al. Position: Trustworthy AI agents require the integration of large language models and formal methods[C]// Forty-second International Conference on Ma-chine Learning Position Paper Track. 2025.

[96] Kong D, Lin S, Xu Z, et al. A survey of llm-driven ai agent communication: Protocols, security risks, and defense coun-termeasures[J]. arXiv preprint arXiv:2506.19676, 2025.

[97] Zheng J, Shi C, Cai X, et al. Lifelong learning of large language model based agents: A roadmap[J]. IEEE Transactions on Pat-tern Analysis and Machine Intelligence, 2026.

[98] Hu X, Xu Z. Large language and vision-language models for robot: safety challenges, mitigation strategies and future direc-tions[J]. Industrial Robot: the international journal of robotics research and application, 2025.

[99] Lan H, Zhou Z, Zhu Q, et al. Heterogeneous Confidential Computing System for Large Language Models: A Survey[J]. ACM Transactions on Architecture and Code Optimization, 2026, 23(1): 1-26.

[100] Kawabata A, Sugawara S. Rationale-aware answer verification by pairwise self-evaluation[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 16178-16196.

[101] Li B, Qi P, Liu B, et al. Trustworthy AI: From principles to practices [J]. ACM Computing Surveys, 2023, 55(9): 1-46.

刘井平（1982– ），男，广东潮州人，中山大学软件工程学院副教授、博士生导师，主要研究方向为软件测试与程序分析、大模型智能体、自然语言处理、知识图

郑子彬（1982– ），男，广东潮州人，中山大学软件工程学院教授、博士生导师，主要研究方向为可信大模型，软件可靠性，程序分析，区块链，智能合约，可信软件。

李卫（2000– ），男，山西忻州人，中山大学软件工程学院博士生，主要研究程序分析与智能软件工程，聚焦软件可靠性、安全与演化。

陈玉轩（1998– ），男，湖南湘潭人，中山大学软件工程学院博士生，主要研究方向大语言模型、代码检索、RAG。

张泽凯（2000– ），男，广东广州人，中山大学软件工程学院博士生，主要研究方向为人工智能与软件工程、大语言模型赋能软件工程。