

LLM 智能体的评估与基准测试：一项综述

Mahmoud Mohammadi mahmoud.mohammadi@sap.com SAP Labs 美国
华盛顿州 贝尔维尤

Jane Lo jane.lo@sap.com SAP Labs 帕罗奥图, 加利福尼亚州, 美国

摘要

基于 LLM 的智能体的兴起为人工智能应用开辟了新的前沿, 然而对这些智能体的评估仍然是一个复杂且尚未成熟的领域。本综述对新兴的 LLM 智能体评估领域进行了深入概述, 提出了一种二维分类法, 从以下两个维度对已有研究进行组织: (1) 评估目标——评估什么, 例如智能体行为、能力、可靠性和安全性; (2) 评估过程——如何评估, 包括交互模式、数据集与基准、指标计算方法以及工具支持。除分类法外, 我们还重点指出了企业级场景特有的挑战, 例如基于角色的数据访问权限、对可靠性保障的需求、动态且长周期的交互以及合规性要求, 这些在当前研究中往往被忽视。我们还识别了未来的研究方向, 包括更加整体化、真实可扩展的评估。本工作旨在厘清智能体评估这一碎片化领域的现状, 并提供一个系统化的评估框架, 使研究者和从业者能够针对真实部署场景对 LLM 智能体进行评估。

CCS 概念分类

· 计算方法学 → 自然语言处理; · 软件及其工程 → 软件验证与确认; · 以人为中心的计算 → 人机交互 (HCI)。

关键词

大语言模型智能体; 智能体评估; 评估分类法; 智能体行为、基准测试、安全性;
企业人工智能

ACM 引用格式: Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and Benchmarking of LLM Agents: A Survey. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD ' 25), August 3–7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3711896.3736570>

本作品采用知识共享署名 4.0 国际许可协议 (Creative Commons Attribution 4.0 International License) 授权。KDD ' 25, 加拿大安大略省多伦多 © 2025 版权所有人为作者本人。ACM ISBN 979-8-4007-1454-2/2025/08 <https://doi.org/10.1145/3711896.3736570>

Yipeng Li yipeng.li@sap.com SAP Labs 美国华盛顿州贝尔维尤

Wendy Yip wendy.yip@sap.com SAP Labs 美国 加利福尼亚州 帕洛阿尔托

1 引言

基于 LLM 的智能体是利用 LLM 进行推理、规划与行动的自主或半自主系统，代表了人工智能领域中一个快速发展的前沿方向 [69, 105]。从客服机器人、编程助手到数字助理，LLM 智能体正在重新定义我们构建智能系统的方式。随着这些智能体从研究原型走向真实应用 [23, 62]，如何对其进行严格评估的问题变得愈发紧迫而复杂。然而，评估 LLM 智能体比单纯评估 LLM 更为复杂。与主要针对文本生成或问答能力进行评估的 LLM 不同，LLM 智能体运行在动态、交互式的环境中。它们进行推理和规划、调用工具、利用记忆，甚至与人类或其他智能体进行协作 [20]。这种复杂的行为以及对真实世界效应的依赖，使得标准的 LLM 评估方法不再适用。打个比方：LLM 评估类似于考察发动机的性能，而智能体评估则需要全面评估一辆汽车在各种驾驶条件下的整体表现。LLM 智能体评估也不同于传统的软件评估。软件测试关注的是确定性的、静态的行为，而 LLM 智能体天然具有概率性并表现出动态行为；因此，对其性能的评估需要采用全新的方法。LLM 智能体评估处于自然语言处理（NLP）、人机交互（HCI）和软件工程的交叉地带，需要引入更多元的视角。尽管该领域受到越来越多的关注，现有综述要么仅聚焦于 LLM 评估，要么仅覆盖智能体的某些特定能力，缺乏整体性的视角 [121]。此外，企业级应用对智能体提出了额外的要求，包括对数据和系统的安全访问、为审计与合规目的所需的高可靠性，以及更复杂的交互模式，而这些问题在现有文献中鲜有涉及 [107]。本综述旨在为智能体评估领域的研究者和从业者提供一份有益的参考。本综述的贡献体现在以下两个方面。

- 我们提出了一种 LLM 智能体评估的分类法，从评估目标（评估什么，例如行为、能力、可靠性和安全性）和评估过程（如何评估，包括交互模式、数据集与基准、指标计算方法、评估工具以及评估环境）对已有研究进行组织。

我们着重指出企业场景下的特有挑战，包括基于角色的访问控制、可靠性保障、长期交互以及合规要求。

本文其余部分的组织结构如下：第 2 节描述本文综述所采用的分类法，用以分析智能体评估领域的整体格局。第 3 节深入探讨分类法的第一个维度——评估目标，聚焦于待评估的智能体各方面特性。第 4 节描述第二个维度——评估过程，聚焦于评估方法。第 5 节讨论在企业环境中评估 LLM 智能体所面临的挑战。第 6 节概述开放性问题与未来研究方向，以指导 LLM 智能体评估下一阶段的工作。

2 LLM-based 智能体评估分类法

我们提出一个二维分类法，从评估目标（评估什么）和评估过程（如何评估）两个轴线来组织 LLM-based 智能体评估的各个方面。该分类法以层次树的形式在图 1 中可视化呈现。评估目标维度关注的是评估的对象。该维度的第一类——智能体行为，聚焦于面向结果层面的特性，如任务完成度和输出质量，反映智能体在多大程度上满足终端用户的期望。其次，智能体能力强调面向过程的胜任力，包括工具使用、规划与推理、记忆与上下文保持，以及多智能体协作。这些能力能够揭示智能体如何达成其目标，以及其对设计规格的契合程度。可靠性评估智能体在相同输入下行为的一致性，以及在输入变化或系统遭遇错误时的鲁棒性。最后，安全与对齐评估智能体的可信度与安全性，涵盖公平性、合规性以及有害或不道德行为的

防范。评估过程维度描述如何对智能体进行评估。交互模式区分静态评估（智能体对固定输入作出响应）与交互式评估（智能体与用户互动）。评估数据涵盖合成数据集、真实世界数据集，以及面向软件工程、医疗健康、金融等特定领域的基准[23, 62]。指标计算方法既包括任务成功率、事实准确性等定量度量，也包括基于人类或 LLM 判断的定性评估。评估工具支撑指用于支撑评估的基础设施，例如 LangSmith、Arize AI 等插桩框架，以及 Holistic Evaluation of Agents 等公开排行榜，它们使可扩展且可复现的评估成为可能。最后，评估上下文定义了开展评估所处的环境，从受控仿真到网页浏览器或 API 等开放世界场景。该分类法既可作为概念框架，也可作为实践指南，能够在广泛的目标、方法学与部署条件下对 LLM 智能体进行系统化的比较与分析。在后续各节中，我们将逐一详述每个维度，着重介绍关键的评估实践与代表性研究。

随着 LLM 智能体被部署到日益多样且复杂的环境中，单轮与多轮交互、多语言能力以及多模态等因素都变得愈发重要。尽管该分类法仍然适用于这些变体场景，但通常需要有针对性的指标与评估策略。我们将在后续相关章节中讨论这些具体的适配方法。

3 评估目标

3.1 智能体行为

智能体行为是指用户所感知到的智能体整体表现，此时将智能体视为一个黑盒。它代表了评估中最高层级的视角，能够最直接地反映用户体验。该范畴涵盖任务完成度、输出质量、延迟和成本等方面。

3.1.1 任务完成：任务完成是智能体评估的基本目标，用于评估智能体是否成功达成给定任务的预设目标[11, 80, 96, 115]。它涉及判断是否达到期望状态，或是否满足为任务成功所定义的特定标准[57, 93]。尽管有时被指出在提供细粒度失败洞察方面存在局限，尤其在大多数模型成功率较低的情况下[64]，任务完成仍然是衡量智能体整体性能的主流且必要的指标[64]。

任务完成通常使用成功率（Success Rate, SR）[11, 57, 64]等指标进行量化，该指标也可称为任务成功率（Task Success Rate）[115]或总体成功率（Overall Success Rate）[57]。其他相关指标还包括任务目标完成度（Task Goal Completion, TGC）[91]和通过率（Pass Rate）[76]。部分评估采用二元指标，例如返回 0 或 1 的奖励函数以表示目标达成情况[42]。 $\text{pass}@k$ 与 pass^k 等指标则进一步考虑了多次试验中的成功情况[104]。

这一关键目标被广泛应用于各类 LLM 智能体评估领域与基准测试中[93]，涵盖编码与软件工程相关任务，例如解决 GitHub issue（SWE-bench [40]）、科学数据分析编程（ScienceAgentBench [11]）、复现研究（CORE-Bench [86]、PaperBench [87]）以及应用中的交互式编程（AppWorld [91]）。它同样广泛用于评估与网页环境交互的智能体，包括通用网页导航（BrowserGym [14]、WebArena [126]、WebCanvas [73]）、多模态网页任务（VisualWebArena [42]、MMInA [124]）以及真实且耗时的网页任务（ASSISTANTBENCH [109]）。

3.1.2 输出质量：输出质量指 LLM 智能体响应所具备的各项特征，是一个涵盖准确性、相关性、清晰度、连贯性以及是否遵循智能体规范或任务要求等维度的统称[80]。即便智能体完成了任务，若交互过程缺乏上述质量，用户的实际体验仍可能不佳。输出质量在对话式智能体的评估中尤为相关，因为用户目标通常需要经过多轮交互才能达成。该类别中的许多指标与 LLM 评估中所用指标存在重叠。例如，流畅性（fluency）指标用于衡量 LLM 输出符合自然语言规范的程度[120]。逻辑连贯性（logical coherence）指标则关注

论证过程的严谨程度[120]。由于 LLM 智能体可能借助工具检索事实性信息并提供具有上下文感知的文本回答，检索增强生成（RAG）系统中的标准指标同样适用，这类指标包括响应相关性（Response Relevance）与事实正确性（Factual Correctness）[21]。

3.1.3 延迟与成本：延迟是智能体行为的关键方面，尤其在用户与智能体进行同步交互的场景中。较长的等待时间会显著降低用户体验并削弱其对系统的信任。此类场景下常用的指标为首 token 时间（Time To First Token, TTFT），用于衡量用户在流式模式下看到 LLM 响应首 token 之前的延迟。对于智能体以异步方式运行的用例，端到端请求延迟（End-to-End Request Latency）——即获得完整响应所需时间——通常更为相关[70]。

虽然成本无法被终端用户直接观察，但它在决定智能体大规模部署的可行性方面起着至关重要的作用。我们将成本作为智能体货币效率的一项度量，其通常依据输入与输出 token 的数量进行估算，而这些数量与大多数 LLM 部署中的按用量计费模式直接相关。

3.2 智能体能力

除外部行为外，评估工作还常常聚焦于 LLM-based agents 中支撑其性能的具体能力。该类别的关键方面包括工具使用、规划与推理、

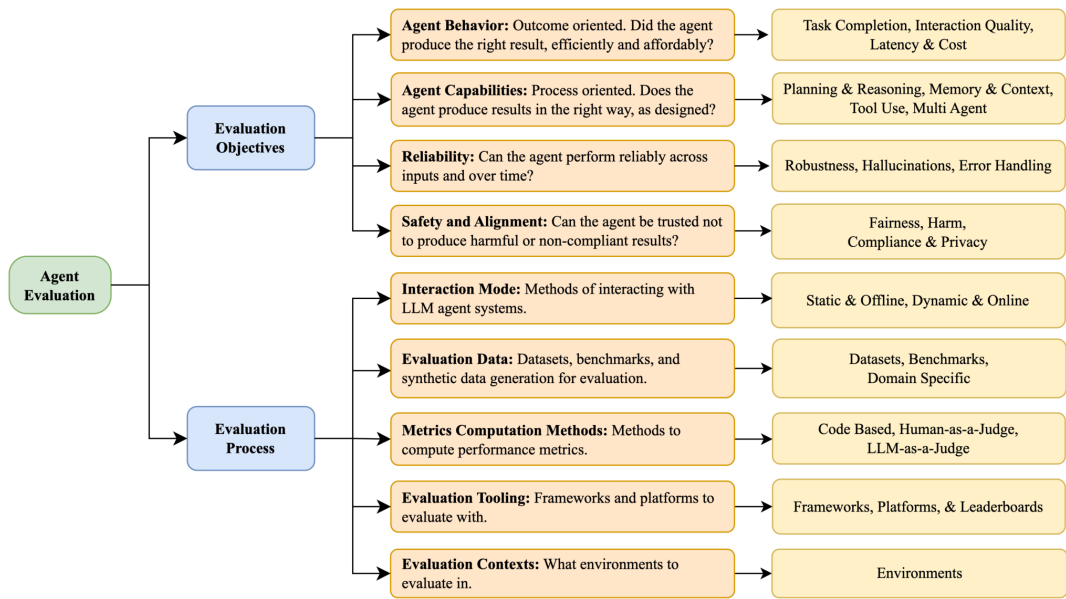


图 1: LLM 智能体评估分类体系

记忆与上下文保持，以及多智能体协作。对这些能力进行评估有助于在更细粒度层面判断智能体的优势与不足。

3.2.1 工具使用：工具使用是基于 LLM 的智能体的核心能力，使其能够获取外部知识、执行操作并与外部环境交互。在本综述中，工具使用涉及对单个工具的调用，并与函数调用通用；关于在复杂任务中确定工具调用顺序的更复杂情形将在 3.2.2 节中讨论。ChatGPT-3.5 及之后的 LLM 等最新进展已原生支持函数调用，这些模型可以自主决定是否调用函数、从候选集中选取合适的函数并生成所需参数。因此，LLM 智能体可以直接利用底层模型的函数能力，从而将许多原本针对 LLM 设计的评估技术应用于工具使用 [47]。工具使用的评估涉及回答若干关键问题。首先，智能体能否正确判断给定任务是否需要调用工具？若需要，能否从预定义的候选集中选取合适的工具？工具选定后，智能体还需准确识别该工具所需的参数，并为每个参数生成合适的取值以确保正确执行。在候选工具集规模较大的情形下，智能体可能还需要根据任务自然语言描述从工具库中检索正确的工具 [83]。

研究人员提出了若干评估上述能力的指标。调用准确率 [54] 用于衡量智能体关于是否调用工具的决策是否正确。工具选择准确率用于衡量是否从工具列表中选出了合适的工具。检索准确率关注系统能否从更大规模的工具集中检索到正确工具，常以排名准确率 k 衡量。在基于排名的评估中，平均倒数排名（MRR）刻画了正确工具在排序列表中的位置；而归一化折损累计增益（NDCG）则反映了系统对所有相关工具的排序质量 [54]。参数相关评估涵盖两个层面。参数名称 F1 分数 [83] 用于衡量智能体能否正确识别给定函数所需的参数名称并为其正确赋值。部分评估依赖抽象语法树（AST）的正确性来判断工具调用在语法层面是否合法，但该方法可能遗漏语义层面的错误，例如参数取值的错误或幻觉，特别是对取值受枚举类型约束的参数 [75]。针对这一局限，Gorilla 等近期工作提出了基于执行的评估方法，即实际运行工具调用并评估其结果，从而为工具使用能力提供更全面、更贴近真实表现的评估 [75]。

规划与推理：规划与推理是 LLM-based agents 的核心能力，对于需要在不确定性下执行多步骤或进行决策的复杂任务尤为重要。规划涉及以恰当顺序选取正确的工具集合；与此同时，推理使 agent 能够在任务执行前或执行过程中根据上下文做出决策 [37]。T-eval [12] 将规划评估形式化为将预测的工具集合与参考集合进行比较。由于工具的顺序与依赖关系同样关键，部分基准采用基于图的表示方式，并引入节点 F1（评估工具选择）以及边 F1 或归一化编辑距离（评估工具调用序列与结构准确性）等指标 [83]。在动态环境中，agent 通常需要交替进行规划与执行，根据不断演化的上下文调整自身行为 [37]。这一模式由 ReAct 范式体现，其中 agent 在推理步骤与工具使用之间反复切换 [106]。对此类自适应推理的评估不能仅止于对比静态计划，而需要能够反映实时决策的指标。T-Eval 框架 [12] 通过引入一种推理指标来应对这一问题：逐步骤衡量 agent 预测的下一工具调用与期望调用之间的契合程度，从而刻画 agent 在工具输出未知时作出明智决策的能力。类似地，AgentBoard [64] 提出了进度率（Progress Rate）指标，将 agent 的实际轨迹与期望轨迹进行对比，精细衡量其朝目标推进的有效程度。当 agent 被要求以生成完整多步程序的方式进行规划时，代码生成领域的评估方法便具有了相关

性。ScienceAgentBench 等基准使用程序相似性指标，将生成的规划与标注参考进行比较[11]。此外，步成功率（Step Success Rate）被提出用于衡量所生成规划中各步骤成功执行的比例，从而在执行过程中对规划质量提供整体视角[28]。

记忆与上下文保持：对于长时间运行的 agent 而言，关键能力在于在多轮交互中保留信息，并将先前上下文应用于当前请求。Guan 等人[31]将多轮对话中的记忆评估按记忆跨度（信息存储时长）与记忆形式（信息表征方式）进行分类。例如，LongEval[43]与 SocialBench[9]是测试 agent 在长对话（40 轮以上）中上下文保持能力的基准。agent 可能会获得一段跨越数十轮交互的对话，并在后续被询问需要回溯对话早期细节的问题。Maharana 等人[65]展示了跨度达数百轮（600 轮以上）的对话评估，Li 等人[50]则引入了记忆增强评估技术，追踪 agent 在长周期任务中维持一致性的能力。这些评估通常使用合成或记录式对话作为数据集，指标包括事实召回准确率（Factual Recall Accuracy）与一致性得分（Consistency Score，即各轮之间不存在矛盾）。针对使用工具的 agent，记忆评估还可考虑工作记忆（即 agent 是否跟踪中间结果）以及遗忘策略（即是否能恰当地遗忘无关细节以避免混淆）。

多 agent 协作：在 LLM-based 系统中评估多 agent 协作需要采用与传统基于强化学习的协同方法不同的方法论[7,48,89]。与依赖预定义奖励结构的传统 agent 不同，LLM agents 通过自然语言、策略性推理与去中心化的问题求解进行协调[32,33]。这些能力在金融决策与结构化数据分析等真实应用中至关重要，在这些场景中，autonomous agents 必须高效地交换信息、协商并同步决策过程[50,55]。面向协作任务的 autonomous agents（Autonomous Agents for Collaborative Tasks）[55]评估了协作效率（Collaborative Efficiency），衡量多个 agent 共享职责与动态分配任务的程度。

3.3 可靠性

可靠性是一项关键目标，尤其是在 LLM agents 被考虑用于企业级和安全关键型应用时。它涵盖 agent 输出的一致性、对抗扰动的鲁棒性以及可信度。与任务性能（可能衡量最佳情况能力）不同，可靠性评估考察的是最坏情况与平均情况下的表现。

3.3.1 一致性：一致性指在多次重复同一任务时输出的稳定性 [54]。由于 LLM 本身具有非确定性，基于 LLM 的智能体在行为上同样会表现出波动。若要将智能体部署于企业或其他高风险场景并获得信任，必须能够在重复执行相同任务时保持稳定表现。此场景下常用的指标是 $\text{pass}@k$ ，衡量智能体在 k 次尝试中至少一次成功的概率。然而，更严格的一致性衡量标准是智能体是否在全局 k 次尝试中均能成功。 τ -benchmark 将其形式化为 pass^τ 指标，能够更准确地反映关键任务部署场景下对一致性的要求。

3.3.2 鲁棒性：鲁棒性指智能体在面对输入变化或环境变化时输出结果的稳定性。为保持有效性与可信赖性，基于 LLM 的智能体必须在一系列具有挑战性的条件下持续提供高质量表现。鲁棒性评估通常通过施加扰动的输入来对智能体进行压力测试，例如改写后的指令、干扰性或误导性上下文，以及拼写错误、方言等语言变体，以考察其是否仍能成功完成任务。例如，鲁棒性评估可能包括对标准提示进行

系统性变换，并衡量任务成功率或输出质量的下降程度。HELM 基准 [51] 显式纳入了此类测试，追踪模型在输入变化下的性能退化情况。鲁棒性还涵盖适应性韧性，即智能体在环境动态变化时恢复的能力。例如，WebLinX [63] 研究了当网页结构在执行过程中发生变化时智能体的行为。在此类场景中，高效的智能体必须能够调整策略，而不是停滞或失败。在使用工具的智能体中，鲁棒性还进一步体现在错误处理能力上。正如 ToolEmu 的评估 [82] 所示，智能体必须能够从容地应对工具调用失败或意外输出。鲁棒性测试可以包括刻意注入的故障，例如 API 错误或空响应，以观察智能体能否恢复（例如重试、切换工具或向用户解释问题）抑或彻底崩溃。一项关键指标是可被恰当处理的注入故障比例，反映智能体在不确定或不完善条件下的可靠性。

3.4 安全性与对齐

安全性涵盖智能体对伦理规范的遵循、对有害行为的规避，以及对法律或政策约束的遵守。随着 LLM agents 能力与自主性的不断增强，诸如生成虚假信息、仇恨言论或不安全指令等非预期负面后果的风险也随之上升，使安全性评估不可或缺。在金融服务、网络安全以及自主决策等领域，这些评估尤为关键，因为智能体的脆弱性可能引发严重后果 [22, 26, 34, 49]。

3.4.1 公平性：AI 智能体在公平性与透明度方面的缺失可能导致有偏的结果、用户信任下降以及非预期的社会后果 [16]。例如在金融应用中，贷款审批或投资策略中的有偏决策会固化系统性不平等（FinCon [111]、AutoGuide [24]）。在多智能体交互中同样存在伦理层面的关切，决策框架必须确保符合规范与社会准则 [67]。可解释性对于增强用户信任至关重要，尤其是在 AI 智能体提供推荐或自动化辅助的交互式系统中。基于准则的决策方法（AutoGuide [24]）以及结构化的透明度机制（MATSA [66]、FinCon [111]）能够为用户提供清晰的推理路径。同时，Rjudge [112] 分析了智能体在自主决策时如何感知风险，强调了 AI 交互中的透明性与可信度。对上述维度的评估能够确保 AI 智能体在契合伦理标准的同时，在其运行情境中保持公平性。

3.4.2 危害、有害性与偏见：安全性的一个方面是确保智能体的输出不包含仇恨言论、骚扰或严重偏见等有害内容。有害性评估通常使用专门的测试集和指标，例如 RealToxicityPrompts 数据集 [27]——一个旨在引出有害内容的提示集合——其响应会通过自动化有害性检测器或人工标注员进行核查。指标包括含有有害语言的响应占比或平均有害性得分（由分类器给出）。HELM [51] 将有害性与偏见指标纳入整体评估范畴，用于反映模型产生冒犯性内容或表现出不良偏见的频率。对于交互式智能体，可以通过提供挑衅性或伦理上具有挑战性的输入（红队测试）来评估它，进而测量其失败率（即以不安全方式响应的频率）。专注于安全的数据集，如 CoSafe，正是针对此目标：Yu 等人提出 CoSafe [110] 来评估对话智能体在旨在诱使其违反安全规则的对抗性提示下的表现（例如，用户隐晦地请求自伤建议或非法指令）。CoSafe 揭示出即使是先进的智能体也存在漏洞，例如易遭受指代消解型攻击（即用户通过含混指代来绕过过滤机制）。评估过程包括在面对这些对抗性查询时监测智能体的响应是否违反策略。借助数值化得分（例如“智能体在 X% 的对抗性案例中产生了违规响应”），可以量化安全性。

3.4.3 合规性与隐私：除避免明显的有害输出外，许多部署场景要求智能体遵守特定的法规或政策限制 [6, 123]。例如，金融聊天机器人不得泄露机密信息或提供特定类型的理财建议，医疗助手则不得偏离既定的医学指南。合规性评估往往高度依赖具体领域，因为它涉及精心设计的场景以检验智能体是否遵守边界（例如，用户向其索取仅限处方药建议——正确、安全的行为是拒绝并建议咨询医生）。在企业环境中，合规性评估可能需要反映实际策略的专有测试用例。一种思路是将这些考量整合到评估框架中：例如，针对企业的 HELM 基准 [51] 旨在为金融、法律等领域纳入领域特定的提示与指标（如金融术语的准确率或响应合规率）。其过程包括收集具有代表性的企业场景（可能包含机密或定制数据），并设计反映真实世界成功标准的评估指标（例如：智能体的响应是否满足了所有法律免责声明的要求？）。例如，TheAgentCompany [97] 在结构化正确性约束下评估企业 AI 智能体，要求其在完成任务时遵守预定义的组织策略。

4 评估流程

4.1 交互模式

LLM 智能体的评估可在多种交互模式下、借助不同工具进行。一个根本性的区分在于离线评估（使用预先生成的静态数据集）与在线评估（涉及反应式仿真、人在环或真实系统的实时监测）。

表 1：评估目标及其指标。

目标类别指标相关文献智能体行为任务完成成功率（SR）、F1 分数、Pass@k、进度率、执行准确率、迁移学习成功率、零样本泛化准确率

AgentBoard [8]、WebShop [103]、AgentBench [58]、Tool Use Evaluation [53]、InformativeBench、SQuAD [79][78][56]、ResearchArena [41]、AgentBoard [8]、AppWorld [91]、TheAgentCompany [97]、MAGIC [99]、Mobile-Env [117]、Re-ReST [19]、XMC-AGENT [59]、SWE-bench [40]。

输出质量

连贯性、用户满意度、可用性、喜爱度、整体质量

PredictingIQ [80]、EnDex [98]、PsychoGAT [101]

延迟与成本延迟、Token 用量、成本Cluster diagnosis [84]、MobileBench [18]、MobileAgentBench [92]、LangSuitE [39]、WebArena [126]、Mobile-env [116]、GUI Agents [94]、GPTDroid [60]、Spa-bench [10]智能体能力工具使用任务完成率、工具选择准确率

ToolEmu [81]、MetaTool [38]、AutoCodeRover [119]

规划与推理推理质量、准确率、细粒度进度率、自一致性、规划质量

AgentBoard [8]、Massive Multitask Language Understanding [36]、LLM-Augmented Autonomous Agents [55]、Cluster diagnosis questions [84]、SimuCourt [35]、Magis [90]——记忆与上下文保持

LongEval [43]、SocialBench [9]、LoCoMo [65]、Optimus-1 [50]多智能体协作

事实准确性召回率、一致性得分

AgentSims [52]、WebArena [126]、MATSA [66]、GAMEBENCH [15]、BALROG [72]、TheAgentCompany [97] 可靠性 一致性 $\text{pass}^k \tau$ -Bench [104]
鲁棒性 扰动下的准确率与任务成功率

Reliability (可靠性) Consistency (一致性) $\text{pass}^k \tau$ -Bench [104]

HELM [51]、WebLinX [63]

WebLinX [63]

CASA [77]、R-Judge [112]、SimuCourt [35]、MATSA [66]、FinCon [111]、AutoGuide [24]

Agent Security Bench (ASB) [118]、AgentPoison [13]、AgentDojo [17]、Backdoor Attacks [102]、SafeAgentBench [108]、Agent-Safety Bench [122]、AgentHarm [5]、Adaptive Attacks [113]、RealToxicityPrompts [27]
合规与隐私

Harm 对抗鲁棒性、提示注入抗性、危害性、偏见检测

Compliance & Privacy (合规性与隐私) Harm (危害) Adversarial Robustness, Prompt Injection Resistance, Harmfulness, Bias Detection
(对抗鲁棒性、提示注入抵御能力、有害性、偏见检测)

R-Judge [112]、Cybench [114]、TheAgentCompany [97]

R-Judge [112]、Cybench [114]、TheAgentCompany [97]

4.1.1 静态与离线评估 离线评估通常作为基准开展，一般依赖数据集与静态测试用例：即由任务、提示或对话构成的集合，用以表征智能体可能面临的挑战。模拟对话可用于辅助构建此类数据，但在不同运行之间其本身是惰性的。离线评估虽然运行与维护成本相对较低且操作简单，但往往缺乏足够的细节，难以全面刻画 LLM agent 可能或被预期给出的各类响应。因此，这类评估更容易产生误差累积，通常也无法准确反映系统的真实性能。

对智能体行为的评估。例如，在 Web 智能体评估中，研究者构建了若干 Web 模拟器 (MiniWoB [85]、WebShop [103]、WebArena [126] 等)，在其中可对智能体的行为 (点击链接、填写表单) 进行编程，以验证其动作序列是否正确。评估驱动开发 (Evaluation-driven Development, EDD) 的概念也被提出 [95]，主张将评估作为智能体开发周期的有机组成部分。该理念倡导对智能体进行持续评估，既包括开发阶段的离线评估，也涵盖部署后的在线评估，从而及时发现性能回退并适应新用例。作者进一步勾勒了一种参考架构，其中 AgentOps 组件负责在生产环境中监控智能体性能，并将洞察反馈给开发者。尽管这一思路仍处于萌芽阶段，但它凸显了一个核心观点：评估并非一次性任务，而是一个持续过程，对会学习或进化的智能体而言尤其如此。

4.2 评估数据

4.1.2 动态与在线评估 与其他机器学习系统类似，在线评估通常在 LLM agent 部署之后进行。它不再依赖合成的、历史性的或人工构造的数据，而是借助模拟环境或真实的用户交互。这类自适应数据对于发现静态测试中难以暴露的痛点与问题至关重要，并且往往蕴含丰富的领域上下文，而这类上下文通过合成或通用基准更难捕捉。动态评估可使用代理来模拟用户或环境，从而以反应式的实时方式进行评估。

对 LLM-based agents 评估日益增长的兴趣，推动了针对第 3 节所讨论的智能体能力而构建的多样化数据集、基准和排行榜的出现。

许多此类资源旨在反映真实世界的复杂性，并由人工标注数据、合成数据和交互生成数据混合构建。例如，AAAR-1.0 [61]、ScienceAgentBench [11] 和 TaskBench [83] 等数据集提供了用于评估研究推理、科学工作流以及多工具规划的结构化、专家标注基准；而 FlowBench [96]、ToolBench [38] 和 API-Bank [47] 等则聚焦于大规模 API 库中的工具使用与函数调用。这些基准不仅包含标准的工具调用序列，还涵盖期望的参数结构，从而支持细粒度的评估。与此同时，AssistantBench [109]、AppWorld [91] 和 WebArena [126] 等数据集在 Web 与应用环境中模拟更开放、更具交互性的智能体行为，强调动态决策、长程规划以及用户与智能体之间的交互。部分基准还支持安全性与鲁棒性测试，例如 AgentHarm [5] 用于评估潜在的有害行为，AgentDojo [17] 用于评估抵御提示注入攻击的能力。Berkeley Function-Calling Leaderboard (BFCL) [100] 和 Holistic Agent Leaderboard [88] 等排行榜则通过提供标准化的测试用例、自动化指标（如 AST 正确率、胜率）以及排序机制，将上述评估加以整合，以支持系统间的横向比较。

4.4 规模，使得需要频繁评估的大规模自动化系统难以实际采用。

在过程维度中，一个值得关注的方面是支持自动化、可扩展且持续化 agent 评估工作流的软件框架与平台的出现。这些工具能够将评估直接集成到开发生命周期中，反映出在 agent 构建中逐步兴起的评估驱动开发（Evaluation-driven Development, EDD）[95] 趋势。OpenAI Evals 是一个开源框架，允许开发者为模型指定评估任务和指标，从而自动执行并报告评估结果（虽然未在学术文献中正式描述，但反映了实际需求）[71]。DeepEval [2]、InspectAI [3]、Phoenix [1] 和 GALILEO [25] 等其他开源或商业工具提供了丰富的分析能力、评估编排与调试功能。此外，Azure AI Foundry [68]、Google Vortex AI [29]、LangGraph [44] 和 Amazon Bedrock [4] 等 agent 开发平台也越来越多地集成评估功能，帮助开发者监控性能、检测回归，并使 agent 适应不断变化的用户需求。Xia 等人 [95] 还进一步提出了一种 AgentOps 架构，用于持续监控已部署的 agent，通过实时反馈和质量控制来打通开发与部署之间的闭环。

4.3 指标计算方法

基于代码的方法是最具确定性和客观性的方法 [8, 53, 57]。[8] 它依赖显式规则、测试用例或断言来验证 agent 的响应是否满足预定义的标准。该方法对于输出定义明确的任务尤其有效，例如数值计算、结构化查询生成或编程任务中的语法正确

性。其主要优势在于一致性和可复现性，使其在基准测试中高度可靠。然而，基于代码的方法通常缺乏灵活性，难以评估开放式或定性的响应，例如自然语言生成或创造性问题求解等正确性具有主观性的任务。尽管如此，它仍然是评估正确性定义明确的结构化任务的基础技术。LLM-as-a-Judge 方法 [125] 利用 LLM 的推理能力，根据定性标准评估 agent 的响应，依据通过指令提供的标准对响应进行评判。由于该方法能够处理具有主观性和细微差异的任务（如摘要、推理和对话交互），因而日益受到关注。该方法的一种近期扩展是 Agent-as-a-Judge [127]，其评估过程涉及多个 AI agent 交互以优化评判，有望提升评估的可靠性。该方法具有高度可扩展性，能够适应复杂任务，因此受到了越来越多的关注 [30, 46]。人在回路（Human-in-the-loop）的评估仍然是主观维度（如自然度和用户满意度）以及安全关键型判断的事实标准。人类评估可采取用户研究、专家审计（由领域专家审查 agent 输出）或众包标注（由标注者沿相关性、正确性和语气等维度对输出进行评分）等形式。该方法在内容生成、战略决策或对话连贯性等开放式任务中具有最高的可靠性。然而，其成本高、耗时长，并且难以大规模开展。

4.5 评估环境

评估环境涉及执行评估所处的环境。类似于软件工程，在更真实（但通常成本更高且安全性可能更低）的环境和更简单、更可控（但通常对最终性能代表性不足）的环境之间存在权衡。系统的评估环境通常由其预期用途所决定：一个不具备编辑权限的较简单 LLM agent 可能直接在真实工作环境中进行测试，而一个旨在与众多相互关联的系统协作并进行修改的 LLM agent 则很可能在模拟 API 或沙箱环境中进行评估。对于耦合度较低的系统，评估环境也可以采取 Web 模拟器的形式，例如 MiniWoB [85]、WebShop [103] 或 WebArena [126]。随着 agent 开发的持续推进，评估环境也会随之演进——从较小的模拟 API 环境逐步过渡到真实部署，这取决于 agent 性能与可信度的确立情况。

5 企业场景特有的挑战

随着 LLM-based agents 从研究演示阶段过渡到企业部署阶段，一系列新挑战正在浮现。企业通常要求高性能，同时还需要可预测的可靠性、法规合规性、数据安全性以及可维护性，而这些方面在评估过程中往往被忽视。为弥补这些差距，我们将讨论以下几个小节中列出的问题，并勾勒未来的研究方向。

5.1 基于角色的访问所带来的复杂性

在企业环境中评估 LLM-based agents 的一项关键挑战在于需要考虑基于角色的访问控制（RBAC），它决定着用户访问数据和服务的权限。在这些环境中，用户依据其角色拥有不同的访问级别，

而代表用户行事的 agent 也必须遵循同样的约束。这为 agent 评估带来了复杂性，因为 agent 检索或操作信息的能力并非一成不变，而是受到用户权限的情境性约束。为应对这一问题，部分评估框架已开始在设计中纳入访问控制约束。例如，IntellAgent[45] 包含了要求对用户身份进行认证的评估任务，并强制实施拒绝其他用户访问信息的策略。通过将角色特定的限制嵌入任务生成过程，这些方法能够更准确地刻画 agent 在权限敏感的企业上下文中的行为。

5.2 可靠性保障

在企业环境中，可靠性保障尤为重要，因为企业期望agent能够在要求确定性或可复现行为且具有可解释性的合规与审计框架内运行。在此类场景下，偶有成功远远不够；agent必须在不同时间和使用场景下保持稳定可靠，方可被视为可投入生产。可靠性的评估并非易事。由于LLM-based agents天然具有随机性，衡量一致性需要对同一任务执行多次并观察结果的变异程度。这会带来显著的评估开销：每个输入运行多轮试验在计算上代价不菲，尤其是在测试涉及工具、记忆或多agent协作的复杂任务时。此外，为得出有意义的结论，基准测试必须包含一个具有代表性的数据集，以反映agent可能遇到的任务类型与运行条件。一些研究已开始应对这一挑战。例如， τ -benchmark[104]显式引入pass@k指标来评估agent的一致性。通过将其应用于零售和机票预订等领域，作者展示了现有agent在一致性方面仍存在明显不足。

5.3 动态与长时程交互

评估LLM-based agents的一大挑战在于衡量其在动态、不断演化的环境中完成长时程任务的能力。与目前多数聚焦于短时片段或单次交互的基准不同，现实中的企业agent往往需要在长时间内持续运行，并与用户、系统和数据不断交互。应对这一挑战对于理解agent随时间推移的行为表现至关重要，特别是在企业环境中，可靠性、适应性和目标对齐贯穿agent的整个生命周期。标准的短期评估无法捕捉诸如性能漂移、上下文保持以及决策对业务结果的累积效应等现象。为着手解决该问题，一些研究工作引入了长期仿真与扩展对话作为评估工具。例如，Park等人[74]在一个持续运行的模拟小镇环境中观察生成式agent，以研究跨多日交互中涌现的行为；类似地，Maharana等人[65]通过600轮对话评估了长期对话记忆，重点关注agent在长对话中维持连贯性与上下文的能力。

5.4 遵循领域特定策略与合规要求

评估企业agent的另一个重大挑战在于确保其能够在领域特定策略与合规约束下运行。企业通常会强制实施严格的操作规范，例如审批流程、数据保留策略、使用配额以及GDPR或HIPAA等法律法规，agent在整个任务执行过程中都必须遵守这些规范。在此类情境下评估agent不仅需要衡量任务是否成功，还要求确认其行为符合形式化的策略约束与法律合规标准。例如，生成财务报告的agent必须避免未经授权访问保密预测数据，并确保所生成的内容遵循监管报告标准。若在评估过程中未对这些约束进行显式建模，即使在传统基准上被判为“正确”的agent仍可能在实际部署中因违反策略或合规风险而失败。

6 未来研究方向

随着基于LLM的智能体在复杂性和应用范围上持续扩展，未来研究必须着力推动更稳健、更实用且更具可扩展性的评估方法论。我们重点强调四个能够显著推动该领域发展的关键方向：

整体性评估框架：当前的评估工作往往聚焦于孤立维度，如任务成功率、规划质量或工具使用能力。然而，在真实应用场景中，智能体必须同时平衡多种能力。未来工作应当构建能够跨多个相互依赖维度评估智能体表现的整体性评估框架。

更真实的评估场景：为弥合实验环境与生产环境之间的差距，智能体评估必须向更贴近真实条件的方向演进。这包括构建融合企业特定要素的评估环境，例如动态多用户交互、基于角色的访问控制以及领域特定知识等。这类场景可通过真实部署试验或模拟企业工作流程的仿真环境来实现。

自动化与可扩展的评估技术：人工评估智能体行为在多层或长时程场景中成本高昂且难以扩展。未来研究应探索自动化评估方法，以减少人力投入并提升可复现性。这包括利用合成数据生成可控测试用例，借助仿真环境模拟任务上下文，以及推进基于 LLM 的评估技术，例如 LLM-as-a-judge 或 Agent-as-a-judge。

时间与成本受限的评估协议：评估必须高效，能够支持智能体的迭代开发。当前的方法——尤其是那些需要重复试验或人在回路评估的方法——往往既耗时又耗费资源。未来研究应致力于构建时间与成本受限的评估协议，在评估深度与效率之间取得平衡。

综上所述，未来研究应聚焦于开发兼具整体性、真实性、可扩展性和高效性的评估方法。这些方向对于大规模构建可靠、可信的基于 LLM 的智能体至关重要。

参考文献

- [23] Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang, Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. 2024. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. doi:10.48550/arXiv.2411.04468
- [24] Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. 2024. AutoGuide: Automated Generation and Selection of Context-Aware Guidelines for Large Language Model Agents. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 119919–119948.
- [25] Galileo. 2025. Introducing agentic evaluations. <https://www.galileo.ai/blog/introducing-agentic-evaluations> Accessed: 2025-05-20.
- [26] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, et al. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. arXiv preprint arXiv:2411.09523 (2024).
- [27] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv:2009.11462 [cs.CL] <https://arxiv.org/abs/2009.11462>
- [28] Luca Gioacchini, Giuseppe Siracusano, Davide Sanvito, Kiril Gashteovski, David Friede, Roberto Bifulco, and Carolin Lawrence. 2024. AgentQuest: A Modular Benchmark Framework to Measure Progress and Improve LLM Agents. doi:10.

48550/arXiv.2404.06411 [29] Google. 2024. Google Vortex AI. <https://cloud.google.com/vertex-ai> [30] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594 (2024). [31] Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and YIzhou Wang. 2024. Riche- lieu: Self-Evolving LLM-Based Agents for AI Diplomacy. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 38. Curran Associates, Inc., 0. [32] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680 (2024). [33] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. LLM multi-agent systems: Challenges and open problems. arXiv preprint arXiv:2402.03578 (2024). [34] Feng He, Tianqing Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S Yu. 2024. The emerged security and privacy of llm agent: A survey with case studies. arXiv preprint arXiv:2407.19354 (2024). [35] Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. AgentsCourt: Build- ing Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation. arXiv:2403.02959v3 (2024). <https://arxiv.org/pdf/2403.02959> [36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] <https://arxiv.org/abs/2009.03300> [37] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. doi:10.48550/arXiv.2402.02716 [38] Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2024. MetaTool Benchmark for Large Language Models: Deciding Whether to Use Tools and Which to Use. arXiv:2310.03128 [cs.SE] <https://arxiv.org/abs/2310.03128> [39] Zixia Jia, Mengmeng Wang, Baichen Tong, Song-Chun Zhu, and Zilong Zheng. 2024. LangSuitE: Planning, Controlling and Interacting with Large Language Models in Embodied Text Environments. arXiv preprint arXiv:2406.16294 (2024). [40] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? doi:10.48550/arXiv.2310.06770 [41] H Kang and C Xiong. 2024. ResearchArena: Benchmarking LLMs’ Ability to

Collect and Organize Information as Research Agents. (2024). [42] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. VisualWebArena: evaluating multimodal agents on realistic visual web tasks. doi:10.48550/arXiv.2401.13649 [43] Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. arXiv:2301.13298 [cs.CL] <https://arxiv.org/abs/2301.13298> [44] LangChain. 2024. LangGraph Platform. <https://www.langchain.com/langgraph-platform> [45] Elad Levi and Ilan Kadar. 2025. IntellAgent: a multi-agent framework for evaluating conversational AI systems. doi:10.48550/arXiv.2501.11067

[1] Arize AI. 2025. Phoenix. <https://github.com/Arize-ai/phoenix> [2] Confident AI. 2025. DeepEval. <https://github.com/confident-ai/deepeval> [3] UK AI Security Institute. 2024. Inspect AI: Framework for Large Language Model Evaluations. https://github.com/UKGovernmentBEIS/inspect_ai [4] Amazon. 2024. Amazon Bedrock Agents. <https://aws.amazon.com/bedrock/agents/> [5] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Due-nas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. 2025. AgentHarm: a benchmark for measuring harmfulness of LLM agents. doi:10.48550/arXiv.2410.09024 [6] Nik Bear Brown. 2024. Enhancing trust in llms: Algorithms for comparing and interpreting llms. arXiv preprint arXiv:2406.01943 (2024). [7] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023). [8] Ma Chang, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhen-zhong Lan, Lingpeng Kong, and Junxian He. 2024. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. *Advances in Neural Information Processing Systems* 37 (2024), 74325–74362. [9] Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. SocialBench: Sociality Evaluation of Role-Playing Conversational Agents. arXiv:2403.13679 [cs.CL] <https://arxiv.org/abs/2403.13679> [10] Jingxuan Chen, Derek Yuen, Bin Xie, Yuhao Yang, Gongwei Chen, Zhihao Wu, Li Yixing, Xurui Zhou, Weiwen Liu, Shuai Wang, et al. 2024. Spa-bench: A comprehensive benchmark for smartphone agent evaluation. In *NeurIPS 2024 Workshop on Open-World Agents*. [11] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi

Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2025. ScienceAgentBench: toward rigorous assessment of language agents for data-driven scientific discovery. doi:10.48550/ arXiv.2410.05080 [12] Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, and Feng Zhao. 2024. T-eval: evaluating the tool utilization capability of large language models step by step. doi:10.48550/arXiv.2312.14033 [13] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. Agent-Poison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases. arXiv:2407.12784 [cs.LG] <https://arxiv.org/abs/2407.12784> [14] Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omid Shayeagan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. 2025. The BrowserGym ecosystem for web agent research. doi:10.48550/arXiv.2412.05467 [15] Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Har- iharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. Gamebench: Evaluating strategic reasoning abilities of llm agents. arXiv preprint arXiv:2406.06613 (2024). [16] José Antonio Siqueira de Cerqueira, Mamia Agbese, Rebekah Rousi, Nannan Xi, Juho Hamari, and Pekka Abrahamsson. 2024. Can we trust AI agents? An experimental study towards trustworthy LLM-based multi-agent systems for AI ethics. arXiv preprint arXiv:2411.08881 (2024). [17] Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. doi:10.48550/ arXiv.2406.13352 [18] Shihan Deng, Weikai Xu, Hongda Sun, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, Bin Wang, Rui Yan, et al. 2024. Mobile-bench: An evaluation benchmark for llm-based mobile agents. arXiv preprint arXiv:2407.00993 (2024). [19] Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. 2024. Re-rest: Reflection-reinforced self-training for language agents. arXiv preprint arXiv:2406.01495 (2024). [20] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. arXiv preprint arXiv:2401.03568 (2024). [21] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2025. Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217 [cs.CL]

<https://arxiv.org/abs/2309.15217> [22] Richard Fang, Rohan Bindu, Akul Gupta, and Daniel Kang. 2024. Llm agents can autonomously exploit one-day vulnerabilities. arXiv preprint arXiv:2404.08144 13 (2024), 14.

[46] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (2024).

[47] Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. API-bank: a comprehensive benchmark for tool-augmented LLMs. doi:10.48550/arXiv.2304.08244 [48] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 1, 1 (2024), 9.

[49] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. arXiv preprint arXiv:2401.05459 (2024).

[50] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. 2024. Optimus-1: Hybrid Multimodal Memory Empowered Agents Excel in Long-Horizon Tasks. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 49881–49913. [51] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. doi:10.48550/arXiv.2211.09110

[52] Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. arXiv preprint arXiv:2308.04026 (2023).

[53] Bing Liu, Zhou Jianxiang, Dan Meng, and Haonan Lu. 2024. An Evaluation Mechanism of LLM-based Agents on Manipulating APIs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida,

USA, 4649–4662. doi:10.18653/v1/2024.findings-emnlp.267 [54] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, Yuheng Cheng, Suyuchen Wang, Xiaoqiang Wang, Yuyu Luo, Haibo Jin, Peiyan Zhang, Ollie Liu, Jiaqi Chen, Huan Zhang, Zhaoyang Yu, Haochen Shi, Boyan Li, Dekun Wu, Fengwei Teng, Xiaojun Jia, Jiawei Xu, Jinyu Xiang, Yizhang Lin, Tianming Liu, Tongliang Liu, Yu Su, Huan Sun, Glen Berseth, Jianyun Nie, Ian Foster, Logan Ward, Qingyun Wu, Yu Gu, Mingchen Zhuge, Xiangru Tang, Haohan Wang, Jiaxuan You, Chi Wang, Jian Pei, Qiang Yang, Xiaoliang Qi, and Chenglin Wu. 2025. Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems. doi:10.48550/arXiv.2504.01990 [55] Wei Liu, Chenxi Wang, Yifei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. 2024. Autonomous agents for collaborative task under information asymmetry. arXiv preprint arXiv:2406.14928 (2024). [56] Wei Liu, Chenxi Wang, YiFei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. 2024. Autonomous Agents for Collaborative Task under Information Asymmetry. In The Thirty- eighth Annual Conference on Neural Information Processing Systems. [57] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. AgentBench: Evaluating LLMs as Agents. doi:10.48550/arXiv.2308.03688 [58] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. AgentBench: Evaluating LLMs as Agents. arXiv:2308.03688 [cs.AI] <https://arxiv.org/abs/2308.03688> [59] Yanjiang Liu, Tianyun Zhong, Yaojie Lu, Hongyu Lin, Ben He, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Zhongyi Liu, Xianpei Han, et al. 2024. XMC-Agent: Dynamic Navigation over Scalable Hierarchical Index for Incremental Extreme Multi-label Classification. In Findings of the Association for Computational Linguistics ACL 2024. 5659–5672. [60] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2023. Chatting with gpt-3 for zero-shot human-like mobile automated gui testing. arXiv preprint arXiv:2305.09434 (2023). [61] Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun

Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. 2025. AAAR-1.0: Assessing AI’ s Potential to Assist Research. doi:10.48550/

arXiv.2410.22394 [62] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292 (2024). [63] Xing Han Lù, Zdeněk Kasner, and Siva Reddy. 2024. WebLINX: Real-World Web- site Navigation with Multi-Turn Dialogue. (Feb. 2024). arXiv:2402.05930 [cs.CL] [64] Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhen- zhong Lan, Lingpeng Kong, and Junxian He. 2024. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. doi:10.48550/arXiv.2401.13178 [65] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. arXiv:2402.17753 [cs.CL] <https://arxiv.org/abs/2402.17753> [66] Puneet Mathur, Alexa Siu, Nedim Lipka, and Tong Sun. 2024. MATSA: Multi- Agent Table Structure Attribution. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 250– 258. [67] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. 2024. Integrity-based explanations for fostering appropriate trust in AI agents. ACM Transactions on Interactive Intelligent Systems 14, 1 (2024), 1–36. [68] Microsoft. 2024. Azure Foundry. <https://azure.microsoft.com/en-us/products/ai-foundry> [69] Yohei Nakajima. 2023. Babyagi. GitHub repository (2023). [70] Nvidia. 2024. LLM Benchmark Metrics. <https://docs.nvidia.com/nim/benchmarking/llm/latest/metrics.htm> [71] OpenAI. 2023. OpenAI Evals. <https://github.com/openai/evals> [72] Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. 2024. BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games. arXiv:2411.13543 [cs.AI] <https://arxiv.org/abs/2411.13543> [73] Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, and Zhengyang Wu. 2024. WebCanvas: benchmarking web agents in online environments. doi:10.48550/ arXiv.2406.12373 [74] Joon Sung Park, Joseph O’ Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology. 1–22. [75] Shishir G. Patil, Tianjun

Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large Language Model Connected with Massive APIs. doi:10.48550/arXiv.2305.15334 arXiv:2305.15334 [cs].

[76] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. ToolLLM: facilitating large language models to master 16000+ real-world APIs. doi:10.48550/arXiv.2307.16789

[77] Haoyi Qiu, Alexander R. Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating Cultural and Social Awareness of LLM Web Agents. arXiv:2410.23252 (2025).

[78] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822 (2018).

[79] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250 [cs.CL]

[80] Merle M. Reimann, Catharine Oertel, Florian A. Kunneman, and Koen V. Hindriks. 2023. Predicting interaction quality aspects using level-based scores for conversational agents. In Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. ACM, Würzburg Germany, 1–8. doi:10.1145/3570945.3607332

[81] Wenjie Ruan, Wei Huang, Xiaowei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. 2024. ToolEmu. In Proceedings of t.

[82] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. 2023. Identifying the risks of lm agents with an lm-emulated sandbox. arXiv preprint arXiv:2309.15817 (2023).

[83] Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. 2024. TaskBench: Benchmarking Large Language Models for Task Automation. doi:10.48550/arXiv.2311.18760

[84] Honghao Shi, Longkai Cheng, Wenli Wu, Yuhang Wang, Xuan Liu, Shaokai Nie, Weixv Wang, Xuebin Min, Chunlei Men, and Yonghua Lin. 2024. Enhancing Cluster Resilience: LLM-agent Based Autonomous Intelligent Cluster Diagnosis System and Evaluation Framework. arXiv:2411.05349 (2024). <https://arxiv.org/pdf/2411.05349>

[85] Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. World of bits: An open-domain platform for web-based agents. In International Conference on Machine Learning. PMLR, 3135–3144.

[86] Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. 2024. CORE-bench: fostering the credibility of published research through a computational reproducibility agent

benchmark. doi:10.48550/arXiv.2409.11363 [87] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. PaperBench: Evaluating AI's Ability to Replicate AI Research. doi:10.48550/arXiv.2504.01848 [88] Benedikt Stroebel, Sayash Kapoor, and Arvind Narayanan. 2025. HAL: A Holistic Agent Leaderboard for Centralized and Reproducible Agent Evaluation. <https://github.com/princeton-pli/hal-harness>. [89] Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. arXiv preprint arXiv:2306.03314 (2023). [90] Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems* 37 (2024), 51963– 51993. [91] Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjana Balasubramanian. 2024. AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents. doi:10.48550/arXiv.2407.18901 [92] Luyuan Wang, Yongyu Deng, Yiwei Zha, Guodong Mao, Qinmin Wang, Tianchen Min, Wei Chen, and Shoufa Chen. 2024. MobileAgentBench: An Efficient and User-Friendly Benchmark for Mobile LLM Agents. arXiv preprint arXiv:2406.08184 (2024). [93] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science* 18, 6 (Dec. 2024), 186345. doi:10.1007/s11704-024-40231-1 [94] Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, et al. 2024. Gui agents with foundation models: A comprehensive survey. arXiv preprint arXiv:2411.04890 (2024). [95] Boming Xia, Qinghua Lu, Liming Zhu, Zhenchang Xing, Dehai Zhao, and Hao Zhang. 2024. An Evaluation-Driven Approach to Designing LLM Agents: Process and Architecture. arXiv preprint arXiv:2411.13768 (2024). [96] Ruixuan Xiao, Wentao Ma, Ke Wang, Yuchuan Wu, Junbo Zhao, Haobo Wang, Fei Huang, and Yongbin Li. 2024. FlowBench: revisiting and benchmarking workflow-guided planning for LLM-based agents. doi:10.48550/arXiv.2406.14884 [97] Frank F Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. 2024. Theagent- company: benchmarking llm agents on consequential real world tasks. arXiv preprint arXiv:2412.14161 (2024). [98] Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nischal Chandra, and Nanyun Peng.

2022. EnDex: Evaluation of Dialogue Engagingness at Scale. 4884–4893. doi:10.18653/v1/2022.findings-emnlp.359 [99] Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See Kiong Ng, and Jiashi Feng. 2023. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. arXiv preprint arXiv:2311.08562 (2023). [100] Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Berkeley Function Calling Leaderboard. https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html. [101] Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. PsychoGAT: A Novel Psychological Measurement Paradigm through Interactive Fiction Games with LLM Agents. arXiv:2402.12326v2 (2024). <https://arxiv.org/pdf/2402.12326> [102] Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. 2024. Watch Out for Your Agents! Investigating Backdoor Threats to LLM-Based Agents. arXiv:2402.11208 [cs.CR] <https://arxiv.org/abs/2402.11208> [103] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems* 35 (2022), 20744–20757. [104] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. doi:10.48550/arXiv.2406.12045 [105] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2210.03629v3> [106] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. doi:10.48550/arXiv.2210.03629 [107] Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on Evaluation of LLM-based Agents. doi:10.48550/arXiv.2503.16416 [108] Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. 2024. SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents. arXiv preprint arXiv:2412.13178 (2024). [109] Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. 2024. AssistantBench: can web agents solve realistic and time-consuming tasks? doi:10.48550/arXiv.2407.15711 [110] Erxin Yu, Jing

Li, Ming Liao, Siqi Wang, Zuchen Gao, Fei Mi, and Lanqing Hong. 2024. Cosafe: Evaluating large language model safety in multi-turn dialogue coreference. arXiv preprint arXiv:2406.17626 (2024). [111] Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur (Suba) Subbalakshmi, GUOJUN XIONG, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 137010–137045. [112] Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. 2024. R-judge: Benchmarking safety risk awareness for llm agents. arXiv preprint arXiv:2401.10019 (2024). [113] Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. 2025. Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents. arXiv:2503.00061 [cs.CR] <https://arxiv.org/abs/2503.00061> [114] Andy K. Zhang, Neil Perry, and Riya Dulepet et al. 2024. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models. arXiv:2408.08926v3 (2024). <https://arxiv.org/pdf/2408.08926> [115] Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2025. Large Language Model-Brained GUI Agents: A Survey. doi:10.48550/arXiv.2411.18279 [116] Danyang Zhang, Lu Chen, and Kai Yu. 2023. Mobile-env: A universal platform for training and evaluation of mobile interaction. CoRR (2023). [117] Danyang Zhang, Zhennan Shen, Rui Xie, Situo Zhang, Tianbao Xie, Zihan Zhao, Siyuan Chen, Lu Chen, Hongshen Xu, Ruisheng Cao, and Kai Yu. 2024. Mobile-Env: Building Qualified Evaluation Benchmarks for LLM-GUI Interaction. arXiv:2305.08144 [cs.AI] <https://arxiv.org/abs/2305.08144> [118] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. 2024. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. arXiv preprint arXiv:2410.02644 (2024). [119] Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. 2024. AutoCodeRover: Autonomous Program Improvement (ISSTA 2024). Association for Computing Machinery, New York, NY, USA, 1592–1604. doi:10.1145/3650212.3680384 [120] Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. LLM Eval: A Preliminary Study on How to Evaluate Large Language

Models. arXiv:2312.07398 [cs.AI] <https://arxiv.org/abs/2312.07398> [121] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A Survey on the Memory Mechanism of Large Language Model based Agents. doi:10.48550/arXiv.2404.13501 [122] Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. 2024. Agent-SafetyBench: Evaluating the Safety of LLM Agents. arXiv preprint arXiv:2412.14470 (2024). [123] Zhiping Zhang, Michelle Jia, B Yao, S Das, A Lerner, D Wang, and T Li. 2023. It’s a fair game, or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. arXiv preprint arXiv:2309.11653 (2023). [124] Ziniu Zhang, Shulin Tian, Liangyu Chen, and Ziwei Liu. 2024. MMInA: benchmarking multihop multimodal internet agents. doi:10.48550/arXiv.2404.09992 [125] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623. [126] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023. Webarena: A realistic web environment for building autonomous agents. arXiv preprint arXiv:2307.13854 (2023). [127] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. 2024. Agent-as-a-judge: Evaluate agents with agents. arXiv preprint arXiv:2410.10934 (2024).