AI 技术每日分析

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 10 月 5 日

摘要

把 Agent 交给企业业务，不能只看它是否答得像样。微软与 Hugging Face 公布 ThinkingBox：507 个合成企业工作流在可重置沙箱中运行，以数据库最终状态和副作用判定结果，还建议重复运行 20 次看稳定性。MLCommons 的 Agent Privacy Risk Taxonomy v0.1 整理出 25 类隐私风险，覆盖数据摄取、聚合共享、跨 Agent 协作、动态同意和责任治理，下一阶段计划在 2027 年一季度前推进可测量的 Benchmark。10 月 4 日刊发的一篇 ServiceNow 实践复盘，把许可证利用、CMDB 质量和目录项维护列为更实际的 Agent 切入口；DeepSeek Harness v0.2 预览版把插件管理、文件与 Diff 审阅和自动化任务带向桌面工作环境。论文《Agents Are Systems, Not Models》则提醒人们：同样的 Agent 配置，重复运行仍可能得到相差很大的结果。验收 Agent 时，模型之外的状态、权限、工具和运行环境都得算进去。

Contents

一、ThinkingBox 把 Agent 评测从“回答正确”推进到“后台状态正确” 3

二、MLCommons 梳理 25 类 Agent 隐私风险：隐私治理开始走向统一“风险字典” 3

三、ServiceNow 实战复盘：企业 Agent 的高价值入口往往比路线图想象得更窄 4

四、DeepSeek Harness v0.2 走向桌面化：Agent 框架竞争从 CLI 扩展到日常工作入口 5

五、“Agents Are Systems, Not Models”：同一个 Agent 配置重复运行，也可能出现巨大差异 5

趋势观察 6

参考资料 6

一、ThinkingBox 把 Agent 评测从“回答正确”推进到“后台状态正确”

微软 Copilot Studio 团队与 Hugging Face 于 10 月 3 日联合发布 ThinkingBox 的 OpenEnv 集成，相关资料在 10 月 4 日进一步传播。ThinkingBox 提供 507 个合成企业工作流，涉及零售、汽车保险、旅行、数字银行和咨询。任务中预设初始后台状态、用户目标、策略约束、MCP 工具，以及能执行的最终状态检查。

一个 Agent 可能工具调用顺畅、回复也像完成了任务，却把数据库字段写错，漏做必要动作，或额外改动了其他内容。ThinkingBox 将执行后的数据库与预设“黄金状态”比较，直接检查业务记录到底有没有按要求改变。

单次通过也不能说明流程稳定。官方建议在同一任务上重复运行；公开分析中，一些模型的单次成功率不低，连续 20 次全部成功的任务比例却明显下降。采购方要问的不是“这次演示成了吗”，而是相同流程能否反复做对。

客服、财务、IT 运维和销售 Agent 的验收，可以据此检查最终记录、审批状态、外部副作用与重复运行结果。自然语言演示仍有用，但不能代替后台状态检查。

二、MLCommons 梳理 25 类 Agent 隐私风险：隐私治理开始走向统一“风险字典”

MLCommons AI Risk & Reliability 项目发布 Agent Privacy Risk Taxonomy v0.1，10 月 4 日再次受到行业讨论。它将 25 类风险归到五个领域：数据摄取与处理、数据聚合/使用/共享、跨 Agent 隐私实践不一致、静态同意模型失效，以及责任与治理。

聊天机器人通常处理眼前对话，Agent 还可能长期观察环境、跨会话保存记忆、调用第三方工具、转交上下文，甚至动态生成代码。单独看每一步都合规，串起来仍可能让敏感信息流向不该去的地方。

一次授权不能解决所有后续场景。Agent 运行时可能遇到新联系人、新资料、新的下游 Agent；用户撤回同意后，已转交的数据该如何处理，也需要相应的运行时规则。MLCommons 特别把这种动态同意问题列了出来。

v0.1 还是草案，并非强制标准。MLCommons 计划与大规模部署方确定优先风险，再制定指标、试点 Benchmark 和缓解实践，目标在 2027 年一季度前完成下一阶段工作。企业可以先按这份分类检查数据最小化、跨 Agent 传递、运行时授权和审计日志，而不只看隐私政策写得是否周全。

三、ServiceNow 实战复盘：企业 Agent 的高价值入口往往比路线图想象得更窄

VentureBeat 在 10 月 4 日刊发 Bytemethod.ai 团队的 ServiceNow Agent 复盘。团队从 ITSM 日常运营找需求，没有先做覆盖整个系统的“万能 IT Agent”：服务目录维护、许可证利用分析、CMDB 健康度和重复记录操作，更耗人力，结果也更容易核算。

这份复盘把选场景的问题说得很具体。功能清单可以列得很长，上线后的收益却常集中在高频、规则清楚、已有数据且能核对结果的少数流程。

团队起初用 Browser Use 操作 Web 界面，后来逐步转向结构化系统连接。浏览器适合迅速验证想法，界面改版、加载时间和隐藏状态却会影响长期运行；API、结构化工具、权限边界与后台状态检查更适合作为持续运行的基础。

这也解释了为什么 ThinkingBox 要核对最终状态。企业优先做哪些

Agent，不妨从能计算节省工时、减少许可证浪费、改善数据质量或缩短处理周期的流程开始，而非先追求覆盖面。

四、DeepSeek Harness v0.2 走向桌面化：Agent 框架竞争从 CLI 扩展到日常工作入口

10 月 4 日前后，多家开发者媒体跟进 DeepSeek Harness v0.2 预览版。它是开源 Agent Harness，可用插件组合模型适配器、工具注册、技能、会话、沙箱、存储、执行循环、调度和界面。

预览版公开资料提到桌面使用形态、插件管理、文件与 Diff 审阅、文档任务和自动化能力。Agent 若不必一直待在终端里，处理知识工作中的多文件任务会更方便。

不过，第三方报道对“官方桌面安装器”的具体版本和平台支持说法不一。现阶段可以确认的是 Harness 的插件化运行时正在向桌面工作环境扩展，不能把社区包装器直接称作官方产品。

团队若要自建工作台，需要一起评估插件生态、沙箱、安全边界、任务调度、文件审阅和跨模型兼容性。模型能否接入只是其中一项，Harness 本身也要单独选型。

五、“Agents Are Systems, Not Models”：同一个 Agent 配置重复运行，也可能出现巨大差异

论文《Agents Are Systems, Not Models: Rethinking Agentic Evaluation》于 10 月 1 日提交，10 月 3 日前后仍在更新和传播。作者主张把 Agent 当作可配置系统评测，而非只记下所用模型与 Prompt。

研究分析四个科学任务中的任务信息、推理配置、自我验证、时间预算和底座模型等因素，发布超过 1.8 万条 Agent 轨迹。作者报告，约 54% 的结果方差来自同一配置的重新运行；任务信息质量的影响甚至超过模

型大小与时间预算。一次跑通，不能代表稳定跑通。

Prompt 里写一句“检查答案”，对实际验证行为的影响有限。系统提供专门的验证工具后，Agent 才更明显地开展验证。想让它核对结果，就得让核对有可执行的入口。

因此企业评测也应记录验证工具、数据约束、重试策略、超时、预算和运行环境。它们会改变结果，不能只在评测表里填一个模型名称。

趋势观察

今天的五条线索指向同一个验收问题：任务最终改对了吗，而且能否持续做对？ThinkingBox 检查数据库状态，MLCommons 整理隐私风险，ServiceNow 实践把投入收束到可核算流程，DeepSeek Harness 扩展工作入口，系统级评测则显示重复运行本身有不小的方差。企业要给 Agent 设定可检查的结果、权限和运行条件，成本也要算得清楚。

参考资料

1. **Microsoft × Hugging Face** | 《The Agent Said It Was Done. The Database Disagreed.》 | 2026-10-03 | 用途：核验 ThinkingBox、507 个工作流、最终状态评测与重复运行方法。
   `https://huggingface.co/blog/microsoft/thinkingbox`

2. **Microsoft / Hugging Face ThinkingBox-Bench** | ThinkingBox-Bench v1.0 | 2026-10 访问 | 用途：核验黄金数据库状态、执行检查和 Benchmark 许可。
   `https://huggingface.co/datasets/microsoft/ThinkingBox-Bench`

3. **AccessAllGPT Research** | 《Microsoft and Hugging Face Bring

ThinkingBox’s 507 Stateful Agent Workflows to OpenEnv》｜2026-10-04 ｜用途：交叉核验 OpenEnv 集成、20 次重复运行和公开结果。
https://www.accessallgpt.com/research/microsoft-hugging-face-thinkingbox-stateful-agent-benchmark-openenv-release

4. **MLCommons** ｜《What Could Go Wrong When Agents Handle Your Sensitive Data? A New Taxonomy Answers》｜2026-10-01 ｜用途：核验五类隐私风险、v0.1 状态和 2027 年工作计划。
https://mlcommons.org/2026/10/agent-privacy-risk-taxonomy-v0-1/

5. **PPC Land** ｜《MLCommons catalogues 25 privacy risks posed by AI agents》｜2026-10-04 ｜用途：补充 25 个风险向量和六阶段事件链。
https://ppc.land/mlcommons-catalogues-25-privacy-risks-posed-by-ai-agents/

6. **VentureBeat** ｜《We built an AI agent for ServiceNow. The real pain points were narrower than our roadmap assumed》｜2026-10-04 ｜用途：企业 ITSM Agent 实战复盘。
https://venturebeat.com/orchestration/we-built-an-ai-agent-for-servicenow-the-real-pain-points-were-narrower-than-our-roadmap-assumed

7. **MarkTechPost** ｜《DeepSeek Harness v0.2 Brings Official Desktop Apps to Its Open-Source Agent Harness》｜2026-10-03 ｜用途：补充 Harness v0.2 桌面化、插件管理与文件审阅能力。
https://www.marktechpost.com/2026/10/03/deepseek-harness-v0-2-brings-official-desktop-apps-to-its-open-source-agent

-harness/

8. **iTechGuides** | 《DeepSeek Harness and Desktop Apps: What's Official, and What v0.2 Means》 | 2026-10-04 | 用途：校验官方 Harness 与社区桌面包装器之间的边界，避免夸大。
   https://www.itechguides.com/deepseek-harness-and-desktop-apps-whats-official-and-what-v0-2-means/

9. **arXiv** | 《Agents Are Systems, Not Models: Rethinking Agentic Evaluation》 | 2026-10-01 | 用途：核验 54% 运行方差、配置因素及 1.8 万余条 Agent 轨迹。
   https://arxiv.org/abs/2610.01618

10. **Fervor AI** | 2026-10-04 Agent Briefing | 2026-10-04 | 用途：背景资料，观察 Agent 社区从“能做什么”转向预算、约束和工作状态治理。
    https://fervorai.dev/

联系我们，请扫描二维码

新质生产力工作委员会

官方公众号

工业智能算网

gyznsw.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址： https://gyznsw.cn