Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

智能体AI驾驭层：

模型之上的层级

首次出版：2026年9月

采编说明			
原文题目	Agentic AI Harnesses: The layer above the model		
中译题目	智能体 AI 驾驭层：模型之上的层级		
作者/机构	澳大利亚信号情报局（ASD）	**原文发布日期**	2026年9月11日
原文出处	https://www.cyber.gov.au/sites/default/files/2026-09/Agentic%20AI%20Harnesses%20-%20The%20layer%20above%20the%20model.pdf		
作者简况			
机构简况	澳大利亚信号情报局（ASD）隶属于澳大利亚国防部，是澳大利亚主要的信号情报与网络安全机构。该机构核心业务包括搜集、分析境外电子通信信号，为国家决策提供情报支撑；统筹国家网络防御工作，为政府和各类机构提供网络安全防护支持；经授权开展网络作战行动，维护国家安全利益。同时，ASD 是“五眼联盟”重要组成部分，长期与其他成员国开展情报信息共享，协同完成情报搜集等相关工作，在澳大利亚国防、对外安全事务当中发挥着关键作用。		
采编方	安天战略情报中心信息采编组	**责编**	卢开思
翻译方式	机翻精校	**机翻平台**	龙译（安天内部平台）
免责声明	本译文为安天战略情报中心信息采编组采集发现，并进行机翻校对。 本文原文来自互联网的公共信源，基于AI平台进行翻译，并进行部分校对，力图符合所获得之电子版本原意进行翻译，但受翻译水平和技术水平所限，不能完全保证译文完全与原文含义一致，同时对所获得原文是否存在臆造、或者是否与其原始版本一致未进行可靠性验证和评价。 本译文对应原文所有观点亦不受本译文中任何打字、排版、印刷或翻译错误的影响。译者与安天不对译文及原文中包含或引用的信息的真实性、准确性、可靠性、或完整性提供任何明示或暗示的保证。译者与安天亦对原文和译文的任何内容不承担任何责任。 安天战略情报中心基于“国家安全是我们的唯一立场和第一视角”的安天企业原则展开工作，采编和翻译本文服务于支撑国家相关工作和企业发展工作需要。原文必然带有相关机构和编写者本身的立场，安天编译本文不代表认同原文表述内容、持有立场和态度。 译者与安天均与原作者与原始发布者没有联系，亦未获得相关的版权授权，鉴于译者及安天出于学习参考之目的翻译本文，而无出版、发售译文等任何商业利益意图，因此亦不对任何可能因此导致的版权问题承担责任。		

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

目录

目的

本出版物阐述了“智能体AI驾驭层”（以下简称“Harness”）的概念，即使大语言模型（LLM）能够与数据、工具和系统交互以执行智能体任务的软件层。本文探讨了Harness设计如何影响智能体AI系统的安全性、治理、可靠性及运营风险。

ASD的《审慎采用智能体AI服务》为智能体AI的安全考量提供了更广泛的指导，包括扩大的攻击面、系统复杂性以及不断演变的安全形势。该指南还概述了权限风险、设计与配置风险、行为风险、结构风险和问责风险等主要风险类别，并为设计、开发、部署和运营安全的智能体AI系统提供了最佳实践指南。本出版物通过聚焦“Harness”（作为组织能够最直接进行治理、保障和控制的组件），对上述指导进行了补充。

受众

本出版物面向正在采用或考虑采用智能体AI的组织，主要面向决策层管理者、首席信息安全官和IT负责人。

执行摘要

“Harness”是一种软件层，它使大语言模型（LLM）能够在现实环境中运行。如果将LLM比作大脑，那么Harness就是身体。LLM以上下文的形式接收信息，对其进行推理，并提出行动方案以获取更多信息或完成任务。“Harness”提供信息（如上下文和记忆），执行操作以访问数据或使用工具，并能对相关数据和工具实施权限管理与控制。这些要素共同构成了一个具有智能体能力的AI系统。

理解这一区别至关重要，因为虽然部分风险源于LLM本身的行为，但许多影响最大的风险往往是在模型通过接口与企业数据、系统和工具连接时产生的。安全、隐私、治理和运营风险通常取决于智能体能够访问哪些内容、执行哪些操作，以及如何与组织系统进行交互。某些风险（包括提示注入）无法仅通过模型本身可靠地解决，而需要在Harness、相关系统以及组织治理流程中实施额外的控制措施。Harness及其相关的集成方案、治理机制和运营流程，很可能代表着组织的一项重大长期投资。虽然LLM将持续演进，并可能随着时间的推移被替换，但Harness生态系统很可能成为组织更具持久性的能力。组织应将该Harness视为其AI系统的关键组成部分，并实施与其带来的风险相称的安全和治理控制措施。

没有任何Harness天生就是安全的。组织应采用成熟的网络安全实践，包括最小权限原则、身份与访问管理、安全设计、监控、人工监督以及与事件响应流程的整合。引入智能体AI应采取分阶段的方法，明确批准的使用场景、对数据进行分类、选择合适的Harness、应用ASD《审慎采用智能体AI服务》中所述的控制措施，并在大规模部署前验证安全控制措施的有效性。

成功的采用需要将模型和Harness与任务相匹配，仅授予执行该任务所需的访问权限，在系统的多层级应用控制措施，并确保对决策、行动和结果保持明确的人员问责制。

关键要点

- 组织控制的是“Harness”，而非LLM
- “Harness”是组织长期价值、治理和投资的不断累积之处。
- “Harness”及其周边的技术生态系统是安全与风险管理的主要关注点

什么是“Harness”

一个智能体AI系统由LLM以及外部工具、外部数据源、记忆和规划工作流组成。这些组件共同使系统能够感知其环境，并在适当的情况下采取行动以实现其目标。

本出版物使用“Harness”一词来描述除LLM本身之外的智能体AI系统的组成部分。Harness是将大语言模型与外部工具、数据源、记忆及规划工作流相连接的软件层。它还负责执行系统的操作和执行权限，并运行控制循环，该循环将持续重复直至任务完成。在《审慎采用智能体AI服务》一文中，LLM是指用于确定应采取何种行动的统计模型。Harness为智能体提供信息输入，授予其工具或服务访问权限，执行其行动和执行权限，管理其记忆和规划工作流，并记录其评估指标。

将Harness明确界定为智能体AI系统的独立组成部分，使组织能够将其与LLM分开进行评估和管理。这一点至关重要，因为许多决定智能体AI系统可靠性、运营成本和安全风险的决策是在Harness中做出的，而非在LLM中。在某些情况下，Harness由组织自行开发；在其他情况下，它作为商业产品或服务的一部分提供。无论其来源如何，组织都应了解Harness的能力、权限、集成点和安全控制措施，并评估其使用相关的风险。

下图总结了LLM与Harness之间的主要区别。

维度	LLM	Harness
核心任务	预测后续单词或文本；基于当前文本进行推理。	决定将哪些文本输入模型，以及如何处理输出结果。
状态	无状态——每次交互均从零开始，不保留记忆或先前交互记录。	有状态——在不同轮次和会话之间持久化上下文、记忆和进度。
行动	仅生成令牌。	执行工具、编辑文件、运行命令、调用API、浏览。
记忆	仅考虑当前上下文窗口。	长期记忆、会话日志、临时存储区、文件。
控制	生成输出，但不直接控制任何由此产生的操作。	权限、沙箱、审批闸门、钩子、策略。
一致性	相同的输入可能产生不同的输出。	添加验证、结构化输出、重试和错误恢复。
故障模式	产生幻觉，给出错误答案。	工具错误、记忆丢失、不安全操作、上下文溢出。
可变更性	一个可插拔的组件，可以进行替换。	一种比任何单一模型都更持久的架构。

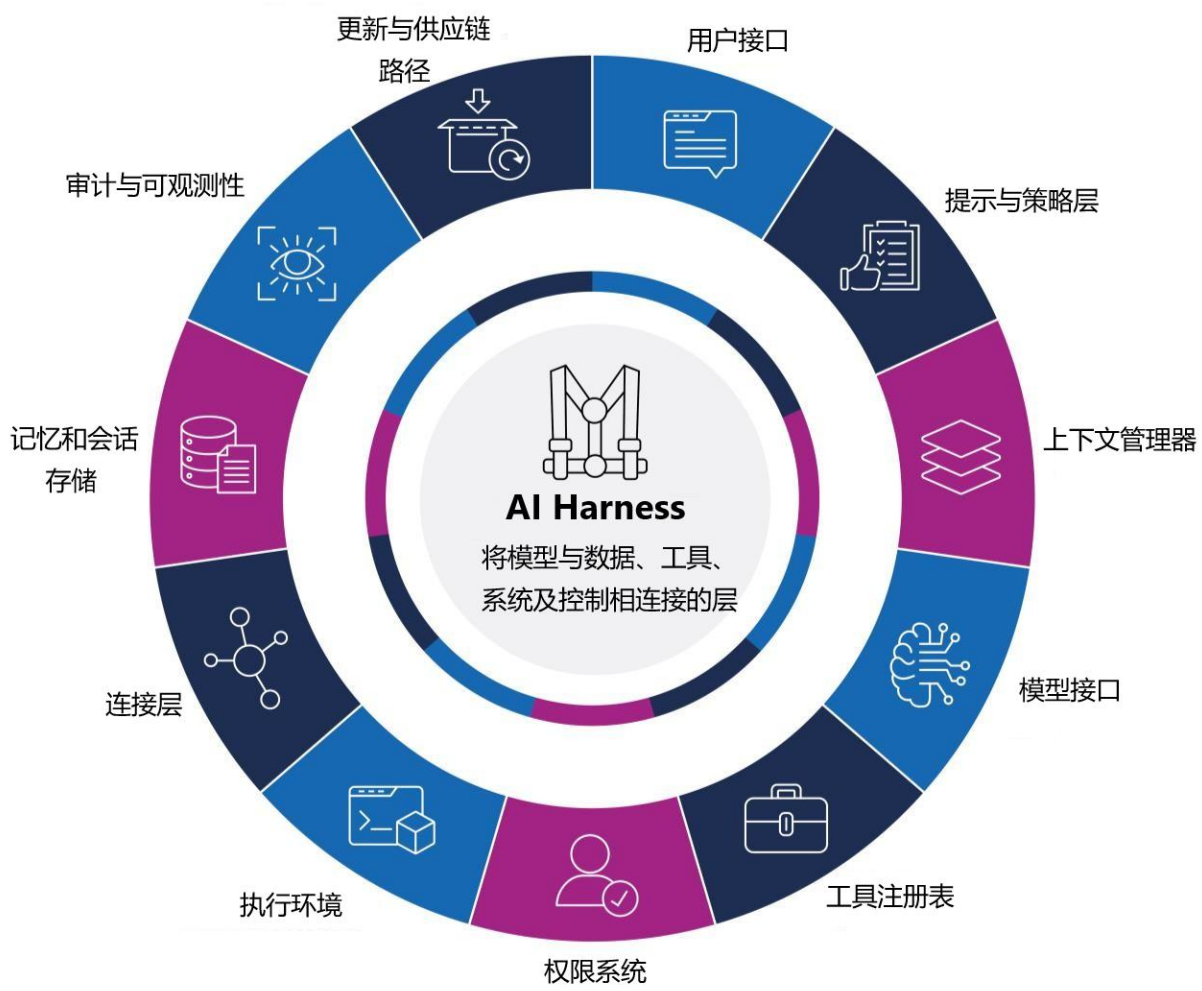
这种区分有两个实际影响。

首先，LLM是可以替换的。一个Harness可以与多个大语言模型配合使用，而设计良好的Harness将比接连不断的模型升级更具持久性。对于大多数组织而言，Harness才是更长期的投资。

其次，许多看似是智能体的故障，其实源于模型与Harness之间的交互，而非模型本身。大语言模型可能会选择不合适的工具、错误使用工具、丢失上下文、超出实际上下文限制，或提出不安全的操作。Harness决定哪些工具可用，执行权限和防护措施，并执行已批准的操作。例如，模型可能会提出一个文件删除命令，但该操作是否执行取决于Harness及其控制机制。当智能体系统表现异常时，组织应同时调查模型的决策过程和Harness的配置，而不是仅依赖于调整提示词。

Harness组件

下文将详细介绍Harness的各组件及其各自的安全意义。



组件	描述
用户接口	确定用户对智能体活动的可见范围，以及可用于审批或干预的控制选项。
提示与策略层	存储系统指令、业务规则和组织策略，以指导智能体的行为。
上下文管理器	选择提供给模型的信息，此过程可能导致敏感信息泄露或遗漏重要上下文。
模型接口	作为测试Harness与组织批准的LLM服务之间的接口，负责发送提示、上下文和模型参数，并接收生成的响应。

组件	描述
工具注册表	定义可用工具及其功能。定义不明确的工具可能会增加运营风险。
权限系统	控制敏感操作的审批和权限。
执行环境	规定操作的执行位置，并可对潜在操作施加限制。
连接器层	将Harness连接到外部数据源、服务和企业系统，包括RAG存储库和支持MCP的工具。
记忆和会话存储	在会话之间保留信息，并可能保存敏感、过时或被篡改的内容。
审计与可观测性	记录活动以供监控、审查和事件响应。
更新与供应链路径	引入了新功能及潜在的供应链风险。

这些组件构成了组织的配置界面：对智能体AI系统的大部分实际控制——例如，哪些工具对外开放、允许执行哪些操作、需要哪些审批以及连接了哪些连接器——都是此列表中的设置。

安全风险与缓解措施

Harness既会扩大智能体AI系统的攻击面，同时也提供了防御攻击的手段。Harness添加的每个外部工具、数据源、连接器和内存存储，都为系统开辟了一条新的入侵路径。与此同时，Harness正是必须实施安全控制的地方，因为智能体AI最显著的弱点无法在模型本身中得到修复。

更广泛的安全考量

Harness既增强了智能体AI系统的能力，也扩大了其攻击面。通过将LLM连接到组织数据、工具、服务和运营环境，它引入了额外的信任关系、依赖关系和集成点，如果未得到适当保护，这些都可能被恶意行为者利用。

智能体AI系统本质上是复杂的。决策和行动可能取决于模型、Harness、工具、数据源和外部服务之间的交互。这种复杂性可能会掩盖故障的根源，并增加故障在互联系统中产生连锁反应的可能性。

组织应在既定的网络安全框架内管理智能体AI安全问题，并应用现有的安全实践，包括“安全设计”原则、深度防御、身份与访问管理、监控及事件响应。Harness是实施和执行这些控制措施的关键场所。

智能体AI的安全风险

当收到查询时，LLM会将指令与信息一并作为上下文接收。因此，模型可能难以区分授权指令与外部内容中的信息，从而为提示注入及其他操纵技术提供了可乘之机。大语言模型处理的内容（包括网页、文档、电子邮件和代码注释）都可能被解释为指令。这一弱点是提示注入攻击的基础，目前尚无完全可靠的技术缓解措施。由于该弱点源于LLM处理上下文的固有机制，因此必须在控制环境中实施缓解措施，通过控制智能体可访问的内容及其被允许执行的操作来实现。

ASD的《审慎采用智能体AI服务》，将智能体AI系统的安全风险分为五类：

- **权限风险：**智能体被授予过度的访问权限，导致单次安全事件便会产生深远影响
- **设计与配置风险：**不安全的架构、集成或设置，在系统投入运行前便已埋下隐患
- **行为风险：**智能体以非预期方式追求目标，或被恶意行为者通过提示注入和数据中毒等技术加以操控
- **结构性风险：**故障在相互关联的组件和多步骤工作流中级联传播，导致难以追溯故障源头
- **问责风险：**智能体系统的复杂性和不透明性使得难以追溯决策、审计行动或明确责任归属。

出于安全考虑，应将多智能体系统视为单一智能体，因为某个组件的被攻破可能通过共享上下文和信任关系向外扩散。

ASD的《审慎采用智能体AI服务》建议在智能体AI系统的整个生命周期中实施深度防御控制措施。关键实践包括：最小权限访问、强身份与访问管理、对工具和外部数据源的受控使用、对智能体输出的验证、对高影响操作的人工监督、持续监控、供应链保障以及智能体能力的分阶段部署。组织应尽可能通过Harness及其周边基础设施来实施这些控制措施，而非仅依赖模型行为或提示指令。

其中许多控制措施是在Harness内实现的，或由Harness强制执行的。

最佳实践

通过使安全控制措施可重复且可执行、将安全测试整合到工作流中、限制不安全智能体行为的影响，以及生成支持监控、调查和持续改进的证据，Harness能够提升组织的网络安全态势。

- 使用由组织控制且经过强化防护的环境，该环境仅包含任务所需的工具和权限。限制对生产系统、敏感数据和网络服务的访问。
- 配置Harness以应用组织的安全编码标准，禁止不安全操作，并验证工具参数和输出结果。对敏感或高影响的操作，必须经人工批准。
- 在将输出结果投入实际运营使用前进行验证。智能体的输出结果和操作在投入运营环境使用前，应经过审查、测试和正式验收，而不能仅因其表述自信就视为正确。Harness可以起草并执行操作，但结果的责任仍由人承担。
- 对API实施成本控制有助于成本监控和安全保障。无论是由托管服务商按代币计费，还是在自建基础设施上作为计算资源消耗，智能体会话的成本都远高于与聊天机器人的一次交互。监控此类消耗可识别出可能表明滥用、系统遭入侵、智能体循环失控或钱包拒绝服务攻击的异常活动。
- 记录提示、响应、工具调用、审批、操作、安全事件和配置变更。保护、保留、审查并独立监控日志，以支持可审计性、安全监控、事件响应、调查和问责
- 选择适合任务的可信模型，并评估其来源、局限性、安全特性和行为。不要依赖模型安全控制来替代由Harness强制执行的控制措施。
- 保持上下文聚焦且相关，仅提供当前任务所需的信息和访问权限。以最低权限原则操作框架。在管理AI工作流时，仅保留与任务相关的信息。若必须精简上下文，应删除过时内容而非对其进行摘要，因为摘要会重写记录，并可能引入新的错误。将持久性事实和决策保存到外部内存或文件中，并用一个包含简明摘要的新会话替换冗长且过时的会话。
- 将探索任务隔离在子智能体中（即Harness为任务的某个限定部分创建的独立智能体实例），并推广其使用。通过应用最小权限原则，组织可将子智能体配置为仅具备特定任务所需的工具、权限和上下文。这既能限制接触不可信内容的风险，又能保持主智能体会话上下文的精简，从而降低安全风险、运营成本及上下文退化。

- 在组织层面维护一个针对Harness的持久性强制规则文件。将标准、禁止区域以及构建和测试命令记录在一个文件中（该文件由Harness在每次会话中读取），可避免每次都需重新查找相同的信息，并使Harness的行为更加一致且更易于审查。

面向高管的治理问题

董事和高级管理人员无需了解智能体AI系统的每个实现细节。然而，他们应确保Harness、其集成方案及其治理控制措施均已得到恰当的设计与实施。

- 该智能体AI系统旨在实现何种业务成果？为何需要采用智能体方法？
- 智能体可以访问哪些数据、系统和工具？是否遵循了最小权限原则？
- 智能体可以执行哪些操作？哪些操作需要人工批准？
- 如何缓解提示注入、数据中毒及其他针对AI的特定攻击？
- 所有重大决策、工具调用和操作能否被监控和审计？
- 我们将如何判断系统是否正在创造价值、安全运行并保持在可接受的风险容忍度范围内？
- 如果Harness遭到破坏、配置错误或被篡改，最坏的结果会是什么，又有哪些控制措施可以防止或限制这种结果？

结论

关于智能体AI的讨论往往聚焦于大型语言模型（LLM），但组织也应同样重视“Harness”。Harness决定了LLM如何与工具、数据和工作流进行交互，并在能力与风险方面都发挥着重要作用。

随着模型的不断演变，Harness很可能成为组织中更具持久性的能力。设计良好的Harness能够在不断更迭的人工智能技术代际中，提供一致性、治理和安全控制。设计不佳的“Harness”甚至会削弱最强大的LLM的效能。

应从低风险用例入手，实施适当的控制措施，保持人工监督，并仅在控制措施日趋成熟后才扩大部署范围。

组织应遵循ASD《审慎采用智能体AI服务》中所述的安全实践，并通过“Harness”及其周边生态系统加以落实。

延伸阅读

英国国家网络安全中心

[提示注入并非SQL注入（其危害可能更大）]

免责声明

本指南中的内容仅供一般参考，不应被视为法律建议，也不应作为在任何特定情况或紧急状况下寻求帮助的依据。对于任何重要事项，您应根据自身具体情况寻求适当的独立专业建议。

澳大利亚联邦对因依赖本指南所含信息而造成的任何损害、损失或费用概不承担任何责任。

版权

国徽的使用

关于国徽的使用条款，详见总理兼内阁部网站（https://www.pmc.gov.au/resources/commonwealth-coat-arms-information-and-guidelines）

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre