TC260-PG-20266A

网络安全标准实践指南

——智能体部署使用安全指引

(v1.0-202607)

全国网络安全标准化技术委员会秘书处

2026 年 7 月

本文档可从以下网址获得：
www.tc260.org.cn/

前　言

《网络安全标准实践指南》（以下简称《实践指南》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织制定和发布的标准相关技术文件，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

本文件起草单位：中国电子技术标准化研究院、国家计算机网络应急技术处理协调中心、清华大学、浦江国家实验室、北京中关村实验室、中央网信办（国家网信办）数据与技术保障中心、浙江大学、阿里云计算有限公司、华为云计算技术有限公司、北京火山引擎科技有限公司、北京航空航天大学、中国移动通信集团有限公司、工业和信息化部电子第五研究所、北京邮电大学、北京升鑫网络科技有限公司、腾讯云科技有限公司、杭州安恒信息技术股份有限公司、方寸跃迁（雄安）科技有限公司、杭州图灵智越科技有限公司、北京神州绿盟科技有限公司、北京奇虎科技有限公司。

本文件主要起草人：姚相振、薄兆一、任奎、徐葳、周睿康、王博、王迎春、崔勇、李琦、郝春亮、王志伟、刘祥龙、秦湛、关振宇、张蕾、张妍婷、徐阳、李子威、孟令宇、赵宇航、崔栋、褚志轩、石桂欣、彭骏涛、耿涛、邵萌、郭建领、袁开国、李筱健、王嘉凯、赵斌、范春鹏、邹权臣、苏锐生、孙立鹏、肖娜、陈星、叶晓虎、郑炎亭、余雄辉、胡皓、陈彦旭、钟宇澄、杨绍波、彭晓军。

声　明

本《实践指南》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《实践指南》的任何部分。凡转载或引用本《实践指南》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。

目　录

1 范围

本文件给出部署使用智能体的安全指引，覆盖评估、准备、部署、使用、停用等阶段。

本文件适用于智能体部署使用的安全风险防范，也可为选择使用商业智能体服务提供参考。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069 信息安全技术 术语

3 术语和定义

GB/T 25069 界定的以及下列术语和定义适用于本文件。

3.1 智能体 agent

具备自主感知、记忆、决策、交互与执行能力的智能系统。

注：本文件中的智能体特指面向个人助手场景、需要用户授予较高权限、基于大模型的人工智能智能体，一般为软件系统。

3.2 工具 tools

赋予智能体调用外部能力的标准化接口，定义了执行某一原子操

作的名称、输入参数和返回结果。

3.3 技能 skills

面向特定任务的标准化指令集合，封装了完成某类任务的最佳实践、操作流程和领域知识。

4 缩略语

API: 应用程序编程接口（Application Programming Interface）

IP: 网际互连协议（Internet Protocol）

5 概述

在部署使用智能体过程中，使用者宜按照本文件第6章至第10章安全指引做好安全防护，并参照本文件附录A进行安全检查。如在安全检查或使用过程中发现存在安全问题，宜及时根据本文件相应章节做好安全加固工作。使用者应遵循国家及所在组织相关规定开展智能体部署使用活动。为应对内部人员部署使用智能体可能带来的各类安全风险，相关组织宜按照本文件附录B安全管理指引，建立相关安全管理制度及配套安全能力。

6 评估阶段

本阶段主要在智能体部署使用前进行评估，包括使用智能体的必要性、开源智能体与商业智能体的适用性，以及拟选择智能体是否具

备基础安全特性要求等方面。

a) 应明确使用目的、业务场景、工作任务，评估采用智能体完成相关任务的必要性与合理性。

b) 应预先了解智能体的技术特性、能力边界、安全风险，以及开源智能体与商业智能体的主要差异。

c) 根据 a）需求分析与 b）知识准备，谨慎选择拟部署使用的智能体，宜优先选择由提供方同时提供智能体和配套安全防护能力的成套方案，慎重选择缺乏安全防护措施的开源项目。

注：选择商业智能体时，应认真阅读提供方发布的安全注意事项，并依据本文件后续章节核对其安全特性，在满足要求时无需重复采取后续安全措施。选择需自行实施安全防护的开源智能体时，应依据本文件后续章节加强安全防护。

d) 应检查拟部署使用智能体项目的维护状态，存在下列情形之一的智能体项目，不应选择：

1) 超过一个月无维护记录，且存在未解决的安全问题；

2) 已公开的安全问题长期未获响应；

3) 存在未修复的高危漏洞；

4) 提供方明确已停止维护。

e) 宜优先选择集成安全沙箱、高风险操作管控、急停控制、可回溯文件系统等机制的智能体；不应选择缺少日志审计、权限管理等基础性安全机制的智能体。

f) 应检查拟部署使用的智能体是否存在自动开放公网接口、强制

回传运行数据等重大安全隐患。

7 准备阶段

本阶段主要对智能体的安装部署进行准备，包括安装物料、部署环境、大模型、安全防护等方面。

a) 选择智能体安装物料时：

1) 应从提供方官方网站、应用商店等具备明确责任主体和发布机制的渠道获取智能体安装文件、容器镜像等安装物料；不应使用来自网盘、论坛、即时通信群组等渠道的修改版、二次分发版、破解版或其他不明来源的安装物料；

2) 安装智能体前，宜通过发布主体提供的数字签名、杂凑值、校验码等方式验证安装物料的真实性和完整性，防范安装物料被投毒或篡改。

b) 根据智能体安装部署方式，采取相应的安全防护措施：

1) 拟在本地环境直接部署的，应选择独立设备安装智能体，不宜与日常生活、工作办公等用途的设备混用。部署前宜对设备内原有文件进行备份；设备中存在敏感数据、隐私文件的，应提前进行清理或迁移；

2) 拟在虚拟机、容器等虚拟化环境部署的，应通过禁止虚拟机访问其宿主机内文件等配置实施严格隔离，防止安全问题扩散至宿主机；

3) 拟在云环境部署的，应选择具备智能体身份管理、访问控制、日志审计和安全告警等能力的云平台。部署前宜向云平台服务提供商确认其提供的安全能力是否满足本文件相关要求。

c) 选择智能体运行所需大模型时:

1) 应选择已通过生成式人工智能服务备案的大模型;

2) 对数据安全、隐私保护有相关要求，且具备相关算力条件和技术基础的，宜优先采用本地化部署的大模型;

3) 调用外部大模型服务的，应使用大模型服务提供方的官方接口，不应通过来源不明的中转站、未授权代理等方式调用大模型;

4) 应根据智能体任务复杂度和安全需求，选择上下文窗口、能力边界、安全水平满足要求的大模型，避免因上下文窗口长度或安全能力不足引发智能体误操作。

d) 应根据安全需求，对所选智能体及大模型已有安全机制进行差距分析。以下安全能力存在不足的，宜通过部署安全工具或接入安全服务等方式进行增强:

1) 输入输出内容安全保护;

2) 智能体行为管控（如运行时行为监控、异常熔断等）;

3) 智能体身份与凭证安全管理;

4) 智能体供应链（如技能、工具等）安全检测;

5) 高风险操作拦截与管理;

6) 基于沙箱等技术的环境隔离;

7) 资源消耗与成本控制。

8 部署阶段

本部分对智能体部署阶段提出安全指引，包括部署方式、插件安装、运行账户、权限配置、网络暴露控制、日志审计、高风险操作管理等方面。

a) 应根据智能体提供方的官方文档或经核验的部署脚本进行部署，不应采用来源不明的一键部署等便捷化脚本。

b) 需要额外安装插件的，应在安装前检查插件来源的可靠性，不应安装来源不明、存在未修复高危漏洞、权限需求异常或与预期任务无关的插件。

c) 不应使用操作系统管理员权限运行智能体。宜仅授予智能体完成预期任务所必需的最小权限，以降低智能体越权操作和被滥用时的影响范围。

d) 应限制智能体的可访问目录范围，不应授予智能体使用者主目录、系统目录等重要位置的默认访问权限。宜为智能体建立专用工作目录，仅将任务必需文件置入该目录中。

e) 应将智能体及其相关服务配置为仅本机可访问，不开放至外部网络，防止未授权的远程调用。确需通过互联网访问智能体的，

应配置加密访问，并限制访问来源。

f) 应开启智能体的日志记录功能，对文件操作、指令执行、网络活动、技能调用等行为进行全量日志记录，宜将记录粒度设置为可支持的最细级别。

g) 应提前建立高风险操作清单，并通过智能体安全机制进行管理，对清单内操作实行二次确认或直接阻断，高风险操作包括但不限于：停止与停用系统服务、终止进程、格式化或复写磁盘、批量删除、权限修改、密钥更改、开放端口、修改防火墙规则、修改系统设置、访问密码管理器等。宜根据智能体使用目的，将涉及转账、支付等影响个人财产安全的操作列入高风险操作清单。

9 使用阶段

本部分对智能体的使用阶段提出安全指引，包括安全复查、技能安全、敏感信息保护、网络访问安全、长期记忆管理、应急处置与更新等方面。

a) 应对智能体安装与配置情况进行复查，包括但不限于：

1) 使用者是否可以有效了解智能体执行状态；

2) 使用者是否可以对智能体所执行的操作进行安全管理，例如配置黑名单、白名单等；

3) 使用者是否可以便捷地急停控制智能体。

b) 加强对技能使用的安全防护，包括但不限于:

1) 应使用来源可靠的技能;

2) 宜对技能进行全面安全测试，使用未经安全测试的技能时，应提前做好风险防范准备。

c) 在使用连入互联网络的智能体时，应注意保护敏感数据安全，包括但不限于:

1) 向智能体提供个人信息时，应坚持最小必要原则，涉及生物特征、家庭隐私等敏感信息的，应审慎提供;

2) 未获得第三方授权时，不应向智能体提供第三方的个人信息及敏感数据;

3) 不应向智能体提供使用者未获得许可的业务数据及涉及知识产权的内容。

d) 通过互联网访问智能体的，应定期检查智能体公网接口必要性及安全性，及时关闭非必要暴露的接口。

e) 使用智能体获取网络信息服务或参与网络社交媒体讨论的，应符网络安全相关法律法规要求。

f) 宜定期开展智能体运行环境内有价值数据的容灾备份。

g) 宜定期手动查看长期记忆文件记录的内容，对不宜存储的个人信息、隐私数据、内部信息及时进行处理，并相应调整长期记忆配置策略。

h) 宜及时回收敏感权限，定期清理不再使用的技能与敏感对话记

录，开启多因素鉴别以保护账户安全。

i) 应及时对已部署安全工具、安全服务的告警信息进行核查处置，并根据获取到的其他安全威胁信息、风险提示等，对智能体部署使用情况进行安全复查，及时消除安全隐患。

j) 应持续关注智能体运营方、安全企业发布的安全公告，当出现影响特定版本的安全事件时，根据安全公告采取相应措施（如升级、降级或版本锁定），避免长期使用存在已知漏洞或已停止维护的版本。

10 停用阶段

本部分对智能体停用阶段提出安全指引，包括安全地终止运行、清理数据与权限、完成环境回收等方面。

a) 应停止智能体主程序，以及所有关联服务与后台进程的运行，确认智能体已完全退出运行状态。

b) 在清理环境前，应对仍需保留的会话记录、配置文件、知识库文件、日志等数据进行容灾备份。

c) 完成必要数据备份后，根据部署环境执行相应的环境清理操作：

1) 本地环境部署的，如智能体提供方配备官方卸载程序，应通过官方卸载程序进行卸载，并在卸载后检查相关进程、端口和网络连接等情况确认无残留；如无官方卸载程序，

应重置操作系统；

2) 云环境部署的，应关闭与该智能体有关的公网访问入口，删除不再使用的工作文件，以及该智能体相关的知识库、日志、长期记忆、插件和技能配置等数据，并同步撤销 API 密钥、服务账号凭据、第三方应用授权等相关凭证和权限；

3) 虚拟化环境部署的，宜直接停用该虚拟化环境，无法停用该虚拟化环境的，应参照 1）中相关指引进行环境清理；

4) 停用智能体后，应及时确认终止相关服务、取消订阅及自动续费，关注大模型接口是否产生异常费用。

附　录　A

（资料性）

安全检查清单

表 1 部署使用安全检查清单

对应阶段	检查项	重要性	是否完成
评估阶段	判断是否需要由智能体去解决目标工作任务	★★★	□
	整体了解智能体的有关知识	★★★	□
	对需要部署使用的智能体进行选择	★★★	□
	选择自带安全防护机制的智能体	★★★	□
	不使用存在显著不合理安全隐患的智能体	★★★	□
	检查拟部署使用的智能体项目近期动态	★★☆	□
准备阶段	从具备明确责任主体和发布机制的渠道获取智能体安装物料	★★★	□
	选择合适的部署方式，并落实相应安全防护措施	★★★	□
	选择合适的智能体运行相关的大模型	★★★	□
	校验智能体安装物料的真实性和完整性	★★☆	□
	通过安全工具、安全服务等方式增强防护能力	★★☆	□
部署阶段	不通过未查明来源的一键部署等便捷化脚本进行部署	★★★	□

	对智能体的网络暴露形式进行安全配置	★★★	☐
	开启智能体的日志记录功能，并对该功能进行配置	★★★	☐
	安装前检查插件来源的可靠性	★★☆	☐
	不使用操作系统管理员权限运行智能体，最小化授予智能体的权限	★★☆	☐
	限制智能体的可访问目录范围	★★☆	☐
	提前定义高风险操作，并通过智能体安全机制进行管理	★★☆	☐
使用阶段	加强对技能使用的安全防护	★★★	☐
	加强对自身个人信息以及敏感数据的安全防护	★★★	☐
	合法合规使用智能体获取网络信息服务或参与网络社交媒体讨论	★★★	☐
	对安装与配置情况进行复查	★★☆	☐
	定期检查智能体公网接口必要性及安全性	★★☆	☐
	定期对智能体所处环境内的有价值数据进行备份	★★☆	☐
	定期手动查看长期记忆文件记录的内容并对其进行管理	★★☆	☐
	及时回收敏感权限，定期清理不再使用的技能与敏感对话记录	★★☆	☐
	关注告警信息并进行核查处置，及时消除安全隐患	★★☆	☐
	关注智能体相关安全公告，及时采取相应措施	★★☆	☐

停用阶段	在卸载前停止运行	★★☆	☐
	卸载前对需要保存的数据进行导出或转存	★★☆	☐
	执行环境清理操作	★★☆	☐

附 录 B

（资料性）

安全管理指引

组织内部存在使用智能体相关情况的，建议围绕以下几方面做好管理工作。

a) 建立内部智能体使用管理制度，对智能体的部署、使用、变更等活动进行管理。管理制度至少包括以下内容：

1) 禁止行为要求。明确智能体使用过程中禁止实施的具体行为，例如不将组织内部数据或系统接口接入未经审批通过的智能体等；

2) 使用边界要求。遵循最小授权原则，明确智能体的使用范围、权限边界和数据处理边界；

3) 审批流程要求。明确部署智能体前的申请、评估、审批和备案流程，提交的申请内容包括使用目的、使用边界、拟配置技能、拟接入工具、安全防护措施等。

b) 建立智能体资产登记表，对已审批智能体进行统一登记，并在发生变更时进行更新。资产登记表至少记录以下信息：

1) 智能体信息，包括智能体名称、用途、业务归属；

2) 部署位置信息，包括主机物理位置、IP 地址、容器名称或其他可唯一识别的智能体标识；

3) 开发者信息，包括部署人、部署时间、责任人；

4) 模型信息，包括使用的大模型名称、版本及接入方式；

5) 工具组件信息，包括安装的技能、插件及其他外部工具列表；

6) 访问控制信息，包括可访问的内部系统、接口和数据范围；

7) 授权信息，包括权限配置情况；

8) 审查记录信息，包括最近一次安全审查时间及审查结果。

c) 对已审批智能体活动进行管理：

1) 对关键活动进行日志记录，包括高权限指令执行、敏感数据访问、内部系统接入、外部数据传输、权限变更、工具调用和异常操作等；

2) 使用人工智能、规则分析或其他技术手段，对智能体行为模式进行风险分析；

3) 周期性对日志记录进行分析排查，及时发现并处置安全问题，形成记录。

d) 宜通过以下手段建立未审批智能体的发现能力，以便于进行进一步的网络隔离、权限降级、进程阻断等处理：

1) 定期扫描组织内网中典型智能体服务端口；

2) 分析网络流量日志，识别异常的智能体通信行为；

3) 识别与已知大模型 API 端点通信的内网主机；

4) 检索安装或运行智能体相关进程的终端。

e) 面向组织内部员工开展安全教育，提升员工对智能体安全风险和安全使用的认知。安全教育内容包括提示词注入攻击、供应链安全、凭证泄露、数据泄露、越权访问、违规使用责任等。