中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

OpenClaw类智能体部署风险管理指南

中国人工智能产业发展联盟安全治理委员会

2026年4月

版权声明

前　言

OpenClaw 作为 2026 年现象级开源 AI 智能体执行框架，可看作连接大语言模型与本地系统、外部工具的执行中枢，具备系统级权限，支持执行 Shell 命令、访问文件系统、操控浏览器等操作能力。从 OpenRouter 平台数据来看，OpenClaw 已成为全球活跃度最高的 AI 智能体框架。当前，OpenClaw 正从开源社区热潮迈向规模化试点部署阶段。国内头部 AI 公司与云厂商普遍推出一键部署镜像与专属托管服务，在金融、制造、政务等领域以私有部署、内网隔离、权限最小化为原则开展试点应用。随着 OpenClaw 及同类智能体（全文统称为 OpenClaw **类智能体**）在各类场景中的加快应用，其在权限管理、工具调用、会话隔离、配置安全、持续运维等方面的风险日益显现。

为帮助个人和企业提前识别部署隐患、建立风险管理基线，中国人工智能产业发展联盟（AIIA）安全治理委员会结合产业实践，研究形成本指南。指南面向 OpenClaw 类智能体服务提供方和使用方，基于技术架构和风险暴露面分析，总结了技术和管理维度的典型风险，并基于此提出**操作可信、权限可控、风险可溯**三大原则，以及覆盖**部署、使用、下线**全过程的安全部署总体框架和自查规范，旨在帮助个人和企业在引入 OpenClaw 类智能体时系统建立风险意识，落实安全管理措施，推动智能体从“能部署”迈向“安全部署、规范使用、持续治理”，以安全底座支撑新质生产力高质量发展。

编写委员会

主要编写单位：中国信息通信研究院、腾讯云计算（北京）有限责任公司、北京百度网讯科技有限公司、华为技术有限公司、科大讯飞股份有限公司、北京智谱华章科技股份有限公司、上海稀宇科技有限公司、上海阶跃星辰智能科技有限公司。

主要参编专家：魏凯、石霖、呼娜英、刘铂、郭苏敏、徐鹏、李滨、刘志高、周启明、张玉峰、康雨辰、朱希、李洋、李岳阳、包沉浮、冯景辉、徐艺溦、柳嘉琪 、郑贵、刘潇、赵瀚霖、覃波、薛柯、高翔、王笑尘、陈荣、乔文斌、张晨、买尔旦、沈俊成、鲍景雨。

目　录

一、概述

智能体技术正加速应用落地，OpenClaw 类智能体凭借系统级权限优势实现多场景试点部署，但其安全风险防控体系尚未完善。本章通过阐释 OpenClaw 与智能体核心内涵、梳理行业应用与治理趋势、分析部署核心风险，为后续风险管理体系构建奠定认知基础。

（一）OpenClaw 与智能体简介

OpenClaw（俗称“龙虾”，曾用名 Clawdbot、Moltbot）是一款**开源、可自托管的多渠道智能体网关与执行框架**，可对接微信、钉钉、飞书等通讯软件，支持多智能体并行调用，具备会话管理、记忆存储、任务分配、工具调用、远程节点连接与双向通信等功能，并可通过插件或技能（Skills）扩展接入多类业务系统与数据资源。其核心为由大语言模型驱动的智能体，可根据目标拆解任务，并按需多步调用外部工具；相较于对话类应用，OpenClaw 类智能体不仅输出文本结果，还可依据环境反馈调整行动策略，形成思考与执行闭环，其核心能力通常包括任务规划、工具调用、持久记忆以及运行状态与上下文管理等。

（二）智能体应用与治理趋势

当前，人工智能正加速**从单一对话工具向多能力自主代理演进**，大语言模型的自然语言理解、规划推理、工具调用能力持续提升，推动智能体技术快速发展并进入产业化应用阶段。其应用形态不断丰富、应用场景持续拓展，已成为千行百业数字化转型的重要抓手。

在**个人应用场景中**，智能体应用快速渗透，典型形态包括通用对话助手、设备端智能体以及垂直场景专用智能体。整体呈现端云协同、多模态

交互、自主规划、跨设备流转等趋势，从被动问答转向主动服务，从通用能力走向垂直深耕，持续融入日常出行、居家、健康、学习与办公等高频场景，逐步成为个人数字生活的核心入口与贴身助手。

在**企业应用场景中**，智能体的典型应用包括客服自动化、知识库问答、业务流程自动化、代码辅助、跨系统数据整合等。其核心价值在于将原本需要人工逐步执行的流程交由模型自主编排执行，缩短业务处理周期，提升企业运营效率。

全球各国及相关机构也加快推进智能体领域治理布局。2025 至 2026 年期间，多个国家和地区陆续发布与智能体及其治理相关的研究报告、框架规范与标准文件。其中，新加坡资讯通信媒体发展局（IMDA）、英国信息专员办公室（ICO）先后出台相关治理框架；我国则依托《网络安全法》《数据安全法》《个人信息保护法》以及《生成式人工智能服务管理暂行办法》等法律法规，初步构建起覆盖智能体技术研发、应用、数据安全与个人信息保护的全方位治理体系。

（三）OpenClaw 类智能体风险概述

与仅具备内容生成和问答能力的传统对话式人工智能应用不同，OpenClaw 类智能体的核心特征是拥有系统级操作权限与工具调用能力，可在对话触发下对终端命令、文件系统、浏览器、业务接口等各类资源进行访问与操作，其风险暴露面大幅扩大，风险边界从内容层面延伸至系统操作、数据流转、业务执行全维度，风险传导性、影响范围、隐蔽性显著增强。

从风险触发场景来看，身份冒用、提示词注入、工具滥用、权限配置不当、第三方组件漏洞等均可能引发安全事件，导致敏感信息泄露、系统配置被篡改，引发业务中断、资产损失、不可逆操作等严重后果，且由于OpenClaw类智能体具备跨系统编排、多工具联动的能力，单一风险点易通过工具链、系统链产生链式放大效应，形成级联失控，进一步扩大风险影响范围。

从风险防控难点来看，OpenClaw类智能体普遍具备多入口接入、插件技能扩展、第三方组件依赖的特性，不仅增加了攻击面与供应链安全风险，还导致其运行过程呈现自主规划、多步骤执行、跨系统交互的特点，传统的日志审计、监控预警手段难以全面覆盖其决策链条与操作链路，易形成观测盲区，增加风险发现与溯源难度。

此外，当前OpenClaw类智能体正处于规模化试点部署阶段，部分主体存在技术架构风险认知不足、安全管控体系不完善、责任边界不清晰等问题，进一步加剧了安全隐患。因此，部署OpenClaw类智能体需将其作为具备系统性影响的基础设施进行全生命周期治理，明确责任边界与信任边界，建立动态的风险识别、管控与处置机制，守住安全合规底线。

二、概念内涵与适用范围

本指南适用于OpenClaw类智能体的部署与运营风险管理。为明确适用范围边界，本章对相关概念进行统一界定。

（一）OpenClaw类智能体定义及类型

OpenClaw类智能体是指以OpenClaw框架或其核心设计理念为基础，具备多入口接入、多智能体编排、工具与插件扩展能力，并可在对话触发

下执行系统级或业务级操作的智能体执行框架，其核心功能是实现任务自动化、场景化智能交互与业务流程赋能。根据开发与改造形态，本指南提及的 OpenClaw 类智能体可分为以下三类：

一是原生部署型，即直接采用官方发布的 OpenClaw 框架及配套组件，未进行任何功能或架构上的修改，直接部署使用；

二是定制扩展型，以官方框架为基础，结合具体业务需求，对框架功能、插件配置、交互逻辑等进行定制化修改、优化或扩展，适配特定场景的使用需求；

三是自研同类型，指未直接采用官方框架，但核心设计理念、功能架构与运行逻辑类似，自主研发的具备同类智能执行能力的智能体。

（二）OpenClaw 类智能体服务使用方

OpenClaw 类智能体使用方是指部署、运行并实际使用 OpenClaw 类智能体的主体，按使用主体类型分为个人用户与企业用户。部署方式上，支持本地部署、云端部署或端云协同部署等形态。个人用户以自然人身份使用智能体服务，多面向轻量化、日常化场景，对自身使用行为、数据安全及操作后果承担直接责任；企业用户以组织形式在内部业务场景中规模化应用，承担场景需求确认、权限与数据边界管理、日常运营维护、内部合规落地等职责，并对实际应用效果及潜在风险负责。本指南后续将其简称为**“使用方”**。

（三）OpenClaw 类智能体服务提供方

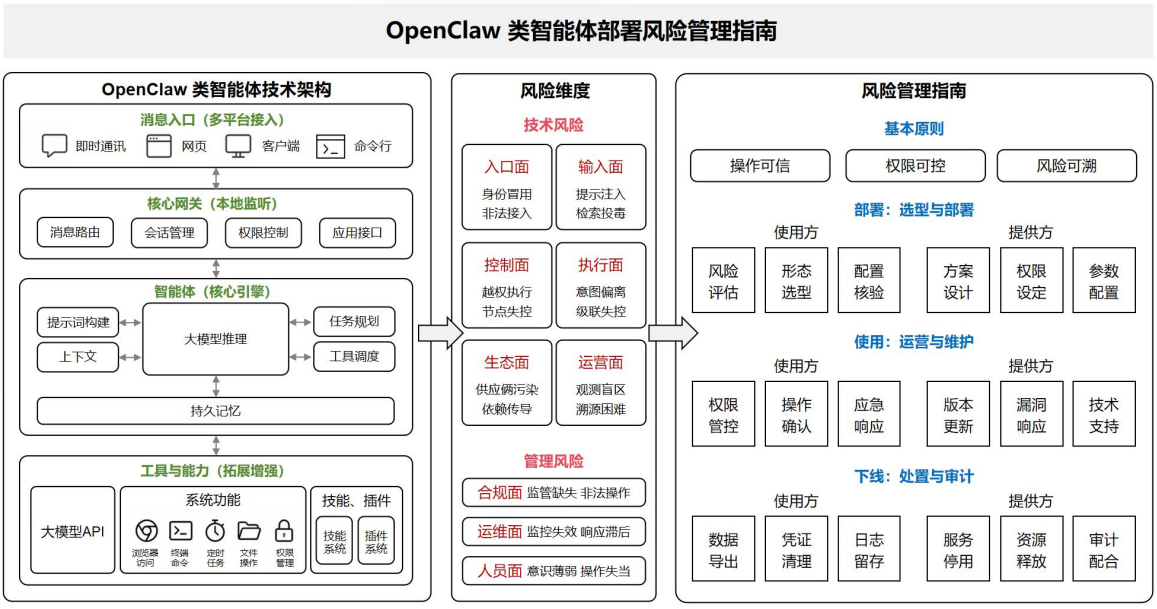
OpenClaw 类智能体服务提供方是指为使用方提供 OpenClaw 及同类相关产品、技术服务或托管服务的企业或组织，如研发供应商、系统集成

商或授权服务商等。服务提供方通常负责交付安装与部署方案、提供运行支撑能力及必要的安全配置建议，并就版本来源、更新机制与服务保障等向使用方提供可核验的信息与材料，保障平台基础能力稳定、安全与可追溯。本指南后续将其简称为**“提供方”**。

三、OpenClaw 类智能体部署风险

（一）技术架构与风险暴露面

OpenClaw 类智能体并非单一问答工具，而是具备系统访问、接口调用与自动化执行能力的复杂代理平台。其技术架构分为消息入口、核心网关、智能体、工具与能力四个层次，每一层级均可能引入新的攻击面。



来源：中国人工智能产业发展联盟安全治理委员会编制

图 1：OpenClaw 类智能体技术架构与风险管理全景图

消息入口支持多平台接入，包括即时通讯、网页、客户端与命令行等多种入口形态，并承担不同渠道的认证与消息格式适配，实现输入输出的统一与标准化。

核心网关负责在本地监听请求与统一集中治理，承担消息路由、会话管理、权限控制与应用接口等关键能力，作为策略下发与控制的集中节点，对入口请求进行校验、编排与调度。

智能体作为核心引擎，完成推理与执行闭环，围绕大模型推理组织提示词构建与上下文管理，并结合任务规划与工具调度完成多步骤执行，同时以持久记忆支持跨轮次状态保持与任务连续性。

工具与能力层覆盖大模型 API、浏览器访问、终端命令、定时任务、文件操作、权限管理，以及技能系统（Skills）与插件系统等，可按场景接入外部能力与业务系统。

表 1：OpenClaw 类智能体架构与风险关联关系

架构层次	核心功能	主要关联风险
消息入口	即时通讯、网页、客户端、命令行等入口接入与认证适配	入口面风险、输入面风险、运营面风险、人员面风险、运维面风险
核心网关	消息路由、会话管理、权限控制、应用接口统一治理	入口面风险、控制面风险、运营面风险、运维面风险
智能体	提示词构建、上下文管理、大模型推理、任务规划、工具调度、持久记忆	输入面风险、生态面风险、执行面风险、运营面风险、合规面风险、运维面风险
工具与能力	大模型 API、浏览器访问、终端命令、定时任务、文件操作、权限管理、技能/插件	执行面风险、生态面风险、运营面风险、合规面风险、运维面风险

（二）技术风险

基于上述架构层次，OpenClaw 类智能体典型技术风险包括：

1、**入口面风险：**身份冒用与非法接入。入口面承担多渠道消息接入与认证适配功能，攻击者通过账号冒用、令牌泄露等方式获取智能体访问权限。多入口暴露面增加，未授权接入风险随通道数量递增。

2、**输入面风险：**提示注入与检索投毒。攻击者将恶意指令夹带在各类文本中，诱导智能体忽略约束执行越权操作，造成隐蔽难查的安全事件。输入违规提示词、模型未经安全评估和模型幻觉等因素均可能导致智能体生成违法违规和不良内容，或在传播中触发舆情与合规问题。

3、**控制面风险：**越权执行与节点失控。智能体权限授权、身份继承与敏感资源边界界定不清，易形成权限蔓延，从读取权限滑向越权执行。智能体连接的执行节点被恶意接管，可能被作为跳板访问内网，引发横向移动与持续控制风险。

4、**执行面风险：**意图偏离与级联失控。用户表达、模型理解与工具执行并非天然一致。智能体执行删除、转账等难以撤回的动作，若缺乏人工确认，错误会被快速放大并造成损失。在多智能体协同环境中，单一智能体被劫持后，攻击者极易利用低权限智能体作为跳板，通过智能体间的信任链条实现级联传播。自主执行过程中具备自动规划、连续调用等特性，系统极易出现循环调用接口、重复触发外部工具或大量消耗推理词元（Token）等异常行为。

5、**生态面风险：**供应链污染与依赖传导。智能体相关框架、插件或部署环境的漏洞被利用，因组件集成多、更新不及时、供应链攻击等，漏洞风险易累积。调用过程中数据传输、留存不当引发数据泄露风险。贯穿

智能体运行全流程，可能在对话、日志、插件等环节暴露个人敏感信息、商业秘密。

6、**运营面风险**：观测盲区与溯源困难。智能体具备自主规划、多步骤执行、跨系统调用的特性，传统日志审计难以覆盖其复杂的决策链条和工具调用链路，形成观测盲区。缺少全链路日志记录与关联分析，攻击行为难以发现；异常调用、越权操作、敏感信息外发等情形无法及时告警；安全事件发生后难以追溯决策过程、定位责任主体。

（三）管理风险

部署 OpenClaw 类智能体的管理风险具体阐释如下：

1、**合规面风险**：因未充分识别或落实数据、网络及行业监管要求，如数据出境合规、个人信息保护、网络安全、数据安全、大模型备案等，可能导致部署形态、数据流向及日志留存不合规，侵犯商业秘密、版权等。

2、**运维面风险**：运维管理机制不完善，缺乏持续监控、告警及应急能力，运维人员无法及时发现智能体异常调用、越权操作等问题。

3、**人员面风险**：部署与使用人员安全意识薄弱、操作不规范、权限管理松散，如共享账号、违规开通高权限等，因智能体具备自动化执行能力，此类疏忽易被放大，引发误操作、数据泄露或被黑客利用。

四、OpenClaw 类智能体部署风险管理实施规范

（一）风险管理基本原则

结合 OpenClaw 类智能体部署运营全流程风险特征，本节提出**操作可信、权限可控、风险可溯**三大核心风险管理原则，适用于系统选型、部署实施、运行维护与下线处置等全生命周期环节，具体阐述如下：

操作可信：保障 OpenClaw 类智能体操作的可验证、可管控机制，有效防范恶意指令注入、未授权操作等风险，提升操作行为的合法性与安全性，夯实智能体执行闭环的信任基础。

权限可控：严格遵循最小权限理念，明确 OpenClaw 类智能体各类操作的权限边界，实现权限分级授予、动态调整与定期复核，严防权限滥用、越权操作等风险。

风险可溯：对 OpenClaw 类智能体部署、运营、停用全流程的操作行为、数据流转、风险事件进行全面记录，确保所有操作可追溯、风险可核查、责任可界定。

（二）部署阶段：选型评估与安全配置

部署阶段是风险管控的第一道关口，使用方与提供方应协同完成选型评估与安全配置。

1、使用方

（1）场景风险评估：结合使用场景和需求，识别智能体部署风险等级，重点明确智能体与本地系统、知识库、数据库的对接边界，需匹配场景和数据的敏感等级与合规要求，为后续部署形态选择、权限配置提供核心依据。

□**场景风险分级**：明确智能体的使用场景，如内部辅助、对外服务或业务决策等。分析智能体决策对业务运行的影响程度、错误容忍度以及是否涉及自动化执行，评估其整体风险。

□**交互风险评估**：分析智能体与外部环境交互的核心边界，包括由内网与互联网构成的网络边界、本地系统构成的系统边界、知识库与数据库

构成的数据边界、外部调用协议构成的接口边界，以及扩展插件、技能构成的工具边界，对各维度交互点开展风险评估。

□**合规要求对齐**：依据具体使用场景，如是否涉及金融、医疗、教育等领域，识别需遵循的法律法规及标准规范。确保智能体建设在满足效率的同时，严格符合法律监管与安全标准。

（2）部署形态选型：结合数据敏感度、业务连续性需求与自身资源条件，选择适配的部署形态，明确各方责任边界。

□**架构选型适配**：依据业务场景、数据安全等级与运维能力，在本地私有化部署、云端托管部署、端云协同部署三类模式中选择适配架构，兼顾安全性、可用性与建设成本。

□**明确数据管控**：确定数据是否允许出域，明确是否与内网、专网隔离，严格限制外部模型服务与第三方能力的接入范围，防范数据泄露与外部风险传导。

（3）权限配置核验：以最小权限为核心原则，对提供方交付的权限配置开展全维度核验，确保权限配置与场景风险等级、合规要求一致，防范权限滥用风险。

□**权限配置核验**：对照场景风险评估结果，核验网络访问、系统接入、数据读写、接口调用等权限配置，确保权限范围与业务需求匹配，无超范围授权。

□**白名单机制核验**：验证工具、接口、插件等白名单配置的有效性，确认高风险工具默认禁用、高权限操作需人工确认，符合安全管控要求。

□**身份凭证核验**：核查账号、密钥、令牌等身份凭证的管理机制，确认无共享账号、明文存储等违规配置，凭证存储、轮换流程合规。

2、提供方

（1）部署方案设计：提供适配场景的产品部署形态与工程化交付方案，明确控制边界与验收要点。

□**信息告知披露**：通过用户协议、隐私政策、安全与风险使用指南或相应合同约定，以显著方式、清晰易懂的语言，真实、准确、完整地向使用方告知个人信息处理事项、系统关键配置、潜在风险分级及双方权责。

□**部署形态设定**：结合使用方场景与需求，确定本地、云端或混合部署形态，根据实际需要将容器、沙盒等虚拟化隔离措施作为关键安全配置纳入部署方案。

□**模型可信评估**：核验基础模型提供商资质、版本标识及服务形态（直供或代理中转），评估模型安全机制、内容安全机制及供应链可追溯性；对微调或蒸馏版本同步开展安全评估；建立模型版本管理制度，支持异常情况下快速回滚；部署输入安全过滤与输出安全审核机制。

□**信源可信评估**：智能体配置的默认联网搜索信源或知识库，应优先采用可信、权威的来源。应公布主要信源列表或选择标准。当智能体需调用第三方接口时，应评估该接口服务商的安全合规水平。对于涉及处理敏感数据的场景应优先选择经安全认证的接口服务。

□**信任边界划分**：当多用户共用同一网关渠道时，应识别其隐含的共享工具授权风险，按信任等级拆分网关实例、访问凭证及执行环境，避免高敏感操作与低敏感场景混用同一权限上下文。

（2）权限边界设定：以最小权限为核心原则，明确智能体可操作范围与权限限制，设置动态白名单审核机制，避免智能体权限滥用引发风险。

□**网络访问权限管控**：明确网络通信边界与访问规则，限定智能体可访问的网络域、IP 段与端口，实施内外网隔离与访问控制，防范外部风险向内网渗透。

□**本地业务系统接入管控**：界定智能体与本地业务系统的接入边界，严格控制可访问系统范围与操作深度，按场景分级授权，禁止未经许可的系统访问与跨域操作。

□**数据读写管控**：识别数据类型与敏感级别，明确数据在收集、存储、使用、加工、传输、提供、公开等数据处理各环节的去向与存储位置；界定智能体对知识库、数据库表、工作区及文件目录的读写、修改与删除权限，实施最小化访问与数据隔离。

□**外部接口调用管控**：梳理外部调用接口清单，建立接口白名单机制，限定调用频率、传输内容与授权范围，防范未授权调用与数据泄露风险。

□**扩展插件与工具管控**：采用白名单策略，对插件、技能及第三方工具实行动态准入审核，明确可加载范围、执行环境与执行权限，禁止高风险工具默认启用，高风险工具执行操作须经人工确认。

□**身份账号与密钥凭证管控**：实行账号管理与身份认证，严禁共享账号；统一管理密钥与凭证，规范存储、使用与轮换机制。

□**记忆安全管理**：界定智能体长期记忆存储边界，实施敏感信息识别与脱敏；建立记忆生命周期管理机制，含访问控制、定期清理与过期淘汰，防止历史会话数据泄露。

（3）关键参数配置：依据部署方案和权限管控需求，配置加固关键参数，确保交付系统具备安全可控能力。

□**访问与鉴权设置**：配置访问控制与鉴权策略，收敛网络入口，对使用方实施身份校验；配置特定安全策略防范非法接入与指令注入风险。

□**人工确认与风险控制**：按业务敏感等级设置人工确认环节，对数据改写、脚本执行等高风险操作构建动态防护；限定智能体授权运行边界。设置多智能体协同安全策略，防范级联失控。

□**沙箱与边界防护**：通过虚拟环境隔离、目录权限约束等方式限制智能体交互范围，禁止越权访问敏感配置；启用网络访问管控，拦截内网敏感目标探测行为，避免敏感数据泄露。对提供的服务以及运行环境进行必要的安全检测与防护。

□**审计与溯源配置**：配置全链路日志审计策略，对服务运行中的相关操作与处理行为进行记录；规范日志存储周期与加密权限，实现高风险操作流程可追溯、不可篡改。

（三）使用阶段：运营管控与应急管控

运营阶段是风险持续暴露和累积的阶段，需要建立常态化的安全运营监控与应急机制。

1、使用方

（1）权限动态管控：以最小权限为核心原则，并根据使用需要循序放开，同时建立持续的权限复核与整改机制，确保权限与职责相匹配。

□**白名单动态运维**：工具、技能、插件、接口、协议（如 MCP、A2A）、文件目录、数据、网络访问范围等应基于场景按需动态调整，避免默认全

开放或一次性授予过高权限。宜制定内部白名单，如上架审核机制，并开展动态审核运维。

□**定期清理与复核**：按月或按季度开展权限复核，清理闲置账号、过期令牌、未使用插件，及时回收超范围权限。

（2）风险操作确认：针对高风险操作设置人工确认环节，做到授权过程可追溯、责任主体可明确、执行行为可核验。

□**风险分级定义**：形成清晰的高风险操作清单与分级标准，明确各等级对应的责任人以及授权有效期限。

□**关键操作确认**：对于重要文件和关键配置删改、资金转移、敏感数据外发等高风险操作，引入人工确认。

（3）应急响应预案：围绕智能体失控、数据泄露与异常操作等场景建立详细可执行的预案。

□**告警触发与分级处置**：明确异常调用、越权操作、外部连接异常、敏感信息外发等各类异常情形的告警触发阈值与判定标准，制定分级处置策略，明确各等级处置责任主体、响应时限。

□**标准化止损操作**：配置专属安全管理员，明确其权限与职责，以及停用智能体、禁用访问凭证、隔离执行节点等关键操作的步骤与优先级，确保突发情况下可快速启动止损流程、遏制风险扩散。

2、提供方

（1）版本更新管理：建立透明的版本更新与变更披露机制，在确保可控的前提下完成升级或回退，保障系统稳定性与可维护性。

□**更新信息透明**：应以变更日志、公告、API 版本号或通过管理后台通知等方式，向使用方清晰告知重大功能更新、安全补丁更新及兼容性变更。对于修复了高危或关键安全漏洞的更新，应提供简要的安全通告。

□**升级回滚可控**：提供灰度升级建议与兼容性验证方法，或者配套完善的回滚方案与验证流程。

□**生命周期管理**：明确各版本支持周期与停止维护时间，提前告知使用方迁移路径，并提供必要的过渡支持与工具。

（2）漏洞响应机制：建立从接收、评估到修复与验证的漏洞处置闭环，确保风险收敛路径明确、进度可跟踪、证据可留存。

□**通报渠道稳定**：建立稳定的漏洞报告渠道与通报机制。发现漏洞后及时向工业和信息化部网络安全威胁和漏洞信息共享平台人工智能产品安全漏洞专业库（简称 CAIVD）报送[1]。同步向使用方告知漏洞及处置进展。

□**分级时效承诺**：依据漏洞严重程度，明确承诺修复或缓解的时限；必要时提供可执行的临时缓解措施，直至漏洞彻底修复。

（3）技术支持能力：提供完善的运维支持与监控能力，确保使用方可快速获得专业支持，智能体运行状态可监控。

□**支持通道畅通**：建立标准化工单与应急响应通道，明确响应时间与升级路径，确保关键问题能够被快速接收与派单。

□**提供可观测与内容安全工具：**实现智能体行为记录与告警通知可视化，对智能体输入（用户提问/指令）与输出（生成内容/执行动作）提供

[1] 网址为 https://ai.nvdb.org.cn

安全监控与过滤机制，具备识别并拦截敏感信息泄露、恶意指令、不当生成内容等风险的能力，记录告警并通知使用方。

（四）下线阶段：处置清理与审计留存

智能体下线阶段是数据生命周期管理的收尾环节，需要确保数据安全、凭证可控、资源合规释放。

1、使用方

（1）数据导出备份：在下线前完成个人和企业所需数据的导出与备份，确保后续能够满足审计取证、迁移复用与复核核验等需求。

□**明确导出范围**：覆盖提示词模板、配置策略、会话记录、生成内容与关键操作记录等必要数据，并形成导出清单。

□**确保导出安全**：导出过程实施权限控制与加密传输，针对含敏感信息的数据进行脱敏处理，避免二次泄露。

□**校验数据可用**：提供可读格式与完整性校验方式，确保导出数据能够被还原、被检索并支撑追溯。

（2）访问凭证清理：在停用过程中撤销账号与令牌并关闭所有遗留入口，避免系统下线后仍然存在被访问或被滥用的风险。

□**识别与列举凭证**：全面梳理相关访问凭证清单，应包括智能体平台自身的 API 密钥、管理员与用户账号、所集成大模型服务的 API 密钥、第三方应用与工具的 OAuth 令牌及 API 密钥、各接入渠道的令牌或密钥，以及所有回调地址与网络白名单配置。

□**执行吊销与解除**：依据清单，在相关管理平台逐一执行吊销密钥、取消授权、解除集成绑定、移出白名单等操作，包括智能体服务商控制台、第三方应用授权页面、企业统一身份管理平台、个人登录平台等。

□**复核清理有效性**：对已清理的凭证进行有效性复核，如尝试使用已吊销的 API 密钥调用接口，验证其是否已被拒绝；检查已解除的集成是否仍能接收回调。确保所有访问路径均已失效。

（3）审计日志留存：按照合规要求保留全链路日志，确保智能体行为与人员操作均可被精准追溯，支持复盘、取证与内部审计。

□**日志结构化存储**：统一采用结构化格式记录日志，覆盖用户指令、会话记录、工具调用与系统操作等核心要素。

□**防篡改与合规留存**：实施日志保全，防止数据被篡改、删除或覆盖，关键日志留存期限应不少于 6 个月。

2、提供方

（1）服务停用清除：使用方决定停用后，提供方应完成对服务的不可逆下线与数据清除，确保无残留运行状态。

□**提供专用停用清除工具**：应提供自动化工具或操作说明文档，支持使用方自主执行或由提供方协助完成全量清除。

□**彻底清除服务与进程**：确保所有相关进程终止、无端口持续监听、无后台任务残留。

□**第三方授权撤销**：对于无法完全自动化清理的第三方授权，需在卸载时向用户提供明确的、逐步操作的指引文档，说明如何前往相关平台手动撤销应用授权。

（2）云端资源释放：在应用卸载的同时，同步释放并清理为该服务分配的所有云资源与网络配置（如有），确保不遗留可访问的端点、存储或计算资源，避免产生持续性费用与安全暴露面。

□**释放并确认资源删除**：删除为服务创建的所有专属资源，包括但不限于计算实例、容器服务、数据库实例、对象存储及相关快照/备份。

□**关闭网络与权限配置**：删除或禁用相关网络访问规则、API 路由及回调配置，同时删除服务运行时使用的所有角色、访问密钥及权限策略。

（3）审计配合支持：在服务停用及后续审计期间，应根据使用方或审计方要求，及时提供必要的证据材料与说明，协助完成合规验证。

□**提供审计材料**：准备数据销毁记录、资源释放记录、版本升级历史与重大事件处置记录等可核验材料，满足取证需求。

□**协助核验闭环**：支持审计问询与证据复核，直至形成可追溯的关闭确认结论并归档。

五、结语

OpenClaw 类智能体部署风险管理是一项系统工程，需要使用方与提供方共同参与、协同治理。本指南提供了全生命周期的风险管理框架，相关主体应根据自身实际情况选择适配的控制措施，在推进人工智能能力建设的同时，守住安全底线。

随着技术发展和风险演进，本指南将持续更新完善，欢迎各方提出宝贵意见。

附录 A：参考来源

[1] 工业和信息化部网络安全威胁和漏洞信息共享平台. 关于防范 OpenClaw（“龙虾”）开源智能体安全风险的“六要六不要”建议. 2026 年 3 月 11 日.

[2] 国家网络与信息安全信息通报中心. OpenClaw 安全风险预警通报. 2026 年 3 月 13 日.

[3] 中国人工智能产业发展联盟 AIIA. 关于防范 OpenClaw 开源 AI 智能体安全风险的提示. 2026 年 3 月 12 日.

附录 B：术语表

表 2：术语及说明

术语	说明
OpenClaw	开源的多渠道智能体网关与执行框架，支持大语言模型对接本地系统、外部工具及多平台通讯接口，实现任务规划、工具调用与自动化执行闭环。
智能体 (Agent)	由大语言模型驱动的软件实体，能够感知环境并在明确目标下进行任务分解与自主执行。
OpenClaw 类智能体	以 OpenClaw 框架为基础，具备多入口接入、多智能体编排、工具与插件扩展能力，并可在对话触发下执行系统级或业务级操作的智能体执行框架。
网关 (Gateway)	OpenClaw 核心组件，负责多渠道消息接入、路由、会话管理与权限控制。
节点 (Node)	可连接的执行环境或远程客户端设备，用于在隔离环境中执行特定任务或代理浏览器操作。
技能/插件 (Skill/Plugin)	可安装至智能体的第三方功能扩展模块，用于扩展外部系统调用或特定计算能力。
模型上下文协议 (Model Context Protocol, MCP)	一种开源标准化通信协议与接口，用于统一大模型与外部数据源、工具及服务之间的交互方式，实现无缝、可扩展的连接与集成。
智能体到智能体 (Agent-to-Agent, A2A)	一种开放的互操作性标准协议，用于使不同平台、不同厂商、不同基础设施上构建的智能体能够相互发现、委托任务并交换结果
提示注入 (Prompt Injection)	攻击者通过恶意输入诱导模型忽略既有安全约束并执行越权操作的行为，包含直接与间接注入。
级联失控	在多智能体协同或跨系统调用中，单一节点失陷被用作跳板，导致异常行为逐级放大的链式安全事件。
通用漏洞与暴露 (CVE)	全球公认的安全漏洞标准化编号体系，用于组件与容器镜像的常态化安全扫描核验。
最小权限原则 (Least Privilege)	系统或用户仅被授予完成当前任务所必需的最低访问权限与动作范围，以最大限度收敛攻击面。
CAIVD	工业和信息化部网络安全威胁和漏洞信息共享平台人工智能产品安全漏洞专业库。
可观测 (Observability)	通过日志采集、行为追踪、状态采集与可视化等方式，呈现智能体全生命周期运行行为与内部状态。

附录 C：自查清单表

本自查清单由使用方与提供方按责任分工，各自对照开展自查。

表 3：OpenClaw 类智能体使用方、提供方自查清单表

阶段	角色	类别	自查项	自查要点
部署	使用方	场景风险评估	□场景风险分级	明确智能体在业务中的使用场景（内部辅助、对外服务、业务决策等），分析决策影响程度、错误容忍度及自动化执行程度，评估整体风险
			□交互风险评估	分析网络边界、系统边界、数据边界、接口边界、工具边界等各维度交互点，开展风险评估
			□合规要求对齐	依据具体使用场景，如是否涉及金融、医疗、教育等领域，识别需遵循的法律法规及行业标准，确保符合法律监管与安全标准
		部署形态选型	□架构选型适配	依据业务场景、数据安全等级与运维能力，在本地私有化部署、云端托管部署、端云协同部署三类模式中选择适配架构
			□明确数据管控	确定数据是否允许出域，明确是否与内网、专网隔离，严格限制外部模型服务与第三方能力的接入范围
		权限配置核验	□权限配置核验	对照场景风险评估结果，核验网络访问、系统接入、数据读写、接口调用等权限配置，确保权限范围与业务需求匹配，无超范围授权
			□白名单机制核验	验证工具、接口、插件等白名单配置的有效性，确认高风险工具默认禁用、高权限操作需人工确认，符合安全管控要求
			□身份凭证核验	核查账号、密钥、令牌等身份凭证的管理机制，确认无共享账号、明文存储等违规配置，凭证存储、轮换流程合规
	提供方	部署方案设计	□信息告知披露	通过用户协议、隐私政策、安全与风险使用指南或相应合同约定，以显著方式、清晰易懂的语言，真实、准确、完整地向使用方告知个人信息处理事项、系统关键配置、潜在风险分级及双方权责
			□部署形态设定	结合使用方业务场景与需求，确定本地、云端或混合部署形态，根据实际需要将容器、沙盒等虚拟化隔离措施作为关键安全配置纳入部署方案
			□模型可信评估	核验基础模型提供商资质、版本标识及服务形态（直供或代理中转），评估模型安全机制、内容安全机制及供应链可追溯性；对微调或蒸馏版本同步开展安全评估；建立模型版本管理制度，支持异常情况下快速回滚；部署输入安全过滤与输出安全审核机制
			□信源可信评估	智能体配置的默认联网搜索信源或知识库，应优

阶段	角色	类别	自查项	自查要点
				先采用可信、权威的来源；应公布主要信源列表或选择标准；评估第三方接口服务商的安全合规水平
			□信任边界划分	当多用户共用同一网关渠道时，应识别其隐含的共享工具授权风险，按信任等级拆分网关实例、访问凭证及执行环境，避免高敏感操作与低敏感场景混用同一权限上下文
		权限边界设定	□网络访问权限管控	明确网络通信边界与访问规则，限定智能体可访问的网络域、IP 段与端口，实施内外网隔离与访问控制
			□本地业务系统接入管控	界定智能体与本地业务系统的接入边界，严格控制可访问系统范围与操作深度，按场景分级授权，禁止未经许可的系统访问与跨域操作
			□数据读写管控	识别数据类型与敏感级别，明确数据在收集、存储、使用、加工、传输、提供、公开等数据处理各环节的去向与存储位置；界定智能体对知识库、数据库表、工作区及文件目录的读写、修改与删除权限，实施最小化访问与数据隔离
			□外部接口调用管控	梳理外部调用接口清单，建立接口白名单机制，限定调用频率、传输内容与授权范围
			□扩展插件与工具管控	采用白名单策略，对插件、技能及第三方工具实行动态准入审核，明确可加载范围、执行环境与执行权限，禁止高风险工具默认启用，高风险工具执行操作须经人工确认
			□身份账号与密钥凭证管控	实行账号管理与身份认证，严禁共享账号；统一管理密钥与凭证，规范存储、使用与轮换机制
			□记忆安全管理	界定智能体长期记忆存储边界，实施敏感信息识别与脱敏；建立记忆生命周期管理机制，含访问控制、定期清理与过期淘汰，防止历史会话数据泄露
		关键参数配置	□访问与鉴权设置	配置访问控制与鉴权策略，收敛网络入口，对使用方实施身份校验；配置特定安全策略防范非法接入与指令注入风险
			□人工确认与风险控制	按业务敏感等级设置人工确认环节，对数据改写、脚本执行等高风险操作构建动态防护；限定智能体授权运行边界。设置多智能体协同安全策略，防范级联失控
			□沙箱与边界防护	通过虚拟环境隔离、目录权限约束等方式限制智能体交互范围，禁止越权访问敏感配置；启用网络访问管控，拦截内网敏感目标探测行为，避免敏感数据泄露。对提供的服务以及运行环境进行必要的安全检测与防护。

阶段	角色	类别	自查项	自查要点
			□审计与溯源配置	配置全链路日志审计策略，对服务运行中的相关操作与处理行为进行记录；规范日志存储周期与加密权限，实现高风险操作流程可追溯、不可篡改
使用	使用方	权限动态管控	□白名单动态运维	工具、技能、插件、接口、协议（如 MCP、A2A）、文件目录、数据、网络访问范围等应基于场景按需动态调整，避免默认全开放或一次性授予过高权限；制定内部白名单，上架审核机制，并动态审核运维
			□定期清理与复核	按月或按季度开展权限复核，清理闲置账号、过期令牌、未使用插件，及时回收超范围权限
		风险操作确认	□风险分级定义	形成清晰的高风险操作清单与分级标准，明确各等级对应的责任人以及授权有效期限
			□关键操作确认	对于重要文件和关键配置删改、资金转移、敏感数据外发等高风险操作，引入人工确认，做到授权过程可追溯、责任主体可明确、执行行为可核验
		应急响应预案	□告警触发与分级处置	明确异常调用、越权操作、外部连接异常、敏感信息外发等各类异常情形的告警触发阈值与判定标准，制定分级处置策略，明确各等级处置责任主体、响应时限
			□标准化止损操作	配置专属安全管理员，明确其权限与职责，以及停用智能体、禁用访问凭证、隔离执行节点等关键操作的步骤与优先级，确保突发情况下可快速启动止损流程
	提供方	版本更新管理	□更新信息透明	应以变更日志、公告、API 版本号或通过管理后台通知等方式，向使用方清晰告知重大功能更新、安全补丁更新及兼容性变更；对于修复了高危或关键安全漏洞的更新，应提供简要的安全通告
			□升级回滚可控	提供灰度升级建议与兼容性验证方法，或者配套完善的回滚方案与验证流程
			□生命周期管理	明确各版本支持周期与停止维护时间，提前告知使用方迁移路径，并提供必要的过渡支持与工具
		漏洞响应机制	□通报渠道稳定	建立稳定的漏洞报告渠道与通报机制；发现漏洞后及时向工业和信息化部网络安全威胁和漏洞信息共享平台人工智能产品安全漏洞专业库报送，同步向使用方告知漏洞及处置进展
			□分级时效承诺	依据漏洞严重程度，明确承诺修复或缓解的时限；必要时提供可执行的临时缓解措施，直至漏洞彻底修复
		技术	□支持通道畅通	建立标准化工单与应急响应通道，明确响应时间

阶段	角色	类别	自查项	自查要点
		支持能力		与升级路径，确保关键问题能够被快速接收与派单
			□提供可观测与内容安全工具	实现智能体行为记录与告警通知可视化，对智能体输入（用户提问/指令）与输出（生成内容/执行动作）提供安全监控与过滤机制，具备识别并拦截敏感信息泄露、恶意指令、不当生成内容等风险的能力，记录告警并通知使用方
下线	使用方	数据导出备份	□明确导出范围	覆盖提示词模板、配置策略、会话记录、生成内容与关键操作记录等必要数据，并形成导出清单
			□确保导出安全	导出过程实施权限控制与加密传输，针对含敏感信息的数据进行脱敏处理，避免二次泄露
			□校验数据可用	提供可读格式与完整性校验方式，确保导出数据能够被还原、被检索并支撑追溯
		访问凭证清理	□识别与列举凭证	全面梳理相关访问凭证清单，应包括智能体平台自身的 API 密钥、管理员与用户账号、所集成大模型服务的 API 密钥、第三方应用与工具的 OAuth 令牌及 API 密钥、各接入渠道的令牌或密钥，以及所有回调地址与网络白名单配置
			□执行吊销与解除	依据清单，在相关管理平台逐一执行吊销密钥、取消授权、解除集成绑定、移出白名单等操作，包括智能体服务商控制台、第三方应用授权页面、企业统一身份管理平台、个人登录平台等
			□复核清理有效性	对已清理的凭证进行有效性复核，如尝试使用已吊销的 API 密钥调用接口，验证其是否已被拒绝；检查已解除的集成是否仍能接收回调，确保所有访问路径均已失效
		审计日志留存	□日志结构化存储	统一采用结构化格式记录日志，覆盖用户指令、会话记录、工具调用与系统操作等核心要素
			□防篡改与合规留存	实施日志保全，防止数据被篡改、删除或覆盖，关键日志留存期限应不少于 6 个月
	提供方	服务停用清除	□提供专用停用清除工具	应提供自动化工具或操作说明文档，支持使用方自主执行或由提供方协助完成全量清除
			□彻底清除服务与进程	确保所有相关进程终止、无端口持续监听、无后台任务残留
			□第三方授权撤销	对于无法完全自动化清理的第三方授权，需在卸载时向用户提供明确的、逐步操作的指引文档，说明如何前往相关平台手动撤销应用授权
		云端资源释放	□释放并确认资源删除	删除为服务创建的所有专属资源，包括但不限于计算实例、容器服务、数据库实例、对象存储及相关快照/备份
			□关闭网络与权限配置	删除或禁用为该服务配置的所有网络访问规则、API 路由及回调配置，同时删除服务运行时使用

阶段	角色	类别	自查项	自查要点
				的所有角色、访问密钥及相关权限策略
		审计配合支持	□提供审计材料	准备数据销毁记录、资源释放记录、版本升级历史与重大事件处置记录等可核验材料，满足取证需求
			□协助核验闭环	支持审计问询与证据复核，直至形成可追溯的关闭确认结论并归档