文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

OpenClaw 智能代理框架安全问题分析与防御策略研究

张亘政　魏亚川
华电青岛发电有限公司
DOI:10.32629/acair.v4i2.20409

[摘　要] 在大语言模型技术持续迭代的当下,具备自主决策、工具调用与任务闭环能力的AI智能代理,已成为人工智能领域重要的发展方向。OpenClaw属于开源通用型智能代理开发框架,凭借部署便捷、扩展性强、支持多类工具联动开发等优势,广泛应用于个人办公自动化、企业业务流程调度、网络运维等各类场景。该框架在功能层面迭代速度较快,但开发过程中忽视了安全体系搭建,原生安全防护能力薄弱,各类安全漏洞与潜在威胁不断暴露。本文结合当下AI Agent安全领域前沿研究成果,以CIK三维分析模型为核心研究手段,从能力、身份、数据知识三个维度,系统梳理OpenClaw运行过程中存在的安全隐患,量化各类攻击方式的危害程度,深度剖析风险形成的内在原因。同时结合实际应用场景,搭建适配该框架的多层次防御体系,提出权限管控、安全审计、分级部署等可落地的优化方案。研究证实,OpenClaw安全短板主要来源于权限配置不合理与输入校验机制缺失,只有从架构、权限、运维多维度协同防护,才能平衡智能化体验与安全可靠性,也为同类开源智能代理框架的安全建设提供参考思路。
[关键词] OpenClaw；AI智能代理；网络安全；数据投毒；权限管控；安全防御
中图分类号：G250.72　文献标识码：A

Analysis of Security Issues in the OpenClaw Smart Proxy Framework and Research into Defence Strategies

Genzheng Zhang　Yachuan Wei
Huadian Qingdao Power Generation Co., Ltd.

[Abstract] With the ongoing evolution of large language model technology, AI agents capable of autonomous decision-making, tool invocation and task completion have become a key area of development in the field of artificial intelligence. OpenClaw is an open-source, general-purpose AI agent development framework. Thanks to its ease of deployment, strong extensibility, and support for the integrated development of multiple tools, it is widely used in various scenarios such as personal office automation, enterprise business process scheduling, and network operations and maintenance. Whilst the framework evolves rapidly in terms of functionality, the development process has overlooked the establishment of a security framework, resulting in weak native security defences and the continuous exposure of various security vulnerabilities and potential threats. This paper integrates cutting-edge research findings in the field of AI agent security, utilising the CIK three-dimensional analysis model as its core research methodology. It systematically identifies security risks present during OpenClaw's operation across three dimensions—capabilities, identity, and data knowledge—quantifies the severity of various attack vectors, and conducts an in-depth analysis of the underlying causes of these risks. Furthermore, by incorporating real-world application scenarios, the paper establishes a multi-layered defence system tailored to this framework and proposes actionable optimisation strategies, including permission control, security auditing, and tiered deployment. The research confirms that OpenClaw's security weaknesses primarily stem from unreasonable permission configurations and the absence of input validation mechanisms. Only through coordinated protection across multiple dimensions—including architecture, permissions, and operations—can a balance be struck between intelligent user experience and security reliability. This study also provides a reference framework for the security development of similar open-source intelligent agent frameworks.
[Key words] OpenClaw; AI-powered agents; cybersecurity; data poisoning; access control; security defences

文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

1 引言

1.1研究背景

人工智能产业的快速升级,让大语言模型不再局限于基础对话交互,逐步向自主智能执行方向转型。AI Agent能够脱离高频人工干预,自主拆解复杂任务、规划执行步骤、调用外部工具完成目标,大幅提升各行业自动化水平。

开源框架的普及降低了智能代理的开发门槛,其中OpenClaw凭借轻量化部署、跨平台适配、插件生态丰富等特点,受到开发者与中小企业的广泛青睐。该框架可实现文件管理、系统指令调用、网络接口访问、自动化数据处理等多元化功能,落地场景愈发广泛。

现阶段行业普遍存在重功能开发、轻安全防护的现象,OpenClaw在初始架构设计中优先保障使用便捷性与任务执行效率,并未建立完善的访问控制、数据校验、行为审计等安全机制。随着大量实例暴露在公网与企业内网环境中,数据投毒、提示词注入、权限越权、恶意插件供应链攻击等安全事件频发,一旦被不法分子利用,极易引发核心数据泄露、系统被劫持、业务瘫痪等严重后果。在此背景下,针对OpenClaw开展系统性安全研究,具有极强的现实应用价值。

1.2研究意义

从理论层面来看,当前国内针对OpenClaw专项安全研究内容较少,现有资料大多局限于零散漏洞描述,缺乏体系化的学术分析与风险归纳。本文引入CIK分析模型完成全维度风险拆解,完善了开源AI智能代理的安全研究体系,填补了相关领域理论研究的空白。

从实践层面来讲,本文提出的防御方案兼顾实用性与可落地性,既适配个人轻量化使用场景,也能够满足企业核心业务的安全防护需求。相关加固思路不仅适用于OpenClaw,对AutoGPT等同类型智能代理框架的安全部署、运维管控也具备借鉴意义,能够有效帮助开发者规避安全风险,推动AI智能代理技术规范化应用。

1.3论文整体架构

本文一共分为五个核心部分：第一部分为引言,阐述研究背景、意义与整体行文结构；第二部分介绍OpenClaw技术特征与CIK安全分析模型的核心原理；第三部分依托分析模型,深度挖掘框架现存各类安全风险并溯源成因；第四部分结合风险痛点,构建全方位防御体系与优化策略；第五部分总结全文研究成果,并对未来AI智能代理安全发展方向做出展望。

2 OpenClaw框架与CIK安全分析模型基础理论

2.1 OpenClaw框架核心技术特征

OpenClaw是面向大模型应用打造的开源自主智能体开发框架,核心优势集中在三个方面。

首先,工具联动能力突出,支持系统命令、代码运行、数据库访问、第三方API对接等各类外部工具集成,可实现多工具链式调度,完成高复杂度的复合型任务。

其次,部署适配性高,开源免费且配置简单,能够兼容本地终端、云端服务器、公网服务等多种运行环境,适配不同用户的开发与使用需求。

最后,自主智能化程度高,无需人工分步指令输入,可根据既定目标自主规划流程、动态调整执行逻辑,实现全流程自动化作业。

也正是因为过度追求易用性与智能化,框架牺牲了安全设计,为后续各类网络攻击与恶意操作留下了可乘之机。

2.2 CIK三维安全分析模型

为实现对OpenClaw风险的全面、客观评估,本文采用迁移优化后的CIK三维安全分析模型,从能力、身份、知识三大维度开展系统性研判：

(1)能力维度：主要指代智能代理拥有的操作权限与资源调用范围,包含文件读写、系统指令执行、内网访问等核心能力,核心风险集中在权限冗余、边界失控；(2)身份维度：聚焦运行会话、用户认证、访问隔离等机制,重点考量不同实例、不同用户之间的权限划分与身份校验能力；(3)知识维度：涵盖智能体接收的外部输入、上下文记忆、交互数据等核心信息,主要风险体现为数据篡改、恶意投毒、上下文诱导。

该模型能够穿透表层功能表现,从架构底层定位安全缺陷,保障风险分析全面无死角。

3 OpenClaw框架核心安全风险深度剖析

3.1知识维度风险：上下文篡改与数据投毒攻击

OpenClaw的决策逻辑高度依赖外部输入与历史上下文数据,框架本身未设置严格的数据清洗与合法性校验流程,这也让数据投毒成为最易实施的攻击手段。

攻击者通过构造隐藏恶意指令的诱导数据,篡改智能代理的判断依据,进而操控其执行非授权操作。相关实验数据表明,在框架原生无防护状态下,恶意指令自然触发概率仅为24.6%；而经过单维度数据投毒改造后,攻击成功概率直接攀升至64%~74%。若叠加多维度联合投毒方式,攻击者可完全掌控智能体的任务执行逻辑,隐蔽性强且普通用户难以察觉。与此同时,插件导入、网络请求接入的外部数据缺乏溯源校验,极易形成长期隐性安全隐患。

3.2能力维度风险：权限冗余越权与工具滥用

权限管理混乱是OpenClaw最核心的结构性缺陷,框架默认配置遵循高权限开放原则,未践行网络安全领域通用的最小权限原则。

为保证任务执行顺畅,智能代理被默认赋予全盘文件读写、系统命令执行、内网资源漫游、跨接口调用等高风险权限。攻击者只需借助简单的提示词注入,即可突破原有操作边界,利用冗余权限实现远程代码执行、敏感文件窃取、内网横向渗透等高危行为。一旦攻击得逞,不仅会造成本地数据丢失、泄露,还会波及关联业务系统,引发规模化安全事故。

3.3原生防御缺陷：防护能力薄弱且存在可用性矛盾

OpenClaw并未内置独立的安全防护模块,抵御恶意攻击完全依赖底层大语言模型自身的内容识别能力,防护可靠性极低。

实测数据显示,即便搭载当前行业主流的轻量化防护手段,恶意操作依旧有着63.8%的触发成功率,面对多轮隐性诱导、复杂混淆指令等高级攻击几乎没有抵御能力。

此外,该框架长期存在安全防护与业务可用性相互博弈的痛点。如果采用高强度文件隔离、全量输入过滤、严格权限封禁等防护方式,虽然可以拦截97%以上的恶意注入和越权操作,但会严重破坏正常的工具调用逻辑,大幅降低自动化作业效率;若保留原有便捷性,又会持续暴露高危漏洞,二者难以平衡。

3.4身份与生态风险：会话隔离缺失与供应链攻击

在身份会话管理层面,OpenClaw缺乏完善的用户认证与实例隔离机制,不同任务、不同用户的运行会话容易出现数据交叉污染,未授权人员可绕过身份校验访问历史数据与操作记录。

在插件生态层面,框架高度依赖第三方开发的技能拓展包,相关插件缺乏统一安全审核机制。现有公开数据显示,生态内近五分之一的第三方插件存在漏洞或恶意后门,用户盲目安装后会直接引入安全隐患,形成典型的供应链攻击,威胁整个运行环境的安全稳定。

4 OpenClaw安全防御体系构建与优化方案

4.1搭建CIK三维动态安全审计机制

结合前文风险分析结果,建立全流程动态审计体系,从源头约束不安全行为：

在能力审计上,实时监控智能代理的工具调用、系统操作、资源访问行为,建立高危操作黑名单,自动拦截违规指令与越权访问;

在身份审计上,引入角色权限管控模式,为不同运行实例、不同使用者分配独立身份标识,完善会话隔离与访问溯源记录;

在知识审计上,增设独立的数据清洗与语义检测模块,甄别剔除恶意投毒数据、篡改指令与隐性诱导内容,保障决策数据安全可信。

4.2落实精细化最小权限管控策略

彻底摒弃默认高权限的配置模式,对所有操作权限进行拆分分级。将框架工具划分为普通查询、常规文件操作、高危系统指令、内网资源访问等不同等级,默认关闭高风险操作权限,仅在人工授权后临时开放。

建立文件与网络访问白名单机制,限定智能代理仅能访问指定目录文件、可信域名接口,从物理边界切断数据泄露与内网渗透路径。同时完善权限操作日志留存,对高危行为全程记录,方便后期风险追溯与安全排查。

4.3打造多层级联动恶意攻击防御体系

脱离对大模型自身防护能力的依赖,搭建独立的前置防御架构。一方面增设专用输入过滤层,通过特征匹配+语义识别双重机制,拦截提示词注入、特殊字符混淆、恶意指令拼接等常见攻击;另一方面引入人在回路审核机制,凡是涉及系统修改、批量文件删除、核心数据外传等高风险操作,必须经过人工确认后方可执行。

同时推广沙箱隔离部署方案,将智能代理运行环境与宿主机核心资源物理隔离,即便出现恶意代码执行行为,也能限制其破坏范围,避免整机或内网沦陷。

4.4推行差异化分级部署安全规范

针对不同应用场景制定适配的防护策略,平衡安全性能与使用体验：

个人日常轻量化部署场景,采用基础输入过滤+普通权限限制模式,配置简单且不影响使用效率,满足基础安全需求;

企业核心业务部署场景,强制启用沙箱隔离、全量行为审计、人工审核三重防护,严格禁用高危系统指令,保障业务数据绝对安全;

公网公开部署场景,叠加防火墙、应用防护网关,收缩对外开放端口,减少外部攻击者的攻击面,降低被探测与入侵的概率。

5 结论与展望

5.1研究结论

通过对OpenClaw智能代理框架的全方位安全研究可以发现,该框架安全问题的核心根源,是高权限默认配置与原生安全校验机制匮乏的结构性矛盾。数据投毒、权限越权、提示词注入、恶意插件供应链攻击等风险普遍存在,现有轻量化防御手段抵御能力不足,无法满足常态化安全运行要求。

本文基于CIK三维分析模型,完成了风险溯源与危害量化,提出的审计机制搭建、最小权限管控、多层级防御、分级部署优化等方案,能够从架构到运维形成纵深防护体系,在保障OpenClaw智能化、自动化使用价值的前提下,显著降低被攻击风险,有效化解安全防护与业务可用性之间的矛盾。

5.2未来展望

AI智能代理属于人工智能未来核心发展方向,随着应用场景不断拓宽,其安全攻防对抗也会更加复杂。后续研究可以聚焦三个方向:一是研发自适应智能防御模型,依靠机器学习自主识别新型未知攻击,摆脱传统规则防护的局限性;二是推动框架源码层面的安全重构,将安全机制内嵌于底层架构,从源头消除设计缺陷;三是建立开源AI Agent统一安全规范与审核标准,规范插件生态与开发流程,推动整个行业安全、健康、可持续发展。

[参考文献]

[1]安全研究团队.Your Agent,Their Asset:A Real-World Safety Analysis of OpenClaw[J].arXiv preprint arXiv:2604.04759,2026.

[2]张宇,李萌.开源AIAgent框架安全隐患与防御技术研究[J].计算机工程与应用,2026.

[3]网络安全产业联盟.AI智能代理框架安全风险评估指南[R].2026.

[4]王浩,陈佳.大语言模型提示词注入攻击与防御技术综述[J].信息安全学报,2025.

作者简介：

张亘政(1982--),男,汉族,山东省青岛市人,本科,网络工程师,研究方向：网络安全。

魏亚川(1999--),女,汉族,浙江省宁波市人,硕士,助理工程师,研究方向：网络安全。