Research on Personal Information Security Risks and Systematic Protection in Generative AI Applications

Shuang Liu

Sichuan University of Architectural Technology, Deyang, Sichuan, 618000, China

Abstract

Generative artificial intelligence technologies centered on large language models are rapidly expanding their application boundaries from traditional text interaction to system-level automated execution scenarios. The phenomenal popularity of executive AI frameworks represented by OpenClaw (commonly known as the "Lobster Agent") has brought about unprecedented challenges to personal information security while unleashing productivity dividends. Based on the new characteristics of generative AI development, this paper deconstructs the core risk causes and transmission chains of individual users in the use of large models and agents, analyzes the nature of risks with the Lobster Agent as a typical case, and constructs a systematic protection path for personal information security from multiple dimensions.

Keywords

generative artificial intelligence; large language model; executive agent; personal information security; data protection

生成式AI应用中个人信息安全风险与体系化防护研究

刘爽

四川建筑职业技术大学，中国·四川 德阳 618000

摘要

以大语言模型为核心的生成式人工智能技术其应用边界从传统文本交互正在快速延伸至系统级自动化执行场景。以OpenClaw（俗称“龙虾智能体”）为代表的执行型AI框架现象级爆火，在释放生产力红利的同时，也让个人信息安全面临前所未有的挑战。本文基于生成式AI技术发展的新特征，解构个人用户在大模型与智能体使用中的核心风险成因与传导链条，以龙虾智能体为典型案例剖析风险本质，从多方面构建个人信息安全的体系化防护路径。

关键词

生成式人工智能；大语言模型；执行型智能体；个人信息安全；数据防护

1 引言

AI技术的普惠化已深度融入办公、学习、生活、生产等全场景，也让个人信息安全进入了全新的风险周期。与传统互联网时代黑客攻击、平台违规收集等被动泄露模式不同，当前个人信息安全的核心风险源，更多来自用户自身的主动行为。龙虾智能体的爆火，更是将这种风险推向了极致——据CNCERT2026年3月发布的风险提示显示，全球已有超23万龙虾智能体实例直接暴露在公网，大量用户因不当配置面临数据泄露、系统被控、密钥被盗的极高风险。

个人用户作为AI工具最广泛的使用主体，其信息安全不仅关乎个人隐私与财产安全，更是国家网络安全体系的基础组成部分。基于此，本文结合当前AI技术发展的最新趋势，解构个人用户在AI使用中的核心安全风险与成因，以龙虾智能体为典型案例分析风险传导链条，并提出可落地的个人信息安全防护体系，为个人用户安全使用AI工具提供实践指引。

2 生成式 AI 时代个人信息安全风险的全新演进特征

生成式 AI 的技术特性与应用模式变革，彻底改变了个人信息安全的风险形态、传导路径与发生逻辑，呈现出与传统互联网时代截然不同的三大核心特征。

【作者简介】刘爽（1990–），男，中国四川德阳人，本科，工程师，从事信息通信工程、智能制造研究。

2.1 风险源：从被动窃取到主动投喂的核心转向

传统互联网时代，个人信息泄露的核心诱因是外部攻击与平台违规收集，用户处于被动受害的地位。而在生成式AI场景中，用户主动向大模型提供隐私数据，已成为个人信息泄露最核心的风险源。为获得更精准、更贴合需求的AI输出结果，用户往往会在对话中无保留地提供完整信息：为撰写劳动仲裁申请，上传完整的工资流水、身份证号、劳动合同；为完成财务工作，提供企业营收数据、客户信息、商业合同；为解决生活问题，披露家庭情况、健康信息、行程轨迹。

大模型的上下文记忆与多轮对话机制，会让用户的隐私信息不断累积，形成维度完整、细节丰富的个人信息画像。一旦AI平台发生数据泄露、模型被恶意攻击，用户的全量隐私信息将一次性完全暴露，其危害程度远超传统互联网时代的单点信息泄露。

2.2 风险载体：从文本交互到全系统权限的渗透扩张

传统对话型大模型的风险边界，仅局限于用户输入的文本对话内容，无法触碰用户的本地设备、系统资源与线下场景，风险具有明确的隔离性。而以龙虾智能体为代表的执行型AI框架，彻底打破了这一边界，实现了从“文本交互”到“系统操控”的权限跨越。

这类执行型AI工具的核心价值，在于打通大模型与本地设备的执行通道，默认需要获取用户设备的全盘文件读写、鼠标键盘模拟、浏览器全程接管、命令行脚本执行、第三方API调用等系统级最高权限。这种权限扩张，让个人信息安全风险从单一的文本信息泄露，升级为全设备、全场景的系统安全风险：AI不仅能获取用户主动输入的内容，还能读取设备内的所有文件、实时监控屏幕显示内容、操控设备执行任何系统操作，一旦出现权限失控，用户的设备将完全被攻击者掌控，造成不可逆的损失。

2.3 风险认知：从技术抵触到盲目信任的深层偏差

用户的认知偏差，是当前个人信息安全最核心的隐性风险。在传统互联网时代，用户对安全软件的全盘扫盘、平台的个人信息收集行为高度敏感、极度抵触，核心原因是用户认为这种行为是“被动强加的、非必要的、带有商业目性的”。但在AI场景中，用户对AI工具索要的更高权限、更多隐私信息，却表现出极高的包容度，甚至主动放弃安全边界。

这种认知偏差的核心，是用户对AI技术的盲目崇拜与信任，将AI工具视为“高效的、无恶意的”，忽视了其背后的技术漏洞、生态风险、数据安全隐患。用户普遍存在“本地部署 = 绝对安全”、“开源 = 无风险” 的认知误区，甚至对监管部门发布的高危预警视而不见。这种认知上的盲区，让用户主动拆除了个人信息安全的第一道防线。

3 执行型 AI 的风险解构：以 OpenClaw（龙虾智能体）为典型案例

龙虾智能体的爆火，是生成式AI从“对话工具”向“执行工具”转型的标志性事件，其暴露的安全风险，正是当前个人用户AI使用中所有问题的集中体现。深入解构其技术逻辑、风险链条与用户认知误区，能清晰揭示生成式AI时代个人信息安全的核心矛盾。

3.1 技术逻辑与用户核心认知误区

龙虾智能体并非通用大语言模型，而是一款开源的、本地优先的执行型 AI 智能体框架，其核心定位是连接大语言模型与本地设备的执行桥梁。其本身不具备思考与决策能力，必须外接云端或本地大模型作为“大脑”才能运行。

在龙虾智能体的普及过程中，用户存在两个根本性的认知误区，这也是风险爆发的核心根源。一是将“本地部署”等同于“绝对安全”，多数用户认为只要将框架安装在本地电脑，就不会出现数据泄露。但实际上，框架本身仅负责执行，决策完全依赖外接大模型。框架在执行任务时会将用户的屏幕截图、文件内容、系统信息等全量数据打包发送至外接的大语言模型解析，用户的隐私数据完全处于失控状态。二是将“开源”等同于“无风险”，开源仅意味着代码可审计，不代表用户会完成审计，大量用户使用的第三方汉化版、破解版、插件合集包，往往被植入恶意代码。

3.2 闭环风险链条的深度解析

结合CNCERT 发布的权威预警，龙虾智能体的安全风险形成了完整的闭环，覆盖权限、生态、数据、操作四个核心维度，对个人信息安全形成全方位威胁。

第一，系统权限失控风险。龙虾智能体的核心功能是实现电脑系统的自动化操作，默认需要获取极高的系统权限，若用户使用管理员权限部署，框架将获得设备的完全控制权。一旦出现提示词注入、模型被劫持、插件投毒等情况，攻击者可直接通过框架获取系统最高权限，实现远程控制、数据窃取、文件加密勒索等恶意操作。

第二，插件生态投毒风险。龙虾智能体的核心竞争力在于其开源社区提供的超 5700个功能插件，覆盖文件处理、浏览器自动化、邮件收发等各类场景。但这些插件无严格审核机制，据CNCERT监测数据显示，其官方插件平台中约10.8%的插件含有恶意代码，17.7%的插件会从不可信第三方服务器获取内容。这些恶意插件可在用户不知情的情况下，窃取API密钥、上传全盘文件、植入木马程序。

第三，数据跨域泄露风险。对于绝大多数普通用户而言，龙虾智能体的默认配置是外接云端大模型，这意味着用户设备中的所有数据，都可能在执行任务的过程中被传输至云端，隐私数据完全脱离管控。

第四，不可逆操作风险。龙虾智能体属于全自动闭环执行架构，主要风险源于复杂场景下的链式偏差：一是多轮任务叠加、上下文冗余干扰，模型在连续执行复杂流程时，容易放大边界判定误差；二是插件脚本自主兜底纠错逻辑激进，部分自动化插件会默认扩大执行范围以“完成任务闭环”。

3.3 爆火背后的盲目信任危机

龙虾智能体的现象级爆火，本质上是个人用户对AI技术盲目信任的集中体现。在相关社区讨论中，绝大多数用户关注的是其功能、效率、插件拓展，极少有人关注背后的安全风险；即使监管部门连续发布高危预警，依然有大量用户按照网上的“新手教程”，将实例直接暴露在公网、安装第三方插件合集、使用管理员权限部署、外接云端大模型处理敏感数据。

这种盲目信任主要源于三个方面：一是对技术的过度崇拜，单方面将AI工具定性为“中立无恶意性”的；二是对效率的极致追求，主动放弃安全边界；三是普遍存在侥幸心理，低估风险发生的概率与危害。

4 生成式 AI 时代个人信息安全的体系化防护路径

针对当前AI技术发展的新特征与个人用户的核心风险，本文从认知纠偏、行为规范、技术防护、应急处置四个维度，构建个人信息安全的体系化防护路径。

4.1 认知纠偏：破除技术崇拜，树立 “最小必要” 核心原则

认知纠偏是个人信息安全防护的基础，只有纠正了对 AI 技术的盲目信任，才能从根源上规避风险。首先，个人用户必须树立“AI工具非中立”的核心认知，明确 AI工具本质上是一款软件产品，同样存在技术漏洞与安全隐患。其次，要彻底破除“本地部署=绝对安全”、“开源=无风险”的认知误区。明确安全的核心，不在于工具部署的位置，而在于数据的流向与权限的管控。

要树立“最小必要”的黄金安全法则，这也是《个人信息保护法》确立的核心原则。无论是向AI工具提供信息，还是授予系统权限，都只提供完成当前任务所必须的、最小范围的信息与权限，绝不提供额外的隐私内容，绝不授予超出任务范围的权限。

4.2 行为规范：构建全流程隐私管控体系

个人用户与AI工具的交互过程，是隐私数据泄露的核心触发点，需搭建覆盖事前、事中、事后的全链条行为管控框架，从操作源头阻断隐私信息外泄的传导路径。

事前管控的核心是数据脱敏。所有向AI工具提交的信息，均需完成前置脱敏处理：身份类信息仅提供任务必需的片段内容；财务类信息屏蔽关键金额、账户编号等核心要素；工作类信息剔除涉密内容与内部敏感数据；严禁向AI工具提供完整银行卡号、支付密码、身份证正反面影像、经营资质证照等核心敏感信息。

事中管控的核心是对话边界管控。首先，需对AI工具的使用场景做严格划分，严禁使用公共开放的AI服务处理涉密信息、非公开数据等敏感内容；其次，需避免在单轮多回合对话中持续累积隐私信息，不同场景、类型的任务需单独创建对话窗口，防止多维度的隐私信息在同一对话中汇聚整合，从而形成完整的个人或主体信息画像；最后，需主动关闭大模型的“对话数据用于模型训练” 等相关功能，从源头规避隐私数据被违规使用的风险。

事后管控的核心是定期清理与审查。定期清理AI平台的对话历史，删除包含隐私信息的对话内容；定期审查自己的API密钥使用情况，注销不再使用的密钥；定期查看AI工具的操作日志，排查异常数据传输行为。

4.3 技术防护：分层隔离部署，降低风险暴露面

针对不同类型的AI工具，要采用差异化的技术防护手段。对于普通对话大模型，优先选择具备端侧推理、隐私计算能力的合规平台，实现数据仅在用户本地设备上处理；必须使用云端大模型的，要选择符合《生成式人工智能服务管理暂行办法》要求、具备合规资质的国内平台。

对于以龙虾智能体为代表的执行型AI工具，必须严格遵循工信部发布的“六要六不要”安全规范，执行相应的技术防护措施：一是仅使用官方发布的最新版本，绝不使用第三方修改的汉化版、破解版；二是绝不将实例直接暴露在公网，关闭外网访问权限，仅本地运行；三是绝不使用管理员权限部署，采用普通用户权限运行，严格限制其可访问的目录与系统资源；四是谨慎使用插件市场，仅安装经官方签名验证的插件；五是优先搭配本地部署的开源大模型使用，实现完全离线运行，若必须使用云端大模型，要严格限制数据传输的范围，绝不传输敏感文件与信息；六是采用容器化部署，将工具封装在托管容器或在虚拟机中运行，与本地系统实现隔离。

4.4 应急处置：建立风险预案，降低泄露损失

个人用户需做好核心数据离线备份，对AI工具的系统权限，设置操作熔断机制，一旦出现批量删除文件、大规模数据外传等异常操作，立即终止其运行权限；一旦发生密钥泄露、设备被控，立即断开网络，终止AI工具的运行，使用专业的安全软件进行查杀，必要时重装系统；若涉及身份信息、金融信息泄露，应及时向公安机关报案，留存相关证据。

5 结语

在AI技术快速迭代的今天，个人信息安全防护，从来不是拒绝技术、远离AI，而是要以理性的态度、审慎的行为、体系化的防护，在享受技术红利的同时，守住个人信息安全的底线。

只有形成个人自律、企业尽责、监管兜底的三方协同防护体系，才能真正实现AI技术健康可持续发展。

参考文献

[1] 龙卫球，郃江丽。生成式 AI 用户交互中的个人信息保护 [J]. 现代法学，2025 (05):3−12.

[2] 郑志峰，袁梦。生成式人工智能个人信息利用的规范体系 [J]. 北京航空航天大学学报 (社会科学版),2025,38 (02):1−9.

[3] 孟小峰。大模型时代的数据安全与隐私保护 [J]. 计算机学报，2024,47 (01):1−22.