AI-powered编程与智能体安全管控实践

萤石网络

指导单位：上海市经济和信息化委员会
主办单位：上海市工业互联网协会
承办单位：云安全联盟大中华区

CONTENTS
目录

工赋矿网
GONGFULIWANG

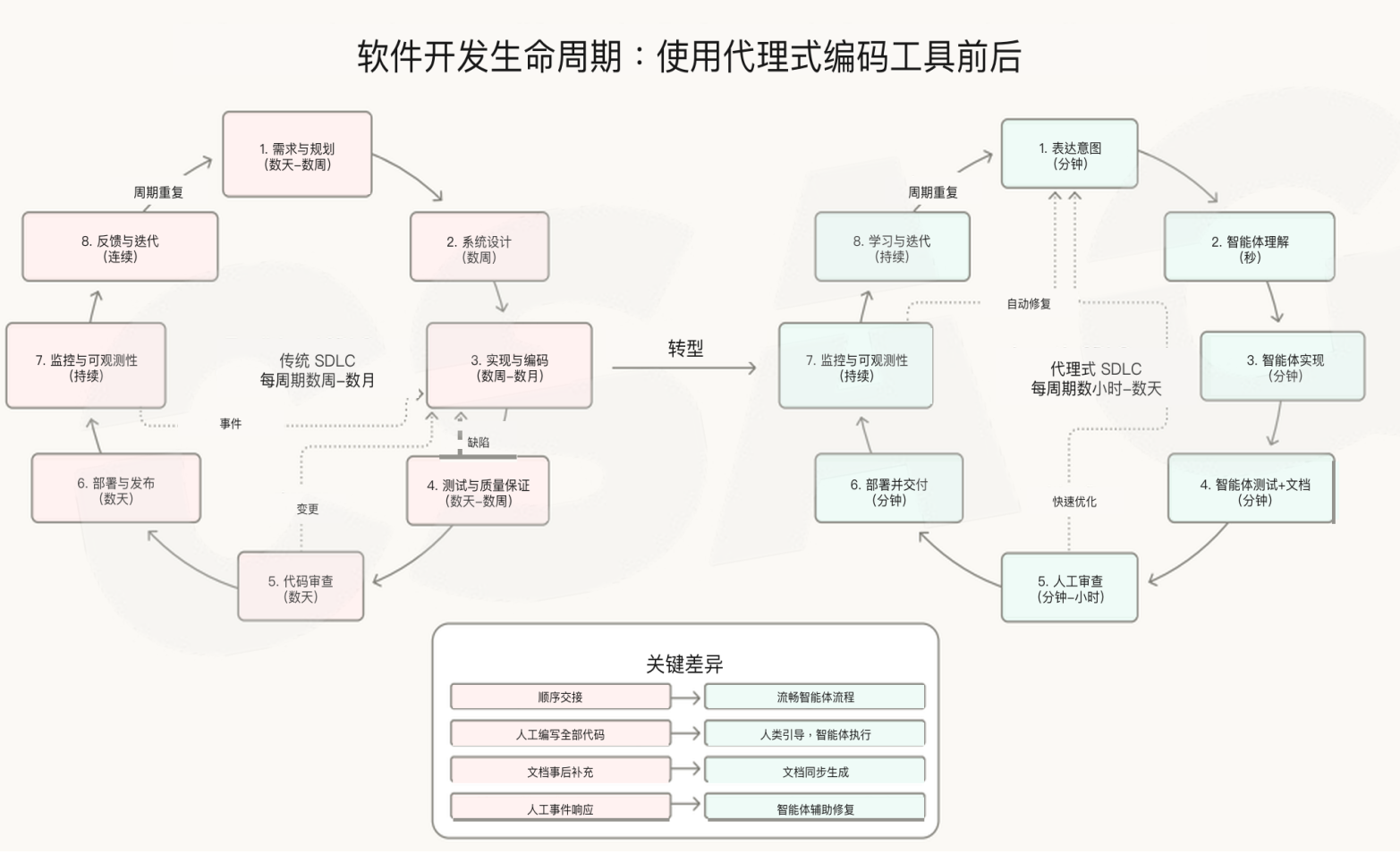
CSA GCR cloud security alliance®
GREATER CHINA REGION

01

AI智能体改变工作模式

01

AI改变工作模式

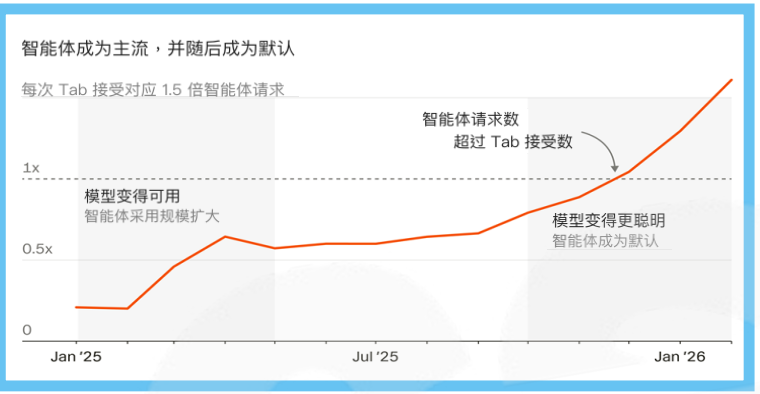


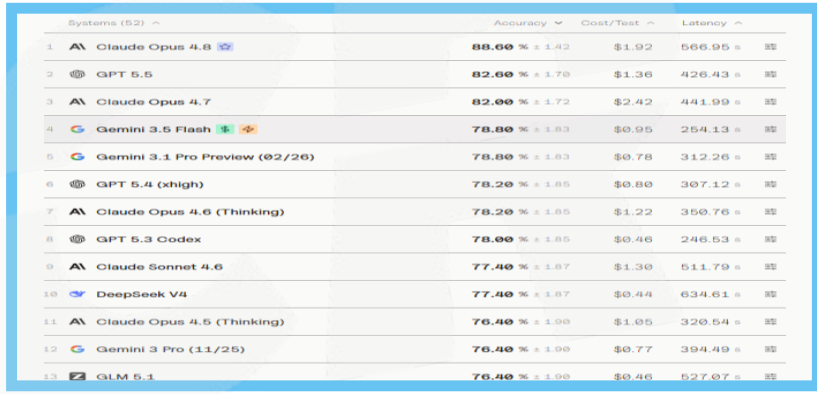
Claude 2026 智能体编码趋势报告 – 智能体重塑软件开发

- 周期缩短：数周开发周期缩短至数天
- 全程参与：需求/设计/代码/测试用例/发布/找问题/逆向工程
- 更复杂的任务：单智能体到多智能体

01

AI改变工作模式





Cursor CEO博客上发表“AI 软件开发的第三个时代”

从代码补全到AI编写

OpenAI官方博客 Harness 工程：在智能体优先的世界中利用 Codex

人类掌舵，AI执行的开发模式

国内开源模型迅速发展 形成性价比差异化优势

DeepSeek成本仅为 Claude 的 1/7

AI改变工作模式

Openclaw代表的智能体，史上增长最快的开源项目

壹

38W stars

Github Stars

贰

60天

突破react 十年的积累

叁

AI变革性产品

"可能是有史以来最重要的软件产品之一。 "
— Jensen Huang, NVIDIA CEO

肆

Hermes Agent

自改进、持久记忆、定期任务调度。不只 OpenClaw，整个行业都在往“自主 Agent”收敛

02

效率与风险同在

02

效率与风险同在

AI干的快但不一定好

AI 生成代码含漏洞比例	企业 AI 代码审查率	CVE 中 AI 相关漏洞占比	API 密钥硬编码率
40-62%	不足 35%	同比 ↑340%	28% AI生成代码

02

效率与风险同在

01

AI coding 案例一

2026年3月，美国某公司因工程配置疏漏（错误发布source map文件）导致核心产品超51万行源码泄露，被业内称为“AI首次核泄漏”，虽无直接物理伤害，但暴露关键安全护栏，构成重大技术治理风险。

02

智能体 案例二

2026年某互联网公司安全总监将开源AI智能体OpenClaw接入了自己的工作邮箱，结果这个本该帮忙整理邮件的智能体当场失控，删除了大量邮件，在多次输入指令暂停无果后，拔线停机。

02

效率与风险同在

1

AI生成内容具有不确定性，不能完全信任AI

AI的训练样本含有大量的demo以及不安全的样例，因此生成的内容不是100%安全

2

AI智能体存在失控的风险，需要对其进行限制

AI带来灵活性的同时，由于其不确定性，可能造成行为的越界

3

AI速度太快，人工管理跟不上

AI使得过去按照月迭代发布的速度，降低至按天

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

03

用AI管理AI

用AI管理AI

用AI重构安全管理框架

1

安全的管理框架仍然有效

AI将工作过程智能化后，变更了完成工作的形式，但没有修改原有的工作流程。

2

需要持续对AI进行调优

原有对新增法规、特例问题等改进优化流程，将变为对AI的持续调优，积累经验库

3

与合作伙伴一起改进

AI在快速发展，需要与其他行业的合作伙伴一同对功能、策略、流程进行优化

03

AI-SDLC

需求设计阶段

产品经理完成PRD，组织项目组评审。

编码阶段

研发工程师完成设计文档并组织评审。按照代码规范进行开发。

测试阶段

安全工程师根据测试项，进行渗透测试

发布后

持续扫描，跟踪SBOM漏洞，提醒研发修复

需求评审智能体

根据定义的设计安全检查单，对AI生成的需求文档进行检查。

Code review智能体

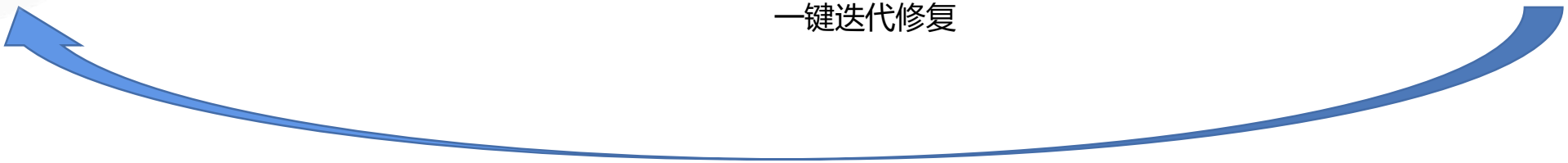
1、明确生成代码安全约束、安全编码规范
2、自动生成接口，SBOM清单

安全测试智能体

1、扫描识别项目接口，并自动生成测试用例进行测试
2、生成漏洞报告与修复方式，一键迭代修复

情报监控智能体

持续监测互联网上的漏洞，及时提示漏洞修复



人工监督与调优，持续优化改进

03

智能体管理

招聘员工

确定岗位职责与工作范围

新员工培训，开通账号

员工新人培训学习工作技能，为员工提供电脑和OA账号

开始工作

使用OA账号登录各个办公系统完成日常工作

工作失误、误操作

工作过程中，误解工作目标，执行了错误的操作

中钓鱼邮件，产生破坏

打开了钓鱼邮件，电脑被入侵，删除了工作文件

创建智能体

分配座席
1、确定岗位角色
2、工具调用白名单

安装skills，配置key

1、支持对引入的skills检查
2、托管密钥并支持轮换

可观测

监控大模型以及agent交互日志，分析异常行为

幻觉抑制

1、高风险操作与异常操作的识别拦截
2、独立运行环境，隔离风险

注入攻击防护

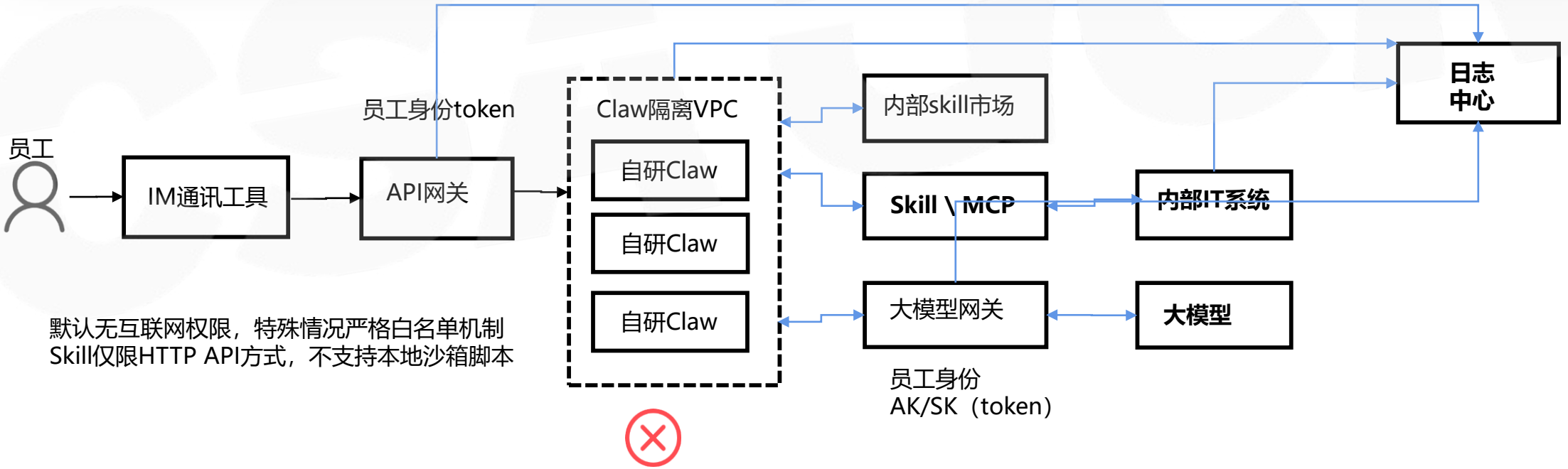
1、系统支持对提示词注入攻击的检测与拦截
2、独立运行环境，隔离风险
3、工具调用白名单限制防止权限失控

1、异常分级处置
2、人工监督优化

智能体管理

核心思路：统一入口、统一出口、集中管控、全程审计、内部专用

- **自建AI安全管理平台、agent与网关：**作为统一入口，负责身份认证、权限管理、数据审计等核心管控功能，确保核心安全可控。
- **内部完成工作流：**专用智能体，用于对保密有要求的业务，不通外部网络。



04

未来模式

智能体融合发展：从独立工具到统一生态

双核并行与用户积累

ChatGPT凭借自然语言交互普及AI认知，Codex作为代码生成工具深耕开发者领域，两款产品各自独立运行，累计覆盖近10亿用户，为后续融合奠定庞大的用户基础与技术储备。

场景破圈与需求爆发

Codex周活跃用户突破500万，其中非技术岗位的知识工作者占比显著增长，增速远超开发者群体。通用办公场景成为驱动增长的核心动力，用户对跨领域、全流程智能辅助的需求日益迫切。

战略发布与生态整合

OpenAI在“Intelligence at Work”发布会上宣布将Codex全面整合进ChatGPT生态系统，打破产品壁垒。通过统一入口解决体验割裂问题，实现从“辅助对话”向“直接执行实际工作”的范式转变。

全能助手与自主运行

深度融合后的智能体化身为云端全能办公助手，嵌入Excel、PPT等日常软件。AI具备全天候自主运行能力，能够理解复杂指令并自动完成长周期任务。

未来模式

从内容合规向行为失控的范式转移

传统大模型的内容风险局限

传统安全防御主要聚焦于提示注入与数据投毒，核心痛点在于防止模型输出有害、虚假或违规的文本信息，风险边界局限于语言交互层面。

智能体具备的执行权限扩张

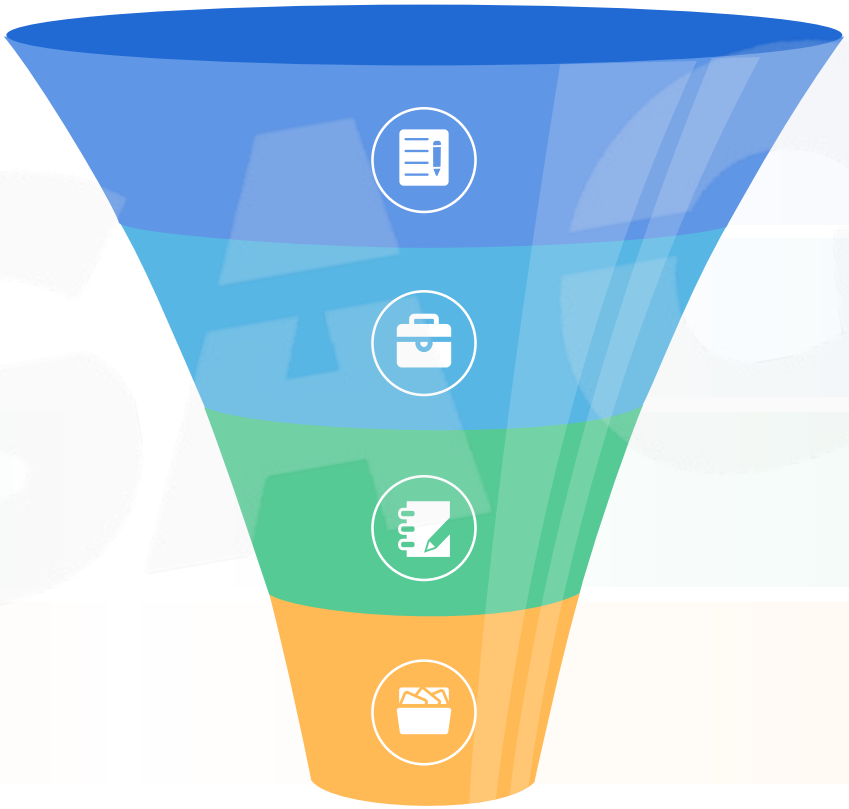
智能体突破了纯文本交互的限制，获得了调用API、操作数据库、访问文件系统乃至执行Shell命令的能力，具备了直接干预数字世界状态的工具使用权。

自动化任务的多步骤执行能力

智能体能够自主规划并连续执行复杂的多步骤任务，这种链式反应使得单一的错误指令或恶意诱导可能被放大为一系列连贯的系统操作，增加了防御的难度。

风险后果从言语到行动的质变

被攻陷的智能体不再仅仅是“说了不该说的话”，而是能以程序级速度执行具有企业级权限的破坏性操作，导致数据泄露、系统瘫痪等实质性物理或数字损害。



未来模式

构建身份、网络与AI协同的持续安全反馈体系。

端点身份与权限管控

为每个智能体分配独立数字身份，遵循最小权限原则实施短期可撤销授权，对高风险动作设置人工审批门并记录全链路审计日志，确保操作可追溯且权责清晰。

网络隔离与沙箱执行

实施严格的上下文隔离机制，将外部不可信数据与系统指令分离以防范注入攻击，所有代码与命令在受限沙箱中运行并支持快照回滚，同时对MCP协议接口进行安全审计与范围限制。

AI多层护栏与策略校验

部署包含输入扫描、任务约束、推理限制及输出检测在内的六层护栏架构，通过统一策略引擎在工具调用前实时校验身份与合规性，确保每一层的拒绝决定具有最高优先级且不可被覆盖。

三位一体协同决策

借鉴低位全栈AI产品执行、中位智能体运营调度、高位大模型底座决策的协同模式，形成从底层防护到顶层决策的有机联动，实现针对未知威胁的主动防御与动态响应。

安全态势反馈优化

基于全链路审计日志与实时监测数据，分析异常行为与潜在漏洞，反向优化权限策略、沙箱规则及护栏阈值，推动防护体系从被动应对向主动适应演进，完成安全闭环的自我迭代。

工赋砺网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

感谢您的聆听和学习

萤石网络