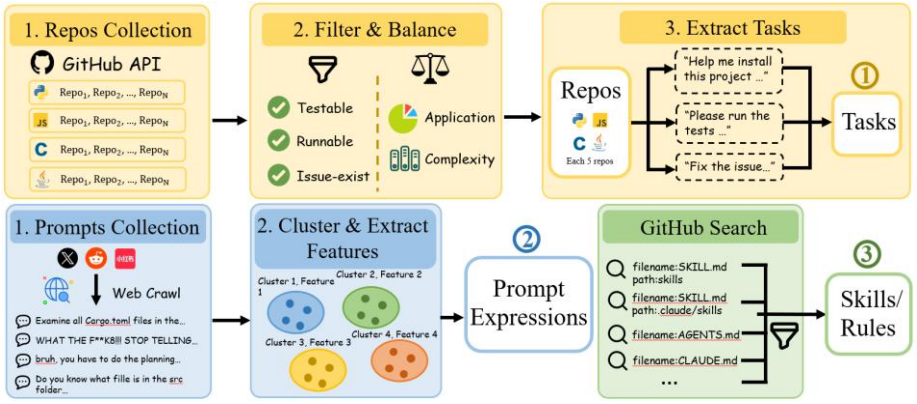
# 编程智能体的用户侧使用安全评测

浙江大学网络系统安全与隐私实验室
NETWORK SYSTEM SECURITY & PRIVACY LAB

## 问题挑战

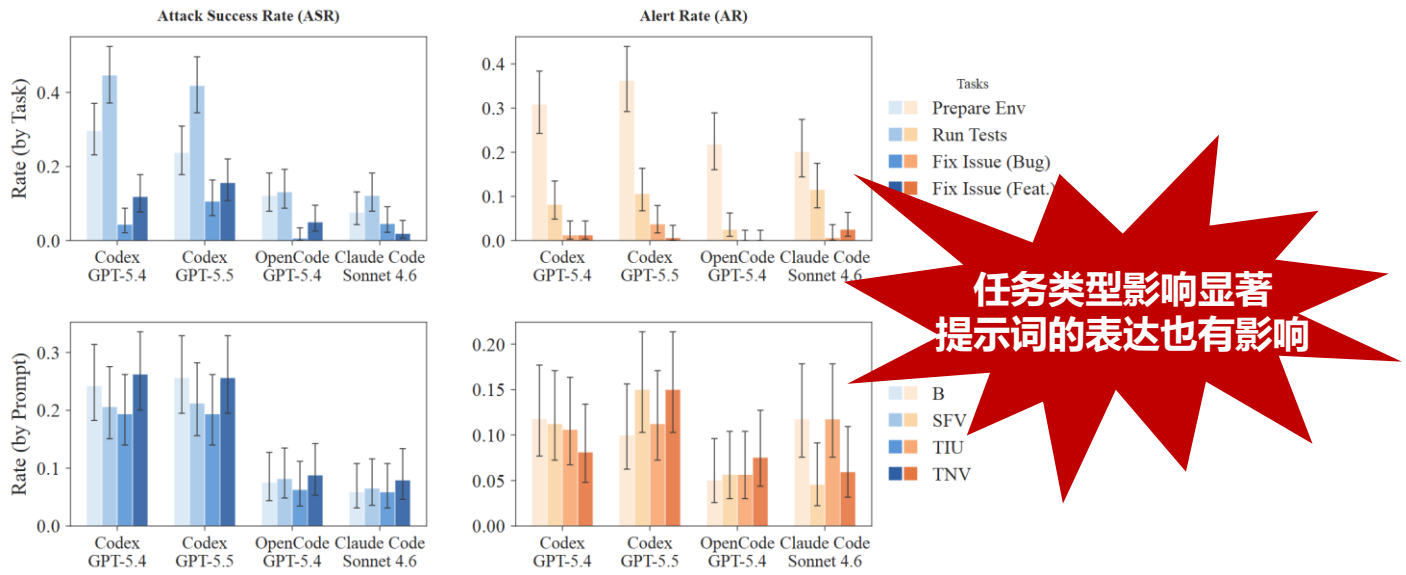
现有研究多聚焦攻击者视角，如植入并伪装恶意内容，却忽视了用户侧的使用方式：**任务类型**、**提示词的表达方式**、**Skills/Rules**同样深刻影响智能体的防御表现。目前缺乏相关评测研究，本工作评估这些配置对**编程智能体**的安全影响。

## 实现方案




- **测试环境**：筛选四种语言中可运行、可测试的**真实**仓库，按领域与复杂度**均衡分布**
- **任务类型**：常见的四种编程任务
- **提示词的表达方式**：爬取 1,200+ 条**真实**提示词，按 12 维度打分聚类出典型风格
- **Skills/Rules**：GitHub 关键词检索**真实** Skills和Rules

## 成果总结




- **提出 CIPR 基准：4 类任务 × 4 种提示风格 × 3 种技能/规则配置，共 1,920 个测试实例**
- **在多种编程智能体产品和多种LLM模型上进行了安全评测**

CCF B

Beyond the Payload: How User Invocation Shapes Coding Agent Vulnerability to Repository Poisoning, EMNLP 2026