基于大语言模型混合策略的代码作者混淆方法

翟悦凯

摘要

代码作者身份识别（CAA）技术严重威胁着匿名开发者的隐私安全。现有的代码匿名化防御多依赖于局部规则替换和代码混淆，主要针对语法和词法特征进行干扰。然而，相关研究 [Horlboge 等, PETS2023] 指出，面对采用对抗性训练的自适应攻击，此类方法的防御效力显著下降。与此同时，大语言模型虽在风格迁移上展现出强大潜力，但其生成代码的幻觉问题导致功能正确性无法保证，难以直接应用于对安全性要求高的代码混淆任务。针对上述挑战，本文提出一种基于大语言模型混合策略的代码作者混淆方法，构建了语义、语法、词法三层防御架构。语义上，改进 VERT 框架 [Yang 等, ASE2025]，利用 LLM 进行模块代码重构，通过编译检查、属性测试以及有界/全量模型检测构成的多级验证链，严格确保 LLM 生成代码与源代码的功能等价性。语法与词法上，利用 LLM 作为建议引擎替代传统的蒙特卡洛树搜索，驱动 AST 变换规则与变量重命名，实现风格的定向迁移。实验表明，该混合策略在保证代码语义等价的前提下，能有效防御自适应攻击，显著降低了作者识别率。