# 基于语言无关特征和LLM融合的开源组件投毒检测项目需求调研报告

## 一、项目概述

**项目名称**：基于语言无关特征和LLM融合的开源组件投毒检测系统  
**项目背景**：

* **背景**：开源软件供应链攻击激增（如PyPI/npm恶意包窃取敏感数据、植入木马），传统检测工具存在语言依赖性强、误报率高（如Bandit4mal误报率86%）、跨语言样本稀缺等问题。
* **目的**：构建智能化检测系统，通过语言无关特征+LLM融合技术，解决跨语言恶意包识别难题，提升开源生态安全性。
* **预期价值**：能够在限定时间（如一周内）检测一定量（60 ~ 100个）的未知恶意包（实际验证），误报率降低60%+，支持自动化监控。

**目标用户**：

* **开发者**（25-40岁，技术从业者）：需快速验证依赖包安全性，避免项目被投毒。
* **安全团队**（30-45岁，企业/机构）：需监控供应链风险，减少人工审核成本。

**· 开源平台维护者**（如PyPI/npm管理员）：需高效下架恶意包，保障仓库安全。  
**核心痛点**：

1. 跨语言检测难：单一工具无法覆盖JavaScript/Python等多生态。

2. 误报率高：传统工具（如OSSGadget）误报率>77%，人工验证负担重。

3. 新型攻击应对不足：混淆代码、依赖混淆攻击等难以识别。

## 二、市场调研

**行业现状**：

* **规模**：2023年开源供应链攻击增长78%，PyPI/npm平台年均发现恶意包超2000个。
* **趋势**：
  + LLM+安全分析成为新方向（如ChatGPT用于代码审计）。
  + 跨语言检测需求激增（60%企业使用多语言栈）。
* **痛点**：现有方案（如华为Donapi）仅支持单语言，且依赖特定特征库。

**用户需求分析**：

* **核心需求**：
  + 实时监控新上传包，自动化检测（响应时间<5分钟）。
  + 跨语言支持（至少覆盖JavaScript/Python）。
  + 误报率<10%（当前竞品平均>50%）。
* **高级需求**：
  + 提供恶意代码定位与解释（辅助人工审核）。
  + 可视化报表（恶意包分布、威胁类型统计）。

## 三、竞品分析

**竞品选择**：华为Donapi、OSSGadget-backdoor

**竞品概述1：华为Donapi**

* **产品名称**：Donapi（恶意npm包检测工具）
* **访问地址**：[论文链接](https://arxiv.org/abs/2403.08334" \t "_blank)
* **产品定位**：面向企业安全团队，专注JavaScript生态恶意包检测。
* **功能对比**：

| **功能** | **Donapi** | **本项目** |
| --- | --- | --- |
| 核心功能 | 基于API序列的恶意行为映射 | **跨语言特征+LLM融合检测** |
| 特色功能 | 高精度JavaScript检测 | **支持JS/Python双生态** |
| 缺失功能 | 无Python支持 | **LLM复筛降低误报率** |

* **用户体验**：
  + 界面设计：命令行工具，无图形界面。
  + 交互体验：需手动配置，无实时监控能力。
* **优劣势**：
  + 优势：JavaScript场景召回率高（89%）。
  + 劣势：跨语言能力缺失，误报率15%。

**竞品概述2：OSSGadget-backdoor**

* **产品名称**：OSSGadget-backdoor
* **访问地址**：[GitHub仓库](https://github.com/microsoft/OSSGadget" \t "_blank)
* **产品定位**：开源社区工具，检测后门代码。
* **功能对比**：

| **功能** | **OSSGadget** | **本项目** |
| --- | --- | --- |
| 核心功能 | 基于规则的后门扫描 | **141个语言无关特征分析** |
| 特色功能 | 支持多包格式解析 | **自动化爬包+实时监控** |
| 缺失功能 | 无机器学习模型 | **XGBoost+LLM双级验证** |

* **用户体验**：
  + 界面设计：基础命令行输出。
  + 交互体验：无状态管理，崩溃率高。
* **优劣势**：
  + 优势：轻量级，易集成。
  + 劣势：误报率77%（本项目3.4%），无法检测新型混淆代码。

## 四、需求分析结论

**核心需求优先级**：

1. **跨语言检测能力**（JavaScript/Python必选，Ruby/PHP扩展）。
2. **低误报率**（XGBoost初筛+LLM复筛保障<5%）。
3. **自动化流水线**（包爬取→特征提取→检测→报告生成）。

**差异化优势**：

* **语言无关特征集**：141个特征覆盖混淆字符串、熵值分析等通用模式。
* **LLM复筛机制**：ChatGPT定位恶意代码片段，解释可信度高（准确率100%）。
* **真实场景验证**：10天检出134个未知恶意包（npm 38个 + PyPI 96个）。

**市场挑战**：

* 用户对LLM黑盒性的信任问题 → 需提供可解释的恶意代码片段输出。
* 小众语言（如Ruby）样本稀缺 → 优先巩固JS/Python生态。

## 五、产品规划

**功能规划**：

| **模块** | **功能描述** | **优先级** |
| --- | --- | --- |
| 包爬取监控 | 实时扫描npm/PyPI新包（**concurrent-couch-follower**库） | 高 |
| 特征提取引擎 | 计算141个语言无关特征（熵值、文件扩展名分布等） | 高 |
| XGBoost初筛 | 跨语言恶意包分类（召回率75.5%） | 高 |
| LLM复筛 | ChatGPT恶意代码定位（输出代码片段+类型） | 中 |
| 可视化平台 | 恶意包分布雷达图、威胁类型饼图 | 中 |

**技术选型**：

* **后端**：Python（PyTorch、XGBoost）、FastAPI。
* **前端**：Vue.js + ECharts（可视化）。
* **LLM集成**：OpenAI ChatGPT API。
* **部署**：Docker容器化 + Kubernetes集群管理。

**迭代计划**：

| **阶段** | **目标** | **周期** |
| --- | --- | --- |
| V1.0 | 实现JS/Python跨语言检测流水线 | 2周 |
| V2.0 | 集成LLM复筛与Web可视化报表 | 1-2周 |
| V3.0 | 扩展Ruby/PHP支持 | 不定，可能无法完成 |

**核心指标验证**：

* 误报率：≤5%（竞品平均>50%）。
* 检出率：10天内未知恶意包≥100个。
* 响应速度：新包扫描≤5分钟。