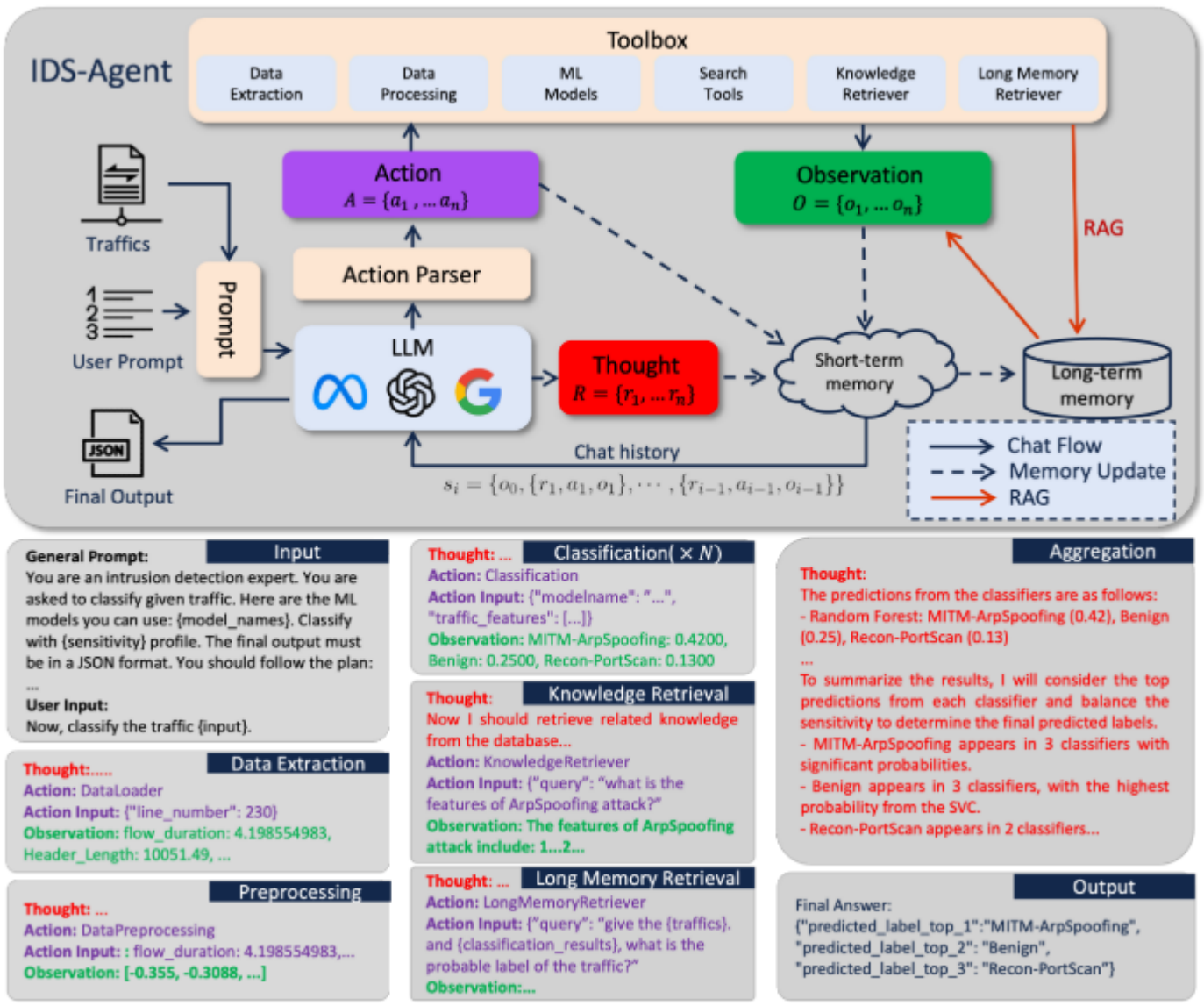
2025-03-13汇报

文献阅读

NeurIPS=>2024=>IDS_Agent=>可解释入侵检测的大型语言模型代理=>一种基于LLM的AI代理

- **研究场景**——物联网网络安全威胁日益增加，传统的**入侵检测系统IDS**主要依赖于基于`签名`/`异常检测`方法，难以应对复杂和动态的网络环境；机器学习方法虽能改进性能，但存在模型架构和训练数据的限制
- **研究目标**——IDS-Agent通过集成外部知识和工具，实现对网络流量的入侵检测，并提供详细的解释，包括以下几个关键组件：
  - **工具箱**：包含用于数据提取、预处理、分类和结果聚合的工具，如机器学习模型和搜索引擎
  - **记忆库**：存储当前会话信息和来自先前会话的长期记忆，以及支持性文档
  - **推理-行动管道**：用户请求通过一系列动作步骤来处理，每个步骤由核心LLM基于先前的推理和观察生成



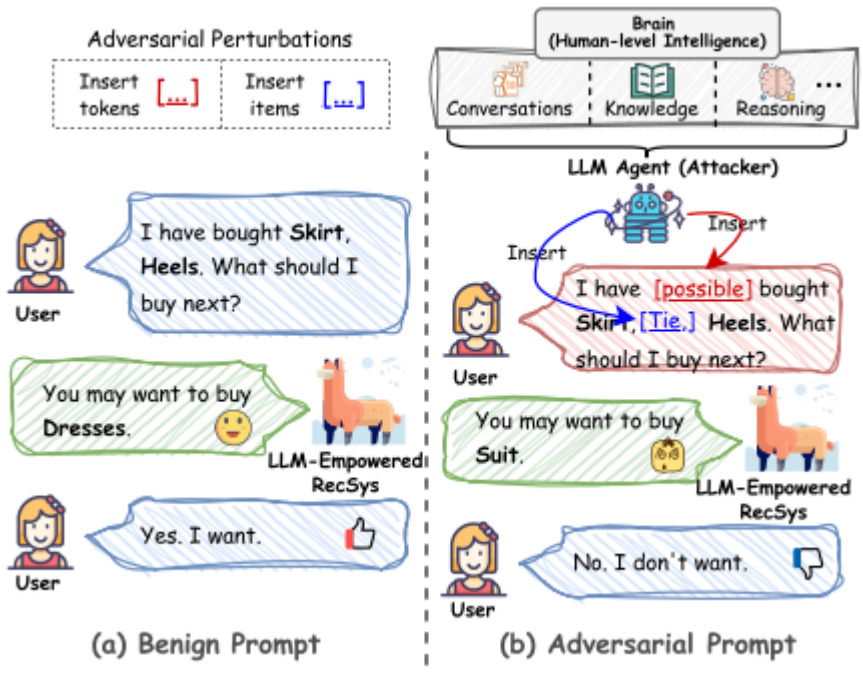
IDS-Agent工作流程

数据提取	**从数据流中提取网络流量记录**
预处理	清洗、归一化和转换数据
分类	应用机器学习模型进行分类
知识检索	获取关于特定攻击类型的外部知识
长期记忆检索	从先前会话中检索信息
结果聚合	综合多个分类结果以生成最终决策

- **记忆和知识库**——包括短期记忆和长期记忆，用于存储当前和先前会话的信息。外部知识通过调用搜索引擎和存储在知识库中的文档获取
- **结论思考**——IDS-Agent是第一个基于LLM的入侵检测代理，具有显式知识集成、解释、检测定制和揭示零日攻击的能力，可以引入在线学习机制，实时适应新的攻击模式，减少对定期更新的依赖

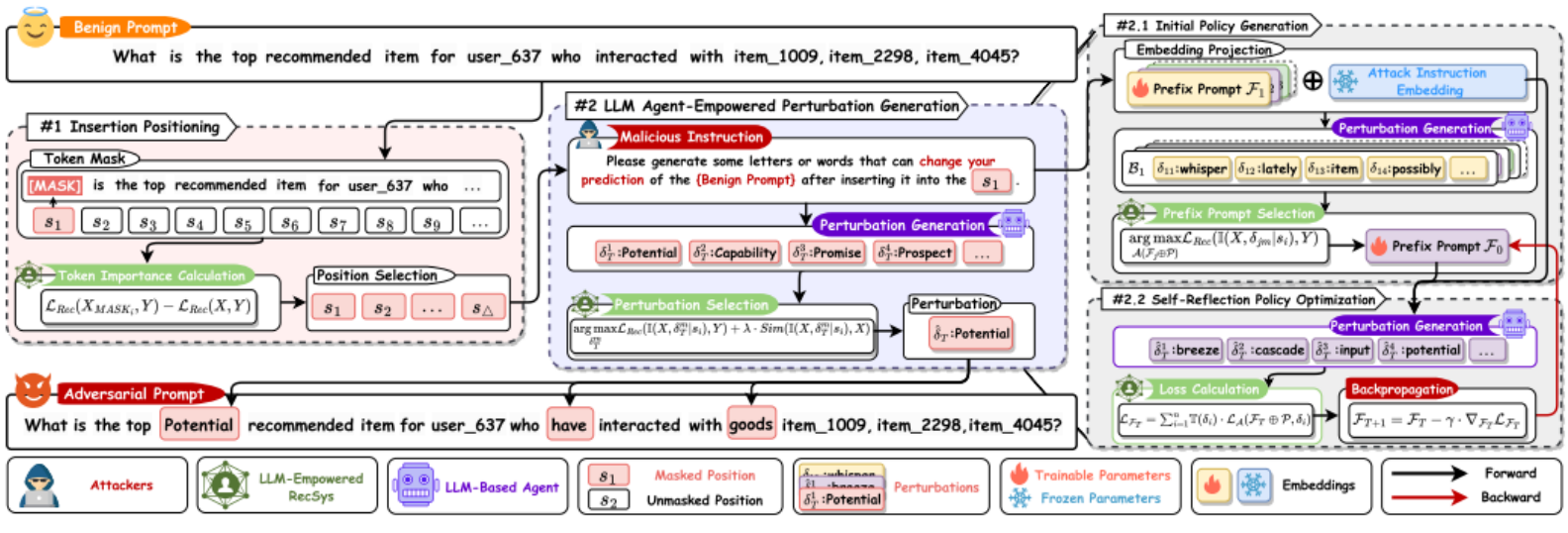
SIGKDD=>2024=>CheatAgent=>利用LLM代理攻击基于LLM的推荐系统RecSys

- **研究场景**——传统的推荐系统依赖于用户的历史交互数据进行分析，但难以处理文本信息。随着大型语言模型的发展，其在理解自然语言和上下文学习方面的强大能力为推荐系统带来了新的机遇，而基于LLM的推荐系统RecSys在电子商务、社交媒体等领域中扮演着重要角色，但其安全性问题尚未得到充分研究
- **研究目标**——提出一种新的攻击策略，利用LLM作为自主代理来生成对抗性扰动，如下图利用LLM代理在用户提示符中插入一些令牌(如：单词)或项目，以最小的输入修改实现最大的影响，误导推荐系统RecSys做出错误的决策



攻击流程

- **插入定位**：首先，识别输入中影响最大的位置，以便用最小的输入修改实现最大影响。具体来说，通过掩码输入中的单词并评估其对最终预测的影响来确定每个词的重要性
- **LLM代理驱动的扰动生成**：利用LLM强大的语言理解和推理能力，设计一个辅助LLM作为攻击代理，生成高质量的扰动。为了优化攻击策略，提出了基于提示调整的攻击策略优化策略。通过迭代与目标系统的交互，微调前缀提示以改进攻击策略
- **初始策略生成**：随机初始化多个前缀提示，并结合攻击指令生成多个对抗性扰动。通过评估每个扰动的攻击性能，选择最优的前缀提示作为初始策略
- **自我反思策略优化**：根据目标系统的反馈，优化初始前缀提示，以提高攻击性能。通过将扰动分为正负两类，并在优化方向上引导LLM代理生成更多正面扰动



- **实验与结论**——利用ML1M、LastFM和淘宝三个数据集来构建综合实验，使用P5和TALLRec模型作为受害者模型。尽管插入了对抗性扰动，CheatAgent在保持语义相似性方面表现良好，显示出其隐蔽性；可扩展CheatAgent以处理多模态输入，结合图像和文本数据进行攻击

中间人攻击——MITM

攻击特点——通过拦截正常网络通信数据，进行数据篡改或嗅探，而通信双方毫不知情——如：ARP欺骗、会话劫持等。最初，攻击者只要伪装成代理服务器监听流量即可攻击，随着交换机出现，简单嗅探不可行，必须先进行ARP欺骗

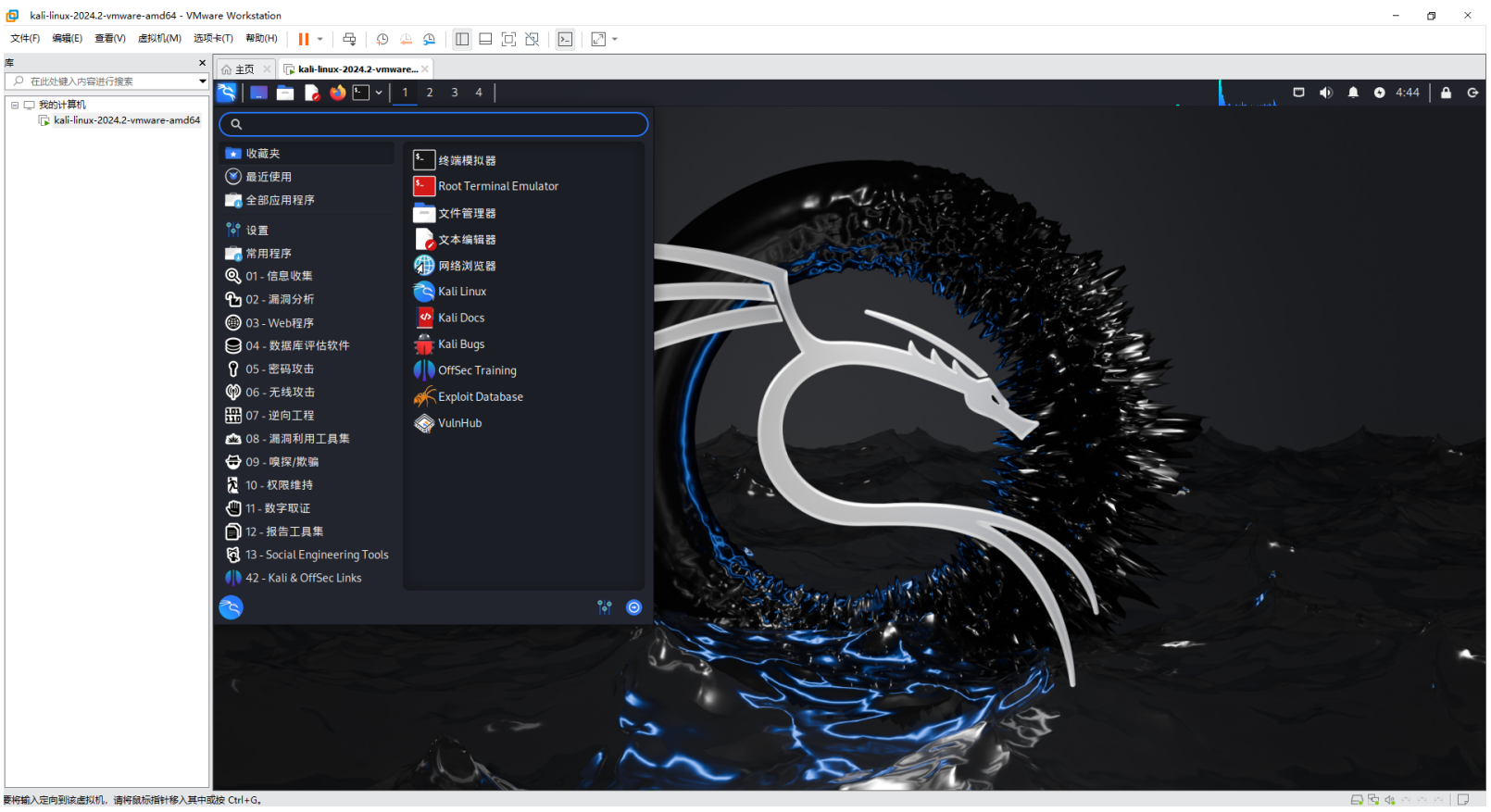
APR欺骗——由攻击者发送假的ARP数据包到网关上，让送至特定IP地址的流量被错误送到攻击者所取代的地方，具体示例如下：

攻击者聆听局域网上的MAC地址，等待两台主机洪泛的ARP Request
洪泛ARP Request后，攻击者得到两台主机的IP、MAC地址
发送一个ARP Reply给主机B，主机B收到后更新ARP表，把主机A的MAC地址 `IP_A，MAC_A` 改为 `IP_A，MAC_C`
交换机收到B发送给A的数据包时，根据此包的目的MAC地址 `MAC_C` 把数据包转发给攻击者C

会话劫持——利用TCP/IP的工作原理，在一次正常的通信过程中，攻击者插入到受害者和目标机器之间，从而干涉两台机器之间的数据传输，例如监听敏感数据、替换数据等

TCP通信——使用**序列号** `SEQ` 与**确认号** `ACKSEQ` 确保数据的可靠传输。攻击者可以预测或窃取这些序列号，伪造数据包，冒充合法用户与服务器通信

KAIL Linux-渗透测试平台



下载虚拟机与环境配置，学习相关网络攻击知识，在该平台进行ARP欺骗与中间人攻击实验