附件 1

测试场景说明

场景 1：AI 驱动攻击的网络安全智能防御

利用 AI 大模型或智能体对流式推送的多源日志进行实时自动化分析，检测攻击行为，并通过跨设备、跨时间窗口的关联分析还原属于同一攻击链的事件。测试数据包括 WAF 日志、服务器 WEB 日志、DNS 日志、主机进程日志、主机认证日志等。按照攻击识别的准确率、漏报率以及攻击链还原的准确率综合评分。

场景 2：网络系统及源代码漏洞智能挖掘

使用 AI 智能体自主完成目标理解、程序分析、漏洞验证、漏洞评估等任务，实现对源代码和网络系统的漏洞挖掘。漏洞挖掘结果需要包含对漏洞位置、利用条件、复现场景等描述，达到评判人员能够复现的程度。按照漏洞识别的准确率、漏报率等指标进行综合评分。

场景 3：网络流量安全威胁检测

使用 AI 技术对互联网流量进行分析处理，识别侦察、初

始访问、持久化和命令与控制等攻击阶段，以及端口扫描、漏洞利用、植入后门、木马回连等攻击技术。测试数据为PCAP 格式的流量数据包文件。按照识别的准确率、漏报率等指标进行综合评分。

场景 4：网络安全告警日志降噪

使用 AI 技术对网络威胁告警日志和终端日志进行分析，从日志中找出攻击成功、可以形成攻击事件的真实告警。按照真实告警识别的准确率、漏报率等指标进行综合评分。

场景 5：大模型安全护栏能力检测

评估大模型安全护栏相关产品的风险识别与防护能力，帮助大模型过滤安全风险，促进提升大模型安全防护水平。测试数据为对大模型的提问内容及其输出内容。按照准确率、漏报率等指标进行综合评分。

场景 6: AIGC 生成图片检测

在海量图片中识别出 AI 技术生成的图片。测试数据为图像数据，按照识别的准确率、漏报率进行综合评分。

场景 7：AI 智能体恶意操作行为检测

对 AI 智能体运行过程中产生的日志进行分析，检测并

识别其中权限获取与提升、敏感数据窃取与外传、破坏性操作等恶意行为。测试数据包括应用日志、主机日志以及 PCAP 流量包 3 类信息。按照识别的准确率、漏报率进行综合评分。

场景 8：广播电视 IPTV 账号异常行为检测

使用 AI 技术对 IPTV 账号的原始行为日志进行分析，识别出内容爬取、刷播放量、账号共享等异常行为。测试数据包括点播记录、页面访问记录、开机记录。按识别异常行为的准确率、漏报率等指标综合评分。