2024 年 CCF-绿盟科技“鲲鹏”科研基金项目申报指南

一、总则

CCF-绿盟科技“鲲鹏”科研基金重点面向国内高校、科研机构的研究人员和团队，旨在以小微课题的方式支持科研人员的研究与创新，推动科研技术成果转化，促进外部科研机构优秀研发能力与公司内部产品价值的深度融合，构建互动合作与创新发展的生态圈，为绿盟科技的产品与解决方案创新赋能。

二、资助对象和条件

1.申请人须是国内高校、科研院所在职的全职教师或研究人员，具有高级专业技术职务，博士毕业，且拥有一定数量的相关领域研究成果，能作为项目的实际负责人并担负实质性研究工作；

2.申请人只能申报一个项目，不能重复申报。

三、资助方式及项目范围

1.项目实施期为 1 年,单项资助额度原则上不低于 8 万元。

2.2024 年，CCF-绿盟科技“鲲鹏”科研基金重点资助的研究领域和方向：

（1）数据安全方向；

（2）人工智能安全方向；

（3）云计算安全方向；

（4）安全对抗方向。

四、项目申请和评审

1.符合条件的研究人员在项目申报规定时间内填写《2024年 CCF-绿盟科技“鲲鹏”科研基金项目申报表》并发送至kunpeng2024@nsfocus.com，每位申请人仅限提交一份申请。

2.申请人在申报前需确认所在高校/科研院所可以作为项目依托单位签署科研合作协议，申请人本人可以作为项目负责人签署项目保密协议等相关承诺文件。任何针对项目申报的问题，请联系尤老师：(010)68438880-5485，技术答疑邮箱：kunpeng2024@nsfocus.com。

3.评审时，CCF和绿盟科技成立联合项目组，共同邀请专家审核申请项目。专家评审时主要考虑以下方面：

（1）申请项目的研究意义，包括国内外研究现状以及市场前景；

（2）申请项目的技术基础，包括技术成熟度、自主知识产权积累、与主流技术对标或产品适配验证等情况；

（3）申请项目的主要研究内容，包括技术路线、研究内容、具体技术指标、创新性、支撑条件需求等；

（4）项目实施计划、预算设计的合理性；

（5）预期交付的成果形式及数量；

（6）申请者能力及团队保障条件，包括领军人物、团队学术水平和科研能力。

4.联合项目组依据专家审核意见，结合公司具体情况，确定资助的研究项目及资助强度等。

五、本期项目时间安排

项目指南发布	2024 年 8 月 21 日
项目申请截止	2024 年 9 月 22 日
答辩时间	2024 年 10 月中旬
结果发布，签署协议	2024 年 10 月下旬
CNCC2024 项目证书授予仪式	2024 年 10 月 24 日
中期检查，提交报告	2025 年 5 月
提交成果，终期答辩	2025 年 11 月

项目进行过程中的具体时间节点，请关注 CCF-绿盟科技“鲲鹏”科研基金项目组通知。

六、项目经费管理

1.基金项目评审结果公布后的 3 个月内，CCF、绿盟科技、项目负责人及其所在单位四方需完成基金项目技术交底及项目合同书的签署工作，以确定各方责任和义务，鲲鹏基金支持的项目将依据项目合同进行管理。

2.合同履行期间，绿盟科技可根据需要委派领域专家（组）或其代表，对受资助人合同履行的情况进行检查、监督。CCF 依据绿盟科技委托，根据确定的项目经费、项目执行检查情况及合同约定，将项目经费分阶段划拨至项目负责人所在单位。项目负责人按阶段提交研究成果和检查报告。

3.CCF-绿盟科技“鲲鹏”科研基金实行专款专用，该经费不得用于发放人员工资（可用于劳务费支出），可用于项目研发产生的相关设备费、材料费、试验加工费、信息资料费、差旅费、管理费等。

七、项目管理

1.项目立项后不可更换项目负责人。在项目研究工作中，如因项目负责人自身原因中断研究工作，而造成项目终止。项目负责人需根据项目合同书的经费使用说明，退回已拨经费。

2.绿盟科技按照合同条款约定定期检查评估全部资助项目，项目负责人需按照合同要求，按时填写提交《中期报告表》，并提交阶段成果。

3.项目完成后，项目负责人填写《结题报告表》，由联合项目组组织检查验收，项目负责人应将结题报告和合同中规定的相关技术成果完整提交给绿盟科技和项目负责人所在单位归档。

4.项目负责人原则上不可放弃基金资助，如有特殊情况，需提交《放弃基金声明》并加盖项目负责人所在单位公章后，由联合项目组存档留备。

八、成果管理

1.项目负责人在项目研究过程中形成的与项目相关的成果的著作权及专利等，包括但不限于论文、著作、源代码、研究报告和数据等，其知识产权权利归属项目负责人及其所在单位和绿盟科技三方共同所有。绿盟科技有权免费优先使用。使用的具体细节以与项目负责人和其所在单位签署的协议为准。

2.在此期间发表的论文及著作需标注“受CCF-绿盟科技‘鲲鹏’科研基金资助”字样。CCF有权将上述论文及著作收入CCF数字图书馆，供CCF会员阅读。

本指南自公布之日起实施。

CCF-绿盟科技“鲲鹏”科研基金项目组

2024年8月21日

2024 年 CCF-绿盟科技“鲲鹏”科研基金申报方向与指标要求

一、人工智能安全领域

1. 深度学习推荐算法投毒攻击与检测技术研究

研究背景

推荐系统是 Google、Baidu、Taobao、YouTube、Amazon 等大型在线服务平台的核心功能之一，能够向用户推荐新闻、商品、金融、政策等各类信息，在政治、民生和金融等发挥着重要作用。推荐算法一旦受到投毒攻击，不仅会造成企业和用户的财产损失，也可能成为谣言传播、负面宣传等危害社会稳定和公共安全，因此增强推荐系统模型的鲁棒性、安全性显得尤为重要。

推荐系统算法正从传统的基于协同过滤的方法向基于深度学习的方向转变。然而，深度学习模型对数据的依赖性强，且其复杂性和不可解释性使其在鲁棒性存在不足。这导致投毒攻击能够通过精心设计的恶意数据干扰模型的学习过程，从而操纵推荐结果。数据投毒攻击是一种危害极大的攻击方式，因此课题研究并探索相关的攻击与防御技术。

研究内容

1.研究适用于基于深度学习的推荐系统投毒攻击算法，在逃避现有防御检测的同时保持较高的攻击效果；

2.投毒攻击会导致训练数据漂移，研究可以引起数据漂移的边界模式数据的定义以及检测方法；

3.研究投毒用户的特征向量化表达，提高检测器对攻击用户分类特征的检测水平，实现对有毒样本的数据清洗。

考核指标

1.基于深度学习的推荐系统投毒攻击及检测算法 1 套；

2.研制上述研究内容的原型系统 1 套，支持多种深度学习的推荐算法投毒攻击与检测，并提供源代码；

3.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

2. 基于大模型的识别钓鱼网站技术研究

研究背景

钓鱼攻击是对互联网用户安全的严重威胁之一。钓鱼攻击在不断进化其攻击方式，倾向于使用更加复杂的伪装手段，动态生成钓鱼网站内容，每个受害者看到的钓鱼页面都不同，避免直接暴露攻击载荷。钓鱼网站的进化和变异对基于静态分析或基于相似性分析的检测方法带来新的挑战，因此课题研究并探索相关的钓鱼网站识别技术。

研究内容

1.研究钓鱼网站的多维度数据收集和预处理技术,包括钓鱼网站的 URLs、视觉特征、HTML 源码、页面内容等;
2.研究可扩展的模型融合技术，基于可信度和置信度分析，整合现有检测模型的识别能力对数据进行标注，构建大规模高质量的钓鱼网站语料库;
3.研究钓鱼网站识别的大语言模型，通过迁移学习、对抗训练、模型微调等适应钓鱼网站识别任务，并通过持续学习不断更新和学习新的钓鱼攻击技术。

考核指标

1.钓鱼网站识别率不低于 95%，单个网站识别时间消耗不得大于 20 秒;
2.识别算法至少 5 种以上，更够绕过钓鱼网站的对抗识别方法，方法种类不少于 6 种;
3.研制上述研究内容的原型系统 1 套,并提供钓鱼网站语料库、源代码;
4.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

3. 基于大模型的多源异构信息融合研判技术研究

研究背景

网络空间威胁检测与狩猎，愈发依赖多源、异构网络数据集的综合线索挖掘。需结合大模型认知推理能力，全面理解和分析网络、终端、蜜罐等综合线索并形成威胁识别和理解能力，需拥有判断研判依据以及相关知识是否充足的能力，如果研判知识缺失，需要能够给出缺失知识列表的能力。

研究内容

研究基于网络遥测日志，实现基于多源异构日志的安全事件阶段与威胁识别能力，包括：

1.提升大模型对网络、终端、蜜等异构安全日志的理解能力；

2.实现结合多类型、多来源日志的综合解读能力与关联分析能力；

3.实现关键攻击线索的识别与抽取，支撑攻击链路、攻击步骤等攻击知识的自动化抽取；

4.实现复杂日志行为语义的基线自动化生成，实现面向未知威胁、潜在风险的异常线索挖掘；

5.实现对研判知识充足性的判断，并且若研判知识不足以支撑确凿的定论，则需要给出研判缺失知识的列表以及描述。

考核指标

1.至少支持典型网络、终端、蜜罐等安全监测日志的研判理解，对于缺失研判知识的告警数据，大模型应能够正确识别并提出所需补充的研判知识，要求整体识别准确率达到90%以上；

2.实现关键攻击线索实体、关联的自动化抽取，识别抽取准确率不低于85%；

3.实现自动化的行为语义基线构建，异常识别精度不低于 80%，误报率不高于 1%；

4.完成基于大模型的综合研判分析引擎一套，并提供源代码和项目开发使用说明文档。对于 LLM 相关的 PROMPT 需要有独立的详细说明文档；

5.给出测试数据集，原型系统和验证数据集至少支持和包含以下攻击中的 6 类：SQL 注入攻击、暴力破解攻击、文件上传攻击、WEB 远程代码执行攻击、系统远程代码执行攻击、远程命令执行攻击、文件读取/下载攻击、权限控制缺失、远程控制、流量代理、Webshell、反弹 shell、扫描探测。告警测试数据量不小于 4 万；

6.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

4. 基于大模型的软件代码安全检测及修复的技术研究

研究背景

随着人工智能和大语言模型技术的飞速发展，软件开发与维护的自动化水平不断提升。然而，随着软件规模和复杂度的增加，代码中的安全漏洞日益增多，导致网络攻击和数据泄露的风险加剧。传统的漏洞检测与

修复方法往往依赖于手动审查和工具扫描，效率低下且容易遗漏问题。基于 LLM 的技术研究能够有效地分析代码，识别潜在漏洞，并生成修复建议，从而提高软件的安全性和开发效率。

研究内容

1.收集和构建包含多种编程语言和漏洞类型的大规模代码数据集，以便为 LLM 的训练和评估提供基础；
2.研究和开发基于 LLM 的漏洞检测模型（工具），分析软件代码（二进制或者源代码）片段和软件代码库中存在的安全漏洞。模型应具备识别常见的安全问题，如 SQL 注入、缓冲区溢出和跨站脚本攻击等；
3.开发基于大模型的二进制文件分析工具，实现 90%以上的反编译代码可通过编译，能够对生成的反编译代码进行漏洞检测，漏洞准确率与误报率明显优于反汇编代码静态扫描，并提供源代码；
4.研究基于 LLM 实现漏洞代码自动修复措施。基于检测到的漏洞，模型将提出适当的修复策略，并自动生成修复代码。

考核指标

1.开发基于大模型的漏洞代码检测并修复工具，具备上传单个代码文件和源代码包的批量检测和修复，其中漏洞检测准确率和修复有效性不低于 80%；
2.识别常见的代码执行漏洞（缓冲区溢出/文件包含漏洞/命令注入/SQL 注入等）、信息泄露漏洞、拒绝服务漏洞、权限提升漏洞、跨站脚本攻击等 5 种以上的漏洞大分类；
3.实现开发语言（java、python、c/c++、golang 等）代码中的两种以上检测和修复，实现二进制代码的漏洞检查流程，并输出对应的结果报告；
4.完成基于大模型的漏洞代码检测并修复工具一套，并提供源代码；
5.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

5. 基于大模型的网络防御策略自主优化研究

研究背景

传统的网络防御机制通常依赖于预定义的规则和策略，这些规则往往基于历史攻击模式，缺乏对新型攻击的应对能力。希望基于大模型的自适应学习能力和复杂问题分析能力，在复杂多变的网络环境中实现防御策

略的自主优化，从而提高网络的整体防御能力和应对新型威胁的能力。

研究内容

1.防御策略风险点分析:利用大模型技术,针对海量行为数据(网、端)、安全设备告警、暴露面监测数据进行多源信息整合分析，结合网络拓扑结构，发现网络安全布防的弱点、风险点（如检测规则的缺失、防御策略不合理、设备部署位置不合理等）；
2.基于大模型的自主策略生成和优化：针对分析结果，生成优化后的检测规则、防御策略，自主调整网络布防，并自主进行防御策略有效性验证。

考核指标

1.设计防御策略“自主评估-自主优化-自主调整”模型，并实现 Agent 和原型系统；调研典型的企业网络拓扑，设计典型网络拓扑数量不少于 5 类，每类拓扑包括典型防御风险点不少于 6 类；在网络仿真环境中验证模型可用、有效；自主生成的防护策略和规则对拓扑中风险点防护有效性不低于 90%;
2.提供模型预训练、微调阶段的标签数据集不少于 800 条，以及与场景相关的 prompt 模板不少于 12 个;
3.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

6. 面向未知威胁的异常检测与仿真智能体技术应用研究

研究背景

当今世界的网络安全形势复杂且严峻，网络攻击频发，高级持续性威胁（APT)攻击日益成为全球网络安全领域的重大挑战。与传统的网络攻击相比，APT 攻击具有高级性、持续性和隐蔽性，通常由资源充足且技术精湛的攻击者（如国家支持的黑客组织）实施，目标包括政府机构、金融机构、科技公司等关键基础设施，不仅威胁到目标组织的机密信息，还对国家安全和社会稳定构成严重威胁。
传统的基于签名的威胁检测方法已难以应对 APT 攻击、0day 攻击等未知高隐匿威胁。近两年，大语言模型技术在网络威胁检测领域取得突破性进展， 特别是在模式识别和异常检测方向表现出巨大潜力。大模型不仅能够分析研判复杂的攻击行为告警，还能够自主的从网络流量、行为数据中提取高价值的特征维度，发现潜在的未知威胁。利用该思路，能够

从海量数据中更加高效的发现新型攻击、高隐匿行为模式。研究如何将大模型技术应用于未知威胁具有重要的理论意义和应用价值。
此外,在模拟仿真方面,基于大模型的 APT 仿真智能体通过模拟真实 APT 攻击的各个阶段,为安全团队提供接近真实的攻击环境,进行安全评估、事件响应演练和防御策略验证，有效提升组织的网络防御能力。

研究内容

研究目标 1：基于大模型的未知威胁检测技术研究

1.大模型技术在未知网络攻击（0day）检测领域的应用方法研究;
2.基于大模型发现高隐匿的异常行为模式:
采用大模型技术，针对海量行为数据进行自主建模，发现高隐匿威胁行为;
3.未知威胁检测模型结果可解释性研究:
研究如何提高大模型在未知威胁检测中的解释性，使得运营专家能够理解和信任模型的检测结果。

研究目标 2：基于大模型的 APT 仿真智能体技术研究

1.研究智能化 APT 攻击策略理论和架构，构建支撑人工智能决策的 APT 攻击知识库，实现针对特定目标场景的 APT 攻击策略生成;
2.研究面向 APT 攻击模拟的智能体框架，将网络杀伤链和 APT 攻击策略有机结合，构建精确的提示词以引导智能体模拟 APT 攻击行为模式;
3.研究基于大语言模型的 APT 仿真智能体构建方法，设计以大语言模型为决策大脑的 APT 仿真智能体,融合大语言模型提示词模拟各攻击阶段,依据 APT 知识库构建攻击框架，实现不同 APT 攻击手法的无缝接入和智能调用，复现 APT 攻击并发现潜在的安全威胁。

考核指标

研究目标 1 考核指标:

1.研发一套模型验证/原型系统，并提供源代码：基于课题思路，设计并实现 Agent 和原型系统，以验证理论模型可落地，并通过实际样例数据验证原型系统针对未知威胁检测的有效性，检测准确率达 90%以上，误报率达 10%以下;
2.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

研究目标 2 考核指标：

1.构建面向 APT 攻击的 LLM 微调数据集，数据集包含 10 个以上战役级别的威胁事件，包含 2 个以上 APT 组织的攻击行为方式；

2.构建一套面向 APT 攻击模拟的智能体行为框架，生成 APT 攻击决策的推理轨迹，并进行模拟仿真，并提供相关原型系统与源代码；

3.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

7. 大模型安全风险评估关键技术研究

研究背景

随着大模型（如 GPT-4、Llama 等）的快速发展，这些模型在各种应用中显示出了极大的潜力。然而，随之而来的风险也不容忽视。大模型可能会生成不当或有害的内容，这对用户的安全和隐私构成了威胁，因此，需要构建大模型业务风险评估标准集，实现对模型的安全能力评估。目前，风险评估提示词主要依赖人工收集与编写，极度依赖领域专家的专业知识以及专业经验，现有风险评估提示词的生成方式，不仅效率低下，而且难以覆盖所有潜在风险，无法快速针对特定风险分类完成提示词的储备工作。为了提高针对大模型风险评估覆盖维度，需要研究基于 self-instruct 等提示词的自动化生成技术，实现在大模型业务风险评估领域相关评估提示词的自动化生成；为了当前的研究主要集中在防御和检测传统的安全威胁，而对提示词攻击的研究较少。为了提升大模型评估的准确性，需要构建分类检测模型，实现目标劫持、越狱攻击以及正常提示词的分类检测。

此外，大模型的快速发展使其在内容生成和处理领域的应用愈加广泛。然而，伴随而来的内容安全风险也日益显著，如虚假信息、暴力和仇恨言论等。现有的内容检测和防御手段常常难以应对大模型生成内容的复杂性和多样性。因此，本课题旨在基于大模型技术与内容安全风险数据集构建防御检测模型，以实现对输入内容以及对话上下文的全面风险判断，确保内容生成和传播的安全性。

最后，多模态大模型在图像、文本、音频等多种模态上的卓越表现备受瞩目，然而，多模态大模型较大语言模型面临着更为复杂、多样的风险挑战，例如跨模态间的影响。因此，针对多模态大模型，需要设计多维度和视角的全面评估体系，更为准确和全面地评估多模态大模型的安全分风险。

研究内容

研究目标 1：智能对话场景下的风险评估提示词生成技术

1.研究针对智能对话场景下的风险分类与建模，构建种子风险提示词，如：政治与军事敏感问题、危害国家主权与形象、恐怖与暴力主义等通用风险，金融敏感问题、公司经营敏感问题等业务风险；
2.研究针对智能业务执行体场景下的风险分类与建模，构建种子风险提示词，如：金融场景下越权交易等风险；
3.研究基于 self-instruct 的风险提示词生成的 Prompt 工程技术；
4.研究基于特定实时信息的风险提示词生成的 Prompt 工程技术；
5.研究风险评估提示词的动态更新技术；
6.自动化大模型业务风险评估提示词系统原型。

研究目标 2：提示词攻击检测分类模型研究

1.研究构建提示词攻击样本库，包括：目标劫持、越狱攻击以及正常提示词；
2.研究提示词攻击特性，以及相关的特征提取与表示方法；
3.研究基于多语言基础模型实现训练，覆盖多种语言的提示词攻击分类；
4.研究设计训练多种分类模型，实现对提示词攻击的自动分类，评估不同模型性能；
5.研究基于最优分类模型，开展分类优化与改进工作，提升检测能力至业界先进水平；
6.研究实现模型的在线更新机制，实现针对新型攻击模式的及时检测。

研究目标 3：大模型内容安全防御检测模型研究

1.研究针对智能对话场景下的风险分类与建模，如：政治与军事敏感问题、危害国家主权与形象、恐怖与暴力主义等通用风险，金融敏感问题、公司经营敏感问题等业务风险；
2.研究积累多种内容安全风险相关的数据集，并研究设计风险防御检测的提示词指令框架；
3.研究针对单轮、多轮对话内容，以及针对风险问题输入的内容安全风险检测机制；
4.研究针对多语言内容安全风险的检测防御技术；
5.实现大模型内容安全防御检测模型原型。

研究目标 4：多模态大模型安全评估技术

1.研究国内外多模态大模型安全评估框架、基准及进展；
2.研究多模态大模型安全的评估维度、任务场景和评估视角，设计评估基准和评估数据集；
3.研究多模态大模型安全较大语言模型安全的异同，以及在多模态大模型中影响安全可信程度的因素；
4.研究不同的防御和安全对齐机制对多模态大模型安全的影响，以及不同对齐方案/可信性增强方案面对多模态大模型时的有效性。

考核指标

研究目标 1 考核指标：

1.智能对话场景与智能业务执行体场景下，风险分类与建模的具体风险项各不少于 10 种，通用风险至少包括政治与军事敏感问题、危害国家主权与形象、恐怖与暴力主义、低俗色情言论、受管制或控制的物质（枪械、毒品等）、诱导与不当言论等 6 类风险。业务风险至少包括金融场景、医疗场景、通信场景、销售场景等 4 种风险场景；
2.自动化风险生成的 Prompt 工程模板不少于 10 种，支持基于种子提示词生成上述 6 种通用风险提示词以及 4 种风险场景提示词，以及基于特定实时上下文信息生成风险提示词模版 1 种；
3.自动化生成风险评估提示词的合格率率不低于 90%；
4.实现自动化的大模型业务风险评估提示词系统原型；
5.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

研究目标 2 考核指标：

1.建立包含不少于 30000 个样本的提示词攻击样本库，覆盖目标劫持、越狱攻击和正常提示词三种类型；
2.研究实现基于多语言下的提示词攻击分类模型，支持语言种类不少于 5 种；
3. 设计并训练分类模型在测试集上的综合检测准确率不低于 90%；
4.实现提示词攻击的实时检测，单次检测时间不超过 1 秒；
5.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

研究目标 3 考核指标：

1. 内容安全防御检测模型中针对相关风险的检测准确率和召回率应分

别达到 90%以上；
2. 模型针对所有测试场景下的检测误报率应低于 5%；
3. 模型应支持至少两种语言（中英文），均达到 90%以上的检测准确率和召回率；
4. 联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份；
5.完成一套原型系统，并提供源代码；
6.检测内容安全风险覆盖，通用风险至少包括政治与军事敏感问题、危害国家主权与形象、恐怖与暴力主义、低俗色情言论、受管制或控制的物质（枪械、毒品等）、诱导与不当言论等 6 类风险。业务风险至少包括金融场景、医疗场景、通信场景、销售场景等 4 种风险场景。

研究目标 4 考核指标：
1) 提供一套多模态大模型安全可信评估框架；
2) 提供一份与评估框架对应的评估数据集和源代码，数据集包含的测试样本不少于 1500 个；
3)在多模态大模型（不少于三种）上进行评估测试，并输出报告一份；
4)联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项。

8. 面向网络流量的智能异常检测与分析技术研究

研究背景

随着网络规模的不断扩大和网络应用的日益增多，网络安全面临着越来越多的挑战。恶意攻击、数据泄露、网络钓鱼等安全威胁不断涌现，特别是恶意软件越来越多地使用了加密通信逃避网络异常检测机制，给网络运行和用户信息安全带来了严重的威胁。
针对加密网络流量和新型网络应用设计通用的识别和异常检测方案，基于全流量实现加密应用的精准识别和异常行为监测。可结合大模型技术，实现具有良好可解释性的加密或非加密流量分析。

研究内容

1.研究全流量中各种加密应用的流量特征和行为模式，提出通用的加密应用识别方法，进而针对特定加密应用提出通用的异常检测方法；
2.研究非加密流量场景中的有监督或无监督的异常流量分类方法；若为监督方法，则需支持将预训练模型快速适配到具体网络环境中；

3.探索大模型技术在流量分析中的应用，能够在如特征表征、模型构建与优化、流量检测、可解释等方面对比现有方法有显著提升。

考核指标

1.支持现网大流量场景下的实时识别，针对各类加密应用识别的召回率98%以上、误报率 1%以内，针对特定加密应用流量进行异常检测的召回率 98%以上、误报率 1%以内；
2.支持现网大流量场景下的实时识别，针对不特定的非加密应用流量的异常检测，召回率 98%以上，误报率 1%以内；
3.大模型相关的部分给出预训练、微调和测试数据集，微调和测试数据集包含 10 种以上实际异常流量数据，涉及到自然语言的部分如可解释相关内容必要时给出 PROMPT 模版，并附详细的说明文档；
4.形成一套原型系统，并提供源代码；
5.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

二、云计算安全领域

9. 微服务安全的关键技术研究

研究背景

微服务安全已成为云原生安全未来重要的发展方向，云原生环境中的微服务通信协议和潜在的攻击行为以及微服务调用链路异常攻击行亟待研究，需针对具体攻击场景设计异常检测模型，并提供相应的异常行为检测能力。

研究内容

1.研究面向云原生微服务与服务网格使用的通信协议（如 gRPC、GraphQL、Websocket、JSON-RPC、MQTT 等）及应用存在的安全风险和相应的威胁，如水平/垂直越权、认证失效、SSRF、DoS、Web 类攻击等；研究可应对前述风险和威胁的安全检测防护机制；
2.研究服务网格中微服务行为模式及基线模型，研究基于图学习的智能微服务异常行为检测机制。

考核指标

1.围绕 OWASP API 安全十大风险实现一套具有实用价值的检测模型和算法，作为异常检测引擎的核心以提升检测的准确率，能够支持检测项包括：
（1）不少于 5 种常见微服务通信协议的攻击行为（支持至少 5 类协议的检测规则，规则应能覆盖 OWASP API 安全十大风险）；
（2）不少于 3 种常见微服务调用链路异常行为，如调用参数异常、调用频率异常、调用前后顺序异常
等（支持至少 3 类检测规则，规则应能覆盖 OWASP API 安全十大风险）；
2.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份；
3.研制上述研究内容的原型系统 1 套，并提供源代码。

10. 面向云原生环境的红队技术研究

研究背景

云原生安全已经从安全建设走向了安全攻防，基于以攻促防的理念，应研究云原生环境下的攻击手法，提出对应的安全建议和加固措施。

研究内容

通过研究云原生环境下的攻击手法，对云原生环境以及云原生安全平台能力进行安全评估，

1.研究集群、镜像仓库、镜像等云原生攻击利用手法、云原生错误配置、漏洞在真实环境中的模拟方式；

2.结合攻击模拟BAS技术，研究云原生安全能力有效性评估技术。

考核指标

1.实现一套云原生攻击模拟的原型系统，具体包括真实攻击模拟、持续攻击、攻击编排、修复建议、安全评分功能，并提供源代码；

2.新增云原生风险无害化攻击手法不少于3个(最新版本或最新暴露)；

3.新增云原生风险（漏洞和错误配置）模拟手法不少于3个（最新版本或最新暴露）；

4.联合发表不少于1篇高水平论文（CCF-B或以上），联合申请发明专利1项，输出技术报告1份。

11. 公有云安全态势关键技术研究

研究背景

攻击者通常会查找租户的错误配置或弱凭证攻击公有云，因而通过研究公有云中的各类资源特性、风险情况，可实现以云凭证为基础的公有云环境安全性评估。

研究内容

1.研究主流大型云厂商的常见云服务的资源统一抽象，实现云服务知识图谱构建；

2.研究各类风险间的上下文关系，实现对风险的排序；

3.研究无代理模式下对工作负载进行漏扫方式；

4.研究基于审计日志的云用户行为分析及权限边界分析。

考核指标

1. 提供一套原型系统，并提供源代码，要求：

（1）支持主流至少三种公有云厂商常见云服务的资源统一抽象，云服务资源类型需包括，云主机、云原生服务、云函数Serverless、IAM访问控制、云网络、云存储（云数据库、对象存储等），最终实现云服务知

识图谱可视化构建;
（2）支持主流至少三种公有云厂商常见攻击路径分析，需至少满足以下攻击场景的攻击路径分析：
场景一：利用泄露的云凭据渗透测试;
场景二：利用实例元数据服务渗透测试;
场景三：利用容器逃逸展开攻击;
场景四：利用错误配置的存储桶展开攻击;
场景五：利用虚拟机逃逸展开攻击;
场景六：对企业内部网络、运维或管理内部网络进行攻击;
场景七：利用 IAM 服务进行渗透;
场景八：Kubernetes 集群中的渗透测试。
需针对风险的上下文进行分析，攻击路径需要以可视化的方式展示出来;
（3）支持主流至少三种公有云厂商无代理模式下对工作负载（至少支持云主机资源）进行漏扫的能力;
（4）支持主流至少三种公有云厂商的基于日志审计的云用户行为分析及权限边界分析（建立分析引擎，建立基线，主要针对 IAM 做分析），如发现挖矿勒索等需要产生事件告警;
2. 联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

三、数据安全领域

12. 新型数据水印技术研究

研究背景

传统数据库水印技术存在数据可用性较差的缺点，此外，当前缺乏针对纯文本类型或 API 水印技术。因此，针对数据库水印，需要孵化能够维持鲁棒性的同时提高数据可用性的创新水印技术，针对纯文本类型和 API 需要孵化可用的数字水印技术。

研究内容

1.针对数据库水印、文本水印、API 水印三种类型中的至少一种类型，研究具有实用价值的数据水印算法；
2.要求所研究的算法具有较好的隐蔽性、鲁棒性、安全性，能够尽可能减少对原数据的修改，提高数据可用性。

考核指标

1. 提供 1 套原型系统，并提供源代码，要求：
（1）支持数据库水印、文本水印或 API 水印中至少一种，构建水印系统；
（2）支持至少一种水印生成算法，能够生成安全性强、隐蔽性高的水印内容；
（3）支持至少一种水印嵌入算法，能够保证在对目标内容嵌入水印后，不降低鲁棒性的前提下保证数据可用性；
（4）支持至少一种水印溯源算法，能够根据疑似泄露数据，快速定位原数据所有者与泄露源，要求溯源算法复杂度尽可能低。
2. 联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

13. 可信数据确权的方法论与技术研究

研究背景

数据确权是数据要素市场化规模化的基础和前提。目前，我国在数据权属认定、流通交易规则、安全能力规范等方面的设计尚处于起步阶段，对政策法规的技术落实与补充完善缺乏系统化的技术解析。

研究内容

1.基于国内外主流数据权属模型(如三权分置模型)，针对典型的数据要素流通场景，系统梳理场景中的数据权属情况，并明确场景中的权属边界模糊点；

2.针对梳理出的数据权属模糊点，提出系统性的权属鉴别方案。可以通过限制流通方法实现，也可以通过在流通场景中扩充审计能力等其它方法实现。该鉴别方案需形成一套可落地的技术路线图,且须与现行法律法规、行业政策相衔接,为数据合规流通提供明确的制度规范和可执行的技术方案。

考核指标

1.提供一份技术调研报告，对至少三种数据要素典型场景中涉及的数据完成权属分析；并形成符合当前法规政策权属模型的分析方法论；

2. 形成一套原型系统，与调研报告相匹配，并提供源代码，要求：

（1）能够包括典型数据要素场景中的数据确权全过程；

（2）能够以技术手段而非约定等方式对场景中的权属模糊点进行分析处理；

3. 联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

14. 数据安全大模型技术与应用研究

研究背景

数据安全合规和业务需求越来越强烈，现有的安全技术能力已无法满足相关要求，需要通过大模型等方式能够提升现有数据安全技术能力。

研究内容

基于大模型进行以下技术研究：

1.研究基于 LLM 的半结构化、非结构化数据识别及分类分级技术；

2.研究基于 LLM 的文档无偏水印生成及溯源技术，不改变语义等情况下进行合同等文档内容水印添加，能保证严谨和合规，具有较强的鲁棒性，能够应对一定程度的文本修改攻击；

3.研究基于 LLM 的自动化脱敏策略管控技术，根据数据分类分级结果、重要数据、用户自定义、数据关联性情况等，生成符合用户数据情况及偏好的自动化的脱敏管控策略；

4.研究基于 LLM 的数据风险评估技术，挖掘数据之间的关联性，进行数

据泄露风险评估，脱敏后由于数据关联导致数据泄露等情况的风险评估；
5.研究基于 LLM 的数据脱敏、加密算法识别技术，可根据预置加密算法列表、脱敏算法列表，进行脱敏后数据算法识别；
6.研究基于 LLM 的图片数据脱敏技术。

考核指标

1.研制一套部署在“绿盟风云卫大模型”上的原型系统，基于大语言模型与 Langchain 等应用框架实现可离线部署的 RAG 与 Agent 系统，并提供源代码；
2.支持多模态的非结构化数据分类分级、脱敏、水印能力，分类分级准确率能达到 90%以上，文档水印能够实现文档语义不影响，合同等保障合规性。数据分级分类对象包括文档、图形、视频、音频等多模态数据，具体的支持的数据格式类别：常见文档：TXT、PDF、WORD、XML、JSON、Email、HTML 等 图像：JPG、PNG、BMP 等 视频：MP4、AVI、WMV 等 ；音频：MP3、WAV 等。图片脱敏技术的对象包括人脸、车牌等常见个人敏感信息等，可以支持自定义图像脱敏方法；
3.能够输出十种以上的风险评估方案，验证技术可行性；
4.能够具备脱敏、管控策略的输出；
5.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

15. 隐私信息梯度泄露检测技术研究

研究背景

随着互联网生活的快速发展，个体生活已经完全融入互联网。然后个体隐私信息在不经意间被各大网络平台、APP、社交网站不同程度的梯度泄露，就单体而言都是符合隐私保护，但经过多个平台的梯度泄露，则导致个体的完整信息被推测还原，甚至导致互联网凭证、金融财产被窃取。因此需要技术发现隐私梯度泄露问题，保证互联网健康生活。

研究内容

1.研究国内 APP、大型网络平台、民生服务平台等个人信息梯度泄露点位、类型等挖掘技术；
2.研究泄露点位信息的自动化获取与和抗干扰技术；
3.根据泄露点位信息进行数据分析和推理预测，还原梯度攻击链路并实现自动预警。

考核指标

1.覆盖国内外 APP 83 个以上，大型网站和民生服务系统 60 个以上；
2.发现国内梯度泄露攻击路径若干，国外至少 20 个以上；
3.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份；
4.研制上述研究内容的原型系统 1 套，并提供源代码。

四、安全攻防领域

16. 智能自动化渗透技术研究

研究背景

随着网络安全威胁的不断演变，红队（Red Team）需要更高效的渗透测试和攻击策略。当前，传统的攻击手段和技术文档更新速度慢，难以应对快速变化的防御措施。本项目旨在研究如何利用大规模语言模型(LLM)自动化决策支持系统，通过提取信息并构建决策模型，自动化选择红队下一步攻击策略，提高渗透测试的效率和效果。

研究内容

1.研究特定攻击场景下的攻击技战法，收集并分析构建技战法知识库，分类和标记每种技战法的特征与适应环境；
2.研究设计决策模型，输入包括当前环境信息、目标系统特性及已使用的攻击手段；
3.研究决策模型输出最优攻击策略推荐，提供多种候选方案并排序；
4.研究落地自动化渗透决策建模系统，集成技战法知识库与决策模型；
5.基于自动化渗透决策系统开展靶场实战验证，分析实际运行案例，改进决策模型，提高系统适用性；
6.鼓励在自动化渗透决策建模系统中接入已有技战法的攻击原子，实现对特定靶场的攻击能力覆盖。

考核指标

1.确定决策模型将覆盖的攻击场景，包含技战法覆盖攻击阶段不少于 4 种（攻击阶段至少覆盖初始访问、横向移动、权限提升、权限维持 4 个阶段），攻击技战法数量不少于 20 个；
2.决策模型推荐策略的准确率不低于 80%；
3.针对指定靶场环境的自动化渗透覆盖率不低于 80%；
4.研制上述研究内容的原型系统 1 套，并提供源代码；
5.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

17. 暗网空间资源要素测绘关键技术研究

研究背景

面向暗网的勒索组织及其勒索事件、重点行业和企业数据泄露、暗网空间黑市交易猖獗等监测和预警需求，行业主管部门急需支持暗网空间数据泄露、勒索组织及其勒索事件、暗网服务节点等监测预警能力产品或解决方案。

研究内容

1.针对当前暗网空间资源要素藏匿、变换等匿名化机制所产生的测绘难题，研究有效的暗网空间资源要素测绘方法，支撑暗网空间威胁情报要素发现和行为识别；
2.研究暗网情报监测、数据和实体关系抽取、情报检索和预警分析等关键技术，具体包括：
（1）研究高效的暗网资源要素监测模型和方法；
（2）研究暗网隐藏服务探查方法；
（3）研究暗网空间威胁情报侦测识别方法。

考核指标

1.研制 1 套暗网数据交易监测原型系统，支持本地化和 SaaS 服务模式部署，并提供源代码；
2.支持 Tor 和 I2P 暗网，暗网服务节点每天不少于 1 万；
3.覆盖全球 1000+重要暗网站点、2000+Telegram 群组和频道，每天新增数据不少于 3 万+，更新频率不少于每天 1 次；
4.具备电信行业数据安全监测和暗网威胁情报查询服务能力；
5.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

18. 大规模网络风险监测技术研究

研究背景

对在互联网中暴露的资产进行监测，识别其中存在的漏洞、不当配置等风险，建立模型评估网络风险总体状况并预测其发展趋势。

研究内容

1、研究大规模网络监测关键技术；
2、研究漏洞识别关键技术；
3、研究不当配置识别关键技术研究；
4、研究网络风险评估模型关键技术研究。

考核指标

1、研制上述研究内容的原型系统 1 套，并提供源代码；
2、支持以插件形式来增强、完善风险识别能力，提供插件 10 个以上；
3、发现真实有效的风险事件 1000 条以上。

19. 基于强化学习的恶意软件协议逆向技术研究

研究背景

知晓恶意软件通信协议是攻击检测和监测的重要前置技术，然而恶意软件不同于常规正常软件，通信协议明确完整，客户端和服务端软件程序健全，在攻击防御过程中往往获得的是单边恶意程序或仅有恶意程序的通信流量。还原恶意软件完整的通信协议就成为当前攻击检测领域面临的首要安全问题。

研究内容

1. 探索强化学习在网络协议分析和协议异常领域的应用技术；
2. 研究基于恶意软件客户端源代码和攻击协议片段，还原通信协议；
3. 研究多智能体协调探测恶意软件或特定服务系统的通信协议格式、交互状态机等通信机制。

考核指标

1.能够还原 5 种以上的恶意软件通信协议（自定义通信协议、在通用协议之上魔改的协议），还原拟真度 80%以上；
2.能够发现 2 个以上常规通信协议的隐匿机制；
3.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份；
4.研制上述研究内容的原型系统 1 套，并提供源代码。

20. 目标指纹探测与识别技术研究

研究背景

在现实攻击事件中攻击基础设施，通常与具体的事件相关联，而多个事件是否并案是否牵连都很难界定，针对现有事件的攻击基础设施的拓线手段也非常有限，为了更好的发现失陷主机，消除安全隐患及时止损，需要研究面向攻击基础设施的发现技术以及相似性识别技术，深入探索

APT 攻击基础设施及其属性，进行探测、同源性分析、组织资产画像绘制等关键技术研究。

研究内容

1. 研究 APT 组织攻击基础设施的指纹特征表征技术；
2. 研究基于主被动流量探测的攻击资产同源性关联与溯源分析方法；
3. 研究基于指纹识别分析的 APT 组织资产追踪拓线方法；
4. 研究 APT 组织攻击资产画像绘制技术。

考核指标

1.支持对主流 APT 组织资产指纹的全面表征，确保对不少于 5 个主流 APT 组织的资产具有高效的指纹识别与探测能力；
2.具备对攻击资产进行指纹相似性分析，明确 APT 组织基础设施的相似性关系。能够在现有 APT 组织资产信息的基础上，拓线不少与 20 个 APT 组织未知资产；
3.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份；
4.研制上述研究内容的原型系统 1 套，并提供源代码。

21. 自动驾驶仿真技术安全技术研究

研究背景

随着智能汽车的普及，以及自动驾驶汽车的商用化推进，针对自动驾驶汽车的基础技术研究以及安全风险研究已经成为研究热点，尤其是在传感器采集、智能决策方面的安全风险，则需要自动驾驶领域专业的模拟环境以及专业能力来支撑。本课题旨在基于自动驾驶仿真模拟技术，实现对软件、网络、传感器及决策算法四方面的攻击场景。

研究内容

1.研究自动驾驶仿真环境，完成对自动驾驶仿真环境的软硬件系统搭建；
2.分析软件环境以及网络环境的安全风险，发现针对软件、网络的漏洞并完成相应的技术研究报告；
3.分析传感器和智能决策算法的安全风险，发现其中的漏洞并完成相应的技术研究报告。

考核指标

1.完成满足要求的软硬件系统原型系统，实物部分需要有雷达、摄像头，其数据可以输入至仿真平台，满足虚实结合，并提供源代码；
2.模拟软件、网络漏洞不少于5个；
3.模拟传感层、智能决策方面的漏洞不少于5个；
4.完成雷达、摄像头两类传感器与仿真平台的数据互通；
5. 联合发表不少于1篇高水平论文（CCF-B或以上），联合申请发明专利1项，输出技术报告1份。

22. 多模态异构数据融合的智能制造融合安全处理关键技术研究与应用

研究背景

工业控制系统资产种类多，各类工业协议复杂，各类工业场景下差异大与业务关联度高的特点，网络安全对于安全事件的处置缺乏基于业务的风险价值判别的问题，容易发生工业系统的攻击，应对相应的工业协议和相关风险进行研究。

研究内容

研究针对工业协议快速识别、深度解析的方法，实现对于工业资产基于业务与安全的画像，实现针对工艺过程要素的安全监测及网络安全波动对业务的影响动态评价系统。

考核指标

1.支持不少于30种工业协议的深度解析，至少包含：ICCP、cip safety[基于CIP协议]、CODESYS、RSSP-1、ANKONG500、IEC-61850-PRES、RSSP-2、DDP、IEC CMS、LSV2、Focas、hls_macs5[t]、hls_macs5[u]、hls_macs6[t]、hls_macs6[u]、FOX、GSK-LINK、GE-SRTP、SUPCON UDP、HOLLYSYS-MACS、LSV2/DNC OPT#18、PCWorx、Profinet（MRP）、Profinet（PTCP）、TRDP、SECS、PROFIsafe、EGD、SRTP、SLMP；
2.至少支持四类主流控制器类别：PLC、DCS、机器人、数控机床，至少支持10个主流品牌控制器类型的识别；安全评价系统需针对关键工艺的风险值给出定量分析，建立网络攻击行为与业务影响的动态关联及可量化的指标项，可适配于石化、汽车制造、冶金等行业的应用场景，相关的安全处置有效性大约90%，业务扰动小于1%；
3.提供相关协议的分析报告1份，提供相关解析代码（python语言支

持）；
4.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。

23. 基于 FPGA 的高性能正则匹配的技术研究

研究背景

当前网络发展迅速，网络流量也越来越大，因此对大流量高性能检测的需求，越来越迫切。其中对大流量的深度内容检测又是非常消耗处理能力的，当前使用传统 CPU 已经无法满足大流量下深度内容的检测性能要求。
因此需要研究其他的方案，来提升深度检测的性能，如使用 FPGA 的方案。

研究内容

研究面向大流量的深度内容检测场景下基于 FPGA 的高性能正则匹配技术。

考核指标

1.设计 FPGA 高速正则匹配的设计方案,能够支持 5 万正则场景下 50Gb/s 检测性能；
2.研制原型系统 1 套，并提供源代码；
3.能够在 bps 等测试环境下（如企业流）达到性能要求；
4.联合发表不少于 1 篇高水平论文（CCF-B 或以上），联合申请发明专利 1 项，输出技术报告 1 份。