基于大模型的工业安全平台

引言/导读

北京金睛云华科技有限公司(以下简称“金睛云华”)，2016 年成立于北京，致力于成为以 AI 智体为核心的新一代网络安全产品与引擎提供商。创始人及核心团队主要来自于清华大学 KEG 实验室、华为、启明星辰、东软等一线 AI 科学家与网络安全专家，在网络安全和人工智能领域有二十余年丰富的经验和技术积累，已申请 100 余项基于人工智能的网络安全发明专利。在北京与沈阳建立了研创中心与安全运营中心，并在北京、上海、广州、深圳等二十余个省市建立了营销与服务网络，已形成研发、创新、营销与服务覆盖全国的战略布局。

金睛云华作为国内最早专注人工智能和 NDR 领域的公司，实施了创新驱动发展战略，形成了以 AI 检测智体为核心的 XDR 解决方案和以 AI 运营智体为核心的智能安全运营(AISecOps)两条业务线，并在两个方向上持续拓展，践行金睛云华在用户数字化转型和智能化升级时代的使命一安全智体平权，践行“普惠安全”。金睛云华打造的产品体系涵盖以安全大脑「CyberCopilot」赋能的工业互联网威胁检测和智能安全运营产品族，产品包括云鉴·高级威胁检测系统(ATD Pro)、云踪·网络流量回溯审计系统(TFS Pro)、云晰·加密流量检测系统(ETD Pro)、云图·网络安全智能中心(CC Pro)以及云智网络安全大脑(BOC)。以大语言模型和智能体技术为核心，金睛云华将以安全大脑赋能高级威胁检测和智能安全运营作为公司未来战略发展方向，完成公司从智能威胁检测解决方案到智能安全运营解决方案的业务模式演进，最终以安全大脑「CyberCopilot」提供商的角色赋能网络安全行业。

智能工业互联网威胁检测系统作为一个基于人工智能技术的网络安全检测工具，实时捕获和分析网络流量，通过机器学习、集成学习、深度学习、大语言模型（LLM）等技术，实时发现网络中异常行为和潜在的安全威胁。系统能够根据新的数据和威胁模式不断更新和优化智能化检测模型，以保持对新兴威胁的有效识别能力。同时，通过自动学习建立流量白模型和识别正常行为模式，用于增强对网络威胁的检测和响应能力。系统提供直观的数据分析和可视化界面，帮助安全分析师快速理解和分析安全威胁，减轻安全团队的工作负担。

随着网络攻击的不断增加和复杂化，网络安全面临着大数据和智能化的挑战，传统的基于规则的安全检测技术难以应对高级持续性威胁、零日漏洞、恶意加密流量的攻击。大语言模型（LLM）对于复杂网络环境理解能力高，通过学习大量的数据、自动提取数据的特征和规律性，发现复杂环境下隐蔽的攻击模式。经过微调后的检测模型特别是对于恶意代码变种、加密流量检测和用户白流量建模能力强大，大大提高了检测的效率和准确性，降低误报和漏报率，提升了安全防护的能力。大语言模型（LLM）也具有强大的适应能力，能不断学习和自动适应新的数据和场景，用于检测新型网络攻击。

一、 关键词

安全大模型，智能体，程序语言大模型、工业互联网智能安全运营

二、 测试床项目承接主体

2.1. 发起公司和主要联系人联系方式

北京金睛云华科技有限公司，胡永亮 18698814130

2.2. 合作公司

【说明以合作伙伴身份参与的公司。】

三、 测试床项目目标

1、本地化工业安全大模型的训练：大语言模型通常具有数亿甚至百亿、千亿级的参数，需要大量的 GPU 计算资源用于本地化训练。如何在资源有限的情况下保证其稳定运行并提供低延迟的服务，是一个巨大的技术挑战。

2、模型的安全性和隐私保护：在处理敏感数据时，如何确保大模型本身不被用于泄露隐私信息或进行恶意操作，以及如何避免模型被训练出偏见或歧视性行为，是必须考虑的安全性问题。

3、智能体的构建与优化：构建能够通过拖拽等可视化界面进行交互的智能 Agent，需要将复杂的任务分解为可由大模型处理的子任务，并且要确保这些子任务能够有效地组合起来完成整体任务，这需要高度的抽象和设计能力。

4、多源数据的整合与统筹分析：需要整合来自不同来源的数据，如情报系统、资产系统等。如何确保数据的准确性、一致性和时效性，以及如何有效地利用这些数据进行综合分析，是系统实现的关键。

5、智能化的工业互联网告警分析与降噪：需要能够智能地分析安全告警数据，区分攻击的真实性和危害程度，并进行有效的降噪处理。要求系统具备深度学习和自然语言理解的强大能力，能够从大量的告警信息中提取关键特征，并进行准确的判断。

四、 测试床方案架构

4.1. 测试床应用场景

1、围绕大语言模型的智能特性形成基于大语言模型技术的智能运营分析系统，以智能体架构为底座的能够对接内置离线安全大模型或第三方大模型的开放式平台，提供对话式可视化界面和 API 调用两种交互方式，能够基于思维链模式通过可拖拽可视化界面交互方式进行智能体构建，智能体调用大模型来构建智能化的安全运营分析。

2、系统具备基于大语言模型智能安全告警数据分析研判功能，通过思维链调用安全大模型，并结合情报系统、资产系统等工具进行数据整合，智能化统筹分析，达成告警深度分析研判，直接给出告警是否攻击成功、攻击失败等研判结论，完成告警降噪。最终，形成告警降噪、网络溯源、知识问答、告警解读、攻击载荷分析、自动化安全报告生成等关键能力。

4.2. 测试床架构

基于大模型的网络安全运营方案可分为 4 个层级和 2 个框架。方案逻辑框架如下图所示：

数据采集层：收集网络流量、日志数据等待分析数据。

数据处理层：对采集的数据进行预处理，用于后续分析。

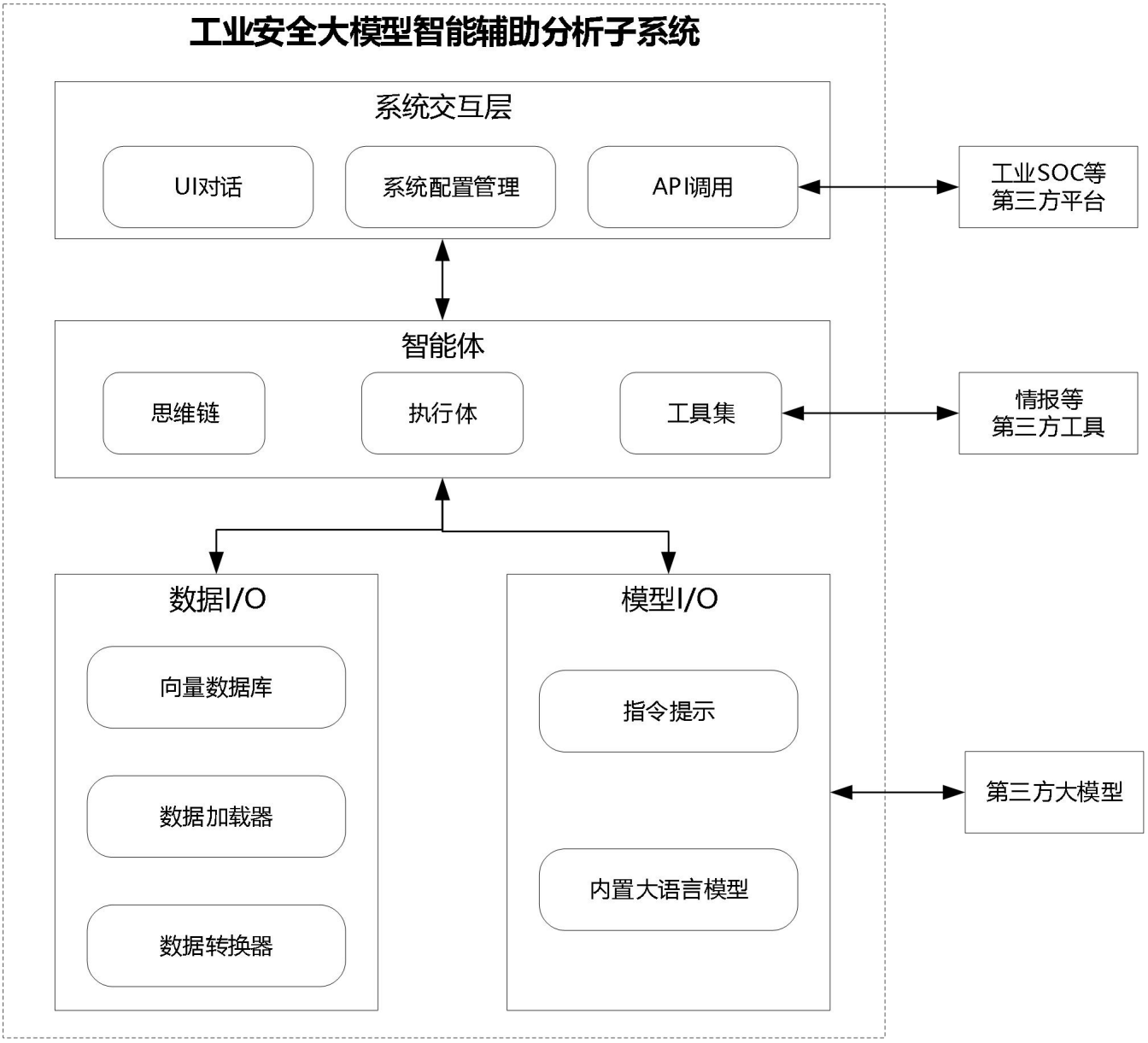
AI 算法模型： 根据业务需求，通过安全运营大模型、大模型深度威胁检测与载荷分析、多场景 AI 检测小模型等功能为上层应用提供计算基础。

业务运营层：通过利对数据进行分析研判，实现告警降噪、攻击溯源、响应处置、态势展示、智能报告等业务场景，也可以自定义业务场景。

容器化流量检测引擎基础框架：构建基于容器技术的流量检测引擎框架，实现不同类型的检测引擎可以灵活更新与扩展。

融合大模型的 Langchain 框架：通过提示词、数据解析器完成大模型对接，通过智能 Agent 和链（Chain）完成各类数据、工具、流程和大模型决策调度整合，实现大模型可以灵活更新于扩展，支撑业务场景的设计与处理流程自定义。

4.3. 测试床方案



架构具体描述：

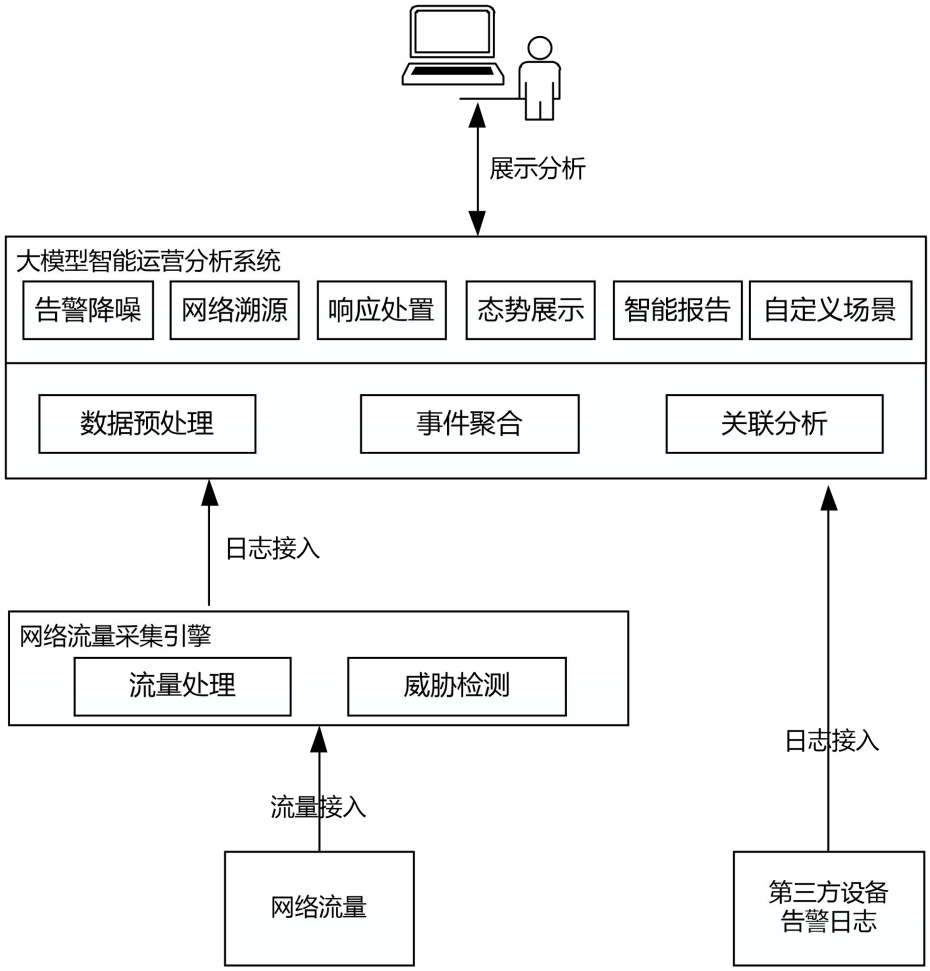
智能体：基于思维链进一步串联工具（Tools），从而将大语言模型的能力和本地、云服务能力结合。对于不同的告警数据处理场景，使用不同的智能体。告警解读从告警的攻击者、受害者、攻击载荷等多维度进行分析，在大模型进行分析过程中，根据不同上下文智能决策思维链流转和工具交互逻辑，完成全面的告警解读场景。

思维链：用于串联 模型 I/O 和数据 I/O 模块，以实现串行化的连续对话、推测流程

模型 I/O：管理大语言模型（Models）及其输入（Prompts）和格式化输出（Output Parsers）。

数据 I/O：主要用于建设私域知识（库）的向量数据存储（Vector Stores）、内容数据获取（Document Loaders）和转化（Transformers），以及向量数据查询（Retrievers）。

基于大模型的网络安全运营方案业务流程如下图所示：



用户可以查看大模型智能运营分析系统相关业务数据并进行交互分析。

网络流量采集引擎接入网络流量，包括在线流量和离线 PCAP，进行检测，检测后产生的日志可以发送至大模型智能运营分析系统进行进一步的关联分析。

同时，大模型智能运营分析系统通过 SYSLOG 等方式接收第三方告警日志等日志类数据，对所有接收的日志数据进行预处理、事件聚合、关联分析。然后基于大语言模型的告警事件研判分析，以实现告警降噪、攻击溯源、响应处置、态势展示、智能报告。

第三方设备可以支持多厂商的安全探针设备，包括但不限于威努特、知道创宇等。

4.4. 方案重点技术

【介绍测试床方案的重点技术，与现有国内外技术对比。】

国际现状是，人工智能在网络安全领域的应用日益重要。通过自动化、智能化的方式，人工智能可以帮助识别、防范和应对各种网络安全威胁。微软推出了基于 OpenAI 的 Security Copilot 系统用于安全数据分析，提升网络安全防御的效果和效率。谷歌云推出 Security AI Workbench，这是业界首套由谷歌安全大模型 Sec-PaLM 提供支持的可扩展平台。这套新安全模型针对安全用例进行了微调，并结合谷歌强大的安全情报，包括谷歌的威胁态势可见性，Mandiant 关于漏洞、恶意软件、威胁指标与恶意黑客行为模式的一线情报。

国内现状是，在人工智能小模型时代，真正将 AI 模型应用到网络安全领域的公司不多，在大模型火热以来，一些大型安全公司才开始思考这领域的技术方案，但还停留在宣传和实验室阶段，还没有几家公司能够将小模型技术大规模产品化，将大模型技术工程化，在具体的安全场景能够有效解决安全问题的公司更是屈指可数。

智能威胁检测子系统通过旁路镜像和高性能采集技术，系统对网络流量进行实时解码和元数据提取，建立完整的日志、协议、数据包全字段索引库。利用 Transformer 架构的大语言模型（LLM）学习大量的数据、自动提取数据的特征和规律性，采用特定或自有的大规模高精度标注的威胁数据进行模型精调，发现复杂环境下隐蔽的攻击模式。基于 Kill Chain 框架，以实现对攻击阶段的全覆盖，发现更多的攻击威胁事件，减少盲点，并将不同阶段的攻击事件进行串联。能够对攻击事件的详细信息进行溯源分析，对攻击源、攻击过程、攻击扩散面、被攻击的业务系统、攻击的恶意软件功能和危害等情况进行深入的分析，帮助安全团队更好的判定攻击的性质、手段和影响，确定合理的应对措施。同时能够与第三方安全防护设备联动响应，实现对威胁的阻断处置。并能够与大数据安全分析子系统联动，实时上传日志、事件等相关信息，为大模型智能辅助分析子系统提供有力的数据支撑。

4.5. 方案自主研发性、创新性及先进性

【介绍测试床方案的自主研发性、创新性及先进性。】

1、本地化大模型的训练问题：大语言模型通常具有数亿甚至百亿、千亿级的参数，需要大量的 GPU 计算资源用于本地化训练。如何在资源有限的情况下保证其稳定运行并提供低延迟的服务，是一个巨大的技术挑战。

2、智能体的构建与优化：构建能够通过拖拽等可视化界面进行交互的智能 Agent，需要将复杂的任务分解为可由大模型处理的子任务，并且要确保这些子任务能够有效地组合起来完成整体任务，这需要高度的抽象和设计能力。

3、多源数据的整合与统筹分析：需要整合来自不同来源的数据，如情报系统、资产系统等。如何确保数据的准确性、一致性和时效性，以及如何有效地利用这些数据进行综合分析，是系统实现的关键。

4、智能化的告警分析与降噪：需要能够智能地分析安全告警数据，区分攻击的真实性和危害程度，并进行有效的降噪处理。要求系统具备深度学习和自然语言理解的强大能力，能够从大量的告警信息中提取关键特征，并进行准确的判断。

5、工业安全平台的大模型组件采用 MOE 架构，通过模型调度路由将不同的输入数据调度给对应的专有模型，通过将大规模参数的单一大模型划分为多个中小规模大语言模型，每个模型负责专一的业务场景，比如工业安全检测大模型、安全运营大模型、工业知识经验大模型等。

4.6. 方案安全风险控制

【方案是否有安全风险应对模型？哪些是最关键部件？哪些是最脆弱的部件？最易受攻击的部件？】

模型的安全性和隐私保护：在处理敏感数据时，如何确保大模型本身不被用于泄露隐私信息或进行恶意操作，以及如何避免模型被训练出偏见或歧视性行为，是必须考虑的安全性问题。模型构建过程中，使用安全围栏（SafeGuard）技术，确保意识形态和社会伦理正确。

五、 测试床实施部署

（正文 小四 宋体。行距 1.5 倍行距）

5.1. 测试床实施规划

【明确测试床实施的时间规划。】

基于大模型的网络安全运营方案整体建设周期为 1 年，建设经费包括硬件成本和软件成本。为避免一次性投入过大，可以采取分阶段、分步骤建设方式，逐步实现基于大模型的网络安全运营与检测建设。

5.2. 测试床实施的技术支撑及保障措施

【说明测试床实施的技术支撑及保障措施。】

5.3. 测试床实施的自主可控性

【说明测试床实施的自主可控性。】

六、 测试床预期成果

（正文 小四 宋体。行距 1.5 倍行距）

6.1. 测试床的预期可量化实施结果

【明确测试床的预期可量化实施结果，针对测试项。】

研制内置安全大模型和安全运营智能体的原型系统，系统内置本地化训练精调的具备专业安全知识的安全大模型。提供支持 10 类以上威胁检测的安全检测大模型，提供具备 5 类以上智能体的安全日志分析研判的运营大模型，运营大模型的参数规模不低于 300 亿；提供分布式的内置安全大模型的原型系统，支持 10wEPS 处理能力，内置安全辅助智能分析助手，智能助手支持安全知识开放问答、告警解读、告警处置、告警关联、载荷分析等，可以进行对话式安全数据分析；系统支持通过拖拽式界面操作完成自定义安全智能体。

6.2. 测试床的商业价值、经济效益

【说明测试床的商业价值、经济效益。】

工业互联网接入的设备类型多，安全产品类别多。产生的告警日志数量庞大，采用基于大模型的工业安全平台能够显著提升运营效率。根据业界经验，一名安全服务工程师 1 天能够分析 500 条日志，对于一天动辄几十万甚至百万的日志的情况很常见。我们以每天 5000 条日志测算，需要 10 人的安服团队。一套内置大模型的工业分析平台 24 小时工作，每秒处理 6 条日志，即：10(人) * 500(条)--VS—6 * 60 * 24（1 套）。

如此推算 1 套基于大模型的安全平台，可产生 100 万的经济价值。（注：每个安服工程师 10 万薪资/年）。

6.3. 测试床的社会价值

【说明测试床的社会价值。】

工业安全平台的应用能够显著提高组织安全性，安全运营平台通过持续监控和分析网络活动，提高组织对安全威胁的防御能力。通过识别和缓解安全风险，减少潜在的财务损失和声誉损害。保护敏感数据不被泄露或滥用，维护个人和企业的隐私权益。同时，提升应急响应能力，在安全事件发生时，能够快速响应并采取措施，减少安全事件的影响。保证业务连续性，确保关键业务系统的稳定运行，减少因安全问题导致的业务中断。

6.4. 测试床初步推广应用案例

【证明测试床的对外服务性。】

当前在东北大学已经进行初步应用。

七、 测试床成果验证

（正文 小四 宋体。行距 1.5 倍行距）

7.1. 测试床成果验证计划

【明确测试床成果的验证计划。】

7.2. 测试床成果验证方案

【明确测试床成果的验证方案，涉及对设备单元、信息系统、关键技术、多样场景等的全面验证。】

八、 与已存在 AII 测试床的关系

【请说明测试床和之前已经审批的测试床的关联关系，以防重复申请。并考虑互操作性。】

（正文 小四 宋体。行距 1.5 倍行距）

九、 测试床成果交付

（正文 小四 宋体。行距 1.5 倍行距）

9.1. 测试床成果交付件

【请确定测试床需要提供的交付件：阶段目标及阶段交付件、最终交付件、测试床成功标准。测试床成果需包含至少一项知识产权，包括专利、商标、版权等。】

提供一套可验证的原型系统，配套提供相关技术方案及相关的专利 1 项。

9.2. 测试床可复制性

【明确测试床可复制于哪些行业、哪些场景。】

可以复制推广到运营商、国家监管单位。用于大范围的关键基础设施的安全防护。

9.3. 测试床开放性

【证明同行企业或者上下游企业的共同参与度。】

和华为人工智能中心在大模型训练方面进行深度合作。

十、 其他信息

（正文 小四 宋体。行距 1.5 倍行距）

10.1. 测试床使用者

【明确非发起方的公司可以使用测试床程度，以及相关的要求和限制条件。】

10.2. 测试床知识产权说明

【请清晰说明谁对测试床的建设、运营以及使用拥有产权。】

10.3. 测试床运营及访问使用

【请说明测试床如何部署、运营以及访问使用。】

10.4. 测试床资金

【请列出预估的资金需求和资金来源。】

基于大模型的工业安全平台方案所需软件包括网络流量采集引擎软件、大模型智能运营分析系统软件，各软件的成本估价如下：

网络流量采集引擎软件成本估价：

名称	功能介绍	数量（套）	单价（万元）	合计（万元）
流量处理模块	具备数据采集、数据过滤、数据还原功能。	1	10	10
威胁检测模块	具备特征检测、行为检测、威胁情报检测、AI 模型检测能力，支持 Shadowsocks 流量、VPN 流量、恶意加密、SQL 注入、Webshell、暗网流量、DGA 域名、DNS/ICMP/HTTP 隐蔽隧道、恶意代码变种等进行威胁检测。			
业务应用模块	具备攻击链分析、关联与溯源、告警通知、设备联动响应、数据外发等能力。			

基于大模型的工业安全平台系统软件成本估价：

名称	功能介绍	数量（套）	单价（万元）	合计（万元）
数据采集模块	负责接收各类设备的网络协议元数据、告警日志等数据。	1	200	200

数据流计算模块	负责对接入的数据做大模型检测、大模型辅助检测等日志、告警检测分析。具备大模型本地化训练能力。			
数据存储模块	负责将各类数据进行存储，并提供数据的检索、更新能力。			
智能体模块	负责通过链(Chains)、工具集等组件完成不同业务分析场景的智能体构建，支撑智能化安全告警辅助运营。			
系统交互模块	负责提供人机交付可视化界面，可以通过对话方式执行智能辅助运营任务，提供 API 接口和其他系统对接，赋能智能辅助运营能力。			

10.5. 测试床时间轴

【请列出测试床的建设、部署和运行时间段。预估关键的时间点。标明是短期项目还是需要多年的研究项目。】

序号	时间节点	工作内容
1	2024 年 6 月-2024 年 10 月	大模型研发，系统研发
2	2024 年 11 月-2025 年 1 月	用户实验局部署测试运行
3	2025 年 2 月-2025 年 4 月	针对试用效果整改完善
4	2025 年 5 月-2024 年 6 月	项目总结，结题

10.6. 附加信息

【请列出其他有价值的信息以便委员会更好的对测试床提案申请进行评估和决策，如可应用复制的行业等。】