智算中心的综合安全体系与方案设计

闫崇喆

（北京飞机维修工程有限公司，北京 100000）

摘 要 智算中心作为人工智能发展的关键基础设施，为AI模型训练、推理等提供强大算力支持，在数字化进程中扮演着举足轻重的角色。但随着AI技术的广泛应用，智算中心面临的安全挑战日益凸显，涵盖数据、模型、部署推理及监管合规等多个层面。智算中心安全管理需覆盖全生命周期，既要强化技术防护能力，也要严格落实监管合规要求。基于此，本文为企业部署安全、合规的AI推理服务提供了具体、可操作的实施指南与治理框架，以期为相关人员提供参考。

关键词 智算中心；安全体系；模型防护

中图分类号：TP3 **文献标志码**：A **DOI**：10.3969/j.issn.2097-3365.2025.36.037

0 引言

当前，全球数字化转型进入深水区，AI已成为驱动业务创新和效率提升的核心引擎。为了在利用公有云海量算力和灵活性的同时，将核心数据和模型资产保留在可控的私有云环境中，混合云已成为部署AI推理服务的首选架构。这种架构将私有云与公有云相结合，通过安全的网络连接进行协同工作。然而，数据和计算任务在不同信任域之间的流转，使得AI推理系统的攻击面显著扩大。本设计聚焦智算中心面临的三大类型关键安全风险，并基于监管合规要求，提出一个系统性的解决方案。

1 智算中心安全趋势和挑战

1.1 训练数据准备阶段

数据侵权问题突出，生成式AI对海量数据需求强烈，部分机构未经许可非法爬取受版权保护的文本、图像等数据，侵犯原创者相关权利。训练数据投毒危害极大，攻击者混入虚假、错误标注或恶意样本，导致模型输出错误有害信息。敏感信息泄露后果严重，智算中心存储的用户隐私、商业机密等数据[1]，可能因防护漏洞被窃取。账号凭据窃取风险不容忽视，攻击者通过暴力破解等手段获取账号密码，非法访问系统窃取或篡改数据，高权限账号泄露可能导致大规模数据损失与破坏。

1.2 模型训练&调优阶段

模型训练与调优作为智算中心核心环节，暗藏多重关键安全风险，直接威胁核心资产与业务安全。不安全插件隐患突出，开发人员引入的未经验证插件可能存在代码漏洞、权限管控缺陷，易被攻击者利用作为入侵突破口，导致训练算法、中间结果泄露或代码被篡改，使模型训练偏离方向。模型参数萃取风险严峻，核心参数承载模型关键知识，一旦被盗取，不仅引发知识产权侵权，还可能被用于构建对抗模型实施攻击。此外，模型文件易遭网络攻击或内部违规盗取，造成核心资产流失与法律风险；API恶意调用则通过无效请求、代码注入等消耗算力，导致训练中断或模型输出错误信息，损害业务连续性与声誉。

1.3 模型部署&推理阶段

AI部署推理阶段面临输出内容不当、过度代理、提示注入攻击、拒绝服务攻击等多重安全风险，严重影响应用合规性与业务连续性。AI输出内容易出现合规问题，受训练数据局限、算法缺陷等影响，可能生成错误信息、偏见言论或违法侵权内容。提示注入攻击难以防范，攻击者通过构造恶意提示绕过安全机制[2]，泄露隐私或诱导模型输出有害内容，聊天机器人场景已出现此类隐私泄露案例。拒绝服务攻击则通过海量恶意请求占用算力、带宽资源，导致服务中断。

1.4 监管合规要求

当前我国已形成以多部规章为核心的人工智能监管体系，智算中心面临严格的合规适配压力。《互联网信息服务算法推荐管理规定》等要求算法公开透明，保障用户知情权与选择权，但智算中心的算法模型往往具有复杂性与保密性，公开范围与程度的界定成为合规难点。部分智算中心存在合规意识薄弱问题，未按要求开展算法备案、安全评估，或未落实生成合成内容的标识义务。

此外，相关法规更新速度快，智算中心的安全管

理体系难以快速适配，部分技术防护措施与监管要求存在差距，导致合规风险持续存在。

2 智算中心安全体系框架

2.1 核心防御理念

网络安全的核心是持续演进的攻防对抗，智算中心安全建设摒弃单一防护模式[3]，深度融合单点防御、纵深防御、主动防御与零信任安全四大理念，构建动态适配的立体防护体系。单点防御作为基础屏障，通过高性能防护设备与DDoS清洗技术，精准抵御网络攻击、流量轰炸等外部威胁，筑牢第一道安全防线。纵深防御依托分区分域分权机制，将智算中心划分为外联区、训练区等多类独立模块，明确访问权限与数据流转边界，在数据存储、模型训练至服务部署全链路层层设防，防范核心资产泄露风险。主动防御以全域安全运营中心为核心，汇聚全场景安全数据，实现态势统一可视可控，打破被动响应模式，提前识别威胁并快速处置。零信任安全贯穿访问全流程，构建远程接入机制，摒弃“内网可信”预设，对所有请求严格校验，搭配API代理实现接口粒度管控，从源头遏制隐患。四大理念协同发力，形成覆盖“防御—管控—预判—溯源”的全周期安全能力，精准适配智算中心复杂安全需求。

2.2 分层防御架构

为适配智算中心多场景、高敏感的复杂安全需求，方案以“分层解耦、全域协同”为核心，构建执行层、管控层、分析层三级架构，形成“检测—研判—处置—反馈”的闭环防护体系。执行层作为“一线执行单元”，聚焦算力底座防御，部署高性能防护设备、DDoS清洗系统等组件，实时监控算力集群、存储节点等异常行为，精准拦截恶意攻击、算力盗用等威胁，同时将安全事件实时上报，为研判提供原始数据支撑。管控层担任“协同调度中枢”，一方面归集计算、存储、网络等域的安全事件与日志，经整合清洗后同步至分析层；另一方面接收处置策略，向执行层推送管控指令，联动资源隔离、加密启停等手段，构建算存网端协同处置能力。分析层作为“大脑决策中心”，整合多源数据后通过AI算法开展关联分析、威胁溯源与风险预判，可视化呈现安全态势，同时制定精准处置策略下发执行，实现全流程管控，确保安全态势可感、可控、可追溯。

3 智算安全方案分层说明

3.1 基础算力层防挖矿

算力防挖矿方案基于威胁图挖矿、动态行为检测技术，覆盖数十种挖矿算法，构建全链路防护体系。方案在通算、智算、推理服务器部署轻量级EDR（终端检测与响应），联动防火墙与EDR管理平台实现检测与阻断闭环；通过DNS请求数据与威胁信息矿池库比对，精准识别恶意连接；监控常见挖矿工具落盘动作与启动参数，从源头阻断执行；监测CPU、GPU等资源占用异常，及时终止挖矿进程保障业务；清除恶意计划任务、系统服务等持久化后门，并提供病毒查杀与文件隔离功能，全方位防范算力盗用与衍生安全风险。

3.2 模型防投毒

模型防火墙藏毒检测依托五大核心技术，构建大模型文件全维度安全防护体系，实现精准高效的恶意代码检测。首先，通过模型识别技术，深度解析流量中的文件内容，精准极速识别onnx、pickle等主流模型文件格式。其次，采用流式解析方案，无需还原完整模型文件，仅流式提取关键检测片段，大幅降低资源占用。静态分析环节中，反病毒引擎对文件内可疑代码深度剖析，通过构建抽象语法树强化反序列化执行代码的检测能力。模拟执行分析则借助Opcode Emulator，还原恶意代码调用流程，精准识别真实攻击意图。最后，通过病毒监测引擎，用更少内存实现数亿级病毒变种的高效覆盖，全方位筑牢模型文件安全防线。

3.3 推理业务高性能边界防护

智算中心面临远程研发、办公人员互联网访问智算资源的场景，存在攻击面暴露、身份不可信、访问越权及数据泄露等核心安全挑战。基于零信任理念的推理业务高性能边界防护方案，替代传统VPN实现多重安全升级：通过应用隐藏技术大幅收缩攻击面，动态认证与授权机制相较静态认证更能抵御身份伪造风险；将零信任与防火墙深度融合，依托国产SOC芯片加速，零信任代理性能显著提升，在强化安全防护的同时，保障远程访问的流畅体验，全方位化解边界访问安全隐患。

3.4 Agent网关防越权

推理防火墙作为兼具API网关与Agent网关核心能力的关键组件，为智算中心推理场景构建精准防护屏障。其核心优势之一是实现推理业务调用全可视，针对机机交互的API调用场景，建立API调用行为基线，通过实时监测偏离基线的异常行为[4]，快速识别潜在攻击风险。同时，该网关可代理用户经智能体对内外系统的访问请求，在传输链路中严格校验权限，有效防范权限滥用与越权访问问题，筑牢访问安全边界。此外，网关支持推理资源弹性调度，通过构建资源池动态分配算力：优先保障VIP用户的访问需求，同时兼顾普通用户访问，实现资源利用最大化。三者协同

发力，既保障了推理业务的安全可控，又提升了资源配置效率。

3.5 模型防诱导

大模型推理阶段因训练优化的性能考量与攻击场景的不可穷举性，易遭受恶意提示词注入攻击，引发不良信息输出等风险。推理防火墙依托千万级对抗样本训练的专属模型，精准过滤越狱、恶意诱导类提示词，从源头阻断攻击。同时，防火墙支持自定义过滤功能，通过正则表达式灵活适配特殊防护需求，弥补通用防护的覆盖盲区。针对提示词中隐藏的 SQL 注入、XSS 等常见攻击模式，具备专项识别与拦截能力，全方位防范多类型注入风险。此外，防火墙可限定大模型的对话领域边界，通过“通用过滤 + 自定义配置 + 攻击模式拦截 + 领域限定”四重防护，保障大模型输出的合规性与安全性。

3.6 数据防泄漏

智算场景下数据使用频繁，权限管控与防泄漏至关重要。本方案通过端网协同架构，实现数据无界安全流转，同时筑牢泄漏与越权访问防护屏障。终端侧部署 EDR（支持 PC 与服务器），对数据进行精细化分级分类：按业务场景划分为研发、人力资源等类别，按敏感程度设定机密、高密、秘密、公开等等级。数据使用或流转时，通过底层协议染色技术为数据打上标签，标签随数据报文全程同步流转。网关与推理防火墙作为关键校验节点，依据数据染色标签与用户访问权限实时匹配，一旦检测到越权访问立即拦截 [5]。同时，端网协同构建多维数据流转行为图谱，借助 AI 算法深度分析异常行为，精准识别潜在数据泄漏风险，实现“分级分类—标签溯源—权限校验—智能预警”全链路防护。

3.7 数据防勒索

智算中心的数据价值非常高，勒索病毒攻击容易造成巨量损失，本方案设计“端网存”联动防勒索体系，实现勒索极低误报、数据不丢失的核心保障。本方案基于勒索病毒攻击全流程，构建“网络 + 存储”多层防护，兼具检测准、处置快、恢复稳三大优势。事前依托 AI 检测引擎与专属算法，对勒索病毒进行智能深度分析 [6]，检测率基于业界常规指标提升至 90% 以上，大幅降低病毒漏检风险。事中通过安全态势感知实时收集全网威胁信息，借助网安协同机制快速阻断威胁扩散，将业界需要数天的人工取证、分析、策略配置流程，压缩至分钟级完成分析与处置策略下发。事后态势感知实时同步勒索告警至存储管理器，联动执行快照恢复、数据隔离、恶意文件拉黑等动作，数据恢复速度提升数倍，全方位筑牢数据安全防线。

3.8 安全运营中心

智算中心日均产生百万级告警，人工处理响应慢且多资产监控分散，难以实现全局安全态势感知与快速处置。安全运营中心基于“AI 对抗 AI”思路 [7]，构建集智能资产管理、安全检测、威胁防御与事件响应于一体的运营管理体系。中心内置安全大数据平台，支持接入 200 多种网络安全产品的 syslog、kafka 等数据，完成全量数据治理与服务。威胁分析与响应环节，结合威胁情报与 AI 技术，通过检测规则实现威胁检测率，借助丰富事件处置模板达成 80% 以上自动执行率。同时支持多种类型安全态势可视化呈现，从攻击链视角分析威胁趋势，实现全网统一态势感知、精准溯源取证与自动化闭环处置，破解智算中心安全运营难题。

4 结束语

在数字化转型浪潮下，AI 推理服务已成为企业提升核心竞争力的必然选择，但随之而来的全链路安全风险对传统安全防护体系提出严峻挑战。本文立足智算中心安全需求，突破传统安全思维局限，提出基于“零信任”与“机密计算”双核心的多维深度防御体系，为算力底座、数据流转、模型应用至推理服务全流程提供系统性安全解决方案。研究核心价值在于重新定义了 AI 安全与创新的关系——安全并非 AI 发展的阻碍，而是保障其健康可持续推进的核心基石。通过落地该综合安全体系，企业既能充分释放 AI 推理服务的技术红利，又能有效规避各类安全风险，实现数字资产的全面保护，最终在技术创新与风险管控间达成动态平衡，持续赢得市场与客户的信任，为数字化时代的 AI 应用安全提供了可参考的实践范式。

参考文献：

[1] 赵果 . 大数据时代下智算中心的数据安全与隐私保护 [J]. 通讯世界 ,2025,32(10):39-42.
[2] 薛婴 , 潘洁 , 侯慧芳 , 等 . 智算环境安全防护体系研究 [J]. 保密科学技术 ,2025(05):37-42.
[3] 郭雪松 , 李雨轩 , 陈军 , 等 . 智算云安全责任共担机制构建 [J]. 通信企业管理 ,2025(03):69-72.
[4] 赵栖平 , 丁飞 , 王诗怡 , 等 . 算力中心云服务架构与关键技术研究 [J]. 信息通信技术与政策 ,2025,51(02):32-34.
[5] 郭莹 , 杨秩 , 蔡幸波 . 智算中心及边缘计算环境下数据动态脱敏技术研究 [J]. 智能建筑电气技术 ,2025,19(01):15-17.
[6] 刘泳妍 , 郭栩浩 . 基于云服务的勒索病毒防范与分析 [J]. 网络安全技术与应用 ,2024(11):58-61.
[7] 金威 , 姚昌华 , 余晓晗 , 等 .AI 对抗中基于位置预测的无人机决策方法 [J]. 计算机仿真 ,2025,42(07):15-20.