

主题：高管视角、课题摘录

交易技术前沿

2026年 第3期 总第72期

——网络安全专刊 (第4期)

ITRDC 证券信息技术研究发展中心(上海)



上海證券交易所
SHANGHAI STOCK EXCHANGE



P02 安全视角下，期货行业人工智能建设的思考
上海期货信息技术有限公司 范宏婷

P09 AI时代证券行业安全风险管理的思考与探索
兴业证券 蒋剑飞

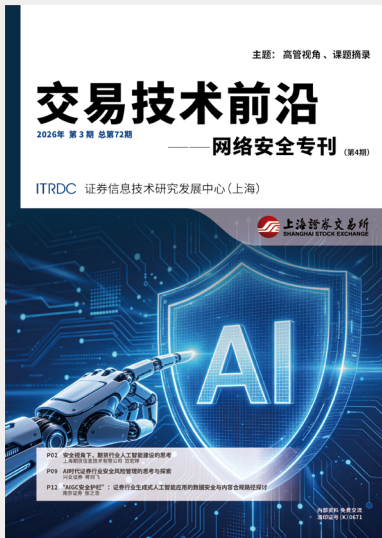
P12 “AIGC安全护栏”：证券行业生成式人工智能应用的数据安全与内容合规路径探讨
南京证券 张之浩

内部资料 免费交流
准印证号(K)0671

交易技术前沿

2026年 第3期

总第72期



总编

邱 勇

副总编

王 泊

本期“安全网络专刊”：

执行总编

道 晟 刘政言

责任编辑

张 涛

运营：

证券信息技术研究发展中心（上海）

网络安全创新实验室

主管、主办：

上海证券交易所



目录

01 高管视角

- P02 安全视角下, 期货行业人工智能建设的思考**
范宏婷/上海期货信息技术有限公司
- P09 AI时代证券行业安全风险管理的思考与探索**
蒋剑飞/兴业证券股份有限公司
- P12 “AIGC安全护栏”: 证券行业生成式人工智能应用的数据安全与内容合规路径探讨**
张之浩/南京证券股份有限公司

02 安全实践

- P17 基于多智能体编排与国产开源模型的邮件安全智能检测系统实践**
李春威/天弘基金管理有限公司
- P20 大模型驱动的证券EDR智能告警降噪与精准研判实践**
邬晓磊、甄明达、陆呈昀/东方证券股份有限公司
赵阳/奇安信科技集团股份有限公司
- P25 基于AI的网络资产风险感知与治理体系建设实践**
沈嗣贤、徐俊超、邱智晖、任思豫、鞠叶/东吴证券股份有限公司
- P31 金融机构小程序安全合规关键技术研究与实践**
陶剑峰、张强、徐孝成/中信建投证券股份有限公司
- P36 攻防实战中的安全运营: 基于AI研判的高效安全运营体系探索**
夏英杰、赵川、严星宇/国联民生证券股份有限公司
李杨、王玥/北京神州绿盟科技有限公司
- P40 证券期货行业规约驱动开发安全编程实践**
蔡子豪、姜娜、邢骁/西部证券股份有限公司

- P45 证券期货业务电子认证电子签名技术应用标准研究**
王恩潭、董延敏、张佳佳、马原、张秀敏/中泰证券股份有限公司
- P53 基于多源数据归集和智能安全编排的勒索防护能力探索**
曹宁、陈其乐、崔毅然、张沁怡、吴鹏/上海证券有限责任公司
- P58 证券行业开源软件供应链安全治理体系落地实践**
黄韦、刘传友、张忍/国泰海通证券股份有限公司
- P64 金融智能体可信运行机制研究：基于全链路风险评估与零信任动态管控**
王建国、徐渊、薛志勇、李江涛、韩晓昱/山西证券股份有限公司
- P67 AI+安全实践新形势下攻防变革的道与术**
饶滔、郭孝军、吴善鹏、刘恣彤、蔡远鹏、刘世杰、蒋琼/中银国际证券股份有限公司
- P71 证券行业数字员工安全问题研究与探索**
梁德汉、范睿华、李家攀、刘亦翔/国投证券股份有限公司
- P76 大模型在证券行业网络安全威胁建模的落地实践**
黄韦、成昊、郑洲豪/国泰海通证券股份有限公司
- P79 在AI浪潮下传统网络安全何去何从**
张旭/上交所技术有限责任公司

03 思考和观点

- P83 软件数字孪生与持续“演化—可信”平衡——五年百例实证研究**
张千里/上海金融期货信息技术有限公司
- P90 向量切片级细粒度管控：证券投研知识库在RAG（检索增强生成）架构下的非结构化数据安全隔离研究**
茆正华/上海派拉软件股份有限公司
- P95 从RSAC看国内网安AI发展方向**
大模型安全研究团队/绿盟科技集团股份有限公司

04 热点解读

- P100 证券行业安全养“虾”探索**
庄飞、江旺/华泰证券股份有限公司
- P103 AI Agent安全视角下的大模型金融应用风险与多层防御体系研究**
陈洪炎/上交所技术有限责任公司
- P108 证券期货业大模型安全围栏的建设与运营思考**
钟蓉、吴佳伟、李鹏/兴业证券股份有限公司
- P113 证券公司实网攻防演习防守实践**
刘志敏、陆修然/国盛证券股份有限公司

05 2024年度实验室优秀课题摘录

- P120 基于安全大模型提升网络安全监测响应能力的应用实践**
胡广跃、王磊、胡沐宇/光大证券股份有限公司
钱峰/深信服科技股份有限公司
- P125 探索集中交易系统的应用数据安全国密改造**
罗清平、刘海伟、邵伟/中国银河证券股份有限公司
夏乔/深圳市金证科技股份有限公司
张玉涛/三未信安科技股份有限公司
- P131 基于人工智能的加密流量威胁检测技术研究**
马冰、吴晨炜、孙佩、秦伟伦/国泰海通证券股份有限公司
- P136 高效特权账号管理体系技术研究与落地实践**
丁安安、夏英杰、丁旦/国联民生证券股份有限公司

01 高管视角

P02 安全视角下, 期货行业人工智能建设的思考

范宏婷

P09 AI时代证券行业安全风险管理的思考与探索

蒋剑飞

P12 “AIGC安全护栏”: 证券行业生成式人工智能应用的数据安全与内容合规路径探讨

张之浩

安全视角下， 期货行业人工智能建设的思考

范宏婷 | 上海期货信息技术有限公司

摘要：在资本市场数字化转型向纵深推进、“人工智能+”战略全面提速的背景下，人工智能凭借高效的数据处理、精准的语义解析等核心优势，正深度渗透至期货行业辅助研发、投研分析、合规风控等多元场景，成为推动期货行业提质增效、助力资本市场高质量发展的关键技术引擎。人工智能的创新应用，在持续拓宽行业服务边界、提升行业运行效能的同时，也带来了算法黑箱、模型失控等新型技术风险。加之期货行业本身具有“高杠杆”、“高波动”、“风险传导迅速”的固有特性，进一步加剧了技术风险的传导进程，甚至可能衍生出更为复杂的跨领域风险。鉴于此，本文系统剖析期货行业应用人工智能面临的核心风险，结合行业监管要求与法律法规体系，探讨适配行业特点的安全治理思路，进而阐明期货行业人工智能建设所需的核心安全能力，构建科学可行的落地实施路径，以期为大模型及各类人工智能技术在期货行业的安全、合规、可靠落地提供理论视角与实践层面的思考，助力行业实现技术创新与风险防控的动态平衡。

关键词：人工智能、期货行业、技术风险、安全治理、风险防控

一、引言

随着“人工智能+”行动的深入推进，人工智能已经从技术创新走向产业赋能的新阶段。然而，在推动产业效率提升与服务能力升级的同时，人工智能固有的技术风险也日益凸显，并以更复杂、更隐蔽的形式衍生出新的风险隐患。

在此背景下，国家层面密集出台相关政策，强调人工智能治理，国务院《关于深入实施“人工智能+”行动的意见》^[1]指出，要强化安全可控，推动模型算法等维度的安全能力建设，提升安全能力水平；《2026年政府工作报告》^[2]明确要求完善人工智能治理。国家互联网应急中心在《人工智能安全治理框架》^[3]中亦强调，应坚持发展和安全并重，以有效防范化解人工智能安全风险为出发点和落脚点。

结合国家人工智能安全治理的政策导向，本文聚焦期货行业人工智能建设，将系统剖析人工智能在期货行业应用中面临的核心风险，在此基础上，围绕人工智能建设的安全治理思路、核心能力与落地路径展开深入探讨，旨在明晰期货行业人工智能应用的安全底线与合规边界，为行业筑牢安全防线、稳妥推进智能化转型提供针对性思考与实践思路。

二、期货行业应用人工智能的风险

分析期货行业应用人工智能的风险，既要识别人工智能自身的内生安全风险，也要立足期货行业专属特性，考虑内生安全风险与行业特性相互叠加后，进一步演化形成的一系列衍生风险。

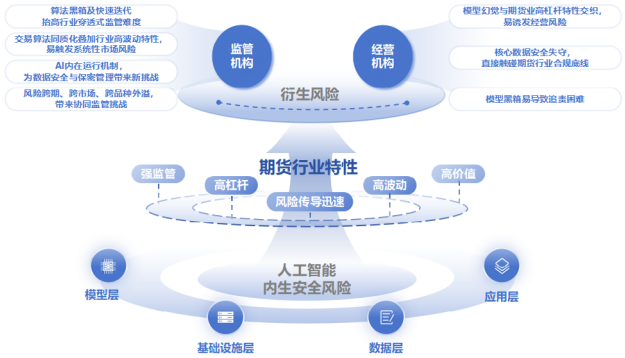


图1 期货行业应用人工智能-风险分析

（一）期货行业特性

期货市场是我国金融体系的重要组成部分，具有发现价格、管理风险、配置资源等功能，与证券市场都具备强监管、高价值等特性。与此同时，期货市场因实行保证金交易、双向开仓等机制，并具有期现联动属性，在运行机制上又呈现出鲜明的自身特征，主要体现在高杠杆、高波动以及风险传导迅速等方面：

高杠杆：期货交易采用保证金制度，投资者只需缴纳合约价值一定比例的资金即可参与交易。这一机制在提升资金使用效率、放大潜在收益的同时，也显著增加了风险。

高波动：杠杆机制会加剧价格波动。在行情剧烈变化时，可能触发集中性的强制平仓，进而引发短期内的价格急剧震荡。

风险传导迅速：期货价格对宏观政策、行业信息和市场情绪极为敏感，风险不仅会在期货合约间快速扩散，还可能迅速传导至现货市场及关联金融领域，形成跨市场风险外溢。

（二）内生安全风险：基础设施至应用层的原生风险

信通院在《人工智能安全研究报告》^[4]中从基础设施层、数据层、模型层、应用层四个角度系统梳理了人工智能内生安全风险。报告中指出，基础设施层风险主要包括软件供应链不可信风险、硬件缺陷被利用风险和软硬集成脆弱性风险等；数据层风险主要包括数据采集阶段的合规性与完整性挑战、数据清洗阶段存在的数据污染风险和敏感信息残留隐患、数据标注阶段的标注质量缺陷等；模型层风险主要包括模型输出幻觉、可解释性黑箱、算法同质化等；应用层风险主要包括小参数模型的逻辑缺陷使得模型输出偏离预期、智能体权限控制风险、通讯协议安全风险等。

（三）衍生风险：行业特性与内生安全风险的叠加

当人工智能逐步渗透至期货行业各业务场景，其内生风险与期货行业高杠杆、高波动、风险传导迅速的核心特性深度交织、叠加共振，进而衍生出一系列行业特有的、影响行业监管和机构经营的新风险。

对于监管侧而言，人工智能在期货行业的应用正催生新型监管挑战与市场运行风险，其内生的复杂性和不可预见性，打破了传统期货市场的监管逻辑，为监管机构开展穿透式监管及风险防控带来全新挑战：一是算法黑箱及快速迭代抬高行业穿透式监管难度。期货行业监管以穿透式监管为重要抓手，人工智能模型的黑箱属性导致监管机构难以穿透识别潜在风险或追踪风险传导路径；同时，算法的快速迭代，使得监管规则难以同步适配，易出现监管空白或滞后，导致借助人工智能技术规避监管、实施违规交易的行为难以被发现。二是交易算法同质化叠加行业高波动特性，易触发系统性市场风险，加剧监管的应对及处置压力。期货市场程序化交易快速发展、大模型技术逐步应用于投研业务，机构间的交易算法呈现同质化倾向，极端行情将会加剧市场非理性波动，甚至引发踩踏效应，显著抬升金融体系发生流动性骤停或交易中断的结构性风险，进一步加剧监管应急响应与风险处置的压力。三是风险跨期、跨品种、跨市场外溢，带来协同监管挑战。期货与大宗商品现货、境外衍生品市场高度联动，数据泄露、模型失效等问题衍生的市场风险，易通过跨期套利、期现联动、跨市场套利、跨品种套利等机制快速传导，形成风险外溢效应，进一步加大跨部门监管协同、风险统一处置的工作难度。四是人工智能内在运行机制叠加行业数据高敏感特性，为行业数据安全与保密管理带来新挑战。相较于经营机构，监管侧数据更为集中，既汇聚了市场交易、持仓、风控、经营数据等高敏感数据，又涵盖了未公开的监管规则、风控策略、政策草案、监控指标等高等级涉密信息。而在行业引入人工智能的过程中，大模型从底层改变了信息获取、处理与传播模式，给监管侧数据安全和保密管理埋下潜在风险：一旦涉密信息进入大模型，攻击者可通过提示词注入等手段窃取涉密敏感信息，造成失泄密事件。这类涉密信息一旦外泄，易被不法分子用于违规套利或预判政策走向等活动，不仅会扰乱期货市场公平交易秩序，严重时甚至可能诱发系统性风险，大幅增加市场监管与风险防控的难度。

对于经营侧而言，人工智能内生安全风险易诱发新的经营风险与合规危机，核心集中在三个方面：一是模型幻觉问题与期货业高杠杆特性交织，易诱发经营风险。模型幻觉可能导致期货经营机构在交易策略制定、行情趋势判断、风控模型构建等方面产生错误决策，叠加期货高杠杆与强平规则，易引发大额交易亏损、穿仓、持仓被强平等严重后果，进而引发客户投诉、资金流失及品牌声誉受损，间接加剧机构整体经营风险。二是核心数据安全失守，直接触碰期货行业合规底线。期货行业客户持仓、保证金余额、交易流水等均属于高敏感数据，《期货和衍生品法》^[5]、《证券期货业网络和信息安全管理办法》^[6]对投资者信息与交易数据保护提出硬性要求，若在采集、清洗或训练等环节，发生数据泄露等安全问题，将直接触碰合规红线，面临监管处罚。三是算法黑箱易导致追责困难。人工智能尤其是大模型，决策过程不透明，决策逻辑难以解释，若经营机构过度依赖此类模型开展投研分析与风控决策，一旦模型出现偏差或失效、引发交易亏损等问题，由于缺乏完善的人机权责界定标准，将难以明确问责主体，极易造成问责边界模糊、事后追责困难等治理问题。

三、期货行业人工智能建设的安全治理思路

针对应用人工智能带来的各类风险，各国政府及机构积极探索构建治理与监管体系，相继出台了一系列法律法规（见表1）。结合上述法律法规，立足期货行业高杠杆、高波动、风险传导迅速的特性，为严守行业安全底线、充分释放人工智能技术在期货领域的应用价值，经初步研究，我们尝试提出以下适配于期货行业人工智能建设的安全治理思路，谨供行业讨论与参考：“严守监管底线，坚持合规先行”、“坚持风险可控，构建防控体系”、“警惕数据风险，保障数据安全”、“强化人机协同，明确权责边界”、“发展与安全并重，分类分级推进”。

来源：网信办官网、发改委官网、工信部官网等

发布时间	国家/地区	全球法律法规/政策文件	说明
2023年	英国	《人工智能白皮书》	指导如何使用人工智能，提出了监管人工智能的方法。
2024年	欧盟	《人工智能法案》	全球首部全面监管人工智能的法规；2026年，欧盟拟针对该法案进行修订。
2026年	韩国	《AI基本法》	确立了人工智能健康发展框架与信任基础，形成了涵盖治理架构、产业促进、安全保障的人工智能治理法律框架。
2026年	美国	《国家人工智能立法框架》	旨在开展监管的同时，为人工智能发展创造有利环境。
2023年	中国	《生成式人工智能服务管理暂行办法》	由国家互联网信息办公室发布，提出促进生成式人工智能健康发展和规范应用。
2025年	中国	《人工智能安全治理框架》	由国家互联网应急中心发布，系统梳理了人工智能安全风险及防治措施。
2026年	中国	《网络安全法》	新修改的网络安全法规定：加强风险评估和安全监管，促进人工智能应用和健康发展。

表1 全球人工智能法律法规/政策文件（部分）

（一）严守监管底线，坚持合规先行

习近平总书记在中共中央政治局第二十次集体学习时强调，要“完善人工智能监管体制机制”。立足期货市场特性与监管实际，期货行业属于强监管领域，在应用人工智能技术时，更要始终严守监管底线、坚持合规先行，构建“监管侧强化引导、经营侧主动践行”的协同机制，监管层面健全适配行业特点的监管制度体系，经营机构主动对标监管要求，严格遵循相关法律法规，确保人工智能应用全程合规可控，以合规创新赋能期货行业高质量发展。

（二）坚持风险可控，构建防控体系

习近平总书记在2023年中央金融工作会议中强调，坚持把防控风险作为金融工作的永恒主题。国务院办公厅转发的《关于加强监管防范风险促进期货市场高质量发展的意见》^[7]明确指出，期货市场要努力实现加强监管、防范风险、促进高质量发展有机统一。

因此，期货行业落地人工智能，应始终坚持风险可控的原则，全面研判数据污染、模型投毒、算法黑箱等潜在安全风险，加快构建适配期货场景的人工智能全生命周期风控体系，实现风险早识别、早预警、早处置，稳步化解人工智能在行业应用中的风险，为服务实体经济高质量发展提供更加安全、高效的技术支撑。

（三）警惕数据风险，保障数据安全

数据是人工智能训练与推理的核心原料，数据层存在的安全风险会直接传导并逐级放大到模型层和应用层，信通院在《人工智能安全研究报告》中强调“数据层的风险既是风险威胁的根源，也是风险传导的核心枢纽，和人工智能时代攻防的关键对象”。

立足上述安全逻辑，结合行业实际来看，期货领域对人工智能数据安全有着更为严苛的现实要求。期货交易及客户数据高度敏感私密，《期货和衍生品法》、《证券期货业网络安全和信息安全管理规定》等法规对信息保密、隐私保护提出刚性要求。在此背景下，期货行业人工智能建设必须树立覆盖人工智能全生命周期的数据安全理念，严格落实数据分类分级管理、权限合理隔离等核心要求，全面防范数据泄露、污染与投毒风险，筑牢核心数据安全防线。

（四）强化人机协同，明确权责边界

欧盟《人工智能法案》^[8]明确提出，高风险人工智能系统的设计和开发，必须确保由自然人进行有效监督，其中提及的“可信AI伦理原则”也强调以人类为主体的人机协同治理模式；国务院《关于深入实施“人工智能+”行动的意见》中也多次提及人机协同。

上述人工智能安全治理理念，对期货行业人工智能建设具有重要借鉴意义。期货行业风险传导迅速、风险后果不可逆，在风控处置、交易确认、资金管理、保证金调整等关键场景下，构建以人为主、人机协同的决策机制，是防范技术失控、守住业务安全底线的必然选择，在落地过程中，应明确人工智能的“辅助”定位，厘清人机权责边界，严格落实人工

复核，避免模型自主决策，有效防范算法失控、越权操作引发的重大风险。

（五）发展与安全并重，分类分级推进

在应用场景选择上，期货行业人工智能创新需秉持“发展与安全并重、分类分级推进”的策略。具体而言，需系统研判各类场景潜在安全风险，严格遵循监管要求，并结合风险影响严重程度评估（详见表2），优先落地风险可控、需求迫切、技术成熟的场景，做到不冒进、不越界，实现创新赋能与风险防控的有效平衡。以“人工智能直接执行交易”场景为例，鉴于期货行业高杠杆、高波动、风险传导快的特点，一旦模型策略失效或发生算法偏差，极可能通过杠杆放大风险冲击、加剧市场波动，导致风险迅速蔓延、亏损敞口扩大，操作风险及跨市场外溢隐患突出，因此，该类场景应纳入严控使用的范畴。

风险等级	典型场景	主要特征	风险影响严重程度	推进建议
低	辅助研发、辅助办公、内部智能问答、投研分析、培训支持等。	不直接接触交易、资金与关键风控决策；以提效为主；人工复核成本低。	低	优先建设。 宜作为首批试点和价值验证场景，尽快形成流程、制度和模板。
中	舆情预警分析、市场异常波动预警、合规风控辅助、异常交易行为监测等。	风控、合规等关键领域；存在一定的风险传导性与合规敏感度。	中	谨慎推进。 须建立严格的全流程管控机制，坚持人工终审底线和全程留痕。
高	直接执行交易、资金划转、自动调整保证金与强平参数等。	一旦失准将直接影响客户权益、市场稳定或监管责任；容错空间极低。	高	严控使用。 可考虑在隔离环境中研究验证，审慎评估。

表2 各应用场景风险评估

四、期货行业人工智能建设的安全核心能力

兼顾期货行业监管侧与经营侧双重安全需求，期货行业人工智能安全风险防治，首要在于控制人工智能应用中的内生安全风险，防范其与期货市场高杠杆、高波动和风险传导迅速等行业特征叠加放大后引发生生风险。

基于此，为落实上述安全治理思路，围绕人工智能建设各环节，综合对标我国现行法律法规、人工智能领域相关标准规范及政策文件要求，期货行业须构建涵盖基础设施安全、数据安全、模型安全、应用安全的多维度人工智能安全核心能力（见图2，该能力体系的关键要素下文将逐一展开说明），从而夯实算力底座运行根基、严控核心敏感数据泄露风险、保障算法模型安全可控，进而有效防范人工智能在期货行业应用过程中的各类衍生风险，筑牢行业安全风控底线，维护期货市场平稳有序运行。

当然，上述能力要求并非一成不变的静态范式，而是需要与行业监管政策同步迭代的动态体系。目前，行业监管机构正密集出台一系列政策制度，加速推进完善行业人工智能安全治理体系建设。对此，需持续跟踪研判最新政策动态与监管要求，结合人工智能应用实践与风险演变趋势，动态优化、迭代完善核心安全能力，确保安全能力建设始终与监管要求同频、与行业发展同步，最终为人工智能在期货行业的安全合规、稳健可控落地筑牢长效根基。

核心能力	关键要素	国内法律法规、标准规范及政策文件等	对期货行业的意义
应用安全	- 内容合规治理 - 动态输入拦截 - 应用权限与通信安全 - 备份与应急响应	《互联网信息服务算法生成管理规定》 《网络安全技术 人工智能生成内容标识方法》 《网络安全技术 生成式人工智能服务安全基本要求》 《证券期货业网络安全等级保护基本要求》 《新一代人工智能发展规划》	保障业务连续与合规： 规范人工智能在行业内的场景落地，防范衍生风险，支撑业务稳定，赋能行业高质量发展。
模型安全	- 模型准入合规管控 - 安全审计与测试验证 - 模型输出监测 - 模型升级动态评估	《个人信息保护法》 《生成式人工智能服务管理暂行办法》 《网络安全技术 生成式人工智能服务安全基本要求》	确保模型可控与安全： 防范模型幻觉等内生风险，保障输出内容的准确与合规，进而防治衍生风险。
数据安全	- 分级及授权管控 - 全流程合规治理 - 输入监测与动态脱敏 - 数据安全销毁闭环	《生成式人工智能服务管理暂行办法》 《生成式人工智能训练和推理训练数据安全管理规范》 《生成式人工智能服务安全基本要求》	防止核心数据泄露： 严防核心敏感数据在AI训练或推理时触权流动与泄露，防范交易数据、客户隐私等。
基础设施安全	- 网络安全防护 - 算力采购与供应链安全 - 可观测与安全运维 - 高功耗下的环境安全	《网络安全法》 《数据中心设计规范》 《证券期货业网络安全等级保护基本要求》 《生成式人工智能服务管理暂行办法》 《生成式人工智能服务安全基本要求》 《新一代人工智能发展规划》 《关于深入实施“人工智能+”行动的意见》	筑牢算力底座： 保障期货行业人工智能应用的稳定运行，防止安全事件影响期货核心交易系统。

图2 人工智能建设所需的安全核心能力

（一）夯实基础设施安全，稳固人工智能安全底座

立足期货行业人工智能发展与监管实际，基础设施安全建设应重点围绕以下四个环节：

其一，夯实网络安全防护能力。

一方面，按照《证券期货业网络安全等级保护基本要求》^[9]等，结合人工智能技术的特性，针对期货行业风控、辅助办公等不同业务场景进行科学规划，分级分类实施精准化、差异化网络防护策略，采取有效措施做好风险隔离，全面贴合期货业网络安全合规建设标准。另一方面，积极响应《新一代人工智能发展规划》^[10]的号召，探索并应用人工智能专属网络安全技术，加固人工智能产品与系统的网络边界防护，确保网络基础设施安全可靠，有效抵御各类网络攻击。

其二，强化算力采购与供应链安全管控能力。

严格落实《生成式人工智能服务管理暂行办法》^[11]等要求，优先采用安全可信的芯片、软件及算力等资源。同时，需全面评估外采资源的供应链风险，防范底层技术后门或断供风险，切实提升核心基础设施的自主可控能力。

其三，健全可观测与安全运维能力。

积极响应《关于深入实施“人工智能+”行动的意见》，建立健全人工智能技术监测、风险预警与应急响应体系。结合期货行业业务特性，搭建人工智能一体化监测平台，将分散的AI基础设施组件（如计算主机、网络、存储、算力等）纳入统一平台进行监控，实现对基础设施运行状态与安全告警的实时感知，构建“监测-告警-处置-复盘”的闭环管理机制，推动人工智能基础设施的应急响应与精细化运维能力全面升级。

其四，强化高功耗下的环境安全能力。

对标《数据中心设计规范》^[12]及关键信息基础设施保护相关要求，结合人工智能算力设备高功耗运行特性，全面升级机房与机柜供电支撑能力，优化机房散热制冷体系，有效规避局部热聚集造成的智算服务器宕机甚至火灾等安全隐患，防范因底层物理节点故障引发上层人工智能应用的中断，为系统的高可用与连续运行提供坚实的物理保障。

（二）筑牢核心数据安全，严守期货行业关键防线

期货行业数据敏感度高、涉密属性强，数据安全是行业

人工智能建设的关键防线，应将全流程数据安全保障融入人工智能应用全周期，重点围绕事前数据管控、合规治理、事中运行防护、事后闭环收尾等环节开展：

其一，构建数据分级及授权管控能力。

建立期货行业专属数据分类分级体系，明确保证金、客户持仓、个人信息等高敏数据在模型训练、知识库构建和推理调用等不同环节中的使用边界与控制要求；按业务角色与应用场景实施精细化权限访问管控，严防数据越权风险，提升高敏数据的安全治理能力。

其二，闭环全流程合规治理能力。

在数据分级的基础上，参照《生成式人工智能预训练和优化训练数据安全规范》^[13]等相关要求，建立覆盖数据采集、评估、记录和使用的全流程管理机制；结合期货行业数据敏感性高、业务关联性强的特点，建立统一的数据标注规范和质量校验机制，细化标注规则与格式标准；同时，在人工智能应用中，针对期货核心交易、结算、风控数据这类高敏感、涉密类核心业务数据可实施加密存储，严防数据泄露触碰行业监管底线或引发重大经营风险。

其三，强化输入监测与动态脱敏能力。

聚焦人工智能应用运行环节，建立持续的模型输入动态监测与拦截机制，防范恶意指令注入等攻击风险的同时，对进入大模型调用链路的涉密数据进行识别拦截或脱敏处理，避免涉密数据在模型训练、推理调用与内容输出环节的泄露风险，全面强化人工智能应用运行环节的数据保护能力。

其四，完善数据安全销毁闭环能力。

建立严格的数据删除管控策略，若涉及租赁智算设备，在设备退租环节，针对数据存储介质，可采用物理销毁等不可逆技术手段，确保数据被彻底删除、无泄露隐患，形成数据全生命周期末端安全闭环，整体提升数据防泄露管控水平。

（三）强化模型安全治理，落实模型全生命周期风险管控

模型的安全治理必须贯穿“准入-部署-运行-升级/下线”的全生命周期。遵循《生成式人工智能服务管理暂行办法》要求，并参考《生成式人工智能服务安全基本要求》^[14]等标准，可考虑重点围绕以下四个环节，加强模型安全能力建设：

其一，构建模型准入合规管控能力。

在模型准入环节，建立健全基础模型准入管理机制，使用经国家网信主管部门备案产品，从源头规避模型来源不可靠、技术路径过度依赖等风险。

其二，强化安全审计与测试验证能力。

在模型部署与集成环节，需针对开发框架及代码开展安全审计和后门存在性检测，重点围绕提示词注入、越权访问、代码风险等问题进行深度安全验证，及时识别并修复潜在漏洞，持续增强模型底层的安全韧性。

其三，建立模型输出监测能力。

在运行监控环节，依托前置护栏机制，对人工智能模型输出的风险预警、交易策略、风控建议开展实时监测，校验输出结果的准确性与业务合规性；对异常输出和错误结论

进行及时拦截、复盘整改，并借助行业知识库等手段持续迭代优化，有效防范模型幻觉误导期货业务决策。此外，针对算法黑箱及其引发的追责困难问题，优化模型解释模块，留存核心决策逻辑、参数依据、计算过程，实现模型决策可追溯、可复盘、可界定责任，解决人机权责模糊、事后追责难的问题。

其四，健全模型升级的动态评估能力。

在迭代升级环节，建立安全评估机制，对于模型重要更新、能力扩展或安全策略调整，开展上线前测试、风险评估及开源软件扫描，并留存变更记录、评估结果和回退方案，确保模型版本演进过程可追溯与可复盘。模型下线应当审慎开展合规与数据安全评估，落实关键数据的备份、归档或清理处置工作，全过程保障数据安全。

（四）完善业务应用安全，守住内容合规与业务连续底线

期货行业强监管的特性以及对业务连续性的高要求，决定了应用层面的安全建设须聚焦于内容合规与业务的连续性，具体应重点落实以下四个关键环节：

其一，强化内容合规治理能力。

严格遵循《网络安全技术 人工智能生成合成内容标识方法》^[15]等规定，对文本、图片、音频等生成内容强制性添加AI水印标识。例如，在内部投研与风控场景中，应通过AI内容标识和操作日志，明确区分AI建议与人工最终决策，从而既能确保决策过程可追溯、权责归属明确，又能避免产生误导。此外，还需在服务显著位置或协议中，明示服务范围、局限性及个人信息使用情况，充分保障投资者的知情权，防范误导风险。

其二，建立动态输入拦截能力。

参照《生成式人工智能服务安全基本要求》第5条、第6条，在用户使用模型过程中，动态检测输入内容，重点防御针对客户隐私与风控规则的恶意探测攻击，若连续或累计输入违法不良信息达到一定次数，则要采取暂停服务等处置措施。

其三，强化应用权限治理与通信安全防护能力。

结合《证券期货业网络安全等级保护基本要求》^[16]及人工智能相关安全规范，针对期货行业智能风控、投研分析等场景，建立健全权限分级体系与精细化访问控制机制。特别是在智能体广泛应用的背景下，应基于最小权限原则，加强智能体权限治理，实现权限动态管控与行为监测，防范因模型幻觉、逻辑偏差引发智能体操作（如越权下单、违规业务操作）风险；同时，聚焦通讯协议安全风险，推行通讯链路加密传输，保障数据传输过程的机密性与完整性，提升整体抗攻击能力。

其四，完善备份与应急恢复能力。

参考《生成式人工智能服务安全基本要求》第6.6条，对数据、模型、开发框架及底层工具，制定备份机制及应急恢复策略，特别是当人工智能应用于风险监控、程序化交易等对业务连续性要求极高的场景时，应制定专项灾备方案，确保在各类突发技术故障下，核心业务不中断，有效支撑期货市场安全稳健运行。

五、期货行业人工智能建设安全落地实施路径

将安全核心能力转化为可持续的治理实践，需要进一步明确安全落地的实施路径与组织保障机制，并将其贯穿于人工智能应用的规划、建设、运行及优化全过程，形成责任清晰的闭环。本节立足于单一机构的实施视角，提出了一套可供期货行业各类主体探讨的落地实施路径（如图3所示），该落地路径可概括为：以组织保障为基石，以制度先行立规范，以审慎试点稳起步，最终依托持续评估机制，实现覆盖人工智能全生命周期的迭代与优化。

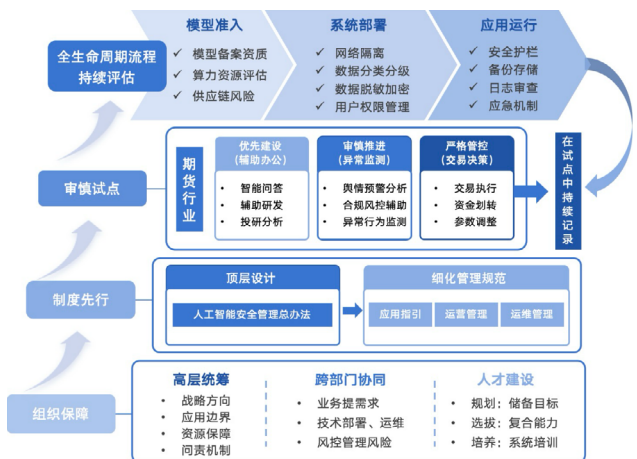


图3 人工智能建设安全落地实施路径

（一）强化组织保障，夯实治理基石

人工智能在期货领域的深入应用不仅是技术架构的升级，更是对现有业务边界与风控体系的系统性重塑，必须依托坚实的组织保障。推进人工智能安全落地，首先需要强化高层统筹，由管理层从战略层面统筹把握建设方向，明确应用边界、资源配置和责任机制，划定大模型在辅助办公、异常监测与交易决策等场景的安全边界与容错底线，确保人工智能建设始终在可控范围内有序推进。其次需要完善部门协同机制，推动业务、技术、风控合规等相关条线有效衔接，形成“需求提出-技术实施-运行维护-风险管理”相互贯通的工作闭环；厘清人工智能建设、使用、运维等相关部门的保密责任，强化权责归属，防止因责任划分不明确而引发协作脱节、管理失控和泄密事件。最后，还需加强人才建设，将人才储备纳入整体规划，选拔和培养既懂期货核心业务规则、又具备人工智能应用能力与安全意识的复合型人才，提升全员人工智能安全治理能力。

（二）坚持制度先行，明确应用边界

人工智能的应用逐步覆盖期货行业业务支持、风险管理、客户服务、合规审查、研究分析和运营管理等多个环节，其价值不仅体现在效率提升，更进一步延伸至业务安全、风险防控和监管合规等关键领域。因此，制度建设必须走在技术落地之前。推进人工智能安全应用，首要在于强化顶层设计。机构需结合自身实际，优先建立企业级人工智能

安全管理办法，将安全合规贯通于智能化建设全过程。在此基础上，进一步细化配套规范。围绕应用指引、运营管理与系统运维三大核心环节，形成分层分类、衔接有序的制度体系。通过将总体的安全要求落实为具体、可执行的操作标准，确保各业务条线与技术部门在实际推进中标准明确、有章可循。

（三）推行审慎试点，稳步梯度推进

鉴于期货行业容错空间低、风险传导迅速等特征，人工智能建设绝不能追求一次性激进铺开，须基于风险影响严重程度实行分层分类、梯度推进的稳妥试点。首先要依托风险分级，落实梯度推进。机构应优先建设智能问答、辅助研发、投研分析等低风险场景，以较小的试错成本，积累初步的合规治理与运营经验；对于涉及舆情预警分析、合规风控辅助及异常行为监测等中风险场景，需在严格叠加人工兜底与复核的前提下审慎推进；而对于直接触及交易执行、资金划转等高风险的交易决策场景，现阶段应当按需考虑实施严密管控。其次要强化全程追踪，建立量化指标。在试点过程中，可建立多维度的观测体系，持续记录运行状态与业务反馈。通过客观且连续的实操运行数据，为后续场景的稳步扩容与全面推广提供科学的决策支撑。

（四）深化持续评估，贯穿生命周期

生成式人工智能技术持续迭代，机构必须确立全局治理思维，将常态化的风险评估深度嵌入人工智能建设的全生命周期，构建系统性的安全闭环体系。首先，在模型准入阶段，要对模型备案资质、算力资源及供应链风险进行持续评估。严格审核拟引入模型的备案资质，坚决守住监管红线；系统评估算力资源的承载韧性，防范基础设施瓶颈；深度排查底层供应链隐患，确保人工智能底座的绝对安全与自主可控。其次，在系统部署阶段，对网络、数据、权限进行动态管控。立足企业级架构安全，严格执行网络隔离，定期回顾网络策略，保障网络安全。紧扣期货行业高敏数据特质，对系统内现有数据分类分级、用户权限分配，动态开展风险评估，从系统底层切断数据泄露与越权隐患。最后，在应用运行阶段，完整记录系统运行状态与安全防护措施落实情况，开展全周期运行监测与能力迭代，持续夯实全链路安全防护体系。同时，应紧跟法律法规和行业监管要求变化，动态优化安全管控规则。结合实际运行反馈，进一步健全数据备份存储、日志审查体系及应急处置预案，提升系统应对极端风险冲击的韧性，确保核心业务持续、稳定运行。

（五）案例与实践：上期所人工智能安全治理实践

基于上述期货行业人工智能安全建设的安全治理思路、能力要求及实施路径，我所结合自身业务实际与监管合规要求，选取辅助研发、辅助办公等场景先行试点，开展人工智能安全治理相关实践。

在组织保障方面，我所组建人工智能专项工作组，统筹

整体建设方向，研判重大事项，推动业务、技术等多条线协同联动，确保人工智能建设全程可控、稳步推进。

在制度构建层面，我所正积极推进人工智能相关管理办法的制定，为各业务条线与技术部门提供明确标准与执行依据，实现“建章立制、规范运行”。

在加强组织保障、完善制度体系的基础上，我所紧扣人工智能安全建设的核心能力要求，聚焦基础设施、数据资源、算法模型、业务应用四大维度同步发力，分层分类落实安全管控举措，系统推进各领域安全能力建设（图4）。

上期所人工智能安全能力建设实践			
基础设施安全		数据安全管控	
• 硬件环境 ：网络分区隔离、定向限源，筑牢底层硬件安全边界。		• 权限隔离 ：人员分级隔离，落实最小授权原则。	
• 平台软件 ：开源软件常态化测评，搭建“监测-溯源-整改”闭环。		• 销毁处置 ：完善全生命周期规范，杜绝数据残留风险。	
模型全生命周期安全		应用多维防护	
• 备案准入 ：规范开源模型引入，严格落实备案审批。		• 事前防范 ：部署安全护栏，抵御恶意指令、违规内容注入。	
• 建立知识库 ：依托内部知识库，有效降低算法幻觉风险。		• 事中监测 ：多维度用户行为监测，实时监测异常调用。	
• 可审计 ：全量日志归集，确保推理过程可追溯。		• 合规约束 ：落实内容显性标识，防范用户误导风险。	

图4 上期所人工智能安全能力建设实践

在基础设施层面，统筹硬件环境与平台软件一体化防护体系建设：依托网络分区隔离、定向限源访问等管控手段，筑牢底层硬件安全边界，从源头规避敏感数据外泄风险；健全开源软件全流程准入管控体系，对人工智能开源组件实施常态化安全测评与风险扫描，精准排查各类潜在漏洞，并搭建“漏洞监测-动态溯源-闭环整改”全链条管理机制，系统性降低软件层面的风险。

在数据安全层面，构建精细化的分级管控体系：严格推行人员权限分级隔离策略，落地数据访问最小授权原则，有效规避越权操作、违规访问引发的失泄密隐患；针对外部租赁智算资源，完善资产全生命周期处置规范，通过物理销毁存储介质等举措，杜绝数据残留与外泄风险。

在模型安全层面，强化源头准入与运行全流程管控：规范开源模型引入标准与审批流程，落实模型备案准入机制，确保各类大模型合规引入与可控运行；依托专属内部知识库，有效降低算法幻觉所带来的风险，遏制错误生成、信息误导等业务风险；同步建设模型全量日志归集能力，完整留存用户推理交互、运算调用关键记录，确保模型推理过程可追溯与可审计。

在应用安全层面，构建事前防范、事中监测、合规约束的多维防护体系：部署常态化安全防护护栏，有效抵御恶意指令注入、违规内容生成等应用层安全威胁；搭建多维度用户行为监测能力，对异常调用行为实时观测与动态管控；严格落实人工智能内容显性标识规范，明晰智能生成内容边界，切实防范因信息披露不充分引发的用户误导风险。

六、结语

期货行业高杠杆、高波动、风险传导迅速的特性，决定了其人工智能建设，绝非单一的技术升级，而是一场必须深度融合风险管控、合规约束、业务连续性与治理有效性的系统性重构。正因如此，安全并不是限制人工智能建设的约束条件，而是其得以推进、应用得以落地的内在前提。

从推进路径来看，期货行业人工智能建设必须守住监管、风险与业务连续性三大底线，既不能因风险而观望，也不宜因热潮而盲目扩张。为此，应依据不同业务场景的风险影响，采取“优先建设-审慎推进-严格管控”的梯度策略，并同步构建覆盖基础设施、数据、模型到应用的全链条安全防护体系。最终，通过组织保障、制度先行、审慎试点与全生命周期评估，方能构建一个稳定、有序、可持续的人工智能推进体系。

人工智能建设的成熟标志，不在于模型被应用于多少场景，而在于行业是否已经形成一套能够承载人工智能长期运行的治理体系。展望未来，期货行业人工智能建设仍需在发展与安全之间持续寻求动态平衡。技术越先进，越需恪守应用边界；应用越深入，越需完善制度支撑；系统越智能，越需压实主体责任。因此，未来建设重心应聚焦于“以安全为前提深化应用质量、以审慎为原则推动稳步扩展”，从而将人工智能从局部提效工具逐步发展为赋能期货行业高质量发展的重要支撑。

参考文献

1. 国务院. 关于深入实施“人工智能+”行动的意见[Z]. 2025.
2. 国务院. 政府工作报告——2026年3月5日在第十四届全国人民代表大会第四次会议上[R]. 2026.
3. 全国网络安全标准化技术委员会. 人工智能安全治理框架(2.0版) [Z]. 2025.
4. 中国信息通信研究院. 人工智能安全研究报告：技术视角下的安全风险梳理与应对[R]. 2025.
5. 全国人民代表大会常务委员会. 中华人民共和国期货和衍生品法[L]. 北京：中国法制出版社，2022.
6. 中国证券监督管理委员会. 证券期货业网络和信息安全管理暂行办法[Z]. 2023.
7. 国务院办公厅. 国务院办公厅转发中国证监会等部门《关于加强监管防范风险促进期货市场高质量发展的意见》的通知[Z]. 2024.
8. European Commission. Artificial Intelligence Act[Z]. Brussels: European Commission, 2024.
9. 中国证券监督管理委员会. 证券期货业网络安全等级保护基本要求[S]. 2021.
10. 国务院. 新一代人工智能发展规划[Z]. 2017.
11. 国家互联网信息办公室等. 生成式人工智能服务管理暂行办法[Z]. 2023.
12. 住房和城乡建设部等. 数据中心设计规范[S]. 北京：中国计划出版社，2017.
13. 国家市场监督管理总局, 国家标准化管理委员会. 网络安全技术 生成式人工智能预训练和优化训练数据安全规范[S]. 2025.
14. 国家市场监督管理总局, 国家标准化管理委员会. 网络安全技术 生成式人工智能服务安全基本要求[S]. 2025.
15. 国家市场监督管理总局, 国家标准化管理委员会. 网络安全技术 人工智能生成合成内容标识方法[S]. 2025.
16. 中国证券监督管理委员会. 证券期货业网络安全等级保护基本要求[S]. 2021.

AI时代证券行业安全风险管理的思考与探索

蒋剑飞 | 兴业证券股份有限公司

摘要： AI已成为推动证券行业高质量发展的重要引擎，为行业数智化转型注入了强劲动能。随着AI算力集群、大模型、自主智能体的规模化应用，同步催生了算力劫持、数据投毒、模型越狱、智能体越权操作等复合型新风险，如何实现“守得住安全、放得开创新”已成为行业关注热点。本文以“拥抱AI、安全护航”为核心思路，系统剖析AI融入证券业务后衍生的安全风险新态势，研究覆盖算力基建、数据、智能体的多层次纵深防御体系；通过强化全员安全认知，建设“以AI评估AI”全链路验证机制和“以AI防守AI”智能化安全运营体系，推动“防AI风险、用AI强安全”双向落地，以安全底座充分释放AI业务价值，护航资本市场高质量发展。

关键词： AI安全、多层次安全防御、数据安全、智能体安全、AI赋能

一、引言：拥抱AI发展机遇，筑牢安全护航底座

在数智化时代，AI技术正以前所未有的速度渗透到各行各业，金融领域尤为明显。AI技术的应用不仅提升了金融服务的效率，还重新定义了传统金融业务的运作模式。AI已成为券商提升核心竞争力、服务实体经济与财富管理的关键抓手。随着AI技术的深入应用，传统网络安全、数据安全防护边界被打破，模型参数泄露、提示词注入、算力恶意占用、智能体自主违规操作等新型风险日益显现，传统静态、边界式安全防护体系逐渐难以适应智能化转型需求。面对新型安全风险，亟需构建适配AI场景的安全风险管理体系，以灵活可控的安全底座为AI应用“保驾护航”，实现“守住底线、放开创新”目标，最大化释放AI业务价值。

二、AI时代证券行业安全风险新态势

AI时代安全风险呈现复合型新态势，传统网络架构引入AI基础设施后带来新型软硬件供应链隐患、算力风险及内容风险，数据治理面临隐私与合规挑战，而AI智能体自主决策则衍生出内容及行为双重失控的风险，三者交织构成拥抱AI技术潮流中安全治理的核心议题。

（一）网络与算力基础设施面临新型安全威胁

证券行业传统网络架构在AI时代新增了攻击面，包括但不限于AI算力、大模型与第三方AI工具链等。一是算力资源集中调度模式可能引发算力劫持、恶意抢占交易和风控等核心业务算力；二是模型使用过程中可能出现的参数窃取、对抗样本注入、记忆泄露、敏感信息外泄等隐患；三是引入开源AI框架和插件带来的软件供应链漏洞、知识产权及开源许可合规风险。三者耦合使传统纵深防御体系效能衰减，亟需构建覆盖“算力-模型-AI工具链”全栈的弹性安全架构，以应

对智能化转型中的复合型风险挑战。

（二）全链路数据泄露与数据投毒风险加剧

人工智能技术在证券行业的深度应用，同时引入了新型数据安全风险。在数据训练阶段，核心风险在于敏感金融数据的违规采集与隐私泄露。大模型预训练及微调需海量语料，若未严格执行《中华人民共和国个人信息保护法》及《中华人民共和国数据安全法》关于“最小必要”与知情同意的要求，易将客户身份信息、交易记录等纳入训练集，导致数据滥用或模型记忆化泄露。进入模型搭建阶段，风险转向算法黑箱与数据投毒。依据《生成式人工智能服务管理暂行办法》，金融机构须确保训练数据来源合法并进行安全评估，但模型内部逻辑不透明可能掩盖偏见或后门攻击。在智能体应用上线阶段，新增数据风险集中于实时交互中的提示词注入与越权访问，智能体在执行自动化投研、客服等任务时，若缺乏有效的输入过滤与权限隔离机制，可能被诱导输出内幕信息或执行非授权操作，违背《证券期货业网络安全等级保护基本要求》中对访问控制与审计追踪的规定。

（三）智能体自主行为失控风险且事后追溯难度大

证券行业在引入AI智能体提升业务效率的同时，衍生出内容风险及行为级安全风险。其自主决策与多步推理能力突破了传统软件确定性执行逻辑，提示词注入、上下文污染或目标函数偏差均可能引发非预期操作，如错误下单、合规误判或客户信息越权访问。智能体间协同及与外部的动态交互则扩大了攻击面，恶意智能体甚至可伪装合法身份实施链式渗透，而现有权限模型难以对其临时生成的工具调用进行细粒度管控。更严峻的是，智能体具备环境感知与策略自适应特性，可能绕过基于规则的风控机制，甚至在对抗性反馈中强化有害行为模式，形成“越用越险”的负向演化。此外，其决策过程缺乏可解释性，也可能导致事后审计与责任追溯困难。

三、AI多层次协同防御体系：以安全底座护航AI规模化应用

面对AI时代的复合型安全挑战，需要构建基础设施-数据全生命周期-智能体行为三层联动、弹性迭代的安全防御体系。在全面防范新型安全风险的基础上，为AI业务应用打开落地空间。

（一）底座升级：AI基础设施安全防护能力跃迁

面对新引入的算力基建风险，治理的核心在于管好AI算力、模型及第三方AI工具链，如表1所示。在算力维度，恶意消耗不仅意味着成本失控，更可能挤占核心业务的计算资源，因此治理思路需从粗放式供给转向基于Token计量的算力精细化运营与弹性调度，保障核心业务算力优先供给，在安全可控前提下开放算力供业务创新使用；在模型层，风险则体现为内容输出的不可控，需构建全生命周期的围栏防护体系，在模型引入前进行严格的合规与版权评估，在运行时通过DLP等技术手段对输入输出进行实时过滤与审计，确保智能应用在既定轨道内安全运行；对于AI相关技术支撑工具而言，开源组件的广泛引入使得软件供应链的复杂性呈指数级上升，安全治理应前移，建立涵盖商业与开源软件的统一资产台账（AI+CMDB），通过严格的网络隔离策略收敛攻击面，并持续追踪漏洞情报，将不可控的外部依赖纳入内部可视、可管的安全域中。

管控维度	核心风险	治理落地思路（兼顾安全与业务赋能）
AI算力	算力恶意占用、资源抢占、成本失控	算力池化Token精细化计量；业务算力分级隔离；异常算力消耗自动告警限流，保障创新场景算力供给。
大模型资产	版权纠纷、敏感数据输出、云上数据出境	上线前合规与安全双评估；运行时输入输出实时过滤围栏；本地数据与云上模型隔离，在安全可控的前提下开放调用权限。
AI开源/商用工具	供应链漏洞、开源版权风险	AI+CMDB全资产台账；内外网隔离收敛攻击面；漏洞情报常态化追踪，及时修复，并尽量减少对业务迭代的影响。

表1 AI基础设施分层防护

（二）数据安全闭环：全生命周期内嵌安全，释放数据要素价值

针对AI时代证券行业数据全生命周期的安全治理，需紧扣数据采集—标注训练—业务使用三大环节，将安全防护内嵌至AI数据流转全链路，做到“数据能用、隐私可控、全程可审”，如表2所示。在采集环节，重点在于强化用户授权合规与隐私协议更新，确保数据来源合法透明；在标注与训练阶段，通过部署专业脱敏设备与加强人员培训，规避敏感信息泄露；在使用阶段，则需严控智能体数据权限并建立全链路审计追踪机制，防范提示词注入等新型攻击。

阶段	核心风险	安全落地路径（赋能+风控并重）
数据采集	未获授权、超范围采集	完善AI专项隐私协议，落实用户授权，依法合规采集模型优化所必需的业务数据
数据标注、训练	敏感数据入模、数据投毒	自动化脱敏设备全覆盖，人员操作全审计，过滤恶意语料，保障训练数据安全可用
数据业务使用	提示词注入窃密、越权调取数据	智能体分级最小权限，对全链路推理日志进行审计，实时拦截异常数据查询，安全开放数据调用能力

表2 AI生命周期数据安全治理策略

（三）智能体防线：多维度的智能体安全管控

智能体作为AI应用的核心载体，其风险已从单一模型层面扩展至记忆、工具与平台等多个维度，需构建针对性治理框架，在发挥智能体自动化、高效率优势的同时，明确操作安全边界。如表3所示，在记忆层，智能体对历史对话的持久化存储易导致敏感信息泄露或上下文被恶意注入，应实施动态记忆脱敏与访问权限分级，确保记忆内容仅在授权范围内使用；在工具调用层，智能体执行外部操作时可能触发越权行为或被诱导执行恶意指令，需建立工具调用的白名单机制与实时行为审计，限制其操作边界并实现异常行为的即时阻断；在智能体平台层，多智能体协同、任务编排与资源调度等核心功能若缺乏安全设计，可能引发权限混乱、资源滥用或单点故障，应引入零信任架构与平台级安全策略引擎，实现对智能体的统一管控与风险隔离，支撑多业务智能体并行创新落地。三者协同防护，方能保障智能体在复杂业务场景中既灵活高效又安全可控。

管控层级	核心风险	治理方案（兼顾智能体灵活性与安全约束）
记忆层	上下文污染、敏感记忆泄露	动态脱敏后存储，分级访问权限，保留对话上下文能力，严控敏感信息留存
工具调用层	越权操作、恶意指令执行	工具调用白名单，实时行为审计告警，自动阻断违规操作，在合规前提下开放业务自动化能力
智能体平台层	认证漏洞、权限混乱、多智能体交叉风险	零信任底层架构，全生命周期统一管控与风险隔离，支撑多场景智能体有序上线和迭代

表3 多维度的智能体安全管控

四、AI安全风险实战：以智治智，护航AI创新

安全与AI是一体两面，互相赋能。通过构建“以AI评估AI”风险验证、“以AI防守AI”安全智能运营、全员认知防火墙三大实战体系，形成攻防闭环，持续优化安全能力。

（一）动态验证：全链路AI风险测评，以安全准入畅通创新通道

为有效检验多层次防御体系的实战效能，建立覆盖智能体立项评估-开发测试-合规备案-安全上线-持续运营的全链路安全闭环管理流程，以自动化AI评测工具辅助人工检测，实现风险早识别、早修复，通过安全评估的智能体可快速准入业务场景，缩短创新落地周期，如图1所示。

第一阶段，建立智能体分类分级管理机制，依据业务场景、数据敏感度对智能体定级，高风险或涉及高敏感数据的交易场景配套更强安全围栏，对轻量化资讯类智能体适当简化评测流程，差异化管控平衡安全与落地效率；

第二阶段，依托AI评测平台及针对Skills（智能体技能或插件）的自动化安全检测工具，对内部智能体应用开展常态化、全链路安全风险检测，通过模拟提示词注入、上下文污染、越权调用等典型攻击模式，快速识别其在意图理解、记忆管理与工具执行中的潜在漏洞，形成可量化的风险基线，在开发阶段完成漏洞修复，避免上线后集中整改影响业务开展；

第三阶段，面向客户的对外智能体严格落实算法、模型监管备案，并开展隐私协议、资质文件的合规审查，满足监管合规要求；

第四阶段，依据智能体特性接入安全围栏防护，并完成AI资产CMDB入库，实现上线即纳管、资产底数清晰；

第五阶段，在常态化AI漏洞情报监测基础上，积极探索AI红队机制，融合高水平安全专家与AI攻击代理，开展高强度、高仿真对抗演练，自主探索复杂攻击路径，结合人工研判精准定位防御盲点，推动安全能力动态进化。



图1 AI全链路的安全管理

（二）安全运营：以AI防守AI，破解传统安全瓶颈

随着一系列先进安全模型及渗透测试智能体的涌现，网络攻击正经历从人工驱动向自动化、智能化与规模化的根本性跃迁。此类智能体具备自主理解业务逻辑、动态生成多模态攻击载荷以及模拟真实用户行为以规避传统检测规则的能力，显著降低了高级持续性威胁（APT）的实施门槛。面对AI驱动的无差别、高频次自动化攻击，证券行业传统安全防护中普遍存在的人力瓶颈与响应滞后问题将愈发凸显，亟需引入AI能力以提升告警分析、研判及处置的效能。为此，需要将AI能力深度融入安全运营（SOC）平台，构建“告警降噪-智能研判-自动处置”三位一体闭环运营体系，用智能防御对抗智能攻击，降低安全管控对业务流程的影响，如图2所示。该全流程防守体系覆盖了流量安全、主机安全、钓鱼邮件防护、数据防泄漏（DLP）及API安全等核心场景，有效提升整体防御体系的韧性及效率。



图2 AI赋能安全闭环运营

（三）认知防火墙：筑牢全员AI安全意识底线，释放一线创新活力

在AI时代，技术防线固然重要，但“人”依然是安全管理

体系中最关键且最具弹性的环节。随着智能体应用的深入，诸如敏感数据违规出域、Skills投毒等新型风险往往源于操作层面的疏忽或认知盲区，这使得用户的安全意识成为防御体系中的核心变量。为应对这一挑战，构建系统化的人员赋能机制至关重要。我们依托内部培训平台，建立常态化的AI安全教育体系，通过投放涵盖AI工具合规使用指南及典型风险案例解析的专项课程，推动从认知层面筑牢防线。这不仅能够提升全员在复杂场景下的风险识别与防范能力，更能为业务的稳健开展与合规运营提供坚实的人才保障与智力支持。

五、总结与未来展望

AI正重构证券行业业务生态，是行业转型升级的核心机遇，但其衍生的算力、数据、智能体复合型风险，也对传统安全治理体系提出颠覆性挑战。安全建设不能陷入“重封堵、轻赋能”的误区，必须坚持安全与发展同步推进的思路：拥抱AI、安全护航。通过“以AI评估AI”动态验证、“以AI防守AI”智能运营两大核心机制实现安全能力自迭代，辅以全员安全认知防线构建完整治理闭环。未来证券行业AI安全治理要持续深化“以智治智、以攻促防”路径：一是持续完善人机协同安全治理框架，细化AI业务权责划分，明确业务、技术、合规、安全多方协同机制；二是紧跟监管政策，同步优化模型备案、数据合规、算法透明度管控流程；三是持续打磨安全与业务融合的轻量化防护工具，让安全管控与AI创新相互促进。唯有守住金融安全底线、放开AI创新空间，实现风险可控与价值释放双向平衡，才能充分发挥AI作为新质生产力重要驱动力的作用，持续赋能资本市场高质量发展，为广大投资者提供更加安全、合规、高效的智能化金融服务。

参考文献

1. Anthropic. (2026). Zero Trust for AI Agents.
2. 中国信息通信研究院. (2025). 人工智能安全治理研究报告.
3. 中国证券投资基金业协会. (2026). T/AMAC 0004-2026基金经营机构大模型技术应用规范.
4. 全国网络安全标准化技术委员会. (2025). GB/T 45654—2025网络安全技术生成式人工智能服务安全基本要求.
5. 国家互联网信息办公室. (2023). 生成式人工智能服务管理暂行办法.
6. 中国证券监督管理委员会. (2021). JR/T 0060—2021 证券期货业网络安全等级保护基本要求.
7. 中华人民共和国. (2021). 中华人民共和国个人信息保护法.
8. 中华人民共和国. (2021). 中华人民共和国数据安全法.

“AIGC安全护栏”：证券行业生成式人工智能应用的数据安全与内容合规路径探讨

张之浩 | 南京证券股份有限公司

摘要：在“人工智能+”战略深度落地的背景下，生成式人工智能（下称AIGC）已成为证券行业数字化转型的核心驱动力，但其在落地过程中暴露出的数据安全、内容合规等风险，成为制约行业AIGC规模化、实战化应用的关键因素。证券行业兼具数据高敏感性、业务高风险性与监管强约束性特征，AIGC应用的安全合规底线不可突破。本文立足证券行业业务实际，结合大模型安全防护技术实践，探讨构建覆盖输入过滤、输出审核、权限管控的AIGC全链路“安全护栏”体系，通过“事前防控、事中管控、事后追溯”的闭环设计，平衡人工智能生产力释放与风险防控需求，为AIGC在证券行业的安全合规落地提供实践路径。

关键词：安全护栏、生成式人工智能（AIGC）、证券行业、数据安全、内容合规

一、引言

2026年1月1日，新修订的《中华人民共和国网络安全法》^[1]已正式施行，从国家层面明确了人工智能技术研发支持、基础设施建设、伦理规范完善与安全监管强化的整体要求，为AIGC的健康发展划定了框架。当前，证券行业正处于数字化转型的攻坚阶段，以大语言模型为核心的AIGC技术，已从创新探索转变为行业核心竞争力的关键支撑，在智能投研、智能客服、合规风控、智慧办公等证券业务全链条实现深度赋能，驱动行业加速场景创新与技术落地。

但技术进步背后，大模型应用的安全合规脆弱性日益凸显，行业典型风险事件接连发生：2025年，国外Grok模型用户聊天记录外泄引发隐私保护争议，Microsoft 365 Copilot提示词注入漏洞暴露应用层安全短板；2025年初，国内DeepSeek大模型因数据库配置错误（未授权访问、未加密），泄露了约百万行日志，含用户聊天记录、API密钥等敏感信息，2026年315晚会则曝光了AI“数据投毒”灰产业链，企业通过虚构产品、伪造测评、杜撰话术操控AI输出，生成虚假宣传内容，误导消费者与市场判断等。与此同时，监管层面的合规约束持续收紧，国家《生成式人工智能服务管理暂行办法》^[2]明确了数据安全与内容合规的刚性要求，《证券期货业网络安全等级保护测评要求》^[3]、《证券期货业人工智能应用监管指引（征求意见稿）》^[4]更是针对性划定了证券行业AIGC应用的安全合规红线，对训练数据来源、生成内容合规性、投资者权益保护等提出精细化要求。

在此背景下，构建适配证券行业特性的AIGC“安全护栏”，成为行业突破AIGC应用瓶颈的核心举措，也是证券机构实现数字化转型与合规发展双赢的必然选择。该安全护栏体系覆盖AIGC全体系应用，包括大模型、AI智能体等关键模块，一体化实现AIGC全链路安全风险与合规管控。

二、证券行业AIGC应用落地的安全与合规挑战

由于证券行业属性与大模型的技术特征相互叠加，证券行业机构在AIGC落地过程中面临多重安全与合规挑战，传统的安全防护体系难以适配，风险防控难度显著提升。

（一）数据安全风险突出，核心资产易遭泄露

客户持仓信息、交易策略、资金流水等是证券行业的核心敏感数据，也是大模型应用全流程的主要风险点。一方面，第三方大模型存在提示词注入、Shadow Escape等新型攻击隐患，攻击者可在无用户主动操作的前提下非法窃取敏感金融数据，攻击手段的隐蔽性让传统防护手段难以识别；另一方面，员工在使用互联网大模型时的不规范数据投喂行为，易导致内部核心资产外泄，给证券机构与客户权益带来直接威胁，而此类人为操作风险的管控难度远高于技术层面的风险。

（二）内容风控难度升级，易引发业务与舆情风险

大模型的“幻觉”特性与对抗性攻击易感性，使其内容生成环节面临双重的风险防控难题。在智能客服、投研分析等核心场景中，模型“幻觉”可能生成错误投资建议、不实市场信息，直接误导投资者决策，引发客户纠纷甚至法律风险；若遭受对抗性攻击，模型可能被恶意诱导输出违规内容，不仅触碰监管红线，还可能引发市场舆情波动，甚至导致交易异常。此外，模型对违法违规内容的过滤机制若存在漏洞，将直接给证券机构带来合规处罚风险，影响机构正常经营。

（三）合规监管要求严苛，全流程安全管控压力大

证券行业作为强监管领域，AIGC应用需满足多重安全政策的合规要求，且监管呈现全流程、穿透式特征。国家《生

成式人工智能服务管理暂行办法》等政策明确要求金融机构落实数据合法使用、算法备案、内容安全管控等义务；行业穿透式监管原则进一步要求大模型应用从训练、部署到运行的全流程安全风险的可追溯、可管控。传统网络安全防护体系聚焦于静态边界防御，难以适配大模型应用精细化、全链条的安全合规要求，如何在技术创新与安全合规达标之间寻求平衡，成为证券机构落地AIGC的核心难题。

（四）技术适配性不足，传统安全防护体系失效

AIGC的分布式训练、动态交互等技术架构与应用模式，与证券行业传统信息系统存在显著差异，依赖静态规则、边界防御的传统安全防护手段效果甚微。无论是数据流过程中的实时安全监测防护，还是内容生成环节的动态审核，亦或是合规要求的技术落地，都需要专门适配大模型特性的安全解决方案。现有安全防护体系的技术适配性不足，无法有效应对AI大模型带来的新型安全与合规风险，成为制约行业技术落地的关键瓶颈。

三、证券行业AIGC安全护栏的核心架构与适配设计

AIGC安全护栏是适配证券行业特性的全链路“AI安全网关”，以“数据安全、应用安全、模型安全”为三重防护核心，覆盖AIGC全体系（含大模型、AI智能体等），通过低侵入网络架构部署、实时检测拦截、精准识别过滤的技术特征，适配证券行业复杂的业务场景与严格的安全合规要求。AIGC安全护栏的管控范围为证券行业AIGC全体系，并非单一大模型应用，核心涵盖大模型、AI智能体等AIGC核心应用组件，实现对各类AIGC应用的全链路、一体化的安全合规管控。其架构设计遵循以下三大核心原则：

- 一是安全合规优先适配，将数据安全、内容合规、全流程可追溯等法律法规、行业监管等刚性要求嵌入架构底层；
- 二是业务影响最小化，采用低侵入式网络部署模式，避免影响交易、投研等核心业务的实时性与可用性；
- 三是安全风险精准防控，针对证券行业敏感数据类型与典型攻击场景，构建精准化、场景化的安全风险防控能力。

AIGC安全护栏的具体架构如下图1所示：

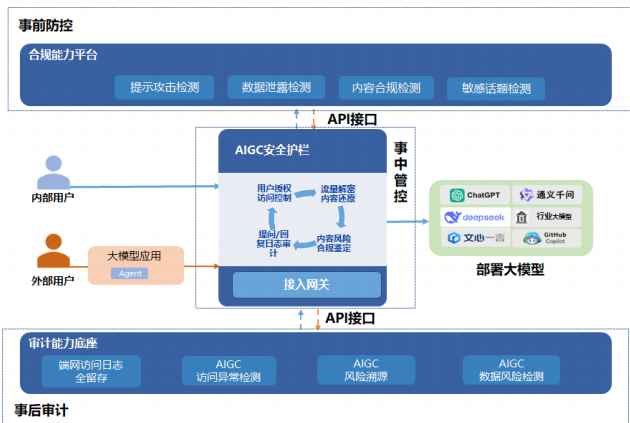


图1 AIGC安全护栏架构图

基于上述原则，AIGC安全护栏构建“事前防控、事中管控、事后追溯”的全流程的安全合规体系，实现对AIGC大模型应用的全链路的安全合规管控。

（一）事前防控：风险前置拦截与合规基线构建

合规能力平台作为AIGC安全防护体系事前防控的核心中枢，聚焦AIGC全流程合规与风险前置拦截，通过多维度精准检测构建证券行业AIGC应用的第一道安全防线，有效落实证券机构的“安全左移”理念。

该平台依托提示攻击检测、数据泄露检测、内容合规检测、敏感话题检测四大核心能力，形成全链路安全合规风险识别闭环：

- （1）提示攻击检测：精准捕捉各类恶意指令注入、语法混淆、提示词劫持等攻击行为，从输入源头阻断模型被诱导违规、越权访问或执行非授权操作的风险，筑牢模型交互入口安全；
- （2）数据泄露检测：实时监测数据交互过程中的敏感信息流转，对客户身份、交易流水、核心商密数据等证券行业关键信息实现自动识别与拦截，防止训练数据、运行数据泄露，匹配证券行业严格的数据保密要求；
- （3）内容合规检测：对照国家法律法规和证券行业监管规范，校验生成内容的合规性，重点识别“承诺收益”“内幕交易”“虚假宣传”“股票推荐”等违规表述，确保AIGC输出内容符合国家及行业监管要求与业务合规准则；
- （4）敏感话题检测：实时感知并拦截涉政、涉敏、违规类话题等风险内容，避免模型生成不当言论，保障AIGC应用在合规框架内稳定运行。

合规能力平台以上四大能力通过API接口与AIGC安全护栏深度联动，实现事前防控与事中管控的无缝衔接，既为训练数据、模型上线前的合规校验提供核心支撑，也为后续运行阶段的安全风险管控筑牢合规基线，全面保障证券行业AIGC全体系（含大模型、AI智能体）应用的安全合规。

（二）事中管控：实时防护与动态权限管控

事中管控以“实时拦截+动态适配”为核心，通过“AIGC安全护栏”构建核心防护枢纽，实现对模型输入输出的全流程安全管控，防范运行阶段的突发风险。

输入侧通过SSL/TLS解密还原问答内容，融合关键词库、规则库与AI语义分析技术，精准识别并过滤提示词注入、恶意指令等攻击行为，同时扫描用户输入中的敏感信息，实时进行脱敏处理或直接拦截，避免敏感数据进入模型链路；输出侧则对模型生成内容开展多维度安全合规性审查，重点拦截虚假投资建议、违规营销话术、敏感数据泄露等内容，对涉及金额、合同条款、合规解释等高风险回复，触发人工复核流程后再推送至用户，确保输出内容的合规性与准确性。

此外，AIGC安全护栏支持流式会话还原与管控，在保留大模型流式输出特性的同时，具备限额限速功能，防范通过超长输入、并发轰炸等方式引发的服务可用性风险，兼顾技术防护与业务体验。

（三）事后追溯：全链路审计与策略迭代优化

审计能力底座作为AIGC安全防护体系事后审计的核心支撑，以“全链路可溯、风险可查、策略可优”为目标，构建证券行业AIGC应用的安全合规闭环运营体系，为风险追溯与防护迭代提供坚实的数据基础。

该底座通过端网访问日志全留存、大模型访问异常检测、大模型风险溯源、大模型数据风险检测四大核心能力，实现全维度风险审计与溯源：

（1）**端网访问日志全留存**：全面采集用户身份、访问时间、输入输出内容、敏感信息检测结果、风险拦截记录等全链路日志数据，满足《网络安全法》180天日志留存合规要求，并通过国密算法签名保障日志不可篡改、可追溯，为合规审计与事件调查提供核心凭证。

（2）**AIGC访问异常检测**：实时分析用户访问行为、调用频次、数据流向等特征，识别越权访问、批量调用、异常数据提取等可疑操作，及时预警潜在滥用与数据泄露风险。

（3）**AIGC风险溯源**：依托全量日志数据，支持按用户、时间、风险类型等多维度精准检索，在发生数据泄露、违规输出等安全事件时，可快速定位责任主体、还原事件全流程，实现风险根源追溯。

（4）**AIGC数据风险检测**：针对日志中沉淀的输入输出数据，持续挖掘敏感信息泄露、违规内容生成等潜在风险，为防护规则迭代与模型优化提供数据支撑。

审计能力底座的上述四大能力通过API接口与AIGC安全护栏深度联动，将审计数据转化为防护策略优化依据，推动安全护栏能力持续迭代，实现证券行业AIGC全体系（含大模型、AI智能体）应用的安全风险长效管控。

四、证券行业AIGC安全护栏的实践与价值

AIGC安全护栏在证券行业的有效落地，需结合大模型应用的不同场景进行针对性部署，目前已在部分机构得到实践验证，主要覆盖内网自建模型自建应用、自建应用对接互联网大模型API、内网访问互联网大模型应用三大核心场景，实现不同场景下的精准防护。

一是内网自建模型场景，安全护栏重点保障模型本身的健壮性与知识库安全，通过提示词注入攻击防护防止模型被恶意误导，对训练数据进行脱敏处理，保护投研成果、交易策略等敏感信息不被泄露，同时规避员工操作过程中的数据泄露风险，实现模型、数据、用户的三重安全保障；二是自建应用对接互联网大模型API场景，安全护栏前置部署实现数据流的解密与解析，对智能客服、智能投研等应用的问答内容进行全审计，过滤敏感信息，同时支持基于业务范围的问答管控，对超范围提问采取拒答或固定答复机制，确保应用交互的合规性；三是内网访问互联网大模型应用场景，安全护栏部署于互联网出口，通过正向代理与SSL/TLS解密实现对员工使用互联网大模型的全流程管控，通过应用发现功能掌握员工使用情况，通过数据防泄漏功能限制核心或敏感数据上传，通过安全合规管控措施，实现不同角色员工的使用

授权，平衡人工智能工具的生产力释放与风险防控。

从实践价值看，AIGC安全护栏为证券行业AIGC的安全合规落地提供了以下三个方面的核心价值：

（1）**合规价值**：有效满足国家《生成式人工智能服务管理暂行办法》等政策要求，实现数据安全、内容合规、全流程可追溯，能够助力证券机构顺利通过监管合规检查，规避合规处罚风险；

（2）**安全价值**：实时阻断敏感数据泄露，有效防御提示词注入、Shadow Escape等各类AI大模型攻击行为，从源头上有效防范数据泄露与内容失范的风险，保障客户与机构核心资产安全；

（3）**业务价值**：采用低侵入式部署模式，无需改造现有业务系统，降低AIGC应用落地成本，在保障安全合规的前提下实现大模型的无门槛应用；同时通过对大模型使用情况的统计分析，为机构优化AIGC应用布局提供数据支撑，提升AIGC技术的业务赋能效率。

五、证券行业AIGC安全合规的未来发展方向

随着AIGC在证券行业的应用不断深化，AIGC安全合规将从单一的“技术防护”向体系化的“安全治理”演进，结合当前行业AIGC技术与监管趋势，预期未来主要呈现三大发展方向。

（一）技术体系持续迭代，打造场景化、智能化防护能力

围绕证券行业核心业务场景优化安全护栏能力，针对量化交易、智能投顾等场景打造专属防护模块，提升对算法策略、投资建议等专业内容的风险防控能力；可以通过融合隐私计算、联邦学习等技术，实现大模型训练数据的“可用不可见”，在保护数据隐私的前提下提升模型效果，破解数据安全与模型优化的矛盾；基于大模型本身构建安全对抗引擎，实现对新型攻击手法的自适应检测与防御，让防护技术跟上攻击手段的迭代速度。

（二）安全治理体系化建设，融入机构整体网络安全体系

将AIGC安全护栏纳入证券机构的整体网络安全运营防护体系，建立大模型应用的全生命周期安全管理流程，从模型选型、训练部署到日常使用进行全流程风险管控；构建基于“技术+制度+人员”的AIGC应用三维安全风险防控体系，通过完善大模型使用管理制度，加强员工安全培训，规避人为操作引发的安全风险，以及将大模型安全指标纳入证券机构的全面风险管理考核体系，推动AIGC应用安全合规从“被动应对”向“主动防御”转变，形成长效的安全管理机制。

（三）推动行业生态协同共建，实现标准化、规范化发展

通过联合行业监管机构、科技企业共同制定证券行业大模型安全应用标准，明确安全合规的技术要求与合规底线，实现行业内安全合规的标准化。推动建立行业级的威胁情报共享机制，实现大模型攻击手法、风险场景的实时共享，提

升行业整体防御能力，形成“联防联控”的安全格局。同时，通过加强与监管机构的沟通协作，探索监管沙盒模式下的大模型安全创新，在合规框架内鼓励技术探索，平衡技术创新与合规监管的需求。

六、结语

AIGC以其强大的数据分析、内容生成与效率提升能力，正深刻重塑证券行业的投研分析、客户服务、合规风控与智慧办公等各类应用场景，成为证券行业数字化转型的核心引擎与未来竞争力的关键支撑。但证券行业作为承载巨额资金流转、关乎市场稳定与投资者权益的核心金融领域，其数据的高敏感性、业务的高风险性与监管的强约束性，决定了AIGC应用的规模化落地必须坚守安全与合规底线，数据安全与内容合规并非“附加项”，而是贯穿AIGC应用全生命周期的“必选项”。

本文提出的AIGC安全护栏体系，以“全生命周期防控+纵深防御”为核心，通过“事前防控、事中管控、事后追溯”的全流程闭环设计，能够有效破解证券行业AIGC应用面临的数据泄露、提示词注入、内容违规、模型滥用等核心痛点，深度适配证券行业的业务特性，实现了“安全、可控、可信、合规”的AIGC应用目标。

面对AIGC技术的快速迭代与监管环境的持续完善，证券机构需以战略高度推进AIGC应用的安全合规体系的落地与优化：一方面，立足自身业务场景与技术架构，加快安全护栏的部署实施，将安全合规嵌入模型训练、部署、运行、迭代的全流程，实现“安全左移”与“全程可控”；另一方面，推动安全技术与管理体的深度融合，构建“技术防护+制度规范+人员培训”的三位一体治理模式，让安全合规理念贯穿组织运营全过程。同时，证券机构应积极参与行业协同，推动建立跨机构的威胁情报共享机制与安全标准体系，共同抵御新型网络攻击与行业共性风险，构建开放共赢的安全生态。

展望未来，随着大模型、AI智能体技术与证券业务的深度融合，AIGC应用的安全合规将向更加智能化、自适应、场景化的方向演进。AIGC安全护栏的规模化应用、大模型和AI智能体在安全审计与风险预判中的创新赋能，将成为行业AIGC应用安全合规的重要趋势。唯有将安全合规内化为核心竞争力，以完善的安全护栏护航AIGC技术应用创新，证券机构才能在AI时代牢牢把握发展机遇，实现业务效率、客户体验与风险防控的协同提升，为行业高质量发展注入持久动力，助力资本市场的稳定健康运行。

参考文献

- [1] 全国人民代表大会常务委员会. 中华人民共和国网络安全法[EB/OL]. (2025-12-29). https://www.cac.gov.cn/2025-12/29/c_1768735112911946.htm.
- [2] 国家互联网信息办公室. 国家发展和改革委员会，教育部，科学技术部，工业和信息化部，公安部，国家广播电视总局. 生成式人工智能服务管理暂行办法[EB/OL]. (2023-07-13). https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
- [3] 中国证券监督管理委员会. 证券期货业网络安全等级保护测评要求[S]. JR/T 0067—2021. (2021-08-30).
- [4] 中国证券监督管理委员会. 证券期货业人工智能应用监管指引（征求意见稿）[EB/OL].

02 安全实践

P17 基于多智能体编排与国产开源模型的邮件安全智能检测系统实践

李春威

P20 大模型驱动的证券EDR智能告警降噪与精准研判实践

邬晓磊、甄明达、陆呈昀、赵阳

P25 基于AI的网络资产风险感知与治理体系建设实践

沈嗣贤、徐俊超、邱智晖、任思豫、鞠叶

P31 金融机构小程序安全合规关键技术研究与实践

陶剑峰、张强、徐孝成

P36 攻防实战中的安全运营：基于AI研判的高效安全运营体系探索

夏英杰、赵川、严星宇、李杨、王玥

P40 证券期货行业规约驱动开发安全编程实践

蔡子豪、姜娜、邢骁

P45 证券期货业务电子认证电子签名技术应用标准研究

王恩潭、董延敏、张佳佳、马原、张秀敏

P53 基于多源数据归集和智能安全编排的勒索防护能力探索

曹宁、陈其乐、崔毅然、张沁怡、吴鹏

P58 证券行业开源软件供应链安全治理体系落地实践

黄韦、刘传友、张忍

P64 金融智能体可信运行机制研究：基于全链路风险评估与零信任动态管控

王建国、徐渊、薛志勇、李江涛、韩晓昱

P67 AI+安全实践新形势下攻防变革的道与术

饶滔、郭孝军、吴善鹏、刘恣彤、蔡远鹏、刘世杰、蒋琼

P71 证券行业数字员工安全问题研究与探索

梁德汉、范睿华、李家攀、刘亦翔

P76 大模型在证券行业网络安全威胁建模的落地实践

黄韦、成昊、郑洲豪

P79 在AI浪潮下传统网络安全何去何从

张旭

基于多智能体编排与国产开源模型的邮件安全智能检测系统实践

李春威 | 天弘基金管理有限公司

摘要：面对数据合规收紧与AI钓鱼威胁，天弘基金基于OpenClaw框架与DeepSeek/Qwen等国产开源大模型，通过私有化部署构建了“规则预筛+本地LLM多智能体深度分析”的混合架构系统。该系统采用分层处理，由四个专业智能体协同执行多模态解析、语义分析、行为与技术检测及风险评级。实践证明，系统上线后对AI钓鱼邮件检测率由30%跃升至80%以上，BEC攻击检测率超80%，且数据全本地处理，严守金融合规底线。

关键词：多智能体编排、国产开源模型、邮件安全检测

一、背景意义

（一）合规背景与行业需求

随着《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》以及《证券期货业网络和信息安全管理办法》等法规的深入实施，金融行业的数据安全治理要求持续收紧。邮件作为企业内外部信息交换的核心通道，承载着大量敏感数据与业务指令，其安全性直接关系到机构的资产安全与合规底线。

与此同时，大语言模型技术的快速发展为邮件安全检测带来了新的可能性与挑战。一方面，大模型具备强大的语义理解与上下文推理能力，能够识别传统规则引擎难以捕捉的隐藏威胁；另一方面，攻击者同样可以利用AI工具批量生成语法自然、语义连贯、上下文高度拟合的钓鱼邮件，使得传统基于特征匹配的检测机制日益失效。而云端大模型API因数据出域约束难以在金融行业广泛应用。在此背景下，天弘基金探索以OpenClaw智能体编排框架为基础，叠加DeepSeek、Qwen等国产开源大模型的私有化部署能力，构建一套全本地、多维度、可演进的邮件安全智能检测系统，在确保数据不出域的前提下，实现对复杂邮件威胁的精准识别与快速响应。

（二）传统方案面临的困境

1. 碎片化检测导致防御盲区

传统邮件防御体系呈现“烟囱式”结构，各检测环节彼此孤立。技术校验（如SPF/DKIM）、内容过滤与附件沙箱各自为战，缺乏“语义+行为+技术”的多维交叉印证机制。这种碎片化架构使得针对性极强的复合型攻击，例如商务邮件欺诈（BEC）、冒充高管的虚假指令等，能够轻易突破单一维度的防御。

2. 规则引擎在AI攻击场景下失效

传统邮件防御体系的检测逻辑建立在“已知威胁特征”的基础之上，通过关键词匹配、正则表达式、固定指纹库等方式进行威胁识别。在大模型驱动的攻击场景中，攻击者可利用AI工具批量生成语法自然、语义连贯、上下文高度拟合的钓鱼邮件，不再具备明显的“低级语法错误”或“固定指纹”，使基于特征匹配的检测机制难以奏效。

3. 云端API分析存在合规风险

通过将邮件正文、附件上传至第三方云端大模型进行分析，虽能获得较强的语义理解能力，但面临严格的数据出域约束与隐私泄露风险。金融行业的邮件内容涉及客户资产信息、交易数据等高度敏感数据，“数据不出域”是刚性约束。

（三）解决思路与目标

基于上述痛点，明确邮件安全能力增强项目的核心目标：通过构建“规则预筛+本地LLM多智能体深度分析”的混合架构，实现以下三重价值：

一是合规保障，邮件数据全程不出域，满足金融行业数据合规要求；

二是检测升级，检测维度从单点技术验证升级为语义、行为、技术三维交叉分析，提升复合型攻击的检出率；

三是持续演进，通过运营反馈闭环实现模型与策略的动态优化，确保检测能力随威胁演进持续更新。

二、主要做法

（一）构建“预筛层+分析层”分析处理架构

该架构设计的核心思想是：用轻量级规则快速处理大量明确安全的邮件，将可疑邮件集中送入多智能体系统进行深度分析，在保证检测精度的同时将算力成本控制在可接受范围内。

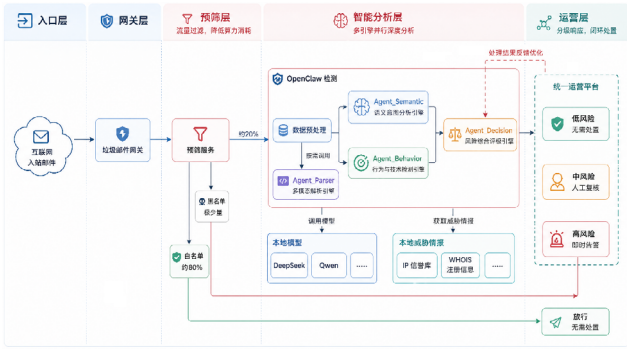


图1 系统架构图

1. 预筛层：基于纯CPU逻辑的快速分流

预筛层部署在邮件网关旁路，通过白名单快速放行与黑名单即时告警两种机制，在毫秒级时间内完成约80%的邮件处理。白名单机制基于已验证合作伙伴、SPF/DKIM/DMARC等多维度条件进行综合判断，能够在不调用任何AI模型的情况下快速放行日常邮件。黑名单机制则基于已知恶意IP、恶意域名、已确认的钓鱼特征等指标，对明确恶意邮件实施即时拦截与告警。预筛层的设计原则是“宁可漏放（进入分析层）也不误拦（阻断正常业务）”，将安全判断的核心职责交给多维度LLM分析。

2. 智能分析层：基于OpenClaw多智能体的深度检测

经过预筛层过滤后，约20%的邮件进入智能分析层，由OpenClaw编排的四个专业智能体协同完成多维度深度分析。智能分析层的核心设计理念是“多引擎并行深度分析”：邮件经数据预处理完成标准化与特征抽取（对于包含图片、文件等富文本的邮件，通过多模态解析引擎辅助解析）后，同时分发语义意图分析引擎和行为与技术检测引擎进行并行分析，最终由风险综合评级引擎对各引擎的分析结果进行综合评估，输出最终判定结果。该架构在保障检测准确性的同时，将单封邮件的平均分析时间控制在5秒以内，充分满足实时检测的业务需求。

（二）构建智能体协同分析流水线

分析层的核心是由OpenClaw编排的四个专业智能体，各智能体职责明确、分工协作，通过结构化的Prompt工程与结果传递机制实现多维度交叉验证。

1. Agent_Parser：多模态解析引擎

Agent_Parser作为多模态解析引擎，定位为预处理阶段的富媒体解析组件，主要负责对包含图片、附件等富文本内容的原始邮件进行辅助解析。邮件预处理服务会按需调用Agent_Parser，将邮件中的多模态相关特征将与邮件基础结构化信息一并整合，形成统一的JSON数据视图，为后续语义分析与技术行为检测等智能体提供标准化输入。

2. Agent_Semantic：语义意图分析引擎

Agent_Semantic是整个检测流水线的核心环节，负责

对邮件内容进行全面的语义分析。它从8个维度对邮件进行深度审查：性质判定（邮件性质与声称角色是否一致）、紧迫性分析（是否利用时间压力、后果威胁等手段催促行动）、指令检测（是否包含转账、下载、点击链接等具体操作指令）、情感操控（是否利用恐惧、贪婪等情感驱动行为）、链接安全性（链接指向是否与邮件声称的来源一致）、附件风险（附件类型与邮件内容是否匹配）、上下文一致性以及诱导性检测。每个维度均输出结构化的评分与理由说明，为后续的综合评级提供详尽的依据。

3. Agent_Behavior：行为与技术检测引擎

Agent_Behavior负责从行为层面对技术层面邮件进行检测，与Agent_Semantic的语义分析形成互补。它从6个维度开展检测：发件人身份验证（校验SPF、DKIM、DMARC记录）、发件人行为画像（分析历史邮件频率、收件人范围、邮件体量异常等）、链接与地址检测（分析域名注册时间、是否为知名域名的变体、是否使用短链接服务）、附件深度检测（分析附件类型、文件名、哈希值）、地理位置异常（分析发件IP地理位置与声称位置是否匹配）以及时间异常检测（分析发件时间是否存在异常）。

4. Agent_Decision：多维风险评级与决策引擎

Agent_Decision作为流水线的最终决策环节，汇聚Agent_Semantic和Agent_Behavior的分析结果，通过加权综合评分与特殊覆盖规则机制输出最终风险等级。风险评级分为三个等级：高风险（自动隔离+即时告警）、中风险（人工复核）、低风险（无需处置）。评级引擎的核心设计原则是：要求语义与行为双高或技术明确恶意才触发自动隔离，有效避免单一维度异常导致的误报。例如，语义分析可能将措辞强硬的外部通知误判为钓鱼（语义高但技术正常），技术检测可能将正常的新供应商邮件误判为可疑（技术异常但语义正常），而Agent_Decision的交叉验证机制能够有效过滤这类误报。

（三）威胁情报服务与运营闭环机制

为提升检测的实时性与准确性，系统集成威胁情报服务与运营反馈闭环机制。

威胁情报服务定期拉取公开的恶意域名、IP地址、钓鱼特征等威胁情报数据，更新至本地知识库，为分析层的技术检测提供数据支撑。情报更新频率为每日一次，确保检测规则能够及时跟踪最新的威胁趋势。

运营反馈闭环是检测能力持续有效的基础。安全运营人员每日对系统自动标记的高、中风险邮件进行复核，将复核结果（确认恶意或误报）反馈至系统。系统根据反馈数据调整Prompt参数与评级权重，形成“检测→复核→优化→验证”的持续迭代闭环。Prompt版本采用Git进行管理，支持版本回滚与A/B对比验证，确保每次调优均可追溯、可回退。

三、实践成效

（一）检测能力显著提升

系统上线以来，邮件安全检测能力实现了质的飞跃。对于传统模板化钓鱼邮件，检测率从原来的90%提升至95%以上；对于AI生成的高质量钓鱼邮件，检测率从原来的30%以下提升至80%以上。特别值得一提的是，对于传统方案几乎无法检测的BEC类攻击，系统的检测率达到80%以上，这得益于多智能体的语义与行为交叉分析能力。在误报率方面，通过Agent_Decision的交叉验证机制，误报率控制在5%以下。

（二）运营效率成效

系统上线后，安全运营效率实现了显著提升。在事件响应方面，高危邮件通过自动隔离与即时告警机制，安全事件从发现到处置的平均时长从原来的数小时缩短至30分钟以内。在日常运营方面，系统每日自动生成检测日报，汇总处理量、风险分布、待复核案件等关键指标，运营交接效率提升超过100%。在人力成本方面，系统上线后安全运营团队的日常手动复核工作量减少约60%，运营人员能够将更多精力投入到高价值的威胁分析与策略优化工作中。

（三）合规与数据安全保障

全部邮件数据在本地环境中完成全流程处理，数据流转路径为：邮件网关→本地预筛节点→本地GPU服务器→本地SQLite数据库，全程不涉及任何外部API调用。相较云端方案，完全消除了数据出域与第三方平台数据留存风险，满足金融行业“数据本地化”的严格监管要求。

四、经验启示

（一）混合架构是平衡成本与精度的最优解

纯规则引擎无法应对AI生成攻击，纯LLM分析成本不可接受。实践证明，“规则预筛过滤大量低风险邮件+LLM聚焦灰名单精准分析”的混合架构，在算力成本与检测精度之间取得了最佳平衡。这种架构设计思路同样适用于其他需要平衡成本与精度的安全检测场景，如网络流量分析、终端检测响应（EDR）等。

（二）多智能体交叉验证是降低误报的关键

单一维度的检测结果容易产生误报。Agent_Decision的风险评级引擎通过“加权综合评分+特殊覆盖规则”机制，要求语义与行为双高或技术明确恶意才触发自动隔离，有效降低了误报率至5%以下。这一经验表明，在多个检测维度之间建立交叉验证机制，是提升检测精度的有效手段。

（三）本地化部署是数据敏感行业的刚性选择

金融行业对数据出域有严格限制，本地化部署虽增加了初期硬件投入与运维复杂度，但从长期合规成本与数据主权角度，是不可替代的选择。国产开源模型（DeepSeek、Qwen）在中文语义理解与安全场景推理上已具备商用能力，为本地化部署提供了技术可行性。因此，本地化部署不仅是技术选型，更是合规底线的刚性要求。

（四）持续运营闭环是检测能力持续有效的基础

攻击者也在不断进化：从模板化钓鱼到AI个性化生成，从单一话术到多模态伪造。固定不变的检测策略效果难以满足实战需求。通过“运营反馈→Prompt调优→A/B验证”的闭环，检测策略能够动态适应威胁演进。Prompt版本采用Git管理，支持快速回滚，确保调优过程的可控性。实践表明，每周进行一次Prompt微调，结合实际案例的复核结果进行参数优化，能够有效保持检测能力的时效性。

五、未来展望

在当前系统的基础上，后续将从以下几个方向持续演进：一是多模态检测能力拓展，引入音频、视频等多模态内容检测能力，重点识别语音仿冒、视频伪造等新型攻击手段；二是智能体自主学习能力升级，探索基于运营反馈数据的自动化Prompt优化，减少人工干预，实现检测策略的自适应迭代；三是跨部门协同防御，将邮件安全检测能力与即时通讯、内部协作平台等场景融合，构建更广泛的社交工程攻击防御体系；四是模型能力持续迭代，紧跟国产开源模型的技术进步，及时评估并引入更强大的基座模型，持续提升检测精度与效率。

参考文献

- 1.《基金管理公司网络和信息安全三年提升计划（2023-2025）》。
- 2.《中华人民共和国网络安全法》。
- 3.《中华人民共和国数据安全法》。
- 4.《中华人民共和国个人信息保护法》。
- 5.《证券期货业网络和信息安全管理暂行办法》。

大模型驱动的EDR智能告警降噪与精准研判实践研究

邬晓磊、甄明达、陆呈昀 | 东方证券股份有限公司

赵阳 | 奇安信科技集团股份有限公司

摘要： 证券交易终端安全直接关系到资金安全、客户数据合规及业务连续性，终端威胁检测与响应（Endpoint Detection and Response，以下简称EDR）虽是关键防护工具，但常受困于海量告警噪声与研判困难。本文结合证券行业监管要求与业务特性，探讨大模型在EDR智能告警降噪、精准研判中的应用路径，通过数据预处理、模型训练、告警分级、攻击链还原等环节，构建智能化终端安全运营体系，结合行业实践案例验证应用成效，为证券机构EDR智能化升级提供严谨、可落地的实践参考，提升行业网络安全防护水平。

关键词： 大模型、EDR、证券行业、告警降噪、精准研判、银狐攻防

一、引言

随着证券行业数字化转型深入，交易终端、柜员终端、行情终端等资产数量大幅增加，同时网络态势发生深刻变化，由以单一简单目标为主转变为以关键基础设施和重点单位的定向行动。

APT攻击、无文件攻击、勒索软件等恶意威胁日趋隐蔽复杂，EDR已成为证券机构终端安全防护的核心支撑。它突破了传统防病毒软件的局限，通过实时监控进程行为、内存操作及网络连接，能够有效揪出“无影无踪”的无文件攻击痕迹，识别凭证窃取与横向移动行为，并阻断隐蔽的C&C通信，为行业构建起关键的主动防御体系。

然而，传统EDR基于规则匹配的告警模式，难以适配证券行业高合规、高可用、高敏感的业务需求，大量无效告警占用安全运营资源，导致高危威胁漏判、处置延迟等问题，严重影响终端安全防护效能。

大模型凭借其强大的自然语言处理、业务上下文关联分析及行为模式识别能力，能够深度理解证券EDR日志的语义特征、业务场景逻辑及恶意行为规律，有效破解传统EDR告警研判的痛点。本文立足证券行业实际，围绕大模型驱动的EDR智能告警降噪与精准研判展开实践探讨，结合行业实践案例，明确技术落地路径与优化方向，助力证券机构提升终端安全运营的智能化、精准化水平，满足证监会、等保三级等监管合规要求。



图1 大模型在证券终端安全的应用需求

二、证券行业EDR运营面临的现实挑战



图2 证券行业运营挑战

（一）告警很多，但真正需要处置的不多

证券机构日常存在补丁分发、软件升级、脚本化巡检、应急排障、远程协助多类操作，这些操作与恶意样本执行、命令解释程序调用、注册表修改在技术特征上几乎一致。终端安全平台只看单条行为时，容易把经过审批的正常操作与真正异常混在一起，导致值守人员首先面对的是大量无意义的告警，长期如此，终端安全运营容易形成“人盯告警”的被动模式，安全资源大量消耗在重复比对、反复确认、频繁跨团队沟通上，整体安全管理能力难以稳定提升。

（二）单点告警看得到，连续风险看不清

在真实入侵中，攻击者很少只触发一条明显告警便停止动作。通常先以邮件、即时通信文件或压缩包完成初始投递，再通过脚本落地、任务驻留、凭证访问、内网试探和外联控制逐步推进。每一个动作单独来看，未必都达到立即封堵的程度，但若将其放入完整时间线观察，往往已经构成连续风险链条。

（三）同一技术动作在不同业务语境下含义不同

同一条命令、同一个工具、同一次远程连接，放在不同资产类型、不同岗位角色之下，风险含义并不相同。如果研

判机制不了解企业的业务语境，就很难准确区分正常业务波动与可疑入侵行为。结果往往是两头落空：一头是正常操作被频繁打扰，另一头是真正异常因缺少背景支撑而产生漏报。

（四）研判和留痕依赖人工，难以支撑长期运营

证券行业终端安全工作不仅要发现风险，还要说明为何升级、依据何在、谁来复核、怎样处置、何时恢复。若研判过程、处置动作和复盘结论主要依赖人工整理，极易出现记录内容不统一、叙事逻辑不完整、关键证据散落在多个系统中的情况。

三、大模型驱动的EDR智能研判建设思路



图3 智能研判建设思路

（一）先把业务语境接入模型，而不是只接日志

我司在建设过程中首先处理的不是模型参数问题，而是业务表达问题。只有先把终端分类、用户角色、维护窗口、审批状态、可信工具、访问路径和历史结论整理为统一口径，大模型对同一条告警的理解才不会随输入来源变化而发生偏移。

围绕这一目标，我司首先明确各类业务终端的常见应用、常见连接、常见运维方式和例外场景。同时，将文件落地、脚本调用、注册表改动、计划任务创建、外联行为关联到具体资产，避免模型只看到抽象事件，不知道事件发生在什么地方、由谁触发、为何出现。

（二）把行为数据汇总为连续事件，而不是孤立提示

仅有业务基线仍然不够，终端安全研判还需要把分散在不同平台的线索放进同一条时间线上。我司在运行实践中，将终端的文件行为、访问行为、服务行为、威胁情报等各类信息统一归集，构建面向事件的研判视图。

这种方式尤其适合证券行业常见的复合型风险场景。许多高风险事件并不是某一个动作特别刺眼，而是多个中等强度线索在短时间内连续出现。模型擅长处理这种前后承接关系，因此更容易在风险尚未扩散前完成升级提示。

（三）把风险策略写入研判过程，而不是停留在制度文件中

证券行业终端安全治理强调分级分类管理。核心交易相关终端重视连续性，普通办公终端重视快速止损，外部接入终端重视访问边界。若这些要求只停留在制度文件中，模型就难以在实际研判中体现管理意图。

为此，我司将风险偏好、处置倾向、审批边界和处置规则直接写入模型研判流程。例如，办公终端使用未签名风险组件则终端关注，研发终端使用未签名风险组件则提前评估“自研研发工具”的可能性，极大避免了处置动作影响业务连续性。

（四）人机优势协同，把模型放在正确位置

在我的建设思路中，大模型承担的是信息汇聚、上下文理解、连续风险串联和标准化叙事整理职责，不直接替代人工做最终裁决。对于涉及终端隔离、账号封禁、策略收紧和业务恢复的关键动作，仍然由值守人员、运维保障人员和业务负责人共同复核。

这种边界安排并不是降低模型作用，而是让模型在最擅长的环节充分发挥价值，在敏感决策上保留必要的人为审慎。实践证明，只有坚持人机协同、边界清晰、责任明确，智能研判机制才能真正进入正式运营链路。

四、从建设到运行的落地实践



图4 运行落地实践

（一）梳理业务终端行为基线

建设初期，我司没有直接追求终端安全能力的提升，而是先做行为梳理。协调本单位安全人员、运维人员、研发人员、业务人员等多方合力，围绕常用应用、运维习惯、业务行为、常见异常案例等内容逐项校核，人员经验转化为可共享、可维护的知识库。

行为梳理特别强调与真实业务场景结合。例如，研究服务终端经常接收外部研究材料和压缩文件，普通办公终端经常处理流程通知和对账附件，运维终端经常触发批量命令和远程连接。只有把这些差异提前说清，模型才知道什么属于合理波动，什么属于需要深挖的偏离。

（二）建立行为数据研判标准

在数据接入阶段，我司重点解决的是同一事件在不同系统中表述不一致的问题。终端行为来自EDR，账号状态来自身份体系，资产归属来自台账，外联特征来自网络侧记录，历史结论来自值守台账。若这些信息无法彼此对齐，研判链路便会在输入环节出现断点。

因此，我司对各类行为信息进行统一处理。使得一线值守人员看到的不是杂乱告警，而是一段顺畅的风险链条。模型则在此基础上解决了“告警很多、信息很乱、判断很慢”的情况。

通过设置大模型研判标准，量身打造企业专属的AI研判能力，如下图：

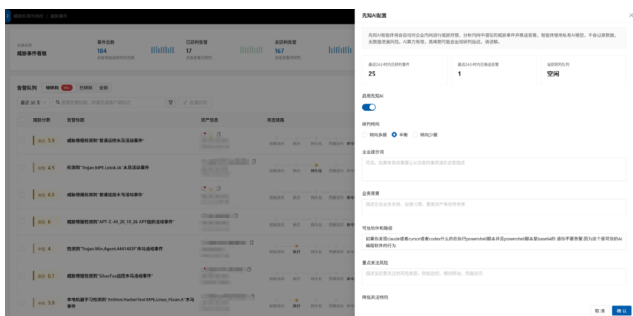


图5

（三）分层处置威胁告警

在告警汇聚阶段，我司没有简单清洗并丢弃原始信息，而是把汇聚过程分为行为聚合、威胁聚合和事件聚合三个环节。首先，对处于审批窗口、命中可信工具清单、符合既往模式的行为做记录；其次，将短时间内围绕同一文件、同一进程链、同一终端组产生的异常行为汇聚成威胁项；最后，对同时出现脚本落地、持久化尝试、异常外联和账号异常使用的事件做重点告警展示。

这种分层机制的好处，在于既能减少一线人员反复阅读近似告警的负担，又能保留对连续异常的敏感度。真正高风险的场景不会因为前面做了降噪而被埋没，反而会因为链路被自动串联而更快暴露。

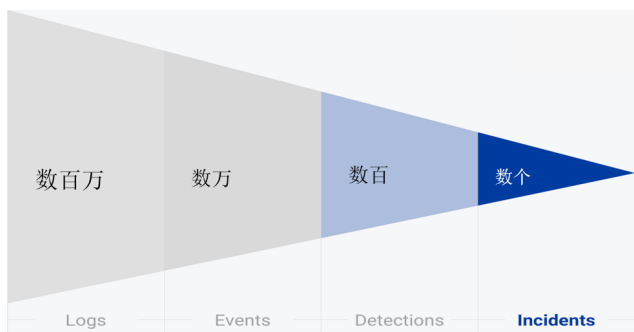


图6 行为信息汇聚

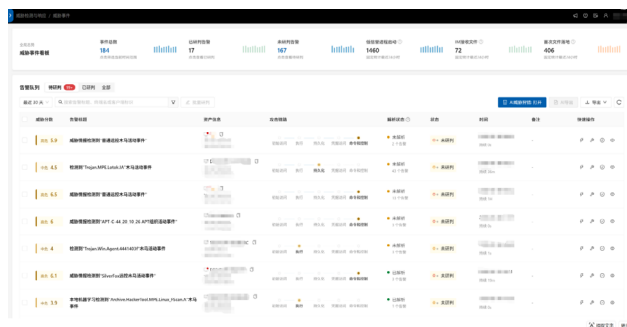


图7 基于大模型汇聚威胁事件

（四）形成运营闭环体系

终端安全能力能否长期稳定，关键不在某次演练效果，而在日常运行是否形成闭环。我司将模型输出的告警与值守汇报、人工复核、应急处置、事后复盘自然衔接，使每一次风险识别都能留下完整脉络。

在值守阶段，模型负责把散乱信息整理为统一叙事；在复核阶段，人工确认关键结论、判断业务影响、决定处置范围；在复盘阶段，再将有效经验、误判原因和新发现的场景补回模型的提示词。如此循环，模型判断会越来越贴近机构真实运行环境，团队经验也会从个体记忆沉淀为制度资产。

这类闭环还带来一个直接好处，就是不同岗位之间更容易形成共同语言。安全岗位看到的是风险脉络，运维岗位看到的是业务影响，合规岗位看到的是程序留痕，管理岗位看到的是总体态势。模型把这些内容汇总为统一记录后，跨团队协作明显顺畅。

五、银狐攻防场景下的实战验证

为了验证机制是否真正具备实战支撑能力，我司没有停留在一般性降噪测试，而是将银狐类攻击链作为重点验证场景。原因在于，此类攻击投递方式贴近日常办公习惯，落地动作具有一定隐蔽性，后续往往伴随远控、凭证访问和横向试探，十分适合检验大模型是否能够在证券办公环境中识别连续风险。



图8 银狐实战验证

（一）观察首个风险信号，识别“业务文件”伪装

银狐类攻击常以制度通知、会议材料、费用报表等工作文件为掩护，通过邮件、即时通信和压缩包投递进入普通办公终端。从单点行为看，用户下载附件、解压文件、打开查看似正常的程序并不罕见，因此这一步最容易与日常办公动作混杂。

我司在验证场景中，重点观察模型能否结合文件来源、终端类型、业务特点历史使用习惯识别首个异常信号。例如，同一岗位长期只处理文档文件，却突然运行带有自解压或安装特征的未签名文件。若这些线索能够被提前串联，后续风险判断就有了良好起点。

（二）串起连续攻击链，复原“脚本落地到违规外联”

首个信号出现后，银狐类攻击的关键不在单一恶意样本识别，而在连续动作是否被及时串联。我司将落地执行、持久驻留、凭证访问、横向试探和外联控制几个环节作为重点验证内容。



图9 银狐实战验证



图10 AI自动生成溯源报告

1. 落地执行阶段，模型识别异常进程关系

当伪装文件被触发后，脚本引擎、动态库加载程序等正常业务组件会随之被拉起。单看其中任一动作，可能只是正常的软件安装；但模型会基于行为特征，将其标记为威胁落地的关键信号。

2. 持久驻留阶段，模型识别异常终端服务

在后续阶段，系统会关注计划任务创建、启动项写入、注册表自启动修改和用户目录异常文件落地一类行为。这些动作在传统规则体系中可能分散为多条中危提示，但大模型会把它们与前序文件投递、脚本调用和同机外联放在一起理解，从而判断其更接近远控驻留，而不是普通软件安装。

3. 横向试探阶段，模型识别风险扩散意图

若终端随后出现浏览器凭证访问、共享目录探测、邻近主机连接尝试或异常远程访问行为，模型会进一步提高风险等级。此时系统关注的不再是单台终端是否受扰，而是攻击者是否已开始寻找更高价值资产。对于证券机构而言，这一步决定了事件能否被限制在普通办公区域，还是可能向研究服务环境、运维环境和核心业务边界继续推进。

4. 外联控制阶段，完成银狐类场景定性

当异常进程链与非常用外联、同源样本指纹、已知投递特征或可疑控制流量相互印证时，模型即可将事件从一般异常上升为银狐类高风险场景。这里的关键不是死板匹配某一条特征，而是综合判断文件伪装、执行链条、驻留动作、凭证访问和外联行为是否构成一条连贯叙事。正因为大模型更擅长处理这种上下文关系，APT高级未知威胁的早期识别能力明显增强。



图11 银狐攻击路径

（三）区分业务特点，形成差异化处置策略

在处置阶段，我司并未对所有终端一刀切执行同一动作。若事件发生在普通办公终端，系统会优先建议隔离主机、联合同源文件、保全样本和日志，力争在最短路径内阻断扩散。

若相似线索触及研究服务终端、运维跳板终端或核心交易相关终端，处置重点则转为“先控关键动作、再评估业务影响、后实施分层阻断”。例如，先终止异常进程、限制异常外联并与业务保障团队同步研判影响范围，避免简单断网或整体隔离对连续运行造成附带冲击。

这种差异化处置方式既体现了证券行业对稳态运行的要求，也体现了大模型输出的实际价值。模型不是盲目说“有风险”，而是全面考量真实业务场景。

（四）大模型“记忆”，把单次事件沉淀为长期能力

银狐场景验证结束后，我司并未将其视为一次独立事件，而是将其中出现的攻击特点、相关业务特点、事件处置特点等内容形成提示词记忆，回馈给大模型。后续再遇到相似线索时，系统便能够更精准给出清晰的处置。



图12

六、应用价值与成效



图13 应用价值

（一）信息化治理水位整体提升

我司的实践体会是，大模型驱动的EDR智能研判首先改变的是工作重心。过去值守人员需要花费大量时间确认重复的告警、研判各类终端行为和人工复原上下文，当前则可以更多精力放在连续风险分析、关键资产守护和复杂场景复盘上，安全团队由此从被动筛查转向主动识别，这种变化并不意味着减少人工责任，而是让人工把关出现在更有价值的环节。

（二）团队经验反哺平台能力

总体来看，大模型进入证券机构的终端安全场景，意义并不局限于引入一种新的研判工具，二是有价值的运营经验，持续转化为统一的研判和运营流程。模型负责整理和提示，人员负责约束和校正，两者结合后，终端安全运营的稳定性、可追溯性和可传承性都会明显增强。

七、结论与展望



图14 结论与展望

下一阶段，我司计划在现有基础上，将大模型这一能力稳步延伸至身份安全、数据访问、研发终端、移动办公、第三方接入、漏洞治理和应急协同场景，进一步推动终端安全与整体风险治理体系深度衔接。只要始终坚持技术建设服从治理目标、模型判断服从制度约束、关键动作服从审慎原则，大模型驱动的智能研判机制就能够持续释放实践价值，为行业高质量安全发展提供更有有力支撑。

基于AI的网络资产风险感知与治理体系建设实践

沈嗣贤、徐俊超、邱智晖、任思豫、鞠叶 | 东吴证券股份有限公司

摘要：金融科技驱动证券行业数字化转型，网络资产规模激增、边界模糊，传统静态台账与人工管理模式已难以应对资产底数不清、权属不明、风险积压等治理困境。本文从证券行业资产治理痛点出发，设计AI贯穿资产风险感知与治理全流程的实践框架：采用分布式探测架构融合多源数据实现全域资产发现，依托大模型完成资产数据清洗、指纹识别与归属绑定；构建融合CVSS、EPSS、CISA KEV的VPT漏洞优先级模型，结合暴露面、威胁情报、业务影响等多维因素，推动风险评级向实战优先级转化；践行安全左移理念，将感知能力与合规要求前置至资产上线环节，向后贯通定责、派单、整改、核验、归档完整处置链路，配套SLA量化考核形成刚性约束，实现风险全生命周期闭环治理。经生产场景验证，该体系清零189项影子、僵尸资产、高危端口，高危漏洞修复周期由15天压缩至6天，月度处置工单闭环率由65%提升至94%，为证券行业资产精细化治理提供了可落地的实践范式。

关键词：资产安全治理、漏洞优先级排序、安全左移、人工智能、证券行业

一、研究背景

（一）背景与痛点

1. 研究背景

证券行业作为资本市场核心枢纽，承载着交易结算、财富管理、投资者服务等关键职能，业务实时性强、数据敏感度高、合规约束严、故障影响面广。近年来，移动证券、智能投顾、量化交易等新业态全面普及，驱动IT架构向分布式微服务、云原生、API互联互通转型。网络资产随之从传统服务器、物理设备，扩展至容器实例、API接口、量化节点、第三方合规对接系统等多元动态载体，资产复杂度与动态性显著提升。

资产形态的快速演化，直接冲击了传统安全体系的根基。证券业务要求交易零中断、数据零泄露、风险零容忍，监管亦明确要求资产全量管控、风险动态感知、隐患闭环整改、合规全程可溯^[1]。但现实困境在于，当前资产“迭代快、类型多、归属散、边界虚”，传统安全体系赖以运行的“边界固定、资产可数、风险静态”逻辑已难以为继，静态防御模式无法适配动态化、开放化的运行场景。

从网络资产风险治理演进来看，行业先后经历了两个阶段：初期“台账登记制”以CMDB为核心，仅能记录已知资产，本质是静态快照；其后“扫描发现制”引入周期性漏洞探测，实现了被动到主动的跨越，但价值止步于报告生成，后续整改、复测、归档仍依赖人工，技术能力与治理流程之间存在断层^[2]。时至今日，多数券商仍沿袭“CMDB台账+人工盘点+离散扫描”的旧模式，工具迭代速度远落后于资产演化节奏，导致大量影子资产、僵尸资产长期游离于管控之外，在攻防场景中极易成为外网入口与入侵跳板^[3]。CNCERT 2023年数据显示，超40%的安全事故源于资产底数不清与

治理闭环缺失^[4]，这一问题已从技术短板上升为行业共性难点。正因如此，本文致力于探索第三阶段治理模式——核心命题从“发现风险”转向“闭环管控、合规化解”，治理体系从单一技术防御，升级为适配证券强合规、高稳定要求的全流程、智能化、体系化治理路径^[5]。

2. 研究痛点

在向智能化资产治理阶段转型的过程中，结合证券行业强合规、高并发、零中断、多协同的业务特性，行业普遍存在三个痛点：

一是技术与治理的断裂。证券资产治理涉及安全、研发、运维、合规等多团队协同，风险发现后需经历工单生成、责任分派、处置跟踪、结果验证等闭环环节。然而，当前多数券商的安全扫描工具与运维管理流程之间缺乏有效衔接，风险信息难以自动转化为可执行的工单任务，责任人指派、进度跟踪、逾期催办等治理动作仍依赖人工协调，导致安全团队陷入“发现越多、积压越多”的困境，发现能力与处置能力之间形成显著失衡^[6]。

二是风险度量脱离证券业务实战。当前券商普遍沿用通用CVSS作为风险处置排序依据，该标准仅衡量漏洞技术层面的严重程度，未贴合证券行业业务优先级与风险危害权重^[7]。通用评分体系导致风险处置优先级错配，有限的运维、安全资源被无效占用，核心高风险资产隐患得不到优先处置。

三是AI应用的浅层化与碎片化。大模型技术当前仅零散应用于资产识别、信息富化等单点场景，未能形成贯穿“感知—研判—处置—验证”全流程的智能能力^[8]，AI技术未真正赋能证券资产安全治理的提质增效。

综上，当前资产安全治理的核心瓶颈，已经不是“看不清”，而是“管不住”。资产测绘与漏洞扫描技术本身已具备相当成熟度，却始终未能与治理机制形成有效联动——技术

与治理之间存在断裂，风险在断裂带上淤积，最终演变为安全事件。因此，亟需构建适配证券行业特性的智能化资产风险感知与治理体系。

（二）研究目标

围绕证券行业业务强合规、高可用、交易零中断的约束，针对行业普遍存在的资产安全短板，确立三项总体建设目标：

一是全域资产底数清晰可控：基于分布式探测底座与AI能力，实现多云、内外网、多业务载体资产自动化识别、分类与异常挖掘，全面清退闲置失管资产，构建动态更新、完整可信的资产台账；

二是风险分级研判贴合实战：整合静态漏洞评分、野外利用情报、业务资产权重搭建VPT研判模型，将理论漏洞等级转化为可落地的处置优先级，解决风险分级失真、治理资源错配问题；

三是体系化闭环治理长效落地：践行安全左移理念，向前延伸至资产上线即纳管、基线异常即发现、合规要求前置生产环境，从源头遏制风险滋生；向后打通整改督办、复测核验、合规审计全链路，实现风险全生命周期可追溯、可管控、可清零；配套分级SLA量化考核与常态化运营机制，构建适配证券公司的智能化资产安全治理长效机制。

二、感知和治理思路与总体方案

（一）总体方案架构

本文以“看清资产—判准风险—管住全生命周期”为治理主线，将治理目标逐层映射为数据采集、数据汇聚、智能分析、业务应用四层架构。其中，数据采集与汇聚层共同承载“看清资产”的感知职能，智能分析层承载“判准风险”的研判职能，业务应用层承载“全生命周期管理”的落地职能。四层纵向贯通，上层需求驱动下层能力构建，下层数据向上层持续输送，共同支撑从发现到清零的全链路治理。本项目采用分层模块化设计，各层级围绕治理思路设定对应职能，避免功能堆叠。整体架构如图1所示。

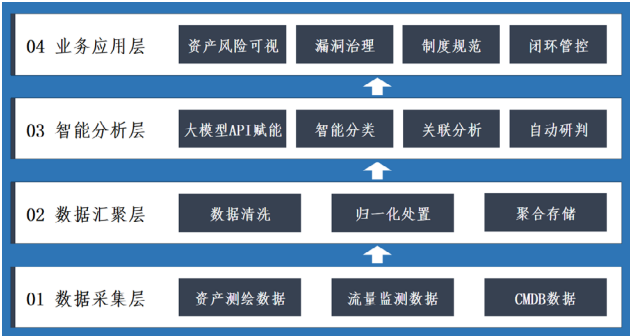


图1 治理总体架构图

数据采集层：负责接入全网资产测绘、流量监测、CMDB台账与业务系统等多源数据，建立内外网资产统一视图，实

现对云主机、容器、API、移动端等多元载体的持续发现与异常识别，为上层研判提供准确的资产底数与归属信息。

数据汇聚层：负责多源异构数据的归一化处理、关联聚合与标准化存储，建立统一的数据模型与访问接口，保障各类数据的格式统一、口径一致、更新及时，为智能分析层提供干净、完整、可用的数据输入。

智能分析层：承载风险量化模型与AI决策引擎，对汇聚后的数据进行融合分析，完成风险等级量化评估、影响范围判定与处置优先级排序，输出标准化的风险研判结论与治理指令。

业务应用层：聚焦安全运营与合规治理场景，提供资产全景管控、风险态势可视化、漏洞全周期治理、工单流程管控、合规自动化审计等核心功能，将研判结论转化为可执行、可跟踪、可考核的治理动作，推动治理流程的标准化、可视化与自动化。

各层之间通过标准化接口进行数据交互与指令传递，上层应用按需调用下层能力，下层数据主动向上层输送，形成“采集—汇聚—分析—应用”的贯通链路。

（二）核心治理理念

基于此，本文将资产探测、漏洞评分、威胁情报、AI分析等技术能力作为底座，在技术层与治理层之间构建完整的连接链路。

核心设计思路是：以标准化闭环重组技术能力，让资产探测成为持续感知、漏洞评分变为动态研判、AI贯穿全流程决策；同时以系统流程承载制度落地，将分工、SLA、升级机制内嵌于平台，让执行有闭环，各环节各司其职，共同指向“将风险真正管住”这一核心目标。

本方案的设计贯穿以下三条核心理念：

(1) 以全面感知与精准归属为治理基础。将资产探测从单点发现扩展为持续感知能力，覆盖多云、内外网及各类业务载体，主动识别未备案资产、异常端口、证书过期等风险隐患；同时将资产权属信息同步纳入感知范畴，通过内外网映射与CMDB联动完成资产确权，确保资产清晰、归属明确、责任可溯。

(2) 以多源融合校准风险研判。将权威漏洞标准、威胁情报、资产重要性、暴露面可达性纳入统一量化框架，使风险排序超越单一维度的技术评级，输出贴合实战的处置优先级，确保高危隐患优先处置、有限资源精准投放。

(3) 以技术落实制度要求。将处置时限、变更同步机制嵌入系统流程，使制度要求转化为不可跳过的执行路径，以技术保障每一项治理动作可追踪、可考核、可追溯。

三、资产风险感知治理体系建设

“风险感知→智能研判→闭环治理”为核心主线，每一环节都对应具体的治理目标，技术细节服务于治理效果。

（一）全域资产获悉：夯实治理基础底数

资产感知是治理体系的源头输入，其质量直接决定后续风险研判与闭环治理的有效性，本节围绕资产发现与资产归属两个环节展开。

1. 全域多源分布式资产测绘

证券行业资产部署呈现多云并存、多地机房、内外网隔离的复杂特征，任一单一扫描工具或数据源均无法独立完成全域覆盖。为此，本方案采用分布式探测架构，在多网络区域分别部署轻量化探测节点，由中心管控节点统一调度任务、汇总结果，实现对隔离网段、云环境、本地机房等各类场景的统一覆盖。该架构具备良好的水平扩展能力，新增网络区域时仅需部署探测节点即可纳入治理范围。

在扫描能力层面，不同工具在操作系统指纹识别、Web应用组件检测、服务版本识别等维度各有侧重。本方案通过统一调度框架集成多款扫描工具，将各工具的探测结果进行交叉验证与去重融合，取各自所长，避免单一工具漏报或误报带来的盲区，提升资产发现的覆盖广度与识别精度。

在数据源层面，资产发现不仅依赖于主动扫描结果，同时整合全网入向流量监测数据与公司CMDB台账，形成主动扫描、被动流量、已有台账的三重校验机制：流量数据可发现未开放端口但存在通信行为的隐式资产，CMDB台账可兜底因扫描条件受限而未能探测到的资产。三者相互补充、交叉印证，最大限度缩减资产漏报空间。

在探测节奏上，结合证券行业交易时段业务不可中断的刚性要求，设计差异化分时段策略：非交易窗口执行全量深度资产测绘，包含端口全量扫描、服务指纹精细识别与漏洞深度检测；交易时段仅开展轻量级资产状态巡检，监测资产在线状态与关键端口变化，在确保持续感知的同时规避对交易业务的干扰。

2. AI驱动的资产精准识别与归属绑定



图2 资产精准识别与归属绑定逻辑图

(1) 多源数据清洗与资产归一化。来自扫描工具、流量数据、CMDB台账的多源资产数据在格式、命名规范、字段完整性上存在显著差异。例如，同一资产在CMDB中记录为主机名，在扫描结果中记录为IP地址，在流量数据中记录为域名，三份数据若不进行关联归并，将被误判为多个独立资产。本方案通过AI对多源数据进行关联分析与字段映射，依据IP、MAC、主机名、域名、资产指纹等多维度特征自动判定资产同一性，完成去重合并，形成以资产为唯一索引的标准化台账记录。

(2) 业务指纹识别与资产分类。传统基于端口+服务字符串的识别方式无法准确判断资产的业务归属。本方案依托大模型多模态识别能力，结合页面内容、响应头信息、SSL证书特征、URL路径模式等多维度特征，构建证券公司专属业务指纹库，将资产自动归类为交易系统、行情系统、中后台等类型，为后续差异化风险评估与处置策略提供精准的资产画像。

(3) 归属确权与CMDB联动。资产感知的核心价值不仅在于发现，更在于明确归属。本方案通过搭建内外网资产映射机制解决外网暴露资产与内网实体资产之间的对应关系：针对探测到的互联网暴露资产，关联出口网络设备地址转换记录，还原其对应的内网真实业务载体，再联动CMDB台账自动归集归属团队、责任人员、所属业务系统、资产用途、上线时间等权属信息。通过该机制，感知阶段即完成资产确权，使每一项资产从发现起即具备明确的责任主体。

（二）多维风险智能研判：支撑分级精准治理

风险是一个多因素共同作用的复合概念，单一维度的评估存在局限性。本文从暴露面、脆弱性、威胁情报、业务影响四个维度分别量化资产风险，最终融合输出综合风险评分。



图3 资产多维风险评估逻辑图

(1) 暴露面评估。资产暴露在互联网上的程度直接决定其被攻击的概率。暴露面评估主要关注资产对外开放了哪些端口和服务、是否存在已知高危端口（如数据库端口直接暴露）、该资产在互联网上已持续暴露多长时间。原则上，端口开放越多、暴露时间越长，被扫描和利用的可能性越大，暴露面维度得分越高。

(2) 脆弱性评估。评估资产自身存在的漏洞与配置缺陷。本维度统一依托VPT多源融合漏洞优先级模型完成，通过对CVSS 3.1（技术危害基准）、EPSS（可利用性概率）、CISA KEV（已知野外利用状态）三类权威数据分层计算，并叠加AI对CVE文本与威胁情报的语义理解，动态补充0day热度权重，将静态漏洞评分转化为贴合实战的动态脆弱性评分。同时覆盖高危端口开放、弱口令、不安全配置等非CVE类缺陷，依据其可利用性与影响程度统一纳入脆弱性评估。

(3) 威胁情报匹配。评估资产当前是否面临活跃攻击。通过联动实时威胁情报库，匹配是否存在针对该资产类型、开放端口或已安装组件的在野攻击事件、漏洞利用工具或APT攻击偏好。同时，考虑不同告警数据源的可信度差异——蜜罐告警可信度高于边界WAF告警——对多源告警进行差异化加权，降低误报对研判结论的干扰。若存在活跃威胁，该维度得分显著上调。

(4) 业务影响评估。评估资产一旦失陷所造成的实际损失。同一漏洞出现在核心交易系统与测试环境，其真实危害天差地别。本维度依据资产所属业务线的重要性（结合等级保护定级）、该资产涉及的用户规模与交易体量、以及安全事件可能导致的品牌与合规损失，综合测算影响范围与损失程度。同时考虑资产的网络访问权限——若资产可横向移动至核心系统，其潜在影响面需向上修正。

上述四个维度分别量化后，综合考量安全事件的最终处置优先级时还需纳入“攻击结论”这一动态因子——即是否存在真实攻击行为或已验证的利用尝试——作为综合评分的修正项。最终采用AHP-熵权法完成多指标动态权重配比，输出统一的资产综合风险评分与等级。其核心逻辑是：低技术风险+高资产价值+高可信告警的事件，处置优先级可能高于高技术风险+低资产价值的事件，使定级结果贴合实战。评估维度与权重可根据行业监管要求与机构实际情况动态调整。

（三）全流程闭环治理：实现风险长效清零

1. 源头治理与风险前置阻断

风险治理的高效方式，是在风险产生之前将其阻断。本节在感知与处置两个环节分别向前延伸，形成源头治理能力。

(1) 上线即纳管，下线即清除。影子资产与僵尸资产的根源在于资产生命周期管理缺失。本体系将资产感知前置到上线环节——新资产完成部署后，CMDB台账同步触发资产登记流程，资产探测引擎在下一探测周期自动验证其纳入情况，未登记资产自动触发预警。同样，资产下线时自动触发互联网映射关系注销确认，避免已下线资产因未清理而继续暴露。通过将感知能力左移至资产上线环节，从源头压缩影

子资产的滋生空间。

(2) 基线驱动的异常主动发现。传统的扫描发现依赖已知特征库，对未知资产或异常变化缺乏识别能力。本体系基于3.1中建立的资产运行基线，持续监测资产状态变化——新资产出现未登记、端口非批准开放、服务类型变更、证书临近过期——这些偏离基线的行为在线上或变更的第一时间即被识别，不等漏洞出现即可前置干预，将治理窗口从“已暴露”前移至“发生变化”。

(3) 合规要求驱动的检查前移。上线环节仅有技术感知仍不够，流程层面的硬约束同样需要前置。通过联动OA、金融科技平台等，将上线备案、变更审核等管理要求转化为自动化检查项——资产未完成前置流程即上线、变更未经审批即生效，系统自动识别并告警拦截，使合规管控在风险发生之前即生效，而非事后追溯补漏。

2. 风险处置与持续改进

在完成资产风险分级研判、明确各类隐患处置优先级后，本体系打通“感知—研判—定责—派单—整改—核验—归档”完整链路，通过技术工具与管理机制深度融合，破解传统模式下权责划分模糊、风险整改滞后、复测核验缺失、资产台账与风险处置脱节等核心治理痛点，搭建一套标准化、自动化、可量化考核的长效闭环治理体系，整体治理执行流程如图4所示。



图4 治理执行流程图

(1) 精准定责与分级派单。风险若无法对应到具体责任人，闭环便无从谈起。研判结论生成后，系统自动匹配该资产对应的责任人（基于3.1.2资产归属绑定），结合风险等级自动触发分级处置工单。严重风险24小时启动处置、72小时闭环；高风险3个工作日内完成整改核验；中风险5个工作日；低风险纳入常态化监测。超时工单自动升级督办，处置记录关联绩效考核。

(2) 差异化修复方案与智能派单。系统调用公司本地部署大模型，结合漏洞CVE特征、攻击机理、资产所属业务类型及运行环境，自动生成差异化修复建议：核心生产资产优先推荐热修复、权限收敛、访问策略隔离等低业务影响方案；测试及闲置资产推荐补丁更新、组件替换、服务下线等彻底处置方案。每一份工单派发时附带专属修复指引，协助解决怎么修的落地难题。

(3) 复测核验与自动归档。责任人完成整改后，工单流转给安全团队同步发起复测，验证风险是否消除。复测通过则工单自动归档，处置记录同步更新资产台账与合规审计日志；复测不通过则工单回退至整改环节，直至风险彻底清

零。若工单超出对应等级处置时限，平台将自动升级至上级负责人和内控合规部门督办，相关处置记录直接关联绩效考核。

(4) 持续改进：治理闭环反哺感知。每一次风险处置的完成，不仅是一个风险的结束，也是感知能力迭代的起点。处置过程中积累的修复数据、误报反馈、漏洞误判记录，持续回传至资产指纹库与VPT模型，优化资产识别精度与风险研判逻辑。同时，高频复发的问題类型驱动管理制度与SLA指标的动态调整，形成“感知越准、治理越实”的正向循环。

四、资产风险感知治理体系实践

(一) 核心模块落地展示

1. 全域资产感知

对应3.1“全域资产感知”的设计，平台在落地层面实现了多源资产统一纳管与互联网暴露面感知。

如图5所示，平台整合了SCMDB信息表、PBI产品表、应用资产列表、机房网络信息等多源资产数据，形成覆盖全网的标准资产台账。资产管理页面支持按资产类型、网络区域、归属团队、数据来源等多维度筛选检索，资产字段涵盖IP、主机名、资产类型、业务归属、责任团队、开放端口、运行状态、探测时间等完整信息。



图5 多资产统一纳管

如图6、图7所示，平台集成外部攻击面管理数据，持续监测互联网资产信息与互联网发布资产，识别暴露在公网的域名、IP及开放端口。同时，基于流量数据发现的URL资产完成纳管，弥补主动扫描盲区，并据此进行主动验证。

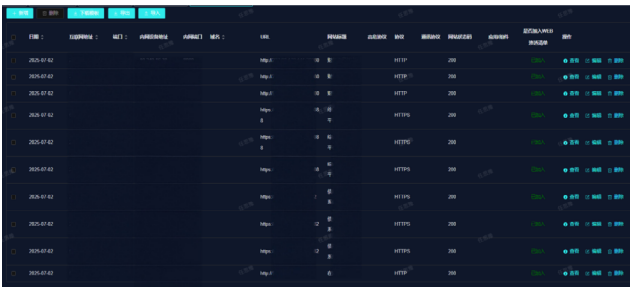


图6 互联网资产信息收集

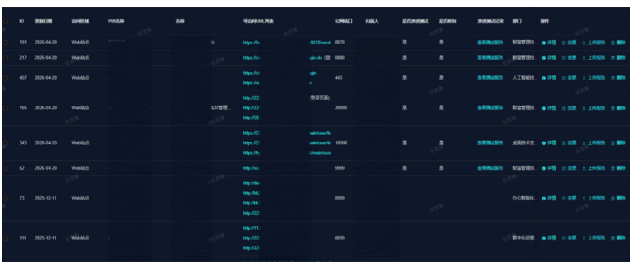


图7 Web渗透资产列表

2. 风险研判与漏洞优先级管理

对应3.2设计，平台在落地层面实现了以下核心功能：

威胁情报与漏洞态势感知。如图8所示，威胁情报看板实时展示全网漏洞情报与野外攻击态势，同步CISA KEV、CNVD等权威漏洞库更新，标记已存在野外利用的高危漏洞，辅助安全团队优先关注真实威胁。



图8 威胁情报界面

脆弱性评估界面以资产视角聚合展示各项资产关联的漏洞清单。每个漏洞条目独立标注CVE编号、CVSS 3.1评分、EPSS可利用性概率、CISA KEV收录状态及VPT动态优先级评分，支持按优先级、状态、业务系统等多维度筛选。安全团队可直观查看每项资产的脆弱性全景，依据VPT优先级有序开展修复工作。

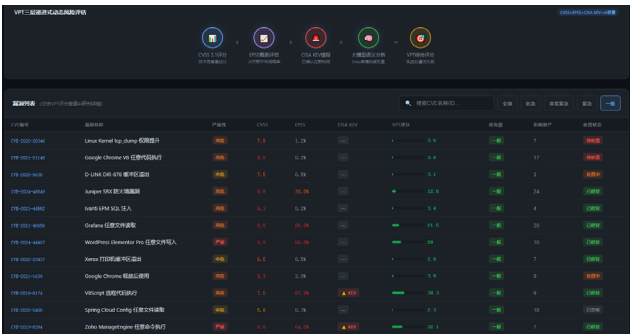


图9 脆弱性评估

3. 工单闭环与治理流程落地

对应3.3中“源头治理与风险处置”的设计，平台在落地层面实现了工单全流程管控。如图10所示，治理工单看板集中展示所有风险工单及其处理状态，涵盖工单编号、风险等级、关联资产、责任人、处置时限、当前状态等关键字段。工单与资产权属信息自动绑定（基于SCMDB信息表），处置状态实时同步，超时工单自动高亮提醒。看板与3.3.2中“精准定责与分级派单”的设计一致——危急风险24小时启动处置、72小时闭环；高风险3个工作日完成整改核验，超时自动升级督办。

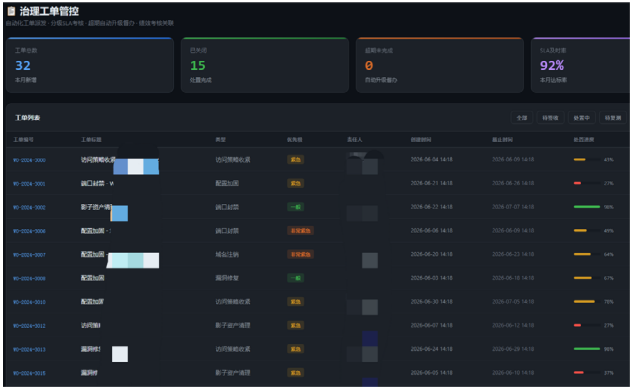


图10 工单管控模块界面

(二) 实践效果量化分析

体系上线后，通过日常轻量级资产巡检、非交易时段深度扫描、漏洞分级闭环处置和定期合规审计复盘，持续排查资产风险、规范治理流程、迭代优化体系能力。在感知层面，分布式扫描集群覆盖公司全域网络，融合SCMDB、PBI产品表、TDP流量、DNS解析记录等多源数据，形成统一的资产台账底座，有效识别并清退一批长期游离于监管之外的影子资产与僵尸资产，资产台账实现动态更新，底数清晰可查。在研判层面，VPT模型已完成多轮评估，研判结论与安全专家人工判定匹配度超92%，单条漏洞评估耗时由传统人工15分钟缩短至30秒以内，推动安全运维资源精准聚焦高风险资产。在治理层面，打通感知、研判、定责、派单、整改、核验、归档全链路，全量风险工单实现流程化闭环处置，SLA考核指标驱动整改效率得到显著提升。

评估维度	优化前	优化后	改善幅度
漏洞平均修复周期	15天	6天	提升60%
月度工单闭环率	65%	94%	提升29个百分点
影子资产/僵尸资产/高危端口	大量堆积	发现清除189项	大幅缩减管控盲区
必要紧急修复项	168	72	减少57%

表1 指标对比

整体来看，体系推动资产风险治理从“被动整改”向“主动预防、长效管控”转型，全面适配证券行业常态化合规治理与风险防控要求。

五、总结与展望

(一) 总结

本文针对证券行业资产底数模糊、风险度量失准、发现处置链路断裂、制度落地松散等核心痛点，立足行业强合规、高稳定、零中断的业务特性，构建了“智能感知—精准研判—闭环治理”一体化的资产风险感知与治理体系。

通过分布式探测架构与多源数据融合（主动扫描+被动流量+CMDB台账），建立覆盖多云、内外网、多业务载体的全域资产感知能力，配合AI驱动资产清洗、指纹识别与归属绑定，使每一项资产在发现之初即完成确权，有效解决了资产底数不清、权属不明的治理基础问题。

突破传统单一CVSS评分体系的局限，构建融合暴露面、脆弱性、威胁情报、业务影响四维度的综合风险量化模型，引入VPT多源融合漏洞优先级排序机制（CVSS定基准→EPSS做修正→KEV强制上浮→AI语义调权），将理论漏洞等级转化为贴合实战的处置优先级，有效解决了风险分级失真、治理资源错配的问题。

践行安全左移理念，将感知能力与合规要求前置至资产上线和变更环节——上线即完成纳管验证，变更即触发基线监测，合规制度转化为自动化前置检查项，在风险形成之前将其阻断；向后打通“定责—派单—整改—核验—归档”完整处置链路，配套分级SLA考核与常态化运营机制，实现风险全生命周期可追溯、可管控、可清零，推动资产风险治理从被动运维向主动防控、从碎片化管控向体系化运营的转型。

体系在实际运行中仍存在若干可优化之处：系统冷启动阶段，因资产数据积累不足，AI研判精度会出现短暂波动；对于IPv6新型资产和加密流量承载资产的识别覆盖率尚有提升空间；大模型API高频调用存在重复问题，消耗AI平台资源。

(二) 展望

后续将围绕现有短板持续迭代优化：一是深化大模型行业场景适配能力，优化专属提示词模板，逐步实现合规报告、修复方案、态势分析和风险研判的全自动生成，进一步提升治理智能化水平；二是拓展供应链资产风险治理，覆盖第三方合作机构关联资产，构建全域供应链风险感知与联防联控体系；三是优化云原生和IPv6新型资产的探测与研判能力，以适配行业IT架构的迭代趋势；四是后续可通过固化高频场景的提示词模板和引入语义缓存，避免同一类漏洞反复调用大模型产生重复资源消耗。

参考文献

[1] 中国证券监督管理委员会. 证券期货业网络和信息安全管理 办法(证监会令第218号) [Z]. 2023-02-27.

[2] 黄施宇. 基于证券行业资产风险视角的漏洞全生命周期的 管理研究[C]//2023年网络安全与数据治理学术会议论文集. 2023.

[3] 腾讯安全. 数字化业务风险主动防御与软件供应链安全 治理核心洞察[EB/OL]. (2026-05-30). <https://cloud.tencent.cn/developer/article/2677904>.

[4] 国家计算机网络应急技术处理协调中心 (CNCERT). 2023 年中国互联网网络安全报告[R]. 北京: CNCERT, 2024.

[5] Gartner. Hype Cycle for Security Operations, 2023[R]. Gartner Research, 2023.

[6] 陈云领, 周轶伦. 广发证券网络安全漏洞管理研究与实践 [J]. 中国金融电脑, 2022(11).

[7] 曹旭博, 钟夫. 一种漏洞优先级评估模型算法[J]. 通信技术, 2023, 56(12): 1411-1417.

[8] 邓胜利, 陈舫, 袁许俊. 人工智能赋能的智能化漏洞治理研 究[J]. 中国信息安全, 2025(11): 27-31.

金融机构小程序安全合规关键技术研究与实践

陶剑峰、张强、徐孝成 | 中信建投证券股份有限公司

摘要： 随着移动互联网的深度普及，小程序凭借其轻量化、跨平台的独特优势，已成为金融机构深化数字化转型与客户触达的关键渠道。然而，金融行业的强监管、高风险及数据高敏感性，对小程序的生命周期安全和合规提出了远超一般行业的刚性要求。特别是伴随大语言模型与智能体技术的演进，宿主平台原生AI助手对小程序的直接操控更是加深了这一影响。文章聚焦于金融机构小程序安全合规的关键技术，系统性的梳理了从设计开发到监控运营面临的核心挑战与实践框架，旨在为金融机构小程序在业务创新与监管合规的动态发展中提供兼具理论深度与工程参考价值的技术范式，保障其稳健与高韧性发展。

关键词： 小程序、安全开发、合规展业、数据安全、个人信息保护、AI助手

一、引言

近年来，以微信小程序、支付宝小程序等为代表的轻量化应用生态，正在深刻重塑金融服务业的交互模式与用户习惯。依托超级App流量优势，小程序生态成功将用户规模转化为多元服务场景，实现商业边界的持续拓展^[1]，涵盖政务、办公、教育等多元场景，其中金融理财类应用凭借其核心业务需求与高频交互的特征，长期位居用户使用频次与停留时长的头部区域。金融机构通过小程序提供包括账户开立、信息浏览、资产查询、财富管理等在内的全栈金融服务，极大地拓宽了普惠金融的覆盖广度与便捷程度，该渠道已由最初的辅助性流量入口蜕变为不可或缺的战略级核心展业渠道。

然而，这种便捷性的背后，隐藏着复杂且严峻的安全与合规挑战。与传统原生应用程序相比，小程序运行于第三方宿主平台提供的沙箱环境中，其技术架构、数据流向及安全边界存在显著差异。开发者对底层沙箱运行环境的控制权天然受限，代码在前端易遭受反编译及逆向工程威胁，且数据的部分传输和存储高度依赖宿主平台开放的API能力。同时，金融业务特有的高资产价值与数据高敏感性，如用户身份信息、资产状况、交易流水等，使其沦为黑客自动化攻击与渗透的重点目标。屡禁不止的虚假宣传、诱导消费、交易欺诈、过度索取权限等违规行为，不仅侵害消费者合法权益，更可能引发金融风险与声誉风险。特别是伴随大语言模型与智能体技术的加速演进，宿主平台原生AI助手（如微信小微、支付宝阿宝）可直接操控小程序内部控制流，这种新型交互范式彻底打破了传统的边界防护体系，带来了诸如提示词注入^[2]与越狱攻击^[3]等前沿安全合规挑战。

金融机构小程序的安全合规建立在《网络安全法》、《数据安全法》、《个人信息保护法》、《网络数据安全管理条例》等法律和行政法规之上。中国人民银行、工信部、公安部、国家市场监督管理总局、国家金融监督管理总局、证监

会等部门持续强化对互联网应用的监管力度，先后出台了多项管理要求，如《App违法违规收集使用个人信息行为认定方法》、《关于加强银行业保险业移动互联网应用程序管理的通知》，明确将小程序纳入移动应用管理，其需要满足网络安全、数据安全、隐私保护、合规展业等要求后方可上架发布^[4]。中国互联网金融协会等行业自律组织亦积极推动标准建设，于2025年正式发布了T/NIFA 33-2025《移动金融小程序安全要求》团体标准，从行业层面系统化规范了安全开发与个人信息收集的技术基线^[5]。此外，微信、支付宝等平台自身针对金融类小程序也有特殊资质审核要求，需要企业进行认证并上传相关许可证和展业牌照。这意味着，各层面已对金融机构小程序的安全技术能力和合规治理水平提出了系统性要求。

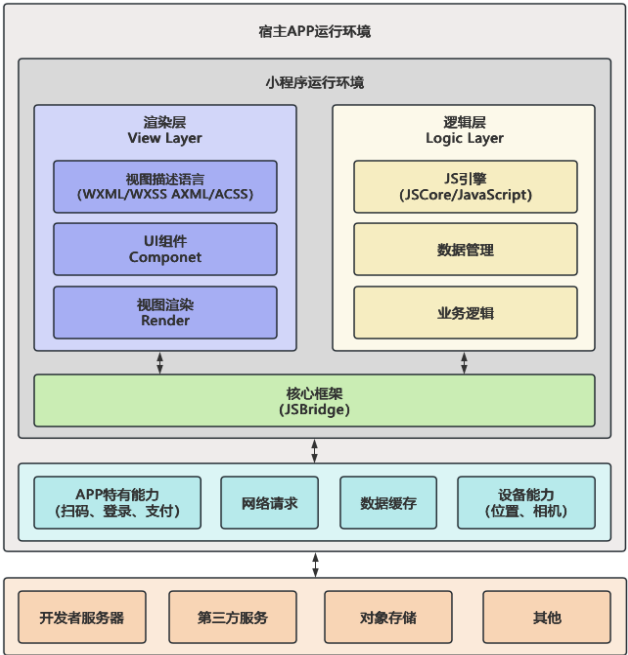


图1 小程序技术架构示意图

本文旨在结合该领域的理论研究成果与行业工程实践，系统性地探讨金融机构小程序从开发到运营全生命周期的安全保障体系构建与合规遵循路径。本文的组织结构如下：第二部分深入分析金融机构小程序的发展现状与核心安全风险；第三部分详细构建金融机构小程序全栈纵深防御的核心技术体系；第四部分讨论目前面临的主要挑战；第五部分总结全文。

二、金融机构小程序发展现状与风险分析

（一）金融机构小程序发展概况

金融机构小程序的业务布局经历了从信息浏览到交易闭环的阶段。早期阶段，银行和券商主要利用小程序发布网点信息和产品介绍，功能和交互模式较为单一。随着平台能力的增强和监管框架的逐步明确，金融机构小程序的功能逐步拓展至流水查询、转账汇款、基金申赎等核心金融业务环节。现阶段，部分领先的金融机构已打破传统的孤岛化服务模式，将小程序打造成为集营销获客、客户服务、场景金融于一体的综合化数字平台。

（二）主要安全风险分类与成因

1. 技术维度

小程序应用天然的混合架构导致其安全边界高度依赖端侧沙箱。首先是端侧源码逆向与逻辑泄露风险。宿主平台在客户端本地文件系统中采用了专有的单体归档缓存机制（如微信生态的wxapkg格式包），攻击者极易通过本地文件提取与自动化反编译工具，复原出近乎明文的前端逻辑代码。这不仅导致应用的核心业务逻辑暴露，更直接造成前端硬编码的令牌或敏感密钥信息的直接泄露。其次是网关接口越权与重放攻击威胁。小程序高度依赖服务端API接口进行数据交互，若服务端缺乏严密的身份鉴权与参数校验，极易导致水平/垂直越权攻击。最后，传统互联网应用面临的安全问题在小程序生态中依然存在，诸如SQL注入、跨站脚本（XSS）以及应用层拒绝服务等攻击。

2. 合规维度

根据中央网信办、工业和信息化部、公安部联合发布的《关于开展2026年个人信息保护系列专项行动的公告》和通报的国家计算机病毒应急处理中心检测结果，71款移动应用存在一项或者多项违法违规收集使用个人信息情况。其中金融类应用因涉及敏感信息而备受关注，金融领域违法违规收集使用个人信息也在此次专项行动的治理中。检测结果通报的典型问题包括：首次运行时未通过弹窗等明显方式提示用户阅读隐私政策等收集使用规则；隐私政策未逐一列出收集使用个人信息的目的、方式、范围；未向用户提供撤回同意收集个人信息的途径、方式；未采取相应的加密、去标识化等安全技术措施等^[6,7]。此外，金融机构小程序易涉及虚假宣传与误导性展示、广告违规、跨境数据传输违规、未成年人保护等问题。这些违规行为极有可能面临监管处罚、产品下架的风险，并导致机构声誉严重受损。

3. 业务维度

小程序中的营销活动，如抽奖、加息券、红包等是营销套利和业务欺诈的重灾区。黑灰产利用批量注册的虚假账号和自动化脚本，在毫秒级别内完成领券、下单等操作，造成金融机构真正的目标客户拿不到福利与财务损失。恶意的网络爬虫针对接口数据的高频抓取，不仅会拖慢服务器的响应速度，还会导致业务数据泄露和用户隐私受损。此外，不法分子构建高相似度的仿冒小程序对企业 and 用户造成的危害是极其深远且致命的，它不仅导致核心业务流量被截流、钓鱼攻击、客户投诉，甚至使企业面临法律责任。

三、金融机构小程序安全合规关键技术体系构建

金融机构小程序的安全合规是一个系统工程，需遵循安全左移、纵深防御、持续反馈、合规贯穿全程的原则，将安全合规理念融入设计、编码、测试、监控和运营的每一个环节。

（一）安全合规设计

金融机构小程序安全防御体系的构建应严格遵循安全左移原则，将安全考量与合规约束深度融入需求分析与系统架构设计的初始阶段。

1. 需求分析和设计

在需求分析和产品设计阶段，企业应该梳理适用法律法规（《网络安全法》、《数据安全法》、《个人信息保护法》等），分析监管文件，识别行业标准要求，明确平台准入资质（金融牌照、ICP备案、监管机构备案等），从而建立合规矩阵，逐条映射至产品功能需求，确保小程序的业务功能从设计源头上完全收敛在安全合规的刚性框架之内。

2. 安全架构设计

在系统架构设计上要建立零信任意识，假定前端的所有代码是完全暴露的、前端发送的所有数据是可被篡改的、前端所处的移动端宿主环境是不安全的。遵循纵深防御原则，从宿主层、应用层、网络层、服务层、数据层构建多层次的安全防御架构。



图2 小程序多层次安全防御架构图

(二) 安全编码实现

安全编码是保障金融移动客户端合规展业的基石，相应的安全规范与最佳实践也日益体系化。然而，鉴于小程序特有的运行特性，开发者在编码实现阶段仍需严守以下关键技术红线。首先，严禁在小程序前端源码中硬编码任何敏感凭证与关键配置参数，诸如小程序密钥AppSecret、各类接口访问令牌、加签私钥以及开发阶段残留的账户信息等。其次，敏感业务逻辑诸如加密算法、交易模型、风控评级模型等高价值商业机密，必须强制部署于受信任的后端服务器侧，前端仅保留UI渲染与人机交互流，从源头上消除核心逻辑的暴露面。再者，前端代码必须实施编译期加固。鉴于小程序前端代码可下载并暴露于客户端，在工程打包阶段必须引入深度混淆机制，综合运用无意义字符替换、控制流平坦化^[8]以及死代码注入等技术手段，最大程度提高逆向工程的难度和成本。最后，收紧软件供应链的安全边界。开发者需建立常态化的第三方SDK与组件插件审查机制，通过构建动态的软件物料清单，持续跟踪并限时修补供应链层面的漏洞与缺陷。

服务端接口必须实现严格的越权防护。首先，系统可以引入基于JWT (JSON Web Token) 的状态令牌机制，严格比对Token中所包含的用户身份与请求资源的归属映射关系，以杜绝水平越权漏洞。同时，构建基于角色的访问控制 (RBAC) 权限模型，严格映射角色与操作权限，从而阻断垂直越权路径。其次，针对所有API请求实施签名校验，通过后端对报文签名一致性、时间戳有效性以及随机数 (Nonce) 唯一性的三重逻辑验证，构建防篡改与防重放的校验机制。针对转账、密码修改、大额支付等关键高风险操作，引入一次性短信验证码 (OTP) 或端侧生物识别等二次身份鉴别机制，以确保关键交易意图的真实性与抗抵赖性。此外，为了保障服务的高可用性，系统可在API网关层实施精细化限流策略，如针对单用户、单IP设定每分钟最大请求阈值，从而有效遏制恶意网络爬虫和应用层拒绝服务攻击。

(三) 数据安全和个人信息保护

1. 数据分类分级方案设计

数据分类分级是数据安全管理的基础性工作。金融机构小程序应参照《金融信息服务数据分类分级指南》(2026年)的框架，首先对小程序所涉及的数据资产进行全面盘点，梳理数据来源、存储位置和流向路径。在此基础上，按照业务属性进行一级分类 (业务数据、用户数据、企业数据)，然后进一步细化为二级和三级分类。分类完成后，依据数据的重要程度和敏感程度以及一旦遭到泄露、篡改、损毁等可能造成的危害程度进行分级，对核心数据和重要数据实施最严格的保护措施，还需要定期复核数据资源和数据分类分级情况，及时更新数据分类分级清单和重要数据目录^[9]。

2. 数据全生命周期安全防护

数据全生命周期安全防护应覆盖采集、传输、存储、使用、共享、删除/销毁等环节。采集环节遵循最小必要原则，仅收集与业务功能直接相关的数据，并向用户明确告知收集目的、方式和范围。传输环节对所有网络请求实施HTTPS/

TLS加密，目前主流小程序平台已要求网络请求必须通过HTTPS协议。存储环节对敏感数据实施加密存储，密钥与数据分离保管。使用环节对敏感信息实施脱敏处理，对数据的访问权限实施最小化控制。共享环节在向第三方提供数据前必须取得用户明确同意。删除环节在数据不再需要时按规范流程彻底删除。

阶段	核心安全要求	具体技术措施
采集	合法、正当、必要、用户知情同意	前置协议弹窗、分项授权、采集内容最小化
传输	机密性、完整性、防篡改、防重放	HTTPS/TLS传输、请求签名、时间戳校验
存储	分级分类、加密存储、异地灾备	国密算法加密、冷热数据分离、数据备份、密钥管理
使用	脱敏展示、访问控制、日志审计	敏感信息中间位遮蔽、RBAC权限模型、操作日志全量记录
共享	合规审批、期限管理、用户单独确认	数据共享审批流程、数据沙箱、隐私计算
删除/销毁	彻底清除、可审计	留存期管理、逻辑删除、物理销毁、销毁记录归档

表1 数据全生命周期安全要求与技术措施

3. 个人信息保护

在软件设计开发中，遵守个人信息保护的相关法律法规和行业规范早已经是强制项，如若失守，将导致监管处罚和法律责任。个人信息保护合规功能设计应紧扣《个人信息保护法》和历年个人信息保护系列专项行动中明确的重点治理问题。一是完善个人信息收集使用规则的公开透明机制，隐私政策需以弹窗等明显方式在用户首次使用时展示，清晰告知收集信息的种类、目的、方式和保存期限。二是优化用户同意机制，杜绝“默认同意”、“强制同意”、“捆绑同意”，为用户提供主动勾选同意的操作路径。三是保障用户权利行使功能，提供个人信息查询、更正、删除以及账户注销的便捷入口。四是加强敏感个人信息的增强保护，对生物识别信息、金融账户信息等实施加密存储和去标识化处理。五是向第三方提供个人信息的，必须向个人告知接收方的名称或者姓名、联系方式、处理目的、处理方式和个人信息的种类，并取得个人的单独同意。

(四) 全面测试与验证

小程序系统测试阶段除了例行常规的功能测试以外，在安全合规方面还需要做大量针对性测试，主要包括了：

(1) 自动化应用安全测试。包括静态应用安全测试SAST，扫描源码中的安全缺陷；软件组件分析SCA，扫描第三方库的已知CVE漏洞；动态应用安全测试DAST，对运行中的API进行黑盒扫描；交互式安全测试IAST，实时监控代码流、调用栈和内部数据流来识别漏洞；以及小程序包专项扫描，反编译后检测敏感信息泄露和危险API调用。大部分工具可以集成到CI/CD流水线中，在代码提交阶段即执行安全扫描。

(2) 渗透测试。覆盖小程序调用的重点API接口，模拟真实黑客进行越权访问、参数篡改、注入攻击、业务逻辑绕过、短信轰炸等测试。建议定期进行全面渗透测试，重大版本变更时追加专项测试。

(3) 压力与可用性测试。模拟多用户高频访问以评估系统的吞吐量与核心性能瓶颈，针对转账、理财申购等核心金融交易场景进行超负荷施压，检测高并发状态下的条件竞争、死锁及内存泄漏问题。主动模拟网络断连、数据库宕机、第三方依赖崩溃等突发异常场景，验证系统的快速容灾切换与降级能力。小程序端性能专项测试可深度检测小程序在低配设备或极端弱网环境下的白屏率、首屏加载时延及宿主主机的资源占用。

(4) 合规性验证测试。这是有别于一般功能测试的特殊测试维度。验证内容主要包括隐私政策一致性，确保实际数据收集范围是否与声明一致。授权机制有效性，验证分项授权弹窗是否在正确时机触发、撤回同意功能是否生效。账号注销完整性，验证在线注销流程是否可用、数据清除是否彻底。内容合规性，查看收益展示是否规范、风险提示是否完整、利率是否明示等。建议企业建设自动化合规测试能力，通过模拟用户操作路径自动验证关键合规功能的可用性。

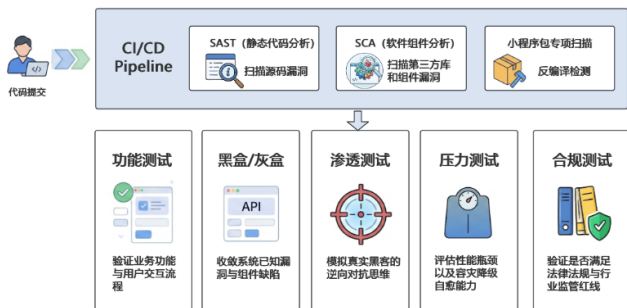


图3 小程序全面测试与验证流程示意图

(五) 运行时监控与风险处置

金融机构小程序上线后面临的安全威胁是持续存在且不断演变的，企业需建立覆盖运行监控、应急响应和仿冒监测的全维度监控处置体系。首先构建全天候、全栈式的动态监控技术矩阵，从三个维度实施监控指标的采集：一是应用可用性与性能指标，实时监测API响应时延、系统吞吐量及异常错误率，保障业务连续性；二是网络与应用层安全威胁监测，深度整合Web应用防火墙拦截日志，精确识别异常API调用模式、自动化Bot流量及应用层拒绝服务攻击态势；三是业务欺诈与合规风险防范，构建针对批量恶意注册、非正常交易流、以及恶意营销套利等高危行为的数字风控防线。

上述多源异构的监控数据可实时汇入大数据态势感知与分析平台，利用智能分析和关联算法，将碎片化的孤立事件编排为可溯源的攻击全链条。系统需配置严密的动态分级响应机制，一旦触发高危安全事件，系统将即时激活预设的安全响应流程，在秒级实施信道熔断或阻断攻击源，并同步成立由分管领导牵头的跨部门应急响应小组进行处理和复盘，从而实现动态感知、智能决策、自动响应的安全防护闭环。

金融机构因其品牌价值和多用户客群，是仿冒小程序的高频目标。企业应建立常态化的仿冒监测机制。定期以品牌关键词在各平台搜索排查，利用图像识别技术比对图标和截图，关注用户投诉中涉及仿冒的举报信息，订阅平台品牌保护服务。发现仿冒后应立即通过平台侵权举报通道投诉或付

诸法律手段，并同步评估和处理受损用户的问题。同时，在小程序首页展示官方认证标识，通过公众号和短信定期推送防骗提示，帮助用户辨别真伪。

(六) 常态化合规运营与持续改进

金融机构小程序的安全合规运营并非静态的，而是一个伴随监管政策演进、业务形态变更以及新型威胁涌现的动态治理过程。针对金融行业强监管、高风险的行业特质，企业必须将合规要求深度嵌入到日常数字化运营流程中。

制度是运营合规的基石，金融机构应制定覆盖全生命周期的核心制度文件，包括安全管理总则、安全开发规范、数据安全管理办法、个人信息保护管理制度、安全事件应急预案及第三方组件管理规定等。业务、技术、风控、法律等多个职能部门要有明确的责任划分，避免责任模糊和响应延迟。流程管控方面，所有重要变更应经过评审，对于涉及隐私协议修改、数据采集范围变更、用户授权流程调整等高风险变更，应增设法律合规部门的前置审批环节。

金融机构应建立定期安全评估和审计机制，评估内容涵盖技术安全、数据安全、合规风险等方面。建议定期参照监管通报中的常见违规情形进行合规自查，发现问题形成整改台账并跟踪至闭环，逐步完善问题发现、评估、修复、验证的全流程运营管理机制。

四、讨论

(一) 多平台开发的安全合规挑战

金融机构的小程序可能存在同时部署在微信、支付宝等多个平台的需求。每个平台的技术架构、审核标准、安全能力、开放接口各不相同。例如微信的code换取session_key机制，以及支付宝的auth_code换取access_token机制。平台差异导致安全与合规策略难以统一实施，金融机构需要针对不同平台分别进行安全适配和合规审查，管理成本显著增加。实践中，企业可采取统一后端与多前端适配的架构策略，将核心业务逻辑和安全策略集中在统一的后端服务中，前端仅负责UI渲染和平台适配，从而在一定程度上降低了开发和管理复杂度。

(二) 小程序与App安全的差异化处理

小程序与App在技术架构上存在本质差异，这决定了其安全策略不能简单照搬。小程序的安全边界受制于宿主平台，金融机构的可控范围有限。App可以实施客户端加固、证书锁定等技术措施，而小程序的安全加固更多依赖代码混淆、通信加密等有限手段。此外，小程序的发布审核由平台方执行，金融机构与平台方的协同配合成为安全管控的重要环节。因此，金融机构在小程序安全管理中需更加重视平台安全能力的研究利用，加强与平台方的沟通协调。

(三) 安全合规与用户体验的平衡

金融机构小程序面临的一个核心矛盾是安全防护力度与

用户体验流畅性之间的平衡，这也是大多数C端互联网应用普遍存在的问题。过于严格的安全策略虽然能有效降低安全风险，但会导致用户操作流程冗长、转化率下降。反之，过度简化安全流程以追求极致体验，则可能留下安全隐患。解决这一矛盾的关键在于对操作的风险等级实施差异化验证策略。对于低风险操作，如浏览产品列表采用无感知的校验。对于中风险操作，如修改个人信息采用轻量级验证（短信验证码、手势密码）。对于高风险操作，如大额转账、修改交易密码采用强验证组合（如短信验证码+人脸识别）。这样既能保障关键操作的安全性，又不显著影响日常使用的流畅度。

（四）人工智能等新技术的安全前瞻

随着大语言模型与智能体技术的爆发式发展，小程序生态正迎来前所未有的双刃剑效应。在攻防对抗端，AI既能赋能防御体系实现智能化的漏洞检测与威胁态势感知，也可能成为黑客实施自动化漏洞挖掘、绕过系统认证的致命武器。近期，随着原生AI助手（如微信小微、支付宝阿宝）全面打通小程序生态，用户通过自然语言即可直接操作小程序。这种新型交互打破了传统的UI前端控制流，导致AI助手在执行接口调用时，可能绕过原有的知情同意、隐私弹窗或上下文状态校验。这极易诱发提示词注入与越狱攻击，导致敏感金融数据在不可信的语义环境中泄露，甚至引发未经二次确认的敏感操作。未来金融机构必须从代码与接口安全合规向语义安全与模型内容与行为对齐^[10]等前沿领域全面延伸。

五、结论

本文针对金融机构小程序在深化数字化转型中所面临的安全与合规压力，系统性地研究并构建了金融机构小程序安全与合规的关键技术体系。企业需将安全与合规提升至战略高度，贯穿于产品设计、技术实现和运营管理的全程。针对当前热点如大语言模型和智能体技术给小程序带来的安全隐患等问题也进行了讨论分析。展望未来，金融机构小程序的安全合规防护范式将迎来较大变革。随着监管框架的持续完善、安全技术的不断创新以及行业生态的日益协同，金融机构需要在坚守安全合规底线的同时，充分释放小程序作为新型数字服务渠道的巨大潜力，为广大金融消费者提供更加安全、便捷、智能化的金融服务，实现商业价值与社会责任的统一。

参考文献

- [1] 2025全景生态流量年度报告[EB/OL]. (2026-02-03). <https://www.questmobile.com.cn/research/report/2018573146077106178>.
- [2] 台建玮, 杨双宁, 王佳佳, 等. 大语言模型对抗性攻击与防御综述[J]. 计算机研究与发展, 2025, 62(3):563-588.
- [3] 袁明, 邹其霖, 袁文骐, 等. 大语言模型提示词注入攻击与防御综述[J]. 信息安全, 2026, 26(3):341-354.
- [4] 关于加强银行业保险业移动互联网应用程序管理的通知[EB/OL]. (2024-09-12). https://www.gov.cn/zhengce/zhengceku/202409/content_6974749.htm.
- [5] T/NIFA 33-2025. 移动金融小程序安全要求[S].
- [6] 关于开展2026年个人信息保护系列专项行动的公告[EB/OL]. (2026-04-02). https://www.cac.gov.cn/2026-04/02/c_1776867645836849.htm.
- [7] 国家计算机病毒应急处理中心检测发现71款违法违规收集使用个人信息的移动应用[EB/OL]. (2026-06-03). <https://www.cverc.org.cn/zxdt/report20260603.htm>.
- [8] 陈沈杰. 基于链式结构的控制流平坦化技术分析[J]. 集成电路应用, 2025, 42(2):124-125.
- [9] 关于印发《金融信息服务数据分类分级指南》的通知[EB/OL]. (2026-06-13). https://www.cac.gov.cn/2026-06/13/c_1782919789934988.htm.
- [10] 张钰莹, 云静, 刘雪颖, 等. 基于反馈的大语言模型内容与行为对齐方法综述[J]. 计算机工程与应用, 2025, 61(20):75-104.

攻防实战中的安全运营： 基于AI研判的高效安全运营体系探索

夏英杰、赵川、严星宇 | 国联民生证券股份有限公司

李杨、王玥 | 北京神州绿盟科技有限公司

摘要：在证券行业数字化转型的浪潮中，网络安全已从传统的合规驱动转向实战化、智能化的新阶段。随着《证券期货业网络和信息安全管理办法》的落地实施以及“证券公司网络和信息安全三年提升计划”的深入推进，行业对安全运营体系的要求不再局限于“有无”，而是聚焦于“好不好用、扛不扛打”。攻防实战演练的常态化开展，使得传统安全运营中心面临的挑战从单一的技术防御转向多维度、高强度的持续对抗。面对日均百万级的告警量、APT组织的定向渗透、内部人员的异常操作、交易系统的实时性等处理要求，仅依靠传统的人工研判模式在实际的攻防实战中已难以为继。AI研判技术的崛起为证券行业安全运营体系的升级提供了关键支点。大语言模型（LLM）智能体协同、多源数据关联分析、安全编排自动化响应（SOAR）等技术的深度融合，能够推动安全运营由“被动告警响应”转向“主动威胁狩猎”，为打造依托AI研判的高效安全运营体系开辟全新路径。

关键词：攻防实战、AI安全运营、AI威胁研判

一、引言：研究背景与现状

（一）网络攻击新趋势

随着人工智能、大数据、自动化技术的快速发展，网络攻击已彻底摆脱传统“零散化、随机化、低技术化”的标签，逐步演进为高度组织化、工具化、自动化的产业级行为，攻击手段的智能化水平持续提升，攻击门槛大幅降低，对行业机构的响应能力与技术水平提出了新的考验。

从攻击技术演进来看，AI技术已成为攻击者提升攻击效能的核心赋能工具，推动攻击模式从“被动试探”向“主动精准”转型。攻击工具的标准化、模块化、开箱即用化，推动攻击门槛持续降低，使得网络攻击从“高技能少数人专属”向“低成本大众化”扩散。

（二）传统安全运营模式的瓶颈与面临的新挑战

面对攻击侧的AI智能化、产业化升级，当前大多数网络安全防御体系仍停留在传统被动防御模式，依托“单点防护设备+人工运营”的架构，难以适配新时代网络安全威胁的发展态势。传统安全运营模式的瓶颈主要体现在以下两个方面：

1. 安全运营效率低下，告警过载与误报问题突出，核心威胁难以精准识别。

随着数字化程度的提升，各类安全设备（防火墙、WAF、EDR、流量分析系统、日志审计系统等）的部署数量持续增加，每日会产生大量的安全告警数据。其中包含大量的误报、重复告警，而真正具有高风险的有效告警占比极低。

2. 依赖人海战术，运营标准化程度低，人力成本高企且难以规模化复制。

多数安全运营工作仍以人工操作为主，从告警研判、漏洞扫描、威胁狩猎，到事件处置、应急响应，每一个环节都高度依赖资深安全运营人员的经验与能力。这种“人海战术”式的运营模式，存在响应速度慢、准化程度低、人力成本高

等诸多弊端。

当前，AI技术已逐步成为攻击者会使用的新型技术工具，其利用AI可大幅提升渗透、隐蔽攻击效率，进一步扩大攻守态势的不对称性。

二、基于AI研判驱动的安全运营体系总体架构

（一）设计理念与核心目标

面对攻击侧的智能化升级与防守侧的现实困境，传统基于规则驱动、特征匹配的被动防御模式已难以持续。防守方须顺应技术发展趋势，尝试用AI对抗AI，以智能化技术提升防御效能，构建“主动感知、智能研判、自动响应、持续进化”的下一代智能安全运营体系。

基于AI研判的高效安全运营体系建设的核心目标包括：

1. 威胁检测方面，具备未知攻击检测能力，精准识别各种攻击变种，提前发现潜在的攻击行为；
2. 威胁研判方面，通过AI技术自动整合多源数据实现告警的自动降噪、研判和调查分析，为处置决策提供支撑；
3. 自动化处置方面，可对研判出的攻击事件进行快速响应与闭环处置，提升处置效率，降低攻击损失。

（二）体系架构设计

我单位态势感知平台以AI大模型智能分析技术为支撑，配套标准化安全运营工作流程，可对各类安全威胁开展实时监测、快速研判分析，实现威胁从发现、分析到主动处置的闭环管理。

AI安全能力平台采用智能体构建设施、智能体中心、运行管理、开发框架及统一接入层的分层架构设计，整体架构以深度落地AI安全赋能、提升安全运营实战效能为核心建设目标。

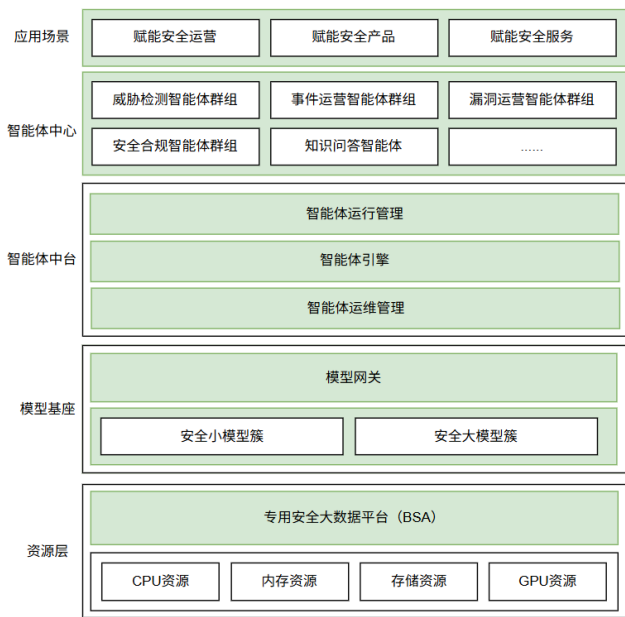


图1 AI安全能力平台架构图

三、AI研判核心技术体系

(一) 智能体驱动的研判引擎

AI智能体 (AI Agent) 为新一代安全运营提供了全新的技术实现路径。与传统AI辅助工具不同,智能体具备自主感知、决策和执行的能力,能够在动态环境中持续运行并完成任务。

威胁检测智能体:能够对网络流量、请求报文进行深度的安全语义和模式分析,突破传统检测基于已知规则和特征的局限性,提升对新型、复杂及隐蔽攻击的发现能力。事件运营智能体:通过降噪研判、调查分析、响应处置和报告生成等一系列原子智能体,构建完整的运营闭环,有效帮助安全人员从海量告警中筛选出高价值威胁,提升响应效率与准确性。

(二) AI降噪分诊

1. AI多手段自动降噪收敛

平台通过挖掘告警载荷数据,从试探性、利用性角度提取相关特征,结合变种SimHash算法生成向量化表示,用以刻画攻击意图并计算载荷相似度;利用深度自编码器依据重构误差识别异常告警,按告警结构化特征分组,在不误过滤异常告警的前提下选取覆盖范围最优的过滤规则,经后处理构建告警降噪基线。

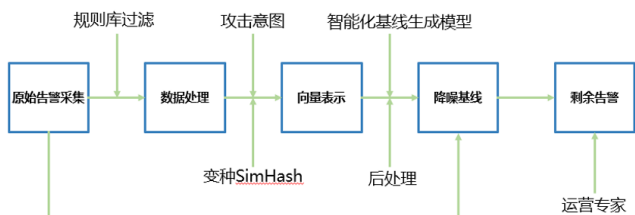


图2 AI多手段自动降噪收敛技术原理图

2. 正常行为识别降噪

企业内部正常的业务访问往往具备复杂度高、行为多样的特点,业务流量在策略规则匹配过程中极易产生误报告警,对风险告警的精准研判造成干扰。为有效对正常业务行为进行识别与降噪,平台以抓取正常业务告警共性特征为核心思路,实现正常业务行为的精准识别与降噪。

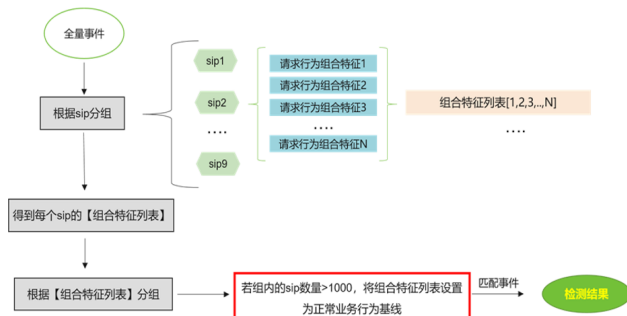


图3 正常行为识别降噪技术原理图

3. AI风险评分与智能分诊推荐

平台基于安全运营最佳实践、AI智能降噪小模型、AI大模型分析研判结果,构建全面的事件风险评分模型,实现全量事件的综合风险评估与分诊推荐。

告警事件智能研判与结果推导模块依托大模型推理分析能力,结合攻防最佳实践构建的专家思维链,自动解析告警上下文、关联日志数据与威胁情报,完成安全事件深度分析。该机制可快速判别攻击意图、验证攻击有效性、推导攻击影响,并输出自然语言形式的告警解读与攻击特征,提升安全研判工作的整体效能。



图4 AI风险评分与智能分诊推荐效果图

四、AI安全运营体系落地探索与实践经验

(一) 传统安全运营体系的升级融合

本次探索并非脱离传统安全架构的全新技术迭代,而是基于现有成熟安全运营基座的智能化升级。我单位主动审视现有安全运营体系短板,依托存量安全设备、日志系统、情报平台等传统技术基座,结合大模型多智能体能力开展针对性调整,探索传统安全运营与AI智能研判技术的融合方式,逐步形成适应网络安全攻防新态势的运营体系。

在基础运营能力方面,经过近几年的经验积累与技术融合,我单位目前已完成安全设备多源日志的统一接入与标准

化治理，解决了传统运营中数据分散、口径不一、联动困难的核心问题。依托统一数据底座，搭建内外部结合的威胁情报驱动机制，将安全告警、攻击特征、行业专项威胁规则进行关联匹配，打破了传统依赖单一设备特征告警的局限，让威胁研判有据可依、有源可溯。

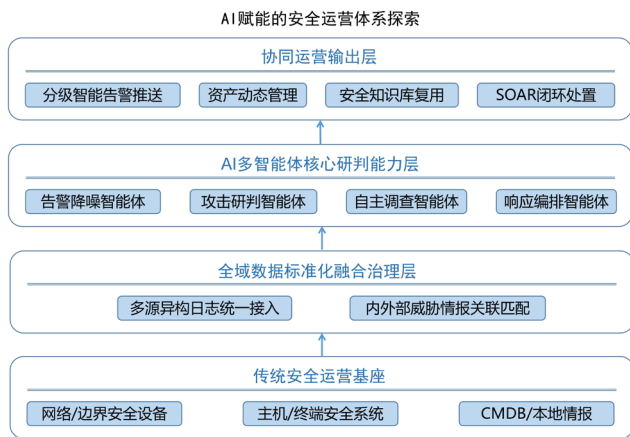


图5 AI安全运营体系探索实践

（一）传统安全运营体系的升级融合

在安全联动响应方面，对内打通重要安全工具自动响应通道，对外建立母子公司联防处置协同流程，在传统人工流转、线下处置的零散流程基础上，结合AI智能研判+自动化编排的模式，优化了以往流程割裂、响应滞后、处置标准不统一的运营痛点。

在精细化运营落地层面，公司安全团队目前已完成企业级安全知识库的搭建，整合近年攻防演练案例、漏洞处置规范、应急处置预案、行业共享攻击处置经验等存量数据，通过大模型训练优化，实现知识智能检索、处置方案智能推荐，弥补了传统安全运营过于依赖核心人员经验、新人上手慢、处置标准不一致的短板。

（二）资产全生命周期精细化管理

网络资产精细化管理是安全运营的技术底座与核心前提，所有威胁研判、风险处置均需以清晰可信的资产数据为支撑。传统安全运营模式下，企业内部安全团队往往会面临资产台账更新滞后、管理边界模糊等痛点问题。在网运行的业务服务器、终端办公设备、中间件、数据库等资产分散属于不同团队管理，人工定期盘点工作量大，影子资产私搭业务、临时测试服务器等极易形成安全管理盲区。安全运营平台产生的告警与资产信息容易出现割裂，发生威胁时无法快速判定受影响资产等级、业务重要程度，只能协同网络团队、资源团队、应用团队全网排查，拉长研判处置周期。

针对以上问题，我单位也在持续开展内外部资产全生命周期管理工作。

首先是打通内部网络访问策略，依托态势平台的资产发现能力每天定期进行资产自动扫描识别；

其次是结合内部CMDB平台获取资产归属团队、所属系统、负责人等关键信息，通过自动化技术手段替代传统

的人工盘点模式，极大地提高了资产信息的准确性和更新时效性；

最后是在资源申请流程方面，通过前置审批和后置安全纳管环节作为兜底管理手段，确保安全人员可及时掌握资产变动、系统上下线等信息。

此外，针对日常探测发现的影子资产、关键属性缺失资产，平台定期形成待确认资产清单，通过网络等团队协同，持续推进与未纳管资产的治理工作。

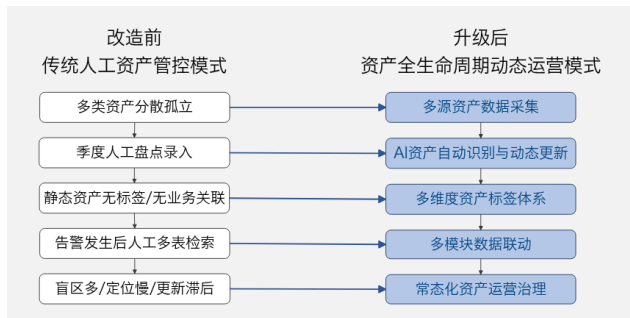


图6 资产全生命周期精细化管理

在此基础上，平台可实现资产与告警、漏洞、威胁情报数据的精准联动，当产生安全告警时系统自动匹配对应资产标签，同步展示资产归属等关键信息，同时优先高亮标识核心业务资产相关风险。通过全流程精细化改造，解决了以往资产底数不清、风险定位不准、管控滞后的运营短板，为AI场景化研判、分级告警推送提供可靠资产底座支撑。

（三）基于语义理解与行为逻辑的场景化分析研判规则

传统安全运营平台多基于源IP、攻击类型、告警事件名称等单一字段进行简单归类合并，存在事件碎片化、研判维度单一等问题。另外安全设备内置规则存在缺少业务特征、同业情报匹配能力不足等问题。为补全研判识别能力短板，我单位采用“通用规则基线+自定义规则+AI场景化规则”三层架构，搭建了基于语义理解与行为逻辑场景化分析的研判规则库。

首先通过对安全设备原生规则进行梳理，完成冗余、高误报规则筛选清理，形成标准化基础规则基线；其次段结合外部监管要求、行业共享情报等，新增金融钓鱼攻击、业务参数遍历、内部账号爆破等数十条专属自定义规则。基于以上规则，可将分散、多维度、跨设备的碎片化告警，按照完整攻击链路进行场景化聚类，最终形成一条具备完整攻击逻辑、可解释、可溯源的研判事件。以下为部分研判场景：

（1）暴力破解联动异常访问场景。系统将单设备多次登录失败告警、异地IP重试行为、后续少量成功访问、敏感接口试探请求等分散告警进行场景聚合，不再单独统计破解次数，而是形成“单IP持续爆破—试探绕行—弱口令尝试—后续微量存活访问”的完整入侵试探事件。

（2）文件上传与后续回连操控场景。整合文件上传告警、外联请求、异常进程启动、恶意域名访问等跨设备离散日志，还原“恶意文件上传落地—触发执行—外联C2—受控探测”的完整植入控制链路。

(3) 业务接口批量扫描试探场。对高频接口请求、异常参数遍历、短时多端点访问等行为进行语义聚合，识别攻击者对内部业务接口的系统性测绘与漏洞试探行为，避免单一访问告警造成的误判与漏判。

(4) 内网横向移动渗透场景。将单主机异常扫描、内网端口探测、跨主机账号登录、内网可疑外联等跨主机告警联动聚合，还原攻击者突破边界后在内网的扩散试探路径。

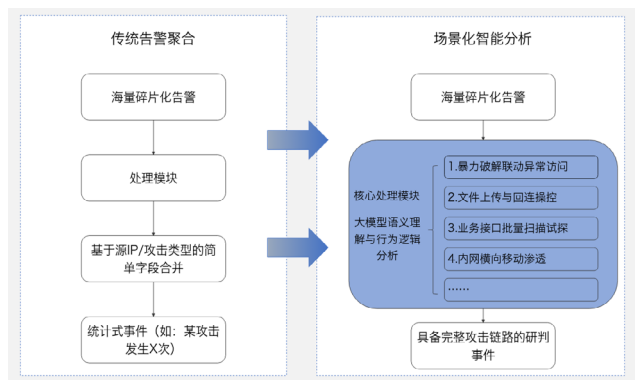


图7 场景化智能分析研判

上述场景化聚合方式突破了传统规则聚合的局限性，将相关网络安全告警事件由“字段统计式合并”转变为“行为逻辑式还原”，能够将大量碎片化噪声告警提炼为有限、清晰、具备完整攻击逻辑的研判事件，有效降低运营人员的事件梳理成本，为后续溯源分析、风险定级、闭环处置提供更加精准、完整的数据支撑。

（四）自动化联动响应运营优化实践

在日常运营工作中，我单位结合行业内外先进机构经验对现有安全运营体系流程持续进行调整优化，构建了分级编排响应处置能力。

针对低危扫描、批量异常访问等实际危害较低的告警，配置全自动处置策略，系统完成AI研判后自动执行IP封禁等操作；同时根据来源IP设置分级解封机制，国外威胁IP封禁14天、国内IP根据其属性（如运营商基站、云服务商、IDC等）封禁2小时至3天；针对Webshell植入、内网横向移动、漏洞利用等中高危威胁，设置“自动处置+人工复核”的处理机制，自动化执行初步隔离操作后，同步推送研判链路、资产影响、推荐处置方案至安全运营人员进行复核确认。

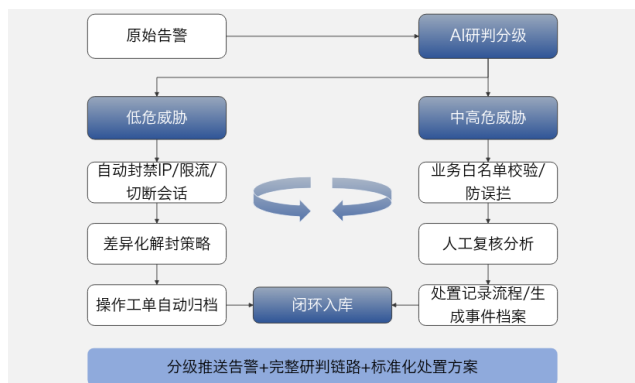


图8 自动化响应分级处理机制

此外还建立有自动化处置校验机制，针对重要业务增设业务白名单，避免自动化策略误拦截影响业务稳定运行，形成“监测-研判-处置-复盘-优化”的可持续迭代的联动响应运营模式。

五、攻防实战场景下的运营成效分析

我单位在近期攻防演习行动中通过态势感知平台与安全大模型的深度集成，充分发挥了大模型多智能体协同能力，全面覆盖告警降噪、事件智能研判、调查取证、自动化响应及智能报告生成等核心运营场景，有效提升了安全运营的研判精度、处置效率与应急响应速度。内部态势平台汇聚海量安全日志数据，经平台事件聚合过滤后对原始日志降噪效果达到95%以上；经AI智能研判后，进一步对事件进行降噪处理，智能研判出少部分高价值安全事件，使安全运营人员能够专注于处理最为准确紧迫的威胁。

AI降噪推荐



图9 AI辅助研判效果展示

六、结束语

海量告警与高误报的行业难题，让传统安全运营始终存在人工分析成本高、事件响应滞后的痛点。我单位以态势感知平台为基础，大小模型与多智能体协同为核心，尝试构建了包含告警降噪、研判、调查、报告、响应五大智能体自主协同的新安全运营体系。经过攻防演练与常态化运营验证，本次改造实践已初步显现运营效能提升价值。该体系依托传统运营基座夯实底层能力，借助AI多智能体能力提升运营上限，有效解决了人工运营效率低、误报率高、经验壁垒强、处置滞后等痛点，为证券行业机构传统安全运营体系智能化升级提供了可参考的实践经验。

证券期货行业规约驱动开发安全编程实践

蔡子豪、姜娜、邢骁 | 西部证券股份有限公司

摘要：基于大语言模型（LLM）的编程正深刻改变软件工程范式，随着Anthropic和OpenAI发布的模型在编程上的优异表现，软件开发的核心正在从代码开发能力转变为意图表达。为解决这个问题，规约驱动开发（Specification-Driven Development，简称SDD）作为一种实践方法被提出。规约驱动开发强调“契约先行”，核心原则是要在编写任何实现代码之前，必须先定义、评审并确认精确的、无歧义的“系统规约”（Specification）。这些规约作为系统的单一事实来源，约束大模型按照规约完成代码实现、测试用例生成以及系统集成。区别于传统SDLC与DevSecOps在各阶段植入安全活动、以流水线检测与签批为主的软件安全开发过程，SDD以规约约束AI实现路径，代码由大模型生成，有助于降低部分实现层常见漏洞的引入。但其标准流程仍在安全层面仍存在风险，现有实践通常只检查规约功能正确性，难以系统承载鉴权、脱敏、审计与抗滥用等行业安全要求。为此，本文提出SDD安全编程实践框架，将安全需求与功能需求同源规约、同权校验，构建安全Skill，在规约（Spec）、方案（Design）、拆分（Split）、实现（Implement）四阶段构建闭环机制，分别执行风险识别与控制生成、技术映射、任务对账、扫描测试，并在各阶段设置安全闸门。该框架将SDL的关键活动前移到规约与任务层，形成“需求—控制—实现—验证”的可追溯证据链，使AI代码生成始终受安全规约约束。以“按客户号查询资金余额”接口为例，实践结果表明，该方法可有效降低越权、注入和敏感信息泄露风险，提升AI编程在高合规场景下的可控性、可审计性与工程落地效率。

关键词：规约驱动开发、安全开发框架、证券行业、AI编程、SDL

一、概述

Andrej Karpathy将软件工程分为1.0到3.0的三次革命性变更^[1]。软件工程1.0范式中，人类通过编程语言编写计算机程序，计算机根据程序指令执行特定的任务。在这个时期，程序员编写显式的代码（如Python或者C++），告诉计算机如何处理每一个逻辑步骤。在软件工程2.0范式，人类已经不需要直接写数据处理逻辑，而是通过训练和优化神经网络，告诉计算机如何处理数据。到了软件工程3.0，程序员通过提供提示词（Prompts）控制模型进行编程，编程的指令从复杂的程序代码，变成了人类的母语，通过大模型实现人类表达的意图。

在软件工程3.0之前，人类生产代码、数据，并验证最终的程序，产生的信息安全漏洞主要包括：1. 编码缺陷类漏洞，如SQL注入、反序列化、缓冲区溢出等；2. 设计与配置缺陷，如访问控制设计不当、密钥硬编码、错误的配置、平行越权等；3. 供应链安全风险，引入包含漏洞的开源组件、第三方库。这些漏洞的共同特点是产生环节明确，分别在需求、设计、编码、构建、部署等环节，存在缺陷。传统的应用安全防护方式，是在软件各活动阶段嵌入安全检查，通过安全需求分析、静态代码检测、安全编码规范检查、软件成分分析、交互式安全检测、黑盒漏洞扫描、配置核查等方式，在软件的开发交付链路上设置门禁检查，对查出的风险进行控制。

在软件工程3.0时代，编程的本质已经改变，人类已经不再是告诉计算机怎么做，而是告诉计算机做什么，在这种新的编程范式中，人类生产提示词（Prompt），大模型理解人类的意图，生产代码和程序。2023年OpenAI发布了GPT-4模型，从这个阶段开始，模型能写代码获得了广泛认可^[2]，但是代码的质量和安全性仍然存在质疑。CodeRabbit于2025年12月发布了《2025年度AI与人类代码生成现状报告》^[3]，通过470个开源GitHub PR进行分析发现，由生成式AI共同撰写的代码，所含重大问题约为人工代码的1.7倍，错误配置增加75%，安全漏洞增加2.74倍。随着模型编程逐渐开始产品化和规模化，技术债不断积累，大多AI编程便会撞上“三月之墙”。

为了解决这个问题，规约驱动开发（SDD）逐渐被采用，它摒弃源代码，将一种机器可读的规格文件作为软件工程中人类交付文件，规定了软件的开发交付结果与绝对事实来源，将模型不可预测的输出代码，转变为可靠的执行。区别于传统应用安全由流程节点与人工闸门保证，SDD模式开发的编写者产出Spec，由大模型每分钟产出数百行代码的AI，相应的安全问题随之显现：

1. 人工审查困难，AI的代码产能远超人工审评带宽，以GitHub Copilot、Cursor等工具为例，其在标准提示下的代码生成速度可达人工编写的5-10倍。在此压力下，末端审查不得不由逐行审查退化为抽样式的概览放行。一个越权检查的缺失可能被一致地复制到多个接口，人工抽样审查难以

捕获此类系统性偏差。

2. 业务语义的理解断层, AI编码工具擅长写“能跑的代码”, 却读不懂“该防止什么样的漏洞出现”。在证券行业, 同样的技术盲区反复出现: 客户信息接口返回了明文身份证号, AI觉得“数据取出来了, 没问题”; 转账接口只验了登录态就放行, AI觉得“用户已登录, 很合理”; 查询持仓时没做用户隔离, AI觉得“SQL查对了, 逻辑通”。这些代码单独看都正确, 但放在证券业务的语境里, 每一条都是安全风险。

3. 安全无法嵌入模型生成代码的过程, 传统SDLC的安全工作通常在阶段活动之后, 如代码编写完成后进行扫描、审计, 系统持续部署完成后开展黑盒扫描。即使是IDE代码检测插件嵌入到了编码过程中, 也需要在人完成代码编写后进行检测并与代码开发人员进行交互。这一模式的本质是事后检测, 而非事前预防。当AI以极高速度产出代码时, 末端检测的成本急剧上升——同一漏洞在多个接口中复制, 末端扫描需重复发现多次; 若安全要求被前置写入规约并嵌入流程, 漏洞在产生之前即有机会被阻断。

在本文中, 我们并不试图去探讨SDD安全开发的顶层方法论, 仅针对以规约驱动开发这样的新AI开发范式, 在SDD各个阶段嵌入安全工作, 探究在证券行业实现SDD安全开发的一种实践方法。

二、背景

规约驱动开发 (SDD) 是一种编程开发方法论, 通过以规格说明文件作为模型编写代码的准则, 使得模型以约定的规格设计、编码、测试和验证程序, 从而有效的实现软件开发。在Addy Osmani的AI编程企业级Skills项目仓库中, 用一句话描述了规约驱动开发的本质: “在写任何代码之前, 先写结构化的规格文件。规格文件是模型与人类工程师之间共享的事实来源, 它定义了我们在构建什么、为什么构建, 以及我们如何知道程序已经完成了”^[4]。

证券行业兼具高数字化与高实时性的技术特征, 注重系统稳定性、合规性与可追溯性。AI编程的介入将研发速度提升了数倍, 但对于模型编码模式带来的不可控性, 产出不可预期、过程不可追溯、质量不可复现的问题, 在证券这种低容错场景中会被急剧放大。如何让AI既能加速开发, 又不触碰合规红线, 成为行业关注的焦点, 规约驱动开发是一种较好的实践方式。

(一) 规约驱动开发流程

规约驱动开发的核心思想是: 先把对错说清楚, 再动手实现。规约是唯一事实来源, 代码是规约的产物, 前一步不达标则不进入下一步。其流程分四步:

Specify (规约): 界定意图与验收标准, 用一份结构化规约描述系统应做什么、什么情况下算做对了。

Plan (方案): 确定技术结构与选型, 用设计文档回答怎么做。

Tasks (拆分): 拆分为可执行任务, 每个任务有明确的

输入输出和验收标准。

Implement (实现): 由AI按任务编码, 产出始终对齐规约。

四阶段之间形成严格的递进关系: 前一阶段达标后方可进入下一阶段, 规约作为唯一事实来源约束全过程。

(二) 规约驱动开发工具介绍

近半年来, 一批工具推动了SDD的实际落地, 尽管工具功能不一, 但都遵循着上述四个流程步骤, 并在每个边界设立明确门禁。

工具	定位	主要能力
Speckit	规约管理	将规约作为核心资产进行版本控制与协作审批
OpenSpec	规约标准	尝试建立跨团队的规约标准格式, 便于工具对接
Superpowers	AI编程增强	引入Task Master思路, 将规约自动拆解为任务并跟踪

表1 SDD相关工具对比

这些工具的共同趋势是将规约从静态文档转化为可执行约束。在四阶段中, 人与AI的角色分工应当明确: Specify和Plan是人主导的设计活动, AI仅作为辅助; Tasks是人机协作; Implement是AI主导的执行活动, 人负责审查结果。这种分工避免了角色混乱: 人做决策, AI做执行。

SDD的价值可从两个场景理解。新系统开发中, 规约是蓝图, 从零描述待建系统, AI依此生成。增量开发中, 规约是围栏, 用变更规约圈定加什么功能、不动什么、影响到哪里, AI只在围栏内修改。但SDD本身解决的是可控问题, 而非安全问题。标准流程默认只规约功能正确性, 不会追问该次调用是否越权、输入是否合法、是否留痕可查。

三、安全控制融入SDD的实践设计

(一) Security Skill构建

在实践中, 我们参考了传统软件开发生命周期 (SDLC) 中SDL (Security Development Lifecycle) 的模式。SDL的核心理念是在开发各阶段植入安全活动——需求阶段做安全需求分析, 设计阶段做威胁建模, 编码阶段做静态扫描, 测试阶段做动态验证, 上线前做安全签批。这一模式的价值已被大量项目验证, 但其编排方式依赖人工逐节点推进, 难以适配AI编码的高产能节奏。

为了让安全要求像功能要求一样被规约化、自动化地嵌入开发流程, 我们将SDL的安全活动拆解为可规则化的约束模板, 与Speckit的规约管理能力相融合。具体做法是: 在Speckit生成接口规约的同时, 自动触发安全约束注入——就像规约中必须包含输入参数和返回值一样, 高风险接口的规约中也必须包含鉴权、脱敏、审计等安全约束。

这一机制在大量业务场景的测试中不断迭代完善。最初的安全规则库仅覆盖输入校验和鉴权两类约束; 在测试过程中, 我们发现资金类接口需要额外的限额控制, 查询类接口需要脱敏处理, 外部对接接口需要熔断保护——这些来自

真实场景的反馈被逐步沉淀为规则库的新条目。经过多轮迭代，Security Skill的规则库已覆盖鉴权、授权、输入校验、数据保护、合规拦截、审计留痕、限流熔断等七大类安全维度，能够根据接口的功能描述自动匹配适用的约束集合。

(二) Security Skill的工作机制

Security Skill本质上是一个基于大模型的安全分析引擎。它不是独立工具，而是一套嵌入SDD流程的规则模板与AI模型的组合。当开发者在Speckit中输入接口的功能描述后，Security Skill自动完成以下三步：

第一步：需求解析与风险识别。Security Skill读取接口的功能描述，识别其中的风险特征——是否涉及个人金融信息、是否触碰资金操作、是否对接外部系统、是否暴露敏感数据等。基于这些特征，调用内置的行业安全规则库，判定接口的风险等级。

第二步：安全风险分析。针对已识别的风险点，Security Skill通过大模型分析可能的攻击路径和安全漏洞。例如，对于“按客户号查询资金余额”接口，分析过程会识别出：客户号可能被篡改（注入攻击）、会话可能被冒用（越权访问）、余额数据可能泄露（敏感信息暴露）、高频查询可能拖垮系统（拒绝服务）等四类风险。

第三步：安全控制生成。根据风险分析结果，Security Skill自动生成对应的安全控制措施——输入校验规则、鉴权策略、数据脱敏方案、审计日志要求、限流熔断配置等。这些控制措施以结构化格式写入规约文档，成为后续阶段必须满足的约束条件。

生成的安全控制不是静态模板，而是基于大模型对当前接口特定场景的动态分析结果。Security Skill将这三步的执行过程和分析依据一并记录，形成可追溯的安全决策链。开发者可以审阅Security Skill的产出，接受、调整或补充其生成的安全控制，但不可跳过——闸门会检查。

Security Skill在SDD开发四阶段中的介入方式存在明确分工：规约阶段以“大模型动态分析”为主导，针对接口特定场景生成个性化安全控制；方案、拆分阶段以“规则模板”为主导，将安全控制标准化地转化为技术要求和执行任务；实现阶段两者结合，大模型生成针对性测试用例，规则模板执行标准化扫描。四阶段的总览如表2所示。

阶段	Security Skill角色	核心动作	产物	闸门判据
① 规约	大模型动态分析主导	风险识别 → 攻击路径分析 → 生成安全控制	含安全约束的接口规约	安全约束完整性
② 方案	规则模板主导	安全控制 → 技术实现要求映射	带安全方案的设计文档	覆盖全部约束
③ 拆分	规则模板主导	安全方案 → 可执行任务拆解 + 对账	含安全任务的清单	标准-任务对账通过
④ 实现	大模型 + 规则混合	静态扫描 + 测试用例生成 + 覆盖率验证	测试报告 + 扫描结果	测试全绿 + 覆盖率100%

表2 Security Skill四阶段介入总览

(三) 规约阶段 (Spec)

四阶段的起点是规约。开发者在Speckit中输入接口功能描述后，Security Skill以“大模型动态分析”为主导介入——读取功能描述、识别风险特征、分析攻击路径、生成个性化安全控制。

场景：开发团队需要新增“按客户号查询资金余额”接口。

安全动作（大模型动态分析）：Security Skill识别该接口涉及个人金融信息和资金数据，判定为高风险。通过大模型分析四条攻击路径（注入、越权、信息泄露、拒绝服务），针对每条路径生成对应的安全控制。

安全产物：带安全约束的接口规约文档

```
【接口】按客户号查询资金余额/api/v1/account/balance
【功能】按客户号查询该客户的资金账户余额
【Security Skill注入 - 大模型风险分析与安全控制】
  风险等级：高（涉及个人金融信息 + 资金数据）
  风险识别：
    - 客户号可能被篡改（注入攻击）
    - 会话可能被冒用（越权访问）
    - 高频查询可能DoS（拒绝服务）
  强制安全控制：
    1. 输入校验：客户号必须符合 /^[0-9]{10}$ /，非法格式直接拒绝（防注入）
    2. 鉴权控制：会话主体必须与查询客户号一致，不一致返回403（防越权）
    3. 数据脱敏：返回余额保留两位小数，不返回账户明细（防信息泄露）
    4. 限流熔断：单客户号60秒内最多10次查询，超限触发熔断并告警
```

图1 规约阶段产物：Security Skill生成的安全约束

安全闸门：安全约束完整性检查 — 缺失任意一项，规约退回补全。

(四) 方案阶段 (Design)

规约通过安全闸门后，进入方案阶段。此阶段Security Skill以“规则模板”为主导——将规约中的安全控制按预设映射规则转化为技术实现要求，不再依赖大模型的动态分析。

场景：规约通过双轨验收后，进入技术方案设计。

安全动作（规则模板映射）：Security Skill将每条安全控制按规则模板转化为技术实现要求：鉴权控制 → JWT验签 + 客户号一致性校验；脱敏控制 → 余额格式化策略；审计控制 → WORM存储 + 哈希链；限流控制 → Redis滑动窗口 + 熔断告警。

安全产物：带安全方案的设计文档

```
【方案】按客户号查询资金余额 - 安全实现
鉴权：JWT Token 验签 + 客户号一致性校验
脱敏：余额字段保留两位小数，不暴露账户明细子项
日志：写入 WORM 存储（一次性写入不可修改），日志条目带 SHA-256 哈希链
限流：基于 Redis 滑动窗口计数，单客户号 60 秒 10 次，超限返回 429
熔断：触发限流后生成告警
```

图2 方案阶段产物：安全约束的技术实现方案

安全闸门：方案必须覆盖规约中全部安全约束，未覆盖则退回。

(五) 拆分阶段 (Split)

此阶段Security Skill仍以“规则模板”为主导——将安全方案中的控制点拆解为可执行的安全任务，并执行标准-任务对账。

场景：方案签批后，安全任务与功能任务并行纳入迭代。

安全动作（规则模板拆解）：Security Skill将五条安全控制拆解为四条可执行的安全任务，每条标记风险等级和验收标准。同时生成对账脚本，统计规约中的安全控制条数与

任务清单中的安全任务条数，确保每条控制至少对应一个任务。

安全产物：含安全任务的任务清单

任务编号	任务类型	任务描述	验收标准
T-001	功能	实现按客户号查询资金余额接口	返回正确的余额数据，保留两位小数
T-002	安全	实现客户号格式校验与鉴权	非法客户号返回400，会话不一致返回403
T-003	安全	实现审计日志写入	操作记录写入WORM存储，带SHA-256哈希校验
T-004	安全	实现限流熔断机制	60秒内超10次触发429并告警

表3 拆分阶段产物：含安全任务的任务清单

安全闸门：安全任务必须与功能任务同优先级，缺失安全任务则拆分退回。

(六) 实现阶段 (Implement)

开发完成全部任务后，进入最后一道安全闸门。此阶段 Security Skill采用“大模型 + 规则模板”混合模式——规则模板执行标准化静态扫描，大模型针对接口特定场景生成针对性测试用例。

场景：编码完成后，进入最终安全验证。

安全动作（混合模式）：规则模板执行标准化静态分析，扫描代码中的注入漏洞、硬编码密钥等通用缺陷；大模型基于规约中的四条攻击路径，生成五条针对性安全测试用例。两者结合验证安全约束覆盖率。

安全产物：安全测试报告 + 代码安全扫描结果

【安全测试用例 — Security Skill 生成】	
用例1：越权访问 — 使用A客户会话查询B客户余额，预期返回403	
用例2：SQL注入 — 客户号输入 ' OR '1'='1，预期返回400	
用例3：脱敏验证 — 查询结果仅返回余额数值，不包含账户明细	
用例4：日志完整性 — 篡改日志文件后哈希校验失败，触发告警	
用例5：限流验证 — 60秒内第11次查询返回429并触发熔断告警	
【代码扫描结果 — Security Skill 静态分析】	
扫描范围：输入校验、鉴权逻辑、脱敏处理、日志写入、限流配置	
发现问题：无高危漏洞，1个中危（日志路径未配置绝对路径）已修复	
安全约束覆盖率：5/5（全部通过）	
安全闸门状态：通过	

图3 实现阶段产物：安全测试用例与扫描报告

安全闸门：安全测试用例全部通过 + 代码扫描无高危漏洞 + 安全约束覆盖率100%，否则退回修复。

(七) 全程追溯

每个阶段的安全产物均纳入版本管理，形成“规约约束 → 方案实现 → 任务分解 → 测试验证”的完整证据链。以“按客户号查询资金余额”接口为例，从Security Skill的风险识别到最终代码合入，全过程留下可追溯的产物：

规约阶段产物：Security Skill的风险分析报告 + 五条安全约束（输入校验、鉴权控制、数据脱敏、审计日志、限流熔断）

方案阶段产物：技术实现方案（JWT鉴权、WORM日志、Redis限流等）

拆分阶段产物：四条安全任务（T-002至T-005）与功能任务（T-001）并行的任务清单

实现阶段产物：五条安全测试用例全部通过的测试报告 + 静态扫描无高危漏洞的扫描报告

审计时可逐条追溯：规约中的“鉴权控制”约束 → 方案中的“JWT Token验签 + 客户号一致性校验”实现 → 拆分中的T-002任务 → 实现阶段的越权访问测试用例通过记录。这条完整的证据链正是证券行业监管审计与事后回溯的核心依据。

(八) 全局流程图

表4展示了从需求输入到代码合入的完整流程。Security Skill在四个阶段分别以不同模式介入，与Speckit的规约管理、开发者的审阅决策、自动化测试工具协同工作，形成闭环。

流程节点	执行主体	Security Skill介入	产出 / 决策
① 需求输入	开发者	—	功能描述（自然语言）
② 风险分析	Security Skill（大模型）	读取功能描述 → 识别风险特征 → 分析攻击路径	风险等级 + 攻击路径清单
③ 安全控制生成	Security Skill（大模型）	针对攻击路径生成安全控制	安全约束集合
④ 规约生成	Speckit + Security Skill	功能规约 + 安全约束合并	双轨规约文档
⑤ 规约审阅	开发者	审阅 Security Skill 产出	通过 / 调整 / 补充
⑥ 规约闸门	自动化检查	完整性校验	通过 → 方案阶段 / 不通过 → 退回④
⑦ 技术方案	Security Skill（规则模板）	安全控制 → 技术实现映射	带安全方案的设计文档
⑧ 方案签批	安全负责人	人工评审	签批 → 拆分阶段 / 退回 → ⑦
⑨ 任务拆分	Security Skill（规则模板）	安全方案 → 可执行任务 + 对账脚本	含安全任务的任务清单
⑩ 任务闸门	自动化检查	标准-任务对账	通过 → 编码 / 不通过 → 退回⑨
⑪ 编码实现	开发者 + AI	按任务清单编码	功能代码 + 安全代码
⑫ 安全扫描	Security Skill（规则模板）	静态分析扫描通用漏洞	扫描报告
⑬ 测试用例	Security Skill（大模型）	针对攻击路径生成测试用例	安全测试报告
⑭ 实现闸门	自动化检查 + 人工复核	覆盖率验证 + 高危漏洞检查	通过 → 合入 / 不通过 → 退回⑬
⑮ 代码合入	版本控制系统	commit-on-green	安全产物归档

表4 SDD-Sec全局流程：从需求输入到代码合入

四、结语

AI编程的挑战在于它抽走了传统安全的隐含前提——安全由速度匹配的末端人工门户把守。规约驱动开发让开发重归可控，却只管功能；可控与安全之间的缺口，需要专门的框架来填补。

本文的核心思想是在规约驱动开发的框架上加入安全：通过Security Skill将安全需求与功能需求同等规约化，在规约、方案、拆分、实现四个阶段各设安全闸门，产出可追溯的安全产物，并与现有SDL安全生命周期融合，形成完整的安全实践架构。它不取代传统SDL的技术内核，而是把这些成熟手段从末端前移、装进规约驱动开发的框架，以适配AI产能暴涨的现实。

本文以“按客户号查询资金余额”接口为例，完整展示了Security Skill从风险识别到安全控制生成、从规约注入到测试验证的全流程。这一实践表明，安全控制的自动化嵌入在工程上是可行的——关键不在于AI有多智能，而在于约束AI的流程足够可靠。

后续工作包括：在更多证券业务场景中验证Security Skill的覆盖度；建立行业特定的安全规则库与风险分析模板；探索安全规约的自动验证与AI辅助威胁建模，进一步降低规约阶段的人工成本。

参考文献

- [1] Andrej Karpathy: Software Is Changing (Again), <https://www.bilibili.com/video/BV19fgrzJEGh/>
- [2] SWE-BENCH: Can Language Models Resolve Real-World GitHub Issues?, <https://arxiv.org/pdf/2310.06770>
- [3] CodeRabbit: State of the AI vs. Human Code Generation Report, <https://www.coderabbit.ai/whitepapers/state-of-AI-vs-human-code-generation-report>
- [4] AddyOsmani: Agent-Skills <https://github.com/addyosmani/agent-skills/blob/main/skills/spec-driven-development/SKILL.md>

证券期货业务电子认证电子签名技术应用标准研究

王恩潭、董延敏、张佳佳、马原、张秀敏 | 中泰证券股份有限公司

摘要： 证券期货业务以网上业务为主，电子认证、电子签名是提供安全保障的关键技术，然而业内对其认知尚不尽相同，有待进一步形成共识、确立标准，加强规范管理。本课题研究着眼更好落实实名制及具体监管要求，从研读法律法规、相关规定、标准、白皮书等参考文献入手，结合分析行业生产实践中存在的具体问题，通过厘清电子认证、电子签名的内涵和实质，梳理适用证券期货业务的电子认证、电子签名方式方法，借助公认、成熟的权威理论与标准框架，探讨构建起证券期货业务的技术应用体系及规则，初步形成了自洽、与现行国标行标相融的标准草案。本课题研究以“先立后破”方法论为指导，将本单位近年部分研究成果，包括“口令+‘三方协同鉴别’”强认证、“数字铅封”式电子签名等，同步纳入标准草案；在技术应用方面，对应现行国标行标落地了部分“分类匹配、分级管理”的具体设想和建议。

关键词： 电子认证、电子签名、eID、实名制、网上业务

本课题研究从研读有关法律法规、相关规定、标准、白皮书等参考文献入手，通过分析行业生产实践中存在的具体问题，结合梳理近年本单位研讨积累及课题组收集的有关最新研究成果等，分阶段推进。

2022年中泰证券立项年度课题《融合多方IT资源保障业务安全提升》，研究期间对生产实践中部分典型问题展开剖析，面向存管银行、电信运营商等开展调研，摸底可以为我所用的IT资源；2023年中泰证券立项年度课题《证券业务主体身份认证与电子签名技术标准》，形成《证券业务电子认证电子签名技术应用指引（草案）》；2024年中泰证券联合中泰期货成功立项证标委年度标准研究课题《证券期货业务电子认证电子签名技术应用标准研究》（BZKT-2024-032），面向整个证券期货行业就研究主题开展了更加广泛和深入的探讨，并就初步研究成果以问卷调查方式向部分行业专家小范围征求了意见，目前已初步成稿全面覆盖证券期货业务的标准草案。

作为本课题研究主要依据的法律法规、相关规定、标准和技术文献等共计30件，详细信息见参考文献[1]-[30]；作为参考的部分本单位前期研究成果多项，主要包括：《“面向客户的强身份认证应用模式”之探讨与实践》《证券业“网上开户”存在的问题及其化解之道》《“非现场销户”规范要义与转销户相关规定的讨论》《也谈证券业务中的电子签名技术应用——基于身份认证强度、施以分类管理，聚焦基本定义探讨“数字铅封”式电子签名》《大力推广应用“客户强身份认证”正当时》《“口令+‘三方协同鉴别’”——一种满足第三/四级等保要求、适用证券业务的，限定条件的强身份认证方法》《浅谈网上证券安全开户模式——完善技术措施，进一步落实开户人身份核验、无纸化签约2个关键环节》，详细信息见参考文献[31]-[37]。

一、背景

（一）国家宏观大背景

1. 《国家标准化发展纲要》指明方向、部署任务

2021年10月，中共中央、国务院印发《国家标准化发展纲要》，从国家层面为标准化工作指明方向，提出2025年和2035年2个阶段性发展目标，部署了“推动标准化与科技创新互动发展”等7项重点任务。

2. 《质量强国建设纲要》对高质量发展提出新要求

2023年2月，中共中央、国务院印发《质量强国建设纲要》，明确了质量强国建设的形势背景、总体要求和重要举措。指出面对新形势新要求，必须把推动发展的立足点转到提高质量和效益上来，坚定不移推进质量强国建设。

（二）证券期货行业背景

1. 法律层面新规定

2019新版《中华人民共和国证券法》通过修改第106条、新增第107条，强化了“账户实名制”规定和要求，不仅强调账户开立要实名，而且要求账户使用也必须实名。2022年8月1日起施行的《中华人民共和国期货和衍生品法》第18条亦有同样的规定和安排。

2. 《指导意见》提出3大目标、6项主要任务

2023年11月，证监会、国家标准委联合印发《关于加强证券期货业标准化工作的指导意见》，提出“标准化支撑作用更加显著”等3方面主要目标，以及“推动业务与标准化深度融合”等6项工作任务。

（三）部分具体监管要求

《证券期货投资者适当性管理办法》（2022年8月，证监会第202号令修正）关于非现场网上业务留痕安排的要求；

《证券投资基金经营机构信息技术管理办法》(2021年1月,证监会第179号令修正)关于电子合同使用管理的要求;《金融机构客户尽职调查和客户身份资料及交易记录保存管理办法》

(2025年10月,人民银行、金融监管总局、证监会[2025]11号令修正)关于客户尽职调查方面的要求;2023年1月,证监会第204号令发布《证券经纪业务管理办法》,要求“切实做好客户身份识别、客户适当性管理、账户使用实名制……”;2023年第5期《机构监管情况通报》,以“加强涉诈风险管控规范开立综合账户”为题,提出四项具体要求、首项便是“加强客户身份识别”……

二、研究目标

(一) 探研技术应用标准赋能高质量发展

以证券行业为例,截至2023年初,百余家券商、1.2万分支机构、35.5万从业人员,服务于2.1亿客户。行业经营机构以非现场业务为主,绝大部分种类的业务和客户服务均在网上开展、线上提供,电子认证、电子签名技术是提供安全保障的关键技术,然而大家对其认知尚不尽相同,有关技术应用缺失统一规范和标准。鉴于此、着眼更好落实实名制及具体监管要求,提出本课题研究的愿景目标,即:对电子认证、电子签名技术及其在行业的应用,进一步形成共识、确立标准,加强规范管理,为行业高质量发展增砖添瓦。

(二) 破题行业生产实践中部分共性疑难

本课题研究的初衷是探讨技术保障措施、解决行业生产实践中遇到的部分共性疑难问题。例如:2013年至今,“网上开户”的行业定义悄然发生了改变,从一开始强调依赖数字证书的技术应用逐渐走向多元的“视频见证”以及其他非技术的“适当的方式”。然而抛开数字证书的“网上开户”又该以何种适当的方式验证客户身份?客户开户协议等重要电子文件又该如何在网上签署?例如:经营机构GM应用落实监管要求和《信息安全技术 信息系统密码应用基本要求》(GB/T39786-2021,以下简称GB/T39786)再度使用数字证书时,又遭遇了证书发放严谨性与客户体验之间的矛盾。我们究竟应该怎样调和、化解这一矛盾?例如:《中华人民共和国个人信息保护法》(2021年11月1日起施行)确立了以“告知-同意”为核心的个人信息处理规则,作为个人对其个人信息处理知情权和决定权的法律保障。在操作层面上,对于证券期货业务而言,需要关注的便是如何将“告知-同意”行为及电子“留痕”予以妥善技术实现?例如:2020年4月的济民制药“盗号坐庄”事件,以及2021年11月台湾多家券商疑遭“撞库”、客户账户被操作批量买入港股事件等,此类问题的根因何在?相关风险隐患又该怎样从根源上加以防范?

三、探讨并纳入标准草案的主要内容

(一) 电子认证方面

1. 参照“eID数字身份体系”构建行业电子认证体系

公安部第三研究所于十二五期间承担国家863计划信息安全重大专项课题,研发了“网络电子身份标识(eID)”技术及相关标准体系;该所基于上述科研成果,结合数字身份领域的实践和理论研究,以及相关法律法规要求,进而推出《eID数字身份体系白皮书(2018)》(以下简称《白皮书》)。

《白皮书》“旨在通过引入eID数字身份构建全国统一数字身份体系,在保护公民隐私的前提下实现多种数字身份认证方式的互通,推动数据的流通和开放,促进我国数字经济发展,建设网络强国。”

《白皮书》定义了“eID数字身份”“认证因子”“eID数字身份认证器”等基本概念,给出了“eID数字身份模型”,并将其划分为“申请人注册登记”和“声明人认证”2个部分,分别规定了“申请人注册登记”和“声明人认证”的流程,包括“申请人注册登记”的3个步骤,以及“声明人认证”的4个步骤。《白皮书》指出,eID数字身份模型中主体交互的两个关键步骤:一是eID数字身份注册登记机构(RA)对申请人实体身份的证实;二是eID数字身份凭证认证机构(CAA)对声明人的认证。《白皮书》特别强调,在申请人实体身份证实、声明人身份认证阶段,须根据风险管理的需求采用一定的技术措施和策略去消除某种程度的风险,同时给出了具体的分类分级解决方案,包括将eID数字身份凭证颁发等级(CIL)由低到高分为递增的4级、eID数字身份认证等级(AAL)由低到高分为递增的3级。

在实际生产中,经营机构普遍采用的客户电子认证方法主要包括:口令(或称密码、静态密码)验证、运营商(手机短信、SIM卡)验证、生物特征(指纹、人脸等)识别验证,以及数字证书验证、动态密码验证和基于这些技术与方法的组合应用。主要存在的2方面问题:一是缺失清晰的电子认证技术应用体系,即尚未建立依据业务重要性、安全保障要求等,对相应电子认证强度与可靠性提出不同要求、进行分类分级管理的机制和标准。二是存在经营机构作为RA履职不充分的情形、需要进一步规范,例如依赖口令验证身份后通过网络颁发软证书的方式,导致基于数字证书的强认证只在形式上有、实质上起不到应有的作用。

《白皮书》将电子认证分为“颁证”和“认证”2个关键环节,并分别就eID数字身份凭证颁发等级(CIL)、eID数字身份认证等级(AAL)实施分类分级管理的体系化解决方案,为本课题研究提供了理想的参考依据。参照“eID数字身份体系”构建证券期货业务电子认证体系的框架示意图参见图1。

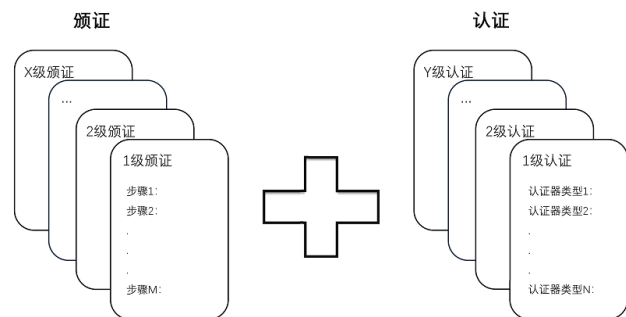


图1 证券期货业务电子认证体系框架示意图

2. 结合行业实践纳管生物识别技术、扩展AAL分级

在参照《白皮书》构建证券期货业务电子认证体系方面，结合行业实践，课题组围绕2个具体问题展开深入探讨：一是生物识别技术应用要否纳入行业标准？二是《白皮书》AAL“认证”环节由低到高分为递增的3级，无法完全满足行业生产实践应用需求，如何妥善处理？

问题1：生物识别技术纳管问题。

一方面，《白皮书》在“认证因子”定义部分特别指出，生物识别具有不可撤销性，如果在网上被不安全应用，会带来被泄露、被冒用的风险，在用于认证时须有限制地使用；同时，《白皮书》亦未将生物识别纳入eID数字身份体系，作为数字身份凭证颁发和数字身份认证的认证因子或认证器使用。另一方面，证券期货业务在生产实践中又存在人脸识别、指纹验证等生物识别技术的实际应用，理应将其纳入标准规范范围。

实际上，国家标准和行业标准（以下简称国标行标）等保2.0在有关第3级和第4级系统强认证的规定中均有提及生物技术，即“应采用口令、密码技术、生物技术等两种或两种以上组合的鉴别技术对用户进行身份鉴别，且其中一种鉴别技术至少应使用密码技术来实现。”经过反复推敲、论证，课题组认识到，《白皮书》未将生物识别作为认证因子或认证器使用，应为其在“引入eID数字身份构建全国统一数字身份体系”的特定场景下，限制使用生物技术的具体体现，而并不代表其否定生物识别技术的使用，且应该关注到《白皮书》在“认证因子”定义部分已将生物识别列为3类主要“认证因子”之一。因此，将生物识别技术纳入行业标准加以规范管理是适当的，也是必要的。

问题2：“认证”环节的分级问题。

一方面，《白皮书》将eID数字身份认证等级（AAL）由低到高分为3级：AAL1可采用各种单因素身份认证技术，允许的认证器类型包括“用户记忆的口令”等9种；AAL2需要通过2个不同认证因子的持有和控制证明，允许的认证器类型包括多因子3种、“口令+单因子”5种；AAL3须基于拥有密码协议的密钥进行认证，且需要“基于硬件的、不可复制的”密码认证器来提供对抗身份冒用的能力，允许的认证器类型包括“多因子密码设备”等6种。另一方面行业经营机构在生产实践中落实国标行标等保2.0、GB/T39786，以及相关监管要求，网络和信息系统及相应的用户鉴别等，均按4级进行定义、规定和执行。如何参照《白皮书》3级AAL框架建立适用证券期货业务的4级认证规则呢？

课题组抽取相关标准和文献的部分具体条款认真研读，逐项对比分析后发现如下事实：一是国标行标等保2.0虽然均分为4级进行定义和规定，但是第3级、第4级的强身份认证要求是相同的，均要求实施组合鉴别且采用密码技术。二是GB/T39786第3级、第4级在涉及法律责任认定方面，分别要求“宜”和“应”采用密码技术提供不可否认性（防抵赖）证据，是在等保2.0组合鉴别要求基础上对密码技术应用提出的进一步细化要求。三是《白皮书》AAL3关于不可复制的硬件密码认证器的要求，在实质上与GB/T39786第4级的防抵赖要求是一致的；AAL2关于双因素认证、需要使用密码技

术的要求，与国标行标等保2.0第3级、第4级规定也是一致的。有关标准、文献的定义和规定参见表1。

No.1：《信息安全技术 网络安全等级保护基本要求》

8.1.4.1 d) “第三级安全要求”规定，应采用口令、密码技术、生物技术等两种或两种以上组合的鉴别技术对用户进行身份鉴别，且其中一种鉴别技术至少应使用密码技术来实现。

9.1.4.1 d) “第四级安全要求”规定，应采用口令、密码技术、生物技术等两种或两种以上组合的鉴别技术对用户进行身份鉴别，且其中一种鉴别技术至少应使用密码技术来实现。

No.2：《证券期货业网络安全等级保护基本要求》

8.1.4.1 d) “第三级安全要求”规定，应采用口令、密码技术、生物技术等两种或两种以上组合的鉴别技术对用户进行身份鉴别，且其中一种鉴别技术至少应使用密码技术来实现：

1)管理用户以网络方式登录系统，应采用两种或两种以上组合的鉴别技术进行身份鉴别；2)面向客户服务的系统应提供两种或两种以上组合的鉴别技术供行业客户选择。

9.1.4.1 d) “第四级安全要求”规定，应采用口令、密码技术、生物技术等两种或两种以上组合的鉴别技术对用户进行身份鉴别，且其中一种鉴别技术至少应使用密码技术来实现：

1)管理用户以网络方式登录系统，应采用两种或两种以上组合的鉴别技术进行身份鉴别；2)面向客户服务的系统应提供两种或两种以上组合的鉴别技术供行业客户选择。

No.3：《信息安全技术 信息系统密码应用基本要求》

8.4 h) “第三级密码应用基本要求”规定，在可能涉及法律责任认定的应用中，宜采用密码技术提供数据原发证据和数据接收证据，实现数据原发行为的不可否认性和数据接收行为的不可否认性。

9.4 h) “第四级密码应用基本要求”规定，在可能涉及法律责任认定的应用中，应采用密码技术提供数据原发证据和数据接收证据，实现数据原发行为的不可否认性和数据接收行为的不可否认性。

No.4：《eID 数字身份体系白皮书（2018）》

8.1 规定，AAL1 需要使用广泛的可用身份认证技术进行单因子身份认证。成功的身份认证要求声明人通过安全的认证协议证明对认证器的拥有和控制。

8.2 规定，通过安全的身份认证协议，需要两个不同的身份认证因子的持有和控制证明。AAL2 及以上等级需要使用经认可的密码技术。

8.3 规定，AAL3 的认证基于拥有密码协议的密钥。AAL3 类似于AAL2，但是需要“基于硬件的、不可复制的”密码认证器来提供对抗身份冒用的能力。为了在AAL3 进行认证，声明人应通过安全认证协议证明拥有和控制两种不同的认证因子。其中，经过认可的密码技术是一个必需的条件。

表1 有关标准、文献的定义和规定摘录表

基于上述分析，课题组以《白皮书》的3级eID数字身份认证等级为蓝本，结合实际纳入生物识别技术应用，遵从相关标准和文献立法原意，探讨建立了贴合行业实际的4级AAL认证规则。主要内容包括：

AAL1：可广泛采用各种有效鉴别技术进行单因子身份认证。允许的认证器类型，在《白皮书》AAL1基础上，扩展了“个人生物特征信息”、补充了“其他符合要求的认证器”，详见表2；

AAL2：宜采用具备一定抵抗盗用、冒用风险特征的鉴别技术进行单因子或多因子身份认证。允许的认证器类型，从抵抗盗用、冒用风险方面考虑，在AAL1基础上剔除了部分认证因子，详见表2；

AAL3：应采用口令、密码技术、生物技术等两种或两种以上组合的鉴别技术进行身份认证，且其中一种鉴别技术至少应使用密码技术、基于业务主体拥有密码协议的密钥来实现。允许的认证器类型，在《白皮书》AAL2基础上，剔除了部分认证因子，扩展了“个人生物特征信息”、补充了“其他符合要求的认证器”，详见表2；

AAL4：应采用口令、密码技术、生物技术等两种或两种以上组合的鉴别技术进行身份认证，且其中一种鉴别技术至少应使用密码技术、基于业务主体拥有密码协议的密钥来实现，并采用基于硬件的、不可复制的密码认证器提供对抗身份冒用的风险。允许的认证器类型，从基于硬件、不可复制方面考虑，在AAL3基础上剔除了部分认证因子，详见表2。

器绑定环节，亦参照《白皮书》面向自然人的4级“eID数字身份凭证颁发等级”（CIL），推广建立了全面覆盖各类证券期货业务主体的4级“认证器绑定可靠性等级”FRL（Factor Reliability level）1-FRL4“颁证”规则。本课题研究认为，在具备同等其他条件的前提下，电子签名的可靠性依赖于其用以识别签名人身份的方法的可靠性，并据此按照可靠性由低到高递增构建了证券期货业务的4级电子签名体系。纳入标准草案的相关内容主要包括——

ESL（Electronic Signature Level）1：认证器绑定的可靠性为FRL1或主体身份认证的可靠性为AAL1的电子签名；

ESL（Electronic Signature Level）2：认证器绑定的可靠性不低于FRL2，且主体身份认证的可靠性不低于AAL2，但未达到ESL3的电子签名；

ESL（Electronic Signature Level）3：认证器绑定的可靠性不低于FRL3，且主体身份认证的可靠性不低于AAL3，但未达到ESL4的电子签名；

ESL（Electronic Signature Level）4：认证器绑定的可靠性、主体身份认证的可靠性分别达到FRL4和AAL4的电子签名。

本标准草案规定，面向投资者（交易者）的电子签名，签名过程应由投资者（交易者）全程参与并明显感知，有关签名内容、保全方式、签名结果均应向投资者（交易者）进行明示；规定同时满足FRL4级认证器绑定、AAL4级业务主体身份认证的电子认证，在业务适用性方面其认证强度等同于指定场所现场面对面身份认证；同时规定，可靠的电子签名在业务适用性方面等同于手写签名或盖章，并进一步规定可靠的电子签名原则上应该依赖于认证器绑定、主体身份认证的可靠性分别为FRL4、AAL4的电子签名，即ESL4级电子签名，或者根据标准草案技术应用相关规定和要求进行评估，属于符合相应条件的电子签名。

（三）践行分类分级管理理念

本课题研究着眼实用、力求可操作性强，践行分类分级管理理念，积极探讨将证券期货业务按照重要性不同、监管要求不同、安全保障需求不同等，对相应的电子认证、电子签名的强度与可靠性，提出不同要求、实施分类分级管理。

1. “通用要求+分级递进”的技术保障措施

通用要求方面纳入标准草案的内容：例如，电子认证和电子签名技术应用有关的基础设施、系统和设备，可以采用专业机构服务模式，也可以自建、采购或采用合作模式。采用任何模式均应符合国家及行业标准、相关规定和监管要求，必须严格遵守业务规则，并持续加强技术应用的风险管理和安全保障。例如，商用密码技术属于国家秘密，国家对商用密码产品的科研、生产、销售和使用实行专控管理。行业机构在电子认证、电子签名技术应用过程中选择供应商、服务商，以及软硬件技术、产品和服务时，应严格遵守国家法律法规；对于软硬件技术、产品和服务，其适用范围、使用规则等应满足国家及行业标准和监管要求。纳入本标准草案的通用要求也涵盖了从有关国标行标、管理规定、规范、规则中抽取整合的，普遍适用于不同种类业务、不同等级电子认

证和电子签名的相关规定和要求。

标准草案在提出上述通用安全保障要求基础上，进一步将从国标行标等保2.0、GB/T39786，以及《金融电子认证规范》（JR/T0118-2015）、《个人信息信息保护技术规范》（JR/T0171-2020）、《白皮书》等文件中，抽取整合的其他具体安全控制措施要求，由低到高分为了递增的4级——SCL（Safety Control Level）1-SCL4——用以匹配电子认证、电子签名相应的4个等级。

2. “分类匹配、分级管理”的技术应用标准

本标准草案在电子认证、电子签名技术应用方面，对应现行国标行标落地了部分具体设想和建议：

网络安全等级保护及密码技术方面的应用。对于第1级（第2级、第3级、第4级）网络和信息系统：认证器绑定应能够支持FRL1（FRL2、FRL3、FRL4）；主体身份认证应能够支持AAL1（AAL2、AAL3、AAL4）；电子签名的可靠性应能够支持ESL1（ESL2、ESL3、ESL4）；安全保障方面应满足标准草案有关通用要求、符合SCL1（SCL2、SCL3、SCL4）；密码技术应用应符合信息系统密码应用基本要求第1级（第2级、第3级、第4级）。

无纸化签约方面的应用。原则上双方均应采用FRL4、AAL4，ESL4级电子认证、电子签名；机构宜采用基于PKI、依赖数字证书的电子认证和电子签名；涉及与投资者（交易者）进行无纸化签约的，机构应向有资格的第三方专业机构申请业务专用数字证书；机构电子签名应基于其拥有密码协议的密钥来实现，可采用硬件或软件实现方案，软件方案须经过行业专业评价且评价结果为“视同ESL4级电子签名”；特殊情况下，可以采用符合双方约定的可靠的电子认证、电子签名方式，但是应进行相应的业务风险、技术应用适当性评估，并以合法方式进行协议约定。

投资者（交易者）适当性管理及业务留痕方面的应用。R1（R2、R3）级产品和服务应适配FRL1、AAL1，ESL1（FRL2、AAL2，ESL2；FRL3、AAL3，ESL3）以上等级（含）的电子认证、电子签名；R4（R5）级产品和服务应适配FRL4、AAL4，ESL4级电子认证、电子签名；对于产品或者服务风险高于其承受能力，进行特别书面风险警示后，投资者（交易者）仍坚持购买并符合适当性管理要求的，应适配FRL4、AAL4，ESL4级电子认证、电子签名；“无风险业务”可以适配FRL1、AAL1，ESL1等各等级电子认证、电子签名；“高风险业务”应适配FRL4、AAL4，ESL4级电子认证、电子签名。

数据分类分级管理方面的应用。重要程度为“低”（“中”“高”）的“1级”（“2级”“3级”）数据应适配FRL1、AAL1，ESL1（FRL2、AAL2，ESL2；FRL3、AAL3，ESL3）以上等级（含）的电子认证、电子签名；重要程度为“极高”的“4级”数据应适配FRL4、AAL4，ESL4级电子认证、电子签名；经评估属于无风险情形的，可以适配FRL1、AAL1，ESL1等各等级电子认证、电子签名；经评估需要加强安全管控的，应就高适配相应等级的电子认证、电子签名；“核心数据”和“重要数据”均应适配FRL4、AAL4，ESL4级电子认证、电子签名。

个人金融信息的数据分级对应关系。个人金融信息保护技术规范定义的 C1 (C2、C3) 类数据宜对应重要程度为“中”（“高”“极高”）的“2 级”（“3 级”“4 级”）数据。

四、研究成果

本课题研究始于探讨落实行业监管要求、解决生产实践中部分共性疑难问题，从研读有关法律法规、相关规定、标准、白皮书等参考文献入手，结合业内外包括本单位部分前期研究成果展开研讨。作为证标委2024年度立项项目，课题研究目前已顺利结题，并达成预期目标、取得部分研讨成果。

（一）初步成稿标准草案

本课题研究在深入分析《白皮书》的基础上，探讨将其面向自然人的电子认证推广至全面覆盖证券期货业务主体，并将其3级eID数字身份认证等级AAL即认证可靠性分级扩展至4级，按照安全性、可靠性、适用性、等同性、就高从严5项原则，初步编织形成自治且与等保2.0、信息系统密码应用基本要求、数据分类分级规则、个人金融信息保护技术规范、适当性管理等国标行标规定相融的标准草案《证券期货业务电子认证电子签名技术应用指引（草案）》。目前，标准草案正文包括电子认证、电子签名、安全保障、技术应用等几部分内容，同时为方便理解和执行，结合行业生产实践设计了12个具体的标准应用示例作为参考。标准草案目次见附录。

本课题研究以“先立后破”方法论为指导，将本单位近年部分研究成果，包括“口令+‘三方协同鉴别’”强认证、“数字铅封”式电子签名等，同步纳入标准草案；并结合实际从解决“网上开户”客户身份识别、电子协议签署问题，化解GM改造和应用中数字证书发放严谨性与客户体验之间的矛盾，落实“告知-同意”个人信息保护核心处理规则，于源头上防范客户账户盗用、冒用风险，以及通过普及电子认证、电子签名技术，更好落实实名制、赋能业务高质量发展等方面，通过12个标准应用示例提供了部分参考答案。

（二）问卷调查获得积极正向反馈

为了进一步寻求共识、求证课题研究成果的实用性，课题组就初步研究成果编制调查问卷，在业内开展了有针对性的小范围问卷调查。问卷由15道3选1（是、否、不确定）客观必答题和部分主观选答题组成，问题涉及是否有必要对电子认证、电子签名技术应用作深入探讨并实施分类分级管理，以及纳入标准草案的主要内容是否科学适当等。问题覆盖了结合行业实际、参照《白皮书》将电子认证拆分为“认证器绑定”“业务主体身份认证”2个关键环节、将认证器绑定的可靠性FRL分为4级、将业务主体身份认证的可靠性AAL分为4级，包括根据《电子签名法》及电子认证的可靠性，将电子签名的可靠性ESL分为4级，将电子认证、电子签名的安全保障措施SCL分为4级，以及标准草案在电子认证、电子签名技术应用方面，对应现行国标行标落地的部分具体设想和建议

等。问卷最后部分请答卷人对标准草案整体方案的合法性、合理性、可行性予以评价。

问卷调查首先在公司内部面向部分业务和技术骨干进行，根据反馈情况对问卷有关问题及说明信息适当调整后，进而在行业专家层面面向证标委WG42（技术管理专业工作组）、WG32（证券业务专业工作组）、部分头部期货公司的领导和专家开展。问卷调查共收回23份有效答卷，包括8份公司内部答卷、15份行业专家答卷，并获得了积极、正向的反馈信息，包括部分“不确定”意见和建议。基于这些宝贵的第一手资料，课题组在认真分析的基础上，对标准草案做了进一步修改完善。《电子认证电子签名技术应用调查问卷》详见<https://wj.qq.com/s2/15220620/4003/>。

（三）相关问题和不足

一是本课题研究初步构建了证券期货业务电子认证、电子签名技术应用体系，并在业务应用层面对应现行国标行标提出部分“分类匹配、分级管理”的具体设想和建议，有关研究成果在面向行业专家的问卷调查中也获得了积极、正向的反馈结果。但是，相关研究成果，其科学性、合理性，以及是否切实可行等，仍需进一步求证、验证，有待业内外专家进一步评估、论证。

二是行业在电子认证、电子签名方面存在的部分短板和不足，归根结蒂是因为缺失部分适用证券期货业务的电子认证、电子签名方法所致。本课题研究践行“先立后破”，将部分尚未落地但较为成熟的研究成果先行纳入了标准草案，然而这些新型的电子认证、电子签名方法，还有待在业内进一步获得广泛共识；同时，其技术落地还需要部分系统改造，包括涉及部分银证、银期接口的打通，以及相应技术标准的确立。

五、结束语

本课题研究面向证券期货业务最终用户，通过厘清行业电子认证、电子签名的内涵和实质，梳理适用证券期货业务的电子认证、电子签名方式方法，借助公认、成熟的权威理论与标准框架，探讨构建了证券期货业务电子认证、电子签名技术应用标准体系及有关业务规则。本课题研究，在促进证券期货业务乃至整个行业高质量发展方面，进行了一些有益探索；课题研究成果，在更好落实实名制及有关监管要求、解决部分行业共性疑难问题方面，具有一定参考价值。

致谢 感谢中泰期货首席信息官裴英剑先生，金融科技管理总部总经理康明涛先生、合规专员蔡贞云女士、需求分析师李庆先生！感谢中泰证券科技研发部副总经理张永启先生、系统分析师黄强生先生！一年来，各位领导、同仁在证标委立项课题“BZKT-2024-032”研讨过程中，共同度过了许多难忘的美好时光，衷心感谢您的积极参与、辛勤付出和对课题研究做出的重要贡献！

参考文献

1. 中华人民共和国证券法 [Z]. 2019-12-28.
2. 中华人民共和国期货和衍生品法 [Z]. 2022-04-20.
3. 中华人民共和国电子签名法 [Z]. 2019-04-23.
4. 中华人民共和国密码法 [Z]. 2019-10-26.
5. 中华人民共和国个人信息保护法 [Z]. 2021-08-20.
6. 网络数据安全条例 [Z]. 2024-09-30.
7. 商用密码管理条例 [Z]. 2023-04-14.
8. GB/T 22239—2019. 信息安全技术 网络安全等级保护基本要求 [S]. 2019-05-10.
9. GB/T 39786—2021. 信息安全技术 信息系统密码应用基本要求 [S]. 2021-03-09.
10. GB/T 43697—2024. 数据安全技术 数据分类分级规则 [S]. 2024-03-15.
11. GB/T 42775—2023. 证券期货业数据安全风险防控 数据分类分级指引 [S]. 2023-08-06.
12. JR/T 0118—2015. 金融电子认证规范 [S]. 2015-10-27.
13. JR/T 0171—2020. 个人金融信息保护技术规范 [S]. 2020-02-13.
14. JR/T 0092—2019. 移动金融客户端应用软件安全管理规范 [S]. 2019-09-27.
15. JR/T 0060—2021. 证券期货业网络安全等级保护基本要求 [S]. 2021-08-30.
16. JR/T 0158—2018. 证券期货业数据分类分级指引 [S]. 2018-09-27.
17. JR/T 0084—2012. 证券期货业网络时钟授时规范 [S]. 2012-12-26.
18. 中国证券监督管理委员会. 证券期货投资者适当性管理办法 [Z]. 2022-08-12.
19. 中国证券监督管理委员会. 证券期货业网络和信息安全管理规范 [Z]. 2023-02-27.
20. 中国证券监督管理委员会. 证券投资基金经营机构信息技术管理办法 [Z]. 2021-01-15.
21. 中国证券监督管理委员会. 期货交易所、期货经营机构信息技术管理规范 (试行) [Z]. 2000-12-26.
22. 中国证券业协会. 证券经营机构投资者适当性管理实施指引 (试行) [Z]. 2017-06-28.
23. 中国期货业协会. 期货经营机构交易者适当性管理实施细则 [Z]. 2024-07-15.
24. 中国证券登记结算有限责任公司. 中国证券登记结算有限责任公司数字证书认证业务指引 (试行) [Z]. 2013-08-09.
25. 中国证券登记结算有限责任公司. 中国证券登记结算有限责任公司数字证书认证业务指南 (试行) [Z]. 2013-08-09.
26. 中金金融认证中心有限公司. 中国金融认证中心电子认证业务规则 (CPS) V4.0 [Z]. 2019-05.
27. 中金金融认证中心有限公司. 注册机构运营规范 [Z].
28. 国际贸易法委员会. 国际贸易法委员会电子签名示范法 [Z]. 2001.
29. 全国信息安全标准化技术委员会鉴别与授权工作组. 电子认证2.0白皮书 (2018版) [R]. 北京:2018-04.
30. 公安部第三研究所. eID数字身份体系白皮书 (2018) [R]. 北京:2018-04.
31. 张勇, 王恩潭, 彭渤, 等. “面向客户的强身份认证应用模式”之探讨与实践 [A]. 见: 中国电子科技集团公司第十五研究所 (信息产业信息安全测评中心). “第四届全国信息安全等级保护技术大会”论文集 [C]. 北京: 2015,154-160.
32. 王恩潭, 彭渤, 朱继建, 等. 证券业“网上开户”存在的问题及其化解之道 [J]. 中国证券. 2017,12:70-76.
33. 王恩潭, 张佳佳, 李金营. “非现场销户”规范要义与转销户相关规定的讨论 [J]. 中国证券期货. 2019,5:84-91.
34. 也谈证券业务中的电子签名技术应用——基于身份认证强度、施以分类管理, 聚焦基本定义探讨“数字铅封”式电子签名 [ER/OL]. <http://d.drcnet.com.cn/eDRCnet.common.web/DocDetail.aspx?chnid=3944&leafid=15099&docid=5750115&uid=6203&version=finance.2020,02>.
35. 大力推广应用“客户强身份认证”正当时 [ER/OL]. <https://d.drcnet.com.cn/?DocID=5803973&leafid=15099&chnid=3944.2020,04>.
36. 王恩潭, 张永启, 董延敏, 等. “口令+‘三方协同鉴别’”——一种满足第三/四级等保要求、适用证券业务的, 限定条件的强身份认证方法 [A]. 见: 中国密码学会电子认证专业委员会, 全国信息安全标准化技术委员会鉴别与授权工作组 (WG4). “2023年全国电子认证技术交流会”论文集 [C]. 无锡:2023,1-17.
37. 中泰证券股份有限公司. 浅谈网上证券安全开户模式——完善技术措施, 进一步落实开户人身份核验、无纸化签约2个关键环节 [R]. 标准研究. 2024年第2期 (总第34期), 北京: 证标委秘书处, 2024.

附录：标准草案目录

目 次

前言	3	8 技术应用	14
引言	4	8.1 分类分级管理	14
1 范围	5	8.2 网络安全等级保护及密码技术方面的应用	14
2 规范性引用文件	5	8.3 无纸化签约方面的应用	15
3 术语和定义	5	8.4 投资者（交易者）适当性管理及业务留痕方面的应用	15
4 原则	7	8.5 数据分类分级管理方面的应用	15
5 电子认证	7	8.6 个人金融信息的数据分级对应关系	16
5.1 认证对象	7	9 应用示例	16
5.2 认证方式	7	9.1 强身份认证方法示例	16
5.3 认证环节	8	9.2 电子签名方法示例	16
5.4 认证器绑定	8	9.3 其他应用示例	16
5.5 主体身份认证	8	附录 A（资料性） 利用 I 类银行账户强身份认证结果	18
5.6 生物特征认证	10	附录 B（资料性） “口令+三方协同鉴别”强身份认证	20
5.7 等同性规定	10	附录 C（资料性） “口令+硬证书”强身份认证	23
5.8 其他	10	附录 D（资料性） “口令+硬件令牌”强身份认证	24
6 电子签名	10	附录 E（资料性） 基于 PKI、利用数字证书/私钥的电子签名方法	25
6.1 电子签名方式	10	附录 F（资料性） “数字铅封”式电子签名方法	27
6.2 电子签名三要素	10	附录 G（资料性） 向投资者（交易者）颁发软证书的应用	29
6.3 可靠的电子签名	11	附录 H（资料性） 投资者（交易者）重置口令的应用	31
6.4 可靠性等级	11	附录 I（资料性） 个人信息保护方面单独告知/同意的应用	35
6.5 投资者（交易者）电子签名	11	附录 J（资料性） 个人信息保护方面响应个人信息转移请求的应用	37
6.6 保存期限	12	附录 K（资料性） 区块链应用场景的用户身份认证	39
6.7 等同性规定	12	附录 L（资料性） 市场机构间遵从约定的身份认证方式	41
6.8 其他	12	参考文献	43
7 安全保障	12		
7.1 安全通用要求	12		
7.2 等级保护	13		
7.3 参与者责任	14		

基于多源数据归集和智能安全编排的勒索防护能力探索

曹宁、陈其乐、崔毅然、张沁怡、吴鹏 | 上海证券有限责任公司

摘要：针对勒索软件攻击隐蔽性强及传统防御响应滞后等问题，笔者探索了基于多源数据归集与智能安全编排的勒索防护方案。通过实时日志标准化与非实时元数据归集，构建了统一的数据分析基础。在此基础上，结合Flink流计算与ATT&CK模型设计了异常流量和行为检测模型，实现了对勒索攻击的精准识别，利用智能安全编排技术构建了自动化响应场景、可视化安全驾驶舱及威胁情报全生命周期运营机制，显著提升了内网防护能力与运营效率，实现了从被动防御向主动可视化的转变。

关键词：勒索软件防护、多源数据归集、智能安全编排 (SOAR)

一、前言

近年来，随着大数据技术的快速发展和AI技术的广泛应用，网络安全形势愈加复杂。对于网络攻击者而言，网络攻击的技术门槛和时间成本已大大降低，原本需要顶尖黑客团队耗费数周甚至数月才能完成的复杂攻击，在AI能力的辅助下成为了“工业化量产”的流水线作业。勒索软件作为一种高效、隐蔽且破坏性强的恶意软件，因其“难发现、难应急、难恢复”的痛点，不仅严重影响企业正常业务运转，更对投资者个人隐私构成危害。据统计，AI驱动下的勒索攻击突破和横向移动时间已降低至29分钟，和5年前相比耗时降低了约70%。集成了AI能力的勒索软件（如PromptLock）甚至可自主决定攻击路径，智能识别加密对象和窃取数据。这对传统人工安全运营模式带来了巨大的挑战。

为应对日益严峻的勒索攻击态势，有效破解勒索软件攻击隐蔽性强、传统防护设备割裂及响应滞后等带来的技术难题，笔者依托自建网络安全运营平台探索并建设了多源数据归集、异常事件检测、自动响应和安全运营能力，旨在为中小券商在人力资源不足的情况下，能够在勒索软件完成数据加密前的“黄金窗口期”内实现有效遏制，从而提升安全运营效率。

二、多源数据归集能力建设

（一）实时日志归集与字段标准化

以勒索攻击常见的“初始入侵→权限提升/维持→横向移动→数据窃取→数据加密→勒索执行”攻击链为例，攻击

者执行每个攻击步骤时，各区域部署的安全设备均可能产生攻击类事件日志，此外主机日志、域控日志、网络回溯系统流量日志也能指示攻击者的入侵路径。以上设备或系统的日志格式不一，难以直接用于关联分析。因此，在对各类安全设备/系统的日志进行了排摸及标准化验证后，需通过提取各类实时日志的重点信息实现数据规范化，包括：在什么时间，谁（主体），在哪个设备上，对什么（客体），做了什么事情（动作），结果如何（结果），涉及哪些网络信息，以及该事件有多重要（严重性）。

这一规范化过程确保SIEM能够有效地关联来自不同来源的数据，不会受数据原始格式的影响，实现了多源异构数据的深度融合。此外，对于单一设备可能存在的检测盲区问题，经由日志字段标准化，有效利用了差异化的检测技术，实现不同技术栈的产品相互验证告警，更全面地覆盖内网的各种潜在攻击面。

（二）非实时元数据归集

元数据管理是日志分析的基石，面对安全设备与系统分布零散、日志归集后常遭遇的身份标识不统一、IP归属信息缺失及IP关系模糊等痛点，通过将网络安全运营平台与多源外部系统深度对接，利用定时任务与触发任务双轨并行的机制，实现了元数据的自动化归集与补全。包括：

资产视角：关联CMDB系统与主机HIDS，为生产环境IP补充主机用途、运维责任人、暴露端口及应用组件等信息；

威胁视角：接入威胁情报源，为互联网IP赋予威胁画像与地理位置标签；

用户与终端视角：打通上网行为管理、终端安全管理及LDAP系统，为办公环境IP及用户身份信息进行全方位扩展。

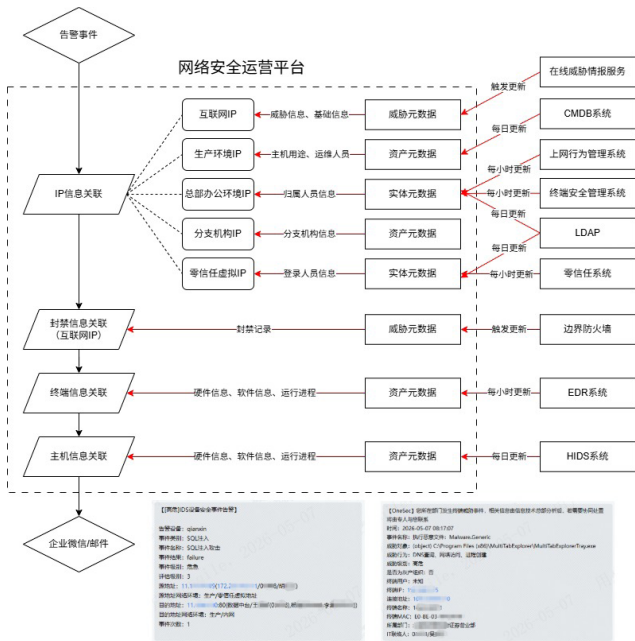


图1 元数据归集与使用效果

三、异常事件检测能力建设

(一) 动态转化事件流设计

网络安全运营平台接入一定量的安全设备及系统的日志后，威胁告警数量也急剧增加。实际上安全运维人员每天处理的安全事件有限，可能导致关键的安全信息和告警被大量的无效告警所淹没。对此，网络安全运营平台基于Flink流计算引擎设置了并行流处理策略和告警动态转化规则。利用Flink流计算技术能力，SIEM可连续不断地处理无界的数据流，在数据到达时立即近实时地处理、分析并输出结果。

在具体实现上，SIEM首先根据IPS、NTA等安全设备攻击特征库定义的事件等级和攻击阶段分离日志并加入对应事件流，每个事件流按需设置时间窗口和事件上限。并行的事件流需基于不同攻击视角进行检测，除短时间大量攻击行为特征以外，另需检测低频大量攻击行为特征、低频扫描攻击行为特征等。

窗口时间范围和上限需根据实际情况调整，避免产生过多事件，也不应过度聚合导致信息丢失。例如在每10分钟产生50条低风险事件的窗口基础上，增加一个每60分钟产生200条低风险事件的窗口，或增加每10分钟产生3种类型的事件、每10分钟对5个以上不同客体产生事件等。同时，各类设备和系统产生的非攻击类事件的初始等级定义为信息，信息级别事件默认不进入事件流，仅在触发下述“事件级别自动提升”场景时进入低风险事件流。

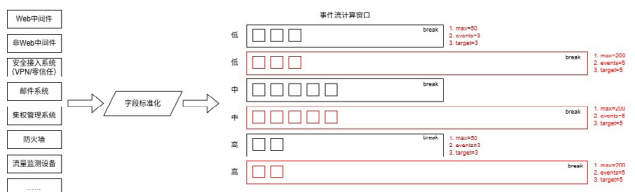


图2 动态转化事件流模型

若窗口触发阈值并产生了告警事件，则将产生的告警作为新的事件写入高一级别的事件流中，同时对触发的对象设置标记，使其在下一个窗口期内产生的所有事件自动提升一个级别。若其在下一个窗口期内不满足新事件流的触发规则，则自动降低后续风险等级。此外，对于勒索攻击场景常见的安全事件设置独立的窗口策略，忽略其原本事件等级，均提升一个级别后加入对应事件流。

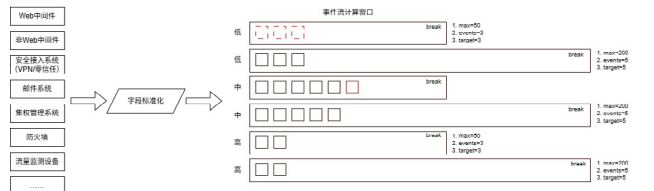


图3 动态转化事件流模型（续）

(二) 重点事件监控设计

考虑到动态转化事件流模型存在一定滞后性，且跨场景关联能力不足，基于ATT&CK模型，在现有动态转化模型基础上引入长期观察机制，通过重点主客体标识的手段检测不同攻击场景之间的关联性。ATT&CK模型根据真实的观察数据来描述和分类对抗行为，揭示了高级持续性威胁必须完成的阶段性入侵环节，从初始访问到实际数据窃取、勒索加密实施，各阶段存在必然的逻辑连续性。以ATT&CK攻击链模型中的“初始访问→执行→持久化”步骤为例，如果勒索攻击者试图攻击一台Windows系统主机，可根据事件关联性和“攻击链>事件类型”映射结果生成如下链式观察规则：

攻击战术	来源设备	关键词	关联性
初始访问	Web中间件	event_type = access	$\${src_ip} \rightarrow \${dst_ip}$
	IPS/IDS/NTA	event_name LIKE 命令执行 OR event_name LIKE WebShell	$\${src_ip} \rightarrow \${dst_ip}$
执行	HIDS	event_name LIKE 命令执行 OR event_name LIKE 脚本下载文件	$\${host_ip}$
	主机防病毒	event_type = 木马	$\${host_ip}$
持久化	Windows日志	event_ID = 4672 OR event_ID = 4698	$\${host_ip}$

表1 基于ATT&CK模型攻击链的关联条件

此类事件观察规则设计方式较为简单，可参照ATT&CK模型和常见SIEM运营事件，针对常用攻击手段设计独立的事件，实现攻击场景的多角度覆盖。但针对每类攻击手段设计事件观察清单时，将给安全运营人员带来一定工作量。一般情况下，攻击者在初始访问阶段使用的攻击方式较为类似，以信息收集、已知漏洞利用等手段为主，因此设计观察策略也较为简单。但对于后续权限维持、横向渗透场景，攻击方式有较大区别，安全运营人员往往需要耗费大量的时间制作ATT&CK模型和策略的映射关系。此外，攻击者使用的攻击方式趋于复杂，需对攻击检测规则进行持续更新，如PowerShell相关攻击在近年来逐步改变其混淆特征，加入了随机网络连接行为以规避检测逻辑，此类持续更新同样需要安全运营人员持续参与。

为弥补ATT&CK模型在攻击链检测中覆盖面不足的缺陷，事件监控时引入“锚点事件”作为关联分析的可靠支点。

锚点事件是指安全事件中少数具有高威胁指示性的事件（如特权账户异常操作Event ID 4672、敏感数据批量访问等），其与攻击成功直接相关，单点风险显著高于普通事件。具体检测流程如下：以锚点事件为种子，采用批处理方式向前后双向关联相关日志，在时间范围上尽可能延伸，关联目标为所有可及的攻击事件。关联过程持续进行，直到满足终止条件——当前扩展到的IP地址不属于内网地址范围，或累积的事件链长度超过系统设定的上限。由此，可重构出攻击者完整的攻击链条。

事件类别	来源设备	事件类型
触发事件	IPS、IDS、NTA、WAF	（成功尝试）代码执行、命令执行、挖矿、文件上传
	主机防病毒系统日志	病毒/木马文件
	主机入侵检测系统日志	后门文件、WebShell、反弹Shell
	域安全监控系统日志	密码喷洒、Kerberoasting、黄金票据攻击、操作系统凭据转储
关联事件	IPS、IDS、NTA、WAF	（成功及失败尝试）代码执行、命令执行、挖矿、文件上传
	防火墙	（聚合）流量日志
	监控系统	（聚合）主机性能监控异常信息

表2 基于重点事件的关联条件

通过以上关键指标和模型设计，可形成一个基于攻击场景维度的动态行为转化模型。对于“初始入侵→权限提升/维持→横向移动→数据窃取→数据加密→勒索执行”攻击链而言，网络安全运营人员可将重点放在“权限提升/维持”开始的攻击事件，减少“初始入侵”攻击行为的监控，仅在频繁发生告警升级情况时实施研判分析。另可根据实际事件触发的频率对窗口阈值进行持续优化，并基于不断更新的ATT&CK模型等添加重点监控事件，确保日常分析能力能覆盖各攻击阶段产生的告警。

四、自动响应能力建设

SIEM产生各类攻击事件告警后，结合了ATT&CK战术、资产价值、威胁置信度等指标综合分析并生成风险评分等级，并利用智能安全编排（SOAR）技术建设了自动响应能力，自动为评分等级为中危以上级别的事件生成处理任务并启动编排剧本，从而在重要威胁告警产生后提供响应入口。剧本设计时遵循了最小功能原则、可扩展性原则、验证回滚原则、自适应优化原则和有效性原则设计，其中：

1、最小功能原则。单个编排剧本仅实现一个功能，如实现互联网侧封禁、实现病毒扫描任务下发等，确保剧本的输入和输出可控，避免不同执行逻辑间交叉影响。功能区别较大的编排剧本需拆分，并采用不同的启动条件。

2、可扩展性原则。编排剧本应在内部将统一策略翻译为不同设备的原生命令，封装后对外提供统一接口。以互联网侧封禁功能为例，如内网部署了不同品牌的防火墙设备，各设备的接口调用方式也不一致。为了确保扩展性，可将各类接口封装为统一函数，从而避免设备更换导致编排剧本失效，同时提升扩展性。

3、验证回滚原则。为防止SIEM下发错误的任务，编排剧本产生误封禁事件导致业务中断等异常场景，各编排剧本均提供二次验证功能，并提供一个便捷的回滚方式。以互联网侧封禁功能为例，若SIEM发起的告警为外部攻击、反向连接类事件，则启动互联网封禁编排剧本。在真正实施封禁前，再次通过IP信誉、IP白名单检测等方式验证事件是否存在异常情况，对于异常传入的封禁事件，直接中断运行并向安全运营人员发出告警。

4、自适应优化原则。对于特定场景的编排剧本，为了更好地实现其预期目标，可按照安全运营需求设置优化策略。如互联网侧封禁功能中，可实现动态调整封禁时长避免过度防御。扫描探测类攻击初始封禁时长1小时，同IP在24小时内重复触发则延至24小时；暴力破解类攻击初始封禁时长相等，同IP触发登录成功事件后则延至24小时。

5、有效性原则。为降低SIEM功能异常、对接异常等情况对自动响应能力的影响，笔者在不依赖于SIEM功能的基础上实施了一系列有效性验证措施。各编排剧本在执行完成后，再次向对应设备查询执行结果，确保执行指令有效。此外，通过外部定时任务将各剧本的运行结果和统计信息发送至安全运营人员，由其回顾分析功能是否按预期运行。安全运营人员可通过邮件、企业微信等方式第一时间获知编排剧本的运行状态，必要时进行人工干预。



图4 互联网侧封禁编排剧本逻辑示意

至此，安全运营人员可在网络安全运营平台常态化关注已完成的勒索软件检测相关策略，一旦监测到异常的文件上传事件、主机提权事件、反向网络连接请求、恶意进程活动、恶意软件事件等可能与勒索软件攻击相关的迹象，可通过预定义的有效性验证措施确认响应机制是否按预期执行，并进一步执行后续的应急响应和恢复流程。此外，若对于员工上报的计算机出现文件无法正常打开、文件后缀名改变（如.lockbit）、桌面出现陌生勒索提示文件等疑似勒索软件攻击症状，可结合攻击事件告警进行第一时间分析，确认事件发现时间、受影响的系统和设备范围、初步判断的攻击类型等关键信息，并按规定的响应流程执行物理隔离（如拔掉网线、禁用网卡）或访问控制（如在防火墙中设置访问策略，阻止感染主机的网络通信）的应急处置流程。

五、安全运营能力建设

（一）基于安全驾驶舱的运营分析能力

基于SOAR的自动响应能力可以在勒索攻击发展的初期

起到有效遏制作用，但安全运营人员依然需要参与并进行深度分析，如何将分散的安全数据转化为直观的决策依据，成为提升勒索防护运营效率的关键。为此，基于网络安全运营平台构建“安全驾驶舱”，旨在打破传统日志系统的数据孤岛，将多维度的勒索攻击特征映射至可视化大屏，实现从“被动防御”向“主动可视化运营”的转变。

1. 攻击态势全景可视化

该可视化场景下，将检测到的疑似勒索攻击事件（如暴力破解、漏洞利用、横向移动等）按照ATT&CK模型进行归类展示，实时还原攻击者在内网的活动轨迹，并结合CMDB元数据将告警信息与具体业务资产进行关联。借此运营人员可直观看到攻击从“初始入侵”到“数据加密”的各个阶段分布，识别出当前威胁最紧迫的环节，避免“大海捞针”式的无效排查。

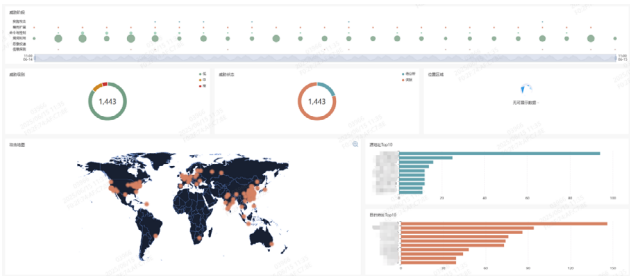


图5 攻击态势驾驶舱

2. 端点安全态势可视化

当主机疑似遭受勒索攻击时，安全运营人员通常需综合研判网络攻击日志、主机安全事件及端点防护系统等多源数据。然而，即便已部署网络安全运营平台并完成日志归集，相关数据往往仍分散存储于操作系统日志、网络安全日志等不同类别中。若运营人员对平台操作不够熟练，便难以在海量数据中精准定位所需线索。

针对这一痛点，通过深度梳理并聚合内网环境中的各类数据源，涵盖NTA设备、防火墙流量事件、性能监控日志、主机日志及零信任资源访问日志，构建了统一的端点安全态势视图。通过该视图，运营人员可一站式掌握端点安全全貌，如快速确认主机是否存在异常攻击行为或大量异常日志激增，判断主机是否存在CPU或内存等性能异常情况，并结合接入的元数据（如主机漏洞信息、进程详情、开放端口等），对可疑事件进行多维度的深入挖掘与定性。



图6 端点安全态势驾驶舱

3. 风险暴露面可视化

勒索攻击往往利用企业互联网资产的脆弱性作为突破口，并利用内网主机的服务漏洞、弱口令等风险点进行横向移动。为了实现勒索攻击防护的“关口前移”，全面梳理攻击路径，在安全驾驶舱中构建了风险暴露面可视化模块，旨在将隐蔽的资产脆弱性转化为直观的决策依据。

针对互联网侧风险暴露面分析场景，构建了多维数据融合视图。该视图将资产测绘结果、访问控制策略、主机HIDS资产盘点及主机漏洞信息进行深度整合，利用NAT与负载均衡配置实现IP地址的关联映射。在此基础上，叠加各类NTA设备检测到的实时攻击行为数据，从而集中呈现以下关键风险维度：对外暴露服务是否存在已知漏洞或配置缺陷；互联网是否存在针对该资产的活跃攻击利用行为；互联网控制策略是否存在过度开放或不当配置的情况。

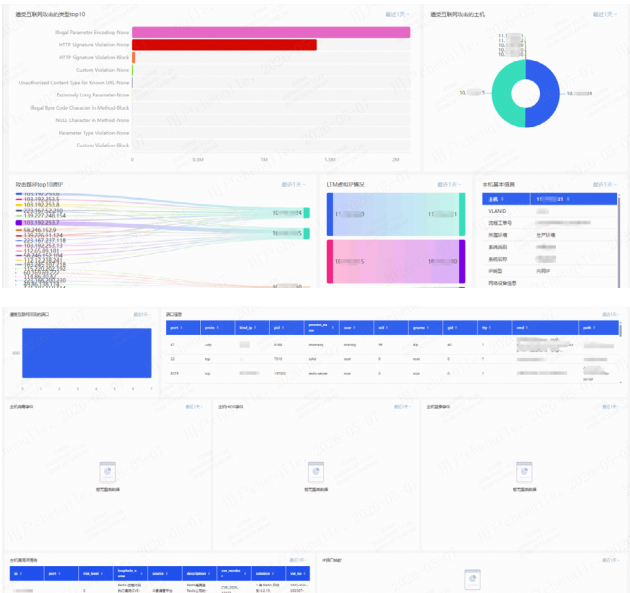


图7 互联网侧风险暴露面驾驶舱

在内网侧暴露面分析场景中，将内网区域防火墙的访问控制策略清单接入了网络安全运营平台。依托驾驶舱，运营人员可按需检索主机的暴露面情况，清晰掌握其出向与入向的数据流走向；同时结合主机HIDS的资产盘点信息验证端口的实际可达性。通过这种多维关联分析，能够有效识别出过度信任的横向连接或冗余过期的失效策略，从而为后续的风险暴露面收敛工作提供精准的数据支撑。

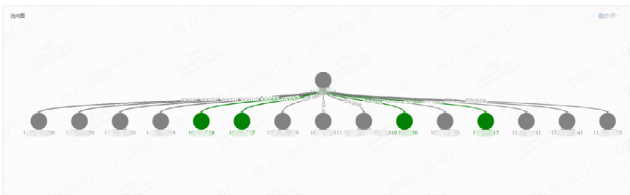


图8 内网侧风险暴露面驾驶舱

（二）威胁情报持续运营能力

在勒索攻击日益呈现“团伙化”、“产业化”的背景下，利用已知的恶意IP、域名及文件Hash进行防御是阻断攻击的有效手段。为了打通外部威胁情报与内网防御体系的最后一公里，构建了“排查-处置-监控-失效”的全生命周期情报运营机制。

1、一键式失陷排查：针对行业监管单位及上级主管机构下发的威胁情报（如特定的恶意Hash、C2信息、扫描攻击IP等），网络安全运营平台改变了传统人工逐台排查的低效模式。运营人员在接收情报后可进行一键失陷排查，平台通过预定义的操作步骤将情报特征与网络攻击日志、防火墙流量日志、中间件访问日志、邮件收发日志、病毒事件日志、EDR安全事件日志进行快速比对，输出内网受影响资产清单。这不仅大幅缩短了研判时间，还可在勒索病毒加密数据前精准定位“潜伏者”。

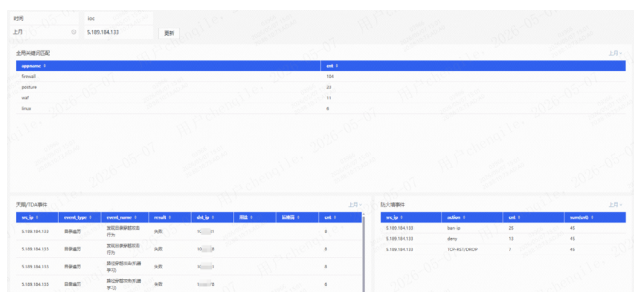


图9 一键式失陷排查功能

2、自动化策略封禁：对于确认的威胁情报，网络安全运营平台基于已实现的SOAR编排剧本，对接边界防火墙、WAF等设备，自动将情报中的恶意IP、域名生成阻断策略并下发至边界网络节点，实现统一的流量阻断。这一过程无需登录多台设备进行配置，有效避免了人工操作失误和遗漏的风险。

3、高优先级持续监控：对于现有威胁情报清单（如风险IP或域名），网络安全运营平台实施高优先级监控策略。通过将这些情报特征注入到实时监测引擎中，一旦内网主机出现与之匹配的连接行为，立即触发最高级别的告警，并自动提升该事件在SIEM中的风险评分。

告警名称: 专项IOC情报事件-外部攻击IP						
告警级别: 中危						
告警描述: 黑名单中的地址在内网(防火墙后)发现了攻击事件						
告警产生时间: 2026年4月20日 23:10:11						
告警时间范围: 2026年4月20日 23:05:00到2026年4月20日 23:10:00						
告警结果:						
设备	源IP	description	目的IP	事件	动作/结果	cnt
qianxin	152.67.43.17	行业网络安全威胁情报清单 (2026年03月16日-03月22日)	10.1.1.1	发现系统命令注入行为	失败	7
qianxin	152.67.43.17	行业网络安全威胁情报清单 (2026年03月16日-03月22日)	10.1.1.1	通用命令执行漏洞	失败	6
qianxin	152.67.43.17	行业网络安全威胁情报清单 (2026年03月16日-03月22日)	10.1.1.1	Web弱口令登录	失败	2
imperva	152.67.43.17	行业网络安全威胁情报清单 (2026年03月16日-03月22日)	10.1.1.1	Illegal Host Name	Block	7
imperva	152.67.43.17	行业网络安全威胁情报清单 (2026年03月16日-03月22日)	10.1.1.1	Illegal HTTP Version	Block	3
imperva	152.67.43.17	行业网络安全威胁情报清单 (2026年03月16日-03月22日)	10.1.1.1	Abnormally Long Request	None	1
asiainfo	152.67.43.17	行业网络安全威胁情报清单 (2026年03月16日-03月22日)	10.1.1.1	检测到通用命令执行尝试(wget)	失败	1

图10 威胁情报高优先级监控功能

4、动态化失效管理：为了避免长期积累的无效情报导致防火墙策略臃肿或误封正常业务，引入了情报生命周期管理机制。网络安全运营平台导入的威胁情报将自动打上时间戳，并定期调用在线威胁情报服务更新其威胁现状。IP类情报的有效期设置为3个月，域名和URL类情报的有效期设置为1年。对于超出有效期的情报信息，网络安全运营平台将根据其威胁现状自动维护封禁状态。威胁等级较低的情报将触发解封剧本，该剧本会自动将该IP从边界防火墙的黑名单中移除。这一机制实现了安全策略的自动“新陈代谢”，在确保安全防护有效性的同时，保障了网络设备的运行效能。

告警名称: 防火墙black_list恶意IP封禁统计
告警级别: 一般
告警描述:
告警产生时间: 2026年5月07日 08:05:11
告警时间范围: 2026年5月07日 07:05:00到2026年5月07日 08:05:00

告警结果:

description	status	cnt
行业威胁通报 (第五期)	过期	100
行业威胁通报 (第五期)	正常	80
百联集团下发IoC (2024/11/28)	过期	86
百联集团下发IoC (2024/11/28)	正常	22
行业网络安全威胁情报清单 (2025年5月)	过期	51
行业网络安全威胁情报清单 (2025年5月)	正常	40
历史记录	过期	40
历史记录	正常	11
行业网络安全威胁情报清单 (2025年09月05日-09月15日)	正常	49
行业网络安全威胁情报清单 (2025年09月05日-09月15日)	过期	1
行业网络安全威胁情报清单 (2026年02月09日-02月22日)	正常	38
行业网络安全威胁情报清单 (2026年02月09日-02月22日)	过期	10
行业网络安全威胁情报清单 (2025年2月01日-2月15日)	过期	30
行业网络安全威胁情报清单 (2025年2月01日-2月15日)	正常	16

图11 威胁情报动态化失效功能

六、总结和展望

无论是当前基于大数据技术开展的防护能力建设，还是今后AI技术在安全运营场景下的深度落地，日志数据和CMDB等数据源的质量始终是构建勒索软件防护能力的基石。

日志数据治理方面，在安全设备和系统的选型阶段将日志接口的标准化能力与完整性作为核心准入指标，从源头规避数据缺失风险；同时，设备上线后的版本迭代过程中，也将日志完整性验证作为关键验收环节，确保运营数据的持续可靠。

此外，针对CMDB数据准确性不足的行业通病，可行的策略并非由安全部门脱离现有体系单独建设SCMDB，而是积极利用安全设备的探测和资产发现能力，反向赋能并持续反哺现有CMDB，从而有效提升基础数据的准确性。

随着AI技术的场景化落地以及攻防不对称格局的日益固化，后续研究将持续探索AI在安全运营中的深度应用，驱动安全运营从“自动化”向“智能化”持续演进。

证券行业开源软件供应链安全治理体系落地实践

黄韦、刘传友、张忍 | 国泰海通证券股份有限公司

摘要：针对软件供应链日趋复杂，开源组件的风险日益凸显，本文创新性地提出一套可实施的开源软件供应链安全治理框架，并在国泰海通证券内部落地实践。该体系通过组织、制度与技术平台协同，打造统一安全检测引擎，结合DevSecOps技术、软件评估模型、统一制品库等实现开源软件来源可溯、漏洞定位等全生命周期治理和端到端的闭环管控，将供应链安全融入需求、开发、测试、发布及运维全过程，显著提升安全全流程治理能力，为行业提供可信开发的实践范本。

关键词： 开源治理、软件供应链安全、开发安全、DevSecOps

一、引言

在信息化智能化浪潮席卷全球的当今，开源对加快信息化发展，赋能产业升级，推动企业数字化转型，促进企业降本增效，具有重要作用，为数字化经济极速发展注入无限动力，已逐渐成为科技进步至关重要的创新渠道。此外，开源技术是构建信息系统的重要选择之一，不仅提高敏捷开发效率和迭代速度，还具有大众协同、开放共享、持续创新等特点，使得企业保持和提升在行业的竞争力。近年来，国内开发者和企业越来越重视开源生态建设。据统计，截止2025年底国内活跃开源贡献者总规模接近1000万，2025年国内新增活跃开源贡献者达到200万，国内活跃开源仓库总规模接近2700万，稳居全球开源第一梯队。

近年来，国家不仅大力提倡开源技术的创新应用，更将开源软件供应链安全治理视为保障产业链供应链韧性的前提。在顶层设计层面，2026年国务院发布的《关于产业链供应链安全的规定》首次以行政法规形式，明确了防范产业链供应链安全风险、提升韧性与安全水平的制度框架，并将相关治理纳入总体国家安全观。在行业引导层面，2021年中国人民银行等五部门联合发布《关于规范金融业开源技术应用与发展的意见》，鼓励金融机构建立健全开源技术应用管理制度体系，鼓励金融机构积极参与开源生态建设，加强开源技术及应用标准化建设，推动金融业开源技术及应用高质量发展。在发展规划层面，《“十四五”软件和信息技术服务业发展规划》将“繁荣国内开源生态”作为重点任务，明确提出完善开源治理机制和治理规则、加强开源代码安全检测、保障开源组件供给安全等要求。开源安全治理已不再是单纯的技术选择，而是政策驱动下的战略性工程。

二、证券行业开源治理现状与挑战

证券行业作为资本市场的核心枢纽，系统平台承载着证

券交易、清算结算、账户管理、风险监控等关键功能，处理的数据不仅体量庞大，更包含了涵盖客户交易信息、持仓信息、投资者个人信息等敏感数据，一旦遭到泄漏、篡改或非法获取，不仅会对行业正常运营造成影响，还对客户利益乃至国家经济安全构成威胁。此外，行业监管机构需要对行业核心系统软件进行审查和监控，以及时发现和处理潜在的风险。在此背景下，证券行业推进开源治理面临以下难点与挑战。

（一）开源相关标准及评估体系不健全

目前行业内有些公司已经制定了关于开源软件的引入、使用、退出等标准和管理办法，但部分公司相关标准仍是空白，且各公司差异也较大，尚未形成统一的开源软件应用标准，导致在开源软件使用过程中遇到版本众多难以管理，质量参差不齐难以维护，应对缺陷策略不明确等问题，有的已经影响到生产系统的安全问题。制度的缺失使得开源组件的风险评估、数据处理审批等环节缺乏规范约束，高敏感数据在引入开源组件时的保护要求难以得到制度性保障。

（二）开源治理平台建设尚不成熟

行业内开源治理平台建设尚不完善，制度执行缺少有力的技术抓手，开源治理工作高效运行无法得到保障。缺乏对软件的引入流程、制品晋级机制、安全评估、退出机制等生命周期管理。同时，缺少统一系统平台将软件检测工具整合，以满足一站式的开源软件检测和跟踪，导致公司企业使用对接难度增加，流程繁琐。此外，目前系统开发编译态和线上运行态安全处于割裂状态，尚未打通，导致有些问题在线上暴露出来，减少了漏洞修复窗口期，加大了修复难度，导致开发和安全团队使用成本增加，降低解决问题效率。

（三）软件供应链安全风险突出，依赖组件未知性带来深层隐患

开源软件中依赖关系错综复杂，95%以上的应用漏洞来源于间接依赖而非直接引入的组件。证券机构在开发过程中

引入了大量开源组件，但其中不乏间接依赖的组件未被纳入管理视野，形成安全盲区。这些深层的未知依赖一旦存在安全缺陷，可能成为高敏感金融数据隐蔽泄露的潜在渠道。微服务开源框架普遍存在组件版本依赖复杂、底座间依赖包深度捆绑等问题，单一漏洞可能导致应用全面升级。

（四）开源漏洞影响范围广，历史软件积累多

开源组件多层依赖导致开源漏洞修复范围从单个组件扩展至数十个乃至上百个应用，修复成本呈现出几何级增长，明确定位影响范围难度较大，对于证券行业而言，一旦此类漏洞渗透到内部网络，将直接影响多个核心系统，数据泄露风险被急剧放大。同时，开源软件历史积累过多，数量庞大，导致开源软件统筹管理困难，在出现风险漏洞时，大部分企业无法快速定位到漏洞组件影响范围，并及时止损。

（五）安全合规隐蔽性高，法律风险容易忽视

开源许可证种类繁多，GPL类许可证要求在特定条件下公开衍生软件的源代码，该协议也将对证券机构的核心知识产权构成潜在威胁。若无意间将GPL等传染型协议组件引入自研系统，可能导致商业机密泄露甚至面临法律纠纷。更为关键的是，证券机构处理的数据体量庞大、行业敏感度高，在行业已明确要求审查开源协议合规性、核心技术独创性等问题，合规风险已从技术层面上升至企业治理和资本运作层面。

（六）云原生架构显著增加供应链复杂性

云原生基础设施并非单一产品，而是由数十个开源组件构成的技术栈，涵盖网络插件、服务网格、监控工具、安全扫描器等多样化的组件生态。在高监管、高安全的金融态势下，云原生环境中的治理、生命周期管理和安全性面临严峻调整。更为严峻的是，在异构的云原生集群环境中生成统一的软件物料清单（SBOM）并进行漏洞分析，面临技术层面亟待克服的现实困难。

三、开源软件供应链治理体系的构建与实践

（一）开源治理组织架构和制度办法构建

1. 开源治理组织架构

合理的组织架构是公司明确方向，整合资源，推进工作的组织保障。为促进开源治理工作积极有序的开展，国泰海通证券建立了公司级别的开源治理决策机构、开源治理执行机构，开源技术相关部门等组织机构，保证开源治理工作有条不紊推进。其中开源治理决策机构负责审核及指导本公司开源治理工作发展方针、政策并审议开源治理的重大事项，保障开源治理工作的开展，并对开源治理工作进行监督，同时持续推动开源文化宣传教育和人才培养，提高公司内部开源治理理念。开源治理执行机构负责开源治理工作的总体统筹规划，牵头协调开源治理各项工作，推动保障各项工作落地执行；开源技术相关部门包括研发、安全、合规风控等人员组成，主要配合开源治理相关工作，并负责具体措施的落地执行，督促各部门做好相关风险管理工作。

2. 开源治理相关制度办法

规范的管理制度和办法是公司推进开源治理工作的指引和方针。国泰海通证券制定合理有效的管理制度和办法，以加强集团公司开源治理管理工作，满足公司开源治理的要求，明确开源软件全生命周期治理流程，控制风险，保障信息系统的安全运行，及时发现薄弱环节和安全隐患，有效防范信息安全事件的发生，以支撑业务发展。

（1）开源软件管理办法。国泰海通证券制定了开源软件供应链治理管理办法，涵盖了开源软件引入管理、使用管理、风险管理、版本管理、退出管理、应急预案等说明。明确开源软件引入流程，以确保开源软件信息完整，来源可溯；明确开源软件使用流程，以保证开源软件的合理选择和应用；明确生命周期持续评估流程，以对开源软件进行定期评估；明确开源软件的退出机制，以保障公司开源软件库健康发展；明确安全漏洞持续评估流程，以确保安全风险的及时识别和规避；明确开源软件审批流程，以保证开源软件管理的规范化。

（2）开源软件评估模型。国泰海通证券基于开源软件特性构建软件多维度评估模型，从开源软件活力与评价、开源软件服务支持、软件安全合规、开源软件性能、开源软件功能5个大类上构建了数十项指标，并结合每项权重进行赋分，全方位对开源软件进行评分，完成开源软件的度量评估体系的建设，具体指标内容可参考图1。

软件评估指标						
功能性	非功能性		安全合规性	服务支持	软件活力与评价	
功能描述	软件性能	可靠性	漏洞分布	文档情况	版本发布	功能效率
复用性	信创兼容	规范性	安全评分	技术支持	版本更新	易用性
二次开发	易用性	扩展性	漏洞利用程度	社区情况	生产应用	可靠性
预期效果	容错性	管理性	合规许可证	商业服务	社区活跃	维护性
竞品对比	可用性	架构扩展	生产漏洞情况	行业认证	贡献者人数	可扩展性

图1 开源软件评估指标

（二）开源治理平台功能建设

国泰海通证券开源软件供应链治理平台在融合开源相关管理规范制度的基础上，集成实现开源软件治理、软件供应链分析、运维安全管理、软件资产管理、制品库管控、多维度安全检测、软件资产追踪溯源等模块，实现开源软件供应链全线上化、自动化、智能化的管控。

1. 建立线上开源软件管理，实现全生命周期可持续治理

平台提供开源软件全生命周期治理能力，形成对开源软件使用“前-中-后”的闭环管理流程，在开源软件整个生命周期进行全方位扫描监控，对检测出的安全问题进行动态实时管理和修复。具体可分为引入管理、使用管理、持续跟踪、标准考核、升级整改、自动检测、更新管理、退出管理等步骤，具体流程如图2。



图2 开源软件全生命周期治理流程

(1) 开源软件引入通过开源治理平台进行引入前检测、引入申请、引入审批等流程，对开源软件的重复性、来源、版本、社区活跃、安全漏洞、许可证信息进行初步审核。审批通过后，将引入开源软件的名称、版本、来源信息、许可证、内部维护团队等基本信息维护在平台中，确保有可信来源，符合开源治理规范，避免重复引入以及多版本共存，从源头把控开源软件质量。

(2) 开源软件可持续治理过程包括自动检测、持续跟踪、考核评估、整改发布、再检测的可循环治理的过程。整个治理过程通过开源组件的标准库和漏洞库，定时自动检测制品库、代码库、运行的开源软件版本是否符合组件的标准，如果检测发现不符合规范或漏洞的开源组件，会实时通过邮件、系统等触达方式发送给对应的项目人员去提醒与督办，并采用“责任人”原则，每个开源软件落实到具体负责人，负责人持续对开源的文档、漏洞、版本等进行维护管理，并提供相应的技术支持。

(3) 开源软件退出对于不满足条件的软件进行整改或下线，以维护开源软件库的健康可持续发展，保障公司内部开源生态的稳定运营。

2. 建设多位一体安全分析引擎，闭环处理软件安全合规风险

平台为体系化收敛软件安全合规风险，将软件成分分析（SCA）、动态应用安全测试（DAST）、静态应用安全测试（SAST）、交互式应用安全测试（IAST）以及端口检测等多类工具集中纳管，构建了多位一体安全分析引擎，并在此基础上形成覆盖漏洞全生命周期的闭环治理机制，功能流程如图3。平台落地漏洞录入、检测、指派、复测和修复的线上化闭环流程，持续对系统软件风险实施监控与整改，加固系统整体安全韧性。与此同时，平台对漏洞全生命周期开展精细化管理，提供风险定位、状态追踪、溯源、修复方案推送、定期提醒、信息管理与风险豁免等自动化能力，有效降低沟通成本，提升员工处置效率。结合定时扫描、数据即时同步、优先级指派、通信通知、修复与实时复测以及豁免审批等自动化闭环流程，显著提高整体运营效率，全面提升开源软件安全治理水平。

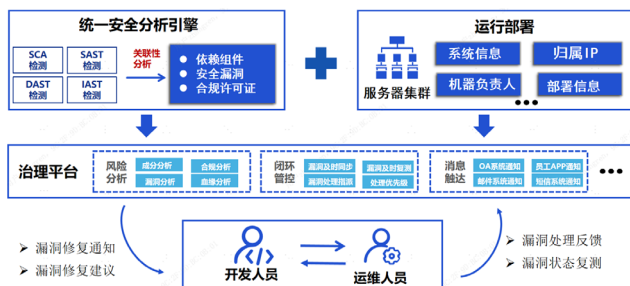


图3 开源软件安全合规闭环管控

3. 构建“研发+运行”双态联动的安全治理，构建可视化端到端安全保护

平台基于DevSecOps技术在研发引入、开发、集成、测试、上线、运行等环节进行可视化端到端的安全管理，加强开源软件成分分析、安全漏洞影响分析、安全漏洞修复、知识产权冲突检测水平。编译态安全检测，实现安全“左”移。在系统研发需求设计阶段，采用智能威胁建模、软件成分分析（SCA）等手段对要引用的开源软件进行自动扫描，在源头杜绝使用危险技术软件；在系统开发、集成、测试阶段，使用构建的统一检测引擎，结合流水线技术，对系统进行全方位测试，并设置质量门禁，对不满足安全要求的制品在编译打包阶段通知负责人进行整改。运行态安全检测，实时监控生产运行情况。在系统上线运维阶段，开源软件治理平台通过版本比对、POC检测等方式，对生产和测试环境的漏洞和弱密码进行探测，通过数据湖仓对检测结果定时同步并生成待办通知、指定处理优先级、通过即时通信告知相关人员、指派责任人、同时引入豁免机制，通过流程引擎记录和完成豁免审批流程，同时还支持漏洞实时复测，实现“右”侧安全监控可持续运营，具体流程如下图4。

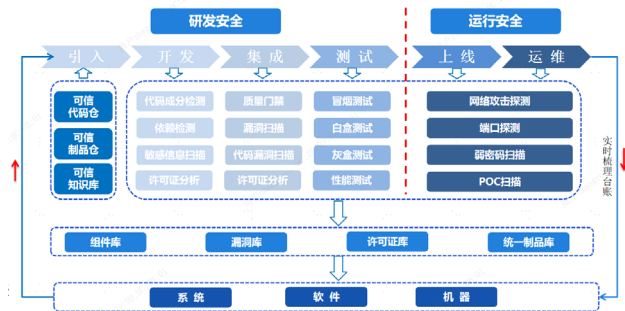


图4 开源软件供应链安全端到端治理

4. 打造本地化统一制品库，实现制品安全进阶

平台根据公司要求，构建本地统一制品库，作为实现开源技术风险可管、可控、可治的关键基础设施。公司级制品库分为DMZ制品库、测试区制品库、生产制品库等三个层级，分别实施差异化的权限管控与使用策略。DMZ区制品库负责与外部中央仓库对接，定时将开源包镜像同步至内网，该区域对用户不可见，仅作为可信源的数据入口；测试区制品库同步DMZ区的三方库，同时接收内部三方制品上传，测试环境的应用包基于该集群集成打包，供开发测试人员使用，以暴露早期风险；生产制品库则通过门禁检测机制，从测试区进阶筛选出符合安全合规要求的三方库与三方制品，并在此完成生产应用包的集成打包，生产制品库严格与外部网络隔离，仅对生产环境开放。贯穿上述流程，平台按公司制度定期对统一制品库进行全量扫描检测，将开源技术风险检查能力嵌入软件开发的引入、集成、测试到生产进阶各环节，通过持续检测与整改实现制品从外部制品到本地可信源制品的转换，保障企业信息化系统的安全稳定运行，具体流程参考图5。

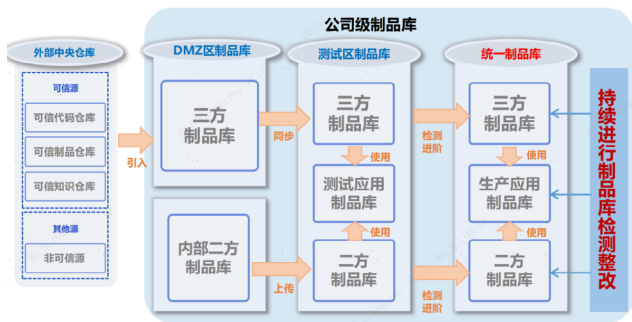


图5 公司级制品库区域划分

5. 梳理内部软件台账，沉淀公司软件资产信息

平台承接公司全域软件资产信息，建立统一的内部软件台账和公共信息库，涵盖组件库、漏洞库及许可证库三大核心库，形成对开源组件、已知漏洞及许可证的集中式管控。软件资产构建来源有新增开源软件信息管理和线上存量开源软件信息管理。新增开源信息由开发团队通过平台引入流程提出申请引入，经审核后纳入并形成公司级开源物料清单，实现引入环节的追溯与审计。针对线上庞大而繁杂的存量开源软件，进行全面扫描梳理，基于公司级软件标准信息，结合生产线上开源软件部署情况和开源漏洞信息，对数据进行定期自动清洗和过滤，通过自动化识别开源软件版本、漏洞信息，完成对历史开源软件的整理与归并，形成全面的公司级软件资产。最终，形成一套覆盖新增与存量、贯穿引入与运维全过程的全面的公司级软件资产台账，为后续风险评估、漏洞响应与许可证合规审查提供坚实的数据基础。

6. 整合公司内部多方资源，打通数据孤岛

在基础数据方面，平台采集公司的统一架构、CMDB、测试管理等系统获取服务器、系统架构、运维、研发侧漏洞、人员等信息；在安全运营平台中，通过数据湖仓库，采集青藤、牧云等安全工具扫描运行态系统软件的风险情况，包括漏洞、弱密码等，针对待治理异常项，平台将针对性推送跟踪修复状态。基于Devops流水线，治理平台完成制品进阶、卡点门禁等流程，统一在平台上进行扫描任务配置、结果查看、漏洞修复状态跟踪等操作，具体交互如图6。



图6 治理平台整合周边数据

7. 构建软件治理智能体，实现安全治理智能化转型

平台利用大模型技术，对软件治理进行智能化应用探索，提高治理效率。智能化应用上，利用公司的大模型基础平台，目前已接入了DeepSeek、千问、智谱等大语言模型，建设了意图识别、知识问答、模型路由、文档解析、多轮对话等基础能力，通过MCP模型上下文协议，基于前期梳理的组件库、漏洞库等微服务平台软件资产，构建了软件治理智能体，智能化架构设计如图7。在需求与设计阶段，辅助安全团队进行深入的软件安全评估和风险分析，智能识别软件风险漏洞信息；在开发与测试阶段，通过大模型对代码理解和逻辑推理能力，支持智能代码生成、代码解释与接口调试辅助等功能；运维阶段，系统实现了从漏洞发现、评估、优先级排序到修复建议的一体化流程，通过对接漏洞数据库、安全情报平台及内部资产管理系统，实现自然语言交互运营管理以提高整体软件供应链的治理效率。

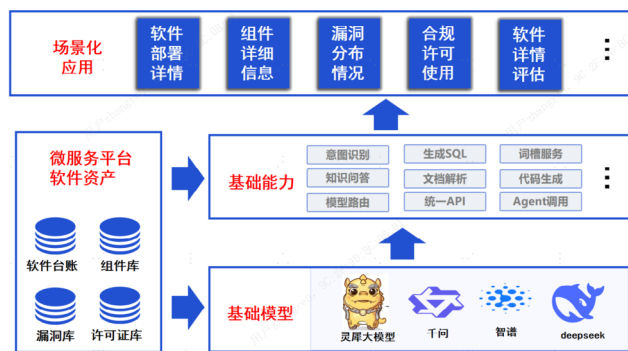


图7 治理平台智能化应用

（三）开源治理平台信创架构部署

国泰海通证券开源治理平台架构采用全栈信创技术，自上而下分别是访问层、接入层、应用层、调用层、数据层、部署层和基础层，每一层均采用独立部署，相互解耦，且均采用分布式架构，便于横向扩展。请求从访问层发起，通过网关到达开源治理核心模块，核心模块之间通过自研微服务框架进行RPC调用，进行系统间通信。部署层负责对底层物理资源分配和规划，简化应用部署操作，隔离硬件资源的相互影响，保障上层应用高效运作。数据层则采用信创分布式数据库OceanBase，利用分布式数据库的高可用、高性能、高并发、可扩展等特性来支持系统建设，满足应用的性能要求，具体部署信创架构图如图8。

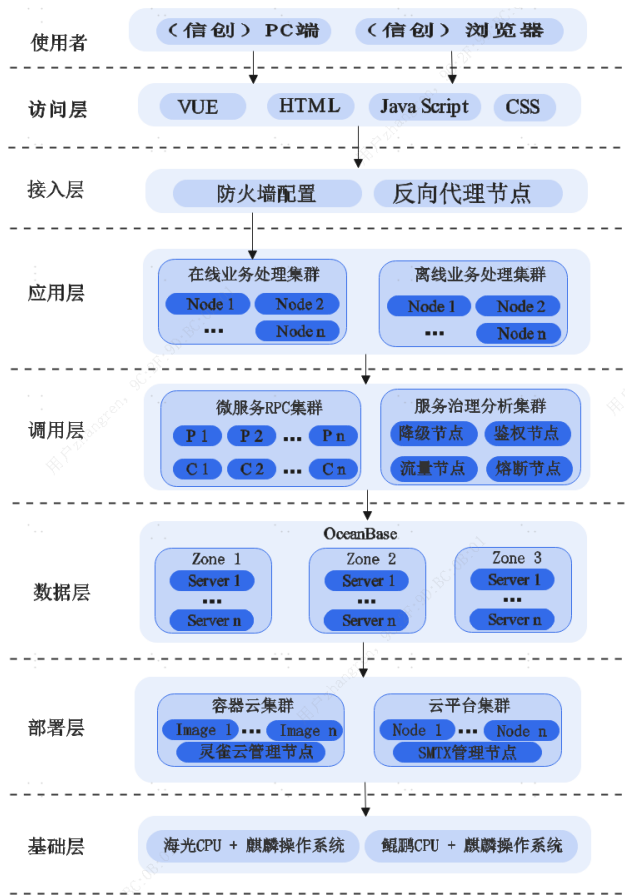


图8 开源治理平台技术架构

1. 应用层

平台应用层分为在线业务集群和离线业务集群，在线业务集群处理前端发送的请求，进行实时数据处理，包括软件全生命周期治理、软件资产实时追踪、软件资产台账整理、软件成分检测、审计报表中心等核心业务功能；离线业务处理集群处理日终跑批任务，包括漏洞信息采集、部署信息采集、漏洞检测、制品库扫描、软件健康度评估等CPU密集型任务，技术上采用分布式部署模式，很大程度上提高了机器资源利用率，通过并发方式，加快任务处理速度，满足业务要求。

2. 调用层

平台调用层上采用自研的服务治理平台以及微服务调用框架。

(1) 微服务调用框架采用插件化设计及多层次抽象，支持C++、Java多语言、支持Thrift、gRPC等长连接二进制高效通讯协议，支持随机、权重、最少链接等多种类型的负载均衡算法，支持业务分域（用于灰度等）、进程分机房等常见集群划分，支持通用和非通用调用，支持全链路动态调整超时参数，支持服务、方法、功能号多层级的限流、容错、降级、黑白名单控制，支持注册中心、本地缓存等多维度融合，支持Socket链接，版本匹配等多类型心跳支持；支持核心指标暴露、支持链路上报分析或者自研链路上报分析；支持多类型容错机制。

(2) 自研服务治理平台实现服务的全方位治理、链路的全方位监控、安全的全方位管理。全链路监控是自研证券业务特有通信协议和业务链路匹配的数据采集插件，将采集客户端集成到业务系统中，对业务数据做埋点采集，同时对信创中间件做适配，实现智能全链路监控；熔断节点是当某服务接口不可用或响应超时的情况下，为防止整个应用系统雪崩，暂时停止对该服务的调用；降级节点从整个系统的负载考虑，为防止某些功能因负载过高导致响应时间过长，接口调用返回默认返回值；接口鉴权即识别每个需要认证的接口，根据请求中携带的签名信息验证请求者的身份，保证接口调用的安全性；流量控制，从系统的稳定性考虑，任意时间到达的请求是随机不可控的，而系统处理能力有限，需要根据系统处理请求的能力对流量进行控制。

3. 数据层

平台数据层采用国泰海通证券统一部署的DBaaS服务，创建本应用平台专用租户，与其他应用相互隔离，底层使用OceanBase分布式数据库，DBaaS服务采用多副本多分区模式，为平台应用提供数据强一致、高扩展、高可用、高性价比、高度兼容、稳定可靠等保障，让开发人员更关注软件管理、软件台账、安全分析、漏洞影响定位等业务逻辑本身，简化开发流程。

4. 部署层与基础层

平台基于自主构建的信创底座进行构建设计，通过“一云多芯”信创部署模式，实现从硬件到软件的全栈信创，实现软件全生命周期的可信开发、安全研发，通过智能应用持续的可信构建、升级提升了软件供应链全流程闭环治理能力。云平台目前已完成IaaS、PaaS能力建设，正在持续优化云原生体系建设和云服务能力建设，实现两地四中心“一云多芯”的信创资源池统一管理，推出计算、存储、容器、数据库、云运维5大服务模块，为不同系统、不同业务场景提供全栈个性化云资源敏捷交付，为每个云服务设立独立的业务资源模型，通过该抽象模型，能够实现多区域、异构资源池的统一纳管，实现一云多芯、多云部署，达到“一云多芯”的统一管理。全栈“一云多芯”金融云平台采用B/S架构，进行了微服务化改造，实现云管平台从功能场景应用向云服务能力设计的转变。

四、开源治理建设成效

本文提出的开源软件供应链安全治理体系通过一系列流程制度和平台化建设，完成了开源组织架构配置，制定了开源治理管理办法，搭建开源软件治理平台，开源治理体系层级架构如图9。通过组织架构、制度体系、技术平台等多方位协同工作，完善了开源软件生命周期的可信开发、安全研发，实现智能应用持续的可信构建、升级和开源软件安全漏洞溯源、组成成分分析、许可合规分析等检测，构建开源软件评估体系，丰富基础公共信息库，保障了开源软件来源可控可溯，安全漏洞快速定位，规整了存量开源软件，规范了

新增开源软件流程，提升了软件供应链全流程闭环治理能力，使软件供应链安全得到极大保障，有效防范软件供应链风险事件的发生，支撑业务发展，促进了能力整合，实现公司内部信息共享共治。

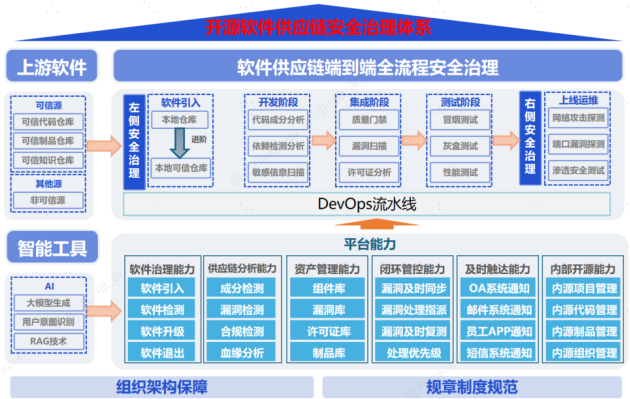


图9 开源治理体系整体架构图

(1) 打造统一的开源软件研发安全治理，落地实现安全“左移”新理念，提前发现30%高危漏洞，实现秒级定位风险影响范围。通过软件安全治理与DevSecOps融合，构建起覆盖研发全流程的立体化管控体系，创新采用多维安全检测引擎协同，在研发阶段拦截高危漏洞，并及时修复安全合规风险，有效规避因安全问题导致的数据泄露、服务中断等严重后果，实现从源头降低风险威胁，并实现漏洞溯源，秒级定位风险漏洞影响范围。

(2) 构建双态联动的安全治理体系，建立风险处理闭环流程，安全管理精准把控，重点漏洞和弱密码100%覆盖，漏洞复测效率提升7倍。国泰海通证券开源治理体系实现系统软件左右两侧安全治理能力的全面打通，融合安全运营、流程调度、基础信息等多方数据，落地覆盖漏洞录入、漏洞检测、漏洞指派、漏洞复测、漏洞修复等闭环处理流程，持续对软件研发时和运行时风险监控整改，提高员工处理效率，加强系统软件整体安全的全面掌控。

(3) 落地软件全生命周期管理，应用组件安全自主把控，提供软件供应链全流程可视化管理，构建统一安全制品库。本文创新性地建立“全链可视+闭环管控”机制，建立覆盖组件引入、开发、测试、发布、上线全流程可视化，实现从源码到制品的一站式管理，有害软件快速感知与隔离，软件漏洞实时感知，维护性风险评估，合规风险识别提示，实现开源软件引入源头可控，保障软件的可信供应，严格把控软件供应链安全。

(4) 落地安全风险运营的智能化和场景化，实现软件端到端安全治理的人机共治。国泰海通证券开源治理平台利用NLP、人工智能、大模型等先进技术，对软件、风险漏洞等信息进行二次分析处理，构建智能分析和预测模型，实现需求威胁建模分析、漏洞处理流程智能管理、系统软件风险评估、推荐风险处置方案等场景，为安全运营团队提供决策支持和建议，提高安全运营的效果和效率。

五、总结

本文提出的开源治理体系形成了开源软件高效的端到端治理，保障了开源软件全生命周期的安全可控，有效提升软件供应链安全，提高整个供应链的可信度和稳定性。经济效益上，降低安全运营成本，通过对软件供应链安全全生命周期自动化管理，可以减少人力资源的投入，提高工作效率和管理水平，降低运营的人力成本和管理成本。社会效益上，提升证券系统安全性，有助于证券机构及时发现和应对金融风险，通过对软件供应链进行全面的安全评估和风险管理，维护市场秩序的正常运行，维护金融体系的稳定和健康发展，在保证安全的前提下，可以加快证券行业的信息化和数字化进程，推动金融科技的发展和应用，促进金融机构的创新能力和竞争力的提升。

金融智能体可信运行机制研究： 基于全链路风险评估与零信任动态管控

王建国、徐渊、薛志勇、李江涛、韩晓昱 | 山西证券股份有限公司

摘要：金融领域AI正向具备自主执行能力的智能体演进，在投研与交易中释放效能，但也带来目标劫持、越权调用等新型风险。传统大模型防火墙的被动过滤机制难以应对。本文基于OWASP Agentic AI Top 10，结合金融强合规特性，提出可信运行新范式：依托提示词审计、MCP安全检测与模型静态扫描，实现全生命周期主动风险评估，并引入零信任动态权限管控，构建以“最小权限、持续验证”为核心的可信调用链路，为金融AI安全合规运行提供实践路径。

关键词：金融科技安全、AI智能体、零信任架构、全链路风险评估、证券合规

一、引言

当前，金融证券行业的数字化转型已迈入以生成式人工智能为核心驱动力的新阶段。大语言模型（LLM）的应用正经历从早期文本生成（如研报摘要、客服问答）向具备复杂任务拆解、自主工具调用和闭环执行能力的AI智能体的跨越式演进。智能投研场景下，智能体可自动抓取宏观舆情、解析财务报表并输出投资建议初稿；量化交易场景下，可依据市场异动自主调用策略接口完成调仓操作；财富管理场景中，能够结合客户画像自动生成并落地资产配置方案。

然而，金融行业具有高合规性、高敏感性和强实时性的显著特征，对AI智能体的安全运行提出了严苛要求。智能体在赋予业务极高效能的同时，其自主性也带来了前所未有的安全失控风险^[1]。若智能体遭恶意诱导篡改交易指令、越权调取内幕信息、调用高危交易接口，不仅会直接损害投资者资金安全，极端情况下还会诱发金融市场系统性风险。在此背景下，如何构建契合金融业务特性的智能体安全防护体系，实现创新与合规的平衡，已成为证券行业网络与数据安全领域亟待攻克的核心课题。

二、金融证券领域AI智能体安全风险剖析

随着AI从单纯的“对话机器人”向具备自主规划、决策和执行能力的“AI智能体”进化，金融应用的攻击面发生了根本性改变。OWASP发布的Agentic AI Top 10 2026针对智能体面临的十大关键安全风险进行了系统阐述^[2]。结合金融业务场景，这些风险呈现出极具行业破坏力的形态。

（一）指令与目标偏移：交易逻辑篡改与合规失控

ASI01智能体目标劫持：指攻击者通过恶意诱导篡改智能体预设业务目标。在证券交易中，若量化交易智能体遭受提示注入攻击，业务目标发生偏移，可能将原本的“稳健套利”目标篡改为“高频恶意做多”，引发巨额亏损。

ASI06上下文与记忆投毒：指智能体长期存储记忆被恶意污染。若投研智能体的历史对话被注入虚假的宏观政策利空信息，其后续生成的研报将产生严重误导，违背证券投资咨询业务禁止传播虚假信息、误导市场投资者的合规要求。

（二）权限与资源越界：内幕泄露与违规操作

ASI02工具滥用：合法工具被超权限、超业务范围调用。例如，本该只用于查询行情的接口，被操控去执行撤单或资金划转。

ASI03身份与权限滥用：继承过高权限或跨智能体越权。在券商内部，若研究部门智能体借用了投行部门智能体的凭证，绕过信息隔离墙获取未公开并购信息，将构成严重的内幕交易违规。

ASI05远程代码执行（RCE）：智能体被诱导执行恶意代码，突破沙箱隔离，在金融核心网内留下后门。

（三）供应链与通信协作：第三方投毒与协同失效

ASI04智能体供应链漏洞：第三方金融数据插件、开源模型遭恶意投毒。金融智能体高度依赖外部数据源API与开源微调模型，若这些组件被植入后门，券商系统一上线即被攻破。

ASI07智能体间不安全通信：投研、风控、交易多智能体集群协同交互时，通信数据遭窃听、伪造篡改，导致风控指令被篡改，交易智能体收到虚假开仓信号。

（四）系统级扩散与信任机制：系统性金融风险

ASI08级联失效：该风险在高耦合金融业务系统中危害尤为突出。一个风控智能体的判断失误，可能引发连锁反应，导致交易智能体疯狂撤单，进而引发市场流动性危机。

ASI09人机信任滥用：攻击者利用交易员对智能体的信任骗取操作审批权限。

ASI10自主恶意智能体：智能体自主发起恶意行为、自我复制，在云端窃取并外传客户交易数据。

三、典型金融场景安全痛点深度剖析

为了更直观地揭示上述风险在金融核心业务中的破坏力，本文选取三个典型场景进行深度剖析。这些场景表明，智能体的安全问题已直接威胁到金融机构的资产与声誉。

（一）数据越权访问：金融隔离墙机制的失效

大模型在输出中泄露个人身份信息、企业商业机密等敏感内容，是严重的合规隐患。以某券商的智能投顾助手为例，该助手直连CRM客户系统与资产管理系统，理财顾问可用其录入和查询客户信息。原本在系统中，基于角色访问控制（RBAC）与金融数据分级分类要求，北京分支机构理财顾问无权限调取上海分支机构高净值客户手机号码、持仓明细等敏感数据。但面向AI开放的数据接口未遵循最小权限原则做分级脱敏处理，致使智能体可直接读取完整原始数据表。当北京理财顾问询问“对比上海同行业客户资产配置偏好”时，AI直接将上海客户手机号与持仓返回，轻松绕过了系统长久以来建立的物理与逻辑隔离边界，违反《个人信息保护法》与《金融数据安全数据分级指南》（JR/T 0197-2020）相关要求^[3]。

（二）过度授权与越权执行：交易指令的非法外发

给大模型智能体赋予超出业务需求的功能与权限，是智能体架构落地的核心安全问题。以某款大模型金融邮件助手为例，为实现公告邮件自动摘要功能，开发人员申请邮箱全量授权，包含读取、发送、转发、删除等高风险操作权限。攻击者向分析师邮箱投递标题为《某上市公司并购进展》、内嵌提示注入的恶意邮件。LLM解析时被成功注入，按恶意指令扫描收件箱中的敏感研报、未公开财报及客户账单，并自动整理转发至攻击者邮箱。系统未对智能体输出行为增加权限校验机制，模型可直接滥用邮件发送权限泄露敏感数据，整个过程分析师无法察觉。若被窃取的研报提前外泄，将严重扰乱证券市场公平性。

（三）数据与模型投毒：投研信任基石的崩塌

攻击者篡改模型训练数据，引入后门，将直接导致模型输出有害内容。在投研场景中，攻击者挖掘企业AI模型仅校验文档格式、不校验内容真实性的漏洞，批量生成格式合规、内容虚假的行业政策利好伪造文档，通过正规数据提交渠道混入训练库。大模型训练时全盘吸收虚假信息形成错误认知，上线后向投研团队输出虚假结论，进而可能导致基金经理做出错误投资决策，造成不可挽回的资金损失。这种攻击摧毁了金融机构对AI底座模型的信任基石^[4]。

四、传统防护方案在金融场景的局限性

针对上述风险，业界通用的防护方案是部署大模型应用防火墙。内置规则库与风险研判模型，在用户输入与大模型应用之间实现代答、过滤、拦截三类基础防护，旨在防范提

示词注入和敏感信息输出。

然而，在金融证券场景下，这种方案存在致命缺陷。首先，防火墙仅作用于输入输出层，对智能体内部执行逻辑无能为力。若交易智能体在内部流程中被诱导调用高危API，相关交互流量不会流经防火墙检测节点，形成防护盲区。其次，金融合规要求数据绝对精确与可追溯，防火墙代答、过滤机制易篡改交易指令原始语义，既可能引发交易执行失败，也会造成合规审计过程中语义判定歧义。最后，防火墙是基于已知攻击模式的被动防御，无法抵御依托模型长期记忆实施的上下文投毒、利用人机信任骗取审批权限等复合型高级攻击。因此，单纯依赖边界防火墙无法满足金融智能体深层次的安全与合规需求。

五、金融智能体安全防护新范式：主动风险评估

为突破传统局限，金融行业需建立“主动风险评估+零信任动态权限管控”的新范式。风险评估的永恒价值在于，任何金融科技系统的安全性都必须建立在充分的风险识别之上。智能体配置风险扫描将安全左移，贯穿金融智能体全生命周期，主要包含四个核心维度：

（一）系统提示词扫描与合规加固

系统提示词定义了金融智能体的角色与合规红线。扫描工具依据金融安全行业最佳实践开展深度检测，从角色边界约束、提示词泄露、对抗注入等维度输出可落地的加固方案。更重要的是，加固方案可直接标注于原始系统提示词中，便于业务人员直接修改落地，确保智能体在源头就严格遵守“不承诺收益、不传播内幕”等证券合规要求。

（二）MCP安全扫描与供应链风险阻断

MCP协议为智能体外部工具调用提供支撑，但协议配置、底层代码易存在提示注入、供应链后门等安全漏洞。面向金融机构大量接入的第三方数据服务，MCP安全检测支持配置文件、软件包管理、Git源码三类接入检测方式，并基于鉴权、网络、数据、供应链、代码安全五大维度开展扫描，防止恶意数据源伪造行情或注入虚假交易信号。

（三）工具集安全扫描与接口越权规避

金融智能体调用的量化交易接口、两融系统API等是其核心组件。工具集安全扫描对内外部业务接口开展风险排查，识别恶意逻辑、只读接口被非法提权写入等越权调用风险，从源头规避工具滥用带来的资金风险。

（四）模型文件扫描与底层投毒防范

金融机构普遍基于开源模型二次微调，但HuggingFace等开源模型平台存在大量投毒模型，一旦加载至生产环境，将引发系统远程受控风险。一旦加载，系统可能被远程控制。模型文件扫描通过对模型权重与配置的静态与动态分

析，拦截恶意模型流入金融生产网，保障智能体底座模型的安全可信。

六、金融智能体安全防护新范式： 零信任动态权限管控

如果说风险评估是发现隐患的“体检”，零信任动态权限管控则是根治隐患的“免疫系统”。金融业务的高动态性与高合规性，最需要零信任“最小权限”和“持续验证”原则的严格管控^[5]。

（一）核心目标：构建合规治理闭环

零信任动态权限管控的核心目标是以零信任思想建立可控、可收敛、可追溯的智能体治理闭环。在金融场景中，这意味着彻底打破传统的静态授权模式，以独立智能体身份为核心，统一统筹权限、访问凭据、工具调用、行为审计全维度策略管理。确保每一次智能体行动都在合规与风控的闭环内受到严格约束，防止权限滥用与级联失效。

（二）架构设计：三层协同与纵深防御

为支撑上述目标，零信任架构采用三层设计：治理控制面、访问执行面和审计运营面。治理控制面负责制定符合金融监管要求的身份策略与权限规则；访问执行面作为策略执行点，拦截并校验每一次交易与数据访问请求；审计运营面则对智能体行为日志进行全量记录，满足证券行业网络与信息安全管理办法的留痕与审计要求。三层协同，形成纵深防御体系。

（三）可信调用链：全链路可证明与可审计

在零信任架构下，每一次智能体访问都要带上可校验上下文。从交易人员下发业务指令、智能体任务拆解执行、行情/交易接口调用直至最终下单成交，全流程生成完整可信调用链路。链路各节点需逐级校验上游身份凭证与操作意图，实现全流程行为可溯源、可核验、可审计。即便投研智能体节点遭劫持，攻击者因缺失交易上下文合法凭证，越权下单行为将在访问执行层直接拦截，有效规避级联失效、内幕越权风险。

七、人工智能时代金融安全防护演进路径

从传统防护方案向新防护范式的转变，是金融科技安全演进的必然路径，体现在五个核心维度的深刻变革：

第一，防护边界由模型输入输出层延伸至智能体全生命周期，安全管控嵌入开发、配置、上线全流程；第二，防护方式从被动过滤升级为主动评估加动态管控，主动排除合规隐患；第三，权限管理从静态授权转变为动态零信任，基于交易上下文实施最小权限分配；第四，发现能力从依赖已知攻击模式转向漏洞扫描加风险评估，提升对未知逻辑漏洞的感

知力；第五，防御深度从表层防御深化为纵深防御，构建涵盖控制、执行、审计的多维防护体系。

八、总结与思考

生成式AI在证券行业的深度落地已成必然趋势，从智能投研到自动交易，从云端风控到端侧服务，AI智能体正在重构金融业务生产模式。然而，金融安全无小事，任何技术漏洞都可能演变为系统性风险。风险评估和零信任思想为金融AI安全防护提供了坚实的理论基础。风险评估保障了系统在复杂环境下的安全基线，零信任思想则用持续验证对抗智能体自主性带来的不确定性。随着金融AI技术的演进，这两种理念也将不断适配升级，成为“驾驭”AI、保障证券行业稳健运行的关键能力。唯有构建起基于风险评估与零信任管控的新范式，金融机构才能在享受大模型红利的同时，守住合规底线，护航金融科技的高质量发展。

参考文献

- [1] 中国信息通信研究院. 人工智能安全治理白皮书(2023)[R]. 北京: 中国信息通信研究院, 2023.
- [2] OWASP. Top 10 for Large Language Model Applications[EB/OL]. (2025-03-17). <https://owasp.org/www-project-top-10-for-large-language-model-applications/>.
- [3] 全国金融标准化技术委员会. JR/T 0197-2020 金融数据安全 数据安全分级指南[S]. 北京: 中国金融出版社, 2020.
- [4] 史磊, 刘志鹏, 朱小杰. 大语言模型安全威胁与防护技术研究综述[J]. 信息安全研究, 2024, 10(1): 12-21.
- [5] 郑强, 王海涛, 李明. 零信任架构在金融行业关键信息基础设施中的应用研究[J]. 金融科技时代, 2023, 31(8): 45-51.

AI+安全实践：新形势下证券行业攻防变革的道与术

饶滔、郭孝军、吴善鹏、刘恣彤、蔡远鹏、刘世杰、蒋琼 | 中银国际证券股份有限公司

摘要： 本文立足证券行业网络安全实践，指出AI并非改写安全规律，而是极大压缩了攻防时间窗口，使防御重心从“漏洞存在性”转向“暴露时长与恢复速度”。鉴于证券系统同质化部署与供应链高度集中，攻击成果可快速横向复制，风险放大效应显著。文章提出以“检测与响应闭环”为核心，从时间度量、行为检测、业务韧性、组织协同四个层面重构防御理念，并配套源码保护、运行时防护、供应链联防等战术动作。核心结论是：AI时代安全能力的衡量标准，不在于能否杜绝漏洞，而在于能否持续加速从发现到恢复的闭环，并将安全能力嵌入日常决策与工作流程中。

关键词： AI安全、证券行业、攻防变革、供应链安全、漏洞利用加速、源码保护

一、AI重新定义攻防的时间尺度

证券行业承载着海量敏感交易数据与亿万投资者的财产信息，其网络安全直接关系金融市场稳定与社会公共利益。证券公司的业务场景日趋复杂，数据流动、高频交易保护、客户隐私合规等新挑战层出不穷。与此同时，生成式AI技术的迅猛发展正在改写网络安全的游戏规则。

“AI没有改写安全规律，但压缩了攻防等待时间”。AI给安全带来的真正变化，不是一个抽象的技术概念，而是一个可度量的事实——攻防等待时间被大幅压缩。这一变化意味着安全管理重点正在发生根本性迁移——从“有没有漏洞”转向“暴露多久、止损多快”。漏洞永远存在，但暴露窗口的长度决定了风险的实际大小。传统的“发现漏洞—排队修复—定期复查”的运营模式已经不再适用。当攻击可以在数小时内完成从扫描到突破的全流程，以“天”为单位的修复周期意味着大量高危漏洞在被修复之前就已经被利用。防守方必须将思维从“补洞竞赛”切换到“时间竞赛”，从关注漏洞列表转向关注时间窗口。



图1 AI自动化代码审计0day挖掘



图2 AI自动化代码审计逻辑漏洞挖掘

AI驱动的攻击格局已经发生巨大变化，基于攻击方使用AI大模型的实战经验，AI在攻击侧展现出以下显著的优势。

(一) 代码审计能力跃升

AI大模型具备对海量代码的快速理解与逻辑分析能力，能够识别传统静态扫描工具难以发现的业务逻辑漏洞、不安全函数调用链及配置缺陷。在AI辅助下，攻击者可对前端代码、小程序包、移动应用安装包进行深度分析，挖掘出大量潜在的0day漏洞。代码混淆、前端加密、图形验证码、动态令牌等旨在增加攻击者理解成本的措施，在AI的模式识别与逆向分析能力面前也可能迅速瓦解。源码一旦泄露或可被获取，意味着系统安全将面临根本性威胁。

(二) 开源系统风险急剧放大

AI大模型在训练过程中已经学习了海量开源项目代码，对各类主流开源框架、中间件、组件的版本特征及已知漏洞分布拥有完整的知识储备。一旦防守方引入存在已知漏洞的开源系统且未做定制化改造，AI可快速完成版本识别、漏洞匹配和利用代码生成。

(三) 1-Day和N-Day漏洞的武器化进程急剧加速

过去，即使是公开已久的已知漏洞，攻击者仍需要具备一定的脚本编写能力、环境适配经验和绕过技巧才能完成稳定利用。现在，AI可以根据公开的漏洞描述或概念验证代码

片段，自动生成完整、稳定、可定制化的利用代码，并能根据目标环境的差异自动调整参数。一个更稳健的判断是：高危0-Day未必会暴涨，但1-Day和N-Day从公告到规模化利用的窗口一定会继续缩短。

（四）内网安全面临的威胁面被急剧放大

AI大模型能够根据攻击者输入的扫描结果、命令回显等碎片信息，分析内网拓扑结构、推荐横向移动路径、生成提权脚本，并以对话方式指导攻击者完成每一步操作。

API数据泄露风险空前加剧。AI能够自动解析API请求与响应结构，识别未授权访问、越权漏洞、敏感数据过度暴露、参数遍历风险。同时可通过分析API命名规律自动推测并探测潜在的高危接口。

二、证券行业特性放大AI攻击风险

如果说AI带来的攻击加速是整个数字世界共同面临的挑战，那么证券行业应用系统建设选型的特殊属性，使得这一挑战表现得尤为尖锐——相似系统的集中度和供应链的共性依赖，让一次成功的攻击模式可以低成本地在多家机构之间横向复制。

证券行业的IT供应链高度集中，头部供应商的系统在行业内占据主导份额。证券行业在长期发展中形成了高度标准化的业务模式和技术架构，核心交易系统、行情系统、风控系统、运维平台等关键基础设施在行业内存在大量同质化部署，一家机构采用的业务应用系统，很可能也是同行业其他机构的选择。这种相似性在传统安全评估中往往被视为“成熟稳定”的标志，但在AI时代却成了风险的放大镜。一旦攻击者利用AI挖掘出某个头部供应商的核心产品漏洞，影响范围可能覆盖数十家甚至上百家证券机构，同样的漏洞模式几乎可以无缝迁移到使用相同应用系统的机构。

供应链集中度风险的可怕之处在于其连锁反应的速度。传统模式下，一家机构发现漏洞后，可以通过行业信息共享机制通知其他机构，大家有时间逐个排查和修复。但在AI时代，从漏洞验证到全行业批量扫描再到自动化利用，中间的时间窗口可能短到让信息共享机制形同虚设。

简单的漏洞会更快暴露，测试环境尤其危险。许多证券机构的测试环境常被当作“非生产系统”对待，却往往存在跟真实生产环境同样的应用、组件、配置等，安全防护级别远低于生产环境，为了方便，可能还存在大量弱口令、通用口令。这些“低垂的果实”会先被攻击者发现并“摘掉”，进而导致生产环境同样问题被利用。AI不创造弱口令、未授权和老组件，只会让这些漏洞更快被发现、验证和利用。

三、解题之道：把检测与响应做成闭环

面对AI加速的攻击节奏和证券行业特有的风险放大效应，防守方必须从根本上调整防御思路。答案不是构筑更高的围墙——因为攻击速度已经让静态围墙失去了意义。真正

的解题之道在于：把检测与响应做成闭环。闭环的核心不是“没有漏洞”，而是“更快地发现、限制、清除和恢复”。

（一）以时间为核心重新定义安全度量

首先需要改变的是衡量标准。我们要处理的不是漏洞总数——那个数字永远让人焦虑——而是三个窗口的长度有没有变短：发现窗口、缓解窗口、恢复窗口。目标不是消灭漏洞清单，而是压缩从攻击发生到业务恢复的整条时间带。

这一转变的意义在于重新定义安全工作的优先级。漏洞的CVSS分数不再是决定修复顺序的唯一依据，更重要的是漏洞所在系统与核心交易链路的距离、暴露在公网的时间长度、以及一旦被突破后的恢复路径复杂度。无法立即修复的漏洞，也要先临时缓解、限权、隔离、检测，并设定到期复盘节点。指标不是为了好看，是为了判断安全投入有没有把闭环变快。

（二）以行为为抓手重构检测能力

传统的特征匹配防御在AI生成的变异攻击面前日趋无力。AI可以自动生成无数种变形和混淆变体，让基于固定特征的规则快速失效。防守方应将检测重心从攻击特征转向异常行为——异常的登录模式、异常的命令执行、异常的数据访问、异常的横向连接。

零日攻击防不住第一次触达，这是必须接受的现实。但防守方必须做到三件事：攻击者不能扩大影响范围——通过最小权限、网络分段、出站控制来限制横向移动的空间；攻击者不能保持隐身——通过行为检测让每一次异常操作都被记录，让攻击链可见；攻击者不能持久驻留——通过隔离、封禁、密钥轮换将攻击者逐出环境。审计日志缺失不是合规小问题，而是看不见攻击链的致命盲区。

（三）以韧性为底线重新定位安全目标

证券行业的底线不是“没有漏洞”，而是交易连续性和客户信任。漏洞只是入口，真正的影响最终会落到交易系统、客户资产、监管评价和品牌声誉。安全建设不能只按漏洞等级排队，还要按业务影响和恢复窗口排队。

这意味着安全部门需要与业务部门建立更紧密的协作机制。当安全事件发生时，判断是否触发业务应急预案、是否需要启动灾备切换、是否需要向监管报告，这些决策不能只基于技术指标，还需要综合考虑业务影响和恢复时间。安全部门的价值定位应从“找出所有漏洞”转变为“保障业务韧性”——让业务在面对攻击时能够持续运行、快速恢复。

（四）以闭环为目标构建组织协同

安全闭环不是安全部门单独能完成的。安全部门能发现问题、能提出建议、能部署工具，但形成真正的闭环——从发现到确认、从确认到缓解、从缓解到修复、从修复到复盘、从复盘到能力沉淀——需要整个组织的速度。

关键问题需要明确：谁负责资产、漏洞、系统、供应商？谁能看到日志、证据、权限、路径？谁来完成缓解、修复、复盘、整改？CMDB的一致性、审计日志的完整性、集权系统的

加固，每一样都要有清晰的责任边界，否则再好的技术工具也无法形成闭环。

闭环能力的度量不是看发现了多少问题，而是看每个问题是否真正走完了从发现到恢复的完整路径。每次事件结束，都应当沉淀成下一次更快的剧本——情报进入时自动关联资产、补齐上下文、触发检测规则、启动响应流程。威胁情报不能只做IP封禁，要进入DNS、应用、身份和出站检测。

四、解决之术：关键战术动作

在上述原则指导下，以下几个战术动作具有高优先级：

（一）源码保护与可反编译资产加固

前端JavaScript、移动应用安装包、小程序代码包必须进行深度混淆、加固或核心逻辑下沉至后端。对Java JAR/WAR包、.NET程序集、Android APK等可反编译资产采取代码混淆、加密关键类、运行时完整性校验等针对性加固。建立源码访问的零信任体系，确保源码不在任何环节意外泄露。

（二）前端零硬编码

所有配置通过后端接口动态下发，设置合理有效期和刷新机制。建立前端代码自动化检测流程，定期扫描硬编码痕迹。

（三）开源系统特征隐藏

部署前修改或替换默认图标、更改静态资源命名规则和路径结构、清除或替换HTTP响应头中的Server和X-Powered-By字段、自定义错误页面、修改默认会话Cookie名称。目标是将开源系统伪装成自研系统，大幅增加AI进行版本识别和漏洞匹配的难度。

（四）运行时检测与响应

在应用层部署ADR/RASP等运行时防护机制，将真实的利用行为纳入检测与响应体系。不能修不等于只能等补丁——先把影响面收住，把行为看见。

（五）AI诱饵密钥反制

在前端代码中埋设人工构造的、看似真实但实际永不使用的伪密钥。后端对该类伪凭证进行专属监控，一旦接收到携带该凭证的请求即可判定为AI自动化攻击行为，执行封禁IP、触发蜜罐引流、记录全量攻击流量等反制措施。甚至可以将AI诱饵升级为“计算陷阱”——诱导攻击者AI执行高复杂度算法以消耗其算力和Token。

（六）WAF行为化升级

对单个源IP在短时间内发起多种不同类型攻击尝试的行为进行综合风险评分并主动封禁，而非逐条拦截。建立AI攻击行为指纹库，对主流AI模型工具的特征进行持续跟踪。

（七）关键系统加固隔离

根据业务角色和数据敏感度将内网划分为独立的信任区域。对集权类系统、关键服务器启用双因子认证和操作审计，管理好账号职责分离，限制访问白名单等。对内网常见服务进行安全加固，禁止默认配置和弱口令。建立内网异常行为检测模型，对异常横向连接、权限提升等行为实时告警和自动阻断。

（八）API全生命周期安全管控

设计阶段强制每个API明确认证方式、授权模型和数据返回范围。开发阶段使用统一API网关进行鉴权拦截。测试阶段开展越权测试和未授权访问测试。运行阶段对API响应内容进行敏感数据自动脱敏，建立API资产清单和调用基线。

（九）主动代码审计

主动使用AI大模型或高级静态分析工具，对自有代码和第三方组件进行常态化、全覆盖的代码审计。重点扫描身份认证、权限校验、SQL操作、文件读写、反序列化等高风险区域。目标是在自己被AI打穿之前，先用AI找到并修复漏洞。

（十）供应链协同防御

针对证券行业供应链集中度风险，需要深入融入监管的联防联控体系，主动获取威胁情报的同时也主动贡献自身的发现，形成行业内持续协同的防御机制。一家机构发现的问题，要快速变成一类系统的排查方法。当某家机构在某个供应商产品中发现漏洞时，信息应同步到行业其他机构、其他使用同类产品的机构以及核心供应商。输出不是一份报告，而是可复用的排查方法——如何确认系统是否受影响、如何检测漏洞是否已被利用、如何在不影响业务的前提下完成修复。同时，应对供应链中的共性产品建立持续监控机制，一旦发现新的漏洞模式，能够快速定位全行业受影响面，在攻击者完成利用之前启动应急响应。

五、以闭环对抗焦虑，以速度赢得底气

AI大模型在攻防领域的应用，显著放大了本就存在的攻防不对称性。攻击方不再依赖少数高技能人员，而是可通过AI赋能实现漏洞挖掘、利用生成及内网渗透的规模化、自动化。证券行业因其系统的相似性和供应链的集中度，面临的风险放大效应尤为突出。

AI时代安全防御的“道”，在于完成四个层面的认知跃迁：以时间为核心，从“有没有漏洞”转向“暴露多久、止损多快”；以行为为抓手，从特征匹配转向异常行为检测；以韧性为底线，从追求无漏洞转向保障业务连续与客户信任；以闭环为目标，从安全部门的单打独斗转向组织级的闭环责任体系。

AI时代安全防御的“术”，则是上述理念的系统落地：暴露面管理覆盖源码保护、开源特征隐藏、API全生命周期管

控；日志与证据链确保每一次异常操作都被记录、每一个攻击步骤都可追溯；运行时检测响应让攻击者不能扩大、不能隐身、不能久留；供应链协同防御让一家发现的问题变成一类系统的排查方法。

AI在赋能安全的同时，其自身也面临严峻的安全挑战。证券行业在推进AI+安全时，必须同步构建“Security for AI”的保障体系——在数据层面防范投毒风险，在模型层面防范提示注入和篡改风险，在部署层面采用私有化确保

数据不出域，在治理层面强化关键环节的人类把关机制。安全部门的价值正在从单次执行转向 workflow 设计和验证体系建设。

AI时代安全能力的核心评判标准是：是否能够更快地形成闭环——不是保证永远没有漏洞，而是在攻击进入机器速度之后，更快地发现、限制、清除和恢复。真正的“道”与“术”，不是追逐技术的神秘感，而是把能力可持续、可验证地放进 workflow 和决策链中。

证券行业数字员工安全问题研究与探索

梁德汉、范睿华、李家攀、刘亦翔 | 国投证券股份有限公司

摘要：随着大语言模型（LLM）与智能体（Agent）技术的发展，数字员工正逐步从传统规则驱动的自动化工具演进为具备自主规划、推理决策和工具调用能力的新型数字劳动力。在提升证券行业运营效率的同时，其自主性和复杂交互能力也带来了新的安全挑战。针对数字员工的身份管理、目标对齐、能力控制和行为约束等问题，基于“身份-目标-能力-行为”四维分析框架，构建覆盖准入、运行和退出阶段的全生命周期安全治理体系。在此基础上，结合证券行业实践，探讨大模型安全网关、SkillsHub、安全沙箱、安全围栏及AgentShield等关键安全措施的设计与应用，为证券行业数字员工安全治理提供参考方案。

关键词：数字员工、智能体、全生命周期治理、大模型安全、主动防御

一、背景和目标

近年来，以大语言模型（LLM）和智能体（Agent）为代表的生成式人工智能技术快速发展，推动证券行业数字化转型进入新的阶段。传统基于固定规则的机器人流程自动化（RPA）正逐步向具备自主规划、推理决策和工具调用能力的数字员工（Digital Worker）演进。

与传统自动化工具相比，数字员工具备更强的自主性和跨系统协作能力，能够在业务、办公、研发、运维及安全等场景中持续执行复杂任务。与此同时，其所具备的自主决策、工具调用和权限继承能力，也对传统安全边界、权限管理模式和责任认定机制带来了新的挑战。

证券行业具有监管要求严格、核心业务关键、客户数据敏感等行业属性，大模型幻觉、提示词注入攻击、工具能力滥用、自主行为失控等原生风险，在证券业务场景中易被进一步放大，可能引发合规处罚、业务中断、数据泄露等严重后果。因此，如何在利用人工智能生产力的同时实现风险可控，已成为数字员工规模化应用过程中亟需解决的问题。本文围绕数字员工全生命周期安全治理展开研究，基于“身份-目标-能力-行为”四维分析框架，并结合证券行业实践探索相应的安全治理体系。

二、数字员工及治理分析

（一）数字员工的定义

数字员工及治理分析与传统软件系统不同，数字员工并非单纯的信息处理工具，而是一种具备组织身份、明确目标、专业能力和自主行为的数字主体。

（二）从RPA到AI Agent演进

数字员工的发展经历了从“规则驱动”向“认知与执行

双驱动”的代际跨越，其核心差异可拆解为表1所示的五个维度：

维度	RPA时代（1.0/2.0时代）	LLM+Agent时代（3.0时代）
驱动核心	固定规则驱动：依赖人工预设条件、脚本和录制的工作流。	大语言模型驱动：依赖大模型的常识推理、语义理解与主动规划。
输入理解	结构化数据：只能识别标准的Excel表格、固定格式的表单，一旦格式微调即报错。	非结构化与多模态数据：能读懂扫描件、PDF财报、自由文本、音视频和统计图表。
异常处理	死机中断：对于环境变化（如网页按钮挪动1像素）缺乏容错能力，直接中断。	自主反思与自纠错：能根据接口或环境反馈自主调整规划路线。
交互范式	低代码/拖拽编码：需要IT专业人员编写逻辑代码或维护预设流转节点。	自然语言交互：真实员工可使用自然语言直接下达复杂任务，实现“即说即办”。
组织角色	数字化搬运工：只负责跨系统的数据复制、粘贴、录入等机械化、体力型重复工作。	智能助手：能主动查阅资料、起草分析报告、撰写研报、执行安全漏洞验证等。

表1 传统数字员工与现代数字员工的核心差异

（三）数字员工治理框架

基于对数字员工“数字主体”本质的认识，从组织管理视角看，数字员工承接了传统员工的部分岗位职责；从智能体（Agent）视角看，其具备任务规划、推理决策、工具调用和自主执行能力；从身份治理视角看，其拥有独立的数字身份、权限边界和审计链路。该框架将数字员工的治理属性拆解为四个互相关联的维度：

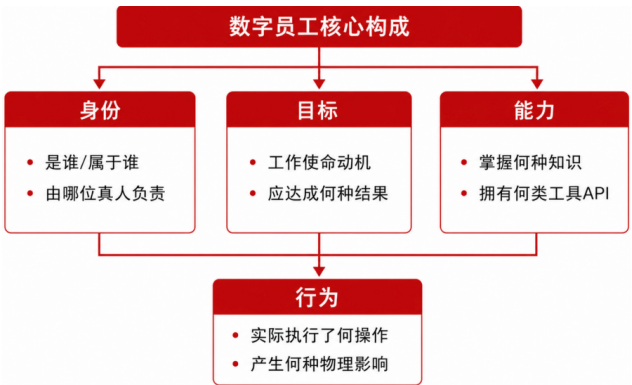


图1 数字员工核心构成图

身份：定义数字员工“是谁、属于谁、由谁负责”。包括其在统一身份管理中的数字标识、组织归属以及绑定的责任人。

目标：目标定义数字员工“为什么工作、应达成什么结果，以及必须遵循哪些业务规则、安全要求和行为边界”。数字员工的推理和决策受到系统提示词、业务规则、安全策略、价值观及监管要求等约束，共同决定数字员工的行为。

能力：定义数字员工“掌握什么知识、具备什么工具和资源”。表现为模型本身的推理能力、外挂的私有向量知识库（RAG），以及注册的API、执行脚本和工具集（Skills/MCP）。

行为：定义数字员工“实际执行了哪些操作、产生了什么影响”。它是行为计划的落地执行，表现为调用外部系统写接口、执行命令行或外发数据。

数字员工的安全风险也相应来源于这四个维度：身份缺失导致权责不清与合规风险，目标偏移导致决策与认知偏差风险，能力滥用导致权限与高危工具失控风险，行为失控导致操作风险，而提示词注入、越狱（Jailbreak）、恶意Skills、MCP攻击、模型工具链及供应链污染等技术攻击则会对上述各个维度产生叠加影响。

因此，数字员工安全治理的核心并非单纯保护模型或代码本身，而是围绕数字员工这一新型数字主体，构建覆盖“身份、目标、能力和行为”且贯穿数字员工全生命周期的治理体系。

三、数字员工的典型应用场景

当前，随着大语言模型与智能体技术的快速发展，数字员工已成为证券行业人工智能应用的重要方向之一。证券行业已围绕投研分析、财富管理、客户服务、运营保障与协同办公等场景开展数字员工的应用探索，形成超级研究员、智能投顾助手、营销助手、尽调助手、办公助手等多种应用形态；网络安全领域也逐步出现面向安全运营、漏洞研判、威胁分析和自动化处置的AI数字员工，并逐步推动数字员工从辅助工具向业务协同角色演进。

在投研领域，部分机构已尝试将大模型与知识检索、智能体规划等技术相结合，实现研究报告辅助生成、公告信息提取、会议纪要整理等功能，提升信息处理效率和研究支持能力；在财富管理与客户服务领域，数字人、智能投顾助手等应用逐步落地，通过自然语言交互、多模态内容生成等能力，为客户提供投资咨询、业务办理指引和市场资讯服务；在协同办公领域，数字员工被广泛应用于知识问答、文档生成、流程辅助和数据查询等场景，有效降低重复性事务工作带来的负担。

总体来看，数字员工正在逐步融入证券行业的业务、管理和技术体系，并呈现出从单点应用向多场景协同、从任务自动化向智能化决策辅助发展的趋势。这种生产力形态的演进在提升运营效率的同时，也使数字员工逐渐具备更强的自主决策和系统交互能力，对传统安全边界、权限管理模式以及风险控制体系提出了新的挑战。

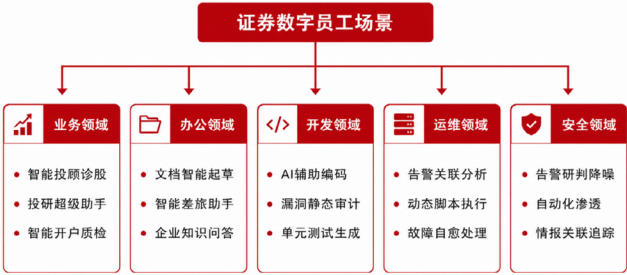


图2 证券数字员工场景

在数字化转型实践中，数字员工已逐步融入证券行业多个核心场景。

(1) 业务领域。应用于智能投顾、投研分析、开户辅助及合规质检等场景，实现客户服务、知识分析和业务流程自动化。

(2) 办公领域。应用于文档生成、知识问答、差旅报销等日常办公场景，提升组织协同效率和知识服务能力。

(3) 开发领域。应用于代码生成、漏洞审计和测试用例编写等软件研发环节，辅助提升研发质量和交付效率。

(4) 运维领域。应用于告警分析、配置变更和故障处置等运维场景，增强系统运行保障和自动化运维能力。

(5) 安全领域。应用于告警研判、漏洞验证、威胁情报分析和恶意样本研判等安全运营场景，提升安全事件发现与响应效率。

总体来看，数字员工正在从单点任务自动化工具逐步演进为贯穿业务、办公、研发、运维和安全等多个领域的智能协作主体，并具备跨系统交互、自主规划和工具调用能力。

四、基于生命周期的攻防视角风险分析

安全威胁贯穿数字员工全生命周期，按运行流程可划分为“准入、运行、退出”三个阶段，每个阶段均存在攻击侧威胁利用与防御侧安全管控的持续对抗，对应攻防矩阵如下：

阶段	步骤	维度	攻击视角（风险）	防御视角（措施）
准入阶段	资产准备	身份/能力	供应链攻击、数据投毒	模型安全扫描、语料清洗、数字员工池隔离
	角色权限分配	身份	身份冒用/责任主体不清、权限过度授予	IAM统一纳管、最小权限控制
	Skills发布上线	能力	恶意Skills、供应链污染	SkillsHub审计与签名准入
	部署与上线	目标	Prompt后门、Jailbreak风险	红队测试、提示词审计、影子沙箱验证
运行阶段	任务交互与输入	目标/能力	Prompt Injection、敏感数据泄露	大模型安全网关、动态脱敏
	推理与逻辑规划	目标	模型幻觉、错误决策、误操作、提供投资建议	运行沙箱、安全围栏
	工具与API执行	能力/行为	MCP攻击、工具链污染	API安全网关、AgentShield
退出阶段	账户注销与清理	身份	孤儿账号、权限残留	自动化身份与权限回收
	环境与资产清理	能力/行为	数据残留、资产遗留	沙箱销毁、审计归档

表2 数字员工全生命周期攻防对抗矩阵

五、基于四维框架的安全风险分析

实践表明，数字员工安全风险并非来源于单一技术缺陷，而是数字身份管理、目标对齐机制、能力边界控制以及自主行为执行等多个因素共同作用的结果。从生命周期视角来看，不同阶段呈现出不同的风险表现形式；而从治理视角分析，这些风险均可归结为数字员工的身份、目标、能力和行为四个核心维度。因此，本文进一步基于“四维框架”对数字员工安全风险进行分类，从风险本质层面识别数字员工面临的关键威胁，为后续安全治理体系设计提供理论依据。

（一）身份维度：身份治理缺失与权责失衡风险

身份是数字员工作为组织数字主体存在的基础，也是责任认定、权限管理和合规治理的核心载体。当数字员工身份管理缺失或责任边界不清晰时，极易引发组织治理失效和监管合规风险。

责任主体缺位及权责边界模糊：在现有法律框架下，数字员工不具备独立法律主体资格。若智能投顾数字员工因决策偏差导致客户损失，或运维、安全数字员工因自主操作引发生产事件，其法律责任与管理责任易在业务部门、开发团队和系统管理者之间出现界定模糊、追溯不清的问题。

数字员工身份管理失控：数字员工在实际运行过程中往往拥有独立账号、访问凭证和系统权限。若其身份生命周期未纳入统一身份与访问管理体系（IAM）进行管理，则可能形成长期有效的“僵尸账号”或“孤儿账号”，成为攻击者进入核心业务系统的重要入口。

（二）目标维度：目标偏移与认知失真风险

目标决定数字员工的行为方向和决策逻辑，是智能体推理规划过程中的核心驱动力。当目标约束机制不足、模型认知出现偏差或目标受到外部恶意诱导时，可能导致数字员工产生与组织预期不一致的行为结果。数字员工不仅可能因认知缺陷产生错误决策，还可能为了完成目标而突破业务规则、安全要求或行为边界，引发业务风险和合规风险。

目标约束失效与目标劫持风险：攻击者可通过提示词注入（Prompt Injection）、越狱攻击（Jailbreak）等方式破坏数字员工既定目标约束，诱导其接受新的任务目标或忽略原有安全要求，使数字员工偏离组织既定目标，执行违规、越权或非预期操作。从本质上看，此类攻击属于针对数字员工目标体系的社会工程攻击。

认知偏差与决策公平性风险：生成式人工智能固有的幻觉问题以及模型训练数据偏差，可能导致数字员工生成错误事实、虚构信息或产生隐性偏见，在客户服务、授信评估、投资分析等高精度业务场景中输出错误结论或不公平决策，影响业务准确性、公平性及监管合规。

（三）能力维度：能力滥用与工具链攻击风险

能力决定数字员工能够访问哪些知识、调用哪些工具以及执行哪些任务。随着数字员工逐步具备外部工具调用和自主执行能力，其攻击面也从模型本身扩展至整个工具生态和

供应链体系。

推理与知识库风险：数字员工的推理能力及其通过RAG等方式接入的外部知识共同影响任务判断与决策结果。若模型推理出现偏差，或知识库内容被污染、篡改、混入不可信数据，可能导致错误判断、错误决策；若知识库中包含敏感数据，且检索、调用或输出过程缺乏有效控制，还可能造成敏感数据泄露。

恶意Skills与工具链污染：攻击者可能通过第三方插件、开源工具或API描述文件篡改等方式污染数字员工能力体系，使其在推理过程中产生错误决策或执行越权操作。

MCP攻击与供应链风险：随着MCP等新型模型交互协议的广泛应用，数字员工与外部系统的连接边界持续扩大。一旦协议实现、底层模型或智能体框架存在安全缺陷，将可能导致敏感数据泄露、身份劫持或远程控制等安全事件。

（四）行为维度：行为失控与越权执行风险

行为是数字员工推理结果最终落地到现实世界的具体体现，也是风险影响业务系统和生产环境的直接表现。当数字员工拥有自主执行能力后，其行为风险往往具有影响范围广、扩散速度快和损失后果严重等特点。

自动化执行导致的系统性风险：与传统RPA不同，数字员工具备自主规划和连续执行能力。一旦出现推理偏差、目标漂移或受到恶意指令操控，可能在短时间内连续调用大量接口，触发业务异常、系统崩溃或声誉风险事件。

内网越权与横向渗透风险：数字员工通常需要跨系统获取数据和执行任务，若运行环境缺乏有效隔离，遭受攻击并被控制的数字员工可能成为内网横向移动的跳板，通过SSRF、权限继承或接口滥用等方式访问敏感资源，进一步扩大攻击范围。

综上所述，数字员工安全风险贯穿其整个生命周期，并最终表现为身份、目标、能力和行为四个维度的治理问题。因此，数字员工安全治理的核心并非单纯提升模型安全能力，而是围绕数字主体建立覆盖身份管理、目标约束、能力控制和行为审计的一体化治理体系，从而实现数字员工能力释放与风险控制之间的动态平衡。

六、数字员工安全设计与实践

针对数字员工在准入、运行和退出阶段面临的不同安全风险，结合证券行业安全要求高、合规约束强、审计要求严的行业特点，需要构建覆盖数字员工全生命周期的安全治理体系。治理体系以数字员工“身份可管理、目标可约束、能力可控制、行为可审计”为总体目标，在数字员工生命周期各阶段建立相应的组织治理机制和技术控制措施，实现风险控制前移、过程可控和责任可追溯。

（一）准入阶段安全治理

数字员工准入阶段的治理重点在于建立可信身份、可信能力和可信运行环境，防止存在安全缺陷或恶意能力的数字员工直接进入生产环境。

身份治理设计方面，以“唯一身份、责任明确、统一管理”为设计原则，将数字员工纳入企业统一身份与访问管理体系（IAM），注册唯一数字身份，并与具备相应业务授权的申请人建立明确责任关系。数字员工在系统中的权限边界、审计责任和业务归属均继承其申请人，实现数字员工身份与组织责任的统一管理。落地实践方面，数字员工由员工本人发起申请，并签署数字员工使用承诺书，明确数字员工使用规范、责任边界及安全要求，经审批通过后方可创建数字员工身份。数字员工统一纳入IAM管理体系，并采用员工账号增加特定标识后缀的方式建立数字员工身份，与申请人形成一一对应关系，实现数字员工身份注册、责任明确、权限继承和生命周期统一管理。数字员工默认运行于安全沙箱环境，确保其运行边界可控。

目标治理设计方面，以“目标明确、约束完整、持续验证”为设计原则，在赋予数字员工业务目标的同时，同步定义其角色职责、安全要求和行为边界，将业务规则、合规要求及安全策略融入数字员工目标体系，确保数字员工始终围绕组织既定目标开展推理和决策，避免因目标约束失效、提示词注入或恶意诱导导致数字员工偏离组织预期。落地实践方面，通过系统提示词定义数字员工角色职责、业务目标及安全约束，并将系统提示词静态分析、提示词红队测试（Prompt Red Teaming）、越狱攻击测试（Jailbreak Testing）等能力纳入数字员工上线流程，对数字员工目标约束机制的有效性、抗提示词注入能力及行为稳定性进行综合验证，确保数字员工满足安全要求后方可进入生产环境。

能力治理设计方面，以“统一纳管、可信接入、最小能力、最小权限”为设计原则，建立企业统一的SkillsHub，对数字员工使用的Skills、MCP工具及第三方能力组件实施集中管理。所有能力组件在上线前均需完成静态代码审计（SAST）、软件成分分析（SCA）、恶意行为检测和数字签名认证，未经审核批准的能力组件禁止加载运行，从源头降低供应链攻击和恶意工具注入风险。落地实践方面，数字员工镜像在上线前需完成静态代码审计（SAST）、软件成分分析（SCA）及安全检测，确保镜像不存在已知安全漏洞、恶意代码或高风险组件，满足生产环境运行要求后方可正式上线。在能力组件管理方面，建立企业统一的Skills Hub，对数字员工使用的Skills实施统一注册、审核、发布和版本管理。所有Skills在上线前均需完成安全扫描和合规审批，未经审核批准的Skills禁止进入生产环境。数字员工仅允许从企业内部Skills Hub安装和调用技能，实现Skills的统一治理和可信接入。在模型能力管理方面，建立统一的Token Hub，对数字员工使用的大模型进行集中管理。数字员工仅允许访问企业统一纳管的大模型，所有模型调用均通过大模型安全网关，对提示词注入、内容安全及敏感数据泄露等风险进行统一检测和过滤，并统一记录访问日志，定期开展安全审计。未经授权的外部大模型及外部Skills Hub默认禁止访问，从能力调用源头实现统一管控。

（二）运行阶段安全治理

运行阶段是数字员工执行任务和产生业务价值的核心

阶段，也是安全风险最为集中的阶段。与准入阶段侧重可信准入不同，运行阶段安全治理更加关注数字员工运行过程中的持续可信，围绕目标、能力和行为建立纵深防御体系，实现数字员工运行全过程安全控制。

目标治理设计方面，以“目标明确、约束完整、持续验证”为设计原则，在数字员工运行过程中持续保护目标约束体系，确保数字员工始终围绕既定业务目标开展推理和决策，避免因提示词注入、越狱攻击等导致目标约束失效而偏离组织预期。落地实践方面，通过在数字员工与大模型之间部署大模型安全网关（LLM Security Gateway），安全网关负责对输入和输出内容进行实时检测与过滤，对提示词注入（Prompt Injection）、越狱攻击（Jailbreak）、恶意载荷（Payload）以及敏感信息泄露等风险进行识别和拦截。同时，在数据进入模型之前，对客户隐私信息、企业敏感数据和核心代码等内容实施动态脱敏处理，降低数据泄露风险。

能力治理设计方面，以“统一授权、按需调用、边界可控”为设计原则，持续控制数字员工运行过程中的能力调用边界，确保数字员工始终在授权范围内调用业务能力。落地实践方面，建立统一API安全网关和个人权限级接口体系，数字员工发起的所有API调用均需经过安全网关统一校验，并强制继承其绑定责任人的个人权限边界。任何超出责任人授权范围的访问请求均被即时阻断，从机制上防止数字员工因提示词注入、工具链污染或逻辑偏移导致的权限扩张和横向越权行为。

行为治理设计方面，以“行为受控、实时监测、快速处置”为设计原则，对数字员工执行过程实施持续监测和动态控制，确保其所有操作均运行在安全可控范围内。落地实践方面，所有代码执行、命令调用及自动化任务均运行于安全隔离沙箱之中。沙箱环境采用容器隔离、资源配额限制和网络微隔离等技术手段，将数字员工活动范围限制在可控范围内。同时，利用确定性规则对高危行为进行强制校验，当检测到高危权限操作、敏感网络操作或高风险业务操作时，联动安全系统直接对数字员工进行关停处理，避免风险进一步扩散。对于涉及资金调拨、授信调整、安全策略变更等高风险业务场景，引入“人在环路（Human-in-the-Loop）”控制机制，数字员工仅负责分析和决策建议生成，最终执行必须经过人工授权，确保关键业务始终处于人工可控状态。

（三）退出阶段安全治理

数字员工退出生产环境后，其遗留身份、访问权限和运行数据仍可能成为潜在安全风险。在退出阶段安全治理中，以“彻底回收、安全销毁、可信退出”为设计原则，建立标准化的数字员工退役治理流程，确保数字员工退出后不再具备持续访问和影响生产环境的能力。

身份治理设计方面，以“身份注销、权限回收、责任解绑”为设计原则，实现数字员工身份和访问权限的自动回收。落地实践方面，IAM系统与人员管理系统联动，当数字员工停用或责任人发生变更时，自动停用数字员工IAM账号，并同步完成访问令牌失效、权限回收以及访问控制列表（ACL）清理，避免形成孤儿账号和长期有效凭证。

运行环境治理设计方面,以“环境销毁、实例失效、能力隔离”为设计原则,确保数字员工退出后不再具备持续运行能力。落地实践方面,数字员工停用后,自动销毁其运行容器,容器内安装的Skills随容器一并清理,运行过程中产生的缓存数据、临时目录和会话状态同步删除。MCP服务及API接口作为企业统一能力组件继续保留,由后续数字员工按授权复用,实现数字员工运行环境的安全退出和能力隔离。

七、总结与展望

(一) 总结

数字员工作为大模型技术的重要应用形态,正逐步从辅助工具演进为具备自主感知、规划、决策和执行能力的新型智能体,其自主性、开放性和持续运行特征,使传统以系统和数据为中心的安全防护体系面临新的挑战。本文结合数字员工的运行特点,提出了面向数字员工的身份、目标、能力、行为四维安全风险框架,并围绕数字员工全生命周期,构建了覆盖准入、运行、退出三个阶段的安全治理体系。落地实践方面,通过建立可信身份体系、目标约束机制、能力治理体系和行为控制机制,实现数字员工的身份可信、目标可信、能力可信、行为可信,为数字员工的安全接入、安全运行和安全退出提供了系统性的治理思路和实践参考。随着数字员工逐步成为企业新的生产力,构建覆盖全生命周期的安全治理体系,将成为企业推动数字员工规模化应用的重要基础能力。

(二) 展望

随着大模型技术持续发展,模型上下文窗口、长期记忆、Agent Harness、工具调用能力以及MCP、API等开放生态将不断成熟,数字员工将逐步具备持续感知、自主规划、多智能体协同和复杂任务执行能力,从“智能助手”逐步演进为能够独立承担业务职责的数字劳动力。

目前,金融、高科技、互联网等行业已陆续开展数字员工建设实践,数字员工正由单点业务试点逐步走向企业级规模化应用。与此同时,人工智能相关法律法规、行业标准和治理体系也将持续完善,为数字员工健康、有序发展提供制度保障和合规依据。技术能力、产业生态与治理体系的不断成熟,将进一步推动数字员工在企业中的广泛应用。

长远来看,企业组织形态与人机协同模式将迎来深刻变革。人类员工将逐步从重复性事务中释放,聚焦于业务研判、AI治理、风险管控与组织管理等高价值职责;可标准化、流程化、规则化的日常工作将交由数字员工自主完成,形成“人负责决策与治理,数字员工负责执行与落地”的权责清晰、安全可控的新型协同模式。

随着数字员工规模持续增长,其将逐步成为企业新的生产力形态,企业安全治理的对象也将由传统的信息系统、账号和数据,扩展至具备身份、目标、能力和自主行为的数字劳动力。未来,数字员工安全治理将从单点技术防护走向体系化治理,更加关注身份可信、目标可信、能力可信和行为可信,并推动企业建立与数字劳动力相适应的安全治理机制和责任体系,实现从“系统安全”向“数字劳动力安全治理”的演进,为数字员工规模化、可信化应用以及企业智能化转型提供基础保障。

参考文献

- [1] 大模型驱动的数字员工3.0建设应用白皮书[R]. 2024
- [2] 智能体驱动的API安全风险管控研究与实践[J]. 交易技术前沿(网络安全专刊), 2025
- [3] 国家互联网信息办公室,等.生成式人工智能服务管理暂行办法[Z]. 2023
- [4] 中国信息通信研究院. 人工智能安全治理研究报告(2025年) [R]. 2026
- [5] Financial and Securities Industry Digital Worker Security Research[R]
- [6] Tencent Zhuque Lab. AI-Infra-Guard:A full-stack AI Red Teaming platform[CP/OL].GitHub, 2026
- [7] TRiSM-aligned architecture for Agentic AI systems (2506.04133v5)[R]. 2025

大模型在证券行业网络安全威胁建模的落地实践

黄韦、成昊、郑洲豪 | 国泰海通证券股份有限公司

摘要：针对证券行业威胁建模高度依赖安全专家的瓶颈，本文提出基于大模型的“泰坦”平台。系统设计需求阶段安全场景自动识别与代码提交持续安全评审两层闭环，实现安全左移。面对大模型输出发散、人机责任冲突及行业特异性三重约束，通过驾驭工程的结构化推理使结果可控，设置人工审核卡口，并内置含200余场景的行业知识库。实践表明，平台已落地核心系统，安全场景识别精确率达92.61%，有效降低对安全专家的依赖与修复成本。

关键词：威胁建模、开发安全、驾驭工程、DevSecOps

一、引言

威胁建模在系统设计阶段识别潜在威胁、分析攻击面并制定缓解措施，是安全左移的理想切入点。然而，这一环节需业务、开发、安全三方协同，对人的依赖极高。在企业中，安全专家是稀缺资源，往往难以全程参与每个项目；在缺乏专家介入时，传统威胁建模方法论在项目组中难以独立运转。

针对这一问题，我司基于大模型构建了“泰坦”威胁感知与应对平台，以私有化部署的大模型为核心，在无需安全专家全程介入、不额外增加研发负担的条件下，实现了威胁建模的自动生成与代码变更的持续安全评审。

二、威胁建模的传统瓶颈

威胁建模是在编码之前对设计方案安全性的审视，是一个老生常谈的安全方法论。

从业界优质的解决方案看，微软SDL威胁建模工具以数据流图为核心输入，基于STRIDE模型自动生成威胁清单；OWASP Threat Dragon提供类似轻量化方案；CSA Cloud Threat Modeling将视角转向云环境资产与业务目标。这些工具均遵循统一模式：输入、分析框架、输出。

输入对于威胁建模至关重要。这一模式的成立依赖一个前提：高质量的分析输入可自动获得，且不依赖安全专家。现实相反。以微软的数据流图为例，绘制合格的图需同时理解系统架构、数据流向及信任边界，这本身就是需要安全思维的技术工作。在我司的实际落地过程中，一开始以安全问卷的形式让项目相关人员，根据自身的需要填写安全问卷来生成威胁建模。然而，现实结果是，产品经理了解业务但不熟

悉架构，开发工程师掌握实现但不了解安全方法论。结果往往无人能完成输入，或是输入质量参差不齐，导致威胁清单遗漏关键风险或充斥无关告警。

核心症结：工具在演进，瓶颈始终在人。“从系统设计到设计图”与“从威胁清单到有效缓解措施”这两步，仍高度依赖安全专家的经验与判断力。

三、大模型的机遇与现实约束

大语言模型在代码理解、安全分析和文档处理方面展现的能力，为上述问题提供了新方向。然而，将大模型置于生产的真实条件下，面临三重现实约束：

第一，如何让大模型的输出结果可控可信。

威胁建模对准确性要求极高——漏报可能留下安全敞口，误报则消耗开发团队的信任。然而大模型本质上是概率生成引擎，其输出天然具有发散性和不确定性。在数据不外传、模型私有化部署的前提下，同样无法寄望于依赖超级模型解决问题。因此如何令模型稳定的产出可控可信的安全分析结果，是本系统面对的首要约束。

第二，人机的责任与效率矛盾。

对安全承担最终责任的依然是人，AI不能直接面向研发人员输出结果，必须经由安全专家作为评审人，进行安全决策的卡口。而模型输出是自动化的、大量的，安全专家相较于大模型低效但是可靠。如何令人与机器达成责任与效率的平衡，这是第二重约束。

第三，证券期货业的特异性安全场景。

行情推送、交易指令下达、实名认证、营销活动等场景的安全需求根植于具体业务流程与监管规范，超出通用安全知识体系的覆盖范围。若不将行业知识注入系统，大模型的产出将仅是一份泛化的安全清单。

四、系统设计

针对上述约束，通过大模型与实际场景结合，做出如下解决方案：以结构化推理补偿模型能力、以人机协同把控最终节点、以内置行业场景库注入领域知识。

（一）两层自动化闭环

系统的整体架构采用两层自动化结构设计，基于大模型的能力，实现安全分析的持续产出，使威胁建模无感知、无门槛地落地到研发的过程中。

第一层：安全场景自动识别。

系统接收原始业务需求及产品经理编写的需求条目，从中自动识别涉及的安全场景，进而从知识库匹配对应的安全设计要求和测试用例。

第二层：代码变更的持续安全评审。

研发人员提交代码后，自动获取变更记录并分析潜在安全风险。发现安全漏洞风险时，生成包含漏洞定位、成因分析和风险描述的说明并推送至对应研发人员。安全评审随每次代码提交持续感知。

两层自动化构成闭环：需求阶段产出的安全设计在代码提交阶段被反向验证，系统不仅检查通用安全漏洞，同时验证代码是否落实了需求阶段提出的安全设计要求。

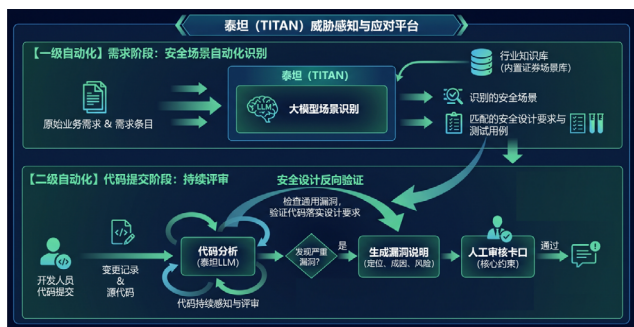


图1 泰坦系统两层自动化闭环示意

（二）基于驭驭工程保障结果可控可信

针对大模型输出发散、不确定的问题，系统引入了驭驭工程范式。其核心思想在于：在不改变模型内部参数的前提下，通过优化模型运行的外部“系统环境”来约束其行为，降低输出的随机性，从而获得稳定、可预期的结果。

在初步实验中，若直接将需求文档输入模型并允许其开放式生成，输出往往面临格式不统一、缺乏推理依据以及存在“幻觉”等问题

为此，本系统将开放式任务转化为受约束的结构化推理，通过“补充上下文”与“设置流程护栏”，设计了三步递进的分析链路：

第一步，进行技术扩写以补充上下文：

引导模型根据原始业务需求，理解可能涉及的开发框

架、通信协议与数据存储等技术实现方案。此步骤旨在完善后续分析所需的基础信息。

第二步，依托场景匹配引入系统护栏：

系统限制模型的输出范围，要求其在预定义的行业安全场景库中进行匹配，并强制标注每项匹配结果的具体依据是源于需求原文，还是衍生自技术推断。预定义的场景库在此起到了规范输出边界的作用。

第三步，执行事实溯源与结果校验：

通过程序化逻辑，剔除仅基于技术推断的场景，严格保留具备需求原文明确支撑的输出。该步骤剥离了模型主观的发散性，从根本上保障了最终结果的客观可信度。

同理，在代码安全评审环节，系统采取了“先广度扫描、后深度分析”的标准化流程。每个步骤的边界明确，模型的职责被限定在单一子任务内。

实践表明，合理的系统结构与环境约束能够有效规避模型在稳定性上的不足。通过系统环境的设计，系统成功将大模型发散的概率生成过程，转化为工程上可控、结果可信的安全分析产出。

（三）人工审核卡口

大模型可以批量化地、自动产出大量的安全分析结果，然而对安全问题进行决策的责任却不可以让渡给大模型。

本系统将安全专家保留在大模型产出之后、推送至研发人员之前的节点上，以人机协作的方式进行对安全问题的审阅。

安全专家面对的是已结构化处理且附带判断依据的信息，核心任务不是从零发现安全问题，而是对大模型的判断进行审核——确认、驳回或补充。场景选择引述需求文档原文，漏洞判定关联至具体代码片段。专家的各项判断回流用于优化场景库与推理参数，人在回路中同时承担把关者与训练者的双重角色。

（四）内置行业场景库

本系统构建了一套涵盖200余个安全场景的行业知识库，对证券业务安全风险进行了体系化建模。

该知识库分为三个维度，形成对系统安全的完整覆盖：

系统架构层：界定“在哪里产生风险”，重点关注系统边界、多端接入面以及二三方系统的交互安全。

业务功能层：回答“什么业务场景面临风险”，沿业务线展开，覆盖核心交易、行情分发、办公内控等核心领域。

应用实现层：指明“以何种技术形态暴露风险”，聚焦代码指令、身份鉴权、数据解析等具体技术点的安全考量。

场景库融合了行业安全标准、历史安全事件的经验抽象以及公司多年的安全评审实践积累。每个场景均深度关联了安全设计建议、测试方法及核查要点。更重要的是，本系统能够持续从日常迭代的需求文档中识别并沉淀新场景，将专家的隐性经验源源不断地转化为系统的显性知识资产。

安全场景	核心风险
● 边界与系统交互 1) 内部微服务与分布式接口间通信调用 2) 与二方集权系统（统一认证、堡垒机、私有仓库）交互 3) 与三方合规接口（实名认证、支付网关）交互 ● 多端接入 1) App、PC端、小程序、IoT及浏览器跨端交互 2) H5本地资源缓存与消息队列中间件数据交换	越权与绕过：边界未授权跨越 凭据泄露：集权系统敏感凭据丢失 数据劫持：三方接口欺诈与中间人篡改 客户端安全：逆向破译与本地缓存污染
● 核心交易与行情分发 1) 交易办理：证券交易生命周期、投资组合、柜台开户与支付风控 2) 行情服务：极速推送加密、行情数据生命周期管理、断流监控告警 ● 内部管理与社会运营 1) 办公内控：大批量报表流转、审批流全链路审计、考勤打卡定位校验 2) 社区营销：UGC内容安全向审计、平台风控运营、营销抽奖与身份验证 ● 用户服务与资讯 账户生命周期管理、多媒体客服安全通信、个性化资讯分发	交易合规风险：交易抵赖、越权操作他人订单、高价值逻辑绕过 业务连续性：行情数据恶意篡改、通讯拒绝服务、监控漏报 内部审计失效：审批流程篡改、恶意宏注入、员工隐私泄露 黑产与舆情：黑产批量薅羊毛、违规涉稳舆情泛滥 账户安全：撞库爆破、凭证欺诈与恶意钓鱼
● 代码指令与Web配置 代码执行：指令解析、系统命令调用、URL重定向与高频正则匹配。 Web防护：安全日志监控、基础安全头部配置、Frame/XSS/CSRF防护 ● 身份鉴权与数据处理 身份控制：多维认证（生物特征/令牌）、OAuth/JWT、RBAC权限控制 数据解析：JSON/XML（XPath/DOM）高危格式深度解析、防重放签名 文件存储：断点续传与安全上传/下载、SQL参数化与Redis缓存交互	Web攻击：XSS跨站脚本、CSRF跨站请求伪造。 权限与数据：越权蔓延、SQL注入、任意高危文件上传。

表1 安全场景库

五、实践效果与反思

（一）迭代收敛过程

本系统的安全场景生成能力经四轮实验迭代逐步收敛：

阶段一“直接输出”：模型能输出威胁点但结构不规范、标准不统一，质量方差不可接受。

阶段二“选择题化”：改为从预定义场景库中选择，但仅输出名称而无理由，审核人员无法判断合理性。

阶段三“选择题加生成依据”：强制附带原文依据，质量显著提升，但对隐含风险存在系统性遗漏。

阶段四“技术扩写+选择题+事实溯源”：在精确率和召回率上均达实用水平。

迭代中发现的实际情况：当模型在特定任务上表现不稳定时，优先排查的不应是模型本身，而是任务是否为模型保留了过多自由度。收缩决策空间、要求可验证依据、将复合推理拆解为递进子任务——这些方法论层面的调整，往往比模型升级更有效。

（二）落地中的现实约束

实际落地中，安全场景识别的质量很大程度上取决于需求文档的文本质量。大量系统架构信息与技术框架选型细节

并不会被写入需求文档——这些信息或存在于架构师的脑中，或分散在设计讨论记录里。需求文档中缺失的技术上下文，直接影响了场景识别的精准度，这是当前仍需持续优化的方向。

为此，在下一阶段的演进计划中引入检索增强机制，构建多层知识库以补充威胁建模的全局背景知识。具体而言，系统将整合各业务线的架构设计文档、API接口定义等技术资产，沉淀历史安全评审记录与漏洞修复案例，与既有的行业安全场景库共同构成多层次的系统知识体系。在安全场景推理阶段，系统自动检索并注入相应层次的上下文材料，从而降低对单一需求文档文本质量的依赖，打破信息孤岛，进一步提升威胁建模的准确性与业务贴合度。

（三）落地实践

本系统已在科技条线落地运行，初步覆盖重点业务系统。在安全场景识别方面，系统日均产出上百个安全设计，整体精确率达到92.61%；研发人员结合实际业务情况对安全设计进行审核采纳，采纳率50%。在代码安全评审方面，系统同步审核需求关联的代码变更，在测试环节前拦截高风险安全漏洞，有效降低修复成本与线上风险敞口。



图2 威胁建模-安全场景识别图

六、结语

威胁建模的困境，本质是安全深度与研发速度之间的长期矛盾。借助大模型，我们将威胁建模与代码安全评审无缝嵌入需求分析和代码提交流程，使安全工作从“专家事后救火”转变为“智能全程伴跑”，在不打断研发节奏的前提下提升建模质量，真正适配敏捷迭代，一定程度化解组织协同的阻力。

AI驱动的安全模式正在逐渐降低甚至消除安全与研发的壁垒，显著降低安全活动对交付效率的冲击，减少返工成本和线上风险事件；通过持续沉淀可复用的安全知识，加速安全左移的文化落地，让安全真正成为业务连续性和数字化转型的助推器，在证券期货行业强监管、高可用的要求下，构筑起值得托付的信任与核心竞争力。

在AI浪潮下 传统网络安全何去何从

张旭 | 上交所技术有限责任公司

摘要： AI正以前所未有的速度重塑网络安全的攻防格局。攻击链全面提速，而防护滞后已成客观现实。但传统安全并未过时，它既是AI赋能的根基，也是信创自主的基石。网络安全从业者当在筑牢传统防线、掌握大模型风险、稳步推进AI赋能三条路径上协同发力，于变局中守住安全防线。

关键词： AI安全、大模型、网络安全、国产化与信创

一、背景与挑战

《国务院关于深入实施“人工智能+”行动的意见》提出在软件、信息、金融、商务、法律、交通、物流、商贸等领域，推动新一代智能终端、智能体等广泛应用，为人工智能金融应用定下了基调。为贯彻落实党中央、国务院关于人工智能发展和做好金融“五篇大文章”的决策部署，行业内各类业务智能体如雨后春笋般冒了出来，但与AI+金融业务蓬勃发展形成鲜明对比的是略显滞后的安全防护。大家在欣喜于AI在业务领域引起的大幅提质增效的同时，好像只有在“小龙虾事件”发生后，才逐步想起在这巨大利好的背后往往也隐藏着巨大的风险。

对于网络安全从业者而言，当下最直观的冲击，不仅源于大模型自身携带的“原生安全缺陷”——黑盒特性带来的不可解释性与不可控性，以及提示词注入带来的新型攻击向量，更有AI赋能下，以Anthropic的Mythos模型为代表的前沿AI力量正以前所未有的速度迭代进化，而0day漏洞的发现与利用也呈现出指数级增长态势。在历年红蓝对抗中，攻击队向来秘而不宣的0day漏洞，如今仿佛不计成本般砸向防守方，这背后，AI辅助的漏洞挖掘与智能化利用工具大幅降低了攻击门槛，使得稀缺资源变得“泛滥”。金融行业素来以严密的边界防护为荣，但在海量0day的轮番冲击下，这道防线被轻易撕开裂口；进入内网后，攻击者的横向移动也不再如以前一样小心翼翼，他们深知传统“告警-研判-处置”链条远跟不上其行动节奏，故而放弃了过去那种小心翼翼、反复试探的渗透策略，转而以快打慢，将时间差转化为压倒性优势。

然而，视角转向防守方，现实却呈现出另一番光景。当前国内大部分企业在“AI+网络安全”方向的布局，仍停留在调研论证、刚刚立项或小范围实验观望的初级阶段。与攻击侧AI能力的迅猛发展形成鲜明对比的是，国央企的防护体系天然承载着更多刚性约束：一是合规审查要求技术选型必须满足等保、关保等监管框架；二是国产化替代则对硬件芯片、操作系统、数据库乃至AI框架的信创适配提出了更高要求；

三是数据不出域、模型需审计等内控红线，更进一步拉长了从技术验证到规模化落地的周期。三重枷锁之下，国央企的AI安全防护往往滞后于攻击手段数个身位，当攻击者已用Mythos类模型自动化挖掘0day时，防守方还在为采购哪款国产AI芯片、适配哪一套信创操作系统而反复权衡。防护滞后所揭示的，其实是攻击与防守两端在AI能力获取和应用速度上的客观距离，如何缩小这一差距，正是传统网络安全转型亟需回答的问题。

二、应对措施及思考建议

那么，这是否意味着传统安全已无足轻重？答案显然是否定的。人工智能在安全领域的效能发挥，从来不是空中楼阁，而是深深扎根于组织已有的安全基建体系之上：事前，需要有成熟完备的代码审计、资产测绘与威胁建模能力；事中，需具备可靠的威胁识别、实时防护、异常检测与联动响应机制；事后，则依赖CMDB等配置管理数据库实现精准的资产定位与漏洞修复全流程闭环。唯有这些“传统功架”足够扎实，AI才能在其上施展智能分析、自动化研判、快速响应的叠加优势，它不是替代传统安全，而是放大传统安全已投入的杠杆。

至于国央企在AI安全布局上的“慢”，亦当放在更宏阔的战略坐标系中审视。国产化、信创、数据不出域等刚性要求，表面看延缓了技术引入的速度，实则从供应链安全与长远发展的根本利益出发，避免关键基础设施在底层技术上被“卡脖子”，防范外部断供带来的系统性风险。这条路径固然曲折，但唯有根基稳固，方能行稳致远。短期的“慢”，换来的恰恰是长期博弈中的自主与从容。

综合来看，在AI驱动的安全新形势下，防守方可围绕以下核心层面，系统性地完善自身防护手段：

1. 查漏补缺，夯实既有安全体系的薄弱环节。

具体而言，应优先补齐流量监测等关键节点的检测覆盖盲区，确保南北向、东西向流量可视可控；其次将内网白

名单策略进一步细化落位，收缩攻击暴露面；再者加速完善CMDB配置管理数据库的资产台账，厘清业务归属与责任边界，为后续应急响应争取宝贵的“找人时间”。

最重要的是，防守方需要清醒地认识到，攻击方过去依赖人工逐台探测、缓慢摸排的侦察阶段，在AI辅助下已实现效率的跃升，攻击者可以更短的时间完成信息收集、更快地锁定突破口。面对这一变化，防守方切不可再抱持“未被发现即安全”的侥幸心理，唯有以更扎实的资产底账、更严密的访问控制，方能在AI加速的攻击链中守住防线。

2. 针对大模型自身的原生安全风险，仍需依托大模型技术本身来实现“以模治模”。

大模型的黑盒特性、不可解释性以及提示词注入等新型攻击面，已超出了传统WAF、防火墙等基于规则匹配的防护能力边界。这些“老办法”面对语义级别的攻击几乎无能为力。唯有通过大模型自身的语义理解与上下文感知能力，对输入输出进行动态风险研判，对异常交互行为进行实时识别与阻断，才有可能在模型内部构筑起有效的防御屏障。这意味着安全团队需要在大模型的全生命周期中嵌入安全检测能力，从训练数据投毒防范、模型行为监控到推理过程的内容安全过滤，形成一套由AI驱动的自适应防护机制。“以模治模”，既是技术路径的必然选择，也是当前应对大模型原生风险最为现实且落地可行的方法论。

3. AI最擅长的正是“与人打交道”，社会工程学领域可成为AI赋能安全的优先突破口。

以钓鱼邮件检测为例，通用大模型凭借其天然的语义理解与上下文关联能力，在识别恶意话术套路、捕捉语气中的紧迫感与异常情绪、研判隐藏链接与附件的潜在风险等方面，表现已显著优于传统基于规则和黑名单的过滤引擎。更为关键的是，通用大模型无需安全垂直领域的专项训练即可快速上手部署，对于现阶段仅拥有通用大模型能力、尚不具备安全定制研发条件的企业而言，这是投入最小、见效最快的务实选择，不必等待漫长的模型微调周期，不必承担高昂的定制开发成本，即可在攻击者最依赖“人性弱点”的入口处构筑起一道智能防线。从钓鱼邮件起步，逐步向短信诈骗、IM钓鱼、敏感词检测和异常登录识别等更广泛的社工场景延伸，这条路径既是对有限资源的合理配置，也是企业在AI安全转型中迈出的扎实第一步。

4. 针对当前零日漏洞频现，智能审计须前置。

以Mythos为代表的AI模型，在数周内发现的零日漏洞数量已接近全球此前全年漏洞发现总量的三分之一，攻防节奏由此被彻底改写。面对这一变化，企业应在事前防护环节积极引入AI能力，将核心代码、开源组件及第三方依赖库优先交由通用大模型或安全垂类大模型进行自动化检测。相比传统静态代码分析工具依赖人工编写规则库、更新滞后的局限，AI模型能够从语义层面理解代码逻辑，发现未知逻辑缺陷与隐蔽后门，帮助企业提前锁定潜在0day风险，为应急响应争取宝贵的预处置时间窗口。将AI代码审计作为日常研发安全流程的固定环节，是当前成本可控、回报明确的防御升级手段。

5. 事中检测环节的智能化升级。

当前企业多采用边界防火墙加主干道HIDS的组合进行运行时防护，但其核心引擎仍依赖规则匹配，这一范式正面临越来越严峻的挑战：规则设置若过于严苛，攻击者可通过大包拆分、编码混淆、变形绕过等手段轻易规避；规则设置若过于宽松，则海量告警蜂拥而至，将安全运营人员淹没在误报洪流之中，待真正攻击被识别时，往往已错过最佳反击时间。部分大型机构先行一步，通过态势感知平台对告警进行聚合降噪，在一定程度上缓解了运营压力，但终究治标不治本，规则匹配的逻辑天花板仍在，被绕过的风险从未消失。

破局之道在于引入AI驱动的检测能力。大模型凭借强大的语义分析、多模态关联与自适应学习能力，能够从流量和日志中识别出隐藏的恶意图，而非仅依赖固定特征签名，这意味着对0day、Nday、Web变种、编码混淆等高对抗性威胁可实现高检出、低漏报，同时显著降低误报带来的人工复核负担。需要指出的是，AI赋能事中检测并非低成本解决方案，在高并发流量下，它对算力资源、模型运维能力和安全运营团队的人员储备均有较高要求，现阶段更适合资源充裕的头部企业率先探索，形成标杆后再向中小型机构辐射。

6. 事后响应环节的自动化探索。

AI可通过关联CMDB资产信息实现受影响资产的自动定位与隔离建议，辅助决策，亦可在大规模攻击场景下自动触发IP封禁等操作，大幅缩短响应时间。但此类自动化手段涉及生产运行安全，任何误判都可能导致业务中断或合法流量被误杀。而大模型自身“不完全可控”的特性，如在测试环境中偶发的“偷懒”行为、指令理解偏差、推理不稳定的“不听话”表现，一旦在生产环境中复现，其后果远超实验室场景的容忍范围。因此，事后环节的AI自动化宜以“辅助决策、人工确认”为底线原则，从非核心业务、低风险场景入手逐步验证，待模型稳定性与可控性得到充分检验后，再审慎地向关键业务延伸，切不可因追求效率而牺牲安全的基本盘。

7. 知识沉淀，构建可持续进化的安全智能底座。

智能安全防护所需的增强能力，不能仅靠大模型的通识能力，而必须基于在实战中积累的各类攻防数据，对大模型进行持续训练与微调，逐步构建面向安全垂直领域的专有模型。不同防守方因业务形态、组织架构、网络拓扑的差异，其所积累的攻防数据天然带有鲜明的个性化特征，这不仅是日常运维的副产品，更是后续模型微调与能力定制的宝贵语料。展望未来，当AI赋能成为行业标配之后，真正的竞争壁垒将不再是模型本身，而是各自手中掌握的知识沉淀深度与数据治理质量。谁的数据更干净、更全面、更具实战代表性，谁就将在下一阶段的智能安全博弈中占据先机。

三、总结与展望

大模型技术仍在高速演进，各行各业争相借势腾飞。在此背景下，网络安全从业者提出安全风险的声音，似乎总显得有些“不合时宜”。然而，事实果真如此吗？任由AI与业务的融合疾驰向前，却忽视对AI自身安全性的审视与加固，无

异于放任一辆高速列车在未知轨道上狂奔，一旦脱轨，后果将是系统性的。那么，传统安全防护能否有效应对AI加持下的新型攻击？答案显然是否定的。但传统安全是否就此被AI浪潮冲刷得无足轻重？答案同样是否定的。

越是技术剧变的时刻，越需要清醒的守夜人。对网络安全从业者而言，这波AI浪潮既是一次严峻考验，更是一次价值重估的契机。当务之急有三：其一，坚守好传统网络安全的基本功，那是所有上层能力的根基；其二，主动学习了解大模型的运行机制与原生安全风险，不回避、不盲从，方能有针对性地布防；其三，积极拥抱新技术，以“小步快跑、稳妥迭代”的方式，将AI能力稳步引入传统安全防护体系中，实现真正的提质增效，既不自缚手脚，也不盲目追随。以上种种，未必周全，权当是一名网络安全从业者面对AI浪潮时的一点尝试与思考。

参考文献

- [1] 国家工信安全中心. 政务大模型应用安全规范[S]. 2025-09-15.
- [2] 国家互联网信息办公室. 人工智能安全治理框架（2.0版）[R]. 2025-09-15
- [3] 刘霞. AI新模型拉响网络安全攻防警报[N]. 科技日报, 2026-04-20
- [4] 央广网. AI新模型拉响网络安全攻防警报[EB/OL]. (2026-04-20). [https://tech.cnr.cn.\[reference:8\]](https://tech.cnr.cn.[reference:8])
- [5] Anthropic Red Team. Mythos Preview System Card & Red Team Report[R]. Anthropic, 2026

03 思考和观点

P83 软件数字孪生与持续“演化—可信”平衡——五年百例实证研究
张千里

**P90 向量切片级细粒度管控：证券投研知识库在RAG（检索增强生成）架构下的
非结构化数据安全隔离研究**
茆正华

P95 从RSAC看国内网安AI发展方向
绿盟科技集团股份有限公司-大模型安全研究团队

软件数字孪生与持续“演化—可信”平衡

——五年百例实证研究

张千里 | 上海金融期货信息技术有限公司

摘要： 软件加速演化与可信要求同步提升，“演化—可信”成为软件产业的核心张力。主流DevSecOps范式以过程合规代理产品可信，在标准化场景中高效可靠，但随软件复杂性与情境差异增加，适用面收窄。本研究历时五年深入一家金融IT链主企业，积累165个现有工具链缺乏有效自动化手段的案例，其共性是样本、特征与关系三类遮蔽所致的认知缺口。据此提出软件数字孪生——以“逆向工程+再工程”持续重建软件整体状态及其演化过程，在过程代理之外构成面向产品本体、与DevSecOps互补的第二条可信认知路径。

关键词： 软件数字孪生、演化—可信、DevSecOps、逆向工程、再工程、软件可信、认知路径

一、前言

我国数字经济规模已占GDP的43.8%^[1]，软件已成为社会运行的数字神经中枢。一方面，数智化推动软件应用频谱拓展、业务耦合加深，演化更难驾驭^[2]；另一方面，科技竞争、新质生产力建设与供应链自主要求，使可信压力同步增强^[3]。“演化—可信”二者并行，张力日益凸显。

这一张力并非外部态势的一时产物，而根植于软件自身两条规律：Brooks指出的本质复杂性源于现实世界概念结构的映射^[4]，软件越贴近现实，情境差异越显著，其复杂性也随之提升；莱斯定理则表明软件语义不存在通用而完备的判定^[5]，可信无法一劳永逸地获得，只能针对具体软件随场景逐案重建。如何让演化不损可信、可信不阻演化，正是本研究的出发点。

DevSecOps将安全与质量活动左移、嵌入持续集成与持续交付流水线，以自动化管控在高速迭代中维系可信，已成为当前主流范式（参考架构见图1）^[6]。其可信保障本质上沿袭制造业“过程受控即产品合格”的逻辑——以过程合规代理产品可信。这一代理在产品高度标准化时高效可靠；但软件须不断迁就异质多变的现实，随着长尾与个性化场景增多，标准化前提与代理关系的适用面将收窄。

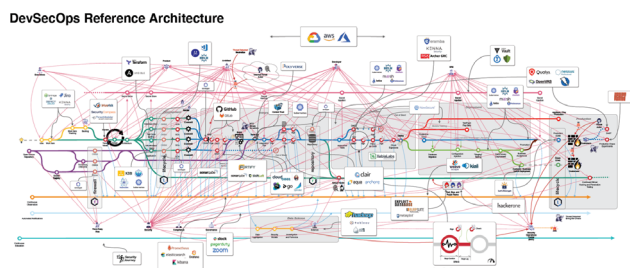


图1 DevSecOps参考架构

这一收窄发生在主流范式覆盖不到的长尾场景，而要看清长尾，唯有下沉到真实现场取样。本研究历时五年深入一家金融IT链主企业，积累了165个现有DevSecOps平台内缺乏有效自动化手段的真实案例。它们揭示的不是工具的缺位，而是对产品本体直接认知的缺位——指向一条与过程代理并列的可信认知路径。本研究据此提出软件数字孪生：如同数字孪生为物理实体构建可计算镜像以供观测与干预^[7]，它为软件持续重建一份涵盖逻辑、组成、调用与环境的本体映射，与DevSecOps并行互补。

二、实证研究的背景与方法

本研究的实证环境，是一家正处数智化转型中的金融IT链主企业，“演化—可信”张力充分：其一，演化强度高，研发约500人、已建成覆盖全生命周期的领先DevSecOps平台；其二，可信压力真，金融IT面临强监管、高可靠与供应链自主的多重要求，失效分析、深度溯源、问题归零是硬约束；其三，角色最完整，它兼具软件生产者与消费者双重身份，令矛盾以丰富场景呈现。

本研究为探索性专项课题，历时五年，以“打破惯性、突破边界、创造价值”为导向，不预设主流范式边界之外的解应是什么，而从一线真实诉求逐个攻关。取样框只纳入一类——现有DevSecOps体系内无成熟工具、流程或自动化手段可直接满足的诉求，已积累165例。它们提供的是一类存在性证据：在过程代理高效覆盖的标准化场景之外，确有一类持续复现、且需直接面向产品本体方能回应的可信诉求。

分析的基本单元是“案例”，即一次具体诉求及其完整求解过程，每例记录触发情境、所涉形态、传统手段受阻之处及创新解法与结果。本研究先从多维度对案例归类标注，再以宏观、中观、微观三层读出规律。

三、实证研究

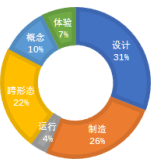
本章按宏观统计、中观分类、微观验证三层展开：先以统计确认DevSecOps边界之外的可信需求真实存在、形态多样，再剖析这类需求受阻的内在机理、定位核心认知缺口，最后以典型案例验证该缺口能否被填补。

（一）宏观统计分析

1. 形态分布：跨形态耦合的新特征

设计形态占比最高（31%），反映一线对代码、配置、声明等非结构化文件的高频内容分析需求；制造形态占26%，反映各类制品的内容分析与流转核验需求上升；运行形态仅4%，因运行时数据含敏感信息、管理权集中于运维团队（图2）。

案例的软件形态分布



案例的软件形态分布 - 按季度

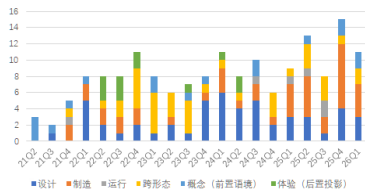


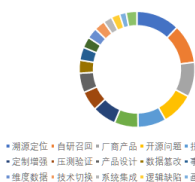
图2 案例的形态分布

跨形态分析占22%：云原生、开源与DevSecOps的普及使同一软件以源码、制品、容器、运行实例多重面貌并存，影响分析、溯源、漏洞利用条件判断普遍需跨形态关联，而现有工具多按单一形态划分职责，这类能力恰落在工具职责的缝隙之间。

2. 需求分布：场景长尾化与主体多元化

可信需求沿两条维度同步分散（图3）。场景维度高度分散：需求散落于20个应用场景类别，异常行为溯源、自研组件召回、厂商产品二开等居前，各类频次相当，无一形成压倒性主流。主体维度多元：需求方涉及22个角色/岗位，覆盖生产线、架构师、安全、开发、运维、测试、合规、厂商、社区等全生态，以生产一线为主导。

案例的场景分布



案例的需求分布

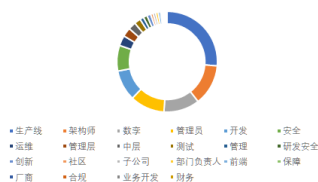


图3 案例的需求分布

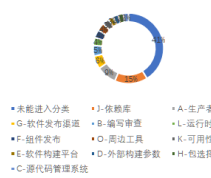
两条分散同源于可信责任的下放——数字化转型与敏捷实践使生产、消费与可信责任不断下沉至各业务单元，可信需求由此来源多元、场景丰富，构成一条有待新能力承接的可信“长尾”。

3. 参考框架：标准收敛，场景发散

以SLSA（锚定代码与构建完整性^[8]）和S2C2F（规范开源组件消费^[9]）对照本研究案例，41%落在SLSA覆盖之外、77%落在S2C2F之外（图4），溢出集中于两框架设计目标未

及之处——运行态、非恶意缺陷，以及构建链之外的人员与环境因素。

案例的SLSA威胁分类



案例的S2C2F实践分类

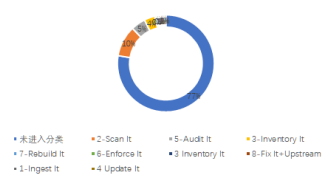


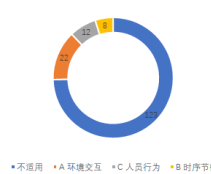
图4 案例的供应链分类分布

标准的本分是收敛可共识化的部分，其扩张节奏自然落在一线即时生长的诉求之后（如SLSA迟至2025年底v1.2才新增源码模块）。

4. 环境影响：多因耦合交互

165例中有42例（25.5%）由软件自身与环境、时序、人员的交互决定，呈梯度分布（图5）：环境交互22例（52.4%）最高频，同版本号异内容、迁移故障反复出现；人员行为12例（28.6%）次之，以画像、基线与身份溯源把“人”纳入刻画；时序节奏8例（19.0%），提供政策、舆情、假期、十年趋势等节律性扰动证据。这42例共同表明，环境、时序、人员并非可忽略的背景噪声。

案例的环境暴露数量分布



案例的环境暴露分布比例

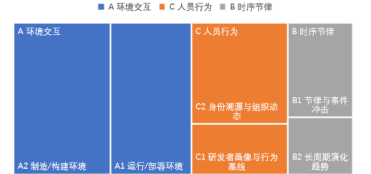
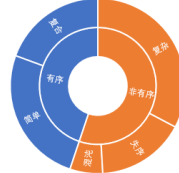


图5 案例的环境暴露分布

5. 复杂性跃迁：规则方法的适用边界

以Cynefin框架^[10]对165例分类（图6）：简单域与复杂域合计45%，因果可事前认知，传统规则化方法依旧有效；复杂域、混沌域与失序域合计55%，因果只能事后回溯，预设规则的适用性随之下降，三类非有序域详见下文中观分类剖析。

案例的CYNEFIN二级分布



案例的cynefin类别分布

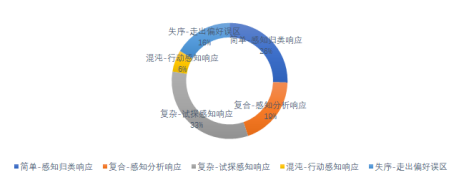


图6 案例的复杂性分布

但有序与非有序的边界并非固定：某类非有序场景一旦被新方法攻克、沉淀为可复用能力，其因果便从事后回溯转为事前可知，重新被规则化方法接管。

6. 小结

六维数据共同指向同一结构性事实：现有DevSecOps体系所依赖的“形态清晰、需求集中、责任专业、标准覆盖、环境可控、行为可规则化”诸前提，已被大量真实诉求所突破，其边界之外的可信需求真实存在且形态多样。

（二）中观分类剖析

本节转入中观：以分类为尺度探索内在机理，提炼一个可填补的核心缺口。

1. 三类决策域：从“规则—执行”到“感知—响应”

55%非有序案例的处置共性,可按Cynefin决策域分类继续细化^[10]。

复杂域（探测—感知—响应）：自研组件召回、开源问题排查、溯源定位等（图7），无法套用预设方案，只能在试探中拓展新做法；因果只能事后回溯，处置本身必须是反复探测的回路。

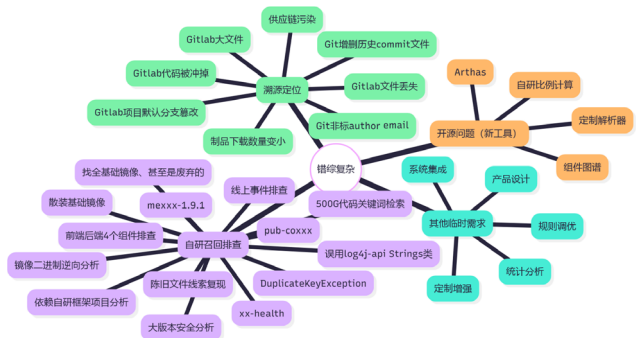


图7 案例的复杂域场景

混沌域（行动—感知—响应）：灾难性中断、超危排查、数据篡改监测等紧急情况（图8），风险最高，现场常在处置中被进一步破坏；无既定规则可循，唯一能依赖的是对产品状态的直接观测与实时决策。

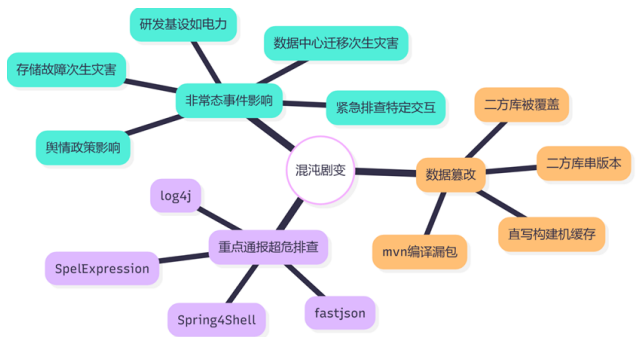


图8 案例的混沌域场景

失序域 (多视角融合): 各角色受限于自身视角, 易将局部当全景 (图9); 它反映的是不同角色认知差异的客观存在, 需要一种融合技术把分散碎片拼成服务所有角色的完整图景。

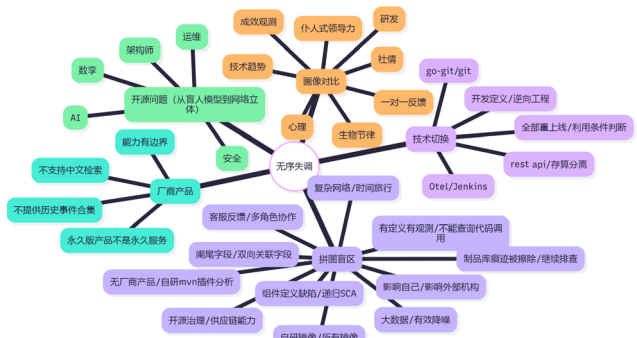


图9 案例的失序域场景

三域并置可提炼出本节最重要的观察：三类域所要求的处置，无一例外是“感知—响应”回路而非“规则—执行”静态方案——感知由此上升为可信保障的核心动作。

2. 三类遮蔽：感知为何失败，认知断层如何形成

从案例库反复出现的受阻场景出发,按认知链条中断的位置归纳,问题集中于三个连续环节——对象是否可获得、是否可理解、信息能否关联——对应形成样本遮蔽、特征遮蔽与关系遮蔽(图10、表1)。

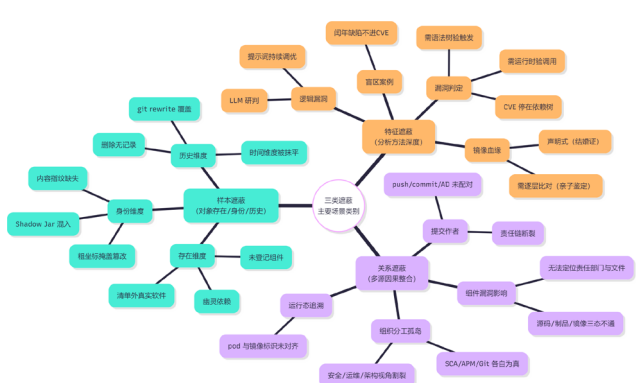


图10 三类遮蔽归纳

三类遮蔽虽表现不同,却同源于软件持续演化与可信要求同步提升所带来的认知压力:当规模、异构程度与变化速度增加,“采不全、读不懂、拼不合”现象增多,催生出全面采集、动态分析与多源整合三类新能力需求。它们已超出流程记录、责任流转与协同治理的范畴,转而直接面向软件产品本体的持续认知。下节结合典型案例验证这些能力能否有效缓解三类遮蔽。

遮蔽类型	典型表现	认知断层	所需能力	认知深度
样本遮蔽	对象采不到、采不全	不知道它是否存在	全面采集	在场
特征遮蔽	对象虽在场但难解析	不知道它是什么	动态分析	内容
关系遮蔽	单点可见但无法关联	不知道意味着什么	多源整合	关联

表1 三类遮蔽及其能力需求

（三）微观案例验证

本节四个案例，以逆向工程还原事实、以再工程重建认知，破解三类遮蔽。每例还隐含一道高门槛——认知断层长期悬置，是填补所依赖的算力与数据近年才成熟可用（技术细节见专利^{[11][12][13][14]}）。

1. 设计形态：漏洞利用的精准排查

内部定制框架滞后社区一代，漏洞爆发后组件依赖分析显示几乎所有线上版本均在影响范围内，整体升级风险过高，必须逐函数判定是否真可被利用。案例以逆向工程从Git代码库全量提取所有提交源码、用ANTLR解析为函数级语法树，直接从本体提取可验证特征，再以利用条件作大数据特征匹配，实现影响范围的全样本、无侵入、精准的确认（图11），语法树分析解决特征遮蔽，全样本采集同时解决样本遮蔽。

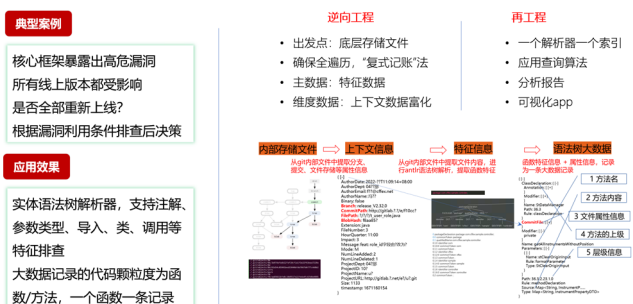


图11 漏洞利用分析的典型案例

其门槛在算力: 对约500GB、2400个代码库、所有历史提交、所有文件的全量遍历、内容的语法树解析、匹配特征代码行提取、数据富化, 需平均800核CPU持续计算5小时, 约相当于一次全球数值天气预报单次算力的一半^[15], 是必须以云计算为地基的场景。

2. 设计形态: 逻辑漏洞排查

另有一类缺陷不违语法、不触规则, 而隐于业务逻辑与上下文——如分页不当导致业务错误、以时间戳作文件名导致并发覆盖、闰年处理错误等, 需跨文件、跨流程乃至跨业务语义研判, 恰是大模型所擅长。

在时间戳文件名案例中, 先据time、now等特征初筛约4万个涉时源文件, 再以提示词作“有缺陷/无缺陷/无法判定”三分类, 经人工复核持续补充正反样本、修正条件, 迭代成闭环; 七轮后疑似并发缺陷占比由约20%收敛至1.7%, 可直接支撑更精准的修复。

大模型显著降低了破解特征遮蔽的门槛, 但其奏效依赖持续的样本积累与人工验证, 是不断精进的过程而非一次到位。其门槛在token与迭代: 单轮分析即耗约5亿token。

3. 制造形态: 容器镜像的祖缘分析

容器镜像作为不可变基础设施继承链复杂, 基础镜像中的漏洞可沿派生关系快速扩散, 而传统治理通常止于Dockerfile的登记式声明, 其作用更接近“出生证”而非“亲子鉴定”, 难以反映镜像实际内容。本研究从镜像注册中心提取目录与文件属性, 依据镜像构建规则逆向重建完整祖缘网络, 破解关系遮蔽。

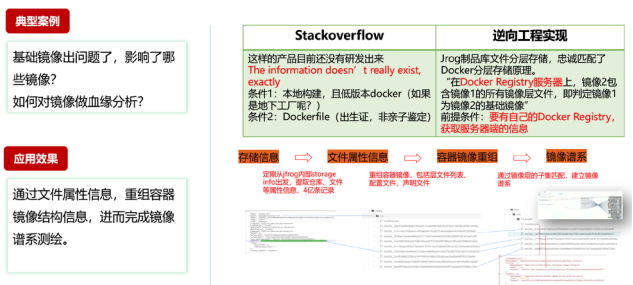


图12 镜像祖缘分析的典型案例

其门槛在于长期的数据积累与中心化计算: 每日两次采集制品库属性, 5年累计数据约1.2TB, 历史文件约200万个; 祖缘计算涉及万亿级关联比对。

4. 多个形态: 时区错误的全链追溯

本案例源于云环境中容器镜像时区配置错误的排查。初

步通过某技术栈的基础镜像祖缘分析, 仅发现局部影响; 进一步结合镜像祖缘分析、全制品库镜像溯源、云上运行时全量检测以及非线上容器运行时检测, 最终确认影响范围较初步评估扩大31倍 (图13)。

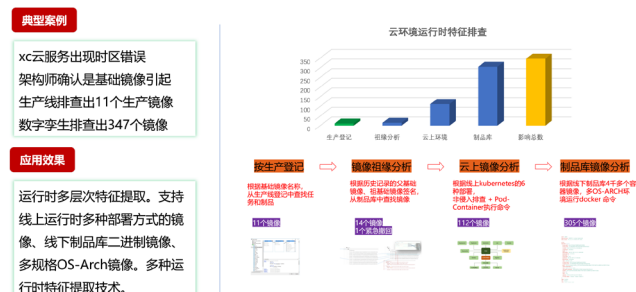


图13 云环境时区设置排查的典型案例

该案例表明, 单一形态分析往往只能获得局部认知, 而跨设计、制造与运行等多个形态的联合分析, 能恢复完整事实链条, 实现全链追溯与影响评估。

四、软件可信的两条认知路径

第3章表明, 即便在覆盖全生命周期的先进工具链之上, 认知断层仍会在边界之外的场景中反复出现。这说明问题并非出在某项工具或某类技术, 而在于其共同遵循的治理逻辑——本章即检视这一逻辑的适用边界。

当前主流软件治理范式承袭自制造业质量管理思想, 其基本逻辑是以过程合规代理产品可信。该逻辑能够成立, 并非因为过程天然可靠, 而是因为制造业产品高度标准化: 当个体差异足够小, 过程状态便能较稳定地反映产品状态, “做对过程”通常意味着“做好产品”。即便如此, 制造业仍保留抽检、计量和探伤等直接测量手段, 作为对过程控制的验证与补充。

软件则不同。软件产品天然具有个体性, 其真实状态不仅取决于本身, 还受到运行环境、时间演化和人员操作等因素影响。在标准化程度较高的场景中, 过程控制依然有效; 但随着场景复杂度和个体差异增加, 过程与产品之间的相关性会逐渐减弱。第3章所呈现的案例, 正是这种相关性收窄后的具体表现。在过程代理难以充分建立可信认知的场景中, 需要补充对软件本体的直接观测, 以恢复对产品真实状态的认识。

由此, 软件可信呈现出两条不同但互补的认知路径: 过程代理通过约束软件形成过程推断产品状态, 本体观测则通过持续测量软件本体认识产品状态。前者依赖规范进行演绎担保, 后者依赖观测进行归纳验证; 二者并非替代关系, 而是分别从不同维度建立可信认知。第3章揭示了过程代理的边界, 也显示出本体观测的必要性。

但其价值并不止于补足边界: 它是一条从产品本体出发、可独立成立并持续生长的认知路径, 如何把它建成支撑软件可信的新基础, 正是下一章的主题。

五、建构软件数字孪生

第3、4章已表明可信认知断层客观存在，需要一种面向软件本体的直接分析能力。本章把五年百例中反复奏效的探索提炼为可复用范式：先给出定义及其让“演化—可信”得以并行的机理，再落为三维参考框架，继而辨明其定位、独立性与落地前提。

（一）定义与目标

与物理产品不同，软件没有具体实体形态：实体产品尚可盘点、丈量、抽检乃至数字孪生，软件却更难建立清单、可视化分析与精准干预。

软件数字孪生，是一种以软件整体状态及其演化过程为研究对象、持续生长的软件可信认知体系。鉴于软件缺乏可供直接丈量的实体形态，它以“逆向工程+再工程”为双引擎：逆向工程从代码库、制品库、运行环境等真相源全量提取样本、特征与上下文，重建软件本体的映射；再工程在该映射上开展跨形态关联与因果推理，面向具体场景产出价值。二者构成回路——新场景一旦超出孪生体的解答能力，所暴露的缺口即转化为新的采集与特征需求，驱动逆向工程补全，再工程随之给出更深、更完整的回答。它不通过软件形成过程推断产品状态，而是直接观测软件实际呈现的状态。旁路解耦于交付管线之外，使可信不阻演化；全样本感知软件的实然样貌，使演化无损可信。受莱斯定理约束，它不以一次性完备判定为目标，只承诺可追溯、可复现、可校准、可生长，在过程代理之外构成直接面向产品本体的第二条认知路径。

（二）三维参考框架：形态广度 × 特征精度 × 关联应用

归纳五年百例可见，尽管场景各异，其分析过程均遵循相似路径：依问题场景选定一种软件形态，全量采集掌握概貌；分析其特征以获得精确刻画；当单一形态不足以回答问题时，增补新的软件形态并研发相应特征提取手段；待多形态获得充分刻画后，再通过特征关联支撑场景满足。这一“形态→特征→关联”过程在百例中反复出现，抽象后凝为三个维度，形成软件数字孪生的三维参考框架（图14）。

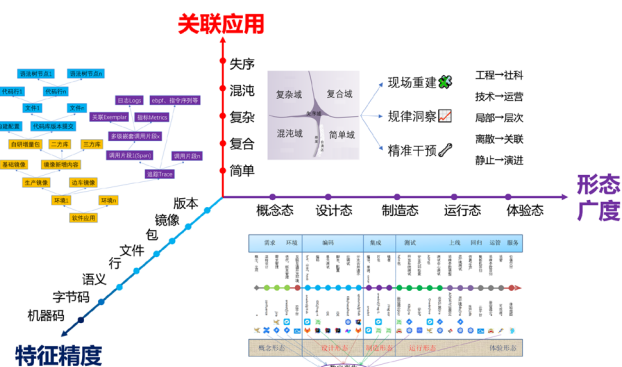


图14 软件数字孪生三维参考框架

形态广度（认识什么·覆盖能力）。回答认识的疆域：把哪些形态纳入视野，并为每种形态找到核心真相源、建起全样本采集渠道。广度只问“看多少种形态、能否够着全量”。

特征精度（认识多深·检测能力）。回答单一形态内部认识的深浅，沿阶梯递进、逐步突破样本、特征与关系三类遮蔽：样本元数据层仅采集坐标与信号，实现全量粗粒度覆盖；内容特征层读取对象内容并提取深度特征；上下文关系层将项目、人员、组织、变更、时间等语境持续叠加富化，使多层特征嵌套、跨形态对象可以索引关联，并为后续连接预留统一坐标。

关联应用（如何关联·用于何处·跨形态因果）。回答各形态达到相应深度后如何跨形态连接并服务场景。其底座是数据编织，即将不同形态的事实依据统一坐标组织为可推理整体；在此之上形成三类能力：现场重建——还原“软件本体×环境×时序×人员”在特定时刻的真实情境，回答“发生了什么、为何如此”，并满足取证科学对客观性、关联性与合法性的要求；规律洞察——跨时间与群体提炼趋势、画像与成效；精准干预——基于跨形态因果，为具体场景提供个性化精准处置方案。

沿三维展开五种软件形态，即得其一次具体展开（表2）。表中形态类型、真相源、特征示例与应用产出仅为本研究对象下的代表性实例，可随研究对象增删扩展，而三维结构本身保持不变。

软件形态	形态广度	特征精度·单形态内 (样本 → 特征 → 关系)	关联应用·跨形态 (重建·洞察·干预)
设计态	Git库；配置中心；Webhook	全量全史检索 → 语法树(CST) 利用条件 → 变更—责任链、变更影响图	重建：还原某次提交、还原代码、分支删除； 洞察：十年技术栈迁移、代码热力图等演化规律。 干预：跨态漏洞验证与应对
制品态	制品库／二三方库／镜像库；编译环境；Webhook	制品多坐标台账、事件流 → 二进制成分还原 → 血缘谱网	重建：制品删除·覆盖·投毒定位。 洞察：活跃老化时序规律、冗余分布。 干预：制品成分交叉核验
运行态	OTel；Arthas；k8s／容器	信号采集 → 运行剖析 → 调用拓扑、span关系网	重建：某次运行异常时的实际调用与部署。 洞察：负载节律、性能趋势与多模式降级规律。 干预：技术异常的业务语义归因
概念态	Jira等工作流；Wiki等知识库	流量、流向统计 → 事件、版本、关键字 → 过程成流（桑基）、跨对象规则	重建：需求／缺陷的完整流转轨迹。 洞察：标准化模板渗透成效 干预：自动判断合规差距
体验态	用户旅程／OTel时间旅行；研发画像（git节律）；反思报告；外部事件	行为事件序列 → 生物节律、压测验证 → 个体旅程、画像组合	重建：以时间旅行重现用户旅程。 洞察：生物节律、个人画像、群体规律。 干预：效能—价值叙事

表2 参考框架实例化示例表

（三）定位：软件可信的第二条认知路径

软件数字孪生正是第4章所述本体观测路径的系统化实现。多年实践中，关于它的定位反复遇到三个“是不是”的追问，对应三个层面：它是不是DevSecOps之下的一件工具？是不是既有技术的改名集成？是不是一件终将完备的通用工具？三问的否定回答，正构成它作为第二条认知路径的定位。

1. 本体观测：与过程代理并行互补

两条路径关注对象不同：过程代理关注软件如何被生产，强调流程、规范与责任；本体观测关注软件最终呈现为何种状态，强调事实、测量与证据。落到软件可信，二者可凝为一对互补的治理模式——过程代理以信息化向上集成、流程驱动，以过程合规预防性维持普遍秩序，是规则之治；软件数字孪生向下挖掘、数据驱动，以对软件本体的直接取证查清实然，是数据之证。把这一分工落到每一环节，二者各成一套完整闭环（表3），是功能分工、互为交叉验证，而非高下之别或路线竞争。

环节/层面	DevSecOps：规则之治 (向上集成、流程驱动)	软件数字孪生：数据之证 (向下挖掘、数据驱动)
数据获取	各阶段工具上报声明与流程记录	非侵入向下挖掘，从Git/Registry/K8s存储层还原原始事实
坐标定位	按流程阶段与责任主体登记	为各形态资产赋全局唯一坐标，跨形态可追溯
特征认识	按预设规则、通用工具扫描	自研引擎下探至语法/指令级，提取长尾特征
验证衔接	串联控制、按流程流转，合规门禁达标即通过	网络化关联，逆向与再工程旁路独立重建
计算载体	阶段性单点工具、服务器插件	云计算、大数据、大模型
认知重心	处理可标准化、可规则化的有序场景，自动管控	应对复杂、混沌、失序场景的认知缺口
可信职能	据事实完成责任归属与流程复盘（管理闭环）	供给全量影响分析、精准根因与证据固化（技术归零）

表3 规则之治与数据之证：分工互补

2. 独立对象：软件整体状态的持续重建

一个新方向能否独立成立，不取决于采用哪些技术，而取决于是否拥有自己的研究对象。软件数字孪生的研究对象是软件整体状态及其演化过程；在这一框架下，逆向工程、程序分析、可观测性、数字取证、大模型等不再是彼此独立的终点，而成为认识软件整体状态的测量手段——独立性来源于统一对象与统一坐标系所产生的整体涌现，而非技术的多寡。

这一独立性可由一道可证伪判据检验：取任一既有领域单独运作，能否围绕软件建立双向、跨形态、可复现的整体状态认知——既回答“同一软件涉及哪些相关因素”，又回答“同一场景影响哪些软件对象”？程序分析以单形态结构为对象、可观测性以运行态信号为对象、数字取证以事后单案为对象，皆未以“软件整体状态”为标的，故单独运作时该判据均不成立。

唯有当局部观测被编排进统一对象与统一坐标系，整体才涌现出局部所不具备的双向追溯、关联推断与场景重建能力——它区别于“再加几个工具”的根本即在于此。

3. 持续生长：认知体系而非一次性工具

最后一问：它是不是一件终将完备、可一次配齐的通用工具？答案是否定的。第3章已表明可信场景显著长尾，不存在一次覆盖全部需求的通用分析器；这一不完备并非工程上的暂时欠缺，而有原理根据——受莱斯定理约束（见第1章）。

这也界定了它的成长方式：在研发与服务的反馈中持续校准。云计算、大数据与大模型等测量手段的近年成熟，使这一持续观测得以从应急之举转为可常态维持的能力，这正是“为什么是现在”的根由。其价值不在一次性回答所有问题，而在建立一个能伴随软件演化、不断积累事实依据并扩展认知边界的认知基础。

（四）规模化成立的条件

软件数字孪生能否从单点实践走向规模化成立，取决于四项现实条件：持续获取代码库、制品库、运行环境等真相源并重建可信事实的数据基础，支撑全量逆向与跨形态关联的计算基础，在统一坐标下汇聚多角色、多生命周期事实的组织基础，以及在研发与服务反馈中持续校准、生长的演化机制。五年百例表明，这些条件已随云计算、大数据、大模型及数智化转型的发展逐步具备。

但软件数字孪生并非必须一步到位的大型工程。不同规模的组织均可从样本层覆盖起步，随着数据、计算与组织能力积累，逐步深入内容特征层与上下文关系层。其本质是将软件实际呈现的状态持续数字化，重建为可计算、可关联、可推理的事实底座，是比信息化更基础的工作^[16]，过程治理、知识沉淀与智能分析，皆可在此基础上展开。

六、总结与展望

本文针对软件产业长期存在的“演化—可信”核心张力，基于一家金融IT链主企业五年165个案例，作出三方面贡献。其一，揭示存在性事实：在DevSecOps高效覆盖的标准化场景之外，仍有一类需直接面向软件本体方能回应的可信需求，其受阻源于样本、特征、关系三类遮蔽。其二，提出软件数字孪生——以逆向工程与再工程持续重建软件整体状态及其演化过程，在“规则之治”之外增加“数据之证”，构成软件可信的第二条认知路径。其三，将该路径凝练为“形态广度×特征精度×关联应用”三维参考框架，支撑现场重建、规律洞察与精准干预三类能力。

本研究为存在性证据，结论源于单一链主企业，但具前瞻参考价值。随着云计算、大数据与大模型成熟，软件数字孪生有望从个案实践成长为软件可信的重要基础设施。为此，诚邀同仁共同探索、积累案例、共享知识，推动持续的“演化—可信”平衡。

参考文献

- [1] 中国信息通信研究院. 中国数字经济发展研究报告 (2025 年) [EB/OL]. 北京, 2026-03-26. <https://www.caict.ac.cn/kxyj/qwfb/bps/202603/P020260323736507370748.pdf>.
- [2] 彭鑫. 大模型时代的软件智能化开发: 我们在哪里? 该往何处走? [EB/OL]. 2024-10-25. <https://mp.weixin.qq.com/s/JZMhWoB1rrOHbfJFGYNyTQ>.
- [3] 中国软件行业协会. 2025中国信创软件高质量发展报告 [R]. 2025-10-24. <https://www.csia.org.cn/content/6351.html>.
- [4] BROOKS F P, Jr. No silver bullet: essence and accidents of software engineering[J]. IEEE Computer, 1987, 20(4): 10-19. DOI:10.1109/MC.1987.1663532.
- [5] RICE H G. Classes of recursively enumerable sets and their decision problems[J]. Trans. Amer. Math. Soc., 1953, 74(2): 358-366. DOI:10.1090/S0002-9947-1953-0053041-6.
- [6] SCHLEEN D J. DevSecOps Reference Architecture[EB/OL]. 2020-01-16. <https://www.sonatype.com/hubfs/DevSecOps%20Reference%20Architecture.pdf>.
- [7] GRIEVES M, VICKERS J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems[M/OL]// KAHLEN F J, FLUMERFELT S, ALVES A, eds. Transdisciplinary Perspectives on Complex Systems. Cham: Springer, 2017. https://doi.org/10.1007/978-3-319-38756-7_4.
- [8] OpenSSF. SLSA specification: Version 1.2[EB/OL]. 2025-11-25. <https://slsa.dev/spec/v1.2/>.
- [9] OpenSSF. Secure Supply Chain Consumption Framework (S2C2F)[EB/OL]. 2022-10-19. <https://github.com/ossf/s2c2f>.
- [10] KURTZ C F, SNOWDEN D J. The new dynamics of strategy: Sense-making in a complex and complicated world[J/OL]. IBM Systems Journal, 2003, 42(3): 462-483. <https://doi.org/10.1147/sj.423.0462>.
- [11] 张千里, 刘智勇, 谢恒. 基于数字孪生的软件供应链分析及构建方法: 202211428158.X[P]. 2022-11-14.
- [12] 张千里, 刘智勇, 应大卫. 一种基于大数据和云计算的容器镜像血缘智能分析系统: 202311717603.9[P]. 2023-12-14.
- [13] 张千里, 金越成, 刘智勇, 应大卫. 一种软件组件依赖关系图谱智能绘制系统: 202311719133.X[P]. 2023-12-15.
- [14] 张千里, 刘智勇, 付煜哲. 一种基于业务语境增强的可观测系统: 202411855100.2[P]. 2024-12-17.
- [15] 肖文名. 筑牢气象数字底座 助力数字气象发展[EB/OL]. CIO时代网, 2024-08-24. https://mp.weixin.qq.com/s/sp_CwrxJJ4XXkM7YXXK_DA.
- [16] 韦青. 狗尾巴的故事—数智时代的第一性原理[EB/OL]. 2022-05-15. https://mp.weixin.qq.com/s/cy4_ADKRYx530lrxBV16qw.

向量切片级细粒度管控： 证券投研知识库在RAG（检索增强生成）架构下的 非结构化数据安全隔离研究

茆正华 | 上海派拉软件股份有限公司

摘要：针对检索增强生成（RAG）知识库中向量检索“只认相关性、不识别访问权限”所带来的敏感信息越权风险，本文提出一种向量切片级的非结构化数据安全隔离方法。其要点是将访问控制贯穿数据全生命周期，并把权限判定前置到向量检索之前——先算清调用者实际可访问的范围，再据此约束向量召回，使越权内容在检索阶段即被隔离，而不依赖模型输出端的事后补救。文中给出该方法的整体框架、落地所需的基础条件与工程实践经验，并通过黑盒测试加以验证：无权限内容无法被检索、每次操作可追溯到人、AI与业务系统权限同源一致。该方法以证券投研场景为例展开，亦可推广至银行、保险、医疗等强监管行业的同类应用。

关键词：检索增强生成（RAG）、向量数据库、细粒度访问控制、切片级数据安全隔离、信息隔离墙、关系型授权（ReBAC）、非结构化数据治理

一、背景意义

证券公司积累了大量高密级的研报报告、内参和未公开重大信息（MNPI，即可能影响股价、尚未对外披露的内幕信息），现在越来越多地把这些资料放进基于“检索增强生成”

（RAG，一种让大模型先从知识库检索资料、再根据检索到的内容来回答问题的技术）的投研知识库和智能助手里。但传统RAG有一个一直被忽视的问题：底层的向量数据库只会判断“哪段内容和问题最相关”，并不知道“谁有权看这段内容”。结果就是，智能助手（AI Agent，通过检索、调用工具来完成任务的AI程序）在一次看起来很正常的问答里，就可能把别的部门的机密、还没发布的研报、甚至高管薪酬一并检索出来回答给用户，让证券行业多年来建立的“信息隔离墙”（指禁止投研、投行、自营等部门之间违规传递敏感信息的合规隔离制度）在AI时代被轻易绕过。

要解决这个问题，不能等大模型生成答案后再去“打补丁”，而要把权限控制贯穿数据使用的全过程，形成一个闭环，包含六个环节：资源发现、分类分级、智能打标、权限策略定义、检索时的“务实过滤”、全链路审计。其中“务实过滤”（指在检索之前就先算清楚用户能看哪些内容、再去检索）是整个闭环的关键：用户发起检索时，先由权限引擎 OpenFGA（一款开源的关系型授权引擎，按“谁和谁是什么关系”来判断权限，即ReBAC模型）实时算出这个用户当前真正有权访问的文档清单（allowed_doc_ids，即“允许访问的文档编号集合”），再把这份清单作为一道硬性的“元数据过滤条件”（metadata filter）提前加到检索里，让没有权限的内容在向量库做相似度计算之前就被排除在外。

本文围绕这个闭环展开，逐个环节讲清楚“做什么、需要

哪些能力、依赖哪些条件、怎么落地最稳妥”，更重要的是，对任何越权提问都能用黑盒测试验证“不该被看到的内容，一条都检索不出来”，从机制上杜绝了RAG的越权风险。

二、从“管住输出”到“贯穿全生命周期的闭环管控”

目前业界大多数RAG安全方案，做法都停留在“管住输出”这一步：要么在大模型生成答案后用安全护栏（Guardrails，指对模型输出做敏感词过滤、脱敏的一道防线）筛一遍，要么在提示词里要求模型“不要回答机密问题”。这两种做法有一个共同的问题——等到这一步时，敏感内容其实早就被检索出来、送进了大模型。所谓的“保护”只是在已经泄露的数据上再遮一层，既容易被提示词注入攻击（Prompt注入，指用精心构造的话术诱导模型绕过限制）绕过，也满足不了监管对“信息真正物理隔离”的硬性要求。

问题的根子在于：向量库不知道“是谁在查”。由此带来三类常见的越权场景：

- (1) “幻觉式回忆”：模型在某次问答里，把它之前在训练或检索时见过的内幕信息（MNPI）又复述了出来；
- (2) 绕过界面直接调接口：智能助手不走前端页面，直接调用后端的检索接口，前端那层权限遮蔽就完全失效了；
- (3) 审计追不到人：审计日志里记录的操作者永远是“系统”（Agent），事后根本查不出“到底是谁让Agent做了这件事”。

本文的核心观点是：要可靠地堵住RAG越权风险，必须让数据和权限在整个生命周期里同步管理，落成前面说的六环节闭环——在检索发生前就把权限条件加进查询里，从机制上挡住越权内容，而不是等模型生成答案后再补救。

如图1所示，这个闭环由六个环节组成：资源发现、分类分级、智能打标、权限策略定义、检索时的务实过滤、全链路审计。六个环节环环相扣：前四个环节在数据“入库”时，为每一段内容打上权限引擎能识别的安全标签和策略；第五个环节在用户“检索”时执行拦截；第六个环节的审计结果又回过头来修正策略和标签，形成持续优化的循环。整个闭环有一条可验证的底线：对任何越权提问，都能验证“不该看到的内容，一段都检索不出来”。



图1 非结构化数据安全治理全生命周期闭环

三、RAG数据安全治理闭环

阶段一·资源发现 (Data Asset Discovery)

这一阶段的前置条件是能访问到这些数据、网络通畅；拿到对象存储、向量库、API网关的接入账号和密钥；以及一套统一的资源登记规范（比如给每个资源起唯一编号 resource_id 的命名规则）。

这一步的目标，是先摸清家底：把分散在各处、将来会被智能助手检索到的数据全部找出来。在投研场景里，研报可能存在文档管理系统里，也可能散落在员工个人网盘、临时数据库、对象存储，甚至一些老旧的内部网站里。任何一个没被纳入管理的“影子数据源”，都可能成为越权访问的后门。

具体做法是“主动扫描 + 被动上报”两条腿走路：被动这一侧，通过资源连接器 (resource-connector，一种嵌入业务系统、用来对接数据的小组件，可用Java等语言的开发工具包接入) 让各业务方主动把自己的数据登记上来；主动这一侧，部署自动扫描工具在全网范围内发现数据。扫描范围覆盖关系型数据库的表结构 (Schema)、向量数据库、对象存储和文件系统里的各类文档、传统的API接口（应用程序接口，即系统之间互相调用的通道），以及新出现的MCP (Model Context Protocol，模型上下文协议，一种让大模型统一调用外部工具和数据的标准接口) 服务；对于老系统，还能用MCP转换器把老旧的Web接口 (RESTful API) 自动包装成大模型能直接理解和调用的资源。最终产出一份统一的“AI数据资产清单”和一张“数据血缘图” (Data Lineage，记录每份数据从哪来、流转到哪去的来龙去脉)。

先让数据“看得见”，再谈“管得住”——家底没摸清之前别急着打标签；用连接器低代码接入已有系统，用MCP转换器把老接口包装起来。如果有漏掉了影子数据源 (员工个

人网盘、临时数据库)，等于直接留了一个越权后门。

阶段二·分类分级 (Classification & Grading)

当有了阶段一摸清的资产清单；以及统一的“分类→定义→分级”三层标准。包括一套分类分级标准引擎、识别个人隐私信息 (PII) 和敏感内容的能力、业务领域划分、分级对照表的管理。

这一步是给找出来的数据定一套统一的标准：每份数据属于什么密级、归哪个业务领域。它是后面所有权限规则的基础。如果底层数据没有清晰、稳定的分类标识，配权限时就只能“一段内容配一条规则”，规则数量很快爆炸，根本管不过来。

具体做法是，用数据安全治理引擎 (ADSG) 对所有数据做一遍自动的语义扫描，识别出其中的个人隐私信息 (PII，如身份证号、手机号等) 和核心商业机密，再对应到四级密级：P0是核心机密，也就是内幕信息 (MNPI)，比如未公开的并购、高管薪酬；P1是部门内参或敏感信息，比如还没发布的研报、策略部门的参数；P2是内部公开信息，比如通用的业务文档、历史研报；P3是对外公开信息，比如已发布的报告、新闻、公告。同时，再按研究域、交易域等把数据划分到不同业务领域。最终产出一棵清晰的数据分类树和一张分级对照表。

标准要先定、字典要统一——先定义一套少而稳定的标签，再大规模去打标；分级的粗细要跟着监管口径走（参照《证券期货业数据分类分级指引》）。如果没有统一标签体系就盲目打标，后期一定会规则爆炸、没法维护。

阶段三·智能打标 (Intelligent Tagging，给每段内容挂上访问控制标签ACL)

这一步的目标，是把安全属性牢牢“贴”在数据的最小单元——切片 (Chunk，指把一份文档切成的一小段内容)——上，让权限有一个最细的落脚点。这正是比传统“按整个文件”或“按一行”来管权限更进一步的地方：一份投研PDF，前半段可能是公开介绍，附录里却藏着内幕信息 (MNPI)。只有把管控细到切片这一级，同一份文档里不同密级的段落，才能各自带上独立的安全标识。

具体做法是搭一条数据“入库流水线”：原始文件先被解析，再按语义切成一段段完整的内容 (智能分块，Smart Chunking)，然后用“多级漏斗”逐层打标签——先用正则表达式 (Regex，一种按固定格式快速匹配文本的规则) 抓出身份证号、银行卡号、股票代码这类格式固定的信息；再用命名实体识别 (NER，自动从文本里识别出人名、机构名、产品名等关键词的技术) 认出人名、机构名、产品代码；接着用深度语义模型判断结合上下文的密级 (比如区分“建议买入”和“未公开的买入建议”)；最后用大模型 (LLM) 兜底处理那些更复杂、更难判断的内容。打完标签后，把访问控制标签 (ACL Viewing Tag，记录“哪些人能看这段内容”的权限标签) 连同密级、文档编号、归属部门、其他标签 (security_level、doc_id、owner_dept、tags) 一起，预先写进每个切片在向量库里的元数据 (metadata，即附在每段内容上的描

述信息)中,再做向量化存储。这条流水线大约能自动发现95%的敏感内容,打标准准确率约90%。图2展示了完整流水线和元数据字段结构。



图2 入库打标流水线与切片元数据 (metadata) 字段结构

一套混合打标技术(正则Regex抓固定格式、命名实体识别NER抓实体、深度语义模型判密级、大模型LLM兜底)、把访问控制标签(ACL)预先写到每个切片上、以及把标签固化到向量库元数据(metadata)里的能力。

按把握程度分流处理——对把握很大的普通研报直接放行,一旦命中内幕信息(MNPI)就强制转人工复核;权限标签(ACL)要和数据一起同步管理。要避免的坑:对密级混杂的文件只按整份文件粗略打标,会导致附录里的内幕信息(MNPI)跟着正文一起被检索出来。

阶段四·权限策略定义(Policy Definition)

这一步的目标,是把人理解的业务规则——信息隔离墙、岗位职责、临时项目组等——翻译成机器能自动执行的动态权限策略。本方案不再走“给每段内容手动配权限”这种低效的老路,而是把三种授权模型组合起来用:TBAC(基于标签授权,按数据贴的标签来判断谁能看)、ReBAC(基于关系授权,按“人和数据是什么关系”来判断)、ABAC(基于属性授权,按时间、地点、设备等动态属性来判断)。

在OpenFGA里,先把企业里的各类对象统一建模成七种实体:用户、组织、岗位、角色、资源、操作、智能助手(User、Organization、Position、Role、Resource、Action、Agent),再通过写入一条条“关系记录”(元组, Tuple, 用来描述“谁和谁是什么关系”的最小单位)搭出层级关系。比如下面几条记录,就能表达“张三属于新能源组、这个组拥有某份研报、新能源组的成员可以使用投研分析助手”。之后引擎会顺着这些关系自动推导出跨部门、多层嵌套组织的权限,不用再一个人一个文件地手动分配。

OpenFGA关系记录(元组)示例:研究域投研场景

```
// 投研部门树: 新能源组 属于 研究所
{"user": "department:research_institute", "relation": "parent", "object": "department:new_energy"}
// 张三 是 新能源组 成员
{"user": "user:zhangsan", "relation": "member", "object": "department:new_energy"}
// 研报归属新能源组
{"user": "department:new_energy", "relation": "owner", "object": "document:report_001"}
// 新能源组全体成员可使用投研分析 Agent
{"user": "department:new_energy#member", "relation": "allowed_by", "object": "agent:research_assistant"}
// 跨墙特批: 合规审批后, 显式赋予交易员李四查看该研报
{"user": "user:lisi", "relation": "viewer", "object": "document:report_001"}
```

信息隔离墙可以用两种方式落地。一种是逻辑隔离(推荐):把研究域和交易域设成两个互不相连的根节点,引擎在“交易员”和“投研文档”之间本来就找不到任何连接路径,隔离墙自然就成立了;万一确实需要跨墙,只要走完合规审批、写入一条明确的授权记录,就能立刻打通。另一种是物理隔离:给研究域、交易域各建一个独立的OpenFGA库(Store, 即相互独立的命名空间),在底层数据库里彻底分开存放,检索时根据数据所属的领域分别到对应的库里去算权限。在这之上,还可以叠加按属性授权(ABAC)的动态条件——交易时段、IP网段、设备是否合规、风险评分等——进一步收紧可访问范围。最终,一个用户实际能看到的内容,是“这个智能助手出厂时被允许的能力范围”“调用者当前的权限”“策略绑定的范围”三者的交集,三重把关。

权限要绑在“组织路径/项目组”上,而不是绑在某个人身上,这样一旦有人换组,只要通过SCIM(全称跨域身份管理系统)实时更新关系记录,权限就能秒级生效;用按标签自动授权(TBAC)来摆脱“一段段配权限”的苦活。要避免的坑:智能助手这一层如果还沿用写死的按角色授权(RBAC),一旦有人换组就得批量改动成千上万条权限记录,期间会有好几个小时的“影子权限”(人已换岗、旧权限还没回收)真空期。

阶段五·向量切片查询务实过滤(Pragmatic Filtering)

这是整个闭环最关键的执行环节,目标是在检索真正发生之前,就从机制上切断越权内容被检索出来的可能。如图3所示,传统的“事后过滤”是先让向量库按相似度找出最像的前K段内容(Top-K),再到业务层一条条判断有没有权限。这样做有三个问题:敏感内容其实已经被读进了上下文(容易引发幻觉泄露);如果找出来的十条恰好都没权限,返回给用户的就是一片空白(翻页空页);而且性能很差。务实过滤的做法正相反——把权限判断提到检索之前。

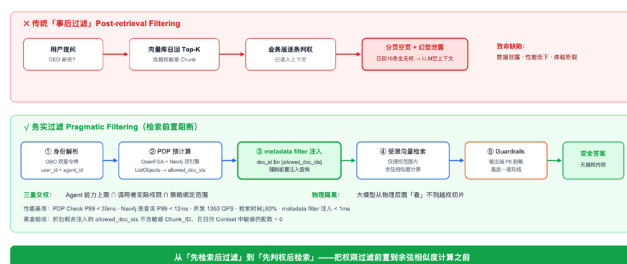


图3 检索侧务实过滤:检索前阻断vs事后过滤

整个检索流程分五步。第一步,解析身份:从OBO双重令牌(On-Behalf-Of,意为“代表某人”,令牌里同时带着“真实用户是谁”和“哪个智能助手在替他操作”两重身份)里解出用户编号(user_id)和助手编号(agent_id)。第二步,由权限决策点PDP(Policy Decision Point,专门负责判断“能不能访问”的组件,这里用OpenFGA加图数据库Neo4j双引擎来加速)一次性算出这个用户当前真正能访问的文档清单(allowed_doc_ids)。第三步,把这份清单作为一道硬性的

元数据过滤条件，提前加到向量检索里。第四步，向量库只在这个授权范围内去做相似度计算。第五步，输出环节再用安全护栏 (Guardrails) 对答案做最后一道个人隐私 (PII) 脱敏兜底。下面的示意代码给出了可落地的调用流程。

检索侧实务过滤示意代码 (Python 开发工具包)

```
def rag_query_with_pragmatic_filtering(user_id, query_text):
    # 1. 检索前置: OpenFGA ListObjects 预计算白名单 (Neo4j 加速, P99 < 20ms)
    allowed = authz.list_objects(user=f"user:{user_id}",
                                relation="can_read", type="document",
                                context={"ip": "10.0.0.5"}) # ABAC 动态因子
    allowed_doc_ids = allowed.get("objects", [])
    if not allowed_doc_ids:
        return "权限不足，您无权查阅相关知识库。"
    # 2. 生成查询向量
    qv = embed_model.encode(query_text)
    # 3. 动态条件注入: doc_id 已建标量索引，保证过滤性能
    results = vector_db.search(vector=qv, top_k=5,
                               filter_expr={"doc_id": {"$in": allowed_doc_ids}})
    # 4. 组装上下文交付大模型
    return llm.generate_answer(query_text, [r.text for r in results])
```

一定要“先过滤再检索”，不要“先检索再判权”——后者必然导致翻页空页、大模型拿不到内容而瞎编 (幻觉)，还拖慢性能；权限字段要建好索引，否则向量搜索会退化成为“全表逐条扫描”。性能基线：加过滤条件的额外耗时小于1毫秒，权限决策 (PDP) 在99%情况下小于35毫秒，检索耗时下降约80%。

阶段六·全链路审计 (Audit & Governance)

这一步的目标，是让AI的每一个动作从看不见的“黑盒”变成责任清楚的“白盒”，形成一条能直接交给监管看的证据链，同时把审计结果回流，持续优化前面几个环节。所有请求经过权限决策点 (PDP) 时——无论放行还是拦截——都会完整记进决策日志，记下五个要素：谁、什么时候、想访问什么资源做什么操作、判定结果、判定原因；同时还带上完整的提问内容 (Prompt)、消耗的Token (大模型处理文本的计量单位，可理解为字词的“块”)、以及OBO令牌里的双重身份 (真实用户human_user_id加智能助手agent_id)，再结合文档编号 (doc_id) 和切片编号 (chunk_id)，做到100%精准追溯。决策日志的核心字段包括：操作主体、追踪令牌、记录时间、资源类型、资源编号、操作、判定结果 (放行ALLOW/拒绝DENY)、命中的策略、触发的条件、风险等级等。日志会异步写入消息队列Kafka (一种高吞吐的数据管道)，再转入冷存储长期留存，保证不可篡改。

在这之上，系统内置一个身份治理稽查智能体 (IGA, Identity Governance & Administration, 即身份治理与管理)，会定期扫描全公司的权限分布：一方面按“互斥规则表”找出像“同一个人既能提单又能审批”这类职责不分离 (SoD, Segregation of Duties, 指关键操作必须分给不同人，防止一手遮天) 的冲突并告警；另一方面，把“给某人分配的权限”和“他实际用到的权限”做对比，对那些长期不用、甚至从来没用过的多余权限，生成“精简/降级建议”并自动回收，从而消除过度授权带来的风险面。这些结构化

的数据，还能一键生成对应等保2.0、ISO 27001 (国际信息安全标准)、SOC 2 Type II (面向服务机构的内控审计报告) 等标准的合规报告，直接作为监管现场检查时的举证材料。

让审计结果回流——用真实调用记录去比对授权情况，自动生成精简建议并触发回收；上线时采用“先观察、后拦截”的灰度方式。要避开的坑：只靠入库时打标，却放弃了输出端和日志监控，大模型可能通过发散推理把零散信息拼出机密，最后陷入“查不清、追不到”的审计黑洞。

四、落地前要先具备的基础条件

向量切片级管控不是一个孤立的小功能，而是一条从数据底座、身份关系到检索引擎的完整链路。要让务实过滤真正跑成闭环，下面六个基础条件必须先到位。

- **数据分类分级体系**：以数据安全治理分类分级标准为核心，给按标签授权 (TBAC) 提供输入，避免“一段段配权限”的规则爆炸。

- **入库打标流水线**：在“解析→切片→入库”过程中自动识别主题、实体和个人隐私信息 (PII)，并预先挂上访问控制标签 (ACL Viewing Tag)，这是权限最细的落脚点。

- **身份与关系建模**：把企业抽象成用户、组织、岗位、角色、资源、操作、智能助手七类实体，用关系记录自动推导出跨部门、多层嵌套组织的权限 (即基于关系的授权 ReBAC)。

- **身份与主数据集成**：通过 SCIM 2.0 标准与人力系统 (HR)、身份认证系统 (IAM, 身份与访问管理) 打通，让员工入职、转岗、离职 (JML, Joiner-Mover-Leaver) 这些事件实时驱动权限的更新和吊销。

- **权限决策引擎 (PDP)**：OpenFGA 负责关系推导、图数据库 Neo4j 负责加速，双引擎配合，经三重把关后输出用户能访问的文档白名单 (allowed_doc_ids)。

- **向量库的元数据过滤能力 (Metadata Filter)**：选用支持高性能字段过滤、主键过滤的向量库 (如 VikingDB 等)，把用户的可访问文档清单 (allowed_doc_ids) 强制提前注入检索，避开“先检索后过滤”的毛病。

五、落地实践与贯穿闭环的工程难点

- **存量历史文档 (数十TB, 即数十太字节) 的回扫重打标**：分批回扫，再按把握程度分流处理。在没有统一标签体系 (分类→定义→分级三层) 之前就盲目打标，后期一定会“规则爆炸”、没法维护。

- **自动打标准确率与人工复核成本的平衡**：按风险分 (RiskScore) 和把握程度 (Confidence) 分流——把握大的普通研报直接放行，一旦命中内幕信息 (MNPI) 就强制转人工复核。既不能完全信任 AI (一旦误判就是合规处罚)，也不能全部靠人工 (会拖垮入库的实时性)。

· **隔离墙跨墙审批与检索低延迟的平衡**：用前置的务实过滤，加OpenFGA与图数据库Neo4j双引擎，毫秒级算出跨墙后的可访问范围再注入检索。绝不能“先检索后过滤”，否则会接连出现三个问题——翻页空页、大模型拿不到内容而瞎编（幻觉）、性能严重下降。

· **研究员频繁换组时权限要秒级生效**：把权限绑在“组织路径/项目组”上，再用SCIM 2.0实时推送入转离（JML）事件，更新“属于哪个组”的关系记录。智能助手这一层如果还用写死的按角色授权（RBAC，权限随角色固定、不随关系变化），换组会造成好几个小时的“影子权限”真空期。

· **向量库元数据过滤（metadata filter）的性能退化**：给权限字段（权限标签Viewing Tag、文档编号doc_id）在库里建好结构化的字段索引。用复杂的正则字符串匹配去代替字段索引过滤，会让向量搜索退化成“全表逐条扫描”。

· **打标出错时的兜底与灰度上线**：用输出端的安全护栏（Guardrails）做脱敏兜底，加全链路审计，按“先观察、后拦截”的方式灰度上线。只靠入库打标、放弃输出端监控，大模型可能靠发散推理把零散信息拼出机密，最后陷入“查不清、追不到”的审计黑洞。

六、对照监管要求与可验证的收益

以下尝试将监管要求直接转化为可落地的技术方案，以便券商技术负责人可直接对照监管条款开展工作。具体对应关系如下：

规范	方案
《证券公司信息隔离墙制度指引》	用OpenFGA的独立命名空间做硬隔离、用关系模型（ReBAC）筑墙
《证券期货业数据分类分级指引》	P0-P3四级自动打标体系
《数据安全法》《个人信息保护法》	用命名实体识别（NER）加正则表达式（Regex）自动识别并脱敏个人隐私信息（PII）
网络安全等级保护2.0（等保2.0）/金融行业标准JR/T 0071-2020	全链路审计加人机协同复核（HITL，Human-in-the-Loop，关键环节由人来把关）
ISO 27001（国际信息安全管理标准）/SOC 2 Type II（面向服务机构的内控审计报告）	策略以代码形式管理、决策路径可回放复现

与“安全成功率99%”这类空泛指标不同，本方案给出的是工程师可以自己复现、验证的实际收益：

一是机制上的拦截：黑盒测试可以验证“无权限的用户去检索敏感问题时，被禁内容一段都检索不出来”。具体做法是，往库里放一段辨识度很高的伪造机密切片，标成L4（只有超级管理员能看），再用普通账号反复拿这个密词去检索，抓包确认算出的可访问清单（allowed_doc_ids）里绝不包含这段内容的编号（chunk_id），并且返回的上下文里一条都匹配不到；

二是责任可追溯：审计记录的四个要素（真实用户human_user_id、智能助手agent_id、文档编号doc_id、切片编号chunk_id）全程可视，能直接满足等保对“责任到人”的要求；

三是性能可实测：加过滤条件的额外耗时毫秒级，并发能力约千级QPS（每秒处理的请求数）；

四是权限一致：AI系统的文档权限和传统业务系统同源（通过SCIM同步），不会出现两套权限各管各、互相打架的情况。

更重要的是它能复制推广：这套从资源发现到审计的完整闭环，并不只适用于证券投研，还能平滑迁移到银行客户经理的客户身份核实（KYC，Know Your Customer，即“了解你的客户”）知识库、保险的核保资料库、医疗的病历库，以及央企的内部研报库。只要是“强监管 + 高密级非结构化数据 + 大模型应用”的组合，都能套用同一套做法。

这是一套可以直接套用到各种主流向量库（如VikingDB、Milvus、Qdrant、PGVector等）的可复用做法，也是一套方法论：把监管条款翻译成OpenFGA的关系模型和切片（Chunk）标签体系，最终输出一条能直接交给监管现场检查的全链路证据链。

参考文献

- [1] OWASP Foundation. OWASP Top 10 for large language model applications[EB/OL]. (2025-01)[2026-06-20]. <https://owasp.org/www-project-top-10-for-large-language-model-applications/>.
- [2] OpenFGA Authors. OpenFGA: relationship-based access control at scale[EB/OL]. Cloud Native Computing Foundation, [2026-06-20]. <https://openfga.dev/>.
- [3] 全国人民代表大会常务委员会. 中华人民共和国数据安全法[Z]. 北京: 全国人民代表大会常务委员会, 2021.
- [4] 全国人民代表大会常务委员会. 中华人民共和国个人信息保护法[Z]. 北京: 全国人民代表大会常务委员会, 2021.
- [5] 中国证券监督管理委员会. 证券期货业数据分类分级指引: JR/T 0158—2018[S]. 北京: 中国标准出版社, 2018.
- [6] 国家市场监督管理总局, 国家标准化管理委员会. 信息安全技术 网络安全等级保护基本要求: GB/T 22239—2019[S]. 北京: 中国标准出版社, 2019.
- [7] 国家市场监督管理总局, 国家标准化管理委员会. 信息安全技术 个人信息安全规范: GB/T 35273—2020[S]. 北京: 中国标准出版社, 2020.
- [8] ISO/IEC. Information security, cybersecurity and privacy protection — information security management systems — requirements: ISO/IEC 27001:2022[S]. Geneva: ISO/IEC, 2022.

从RSAC看国内网安AI发展方向

大模型安全研究团队 | 绿盟科技集团股份有限公司

摘要： 本文结合2026年RSAC展会观察，指出网络安全重心已开始向AI代理安全治理进行转变，本届创新沙盒企业聚焦影子代理管控、AI反欺诈、代理动态权限管理等AI原生安全赛道，形成AI对抗AI、全生命周期治理的行业转向。对比国内产业，我国在大模型合规防护具备先发优势，但在AI代理全生命周期治理、开发流程无缝安全嵌入、安全运营韧性、AI社会工程欺诈防御上等方面存在一些短板。文章提出国内网安三大发力方向：打造AI代理完整治理闭环、布局基于认知心理学的AI反诈技术、依托监管合规红利深耕全链路可信溯源，跳出对标思维，在新的网络安全市场机遇来临时，抢占AI安全新赛道话语权。

关键词： RSAC 2026、创新沙盒、AI安全、AI代理、Agent安全、影子代理、AI原生安全、AI反诈、社会工程防御、大模型安全、安全治理

说实话，干咱们网络安全这行的，每年去RSAC已经有点像“朝圣”了。但今年2026年这届感触特别深，甚至可以说有点“震撼”。不为别的，就因为走进创新沙盒会场那一刻，你闻到的全是“AI”的味道——不是前两年那种把AI当个功能挂件、当个提效工具的玩法，而是AI本身成了安全的“新问题”和“新战场”。



一、RSAC 2026告诉我一件事：安全重心正在从“人”转向“Agent”

这届创新沙盒十强，所有入围企业全都深度整合了AI技术。但比“全覆盖”更值得琢磨的是，十家里面差不多有一半——包括最后拿了冠军的Geordie AI，还有Token Security、Realm Labs——干的都是同一件事：管住AI代理（Agent），或者为非人类身份建一套新的信任体系。

你如果去现场逛一圈，会发现一个现象：CrowdStrike在讲AI代理发现，Astrix在新添代理发现方法，BeyondTrust在给AI代理搞权限强制执行……整个会场约18.5%的参展商业务跟AI安全直接相关，其中73家把“Agentic”写进了自己的核心叙事。

这说明了啥？说明“给AI当保安”这个赛道，已经从边缘变成了主场。2023年大家还在聊大模型本身好不好用，到2026年，AI开始像人一样去操作业务、去调用API、去写代码、去做客服。它已经不再是单纯的“工具”，而是能自主行

动的“数字员工”。当这些Agent数量几百上千地爆发，它带来的身份混乱、权限滥用、行为不可视、多代理互信，就成了一套完全不同于以往的安全难题。

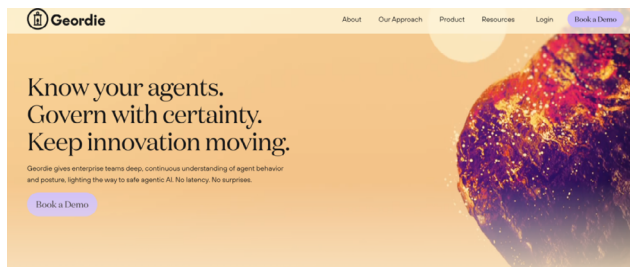
这时候，你得用一个全新的脑子去想安全——不是“怎么防黑客攻我的AI”，而是“怎么在组织里管好一群非人类但又能干坏事的高级数字生物”。

二、从十强看门道：他们具体在做什么，做得怎么样？

咱们挑几个重点的聊一聊，比光看榜单有意义。

冠军Geordie AI，干的就是替企业看住所有“影子代理”

Geordie AI给我留下最深印象的，不是它的技术多玄幻，而是它的切入角度太刁钻了。现在企业内部，员工偷偷在用各种AI代理——在GitHub上装个编码助手Agent，在VSCode里跑个自动化测试Agent，或者市场部自己部署一个邮件营销Agent。这些没经过IT和安全团队审批的“影子代理”，是巨大的盲区。

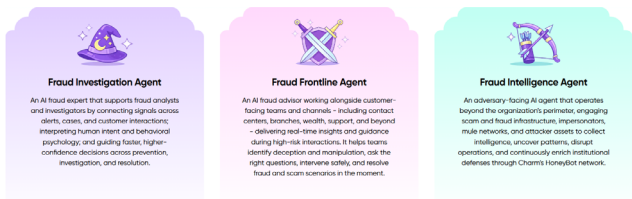


Geordie AI干的就三步：第一，帮你全网扫出来，到底有多少个这样的代理在跑，在哪个环境里，用什么身份。第二，持续盯着它们的行为和权限使用，看有没有异常。第三，再落到缓解措施上。

他们CEO Henry Comfort是从Darktrace出来的, CTO Benji Weber原来是Snyk和Mindgard的工程负责人。这团队组合有个特点: 既懂AI怎么在现实里被滥用, 又懂怎么从开发侧的源头去做工程化防护。所以他们去年9月拿完650万美元种子轮, 产品发布俩月营收就涨了十倍。你不得不服, 人家踩点踩得太准了。

Charm Security: 用心理学来反AI诈骗, 思路很清奇

Charm Security做AI反欺诈, 但不是简单地去认证一个声音是不是Deepfake。他们搞了一套叫HVE (人类漏洞评估) 的模型, 把心理学的认知偏差——比如“权威服从”“紧迫性偏见”“稀缺效应”——给数字建模了。

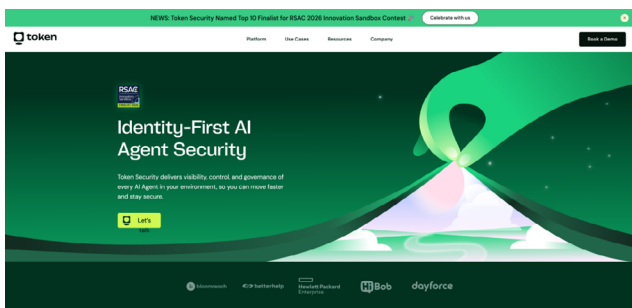


什么意思呢? 就是一个AI骗子在跟你聊天、打电话的时候, 它会实时分析话语里的心理学操控痕迹, 比如“你是不是在用命令式口吻”“对方是不是在制造时间压迫感”“话术里有几个情感勒索的信号”, 然后结合语音语调、对话节奏, 判断这是一起社会工程攻击。

你会感慨, 这才是真正的“用AI对抗AI”。过去咱们防诈骗, 总是在等样本、等规则, 但AI驱动的欺诈是千人千面、实时生成的。只能靠理解“人性漏洞”去反制。这条路国内做得还远远不够。

Token Security: 管代理权限, 不能只看“是谁”, 要看“想干嘛”

Token Security提了个概念——基于意图的权限管理。过去我们做身份安全, 中心思想是“这个人/这个账号是谁, 该给他啥权限”。但到AI代理这儿, 一个授权可能就是一个Token、一个API密钥。代理可能就运行几分钟, 完成一个任务就销毁。你用静态的权限去管, 根本管不住。

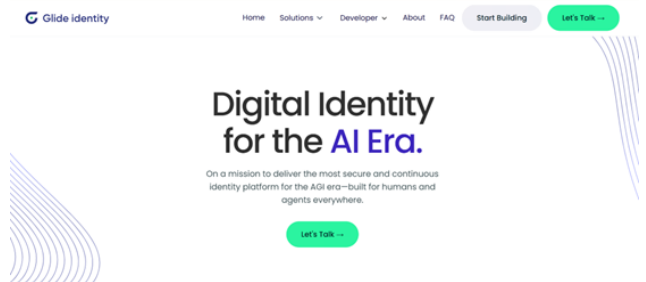


他们做的, 是让系统去理解“代理现在想干什么、为什么要这么干”, 动态赋予最小权限, 干完活儿密钥立刻轮换。这等于是把传统IAM那一套, 从面向人类用户, 彻底改成了面向非人类身份的、具备持续验证能力的体系。

在现场你能看到, CISO们问得最多的问题就是: “我们连自己有多少AI代理都数不清, 怎么才能在它失控之前把它

管住?” 这恰恰说明需求已经非常刚性了。

其他的入围厂商, 也各有看点。比如Clearly AI, 目标是当你的“AI安全工程师”, 自动读你的需求和设计文档, 完成威胁建模和代码风险分级; Crash Override搞了个“工程关系管理”, 给每一行代码加上防篡改的溯源标签, 把供应链追溯做到极致; ZeroPath专做AI原生应用安全测试; Fig Security盯着你的安全运营栈, 发现哪个检测规则失效了、哪个自动化断线了, 给你修复; Glide Identity在无密码身份认证上走得很深, 用电信网络情报和设备信任来替代短信验证码。



其他的入围厂商, 也各有看点。比如Clearly AI, 目标是当你的“AI安全工程师”, 自动读你的需求和设计文档, 完成威胁建模和代码风险分级; Crash Override搞了个“工程关系管理”, 给每一行代码加上防篡改的溯源标签, 把供应链追溯做到极致; ZeroPath专做AI原生应用安全测试; Fig Security盯着你的安全运营栈, 发现哪个检测规则失效了、哪个自动化断线了, 给你修复; Glide Identity在无密码身份认证上走得很深, 用电信网络情报和设备信任来替代短信验证码。

把这些串起来, 我们可以看到三个特别明显的转向:

- 安全产品不再是为“人”设计的, 而是为“Agent”设计的, 这叫“AI原生安全”。
- 攻防对抗的主导权, 正从人交给AI。防守必须做到“AI对抗AI”, 人退到决策和监督位置。
- AI安全不再是一个个孤立检查点, 而是覆盖“发现-评估-管控-恢复”的全生命周期治理。

三、回头看看国内: 咱们走到哪儿了? 差在哪儿? 强在哪儿?

拿RSAC这面镜子照照国内, 实话实说, 有底气, 也有心虚的地方。

AI代理安全与治理, 我们跟得挺紧, 但需要更深层的技术重构

国内反应最快的几家, 其实去年底就已经在布这个局了。今年某老牌安全公司推出了AI安全一体机, 走的是软硬一体、多安全围栏管行为的路线, 重点就是帮客户看见“影子AI”, 用无侵入的方式建起输入输出、检索、执行四道围栏。某互联网大厂联合清华大学开源了ClawAegis, 去防OpenClaw这类代理框架的恶意指令注入。还有的安全公司

在讲“像保护数字员工一样保护AI应用”，覆盖准入、身份、权限、行为等方面。



但坦诚地讲，跟Geordie AI相比，我们多数产品还是偏“加固”思维——相当于知道有坏东西要进来，我建墙、我检查。而Geordie AI更偏向“治理”思维，它假定这些代理本身就要在企业里存在、要干活，我的任务不是不让它进来，而是让它安全地干活。这个从“安全防护”到“安全治理”的升级，我们还需要时间去做深，特别是在代理行为意图的理解、多代理联动风险的建模上，国内厂商差距还比较大，得补课。

代码和供应链安全，国内专精但不“丝滑”

还有的安全公司发布AI赋能代码审计工具，思路都是用多智能体协同来做漏洞检测和复现。咱们在漏洞挖掘的深度上其实差距还比较大，有些厂商甚至给Agent加了个SafeSkill技能平台，防Agent自己执行危险操作，这个想法很前卫，但深度做的不够。

我们缺的是什么？缺的是和开发者日常工作流的“无缝感”。你看Clearly AI和Crash Override，是完全嵌在GitHub、CI/CD管道里的，开发者几乎感觉不到它的存在。国内企业级产品往往部署重、配置复杂，这让习惯了快节奏的开发团队不太容易接受。如何让安全能力“低摩擦”，是下一步产品化要突破的。

大模型安全防护，中国其实走在了前面

这一点不用谦虚。从基于双模型底座的多重语义分析拦截，到智能安全卫士，再到安全大脑类的大模型防护体系，国内在防提示注入、防模型越狱、内容合规检测这些方面的投入和产品成熟度，整体是不输于这届RSAC上展示的同类的。原因很直接——咱们的监管环境。中国2023年就出台了生成式AI管理办法，企业有刚性的合规需求，市场被催熟得早。央视3·15晚会今年还专门曝光了GEO技术给大模型“投毒”，直接拉动了采购需求。这是我们一个很明显的先发优势。

AI赋能安全运营，广度够了，但缺“韧性”视角

从国内安全公司产品体系上看，从AI赋能系列产品及自动化渗透测试，到安全领域的AI助手，到智能安全运营中心，再到AI驱动的扩展检测响应，咱们基本上把安全运营全流程都用AI武装了一遍。但RSAC上Fig Security给了一个不一样的启发：它不解决“怎么用AI检测”，它解决“当你的检测和响应资产越来越多，怎么保证它们不‘腐烂’、不失效”。

细想一下，我们很多安全运维痛点，不是没有告警，而是告警规则老化、日志源掉线、自动化剧本跑不通了没人发现。做一个能自愈、能保持安全运营栈“韧性”的产品，在国内还几乎空白，这个方向，值得做。

最大的短板：用AI防御社会工程攻击

这块是感觉最“扎心”的地方。Charm Security和Humanix干的事儿，在国内找不到严格意义上的对标产品。咱们面对深度伪造的AI换脸、拟声诈骗，更多还是被动防御——提醒用户小心、延迟转账，极少能做到在诈骗进行中，实时识别话术里的心理学操控并介入干预。

你可以说是因为国内外诈骗场景不同、通信数据隐私保护要求高。但别忘了，现在的AI能合成任何人的声音，能生成带有亲属情绪色彩的诈骗话术。以后骗你爸钱的，可能就是一个完全模拟你声音和口气的AI。我们如果不去做基于行为心理学、多模态实时分析的防御，这个攻防缺口会被越拉越大。

四、聊点实在的：下一步国内AI安全该往哪发力？

结合上面的这些观察，如果让我站在厂商和解决方案的角度给点建议，我会说三个值得重点押注的方向。

第一，尽快把“AI代理治理”做成完整闭环，而不是一个个功能点

现在大家都在提“影子代理发现”“代理权限管控”，但真正要落地，得是发现、测绘、行为基线、异常检测、自动响应一套打通，而且要能和客户已有的身份系统、SOC对接。国内某互联网公司在做的多代理互信安全框架（基于可信身份、可信连接、可信意图、可信授权）是很有远见的，同时，某安全公司也在通过AI安全一体机的形态去降低部署门槛。

但更关键的，是要敢于跳出“人管机器”的思维，构建“机器管机器”的治理模型。要让AI代理自己也能承担部分安全责任，比如代理在请求高敏API前，自己能去做一次意图的自检和权限的合理性声明。这个方向，既有技术挑战，也有标准话语权的机会——谁先定义出AI代理的安全交互协议，谁就占据主动权。

第二，补上“社会工程AI防御”这一课，刻不容缓

咱们不怕跟国外比技术，怕的是连赛场都没进。建议有积累的安全企业——比如有多模态内容分析能力的、有反欺诈业务基础的——可以考虑联合心理学研究机构，打磨类似Charm Security那样“认知偏差检测+社交工程对话模式分析”的引擎。不一定非要做成一个独立产品，可以先嵌入到现有电信反欺诈平台、银行风控系统里去。

换个角度想，这其实是一片蓝海。国内银行业、运营商对反欺诈的投入极大，但工具还停留在规则和特征层面。谁能给他们一套能实时识别AI心理操控的检测能力，谁就能打开一个新的市场。

第三，走“合规+技术”双轮驱动的路，把先发优势放大

中国AI安全有个得天独厚的条件是监管推动。今年全球AI治理从原则倡导迈向制度落地，我们在政务、金融、央企这些领域，有大模型安全评测、AI供应链溯源、内容合规检测的刚性需求。这给我们提供了稳定的粮仓。

但咱们不能一直吃合规的红利，得借这个势，把底层技术打磨扎实。像一些领先实践那样做到每行代码的防篡改追踪，我们是不是可以在信创环境里，做出从模型训练数据、微调过程到推理服务的全链条可信溯源？这不仅是技术挑战，更是建立信任壁垒的好机会。国内一些安全厂商在双模型底座的实践中，就在尝试让合规检测和安全能力本身变得AI原生化，用模型去治理模型，这条路相信国内厂商会坚持走下去。

写在最后

RSAC 2026，我最大的感受不是“狼来了”，而是“规则变了”。过去，安全是跟在业务后面走的，业务上线什么，我们就保护什么。但在AI时代，安全有机会、也必须跑到前面，去定义AI该怎么被负责任地使用。

中国网安产业从来不缺勤奋和聪明，缺的可能是对规则变化做出快速反应的那股“狠劲”和“想象力”。AI代理的安全治理、对抗AI欺诈的心理学防御、安全运营的韧性保障，这些都是可以拉开身位的新赛道。作为从业者，我很期待看到国内能有更多厂商不再满足于“对标”，而是敢于去“定义问题”——就像RSAC 2026上那些十强公司做的那样。

毕竟，在AI深度介入安全攻防的今天，谁先建好为“新数字物种”服务的信任体系，谁就拿下了下一个十年。

04 热点解读

P100 证券行业安全养“虾”探索

庄飞、江旺

P103 AI Agent安全视角下的大模型金融应用风险与多层防御体系研究

陈洪炎

P108 证券期货业大模型安全围栏的建设与运营思考

钟蓉、吴佳伟、李鹏

P113 证券公司实网攻防演习防守实践

刘志敏、陆修然

证券行业安全养“虾”探索

庄飞、江旺 | 华泰证券股份有限公司

摘要：随着大模型技术在证券行业落地，具备系统级任务执行能力的AI智能体已成为提升业务效能的核心工具。然而，以OpenClaw为代表的“龙虾”类智能体，在带来效率跃升的同时，也引入了数据泄露、恶意插件注入、远控劫持等新型安全风险。本文介绍了生产环境专用龙虾建设方案以及办公桌面龙虾安全防护思路，通过沙箱隔离、OpenClaw专用网关、本地大模型平台、内部SkillHub等管控措施，降低企业养“虾”的安全风险。

关键词：大模型、OpenClaw、智能体、Skill

一、前言

随着大模型应用落地加速，具备系统级任务执行能力的AI智能体成为企业及个人提升效率的核心工具，以OpenClaw为代表的“龙虾”类智能体，被广泛应用于业务提效、办公自动化、开发辅助、运维监控等场景，可实现从意图理解到任务执行的全闭环。证券行业很多企业也开始探索使用Claw智能体赋能业务，如投研分析、行情资讯解读等。然而，当前Claw智能体的部署使用仍然存在诸多安全隐患：

一是数据泄露风险。Claw智能体在执行任务时，会处理的文件、输入文本及任务相关数据一并发送给大模型，若调用外部公有云大模型，内部敏感数据将直接流出企业边界。此外，攻击者还可能通过控制智能体绕过现行安全策略，借助微信、QQ、飞书等远控端直接读取或外发内部文件，进一步放大数据泄露风险。

二是恶意Skill引发的后门与远控风险。Claw智能体在访问网站、处理文件或安装外部Skill的过程中，可能在用户不知情的情况下自主的访问恶意网站或下载安装恶意程序，进而导致终端被控风险。

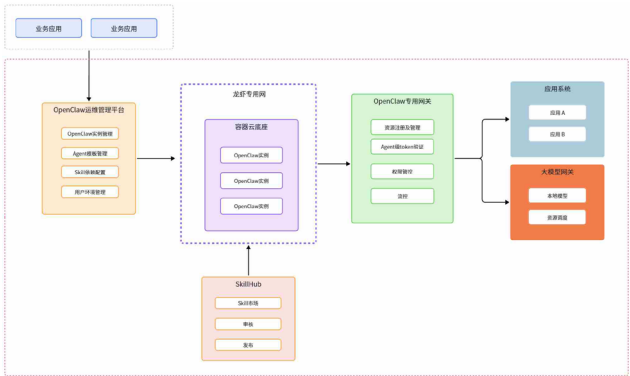
三是在Claw智能体带来的内网攻击风险。Claw智能体本身若存在系统漏洞或暴露不必要的网络服务，则在内网部署后容易因智能体被控进而导致终端被控，同时，一旦远控端账号（微信/QQ/飞书等）失陷，攻击者即可直接操控智能体实施内网横向攻击，显著增加了内网被恶意攻击风险。

为了应对上述Claw智能体带来的安全风险，同时兼顾业务智能体的灵活性需求，本文从生产环境专用龙虾建设、办公桌面龙虾安全防护、龙虾安全管控能力建设三个维度展开讨论，探索一种完整的安全养“虾”方案。

二、生产环境专用龙虾建设方案

某些业务场景，需要在生产环境部署OpenClaw，利用Claw智能体处理业务数据，涉及高敏感业务信息，因此生产环境的龙虾需要从严管控。为了提升生产环境Claw智能体安全风险防护能力，基于零信任原则，设计了生产环境专用

龙虾建设方案，将OpenClaw实例集群部署于独立网络区域中的容器云底座，所有外部访问经由专用网关统一代理与鉴权，模型侧强制接入内部大模型，Skill侧通过内部SkillHub统一管理分发，并在全链路构建完善的日志审计与漏洞运营体系。



本方案的安全管控机制主要包括：

(1) 专用隔离环境。核心网建设OpenClaw实例运行专用网络区域，使用容器隔离沙箱机制，OpenClaw实例集群部署于容器云底座提供的独立网络空间，用于承载实际的业务负载与会话上下文。其内置的OpenClaw-Gateway负责接收上层业务应用的路由分发请求。在实例内部，利用动态挂载用户级环境、使用Podman沙箱作为Agent运行时等方案来实现资源与数据的隔离。使用经过安全加固的OpenClaw版本，禁止使用来源不明的第三方版本与镜像。容器隔离遵循最小权限原则，文件权限层面采用容器内目录管控策略，开启容器逃逸防护，容器仅挂载必要的工作目录，设置目录权限仅为使用者可读写，禁止挂载宿主机敏感目录。

(2) 定制化OpenClaw。基于OpenClaw开源版本，定制化专用OpenClaw版本，并进行安全加固，加固措施包括监听地址限制，将OpenClaw Gateway等服务的监听地址绑定到127.0.0.1或专用内网IP，禁止监听在0.0.0.0；安全的访问入口，禁用所有非必要的容器端口直接映射，如需从外部网络访问管理界面，必须通过反向代理进行转发，并在代理层配置强制HTTPS、IP白名单、会话超时及额外的身份认证；

访问令牌加固，必须修改默认或随机的访问令牌，设置符合企业密码策略的高强度、复杂令牌，并定期（建议每1-3个月）轮换。令牌需通过安全方式分发；日志配置，开启debug等详细日志等。

(3) OpenClaw专用网关。OpenClaw所有访问外部生产系统的请求必须通过OpenClaw专用网关转发，禁止本地直连生产资源。网关代理所有核心路由，执行单元级Token验证、资源注册管理与细粒度权限管控。网关仅能挂载经过安全加固之后的业务系统API，并配置访问鉴权，网关具有流控功能，防止OpenClaw对生产系统发起DDOS攻击。

(4) 本地大模型。OpenClaw只允许访问企业内部部署的私有化本地大模型，禁止直连互联网SaaS大模型，从根本上杜绝业务数据经由大模型调用外流。

(5) 内部组件仓库。由于业务引入的Skills，有的涉及安装第三方依赖，需要依赖npm以及pip源。禁止OpenClaw直接访问互联网安装第三方依赖，而是在生产环境搭建内部npm仓库和pip仓库，Skills第三方依赖包需要在研发测试环境验证后，上传到研发测试环境的仓库，再同步到生产环境，实现可控分发。

(6) 互联网访问管控。外部引入的Skill，存在访问互联网需求，如访问外部行情数据。对于此类Skill，设置准入条件，仅允许访问境内固定数据源，通用搜索类访问不予支持。在DMZ区域搭建OpenClaw专用流量网关，所有互联网请求经由该网关转发，按需开通出站策略，并对访问行为进行完整记录。

(7) 内部SkillHub。建设企业自有SkillHub，作为OpenClaw生态的Skills市场与分发平台，承担Skill包的上传、存储、审核与分发职能。建立全流程Skill安全管控机制，Skill需要经过安全扫描和人工审核环节才能入库和发布，排查恶意代码、权限越界、数据外发等风险。所有Skill需统一从SkillHub获取，禁止用户私自安装来源不明的第三方Skill。Skill安装、更新均需经过人工审批后方可执行。

(8) 全链路日志审计。构建全链路日志审计追溯体系，覆盖OpenClaw运行全生命周期，完整记录用户操作指令、AI决策逻辑、插件调用行为、系统资源访问以及外部接口请求等全维度数据，日志留存时长满足合规追溯要求。所有相关日志均汇聚至安全态势感知平台，针对批量文件读写、高频敏感接口调用、异常数据外传等风险行为配置实时告警规则，一旦发现异常，立即中断服务并启动溯源处置流程；同时定期开展审计日志分析，排查潜在的权限滥用、恶意注入等未被拦截的安全风险。

(9) OpenClaw运维管理平台。OpenClaw运维管理平台作为全局管控中枢，负责OpenClaw实例的全生命周期编排与系统级配置管理，不仅涵盖实例的创建、资源分配与回收以及Agent模板的统一管理，还负责管理用户级隔离环境配置并统一下发Skill依赖策略。此外，平台承担专用网关鉴权身份的创建与实例身份管理职能，并提供全局的调用统计查询与链路跟踪查询服务。

(10) OpenClaw版本更新。版本更新，遵循“按需引入”原则，满足重要漏洞修复、核心功能发布或业务使用方存在需求三个条件之一的优先考虑。核心组件（如OpenClaw-

Gateway）的升级及依赖策略的统一下发，采用灰度发布策略，生产集群按比例（如10%->50%->100%）执行灰度发布，并持续观测大模型网关的调用异常情况。实施版本变更前，系统自动执行实例级配置快照与状态备份；若升级发布后触发严重安全告警、权限越权或业务阻断，OpenClaw运维管理平台需具备分钟级的自动化容灾回滚能力，确保智能体平台的高可用性底线。

(11) 漏洞运营。建立OpenClaw漏洞应急响应机制，按照不同漏洞等级及时进行预警及响应。针对高危及以上级别安全漏洞，触发紧急修复标准流程，一是紧急隔离，在官方补丁或修复代码就绪前，第一时间联动OpenClaw专用网关配置临时访问拦截规则，在SkillHub中下线受影响的技能，从网络与业务入口阻断风险；二是紧急验证测试，修复方案（包括补丁升级或代码修复）须优先在测试环境完成功能及回归验证，确保不影响核心逻辑；三是紧急发布与复测，开启紧急发布通道，跳过常规发版窗口向生产集群全量推送补丁，并开展安全复测，确认漏洞修复且核心功能无损。对于中低危漏洞，依据企业安全服务等级协议纳入OpenClaw的常规敏捷迭代版本中进行集中修复与发布，邮件通知后在业务流量低谷期后台升级。

三、办公桌面龙虾安全防护思路

企业员工为了探索AI赋能业务，也会在办公桌面安装龙虾（如QClaw、WorkBuddy等），用于数据处理及报告自动化生成等场景，然而，桌面龙虾应用种类繁多，本身也处于快速迭代阶段，管理成本和难度相对较高，因此为了提升办公桌面龙虾安全风险防范能力，应从统一使用龙虾种类、建立安全基线、补齐自有技术支撑能力等方面进行防护。

(1) 统一桌面龙虾使用种类。统一企业内部办公桌面使用龙虾的种类范围，以便集中资源开展龙虾应用支撑和安全防护。制定龙虾应用安全基线，规范龙虾账号口令配置、敏感数据存储、日志存储、端口开放、接口加密、漏洞修复以及Skill下载安装等操作，降低因用户使用不熟练或龙虾配置缺陷原因导致龙虾被网络攻击风险。

(2) 建立内部大模型支撑能力。内部大模型建立对接龙虾能力，将桌面龙虾的模型调用引导至公司内部大模型，限制龙虾外连公有云大模型行为，从源头防范内部敏感数据通过大模型调用外泄。同时逐步完善大模型调用的日志记录体系，以及探索利用大模型对龙虾严重违规行为实施拦截等，进一步降低使用龙虾开展违规操作的风险。

(3) 建立桌面龙虾安全Skills。研发龙虾专属安全Skills，通过高危行为强制二次确认、网络下载Skills开启安全扫描、安装外部程序扫描等方式，降低龙虾自主安装恶意程序引发的被攻击被远控风险。

(4) 建立内部IM工具远控桌面龙虾能力。由内部IM工具团队开发支撑龙虾远控功能，使员工能够通过企业内部IM工具远程操控桌面龙虾，降低因为使用微信、QQ、飞书等外部工具而导致的数据泄露风险，以及因控制端账号丢失导致龙虾被恶意控制风险。

(5) **明确桌面龙虾使用红线。**制定并发布桌面龙虾使用红线，如禁止使用龙虾处理涉密信息、禁止利用龙虾故意绕过管控措施等，并加强龙虾使用安全教育培训，提升全员安全意识和技能水平。同时针对发现的违规行为，采取限制龙虾使用及开展安全考核等措施，降低龙虾使用安全风险。

(6) **统一接入内部SkillHub。**限定桌面龙虾只允许安装使用内部SkillHub中已通过安全审核的Skills，避免因使用外部Skills仓库引入木马后门等恶意程序。

四、龙虾安全管控能力建设

在落地上述防护方案的同时，需同步建设配套的技术支撑与流程机制，提升整体安全管控的主动感知与快速响应能力。

(1) **提升龙虾安全监测及应急处置能力。**采集龙虾相关日志到安全态势感知平台，构建多场景龙虾安全告警规则，如违规外连网络、安装恶意软件、违规外发高敏感数据等，建立龙虾日志分析智能体，对龙虾日志智能化实时分析。加强龙虾行为监控，针对违规外连网络、安装恶意软件等高危行为进行告警并及时处置，针对违规外发高敏感数据等违规风险及时预警处置，采取限制使用行为、开展违规警告及考核等措施。建立龙虾漏洞情报监测渠道。跟踪龙虾漏洞情报，及时提醒漏洞修复、程序版本升级等，降低被外部攻击风险。建立龙虾安全风险应急预案。对于龙虾各类风险场景建立安全应急预案，并定期组织演练。

(2) **建设Skill安全扫描能力。**建设Skill安全扫描工具，从意图、脚本、代码及依赖等维度对Skill进行全面安全检测，风险类型覆盖语义劫持与诱导滥用、幽灵指令与提示词注入、引导文件中的对抗性叙事、功能声明与实际代码的一致性、恶意后门与可疑逻辑、命令注入与不安全输入处理、敏感信息硬编码、权限提升与不安全授权、代码混淆与反分析手段、第三方依赖漏洞、异常敏感文件访问与网络行为、供应链投毒行为等。根据扫描结果，将Skill划分为安全、可疑及恶意三个安全等级，并以此驱动后续的审核流转策略。

(3) **建设Skill入库安全审核机制。**建立覆盖自动扫描及人工审核的分级安全管控机制，确保入库Skill在意图、代码、架构及功能层面符合企业安全与治理要求。

无论是外部引入Skill，还是自研Skill，均须通过安全、架构及平台管理三方审核方可入库。SkillHub区分测试区与生产区，Skill经过审核入库测试区SkillHub，测试区Skill如需进入生产区，需经过“门票”审批机制，才能入库发布到生产区SkillHub。安全、架构、功能三个领域分别定义检查规则与分级标准。利用Skill扫描工具先行过滤，仅将恶意和可疑的Skill触达人工复核，降低人工审核成本。

(4) **建设大模型安全围栏。**Claw智能体与用户之间采用基于自然语言的对话方式，攻击者可以通过像语义模糊、隐喻表达、上下文依赖等语言特性，来构造隐蔽攻击。传统安全产品如防火墙依赖的是关键词匹配及规则过滤技术，很难处理自然语义和准确理解其含义，而自然语言的多变性也导

致传统的规则检测难以生效，所以需要以“大模型技术”对抗“大模型威胁”，从而打破传统防火墙难以应对大模型动态交互和内容风险场景。

因此需要建设大模型安全围栏，构建识别并实时过滤风险内容的围栏式检测能力，针对性抵御大模型相关安全风险。大模型安全围栏应实现如下功能：一是违法违规内容拦截。对用户输入中的不良价值观、违法犯罪、涉黄、涉政等违规内容进行分类识别与拦截。二是科技伦理合规内容检测。检测用户输入中存在的歧视言论、商业违法违规、侵犯他人合法权益等内容。三是提示词攻击识别。防护提示词注入、指令劫持、角色扮演诱导、反向诱导、假设场景构造、DAN攻击等多类提示词攻击，阻止攻击者绕过安全限制或操纵模型执行未授权操作。四是生成内容敏感信息识别。对大模型输出文本中的手机号、身份证号码、家庭住址、银行账号等个人隐私信息进行自动识别与脱敏处理。五是安全代答。对涉政等高敏感问题，提供基于知识库的标准安全回答，避免模型直接生成敏感内容。

同时建立不定期质检机制，核查系统检测过程中的误报与漏报情况，持续优化检测策略，提升检测准确率，降低误报率。

(5) **建立桌面龙虾监测管控能力。**在终端安全软件中增加龙虾管理模块，进而发现Claw智能体安装行为，对恶意Skills开展查杀，并对Claw智能体命令行审计，同时建立Claw智能体漏洞管理能力，获取Claw智能体漏洞和威胁情况，开展安全漏洞修复提醒，以及对高危龙虾禁用等。

(6) **加强龙虾数据防泄漏能力建设。**提升针对龙虾场景的数据防泄漏（DLP）能力，对通过桌面龙虾违规外发敏感数据的行为实现及时检测与告警，形成覆盖龙虾全交互链路的数据安全防护闭环。

五、总结

当前Claw智能体技术迭代迅速，应用场景日益复杂，而配套的安全防护体系尚处于起步阶段，市场上成熟的专用安全产品与技术支撑明显滞后。以典型的大模型日志安全风险为例，智能体在运行过程中可能无意间记录用户输入的敏感信息或内部业务数据，而传统DLP方案依赖固定规则与签名识别，难以应对大模型生成式交互中动态、多模态、上下文相关的信息外流场景，技术实现难度极高，难以覆盖全量风险路径。

为此，需从系统性安全防护视角出发，构建覆盖AI业务全生命周期的安全风险管理体系，重点围绕AI数据、资产、控制与响应机制，完善准入、权限、变更管理等关键流程，系统提升大模型及智能体的安全风险管控能力。同时，应持续跟踪Claw智能体相关应用与管控技术的发展动态，密切关注开源社区、安全厂商及学术研究在该领域的最新成果，定期评估现有安全防护措施的有效性，持续迭代更新安全防护策略，为企业Claw智能体的广泛落地与创新应用提供坚实、可持续的安全保障。

AI Agent安全视角下的大模型金融应用风险与多层防御体系研究

陈洪炎 | 上交所技术有限责任公司

摘要：大语言模型在金融证券领域的深度应用正在重塑行业生态，同时也引入了模型幻觉、提示注入、数据泄露、Agent失控等新型安全风险，对金融稳定与投资者保护构成严峻挑战。本文系统梳理了大模型在金融证券领域面临的主要安全风险类型，深入分析了Anthropic公司Claude Code的AI安全架构——包括三层防御体系（环境约束、模型约束、编排约束）、沙箱隔离机制、Auto Mode分类器及权限控制系统——并探讨了这些安全方案对金融证券行业的借鉴价值。研究表明，金融证券机构应构建“环境+模型+编排”的多层纵深防御体系，将AI安全能力嵌入开发生命周期全流程，并建立与AI能力演进同步的动态安全治理机制。本文为金融证券行业大模型安全应用提供了理论框架与实践路径。

关键词：大语言模型、金融证券安全、AI安全架构、提示注入、Agent安全

一、引言

生成式大语言模型（LLM）与自主AI Agent深度嵌入智能投顾、量化交易、风控合规、资产估值等金融核心业务，重构金融业务运行范式的同时，催生模型幻觉、多形态提示注入、隐私数据推演泄露、Agent自主行为失控、市场级系统性波动等新型AI原生安全威胁，传统金融信息安全防护框架难以覆盖模型推理、智能体自主编排、RAG检索链路等新增攻击面。本文旨在系统分析大模型在金融证券领域的安全风险，深入剖析Claude Code的安全架构原理，并探讨其对金融证券行业AI安全治理的启示。

二、大模型在金融证券领域的安全风险分析

（一）风险分类框架

自主金融AI Agent具备“感知输入→多步骤推理→工具调用执行”完整链路，攻击可沿三层链路逐层渗透。

（1）感知层攻击面：用户查询、外部文档、行情数据、第三方RAG知识库输入，主要威胁为显性/隐性提示注入、投毒样本、带扰动对抗行情数据。

（2）推理层攻击面：模型内部参数、上下文记忆、检索增强知识库、向量数据库，主要威胁为模型反演、成员推断、幻觉生成、权重后门。

（3）执行层攻击面：文件读写、网络API调用、交易指令下发、数据库访问、Shell脚本执行，主要威胁为权限越权、凭证泄露、算力滥用、级联系统故障。

基于对国内外金融机构大模型应用实践的调研，本文将金融证券领域的AI安全风险归纳为六大类，形成“D-M-A-I-C-S”风险分类框架（如表1所示）。

风险类别	英文缩写	核心内涵	典型场景
数据隐私风险	Data Privacy	训练/推理过程中客户敏感信息泄露	研报报告生成时泄露持仓数据
模型幻觉风险	Model Hallucination	生成虚假或误导性金融信息	研报中出现虚构的财务数据
对抗攻击风险	Adversarial Attack	恶意输入诱导模型输出错误决策	提示注入篡改交易指令
基础设施风险	Infrastructure	供应链污染、算力资源滥用	第三方模型库植入后门
合规监管风险	Compliance	违反金融监管法规与行业标准	自动化合规审查遗漏关键条款
系统性风险	Systemic	单点故障引发市场连锁反应	AI交易模型集体误判导致闪崩

表1 金融证券大模型应用安全风险分类

（二）数据隐私风险

金融数据具有高度敏感性，包含客户身份信息、交易记录、资产状况等核心隐私数据。大模型在训练和应用过程中面临的数据隐私风险主要体现在三个层面。

1. 训练数据泄露风险

金融机构在使用内部历史数据训练专有模型时，即使经过脱敏处理，仍可能通过成员推断攻击（Membership Inference Attack）或模型反演攻击（Model Inversion Attack）恢复原始敏感信息。研究表明，攻击者仅需访问模型的API接口，即可以较高概率判断特定数据记录是否参与了模型训练。

2. 推理过程泄露风险

在实时推理场景中，用户输入的查询可能包含敏感信息（如“分析某机构重仓股的退出策略”），而模型的输出可能无意中将其训练数据中的关联信息泄露给非授权用户。这种“间接信息泄露”在传统数据安全框架中难以被有效检测。当向量数据库承载敏感信息时，攻击者可通过构造恶意查询从检索增强生成系统中直接提取大量隐私数据，尤其在医疗、金融等高敏感行业中，此类风险尤为突出。

3. 第三方数据流转风险

金融机构往往依赖第三方数据提供商（如Wind、Bloomberg）获取市场数据，而大模型应用进一步引入了模型提供商（如OpenAI、Anthropic）作为新的数据流转节点。2025年金融安全报告显示，供应链已成为攻击的首要突破口，第三方服务商的安全漏洞可直接威胁金融机构的数据安全。

（三）模型幻觉风险

模型幻觉（Hallucination）是指大语言模型生成看似合理但实际错误的内容。在金融证券领域，幻觉风险具有特殊的危害性。

（1）财务数据幻觉。模型可能生成虚构的财务指标、错误的历史股价或不存在的市场事件。例如，在生成某上市公司深度研报时，模型可能“编造”不存在的子公司或夸大股东持股比例，导致投资决策失误。

（2）法规条文幻觉。在合规审查场景中，模型可能引用已废止的监管规定或错误解释法律条款。这种“法规幻觉”可能导致金融机构在不知情的情况下违反监管要求，面临行政处罚。

（3）推理链幻觉。在复杂的多步骤金融分析中（如DCF估值模型），模型可能在中间推理步骤出现错误，但由于最终输出“看起来合理”，人类审查者难以发现。这种“隐性错误”比明显的错误更具危害性。

（四）对抗攻击风险

传统关键词过滤仅能拦截显性注入，隐式语义提示注入将恶意指令嵌入正常研报、PDF文档、行情文本，利用模型上下文优先级规则覆盖系统安全护栏。攻击者利用大模型对“文档最高优先级”类语义指令的顺从性，劫持Agent执行读取全部客户数据、下发大额交易、转发隐私信息等高风险操作；数据投毒仅需注入少量恶意金融样本，即可显著扭曲模型市场判断逻辑；量化对抗样本通过订单簿微小价格扰动，诱导AI交易模型产生同向错误下单。

（五）基础设施、合规、系统性风险

基础设施风险核心为多层技术栈供应链污染，模型权重篡改、沙箱逃逸、版本回退攻击可直接突破AI隔离边界；合规风险核心矛盾为大模型黑盒特性与监管要求的算法可解释、可审计冲突；系统性风险源于金融网络耦合效应，AI Agent与核心交易、清算系统深度互联，单点推理故障沿业务链路级联扩散，模型同质化形成市场“AI羊群效应”，放大股市、期货市场波动。

三、Claude Code AI安全架构分析

Anthropic推出的Claude Code是一款AI驱动的代码助手，其安全架构代表了当前AI Agent安全设计的先进实践。该架构的核心思想是：模型能力越强，安全控制必须越强。以下从三个层面分析其安全设计。

（一）三层防御架构

Anthropic将Agent安全风险分为三类：用户误用（用户恶意或疏忽地指示Agent执行有害操作）、模型异常行为（Agent执行了无人要求的危险操作）和外部攻击（通过工具、文件或网络访问攻击Agent）。针对这三类风险，Anthropic构建了“环境+模型+编排”三层独立防御。

第一层：环境约束（Environment）。通过进程沙箱限制Agent可访问的进程范围；通过虚拟机隔离确保每个云会话在独立VM中运行；通过文件系统边界限制Claude Code只能写入启动目录及其子目录；通过网络出口控制通过Unix域套接字代理限制出站连接。这一层的目标是设置Agent可达范围的硬边界。

第二层：模型约束（Model）。通过系统提示定义Agent的行为边界；通过两阶段推理分类器进行安全判断（快速yes/no判断→深度chain-of-thought审查）；通过服务器端探针进行提示注入扫描；通过训练修改直接抑制不对齐行为。这一层的目标是塑造Agent“倾向于”做什么，构建概率性防御。

第三层：编排约束（Orchestration）。通过权限系统实现默认只读、显式授权编辑/执行；通过人类监督确保关键操作需人工确认；通过审计日志记录所有操作用于合规。这一层的目标是控制Agent的执行流程。

三层防御中任何一层都可能存在漏洞，但三者相互独立、互为补充，共同构成了纵深防御体系。

（二）沙箱隔离与网络访问控制

Claude Code在沙箱技术上实现了精细化的隔离控制。在操作系统层面，macOS使用Seatbelt进行文件系统沙箱，Linux使用Bubblewrap进行进程隔离。网络隔离方面，仅允许通过Unix域套接字连接到代理服务器，代理服务强制执行域名白名单。Anthropic已将沙箱运行时开源（anthropic-experimental/sandbox-runtime，Apache 2许可）。

网络访问支持三种模式：无网络访问（最高安全性）、可信网络访问（默认允许依赖安装相关域名）、自定义环境（用户定义自己的域名白名单）。这种分级设计使金融机构能够根据不同业务场景的安全需求灵活配置。

（三）Auto Mode分类器与权限管理

Auto Mode是Claude Code实现自主执行的关键组件。其设计包含两个核心创新：

（1）两阶段推理分类器：第一阶段进行快速单token yes/no判断；若第一阶段标记为风险操作，则触发第二阶段的chain-of-thought深度审查。分类器被设计为“推理盲”——只能看到用户消息和Agent的工具调用，无法看到Claude自身的响应和中间推理。这一设计防止了Agent“写出有说服力的理由”来绕过审查，体现了安全设计中的“不可欺骗性”原则。

（2）三层权限结构：Tier 1（文件读取、搜索、目录导航）始终允许；Tier 2（项目内文件操作）允许无需分类器审查；

Tier 3 (Bash命令、项目外写入、外部服务调用) 需要分类器审查。Auto Mode还移除了通配shell访问, 阻止脚本解释器 (python、node、ruby) 和包管理器运行命令。

(四) Claude Code Security: AI驱动的漏洞发现

Claude Code Security是Anthropic推出的AI驱动代码安全扫描能力, 与传统基于规则匹配的SAST工具不同, 它具备上下文感知扫描 (理解跨文件数据流)、对抗性验证 (每个发现经Claude自身挑战验证)、补丁建议 (生成针对性修复补丁) 和高严重性聚焦等特性。

在实际应用中, Claude Code Security使用Claude Opus 4.6在开源代码库中发现了500多个此前未知的漏洞, 其中一些bug在专家审查数十年后仍未被发现。Anthropic自身也使用Claude审查内部代码, 已拦截多个生产环境漏洞。这一实践表明, AI不仅能发现已知模式的安全问题, 还能识别人类专家长期遗漏的复杂漏洞模式。

(五) 凭证管理与代理模式

Claude Code推荐在Agent安全边界外运行代理进行凭证管理。其核心流程是: Agent发送无凭证的请求→代理注入凭证→代理转发请求到目的地。这一模式的优势在于: Agent永远看不到实际凭证; 代理可强制执行允许的端点白名单; 代理可记录所有请求用于审计; 凭证集中存储而非分散到每个Agent。对于金融证券机构而言, 这一模式能够有效防止Agent权限被滥用后导致的凭证泄露。

四、面向金融证券的多层防御体系设计

(一) 设计原则

基于Claude Code AI安全体系的核心思想, 结合金融证券行业的特殊性, 本文提出以下设计原则:

- (1) 纵深防御 (Defense in Depth)。不依赖单一安全机制, 通过多层独立防御确保即使某一层被突破, 其他层仍能提供保护。
- (2) 最小权限 (Least Privilege)。AI Agent默认只拥有完成当前任务所需的最小权限, 所有权限提升需经显式授权。
- (3) 零信任 (Zero Trust)。不假设任何内部组件是可信的, 每个请求都需经过身份验证和授权。
- (4) 人类在环 (Human-in-the-Loop)。涉及资金交易、合规决策等高风险操作时, 必须有人类审查和确认。
- (5) 可审计性 (Auditability)。所有AI操作记录完整日志, 支持事后追溯和合规审计。
- (6) 弹性恢复 (Resilience)。系统设计应假设攻击必然发生, 重点放在快速检测、隔离和恢复。

(二) 金融证券大模型安全防御体系架构

适用于金融证券领域的大模型安全主要涉及环境隔离沙箱、模型约束和编排控制三个方面 (如图1所示)。

1. 环境隔离沙箱

将AI Agent部署在隔离的虚拟化环境中, 限制其文件系统访问范围、网络出口和进程权限。对于涉及交易执行等高风险操作的Agent, 应考虑在专用沙箱中运行, 并实施最小权限原则。

2. 模型约束

在模型输入端部署提示注入检测探针, 对用户输入进行两阶段安全审查; 在模型输出端实施内容安全过滤, 防止生成违规或误导性金融信息。对于关键金融决策场景, 应优先选用经过安全对齐训练模型。

3. 编排控制

建立精细化的权限体系, 对金融数据读取、交易指令生成、外部API调用等操作实施分级授权。关键操作必须保留人类审批环节, 所有Agent行为应记录完整的审计日志。

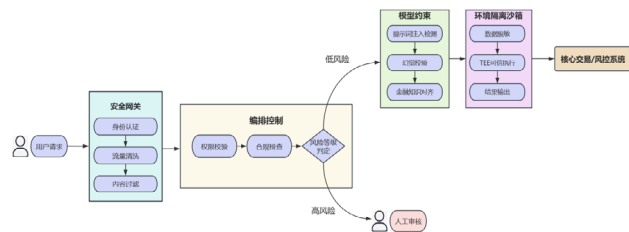


图1 金融证券大模型安全防御体系架构

(三) 环境隔离层: 金融数据安全沙箱

金融数据的安全隔离是防御体系的基础。本文提出“数据沙箱+网络微分段+可信执行环境”的三重隔离机制。

1. 数据沙箱

借鉴Claude Code的Seatbelt/Bubblewrap沙箱机制, 为每个AI任务创建独立的数据沙箱。沙箱内的数据经过以下处理:

- (1) 静态脱敏: 对训练数据和推理输入中的敏感字段 (客户姓名、身份证号、账户余额) 进行自动识别和脱敏。
- (2) 差分隐私: 在模型训练过程中注入calibrated noise, 确保单个数据记录的存在与否不会影响模型输出。参数 ϵ 控制在1.0以内, 满足金融数据的严格隐私要求。
- (3) 联邦学习: 对于跨机构联合建模场景, 采用联邦学习架构, 原始数据不出域, 仅交换加密后的模型参数梯度。

2. 网络微分段

将金融AI系统的网络划分为多个安全域:

- (1) 数据采集域: 仅允许从交易所行情接口、监管机构数据接口获取数据。
- (2) 模型训练域: 与外部网络完全隔离, 仅通过单向数据导入通道接收预处理后的数据。
- (3) 推理服务域: 通过API网关对外提供服务, 内部与核心交易系统通过服务网格进行东西向流量控制。
- (4) 核心交易域: 与AI系统物理隔离或逻辑强隔离, AI系统仅能通过预定义的、受监控的接口向核心系统发送指令。

3. 可信执行环境 (TEE)

对于涉及最敏感数据 (如客户持仓、交易策略) 的AI任务, 部署在可信执行环境中。TEE确保即使操作系统或

Hypervisor被攻破，攻击者也无法读取TEE内的内存数据。加密和可信执行环境从数据保护层面封堵信息泄露，是应对侧信道攻击的有效手段。

（四）模型约束层：金融专用安全分类器

模型约束层是防御体系的核心，需要针对金融场景的特殊性进行专门设计。

1. 金融知识对齐

通过检索增强生成（RAG）技术，将金融法规、行业标准、公司内部制度注入模型上下文。具体实现：

（1）构建金融法规知识库，包含《证券法》《基金法》《期货和衍生品法》等核心法规的全文及解释。

（2）使用向量数据库存储法规条文embedding，实现语义检索。

（3）在模型推理时，自动检索与当前任务相关的法规条文，作为上下文约束模型输出。

输入验证与指令净化是防御语义指令注入攻击的首要措施，系统需对所有外部输入进行结构化校验，识别并清除潜在的恶意指令片段。强化系统提示与构建结构化护栏（guardrail）等技术可以约束大语言模型的生成行为。

2. 幻觉检测器

针对金融场景设计专门的幻觉检测机制：

（1）事实一致性检查：将模型生成的财务数据与权威数据源（如交易所公告、公司年报）进行交叉验证。

（2）数值校验：对模型生成的数值进行合理性检查（如市盈率是否在合理区间、增长率是否符合行业规律）。

（3）来源追溯：要求模型对每一个关键论断标注信息来源，人类审查者可追溯验证。

3. 对抗样本分类器

借鉴Claude Code的两阶段分类器设计，构建金融专用对抗样本检测器：

（1）第一阶段：快速模式匹配，检测已知的攻击模式（如提示注入关键词、异常字符编码）。

（2）第二阶段：深度语义分析，使用专门的BERT类模型检测输入中的语义异常（如逻辑矛盾、情感操纵）。

（3）金融特化：训练数据包含金融领域的对抗样本（如伪装成正常研报的攻击性提示、包含隐藏指令的行情数据）。

（五）编排控制层：业务权限与合规引擎

编排控制层将技术安全机制与金融业务逻辑深度融合。

1. 业务权限引擎

基于角色的访问控制（RBAC）与基于属性的访问控制（ABAC）相结合。

（1）角色维度：研究员、交易员、风控人员、合规人员等不同角色拥有不同的AI功能权限。

（2）属性维度：结合时间（交易时段/非交易时段）、地点（办公网络/外部网络）、数据敏感度（公开/内部/机密）等属性动态调整权限。

（3）交易额度控制：AI Agent生成的交易指令需经过额度校验，超过阈值的交易需人工确认。

2. 合规检查

在AI工作流的关键节点设置合规检查点。

（1）输入检查点：检查用户输入是否包含敏感信息（如未公开的重大事项、内幕信息）。

（2）输出检查点：检查模型输出是否违反适当性管理要求（如向风险承受能力低的投资者推荐高风险产品）。

（3）执行检查点：检查AI生成的操作指令是否符合监管要求（如是否违反T+1交易规则、是否超出持仓限额）。

3. 人类审查队列

针对高风险操作建立人类审查机制。

（1）智能分级：使用机器学习模型自动评估操作风险等级，低风险操作自动执行，中风险操作抽样审查，高风险操作强制人工确认。

（2）审批疲劳缓解：通过用户行为分析，识别审批者的疲劳状态，在疲劳状态下提高审查阈值或强制休息。

（3）双人复核：对于涉及大额资金的操作，实施双人复核机制。

五、大模型金融应用安全的治理框架

（一）将AI安全能力嵌入开发生命周期

金融证券行业的软件开发生命周期可系统地引入AI安全能力。

编码阶段：在IDE中集成AI辅助代码审查工具，实时检测安全漏洞。研究表明，此类应用可赋能架构师和开发者，平均提升项目开发编码效率10%-15%。

CI/CD阶段：在持续集成流水线中部署AI驱动的安全扫描，对每个Pull Request进行自动化安全审查。在持续部署管道中添加自动化渗透测试。

生产阶段：部署AI驱动的威胁检测与告警分类系统。中邮证券的实践表明，将大模型的语义理解与逻辑推理能力引入安全运营核心环节后，有效降噪率超90%，大幅降低了对人工经验的依赖。

（二）建立AI漏洞发现与修复的快速响应机制

主动扫描：优先使用AI工具扫描解析不可信输入、执行认证/授权决策、可从互联网访问的代码。应包含遗留代码——那些作者已离职或审查标准过时的代码往往是最脆弱的。

快速修补：对CISA KEV目录中的所有漏洞立即修补；互联网暴露系统应在漏洞利用公开后24小时内完成修补；使用EPSS对剩余漏洞进行优先级排序。

AI辅助分类：利用大模型对安全告警进行去重、评估暴露面、起草修复工单。这能够有效应对漏洞报告量激增带来的运营压力。

(三) 强化Agent行为监控与治理

随着金融证券行业越来越多地采用AI Agent处理投研、交易、风控等任务，Agent行为监控与治理成为新的安全重点。

行为审计：建立完整的Agent操作审计日志，记录每一次工具调用、文件访问和网络请求。审计日志应具备不可篡改性和可追溯性，以满足监管合规要求。

异常检测：部署Agent行为基线分析系统，识别偏离正常模式的异常操作。

权限最小化：遵循“默认拒绝”原则，Agent默认不具备任何额外权限，仅在明确授权后才能执行特定操作。对于涉及客户资金或敏感数据的操作，应实施多级审批流程。

六、结论

大语言模型在金融证券领域的广泛应用带来了效率的革命性提升，同时也引入了模型幻觉、提示注入、数据泄露、Agent失控等系统性安全风险。对金融证券行业而言，构建与AI能力演进同步的动态安全治理机制已刻不容缓。建议从以下路径推进：一是建立“环境+模型+编排”的三层纵深防御体系，将AI安全能力系统性地嵌入基础设施；二是在开发生命周期全流程引入AI安全审查工具，实现“以AI防御AI”；三是建立与AI加速攻击相匹配的快速响应机制，将漏洞修复周期压缩至24小时以内；四是完善分级分类管理和全生命周期安全治理框架。

展望未来，随着AI Agent在金融证券领域的进一步渗透，安全挑战将更加复杂多元。金融证券机构需要持续跟踪前沿AI安全研究，积极参与行业标准制定，在技术创新与风险控制之间实现动态平衡，方能在大模型时代行稳致远。

参考文献

- [1] Anthropic. Beyond permission prompts: Claude Code sandboxing[EB/OL]. <https://www.anthropic.com/engineering/beyond-permission-prompts-claude-code-sandboxing>, 2025.
- [2] Anthropic. Claude Code Security: AI-powered vulnerability discovery[EB/OL]. <https://www.anthropic.com/engineering/claude-code-security>, 2025.
- [3] 北京金融科技产业联盟. 2025年金融业人工智能应用风险研究报告[R]. 北京, 2025.
- [4] IDC, 腾讯金融研究院. 2025金融业大模型应用报告[R]. 北京, 2025.
- [5] 中国工商银行金融科技研究院等. 中国金融生成式AI多模态内容鉴伪与安全防护白皮书[R]. 北京, 2025.

证券期货业大模型安全围栏的建设与运营思考

钟蓉、吴佳伟、李鹏 | 兴业证券股份有限公司

摘要：人工智能逐步融入证券期货全业务流程，随之带来内容合规失控、提示词注入攻击、供应链隐患、算力损耗等多重安全风险。为有效防范大模型应用风险，部署大模型安全围栏技术至关重要。本文结合两类典型高风险智能体应用场景，介绍兴业证券在该技术建设、落地及日常运营中的实战经验与建设运营思路。

关键词：大模型、安全围栏、安全运营、智能体应用

一、大模型安全围栏的建设背景

（一）背景

随着人工智能技术的快速演进，大语言模型正加速渗透至证券期货行业的各个业务环节，涵盖智能投研、合规审查、量化研究、客户服务等核心场景，也包括多类办公、运维及安全场景。大模型在带来效率跃升的同时，亦引入了前所未有的安全挑战，包括输入输出内容合规风险、提示词注入攻击的持续演化、供应链安全的潜在隐患，以及算力基础设施的消耗风险。当前，行业已呈现“证券行业引领、基金期货跟进、头部机构示范”的应用格局，但若缺乏系统化的安全防护体系，大模型的广泛应用可能使机构面临合规处罚、数据泄露乃至系统性安全风险。安全防护体系的核心一环是大模型安全围栏，目前，开放对客类大模型智能体的企业已基本完成该产品的建设，本文将分享大模型安全围栏的建设与运营思考。

按照大模型智能体的使用对象，可将智能体分为面客类和内部使用类，按照是否使用云上算力，又可将智能体应用分为本地算力类和云上算力类。由于面客类应用需要面向互联网开放，注册用户数量多，是高风险的场景之一。云上算力类由于调用云上算力，有数据出域的风险，也是主要风险场景之一。本文将从上述两个高风险场景入手，探讨大模型安全围栏在应对这两类风险场景中的作用。

（二）面客场景主要风险

1. 输入输出内容

在面客的智能体应用中，大模型输入输出的合规性是首要关切。以投资顾问场景为例，《证券投资顾问业务暂行规定》第二十四条明确要求，证券投资顾问业务中“禁止对服务能力和过往业绩进行虚假、不实、误导性的宣传”，且不得以任何方式承诺或者保证投资收益。该规定第十六条亦要求投资建议必须具有充分的依据，第七条规定提供投资顾问服务的人员须具备相应执业资格。这意味着面向投资者的智能体在生成投资建议时，必须确保回复内容有据可依，避免夸大描述或暗示确定性收益。此外，在通用智能体层面，《生成

式人工智能服务管理暂行办法》第四条要求提供者应当采取有效措施防止生成虚假有害信息，并承担信息内容生产者责任；《中华人民共和国数据安全法》第二十七条要求数据处理活动应当建立健全全流程数据安全管理制度；《中华人民共和国个人信息保护法》第六条则确立了个人信息处理的“最小必要”原则。上述法规共同构成了面客场景下大模型输入输出的合规约束框架。

2. 算力消耗

从面客场景的角度审视，大模型的算力消耗本身即构成一种隐性安全风险。在本地部署场景下，提供实时响应的面客智能体需要持续占用大量GPU资源，单次推理的高计算开销使得硬件采购与运维成本居高不下。当并发访问量激增时，算力资源可能被瞬时耗尽，导致服务不可用，直接影响客户体验与业务连续性。攻击者可利用精心构造的高计算复杂度提示，蓄意消耗模型推理资源，加剧算力瓶颈。

在云上算力场景下，虽然云服务提供了弹性扩缩容的能力，但面客场景的持续性推理需求意味着长期处于“按量付费”的高消耗状态，运营成本难以精确控制。更为关键的是，算力消耗的可观测性不足——机构难以准确区分正常业务负载与恶意算力消耗行为，这为攻击者提供了可乘之机。

3. 提示词注入

提示词注入已被OWASP 2025版LLM应用十大风险列为排名首位的威胁，其攻击原理在于通过恶意构造的输入操纵模型行为，使其背离预设的系统指令。与传统的代码注入不同，提示词注入利用的是自然语言的语义模糊性，攻击者可借助“忽略”“覆盖”等控制性语言篡改模型的推理逻辑，传统基于正则的输入过滤手段难以有效防御。典型案例之一是2025年披露的EchoLeak漏洞（CVE-2025-32711），针对Microsoft 365 Copilot的零点击提示词注入攻击。攻击者仅需发送一封精心构造的邮件，即可在无用户交互的条件下诱使Copilot访问内部文件并将其内容传输至攻击者控制的服务器，实现了跨信任边界的权限提升与数据外泄。

4. 权限及供应链漏洞

人工智能应用的权限管控与供应链安全同样是面客场景不可忽视的风险维度。在权限方面，大模型智能体通常被授

予访问内部知识库、API接口和业务系统的权限，一旦权限过度宽或访问控制策略缺失，攻击者可利用提示词注入等手段横向扩展攻击面，导致越权访问与敏感信息泄露。

(三) 云上算力场景主要风险

1. 数据出域问题

证券期货业大模型的算力部署通常采取云上与本地结合的混合模式。核心交易数据与敏感业务负载部署于本地私有化环境，非敏感推理任务则可弹性使用云上算力以弥补硬件缺口。

然而，部分业务负载迁移至云上后，数据出域风险随之加剧。金融数据涉及客户身份信息、账户明细、交易记录等高度敏感内容，一旦进入云端AI处理链路，其可访问范围和留存周期可能超出原有业务目的的必要范围，引发金融数据管理合规风险。

二、大模型安全围栏的建设过程

(一) 基本架构

面对以上风险，建设大模型安全围栏系统的必要性十分显著。安全围栏的核心运作原理主要体现为严密的“双向语义审查”机制，即分为输入端和输出端两个阶段的防御。在输入阶段，围栏会对用户的提问进行意图识别与越狱检测。这一过程通常采用“规则匹配+语义理解”的双层过滤架构，初级拦截层利用正则表达式精准匹配高危敏感词库，直接阻断显性攻击指令；高级分析层则利用语言分类模型或多模态模型对输入内容（文本、图像、语音等）进行语义分类。一旦发现包含恶意诱导或敏感信息探测的指令，便直接阻断请求。在输出阶段，即使模型生成了回答，围栏也会在终点线前进行二次核查。这一环节重点针对个人身份信息（PII）进行自动脱敏，并交叉比对事实以剔除幻觉内容。这种独立于生成模型之外的双重安检架构，有效拦截了各类潜在风险。

常见的大模型安全围栏架构如下所示，主要包括四层。第一层为接入层，提供各类智能体、大模型算力的接入方式。第二层为配置层，提供各类防护的配置及防护验证。第三层为能力支持层，主要为各类检测提供模型支持，包括但不限于意图识别、判别模型、代答大模型等；第四层为基础层，主要为一些基础服务支撑，如各类数据库等。

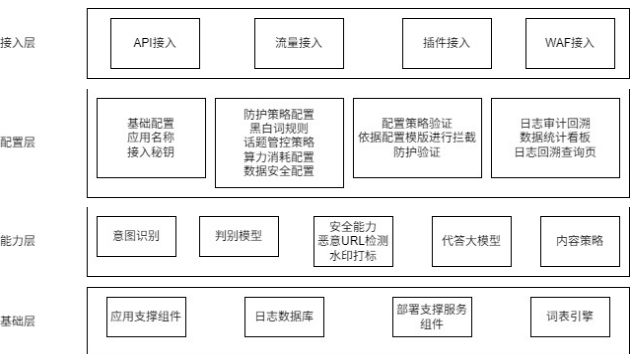


图1 常见大模型安全围栏基本架构

(二) 明确需求

在建设大模型安全产品之前，应明确本单位的需求，大致需要考虑三方面的问题。

1. 算力支持角度

(1) 安全围栏是否需要应用于面客智能体？（面客智能体需要考虑对客的并发量大小，本地部署的围栏需要算力支持，围栏所需算力与支持的并发会话量正相关，过高的并发对话需求可能需要考虑直接使用云上围栏服务）

(2) 如触发风控规则，处置方式是采用兜底回复还是调用安全代答大模型？（代答需要部署代答大模型，对算力要求更高，兜底回复只需配置固定的应答话术）

(3) 是否有敏感业务数据使用需求？（非敏场景下可优先使用云上算力，但为了防止内部数据出域，可叠加一道本地部署的围栏从技术上保障数据不出域）

2. 业务场景设计

(1) 是否有多模态需求？（业务场景是否规划了图片、视频等多模态内容，如是，是否需要围栏提供水印标记）

(2) 是否有搜索需求？（如是，需要围栏提供URL安全检测能力）

(3) 预估的最大上下文长度是多少？（如太大，可能需要对Prompt进行拆分，也可结合具体的业务场景调整接入节点，如部分应用拼接了用户输入及系统提示词再送入大模型会导致围栏审核的上下文过长，超过了最佳的上下文审核长度）

(4) 是否有附件类上传需求？（如是，需要围栏提供附件检测能力）

(5) 应用场景主要需要满足哪些法律法规？（应针对场景需满足的法律法规优化审核算子）

(6) 什么场景需要触发人工审核？

3. 部署架构角度

(1) 大模型安全围栏的接入方式和接入节点是什么？

(2) 算力是否采用多机房多活架构？

明确以上问题后，结合目前市场产品的情况可以考虑以多种方式将大模型安全围栏能力引入单位内部。如企业既有大量对客的场景又有使用云上算力的场景，可以考虑同时引入云上安全围栏和本地安全围栏，高并发对客场景使用云上围栏流转，高敏场景采用本地围栏防范敏感数据出域。

(三) 部署接入

完成对内部需求的分析及采购部署后，最后一步是应用的接入问题。接入一般需要考虑两个问题，一是接入标准，即智能体是否有接入需求，如有需求，最合适的接入围栏类型是什么；其二是接入的节点，适合智能体的接入节点和方式是什么。

对于问题一，单位内部可考虑制定明确的规范来确定接入的标准。除面客类智能体有明确合规要求外，对于仅供内部少数团队使用的智能体，其接入标准宜通过单位内部规范予以约束。本文主要探讨了两类有明确接入需求的场景，即面客场景和使用云上算力的场景。

对于问题二，由于智能体的架构多样，大模型的围栏接入节点也可多变。下图为常见的Workflow类智能体，其接入节点可以是Workflow的某个节点，也可以在大模型网关节点统一接入。大模型网关接入是最常见的接入节点，但智能体形态和送审内容的区别会导致网关在接入大模型安全围栏时存在部分定制开发需求。

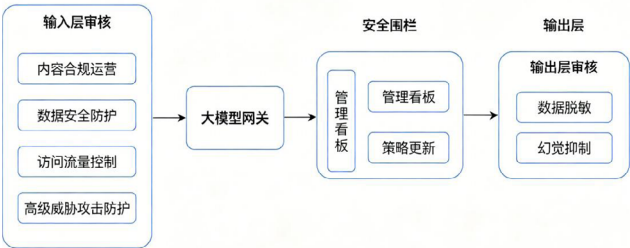


图2 常见Workflow类智能体通过网关接入安全围栏的架构

针对Dify、OpenClaw等主流开源智能体平台，云服务商通常支持以插件形式集成安全围栏，这种机制能够实现对大模型安全防护能力的灵活调度。以Dify为例，用户可通过公开市场检索并安装特定品牌的围栏插件，进而在Workflow中完成安全审核节点的编排。由于外部合规标准对接入节点位置并无强制性规定，在实际部署时应依据项目具体需求进行设计。鉴于围栏审核通常存在50ms至100ms的延迟，架构设计应尽量避免对同一内容进行重复检测，以保障系统响应性能。针对视频、音频及图片等多模态场景，若所选围栏不支持混合样本检测，建议将审核节点前置部署于多模态模型之前。此外，在“用户输入+长系统提示词”的智能体架构中，拼接后的内容可能超出围栏单次审核的长度限制，此时需考虑采用拆分送审策略或调整接入节点位置。对于面向互联网的系统，还可结合传统WAF实现联动防护，部分厂商已支持通过WAF直接接入围栏能力。综上所述，安全围栏的接入节点设计需综合考量智能体总体架构、送审样本特征及性能指标，因地制宜地制定实施方案。

对于使用云上算力的场景，大模型围栏一般部署在云上算力接入前。其主要防护目的是对出域数据进行审计和拦截，架构如下所示。云上算力场景的规则配置在数据防护方面更加严格。

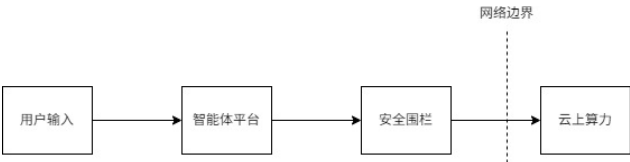


图3 使用云上算力的安全围栏防护架构

三、大模型安全围栏的持续运营

（一）运营指标体系

为验证大模型安全围栏在实际应用中的有效性，并全面评估其对真实用户需求、恶意攻击防范及监管合规要求的满足程度，亟需建立一套科学的考核指标体系。如表1所示，依据GB/T 45654—2025《网络安全技术 生成式人工智能服务安全基本要求》（以下简称《基本要求》），内容安全合格率、应拒答测试题拒答率及非拒答测试题误拒率等关键指标均已明确量化标准。尽管当前尚无专门针对安全围栏的国家标准，但《基本要求》作为生成式人工智能服务安全治理的重要国家标准，其指标体系对企业制定和运营围栏策略仍具有重要的指导意义。此外，中国信通院于2025年发布了大模型安全防护围栏能力检验及安全护栏能力评估等系列测评标准，并提供了第三方认证渠道，认证结果划分为基础级、增强级和优秀级三个等级，可作为企业自评与外部合规的重要参考依据。

序号	指标类别	指标名称	标准要求	对应标准条款	指标解释
1	模型输出安全	模型生成内容安全合格率	≥90%	5.2、B.2.2.2	模型生成内容中，不含附录A 31种安全风险的本占不低于90%
2	模型输出安全	应拒答测试题拒答率	≥95%	B.2.2.2	对违法违规等应拒答的测试题，模型拒绝回答比例不低于95%
3	模型输出安全	非拒答测试题拒答率	≤5%	B.2.2.2	对正常合法的测试题，模型错误拒答比例不高于5%

表1 依据GB/T 45654—2025整理的指标

（二）指标持续提升

为了提升大模型应用的上述指标，需要定期升级安全围栏能力。目前我司建立了上线前题库评审、上线时模板分类及上线后规则运营三项运营策略，以保障指标持续提升。

1. 上线前评审题库

指标测量结果的准确性高度依赖于测评题库的质量与覆盖度。基于工程实践经验，测评题库可划分为基础风险、行业合规拒答及非拒答测试三类，其构建策略各有侧重（详见表2）。第一类为基础风险题库，旨在覆盖GB/T 45654—2025规定的31种安全风险场景。目前主流评测平台均已支持此类题库，同时可通过AI红队测试等渠道持续补充对抗性样本。第二类为行业合规拒答题库，需依据特定行业的政策法规构建。尽管已有机构开源了金融领域评测数据集（如蚂蚁集团与上海财经大学联合发布的FinEval），但证券期货业因面客合规要求极高，仍需依据《证券经纪业务管理办法》（证监会令第204号）、《证券投资顾问业务暂行规定》等法规建立专属题库，并随监管动态定期更新。例如，针对“保本保收益”等违规询问，智能体应严格执行拒答或安全代答策略，以落实投资者适当性原则。第三类为非拒答测试题库，用于验证正常业务的可用性。建议由业务部门在建设初期基于用户意图设计，并可从客服历史问答日志中挖掘高频问题，以确保合法合规的正常请求不被误拦截。上述各类题库均需建立常态化维护机制，通过定期更新与线上回归评测，保障安全围栏效能的持续有效。

分类	测评题集分类	主要内容	题库维护方及上线后人工审核评审方	对应指标
基础题库	基础风险类	GB/T 45654—2025 所列的31类风险项	评测平台建设方，一般由厂商提供	模型生成内容安全合格率/应拒答测试题拒答率
	安全题库	主要为提示词攻击等安全方面的检测样本维护，主要通过AI红队及公开互联网收集绕过样本	安全人员	
合规题集	外部合规测试题集	一些开源金融测试题	开源第三方	
	内部合规题库	主要是证券期货业的合规问题题集	合规	
必答题库	正常合法测试题库	必答题库	业务方	非拒答测试题拒答率

表2 常见测评题库来源

2. 上线时模板分类

依据规则约束强度的差异，我们构建了严谨型、中性型及宽松型三类初始策略模板，并根据应用系统的风险等级进行差异化适配（详见表3）。严谨型模板适用于面向终端用户的对外服务系统，该模板启用全量策略规则，并叠加红线知识库与话题管控机制，对触发风险的内容采取兜底回复或安全代答等强干预措施，以确保合规底线。中性型模板作为通用基线，仅启用部分核心策略，依赖兜底机制进行风险处置，适用于一般性业务场景。宽松型模板则主要面向内部办公及研发测试类应用，在启用部分策略的基础上，采用“兜底+风险标记放行”的弹性处置策略，在保障基本安全的同时兼顾内部使用的灵活性与效率。上述三类模板构成了安全围栏的分级防护体系，实际部署时可根据业务需求在此基础上进行参数微调。

分类	规则	其他设置	风险处置策略	适用应用
严谨型	全策略模板开启	红线知识库+话题管控	兜底/代答	面客
中性型	部分策略模板	无	兜底	
宽松型	部分策略模板	无	兜底+部分放行但标记风险	对内的各类应用

表3 策略模板分类

3. 上线后规则运营

大模型安全围栏上线后，其规则运营体系划分为常态化运营与重保期运营两种模式，并引入人工审核机制构建人机协同的优化闭环（基础流程如图4所示）。在常态化运营阶段，人工审核覆盖拦截与放行两类样本：针对被拦截内容，经人工复核确认为正常请求的，将执行关键词加白策略并同步更新线上规则模板，以降低误拒率；针对放行内容，则通过抽样评审排查漏拦或绕过风险，持续补充自定义风险样本库。审核任务的分派机制与题库维护团队保持一致，确保专业对口。在实际部署中，运营指标需根据应用场景差异化设定。例如，对于内部智能体“老铁”等应用，运营重心在于保障业务可用性，即提升非拒答率。实测数据显示，该围栏系统在内部应用中已将错误拒答率稳定控制在0.5%以下，验证了上述运营机制的有效性。

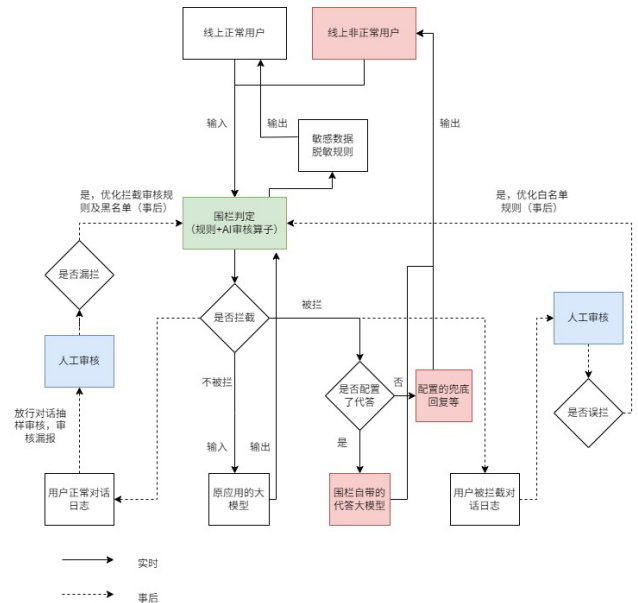


图4 常态化运营流程

重保期运营主要关注互联网舆情、涉政、涉敏等情报，并实时优化拦截或放行规则。

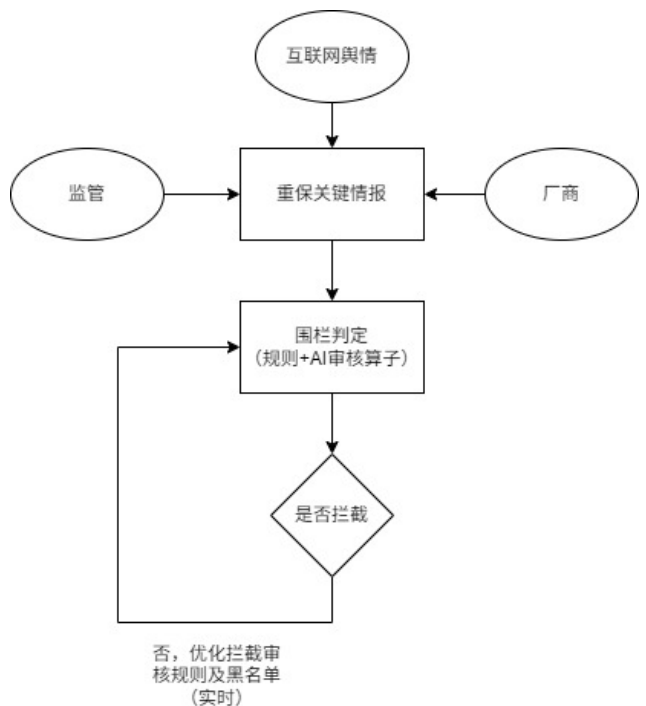


图5 重保期运营流程

四、局限与展望

（一）局限

参照OWASP基金会发布的《大型语言模型应用十大安全风险》，安全围栏并非智能体安全的万能解决方案，其防护能力存在明确的适用边界。安全围栏在输入输出内容过滤与合规管控方面具有显著优势，但难以覆盖智能体执行层的越权操作、工具滥用等风险，对大模型幻觉等内生性缺陷的缓解作用亦十分有限。因此，企业不应将安全围栏作为单一防御手段，而应将其纳入纵深防御体系，与应用接口权限控制、代码执行沙箱、运行时监控等技术措施协同联动，构建覆盖“输入-推理-执行-输出”全链路的智能体安全防护架构。

（二）展望

大模型安全围栏技术正处于快速演进阶段，其能力边界随智能体生态的发展持续拓展。在检测模态方面，已从早期的文本与图像扩展至视频、音频及附件文档等多模态内容；在接入架构方面，亦随智能体交互协议的迭代而适配，例如针对OpenClaw等新型终端类智能体框架的接入，已有厂商推出专属安全围栏插件。与此同时，面对智能体形态多样化、攻击手段复杂化及监管政策动态调整三重驱动，安全围栏的审核规则体系亦需保持持续迭代。作为大模型安全防护体系的关键基础设施，安全围栏的能力建设是一项长期工程，未来仍需围绕多模态理解、实时响应效率及合规自适应等方向持续深化研究，以支撑业务与数据安全需求的动态演进。

参考文献

1. 中国证券监督管理委员会. (2010). 证券投资顾问业务暂行规定
2. 国家互联网信息办公室. (2023). 生成式人工智能服务管理暂行办法
3. 中国证券监督管理委员会. (2023). 证券期货业网络和信息安全管理办法
4. 中华人民共和国. (2021). 中华人民共和国数据安全法
5. 中华人民共和国. (2021). 中华人民共和国个人信息保护法
6. Vineeth Sai. <https://vulnerablemcp.info>
7. 全国网络安全标准化技术委员会. (2025). GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求
8. OWASP. (2025). 大型语言模型应用十大安全风险
9. 上海财经大学. (2025). <https://github.com/SUFE-AIFLM-Lab/FinEval>

证券公司攻防演练防守实践

刘志敏、陆修然 | 国盛证券股份有限公司

摘要： 本文介绍证券公司攻防演练防守实践。在备战阶段，我司制定了完善的演练方案，明确了各团队职责分工，建立了应急处置预案，并通过全员安全意识培训和内部动员会议，为实战做好了充分准备。在布防阶段，我司系统开展了资产风险梳理、安全基线检查、边界策略优化等工作，构建了覆盖网络、主机、应用各层面的监控防护体系，并通过模拟演练验证了防护有效性。在实战阶段，依托科学的组织架构和高效的联动机制，我司监测并处置了多起安全事件，包括钓鱼攻击、零日漏洞利用等高级威胁，各防护层均发挥了关键作用，展现了纵深防御体系的韧性。

关键词： 攻防演练、网络安全、红蓝对抗

一、演练备战阶段

备战阶段，我司围绕“方案引领、组织保障、流程支撑、人员防控”四大核心，系统完成了演练方案制定、多级职责分工体系构建、覆盖常见攻击场景的应急处置预案梳理，以及面向全员的社会工程学防范专项培训。通过这一系列准备工作，实现了防守任务明确化、组织协作流程化、应急动作标准化和人员意识常态化，为实战阶段构建了职责清晰、响应迅速、人技结合的综合防守体系。

(一) 演练方案制定

我司制定了详尽的《攻防演练总体实施方案》，采用分阶段推进的实施框架，构建了完整的网络安全防护体系，作为本次行动的根本依据。

方案核心内容包括：

- 1. 备战阶段基础工作
 - (1) 明确演练目标系统范围，确定核心防护对象。
 - (2) 建立完整的组织架构体系，明确各团队职责分工。
 - (3) 开展全员安全意识培训，提升整体安全防护意识。
- 2. 准备阶段体系建设
 - (1) 系统开展资产风险梳理工作，完成全面安全加固。
 - (2) 进行安全边界梳理，优化网络访问控制策略。
 - (3) 完善安全防护体系，提升整体防护能力。
 - (4) 建设安全监测体系，增强威胁发现能力。
 - (5) 建立情报共享机制，实现威胁信息互通。
- 3. 演练阶段验证优化
 - (1) 通过模拟攻防演练验证防护体系有效性。
 - (2) 根据演练结果进行总结及针对性加固。
- 4. 实战阶段应急处置
 - (1) 建立完善的威胁监测机制。
 - (2) 制定标准的攻击分析流程。
 - (3) 规范事件处置程序。
 - (4) 建立溯源分析方法。
 - (5) 明确事件上报路径。

- (6) 落实情报共享机制。
5. 战后阶段持续改进
- (1) 开展全面总结评估。
 - (2) 推动战时机制向常态化运营转化。
- 本方案通过这五个阶段的系统化设计，确保演练工作有序推进，各环节紧密衔接，形成完整闭环。

(二) 职责与组织分工

- 1. 网络安全责任落实：制定《信息技术管理办法》，明确公司党委对网络安全工作负主体责任，党委书记是第一责任人，分管领导是直接责任人；党委负责审议网络与信息安全管理目标；经营管理层负责落实信息技术管理目标，履行完善绩效考核和责任追究机制职责，下设信息技术委员会负责信息技术治理和安全管理相关工作；信息技术部是信息技术统一归口管理部门。
- 2. 本次网络安全攻防演练由我司信息技术部负责组织实施，为确保本次攻防演练任务的顺利完成，我部成立演练工作组，公司首席信息官任演练领导工作组组长，信息技术部副总经理任副组长，演练工作统筹安排人为网络安全部副总经理，下设五个工作小组，分别为监控研判组、应急处置组、溯源反制组、沟通协调组、演练保障组。
- 3. 防守组织机构如下：

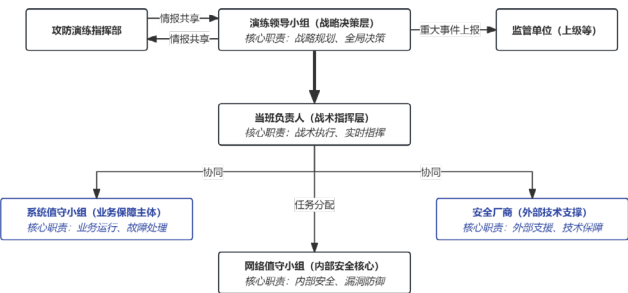


图1 防守组织架构图

角色	核心职责	具体工作
演练领导小组（战略决策层）	作为防守工作的最高决策机构，负责演练期间的总体指挥与协调。	1.领导与授权：全面领导攻防演练工作，对重大事件处置方案、资源调配等关键决策进行审批授权。 2.对外协同：作为唯一接口，负责向监督单位（上级单位等）进行重大事件的上报与沟通协调。 3.态势掌控：接收由当班负责人提交的每日工作简报，掌握整体安全态势与工作进展。
当班负责人（战术指挥层）	作为现场指挥中心的指挥，负责7x24小时值守工作的具体统筹与调度，确保安全事件得到及时有效处置。	1.统筹协调：统一管理安全厂商、网络值守小组和系统值守小组，确保各方力量协同作战。 2.闭环管理：监督并推动完成从“事件监控→分析研判→处置实施→结果反馈”的全流程闭环。 3.信息枢纽：汇总所有安全事件信息，编制每日简报，向演练领导小组汇报；同时将领导小组的决策和指令下达给执行团队。
网络值守小组（内部安全核心）	负责内部安全监控、分析、研判和部分处置工作，是防守体系的中坚技术力量。	1.主动监控：利用各类安全平台进行主动威胁狩猎，发现安全事件后第一时间上报至当班负责人。 2.联动枢纽：接收并验证安全厂商推送的安全告警，将其转化为内部处置工单；同时，将内部发现的问题同步给相关厂商进行深度分析。 3.处置执行：负责或协助系统值守小组进行安全事件的内部处置，如隔离主机、修补漏洞等，并将处置结果同步给当班负责人和厂商。
安全厂商（外部技术支撑）	威胁发现：通过专业设备（如WAF、NDR、EDR）支援演练工作开展。	1.威胁发现：通过专业设备（如WAF、NDR、EDR等）监测攻击，将精准告警上报至当班负责人和网络值守小组。 2.技术支援：协助网络值守小组进行深度事件分析、攻击溯源，并提供专业的处置方案。 3.进展同步：在完成所指派的事件分析或处置任务后，及时将进展与结论同步给相关负责人和团队。
系统值守小组（业务保障主体）	负责保障业务系统的稳定运行，是安全处置动作的最终执行者之一。	1.快速响应：根据网络值守小组或当班负责人的指令，对涉及的具体业务系统进行应急处置，如应用下线、漏洞修复、配置变更等。 2.保障业务：在安全处置的同时，最大限度地减少对业务连续性的影响，确保演练期间业务稳定。

（三）应急处置预案

1. 为应对演练中可能出现的各类安全事件，我司制定了分级分类的《安全事件应急处置预案》，旨在实现“及时发现、快速响应、有效处置、及时恢复”。

2. 建立事件分级标准：依据事件的影响范围、业务破坏程度等维度，将安全事件划分为“重大、严重、一般”三个等级，并对应不同的响应流程与上报机制。

3. 设计标准化处置流程：为常见事件类型（如Web入侵、恶意文件、数据泄露、横向移动等）设计了标准处置流程（SOP），明确每个环节的操作指令、责任团队与交付物。

4. 明确指挥与协作机制：固化以“当班负责人”为指挥中枢，网络安全值守小组、系统值守小组协同作战的应急响应团队架构，确保指令清晰、协同高效。

5. 预设授权与决策通道：针对需快速决策的场景（如业务下线、核心资产隔离等），预先明确了上报至“演练领导小组”的快速审批通道，确保关键时刻决策及时。

6. 该预案确保我司在面对真实攻击时，能够沉着、有序、高效地进行应对，将事件影响降至最低。

（四）员工安全意识培训

1. 为进一步提升公司员工网络安全意识和防护技能，结合“国家网络安全宣传周”活动的工作部署，信息技术部在攻防演练前邀请内部网络安全专家，采用线上+线下内部视频会议方式对公司全体员工进行了网络安全意识教育培训，培训内容包括网络安全形势、典型网络安全事件分析、数据安全、钓鱼邮件、弱口令等网络安全意识。信息技术部将通过公司培训系统组织网络安全意识考试，对全体员工进行网络安全考核。

2. 通过此次专项培训，学习掌握习近平总书记关于网络安全的重要思想和对网络安全工作“四个坚持”的重要指示精神，宣传普及网络安全知识，增强全体员工网络安全意识，提升公司网络安全管理水平，营造安全良好的网络环境，助力公司安全生产和高质量发展。

（五）内部动员会议

1. 在收到演练通知后，随即召开内部动员会议，覆盖信息技术部全员，由信息技术部网络安全部宣讲网络安全攻防演练活动事项及意义、保障方案、报告交付要求及注意事项，公司首席信息官发表重要指示讲话，并成立演练工作组，制定演练工作方案，同时下发《信息技术部关于加强各职场及分支机构网络安全管理工作的通知》，要求公司总部及各分支机构所有人员按照通知要求排查安全隐患，增强全员网络安全意识。

二、演练布防阶段

1. 为在攻防演练中构筑有效防御体系，我司在演练前围绕“摸清家底、消除隐患、构建防线、验证能力”的核心原则，系统开展了全方面备战工作。

2. 通过深入梳理互联网与内网资产，全面排查敏感信息泄露与系统漏洞，我司精准识别并处置了多类安全隐患；在此基础上，完成了安全基线统一核查与边界防护策略优化，显著收敛了整体攻击面。同时，着力构建覆盖网络、主机、应用各层面的立体化监控防护体系，并通过组织多轮攻防对抗演练与全员安全意识培训，切实提升了团队的协同响应能力与整体防御韧性。这一系列系统的布防工作，为演练的顺利开展奠定了坚实基础。

（一）资产风险梳理

1. 互联网资产发现

结合主动发现和被动发现两种手段进行互联网资产发现：

（1）主动发现：利用子域名爆破工具（如SubFinder）对主域名进行枚举；通过证书透明度日志（CT Log）查询、反向DNS查询、搜索引擎语法等手段，发现潜在的关联域名与主机；使用端口扫描工具（如Nmap、Masscan）对获取的IP地址段进行全端口扫描，识别开放的服务（如Web、SSH、RDP、数据库等），如下图，如下图2通过第三方网站进行实时互联网暴露资产测绘，如下图3通过监控功能实时监测公网IP暴露端口，如有更新开放端口设置推送发至邮箱。



图2 资产测绘-X社区

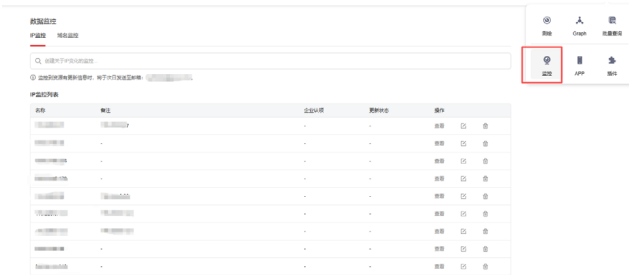


图3 监控端口暴露面-X社区

(2) 被动发现：一方面采集并分析NDR、全流量探针、WAF、负载均衡等网络边界设备的流量日志，通过分析访问日志中的Host头字段、目标IP，以及建立连接的SNI（服务器名称指示）信息，发现正在被实际访问的互联网域名和IP；另一方面系统性地梳理出口防火墙及网络设备上的NAT策略、端口映射规则，反向推导出所有被映射到公网IP的内网服务，以发现在资产清单中遗漏的或仅通过特定路径访问的资产。

2. 内网资产梳理

(1) 内网资产梳理工作遵循“分层分类、责任到人、技术与管理并重”的原则，依据业务属性对资产进行分层分类（如图4），开展内网端口与服务发现工作，重点识别数据库、文件共享、远程管理等高风险服务。同时，结合EDR（终端检测与响应）agent的装机列表，与扫描结果进行比对，确保无遗漏。

(2) 同样，将发现的内网资产进行统一管理，对识别出的非业务必需的高危服务进行关闭或访问限制；实施网络微隔离策略，根据不同业务区域划分安全域，严格限制横向流量的访问范围，防止攻击者突破后的肆意横向移动。

图4 攻防演练防守行动任务分解表

3. 敏感信息泄露发现

在敏感信息泄露发现方面，我司构建了覆盖四大渠道的系统化监测体系。针对文库平台，采用关键词组合+站内搜索技术，重点排查企业内部文档、技术手册及项目计划书等非公开资料。在代码平台（如GitHub、GitLab等），通过自动化监控脚本，结合企业密钥特征规则，对代码仓库、提交记录及Gist片段进行深度扫描。对于网盘渠道，利用文件哈希值比对及元数据分析技术，监测以企业标识命名的非公开共享文件。

4. 漏洞挖掘

我司采用“工具扫描+人工渗透”的双轨模式，首先使用

自动化漏洞扫描器对目标资产进行批量筛查，发现常见的中低危漏洞；随后，由网络安全专家对重要系统进行手动渗透测试，深度挖掘业务逻辑漏洞（如越权访问、支付绕过）、架构缺陷及工具无法发现的复杂安全风险。

(二) 安全基线检查

1. 网络架构评估

针对目标系统开展网络架构评估工作，以评估目标系统在网络架构方面的合理性，网络安全防护方面的健壮性，是否已具备有效的防护措施；

2. 网络安全策略检查

针对目标系统所涉及的网络设备进行策略检查，确保目前已有策略均按照“按需开放，最小开放”的原则进行开放；确保目标系统所涉及的网络设备中无多余、过期的网络策略，最终形成网络安全策略检查报告并与各系统部门进行核对；

3. 网络安全基线检查

针对目标系统所涉及的网络设备进行安全基线检查，重点检查多余服务、多余账号、口令策略，禁止存在默认口令和弱口令等配置情况；

4. 安全设备基线检查

针对目标系统所涉及的安全设备进行安全基线检查，重点检查多余账号、口令策略、策略启用情况、应用规则、特征库升级情况，禁止存在默认口令和弱口令等配置情况；所有设备配置白名单登录内网管理IP、SNMP可读取白名单服务器IP；演练前将所有网络安全设备特征库均更新至最新版本。

(三) 监控与防护体系建设

为有效应对攻防演练中多样化的攻击手法，我司并未依赖单点防护，而是秉承“纵深防御、主动监控、智能响应”的核心思想，构建了一套覆盖“网络—边界—主机—应用—数据”全链条的立体化监控与防护体系。该体系通过各类安全能力的有机组合，实现了从外部攻击拦截到内部威胁狩猎的全流程覆盖，确保能够“看得全、防得住、查得清”。

1. 防守体系

如图5监控防护体系可划分为四个核心层次：

第一层：互联网边界与应用入口防护。在互联网边界，通过下一代防火墙和入侵防御系统（IPS）实施严格的访问控制与威胁过滤。针对最主要的Web攻击面，部署了Web应用防火墙并结合CDN的加速与安全能力，有效抵御SQL注入、XSS等OWASP Top 10攻击，同时利用安全互联网网关对全网互联网出口流量进行管控。

第二层：网络内部与邮件通道监测。为应对边界突破后的横向移动，部署了网络检测与响应（NDR）平台，对东西向流量进行全流量分析，深度检测高级威胁与异常行为。同时，通过邮件安全网关与沙箱联动，拦截钓鱼邮件、恶意附件等通过邮件渠道投递的威胁，并利用DNS解析防护设备提供快速的域名信誉查询与拦截能力，切断恶意软件的C2通信。

第三层：主机终端精准防护与响应。在终端层面，部署了

终端检测与响应 (EDR) 系统和主机入侵检测系统 (HIDS)，结合EDR与杀毒设备等轻量级Agent，实现对服务器和办公主机的进程行为、文件操作、系统日志等进行精细化监控，确保对终端层面的威胁实现精准定位和快速遏制。

顶层中枢：安全运营中心统一调度。以上所有安全组件产生的日志、告警信息均汇集至安全运营中心 (SOC) 平台。SOC通过关联分析引擎将分散的威胁线索进行聚合与研判，实现跨层级的威胁狩猎与事件响应闭环管理，大幅提升整体安全运营效率与协同作战能力。

本监控防护体系作为一个有机整体，以SOC为大脑，打通各安全组件的信息孤岛（如NDR发现内网可疑IP，可自动通知防火墙封禁、EDR进行主机排查，实现协同响应），构建了“网络—主机—应用”多层递进的防御纵深，具备事前预防、事中检测、事后响应的完整能力，为本次攻防演练的防守任务奠定了坚实的技术基础，可显著提高攻击者的成本和难度。

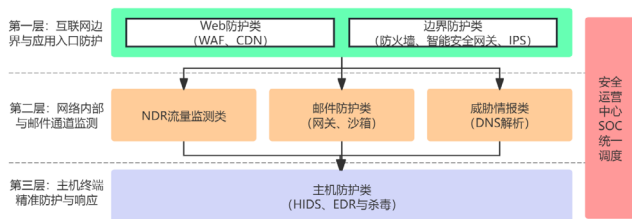


图5 监控防护体系

2. 防护能力

(1) 边界威胁监测与防护

通过边界安全网关独有的智能协议识别技术，在结合端口信息的基础上，通过流量特征来进行协议识别，提供从网络层、应用层到内容层的全面协议识别，即使是通过动态端口或隧道实施的恶意入侵行为也能被准确地检测和防护。

(2) 4-7层入侵检测与防护

通过基于攻击模型的攻击检测能力，实现各类4-7层威胁如SQL注入攻击、扫描探测、病毒利用、漏洞利用、非授权访问等恶意行为的威胁检测；同时利用IP碎片重组和TCP流汇聚能力，进行任意分片方式进行的攻击检测，进行4-7层的威胁检测与防护。

(3) 威胁情报监测与防护

通过威胁情报更新引擎和检测引擎，实现攻击威胁情报的实时同步和攻击检测；同时基于情报分析和生产引擎，实现全网攻击IP的快速捕获和更新，有效防护攻击队的探测扫描、入侵攻击等行为，提高演练期间的防守效率和攻防对抗过程中的主动防御能力。

(4) 威胁联动阻断

平台通过开放接口，可以实时获取SOAR等平台下发的封禁策略，可以实现黑名单及IP和域名的封禁能力。

(5) 全流量威胁监测

通过全流量威胁监测NDR平台，通过双向流量检测和全包协议识别作为检测基础，以持续高级威胁手段研究和威胁情报赋能作为加持，实现基于流量的实时威胁检测。

(6) 终端威胁监测

通过对终端行为数据的采集和应用大数据分析、机器学习、Graph图计算等新兴技术，采集操作系统日志行为数据，利用多个维度进行检测，并结合机器学习、Graph图计算、事件聚合、关联分析等技术综合评判威胁事件与进展，展示完整的进程上下文，并还原攻击全过程，溯源风险点。实现覆盖基础终端威胁以及高级终端威胁的终端威胁检测，实现无文件攻击、内网横移、0day漏洞等攻击的发现。

(7) SOAR处置

通过SOAR对日常安全告警处置和安全事件响应流程进行自动化编排，目前已覆盖告警研判、响应处置、协同联动、策略更新等多个场景，其中典型案例如下：

攻击IP封禁。安全大数据中台实时采集各类型检测设备攻击告警，基于攻击IP、攻击类型等维度进行实时统计和归并，将告警数量达到设定阈值的IP推送至SOAR，SOAR判定IP是否为数据报备白名单IP或我司业务对接第三方企业出口IP，若是则自动派发通知邮件，IP所属机构类型、历史攻击情况等多个维度进行威胁评级，针对不同级别的IP设定通知相关进行排查确认。否则将结合外部机构和行业情报共享等渠道获得的威胁情报，通过IP属地、不同封禁时长，并将信息自动推送至网络防火墙、旁路阻断设备和CDN，实现攻击IP的自动封禁。

SOAR平台在实施攻击IP自动封禁机制时，其威胁评估算法基于以下三个维度进行综合计算：

1) 攻击次数：根据攻击的次数来计算威胁值，平均每个资产受到的攻击次数越多，则威胁值越高。具体算法如下：

$$T_a = \frac{\sum(N_1 + \dots + N_{30})}{M_1} \times W_1 + \dots + \frac{\sum(N_1 + \dots + N_{30})}{M_n} \times W_n$$

T_a 为该攻击IP的威胁值， $\sum(N_1 + \dots + N_{30})$ 表示该攻击IP近30天内在我司的攻击告警总数，M为每个数据中心的资产数量，W为同一攻击事件在各数据异构网络安全设备检出频率的权重，同一攻击事件同时在多个网络安全设备上检出，告警准确性较高，权重也相对较高。

2) 攻击范围：根据情报中攻击IP攻击数据中心机房数量、IP或域名数量来计算威胁值，如果一个IP攻击数据中心机房数量和目标系统越多，范围越大，则威胁值越高。具体算法如下：

$$T_b = \log_a(N \times W_1 + S \times W_2 + D \times W_3)$$

其中 T_b 为该攻击IP的威胁值，a为对数算法参数，N为受攻击数据中心数量，S为受攻击的IP数量，D为受攻击的域名数量，W为攻击范围权重参数。

3) 攻击行为：根据攻击的类型来计算威胁值。如果攻击IP主要是信息收集、扫描探测类行为，则威胁值相对较低；如果有命令执行、木马上传、SQL注入、敏感文件爆破等攻击行为，则威胁值相对较高。具体算法如下：

$$T_c = B_1 \times W_1 + \dots + B_n \times W_n$$

其中 T_c 为该攻击IP的威胁值，B代表该IP使用的攻击类型，W为每类攻击行为的权重，攻击行为的风险和危害越大，则权重越高。

三、演练实战阶段

在为期多天的攻防演练中，本阶段工作以协同作战、纵深防御、主动应对为核心指导思想，通过四大维度的有机联动，成功构建了动态、高效的整体防守态势。

首先，在组织协同体系方面，我司依托前期建立的指挥框架，实现了多方高效联动。通过“每日例会、急事急报、重大决策”的机制，确保了防守指令清晰传达、事件信息无缝流转与应急决策迅速落地，形成了统一指挥、分工明确、协同响应的作战格局。

其次，在安全设备值守层面，以前期部署的WAF、NDR、EDR等设备为依托，执行7x24小时全天候监测。通过对海量告警的实时分析、聚合与研判，实现了对攻击行为的分钟级发现与有效拦截，构筑了坚实的技术防线。

同时，威胁情报的深度应用为防守工作赋予了“预见性”。通过引入多源威胁情报（IP、域名、攻击特征等），我司实现了对已知恶意行为的精准封堵与对新型攻击的早期预警，显著提升了威胁狩猎的效率与精准度。

最终，在重点事件监测处置上，我司聚焦于钓鱼攻击、零日漏洞利用等高风险场景，设立专项分析流程。通过深度溯源与快速封控，验证并强化了针对高级持续性威胁（APT）的应对能力。

实战阶段通过四大维度的紧密配合，将人员、技术、流程整合为一个有机整体，通过常态化协同、精细化分析、主动化封堵和闭环化管理，实现了“人机结合、情报驱动、闭环管理”的主动防御模式，有效保障了目标系统的安全稳定。

（一）攻击监测预警情况

1. 整体安全防护韧性

为确保安全防护体系持续有效运转，我司对关键安全设备实施了7×24小时全天候值守监控。通过对各类安全设备告警数据的实时分析、攻击类型统计与趋势研判，构建了完整的攻击感知与响应闭环，为防守工作提供了有力支撑。

防御层面	数据源	核心指标
互联网入口防御层（挡住了多少“门外”的噪音攻击？）	CDN、WAF	CDN累计拦截恶意请求/攻击次数
	邮件网关、沙箱	Web攻击拦截
	智能安全网关	钓鱼邮件与恶意附件拦截量
	DNS解析防护设备	累计封禁IP数
内网流量监测层（有多少攻击溜进了内网，并被及时发现？）	NDR	恶意域名请求拦截量
		内网横向移动/异常连接告警数
		可疑外联检测次数
主机终端防护层（挡住了多少“最后一公里”的攻击？）	EDR与杀毒设备	通过流量分析溯源发现的失陷主机数量
		恶意进程/行为阻断次数
	HIDS	确认并处置的（疑似）失陷主机数量
安全运营中心	SOC	攻击成功事件数量
		告警降噪比
		通过剧本化/自动化流程处置的事件占比
		指挥完成跨层级联动处置的事件数量

（二）事件分析研判及应急处置情况

在全面监控的基础上，本次攻防演练聚焦于高隐蔽性、高危害性的核心威胁场景，我司重点围绕钓鱼攻击、零日漏洞利用、互联网直接入侵三大核心攻击路径，构建了覆盖“事前预警、事中检测、事后处置”的全链条防守能力。通过主动狩猎、深度分析与快速响应，成功化解了多起高级别攻击尝试，有效验证并提升了针对关键威胁的实战应对水平，如下列举攻防演练中遇到的钓鱼攻击案例进行剖析。

案例一：

如图6，红队针对我司在线客服发送恶意钓鱼链接，得益于前期常态化对客服中心全员进行多次网络安全意识培训，客服中心人员未点击钓鱼链接且第一时间上报至网络安全部。

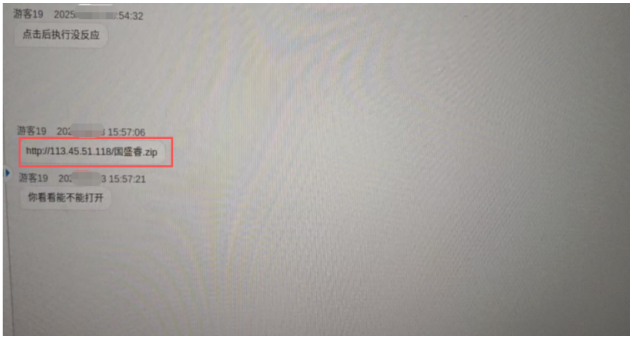


图6 红队钓鱼攻击

1. 红队木马样本

捕获到样本程序，属于CS木马家族的变种，如图7-10，通过mailslot的方式规避检测，最终实现远程命令执行和敏感数据窃取等恶意功能，主要影响Windows系统。

（1）样本信息



图7 红队攻击样本信息

05 2024年度实验室优秀课题摘录

P120 基于安全大模型提升网络安全监测响应能力的应用实践

胡广跃、王磊、胡沐宇、钱峰

P125 探索集中交易系统的应用数据安全国密改造

罗清平、刘海伟、邵伟、夏乔、张玉涛

P131 基于人工智能的加密流量威胁检测技术研究

马冰、吴晨炜、孙佩、秦伟伦

P136 高效特权账号管理体系技术研究与落地实践

丁安安、夏英杰、丁旦

基于安全大模型提升网络安全监测响应能力的应用实践

胡广跃、王磊、胡沐宇 | 光大证券股份有限公司

钱峰 | 深信服科技股份有限公司

摘要：传统网络安全防护手段主要基于传统安全设备的现有规则进行监测响应，但在应对高对抗性的网络攻击上，尤其0/1DAY，仍存在检测效果和研判效率不足的情况。本课题在现有纵深防御体系基础上，引入了大模型技术来构建“1+N”安全监测能力。“1”指大模型技术，对现有安全设备采集到的原始代码数据像安全专家一样逐条研判，“N”指现有监测机制，通过多种安全设备的检测规则来识别并拦截超过90%的已知网络攻击，通过“1+N”模式的相互协作与补充，有效提升检测能力和告警研判效率。在某次高强度对抗演练中，该实践机制成功捕获10起0/1DAY漏洞攻击，并识别出23个高价值事件，大幅提升人员处理告警信息的效率与精准度，充分验证了安全大模型在监测响应上的实际应用效果。

关键词：网络安全、0/1DAY漏洞、大模型技术、“1+N”安全监测

一、绪论

（一）课题研究背景

随着网络攻防对抗的持续升级、攻防不对等的现象日益严重，当前网络攻击手段日趋复杂，呈现三大特征：一是弱特征高级威胁成为主流，攻击由明显特征向弱特征、组合化体系攻击演变；二是0day高危漏洞数量增多、漏洞利用速度加快；三是攻击者普遍使用流量对抗、流量绕过技术规避检测。证券行业虽已落地各类安全设备，建成基础防护体系，但整体防护依托固定规则，难以抵御上述新型高级攻击。

生成式AI大模型快速发展，依托泛化分析能力弥补规则防护短板，为优化证券网络安全监测体系提供新路径，实现新旧防护能力互补增益。

（二）课题研究内容

现阶段证券行业已部署防火墙、WAF、态势感知等传统安全防护设备，配备专职安全人员值守运维，但面对0/1DAY漏洞、新型免杀钓鱼等攻击仍存在明显短板：其一防护依赖预设规则库，新型变异攻击极易绕过防护形成漏报；其二业务扩张带来海量告警，安全人员无法全量深度分析处置，安全投入与实际防护效果不匹配。

课题落地“1+N”智能化安全监测架构，整体分为数据采集端、安全大模型平台端。采集端依托现有态势感知与各类安全产品完成全量流量、日志采集；平台端依托专属GPU服务器部署本地化训练后的安全大模型，针对存量安全数据开展智能分析，补齐传统设备检测盲区，从海量告警中筛选关键安全事件，辅助运营人员研判处置。

（三）课题研究价值

证券行业现有安全运营模式依靠规则告警+人工处置，新型绕过攻击易造成防护漏过，海量告警也带来极高人工研

判成本，致使安全建设投入和实战成效失衡。

本课题依托安全大模型天然的泛化理解能力，模拟安全专家解析攻击代码与流量，跳出规则库约束，补充现有安全监测短板，实现整体防护效能1+1>2的建设价值。

（四）课题研究目标

1. 依托全量HTTP流量、系统日志、恶意代码样本完成大模型预训练，打造具备流量、代码、攻防知识、证券业务认知的虚拟安全专家，强化全维度流量检测能力；

2. 大模型对原有告警做二次校验并补充新增研判告警，支持人员通过平台对话解析数据包，提升告警自动化研判与事件响应效率；

3. 依托统一物理算力底座，按需迭代扩充流量检测、安全运营等相关安全能力。

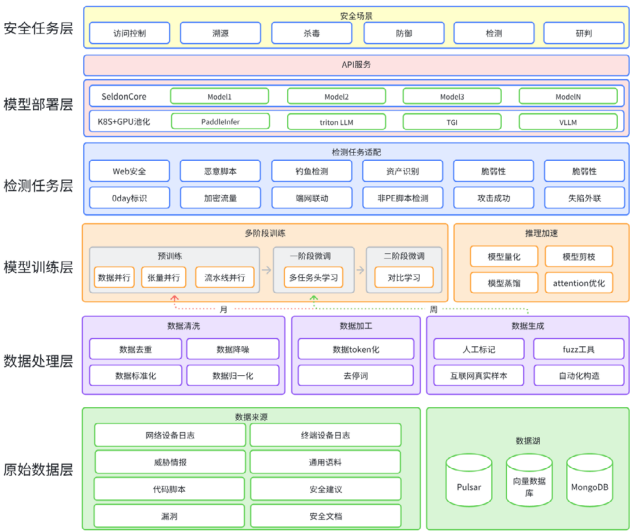
（五）课题研究方法

课题从检测精准率、高威胁事件检出、研判效率三个维度开展落地效果量化评估。检测精准率由安全人员抽样复测大模型检测结果进行核算；高威胁事件检出统计0/1day、高对抗、弱特征类攻击捕获数量；研判效率对比同等工时下，纯人工研判、大模型辅助研判的告警处理总量。

二、基于安全大模型提升网络安全监测响应能力的架构设计

系统架构划分为数据采集端、安全大模型平台端两部分，采集端复用现网存量安全设备，平台端搭载自研安全大模型并分层设计，整体架构自上而下分为六层：原始数据层、数据处理层、模型训练层、检测任务层、模型部署层、安全任务层。

由于数据采集端目前主要为通用的态势感知平台类设备,因此本文不进行赘述,重点介绍安全大模型平台端,其系统架构图如图1所示:



上述架构中,各层技术原理及功能如下:

- 原始数据层: 汇聚网端日志、终端日志、威胁情报、安全文档、通用语料等多源原始数据,为全链路处理提供数据源;
- 数据处理层: 完成数据清洗、去重、降噪、标准化、标记化等预处理,优化数据集质量,降低脏数据对模型训练的干扰;
- 模型训练层: 采用多阶段训练策略完成模型预训练与微调,搭配知识蒸馏、模型量化、剪枝、Attention优化四种技术优化推理性能,保障模型可落地实时在线检测;
- 检测任务层: 面向Web安全、恶意脚本、钓鱼识别、0day标识、加密流量分析等场景开展实时检测,并结合人工反馈持续迭代调优;
- 模型部署层: 基于K8S+GPU资源池,借助Seldon Core、Triton、VLLM等推理组件实现模型容器化部署与全生命周期管理;
- 安全任务层: 将大模型检测能力落地至边界访问控制、内网溯源、终端防御、威胁检测等实际安全运维场景。

三、系统测试情况分析

(一) 测试目标

本课题系统开发完成后,在实战攻防过程中预期达到如下目标:

- (1) 形成对现有安全设备检测能力的有效补充,能够检出攻击队使用的多类高对抗攻击,提升公司整体的网络安全监测响应能力;
- (2) 对现有的安全设备产生的误报进行纠正,降低需要关注的告警数量;

(3) 实现安全运营的“提质增效”,安全运营人员借助安全大模型能够快速对安全告警进行二次研判和处置。

(二) 课题环境和拓补

1. 测试说明

本次课题研究中,新部署了3台数据标准化平台和1台安全大模型设备,同时通过复用原有安全设备如态势感知平台、终端安全平台等,进行原始安全数据的采集。

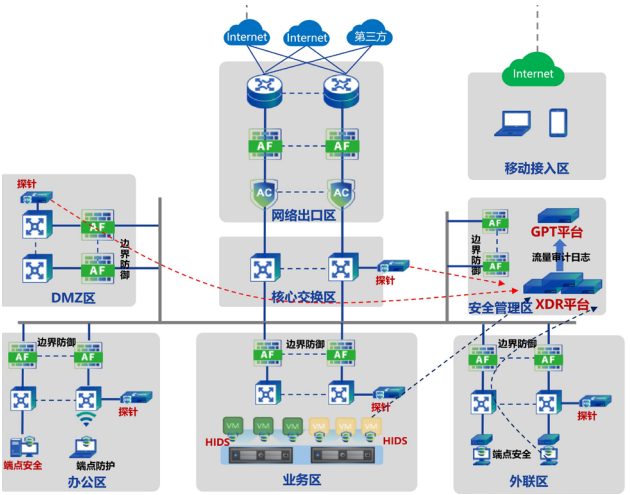


图2 光大证券安全大模型部署逻辑拓补

(三) 典型成果

在2024年某个全国范围内的较长时间攻防演练过程中,对安全大模型的检测效果进行了验证,部分成果如下所示:

1. 成果1: 反序列化远程代码执行0day漏洞攻击的检测

概况说明: 光大证券的安全大模型平台在对抗期间,持续对流量探针采集到的原始流量日志进行解读分析,并发现一起JAVA反序列化攻击事件,被攻击资产为仿真蜜罐。经人工进一步研判确认,该攻击为反序列化远程代码执行0day漏洞。技术实现过程如下:

过程一: 探针采集到原始数据包并上报



图3 态势感知平台采集原始流量数据

过程二：安全大模型平台不基于检测规则，而是像专业安全专家对原始的数据包进行专业化解读



图4 安全大模型对数据包进行解读

过程三：原始数据包经过gzip算法加密，安全大模型平台自动使用gzip算法解密后还原出关键payload，同时判断其具备java反序列化漏洞攻击特征

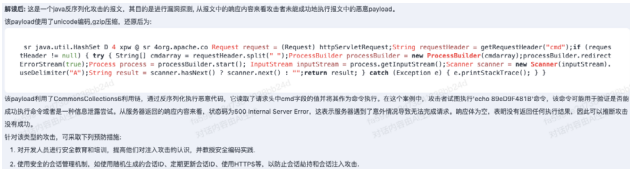


图5 安全大模型对数据包进行研判分析

过程四：大模型平台确认为java反序列化攻击后，形成告警同步到现有安全运营平台，作为现有安全检测体系的能力补充。

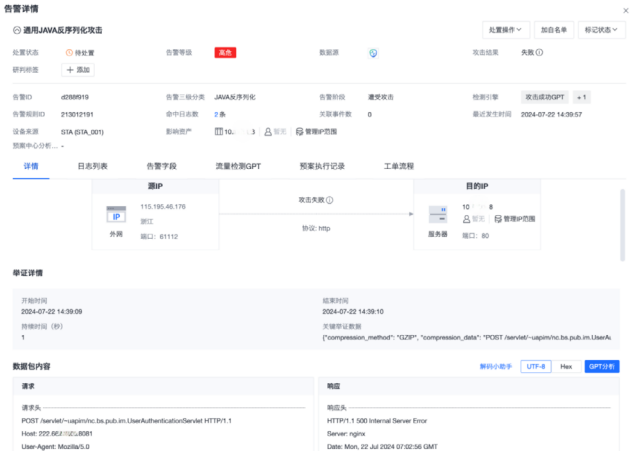


图6 安全大模型对攻击形成告警并同步到现有安全运营平台

过程五：通过人工进一步研判确认，确认为反序列化远程代码执行0day漏洞。在7月22日通过互联网及漏洞平台进行查询，未发现该漏洞被披露；在7月26日，此漏洞才被正式公开为高风险漏洞。

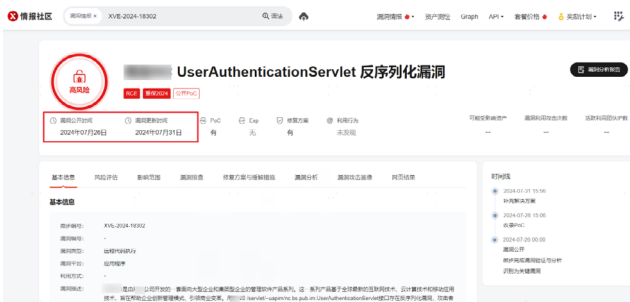


图7在漏洞平台上对此漏洞进行查询

2. 成果2: Apache Struts OGNL注入漏洞攻击自动化研判

融入最新DeepSeek能力后，安全大模型具备完整的安全告警研判能力，可自动化分析并处置遭受的网络攻击。以遭受Apache Struts OGNL注入漏洞攻击为例。

安全大模型的思维链第一步为解释OGNL注入漏洞攻击的原理，第二步通过对网络流量设备上采集的数据包进行解读并判断告警为真实的攻击，且已经攻击成功，第三步对告警发生的主机进行查询，发现有同样的告警，验证了此主机确实遭受了攻击。

安全大模型对告警的思考的过程与一个约有5年经验安全专家对告警进行分析的过程接近一致。



图8 安全告警智能研判过程

(四) 测试结论

在研究实践中，为评估安全大模型的整体效果，设定了检测精准率、高威胁事件检出、研判效率三大维度对安全大模型的效果进行评估，同时与传统安全检测方案进行对比，整体结论为：

在2024全国范围的高对抗攻防演练期间，安全大模型通过对原始流量审计日志进行自动研判，研判告警准确率在98%以上，同时安全大模型成功检出高混淆攻击、弱特征攻击事件、0/1DAY共计33个，驻场安服人员通过安全大模型对告警进行解读，从原有的每天完成500+告警研判增加至每天研判2000条以上，研判效率提升至原有的300%，以上数据有效证明了安全大模型平台有效提升了公司的安全防护体系检测能力。

指标	子类	统计方法	基于安全大模型	基于现有检测体系	效果提升百分比
检测精准率	精准率	精准率=正报告告警数量（经过人工二次研判）/安全检出告警数量*100%	98%	65%	50.7%
高混淆攻击					
高威胁事件检出	弱特征攻击事件	数量统计	33	15	120%
	0/1DAY				
研判效率提升	安全告警研判效率	单个安全工程师人工每天研判的大致告警数量	2000	500	300%

表1 安全大模型整体效果评估统计表

四、应用推广

基于安全大模型提升网络安全监测响应能力的课题，自2023年9月开展安全大模型技术研发以来，不断打磨完善，并于2024年5月正式上架，同时对接态势感知平台数据。2024年7月初基于前期测试结果进行二次调优，同时接入端点侧安全数据。目前已经完成全公司所有安全数据的接入。为支持安全大模型能够在证券行业顺利落地及推广，本文总结几大核心经验要点如下：

1. 兼容性，数据解耦

大模型兼容现有全品类安全设备数据源，通过标准API对接实现检测结果回传原有安全平台，不用改造现有安全架构，规避重复建设。

2. 普适性，资源易得

安全大模型作为垂域模型，应提前通过知识蒸馏、模型剪枝等手段减少模型参数量，降低模型复杂度，从而降低实践落地过程中的资源消耗。

3. 合规性，数据可控

应满足证券行业特性要求，完全采用本地化算力搭载模型执行推理，模型本地更新，数据严格控制在本地。模型训练数据集也应进行加密及脱敏。

4. 开放性，与时俱进

安全大模型在前期架构设计的过程中，应提前考虑模型自身开放性，可按需融合最新开源大模型能力，如DeepSeek-R1、Qwen系列等，保证模型底层能力的先进性。

五、课题研究成果

(一) 优化大模型在安全监测垂直领域能力

大模型是当下科技创新的关键驱动力，但在安全监测领域，尤其是证券行业的网络安全方面，仍存在较大的技术和应用空白。大模型时代，数据的丰富程度和数据量是决定最终模型效果的重要因素之一，光大证券自身长期积累了大量的行业级的流量样本，IP、URL、漏洞等情报数据，为大模型提供了不同类型的真实数据，结合安全人员研判知识，对业界优秀大语言模型基座进行二次预训练和微调，构建安全检测大模型。安全检测大模型能够发现混淆、编码类高绕过流量，并针对Web漏洞利用攻击有良好检出效果，具有较强

Web 0/1day漏洞检测能力，同时针对攻击是否成功的研判具有较高准确率。

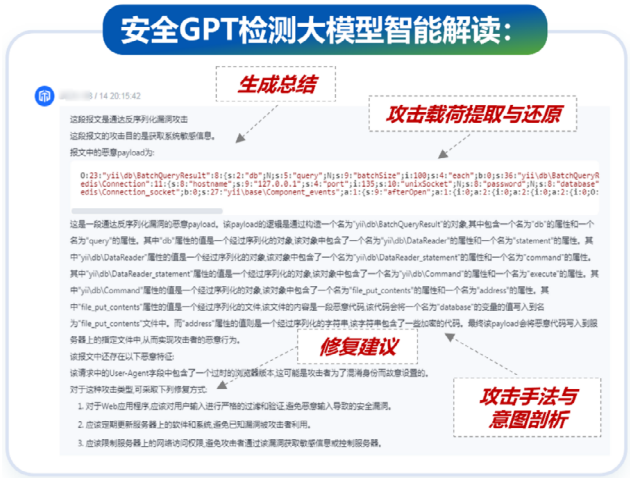


图9 光大证券GPT平台对攻击载荷的解读展示

(二) 完成大模型技术在证券行业安全检测领域的工程落地

理论研究到技术落地之间通常存在不小的鸿沟。本次课题落地实践中，将“不对现有安全架构进行大幅改造，同时有效赋能现有安全体系”作为核心原则，最终探索出“统一采集—并行检测—统一呈现”的方式，实现大模型技术的成功落地。具体落地模式为：流量探针统一采集流量数据，并传输至传统流量检测平台与大模型平台并行进行检测，将检测结果统一传输至安全运营监测平台进行聚合、呈现。本项目在大模型技术研发创新的基础上，为证券行业创新落地提供了良好的实践基础和案例。

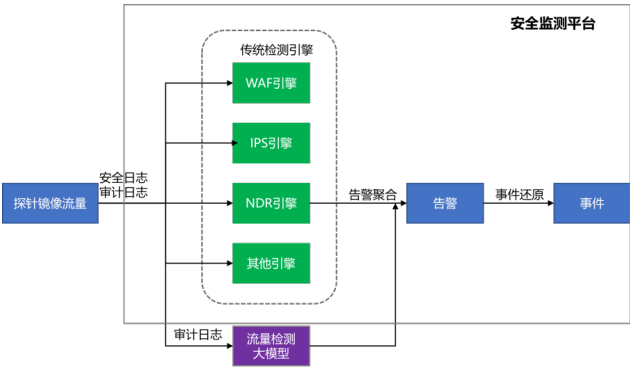


图10 光大证券GPT平台整体检测流程

(三) 基于大模型技术实现安全运营“提质增效”

网络安全领域的提质增效一直是各大证券行业机构长期关注的事项。一方面，网络安全设备的告警需要大量的安全工程师进行分析研判；另一方面，部分告警对安全工程师的专业能力提出了很高的要求。这两者叠加导致在日常安全运营中的投入成本非常高。而安全大模型更像是一位“既懂安全，又懂代码开发”的安全专家，它能够对原始的流量数据进行解读，判断代码的攻击意图，从而实现大量告警的自动化处理，仅将少量真实的安全事件传递给安全工程师进行

二次判断。此外，安全工程师可以通过安全大模型平台使用自然语言进行持续对话研判，从而显著提升日常安全运营的效率。



图11 安全大模型平台自动化处理告警的展示

（四）基于大模型底座完成安全能力的叠加与扩展

目前通用大模型在细分安全领域的能力存在不足，因此光大证券此次在探索实践的过程中，基于统一、通用的大模型基础算力底座，在此底座上不断扩展各类安全应用能力，目前已经完成检测能力与网络安全运营能力的扩展，后续还将依托该底座持续拓展，最终实现“一个底座支撑、多类垂直网络安全能力持续生长”的格局。

六、课题总结

基于安全大模型提升网络安全监测响应能力课题的研究与实践历经近2年的时间，完成了大模型系统的技术研发与实际落地，目前系统已稳定运行。研究成果达到了预期目标，通过对传统大模型的语料输入和不断训练，成功优化了大模型在安全监测响应垂直领域能力，能够有效检测出传统网络安全设备难以检测的攻击，同时在证券行业率先完成了安全大模型的落地实践，并在实战中取得良好的成效。另外通过大模型实现了与现有安全运营流程的紧密结合，大幅提升了安全运营人员日常的研判效率，降低了人员的精力投入。同时，在现有安全大模型的底座上，以模块化的方式集成了检测和运营的能力，后续也将继续探索大模型在其他网络安全细分领域的应用，通过大模型为网络安全领域赋能带来新的篇章。

参考文献

[1] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." arXiv preprint arXiv:2104.08821 (2021).

[2] Crawshaw, Michael. "Multi-task learning with deep neural networks: A survey." arXiv preprint arXiv:2009.09796 (2020).

[3] Yang, Enneng, et al. "Adamerging: Adaptive model merging for multi-task learning." arXiv preprint arXiv:2310.02575 (2023).

[4] Bengio, Yoshua, et al. "Curriculum learning." Proceedings of the 26th annual international conference on machine learning. 2009.

[5] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

[6] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

探索集中交易系统的应用数据安全国密改造

罗清平、刘海伟、邵伟 | 中国银河证券股份有限公司

夏乔 | 深圳市金证科技股份有限公司

张玉涛 | 三未信安科技股份有限公司

摘要： 课题组将国产密码技术引入到集中交易系统改造过程中，探索国产密码技术在各环节中如何正确有效的应用，并且尽量降低对系统业务连续性的影响。积累金融领域证券期货业国密改造经验，对行业内核心交易系统进行国密改造有引导作用，进一步推进和规范国产密码技术在行业内的应用。主要针对身份鉴别、访问控制信息完整性、重要信息资源安全标记完整性、重要数据传输与存储的机密性和完整性等多个方面在技术、设备、标准等方面进行研究。

关键词： 集中交易系统、国产密码、应用数据安全

一、课题概况

（一）研究背景

当前国内外网络空间安全形势复杂严峻，安全事件频发。我国从国家安全战略层面出发，已通过多项法律法规明确要求应用国产密码技术，监管部门同步出台实施规划，推动行业安全自主可控。密码技术是网络与信息安全的核心支撑，在身份鉴别、安全隔离、数据加密、完整性保护、抗抵赖等方面发挥关键作用。金融信息安全是国家信息安全的重要组成部分，加快金融领域国密应用、提升信息系统自主可控能力已十分迫切。按照《证券公司网络和信息安全三年提升计划（2023-2025）》要求，需实现国密技术在证券期货行业的全面覆盖，强化行业安全保障能力。

本课题探索在集中交易系统的各业务运行过程中正确、有效的使用国产密码技术，对重要流程、重要数据加强安全防护，充分发挥国产密码在保障证券期货行业信息安全中的重要作用，促进资本市场业务安全稳定发展。

（二）研究思路

课题采用自顶向下的分析方法对业务过程分类，结合对业务数据的分类，研究如何在身份认证、重要数据传输、重要数据存储等交易系统应用各个环节安全加固，探索更适用于行业交易环节的国产密码应用模式。

本课题率先将国产密码技术引入到集中交易系统改造过程中，探索国产密码技术在各环节中如何正确有效的应用，并且尽量降低对系统业务连续性的影响。课题探索一种可复制的集中交易系统国产密码应用改造(以下简称：国密改造)模式，积累金融领域证券期货业国密改造经验，对行业内核心交易系统进行国密改造有引导作用，进一步推进和规范密码在行业内的应用。

（三）进度计划

2024.11—2025.1：完成国密改造方案技术验证与路线选型，开展关键安全能力验证及国密设备选型、功能性能与认证测试。

2025.2—2025.5：完成国密功能改造与业务验证，覆盖身份鉴别、权限完整性、敏感数据保护及交易业务验证。

2025.5—2025.6：搭建原型环境并完成全流程测试。

2025.6：课题结题，输出数据、优化方案并形成报告。

（四）预期成果

本课题预期形成一套完整的集中交易系统国密改造方案，并按方案开展验证与全流程测试，评估国密技术在业务流程中的应用有效性，输出测试报告，为行业集中交易国密改造提供可行性参考。改造完成后，可有效完善等保三级及以上系统的数据与安全防护能力，防范敏感信息泄露、越权操作、滥用及恶意攻击等隐患，显著降低金融行业网络安全事件风险与社会影响。

二、课题研究内容

（一）系统现状

1. 集中交易服务

集中交易系统作为证券公司的核心交易系统，基于客户端、通讯中间件、业务中间件和数据库四层架构，为证券公司提供了易维护、安全可靠的系统，对证券公司经纪业务提供全方位服务，提升了客户体验。集中交易系统紧跟交易所、中登和基金公司脚步，支持沪深股债基的普通买卖，大宗交易，债券转股回售，市值配售、质押回购、股票质押、约定购回、自主行权，融资行权、债券协议回购、债券借贷、H股全流通，港股通、股转北交所、沪深京固收平台、银行间债券和场外基金等交易所全业务的交易和清算簿记，同时也支持交

易所报送、投保报送、证金报送等功能。

2. 统一接入服务

统一接入服务定位为外围用户与服务端的中间层，用于隔离外网客户端与内网服务端，降低耦合性，提供高性能、高可用的API服务；提供路由请求、鉴权、监控、缓存、限流等策略，避免外围直接对服务端造成压力，增强服务端的稳定性，提供统一的接入规范和服务接口，降低接入成本与技术复杂度。

3. 统一账户服务

统一账户服务提供投资者、柜员等账户全生命周期管理，集中管理投资者的账户信息、敏感信息和柜员权限、密码等相关数据。涉及到客户的开户、权限变更、密码管理、资料变更等功能。

(二) 密码产品调研选型

1. 服务器密码机

提供数据转加密、内部密钥加解密、外部密钥加解密等功能，生成会话密钥、保障应用层数据机密性、完整性。

涉及业务功能：应用系统数据的签名/验证，敏感信息（证件号码、手机号码、地址、邮箱）加密/解密。涵盖登录、开户、密码管理等相关业务。

2. 协同签名服务器

提供符合国标规范的数据签名功能，客户端与服务器端分别根据自身密钥因子完成中间签名，最后由协同签名合成符合规范的电子签名。协同签名系统提供标准数据签名的验签功能，系统可以利用CA颁发的数字证书完成对数据签名值的验证功能，能够根据签名结果判断数据的真实性，可靠性。

涉及业务功能：1.与柜员协签客户端联动提供合规的数字签名能力，实现对柜员类终端的应用层身份鉴别。2.外围请求会话提供合规的数字签名能力，对访问会话层提供身份鉴别。

3. 密钥管理服务

本研究中提供密钥、PIN的生命周期管理。

4. SSLVPN

符合国标规范的网络通道，主要应对网络传输层面的机密性、完整性要求，保障网络中数据传输的安全性。

本研究中为了保证网络通道的安全性的基本保障，无需业务系统等进行相关适配改造。

(三) 系统重要数据甄别

1. 鉴别数据

主要是柜员登录口令、资金口令、账户口令、柜员口令等信息，这类信息在登录、业务操作等业务过程中参与用户身份或操作权限的鉴别，需要使用密码技术在存储、传输、登录等过程中进行机密性保护。

2. 个人信息

保存在集中交易系统数据库中的投资者身份证号、投资者联系电话、投资者家庭住址等，需要使用密码技术在存储、传输、登录等过程中进行机密性保护。

3. 重要操作及交易日志数据

主要是集中交易系统产生的业务日志文件，对每日打包后的文件进行完整性进行保护。不涉及到传输方面的加密。

4. PIN码

对柜员登录进行二次认证，输入口令和PIN码来共同认证柜员的合法性。

三、成果环境建设

(一) 系统架构

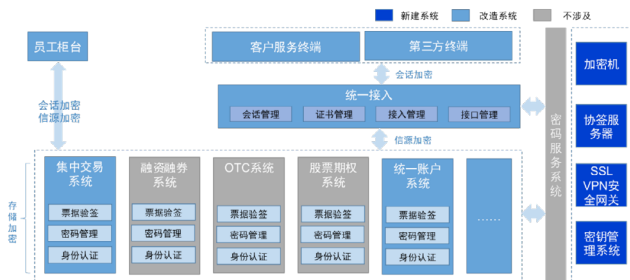


图1 系统整体架构

该课题涉及到的整体架构如图1，分为三层，外围终端、承接外围接入的统一管控的接入层和本课题要研究国密改造的集中交易系统后台业务系统。

对于传输数据的加密处理，我们引入三个概念，首先对于外围终端到统一接入，这一层我们定义为会话层的加密；对于从统一接入到后台业务系统的加密，我们定义为信源加密；对于业务系统中的存储数据的加密，我们定义为存储加密；

加密处理我们引入了加密机硬件，以集群的方式进行部署，需要跟统一接入、业务系统进行连接交互。

(二) 系统部署

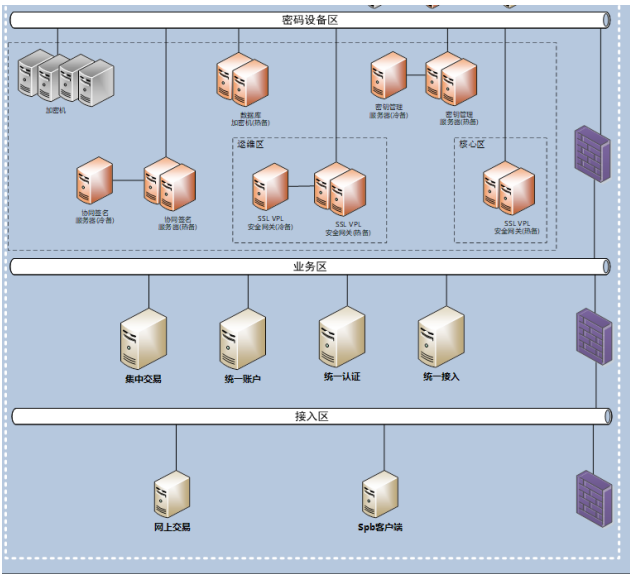


图2 系统部署架构

该课题涉及的系统主要如下:

业务系统: 集中交易、统一接入、统一账户

国密硬件: 服务器密码机、协同签名服务器、SSL VPN 安全网关、密钥管理服务、数据库加密机、运维堡垒机

其他工具: 密码设备相关CA证书、接入客户端(网上交易)

(三) 课题亮点

1. 针对数据传输安全的设计

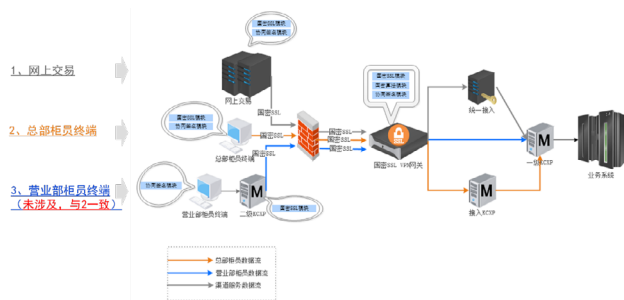


图3 数据传输

(1) 网络通信层身份鉴别: 通过部署国密SSL VPN安全网关, 由CA系统签发证书, 按接入对象集成SSL SDK或客户端, 建立国密SSL通道, 实现渠道服务端、内网终端、柜员终端三类对象的身份认证。

(2) 通信数据完整性: 采用符合GM/T 0025-2014的SSL VPN网关, 依托会话MAC密钥进行HMAC运算, 保障报文传输完整。

(3) 通信数据机密性: 基于国密SSL协议, 使用SM2/SM3/SM4算法加密数据包, 保护敏感字段与报文机密性。

(4) 网络边界访问控制完整性: 通过网关内置密码硬件模块调用SM3算法, 保障访问控制信息不被篡改。

2. 集中交易核心业务全流程国密改造

本课题对集中交易核心业务全流程国密改造, 涉及到柜员终端的业务流程、外围访问集中交易系统加密流程, 这两个场景为核心业务场景的主要来源, 做到了业务上的国密改造, 核心场景逻辑如下:

(1) 柜员终端身份认证

该部分要包含两部分内容:

第一部分为PIN码应用流程(见图4)。

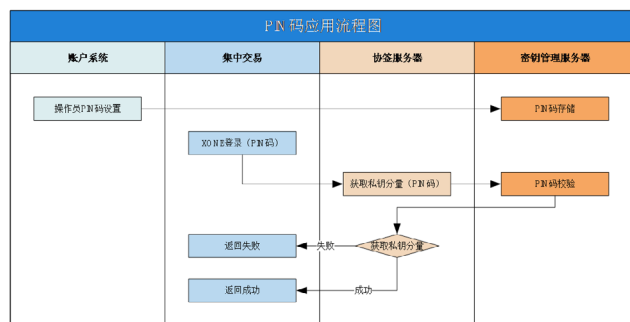


图4 PIN码应用流程图

详细流程说明如下:

账户客户端: PIN码通过账户系统前端发起, 账户KCBP调用对应XA, 同步到对应的密钥管理服务。

交易客户端: 集中交易前端客户端登陆, 输入客户PIN码及客户密码。客户端输入对应的PIN码, 从协同签名服务器获取对应的私钥分量。协同签名服务器根据送入的客户号及PIN码, 访问密钥管理服务, 同时校验PIN码的正确性。PIN码不正确则直接报错, 否则获取私钥分量成功。

第二部分为集中交易系统柜台柜员登录流程(见图5)。

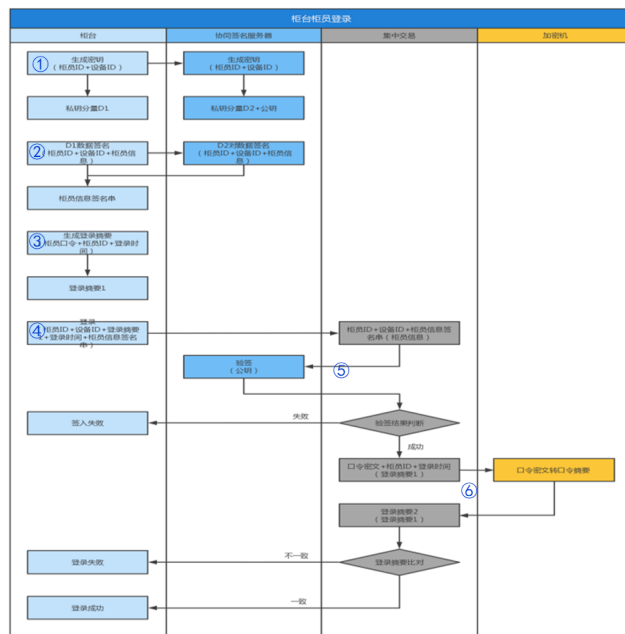


图5 柜台柜员登录

详细流程说明如下:

① 密钥对生成: 柜员终端调用协同签名SDK生成密钥对, 同时协同签名服务器也生成密钥对; 公钥ID是设备ID;

② 生成柜员信息签名串: 柜员终端用国密算法对柜员信息进行签名, 同时协同签名服务器也进行签名;

③ 生成登录摘要1: 柜员终端生成登录摘要1, 再发起登录请求;

④ 柜台发起登录: 将柜员ID、设备ID、登录摘要1、登录时间、登录信息签名串传到服务端;

⑤ 后台登录验证签名串: 服务端使用收到的柜员信息串、柜员ID、设备ID调用验签接口, 对柜员签名串进行验签, 返回验签结果;

⑥ 验证柜员口令: 验证通过后, 业务系统调用加密机对口令密文(数据库)+柜员ID(登录请求串)+登录时间(登录请求串)生成登录摘要2, 与柜台提供的摘要1进行对比, 确定柜员登录是否成功。

(2) 外围访问集中交易系统加密流程

外围(网上交易)通过统一接入对集中交易系统进行访问, 需要与统一接入先完成会话申请, 再进行业务流程功能调用。

第一部分为外围访问会话层加密流程(见图6)。

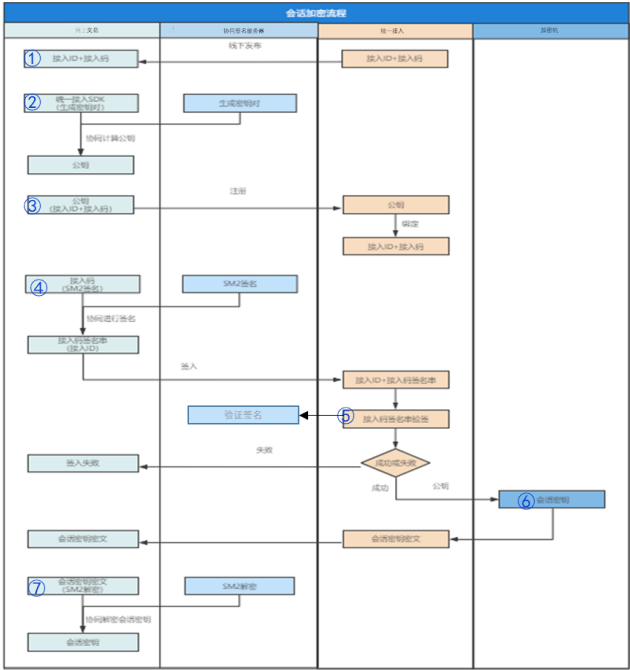


图6 会话加密流程

详细流程说明如下：

- ① 统一接入生成接入ID与接入码，线下分发至网上交易。
- ② 网上交易通过统一接入SDK生成密钥对，SDK同步调用协同签名服务器生成密钥对，两方协同形成公钥。
- ③ 网上交易使用公钥向统一接入发起注册，统一接入将公钥与接入ID、接入码绑定。
- ④ 网上交易与协同签名服务器分别对接入码签名，合并后形成接入码签名串，网上交易据此发起签入请求。
- ⑤ 统一接入对签名串验签，验签失败则拒绝接入。
- ⑥ 验签通过后，统一接入调用加密机生成并加密会话密钥，返回密文至网上交易。
- ⑦ 网上交易与协同签名服务器分别解密会话密钥密文，合并结果后完成会话密钥保存。

第二部分外围调用集中交易做业务请求，也是我们所说的信源加密的整体流程（图7）。

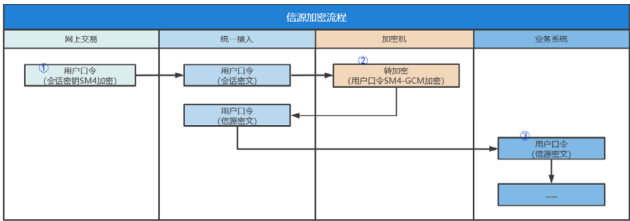


图7 信源加密流程

详细流程说明如下：

- ① 用户口令加密：网上交易使用会话密钥对用户口令进行加密，将用户口令密文（会话密文）发往统一接入；
- ② 会话密文转加密：统一接入收到用户口令密文（会话密文），使用加密机进行转加密，转加密完成返回信源密文，并发往业务系统；

③ 业务系统收到信源密文，再进行后续数据处理。

3. 应用和数据安全国密密码应用

(1) 身份鉴别采用统一验证机制，账户系统线下发用户ID、登录口令及协签PIN码。用户以ID+口令+数字签名完成登录认证，流程如下：

- 1) 用户输入ID、口令、PIN码，协签系统校验PIN码有效性；
- 2) 终端按ID+口令+时间盐值计算SM3摘要，通过协同签名组件完成SM2数字签名；
- 3) 终端提交ID、设备ID、摘要、签名等信息；
- 4) 系统调用协签接口验签，核对摘要与口令合法性，通过则完成认证；
- 5) 访问控制信息完整性通过服务器密码机的SM3-HMAC实现：系统先计算访问控制列表的MAC值作为基线BaseMAC；运行中周期性重算得到TempMAC并与基线比对，校验是否被篡改；信息修改后重新计算并更新基线，保障完整性。

(2) 重要数据传输的机密性与完整性保护如下：

- 1) 集中交易系统与网上交易等外围系统通信时，需保护账户口令、身份证号、手机号、家庭住址等重要数据。渠道身份鉴别通过后，统一认证系统使用对端公钥加密下发会话密钥，并由密码机完成密钥与时间信息签名；会话期间采用SM4-GCM加密关键数据，保障传输安全。
- 2) 集中交易系统与柜员间的管理类通信不含敏感信息，应用层不额外加密，依托国密SSL安全通道实现传输保护。
- 3) 重要数据存储按类型分级保护：
 - 重要数据（账户口令、身份证号、手机号、住址、邮箱等）采用SM4-GCM算法，通过服务器密码机实现写入前加密、读取时解密，保障存储机密性与完整性。
 - 一般数据（交易记录、日志）因实时性要求高，为保证交易性能，不做加密处理。

四、课题成果测试

（一）功能测试

1. 测试目的

银河证券集中交易系统启用了国密功能，为了确保银河证券启用国密功能后涉及到的相关业务平稳运行，测试相关业务，测试覆盖集中交易柜台敏感信息转换功能及外围接口的密码校验逻辑正确性。

2. 测试设备清单

设备用途	设备数量
加密机	2台
协同签名服务器	1台

表1 国密设备信息表

设备用途	设备数量	操作系统	应用组件版本
部署集中交易	1台	WindowsServer 2016	KCBP6491
部署统一账户	1台	WindowsServer 2016	KBSSACCT(WIN版)2.7.4.3_upd002_银河证券02(R)
部署数据库	1台	WindowsServer 2016	SQLserver2014
部署统一接入	1台	Kylinv10sp2	KGMS2504

表2 应用设备信息表

3. 测试工具

性能测试工具: Jemter

4. 测试结论

本次接口测试覆盖测试范围:

(1) 国密模式 (gmhardware=1, 启用): 验证410411、410611委托接口, 410301登录接口, 410605银证转账接口, 410753资金密码校验接口的密码校验逻辑。

(2) 回退模式 (gmhardware=2, 回退): 对上述相同接口开展密码校验逻辑测试。

国密开启模式下, 走统一接入模式时接口功能的正确性;

本次功能测试覆盖测试范围:

(1) 国密模式 (gmhardware=1): 验证柜台国密密码升级功能的正确性;

(2) 国密模式 (gmhardware=1): 验证密码升级后柜台敏感信息查询解密、新增信息加密的正确性;

(3) 回退模式 (gmhardware=2): 验证柜台敏感信息查询功能正常;

(4) 回退模式 (gmhardware=2): 验证柜台新增敏感信息以明文存储, 不经过加密机处理。

测试结果如下:

(1) 国密模式下, 外围接口按国密票据>普通票据>国密密码>普通密码优先级校验, 有值则验、通过即止, 为空不验。

(2) 委托接口二次校验: 410411按gmtrdpwd>trdpwd校验; 410611仅校验trdpwd, 均为空则不验。

(3) 回退模式下, 接口仅校验普通票据与普通密码, 普通票据优先级更高。

(二) 性能测试

1. 测试目的

测试评估集中交易系统在国密改造前后的运行状况, 通过对被测系统加压获得系统处理交易请求的能力, 系统的响应时间、TPS、服务器资源消耗等方面的数据, 验证国密改造后性能的表现情况。

2. 测试设备清单

设备用途	设备数量
加密机	2台
协同签名服务器	1台

表3 国密设备信息表

设备用途	设备数量	配置	操作系统	应用组件版本
部署集中交易	1台	32C32G	WindowsServer 2016	KCBP6491
部署数据库	1台	32C32G	WindowsServer 2016	SQLserver2014

表3 应用设备信息表

3. 测试工具

性能测试工具: KCBPTest

4. 测试结论

	TPS (笔/秒)	平均时延 (毫秒)	压测时长 (秒)
非国密	1056	14	180
国密	938	15	180
相对损耗	-11%	-7.1%	

表5 测试场景一: 使用15线程进行普通委托接口压测

	TPS (笔/秒)	平均时延 (毫秒)	压测时长 (秒)
非国密	7154	2	180
国密	7154	2	180
相对损耗	/	/	

表6 测试场景二: 使用15线程进行委托查询接口压测

	TPS (笔/秒)	平均时延 (毫秒)	压测时长 (秒)
非国密	8346	1	180
国密	7597	1	180
相对损耗	-9%	/	

表7 测试场景三: 使用15线程进行股东查询接口压测

	TPS (笔/秒)	平均时延 (毫秒)	压测时长 (秒)
非国密	8114	1	180
国密	7665	1	180
相对损耗	-6%	/	

表4 测试场景四: 使用15线程进行外国查询客户资金账户信息接口压测

注: 集中交易在国密改造之后由于涉及到硬件加密链路增加的原因, 对于普通委托的业务时延有少量增加, TPS有少量降低; 针对用户相关信息查询时延变化不大, TPS有少量降低。性能损耗在接收范围内, 符合课题预期目标。

五、课题完成情况以及成果应用

(一) 集中交易国密改造相关功能开发完成

模块	功能	完成情况
柜台功能测试	国密密码升级	已完成
	柜员访问权限设置	已完成
	柜员特殊权限设置	已完成
	柜员权限完整性检查	已完成
	三方存管联合开户	已完成
	三方存管账户开户	已完成
	资金信息修改	已完成
	普通委托	已完成
	市价委托	已完成
	上海大宗委托	已完成
	深圳大宗委托	已完成
	公司查询	已完成
	客户查询	已完成
	柜员登录	已完成
	客户密码修改	已完成
涉及敏感信息菜单	客户普通信息修改	已完成
	柜员管理	已完成
	客户管理	已完成
	资金账号	已完成
	证券帐号	已完成
	代理人管理	已完成

模块	功能	完成情况
涉及敏感信息菜单	三方存管	已完成
	资金存取划拨控制	已完成
外围接口	证券交易	已完成
	委托	已完成
	登录	已完成
	转账	已完成

表5 功能完成情况表

注：集中交易国密改造已于2025年05月完成搭建并全流程测试完成

（二）数据传输安全结合应用和数据国密密码应用方案

本课题预期形成一套完整的集中交易系统国密改造方案并完成验证测试。研究重点围绕数据传输安全与数据本体国密应用两大维度：采用国密SSL协议实现传输层身份鉴别、数据机密性与完整性保护，以及网络边界访问控制信息完整性保障；通过国密密码技术完成集中交易核心业务全流程改造，实现柜台操作与外围接口访问的多渠道安全防护。

（三）成果应用

经验证，本课题采用国密SSL VPN保障传输安全、服务器密码机保障应用数据机密性与完整性、协同签名服务器实现应用层与会话层身份鉴别的全流程硬件化国密改造方案，在集中交易系统中可行、稳定、可靠。该方案可直接复用至融资融券、股票期权等同类核心交易系统的国密改造。

研究成果对银河证券自身集中交易改造具有直接落地价值，对全行业、尤其金证科技承建的核心交易系统国密改造具备较强借鉴意义，可显著降低开发工作量、国密设备适配成本及功能与性能测试投入。后续将持续优化技术方案，进一步降低国密改造对交易系统的性能影响。

（四）行业引导价值

金融安全关乎国计民生，在数字化竞争加剧的背景下，密码技术已成为金融行业主动防御、提升核心竞争力的关键支撑。我国高度重视国产密码应用，相关法律法规与行业监管要求持续强化，密码技术既是守护行业安全的“盾”，也是参与国际数字竞争的“矛”，具备重要战略价值。

本课题通过集中交易系统国密改造，全面应用SM2/SM3/SM4等国产密码算法，构建身份鉴别、数据传输加密、存储加密、访问控制完整性防护体系，从根本上摆脱对国外密码技术的依赖，可有效抵御数据窃取、交易篡改、金融欺诈等风险，显著提升核心交易系统安全可控水平。改造方案兼顾安全与性能，实现敏感信息全生命周期保护，保障了交易传输的机密性、完整性与不可抵赖性。

参考文献

[1] JR/T 0158-2018 证券期货业数据分类分级指引

[2] JR/T 0071.2-2020 金融行业网络安全等级保护实施指引第2部分：基本要求

[3] JR/T 0171-2020 个人金融信息保护技术规范

[4] JR/T 0197-2020 金融数据安全数据安全分级指南

[5] GB/T 32905 SM3 密码杂凑算法

[6] GB/T 32907 SM4 分组密码算法

[7] GB/T 32918 SM2 椭圆曲线公钥密码算法

[8] GB/T 36322 密码设备应用接口规范

[9] GB/T 39786-2021 信息安全技术 信息系统密码应用基本要求

基于人工智能的加密流量威胁检测技术研究

马冰、吴晨炜、孙佩、秦伟伦 | 国泰海通证券股份有限公司

摘要： 全球网络走向全面加密已经是大势所趋，近年来金融证券期货行业业务服务越多采用加密方式进行通信。与此同时，为了隐藏攻击痕迹，大多数威胁攻击都开始采用通信加密技术掩盖自身的攻击特征，而基于规则的传统检测手段对于加密后的流量，无法有效识别隐藏其中的恶意行为，已无法满足金融机构全面防护的需求。本课题研究将基于人工智能技术围绕加密流量高对抗攻防在技术、工程等方面的核心问题展开研究，通过基于各类机器学习算法、传统深度学习算法构建加密流量威胁检测多个独立模型，检测已知/同源变种威胁并进行综合决策，形成基于人工智能的加密流量威胁检测系统。以充分挖掘AI潜力，综合有效运用AI能力来应对加密攻击高对抗攻防态势。

关键词： 特征工程构建、模型训练迭代、加密攻击高对抗、隐匿加密通信检测

一、研究背景和目标

（一）研究背景

随着全球信息化进程加速，全球网络走向全面加密已经是大势所趋，近年来金融证券期货行业业务服务越多采用加密方式进行通信，与此同时，网络攻击形势也发生了深刻变化。网络攻击已不再仅仅是黑客个人行为，而是呈现出组织化、国家化的趋势。为了隐藏攻击痕迹，大多数恶意软件及APT攻击都会采用通信加密技术掩盖自身的攻击特征。同时，伴随着智能化发展大潮，网络攻击在人工智能技术的加持下，攻击质量和数量双双增长，导致网络安全风险日趋严峻。在此背景下，网络攻击手段变得更加复杂，也越来越多的采用加密的攻击方式或手段。这使得安全防御在面临传统网络攻击的同时，也要应对加密流量中的未知威胁，让传统安全防御体系出现严重瓶颈，同时也给新的网络安全防护体系建设带来前所未有的挑战。

目前，新型网络攻击呈现如下趋势：

1. 网络攻击呈现组织化、国家化

近年来，网络攻击事件频发，其背后组织者和实施者往往具有明确的政治、经济目的。这些攻击行为不再是孤立的个体行为，而是有组织、有预谋的集体行动。国家力量介入网络攻击的现象也日益明显，网络空间已成为各国争夺的新战场。在这种组织化、国家化的网络攻击背景下，高对抗攻防态势已逐渐形成。

2. 攻击手段越发复杂化

在网络攻击场景下，攻击手段呈现出以下特点：

- 加密通信：攻击者采用加密技术对攻击数据进行加密，使得攻击流量难以被检测。加密入、加密出的攻击方式，增加了网络安全防护的难度。
- C&C隐匿：攻击者通过隐藏命令和控制（C&C）服务器，使得追踪攻击源头变得更加困难。
- 快速迭代：攻击者针对防护措施，迅速调整攻击策略，实现持续对抗。

上述趋势让网络攻击逐渐形成加密通信高对抗攻防态势，并在微观维度和宏观维度呈现出一些显著特点，如微观维度单事件攻击链条各环节攻击质量全面升级，传统威胁检测手段出现严重瓶颈；宏观维度加密技术被黑客利用，导致加密攻击流量占比越来越高，传统威胁分析能力出现严重不足。鉴于高对抗下攻击呈现加密化趋势，传统安全防御体系出现瓶颈，急需引入新型加密流量威胁检测技术来解决面临的加密威胁攻击问题。

（二）研究范围及目标

AI已在NLP（自然语言处理）和CV（计算机视觉）等领域展现重要价值。在网络安全领域，人工智能与加密流量检测技术的融合是机遇也是挑战。针对加密攻击高对抗特点，本课题拟开展基于人工智能的加密流量威胁分析检测方法研究，研发基于人工智能的加密流量威胁检测安全防御系统并试点应用。

总体技术研究目标及研究范围概括如下：

1. 开展基于人工智能的加密威胁分析及检测的方法研究：研究基于多层面细粒度特征工程构建技术方法；研究基于AI的加密攻击（解决如何检测隐蔽入、隐蔽出、C2隐匿这三大问题）检测技术方法。
2. 研发基于人工智能的加密流量威胁检测系统：构建技术框架、研发原型系统并确定核心技术指标。
3. 开展课题成果落地工作，并在行业示范：完成试点应用，并在金融证券期货行业内形成示范效应。

二、研究主要方法及内容

（一）研究方法

本项目研究中采用了如下几种研究方法：

1. 文献调研：通过收集和分析国内外相关研究成果，了

解当前人工智能在网络安全领域的应用现状和发展趋势，为项目提供理论基础和技术参考。

2. 算法研究与设计：结合机器学习和深度学习算法，研究并设计能够有效识别和防御加密流量网络攻击的模型和算法。

3. 系统开发与实现：基于设计的研究成果，开发一套集成了人工智能技术的网络安全加密流量监测防御系统，实现对网络加密威胁的实时监测和预警。

4. 实验验证：通过搭建实验环境，模拟现实网络攻击场景，对研发的系统进行验证和测试，评估其在实际应用中的性能和效果。

5. 技术迭代与升级：紧跟网络安全和人工智能领域的最新发展动态，不断更新和升级系统的技术和功能。

6. 试点应用示范：在实际业务场景上试点应用，实现研究成果的快速转化，并在行业内示范。

通过上述研究方法的综合运用，本项目实现了基于人工智能的加密流量威胁检测技术的突破和创新。

(二) 研究思路及总体方案

1. 研究思路

加密流量高对抗场景下攻击手段复杂，攻击方对抗能力强，加密入、加密出、C&C隐匿等新技术使得攻击被发现难度增加，而行业网络安全检测体系却存在误报多、分析能力不足等问题，尤其前端检测设备针对加密网络流量攻击的识别发现能力亟待提升。可从不同角度来发现针对加密流量高对抗攻击特点。

基于攻防技术视角的特征分析

对比项	对应攻击链	高对抗攻击	常规攻击
解决“怎么入”的问题	信息侦察及武器化阶段	储备大量魔改的攻击工具	常规攻击工具
	初次入侵阶段	加密扫描爆破、加密webshell	明文扫描爆破、HTTP webshell
解决“怎么隐匿”的问题	C&C构建阶段	利用CDN/云函数/域前置等白站进行中转实现C&C隐匿	C&C使用真实域名或服务器
	横向移动阶段	加密方式扫描爆破	明文方式扫描爆破
解决“怎么出”的问题	窃密回传阶段	上行采用TLS加密或利用DNS/ICMP/HTTP隐蔽隧道方式穿越边界	上行采用明文方式回传数据
	逃逸销毁阶段	下行采用TLS或各种隐蔽隧道进行加密通信	一般不涉及

表1 高对抗攻击与常规攻击对比

基于行业工程视角的检测短板分析

目前针对网络流量监测已经广泛部署了前端流量检测设备，但在加密流量威胁检测应对方面存在诸多短板：

- (1) 入联复杂加密攻击检测能力不足。
- (2) 检测攻击方隐匿方式支撑能力较弱。
- (3) 人工智能技术应用在新型威胁检测能力缺乏。

从上述分析可以看出，传统安全检测技术手段已经无法适应加密和隐匿攻击的检测需要，因此需要引入新型安全检测技术，提升对加密及隐匿攻击检测的能力。

2. 研究内容

基于加密流量高对抗场景所存在的问题，本项目研究全

面围绕加密流量攻击检测和防御展开研究，主要包括以下内容：

- (1) 开展基于多层面细粒度特征工程构建技术研究；
- (2) 开展基于增量学习的入联高对抗攻击检测技术研究；
- (3) 开展基于多流交互行为特征的隐匿攻击检测技术研究；

究；

- (4) 开展基于集成学习的出联高对抗攻击检测技术研究。

3. 主要技术路线

为解决上述技术问题，实现本课题的主要研究目标，本项目采用以下技术路线：

- (1) 深入分析网络加密攻击特点和现有技术不足，确定研究方向；
- (2) 研发创新的算法和模型，进行加密攻击检测和防御技术研究；
- (3) 搭建实验环境，对研究成果进行实证测试和性能评估
- (4) 不断完善和优化系统，提高检测准确率和防御效果。

(三) 加密流量威胁检测体系建设技术框架

本课题所研究内容基于人工智能的加密流量威胁检测体系技术框架，可以有效挖掘AI潜力，综合有效运用AI能力来应对加密攻击高对抗场景。根据AI领域技术成熟度，分析网络安全领域应用场景：深度应用小模型应对已知/同源变种攻击，全面、精准、深度嵌入到安全防御体系中。整个体系的技术框架如下图所示：

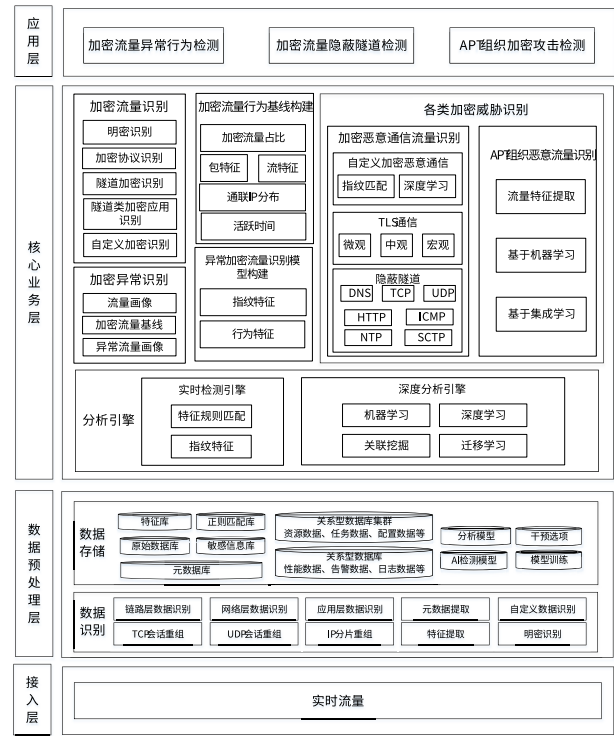


图1 系统技术框架

(四) 主要研究内容

1. 开展基于多层面细粒度特征工程构建技术研究

在构建人工智能模型用于加密流量威胁检测之前，有一部关键步骤为特征工程即加密流量数据的特征提取和分类。

然而在实际特征生成的过程中, 存在加密流量样本差、特征工程构建质量差的问题。为解决这一问题, 本课题研究综合利用现网采集、样本清洗、多层次细粒度的特征工程构建能力来解决。

研究针对加密会话, 从时空特征、握手特征、证书特征、背景流量特征四个方面进行特征提取分析并构建特征工程。核心研究思路通过全面提取加密流量“黑”“白”差异化有效特征, 支撑多维度模型训练。

基于多层次细粒度特征工程构建技术路线研究分别从以下方面的具体维度展开分析: ①时空特征: 时间特征和空间特征, 如流时长、包间隔、包大小、转移矩阵、熵值等基础元数据和基于这些元数据计算得出的特征。②握手特征: 客户端和服务端在握手阶段的特征, 如协议版本、支持的扩展项等基础元数据和基于这些元数据计算得出的特征。③证书特征: 如证书链长度、有效期、证书正常度、公钥算法等基础元数据和基于这些元数据计算得出的特征。④背景流量特征: 客户端和服务端在加密会话过程中所关联的DNS、HTTP等其他协议元数据和基于这些元数据计算得出的特征。

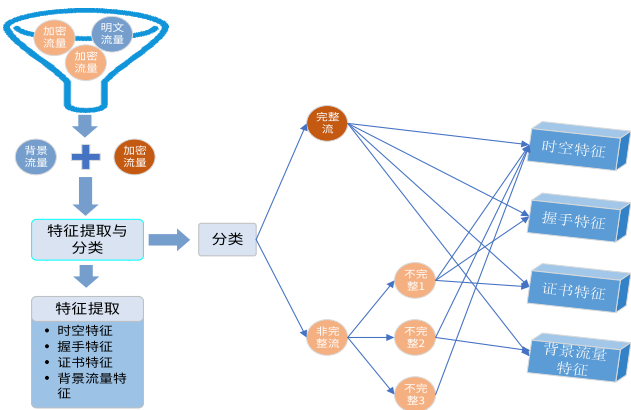


图2 特征工程构建技术路线

2. 开展基于增量学习的入联高对抗攻击检测技术研究

入联高对抗攻击类型常见类型包含0day漏洞利用、加密webshell攻击、加密扫描爆破等攻击类型等。当前防御体系在检测上述入联高对抗攻击过程中, 存在如下问题: 当前主要通过机器学习固化模型等方式进行检测, 由于现网流量不断变化, 所以存在高误报的问题。为解决这一问题, 需要引入基于增量学习(基于欠采样提取白样本、基于过采样增广黑样本)的自适应AI技术能力。

该技术需要在现网采集大量样本, 对固化模型进行增量训练, 得到自适应检测模型, 利用自适应模型完成相关加密攻击流量的检测, 降低固化模型检测所存在的误报率高的问题。然而在样本采集过程中, 工程人员面临着白样本数量过多, 而黑样本数量过少, 导致样本分布失衡的情况。为解决这一问题, 本课题研究基于自适应增量学习技术动态更新检测模型。采用欠采样获取现网白样本, 利用过采样增广黑样本, 实现对入联复杂加密攻击(加密webshell攻击、加密扫描爆破)的检测。

本次研究引入了增量学习技术, 在不投入额外资源的前提下, 对原有固化模型能够进行“滚雪球方式”的增量学习

处理, 在此过程中, 原有固化模型能学习到最新的业务流量变化, 从而具备自适应检测能力。

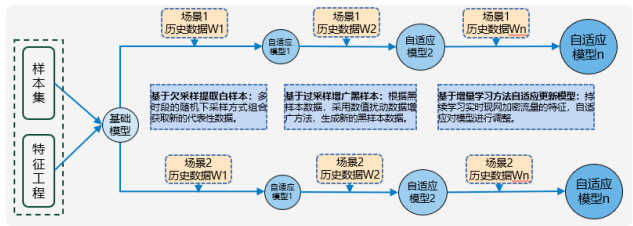


图3 增量学习技术模型迭代

自适应增量学习的检测流程如下图所示:

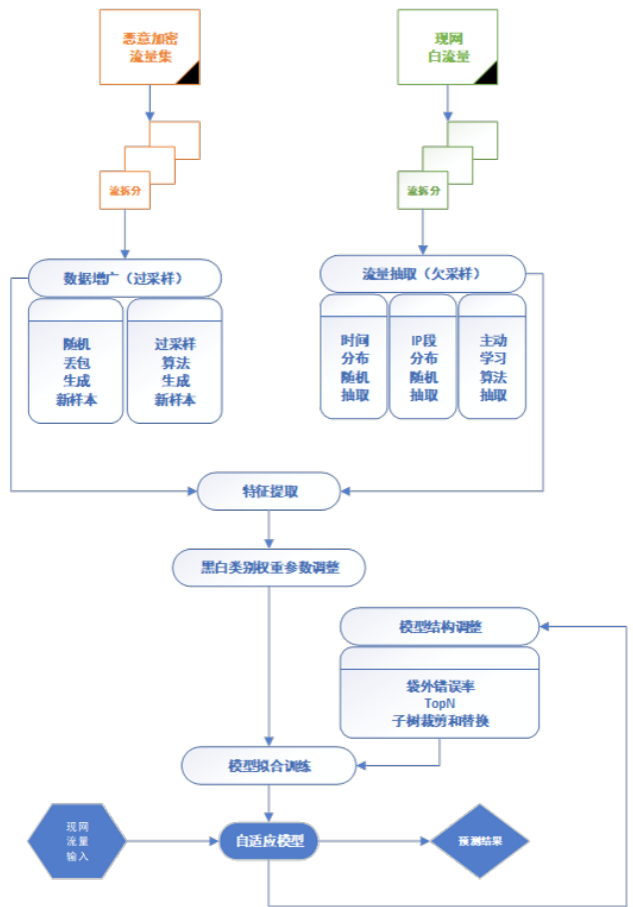


图4 自适应增量学习检测流程

3. 开展基于多流交互行为特征的隐匿攻击检测技术研究

近年来攻击者大量使用攻击工具利用CDN/云函数/域前置实现身份伪装, 其伪装方式以及所带来的问题如下: ①隐匿攻击方式: C&C构建时会采用CDN/云函数/域前置等白站中转方式, 导致恶意流量在协议特征上与正常流量无异, 实现隐匿效果。②当前防御体系问题: 由于IP、域名、SNI和证书都是白站信息, 所以传统基于IOC方式的规则、传统机器学习等检测手段完全失效。

经研究发现, 各种黑客工具、恶意软件在使用隐匿攻击时, 其依然包含攻击载荷下载、控制指令下发、数据回传等过程, 这些过程中的数据虽然被加密, 但是在流量的时间和空间行为上却存在行为特征, 可通过分析时间维度及空间维度

的恶意行为特征，利用不基于IOC的加密流量要素建立交互行为模型进行检测，实现对采用云函数、域前置、CDN等方式进行C&C隐匿加密攻击的威胁检测。

该检测技术路线的核心研究思路为：研究提取时间分布及空间分布两大维度的关键特征要素，识别上线、心跳、下发指令、数据回传等多种恶意交互行为，这些行为在时间和空间分布上存在如下特征：①时间分布维度：研究心跳会话之间、心跳会话与回传会话之间的间隔特征；②空间分布维度：研究HTTPS请求头及请求体、响应头及响应体的长度特征。

为了验证上述分析研究成果，研究中对通过对正常应用会话间隔和CS会话间隔通过构建时空分布图的方式进行了对比，分布图展示如下：

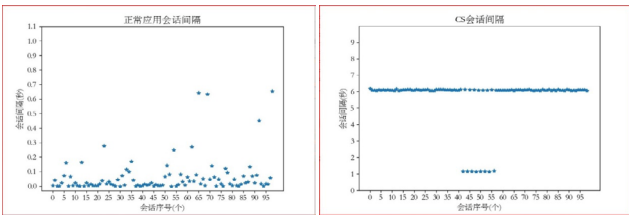


图5 正常应用会话间隔与CS会话间隔对比

从分布图中可以清晰看到正常应用会话间隔和CS会话间隔存在时间和空间上巨大差异。

基于上述研究成果，原型系统制定了如下的隐匿攻击检测流程：

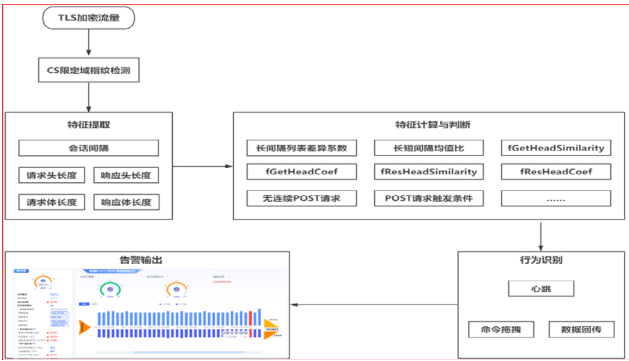


图6 CS 隐匿攻击检测流程

4. 开展基于集成学习的出联高对抗攻击检测技术研究

在进行高对抗攻击检测过程中，研究发现现网加密流量多变，导致单模型泛化性较差，从而出现单一检测技术和模型容易产生误报的问题，因此经研究后，本课题通过利用集成学习思想，由多个独立模型检测后进行综合决策，实现恶意加密流量的检测能力。另外，本课题还同步研究了对使用加密通信方式的恶意软件家族及APT组织的归因解决方案。

本课题基于集成学习思想，借鉴集成学习的思想，通过对加密流量进行深度分析，将加密流量所能提取的特征相对独立的分为：时空特征、握手特征、证书特征、背景流量特征，分别构建和训练单独的子模型进行检测。不同类型的SSL加密流量，根据所包含的信息，调用不同的子模型进行预测，最终在子模型输出的基础上，再进行加权平均得到最终预测结果。

由多个独立模型检测后进行综合决策，可有效实现对出联高对抗攻击的加密流量检测。

多模型工作原理如下图所示：

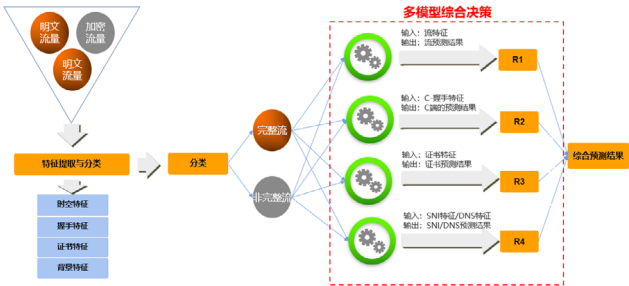


图7 多模型综合决策检测流程

针对不同的子模型准备相应的黑白样本并进行特征提取后：流特征训练固化为流模型、握手特征训练固化为握手模型、证书特征训练固化为证书模型、背景流量包含SNI特征与DNS特征训练固化为DNS模型。

模型固化过程使用不同算法进行对比，选取最优算法进行固化。采用多模型集成学习的方式，通过基于多模型机器学习综合检测技术对加密流量进行检测分析，极大的提高对复杂现网环境的适应能力，能够满足在各种不同复杂现网环境对恶意加密威胁的检测需求，检测的准确率远高于现有单一模型的检测能力。

三、研究成果和应用

(一) 研究成果

本课题的研究工作及核心成果紧紧围绕新技术的战略性和前瞻性研究、新技术新业务评估展开，为行业内针对加密流量威胁检测提供强有力的技术支撑。核心成果可以分为以下几个层面：

(1) 提出并充分验证了加密流量攻击高对抗场景的检测技术

根据加密流量高对抗攻击的变化趋势，提出了从入联、C&C隐匿、出联三个维度的基于人工智能的加密流量威胁检测技术。

(2) 提出了加密流量威胁检测AI应用及核心技术指标

嵌入人工智能各方向（监督机器学习、监督深度学习等），有机整合、扬长避短，构建以人工智能为内核的加密流量威胁检测系统。

本课题已形成一系列具有实践和指导意义的研究成果，如下：

形式类型	数量	研究成果类型
研究报告	1	基于人工智能的加密流量威胁检测技术研究课题报告
原型系统	1	原型系统1套（基于人工智能的加密流量威胁检测系统）
试点应用	1	在我方企业现网出口位置部署应用
行业示范	/	通过在现网部署试点应用后，验证成果形成落地实践，拟在证券期货行业形成示范

表2 研究成果

构建的基于人工智能的加密流量威胁检测原型系统，能够解决在加密流量攻击高对抗在技术、工程、实战方面的几个核心问题：①技术问题：入方向加密攻击检测误报高；出方向加密通信无法审计内容；C&C隐匿攻击缺乏检测手段。②工程问题：高对抗样本少且质量差；特征工程构建质量差；大网网络质量差导致检测容错性差。③实战问题：APT加密攻击发现能力差；CS等重要威胁监测能力差；webshell等加密攻击识别能力差。

通过嵌入人工智能各研究方向，包括监督机器学习、监督深度学习等，有机整合、扬长避短，构建以人工智能的加密流量威胁检测安全防护原型系统。构建本系统的核心要求是实现人工智能的全面集成、精准检测、深度融入。

该原型系统具备入联复杂加密攻击检测能力，具备检测CS隐匿加密通信能力，具备恶意加密通信流量检测及家族组织归因能力，具备对隐蔽隧道通信检测能力，具备APT加密攻击检测能力。

原型系统可应用的实战对抗场景如下：①可部署在业务互联网及办公网的边界出口位置，旁路的方式镜像接入流量，单台设备接入性能可达10Gbps。②全面监测网络中各类重要加密威胁事件（恶意软件家族加密通信、隐蔽隧道加密通信等）。③监测预警APT组织加密攻击流量，包括方程式、图拉、APT28奇幻熊、摩诃草、拉撒路、蓝宝菇、海莲花等。

（二）主要创新

1. 技术创新

基于增量学习的入联攻击检测技术解决当前固化模型误报高的问题；基于集成学习的出联加密流量检测技术解决当前恶意加密流量无法审计内容的问题；基于恶意交互行为的C&C隐匿攻击检测技术解决当前完全没有检测手段的问题。

2. 工程创新

综合利用现网采集、样本清洗、多层面细粒度的特征工程构建能力解决当前高对抗样本差、特征工程构建质量差的问题。

3. 实战创新

基于深度学习进行分类识别，实现APT加密协同攻击检测能力；基于交互行为精准特征，实现现网CS、Sliver等重要威胁监测能力；基于自适应AI能力，具备加密攻击识别能力。

（三）实践验证

完成本原型系统能力构建后，针对其进行测试应用实践验证，采用离线流量包回放验证、现网验证方式。

（1）离线包回放验证

离线方式快速测试产品核心功能的完备性及能力，对加密流量威胁高对抗攻击的检测能力进行快速验证，重点包含专项威胁验证：远控木马、隐蔽隧道、加密webshell、代理工具等。

（2）现网验证

部署在业务互联网及办公网的边界出口位置，接入现网流量，验证原型系统的实战能力。

验证中主要用的组网方式如下：

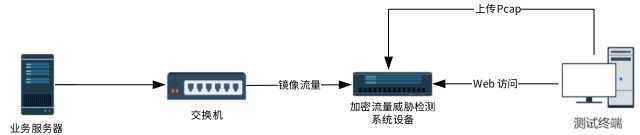


图8 实践验证组网

设备部署在现网完成安装调试，接入现网镜像流量，同时使用测试终端上传离线流量包进行测试。

离线流量包回放验证提供6类攻击类型，涵盖加密流量高对抗攻击各种重要类型，共82个第三方攻击样本。针对不同加密流量场景，具体如下表所示：

序号	加密流量高对抗攻击	威胁类型	样本数量
1	重要专项攻击（CS、银狐）	CS隐匿攻击	11
2	远控木马	TLS加密回连	26
3	Webshell	哥斯拉、冰蝎、蚁剑	16
4	隐蔽隧道	ICMP、DNS、HTTP	10
5	代理工具	加密通信	11
6	APT加密攻击	加密通信	8

表3 加密流量高对抗攻击离线流量测试验证

本原型系统检测告警总数82项，实现全部检出。经过测试验证，本原型系统对恶意软件、隐匿攻击、隐蔽隧道、加密webshell以及APT等加密威胁均有良好检出能力。同时本原型系统在现网部署在业务互联网及办公网的边界出口位置，接入现网流量，应用监测验证实战能力。

当前已初步研制出基于人工智能的加密流量威胁检测技术防御原型系统，初步形成加密流量检测解决方案，经测试验证，能有效检测预警远控木马、隐蔽隧道、加密webshell、代理工具等类型加密流量攻击。

（四）价值总结

1. 提升网络安全能力

通过本项目的实施，研制基于人工智能的加密流量威胁检测原型系统，能够有效提升网络安全体系在加密流量高对抗网络攻击场景的防御能力，这将有助于我司提升网络安全防护水平，降低网络安全威胁带来的损失。

2. 推动技术研究进步

在项目研究过程中，已研发出创新的算法和模型，这些技术和方法将推动人工智能在网络安全领域的应用深度和广度，促进相关技术的发展。

3. 提升行业影响力

通过本项目的研究成果，预期能够为金融证券期货行业在网络安全领域提供新的针对加密流量威胁检测技术和方法，影响和提升行业内在高对抗网络攻击的体系建设和防御能力。

高效特权账号管理体系技术研究 与落地实践

丁安安、夏英杰、丁旦 | 国联民生证券股份有限公司

摘要：在数字化转型中，企业数字资产安全面临严峻挑战。特权账号作为掌控核心系统与数据的“关键钥匙”，其跨系统、高权限特性使特权账号失管或滥用可能引发数据泄露、业务中断等重大风险。伴随云计算、大数据等技术应用，企业IT架构复杂化导致特权账号数量激增，传统静态管理模式已难以应对动态安全需求。构建现代化特权账号管理体系需从四个维度着手：顶层设计层面制定全生命周期安全策略；融合智能技术实现自动化管控与风险预警；优化跨部门协作流程确保策略落地；强化人员安全意识培训筑牢最后防线。通过剖析本机构实践，本文提炼出覆盖“账号管理-权限权利-密码管理-审计分析-安全培训”的全生命周期特权账号管理体系，助力企业构建适应动态环境、符合合规要求的安全体系，在保障业务连续性的同时实现安全与效率的平衡，为数字化转型筑牢安全根基。

关键词： 特权账号、密码管理、权限管控、运维安全审计

一、课题研究背景

（一）课题研究背景

特权账号是企业信息技术系统中具有特殊权限的用户账号，赋予持有者系统维护、权限增加、数据修改删除、导出等高级权限，通常用于连接企业核心资源，对关键系统、数据库、网络设备和应用程序进行高权限操作。正是由于其特殊性，证券行业相关管理规范以及审计指南均对特权账号管理提出了具体要求：

1. 《证券期货业信息系统运维管理规范JR/T 0099—2012》：应每季度对管理员口令进行修改，更新的管理员口令至少5次内不能重复；应明确责任人，负责统一保管、安全存放管理员口令。

2. 《证券期货业信息系统审计指南第5部分：证券公司JR/T 0146.5—2016》：是否授予不同账户为完成各自承担任务所需的最小权限；是否实现操作系统和数据库系统特权用户的权限分离；应在保证系统运行安全和效率的前提下，启用系统审计或采用第三方安全审计产品实现审计要求。

二、证券行业特权账号管理问题分析和 课题研究目标

在本课题前期准备阶段，我单位也对行业内特权账号管理的情况进行了小范围的调研：

1. 约18%的单位部署有特权账号管理系统，其余均未在特权账号管理方面做额外投入，但均有堡垒机系统以满足运维操作和审计的基本要求。2. 统一管理与最小权限原则成为行业共识，但落地与执行仍存在挑战。3. 系统化解决方案是未来大的趋势，但集成与扩展性仍需提升。

经过调研发现，在特权账号管理方面，目前行业机构可能存在以下共性问题：

1. 系统建设和人员投入有限。2. 密码管理方式相对传统且分散，对人工依赖程度较高。3. 定期改密的操作成本和附加运行风险较高。4. 密码复杂度无法保证，或者多台机器使用相同的密码。5. 操作系统和数据库特权用户的权限分离难以实现。6. 缺乏运维安全运营体系。

本课题的研究目标是尝试探索一种综合考虑成本、安全合规、可落地等因素的特权账号管理体系，在多方需求之间寻找平衡点，提升行业机构的整体安全防护能力，为企业系统安全稳定运行提供保障。

三、特权账号管理体系研究

经过深入的前期理论与丰富的技术积累，课题研究小组认为特权账号管理是一项复杂而系统的工程，它不仅涉及多个环节的紧密配合，更要求每个环节之间形成相互关联、相互促进的闭环体系。特权账号的闭环管理包括账号集中管理、权限统一分配、定期改密、运维操作审计、特权账号行为分析、周期性检查与评估、定期安全培训等多个环节。



图1 特权账号管理体系

1. 账号集中管理：搭建统一管理平台，集中存储、管控、检索特权账号，规范账号全生命周期流程，做到操作留痕，提升管理效率与安全水平。

2. 权限统一分配：依照岗位角色划分权限，严守最小权限原则，依托细粒度访问策略，结合业务与合规要求动态授权，防范权限滥用。

3. 定期改密：推行强密码规则，要求定期更换密码，借助自动化改密工具简化管理，强化密码安全。

4. 运维操作审计：全程记录账号操作信息，结合实时监控与事后分析，留存审计日志、定期汇总结果，实现行为可追溯、风险早排查。

5. 特权账号行为分析：运用智能技术搭建行为模型，对比历史数据识别异常操作，主动预警安全风险。

6. 周期性检查与评估：定期全面核查各管理环节，排查问题并落地整改和优化。

7. 定期安全培训：常态化开展安全培训，配套考核机制，强化人员安全意识与实操能力。

四、特权账号管理技术实践

经过前期调研和研究，结合我单位实际技术情况，我单位逐渐摸索出了一套相对完整、但仍需不断完善优化的特权账号管理思路。

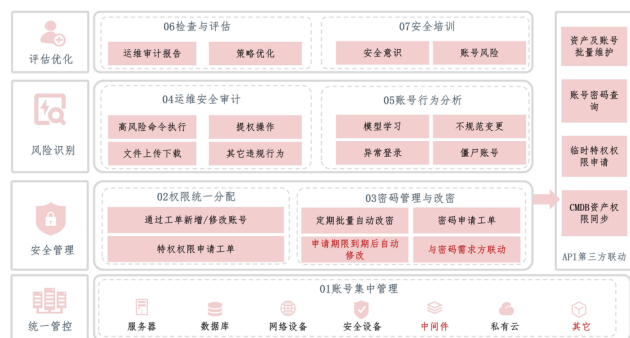


图2 特权账号管理设计和技术实践

（一）账号集中管理

行业内机构普遍部署有堡垒机系统，作为统一运维入口、满足基础审计需求。机构可依托现有体系拓展纳管范围，统一管理服务器、数据库、网络及安全设备、中间件等资源，搭建一体化特权账号管理底座。针对各类应用来源多样，部分需高权限账号运行的问题，我方尝试了新的思路，针对可降权运行的应用统一配置普通账号，普通与特权账号全部纳入堡垒机管理。另外搭配主机安全探针，常态化排查系统账号风险，持续推进风险整改。

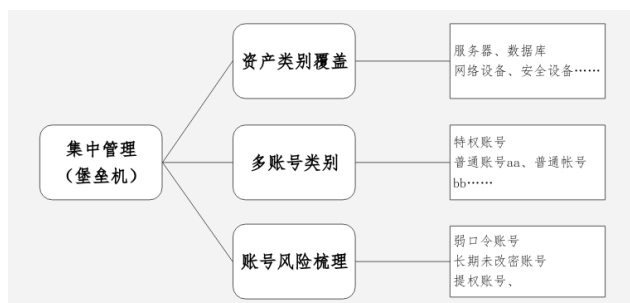


图3 账号管理体系

（二）权限统一分配

账号集中管理后，堡垒机作为统一运维登录入口在权限管理方面具有先天优势。在管理上可为用户（员工）默认只分配分配普通账号登录操作权限，并辅以完善、便捷的权限申请流程，方便用户按需申请短期临时特权账号权限。

常见需申请特权权限操作的场景如下：

1. 安装“明确需要使用特权权限安装”的第三方软件。
2. 执行计划中的变更升级。
3. 基础资源维护，如挂载新硬盘等。
4. 操作为普通账号根据最小化权限原则赋予所需权限。

（三）密码管理与定期改密

堡垒机系统通过RDP/WinRM/SSH等协议，按预设策略自动修改目标设备特权账号密码，新密码由系统生成并同步至密码库，实现运维人员免密登录且全程无感知。管理流程标准化：业务团队提交改密申请（含资产范围、周期等），经审批后自动配置执行任务，结果通过加密邮件反馈，逐步弱化人工接触密码。

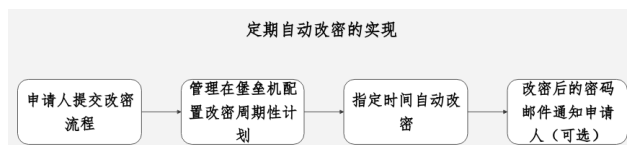


图4 定期改密过程

系统提供密码查询API接口，可与运维工具集成，构建全生命周期管理体系。典型应用场景包括：

1. 普通账号提权操作。
2. 资源管理员带外维护。
3. 安全扫描器自动登录检测。
4. DevOps平台资源初始化部署。

该机制通过自动化密码管控，在提升维效率的同时，显著降低特权账号泄露风险，强化企业安全合规能力。

（四）特权账号运维安全审计与账号行为分析

特权账号运维安全审计的落地需以全链路技术闭环为核心，通过日志接入模块实现特权操作日志的无缝采集、经过多种实时处理引擎的分析，最终进行风险预警与结果可视化。

1. 日志接入-多源异构安全审计日志的统一采集

运维安全审计系统的日志接入模块作为核心数据入口，需实现多源异构日志的高效采集、清洗与标准化。该模块支持多设备类型及场景的日志接入，涵盖结构化、半结构化及非结构化数据的解析需求。

架构采用“采集器+中心化处理引擎”混合模式：在数据源端，通过轻量级代理（如Filebeat/Fluentd）实现主机日志采集，并配置资源限制参数确保业务稳定；对网络设备则通过Syslog协议+TLS 1.3加密传输实现无代理接入。

中心处理层基于Apache Nifi构建可视化流水线，执行三阶段处理：首先通过正则匹配与机器学习模型进行日志分类；其次采用AES-256算法对敏感字段脱敏加密；最终通过时间窗口算法将高频日志聚合为批处理事件。

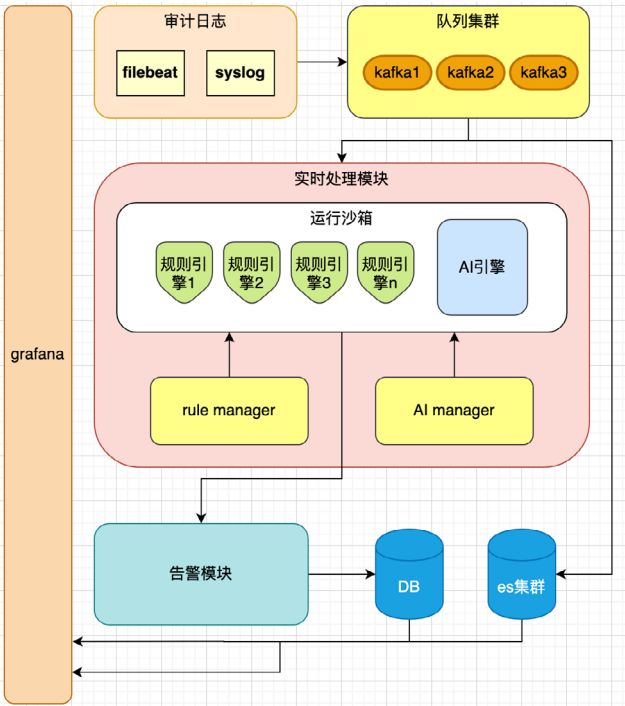


图5 多元日志接入及处理架构图

2. 日志实时处理-双引擎驱动的智能决策体系

规则引擎：基于确定性逻辑的快速响应机制

规则引擎的核心在于将安全专家的领域知识转化为可执行的检测逻辑，其架构分为规则管理层、规则编译层、运行时执行层三个层次。在规则管理层，采用声明式规则描述语言（DSL）定义检测策略。

```
rule "异常文件访问检测" {
  when:
    $log: LogEvent(type=="FILE_ACCESS",
      user not in ["admin", "auditor"],
      extension in [".conf", ".key"],
      access_time not in ["09:00-18:00"])
  then:
    triggerAlert(LEVEL.HIGH, "非授权时段访问敏感文件");
}
```

图6 规则设计示意图

规则库采用版本控制管理，支持灰度发布与热加载。规则编译层通过JIT（即时编译）技术将DSL规则转换为底层字节码，利用有限状态机优化匹配效率。

AI引擎：基于机器学习的深度威胁发现

AI引擎采用“离线训练+在线推理”的双层架构，核心技术栈涵盖特征工程、模型服务化、持续学习三大子系统。构建多维日志向量空间，涵盖时序、空间、行为三类特征。模型分为监督、无监督两类：LightGBM识别已知攻击，准确率98.7%；深度自编码器检测未知异常，F1值92.4%。基于KServe搭建推理服务，千维特征计算耗时低于50ms。系统每月自动重训，借助GAN生成攻击样本，模型误报率控制在0.5%以内。

持续学习子系统工作原理：

1. 新数据收集：累积生产环境中的预测结果和反馈数

据。2. 样本增强：通过对抗训练生成边界样本。3. 增量训练：采用参数微调方式更新模型权重。4. 模型验证：基于时间窗口划分验证集评估性能。5. 部署上线：通过金丝雀发布策略逐步替换旧模型。

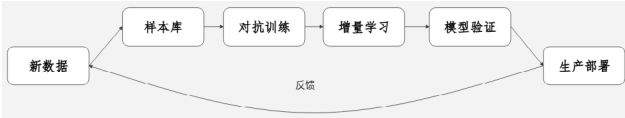


图7 持续学习子系统原理

双引擎协同机制

规则引擎与AI引擎通过事件总线实现深度协同：AI引擎输出的风险评分会动态调整规则引擎的检测阈值（如将高风险IP的登录失败告警阈值从5次降至3次），而规则引擎筛选出的可疑事件会优先送入AI模型进行深度分析。这种“规则过滤+AI精判”的级联架构，使系统在保持每秒百万级事件处理能力的同时，有效地提升了告警检出率。

3. 告警-闭环式安全事件管理

告警模块构建了从事件生成到处置反馈的完整生命周期管理体系。在事件生成层，采用多级降噪策略：首先通过时间衰减算法抑制重复告警（如10分钟内相同IP的多次登录失败仅触发一次告警），告警通知采用分级响应机制：低风险事件自动生成工单进入数据库，中高风险事件通过调用公司cmdbapi接口，推送到公司统一告警平台。

4. 数据存储与可视化-多模态数据的统一洞察

Elasticsearch集群采用Hot-Warm架构实现成本优化：热节点使用NVMe SSD存储最近7天数据，提供每秒10万次的实时查询；温节点使用HDD存储历史数据，通过ILM（索引生命周期管理）自动执行段合并优化。在Grafana可视化层，构建了二级仪表盘体系：

态势感知：实时滚动的事件风暴图，通过桑基图展示攻击路径的扩散趋势。

威胁分析：支持将任意告警事件下钻到原始日志上下文，通过时间轴展示威胁信息的实时情况。

5. 系统演进-从合规审计到主动防御

当前系统未来将向两个方向演进：首先构建“数字孪生”沙箱，对高危操作进行模拟推演；其次是构建自主响应能力，当检测到异常操作时，真正实现从“被动审计”到“主动防护”的质变。

6. 部分审计成果展示

(1) 不规范变更行为

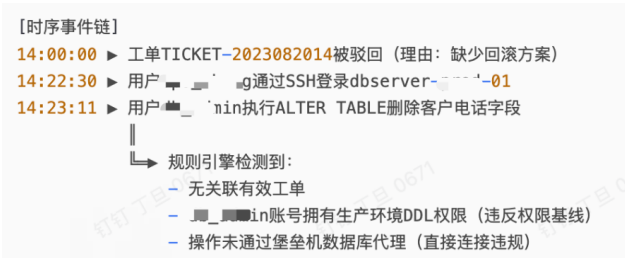


图8 不规范变更行为为审计示意图

根据多种日志内容,进行cep组合定位。例如监测命中相关数据库变更语句后,进行上下文定位。确定是否具备相关工单,相关审批状态是否为success,变更执行时间是否符合工单规定的执行时间段等。

(2) 运维执行高风险命令

根据操作系统的审计日志扫描,匹配基础规则

```
# 高危命令特征库 (正则表达式)
high_risk_patterns = [
    r"rm\s+--rf\s+var/log",      # 递归强制删除日志目录
    r"dd\s+if=/dev/urandom",    # 磁盘覆盖操作
    r"chmod\s+777\s+etc",       # 敏感目录权限变更
    r"mysqldump\s+--all-databases" # 全库导出
]

# 实时命令流扫描
def rule_engine(command):
    for pattern in high_risk_patterns:
        if re.search(pattern, command):
            return True, pattern
    return False, None
```

图9 高风险命令执行检测逻辑

检测过程推演

```
[原始日志流]
15:29:50 ▶ 用户 d...通过堡垒机,建立SSH会话SSH-20230820-1530
15:30:00 ▶ 执行命令序列:
          sudo -i & cd /var/log/nginx & ls -al & df -h & rm -rf *
15:30:06 ▶ 文件系统检测到/var/log/nginx下1.2TB日志被删除

[检测引擎处理]
1. 规则引擎命中高危模式: rm -rf /var/log/*
2. 上下文分析:
   - 前序操作包含权限提升(sudo -i)和空间检查(df -h)
   - 操作时段 (15:30) 非预设维护窗口 (00:00-03:00)
3. 基线比对异常:
   - 用户 d...历史rm命令占比仅0.3%
   - 首次对web-...~01执行大规模删除
4. 综合风险评估: 40(敏感目录)+30(root权限)+20(空间检查)+10(时段)=100/100
```

图10 检测过程展示

(3) 违规上传下载文件

主要日志来源

```
堡垒机文件传输日志:
log
2023-09-15T11:22:33.456Z | user: "d..." | session_id: "SFTP-20230915-1122" |
operation: "PUT" | local_path: "/data/export/customers.csv" | remote_path: "sftp://
an.baidu.com/upload" |
file_size: 128MB | hash: "sha256:9f86d081884c7d659a2feaa0c55..."

数据防泄漏 (DLP) 扫描日志:
log
2023-09-15T11:22:35.001Z | file: "customers.csv" |
matched_rules: ["DLP-003 客户身份证号", "银行卡号"] |
sensitive_count: 15,328 | risk_level: "CRITICAL"

网络代理访问日志:
log
2023-09-15T11:22:34.789Z | src_ip: "10.5.1.1" | dst_ip: "45.76.139.11" |
domain: "pan.baidu.com" | method: "POST" | url: "/upload" |
user_agent: "WinSCP/6.1.2" | bytes_uploaded: 134217728
```

图11 日志来源

部分文件内容分析逻辑

```
# DLP引擎检测流程
def dlp_scan(file_path):
    # 步骤1: 文件类型识别
    file_type = magic.from_file(file_path) # 返回: "CSV text"

    # 步骤2: 敏感模式匹配
    patterns = {
        "身份证号": r"\b[1-9]\d{5}{18|19|20}\d{2}{0[1-9]|1[0-2]}(0[1-9]|[12]\d|3[01])\d{3}[0-9Xx]\b",
        "银行卡号": r"\b(62[0-5]|6[6-9])\d{14,17}\b"
    }

    # 步骤3: 内容抽样扫描
    with open(file_path, "r") as f:
        sample = f.read(1024*1024) # 扫描首1MB内容
        for field, regex in patterns.items():
            if re.search(regex, sample):
                return field, len(re.findall(regex, sample))

    return "CLEAN", 0
```

图12 文件内容分析逻辑

主要检测过程推演

```
[事件时间线]
11:22:30 ▶ 用户 d...通过堡垒机SFTP连接生产服务器10.20.1.10
11:22:33 ▶ 发起文件上传: /data/export/customers.csv -> pan.baidu.com
11:22:34 ▶ 网络代理检测到百度网盘域名访问 (触发规则ID: NET-107)
11:22:35 ▶ DLP引擎实时截获传输流, 识别15,328条敏感数据
11:22:36 ▶ 风险判定完成并阻断传输

[多源日志关联]
1. 堡垒机日志 -> 识别SFTP高危操作 (PUT到外部地址)
2. 网络日志 -> 确认公网文件传输 (45.76.139.11为百度网盘IP)
3. DLP日志 -> 验证文件含身份证号/银行卡号 (命中DLP-003/DLP-007)
4. 威胁情报 -> 标注目标为个人网盘 (风险等级CRITICAL)
```

图13 检测过程示意1

(五) 检查评估与安全培训

通过自动化工具定期生成涵盖风险事件与操作合规性的审计报告,由IT审计部门与安全团队协同开展事件审查和复核。对确认的不规范操作,推动责任部门落实整改并完善运维规范;针对模型误报问题,基于实际业务场景持续优化行为分析阈值与审计规则集。同时深化与变更管理流程的联动,建立操作命令白名单机制,通过工单系统前置审批关键操作,从源头减少事后审计误判。

在人员能力建设方面,将密码安全、账号风险管理等内容纳入年度网络安全意识培训体系,采用案例教学与实操演练相结合的方式,每年至少开展1次全员覆盖的专项培训,重点强化一线运维人员的合规操作意识与风险处置能力,形成“技术防控-流程管控-人员意识”三位一体的治理框架。

(六) 接口管理及自动化

在实践中可充分利用相关系统现有接口和自研能力把简单重复的工作交给程序代码来完成,减少管理人员日常运维操作工作。

1. 堡垒机系统可与CMDB联动,定时从CMDB获取(或CMDB主动推送)系统列表以及系统所属负责人员信息,包括对应应用系统下的服务器资产清单。



图14 CMDB资产归属同步

2. 新申请计算资源（服务器、虚拟机）根据资源申请流程中所填写的信息，调用堡垒机API接口自动将资产及账号添加到堡垒机中，资产自动归属到一应用系统中并为运维人员自动分配非特权账号登录权限。

3. 当运维人员需要使用特权权限执行系统升级、磁盘挂载等特殊操作时，可提交临时权限申请流程，通过调用堡垒机接口，管理人员可仅通过点击审批流程实现多人员、多资产账号、不同有效期限等复杂权限需求场景的一键设置。

（七）课题研究实施过程中遇到的问题及解决方法

在课题推进过程中，我们确实遭遇了若干技术瓶颈与实施阻碍。但正是在公司技术部门领导的鼎力支持下，以及课题组成员齐心协力、攻坚克难的持续努力下，项目才得以突破重重困境，最终取得了现阶段的阶段性成果。主要问题及解决方法如下：

问题1. 特权账号管理系统建设成本投入较高，短期内IT预算有限

在需求调研阶段，我单位也对市场上主流的特权账号管理系统进行了调研，发现系统建设成本对于一般中小型企业来说较高。最终通过对现有堡垒机平台进行定制化开发，增加权限分配、密码查询、资产账号配置等API接口，以较低的投入实现了对特权账号进行集中、细粒度管控的核心需求。

问题2. 账号管理的边界如何确定

通过制定内部信息系统账号及权限管理细则，将实际工作中需要使用的各类账号进行分类，明确特权账号、运维管理账号、受限账号、赋权账号和普通账号的分类及管理要求，将操作系统登录相关的账号纳入本课题研究及后续推行实施的范畴。课题小组重点聚焦整体框架、解决方案的研究，而非普通账号应该有哪些权限这样的具体标准。

问题3. 存量信息系统特权账号管理推进

对于存量信息系统，首先确定相关应用是否支持以非特权权限运行。对于支持以非特权权限运行的自研应用和第三方采购应用，由系统管理员补充创建普通账号并纳入到堡垒机平台统一管理。制定特权账号权限回收计划，最终在堡垒机上收回特权账号权限，实现“默认仅分配普通权限，按需申请特权权限”的预期目标。

（八）特权账号管理极端情况应对

根据公司内部相关制度要求，系统管理员账号密码会定期进行封存，特殊情况下可启用封存的账号密码以应对极端情况。

堡垒机系统本身为多节点集群部署，2个节点出现异常时仍不影响对外提供服务，系统全量配置文件每日会进行自动本地和异地远程双备份；堡垒机系统支持配置密码管理员、配置管理员角色，特殊情况下可由配置管理员为密码管理员赋权、由密码管理员打印所需资产账号和密码，实现双人复核下的手工密码查询。计算资源管理员也可以通过控制台、带外等方式使用非对外交付的保留账号登录进行资源层面的应急处置操作。

五、特权账号管理未来展望

在数字化转型与网络威胁加剧的双重驱动下，特权账号管理（PAM）已成为企业安全架构的核心支柱，呈现三大技术演进方向：

场景化密码管理向新兴领域深度延伸，从传统IT资产覆盖至云原生环境、DevOps工具链及物联网设备。例如，与Kubernetes集群集成可实现容器化应用动态凭证管理，践行最小权限原则。API标准化成为关键突破口，企业可通过RESTful/GraphQL接口将PAM能力嵌入自动化流程：在CI/CD管道中调用短期权限令牌，规避硬编码风险；联动SIEM/SOAR平台构建安全响应闭环，实现异常行为触发密码轮换或会话终止。

智能自动化技术加速融合，AI驱动的权限分析引擎通过机器学习优化访问策略，自动识别异常并动态响应。利用RPA技术简化特权任务执行，形成“智能决策-自动执行-全程留痕”闭环，降低人为风险并强化审计追溯。

零信任架构与PAM深度耦合，突破传统账号管理范畴，成为动态安全边界的关键节点。通过持续验证设备状态、用户身份及上下文信息，构建“动态授权+最小权限”体系。结合UEBA技术实时比对行为基线，异常时触发二次认证或权限降级，推动安全防护向动态信任评估转型，形成覆盖“身份-权限-行为”的全周期管控能力。

参考文献

- [1] JR/T 0099—2012《证券期货业信息系统运维管理规范》
- [2] JR/T 0146.5—2016《证券期货业信息系统审计指南第5部分：证券公司》
- [3] JR/T 0060—2021《证券期货业网络安全等级保护基本要求》

投稿邮箱：
wsil@sse.com.cn

ITRDC

ITRDC证券信息技术研究发展中心(上海)



中国上海市杨高南路388号

邮编: 200127

公众咨询服务热线: 4008888400

网址: <https://www.sse.com.cn/>

内部资料 免费交流

本资料仅为内部交流使用, 本期印200册, 编印单位为上海证券交易所, 面向证券期货行业发送, 印刷时间为2026年9月, 印刷单位为上海邦达彩色包装印务有限公司。

部分图片或文字来源于互联网等公开渠道, 其版权归属原作者所有, 如有版权相关事宜, 请发送邮件至wsil@sse.com.cn。