2025年中国安全大模型行业概览

AI重构网络安全：大模型如何颠覆攻防博弈？

China large Language Model for Security Industry

中国安全大モデル産業

头豹研究院

摘要

网络安全大模型作为新兴力量，显著增强了防护能力，但其安全性亦需全面保障。AI在网络安全领域作用凸显，但不当应用亦带来新风险，需警惕其在网络攻击中的潜在增强作用及数据泄露风险。大模型发展面临数据投毒、隐私泄露等多重挑战，要求构建涵盖基础设施、能力支撑、安全服务及运营管理的全方位架构体系。中国凭借领先的算力与数据资源，推动安全大模型从小至大演进，实现了对安全风险的快速预测与检测，并促进了定制化安全解决方案的发展。技术层面，预训练-微调范式、RLHF技术及LLM智能体构建等，共同提升了AI在复杂安全任务中的高效应用能力。目前，AI大模型在网络安全中已提升自动化水平，未来有望全面替代传统产品，成为核心力量。然而，基础大模型在安全领域应用受限，需通过增量预训练注入行业知识以增强实用性。安全大模型通过自动化处理、多源数据整合与持续学习，优化了威胁分析与防御策略，为组织提供长期安全保障，优化安全运营并提升用户体验。

- **网络安全大模型提升防护力，但需全面保障安全，以应对多重挑战**

网络安全大模型通过融合知识、技术与数据，显著提升了防护能力。目前，大模型技术能实现对安全风险的快速预测与检测，准确率高达95%以上。然而，其发展也面临数据投毒、隐私泄露等严峻挑战，如去年全球范围内发生的数据泄露事件中，涉及大模型的占比高达30%。因此，全面加强安全防护和监管，确保技术健康稳定发展至关重要。中国作为领先者，其安全大模型行业已历经小模型到大模型的演进，为行业树立了典范。

- **中国算力领先，AI大模型需高质量数据支撑，技术与应用正加速网络安全升级**

中国在全球算力领域占据领先地位，AI算力支出稳居榜首，为安全大模型的发展奠定了坚实基础。然而，构建高效安全大模型需依赖高质量数据，包括多源采集、精细筛选与持续清洗等环节，以确保模型精度与国际竞争力。目前，AI大模型在网络安全中的应用已显著提升自动化水平，预计未来几年将全面替代传统产品，实现全面网络防御功能。通过预训练-微调范式及RLHF技术优化，大模型性能不断提升，威胁分析效率提高30%，防御策略精准度提升25%。同时，基础大模型在安全领域的应用受限，需通过增量预训练注入行业知识，以增强其实际应用效果。

内容目录

研究目标

研究目的

了解安全大模型的技术演变、探析产业链生态图谱，洞察厂商商业模式并探析安全大模型行业、业务场景以判断行业发展趋势。

研究目标

- 了解中国安全大模型的背景、定义
- 探究中国安全大模型技术演变
- 探析中国安全大模型行业产业链情况
- 分析中国安全大模型的行业应用场景
- 预判中国安全大模型行业发展态势

本报告的关键问题

- 参与者：中国打造安全大模型的企业有哪些？哪些企业更具备发展潜力与优势？
- 发展趋势：中国安全大模型发展面临哪些机遇与挑战？发展驱动力有哪些？

名词解释

- **模型窃取：**模型窃取是指攻击者通过多次查询目标模型，获取其全部或部分信息（如模型参数、结构等），以构建功能相近的替代模型，进而利用该模型提供服务或进行对抗样本攻击，对机器学习即服务场景构成安全威胁。
- **API接口：**API接口是一组预定义的方法和协议，用于不同软件组件或系统之间的交互。它定义了应用程序如何请求和接收数据，简化了软件组件的复杂性，使不同应用程序能够高效、安全地访问服务。
- **模型精调（Fine-Tune）**：模型精调是指对预训练的AI模型进行再训练，以更好地适应特定任务或数据集。通过精调，可以显著提高模型在特定领域的表现，减少训练时间，并优化模型性能。
- **非私有化预训练：**非私有化预训练是指在公共或大规模数据集上对模型进行初步训练，使其具备基础的特征表示能力。这种训练方式允许模型在后续任务中通过微调快速适应，但训练数据和模型本身不局限于单一组织或个体。
- **安全任务编排：**安全任务编排是将企业或组织在安全运营中涉及的不同系统或组件的安全功能，按照一定的逻辑关系组合起来，以自动化、高效地响应和处理安全事件。它旨在提高安全响应的准确性和及时性。
- **安全Copilot：**安全Copilot是一种基于生成式AI的安全解决方案，为安全专业人员提供自然语言辅助助手体验。它集成于安全产品组合中，帮助分析人员处理安全信号，以机器速度和规模改善安全结果，提高防御效率。
- **SIEM（安全信息与事件管理系统）**：SIEM是一种集成了安全信息管理和安全事件管理功能的安全技术，能够实时监测、检测、分析和报告安全事件，帮助企业或组织保护其信息系统和数据免受安全威胁。
- **安全领域垂直站点**：安全领域垂直站点是专注于安全领域的专业网站或平台，提供深入的安全技术、产品、解决方案及行业动态等信息。这些站点通常面向特定行业或领域的安全从业者，提供垂直化的服务和资源。
- **模型微调（Model Tuning）**：模型微调是在模型精调的基础上，对模型参数进行细微调整以优化其性能的过程。通过微调，可以进一步改善模型在特定任务上的表现，使其更加适应实际应用场景的需求。

Chapter 1
行业综述

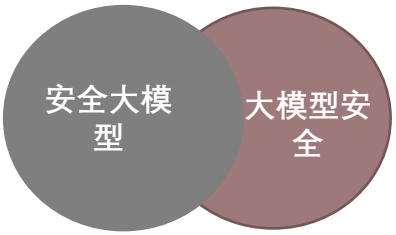
- 行业认知：网络安全大模型提升防护力，大模型安全需全面保障。厂商积极应用，融合知识技术数据，产品以安全智能体和聊天机器人为代表，强化安全防护与模型安全
- 市场背景：AI在网络安全中作用显著，但应用不当也带来风险。需加强AI在保护系统免受攻击方面的应用，同时警惕其在网络攻击中的增强作用，并重视大模型带来的数据泄露和劫持风险
- 发展痛点：大模型发展面临数据投毒、隐私泄露、记忆风险、内容合规性、透明度不足及数据传输安全等多重挑战，需全面加强安全防护和监管，确保技术健康稳定发展
- 架构体系：安全大模型平台架构包括基础设施、能力支撑、安全服务和服务与运营四层。基础设施涵盖底层设备，能力支撑提供关键技术支持，安全服务层核心保护数据和网络，服务与运营层关注整体平台运作与管理
- 发展历程：中国安全大模型行业历经小模型到大模型的演进，大模型技术以其强大的学习和推理能力，实现了对安全风险的快速预测与检测，并推动定制化安全解决方案的发展，构建了全方位的安全防护体系

中国安全大模型行业综述——行业认知

- 网络安全大模型提升防护力，大模型安全需全面保障。厂商积极应用，融合知识技术数据，产品以安全智能体和聊天机器人为代表，强化安全防护与模型安全

中国安全大模型行业认知

安全大模型是针对特定安全垂直领域的大型语言模型，通过大量的数据训练和复杂的算法设计来满足特定的安全需求，旨在解决各类复杂的安全应用场景需求，如网络安全攻防、安全运营等。



通常指的是涉及大规模数据集、复杂算法模型以及广泛应用场景下的AI系统安全，不仅包括了模型本身的保护，还涉及到数据的隐私保护、模型的鲁棒性以及合规性等多个方面。

- 安全大模型指的是“网络安全大模型”，通常指的是利用大规模数据集和复杂算法（如深度学习、机器学习模型）来提升网络安全水平的方法和技术。这种模型利用大量的网络流量、用户行为、设备活动等数据进行分析，以识别潜在的安全威胁、异常行为和攻击模式。网络安全大模型能够帮助检测未知的威胁、预测未来的攻击趋势，并提供自动化的响应机制
- 大模型安全则是涵盖了所有关于大规模数据处理和机器学习模型的安全性问题。大模型安全不仅包括了模型本身的保护，还涉及到数据的隐私保护、模型的鲁棒性、公平性、透明度以及合规性等多个方面。大模型安全关注的是确保这些复杂模型在整个生命周期中的安全性，从数据收集、模型训练、测试、部署到维护的每一个环节。
- 网络安全大模型也要保证模型本身的安全。大模型本身的安全性问题包括隐私数据泄露、算法模型可解释性难度大、模型可靠性问题等。这些问题不仅影响模型的正常运行，还可能引发严重的安全威胁，如信息泄露、恶意攻击等。为了应对这些挑战，业界已经提出了多种解决方案和框架。例如，360公司提出了大模型的安全四原则：可靠、可信、向善、可控，旨在从顶层设计上全面保障大模型的安全。

中国部分安全大模型应用效果

来源：头豹研究院

- 各大网络安全厂商积极推动大模型技术在安全领域的应用，通过整合行业知识、技术和数据，提升安全防护能力，**产品形态主要包括安全智能体和聊天机器人**

360安全大模型结合安全大脑，在实际应用中提升告警降噪效果近10倍、事件研判与溯源调查时间缩短近100%、事件自动化处置率提升900%、运营人效增长300%。

在实测中，深信服安全大模型检测了3万高对抗钓鱼邮件与100万白邮件。检出率从15.7%飞升至91.4%，误报率从0.15%降低至0.046%，超越传统终端安全和邮件网关产品。

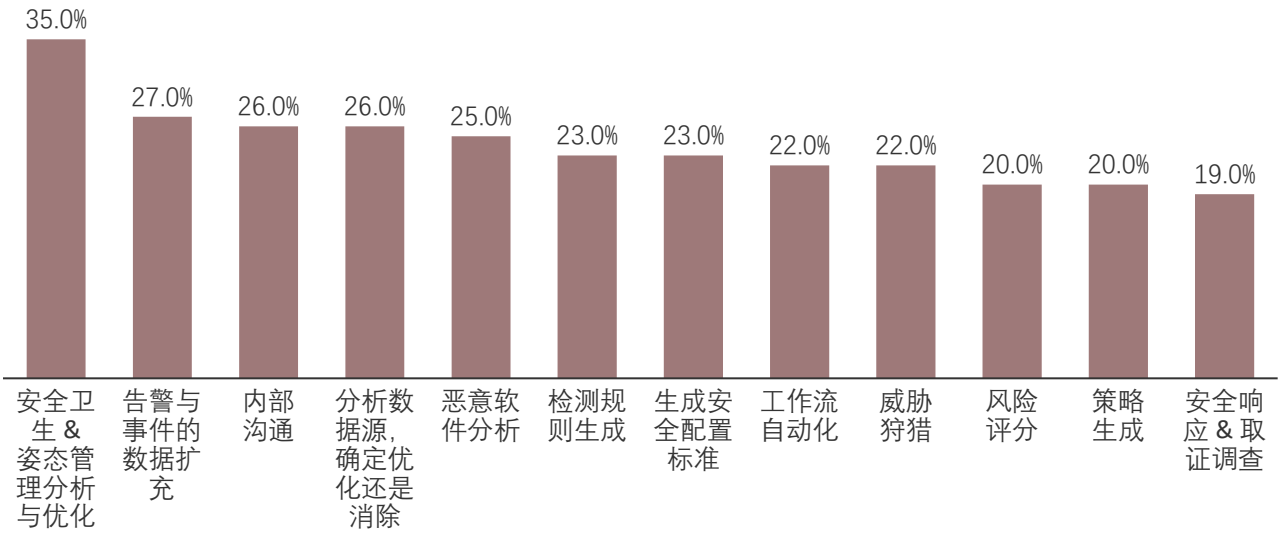
奇安信Q-GPT安全机器人80%以上告警自动化研判；腾出90%以上安全人员时间；单一威胁事件的平均处理时间减少98%。

微步在线的“情报智脑XGPT”能够实时关联超过100个数据源和8大分析引擎，提供精准的知识问答与威胁分析功能；专为应对新型网络威胁而设计；全面开放至微步x安全情报社区。

中国安全大模型行业综述——市场背景

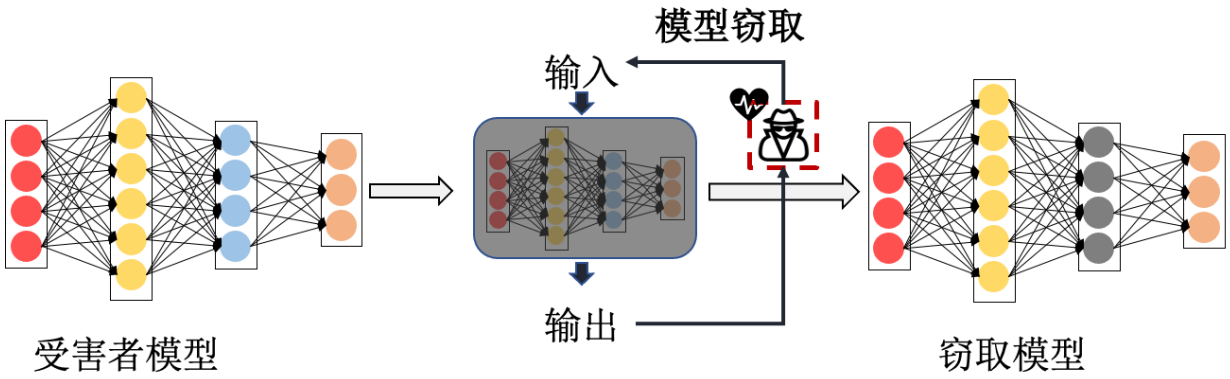
- AI在网络安全中作用显著，但应用不当也带来风险。需加强AI在保护系统免受攻击方面的应用，同时警惕其在网络攻击中的增强作用，并重视大模型带来的数据泄露和劫持风险

AI在网络安全中的应用



- AI在网络安全中的应用非常广泛，涵盖了多个方面。其中，“安全卫生&姿态管理分析与优化”应用占据了最高的比例，达到了35.0%，这表明在网络安全中，AI在保护系统免受攻击和入侵方面发挥着重要作用。AI通过实时分析网络流量、系统日志等数据源，能够及时发现潜在的安全威胁，评估网络的安全状态，并根据分析结果调整和优化安全策略。此外，AI还在“告警与事件的数据扩充”（27.0%）、“内部沟通”（26.0%）、“分析数据源，确定优化还是消除”（26.0%）和“恶意软件分析”（25.0%）中应用比较深入，这些应用表明AI在网络安全中的实时监控、内部威胁检测、数据分析以及恶意软件识别等方面也扮演着重要角色。

AI在网络攻击中的应用（模型窃取攻击）



来源：专家访谈、头豹研究院

（接上页——市场背景）

- 人工智能在网络攻击中的应用同样广泛。AI技术不仅能够增强网络攻击的能力，还能通过自动化和规模化的方式显著提高攻击的成功率和效率。

例如模型窃取攻击。模型盗用攻击旨在通过特定策略，非法获取一个与{目标模型}在功能及性能上高度相似的复制模型，以此规避高昂的模型训练成本并谋取不当利益。此类攻击构成对人工智能模型知识产权的严重侵犯。模型盗用过程可隐蔽地在对目标模型进行交互时实施，攻击者利用有限的黑盒访问权限，反复调用目标模型的API接口，通过精心设计的多样化查询样本输入，细致观察并分析模型输出的响应变化。在此过程中，攻击者不断迭代优化其查询策略，以更精准地探索并揭示目标模型的决策边界细节。基于这些交互数据，攻击者能够运用模仿学习技术，构建出一个与目标模型行为高度一致的盗用模型，或更直接地，通过技术手段尝试提取目标模型的内部参数、超参数等敏感信息，从而实现对目标模型知识产权的非法侵占。

来源：专家访谈、头豹研究院

中国安全大模型行业综述——发展痛点

- 大模型发展面临数据投毒、隐私泄露、记忆风险、内容合规性、透明度不足及数据传输安全等多重挑战，需全面加强安全防护和监管，确保技术健康稳定发展

中国安全大模型发展痛点

外部攻击安全问题
包括数据投毒、模型后门、对抗样本、数据泄露、模型窃取、软件漏洞等安全隐患。

模型记忆风险
经过模型的训练和推理后，模型可能会形成对特定输入的“记忆”，这可能导致模型输出偏见或错误。

新兴威胁场景
随着AI技术的发展，新的威胁场景不断出现，如智能生成虚假新闻等，这些都需要在安全策略中加以考虑。

模型流转/部署过程中的安全问题
在大模型的非私有化预训练、精调和推理服务过程中，数据需要在不同的主体或部门之间传输，这增加了传输截获的风险。

数据安全与隐私
大模型在训练和部署过程中需要处理大量敏感数据，如何保护这些数据不被未授权访问或泄露是一个重大挑战。

内容合规
AIGC的内容合规性也是一个重要问题，需要确保生成的内容符合法律法规和社会伦理标准。

算法透明度和可解释性
由于AI模型通常是“黑箱”式的，其决策过程缺乏透明度，这可能导致用户对AI系统的信任度下降。

- **安全大模型发展面临一系列复杂而严峻的痛点，这些问题不仅关乎技术本身，更涉及数据安全、隐私保护、算法透明度等多个维度**

首先，数据投毒、模型后门等攻击手段层出不穷，对模型的安全性和稳定性构成了严重威胁。这些攻击手段往往能够巧妙地绕过模型的防御机制，导致模型在不知不觉中受损，进而产生错误或偏见的输出。

其次，大模型在训练和部署过程中需要处理大量敏感数据，如何保护这些数据不被未授权访问或泄露是一个重大挑战。数据泄露、模型窃取、软件漏洞等安全隐患时刻威胁着大模型的稳定运行和数据的安全。

此外，模型记忆风险也是大模型发展中的一个重要问题。经过长时间的训练和推理，模型可能会形成对特定输入的“记忆”，这种记忆可能导致模型在处理相似输入时产生偏见或错误的输出。

同时，随着AI技术的发展，智能生成的内容越来越丰富和多样，但同时也带来了内容合规性的挑战。如何确保生成的内容符合法律法规和社会伦理标准，是当前大模型发展中亟待解决的问题之一。

此外，随着AI技术的发展，新的威胁场景不断出现，对大模型的安全性和可信度提出了更高的要求。同时，由于AI模型通常是“黑箱”式的，其决策过程缺乏透明度，这也可能导致用户对AI系统的信任度下降。

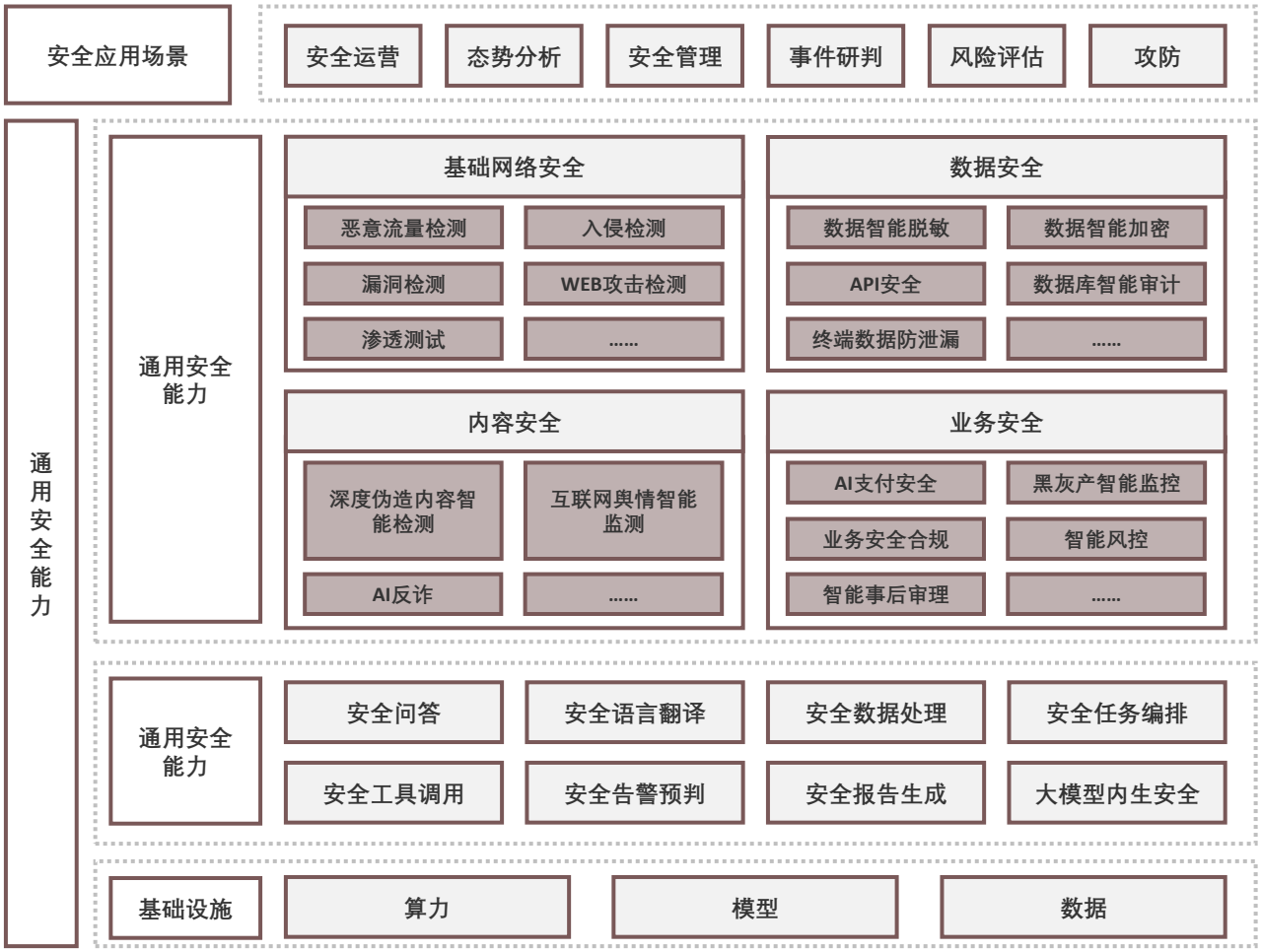
最后，由于大模型的非私有化预训练、精调和推理服务过程中，数据需要在不同部门之间传输，这增加了传输截获的风险。

来源：专家访谈、头豹研究院

中国安全大模型行业综述——架构体系

- 安全大模型平台架构包括基础设施、能力支撑、安全服务和服务与运营四层。基础设施涵盖底层设备，能力支撑提供关键技术支持，安全服务层核心保护数据和网络，服务与运营层关注整体平台运作与管理

安全大模型架构



- 安全大模型的架构设计旨在应对各种复杂的安全应用场景需求，其体系结构分为基础设施层、通用安全能力和安全原子能力三个层次。基础设施层是整个架构的基础，它提供了大模型运行所需的算力支持、模型框架以及数据资源。在此之上，通用安全能力层集成了多种功能，例如安全问答、安全语言翻译、安全数据处理、安全任务编排等功能模块，还包括了安全工具的调用机制、安全告警的研判分析以及安全报告的自动化生成，同时还涵盖了大模型自身的安全防护能力。最上层的安全原子能力层则进一步细化，包含了基础的网络安全能力、数据安全能力、内容安全能力和业务安全能力等基本要素，这些能力可以通过不同的组合方式来满足特定的安全应用场景需求。安全大模型的应用场景覆盖了安全运营、态势分析、安全管理、事件研判、安全技术防护、风险评估和攻防演练等多个方面。通过这三个层次的协同工作，安全大模型能够在保障数据隐私和网络安全的同时，有效地提升安全防护水平，为企业和机构提供全面的安全保障方案。

来源：新华三、头豹研究院

中国安全大模型行业综述——发展历程

- 中国安全大模型行业历经小模型到大模型的演进，大模型技术以其强大的学习和推理能力，实现了对安全风险的快速预测与检测，并推动定制化安全解决方案的发展，构建了全方位的安全防护体系

首席分析师：oliver.yuan@leadleo.com

主笔分析师：ruowei.lin@leadleo.com

■ **中国安全大模型行业的发展可以划分为两个阶段：AI1.0/小模型阶段与AI2.0/大模型阶段**

在AI1.0/小模型阶段，行业主要聚焦于行为分析、小模型以及机器学习等核心技术的研发与应用。这些技术虽然具有一定的检测和分析能力，但受限于技术本身的局限性，它们主要被用于单一的检测场景。这意味着在面对复杂多变的网络安全威胁时，小模型的效能可能无法完全满足实际需求。尽管如此，这一阶段的探索为后续的技术升级奠定了坚实的基础。

随着技术的不断进步，AI2.0/大模型阶段应运而生。在这一阶段，生成式AI和大模型技术成为了行业的新宠。大模型凭借其强大的学习和推理能力，能够迅速适应并应对未知威胁，实现了对安全风险的快速预测和检测。此外，大模型还具备了协助检测和分析的能力，使得安全人员在处理安全事件时能够事半功倍，达到更为高效的效果。同时大模型具有场景化安全应用的能力，能够根据不同行业和场景的需求，提供定制化的安全解决方案，从而进一步提升安全防护的针对性和有效性。

在发展的过程中，行业还涌现出了一系列重要安全产品，如态势感知、NDR、TIP、安全大脑、SIEM、安全Copilot以及XDR等。这些产品各自在安全防护的不同环节发挥着重要作用，共同构建了全方位、多层次的安全防护体系。

来源：专家访谈、头豹研究院

Chapter 2
产业链分析

- 上游分析：中国算力全球领先，AI算力支出居首；数据资源丰富但安全大模型需高质量数据，构建流程需多源采集、精细筛选与持续清洗，以提升国际竞争力
- 技术分析：安全大模型通过预训练-微调范式提升防护力，RLHF技术优化AI性能，智能体基于LLM构建，强调协同与适应。这些技术共同推动AI在安全与复杂任务中的高效应用
- 产品分析：AI大模型在网络安全中的应用已提升自动化水平，目前处于工业应用层级，未来有望全面替代传统产品，实现全面网络防御功能，成为网络安全领域的核心力量
- 大模型应用现状：基础大模型在安全领域应用受限，因缺乏专业知识与数据支持。需通过增量预训练注入行业知识，避免“幻觉”现象，确保模型在实际应用中有效
- 应用场景：安全大模型通过自动化处理、多源数据整合与深度学习，显著提升威胁分析效率与防御策略精准度，具备持续学习能力，为组织提供长期安全保障，优化安全运营，增强响应速度并提升用户体验

中国安全大模型产业链分析——产业链图谱

- 中国安全大模型行业的产业链上游包括硬件、软件和技术供应商，中游则是安全大模型行业产品和解决方案提供商，下游为安全大模型行业的应用领域和终端用户

中国安全大模型行业产业链图谱

上游

硬件提供商

DELL inspur 浪潮 SMIC NAURA 北方华创 兆芯 WILLSEMI 紫光国微 GUOXIN MICRO Rockchip

软件提供商

KYLINSOFT 麒麟软件 KING BASE GBASE 东方通 BES 达梦数据库 HUAWEI 中科方德 海量数据 VASTDATA 神舟通用 统信软件 UNIONTECH

技术提供商

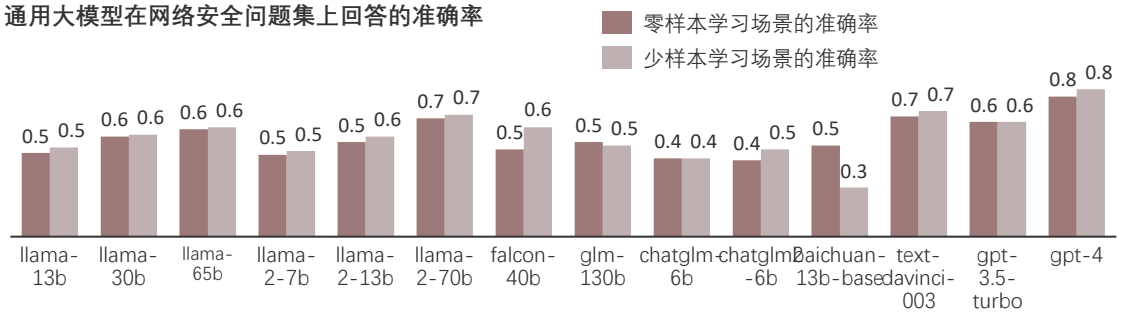
卡亭科技 CHAITIN NSFOCUS 安恒信息 DAS-security 腾讯安全 Tencent Security 天融信 TOPSEC Hillstone 山石网科 知道创宇

中游

微步在线 NSFOCUS SANGFOR 深信服科技 中国移动 China Mobile AsiaInfo 亚信安全 腾讯安全 Tencent Security 奇安信 安恒信息 DAS-security Venus 启明星辰 卡亭科技 CHAITIN Hillstone 山石网科 天融信 TOPSEC 知道创宇 金睛云华 360 美亚柏科 MEIYA PICO 吉大正元 永信至诚

- AI大模型在网络安全中的应用已提升自动化水平，目前处于工业应用层级，未来有望全面替代传统产品，实现全面网络防御功能，成为网络安全领域的核心力量
- 安全大模型在网络安全领域潜力巨大,通过提升处理效率与智能化水平，为复杂安全问题提供新解决方案。展望未来，安全大模型将发挥更重要作用，为网络安全保驾护航

下游



- 因缺乏专业知识与数据支持，基础大模型在安全领域应用受限。目前，通用大模型在解答安全相关问题时，其准确率普遍偏低，往往不足50%，对于中文问题的解答准确率更是进一步下降

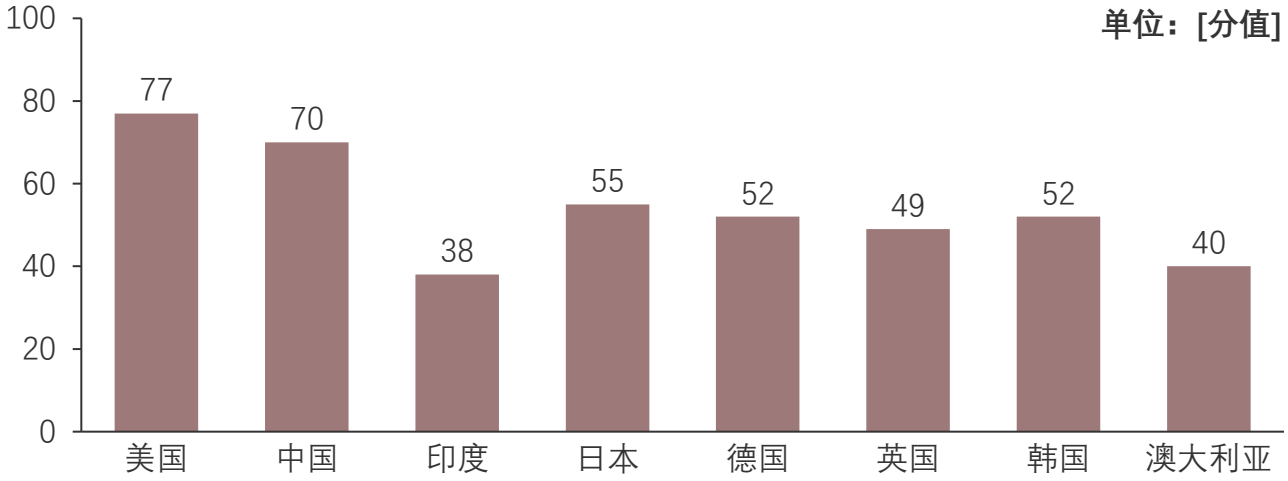
注：图谱中所展示logo顺序及大小无实际意义，不涉及排名，仅展示部分行业代表性企业

来源：专家访谈、头豹研究院

中国安全大模型产业链上游分析——算力与算据

- 中国算力全球领先，AI算力支出居首；数据资源丰富但安全大模型需高质量数据，构建流程需多源采集、精细筛选与持续清洗，以提升国际竞争力

世界范围算力现状概览

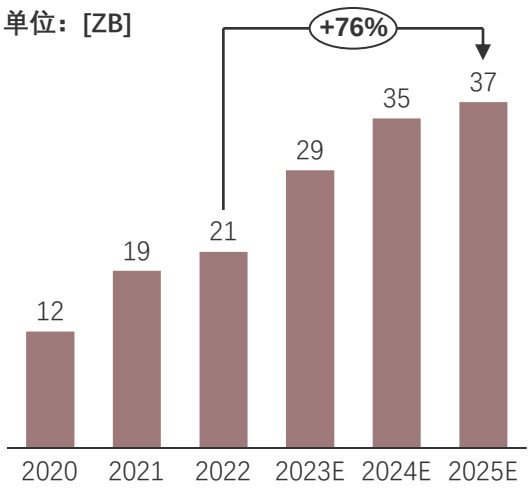


中国计算力全球第二，增速全球第一，AI算力支出首超美国居首

清华大学最新发布的全球计算力指数评估报告显示，中国在2022年的计算力总量已跃居世界第二位，仅次于美国，且两者差距趋于缩小。这一成就彰显了中国在算力基础设施建设上的强劲动力与显著成效。近年来，中国持续加大算力投资，数据中心规模实现年度飞跃，同比增长超过30%。值得一提的是，中国总算力年增速高达11.5%，位居全球首位。具体在AI算力方面，中国在AI算力发展方面表现尤为突出，其服务器支出规模同比大幅增长44.5%，首次超过美国位列全球第一。

安全大模型数据来源分析

中国数据总量



安全领域数据的数据来源

数据源	数据量级	数据质量
Common Crawl	TB	低
书籍	TB	高
安全站点	GB	中
Arxiv	GB	高
百科	GB	高
开源数据	TB	中

来源：专家访谈、头豹研究院

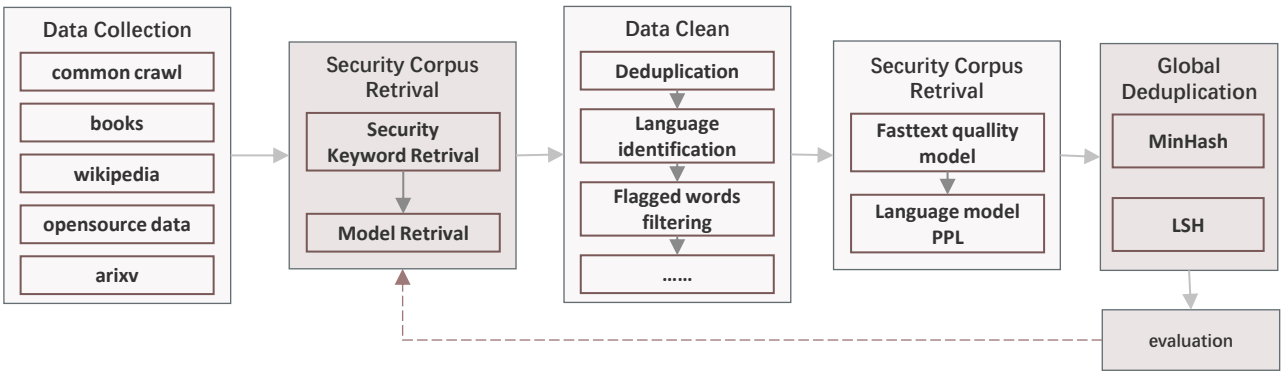
（接上页——算力与算据）

■ **中国数据增长迅猛，但安全大模型训练需高质量数据集。来源需精选，包括书籍、安全站点、Arxiv、百科及开源社区数据，需严格筛选与清洗，以提升国际竞争力**

中国的数据量近年来增长迅猛，数据中心市场规模从2018年的592.3亿元增长至2022年的2,032.5亿元，数据总量则是以30.3%的年复合增长率高速增长至2022年的20.7ZB。此外，随着“东数西算”工程的推进，预计到2025年，中国数据中心机架规模将达到1400万架，总增量投资约达2万亿元人民币。

尽管中国的数据资源丰富且增长迅速，但用于安全大模型训练的数据集对质量要求较高，这很大程度决定了其与国际先进模型的竞争力。具体从安全大模型训练的数据来源来看，Common Crawl等网页资源，虽蕴藏丰富的文本数据，但普遍质量参差，亟需执行详尽的数据清洗步骤及严格的领域针对性筛选，具体流程可参照既定网页数据清洗规范。相比之下，书籍数据以其较高的质量脱颖而出，需重点聚焦于安全主题内容的精准筛选与后续清洗。安全领域垂直站点，如Hacknews，其数据自然具备高度的领域契合度，可直接纳入考量。同时，Arxiv与百科类资源，凭借卓越的数据质量，亦需细致筛选其中与安全领域紧密相关的部分。最后，开源社区作为另一重要数据来源，其数据亦需经过安全领域的精心筛选，以构建全面且专业的安全数据体系。

安全领域数据构建流程



■ **安全领域数据构建流程具体如下：**

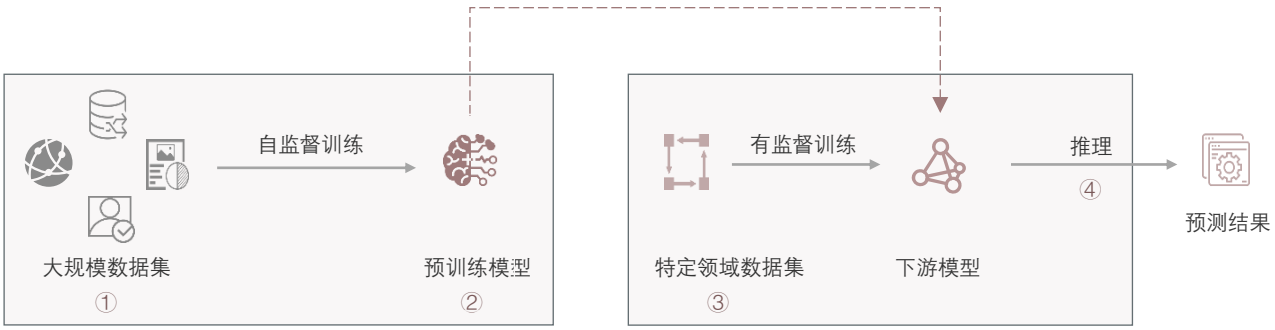
首先，广泛采集多源数据，随后利用安全文本召回技术提取关键信息。接着，实施数据清洗，包括去重处理、语言识别（专注于中英文）及针对不同格式的脏数据过滤。之后，通过自定义规则过滤与高质量文本筛选，确保数据质量。最终，全局去重合并后，数据进入训练评估阶段，并在此过程中实施递归清洗，持续优化，为安全领域大模型的训练提供坚实的数据基础。

来源：专家访谈、头豹研究院

中国安全大模型产业链上游分析——技术分析

- 安全大模型通过预训练-微调范式提升防护力，RLHF技术优化AI性能，智能体基于LLM构建，强调协同与适应。这些技术共同推动AI在安全与复杂任务中的高效应用

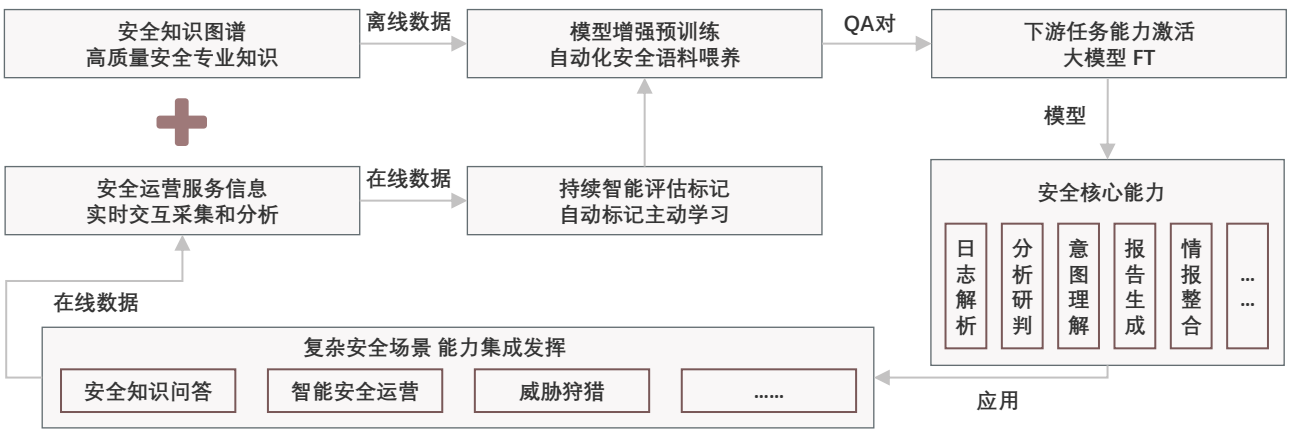
安全大模型预训练-微调的训练范式



■ **安全大模型采用预训练-微调范式，虽复杂但非安全威胁唯一因素。同质化、多模态挑战增加威胁，但大模型参数多、需敏感数据少，也增强防护力，保障广泛应用**

安全大模型在训练策略上与传统端到端模型有着较大的差异。安全大模型采用了先进的预训练-微调范式，这一过程首先依赖于海量未标注数据进行广泛而深入的预训练，使模型能够捕获到数据的普遍规律与特征。随后，在特定下游任务的标注数据上进行精细微调，从而生成适用于该领域的垂直模型。尽管复杂的训练过程与模型结构往往与更高的安全风险系数相关联，但安全大模型所面临的安全威胁并非仅由此简单决定。值得注意的是，同质化现象及多模态数据对齐的挑战可能使大模型暴露于更多元化的安全威胁之下。然而，大模型独有的优势，如其庞大的参数规模及在微调阶段对敏感数据需求的减少，也在一定程度上构筑了抵御对抗样本攻击与数据隐私泄露的防线，为安全大模型的广泛应用提供了更为坚实的保障。

安全行业大模型的关键技术-数据增强



来源：专家访谈、头豹研究院

（接上页——技术分析）

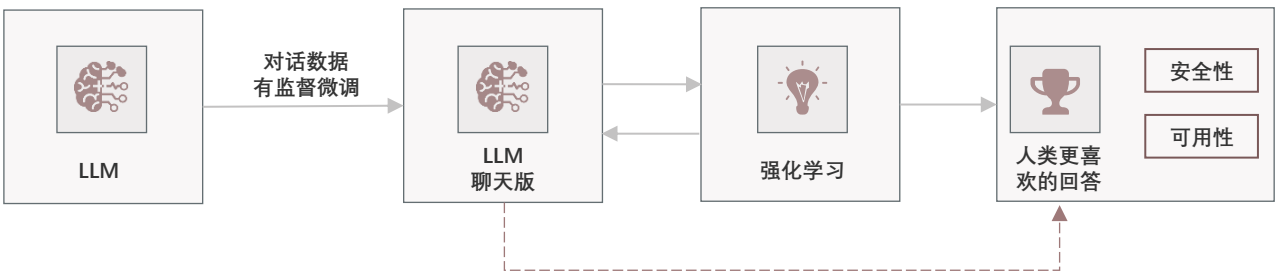
- **安全大模型因缺乏安全专业知识受限，通过构建高质量安全语料库并应用双数据飞轮机制，持续学习优化，以提升在安全领域的实际应用效果与适应性**

通用大模型LLM在安全领域的应用面临的主要障碍是安全专业知识的缺失。这涵盖了从安全术语、复杂的攻防策略到应急响应方案、漏洞利用细节及攻击特征等多个维度。特别是在安全事件分析中，对网络攻击流量的精准解析与海量告警日志的有效降噪，均依赖于深厚的安全知识积累，包括漏洞知识库、攻击技术识别及威胁情报数据等。然而，当前多数LLM因基于通用语料库预训练，难以直接提供精准的安全领域知识，从而限制了其在安全行业的实际应用效果。

为克服这一局限，一些安全大模型在构建过程中特别强调了安全语料库的重要性。它们采用双数据飞轮自运转机制，通过离线与在线两条路径持续收集并处理安全数据，以构建并优化高质量的安全语料库。离线方面，这些模型利用历史积累的安全知识图谱，整合多源异构的网络安全数据，提供坚实的安全专业知识基础。在线方面，则依托实时安全服务反馈的数据，这些数据经过安全专家的严格审核与处理，确保了数据的可靠性与置信度。

通过这一机制，安全大模型能够不断吸纳新的安全语料，并通过增量训练进行模型优化。同时，它们还能从复杂安全场景的反馈中持续学习，逐步提升处理安全问题的能力与效果。双数据飞轮的持续运转，不仅丰富了安全大模型的知识储备，还增强了其适应变化需求与新应用场景的能力，从而推动了安全行业大型模型性能与效果的全面提升。

安全行业大模型的关键技术-人类反馈强化学习



- **RLHF技术融合人类反馈与强化学习，优化AI模型性能，提升安全性与可用性**

人类反馈强化学习（RLHF）技术作为一种前沿方法，将强化学习框架与人类专家的直接反馈深度融合，旨在优化人工智能模型的性能表现。该技术的精髓在于利用人类的智慧与经验，构建一个更为强大且智能的学习机制，通过精心设计的奖励信号来引导模型的学习路径。

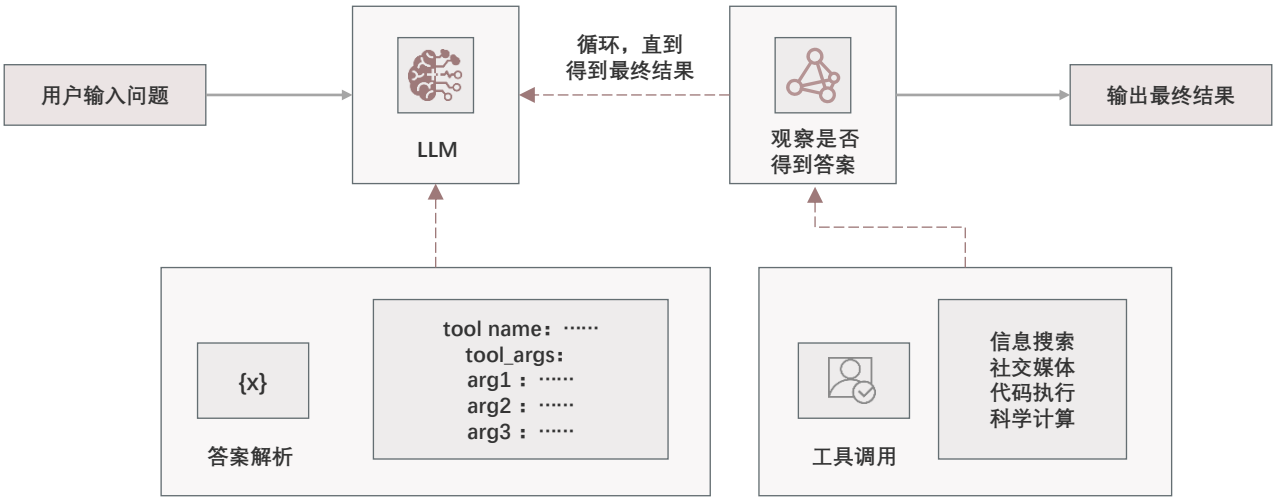
在实际应用中，RLHF技术对于提升大型AI模型的安全性与可用性具有显著作用。安全性方面，该技术确保模型在面对非法或不适宜的查询时，能够自主判断并拒绝回答，从而保护用户权益与信息安全。而可用性则体现在模型能够生成准确、有用且符合用户需求的回答，提升用户体验。以ChatGPT等模型为例，RLHF技术的应用极大地增强了人机互动的自然流畅度与有效性。

然而，安全性和可用性之间往往存在微妙的平衡关系。过度强调其中一方面可能会导致另一方面的性能受损。因此，持续进行强化学习实验，不断调整优化策略，以寻求两者之间的最佳平衡点，是RLHF技术发展的关键所在。此外，RLHF技术还面临诸多挑战与限制，如高效收集与处理人类反馈的难题、设计稳定且广泛适用的奖励模型等。针对这些问题，未来的研究将聚焦于提升反馈收集效率、增强奖励模型的鲁棒性与通用性等方面，以推动RLHF技术向更加成熟、高效的方向发展。

来源：专家访谈、头豹研究院

（接上页——技术分析）

安全行业大模型的关键技术-智能体



- **智能体基于LLM构建，通过分解问题、利用工具、记忆与规划实现目标，具备环境感知与自主理解力，强调组件协同与动态适应**

智能体（Agent）是一种基于大型语言模型（LLM）构建的复杂系统，其设计目的是通过逐步思考和规划来实现特定的目标任务。该系统的运作原理在于将复杂问题分解成多个可管理的子模块，并利用大模型生成的结果来解析出调用工具的具体指令，进而使用多种工具逐步完成目标。

智能体通常包含四个主要组成部分：大模型、工具、记忆模块以及规划模块。大模型负责提供基本的推理和理解能力；工具用于执行具体的任务操作；记忆模块则存储与环境交互的历史记录，帮助智能体积累经验和做出更明智的决策；规划模块则承担着制定合理行动计划的角色，确保智能体能够有效地朝着目标前进。此外，智能体具备环境感知和自主理解能力，使其能在复杂多变的环境中灵活应对各种情况。例如，在处理需要长时间序列上下文任务时，智能体必须能够维持对上下文的理解，确保在规划和调用工具的过程中，始终保持与用户目标的一致性。这一特性对于确保智能体在执行任务时的连贯性和准确性至关重要。

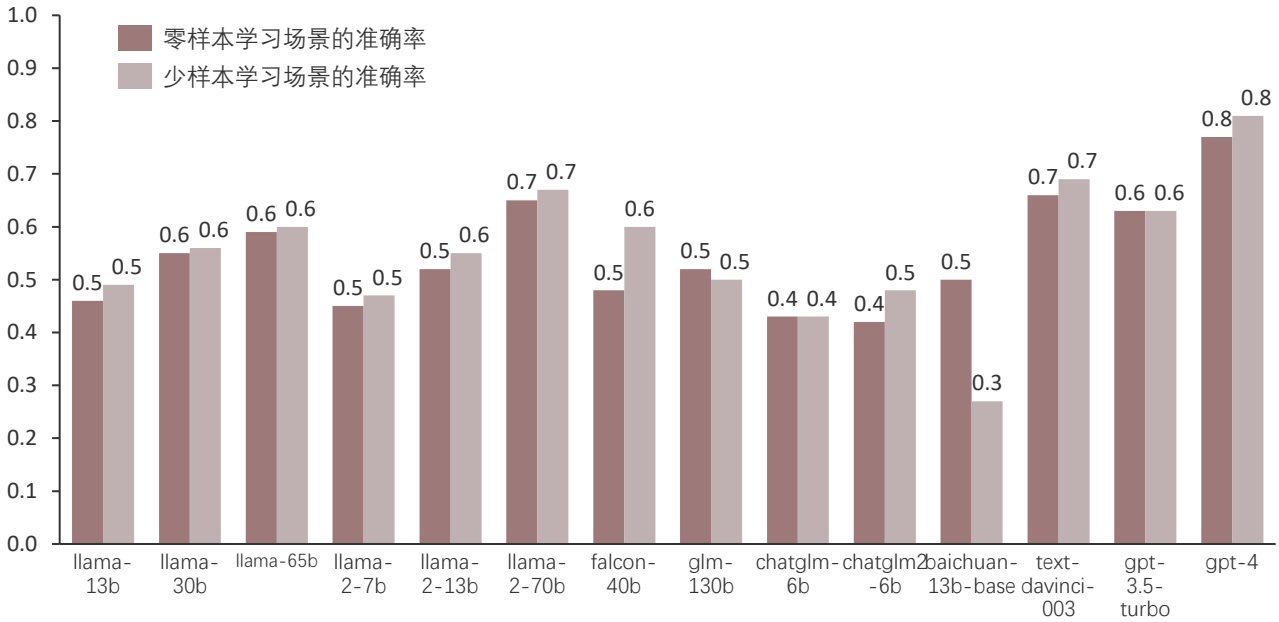
智能体的设计不仅强调了各组件之间的协同工作，还重视在整个任务执行过程中保持对环境的动态适应。通过不断的学习与调整，智能体可以在处理复杂任务时展现出更高的灵活性和适应性，从而更好地服务于特定的应用场景。

来源：专家访谈、头豹研究院

中国安全大模型产业链下游分析——大模型应用现状

- 基础大模型在安全领域应用受限，因缺乏专业知识与数据支持。需通过增量预训练注入行业知识，避免“幻觉”现象，确保模型在实际应用中有效

通用大模型在网络安全问题集上回答的准确率



■ **通用大模型安全应用受限，缺专业知识与数据。需增量预训练，丰富知识库，提升应用效果**

基础大模型在安全领域的实际应用中，其在性能上受到缩放定律的限制，这意味着尽管基础大模型在理论上具备强大的处理能力，但在实际操作中往往难以达到预期的流畅与高效。核心问题在于，大模型通常基于广泛但非特定的通用语料库进行训练，而安全行业的数据则具有高度的专业性和稀缺性。这一差异导致大模型在安全领域内缺乏必要的专业知识和数据支持，进而在理解和处理特定安全问题上显得力不从心，难以给出准确或深入的解答。

目前，通用大模型在解答安全相关问题时，其准确率普遍偏低，往往不足50%，对于中文问题的解答准确率更是进一步下降。这一现象凸显了直接将通用大模型应用于安全领域的局限性。为了克服这一障碍，安全行业开始探索通过增量预训练（CPT）等技术手段，向大模型中注入安全行业的特定知识和数据。这一过程被视为提升大模型在安全领域内应用效果的关键步骤。值得注意的是，如果未能经过充分的增量预训练，而直接对大模型进行针对安全任务的监督微调（Supervised Fine-tuning，SFT），很可能会因为模型内部知识结构的不足，导致其在面对未知或复杂安全问题时产生“幻觉”现象，即给出看似合理实则错误或无关紧要的回答。因此，对于希望将大模型应用于安全行业的组织而言，实施有效的增量预训练策略，以丰富模型的安全知识库，是确保模型在实际应用中能够发挥应有价值的重要前提。

来源：专家访谈、头豹研究院

中国安全大模型产业链下游分析——应用场景

- 安全大模型通过自动化处理、多源数据整合与深度学习，显著提升威胁分析效率与防御策略精准度，具备持续学习能力，为组织提供长期安全保障，优化安全运营，增强响应速度并提升用户体验

安全大模型的应用

智能威胁分析

多源数据整合 | 非结构化数据处理 | 深度学习与推理
威胁情报自动化 | 生成可读报告 | 增强防御策略

- **自动化处理：**能够自动化处理复杂的威胁分析过程，显著提高了分析效率。相比传统方法，大模型能够同时处理大量数据，快速识别潜在威胁，减少人工干预和错误。
- **整合数据和关联分析：**能够整合来自不同源的数据，包括威胁情报、日志、网络流量等，通过关联分析构建出完整的攻击链。
- **持续学习：**具备持续学习的能力，能够不断吸收新的威胁知识和数据，保持与最新攻击技术的同步。这意味着模型能够随着时间的推移变得更加智能和准确，为组织提供长期的安全保障。
- **优化防御策略：**基于智能威胁分析的结果，安全团队可以更加精准地调整防御策略。大模型能够预测潜在的威胁趋势和攻击模式，帮助组织提前应对风险，降低安全事件发生的概率和影响。

智能安全运营助手

自动化值守 | 智能辅助 | 多模态数据融合
上下文理解 | 高效研判与处置 | 持续学习与优化

- **提升运营效率：**通过大模型的复杂推理和语义理解能力，助手能够迅速识别潜在威胁，减少人工干预和重复劳动，使安全团队能够更专注于高价值的任务。
- **增强安全响应速度：**利用大模型的实时分析能力，智能安全运营助手能够在极短的时间内对安全事件进行研判和处置，大幅缩短了事件的响应时间。
- **提升用户体验：**智能安全运营助手通过自然语言交互的方式，为安全人员提供了更加便捷、高效的工作界面。安全人员可以通过简单的对话和指令，获取所需的安全信息和支持，从而提高了工作效率和满意度。

高级恶意软件分析

样本预处理 | 特征转换 | 大模型分析
变种检测 | 行为建模 | 报告生成

- **处理复杂行为：**对于具有复杂行为模式的恶意软件，大模型能够深入解析其行为特征，揭示其潜在的恶意目的。
- **变种识别：**大模型具有强大的泛化能力，能够识别出恶意软件的变种和变种间的相似性，即使面对未知的恶意软件也能进行有效分析。
- **适应新型威胁：**随着恶意软件技术的不断发展，大模型能够持续学习新的威胁知识，保持与最新攻击技术的同步。

来源：专家访谈、头豹研究院

（接上页——应用场景）

安全大模型的应用

智能安全策略管理

策略文档解析 | 合规要求分析 | 差距分析
策略生成与更新 | 人机协作 | 策略执行与评估

- **智能化分析：**通过NLP技术和深度学习算法，大模型能够智能地识别合规要求、分析策略差距，并生成符合组织需求的安全策略，使策略管理更加智能化。
- **实时更新：**大模型能够实时跟踪和分析最新的合规要求，确保安全策略与最新的法律法规、行业标准保持一致，降低合规风险。
- **精准定位：**通过差距分析，大模型能够精准定位现有策略中的不足和需要改进的地方，为策略优化提供明确的方向。

高级威胁狩猎

智能分析 | 威胁情报融合 | 行为模式识别
自动化狩猎 | 持续监控与响应 | 证据链生成

- **深度学习能力：**安全大模型通过深度学习技术，能够分析海量的安全数据，从中发现传统方法难以察觉的高级威胁。这种深度分析能力大大提高了威胁检测的准确性。
- **持续监控：**安全大模型具备持续监控网络环境的能力，能够实时分析安全数据，及时发现并响应高级威胁。这种实时响应能力有助于组织在威胁发生初期就采取措施，防止事态扩大。
- **资源优化：**通过自动化和智能化的威胁狩猎流程，大模型能够减少人工干预和降低对专业人员的依赖程度，从而优化安全资源的配置。

高级欺诈检测与防护

实时数据分析 | 捕捉隐蔽欺诈行为 | 行为分析
设备指纹识别 | 社交网络分析 | 智能决策与响应

- **复杂模式识别：**通过深度学习技术，能够识别出复杂的欺诈模式，这些模式往往隐藏在大量的交易数据中，传统方法难以发现。大模型能够学习到这些模式的特征，并在实时检测中快速识别出潜在的欺诈行为。
- **减少误报与漏报：**相比传统方法，大模型在欺诈检测中能够显著降低误报率和漏报率。误报率的降低意味着减少了正常交易的干扰，提高了用户体验；而漏报率的降低则意味着能够更有效地防止欺诈行为的发生，减少组织的经济损失。
- **个性化服务：**通过对用户行为的分析，大模型还可以为用户提供个性化的安全服务。例如，根据用户的交易习惯和风险等级，提供定制化的验证方式和安全建议。

来源：专家访谈、头豹研究院

业务合作

会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

行研训练营

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

报告作者

袁栩聪
首席分析师
oliver.yuan@leadleo.com

林若薇
行业分析师
ruowei.lin@leadleo.com

业务咨询

- **客服电话：400-072-5588**
- **官方网站：www.leadleo.com**

深圳办公室

广东省深圳市南山区粤海街道华润置地大厦E座4105室

邮编：518057

上海办公室

上海市静安区南京西1717号会德丰国际广场 2701室

邮编：200040

南京办公室

江苏省南京市栖霞区经济开发区兴智科技园B栋401

邮编：210046

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明