TRUSTED SECURITY 可信安全

云计算开源产业联盟
OpenSource Cloud Alliance for IndustRy,OSCAR

云上人工智能安全发展研究报告

（2025年）

报告愿景及目标读者

在数字化浪潮推动下，云计算与人工智能成为各行业变革的核心动力。云计算以强大算力和灵活资源为人工智能发展提供坚实基础，人工智能则赋能云服务实现智能化，两者深度融合，引领智能云时代。

随着云计算技术成熟及市场快速扩张，全球云服务提供商持续扩大数据中心规模，优化架构以满足数字化转型需求。与此同时，人工智能在自然语言处理、计算机视觉、机器学习等领域取得突破，广泛渗透金融、医疗、制造、交通等行业，创造巨大商业价值。

然而，云上人工智能的普及也带来了显著安全挑战。数据在存储、传输和处理过程中面临泄露、篡改和滥用风险；模型训练和应用遭受窃取、中毒及对抗攻击等威胁；云基础设施的漏洞及多租户环境中的隔离不足也增加系统风险。这些安全隐患不仅威胁企业数据和业务连续性，还可能对个人隐私、公共安全甚至国家安全造成严重影响。

本报告从云上人工智能安全概述、云上大模型安全风险分析与防护体系建设、云上人工智能安全发展趋势展望等方面进行深入解读和细致分析，全面覆盖云上人工智能的安全现状与挑战。通过剖析大模型在云环境中的安全风险及防护实践，探讨多方协同与技术创新驱动下的安全治理新模式，为云上人工智能安全建设提供系统性解决方案与发展参考。

主要撰稿人

马铭洋、郭雪、卫斌、李忠权、贾金鹏、喻洪莲、王浩硕、宋志明、王锴、周飞、何亮忠、常三强、李军华、江为强、郭中元、程艳、张然、姚杨、曾祥楷、饶飞、徐正军、王龑、唐佳伟、王晨迪、吴剑刚、张暴、肖鹏、靳莉、王睿超、梁雷、马乐乐、岑黎光、孙侠、荆昆仑、张建伟、梅瑞、郑淼、蔚永强、李志伟、王新洋、刘冰、毛昱、蔡义祥、胡毅勋、陈希宁、潘文伦、柳晶、冯子祥、张俊强、杨柳、栗伟清、刘涛、金士英、夏营、廖文、徐钟豪、刘宇馨、刘岩

致　谢

本报告在撰写过程中得到了以下单位和个人的支持和帮助，在此表示感谢。由于撰稿时间有限，行业态势变化快，如有疏忽和纰漏，欢迎批评指正。

中国信息通信研究院、阿里云计算有限公司、中国移动通信集团、天翼云科技有限公司、中移（苏州）软件技术有限公司、中国平安人寿保险股份有限公司、科大讯飞股份有限公司、华为云计算技术有限公司、腾讯云计算（北京）有限责任公司、浪潮云信息技术股份公司、蚂蚁科技集团股份有限公司、奇安信科技集团股份有限公司、启明星辰信息技术集团股份有限公司、北京天融信网络安全技术有限公司、瑞数信息技术（上海）有限公司、北京百度网讯科技有限公司、中兴通讯股份有限公司、北京知道创宇信息技术股份有限公司、中电云计算技术有限公司、北京兴云数科技术有限公司、北京金山云网络技术有限公司、北京海泰方圆科技股份有限公司、北京神州绿盟科技有限公司、北京云起无垠科技有限公司、上海斗象信息科技有限公司

（排名不分先后）

目 录

表目录

一、云上人工智能安全发展概述

（一）云上人工智能产业生态深度演进，服务模式多元化与安全挑战并存

在全球数字化转型的浪潮中，云计算与云上人工智能的深度融合已成为驱动产业变革的核心力量。云计算凭借弹性算力供给与资源高效调度，为云上人工智能的模型训练与推理提供坚实的基础设施支撑。而云上人工智能则通过智能算法优化，赋予云服务自主决策与智能交互能力，二者相互赋能，构建起云智算生态体系，正以前所未有的速度重塑各行业发展格局。与此同时，随着应用场景的不断拓展，安全问题逐渐凸显，成为制约产业健康发展的关键因素，构建全方位的安全防护体系迫在眉睫。

1．产业生态持续发展，市场驱动与技术赋能协同共进

一是云上人工智能产业规模持续增长，成为数字经济核心引擎。近年来，云上人工智能市场保持着持续高速增长，成为全球数字经济发展的重要引擎之一。根据 Verified Market Research 发布的《Global Cloud AI Market Size and Forecast》[1]报告显示，2024 年全球基于云计算的人工智能市场规模已达到 482.2 亿美元，预计在 2025 年至 2032 年期间将以 30.1%的复合年增长率快速扩张，到 2032 年市场规模将达到 3934.4 亿美元。这一增长主要受企业加快数字化转型、云基础

[1] 来源：Verified Market Research，PR Newswire，2025 年 4 月 11 日

设施广泛采用以及基于云的人工智能方案和工具兴起等因素驱动。**第一，**作为新基建与数字化转型的关键支撑，其被金融、制造、医疗等多行业广泛采用，推动智能技术与云计算深度融合，形成算力供给与产业赋能的闭环；**第二，**云上人工智能服务模式加速技术普惠，市场渗透从头部企业向中小企业延伸，成为数字经济增长极。

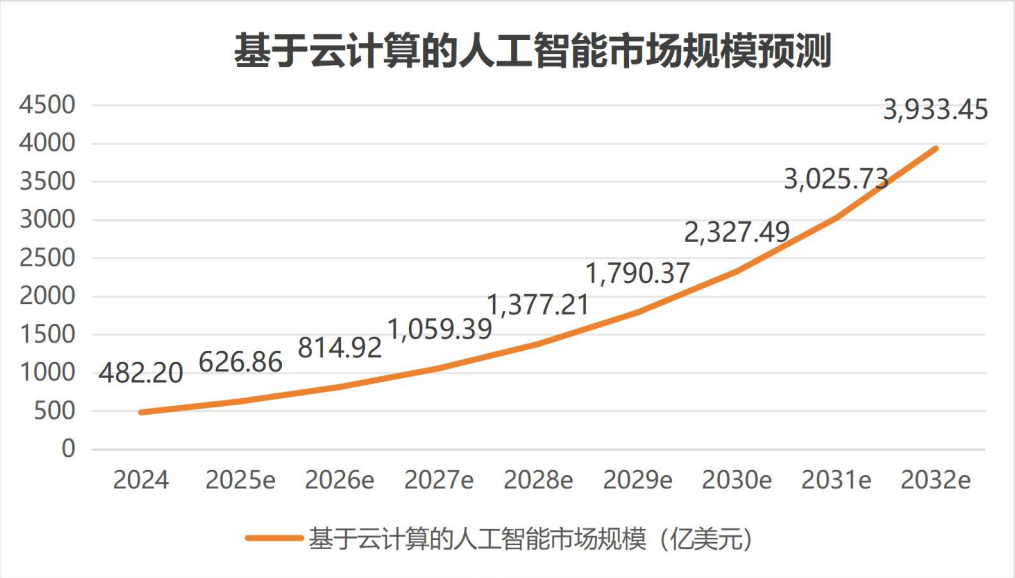

图 1 基于云计算的人工智能市场规模

数据来源：Verified Market Research,《Global Cloud AI Market Size and Forecast》

二是产业技术底座加速构建，云平台通过核心功能体系夯实技术协同基础。云平台作为云上人工智能产业的技术底座，通过整合算力、开发、服务三大核心能力，形成系统化的技术协同支撑体系。**其一，算力资源的集约化管理与弹性供给机制。**云平台通过分布式计算架构实现算力资源的池化管理，将异构算力与存储资源进行统一调度，既能够支撑大规模模型训练的算力需求，又能根据业务流量动态调整资源分配，解决算力供给与需求的匹配问题。**其二，全流程开发工具链的一体化集成与能力下沉。**云平台整合数据预处理、模型训练、部署推理等全环节开发工具，构建低代码、模块化的人工智能开发平台。

通过可视化编程界面、自动化调参与模型优化工具，平台将专业算法开发能力转化为普惠化工具，降低技术应用门槛。这种能力下沉机制促进多领域协作开发，推动技术创新从头部企业向全行业渗透。**其三，智能服务的标准化输出与场景化适配体系。**云平台将人工智能能力封装为标准化服务接口，以模型即服务模式向企业输出自然语言处理、计算机视觉等功能。一方面通过领域数据微调与策略优化，使通用模型具备行业特性；另一方面集成数据脱敏、合规审计等多个模块，满足金融、医疗等领域的特殊要求。标准化与定制化结合的模式，推动人工智能技术从通用能力向产业专属解决方案转化。

2．服务模式体系多元适配行业需求，面临差异化安全风险挑战

一是公有云托管模式以标准化服务体系实现算力资源的普惠供给与弹性配置。第一是模式体系方面。借助公有云平台向用户提供SaaS、PaaS、IaaS等全栈人工智能服务，支持以按需付费方式灵活获取算力资源与模型能力，形成开箱即用的技术供给体系。**第二是核心优势方面。**其一，通过弹性算力调度机制，实现业务流量动态扩容，有效降低企业硬件初始投入。其二，依托标准化服务接口，简化技术集成流程，为中小微企业快速落地人工智能应用创造条件。其三由云服务商承担集中化运维工作，持续提供技术更新与安全防护支持，降低企业管理成本。**第三是安全风险方面。**其一，多租户架构存在数据隔离失效隐患，可能导致跨企业数据泄露风险。其二，服务控制权依赖第三方云服务商，易出现API滥用与模型能力非法调用问题。其三，

跨境数据流动面临不同司法辖区合规监管要求冲突等问题。

二是私有化部署模式依托物理隔离架构构建数据全链路的更高安全方案。第一是模式体系方面。将人工智能平台部署于企业自有数据中心或专属云环境，通过物理隔离手段实现数据采集、存储、处理全链路管控，重点服务政务、金融、能源、医疗等数据安全高敏感领域。**第二是核心优势方面。**其一，物理隔离架构确保核心数据不出本地，严格符合合规要求。其二，企业可掌控模型训练权限与数据访问流程，有效防范核心算法与商业数据外泄风险。其三，本地化运行特性降低对网络连通性依赖，保障业务连续性。**第三是安全风险方面。**其一，自建基础设施安全防护水平参差不齐，易存在设备漏洞与管理盲区。其二，本地化运维对企业技术能力要求较高，配置失误可能引发安全漏洞。其三，模型更新与算力扩容依赖企业自主投入，技术迭代速度或滞后于市场服务商。

三是混合云协同模式通过公有云私有云资源整合实现灵活性与安全性的动态平衡。第一是模式体系方面。有机融合公有云弹性算力与私有云安全优势，将非敏感业务部署于公有云，核心数据与关键模型留存私有云，通过专线网络实现跨平台协同作业。**第二是核心优势方面。**其一，敏感数据本地存储计算，显著降低公有云环境下的数据泄露风险。其二，非核心业务借助公有云资源，避免自建基础设施资源闲置。其三，动态资源调度机制支持业务峰值算力扩展，优化整体成本效益。**第三是安全风险方面。**其一，公有云私有云数据交互过程存在传输加密失效风险，易遭受中间人攻击。其二，跨平台架构安全

策略统一难度大，易形成防护薄弱环节。其三，混合云管理复杂度较高，权限划分模糊可能导致数据访问失控。

四是大模型一体机模式凭借软硬一体化集成提供高效安全的本地化部署路径。第一是模式体系方面。以一体化设备集成高性能算力集群、预优化大模型及安全管控模块，提供从模型训练到推理的全流程本地化解决方案，适用于对算力性能与部署效率多重要求的场景。**第二是核心优势方面。**其一，高算力配置支持大模型的快速训练与本地推理，显著提升计算效率。其二，预集成行业适配模型与数据脱敏、模型水印等安全机制，实现技术能力与安全防护同步部署。其三，设备化交付模式简化部署流程，降低企业技术实施门槛与运维复杂度。**第三是安全风险方面。**其一，一体机硬件架构可能存在供应链安全隐患，面临后门或固件漏洞风险。其二，本地化模型更新依赖企业自主操作，易因版本管理不当引发安全漏洞。其三，设备内数据存储与传输若缺乏有效加密，可能导致敏感数据泄露。

五是模型即服务（MaaS）模式以标准化接口推动人工智能能力的低门槛普及与规模化应。第一是模式体系方面。通过 API/SDK 接口开放预训练模型能力，提供涵盖通用领域与行业定制的标准化服务，助力用户在无需掌握算法开发技术的前提下集成人工智能功能。**第二是核心优势方面。**其一，标准化接口显著降低技术应用门槛，加速人工智能技术在中小企业的普及进程。其二，模型服务商、云服务商持续优化模型性能，使用户及时获取最新技术成果；其三按调用量付费模式降低初期投入成本，便于业务场景快速验证。**第三是安全风险方**

面。其一，模型特性导致输出结果可解释性不足，易引发决策偏差与合规争议。其二，API 调用过程存在密钥泄露风险，可能被用于恶意内容生成或数据窃取。其三，模型服务版本迭代可能引发兼容性问题，影响业务系统稳定运行。

（二）云上人工智能安全面临多维度风险困境，需以协同创新筑牢防护屏障

1．面临系统性安全风险，亟需构建全栈协同防护体系

一是人工智能全生命周期安全风险凸显。第一，开发训练阶段易受数据投毒与后门植入攻击，恶意篡改训练数据或算法逻辑，导致模型基础决策能力失效。**第二，**部署推理阶段面临对抗样本干扰与模型窃取风险，攻击者通过特制数据误导模型输出，或逆向解析核心算法，威胁企业技术资产安全。**第三，**迭代更新阶段存在版本管理漏洞，新旧模型切换时若未全面验证，可能引入新的安全隐患，影响业务连续性。

二是人工智能系统应用风险向产业生态蔓延。第一，关键领域应用场景中，模型决策失误或被恶意操控将直接引发严重后果，如金融风控误判导致资产损失、自动驾驶系统错误指令威胁公共安全。**第二，**人工智能生成内容的滥用加剧信息安全风险，深度伪造技术被用于虚假新闻传播、身份欺诈等，冲击社会信任基础。**第三，**跨企业数据与模型共享协作过程中，若缺乏有效隔离与权限管控，易造成敏感信息泄露或技术成果非法复用，扰乱市场竞争秩序。

三是安全治理体系与技术发展存在代差。第一，企业安全认知与防护能力滞后，仍沿用传统网络安全方案应对人工智能新型风险，对模型安全、算法可信等核心环节缺乏专项防御策略。**第二，**监管标准尚未统一，数据跨境流动、算法伦理等合规要求存在地域差异，企业难以形成普适性安全管理框架。**第三，**产学研协同机制不完善，技术创新与安全研究协同不够紧密，新型攻击手段与防御技术发展失衡，亟需通过多方联动构建动态防护体系。

四是人工智能供应链安全风险加剧。第一，开源框架与第三方组件存在安全隐患，未经充分审计的代码库可能携带漏洞或恶意脚本，随着系统集成扩散风险。**第二，**硬件供应链的不可控因素威胁系统安全，芯片、传感器等关键设备若存在后门或生产缺陷，易被攻击者利用实现远程控制或数据窃取。**第三，**服务供应商的安全管理漏洞会产生连锁反应，模型训练外包、云计算服务租赁等合作模式下，一旦服务商防护体系薄弱，将导致客户数据与业务系统面临被入侵风险。

2．构建多元安全技术协同创新体系，筑牢产业高质量发展安全屏障

一是人工智能赋能安全，革新云上安全防护模式。第一是实现数据安全的智能化管理。机器学习算法能够对云上海量数据进行实时扫描与分析，例如通过聚类算法精准定位敏感数据存储位置，利用分类模型自动标记高风险数据资产。在医疗领域，自然语言处理技术可识别电子病历中的患者隐私信息；在制造业，该技术能检测生产流程数

据中的核心工艺参数;在政务系统中,可筛查公文档案中的机密信息。一旦发现异常数据访问行为,系统便及时发出预警,有效预防数据泄露风险。**第二是**提升威胁检测与响应效率。基于深度学习的异常检测模型,可通过对历史业务数据的持续学习,动态构建正常业务行为模型。当新型网络攻击如 DDoS 攻击变种、零日漏洞攻击出现时,模型能在毫秒级时间内识别异常流量与操作,并且人工智能驱动的自动化响应系统会迅速执行阻断攻击源、隔离受感染服务器、修复系统漏洞等操作,将安全事件的影响范围缩至最小。**第三是**强化身份认证与访问控制。借助生物识别技术与人工智能算法,可实现多因素身份认证的动态化升级。例如通过行为分析模型,持续监测用户操作习惯、设备环境等特征,对异常登录行为进行二次验证,有效防范身份盗用风险;同时,基于权限动态管理策略,根据用户实时行为与业务需求,智能调整访问权限,避免权限滥用。**第四是**推动安全运维自动化。智能运维系统可利用人工智能技术,自动分析系统日志、性能指标等数据,提前预测硬件故障、软件冲突等潜在风险,并触发自动化修复流程。此外,通过生成式人工智能技术,还能自动生成安全配置策略与应急预案,大幅降低人工运维成本与配置失误风险。

二是应对人工智能安全风险,需多技术协同防护。第一是抵御数据攻击,保障数据全生命周期安全。其一在数据采集阶段,可运用差分隐私技术向原始数据添加可控噪声,保护个体隐私信息,防止恶意设备或篡改程序获取敏感数据。例如在智能医疗设备采集患者健康数据时,模糊化敏感细节。其二在数据存储与传输阶段,可采用同态加

密技术，实现数据加密状态下的计算处理，并建立安全传输通道，杜绝数据窃取、篡改风险。其三是在模型训练阶段，借助联邦学习技术打破数据孤岛，实现数据不动模型动；同时部署数据投毒检测技术，识别并剔除污染数据，保证训练数据真实性。其四可建立数据使用审计机制，对数据访问、共享行为进行全程记录与监控，确保数据使用合规。**第二是防范模型攻击，构建模型全链条防护体系。**模型开发阶段，实施对抗训练技术，主动引入对抗样本增强模型鲁棒性；同时利用代码审计工具对模型代码进行漏洞扫描，从源头提升安全性。模型部署后，应用模型水印技术在参数中嵌入独特标识，用于版权追溯；搭建实时监测系统，结合对抗样本检测技术，抵御恶意攻击。模型更新与维护阶段，建立严格的版本控制机制，确保更新过程可追溯；持续运行模型投毒检测，监控数据质量，防止模型被恶意篡改，保障迭代安全。同时，定期对模型进行安全评估，模拟攻击场景测试模型防御能力，及时优化防护策略。**第三是应对应用层攻击，夯实平台与服务安全基础。**可部署 API 网关统一管理人工智能服务接口，提供流量监测、漏洞扫描、访问控制等多重保护，约束 API 调用规则，防止接口被恶意利用。同时，可构建安全审计系统，全面记录用户操作、数据访问、模型调用等行为；基于大数据与人工智能搭建态势感知平台，实时监测平台安全状况，及时发现并预警潜在风险。**第四是建立应急响应机制，**针对应用层攻击制定标准化处置流程，确保快速恢复服务环境完整性与稳定性。

3．多元主体协同发力，构建健全云上人工智能安全生态

一是多元主体分工协作，搭建多层次安全产品体系。第一，云服务商构建底层算力安全与平台级安全能力体系。通过模型托管平台实现对云上模型部署的全流程管控，利用访问控制系统与审计机制，保障模型调用的合规性与可追溯性；同时提供计算资源隔离、动态权限管理等基础安全服务，降低底层设施引发的安全风险。**第二，**人工智能服务商专注于模型本体安全防护。通过部署提示注入防御机制，防止恶意指令诱导模型输出有害内容；构建对抗攻击评估体系，提升模型在对抗样本下的鲁棒性；运用模型水印嵌入技术，实现对模型知识产权的保护及非法使用行为的追踪。**第三，**安全服务企业凭借行业经验，提供定制化安全解决方案。例如若针对政务领域的严格合规要求，开发人工智能合规审查系统，实时监测模型调用行为；若面向金融行业的数据安全需求，推出人工智能 API 网关解决方案，精准限制模型调用频率、过滤非法参数注入，防范数据泄露与业务逻辑篡改。

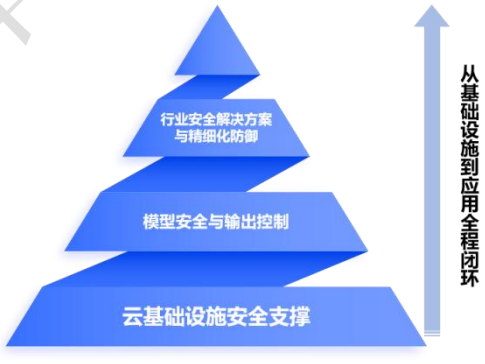

图 2 多层次安全产品体系

二是多方主体协同联动，构建人工智能安全发展新生态。多方主体协同联动，从机制、技术、人才、生态等维度构建系统化安全治理体系。**第一在安全评估治理机制上，**国内以产学研协同为核心，头部

企业将实践经验转化为技术规范，为标准制定提供支撑，中国信息通信研究院等第三方机构牵头制定人工智能安全标准并开展评估验证，推动标准在重点行业落地，高校与科研机构则从理论层面创新评估方法以量化模型风险。**第二在技术创新领域，**企业与科研机构聚焦人工智能安全关键技术联合攻关，通过整合产业资源与科研力量加速新技术转化。**第三是人才培养方面，**企业与高校合作开设专业课程、共建实训基地，行业协会通过技能竞赛与培训讲座等活动，共同构建多层次人才培育体系。第四是生态共建层面，产业联盟汇聚产业链上下游企业共探安全发展路径，开源社区通过吸引全球开发者贡献技术方案，以开源协作推动人工智能安全领域创新生态繁荣。

（三）云上人工智能安全相关法律法规不断完善，标准化治理加速推进

我国在人工智能安全领域坚持“发展和安全并重、促进创新和依法治理相结合”的原则，依托《网络安全法》《数据安全法》《个人信息保护法》等基础法律，逐步完善涵盖算法治理、数据安全、伦理审查和内容标识的法律法规体系。

2022 年 3 月，由国家互联网信息办公室、工业和信息化部、国家市场监督管理总局联合发布的《**互联网信息服务算法推荐管理规定**》正式实施，加强了对算法推荐行为的规范与监管。

2022 年 11 月，国家互联网信息办公室、工业和信息化部、公安部联合发布了《**互联网信息服务深度合成管理规定**》，明确深度合成

技术的管理要求，提升技术应用的安全性和合规性。

2023 年 7 月，国家互联网信息办公室联合工业和信息化部等七部门共同发布了《**生成式人工智能服务管理暂行办法**》，推动生成式人工智能的规范化发展。

2023 年 9 月，针对人工智能技术研发的伦理问题，实施了由相关部门牵头制定的《**科技伦理审查办法（试行）**》，强化科研和技术开发的伦理监管。

2025 年 3 月，国家互联网信息办公室、工业和信息化部、公安部、国家广播电视总局联合发布了《**人工智能生成合成内容标识办法**》，明确人工智能生成内容的标识要求，推动内容透明度和责任追溯。

上述法律法规的陆续出台，构建了覆盖广泛、层次清晰的人工智能安全治理制度体系，为构建安全、可信的人工智能发展环境提供了坚实保障，并加快了标准化治理的步伐。

（四）云上人工智能安全发展意义重大，多维价值支撑产业前行

1．筑牢企业业务安全防线，夯实高质量发展根基

一是降低安全事件损失，守护企业核心利益。在数字化转型深化背景下，企业对人工智能技术的依赖度持续提升，安全漏洞可能引发系统性风险。部分企业的人工智能业务系统遭恶意攻击后，出现数据篡改、算法偏差等问题，不仅面临高额经济损失，更因客户信任流失导致市场份额萎缩。相关统计显示，近年因人工智能安全事件导致企

业面临巨额损失的案例屡见不鲜，人工智能技术安全已成为企业核心竞争力的重要组成部分。

二是保障业务连续性，避免服务中断风险。云服务的稳定运行是企业维持经营的关键前提，而安全防护缺失可能导致服务链路断裂。部分依赖人工智能的服务平台因安全防护薄弱，遭遇网络攻击后出现系统瘫痪，直接影响用户服务体验，引发企业损失。实践表明，构建全链路安全防护体系，是保障企业业务连续性的必要条件。

2．激活技术创新发展动能，拓展产业应用边界

一是增强用户信任基础，破除技术应用壁垒。人工智能系统的安全可靠性直接影响用户技术认同度。当人工智能应用出现数据泄露、算法被篡改等安全事件时，用户对技术的抵触情绪显著上升，甚至导致大规模用户流失。调研数据显示，超七成用户会因人工智能安全问题减少或停止使用相关服务，技术安全已成为制约人工智能普及的核心瓶颈。安全是人工智能技术广泛应用的基础，只有保障安全，才能增强用户对人工智能技术的信任，推动人工智能技术的健康发展。

二是推动关键领域应用拓展，释放产业价值潜力。安全保障是人工智能技术进入金融、医疗等敏感领域的必要条件。例如：在金融行业，人工智能风控系统需满足数据加密、算法可解释性等严格要求，才能规避系统性风险；在医疗领域，人工智能诊断工具必须通过安全认证，确保诊疗决策的准确性与可靠性。合规安全体系的完善，为人工智能技术在关键领域的深度应用开辟了路径。

3．筑牢社会公共安全屏障，维护数字时代稳定

一是防范人工智能恶意滥用，遏制技术风险外溢。人工智能技术若被用于恶意场景，可能对社会秩序造成冲击。近年来，利用人工智能合成的虚假信息在网络空间快速传播，影响公众认知与社会稳定；同时，人工智能赋能的自动化攻击技术被用于网络犯罪，导致钓鱼攻击、数据窃取等安全事件数量显著上升。加强人工智能应用的安全监管，是防范技术滥用风险的必然要求。

二是守护关键领域安全，保障社会运转基石。政务、金融、交通、能源等关键基础设施的人工智能安全直接关系公共利益。自动驾驶系统若遭黑客攻击可能引发重大安全事故；智能电网的调度系统漏洞可能导致区域性能源供应中断。实践表明，构建关键领域人工智能应用的安全冗余机制，是守护关键领域运转的重要保障。

4．提升企业安全防线，夯实高质量发展根基

一是技术革新驱动安全升级的必然要求。第一是融合发展催生安全新挑战。人工智能与云计算的深度融合，在释放技术红利的同时加剧了安全风险的复杂性。一方面，多模型共享云资源模式下，数据泄露风险呈指数级增长，攻击者可能通过单点突破获取跨系统敏感信息；另一方面，人工智能技术的智能化特性使攻击手段更趋隐蔽化、自动化，传统安全防护体系已难以应对新型威胁，亟需构建动态自适应的安全防护体系。**第二是新兴技术带来安全威胁迭代。**量子计算等前沿技术的突破，对现有加密体系形成颠覆性挑战。理论上，量子计算机

可在短时间内破解加密算法，导致系统中数据存储、传输及模型保护机制失效。若关键领域人工智能应用的加密防线被突破，将直接威胁经济安全与社会稳定，亟需提前布局抗量子加密技术及安全方案。**第三是技术架构复杂化放大安全隐患。**人工智能系统技术架构日趋复杂，分布式部署、模块化设计等特性显著扩大了攻击面。系统组件间的接口暴露、调用逻辑漏洞，以及第三方依赖库的安全缺陷等问题频发，成为黑客攻击的潜在切入点。一旦关键组件被利用，极易引发数据泄露、模型篡改等连锁反应。

二是市场需求倒逼安全体系完善的现实选择。第一是企业数字化转型的安全刚需。随着企业加速上云与人工智能应用落地，数据泄露、服务中断等风险成为制约转型的关键因素。关键基础设施行业因安全事件导致的年均损失达数亿元。构建覆盖数据全生命周期、人工智能模型全链条的安全防护体系，已成为企业保障核心竞争力的必答题。**第二是用户安全诉求升级的市场倒逼。**用户数据安全与隐私保护意识的提升，重塑了云服务市场竞争格局。个人用户对云存储、智能终端的隐私保护要求不断提高，企业客户则更关注人工智能系统的合规性与抗攻击性。安全能力在某种意义上已超越价格因素，成为用户选择云服务的首要决策指标，倒逼服务提供商持续提升安全服务水平。

二、云上人工智能安全风险分析与防护体系建设

（一）云上人工智能安全风险分析，全流程暴露安全隐患

在大模型应用的全生命周期中，从准备到应用的各个环节均面临系统性安全风险挑战，这些风险贯穿数据合规性、基础设施脆弱性、模型自身缺陷及应用场景复杂性等多个维度，亟需构建覆盖全生命周期的安全防护体系，实现对大模型应用风险的主动防御与系统性治理，为人工智能技术的健康发展提供坚实保障。

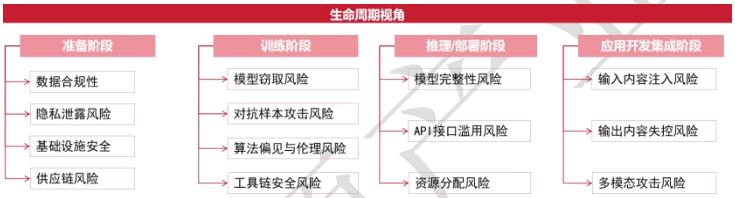

图 3 云上人工智能安全风险概览

1．准备阶段：数据与基础设施风险

一是数据合规性与隐私泄露风险。云上大模型的训练数据体系构建过程中，数据来源的合法性与隐私保护构成核心挑战。**其一，**训练数据来源可能包含第三方数据集，存在恶意污染或偏见数据被系统性引入的可能性。攻击者通过篡改数据标签、插入虚假样本或干扰语义结构，可能破坏模型的认知逻辑与泛化能力，导致输出结果失真。**其二，**多源异构数据的聚合过程可能引入未授权的敏感信息，如个人身份标识、商业机密或受版权保护内容等敏感内容，此类数据若未经严格合规审查即用于模型训练，将可能直接违反数据保护法规，引发相关法律纠纷与声誉损失，形成数据合规性与业务连续性的双重风险叠

加。

二是基础设施安全与供应链风险。云平台作为大模型训练的核心载体，其架构开放性与资源分布特性天然存在安全隐患。存储、网络与算力资源的跨域调用模式在提升系统弹性的同时，也扩大了攻击暴露面。**其一，**在多租户环境下，若虚拟化隔离机制存在设计缺陷，攻击者可能通过虚拟机逃逸漏洞突破虚拟机边界，直接访问宿主机或其他租户的虚拟机资源，导致训练数据、模型参数及用户隐私等敏感信息泄露。**其二，**模型的训练和推理依赖海量用户隐私信息或商业机密的数据集，而云服务通常需要动态管理大量配置项，包括网络配置、存储配置、安全组规则等。错误的配置可能导致数据被未经授权的用户访问，进而引发大规模数据泄露事件。**其三，**供应链层面的系统性隐患，如开源框架的版本依赖漏洞、第三方组件的隐蔽后门植入以及硬件设备的固件级缺陷，均可能被恶意利用构建持续性威胁通道，造成云上大模型生态系统的结构性安全危机。

2．训练阶段：模型安全与对抗攻击风险

一是模型窃取风险。在当前云计算环境中，模型训练过程面临多重数据泄露隐患。**其一，**攻击者可通过高频访问接口获取模型迭代过程中的梯度向量和中间激活值，结合差分隐私攻击技术逐步重构模型架构。典型场景如联邦学习框架下，若未采用差分隐私机制或安全聚合协议，恶意参与方可能通过逆向工程解析本地模型更新中的敏感特征。**其二，**训练过程中产生的模型快照或中间层输出若缺乏访问控制，

可能被恶意节点截获并用于构建近似模型，此类知识产权侵犯行为可能引发后续供应链污染风险。

二是对抗样本攻击风险。对抗样本攻击已形成系统性威胁，其核心在于通过微扰动实现对模型决策边界的操控。攻击者可采用白盒、黑盒混合策略，在训练集中植入经过精心设计的噪声样本，这些样本在常规数据预处理中难以被识别，却能引导模型建立错误的特征映射关系。如在图像分类任务中，对抗样本可能导致模型对特定类别的判别能力显著下降。这种攻击的隐蔽性在于常规数据清洗难以识别扰动样本，且其影响具有滞后性。模型在训练完成后的推理阶段才显现异常输出，同时对抗样本的可迁移性可能波及多个模型，形成系统性脆弱性。

三是算法偏见与伦理风险。训练数据中的结构性偏差可能被模型指数级放大，形成系统性歧视。**其一，**当训练集在性别、种族等维度存在样本分布失衡时，模型可能通过注意力机制强化既有偏见，导致决策结果偏离社会公平准则。例如若训练数据中某类群体的样本占比失衡，模型可能在训练过程中形成偏差性权重分配，进而影响决策公平性。**其二，**模型训练过程的黑箱属性使得偏见溯源与干预手段受限，可能削弱公众对人工智能技术的信任基础，并引发监管机构对训练数据透明度的强制要求。

四是工具链安全风险。训练大模型所依赖的开源框架、依赖库及第三方工具链存在潜在安全漏洞。例如训练框架的版本更新滞后可能引入已知漏洞，依赖库的供应链污染可能嵌入恶意代码，硬件加速库

的后门程序可能被用于窃取模型参数或训练数据。此类风险在训练阶段尤为致命，攻击者可能通过工具链漏洞篡改训练逻辑，导致模型收敛至恶意目标。同时，分布式训练环境中的节点异构性进一步放大风险，攻击者可能利用节点间的通信漏洞实施中间人攻击，破坏训练数据的一致性与完整性。

3．推理部署阶段：模型完整性与访问控制风险

一是模型完整性风险。模型在部署过程中若缺乏严格的版本控制与验证机制，可能面临被篡改或替换为含后门版本的威胁。**其一，**攻击者可能通过供应链污染手段，在模型文件中植入恶意逻辑，例如在权重矩阵中嵌入基于模式匹配的触发条件，当输入包含特定特征组合时，模型将输出预设响应，一旦部署即可能对下游应用造成不可逆破坏。**其二，**模型更新机制的安全性同样值得关注，若更新通道未采用端到端加密或数字签名验证，攻击者可能通过中间人攻击替换合法模型，进一步扩大攻击范围。

二是 API 接口滥用风险。开放的 API 接口若缺乏强身份验证与访问控制策略，可能被恶意调用以窃取模型参数或实施服务中断攻击。**其一，**攻击者可通过暴力破解 API 密钥、枚举接口路径或构造对抗性输入绕过内容过滤机制，从而获取敏感信息或触发模型异常行为。例如，通过高频请求耗尽计算资源导致服务不可用，或通过设计的输入样本诱导模型输出非预期结果。**其二，**API 接口的开放性可能被用于横向渗透攻击，攻击者通过利用 API 漏洞作为跳板，进一步入侵云平

台或关联系统，形成链式安全风险。

三是资源分配风险。在多用户共享算力的部署环境中，攻击者可能通过资源调度机制实施攻击。恶意用户可能通过异常资源请求例如伪造高优先级任务引发算力资源的非授权占用，导致其他用户的服务请求出现显著延迟甚至完全中断，形成针对性的拒绝服务攻击。该类资源竞争风险可能引发跨租户的数据隔离机制失效，关键服务的系统稳定性受损以及共享计算资源的可用性遭受持续性破坏，对整个云计算基础设施的安全性构成系统性威胁。

4．集成应用阶段：内容安全与多模态攻击风险

一是输入内容注入风险。在人机交互场景中，用户输入若缺乏严格的安全校验机制，可能成为恶意攻击的切入点。**其一，**攻击者通过对抗性提示诱导模型生成违规内容，例如利用上下文嵌套技巧绕过内容过滤规则，或通过语义伪装规避关键词检测，表面上符合常规交互模式，实则通过模因化表达或跨语言转换传递恶意意图。**其二，**攻击者可能通过构造多轮对话链逐步诱导模型输出非法指令，或利用模型对模糊语义的容忍特性，生成具有危害性的文本内容。同时，可能通过自动化工具批量生成攻击样本，形成规模化渗透，对内容安全防护体系构成挑战。

二是输出内容失控风险。模型生成的内容若缺乏动态审核与实时干预机制，可能因训练数据偏差、算法逻辑缺陷或对抗性输入导致内容失控。**其一，**生成的文本可能面临虚假信息风险，如伪造的新闻报

道、虚构的学术成果；隐私泄露风险，如意外输出训练数据中的个人信息；违法内容风险，如违反社会公序良俗的表述。**其二，**模型的黑盒属性使其难以追溯生成路径，即使发现异常输出也难以定位具体触发条件或责任主体。同时，生成内容可能被滥用，进一步放大社会危害，可能直接引发法律纠纷或信任危机。

三是多模态攻击风险。在视觉—语言、语音—文本等多模态应用场景中，攻击者可能通过跨模态输入诱导模型生成有害文本。在图像识别任务中，攻击者可能在图像中嵌入隐蔽噪声模式或语义触发器，使模型输出与图像无关的极端主义言论；在语音—文本交互场景中，攻击者可能通过特定频段的音频信号干扰模型对语音内容的理解。此类攻击的复杂性在于攻击向量跨越不同模态，传统单模态审核机制难以覆盖。这种跨模态的攻击传播特性增强了隐蔽性，可能形成多级联锁反应，使攻击效果在不同模态间持续放大。

（二）云上人工智能安全防护体系建设实践，全链条筑牢安全防线

1．明确云上人工智能安全防护需求，锚定防护目标方向

模型应用起步于基于自然语言的问答式交互模式，大量安全防护聚焦模型本体。大模型应用自以 ChatGPT 为代表的问答式交互兴起，在该业务形态下，用户通过提问与大模型互动，模型基于训练数据及公开信息生成响应。此阶段，安全建设的核心集中于模型本身的风险防控，且相关安全责任主体主要为模型构建团队。同时，云上大模型

的训练与推理环节逐步成熟，涵盖大规模分布式计算资源管理、弹性算力调度以及模型参数的安全加密存储，这些云基础设施与服务的安全保障也成为防护体系的重要组成。

图 4 基于自然语言的问答式应用

基于知识库和精调大模型的知识引擎式应用拓展了云上大模型的安全边界。随着通用大模型在知识准确性方面存在“幻觉”问题，特定行业业务场景开始通过私有知识库、增强检索及模型微调构建更精准的行业大模型，并广泛应用于智能客服、文档管理等领域。云端多租户环境下，数据隔离、权限控制与访问审计成为防护的要点，数据安全和内容安全的防护责任由单一模型团队扩展至涵盖数据管理与安全运维团队，形成覆盖训练数据、知识库静态存储及在线访问的多层安全保障。

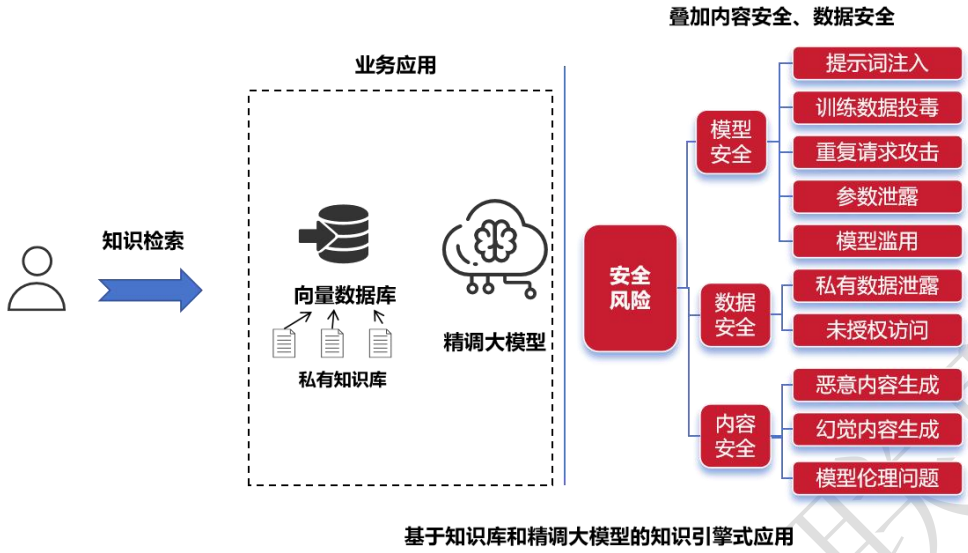

图 5 基于知识库和精调大模型的知识引擎式应用

云上智能体应用与 MCP 协议的发展，再度提升云上大模型安全治理难度。智能体通过理解用户指令、形成决策链并调用多样化的工具完成任务，实现从推理到执行的闭环。MCP 协议快速推进，降低智能体工具接入门槛，催生繁荣的工具生态环境。此过程中，安全边界进一步扩大，除了模型安全、内容安全、数据安全外，工具安全、协议安全和云上运行环境的安全风险日益突出。尤其云上环境中容器安全、网络隔离、身份认证及权限管理需配合智能体的动态调度机制。MCP 协议的安全范式尚未完全定义，给云环境中的多方协同安全带来挑战，且参与方从单一团队扩展到多个内外部工具提供商，管控复杂度大幅提升。

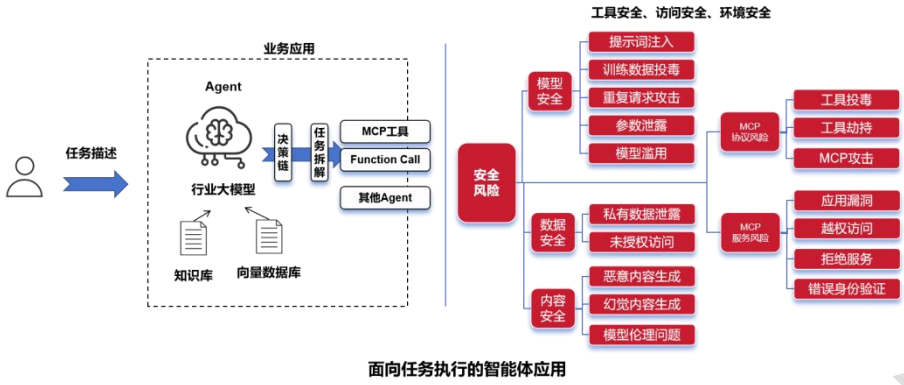

图 6 面向任务执行的智能体应用

云上人工智能应用安全防护呈多层次多阶段协同的复杂态势。当前安全团队在大模型输入阶段通过提示词安全规则、语义与关键词过滤，实现基础的风险筛查；但智能体任务执行阶段的隐蔽攻击如调用工具获取敏感凭证等，传统文本过滤难以覆盖，依赖模型内置安全小模型与伦理审查机制的风险防控在保障灵活性的同时，缺乏对完整工具链的安全掌控。任务执行环节主要依赖传统 WEB 安全防御措施，但面对工具投毒、劫持、伪造及权限诱导越权等新兴攻击，防护能力明显不足。云上分布式架构下，任务调度和容器编排的安全加固也是亟待加强的方向。

构建系统化、全链条的云上大模型安全防护体系势在必行。通过实现训练、推理、开发及部署各环节的安全信息共享与数据互通，构建分层阻断和协同联动的安全壁垒，打造与云上大模型业务形态深度融合的安全模式。在业务流程中覆盖数据采集、训练环境、模型推理、决策链条以及工具调用等多源信息，实现安全判断的全局视野。依托安全大模型、安全小模型和安全规则集的协同能力，动态识别和应对针对模型的攻击，结合经验规则防范传统威胁。在云端实现多阶段精

准阻断，最大限度释放安全风险，筑牢云上大模型的安全防线。

2．构建全链覆盖人工智能风险治理框架，统筹安全治理路径

随着云计算与大模型技术的深度融合，人工智能服务呈现多样化发展态势，涵盖公有云托管、私有化部署、混合云协同、模型一体机、模型即服务（MAAS）等多层次、多形态的服务模式。不同服务模式在权限管理、资源隔离与运维管理等方面具有各自特点，伴随各阶段生命周期环节的不同，也带来了多样化的安全风险和挑战。为此，需构建包含防护边界与防护关键点在内的整体视角云上人工智能风险防护框架。

图 7 云上人工智能风险防护边界

一是划分防护边界。防护边界是所有用户和服务都需关注的防护重点，可依照以下两个维度进行划分和实施：**其一是服务模式维度。**针对公有云托管、私有化部署、混合云协同、模型一体机和模型即服务（MAAS）等多样化服务形态，分别制定差异化的边界安全策略，明确权限管理、资源隔离和运维安全等关键防护措施。**其二是生命周期维度。**构建覆盖准备、训练、推理部署、应用开发等各个环节的具

体安全要求，调配相应的边界防护能力，确保安全策略贯穿云上人工智能产品和服务的全生命周期。

二是聚焦防护关键点。用户可基于自身业务特点，针对性聚焦五大关键安全领域，即基础设施安全、模型安全、数据安全、工具安全和内容安全；基于防护边界制定具体防护措施落地路径，覆盖云上人工智能应用不同生命周期中的重点风险点与管理重点，根据自身情况制定差异化防护关键点，保障云上人工智能安全平稳运行。

防护边界与防护关键点相互支撑、协同运作，构建起多层次、多维度的风险防控体系，既确保了不同服务模式和生命周期环节的边界安全，也实现了对核心安全领域的精细化治理，为云上人工智能服务的安全、稳定和高效运行提供坚实保障。

图 8 云上人工智能安全防护体系框架

（1）云上人工智能系统防护边界，划定安全防护范围

防护边界作为风险治理的第一道防线，其作用主要体现在两个方面：**一是强调根据不同服务模式的差异，合理聚焦并部署防护关键点。**不同的服务模式（如公有云托管、私有化部署、混合云协同、模型一体机和 MAAS 模式）在权限管理、资源隔离和运维管理等方面存在

显著差异，这直接影响相关的安全重点和防护策略。**二是结合云上人工智能关键生命周期环节，精准识别风险侧重点。**不同环节面临的安全挑战和风险特征各异，因此需精准识别生命周期环节重点风险，制定包括准备、训练、推理部署及应用开发等在内的差异化风险防护关键。

企业应结合自身所处的服务模式和生命周期环节，明确安全需求，并有针对性地选择和强化防护关键点。防护关键点主要包括基础设施安全、模型安全、数据安全、工具安全及内容安全，通过这种方式，防护边界不仅统一了安全防护的框架视角，也为企业量身定制风险治理方案提供指导，保障安全资源与能力的高效配置，持续提升云上人工智能的安全韧性与合规水平。

（2）云上人工智能系统防护关键点，锁定核心防护对象

结合防护边界对服务模式和生命周期的差异化指导，聚焦防护关键点，实现针对性和精细化的安全防护，为云上人工智能应用的健康发展筑牢坚实基础。

一是基础设施安全。作为云计算环境中大模型安全的核心支撑，基础设施的稳健性直接决定模型训练与推理的可靠性和效率。其风险主要涵盖存储、网络、系统及算力调度四大关键领域。**第一是存储安全。**需严防数据泄露、篡改及备份风险，建立多层加密体系和细粒度访问控制机制，保障数据在静态和传输过程中的安全性，同时完善数据备份及灾难恢复方案，确保数据完整性和业务连续性。**第二是网络**

安全。作为数据传输和资源连接的关键环节，应采用虚拟私有云、微分段及零信任架构，结合先进防火墙、入侵检测与端到端加密技术，有效抵御非法入侵、中间人攻击等威胁，确保网络链路的安全可靠。**第三是系统安全。**涵盖操作系统、虚拟化平台与容器环境，需强化系统加固、及时修补漏洞及严格权限管理，通过应用隔离和访问控制，阻断潜在恶意行为，保障运行环境的纯净与稳定。**第四是算力调度安全。**关注计算资源的合理分配与防护，需实施严格身份认证、多维权限管理和实时监控，实现多租户环境下算力资源的公平隔离与弹性伸缩，并结合流量控制和异常检测机制，防范拒绝服务攻击，保障算力服务的持续稳定。通过构建覆盖存储、网络、系统及算力四大领域的全链路安全防护体系，融合先进技术与科学管理手段，可有效云上人工智能服务的安全可靠性。

二是模型安全。作为人工智能系统的核心资产，模型安全直接决定整体系统的可信度和商业价值，面临模型窃取、对齐失败、对抗攻击及开发流程安全等多重复杂风险，亟需建立系统化、全生命周期的安全治理框架。**第一是模型窃取防护。**涉及知识产权和核心技术泄露。攻击者可能通过频繁接口调用、反向推断等手段窃取模型结构和参数，甚至复制模型谋取非法收益。对此，应通过限制接口访问频率和调用行为，采用模型水印与指纹技术实现溯源与篡改检测，以及利用输出扰动技术混淆推断路径，降低模型被逆向工程的风险。**第二是模型对齐。**旨在确保模型输出符合伦理规范和业务预期。面对模型可能存在的偏见、误判等问题，应持续监控模型行为，结合先进对齐方法对模

型进行动态训练与调整，确保决策合理、可靠，符合法律和社会伦理要求。**第三是对抗攻击的防护。**对抗样本通过微小扰动诱导模型误判，危及模型鲁棒性。防御措施包括对抗训练以增强抵抗力、多样化数据提升泛化能力，定期开展鲁棒性验证和安全测试，及时调整防护策略应对新型攻击。**第四是模型开发安全。**关系安全治理的起点。开发过程中代码漏洞、后门植入及版本管理不善可能引发重大风险。应构建标准化安全开发流水线，严格代码审计、版本控制和动态安全检测，强化开发人员安全意识和权限管理，尤其在协同开发环境下实施多方职责分离，保障开发过程透明合规。总体而言，模型安全防护需融合技术创新与科学管理，覆盖模型设计、训练、部署及维护全生命周期，构筑稳健可信的安全防线。

三是数据安全。作为大模型训练与推理的基础，数据的质量与安全水平直接影响模型性能和系统合规性，涵盖隐私保护、数据质量管理及多源数据的安全整合，科学的数据安全治理是保障大模型安全稳定运行和风险合规的关键支撑。**第一是数据清洗。**专业化和自动化的数据清洗是确保数据质量的前提。训练数据中的冗余、错误或异常值不仅引入噪声，还可能成为恶意攻击的载体，损害模型性能与行为准确性。应部署高效的自动化清洗工具，结合机器学习技术精准识别异常数据，及时剔除或修正不合规样本，提升数据可信度。**第二是数据补全。**通过插值、生成模型等技术修复缺失数据，维护训练集的完整性和多样性，防止样本偏差提升模型泛化能力和稳定性。**第三是数据增强。**利用多样化变换和合成技术扩充数据空间，增强模型对复杂环

境和多样输入的适应性。尤其是对抗样本生成，不仅丰富训练场景，更强化模型鲁棒性，成为行业主流手段。**第四是数据融合。**多源异构数据融合面临隐私保护和安全合规挑战。需依托联邦学习、同态加密等先进隐私保护技术，结合严格权限管理和合规审计，保障数据融合过程的透明安全。总体来看，云上人工智能数据安全治理应兼顾数据质量与隐私保护，构建系统完备的治理体系，支撑高质量大模型训练，确保系统决策的可信性和合规性。

四是工具安全。作为人工智能模型开发与运维的重要环节，工具链的安全性直接影响整体系统的纯洁性和可靠性。其风险点主要涵盖代码安全、安全后门、配置管理、供应链稳定性及 API 安全，需通过全面的安全管理措施加以防范。**第一是代码安全。**代码中存在的漏洞或恶意模块可能导致模型泄露或异常行为。应结合静态与动态代码分析工具进行严格审计和风险评估，及时发现并修复安全隐患，保障代码质量和安全。**第二是安全后门风险。**尤其在第三方组件中隐蔽且难以检测。需强化供应链安全审计，实行代码签名验证和行为监控，确保软硬件组件来源可信，防止恶意植入。**第三是配置安全。**误操作或权限不足常引发安全事件。有效配置管理需落实版本控制、自动化校验及权限分离，降低人为错误和配置篡改风险。**第四是供应链中断风险。**影响开发与运维的连续性。应构建多源备份体系和应急预案，结合风险预警机制，提升工具链及供应链的稳定性和鲁棒性。**第五是 API 安全。**作为系统开放接口，风险直接影响整体安全态势。通过严格身份认证、访问权限控制、访问频率限制和行为审计，减少滥用和

恶意攻击，缩小攻击面。总体而言，云上大模型工具安全必须适应现代软件开发的敏捷性和生态复杂性，建立涵盖研发至运维的全流程安全保障体系，确保大模型开发与运行环境的安全可靠。

五是内容安全。作为云上大模型整体安全风险治理的关键环节，内容安全涵盖模型输入、训练与输出的全生命周期。随着模型规模和应用场景的多样化，内容安全风险愈发复杂且动态多变，直接影响平台的合规性和社会责任履行。**第一是输入内容检测。**作为防范异常及恶意输入的首要防线，应采用多层次、多维度的过滤机制，精准识别并拦截恶意文本、非法代码及其他高风险输入，保障模型推理的稳定性与安全性。**第二是输出内容审核。**确保模型生成结果合法合规且具备真实性。应结合先进的自动化筛查技术与人工复核机制，利用深度语义理解与风险识别算法，有效防控有害信息传播，降低法律合规风险，强化企业的社会责任感与风险防御能力。**第三是训练内容安全。**严格把控训练数据的伦理合规性，去除潜在的偏见、歧视及违法信息，防止模型产生负面偏向。建设完善的数据审查和质量治理体系，是确保模型公平性和可信性的基础。**第四是内容安全运营管理。**侧重于建立机制化的风险监控与及时响应流程，制定完善的应急预案与处置规范，实现对安全事件的快速识别与闭环处理，确保内容安全事件的实时应对与有效缓解。整体而言，内容安全体系须紧密结合技术创新与管理规范，协调技术、合规与运营多维度资源，系统化推进云上大模型的健康、合规与可持续发展。

3．建立云上人工智能风险识别与预警机制，提升安全防御能力

在云计算环境下，人工智能系统尤其是大模型的应用日益复杂，其风险类型多样且动态演变。为保障系统安全与业务稳定，需构建科学高效的风险识别与预警机制。

一是主动识别大模型内生风险。大模型内生风险主要源自训练数据偏差、模型设计缺陷、参数超调及“幻觉”现象等因素。云上环境汇聚大量异构数据和模型版本，须建立多维度风险识别框架，结合数据质量监控、模型性能评估和行为分析，对训练与推理阶段潜在风险进行实时感知和预警。借助模型可信度评估技术，自动检测异常输出和不合理决策，对内生风险进行动态识别，防止风险在系统内部累积和扩散。

二是加强对衍生风险的动态监控。衍生风险主要包括模型与外部工具、API 及云基础设施交互中产生的安全隐患，如接口协议漏洞、工具调用越权、环境配置失控、数据泄露、身份冒用和权限滥用等。为防范这些风险，可以采用多层次的技术手段，包括日志审计、异常行为检测、访问控制、权限分离与最小权限原则、多因素认证、漏洞扫描及补丁管理等。此外，通过持续安全监控和漏洞管理，结合威胁情报共享与应急响应机制，能够及时发现并缓解潜在风险，保障云上大模型及其配套生态的安全稳定运行。

三是利用人工智能技术实现对风险的智能识别与预测。一方面是传统云安全风险。包括对云基础设施、网络流量、访问权限、身份认证、API 调用及异常行为等的全面监控和防护，旨在保障云平台及其

运行环境的整体安全，防范网络攻击、权限滥用和系统漏洞等常见威胁。**另一方面是大模型业务相关的安全风险。**特别关注模型应用过程中的内容安全问题，如不当内容生成、偏见与歧视风险、数据隐私泄露、模型滥用以及合规伦理风险。针对这些业务层面风险，可依托模型行为分析、内容审核、异常使用检测等专业技术，实现动态监测和实时预警。通过融合机器学习、自然语言处理和异常检测算法，智能风险识别系统能够对海量日志、告警及业务数据进行深度挖掘，准确定位潜在威胁，并基于预测模型预判风险演变趋势，形成预警。此类人工智能驱动的风险识别与预测不仅提升了响应速度和准确性，也促进安全策略的动态调整，助力云上大模型环境实现智能化和闭环式安全治理。

4．提升云上人工智能风险系统综合治理能力，强化整体防控效能

一是加强企业内部多部门角色协作，建立标准化流程规范。人工智能系统的风险治理涉及数据、安全、法律、合规、研发、运维等多个部门，企业需打破部门壁垒，形成协同联动的治理体系。通过明确各职责边界和信息共享机制，制定涵盖需求分析、模型训练、推理服务、工具接入及持续监控的标准化流程，保障风险管理贯彻到每一环节。标准流程不仅规范操作行为，还提升跨团队响应效率，确保安全事件快速定位和有效应对，适应人工智能技术及应用的快速迭代。

二是围绕人才培养与意识培训持续发力，夯实治理基础。人才是

人工智能风险系统治理的根基。组织需结合云计算和大模型技术特点，构建多层次人才培养体系，涵盖安全专家、研发工程师及管理者，注重安全意识和技术能力双提升。通过定期开展安全专题培训、模拟演练、案例分享和技术工具实操，强化员工对安全风险的识别与防范能力。良好的安全培训体系能够激发全员安全责任感，形成积极主动的风险治理文化，为复杂环境下的安全保障提供坚实支撑。

三是推动外部生态共建，构筑健壮的产业攻坚合力。人工智能风险治理不仅依赖单一组织，还需借助开放的产业生态。企业应主动与云服务提供商、工具厂商、合规机构等外部主体建立深度合作，推动安全标准化与规范化进程。通过联合开展安全评测、共享威胁情报和风险控制经验，实现资源互补和能力提升，构筑多方协同的防护网络。此外，依托产业联盟等业内组织，共同编撰安全策略和技术规范，保障整个生态环境的稳定与合规，提升整体抗风险能力。

5．健全标准建设与模型应用风险评估，规范安全评估流程

一是完善云上人工智能安全相关标准建设。针对模型训练准备、训练、部署推理、应用开发等关键环节，制定涵盖基础设施安全、模型安全、数据安全、工具安全及内容安全在内的技术标准和评估方法。同时，形成统一的能力标准、测试指标及风险等级划分，提供有效参考，助力规范化评估和风险管控工作。

二是强化模型自身安全评估。借助相关模型评估平台和测试问题集，对基础模型的安全属性进行更为全面的测试。评估内容可包含模

型偏差、漏洞、对抗样本易感性、信息泄露风险、性能稳定性、鲁棒性、隐私保护及内容安全等多个方面，力求对模型的安全特性进行全面覆盖和深入分析。通过定期更新的问题集，持续跟进最新风险并推动相关改进，最终提升模型自身的安全水平。

三是推动模型应用安全评估。鼓励引入第三方专业机构开展独立测试和评估服务，全面识别和防范模型在具体应用场景中可能面临的安全风险。模型应用环节涉及数据处理、用户交互、权限管理、内容生成及业务逻辑等多个层面，易受到内容安全、隐私泄露、滥用风险及合规性等挑战。通过系统化的安全评估，能够及时发现应用中的潜在漏洞与风险，促进模型应用方完善安全控制措施与治理策略，提升模型服务在不同业务环境下的安全性、可靠性和合规水平。

四是通过风险评估促进模型服务商和应用方强化安全弱点。完善自身安全管理体系和技术防护措施。评估结果不仅为风险治理提供科学依据，还助力优化研发流程和安全策略，推动整个产业链形成健康、可信赖的安全生态环境，保障云上大模型及其应用的长期稳定与安全运行。

6．完善云上人工智能风险响应与恢复方案，筑牢安全兜底保障

一是加强模型自身的风险预防与响应。在安全设计与系统加固方面，聚焦内容安全、模型盗用及滥用等重点风险，提升对抗攻击防护能力，建立多层次防护体系，最大限度降低潜在威胁。配合持续的风险监测与快速响应手段，利用自动化处置工具和智能决策支持，实现

对模型安全隐患的及时发现和化解。

二是完善模型应用的应急响应流程。明确职责分工和协同机制，推动建立快速响应体系，确保在安全事件或业务异常时可迅速启动应急措施。通过定期开展预案演练，检验响应流程的科学性和实操性，强化业务连续性保障能力。

三是构建稳健的恢复保障体系。完善数据和系统备份机制，定期组织灾备演练，保障关键业务和信息的高可用性与完整性。确保在遭受攻击或故障后，云上人工智能系统能实现高效恢复和稳定运行。

三、云上人工智能安全发展趋势展望

（一）技术创新驱动安全升级

随着云计算与人工智能技术的深度融合，云上人工智能安全面临着全新的挑战与机遇。技术创新成为推动安全升级的核心动力，从多个维度重塑安全防护体系。

1．人工智能赋能安全实现主动智能防御，以智能技术反制风险

人工智能技术的深度应用正重塑安全防护范式，以智能化手段构建主动防御体系。

一是机器学习算法赋予安全系统强大的数据分析能力。能够实时处理海量网络数据，自动识别异常行为和潜在威胁。例如，基于深度学习的入侵检测系统突破传统规则匹配局限，可精准识别新型网络攻击，大幅提升检测准确率与效率。

二是人工智能在威胁预测领域可被广泛应用。通过整合历史数据与实时情报，对安全事件进行前瞻性预判，促使企业从被动应对转向主动防御。

三是自然语言处理技术深度赋能安全运营。能够智能解析安全日志与报告，快速定位问题根源，显著提升安全团队响应处置效能，全面增强云上人工智能系统的安全性与稳定性。

四是智能问答技术人工智能为安全服务带来全新形态。基于自然语言处理与知识图谱技术构建的智能问答系统，能够快速理解用户安全咨询需求，自动检索安全知识库，精准提供安全问题解决方案。例如，在企业安全运营场景中，员工遇到网络安全风险、数据泄露等问题时，可通过智能问答系统即时获取专业解答，无需等待安全团队人工响应，有效提升安全服务效率；同时，智能问答系统还可通过对大量用户咨询数据的学习分析，不断优化自身知识库，提前预判用户潜在安全需求，主动推送安全防护建议，实现从被动答疑到主动服务的进阶，进一步完善人工智能赋能的安全防护体系。

2．人工智能安全技术创新，构建协同发展新生态

一是深化产业链技术协同，发挥差异化技术优势。一方面发挥头部企业引领作用。头部科技企业凭借算法优化、算力调度、模型训练等核心技术优势，主导研发先进的人工智能安全技术解决方案。另一方面是激发中小企业创新活力。中小企业聚焦细分场景，基于自身专业技术积累，开发差异化的安全技术模块，双方通过技术接口对接、

数据共享与联合开发，实现优势互补，构建协同创新体系。

二是推进多技术融合创新，提升安全防护智能化水平。以融合创新催生技术升级。聚焦人工智能与物联网、大数据、区块链、边缘计算等技术的深度融合，通过跨领域技术的碰撞与整合，突破传统安全防护技术瓶颈，孵化出具备主动防御、智能识别等特性的新型安全防护技术与产品，为人工智能安全领域注入创新动能。

三是拓展多元技术应用场景，驱动产业技术创新发展。围绕金融、医疗、能源、交通、政务等领域复杂多样的安全需求，深入剖析各行业业务特性与安全痛点，定制适配不同场景的人工智能安全技术解决方案。金融领域可运用联邦学习、图计算技术实现跨机构数据协同分析；医疗行业可借助差分隐私、同态加密技术保护患者敏感数据；能源领域可利用人工智能与物联网技术保障智能电网安全；交通行业可通过车联网与人工智能技术结合筑牢智能交通系统防线；政务领域可运用区块链与人工智能技术实现数据可信共享。将这些定制化方案投入行业实践，在应用中持续验证优化技术，达成数据协同、隐私保护、系统防护等目标，以实际应用成果驱动产业规模扩张，推动人工智能安全技术迭代升级，全方位提升产业竞争力与创新力。

3．开源技术驱动人工智能安全创新与产业协同发展

开源技术凭借开放共享的特性，成为推动云上人工智能安全创新与产业协同发展的核心力量。

一是技术开源激活人工智能安全创新动能，加速安全能力升级。

技术开源打破技术壁垒，为人工智能安全领域带来显著优势。一方面，开发者能够便捷获取开源代码、算法与工具，快速掌握前沿人工智能安全技术，大幅降低企业在人工智能大模型安全研发上的门槛与成本，加速自身安全能力的迭代升级。另一方面，开源社区汇聚全球开发者智慧，形成协同创新的良好氛围，促使人工智能大模型安全解决方案不断优化迭代，加速技术创新进程，推动人工智能安全行业技术水平的提升。此外，开源技术还能促进统一技术标准与规范的形成，增强行业技术的兼容性与通用性，提升人工智能大模型安全防护的整体水平。

二是应对人工智能开源应用风险，构建人工智能安全治理体系。人工智能开源在带来机遇的同时，也伴随着代码漏洞、隐私泄露、恶意篡改等应用风险。对此，需构建完善的人工智能安全治理体系加以应对。首先，建立开源代码审查机制，通过专业团队与社区力量结合，对开源代码进行审查，及时发现并修复安全漏洞；其次，强化数据隐私保护，制定开源项目数据使用规范，明确数据收集、存储、共享的边界，运用加密、脱敏等技术保障数据安全；再者，完善开源许可证管理，规范开源技术的使用、修改与分发，避免知识产权纠纷与法律风险；最后，构建开源社区安全协作机制，鼓励开发者共同参与漏洞发现与修复，形成风险预警与快速响应体系，提升人工智能开源生态的安全性与可靠性。

三是推动开源闭源协同发展，促进人工智能安全产业发展。其一是资源互补，开源技术提供人工智能安全基础框架、核心算法与创新

思路，降低闭源人工智能研发成本与技术难度；闭源人工智能结合企业业务需求与商业目标，对开源成果进行个性化定制，实现人工智能安全技术商业价值转化。**其二是**创新互促，开源社区为闭源研发提供创新灵感与技术方向，闭源实践则为开源技术反馈实际应用中的问题与优化建议，加速人工智能安全技术迭代升级。**其三是**生态共建，开源技术通过开放共享吸引更多开发者参与，扩大人工智能安全技术生态；闭源人工智能借助商业化推广，提升技术影响力，两者共同构建起繁荣、可持续的人工智能安全产业生态，推动产业高质量发展。

（二）多方协同联动构建全链条安全防护生态

面对云人工智能安全风险复杂化趋势，多方协同治理是构建全链条防护的关键。企业、科研机构、高校与用户等主体联动，整合技术、资源与人才，覆盖研发、应用、运维全周期，提升整体安全防护能力。

1．强化企业主体协同，夯实安全治理技术根基

一是深化产业链技术协作。芯片厂商、云服务提供商、大模型服务商与多方主体紧密配合，在硬件安全、基础设施安全、应用安全等环节实现资源共享与优势互补。各方凭借在算法安全优化、硬件研发等领域的专业经验，联合提升硬件算力安全适配性，共同构建模型训练与部署的全链路安全防护，最终输出适配相应环境要求的人工智能安全防护产品。

二是推动同行业协同防御。建立常态化安全攻防演练与经验共享

机制，围绕新型安全威胁开展联合研究。行业内各方可共同搭建安全威胁情报共享平台，模拟对抗数据泄露、大语言模型恶意指令攻击等场景，同步分享防护技术和策略，通过案例复盘与技术迭代，持续提升整体防御水平。

三是加大企业自主研发投入。聚焦自身业务场景组建专业安全团队，研发定制化安全解决方案。企业需重点突破模型对抗攻击检测、数据隐私增强学习等关键技术，同时积极参与开源社区建设，推动行业安全技术创新。

2．深化产学研用联动，增强安全治理创新动能

一是聚焦前沿技术研究。科研机构与高校肩针对量子安全、对抗样本防御等人工智能安全前沿领域开展深入且系统的研究。通过大量的实验、数据分析与理论推导，不断探索新技术、新方法，为人工智能安全领域提供坚实的理论基础和充足的技术储备，以应对未来可能出现的各种安全威胁，确保人工智能技术能够安全、可靠地发展。

二是加速科研成果转化。企业作为市场的主体，可与科研院校建立联合实验室、实习基地等多元化合作机制。不仅能将科研院校的创新成果快速引入企业生产实践，还能结合企业的实际需求对科研成果进行优化和完善，使其迅速转化为实际生产力。

三是构建创新闭环。引导用户深度参与人工智能安全技术研发是推动技术进步的关键环节。在模型训练、数据处理、算法应用等实际场景中，用户可基于自身业务场景的安全需求和使用体验，及时反馈

诸如数据泄露风险、模型对抗攻击漏洞、算法偏见等安全问题。通过收集这些需求反馈和安全风险报告，研发团队能够结合安全检测、防御、审计等技术，更准确地定位人工智能系统的安全短板，进而有针对性地优化人工智能安全防护策略、升级安全技术架构，形成“研发、应用、反馈、创新”的良性循环。

3．推进企业自律与行业协同，共筑安全治理防线

一是健全企业自律机制。企业主动制定涵盖人工智能安全的内部安全规范与操作流程。在数据使用环节，重点保障训练数据的真实性、完整性和隐私合规，防范数据投毒、数据偏见等风险；在算法设计阶段，嵌入可解释性、鲁棒性技术要求，避免算法黑箱与对抗样本攻击。同时，定期开展针对人工智能系统的内部安全审计与合规检查，确保模型全生命周期的安全可控。

二是加强行业协同攻关。针对数据隐私保护、模型对抗攻击等共性安全问题，企业可联合企业、研究机构、科研院所组建研究小组，加大安全研发投入，聚焦自身业务场景研发定制化安全解决方案，共享技术资源与研究成果，联合开展技术攻关。

（三）完善标准应用筑牢产业规范发展根基

在云上人工智能产业高速发展期，完善行业标准应用是规范产业秩序、保障技术安全、推动高质量发展的核心路径。

1．健全标准体系架构实现全链条覆盖

一是聚焦核心技术领域。针对云上人工智能基础设施安全、模型安全、内容安全、数据安全、智能体安全、接口安全、平台安全、工具安全等关键环节，制定专项技术标准。标准将系统规范云上人工智能全链条安全管理，明确各环节安全技术要求，填补技术标准空白，通过建立统一规范的安全框架，有效防范潜在风险，提升人工智能产业整体安全水平，为产业健康发展筑牢技术标准防线。

二是完善应用场景标准。结合金融、医疗、交通、政务等重点行业特性，制定行业专属人工智能安全应用标准。如金融领域的人工智能金融风控标准、医疗行业的人工智能数据隐私保护相关标准，交通行业的智能交通系统数据交互与安全管理标准，政务领域的政务服务人工智能安全体系标准，确保标准贴合行业实际需求。

三是构建人工智能安全防护产品标准体系。针对人工智能内容安全产品，明确违规内容识别准确率、响应速度等性能指标，规范文本、图像、视频等多模态数据的检测流程与过滤规则，要求产品具备实时监测、自动拦截与溯源分析功能，满足内容发布平台、社交网络等场景的安全需求；对于人工智能系统安全产品，制定系统漏洞检测覆盖率、防护策略有效性等评估标准，规范入侵检测、恶意代码防护、系统加固等功能模块的技术要求，保障人工智能运行环境的稳定性与安全性；针对人工智能防火墙类新兴安全防护产品，规范基于人工智能算法的威胁识别、访问控制、安全策略动态调整等功能标准，有效抵御网络攻击与数据泄露风险。通过明确各类产品的功能要求、性能指

标和检测方法，推动人工智能安全防护产品的规范化、标准化发展，提升产品质量和可靠性。

2．强化标准推广应用提升产业执行效能

一是加强标准宣贯。开展人工智能安全标准解读与培训相关活动，帮助企业准确理解标准内涵与实施要点，快速掌握安全防护技术标准能力要求，提升标准应用能力，减少因标准理解偏差导致的安全漏洞，从源头降低安全风险，为产业安全发展提供人才支撑。

二是建立评估验证机制。对标准应用效果进行量化评估与验证。通过第三方检验检测机构对企业进行人工智能安全相关评估验证，开展标准试点、测试评估和定期审查，推动形成覆盖行业差异化场景的安全合规体系，提升重点行业人工智能系统的安全水平，加快安全技术成果的产业化应用进程。

中国信通院可信安全团队持续推进云上人工智能安全相关标准建设与评估检验工作，充分发挥在云计算智能化领域的综合优势和行业积淀，积极牵头和参与国内外云上人工智能安全标准的编纂工作，推动形成相关标准体系。围绕行业实际应用场景，通过开展基础设施安全、模型安全、内容安全等领域的标准试点评估、技术能力检验和安全性能测试等工作，加快推动云上人工智能安全技术的落地应用与产业化进程。

（四）多层次的人工智能安全治理体系建设

1．深化产业协作，凝聚多方安全治理合力

一是压实企业主体责任。引导企业建立健全内部人工智能安全管理制度，加大安全研发投入，组建专业安全团队，对内部系统全生命周期进行安全管控，从源头降低安全风险。**二是发挥行业组织作用。**充分发挥行业协会、联盟的桥梁纽带功能，搭建资源共享、经验交流、技术合作平台，组织企业联合开展安全技术攻关与最佳实践推广，提升行业整体安全防护水平。**三是推动产学研用融合。**构建企业提需求、高校育人才、科研机构做支撑的协同创新模式，围绕实际应用场景中的安全痛点，共同探索解决方案，实现多方资源共享、优势互补，推动产业安全协同发展。

2．健全风险防控，提升应急处置能力

一是建立风险威胁信息共享机制。整合企业、行业组织等多方资源，搭建人工智能安全风险信息共享平台，实现安全漏洞、攻击手段、防御策略等信息的及时互通与协同分析。**二是完善应急处置机制。**制定标准化的应急响应流程与预案，明确各方在风险事件中的职责分工，定期组织应急演练，提升对安全事件的快速响应与处置能力。**三是强化人才实战培养。**通过模拟攻防演练、应急案例研讨等方式，提升安全人才的风险识别、应急处理与协同作战能力，为风险防控工作提供坚实的人力保障。