人工智能安全治理白皮书（2025）

中国联合网络通信有限公司

华为技术有限公司

北京百度网讯科技有限公司

三六零数字安全科技集团有限公司

超聚变数字技术有限公司

奇安信科技集团股份有限公司

恒安嘉新（北京）科技股份公司

北京汽车研究总院有限公司

浙江大学

北京邮电大学

南开大学

互联网教育智能技术及应用国家工程研究中心

2025 年 7 月

人工智能产业链联盟

星主： AI产业链盟主

知识星球

微信扫描预览星球详情

版权声明

目录

前言

人工智能（Artificial Intelligence，简称 AI）技术的飞速发展给人类和经济社会的发展带来了翻天覆地的变化，是驱动第四次工业革命和经济社会数字化转型的先进生产力。近年来，从智能体应用到具身智能再到智能物联网，从智慧车间到智能驾驶再到自动化医疗诊断，人工智能技术正在重塑全球产业格局。然而，在人工智能技术跃迁的背后，潜藏着数据、模型、基础设施与应用的多重安全风险。随着各行各业纷纷布局人工智能模型应用，这些风险正从理论推演演变为现实威胁，迫使全球各国、组织和企业重新审视对人工智能技术的安全治理。

面对人工智能安全治理的新态势，中国联通联合多家单位，在《人工智能内生安全白皮书（2024）》的基础上发布《人工智能安全治理白皮书（2025）》，在动态演化的技术生态中主动升级治理体系。

本白皮书以建立安全可靠、公平可信、智能向善的人工智能系统为目标，围绕人工智能安全治理，重点介绍了人工智能安全治理风险，构建了人工智能安全治理体系，明确了人工智能安全治理的要求和目标，归纳了人工智能安全治理技术体系的建设要点。在人工智能安全治理框架下，本白皮书从 AI 基础设施、数据、模型、应用等多个维度提供了解决方案和案例，并从 AI 安全治理法律法规、标准体系建设、AI 安全前沿技术探索和 AI 安全人才培养与产学研协同创新等多个维度提出人工智能安全治理发展建议，推动我国人工智能技术健康、安全、可持续发展。

本白皮书由中国联通研究院主笔，华为技术有限公司、北京百度网讯科技有限公司、三六零数字安全科技集团有限公司、超聚变数字技术有限公司、奇安信科技集团股份有限公司、恒安嘉新（北京）科技股份公司、北京汽车研究总院有限公司、浙江大学、北京邮电大学、南开大学、互联网教育智能技术及应用国家工程研究中心联合编写。

编写组成员（排名不分先后）：

总策划：谢攀、叶晓煜、郑涛、徐雷、陶冶、吴琮、赵平、于天水、李加赞、冯运波、李志伟、邹荣新、姜伟生、杨满智、任奎、秦湛、张熙、刘哲理

编委会：阿曼太、安宏亮、陈泱、陈小华、陈晓光、程莉莉、董航、董玉强、高华、管铭、韩文峰、华佳烽、贾雅清、姜福利、李慧芳、李兴成、李朝卓、梁晨、刘兆峰、刘钢、刘东、卢宇荣、马一男、祁彬斌、尚程、尚煜茗、盛杰成、孙世丁、孙浩文、孙学磊、孙忠阁、孙佳、孙禹、唐文、滕义、王庆龙、王一、许小兵、徐积森、杨必琨、杨臻、张钊、张先鹏、张景龙、张小梅、张越威

一 AI 概述

1.1 AI 技术发展历程

AI 技术历经从符号规则、机器学习（Machine Learning，简称 ML）再到深度学习（Deep Learning，简称 DL）的持续演进，其发展技术脉络在 2024 年发布的《人工智能内生安全白皮书（2024）》中已有系统梳理。近年来，AI 技术和应用更是呈现出爆发式增长态势，尤其是以大模型（Large Language Model，简称 LLM）为核心的新一代人工智能技术，已成为推动科技进步、产业升级的关键性技术之一。自 OpenAI 发布 ChatGPT 大模型之后，国内外各大公司均开始拓展大模型在商业化场景中的应用。在国际上，OpenAI（GPT 系列）、Google（Gemini 系列）、Anthropic（Claude 系列）、Meta（Llama 系列）、Microsoft（Phi 系列）以及 xAI（Grok 系列）等为代表的科技公司相继发布与迭代更新了一系列自研大模型，在语义理解与生成、图像视频生成、代码生成等方面展现出了优秀的能力。

国内大模型领域也呈现出蓬勃的发展态势，众多大模型公司在技术研发和应用拓展方面取得了显著进展，通过不断创新和优化，推动了大模型在多个行业的广泛应用，为人工智能技术的发展和产业智能化升级做出了重要贡献。国内以百度（文心一言系列）、阿里巴巴（通义千问系列）、腾讯（混元系列）、字节跳动（豆包系列）、华为（盘古系列）等为代表的大型公司发布了自家研发的大语言模型，凭借其所积累的大规模用户数量与广泛的内外部业务场景，广泛应用于智能办公、电商直播等多个行业。与此同时，杭州深度求索人工智能基础

技术研究有限公司开发的大模型 DeepSeek 备受关注。DeepSeek 模型性能比肩国际领先水平，训练成本极低，同时采用开源策略，极大地推动了 AI 技术的普及。更重要的是，DeepSeek 的崛起对整个 AI 行业格局产生了深远影响，为未来 AI 技术发展提供了新思路和巨大潜力。

未来，大模型技术将持续深化发展。在技术架构层面，通过并行计算、软硬件协同等技术支撑，实现计算效率的跃升。在应用拓展层面，大模型服务将走向多领域应用，具备更强的泛化能力和自我进化能力。同时，随着各行各业纷纷布局大模型应用服务，安全性、可靠性、可控性也将成为大模型能力发展的重要考量。

1.2 AI 技术应用发展趋势

1.2.1 AI 智能体

2025 年，AI Agent 产业正处于加速发展的关键时期，从产业链来看，产业上游由算力提供商、数据供应商以及模型开发商主导；产业中游聚焦于 AI 智能体的集成和平台化服务；产业下游则围绕行业垂直应用和通用智能体的开发与推广，逐渐呈现多样化的发展趋势。AI Agent 的关键技术包括模型推理、记忆存储、工具执行和状态管理，其所展现出的自主性和环境适应性，能够透彻理解和快速响应用户的需求，提供个性化的服务和建议。AI Agent 的技术演进与行业应用正加速融合，呈现出多模态交互升级、多 Agent 协同工作、垂直领域 Agent 专业化的趋势。

1.2.2 具身智能

当前，具身智能在大模型技术的推动下成为科技产业的热点。具身智能的关键技术包括机器人 AI 芯片的研发、高性能仿生多指灵巧手的研制、具身智能基座模型的构建，以及具身智能本体的控制。未来，具身智能机器人将广泛应用于工业制造、家庭服务和商业服务等多个领域，有望解决工厂劳动力短缺、社会老龄化等问题，把人类劳工从枯燥、高危劳动中解放出来，进一步推动全社会生产关系的改变与重塑。

1.2.3 端侧 AI

近年来，端侧 AI 技术取得了显著进展，终端设备的算力大幅提升以及多模态大模型、文生视频等需求的增长推动了端侧 AI 硬件能效的持续提升，百亿参数大模型有望端侧落地，端云协同的混合 AI 架构逐渐成为主流，越来越多的企业和机构开始布局端侧 AI 市场。端侧 AI 的关键技术包括高性能芯片模组的研发、模型压缩与优化和异构计算支持。端侧 AI 算力将持续泛化普及，从手机、PC 等主流终端设备向更多产品渗透，如智能穿戴设备、智能家居产品以及新型终端（人形机器人等），AI 终端的生产力化将加速价值变现甚至赋能工业变革。

1.3 AI 安全治理体系

针对人工智能面临的安全风险与挑战，本章提出了人工智能安全

治理体系，明确了人工智能安全治理的目标，归纳了技术体系和监督与管理体系的建设要点，如图 1.1 所示。

图 1.1 人工智能安全治理体系

在监督与管理方面，在人工智能技术快速走向应用落地的背景下，形成“法规政策、标准规范、管理制度、保障措施”四位一体的监督与管理体系，有力推动人工智能技术朝着安全、可靠、可控的方向发展。在技术体系方面，AI 基础设施安全、AI 模型算法安全、AI 数据安全以及 AI 应用安全是确保整个 AI 系统安全性的四个关键方面，它们之间有着紧密的联系，并且相互促进。任何一个环节出现问题都可

能导致整个 AI 系统的脆弱性增加，只有当所有层面都得到充分重视并采取有效措施时，才能构建起一个真正安全可靠的 AI 生态系统。

1.3.1 AI 基础设施安全

AI 基础设施为模型训练和推理提供计算、存储等服务，确保其安全、稳定、高效地运行。AI 基础设施安全是模型安全、数据安全和应用安全的基础。2023 年 10 月，工业和信息化部等六部门印发《算力基础设施高质量发展行动计划》，旨在推进算力基础设施高质量发展，充分发挥算力对数字经济的驱动作用。2023 年 12 月，国家发展改革委等五部门联合印发的《关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见》，提出加快构建联网调度、普惠易用、绿色安全的全国一体化算力网。在 AI 基础设施安全方面，需要针对智算硬件设备、智算云、智算 MaaS 平台、智算算力网络构建防护能力。

1.3.2 AI 数据安全

数据是人工智能技术发展的基础资源和重要驱动力。2021 年是我国数据安全立法元年，我国正式颁布了《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》。2023 年 10 月，国家数据局正式挂牌成立，负责协调推进数据基础制度建设。2025 年 1 月 1 日起《网络数据安全管理条例》施行，对网络数据处理活动进行了更详细的规定，加强了网络数据安全保护。

在 AI 时代，大模型技术迅猛发展，AI 数据安全保护体系亟待建立与完善。一方面，需要构建通用基础的数据安全与隐私保护能力。另一方面，需针对大模型数据全生命周期建立起完善覆盖训练数据、微调数据、推理数据以及知识库数据等全流程的 AI 数据安全防护体系，充分保障大模型数据全生命周期的安全。

1.3.3 AI 模型算法安全

人工智能模型算法是人工智能的核心，保障人工智能模型算法的安全，是推动 AI 技术进一步赋能产业的必然要求。2025 年 1 月，为完善人工智能安全标准体系建设，全国网络安全标准化技术委员会组织编制了《人工智能安全标准体系（V1.0）》。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《人工智能生成合成内容标识办法》，自 2025 年 9 月 1 日起施行，旨在促进人工智能健康发展，规范人工智能生成合成内容标识，保护公民、法人和其他组织合法权益，维护社会公共利益。

1.3.4 AI 应用安全

AI 技术在医疗诊断、编程辅助、自动驾驶和智能物联网等领域的广泛应用极大地改变了人们的生活方式与工业生产方式。然而，这一进步也带来了显著的安全挑战。AI 应用的开发者和使用者需要采取一系列措施来保障 AI 系统的安全性，如加强数据保护、建立有效的监管框架和技术标准。这不仅有助于保护用户的利益，也有助于促

进 AI 技术的可持续发展、提升社会接受度，为各行业的创新提供坚实的基础，推动社会向着更加智能化的方向前进。

二 AI 安全治理风险分析与挑战

AI 技术在推动千行百业实现数字化转型的同时也带来了新的风险与挑战。人工智能安全治理风险既包括人工智能自身存在的脆弱性，也包括人工智能应用运行时面对的外部威胁。具体而言，人工智能安全治理风险可以进一步分为 AI 基础设施安全风险、AI 数据安全风险、AI 模型算法安全风险以及 AI 应用安全风险，如图 2.1 所示。

图 2.1 AI 安全风险

2.1 AI 基础设施安全风险

AI 基础设施安全风险主要包含智算硬件设备安全风险、智算云安全风险、智算 MaaS 平台安全风险以及智算算力网络安全风险。

2.1.1 智算硬件设备安全风险

智算硬件设备包括服务器、智算卡等设备，其面临的安全风险主要是由于硬件设备的设计、部署、使用不当所引发的多维度攻击，包括：攻击者通过非授权接触设备本体实施入侵的物理攻击，如打开机箱使用探针读取总线数据等；在不打开设备前提下，通过外部接口发起的硬件接口攻击，如通过一些面板调试端口篡改系统软固件，绕过设备安全机制或篡改产品配置；利用设备软固件的安全漏洞发起的软件攻击，攻击者可以通过探测、扫描等方式获知并利用这些漏洞发起攻击。此外，针对智算硬件设备的攻击方式还包括侧信道攻击、故障注入攻击等。

2.1.2 智算云安全风险

云底座作为智能计算的基石，其安全性直接影响到整个系统的稳定性和数据的保密性。特别是云底座的存储和基础服务，是数据泄露的主要目标，这对于使用庞大且通常敏感数据集的 AI 计算而言尤为重要。同时，云底座中未经安全管理的 API 也可能会被攻击者利用来绕过安全控制并获得对数据和服务的未经授权的访问。此外，内部威

胁是更难检测的安全威胁之一，授权用户可能会利用其访问权限进行恶意活动。

智算云安全还可以进一步分为云操作系统安全和容器镜像安全。云操作系统面临安全漏洞、错误配置、访问控制不足、缺乏可见性与监控等风险。运行带有已知但尚未修补的漏洞的云操作系统会为攻击者提供入侵的切入点。而容器镜像作为智能计算应用程序的部署单元，其安全性同样重要。容器镜像通常基于基础镜像构建并包含第三方库，如果这些组件过时或不安全，则可能继承已知的漏洞。使用被恶意软件感染或来自不可信来源的容器镜像存在遭受供应链攻击的风险。

2.1.3 智算 MaaS 平台安全风险

智算 MaaS（Model as a Service，简称 MaaS）平台是一种以 AI 模型为核心的服务模式，旨在为用户提供更便捷、更可靠、更高效的 AI 模型开发、部署和管理服务。通过模块化设计（包括模型托管与部署、模型训练与优化、API 服务与调用、资源管理与调度、模型市场与社区等）和 AI 应用的全生命周期支持，智算 MaaS 平台降低了 AI 开发门槛，提升了模型开发和部署效率，广泛适用于图像识别、自然语言处理、推荐系统等场景。

智算 MaaS 平台存在以下安全风险：（1）模型知识产权风险，攻击者通过逆向工程或权限滥用提取模型参数（如嵌入层权重），复现或篡改 AI 模型。例如，某开源模型托管平台因权限控制漏洞，导致企业定制模型被第三方非法商用。（2）拒绝服务攻击风险，攻击者

针对模型 API 接口发起流量攻击，导致服务中断。（3）资源滥用风险，恶意用户通过批量调用低功耗模型占用 GPU 资源，导致平台算力成本激增。

2.1.4 智算算力网络安全风险

智算算力网络是一种以算力为核心资源的新型基础设施，旨在通过网络将分散的算力资源进行高效整合和调度，满足人工智能应用对大规模计算的需求。算力网络不仅提升了算力的利用效率，还降低了使用门槛，推动了 AI 技术的普及和创新。由于算力网络涉及多源、泛在算力节点，无法保证每个节点都能做到安全可靠，同时数据分散到多方算力节点进行计算，会导致受攻击风险增加。具体而言，智算算力网络存在以下安全风险：

（1）算力编排管理安全风险：在基础设施层面，算力网络具有算力泛在、灵活接入等特点，算力的动态调度对跨系统、跨域甚至跨境的多场景点对点网络连接机制提出了新的需求，也为攻击者提供了更多的攻击路径。同时，相较于传统网络架构，算网新型架构新增了网元如算网感知单元和算网控制单元等，这些网元的引入将导致算网全网安全的管理复杂度提升，安全风险也随之增加。算网信息在编排管理层汇聚，算力信息的正确性、完整性、安全性将影响算力网络正常编排调度服务的开展，一旦节点被攻击或仿冒，造成虚假算力信息上传，将严重影响算网的可靠性。

（2）算力运营服务安全风险：算力服务是端到端服务，用户群

体庞大，分布式资源节点数量较多，数据信息管理起来较为繁杂，因此存证溯源的复杂度加大，出现安全问题时难以快速定位安全威胁源。此外，攻击者或网络中的恶意节点还可能发起数据窃取等网络攻击行为，对服务稳定性、数据安全性等造成威胁。

2.2 AI 数据安全风险

AI 数据安全风险可以分为：通用数据安全风险与 AI 数据生命周期风险。通用数据安全风险通常包括：数据合规、数据泄露、数据篡改等安全风险，聚焦于通用、共性数据安全风险；AI 数据生命周期安全风险分为训练数据安全风险、微调数据安全风险、推理数据安全风险和知识库数据安全风险，聚焦于 AI 模型训练与推理所面临的数据安全风险。

2.2.1 通用数据安全风险

通用数据安全风险涉及多个方面，包括数据合规、数据泄露、数据篡改及数据质量等核心领域。数据合规风险主要关注在处理个人信息时是否遵守相关法律法规，如未遵循《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》等法规要求，可能导致法律制裁及信誉损失。数据泄露风险是指未经授权的个体通过网络攻击、内部失误或第三方服务提供商的安全漏洞获取敏感信息，进而导致隐私泄露和经济损失。数据篡改则是指未经授权修改数据的行为，破坏了数据的完整性，可能导致决策错误或财务损失。数据质量问题同样不容

忽视，它涉及到数据的准确性、完整性和一致性，低质量的数据会影响训练 AI 模型的性能，导致 AI 模型做出错误决策。

2.2.2 训练数据安全风险

AI 训练数据的安全风险主要涉及数据来源、内容安全、数据中毒、数据质量等多个层面。首先，数据可能因违规获取或存储传输环节的漏洞导致泄露，攻击者可通过逆向工程从模型输出中推断敏感信息、隐私和商业机密。其次，训练数据中可能包含违法、歧视性内容或侵犯知识产权的信息，导致模型输出有害或误导性结果，例如虚假信息、偏见或违反社会主义核心价值观的内容。此外，训练阶段面临的数据风险还包括数据投毒，攻击者在数据收集或预处理阶段注入恶意样本以篡改模型行为，使其在特定场景下产生错误输出或后门漏洞。同时，数据质量低下（如噪声或过时信息）也会削弱模型鲁棒性继而造成模型效果不佳。这些训练数据安全风险不仅威胁模型性能，还可能引发法律合规问题，需通过分类分级管理、加密监控及技术防护措施加以应对。

2.2.3 微调数据安全风险

大模型微调是利用特定领域的数据集对已训练的大模型进行进一步的训练过程，旨在进一步优化模型在特定任务的性能，使模型能够更好地适应和完成特定领域的任务。微调适合垂直领域任务增强和任务特定化的场景，如增强通用大模型在医疗、法律、金融等领域的

特定知识，或者针对问答系统、文本分类、命名实体识别等场景任务进行优化。由于微调可能基于私有而非公开数据进行模型微调，若存储或传输环节存在漏洞（如未加密、权限管理不当），攻击者可通过逆向工程从模型输出中反推训练数据，因此存在隐私泄露的风险。

2.2.4 推理数据安全风险

随着大模型的产业落地和规模化应用，其所承载的业务价值越来越大，大模型将面临各种安全攻击的威胁。在推理环境下，需要综合考虑高价值模型（采用闭源模型或微调的专业模型）、关键的行业数据等安全需求。在未受保护的推理应用环境中，攻击者可能是恶意的系统运维人员，或者通过漏洞、恶意软件等取得较高访问权限的黑客。攻击者利用已获得的服务器访问权限，可以绕过存储态加密保护，直接从内存中导出模型，从而实现模型窃取。同时，攻击者还可在服务器端 API 等关键点插入恶意代码，进一步实现对用户隐私推理数据的窃取。

2.2.5 知识库数据安全风险

基于检索增强生成（Retrieval-Augmented Generation，简称 RAG）的 AI 应用通常采用“前端+业务应用+模型后台+RAG 后台”的技术架构，其知识库数据面临着多种安全风险，包括：（1）外部攻击者可能会利用网络漏洞，通过 SQL 注入、恶意软件攻击等手段，非法获取数据库中的敏感信息如用户数据、商业机密等；（2）内部授

权用户可能会因误操作、恶意行为或受到社会工程学攻击而导致数据泄露。例如，内部人员可能会将敏感数据发送给未经授权的人员，或者在不安全的环境中处理数据，使得数据容易被窃取；（3）数据库服务器可能会受到恶意软件的感染，如病毒、木马等，这些恶意软件可能会窃取数据、破坏系统或为攻击者提供远程控制权限。

2.3 AI 模型算法安全风险

AI 模型算法的安全管理与防护至关重要。从 AI 模型算法生命周期的角度可以将 AI 模型算法面临的风险分为：模型训练风险、模型微调风险、模型推理风险和模型部署风险。此外，依据 AI 模型的性质，可以将模型分为决策式 AI 模型与生成式 AI 模型，因此还可以将 AI 模型算法安全风险分为通用 AI 模型安全风险与生成式 AI 模型安全风险。

2.3.1 AI 模型算法生命周期安全风险

（1）模型训练风险

模型训练风险主要涉及数据隐私、版权侵犯、数据偏见以及对抗攻击等。在训练 AI 模型时，需要大量的数据集，这些数据可能包含个人隐私信息或受版权保护的内容，如果未经适当处理或授权使用，可能会导致法律纠纷。此外，训练数据中的偏见可能导致模型产生有偏见的预测结果，影响公平性和准确性。同时，模型还可能遭受对抗攻击，即通过精心设计的数据输入来误导模型，使其输出错误的结果。

（2）模型微调风险

模型微调通常是为了让预训练模型更好地适应特定的任务或领域。然而，这种过程带来了新的风险，例如：微调过程中使用的特定领域数据可能含有敏感信息，若不加以妥善保护，可能会造成数据泄露。同时，研究表明微调后的模型更容易受到越狱攻击，即通过特殊指令使模型偏离其原本设定的安全限制，执行未授权的行为。此外，不当的微调策略还可能导致模型性能下降，增加大模型产生幻觉现象的发生率。

（3）模型推理风险

在模型推理阶段，也面临多种风险。例如：对抗攻击同样可以在推理阶段发生，攻击者可以通过输入经过轻微扰动的数据样本，诱使模型做出错误的决策；由于模型训练所用的数据和环境与实际应用场景存在差异，可能导致模型在新环境中表现不佳，出现“数据漂移”现象；此外，模型可能会遇到未曾见过的新数据，从而导致预测失败或产生误导性结论。

（4）模型部署风险

模型部署到生产环境后，会面临一系列挑战和风险。首先是模型窃取问题，包括未经授权访问模型及其数据、传输过程中数据泄露等。其次是模型维护和更新的风险，随着时间推移，模型可能需要根据最新数据进行调整以保持其有效性，但这同时也增加了引入新漏洞的可能性。再者，模型的可解释性和透明度也是一个重要考量点，尤其是在金融、医疗等领域，用户和服务提供者都需要理解模型决策背后的

逻辑以防范业务疏漏。最后，模型部署还需要考虑合规性要求，确保遵守相关的法律法规，避免法律风险。

2.3.2 通用 AI 模型算法安全风险

（1）模型鲁棒性弱

AI 模型的鲁棒性弱是指模型在面对数据中的噪声、干扰、异常值或环境变化等不确定因素时，未能保持良好的性能和稳定性。增强模型的鲁棒性对于提升系统稳定性和安全性至关重要，尤其在自动驾驶、金融交易等高风险领域。鲁棒性与模型的泛化能力密切相关，鲁棒性强的模型通常具有更强的泛化能力，能够在未见过的数据上保持良好的性能，确保 AI 系统在复杂环境下的稳定性和可靠性。

（2）模型泛化性差

AI 模型的泛化性差风险是指模型从训练数据中学到的知识，未能有效地应用到未见过的数据上。过拟合是泛化性风险的一个典型例子，当模型过于复杂以至于学习到了训练数据中的噪声和细节而不是其背后的模式时，就会发生这种情况。解决这一问题的方法包括使用更多的训练数据、采用正则化技术以及进行交叉验证以评估模型在不同子集上的表现。此外，通过设计更加简单的模型结构也可以有助于减少泛化误差。

（3）模型可解释性差

AI 模型的可解释性风险涉及到模型决策过程的透明度问题。对于深度学习模型而言，由于其复杂的内部结构，理解它们如何做出决

策变得非常困难。这种“黑盒”特性不仅影响了用户的信任度，也可能阻碍了对潜在问题的诊断与修正。

（4）模型偏见与歧视风险

AI 模型的偏见与歧视风险是指训练数据中对某类群体或个体存在偏见与歧视的现象，从而导致训练的模型存在偏见，无法保证决策的公平性。例如，如果面部识别系统中的 AI 模型主要利用白人面孔的数据集进行训练，则该系统可能在识别有色人种时准确性较低。为减轻此类风险，必须仔细选择和处理训练数据，并采用公平算法设计原则。

（5）模型逆向工程风险

AI 模型逆向工程风险是指攻击者通过模型输出来推断其内部结构或训练数据的过程，从而利用这些信息进行对抗攻击或者模型复制。为了防范这种风险，可以采取诸如差分隐私、密码学等技术，确保即使是在查询模型的过程中也能够保护敏感信息的安全。

（6）模型对抗攻击风险

对抗攻击风险是指通过向 AI 模型的输入添加细微且不易察觉的扰动，导致模型输出错误结果的安全威胁。对抗攻击利用了深度学习模型对特定输入变化的高度敏感性，即使这些变化对于人类来说几乎不可感知。对抗攻击的危害包括 AI 系统性能下降、隐私泄露、关键决策失误以及可能引发的信任危机和社会影响，这些问题在医疗诊断、自动驾驶和金融交易等依赖 AI 准确性的领域尤为严重。因此，确保 AI 系统的安全性以抵御此类攻击至关重要。

2.3.3 生成式 AI 模型算法安全风险

生成式 AI 模型安全风险主要包括大模型用户不当输入风险和大模型生成内容合规风险两大类。

（1）大模型用户不当输入风险

1）提示词攻击风险：大模型在推理阶段可能面临恶意提示词注入攻击的风险。攻击者通过提示词工程技术，与大模型工具模块进行异常交互，从而非法调用大模型后台工具；或者迫使大模型脱离其内在安全机理，引发违规内容生成、敏感数据泄露及信息篡改等安全问题。典型的提示词攻击手法包括：角色扮演攻击、目标劫持攻击、反向诱导攻击、上下文操纵攻击等。

2）大模型接口攻击/频率突破：API 交互是大模型应用服务重要的呈现范式，因此面临接口攻击或者频率突破等网络安全威胁。例如恶意用户通过对大模型应用 API 发起大量请求，导致接口崩溃或服务中断。

（2）大模型生成内容合规风险

1）侵犯他人合法权益：在生成式人工智能服务过程中，前端交互操作、数据收集处理或内容生成反馈等任一环节若未经规范管理，就可能对他人合法权益造成侵害。例如，在训练多模态大模型时，所使用的图像、视频等内容若未经授权，就可能侵犯他人的肖像权。此外，生成式人工智能在训练阶段若未经许可使用他人作品，也可能构成对著作权的侵权。在内容生成阶段，若生成的内容与已有作品高度相似，或使用了受保护的作品片段，也可能构成对原作品的侵权。

2）违反核心价值观：大模型生成的内容可能会涉及到一系列敏感问题，可能涉及意识形态类内容，如特定的政治立场、宗教信仰等，这些都可能引发社会争议和冲突，并对社会稳定与安全构成威胁；同时，也可能产生涉及黄色、暴力、恐怖主义以及毒品等违法犯罪类不良信息，严重损害用户身心健康并对社会造成负面影响。此外，大模型还可能无意间生成含有性别、种族、年龄、职业或地域歧视的言论或观点，不仅会对特定群体造成伤害和不适，还可能加剧已有的社会矛盾和冲突。因此，确保大模型生成内容的安全性、合法性和伦理性成为当前技术发展中的重要挑战。

3）模型幻觉：大模型在推理阶段可能会遭遇模型幻觉问题，导致其输出具有不确定性和不可预测性。模型幻觉主要分为两类：首先是事实性幻觉，指的是当大模型生成涉及事实性知识的内容时，可能会捏造或错误解释某些概念、事实或数据，从而向用户提供错误信息；其次是忠实性幻觉，表现为生成内容与用户指令或提供的上下文存在偏差，或是生成内容内部自相矛盾，比如回答与用户提问无关；此外，在多模态大模型处理图像和文本等类型输入时，生成的内容还可能与实际输入在逻辑、内容或跨模态关联上出现不一致或矛盾。

4）思维链（Chain of Thought，简称 CoT）安全推理风险：一种典型的思维链攻击方式是攻击者首先通过诱导模型对高危请求进行拒绝，获取其内部安全审查的完整推理链，进而分析模型的安全策略。然后，攻击者针对性地修改模型输出的思维链内容，如将拒绝理由替换为符合安全策略的陈述，或直接跳过安全论证阶段，从而引导

模型误判请求的正当性、绕过安全防线。

2.4 AI 应用安全风险

AI 应用所面临的安全风险主要包括：AI 模型算法滥用风险、AI 应用开发安全风险以及 AI 垂直行业应用风险。

2.4.1 AI 模型算法滥用风险

（1）虚假有害信息引发舆情风险

大模型生成的内容可能会被用于传播虚假信息，进行误导公众、操纵舆论或欺诈活动。

（2）多模态深度伪造风险

多模态深度伪造风险是通过融合视频、音频、文本等多种模态数据生成高度逼真的虚假内容，进而从事非法活动。此类攻击可能造成直接经济损失，并破坏社会信任体系。

（3）模型透明性不足风险

在 AI 与用户交互的过程中，透明性不足的问题正变得愈发显著。一方面，由于 AI 技术的局限性，往往未能对决策结果或生成内容进行充分的告知与解释，导致用户难以理解其决策机制以及潜在的问题，这无疑增加了使用风险。另一方面，在 AI 应用的运行过程中，有时缺乏明确的身份提示，用户可能会误将其当作人类或具有高度智能的实体，从而在交互过程中产生过度依赖。同时，部分 AI 系统前端交互的“拟人化”设计，虽然在一定程度上提升了用户体验，但也可能

进一步强化用户的信任感。在情感支持、心理咨询等敏感领域，这种过度信任可能导致用户在情感和心理上过度依赖 AI，而忽视了其作为科技工具的本质，进而引发一系列潜在伦理道德问题。

2.4.2 AI 应用开发安全风险

（1）端侧 AI 安全风险

端侧设备受限于低功耗、小内存和有限算力，需对 AI 模型参数量进行压缩和优化，然而此类优化可能会牺牲模型鲁棒性和安全性，出现“安全税”的现象，从而造成模型性能的下降。同时，端侧部署往往意味着需在设备端实现实时推理，并依赖云边协同架构进行模型更新和任务调度。这要求分布式计算框架具备动态负载均衡能力，但异构硬件的兼容性问题以及因网络延迟或中断导致的任务切分与调度失误都可能引发系统失效。

（2）智能体安全风险

AI 智能体是一个利用大模型作为其核心计算引擎的人工智能系统，可进行一定程度的自主行为以完成超出文本或图像生成的复杂任务。AI 智能体除了继承大模型固有的安全风险，还有因其与外界环境不断进行交互产生的特定安全风险：一方面，模型上下文协议（Model Context Protocol，简称 MCP）、代理对代理协议（Agent to Agent Protocol，简称 A2A）等通过标准化接口允许智能体和第三方工具间通信协作，但这类协议本身可能存在设计缺陷使攻击者可篡改工具描述，植入隐藏指令或后门；另一方面，智能体具备自主规划、

分解任务和长期运行的能力，这种自主决策链路的不可预测性导致其行为决策可能因环境反馈或对抗攻击逐步偏离原始目标。此外，智能体决策辅助所需的知识、规则、记忆等也存在漂移或被篡改后导致业务决策错误风险。

（3）具身智能安全风险

具身智能在融入现实世界的过程中面临多重安全风险，其安全隐患主要体现在数据隐私、物理安全、系统漏洞及法律伦理等方面。首先，通过摄像头、麦克风等传感器设备收集的个人信息（如生物特征、行为习惯等）存在未授权收集、泄露、窃取或滥用的风险。其次，由于具身智能体具备物理行动能力，管控平台的漏洞和管理脆弱性可能被攻击者利用，导致其动力系统和执行器被恶意攻击或算法决策失误，将会引发人身伤害或财产损失，例如服务机器人操作失控、自动驾驶事故等。此外，相关的法律与伦理规范尚未完善，人机伦理冲突时的责任归属问题仍未解决。这些风险相互交织，需要从技术防御、法律约束、伦理治理等多维度协同应对，以确保具身智能的可控性和安全性。

（4）智能物联网安全风险

智能物联网（Artificial Intelligence of Things，简称 AIoT）的风险核心在于其融合了 AI 的算法脆弱性与 IoT 的物理暴露性，需要在在资源受限的边缘环境中实现复杂协同。相较于通用 AI 系统，AIoT 设备部署在物理边缘场景，需应对传感器噪声干扰、物理攻击以及复杂环境干扰对数据完整性的影响。而相较于通用 IoT，AIoT 不

仅面临传统 IoT 的数据泄露风险，还需应对 AI 特有的威胁，如对抗样本攻击、训练数据投毒以及模型窃取。

2.4.3 AI 垂直行业应用风险

（1）AI+医疗

人工智能在医疗行业的应用带来革新的同时也引入了技术和伦理方面的多重风险。例如，手术机器人依赖算法控制机械臂的精准度，但训练数据偏差或系统漏洞可能导致手术中出现错误操作。医学图像识别算法也可能因数据质量问题（如样本分布不均或标注错误）产生漏诊或误判。此外，由于医学图像和手术数据包含患者敏感信息，AI 系统若缺乏加密措施或权限管理，可能被非法获取导致患者隐私被侵犯。

（2）AI+新闻

大模型生成的内容可能会被用于传播虚假信息以误导公众、操纵舆论或进行欺诈活动。例如，一些不法分子可能会利用大模型生成虚假的新闻报道、虚假证据或虚假视频，以制造恐慌、引发混乱。

（3）AI+金融

金融领域一般通过图像、声音、文本等多维度信息对用户进行身份验证。攻击者可利用多模态深度伪造技术伪造人脸、声纹和行为特征，通过金融系统的多模态身份核验，实施盗刷、恶意注册等欺诈行为。

（4）AI+编程

AI 辅助编程在提升开发效率的同时，也潜藏着多重风险。首先，AI 生成的代码可能存在安全隐患，例如跨站脚本攻击（XSS）、SQL 注入等常见漏洞。其次，AI 生成的代码往往缺乏整体架构设计，导致维护困难。此外，用户可能因信任 AI 而减少人工审查，进一步放大安全风险。

三 AI 安全治理体系现状

3.1 AI 安全治理法律法规政策现状

在人工智能技术蓬勃发展的时代背景下，全球各国和组织制定了一系列政策，以引导人工智能技术安全发展。

（1）欧盟

2024 年 3 月，欧洲议会通过了《人工智能法案》，2024 年 8 月 1 日正式生效。这是欧盟范围内针对人工智能治理的统一的监管和法律框架，也是全球首部全面监管 AI 的法规。该法案对人工智能系统风险进行了分类，并提出通过识别人工智能风险类别来制定监管制度。该法案提出严格禁止“对人类安全造成不可接受风险的人工智能系统；同时，该法案要求人工智能公司需要对其算法模型进行人为控制，为“高风险”应用建立风险管理系统。总的来说，该法案有助于确保人工智能技术的安全和可控，防止误用和滥用。

2024 年 12 月欧盟数据保护委员会发布关于使用个人数据开发和部署人工智能模型的指导意见，提供了基于合法利益使用个人数据开展算法训练的要求。未来，欧盟执法机关是否会基于此意见对算法训

练的立场有所松动值得持续跟踪。

（2）美国

2022 年 10 月，美国白宫发布《人工智能权利法案蓝图》，提出五项基本原则，包括安全有效的系统、算法歧视保护原则、数据隐私原则、通知和解释原则、以及人工选择、考虑和退出原则。

2023 年 1 月，美国商务部下属机构美国国家标准与技术研究院（NIST）发布《人工智能风险管理框架》，该框架对人工智能存在的风险进行了分级与分类，旨在对人工智能进行有效风险评估。

2023 年 5 月，美国白宫发布《人工智能研究和发展战略计划：2023 更新版》，该计划强调需要对可信人工智能技术，以及对人工智能所引起的伦理、法律和社会问题进行重点研究。

2023 年 6 月，美国提出《国家人工智能委员会法案》，拟设立国家 AI 委员会，通过监管减少人工智能技术带来的风险与危害，并主导 AI 法规的建立。

2024 年 4 月，美国两党参议员提出了《2024 年人工智能创新未来法案》，旨在通过制定标准、评估工具、测试平台和加强国际合作，推动 AI 技术的创新和发展。法案强调了 AI 政策应最大化 AI 的潜力，以造福所有私人和公共利益相关者。

2024 年 9 月，美国加利福尼亚州通过了《人工智能透明度法》，要求 AI 生成内容以明示或隐式方式添加水印，旨在赋予消费者辨识内容是否由 AI 生成的能力。

2025 年 1 月，美国国会参议院情报委员会的主席 Josh Hawley

提出了一项名为《美中人工智能能力脱钩法案》的新法案，旨在全面切断美国与中国在人工智能（AI）领域的合作与技术交流。法案的通过将意味着美国和中国在 AI 技术及其相关知识产权的双向流动将会终止。

2025 年 1 月 13 日，美国商务部工业与安全局（BIS）发布了《人工智能扩散框架》的临时最终规则，并于发布后立即生效。该举措标志着美国政府对先进计算集成电路和闭源 AI 模型权重的出口控制力度进一步加大，并引入了全新的强制性全球许可制度。这项政策意在限制第三方国家与中国之间的正常贸易往来。

（3）中国

2019 年 6 月，国家新一代人工智能治理专业委员会发布的《新一代人工智能治理原则——发展负责任的人工智能》，强调加强人工智能系统可解释性、可靠性和可控性的重要意义。2021 年 9 月，国家新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》，该规范强调伦理道德在人工智能全生命周期中具有重要地位，可以促进人工智能技术公平、安全地发展。

2022 年 1 月，国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局为了规范互联网信息服务算法推荐活动，联合发布《互联网信息服务算法推荐管理规定》，旨在促进互联网信息服务健康有序发展。

2022 年 11 月，为规范深度合成技术的应用，国家互联网信息办公室、工业和信息化部、公安部发布《互联网信息服务深度合成管理

规定》，规定不得使用深度合成技术从事违反法律、行政法规的活动。

2023 年 7 月，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、国家广播电视总局发布《生成式人工智能服务管理暂行办法》，该办法强调国家坚持发展和安全并重，推动生成式人工智能创新发展的同时对生成式人工智能服务实行分类分级监管。该办法也指出生成式人工智能服务应当遵守法律法规，尊重社会公德和伦理道德。

2023 年 10 月，习近平总书记在第三届“一带一路”国际合作高峰论坛开幕式中提出《全球人工智能治理倡议》，围绕人工智能发展、安全和治理三个方面系统阐述了人工智能治理的中国方案。这是中国积极践行人类命运共同体理念，落实全球发展倡议、全球安全倡议、全球文明倡议的重要举措。

2024 年 7 月，李强总理出席 2024 世界人工智能大会暨人工智能全球治理高级别会议。大会发表《人工智能全球治理上海宣言》，该宣言系统性地提出推动全球人工智能健康有序发展的治理方案，彰显了中国积极参与人工智能全球治理的决心和努力，为全球人工智能治理注入了强劲动力。

2024 年 9 月，全国网络安全标准化技术委员会发布《人工智能安全治理框架》1.0 版，提出了包容审慎、确保安全，风险导向、敏捷治理，技管结合、协同应对，开放合作、共治共享等人工智能安全治理的原则。针对各类人工智能安全风险提出了相应的技术应对措施和综合治理措施，为促进人工智能健康发展和规范应用提供了基础性、

框架性的技术指南。

2025 年 3 月，为规范人工智能生成合成内容标识，保护公民、法人和其他组织合法权益，维护社会公共利益，国家互联网信息办公室、工业和信息化部、公安部、国家广播电视总局发布了《人工智能生成合成内容标识办法》，该办法将于 2025 年 9 月 1 日正式施行。

全球范围各国和组织都在积极推进针对 AI 技术监管的政策制定，防止利用 AI 技术从事违法、违背社会公德等滥用行为，促进 AI 技术更好服务人类。除了针对当下产生的 AI 安全问题进行政策制定，提升监管的完备性，还需要积极探索未来可能出现的潜在 AI 安全问题，提前布局相关法律法规与政策，防范可能出现的重大 AI 安全问题，改善监管的滞后性，避免不必要的损失与危害。

3.2 AI 安全治理标准规范现状

2017 年 10 月，国际标准化组织（ISO/IEC）成立人工智能分委员会，设立了人工智能安全工作组，并从人工智能可信度、鲁棒性、伦理道德等角度进行了一系列标准制定工作，如表 1 所示；国际电信联盟（ITU）从 AI 安全管理和应用服务等角度开展多项标准工作，如表 2 所示；电气与电子工程师协会（IEEE）从 AI 伦理安全风险、可解释 AI、负责任的 AI 等角度制定了多项标准，如表 3 所示。

表 1、ISO/IEC 代表性 AI 安全治理国际标准规范

	标准名称	阶段	发布时间
ISO/IEC	Information technology - Artificial intelligence -	已发布	2020-5

	Overview of trustworthiness in artificial intelligence		
	Information technology - Artificial intelligence - Overview of ethical and societal concerns	已发布	2022-8
	Information technology - Artificial intelligence - Guidance on risk management	已发布	2023-2
	Security and privacy in artificial intelligence use cases - Best practices	已发布	2023-5
	Information technology - Artificial intelligence - Management system	已发布	2023-12
	Artificial intelligence - Functional safety and AI systems	已发布	2024-1
	Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Guidance for quality evaluation of artificial intelligence (AI) systems	已发布	2024-1
	Cybersecurity and Privacy - Artificial Intelligence - Privacy protection	在研	-
	Cybersecurity — Artificial Intelligence — Guidance for addressing security threats to artificial intelligence systems	在研	-

表 2、ITU-T 代表性 AI 安全治理国际标准规范

	标准名称	阶段	发布时间
ITU-T	Guidelines for security management of using artificial intelligence technology	发布	2022-5
	Security requirements for AI systems	在研	-
	Security Guidelines for Generative Artificial	在研	-

	Intelligence Application Service		
	Security threats and requirements for data annotation service of generative artificial intelligence	在研	-
	Artificial intelligence generated content: General framework and requirements	在研	-
	Guidelines for data security using machine learning in big data infrastructure	发布	2025-4
	Requirements and evaluation methods of artificial intelligence agents based on large scale pre-trained model	发布	2025-3
	Security Requirements and Guidelines for Artificial Intelligence Agent	在研	-
	Security guidelines for synthetic data in the context of AI systems	在研	-
	Guidelines for Artificial Intelligence-generated content detection	在研	-
	Security guidelines for fine-tuning generative AI model	在研	-

表 3、IEEE 代表性 AI 安全治理国际标准规范

	标准名称	阶段	发布时间
IEEE	Guide for an Architectural Framework for Explainable Artificial Intelligence	发布	2024-8
	Recommended Practice for Privacy and Security for Federated Machine Learning	发布	2024-4
	Guide for Threat Intelligence Retrieval Framework Based on Large Language Model	在研	-
	Standard for Large Language Model Evaluation	在研	-

	Standard for the Functional Requirements for Cybersecurity-Specific Large Language Models	在研	-
	Standard for Evaluation Method of Machine Learning Fairness	发布	2025-5
	Guide for Framework for Trustworthy Federated Machine Learning	发布	2024-12
	Recommended Practice for Framework and Process for Deep Learning Evaluation	发布	2023-4
	Robustness Evaluation Test Methods for a Natural Language Processing Service That Uses Machine Learning	发布	2024-8

2020 年 7 月，为加强人工智能领域标准化顶层设计，我国国家标准化管理委员会、中央网信办、发展改革委、科技部、工业和信息化部联合印发了《国家新一代人工智能标准体系建设指南》，旨在推动人工智能产业技术研发和标准制定。2024 年 6 月，工信部、中央网信办、国家发改委、国标委对外发布《国家人工智能产业综合标准化体系建设指南》（2024 版），旨在强化全产业链标准工作协同，为推动我国人工智能产业高质量发展提供坚实的技术支撑。为积极响应《全球人工智能治理倡议》，支撑落实《人工智能安全治理框架》，充分发挥标准对人工智能技术应用和产业发展的规范引领作用，持续完善人工智能安全标准体系建设，全国网络安全标准化技术委员会组织编制了《人工智能安全标准体系（V1.0）》（征求意见稿）。全国信息安全标准化技术委员会、全国信息技术标准化技术委员会等已相继制定了多项有关人工智能安全治理的标准，如表 4 所示。

全球范围各个国家和组织针对人工智能安全问题，从不同层次制定了一系列标准，这有利于推动人工智能技术更加安全、稳定、高效地应用于各个场景。人工智能是一个技术快速迭代的领域，需对人工智能安全治理标准进行持续研究与探索，不断完善人工智能安全治理标准体系。

表 4、代表性 AI 安全治理国家标准规范

组织	标准名称	阶段	发布时间
全国网络安全标准化技术委员会	信息安全技术 机器学习算法安全评估规范	已发布	2023-8
	网络安全技术 人工智能计算平台安全框架	征求意见稿	-
	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	已发布	2025-4
	网络安全技术 生成式人工智能数据标注安全规范	已发布	2025-4
	网络安全技术 生成式人工智能服务安全基本要求	已发布	2025-4
	网络安全技术 人工智能生成合成内容标识方法（强制性国家标准）	已发布	2025-2
	网络安全技术 人工智能代码生成服务安全要求	制定中	-
	网络安全技术 互联网信息服务深度合成安全规范	制定中	-
全国信息技术标准化技术委员会	人工智能 管理体系	已发布	2024-11

四 AI 安全治理技术解决方案

AI 安全治理技术解决方案主要包括四方面，分别为：AI 基础设施安全解决方案、AI 数据安全解决方案、AI 模型算法安全解决方案以及 AI 应用安全解决方案。只有建立全方位、综合性的 AI 安全技术解决方案，才能有力推动 AI 技术朝着高效、安全、可靠的方向发展。

4.1 AI 基础设施安全解决方案

AI 智算基础设施安全解决方案主要包括：智算硬件设备安全解决方案、智算云安全解决方案、智算 MaaS 平台安全解决方案以及智算算力网络安全解决方案。

4.1.1 智算硬件设备安全解决方案

（1）硬件安全防护

严格管理机房和设备的物理接触权限，采用多级身份验证（如生物识别、门禁卡）和实时监控系统，防止非法人员接触硬件设备。同时，对设备自身设计抗物理攻击能力，如防拆卸、防篡改机制。

（2）硬件固件安全防护

一方面，部署安全启动流程，结合硬件安全芯片验证固件和系统的完整性，防止恶意代码植入或篡改。同时定期检查硬件运行状态，确保其符合可信计算标准。另一方面，建立固件和软件的自动化更新流程，及时修补已知漏洞。

4.1.2 智算云安全解决方案

（1）保护云底座

1）网络区域划分与访问控制：根据业务类型划分不同网络区域（如管理区、计算区、存储区），并在区域之间建立访问控制策略，限制非必要通信。例如，禁止外部直接访问核心计算资源，仅允许授权 IP 或设备接入。

2）利用云安全态势管理（CSPM）工具：CSPM 工具可以帮助识别和修复错误配置、确保合规性并提高整体云安全态势。该工具可持续监控云环境中与智能计算服务密切相关的错误配置，如用于训练数据的安全存储配置。

（2）加固云操作系统

1）实施补丁管理策略：云操作系统应当及时应用安全补丁和更新，可以使用云操作系统实例自动化补丁管理，优先处理安全更新并在非生产环境中进行测试。

2）应用操作系统加固技术：禁用不必要的服务、关闭未使用的端口、删除默认账户以及配置强大的安全设置。通过利用 CIS 基准或其他行业认可的加固指南来保护运行智算工作负载的操作系统。

3）强制执行最小权限原则：在操作系统配置用户账户和权限，以确保用户和进程仅拥有执行其任务所需的必要权限。

4）实施安全监控和日志记录：启用全面日志记录，使用安全监控工具来监测操作系统的可疑活动和潜在安全风险事件，并将这些日志与集中的安全信息和事件管理系统集成。

（3）保护容器镜像

1）使用可信的基础镜像：仅使用来自可信注册表的基础容器镜像，并维护一个内部批准的基础镜像目录，供智算项目使用。

2）实施容器镜像扫描：使用容器扫描工具在部署前识别镜像中的漏洞，并持续监控注册表中的镜像。同时删除不必要的软件包、禁用未使用的服务以及确保最小化镜像大小来减少容器镜像攻击面。

3）容器镜像签名验证：对容器镜像进行数字签名以确保其完整性和真实性，并在部署前验证数字签名，以确保仅将可信且未更改的镜像部署用于智算工作负载。

4.1.3 智算 MaaS 平台安全解决方案

（1）数据加密与访问控制

MaaS 平台需采用端到端加密技术保护数据在传输和存储过程中的安全性，结合身份认证（如多因素认证）和最小权限原则，严格限制用户对敏感数据的访问权限。通过 API 网关实现动态访问控制，确保只有授权用户才能调用模型服务。

（2）安全审计与持续监控

建立实时监控系统，跟踪模型调用、资源使用及异常行为，结合日志分析与威胁情报，及时发现潜在攻击。定期进行安全审计，验证访问控制策略的有效性，并通过自动化工具修复配置漏洞。

（3）供应链安全管理

严格评估第三方组件（如依赖库、框架）的安全性，要求供应商

提供漏洞修复记录和合规证明。通过代码签名和镜像扫描技术，防止恶意代码注入 MaaS 平台。

4.1.4 智算算力网络安全解决方案

（1）编排管理安全

编排管理安全包括安全需求感知、编排安全保障、智能安全调度、算力安全管控，是算力网络安全体系架构的管控核心。

1）安全感知：算力网络安全感知的对象主要包括计算任务安全需求和计算节点的安全信息。安全感知信息作为安全资源编排的基础，在算力网络安全感知过程中，需要对算网节点身份进行统一标识，便于进行统一身份认证、安全风险溯源及安全事件及时处置。同时，可以考虑增加安全标识信息，便于算力资源的安全接入和安全分级分类管理。

2）安全编排：算力网络安全编排过程中，应该通过安全通道收集网络和算力资源信息数据，确保信息数据的机密性、完整性和真实性。对于用户信息、任务数据、算力资源分布信息等算力敏感数据，需时刻对其进行监控和防护，保障数据安全。另外，编排管理行为应受到监控，对编排管理系统的访问权限应分级授权，防止非法用户越权调度算力网络资源。

3）安全调度：算力和网络跨域拉通时，建立管理通道和控制通道应进行双向认证，加强访问控制。同时根据任务需求，结合算力用户、算力任务、算力节点以及算力资源的安全等级标识，将任务调度

到具有相应安全等级的资源节点处，实现算力的安全调度。在具体调度过程中，需要根据实际情况对用户信息和任务数据进行加密传输，采用动态监测和节点验证等手段保障安全。

4）安全管控：安全管控需要从被动防御转变为自主检测和主动防御。这一转变体现在算力网络安全管控的各个环节中，涵盖了算力注册、算力资源全生命周期管理、算力统一证书管理、动态鉴权等关键环节，确保了算力资源的可信接入和管理。

（2）运营服务安全

运营服务安全包括安全监控、安全交易、安全审计等，是保障算力网络运营服务安全的有效手段。

1）安全监控：可采用安全日志管理和入侵检测等技术，实现对算力资源安全状态的实时监测，在检测到安全状态异常时将触发告警或其它自动响应机制，保障交易中数据和信息的安全。同时通过对算力资源进行访问控制，确保只有授权用户和程序能够使用所辖算力资源。此外也可通过限制算力用量、拒绝算力请求或降低算力用户信用等措施对非法算力使能行为进行管控。

2）安全交易：算力交易过程需要确保交易参与方的真实可信、交易过程的安全可控，做到安全事件可追溯，交易过程可追溯，安全风险可防范。

3）安全审计：安全审计包括对算力交易流程进行识别、记录、归档整理以及分析，对于重要记录需进行备份，确保出现问题时有据可查。同时，基于数据治理和访问控制规则，对算力网络中重要数据

和访问行为进行记录和审计，追溯数据的各处理环节，提升算力网络数据处理环节的公信力。

4.2 AI 数据安全解决方案

AI 数据安全解决方案需围绕全生命周期构建多维防护体系，从数据采集、训练、微调、推理到知识库管理的每个环节，均需融合技术手段与管理机制，确保数据从收集到销毁的全生命周期内都受到妥善管理。同时，采用先进的加密技术来保护数据安全，实施定期的数据质量检查和风险评估，及时发现并解决潜在问题。

4.2.1 通用 AI 数据安全解决方案

针对通用 AI 数据安全风险，可以采取一系列综合性的措施来保护数据的机密性、完整性和可用性，在 AI 数据采集、训练、微调、推理等多个阶段均会涉及。

（1）数据合规防护

1）数据脱敏：可以使用自动化工具和技术，如自然语言处理算法和机器学习算法自动识别敏感数据。敏感数据检测不仅包括识别个人身份信息，还包括商业秘密等类型的敏感信息。同时，可以使用数据脱敏技术，对敏感数据进行脱敏处理，以减少敏感信息泄露的风险。

2）数据合规清洗：确保数据满足法律、法规及业务规则，并对敏感信息进行保护。同时实施加密和访问控制等安全措施，从而满足企业的合规需求并保障数据的安全性和可用性。

（2）数据泄露防护

1）数据加密：无论是静态数据（如数据库中的信息）还是动态数据（在网络中传输的信息），都应采用强加密算法进行加密。

2）访问控制与权限管理：实施严格的访问控制策略，确保只有经过授权的用户才能访问特定的数据。

（3）数据篡改防护

1）数据投毒检测：建立严格的数据验证机制，在模型训练之前对数据集进行中毒检测，如数据分布一致性验证、异常样本检测等，识别出可能存在的恶意、异常的数据样本。同时，建立严格的访问控制机制，限制数据存储库访问与修改权限。

2）数据备份与恢复：定期备份数据，并确保能够迅速恢复以应对数据丢失或系统故障的情况。例如数据备份采用“3-2-1”备份原则，即至少保留三份数据副本，使用两种不同的存储介质，并至少有一份备份位于异地。

（4）数据质量防护

1）数据完整性验证：保证数据未被未经授权的方式修改或删除至关重要。常用的方法包括使用哈希效验技术检查文件的一致性，或者在数据库中通过日志追踪修改历史。

2）数据清洗与标准化：数据集通常包含损坏、缺失等问题，因此需对数据集的错误、重复和不完整等问题进行清洗，从而提高数据的准确性与质量。同时，需建立统一的数据标准格式，规范数据处理流程，保证数据的一致性和正确性。

4.2.2 训练数据安全解决方案

（1）数据脱敏和加密

在收集和处理数据时，可以采取脱敏和匿名化技术，确保不泄露敏感信息；在使用现有数据训练模型时，明确版权声明和许可协议，避免侵犯他人版权；使用机密计算技术（如 TEE 可信执行环境），在隔离环境中完成数据训练，确保外部无法访问原始数据。

（2）访问控制与安全认证

实施最小权限原则，通过多因素认证（MFA）和基于角色的访问控制（RBAC）限制数据访问；加强模型服务器的安全认证管理，防止未经授权的访问和服务中断。

（3）数据质量管理

通过对抗性训练提高模型的鲁棒性，防止数据投毒和对抗攻击；建立数据质量检测机制，通过去重、去噪和标注校验等手段提高数据质量。

4.2.3 微调数据安全解决方案

（1）安全微调环境

构建封闭的沙箱环境进行微调，限制外部网络连接以防止数据泄露。同时，可以使用训练推理一体机内置的安全方案，完成数据合规清洗并添加水印，防止数据篡改。

（2）权限管理和数据最小化暴露

在微调阶段实施细粒度权限控制，例如仅允许特定角色修改模型参数，并记录操作日志；在微调时仅使用与任务目标相关的必要数据，并利用数据裁剪等技术手段减少敏感信息暴露。

4.2.4 推理数据安全解决方案

（1）输入输出安全防护

输入端建立对抗样本检测机制，过滤异常输入数据；输出端对推理结果进行脱敏处理，例如隐藏敏感字段或返回模糊化结果。

（2）运行时安全监控

采用哈希校验和数字签名技术，确保推理过程中模型文件与推理数据的完整性；部署异常检测模块，识别并阻断恶意查询请求。

4.2.5 知识库数据安全解决方案

（1）数据库安全访问控制

实施数据分级分类管理，基于数据敏感度动态调整访问权限；建立版本控制和审计追踪机制，记录知识库的所有修改与访问行为。

（2）数据脱敏与隔离

对结构化数据如手机号、证件号进行加密脱密，对非结构化数据进行语义级脱敏，保证数据使用时仅暴露必要信息；通过零信任架构隔离核心数据资产，限制数据流转。

4.3 AI 模型算法安全解决方案

AI 模型算法安全解决方案包括：通用 AI 模型算法安全解决方案和生成式 AI 模型算法安全解决方案。

4.3.1 通用 AI 模型算法安全解决方案

（1）模型鲁棒性增强

为了增强模型的鲁棒性，可以采取多种策略，如对抗训练、数据增强、正则化、集成学习等技术。例如：对抗训练技术通过在训练过程中向输入数据添加微小但精心设计的扰动来模拟攻击，从而使得模型学习到如何正确分类这些被扰动的数据；数据增强技术可以通过对训练数据进行变换（如旋转、平移、缩放等）增加数据集的多样性，使模型更具有泛化能力；正则化技术（例如 L2 正则化）能够减少过拟合现象，确保模型不会过度依赖于特定的数据特征；集成学习策略结合多个模型的结果可以进一步提高整体模型的稳定性，因为不同模型可能对不同的干扰有不同的抵抗力。

（2）模型泛化性增强

常见的增强 AI 模型泛化性方法包括：扩大数据集规模、增加数据样本的多样性、正则化技术、优化模型结构等。例如：在神经网络模型中使用 dropout 正则化技术可以帮助防止模型过度拟合训练数据，从而提升模型在未见过的数据上表现；扩大数据集规模并增加样本的多样性是一种更为直接、简单且有效的提升模型泛化性的手段。

（3）模型可解释性增强

提高 AI 模型的可解释性对于理解模型的决策过程至关重要，尤其在医疗、教育、金融等领域。常见的方法是进行特征重要性分析，通过识别对于决策结果更为关键的重要特征，提升模型决策可解释性。可解释性技术不仅帮助用户更好地理解模型的工作原理，也增加了模型的透明度和可信度。

（4）模型偏见与歧视减轻

为了避免模型中出现偏见和歧视，可以在数据收集阶段注意数据的代表性和公平性，确保训练数据涵盖了广泛的社会群体，避免某些群体被低估或忽视。此外，在模型设计阶段还可以引入公平性约束条件，确保不同群体间的错误率相等或者机会均等，并且部署前应进行全面的公平性测试，检测模型是否存在系统性偏见，并及时调整模型参数或算法，以保证模型的公正性和准确性。

（5）模型逆向工程防护

保护模型免受逆向工程攻击涉及多个层面的安全措施。例如：通过使用模糊模型输出技术，可以使攻击者难以直接从输出推断出模型的具体参数或结构；通过使用加密技术保护模型参数的安全性，可以确保在网络传输过程中，模型参数也不会被未经授权的人访问或篡改。此外，还可以考虑采用混淆技术隐藏模型内部结构，增加模型逆向工程的难度。

（6）模型对抗攻击防御

对抗攻击防御策略可以分为主动防御和被动防御两大类。主动防御技术包括对抗训练、多模型融合等方法，旨在通过提前预防和自我

进化来增强模型对对抗攻击的防御能力。被动防御则侧重于事后检测与响应，如使用异常检测算法识别对抗样本。通过综合运用这两种防御方式，可以构建主被动相结合的对抗攻击防御屏障，提升模型的安全性。

4.3.2 生成式 AI 模型算法安全解决方案

（1）大模型安全对齐

大模型安全对齐是指通过强化学习与人类反馈对模型进行再训练，使大模型生成内容的价值观与人类价值观保持一致，避免生成有害内容或错误决策。大模型安全对齐技术通常可以分为以下几步：

1）数据处理阶段筛选训练数据：在训练前对数据进行严格筛选，标记有害内容（如暴力、偏见信息），确保数据符合法律和社会规范；

2）训练阶段进行对齐优化：可以使用强化学习技术并结合人类反馈和偏好优化平衡模型的有害性与有用性。还可以采用潜空间特征对齐方法生成对齐向量，在不显著修改模型参数的情况下调整输出方向，兼顾安全性与性能；

3）推理阶段加入后置修正：在模型输出端部署后置对齐模块等安全修正模块，实时拦截并修正模型输出的有害内容。

随着多模态大模型在输入维度（图像、音频、视频、传感数据等）和输出形式（文本生成、语音合成、行为控制等）上的能力不断增强，其潜在的安全风险也更为复杂、多样化。相比传统大语言模型，多模态模型存在跨模态内容审查的特殊复杂性，需同时处理文本、图像、

音频等多模态信息的有害关联。例如，检测图文组合的隐含歧视内容，或拦截语音指令中的隐蔽攻击。因此，针对多模态大模型的安全对齐一方面需要开发跨模态对齐向量，确保不同模态的输出在价值观上一致，如文本描述与生成图像均符合伦理规范；另一方面要引入多模态对齐机制，协调不同模态处理模块的决策逻辑，避免局部安全而整体违规的风险。

（2）大模型幻觉减轻技术

1）训练阶段的幻觉减轻技术

① 构建高质量的预训练与微调数据集：通过数据过滤技术，选择高质量的数据，确保数据的事实准确性，减少预训练数据集中的错误信息、偏见等。随着数据量规模的不断扩大，当前数据过滤方法的效率和可扩展性面临重大挑战，忽略了由大模型生成的内容所带来的影响，需要开发更加高效的自动化数据过滤算法。

② 模型编辑：模型编辑提供了一种精确的方法来缓解由特定错误信息引起的幻觉现象，无需全参数训练而是通过定向编辑模型参数来注入最新的知识，从而纠正模型行为。当前的模型编辑技术可以分为两类：定位-然后编辑和元学习。定位-然后编辑方法首先定位模型参数中的“错误”部分，然后对其进行更新以改变模型的行为。元学习方法是通过训练一个外部超网络来预测原始模型的权重更新。然而，当持续使用模型编辑技术注入新的知识时，可能会导致模型的整体性能下降。

2）推理阶段的幻觉减轻技术

① 事实性增强解码策略：旨在通过优先考虑生成信息的事实性来提高大模型输出的可靠性，侧重于使模型输出与既定的现实世界事实紧密对齐。事实性增强解码策略试图利用大模型的自我纠正能力来精细化控制生成的内容，并且不依赖外部知识库。事实性解码策略具有即插即用特性，适用于在不需要进行计算密集型训练的情况下使用。

② 忠实性增强解码策略：该策略侧重考虑大模型生成的内容与提供的上下文保持一致，并强调增强生成内容内部的一致性。受到人类思维过程的启发，思维链被引入大模型推理过程中，将复杂问题分解为明确的中间步骤，从而增强推理过程的可靠性。忠实性增强解码策略显著推进了大模型生成内容与所提供上下文的一致性，并增强了生成内容的内部一致性。

③ 检索增强生成（RAG）技术：该技术利用外部非参数数据库对模型进行知识补充。首先从外部数据源检索相关知识，然后基于用户查询和检索到的文档，由大模型生成最终响应。通过将外部知识与大模型分离，RAG 可以有效缓解由知识差距引起的幻觉现象，而不会影响大模型的其他能力表现。RAG 技术具有模块化、灵活性等优势，将外部知识库视为插件，可以根据需要进行替换或修改，从而减轻由知识差距引起的幻觉，适用于任何领域。

4.4 AI 应用安全解决方案

AI 应用安全解决方案包括：AI 应用开发安全解决方案和 AI 垂直行业应用安全解决方案。

4.4.1 AI 应用开发安全解决方案

（1）智能体安全解决方案

AI 智能体不仅继承了大模型所面临的安全风险，还需要面对智能体所具有的特定安全风险，这些安全风险是由于智能体与外部交互而产生的，因此需要采取一系列措施防范这些安全威胁。

1）决策链路追踪：记录智能体从数据访问到模型推理的完整操作过程，确保所有活动都有迹可循。

2）安全审计：验证模型来源，定期进行安全性审查和风险评估。

3）安全阻断：针对智能体代理功能性操纵风险，应采取主动安全措施。例如在使用第三方 LLM 代理时，应保护个人隐私并警惕第三方过度的数据请求；严格控制用户数据的共享范围，在交互过程中严格防止个人敏感信息的泄露。

4）建立安全通信协议：设计安全的通信协议，明确智能体之间交互的规则和格式，确保智能体之间的通信安全可靠，防止信息在传输过程中被窃取或篡改。

（2）具身智能安全解决方案

具身智能安全解决方案需围绕“感知—认知—行为”的闭环结构，构建多层次防御机制。一方面，采用多模态冗余感知、对抗鲁棒性训练与行为可解释架构等技术，保障在复杂动态环境下的感知稳定性与策略可靠性。另一方面，需引入社会价值观对齐与语境适应机制，确保系统在跨文化、跨场景应用中的行为合理性与伦理合规性。此外，

针对多模态联合攻击与异构平台部署风险，需强化模态一致性检测、策略迁移验证与人机协同接口，构建具备自诊断、自恢复与风险预警能力的安全闭环。具身智能安全的根本目标在于实现“可信、可控、可解释”的智能行为，支撑其在高风险场景与民生应用中的大规模、可持续部署。

4.4.2 AI 垂直行业应用安全解决方案

（1）AI+医疗安全解决方案

对于手术机器人这样强场景且软硬件结合的 AI 应用，需要降低算法偏差和系统漏洞风险。一方面，可采用对抗性训练增强算法鲁棒性，在模型的训练数据中引入对抗样本，并通过强化学习等技术，在仿真环境中提升 AI 系统应对突发状况的能力。另一方面，手术中可通过高精度传感器与光学定位技术对机械臂运动轨迹进行动态监控，结合术前三维重建模型与术中实时影像数据等，对算法策略进行实时校验，当检测到运动偏差超过预设阈值时自动触发紧急制动。

（2）AI+编程安全解决方案

针对代码生成漏洞隐患，可以通过自动化安全检测与加固机制来加强防护。具体而言，即在模型训练阶段通过对抗性训练优化模型使其主动规避漏洞模式，在应用输入端上加入护栏机制进行输入净化、参数化查询等安全约束，在应用输出端集成静态分析、动态测试及形式化验证工具来自动化筛查 XSS、SQL 注入等漏洞。针对架构设计缺失与人工审查不足的问题，可以部署辅助审查工具，用来标注安全风

险点，并在核心模块的方案设计过程中，配合专家进行架构方案的复核。此外，还可以建立代码溯源机制追踪漏洞根源，形成自动化与人工审查的闭环防御。

五 AI 安全治理行业案例

5.1 中国联通人工智能安全治理一体化解决方案

图 5.1 中国联通人工智能安全治理一体化解决方案

中国联通全栈自主创新研发“人工智能安全治理一体化解决方案”，通过智能化安全评测技术精准识别风险，依托可信增强技术提升模型安全性，实现多模态大模型主动防护，助力教育、能源、农业、制造等垂直行业的人工智能安全应用。

5.1.1 大模型内容安全评测与增强

中国联通自主创新构建了一套覆盖多维度的大模型安全评估体系，严格遵循国家标准《生成式人工智能服务安全基本要求》，形成包含 5 大类 31 小类的百万级评测数据集。重点针对医疗、教育、制

造、交通、农业等关键行业及政治敏感领域，建立具有针对性的攻击样本库，通过模拟多样化攻击手段进行对抗测试，形成了具有高攻击成功率的大模型安全评测数据集。依托自研的裁判大模型和智能评测策略，实现了对生成内容安全性的高精度检测，有效防范潜在风险。基于内容安全评测结果，中国联通建立了多层级的模型算法安全增强机制，通过细粒度分析政治、伦理、暴力等风险类型，研发有针对性的价值观对齐算法，在微调阶段植入“以人为本、智能向善”的安全准则。该方案创新性地结合对抗训练、强化学习等技术手段，并采用动态权重调节技术，在提升大模型安全性的同时保障大模型业务性能。

5.1.2 大模型中文幻觉评估与减轻

针对金融、医疗、政务等关键领域的大模型应用，中国联通打造了基于垂直行业定制化的中文幻觉评估解决方案。该方案通过构建多维度评测指标体系，结合细粒度分析方法，精准识别模型在知识准确性、逻辑一致性等方面的缺陷。通过构建行业知识图谱基准库和事实验证引擎，实现对模型输出的精准量化评估，为高风险领域的人工智能应用提供了可靠的质量保障。此外，根据幻觉评估结果，中国联通还建立了从数据源头到生成输出的全流程治理体系。在数据层，开发了训练数据安全审核算法，通过知识可信度验证技术过滤低质噪声数据；在推理层，创新应用思维链引导技术，确保生成内容可信、完整可靠。同时，通过事实验证模块与知识图谱的实时校验，结合动态监测模块自动触发潜在幻觉风险复核，显著提高输出可靠性。

5.1.3 大模型可解释性评估与可控增强

中国联通创新研发的大模型可解释性综合评测工具集，包含鲁棒性测试与可解释性分析两大模块。该工具集通过对抗攻击模拟和噪声干扰测试评估模型抗干扰性能，同时运用神经元激活分析、特征重要性量化等技术解析模型决策逻辑，不仅提升了人工智能系统的透明度，还通过量化指标支持模型优化，为金融、医疗、制造等强合规场景提供了可信赖的评估方案。依托可解释性分析结果，自主创新研发知识注入算法，有效解决知识缺失问题，通过自适应检索增强生成技术动态补全模型知识盲区；针对价值偏差，开发多模态价值观对齐算法，在文本、图像等生成任务中植入伦理约束；针对公平性问题，创新提出分层偏见修正算法，有效消除性别、地域等敏感属性歧视。

基于人工智能安全治理一体化解决方案，中国联通自主创新研发人工智能安全治理服务平台，有效解决了大模型在行业应用中面临的有害内容风险、事实偏差和决策透明度低等关键问题。在教育、医疗、金融、制造等重点领域，依托百万级评测数据集精准识别风险，通过价值观对齐算法和幻觉减轻技术，有效提高大模型安全性，实现了从模型训练到应用落地的全流程治理，既确保生成内容的安全合规，又保障专业场景下的输出可靠性。除此之外，中国联通持续引领人工智能安全国际标准，有效提升我国网络安全、人工智能安全国际影响力和话语权，为构建全球人工智能安全治理生态提供联通方案。

5.2 华为推理数据与知识库安全解决方案

5.2.1 推理数据安全解决方案

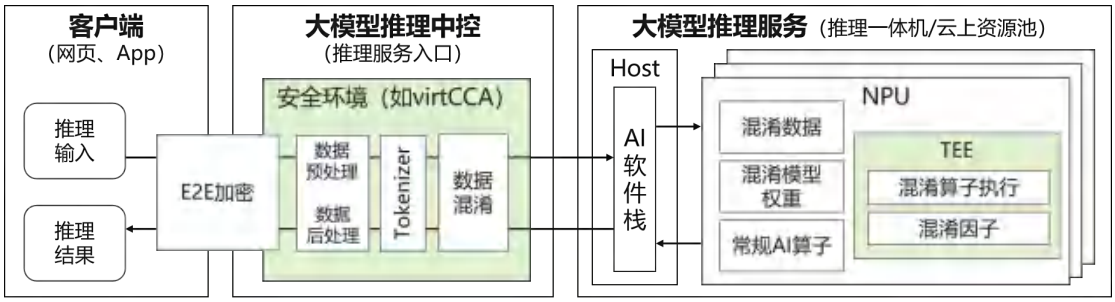

图 5.2 华为 PMCC 机密计算方案

针对大模型推理应用场景中模型与数据隐私保护的安全需求，华为提出了基于硬件可信执行环境 TEE 的 PMCC（Privacy and Model Confidential Computing）轻量级 AI 机密计算方案。

PMCC 安全防护方案以昇腾 NPU 的可信执行环境 TEE 为基础，在传统存储态保护的基础上，通过在 TEE 中利用混淆因子对模型权重参数及推理数据进行混淆变换的方式，将防护能力进一步扩展到了运行态。PMCC 可以确保混淆化模型处理混淆化数据后，仍然保证正确的推理结果，同时推理过程仍可以被 AI 核心正常加速。在模型使用过程中，所有在开放环境中出现的模型权重和推理数据均进行了混淆化处理。因此，即使攻击者获得了较高的访问权限，也无法直接从 CPU 或 NPU 中获取到模型和数据的明文信息。

除了保护模型权重参数，PMCC 方案还能够在推理、微调中同步保护用户的敏感数据。由于 PMCC 的混淆变换是在昇腾 NPU TEE 的高安环境中执行的，因此可以实现变换因子与硬件的绑定，避免传统软件安全方案可被拷贝到其他硬件上执行的缺陷。

5.2.2 知识库安全解决方案

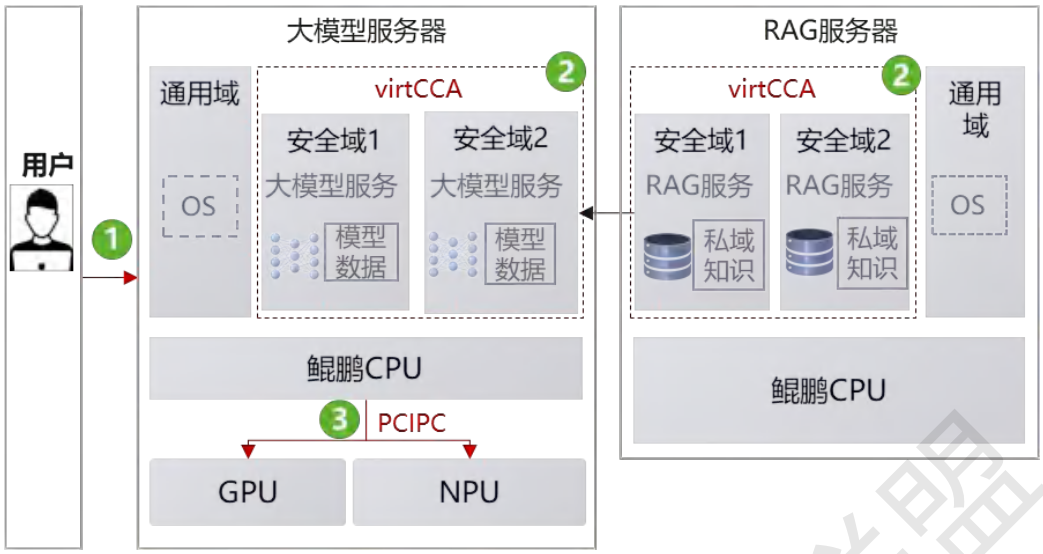

图 5.3 鲲鹏 AI+RAG 安全方案

基于鲲鹏独有的 virtCCA 与 PCIPC 异构机密计算组件构建的 AI 和 RAG 安全方案，通过用户认证和访问鉴权机制实现了敏感数据访问的分权分域管理。利用 virtCCA 技术，将高安全性应用与敏感数据分离部署，确保各应用间存在物理隔离的安全域，从而加强了数据保护。此外，利用 PCIPC 能力能有效将 CPU 安全域平滑拓展至 XPU 设备，无需修改现有训练框架、模型和设备驱动，在提升安全 IO 性能的同时，实现即插即用即安全。基于 openGauss 向量数据库+RAG 安全网关实现高危命令和拖库检测，防范数据窃取和注入安全风险。

5.3 百度大模型安全解决方案

5.3.1 百度智能云千帆大模型数据安全解决方案

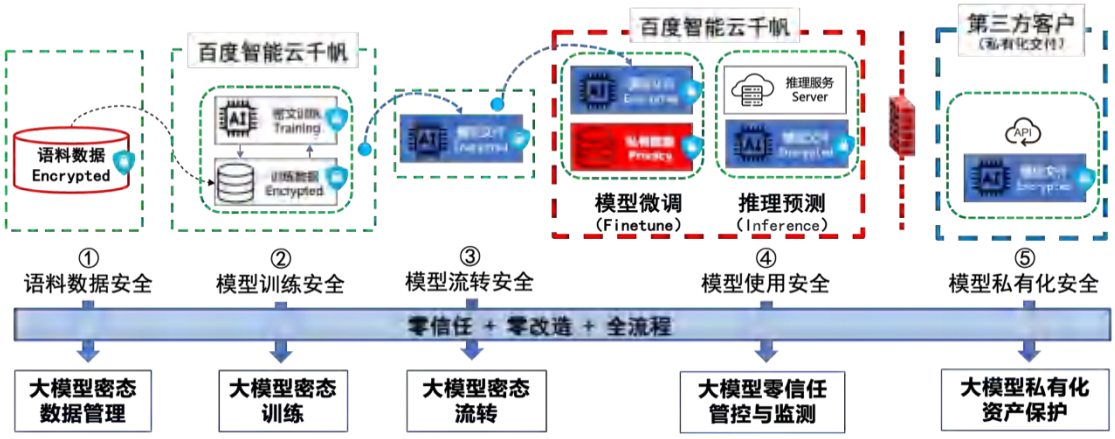

图 5.4 Baidu AI Realm 大模型数据安全技术框架

百度为防范大模型全生命周期各阶段相关数据安全风险，将领先前沿数据安全与隐私保护技术与大模型生态相结合，形成 Baidu AI Realm 大模型安全技术框架，为百度智能云千帆大模型业务提供端到端的数据密态管控与数据安全合规能力，覆盖大模型语料数据安全管理、大模型训练数据安全管控、大模型微调数据安全管理、大模型推理安全服务、大模型私有化数据资产保护等大模型生命周期各环节。

（1）大模型语料数据安全管理

大模型业务开展过程中涉及大量数据安全管理工作，Baidu AI Realm 大模型数据安全技术不仅为大模型数据安全提供基于文心大模型的智能分类分级管控，同时还提供了全流程密态管理方案，进一步加强大模型语料数据、标注数据和日志数据中的敏感数据和机密数据全流程安全保护。

（2）大模型训练数据安全管控

大模型训练过程中涉及多个部门和大量员工的分工协作，极容易出现大模型文件、大模型参数、大模型语料配方等大模型生产工艺的

数据泄漏风险。Baidu AI Realm 为大模型训练环节提供全流程安全管控措施，通过数据安全风险评估等机制，有效防范模型文件泄漏风险，保护大模型生产工艺等知识产权秘密，助力大模型业务健康有序发展。

（3）大模型流通安全服务

在完成大模型的训练后，企业将大模型分发给内外部的多个业务部门进行部署和使用。然而在大模型分发流转过程中，大模型数据流通安全管控至关重要。Baidu AI Realm 为大模型在企业内部流通使用提供一整套安全管控手段，结合百度数据管理平台，为大模型流通安全管控提供全流程密态管控，确保模型资产在流转流通过程中始终处于加密状态，有效防范明文模型资产的泄漏。

（4）大模型微调推理数据安全管理

大模型训练完成后，将大模型部署在智算中心的服务器上，为业务部门提供模型推理和模型微调服务，是大模型常见的应用方式。

大模型推理安全方面，Baidu AI Realm 针对日志类型敏感数据文件提供 FUSE 文件透明加密方案，使业务部门在“零改造”的情况下低成本实施敏感数据加密，降低业务数据安全合规成本。

模型微调安全方面，Baidu AI Realm 为大模型微调提供基于零信任的隔离管控措施，并结合百度数据管理平台为大模型微调提供数据清单、环节清单、程序清单等一系列数据安全隔离管控方案，在保护基座大模型资产的同时，也保护客户数据隐私安全。

（5）大模型私有化数据资产保护

私有化部署是大模型商业业务中一种重要的服务方式，Baidu AI Realm 为大模型私有化部署提供端到端安全方案，将大模型私有化场景的模型、数据和程序实施一体化防护，充分保护企业数据资产和知识产权。Baidu AI Realm 的大模型私有化保护方案包括模型文件加密保护、关键参数加密保护、私有化 License 安全管控等安全方案，有效防范大模型私有化过程中的数据资产和知识产权泄漏风险。

Baidu AI Realm 大模型数据安全技术框架为百度智能云千帆大模型数据安全提供全流程数据安全管控方案，有效防范百度智能云千帆大模型全生命周期各环节的数据安全风险。

5.3.2 百度多模态审核大模型安全建设方案

百度多模态内容审核流程正从“小模型串联式”的多模态识别方案，演进为基于统一大模型的“理解驱动式”多模态审核架构，图 5.5 展示了当前多模态内容审核系统从传统小模型拼接流程向统一多模态大模型架构的演进路径，明确了两个阶段的审核流程逻辑差异与能力提升重点。

基于统一多模态大模型的智能审核框架是新一代多模态大模型审核流程，主要特征是利用融合图文理解能力的大模型进行整体风险判定。其流程逻辑如下：

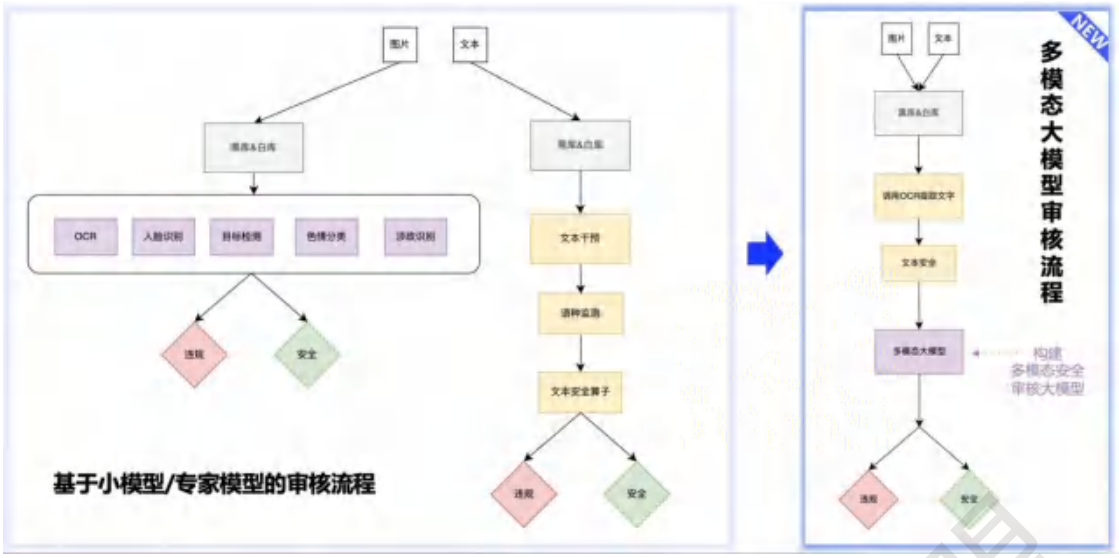

图 5.5 模型审核流程演进

（1）统一抽取模态内容：图像和文本仍作为原始输入，但在处理前，会调起 OCR 从图片中抽取文字信息，实现视觉与语言模态的合并。

（2）融合语言处理：将抽取到的文字信息与原始文本一并送入统一的“文本安全”模块，进行初步语义筛查。

（3）构建多模态审核大模型：核心步骤是将图像本身与其对应的文字信息（含图中提取文本）一同送入一个具备图文对齐和语义理解能力的多模态模型中。该模型通过对图像内容与文本语义的联合建模，识别其中潜在的风险内容或语义诱导问题。

（4）一体化风险判定：由大模型直接输出风险判断结果，实现对图文共现、语义隐喻、模态协同攻击等复杂场景的审查能力。

多模态安全审核大模型，在训练阶段从 prompt 构建到全量微调，分为几个关键阶段：1）首先是 Prompt 构建，需要设计构建适合安全审核任务的提示词模板，引导模型更好地理解和输出内容，通过此

方法可以提高模型在内容审核任务中的表达效率，减少输出不确定性；2）其次是输入尺寸优化，通过多次不断的优化多模态输入的尺寸，提高模型训练效率和表现，核心目标是减少显存占用，同时不丢失关键信息，提高收敛速度；3）然后进行辅助任务训练，用多种子任务辅助主任务（如 OCR 违规内容判别、理解等），用来提升模型对复杂场景中“文字+目标+语义”结合问题的理解能力；4）最后进入全量微调，在主任务数据上做完整训练，解决多任务冲突，输出最终模型，训练出在审核场景下泛化能力强、稳定性高的最终模型。

多模态安全审核大模型通过融合语言与视觉理解能力，具备对图文共现语义的深度建模与风险识别能力，能够有效识别如图文配合诱导、视觉文本隐喻、跨模态攻击等复杂风险场景。同时，大模型具备更强的上下文记忆能力，可结合对话历史、指令意图等进行动态判定，显著提升审核智能化水平。相比传统方案，大模型审核体系不仅能提升识别准确率、降低误判漏判，还能在模型架构上实现能力的一体化融合，降低系统维护成本，提高安全体系的灵活性与演进空间。

5.4 360 大模型安全解决方案

5.4.1 360 大模型内容安全护栏系统

360 智盾大模型内容安全护栏系统针对大模型应用过程中产生的内容安全问题，建立了一套大模型内容安全护栏系统，所有用户输入的内容都会经过风险检测模块，根据识别的标签对输入内容进行分级分类。其中，红线类问题直接拒答，敏感类问题流转给安全代答大模

型生成安全回复，完全无害的问题流转给主干大模型进行回答。模型输出的回答再经过一次风险检测引擎，检测无害后正常返回给用户，检测违规及时中断回复，确保模型回答内容的安全性。

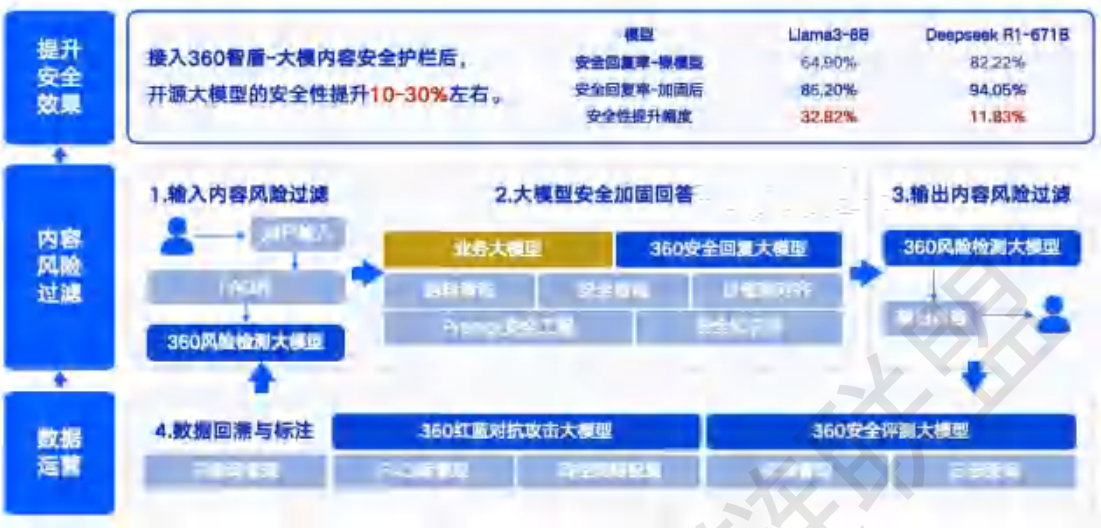

图 5.6 360 智盾大模型内容安全护栏系统

5.4.2 360 智鉴-大模型系统安全检测平台

360 智鉴大模型系统安全检测平台分别针对算力基础设施、业务开发环境、大模型在线服务、智能体应用的安全性开展评测。智鉴系统能够自动识别不同具备网络接口的大模型服务，如 Ray、Ollama 等，并进行安全性验证。目前大模型服务指纹库条目 30+，PoC 验证库条目 70+。针对大模型服务开发环境，智鉴系统能够收集项目中的大模型组件信息。基于已识别的组件清单，扫描其中已知的安全漏洞，并提供详细的检测结果和修复建议。目前支持大模型组件库种类 1000+，漏洞库条目 500+。

图 5.7 360 智鉴-大模型系统安全检测平台

5.5 奇安信大模型安全技术解决方案

面对伴随大模型应用出现的新问题，奇安信提出了一套针对性的多维度纵深防御管控体系，包括了拦攻击、审数据、控权限、强基础、清风险五大核心防护战略，构建起构建覆盖全场景的安全防护体系。

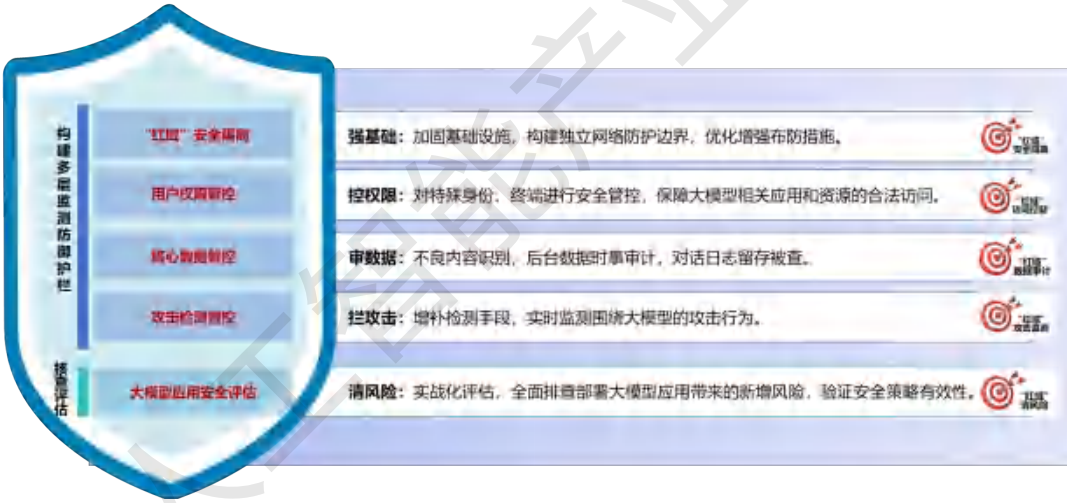

图 5.8 奇安信大模型安全解决方案

（1）“红域”安全隔离，针对大模型应用，打造独立的“红域”安全隔离区，对算力资源、模型数据等核心资产实施分层防护，同时对专用终端、服务器等基础设施进行专项安全加固和优化。

（2）用户权限管控，通过落地零信任架构以及终端安全空间隔

离措施，严格限制开发、训练、运维等特权人员的操作权限，有效降低内部人员误操作或恶意窃取、滥用大模型相关软硬件资源的风险。

（3）核心数据管控，针对大模型训练微调、能力增强相关的知识库数据集，以及所有用户与大模型业务应用的对话内容，进行全量的安全审计和分析监测。严格审查所有输入输出内容的安全性，确保内部数据和敏感信息不被窃取或滥用。

（4）攻击监测管控，通过大模型卫士和天眼威胁监测系统，实时拦截针对大模型的新型恶意攻击，比如提示词注入、数据爬取等，有效抵御外部入侵。

（5）大模型应用安全评估，奇安信安全实验室积累了大量的 AI 安全攻防经验，可以通过实战化的形式为企业提供专项的风险评估服务，全面排查大模型部署带来的新增风险，也可以用来对现有安全防护策略进行有效性验证，确保新型风险提早预防、新增缺陷提早修复。

奇安信大模型应用安全防护方案，在兼顾传统风险的同时，针对大模型应用特有的数据泄露、模型攻击、内部滥用等问题，提供从终端、网络、应用、模型、数据、主机的全链条防护体系，涵盖了大模型应用的训练、评估、部署、上线和日常运营等各类风险场景。而且，依托于奇安信庞大的威胁情报网络，及时获取大模型安全相关的最新威胁和攻击特征和趋势，专家团队也已经积极参与到大量真实的大模型应用安全事件应急中，积累了丰富的实战经验，为方案整体的落地效果提供坚实的保障。

5.6 超聚变 xRAY 智能服务一体机解决方案

超聚变 xRAY 智能服务一体机解决方案基于 AI 软硬件资源为企业各业务场景提供 AI 服务，依托 MaaS 大模型平台对外提供 AI 能力，实现对话型应用、文本生成应用、Copilot 及 ChatToX 形式的多种落地模式，为企业各场景提供 AI 助力。平台总体框架如下图所示：

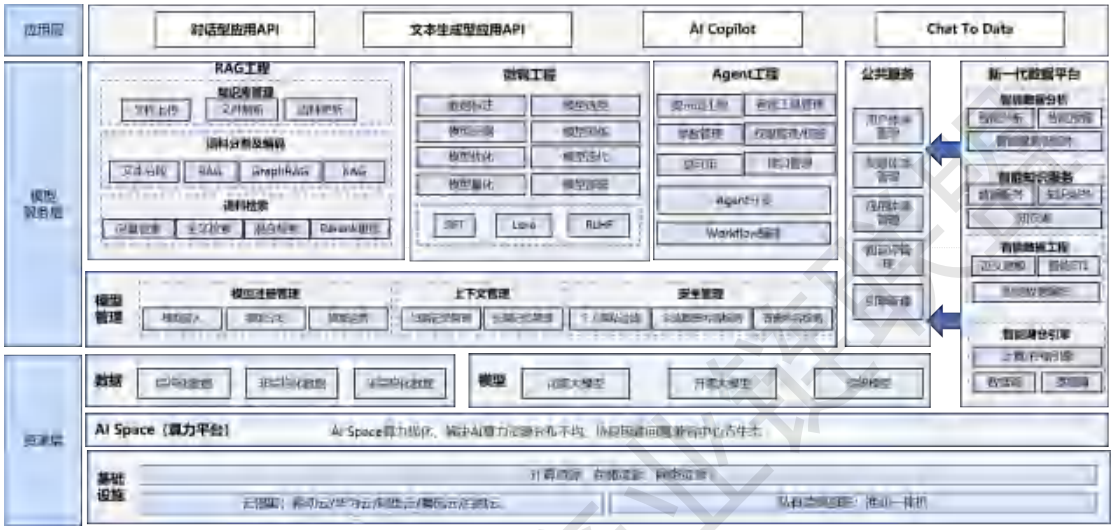

图 5.9 超聚变 xRAY 智能服务一体机

（1）资源层包括基础设施、数据和模型三个层次：

1）基础设施层为超聚变 AI 大模型解决方案整体应用的硬件设备，用于提供不同方式的大模型能力接入，包括算力服务器、网络资源、云部署方式、本地部署训推一体机等；

2）数据是贴合业务场景的前提，对企业内部业务场景来讲，大模型需基于已有业务数据提供 AI 服务，业务数据包括企业的结构化数据、非结构化数据及半结构化数据；

3）模型包括大模型与传统模型，为上层提供 AI 能力，超聚变 AI 大模型解决方案可支撑各类大模型接入提供服务，包括开源大模

型、闭源大模型及传统小模型，提供大小模型结合的AI服务能力。

（2）MaaS平台主要包括Foundation、模型管理、RAG工程、微调工程、Agent工程等：

1）Foundation：是超聚变AI大模型平台的基础，主要进行平台基础用户、权限、应用、数据源及执行引擎等基础模块管理；

2）模型管理：通过模型管理模块对外部大小模型进行封装，主要包括内外部模型的注册分发、各模型的上下文记忆管理、及面向应用输入输出内容的安全过滤模块；

3）RAG工程：主要用于文本生成场景中利用内部知识库或文档库，以提高生成结果的准确性和丰富性。主要包括知识库创建管理、知识库语料分割、语料内容检索模块，使得在生成文本时引入已有内部知识；

4）微调工程：模型微调主要在预训练模型基础上，用业务的少量数据进行训练，以提升模型在业务场景中的表现，包括数据标注、模型选型、模型评测等模块，对后续应用层的结果准确率至关重要；

5）Agent工程：对复杂的业务场景任务，需要利用Agent能力来设计、构建和管理对应的AI能力，主要包括基于大模型的提示词工程、复杂场景的可视化工作流编排、agent低代码开发等。

（3）应用层：面向不同类型的业务场景，超聚变AI大模型平台提供了多种场景落地形式，包括问答式的对话型应用、预设提示词的文本生成型应用、与业务系统融合的AI Copilot、基于数据处理场景的Chat To X场景；可基于场景用户触点提供不同的落地形态。

超聚变 xRAY 智能服务一体机解决方案贯穿大模型全生命周期的安全管理，旨在从基础安全措施、数据集管理、开发流程以及运行监控等多维度降低应用风险，保障模型安全可靠运行。其中模型安全管理涉及个人隐私过滤、企业敏感内容检测、有害内容合规检测等。个人隐私过滤通过关键词、正则表达式、自然语言处理和模糊匹配等技术，精准识别隐私数据；企业敏感内容检测则借助基于规则的检测、机器学习、上下文分析和灵活策略制定，守护企业数据安全；有害内容检测则依靠内容合规检测、语义理解及实时监控与审计，过滤有害信息。这些技术和工具共同构建起 AI 安全治理体系，有效管理 AI 应用中的安全风险，保护个人隐私和企业敏感信息，确保生成内容符合道德和法律标准。

5.7 恒安嘉新大模型安全监测解决方案

在大模型应用迅速发展的背景下，针对大模型 API 调用的审计与风险检测成为保障业务稳定、合规运营的关键。恒安嘉新大模型 API 安全监测网关产品，专注于大模型 API 调用审计与风险检测，旨在帮助企业全面掌握 API 调用情况，及时发现并处理潜在风险，确保业务的稳定运行与合规发展。

恒安大模型安全监测网关产品整合了全面的审计功能、精准的风险检测能力、强大的统计分析功能以及卓越的技术性能，为企业、科研机构、大模型 API 服务提供商等各类用户在大模型应用的审计与风险管控方面提供了全方位的解决方案，能够有效提升业务稳定性、保障合规运营、优化资源分配、增强决策科学性，在充分发挥大模型技

术优势推动业务创新的同时，确保业务在安全、合规、高效的轨道上持续发展，从容应对大模型应用过程中的各类挑战与风险，把握数字化转型的时代机遇。

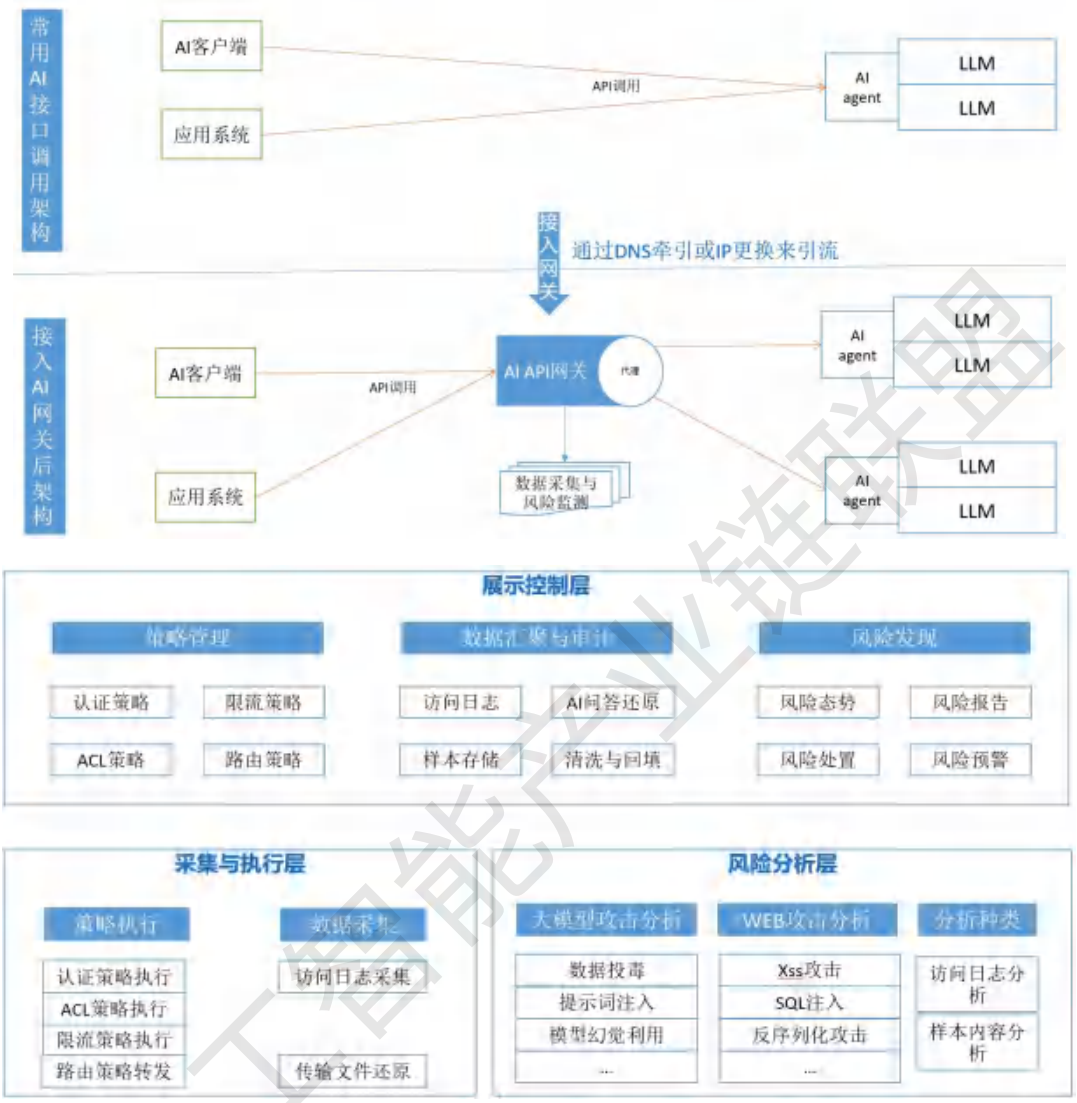

图 5.10 恒安大模型安全监测网关架构

5.8 浙江大学 AI 安全应用实践

5.8.1 人工智能安全评测平台 Aicert

Aicert 是在科技部科技创新 2030-“新一代人工智能”重大项

目、国家重点研发计划青年科学家项目、国家自然科学基金委区域创新发展联合基金重点项目等多个国家级/省部级项目以及浙江大学区块链与数据安全全国重点实验室的共同支持下研发的 AI 系统多个层面全方位覆盖的全栈安全评测平台，Aicert 的评测范围广泛，覆盖了数据、算法、模型、框架和系统等多个关键层面，为人工智能应用的稳定运行提供安全技术底座。

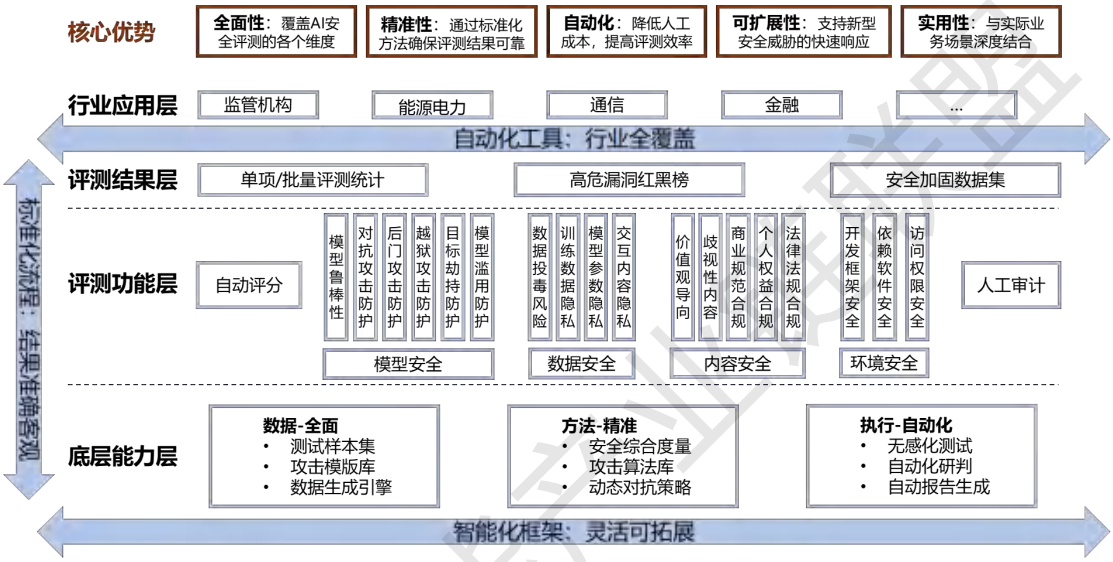

图 5.11 浙江大学人工智能安全评测平台 Aicert

AI 系统具有规模庞大、结构复杂、可解释性差等问题，同时 AI 安全评测面临风险来源多、模型架构/模态异构程度高等技术挑战。以智轨交通风险评估为例，中车株洲所在列车无障碍检测、矿卡无人驾驶、智轨交通系统等领域广泛应用 AI，但面临攻击测试数据不足、安全漏洞种类多和算法鲁棒性差等问题。Aicert 针对性的为列车轨道驾驶、矿卡室外作业等 10 余个场景增加了 10 余万测试样本，为多种任务生成 20 万攻击用例，针对轨道系统挖掘出 26 个框架漏洞，其中 7 个为高危漏洞；服务于网信办针对 16 个大模型开展安全评测，

覆盖语言、视觉等不同模态大模型，设立 6 大类安全评估指标，其安全性直接关系到产业发展和用户权益保护；针对南方电网所属的输电域大模型、市场营销域大模型、安监域关键工序检测模型进行了安全评估，发现部分模型在攻击环境下鲁棒性不足，可能会造成严重的生产安全风险，相关报告已经提交南方电网。

5.8.2 图像深度伪造检测平台 DFScan

图 5.12 浙江大学深度伪造评测平台 DFScan

DFScan 是在浙江大学区块链与数据安全全国重点实验室的支持下研发的图像深度伪造检测平台，基于千万级伪造图像数据底座，覆盖全脸合成、人脸替换、表情驱动、文生图/视频、音频驱动视频等全方位深度伪造内容检测，旨在促进生成式人工智能的安全稳定发展。平台聚焦现有检测算法泛化性差、鲁棒性不足等问题，构建千万级图像伪造数据库，囊括不同人种、性别、年龄、拍摄角度光照等差异化数据，支持对现实世界检测任务的场景化数据增强，提出增量学习框架、高精度换脸检测、伪造特征解耦等自研检测技术，支持省公安厅

开展电信反诈、物证鉴伪等信息网络犯罪预警防控工作。

5.8.3 大模型内容水印标识 GCmark

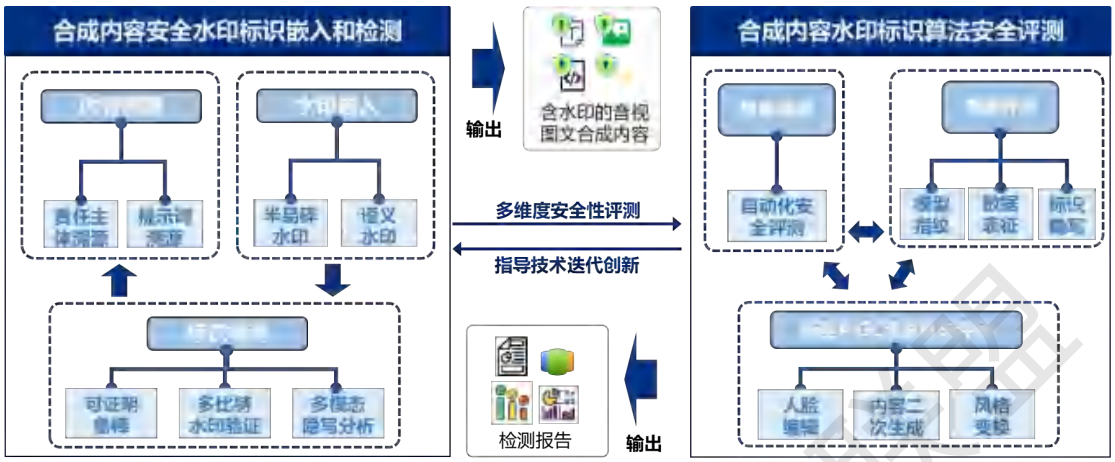

图 5.13 浙江大学大模型内容水印标识 GCmark

GCmark 是一款大模型合成内容水印标识产品，为政企提供基于数字水印的合成内容滥用问题解决方案。GCmark 汇总了相关政策标准文件和实践指南，主要为生成合成服务提供者和内容传播服务提供者提供编码辅助生成、编码规则校验、多模态标识合规性检测等服务功能。平台提供针对人工智能合成内容标识的合规性预检测服务，支撑监管部门对使用人工智能生成合成内容的制作与传播平台进行评测与监督，以及人工智能生成合成文本、音频、图像及视频内容高鲁棒、防篡改的标识与提取服务，帮助人工智能生成合成内容的制作与传播平台进行合规性自查。GCmark 服务政府和国安相关部门、企业、媒体与直播平台及模型开发者多方需求，构建 AIGC 安全协同生态；面向企业提供标识预检测与合规打标两类产品，提供平台和 API 两种形态，支撑企业满足合成内容标识强制性标准要求。

六 AI 安全治理发展建议

在人工智能安全治理体系下，本白皮书从完善 AI 安全法律法规、及 AI 安全标准建设、持续展开 AI 安全技术攻关以及 AI 安全人才培养等多个维度提出人工智能安全治理发展建议。

（1）完善法律法规及标准治理体系

通过完善 AI 安全治理的法律法规，明确 AI 应用开发的全生命周期的安全责任主体与合规要求，建立分级分类的安全审查机制和风险预警体系，尤其是针对垂直应用场景，构建风险分析框架与分类分级标准，为 AI 产业健康发展提供坚实的制度保障。同时，建议加快建立涵盖数据、算法、模型到应用场景的全流程 AI 安全标准，重点针对智算基础设施、模型训练与推理等关键环节制定可量化评估指标和可参考的研发应用流程，通过产业协同形成具备共识的通用安全基线和针对各垂直领域的安全应用指南，促进 AI 技术研发与产品落地的规范化。

（2）围绕 AI 安全前沿技术开展系统攻关

通过围绕 AI 安全治理的核心前沿技术开展系统性攻关，一方面重点突破模型鲁棒性与泛化增强、对抗样本检测与防御、大模型安全对齐、大模型幻觉减轻等关键技术，构建面向真实场景的动态防御体系；另一方面构建全生命周期的 AI 安全监测体系，融合模型水印和溯源技术，形成从研发到部署的闭环防护能力，为 AI 安全治理提供坚实的安全技术支撑。

（3）强化复合型人才培养与产学研协同创新

通过设立 AI 安全交叉学科，以“AI 技术+AI 安全”双核能力为导向，培育兼具 AI 算法开发与安全防御思维的复合型人才；推动高等院校、研究所与企业共建联合实验室和测试验证平台，加速攻防演练、AI 系统安全评估等技术的产学研转化。

缩略语

缩略语	英文全称	中文释义
AI	Artificial Intelligence	人工智能
AGI	Artificial General Intelligence	通用人工智能
DL	Deep Learning	深度学习
ML	Machine Learning	机器学习
LLM	Large Language Model	大语言模型
RAG	Retrieval-Augmented Generation	检索增强生成
API	Application Programming Interface	应用程序接口
AIGC	Artificial Intelligence Generated Content	人工智能生成内容
GPU	Graphics Processing Unit	图形处理单元
APP	Application	应用程序
MFA	Multi Factor Authentication	多因素验证
IoT	Internet of Things	物联网
AIoT	Artificial Intelligence of Things	智能物联网
CoT	Chain of Thought	思维链
MCP	Model Context Protocol	模型上下文协议
A2A	Agent to Agent Protocol	代理对代理协议

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems, 2012, 25.

[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing systems, 2017, 30.

[4] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[5] Kaur D, Uslu S, Rittichier K J, et al. Trustworthy artificial intelligence: a review[J]. ACM Computing Surveys, 2022, 55(2): 1-38.

[6] Li B, Qi P, Liu B, et al. Trustworthy AI: From principles to practices[J]. ACM Computing Surveys, 2023, 55(9): 1-46.

[7] Wang J, Li H, Wang H, et al. Trustworthy Machine Learning: Robustness, Generalization, and Interpretability[C] Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023: 5827-5828.

[8] Liu H, Wang Y, Fan W, et al. Trustworthy ai: A computational perspective[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 14(1): 1-59.

[9] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[10] Yu M, Meng F, Zhou X, et al. A survey on trustworthy llm agents: Threats and countermeasures[J]. arXiv preprint arXiv:2503.09648, 2025.

[11] He F, Zhu T, Ye D, et al. The emerged security and privacy of llm agent: A survey with case studies[J]. arXiv preprint arXiv:2407.19354, 2024.

[12] Xing W, Li M, Li M, et al. Towards robust and secure embodied ai: A survey

on vulnerabilities and attacks[J]. arXiv preprint arXiv:2502.13175, 2025.

[13] Sun L, Huang Y, Wang H, et al. Trustllm: Trustworthiness in large language models[J]. arXiv preprint arXiv:2401.05561, 2024, 3.

[14] Wang K, Zhang G, Zhou Z, et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment[J]. arXiv preprint arXiv:2504.15585, 2025.

[15] Sapkota R, Roumeliotis K I, Karkee M. Vibe Coding vs. Agentic Coding: Fundamentals and Practical Implications of Agentic AI[J]. arXiv preprint arXiv:2505.19443, 2025.

[16] Shi D, Shen T, Huang Y, et al. Large language model safety: A holistic survey[J]. arXiv preprint arXiv:2412.17686, 2024.

[17] 陈钟;谢安明.人工智能安全挑战及治理研究[J].中国信息安全,2023(05):32-35.

[18] 中国人工智能学会.中国人工智能系列白皮书——大模型技术(2023 版)[R].2023.

[19] 清华大学，中国信息通信研究院，蚂蚁集团.可信 AI 技术和应用进展[R]. 2023.

[20] 蚂蚁集团.人工智能安全白皮书（2020）[R]. 2020.

[21] 之江实验室.生成式大模型安全与隐私白皮书[R]. 2023.

[22] 全国信息安全标准化技术委员会.人工智能安全标准化白皮书（2019 版）[R]. 2019.

[23] 全国信息安全标准化技术委员会.人工智能安全标准化白皮书（2023 版）[R]. 2023.

[24] 中国智能算力产业联盟，人工智能算力产业生态联盟，商汤科技智能产业研究院.新一代人工智能基础设施白皮书[R]. 2023.

[25] 中国科学技术信息研究所.人工智能计算中心发展白皮书[R]. 2020.

[26] 中国科学技术信息研究所.人工智能计算中心发展白皮书 2.0 — 从人工智能计算中心走向人工智能算力网络[R]. 2020.

[27] 新华三集团，中国信息通信研究院.2023 智能算力发展白皮书[R]. 2023.

[28] 中国信息通信研究院.中国算力发展指数白皮书（2023 年）[R]. 2023.

[29] 工业和信息化部，中央网络安全和信息化委员会办公室，教育部，国家卫生健康委员会，中国人民银行，国务院国有资产监督管理委员会.算力基础设施高质量发展行

动计划[R]. 2023.

[30] 国家信息中心，阿里云. 人工智能 2.0 时代的公共智算服务发展指南[R]. 2023.

[31] 中国智能计算产业联盟. 国家“东数西算”工程下算力服务发展研究报告[R]. 2023.

[32] 国家信息中心信息化和产业发展部. 智算中心规划建设指南[R]. 2020.

[33] 国家信息中心. 智算中心创新发展指南[R]. 2023.

[34] 首届全球人工智能（AI）安全峰会. 布莱切利宣言[R]. 2023.

[35] 深圳市科学技术协会. 生成式人工智能安全与全球治理报告[R]. 2024.

[36] 中国信息通信研究院,腾讯科技有限公司. 数据安全治理与实践白皮书[R]. 2023.

[37] 金杜律师事务所，上海人工智能研究院等. 大模型合规白皮书[R]. 2023.

[38] 百度安全. 大模型安全解决方案白皮书[R]. 2023.

[39] 中国联通研究院. 中国联通算力网络安全白皮书（2024）[R]. 2024.

[40] 清华大学,中关村实验室,中国信息通信研究院,蚂蚁集团. 大模型安全实践(2024)[R]. 2024.

[41] 腾讯研究院，清华大学，浙江大学. 大模型安全与伦理研究报告 2024[R]. 2024.

[42] 中国科学院，公安部第三研究所，蚂蚁安全实验室. 生成式大模型安全评估白皮书.[R]. 2024.

[43] 中国信息通信研究院，中国人工智能产业发展联盟. MaaS 框架与应用研究报告（2024 年）[R]. 2024.

[44] 安永（中国）企业咨询有限公司，上海市人工智能与社会发展研究会. 可信人工智能治理白皮书[R]. 2025.

[45] 方滨兴. 人工智能安全[M]. 电子工业出版社，2020.

[46] 王琦,朱军,王海兵. 人工智能安全:原理刨析与实践[M]. 电子工业出版社,2022.

[47] 腾讯朱雀实验室. AI 安全：技术与实战[M]. 电子工业出版社，2022.

[48] 陈左宁，卢锡城，方滨兴. 人工智能安全[M]. 电子工业出版社，2022.

[49] UC Berkeley, Google, OpenAI. Unsolved Problems in ML Safety[R]. 2022.

中国联通研究院是根植于联通集团（中国联通直属二级机构），服务于国家战略、行业发展、企业生产的战略决策参谋者、技术发展引领者、产业发展助推者，是原创技术策源地主力军和数字技术融合创新排头兵。联通研究院致力于提高核心竞争力和增强核心功能，紧密围绕联网通信、算网数智两大类主业，按照 4+2+X 研发布局，开展面向 C3 网络、大数据赋能运营、端网边业协同创新、网络与信息安全等方向的前沿技术研发，承担高质量决策报告研究和专精特新核心技术攻关，致力于成为服务国家发展的高端智库、代表行业产业的发言人、助推数字化转型的参谋部，多方位参与网络强国、数字中国建设，大力发展战略性新兴产业，加快形成新质生产力。联通研究院现有员工 700 余人，85%以上为硕士、博士研究生，以“三度三有”企业文化为根基，发展成为一支高素质、高活力、专业化、具有行业影响力的人才队伍。

战略决策的参谋者
技术发展的引领者
产业发展的助推者

态度、速度、气度

有情怀、有格局、有担当

中国联合网络通信有限公司研究院
地址：北京市亦庄经济技术开发区北环东路 1 号
电话： 010-87926100
邮编： 100176

AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩

- Zero
  - AI音乐创作人
  - 水墨动漫联盟创始人
  - 百脑共创联合创始人
  - 人工智能产业链联盟创始人
  - 中关村人才协会秘书长助理
  - 河北北大企业家分会秘书长
  - 墨玫星辰智能科技有限公司CEO
  - 河北清华发展研究院智能机器人中心线上负责人
  - 中关村人才协会数字体育与电子竞技专委会秘书长助理
- 主要业务:AI商业化答疑及课程应用场景探索，各类AI产品学习手册，答疑及课程

欢迎扫码交流

提供：学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态

人工智能产业链联盟创始人

邀请你加入星球，一起学习

人工智能产业链联盟报告库

星主：人工智能产业链联盟创始人

每天仅需0.5元，即可拥有以下福利！
每周更新各类机构的最新研究成果。立志将人工智能产业链联盟打造成市面上最全的AI研究资料库，覆盖券商、产业公司、研究院所等...

知识星球

微信扫码加入星球 ▶