大模型
安全白皮书

安全·向善·可信·可控

让AI世界更安全更美好

引言：
外筑内固，构建大模型安全全链路防护体系

据IDC预测，到2030年，人工智能将为全球带来22.3万亿美元经济收入，大模型正驱动百行千业智能化变革，而开源大模型在性能提升、部署成本降低的推动下，加速向政务、金融、能源等重点行业落地，安全风险也随之渗透到全生命周期：从模型层的提示注入、越狱攻击，到数据层的敏感信息泄露，再到应用层的智能体越权，任何一处漏洞都可能威胁个人、企业国家的财产安全，甚至是生命安全。清晰的网络安全边界正不断消融、趋于模糊，并延伸至大模型的全链路安全，传统基于规则和特征匹配的防御体系已然失效。

我们正在从传统的“网络安全”时代，迈入以“大模型安全”为核心挑战的新阶段。当前威胁呈现指数级演化态势：一方面，攻击面急剧扩大，针对算力基础设施的劫持、供应链中的恶意模型文件、以及利用LangChain等框架漏洞的新型攻击，表明风险已深度嵌入技术底座。另一方面，攻击主体高度“平民化”，自然语言取代专业代码成为攻击武器，提示注入、越狱攻击让“全民黑客”从概念走向现实，极大降低了网络犯罪门槛。与此同时，模型固有的“幻觉”问题、智能体在工具调用中的越权风险、以及数据泄露与知识污染，共同构成了一个多维、动态、交织的复杂威胁矩阵。

面对这一全球性挑战，中国开创了“发展与安全并重”的敏捷治理之路，以《生成式人工智能服务管理暂行办法》为代表的“包容审慎、分类分级”原则，为技术创新与风险防控提供了动态平衡的框架。

在此背景下，360主张：必须采用以AI对抗AI、原生融合安全的新范式来应对大模型时代的安全挑战。我们提出“外筑‘以模治模’动态屏障，内固‘平台原生’安全底座的核心理念，将安全能力内嵌于大模型的构建、训练、部署与运营的全过程。通过构建一个纵深防御、动态演化的全景框架，从基础设施安全、内容与价值对齐、幻觉缓解到智能体行为管控，实现从“网络安全”到“大模型安全”的范式升维，为人工智能时代提供一个“安全、向善、可信、可控”的底座。

核心观点

本白皮书系统性地论证了大模型安全正经历从传统网络边界防御到原生、全栈、智能动态防御的范式转移，并提出“外筑"以模治模"动态屏障，内固"平台原生"安全底座”的核心理念，覆盖“安全、向善、可信、可控”四大支柱的全景安全框架，构建智能时代的核心免疫系统。

1. 风险之变：从边界防御到全栈免疫

大模型安全风险是系统性、全栈式的。它贯穿基础设施、模型层、数据层、智能体层及用户端五大层次，具体表现为算力劫持、供应链投毒、内容越狱、模型幻觉、知识污染、隐私泄露、行为失控以及工具滥用等诸多方面，共同构成了一个传统安全方案无法应对的复杂威胁立体空间。

2. 攻击之变：从专业黑客到全民黑客

攻击技术持续向高端化演进，例如针对框架和基础设施的深度利用；与此同时，攻击主体则日趋平民化，“自然语言黑客”的出现极大降低了攻击门槛。这导致攻防不对称性加剧，必须发展出同等智能、动态感知的防御体系。

3. 治理之智：中国特色的“发展与安全”平衡术

中国的治理模式采用“发展与安全并重”的动态平衡策略。通过“包容审慎、分类分级”的监管框架，既划定安全底线，又为技术快速迭代预留了弹性空间，为产业创新提供了关键的政策窗口期。

4. 应对之道：以模治模+平台原生安全

360提出的“以模治模+平台原生安全”是应对新范式的技术必然。它通过专用安全大模型，例如风险检测、幻觉纠正与红蓝对抗模型，对抗通用大模型风险，它既具备“外挂式”插件的灵活快速，又兼具“原生式”的深度安全能力，实现了双向赋能的协同防御，构筑起从模型、数据、内容到行为的全链路深度防护体系。

5. 生存之道：开放共生，生态共治

大模型安全的复杂性、全局性，决定了仅靠单一力量的技术或资源难以实现全链路、全场景的大模型安全治理。必须通过标准共建、产学研协同——例如开源安全模型、共建联合平台，以及组建产业联盟等形式，汇聚各方力量，共同打造智能时代可信、向善的安全基底。这既是产业发展的必然要求，也是国家层面的战略需求。

法律声明

三六零数字安全科技集团有限公司（或称“360”）提醒您在阅读或使用本文档之前，仔细阅读、充分理解本法律声明的全部内容。您对本文档的任何阅读或使用行为，即视为您已认可并同意接受本声明的全部约束。

1.文档获取与使用

您应通过360官方指定网站或360授权的其他正式渠道下载、获取本文档。本文档仅可为自身合法、合规的非商业性活动之目的而使用。

2.知识产权

未经360事先书面许可，任何单位或个人不得擅自对本文档的任何内容（包括但不限于文字、图表、数据、架构设计）进行包括但不限于篡改、翻译、复制、发行、或以其他任何形式传播。本文档所涉及的所有内容，包括但不限于商标、专利、著作权、商业秘密等知识产权，均归360或其关联公司依法所有。

360保留本文档中未明确列明的所有权利。

3.文档更新

鉴于技术、产品、法律与政策的持续演进，360保留在不事先通知的情况下，对本文档内容进行更新、修订或替换的权利。

4.免责声明

本文档仅作为技术参考与指引提供，所有内容按“现状”、“包含可能缺陷”及“当前功能”状态呈现。尽管360已力求文档内容的准确性与可靠性，但不对其准确性、完整性、适用性、及时性作任何明示或默示的保证。任何单位或个人因依赖或使用本文档而直接或间接遭受的任何损失（包括但不限于数据、收入、商誉损失），360及关联方均不承担法律责任。

本文档内容仅供参考，不构成法律、政策建议；不构成投资、商业决策依据；本文档引用的数据和观点不代表360立场；360不对引用资料的准确性、完整性承担保证责任。

5.遵守法律

您在使用本文档及其中所述技术时，应严格遵守《网络安全法》、《数据安全法》、《个人信息保护法》及生成式人工智能相关法规等中华人民共和国法律法规，并承担因使用不当所引发的一切法律责任。

6.反馈与联络

如您发现本文档存在任何错误、疑问或可能的侵权内容，请通过官方指定渠道与我们联系：
service-tech@360.cn

01

范式迁移：
从网络安全到大模型安全的时代挑战

02

大模型安全
威胁全景透视

03

源于实战：
大模型安全的应对新思路

04

360解决方案：
可全链路的安全防护

05

生态共治：
构建可信AI生态

目录

第一章

01

范式迁移：从网络安全到大模型安全的时代挑战

1.1 安全升维：
安全边界从网络扩展到模型全栈

- 攻击面扩大：大模型安全漏洞呈指数级增长，智能体成为新的攻击对象

“人工智能+千行百业”将带动新一轮工业革命，为高质量发展注入强大动能，引领人类社会进入智能化时代，为生产、生活方式带来巨大变革。大模型作为目前人工智能技术的核心引擎和技术底座，重塑着各行业的应用生态，其安全内涵已发生根本性转变。我们正在经历一场从传统“网络安全”向新型“大模型安全”的范式迁移。据ISC.AI 2025大会披露的数据，“大模型安全漏洞呈指数级增长”已成为现实。2025年9月16日，第22届中国网络安全年会暨国家网络安全宣传周网络安全协同防御分论坛活动中，国家计算机网络应急技术处理协调中心发布了国内首次针对AI大模型的实网众测检验结果，累计发现各类安全漏洞281个，其中大模型特有漏洞177个，占比超过60%。这充分表明，当前AI大模型产品面临着大量传统安全领域之外的新安全风险。针对大模型的新型攻击手段层出不穷。已从技术层面的单一威胁演变为系统性风险，包括提示注入攻击、敏感信息泄露、供应链组件风险、模型中毒攻击、模型拒绝服务等多样化攻击。

- 攻击者平民化：“全民黑客”时代来临，传统防御体系失效

当前，攻击者从专业黑客变为全民黑客，攻击主体的开始趋向平民化，以往需要精通机器语言的专业黑客，如今只需通过自然语言指令即可让大模型自动生成攻击代码、设计钓鱼邮件、实施社会工程学攻击。另一方面，企业为了实现大模型和智能体能够真正结合实践生产，往往需要将企业的数据知识训练到大模型、知识库中，而用户或员工仅需要“套话”的方式就能将企业核心数据套走。360集团创始人周鸿祎曾指出，“如今，从前台小文秘也能欺诈后台大模型”，普通人无需技术背景，仅凭简单的提示词就能发动专业级攻击，真正实现了“有手就行”的攻击平民化。这种“全民黑客”现象正使网络犯罪从技术壁垒走向大众化、产业化，给安全防御带来前所未有的挑战。大模型幻觉问题严重影响生成内容可靠性。由于大模型固有的技术特性带来的忠实性和事实性模型幻觉问题，会侵蚀生成内容可靠性基石，进而引发决策失误、信任危机并阻碍其在关键领域的深度应用。智能体安全风险进一步放大了威胁范围。涵盖了从底层模型到高层行为的多个维度，主要包括记忆篡改、提示词注入、敏感数据泄露、Agent越权与失控风险、工具调用风险，以及智能体仿冒、中间人劫持等多智能体风险，这些风险相互交织，形成了传统安全体系难以应对的复杂威胁矩阵。

1.2 挑战交织：
技术、数据与国际化构成安全核心阵地

中国在人工智能领域发展迅猛，但在技术、数据、国际化等多维度仍面临严峻挑战，这些挑战相互交织，构成了大模型安全发展的核心困境。

- 安全维度上

人工智能为网络攻击提供了新型渗透载体和手段，大模型本身成为"双刃剑"—既可作为防御工具，也可能被恶意利用为攻击平台。大模型需应对提示注入、模型越狱、RAG篡改等新型攻击手段，而AI技术大幅降低了网络攻击门槛，使"全民黑客"时代加速到来，传统安全防御体系难以应对AI大模型增强的规模化、自动化攻击。

- 数据维度上

人工智能进一步加大了维护数据安全和保护个人信息的难度，大模型训练需要海量数据但面临合规采集困境，模型可能无意泄露训练数据中的敏感信息，内容安全过滤难度显著增加，同时需平衡《网络安全法》、《数据安全法》、《个人信息保护法》等法规要求与技术创新需求。

- 技术维度上

人工智能为发达国家实施技术封锁提供了新机会，高端AI芯片受限、基础模型架构受限、开源生态受阻等问题凸显，中国大模型在算力资源、核心算法、训练数据等方面面临"卡脖子"风险。

- 国际化维度上

中国大模型出海需同时满足国内法规要求与目标市场监管环境，既要符合我国"安全与发展并重"的治理原则，又要适应欧盟GDPR、美国出口管制等多元监管体系，面临"双合规"压力与市场准入壁垒。这些挑战相互交织，要求中国大模型产业必须构建"内生安全"能力，突破核心技术瓶颈，建立自主可控的产业生态，同时积极参与全球AI治理，方能在保障安全的前提下实现高质量发展。

1.3 治理路径
中国走出发展与安全并重的敏捷治理之路

国家层面已明确提出“安全与发展并重”的原则，强调人工智能治理应“以人为本、智能向善”，既要防范安全风险，也要促进技术进步与应用。

- 多部门出台生成式人工智能相关法规，建立“包容审慎、分类分级”的监管体系

近年来，国家网信办等多部门联合发布了《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》、《生成式人工智能服务管理暂行办法》（以下简称《办法》）、《人工智能生成合成内容标识办法》等法规，明确了服务提供者主体责任，在算法推荐、内容合成、模型管理等方面提出规范要求。其中，《办法》建立了“包容审慎、分类分级”的监管体系，要求服务提供者履行安全评估、内容过滤、算法备案、模型备案以及产品登记等责任，并承担包括坚持社会主义核心价值观、防止生成歧视性内容、尊重知识产权、尊重他人合法权益等多项义务，该体系在为技术创新预留充分空间的同时，也充分体现了“安全与发展并重”的治理理念。

2025年10月28日，第十四届全国人民代表大会常务委员会第十八次会议通过了关于修改《中华人民共和国网络安全法》的决定，其中第二十条明确提出了“完善人工智能伦理规范，加强风险监测评估和安全监管，促进人工智能应用和健康发展”。

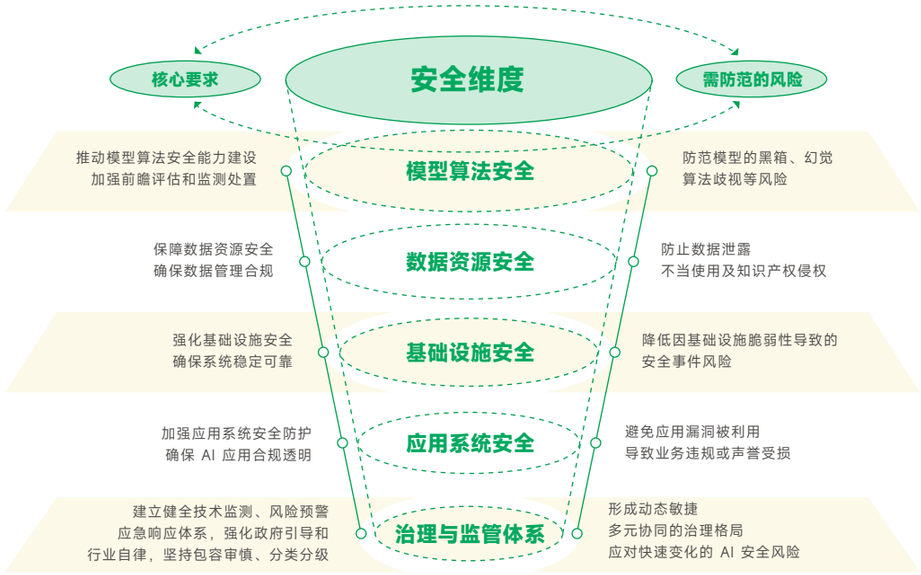
- 监管部门构建“事前-事中-事后”的全生命周期监管技术体系，开展大模型治理专项行动

2025年4月30日，中央网信办启动“清朗·整治 AI 技术滥用”专项行动。第一阶段强化AI技术源头治理，清理整治违规AI应用程序，加强AI生成合成技术和内容标识管理，推动网站平台提升检测鉴伪能力。第二阶段聚焦利用AI技术制作发布谣言、不实信息、色情低俗内容，假冒他人、从事网络水军活动等突出问题，集中清理相关违法不良信息，处置处罚违规账号、MCN机构和网站平台。多地网信部门对辖区大模型生成内容开展实时监测、内容安全评测等工作，利用大模型监管生成式人工智能服务，通报生成式人工智能服务存在的问题，以技术应对技术，以智能管理智能。

多地监管主管部门根据相关法规和职责，开展前瞻性先行先试，积极构建“事前-事中-事后”的全生命周期监管技术体系，在事前阶段对自研模型、微调模型开展上线备案审查，对调用大模型的AI产品开展登记备案审查，对大模型基础设施安全检测；在事中阶段，开展大模型内容安全技术监测、风险预警、攻防演练等，评估整体安全态势，跟踪变化趋势，对大模型进行动态管理；在事后阶段，开展研判分析、通报处置等工作，实现对大模型全生命周期的安全监管和治理。360大模型卫士系列产品能够满足监管用户实现对大模型全生命周期的安全监管和治理。

- 各产业与国家战略的同频共振

2025年8月21日，国务院印发《关于深入实施“人工智能+”行动的意见》（以下简称“意见”），在第十四条强调统筹重点领域的发展与安全，体现了国家加快创新与筑牢安全底线的战略考量，既重视技术风险防范，又推动治理机制完善，为各地落实“人工智能+”行动提供了明确指引，确保人工智能发展始终安全可控。



各区域和产业迅速响应，因地制宜制定实施方案，推动人工智能发展与安全统筹兼顾。河北省在《河北省推动“人工智能+”行动计划（2025—2027年）》中提出“一体推进研发攻关、应用迭代和生态培育”，在钢铁、化工、汽车制造等八大产业推动行业大模型应用并建立安全评估机制。国家发改委、能源局联合发布的《关于推进“人工智能+”能源高质量发展的实施意见》也强调，在推动电网、发电等场景智能化的同时，提升能源领域人工智能技术安全应用水平。

各区域和产业落实《意见》，坚持“统筹发展与安全”，结合区域和产业特点，通过建立安全评估体系、强化风险预警、完善标准规范，确保人工智能赋能产业的同时守牢安全底线，实现国家战略与地方需求的同频共振，为培育新质生产力和高质量发展提供了坚实支撑。

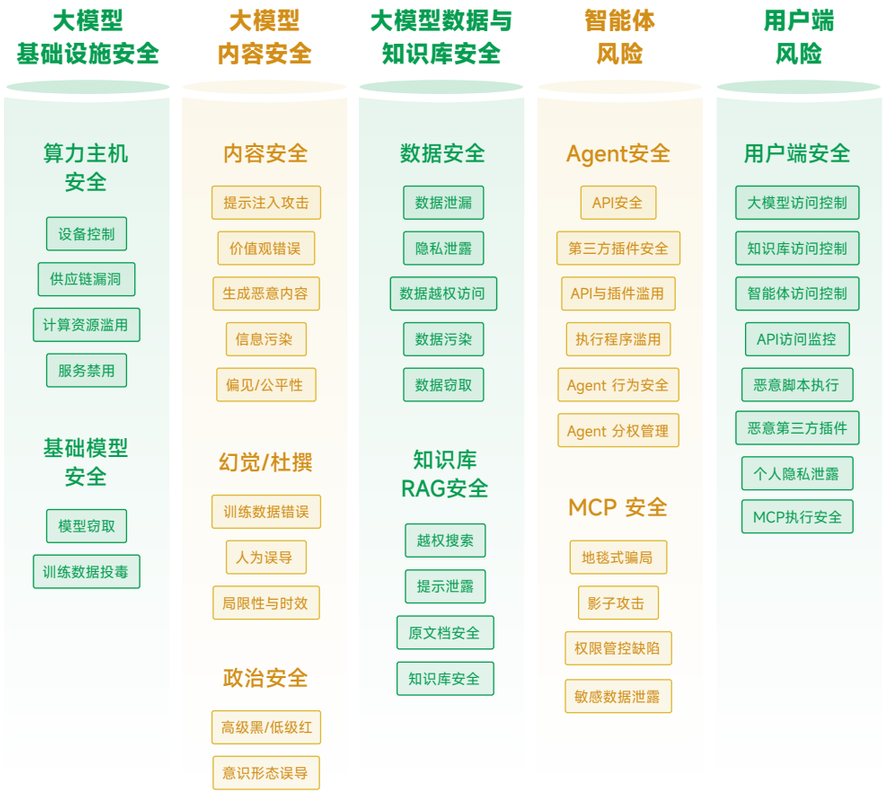
第二章

02

大模型安全
威胁全景透视

2.1 大模型安全呈多维复杂态势 挑战超越传统安全边界

在AI技术，特别是大模型快速发展与广泛应用的背景下，大模型安全风险已形成多维度的复杂体系。除了传统的网络安全与数据安全解决方案能够覆盖的大模型应用环境上的安全问题以外，大模型在运行时的安全风险尤其具有其独特性。大模型运行时安全风险主要涵盖以下五个关键风险点：一是大模型基础设施安全，涉及算力主机的设备控制、供应链漏洞及基础模型的窃取与数据投毒等问题；二是大模型内容安全，包含内容层面的提示注入、恶意生成，幻觉杜撰带来的信息误导，以及政治维度的意识形态风险；三是大模型数据与知识库安全，聚焦数据泄漏、隐私侵犯、知识库越权访问与信息污染；四是智能体安全，涉及 Agent 的 API 滥用、行为安全及 MCP 的投毒攻击、权限缺陷等；五是用户端安全，涵盖大模型、知识库、智能体的访问控制，API 监控及恶意插件、隐私泄露等风险。这些维度共同构成了AI安全治理需重点关注的全域图景。对于上述风险，传统的安全厂商尚未提供有效的解决方案。



2.2 大模型基础设施层风险： 算力与框架的“地基”隐患

大模型基础设施的安全风险是一个贯穿大模型生命周期的多层次威胁体系，它不仅继承了软件供应链、云服务、身份认证等的安全风险（如软件漏洞、沙箱逃逸、身份权限配置错误、API密钥泄露等），更因其特有的软件生态系统而引入了全新的、高价值的攻击向量。针对大模型基础设施的攻击已经从理论变为现实，具体可以分为三种类型：针对算力基础设施的攻击、针对开发环境的攻击以及针对在线智能服务的攻击。攻击者正积极地利用不同维度的安全缺陷，发起多起备受瞩目的安全事件并造成了重大的经济损失。

- 针对算力基础设施的攻击：算力劫持、资源滥用

“ShadowRay”攻击活动是最具代表性的算力基础设施入侵事件之一。攻击者利用了AI分布式框架Ray的一个关键架构缺陷（CVE-2023-48022）——其仪表盘和作业提交API在错误配置下无需身份验证即可公网访问。此次事件的后果极为严重，攻击者在全网扫描并入侵了数千台暴露的服务器，最直接的损失是大规模的计算资源劫持。他们部署了XMRig等加密货币挖矿软件，窃取了企业用于AI训练和推理的昂贵GPU算力（包括A100/H100），造成了巨额的经济损失。更深远的影响在于数据和知识产权的泄露，攻击者获得了服务器的完全控制权，使他们能够窃取专有的训练数据集、模型权重、以及AWS、GCP等云平台的API密钥，为进一步的内部渗透和企业间谍活动打开了通道。

- 针对开发环境的攻击：供应链投毒、恶意模型文件（以Hugging Face事件为例）

此外，模型供应链也成为针对大模型开发环境攻击的重灾区。Hugging Face平台上持续发现的“特洛伊木马”模型便是力证。这些攻击利用了大模型框架中长期存在的“不安全反序列化”漏洞（即pickle格式的滥用），实现对受害者系统“零点击”式的入侵，当开发者或自动化MLOps管道下载并使用恶意模型文件时，植入的恶意代码立即执行。之后攻击者可以在受害者的开发工作站或生产推理服务器上获得了持久化的后门，并进一步窃取敏感的内部数据、源代码、以及该环境中的其他专有模型。这类事件严重破坏了开源模型生态系统的信任基础，迫使Hugging Face等平台加强安全扫描。此外，当开发者在开发大模型业务时，通过加载开源社区提供的提示词模版（CVE-2023-36281）、IDE配置等文件时，同样可能会受到攻击。

- 针对在线智能服务的攻击：框架组件漏洞、API滥用（以LangChain漏洞为例）

针对在线智能服务的攻击同样日益增多，攻击者无需精心构造攻击数据，而只需要通过自然语言描述的方式实现攻击，如主流的服务构建框架LangChain成为了重灾区。2024年底至2025年间披露的多个漏洞（如CVE-2024-8309）显示，在LangChain的特定组件中触发经典的SQL注入或远程代码执行漏洞。这表明攻击者的策略正在发生转变，不再局限于攻击LLM模型本身，而是更精明地攻击那些用于连接LLM与数据库、API等外部工具的“胶水代码”和框架。

2.3 大模型内容风险：大模型的失控与越轨

随着大模型在各行业的深入应用，其潜藏的内容安全风险愈发突出。2025年2月，在思科旗下 Robust Intelligence 开展的安全评估中，DeepSeek R1在50条恶意提示测试中“全失守”，凸显了大模型在越狱攻防中的脆弱性。2024年5月，谷歌 Med-Gemini 在医学影像场景中“编造”不存在的解剖结构，属于典型的幻觉问题，若直接用于临床决策，可能造成严重误判。

这些案例共同揭示了当前大模型内容安全的三大核心挑战：内容合规风险（符合社会主义核心价值观、不包含歧视性、商业违法违规、侵犯他人合法权益等内容）、幻觉风险（内容不准确、不可靠）、越狱攻击风险（安全策略被绕过与滥用）。这要求政企在大模型应用中强化安全审查机制，建立检测、评估与防护并行的全链路内容安全防护体系。

2.4 大模型数据与知识库风险：知识“源泉”的污染与泄露

随着大模型在医疗、科研、日常服务等领域的深度渗透，其数据与知识库所潜藏的风险日益凸显，成为制约大模型安全应用的关键瓶颈，其中数据泄漏、知识库越权搜索、内容不可信三大风险尤为突出，需重点警惕。

- 数据泄露：训练数据、用户对话中的敏感信息泄露

数据泄漏风险频发，已成为大模型中的“心腹之患”。部分用户或机构因操作不当，如将涉密实验数据、个人隐私信息违规输入 AI 工具，或 AI 平台自身存在数据存储漏洞，导致核心技术参数、用户聊天记录、地理位置等敏感信息外泄。这些信息一旦流入黑市，不仅会让个人面临诈骗、骚扰等威胁，还可能致使企业核心算法被盗、国家关键领域数据失防，对个人权益、企业发展乃至国家安全造成严重冲击。应对此风险，需从技术与管理两方面发力，通过端到端数据加密、建立严格的 AI 数据使用审批机制，堵住数据泄漏的漏洞。

- 知识库越权：RAG场景下的未授权访问与数据窃取

知识库越权搜索则打破了数据访问的“安全边界”。在 RAG 等大模型应用场景中，企业为搭建私有知识库，会将分散在员工手中的文档集中整合，这虽提升了数据利用效率，却也放大了越权访问的风险。攻击者常利用大模型指令的模糊性，绕过常规防护机制，非法获取医疗病历、政务户籍、企业商业机密等敏感数据；部分内部人员也可能因权限管理疏漏，违规查阅、下载核心信息。此类行为不仅破坏机构数据管理秩序，还可能引发身份盗用、商业泄密等连锁问题，需通过建立动态权限绑定系统、完善操作审计日志追溯机制，筑牢知识库的“防护墙”。

- 内容不可信：输入错误知识导致生成垃圾信息

内容不可信问题则严重冲击大模型的“公信力”。由于大模型训练数据可能掺杂错误信息、过时内容，且生成逻辑存在固有漏洞，大模型时常出现“一本正经地胡说八道”的伪专业输出。在医疗场景中，大模型依据过时医学指南给出错误用药建议，可能延误患者治疗；在法律领域，大模型生成的合同条款忽略地域法规差异，会让企业陷入巨额纠纷；在新闻传播中，大模型编造的虚假信息短时间内就能扩散，引发社会恐慌。要解决这一问题，需从源头优化训练数据质量，建立大模型内容溯源机制，同时加强行业监管，提升公众对大模型输出内容的辨别能力，让大模型输出更可靠、更可信。

2.5 智能体行为风险：失控的“数字员工”

智能体（Agents）的安全应用需警惕多维度风险，这些风险相互交织，若缺乏有效管控，将严重威胁其运行安全与数据保护。

- 工具调用

在工具调用层面，智能体依赖的 Web 搜索、第三方插件、代码执行、API、MCP 等工具均暗藏隐患。第三方插件可能存在安全漏洞，成为恶意攻击的入口；代码执行功能可支持任意操作，若被滥用易引发系统破坏；API 调用若权限管控不当，可能造成敏感数据泄露或违规访问外部资源；而 MCP 协议本身也存在投毒、恶意代码植入等安全问题，进一步放大工具调用风险。

- 权限与行为

权限与行为层面的风险同样突出。部分智能体运行权限过大，未遵循“最小权限”原则，易出现违规调用工具、窃取企业核心数据等越权行为；同时，大模型可能因对任务指令理解错误产生“幻觉”，导致智能体偏离预设目标，错误执行关键操作，造成不可逆损失。

- 全流程与数据权限

全流程与数据权限维度也存在漏洞。智能体从开发、发布、审核到应用的全流程涉及开发者、审核员、使用者等多类角色，若未做好分权管理，可能出现未审核即上线、非授权人员篡改配置等问题；此外，智能体调用数据库、知识库时，开发者与使用者的权限边界模糊，未明确界定“谁能看、谁能改、谁能调用”，易导致数据访问失控，引发信息泄露风险。

- MCP安全

MCP 虽具备强大能力，但其应用过程中潜藏着三类高风险攻击场景，需重点警惕：一是投毒攻击，攻击者会将有害命令隐秘嵌入 MCP 工具的描述信息中，这些指令对用户完全不可见，却能暗中诱导 AI 模型执行诸如数据篡改、系统入侵等危险操作；二是地毯式骗局，MCP Server 在初期应用规模较小时，会展现正常功能以获取信任，一旦用户量或应用范围扩大，便会悄然植入恶意代码，实施批量攻击；三是影子攻击，即 MCP 自身虽无恶意设计，却因调用了存在安全隐患的第三方服务而间接引发风险。

2.6 用户端与入口风险：最后一道防线的失守

AI 用户端在应用过程中面临多维度安全风险，若不加以有效管控，将对用户数据、系统稳定及业务安全造成严重威胁。

- 访问控制

在访问控制维度，存在多重隐患：大模型访问方面，若身份验证机制存在漏洞，易被攻击者利用实施未授权调用，不仅引发计算资源滥用，还可能导致企业敏感数据泄露；知识库访问控制不足时，恶意用户可绕过权限限制，非法获取机密知识或对知识库进行数据污染，破坏知识体系的完整性与可靠性；智能体访问控制缺失则可能造成权限失控，使智能体被恶意操纵，执行窃取数据、破坏系统等危险任务；API 访问监控薄弱会导致异常流量难以被及时识别，为数据窃取、恶意攻击等行为提供可乘之机。

- 执行环境的安全

在执行环境安全层面，风险同样突出：恶意脚本可利用客户端运行环境的漏洞，注入钓鱼、勒索等恶意代码，劫持客户端功能或窃取用户信息；恶意第三方插件因缺乏严格的审核机制，可能携带后门程序或捆绑恶意软件，一旦安装将威胁整个系统的安全稳定。

- 隐私与协议安全

隐私与协议安全也不容忽视：个人隐私方面，若数据采集、存储和传输环节防护存在缺陷，用户的身份信息、行为习惯等敏感数据易被非法获取，侵犯用户隐私权；MCP 协议执行过程中，可能遭遇命令注入、权限逃逸等攻击，破坏客户端运行逻辑，进而危害整体系统的稳定性与数据安全。

综上，需从访问控制强化、执行环境加固、隐私保护升级、协议安全优化等多个角度构建防护体系，才能有效化解 AI 客户端的多维度安全风险，保障其安全可靠运行。

第三章

源于实战：
大模型安全的应对新思路

3.1 核心理念：外筑“以模治模”动态屏障，内固“平台原生”安全底座

面对 AI 大模型从基础设施到配套组件的全链路安全挑战，360 通过“以模治模”打造动态防御的外部屏障，以“平台原生安全”筑牢组件自带的安全底座，二者协同形成覆盖大模型全生命周期的安全防护能力，确保大模型“安全、向善、可信、可控”。

- 以模治模：用大模型对抗AI安全风险

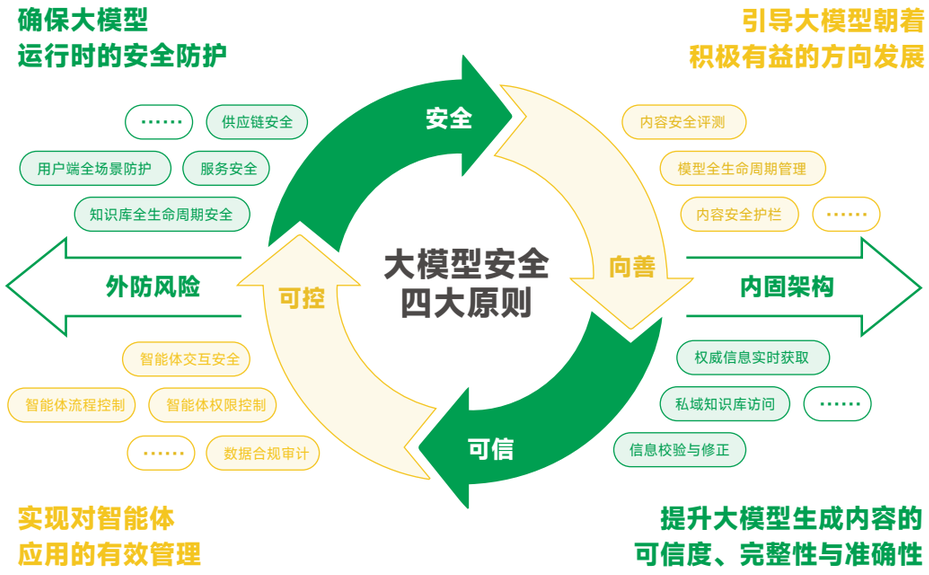
明确以大模型技术对抗大模型风险的创新路径，即以AI大模型之力反哺大模型安全防护，构建纵深防御与动态演化的全链路大模型安全防护体系。

- 平台原生安全：让大模型配套组件自带安全基因

强调将安全能力融入知识库、智能体等配套组件的底层架构，从设计阶段规划安全防护能力，夯实大模型整体安全的基础前提。

3.2 安全四原则：安全、向善、可信、可控

360作为网络安全领域的重要力量，提出大模型安全“安全、向善、可信、可控”四大原则，为大模型安全发展筑牢核心竞争力，护航AI时代稳健前行。



- 【安全】：“安全”原则确保大模型运行时的安全防护。

在数字世界，网络攻击、数据泄漏、个人隐私泄漏等风险如影随形，大模型系统也面临着诸多安全挑战。360通过保证大模型运行时安全，降低各类安全风险，同时提升攻击防护能力，为大模型打造坚固的安全壁垒；将安全基因融入知识库，提供数据采集、处理、存储、应用、流转至销毁的全生命周期防护方案，规避高频知识调用中的数据和知识泄露风险；基于20年终端安全技术积累，打造集成智能体沙箱、行为隔离、威胁管控、MCP 安全及身份认证的智能体客户端，构建用户使用智能体应用时的全场景防护。

- 【向善】：“向善”原则引导大模型朝着积极有益的方向发展。

一方面，要提升应对提示注入攻击的能力，防止恶意人员通过恶意提示诱导大模型生成有害内容。另一方面，从内容和能力两方面确保大模型“向善”，内容上确保生成内容符合社会道德伦理和法律要求，能力上避免大模型被用于生成违规内容、伪造图片视频、恶意代码、钓鱼邮件等。这使得大模型不只是技术工具，更成为符合社会价值规范的“向善”力量，在为人们提供服务时，始终传递积极、合法、符合道德的价值 。

- 【可信】：“可信”原则致力于提升大模型生成内容的可信度、完整性与准确性。

大模型天生会出现“幻觉”问题，生成与事实不符的内容，这会极大影响其可用性与可信度。360聚焦内容可信与完整可用，着力降低大模型“幻觉”问题，让大模型生成的内容更可靠、更准确。无论是用于信息获取、内容创作还是决策辅助，可信的内容能让用户更放心地依赖大模型，推动大模型在各个领域发挥更有效的作用，成为人们可以信赖的智能伙伴。

- 【可控】：“可控”原则实现对智能体应用的有效管理。

通过Agent框架安全控制，保障智能体在交互等场景下的安全；确保人在决策回路，避免出现“不可撤销”的后果，让人类能对智能体的关键行为进行干预与决策；全程审计则对智能体的行为做全过程监控审计，及时发现并纠正可能出现的问题。这一系列措施让智能体始终处于安全、合规的管控之下，使其发展与应用能更好地契合人类的需求与社会的规范。

3.3 全景框架：“外挂式安全 + 平台原生安全”的双轨安全防护体系

在人工智能飞速发展的今天，大模型的安全问题愈发受到关注。360推出大模型卫士，以“以模治模”为核心，构建起全面的安全保障体系，为AI安全保驾护航。针对大模型算力主机、模型基础设施、模型内容等核心环节的风险，可通过大模型安全类产品直接解决，这类问题属于“外部可干预”的安全范畴，具备明确的产品化应对路径；大模型应用中涉及的知识库、智能体、客户端等配套组件，其安全无法依赖外部大模型安全产品解决，组件需自身具备原生安全能力，才能从源头规避漏洞（如数据泄露、权限失控等），这是保障大模型整体安全的基础前提。360 凭借业内独有的“AI + 安全”双重基因，采用“外挂式安全 + 平台原生安全”双轨策略，实现“外防风险、内固架构”的全面防护。

（1）外挂式安全：以“以模治模”构建大模型外部防护屏障

作为大模型的“外部安全屏障”，外挂式安全以“以模治模”为创新核心，聚焦大模型运行的“基础设施层”与“内容安全”：

- 针对算力基础设施（硬件及软件设施），通过专用算力主机安全系统监测主机运行风险，规避硬件故障、非法入侵操作系统等问题；
- 针对大模型内容安全，利用 AI 模型对输出内容、输入数据进行实时检测，防范恶意指令、敏感信息泄露等风险，为大模型搭建“即时响应”的外部防护网。

（2）平台原生安全：从底层架构夯实内部安全基础

原生安全深度融入大模型运行所需组件中，聚焦“配套组件安全”与“全流程管控”，解决“组件自身安全 + 全链路合规”问题：

- 组件安全能力适配：支持用户端的异常行为控制、身份认证和智能体沙箱隔离等安全功能；支持知识库、智能体的全生命周期风险管理，内置数据泄露监控、分级分类权限等功能，确保配套组件从设计阶段就具备安全属性；
- 全流程管控机制：通过多角色分权管控、操作行为审计等功能，覆盖知识库和智能体的全流程，实时追溯风险行为，从底层架构杜绝越权访问、数据污染等隐患。

第四章

04

360解决方案：
可全链路的安全防护

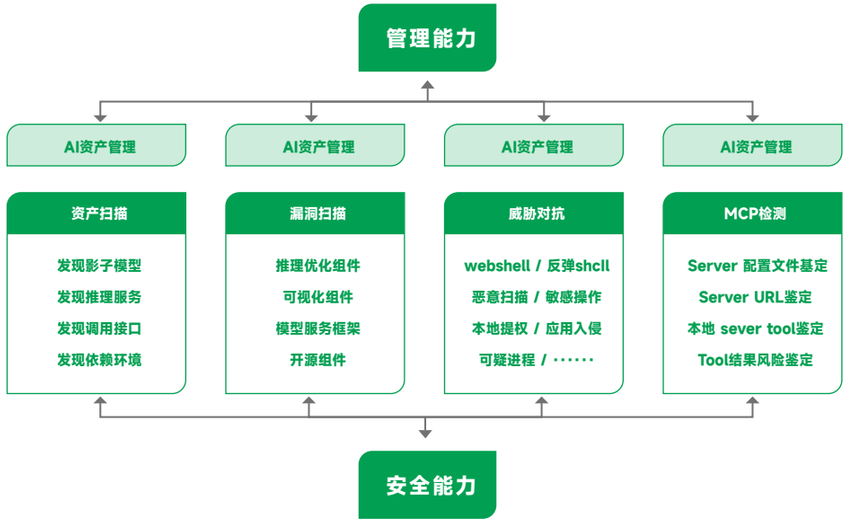
4.1 “外挂式”安全：以“以模治模”构建大模型外部防护屏障

“外挂式”安全聚焦大模型运行基础设施层与内容交互层的风险，这类问题的共性是“需外部工具实时监测、干预才能规避”。“外挂式”安全的设计核心是“不侵入大模型原生架构，通过外部工具实现灵活防护”，其必要性在于：一是能作为独立模块快速对接不同企业的大模型架构与运行环境，避免嵌入原生平台的重复开发高成本，适配更灵活；二是可通过独立监测引擎与响应机制实现毫秒级风险识别干预，满足基础设施攻击、内容风险等问题的实时防护需求。

4.1.1 360大模型卫士算力主机安全系统

【概述】

在生成式 AI 与大模型技术深度渗透企业核心业务的当下，大模型主机作为 AI 算力输出与业务运行的核心载体，正面临传统主机安全威胁与 AI 原生风险的双重夹击。360 大模型卫士算力主机安全系统面向大模型服务器场景打造的轻量化终端级防护软件，精准契合企业在 AI 规模化部署中对资产安全、运行可靠、合规可控的核心需求。为企业提供覆盖AI资产全生命周期的安全防护，实现从威胁感知到主动防御的一体化解决方案。系统的核心能力聚焦于四大维度：AI资产探测与画像，AI入侵检测与防御、AI漏洞检测与评估以及MCP检测与防御。



【产品能力】

(1) AI 资产发现专家

AI 资产发现并非单一动作，而是通过“人工补全 + 自动探测”的组合方式实现全面覆盖：一方面，由 AI 运维或开发团队按平台指定格式（如 Excel 表格、API 接口）手动导入内部已登记维护的 AI 资产信息；另一方面，在服务器、终端设备或云环境中部署轻量化 AI 探针，无需人工干预即可实时扫描设备资源、网络流量与进程活动，自动识别已部署的 AI 模型资产并发现“影子 AI”（含模型权重、推理服务、调用接口、依赖环境等），为后续防护提供资产可见性。

(2) AI 漏洞检测与评估专家

通过AI探针可自动扫描识别系统中 AI 服务组件（如 TensorFlow、PyTorch）及其直接 / 间接依赖库，实时检测已知安全漏洞（覆盖 CVE、CNNVD 等公开库及 360 安全运营中心漏洞）与过低版本（停维护或含漏洞的过期版本），并为每个风险项生成含漏洞编号、等级、影响范围等信息的提示，提供升级版本等针对性修复建议，同时支持持续监控与定时扫描，新依赖引入或版本变更时自动触发评估，保障 AI 应用全生命周期安全合规。

(3) MCP检测与防御专家

提供“运行前检测+运行防护”双模引擎，对每一次调用进行实时安全裁决，确保MCP协议的安全使用。在运行前对MCP环境进行监测，包括 MCP Server URL、Server配置文件、Server tools 等。在运行过程中监测 Client 对 Server 的访问风险、工具返回内容风险、工具执行风险等。

(4) AI入侵检测与防御专家

在AI主机侧部署安全检测与响应插件，对进程、系统调用进行7×24小时实时守护，自动对注入、非法调用、命令执行等链式入侵行为进行监测与防御。构建“实时感知—精准拦截—闭环响应”的纵深防御，帮助管理员即时发现威胁、降低风险并满足合规审计。

4.1.2 360大模型卫士检测系统

【概述】

当前大模型落地加速，安全风险呈“全链路渗透、高隐蔽性、强破坏性”特征，OWASP LLM Top10 风险高发，且政策强制要求企业落实 AI 安全责任，传统检测工具适配性不足，企业陷入“风险看不见、漏洞查不出、合规跟不上”困境。360 大模型卫士检测系统以“以模治模”为核心思路，通过专项训练的检测模型对抗大模型安全风险，全面覆盖 OWASP LLM Top10 安全威胁。该系统具备三大核心能力：一是大模型资产识别，精准梳理企业内部模型部署情况、版本信息及关联业务；二是全维度漏洞检测，从模型层（后门、隐私泄露漏洞）、应用层（接口权限漏洞）到业务层（输出合规风险）一站式检出隐患；三是交互式审计，支持对模型交互过程实时监控与风险追溯。最终通过全链路检测能力，帮助企业满足监管合规要求，为 AI 业务安全落地保驾护航。

【产品能力】

（1）模型梳理，理清模型使用现状

通过资产探测与数据对接，可全面发现组织内部模型使用情况以摸清家底，还能梳理大模型实际应用现状，包括识别 ollama 等近 200 个模型应用指纹、500 + 智能化业务指纹，以及自动化发现模型对外开放情况。同时按模型层、模型应用层、模型业务层分类，智能化梳理出清晰的模型应用概览，并记录安全措施与备案现状，满足监管要求，还可快速导出各类审计报告。

（2）专项检测，全面排查大模型服务及应用漏洞

专项检测可全面排查漏洞，含 200 + 大模型服务及应用专项 POC、5000 + 智能化应用及组件 POC，还能智能化编排任务提升扫描速率与检测能力，重大漏洞平均 8 小时内、HW 重保期间 5 小时内发布专项 POC。情报碰撞支持多手段输入模型组件，结合资产识别结果规避供应链风险。且依托 360 漏洞情报体系，内置 32w + 覆盖 CNVD、CVE 等主流库的漏洞情报。

（3）旁路检测，实现交互式大模型安全审计

基于OWASP LLM TOP10风险为核心，采用交互式方式进行旁路审计，避免对模型业务造成影响。全面检测模型从输入到输出的全链条安全问题，覆盖输入控制缺陷、供应链风险、知识库控制风险、输出控制缺陷等，保障模型自身安全。以模治模，基于攻击模型生成审计规则，基于裁判模型审计模型输出结果，显著提升审计效率和准确度。

4.1.3 360大模型卫士防护系统

【概述】

随着大模型在各行业广泛应用，输入敏感信息、输出违规内容、提示词注入攻击等内容安全风险日益凸显，加之国家监管政策对模型合规提出明确要求，企业正面临“上线即风险”“防护滞后”等挑战，亟需体系化、自动化、可闭环的内容安全解决方案。360大模型卫士防护系统致力于解决AI内容合规与安全问题，以“以测促防、以模治模”为设计理念，依托专项训练的风险检测、评测裁判与安全代答三大模型，结合国标合规基线与实战攻防经验，构建起覆盖“事前评测 - 事中拦截 - 事后优化”的全链路防护体系；该系统不仅具备输入输出内容风险检测、安全代答、安全改写等核心能力，还集成越权攻击对抗靶场与内容安全评测平台，通过评测与防护一体化、模型与策略联动、合规与实战结合的模式，为企业提供从建设到运营的全流程内容安全保障，助力企业在 AI 应用中实现内容安全的可知、可管、可控，达成“安全可控、合规发展”的目标。

内容评测·风险前移　　**多重检测·精准防护**　　**策略调优·能力迭代**

【产品能力】

（1）内容安全评测数据集

- 合规数据集：一方面覆盖国标《网络安全技术生成式人工智能服务安全基本要求》标准规定的5大类31小类风险评测，重点针对标准附录A5中所涉及的不可靠、不准确内容等幻觉风险进行专项评估，支撑合规备案；另一方面提供金融、医疗、政务等垂直行业模型合规评测数据集及专属合规建议。
- 对抗性攻击数据集：面向典型提示词攻击场景，提供高对抗性的良性与恶意提示词样本，全面覆盖目标劫持、提示泄露等 7 类典型攻击类型，且每个攻击类型下细分不同难度等级样本，辅助企业模拟真实攻击环境，识别并提升模型抗攻击能力。
- 自定义数据集：开放高灵活度的自定义功能，支持手动上传自有数据集，适配垂直行业合规验证、特定攻击场景模拟等个性化评测需求。

（2）内容安全评测

- 提供自动化、体系化的评测能力，覆盖模型上线前风险评估、备案自评与常态化自检。
- 评测裁判大模型实现多维度风险量化评分，人工一致率超过95%，大幅降低人工审查成本。

（3）内容安全护栏

- 实时检测用户输入与模型输出，对风险内容进行拦截或安全代答，保障回复内容安全可信。
- 支持干预回复库快速配置，实现风险策略动态更新，形成防护闭环。

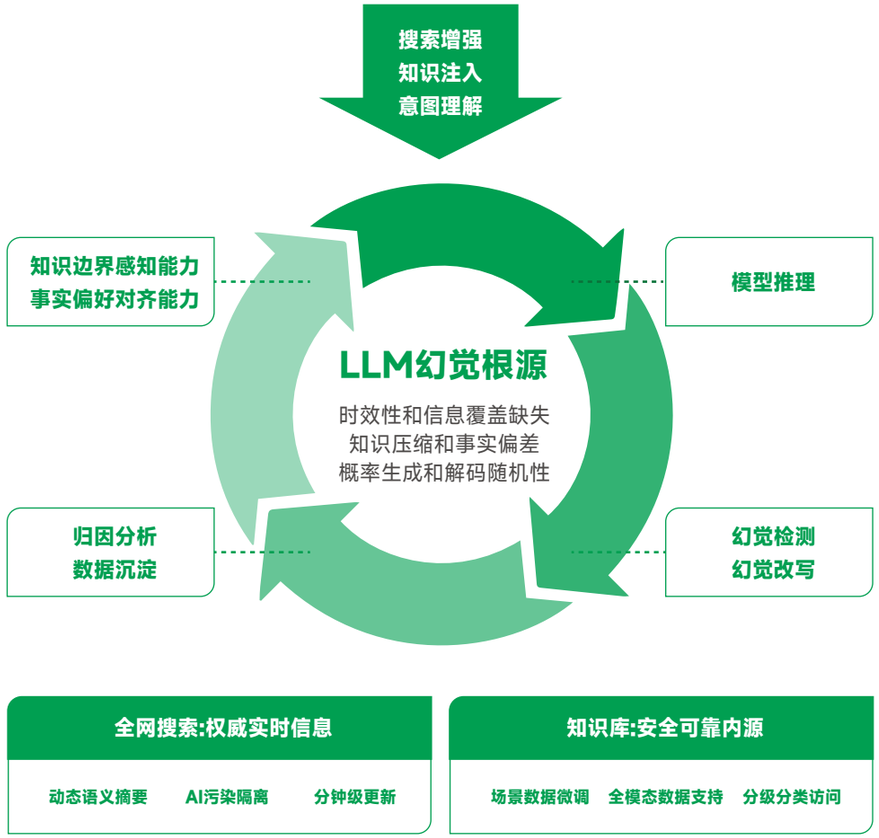
（4）模型全生命周期管理

- 支持多类型模型快速接入与资产统一管理，构建“接入-资产-漏洞”一体化视图。
- 结合漏洞扫描与态势大屏，实现模型风险可视、可溯、可治。

4.1.4 360大模型幻觉检测与缓解系统

【概述】

360大模型幻觉检测与缓解系统，是一套致力于提升大模型内容生态可靠性、准确性与安全性的综合解决方案。该系统以检索增强技术、幻觉检测与缓解智能体为核心，构建起人机协作时代下内容生态的可信防线。这套系统通过构建动态反馈闭环，能够持续学习优化，持续提升大模型的能力，这种自我完善的机制最终将推动大模型从“概率生成”向“可信、准确推理”演进，重塑人机协作的信任边界。



360大模型幻觉检测与缓解系统

【产品能力】

（1）权威信息实时获取

360大模型幻觉检测与缓解系统通过智能搜索接口提升大模型全网信息获取能力。智能搜索接口是在传统搜索接口基础上，专为大模型 RAG 应用场景优化打造的 AI 检索增强接口，侧重全文语义匹配与精品知识召回，能为大模型生成提供更优输入内容；为适配多场景应用，它还提供网页搜索、精品知识库搜索、图片搜索、新闻搜索等多种接口，以 SaaS 方式对外提供在线调用服务。搜索接口具备定向抓取和及时补录能力，时效性内容更新效率快至分钟级，支持段落级语义查询、旅游和健康医疗等领域查询，可指定可信内容源、站点及时间范围查询，能提供天气、汇率、股票、油价、期货、门票等实时信息。

（2）私域知识库访问

360 大模型幻觉检测与缓解系统通过接入质量可控、内容安全的企业级知识库和云端SaaS化知识库，强化对专业领域及私域知识的理解能力，有效识别并应对领域性幻觉。云端SaaS化知识库采用全场景、开放式、模块化的RAG综合架构，深度融合场景化的多路召回机制与DeepSearch深度搜索技术，构建覆盖全场景的综合召回解决方案。全面支持60余种数据格式，涵盖文本、表格、图像、音视频、网页链接及代码等多种模态，真正实现全模态数据处理与支持。私有化企业级知识库能力后续有专门章节介绍。

（3）信息校验与修正

360大模型幻觉检测与缓解系统的幻觉检测与缓解智能体融合搜索增强与知识库增强双路信息源，运用幻觉检测大模型构建多源协同的内容校验和修正机制。智能体不仅能够精准识别潜在幻觉，更具备内容改写与自动修复能力，并通过反馈与数据沉淀反哺基础模型训练，提升其知识边界感知与事实偏好对齐能力，实现对基础模型事实一致性与逻辑可靠性的持续提升。

4.2 平台原生安全：从底层架构夯实内部安全基础

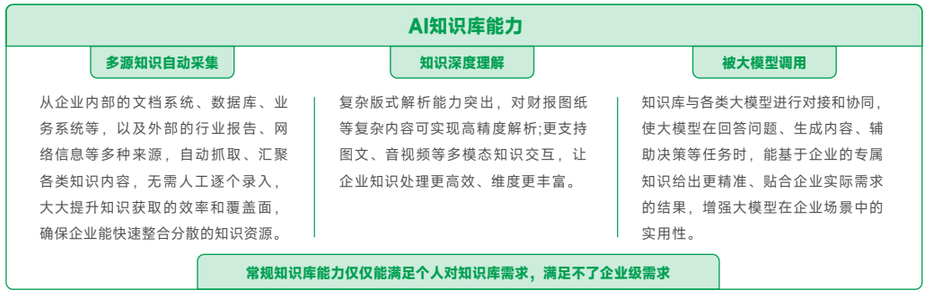
数据与知识应用、智能体、用户端作为智能体应用的核心组件，其安全无法依赖外部外挂式方案解决，需依托业内现有实践的知识库平台、智能体构建与运营平台及配套智能体用户端产品，从底层架构赋予原生安全能力，唯有让这些核心平台与产品在设计之初就嵌入安全属性，才能从根源保障智能体应用全链路安全，夯实内部安全基础。360 提供的 AI 企业级知识库、智能体构建运营平台及智能体客户端，能够全面满足国家与行业对大模型建设的安全合规要求。

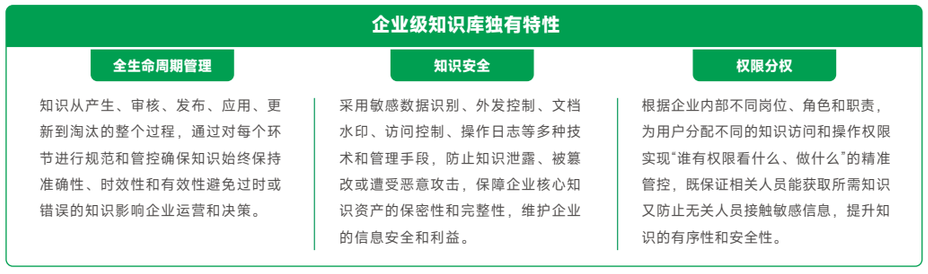
4.2.1 企业级知识库

【概述】

随着 AI 技术变革逐步迈入“深水区”，企业对数据和知识应用的需求已发生本质性升级 —— 过去仅用于文档检索的通用 RAG 工具，早已无法满足“知识驱动业务创新”的核心诉求。当下企业真正需要的，是能串联“文档、数据、业务、Agent 应用”的企业级知识库：它既是企业数据持续、高效转化为 AI 新质生产力的中间枢纽，也是智能体落地的核心 AI 基础设施，更是打通企业 AI 转型“最后一公里”的关键载体。

而要实现这一价值，“安全”是企业级知识库不可逾越的核心壁垒 —— 与通用 RAG 工具仅需基础数据加密不同，企业级知识库承载的是企业核心知识资产（如商业机密、技术专利、客户隐私数据），其安全能力需覆盖知识全生命周期，构建多维度防护体系。具体而言，首先需具备精细化权限管控能力，通过数据分级分类（按敏感程度划分为公开、内部、机密、绝密等级）与角色分权管理，确保不同岗位员工仅能访问权限范围内的知识，避免越权查看导致的机密泄露；其次需强化全链路风险监控，从数据采集阶段的敏感信息识别，到存储阶段的文件存储安全（如加密存储、容灾备份），再到调用阶段的数据泄露监控（如异常访问预警、调用日志追溯），实现“事前预防、事中干预、事后追溯”的全流程管控；此外，还需具备基础安全防护能力，通过文件数据杀毒、恶意访问拦截等功能，防范外部攻击或恶意软件对知识中枢的破坏，保障核心知识资产的完整性与可用性。





【原生安全能力】

360 AI 企业知识库一款专为企业AI转型和高效落地智能体而打造的“企业级智能体应用知识中枢”产品，是企业AI核心基础设施。具备“AI知识库+企业级知识管理”双重特性，既能有效支撑智能体落地，又让企业知识管理“不混乱、更高效、更安全”。360AI 企业知识库以“事前定策略、事中保安全、事后可溯源”为核心，构建覆盖知识全生命周期的 360° 安全管控体系，从权限、分类、存储、审计到管理多维度保障企业知识安全。

具体安全功能可归纳为五大核心模块：

（1）精细化权限与分类管控

通过“基于角色的 AI 问答权限控制”实现“千人千面”问答，结合用户身份与角色匹配知识片段，避免涉密信息泄露；“分类分级 Agent”按企业规则自动完成知识分类与密级判定，联动权限系统精准匹配访问权限；同时支持用户与文件密级管理，密级随人员、文件流转，确保全周期安全，还可批量设置密级属性，提升管理效率。

（2）多场景安全防护

数据和文件层面，按密级与用户身份生成含专属信息的水印，防止截图外传；上传环节嵌入云查毒引擎，扫描通过方可存储，检测到病毒自动隔离，杜绝云端病毒扩散；敏感词检测依托“内置 + 自定义”双库，实现敏感信息实时识别、文件拦截与全盘扫描，降低合规风险。

（3）全链路数据安全保护

存储端以高安全架构为基础，通过多重备份、分散存储与数据加密防单点故障，搭配容灾机制应对极端场景；传输端采用 HTTPS 协议封装、AES CTR 256 算法加密与无落地存储，全程杜绝数据窃取；登录端从设备、IP、时效三维防护，支持设备限制、IP 白名单、自动登出与登录提醒，锁定账号风险。

（4）知识全生命周期管理

体系化的知识管理能力，可以涵盖“知识生产、知识存储、知识处理、知识发布、知识审核、知识应用、知识运营、知识销毁”知识全生命周期各个流程环节，可以全方位支撑企业知识从产生到价值转化的全周期需求。

（5）全行为日志审计

详细记录文件使用与用户操作的全量日志，包含操作人、时间、内容等关键信息，日志可实时调取，为审计合规与风险追溯提供完整依据，实现 “每步操作有记录，每次风险可溯源”。

4.2.2 智能体构建和运营平台

【概述】

在企业智能体应用落地进程中，智能体构建与运营平台作为智能体应用生产和运营中心，其原生安全能力直接决定智能体业务的合规性与稳定性。智能体构建与运营平台需从功能、认知、权限、协议四维度强化安全：

（1）功能调用管控：

建立第三方插件、MCP、API 等工具安全准入审核，经漏洞扫描、权限审计后才可接入；对代码执行功能做权限隔离。同时对工具调用全流程日志审计，异常调用立即告警中断。

（2）认知执行保障：

在任务规划环节引入多轮校验，构建幻觉样本库优化大模型理解能力；对关键操作设置人工或自动化校验关卡。

（3）权限分级隔离：

搭建“开发者 - 审核者 - 使用者”多角色权限体系，细化数据库、知识库访问权限；为不同智能体提供独立运行环境，隔离资源调用。

（4）MCP 协议强化：

审计加固 MCP 协议，通信引入 TLS 加密、身份认证；管控智能体 MCP 调用权限，检测传输数据防止恶意指令与数据窃取。

【原生安全能力】

360 基于对智能体原生风险（如工具滥用、行为失控）与 MCP 协议安全隐患（如投毒攻击、未授权访问）的深度洞察，打造了“Agent 安全防护 + MCP 安全管控”双核心安全体系，既覆盖智能体从设计到执行的全生命周期风险，又破解 MCP 协议交互中的安全痛点，为企业智能体规模化落地提供全方位安全保障。

（1）Agent 安全防护：原生可控 + 动态沙箱

针对智能体在设计阶段的工具固有威胁与执行阶段的行为失控风险，平台通过“原生框架嵌入 + 动态沙箱拦截”的组合策略，实现从根源到过程的全维度防护。

- 原生安全可控框架

在智能体设计阶段即原生嵌入安全机制，如同为大模型的决策与执行能力装上“安全枷锁”，通过四重管控实现大模型能力的可管可控：

任务规划监督：

拆解任务失控概率与危害程度，关键任务强制人工审批。

工具调用监控：

实时监控 API、数据库等工具调用，触发安全规则立即拦截告警。

调用策略分级：

对工具服务调用分类分级，采取禁止启动、人工审批等差异化措施。

关键动作审批：

高风险关键动作（如敏感数据导出）严格人工审批并持续验证合规性。

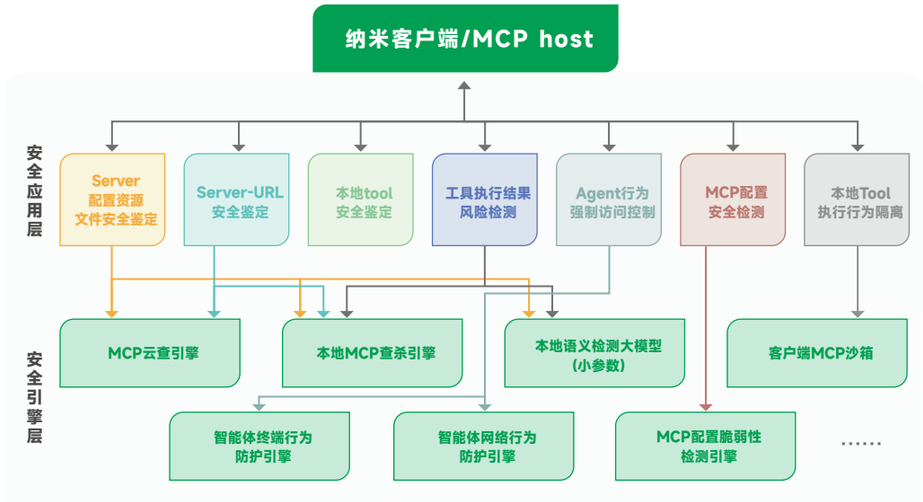
- Agent 行为沙箱

聚焦智能体执行阶段的动态风险，通过模拟真实业务场景构建“安全试验场”，实现风险的提前识别与拦截。

- 精准识别并拦截恶意行为，如检测到“rm -rf /”等恶意删除系统文件指令、发送含客户隐私的错误邮件、擅自关闭安全防护组件等操作时，立即中断执行并触发告警。
- 内置大模型行为评价模块，区分正常与恶意操作，实现主动风险免疫。

(2) MCP 安全管控：运行全过程的安全防护

针对 MCP 协议在连接大模型与外部工具时存在的投毒攻击、地毯式骗局、影子攻击等风险，360从MCP Client 的视角做全过程的安全防护，可划分为运行前风险评估 、 运行时安全防护。



360MCP防护原理图

- MCP运行前风险评估：前置筛查，多维度校验鉴定

以“准入校验”为核心，通过安全应用层和安全引擎层进行多维度安全鉴定，提前识别仿冒、篡改、脆弱性等潜在威胁，从接入环节阻断风险入口，确保后续交互的基础安全性。

安全应用层：

聚焦 Client 交互的核心风险点，安全应用成具备七大风险检测能力 ——Server 配置资源文件安全鉴定（防范篡改配置引发的风险）、Server-URL 安全鉴定（拦截恶意 URL 连接）、本地 Tool 安全鉴定（校验工具合法性）、工具执行结果风险检测（识别执行后输出的违规内容）、Agent 行为强制访问控制（限定 Agent 操作范围）、MCP 配置安全检测（排查配置漏洞）、本地 Tool 执行行为隔离（避免工具异常影响 Client 整体安全），MCP配置脆弱检测（对 MCP Client 与 Server 之间的配置（如超时、重试策略、凭证配置）进行脆弱性扫描与评估），从应用交互层面进行风险评估鉴定。

安全引擎层：

依托多类专业引擎提供底层安全支撑，包括 MCP 云查引擎（同步云端威胁情报）、本地 MCP 查杀引擎（实时检测本地恶意文件）、小参数语义检测大模型（快速识别恶意指令）、客户端 MCP 沙箱（模拟执行环境提前发现风险）、MCP 配置脆弱性检测引擎（扫描配置缺陷），以及智能体终端 / 网络行为防护引擎（监控终端操作与网络传输），全面覆盖配置、执行、本地存储、网络通信等安全场景。

- MCP运行时安全防护：动态管控，守住安全底线

聚焦“执行环节”的实时防御，通过强制访问控制、执行结果检测、沙箱隔离等机制，动态监控工具调用行为、数据交互过程，及时拦截越权操作、恶意返回值等风险，实现交互全流程的安全闭环管控

Agent 行为强制访问控制：

基于强制访问控制策略，工具调用前结合当前任务 MCP Client 风险级别，检测并拦截其调用风险级别高于自身的工具。

工具执行结果风险检测：

对 MCP 工具执行返回的结果开展内容安全检测，识别并拦截包含恶意脚本、敏感信息泄露或异常输出的返回值。

本地 tool 执行行为隔离：

将本地 MCP 工具进程运行在沙箱中，限制其文件系统、网络和系统调用范围，避免对本地环境造成恶意破坏。

4.2.3 智能体客户端

【概述】

智能体客户端是一款整合 AI 相关核心功能与安全防护能力的一体化软件工具，旨在为用户提供便捷的 AI 服务访问入口，同时聚焦 AI 客户端在访问大模型、执行智能体过程中面临的各类安全问题，通过集成多种安全防护功能，解决 AI 应用过程中的安全与合规风险，为 AI 应用提供全方位、全流程的原生安全保障。

（1）企业AI的统一入口，满足用户AI使用的需要

- 能够访问、切换企业的多个大模型，简化模型评估与选择，能访问企业知识库；
- 能运行企业搭建的智能体，能对MCP工具等进行管理，降低工具集成成本。
- 能通过沙盒执行代码、视频生成、报告生成等运行过程，可实现快速启动、弹性伸缩、故障自愈，保障复杂任务的稳定执行
- 满足智能体高交互的需求： 实现“人在回路”闭环，（即人类参与关键决策）可通过标准化工具调用事件发起请求，用户只需在客户端完成确认、修改、选择等操作，即可无缝衔接后续流程，让智能体的结果更可靠。
- 使智能体具备更强的执行力:客户端可承载Computer Use、Browser Use 和 Soft IM Use 技术的运用，凭借自动化处理、跨场景适配、高效协作与低门槛交互的综合优势，简化了用户操作，大幅强化了客户端的执行能力，也提升了执行效率。

（2）满足企业AI应用的安全合规需求，对智能体客户端侧的风险进行安全保障

智能体客户端是企业AI应用落地的“最后一公里”，也是攻击的“最前线”，长期暴露在复杂多变的环境中，成为黑客突破防御的首要目标。因此，客户端的原生安全能力至关重要，需从底层筑牢安全防线，为 AI 应用的安全、稳定运行筑牢根基。智能体客户端安全主要考虑两个方面问题，一是在访问大模型时的行为是否安全；二是客户端在执行大模型或智能体时，是否会对本地造成威胁或导致隐私数据及文件泄露。具体涉及到以下五个要点：

大模型及智能体访问控制：

具备精细化的组织、身份、应用的权限管理能力，对智能体应用进行用户授权、访问控制和审计；

API 访问监控：

客户端需具备API访问控制与监测能力，保障智能体数据与应用调用的安全；

恶意智能体本地执行：

攻击者获取Agent 平台权限后，恶意构建智能体并诱导客户端执行，最终实现获取用户权限、获取用户文件、加密文件、删除文件、下载恶意文件，并对计算机进行恶意破坏等操作。

个人隐私数据防护：

构建智能体的开发者，误用了其他恶意插件，或者本身的目的就是采集客户端用户数据，采集文件，个人隐私等问题，需要采取防护措施。

独立的沙盒执行空间：

通过独立环境执行智能体，对有风险的智能体进行安全隔离，避免扩散。

【原生安全能力】

【360智能体客户端：双重原生防护，保障业务与AI数据安全】

360智能体客户端以兼容为特色、安全为基石、AI为引擎，构建连接“人-业务-AI”的智能办公生态，成为企业AI客户端侧的核心基础设施。以企业数据安全为核心前提，实现AI能力的高效分发与场景化落地，同时筑牢业务+AI数据安全双重屏障，一是进行全生命周期的业务安全保障，包含自身安全、接入安全、行为安全和数据安全，二是平台具备原生的访问和权限管控能力，全面保障客户端的原生安全。

（1）全生命周期的业务安全保障

文件和隐私数据安全：

对客户端进行文件的上传、下载场景时针对目标文件进行扫描查杀；在访问管理员定义的高敏感度业务应用时自动调用环境安全检测。

客户端环境安全：

客户端启动环境检测、进程注入防护、反编译防护、客户端程序完整性校验、文本控件溢出防护、以及第三方服务和插件加载等判断等，为客户端安全运行提供准备。

异常行为管控：

页面水印覆盖、落盘数据加密储存、截图保护、禁止下载等对用户行为、访问控制、访问监管和审计等安全管控功能，严防办公数据泄露。

智能体沙盒隔离执行：

360智能体客户端沙盒生态，打造“音视频生成、代码运行、报告生成、可视化图表渲染、浏览器运行、网页运行”等多类专业沙盒，既满足不同场景下的智能体运行需求，通过安全隔离技术，将各类操作风险牢牢锁定在沙盒内。

动态身份验证：

客户端融入零信任SDP能力对业务系统进行零信任接入控制、环境感知、动态和最小授权访问，从接入防护到业务上层操作风险拦截，为业务运行加注多重保护。

（2）安全访问、精准调度、权限管控三位一体的智能体客户端治理体系

安全可控的AI访问：

通过在客户端中限制外部生成式AI的使用、加强数据传输审计，并提供合规可控的AI替代方案，最大限度降低数据泄露风险；

私有化AI精准调度：

支持企业私有化部署的大模型及智能体，按业务需求智能匹配调用，如销售场景自动分配CRM分析智能体，eHR使用场景调用人力智能体；

权限精细化管控：

通过身份认证与模型权限策略结合，限制AI返回内容的可见范围，避免信息越权访问。

第五章

05

生态共治：构建可信AI生态

5.1 生态力量：
以标准共建与产业联盟夯实安全基础

360深度参与了多项生成式人工智能安全国家标准的制定工作，包括《GB/T 45654- 2025 网络安全技术 生成式人工智能服务安全基本要求》、《GB/T 45674-2025 网络安全技术 生成式人工智能数据标注安全规范》和《GB/T 45652-2025 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》等，为行业安全合规和技术标准化提供了系统支撑。

在人才生态建设方面，360携手国内顶尖高校与科研机构，聚焦大模型推理能力及大模型安全领域，深度开展产学研合作，并建立了覆盖本科至博士后的全方位人才联合培养机制。合作成果斐然：形成多篇学术论文，发表在AAAI、ACL、EMNLP等国际顶尖AI会议上，同时还开源了tiny-r1等性能位居第一梯队的前沿大模型。基于此，360系统构建了从理论研究、技术实践到产业应用的AI安全人才培养闭环，正全力打造可持续的安全人才供给链。

同时，360携手多家大模型厂商、科研院所、算力与安全伙伴，共同发起成立大模型安全联盟。该联盟汇聚产、学、研多方力量，致力于打造资源共享、共创共赢的大模型安全生态集群，推动安全标准共建、安全技术创新与安全能力提升，共同探索大模型安全治理的新范式。

5.2 联合实践：
通过产学研协同将安全融入技术生命周期

在大模型安全领域，360始终坚持产学研协同创新，联合产业与学术力量，共同探索安全、可控、可信赖的人工智能发展路径。

360与北京大学联合研发的 TinyR1-32B 模型，聚焦开源大模型“安全性不足”的痛点，在安全对齐与推理性能上实现“双突破”。该模型采用创新的 Control Token 技术，可根据内容安全检测信号动态切换工作模式，在安全与有用性之间实现自适应平衡。TinyR1-32B 仅以 DeepSeek-R1-0528 约5%的参数量，在安全能力上超越 Qwen3-32B 25分、DeepSeek-R1-0528 17分，并在数学、科学、代码等任务中达到后者 93% 的推理性能。目前模型已全面开源，可快速应用于安全审核、科学问答、代码生成等场景，为安全可信的开源大模型提供了可复制的技术范式。

360联合中国信息通信研究院等单位共同建设的工信部人工智能大模型公共服务平台在2024年10月正式上线试运行，目前已有注册用户5000余人，累计访问量20万余次。依托平台安全检测能力，累计为100余家企业的大模型提供了模型安全等方面测试工作，支撑7个批次的AI Safety Benchmark大模型安全基准测试工作，挖掘出国内外主流开源大模型，以及部分商用模型的安全风险，有效支撑生成式人工智能安全治理工作。

在 2025 年 9 月举办的第 22 届中国———东盟博览会上，360作为安全技术与服务合作方，为展会期间的大模型应用与安全运营提供了全方位支撑。依托模型安全、内容安全与平台管控等成熟能力，360为博览会AI应用场景提供了实时风险监测、智能内容过滤、身份与权限控制等关键安全保障。在为期五天的展会期间，系统累计防护89097次，拦截恶意请求 117 次、识别攻击行为 571 次，安全检测准确率超过 95%，确保了大型公共活动中AI系统的稳定、安全与合规运行。

5.3 未来倡议：携手产业与用户共建可信大模型生态

大模型安全的复杂性与全局性，决定了这绝非任何单一主体能够独立承担的任务。构建安全、向善、可信、可控的大模型生态，是一项需要企业、行业组织、与用户等多方合力共筑的系统工程。

- 致AI企业：坚守安全底线，践行原生安全

我们呼吁所有大模型技术与服务提供商，将“安全、向善、可信、可控”奉为发展的生命线。唯有坚守这一底线，AI创新才能行稳致远。核心在于践行“以模治模、原生安全”理念。需将安全能力深度融入大模型全生命周期，让安全成为每个环节的固有属性，而非事后补充。这不仅是风险防控的必然要求，更应成为企业的核心产品竞争力。

- 致行业用户：建立评估机制，优选可信服务

倡议行业用户建立常态化的AI安全能力评估与审计机制。在采购或自建大模型应用时，应将服务商的安全资质、模型的抗攻击能力、数据的隐私保护方案作为核心考核指标，优先选择并激励那些公开承诺且能验证其安全实践的可信大模型服务。

最终，一个安全、向善、可信、可控的大模型生态，将成为“数字中国”宏伟愿景中最稳固的基石。