数世咨询
Digital World Consulting

AI 安全系列研究报告

安全优先的大模型

(Security-Prioritized Foundation Models)

数世咨询
Digital World Consulting

AI 安全系列研究报告

安全优先的大模型

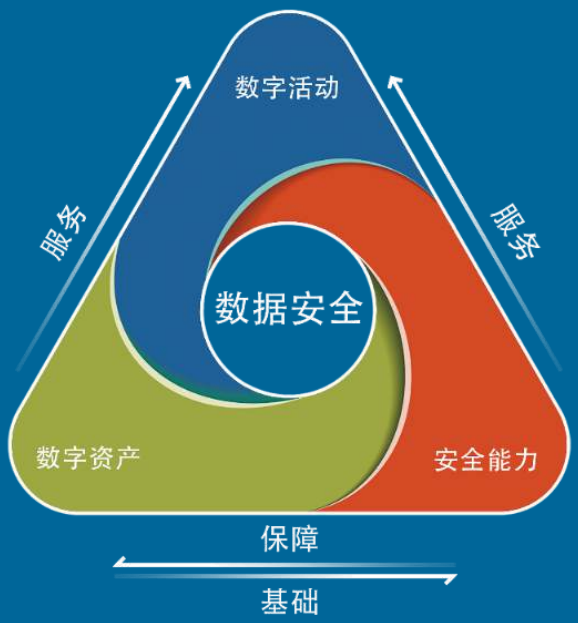
（Security-Prioritized Foundation Models）

数字安全是指，在全球数字化背景下，合理控制个人、组织、国家在各种活动中面临的数字风险，保障数字社会可持续发展*的政策法规、管理措施、技术方法等安全手段的总和。

这里的风险，不再局限于围绕数字化资产的攻防对抗，还包括了数字资产所承载业务的稳定性、连续性和健康性。这里的安全不再特指有意还是无意，天灾还是人祸，保安还是保险，而是更为广义的安全状态（SecSafe）。

*“世界环境与发展委员会出版的《我们共同的未来》报告中，将可持续发展定义为：“既能满足当代人的需要，又不对后代人满足其需要的能力构成危害的发展。”

——数世咨询，2023 年 11 月



以安全能力、数字资产和数字活动为三元素，以数据安全为核心目标，即三元一核的“数字安全三元论”。

“数字安全三元论”由“网络安全三元论”（数世咨询于 2020 年提出）更新迭代而来，旨在匹配数字中国建设的进程，保障数字基础设施稳定、可持续运行，保障数据有效流动、激发数据要素价值。

数世咨询作为国内独立的第三方调研咨询机构，为监管机构、地方政府、投资机构 ．网安企业等合伙伙伴提供网络安全产业现状调研，细分技术领域调研、投融资对接、技术尽职调查、市场品牌活动等调研咨询服务。

报告编委

报告类别：研究报告

报告名称：安全优先的大模型

主笔分析师：靳慧超 数世咨询·战略分析师

分析团队：数世咨询·数字安全研究院

报告审核：李少鹏 数世咨询·首席分析师

目 录

本报告调研的入选标准 1
本报告调研的安全能力供应商 1
报告背景 2
关键发现 3
一、概念定义 4
1.1 定义 4
1.2 定义解读 4
二、发展潜力 6
2.1 市场发展驱动力 6
2.1.1 大模型内生“缺陷”转化为安全原生需求 6
2.1.2 大模型强监管态势夯实安全合规基础 7
2.2 安全能力核心逻辑 8
2.2.1 以“模”制“模” 8
2.2.2 以权限和身份管理重塑业务 9
2.2.3 以安全工程保障系统工程 9
2.3 未来趋势 11
2.3.1 大模型重塑数字生活，人工智能治理道阻且长 11
2.3.2 端侧大模型需求激增，安全能力需要新突破 11
2.3.3 后训练是应用关键，数据价值再次攀升 11
2.3.4 公众模型成为基础设施，智能体百家争鸣 12
三、推荐供应商 13
四、安全能力 17
4.1 大模型业务系统安全风险 18
4.2 安全优先的大模型能力图谱 20
五、解决方案/典型案例推荐 22
5.1 大模型安全解决方案 22
方案背景 22
方案概述 22
解决方案核心价值 23
AI 安全产品详解 24
5.2 数字政府智算服务一体化安全保护案例 28

项目背景简介 28
大模型业务系统安全保护需求 28
整体解决方案 28
5.3 科技制造业大模型安全防护案例 33
项目背景简介 33
大模型业务系统安全保护需求 33
整体解决方案 34
核心安全能力 35
5.4 联想携手火山引擎推出可信个人云案例 37
项目背景简介 37
大模型业务系统安全保护需求 38
整体解决方案 38
核心安全能力 39
5.5 央企大模型综合治理案例 41
项目背景简介 41
整体解决方案 41
核心安全能力 42
5.6 杭州市数据资源局大模型安全防护案例 45
项目背景简介 45
大模型业务系统安全保护需求 45
整体解决方案 46
核心安全能力 46

本报告调研的入选标准

- 具有 AI 研究能力、大模型安全保护产品具备自主知识产权，通过 SaaS 或私有化部署方式，为企业用户提供保护大模型业务应用的产品、服务、解决方案的安全厂商。
- 本报告调研的安全能力，不包括通用安全，只针对明确的大模型合规要求或大模型业务系统特有安全需求。
- 产品或服务可被完整交付，解决方案有实际场景或落地案例为支撑。
- 接受数世咨询的调研与访谈，并承诺提供数据的真实性。

本报告调研的安全能力供应商

（按调研顺序排序）

报告背景

人工智能技术已经在全球范围内得到了普及，而大模型作为人工智能的复杂应用，在以中美为核心引领的环境下，已经为人类带来了极大的震撼。随着人工智能技术的不断发展，在可预见的未来，通用大模型将成为数字智能的基础设施。

当前，通用大模型竞争格局逐渐明朗，人们也越来越清晰的认识到，人工智能的价值并不在于模型本身，而在于其深度融合并改造业务场景的过程。以专业领域大模型为核心的大模型业务系统（包括智能工作流和智能体）和具身智能接棒开启了新一轮的白热化竞争。

为了推动大模型业务系统和具身智能等人工智能应用的持续发展，更为了用户可以合规、安全的使用大模型赋能业务，数世咨询特开展了本次调研工作。

希望本报告可以帮助使用大模型的用户，了解大模型业务系统中的安全风险和相应的安全能力供应商，在大模型业务系统建设和运营时提供有益的参考。

关键发现

- ✓ 大模型安全保护市场，自 2025 年开始加速进入需求爆发期，现阶段以合规为核心驱动。随着大模型稳定性以及数据要素价值的升高，未来以“合规+业务”为双轮驱动。

- ✓ 大模型本身不等于大模型业务系统（包括使用大模型的工作流和基于大模型的智能体），前者是后者的子集，后者需要通过系统化的安全保障能力满足业务系统安全需求。

- ✓ 大模型业务系统安全与数据安全的保护理念是一致的，都需要深度融入业务流程和数据流向之中，对数字安全产业来说既是挑战也是机会。

- ✓ 现阶段面向公众提供服务的大模型，其核心需求是备案全流程服务，上线后则侧重内容风控。为企业经营赋能的大模型，其核心挑战是引入大模型后对原有业务流程和访问控制的重塑，关键点是数据泄露防护。

- ✓ 现阶段，用户需求较为集中且安全供应商可完整交付的、较为成熟的安全产品和服务主要有大模型安全围栏、内容风控、风险评测与备案服务。

一、概念定义

1.1 定义

数世咨询将安全优先的大模型定义为：

由于大模型原生安全缺陷和业务系统内生安全风险无法避免，为了有效控制安全风险为企业带来的经营风险、更为了实现较高水平的社会治理，在建设、运营、监管大模型业务系统过程中的一种思想，即安全优先。

1.2 定义解读

这里的安全并不特指网络安全，而是由于科学技术的应用可能给社会带来的潜在影响，由于必须保障这种影响是积极、可控的，所以安全性是必须优先考虑的。

国家层面，大模型的应用在军事（如认知域作战）、生物（如蛋白质结构）、医疗（如影像诊断学）等方面已经展现出强大推动力，但生成内容的准确性、系统的鲁棒性都是必须优先解决的关键问题，如处理不当将会造成不可预估的颠覆性灾难。

社会层面，大模型正在对人们的生活产生潜移默化的影响，如搜索方式的转变（搜索引擎到大模型应用）、内容创作的转变（人的独创到人与大模型的交互）等，但大模型应用对个人信息的滥用、对流程化工作岗位的取代等社会现实问题已经成为热点讨论话题，如处理不当将引发生群体性事件导致社会动荡。

企业层面，大模型可以赋能数字化应用从而促进核心业务发展，如商业数据分析（突发性、创意性数据分析需求）、产品智能化升级（自动驾驶路径规划）、自动化安全运营（7*24 小时告警降噪）等，但敏感信息和业务数据泄露以及知识产权保护等问题都与企业经营息息相关，如处理不当将使企业遭受巨大经济损失。

综合来看，大模型应用安全风险所造成的负面影响通常是无法被接受的，这也就直接导致了大模型应用畏首畏尾的局面，究其根本原因是对大模型的不信任。

而安全能力通过针对性和体系化的保障手段可以间接提高信任度和满足合规要求，所以安全优先的大模型可以有效推动大模型应用发展。

二、发展潜力

大模型的特异性来源于人工智能算法、模型权重和训练数据，大模型的业务系统依托于基础设施和供应链，大模型的应用价值靠高质量数据集和业务的互动来实现。

所以实现安全优先的大模型是一项系统性工程，它包含了国家安全、社会治理以及企业的网络与数据安全。

2.1 市场发展驱动力

“安全优先的大模型”真正实现了业务驱动的逻辑闭环，数字安全产业自此正式开启“以合规为基、以业务为柱”的新价值时代。

网络安全领域的发展主要以监管合规的要求为核心（产值贡献 80%以上）驱动，虽然安全保障也涉及业务连续性方面，但更多的原因是关键信息基础设施发生安全风险会对国家安全、社会治理带来重大威胁。

数据安全领域的发展虽然本质上是合规和业务双轮驱动的，但在现阶段我国数据流通基础设施尚未完善、数据交易体系尚未健全的情况下，数据要素价值还没有找到充分释放的场景，数据安全仍然以合规监管为核心驱动。

然而人工智能安全却在诞生之初就具备业务驱动的逻辑闭环，真正实现了“以合规为基、以业务为柱”的驱动形态。

2.1.1 大模型内生“缺陷”转化为安全原生需求

大模型的突破性进展催动人类加速步入 AGI 时代，在人工智能逐渐成为数字化基础设施的这一背景下，对于国家、社会、企业来说已经无需再探讨是否使用人工智能的话题，关键是解决如何利用好人工智能的问题。对于企业来说，人工智能所带来的高效性是数字时代商业竞争的核心支撑，不使用人工智能的企业终将丧失竞争力，彻底出局。

但大模型自身安全问题无法彻底解决，如幻觉、数据漂移、非预期行为等，更为关键的是大模型应用安全风险全部来源于具体业务系统的风控需求，这不仅仅是基础设施层面的安全可靠性保障，而是业务应用层面的价值保障。

由于大模型的性能和创造力与安全对齐的强度是成反比的，不能本末倒置的为了追求安全性而降低大模型的应用价值，只能通过后期工程化的方式用系统性的安全能力满足各类应用场景的安全需求。

所以这种大模型的内生“缺陷”就决定了大模型业务系统与安全能力的孪生属性，安全能力就成为了大模型应用的原生需求。

大模型业务系统支撑企业的数字化业务，数字化业务的发展决定了安全优先的大模型市场规模的高度，业务驱动则成为了大模型安全的支柱。

2.1.2 大模型强监管态势夯实安全合规基础

自人工智能技术诞生之初，人工智能治理的概念在全球范围内就得到了共识。我国作为人工智能强国，在 2023 年“一带一路”峰会上，由习近平主席发布了全球人工智能治理倡议，倡议人工智能的发展要以人为本，建立健全法律和规章制度。在 2025 年世界人工智能大会上，由李强总理发布了人工智能全球治理行动计划，强调把握机遇共同发展，并开展人工智能安全治理。

全球范围内对人工智能，尤其是生成式大模型应用都处于高位监管态势，我国陆续发布了《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》、《人工智能生成合成内容标识办法》以及国家标准《生成式人工智能服务安全基本要求》，从算法安全、语料安全、模型安全、应用安全以及模型上线等过程均有高强度监管要求。

除此之外，中央网信办还开展了“清朗·整治 AI 技术滥用”专项行动，统筹协调全国各地对 AI 技术滥用、AI 管理缺失等现象进行整治，成果颇丰。

在人工智能强监管态势的确定性环境中，深入实施“人工智能+”行动将继续促

动人工智能的发展，从而进一步夯实人工智能安全合规基础。

2.2 安全能力核心逻辑

由于实现安全优先的大模型需要依靠技术和管理手段，所以相应的大模型安全保护产品、解决方案和服务也就应运而生。

实现这些安全能力的核心逻辑有三点，分别为以“模”制“模”、以权限和身份重塑业务、以安全工程保障系统工程。

2.2.1 以“模”制“模”

以“模”制“模”的本质是基于大模型性能与安全性无法平衡的根本属性（安全对齐强度与创造力成反比），用安全专业“小”模型消减大模型输入风险、审核大模型输出内容，实现最高的投入产出比。

但用户在选择产品时需要注意分辨，其中最重要的语义分析不是关键字匹配，而是意图推测和多轮对话的上下文关联分析，有些厂商会混淆概念以夸大自身能力。如果内容安全控制方面存在多模态需求，更需要进一步甄别，多模态识别能力与人工智能研究能力强相关，不同供应商之间差别较大。

✓ **对抗性攻击防护**

提示词注入、模型规避（Model Evasion Attacks）等对抗性攻击，有效的解决方法是对模型进行代码调整，但其花费的时间和金钱成本较高，而这些攻击又相当于软件的零日漏洞，是不可计量、无法预测的。

前置语义检测安全大模型是简单、有效、低成本的最佳方法，对于已经发现的对抗性攻击类型可以直接进行防护。对于未发现的对抗性攻击类型，只需要对安全大模型进行少量调整即可，而安全大模型的代码调整、更新部署是极其快速和简便的。因为其本质是由大模型蒸馏而来的“小”模型，并且其更新与业务系统和流程不产生直接影响。

✓ **输出审核**

大模型幻觉问题至今无法有效解决，价值观偏见、不安全的输出也会随着数据漂移、数据投毒等问题逐渐失效。还有，不同用户对敏感数据的定义也不尽相同，在输出内容的控制上无法通过模型自身满足不同需求。

解决这些问题同样需要从模型训练和数据入手，其花费的时间和金钱成本较高。而通过安全大模型对输出的内容进行审核或代理回答，既灵活又简单，通过自定义的输出内容审核规则，可以满足不同用户的个性化需求。

2.2.2 以权限和身份管理重塑业务

权限和身份重塑的本质是为了解决业务系统引入大模型后，由于业务系统流程逻辑变更从而导致的原有身份和权限控制失效，致使发生商业数据泄露、信息泄密、敏感信息泄露、知识产权受损等事件。

如引入大模型的人力资源系统、文档管理系统，原有控制是通过身份来设置的数据访问权限，用户通过应用系统访问数据库。由于大模型的引入改变了用户与系统的交互方式，用户通过大模型访问数据库，任意员工均有可能通过与大模型的交互绕过原有身份权限获取更大范围的数据、信息。

有效应对的方法是通过业务具体控制需求和流程，以模型权限、用户身份管理为核心，辅助 UEBA、API、数据分类分级以及模型交互审查等技术，重塑业务系统安全控制体系。

2.2.3 以安全工程保障系统工程

系统保障的本质是为了解决大模型应用风险的传递性，通过安全系统对信息系统的全生命周期、数据处理的全流程进行整体性、体系化的安全保障，利用技术和管理手段构建全方位、多层次的安全能力。

由于大模型的本质是软件，在应用过程中扩展为信息系统，涉及网络与数据基础

设施、软件与模型供应链、数据管理、应用管理等，此时的大模型安全风险已经从大模型自身扩展到了大模型应用系统，每一个环节的安全风险都可能会引起大模型应用的安全事件，如非授权访问、敏感数据泄露，而这些安全风险也会通过系统的传递性间接影响大模型自身，最终产生各类安全问题。

✓ **数据安全**

大模型应用涉及预训练数据、后训练数据、RAG 数据、用户输入数据、模型输出数据等，在数据处理的各环节都有相应的安全风险，如数据投毒、数据窃取、个人信息保护、敏感信息泄露等。

对于这些数据处理的全流程都需要进行安全保障，在通用数据安全和个人信息保护能力之上，还需要数据标注、数据清洗、数据聚合泄密、信息推断泄露等大模型专有数据安全保障需求。

✓ **供应链安全**

大模型应用的供应链涉及算法、模型、框架、部署与推理工具、集成组件，在供应链上的每一个安全风险都可能影响整个大模型应用系统的安全。如利用部署与推理工具 Ollma 的安全漏洞进行模型窃取，在 GitHub 上传恶意组件包预留后门。

对于供应链安全风险，与通用软件供应链安全保障思路一致，目前并未发现大模型专有安全保障需求。

✓ **基础设施安全**

大模型系统的基础设施安全需求总体上与通用信息系统一致，安全保障技术和思路延用通用网络安全来构建体系化的保障能力。

唯一需要注意的是，在可预见的未来，智能手机、智能设备以及具身智能会成为新的大模型主要运行环境，而且都有其各自的特性，比如操作系统、存储类型、交互方式等，需要考虑安全防护能力在算力、存储等方面的限制，进行轻量化、针对性设计。

2.3 未来趋势

2.3.1 大模型重塑数字生活，人工智能治理道阻且长

大模型的应用已经悄然改变了互联网搜索的交互方式，随着应用的深入，会有更多的传统数字化应用交互方式被改变，未来还会出现全新的数字化交互方式重新塑造数字生活。

在面对一个全新的数字化社会形态时，尤其是在大国竞争转为贸易战和科技战的背景下，人工智能作为可以赋能第一二三产业的全面型应用，势必会受到额外的重视。

价值越高，风险越大，未来的人工智能治理道阻且长，需要国家、社会、企业共同参与，贡献自己的力量。

2.3.2 端侧大模型需求激增，安全能力需要新突破

模型蒸馏使得大模型轻量化成为可能，目前已经出现了内置大模型的 PC 和智能手机，而且这种趋势必将快速演进。

随着端侧大模型需求激增，对大模型的安全保护又有了新的要求。现在的大模型都部署在云环境或者一体机中，用户应用大模型都需要通过网络来完成，而端侧则由用户直接与大模型进行交互，通过网络边界提供的安全能力立即失效。

端侧大模型的保护需要安全能力进行针对性设计，不仅要适配端侧算力和存储的要求，最重要的是提供离线使用控制能力，在不损失大模型应用价值的同时确保大模型和输出内容的安全与合规。

2.3.3 后训练是应用关键，数据价值再次攀升

全球范围内，当前通用大模型参数规模已经突破 3000 亿，且发展趋势由预训练转为强化学习主导的后训练。由于距离实现通用人工智能还有很长的路要走，大

模型基础研究依然十分重要。

但后训练涉及的微调、强化学习和规模扩展等技术，其核心之一就是高质量数据。高质量数据通常是由实际工作环境中获得或通过其他高质量数据合成，但由于受生成合成数据的模型自身稳定性的影响，合成数据会具有更高的风险。

基于此，获取高质量数据的关键还在行业真实环境，数据的价值因为人工智能再一次得到升级，由数据资产上升成为知识产权。相应的，对于数据的安全保护难度也再一次升级。

2.3.4 公众模型成为基础设施，智能体百家争鸣

虽然大模型基础研究十分重要，但当前人们已经广泛的认识到，人工智能的价值并不在于模型本身，而在于其深度融合并改造业务场景的过程中。而智能工作流、智能体和具身智能是当前可充分发挥大模型价值的载体。

随着通用大模型竞争逐渐明朗，通用大模型将成为数字化的基础设施，而以领域大模型为核心的智能工作流、智能体和具身智能将展开新一轮的白热化竞争态势，呈现出百家争鸣的现象。

由于智能体应用的发展，将会出现越来越多的具体安全需求，也会相应诞生与之相匹配的大模型业务系统安全解决方案。届时，安全优先的大模型概念必将深入人心，大模型安全保护产品、服务和解决方案也将迎来真正的爆发。

三、推荐供应商

在本报告调研的过程中，数世咨询发现，现阶段可提供大模型安全保护产品、服务、解决方案的供应商共 30 家（不包括通用安全能力）左右，参与本次调研工作的共 23 家。

根据调研数据，结合各供应商在安全优先的大模型领域中的资源投入、AI 研究能力、产品能力、服务水平以及分析师评价，数世咨询评选出 6 家“安全优先的大模型推荐供应商”。

✓ **奇安信**

作为国内网络安全行业领军企业，奇安信集团密切关注人工智能大模型及应用系统全生命周期的主要威胁，深度参与国家相关标准与规范的制定与起草工作。奇安信是信通院“云上大模型安全推进方阵”成员单位、生成式人工智能服务安全应急响应指南《网络安全标准实践指南》起草单位、大模型安全测评标准参编单位、安全大模型能力要求与评估方法核心参编单位。

目前，奇安信针对大模型安全提供涉及安全开发、安全合规测试、安全评估、安全运行防护、智能安全运营与安全响应在内的多项产品与服务，包括大模型安全评估服务、大模型安全卫士、零信任访问控制、数据安全网关、特权卫士、代码

/开源卫士等，致力于为人工智能大模型及应用系统提供全生命周期的安全保障，确保广大政企机构的智能化转型安全顺畅。

✓ 绿盟科技

绿盟科技依托二十余年网络安全深耕与十余年 AI 安全研究，已形成覆盖“研究—产品—运营”全栈的大模型安全能力。公司设有星云、天枢两大 AI 安全实验室，累计发布《大模型安全风险矩阵》《SecLLM 技术白皮书》等权威报告，并参与制定国内首个《云上大模型安全参考架构》，成为“云上大模型安全推进方阵”首批成员。

面向产业落地，绿盟推出“AI-UTM 安全一体机”与“大模型安全围栏”双轮产品：一体机集成“AI-Scan”、“AI-AFW”、“AI-CONT”、AI-DLP”四大引擎，形成“评估+加固、阻断+代答、审计+回溯”三道纵深防线。其中，AI-Scan 贯穿“训练-部署-运营”全周期，内置百余种对抗样本模板与自动化变异算法，可在分钟级完成提示注入、越狱攻击、幻觉诱导等 20 类风险场景的红队测试，并输出 CVSS-AI 评分及修复建议。围栏则以意图识别、提示词过滤、算力熔断为核心，解决 API 滥用、投毒、幻觉等场景化痛点。

在合规与供应链维度，绿盟建立 RAI 负责任 AI 框架，形成覆盖基座、数据、模型、应用、身份五大域、56 子域的评估体系，已为金融、运营商、政务等头部客户完成十余个大模型的合规备案与对抗测试。同时，绿盟开放 AI 安全生态社区，携手合作伙伴持续输出威胁情报、最佳实践与人才培养计划，实现大模型安全的可持续演进。

✓ 联通数科

联通数科推出“智盾·智算安全防护体系”，助力全面识别并应对智算服务中的潜在安全风险，打造端到端的智算安全产品能力，为各类智算应用提供内生式、一体化的安全解决方案。

围绕智算基础设施和模型应用两大核心方向，提供一体化的安全防护方案。基础

设施层面，平台聚焦网络、负载、管理三大关键节点，整合联通 DDoS 防御、网络入侵防御、主机容器安全、平台安全管理审计等系列产品，构建起可防御、可管理、可审计、可溯源的立体化安全防护体系。模型与应用层面，平台重点针对模型、数据、应用三个维度，结合大模型风险评估、内容安全围栏、大模型防火墙、数据清洗审计等工具，形成覆盖智算服务全生命周期的安全保障机制。

多层级联动防护，提升智算公共服务及私有化交付场景的安全效能。融合联通智算基础设施和服务能力，采用平台化思路整合原子级安全能力，实现全局视角的风险评估、分析研判、响应处置与事件溯源，构建起涵盖事前安全评估、事中主动防御、事后追踪溯源的一体化安全架构，整体安全防护效率提升 30%以上。联通数科智盾智算安全防护体系切实保障人工智能技术应用过程中的安全可靠、内容可信、风险可控，是支撑国家安全体系建设的生动实践。

✓ **火山引擎**

火山引擎是字节跳动旗下云和 AI 服务平台，将字节跳动快速发展过程中积累的增长方法、技术能力和应用工具开放给外部企业，通过云和智能技术帮助企业构建体验创新、数据驱动和敏捷迭代等能力，推进企业 AI 转型，激发增长潜能。

火山引擎云安全依托字节跳动在安全技术上的实践沉淀，面向互联网、金融、汽车、大消费等行业输出云上安全能力，保障企业用户网络、数据、云原生、终端、大模型等的安全。同时，紧贴客户需求，重点布局大模型安全、数据隐私安全、AI 安全智能体等领域，致力于在 AI 时代，为企业大模型应用提供最全面的云上安全防护方案。

✓ **安泉数智**

安泉数智深度参与国家人工智能安全顶层设计，参与和起草多项行业标准。以“AI 对抗 AI”的理念推出“大模型安全综合治理平台”。业界首创“RAPAO”大模型安全五步闭环管理模型，从模型训练、部署和运行，覆盖大模型全生命周期十个方面安全问题。该方案包括大模型资产台账系统、人工智能模型评测平台、人工智能增强平台和大模型审计与配置系统以及大模型安全运营系统。

不仅涵盖数据安全、模型鲁棒性与算法合规性等风险，深入大模型安全机理研究，构建起实时溯源、可控的风险监控与资源调度体系。通过有害内容拦截、越狱攻击检测、敏感词过滤等输入管控功能，以及内容合规检查、有害内容替换、隐私数据脱敏等输出管控功能，确保大模型生成内容的安全性与合规性，为开发者与企业提供高效、可靠的安全防护解决方案。“配置审计”精准纠偏，梳理关键配置项，制定基线标准，保障模型配置合规；“安全运营”持续改进，通过实时监测、快速响应与持续优化实现动态闭环，提升风险事件响应速度。助力应对 AI 全生命周期安全挑战，通过持续技术创新，助力 AI 造福人类。

公司通过"资产-评测-防护-治理-运营"五位一体架构，将技术能力与行业需求深度融合，为监管、能源、金融、政务等领域提供从风险预警到主动免疫的闭环防护，树立大模型安全可信应用标杆。

✓ **360 数字安全**

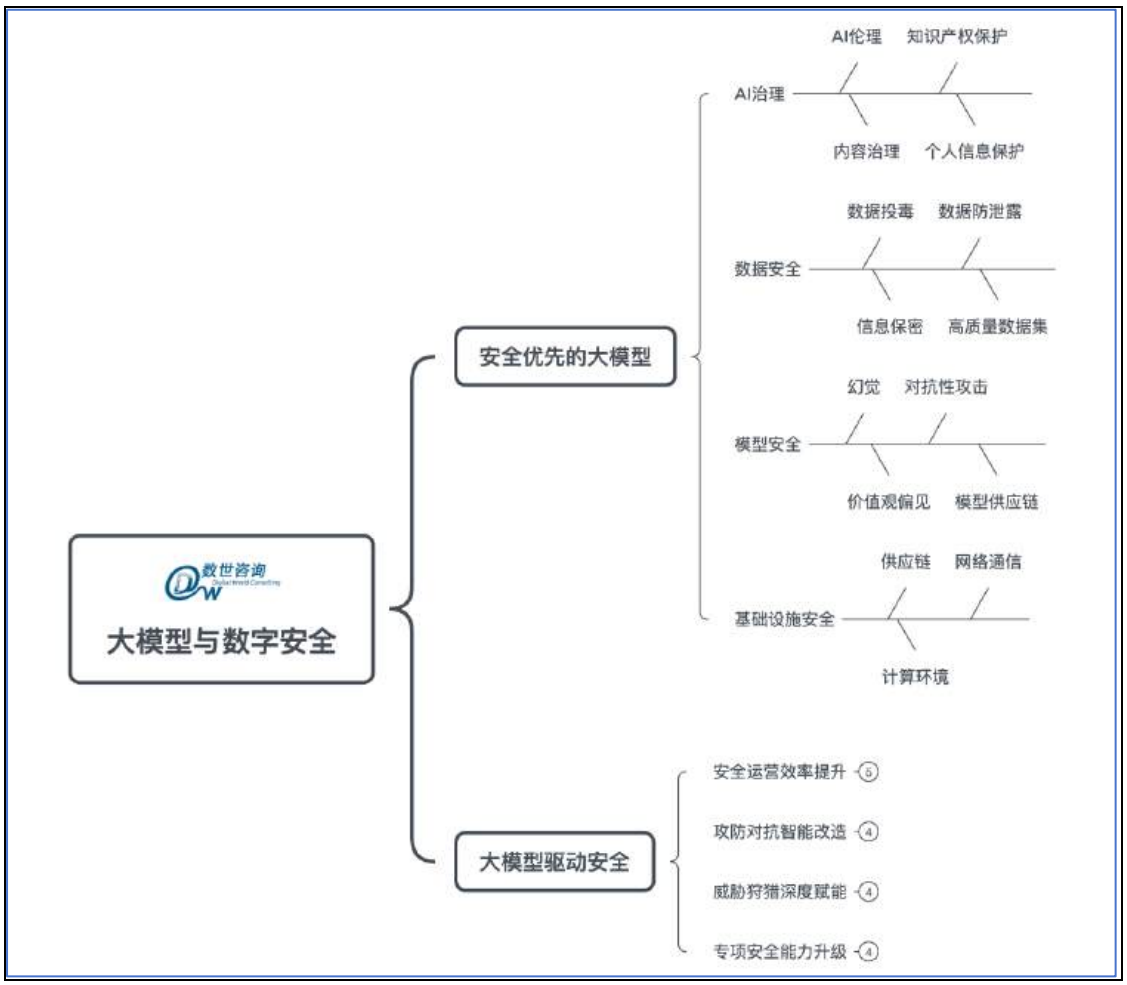
360 依托二十年来网络安全领域的深厚积累、AI 领域的技术深耕以及 AI 业务（360 智脑、纳米搜索等）的安全实践，形成了“懂 AI 更懂安全”的跨领域优势，提出“以模制模”新范式，即利用人工智能技术自身优势对抗 AI 安全风险，打磨出 360 大模型安全卫士，解决 AI 自身安全“可靠、可信、向善、可控”四大核心问题。“可靠”指聚焦模型基础安全问题，智能识别供应链开源软件漏洞和 AI 自身缺陷，实现模型资产闭环管理，保障系统环境安全；“可信”与“向善”则针对模型原生安全挑战，借助幻觉抑制、内容安全防护等技术，保障输出内容真实可信、符合社会良善导向，应对误导、违规风险；“可控”强调智能体执行安全，通过身份认证、权限管控、异常识别等，防范数据泄露与越权操作，确保 AI 行动能力可控。整套体系通过 AI 对抗 AI 的闭环设计，实现了从源头上化解威胁的全局方案。

同时，360 积极配合推动 AI 安全行业发展，作为国家人工智能标准化大模型专题组联合组长单位，公司积极参与国标与安全框架的制定，并牵头发起大模型安全联盟，打造资源共享、共创共赢的生态集群。未来，360 将深化生态协作，致力于为 AI 时代的可持续发展注入强劲动力，助力构建更可信赖的智能世界。

四、安全能力

数世咨询持续关注并研究人工智能领域，在大模型安全方面，已经发布了《LLM 驱动数字安全》（关注“数世咨询”公众号，回复“安全大模型 2024”下载）调研报告，该报告核心方向为利用大模型赋能安全运营。

大模型安全分为两方面，一是利用大模型做安全，一是保护大模型安全。本报告所介绍的安全能力方向为保护大模型安全，即安全优先的大模型。



在理论层面上，大模型的安全保护关注 4 个方向，即 AI 治理、数据安全、模型安全和基础设施安全，如上图所示。

在实际应用的过程中，站在企业用户的角度上，现阶段对于大模型业务的安全保

护则主要围绕大模型业务系统以及数据处理全流程来展开。

4.1 大模型业务系统安全风险

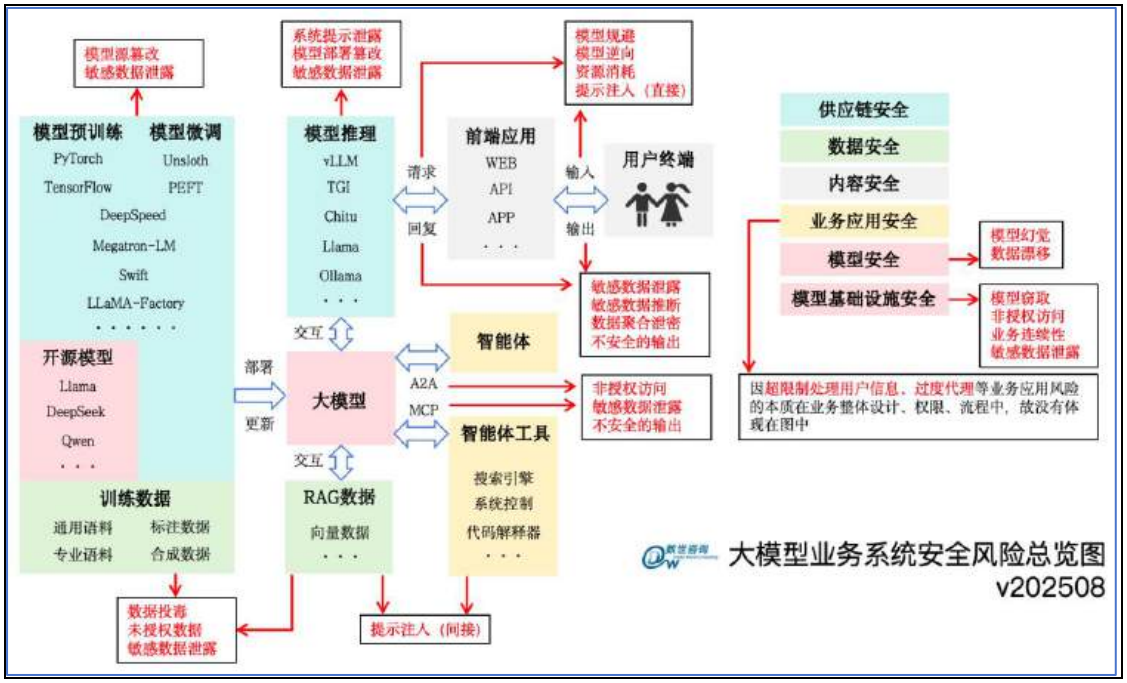
大模型业务系统包括大模型赋能的工作流和基于大模型的智能体，在业务运营的过程中，各个阶段都存在着大模型特有的安全风险以及系统安全风险。

企业用户的大模型业务大致可分为以下 4 个主要阶段：

- 预训练：大型科技企业或科研机构会通过预训练的方式生成自己的通用大模型，其他企业通常会以开源通用大模型或者商业通用大模型作为自己的基础模型代替预训练阶段。
- 后训练：通用大模型的核心是思维推理能力，而解决具体问题的能力是由后训练提供的。后训练包括更加深入的领域知识强化学习以及价值观对齐，企业用户可以自己进行模型微调或者购买专业的领域大模型。
- 开发与部署：拥有领域大模型后就可以进行业务系统的开发、部署，目前较为成熟的大模型业务模式为智能工作流和智能体两类，由于业务的多样性可能涉及与各类应用工具或其他智能体交互。为了使大模型业务应用输出更加准确，通常还会与 RAG 数据进行交互，获取实时数据。
- 推理与运营：大模型业务系统部署上线后即可进行推理任务，与其他信息系统一样，需要持续性的运营工作以保证其业务正常开展。

在大模型业务运营的不同阶段中，由于安全能力不足没有执行相应的管理和控制措施，将会发生诸多类型的特定安全事件对业务产生负面影响，这些负面影响如下图所示：

而这些相应的大模型业务系统安全风险，散布在大模型业务的不同关键节点中，如下图所示：



根据大模型业务系统所面临的安全风险，经分析后整理成 6 种类型，如上图所示：

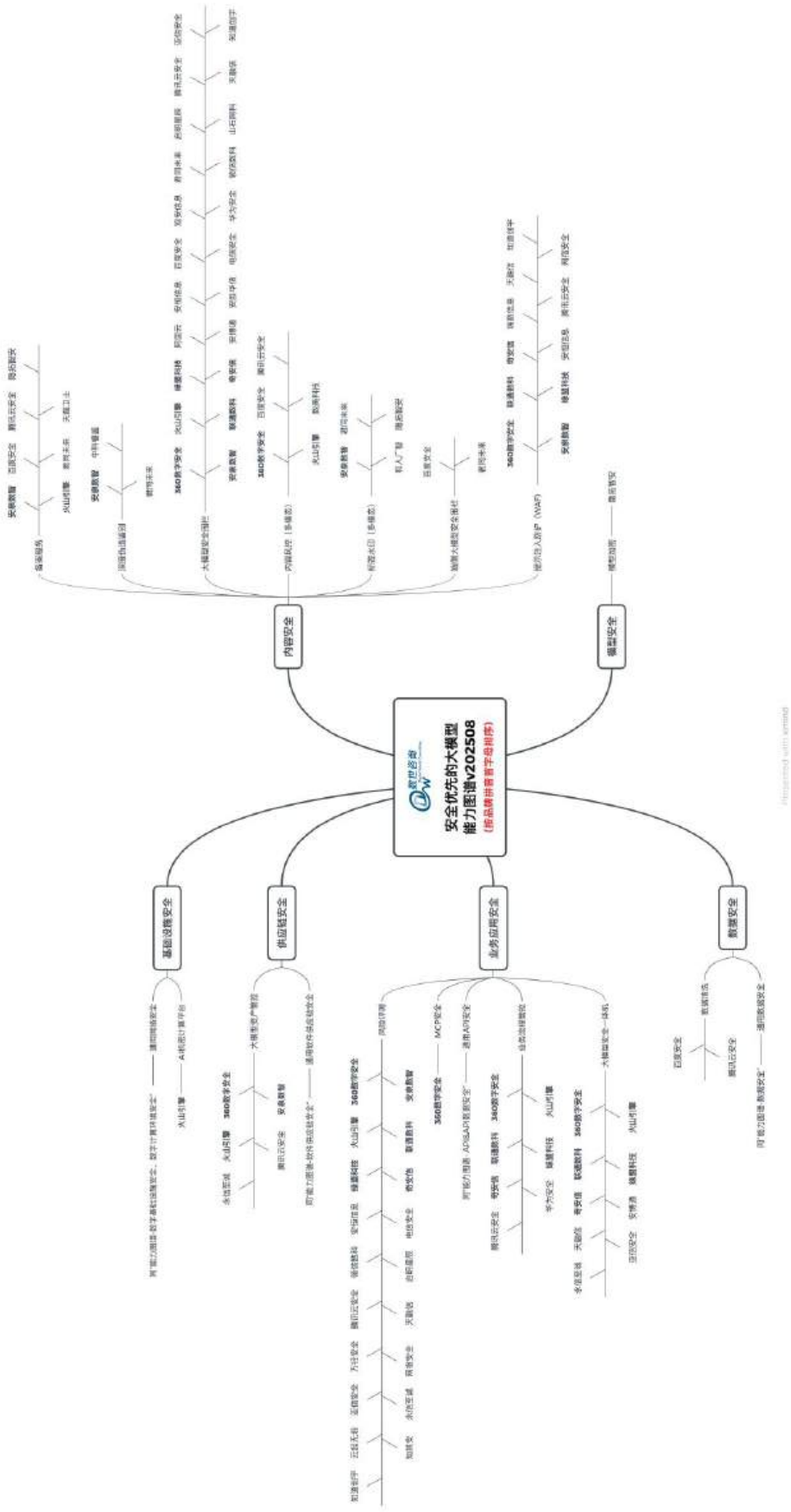
- 供应链安全风险：主要存在于模型预训练、后训练、模型部署、模型推理所使用的各类框架、工具、类库之中，并存在安全风险的传递性，可能发生模型源篡改、敏感数据泄露等安全事件。
- 数据安全风险：存在于数据处理的全流程，集中表现在预训练数据、后训练数据、RAG 数据上，可能发生数据投毒、敏感数据泄露等安全事件。
- 内容安全风险：大模型业务应用特有的安全风险，存在于输入与输出内容之中，可能发生对抗性攻击、敏感数据泄露等安全事件。
- 业务应用安全风险：存在于大模型与工具或其他智能体的交互过程中，目前主流交互方式有 API、MCP、A2A 等协议，可能发生非授权访问、敏感数据泄露等安全事件。
- 模型安全风险：由模型自身特异性决定，不受外部威胁影响即可发生模型幻觉、数据漂移等相应安全事件。
- 基础设施安全风险：存在于大模型业务系统的存储、运行环境中，可能发生模型窃取、非授权访问等安全事件。

4.2 安全优先的大模型能力图谱

为了有效应对大模型业务系统的安全风险、降低安全事件产生的负面影响，数世咨询根据调研信息绘制了“安全优先的大模型能力图谱”，旨在为开展大模型业务的用户，在供应商选择和产品选型方面提供有益的参考。

“安全优先的大模型能力图谱”根据大模型业务系统安全风险，一一对应分为六类，同样为供应链安全、数据安全、内容安全、业务应用安全、模型安全和基础设施安全。

能力图谱中，不同的安全能力即可有效应对大模型业务系统中各个关键节点可能面临的安全风险，完整图谱如下。


安全优先的大模型
能力图谱v202508
(按品牌首字母排序)
内容安全
模型安全
基础设施安全
供应链安全
业务应用安全
数据安全

五、解决方案/典型案例推荐

5.1 大模型安全解决方案

方案背景

随着人工智能技术的迅猛发展，大模型在政务、金融、运营商、医疗、制造等众多领域得到了广泛应用。然而，大模型特有的安全风险和日益严格的合规要求正成为制约其发展的关键因素。

大模型面临的安全风险表现在几个方面：

- 生成内容的不可控性：大模型在生成内容时可能存在偏见、虚假信息（幻觉现象）、道德争议性内容等问题，难以完全预测和控制输出结果。
- 大模型应用下新的攻击方式：恶意用户可以通过设计特殊的输入（Prompt 注入），绕过模型的安全规则，使其生成敏感，或通过注入达到入侵业务系统的目的。
- 模型算力耗尽导致业务连续性中断：攻击者可能通过诱导模型执行复杂推理链或无限任务循环，从而引发算力耗尽型拒绝服务（Compute-DoS）攻击。

方案概述

绿盟科技凭借多年网络安全领域的技术积累，推出"绿盟大模型安全解决方案"，该方案由大模型安全评估系统（AI-SCAN）、AI 安全一体机（AI-UTM）两款产品组成，形成覆盖大模型全生命周期的安全评估和防护体系。

在模型训练和微调阶段，大模型安全评估系统（AI-SCAN）发挥着关键作用。该系统基于《大模型系统安全测评要求》等标准规范，对大模型进行全方位"体检"。通过内置的 10 万+测试用例库，系统可模拟提示词注入、数据投毒等 21 类攻击手法，检测模型在内容合规性、对抗防御能力等方面的薄弱环节。特别是在供应链安全方面，AI-SCAN 能深度扫描.pb、.h5 等 15 种模型文件格式，识别后门植入风险，并对 Ollama、Ray 等 450 多个大模型组件进行漏洞检测，从源头保障模型安全。

在模型部署和应用阶段，AI 安全一体机（AI-UTM）提供关键的运行安全保障。该产品采用独特的"三体防护"架构，在内容安全方面建立三级过滤机制：基于 30 万+敏感词的词法检测实现毫秒级响应；通过自研风云卫模型进行语义理解，识别变体违规内容；利用 128K tokens 的上下文记忆窗口确保多轮对话中的精准判断。在算力安全方面，其可将算力资源划分为保障级、普通级和限制级，通过预测算法防止 Token 耗尽攻击，保障模型服务安全稳定。

在应用和智能体运行阶段，AI 安全一体机（AI-UTM）针对大模型特有的漏洞攻击场景，构建了多维度、智能化的防护体系，通过深度语义分析及动态检测引擎，精准拦截 SQL 注入、XSS、SSRF 等传统 Web 攻击，防止攻击者利用漏洞入侵大模型后端服务或窃取数据。通过多维度检测机制（如关键词过滤、上下文语义分析、异常输入模式识别），阻断恶意构造的提示词输入，避免模型被诱导输出违规内容或泄露训练数据隐私，保障模型应用安全。

解决方案核心价值

✓ **全生命周期安全防护**

◆ 训练和微调阶段：通过 AI-SCAN 进行模型安全评估，识别训练数据投毒、后门植入等风险

◆ 部署和应用阶段：利用 AI-UTM 提供内容安全防护、算力资源管控和数据泄露防护

◆ 模型和智能体应用阶段：通过 AI-UTM 实现业务应用/API 暴露面的软件漏洞及应用层攻击防护

✓ **多维安全能力融合**

◆ 安全运营：组件间联动运营，AI-SCAN 评估的风险可输入给 AI-UTM、AI-UTM，从而生成安全防护策略，形成 AI 安全运营和闭环

◆ 合规性保障：满足 TC260-003 技术标准、大模型备案等合规要求

◆ 全面风险识别：覆盖提示注入、越狱攻击、敏感信息泄露等 21 类对抗风险

AI 安全产品详解

✓ 大模型安全评估 AI-SCAN

AI-SCAN 是一款专业的大模型安全评估工具，凭借专业人员精心筛选和校准的高级知识库，该系统可高效精准地检测大模型在生成内容安全、对抗防御能力以及供应链安全三方面可能存在的隐患，并且可通过自定义导入企业内部风险库进行针对性的大模型安全风险智能化评估，最后通过详尽的可视化风险评估报告为用户提供深刻洞见。

➢ 核心功能

◆ 内容合规评估：严格依据 GB/T 45654-2025 标准，通过多维度评估引擎，实现对模型输出内容的全面安全合规验证

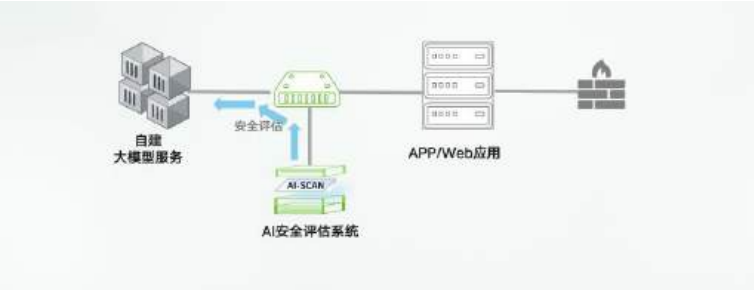
◆ 对抗防御评估：覆盖模型越狱、Prompt 泄露、角色逃逸、反演攻击等 7 大类 22 小类对抗安全风险

- 模型后门检测：提供先进的恶意模型后门检测分析技术，覆盖 15+种主流 AI 模型文件格式的后门风险检测
- 模型组件漏洞扫描：覆盖数据处理访问、训练部署、ML Ops 等 13 个大模型全生命周期中涉及的组件及 Web 应用服务的漏洞检测.漏洞数量 3000+。
- 自定义题库智能评估：行业特色或特定场景化题库快速导入，内置匹配类、智能评估类、拒答类等多种评估器灵活适配不同题库场景

➢ 技术优势

- 全面性：覆盖伦理对齐、对抗攻击防护、供应链检测等多个维度
- 创新性：采用"以模治模"、"高效匹配"、"拒答判断"等多种评估方式
- 高效性：支持并行处理，单任务评估时间<30 分钟
- 兼容性：全新大模型分钟级适配接入，简单快捷
- 简洁性：跟踪说明风险检测的全过程，采用易读易懂的方式展示每条风险详情
- 灵活性：除内置多种题库外，可灵活增加其它特定场景题库评估

➢ 典型部署场景



AI 安全评估系统旁路部署，生成多样化的对抗攻击样本和内容合规风险样本，用于评估各版本大模型在不同应用场景中的输出内容安全性

✓ AI 安全一体机 AI-UTM

AI 安全一体机是专为大模型场景设计的新一代安全网关，采用创新的"三体防护

″架构，深度融合规则引擎与 AI 算法，提供内容安全防护、算力资源管理、数据泄露防护和大模型安全评估四大核心能力。为大模型基础组件安全、大模型自身安全、大模型应用安全、大模型数据安全，提供分层递进的防护能力。

➢ 核心功能

◆ 内容安全防护：三级内容过滤体系（词法、语义、上下文），128K tokens 记忆窗口

◆ 提示词加固：对传递给大模型的指令，可配置策略限制提示词语境环境，有效降低大模型自身安全风险。

◆ 算力资源管理：三级优先级动态分配策略，智能预测算法防止系统过载

◆ 数据泄露防护：敏感信息识别准确率超过 99%，支持文本、图片等多模态内容识别

◆ 全链路审计：支持智能体应用、大模型 API 输入输出的全链路安全审计

➢ 技术优势

◆ 场景接入灵活、全面：同时防护大模型流量+传统 Web 流量，快速兼容各类大模型应用和传统 web 应用藕合的客户场景

◆ 高性能：毫秒级实时响应，支持流式检测、不影响模型业务模式

◆ 高可靠：可集群部署，服务可用性≥99.9%

◆ 易管理：可视化控制台，多维策略配置集中管理

➢ 部署模式（部署在模型前）

◆ 典型场景 1

AI 安全一体机部署在大模型服务的 API 接口前，隐藏大模型服务真实 API，实现 AI 网关、key 和 token 管控、内容安全合规、提示词攻击防护、算力攻击防护、智能体级对话审计等功能，提供一站式大模型服务安全防护

◆ 典型场景 2

AI 安全一体机部署在调用大模型能力的业务应用/Web 服务前，对外隐藏业务应用，实现内容安全合规、提示词攻击防护、算力攻击防护、数据泄露防护、用户级对话审计等功能，提供一站式大模型服务及智能体应用安全防护。

本方案由绿盟科技提供

5.2 数字政府智算服务一体化安全保护案例

项目背景简介

某省政府已完成能智能支撑平台建设，平台采用“115+N”架构进行建设，打造 1 个知识中心、1 个模型中心、5 个智能化支撑平台，为 N 个场景智能化建设提供支撑服务。平台及其生产的智能应用（智能问答、智能写作、智能搜索等），在智算基础设施、模型、应用、内容层面需要一体化安全能力建设，以满足规划要求以及内生安全建设需求。

大模型业务系统安全保护需求

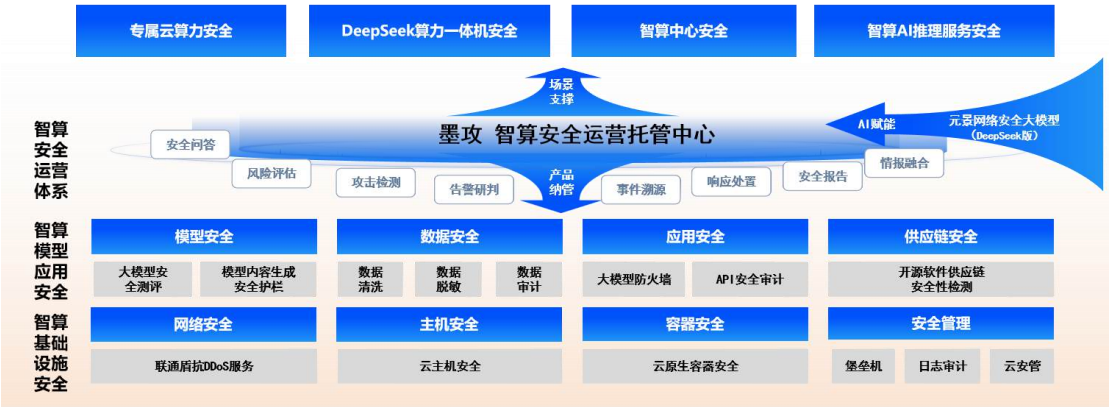
由于平台所生产的智能应用，需要面向公众提供 AI 服务，所有智能体均采用统一平台构建，对于 AI 服务在模型应用以及内容安全层面的安全风险和防护需求，主要集中在三个层面：

- 合规性内生安全：以《生成式人工智能服务管理暂行办法》为依据，针对对公服务的智能化应用开展备案前的风险评估以及日常风险巡检，及时发现智能化应用的潜在风险问题，联合其他工具进行针对性的安全防护建设；
- 应用层统一防护：省级智能化应用存在海量的潜在用户，需要进行一体化的应用层安全防护能力建设，规避 Web 攻击、API 异常调用、提示词注入、算力消耗等攻击风险，同时需要以低延时的效果提升用户的交互体验；
- 多模态内容安全：智能化应用涉及到 AIGC 的多个领域，对于输出的多模态内容需要进行安全过滤，同时针对用户提交的敏感问题，需要智能化安全代答，在保证输出内容合规的前提下，对用户进行向善引导；

整体解决方案

联通智盾·智算安全防护体系，依托联通“国芯+国算+国模+国盾”四位一体的战略布局，基于运营商云网数智资源禀赋，在网、端、管、控等基础设施安全能

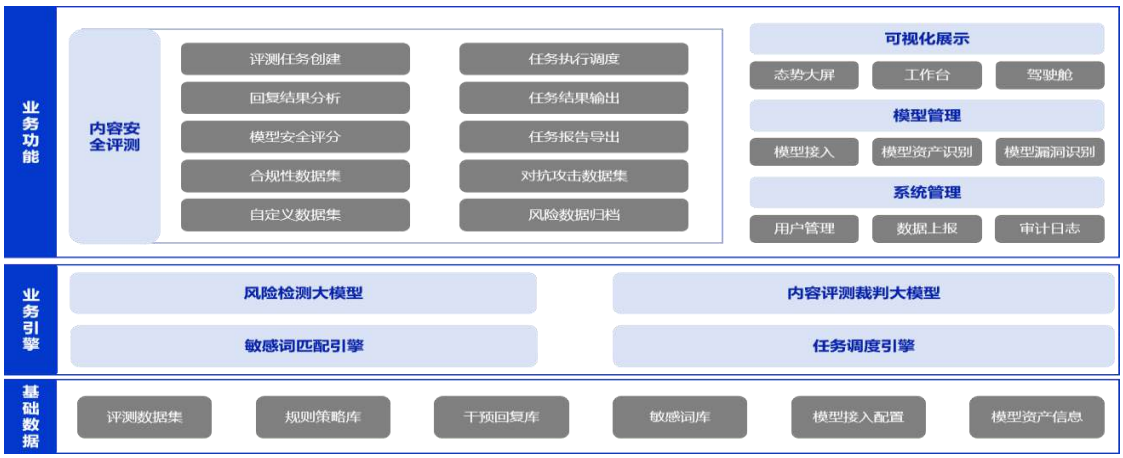
力之上，深挖智算服务在模型、应用、内容层面的安全风险，通过大模型风险评估、大模型防火墙、大模型安全围栏等产品，在智算安全运营支撑体系下，切实保障人工智能的安全可靠、内容可信、风险可控。



✓ 大模型风险评估

大模型风险评估评估是一款专注于大模型安全评测的自动化工具，提供一站式的模型接入、数据管理、安全评测、任务管理及结果分析能力。通过预置的海量安全评测数据集和大模型自动泛化生成的攻击数据集，对模型供应链、应用层漏洞、模型内容安全、合规性满足、对抗攻击防御等方面进行综合评估。

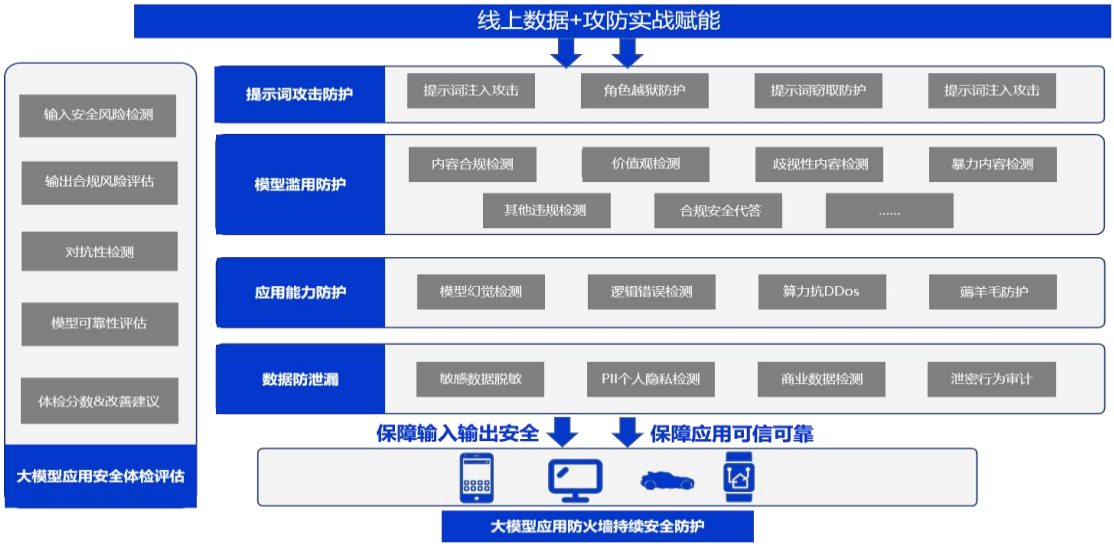
- ◆ 多维度测评数据集：内置海量行业标准安全测评数据集，覆盖国家安全、公共安全、伦理安全等多个评测维度，可通过大模型自动泛化生成攻击样本的能力，确保测评的全面性和时效性。
- ◆ 智能化精准测评：通过优质题库+专业裁判大模型，采用先进的自然语言处理技术和智能算法，可实现高效、精准的安全检测。
- ◆ 模型资产风险管理：提供模型基础设施安全、模型组件安全、模型应用安全等扫描能力，可覆盖服务开发、中间件、向量数据库等 35 种模型组件，可识别 22 种常见安全问题。



✓ 大模型防火墙

大模型防火墙，整合传统 Web 应用防火墙能力，针对智算服务使用场景，增加对于内容安全、提示词防护相关能力，实现对文本内容的输入攻击检测、输出安全过滤、敏感问题安全代答等功能，供给 All in one 的智算应用安全事中防护能力。

- 一站式安全能力：整合基础安全防护组件，实现对传统 DDoS 和网络攻击的全方位防护，支持多模态注入攻击和有害内容检测，通过端侧水印技术实现高精准防薅羊毛与防爬虫，极大避免了因 tokens 盗用带来的经济损失。同时，将 Web 漏洞防护、抗 D 能力默认统一接入，一站式解决所有大模型 API 安全问题。
- 安全防护低延时：风险监测延迟普遍在 150ms 以内，成功接入优化后，延迟稳定在 100ms 以内，可为后续高并发场景奠定性能基础。
- 流式安全检测：具备业内领先的流式输出检测和拦截能力，由专业化和持续迭代的专业安全模型来识别恶意提示词，并对敏感问题进行代答和正向引导。



✓ 大模型安全围栏

大模型安全围栏是专门为大模型服务提供方打造的多模态内容安全防护系统，通过风险内容检测、敏感问题代答等能力，帮助大模型过滤有害输入和输出内容，防止大模型生成不良信息。

- ◆ 多模态内容检测：基于深度学习和大语言模型技术，可检测文本、图像、音频、视频、代码等多模态输入/输出内容，覆盖政治敏感、暴力违禁、虚假信息等超 100 种风险类型
- ◆ 智能化安全代答：对于敏感非拒答问题，通过干预库和安全回复大模型两个模块实现智能化安全代答，既能保证回答的广覆盖，也能提供精准匹配回答，引导输出内容安全向善。
- ◆ 一体化风险运营：风控运营系统还通过规则引擎提供了细粒度的风控尺度调控，支持不同业务场景下不同的风控松紧度，提供风控数据统计功能，量化业务侧风险水位、防护效果和护栏的价值。


应用场景
文生文
文生图
文图生图
文生音视频
多模态大模型
能力层
大模型输入输出内容安全检测
100+检测子类，依据场景快速配置
大模型安全代答
百万级代答库，高度安全对齐大模型
模型层
文本安全检测
关键词识别/特征匹配
情感立场分析
深度语义分析
对抗攻防
图像安全检测
图像内容分析
图像分割检索
特定人识别
OCR识别
音频安全检测
ASR语音转文本
语义分析
声纹识别
背景音检测
视频安全检测
关键帧检测
特定人识别
视频内容分析
视频匹配
数据层
商业Logo库
特定人物库
特定声纹库
敏感标识
谣言库
敏感关键词库
有害违规库
图像黑白名单
敏感特征库
高质量测评集
政策法规
违反社会主义核心价值观
歧视性内容
商业违法违规
侵犯他人合法权益
无法满足特定安全要求
《生成式人工智能服务管理暂行办法》
《生成式人工智能服务安全基本要求》

本案例由联通数科提供

5.3 科技制造业大模型安全防护案例

项目背景简介

作为一家全球化布局的高科技制造企业，该组织在国内外拥有多个分支机构。公司高度重视信息安全，并严格遵循各业务所在国的监管合规要求。该组织积极推动生成式人工智能（GenAI）在内部运营与生产系统的深度应用。

当前重点聚焦于

- 跨部门知识管理与共享： 利用 GenAI 提升内部知识沉淀、检索与流转效率，赋能跨部门协作。
- 智能化文档处理： 应用 GenAI 技术实现文档的自动生成、摘要、翻译与关键信息提取，优化办公流程。

大模型业务系统安全保护需求

在上述应用实践中，确保以下方面至关重要：

✓ **严格的信息安全**

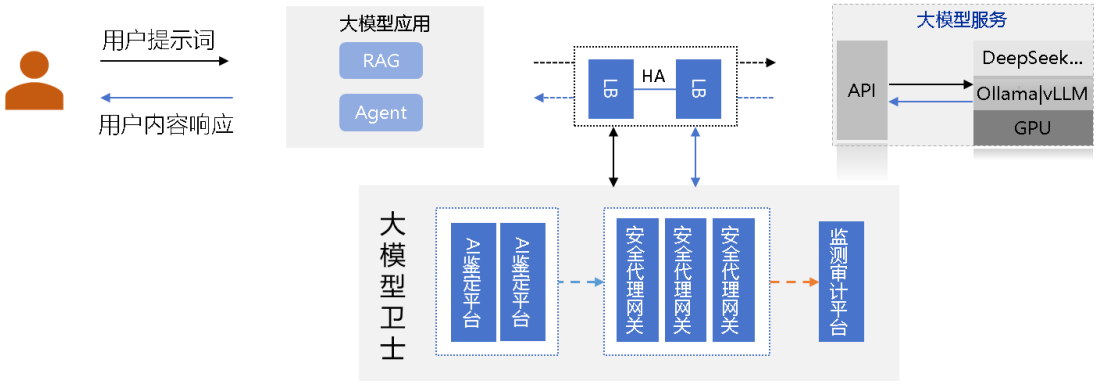
保障公司敏感数据、知识产权及员工隐私在 AI 应用全生命周期的安全防护。组织已经在内部网络部署了多个生成式人工智能大模型，并为内部的多个应用提供模型服务。大模型的引入不仅增加了应用与数据的暴露面，攻击者同时有了新的攻击方式，例如利用提示词注入攻击操纵模型、利用向量与嵌入漏洞越权访问数据等，现有的安全防护手段无法有效发现和处理这些新型攻击。

✓ **属地化人工智能合规**

满足各分支机构所在国家/地区关于数据主权、跨境传输及人工智能使用的特定法律法规要求。例如：欧盟《人工智能法案》、GDPR、《GPAI 行为准则》、《生成式人工智能服务管理暂行办法》等。

用户体验与连续性：安全措施的部署和运行，应最大限度的减少对用户使用大模型应用的最终体验的影响，例如请求处理和答复响应的时间延迟；方便进行安全能力的扩展以适应业务处理能力的变化，并满足业务对安全组件的可靠性要求。

整体解决方案



✓ 模型使用可见

通过旁路 API 访问代理引流的方式，对应用系统与模型服务之间的 API 请求和响应及交互的内容进行解密、完整可见与内容记录。为合规与安全访问审计提供全面的数据支撑。

✓ 模型输入检查

通过协议解析对应用系统调用模型的提示词内容进行分离和多维度的检测，使用多种检测技术包括关键词引擎、规则引擎、语义分类引擎，发现输入内容中存在安全与合规风险包括：提问内容违规、敏感数据泄露、隐私数据违规、提示词注入攻击等。对发现的潜在风险根据预设的处置策略进行代答、告警等处置动作。

✓ 模型输出过滤

对模型生成的内容进行输出合规与安全检查，确保模型生成的内容符合组织内外部的合规与安全要求。基于安全策略对发现风险的内容进行事件告警、内容改写等处置动作。

✓ 体验与业务连续

通过优化引擎检测技术及组合设计多种检测与处置的协同模式适应对用户体验的要求；通过安全能力的集群化部署配合负载均衡设备进行大量请求负载的动态处理。

核心安全能力

✓ 组件化分离架构

通过 AI 鉴定平台、安全代理网关（SWG）、监测审计平台三个组件实现，输入输出全链路防护。组件架构为性能扩展、环境部署、应用集成提供了极大的便利性。

✓ 高效的检测引擎

- 采用“多引擎协同+动态检测”技术：内置多个风险鉴定引擎。首创分层拦截架构，实现“字符级过滤→攻击模板识别→意图分析”三级防护。可以有效检测针对大模型的新型攻击。
- 自研的安全对抗防御引擎：基于 transformer 的预训练检测引擎，可实时拦截 70+类攻击手法（例如：提示词注入、模型对抗攻击）；
- 流式风险评估：支持上下文感知机制，实现流式 Token 实时风险评分，实时拦截恶意 Prompt、实时中断有害内容输出，避免有害信息扩散。
- 合规性检测：内置 TC260《生成式人工智能服务管理暂行办法》合规检测能力，覆盖数据隐私、法律合规等主要风险场景；
- 敏感数据检测算法 ：运用先进的敏感数据检测技术，对大模型的输入数据进行实时扫描和分析，能够准确识别并拦截包含企业核心商业机密、个人隐私信息等敏感数据的投喂行为。

✓ 全面的兼容性

与主流的大模型应用和技术架构具有良好的兼容性，可无缝集成到企业现有的 IT 环境中，降低企业的部署成本和复杂度，快速实现对大模型的安全防护升级。

✓ 主动防御与智能分析

利用 AI 技术的自我学习和进化能力，对大模型的安全风险进行主动预测和分

析，提前发现潜在的安全威胁并采取相应的防护措施。同时，通过对大量安全数据的挖掘和分析，不断优化和完善安全防护模型，为企业提供更智能、更精准的安全防护服务。

✓ **安全能力的持续更新**

丰富的攻击预训练样本及红蓝对抗测试验证。持续收集来源红队、自有情报等各方面，基于自然语言规则的数百万条风险样本。组织大规模人工渗透测试，验证系统防御能力。

本案例由奇安信提供

5.4 联想携手火山引擎推出可信个人云案例

项目背景简介

2023 年以来，大模型技术快速成熟，特别是 DeepSeek、豆包大模型等国产模型的崛起，推动中国人工智能进入高速发展期。大模型的引入显著提升了三大核心能力：

- 意图理解维度：通过千亿级参数对自然语言的深度解析，使智能终端能精准捕捉用户模糊需求（如多轮对话、隐含语义识别）；
- 服务泛化能力：单一模型可同时支撑搜索、内容生成、设备控制等跨场景任务，大幅降低传统 AI 的场景定制开发成本；
- 持续进化特性：基于在线学习的模型迭代机制，使终端服务能动态适应用户行为模式演变。

与此同时，这些能力提升也带来了新的安全挑战：模型训练依赖的海量数据包含敏感信息，推理过程的实时交互需求迫使部分计算前移至终端，传统基于边界防护的安全架构已无法满足"数据不动模型动"的新型范式。行业亟需构建"智能动态防御"与"大模型安全防护"双体系并行的新一代安全架构。

联想作为 AI PC 领域的先行者，始终将安全视为智能体验的核心基石。2023 年 4 月推出的全球首款真正意义上的 AI PC，即以端侧数据隐私保护作为五大核心特征之一;2024 年 5 月发布的天禧个人超级智能体,进一步通过"端-云"混合架构，以用户数据在端、云之间传输和处理过程中的绝对安全为目标。

作为国内 PC 领域首个可信个人云方案，联想个人云基于火山引擎 Jeddak AICC 平台打造，旨在构建严密可信的云上计算环境，全面保障大模型推理、RAG 检索增强生成、AI Agent 等核心能力的数据安全，并以更强的开放性与软硬适配能力，支撑多样化的企业部署场景。

大模型业务系统安全保护需求

随着端云协同成为智能终端发展的主流方向，大模型服务正从传统的云端集中式部署向终端本地化预装演进。这一趋势在提升 AI 体验的同时，也对安全与性能提出了更高要求：

✓ **安全需求：端云协同下的数据隐私保护**

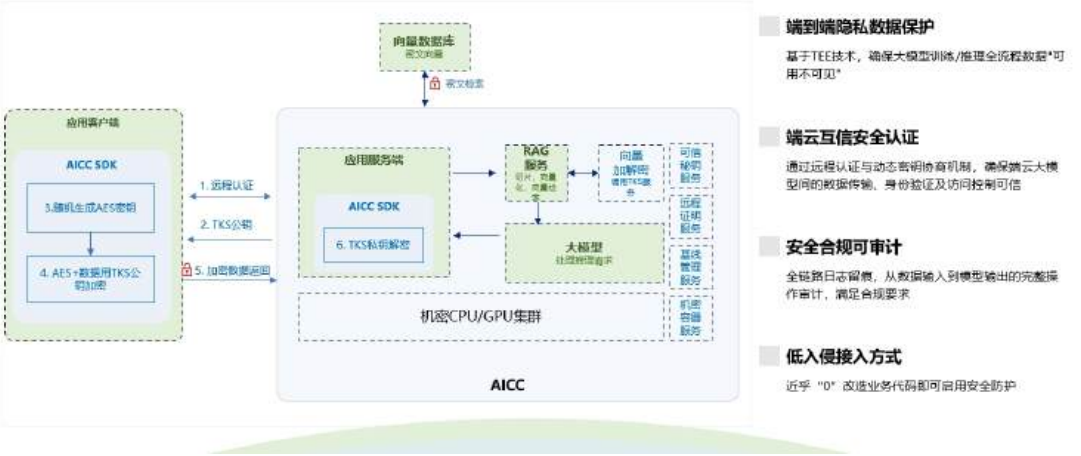
在智能办公场景下，用户依赖大语言模型（LLM）完成文档智能总结、交互式问答等高阶任务，但云端模型处理需频繁上传文档数据，存在泄露敏感信息的风险。例如，企业会议纪要、个人隐私文件等一旦在传输或云端存储过程中被窃取，将造成严重的安全隐患。因此，联想 AIPC 亟需构建端云协同的可信安全架构，确保数据在本地预处理、加密传输及云端计算的全链路安全，真正实现“数据可用不可见”。

✓ **性能需求：安全防护不影响流畅体验**

作为办公、学习及娱乐的核心设备，PC 需在 AI 文档处理、语音助手、实时搜索等场景下提供毫秒级响应，而传统云端安全校验机制（如数据回传、鉴权延迟）可能成为性能瓶颈。联想需优化端侧 AI 算力调度，结合轻量化模型本地推理，在保障安全的同时，提升 AI 助手的交互流畅度，从而增强用户粘性，巩固联想 AI PC 的市场竞争力。

整体解决方案

联想个人云以火山引擎 AICC 方案为基础，充分发挥其全链路 100%加密保障、可自证清白的透明服务以及良好架构实现的能效平衡等优势，构建严密可信的云上计算环境，提供更强的开放性、适应性，构建 AIPC 应用坚实、可信的算力底座，全面保障大模型推理、RAG 检索增强生成、AI Agent 等核心能力的数据安全。



基于个人云安全方案，联想在知识库构建等典型应用中，已实现从内容创建、密态存储到加密检索与解密输出的全流程端到端隐私数据链路流程闭环。用户无需改变操作习惯，即可获得快速响应、可信输出的智能反馈，实现“安全无感”的日常体验，让 AI 服务真正成为可感知、可信赖、可持续的终端能力。

除全链路加密等核心安全能力以外，该方案在设备兼容性方面也展现出高度适配性：不仅支持 PC 场景，也面向 ARM 架构进行了深度优化，覆盖手机、平板等多形态终端，并通过在私密云中集成 NVIDIA NVLink 与 NVSwitch 等高带宽互联技术，实现跨设备的 AI 能力流转与数据安全统一调度。

核心安全能力

火山引擎 AICC 机密计算平台基于 TEE（可信执行环境）等前沿机密计算技术，为企业构建云端大模型的“安全计算空间”，从根源上消除数据在云端处理时的泄露风险，让企业真正“敢上云、敢用云”。

核心功能包括：

- 芯片级硬件隔离方案：在 AICC 环境中可对隐私数据进行计算和处理，全程外界无法查看原始数据内容，确保敏感信息不泄露。
- 全链路密文流转：数据上云传输和计算过程中，始终以加密形式存在，确保数据在不可信环境中的安全性和隐私性。
- 数据即用即销毁：计算完成后自动彻底删除原始数据及中间结果，不留存副本，杜绝数据在计算过程中留存的风险。
- 安全可信可证明：可信证明服务确保计算环境、过程及结果的可信性与透明度。

本案例由火山引擎提供

5.5 央企大模型综合治理案例

项目背景简介

某央企作为电力能源行业的领军企业，业务覆盖煤炭开采、电力生产、油气输送、新能源开发等全产业链环节，其核心业务系统承担着能源生产调度、设备运维管理、客户能源供应服务及安全生产监管等关键职能。2025 年以来，该企业全面拥抱大模型，围绕集团战略规划、市场营销、工程建设、生产运维、安全环保和智慧应用等领域，构建了上百类智慧模型和智能体应用，促进整个集团数智化和智能化建设，实现核心业务能力的智能和效率。

整体解决方案

针对大模型安全风险整体情况，安泉数智联合企业共同围绕大模型训练、部署和运行三个阶段，总结出了十个方面的问题，并针对提出全生命周期的完整解决方案。

大模型资产台账系统为模型资产和训练数据提供一站式档案管理，为应用上架和管理提供全维度数据。人工智能模型评测平台通过自动化问答机制，评估目标大模型的输出内容安全性、数据泄露风险等，并提供整改建议。人工智能增强平台（即大模型防火墙）作为一道屏障和代理，抵挡在目标大模型之前，进行问答内容输入输出的管控，防止大模型的幻觉，或者回复恶意信息、被引导信息泄露。

大模型审计与配置系统能给大模型配置合规和运行提供全面风险监测和审计。大模型安全运营系统为模型和智能应用提供多维度、实时安全态势和运营情况，为模型攻击和防御处置提供决策参考。

整体架构图如下：

核心安全能力

✓ 资产管理：摸清家底

通过构建“全生命周期模型资产库”，实现了从训练、部署到运行的全链路精细化管理。通过模型和智能体台账，对基础大模型、微调模型、衍生智能体行唯一标识与元数据登记（包括版本号、训练数据来源、适用场景、责任人），确保“底数清、权属明”；通过部署模型版本控制与访问权限管理系统，记录每一次修改、分发与部署的操作日志，防止未经授权的篡改或扩散。该体系为企业模型技术的合规应用与知识产权保护提供了坚实基础。

✓ 风险评测：高效预警

平台以自动化与智能化为核心，覆盖合规性、鲁棒性、对抗性、隐私性等维度，融合自动化测试工具与机器学习算法，自动生成多样化测试用例，覆盖大模型全

环节潜在风险，智能分析结果以快速定位漏洞，并借助并行处理缩短测评周期。同时构建全生命周期测评闭环，在安全开发阶段提供规范指导与环境扫描，训练阶段实施实时监控与性能评估，应用阶段开展常态化监测与快速响应，持续优化安全策略。

✓ **风险防护：智能拦截**

以“AI”对抗“AI”为指导思想，训练了 9 类小模型和 2 个安全垂域大模型，能融合实时监测、智能分析与动态拦截技术，具备实时流量分析能力以识别异常行为，集成深度学习模型提升未知威胁识别能力，支持自定义防护策略。通过输入过滤、输出把关、运行防护建立覆盖部署、运行、升级全流程的动态自适应机制，部署阶段进行前置合规检查，运行阶段实时监测并拦截攻击，升级阶段同步更新防护规则，有效筑牢了技术应用的“安全护城河”。

✓ **配置审计：合规保障**

模型配置的合理性直接影响安全策略的执行效果，而配置偏差往往是引发风险的“隐性漏洞”。梳理模型运行的关键配置项（如内容过滤规则的严格等级、API 访问白名单范围、日志记录的详细程度），制定《模型安全配置基线标准》，明确不同业务场景（如面向公众的开放服务、企业内部的专用工具）的推荐配置模板；通过自动化扫描工具定期检查实际配置与基线标准的差异，并结合日志分析验证配置的实际执行效果，为模型配置的“精准合规”提供了制度与技术双重保障。

✓ **安全运营：持续改进**

安全运营是将静态能力转化为主动防御的动态安全体系，通过实时监测、快速响应与持续优化实现安全防护的动态闭环。在监测层，整合模型资产管理、风险评测、防御增强等模块的数据，基于 AI 算法对异常行为进行智能研判，按照风险等级自动推送告警信息至安全运营团队，并通过可视化大屏展示全局风险态势；在处置层，制定标准化应急预，联动多部门协同响应，确保风险事件“发现即处理”；企业模型安全事件的平均响应时间从小时级缩短至分钟级，真正实现了从“被动救火”到“主动护航”的能力跃升。

该案例通过“模型资产管理”夯实基础、“风险评测”把好入口、“防御增强”主动免疫、“配置审计”精准纠偏、“安全运营”持续进化，构建了覆盖大模型与智能体全生命周期的体系化安全核心能力，不仅为企业提供了从技术到管理、从预防到处置的整体防护方案，更通过动态闭环机制推动安全能力与业务发展同频共振，为人工智能技术的可信应用树立了标杆实践。

本案例由安泉数智提供

5.6 杭州市数据资源局大模型安全防护案例

项目背景简介

✓ 业务系统功能简介

杭州市数据资源管理局是杭州市人民政府下属的核心部门，负责全市政务数据的统一管理、开放共享和应用推广。2024 年，随着杭州“市政大模型”项目的落地，数据局承担了统一部署与对外服务的职责，政务系统内的大模型应用（包括政务咨询、智能客服、政务信息检索、政策解读等）均需通过数据局提供的统一接口进行调用。

✓ 大模型的能力提升

大模型的引入显著提升了政务系统的智能化水平：

- 自然语言交互能力增强：市民可通过自然语言与政务服务系统交互，大幅降低了使用门槛。
- 知识覆盖面广：大模型可对政务政策法规、公共服务内容进行快速解答，提高政务信息服务效率。
- 服务自动化程度提升：减少人工客服压力，提升政务服务的响应速度与市民满意度。

大模型业务系统安全保护需求

✓ 引入大模型带来的安全风险

政务大模型在应用过程中面临多类安全挑战：

- 输入风险：市民可能无意或恶意输入包含敏感、违规、涉政涉恐或违法不良信息的内容。
- 输出风险：模型可能生成不当言论、虚假信息或不符合政务导向的回答，导致舆情风险。
- 攻击风险：存在越狱提示注入、敏感信息窃取等新型对抗风险。

✓ 合规需求

为确保政务应用的合规性与可靠性，大模型服务需满足以下要求：

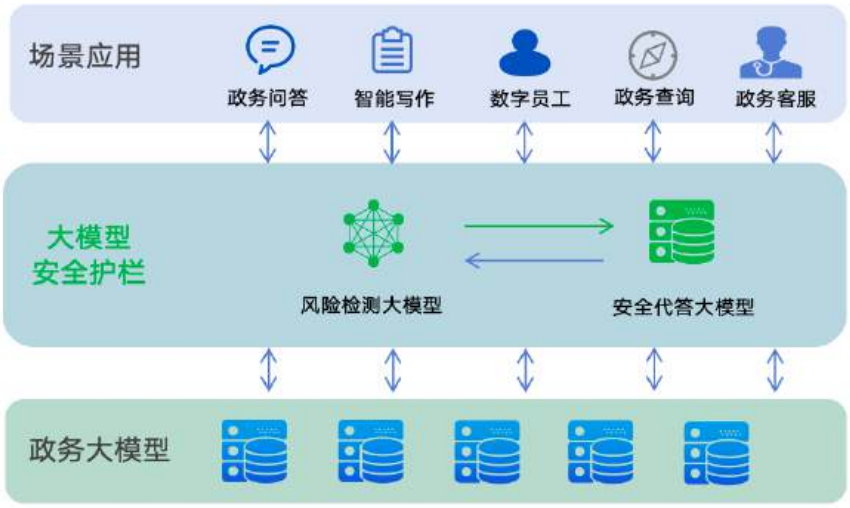
- 符合国家《生成式人工智能服务管理暂行办法》要求。
- 满足网信办等监管部门关于政务服务场景的合规规范。
- 实现对违法不良信息的及时识别与处置，确保政务系统服务过程中的 可控、可靠、安全。

整体解决方案

✓ 整体思路

杭州市数据局在政务大模型部署过程中，引入 360 智脑大模型安全护栏作为核心安全防护组件。该系统通过实时检测和安全代答机制，实现对大模型输入与输出全链路的风险管控，形成 “前置过滤—过程检测—结果处置” 的闭环安全防护体系。

✓ 方案框架图（示意）



核心安全能力

✓ 风险检测与智能识别

- 基于海量敏感词库与风险检测大模型双引擎，动态更新风险库，覆盖面广，识别准确率达 99% 以上。

- 拒识率低于 0.1%，在保障安全性的同时兼顾可用性。

✓ **安全代答机制**

- 对高敏感度问题提供 预置安全答案，避免大模型直接生成潜在风险内容。
- 每日支持 900 余条安全代答请求，确保用户体验与服务连续性。

✓ **高性能与可扩展性**

- 部署规模：单台服务器，3 张 GPU 显卡，支持 40 并发请求。
- 支撑杭州市政务系统的多业务场景应用，满足高并发调用需求。

✓ **客户收益与系统效果**

- 业务安全性：业务回复安全率达 99.9%，政务服务可信度显著提升。
- 防护效果：日均检测违规不良内容 1000 余条，拒答 100 余条，安全代答 900 余条，日均防护超万次。
- 综合提升：大模型回复安全性提升 30% 以上，有效降低舆情与合规风险。

本案例由 360 数字安全提供

360数字安全
数 字 安 全 的 领 导 者

北京数字世界咨询有限公司（以下简称“数世咨询”）是国内数字化领域独立第三方调研咨询机构，主营业务为网络安全产业领域的调查研究、资源对接与行业咨询。在国内网络安全产业的调查研究领域，无论是专业性还是资源丰富性，均处于业界领先地位。

调查研究方面，撰写发布《中国数字安全大事记》、《中国数字安全能力图谱》、《中国数字安全100强》、《中国数字安全产业年度报告》等业内影响力巨大的公开报告。同时，还为监督机构、国家部委、大型国企等单位提供各种定制化的内部调研报告。

资源对接方面，数世咨询目前已对接国内网络安全企业700余家，以及150余家网络安全投资业务的资本方，建立了频繁且良好的沟通合作关系，包括共同举办会议活动、投资对接，安全产品与企业推荐，企业资源整合等

行业咨询方面，经常性的为监管部门、国家部委、安全企业、安全用户、一二级市场投资机构提供建议、企业培训及专家评审等咨询服务。

公司地址：北京市东城区天鼎218文化金融园东外110号 网安小酒馆
官方网站：www.dwcon.cn
联系邮箱：dw@dwcon.cn

小酒馆

数世咨询
Digital World Consulting
数字安全领域独立第三方调研机构