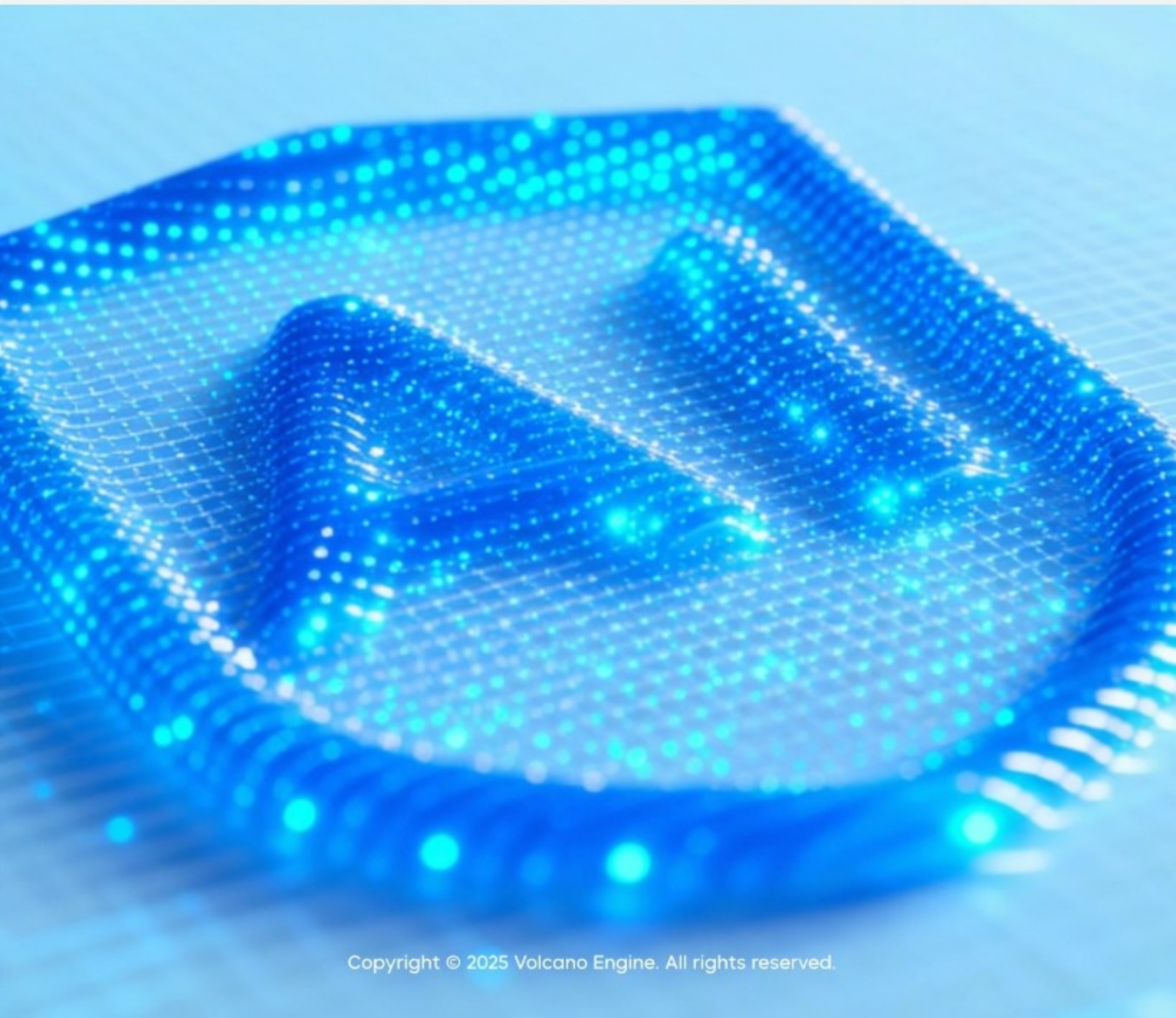
火山引擎

VOLCENGINE.COM

生成式 AI 安全白皮书

Volcano Engine

Generative AI Security White Paper



生成式AI安全白皮书

CONTENTS

Volcano Engine

Generative AI Security White Paper

01 序言

1.1 产业轨迹与拐点：从模型到业务的全面跃迁

1.2 生成式 AI 安全的核心问题与现实挑战

1.3 火山引擎的 AI 安全主张：可信、可控、合规的AI云原生基座

02 生成式AI安全风险

2.1 监管合规风险

2.2 数据隐私风险

2.3 生成式AI安全风险

03 火山引擎生成式AI服务安全保障体系

3.1 生成式AI浪潮下的安全责任

3.2 合规资质与认证

3.3 数据安全与隐私保护设计理念

3.4 生成式AI安全技术保障体系

04 总结

4.1 生成式 AI 行业安全展望

4.2 火山引擎致力于保障生成式 AI 安全

1. 序言 Introduction

1.1 产业轨迹与拐点：从模型到业务的全面跃迁

基础模型的能力边界快速拓展

从文本到图像、语音、视频的多模态表达，从“调用型”向“智能体化”工作流演进。模型不再是外置的试验工具，而是能够被嵌入到知识管理、研发协作、客服运营、风险控制等关键流程，形成可复用的“技能栈”。这种可工业化的能力，要求企业把模型服务、数据治理、权限体系、合规审计放到同一工程体系下统一管理，而不是零散的功能试点。

企业正从“单点试验”转向“平台化建设”

一方面，公有云与私有化部署需要在性能、合规、成本、可运维性之间找到动态平衡；另一方面，模型的选择从“追最新”转向“适配业务”，强调稳定性、可控性与治理可视。

1.2 生成式 AI 安全的核心问题与现实挑战

模型层

对抗、失真与滥用的攻防拉锯

在模型层，提示词注入、越狱攻击、对抗样本与模型投毒带来输出失真与能力滥用的风险。安全不再依赖简单的“黑白名单”，而是由红队评测、威胁建模、策略护栏、推理时检测与响应等机制协同构成的系统化治理方案。企业需要建立“上线即运营”的安全评测体系，形成从开发到部署的持续检测与反馈的完整链路。

数据层

从“可用”到“可信”的治理升级

训练与推理数据的污染、隐私泄露与越权访问，是生成式系统的核心风险源。数据血缘、分级分类、最小必要使用、脱敏与匿名化等能力需要与模型管理深度绑定，确保从采集、标注、训练、后训练到推理的每一步都可审计、可追溯、可复盘。

应用层

插件、工具与外部调用的安全新面貌

智能体应用的插件体系、函数调用与外部工具执行，扩大了攻击面：从凭证泄露到指令劫持，从跨租户数据穿透到供应链风险。治理重点在“意图识别与动态授权”：让每一次调用都在可见、可控的权限域内发生，并形成异常行为审计与隔离能力。

治理与合规

把“可解释、可审计、可问责”嵌入产品

生成式系统不仅是技术工程，更是治理工程。企业需要将政策、红线、行为准则固化到模型与应用的运行时：以可解释与可审计的机制支持人的监督，明确责任边界，沉淀成组织的“安全运营语言”。

1.3 火山引擎的 AI 安全主张：可信、可控、合规的AI云原生基座

火山引擎将自身定位为 **AI 云原生的可信安全基础设施提供者**，以“安全即服务”的方式，承载企业的 AI 工作负载与治理能力，建立客户信任与透明度的长期机制。

火山引擎构建**“技术领先、治理完善、生态开放”**的AI安全能力。在架构与算法层保持 AI 原生的安全创新，在合规与治理层构建全生命周期的框架与支持，在生态层以标准化接口与开放协作促进企业集成与扩展。

2. 生成式AI安全风险

生成式人工智能作为人工智能领域的重大突破，正深刻改变着技术范式。生成式人工智能凭借其强大的生成式能力，已广泛应用于游戏、汽车、智能终端、教育等多个领域，显著提升了信息处理的效率和智能化水平。随着生成式人工智能服务不断创新和应用场景的日益深入，其面临的安全风险也日益凸显。企业用好生成式人工智能服务的同时，如何处理好监管合规风险、隐私数据安全风险、以及生成式AI自身安全风险，也成为一个极具挑战性的话题。

2.1 监管合规风险

随着人工智能技术的迅猛发展，伦理、偏见、歧视等问题日益凸显。如何确保人工智能行业在符合社会价值观的框架下实现健康发展，已成为全球监管部门首要关注的问题，当前各国正加快构建针对人工智能领域的法律法规要求与合规监管框架。对于人工智能服务提供者和使用者而言，严格遵守法律监管要求至关重要。

在全球范围内，欧盟于 2024 年 8 月正式生效《人工智能法案》，作为全球首部全面针对人工智能的法案，该法案采用四级风险模型，为欧盟内人工智能系统的开发、市场投放和使用制定了统一规则，禁止违背欧盟价值观、有害的人工智能服务发展；美国推出《人工智能创新未来法案》强调了国际标准的制定、数据共享和安全性研究的重要性。

目前中国的人工智能监管体系建立在《网络安全法》、《数据安全法》、《个人信息保护法》三大法律基石之上，为人工智能领域合规管理提供了坚实的法律基础，在此基础上，各部委陆续出台**《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》**等法规要求，确立了服务提供者主体责任，明确内容合规与算法公平性等要求，并建立了人工智能安全评估和备案管理制度，为人工智能行业发展提供了明确的标准与指导。因此，在中国境内提供、使用生成式人工智能服务的企业，需要依据法律法规履行备案义务，保障用户权益、以及内容、算法安全。

此外，各国也在不断完善人工智能相关法律法规体系中。2025 年 8 月，中国国务院印发《关于深入实施“人工智能 +”行动的意见》中特别强调应**加强政策法规保障**，完善人工智能法律法规、伦理准则、推进人工智能健康发展相关立法工作、优化人工智能相关安全评估和备案管理制度。25 年 7 月，欧盟发布《通用 AI 行为准则》、《通用 AI 模型提供者指南》、《数据训练摘要模板》作为《人工智能法案》的核心配套措施，构建欧盟人工智能合规观框架。作为生成式人工智能服务的提供者与使用者需要持续关注行业法律法规建设，保障人工智能服务合法合规。

2.2 数据隐私风险

数据是每个企业的核心资产，近年来数据安全事件层出不穷，给企业带来巨大商业秘密泄露风险的同时，用户个人的隐私权也可能因此而受到威胁。生成式人工智能的发展高度依赖海量数据，但在数据大规模收集、存储、训练、推理等过程中，势必会伴随着复杂的数据与隐私安全风险。

数据收集阶段

生成式人工智能依赖海量训练数据，这些数据来源广泛，如果数据收集过程不当，可能会包含个人信息、甚至敏感个人信息，在未明确获得用户授权情况下，存在违规使用个人信息的风险。若训练数据中包含商业秘密信息，也会导致核心信息泄露的风险；此外，训练数据中也可能会包含虚假、不真实、偏见信息，这些内容会导致训练数据被污染，甚至影响输出内容的合法性、公平性、中立性等等；

数据存储阶段

用户在使用生成式人工智能服务时会涉及以下关键数据资产，如，会话数据、训练数据、以及精调后的模型等。若未采取适当的安全保障措施，数据存储阶段存在安全漏洞，最终可能导致核心数据被批量泄露；

模型训练阶段

基于生成式人工智能的技术特性，数据记忆会导致作恶分子通过特定输入触发模型的“记忆”，致使模型训练时的数据可能被提取。数据记忆是提取攻击、成员推理攻击的前提，模型对训练数据的记忆越深刻，攻击者就越容易通过设计输入信息以“唤醒”这些记忆，进而实施更精准的数据窃取；

模型推理阶段

用户在使用生成式人工智能服务时的输入与输出环节，可能因为数据传输、存储、API接口存在的安全漏洞，导致数据被第三方非法获取，从而造成数据泄露问题。

除了上述问题外，内部人员违规操作或者人为疏漏也是常见的数据与隐私安全风险的诱因。

2.3 生成式AI安全风险

生成式 AI 正在快速嵌入企业生产力、开发运维与对外服务。安全风险不再停留在传统应用层，而是沿着“AI 基础设施 → 大模型 → 智能体”链条相互作用、彼此放大。

AI 基础设施安全风险

算力滥用：当 GPU/TPU 与训练集群缺乏精细的配额与准入控制，未授权调用会造成经济损失，甚至被用于非法挖矿或异常训练。

网络隔离薄弱：资源直连公网、入/出站流量缺乏分级管控，导致暴露面扩大，横向移动更容易。

供应链漏洞：开源框架、驱动与容器镜像成为常见入口，版本污染或镜像被植入会在训练/推理链路中纵深扩散。

访问控制缺陷：IAM 策略误配、长效 AK 凭证泄露，使攻击者轻易绕过控制面直达算力与数据。

模型与平台安全风险

模型泄露：参数提取、逆向推断或错误发布导致权重外泄，直接损害资产价值。

数据隐私泄漏：模型在推理中“记忆”敏感信息，一旦遭遇 Prompt 注入，可能被诱导输出个人或企业机密。

对抗攻击：恶意输入触发异常行为，造成错误回答、策略绕过或安全审计失效。

后门与中毒：训练或微调阶段的污染样本，使模型在特定触发词下被操控，风险在生产环境中隐蔽显现。

内容安全风险：模型在用户输入引导下，生成违反法律法规、公序良俗或存在安全隐患的内容。

传统 Web 安全风险：传统Web漏洞，认证鉴权的缺失，访问控制不当，会造成模型平台的失陷，造成模型和用户数据泄漏

AI智能体安全风险

Prompt 注入：精心构造的指令让模型执行非预期任务，典型表现为越权调用 API 或读取敏感数据，泄漏系统提示词。

工具滥用：具备代码执行、数据库访问与外部系统调用能力的 Agent，若缺少最小权限与隔离，将造成严重泄露与破坏。

供应链安全风险：接入的第三方插件与 API 成为新攻击面，依赖的生态漏洞被复用扩散。

隔离机制失效：多租户场景中，未对网络和数据进行隔离，导致租户间的资源、数据或操作边界被打破。

传统 Web 安全风险：传统 Web 漏洞，认证鉴权的缺失，访问控制不当，会造成智能体失陷，造成用户数据泄漏。

3. 火山引擎生成式AI服务安全保障体系

3.1 生成式AI浪潮下的安全责任

随着生成式人工智能（Generative AI）技术的广泛应用，火山引擎致力于为人工智能服务使用者提供安全、合规的人工智能服务。然而，如同云服务责任共担体系一样，在人工智能平台上部署的AI工作负载，其安全、稳定运行需要使用者与服务提供者共同关注并维护。当然，根据您所选择服务类型的不同，您所需承担的安全责任也存在相应差异。例如，基于机器学习平台（AML）构建AI工作负载，您需要关注模型训练、模型部署等全生命周期工作流的安全合规责任；如您选择豆包大模型搭建生成式人工智能服务，模型的安全合规则由火山引擎与您共同承担。以下将从合规、安全、数据隐私三个方面分别阐述生成式人工智能场景下的责任体系。

合规责任：恪守法规、共筑健康生态

人工智能行业健康发展首先需要人工智能服务提供者、使用者严格遵守法律规范，恪守合规底线。合规方面首要关注备案合规与内容合规（见：图1）：

备案合规

火山引擎为客户提供的豆包大模型已完成算法备案和生成式人工智能服务备案。当您基于火山引擎豆包大模型对公众提供具有舆论属性或者社会动员能力的生成式人工智能服务，则建议开展算法备案，并按照属地网信部门要求进行生成式人工智能服务备案或登记。

内容安全合规

火山引擎根据《生成式人工智能服务管理暂行办法》、《网络信息内容生态治理规定》等法律法规要求，针对模型全生命周期建设了内容安全策略，对豆包大模型生成内容进行严格管控。当您搭建生成式人工智能服务时，尤其是基于三方模型、对预训练模型进行精调或针对特殊场景（例如未成年人）的情况下，需要根据自身业务需求额外建设内容审核能力；如果您在机器学习平台、GPU 云服务器搭建的 AI 工作负载，则需要由您满足相应监管要求。

内容标识合规

依据《人工智能生成合成内容标识办法》要求，若您向公众提供人工智能生成合成内容服务，需满足显式标识、元数据隐式标识等内容标识要求，在火山方舟场景下可以为您提供相关能力以满足监管要求。

	GPU云服务器	机器学习平台	火山方舟	豆包大模型
内容标识合规	客户若向公众提供人工智能生成合成内容服务，需满足显式标识、元数据隐式标识等内容标识要求；	客户若向公众提供人工智能生成合成内容服务，需满足显式标识、元数据隐式标识等内容标识要求；	客户若向公众提供人工智能生成合成内容服务，需满足显式标识、元数据隐式标识等内容标识要求；	客户若向公众提供人工智能生成合成内容服务，需满足显式标识、元数据隐式标识等内容标识要求；
内容安全合规	客户若向公众提供人工智能生成式人工智能服务，需采取有效措施防止产生违法违规内容；	客户若向公众提供人工智能生成式人工智能服务，需采取有效措施防止产生违法违规内容；	若您对公众提供生成式人工智能服务，尤其基于三方模型、对预训练模型进行精调或针对特殊场景，需要根据自身业务需求额外建设内容审核能力； 火山引擎针对模型全生命周期建设了内容安全策略，对豆包大模型生成内容进行严格管控；	若您对公众提供生成式人工智能服务，尤其基于三方模型、对预训练模型进行精调或针对特殊场景，需要根据自身业务需求额外建设内容审核能力； 火山引擎针对模型全生命周期建设了内容安全策略，对豆包大模型生成内容进行严格管控；
备案合规	客户需要关注选择的大模型是否具备算法备案和生成式人工智能服务备案； 若您对公众提供生成式人工智能服务，则建议以服务提供者的角色开展算法备案，并按照属地网信部门要求进行生成式人工智能服务备案。	客户需要关注选择的大模型是否具备算法备案和生成式人工智能服务备案； 若您对公众提供生成式人工智能服务，则建议以服务提供者的角色开展算法备案，并按照属地网信部门要求进行生成式人工智能服务备案。	若您对公众提供生成式人工智能服务，则建议以服务提供者的角色开展算法备案，并按照属地网信部门要求进行生成式人工智能服务备案。 火山引擎提供的豆包大模型服务均已完成算法备案和生成式人工智能服务备案；	若您对公众提供生成式人工智能服务，则建议以服务提供者的角色开展算法备案，并按照属地网信部门要求进行生成式人工智能服务备案。 火山引擎提供的豆包大模型服务均已完成算法备案和生成式人工智能服务备案；

（图1） 客户责任 火山引擎责任

隐私责任：尊重隐私，共建可信AI

保护隐私安全是火山引擎与客户的共同责任。根据构建AI工作负载的方式不同，对训练和推理数据的掌控程度会存在相应差异，所需承担的安全责任也有所不同（见：图2）。

训练数据合规

如果您直接使用豆包大模型构建生成式人工智能服务，豆包大模型基于国家标准《GB/T 45652 生成式人工智能预训练和优化训练数据安全规范》对训练数据的来源进行审核，涉及个人身份或敏感属性的内容，默认脱敏与去标识化，保障训练数据安全合规；若您基于GPU云服务器、机器学习平台训练自有模型或使用火山方舟对模型进行精调，您需要保障训练数据集的安全性、合规性，避免训练数据中包含风险数据；

客户数据安全

火山引擎通过安全互信计算架构，提供应用层会话加密、网络层传输加密、多维度环境隔离以及多类别审计日志等安全能力，为客户提供多维度、全方位的数据安全增强保护。若您基于GPU云服务器、机器学习平台构建AI工作负载，您需要对数据的来源及内容负责，火山引擎则保障产品安全性，且确保不会访问您的数据，确保数据唯您可见，唯您所用，唯您所有。

	GPU云服务器	机器学习平台	火山方舟	豆包大模型
训练数据合规	客户需对数据的来源及内容负责	客户需对数据的来源及内容负责	精调模型：如果您基于方舟平台上基础模型进行精调，需要保障训练数据安全合规 预训练模型：豆包大模型基于国家标准 GB/T45652 对训练数据的来源进行审核，保障安全合规	豆包大模型基于国家标准 GB/T45652 对训练数据的来源进行审核，保障安全合规
客户数据安全	客户需要配置安全能力保护数据生命周期安全 火山引擎保障产品安全性，确保未经客户授权不会访问客户数据	客户需要配置安全能力保护数据生命周期安全 火山引擎保障产品安全性，确保未经客户授权不会访问客户数据	客户可以按需配置访问权限，定期开展审计 火山引擎通过加密和安全沙箱等技术，为客户提供安全可信、会话无痕的执行环境	客户可以按需配置访问权限，定期开展审计 火山引擎通过加密和安全沙箱等技术，为客户提供安全可信、会话无痕的执行环境

客户责任 火山引擎责任

（图 2）

安全责任：多层防护、共担安全使命

生成式人工智能安全需由服务提供者与开发者共同维护，安全责任与服务类型、模型构建方式密切相关，以下将从基础设施安全、模型安全两个方面介绍安全责任划分（见：图3）：

基础设施安全

当您选择在火山引擎平台上构建您的 AI 工作负载，无论是直接调用豆包大模型 API 还是基于机器学习平台构建自训练模型，火山引擎将通过平台强大的系统承载力，全力保障大模型落地安全性、可靠性，为您提供安全可信的执行环境；

模型安全

当您使用火山方舟与豆包大模型构建生成式人工智能服务，火山方舟将通过稳定可靠的安全互信架构，提供链路全加密、数据高保密、环境强隔离、操作可审计的安全能力，保障服务全生命周期安全；如果您基于 GPU 云服务器、机器学习平台方式构建 AI 工作负载，则需要关注全生命周期安全，包括模型选型、训练部署、上线运行等环节的安全性；

	GPU云服务器	机器学习平台	火山方舟	豆包大模型
模型安全	要关注全生命周期安全，包括模型选型、训练部署、上线运行等环节的安全性	要关注全生命周期安全，包括模型选型、训练部署、上线运行等环节的安全性	火山方舟将通过稳定可靠的安全互信架构，保障服务全生命周期安全	火山引擎定期对模型进行巡检，验证端到端模型的安全性，保障模型安全
基础设施安全	火山引擎全力保障 AI 基础设施稳定可靠、安全可信	火山引擎全力保障 AI 基础设施稳定可靠、安全可信	火山引擎全力保障 AI 基础设施稳定可靠、安全可信	火山引擎全力保障 AI 基础设施稳定可靠、安全可信

客户责任 火山引擎责任

（图 3）

3.2 合规资质与认证

2021年以来，我国陆续发布《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》等大模型服务相关的法律法规，形成一套完备的生成式人工智能服务监管体系。火山引擎为保障平台安全合规，为客户提供服务的大模型均以服务技术支持者的角色完成算法备案与生成式人工智能服务备案，并且针对大模型平台单独开展网络安全等级保护测评，以证大模型平台在网络安全技术能力、安全管理体系等方面充分满足国家安全合规要求，为用户数据安全和业务稳定运行提供了坚实保障。同时，火山引擎致力于贡献安全实践促进行业安全生态建设，在人工智能、数据安全、个人信息保护等领域积极参与国家标准、行业标准的制定，参与包括全国网安标委、工信部人工智能标委会、全国通信标准化委员会等多个权威标准化组织，贡献GB/T 45958 人工智能计算平台安全框架、GB/T 35274 大数据服务安全能力要求等多项国标、行标，为行业标准化建设贡献力量。

在满足法律法规要求的基础上，火山引擎为了向客户提供更高质量的大模型服务，也积极参与行业权威认证，通过国际、国内独立第三方专业机构验证大模型相关产品安全合规能力。2025 年 2 月，火山引擎正式通过国际权威认证机构 DNV 的严格审核，成为全球首批获得欧盟授信机构 RVA 认可 ISO/IEC42001 人工智能管理体系认证的企业，标志着火山引擎在 AI 治理领域的技术实力与合规能力达到国际最高标准。截至目前，火山引擎云平台以及大模型服务已经通过 ISO/IEC 27001 信息安全管理体系、ISO/IEC 20000 信息技术服务管理体系、ISO22301 业务连续性管理体系、ISO/IEC 27701 隐私安全管理体系和 ISO9001 质量管理体系等多个管理体系认证，旨在向客户提供最高质量的合规性保障体系。

ISO42001人工智能管理体系认证	ISO9001质量管理体系认证
ISO27001信息安全管理体系认证	ISO27017云服务信息安全管理体系认证
ISO27018公有云个人身份信息保护管理体系认证	ISO27701隐私信息管理体系认证
ISO22301业务连续性管理体系认证	ISO27040数据存储安全管理体系认证
ISO29151个人身份信息保护管理体系认证	ISO14001环境管理体系认证
ISO45001职业健康安全管理体系认证	BS10012个人信息管理体系认证
CSA STAR云安全管理体系认证	SOC1、2、3审计
网络安全等级保护三级-大模型平台	信通院可信AI认证
信通院可信云认证	中国电子技术标准化研究院-大模型国标符合性测试
中国软件测评中心-大模型安全服务能力认证	第三方专业机构安全评估报告

3.3 数据安全与隐私保护设计理念

设计理念

和传统AI数据安全和隐私保护相比，大模型或者生成式AI的数据与隐私安全的关键挑战在于：

云上大模型的数据安全问题

相较于传统AI模式，很多用户为了确保数据安全将模型部署在私有化环境里；但考虑到大模型迭代速度，如果用户想使用市场上最新、最强的模型能力，往往会选择云上的大模型服务，从而带来数据上云的安全挑战；

模型记忆和数据提取攻击问题

相较于传统 AI 因参数规模较小、记忆能力相对较弱的特点，大模型因参数规模庞大，对训练数据的记忆能力显著增强，会带来数据保密性的挑战；

黑盒模型的可解释性问题

大模型的黑盒特性使其决策过程难以追溯，人员的恶意行为也会增大此类风险，从而带来如何确保大模型操作透明化的挑战。

在这样的挑战下，我们认为在生成式AI时代，需要打造一套全周期的安全可信方案，全方位保障客户数据和隐私安全，实现会话无痕，保障数据唯客户所见、唯客户所用、唯客户所有。

技术保障体系

围绕生成式人工智能服务全流程中的数据和隐私安全风险，火山引擎方舟可信人工智能系统（Ark TrustAI System，以下简称“方舟”）提出生成式人工智能安全互信计算框架，其旨在结合隐私增强、可信计算等安全计算技术，实现云上模型推理和训练过程中数据和模型的数据安全和隐私保护能力。

相关保护方案具有以下技术特点：

01 链路全加密： 在传输安全方面，在用户到方舟安全计算环境之间建立端到端加密通信通道，防止用户数据在传输链路中被截获。用户请求基于火山引擎标准API网关接入之后，在所有内网通信均采用mTLS（Mutual Transport Layer Security）传输，保证全链路数据安全；

02 数据高保密： 在加密存储方面，方舟采用多层级数据加密方案，训练集与精调数据集只有在安全沙箱内存中被解密。同时支持用户使用自有密钥（Hold Your Own Key，HYOK）进一步提升数据保护水平，实现对用户数据的机密性保护，保证用户数据非本人不可见。通过EFS(Encrypted File System)透明加密本地文件系统技术，实现方舟安全沙箱数据落盘即加密。EFS使用火山引擎标准的秘钥管理服务(KMS)进行秘钥托管，保证秘钥安全性。在加解密速度方面，EFS提供了显卡

加速方案，可实现沙箱内部解密速度超100GBps，解密带来的延迟对于推理任务启动几乎无影响。沙箱内EFS加密之后的密文数据通过服务组件打通多种标准云存储服务，保证密文数据存储安全性；

03 环境强隔离： 火山引擎通过领先的容器沙箱、网络隔离、可信硬件等多维度强制隔离技术，杜绝外部风险入侵和内部数据泄露。

- 容器沙箱方面，云原生容器沙箱技术采用开源vArmor，通过Linux AppArmor LSM（Linux Security Modules），BPF LSM和Seccomp技术实现强制访问控制器，从而对模型运行时环境进行安全加固。它可以用于增强容器隔离性、减少内核攻击面、增加容器逃逸可成横行移动攻击的难度与成本。
- 网络隔离方面，同时开启VPC（Virtual Private Cloud）网络以及RDMA（Remote Direct Memory Access）网络隔离，保证单个任务内运行环境之间可通信、跨任务严格隔离。在VPC网络中，主要基于Kubernetes的NetworkPolicy对任务的主网卡进行隔离，防止不同任务之间互相通信；另外精调或者推理任务还会使用辅助网卡进行RDMA相关的通信，对于这部分采用自研的技术对不同的任务进行分组隔离，保证不同任务之间RDMA通信被阻断。

04 操作可审计

- 在访问控制方面，对于运行期间需要访问的外部服务，方舟会进行严格的审查，制定对应的访问控制策略，并且通过服务组件进行访问代理和策略实施。对于精调训练获得的更新的模型，借助加密存储将更新的模型保存到训练平台对象存储，保证精调模型机密性、模型提供方和训练数据提供方产权共享的同时，也提升了精调模型存储的安全性；
- 在可信运维方面，方舟基于互信计算框架，对于进出安全执行环境的出入流量、数据读写均有严格的管控。同时，基于火山引擎标准的堡垒机产品，在经过审批授权之后，允许进入安全沙箱进行运维，并严格限制了安全沙箱内的危险操作，对全程进行录屏操作，以便对用于后续追溯，降低内部作恶的风险。

在标准通用的安全保护方案基础上，对于进阶安全需求方舟安全计算环境提供基于硬件可信技术的机密部署模式，包括：

芯片级隔离	高可信推理	安全可信可证明
构建从物理芯片（GPU/CPU）到容器的全链路的安全隔离方案，杜绝方舟的云服务商或者生成式人工智能供应商接触数据，让隐私数据在云端获得比本地更安全的计算保障；	基于 TKS（Trusted Key Service）密钥提供会话加密，并提供可信验证报告，证明密钥提供服务可信。数据在安全沙箱中解密，解密后的数据由硬件提供机密性和完整性保护，进行密态运算；	提供远程证明服务，随时为客户验证推理点部署环境的安全性，从三方身份和技术上解决了生成式人工智能供应商、云服务厂商的信任问题。

总而言之，方舟强调云端AI计算的动态安全加固和透明可信体验感的增强，致力于通过安全、合规、可信的保护方案，实现用户会话零保留、平台违规零容忍，保障数据与隐私安全。

3.4 生成式AI安全技术保障体系

火山引擎基于AI业务，构建了一套以透明可信为核心的“三层级”生成式AI安全保障体系。“三层级”涵盖了AI基础设施安全、模型与平台安全、AI智能体安全。AI 安全研究则是不断发掘新兴安全风险探索防护方案，并为“三层级”安全能力提供技术支持和方向指引。

（图4）

3.4.1 AI 基础设施安全

AI 基础设施是云平台的底座，承载 IaaS/PaaS 主干与对外服务。要实现“可用、可信、可控”的算力与数据底座，需将平台基础安全和增强安全方案组合成体系化的安全能力。

平台基础安全

01 治理架构与规范体系

“安全风险治理飞轮”将安全文化、组织、流程、工具与最佳实践融入业务，形成贯穿产品全生命周期的治理闭环。以“安全分”度量安全水位，通过持续投入与能力补齐提升分数，已纳入考核与精细化运营。

内外合规	默认安全	客户信任
内部威胁强管控，外部监管不违规 上云即合规	高危严重风险不上线，数据不丢 上云即安全	对外透明，对内可信 上云即可信

02 产品安全保障

发阶段（上线前）	运行阶段（运营防护）
代码 diff 风险评测与巡检	CloudTrail 与事件驱动检测
FaaS 函数自动化扫描	AI+SOC 联动多 agent、MCP 工具
人工聚焦复杂场景，大模型兜底简单场景	告警研判与入侵调查智能化

03 平台基础防护

原生DDoS 防护：在出口部署攻击检测与清洗系统，过滤流量型与应用层攻击，正常流量无损回源；结合运营商黑洞等能力，在大流量场景快速封禁，确保业务持续稳定。

WAF 应用层防护：结合特征规则、AI 智能检测与行为分析，防御常见 Web 攻击（SQL 注入、XSS、文件上传漏洞等），并支持 CC 攻击缓解、Bot 管理与精准访问控制，提供实时监测与可视化报表。

04 威胁情报与供应链

漏洞情报：漏洞情报平台对数据进行清洗，专家复核并评估危害与影响面；联合安全扫描治理漏洞，联动安全防护能力收敛风险。

灰黑产情报：围绕火山引擎业务场景，构建“场景化采集 + 多维度关联”的采集架构与“规则匹配 + 行为建模 + 团伙关联”的分析体系，并以“分级预警 + 场景化响应”联动云产品能力，实现主动防御与风险收敛。

05 攻防演练与外部验证

红蓝演练：覆盖外部渗透与内鬼模拟，针对薄弱点开展治理与加固。

漏洞奖励计划：与全球白帽社区共建安全生态，持续专项测试、演练与验证，提升产品安全水位与可信度。

增强安全方案

固件资产管理与漏洞响应

以资产为锚点、以情报为触发，形成“精准识别—批量升级—对客提示”的闭环。

- 构建覆盖CPU/GPU/BMC/网卡的固件资产台账，并与租户账号关联，支持批量固件升级与对客风险提示。
- 结合供应链漏洞情报，对高危漏洞实施补丁升级与版本卡点控制；如 BMC 高危漏洞 CVE-2023-34335，通过对新增固件版本卡点与存量机器在线升级，实现全量修复，保障客户与平台侧不受影响。

硬件可信根

以硬件为信任锚，确保整机平台完整性与设备身份可信，贯穿启动、升级与运行时。

- **整机平台完整性保护：**启动阶段验证 BMC/BIOS 固件签名；升级后在 BMC/BIOS 重启流程中复核完整性；运行时监听受保护区域访问，防篡改。
- **设备身份与生命周期可信管理：**基于硬件因子派生设备私钥，产线收集公钥作为设备身份凭证，以设备身份为基础构建双向鉴权信道，并按生命周期轮转管理。

- **动态度量可信：**设备启动或更新时记录关键信息完成可信度量并上报，结合远程证明能力，确保启动和升级过程中的机密性与完整性。

可信执行环境

以硬件级隔离为核心，为敏感代码与数据提供运行时的保密性与完整性。结合远程证明与密钥管理，可在云侧与虚拟化环境中建立可验证的信任链。

- **Intel TDX 能力与适用：**通过在硬件层部署信任域（TD），保护敏感数据与应用免受未经授权访问，并确保**完整性、保密性与真实性**。其软件模块在新的 CPU 安全仲裁模式（SEAM）中启动，配合现有虚拟化基础设施完成 TD 的进入/退出。
  - 隔离信任域（TD）：以硬件为根的隔离执行，限定访问边界。
  - 内存加密与状态保护：使用安全扩展页表（EPT）、GPA 共享位、物理地址元数据表与 Intel TME-MK 多密钥内存加密，保护 TD 的 CPU 状态与内存不被非 SEAM 模式窥探。
  - 远程证明：基于经 CPU 度量的 TDX Module 启动，支持对 TD 可信状态进行远程认证。
  - 虚拟化友好：与既有虚拟化基础设施协同，降低工程集成成本。
- **Intel SGX 能力与适用：**Intel SGX 通过在进程内创建**安全区（Enclave）**，为敏感代码与数据提供极细粒度的隔离。除安全区与 CPU 外，操作系统、虚拟化管理程序、SMM、BIOS 等特权软件**不在信任边界**，即便底层平台受恶意软件影响，安全区内信息仍保持机密。
  - 进程级安全区：面向关键模块的高强度隔离与最小信任面。
  - 强韧的机密计算环境：以硬件防护抵御特权层面窥探与干扰。
  - 适用场景：机密计算、数据加密应用、数据共享与计算、对高安全可信有明确要求的任务。
  - 云侧形态：火山引擎 embg2t 弹性裸金属（第三代至强），单实例最高含 256G 加密内存。
- **AI机密计算：**基于机密计算、密码学应用、信息流安全等隐私保护技术，面向公有云环境，实现敏感数据流转与应用安全的通用能力。
  - 数据在用保护：在推理/训练过程中保护机密数据与模型资产。
  - 端云协作安全：端侧受限场景下，以云侧算力承载并保持透明可信。
  - 隐私保护技术栈：机密计算 + 密码学应用 + 信息流安全的组合式能力。

（图 5）

3.4.2 AI 模型与平台安全

火山方舟是火山引擎推出的大模型服务平台（MaaS，Model-as-a-Service），面向企业与个人开发者提供模型精调、推理、评测等全方位功能与服务，以及丰富的插件生态和 AI 原生应用开发服务。

根据国际数据公司（IDC）发布的《中国大模型公有云服务市场分析，2025H1》报告，2025 年上半年，中国公有云上大模型调用量达 536.7 万亿 Tokens（统计口径：各大云厂商对外部客户提供的大模型公有云服务调用量，不包含自有业务调用），火山引擎以 49.2% 的份额占比位居中国第一。火山方舟通过安全可信的基础设施、专业的算法技术服务，全方位保障企业级 AI 应用落地。

（图 6）

模型安全

模型安全是一条贯穿数据进入、模型训练到发布服务的治理生产线。其目标是确保数据可用与可解释、训练过程稳健与可追溯、上线前后可审核与可回滚，从而在满足合规要求的同时保障产品与工程的持续稳定运行。

01 安全原则

- **数据来源与分级：**严格落实GB/T45652对训练数据的来源审核要求，来源与分级直接影响后续清洗强度与访问权限边界。
- **最小必要与隐私保护：**围绕训练目标控制数据范围：不采“不必要”，不存“或然需要”。涉及个人身份或敏感属性的内容，默认脱敏与去标识化，兼顾质量收益与合规成本的平衡。
- **权限与环境隔离：**清洗、训练、评测分区运行；跨区访问需审批并记录审计。高风险数据集实施更严格的访问策略与更短的会话时长。
- **审计与可追溯：**从采集到上线在关键环节记录日志并保存：谁、何时、对什么、做了什么、为什么。清洗与过滤规则需版本化管理，保障训练再现性与变更透明度。
- **流程门禁与报备：**将“可否采、可否用、可否上”嵌入流程门禁：需求接入与法务评估、训练前的数据治理评审、上线前的安全门禁与数据安全报备，实现事前阻断与事后复核。

02 模型生命周期安全

从“数据标注 → 预训练 → 后训练与上线”，各阶段侧重点不同，但遵循统一治理原则与证据化要求。

数据标注	预训练	后训练与上线
• 以**流程与证据**为主：需求接入→立项评估→资源准入→作业→质检→交付→结项。 • 明确**法务评估**与外包合规，控制信息外泄与成本暴露。 • 标注**平台环境隔离、权限分层**；质检全覆盖与日志留存，问题打回并复核改进。	• **来源筛选**遵循 GB/T45652 对训练数据的相关安全要求 • **风险过滤与质量评估**：周级规则更新；风险分级 P0/P1/P2；质量 0/1/2 打分；中英双语过滤。 • **环境与留痕**：边界防护、堡垒机、加密传输与存储、日志保留 6 个月；清洗与采样可追溯。	• **安全对齐**：红线 / 高危 / 灰色分层策略；结合权威口径与价值观引导。 • **评测与红队**：固定节奏攻防演练与问题回灌；标准按周迭代。 • **门禁与报备**：发布前安全回扫、上线门禁、数据安全报备与灰度监控；异常可快速回滚。

数据标注阶段

数据标注阶段将“可执行的流程”与“可证明的合规”结合。需求接入后完成合规与可行性评估，明确是否可外发及合同约束；随后进行资源与人员准入，确保对象、流程与工具在受控环境运行。标注数据进入质检队列，未达标打回并记录，形成“问题定位—修复—复核”的完整流程。

- 流程与门禁：法务评估明确可外发与披露范围；招采与商签单管理供应商与合同；人员准入与实施方案评审确保“人员与方法”合规。
- 隐私与最小必要：标注范围遵循最小必要，模板与操作默认脱敏。
- 平台与访问：标注平台提供隔离环境与权限分层；作业与风控日志在工具内留存，可导出追溯。
- 质检与留痕：质检团队二次把关；关键节点留痕（模板、队列、数据集变更），交付可验收、问题可定位。

预训练阶段

预训练数据治理将“来源合规、风险过滤、质量提升”组织成一条可验证的流水线。数据来源遵循GB/T45652对训练数据的相关安全要求；英文与垂直领域资源采用专用风险过滤模型与敏感词机制，确保不含违禁与低质内容。清洗规则按周更新，覆盖涉政、色情、违法违规等重点方向，同时对劣质样本进行压制。每次训练批次进行抽检与定向攻击评测并留存报告，形成稳定的安全保障链路。

- 来源与合规：数据来源严格遵循GB/T45652对训练数据的安全要求。
- 风险过滤与质量评估：风险分级 P0/P1/P2 与质量 0/1/2 打分；中英双语过滤模型与敏感词表叠加；低质模型过滤后再投训。
- 访问与存储安全：数据访问控制严格管控，操作日志保留 6 个月。
- 版本化与追踪：清洗规则与采样过程版本化管理；抽检报告与评估记录可回溯，支持训练再现性。

后训练与上线

后训练关注输出安全与行为一致性。将问题分为红线、高危与灰色，分别采用兜底话术、权威正确口径与价值观降险话术，形成策略兜底与安全对齐。发布前执行安全回扫、上线门禁与数据安全报备；采用灰度与分层监控，在突发或重大敏感节点升级审核与专班应对；异常场景按预案回滚。

- 安全对齐与策略：红线命中兜底，高危输出权威正确口径，灰色在正确口径基础上追加降险与价值观导向。
- 评测与红队：依据国家标准《网络安全技术 生成式人工智能服务安全基本要求》，定制了多套题库验证模型的安全性，蓝军攻击团队与审核标注团队协同，问题即时回灌训练与策略。
- 报备与门禁：上线前完成数据安全回扫与数据安全报备；灰度发布与多维监控；突发事件节点临时升级策略与队列。

*特别说明：方舟平台仅对三方模型进行基础安全测试，但不对其安全性进行承诺。如用户选择选择和使用开源模型，需对模型生成内容的安全和侵权风险自行评估确认。
详见 火山方舟大模型服务平台专用条款 3.7.3。

平台安全 安全互信计算架构

火山方舟通过稳定可靠的安全互信方案，保障模型服务提供方的模型安全与模型使用者的信息安全。方舟安全互信计算架构结合云原生安全沙箱、加密存储、网络隔离以及加密传输等技术，针对大模型数据预处理、推理、精调以及评测等场景提供了多维度全方位的数据安全增强。

安全设计	安全能力	实现说明
链路全加密 在用户到方舟安全计算环境之间建立端到端加密通信通道，防止用户数据在传输链路中被截获 	应用层 会话加密	• 用户自定义开启，一行代码即可免费使用 • 基于用户和安全沙箱的双向身份认证，进行再次密钥协商，建立用户和安全沙箱之间的互信连接，保证会话数据只能在正确身份的安全沙箱内被解密
	网络层 传输加密	• 全局默认开启，无需用户操作 • 外网传输使用HTTPS，内网传输使用mTLS，内网跨VPC通信信使用Private Link
数据高保密 实现对用户数据的机密性保护，保证用户数据非本人不可见 	多层级 数据加密	**精调场景支持客户使用自有密钥（HYOK），获得云环境中数据安全的“最高级别”控制权** • 训练集与精调模型落盘即加密，只能在安全沙箱内存中被解密 • 基于FUSE的透明加密文件系统，保证数据和模型从安全沙箱内存直写至分布存储时无缝加密 • 密钥由用户通过密钥管理系统（KMS）管理，并在安全沙箱内进行安全访问 • 支持GPU加、解密，保证数据访问效率
环境强隔离 通过领先的可信容器沙箱、网络隔离等多维度强制隔离技术，杜绝外部风险入侵和内部数据泄露 	多维度 环境隔离	**任务级动态安全计算环境，机密计算保护运行中数据** • **安全沙箱**：为动态运行时实施任务级隔离策略。采用vArmor增强容器安全，确保程序以非特权模式启动，动态阻断高危的系统调用。保证攻击者无法利用当前任务漏洞横移到其他任务，影响其他任务的安全 • **芯片级隔离**：分离式部署构建从物理芯片（GPU/CPU）到容器的全链路的安全隔离方案，杜绝方舟的云服务商或者大模型供应商接触数据，让隐私数据在云端获得比本地更安全的计算保障 • **网络隔离**：针对VPC和RDMA网络采用任务级网络隔离，支持配置动态隔离策略。保证即使是处于同一网络(如VPC)中的不同任务，也不能直接通信 • **可信数据访问代理**：针对沙箱外部服务(如KMS)，只能通过方舟代理访问，叠加严格的权限检查和访问请求检查，防止沙箱内的进程将数据发送到外部
操作可审计 影响用户数据资产的所有操作均有日志记录，验证安全策略生效，识别潜在风险 	多类别 审计日志	**通过OpenAPI以及方舟控制台，获取指定推理接入点和精调任务的审计日志，以监控安全策略生效和潜在风险** • **沙箱登录日志**：记录通过远程或本地方式登录沙箱的行为 • **沙箱连接日志**：记录沙箱是否通过对外网络连接发送数据 • **沙箱容器逃逸日志**：记录是否有超越沙箱权限的执行命令 • **vArmor拦截日志**：记录被vArmor拦截执行的用户非法命令 • **KMS访问日志**：记录系统对KMS的访问,验证是否执行加解密 （更多审计日志请查看控制台）

（图 7）

◉ **推理会话数据零留存：**“数据零留存”是火山方舟的一项重要安全承诺和数据管理策略，指在训练、推理、评估等任务完成后，平台将从内存和持久性存储中擦除相关模型、样本及临时文件，确保非授权不留存任何用户数据。

- API 网关（用户通过API调用方舟大模型服务）、安全节点均不存储用户任何推理会话内容；
- 用户调用第三方插件，平台侧不存留会话内容；
- 方舟对用户授权收集/利用的数据先脱敏（假名化）再加密存储，且支持还原（重标识）；
- 安全沙箱通过隔离手段绑定任务资源，严格审计传出信息，任务结束后沙箱销毁，数据不可恢复；
- 用户可选择平台原生提供的机密推理服务，通过把推理任务运行在资源独享、高度隔离、严格保护的“云端安全屋”中，进一步提升推理数据安全保护等级。

*方舟严格执行用户隐私管理，非获得用户授权或法规要求不留存用户数据，详见 火山方舟大模型服务平台专用条款 与 火山方舟平台免责声明和体验服务规则说明。

◉ **自持密钥用户完全自主可控：**密钥是各项加解密工作的核心，很大程度上影响着数据安全信任感。火山方舟在行业内率先支持 MaaS 原生的 HYOK（Hold Your Own Key，自持密钥）能力，数据集与精调模型的传输、存储、调用过程全部支持使用用户自持密钥。使用 HYOK 后：

- **密钥可信：**密钥在用户信任环境下生成并保存，并由用户自行管理密钥的生命周期。
- **访问可控：**方舟仅在用户明确授权下执行短暂且受控的加解密操作，用户可随时撤销方舟对密钥的访问权限。
- **审计可验：**密钥绝不会以明文形式传输或存储在方舟，仅在必要计算时访问，用户可通过第三方密钥日志交叉验证。

◉ **机密推理塑造“硬件级”安全信任根基：**将机密计算技术原生内置于 MaaS 平台，火山方舟在行业内率先推出 MaaS 原生的机密推理服务。它不仅能进一步解决运行时数据的安全保护难题，还向用户提供可自行验证的远程证明文件，带来“透明可验证”的信任。使用机密部署后：

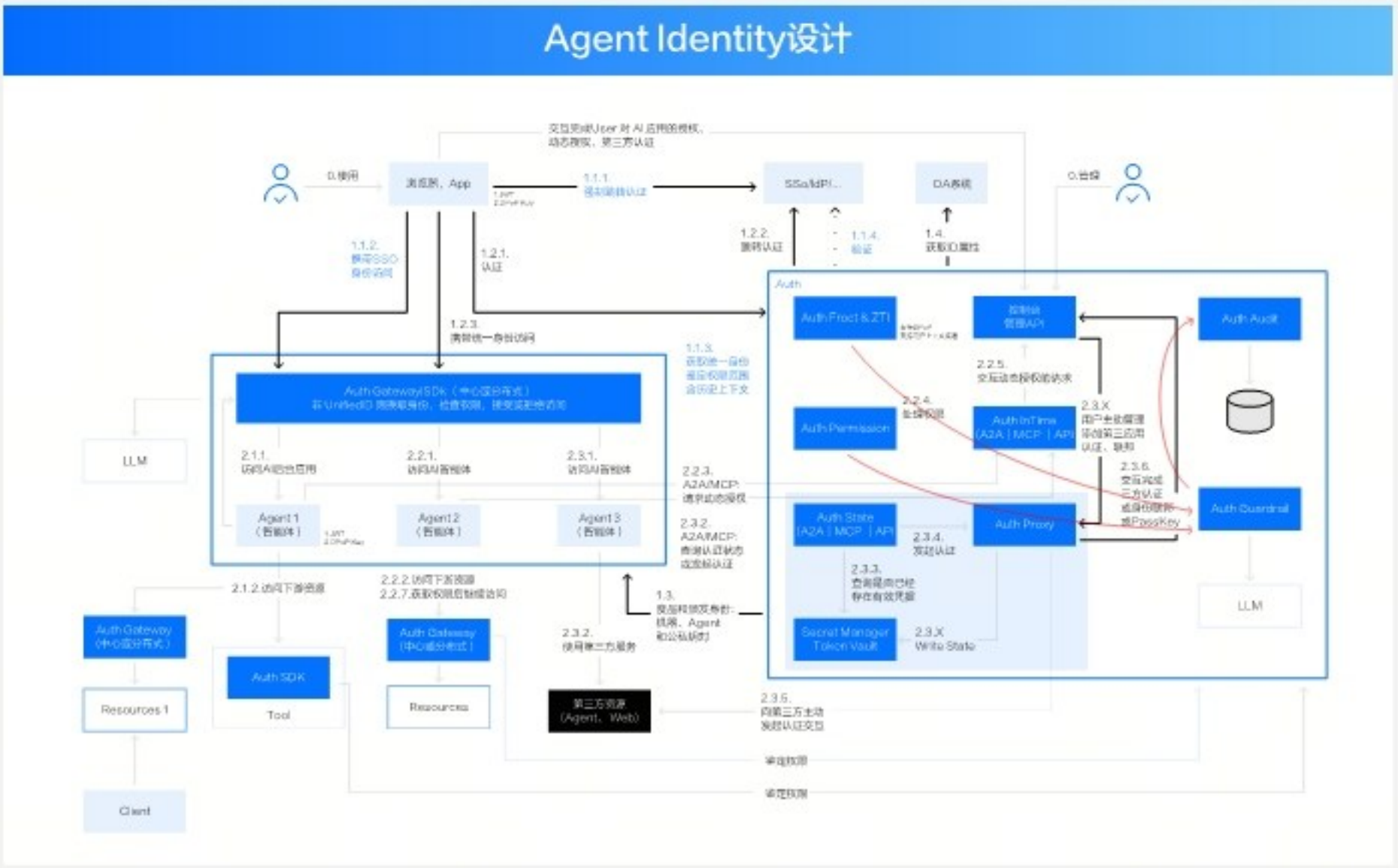
- **芯片级安全隔离：**从物理芯片（GPU/CPU）到容器的全链路安全隔离，即使底层基础设施被入侵，数据也无法被窃取或篡改，从根本上杜绝云服务商、大模型供应商以及其他非授权人员接触数据。
- **端到端数据保护：**用户在平台会话中的明文数据（如 Prompt、Response）仅在 TEE 内部可见，数据进出硬件隔离环境时，执行严格校验并加密传输；Prompt、Response 等数据不会被任何非用户的第三方触碰获取。
- **硬件级密钥托管：**基于可信硬件执行环境构建密钥管理系统，确保密钥全生命周期始终运行在可信边界内。
- **可验证透明信任：**提供可下载的远程证明报告，用户可离线验证部署环境的可信性，实现从“契约信任”到“技术可验证信任”的转变。

基于原生内置，火山方舟机密推理服务默认支持 PD 分离（Prefill-Decode disaggregation）高性能推理框架，在提供更高等级安全防护的同时，保证推理效率。

3.4.3 AI 智能体安全

身份与权限管理

火山Agent Identity核心目标是以统一身份为基础，贯穿 2LO/3LO 的细粒度授权、可信调用链（TIP）、与云平台 IAM 的深度集成，在零信任与全链路加密下，为人—Agent—工具—云资源建立清晰的权限边界与可审计的访问轨迹。



（图 8）

基本原则

- **身份统一且可验证：** Agent 与工具拥有独立、可核验的工作负载身份；不以用户身份“扮演”运行。
- **授权可分解且可审计：** 令牌承载链路信息，双重校验与最小权限在云—Agent—用户三方一致执行。
- **凭据可控且可轮换：** 静态 KMS 加密，传输 mTLS，短期临时凭证与自动轮换降低暴露窗口。

安全特点与设计

- **统一身份与可信调用链（TIP）：** 体系为 Agent 与其托管工具提供独立且一致的工作负载身份，跨容器、虚拟机与 Serverless 环境保持统一凭证与标识。每一次下游访问都会同时携带用户身份与 Agent 身份向身份服务请求令牌，返回的令牌包含链路信息，可被审计与回溯。这种“身份不扮演用户、而是以自身身份运行”的设计，天然划清边界，减少“代理权限被用户借用”的路径。

- **双重校验与最小权限：** 通过与云平台 IAM 的集成，Agent 在代入角色获取临时凭证时，云侧策略同时校验两类条件：请求是否来自受权的 Agent 角色，以及会话上下文中的用户是否对目标资源有访问权。只有两者同时满足才放行，既防越权也防“偷梁换柱”。临时凭证有效期短、绑定到具体会话，辅以最小权限策略，显著降低泄露时的影响面。

- **凭据安全与传输保护：** 凭据统一托管在 Token Vault，支持 OAuth 2.0、API Key、用户名密码与 STS 等类型，提供自动续约与轮换能力（包括数据库密码等）。所有静态数据使用 KMS 加密，传输通道采用基于零信任证书的 mTLS，且每次请求均进行签名、有效期与作用域的逐次校验，避免信任域内横向移动。

- **入站与出站授权：** 入站侧由合法 IdP 签发与校验用户令牌，并将用户的身份声明映射到访问控制与审计链路。出站侧按 Agent 的工作负载身份（TRN）控制外部凭证的发放范围；即便网络可达，若 Agent 身份不在授权名单，令牌不会签发、访问被阻断。TIP 与策略引擎共同构成出站调用的逐次约束。

入站（用户→Agent）

- 用户令牌由合法 IdP 签发与校验。
- 身份声明映射到访问控制与审计链路。
- 认证失败按策略跳转，避免未授权入口。

出站（Agent→工具/云资源）

- 按 Agent 身份控制凭证发放与范围。
- TIP + 最小权限 + 临时凭证共同收敛风险敞口。
- 未授权拒发令牌与访问，阻断越权调用。

- **审计与追溯：** 体系记录用户鉴权、工作负载令牌签发、外部令牌获取与代入角色等关键事件，包含时间、Agent 身份、用户 ID 与目标资源/操作等字段，形成完整调用链，可满足合规审计与异常溯源；在云侧与平台审计链路联动，便于统一取证与对账。

工具管理与准入

火山依托AI智能体，提供标准化的解决方案以及多种云原生工具，其中涵盖通过MCP HUB直接下载并使用火山云原生MCP工具以及第三方MCP工具。

多租户体验模式

体验模式强调以临时、受控的身份访问为主，兼顾安全与易用：

- **认证与权限：** 访问 MCP Server 需提供 **48 小时有效的 OAuth Token**，并通过 MCP 网关兑换为 **火山引擎 STS 临时凭证**，实现用户身份与权限的严格隔离。
- **网络与部署隔离：** MCP 网关与各 MCP Server 之间采用 **VPC 点对点单向打通**，基于账号进行网络层隔离；Server 部署在 **无公网 IP 的隔离环境** 中，降低外部暴露面。
- **高风险操作控制：** 火山侧 MCP Server 的工具能力经严格审查，**默认禁止高风险控制面操作**，避免误删、误改等非预期行为。
- **数据不驻留：** MCP 网关 **不保存租户数据**，准入过程也 **不允许 MCP Server 存储租户数据**，降低数据泄露与合规风险。

单租户部署模式

部署模式在租户自有 VPC 内运行，强调可控与兼容：

- **认证方式：** 允许使用 **长效 API Key** 进行认证，便于与既有系统集成。
- **访问控制：** 提供基于 **IP 黑白名单** 的准入控制能力。
- **部署便利性：** 支持 **一键将本地（Local）MCP Server 转为远程（Remote）**，提升交付效率。

工具自动化准入

MCP Server 上架至 Hub 必须通过 **自动化安全扫描** 与审批流程，覆盖常规Web安全风险与MCP新型安全风险，从源头提升MCP Server 安全性与合规性。

纵深防御与加固

针对智能体边界保护和工具集成常见风险，提供纵深防御能力。

（图 9）

输入侧防护

面向进入模型的请求，识别与拦截影响可用性与安全性的风险。包括：

- **算力抗 DDoS**：对消耗算力的恶意 prompt 进行识别与拦截，缓解“薅羊毛”和服务不可用风险。
- **提示词注入与越狱防护**：检测并阻断指令注入、越狱（jailbreak）等绕过规则的攻击路径。
- **敏感信息识别与脱敏**：对请求中的个人身份信息（PII）与业务敏感数据进行识别、脱敏与平行脱敏，降低出域泄露风险。

输出侧治理

面向模型响应的合规与质量控制。包括：

- **恶意与不良内容过滤**：识别仇恨、暴力、性、自残等不当主题，满足输入输出合规要求。
- **上下文不一致幻觉检测（preview）**：基于请求中提供的参照知识进行冲突检测，减少与源信息不一致的回答。

Jeddak AgentArmor

针对Agent的工作原理存在4个安全缺陷：过度依赖不可信环境输入、以过高权限访问用户资源、自然语言媒介的模糊二义、对外输出缺乏有效管控。使得Agent易遭受目标劫持、工具滥用等多方面威胁，面临数据安全破坏等风险挑战，此即AgentArmor目标覆盖的威胁模型。

AgentArmor 是 " 行为零信任 " 的运行时安全加固能力 （图 10）

运行时策略执行

运行时保护是 AgentArmor 的基础能力：在智能体实际执行过程中持续校验其行为是否符合安全策略，既可实时干预，也支持旁路离线模式进行事后审计。

上下文感知与动态策略

- 上下文感知让策略与具体任务场景贴合：根据不同智能体、不同工具与不同数据的上下文差异，给出更精确的威胁识别与风险评估。
- 动态策略调整在智能体获取新观察 (observation) 后，持续吸收变化并调整策略，降低漏放与误判，兼顾严格与宽松的平衡。

确定可验证与透明可解释

- 确定可验证t使结论可追溯：安全分析采用确定性方法，威胁检测与风险评估的依据可被复核与验证。
- 透明可解释在通过图的中间表示展示攻击路径与判断过程，用户能够理解“为何拦截、为何放行”，为内审与合规提供依据。

行为建模与信息流安全

AgentArmor 将智能体执行过程抽象为程序依赖图，并以类型系统表达安全等级与策略约束：

- **图构建器：** 把线性的运行轨迹转化为结构化的程序依赖图，捕捉控制流与数据流，让“思维链”“行为链”一目了然。
- **属性注册表：** 在为图中各节点（工具、数据、MCP、三方服务）附加安全属性；对未知工具与服务自动挖掘其数据操作流程并生成安全等级。
- **类型系统：** 以类型表示“安全等级”，在依赖图上自动推导并进行策略校验；在风险行为发生前给出升密、降密、告警、拦截等响应建议，保障机密性与完整性。

防护链路与零信任范式

AgentArmor 的防护链路由策略执行点与策略决策点双向联动构成，形成闭环：

- **策略执行点**作为执行枢纽，镜像 LLM 调用流量并采集上下文；依据决策结果允许可信调用通行，阻断或缓解不可信行为。
- **策略决策点**作为智能判断核心，汇聚行为轨迹并完成表示转化；结合动态策略生成与行为安全分析，输出安全决策供执行点落实。
- **运行时工作流**遵循“行为采集 → 安全判断 → 行为干预”的节奏，覆盖用户交互、LLM 调用与环境调用的全链路，支撑快速响应业务变化与新型攻击，形成“可感知、可干预、可进化”的安全共生体。

前沿技术研究

火山引擎定义了模型/Agent风险发现与评估方法，包括风险评估模型、安全性指标与度量方法、Agent安全相关的jailbreak手法、自动化安全评估，能够全面准确的评估Agent安全风险。火山引擎针对复杂Agent开展攻击面的梳理与挖掘工作，目前已发现多个复杂Agent存在高危漏洞，且借助大模型能力实现了风险的自动化发现。火山深度参与行业标准的制定，多次参与机密计算、智能体安全、MCP等领域的相关工作。

针对访问全链路进行安全加固，涵盖流量检测、行为审计、API安全、终端加固和合规保障等，确保整个访问链路的安全性。

4. 总结

4.1 生成式 AI 行业安全展望

生成式 AI 正在以前所未有的深度与广度重塑千行百业，但其安全问题也日益成为行业持续健康发展的关键瓶颈。我们预见，未来的 AI 安全将呈现三大趋势：

- **安全左移与 AI-Native 安全开发运维（DevSecOps）成为共识**安全不再是模型上线后的“附加品”，而是贯穿于数据采集、模型训练、应用开发到部署运维的全生命周期。从源头构建可信数据供应链，在训练中注入安全与对齐能力，在开发时提供默认安全的框架与工具，将成为企业 AI 应用的必选项。
- **从“单点防御”走向“体系化、智能化”**面对日益复杂的攻击手段，孤立的安全工具难以应对。企业需要构建覆盖基础设施、模型、平台到智能体的纵深防御体系。AI 技术也将被更广泛地应用于安全运营，通过智能化的威胁检测、风险研判与事件响应，提升安全运营的效率与精准度。
- **开放生态与责任共担成为主流**AI 安全生态的建设离不开开放协作。云厂商、模型提供方、应用开发者与最终用户需共同承担安全责任。标准化的安全接口、可信的数据共享、开放的威胁情报将促进生态各方高效协同，共同应对安全挑战。

4.2 火山引擎致力于保障生成式 AI 安全

作为领先的云服务提供商，火山引擎始终将安全视为生命线，致力于为企业提供可信、可控、合规的AI云原生基座。我们深知，生成式 AI 的安全之旅道阻且长，行则将至。火山引擎愿与广大客户及合作伙伴携手，共同迎接挑战，护航智能时代的到来，让每一个企业都能安心拥抱生成式 AI 带来的无限机遇。

生成式AI 安全白皮书

service@volcengine.com

400-850-0030（周一至周五 10:00-18:00）