HUAWEI

AI 安全解决方案白皮书

——构建面向 AI 基础设施的安全架构

商标声明

和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com

导读：AI 时代的安全架构：企业 CSO 需要了解和行动的 7 件事

1. 改变安全的定位：从附属成本→产业化核心底座能力

AI 技术的本质特性、风险内生性，使安全成为 AI 产业化发展的前置条件。安全与 AI 系统同构、同源、同生命周期。安全与 AI 发展并重，成为基础设施底座能力。

2. 规划新的安全目标：从防入侵、防泄漏→防失控、防认知颠覆

安全风险从外部入侵到全栈内生风险涌现，核心目标从封堵系统漏洞、抵御外部网络入侵，转向防范 AI 模型与智能体出现行为偏离、决策失准、权限越界等失控问题，以确保 AI 始终在人类设定的框架内运行。

3. 设计全新防护思路：从边界围墙防护→全栈内生免疫

AI 时代数据在云边端流动，模型通过 API 开放调用，智能体可跨系统执行任务，安全边界消失。全新的设计思路是安全要嵌入 AI 基础设施每一层：从芯片级的可信计算、算力调度的隔离机制，到模型训练、推理过程的数据投毒、恶意任务检测，再到智能体的身份、沙箱与权限，构建全栈内生免疫的安全系统。

4. 构建跨域的安全信任链：安全信任链成为 AI 底座

企业选择混合架构的 AI 基础设施，以应对 AI 的快速变化。新的安全架构需要构建全密态为核心的信任链，端到端加密促进数据流通，机密计算保障训推安全。

5. 降低智能化中枢不确定性：从训练开始治理模型安全

大模型本质上是一个概率性系统，其输出结果具有不确定性。当前企业智能中枢的开发模式从训练生产大模型，转变为基于基础模型后训练微调。从训练态安全治理开始，到推理态护栏，以降低模型不确定性的概率。

6. 抑制智能体意图偏离威胁：管控 Agent 全生命周期

企业的数字员工逐步进生产，自主执行操作失控和群体性失控的风险剧增。应像管理员工一样，以识别智能体危险意图为基础，构建 Agent 全生命周期的安全能力。

7. 用安全智能体重塑安全运营：AI 驱动事前制衡、动态治理

模型与智能体的快速发展，大幅降低了攻击门槛。AI 时代的安全运营，必须具备以 AI 制衡 AI 的能力。从人工处置到智能体自主执行，安全形态已逐步从事后补救、静态合规到事前制衡、动态治理。智能体将取代人工完成 90%以上的安全运营工作。

目录

1.AI 业务趋势分析

当前，人工智能产业发展范式正发生深刻变革，逐步从传统封闭、自建自用的私有生态，迈向混合云协同部署、开源基座主导、智能体规模化落地、数实深度融合的全新发展阶段。伴随 AI 算力架构、模型开发模式、应用形态的全面迭代，传统依托物理边界、静态策略、被动防御构建的安全体系已难以适配新时代 AI 业务的发展特征。AI 安全风险由单一单点问题转向基础设施、模型算法、智能应用的全栈式、体系化的风险。传统安全边界持续消融、新型安全威胁集中涌现，成为制约人工智能产业规模化、安全化、高质量发展的核心瓶颈。结合当前 AI 产业发展趋势，梳理全链条安全的挑战：

1.1 混合架构 AI 基础设施：跨域融合带来安全边界模糊风险

AI 算力建设模式已从企业重资产自建、封闭独享的私有算力架构，迭代为公私混合、云边协同的混合算力架构。混合架构依托公有云弹性算力资源与私有云核心资产沉淀的组合优势，有效降低了企业 AI 建设成本、提升算力资源复用效率。但多架构融合部署模式彻底打破了传统固化的安全边界，企业核心模型、训练数据、推理业务请求需在私有云、公有云、边缘节点之间高频跨域流转，数据存储、传输、计算全链路脱离单一可控边界，极易出现数据泄露、非法窃取、链路劫持等安全隐患，算力基础设施的全域安全管控难度显著提升。

1.2 公有云推理场景：规模化推理引发数据隐私保护难题

AI 算力结构呈现结构性迭代，推理算力已成为 AI 基础设施的核心支出主体，公有云推理服务凭借弹性伸缩、按需调度、低成本高效能的优势，成为企业大模型推理业务的主流承载形态。不同于封闭的本地化训练场景，公有云推理具备强开放性、多租户共享、业务高并发特征，企业开展云端推理业务时，需将包含商业机密、用户隐私、行业敏感信息的业务数据上传至公有云完成计算处理。公开共享的云端运行环境，使得用户隐私数据、业务核心数据面临窃取、留存、滥用、泄露等多重风险，规模化推理场景下的数据隐私保护压力持续加剧。

1.3 公共云战略升级：软硬芯全栈信任链可信验证挑战突出

随着 AI 算力成为数字经济核心战略资源，公共云已成为国家及行业 AI 基础

设施建设的核心载体，承载海量关键业务与核心算力服务。公共云的共享复用、全域开放特性，使其成为高级网络攻击的重点靶向，算力供应链安全风险沿芯片、固件、底层系统、云平台、上层应用全栈传导。软硬件任一环节的漏洞、后门与缺陷，可能引发全域算力集群失稳、核心业务瘫痪、关键数据失窃等重大风险。当前，如何构建覆盖芯片、硬件、基础软件、云平台服务的全栈可信的信任体系，实现基础设施可验证、可溯源、可信赖，成为政企 AI 基础设施安全建设的核心痛点。

1.4 模型后训练阶段：数据污染与概率输出带来内生安全风险

大模型研发模式已完成范式迭代，从全自研预训练的重资产模式，转变为依托开源基座模型进行企业适配，二次微调、增量训练的轻量化模式，大幅降低了大模型落地门槛与研发成本。但后训练微调所用行业数据来源繁杂、质量参差不齐，普遍存在数据清洗不充分、内容审核不完善等问题，极易混入错误信息、偏见数据、恶意投毒样本，造成模型数据污染。污染数据将被模型内化学习，引发模型幻觉、输出偏差、逻辑谬误等概率性安全问题，导致模型输出不可控、不可信，产生内容合规、业务决策、风险判断等多维度模型内生安全隐患。

1.5 数字员工智能体：自主化作业引发行为失控与滥用风险

AI 应用形态从被动应答式交互工具，快速演进为具备自主感知、智能决策、闭环执行能力的数字员工智能体。智能体可依托独立身份与权限体系，跨系统、跨场景自主完成业务处理、工具调用、数据读写、流程执行等复杂工作，极大提升业务智能化效率。但高度自主的运行特性，打破了传统人工干预、流程约束的安全管控模式，智能体极易因算法缺陷、指令劫持、策略漏洞出现越权访问、批量操作、数据窃取、恶意调用等失控行为，引发业务风险与数据安全事件，新型自主化 AI 应用的行为管控难度持续攀升。

1.6 数实融合具身智能：虚实联动带来全域安全失控隐患

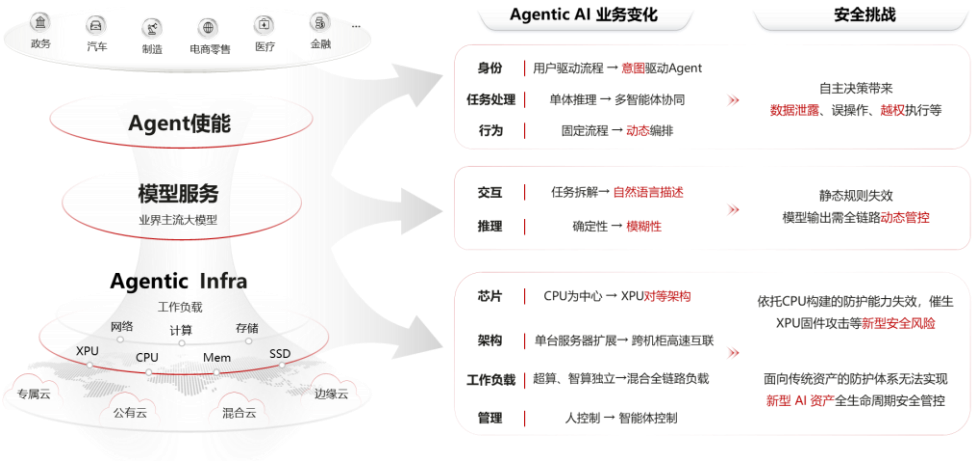
AI 技术正加速从纯数字化虚拟场景，向人机物融合的具身智能场景落地，工业机器人、智能装备、人形终端等具身智能体实现了数字智能与物理执行能力的深度融合。不同于纯软件 AI 应用仅作用于数字空间，具身智能可直接感知、干预、改造物理世界，使得传统 AI 数字安全风险全面向物理空间延伸扩散。虚实

融合的业务形态，让安全风险突破单一网络边界，形成数字风险、物理风险、场景风险叠加的全域安全隐患，传统安全防护体系无法覆盖数实融合场景的新型失控风险。

2. AI 安全总体框架

2.1 AI 业务变化与安全挑战

Agentic AI 正在赋能千行百业，深刻重塑数字基础设施的运行底座。这一变革不仅改变了业务的交付方式，更从根本上打破了传统安全领域在身份管理、攻击面管理、跨域信任及运行态防御上的基本假设。



图：AI 业务变化与安全挑战

Agentic AI 的自主性、动态协作、不确定性等特征，击穿了基础设施在身份管理、攻击面管理、跨域信任、运行态防御等的安全假设。

Agentic Infra：底层基础设施层正在经历算力底座的范式重塑。芯片架构已由 CPU 中心化迈向 XPU 对等架构；系统部署从单台服务器扩展升级为跨机柜的高速互联矩阵；工作负载完成了从超算与智算独立运行向混合全链路负载的融合；管理主体也由人工运维转向智能体自主控制。

安全挑战：架构的极速演进导致了旧有防御体系失效。依托传统 CPU 构建的硬件防护能力难以覆盖 XPU 固件级攻击及侧信道威胁；面向传统静态资产的防护体系，无法满足新型 AI 资产全生命周期的动态管理需求，形成安全盲区。

模型服务：作为智能中枢的模型服务层，其交互与处理方式发生了质的飞跃。业务交互不再依赖繁琐的任务拆解，而是通过自然语言描述即可完成复杂调用；

推理结果也从确定性的程序输出转变为具备概率性的模糊决策。

安全挑战：传统的静态过滤规则难以应对这种不确定性。模型输出的生成式特征，使任何固化的黑白名单机制都会面临失效。安全防护必须进化为全链路的动态管控机制，实时监控 Token 输入与推理过程，确保在复杂上下文交互中识别潜在的恶意诱导与内容合规偏差。

Agent 使能：Agentic AI 实现了业务逻辑的深层蜕变，业务交互范式由原本的用户驱动流程转向意图驱动 Agent；任务处理模式从单一推理演进为多智能体协作；系统行为也从预设的固定流程转化为基于环境感知的动态编排。

安全挑战：高度的自主决策能力直接引发了失控风险。在缺乏有效管控的情况下，Agent 可能导致敏感数据泄露、越权操作或执行偏离预期任务的行为。这要求建立能够识别管理智能体身份，并约束智能体意图的安全治理框架。

2.2 AI 安全解决方案理念

AI 大模型与自主智能体的广泛应用，并非简单的技术升级，而是深刻改变了攻防逻辑，传统安全体系面临三方面结构性挑战：

自主性引发行为失控：自主决策链使行为边界从预设转为实时推理，非预期的安全偏差概率显著上升。

概率性输出放大不确定性：AI 以概率为基础的输出特性，导致传统规则匹配式防御难以覆盖模型内生潜在的安全风险。

跨多系统导致信任域边界模糊：跨模型、跨系统、跨服务的复杂调用链路使传统安全域划分失效，威胁路径呈现高度隐蔽性与穿透性。以封堵和静态边界防护为核心的传统安全无法应对，需系统性重构。

图： AI 安全解决方案理念

从以隔离、访问控制为中心的集中式模式，转向以意图自治、观测管控、协同共治为中心的分布式模式：

以往的安全体系依赖安全大脑进行集中式决策、统一策略下发与全局态势掌控，所有安全事件需上报至中心节点进行分析与响应，其本质是中心管控、边缘执行的树状结构。在 Agentic AI 的分布式、高动态环境下，这种模式面临单点瓶颈、响应延迟、决策失焦等固有缺陷。

新的范式转向“监督管理、统一规范准则”，中心不再介入每一次具体的访问决策与行为许可，而是负责制定统一的安全规范准则、确立意图边界、建立可观测性基线，并对各分布式节点的自主运行进行监督审计。各 Agent、模型服务与基础设施节点在统一准则框架内，具备自主治理与协同共治的能力，能够实时做出符合规范的安全决策，同时接受中心的合规校验与异常追溯。

安全五大核心主张

释放智能交互边界：安全不再是约束，安全是 AI 业务的核心驱动力。安全前置"的设计理念重构交互范式，建立可解释、可审计、可干预的安全闭环，确保每一次交互都在可控范围内创造价值。

构建可控的自主系统：依托软硬芯协同的 AI 基础设施，确保自主代理的行为逻辑全程受控，杜绝越权与违规操作。

意图驱动的内生安全：将安全逻辑内嵌于认知层，通过深度解析交互意图，

从源头识别 AI 安全风险。

统一安全，协同共治：构建跨系统联动的安全防御生态，打破基础设施、模型、数据与应用间的壁垒，实现威胁情报实时共享与自动化响应。

全链可感、可追溯、可审计：建立面向 AI 推理和智能体应用过程、数据流向及决策依据的全链路观测机制，确保每一次生产力释放均符合合规与安全审查要求。

2.3 AI 安全解决方案架构

华为云 AI 安全解决方案以“内生安全+安全产品+统一安全运营”为核心框架构建的安全防护体系。防护范围自上而下芯片安全、硬件安全、AI Infra 安全、Agent Infra 安全、模型安全、智能体安全全层级。通过体系化的建设可有效抵御传统网络威胁、AI 基础设施、大模型、智能体等新型 AI 安全风险，保障业务稳定可靠运行，为企业 AI 规模化落地夯实安全根基。

图：华为云 AI 安全解决方案架构

芯片安全：作为大模型训练与推理的算力底座，AI 专用芯片（GPU、NPU 等）不仅承载着万亿级参数的计算任务，更掌控着模型知识产权、敏感训练数据和推理结果的全生命周期。一旦芯片存在硬件木马、侧信道漏洞或被供应链

植入后门，上层所有软硬协同、以白检黑的安全防护体系将彻底失效，攻击者可绕过软件层直接窃取模型权重、篡改推理输出、劫持 AI 决策流程，甚至通过大规模 AI 集群发起毁灭性攻击。芯片安全已不再是单一的技术问题，而是决定 AI 产业自主可控、国家数字主权和社会公共安全的根本性问题。华为云 AI 安全解决方案深度设计鲲鹏（CPU）和昇腾（NPU）的安全根技术。基于鲲鹏（CPU）的芯片安全技术包括 ARM CCA、机密计算、硬件可信根（RoT）等；基于昇腾（NPU）也包括机密推理、硬件可信根等根技术。

硬件安全：AI 时代的硬件安全已从传统 IT 架构中可被替代的边缘防护环节，跃升为整个智能社会安全体系不可动摇的信任根和最后一道防线：作为大模型训练推理、多智能体协同、具身智能执行的物理载体，硬件不仅承载着所有 AI 计算任务，更掌控着模型权重、敏感训练数据和推理决策的全生命周期。华为云 AI 安全解决方案在硬件安全层提供了，利用 UEFI 安全自动、固件完整性校验、硬件秘钥保护、分布式可信根以及擎天架构。构建基于硬件的信任根。硬件级安全能力具有不可篡改、不可绕过、性能损耗低的特点，能够防止攻击者通过篡改系统内核、BIOS 或固件来绕过安全防护，为 AI 时代的上层的 AI 业务提供了安全的根基。

AI Infra 安全：AI 基础设施（AI Infra）安全已成为 AI 时代整个安全体系的 "基石中的基石"，是所有大模型训练推理、多智能体协同、智能应用落地的前提与根本保障。它覆盖从 GPU/NPU 算力集群、分布式存储、容器编排系统，到 AI 框架、训练平台、推理服务的全栈技术栈。华为云 AI 安全解决方案 AI Infra 安全层主要包含围绕训练/推理的 AI 机密计算、A+K 异构机密计算（鲲鹏 CPU 信任域、昇腾 NPU 信任域）、可信智能计算、云技术设施安全等。

Agent Infra 安全：Agent Infra 安全层是专门针对多智能体运行基础设施的全栈安全防护体系，是多智能体系统安全的核心底座。它覆盖智能体从创建、调度、通信、工具调用到销毁的全生命周期，重点防护智能体身份伪造、跨智能体越权攻击、工具调用劫持、恶意指令注入、数据泄露和决策篡改等独特风险，确保所有智能体在可信环境中安全、可控地执行任务。华为云 AI 安全解决方案 Agent Infra 层主要提供 Agent 安全网关、Agent 身份安全、Agent 环境隔离

沙箱、Agent Harness 安全四大关键 Agent 安全基础设施。为 Agent 稳定运行提供保障。

模型安全：模型安全以全生命周期防护为核心，围绕模型生产和模型推理两大关键阶段构建分层、闭环的安全解决方案。目标为了从不确定性到安全可控的智能化中枢构建模型安全层的安全防护能力。在模型生产阶段，安全防护贯穿数据采集、清洗、标注、训练、微调、评估和发布的全过程，重点防范训练数据投毒、后门植入、样本污染和知识产权泄露等风险。在模型推理阶段，安全防护聚焦于模型部署后的运行环境和交互过程，重点抵御对抗样本攻击、模型窃取、推理数据泄露和恶意指令注入等威胁。

智能体安全：智能体安全是多智能体系统可信运行的核心保障，华为云 AI 安全解决方案智能体安全层围绕 Agent 资产安全、Agent 运行时安全、Agent 执行安全三个核心维度构建全链路、立体化的安全能力体系。Agent 资产安全聚焦于智能体身份凭证、模型权重、专属知识库、工具调用权限、配置策略等核心数字资产的全生命周期保护；Agent 运行时安全则覆盖智能体从创建、调度、通信到销毁的完整运行过程，通过内存隔离、微隔离、运行态行为监控和异常检测技术，抵御身份伪造、跨智能体越权、进程劫持和恶意代码注入等攻击；Agent 执行安全重点管控智能体与外部环境的交互环节，通过工具调用白名单、指令合法性校验、决策逻辑审计和执行结果验证机制，防范恶意指令注入、工具滥用、有害行为执行和决策结果篡改等风险。

2.3.1 内生安全：基于业务意图进行安全检测与防护

AI 时代安全范式将发生根本性的变革，构建内生安全能力，基于业务意图进行安全检测与防护的安全理念彻底颠覆了传统 "外挂式、边界式、特征式" 的安全理念，安全能力不应是业务系统的附加组件，而应从设计之初就深度融入系统架构的每一层，以业务本身的正常意图和合法流程为唯一判断标准，实现安全与业务的原生融合与同频共振。

图：基于业务意图进行安全检测与防护

AI 时代应该构建软硬协同、以白检黑、基于业务的运行态保护，软硬协同为业务意图的安全执行提供了不可绕过的物理底座，通过硬件可信根确保安全策略不会被攻击者篡改或绕过；以白检黑则将检测逻辑从 "防御已知威胁" 转变为 "验证业务意图是否正常执行"，通过学习业务的正常行为基线，精准识别任何偏离业务意图的异常操作。

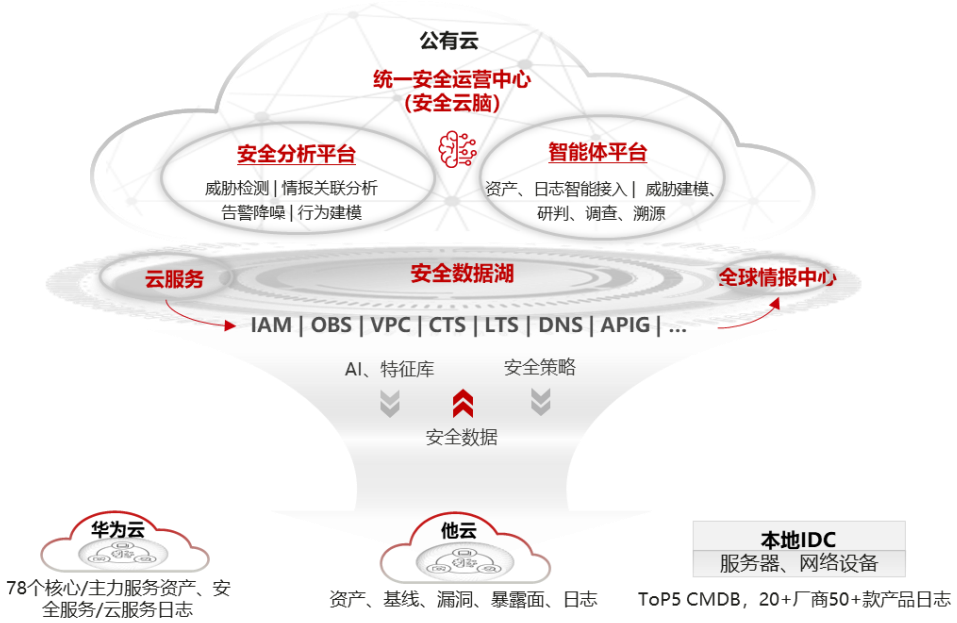
安全可信的产品内核是 AI 安全的基础，内生安全必须建立在安全可信的产品内核之上。没有可信的内核，所有基于业务意图的检测与防护都将成为空中楼阁。AI 时代的产品内核不仅包括传统的 CPU、操作系统和数据库，还涵盖了 AI 芯片、AI 框架、智能体运行时环境等核心组件。从芯片设计、代码编写到系统集成的每一个环节都植入安全基因，构建起从硬件到软件的完整信任链，才能确保业务意图在执行过程中不会被恶意篡改，实现真正意义上的内生安全。

Anthropic 于 2026 年 5 月 22 日发布的 "玻璃翼计划" 首月战报中，这组震撼的数据彻底改写了全球网络安全的行业认知：Claude Mythos Preview 模型仅用数月时间扫描 1000 多个核心开源项目，就发现了 23019 个各类漏洞，其中 6202 个被判定为高危或严重级别，占比高达 26.9%，这意味着每 4 个被发现的漏洞中就有 1 个可能直接导致系统被入侵、数据泄露或服务中断；然而更令人担忧的是，截至报告发布时，这些高危漏洞的整体修复率不足 1.6%，高危

漏洞从发现到补丁落地的平均周期仍长达 34-38 天。AI 时代逻辑漏洞的比例占到 5%-10%，内存安全的比例占到 70%，输入/注入占到 5%-10%。针对 AI 时代的逻辑漏洞以白检黑检测的核心思路，通过内存检测+进程检测+文件检测三重手段实现安全检测；针对内存中的安全威胁，比如做到软硬协同的漏洞防利用；针对输入/注入类威胁，基于业务的运行态保护 RASP/漏洞拦截补丁是有效的手段。攻击快速变化，通过内生安全构建产品韧性，是确定性的风险消减手段。

2.3.2 统一安全：全栈智能威胁监测能力，打造云内生协同防御体系

全栈智能威胁监测能力与云内生协同防御体系，是面向多云混合时代打造的新一代安全运营核心解决方案，它彻底打破了传统安全 "烟囱式建设、分散式管理、被动式响应" 的困局，构建了从数据采集、威胁检测、协同响应到价值呈现的完整闭环。



多云安全统一管理是体系的基础能力，它实现了公有云、私有云、混合云及传统 IDC 安全数据的全域统一纳管与标准化处理，依托大模型驱动的 AI 智能建模技术与全球实时威胁情报的深度融合，能够精准识别已知威胁、零日漏洞、APT 攻击等全类型安全风险，同时通过云、网、端三层安全能力的原生协同联动，将威胁响应时间从传统的小时级压缩至秒级，实现威胁的快速发现与

精准处置。

安全价值可观测是体系的重要突破，它解决了长期以来安全工作 "只做不说、价值难显" 的痛点，通过提供分层级的角色化看板与高度定制化的可视化大屏，将复杂的技术风险转化为不同管理层级能够理解的业务语言：SOC 专家可查看详细的攻击链分析与处置建议，CSO 可掌握全局安全态势与合规风险，CEO 可直观了解安全投入对业务的保护成效，让安全价值真正可量化、可感知、可汇报。

运营智能提效是体系的核心竞争力，它通过部署安全智能体实现了安全运营全流程的自动化与智能化，覆盖从资产自动发现、日志标准化接入，到威胁自动研判、事件自动处置的各个环节，其中核心的威胁研判准确率可达 95% 以上，大幅降低了人工误报率与重复工作量，将安全团队从繁琐的日常运维中解放出来，聚焦于高价值的风险治理与战略规划工作。

3.AI 安全建设指南

3.1 AI 基础设施安全—混合架构下可信赖的 AI 安全底座

3.1.1 AI 基础设施建设客户痛点

随着人工智能技术在政务、金融、能源、制造等政企领域深度落地，AI 基础设施的建设模式、算力重心与资源布局迎来根本性变革。如今，企业核心数据、行业知识库、自研模型持续向云端迁移，AI 基础设施已逐步成为政企全新的核心生产平台。在此背景下，行业面临的核心难题不再是算力资源短缺，而是全域安全管控能力不足。当前政企在当前发展趋势情况普遍面临以下核心顾虑：

- 核心数据上云后是否会泄露；
- 云厂商是否能够看到企业训练数据和自研模型；
- 数据在训练和推理过程中是否存在失控风险；
- AI 平台资产复杂，缺乏统一可视和治理能力；
- 模型权重文件是否可能被窃取或非法复制；
- 云上 AI 平台与私有云、IDC 之间的数据流动是否安全可控；
- 高敏业务是否能够实现独立隔离运行；
- 面对大规模的攻击，业务是否能正常运转。

3.1.2 建设思路—构建混合架构下“内生可信、密态运行、纵深防御”的算力基础设施

面向混合云 AI 基础设施，安全建设应遵循“内生可信、密态运行、纵深防御、持续验证”的总体建设思路，构建可信算力底座 + 全密态企业专属空间 + 云原生纵深防御三位一体的安全体系。

图：内生可信、密态运行、纵深防御的 AI 基础设施安全

- **以软硬芯协同构建 AI 内生安全算力底座。**

  从芯片级可信根出发，建立覆盖 CPU、NPU、服务器直至工作负载的全栈信任链，确保 AI 基础设施从底层起即可信运行，为上层业务提供不可篡改的安全锚点。

- **以全密态数据上云打造企业专属安全空间。**

  通过端到端加密、企业自持密钥（BYOK）以及机密计算等核心技术，实现数据与模型“可用不可见”，确保数据主权始终掌握在企业手中，满足合规与隐私保护的双重诉求。

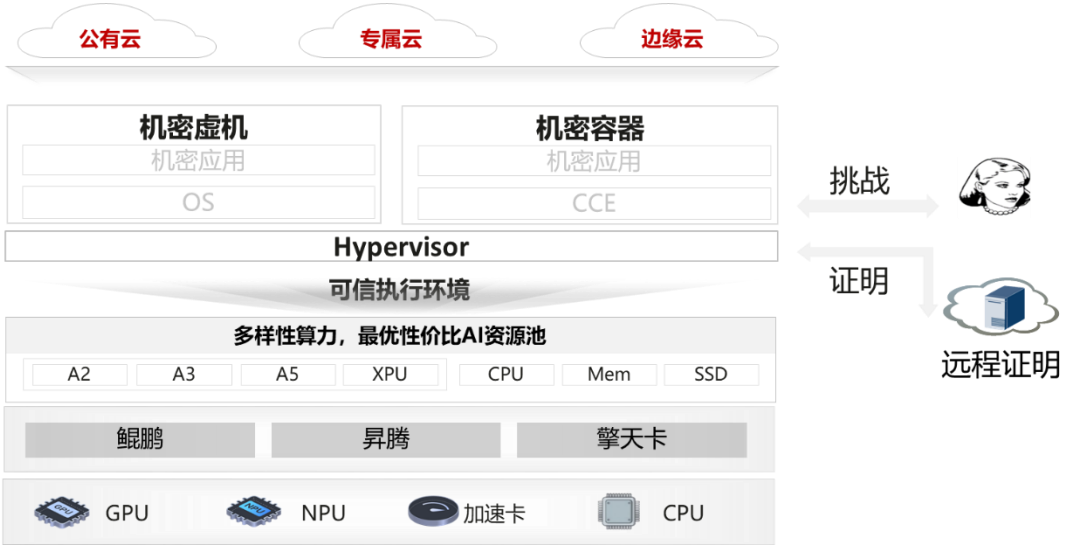
- **以“攻击不瘫、数据不丢”为目标构建云原生纵深防御体系。**

  依托身份、网络、应用、主机四层安全能力的深度协同，形成主动免疫、弹性自愈的防护架构，保障 AI 业务在高对抗环境下的持续稳定运行。

3.1.3 建设重点—软硬芯协同的 AI 内生安全算力底座

伴随 AI 算力资源从私有化部署向多形态公共云的战略转型，AI 基础设施的信任模型已从封闭可控的内部系统扩展为开放分布的多边算力体系。当算力资源不再由单一实体独占，供应链中的任意环节——从芯片硬件底层固件、网络传输链路、虚拟化管理层到 AI 应用容器——都可能成为攻击的突破口。在这一前提下，传统的信任链模型（将信任根植于操作系统或云平台管理层）已难以满足 AI

基础设施全域全栈的安全需求。政企应建设以“软硬芯协同、硬件根信任、全栈可证明”为核心的新一代信任链体系。



图：软硬芯协同的 AI 安全底座

3.1.3.1 专属 AI 资源物理隔离，从逻辑隔离到硬件边界的安全跃迁

专属 AI 资源物理隔离是面向高安全等级 AI 业务的硬件级边界防护技术。该技术对 AI 服务器、NPU 加速卡、网络端口、存储设备进行物理分区部署，将核心算力与公有云共享资源彻底物理断连，摒弃传统虚拟化逻辑隔离方案。针对政企核心算力、高密推理任务，可彻底规避多租户环境下的侧信道攻击、资源串扰、非法接入等风险，满足高密级业务的硬件隔离合规要求。

3.1.3.2 NPU 机密计算，AI 加速器的硬件隔离与数据保护

NPU 机密计算是面向 AI 加速芯片的硬件可信执行技术。依托 NPU 内置安全引擎，在芯片内部构建独立于宿主机操作系统、虚拟化层的专属可信执行环境，AI 模型、推理算子、密文数据均在该隔离域内完成运算，芯片缓存、中间计算数据全程对外不可访问，云厂商运维程序、租户进程均无法窥探内部数据。该能力有效防护推理过程中的数据窃取、模型泄露等威胁，为 AI 加速算力提供硬件级数据与资产保护。

3.1.3.3 CPU 机密计算，从虚拟机到容器的信任隔离基座

CPU 机密计算以通用处理器硬件安全能力为基础，为整机系统、虚拟机、容

器构建硬件级可信隔离域。该能力可为上层机密虚拟机、机密容器提供统一安全运行底座，并与 NPU 机密计算形成算力协同架构：通用业务负载运行于 CPU 可信域，AI 加速任务运行于 NPU 可信域，两类环境双向校验安全状态。有效防范虚拟化逃逸、内存数据泄露等风险，避免单点算力节点被攻破后引发全栈安全失陷。

3.1.3.4 推理混淆保护，抵御模型窃取与逆向工程的全生命周期防护

推理混淆保护是运行态下保障模型与数据安全的增强防护技术。该能力深度依托昇腾 NPU 机密计算环境，在硬件可信域内生成、存储混淆因子并与硬件设备强绑定，在不改动业务逻辑与模型结构的前提下，对模型权重、推理输入输出数据进行动态混淆处理，实现数据与模型 “可用不可见”。即便攻击者获取系统高权限，也无法还原原始明文与模型逻辑，以轻量化、易集成的方式，全方位抵御模型逆向、数据扒取、运行态窃密等攻击。

3.1.3.5 灵衢高性能加密通信，大规模数据交换性能零损耗

灵衢总线高性能加密通信是面向智算集群场景下多卡异构互联的全链路安全传输能力。通过硬件级加密引擎与智能协议栈的协同优化，实现数据在传输层的无缝加密与解密。在技术实现上，依托芯片内生的国密算法硬件加速模块，结合高性能流加密算法，在总线层级直接嵌入加密引擎。数据从显存或内存发出时，即被实时分段加密，并在接收端通过硬件流水线并行解密，从而实现了“传输即加密”的零拷贝模式。通过这种芯片级加密与流式处理机制，灵衢总线将原本仅覆盖单卡的安全域（即单张加速卡内的数据保护）无缝扩展至多卡互联的超大规模计算集群，最终达成多卡间性能零损耗的安全互联效果。

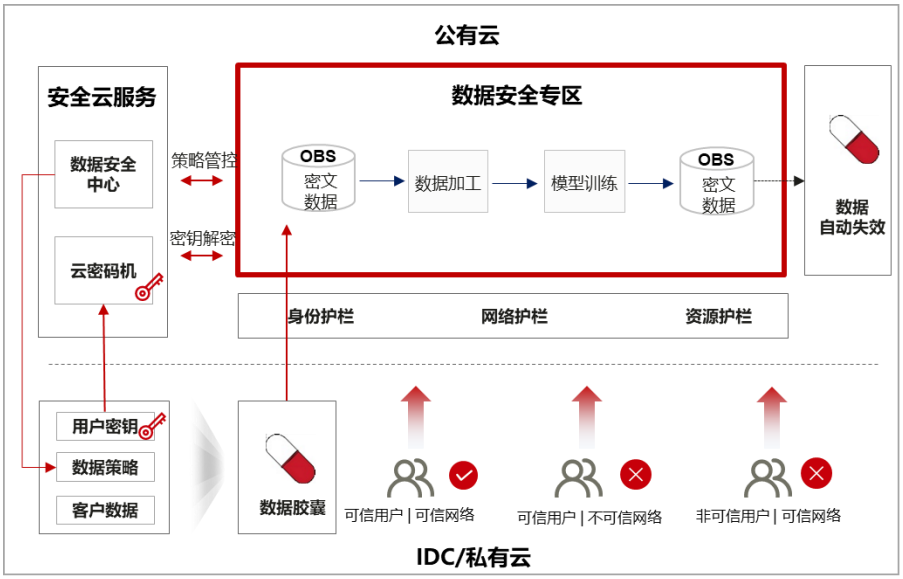
3.1.3.6 远程证明，全栈信任链的可验证闭环

远程证明是实现算力节点可信状态对外可校验、可举证的核心能力。用户可通过标准远程接口，对私有云、公有云、边缘云等全域算力节点发起可信核验，系统自动逐级校验芯片完整性、硬件隔离状态、机密虚拟化环境、安全策略及网络配置等全链路基线。一旦检测到固件篡改、环境异常、基线偏离等问题，将立即判定节点不可信并触发告警。该能力打通内部可信状态与外部合规举证的链路，

满足安全审计、第三方测评、跨云信任互认等监管与业务要求。

3.1.4 建设重点—全密态数据上云，企业专属安全空间

结合政企 AI 基础设施向混合云、公有云推理及全域公共云演进的总体趋势，针对前文提出的数据跨域管控、推理隐私保护、全栈可信等核心安全挑战，围绕混合云典型建设场景，落地“全密态数据流转 + 客户自持密钥”核心安全架构。该架构依托端到端加密、HYOK 客户自持密钥、数据出域自销毁、数据流转可观测、机密推理五大关键技术能力，构建覆盖数据全生命周期与全流转链路的安全防护体系，同时兼顾业务弹性、算力效能与合规要求。



图：全密态数据上云，企业专属安全空间

3.1.5 端到端加密，构建全链路数据机密性闭环

端到端加密是保障混合云 AI 数据安全的基础能力，覆盖数据存储、传输、计算全生命周期环节。针对 AI 混合云场景，实施差异化加密策略：静态数据（数据集、模型、配置文件）采用强加密算法实现存储加密；实时推理请求与边缘采集流数据使用轻量化加密协议，在控制时延开销的前提下完成传输加密。加密策略在私有云、公有云及边缘云之间统一联动，消除跨云边界带来的明文暴露窗口。该技术从根本上防范数据窃听、劫持与明文泄露风险，为 AI 业务提供全链路机密性保障。

3.1.5.1 HYOK（企业自持密钥），实现数据主权与云服务能力的解耦

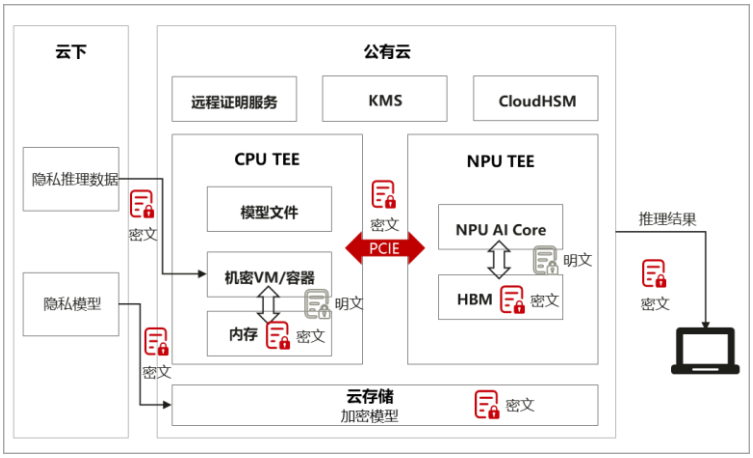
企业自持密钥（HYOK）要求所有加密密钥（根密钥、数据密钥、会话密钥）由政企客户独立生成、自主存储并全权管控，云服务商及任何第三方无权访问或篡改密钥。技术实现上，密钥可存放于客户自建的硬件安全模块（HSM）、本地密钥管理平台或符合合规要求的第三方系统中，云平台仅持有数据密文，无法独立解密。该技术有效应对多云环境下的数据主权旁落与内部人员越权访问风险，使政企客户在享受云服务弹性的同时，实现密钥层面的物理主权控制。.

3.1.5.2 数据胶囊，数据出域自销毁

数据出域自销毁是一种赋能敏感数据具备“离域自主合规管控”的主动防御技术。系统可以预先定义可信安全域，并配置基于数据分级、流转范围、访问权限及留存时效的精细化规则。当加密数据未经授权流出预设安全域、超出允许流转范围或达到留存期限时，自动触发销毁。该技术解决了混合云架构中数据脱离本地控制后可能被滞留、滥用或二次流转的难题，确保数据“用完即焚、离域失效”，实现数据生命周期终点的确定性安全。

3.1.5.3 机密推理，在推理服务中保护用户数据与模型资产

机密推理是一种依托硬件机密计算技术，确保 AI 推理计算在运行期内存中不发生“明文裸奔”的防御机制。在混合云设备的物理内存中构建隔离的硬件级可信执行环境（TEE），将 AI 推理模型、运算逻辑与加密数据封闭运行，与云平台操作系统、虚拟化层及其他租户环境完成物理与逻辑隔离，全程实现密文输入、密文计算、密文输出。该能力可有效规避推理环节的数据窃取、模型窥探、租户数据串扰等风险，使政企客户能够安全地将敏感推理任务部署于云端，兼顾业务弹性与资产安全。



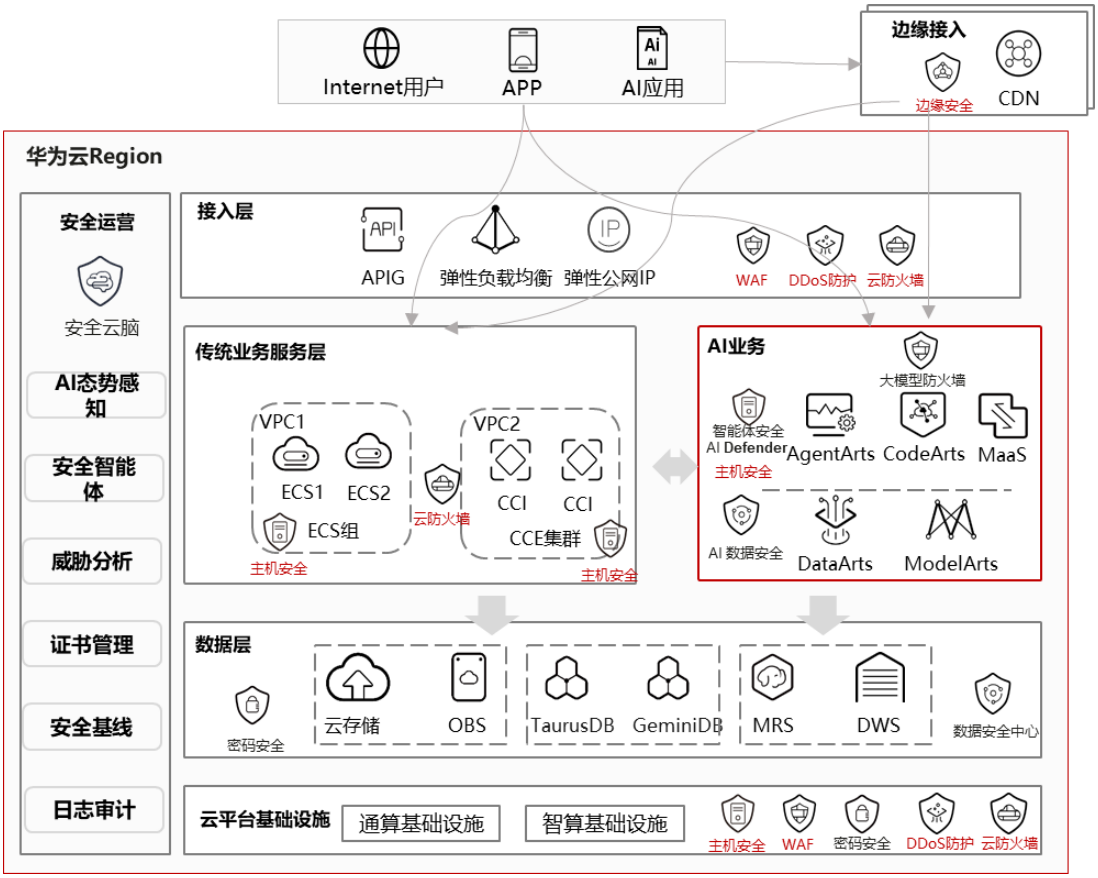
图：端云协同机密推理架构

3.1.5.4 数据流转可观测，实现跨域数据活动的全链路可见与溯源

数据流转可观测通过统一平台捕获数据在混合云环境中的全链路活动。该技术打通私有云、公有云、边缘云的日志、流量与算力调度数据，对每一条加密数据的产生节点、传输路径、访问主体、计算行为及流转去向进行全程记录与关联分析，形成可追溯的数据血缘图谱。解决了混合云多节点数据流转轨迹混乱、行为不可查、事件难溯源的问题，实现数据活动可视化、操作可审计、风险可追溯，提升跨域数据安全运营与事件处置能力。

3.1.6 建设重点—攻击不瘫，数据不丢的 AI 安全防线

AI 基础设施正从辅助算力供给演进为核心生产系统的关键基座，其安全防护能力直接决定业务连续性与数据资产安全。面对身份权限泛化、网络攻击高频化、应用接口暴露面扩大、主机环境复杂化等多维风险，传统单点防御已无法满足 AI 场景下动态、弹性、大规模的防护需求。必须围绕身份、网络、应用、主机四个维度构建纵深协同的主动防御体系，实现“攻击不瘫、数据不丢”的确定性安全目标，为 AI 训推业务提供全链路、全栈式的可信运行环境。



图：攻击不瘫，数据不丢的 AI 安全防线

3.1.6.1 身份防线，精细化权限管控消除人机+机机账号安全隐患

基于云上统一身份认证与权限平台，对 AI 算力、模型资源、数据资产实现精细化分权分域管控，严格收敛各类账号、接口、服务的访问权限；通过多因素认证、异地异常核验、高频调用风控等能力，抵御账号暴力破解、身份劫持、密钥泄露滥用等风险；结合权限动态回收与全量行为审计溯源机制，实时识别越权访问、批量爬取、异常调用等高危行为。从根源防范身份冒用、权限溢出、资源越权窃取等攻击隐患，保障 AI 核心资产仅支持合法主体合规访问，实现身份层面无越权、无冒用、无滥用，筑牢数据与资源安全的第一道防线。

3.1.6.2 网络防线，利用弹性网络架构防御超大规模流量攻击

针对 AI 训推业务流量体量巨大、接口暴露面广、端口固定、易遭受 DDoS、CC 高频冲击及链路嗅探入侵等痛点，依托全域组网资源、弹性流量清洗、智能攻击识别、网络微隔离技术，构建高韧性网络防护体系。通过全网流量实时监测

与 AI 业务特征智能识别，精准区分正常训推流量与恶意攻击流量，依托超大弹性清洗带宽，实现大流量攻击秒级引流与净化；通过云防火墙、安全组与 ACL 访问控制机制，完成算力集群、训推服务、存储资源的分层分域隔离，有效阻断横向渗透、端口扫描、链路劫持等网络攻击；结合攻击场景下的智能流量调度与冗余链路切换能力，在海量攻击、高频冲击场景下，持续保障网络链路通畅、业务端口稳定可用，解决传统自建网络带宽有限、防护固化、抗冲击能力弱的短板，实现攻击冲击下网络不中断、链路不瘫痪、业务不卡顿。

3.1.6.3 应用防线，构筑云上 AI 业务前端智能化安全屏障

面向 AI 训推接口、服务网关、前端应用等核心业务入口，依托 AI 专属应用防护、接口安全管控、高频访问限流、恶意请求智能拦截能力，精准防御 AI 场景专属定制化攻击。通过适配大模型训推业务特征的防护规则体系，精准拦截提示词注入、恶意参数构造、接口越权调用、批量爬虫、服务重放等高频攻击；结合语义识别与用户行为风控能力，精准甄别异常访问、批量薅取模型、窃取推理隐私数据等恶意行为，实现恶意请求实时阻断、正常业务无损通行；同时具备应用层攻击态势感知、攻击链路溯源、防护策略自适应迭代能力，持续适配新型攻击手法。全方位守护 AI 训推服务稳定运行，规避应用层攻击引发的服务宕机、接口瘫痪、数据泄露等风险，实现业务入口可防、攻击可拦、数据可保护。

3.1.6.4 主机防线，标准化建设算力集群环境安全底座

针对 AI 算力服务器、GPU 节点、训练集群等核心算力资源，通过安全基线标准化加固、漏洞动态自愈、恶意程序实时查杀、运行态动态防护、集群多副本冗余备份等技术，构建全时段、全覆盖的主机安全防护能力。统一固化 AI 专属主机安全基线，对系统权限、开放端口、内核参数进行标准化加固，消除弱配置、漏洞暴露等原生安全隐患；依托云端主机安全监测能力，实时感知算力节点异常进程、恶意脚本、暴力破解行为，自动查杀木马、挖矿、后门等恶意载荷；针对训推运行环境实现动态防护，精准拦截恶意注入、程序篡改、环境破坏等攻击行为。同时配套节点冗余、数据多副本、定时快照与快速恢复机制，即便单节点遭遇攻击沦陷、程序损毁、系统异常，仍可快速还原数据、恢复业务，实现算力节点抗入侵、运行环境不篡改、核心数据不丢失、集群服务不瘫痪。

3.2 模型安全—从不确定性到安全可控的智能化中枢

3.2.1 模型安全建设客户痛点

政企客户在基础大模型上，通过 SFT 监督微调以及 DPO、RLHF 等偏好学习与强化学习方法，基于其行业数据与业务知识持续构建专属企业模型能力，模型从通用能力演进为企业的核心生产力。在这基础模型后训练过程中，训练数据、对齐策略、行业知识以及模型权重本身，成为企业关键数字资产。

同时伴随 Agentic AI 快速发展，大模型正在从内容生成工具演进为具备任务规划、工具调用与自主执行能力的智能系统。模型不仅是用户的输入输出处理，而且是同步覆盖用户请求、上下文、多轮对话、RAG 检索与 Tool 调用的复杂调度和执行系统。

当前企业普遍面临以下核心顾虑：

- 后训练是否引入数据污染、隐私泄露与对抗样本，导致模型行为失控？
- 对齐策略、LoRA 与行业知识是否会被篡改、窃取或绕过安全约束？
- Prompt 注入、越狱攻击与上下文污染，是否会诱导模型偏离任务目标？
- RAG 与 Tool 调用链路是否存在知识污染、越权访问与危险执行风险？
- 是否存在算力滥用、数据泄露、恶意内容生成与内容合规风险？
- 能否观测模型为什么这样决策、执行了哪些行为，以及风险从哪里产生？

这些问题共同指向一个核心矛盾：模型行为的不确定性，与企业对系统可控性的要求之间的结构性冲突。

3.2.2 建设思路—安全左移，贯穿生成与推理的连续安全架构

图：安全左移，贯穿生成到推理的模型防护

与传统 Web 安全围绕静态请求与响应构建物理边界的范式不同，大模型智能化中枢安全的安全对象是持续生成的 Token 流、动态演进的知识访问以及非确定性的实时工具行为。

因此从用户的业务流出发，从基础模型后训练生成，再到运行推理，采用后训练内生对齐、运行时安全护栏双重安全建设思路：

内生对齐，通过后训练监督微调和偏好对齐双层机制，在模型参数空间内深度刻画基础安全边界与默认风险偏好，使模型具备默认不作恶的参数级内生安全能力。

安全护栏，围绕 Token 推理流、RAG 链路与 Function Call 架构构建运行时动态安全护栏，对提示词注入、间接投毒、越权检索、参数污染及工具滥用等运行态复合风险实施实时治理与外在约束。

通过上述双层机制，实现模型安全能力从静态规则控制，向运行态持续治理的演进。

3.2.3 建设重点—内生对齐，构建后训练模型内生安全机制

内生对齐是模型安全体系的基础能力，通过后训练机制，将安全能力内化为模型自身的行为约束与决策偏好，使模型在生成过程中天然拥有安全边界。方案围绕数据安全处理、安全对齐、安全测评与模型保护四个核心阶段，构建模型从后训练到交付的全生命周期内生安全强化链条，将模型锻造成具备高度安全自律的可信生产力底座。

图：内生对齐，构建后训练模型内生安全机制

3.2.3.1 数据处理，企业数据脱敏和防投毒

后训练阶段的数据安全，与预训练处理海量公开语料有本质区别。企业注入的是私有、高敏、小样本数据，风险焦点从语料偏见转向隐私泄露与定向投毒。通过对训练数据实施安全处理准入机制，确保进入 SFT 微调的数据经过脱敏与安全过滤。

数据脱敏与防记忆，自动化擦除训练语料中的姓名、身份证号、联系方式等 PII 个人信息，以及客户账号、交易流水等企业核心业务字段。同时引入成员推断攻击验证机制，检测模型是否会过度记忆脱敏前的原始隐私，防止攻击者通过对抗性提示词、推理诱导等方式逆向推导训练集敏感内容。

内容安全过滤，对每一批训练数据进行内容安全净化，通过毒性识别与风险语义分析能力，过滤涉黄、暴恐、极端导向及其他高风险违规内容，降低模型学习有害行为模式的风险。针对第三方数据源及众包标注数据，还需建立异常样本检测与偏好分布分析机制，识别潜藏的后门触发器、异常标签关联及隐蔽投毒行为，防止攻击者通过少量恶意样本对模型植入特定触发响应或隐式行为偏置。

3.2.3.2 安全对齐，行为约束与过度拒绝控制

安全监督微调 SFT 与基于反馈的强化学习 RLHF/DPO 等是模型行为对齐的核心环节。需摒弃单一、死板的微调模式，采用多维数据集组合与按需渐进式对齐的弹性架构，在恶意输出防范、隐私意识内化与业务高可用保持之间实现多目标最优解。

分层可组合的数据策略

通用安全 SFT 数据集，覆盖直接越狱、多轮诱导、角色扮演、间接注入等典型对抗模式，配置标准拒绝回答样本。当基座模型安全退化时，作为安全锚点注入，为后续精细化对齐提供冷启动加速。

场景级隐私合规样本集，针对静态 PII 脱敏无法完全覆盖的上下文关联隐私，如罕见病历组合、特定业务高敏关联信息等，提供 安全-不安全 偏好对照样本，训练模型在复杂语境下的隐私感知与柔性保护能力。该项训练建议为可选，仅在静态脱敏不足以消除隐私风险时启用。

企业自有业务数据集，融合特定行业合规数据集与真实业务场景下的正常问答对、边界案例，用以维系模型在特定业务领域的语义准确性与泛化能力。

多阶段可配置的渐进式对齐

轻量化场景，直接采用业务 SFT 微调到 DPO 直接偏好优化的极简链路，满足高泛化、低开销需求。

高安全敏感场景，部分场景如医疗、金融、企业关键业务，启用多阶段渐进式对齐。先利用通用安全及隐私增强数据集实施轻量级监督微调，强制模型锁

定全局安全基本面；再引入业务数据集实施领域适配；最后接入 DPO 或 RLHF 基于人类和 RLAIF 基于 AI 反馈的强化学习框架，借助安全奖励模型引导模型在高维参数空间进行多轮博弈优化。

灰色请求的引导与过度拒绝收敛

在阻断明确违规请求并输出标准化拒答模板的同时，针对边界模糊的灰色请求，如特定行业的合规边缘问答，输出合规的引导性回答与风险释义，有效收敛因过度对齐而引发的过度拒绝缺陷，确保模型在复杂对抗环境下依然具备业务可用性。

3.2.3.3 安全测评，自动化评估模型安全性

通过模拟真实世界中黑客的长序列、复合型高频语义攻击，开展模型系统化安全测试，为上线提供可量化的评估依据。模型测评覆盖安全抗性、行为鲁棒性与合规风险三个维度，采用自动化红队与专家人工会审的双轨评估机制。

安全抗性测评，基于多智能体的自动化红队平台，利用攻击模型对待测模型实施长序列、变体组合的提示词注入与越狱探测，识别因参数对抗衍生的安全漏洞与隐私泄露风险。

行为鲁棒性测评，模拟实际生产中的复杂输入与长尾极端场景，通过对模型输出概率的统计学分析，量化评估模型在特定业务语义下的回答准确率、逻辑收敛性与运行稳定性。

法律合规测评，参照国家网信办大模型备案要求与行业特定合规规范，动态审查模型输出是否存在违规内容等合规红线问题。

测评平台对测试流量与评测结果进行自动化打分与分类标注，最终生成漏洞分布热力图与量化安全得分，在模型上线前识别并修复防御盲区。

3.2.3.4 模型保护，构建权重混淆的可信执行环境

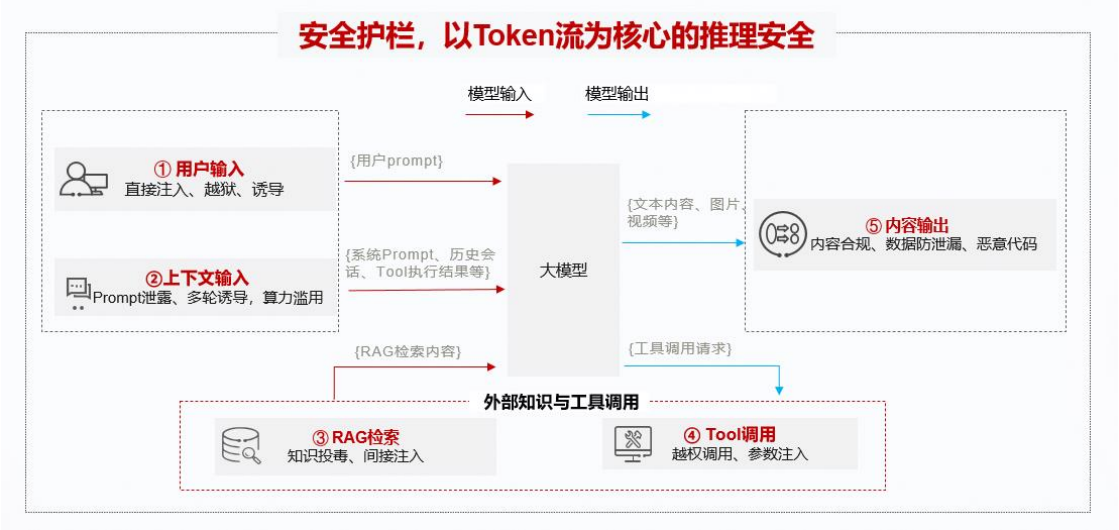
采用软件层混淆与硬件可信执行环境相结合的双重防御体系，守护核心资产的机密性与运行安全，防止模型权重文件在部署与运行过程中被窃取、逆向分析或遭受内存转储攻击。

模型混淆，部署阶段，引入权重保护与参数变换机制，对模型参数进行结构化扰动与映射转换，在不改变模型功能可用性的前提下，提升模型结构被直接逆向分析的难度与成本。推理阶段，通过输入侧编码扰动与推理过程隔离机制，使云端仅处理经过变换后的输入表示，从而降低原始语义信息在云端环境中的直接暴露风险，并在可信侧完成结果还原。

可信执行，在高安全等级场景中，利用支持 TEE 的机密计算能力，将模型权重、推理数据及关键密钥封装在受保护的硬件执行环境中运行。通过内存加密与访问隔离机制，实现计算过程与宿主操作系统的逻辑隔离，从而在基础设施可能存在风险的情况下，降低模型资产被直接访问或泄露的可能性。

3.2.4 建设重点—安全护栏，以 Token 流为核心的推理安全

Token 在运行过程中，会持续穿越输入解析、上下文构建、知识检索、工具调用与内容生成等多个阶段，因此运行时安全并非孤立模块防护，而是围绕 Token 生命周期的连续治理。



图：安全护栏，以 Token 流为核心的推理安全

3.2.4.1 用户输入安全，拦截提示词注入与越狱攻击

用户输入是大模型推理的最外层交互，是模型推理的前置语义审计与可信边界。防护重心是在正式消耗算力进行内核推理与自主规划之前，对外部原始请求实施语义过滤，御敌于国门之外。

用户语义识别，通过关键词结合轻量化、高性能的语义对抗模型，在保障 TTFT 模型首字延迟的前提下，对外部原始请求实施 Token 级安全扫描，识别并实时截断直接越狱与显性提示词注入攻击。

显性注入与越狱拦截，灵敏识别直接越狱与显示提示词注入，覆盖目标劫持、角色扮演、反面诱导、场景任务、对立响应、虚拟对话、DAN 攻击、代码注入等，检测概率触发预设阈值，实时截断当前请求，禁止其转化为计算张量。

角色边界语义硬隔离，在用户输入格式化与 Prompt 拼接阶段，强行隔离用户与系统的角色边界，确保输入意图的纯净性，从源头防范用户恶意指令篡改或覆盖默认系统设定。

3.2.4.2 上下文安全，防范多轮诱导与算力滥用

随着攻击手段从单轮恶意 Prompt 欺骗演进为多轮上下文诱导、角色边界绕过、隐式语义操控等，防护边界必须向内扩展至整个上下文构建链路。聚焦于长文本语义中的越狱变体识别、多轮历史对话的动态审计以及算力资源的保护。

多轮对话语义风险检测，识别长文本语义上下文的越狱变体，锁定系统 Prompt 与初始角色边界，防止攻击者通过多轮渐进式催眠、角色扮演或社会工程学手段诱导模型误判。同时识别并阻断工具返回中的隐藏指令、RAG 间接注入等攻击向量，防范模型发表违规言论、盗取核心提示词或执行恶意操作。

防止算力滥用，引入推理深度控制与 Token 级速率约束，识别循环规划、无意义多轮推理与对抗性输入触发的计算放大攻击，第一时间熔断对抗性欺骗请求，防止昂贵的算力资源被无端消耗。

3.2.4.3 RAG 检索安全，防范知识投毒与间接注入

模型通过 RAG 获取外部知识，安全风险升级为知识投毒与间接注入。任何通过 RAG 检索并拼接进大模型上下文窗口的第三方文本，均视为外部不受信流量，需实施严格的知识检测和治理。

知识库静态扫描过滤，通过敏感数据识别、违规内容检测与知识库扫描能

力，对知识文档进行全量安全检测，防止恶意文档、知识投毒与隐式 Prompt 注入进入模型上下文，扫描内容覆盖文本、元数据及嵌入向量本身。

检索行为审计与异常拦截，通过向量检索行为分析与检索攻击检测能力，对检索注入、高频高开销向量检索 DoS 以及偏离正常用户行为的异常查询行为进行实时识别与拦截。

上下文可信度评估与注入控制，对检索结果进行动态可信度评估与结构化安全标记，对高风险知识切片实施降权与上下文隔离，限制其对系统提示词与核心指令的影响范围，使外部知识仅作为事实性参考信息参与推理，从机制上防止间接 Prompt 注入与语义污染。

3.2.4.4 工具调用安全，约束越权调用与危险参数

Function Calling / Tool Using 是大模型连接外部业务系统的重要机制，其安全风险贯穿于模型对任务意图的理解与工具调用结构生成全过程。在模型层安全体系中，仅对工具调用的生成阶段进行约束，以确保模型输出的工具调用意图符合预设安全与结构规范。

模型层安全的核心目标，是对工具调用意图生成进行结构化治理，使模型始终在受控的语义与格式边界内生成可执行描述，而不涉及实际执行权限管理。

调用生成阶段，对模型生成的 tool_calls 结构进行语法约束与语义校验，识别异常调用意图、不符合安全策略的操作请求以及潜在越权行为模式，确保输出符合预定义工具接口规范。

参数生成阶段，对 SQL、Shell、API 等工具参数进行安全约束与内容校验，识别注入式攻击、危险命令拼接及非法参数组合，防止模型生成具有潜在风险的执行参数。

需要强调的是，模型层仅负责工具调用的生成约束与结构安全，不涉及工具执行权限控制与业务系统访问策略管理。工具执行侧的权限校验、调用审批与业务安全控制属于 Agent 运行层职责范畴。

3.2.4.5 输出内容安全，防止信息泄露与内容违规

模型输出已经不再只是聊天文本，而正在成为下游业务系统的执行参数、自动化决策依据与智能体行为输入。必须建立高性能低延时检测能力，规避错误信息、高敏数据或非预期内容向用户侧与企业下游生产环境的传导。能力覆盖模型输出 Token 流的实时滑动窗口、出站前的文本缓冲区、动态脱敏引擎以及阻断后的重定向通道。

流式内容合规检测，通过高精度、多标签的合规分析集群，对模型生成的文本流实施字词级与语义块的双重实时流式滑动窗口扫描，无需等待整句生成完毕，识别并拦截涉政、涉黄、暴恐及行业特异性违规内容，重定向至安全提示语，确保输出结果满足基础合规要求与监管标准。

企业敏感数据保护，对即将吐出的文本进行深度上下文推理，自动识别并对其中的企业核心代码、高敏财务数字、用户身份证、手机号以及 API Key 等凭据密钥实施动态语义遮蔽或脱敏替换，防止企业敏感数据泄露。

3.3智能体安全—智能体全生命周期安全管控

3.3.1 智能体安全建设客户痛点

政企用户将 AI 智能体引入办公、生产环节，要进行运行环境和企业知识库建设、同时对接常用工具、基于使用者角色设定智能体的角色和权限范围，经过测试后与生产/办公系统工作流集成，上线运行后持续扩展技能，最终成为团队的工作助手，支持政企单位探索数字员工、数字部门建设。

智能体当前最大的价值是基于对用户任务意图的理解，自主规划、分解任务并执行，最终给出用户期望的任务结果。但智能体的自主决策执行、自行编写代码、工具调用/Skill 加载将导致攻击面扩大，同时由于大模型的任务分解机制通常基于概率分布进行，多智能体执行复杂任务时可能因规划错误导致行为失控，对生产/办公 IT 环境及数据造成不可逆破坏。

当前政企用户主要有以下顾虑：

1、担心生产、办公及其他数据被智能体违规访问、破坏或泄露，并且无法回溯

- 智能体行为超出权限边界导致异常操作
- 提示词注入攻击导致系统破坏或数据泄露
- 智能体异常操作无法及时溯源定位

2、担心 IT 环境因智能体失控遭攻击、集群宕机

- 云基础设施被横向攻击，集群运行安全稳定遭破坏
- 不受控影子智能体穿透安全防护体系
- 智能体调用技能工具时权限越界
- 不安全输出引发级联拒绝服务攻击和异常算力消耗

为确保智能体持续受控，结合智能体部署使用过程，政企单位性需要建设安全可信的智能体运行环境，避免系统被攻破、智能体运行失控；同时要持续检测、过滤恶意任务，减少暴露面。

3.3.2 建设思路—构筑“环境内生可信、意图威胁检测”的保护机制

传统的、基于静态边界与滞后补丁的防护模式，在面对由自然语言驱动、具备“高行动权”的自主实体时面临系统性失效。建设智能体安全，必须打破传统外挂审计的局限，以“可信运行环境”作为硬性约束底座，以“动态行为管控”作为实时反向约束边界，将安全内嵌于智能体开发、部署、运行、决策执行的全生命周期，锻造出“全时可信、行迹可控、防线可管”的自适应安全防护体系。

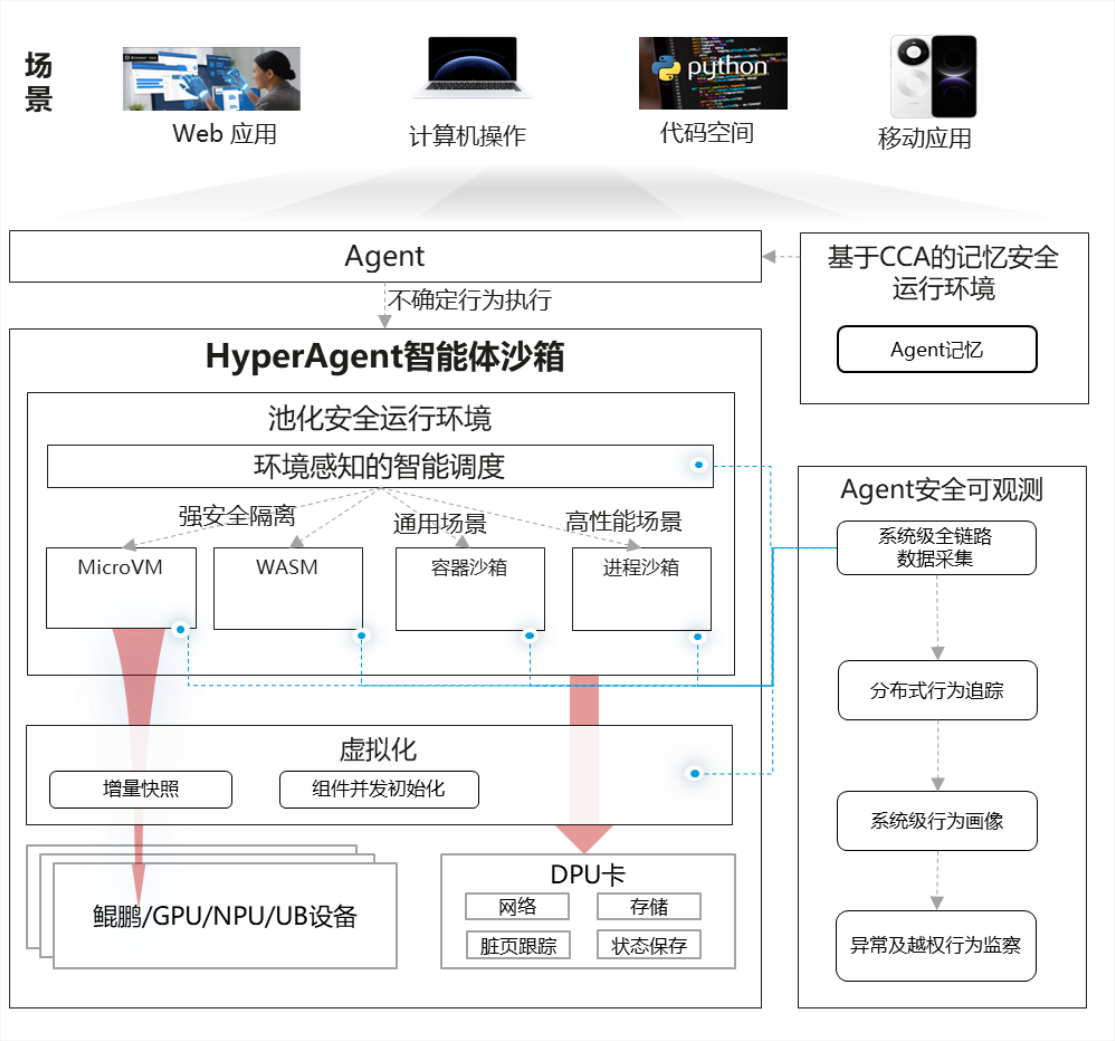
智能体部署前，政企用户需要进行可信运行环境建设，保护智能体系统和数据的机密性和完整性，防止智能体系统和数据被篡改或劫持，出现异常情况可及时追溯；确保基础设施层有足够的安全能力对智能体行为进行硬性约束。建设内容包括智能体资产漏洞识别与管理、安全沙箱、机密计算环境、全流程审计能力

智能体部署运行后，对任务输入、决策推理、工具调用操作执行过程中进行实时监测和异常干预，确保智能体攻击面持续收窄、风险任务持续受控，防止因外部攻击、推理错误等原因导致智能体行为失控，同时保护算力不被异常消耗。建设内容包括针对智能体的身份认证和动态权限控制、基于企业角色的行为基线

控制、用户 Prompt 恶意任务输入过滤、恶意意图任务和高危操作威胁检测能力

3.3.3 建设重点—软硬协同的可信 Agent 基础设施运行环境

构建安全可信的智能体开发运行基础设施，以基础设施保障硬件和系统层的安全策略强制执行，杜绝智能体应用和数据被非授权访问和篡改，为上层业务提供可靠的信任支撑。



图：可信 Agent 基础设施运行环境

3.3.3.1 自动化资产管理，杜绝 IT 环境中的影子 AI

政企单位在将智能体引入生产办公环境时，必须构建完善的资产治理机制。首先需对全网 IT 系统中的“数字员工”进行全面的普查与合规登记，严防未纳

入监管的 AI 智能体 系统沦为外部攻击的跳板。依托云主机安全防护能力，对智能体运行时环境开展全天候、自动化的资产动态测绘，精准识别其依赖的软件组件及对应版本存在的潜在漏洞。通过建立体系化的漏洞收敛机制，确保中高危漏洞得到及时修复、低危漏洞处于受控且无法被利用的状态，从源头彻底杜绝 AI 资产失控风险。

3.3.3.2 智能体沙箱，限制智能体恶意行为

针对智能体自主执行系统级操作与复杂任务时带来的未知风险，系统摒弃了单一的隔离模式，构建了基于风险动态感知的多级沙箱架构。该架构能够实时评估智能体行为（如调用外部工具、处理敏感数据或执行动态生成代码）的风险评级，并自动调度、动态切换至相应等级的沙箱环境中运行。通过多级动态流转机制，在隔离不可信代码并精准阻断攻击行为的同时，确保仅将安全的校验结果返回给用户，实现安全性与云计算集群性能的平衡。

3.3.3.3 Agent 记忆机密运行环境，防止记忆数据被违规访问、恶意篡改

智能体 记忆数据在传输与存储阶段均采用高强度加密算法，实现全生命周期的密文保护，仅在运行时进行即时解密。通过引入机密计算架构（CCA），将数据的解密与运算严格限制在机密计算环境中进行。基于 CCA 的强隔离特性，即便是云平台的超级管理员也无法访问或篡改运行中的记忆数据，从根本上解决了高权限用户越权带来的安全威胁，确保用户隐私与企业核心机密绝不泄露，同时使云平台服务商更加可信。

3.3.3.4 全流程审计溯源，操作行为全集记录、出现问题可回溯

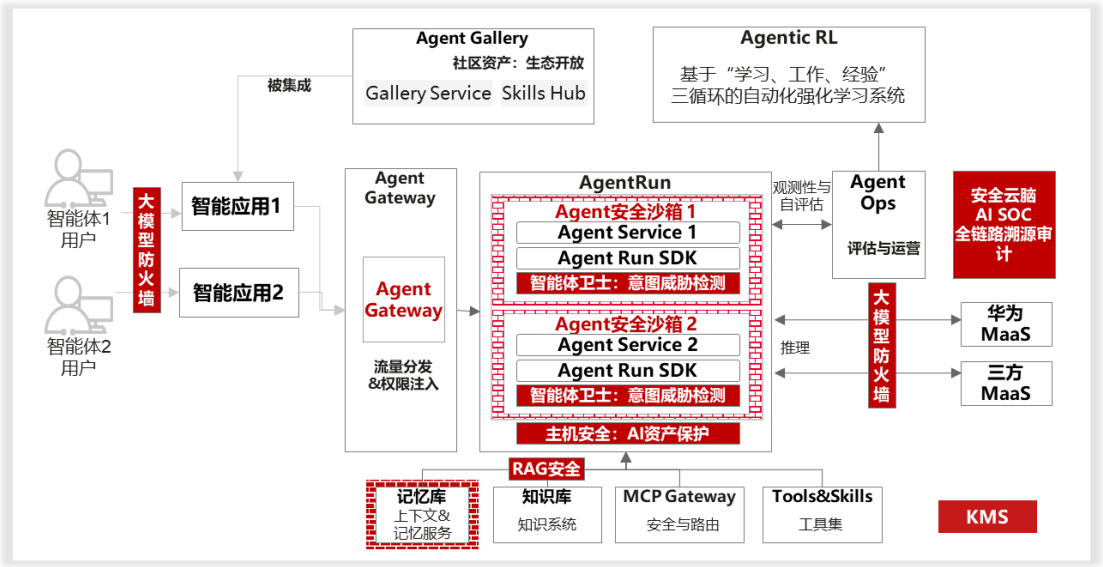
为应对智能体在自主决策与执行过程中可能引发的行为失控、越权操作及责任追溯难题，企业需构建覆盖指令输入、工具调用、环境感知、结果输出的全量日志采集体系。通过行为基线建模与异常操作（如权限逃逸、恶意工具调用）实时告警，实现问题快速定位与责任追溯，保障业务合规性。

3.3.4 建设重点—防止智能体在生产环境失控

为确保 AI 智能体在企业环境中安全可控地运行，需构建多层次、全链路的

安全防护体系。基于智能体的实际运行环境和业务约束，需明确其角色与权限边界，实时监测并阻断异常行为。

具体建设重点包括以下四方面：



图：智能体安全解决方案

3.3.4.1 绑定用户和智能体身份，细粒度认证鉴权

在智能体启用时，每个用户与智能体必须完成唯一身份认证，并基于其“数字员工身份”进行最小权限授权。认证过程需进行多因素身份认证，确保身份不可伪造。初始化阶段，系统将根据预设的岗位角色模板（如管理员、审计员、操作员），自动分配最小必要权限集，禁止开放未授权接口或数据访问权限。通过身份与权限的绑定，可有效防止身份冒用、权限滥用及越权访问，为后续所有操作提供安全基线。

3.3.4.2 企业角色限定，按工作角色限定智能体行为基线

未来，每个智能体将具备唯一的“数字员工身份”，其日常操作需严格遵循岗位职责边界。例如，财务类智能体禁止访问核心代码库，编码类智能体不得执行资金操作。通过岗位授权机制，可有效约束智能体的自主行为，减少越权或危险操作的风险。

3.3.4.3 内核级意图威胁检测，严控 AI 智能体高危操作

智能体在自主执行任务时，可能因语义误解或恶意诱导触发数据删除、权限变更等高危操作。需构建智能体高危行为实时监测系统，出现违规/高危操作时及时拦截异常行为、转由人工审批，确保任何敏感操作均需人工授权后方可执行，形成“机主防、人主控”的协同防御机制。

3.4 智能化安全运营—多智能体全域自治，威胁自主闭环

3.4.1 智能化安全运营建设客户痛点

面向混合云 AI 基础设施、模型资产与智能体业务规模化落地场景，传统碎片化、人工化的安全运营模式，已无法适配 AI 业务高速迭代、威胁形态多样、行为动态多变的特征。从企业整体安全建设视角，当前 AI 安全运营存在四大结构性短板。

- **传统人工运营无法应对 AI 高速攻防：**当前安全运营依赖人工审计、人工研判、人工处置，效率极低。人工防御完全跟不上 AI 攻击迭代速度，出现“攻击秒级发生、处置滞后失效”的问题。
- **AI 新型攻击特征隐蔽无规则：**传统基于固定特征、黑白名单的防护机制，无法识别 AI 生成的变异型、隐匿型攻击等新型对抗行为，误报、漏报率极高。
- **静态防御无法适配动态 AI 业务迭代：**大模型持续增量微调、模型版本快速迭代、业务场景动态拓展，固定安全策略无法随模型能力、业务场景动态更新，极易产生防御盲区。
- **缺乏 AI 对抗 AI 的主动防御能力：**现有安全体系仅能实现被动拦截，无主动溯源、主动诱捕、主动对抗能力，无法预判新型 AI 攻击手法，无法提前挖掘模型隐性风险，整体处于被动挨打格局。
- **无法实现智能化闭环迭代：**攻防数据、模型风险数据、对话交互数据无法沉淀复用，安全策略无法自主学习、持续优化，无法形成“监测-研判-处置-迭代”的智能化运营闭环。

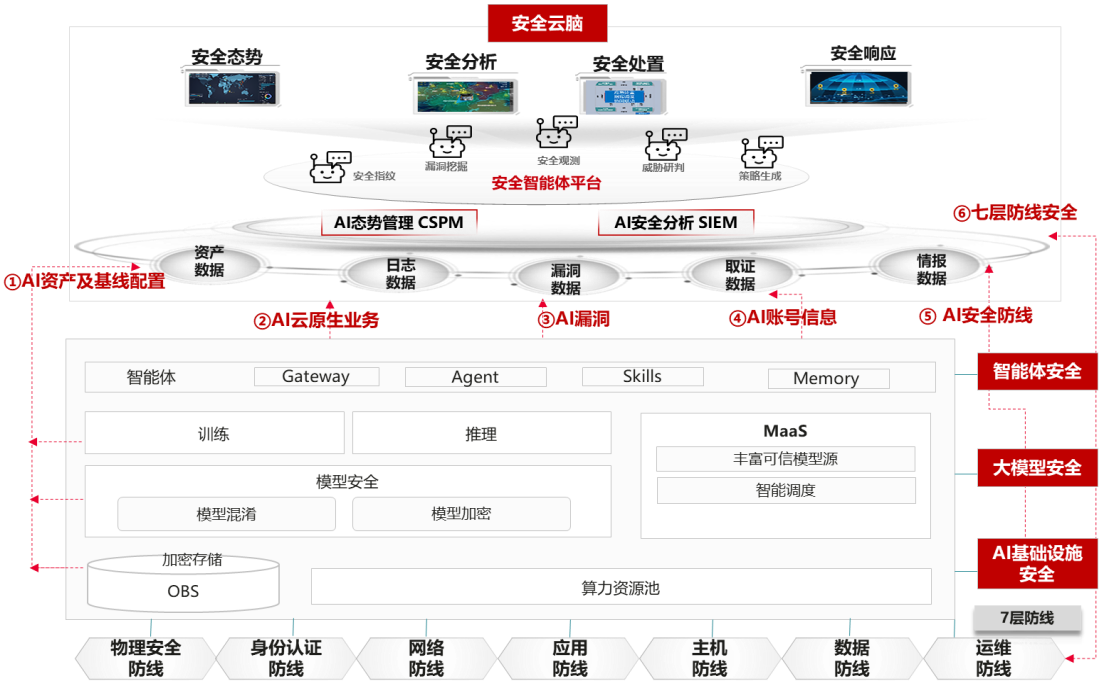
3.4.2 建设思路—从“人工运营”走向“AI 驱动的智能化安全运营”

面向 AI 时代的安全运营建设，应遵循 “全域可观测、风险可研判、处置可闭环、能力可进化” 的总体思路，通过“AI 安全态势管理 +Agentic 智能运营”双轮驱动，构建覆盖 AI 资产全生命周期的智能化安全运营体系，实现全域可观测、风险可研判、处置可闭环、能力可进化，推动安全运营从被动响应向主动防御、从人工运营向 AI 运营升级。

一、构建 AI 安全态势管理体系，实现全域风险“一屏总览”。针对 AI 资产隐蔽难管、风险分散不可见、新型攻击发现滞后等建设痛点，统一搭建 AI 场景专属可观测能力，实现基础设施、模型资产、智能体业务、数据流转的全景可视、风险可查、攻击可溯源，形成全局、实时、精准的 AI 安全态势感知能力，为安全决策、风险预警、合规治理提供统一可视底座。

二、打造多智能体协作的安全卫士，实现“以 AI 对抗 AI”。深度利用大模型原生安全能力，重组并部署安全指纹、漏洞挖掘、安全观测、威胁研判与策略生成五大专属安全智能体（Security Agents）集群，建立分布式多智能体协同网络，替代传统人工单点运营模式，实现风险自主挖掘、威胁智能研判、防护策略自适应生成、安全事件自动化闭环处置，建成以 AI 对抗 AI、以智能抵御智能的主动防御运营体系。

3.4.3 建设重点—AI 驱动的智能化安全运营



图：AI 驱动的智能化安全运营

AI 安全运营中心，针对 AI 资产与 AI 智能体应用的新型安全风险，在传统安全运营（一个中心-安全云脑+七层防线-物理防线/身份防线/网络防线/应用防线/主机防线/数据防线/运维防线：）的基础上，构建适配 AI 场景的全方位安全运营体系，打破传统安全运营仅覆盖网络、终端、系统的局限，聚焦 AI 资产专属安全特性，通过 AI 安全态势管理与 Agentic 智能体自动化分析两大核心体系进行安全技术建设和持续运营，实现 AI 资产全生命周期、多维度的安全防护与运营管控，有效解决传统防护手段对 AI 资产漏洞、训练与推理攻击、AI 专属威胁识别不足的痛点，全面升级智能化安全运营模式。

3.4.4 全域安全态势可视化， AI 资产易管控

在安全态势管理层面，通过平台实现 AI 资产全维度防护可视化能力，构建一体化 AI 安全态势感知界面，彻底解决 AI 资产风险隐蔽、不可视、难管控的问题。

（1）环境安全可视化：全面覆盖 AI 训练、部署、运行的服务器集群、主机环境、容器环境、网络边界等基础设施，实时监测环境漏洞、非法访问、资源异

常占用、边界攻击行为等风险，动态展示环境安全评分与风险分布。

（2）大模型推理安全可视化：推理业务安全聚焦于大模型服务对外提供预测/生成能力时的全链路风险管理，其核心目标是保障服务可用性、输出安全性和业务合规性。其核心风险场景包括：提示词注入攻击、数据泄露、生成虚假信息/恶意代码、API 高频调用导致服务拒绝、生成内容违反地域法律等。

（3）大模型训练安全可视化：大模型依赖训练数据，可能生成虚假或意识形态偏差内容，需通过安全测试管控风险，保障输出安全。针对大模型训练与推理过程中所使用的语料数据集（包括原始数据、标注数据、增强数据等），通过系统化的风险管理手段，确保数据全生命周期的安全性、合规性和可用性。语料数据中的敏感信息泄露、版权侵权、偏见注入、毒化攻击、数据污染等威胁，通过接入对日志信息、新增敏感数据标记、内容风险扫描等形成相关告警；对接云 Web 应用防火墙过滤恶意爬虫窃取语料的行为、自动化识别语料来源。

（4）AI 资产漏洞及攻击风险可视化：自动梳理 AI 模型、AI 智能体、插件、向量数据库等资产台账，实时汇总各类漏洞风险、外部攻击溯源、高危威胁告警，以态势图表、事件轨迹图等形式直观展示，须实时掌握整体 AI 安全态势，实现风险早发现、早预警、早处置。

3.4.5 多安全智能体协同，风险自主发现与闭环处置

在智能分析防护层面，安全运营中心深度落地 IPDRR（识别、防护、检测、响应、恢复）安全运营核心理念，针对 AI 资产与 Agent 应用的新型安全风险，通过五大安全运营智能体（安全指纹智能体、漏洞挖掘智能体、安全观测智能体、威胁研判智能体与策略生成智能体），构建起自动化、智能化、自主化的 AI 安全分析与防护体系，全面解决传统人工运营效率低、盲区多、响应慢、处置弱的短板，实现从安全左移前置防控到安全风险闭环修复的全链条防护。

（1）安全指纹智能体：针对 IPDRR 识别环节，能够自动遍历全网 AI 资产，精准识别大模型、智能体程序、向量数据库、AI 插件、训练推理服务等各类资产，自动采集资产属性、运行状态、权限配置、接口调用特征，建立统一规范的 AI 资

产指纹台账，彻底肃清隐形资产、闲置资产、未备案资产带来的管理盲区，为后续安全防护筑牢资产底数基础。

（2）漏洞挖掘智能体：坚守安全左移理念，聚焦 AI 场景特有风险，针对模型代码缺陷、Agent 代码缺陷、训练数据漏洞、插件后门、推理权限漏洞、模型投毒隐患等开展常态化自动化检测，在 AI 开发、训练、部署前期提前挖掘潜在风险，从源头降低 AI 资产安全隐患。

（3）安全观测智能体：承担全天候监测职责，7×24 小时不间断捕捉 AI 资产运行行为、模型推理记录、工具调用轨迹、数据流转路径，实时监测提示注入、越权操作、数据泄露、模型越狱等异常行为，全方位感知 AI 场景动态风险。

（4）威胁研判智能体：依托 AI 大模型和大数据分析能力，对海量监测告警进行聚合去重、关联分析、溯源定级，精准区分误报、低危告警、中危告警与高危威胁，清晰梳理攻击链路与风险根源，为安全处置提供精准依据。

（5）策略生成智能体：聚焦风险处置与闭环防护，结合研判后的安全风险，自动适配生成合规、可落地的安全防护策略、权限整改规则、风险修复方案与应急处置策略，同时支持策略动态更新与一键下发。

五大安全运营智能体各司其职、协同联动，完整覆盖 IPDRR 安全运营的全流程，贯穿 AI 资产开发、训练、部署、运行、迭代全生命周期，真正实现安全风险事前预防、事中监测、事后修复的一体化安全管控，全面、高效守护 AI 资产及各类智能体应用的运行安全与业务安全。

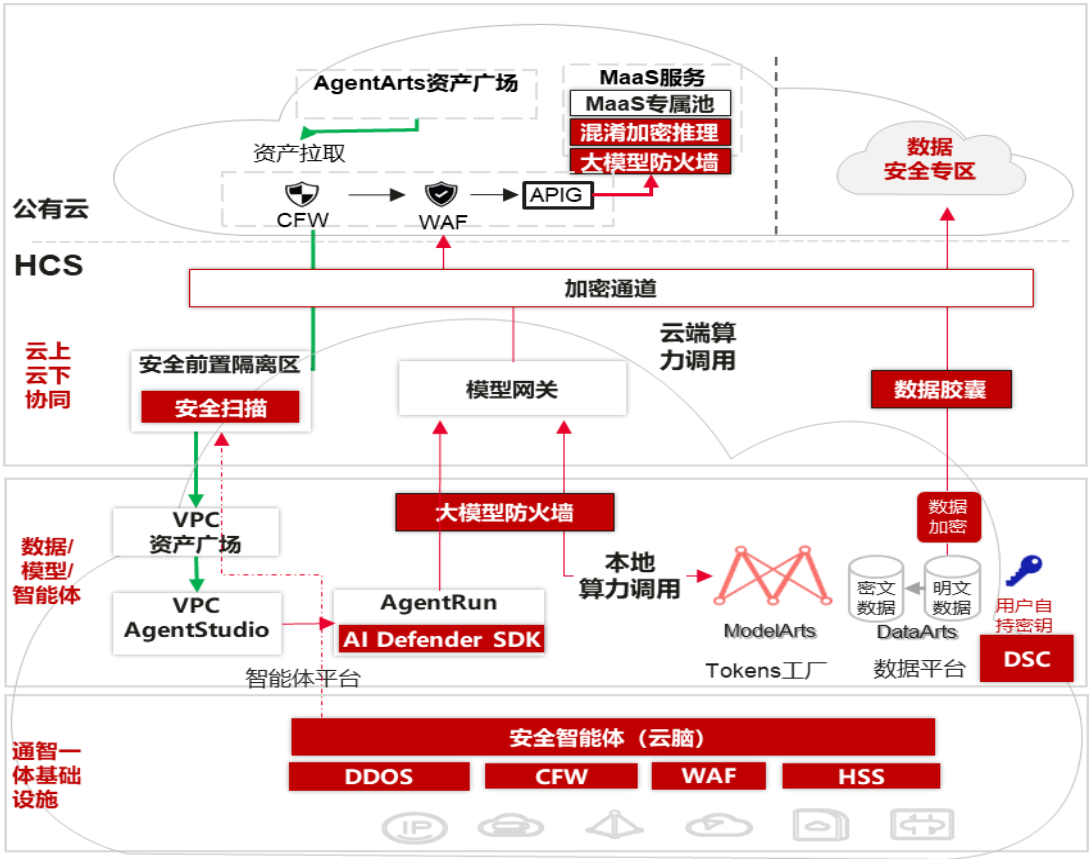
4.典型场景化应用

4.1 混合云场景下智能体安全防护

4.1.1 安全需求

某大型集团性企业于 2025 年中启动“某运营助手”项目，计划在混合云架构下部署三类智能体：办公智能体（辅助集团总部员工处理邮件、日程）、代码智能体（协助研发团队生成与审查生产调度系统代码）、数据查询智能体（对接核心生产数据库，为业务部门提供实时能效报表）。项目上线仅一个月，安全团队便发现多重隐患：

- 数据跨域泄露风险：代码智能体需调用公有云大模型进行代码补全，但在调试过程中，智能体将包含关键能源设施参数的测试数据片段误传至云端推理日志，险些酿成数据泄露事件。
- 提示词注入攻击：攻击者通过邮件附件中的恶意 Prompt，成功诱导办公智能体读取内部通讯录并外发，暴露出“自然语言即指令”的语义攻击缺口。
- 意图劫持：恶意内部人员将数据查询智能体的合法任务“查询本月各电厂发电量”拆解为“获取设施参数表结构→导出敏感坐标数据→发起远程指令修改阈值”的步骤链，各步骤单独合规，串联后构成严重破坏。
- 资产暴露盲区：安全团队发现，超过 30%的 AI 资产未纳入统一资产清单，部分测试模型直接暴露于公网，存在被投毒或窃取的风险。



4.1.2 解决方案

针对集团混合云智能体安全管控痛点，项目落地云上云下协同安全、模型数据与智能体安全、通算一体基础设施安全三位一体纵深防护体系，适配其多云协同、权限分级、合规严控的建设需求。

一、 云上云下协同安全

全密态流转与 HYOK：所有数据（含模型文件、推理输入输出）在私有云与公有云之间端到端加密，客户通过自持密钥（HYOK）实现“专钥专用”，本地代码智能体的核心模型经混淆加密后再上云推理。

数据胶囊与离域自销毁：为每份出域数据绑定统一策略——仅限私有云+公有云可信专区使用，任务完成后或超出区域即自动销毁，杜绝数据滞留。

安全前置隔离区：所有 Skill、Agent 代码在上云前须经隔离区病毒扫描、代码审计与依赖链分析；部署专职“安全智能体”持续巡检 API 调用，主动阻断异常请求。

二、 模型、数据与智能体安全

提示词攻击防护：在推理入口部署多层过滤器——复杂度监测拦截高消耗恶意 Token，规则与机器学习模型对抗直接/间接注入；本地模型加密混淆后上云，减少 Token 成本并保护知识产权。

意图威胁检测：构建全任务链行为序列分析引擎，实时识别“读取→查询→远程指令”等恶意步骤组合，在关键操作节点（如生产数据库导出、工业控制接口调用）设置沙箱门禁，自动挂起可疑任务并通知安全中心。

企业 Agent 角色限定：严格遵循最小权限——办公智能体仅限日程与邮件分类，无权访问代码仓库；代码智能体禁止推送生产分支；数据查询智能体只能访问脱敏生产指标且不得跨部门。行为识别模型一旦检测到“文员请求编译代码”或“工程师发起远程指令修改”等偏离，立即拦截并动态降权。

三、 通算一体基础设施安全

AI 资产保护与可视化：统一 AI 资产图谱，自动纳管全部分布式模型、数据集，持续扫描后门漏洞与 CVE；联动数据流转监控，发现异常外发（如模型文件被复制至未知存储桶）即自动隔离并回收凭证。

沙箱安全环境：基于华为昇腾/鲲鹏硬件 TEE 与安全启动链，构建毫秒级轻量隔离沙箱，所有智能体代码在沙箱中执行，即使被攻破也无法影响主机系统。

基础设施综合防护：覆盖虚拟化层漏洞扫描、容器运行时防病毒、OWASP LLM Top 10 应用层防御（如不安全输出过滤），并引入安全智能体驱动的自学习威胁检测，形成闭环。

4.1.3 落地成效

一是数据流转合规可控，全密态流转+客户自持密钥彻底解决数据违规离域、明文泄露问题，完全适配《数据安全法》、等保及国资数据分级管控要求；二是智

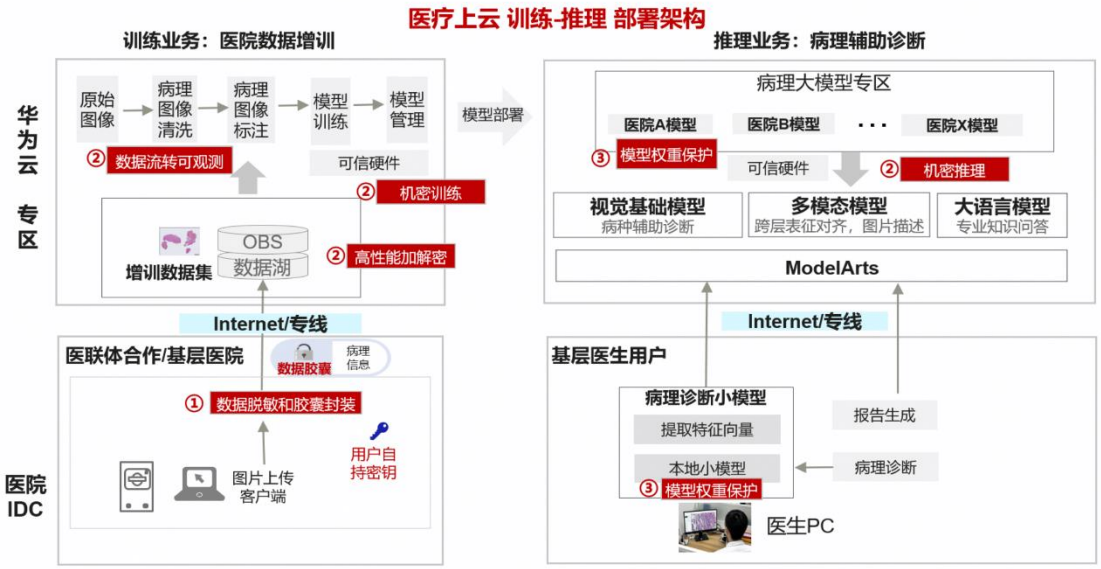
能体行为全程可防，实现提示词攻击、恶意意图全量拦截，彻底杜绝越权操作与 AI 失控风险，消除“合规内鬼”隐患；三是底层底座安全可信，实现 AI 资产可视可管、漏洞闭环处置，硬件沙箱零逃逸，基础设施高级威胁防护能力显著提升；四是安全业务兼容共生，在低时延、不影响自动化业务效率的前提下，支撑全场景智能体规模化、合规化上线运行。

4.2 医疗云上训练和推理

4.2.1 安全需求

随着数字病理与 AI 技术发展，病理大模型可显著提升阅片效率与辅助诊断能力，成为推动优质医疗资源下沉的重要基础设施。医疗病理大模型推广过程中，主要面临两大核心挑战：

- 医疗数据涉及患者隐私，必须满足“原始数据不出院”的监管要求；
- 传统 AI 病理建设成本高、周期长，基层医院难以快速落地。



4.2.2 解决方案

针对上述问题，华为云构建“数据专区 ”医疗 AI 安全方案，实现病理 AI 能力安全上云。

- 本地脱敏与胶囊封装

医院的原始病理图像（WSI 文件）通过本地预脱敏工具进行处理。经过严格的去标识化处理后，原始数据始终保留在院内。

脱敏后的数据被封装成一颗“数据胶囊”， 写入授权范围、有效期等胶囊元信息，并可通过医院自持私钥加密，实现数据全生命周期管理。

- 医疗数据专区与机密计算

处理后的胶囊数据通过加密通道传输至华为云医疗数据专区，专区与公有云环境逻辑隔离，不同医院数据独立存储、分类加密，用户可采用自持密钥加密。

模型训练与推理运行于 TEE 可信执行环境中，数据仅在加密内存中解密处理，云管理员及操作系统均无法访问，实现“数据可用不可见”。在 TEE 可信执行环境中，用户可建立全链路访问审计，整体使用过程可观测。

- 模型安全保护

医院训练出来的模型，采用 TEE 隔离加载、密钥管理、API 鉴权及访问审计等机制，防止模型被窃取或滥用；

医院本地 PC 加载的轻量化小模型则通过模型加密、代码混淆及防篡改机制保障运行安全。

4.2.3 落地成效

整套架构从数据采集、传输到 AI 运算全流程闭环防护，从技术落地层面刚性满足原始数据不出院、不外流的合规管控要求，既保障业务数据可用不可见，满足行业监管合规准则，又破除数据流通与安全管控的矛盾，支撑企业安全开展跨域 AI 训练、智能体业务落地，规避数据泄露、违规出境带来的合规处罚与资产损失。

4.3 智驾上云训练

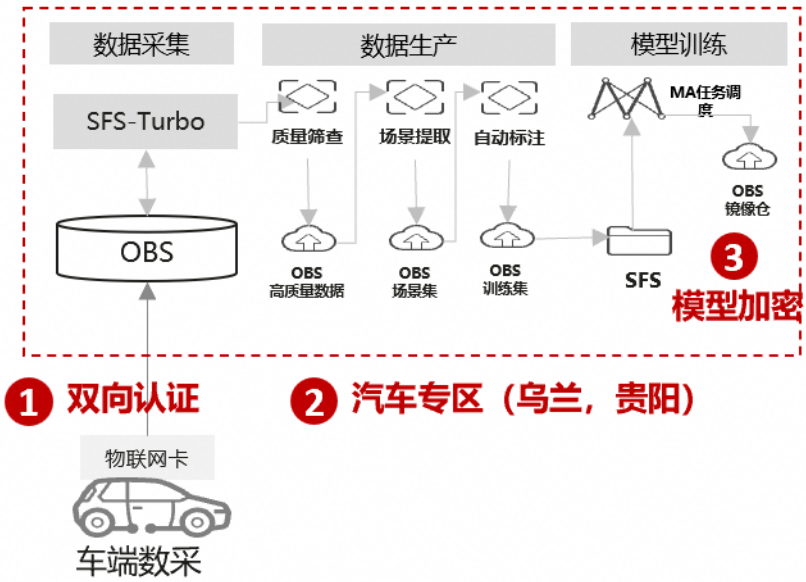
4.3.1 安全需求

当前自动驾驶正处于从“辅助驾驶（L2）”向“高阶智能驾驶（L3 及以上）”过渡的核心爆发期。一个智能驾驶模型版本从训练到发布，大致需要经过数据采集->数据生产->模型训练->仿真测试->量产运营这五大环节，高质量的数据是模型训练的重要生产原料，当前各大车企用于训练的车辆采集的数据规模已经从 TB 级跨入 PB 级，如何保障数据在智驾模型训练过程中的安全及合规的使用是车企迫切需要解决的问题。智驾训练数据主要面临以下三大核心挑战：

- 数据采集安全上云：传输过程不丢数据，存储端数据不泄露。
- 数据云上合规处理：敏感数据脱敏，数据安全合规使用。
- 模型权重文件保护：模型权重参数泄露导致商业竞争力下降，模型违规使用等风险。

4.3.2 解决方案

针对上述问题，华为云通过启用客户端证书双向认证，构建汽车专区，和针对模型文件的高性能加解密来进行应对：



- **OBS 双因子认证**

通过在车端和云端 OBS 对象存储服务启用双向证书认证以及通过启用 OBS 桶临时 AK/SK 的双因子认证访问方式来保证车上数据安全上云。即使车企内部员工或者黑客拿到 OBS 桶的访问凭据 AK/SK，由于没有客户端证书，也无法从 OBS 桶中获得汽车业务采集数据，从而可以防止人为恶意泄露汽车业务数据，避免发生汽车敏感数据泄露、避免勒索造成负面舆论。

- **华为云汽车专区**

华为云汽车专区采用 3 分区+7 层安全防护合规架构，同时引入了多家图商安全合规合作伙伴，为车企客户提供全流程的自动驾驶合规服务。

- **模型权重文件保护**

模型权重文件是车企在模型训练过程中产生的核心数据产物，记录了模型学习到的所有参数，通过对模型权重文件进行高性能加解密来保障权重文件即便被窃取外泄也不可用。

4.3.3 落地成效

满足智驾训练数据采集安全上云，并在华为云汽车专区完成数据合规处理，满足等保及相关行业政策法规的要求。训练阶段重点保护模型权重文件，即便发生外泄也不可用。

4.4 企业代码智能体防护

4.4.1 安全需求

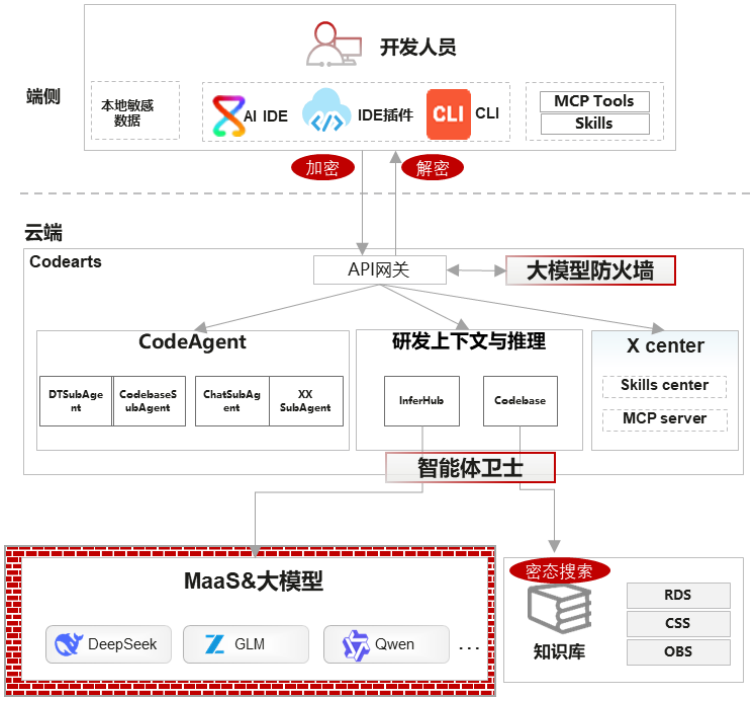
某企业推进 AI 辅助编码试点应用，以 Tokens 消耗量牵引全员用 AI，旨在提升软件开发效率。

试点过程中将代码智能体作为关键工具深度集成至开发流水线，典型使用场景如下：程序员在本地 IDE 中编写或审查代码时，通过自然语言提出需求——如“解释这段复杂逻辑”“为当前函数生成单元测试”或“定位潜在内存泄漏”。AI IDE 将用户指令、当前代码上下文及相关注释打包，提交至云端部署的 AI 辅助编码智能体，调用 MaaS 平台部署的 GLM 大模型进行编码推理，生成代码

建议、逻辑分析或缺陷修复方案，回传至 IDE 界面，供程序员逐行检查、修改并决定是否合入代码库。　　内部调研让该企业快速意识到，AI 能力的深度嵌入增加效率的同时也引入了新的安全风险。比如，若智能体被恶意引导生成含漏洞的代码、越权访问敏感数据，或执行高危操作（如删除生产环境配置），将直接威胁项目安全。

4.4.2 解决方案

为解决上述顾虑，该企业在代码辅助智能体设计与部署阶段嵌入四层防护机制，保护代码智能体不失控、企业核心代码资产不被违规访问、出现异常可回溯。



身份验证及企业角色限定：用户通过认证后与智能体进行单点登录，智能体识别用户身份（如开发工程师、测试人员），并依据其权限范围动态限制可访问的代码库、数据表及系统接口。例如，实习生智能体无法扫描核心加密模块代码，运维角色仅能在其负责的服务范围内请求生成部署脚本。

云端机密推理：公司程序员提交的提示词、代码上下文、RAG 知识库片段在云上推理时均在可信环境中执行，未经用户授权超级管理员也无法访问。生成

的代码在沙箱中运行，对生成代码中检测到的恶意行为（如系统调用、文件删除命令）进行告警。

意图威胁检测与防护：输入侧通过安全护栏监测和拦截违规/恶意指令（如“绕过登录验证生成代码”），输出侧扫描生成内容中夹带的风险行为，一旦检测到疑似危险操作（如直接操作数据库、调用未授权 API），智能体将自动阻断执行，并触发人工审批流程。

全链路行为审计：从用户输入、模型推理到代码合入，所有代码智能体交互过程均记录日志，确保行为可追溯。同时，通过 Tokens 消耗监控识别异常频次或规模的请求，防止恶意用户通过大量生成低质代码消耗计算资源，保障服务稳定性。

通过以上机制，可以确保代码辅助智能体在提升编码效率的同时，其行为始终处于权限可控、风险可管、操作可溯的框架内，为企业研发安全筑牢防线。

4.4.3 方案成效

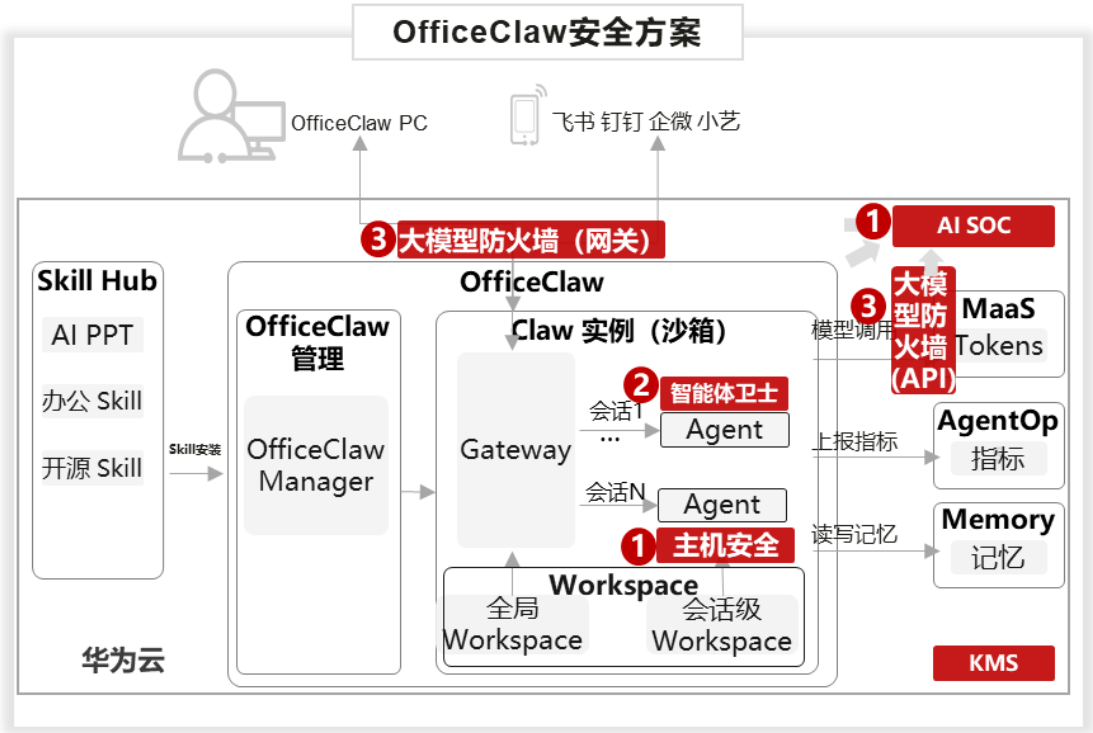
通过最小授权原则与异常行为告警，有效防范非编码类任务滥用与代码泄露风险；借助代码沙箱隔离与高危操作人工确认机制，拦截恶意攻击扩散；依托机密计算能力，确保企业代码资产与用户隐私数据全程加密。同时，基于全链路审计能力，可精准溯源违规代码生成根源（如外部攻击、模型生成异常或人为操作失误），及时阻断问题代码合入流程，全面保障智能编码助手的安全合规运行，显著提升研发流程的安全性与可控性。

4.5 企业办公助手防护

4.5.1 安全需求

某制造业用户期望引入 AI 辅助办公系统，用于替代人工重复性工作，提升员工工作效率。员工通过自然语言指令（如"生成会议纪要摘要"“提取合同关键条款”）提交处理需求，智能体可调用企业内部文档、会议录音等资源，自动生成结构化摘要或执行格式化输出。

办公智能体显著提升了信息处理效率，但同时也引入了新的安全挑战。由于办公智能体处理内容常涉及商业机密、员工个人信息等敏感数据，且部分早期高权限账户可能具备生产系统访问能力，智能体越权访问可能影响生产办公业务，需要特殊保障。



4.5.2 解决方案

针对上述问题，在云上部署办公智能体办公智能体场景，华为云构建办公智能体安全方案，对办公智能体提供资产发现和漏洞管理、办公角色行为限定、外部攻击及高危操作拦截及全链路安全监测能力，保障办公智能体行为可控、数据安全、办公业务合规稳定运行。

AI 资产和漏洞管理：对办公智能体进行全面盘点，保“数字员工”全部登记在册，识别漏洞并进行闭环管理。新增资产类型在 48 小时内完成适配更新，避免办公智能体资产遗漏、有漏洞未被保护，防止被黑客攻击成为跳板。

办公智能体角色限定：确保“数字员工”日常行为与角色权限相匹配：基于员工在企业中的身份和权限进行行为基线建模，限定办公智能体可接收和执行

的任务边界范围，避免办公智能体执行工作外的任务(如炒股等)以及其他违规任务（如访问权限外系统数据等）

异常访问数据意图威胁熔断：办公智能体接收到恶意任务或理解错用户指令、规划执行高危操作时，拦截高危操作，由办公人员最终确认是否执行；若人工确认的任务中存在不可信输入和不可信程序，数据和程序将在沙箱中运行，仅输出最终清洁数据结果给员工，避免办公智能体被攻击/诱导执行修改/删除原始数据等高危操作，规避数据损失。

全链路安全审计：监控“数字员工”执行任务过程，如果办公智能体存在异常数据权限获取/访问行为和高危操作执行任务，需要通过全链路安全监测，回溯任务来源，支撑应急响应，确保办公智能提执行恶意操作时“动手必被抓”。

4.5.3 方案成效

办公智能体安全方案可保障云上部署办公助手场景中的全链路安全管控与合规运营。以资产与漏洞管理进行初始保护，防止影子 AI 出现，避免办公智能体成为黑客跳板；构建办公领域行为基线，有效拦截非办公话题、超权限任务，防止算力浪费与数据窃取；实现高危操作拦截及人工决策介入，不可信数据和程序在安全沙箱中运行，防止办公智能体失控；全链路溯源审计，保障办公智能体行为可控、数据安全、业务合规稳定运行。

5.未来展望

面向未来，AI 安全正迎来深刻的范式变革，迈向 Security（网络/系统安全）与 Safety（物理/行为安全）内生融合、人本优先的新阶段。我们清醒地认识到："没有 Security 的 Safety 是空谈，没有 Safety 的 Security 是盲视"。为此，未来的 AI 治理将紧紧围绕六大安全治理方向，牢牢锚定技术向善，全面守护人类的主体价值：

- **能力治理升级：**针对未来高阶的自主智能与涌现能力，我们将彻底告别静态防御，构建起具备动态感知能力的新型安全管控框架，实时捕捉并化解未知风险。
- **算法边界约束：**我们将坚持立法与技术并重，不仅致力于破除信息茧房，更将通过划定硬性红线，坚决防止算法对人类理性和行为的隐性支配。
- **破解黑盒难题：**未来的智能系统将全面落地"全链路可解释 AI"，让每一次决策过程皆可溯源、结果皆可复盘，用透明度重建人机信任。
- **平衡效率与文明：**我们将跳出单一的"效率至上观"，用完善的安全机制对冲高度自动化带来的社会风险，全力护航人类文明的可持续发展。
- **严控数据滥用：**通过严格规范全景监控与跨境数据挖掘，全面应用隐私计算技术，切实保障个体的选择权，捍卫数字世界的伦理底线。
- **AIGC 全链条治理：**面对真假难辨的信息环境，我们将建立覆盖内容全流程的溯源与鉴伪体系，从源头到传播终端筑起遏制虚假信息的坚实盾牌。

未来的智能时代，安全不再是发展的绊脚石，而是文明的孵化器。通过这场全方位的治理升级，我们有信心让 AI 真正成为拓展人类福祉、而非反噬人类主体性的向善力量。