涂鸦智能 AI 安全合规白皮书

践行负责任 AI，构建可信智能生态

版本号： V1.0

发布日期： 2025.12.29

目录

前言

智能新时代的机遇与责任

我们正身处一个由人工智能驱动的万物互联智能新时代。智能语音助手、AI 驱动的视觉识别、个性化的场景推荐，正以前所未有的深度和广度融入全球用户的家庭、工作与生活。作为全球领先的 AI 云平台服务提供商，涂鸦智能致力于通过技术创新，赋能万千设备具备智能，连接世界各地的品牌、制造商、开发者和最终用户。

然而，AI 技术在提升生活便利与效率的同时，也引入了复杂且独特的挑战：**数据隐私、算法公平、系统可靠性与全球合规**。当 AI 能够“听见”我们的声音、“看见”我们的环境、“预测”我们的行为时，建立并维系全球用户对技术的信任，便不再是一种选择，而是我们业务可持续发展的基石。

浪潮之巅：全球 AI 发展态势与监管格局

AI 技术的全球竞赛已经打响，与之相伴的是日益严密和复杂的监管框架。理解这一宏观背景，是构建可信 AI 的基石。

- **技术发展态势**：
  - **融合化**：AI 与物联网、边缘计算、大数据技术深度耦合，催生出能实时感知、决策和行动的智能实体（如 AI 音箱、智能家居机器人）。
  - **普惠化**：生成式 AI、大模型技术正迅速下沉到各类终端设备，带来极致个性化体验的同时，也显著扩大了数据收集和模型推理的攻击面。
  - **边界拓展**：AI 从虚拟世界走向物理世界，其决策开始直接影响人身安全（如智能安防、老人看护）和关键基础设施。
- **全球立法与监管趋势**：
  - **欧盟《人工智能法案》**：开创性地采用“基于风险”的监管方法，对 AI 系统进行分级管控，**全面禁止**被认为具有“不可接受风险”的 AI 应用（如社会评分），并对高风险 AI 系统（如关键基础设施、医疗设备）设定了极为严格的义务。
  - **中国 AI 治理框架**：中国已出台《生成式人工智能服务管理暂行办法》等法规，

强调**包容审慎**和**分类分级**监管。要求 AI 服务提供者承担网络信息安全、数据安全和个人信息保护责任，并建立内容治理机制。

- **美国与全球其他地区：** 同时，全球各主要经济体也正积极探索和建立**本地的 AI 治理体系**。尽管在监管路径和节奏上各有侧重，但**安全、透明、公平和问责**已成为普遍遵循的核心原则，全球 AI 治理格局正呈现出多元框架并存、核心理念趋同的发展态势。

在此背景下，合规已不仅是法律要求，更是**通往全球市场的准入许可证和核心竞争优势**。

涂鸦智能：全球 AI 云平台服务提供商

- **涂鸦的 AI 战略是赋能开发者，加速 Physical AI 创新，放大平台的赋能效应。** 为此，基于 **TuyaOpen 开源开发框架和 AI Agent 开发平台**等通用引擎，为开发者提供便捷的 AI 技术与生态支持，以显著降低 AI 开发门槛，助力打造各类生活 Agent，加速 AI 技术与物理世界的融合，推动 **Physical AI** 的规模化落地。

- 涂鸦通过构建以人为本、合规透明、风险可控且负责任的 AI 系统，为 AI 硬件产品提供 AI 赋能，以期**持续为开发者输出差异化价值，并为用户打造更安全、更便捷、更智能的未来生活方式**。

涂鸦智能的 AI 安全合规愿景

涂鸦智能的 AI 安全合规愿景是成为**受信赖的智能生态构建者**。我们坚信，安全与合规并非创新的约束，而是创新的前提和核心驱动力。我们的承诺是：

- **对用户负责：** 确保用户数据得到最高级别的保护，赋予用户对其数据和设备充分的知情权与控制权。

- **对伙伴负责：** 为我们的客户（OEM/ODM、品牌方）提供安全、可靠、合规的 AI 创新产品与技术平台，助力其快速进入全球市场并最大化规避业务风险。

- **对社会负责：** 以伦理为先导，负责任地开发和部署 AI 技术，促进科技向善，防范技术滥用。

为实现这一愿景，涂鸦智能打造了‘**Titan Matrix**’一体化智能安全品牌，贯穿 AI 业务全生命周期，为全球开发者与用户提供可靠的安全基座。

涂鸦Titan Matrix提供安全保障

全面推进系统化、透明化、可核查、可审计的安全防护矩阵

AI安全生态

研发安全：
- 安全检测工具（SAST/DAST/IAST）
- 流程与安全运营
- 安全合规白皮书

安全中心：
- 安全开发组件与服务
- 商业安全 & 风险管控
- 数据安全 & 终端安全

防护系统：
- 网络安全防护能力（NGFW，WAF，RSAP）
- 云原生安全防护能力（CNAPP，CSPM，Runtime Security，Container Security）
- 运营安全（SOC，SIEM，SOAR）
- IT安全（NGFW，SASE，XDR）

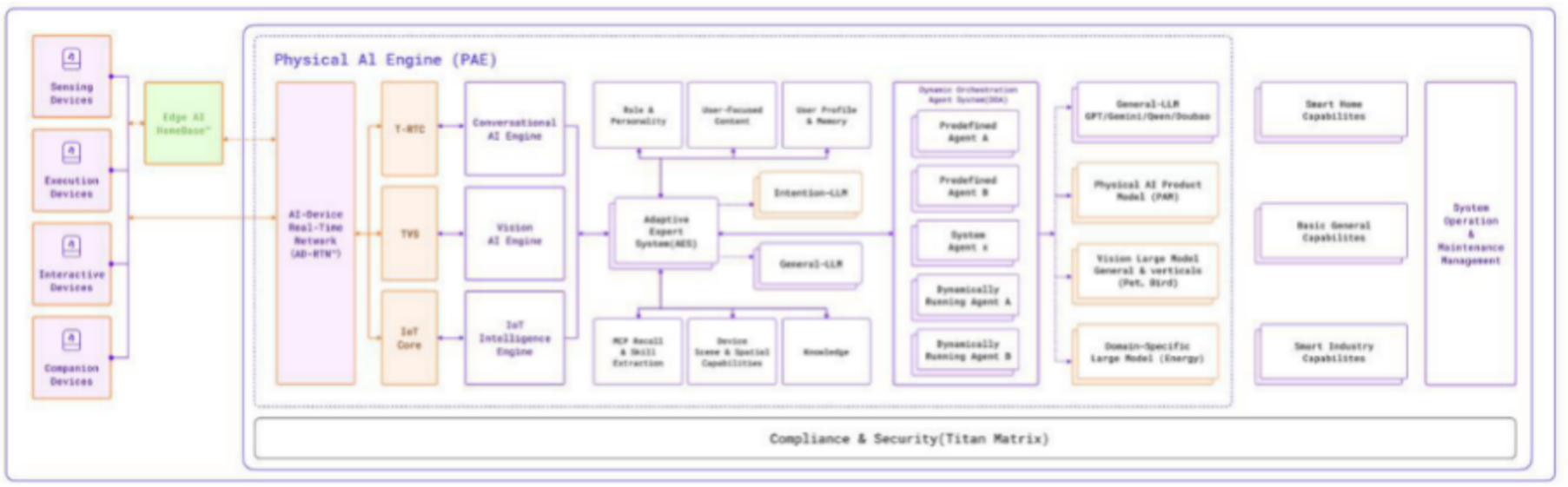
AI安全：
- 大模型安全 & 内容安全
- AI Agent & MCP Server 安全管控
- AI安全运营

隐私 & 合规：
- 安全认证（ISO、SOC、GDPR）
- 合规运营

第一章：涂鸦智能 AI 全景与负责任 AI 框架

1.1 涂鸦 AI 业务全景图

基于涂鸦智能深厚的 IoT 与 AI 融合能力，涂鸦 AI 致力于打造一个无缝融入现实生活的智能生态系统。我们的核心是“AI 生活助手（Your AI Life Assistant）”——以 Hey Tuya 为统一对话入口，打破设备与场景的边界，实现“多端协同，一呼即应”。它不仅是语音与文字的交互界面，更是一个拥有记忆与理解能力的超级 Agent，能够深入家庭的安全、节能、健康、陪伴与效率等核心场景，提供从安防守护、能耗管理到健康建议、办公提效的全场景主动服务。



为实现这一愿景，我们构建了从硬件创新到系统服务的完整技术栈。面向开发者，我们提供AI 硬件创新开发平台，通过开放的 TuyaOS 与 TuyaOpen，赋能全球开发者快速打造 AI 耳机、AI 学习机、AI 机器人等创新硬件。在系统层，我们自主研发的 Physical AI Engine（PAE）智能引擎，集成了设备实时网络、音视频通信、IoT 核心连接、对话与视觉 AI 引擎，并通过自适应专家系统与智能体编排平台，灵活调度行业大模型与专属技能，让 AI 不仅能听懂、看懂，更能主动理解用户意图，调度合适的服务。我们正从“对话式智能”迈向“场景主动服务”，致力于成为每个家庭中真正懂你的“实体贾维斯”，让智能从连接走向思考，从执行命令进化为预见需求。

核心模块和组件如下：

- AI 基座(包含 Conversational AI Engine、Vision AI Engine、AES、DOA)：低时延、高稳定、高自主的软硬一体的开发者 AI 解决方案，提供了统一的模型生命周期管理与基础 Agent 服务。包含了终端 SDK、推流传输服务、云 AI 模块和开发者平台。

- **AI Agent 开发平台**：集成全球主流的大语言模型，为开发者提供高效而灵活的智能体管理功能。开发者可以通过配置和调试，轻松部署和运行智能体相关应用。

- **TuyaOpen**：TuyaOpen 是一款前沿的开源 AI 与物联网开发框架，能够助您快速打造智能互联产品。该框架支持多种芯片平台及类 RTOS 操作系统，可轻松集成大型语言模型与多模态 AI 功能——包括语音、视觉及传感器处理。为您的硬件注入新一代人工智能的强大动力。当前基于 TuyaOpen 实现的 AI 智能终端，包括了 AI 玩具、AI 摄像头、AI 门锁、AI 数码相框、AI 学习机、AI 智慧屏、AI 手表。

- **Hey Tuya（Tuya 超级 AI 助手应用）**：Hey Tuya 是涂鸦智能推出的，基于涂鸦强大的 IoT 与 AI 融合能力打造的超级生活 AI 助手。TA 是家庭的守护者、生活的规划师、工作的搭档。Hey Tuya 通过与你对话、理解你的需要、记住你的习惯，并与各种物理 AI（Physical AI）设备协同工作，带来自然、主动、智能的陪伴。

1.2 Tuya Titan Matrix：全生命周期治理与智能防御体系

涂鸦智能构建了融合 AI 安全、隐私保护与合规治理的一体化可信架构“**TitanMatrix**”。该架构融合 AI 安全、隐私保护与合规治理于一体，通过‘**内生安全、合规驱动、智能防御、持续改进**’四大支柱，确保 AI 系统的安全、可靠与合规。

1.2.1 内生安全：产品级安全隐私防护

我们将安全与隐私要求深度融入 AI 产品研发流程，实施覆盖数据、算法、系统的端到端防护：

- **数据安全**：通过数据分类分级、匿名化处理、加密传输与存储，确保训练数据与应用数据全流程受控。
- **算法可靠**：建立模型风险评估、公平性测试、对抗样本防御机制，保障 AI 决策的可解释性与鲁棒性。
- **系统安全**：强化 AI 服务接口鉴权、推理环境隔离等基础设施安全控制。

1.2.2 合规驱动：全球化 AI 合规体系

我们建立了以法规遵从为导向的 AI 合规治理框架，确保产品与服务符合目标市场要求：

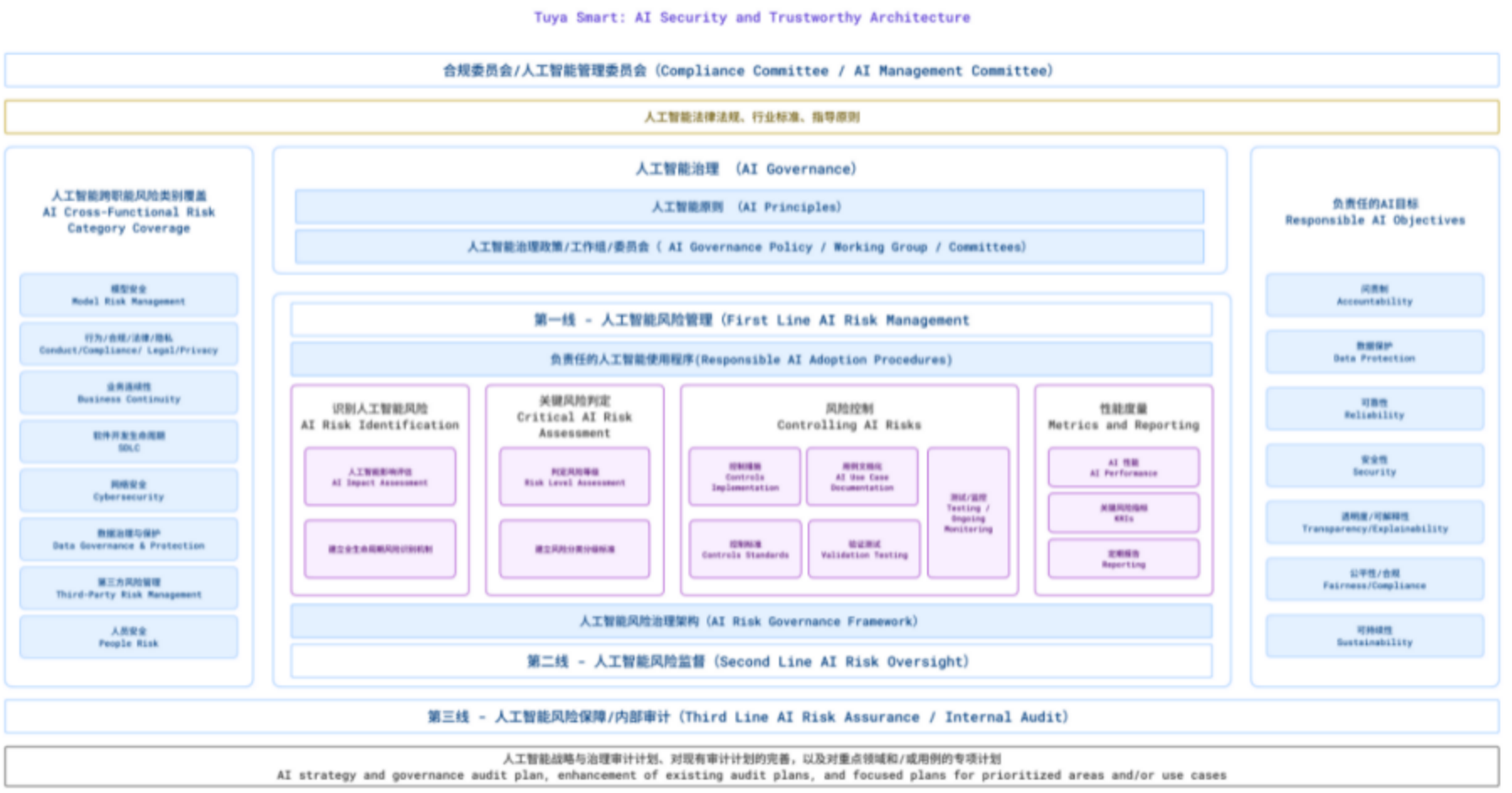
- **法规映射**：系统识别并落地欧盟《AI 法案》、GDPR、中国《生成式 AI 服务管理暂行办法》等关键法规要求。
- **合规嵌入**：将数据主体权利保障、算法备案、高风险 AI 系统监管等合规要求转化为可执行的产品控制点。
- **审计就绪**：维护完整的 AI 系统文档、数据处理记录与合规证据链，支持内外审计与监管问询。

1.2.3 智能防御：AI 赋能的安全运营

我们创新性地利用 AI 技术增强整体安全防御能力，实现安全运营的智能化升级：

- **智能威胁检测**：应用机器学习算法分析网络流量、用户行为与系统日志，实时识别高级持续威胁与零日攻击。
- **自动化响应**：构建基于 AI 的安全编排与自动化响应（SOAR）流程，实现安全事件的快速研判与处置。
- **预测性防护**：通过大数据分析与威胁情报挖掘，预测潜在攻击路径并实施前瞻性防护策略。

1.2.4 持续改进：可信 AI 治理生态



我们建立了 AI 安全可信的持续改进机制：

- **责任体系**：明确 AI 系统所有者、开发者、运营者的安全责任边界。
- **风险评估机制**：针对高风险 AI 应用，我们通过合规委员会主导进行 AI 影响评估和信息安全风险评估，确保在设计和部署阶段识别并管控潜在风险。
- **透明度建设**：通过模型水印、模型全链路日志监控、AI 白皮书或报告、开发者文档和用户告知等方式增强 AI 系统透明度。

该架构形成了“**以合规为底线、以安全为基石、以智能为引擎**”的 AI 可信体系，既满足当前监管要求，又为未来 AI 技术演进预留了安全扩展能力，为涂鸦智能的全球化 AI 业务提供了坚实可信的基础支撑。

1.3 我们的负责任 AI 核心原则

核心安全原则：PREPARE 框架

为实现这一愿景，涂鸦智能在整个 AI 产品生命周期中，严格遵循以下 **PREPARE** 核心安全原则：

- **P - Privacy by Design（隐私保护设计）**： 从产品构思之初，隐私保护便嵌入设计与架构的每一个环节，而非事后补救。
- **R - Resilience & Robustness（韧性性与鲁棒性）**： 确保 AI 系统能够抵御对抗性攻击、数据投毒等威胁，并在异常情况下保持核心功能的稳定与可靠。

- **E - Equity & Fairness（公平与无偏）：** 通过多样化的数据集和持续的偏差检测，致力于消除算法偏见，确保 AI 决策的公平与公正。

- **P - Proportionality & Minimalism（适度与最小化）：** 严格遵循数据收集的“最小必要”原则，仅处理实现特定目的所必需的最少数据。

- **A - Accountability & Governance（问责与治理）：** 建立清晰的 AI 安全责任制与治理结构，确保所有 AI 活动可追溯、可审计、可问责。

- **R - Reliability & Safety（可靠与安全）：** 确保 AI 功能，尤其是在控制物理设备（如电工照明、家用电器等）时，不会因错误或恶意指令导致人身或财产风险。

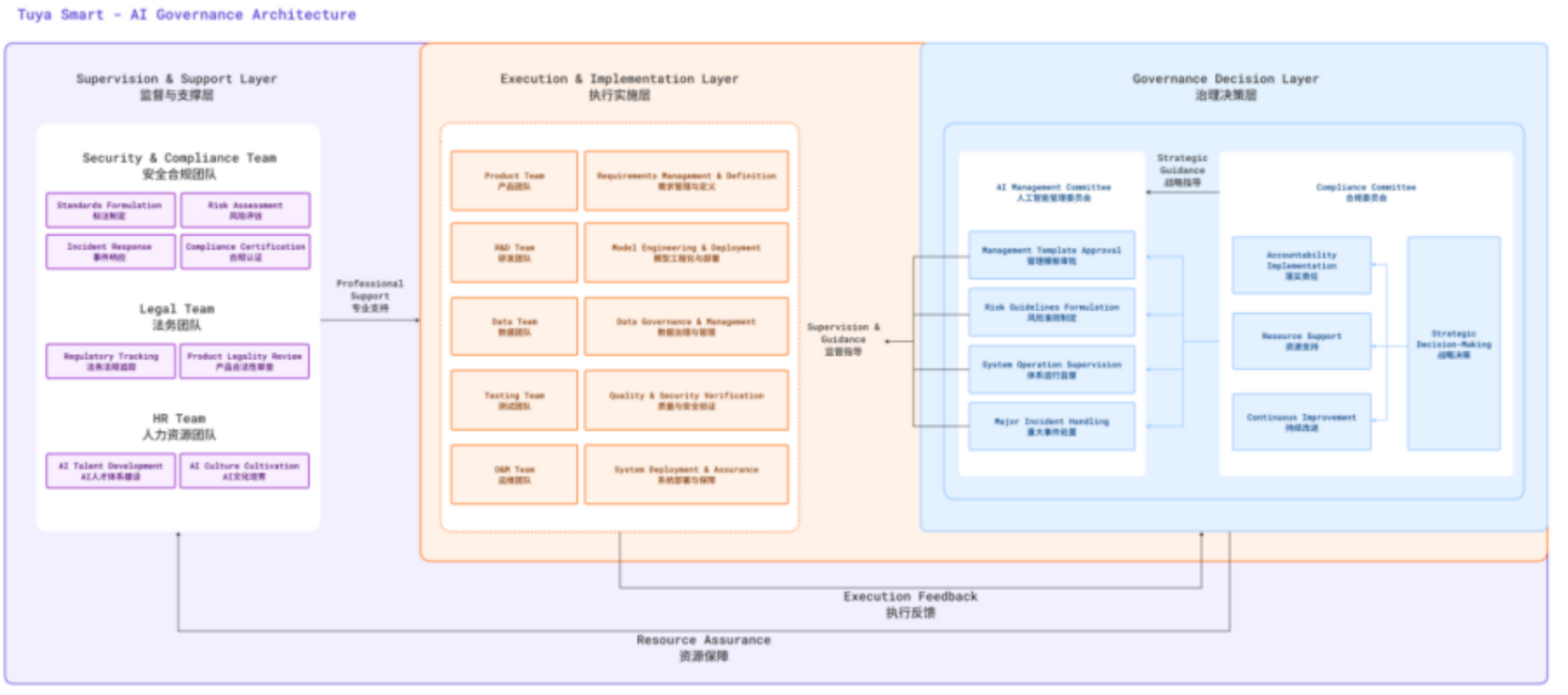
- **E - Explainability & Transparency（可解释与透明）：** 在技术可行的范围内，力求 AI 决策过程对用户透明，并提供清晰、易懂的用户告知。

第二章：坚实的 AI 治理与组织保障

涂鸦的人工智能管理体系是基于欧盟人工智能法案、人工智能相关的国际标准，结合涂鸦在云计算与 AIoT 领域的战略定位而建立。其根本宗旨在于系统化地管理人工智能相关机遇与风险，确保人工智能技术的应用与开发符合“创新引领、促进 AI 安全可信；合法合规使用 AI，确保公平、透明和可解释性”的方针，支撑集团 Physical AI 战略的稳健落地。

2.1 人工智能风险治理架构

涂鸦智能建立了具备明确职责分工及配套政策/流程的人工智能风险治理架构，并配备高效运作的人工智能风险管理体系，以确保全球化 AI 产品与服务的安全、可靠、合规，赢得全球客户与终端用户的信任。



1. 治理决策层：战略引领与顶层监督

该层级是 AI 治理的最终责任主体，负责战略决策与资源保障。

- 合规委员会：作为最高决策机构，确保 AI 方针与公司战略一致，审批关键资源投入，并对重大 AI 风险决策负责。
- 人工智能管理委员会：作为跨职能的管理核心，负责审批 AI 管理目标、风险准则，并监督整个 AI 风险管理体系的有效运行。

角色	职责描述
合规委员会	涂鸦合规委员是人工智能治理的最终责任主体，通过以下行动展现其对人工智能管理体系的领导作用与承诺： ● **确保方向一致**：确保人工智能方针与战略目标持续与集团 AI 战略方向保持一致 ● **落实责任整合**：将人工智能管理要求全面融入集团的组织架构、业务流程及技术实践中，确保治理与实践的一体化 ● **保障资源供给**：确保为建立、实施、维护和持续改进人工智能管理体系提供所必需的资金、基础设施及人力资源，特别是保障平台侧 AI 安全与合规能力的建设 ● **驱动持续改进**：主持管理评审，推动人工智能管理体系有效性的持续改进，并对关键人工智能风险的处置与接受做出决策
人工智能管理委员会	涂鸦设立跨职能的安全合规委员会，作为人工智能治理的管理与协调组织，其主要职责包括： ● 审议并批准本办法的配套实施细则与标准 ● 审批集团级 AI 管理目标与风险接受准则 ● 监督 AI 影响评估与风险管理活动的有效性 ● 协调处理重大 AI 事件与争议

2. 执行与实施层：产品研发与落地

该层级由各业务与技术部门构成，负责将 AI 产品从概念推向市场，是 AI 治理要求的直接实践者。

- 产品、研发、数据、测试、运维团队各司其职，形成了从需求提出、模型开发、数据管理、质量测试到系统部署上线的完整闭环，共同保障 AI 系统在技术上的可靠性、安全性与性能。

3. 监督与支持层：合规保障与体系建设

该层级为 AI 治理提供专业的支持、监督与保障，确保公司行为符合内外部规范。

- 安全合规团队是核心监督力量，负责制定安全标准、执行风险评估与测试，并主导事件响应与合规认证。
- 法务团队负责追踪法律法规，确保 AI 产品的合法性。
- 人力资源团队负责构建 AI 人才体系，培育公司的 AI 文化，为治理架构提供人才支撑。

2.2 AI 管理体系认证

杭州涂鸦获颁国际权威认证机构 DNV 的 ISO/IEC 42001:2023 人工智能管理体系认证证书，成为全球首批获此证书的 AI 平台型企业该认证标志着涂鸦智能在 AI 研发、应用和管理的全生命周期中已建立起**系统化、规范化的治理框架**，为全球客户提供更安全、可靠、可信赖的 AI 产品和服务奠定了坚实基础。

ISO/IEC 42001 由国际标准化组织（ISO）和国际电工委员会（IEC）联合制定，旨在帮助组织在开发、实施和维护 AI 技术时建立有效的管理体系，涵盖了 **AI 系统的规划和战略、AI 项目管理、AI 需求和设计、AI 实施和操作、AI 监测和评估**等多个方面。

在认证过程中，涂鸦智能接受了审核机构对人工智能管理体系的**全面评估**，涵盖了从 AI 战略规划、风险管理到实施运营的全流程。

评估内容显示，涂鸦智能在多个维度表现卓越：

- **治理架构**：建立了自上而下、权责清晰的 AI 治理架构，确保对 AI 活动的全面管理与有效监督。
- **生命周期管理**：执行覆盖 AI 系统全生命周期的系统性管理，从需求设计到最终退役的每一个环节均满足合规性、伦理性与安全性要求。
- **数据安全**：严格遵循数据最小化、匿名化原则，全面保障用户隐私与数据安全。
- **透明度**：在技术可行范围内，力求 AI 决策过程对用户透明，并提供清晰、易懂的用户告知。

涂鸦智能始终将安全、合规和伦理置于 AI 技术研发与应用的核心位置。

ISO/IEC 42001 是全球首个可认证的人工智能管理体系，通过此次认证不仅佐证涂鸦 AI 相关产品和服务的开发和使用是当前行业最佳实践，符合国际市场规范，也充分彰显涂鸦智能在人工智能技术上的行业引领者地位。

2.3 制度化与流程化

健全的管理体系是 AI 技术可持续发展的基石，涂鸦率先在行业内构建起**完整的 AI 安全合规管理制度体系**，通过制度与流程化的严格管控，将 AI 治理理念全面融入组织架构和业务

流程，为全球用户提供安全、可靠、可信赖的 AI 产品与服务。

构建三层治理架构，明确权责划分

涂鸦智能以《涂鸦人工智能管理章程》为顶层设计，确立了**自上而下、权责清晰的三层 AI 治理架构**。治理决策层负责战略引领与顶层监督，执行与实施层专注产品研发与落地，监督与支持层提供合规保障与体系建设。这一架构确保了 AI 治理要求在全组织范围内的有效贯彻，实现了从战略到执行的无缝衔接。

建立全生命周期管理体系，贯穿 AI 产品始终

依据《AI 系统全生命周期策略规范》，涂鸦智能将隐私保护、安全性和伦理考量嵌入从概念设计到系统退役的每一个环节。我们坚持“**设计即安全、设计即合规**”的理念，在规划阶段即进行初步风险评估，明确项目的合法性、正当性与必要性；在开发阶段严格遵循安全编码规范；在部署前执行全面的影响评估；在运营阶段实施持续监控；在退役阶段确保平稳有序下线。这一全过程管理确保了 AI 产品质量与风险的全周期可控。

实施分级风险评估，精准防控 AI 风险

《AI 系统评估和风险评估管理制度》为涂鸦智能提供了系统性的风险评估框架。我们要求在所有 AI 系统的关键阶段进行评估，包括新系统引入、重大变更、场景拓展等情形。通过建立**风险等级判定标准**，针对不同级别的风险采取相应的缓解措施，有效识别并防范 AI 偏见、安全和个人信息风险，确保在技术创新与风险防控之间取得最佳平衡。

强化数据源头治理，保障训练数据质量

数据是 AI 的“燃料”，其质量直接决定模型的可靠性与公平性。涂鸦平台上的所有业务和服务严禁使用开发者或用户未授权的数据进行训练，同时，《AI 系统训练数据管理规范》对数据的采集、清洗、标注、存储、使用和销毁全过程制定了明确要求。我们特别强调**数据来源的合法性、最小必要性原则**，通过严格的数据质量管理流程，从源头减少数据偏见和安全风险，为模型训练提供高质量的“营养”。

完善应急响应机制，提升安全事件处置能力

《人工智能系统安全事件管理办法》建立了一套规范、高效的 AI 安全事件应急响应流程。该制度明确了安全事件的定义、分类分级标准、上报路径和处置流程，确保在发生数据泄露、模型滥用、系统攻击等事件时，能够**快速响应、最小化损失**并满足监管报告要求。通过定期演练和持续优化，不断提升组织的安全事件应对能力。

优化资源管理效能，实现可持续发展

《人工智能系统资源管理规定》对 AI 系统生命周期中所依赖的计算资源、模型资源和数据资源进行统一管理。通过规范云计算资源的申请、分配、监控与成本优化，以及模型文件的版本管理和存储规范，涂鸦智能在保障 AI 系统高效运行的同时，**避免资源浪费，控制运营成本**，实现经济效益与社会效益的统一。

严格供应商管理，构建可信 AI 生态

《人工智能供应商管理规范》将 AI 风险管理延伸至供应链环节。我们对所有提供人工智能相关产品、服务、工具、模型或解决方案的供应商实施统一的安全合规要求，确保整个 AI 生态系统的**可信度与安全性**。这一制度有效管控了外部引入的 AI 应用风险，保障了数据安全与服务质量的持续稳定。

涂鸦智能通过这七大制度的协同运作，建立了**标准化、流程化、制度化**的 AI 治理体系。该体系不仅符合 ISO/IEC 42001 国际标准要求，更在实践中不断优化完善，为公司在全球范围内开展 AI 业务提供了坚实的制度保障。

未来，涂鸦智能将继续完善 AI 治理框架，推动制度要求与业务实践的深度融合，以体系化的管理能力护航 AI 技术创新，为全球人工智能产业健康有序发展贡献涂鸦智慧。

第三章：融入血脉的 AI 安全开发生命周期

涂鸦执行所有人工智能系统实施覆盖其全部生命周期的系统性管理，从需求设计到最终退役的每一个环节均满足合规性、伦理性与安全性要求。

涂鸦的 AI 应用开发生命安全周期，是‘Titan Matrix’研发安全管理支柱的实践核心。其全面涵盖了系统开发生命周期的各个阶段。并且通过安全管理平台进行统一的项目 SDLC 实施监控和管理，基本实现全自动化的流程跟踪和自动化安全评级。

3.1 规划与设计阶段

● **需求定义和边界控制**：在 PRD 中明确定义业务范围、系统边界和设计运行域，详细规定功能、性能、安全、合规及伦理需求。通过清晰界定适用场景与不适用场景，从源头控制 AI 系统的操作边界，防止范围蔓延和场景滥用风险。同时列明技术选型、资源预算等约束条件，为后续开发建立明确框架。

● **风险评估**：联合法务、合规、安全、技术部门开展 AI 系统影响评估，系统识别高风险场景并进行评级。重点分析涉及个人信息处理、自动化决策、生成式 AI 应用等场景对个人权利和社会公共利益的潜在影响，确保在需求阶段即识别并评估合规与伦理风险的

性质、可能性与严重程度。

- **技术可行性评估**：从数据、模型、工程运维及供应链四个维度进行技术可行性分析。评估训练数据的可及性、合法性与质量，算法选型的合理性与合规实现路径，系统集成的技术复杂度，以及对第三方组件的依赖风险。通过全面技术评估识别主要技术风险并提出缓解策略。

- **平台流程支撑**：基于项目的应用场景和合法依据，通过内部的安全运营平台和合规管理平台，以及安全合规部门持续的大模型与大模型业务的风险知识沉淀，提供半自动化的项目安全与隐私合规评估，确保在规划与设计阶段能够最大化识别可能存在的风险。

3.2 数据准备与模型开发阶段

- **数据管理**：遵循《涂鸦 AI 系统训练数据管理规范》，确保用于训练和测试的数据集在采集、标注、存储和处理过程中符合数据安全与隐私保护规定，提升数据的质量、代表性和多样性，有效识别并缓解数据偏见。同时，在涉及处理个人数据的 AI 相关的业务中，严格践行数据最小化、匿名化原则。

- **模型设计与训练**：采用适当的技术与架构，致力于提升模型的稳健性、安全性和可解释性。开发过程中持续评估并记录模型的性能与潜在偏差。

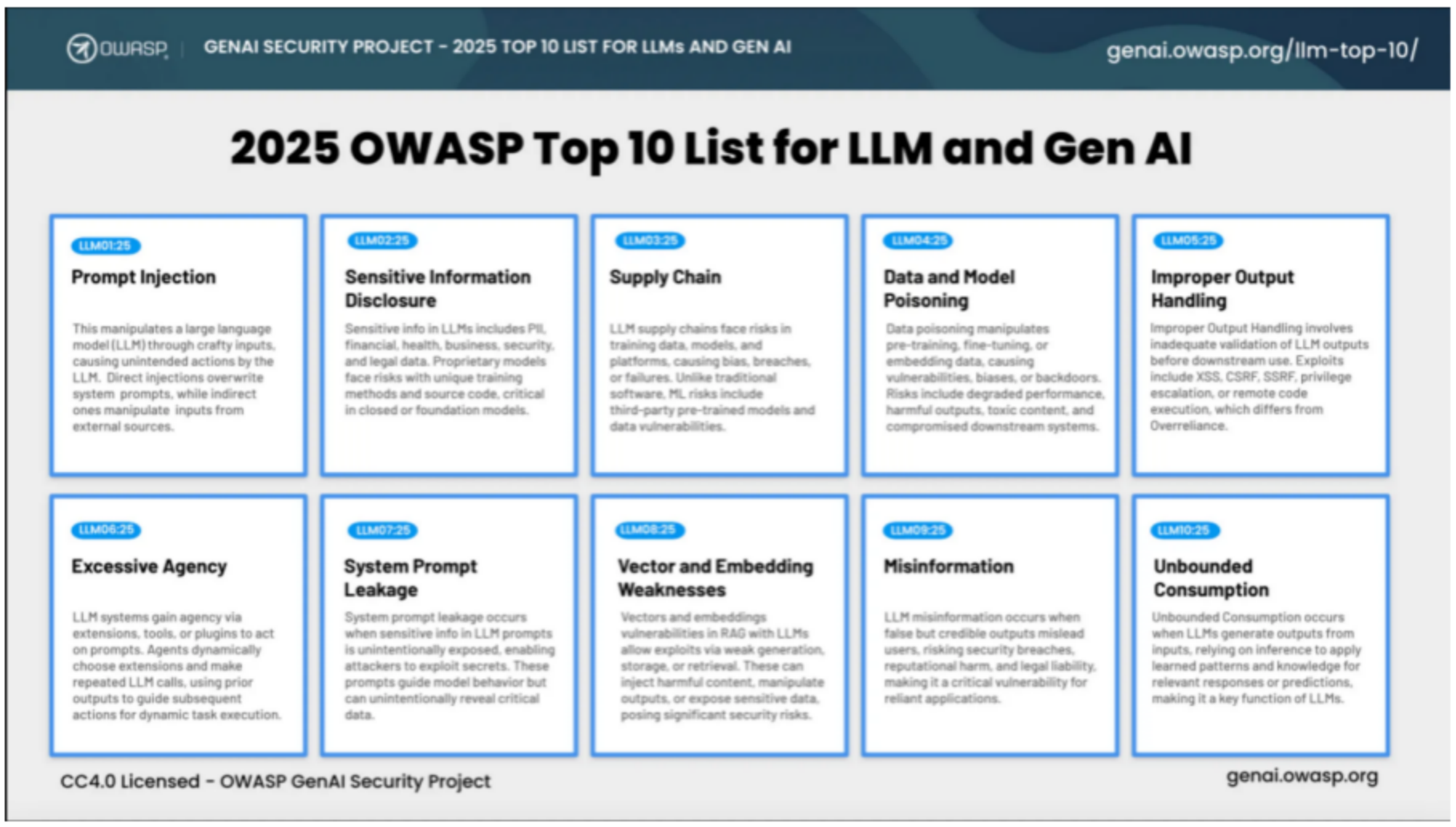
- **供应链安全合规管理**：依据《涂鸦人工智能供应商管理规范》，对使用的模型、API 进行安全评估。

- **开发培训和教育**：为确保开发人员充分理解并遵循《涂鸦 AI 安全编码规范》，涂鸦安全团队提供了全面的培训和教育课程。这些课程涵盖了安全编码的基本概念、最佳实践以及如何在具体项目中应用这些知识。

- **模型开发**：研发团队严格遵守《涂鸦 AI 安全编码规范》，基于风险场景实施安全部门提供的专用解决方案，比如在模型的所有输入和输出场景中需要严格接入安全部门提供的 AI 安全护栏的 AI 注入攻击检测和内容安全合规研发组件 SDK，除此外，安全部门提供了各种安全漏洞防护组件和安全能力组件。

● **测试与验证**：在模拟环境及受控的真实环境中进行充分测试，验证其在预设及边缘场景（如物联网设备面临的网络不稳定、数据异常等情况）下的表现。

● **安全对抗性攻击防护**：研发团队和测试团队需要基于安全团队提供的工具和用例进行 AI 服务和相关接口的越权测试，然后安全团队基于 OWASP Top 10 for LLMs 标准通过持续的沉淀 AI 安全测试用例，执行自动化安全扫描和审计，包括但不限于供应链安全测试和代码审计等，以及专家手工的安全评估，持续性校验大模型的安全性。



3.3 部署与上线阶段

● 部署前，依据《涂鸦人工智能系统影响评估和风险评估管理制度》执行全面的**影响**

评估，系统性地分析其对个人、群体、社会及环境的潜在后果，并记录评估结果与风险处置措施。

- 部署环境需要经过最后的**安全审计验证**，包括部署的主机和镜像都安全性验证，以及其他的基线安全，包括配置、版本漏洞、合规基线等检测，确保环境安全基线满足发布要求。

- 基于安全平台和合规平台中的评审和测试报告的结论，对项目的部署和上线进行最终决策，通过评审和测试后，才能进行发布和部署。

- 部署后，有专门的**人类监督角色与介入机制**，确保在关键决策或系统异常时，人类能够有效覆盖、中断或修正 AI 的输出。

3.4 运营与监控阶段

- 涂鸦的**大模型风险监控体系**，利用可观测的监控工具，实时监测模型的预测漂移、性能衰减及异常行为。

- 涂鸦也部署了**大模型安全漏洞扫描工具**，对大模型、AI Agent、MCP 服务等相关基础设施和接口进行全天候的安全扫描和评估。

● 涂鸦拥有**完善的日志记录机制**，详细记录了 AI 系统的关键输入、决策过程及输出结果可审计、可追溯，以满足故障排查、客户质询与合规审查的需要。

● 通过制定并**演练应急预案**，能够及时响应和处理 AI 系统相关事件。

● 涂鸦开放的**安全生态**，通过安全响应中心（SRC）接收白帽子对平台的安全与合规性风险的提报和提供悬赏，同时，定期执行第三方安全公司对安全服务和能力进行安全评估以及安全部门内部针对 AI 场景和业务的攻防演练对抗。为了推进 AI 安全生态建设，涂鸦 SRC 针对 AI 相关服务和应用，执行两倍赏金奖励。

3.5 退役与归档阶段

- 制定明确的**系统退役计划**，确保服务平稳、有序下线，并通知相关用户与客户。

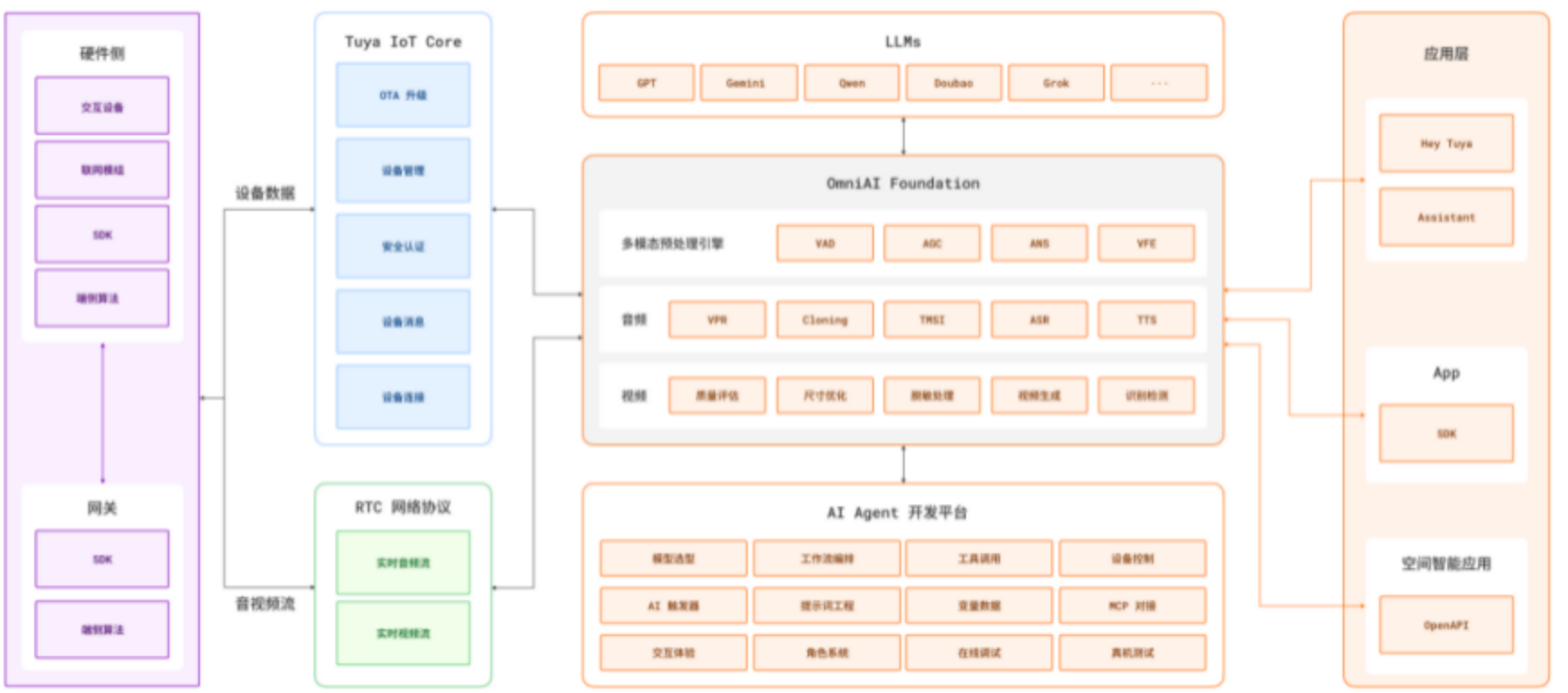
- 依据**数据保留政策**，对系统相关的模型、数据及日志进行安全处置或归档，防止遗留安全与隐私风险。

第四章：聚焦核心业务场景的 AI 安全实践

4.1 AI 云平台：安全、开放的中立生态

关键词：提示词注入、内容安全、AI Agent 安全管控、MCP 网关和数据权限、大模型攻防

涂鸦 AI 平台致力于打造互联互通的开发标准，连接品牌、OEM 厂商、开发者、零售商和各行业的智能化需求。基于全球公有云，涂鸦开发者平台实现了智慧场景和智能设备的互联互通，承载着每日数以亿计的设备请求交互。平台服务涵盖硬件开发工具、物联网云服务、智慧行业开发三方面，打造一站式产品智能化和物联网应用开发体验，为开发者提供从技术到营销渠道的全面支撑。



当前平台集成了多种语言模型，为用户提供高效而灵活的智能体管理功能。用户可以通过配置和调试，轻松部署和运行智能体相关应用。智能体能够对接插件功能允许调用各种工具或 API，例如设备查询、设备控制和场景控制等。插件扩展了智能体的能力，使其能够执行更多样化和复杂的任务。同时，智能体也支持对接涂鸦官方或客户自己的知识库，以便更好地回答用户的问题或执行任务。

在‘Titan Matrix’安全品牌的护航下，涂鸦 AI 平台构建了以下安全实践：

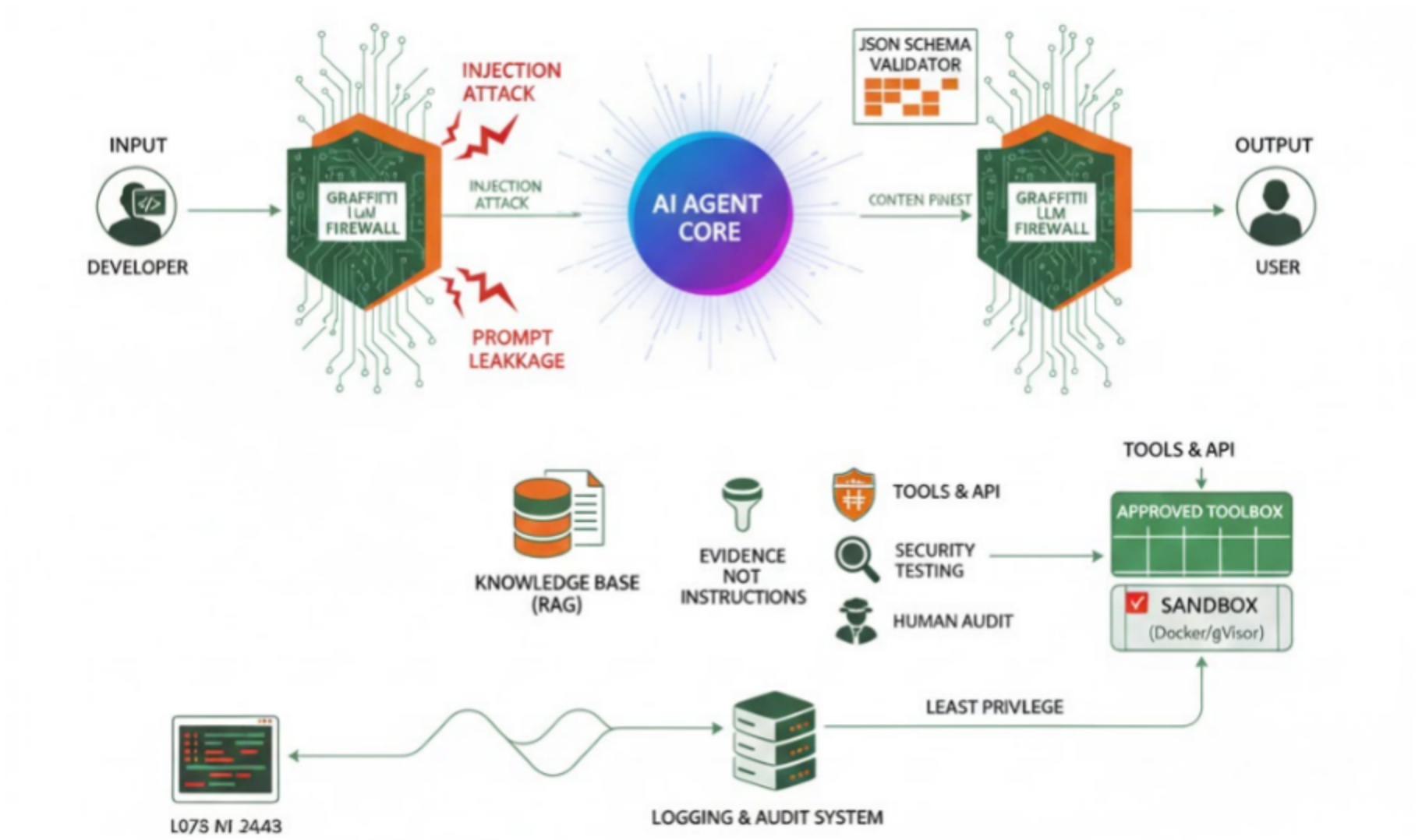
提示词攻击检测

大模型防火墙是‘Titan Matrix’安全中心提供的核心能力之一。涂鸦的大模型防火墙专业防御针对生成式 AI 的**注入式攻击，精准识别越狱指令、角色扮演诱导、系统指令篡改等对抗性攻击行为，构建 AI 系统的“免疫防线”**。目前在 AI 对话的输入输出、AI Agent 的指令交互的输入输出等场景均进行严格的检测和拦截。

内容安全检测

针对内容安全合规，依托于‘Titan Matrix’安全中心的能力，涂鸦实施内容合规检测和敏感内容检测，对生成式 AI 输入输出的文本内容进行**多维度合规审查，覆盖涉政敏感、色情低俗、偏见歧视、不良价值观等风险类别**，确保 AI 生成内容符合法律法规与平台规范。同时，深度检测 AI 交互过程中可能泄露的隐私数据与敏感信息，支持涉及个人隐私、企业隐私等敏感内容的识别，防范训练数据泄露与对话信息外溢风险。目前在所有涉及 AI 对话、智能客服、知识库问答等环节均强制启用检测。

AI Agent 配置审计



针对开发者配置的 AI Agent 的输入输出也同时接入了涂鸦大模型防火墙，**防止注入攻击、提示词泄露**和实施针对性**内容安全过滤**。同时针对**输出强制符合预定义的 JSON Schema**，防止输出被篡改或执行意外操作。针对知识库 RAG 的挂载，**强制去指令化**，明确告知模型引

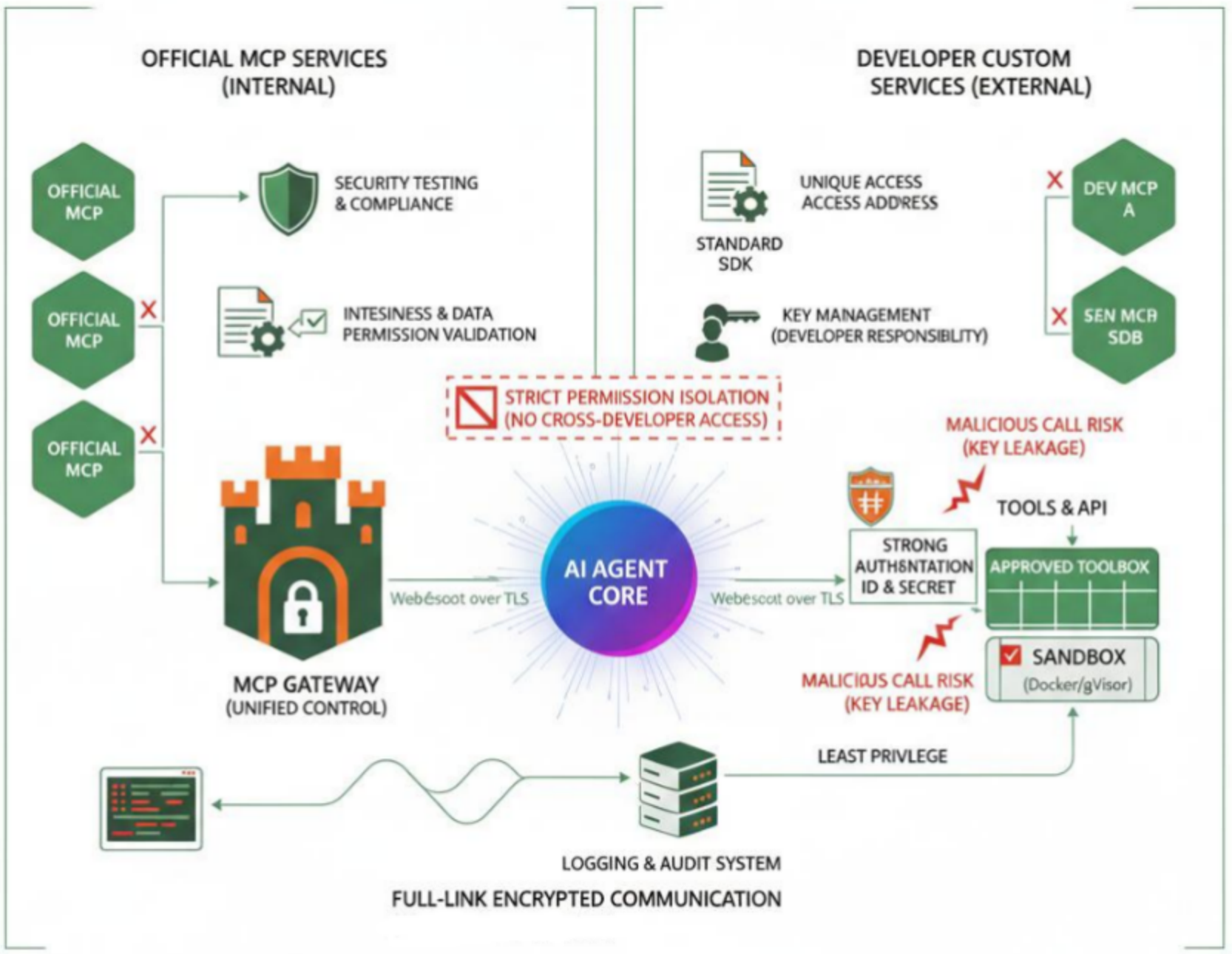
用的文档是“证据”而非“指令”，防止间接注入。

同时，使用针对平台开放的官方工具上架前需要进行严格**安全测试和风险评估**，需要通过安全团队的人工审计和确认。确保为每个 Agent 工具分配完成任务所必需的**最小权限**。如果，存在高风险的操作必须在**隔离环境（如 Docker, gVisor）**中运行。

最后，涂鸦的服务和应用均严格记录 Agent 的关键决策、工具调用和异常事件，便于事后追溯与分析，确保能够持续监控和审计三方 AI Agent 的风险。

MCP 服务审计

涂鸦的 MCP 服务分为官方 MCP 与开发者自定义 MCP 两类。针对官方 MCP，我们通过自研的 **MCP 网关**实施统一管控。所有内部 MCP 服务上线前均须通过严格的安全测试与合规评估，并强制执行**接口调用白名单**机制——每个接口必须通过安全评审方可配置。同时，所有服务均需接入**业务权限与数据权限校验**，从根本上防范越权行为。



对于开发者自定义 MCP 服务，平台提供标准 SDK 以供集成。每个服务将获得由平台生成的唯一接入地址、Access ID 与 Access Secret，并实行**强认证机制**，仅认证通过的请求可

访问对应 MCP 服务。开发者需妥善保管密钥，以防恶意调用。平台实行严格的**权限隔离策略**，开发者仅可配置自有 MCP 服务至其名下的 AI Agent，确保不同开发者间的服务互不干扰，有效防范越权配置与恶意污染。

所有 MCP 服务的通信均基于 **WebSocket over TLS** 协议，构建安全可靠的全链路长连接，保障数据在传输过程中的机密性、完整性与可用性。

AI 服务数据权限设计

涂鸦平台对所有业务数据实施**分级分类**（如公开、内部、机密、个人敏感信息等）。不同级别的数据，对应不同的访问策略，确保高敏感数据始终处于最高防护等级。

同时，所有的 AI 服务均无权直接调用涂鸦内部底层服务，如 IoT Core、数据库、大数据服务等，防止业务越权和数据越权，通过**统一的接口服务网关**，网关会严格调用**鉴权服务**，对应用和服务，以及承载的开发者或 C 端用户，及其对应的业务和数据权限进行严格的校验。权限逻辑，在 RBAC（基于角色的访问控制）基础上，引入环境、数据敏感性等动态属性，实现有效的鉴权决策。

AI 在交互过程中，如展示数据，会对身份证、手机号等敏感字段进行**自动掩码处理**。通过静态脱敏进一步保护用户隐私。

大模型安全审计

对于所有涂鸦平台接入的三方大模型，均经过涂鸦安全合规部严格的安全评估与合规审计。流程覆盖服务全生命周期的安全管理策略。

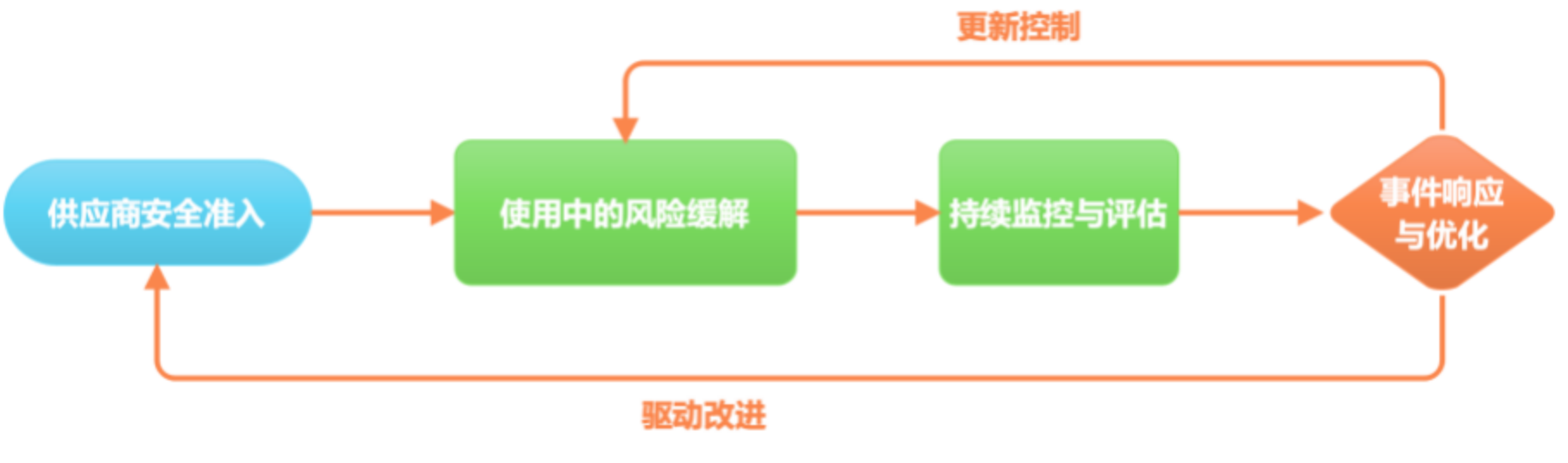
1. 在采购或接入任何第三方大模型之前：

- **安全合规问卷**：通过向供应商发放详细的问卷，涵盖其数据安全与隐私保护政策（如何加密训练数据和用户输入、数据是否会被用于持续训练、保留时限）、模型安全能力（是否内置内容过滤、对抗攻击防护）、合规认证（是否通过 SOC 2 Type II、ISO 27001、等保测评等）以及安全事件响应流程。

- **合同与法律约束**：在服务协议中明确数据处理协议，确保供应商承担数据控制者或处理者的法律义务。同时，通过合同条款严格约定数据所有权、保密性、禁止用于模型训

练的要求，并明确安全违规的处罚与问责机制。

● **大模型安全测试**：安全团队对大模型进行手工专家级安全测试，如存在风险需要确保完成修复后，才能完成采购交付。



2. 使用中的风险缓解与控制：

● **严格的 API 密钥管理**：使用安全的密钥管理系统存储和轮换 API 密钥，避免硬编码在代码中。

● **安全代理网关**：内部开发了大模型统一安全网关。所有请求都通过该网关转发，在此集中实施：内容安全过滤，速率限制与配额管理以及完整的审计日志。

3. 持续监控与定期评估：

● **定期安全复评**：每半年或一年重复第一阶段的评估流程，审查供应商的安全状况是否有变化或出现新的安全事件。

● **红队演练与对抗测试**：定期组织内部安全团队或聘请第三方对使用的 AI 服务进行模拟攻击。

4.2 Hey Tuya：安全可靠的家庭 AI 生活助手

关键词：隐私数据保护、内容安全过滤、数据安全保障、移动应用安全

Tuya 致力于打造家庭 AI 生活助手（Your AI Life Assistant at Home）。Hey Tuya 是其核心 AI Agent，支持文字、语音、视频等多种对话形式，可跨终端使用——手机、中控屏、音箱等。围绕日常生活核心场景（安全、节能、效率、健康、陪伴），Tuya 打造了一系列 AI Agent 产品矩阵，通过多 Agent 协同架构，为用户提供全场景智能生活体验。

隐私保护声明

涂鸦的 APP 严格遵守各国法律法规和全球主流信息安全和隐私保护规范。我们拥有完善的**用户个人信息保障方案**，包括但不限于公开收集使用规则、完善的流程保障、先进的隐私设计（PbD）规范、明示收集使用个人信息的目的方式和范围、完善的用户同意解决方案以及详细的投诉举报或提供给用户反馈的人工处理渠道等措施来保护用户隐私权益。

同时，隐私声明中针对 AI 服务对隐私数据的采集和使用，有清晰明细地功能列举、可能采集的数据，以及涂鸦的承诺包括但不限于不会将任何用户输入的内容用户模型训练等。同时，涂鸦也在应用内明示展示了算法和模型在中国境内的**备案**信息，列举了**算法基本原理、运行机制和应用场景**。此外，针对所有使用的第三方 AI 服务，均有详细地列举，包括信息采集的内容和具体使用的场景。

内容安全过滤

涂鸦对 Hey Tuya 服务的交互内容实施全链路合规检测，**覆盖涉政、色情、歧视及价值观等风险维度**，确保输出合法合规。同时，平台深度筛查并识别交互中可能泄露的个人与企业隐私信息，并对所有 AI 输出中的疑似敏感数据执行强制脱敏，从根本上杜绝信息泄漏风险。

数据安全保障

APP 与云端的 AI 服务交互过程全链路均基于 **WebSocket over TLS** 协议，通过安全可靠的长连接链路，保障数据在传输过程中的机密性、完整性与可用性。同时，针对用户上传

的图片和视频在涂鸦云存储时均强制使用 AES128 强制进行文件加密处理。

移动应用安全

涂鸦智能的移动应用采用了严格的**加固**，包括但不限于防篡改、代码混淆、模拟器拦截检测、Root 环境检测、反调试、界面劫持保护等，能够有效地保护客户端程序。同时，针对客户端产生的本地数据，使用安全的**加密方式存储**，并严格限制读写权限，特别针对密钥等信息，通过移动操作系统的**密钥存储服务**实施安全存储等。

4.3 AI 玩具：面向特殊群体的守护

关键词：儿童隐私保护、儿童内容安全、终端通讯安全、智能终端安全

涂鸦智能 AI 玩具解决方案，是把大模型嵌入玩具，实现玩具的智能化升级，通过涂鸦智能的 AI 玩具方案实现低成本、高效率、快速上线、统一品牌形象。

儿童隐私保护

涂鸦智能在儿童隐私保护上，满足 COPPA 在内的主流国家针对儿童隐私保护相关法律的要求。包括：**1.年龄验证与家长授权机制；2.严格的儿童数据保留限制；3.最小化收集与用途限定，确保儿童数据安全与合规**。

目前，在 APP 上针对儿童隐私有专门的隐私保护声明。同时，在产品包装和说明书上，均有儿童风险提示。

内容安全分级

内部针对儿童玩具相关的 AI 服务进行了严格的内容安全分级，包括涉及法律红线的内容、影响儿童心理健康的内容和涉及价值观塑造的内容，均有特定的过滤和处置方案。

同时，除了使用大模型防火墙针对大模型对输出的检测和拦截以外，涂鸦智能内部也通过持续地收集和整理相应的内容安全合规的检测用例，对 AI 服务进行常态化验证和实施持续性的对抗，保证 AI 服务输出内容的安全与合规。

AI 服务点对点通讯安全

信令通讯安全：第一层基于 TLS 1.2 协议建立传输通道，通过 ECDHE 密钥交换实现前向保密特性，确保通信过程密钥的临时性。第二层，业务数据采用 AES-128-GCM 算法进行加密，该算法通过 NIST 认证的 AEAD 模式同时实现数据机密性与完整性保护，密钥采用基于设备和用户身份生成，实现设备身份与用户凭证的双因子密钥派生，确保了密钥的唯一性。

链路端到端认证：除了依托于可靠的信令通讯链路，端到端的过程双端均需强制验证基于设备唯一密钥和权限信息生成的临时用户名和密码。同时，终端会在本地生成自己的本地认证信息，该认证信息构成端到端的第二道安全认证，完成两道端到端的安全认证，才能够创建 TVS 链路。

链路端到端加密：采用“一机一密”基础密钥（EK）与临时会话密钥（TSK）双重保护机制，每次会话终端上会生成新的随机加密密钥来保证通讯加密安全，该密钥由端到端派发，云服务中无法存储和感知。同时，传输数据基于 HMAC-SHA256 的消息认证机制，防范中间人攻击。

智能终端安全

设备认证：涂鸦智能硬件方案在生产的时候，会写入一对设备认证信息，这些信息在所有设备中都是唯一的，与模组的环境绑定，包括芯片 ID 和 MAC 地址等，在每个会话请求中，签名数据包的时候加签了这些信息。每个通讯都必须保证模组的环境信息和设备认证信息准确，才能够有效通讯。

OTA 安全：涂鸦针对固件升级过程采取了多重保护手段进行保护。1.在生成固件包时，

打包工具会生成一个固件完整性校验信息，该信息由多个变量组成；2.客户端请求固件时，服务端会下发一个固件下载信息和固件校验信息。该固件校验信息采用安全的 HMAC 签名算法，并且加签设备唯一的身份密钥信息，保证传输过程中固件的合法性且无法被篡改；3.客户端获取固件后，需要计算固件校验信息，并和服务端提供的固件校验信息进行对比，同时解压缩的时候还需要校验打包工具在固件内计算的完整性校验信息。只有完成固件双重校验后，才允许写入固件；4.固件如果写入失败，或写入后无法正常使用，会自动恢复到原有的固件。

数据安全：为了保障核心数据的安全，智能终端本地存储的重要信息，会进行 AES 加密后，存放。加密的密钥每个芯片初始化的时候随机生成，并安全存储，仅本地加密使用，不用于任何业务处理或进行任何交互。

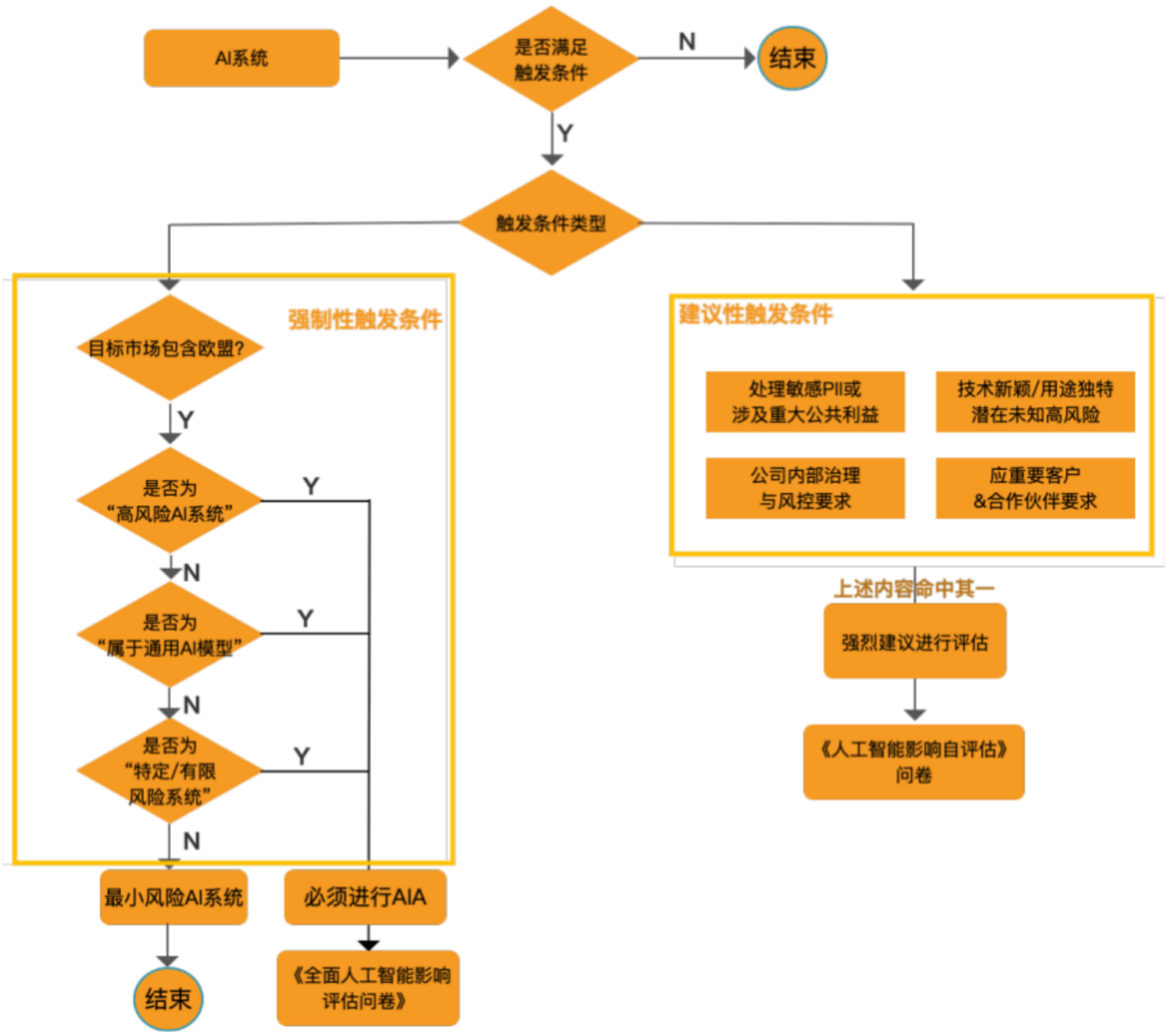
第五章：主动的风险监测与防御体系

涂鸦致力于建立前瞻性、系统化的人工智能风险管理框架，将其深度整合至企业风险管理体系中，以确保 AI 风险可知、可控、可承受。

5.1 风险评估机制

涂鸦制定并维护《AI 系统影响评估和风险评估管理制度》，明确风险识别、分析、评价及更新的标准化流程，当前评估分为周期性评估和事件驱动评估。

人工智能系统影响评估流程



定期评估：每年至少开展一次全面的 AI 专项风险评估，覆盖所有已部署及在研的高风险 AI 系统。

事件驱动评估：当发生重大技术变更、业务模式调整、法律法规更新或安全事件时，须立即启动针对性的风险评估。

5.2 风险处置与接受

依据《人工智能系统安全事件管理办法》，执行 AI 安全事件应急响应流程。

依据风险评估结果，制定并执行风险处置计划，优先采取技术性和管理性措施以降低风险。

对于无法经济有效地消除或降低的残余风险，须明确记录并提交至合规委员会，依据集团既定的风险接受准则进行审批。

5.3 合规性保障

所有 AI 活动必须严格遵守运营所在地关于人工智能、数据安全、个人信息保护及行业监管的现行法律法规与监管要求。

涂鸦法律与合规部门动态跟踪立法动向，并定期组织合规性评审，确保集团 AI 业务的前瞻性合规。

涂鸦内部对于在已出台严格人工智能监管法规的区域（例如欧盟）运营的 AI 系统，在部署前通过法律与合规部门的强制性合规审查，并获得其书面批准。

5.4 安全性保障

涂鸦依托‘Titan Matrix’强大的防护体系，构建了成熟的**纵深防御架构**，该体系包含了从**网络、服务器、应用到代码层面全方位**，细粒度的防护，相关安全能力包含 **WAF（WEB 应用防火墙）、RASP（应用自适应安全防护系统）、CNSPs（云原生安全平台）、HIDS（主机入侵检测系统）、HoneyPot（蜜罐）、SIEM（安全事件分析平台）**以及诸多业务安全应用和组件等多重防护机制，以在技术架构上多层次多维度识别和响应来自互联网的各种威胁。

当前 AI 的攻防能力已经完全融入涂鸦的纵深网络安全防护架构中，包括 **AI 大模型防火墙、AI 蜜罐、AI 业务日志安全审计**等。其中 AI 大模型防火墙能够实现内容安全、内容合规、提示词注入攻击拦截等功能；AI 蜜罐除了模拟常规的业务服务，也涵盖了部分大模型基础设施，包括 MCP 蜜罐；AI 业务日志审计，则基于应用的日志流全面接入到涂鸦 SIEM 中，对

可能存在的攻击行为进行检测和风险评估，并能够联动 SOAR 等其他安全防护系统进行风险封堵。

5.5 持续监控与报告

涂鸦通过构建或利用现有的统一可观测性平台，将 AI 系统的**日志（Logs）、指标（Metrics）、追踪（Traces）以及专有的安全事件数据进行关联存储与分析**，实现了统一可观测性平台的数据融合。这打破了数据孤岛，使得在排查问题时，能从一个异常指标快速下钻到相关的错误日志和具体请求链路，极大缩短根因定位时间。

涂鸦还引入前沿的**智能分析**技术，实现对 AI 系统风险的深度感知和前瞻预警。其中，对监控中识别的风险事件进行实时分级（如基础级、关注级、管控级）。通过**构建“技术—应用—业务”风险关联图谱**，实现从孤立告警到系统性风险态势感知的转变。例如，当同时检测到模型推理异常、关联数据源访问激增和网络异常流量时，系统应能自动关联并提升整体风险等级。

在自动化风险响应和闭环处理层面，涂鸦建立与监控系统深度联动的**自动化响应**能力，将风险处置从“人工响应”升级为“系统自愈”。

1）**与 SOAR 深度集成的自动化剧本**：将安全编排、自动化与响应平台深度整合至 AI 风险监控体系。针对预设的 AI 风险场景（如提示词注入攻击、模型输出内容违规、资源滥用等），预制自动化响应剧本（Playbook）。一旦监控系统触发告警，SOAR 剧本可自动执行一系列动作，例如：隔离受损的模型实例、调用防火墙 API 阻断攻击源 IP、临时下线高风险模型服务进行审查，并自动生成事件工单。

2）**资产与供应链的持续监控**：利用专项工具对内部 AI 资产（包括私搭乱建的模型服务）和 AI 供应链（如第三方模型、数据集、开源组件）进行持续扫描和监控。自动识别组件中的已知漏洞（如特定框架的未授权访问漏洞）、许可合规问题，并与漏洞库和威胁情报进行实时比对，实现风险的前置化发现与管理。

最后，通过建立定期的风险报告机制，将 AI 风险管理状况纳入集团整体的风险报告体系，并向合规委员会进行汇报。

第六章：遵循全球法规的合规性实践

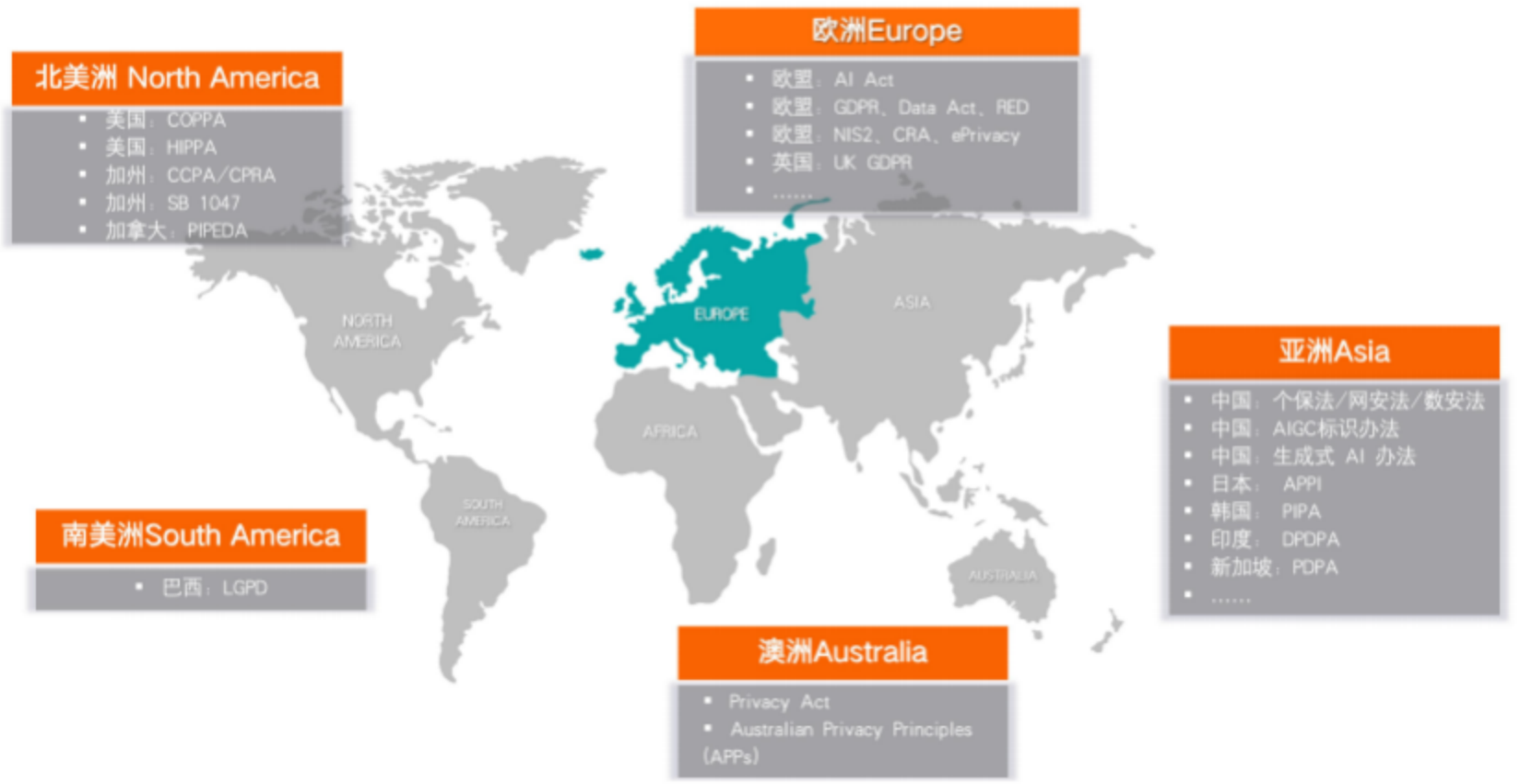
6.1 国际标准体系

涂鸦‘Titan Matrix’在隐私与合规支柱的全球实践中，获得了 **ISO42001 AI 管理体系认证，ISO27001 安全管理体系认证，ISO27701 隐私信息保护管理体系认证，ISO27017 云服务信息安全管理体系认证等 ISO 认证矩阵**，证明了涂鸦提供的 AI 能力在产品与服务层面，拥有了全球领先的安全合规能力保障。同时，涂鸦也拥有三方权威的 **SOC2&SOC3 审计报告，GDPR、CCPA、PIPEA 等各国法规审计报告，还有攻防演练和针对涂鸦产品和服务的安全评估报告**等，这些审计证明了涂鸦坚实有力地持续执行着先进的安全与合规的承诺，都是‘Titan Matrix’能力获得的权威鉴证。

6.2 全球法规映射

涂鸦密切关注全球人工智能监管趋势，持续对照国际主流法规体系进行实践落地。我们在 AI 全生命周期设计、开发、测试、部署及运维过程中，严格遵循各司法辖区的法律与监管要求，确保产品与服务的安全性、透明性与合规性。基于“全球统一基线 + 地区化适配”的治理策略，涂鸦构建了覆盖 AI、安全、隐私和数据的全球法规映射矩阵。

全球 AI & 数据保护法规



1. 欧盟：EU AI Act 的风险分级与合规实践

EU AI Act 将于 2026 年 8 月全面生效，是全球首个覆盖全生命周期的人工智能综合法规，依风险将 AI 系统分为不可接受风险、高风险、有限风险和低风险。涂鸦从制度与实践两方面进行落地：

（1）禁止涉入不可接受与高风险类别

- 涂鸦承诺不开发、提供或销售 EU AI Act 明确列为**不可接受风险**的系统。
- 原则上不涉及**高风险 AI 系统**。如确需例外，必须经过高级别风控审批及完整技术文档化。

（2）内部风险分级制度化落地

涂鸦制定《AIoT 设备控制风险分级模型与管控规范》，依据“可预见风险—严重性—发生概率”原则对 AI 可控制的设备和指令进行分级，明确 AI 控制要求与人工确认要求，保障用户与设备安全，确保 AI 决策链路安全、可控、可审计。

（3）透明度、可解释性与用户知情权保障

针对有限风险及低风险场景：

- 在交互界面显著标识 AI 身份
- 在隐私政策和文档中披露算法逻辑、数据使用目的、潜在影响
- 为用户提供可解释性说明

- 提供人机协同机制

这些要求均与 EU AI Act 第 52 条透明度义务对齐。

（4）供应链责任与模型治理

涂鸦还建立了对第三方模型、插件和供应商的合规审查流程，包括：

- 模型来源与训练数据合规性评估
- AI 系统风险评估与影响评估
- 输出质量与安全合规测试
- 供应链安全合规问卷审查

2. 中国：生成式人工智能与深度合成监管

中国在 AIGC 和算法治理方面形成了成体系的监管框架。涂鸦从制度、产品能力和运营流程三方面同步落实。

《人工智能生成合成内容标识办法》于 2024 年 9 月生效，涂鸦进行了深入解读和全员合规培训，已按要求落实显式及隐式标识。同时遵循《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》《互联网信息服务算法推荐管理规定》等要求，在内容安全、算法安全和数据合规方面全面加强管理。

3. 涂鸦跟踪的全球 AI 与数据保护法规

涂鸦构建了持续更新的全球法规库，用于指导产品设计、算法治理、安全评估和隐私合规。

（1）欧盟

EU AI Act、EU Data Act、GDPR、RED、NIS2、Cyber Resilience Act、ePrivacy Directive

（2）美国

加州 SB 1047、CCPA/CPRA、COPPA、HIPAA

（3）中国

AIGC 标识办法、深度合成规定、生成式 AI 办法、算法推荐规定、App 必要信息规定、网络安全法、数据安全法、个人信息保护法

（4）其他重要法域

- 印度：DPDP + DPDP Rules（2025）
- 韩国：PIPA、AI 基本法（草案）
- 巴西：LGPD、Brazilian AI Act（草案）
- 新加坡：PDPA
- 加拿大：PIPEDA、Quebec Law 25
- 英国：UK GDPR、PSTI
- 澳大利亚：1988 隐私法
- 越南：网络安全法
- 沙特：PDPL

涂鸦将持续跟踪全球 AI 及数据保护法规，持续完善合规体系，保障全球用户权益，并协助客户在各市场安全合规运营。

6.3 产品合规-by-Design

在涂鸦智能的技术体系中，“合规-by-Design”并非一个独立的环节，而是一套贯穿产品硬件、软件、通信与云端的系统性技术架构。我们通过将安全与合规的核心要求转化为可执行、

可验证的技术规范和自动化工具，确保合规性在产品诞生之初便已内嵌其中。

● 在设备端，我们致力于**从物理面建立不可篡改的安全基础**。我们所提供的预认证硬件模组，其技术价值在于其内部固件已深度集成了**符合目标市场法规的安全协议栈**。例如，为**满足欧盟 RED 指令、英国 PSTI** 等主流市场的网络安全要求，模组固件层面默认启用并正确配置了必要的安全功能，如无线通信加密强度、设备唯一性标识和安全的软件更新机制。

● 在通讯与数据层，所有数据传输均通过高强度 **TLS 1.2/1.3** 加密通道进行，同时内容采用 AES 额外一层安全加密，确保数据在传输过程中的机密性与完整性。同时，对于高敏感业务，实行**端到端加密，密钥端到端动态协商和生成**，仅用户与用户终端可执行解密。

● 在云端和平台层，提供**原生安全能力**，开发者能够基于需求对 APP 和终端设备进行按需安全配置，并且支持安全的 **OTA 升级**机制。

● 在隐私保护层面，我们将隐私设计原则转化为具体的技术控制措施。包括执行数据最小化与匿名化处理：在数据上报链路中，平台支持对设备日志和运行数据进行字段级的脱敏和匿名化处理，确保在业务功能不受影响的前提下，避免不必要的个人标识符上传与存储。精细的权限访问控制：平台实现了基于角色的访问控制模型。开发者可以为其应用的用户定义精细的数据和设备操作权限，确保用户只能访问其被授权范围内的设备和数据，从技术层面强制执行了权限分离原则。

第七章：透明沟通与未来展望

7.1 我们的承诺：持续透明与改进

我们坚信，透明是信任的基石。我们承诺：

1. 主动披露：定期发布 AI 安全与治理透明度报告，公开关键安全指标、合规进展与伦理审查摘要。

2. 开放参与：通过“Titan Matrix 安全生态联盟”，邀请客户、开发者及学术机构共同参与安全标准研讨与威胁情报共享。

3. 响应反馈：保持开放的沟通渠道，对每一份来自用户、合作伙伴或研究者的安全问询与建议，给予及时、负责地回应。

我们将持续改进，确保我们的行动始终与全球最前沿的负责任 AI 实践同步。

7.2 未来之路：从安全合规的践行者，到可信智能生态的构建者

展望未来，涂鸦智能将依托 “Titan Matrix”一体化安全基座，引领 Physical AI（物理人工智能）的安全可信发展：

定义下一代 AI 原生安全

我们将持续投入前沿安全技术研发，重点布局：

- **边缘智能安全**：研发轻量化、高可靠的端侧 AI 安全模块，为海量 IoT 设备提供本地的实时隐私计算与威胁防御能力，降低对云端协同的依赖与延迟。

- **深度伪造与多模态攻击防御**：构建针对 AI 生成的伪造音视频、跨模态（图文、语音指令）组合攻击的专项检测与溯源体系，捍卫数字世界真实性。

- **自主进化的安全智能**：深化 AI 在安全运营（AISecOps）中的应用，打造能够从攻

击中自主学习、动态调整防御策略的“自适应免疫系统”。

生态共赢：将企业级安全，化为开发者的天然底气

我们深信，最极致的安全，是让开发者无需为安全而分心。因此，我们绝不将安全作为额外的选项或复杂的 API 交付，而是致力于将 “Titan Matrix” 所承载的企业级安全能力，深度内嵌至涂鸦平台每一个基础架构与服务中。

- **原生融合，无感防护**：当开发者使用涂鸦的 AI Agent 开发平台配置一个智能体时，提示词注入攻击检测与内容安全过滤已自动生效；当调用设备控制接口时，严格的权限校验与行为风控已在底层默默护航。安全不再是需要单独集成的外挂模块，而是如同平台提供的计算、存储一样的基础能力。

- **统一基座，一致体验**：从 TuyaOpen 开源框架到云端 API 服务，我们为所有开发路径预置了统一的安全基座。这意味着，无论开发者从何种入口构建 AIoT 应用，都能自动获得符合全球主流法规要求的安全与隐私保护设计，极大降低了合规门槛与潜在风险。

通过这种方式，我们赋能全球开发者：**无需成为安全专家，即可打造出具备顶尖安全基因的创新产品**。涂鸦的目标，是让每一位开发者在专注于创造业务价值的同时，拥有一份来自平台层的、坚实可靠的安全底气。

我们的决心：涂鸦智能的目标，不仅是成为最安全的 AI 平台，更致力于成为**可信智能生态的基础构建者**。我们将以“Titan Matrix”为引擎，以透明、开放为准则，携手全球伙伴，共同迎接一个安全、可靠、以人为本的万物智能未来。

附录

A. 术语表

Titan Matrix：涂鸦智能旗下的一体化智能安全品牌，涵盖研发安全管理、安全中心、立体防护体系、AI 安全、隐私与合规五大支柱，为涂鸦 AIoT 生态及开发者提供全栈、全生命周期的安全能力。

PREPARE 框架：涂鸦智能遵循的负责任 AI 核心原则，由 Privacy（隐私设计）、Resilience（韧性）、Equity（公平）、Proportionality（适度）、Accountability（问责）、Reliability（可靠）、Explainability（可解释）的首字母组成。

AI 全生命周期管理：指对人工智能系统从规划、设计、开发、部署、运营到退役归档的全过程进行系统性、一体化的安全管理与合规控制。

算法备案：根据中国《互联网信息服务算法推荐管理规定》等法规，要求算法提供者通过官方渠道提交算法基本信息，以增强透明度的一项合规义务。

残余风险：在采取所有合理可行的风险处置措施后，仍然存在且需要被主动管理和接受的剩余风险。

纵深防御架构：一种网络安全设计理念，通过在网络、主机、应用、数据等多个层面部署重叠的安全防护措施，以防范单一防线失效的风险。

AI Agent：能够感知环境、进行决策并执行行动以实现特定目标的智能体。常指基于大语言模型、能调用工具（如控制设备）的智能应用程序。

MCP：模型上下文协议，是一项允许大语言模型安全、标准化地访问外部数据源和工具的

开放协议。涂鸦平台提供官方的 MCP 网关对相关服务进行安全管理。

可观测性：通过收集和分析系统的日志、指标、追踪三类数据，从而能够深入理解其内部状态与性能的能力，是持续监控的技术基础。

SOAR：安全编排、自动化与响应。指通过平台将安全工具连接起来，并预定义剧本，以实现安全事件响应流程的自动化与标准化。

WAF：Web 应用防火墙，用于监控、过滤和拦截进出 Web 应用的 HTTP 流量，防御常见的 Web 攻击。

SIEM：安全信息与事件管理平台，负责集中收集、关联分析各类安全日志和事件，用于威胁检测和调查。

提示词注入攻击：一种针对大语言模型的安全攻击，攻击者通过精心构造的输入（提示词），诱导模型绕过原有的安全限制或指令，执行非预期的操作。

对抗样本：经过特殊构造的、旨在导致机器学习模型做出错误判断的输入数据。防范此类攻击是模型鲁棒性的重要组成部分。

数据投毒：在模型训练阶段，通过向训练数据中注入恶意样本，从而影响模型学习过程、最终破坏模型性能或行为的攻击方式。

AI 大模型防火墙：专门针对大模型应用部署的安全防护系统，核心功能通常包括提示词注入检测、输入输出内容安全过滤等。

红队演练：一种主动安全评估方法，通过模拟真实攻击者的战术、技术和流程来检验系统防御的有效性。

Physical AI：物理人工智能，指与物理世界（通过物联网设备）深度交互、并能产生实际影响的 AI 系统，是涂鸦 AI 战略的核心方向。

B. 涂鸦核心 AI 管理制度列表

- 《涂鸦人工智能管理章程》

- 《人工智能系统安全事件管理办法》
- 《人工智能系统资源管理规定》
- 《AI 系统评估和风险评估管理制度》
- 《AI 系统全生命周期策略规范》
- 《AI 系统训练数据管理规范》
- 《人工智能供应商管理规范》

C. 已获得的安全与隐私认证或审计列表

全球性：

- ISO/IEC 42001:2023 AI 管理体系认证
- ISO/IEC 27001:2022 信息安全管理体系认证
- ISO/IEC 27017:2015 云服务信息安全管理体系认证
- ISO/IEC 27701:2019 隐私信息管理认证
- CSA STAR Cloud Security 云安全管理体系认证
- SOC 2 Type II 美国注册会计师协会 SOC2 审计报告（涵盖安全性、可用性、处理完整性、保密性和隐私性方面的控制措施评估）
- SOC 3 美国注册会计师协会 SOC3 审计报告
- Enterprise Privacy Certification 企业隐私认证
- CMMI 3.0 CMMI 软件能力成熟度评估认证 3 级
- ISO/IEC 9001:2015 质量管理体系认证
- PSA Certified Level 1 物联网设备安全性评估认证
- ioXt Manufacture Certified 物联网安全认证

区域性：

- 【欧盟】GDPR 验证性报告
- 【欧盟】RED DA：欧盟 RED 认证
- 【欧盟】ETSI EN 303645：欧盟消费类物联网产品网络安全标准认证
- 【欧盟】CE 认证：欧盟安全、健康和环保要求
- 【欧盟】RoHS 认证：欧盟环保认证
- 【欧盟】REACH 认证：欧盟化学品管理法规
- 【美国】CCPA 验证性报告
- 【美国】NISTIR 8259：美国 NIST IoT 安全基线认证
- 【美国】FCC 认证：美国联邦通信委员会电磁兼容性标准
- 【中国】等保 3 级：网络安全等级保护备案 3.0
- 【中国】CQC 认证：中国质量标准认证
- 【中国】CCC 认证：中国安全与质量认证
- 【英国】PSTI：英国产品安全和电信基础设施法案
- 【加拿大】PIPEDA & Quebec Law25：加拿大隐私保护评估报告
- 【加拿大】IC 认证：加拿大电磁兼容性要求
- 【印度】DPDPA 合规评估报告：印度数字个人数据保护法合规性评估报告

D. 联系我们

涂鸦漏洞悬赏平台：https://src.tuya.com

Titan Matrix 安全与合规支持：sec@tuya.com