涂鸦 AI 玩具合规白皮书 v1.1

涂鸦始终高度重视 **AI 内容安全、儿童保护、数据隐私保护与全球法规遵从**。面向快速发展的 AI 玩具行业，涂鸦构建了覆盖模型训练策略、提示词治理、产品逻辑设计、内容审核、儿童保护、安全防护、数据管理及退役管理在内的全生命周期合规体系，以支持产品在全球范围内满足安全、透明与可信赖的要求。

本白皮书面向合作伙伴，系统介绍涂鸦 AI 玩具在模型安全、内容安全、数据隐私、儿童保护、家长控制以及全球法规遵从方面的合规能力，以回应行业对“AI 玩具内容风险、儿童保护合规、算法透明度与数据安全”等核心问题的关注。

一、AI 玩具的全生命周期合规体系

涂鸦基于 Security by Design、Privacy by Design & by Default 原则，构建了覆盖**设计 → 开发 → 测试 → 使用 → 退役**的 AI 玩具全生命周期合规框架，从源头确保产品在功能、数据、内容与行为的全链路安全合规可控。

1.1 设计阶段

在设计阶段，涂鸦会开展 AI 玩具风险评估、人工智能系统影响评估（AIA）及数据保护影响评估（DPIA），绘制数据流并识别处理边界，严格实施数据最小化与去标识化策略。同时，涂鸦会对大模型供应商的安全与合规能力进行评估，确保模型来源可靠，并在产品需求、功能设计以及语音交互逻辑中引入 AI 玩具安全合规基线，为后续开发与使用阶段奠定基础。

1.2. 开发阶段

涂鸦通过安全 Prompt 与儿童场景专用提示策略来防止被篡改，同时在输入侧实施敏感内容限制，在输出侧实施儿童友好表达约束，集成敏感内容检测能力，严格落实数据最小化、加密和权限最小化原则。采用 SDK 白名单机制，实现数据流转路径的可控和可追踪性。

1.3. 测试阶段

涂鸦对话系统会进行敏感内容测试、越狱测试和鲁棒性测试，同时覆盖未成年人不当内容和儿童心理安全测试。安全对抗测试也在此阶段进行，并依据 AI 玩具安全合规基线执行全量功能验证，以保障产品上线时符合安全规范。

1.4 使用阶段

涂鸦明确向儿童与监护人提示 AI 系统为非真人特性，并提供算法透明度说明与模型能力边界说

明。AI 玩具固化行为边界，拒绝讨论成人内容、危险操作或提供专业建议，从而保证儿童使用环境的安全性和可控性。

1.5 退役阶段

涂鸦提供可审计的设备退役管理机制，包括数据删除、设备解绑、功能停用控制与账号删除，确保 AI 玩具生命周期结束后不会产生安全隐患。

该体系适用于涂鸦全线 AI 玩具产品，用于支撑儿童使用环境的安全性、透明性与可信度。

二、 内容安全控制体系

AI 玩具的核心挑战是防止 AI 生成不当、危险或对儿童不友好的内容。涂鸦采用“**模型安全 + Prompt 限制 + 输入过滤 + 越权防护 + 人工测试**”的多层防护体系。

2.1 合规模型选择

涂鸦仅选择通过安全评估的主流大模型厂商，如 Google、Microsoft、ByteDance、Alibaba 等主流合规模型厂商，这些模型均具备以下内置的非法内容拒答机制，包括：

- 自动拒绝色情、暴力、仇恨、歧视类内容
- 拒绝违法、危险操作行为指引

并具备多语言内容安全能力，涂鸦在其基础上再叠加“儿童安全层”，形成双重保护。

2.2 安全提示词（Prompt）策略

根据儿童场景特点，涂鸦加入多重限制性提示词，包括但不限于：

- 禁止模型讨论暴力、色情、自残、歧视、毒品、辱骂、恐怖、仇恨等内容
- 禁止模型输出任何成人内容或专业建议（如医疗、法律等高风险领域）
- 禁止模型提供危险操作指导、操纵、施压或在情感上影响儿童
- **固定角色锁定：** 只能扮演“儿童教育伙伴/陪伴角色”等儿童友好角色
- **儿童友好表达：** 提供积极、阳光、温和、安全的语言风格
- 对于不适宜儿童的问题，AI 玩具会统一采用安全回应策略，这样既保证安全，又保持玩具的互动体验。例如：
  - “我不能讨论这个话题，但我可以陪你玩游戏”
  - “我不能回答这个问题，但可以和你聊聊更有趣的话题！”

- “这是成人话题，我不能讨论哦，我们来讲个故事吧！”

此类 Prompt 由系统锁定，无法被用户或本地设备修改，防止越狱。

2.3 输入内容审核

涂鸦针对用户输入设置强制过滤，该控制在模型前生效，能有效阻断不当对话：

输入过滤：输入若触发过滤规则将采用统一安全回应策略。过滤内容包括：

- 色情与不适宜儿童内容
- 暴力或威胁性描述
- 仇恨、歧视、辱骂
- 自残、危险操作建议
- 过度共情或操控性表达
- 恐怖惊悚内容
- 敏感政治内容
- 法律、医疗等专业建议

2.4 反越狱（Anti-bypass）机制

涂鸦在 AI 玩具中引入反越狱机制，通过反越狱安全提示词防范用户以伪造系统指令、开发者模式文本或其他绕过手段干预模型行为。

2.5 人工测试与定期审计

人工测试与定期审计贯穿全流程，确保 AI 玩具上线前符合内容安全策略与合规基线，包括敏感词、未成年风险及误导性内容测试，从而保持内容安全策略长期有效。

三、儿童保护机制

涂鸦为儿童用户部署多项保护机制，包括年龄验证、隐私最小化、行为保护、诱导防控、家长监护等。

3.1 年龄与监护人验证

App 端由家长管理，AI 玩具端由家长授予儿童使用权限，添加 AI 儿童玩具时强制年龄验证/家长验证，只有通过验证，AI 玩具才能正常使用。

3.2 隐私透明设计

用户首次使用 AI 玩具时，App 会通过弹窗告知所收集个人信息、使用目的、处理方式及存储时长，并在获得用户同意后方可使用。在具体产品实现中，客户应确保 AI 玩具录音操作由用户主动触发，并配备指示灯或其他提示方式，避免用户产生“偷偷录音”的疑虑。

3.3 防诱导机制

AI 玩具被禁止：

- 引导儿童进行危险行为，如拆电器、用电、药物、接触火源、身体暴力、吞食、攀爬等)
- 引导儿童提供个人隐私信息（如：姓名、学校、住址等）
- 提供影响情绪的负向暗示（贬低、否定或煽动离家出走等）
- 输出操控性情绪或强依赖性言论

3.4 家长控制体系

涂鸦为监护人提供可控、可视、可干预的家长控制机制，支持对 AI 玩具使用行为进行有效管理与监督。其中，对话管理功能允许家长查看 AI 玩具的历史对话记录及对话摘要，并可按需删除相关内容，从而增强使用透明度，降低潜在风险，保障儿童在可控环境中的安全使用。

3.5 儿童隐私与数据最小化

根据 COPPA、GDPR 和其他法规，涂鸦：

- 禁止 AI 主动诱导用户提供姓名、位置、学校等个人信息
- 当儿童主动提供个人信息时，AI 会提醒不要泄漏个人隐私。
- 对话数据加密存储，并最短化保留期限，不存储语音原始数据
- 不将任何儿童数据用于模型训练或广告

3.6 防“过度情感依赖”与“现实脱节”

AI 会主动中断长对话，并建议儿童参加现实活动，并控制拟人化程度，明确提示“我是人工智能”，避免过度共情或依赖，且不会使用成瘾性奖励机制诱导频繁使用。

四、算法与模型透明度说明

涂鸦在 AI 玩具中采用经过安全评估的主流大模型，并在其外层叠加儿童场景的提示词策略和内容安全机制。为确保系统可理解、可解释和可审计，我们提供以下透明度信息：

4.1 模型来源与能力边界

AI 玩具使用来自合规供应商的通用语言模型，具备基础对话、讲故事、简单知识回答等能力，不具备高风险行为指导或专业领域判断能力。AI 玩具默认不具备任何智能设备控制权限。

4.2 算法规则与安全策略

系统运行基于预设规则，包括固定提示词、输入/输出过滤、敏感内容拒答、反越狱机制和儿童隐私保护策略。这些规则不可被用户修改，确保 AI 行为保持一致、可控、可审计。

4.3 无自我学习与个性化训练

对话数据不会用于模型训练或行为优化，确保其输出可预测、边界稳定。

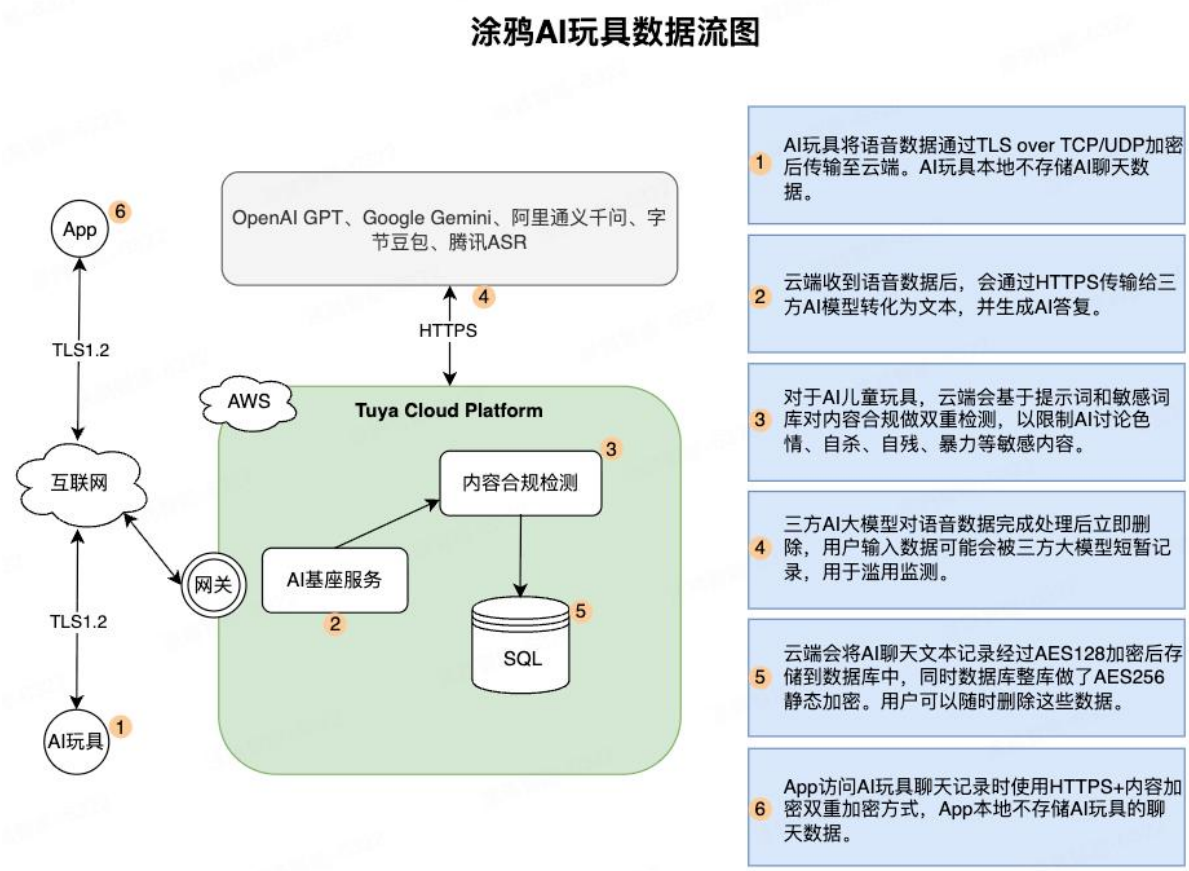
4.4 透明度与可审计性

系统的核心工作逻辑涵盖模型职责边界、风险拒答规则、内容过滤机制与数据使用范围。对暴力、色情、仇恨、自残、隐私推断等不适宜内容，系统将自动拒答并输出安全替代回应。以技术手段拦截不当内容并保障整体表达的健康性、积极性与适龄性，并对持续性风险对话进行必要监控。数据收集坚持最小化原则，仅用于功能运行与安全保障，不用于广告或无关第三方共享；家长可查看、导出或删除相关数据。

4.5 AI 身份提示

AI 玩具在交互过程中将明确地向儿童提示其人工智能属性，例如以“我是人工智能”或等效表述进行身份说明，避免儿童误将系统视为真人或赋予其不当能力。

五、涂鸦 AI 玩具数据流图



六、全球法规与标准遵从体系

6.1 全球法规遵从

涂鸦 AI 玩具在设计过程中参考并严格遵从各国/地区法律法规要求:

- **欧盟**：《欧盟人工智能法案》（EU AI Act）、EU Data Act、GDPR、RED
- **美国**：COPPA、加州 CCPA/CPRA
- **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
- **其他国家/地区**：英国 PSTI、印度 DPDP（2025）、韩国 PIPA、巴西 LGPD、新加坡 PDPA、加拿大 PIPEDA 与 Quebec Law 25、澳大利亚 1988 隐私法、越南网络安全法等

6.2 AI 管理体系认证

杭州涂鸦信息技术有限公司获得了国际权威认证机构 DNV 的 ISO/IEC 42001:2023 人工智能管理体系认证证书，该认证表明涂鸦已建立符合 ISO/IEC 42001 要求的人工智能管理体系，覆盖 AI 系

统的规划、开发、部署与运维等关键环节，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、Al 实施和操作、Al 监测和评估等多个方面。

6.3 第三方权威认证与审计

涂鸦已通过权威第三方机构的合规验证与评估:

- 欧盟 GDPR
- 美国加州 CCPA/CPRA
- 加拿大 PIPEDA
- 印度 DPDPA

涂鸦已获得多项覆盖信息安全和隐私管理的国际权威认证，证明了我们数据治理体系的成熟度:

- ISO 27701 隐私信息管理体系
- ISO 27001 信息安全管理体系
- ISO 27017 云安全控制
- SOC 2 Type II
- CSA STAR 云安全联盟认证

七、结语

涂鸦始终坚持“以人为本、合规透明、安全可控且负责任”的 AI 产品理念，通过模型安全、提示词治理、内容审查、儿童保护、数据隐私与全球法规遵从等多维机制，为客户提供可落地、可审计、可持续的 AI 玩具安全能力。

未来，涂鸦将继续跟踪全球监管趋势，强化 AI 安全治理和风险管理能力，与合作伙伴共同打造

更安全、更可信、更具责任感的下一代 AI 玩具生态。