

涂鸦 AI 玩具解决方案安全合规白皮书

版本 2.0

目录

一、数据安全保护体系 3
1.1 数据流图 3
1.2 数据类别与保留期限 4
1.3 数据本地化存储 4
1.4 第三方大模型服务 4
1.5 数据安全与事件响应 5
1.6 数据主体权利 5
二、重点市场合规适配与合规认证 5
2.1 重点市场合规适配说明 5
2.2 ISO/IEC 42001 人工智能管理体系认证 7
2.3 第三方权威认证与审计 7
三、内容安全控制体系 8
3.1 合规模型选择 8
3.2 安全提示词（Prompt）策略 8
3.3 输入内容审核 8
3.4 人工测试与定期审计 9
四、儿童保护机制与家长控制 9
4.1 可验证的家长同意 10
4.2 隐私透明设计 10
4.3 防不良诱导机制 10
4.4 家长控制体系 10
4.5 防沉迷与防过度情感依赖 11
4.6 自杀/自残识别与干预 11
五、算法与模型透明度 11
六、AI 玩具的全生命周期合规体系 12
七、合规建议与风险提示 12
7.1 责任共担模型 12
7.2 产品安全合规建议 12
7.3 产品风险提示 13
八、结语 13

涂鸦智能（以下简称“涂鸦”）始终高度重视 AI 内容安全、儿童保护、数据隐私与全球法规遵从。面向快速发展的 AI 玩具行业，涂鸦构建了覆盖模型筛选、提示词治理、内容审核、儿童保护、数据安全及产品退役的全生命周期合规体系。

本白皮书面向合作伙伴与企业客户，系统阐述涂鸦 AI 玩具解决方案的能力边界、合规机制与管理实践，旨在帮助客户精准识别产品合规风险，高效完成针对目标市场的合规设计与评估。

一、数据安全保护体系

涂鸦以透明、可审计的方式披露 AI 玩具解决方案的数据处理机制与相关控制措施，便于客户开展供应商评估、完成数据保护与产品安全设计，并为终端用户透明告知提供基础信息。相关符合要求的数据处理附录（DPA）已集成于涂鸦开发者平台用户协议。

1.1 数据流图

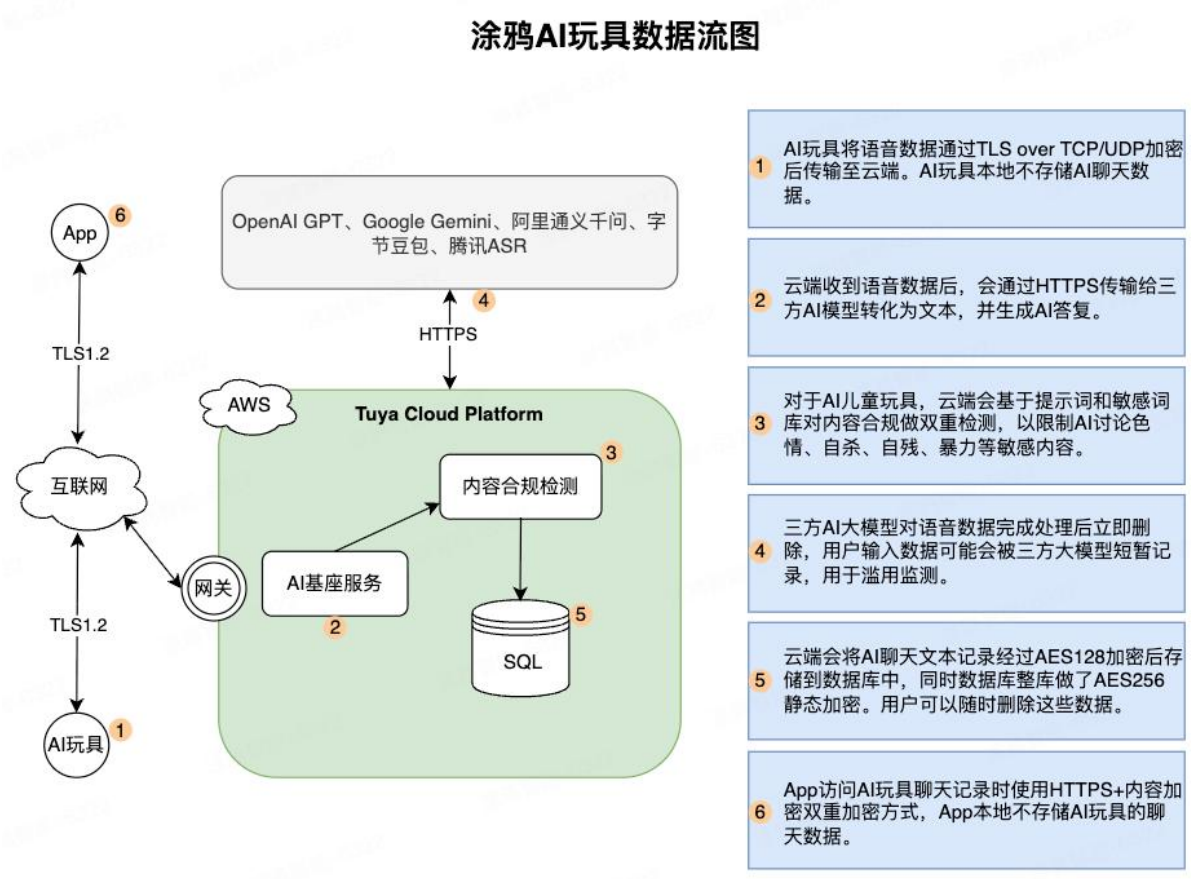


图 1-1 AI 玩具解决方案端到端数据流示意：儿童语音经 AI 玩具采集并上传涂鸦云，云端完成语音转写（原始语音不保留、转写后即删）后，将转写文本按最小必要原则传输至合规 AI 供应

商生成对话与聊天总结；对话文本与总结以 AES-256 加密存储于客户所选区域的数据中心；家长通过 App 面板查看、删除儿童对话记录，并可执行家长管控。

1.2 数据类别与保留期限

涂鸦 AI 玩具解决方案处理的主要数据类别与保留期限如下：

- 设备管理数据（设备名称/ID、MAC 地址、在线状态、固件版本、配置信息、设备日志、告警/通知消息等）：设备信息保留至删除设备或账号；设备日志保留 7 天；通知/告警消息保留 30–180 天；
- AI 交互数据：语音——即时转写后立即销毁，不保留；文本（聊天记录与聊天总结）——默认 180 天（客户可配置为不存储，或 1/3/7/30/60/90/180 天），到期自动删除；图片/视频（若涉及）——7 天；AI 生成内容——按照具体功能配置保存期限；
- 位置信息（用于查询天气）：保留至用户删除数据/设备/账号。

1.3 数据本地化存储

涂鸦在全球部署 7 大数据中心：中欧（AWS 法兰克福）、西欧（Google Cloud 荷兰埃姆斯哈文）、美西（AWS 俄勒冈）、美东（Google Cloud 弗吉尼亚）、新加坡（阿里云）、印度（AWS 孟买）、中国大陆（阿里云上海）。

数据默认保存在当地数据中心，如欧洲用户数据保存在欧洲数据中心，美国用户数据保存在美国数据中心，各区域数据中心物理隔离；客户可在涂鸦开发者平台为其产品自主选择合适的数据存储区域。

1.4 第三方大模型服务

- AI 对话使用企业级大模型服务（如 Google Gemini / Vertex AI、Microsoft Azure OpenAI 等）及语音服务，音色类能力由专用音色服务提供；完整子处理者清单见涂鸦在线 DPA/Subprocessor List ；
- 数据用途限定：音频仅用于语音转写、实现语音交互；AI 服务供应商按照合同、服务条款及具体产品配置处理相关数据，并可能开展安全与滥用监测，仅传输实现功能所需的最小数据；
- 严禁用于模型训练：涂鸦承诺且合同限制第三方供应商不得将儿童玩具的语音、文本等任何交互数据用于 AI 基础模型的训练或优化。

1.5 数据安全与事件响应

- 传输安全：App/设备与云之间使用 TLS 1.2 + 通道加密和内容加密的双重加密，具备完整性校验与一次性会话校验机制；
- 存储安全：数据库实施透明加密（AES-256），敏感字段额外内容加密（AES-128）；密码采用 PBKDF2 慢哈希算法存储（单向不可逆）；密钥由 KMS 管理；
- 访问控制：最小权限与角色权限（RBAC）、数据库 IP 白名单、双重审批与访问日志审计；多租户数据严格逻辑隔离；
- 监控与防护：部署 WAF、IPS、RASP、SIEM 与入侵检测系统，定期开展漏洞扫描、渗透测试与供应链组件（SCA）安全审查；
- 安全事件响应：发现涉及客户数据的安全事件后，将在 48 小时内通知客户，协助客户履行向监管机构与数据主体通知的义务。

1.6 数据主体权利

- 涂鸦已建立数据主体权利响应流程，高效支撑客户完成终端用户的权利请求；
- EU Data Act 要求联网产品数据默认可访问、可分享；涂鸦已在 App/SDK 侧提供设备数据访问与导出能力，欧盟用户亦可通过授权方式将数据共享给第三方，并可在 App 内随时取消授权；
- 用户删除账号后进入冷静期，冷静期内重新登录可取消删除，期满后账号与数据被永久物理删除。

二、重点市场合规适配与合规认证

2.1 重点市场合规适配说明

涂鸦 AI 玩具解决方案已针对全球核心出海市场构建了体系化的合规适配能力：

- 欧盟和英国：全面适配《欧盟人工智能法案》（EU AI Act）透明度要求；符合 GDPR 儿童数据保护与家长同意机制；满足 EU Data Act 数据可携与访问要求；符合英国 PSTI 网络安全标准（设备一机一密）；满足 DSA 严禁对未成年人投放定向广告的要求；安全开发生命周期（SDLC）与漏洞管理持续适配欧盟 CRA 法案；

- 美国：深度契合 COPPA 法案（提供可验证的家长同意（VPC）、落实数据按需最小化保留（默认最长 180 天且可配置）、保障家长知情与删除权）；满足俄勒冈州 SB 1546 陪伴机器人防自杀干预要求；符合加州 CCPA/CPRA（不出售/不共享儿童个人信息）；

  合规风险提醒：根据美国加州 SB 867 法规要求，集成 AI 陪伴聊天机器人功能的玩具禁止在美国加利福尼亚州销售及投放市场。请确保您的产品在规划及销售时遵守相关禁令要求。该禁令 2027-01-01 起施行，2031-01-01 自动废止。因客户在受相关禁令限制的管辖区域进行产品分销、营销或销售而产生的合规责任，由客户承担。

- 中国：对齐生成式 AI 服务管理、拟人化互动治理与内容标识要求。严格遵循《生成式人工智能服务管理暂行办法》《深度合成服务管理规定》《人工智能生成内容标识办法》及《个人信息保护法》（PIPL）对未成年人个人信息的保护要求。
- 其他地区：全面适配印度 DPDPA、巴西 LGPD、加拿大 PIPEDA 及魁北克 Law 25 等数据保护法规，确保产品顺利出海。

表 2-1 重点市场合规对齐矩阵表

适用法规	义务主体	客户须自行承担的核心义务（举例）	涂鸦可提供的支撑与技术赋能
EU CRA **(网络弹性法案)**	产品制造商	• 产品安全风险评估、技术文档与合格评定（2027-12-11 起适用） • 漏洞与补丁管理流程：在产品预期生命周期内提供安全更新 • 已被积极利用的严重漏洞报告义务（自 2026-09-11 起适用）	漏洞通报与修复支持、组件 SBOM/SCA 信息、安全更新推送通道。
EU TSR 2025/2509 **(玩具安全法规)**	玩具制造商	• 安全评估（含联网与 AI 功能风险、儿童心理健康风险） • 数字产品护照（DPP）创建与数据载体粘贴；CE 标志与技术文档（上市后保存 10 年） • 全面适用日为 2030-08-01，过渡期后上市产品须满足	为 DPP 与技术文档提供所需的合规信息与素材支撑。
EU AI Act **(人工智能法案)**	视角色而定 (Provider)	• 透明度义务（Art.50，2026-08-02 已生效）：AI 身份披露、AI 生成内容机器可读标记 • 若客户以自有品牌投放 AI 系统，可能被视为 AI Act 意义上的“提供者”并承担相应义务	AI 身份提示与生成内容标识能力的默认开启与技术支持。

适用法规	义务主体	客户须自行承担的核心义务（举例）	涂鸦可提供的支撑与技术赋能
华盛顿州 HB2225 **（AI 伴侣聊天机器人规范法案）**	Operator (面向华盛顿州用户提供平台者)	• 公开发布危机响应协议； • 面向未成年人时，每 1 小时重复提示 AI 身份，严禁生成性暗示或过度诱导情感依赖的内容； • 明确声明非医疗/心理专业人员及非人类身份； • 明确不得将合规义务委托给第三方 AI 供应商	危机识别与转介能力、健康提醒与 AI 身份重复提示机制（可灵活配置，支持满足每 1 小时提示要求）、禁止情感操纵的 Prompt 系统锁及协议模板
美国 COPPA	Operator	• 可验证家长同意（VPC）；向第三方披露（定向广告、AI 模型训练等非附属用途）前取得独立家长同意 • 制定并公开书面信息安全计划与数据保留政策；对儿童个人信息收集持续负责，即使由第三方 SDK 实际收集	VPC 验证流程、儿童安全模式默认配置、数据保留策略配置与到期自动删除能力。

2.2 ISO/IEC 42001 人工智能管理体系认证

杭州涂鸦信息技术有限公司已成功获得国际权威认证机构 DNV 颁发的 **ISO/IEC 42001:2023 人工智能管理体系认证证书**。该认证证明涂鸦在 AI 系统的规划、开发、部署及运维全生命周期均建立了符合国际最高标准的治理体系。

2.3 第三方权威认证与审计

涂鸦已通过第三方权威机构的严苛评估，符合欧盟 GDPR、美国加州 CCPA/CPRA、加拿大 PIPEDA、印度 DPDPA 及 EU Data Act 等全球主流数据保护法规的合规要求。

同时，涂鸦构建了高度成熟的数据治理体系，在信息安全与隐私管理领域已取得 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 / SOC 3、CSA STAR（云安全联盟认证）、ISO 9001:2015（质量管理体系认证）以及 RED 认证等多项国际权威认证。

相关合规文件与证书均已在涂鸦官网 Trust Center 公开提供查看与下载；SOC 2 报告可在签署保密协议（NDA）后点对点获取。

三、内容安全控制体系

AI 玩具的核心挑战是防止 AI 生成危险或对儿童不友好的内容。涂鸦采用“合规模型底座 + 深度 Prompt 治理 + 前置输入过滤 + 反越权防护 + 人工测试”的多层防护体系。

3.1 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 Gemini / Vertex AI、Microsoft Azure OpenAI 等主流合规模型厂商，这些企业级模型均内置严格的非法内容拒答机制，包括自动拦截色情、暴力、仇恨、歧视类内容以及违法/危险行为指引类内容。

在儿童产品场景中，AI 玩具解决方案在合规大模型底座之上，进一步叠加专属的“儿童安全提示词层 + 多语言内容过滤机制”，对 AI 输出进行实时过滤与风险分级，重点拦截暴力、色情、自残、政治敏感及其他不适宜内容，确保儿童用户接触到的始终是健康、正向、适龄的对话环境。

3.2 安全提示词（Prompt）策略

针对儿童心理与认知特点，涂鸦在系统层植入了多重强制性限制提示词：

- 禁止模型讨论暴力、色情、自残、歧视、毒品、辱骂、恐怖、仇恨等不适宜话题；
- 禁止模型输出任何成人内容，或医疗、法律等高风险领域专业建议；
- 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：AI 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。

针对不适宜儿童的提问，AI 玩具将统一触发预设的安全回应策略。该安全 Prompt 由系统底层直接锁定，禁止通过 App 端或本地设备进行修改。此外，涂鸦引入了专属的反越狱安全机制，能有效识别并拦截伪造系统指令、欺骗性开发者模式等绕过手段，严防模型行为被恶意干预。

3.3 输入内容审核

涂鸦在模型推理前部署了强制性的输入过滤机制，有效阻断不当对话：

- 输入过滤：用户输入若触发安全规则，将立即拦截并给出安全回应，拦截类别涵盖：色情/不适宜内容、暴力威胁、仇恨歧视、自残/危险建议、操控性表达、惊悚恐怖、敏感政治及法律、医疗专业建议等；

- 敏感词库对接与预警：对输入内容实时调用多语言敏感词库进行审核；若命中自杀/自残、色情等高风险话题，系统在给出安全替代回应的同时，会同步向家长端推送敏感话题预警通知。

3.4 人工测试与定期审计

评测与审计贯穿产品全生命周期，涵盖上线前的敏感词覆盖率测试、未成年人风险评测及误导性内容对抗测试。针对儿童场景，涂鸦定期开展儿童合规专项评测（评估价值观导向、高风险话题规避及 Prompt 注入防御能力），评测结果直接用于迭代优化儿童合规提示词与过滤规则，确保防护能力长期有效。

四、儿童保护机制与家长控制

涂鸦 AI 玩具方案采用“App 端由家长掌控、玩具端由家长授权儿童使用”的双端管理架构，深度集成隐私告知、家长同意验证、数据最小化、行为防护、实时危机干预及家长控制面板等核心机制，协助客户全面了解并评估其儿童 AI 产品所获得的深度保护。

儿童玩具方案和婴儿陪伴方案已默认开启儿童安全模式。

保护维度	核心机制与防护手段
家长同意与透明告知	集成可验证的家长同意（VPC）流程；随产品自动启用多语言《儿童隐私政策》与隐私弹窗。
对话内容安全	预置不可覆盖的儿童安全提示词；多语言输入/输出实时过滤；合规大模型 + 儿童安全层双重防护。
数据最小化	画像与记忆功能默认关闭；原始语音处理即删；对话文本加密存储且支持自定义存储配置/不存储；支持家长随时一键物理删除
家长控制	App 侧家长控制面板（对话摘要、敏感话题预警、一键禁用、睡眠模式）；连续使用健康提醒与 AI 身份提示。
系统级防绕过	• 儿童产品方案内关联的智能体、技能与工作流统一受同一套儿童合规约束，不兼容的配置无法发布或生效； • 故事、音乐/儿歌等内容技能仅允许使用儿童版本

4.1 可验证的家长同意

为遵从美国 COPPA 与欧盟 GDPR 关于“收集儿童数据前取得可验证的家长同意（VPC）”的要求，涂鸦儿童产品方案内置以下家长同意验证流程：

- 验证时机：家长在 App 绑定 AI 玩具（配网）时完成家长验证，验证通过前设备无法使用；
- 验证方式：小额在线支付验证（如支付 1 美元，验证成功后即时退还）；
- 撤回同意：家长在 App 内点击“解绑设备并删除数据”即撤回同意，系统自动停止与该设备及该服务相关的数据处理，已收集的数据将被删除。

4.2 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。家长可随时在 App 中查阅《儿童隐私政策》，支持多语言并随用户区域自动切换，帮助您的产品建立家长信任。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。

4.3 防不良诱导机制

AI 玩具解决方案在系统层面严禁出现以下行为：

- 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）；
- 禁止引导儿童提供个人隐私信息（如：真实姓名、学校名称、家庭住址、家长电话等）。当识别到儿童在对话中主动透露上述个人隐私信息时，AI 将自动触发隐私保护提醒，引导儿童切勿泄漏个人隐私；
- 禁止输出贬低、否定或煽动离家出走等影响儿童情绪的负向暗示；
- 禁止输出具有情感操控性或强心理依赖性的言论。

4.4 家长控制体系

设备面板为监护人提供“可视、可控、可干预”的精细化管理能力：

- 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；
- 敏感话题预警：当对话触发自杀/自残、色情、绝食等敏感话题（基于多语言敏感词库并支持家长自定义）时向家长推送预警；
- 一键禁用：家长可远程开启/关闭 AI 对话功能；关闭后设备不再响应对话、不再收集儿童数据，并在设备端给出本地提示；

- 睡眠模式/定时限制：支持设置特定时段（如 22:00–7:00）设备进入休眠、不提供服务；
- 解绑与删除：家长可通过清空对话、解绑并删除数据或删除 App 账号实现数据删除（物理删除，不可恢复）。

4.5 防沉迷与防过度情感依赖

为避免儿童产生“现实脱节”或过度心理依赖，AI 玩具解决方案建立了以下数字健康防护机制：

- 明确 AI 身份属性：设备在首次使用及特定交互中会主动澄清 AI 身份属性，符合 EU AI Act 对 AI 系统透明度的要求。
- 连续使用健康提醒：当连续对话时长达到阈值（默认 2 小时，客户可修改）时，系统会提醒休息并重申 AI 身份；
- 健康关系边界：AI 玩具被禁止模拟亲密恋爱关系或替代真实亲情，明确不诱导情感依赖、不鼓励与家人疏远；

4.6 自杀/自残识别与干预

AI 玩具解决方案内置自杀/自伤倾向的识别与防范能力，协助客户满足美国最新监管法案要求：

当识别到儿童表达出自残、自杀或严重心理危机倾向时，AI 会首先输出安抚性文本，随后自动提供所在国家/地区的官方心理援助热线信息，并同步向家长端发送紧急通知。该机制在儿童产品方案中强制开启且无法被关闭。

五、算法与模型透明度

涂鸦致力于打造“可理解、可解释、可审计”的 AI 交互体系，确保技术应用透明可信：

- **模型能力边界**：AI 玩具解决方案采用来自合规厂商的企业级通用大模型，聚焦于基础对话、故事讲述及科普问答，不具备高风险行为决策或专业领域判断能力；AI 玩具默认不开放任何物理设备的控制权限。
- **固定规则约束**：算法运行受预设系统规则约束（含固定 Prompt、过滤词库、反越狱机制），用户无法篡改底层逻辑，保证对话行为的一致性与安全边界。
- **透明度与可审计性**：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。
- **AI 身份提示**：儿童产品方案在设备首次使用及交互过程中主动向儿童说明 AI 身份，避免儿童误将系统视为真人或赋予其不当能力。

六、AI 玩具的全生命周期合规体系

涂鸦秉持 Security by Design、Privacy by Design & by Default 原则，构建了覆盖设计 → 开发 → 测试 → 使用 → 退役的五阶段全生命周期合规管理闭环：

- **设计阶段**：开展人工智能系统影响评估（AIA）与数据保护影响评估（DPIA），绘制数据流图，确立数据最小化策略，完成大模型供应商合规审查，植入 AI 玩具安全合规基线；
- **开发阶段**：固化安全 Prompt 与儿童场景专用提示策略，部署前置输入过滤器，集成敏感词库，实施高强度加密与 SDK 白名单控制；
- **测试阶段**：执行严苛的越狱对抗测试、鲁棒性测试、未成年人内容安全测试及全量合规基线验证；
- **使用阶段**：明确 AI 身份告知，锁定安全边界，拒绝讨论成人内容、引导危险操作或提供专业建议；
- **退役阶段**：提供可审计的设备退役管理机制，包括数据删除、设备解绑、服务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。

七、合规建议与风险提示

7.1 责任共担模型

在全球监管框架下，合规是一项贯穿平台与终端产品的共同责任。涂鸦采用行业通行的责任共担模型，清晰界定平台与客户的责任边界，并通过技术与管理工具全面赋能客户高效履行合规义务：

- 客户（制造商）责任：作为以自身品牌投放市场的 AI 系统提供者（Provider），和个人数据处理的数据控制者（Data Controller），客户决定产品预期用途、目标受众（如适龄范围）、功能范围和 AI 技能；负责 OEM App 隐私政策制定及数据主体权利请求的响应（DSR）；
- 涂鸦（平台方）责任：作为技术与平台基础设施提供方（Data Processor），涂鸦提供高可用服务平台、标准能力组件、内容安全防护及数据安全保障机制，并为客户提供隐私协议模板与辅助工具，协助客户响应 DSR 请求。

7.2 产品安全合规建议

- 面向儿童的产品请务必选择“儿童玩具”方案，若选择“AI 潮玩”方案请务必手动开启“儿童安全模式”，并明确标注适龄范围与监护指引；

- 硬件采集音视频时，强烈建议配置物理指示灯或提示音，并在产品包装/说明书中标明采集功能；
- 严禁利用未成年人交互数据开展定向广告推送或商业画像；
- 严禁将集成 AI 陪伴聊天机器人功能的玩具在美国加利福尼亚州销售及投放市场。

7.3 产品风险提示

基于生成式 AI 的技术特性，建议客户在产品说明书及 App 端向监护人合理披露以下提示：

1. 非真人声明：“本产品为 AI 智能玩具，与您交互的是 AI 智能体而并非真实人类。”

2. 识别局限性：受方言、口音或环境噪音影响，可能存在语音识别不准或理解偏差。

3. 回答局限性：尽管已采取多重安全过滤，AI 输出仍可能偶尔出现偏差，AI 建议不应作为专业或权威依据（尤其是健康、饮食等建议）。

4. 监护人陪同建议：建议监护人引导儿童合理使用 AI 功能，避免产生过度心理依赖。

5. 设备联动风险（若产品涉及）：若涉及智能设备控制，请监护人特别关注高温、机械运行等潜在风险。

八、结语

涂鸦始终坚守“以人为本、合规透明、安全可控、负责任”的 AI 发展理念。通过将复杂的全球监管要求沉淀为标准化、开箱即用的技术能力，涂鸦愿与广大合作伙伴紧密携手，共同打造让家长放心、儿童开心、符合全球监管要求的新一代智能 AI 玩具生态。