Tuya AI Toy Compliance White Paper v1.1

Tuya places the highest priority on AI content safety, child protection, personal data privacy, and global regulatory compliance. In response to the rapid development of the AI toy industry, Tuya has established a full-lifecycle compliance framework covering model selection strategies, prompt governance, product logic design, content moderation, child protection, security safeguards, data management, and product decommissioning. This framework is designed to support Tuya's AI toy products in meeting global expectations for safety, transparency, and trustworthiness.

This white paper is intended for Tuya's partners and provides a systematic overview of Tuya's compliance capabilities in the areas of model security, content safety, data privacy, child protection, parental controls, and global regulatory compliance. It addresses key industry concerns related to AI-generated content risks, child protection compliance, algorithmic transparency, and data security in AI toys.

---

1. Full-Lifecycle Compliance Framework for AI Toys

Guided by the principles of **Security by Design** and **Privacy by Design & by Default**, Tuya has established a comprehensive AI toy compliance framework spanning the entire product lifecycle—from design, development, and testing to use and decommissioning. This framework supports end-to-end control over functional behavior, data processing, content generation, and interaction safety.

1.1 Design Phase

During the design phase, Tuya conducts AI toy risk assessments, Artificial Intelligence Impact Assessments (AIA), and Data Protection Impact Assessments (DPIA). Data flows are mapped to identify processing boundaries, and strict data minimization and de-identification measures are applied. Tuya also evaluates the security and compliance capabilities of large model providers to ensure the reliability of model sources. AI toy safety and compliance baselines are incorporated into product requirements, functional design, and voice interaction logic, laying a solid foundation for subsequent development and operational phases.

1.2 Development Phase

During development, Tuya implements secure prompts and child-specific prompt strategies to prevent tampering. Sensitive content restrictions are enforced on the input side, while child-friendly expression constraints are applied to outputs. Integrated content detection mechanisms are used alongside strict enforcement of data minimization, encryption, and least-privilege access principles. An SDK whitelist mechanism is adopted to ensure that data flow paths remain controllable and traceable.

1.3 Testing Phase

Tuya's conversational systems undergo sensitive content testing, jailbreak testing, and robustness testing, as well as assessments covering inappropriate content for minors and child psychological safety. Security adversarial testing is also conducted at this stage. All functions are validated against AI toy safety and compliance baselines to support compliance with safety requirements prior to product release.

1.4 Usage Phase

Tuya clearly informs children and guardians that the AI system is not a real person and provides transparency disclosures regarding algorithm logic and model capability boundaries. AI toys operate within fixed behavioral boundaries and refuse to engage in adult content, dangerous operations, or the provision of professional advice, thereby supporting a safe and controllable environment for child users.

1.5 Decommissioning Phase

Tuya provides auditable device decommissioning mechanisms, including data deletion, device unbinding, feature deactivation, and account deletion, to mitigate security risks after the end of an AI toy's lifecycle.

This framework applies to Tuya's full range of AI toy products and supports the safety, transparency, and trustworthiness of children's usage environments.

---

2. Content Safety Control Framework

A core challenge for AI toys is preventing the generation of inappropriate, harmful, or child-unfriendly content. Tuya adopts a multi-layered protection approach combining **model safety, prompt restrictions, input filtering, anti-bypass mechanisms, and human testing**.

2.1 Compliant Model Selection

Tuya selects only mainstream large model providers that have passed security assessments, including—but not limited to—Google, Microsoft, ByteDance, and Alibaba. These models incorporate built-in mechanisms to reject illegal content, including:

- Automatic refusal of pornographic, violent, hateful, or discriminatory content
- Rejection of instructions involving illegal or dangerous activities

These models also support multilingual content safety. Tuya further overlays an additional "child safety layer" to provide dual-layer protection.

2.2 Secure Prompt Strategy

Based on child-specific usage scenarios, Tuya applies multiple restrictive prompt rules, including but not limited to:

- Prohibiting discussions of violence, pornography, self-harm, discrimination, drugs, abuse, terrorism, or hatred
- Prohibiting adult content or professional advice (e.g., medical or legal guidance)
- Prohibiting dangerous operational instructions, manipulation, coercion, or emotional influence over children
- Fixed role locking: limiting the AI to child-friendly roles such as "educational companion"
- Child-friendly expression: maintaining a positive, gentle, and safe tone

For questions unsuitable for children, AI toys adopt standardized safe response strategies to maintain safety while preserving interaction quality, for example:

- "I can't talk about that, but I can play a game with you."
- "I can't answer that, but we can talk about something more fun!"
- "That's an adult topic. Let's tell a story instead!"

These prompts are system-locked and cannot be modified by users or local devices, preventing jailbreak attempts.

2.3 Input Content Review

Tuya enforces mandatory filtering on user inputs prior to model invocation to block inappropriate interactions. When filtering rules are triggered, a unified safe response strategy is applied. Filtered content includes:

- Pornographic or child-inappropriate content
- Violent or threatening descriptions
- Hate speech, discrimination, or abuse

- Self-harm or dangerous operation suggestions
- Excessive emotional manipulation or coercive expressions
- Horror or frightening content
- Sensitive political topics
- Legal, medical, or other professional advice

2.4 Anti-bypass Mechanisms

Tuya implements anti-bypass mechanisms in AI toys through secure anti-jailbreak prompt strategies to prevent users from interfering with model behavior via forged system instructions, developer-mode text, or other circumvention techniques.

2.5 Human Testing and Periodic Audits

Human testing and periodic audits are conducted throughout the lifecycle to verify compliance with content safety strategies and baselines prior to product release. These include sensitive content tests, minor risk assessments, and misleading content evaluations to maintain the long-term effectiveness of content safety controls.

---

3. Child Protection Mechanisms

Tuya deploys multiple protection measures for child users, including age verification, privacy minimization, behavioral safeguards, inducement prevention, and parental supervision.

3.1 Age and Guardian Verification

The app is managed by parents, and children's access to AI toys must be authorized by guardians. Age verification and parental verification are mandatory when adding an AI toy, and only verified devices may be used.

3.2 Privacy Transparency by Design

Upon first use, the app informs users via pop-up notices of the personal data collected, purposes of use, processing methods, and retention periods. In specific product implementations, customers are required to ensure that audio recording is user-initiated and accompanied by indicator lights or equivalent prompts to avoid covert recording concerns.

3.3 Anti-inducement Safeguards

AI toys are prohibited from:

- Encouraging dangerous behaviors (e.g., dismantling appliances, electrical use, medication misuse, fire exposure, physical violence, ingestion, or climbing)
- Inducing children to provide personal information (e.g., name, school, address)
- Delivering negative emotional cues (e.g., belittlement, denial, or encouraging running away)
- Producing manipulative or dependency-inducing emotional expressions

3.4 Parental Control Framework

Tuya provides guardians with controllable, visible, and intervention-capable parental controls to manage and supervise AI toy usage. Conversation management features allow parents to view conversation histories and summaries and delete content as needed, enhancing transparency, reducing potential risks, and supporting a safe usage environment.

3.5 Child Privacy and Data Minimization

In accordance with COPPA, GDPR, and other applicable regulations, Tuya:

- Prohibits AI from actively inducing children to provide personal information
- Reminds children not to disclose personal data when such information is voluntarily provided
- Encrypts conversation data with minimized retention periods and does not store raw audio data
- Does not use child data for model training or advertising

3.6 Prevention of Excessive Emotional Dependence

AI systems proactively interrupt prolonged conversations and encourage children to engage in real-world activities. Anthropomorphism is limited, and the system explicitly identifies itself as artificial intelligence to avoid excessive emotional reliance. No addictive reward mechanisms are used to encourage frequent usage.

---

4. Algorithm and Model Transparency

Tuya employs mainstream large models that have undergone security evaluations and overlays child-specific prompt strategies and content safety mechanisms. To support system understandability, explainability, and auditability, the following transparency information is provided.

4.1 Model Sources and Capability Boundaries

AI toys use general-purpose language models from compliant providers, capable of basic conversation, storytelling, and simple knowledge responses. They do not provide high-risk behavioral guidance or professional judgment and do not possess any default smart device control capabilities.

4.2 Algorithm Rules and Safety Policies

System operation is governed by predefined rules, including fixed prompts, input/output filtering, sensitive content refusal, anti-bypass mechanisms, and child privacy protection strategies. These rules cannot be modified by users, ensuring consistent, controllable, and auditable AI behavior.

4.3 No Self-learning or Personalized Training

Conversation data is not used for model training or behavioral optimization, ensuring predictable outputs and stable capability boundaries.

4.4 Transparency and Auditability

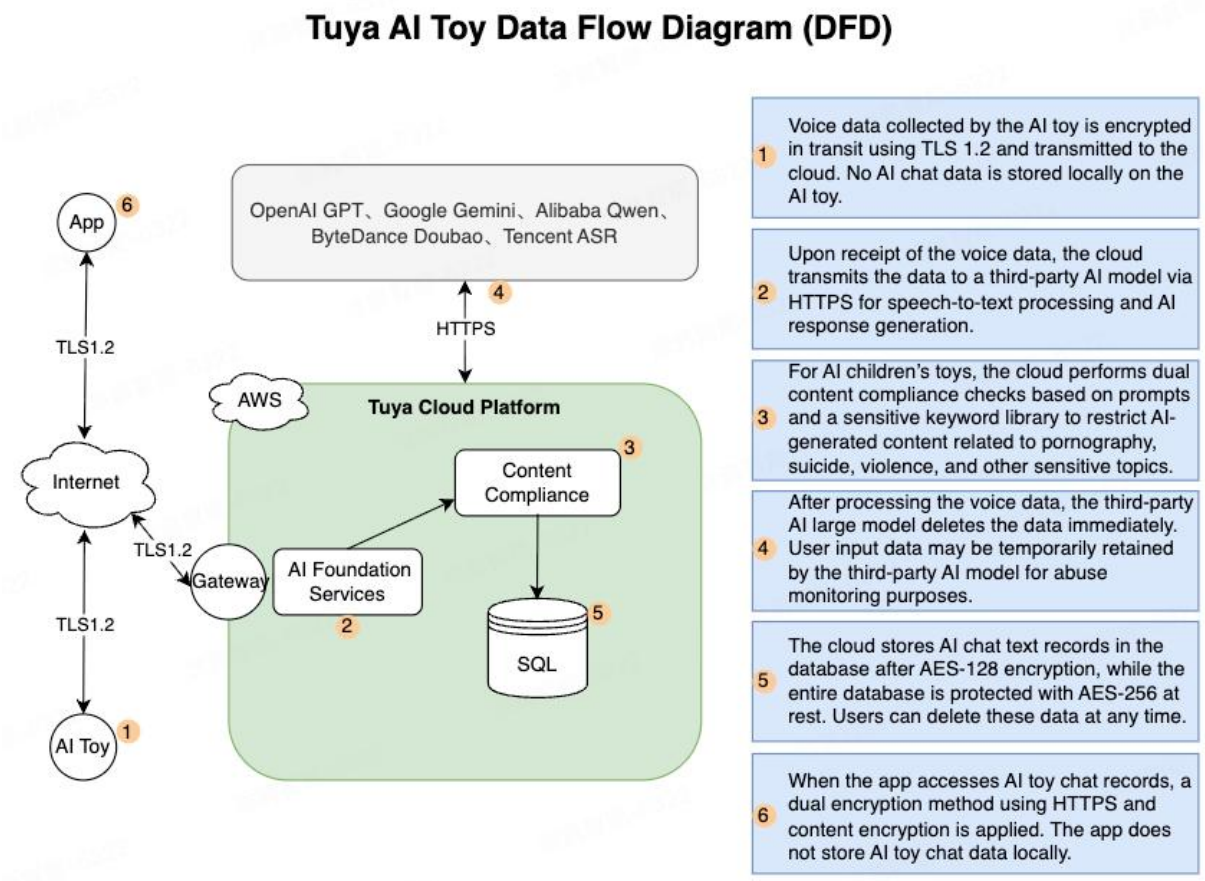
Core system logic covers model responsibilities, risk refusal rules, content filtering mechanisms, and data usage scopes. Inappropriate content such as violence, pornography, hatred, self-harm, or privacy inference is automatically rejected with safe alternative responses. Technical safeguards ensure healthy, positive, and age-appropriate outputs, with monitoring applied to sustained risk scenarios. Data collection follows strict minimization principles and is used solely for functionality and safety purposes, not advertising or unrelated third-party sharing. Guardians may view, export, or delete relevant data.

4.5 AI Identity Disclosure

During interactions, AI toys clearly disclose their artificial intelligence identity—such as through statements like “I am an AI”—to prevent children from mistaking the system for a human or attributing inappropriate capabilities.

---

5. AI Toy Data Flow Diagram



6. Global Regulatory and Standards Compliance

6.1 Global Regulatory Compliance

Tuya AI toys are designed with reference to and in compliance with applicable laws and regulations, including:

- **European Union:** EU Artificial Intelligence Act, EU Data Act, GDPR, RED
- **United States:** COPPA, California CCPA/CPRA
- **China:** AIGC Labeling Measures, Deep Synthesis Provisions, Generative AI Measures, Algorithm Recommendation Regulations, App Necessary Information Rules, Cybersecurity Law, Data Security Law, Personal Information Protection Law

- **Other Regions:** UK PSTI, India DPDP (2025), South Korea PIPA, Brazil LGPD, Singapore PDPA, Canada PIPEDA and Quebec Law 25, Australia Privacy Act 1988, Vietnam Cybersecurity Law, among others

6.2 AI Management System Certification

Hangzhou Tuya Information Technology Co., Ltd. has obtained ISO/IEC 42001:2023 Artificial Intelligence Management System certification from the internationally recognized certification body DNV. This certification indicates that Tuya has established an AI management system aligned with ISO/IEC 42001 requirements, covering key stages of AI system planning, development, deployment, and operation, providing a strong foundation for delivering safe, reliable, and trustworthy AI products and services globally.



ISO/IEC 42001 is jointly developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) and is intended to help organizations establish effective management systems for AI development, implementation, and maintenance, covering areas such as AI strategy, project management, system design, implementation, operation, monitoring, and evaluation.

6.3 Third-party Certifications and Audits

Tuya has passed compliance assessments by authoritative third-party institutions, including:

- EU GDPR
- California CCPA/CPRA
- Canada PIPEDA
- India DPDPA

Tuya also holds multiple internationally recognized certifications covering information security and privacy management, demonstrating the maturity of its data governance framework:

- ISO 27701 Privacy Information Management System
- ISO 27001 Information Security Management System
- ISO 27017 Cloud Security Controls

- SOC 2 Type II
- CSA STAR Certification

---

7. Conclusion

Tuya adheres to a human-centric, compliant, transparent, safe, controllable, and responsible AI product philosophy. Through integrated mechanisms covering model security, prompt governance, content moderation, child protection, data privacy, and global regulatory compliance, Tuya provides practical, auditable, and sustainable AI toy safety capabilities for its partners. Looking ahead, Tuya will continue to monitor global regulatory developments, strengthen AI safety governance and risk management, and work with partners to build a safer, more trustworthy, and more responsible next-generation AI toy ecosystem.