Trust Management for On-Device Agentic AI: The vivo Approach

Guangming Liu

Director of AI Security

vivo Mobile Communication Co., Ltd.

vivo at a Glance: Global Reach and Full-Stack Device Innovation

AI for Good

Global Users' Choice & Honors and Awards

Global Monthly Active Users
500 Million+

Countries & Regions
60+

Annual Production Capacity
200 Million

Top 20 Chinese Global Brands
2024 BrandZ Top50 Chinese Global Brands

2024 Inspirational Star of Innovation
Kantar BrandZ

Top 50 Most Valuable Chinese Brands
2024 Kantar BrandZ Most Valuable Chinese Brands

Global Talents
Team Members Come from Top Art and Design Institutions

Meticulous Design
Nearly 1,000 Prototypes are Tried Out for Each Work

vivo | ZEISS
Global Imaging Partners

vivo BlueImage
Encompasses self-developed sensor technology, imaging chips and algorithm matrix.

Self-Developed BlueLM
vivo's BlueLM Leads Chinese LLM Rankings on Various Global Lists*

Self-Developed BlueOS
Industry's 1st Rust Programming Language-Based Operating System

vivo BlueChip Tech.
Brings exceptional user experience in performance, gaming, power efficiency, etc.

vivo BlueVolt Battery
Super Slim, Super Large, and Super Fast Charging

vivo

From AIGC to Agentic AI: Redefining Trust Challenges

AI for Good

Navigating Trust in the Shift from AIGC to Agentic AI

AIGC

Content Generation
Single-turn Interaction
Limited Execution
Model-centric

Core Trust Priorities

- Prompt Injection
- Hallucination & Misinformation
- Content Safety
- Copyright & Compliance
- Training Data Leakage

Capability Shift

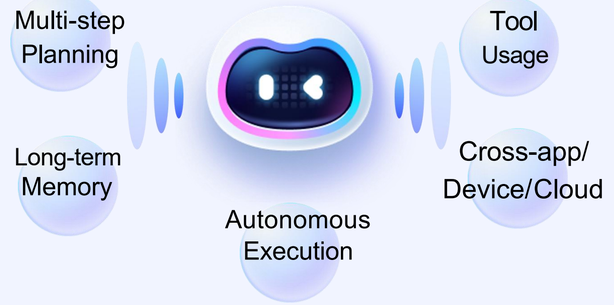
Generate	Perceive/Reason/Act
Reactive Response	Proactive Execution
Single Model	System Orchestration
Static Permissions	Dynamic Authorization

Risk evolves from model issues to operational and ecosystem risks.

What's New?

- From content risk to action risk
- From single prompts to multi-step workflows
- From static access to dynamic authorization
- From one-time input to continuous context

Agentic AI



Evolving Trust Risks

- Privilege Escalation/ Boundary Violations
- Continuous Data Collection
- Cross-App/Cross-Device Data Exposure
- Persistent Memory Leakage
- Tool/Plugin Supply Chain Risk
- Identity & Authentication Abuse
- Autonomous Action Errors
- Runtime Governance Gaps
- Multi-step Attack Chaining
- Audit & Accountability Challenges

vivo

Potential Trust Risks in On-Device Agentic AI during Runtime

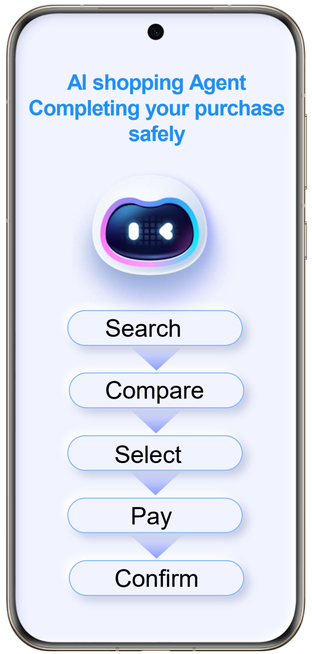
Investigating Runtime Trust Risks in On-Device Agentic AI: Exemplified by Mobile Commerce

Identity Security Risks

- Agent Identity Spoofing and Rogue Agent Infiltration
- Owner-Agent Binding Hijacking and Unauthorized Privilege Escalation
- Session Identity Persistence and Cross-Session Data Leakage

Privilege Escalation and Security Boundary Breaches

- Over-Privileged Agent Scopes and Permission Abuse
- Financial Limit Bypasses via Split Payments
- Unauthorized Tool Infiltration and Rogue API Exploits



Execution Environment Risks

- Malicious Sandbox Bypasses by Rogue Agents
- Unauthorized Access Across Payment and PII Boundaries
- Uncontrolled Network Connections and Data Leakage

Autonomy Risks and Behavioral Deviation

- Unauthorized Agent Spending and Payment Exploits
- Unauthorized High-Value Financial Transactions
- Loss of Agent Control and Kill-Switch Override Threats

Traceability Risks and Accountability Breakdowns

- Untraceable Agent Intent and Masked Actions
- Log Evasion and Forensic Accountability Breakdown
- Unstable Rollbacks and Undetected Behavioral Anomalies

vivo's Approach

On-Device Trust Control Plane for Agentic AI

AI for Good

On-Device Agent Runtime System Architecture

with Integrated Trust Control Plane and Agent Runtime Security Infrastructure(ARSI)

- Observe
- Decide
- Enforce
- Audit

User&Intent
User request,goals,preferences,context

Agent Orchestrator
Planning,reasoning,task decomposition,routing

Trust Control Plane
Agent Runtime Security Infrastructure(ARSI)

- Agent Identity
- AI Firewall
- Trusted Traceability
- Privacy&Data Minimization
- Model Security
- Runtime Safety

Model Runtime
On-device model,LLM inference, context window,caching

Tool&App Execution
Tool use,app interactions, API calls,external services

OS/ Hardware Foundation
Operating system,Secure Enclave/TEE,storage,hardware,network

Protected Outcomes

- Safe Execution
- Transparent Actions
- User Control
- Reduced Data Exposure

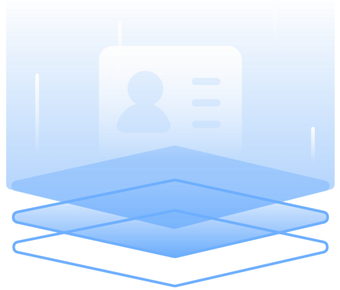
vivo

vivo's Practice for Agent Identity

Verifiable Identity for Accountable AI Actions

AI for Good

Agent Identity provides verifiable identity assurance for AI agents and links each authorized action to an auditable execution trace.

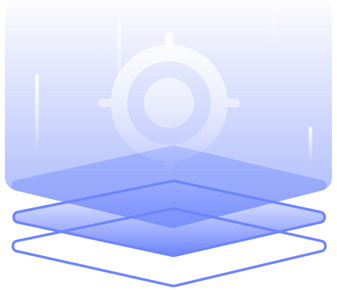


Agent Identity Anchor

Agent Digital Identity Certificate

Each agent has one trusted identity anchor

- Operation traceability and accountability
- Multi-agent identity verification
- Agent permission control

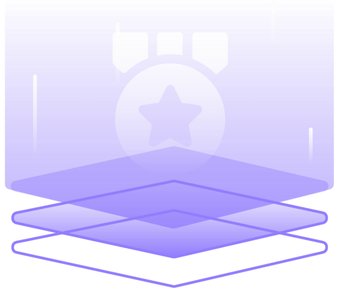


Agent Behavior Tracing

Provable Action Trajectory

A complete causal chain from planning to execution

- First-use authorization before execution
- Step-by-step action visibility during execution
- Manual confirmation for sensitive actions
- Agent action record after execution



Trust Assurance

AI agents act with verified identity, scoped authority, and traceable accountability

Moving users from reluctance to confidence

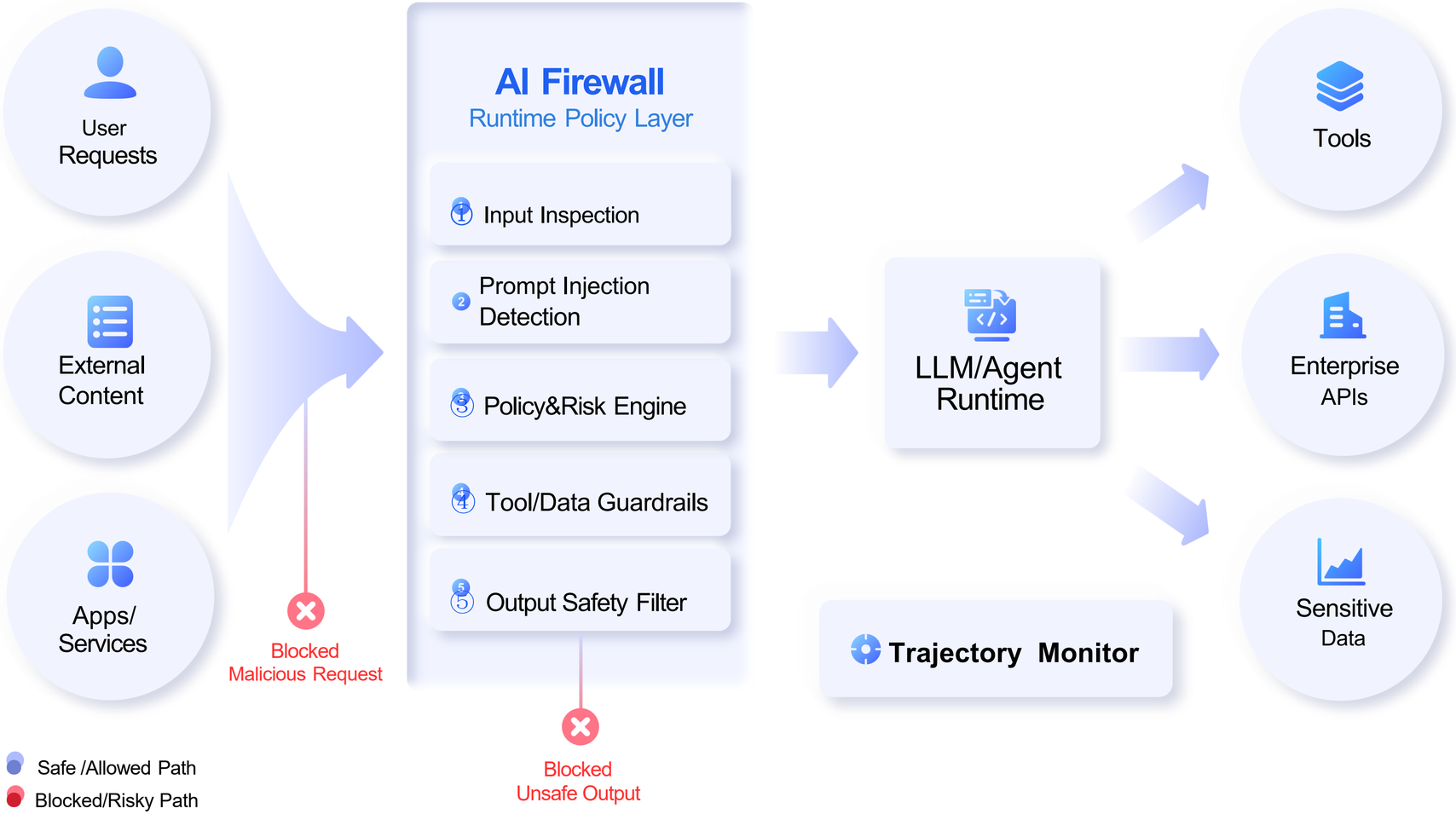
- AI actions are named and accountable
- Unauthorized AI actions are blocked immediately

vivo

vivo's Practice for AI Firewall

Guarding Agent Execution from Prompt Injection,Tool Misuse,and Data Leakage

AI for Good



Architecture Intent

Position — a runtime layer inserted between users,agents,tools,external content and backend services.

Core controls — prompt injection defense,misuse detection,data-leakage prevention and permission enforcement.

Decision logic — converts every request, content item, and tool call into an allow / constrain / block decision.

Security outcome — agent execution becomes traceable, policy-bound and auditable across the full runtime path.

From point defenses to a trust control plane for agentic AI execution.

vivo

vivo's Practice for Agent Behavior Tracing

Visible,Controllable,and Auditable Agent Execution

AI for Good

Agent Behavior Tracing Framework

A user-facing control layer that makes agent execution visible,controllable,and auditable.

Control objectives

- Visible
- Controllable
- Auditable

Lifecycle timeline

Before execution

Task summary
Permission scope
Allow/Cancel

During execution

Live operation steps
stop control
sensitive confirmation

After execution

Result summary
Action record
Audit trail

Deviation detection

Intent-alignment monitoring
deviation alerts
confirm / block

Agent Behavior Tracing

At a glance: What the agent did

Use Case: AI Dining Assistant

12:00:31	S	**Search Restaurants** Called map service to find nearby restaurants
12:00:33	N	**Plan Route &Navigate** Generated route and started navigation
13:12:22	M	**Create Message** Called messaging service to draft messag
13:12:25	√	**Send Message** Sent after user confirmation
13:12:26	C	**Task Complete** Logged result and completed the task

vivo

Standards Relevance for ITU

The ITU can define trust management standards for Agentic AI that fulfill business requirements, ensuring that accountability for every autonomous action is unambiguously attributed to the accountable entity.

Agent Identity

- Identity Management
- Authentication
- User-Agent Binding
- Multi-Agent Delegation

Agent Authorization

- Intent-Bound Authorization
- Temporal and Session-Scoped Constraints
- Least-Privilege Propagation
- Cryptographic Policy Attestation

Agent Behavior Tracing

- Intent-to-Action Mapping
- Prompt Injection Forensic Trail
- Cascading Action Lineage
- Parent-Child Agent Tracking
- Owner-Agent Cryptographic Co-Signing

SDOs Collaboration

- ETSI SAI
- OECD
- IEEE AIS Trust and Agency Committee
- ISO/IEC JTC1 / SC 27 / WG4 & WG5
- ISO/IEC JTC1 / SC 42 / WG3
- IETF IAB, IESG, SEC, ART

THANK YOU