

从感知到授权

From Perception to Permission

多模态、长程任务中的用户目标完整性

Preserving User Goal Integrity in Multimodal, Long-Horizon Tasks

环境内容：证据，也可能包含对抗性指令
Environmental content: task
evidence or adversarial instructions

环境内容不构成授权

Environmental Content Is Not Authorization

用户目标

USER GOAL

环境感知

ENVIRONMENTAL CONTENT

任务规划

PLANNING

工具调用

TOOL USE

外部行动

EXTERNAL ACTION

MiniMax实践：在动作执行前完成授权检查

MiniMax Practice: Check Authorization Before Execution

01

来源追踪

Provenance Tracking

区分用户指令、系统规则、环境内容与工具返回

Separate user, system, environmental, and tool inputs

02

规划与授权分离

Planning ≠ Authorization

模型提出动作，运行时检查目标、范围与外部影响

The model proposes; the runtime checks target, scope, and impact

03

工具隔离与最小权限

Tool Isolation & Least Privilege

将权限限定到工具、数据、对象和参数

Scope access to tools, data, targets, and parameters

04

长程任务中的目标校准

Goal Recalibration

计划、对象或外部影响变化时，重新确认授权

Re-authorize when the plan, target, or impact changes

AUTHORIZATION CHECK

执行前检查

依据动作对象、影响和可逆性分级

Assess target, impact, and reversibility

可自主执行 / PROCEED

读取 · 检索 · 整理 · 起草

Read · Search · Organize ·
Draft

需要确认 / CONFIRM

发送 · 付款 · 删除 · 改权限

Send · Pay · Delete · Change Access

环境内容可以支撑决策，但不能授予权限。

Environmental content can inform a decision; it cannot grant authorization.

安全评测既看任务结果，也看授权边界

Evaluate Task Outcomes and Authorization Boundaries

安全评测 / SAFETY EVALUATION

- 跨模态对抗性指令
Cross-Modal Adversarial Instructions
- 工具与参数越权
Tool and Parameter Overreach
- 长程执行中的目标漂移
Goal Drift in Long-Horizon Execution
- 高影响动作绕过确认
High-Impact Actions Without Confirmation

MINIMAX / ENGINEERING PRIORITIES

下一步：将授权检查做进系统

- 01 多模态来源追踪
Multimodal Provenance
- 02 权限感知运行时
Permission-Aware Runtime
- 03 参数级最小权限
Parameter-Level Least Privilege
- 04 长程安全评测
Long-Horizon Safety Evaluation

任务完成

+

目标与授权完整性

=

可信执行

TASK SUCCESS + GOAL & AUTHORIZATION INTEGRITY = TRUSTWORTHY EXECUTION