ICS 35.240
CCS L70

团　　体　　标　　准

T/ISC XXX—XXXX

端侧多模型资源隔离与调度安全要求

Security Requirements for Resource Isolation and Scheduling of Multiple Models on Terminals

在提交反馈意见时，请将您知道的相关专利与支持性文件一并附上。

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中 国 互 联 网 协 会　　发布

目　　次

前言 II
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 缩略语 1
5 概述 2
6 资源隔离安全要求 2
6.1 计算资源隔离 2
6.2 存储资源隔离 2
6.3 数据资源隔离 2
6.4 权限隔离 3
6.5 密钥与凭证隔离 3
7 模型调度安全要求 3
7.1 调度策略安全 3
7.2 资源分配安全 3
7.3 多模型并发安全 3
7.4 调度审计与可追溯 4
参考文献 5

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国互联网协会提出并归口。

本文件起草单位：××××、××××。

本文件主要起草人：×××、×××。

端侧多模型资源隔离与调度安全要求

1 范围

本文件规定了智能终端设备（包括但不限于智能手机、车载终端、物联网终端等）上同时部署运行多个人工智能模型在资源隔离与调度方面的安全技术要求。

本文件适用于智能终端设备上多模型的部署、运行与调度场景，可为端侧模型服务提供者的设计、开发、部署和运维活动提供参考，也适用于第三方评估机构等组织进行安全评估。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069—2022 信息安全技术 术语

GB/T 39786—2021 信息安全技术 信息系统密码应用基本要求

GB/T 41388—2022 信息安全技术 可信执行环境 基本安全规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

多模型 multiple models

在同一端侧终端上同时部署运行的两个及以上人工智能模型。

3.2

资源 resource

执行所要求的操作而必需的端侧系统的任何组成部分。端侧多模型资源包括系统和计算资源、数据资源、工具资源。

[来源：GB/T 25069—2022，3.807，有修改]

3.3

可信执行环境 trusted execution environment

基于硬件级隔离及安全启动机制，为确保安全敏感应用相关数据和代码的机密性、完整性、真实性和不可否认性目标构建的一种软件运行环境。

[来源：GB/T 41388—2022，3.3]

3.4

调度器 scheduler

负责模型任务排队、资源分配、并发控制、状态监控和异常处理的系统组件。

3.5

安全沙箱 security sandbox

为模型运行提供的隔离环境，限制模型对系统资源的访问范围和能力。

4 缩略语

下列缩略语适用于本文件。

API：应用程序编程接口（Application Programming Interface）

CPU：中央处理器（Central Processing Unit）

GPU：图形处理器（Graphics Processing Unit）
NPU：神经处理器（Neural Processing Unit）
TEE：可信执行环境（Trusted Execution Environment）

5 概述

端侧多模型系统在智能终端上同时部署运行多个人工智能模型，面临资源竞争、数据泄露、权限滥用等安全风险。为确保多模型环境的安全可靠，本文件从资源隔离和调度安全两个维度提出技术要求。其中，资源隔离涵盖计算资源、存储资源、数据资源、权限及密钥凭证等方面，确保各模型运行时资源相互隔离、互不干扰；模型调度涵盖调度策略安全、资源分配安全、多模型并发安全及调度审计与可追溯等方面，保障调度过程的安全可控与可审计。

6 资源隔离安全要求

6.1 计算资源隔离

计算资源隔离要求包括：

a) 不同模型应在相互隔离的执行上下文中运行。执行上下文可通过物理分区、虚拟化、容器、进程隔离或经验证的硬件资源分区实现。不得共享可变运行时状态、临时数据和未授权缓存；
b) 关键模型或处理敏感数据的模型应运行在可信执行环境中；
c) 每个模型应运行在独立的安全沙箱中，沙箱应具备进程隔离、内存隔离和系统调用过滤能力；
d) 不同模型的推理任务不应共享同一运行时上下文；
e) 应支持模型的动态加载与卸载，模型卸载时可彻底清理其占用的所有计算资源。

6.2 存储资源隔离

存储资源隔离要求包括：

a) 各模型的参数文件、配置文件等应分别存储在不同目录，并设置独立的访问控制权限；
b) 各模型的运行时缓存数据应独立存储；
c) 模型权重、激活值的内存地址应空间隔离；
d) 调度器应具备检测模型异常申请内存的能力；
e) 分页空间应进行加密，防止模型运行时数据被写入非加密交换空间；
f) 模型文件、索引、缓存、临时文件、日志、向量库应分域存放，资源回收后应立即清理残留；
g) 模型文件在端侧存储时宜进行加密，密钥与模型文件分别管理，模型文件的加密存储与密钥应用的密码技术要求按照 GB/T 39786—2021 执行；
h) 对于加密存储的模型，每个模型应使用独立密钥；
i) 对于加密存储的模型，密钥应存储在硬件安全模块、TEE 或系统安全密钥库中；
j) 密钥不得以明文写入普通文件、日志或命令行参数；
k) 对于加密存储的模型，应支持密钥轮换、吊销和模型版本绑定；
l) 模型文件应支持完整性校验，防止替换和回滚；
m) 模型权重、激活值、中间结果在释放后应进行清零；
n) 异常退出、进程崩溃和设备重启应具备残留数据清理机制。

6.3 数据资源隔离

数据资源隔离要求包括：

a) 不同模型的用户输入数据应在独立的缓冲区处理；
b) 各模型的推理结果应严格隔离，禁止未授权访问；
c) 模型推理过程中产生的中间计算结果应及时清理，不应被其他模型或进程获取；
d) 输入数据、输出数据、中间结果、上下文、共享记忆、工具返回内容应分级隔离，跨模型共享前应脱敏和最小化；

e） 应对不同模型处理的数据进行分类分级，高敏感数据应在隔离的数据分区中处理，禁止不同安全等级的数据混合存储和处理；
f） 应对不同模型处理的数据进行分类分级，高敏感数据应在隔离的数据分区中处理，禁止不同安全等级的数据混合存储和处理；
g） 模型推理的输入和输出数据应进行来源标识和用途标记，确保数据流向可追溯。

6.4 权限隔离

权限隔离要求包括：
a） 各模型应分配独立的最小权限，不应继承其他模型的权限，模型、调度器、运维代理、插件、工具链、系统服务分别授权；
b） 模型对端侧系统资源（如摄像头、麦克风、存储、网络等）的访问应经过鉴权；
c） 若需支持模型间协同，应建立受控的通信通道，并对通信内容进行审计；
d） 应建立模型权限清单，记录各模型的权限范围、接口访问权限和数据访问权限，并定期审查；
e） 模型完成推理任务后，调度器应立即回收其运行时权限；检测到权限越权行为时，应终止模型运行并告警。

6.5 密钥与凭证隔离

a） 各模型调用云端 API 或访问本地敏感资源时使用的密钥、令牌和证书应独立管理，禁止不同模型共享凭证；
b） 密钥应存储于安全可信区域，禁止明文存储或硬编码在模型代码中；
c） 应建立密钥轮换和吊销机制，当模型被卸载、安全评估不通过或检测到凭证泄露时，应及时轮换和吊销相关凭证；
d） 模型间通信不应直接复用全局共享密钥，应通过密钥派生机制为每对模型生成独立的通信密钥。

7 模型调度安全要求

7.1 调度策略安全

调度策略安全要求包括：
a） 调度策略应支持实时任务保障，对时延敏感型模型（如实时交互模型）应提供实时调度通道或优先级保障；
b） 调度策略应具备防止调度拒绝服务攻击的能力，限制单一模型的调度请求频率；
c） 调度决策过程应可审计，记录调度依据和决策结果。

7.2 资源分配安全

资源分配安全要求包括：
a） 应支持对模型使用的 CPU、GPU、NPU、内存、存储、网络带宽等资源分别设定配额；
b） 在分配资源前应校验目标模型的安全等级与目标执行环境的安全能力是否匹配；
c） 模型任务完成后或异常终止时，应及时回收分配的资源，并清理残留数据；
d） 应建立资源突发控制机制，限制模型在峰值负载下的资源占用上限和持续时间，防止资源耗尽攻击。

7.3 多模型并发安全

多模型并发安全要求包括：
a） 多模型同时运行时，对共享资源（如存储、网络等）的并发访问应有同步机制，防止竞态条件引发的安全问题；
b） 应具备死锁检测与恢复能力；
c） 应支持正常、限流、降级、隔离、熔断和恢复等状态。熔断触发条件、持续时间、恢复条件和人工解除机制应在安全策略中明确；

d） 应对各模型的资源使用情况进行实时监控，对异常的资源使用进行告警。当某一模型连续出现推理超时、输出异常时，调度器应具备自动降级或熔断能力；
e） 应设定最大并发模型数量上限，超过上限时新模型任务应排队等待；
f） 普通优先级模型应具备保底资源或保底时间片，在系统资源达到设定高负载阈值时，普通优先级模型仍应获得不低于其配置保底值的计算时间片或资源配额；
g） 应具备优先级老化机制；
h） 应具备模型资源超限时的排队、拒绝、降级和回收策略。

7.4 调度审计与可追溯

调度审计与可追溯要求包括：

a） 应完整记录所有模型调度决策、资源分配和状态变更日志；
b） 应支持跨层安全事件联动与追溯，实现调度层、运行层、存储层安全事件的关联分析，当任一层发生异常时能够触发关联告警；
c） 调度日志应加密存储，阻止篡改和未授权访问；
d） 调度记录应支持事后审计；
e） 调度日志保留期限不应少于 6 个月；涉及安全事件的调度日志保留期限不应少于 3 年；
f） 调度日志应采用标准化格式，并支持防篡改校验。

参　　考　　文　　献

[1] GB/T　22239—2019　信息安全技术　网络安全等级保护基本要求
[2] GB/T　45654—2025　网络安全技术　生成式人工智能服务安全基本要求
[3] GB/T　45288.1—2025　人工智能　大模型　第1部分：通用要求