ICS 33.050
CCS M 30

团　体　标　准

T/TAF 310—2025

移动智能终端端侧大模型安全实施指南

Security implementation guidelines for large models deployed on the mobile terminal

2025-08-11 发布　　2025-08-11 实施

电信终端产业协会　发布

版权声明

目　　次

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。

本文件由电信终端产业协会（TAF）提出并归口。

本文件起草单位：维沃移动通信有限公司、中国信息通信研究院、小米通讯技术有限公司、OPPO广东移动通信有限公司、荣耀终端股份有限公司、蚂蚁科技集团股份有限公司、联想（北京）有限公司、北京快手科技有限公司、高通无线通信技术（中国）有限公司、安谋科技（中国）有限公司、北京微梦创科网络技术有限公司、上海得物信息集团有限公司、北京零一万物科技有限公司。

本文件主要起草人：赵盈洁、王艳红、武林娜、王淞鹤、李方圆、周楚杰、苏涛、吴越、李根、李辰淑、赵晓娜、林冠辰、李欣、高博雅、谷晨、王江胜、王骏超、任资政、康宇、杨欢、王海棠。

移动智能终端端侧大模型安全实施指南

1 范围

本文件规定了提出了移动智能终端端侧大模型的框架、风险、安全相关的实施指南。

本文件适用于移动智能终端端侧大模型服务提供者的设计、生产活动，也适用于主管部门、第三方评估机构等组织对移动智能终端端侧大模型的安全进行评估。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

移动智能终端 smart mobile terminal

能够接入移动通信网，具有能够提供应用软件开发接口的操作系统，具有安装、加载和运行应用软件能力的终端。

[来源：YD/T 2407—2021,3.1.1]

3.2

大模型 large model

使用大规模数据集通过无监督学习并进行有监督微调得到的，具备大参数量的可以执行广泛任务的深度学习模型。

3.3

移动智能终端端侧大模型 on-device large model for smart mobile terminal

能够在智能手机、平板电脑等移动智能终端 OS 上本地进行部署、运算和推理的一种大模型，基础模式为离线模式，在该模式下可离线进行推理，扩展模式为在线模式，利用云侧协助完成更复杂的推理任务，本文件简称端侧大模型。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

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

APP：移动应用程序（Application）

CPU：中央处理器（Central Processing Unit）

CVE：通用漏洞披露（Common Vulnerabilities & Exposures）

DoS：拒绝服务（Denial of Service）

GPU：图形处理器（Graphics Processing Unit）

ID：身份标识号（Identity Document）

OS：操作系统（Operating System）

SSRF：服务端请求伪造（Server-Side Request Forgery）
URL：统一资源定位系统（Uniform Resource Locator）
XSS：跨站脚本攻击（Cross Site Scripting）

5 端侧大模型概述

5.1 端侧 AI 架构

端侧大模型的部署存在两种情况，一是嵌入终端OS中，二是与终端OS有一定耦合或集成，端侧大模型会与终端上搭载的APP进行相应交互。同时可能存在端侧大模型接入云侧，利用云侧能力协助完成更复杂的任务的情况。

在上述场景中，需要着重考虑模型安全、业务安全以及相应安全保障的内容，整体端侧AI架构见图1。

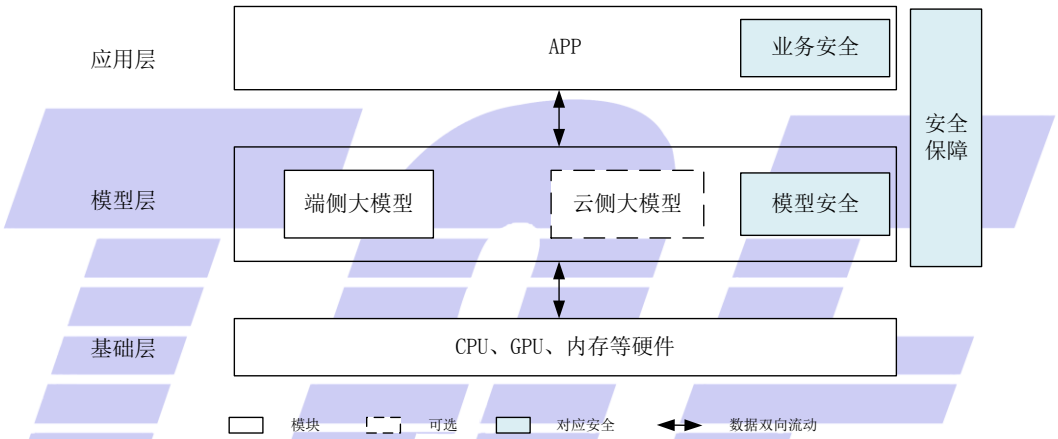

图1 端侧AI架构

5.2 端侧大模型模式分类

根据是否支持云侧接入分为离线模式与在线模式两种模式。

a） 离线模式：
无需联网，仅在本地运行和处理数据，依靠本地算力提供服务。

b） 在线模式：
端侧接入云侧，利用云侧能力协助，可支持更复杂的服务场景，完成更为复杂的任务。

5.3 应用场景

端侧大模型的离线模式可具备文本总结、图像修改、语音识别等能力，在文本创作、图像创作等场景下提供服务；在线模式可具备语言理解能力，在文本创作、图像创作、角色扮演、知识问答、自然对话、逻辑推理、任务编排等场景下提供服务。

5.4 安全威胁

端侧大模型面临的安全威胁见附录A，主要分为两大类，分别是端侧大模型整体面临的安全威胁和端侧大模型中算法面临的安全威胁，具体描述如下：

——在端侧大模型整体层面，安全威胁主要包括模型窃取、模型泄露、模型篡改、数据泄漏等；

——在端侧大模型算法层面，安全威胁主要包括数据投毒攻击、模型后门攻击、成员推理攻击等。

6 端侧大模型安全实施框架

端侧大模型的生命周期可分解为训练、部署、推理结果三个阶段，其中对应的安全主要涉及数据安全、模型安全、业务安全以及贯彻其中的安全保障，整体架构如图2。

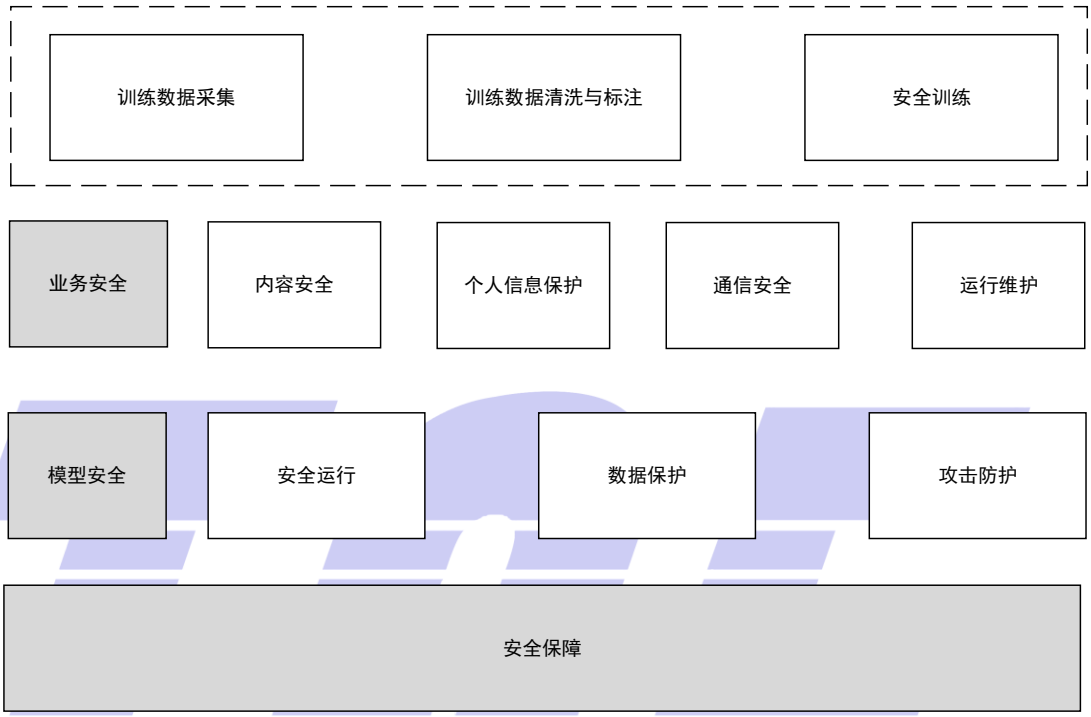

图2 端侧大模型安全架构

注：因本文件中的端侧大模型实施不涉及数据训练的过程，相应数据安全可参考其他标准。

7 模型安全

7.1 安全运行

服务提供者采取的安全措施可包括以下内容：

a) 端侧大模型启动、运行时，终端采用代码签名和完整性校验技术校验模型机密性、完整性、可用性，以及对启动、运行环境进行安全检测，例如终端是否被 root、模型进程是否被进行注入攻击，确保模型在运行时未被篡改；
b) 通过安全通道对端侧大模型进行传输，保证传输的机密性和完整性。
c) 云侧在协助端侧大模型进行推理分析时，仅将推理结果返回给终端，并采用安全通道进行数据传输，保证传输机密性和完整性；
d) 终端 OS 和端侧大模型及时进行安全更新，修复已知漏洞。

7.2 数据保护

服务提供者采取的安全措施可包括以下内容：

a) 对模型数据和参数的访问实施严格的访问控制，只有经过身份验证和授权后主体才可以访问；

b） 端侧模型和权重文件加密存储，并存储于私有目录下，若终端不需要更新模型，则模型和权重文件设为只读；
c） 建立过期模型销毁机制，防止攻击者获取过期模型。

7.3 攻击防护

服务提供者采取的安全措施可包括以下内容：
a） 对端侧大模型关键结构和参数进行混淆和隐藏；
b） 对模型运行文件采用加壳、反调试等加固技术，防止模型被恶意破解或反编译；
c） 对模型返回用户的结果形式加以约束，不打印业务有关信息（例如置信度数值等），仅将用户需求的图像、文本、音频等 AI 推理结果返回，防止信息过度反馈导致逆向攻击；
d） 系统不向用户提供内存 DUMP 工具，以防止用户窃取模型明文；端侧大模型进行升级时，终端固件具有验证升级包的真实性和完整性的能力，当验证失败时能终止升级流程并进入到安全状态。

8 业务安全

8.1 内容安全

服务提供者采取的安全措施可包括以下内容：
a） 建立端侧内容审核能力，对用户的输入内容和生成内容进行审核：
  1） 能够发现内容中包含的违反社会主义核心价值观内容、歧视性内容、侵犯他人合法权益内容等；
  2） 能够对其他关键内容进行识别，如人物、场景、物体等；
  3） 能够与云侧协同进行审核。
b） 建立本地敏感关键词库，支持 and、or 逻辑组合式关键词，变形关键词等，且本地词库定期更新；
c） 建立对模型输入输出的不良内容的处置机制，如对模型输出进行阻断、向用户发送警告或禁止用户访问等；
d） 对端侧审核模型参考第 7 章进行安全保护；
e） 文字、图片、音频、视频等生成内容标识方面，满足国家相关规定以及标准文件要求。

8.2 个人信息保护

服务提供者采取的安全措施可包括以下内容：
a） 采集用户个人信息之前，明确告知用户个人信息采集的目的、方式、用途和范围，并取得用户的同意；
b） 按照最小必要原则收集用户个人信息，不收集与使用目的无关的信息；
c） 在处理个人信息时，优先进行端侧处理；
d） 仅端侧进行推理分析时，将用户个人信息进行加密存储或将其存储在终端安全存储区域，且使用过程中数据不出终端；
e） 采用端云结合方式进行推理分析时，除业务所必须外，端侧对用户个人信息进行去标识化或匿名化处理后再进行云侧分析；
f） 采用端云结合方式进行推理分析时，端侧仅上传最小必要的数据至云侧；
g） 端侧大模型与应用或服务进行数据交互过程中，必要时采用去标识化等技术对个人信息进行脱敏处理。

8.3 通信安全

服务提供者采取的安全措施可包括以下内容：

a） 云侧与端侧大模型通信过程中，对用户输入的URL进行校验，以防攻击者通过SSRF攻击访问到服务提供者的内部网络资源；

b） 在通信过程中，采取安全的通信认证方式或通信协议版本，在模型下载过程中，通过云侧和终端侧之间的安全通道将端侧大模型下发至终端设备；

8.4 运行维护

服务提供者采取的安全措施可包括以下内容：

a） 记录用户使用端侧大模型相关信息，例如用户使用记录等，并加密存储在端侧私有目录内；

b） 建立端侧安全监测系统，实时监控并及时处置端侧大模型产生的异常结果、中断服务等不安全问题；

c） 建立有效的用户投诉反馈渠道，收到用户反馈后，及时评估和处置。

9 安全保障

服务提供者采取的安全措施可包括以下内容：

a） 建设大模型安全合规制度，可包括数据采集规范、用户管理规范、内容安全规范、模型安全规范等，明确端侧大模型安全保护要求；

b） 定期对大模型开发、数据标注、内容审核等相关人员进行安全培训和考核；

c） 定期对大模型算法、效果及相关安全措施进行评估和验证，包括但不限于对大模型平台安全审计，对第三方模块安全检测，采用工具扫描、模糊测试等方式验证模型安全性；

d） 建立安全应急响应制度，明确安全事件分级、事件处置方式、应急演练等内容，并定期进行演练，记录演练效果。

附 录 A
（资料性）
端侧大模型安全威胁分析

端侧大模型安全威胁分析见表A.1。

表A.1 端侧大模型安全威胁分析

安全威胁种类	具体安全威胁描述
仿冒	攻击者仿冒端侧大模型包名拉起应用
	攻击者仿冒应用身份请求大模型接口、技能
	攻击者伪造账号登录
	攻击者通过XSS，仿冒用户身份进行操作，如窃取cookie，进行会话劫持
篡改	攻击者通过XSS，注入恶意链接，导致用户提交的数据被篡改
	攻击者通过话术攻击，篡改预置话术，导致大模型“越狱”，输出违规内容
	恶意用户修改或清除端侧数据或记录
抵赖	攻击者仿冒用户身份登录，构成行为抵赖
	攻击者通过构造伪造请求使用户在不知情的情况下执行操作，事后用户会因为无法证明自己未进行该操作产生抵赖问题
信息泄露	攻击者通过获取用户的浏览器缓存、本地存储或其他敏感信息实现信息泄露
	调用三方大模型API接口服务时，三方供应商获取到用户会话信息
	攻击者通过海量请求，遍历出大模型预设答复策略
	攻击者通过获取审核模型推理置信度，窃取审核模型
	攻击者针对端侧的离线模型、敏感词库等文件进行攻击并窃取
	攻击者下载未加密的大模型、审核模型文件，剽窃知识产权
	攻击者通过接口滥用的漏洞，盗刷模型能力
拒绝服务	攻击者通过利用无校验测试接口发起海量请求
	攻击者利用跨站脚本来大量触发某种消耗资源的操作，造成服务器性能下降甚至DoS
权限提升	攻击者通过修改客户端版本进行无校验旧版本，越权使用大模型功能
	攻击者通过XSS漏洞获取更高的页面权限，例如执行JavaScript代码读取或修改高权限区域的内容
	攻击者通过精心构造数据，触发代码依赖组件中的CVE问题导致权限提升
	攻击者通过端侧修改系统时间、清除缓存、改ID等方式绕过端侧违规处罚策略
	攻击者通过 SSRF 攻击访问到服务提供者的内部网络资源

参 考 文 献

[1] YD/T 2407—2021 移动智能终端安全能力技术要求

电信终端产业协会团体标准

移动智能终端端侧大模型安全实施指南

T/TAF 310—2025

*

电信终端产业协会发布
地址：北京市西城区新街口外大街 28 号
电话：010-82052809
电子版发行网址：www.taf.org.cn