ICS 33.050
CCS M 30

团　体　标　准

T/TAF 150—2023

移动互联网应用人工智能模型安全指南

Security guidelines for artificial intelligence model used with application software

2023-02-08 发布　　2023-02-08 实施

电信终端产业协会　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 2
5 AI 模型威胁分析 2
6 AI 模型安全指南 2
6.1 概述 2
6.2 AI 模型训练算法安全 2
6.3 AI 模型下发和部署 3
6.4 AI 模型数据采集 3
6.5 AI 模型推理和计算 4
6.6 AI 模型结果数据传输 4
6.7 AI 模型更新和退役 5
6.8 AI 模型运行环境和应用安全 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。

本文件由电信终端产业协会提出并归口。

本文件起草单位：蚂蚁科技集团股份有限公司、荣耀终端有限公司、中国信息通信研究院、郑州信大捷安信息技术股份有限公司。

本文件主要起草人：辛知、林冠辰、张璇、彭晋、顾婉玉、吴超、傅欣艺、孟昌华、王维强、罗广文、韩业飞、李志超、张朋、赵晓娜、傅山、王嘉义、魏凡星、刘为华。

移动互联网应用人工智能模型安全指南

1 范围

本文件提出移动应用的人工智能模型的威胁分析、模型安全指南等。

本文件适用于移动应用的人工智能模型研发、设计、测评、应用等活动。

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 41388-2022 信息安全技术 可信执行环境 基本安全规范

GB/T 41867-2022 信息技术 人工智能 术语

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能 artificial intelligence

针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统的相关机制和应用的研究和开发。

[来源：GB/T 41867-2022,3.1.2&3.1.8，有改写]

3.2

可信执行环境 trusted execution environment

移动智能终端上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。

注：硬件级隔离是指基于硬件安全扩展机制，通过对计算资源的固定划分或动态共享，保证隔离资源不被富执行环访问的一种安全机制。

[来源：GB/T 41388-2022，3.3，有改写]

3.3

富执行环境 rich execution environment

移动智能终端上为应用程序提供基础功能和计算资源的一种软件软件环境。

注：富执行环境是相对可信执行环境独立存在的运行环境。

[来源：GB/T 41388-2022，3.4，有改写]

3.4

可信应用 trusted application

运行在可信执行环境下的应用。

4 缩略语

下列缩略语适用于本文件。

AI ：人工智能（Artificial Intelligence）
FGSM：快速梯度符号方法（Fast Gradient Sign Method）
OTA：空中下载技术（Over-the-Air Technology）
PGD：投影梯度下降（Projected Gradient Descent）
REE：富执行环境（rich execution environment）
TA ：可信应用（Trusted Application）
TEE：可信执行环境 trusted execution environment
ZOO：零阶优化（Zeroth Order Optimization）

5 AI 模型威胁分析

移动互联网应用AI模型部署在移动智能终端上，其安全威胁可以分为系统层面的安全威胁和模型算法层面的安全威胁，具体描述如下：

a) 系统层面的威胁主要是指 AI 模型作为智能终端上的一种数据资产面临的安全威胁，这种威胁通常通过攻击智能终端软硬件系统和应用造成的模型数据安全风险，包括模型窃取、模型泄露、模型漏洞攻击、模型后门攻击等等；
b) 模型算法层面的威胁主要是指通过攻击 AI 模型算法，使模型输出和实际或者预期不一致的安全风险，包括数据噪声、数据投毒攻击、后门攻击、预处理攻击以及成员推理攻击、模型提取攻击、对抗攻击等。

从AI模型生命周期角度看，AI模型全生命周期包括模型数据采集、模型训练、模型下发、模型部署、模型推理和计算、模型退役等阶段。模型训练阶段主要面临的安全威胁为模型算法层面的威胁，模型下发、部署阶段主要面临系统层面的安全威胁，模型推理和计算阶段面临的安全威胁包括系统和算法层面的威胁，模型退役阶段面临的威胁主要为系统层面的威胁。

6 AI 模型安全指南

6.1 概述

AI模型整体技术架构如图1所示。移动应用AI模型安全包括模型训练、模型下发和部署、数据采集、模型推理和计算、模型结果数据传输、模型更新和退役、模型系统和应用等过程或阶段的安全。

6.2 AI 模型训练算法安全

6.2.1 AI 模型算法处理的样本类型

AI模型算法处理的样本类型可能包括图像、音频、文本、序列和表格等。

6.2.2 AI 模型算法安全能力

AI模型算法安全能力能够防御训练阶段的数据投毒攻击、后门攻击等攻击手段，以及模型推理阶段的数据噪声、预处理攻击、成员推理攻击、模型提取攻击、对抗攻击、模型逆向攻击等攻击手段。按照攻击防御的难易程度，从容易到难可分为防御黑盒攻击和防御白盒攻击，对应能力可防御的攻击方式描述如下：

a) 防御黑盒攻击，即限定查询次数和扰动大小和时间，在仅能获取模型决策或打分的条件下进行攻击，例如：ZOO 等；
b) 防御白盒攻击，即在允许获取模型的全部信息的条件下进行攻击，例如：FGSM，PGD 等。

6.3 AI 模型下发和部署

AI模型下发和部署流程如下：

a) 基于系统提供方提供的 TEE 等安全机制开发可运行 AI 计算引擎的可信应用（服务）；
b) 利用系统提供方提供可信应用下载平台在终端设备上进行 AI 计算可信应用（服务）的下载、安装和运行；
c) AI 算法提供方训练生成 AI 模型，并使用系统提供方提供的可信环境相关证书进行签名，通过安全通道将 AI 模型加密下发至终端设备；
d) AI 模型下发时应通过服务器和端侧之间的安全通道下发，保障 AI 模型下发过程的机密性和完整性，防止模型被窃取和篡改；
e) 采用数字证书等安全手段来进行 AI 模型的合法性校验，用于 TEE 验证验证模型数据的完整性，证书私钥需要使用安全存储，不能暴露于非安全域；
f) AI 模型通过安全存储保存。

6.4 AI 模型数据采集

6.4.1 AI 模型数据采集流程

AI模型数据采集和存储流程安全指南如下：

a) REE 应用获取用户授权后发起数据采集请求；
b) 如果支持安全传感器，REE 通过 TEE 中的传感器驱动采集数据，TEE 采集到数据通过加密加签保证数据的机密性和完整性，传输给 REE 应用；

示例：安全传感器如基于 TEE 或 SE 安全机制进行数据采集的安全摄像头等。

c) 如果需要在终端存储采集到的数据，则对应用隔离，加密存储，用于加密的密钥保存于 TEE 或 SE 安全存储环境内。

6.4.2 采集过程个人信息主体权益保护

App 的AI模型数据采集遵循用户授权后方可采集的原则，不从未知来源收集信息，不同应用采集的数据应该相互隔离，采集数据应保证其完整性。AI模型数据存储，如涉及到个人信息，存储前应获得个人的授权同意，并且加密存储。

6.4.3 采集数据的完整性

为了保证推理结果的正确性，需要对计算输入数据进行完整性校验，保证数据的源头可信，如可以使用安全传感器外设采集链路。

6.4.4 采集数据去标识化或匿名化

采集的数据如需上传云端服务器，需要在终端进行去标识化或匿名化处理，删除数据和个人信息主体身份的关联。

6.5 AI 模型推理和计算

6.5.1 AI 模型推理和计算流程

AI模型推理和计算流程如下：

a) 采集计算所需数据，需要保证数据完整性，有条件可采用安全传感器外设采集链路；
b) 解密加载 AI 模型到安全内存；
c) 通过 AI 计算可信应用（服务）对采集数据进行推理计算；
d) 推理结果输出需要保证机密性和完整性，可以在终端直接进行分析，也可进行后续上云流程。

6.5.2 AI 模型计算与分析

AI模型计算包括脱敏、密码学计算和AI计算等。AI模型计算属于终端上较为复杂的数据处理分析方法，涉及AI模型和运算引擎的部署，以及运算过程和结果的可信，需要满足以下要求：

a) AI 推理需要在 TEE 下运行，保证推理过程的机密性，防止模型被窃取，以及推理过程被恶意注入，推理结果被篡改；
b) AI 算法提供方需要提供模型生命周期管理机制，包括终端所运行模型的下发、更新和注销；
c) 终端数据采集过程需要保证数据完整性，避免 AI 计算数据源被篡改；
d) AI 计算作为可信应用或可信服务，系统提供方需要提供可信应用和可信服务的生命周期管理机制，包括可信应用和可信服务的下载、安装、更新和注销；
e) AI 数据计算也需在安全环境完成，为了防止信息过度反馈，导致逆向攻击，在保证最小够用的情况下，应减少计算结果反馈到非安全域。

6.5.3 AI 模型推理过程的机密性与完整性

为了保护AI模型数据的机密性，模型从AI算法服务平台下发到终端，需要使用对称密钥进行加密传输，该对称加密密钥使用终端TEE或SE生成的公私钥进行加密交换，每次下发操作前需要更新对称密钥。该对称密钥需要使用安全存储，不能暴露于非安全域。

6.5.4 AI 模型推理结果的完整性

AI模型结果数据传输安全要求如下：

a) AI 推理结果需要保证其完整性，如对推理结果进行加密传输，加密密钥由 TEE 或 SE 生成，需要安全存储，防止推理结果被篡改；
b) AI 推理结果需要遵循最小输入原则，避免通过输出结果逆向模型。

6.6 AI 模型结果数据传输

AI模型结果数据传输安全要求如下：

a) 终端与云端服务器通过安全通道进行密钥协商与交换；
b) 终端应用如果需要直接上传采集数据，则需要对数据进行脱敏、密码算法等去标识化或匿名化处理；如果仅需上传数据分析的中间结果，则仅需要加密加签保证数据机密性和完整性，上传至云端服务器；

c） 通信安全包含两方面：REE 和 TEE 的安全通信和端-云的安全通信。为了保证数据传输的机密性和完整性，需要在安全环境中进行密钥预置；加密与签名密钥需要通过安全通道进行协商与交换；密钥需要存储于安全环境，应用之间密钥相互隔离。

6.7 AI 模型更新和退役

AI模型退役安全指南如下：

a） 模型更新时，新模型下发和部署应符合 6.4 的要求；

b） 模型退役且不再使用相关模型时，涉及个人信息数据应及时删除或者匿名化处理。

6.8 AI 模型运行环境和应用安全

为保障AI模型的运行环境和模型的应用安全，系统和应用程序层面需符合如下要求：

a） 安全启动：终端固件内存在唯一的可信根，具备验证 REE、TEE 二进制代码真实性、完整性的能力，REE 具备验证应用程序二进制代码真实性、完整性的能力，TEE 具备验证可信应用程序二进制代码真实性、完整性的能力，当验证失败时能终止启动流程并进入到一个安全状态；

b） 安全隔离：可信应用之间、可信应用与 TEE 之间使用的密钥相互隔离、内存隔离、存储数据隔离；

c） 可信应用安装和数据验证：REE 验证应用程序安装包中的签名信息，若没有签名信息或签名信息由未授权开发者制作，则终止应用安装流程并进入到一个安全状态；TEE 验证可信应用安装包中的签名信息，若没有签名信息或签名信息由未授权开发者制作，则终止可信应用安装流程并进入到一个安全状态；AI 可信应用能够在 AI 模型数据预置或 OTA 安装过程中验证 AI 模型数据的真实性、完整性、机密性；

d） 系统升级：终端固件具有验证 REE 的 OTA 升级包、TEE 的 OTA 升级包真实性、完整性的能力，当验证失败时能终止 OTA 升级流程并进入到一个安全状态；

e） 调试接口：REE、TEE 的调试功能需被禁用，或经过适当的授权验证后方能开启调试接口；

f） 通信安全：不使用不安全的通信认证方式或通信协议版本；

g） 应用加固：使用例如代码加壳、代码混淆、检测调试器等手段对应用进行安全保护；

h） 安全审计：TEE 具备安全审计功能，能够为可信系统、可信应用相关可审计事件生成审计记录。

电信终端产业协会团体标准

移动互联网应用人工智能模型安全指南

T/TAF 150—2023

*

电信终端产业协会印发
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