ICS 35.xxx
CCS Lxx

团　　体　　标　　准

T/ISC XXX—XXXX

端侧大模型更新与运维技术规范

Technical Specification for Update and Operation Maintenance of On-Device Large Models

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

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。

本文件是T/CIITA x《……》第i部分。T/CIITA x已经发布了以下部分：

——第1部分：……；

——……；

——第n部分：……。

本文件替代T/CIITA x.i《……》，与T/CIITA x.i相比，除结构调整和编辑性改动外，主要技术变化如下：

a)

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：

本文件及其所代替文件的历次版本发布情况为：

——

引　言

随着人工智能技术的快速发展，大模型向端侧下沉成为产业发展的重要趋势，端侧大模型在智能移动终端、智能穿戴、智能车载等各类终端中得到广泛应用，为终端赋予了更强的自主推理、智能交互能力。

但当前端侧大模型的更新与运维环节缺乏统一的行业技术规范，不同企业的模型更新方式各异、版本管理体系不统一、生命周期管控流程缺失，且性能监控指标与故障处理机制缺乏统一标准，导致端侧大模型运行稳定性不足、资源利用效率参差不齐、用户使用体验差异较大，同时也制约了端侧大模型产业的标准化、规模化发展。

为规范端侧大模型的更新与运维行为，明确模型更新、版本管理、生命周期管理及性能监控的技术要求和实施流程，提升端侧大模型运行的稳定性、安全性和资源适配性，降低企业研发与运维成本，保障终端用户合法权益，推动产业高质量发展，特制定本文件。

本文件的制定结合端侧大模型的技术特征和应用场景，规定了端侧大模型全生命周期更新与运维的核心技术要求，可为大模型生产企业、研发机构提供技术依据，也可为相关产品测试、行业监管提供参考。

端侧大模型更新与运维技术规范

1 范围

本文件规定了端侧大模型的更新方式、版本管理、生命周期管理、性能监控的技术要求和实施规范。

本文件适用于搭载大模型的智能移动终端、智能穿戴设备、智能车载终端、智能家居终端等终端设备，其他呆在端侧大模型的终端可参照执行。

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 45288.1 人工智能 大模型 第1部分：通用要求

3 术语和定义

GB/T 45288.1界定的以及下列术语和定义适用于本文件。

3.1

端侧大模型 large model on terminal side

部署在终端本地，具备大参数、强推理能力，可实现本地智能感知、分析、决策的人工智能模型。

3.2

模型增量更新 incremental update of model

仅对端侧大模型的更新差异部分进行下载和部署的更新方式。

3.3

模型残留清理 residual cleaning of model

端侧大模型卸载后，对终端本地残留的模型数据、配置文件、缓存文件等进行彻底清除的操作。

4 总体概述

本文件围绕端侧大模型全生命周期管理核心需求，构建端侧模型更新与运维一体化技术规范体系，明确各环节的技术要求、实施流程和操作准则。

本文件以安全性、稳定性、兼容性、用户可控性为核心原则，针对端侧大模型更新方式不统一、版本管理无规范、生命周期管控缺失、性能监控指标模糊等行业痛点，依次规定模型增量、全量、差分三种更新方式的技术要求与执行流程，统一版本命名、追溯、回滚、兼容的管理规范，明确模型部署、升级、卸载、残留清理的全生命周期操作要求，制定资源占用监控、错误上报、数据管理的性能监控标准，

为端侧大模型研发、生产、运维提供统一技术依据，保障端侧大模型稳定、安全、高效运行，提升终端用户使用体验。

5 模型更新方式

5.1 基本要求

模型更新应遵循安全性、稳定性、兼容性、用户可控原则，更新过程中不得损坏终端原有系统及数据，更新后模型应能与终端硬件、操作系统及其他应用正常兼容运行。

更新触发前应向用户明确告知更新内容、更新包大小、预计耗时及权限要求，由用户自主选择是否进行更新；仅紧急安全更新可在推送提示并说明紧急原因后，由用户确认再执行，禁止无提示自动更新。

5.2 更新模式及技术要求

5.2.1 增量更新

适用于模型小版本迭代、功能优化场景，更新包大小应不超过全量模型包的30%；更新过程应支持断点续传，网络中断恢复后无需重新下载全部差异包；更新完成后应自动校验差异包完整性，校验失败则回滚至更新前版本。

5.2.2 全量更新

适用于模型大版本迭代、核心架构升级场景；全量更新包应进行加密处理，传输过程采用安全加密协议；更新完成后应进行模型完整性和可用性检测，检测通过后方可投入使用。

5.2.3 差分更新

适用于多版本模型的精准更新场景，应支持基于基线版本与目标版本的差分计算，差分文件生成应保证低冗余、高适配；更新过程应支持版本溯源，可快速回滚至任意基线版本。

5.3 更新执行流程

a)终端检测模型更新包，获取更新信息并向用户推送提示；
b)用户确认后，终端下载更新包并进行完整性校验；
c)执行模型更新部署，部署过程中提供进度展示；
d)部署完成后自动进行模型功能、性能测试；
e)测试通过则完成更新，测试失败则自动回滚并向用户推送失败提示及解决方案。

6 模型版本管理

6.1 版本命名规则

模型版本采用主版本号。次版本号。修订号三级命名规则，具体要求：

a)主版本号：模型核心架构、大参数调整时递增，取值为正整数；
b)次版本号：模型功能模块新增、优化时递增，取值为正整数；
c)修订号：模型BUG修复、小范围优化时递增，取值为正整数。版本命名应唯一，且与模型更新内容一一对应。

6.2 版本追溯与归档

应建立端侧大模型版本全生命周期追溯体系，记录每个版本的研发信息、更新内容、发布时间、适用终端型号等；

已发布版本应进行归档存储，归档文件包括模型包、版本说明、测试报告等，归档期限不低于该模型停止服务后2年。

6.3 版本回滚

应支持模型版本快速回滚功能，回滚触发条件包括：更新后模型运行异常、用户主动申请、性能未达预期等；

回滚过程应在5分钟内完成（不含数据备份时间），回滚后终端恢复至原版本运行状态，且不丢失用户原有模型使用数据。

6.4 版本兼容

高版本模型应向下兼容低版本模型的核心功能接口；不同版本模型在终端上的部署目录、运行资源应相互独立，避免版本冲突。

7 端侧模型生命周期管理

7.1 模型部署

7.1.1 部署前检测

部署前应检测终端硬件配置（算力、存储、内存）是否满足模型运行要求，检测不通过则拒绝部署并提示终端适配要求；

部署前应检测终端系统版本是否与模型适配，系统版本不兼容则提示用户升级系统或放弃部署。

7.1.2 部署实施

模型部署应采用轻量化部署方案，降低对终端系统资源的占用；模型包在传输及本地存储阶段进行加密处理，部署过程中对模型包的解压、加载环节进行防篡改校验，防止模型包被篡改、窃取；

部署完成后应自动生成部署报告，记录部署时间、部署路径、资源占用情况等。

7.2 模型升级

模型升级应遵循本文件第5章的更新方式要求；升级过程中应备份当前模型版本及用户数据，备份文件存储至终端安全目录；

升级完成后应自动清理升级过程中产生的临时文件。

7.3 模型卸载

7.3.1 卸载触发

模型卸载支持用户主动触发和终端合规触发（如模型停止服务、终端硬件不兼容），触发前应提示用户备份重要模型使用数据，由用户确认后再执行卸载操作。

7.3.2 卸载实施

卸载过程应彻底删除模型主程序、运行库、配置文件等核心文件；卸载完成后应向用户反馈卸载结果，确认是否完成全部文件删除。

7.4 模型残留清理

7.4.1 清理范围

残留清理应覆盖以下内容：模型缓存文件、日志文件、临时数据、注册表项、权限配置信息等；
清理范围应明确，不得误删终端系统文件和其他应用数据。

7.4.2 清理要求

模型卸载后应自动执行一次基础残留清理；支持用户触发深度清理，深度清理应扫描终端全目录，确保残留文件清除率≥99%；
清理过程应在不影响终端正常运行的前提下进行，清理耗时不超过3分钟。

8 性能监控

8.1 资源占用监控

8.1.1 监控指标

应实时监控模型运行过程中的核心资源占用指标，指标要求及阈值如下。

表 1 模型运行过程中监控指标及参考阈值

监控指标	监控精度	合理阈值
算力占用	≤5%	不超过终端最大算力的70%
内存占用	≤10MB	不超过终端可用内存的60%
存储占用	≤50MB	不超过终端可用存储的10%
电量消耗	≤1%/小时	模型后台运行时，电量消耗不高于终端待机功耗的2倍

8.1.2 监控频率与预警

资源占用监控频率不低于1次/min；当指标超过合理阈值时，应立即向用户推送预警提示，并自动采取资源优化措施（如降低模型运行精度、暂停非核心功能）。

8.2 错误上报

8.2.1 错误分类

模型运行错误分为致命错误、严重错误、一般错误、轻微错误，具体分类要求：
致命错误：模型崩溃、无法启动、导致终端死机等严重影响终端运行的错误；
严重错误：模型核心功能失效、数据丢失、资源占用异常过高的错误；
一般错误：模型非核心功能异常、运行卡顿、结果偏差的错误；
轻微错误：模型日志记录异常、无关提示信息错误等不影响使用的错误。

8.2.2 上报要求

模型运行过程中采集的错误信息均应先向用户展示错误类型、错误描述及上报信息内容，由用户自主选择是否上报，禁止无提示自动上报；

上报信息应做脱敏处理，不得包含用户隐私信息、终端敏感数据；上报信息仅包括错误代码、错误描述、模型版本、终端硬件基础信息（不含设备唯一标识）、运行场景等必要内容。

8.2.3 错误处理

应建立错误分级处理机制，致命错误应在24小时内给出应急解决方案，严重错误应在48小时内完成修复，一般错误和轻微错误应在后续版本迭代中逐步修复；

所有错误的采集、用户上报确认、处理、修复结果应形成闭环记录，供版本优化参考，记录内容不得关联用户身份信息。

8.3 性能监控数据管理

性能监控数据应优先本地存储，支持用户选择是否进行云端备份；本地存储期限不低于 7 天，云端备份期限不低于 30 天；

监控数据应做脱敏处理，不得包含用户隐私信息、终端唯一标识等敏感内容；支持用户随时查看、导出、删除自身终端的模型监控数据，用户删除后应立即清除对应数据。