ICS 35.030
CCS L 70

团　体　标　准

T/TAF 374—2026

基于端云协同人工智能推理的隐私安全技术要求

Technical requirements for privacy and security based on end-cloud collaborative AI inference

2026-08-26 发布　　2026-08-26 实施

电信终端产业协会　发布

版权声明

本文件的版权属于电信终端产业协会，任何单位和个人未经许可，不得进行技术文件的纸质和电子等任何形式的复制、印刷、出版、翻译、传播、发行、合订和宣贯等，也不得未经允许采用其具体内容编制本团体以外各类标准和技术文件。如有以上需要请与本团体联系。

邮箱：tafrb@taf.org.cn

电话：010-82052809

目　　次

前言 II
引言 III
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　概述 2
5.1　端云协同概述 2
5.2　端云协同隐私数据保护原则 2
6　总体架构 3
7　终端隐私安全技术要求 3
7.1　基本原则 3
7.2　安全环境 3
7.3　模型安全 4
7.4　远程证明 4
7.5　密钥管理 4
8　云端隐私安全技术要求 4
8.1　基本原则 4
8.2　安全环境 4
8.3　模型参数安全 5
8.4　数据安全 5
8.5　密钥管理 5
8.6　远程证明 5
8.7　透明日志 6
9　端云数据传输隐私安全要求 6
9.1　数据传输安全 6
9.2　密钥协商 6
9.3　数据传输隐私保护 6
10　端云协同人工智能推理隐私安全能力分级 7
10.1　分级原则 7
10.2　分级方法 7
参考文献 9

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：华为终端有限公司、中国信息通信研究院、OPPO广东移动通信有限公司、北京火山引擎科技有限公司、荣耀终端股份有限公司、高通无线通信技术（中国）有限公司、安谋科技（中国）有限公司、维沃移动通信有限公司。

本文件主要起草人：阮玲宏、袁琦、王键、刘玭娉、张悦、李根、张尧、吴博峰、郭建领、王江胜、王俊超、赵晓娜、赵盈洁。

引　　言

随着大模型技术的快速发展，终端设备受限于自身算力与存储能力，往往需依托端云协同架构（即终端设备与云端服务协同计算）以提供智能化服务。然而，在该架构下，数据需在终端与云端之间传输与处理，致使用户隐私泄露风险显著增加，具体表现为敏感数据明文传输、云端过度采集信息、算法决策不透明等。当前标准尚缺乏对端云协同场景的针对性约束，难以有效应对相关隐私挑战。

端云协同架构已被广泛应用于智能家居、车载系统、穿戴设备等多个领域，但其复杂的数据链路也导致用户隐私泄露事件频发，例如家庭监控数据在云端发生泄露、语音指令未经授权被用于模型训练等。本文件通过规范技术实现路径，降低用户数据在处理过程中的暴露风险，增强公众对人工智能技术的信任，推动数字社会可持续发展。

基于端云协同人工智能推理的隐私安全技术要求

1 范围

本文件规定了人工智能推理场景下端云协同人工智能推理的隐私安全要求，包括终端隐私框架要求、云端隐私框架要求以及端云数据传输的隐私安全能力分级要求，旨在确保在利用云端资源有效提升终端人工智能应用性能的同时，充分保障用户数据的安全与隐私。

本文件适用于端云协同推理计算隐私保护的设计、开发和测试。

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

端云协同 end-cloud collaborative

通过整合终端设备（End）和云端（Cloud）的协同计算与资源调度能力，实现数据、算力和算法的动态优化分配。

3.2

机密计算 confidential computing

在可信的硬件基础上,通过隔离、加密、证明、封装等机制,保护使用中数据安全的计算模式。

[来源：GB/T 45230—2025，3.8]

3.3

会话密钥 session key

会话密钥特指一次隐私计算任务的会话密钥,用于加密一次隐私计算任务中终端与云端之间的通信数据。

3.4

远程证明 remote attestation

对机密计算环境和机密计算应用程序进行完整性校验和身份认证的过程,向验证者提供可信的状态报告。

3.5

盲化签名 Blind Signatures

盲化签名是一种密码学协议，允许签名者（如服务器）对客户端提交的盲化后的消息进行签名，而无需知晓原始消息内容。签名完成后，客户端通过去盲化操作将签名还原为对原始消息的有效签名，且该签名可被第三方公开验证。

3.6

盲化因子 Blinding Factor

盲化因子是一个随机生成的加密参数（通常为大整数），用于在盲签名过程中对原始消息进行数学变换（如乘性盲化），确保签名者无法追踪消息来源。

3.7

匿名传输协议　Oblivious HTTP

将客户端身份和请求内容分开，并通过可信中继服务转发加密的请求。通过移除客户端的元数据（如IP地址），匿名传输使目标服务器无法识别客户端，从而防止对其进行跟踪。此协议使用端到端加密来隐藏请求细节，并在客户端和目标服务器之间使用中继来匿名化流量。

4　缩略语

下列缩略语适用于本文件。

Guest OS：客户机操作系统（Guest Operating System）

Host OS：宿主机操作系统（Host Operating System）

TEE：可信执行环境（Trusted Execution Environment）

TPM：可信平台模块（Trusted Platform Module）

OHTTP：匿名HTTP（Oblivious HTTP）

TLS：传输层安全协议（Transport Layer Security）

ECDH: 椭圆曲线迪菲-赫尔曼密钥交换(Elliptic Curve Diffie-Hellman)

IAK：身份认证密钥（Identity and Access Key）

KMS：密钥管理系统（Key Management System）

MD5：信息-摘要算法5(Message-Digest Algorithm 5)

TSL：传输层安全协议（Transport Layer Security）

SM2：SM2椭圆曲线公钥密码算法（SM2 Public Key Cryptographic Algorithm）

5　概述

5.1　端云协同概述

随着大模型技术的快速发展，终端设备受限于自身算力与存储能力，需依托端云协同架构（即终端设备与云端服务协同计算）以实现更复杂的推理功能（如动态模型切换、多任务并行处理），同时又能借助云端弹性扩展能力，以高性价比处理图像识别、语音转写等大规模人工智能推理任务，应对突发流量对系统稳定性的冲击。然而，端云协同推理模式在数据安全与隐私保护方面面临独特挑战。

a）　环境安全风险：人工智能推理需用户上传数据至云端处理，若云端存储或访问权限控制不当，易导致用户数据泄露。

b）　传输安全风险：推理请求与结果的传输过程面临中间人攻击、数据篡改等威胁，可能泄露用户信息，危害隐私与系统安全，因此保障传输安全是端云协同服务可信的关键。

c）　审计不透明风险：传统推理模型因其“黑盒”特性，数据处理逻辑难以自证透明，加剧用户不信任感，并掩盖潜在的数据滥用风险，亟需引入第三方进行可审计性验证。

5.2　端云协同隐私数据保护原则

在使用端云协同架构向用户提供推理服务时应遵循以下基本原则。

a）　数据加密传输：用户数据需经过加密处理才能出终端，数据加密密钥随机生成并由硬件保护，且只用于一轮交互；数据流经开放网络时采用安全的传输方式，保障全链路端到端隐私安全；

在云端计算节点进行处理时处于可信执行环境中，严格遵循用户定义策略进行数据处理。

b） 隐私保护：用户调用推理服务时，需提前获取匿名凭证，在服务端对用户进行匿名校验时，仅授权用户可使用的推理服务，云端服务仅看到匿名请求，无法关联真实用户身份或构建用户行为画像；网络采用隐私中继协议隐匿传输路径，无法从云端逆向溯源至发起请求的用户，防范对特定用户的针对性攻击。

c） 可追溯性与透明性：在提供推理服务前，需证明节点处于可信状态，用户只有在验证节点可信的情况下才会发起推理请求；推理计算节点遵循最小权限和最小攻击面原则，提供可信运行环境、可观测接口、及透明日志，实现操作的可见、可查与可证。

6 总体架构

本文给出端云协同人工智能推理的隐私安全参考框架如图 1 所示，由终端隐私计算区、安全会话通信、云端隐私计算区三部分组成。

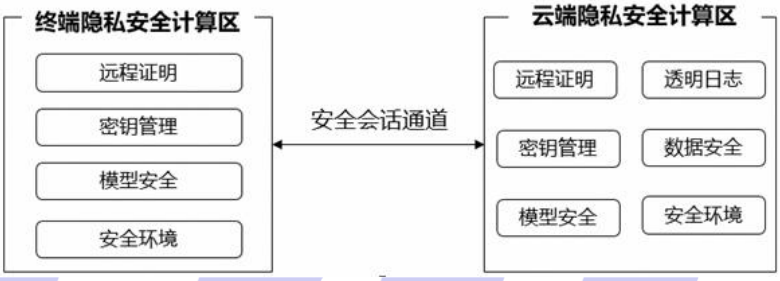

图 1 端云协同人工智能推理隐私安全架构

各部分应分别实现以下能力。

a） 终端隐私安全计算区：负责实现设备终端和推理节点之间，远端服务器的身份真实性的认证，可信连接的建立，以及与远端服务器安全会话密钥的协商和保护。

b） 隐私安全通信：基于 TLS 等安全传输通道建立会话，通过数据脱敏、身份隐匿和通信混淆等技术，通信双方无法通过元数据推断真实身份，交互链路具备抗流量分析能力。

c） 云端隐私安全计算区：负责提供自底向上可验证的隐私计算环境，如提供可信软硬件以及隐私计算服务提供方所使用的接口，同时保障数据安全和推理安全，通过最小化运行时等最大化保障数据隐私安全不泄露。

7 终端隐私安全技术要求

7.1 基本原则

遵循终端处理优先的原则，对于生物识别信息（如指纹、人脸、声纹）、地理位置轨迹、健康数据等敏感信息，需优先采用终端处理模式。应依托终端安全环境（如可信执行环境、安全芯片等）完成数据的解析与运算，以保护敏感信息及终端模型的机密性和完整性。例如，在语音指令识别过程中，原始音频数据应在终端完成转换后，仅将处理后的文本指令上传至云端，原始音频数据在终端及时销毁，确保数据生命周期的全程安全。

7.2 安全环境

终端在处理用户隐私数据时，应满足以下环境要求：

a) 终端应支持基于处理器架构的特权级隔离机制，如TEE等，构建独立于通用执行环境的安全执行区域，实现核心安全功能与普通业务的基础隔离；
b) 终端安全环境镜像应支持全生命周期确保来源合法性和完整性，包括启动、升级；
c) 终端安全环境镜像应支持防逆向设计，如镜像加密、符号表混淆；
d) 支持物理级隔离的设备，如集成或独立的安全芯片，应支持安全核具备独立的计算资源。

7.3 模型安全

终端在使用推理模式时，应满足以下要求：

a） 终端应对终端模型实施严格的访问控制策略，确保只有经过授权的人工智能应用才能访问和使c用终端模型进行推理，防止未授权的访问、滥用、逆向分析和模型窃取；
b） AI模型文件、模型参数、权重及相关配置信息的应加密存储，密码算法应符合国家密码管理机构的要求和相关密码算法标准；
c） 系统应利用可信执行环境（TEE）保障终端侧模型运行的安全性。

7.4 远程证明

终端应在可信执行环境中验证云端提供的远程可信证明报告，确保云端推理服务环境安全性和完整性得到验证，防止假冒或篡改的云端服务对用户数据进行处理，远程证明验证报告应包含以下内容：

a) 应对令牌发行者证书链的校验；
b) 应包含对可信计算服务器凭据证书的验证；
c) 应包含周期性会话公钥来源证明；

7.5 密钥管理

终端用户的数据、推理模型等加解密的密钥应满足以下要求：

a) 密钥应支持分层管理，根密钥不直接用于业务数据加密，用于派生保护会话密钥，如：通过根密钥派生密钥加密密钥，通过密钥加密密钥保护本地加密密钥；
b) 根密钥应存储并运行于安全区域，由安全芯片保护，仅安全芯片的软件或特定的硬件引擎（如加密引擎）可访问，无法被外部获取；
c) 密钥全生命周期管理应由可信执行环境实现，被托管的密钥明文不出可信执行环境；
d) 用于加解密的私钥或对称密钥禁止硬编码在代码中；
e) 高安全要求设备的密钥应部署在安全芯片上；
f) 数据加密密钥应由硬件保护，且只用于一轮交互。

8 云端隐私安全技术要求

8.1 基本原则

云端机密计算环境的隐私保护技术，旨在为用户提供安全、私密的计算服务。通过加密技术和数据隔离机制，确保用户的敏感信息在云端环境中不被未授权的第三方访问或泄露；采用端到端加密和严格的身份验证机制，确保数据在传输和存储过程中的安全性；所有敏感数据均在可信执行环境中处理，防止数据被窃取或篡改，同时，支持操作日志记录和审计功能，确保所有数据处理操作可追溯，并提供透明的隐私政策说明，让用户清楚了解数据的使用方式。

8.2 安全环境

不具备机密计算环境的要求时，云端模型可部署在安全沙箱或者安全容器中，实现模型运行环境与主机操作系统、其他应用/模型的强隔离，应分别满足下述安全要求。

a） 内核应进行最小化裁剪与安全加固，移除不必要的系统调用、驱动程序及服务，关闭冗余端口与通信协议，降低攻击面；应基于最小化基础镜像构建，移除冗余组件与敏感信息，降低安全风险。

b） 运行时应采用权限最小化原则，禁止以root用户运行，严格限制系统调用权限；应实现进程级、网络级、存储级隔离，禁止非授权通信与资源访问。

8.3 模型参数安全

云端大模型参数应满足下述安全要求。

a） 应支对大模型核心参数（权重、偏置）进行加密存储。

b） 大模型加载后，核心参数应在安全内存中采用动态加密机制（加密密钥支持动态更新），中间推理结果采用一次性会话密钥加密存储，会话结束后立即销毁密钥。

8.4 数据安全

构建细粒度、多层次的数据隔离体系，保障大模型运行过程中数据安全。数据安全应满足下述要求：

a） 应支持按照数据类型的敏感等级分类，不同等级数据存储于不同安全区域，核心参数存储于硬件加密存储模块，推理输入/输出数据存储于加密文件系统，中间计算结果仅在安全内存中临时存储；

b） 大模型推理节点应与其他业务节点部署在隔离的物理网络环境，推理数据传输应采用专用加密通道；

c） 大模型推理进程应与系统及其他应用进程实施安全隔离；采用进程白名单机制，仅允许授权的大模型推理相关进程运行，并禁止无关进程的启动与内存访问；

d） 应支持“用完即删”的能力，推理过程中的数据零残留，会话结束后应立即删除的内容包括：推理输入数据（明文及其加密副本）、中间计算结果、会话临时密钥、会话日志、内存参数缓存；

e） 支持在推理会话正常结束后的指定时间窗口内（如30s内），应能自动化触发数据删除流程。

8.5 密钥管理

密钥管理应满足下述要求：

a） 加密密钥采用密钥管理系统（KMS）分布式管理，密钥分片存储于不同安全节点，访问需多重权限校验；

b） 大模型加密密钥应与模型数据分域存储，密钥生成、分发、使用、销毁全程处于安全执行环境，支持密钥生命周期自动化管理，定期轮换密钥。

8.6 远程证明

为确保云端可信计算节点是一个合法节点，构建端云协同的信任基础，确保端云通信与交互的安全可靠。云端计算节点远程证明应满足以下要求：

a） 远程证明应基于硬件信任根（如TPM）实现，证明结果由硬件安全模块存储保护；

b） 应包含虚机/容器实例的完整性证明结果；

c） 应包含安全环境的度量验证结果；

d） 应具备抵抗防重放攻击机制（如采用时间戳/Nonce值校验）；

e） 从硬件平台、Host OS至Guest OS的逐级可信度量；

f） 支持对计算集群内所有节点设备的批量证明能力。

8.7 透明日志

具备的安全要求包括：

a） 透明日志应包含远程证明验证基线、远程证明结果、远程证明证据；

b） 透明日志记录的日志应满足一致性、完整性、不可删除、不可篡改，可追溯，可查询；

c） 透明日志应包含 IAK 证书链、完整性校验证书链、远程证明服务器证书链。

9 端云数据传输隐私安全要求

9.1 数据传输安全

具备的安全要求包括：

a） 应使用安全可靠的传输协议，如SSL/TLS等，防止数据在传输过程中被篡改、窃取或恶意攻击，确保数据的完整性和机密性；

b） 采用加密算法对敏感数据进行加密处理，使用的密码算法应符合国家密码管理有关标准规范，禁止使用不安全算法（如：MD5等），确保数据在传输过程中的安全。

9.2 密钥协商

具备的安全要求包括：

a） 会话密钥应支持周期性更新，应对每个通信会话或特定操作仅使用一次密钥，确保长期密钥仅用于密钥协商，而非直接参与数据加密，从而降低长期密钥泄露后对历史通信数据的追溯风险；

b） 密钥协商应满足下述要求：

1） 应在终端对云端完成远程证明后进行应用层密钥协商（如ECDH、SM2等）；

2） 应在端云两侧的高安计算环境内完成应用层密钥协商、使用和销毁；

3） 应仅支持单次隐私计算会话，会话结束后应用层会话密钥立即销毁；

4） 应具备抵抗重放攻击的能力。

9.3 数据传输隐私保护

9.3.1 账号匿名化要求

采用盲化签名技术实现身份标识匿名化，减少用户身份暴露，应满足下述要求：

a) 在模型交互、数据传输等流程中，应替换为一次性或短期有效（最长不超过业务会话周期）的匿名 token；

b) 服务器侧不应将匿名 token 与用户行为数据、模型使用记录进行长期关联存储；

c) 匿名化后的 token 仅用于会话关联，不应用于模型训练、推理结果追溯；

d) 客户端应生成盲化因子对会话请求进行加密处理，服务端对盲化请求签名后返回客户端，客户端解盲处理获得匿名化令牌，作为业务会话凭证。

9.3.2 设备匿名化要求

为了防止设备标识符被追踪、泄露等安全风险，应满足下述要求：

a) 设备唯一标识符不应在非安全通道明文传输，不应在通用存储区域明文持久化存储，终端仅允许安全芯片留存原始设备唯一标识符，服务器侧仅接收匿名化设备标识；

b） 在端云进行设备身份认证时，应采用匿名化后的设备证书；
c） 服务器不应基于匿名化设备标识进行设备画像与精准关联。

9.3.3 IP 匿名化要求

为确保终端设备IP地址被追踪，应满足下述要求。

a） 为确保访问来源信息的有效隔离与隐私保护，在发送端设备的中间件应清除可能泄露原始 IP 的请求头字段，目标服务器仅能获取代理服务器出口 IP，无法追溯设备真实 IP。
b） 终端设备在与云端交互时，应部署支持 Oblivious HTTP（OHTTP）协议（RFC 9458）的可信中继服务，终端请求经可信中继服务，减轻利用 IP 地址和连接信息等元数据来识别客户端的情况，中继服务剥离客户端 IP 等网络层标识信息，源服务器仅能识别中继服务身份，无法关联多请求至同一客户端，减少客户端 IP 类身份暴露。

10 端云协同人工智能推理隐私安全能力分级

10.1 分级原则

端云协同人工智能推理的隐私安全能力分级原则：

a） 场景适配原则：分级应充分考虑端云协同架构中数据流向的差异性，端侧安全处理、端云安全传输、云端安全处理等不同场景下的隐私安全需求；
b） 能力递进原则：分级从基础级到增强级逐级递进，高等级在满足低等级全部要求的基础上，增加更严格的隐私安全能力要求；
c） 风险导向原则：分级应以数据泄露、隐私暴露等风险程度为核心依据，重点考量端云协同场景中敏感数据传输泄露、云端过度采集、算法决策不透明等特有风险；
d） 技术可行原则：分级要求应结合当前端侧算力水平与云端技术能力设定，既体现技术引领性，又确保分级要求具备产业落地可行性；
e） 保护深度原则：隐私安全能力应贯穿数据全生命周期，涵盖采集、传输、处理、存储、销毁各环节，实现纵深防御。分级要求从单点防护向体系化防护演进。

10.2 分级方法

根据端云协同人工智能推理隐私安全能力的程度，分为基础级和增强级，针对不同的等级定义相应的隐私安全能力的最小集合，具体的等级要求详见表 1。

表1 基于端云协同推理的隐私安全能力分级

技术要求		基本级要求对应章节号	增强级要求对应章节号
终端隐私框架要求	安全环境	7.2 a)～c)	7.2 a)～d)
	模型安全	7.3 a）	7.3 a）～c)
	远程证明	7.4 a）～b)	7.4 a）～c)
	密钥管理	7.5 a）～d)	7.5 a）～f）

表 1　基于端云协同推理的隐私安全能力分级（续）

技术要求		基本级要求对应章节号	增强级要求对应章节号
云端隐私框架要求	安全环境	8.2 a）～b)	—
	模型参数安全	8.3 a）～b)	—
	数据安全	8.4 a）～d)	8.4 a）～e）
	密钥管理	8.5 a）～b)	8.5 a）～b)
	远程证明	8.6 a）～d)	8.6 a）～f)
	透明日志	8.7 a）～b)	8.7 a）～c)
端云数据传输隐私安全要求	账号匿名化要求	—	9.3.1
	设备匿名化要求	—	9.3.2
	IP 匿名化要求	—	9.3.3

参 考 文 献

[1] GB/T 25069—2022 信息安全技术 术语
[2] GB/T 45230—2025 数据安全技术 机密计算通用框架
[3] RFC-9474 RSA Blind Signatures
[4] RFC-9458 Oblivious HTTP

电信终端产业协会团体标准

基于端云协同人工智能推理的隐私安全技术要求

T/TAF 374—2026

*

版权所有 侵权必究

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