ICS 35. xxx
CCS Lxx

团　　体　　标　　准

T/ISC XXX—XXXX

人工智能终端模型训练数据集建设通用指南

General Guidelines for the Construction of Training Datasets for AI Terminal 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

随着人工智能技术由云端向端侧深度演进，人工智能手机、个人电脑、智能座舱及物联网设备等人工智能终端模型的应用日益广泛。人工智能终端模型的训练与部署是一项跨越云端与边缘端、涉及复杂软硬件协同的系统工程。与传统云端大模型相比，终端模型不仅面临严苛的内存、存储及系统输入输出瓶颈，且其数据流转涉及云端原生、由云向端、由端向云及端侧原生驻留等复杂物理边界，这对数据的高效封装、特征对齐以及隐私合规管控提出了前所未有的挑战。

为有效支撑人工智能终端模型的规模化落地，迫切需要建立一套贴合终端模型由云向端演进动态链路的数据集建设规范。本文件基于终端模型演进的物理规律，提出了端培育期、轻量化迁移期、多端协同期、端侧微调期四个递进阶段的总体建设架构。

本文件旨在明确基础与通识预训练、知识蒸馏与对齐、量化校准、联邦协同、端侧微调五类终端训练数据集的形态构成与核心用途。此外，结合跨端数据流转的隐私合规边界与终端受限资源的底层约束，从安全要求、构建方法与格式规范三个核心维度，阐述了各类数据集在脱敏隔离、特征抽取及轻量化封装等方面的处理准则。本文件的发布将为产业界在终端模型训练数据集的建设、处理与管理方面提供通用指南，助力构建安全、高效、合规的端侧人工智能生态。

人工智能终端模型训练数据集建设通用指南

1　范围

本文件给出了人工智能终端模型训练数据集建设的通用指南，包括练数据集建设的总体框架、分类方案、安全要求、构建方法及格式规范。

本文件适用于面向人工智能终端模型各训练演进环节，数据集建设、处理及安全管控相关工作。本文件不适用于专门用于模型性能评测的测试数据集建设，相关要求见及其他专项评测数据集指南。

2. 引言本文件可供开展终端人工智能模型研发的各类企业、科研机构、终端设备制造商，以及第三方人工智能数据服务提供商在进行端侧数据工程实践时参考使用。

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 41867—2022　信息技术　人工智能　术语

3　术语和定义

GB/T 41867—2022界定的以及下列术语和定义适用于本文件。

3.1

人工智能终端模型　artificial intelligence terminal model

部署于智能手机、个人电脑、智能座舱及物联网节点等边缘侧物理设备上，受限于端侧算力、内存与存储资源，在本地执行推理或轻量化微调的人工智能模型。

3.2

知识蒸馏　knowledge distillation

利用特定数据集将复杂大模型所具备的泛化能力、逻辑推理过程或概率分布特征，转移至适合在终端运行的轻量级小模型中的能力迁移技术。

3.3

训练后量化　post-training quantization

在模型训练完成后，利用包含极小规模无标注特征集合的校准数据集统计激活范围，将模型权重或激活值从高精度浮点数向低比特精度转换，以适配终端底层硬件算力的模型压缩技术。

3.4

联邦协同　federated collaboration

在各物理终端敏感原始数据不出域的安全前提下，通过利用本地切分数据进行计算，并仅与云端交互加噪后的参数或梯度差值，从而进行全局模型联合训练与权重迭代的技术范式。

3.5

端侧微调 on-device fine-tuning

模型在终端本地闭环部署后，利用设备产生的私有数据与交互日志，在受限的终端资源下进行的轻量化参数更新，以实现端侧模型的持续个性化迭代。

3.6

非独立同分布特征 non-independent and identically distributed characteristics

在终端协同场景中，由于各物理设备使用者的习惯差异，导致分散驻留在不同终端本地的数据样本呈现出极端数据倾斜与显著统计学差异的数据分布状态。

3.7

差分隐私加噪 differential privacy noise addition

在终端向云端外传模型参数、梯度差值或中间特征前，于终端物理边界内部通过算法注入数学噪声的隐私保护方法，旨在防止攻击者通过模型反转攻击推断出本地数据的原始特征。

4 符号和缩略语

本文件没有需要界定的符号和缩略语。

5 总体框架

5.1 建设架构

人工智能终端模型训练涉及云端算力与端侧物理设备的协同。由于模型训练在云端与终端面临着计算资源、隐私边界等差异，宜分阶段构建具有特定形态与合规属性的训练数据集，以满足训练模型不同阶段的特定需求。人工智能终端模型训练数据集建设总体框架如表1所示。

模型训练阶段	**云端培育期**	**轻量化迁移期**		**多端协同期**	**端侧微调期**
数据集类型	**基础通识数据集**	**知识蒸馏与对齐数据集**	**量化校准数据集**	**联邦协同数据集**	**端侧微调数据集**
安全要求	• 敏感信息滤除 • 价值观与合规对齐 • 知识产权保护	• 记忆化隐私防范 • 提示词注入防御	• 物理形态阻断 • 防逆向工程	• 不出域底线 • 跨端交互加噪	• 本地沙箱隔离 • 动态汰换与安全销毁
构建方法	• 多源采集与汇聚 • 质量清洗与去重 • 分词与序列化	• 指令模板构造 • 逻辑与特征提取 • 映射对合成	• 代表性子集采样 • 前向激活提取 • 离群极值保留	• 本地日志感知 • 异构特征切分与对齐 • 差值计算与封装	• 偏好意图转化 • 资源适配裁剪 • 动态滑动管理
格式规范	列式存储或流式分块封装格式	多层嵌套的结构化封装格式	多维张量的二进制文件格式	集成安全载荷的通信包封装格式	轻量化二进制序列化框架

表 1 人工智能终端模型训练数据集建设参考指南

5.2 模型训练阶段与数据分类

数据集的建设宜贴合终端模型由云向端的实际训练与落地步骤，按需提供相应的数据支撑。本架构将数据集建设划分为四个递进的工程阶段：

人工智能终端模型的训练可划分为四个递进的工程阶段。数据集的建设宜贴合以下模型阶段的特定场景，对应构建五类专属形态的数据集：

a） 云端培育期：对应基础通识数据集。在云端算力集群训练基础大模型阶段，重点建设海量、多模态的通识语料与数据。

b） 轻量化迁移期：对应知识蒸馏与对齐数据集和量化校准数据集。在模型由大变小、向终端设备下发阶段，重点建设这两类高特征代表性数据集，以保障模型压缩后的精度与性能。

c） 多端协同期：对应联邦协同数据集。在跨设备联合建模阶段，重点建设该类数据集，支撑多设备数据在“不出域”前提下的本地切分与联邦对齐。

d） 端侧微调期：对应端侧微调数据集。在模型本地闭环部署后，重点建设该类数据集，规范端侧个性化微调数据的生成与动态淘汰，实现端侧模型的持续迭代。

5.3 数据集建设要素

针对上述各类数据集，从以下三个维度的核心要素提供建设指南：

a） 安全要求：依据不同的跨端数据流向，为各类数据集设定专属的隐私边界、出域防范与脱敏底线。。

b） 构建方法：在满足合规底线的前提下，规范合成数据、特征抽取、联邦切分、滑动淘汰等终端特有数据构建方法。

c） 格式规范：针对终端受限算力与存储特性，规范数据集的张量对齐要求与内存友好型二进制封装格式。

6 终端模型训练数据集分类

6.1 基础通识预训练数据集

a） 数据构成：通常由超大规模、知识覆盖面广且无特定终端场景指向性的通用语料和跨领域多模态数据构成。该类数据宜经过清洗、去重与无害化过滤，并以标准的通用结构化格式存储。

b） 核心用途：为云端大模型注入广泛的世界知识储备与基础认知泛化能力，为后续向终端轻量化迁移提供强大的知识底座。

6.2 知识蒸馏与对齐数据集

a） 数据构成：通常由云端大模型生成并向下传递的数据集合，通常包含高质量的指令输入提示词、云端模型的输出概率分布、思维链推理路径、高质量的合成回复对等。

b） 核心用途：通过模拟和提取云端大模型的认知逻辑，指导终端轻量化模型进行对齐训练，保障模型在参数被大幅压缩后，能维持较高的推理精度。

6.3 量化校准数据集

a） 数据构成：由从通用数据或真实终端场景中抽取的极小规模无标注特征集合构成。通常直接提取为具有统计分布特征的底层张量或激活特征，形成小规模、高代表性的标定子集。

b）核心用途：在模型由浮点运算向终端受限的定点运算转化时，用于校准激活值的动态范围并保留关键离群极值，降低量化过程造成的精度损失。

6.4 联邦协同数据集

a） 数据构成：由端侧感知的多元交互日志、具备非独立同分布特征的本地切分数据，以及加噪脱敏处理后的参数和梯度差值加密包等构成，驻留于各物理终端本地，通常以参数或梯度的形态参与云端交互，不被中心化汇聚。

b）核心用途：在敏感数据不出域的安全前提下，支撑终端与云端协同进行全局模型的联合训练与权重迭代，提升模型在长尾场景下的鲁棒性。

6.5 端侧微调数据集

a） 数据构成：由终端用户日常交互行为自动生成。包含用户偏好记录、高频应用交互快照及精简后的低资源指令模板，全生命周期闭环驻留于设备本地。

b） 作为终端专属的个性化经验库，支撑模型在本地执行高效的轻量级微调，实现端侧模型越用越懂用户的持续个性化迭代。

7 安全要求

7.1 总体要求

终端模型训练涉及云端算力与端侧物理设备的跨界交互，数据集的建设宜遵循“隐私边界与流向管控”的底层原则。针对基础通识、知识蒸馏与对齐、量化校准、联邦协同和端侧微调五类特定数据集，建设方宜根据其数据流转方向的差异，从数据采集清洗、跨端流转防泄露、端侧安全存储与动态销毁等方面保障数据集全生命周期的安全性。

7.2 基础通识数据集安全要求

a）敏感信息滤除：在云端汇聚处理阶段，宜采用自动化与人工抽检结合的方式，清洗语料中的个人身份信息、敏感地理位置及未授权的商业机密。

b）价值观与合规对齐：宜建立无害化过滤机制，拦截并剔除包含暴力、色情、歧视及违反法律法规的有害内容。

c）知识产权保护：汇聚公开网页与代码数据时，宜遵循相关开源协议与版权声明，规避侵权风险。

7.3 知识蒸馏与对齐数据集安全要求

a）记忆化隐私防范：鉴于云端预训练模型可能存在隐私数据记忆现象，在生成指令提示词及合成回复对向下传递前，宜进行二次脱敏审查，防止云端训练阶段的敏感碎片数据通过蒸馏过程向下泄露。

b）提示词注入防御：宜对生成的指令输入提示词进行清洗与安全性评估，过滤掉具有诱导性、尝试绕过安全策略的恶意指令样本。

7.4 量化校准数据集安全要求

a）物理形态阻断：为防范原始隐私泄露，量化校准数据集宜避免使用具有明确语义的原始多模态文件，仅提取并保留无语义指向的底层张量或激活特征。

b）防逆向工程：提取的底层张量或激活特征宜满足不可逆要求，防止攻击者通过特征反向重构或恢复出原始的用户敏感数据。

7.5 联邦协同数据集安全要求

a）不出域底线：宜建立系统级隔离机制，确保构成数据集的多元交互日志与本地切分数据驻留于终端物理设备本地，避免将其明文或原始文件上传至云端或进行中心化汇聚。

b）跨端交互加噪：在参与云端模型交互时，宜对外传的参数、梯度差值或加密中间特征实施差分隐私加噪、同态加密或安全多方计算等保护措施，防止攻击者通过模型反转攻击推断出本地切分数据的原始特征。

7.6 端侧微调数据集安全要求

a）本地沙箱隔离：宜存储于受终端操作系统管控的可信执行环境或安全沙箱中，配置访问控制权限，阻断未经授权的第三方应用或越权进程的读取请求。

b）动态汰换与安全销毁：针对闭环驻留于设备本地的极小规模数据集，宜结合终端存储限制建立滑动窗口淘汰机制。被淘汰的数据宜执行物理级安全覆写与销毁，防止数据残余被恶意恢复。

8 构建方法

8.1 总体原则

数据集的构建宜贯穿从数据准备、模型训练到端侧部署的完整生命周期。建设方宜根据特定数据集的用途、终端物理硬件的算力存储约束以及安全隐私管控边界，统筹实施差异化的数据采集、特征抽取、切分与合成工艺。

8.2 基础通识数据集构建方法

a）多源采集与汇聚：宜从公开网络、开源社区及商业授权渠道，广泛采集覆盖多语言、多领域及多模态维度的原始语料。

b）质量清洗与去重：宜采用启发式规则结合分类模型，剔除乱码、低信息密度及无意义的重复片段，提升语料的信噪比。

c）分词与序列化：针对文本与多模态序列，宜采用统一的子词分词算法进行切分，并转换为模型可读取的索引序列。

8.3 知识蒸馏与对齐数据集构建方法

a）指令模板构造：宜面向终端常见交互场景，构建覆盖问答、摘要、翻译等多任务类型的多样化输入指令，支撑终端模型应对各类长尾调用意图。

b）逻辑与特征提取：将输入指令输入云端模型，宜完整记录并提取其前向推理过程中的中间层输出概率分布及多步推理路径。

c）映射对合成：整合指令输入与云端模型的高质量输出，构建结构化的问答映射集，作为终端模型对齐训练的参考依据。

8.4 量化校准数据集构建方法

a）代表性子集采样：宜从真实通用语料中采用聚类或特征空间均匀采样法，抽取极小规模的无偏子集。针对训练后量化，可省略目标标签信息仅保留输入特征。数据量宜控制在128至1024个批次等特定规模内。

b）前向激活提取：宜将采样数据输入全精度模型进行前向传播，计算并截取特定网络层的底层激活特征。

c）离群极值保留：在激活特征提取与标定过程中，宜统计算法识别并完整保留对模型精度影响显著的长尾离群极值。

8.5 联邦协同数据集构建方法

a）本地日志感知：宜依托操作系统底层接口，持续记录用户真实交互序列与反馈。端侧数据的感知切分宜在设备处于充电、休眠且具备网络连接的非活跃状态下动态执行。

b）异构特征切分与对齐：针对终端设备间数据分布极度不均衡的特性，宜采用本地聚类、标签平滑或分布重采样算法进行合理划分。此外，宜在本地保留少量全局通用分布的代理数据，防止发生灾难性遗忘。

c）差值计算与封装：在本地完成一轮模型迭代后，宜计算更新前后的参数矩阵差异，并将其封装为待加噪的差值特征包。

8.6 端侧微调数据集构建方法

a）偏好意图转化：宜实时解析终端高频应用界面的操作快照与触控事件，将其结构化转化为具备特定意图的本地微调指令。

b）资源适配裁剪：针对受限内存与算力上限，微调指令模板应进行高度压缩与结构化，剔除冗余自然语言修饰词，执行文本摘要或长度截断，降低上下文窗口开销。

c）动态滑动管理：宜在终端本地分配固定存储区块，建立基于时间轴的先进先出滑动窗口动态淘汰陈旧数据。针对高价值核心意图数据，宜建立重要度评分机制予以豁免淘汰并持久化固化。

9 格式规范

9.1 通用规范

针对终端设备算力受限、内存资源紧张及异构计算硬件的物理特性，各类数据集的存储与读取格式宜满足轻量化、低开销与硬件友好的要求。各类数据集宜在底层格式封装时统一写入结构化的扩展元数据标签：

a）数据出域权限标签：宜明确标定数据集的云端原生、由云向端或端侧原生驻留等空间流转属性。对于标记为端侧原生驻留的微调或联邦数据，操作系统底层安全组件宜识别该标签，并在系统级网络输入输出层主动实行阻断，从物理层规避越权外发风险。

b）硬件算力与精度对齐标签：宜明确目标部署物理硬件的架构标识与数据的低比特量化精度类型标记，以指导终端进行底层算子调度与内存预分配。

用于模型评测的测试数据集，其元数据字段可在本通用标签基础上扩展模态专用字段，扩展方式宜与本文件保持兼容。

9.2 基础通识数据集格式规范

宜采用业界通用且高效的列式存储或流式分块封装格式，以支持云端训练时的按需流式读取，并适配快速检索引擎。

9.3 知识蒸馏与对齐数据集格式规范

宜采用多层嵌套的结构化封装格式。格式设计宜支持将用户指令、思维链推理路径、输出概率分布张量等核心维度作为独立字段进行序列化，支撑终端按需提取特定字段。

9.4 量化校准数据集格式规范

直接封装为多维张量的二进制文件。文件中张量的批量、通道、高度、宽度等维度排布顺序宜与终端底层加速算子要求的显存排布格式保持一致。数据封装宜支持内存映射技术，以实现物理存储与计算单元之间的零拷贝传输。

9.5 联邦协同数据集格式规范

宜采用集成安全载荷的通信包封装格式。封装头信息宜包含隐私预算等差分隐私噪声分布参数，内容体宜采用安全聚合协议封装二进制权重差值，并在传输层附加哈希校验码。

9.6 端侧微调数据集格式规范

宜采用轻量化二进制序列化框架进行封装。存储格式宜原生支持内存映射技术，确保终端模型在读取数据时直接执行零拷贝寻址计算，降低处理器负载与内存分配开销。

参 考 文 献

[1] 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
[2] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程
[3] GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范
[4] GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范