

《人工智能终端模型训练数据集建设 通用指南》标准编制说明

标准起草小组

1. 标准范围

本标准确立了人工智能终端模型训练数据集建设的总体框架，规定了数据集分类体系、特殊训练范式数据集构建方法、以及隐私计算与安全合规等方面的通用技术要求。

本标准适用于面向AI手机、AI PC、智能座舱、物联网等终端设备的人工智能模型，在知识蒸馏、训练后量化、联邦协同、端侧微调等训练场景下的数据集建设、处理、管理与评测。可供相关互联网企业、科研机构、终端设备制造商及第三方数据服务商参考使用。

2. 工作简况

当前，全球人工智能技术正经历由“云端大模型”向“端侧轻量化模型”深度演进的关键拐点。我国智能终端产业链虽具备极强的硬件先发优势，但在实际工程落地中，国内企业普遍受制于端侧资源瓶颈、多端分布式数据异构以及跨端隐私合规挑战。针对这些边缘计算范式，国内外尚未有专门针对端侧模型训练的数据集建设通用标准，底层数据处理与标准化构建处于“孤岛”状态，产业界亟需一套打破硬件壁垒的技术指南。为有效支撑人工智能终端模型的规模化落地，本标准的制定旨在填补缺乏“端侧专用训练数据集规范”的空白，统一终端模型特定演进阶段的数据集分类体系、构建方法及隐私安全底线。

3. 标准编制原则和确定标准主要内容

本标准依据《标准化工作导则 第1部分：标准化文件的结构和起草规则》（GB/T 1.1—2020）编制。标准编制原则包括：

科学性：紧扣人工智能终端模型全生命周期演进路线，构建面向云端培育期、轻量化迁移期、多端协同期、端侧微调期四个递进阶段的标准化数据集处理总体技术架构。

系统性：从数据集分类、安全要求、构建方法与格式规范四个核心维度，系统规范基础通识预训练、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、核心用途及处理准则。

可操作性：针对终端受限算力与存储特性，明确激活特征提取、离群极值保留、异构特征切分、差值特征包封装、指令资源裁剪及滑动窗口动态淘汰等具体可落地的技术要求。

本标准核心内容包括：

（1）确立端侧演进总体架构。明确四个递进工程阶段与五类专属形态数据集的对应关系，以及安全要求、构建方法、格式规范三大建设要素。

（2）规范五类终端训练数据集分类。围绕终端硬件受限及分布式环境特征，明确基础通识预训练数据集、知识蒸馏与对齐数据集、量化校准数据集、联邦协同数据集、端侧微调数据集的具体形态构成与核心用途。

（3）制定特殊范式构建方法。针对终端轻量化算法，深度规范激活特征提取与离群极值保留要求；针对分布式协同，规范本地日志感知、非独立同分布数据对齐及差值特征包封装；针对端侧微调，明确指令资源裁剪与陈旧数据滑动窗口动态淘汰机制。

(4) 设定端云交互安全约束。聚焦端侧数据的高敏感隐私属性，明确原生数据流转过程中的物理隔离与脱敏底线，并严格规范跨端分布式协同场景下差分隐私加噪的技术防范指标。

(5) 统一数据格式与元数据。针对终端极其严苛的算力与系统I/O瓶颈，统一数据集的二进制高压压缩与张量封装对齐标准，强制要求配置“数据出域权限”与“算力精度对齐”等扩展元数据标签。

4. 主要试验(或验证)的分析、综述报告

起草小组选取了国内多家代表性终端设备厂商与大模型企业的端侧模型训练数据流水线场景进行了模拟验证。验证结果显示，本标准提出的数据集分类体系与构建方法能够有效支撑终端模型在受限硬件上的落地，所规范的安全要求与格式封装标准与行业实际工程实践基本吻合。

5. 标准在起草过程中遇到的问题及解决办法；重大分歧意见的处理经过和依据；有无重要技术问题需要说明

在本文件的修订过程中，无重大分歧意见和技术问题。

6. 与国外标准的关系：包括：采用国际标准和国外先进标准的程度，与国外标准主要技术内容的差异

该项目没有对应的国际标准或国外先进标准。

7. 修订标准时，说明与标准前一版本的重大技术变化，并列所涉涉及的新、旧版本的有关条款(可引用标准前言的内容)；废止/代替现行有关标准的建议

不涉及。

8. 说明标准与其他标准或文件的关系(可引用标准前言的内容)，特别是与有关的现行法律、法规和强制性国家标准的关系

符合现行法律、法规要求。

9. 标准作为强制性标准或推荐性标准的建议

建议本文件作为推荐性标准。

10. 贯彻国家标准的要求和措施建议(包括组织措施、技术措施、过渡办法等内容)；标准发布后，对国内外业界可能产生的影响

建议本文件发布后，优先在人工智能终端联盟和人工智能终端工作组中推广应用。

本标准的实施将有助于规范市场竞争环境，统一人工智能企业的界定与评价标准。为投融资机构、合作伙伴及政府监管部门提供科学的选型参考与决策支撑。引导企业找准产业链定位，通过对标评估发现短板，推动人工智能与实体经济的深度融合。

建议本文件发布后，优先在中国互联网协会会员单位中推广应用。

本标准的实施将有助于统一终端模型训练数据集的建设规范，打通人工智能大模型从“云端通识”向“端侧个性化”演进的数据链路，在保障数据绝对安全合规的前提下，有效加速端侧轻量化模型的规模化部署。同时，可为投融资机构、

合作伙伴及政府监管部门提供科学的技术参考与决策支撑,引导产业界向“高优、精简、算力对齐”的高质量方向发展,助力我国抢占全球端侧人工智能标准化先机。

11. 标准是否涉及知识产权的情况说明; 如标准中含有自主知识产权, 说明产品研发程度、产业化基础及进程

本文件未涉及。

12. 其他应予说明的事项

本文件未涉及。