



中华人民共和国国家标准

GB/T 47849—2026

人工智能 边端设备模型部署工具链功能要求

Artificial intelligence—Functional requirements of edge and
terminal device model deployment toolchain

2026-07-02 发布

2027-02-01 实施

国家市场监督管理总局
国家标准化管理委员会 发布

目 次

前言 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 1

5 组成框架和流程 2

 5.1 框架组成 2

 5.2 工具链流程 3

 5.3 模型压缩模块 3

 5.4 模型转换模块 3

 5.5 推理加速模块 4

 5.6 辅助功能模块 4

6 功能要求 5

 6.1 模型压缩模块 5

 6.2 模型转换模块 6

 6.3 推理加速模块 6

 6.4 辅助功能模块 8

附录 A（资料性） 模型压缩算法适配层技术选型与应用示例 10

附录 B（资料性） 多模式训练接口层不同训练模式配置示例 11

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。

本文件由全国信息技术标准化委员会(SAC/TC 28)提出并归口。

本文件起草单位：中国电子技术标准化研究院、美的集团(上海)有限公司、北京人形机器人创新中心有限公司、上海商汤智能科技有限公司、海信集团控股股份有限公司、深圳市优必选科技股份有限公司、北京航空航天大学、浙江大学、深圳云天励飞技术股份有限公司、青岛港国际股份有限公司、美的集团股份有限公司、深圳市矽赫科技有限公司、北京嘀嘀无限科技发展有限公司、武汉大学、杭州海康威视数字技术股份有限公司、广电运通集团股份有限公司、中国科学院计算技术研究所、浪潮云信息技术股份公司、中国移动通信集团有限公司、上海人工智能创新中心、北京宝兰德软件股份有限公司、北京智芯微电子科技有限公司。

本文件主要起草人：董建、徐洋、孙传兴、唐剑、刘宁、赵东宇、蔡亚森、何伟宏、龚睿昊、吴庚、刘常昱、梁乔玲、郭晋阳、吴飞、张圣宇、马骋昊、沈芷月、蔡万伟、毕涛、吴宇震、郭乙运、脱立恒、洪鹏辉、吴国斌、江佳伟、章放、杨波、张宗帅、王功明、张站朝、蒋慧、詹年科、刘勇。

人工智能

边端设备模型部署工具链功能要求

1 范围

本文件确立了边端设备模型部署工具链的组成框架和流程,规定了功能要求。
本文件适用于边端设备模型部署工具链的设计、研发和应用。
本文件不适用于仅面向云端部署的模型工具链。

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
GB/T 46345—2025 人工智能 深度学习编译器接口

3 术语和定义

GB/T 41867—2022 和 GB/T 46345—2025 界定的以及下列术语和定义适用于本文件。

3.1

边端设备 edge and terminal device

部署在网络边侧的硬件设备(即边侧设备)和直接面向应用场景或用户的末端硬件设备(即终端侧设备)的总称。

注 1: 常见边侧设备包含边缘服务器、边缘网关、智能工业网关、边缘计算盒子等。
注 2: 常见终端侧设备包含可穿戴设备、智能相机、机器人、物联网设备等。

3.2

模型部署 model deployment

将深度学习模型适配并运行至硬件设备的过程。
注: 包括模型压缩、模型转换、计算图优化、编译优化等过程。

3.3

工具链 toolchain

对深度学习模型进行模型压缩、模型转换、图优化、编译优化等流程处理的软件工具组合。

3.4

模型量化 model quantization

模型参数由高精度数值转换为低精度数值的过程。

4 缩略语

下列缩略语适用于本文件。