



中华人民共和国国家标准化指导性技术文件

GB/Z 252—2026

人工智能 大模型一体机技术要求

Artificial intelligence—Technical requirements for large-scale model
all-in-one machine

2026-08-27 发布

国家市场监督管理总局 发布
国家标准化管理委员会

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 大模型一体机架构 2

 5.1 概述 2

 5.2 技术架构 2

6 基础硬件平台功能要求 3

 6.1 推理一体机 3

 6.2 训推一体机 4

7 基础软件平台功能要求 4

 7.1 计算加速平台 4

 7.2 资源管理平台 4

 7.3 大模型开发与使能平台 4

8 应用开发与使能平台要求 5

 8.1 输入输出控制 5

 8.2 智能化功能开发 5

 8.3 业务接入 5

 8.4 知识增强 6

 8.5 智能体开发及调用 6

 8.6 任务调度 6

9 管理系统 6

 9.1 监测 6

 9.2 运维诊断 7

 9.3 网络安全 7

附录 A（资料性） 大模型一体机应用配置参考 8

参考文献..... 9

前 言

本文件为规范类指导性技术文件。

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、上海燧原科技股份有限公司、阿里云计算有限公司、上海阡视科技有限公司、京东科技信息技术有限公司、上海文镭信息科技有限公司、浪潮通信信息系统有限公司、天翼云科技有限公司、中通服咨询设计研究院有限公司、北京软件产品质量检测检验中心有限公司、中国电力科学研究院有限公司、北京智芯微电子科技有限公司、浪潮电子信息产业股份有限公司、自然语义(青岛)科技有限公司、华为技术有限公司、上海壁仞科技股份有限公司、平头哥(上海)半导体技术有限公司、中科寒武纪科技股份有限公司、超聚变数字技术股份有限公司、浙江大华技术股份有限公司、河南昆仑技术有限公司。

本文件主要起草人：董建、王思善、张琦、徐洋、梅敬青、蒋燕、于小博、周麒、贺皓、张金柱、仲凯韬、梁秉豪、肖红梅、王小乾、王帅、李亚玲、张承明、杨雨泽、孔昊、刘晓璇、李琰、郑哲、崔文朋、李岚泊、胡德鹏、邸贺亮、沈芷月、孙燕群、孙伟伟、聂简荻、包振忠、邢冯、张艺伯、黄岩哲、张同剑、刘照华、孔维生、刘劲楠。

人工智能 大模型一体机技术要求

1 范围

本文件确立了大模型一体机的技术架构,规定了基础硬件平台、基础软件平台、应用开发与使能平台、管理系统的功能要求等。

本文件适用于大模型一体机的规划、设计、开发、测试、应用。

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

大模型一体机 large-scale model all-in-one machine

具备大模型及相关应用支持能力的一种高度集成化的智能计算设备。

注:包括基础硬件平台、基础软件平台、应用开发与使能平台和管理系统等部分。

4 缩略语

下列缩略语适用于本文件。

API:应用程序编程接口(Application Programming Interface)

BF:布瑞恩浮点数(Brain Floating-point)

CPT:继续预训练(Continued Pre-Training)

DPO:直接偏好优化(Direct Preference Optimization)

FP:浮点数(Floating-Point)

GRPO:群体相对策略优化(Group Relative Policy Optimization)

JPEG:联合图像专家组(Joint Photographic Experts Group)

LLM:大语言模型(Large Language Model)

LoRa:低秩适应(Low-Rank Adaptation)

PNG:便携式网络图形(Portable Network Graphics)

RAG:检索增强生成(Retrieval Augmented Generation)

RPC:远程过程调用(Remote Procedure Call)