



中华人民共和国国家标准化指导性技术文件

GB/Z 254—2026

人工智能 语言大模型云端推理部署工具链 技术要求

Artificial intelligence—Technical requirements for cloud-based inference
deployment toolchain of large language models

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 2

5 语言大模型云端推理部署工具链框架 3

6 功能要求 3

 6.1 硬件环境 3

 6.2 压缩工具 4

 6.3 推理引擎 5

 6.4 推理服务 6

7 性能要求 7

 7.1 硬件环境 7

 7.2 推理引擎 7

 7.3 推理服务 8

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。

本文由全国信息技术标准化技术委员会(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、上海商汤智能科技有限公司、北京航空航天大学、上海库帕思科技有限公司、北京人形机器人创新中心有限公司、海信集团控股股份有限公司、北京百度网讯科技有限公司、中国南方电网有限责任公司超高压输电公司贵阳局、天翼云科技有限公司、青岛理工大学、企知道科技有限公司、北京智源人工智能研究院、浪潮通信信息系统有限公司、科大讯飞股份有限公司、杭州海康威视数字技术股份有限公司、广东中科凯泽信息科技有限公司、西安汇智信息科技有限公司、深圳市优必选科技股份有限公司、山东省人工智能研究院、深圳市迪博技术有限公司、天津大学、深信服科技股份有限公司。

本文件主要起草人：徐洋、龚睿昊、吴庚、林达华、李斌斌、郭晋阳、孙传兴、刘祥龙、山栋明、施佳樑、周彩虹、刘宁、蔺向楠、于佃海、党青青、张治然、王帅、康鲁迪、张玉柱、刘颖、肖红梅、梁秉豪、王萍萍、章放、张凯、吴军、惠晓滨、梁乔玲、单珂、袁吉、郑喜、吕琪、高宇、徐金涛。

人工智能 语言大模型云端推理部署工具链 技术要求

1 范围

本文件确立了云端语言大模型推理部署工具链框架,规定了其功能要求和性能要求。
本文件适用于云端语言大模型推理部署工具链的设计、研发和应用。

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

GB/T 45401.1—2025 人工智能 计算设备调度与协同 第1部分:虚拟化与调度

GB/T 45401.2—2025 人工智能 计算设备调度与协同 第2部分:分布式计算框架

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

语言大模型 large language models

通过大规模的语料数据经过预训练及微调等操作,得到的具有理解、推理、生成等能力的深度学习模型。

3.2

推理 inference

从给定的前提进行论证并得出结论。

注1:在人工智能领域中,一个前提是一个事实、一个规则、一个模型、一个特征或原始数据。

注2:术语“推理”既指过程也指结果。

[来源:GB/T 41867—2022,3.2.30]

3.3

模型部署 model deployment

将经过训练的神经网络模型进行格式转换、结构优化、资源配置、模型装载和服务发布,使其能够在目标硬件或云端运行环境中执行推理任务的过程。

注:模型部署通常包括模型压缩、模型转换、计算图优化、编译优化、运行时装载、资源分配、服务启动和接口发布等环节。

3.4

工具链 toolchain

针对神经网络模型进行模型压缩、模型转换、图优化、编译优化、部署、调度、监控等过程的工具