



中华人民共和国国家标准化指导性技术文件

GB/Z 201—2026

人工智能 大模型选型和应用指南

Artificial intelligence—Large model selection and application guideline

2026-07-30 发布

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

引言 IV

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 选型与应用的原则与流程 2

 5.1 原则 2

 5.2 流程 2

6 选型准备 3

 6.1 策略规划 3

 6.2 组织保障 3

 6.3 人员保障 4

 6.4 现状评估 4

 6.5 模型初选 5

7 数据准备 5

 7.1 数据采集 5

 7.2 数据存储 5

 7.3 数据处理 6

 7.4 数据集建设 6

8 选型与优化 6

 8.1 模型选型 6

 8.2 模型训练 7

 8.3 模型调优 7

9 部署应用 8

 9.1 部署模式选择 8

 9.2 模型部署 8

 9.3 模型集成 8

 9.4 安全运维 8

10 应用评估..... 9

 10.1 模型评估 9

 10.2 效益评估 9

11 风险管理..... 9

11.1 组织能力建设 9

11.2 风险管理目标 9

11.3 风险识别及评估 9

11.4 风险应对 10

附录 A （资料性） 大模型选型参考信息 11

 A.1 算力估算 11

 A.2 模型分类 11

 A.3 模型参数 11

 A.4 模型性能 12

 A.5 模型压缩 12

参考文献 14

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、南方电网数字电网集团有限公司、阿里云计算有限公司、华为云计算技术有限公司、蚂蚁科技集团股份有限公司、国能信息技术有限公司、华北电力科学研究院有限责任公司、中国海洋石油集团有限公司、中国民航信息网络股份有限公司、北京智源人工智能研究院、深信服科技股份有限公司、烽台科技(北京)有限公司、中煤科工集团上海有限公司、中国铝业集团有限公司、浪潮软件集团有限公司、浪潮通信信息系统有限公司、中国移动通信集团有限公司、中国国际货运航空股份有限公司、北京中科金得助智能科技有限公司、中国南方电网有限责任公司超高压输电公司、普元信息技术股份有限公司、青岛海信信息科技股份有限公司。

本文件主要起草人：董建、张士宗、徐洋、高歌、李晨、李晋伟、徐小天、杨旭、李劲松、袁帅、苗宗利、彭骏涛、郑子木、彭晋、刘颖、叶润国、肖红梅、曹汐、张逸群、汪萌、王源涛、李蕾红、张丹丹、李学武、孟建、赵玉霞、李峰风、蔡洛维、于明亮、孙钦平、王杰、牛拓蒙。

引 言

随着人工智能技术的快速发展,大模型在多个领域展现出巨大潜力,但其复杂性、多样性和潜在风险也对选型与应用提出了更高要求。本文件围绕选型原则与流程、选型准备、数据准备、选型与优化、部署应用、应用评估、风险管理建立实施框架,支持各类使用方根据实际业务需求和技术环境作出合理决策。本文件指导使用方进行人工智能大模型技术应用实施的选型准备、数据准备、选型与优化、部署应用、应用评估、风险管理全流程技术指引,旨在科学地、规范地提升使用方应用人工智能的水平。

人工智能 大模型选型和应用指南

1 范围

本文件确立了人工智能大模型选型原则和应用流程,包括选型原则与流程、选型准备、数据准备、选型与优化、部署应用、应用评估和风险管理,为各类使用方提供了全生命周期的应用指导。

本文件适用于人工智能大模型技术的选型和应用。

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 23011—2022 信息化和工业化融合 数字化转型 价值效益参考模型
GB/T 35273—2020 信息安全技术 个人信息安全规范
GB/T 36073—2025 数据管理能力成熟度评估模型
GB/T 37964—2019 信息安全技术 个人信息去标识化指南
GB/T 45081—2024 人工智能 管理体系
GB/T 45288.1—2025 人工智能 大模型 第1部分:通用要求
GB/T 45288.2—2025 人工智能 大模型 第2部分:评测指标与方法
GB/T 45288.3—2025 人工智能 大模型 第3部分:服务能力成熟度评估

3 术语和定义

GB/T 23011—2022、GB/T 36073—2025、GB/T 37964—2019、GB/T 45081—2024、GB/T 45288.1—2025、GB/T 45288.2—2025、GB/T 45288.3—2025 界定的术语和定义适用于本文件。

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)

CPU: 中央处理器(Central Processing Unit)

DCMM:数据管理能力成熟度评估模型(Data Capability Maturity Model)

FLOPS:每秒浮点运算次数(Floating-point Operations Per Second)

GPTQ:生成式后训练 Transformer 量化(Generative Pre-trained Transformer Quantization)

GPU:图形处理器(Graphics Processing Unit)

IT: 信息技术(Information Technology)

LoRA: 低秩适应(Low-Rank Adaptation)

MaaS:模型即服务(Model as a Service)