ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T 45288.1—2025

人工智能　大模型
第1部分：通用要求

Artificial intelligence—Large-scale model—Part 1：General requirements

2025-02-28 发布　　2025-02-28 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

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。

本文件是GB/T 45288《人工智能　大模型》的第1部分。GB/T 45288已经发布了以下部分：

——第1部分：通用要求；

——第2部分：评测指标与方法；

——第3部分：服务能力成熟度评估。

请注意本文件的某些内容可能涉及专利。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、上海人工智能创新中心、华为云计算技术有限公司、蚂蚁科技集团股份有限公司、清华大学、中国科学院自动化研究所、北京中关村实验室、北京百度网讯科技有限公司、中国铁建股份有限公司、北京奇虎科技有限公司、中国南方电网有限责任公司、中国移动通信有限公司研究院、国家能源投资集团有限责任公司信息技术分公司、杭州联汇科技股份有限公司、北京智源人工智能研究院、阿里云计算有限公司、深圳市腾讯计算机系统有限公司、科大讯飞股份有限公司、国网湖北省电力有限公司、华为技术有限公司、天津大学、中铁第五勘察设计院集团有限公司、上海商汤智能科技有限公司、北京航空航天大学、浪潮云信息技术股份公司、上海市人工智能行业协会、哈尔滨工业大学、西南科技大学、北京大学、西安电子科技大学、北京赛西科技发展有限责任公司、中国科学院软件研究所、北京大学武汉人工智能研究院、北京大学长沙计算与数字经济研究院、青岛海尔科技有限公司、北京格灵深瞳信息技术股份有限公司、北京工业大学、中山大学、中国电信集团有限公司、北京软件产品质量检测检验中心有限公司、北京小米移动软件有限公司、北京智芯微电子科技有限公司、北京世纪好未来教育科技有限公司、杭州海康威视数字技术股份有限公司、昆仑数智科技有限责任公司、浪潮电子信息产业股份有限公司、青岛海信电子技术服务有限公司、北京中关村科金技术有限公司、天翼云科技有限公司、浪潮软件科技有限公司、上海燧原科技股份有限公司、马上消费金融股份有限公司、上海天数智芯半导体有限公司、咪咕文化科技有限公司、平头哥(上海)半导体技术有限公司、麒麟合盛网络技术股份有限公司、上海文鳐信息科技有限公司、深圳前海微众银行股份有限公司、深圳思谋信息科技有限公司、云知声智能科技股份有限公司、山东浪潮科学研究院有限公司、山东省人工智能研究院、上海计算机软件技术开发中心、上海人工智能研究院有限公司、同方知网数字出版技术股份有限公司、安徽大学、西门子(中国)有限公司、云从科技集团股份有限公司、浙江大华技术股份有限公司、中电信数智科技有限公司、南方电网人工智能科技有限公司、中国移动通信集团有限公司、中移互联网有限公司、湖南科创纺织股份有限公司、中移(苏州)软件技术有限公司、西北工业大学。

本文件主要起草人：董建、徐洋、叶珩、乔宇、鲍薇、曹晓琦、孙曦、陶建华、刘静、王嘉凯、张军、李栋、张向征、余芸、刘伟东、经迪春、赵天成、林咏华、马珊珊、马骋昊、王莞尔、李建欣、熊德意、吴涛、黄超、王士进、彭祥礼、郑中、郑子木、蒋慧、刘祥龙、汪群博、郑佳佳、高东辉、马同森、张天霖、黄现翠、孙传兴、何逸楠、赵春昊、杨沐昀、俞文心、杨超、何刚、郝文建、薛云志、刘艾杉、吴玺宏、刘尚、余甜、刘颖、陈曦、郑若琳、沈芷月、聂简荻、王先庆、王金桥、胡全一、朱贵波、韩红桂、潘恩荣、武姗姗、孔昊、于磊、郑哲、刘子韬、朱江、陈宏志、范宝余、刘微、崔明飞、高鹏军、张峰、梅敬青、曾定衡、宋煜、赵磊、高慧、张旭、仲凯韬、李斌、刘枢、梁家恩、魏子重、舒明雷、陈敏刚、孟令中、王资凯、刘长欣、范存航、生若谷、孙进、孔维生、陈利明、郑桦、赵晓玮、冯俊兰、杨玉宽、孙文庆、朱林、曾杰、钱岭、张涛。

引　言

大模型已成为人工智能发展的重要技术手段，在引领产业变革中发挥重要作用，国内外人工智能相关机构相继研究开发百余种大模型产品和评测榜单，导致用户难以有效评价人工智能产品的技术水平和服务能力。GB/T 45288《人工智能　大模型》旨在规定通用大模型的技术要求、评测指标和服务能力，拟由五个部分构成。

——第1部分：通用要求。目的在于确立大模型的参考架构，规定通用技术要求。

——第2部分：评测指标与方法。目的在于确立大模型的评测指标，描述评测方法。

——第3部分：服务能力成熟度评估。目的在于给出大模型服务能力成熟度等级及评估方法。

——第4部分：计算机视觉大模型。目的在于定义计算机视觉大模型的概念和功能，规定技术要求和测试方法。

——第5部分：多模态大模型。目的在于定义多模态大模型的概念和功能，规定技术要求和测试方法。

人工智能　大模型
第1部分:通用要求

1　范围

本文件确立了大模型的参考架构,规定了大模型的通用要求。

本文件适用于大模型开发、制备、部署和应用。

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 42018—2022　信息技术　人工智能　平台计算资源规范

GB/T 42755—2023　人工智能　面向机器学习的数据标注规程

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

3　术语和定义

下列术语和定义适用于本文件。

3.1

大模型　large-scale model

大规模深度学习模型　large-scale deep learning model

基于大量数据训练得到,具有复杂计算架构,能处理复杂任务,且具备一定泛化性的深度学习模型。

注:大模型的参数量由其功能和模态决定,一般不低于1亿。大模型训练使用的数据总量受参数量的影响,达到收敛的大模型的参数量的对数与其训练数据总量的对数成正比。

3.2

大模型服务　large-scale model service

开发、应用大模型及大模型系统的服务,以及以此为手段提供支持需求方业务活动的服务。

注:常见大模型服务内容包括大模型平台服务、大模型开发定制服务、大模型推理及运营服务。

3.3

任务　task

被调度的训练或推理对象。

注:任务用于完成一个相对独立的业务功能。一个任务属于且仅属于一个作业。

[来源:GB/T 25000.23—2019,4.12,有修改]

3.4

微调　fine-tuning

为提升机器学习模型预测准确性,使用专门领域数据在大模型上继续训练的过程。

注1:专门领域数据一般是特定场景的生产数据或合成数据。

注2:常用的微调方法包括提示词微调、全参微调、参数高效微调等。

[来源:GB/T 41867—2022,3.2.31,有修改]

3.5

提示词　prompt

提示语

使用大模型进行微调或下游任务处理时,插入到输入样本中的指令或信息对象。

4　参考架构

功能视角下的大模型参考架构见图1,包括资源池、工具、数据资源、模型、行业应用和服务平台/组件等。其中:

——资源池包括计算资源、存储资源、网络资源等硬件资源,以及资源虚拟化/调度等软件资源;

——工具包括数据工具、模型工具;

——数据资源包括通用数据、领域数据、私有数据;

——模型包括基础大模型、定制化大模型,其中,基础大模型可支持单模态或多模态数据,定制化大模型是依据用户需求,对基础大模型进行微调,定制适用于生产环境的大模型;

——行业应用为各行业场景用户提供大模型下游任务匹配服务;

——服务平台/组件贯穿各层次,提供支持大模型和相关服务的编排、部署、模型推理、运维和管理。

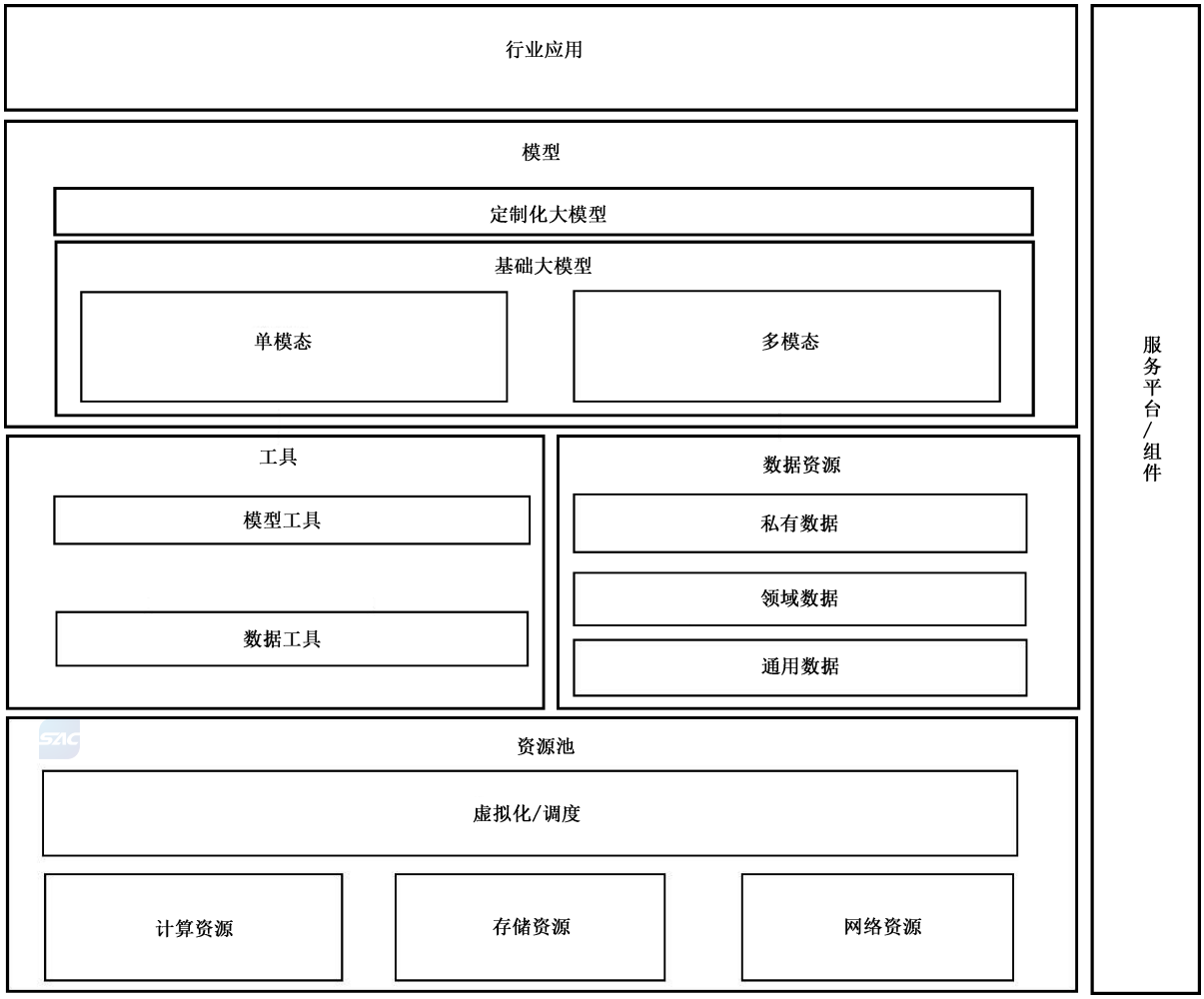

图 1　功能视角下的大模型参考架构

5　通用要求

5.1　资源池

5.1.1　计算资源

为模型训练和推理提供计算和数据处理等能力的物理计算设施用器件[如中央处理器(CPU),图形处理器(GPU),现场可编程门阵列(FPGA),神经网络处理器(NPU),张量处理器(TPU)]或虚拟计算设备。

a)　应执行至少 1 种模态(如文本、图像、语音)的模型训练或推理。

b)　应具备人工智能计算的硬件加速功能,配备分布式训练和推理计算加速库:

1)　训练服务器:应具有不少于 4 个 100 GE 网口;电源模块、风扇模块应具备热插拔和备份功能(如 2+2 冗余,$N+1$ 冗余等);

2)　推理服务器:内存总带宽应不小于 800 GB/s;应具有不少于 2 个高速串行计算机扩展总线标准(PCIe)扩展槽位;电源模块、风扇模块应具有热插拔和备份功能(如 1+1 冗余,$N+1$ 冗余等)。

c) 宜具备基于硬件加速的预处理功能(如图像、视频编解码)。
d) 应具备键值对缓存功能。

5.1.2 存储资源

适用于大模型训练和推理的存储资源,包含存储服务器等。存储资源用于提供数据存储和模型存储:

a) 应具备数据集的分布式存储与访问功能,和冗余备份机制;
b) 应具有标准文件系统接口,如可移植操作系统接口(POSIX);
c) 存储带宽宜不小于 200 GB/s,输入/输出操作次数(IOPS)宜不小于 200 万次/秒;
d) 宜具备内存计算功能;
e) 宜以存储服务器或硬磁盘为单元创建存储池,存储池宜能识别、管理固态盘、硬磁盘等不同类型存储媒体。

5.1.3 网络资源

适用于大模型训练和推理的网络资源,包含集群内交换机和路由器:

a) 应适配高速网络通信协议,如 100 GE,基于融合以太网的远程直接内存访问(RoCE)等;
b) 转发包率宜不小于 4 000 Mpps;
c) 应具备负载均衡功能;
d) 宜具备可靠性组网方案,如链路聚合、跨设备链路聚合组双活(M-LAG)等;
e) 宜具有服务器集群内 40 GE、100 GE、200 GE 或 400 GE 的全连接网络;
f) 宜具备物理交换机与逻辑交换机之间的映射功能,单台物理交换机故障不影响训练、推理任务执行。

5.1.4 虚拟化及调度

应符合 GB/T 42018—2022 中 6.2 与 GB/T 45401.1—2025 的规定。

5.2 工具

5.2.1 数据工具

5.2.1.1 数据采集工具

用于数据采集的数据工具功能包括:

a) 应提供数据采集任务设定功能,包含采集需求、数量、数据源、所采集数据的类别(如文本、语音、图片和视频等)和范围(如话题、内容等);
b) 应具备多种类型数据的采集功能,包括文本、视频、图像、音频等;
c) 应具备采集数据来源、时间和采集方式的记录功能;
d) 应具备结构化、半结构化、非结构化的数据接入功能;
e) 宜具备数据质量检测和数据清洗功能。

5.2.1.2 数据准备工具

用于数据准备的数据工具功能包括:

a) 数据标注流程应符合 GB/T 42755—2023 中第 6 章和第 7 章的要求;
b) 应具备数据清洗功能,包括文本数据的敏感词与特殊符号过滤、图像数据重建与去模糊、视频与音频数据的特定片段截取等;

c） 应具备数据重组、数据标签格式转换功能；
d） 应具备数据检索、分析等功能；
e） 应具备数据增强及扩充功能（如添加扰动产生新数据）；
f） 应具备数据质量检测功能。

5.2.1.3 数据存储工具

用于数据存储的数据工具功能包括：
a） 应具备分布式存储功能；
b） 应具备在线弹性扩展功能；
c） 应具备向量储存功能。

5.2.1.4 数据管理工具

用于数据管理的数据工具功能包括：
a） 应具备数据集要素管理功能（数据集要素包括数据集名称、版本、标注类型、标注标签、数据量、数据来源、创建时间等）；
b） 应具备数据集操作功能，包括创建、查询、修改、删除、导入、导出、发布等；
c） 应具备数据集状态信息查询功能，状态信息包括数据集名称、标注类型、数据量、导入状态、已标注状态和版本等。

5.2.2 模型工具

5.2.2.1 模型设计工具

模型设计工具：
a） 应提供预定义的模型元素（如算子、模块等）和架构；
b） 宜具备特定人工智能加速器上模型的训推模拟和性能分析。

5.2.2.2 模型训练工具

模型训练工具：
a） 应具备分布式训练能力，适配数据并行、模型并行、混合并行等分布式训练方法；
b） 训练集群在训练中出现节点故障（如宕机）时，应具有断点继续训练能力；
c） 应具备基于训练数据的整体或部分特征，构建预训练任务的功能；
d） 应具备模型历史版本和微调迭代过程的日志及查询功能，包含准确率、损失、参数更新等信息；
e） 宜提供多种并行策略，包括算子切分、算子自动并行等，宜支持自定义算子。

5.2.2.3 模型优化工具

模型优化工具符合以下规定。
a） 应具备模型压缩功能（如剪枝、量化、知识蒸馏等），宜适配组合策略压缩方案。
b） 模型微调：
1） 应具备基于文本、语音、图像或视频等数据的微调能力；
2） 应具备单模态、多模态大模型的微调能力；
3） 应具有模型评估指标体系，如准确率等；
4） 宜具备基于用户反馈的微调功能（如基于用户反馈的强化学习）。
c） 宜具备混合精度训练（自动精度混合、手动精度混合）、参数分片、梯度分片等优化训练方法（使

用的精度如半精度浮点，四分之一精度整型或单精度浮点等）。

5.2.2.4 模型验证工具

模型验证工具：

a） 应具备大模型功能（如自然语言处理、计算机视觉、多模态等）效果评估能力；
b） 应提供自动化测试功能；
c） 应具备自定义测试指标功能；
d） 宜具备模型性能实时监测和日志记录功能。

5.2.2.5 模型部署与推理工具

模型部署与推理工具：

a） 应提供服务器本地部署、云上部署功能，宜提供边缘侧（如边缘服务器）和端侧（如移动通信终端）的模型部署功能；
b） 应提供低时延推理实现机制；
c） 应具备模型推理过程的监控和日志记录功能；
d） 宜提供工具链，实现基于自然语言处理模型、视觉模型、多模态模型、科学计算模型等模型的下游任务构建；
e） 宜具备检索增强生成功能。

5.3 数据资源

5.3.1 通用数据

通用数据应符合隐私保护法规，应具备来源多样性、高质量、覆盖面广、完整性和真实性，宜覆盖各类应用场景。

5.3.2 领域数据

领域数据应符合隐私保护法规，应具备领域特征，宜覆盖领域中的使用场景。宜提供定制数据库，包含开源领域数据，具有专业性标注且在本领域具有多样性和覆盖性。

5.3.3 私有数据

私有数据应符合隐私保护法规。数据所有者应对数据使用具备控制权，包括访问权限管理和使用审计。应使用数据工程手段提升数据质量。

5.4 模型

5.4.1 基础大模型

5.4.1.1 单模态

单模态大模型或系统：

a） 应提供单模态数据的特征提取；
b） 应具备至少 1 种单模态理解能力，单模态理解能力见 GB/T 45288.2 中给出的理解能力维度；
c） 宜具备至少 1 种单模态生成能力，单模态生成能力见 GB/T 45288.2 中给出的生成能力维度。

5.4.1.2 多模态

多模态大模型或系统，符合以下要求：

a） 应具备至少1种多模态理解能力，多模态理解能力见GB/T 45288.2中给出的理解能力维度；
b） 应具备至少1种多模态生成能力，多模态生成能力见GB/T 45288.2中给出的生成能力维度。

5.4.2 定制化大模型

基于大模型，定制应用环境所需的模型及系统：
a） 应具有至少2种微调方法，如提示词微调、全参微调、参数高效微调等；
b） 应提供并运营大模型库，实现用户上传、微调和使用模型。

5.5 行业应用

对每种大模型（自然语言处理，计算机视觉，多模态等），宜至少匹配2个下游任务。

5.6 服务平台/组件

大模型服务平台/组件，宜。
a） 具备大模型插件开发能力，提供开发协议以规定插件的规则和接口，如模型接口、输入输出数据格式、插件元数据和插件运行状态码等。
b） 具备部署服务升级、回滚能力。
c） 具备计算资源弹性伸缩能力。
d） 具备大模型灰度发布（逐步将新功能或修改过的版本推向部分用户，以测试和验证其在更大范围内的表现的软件或产品更新方法）和A/B测试（通过为用户随机提供在单个变量上有所不同的产品或服务，收集并分析对应行为数据，验证不同版本之间的差异是否具有统计学意义。测试中将一个总体样本随机分成A组和B组两组，分别给予不同的处理或方案，通过对比两组结果，确定哪种处理或方案更优）功能。
e） 大模型组件具备自动检测和修复能力。
f） 具备插件运行监控和日志记录功能。

参 考 文 献

[1] GB/T 25000.23—2019 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第23部分:系统与软件产品质量测量

[2] GB/T 41867—2022 信息技术 人工智能 术语

[3] GB/T 45288.2 人工智能 大模型 第2部分:评测指标与方法