ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T 45288.2—2025

人工智能　大模型
第2部分：评测指标与方法

Artificial intelligence—Large-scale model—
Part 2：Testing and evaluation for metrics and methods

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2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。

本标准起草单位：中国电子技术标准化研究院、上海人工智能创新中心、中国科学院自动化研究所、蚂蚁科技集团股份有限公司、北京航空航天大学、清华大学、杭州联汇科技股份有限公司、中国铁建股份有限公司、北京百度网讯科技有限公司、中国南方电网有限责任公司、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、华为云计算技术有限公司、上海商汤智能科技有限公司、阿里云计算有限公司、深圳市腾讯计算机系统有限公司、北京奇虎科技有限公司、北京智源人工智能研究院、中铁第五勘察设计院集团有限公司、北京智谱华章科技有限公司、浪潮云信息技术股份公司、科大讯飞股份有限公司、中国电力科学研究院有限公司、天津大学、中国电信股份有限公司研究院、中央广播电视总台、北京百川智能科技有限公司、同方知网数字出版技术股份有限公司、北京中关村实验室、上海市人工智能行业协会、南方电网科学研究院有限责任公司、西安电子科技大学、西南科技大学、哈尔滨工业大学、中国科学院软件研究所、北京大学武汉人工智能研究院、青岛海信电子技术服务有限公司、北京格灵深瞳信息技术股份有限公司、北京工业大学、南方电网人工智能科技有限公司、中国电信集团有限公司、天翼云科技有限公司、北京软件产品质量检测检验中心有限公司、北京世纪好未来教育科技有限公司、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、中国移动通信集团有限公司、云知声智能科技股份有限公司、北京中关村科金技术有限公司、青岛海尔科技有限公司、杭州海康威视数字技术股份有限公司、京东方科技集团股份有限公司、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、浪潮软件科技有限公司、马上消费金融股份有限公司、鹏城实验室、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、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、北京安声科技有限公司、上海燧原科技股份有限公司、上海天数智芯半导体有限公司、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、西北工业大学、西门子(中国)有限公司、云从科技集团股份有限公司、上海文鳐信息科技有限公司、浙江大华技术股份有限公司、万达信息股份有限公司、上海玄武信息科技有限公司、中移互联网有限公司、四川长虹电子控股集团有限公司。

本标准主要起草人：董建、徐洋、鲍薇、陈恺、汪群博、马骋昊、孙曦、宋文林、刘祥龙、陶建华、赵天成、黄现翠、孙传兴、马珊珊、李栋、于佃海、龙云、刘伟东、经迪春、郑子木、蒋慧、彭骏涛、胡智超、张向征、杨熙、郑中、冯涛、郑佳佳、刘聪、周飞、陈晰、李建欣、熊德意、杨明川、王峰、梅剑平、陈炜鹏、张宏伟、张松阳、彭晋、刘静、刘艾杉、王嘉凯、高东辉、马同森、张天霖、高铁柱、陈曦、梁志宏、何刚、俞文心、杨沐昀、孟令中、朱贵波、王金桥、郑若琳、沈芷月、聂简荻、任海峰、石羡、吴玺宏、刘尚、刘卫卫、石聪聪、丁鹏、刘小欧、项超、薛德军、王龙跃、刘微、胡全一、孙浩源、孙林、赵必美、玄日成、赵春昊、索思亮、陈立明、蒋屹新、武姗姗、高鹏军、孔昊、薛云志、刘子韬、于磊、郑哲、邓超、梁家恩、崔明飞、鄂磊、任烨、

张志刚、陈宏志、吴韶华、王珂琛、冯月、李睿、李晋伟、龙震岳、高慧、张旭、段强、单珂、陈敏刚、宋海涛、刘益帆、王思善、余雪松、李斌、张驰、张涛、生若谷、孙进、芮子文、孔维生、童庆、杨登峰、孙文庆、朱林、杨兰。

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测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2部分：评测指标与方法

1　范围

本文件确立了人工智能大模型的评测指标，描述了人工智能大模型的评测方法。

本文件适用于模型提供者、应用服务者和应用消费者等对大模型能力进行评估与测试，也适用于指导大模型的设计、开发、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 42755—2023　人工智能　面向机器学习的数据标注规程

GB/T 45288.1　人工智能　大模型　第1部分：通用要求

3　术语和定义

GB/T 45288.1 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：应用编程接口（Application Programming Interface）

BLEU：双语评估替补（Bilingual Evaluation Understudy）

5　评测指标

5.1　理解能力评测指标

5.1.1　概述

大模型理解能力评测主要分为单模态维度和多模态维度，单模态维度主要包括文本、图像、音频3个二级维度。多模态维度主要包括图文、文音、图音、图文音4个二级维度。理解能力评测维度和典型任务见表1。