ICS 35.240
CCS L 70

NATIONAL STANDARD OF THE PEOPLE'S REPUBLIC OF CHINA

GB/T 45288.2-2025

Artificial intelligence - Large models - Part 2: Evaluation indicators and methods

人工智能 大模型 第2部分：评测指标与方法

Issued on: February 28, 2025 **Implemented on: February 28, 2025**

Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC

Table of Contents

Foreword

This document is in accordance with the provisions of GB/T 1.1-2020 "Guidelines for standardization work Part

1.Structure and drafting rules for standardization documents" Drafting. This document is Part 2 of GB/T 45288 "Artificial Intelligence Big Model". GB/T 45288 has been published in the following parts.

1.General requirements;

2.Evaluation indicators and methods;

3.Service capability maturity assessment. Please note that some of the contents of this document may involve patents. The issuing organization of the document does not assume the responsibility for identifying patents. This document was proposed and coordinated by the National Information Technology Standardization Technical Committee (SAC/TC28). This standard was drafted by: China Electronics Technology Standardization Institute, Shanghai Artificial Intelligence Innovation Center, Institute of Automation, Chinese Academy of Sciences, Ant Group Co., Ltd., Beijing University of Aeronautics and Astronautics, Tsinghua University, Hangzhou Lianhui Technology Co., Ltd., China Railway Construction Corporation Co., Ltd., Beijing Baidu Netcom Technology Co., Ltd., China Southern Power Grid Co., Ltd., China Mobile Communications Co., Ltd. Research Institute, China Energy Investment Group Information Technology Co., Ltd., Huawei Cloud

Computing Technology Co., Ltd., Shanghai SenseTime Intelligent Technology Co., Ltd. Alibaba Cloud Computing Co., Ltd., Shenzhen Tencent Computer Systems Co., Ltd., Beijing Qihoo Technology Co., Ltd., Beijing Zhiyuan Artificial Intelligence Co., Ltd. Research Institute, China Railway Fifth Survey and Design Institute Group Co., Ltd., Beijing Zhipu Huazhang Technology Co., Ltd., Inspur Cloud Information Technology Co., Ltd., iFlytek Co., Ltd., China Electric Power Research Institute Co., Ltd., Tianjin University, China Telecom Research Institute, China Central Radio and Television China Central Television, Beijing Baichuan Intelligent Technology Co., Ltd., Tongfang Knowledge Network Digital Publishing Technology Co., Ltd., Beijing Zhongguancun Laboratory, Shanghai Harbin Artificial Intelligence Industry Association, China Southern Power Grid Research Institute Co., Ltd., Xidian University, Southwest University of Science and Technology, Harbin University of Science and Technology, Institute of Software, Chinese Academy of Sciences, Wuhan Institute of Artificial Intelligence, Peking University, Qingdao Hisense Electronic Technology Service Co., Ltd., Beijing DeepGlint Information Technology Co., Ltd., Beijing University of Technology, China Southern Power Grid Artificial Intelligence Technology Co., Ltd., China Telecom Group Co., Ltd., Tianyi Cloud Technology Co., Ltd., Beijing Software Product Quality Testing and Inspection Center Co., Ltd., Beijing Century Good Future Education Technology Co., Ltd., Beijing Xiaomi Mobile Software Co., Ltd., Beijing Zhixin Microelectronics Technology Co., Ltd., China Mobile Communications Group Co., Ltd., Cloud Zhisheng Intelligent Technology Co., Ltd., Beijing Zhongguancun Kejin Technology Co., Ltd., Qingdao Haier Technology Co., Ltd., Hangzhou Hikvision Digital Technology Co., Ltd. Digital Technology Co., Ltd., BOE Technology Group Co., Ltd., Kunlun Digital Intelligence Technology Co., Ltd., Inspur Electronic Information Industry Co., Ltd., Inspur Software Technology Co., Ltd., Mashang Consumer Finance Co., Ltd., Pengcheng Laboratory, Pingtouge (Shanghai) Semiconductor Technology Co., Ltd., Qilin Hesheng Network Technology Co., Ltd., Shandong Inspur Science Research Institute Co., Ltd., Shandong Artificial Intelligence Research Institute Institute, Shanghai Computer Software Technology Development Center, Shanghai Artificial Intelligence Research Institute Co., Ltd., Beijing Ansheng Technology Co., Ltd., Shanghai Suiyuan Technology Technology Co., Ltd., Shanghai Tianshu Zhixin Semiconductor Co., Ltd., Shenzhen Qianhai Weizhong Bank Co., Ltd., Shenzhen Simo Information Technology Co., Ltd., Northwestern Polytechnical University, Siemens (China) Co., Ltd., CloudWalk Technology Group Co., Ltd., Shanghai Wenyue Information Technology Co., Ltd. Company, Zhejiang Dahua Technology Co., Ltd., Wanda Information Co., Ltd., Shanghai Xuanwu Information Technology Co., Ltd., China Mobile Internet Co., Ltd., Sichuan Changhong Electronics Holding Group Co., Ltd. The main drafters of this standard. Huang Xiancui, Sun Chuanxing, Ma Shanshan, Li Dong, Yu Dianhai, Long Yun, Liu Weidong, Jing Dichun, Zheng Zimu, Jiang Hui, Peng Juntao, Hu Zhichao, Zhang Xiangzheng, Yang Xi, Zheng Zhong, Feng Tao, Zheng Jiajia, Liu Cong, Zhou Fei, Chen Xi, Li Jianxin, Xiong Deyi, Yang Mingchuan, Wang Feng, Mei Jianping, Chen Weipeng, Zhang Hongwei, Zhang Songyang, Peng Jin, Liu Jing, Liu Aishan, Wang Jiakai, Gao Donghui, Ma Tongsen, Zhang Tianlin, Gao Tiezhu, Chen Xi, Liang Zhihong, He

Gang, Yu Wenxin, Yang Muyun, Meng Lingzhong, Zhu Guibo, Wang Jinqiao, Zheng Ruolin, Shen Zhiyue, Nie Jiandi, Ren Haifeng, Shi Xian, Wu Xihong, Liu Shang, Liu Weiwei, Shi Congcong, Ding Peng, Liu Xiaoou, Xiang Chao, Xue Dejun, Wang Longyue, Liu Wei, Hu Quanyi, Sun Haoyuan, Sun Lin, Zhao Bimei, Xuan Richeng, Zhao Chunhao, Suo Siliang, Chen Liming, Jiang Yixin, Wu Shanshan, Gao Pengjun, Kong Hao, Xue Yunzhi, Liu Zitao, Yu Lei, Zheng Zhe, Deng Chao, Liang Jiaen, Cui Mingfei, E Lei, Ren Ye, Zhang Zhigang, Chen Hongzhi, Wu Shaohua, Wang Kechen, Feng Yue, Li Rui, Li Jinwei, Long Zhenyue, Gao Hui, Zhang Xu, Duan Qiang, Shan Ke, Chen Mingang, Song Haitao, Liu Yifan, Wang Sishan, Yu Xuesong, Li Bin, Zhang Chi, Zhang Tao, Sheng Ruogu, Sun Jin, Rui Ziwen, Kong Weisheng, Tong Qing, Yang Dengfeng, Sun Wenqing, Zhu Lin, Yang Lan.

Big models have become an important technical means for the development of artificial intelligence and play an important role in leading industrial transformation. Relevant institutions have successively researched and developed more than 100 large-scale model products and evaluation lists, making it difficult for users to effectively evaluate the technical level of artificial intelligence products. GB/T 45288 "Artificial Intelligence Big Model" aims to specify the technical requirements, evaluation indicators and service capabilities of general big models. Force is proposed to consist of five parts.

1.General requirements. The purpose is to establish a reference architecture for large models and specify general technical requirements.

2.Evaluation indicators and methods. The purpose is to establish the evaluation indicators of large models and describe the evaluation methods.

3.Service capability maturity assessment. The purpose is to provide the large model service capability maturity level and assessment method.

4.Computer vision big model. The purpose is to define the concept and function of the computer vision big model and specify the technical requirements and testing methods.

5.Multimodal large models. The purpose is to define the concept and function of multimodal large models, specify technical requirements and tests method. Artificial Intelligence Big Model Part

1 Scope

Part 2 of China's national standard series on large models, covering evaluation indicators and methods. Evaluating a large model is genuinely hard in a way that evaluating conventional software is not. The public benchmarks are contaminated - their answers are in the training data - and they measure a narrow set of tasks that correlate loosely with what a deployed model does. Human evaluation is expensive and inconsistent. And the properties a buyer actually cares about are the ones hardest to score: whether the model refuses what it should, whether its confident answers are correct, whether it behaves the same next month. This part establishes evaluation indicators for AI large models and

describes the methods for applying them, for use by model providers, application servers and application consumers to evaluate and test model capabilities, and to guide the design, development and application of large models.

This document establishes the evaluation indicators for large AI models and describes the evaluation methods for large AI models. This document is applicable to model providers, application servers, and application consumers to evaluate and test the capabilities of large models. Lead the design, development and application of large models.

2 Normative references

The contents of the following documents constitute essential clauses of this document through normative references in this document. For referenced documents without a date, only the version corresponding to that date applies to this document; for referenced documents without a date, the latest version (including all amendments) applies to This document.

GB/T 42755-2023 Artificial Intelligence Data Labeling Procedure for Machine Learning

GB/T 45288.1 Artificial Intelligence Large Model Part

3 Terms and definitions

The terms and definitions defined in GB/T 45288.1 apply to this document.

4 Abbreviations

The following abbreviations apply to this document. API. Application Programming Interface BLEU. Bilingual Evaluation Understudy

5.1 Comprehension Ability Evaluation Indicators

5.1.1 Overview The evaluation of large model understanding ability is mainly divided into single-modal dimension and multi-modal dimension. The single-modal dimension mainly includes text, image, and audio. The multimodal dimension mainly includes four secondary dimensions. picture and text, text and sound, picture and sound, and picture and text and sound. The types of tasks are shown in Table 1.