ICS 35.020
CCS L 70

# 中华人民共和国国家标准

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

**Artificial intelligence—Scheduling and cooperation for computing devices—Part 1：Virtualization and scheduling**

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 45401《人工智能　计算设备调度与协同》的第1部分。GB/T 45401已经发布了以下部分：

——第1部分：虚拟化与调度；

——第2部分：分布式计算框架。

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、华为技术有限公司、北京航空航天大学、中国科学院软件研究所、华为云计算技术有限公司、阿里云计算有限公司、北京百度网讯科技有限公司、浪潮电子信息产业股份有限公司、上海商汤智能科技有限公司、北京大学武汉人工智能研究院、上海市人工智能行业协会、中国移动通信集团有限公司、中国科学院计算技术研究所、科大讯飞股份有限公司、北京大学、深圳云天励飞技术股份有限公司、上海天数智芯半导体有限公司、北京壁仞科技开发有限公司、杭州海康威视数字技术股份有限公司、南方电网人工智能科技有限公司、龙芯中科技术股份有限公司、苏州登临科技有限公司、浙江大华技术股份有限公司、蚂蚁科技集团股份有限公司、国科础石(重庆)软件有限公司、中国南方电网有限责任公司、广电运通集团股份有限公司、上海计算机软件技术开发中心、上海文鳐信息科技有限公司、京东方科技集团股份有限公司、天津(滨海)人工智能创新中心。

本文件主要起草人：范科峰、杨雨泽、李斌斌、于超、徐洋、王莞尔、曹晓琦、董建、鲍薇、栾钟治、朱毅鑫、董乾、孟令中、郑子木、吴涛、田晓利、张亚强、马珊珊、马骋昊、赵春昊、吴庚、曹汐、王煜炜、吴婷、杨超、王志芳、余雪松、丁瑞全、叶挺群、卢志良、马莞悦、代君、孔维生、郭智慧、罗勇军、梁志宏、巫伟南、杨波、陈敏刚、牛科科、仲凯韬、姜幸群、史殿习。

# 引　言

随着人工智能计算形态的不断发展，承载人工智能应用的计算设备的部署和使用呈现分布式、全场景的趋势。同一人工智能计算任务往往需要多种形态的计算设备协作完成，为不同地域、类型的用户提供服务。需要对不同形态的计算设备资源合理利用及分配，明确必要的技术架构、能力要求以及接口等，为产品提供参考框架以及评价体系，缓解不同形态人工智能计算设备横向协同割裂的现状。

GB/T 45401《人工智能　计算设备调度与协同》拟由两个部分组成。

——第1部分：虚拟化与调度。旨在确立人工智能计算设备虚拟化与调度系统的架构，规定技术要求及对应的测试方法。

——第2部分：分布式计算框架。旨在确立人工智能计算设备分布式计算的架构，规定功能和性能技术要求，定义分布式计算协同接口。

# 人工智能　计算设备调度与协同
# 第1部分:虚拟化与调度

## 1　范围

本文件给出了人工智能计算设备虚拟化与调度的架构,规定了技术要求,描述了测试方法。

本文件适用于人工智能计算设备虚拟化与调度的系统设计、研发和测试。

## 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　信息技术　人工智能　术语

GB/T 45087—2024　人工智能　服务器系统性能测试方法

## 3　术语和定义

GB/T 41867 界定的以及下列术语和定义适用于本文件。

3.1

**人工智能计算单元　artificial intelligence computing unit**

执行人工智能计算任务所必要的部件的最小集合。

注:人工智能计算单元一般封装在人工智能加速器或加速卡中。

3.2

**人工智能加速[处理]器　artificial intelligence accelerating [processor]unit**

人工智能加速芯片　artificial intelligence accelerating chip

具备适配人工智能算法的运算微架构,能完成人工智能应用运算处理的集成电路元件。

3.3

**人工智能加速卡　artificial intelligence accelerating card**

专为人工智能计算设计、符合人工智能服务器硬件接口的扩展加速设备。

注:人工智能加速卡按适用场景分为人工智能训练加速卡、人工智能推理加速卡等。

3.4

**人工智能计算实例　artificial intelligence computing instance**

执行人工智能计算任务的虚拟化对象。

3.5

**虚拟化　virtualization**

用于表示与潜在的物理资源解耦的资源表示形式。

[来源:ISO/IEC 17826:2022,3.55]

3.6

**[异构]资源池　[heterogeneous] resource pool**

由不同架构的人工智能计算资源集合形成的抽象实体。