ICS 35.160
CCS L 61

中华人民共和国国家标准

GB/T 45087—2024

人工智能　服务器系统性能测试方法

Artificial intelligence—Performance testing methods for server systems

2024-11-28 发布　　2024-11-28 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、浪潮电子信息产业股份有限公司、英特尔(中国)有限公司、平头哥(上海)半导体技术有限公司、科大讯飞股份有限公司、新华三信息技术有限公司、超威半导体产品(中国)有限公司、北京航空航天大学、中科寒武纪科技股份有限公司、南京南瑞瑞腾科技有限责任公司、中国南方电网有限责任公司超高压输电公司、石化盈科信息技术有限责任公司、中国电信股份有限公司广东研究院、上海燧原科技股份有限公司、中国科学院软件研究所、北京壁仞科技开发有限公司、上海阡视科技有限公司、上海超级计算中心、上海文鳐信息科技有限公司、美的集团(上海)有限公司、国科础石(重庆)软件有限公司、上海人工智能研究院有限公司、四川华鲲振宇智能科技有限责任公司、深圳鲲云信息科技有限公司、中国铁建股份有限公司、中铁第五勘察设计院集团有限公司、西南科技大学。

本文件主要起草人：董建、徐洋、张琦、王莞尔、曹晓琦、黄剑彬、梁朝明、鲍薇、吴韶华、王海宁、林晓东、马珊珊、高慧、张艺伯、陶玉梅、杨雨泽、郑会平、刘如冰、李岚泊、纪拓、栾钟治、程归鹏、黄现翠、牧军、石超、叶珩、王宁、刘东庆、李先绪、师春雨、梅敬青、孟令中、丁瑞全、程秋林、吴庚、郁华真、张丹丹、仲凯韬、任沛、傅欣杰、胡艳玲、宋海涛、白士玉、刘东、栾丽红、李栋、郑中、俞文心。

引　言

人工智能服务器系统包含人工智能服务器、集群和高性能计算设施等形态，是各类深度学习模型（包含大规模预训练模型）训练和推理的核心载体，是各行业应用人工智能技术提高生产效率的核心工具。人工智能服务器系统专为处理人工智能计算任务设计，在架构、运算方式和用途用法上，与通用服务器系统有较大差别，其测试过程、负载和指标等，皆有独特性。本文件提出人工智能服务器系统性能基准测试的方法，并对基准测试工具的功能和公平性提出要求。

本文件的发布机构提请注意，声明符合本文件时，可能涉及 7.4.2、7.7.1 与人工智能服务器系统性能测试方法相关专利的使用。

本文件的发布机构对于该专利的真实性、有效性和范围无任何立场。

该专利持有人已向本文件的发布机构承诺，他愿意同任何申请人在合理且无歧视的条款和条件下，就专利授权许可进行谈判。该专利持有人的声明已在本文件的发布机构备案，相关信息可以通过以下联系方式获得。

专利持有人：中国电子技术标准化研究院

地址：北京市东城区安定门东大街 1 号

请注意除上述专利外，本文件的某些内容仍可能涉及专利。本文件的发布机构不承担识别专利的责任。

人工智能　服务器系统性能测试方法

1　范围

本文件界定了服务器系统性能测试模式，描述了人工智能服务器系统训练性能和推理性能测试方法。

本文件适用于人工智能服务器系统的性能测试与评价。

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—2022　信息技术　人工智能　术语

3　术语和定义

GB/T 41867—2022 界定的以及下列术语和定义适用于本文件。

3.1

被测系统　system under test

处理测试者给出的测试作业，并返回符合要求结果的系统。

注：被测系统由人工智能服务器系统硬件、算子实现库、机器学习框架软件、模型编译组件和其他必要软硬件组成。

3.2

被测者　tested party

提供被测系统和测试信息，并协助测试实施的机构或个人。

3.3

参考模型　reference model

用于定义系统测试要求的标准化的模型。

[来源：ISO/IEC 14776-414：2009，3.1.87，有修改]

3.4

计时　timing

获取并返回被测系统当前时间戳。

注：假设**被测系统**（3.1）各节点时间一致。

3.5

人工智能服务器　artificial intelligence server

信息系统中能为人工智能应用提供高效能计算处理能力的服务器。

注 1：人工智能服务器含有专为人工智能计算设计的计算模块，为人工智能应用提供专用加速计算能力。

注 2：以通用服务器为基础，配备人工智能加速卡后，为人工智能应用提供专用计算加速能力的服务器，称“人工智能兼容服务器”。

注 3：专为人工智能加速计算设计，提供人工智能专用计算能力的服务器，称“人工智能一体机服务器”。

[来源：GB/T 41867—2022，3.1.3，有修改]