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]

3.6

人工智能服务器集群　artificial intelligence server cluster

由通过高速互联网络(或协议)连接的若干人工智能服务器组成,遵循统一控制和调度,对外提供人工智能计算的系统。

注:简称"集群"。

3.7

人工智能服务器系统　artificial intelligence server system

由人工智能服务器(含集群)和其他必要的计算、存储设备、操作系统等组成,承担人工智能运算任务的计算系统。

3.8

测试数据　test data

测试集　test dataset

用于测试最终机器学习模型功能的数据。

[来源:ISO/IEC 22989:2022,3.2.14]

3.9

测试者　tester

组织、实施测试的机构或个人。

3.10

测试系统　test system

测试实验室执行测试方法所采用的硬件、软件和数据。

注:测试系统不是被测系统中的机器学习框架软件或加速库。

[来源:GB/T 16656.34—2002,3.5.9,有修改]

3.11

作业　job

含有测试样本的数据包。

注 1:1 个作业通常含有 1 个或多个测试样本。

注 2:1 次测试任务至少含有 1 个作业。

3.12

性能　performance

人工智能服务器系统运行计算任务时可被测量的特性。

注:性能通常基于 1 个或多个参数(如时间、功耗、实际吞吐率、资源利用率、弹性、承压力和视频分析最大路数等)的测试或计算获得,以表示在某机器中运行的某技术过程的行为、特性和效率。

3.13

训练数据　training data

训练集　training dataset

用于训练机器学习模型的输入样本子集。

[来源:ISO/IEC 22989:2022,3.3.16]

3.14

验证数据　validation data

验证集　validation dataset

用于验证机器学习模型训练效果的输入样本子集。

[来源:ISO/IEC 22989:2022,3.2.15,有修改]

3.15

布瑞恩浮点数　brain floating-point

包含1位符号、8位指数和7位尾数的浮点数表示方式。

注：与FP32(8个指数位和23个小数位)能表达的范围大小相同，比FP16(5个指数位、10个小数位)能表示的范围更大，不易发生数值上溢或下溢，更适合大模型的训练和推理。

3.16

节点　node

人工智能服务器系统中，能独立完成训练或推理计算，且其性能参数能被独立计量的组件。

3.17

试验次数　number of rounds for a test

按试验的要求，完成相同试验过程或重复处理相同数据的次数。

注：训练测试中，试验次数是从模型的初始化状态训练模型达至准确率门限或训练执行的训期数量(含使用验证集获得准确率)。

3.18

训期　epoch

引入到神经网络中的训练模式序列。

注1：训练过程完整遍历1次训练集即1个训期。

注2：对分布式训练，所有训练节点的本地训练过程遍历处理1遍本地训练集，为1个训期。

[来源：GB/T 5271.34—2006，34.03.19，有修改]

3.19

性能指标　performance indicator

用于评估人工智能服务器系统实现效果的度量。

注：本文件中，在不引起误解的语境中，将人工智能服务器系统性能指标简称为“指标”。

[来源：GB/T 22454—2008，3.1.62，有修改]

3.20

语素　token

用于表示文本数据的最小单位。

注：如单词、词组或字符。

[来源：ISO 23952:2020，3.3.11，有修改]

4　缩略语

下列缩略语适用于本文件。

AUC：曲线下面积(Area Under Curve)

AUTOML：自动机器学习(Automated Machine Learning)

BF16：布瑞恩浮点数(Brain Floating-point)

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

CPU：中央处理器(Central Processing Unit)

FP16：半精度浮点数(Half-precision Floating-point Format)

FP32：单精度浮点数(Single-precision Floating-point Format)

FP64：双精度浮点数(Double-precision Floating-point Format)

GPU：图形处理器(Graphics Processing Unit)

ID：序号(Identity Document)

INT4：4位整型数(4-bit Integer)

INT8:8 位整型数(8-bit Integer)
mAP:平均准确率均值(Mean Average Precision)
mIOU:平均交并比(Mean Intersection Over Union)
NFS:网络文件系统(Network File System)
NPU:人工智能加速器(Neural Processing Unit)
OCR:光学字符识别(Optical Character Recognition)
PCIe:外围组件互连快速总线(Peripheral Component Interconnect Express)
SUT:被测系统(System Under Test)
TF32:张量单精度浮点数(Tensor Floating-point)
UINT4:4 位无符号整型数(4-bit Unsigned Integer)
UINT8:8 位无符号整型数(8-bit Unsigned Integer)
WER:错词率(Word Error Rate)

5 测试模式

5.1 封闭模式

5.1.1 封闭模式训练测试

给定训练集、目标模型结构和精度,在被测人工智能服务器系统上,运行建模和优化算法得到目标模型,结果应符合精度和给定测试集上的准确率要求。

5.1.2 封闭模式推理测试

给定模型(参考实现)、精度和测试集,在被测人工智能服务器系统上,运行模型定义的推理过程,输出推理结果,结果应符合精度和给定测试集上的准确率要求。

5.2 开放模式

5.2.1 开放模式训练测试

给定训练集和精度,被测者选择模型、数据预处理方式及算法优化策略,在被测人工智能服务器系统实施训练,结果应符合精度和给定测试集上的准确率要求。

5.2.2 开放模式推理测试

给定测试集,被测者提供已训练好的模型,在被测人工智能服务器系统运算并输出推理结果,结果应符合精度和给定测试集上的准确率要求。

6 训练性能测试

6.1 测试过程

训练测试过程包含以下步骤。

a) 信息准备,被测者应向测试者提供测试信息,包含但不限于以下内容:
   1) 组织名称或个人姓名;
   2) 测试 ID(用于标识测试);
   3) 测试模式(0 表示封闭模式、1 表示开放模式);

4） 通用场景或专用场景(0 表示通用、1 表示专用)；
5） 任务类型(0 表示推理、1 表示训练)；
6） 数据集类型(0 表示固定数据集、1 表示随机数据集)；
7） 随机数据集用于(重)训练的训期数(零次样本 0-zero shot，单样本 1-one shot，少样本 N-few shot)；
8） 模型序号(对封闭模式有效，开放模式时提供模型名和版本号)；
9） 提交时间(格式为[yyyy:MM:dd HH:mm:ss])；
10） 测试对象类型(0 表示单机、1 表示集群/计算中心)；
11） 节点数(当“测试对象类型”不为“0”时有效)；
12） 每台服务器信息[型号；标称计算能力；实施人工智能加速卡或加速芯片的数量；CPU 型号、核数、主频；CPU 路数；加速卡信息(推理卡、训练卡或训推一体卡，是否需外接电源和接口类型)；存储信息(存储设备接口类型、协议、数量和总容量)；内存信息(型号、条数、单条容量和总容量)；总线信息(PCIe 协议版本和接口形态如 x4、x8 或 x16)]；
13） 节点间通信协议和带宽；
14） 节点间组织关系(0 表示单节点、1 表示主从、2 表示环形、3 表示树状、4 表示其他)；
15） 操作系统信息(名称、内核版本号)；
16） 机器学习框架信息(名称、版本号)；
17） 是否应用虚拟化技术(0 表示不使用、1 表示使用)；
18） 虚拟化组件信息(名称、版本号)；
19） 批大小可变标识(0 表示不可变、1 表示可变)；
20） 批大小的值(正整数，当不为 0 时有效)；
21） 优化器声明(算法名)；
22） 是否混合精度训练(0 表示不使用、1 表示使用，附加精度列表)；
23） 是否使用 AUTOML 完成测试(0 表示不使用、1 表示使用，附加 AUTOML 算法名称)；
24） 是否使用并行训练完成测试(0 表示不使用、1 表示模型并行、2 表示数据并行、3 表示混合并行、4 表示其他并行算法并附加算法名称)；
25） 并行训练时，是否采用异步参数更新 [0 表示不使用(即同步参数更新)、1 表示使用]。

b） 数据准备，被测者于测试前，取得训练集和验证集；如需要，被测者可对数据进行必要的格式转化或封装。
c） 测试运行，被测者按测试内容，编写并运行必要的训练代码(包含数据预处理、数据读入、训练、结果模型格式转化与持久化)，得到结果模型；训练期间记录过程数据，计算指标值，记录日志，生成结果信息。
d） 结果报送，被测者发送训练结果给测试者。
e） 结果审核，训练测试结果应符合 6.3 的要求。

6.2 训练测试要求

训练测试符合以下要求。

a） 训练测试不应实施以下操作：
1） 在测试过程中进行硬件或软件改配；
2） 使用本文件规定之外的训练集进行模型训练，实施模型预训练和迁移学习策略(大模型负载除外，按表 1 和表 12 的规定执行预训练或迁移学习)；
3） 训练测试过程中，对已实现的指标测量函数或测试流程控制函数实施改动、继承或重载(要求被测者实现的方法除外)；

4) 在数据准备过程中：替换数据集；减少数据集中的样本（封闭模式有效，除不足 1 批的残余数据之外）；除 6.2d）规定的操作生成的样本外，增加数据集中的样本（封闭模式有效）；分析数据规律或预先提取、编码和保存样本特征（封闭模式有效）；对数据做排序、索引或拆分操作（封闭模式有效）；

5) 在训练过程中改变指定的优化方法（封闭模式有效）。

b) 应编制并运行的训练测试代码：实现必要接口（如日志报送接口，准确率计算接口等），以采集用于计算 6.6 中指标项的参数值。

c) 测试应使用工具进行日志记录、数据采集、指标项计算，人工智能服务器系统性能测试工具示例见附录 A。

d) 数据准备时：训练数据规格不同或不符合模型需要时，可实施规格调整操作；在不改变输入图像（对视觉类场景）像素值的情况下，可实施插值操作，包含但不限于线性插值、双线性插值和区域插值等；训练集、验证集和测试集的划分比例，默认为 75％、10％和 15％。

e) 训练过程中：可使用可变学习率，学习率改变方法由训练算法确定；权重和偏置应以常量或随机值初始化；试验次数应符合场景要求。

f) 实施分布式训练时：可使用并行训练，方式可包含但不限于模型并行、数据并行和混合并行；可使用分布式文件系统（如 NFS）或存储服务器存放或使用训练数据。

g) 使用 AUTOML 训练的还应符合附录 B 的 B.1 的要求。

6.3 训练测试结果

训练测试结果符合以下要求。

a) 训练结果模型与参考模型一致，训练结果模型精度应符合表 1 和表 3 的要求，loss（损失函数）曲线应按预期保持收敛趋势。

b) 封闭模式下，训练模型脚本与参考模型脚本（见表 1 和表 3）应符合一致的网络结构，训练模型脚本不应导致以下情况的发生：

1) 多余或缺失的层；

2) 多余或缺失的神经元；

3) 改变的激励函数（对应层之间）；

4) 多余或缺失的跨层连接；

5) 改变的池化方法（对应层之间）。

c) 结果应包含以下信息：

1) 6.1 要求的配置信息；

2) 测试场景要求的指标（见表 1 和表 3）；

3) 训练测试代码公开规则应符合附录 C 的要求；

4) 训练日志（非 AUTOML 训练），日志应按每个训期输出，每个训期对应的格式为：[yyyy:MM:dd HH:mm:ss]-[trial_number]-[epoch_number]-[accuracy]，其中：

——[yyyy:MM:dd HH:mm:ss]：日志输出时的时间戳；

——[trial_number]：（训练）试验次数，取值为正整数，不满 1 次完整训练的记录为 1；

——[epoch_number]：训期数，取值为正整数；

——[accuracy]：当前测试集上的准确率，当机器学习框架无法实现训期结束时输出准确率时，则记录为“--”，但训练退出前应输出准确率。

注：机器学习框架软件无法实现时，对应项目记录为“--”。

5) 结果模型文件（含权重和结构信息；AUTOML 训练为最终结果模型文件）。

d) AUTOML 训练的日志应符合 B.2 的要求。

6.4 测试场景

6.4.1 通用测试场景

6.4.1.1 通用封闭模式测试应从表 1 所列测试场景中选择，测试场景类型说明见附录 D。

表 1 通用训练性能测试场景（封闭模式）

序号	类型	场景说明	
1	图像识别	模型	Resnet50_v1.5
		数据集[a]	Imagenet2012
		门限[b]	Top1-准确率＞74％
		优化方法	SGD＋Momentum
		试验次数	5 次
		结果模型精度	FP16 /FP32/BF16
		损失函数	Softmax＋Cross-Entropy Loss
2	图像识别	模型	Inception v3
		数据集[a]	Imagenet2012
		门限[b]	Top1-准确率＞ 78.06％
		优化方法	Adam
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
		损失函数	Cross-Entropy Loss
3	语义分割	模型	Deeplab_v3[c]
		数据集[a]	Cityscapes
		门限[b]	mIOU＞77.98％
		优化方法	Adagrad
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
		损失函数	Pixel-wise softmax＋Cross-Entropy Loss
4	语音识别	模型	Wav2vec2_0
		数据集[a]	Aishell-1
		门限[b]	WER＜5.5％
		优化方法	Adam
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
		损失函数	CTC Loss
5	推荐	模型	Wide&deep
		数据集[a]	Criteo(Kaggle Display Advertising Challenge Dataset)
		门限[b]	AUC＞72％

表 1　通用训练性能测试场景(封闭模式)(续)

序号	类型	场景说明	
5	推荐	优化方法	Wide:FTRL; Deep:Adagrad
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
		损失函数	Logistic loss Label smoothed cross entropy loss
6	推荐	模型	DLRM
		数据集[a]	Criteo(1TB Click Logs)
		门限[b]	AUC>0.8025
		优化方法	SGD
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
		损失函数	BCELoss[c]
7	物体检测	模型	Yolo5s-6-0
		数据集[a]	Coco2017
		门限[b]	mAP@0.5:64.2%
		优化方法	SGD
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
		损失函数	CIoU[c]+BCEWithLogitsLoss
8	自然语言处理	模型	Bert-large[c,d]
		数据集[a]	En-wiki
		门限[b]	Mask_lm accuracy>0.7
		优化方法	Lamb
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
		损失函数	Softmax+Negative maximum likehood loss
9	自然语言处理	模型	GLM v2 6B
		数据集[a]	微调任务:ADGEN 预训练任务:En-wiki
		测试终止条件	2 000 个 step
		优化方法	Adam
		试验次数	3 次
		结果模型精度	FP16/FP32/BF16
		损失函数	Cross entropy

表 1　通用训练性能测试场景(封闭模式)(续)

序号	类型	场景说明	
10	自然语言处理	模型	LLaMa2-13B
		数据集[a]	微调任务:Moss 预训练任务:En-wiki
		测试终止条件	2 000 个 step
		优化方法	Adam
		试验次数	3 次
		结果模型精度	FP16/FP32/BF16
		损失函数	Cross entropy
11	自然语言处理	模型	LLaMa2-70B
		数据集[a]	微调任务:Moss 预训练任务:En-wiki
		测试终止条件	200 个 step
		优化方法	Adam
		试验次数	3 次
		结果模型精度	FP16/FP32/BF16
		损失函数	Cross entropy
12	多模态	模型	Stable Diffusion v2.1
		数据集[a]	微调任务:Pokemon 预训练任务:LAION-5B
		测试终止条件	2 000 个 step
		优化方法	AdamW
		试验次数	3 次
		结果模型精度	FP16/FP32/BF16
		损失函数	MSE

注:“/”表示“或”。

[a] 训练数据的格式,没有统一限定,被测者可根据本地系统组成实施必要的格式转换,格式转换过程不应改变数据的值(如图像像素值),数据格式转换过程不计时。

[b] 门限为参考值,测试实施时可作调整,在多系统对比测试时应使用相同门限值。

[c] Bert-large 测试项中,Sequence-length=512。

[d] 测试终止条件为参考值,测试实施时可作调整,在多系统对比测试时应使用相同测试终止条件。

6.4.1.2　通用开放模式测试应从表 2 所列测试场景中选择,测试场景类型说明见附录 D。

表 2　通用训练性能测试场景(开放模式)

序号	类型	场景说明	
1	图像识别	数据集	Imagenet2012
		门限[a]	Top1-准确率＞75%
		结果模型精度	FP16/FP32
2	物体检测	数据集	Coco2017
		门限[a]	mAP＞35%
		结果模型精度	FP16/FP32/BF16
3	语义分割	数据集	Coco2017
		门限[a]	mIOU＞85%
		结果模型精度	FP16/FP32/BF16
4	推荐	数据集	Criteo(Kaggle Display Advertising Challenge Dataset)
		门限[a]	AUC＞72%
		结果模型精度	FP16/FP32/BF16
5	推荐	数据集	Criteo(1TB Click Logs)
		门限[a]	AUC＞0.8025
		结果模型精度	FP16/FP32/BF16
6	自然语言处理	数据集	WMT18 英->德、英->中
		门限[a]	BLEU＞24%
		结果模型精度	FP16/FP32/BF16
7	自然语言处理	数据集	Cn-wiki
		门限[a]	Mask_lm accuracy＞ 0.7
		结果模型精度	FP16/FP32/BF16
8	语音识别	数据集	Aishell-1
		门限[a]	WER＜5.5%
		结果模型精度	FP16/FP32/BF16
注：“/”表示“或”。			
[a] 准确率门限，按封闭模式场景定义，在测试时可由测试者调整或确定指标和取值。			

6.4.2　专用测试场景

6.4.2.1　专用封闭模式测试应从表 3 所列测试场景中选择，测试场景类型说明见附录 D。

表 3　专用训练性能测试场景(封闭模式)

序号	类型	场景说明	
1	OCR(无预分割)	模型	DBNET
		数据集[a]	Icdar2015
		优化方法	BalanceCrossEntropyLoss+MaskL1Loss+DiceLoss
		门限	Precision>0.896
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
2	人脸识别	模型	FaceNet
		数据集[a]	LFW
		优化方法	Lars/SGD+Momentum
		试验次数	5 次
		结果模型精度	FP16/FP32/BF16
3	语音识别	模型	Conformer(ESPnet2)
		数据集[a]	Aishell-1
		优化方法	SGD+Momentum
		试验次数	5 次
		门限	Precision overall:95.02%
		优化方法	CTCloss+attentionloss(LabelSmoothingLoss)
		结果模型精度	FP16/FP32/BF16
注:"/"表示"或"。			
[a] 训练数据的格式,没有严格的限定,被测者可根据本地机器学习框架进行格式转换,格式转换过程不应改变数据的值(如图像像素值),数据格式转换过程不计时。			

6.4.2.2　专用开放模式测试应从表 4 所列测试场景中选择,测试场景类型说明见附录 D。

表 4　专用训练性能测试场景(开放模式)

序号	类型	场景说明	
1	无预分割(OCR)	数据集	Icdar2015
		结果模型精度	FP16/FP32/BF16
2	人脸识别	数据集	LFW
		结果模型精度	FP16/FP32/BF16
注:"/"表示"或"。			

6.5　测试场景配置要求

针对测试目标的不同,训练性能测试分为通用测试和专用测试:

a)　通用测试是指针对共性问题,使用公共可获得的模型和数据集,完成训练测试;

b） 专用测试是针对行业领域问题，使用专用模型和数据集，完成训练测试。

训练测试场景可变要素配置要求见表5。

表5 训练测试场景可变要素配置要求

可变要素	通用封闭	通用开放	专用封闭	专用开放
训练集	不可变	不可变	不可变	可变
验证集	不可变	不可变	不可变	不可变
测试集	不涉及	不涉及	不涉及	不涉及
数据预处理	不可变	自选或可变	不可变	自选或可变
训练过程中数据预处理(训练算法自带)	不可变	自选或可变	不可变	自选或可变
模型结构	不可变	自选或可变	不可变	自选或可变
优化方法	不可变	自选或可变	不可变	自选或可变
目标模型精度	不可变	不可变	不可变	自选或可变
机器学习框架	自选或可变	自选或可变	自选或可变	自选或可变
混合训练精度	不可变	自选或可变	不可变	自选或可变

6.6 指标项及测试方法

6.6.1 通则

人工智能服务器系统训练性能测试：

a） 时间(见6.6.2)和实际吞吐率(含有效计算能力，见6.6.4)为基础性能指标项；

b） 功耗(见6.6.3)和资源利用率(见6.6.5)表示训练代价；

c） 能效及效率指标项和测试方法见附录E的E.1。

6.6.2 时间

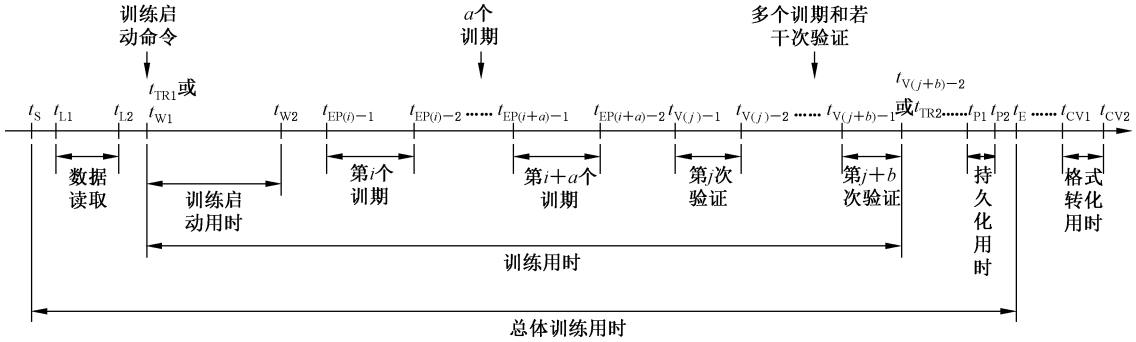
时间单位为毫秒(ms)。训练时间指标项和测试方法见表6，时间采集点见图1。

表6 训练时间指标项和测试方法

指标项	测试方法	说明
总体训练用时(T_T)[a]	a） 在读入训练数据命令前，紧邻该命令计时，获得时间点t_S； b） 在输出模型持久化完成后，串行并紧邻调用计时命令，获得时间点t_E c） 计算总体训练时间：$T_T=t_E-t_S$	从训练开始读入数据，到模型训练完毕且完成在非电易失性存储器上的持久化，所使用的总时长
数据读入用时(T_L)[b]	a） 在读入训练数据命令前，紧邻该命令计时，获得时间点t_{L1}，t_{L1}可等于t_S； b） 在训练数据读取完成时，串行并紧邻调用计时命令，获得时间点t_{L2}； c） 计算数据读入时间：$T_L=t_{L2}-t_{L1}$	为训练目的，训练数据被读入加速器内存(使用通用计算环境时，加速器缺少或不配置存储时，可为主存)，达至可用状态，所使用的时间

表 6　训练时间指标项和测试方法（续）

指标项	测试方法	说明
训练启动用时（T_W）	a）　训练开始前，串行并紧邻调用计时命令，获得时间点 t_{W1}； b）　在每个加速器进入训练状态时，取时间点，直到最后一个加速器进训练状态，获得时间点 t_{W2}； c）　计算训练启动用时：$T_W = t_{W2} - t_{W1}$	多加速器训练时，从训练开始指令到所有加速器都被分配并开始执行训练任务所经历的时长
训练用时（T_{TR}）[c,d]	a）　训练开始前，串行并紧邻调用计时命令，获得时间 t_{TR1}，t_{TR1} 可等于 t_{W1}； b）　训练退出时，串行并紧邻调用计时命令，获得时间点 t_{TR2}； c）　计算训练用时：$T_{TR} = t_{TR2} - t_{TR1}$	从训练开始命令调用到训练退出之间的时间间隔
第 i 个（i 为正整数）训期运行用时［$T_{EP(i)}$］	a）　第 i 个训期开始前，串行并紧邻调用计时命令，获得时间 $t_{EP(i)-1}$； b）　第 i 个训期结束后，串行并紧邻调用计时命令，获得时间 $t_{EP(i)-2}$； c）　第 i 个训期用时：$T_{EP(i)} = t_{EP(i)-2} - t_{EP(i)-1}$	训练过程第 i 次遍历（使用）训练集所用的时间
第 j 次（j 为正整数）验证用时［$T_{V(j)}$］	a）　第 j 次验证开始前，串行并紧邻调用计时命令，获得时间 $t_{V(j)-1}$； b）　第 j 次验证结束后，串行并紧邻调用计时命令，获得时间 $t_{V(j)-2}$； c）　第 j 次验证用时：$T_{V(j)} = t_{V(j)-2} - t_{V(j)-1}$	第 j 次使用验证数据集试运行当前模型，得出当前模型准确率等指标值的过程
模型格式转化用时（T_{CV}）	a）　模型格式转化前，串行并紧邻调用计时命令，获得时间 t_{CV1}； b）　模型转化完毕后，串行并紧邻调用计时命令，获得时间 t_{CV2}； c）　模型格式转化用时：$T_{CV} = t_{CV2} - t_{CV1}$	训练完毕后，将结果模型转化为要求格式所耗费的时间
模型持久化用时（T_P）	a）　模型持久化前，串行并紧邻调用计时命令，获得时间 t_{P1}； b）　模型持久化后，串行并紧邻调用计时命令，获得时间 t_{P2}； c）　模型持久化用时：$T_P = t_{P2} - t_{P1}$	将加速器内存中的模型读出，并完整写入非电易失性存储所用的时间
节点间通信时延（T_{NC}）	a）　在发送数据前，串行并紧邻调用计时命令，获得时间 t_{NC1}； b）　在完整接收数据后，串行并紧邻调用计时命令，获得时间 t_{NC2}； c）　节点间通信时延 $T_{NC} = t_{NC2} - t_{NC1}$	源节点开始发送数据至目标节点完全接收数据的用时
注：假设训练数据已封装为机器学习框架能处理的格式。		
[a] 数据并行时，数据读入用时为数据读入开始至所有工作节点都完整获得所需数据的总用时（含网络传输用时）。 [b] 数据读入过程可伴随训练同步发生，时间计入训练用时。 [c] 训练任务的用时包含数据预处理用时。 [d] 受测系统无法统计的时间，不作要求。		



注 1：训练时间按“训练用时”计。

注 2：数据读入过程可伴随训练同步发生。

注 3：训期表示讲训练数据集中的所有样本都处理一遍的训练过程。

图 1　训练时间序

6.6.3　功耗

训练功耗单位为瓦(W)。训练功耗指标项和测试方法见表 7。

表 7　训练功耗指标项和测试方法

指标项	测试方法	说明
人工智能服务器单机训练平均功率	a)　在 SUT,配套使用功率计； b)　在训练用时中(见图 1),周期性测试整机的负载功率(每秒采样 1 次),并求均值 P_{TR}	单台人工智能服务器在某次训练用时内(T_{TR})的平均功率
人工智能服务器单机训练瞬时峰值功率	a)　在 SUT,配套使用功率计； b)　在训练过程中,周期性测试整机的负载功率(每秒采样 1 次),记录最大负载功率计量值 P_{TRmax}	单台人工智能服务器在某次训练全程(T_{TR})中,服务器正常工作状态下的最大瞬时功率
人工智能服务器集群训练平均功率	a)　在 SUT 各服务器配套使用功率计； b)　按单机训练平均功率测试方法实施,测得每服务器 i(i 为正整数)的平均功率 P_{TR-i}； c)　求和得到集群平均功率： $P_{CTR}=\sum_i P_{TR-i}$	人工智能服务器集群,在某次训练全程(T_{TR})中的平均功率

6.6.4　实际吞吐率

实际吞吐率代表人工智能服务器系统对特定训练作业的有效计算能力,提高有效计算能力可达到硬件系统扩容的同样效果。对视觉类测试,单位为图片数每秒(图片数/s);对自然语言处理类测试,单位为句数每秒(句数/s);对自然语言语句生成类测试,定长输入(句中单词或字的个数)或输出条件下,单位为语素数每秒(语素数/s)。训练实际吞吐率指标项和测试方法见表 8。

表 8　训练实际吞吐率指标项和测试方法

指标项	测试方法	说明
人工智能服务器训练实际吞吐率(Th)	a)　统计每个训期 i(i 为正整数)所使用的时间 $T_{EP(i)}$； b)　基于 a)的结果，统计每训期平均用时 T_{EP}； c)　计算训练实际吞吐率[a]： $$Th=\frac{numberof(训练集)}{T_{EP}}$$ 对文本生成类的训练负载，训练实际吞吐率为： $$Th=\frac{numberoftokens(训练集)}{T_{EP}}$$	人工智能服务器系统在训练过程中，每个训期处理的数据量与时间的比值
人工智能服务器集群训练实际吞吐率(Th_n)	a)　在集群每个节点 n(n 为正整数)上，计算该节点训练吞吐率 Th_n； b)　计算集群训练实际吞吐率(Th_{CL})[a]： $$Th_{CL}=\sum_n Th_n$$；	
人工智能服务器系统训练有效计算能力(人工智能服务器系统训练吞吐率综合加速比)($\overline{Th}$)	a)　对于给定的训练场景集合 S，对每个场景负载 $s\in S$，使用某特定参照计算系统，在 s 上测得吞吐率 Th_s^*，作为基线； b)　设 SUT 在 s 上测得的训练实际吞吐率为 Th_s，则训练综合相对吞吐率，由 $\frac{Th_s}{Th_s^*}$ 在 s 上的加权几何平均计算，其中 τ_s 表示每个场景负载 s 所对应的加权值(0～1 区间)，用来表示不同任务场景的价值权重的差异，全集 S 对应的所有 τ_s 累加求和为 1，α 为调整系数($\alpha>0$，$\alpha\in R^+$)，默认为 100.0： $$\overline{Th}=\alpha\cdot\sqrt[\sum\tau_s]{\prod_s\left(\frac{Th_s}{Th_s^*}\right)^{\tau_s}}$$	人工智能服务器系统在给定任务集合 S 上，实际吞吐率与每任务基线吞吐率之比的加权几何平均[b]

[a] numberof(·)表示计量特定数据集合所含样本的数量，numberoftokens(·)表示计量特定数据集合所含的语素数量。

[b] 基线吞吐率是参考计算系统在特定场景上的吞吐率，τ、α 和参照计算系统由测试者按实测需求确定。

6.6.5　资源利用率

资源利用率包含加速器利用率(%)。训练资源利用率指标项和测试方法见表 9。

表 9　训练资源利用率指标项和测试方法

指标项	测试方法	说明
人工智能服务器加速器资源利用率[a]	a) 在每个训期 i 内(假设一次训练过程有 I 个训期),对每个加速芯片 k(假设有 K 个加速芯片),采样 N 次使用率 p_{k-n}(i,k,N,n,K,I 为正整数,$N \geqslant 3$,$0 \leqslant p_{k-n} < 1$,为正实数,精确到 0.01); b) 对每个加速芯片 k,求出在 i 的平均利用率:$$p_{k-i} = \frac{\sum_n^N p_{k-n}}{N}$$	训练期间(T_{TR}),服务器上所有指定参与训练任务的加速芯片的平均利用率
人工智能服务器集群加速器资源利用率[a]	c) 对每个训期 i,求出多芯片平均利用率(如 $K=1$,本步骤忽略):$$p_i = \frac{\sum_k^K p_{k-i}}{K}$$ d) 对所有训期求平均,得出训练阶段人工智能服务器加速器资源利用率	训练期间(T_{TR}),服务器集群上所有指定参与训练任务的加速芯片的平均利用率
[a] 数据传输芯片利用率不含在加速器资源利用率计算范围内。		

6.7　训练用测试系统要求

6.7.1　功能要求

测试系统功能符合以下要求。

a) 能自动检测或手动写入被测系统软件和硬件信息,符合 6.1 的要求。
b) 能使用机器学习框架、被测系统提供的使能软件函数库和其他必要信息,完成 6.6 要求指标项的测试,提供指标项计算函数。
c) 至少能实施 6.4.1 要求的场景的测试;对 6.4.2 要求的场景,可实施改配或必要编码。
d) 至少实现以下计算精度中的一种:
   1) FP64;
   2) FP32;
   3) TF32;
   4) FP16;
   5) BF16;
   6) INT8;
   7) UINT8;
   8) INT4;
   9) UINT4。
e) 实现配置了容器或使用虚拟化组件的人工智能服务器系统的性能测试。
f) 测试完成后能完全卸载,不残留任何测试组件(不含测试数据)。
g) 提供日志函数,日志所含内容及格式符合 6.3c)4)的要求。
h) 实现测试者对测试过程的管理和监测,包含但不限于:
   1) 训练过程子阶段开始或完成事件,包含训练测试开始,每次训练的开始和结束,训练测试整体进度,训练测试整体进度,训练测试结束和训练结果上传完成;
   2) 训练结果信息,符合 6.3c)的要求;

3） 测试者对重测的允许和次数控制；
4） 能提供证据辅助测试者实施训练结果的有效性判定，或自动判定。

i） 在提前获得测试项目授权后，被测者可在测试期内的任意时间发起测试。
j） 可为不同测试项维护独立的训练结果目录。
k） 可实现本地测试（测试者不介入的测试，如预测试或系统调试等）和网络测试（测试者介入）。

6.7.2 公平性保障要求

测试系统应提供方案和实现保障公平性：
a） 防止对指标项计算函数的修改；
b） 防止测试时对指标计算函数的替代使用；
c） 防止在训练结果上传前对训练结果的修改；
d） 防止在测试开始后，结果上传完毕之前对测试代码的修改；
e） 防止除测试系统外的其他进程向被测者传输训练过程和训练结果；
f） 实施网络测试时，关于测试者授权的鉴别；
g） 测试过程中测试者与被测者通信的加密，信息完整性检查。

7 推理性能测试

7.1 测试过程

推理测试过程含以下步骤。
a） 信息准备，被测者应向测试者提供 6.1a）1）～20）及以下内容：
1） 是否使用稀疏化（0 表示不使用、1 表示使用，附加方法名称）；
2） 是否使用量化（0 表示不使用、1 表示使用，附加量化方法名称）。
b） 测试准备，被测者向测试者发送测试请求，取得测试集；测试者指定测试数据集，告知获取方法；被测系统下载数据集，检验合规性。
c） 测试运行，被测者按测试内容，载入模型（可预先准备好）和数据集；被测者运行测试；记录过程数据，计算指标值；结果合规性检查。
d） 结果报送，被测者发送推理结果给测试者。
e） 结果审核，推理结果应符合 7.3 的要求。

7.2 推理测试要求

推理测试过程，符合以下要求。
a） 合规性要求：
1） 推理测试源码：应实现必要接口（数据准备、输入和输出）；应使用测试系统提供的指标计算方法；应使用测试系统提供的日志记录方法；不应对已实现的指标测试函数或测试流程控制函数实施改动、继承或重载（要求被测者实现的函数或接口除外）；
2） 测试应使用工具进行日志记录、数据采集、指标项计算，人工智能服务器系统性能测试工具示例见附录 A；
3） 推理过程：模型编译或部署时，不应使用其他模型替换测试模型；测试前，除数据集封装格式转化外，不应浏览或记录数据、修改数据（非预处理）、浏览数据或复制数据，不应分析、提取或缓存数据特征；
4） 测试过程中，不应使用推理测试进程之外的任何进程，修改或记录日志；不应使用推理测试进程之外的任何进程，存取测试输入或输出数据；不应缓存或复用输入、输出和过程（预

处理结果或后处理输入)数据;不应修改内存中模型参数;不应保存或缓存后处理过程输入数据;不应记录、分析或使用作业到达模式来预测某时段内的作业量;不应根据测试过程中的准确率、丢失率等指结果,故意忽略待处理数据。

b) 封闭模式推理时,模型压缩,不应实施以下操作:
   1) 删除非零权重;
   2) 使用剪枝或其他改变模型结构的方法;
   3) 实施模型蒸馏。
c) 封闭模式推理时,模型量化符合以下要求:
   1) 不同场景下量化的模型对象应与表 11 一致;
   2) 量化结果不应出现 7.2a)3)列出的情况。
d) 推理精度应符合表 11～表 14 的要求。
e) 应声明推理所用批大小的信息,符合 6.1a)20)的要求。

7.3 推理测试结果

推理测试结果应包含以下信息:

a) 7.1a)规定的配置信息;
b) 推理作业到达模式编号(见表 10);
c) 推理使用的实际精度;
d) 场景要求的指标值(见表 11 和表 13);
e) 推理测试代码公开规则符合附录 C 的要求;
f) 推理日志,日志周期性输出,每条日志的格式为[yyyy:MM:dd HH:mm:ss]-[accuracy]-[已处理作业数]-[已处理样本数]-[样本丢失数],其中:
   1) 第一项为本条日志输出时的时间戳;
   2) 第二项为当前累计的准确率;
   3) 第三项为当前已返回结果的作业数;
   4) 第四项为当前已返回结果的样本数;
   5) 第五项为当前未能在超时范围内处理的样本数,即丢失样本数。

7.4 测试场景

7.4.1 推理作业要求

推理作业应符合以下要求:

a) 作业从测试系统发往被测系统,结果从被测系统发送回测试系统;
b) 每个样本仅含有推理模块要求的必要(输入)参数,不含有额外信息;
c) 推理作业采用特定的到达模式,符合 7.4.2 的要求;
d) 作业丢失指被测系统无法在超时控制门限内返回结果的情况;
e) 超时控制门限指测试者从发送作业到收到对应结果之间允许的最大时间间隔;
f) 按特定推理测试负载的要求,单个样本包含视觉、自然语言和声音等 1 个或多个模态的数据;
g) 使用多模态场景负载测试时,按推理负载的定义,将每个样本按模态占比分为输入和期望的输出;
h) 零次样本(zero-shot)推理,使用随机数据集(生成方法见 7.4.3);
i) N 次样本(N-shot,N 是自然数)推理(包含单样本或少样本的情况),使用随机数据集训练模型 N 个训期后,执行推理测试。

7.4.2 推理作业到达模式

推理作业到达模式应从表 10 中选择。

表 10 推理作业到达模式

到达模式	编号	作业发送方法	作业可缓存（是/否）	运行次数（次）	超时控制门限 1（s）	超时控制门限 2[a]（s）
连续（单一）到达	0	第 i（i 为正整数）个作业在第（$i-1$）个作业完成后紧邻到达。作业（$i-1$）未完成或超时控制门限未达到时，作业 i 不发送	否	1	2	10
固定周期到达	1	作业以固定周期（T）到达，一次到达 n 个作业（n 为正整数）	是	1	4	20
泊松分布到达	2	作业以泊松分布到达：$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!}$$ 式中： k ——某单位时间内到达的作业数，k 为正整数； λ ——是单位时间（如每秒）作业平均到达次数，λ 为正整数	是	1	4	20
高峰到达	3	泊松分布到达模式中，有 j 个短周期，每周期内有突发性大量作业，周期持续一定时长 T_G（如 5 s～10 s），并维持一定并发度水平 σ（σ 为正整数，如 $\sigma>2^{10}$ 个作业/s），短周期内的作业到达，符合固定周期到达模式（T 与 n 可在测试时结合需要选取）	是	1	60	240
离线	4	一次性全部到达	是	1	不涉及	不涉及
混合作业到达	5	在连续到达、固定周期到达、泊松分布到达、高峰到达或离线到达模式中，加入与当前测试场景不同的作业	是	1	取对应超时控制门限值	取对应超时控制门限值
[a] chatGLM V2 6B、Llama2 13B、GLM 130B 和 Stabdiffusion V2.1 符合超时控制门限 2，其余模型符合超时控制门限 1。						

7.4.3 随机数据集生成方法

推理测试使用随机数据集时，应按以下方法生成：

a） 建立空数据集 D；

b） 建立空样本 d；

c） 按 7.4.4 的规定生成特定模态的数据 1 个，加入 d；

d） 如随机数据集含多模态样本，基于 c）的结果生成所需模态的数据，加入 d；

e) 如模态要求不符合，则选择模型并重复执行 c)；如已符合模态要求，则将样本 d 加入数据集 D；

f) 如数据集的样本数量不符合要求，执行 b)；如已符合要求，返回数据集 D。

7.4.4 模态数据生成方法

7.4.4.1 自然语言语句生成方法

按以下方法生成自然语言语句：

a) 生成最大 M 个(M 为自然数)随机的名字或动词，作为关键词；

b) 利用关键词，使用特定模型(按测试需要选择)生成包含 L 个词(L 为自然数，$L>M$)的句子。

7.4.4.2 图像生成方法

按以下方法生成图像数据。

a) 按 7.4.4.1 生成自然语言语句，以语句为输入，使用特定图像生成模型(按测试需要选择)生成图像。

b) 使用自然语言语句的关键字(名词)作为图像类别。

c) 以图像生成相似图像时，使用特定图像数据集和特定图像生成模型(按测试需要选择)生成图像，使用表 11 中 resnet50_v1.5 将原图与生成图像归类并筛选同类图像，作为相似图像。必要时，可再由人工抽检并标注所生成图像与原图的相似性。

7.4.5 通用测试场景

7.4.5.1 通用封闭模式测试应从表 11 所列测试场景中选择，测试场景类型说明见附录 D。

表 11 通用推理性能测试场景(封闭模式)

序号	类型	场景说明	
1	图像识别	模型	inception_v3
		测试集来源	imagenet2012
		作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达，Top1-准确率＞77.3%
2	图像识别	模型	resnet50_v1.5
		测试集来源	imagenet2012
		作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达，Top1-准确率＞74%
3	物体检测	模型	yolo_v5s-6-0
		测试集来源	coco2017
		作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达，mAP@0.5＞55.9%
4	语义分割	模型	deeplab_v3
		测试集来源	coco2017
		作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达，mIOU＞85%

表 11　通用推理性能测试场景(封闭模式)(续)

序号	类型	场景说明		
5	推荐	模型		wide&deep
		测试集来源		Criteo(Kaggle Display Advertising Challenge Dataset)
		作业到达模式及参数[a,b]		连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达，AUC＞72%
6	推荐	模型		DLRM
		测试集来源		Criteo(1TB Click Logs)
		作业到达模式及参数[a,b]		连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达，AUC＞80.25%
7	自然语言处理	模型		Bert-large
		测试集来源		SQuAD1.1
		作业到达模式及参数[a,b]		连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达，exact_match＞83.57%，f1 score＞90.75
8	自然语言处理	模型		chatGLM V2 6B
		精度测试	任务	BoolQ
			门限	AvgScore＞74
		性能测试	数据集	构造数据集[c]
			作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、高峰到达或离线到达
9	自然语言处理	模型		LLaMa2-13B
		精度测试	任务	BoolQ
			门限	AvgScore＜77.4
		性能测试	数据集	构造数据集[c]
			作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、高峰到达或离线到达
10	自然语言处理	模型		GLM 130B
		精度测试	任务	C-Eval
			门限	AvgScore＜39
		性能测试	数据集	构造数据集[c]
			作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、高峰到达或离线到达
11	多模态	模型		Stable Diffusion V2.1
		精度测试	任务	Parti
			门限	CLIP score＞0.369
		性能测试	数据集	构造数据集[c]
			作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、高峰到达或离线到达

表 11　通用推理性能测试场景(封闭模式)(续)

序号	类型	场景说明	
12	语音识别	模型	wav2vec2_0
		测试集来源	Aishell-1
		作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达(j=2),WER<2.96%
[a] 未注明时,泊松分布或固定周期到达模式涉及的参数,参考值为 $\lambda=5$,$T=500$ ms,$n=1$。k 值由测试方给出,但同批次测试的 k 值应一致。 [b] 准确率门限的值为参考值。 [c] 未注明时,默认构造数据集分布采用[输入序列长度,输出序列长度]∈{[256,256],[512,512],[1024,1024],[2048,2048]}。			

7.4.5.2　通用开放模式测试应从表 12 所列测试场景中选择,测试场景类型说明见附录 D。

表 12　通用推理性能测试场景(开放模式)

序号	类型	场景说明	
1	图像识别	测试集来源	Imagenet2012
		门限	Top1-准确率>75%
		作业到达模式及参数[a,b]	连续单一、固定周期到达($T=500$ ms)、泊松分布到达($\lambda=5$)、离线到达或高峰到达
2	物体检测	测试集来源	coco2017
		门限	mAP>57%
		作业到达模式及参数[a,b]	连续单一、固定周期到达($T=500$ ms)、泊松分布到达($\lambda=5$)、离线到达或高峰到达
3	语义分割	测试集来源	coco2017
		门限	mIOU>85%
		作业到达模式及参数[a,b]	连续单一、固定周期到达($T=500$ ms)、泊松分布到达($\lambda=5$)、离线到达或高峰到达
4	推荐	测试集来源	Criteo(Kaggle Display Advertising Challenge Dataset)
		门限	AUC>72%
		作业到达模式及参数[a,b]	连续单一、固定周期到达($T=500$ ms)、泊松分布到达($\lambda=5$)、离线到达或高峰到达
5	推荐	测试集来源	Criteo(1TB Click Logs)
		门限	AUC>80.25%
		作业到达模式及参数[a,b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达
6	自然语言处理	测试集来源	cn-wiki
		门限	mask_lm accuracy>0.7
		作业到达模式及参数[a,b]	连续单一、固定周期到达($T=500$ ms)、泊松分布到达($\lambda=5$)、离线到达或高峰到达

表 12　通用推理性能测试场景(开放模式)(续)

序号	类型	场景说明	
7	自然语言处理	测试集来源	WMT18 英→德、英→中
		门限	BLEU＞24％(适用于二种翻译)
		作业到达模式及参数[a,b]	连续单一、固定周期到达(T＝500 ms)、泊松分布到达(λ＝5)、离线到达或高峰到达
8	语音识别	测试集来源	Aishell-1
		门限	WER＜7％
		作业到达模式及参数[a,b]	连续单一、固定周期到达(T＝500 ms)、泊松分布到达(λ＝5)、离线到达或高峰到达

[a] 未注明时,泊松分布或固定周期到达模式涉及的参数,参考值为 $\lambda=5$,$T=500$ ms,$n=1$。k 值由测试方给出,但同批次测试的 k 值应一致。

[b] 准确率的具体数值为参考值。

7.4.6　专用测试场景

7.4.6.1　专用封闭模式测试应从表 13 所列测试场景中选择,测试场景类型说明见附录 D。

表 13　专用推理性能场景(封闭模式)

序号	类型	场景说明	
1	OCR(无预分割)	模型	DBNET
		测试集来源[a]	icdar2015
		作业到达模式及参数[b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达,mAP＞0.88
2	人脸识别	模型	FaceNet
		测试集来源[a]	LFW
		作业到达模式及参数[b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达
3	语音识别	模型	Conformer(ESPnet2)
		测试集来源[a]	Aishell-1
		作业到达模式及参数[b]	连续单一、固定周期到达、泊松分布到达、离线到达或高峰到达,err＜5.4％

[a] 推理数据的格式,没有严格的限定,被测者可根据本地机器学习框架进行格式转换,格式转换过程不应改变数据的值(如图像像素值),数据格式转换过程不计时。

[b] 未注明时,泊松分布或固定周期到达模式涉及的参数,参考值为 $\lambda=5$,$T=500$ ms,$n=1$。k 值由测试方给出,但同批次测试的 k 值应一致。

7.4.6.2　专用开放模式测试应从表 14 所列测试场景中选择,测试场景说明见附录 D。

表 14　专用推理测试场景(开放模式)

序号	类型	场景说明	
1	OCR(无预分割)	测试集来源[a]	金融行业测试集[b]
		作业到达模式[c]及参数	连续单一,固定周期到达,泊松分布到达,离线到达,高峰到达
2	人脸识别	测试集来源[a]	LFW
		作业到达模式[c]及参数	连续单一,固定周期到达,泊松分布到达,离线到达,高峰到达

[a] 推理数据的格式,没有严格的限定,被测者可根据本地机器学习框架进行格式转换,格式转换过程不应改变数据的值(如图像像素值),数据格式转换过程不计时。

[b] 未确定的模型及数据集,可在具体测试前,由测试者按专用系统的要求统一确定。

[c] 未注明时,泊松分布或固定周期到达模式涉及的参数,参考值为 $\lambda=5$,$T=500$ ms,$n=1$。k 值由测试方给出,但同批次测试的 k 值应一致。

7.5　场景配置要求

针对目标的不同,推理性能测试分为通用测试和专用测试:

a)　通用测试是指针对共性问题,使用公共可获得的模型和数据集,完成推理测试;

b)　专用测试是针对行业领域问题,使用专用模型和数据集,完成推理测试。

推理测试场景可变要素配置要求见表 15。

表 15　推理测试场景可变要素配置要求

可变要素	通用封闭	通用开放	专用封闭	专用开放
训练集	不涉及	自选或可变	不涉及	自选或可变
验证集	不涉及	自选或可变	不涉及	自选或可变
测试集	不可变	不可变	不可变	不可变
数据预处理算法	不可变	自选或可变	不可变	自选或可变
数据后处理算法	不可变	自选或可变	不可变	自选或可变
模型结构	不可变	自选或可变	不可变	自选或可变
模型格式	不可变	自选或可变	不可变	自选或可变
模型压缩方法	不可变	自选或可变	不可变	自选或可变
压缩后精度	不可变	自选或可变	不可变	自选或可变

7.6　指标项及测试方法

7.6.1　通则

人工智能服务器系统推理性能测试:

a)　时间(见 7.6.2)和实际吞吐率(含有效计算能力,见 7.6.4)为基础性能指标项;

b)　功耗(见 7.6.3)表示推理代价;

c) 弹性(见 7.6.5)、承压力(见 7.6.6)和视频分析最大路数(见 7.6.7)表示适用性;

d) 能效和效率指标项和测试方法见 E.2。

7.6.2 时间

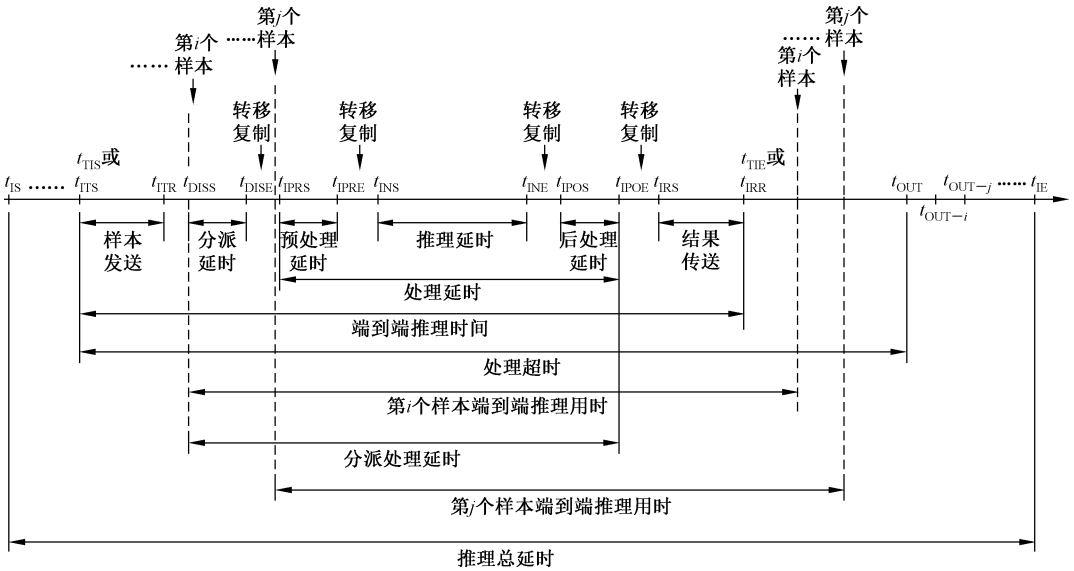
时间单位为毫秒(ms)。推理时间指标项和测试方法见表 16,时间采集点见图 2。

表 16 推理时间指标项和测试方法

指标项	测试方法	说明[a]
推理总延时(T_{I})[b]	a) 测试者在发送第 1 个样本的第 1 字节前,紧邻计时,得到时间点 t_{IS}; b) 测试者在接收到所有样本的最后 1 字节后,紧邻或在最后一个处理超时时间点计时,得到时间点 t_{IE}; c) 计算得到推理总延时 $T_{\mathrm{I}}=t_{\mathrm{IE}}-t_{\mathrm{IS}}$	多次连续推理端到端总延时
端到端推理延时(T_{TI})	a) 测试者在发送某样本第 1 字节前,紧邻计时,得到时间点 t_{TIS}; b) 测试者在接收完该样本返回结果的最后 1 字节后,紧邻计时,得到时间点 t_{TIE}; c) 计算端到端推理延时:$T_{\mathrm{TI}}=t_{\mathrm{TIE}}-t_{\mathrm{TIS}}$	测试者发送样本时间与收到结果时间的差
样本发送延时(T_{IT})	a) 测试者在发送某样本第 1 字节前,紧邻计时,得到时间点 t_{ITS}($t_{\mathrm{ITS}}=t_{\mathrm{TIS}}$); b) 被测者在收到样本最后 1 字节后,紧邻计时,得到时间点 t_{ITR}; c) 计算作业发送延时:$T_{\mathrm{IT}}=t_{\mathrm{ITR}}-t_{\mathrm{ITS}}$	测试者发送样本时间与被测者收到样本时间的差
结果传送延时(T_{IR})	a) 被测者在发送结果第 1 字节前,紧邻计时,得到时间点 t_{IRS}; b) 测试者在收到结果最后 1 字节后,紧邻计时,得到时间点 t_{IRR}($t_{\mathrm{TIE}}=t_{\mathrm{IRR}}$); c) 计算结果传送延时:$T_{\mathrm{IR}}=t_{\mathrm{IRR}}-t_{\mathrm{IRS}}$	被测者发送结果时间与测试者收到结果时间的差
任务分派延时(T_{DIS})	a) 被测者收到样本最后 1 字节后,紧邻计时,得到时间点 t_{DISS}; b) 被测者开始处理前,紧邻计时,得到时间点 t_{DISE}; c) 计算任务分派延时 $T_{\mathrm{DIS}}=t_{\mathrm{DISE}}-t_{\mathrm{DISS}}$	被测者收到样本时间到处理前时间的差
预处理延时(T_{IPR})	a) 被测者对某样本的预处理开始前,紧邻计时,得到时间点 t_{IPRS}; b) 被测者对某样本的预处理结束后,紧邻计时,得到时间点 t_{IPRE}; c) 计算预处理延时 $T_{\mathrm{IPR}}=t_{\mathrm{IPRE}}-t_{\mathrm{IPRS}}$	被测者对某样本预处理的开始时间与结束时间的差
推理延时(T_{IN})	a) 被测者针对某样本推理开始前,紧邻计时,得到时间点 t_{INS}; b) 被测者针对某样本推理结束后,紧邻计时,得到时间点 t_{INE}; c) 计算推理延时 $T_{\mathrm{IN}}=t_{\mathrm{INE}}-t_{\mathrm{INS}}$	被测者对某样本推理的开始时间与结束时间的差
后处理延时(T_{IPO})	a) 被测者对某样本的后处理开始前,紧邻计时,得到时间点 t_{IPOS}; b) 被测者对某样本的后处理结束后,紧邻计时,得到时间点 t_{IPOE}; c) 计算后处理延时 $T_{\mathrm{IPO}}=t_{\mathrm{IPOE}}-t_{\mathrm{IPOS}}$	被测者对某样本后处理的开始时间与结束时间的差
样本处理延时(T_{IP})	a) 被测者对某样本的处理开始前,紧邻计时,得到时间点 t_{IPS}($t_{\mathrm{IPS}}=t_{\mathrm{IPRS}}$); b) 被测者对某样本的处理结束后,紧邻计时,得到时间点 t_{IPE}($t_{\mathrm{IPE}}=t_{\mathrm{IPOE}}$); c) 计算样本处理延时 $T_{\mathrm{IP}}=t_{\mathrm{IPE}}-t_{\mathrm{IPS}}$	被测者处理样本的开始时间与结束时间的差。处理延时约是预处理、推理和后处理时间的总和

表 16　推理时间指标项和测试方法（续）

指标项	测试方法	说明[a]
分派处理延时（T_{DIP}）	a）　被测者收到样本最后 1 字节后，紧邻计时，得到时间点 t_{DIPS}（$t_{DIPS}=t_{DISS}$）； b）　被测者对某样本的处理结束后，紧邻计时，得到时间点 t_{DIPE}（$t_{DIPE}=t_{IPE}$）； c）　计算分派处理延时 $T_{DIP}=t_{DIPE}-t_{DIPS}$	被测者完整收到样本的时间与处理结束时间的差
处理超时（T_{OUT}）	计算某样本处理超时：$T_{OUT}=t_{TIS}+t_{OUT}{}^{*}$，$t_{OUT}{}^{*}$ 为常量	测试者从发送样本到收到对应结果的允许的最大时间间隔
首语素延时（T_{ftoken}）	a）　被测者在收到样本最后 1 字节后，紧邻计时，得到时间点 t_{ITR}； b）　被测者在发送结果第 1 字节前，紧邻计时，得到时间点 t_{IRS}； c）　计算作业延时：$T_{ftoken}=t_{IRS}-t_{ITR}$	被测者收到样本时间与被测者发送第一个语素时间的差
下个语素平均延时（T_{ntoken}）	a）　测试者在发送某样本前，紧邻计时，得到时间点 t_{IRS}^{n}； b）　测试者在发送某样本下一个样本前，紧邻计时，得到时间点 t_{IRS}^{n+1}； c）　计算平均作业延时：$T_{ntoken}=\mathrm{Average}(t_{IRS}^{n+1}-t_{IRS}^{n})$	被测者发送某一个语素与发送下一个语素时间差的平均值
注：时间的差为绝对值。		
[a] 因作业到达模式不同，推理总延时 T_1 可能包括被测者等待作业的间隔时间。 [b] 处理前时间的计法为：存在预处理时，以预处理开始时间计；如不存在，以推理开始时间计。		



注 1：以 i，j 表示并发的样本及处理过程，但推理并发处理不是人工智能服务器系统的必要功能。

注 2：推理的中间结果，可在人工智能服务器系统内部转移或复制，以便处理。

图 2　推理时间序

7.6.3 功耗

推理功耗以功率计算，单位为瓦(W)。推理功耗指标项和测试方法见表17。

表17 推理功耗指标项和测试方法

指标项	测试方法	说明
人工智能服务器单机推理平均功率	a) 在SUT，配套使用功率计； b) 在推理延时(T_{IN})期间，周期性测试整机的负载功率； c) 求均值	单台人工智能服务器在某次推理全程中的平均功率
人工智能服务器数据预处理平均功率	a) 在SUT，配套使用功率计； b) 在数据预处理延时(T_{IPR})期间，周期性测试整机的负载功率； c) 求均值	单台人工智能服务器在某次推理全程中，数据预处理阶段的平均功率
人工智能服务器推理峰值功率	a) 在SUT，配套使用功率计； b) 在数据预处理延时(T_{IN})期间，周期性测试整机的负载功率； c) 取最大值	单台人工智能服务器在某次推理全程中，服务器正常工作状态下的最大瞬时功率
人工智能服务器集群推理平均功率	a) 在SUT各服务器配套使用功率计； b) 在相同时间点，周期性测试每个服务器的负载功率； c) 相同时间点各服务器功率加和为集群瞬时负载功率； d) 求均值	人工智能服务器集群，在某次推理全程中(T_{IN})中的平均功率

7.6.4 实际吞吐率

实际吞吐率代表人工智能服务器系统对特定推理作业的有效计算能力，提高有效计算能力可达到硬件系统扩容的同样效果。对视觉类测试，单位为图像数每秒(图像数/s)；对自然语言处理类测试，单位为句数每秒(句数/s)；对自然语言语句生成类测试，定长输入(句中单词或字的个数)或输出条件下，单位为语素数每秒(语素数/s)。推理实际吞吐率指标项和测试方法见表18。

表18 推理实际吞吐率指标项和测试方法

指标项	测试方法	说明
人工智能服务器系统推理实际吞吐率	a) 在整个推理测试过程中(T_{IN}内)，累计所有实际发送的样本，及实际返回结果，计算样本数量； b) 计算其与实际分派处理延时总覆盖时间的比值	人工智能服务器系统在单位时间内，对于特定任务负载，完整处理的样本数量
	a) 在推理测试过程中(T_{IN}内)，累计所有实际发送的样本，及实际返回结果，计算样本数量； b) 对每个样本，累计语素数量； c) 计算语素数量与实际分派处理延时总覆盖时间的比值； d) 首语素可单独计算延时，并可不计入平均	人工智能服务器系统在单位时间内，对语言生成类负载，完成处理的语素数量

表 18　推理实际吞吐率指标项和测试方法（续）

指标项	测试方法	说明
人工智能服务器系统推理有效计算能力（人工智能服务器系统推理吞吐率综合加速比）	a）　对每个场景负载 $s \in S$，使用某特定参照计算系统，在 s 上测得吞吐率，作为基线； b）　对每个场景负载 $s \in S$，使用 SUT，在 s 上测得推理实际吞吐率； c）　使用表 8 中人工智能服务器系统训练吞吐率综合加速比的公式计算	人工智能服务器系统在给定任务集合 S 上，实际吞吐率与每任务基线吞吐率之比的加权几何平均

7.6.5　弹性

推理弹性单位为百分率每兆字节（%/MB）。推理弹性指标项和测试方法见表 19。

表 19　推理弹性指标项和测试方法

指标项	测试方法	说明
人工智能服务器系统推理弹性	a）　使用高峰模式； b）　被测者记录，每单位时间内，收到的样本数据总量及对应的最大分派处理时间； c）　当第 $i+1$ 个单位时间收到的样本数据总量＞第 i 个单位时间内收到的样本数据总量时，计算区间 $(i, i+1)$ 的推理弹性[a]： $$\mathrm{EL}(i,i+1)=\frac{\dfrac{\max_{i+1}(T_{\mathrm{DIP}})-\max_{i}(T_{\mathrm{DIP}})}{\max_{i}(T_{\mathrm{DIP}})}}{\dfrac{\mathrm{sizeInDuration}(i+1)-\mathrm{sizeInDuration}(i)}{\mathrm{sizeInDuration}(i)}}\times\mathrm{sizeInDuration}(i)$$ 式中： T_{DIP}——被测者完整收到样本的时间与处理结束时间的差。 d）　计算推理弹性： $$\mathrm{EL}=\frac{\sum_i \mathrm{EL}(i,i+1)}{N}$$ 式中： N——c）中的区间数	被测人工智能服务器系统所处理的数量增加时，分派处理时间的变化
注：假设被测系统能根据样本数量调整处理能力。		
[a] sizeInDuration(·)表示计量特定时间段内，获得、输出或持有的数据量。		

7.6.6　承压力

推理承压力的单位为兆字节每秒（MB/s）。推理承压力指标项和测试方法见表 20。

表 20　推理承压力指标项和测试方法

指标项	测试方法	说明
人工智能服务器或集群推理承压力	a)　使用高峰模式； b)　被测者周期性获取并发度，记录其大于并发度压力门限的总时长，并记录期间已处理的样本数据总量； c)　计算推理承压力：数据总量与总时长的比值	被测人工智能服务器系统在并发压力门限[单位为兆字节每秒(MB/s)]以上运行时的实际吞吐率
注：针对不同场景，实际吞吐率使用 7.6.4 规定的指标。		

7.6.7　视频分析最大路数

视频分析最大路数，单位为路。视频分析最大路数指标项和测试方法见表 21。

表 21　视频分析最大路数指标项和测试方法

指标项	测试方法	说明
人工智能服务器视频[a]分析最大路数	a)　在视觉类模型场景的测试时，在实际推理前使用解码器(软件或硬件实现)； b)　使用固定周期到达模式[b]，用 n 模拟视频路数，n 初始值为 1，每个作业含有 1 帧(1 个图像样本)； c)　并行发送 n 路视频，帧率：$f=\frac{1}{T}$ d)　如被测系统能按表 10 规定的超时门限输出处理结果，则将 n 的值调整为($n+1$)； e)　重复步骤 a)和 b)，直至有任 1 路视频无法在规定的超时门限返回处理结果为止，则视频分析最大路数为($n-1$)路	被测人工智能服务器系统，在给定响应超时门限下(表 10 中编号 1)，分析视频流，能承受的最大路数
[a] 可使用视觉类模型，如表 11～表 14 规定的图像识别、物体检测、语义分割、OCR 或人脸识别模型。 [b] 到达周期(T)可按实际测试需求确定(如 25 帧/s，30 帧/s 等)。		

7.7　推理用测试系统要求

7.7.1　功能要求

测试系统功能应符合以下要求。

a)　能自动检测或手动写入被测系统软件和硬件信息，符合 7.1 的要求。

b)　能自动发送作业，作业到达模式符合表 10 的要求。

c)　能接收推理结果并为完成计算指标实现必要的功能，包含：

1)　超时作业识别及统计；

2)　作业超时(丢失)率门限检查。

d)　能使用机器学习框架，人工智能服务器系统提供的使能软件函数库和其他必要信息，完成 7.6 要求指标项的测试，提供指标项计算函数。

e)　至少能实施 7.4.5 要求的场景的测试；对 7.4.6 要求的场景，可实施改配或必要编码。

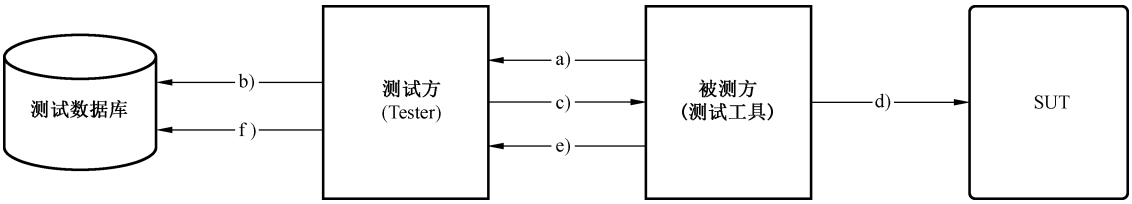
f) 数据类型符合 6.7.1 d)的要求。
g) 实现配置了容器或使用虚拟化组件的人工智能服务器系统的性能测试。
h) 测试完成后能完全卸载,不残留任何测试组件(不含测试数据)。
i) 提供日志函数,日志所含内容及格式符合 7.3 f)的要求。
j) 测试者可对测试过程的管理和监测,包含:
   1) 推理过程子阶段的开始或完成事件,包含推理测试开始;推理作业下载开始;推理作业下载完成;推理测试结束;推理结果上传完成;
   2) 推理结果,符合 7.3 的要求;
   3) 测试者对重测的允许及次数控制;
   4) 能提供证据辅助测试者实施推理结果的有效性判定,或自动判定。
k) 在提前获得测试项目授权后,被测者可在测试期内的任意时间发起测试。
l) 可为不同测试项维护独立的推理结果目录。
m) 可实现本地测试(测试者不介入的测试,如预测试或系统调试等)和网络测试(测试者介入)。

7.7.2 公平性保障要求

测试系统应提供方案和实现 6.7.2 的要求。

附 录 A
（资料性）
人工智能服务器系统性能测试工具示例

人工智能服务器系统性能测试工具套件（如，AISBench 等）是人工智能服务器系统性能测试工具典型实现，适用于人工智能服务器、人工智能服务器集群或人工智能计算中心的性能测试，能兼容主流人工智能加速器（如 CPU、GPU 或 NPU 等）和深度学习框架，使用流程见图 A.1。



标引符号说明：

a)——被测方确立测试项提交给测试方；

b)——测试方在本地测试数据库注册测试项，同时生成测试 ID；

c)——测试方返回每个测试项的配置文件，包含测试 ID 等，在正式测试脚本中使用测试 ID；

d)——被测方编辑和调试被测系统的测试代码，启动正式测试，测试工具监督测试，统计测试结果；

e)——被测方返回测试结果等信息；

f)——测试方在本地测试数据库写入测试结果等信息。

图 A.1 AISBench 使用流程

附 录 B
（规范性）
AUTOML 训练测试要求

B.1 训练要求

使用 AUTOML 实施训练时：

a） 整个训练过程中不应变更模型变异算法；

b） 应按 6.2b）～6.2d）的要求执行；

c） 应区分模型结构生成（变异）阶段和训练（针对某一代变异调整模型参数）阶段，至少在模型结构生成（变异）过程执行前后和训练开始前后，分别记录时点；

d） 模型变异和搜索空间应是有限并确定的。

B.2 训练结果日志要求

AUTOML 训练结果日志符合以下要求。

a） 日志按每次模型变异及对应训练过程输出，每次模型变异后，输出变异信息，格式为[yyyy：MM：dd HH：mm：ss]-[yyyy：MM：dd HH：mm：ss]-[generation_number]-[number_of_neuron s]，其中：

1） 第一项为变异开始时间；

2） 第二项为变异完成时间；

3） 第三项为变异代次计数；

4） 第四项为当前变异结果模型的神经元数（对初始化模型的训练，变异起止时间为空，代次记为 0）。

b） 对变异后模型的训练，日志按每个训期输出，格式应符合 6.3c）4）的要求。

附 录 C
（规范性）
测试代码公开规则

C.1 通则

测试代码公开，应按以下规则执行。

a） 测试代码公开流程，包含：
   1） 公开条件检查：测试者确认测试结果有效性，应符合 6.3 和 7.3 的要求；
   2） 公开协议检查：代码公开前，按协议检查并实施公开事项；未签署协议的，按本文件的规定实施；
      注：测试者与被测者通常在测试前或测试后签订的代码公开协议。
   3） 代码公开：公布于测试者与被测者商定的场所或互联网地址；
   4） 结束公开：在规定的公开周期后，结束公开，原场所或互联网地址，代码将不可访问；代码结束公开时，代码公开协议即告结束。

b） 公开义务：
   1） 测试代码可向测试者公开，具备合法访问权限时，应能浏览和下载；
   2） 被测者不负责向测试者及组织成员之外的机构、团体、企业和个人解释代码原理或实施结果复现事项；
   3） 已达成协议的不公开部分，不应公开；
   4） 测试代码可不公开被测者私有的工具源码（如模型格式转化或部署），该源码功能不含 a）提出的项目；可不公开被测者使用的公共网络可见的程序源码，但需在测试代码中注明（如：//ref:[源码包名_版本，地址]）。

C.2 训练测试代码公开规则

训练测试代码公开，应在符合 C.1 规定的基础上，包含以下功能的实现：

a） 网络构造；
b） 测试工具函数调用（含指标计算、计时、日志、测试起止和校验等）；
c） 日志生成；
d） 训练数据获取；
e） 训练数据读入；
f） 训练数据预处理；
g） 训练启动过程（含学习率调整）；
h） 训练过程[含训期循环、损失函数调用、精度转化（如实施）、模型和数据（在被测系统内）传输指令等]；
i） 配置文件；
j） 模型保存。

C.3 推理测试代码公开规则

推理测试代码公开，应在符合 C.1 规定的基础上，包含以下功能实现代码：

a） 测试工具约定的待实现部分（如功耗计量、数据提供和结果取出等）；
b） 测试工具函数调用（含作业到达模式、计时、日志、测试起止和校验等）；

c) 计算和存储资源管理(如资源申请和释放);
d) 推理过程;
e) 日志生成;
f) 测试集获取;
g) 数据预处理(如实施了预处理);
h) 数据后处理(如实施了后处理);
i) 配置文件;
j) 模型格式转化(至少应公开调用语句);
k) 模型部署(至少应公开调用语句);
l) 推理结果保存。

附 录 D
（资料性）
测试场景类型说明

D.1 图像识别

图像识别是利用计算机处理、分析和理解图像的过程，以识别图像中的目标和对象。图像识别过程的输入一般是特定格式的图像，输出可包含图像的类别（假设已有预先定义的类别集合），特性（如物体的颜色，人的性别、年龄等）或其他业务逻辑所关心的信息。图像识别广泛应用于各类视觉系统（如安检、工业制造流水线、农业养殖、电力巡检、医疗诊断等）。人工智能服务器系统对图像识别过程的加速能力，可提高视觉系统的应用效率。

D.2 物体检测

物体检测是计算机对给定的图片或视频帧，自动识别已知物体并标识物体在图像中的位置（一般使用矩形框和坐标）的过程。物体检测的输入一般是特定格式的图像或视频帧，输出可为已知物体位置信息。物体检测广泛应用于各类视觉系统（如交通、航空拍摄、分拣流水线等）。人工智能服务器系统对物体检测过程的加速能力，可提高视觉系统检测相关应用的效率。

D.3 语义分割

语义分割结合了图像分类、目标检测和图像分割，即将图像分割为具有特定语义的区域，并识别每个区域的内容或类别，最终获得具有逐像素语义标注的分割图像的过程。语义分割广泛地应用于自动驾驶、地质检测、测绘、医学影像病灶分离、服饰辅助设计和农业养殖等系统。人工智能服务器系统对语义分割过程的加速能力，可提高视觉检测系统和自动分析相关应用的效率。

D.4 推荐

推荐是利用计算机对特定数据集合搜索及结果排序的过程。推荐的输入一般为特定格式的查询条件或关键字，输出为有序的结果集合。推荐广泛地应用于各类电子系统（如电子商务、搜索引擎、营销、辅助设计、医疗处置方案辅助等）。人工智能服务器系统对推荐过程的加速能力，可提高相关应用的效率。

D.5 自然语言处理

自然语言处理是以人类语言为对象，利用计算机技术分析、理解和处理自然语言的过程，可分为自然语言理解及自然语言生成两类。生成式自然语言处理在问答和对话等场景下被广泛应用。自然语言处理广泛地应用在机器翻译、语言数据挖掘、搜索引擎等系统。其中，机器翻译是利用规则、统计或神经网络，将一种自然语言（源语言）的文本转换成另一种语言（目标语言）的过程。人工智能服务器系统对自然语言处理过程的加速能力，可提高翻译质量和语言数据挖掘等系统的效率。

D.6 语音识别

语音识别（或称“自动语音识别”）是用计算机将人类自然语言的语音内容转换为相应文字的过程。语音识别技术广泛应用在语音拨号、语音导航、室内设备控制、语音文档检索、语音听写等系统。语音识别技术，在各行业的应用已较为普遍（如金融领域智能客服、语音导航、智能终端语音输入等）。人工智能服务器系统对语音识别过程的加速能力，可提高语音控制、检索、听写等系统的效率。

D.7 光学字符识别

光学字符识别是指对文本资料的图像文件进行分析、识别、获取文字或版面信息的过程。光学字符识别的输入可为带有特定文字及布局信息的图像，输出可为图片上的文字内容或(用户定义的)布局表示。光学字符识别广泛地应用在各行业业务系统中(如金融智能终端证照识别或交通路牌内容识别等)。人工智能服务器系统对光学字符识别过程的加速能力，可提高各行业文字识别应用的效率。

D.8 人脸识别

人脸识别是用计算机系统从人脸图像，基于人脸部特征，识别人物身份的技术。人脸识别技术广泛地应用在各行业业务系统中，如通行、安检等身份核验子系统。人工智能服务器系统对人脸识别过程的加速能力，可提高各行业身份核验子系统的运行效率。

D.9 多模态

多模态能处理图形图像和文本信息，并根据提示或要求生成相应图形图像或文本。这类模型通常基于深度学习技术，并融合了计算机视觉和自然语言处理的方法。它们能理解输入的文本描述，并根据描述生成对应的图像或视频；或者根据输入的图像或视频生成相关的文字描述。

附 录 E
（资料性）
能效及效率指标项和测试方法

E.1 训练

E.1.1 训练能效

训练能效是指人工智能服务器系统单位功耗处理的训练数据量，单位为兆字节每焦耳(MB/J)或兆字节每千瓦时[MB/(kWh)]。训练过程能效指标项和测试方法，见表 E.1。

表 E.1 训练过程能效指标项和测试方法

指标项	测试方法	说明
人工智能服务器训练能效(E_S)	a) 对任意一次训练，测试每个训期的平均功率 P_{EP}； b) 测试每个训期的平均用时 T_{EP}； c) 计算人工智能服务器训练能效 E_S： $$E_S=\frac{\text{sizeof(训练集)}}{(T_{EP}\times P_{EP})}$$ 式中： sizeof(训练集)——训练集大小，单位为兆字节(MB)	人工智能服务器单位时间内消耗单位功耗消化的训练数据量
人工智能服务器集群训练能效(E_{CL})	a) 算出每台人工智能服务器，每训期的平均功率 $P_{EP\text{-}i}$(i 为正整数)和用时 T_{EP}； b) 计算人工智能服务器集群训练能效 E_{CL}： $$E_{CL}=\frac{\text{sizeof(训练集)}}{\sum_i(T_{EP-i}\times P_{EP-i})}$$	人工智能服务器集群单位时间内消耗单位功耗消化的训练数据量

E.1.2 训练效率

训练效率是指人工智能服务器系统训练得到某模型，其预测准确率与训练代价的比值，单位为每千瓦时[1/(kWh)]。训练过程效率指标项和测试方法，见表 E.2。

表 E.2 训练过程效率指标项和测试方法

指标项	测试方法	说明
人工智能服务器训练效率	a) 训练结束时，记录模型在测试集上的实际准确率(具体指标见 6.4.1 和 6.4.2 中各类任务的门限值)； b) 记录 T_{TR} 时间内的实际能耗 P_S； c) 计算训练效率： $$\frac{\text{准确率指标值}}{P_S}$$	人工智能服务器训练得到某模型，其实际判别准确率[a]与训练能耗[b]的比值[c]

表 E.2　训练过程效率指标项和测试方法（续）

指标项	测试方法	说明
人工智能服务器集群训练效率	a）　训练结束时，记录模型在测试集上的实际准确率（具体指标见6.4.1 和 6.4.2 中各类任务的门限值）； b）　记录 T_{TR} 时间内的集群实际能耗 P_S； c）　计算训练效率： $\frac{\text{准确率指标值}}{P_S}$	人工智能服务器集群训练得到某模型，其实际判别准确率[a]与训练能耗[b]的比值[c]
注 1：训练效率定义见 GB/T 17166—2019。 注 2：人工智能服务器集群训练效率，见 GB/T 25000.22—2019 中的能源效率。		
[a] 当准确率指标（设值为 a，$0 \leqslant a \leqslant 1$）为负向指标时（如 WER），以（$1-a$）计。 [b] 训练能耗是训练模型过程中特定时间段内消耗的电量。 [c] 如实施多次训练试验，则应使用平均准确率及平均训练时长。		

E.2　推理

E.2.1　推理能效比

推理能效比以额定工作情况下能效比计算。推理过程能效比指标项和测试方法，见表 E.3。

表 E.3　推理过程能效比指标项和测试方法

指标项	测试方法	说明
视觉任务能效比	a）　被测者在整个推理测试过程中（T_I 内），在每个端到端推理时间内，周期性读取功率计测试值，求出平均功率 $\overline{P_I}$； b）　测试者累计返回结果的任务图像（帧）数 N； c）　测试者累计实际分派处理延时总覆盖时间 T_{DIP}； d）　计算视觉任务能效比： $\frac{\frac{N}{T_{DIP}}}{\overline{P_I}}$	单位为消耗每焦耳或每千瓦时能量处理的图像（帧）数［图像（帧）数/J］或［图像（帧）数/kWh］
自然语言任务能效比	a）　在整个推理测试过程中（T_I 内），在每个端到端推理时间内，周期性读取功率计测试值，求出平均功率 $\overline{P_I}$； b）　累计返回结果的单词数 W； c）　累计实际分派处理延时总覆盖时间 T_{DIP}； d）　计算自然语言处理任务能效比： $\frac{\frac{W}{T_{DIP}}}{\overline{P_I}}$	单位为消耗每焦耳或每千瓦时能量处理的单词数［单词数/J］或［单词数/kWh］

表 E.3　推理过程能效比指标项和测试方法（续）

指标项	测试方法	说明
语音任务能效比	a）　在整个推理测试过程中（T_I内），在每个端到端推理时间内，周期性读取功率计测试值，求出平均功率 $\overline{P_I}$； b）　累计返回结果的句子数 S； c）　累计实际分派处理延时总覆盖时间 T_{DIP}； d）　计算语音任务能效比 $\frac{S}{\frac{T_{DIP}}{\overline{P_I}}}$	单位为消耗每焦耳或千瓦时能量处理的句数[句数/J]或[句数/kWh]
行业任务能效比	按视觉任务能效比或自然语言任务能效比测试方法执行	按视觉或自然语言任务能效比计算

E.2.2　推理效率

推理效率是人工智能服务器系统完成推理任务与代价的比值，单位为每秒千瓦时[1/(kWh)]。推理过程效率指标项和测试方法，见表 E.4。

表 E.4　推理过程效率指标项和测试方法

指标项	测试方法	说明
人工智能服务器推理效率	a）　推理结束时，记录模型在测试集上的实际准确率（具体指标见 7.4.5 和 7.4.6 中各类任务的作业到达模式及参数）； b）　记录 T_{DIP}内的实际能耗 P_S； c）　计算推理效率： $\frac{\text{准确率指标值}}{P_S}$	人工智能服务器实际推理准确率[a]与推理能耗[b]的比值
人工智能服务器集群推理效率	a）　推理结束时，记录模型在测试集上的实际准确率（具体指标见 7.4.5 和 7.4.6 中各类任务的作业到达模式及参数）； b）　记录 T_{DIP}内的集群实际能耗 P_S； c）　计算推理效率： $\frac{\text{准确率指标值}}{P_S}$	人工智能服务器集群实际推理准确率[a]与推理能耗[b]的比值
注：人工智能服务器和集群推理效率的定义，分别见 GB/T 17166—2019 和 GB/T 25000.22—2019。		
[a] 当准确率指标（设值为 a，$a \geqslant 0$ 且 $a \leqslant 1$）为负向指标（如 WER）时，则以（$1-a$）计。 [b] 推理能耗是推理过程中特定时间段内消耗的电量。		

参 考 文 献

[1] GB/T 5271.34—2006 信息技术 词汇 第34部分:人工智能 神经网络

[2] GB/T 17166—2019 能源审计技术通则

[3] GB/T 22454—2008 企业集成 企业建模构件

[4] GB/T 25000.22—2019 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第22部分:使用质量测量

[5] ISO 10303-34:2001 Industrial automation systems and integration—Product data representation and exchange—Part 34: Conformance testing methodology and framework: Abstract test methods for application protocol implementations

[6] ISO 23952:2020 Automation systems and integration—Quality information framework (QIF)—An integrated model for manufacturing quality information

[7] ISO 19440:2020 Enterprise modelling and architecture—Constructs for enterprise modelling

[8] ISO/IEC 14776-414:2009 Information technology—Small Computer System Interface(SCSI)—Part 414:SCSI Architecture Model-4(SAM-4)

[9] ISO/IEC 22989:2022 Information technology—Artificial intelligence—Artificial intelligence concepts and terminology