ICS 25.040
CCS N 10

# T/CAMETA

## 中国机电一体化技术应用协会团体标准

T/CAMETA xxxxxx—2025

# 工业智能模型测试与评价

Industrial Intelligent Models—Testing and Evaluation

（征求意见稿）

2025 — xx— xx 发布　　2025 — xx — xx 实施

中国机电一体化技术应用协会　　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。

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这些专利的责任。

本文件由中国工业互联网研究院提出。

本文件由中国机电一体化技术协会归口。

本文件起草单位：中国工业互联网研究院、XXX等

本文件主要起草人：

# 工业智能模型测试与评价

## 1 范围

本标准规定了工业智能模型，特别是基础算法与工业机理模型，在工业应用场景下的测试流程、测试方法与性能评价指标，适用于模型功能验证、性能分析、可靠性评估及鲁棒性测试等多个阶段。本标准适用于部署于工业互联网平台中面向设备预测维护、质量评估、流程优化等任务的模型系统，包括部署于云端、边缘或端侧的多类型算法模型。

## 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 44067.1-2024《工业互联网平台 技术要求及测试方法 第1部分：总则》

AII/001-2019《工业互联网平台测试验证》

AII/015-2022《工业终端设备信息模型测试规范》

T/CI 093-2023《公共基础服务业云边协同技术标准 第1部分：总则》

T/CI 094-2023《公共基础服务业云边协同技术标准 第2部分：安全协同》

T/CI 095-2023《公共基础服务业云边协同技术标准 第3部分：数据协同》

T/CI 096-2023《公共基础服务业云边协同技术标准 第4部分：算力协同》

## 3 术语和定义

GB/T 44067.1-2024界定的以及下列术语和定义适用于本文件。

### 3.1 工业智能模型

指在工业互联网平台中用于实现数据驱动的感知、分析、预测、优化或控制任务的人工智能模型，包括基于深度学习、图神经网络、时序分析、物理建模等方法构建的模型。

### 3.2 基础算法

指具备通用功能性、可复用性强的核心计算逻辑或神经网络模块，常作为工业智能模型中的通用功能组件，例如分类器、回归器、编码器、注意力模块等。

### 3.3 工业机理模型

指依据工业设备运行物理规律或工艺特征，采用微分方程、传递函数或混合建模方式所构建的解释性强、适应性高的模型，通常用于仿真、预测和诊断。

### 3.4 测试验证

指对工业智能模型的结构配置、功能表现、性能参数及鲁棒性进行系统性测试与验证的过程。

### 3.5 鲁棒性

指模型在输入扰动、场景变化或部分信息缺失情况下依然保持性能稳定的能力。

## 4 系统组成与测试环境

### 4.1 系统组成

模型测试系统包括如下部分：

(1) 云端服务器平台（用于集中模型训练、统一管理、任务调度）；

(2) 边缘计算节点（用于快速响应推理、部分本地更新）；

(3) 工业设备端（传感器接入、原始数据采集）；

系统部署参考如下：云端具备高性能GPU节点（如NVIDIA AH100集群），边缘节点采用Jetson Orin NX/Orin Nano等设备，设备端支持多协议工业网关。

### 4.2 软件与通信配置

云侧运行OS等系统，宜支持Kubernetes+Docker容器编排，边缘节点可启用JetPack SDK与ONNX Runtime。通信协议可采用MQTT+TLS或gRPC，数据格式可标准化为JSON/ProtoBuf结构。

### 4.3 数据准备

用于模型测试的数据集需满足多样性、可重复性和可追溯性，宜使用以下公开或工业采集数据，如：

(1) CWRU轴承故障数据集（时序数据）；

(2) SECOM半导体过程监控数据（多变量质量数据）；

(3) Real-IAD缺陷检测图像集（图像数据）。

每个数据样本宜包含：采集时间戳、传感器ID、原始测量值、设备工况标签等字段，确保输入输出对齐并可用于评估分析。

## 5 测试流程与方法

工业智能模型的测试流程宜覆盖模型开发周期的全生命周期，包括部署验证、功能测试、性能评估、鲁棒性验证、模型可解释性检查等环节。

### 5.1 模型部署一致性验证

在模型部署阶段，宜验证以下内容：

(1) 模型结构完整性：导出结构是否与训练结构一致；

(2) 权重文件匹配性：参数是否正确加载；

(3) 接口兼容性：部署环境支持模型调用的推理接口，如ONNX或TensorRT；

(4) 容器化部署验证：镜像封装是否稳定，是否支持在Docker/Kubernetes环境中无误运行。

### 5.2 功能性测试

功能测试目标在于验证模型是否完成预期任务输出。对于分类/识别/预测类模型，宜进行以下内容测试：

(1) 模型输入输出数据格式正确性；
(2) 输出值范围是否合理，如分类标签是否出现在候选集中；
(3) 功能逻辑正确性，如预测结果是否随输入变化合理波动；
(4) 批处理能力，如正常运行中能支持的最大并发数。

### 5.3 性能评估方法

性能测试应包括运行效率和资源消耗两方面：
(1) 推理时间：以毫秒为单位，记录每轮推理所需时间；
(2) 吞吐量：单位时间内可处理的数据样本数；
(3) CPU、GPU、内存占用：资源利用率宜保持在指定阈值内；
(4) 能耗测量：通过外接测量模块记录运行过程中的功耗波动。
测试前需进行硬件预热，例如进行多轮前向计算，使计算硬件提前分配显存或内存，使计算核心进入稳定工作状态。推理时间和吞吐量宜使用多轮推理的平均值。

### 5.4 鲁棒性与错误处理测试

鲁棒性是工业智能模型的重要质量属性，宜通过以下方式验证：
(1) 多源异构数据兼容性测试：验证模型对不同采样率、缺失率、噪声类型的适应能力；
(2) 输入扰动测试：引入不同程度的随机噪声，观察准确率变化；
(3) 异常输入：测试非法数据或缺失数据时系统处理方式，确保不崩溃且给出错误提示。

### 5.5 模型可解释性测试

宜对模型的重要决策路径进行可视化分析，可采用如下方式之一：
(1) CAM（Class Activation Mapping）热力图显示决策区域；
(2) Attention可视化展示特征聚焦；
(3) 中间层输出分析：逐层跟踪神经网络中的特征演化。

## 6 测试任务案例

测试任务案例设计宜覆盖工业智能模型在典型工业场景中的关键应用，包括分类、检测、预测等常见任务。每个案例宜包含背景说明、模型目标、输入输出要求、测试步骤、评价指标与期望结果等。

### 6.1 设备智能故障诊断

**背景说明：**
现代制造中，设备的运行稳定性直接影响产线效率。通过部署故障诊断模型，可实现对关键部件（如轴承、电机、风扇等）的状态监测与早期故障预警。
**模型目标：**
根据振动信号、多传感器时序数据，识别设备是否存在异常，判断故障类型与位置。
**输入数据格式：**
多通道时序振动信号，样本长度固定（如2048点），附带工况标签。
**输出数据格式：**
故障类型分类结果（多类），如：正常、内圈故障、外圈故障。
**测试流程：**
(1) 使用CWRU轴承数据集或自采集传感器数据进行样本预处理；

(2) 在Jetson Orin NX平台加载模型执行推理；
(3) 比较不同工况下模型准确率与召回率变化；
(4) 模拟边缘断电场景，测试模型断点恢复与状态保持能力。

**评价指标：**

(1) 分类准确率；
(2) 单样本推理延迟；
(3) 异常样本召回率；
(4) 稳态能耗变化。

### 6.2 工业表面缺陷检测

**背景说明：**

表面缺陷是影响产品质量的重要因素，如铝型材、钢板、玻璃表面等存在划痕、裂纹、斑点等缺陷，传统人工检测效率低且易漏检。

**模型目标：**

对图像中存在的缺陷进行检测、定位并分类。

**输入数据格式：**

RGB图像，分辨率统一（如512×512），图像内可能包含多种、多个缺陷目标。

**输出数据格式：**

1）缺陷类别、位置边界框（bbox）及置信度分数，或2）缺陷与背景的像素分割图。

**测试流程：**

(1) 使用Real-IAD数据集或工业产线上图像数据作为输入；
(2) 模型部署在云端训练，边缘推理进行实时判断；
(3) 测试不同瑕疵类型下模型检测率与误检率；
(4) 统计批量推理的时间成本与系统资源利用率。

**评价指标：**

(1) mAP（平均精度）或mIoU（平均交并比）；
(2) 推理帧率；
(3) 错误率；
(4) GPU利用率。

### 6.3 能耗预测与异常能耗识别

**背景说明：**

在大型工业园区或智能工厂中，能耗波动可能暗示潜在故障或系统调度异常，通过建模能耗行为实现预测和异常识别，可提高能源管理效率。

**模型目标：**

对未来一段时间的电力/气体消耗趋势进行预测，并标注异常点。

**输入数据格式：**

时间戳、用电量、设备状态、环境变量（如温湿度等）。

**输出数据格式：**

多步预测值序列与异常评分列表。

**测试流程：**

(1) 构建以滑动窗口方式生成的训练与测试样本；
(2) 部署Transformer时序预测模型；
(3) 人为植入异常点，测试模型灵敏度与响应能力；

(4) 比较模型在不同部署位置（云/边）的延迟与负载表现。

**评价指标：**

(1) 平均绝对误差；

(2) 异常检测F1值；

(3) 推理延迟；

(4) 资源使用率。

## 7 数据输入输出规范

本标准推荐统一的数据结构，提升模型部署与平台对接的一致性。数据字段宜遵循下列规范：

### 7.1 输入数据字段说明（通用格式）

| 字段名称 | 类型 | 描述 | 示例值 |
|---|---|---|---|
| sensor_id | String | 采集传感器唯一编号 | "TEMP_SENSOR_01" |
| timestamp | Datetime | 数据采集时间（UTC） | "2025-06-17T12:30:00" |
| value | Float | 采集值 | 32.5 |
| unit | String | 单位 | "℃" |
| status_code | Integer | 状态码（0正常，1异常） | 0 |

### 7.2 输出数据字段说明（按模型类型分类）

#### 7.2.1 分类模型输出

| 字段名称 | 类型 | 描述 |
|---|---|---|
| class_label | String | 分类结果，如“正常” |
| confidence | Float | 分类置信度（0~1） |

#### 7.2.2 回归/预测模型输出

| 字段名称 | 类型 | 描述 |
|---|---|---|
| predict_value | Float | 预测值 |
| std_dev | Float | 标准差（可选） |
| timestamp_pred | Datetime | 对应预测时间 |

#### 7.2.3 检测模型输出

| 字段名称 | 类型 | 描述 |
|---|---|---|
| defect_type | String | 缺陷类别，如“划伤” |

| | | |
|---|---|---|
| bbox | List | 位置坐标[x, y, w, h] |
| score | Float | 置信度分数（0~1） |

## 8 性能评价指标

性能评价指标是对工业智能模型实际部署效果的综合衡量。需综合考虑模型准确性、效率、资源使用率及部署适应性等多维度指标。

### 8.1 功能类指标

适用于分类、检测、回归、预测等类型模型的性能度量。

(1) 准确率（Accuracy）：分类任务中模型正确预测结果占比：

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN}$$

(2) 召回率（Recall）与精确率（Precision）：适用于不平衡样本情形下的分类/检测模型：

$$Recall = \frac{TP}{FN+TP}$$

$$Precision = \frac{TP}{FP+TP}$$

(3) F1 分数（F1-score）：综合考虑召回率与精确率，取其调和均值，适合用于异常检测类任务：

$$F1score = \frac{2TP}{2TP+FN+FP}$$

(4) 均方误差（MSE）、平均绝对误差（MAE）：适用于预测类模型衡量数值误差：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|$$

(5) 全类平均正确率（mAP）：目标检测模型在多类别下的综合识别能力：

$$MAP = \frac{1}{C}\int_0^1 P_c(r)\mathrm{d}r$$

(6)全类平均交并比（mIoU）：图像分割模型在多类别下的综合解析能力：

$$mIoU = \frac{1}{K}\sum_{i=0}^{K}\frac{TP}{TP + FP + FN}$$

### 8.2 系统性能指标

衡量模型在不同运行环境下的效率表现：

(1) 推理时延（Latency）：模型从接收输入到输出预测所需的平均时间（ms）；
(2) 吞吐量（Throughput）：单位时间内可处理的任务数量（如样本数/秒）；
(3) 并发响应能力：在高并发输入请求下，模型的稳定响应能力；
(4) 启动加载时间：模型初始化加载并准备好服务的时间。

### 8.3 资源消耗指标

(1) CPU/GPU利用率：模型在运行过程中计算资源占用比例，需保持在配置阈值内；
(2) 内存占用：模型运行时的平均与峰值内存使用量；
(3) 功耗变化：在边缘设备等对能耗敏感场景下，统计模型推理前后的瞬时功率变化，单位W；

### 8.4 鲁棒性与可恢复性指标

(1) 输入扰动鲁棒性：微小扰动（如加噪、漂移）对模型输出影响大小；
(2) 异常数据容错能力：非法输入、缺失值等场景下，模型是否崩溃及处理机制；
(3) 模型断点续推能力：在中断后是否能自动恢复，且结果无重大偏差。

### 8.5 工业适应性与部署友好性

(1) 跨平台部署能力：支持主流边缘设备/服务器架构（如x86、ARM）；
(2) 多格式模型兼容性：支持ONNX、TorchScript、TensorRT、PMML等多种模型格式；
(3) 容器集成能力：可封装为Docker镜像运行，支持CI/CD集成部署；
(4) 日志可追溯性：模型日志具备唯一任务编号、时间戳、输入输出摘要等，便于追责审计。

## 9 通用要求与合规性建议

工业智能模型在测试验证阶段，除性能与功能指标外，还宜考虑以下合规性与标准化要求。

### 9.1 日志与数据记录规范

(1) 每次测试任务完整记录测试配置（模型版本、输入参数、运行平台）；
(2) 所有结果文件按统一命名规范保存，日志文件包含运行时间、设备ID、关键警告信息等；
(3) 建议采用结构化日志（如JSON格式）记录测试任务过程，便于后期分析与审计。

### 9.2 安全性建议

(1) 模型部署过程中验证软件包签名、镜像完整性；

(2) 推理过程中涉及的数据通信采用加密协议（如HTTPS、MQTT+TLS）；
(3) 模型服务设置访问鉴权，禁止匿名调用接口；

### 9.3 可复现性要求

(1) 模型测试用例及输入样本可复现，包括随机种子设置、初始化参数、版本号；
(2) 推荐使用固定版本的依赖库及容器封装，确保跨测试节点一致结果；
(3) 对关键流程如预处理、模型后处理等附说明文档。

# 《工业智能模型测试与评价》

## 编制说明

2025 年 10 月

# 一 工作简况

## （一） 任务来源

随着我国工业互联网与智能制造的深入发展，工业领域中人工智能模型在设备预测维护、质量检测、能耗优化等环节的应用不断深化。然而，当前行业缺乏统一的模型测试与评价标准，导致算法性能指标不一致、测试流程不规范、结果复现性差等问题，影响了工业智能技术的推广与落地。

为贯彻落实《国家标准化发展纲要》和《“十四五”智能制造发展规划》，中国机电一体化技术应用协会于 2025 年正式立项《工业智能模型测试与评价》团体标准编制任务。本标准由中国工业互联网研究院牵头，联合相关科研院所、行业协会及企业共同制定，旨在规范工业智能模型的测试流程、方法与评价体系，提升工业智能模型在不同部署环境下的可靠性与可比性。

任务来源具体包括以下三方面：

**国家战略需求：**《中国制造 2025》、《“十四五”智能制造发展规划》、《国家标准化发展纲要》均明确提出，要加快智能制造核心技术标准体系建设，推动工业互联网与人工智能的深度融合。工业智能模型是实现制造业智能决策和自适应优化的关键基础，但目前缺乏统一的性能验证与测试评价体系，难以满足国家对智能制造系统安全、可靠、可控的要求。制定本标准，有助于完善我国工业智能标准体系，支撑制造业数字化、网络化、智能化转型目标的落实。

**行业迫切需求：**当前工业领域部署的大量人工智能模型存在“黑箱化”问题，模型性能、鲁棒性与能耗指标难以量化比较，导致企业间算法能力差距大、部署效果不透明。特别是在设备故障诊断、缺陷检测、能耗预测等典型场景中，模型测试结果缺乏统一标准，行业间测试方法与评价指标不兼容，严重影响模型的迁移与复用。企业迫切需要一套覆盖测试流程、数据规范、性能指标和合规性建议的系统性标准，以实现模型测试的可复现、可对比、可追溯，推动算法能力在工业环境中的可靠落地。

**技术创新推动：**近年来，国内在深度学习、图神经网络、时序预测、物理机理建模等方向取得显著进展，并形成了多种融合算法与多层部署架构。然而，这些创新技术尚未形成统一的测试与评价方法体系，缺乏对新型模型性能、鲁棒性及工业适应性的标准化验证手段。通过制定《工业智能模型测试与评价》标准，可将行业成熟经验与前沿研究成果转化为通用规范，推动技术创新成果的标准化固化与共享应用，加快工业智能模型从实验室

走向工程化与规模化部署。

## （二） 国内关于工业智能模型测试与评价的标准制定情况及最新要求

目前国内外在工业智能模型测试与评价领域的标准体系尚处于起步阶段。虽然在工业互联网平台、设备信息模型及云边协同等方面已有若干相关标准，但针对人工智能模型在工业场景下的功能验证、性能测试、鲁棒性评估与可解释性分析等关键环节，尚缺乏系统性、可操作的技术规范。为此，本标准的制定旨在填补工业智能模型测试与评价方面的标准空白，构建支撑工业智能化发展的技术基础体系。相关标准如下：

GB/T 44067.1-2024《工业互联网平台 技术要求及测试方法 第 1 部分：总则》

AII/001-2019《工业互联网平台测试验证》

AII/015-2022《工业终端设备信息模型测试规范》

T/CI 093-2023《公共基础服务业云边协同技术标准 第 1 部分：总则》

T/CI 094-2023《公共基础服务业云边协同技术标准 第 2 部分：安全协同》

T/CI 095-2023《公共基础服务业云边协同技术标准 第 3 部分：数据协同》

T/CI 096-2023《公共基础服务业云边协同技术标准 第 4 部分：算力协同》

综上可见，现有标准体系主要聚焦于工业互联网平台的总体架构和通信、数据、安全、算力等要素，对工业智能模型测试与评价这一关键环节仅作原则性描述，缺乏统一的测试流程、指标体系与性能评估方法。国内尚未建立覆盖模型部署一致性验证、功能测试、性能与能耗评估、鲁棒性验证及可解释性分析等内容的标准化技术体系，亟需通过标准制定实现系统性突破与方法规范化。

## （三） 标准编制的目的、意义

《工业智能模型测试与评价》标准的编制，旨在通过系统化、科学化的技术规范，填补我国在工业智能模型测试与评价领域的标准化空白，推动工业人工智能技术在生产制造、设备管理、能耗优化等场景中的安全、可靠、可控应用，促进智能制造的高质量发展。

当前，随着人工智能技术在工业互联网平台中的广泛应用，工业智能模型已成为支撑设备预测维护、产品质量检测、工艺优化与能源管理的重要技术基础。然而，由于缺乏统一的测试与评价标准，行业内普遍存在测试流程不规范、性能评价指标不一致、结果复现性差等问题，导致模型部署难以规模化推广，算法能力难以客观比较，严重制约了工业智能技术体系的完善与产业化进程。

本标准的制定，首先着眼于解决工业智能模型应用中的规范性难题，通过明确测试流程、数据输入输出规范、性能评价指标、鲁棒性与可解释性测试方法等核心内容，为企业

与科研机构提供系统、可操作的技术依据，提升模型验证的可靠性与一致性，减少重复验证与资源浪费。其次，标准将前沿研究成果与工程实践经验相结合，将模型全生命周期测试、云边端协同验证、能耗与算力评估等创新方向纳入标准框架，引导行业从经验驱动向科学化、体系化的测试验证模式转变，为工业人工智能的持续创新与工程落地提供制度保障。此外，标准注重与国际先进标准体系的对接，通过引入 ISO/IEC、IEEE、NIST 等组织的相关测试理念，建立与国际接轨的工业智能模型评价体系，提升我国在工业智能领域的国际话语权与竞争力。

从行业生态层面看，本标准的实施将促进产学研用深度融合，加速人工智能技术成果在工业制造中的转化与落地；同时，通过规范模型测试管理、性能评估方法及平台适配要求，推动工业智能技术体系的整体升级，助力我国制造业实现智能化、绿色化、可持续化发展。长远来看，该标准不仅为工业智能模型的高质量应用提供了技术支撑，更将通过标准化手段提升我国工业智能领域的技术自主可控水平，为建设制造强国和数字中国提供坚实基础。

编制本标准的意义非常重大，原因在于：（1）工业智能模型测试与评价领域亟需统一标准指导，以实现模型验证与性能评估的科学化和规范化；（2）目前国内尚无系统覆盖模型功能、性能、鲁棒性与可解释性评价的标准体系，制定本标准对促进工业人工智能技术可信落地具有重要现实意义。

**（四） 标准特点**

本标准主要涵盖工业智能模型测试与评价的术语与定义、系统组成与测试环境、测试流程与方法、典型任务案例、数据输入输出规范、性能评价指标及通用合规性要求等内容。该标准以工业智能模型在云、边、端多层部署环境下的测试需求为导向，系统构建了涵盖功能验证、性能评估、鲁棒性检测和可解释性分析的完整测试体系，突出系统性与可操作性。标准明确了模型输入输出结构、评价指标计算方法与测试过程规范，确保不同类型模型在不同工业场景下具有可比性和可复现性。同时，本标准注重测试过程的安全性与合规性要求，强化数据安全、通信加密与日志追溯机制，确保模型验证过程符合国家相关法规。该标准的制定填补了工业智能模型测试与评价领域的标准化空白，为工业人工智能模型的验证、评估与应用提供了统一依据，旨在推动工业智能模型测试工作的规范化、体系化和科学化，提升我国智能制造技术的整体水平与国际竞争力。

**（五）主要工作过程**

1. 编制准备阶段

2025 年 3 月-9 月。主编单位接到编制任务后，组织专业技术人员成立编制组，开展大量的资料收集和前期调研工作，编写完成标准大纲、标准初稿等。

2.征求意见阶段

2025 年 10 月完成标准草案的完善，并小范围内部征求意见，根据反馈意见修改形成征求意见稿，全面公开征求意见。

3.送审阶段

将接受专家审查，并根据专家审查意见修改送审稿，最终形成报批稿。

4.报批阶段

准备报批。

## 二 标准编制原则

（一）科学性原则：本标准的编制以工业人工智能模型的理论研究成果与工程实践经验为基础，充分考虑不同模型类型的技术特征，确保标准中的测试流程、性能指标和评价方法具有科学性和可验证性，能够真实反映模型在工业场景中的运行性能与可靠性。

（二）统一性原则：本标准的编制充分结合现行的工业互联网、云边协同及信息模型相关标准，统一了模型测试的基本要求、数据格式、评价指标及结果表达方式，确保不同企业、平台和研究机构在工业智能模型测试与评价过程中具备统一的技术依据和对比标准。

（三）公正性原则：本标准编制过程坚持公开、公平、公正原则，广泛征求产业界、学术界和测试机构的意见，综合各方专家建议，确保标准内容的客观性与中立性，避免因利益差异造成标准偏向，保证标准的科学权威与行业公信力。

（四）可操作性原则：本标准在编制过程中注重测试方法与流程的工程可实施性，确保测试要求、环境配置及评价指标能够在实际工业场景中落地应用，避免标准过于理论化或脱离实际，使标准既具指导性又具可执行性。

（五）合规性原则：本标准严格遵循国家相关法律法规和强制性标准要求，确保模型测试过程中的数据安全、算法透明与可追溯性要求，保证标准的合法性、规范性与可持续适用性。

## 三 标准主要内容

1、 范围：说明本标准适用的对象、边界与应用场景，明确工业智能模型在设备预测维护、质量检测、能耗分析等工业任务中的测试与评价适用范围。

2、 规范性引用文件：列出本标准编制过程中所依据的国家标准、行业标准及相关团体标准文件，为标准的引用和执行提供依据。

3、 术语与定义：对本标准中涉及的核心术语（如工业智能模型、基础算法、工业机理模型、鲁棒性、可解释性等）进行统一定义，确保标准内容的一致性与可理解性。

4、 系统组成与测试环境：规定模型测试系统的组成，包括云端服务器平台、边缘计算节点和设备端的硬件与软件配置要求，明确测试环境的通信协议、数据格式及安全要求。

5、 测试流程与方法：规范工业智能模型从部署验证、功能测试、性能评估、鲁棒性验证到可解释性分析的全流程测试方法，明确各阶段的输入条件、测试步骤及输出结果要求。

6、 测试任务案例：给出典型工业应用案例，包括设备智能故障诊断、工业表面缺陷检测和能耗预测与异常识别等任务，阐述测试目标、数据格式、执行流程与评价指标，指导实际应用。

7、 数据输入输出规范：统一模型输入与输出数据的字段结构、类型与格式，确保不同平台与模型间的兼容性、可追溯性与可复现性。

8、 性能评价指标：规定功能类指标（如准确率、召回率、F1 分数）、系统性能指标（如时延、吞吐量、响应能力）、资源消耗指标（如 CPU/GPU 利用率、能耗变化）及鲁棒性指标等，构建多维度性能评价体系。

9、 通用要求与合规性建议：提出模型测试过程中日志记录、安全性、数据加密、访问鉴权、版本控制与结果复现性等方面的要求，保障测试过程规范、安全与透明。

10、 附录与参考信息：包括典型测试样例、指标计算示例及推荐测试工具说明，供标准执行单位参考使用。

## 四 预期经济效果

《工业智能模型测试与评价》标准的实施，预期将产生显著的经济和社会效益。当前，工业智能模型在制造业中的应用规模不断扩大，但由于缺乏统一的测试与评价体系，模型验证成本高、测试结果不可比、工程部署风险大，企业在研发和推广过程中往往面临重复验证、资源浪费及周期延长等问题。通过制定并实施本标准，可有效降低模型开发与验证的经济成本，提升研发效率与测试准确性。

标准的推广应用将为工业企业和科研机构提供统一的技术依据与评价框架，使模型测

试环节由经验驱动转向标准化、自动化。通过规范数据结构、测试流程和评价指标，可显著减少重复建设与试验开支，缩短模型部署周期，提升工业智能系统的整体开发效率。同时，统一的测试体系将促进不同算法和平台间的性能对比与互认，推动产业上下游的协同创新，减少企业在模型迁移和跨平台部署中的适配成本。

此外，本标准的实施将有助于建立工业智能模型的测试与认证机制，为模型产品化与市场化提供技术支撑，提升企业核心竞争力和品牌信誉度。长远来看，该标准的推广将带动智能制造装备、工业互联网平台及测试验证服务等相关产业的协同发展，形成明显的经济带动效应，为我国制造业的数字化、智能化、绿色化转型提供持续动力。

## 五 采用国际标准和国外先进标准情况

在《工业智能模型测试与评价》标准的编制过程中，编制组充分研究了国际上与人工智能系统评测、工业互联网平台测试及智能制造相关的标准和研究成果，重点参考了 ISO、IEC、IEEE、NIST 等国际组织发布的相关标准文件。例如，参考了 ISO/IEC 23053:2024《Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)》、ISO/IEC 25023《Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of system and software product quality》等内容。这些国际标准对人工智能系统的性能、可靠性、安全性及可解释性等方面提出了指导性要求，为本标准的框架设计和指标体系构建提供了重要参考。

在借鉴国外先进标准的基础上，本标准结合我国工业互联网与智能制造体系的实际发展特点，构建了以“统一流程、量化指标、可追溯验证、合规保障”为核心的工业智能模型测试与评价框架。通过与国际标准的对标与融合，本标准不仅能够满足国内企业在智能模型开发与应用中的实际需求，还将有助于推动我国工业智能模型测试体系与国际标准接轨，提升我国在工业人工智能标准化领域的国际影响力和话语权

## 六 与有关的现行法律、法规和强制性国家标准的关系

在编制工业智能模型测试与评价规范标准过程中，我们严格遵循了相关的现行法律、法规和强制性国家标准，确保标准的合规性和权威性。同时，我们也充分考虑了工业智能模型测试与评价规范标准的发展趋势和应用需求。

## 七 重大分歧意见的处理经过和依据

本标准在起草过程中未出现重大分歧意见。

## 八 标准性质的说明

建议本标准为推荐性标准。

## 九 有关专利的说明

本文件在制定过程中，未参考或引用任何专利技术，编制内容均为行业通用技术方法、测试流程及评价指标的总结与规范化，不涉及特定专利的实施或保护。若后续标准实施中发现与第三方专利存在关联的可能性，建议使用者在具体应用中自行评估专利合规性，本文件发布机构不承担识别相关专利的责任。

## 十 贯彻标准的要求和措施建议

本标准经征求各相关方意见，已形成共识，标准实施之日起，各相关方将遵照执行。

## 十一 废止现行有关标准的建议

无。