ICS 国际标准分类号
CCS 中国标准文献分类号

团　　体　　标　　准

T/CES XXX□XXXX

电力人工智能模型场景化验证及评价体系构建规范

A scenario model verification and application evaluation standard

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会　发布

目　　次

前　言

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。

本文件由国网信息通信产业集团有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：国网信息通信产业集团有限公司、北京国网信通埃森哲信息技术有限公司、国网湖北信通公司、国网河南省电力公司驻马店供电公司。

本文件主要起草人：李强、赵峰、赵林林、刘茂凯、许中平、谢可、罗弦、黄俊东、赵智勇、李卫军、王誉博、张朔、安丽丽、吴晓峰、邱镇、黄晓光、王兴涛、白景坡、李炳森等人。

本文件为首次发布。

1. 编制背景

电力人工智能模型场景化验证及评价体系的标准化构建，不仅可以有效地解决电力决策中的复杂性和主观性，还可以提高评价的可靠性和有效性，为实际问题的解决提供有益的参考。因此，该方法已经成为电力领域决策分析中的重要工具之一。可以帮助决策者更加科学、准确地进行决策，为电力领域的可持续发展做出贡献。

2. 编制主要原则

本文档主要遵循以下原则编制：

1) 目标明确：明确研究目标和应用场景，确保评价指标和层次结构的准确性和全面性。

2) 方法科学：遵循科学的、系统的和精细的方法，确保评价结果的可靠性和有效性。

3) 专家参与：加强专家经验的挖掘和应用，注重专家对指标的评价和比较，提高决策的合理性。

4) 数据准确：采集和使用可靠的数据，确保评价结果的准确性和客观性。

5) 统计分析：使用有效的数学模型和统计分析方法，进行数据处理和结果分析，得出科学、准确的决策建议。

6) 实用性强：考虑到决策者的需求和实际应用，确保评价结果的可操作性和实用性。

3. 与其他标准文件的关系

本文档与相关技术领域的国家现行法律、法规和政策保持一致。同时，本文档还与 ISO 9001（质量管理体系标准）、ISO 14001（环境管理体系标准）、ISO 50001（能源管理体系标准）等相关的国家标准和规范相一致。

4. 主要工作过程

本文档的编制借鉴了国家相关标准，着重结合平台实际操作流程和使用需求。编制过程主要包括：

1) 确定评价目标和内容:通过深入了解电力系统及其特点，明确电力系统优化和安全保障的目标和内容，为后续评价提供指导。

2）构建评价体系:构建评价体系是将评价目标进一步细化，根据电力系统的特点提出相应的评价指标，从而形成评价体系。

3）确定指标权重:利用专家主观判断或因子分析法等客观赋权方法，对评价体系中的各项指标进行权重计算，建立判断矩阵，确定各项指标的重要程度。

4）数据采集和处理:该步骤是对电力系统的相关数据进行采集和处理，以便模型的建立和验证。

5）场景模型建立:通过对采集的数据进行处理，建立场景模型，为后续的计算和分析提供基础。

6）场景模型验证与评估:基于场景模型，进行计算和分析，验证评价体系的有效性和准确性，形成数据分析报告。

7）评价结果分析:针对评价结果进行分析，找出影响电力系统优化和安全的主要因素，并提出改进意见，为后续的优化和安全保障提供指导。

5. 标准结构与内容

本文档包括以下内容：

1) 引言：介绍该标准的背景和适用范围。

2) 规范性引用文件：列出该标准所需的相关规范性文件。

3) 术语和定义：解释相关的术语和定义，以确保对该标准的理解和应用达成一致。

4) 场景化电力模型验证及评价标准：包括以下内容：

- 目标层：明确场景化电力模型验证及评价体系的总体目标。
- 准则层：列出用于评价场景化电力模型验证及评价体系的准则（例如：可靠性、安全性、稳定性和可持续性）。
- 指标层：对每个准则列出具体可量化的指标并进行权重分配。
- 验证与评价方法：根据指标的权重分配，建立评价系统，并根据实际场景进行验证与评价。

6. 条文说明

本文档中的各条目内容均和平台使用和模块操作规范紧密相关。 用户在使用平台时，应严格遵守本文档的各项规定。 对于本文档中未涉及的问题，用户应咨询相关技术人员或运营支持人员。

引　　言

随着科技的不断发展，人们的生活和工作中的各种场景需要进行不断的优化和改进，以达到更好的效果和效率。在这个过程中，我们需要一种有效的方法来验证模型的可靠性和评价方案的优劣。结合电力人工智能技术评价指标及层次分析法（Analytic Hierarchy Process, AHP）的场景化模型验证及评价体系，就是一种有效构建电力人工智能模型场景化验证及评价体系的标准方法。

AHP 是一种定量分析方法，旨在协助决策者在复杂多变的环境下进行决策。它的基本思想是将一个复杂的问题层次化，将问题划分为若干层次，每一层次包含若干个因素，然后通过一系列比较，得出各个因素之间的权重，最终得出整体的权重。其主要思路是将场景化模型进行层次化，将其划分为若干层次，包括目标层、准则层、方案层等，然后通过专家的意见和经验，进行一系列比较，得出各个因素之间的重要性权重，最终得出整个方案的优劣程度。

在此基础上，我们可以通过场景化模型验证及评价体系构建方法，对场景化模型进行验证和评价。例如，在工业生产领域中，我们可以对生产线进行场景化仿真建模，然后对不同的生产方案进行比较和评价。通过这个方法我们可以得出最优的生产方案，从而提高生产效率和质量，可以帮助决策者在复杂的场景下进行决策，提高决策效率和准确性。

电力人工智能模型场景化验证及评价标准

1 范围

本文件规定了电力人工智能场景化模型验证及评价体系构建及评估目标的确定、选择评估指标选取、设定权重、收集数据以及分析数据。适用于电力负荷预测、电力市场价格预测、电力设备故障诊断、电力系统优化等预测诊断模型在性能、可靠性、准确性等方面表现的好坏，帮助评估模型的优劣并提高模型的性能和应用效果。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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。

ISO 9001 质量管理体系标准 质量管理框架

ISO 14001 环境管理体系标准 环境管理体系

ISO 50001 能源管理体系标准

3 术语和定义

下列术语和定义适用于本文件。

AHP（Analytic Hierarchy Process）分析层次过程：

一种用于解决决策问题的系统性方法，通过对问题进行结构化分解，将复杂问题划分为易于处理的层次结构，然后对每个层次结构进行成对比较和计算来确定最佳决策方案。

3.1 场景化模型 Scenario-based model：

将实际场景中的特定问题建模为可计算和可分析的数学模型，以支持决策制定和问题解决。

3.2 验证 verification:

利用统计方法和实验数据，对模型的有效性和可靠性进行确认和核实。

3.3 评价体系 evaluation system：

对特定目标、问题或场景进行全面评价的框架和方法。

3.4 权重分配 weight allocation：

根据比较结果为每个因素分配相应的权重值，以反映其对最终决策方案的重要性。

3.5 层次结构 hierarchical structure：

将复杂问题分解为多个层次，从全局到局部逐步分析和解决问题的方法。

3.6 比较矩阵 Comparison matrix：

用于记录因素之间的两两比较结果，以计算其相对权重的矩阵。

3.7 最大特征值 eigenvalue of maximum：

比较矩阵的最大特征值用于确定权重向量，并衡量相对重要性。

3.8 一致性比率（CR）：

对比较矩阵中的一致性进行度量和评估的方法，其值应小于 0.1 以保证一致性。

4 符号、代号和缩略语

下列符号、代号和缩略语适用于本文件。
本文未定义符号、代号和缩略语。

5 业务场景使用规范

AHP（层次分析法）是一种常见的多标准决策分析方法，它可以用于对复杂问题进行结构化、分层和优先级排序；它能够将复杂的决策问题分解为一系列层级结构，在不同层次上分析决策因素的重要程度，并最终得出决策结果。如在碳计量中常见的技术中有以下几个场景使用到了AHP技术：

（1） 碳排放因子选择：在碳计量中，计算碳排放量需要用到碳排放因子，而不同的排放因子对应不同种类的活动，因此需要对不同的碳排放因子进行优先级排序，以便选择最适合特定活动的排放因子。通过运用AHP方法，可以根据多个因素如可靠性、数据可得性、地区和行业特点等来评估不同的排放因子。

（2） 能源消耗分析：对企业的能源消耗进行分析，通过对能源消耗的分类和评估指标的定义，将不同形式的能源消耗转化为统一的碳排放量表达方式，提高碳排放量的准确度。在此过程中， AHP可以被用于对各种能源消耗类型进行排序和评估。

（3） 碳减排方案比较：对企业实施减碳方案的选择决策。AHP可以用于在参考多个因素（如经济、技术可行性等）的情况下确定最佳方案。通过使用AHP方法，可以将各个因素权重进行比较和归一化，确定最佳的减排方案和最佳的投资组合。

（4） 供应链碳足迹计算：一个清晰的供应链碳足迹计算功能, 需要对整条供应链中的各个最小碳排放环节进行监测和计算。AHP可以用于选择最适合的监测技术和方法, 来提高精度和便利度。
在总体上，AHP可以用于许多应用场景，如对碳排放数据的分析、碳减量方案的制定、企业能源管理、供应链管理等

6 电力人工智能模型场景化验证及评价标准

6.1 人工智能 artificial intelligence

利用数字计算机或者由数字计算机控制的机器，模拟、延伸和扩展人类的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术和应用系统。

6.2 人工智能模型 artificial intelligence model

通过学习海量样本数据中的内在规律和表现层次，实现包括分类、回归、聚类、目标检测、光学字符识别、自然语言处理、语音识别等任务。

6.3 分类模型 classificationmodel

判断一个新的观察样本所属的类别的模型。

6.4 回归模型 regression model

预测一个新的观察样本的连续型目标值的模型。

6.5 聚类模型 clusteringmodel

划分无标记的数据集为多个类别的模型。

6.6 目标检测模型 object detection model

确定图像上目标类别和位置属性信息的模型。

6.7 光学字符识别模型 optical character recognition model

将图片、照片上的文字内容转换为直接可编辑文字的模型。

6.8 自然语言处理模型 natural language processing model

人与计算机之间用自然语言进行有效通信的模型，用于人类与机器的交互。

6.9 语音识别 speech recognition

将人类的声音信号转化为文字或者指令的过程。

6.10 语音识别模型 speech recognition model

将语音数据转换为机器可读数据（文本内容、二进制编码、字符序列等）的模型。

6.11 声音检测类 sound detection and recognition

主要以检测连续或者孤立语流中的特定命令或关键词为主要目的的任务。

6.12 文字输入类 text input

以文字录入为主要目的，要求把语音转化为文字的任务。

6.13 训练数据集 training set

模型构建过程中使用的数据集合。

6.14 测试数据集 testing set

评估模型构建质量的数据集合。

6.15 交并比 intersection over union

计算两个矩形框交集与并集的比值，用于评价两个矩形框的相似度。

6.16 真正例 true positives

模型正确判定的正类样本。

6.17 假正例 false positives

被模型判定为正类的负类样本。

6.18 真负例 true negative

模型正确判定的负类样本。

6.19 假负例 false negative

被模型判定为负类的正类样本。

6.20 真正例率 true positive rate

模型正确判定的正类样本占所有正类样本的比例。

6.21 假正例率 false positive rate

被模型判定为正类的负类样本占所有负类样本的比例。

6.22 正确率 accuracy

模型判断正确的样本占所有样本的比例。

6.23 精确率 precision

模型正确判定的正类样本占所有模型判定的正类样本的比例。

6.24 召回率 recall

模型正确判定的正类样本占所有正类样本的比例。

6.25 F1 值 F1-score

精确率和召回率的调和平均。

6.26 对数损失 logloss

模型决策所包含的信息量。

6.27 P-R 曲线 precision recall curve

以精确率为纵轴、召回率为横轴作图得到的曲线。

6.28 平均精确率 mean precision

模型判定所有类别的精确率的均值。

6.29 平均召回率 mean recall

模型判定的所有类别的召回率的均值。

6.30 平均精度 average precision

在P-R曲线下，召回率从0到1各个点的精确率的均值，即P-R曲线下的面积。

6.31 平均精度均值 mean average precision

模型判定的所有类别的平均精度的均值。

6.32 平均绝对误差 mean absolute error

模型预测结果与目标值的差的绝对值的平均值。

6.33 均方误差 mean square error

模型预测结果与目标值的差的平方的平均值。

6.34 均方根误差 root mean square error

模型预测结果与目标值的差的平方的平均值的根。

6.35 决定系数 r-squared

描述回归方程与真实样本输出之间的相似程度。

6.36 校正决定系数 adjusted r-squared

描述回归方程与真实样本输出之间的相似程度，基于决定系数的调整。

6.37 兰德系数 rand index

模型划分正确的样本对数占所有样本对数的比例。

6.38 调整兰德系数 adjusted rand index

兰德系数的去均值归一化。

6.39 互信息 mutual information

描述两个变量之间重叠的信息量。

6.40 调整互信息 adjusted mutual information

一种基于互信息的聚类效果评价方法。

6.41 轮廓系数 silhouette coefficient

结合簇内凝聚度和分离度的一种聚类效果评价方式。

6.42 平均编辑距离 average edit distance

模型识别的字符串变换到标准字符串进行的插入、删除、替换操作次数的均值。

6.43 字符识别准确率 character recognition accuracy

模型正确识别字符数占所有识别字符数的比例。

6.44 字符识别召回率 character recognition recall

模型正确识别字符数占所有标准字符数的比例。

6.45 文本行定位准确率 text line positioning accuracy

模型正确定位的文本行数占所有文本行数的比例。

6.46 文本行定位召回率 text line positioning recall

模型正确定位的文本行数占所有标准文本行数的比例。

6.47 词错误率 word error rate

模型输出词序列与标准词序列的标准编辑距离占标准词序列中所有词语的比例。

6.48 字错误率 character error rate

模型输出字序列与标准字序列的标准编辑距离占标准字序列中所有字的比例。

6.49 句错误率 sentence error rate

模型输出错误句子的个数占所有句子的比例。

6.50 双语评估替换 bilingual evaluation understudy

用于评估自然语言处理领域生成类文本的质量，简称BLEU。

6.51 鲁棒性 robustness

描述扰动、异常和危险情况下模型的工作能力。

6.52 时间效率 time efficiency

模型对给定的数据进行运算并获得结果所需要的时间。

6.53 空间效率 space efficiency

模型运行过程中显存最大占用率。

6.54 完备性 completeness

模型具有算法框架、开发语言、模型版本、运行环境等完整信息以及模型文件及附属源信息齐全等。

6.55 受试者特征曲线 receiver operating characteristic curve

以真正例率为纵轴、假正例率为横轴作图得到的曲线。

6.56 受试者特征曲线下面积 area under receiver operating characteristic curve

在ROC曲线下，假正例率从0到1各个点的真正例率的均值，即ROC曲线下的面积。

6.57 KS 曲线 kolmogorov-smirnov

用于评估模型风险区分能力，指标衡量的是好坏样本累计分部之间的差值。

6.58 黑盒攻击 black box attack

攻击者未知攻击模型的内部结构，训练参数，防御方法，通过一定规则构造攻击样本以完成攻击。

6.59 白盒攻击 white box attack

攻击者已知攻击模型的内部结构，训练参数，防御方法，构造特定的攻击样本以完成攻击。

6.60 快速梯度符号法 fast gradient sign method

基于模型梯度获得攻击样本的一种白盒攻击方法。

6.61 投影梯度下降法 project gradient descent method

基于模型梯度多次迭代获得攻击样本的一种白盒攻击方法。

7. 评价指标与计算

7.1 功能性

被评价模型如涉及光学字符识别、自然语言处理、目标检测、语音识别相关功能，宜优先选用本导则中光学字符识别、自然语言处理、目标检测、语音识别模型功能性指标进行评价。

7.1.1 分类模型功能性指标

用于评价电力人工智能分类模型实现的功能是否满足要求，宜包括下列内容：

a) 正确率 *Accuracy*，按式（1）计算：

$$Accuracy = (TN + TP)/(TN + TP + FP + FN) \tag{1}$$

b) 精确率 *Precision*，按式（2）计算：

$$Precision = TP/(TP + FP) \tag{2}$$

c) 召回率 *Recall*，按式（3）计算：

$$Recall = TP/(TP + FN) \tag{3}$$

d) *F1* 值，按式（4）计算：

$$F1 = (2 \times Precision \times Recall)/(Precision + Recall) \tag{4}$$

e) 对数损失（*Logloss*），按式（5）计算：

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\, log(p_{ij}) \tag{5}$$

其中，*TP*表示预测正确的正样本，*TN*预测是正确的负样本，*FP*表示预测是错误的正样本，*FN*表示预测是错误的负样本，*N*为实例总数，*M*为类别总数，y_{ij}是一个二值指标，表示第*i*个输入实例是否是*j*类别（$y_{ij}=1$表示是，反之为否），p_{ij}为分类模型预测出的第*i*个实例属于*j*类的概率。

f) *AUC*，按式（6）和（7）计算：

$$AUC = \frac{\sum I(P_{正样本}, P_{负样本})}{M*N} \tag{6}$$

$$I(P_{正样本}, P_{负样本}) = \begin{cases} 1, P_{正样本} > P_{负样本} \\ 0.5, P_{正样本} = P_{负样本} \\ 0, P_{正样本} < P_{负样本} \end{cases} \tag{7}$$

其中，$P_{正样本}$表示分类模型预测的正样本的概率，$P_{负样本}$表示分类模型预测的负样本的概率，M表示测试集中正样本数量，N表示测试集中负样本数量。

7.1.2 回归模型功能性指标

用于评价电力人工智能回归模型实现的功能是否满足要求，宜包括下列内容：

a) 平均绝对误差 MAE，按式（8）计算：

$$MAE = \sum |(\hat{y} - y)| / n \tag{8}$$

b) 均方误差 MSE，按式（9）计算：

$$MSE = \sum (\hat{y} - y)^2 / n \tag{9}$$

c) 均方根误差 RMSE，按式（10）计算：

$$RMSE = \sqrt{\sum (\hat{y} - y)^2 / n} \tag{10}$$

d) 决定系数R^2，按式（11）计算：

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \tag{11}$$

其中，y表示真实值，$\hat{y}$预测值，$\bar{y}$表示全部预测值的平均值。

e) 校正决定系数$R^2_adjusted$，按式（12）计算：

$$R^2\_adjusted = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \tag{12}$$

其中，式（8）-（12）中，y表示真实值，$\hat{y}$预测值，$\bar{y}$表示全部预测值的平均值，n表示测试集样本数量，p表示特征数量。

7.1.3 聚类模型性能指标

用于评价电力人工智能聚类模型实现的功能是否满足要求，宜包括下列内容：

a) 调整兰德系数*ARI*，按式（13）计算：

$$ARI = (RI - E(RI)) / (\max(RI) - E(RI)) \tag{13}$$

b) 调整互信息*AMI*，按式（14）、（15）计算：

$$AMI = \frac{MI - E(MI)}{\max(H(U), H(V)) - E(MI)} \tag{14}$$

$$MI = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j) \log\left(\frac{P(i,j)}{P(i), P'(j)}\right) \tag{15}$$

其中，U、V是N个样本标签的两种不同分配情况，注：$P(i)$表示U中类别为U_i的样本的概率，即$P(i)=\frac{|U_i|}{N}$，$P(j)$表示V中类别为V_j的样本的概率，即$P(j)=\frac{|V_j|}{N}$，$P(i,j)$表示在U中类别为U_i，在V中类别为V_j的样本的概率，即$P(i,j)=|U_i \cap V_j|/N$。$H(U)$指的是数据集U的信息熵，$H(U)=-\sum_{i=1}^{|U|}P(i)\log(P(i))$，$H(V)$指的是数据集V的信息熵。

c) 轮廓系数SC，按式（16）计算：

$$SC=\frac{b(\mathrm{j})-a(i)}{\max\{a(i),b(j)\}} \tag{16}$$

其中，$a(i)=average(i)$，i表示向量到所有它属于的簇中其它点的距离，计算$b(j)=min\ (j)$，j表示向量到某一不包含它的簇内的所有点的平均距离

7.1.4 目标检测模型性能指标

用于评价目标检测模型的实现的功能是否满足要求，宜包括下列内容：

a) IoU 大于或等于预设值，判断为真正例；IoU 小于预设值，判断为真反例。IoU 预设值推荐为0.5。

b) 平均精确率MP，按式（17）计算：

$$MP=\frac{\sum Precision}{N} \tag{17}$$

其中，N表示目标类别数，$Precision$表示每类目标的精确率。

c) 平均召回率MR，按式（18）计算：

$$MR=\frac{\sum Recall}{N} \tag{18}$$

d) 平均精度AP，按式（19）计算：

$$AP=\int_0^1 p(r)dr \tag{19}$$

其中，$p(r)$表示模型的P-R曲线函数。

e) 平均精度均值MAP，按式（20）计算：

$$MAP=\frac{\sum AP}{N} \tag{20}$$

其中，N表示目标类别数。

7.1.5 光学字符识别模型性能指标

用于评价光学字符识别模型实现的功能是否满足要求，宜包括下列内容：

a) 平均编辑距离AED可按式（21）计算：

$$AED=\frac{\sum insert(result,labeL)+delete(result,label)+replace(result,label)}{N} \tag{21}$$

其中，N表示测试数据集中的文本行数，$result$表示模型识别出的一行文本，$label$表示标准文本，$insert(result,label)$表示将$result$编辑为$label$需要执行的插入操作次数，$delete(result,label)$表示将$result$编辑为$label$需要执行的删除操作次数，$replace(result,label)$表示将$result$编辑为$label$需要执行的替换操作次数。

b) 字符识别准确率CRA，按式（22）计算：

$$CRA=\frac{\sum C_{right}}{C_{all}} \tag{22}$$

其中，C_{right}表示识别正确的字符数，C_{all}表示总识别出的字符数。

c) 字符识别召回率 CRR，按式（23）计算：

$$CRR = \frac{\sum C_{right}}{Ctruth} \tag{23}$$

其中，C_{right} 表示识别正确的字符数，C_{truth} 表示标准字符数。

d) 文本行定位准确率 $TLPA$，按式（24）计算：

$$TLPA = \frac{\sum T_{right}}{T_{all}} \tag{24}$$

其中，T_{right} 表示定位正确的文本行数，T_{all} 表示定位出的文本总行数。

e) 文本行定位召回率 $TLPR$，按式（25）计算：

$$TLPR = \frac{\sum T_{right}}{T_{truth}} \tag{25}$$

其中，T_{right} 表示定位正确的文本行数，T_{truth} 表示标准文本行数。

7.1.6 语音识别模型功能性指标

设正确文本字数为N，识别结果文本字数为M，按照识别结果文本与正确文本根据“最小代价匹配”原则运用动态规划算法，得到正确识别字数Mc、删除错误字数D、插入错误字数I、替换错误字数R、出错句子数S和句子总数Sn，则有：N=Mc+R+D,M=Mc+R+J。

定义以下性能指标：

a) 字错误率 CER，按式（26）计算：

$$CER = (I + R + D) / N \times 100\% \tag{26}$$

b) 句错误率 SER，按式（27）计算：

$$SER = S / S_n \times 100\% \tag{27}$$

c) 字匹配率 MCR，按式（28）计算：

$$MCR = M_c / N \times 100\% \tag{28}$$

d) 字准确率 WCR，按式（29）计算：

$$WCR = (M_c - 1) / N \times 100\% = 100\% - CER \tag{29}$$

(2)用于评价声音检测识别类语音识别模型实现的功能是否满足要求，宜包括下列内容：

该类型模型应兼顾动作的可靠性问题，指标定义如下：

假设关键词表的词汇量为KW(个)，检测语音长度为HR(小时)，出现关键词N(次)，C为每小时。每个关键词最大容忍的误报个数(一般取10)，系统报出关键词M(个)，其中，正确FD(个)，错误FA(个)，FD+FA=M。则：

a) 误报率 Far，按式（31）计算：

$$Far = FA / (\mathrm{KW} \times HR \times C) \times 100\% \tag{30}$$

b) 漏报率 Frr，按式（32）计算：

$$Frr = (N - FD) / N \times 100\% \tag{31}$$

c) 检出率 Fdr，按式（33）计算：

$$Fdr = FD / N \times 100\% \tag{32}$$

等错率 EER：DET曲线上Far=Frr时，Far或Frr的值。

质量因数 FOM：ROC曲线上Far=0%，10%，20%，40%，60%，80%，100%时，Fdr的算术平均值一般以EER或FOM值度量系统性能指标。

(3)用于评价语音识别类模型响应时间，宜包括下列内容：

假设识别语音用时 Tr，语音实际时常 Ts，则：

语音识别模型响应系数 RF，按式（34）计算：

$$RF = T_r / T_s \tag{33}$$

7.1.7 自然语言处理模型功能性指标

（1）针对情感分析、词义消歧、词性标注等分类任务，可根据样本的预测结果参照使用式（1）、（3）、（4）对模型的精确率、召回率和F1值进行评价。

（2）对于机器翻译、摘要抽取等生成式任务，还可以使用BLEU等相对指标对模型性能进行评估，具体计算公式如下：

$$bleu_n = \frac{\sum_{c\in candidates}\sum_{n-gram\in c} Count_{clip}(n-gram)}{\sum_{c'\in candidates}\sum_{n-gram'\in c'} Count(n-gram')} \tag{34}$$

其中，canditiate表示模型生成句子的集合，reference表示给定的标准译文，n-gram表示长度为n的连续单词切片，对于分子，其第一个求和符号处理模型生成的所有句子，第二个求和符号处理生成句子中的每个n-gram，$Count_{reference}(n-gram)$表示n-gram切片在reference中的个数。故分子的含义为在给定句子中有多少n-gram出现在标准译文中。分母的含义与分子相同，其统计了所有生成句子中的n-gram总数。BLEU可根据n-gram的不同划分为多种评价指标，常见的有BLEU-1、BLEU-2、BLEU-3、BLEU-4四种，其中BLEU-1衡量的是单词级别的准确性，更高阶的BLEU可以衡量句子的流畅性。

7.2 安全性

用于评价电力人工智能模型的安全程度，宜包括但不限于下列内容：

a) 通过黑盒攻击算法构建攻击样本数据集。
b) 通过白盒攻击算法构建攻击样本数据集。

7.3 鲁棒性

用于评价电力人工智能模型是否在样本变动时仍能维持性能，应包括但不限于下列内容：

a) 图像类样本应经过旋转、裁剪、平移、模糊、加噪、缩放构建新的测试数据集，计算模型在新测试数据集上的功能性指标。
b) 文本类样本应扩展、插值、交换、删除、分隔、词汇替换构建新的测试数据集，计算模型在新测试数据集上的功能性指标。
c) 语音类样本应经过扩展、加噪构建新的测试数据集，计算模型在新测试数据集上的功能性指标。
d) 目标检测类模型的鲁棒性评价样本应包含光照或角度变化、相似物与遮挡物干扰。
e) 光学字符识别模型宜包含字体变换，手写与印刷字符变换等，可根据具体业务应用。
f) 自然语言处理模型鲁棒性评价样本应包含同义词、近义词和反义词干扰，停用词干扰。
g) 语音识别类模型的鲁棒性评价样本应包含语速、语调、声调、口音、表达方式变化。

7.4 效率性

用于评价电力人工智能模型运行效率是否满足要求，宜包括但不限于下列内容：

a) 优化算法：可以通过优化算法来提高模型的算力效率。例如使用并行计算技术，将计算任务分配给多个处理器同时进行计算。
b) 硬件设备升级：对于计算资源有限的场景，可以考虑升级硬件设备。使用更快的 CPU、更多的内存或更高效的 GPU 等。
c) 数据处理优化：考虑使用更高效的数据存储和处理方法来优化模型的计算效率，例如将数据存储在高速缓存中，采用分布式存储和处理等。
d) 资源合理规划：如果遇到计算任务较多的场景，可以通过任务调度等方法来合理分配计算资源，以保证计算效率最大化。

优化算法、硬件设备升级、数据处理优化、资源合理规划等手段应作为提高模型效率性的方法，具体指标需根据业务实际需求确定。

8. 模型评价流程

人工智能模型评价流程包括模型完备性评价、评价测试集选取、选择模型评价指标等五个步骤。

8.1 模型完备性评价

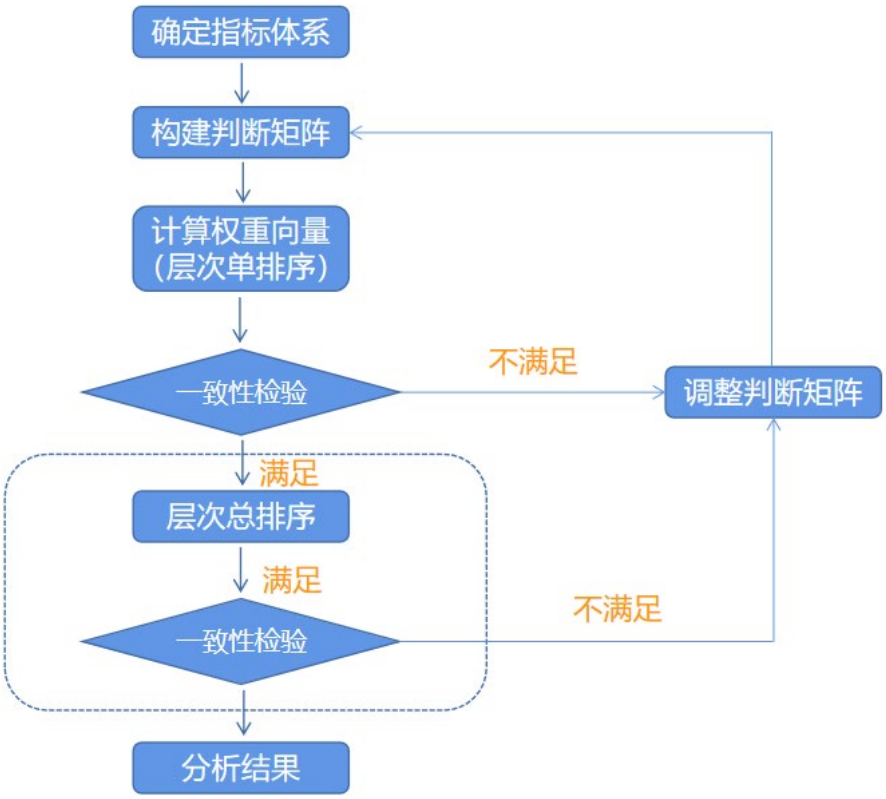

图1　模型评价流程

具备完备性的人工智能模型应具备以下条件：

a) 模型应有对应的模型描述文件，具体应包含模型名称、模型用途、运行模式、模型类型、模型运行环境、开发语言、开发框架、模型版本、模型提供单位、模型训练数据集规模等基本描述信息；

b) 模型宜提供模型源文件和模型相关附属源文件等。

8.2 评价测试数据集选取

a) 测试数据集应与训练数据集具有互斥性，即测试数据集与训练数据集不重合；

b) 测试数据类型为图像文件时，图片宜为 RGB 三通道彩色图像，图片分辨率宜不低于 500*500 像素，图片格式宜为 jpg、png、jpeg、bmp、tif 等；

c) 分类模型与聚类模型测试数据集中所有类别样本数量比例建议相同，可根据具体业务需求进行调整；

d) 目标检测模型测试数据集中包含目标样本与不包含目标样本比例建议为 7:3，且每一个目标类别的样本数量宜不少于 500 张；

e) 测试数据集样本标注信息应完备并准确无误，且应避免被人为添加的恶意数据污染。

f) 语音测试样本格式宜为 cpm、speex、speex-wb、mp3 等，音频采样率宜不低于 8k Hz，音频长度宜不少于 3 秒且不超过 180 秒，语言种类包括中文、英文、地方方言等；

g) 自然语言处理模型的测试数据，宜为 UTF-8 纯文本格式文件，单次文本长度宜不超过 5000 字符（一个汉字、英文字母、标点符号，均记为一个字符），文件内宜根据具体需求涵盖单句、段落、文章、诗词等多种文学结构的文本。

8.3 选择模型评价指标

应结合具体业务应用场景和模型类型选择模型评价指标。具体选取规则如下：

a) 模型评价宜包含功能性、安全性、鲁棒性和效率性等内容；

b) 功能性评价至少应包含目标监测模型性能指标、光学字符识别模型性能指标、语音识别功能性指标、自然语言处理模型功能性指标；

c) 安全性评价：基于模型安全测试样本，计算功能性指标，观察指标变化评价模型安全性，

d) 鲁棒性评价：基于模型鲁棒性测试样本，计算功能性指标，观察指标变化评价模型鲁棒性。

8.4 评估指标确定

a) 确定层次结构：将复杂的决策问题分解成若干层次，并明确每个因素的关系和作用

b) 确定比较矩阵：将同一层次的因素两两进行比较，构建成一个比较矩阵。比较矩阵中的元素代表比较两个因素重要性的权重比例，通常用 1-9 的数字表示，其中 1 表示两个因素同等重要，9 表示一个因素比另一个因素重要程度是极大的差异。如果两个因素之间的重要性不能确定，则取介于 1 和 9 之间的插值数。

c) 计算权重向量：通过计算比较矩阵的特征向量，得到每个因素的权重向量。特征向量是指矩阵中的一个向量，使该向量与矩阵相乘后，得到的向量与原向量具有相同的方向。特征向量的长度是任意的，但是可以通过对其进行归一化来得到权重向量。

d) 一致性检验：为了验证比较矩阵的一致性，需要计算一致性指标和一致性比率。如果一致性比率小于 0.1，即认为比较矩阵是一致的。

e) 建立判断矩阵：判断矩阵是指对于层次结构中的每一层，将各个因素两两比较得到的矩阵。对于每个比较.判断矩阵是 AHP 方法得以实现的基础。

8.5 权重分配方法

a) 计算权重向量：通过计算判断矩阵的特征向量和特征值，可以得到每个因素的权重向量。权重值越高，该因素在决策中的作用越大。例如本标准可引用的指标有：发电成本、负载稳定性、系统安全、能源供应稳定性、能源供需平衡。

b) 一致性检验：AHP 方法中的一致性检验是为了验证判断矩阵不出现矛盾信息的程度。利用计算出的特征向量、特征值，计算一致性指标和一致性比例。若一致性比例接近于 1，即代表该判断矩阵在权重分配过程中是一致的。

c) 敏感性分析：由于 AHP 方法中涉及到多个指标的权值分配和影响力较大的因素之间的比较，通常会运用敏感性分析来分析权重选择下的决策效果。

d) 选择最优决策选取最优决策方案：计算出每个因素的权重后，将各因素的结果进行加权求和，可以得到不同方案之间的比较结果，从而选出最优决策方案。

9. 模型功能性等级判定

9.1 一般规则

a) 本导则适用于模型部署应用前的入网评价和部署应用后的应用效果评价；

b) 本导则对模型的功能性进行等级判定，实际应用中应考虑模型安全性、鲁棒性、效率性等其他因素。

c) 本导则的模型等级是对业务应用功能相近的模型进行归类，不对模型的可用性进行定义，被测模型最终评价结果宜参考具体业务场景的相关规范。

9.2 分类模型等级判定

功能性评价评价等级参考以下规则：

表1 分类模型评价值计算

指标判定	模型等级
准确率≥95%，精确率≥95%，召回率≥95%，F1值≥0.95，对数损失≤0.7，AUC≤0.7	C1
准确率≥85%，精确率≥85%，召回率≥85%，F1 值≥0.85，对数损失≤0.75，AUC≤0.75	C2
准确率≥80%、精确率≥80%，召回率≥80%，F1 值≥0.8，对数损失≤0.8，AUC≤0.8	C3
准确率≥75%、精确率≥75%，召回率≥75%，F1 值≥0.75，对数损失≤0.85，AUC≤0.85	C4
准确率≥70%、精确率≥70%，召回率≥70%，F1 值≥0.7，对数损失≤0.95，AUC≤0.95	C5

9.3 回归模型等级判定

功能性评价等级参考以下规则：

表2 回归模型评价值计算

指标判定	模型等级
决定系数≥0.90	C1
决定系数≥0.85	C2
决定系数≤0.80	C3
决定系数≤0.75	C4
决定系数≤0.7	C5

9.4 聚类模型等级判定

功能性评价等级参考以下规则：

表3 聚类模型评价值计算

指标判定	模型等级
调整兰德系数≥0.95，调整互信息≥0.95，轮廓系数≥0.95	C1
调整兰德系数≥0.85，调整互信息≥0.85，轮廓系数≥0.85	C2
整兰德系数≥0.8，调整互信息≥0.8，轮廓系数≥0.8	C3
调整兰德系数≥0.75，调整互信息≥0.75，轮廓系数≥0.75	C4
调整兰德系数≥0.7，调整互信息≥0.7，轮廓系数≥0.7	C5

9.5 光学字符识别模型等级判定

功能性评价等级参考以下规则：

表4 光学字符识别模型评价值计算

指标判定	模型等级

平均编辑距离，字符识别准确率≥95%，字符识别召回率≥95%，文本定位准确率≥95%，文本定位召回率≥95%	C1
平均编辑距离，字符识别准确率≥85%，字符识别召回率≥85%，文本定位准确率≥85%，文本定位召回率≥85%	C2
平均编辑距离，字符识别准确率≤80%，字符识别召回率≥80%，文本定位准确率≤80%，文本定位召回率≤80%	C3
平均编辑距离，字符识别准确率≥75%，字符识别召回率≥75%，文本定位准确率≥75%，文本定位召回率≥75%	C4
平均编辑距离，字符识别准确率≥70%，字符识别召回率≥70%，文本定位准确率≥70%，文本定位召回率≥70%	C5

9.6 目标检测模型等级判定

目标检测模型 IOU 值宜设置为 0.5，功能性评价值计算参考以下规则：

表5 目标检测模型评价值计算

指标判定	模型等级
平均精确率≥95%，平均召回率≥95%，平均精度≥0.95，平均精度均值≥0.95	C1
平均精确率≥85%，平均召回率≥85%，平均精度≥0.85，平均精度均值≥0.85	C2
平均精确率≥80%，平均召回率≥80%，平均精度≥0.8，平均精度均值≥0.8	C3
平均精确率≥75%，平均召回率≥75%，平均精度≥0.75，平均精度均值≥0.75	C4
平均精确率≥70%，平均召回率≥70%，平均精度≥0.7，平均精度均值≥0.7	C5

9.7 语音识别模型评价体系

功能性评价等级可参考以下规则：

表6 语音识别模型评价值计算

指标判定	模型等级
字精确率≥95%，巨错误率≤5%，响应系数≤1.1	C1
字精确率≥85%，句错误率≤10%，响应系数≤1.2	C2
字精确率≥80%，句错误率≤15%，响应系数≤1.3	C3
字精确率≥75%，句错误率≤30%，响应系数≤1.4	C4
字精确率≥70%，句错误率≤35%，响应系数≤1.5	C5

9.8 自然语言处理模型评价体系

对于自然语言处理的分类任务，功能性评价等级可参考以下规则：

表7 自然语言处理模型评价值计算

指标判定	模型等级
准确率≥95%，召回率≥95%，F1值≥0.95，	C1
准确率≥85%，召回率≥85%，F1值≥0.85，	C2
准确率≥80%，召回率≥80%，F1值≥0.8，	C3
准确率≥75%，召回率≥75%，F1值≥0.75，	C4
准确率≥70%，召回率≥70%，F1值≥0.7，	C5

对于自然语言处理领域的生成式任务，因文本质量难以量化，功能性评价可参考 BLEU 等生成式任务评价指标，与其他同类型模型进行横向对比。