ICS 号
中国标准文献分类号

T/CES

团 体 标 准

T/CES XXX-XXXX

电力人工智能平台样本规范

Sample specification of electric artificial intelligence platform

（征求意见稿）

XXXX-XX-XX 发布 XXXX-XX-XX 实施

中国电工技术学会发布

目　　录

前　言

为规范人工智能图像视频、语音、文本类样本的样本基本要求、样本标注要求和样本标注流程，为人工智能样本标注工作开展提供指导规范，制定本文件。

本标准按照 GB/T1.1—2009《标准化工作导则 第 1 部分 标准的结构与编写》给出的规则起草。

本标准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。

本标准起草单位：国网信息通信产业集团有限公司、福建亿榕信息技术有限公司、北京国网信通埃森哲信息技术有限公司、安徽继远软件有限公司、国网重庆市电力公司电力科学研究院、四川大学、四川中电启明星信息技术有限公司、国网重庆市电力公司、中国电力科学研究院有限公司。

本标准主要起草人：李强、邱镇、赵峰、刘迪、廖逍、李炳森、黄晓光、刘永清、向辉、许中平、谭洪恩、苏少春、杨迎春、周孔均、王晓东、钟加勇、彭舰、王秋琳、黄飞虎、王金策、田鹏、吕小红、厉仄平、苏江文、费长顺、宋卫平、赵灿灿、张琳瑜、崔迎宝、刘璟、宫晓辉、尹玉、周伟、王蓓、梁翀、李温静、王卫卫、伍臣周、王晓辉。

本标准为首次发布。

电力人工智能平台样本规范

1 范围

本文件规定了人工智能样本的基本要求、标注要求、标注流程，其中人工智能样本类型包括图像视频、语音和文本。

本部分适用于人工智能平台样本库的规划、设计、开发、建设和运维。

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。

GB/T 5271.28-2001 信息技术 词汇 第28部分;人工智能 基本概念与专家系统

T/CESA 1040-2019 信息技术 人工智能 面向机器学习的数据标注规程

GB/T 13715—92 信息处理用现代汉语分词规范

GB/T 5271.29—2006/ISO/IEC 2382-29:1999 信息技术 词汇规范：第29部分人工智能 语音识别与合成

3 术语和定

下列术语和定义适用于本文件。

3.1 人工智能 artificial intelligence

一门交叉学科，是自动化和计算机两大学派，研究表现出与人类智能（如推理和学习）相关的各种功能的模型和系统。

[GB/T 5271.28—2001,定义28.01.01]

3.2 样本数据 sample data

其具备的特征能够反映总体数据情况的一部分个体数据。

[Q/GDW 12118.1—2021,定义3.5]

3.3 标注 corpus annotation

采用人工或计算机自动方式对样本的属性或特征进行描述。

[Q/GDW 1906—2013,定义3.5]

3.4 图像分辨率 resolution

图像分辨率指图像中存储的信息量，是每英寸图像内有多少个像素点，分辨率的单位为PPI(Pixels Per Inch)，通常叫做像素每英寸。

[Q/GDW 12118.3—2021,定义3.1]

3.5 视频码率 video code rate

数据传输时单位时间传送的数据位数，单位时间内取样率越大，精度就越高，处理出来的文件就越接近原始文件。（采用“注”的形式）

3.6 采样率 sample rate

录音设备在一秒钟内对声音信号的采样次数，采样频率越高声音的还原就越真实越自然。本文件中除非说明，采样率为音频采样率。目前语音识别服务支持 16000 赫兹和 8000 赫兹两种采样率，其中电话业务一般使用 8000 赫兹，其余业务使用 16000 赫兹。（采用“注”的形式）

3.7 无效音 invalid voice

无实际使用价值的音频。包括音频中只存在背景噪声或者音乐，或背景噪声和音乐声音过大影响识别说话内容；语音为与普通话相差较大的方言或唱歌；语音只存在语气词，以及无意义词。例如：嗯、呃、啊、好、对、是的等；语音过小或者发音模糊，无法确定语音内容。

3.8 标签 label

标识数据的特征、类别和属性等内容，可用于建立数据及深度学习训练要求所定义的机器可读数据编码间的联系。

[T/CESA 1040-2019 定义 3.2]

3.9 发音 utterance

用户输入的一个语音单词，可以是词、短语或者句子。语音单元之间需有有意、明显停顿。

[GB/T 13715-92，定义 3.5]

4 缩略语

下列缩略语适用于本文件。

BIOES：BIOES标注模式 (B-begin,I-inside,O-outside,E-end,S-single)

COCO：上下文中公共对象（Common Objects in Context）

JPEG：联合图像专家组(Joint Photographic Experts Group)

Json：JS对象简谱(JavaScript Object Notation)

PCM：脉冲编码调制(Pulse Code Modulation)

PNG：便携式网络图型（Portable Network Graphics）

VOC：视觉对象类（Visual Object Classes）

XML：可扩展的标记语言(Extensible Markup Language)

RAW：原始音像资料（Raw Sound Data）

5 样本基本要求

本文件制定规范了人工智能图像（含视频）、语音、文本类样本数据基本要求、样本标注要求和样本标注流程，样本标注后汇总到电力人工智能平台中进行管理。

5.1 图像（含视频）类样本基本要求

5.1.1 图像（含视频）文件存储格式要求

图像文件常用的存储格式应该以 jpg、jpeg、png 常用格式，视频文件常用的存储格式应该以 mp4 常用格式。

根据业务需求对视频文件进行部分截取时，截取的图片存储格式应该为 jpg、jpeg、png 常用格式。

5.1.2 图像（含视频）文件命名要求

图像样本名称应有：项目命名词或图像来源；当前图像（视频）专业信息；当前文件的日期，日期格式：年+月+日；文件唯一性编号，从 1 开始计数等部分组成。

视频样本根据业务需求需对其部分帧进行截取时，截取图像名称应由：源视频样本名称；文件唯一性编号，从 1 开始计数，两个部分组成。

5.1.3 图像（含视频）类样本质量要求

图像视频的样本质量按照分辨率应满足：图像样本与视频抽帧样本分辨率应为 1920*1080 像素及以上，识别主题内容应不少于 15*15 像素要求。识别主题内容边缘清晰，无严重重影、遮挡范围不超过主题三分之二；

视频样本码率应满足：视频包含业务相关内容，流畅、清晰，满足编解码格式需求。分辨率-码率宜为 1920*1080 像素-5Mb/s；

样本目标物体的完整性：目标区域在整个图像样本中的占比应不低于 40%。

5.1.4 图像（含视频）样本详细描述要求

样本集描述文件存储格式应为txt格式；命名应有项目命名词或样本来源、本文件创建的日期，日期格式：年+月+日、文件唯一性编号，从1开始计数，三个部分组成；文档内容应描述本样本集的基本信息，应包括样本所属项目、样本来源、创建日期、样本上传单位及联系人、样本量、样本用途等信息，具体内容格式参考附录A.1。

5.1.5 图像（含视频）标注要求

5.1.5.1 标签信息要求

1) 整个样本集中同一类目标物体的标签信息命名应保持一致；
2) 输电图像样本添加标签信息应包括：输电区域名称、线路名称、电压等级、杆塔号、巡检时间、巡检人员、缺陷内容等信息；
3) 变电图像样本添加标签信息应包括：变电区域名称、变电站名称、设备名称、电压等级、巡检时间、巡检人员、缺陷内容等信息；
4) 配电图像样本添加标签信息应包括：配电区域名称、线路名称、电压等级、杆塔号、巡检时间、巡检人员、违规原因等信息；
5) 其他图像样本添加标签信息应包括：创建时间、创建者、图像用途等信息。

5.1.5.2 样本标注规则

视频样本无其他标注要求，主要对截取的图像样本按照图像标注的要求进行标注即可。

图像样本标注规则应按照：

1) 图像中所有目标物体应全部标注；
2) 采用最小标注框对目标物体进行标注；
3) 目标物体前端遮挡不宜超过 2/3，若目标物体存在过大比例（大于 2/3）的遮挡情况，应进行目标物体轮廓标注后，对被遮挡的目标物体添加“遮挡”标签；
4) 标注时使用 COCO 或 VOC 格式存储标注内容；

5) 图像分类标注，同一类图像应使用相同的分类名；
6) 使用矩形框对图像样本进行标注时，目标物体应全包含在标注框内，除非目标物体有过于细长伸出的部位（伸出部分占像素比例小于 5%）、矩形框应将目标物体全部包括，矩形框边缘与目标物体的距离应小于矩形框边长的 3%；
7) 使用边界框对图像样本进行标注时，边界框应将目标物体全部包括，边界框边缘与目标物体的距离应小于边界框边长的 3%；
8) 使用旋转矩形框对图像样本进行标注时，标注信息内容应包含矩形框中心点坐标、宽、高和沿垂直方向顺时针旋转的角度、旋转角度应采用弧度制，同一批数据集标注时应采用同一种旋转方式。对于规则的目标物体，旋转方向宜尽量与设备轴向方向保持一致，标注方向与目标物体轴向方向的角度偏差不宜超过 5%等。

5.1.5.3 标注记录文件命名与存储规则

标注文件应与对应标注图像命名一致，应保存为xml或Json等常见格式文件，存储于指定位置标注数据文件夹内。该文件记录信息应包含对应图像（含视频）的基本信息、目标设备基本信息、缺陷情况信息等。

5.2 语音类样本基本要求

5.2.1 语音文件存储格式要求

语音文件存储格式应为：mp3、pcm、raw 等常见格式。

5.2.2 语音文件命名要求

每个省份应以省名称每个字的拼音首字母编号，发音人性别应用英文 female 和 male 的英文首字母 F、M 编号，设备类别应用英文首字母编号。

5.2.3 语音类样本质量要求

语音样本质量应按照：
1) 录音环境应选择安静、无噪音干扰的环境；
2) 不得有文字错误，有一个或一个以上，该条语音就不达标；
3) 整段语音语速应保持在 150-200 音节/分钟；
4) 语音主体内容必须能有效辨识，语音分贝应不低于 25 分贝；
5) 语音样本无效音占总样本语句应不超过 40%等要求内容。

5.2.4 语音样本描述文件

语音样本描述文件应包含：描述文件记录声源的信息、描述文件记录语音样本的信息等两个文件。

1) 记录声源信息的描述文件命名为：语音文件名+声源信息.txt，内容应包含：声源信息、系统信息。
2) 记录语音样本信息的描述文件命名为：语音文件名+Info.txt，内容应包含：标注规范、料库名、录音文件夹编号、录音日期、录音时间点、录音格式、通道数、发音人 ID、录音地点、环境信息等信息。

5.2.5 语音类样本标注要求

5.2.5.1 语音切分要求

长语音需要切分成小分句应按照：

1) 通过时长作为语音的切分依据；
2) 切分点应落在说话停顿处；
3) 切分点应位于音频波形有明显静音段的地方；
4) 每个切分后的小分句语音，应在 5 秒至 6 秒之间。

5.2.5.2 语音标注规则

语音类样本标注应按照：

1) 标注语音文本时，内容应和听到的语音完全一致，不可多字、少字、错字。对于感叹、停顿的词（例如“嗯”、“啊”、“呃”等）应标注对应的汉字。存在口误、结巴、不流利的内容应完整地标注对应的汉字；
2) 对于语音中出现的阿拉伯数字应写成汉字形式；
3) 标注中应只含有中文、英文以及英文中特殊符号；
4) 在语音转写内容的完整性应与实际发音一致，不可删减。语音中听不清的词应用“*”替代，但在一段语音中出现“*”的概率不可高于 5%；
5) 对于有口音的词组时，应按照普通话的相应词组来标注；
6) 标注文本应由语音对应时间戳和标注内容两部分构成，用半角冒号隔开（:）隔开。标注文件内容格式：“音频语句开始时间戳-音频语句结束时间戳:标注文本内容”。例：“00:02:35-00:02:40:查一下我的定期存款什么时候到期”。

5.2.5.3 标注记录文件命名与存储规则

标注文件应和对应的标注语音文件命名一致，应保存为txt格式。

5.3 文本类样本基本要求

5.3.1 文本文件存储格式要求

文本数据存储格式应采用txt、csv、Json、xls、xlsx、xml等常见格式。

5.3.2 文本文件命名要求

文本文件名称应由：第一部分为项目命名词或文本来源；第二部分为当前文本文件的专业信息；第三部分为当前文本文件的日期，日期格式：年+月+日；第四部分为文件唯一性编号，从1开始计数等组合而成。

5.3.3 文本类样本质量要求

文本类样本质量应按照：

1) 需支持计算机正常读取，文本内容无乱码；
2) 内容要满足相关业务需求；
3) 应该使用 UTF-8 编码格式。

5.3.4 文本样本描述文件

文本样本集的描述文件应按照：

1) 文本存储格式应为 txt 格式；

2) 命名应由：项目命名词或样本来源；本文件创建的日期，日期格式：年+月+日；文件唯一性编号，从 1 开始计数等三个部分组成。
3) 文档内容应描述文本样本集的基本信息，包括样本所属项目、样本来源、创建日期、样本上传单位及联系人、样本量、样本用途等多样信息，具体内容格式参考附录A.1。

5.3.5 文本类样本标注要求

5.3.5.1 基本要求

应按照标注对象范围、标注方式、标注文件命名要求。
1) 文本类样本标注应包括词、句子、整个文本等不同规范的标注；
2) 文本类样本标注应有序列标注、指针标注、多头标注等多种标注形式；
3) 对于序列标注时，应采用 B、I、E、O、S 等标签列表，应采用 BIO、BIOES 标签方案进行标注。

5.3.5.2 单类词词性标注要求

在标注时针对单类词应按照《语法信息词典》确定其词性。

5.3.5.3 多类词词性标注要求

在标注时针对多类词，应按照n-q、a-v、v-b、p-v、p-c等多种规则多类词对其标注，详细规则见附录A.2。

5.3.5.4 实体抽取样本标注要求

实体抽取样本标注须符合：定义实体语义类型，包含实体名称与层次结构，需在样本标注前进行；如果实体内存在属性，应定义属性名称与属性值。

5.3.5.5 标注记录文件命名与存储规则

标注文件应由：与对应标注文本命名一致、为“-bz”，应保存为txt格式这两部分组成。

6 样本标注流程

样本标注流程有：样本检查、标注工具选择、标注任务开展、标注结果收集和标注结果检查等环节。

6.1 样本检查

样本标注时，需提前按照样本基本要求对需要标注的样本集进行检查，可根据样本数量或业务需求进行全面检查或随机抽查。
1) 全面检查：需要对指定的样本集范围内的所有样本数据进行逐条检查。
2) 随机抽查：可按照随机抽样和分类抽样。随机抽样指针对不同的业务类型样本数据进行随机检查。分类抽样指针对同一个业务类型的样本数据，根据类型进行分类检查。

6.2 标注工具

应使用电力人工智能平台标注工具或与其格式相兼容的标注工具进行标注。

6.3 标注任务开展

需根据标注任务的难易程度和业务需求来选择半自动化标注和人工标注等两种方式。

1) 半自动化标注：应按照样本构建、模型构建、批量标注顺序执行。
   a) 样本构建：从需要标注的样本中抽取测试样本和训练样本，应采用随机抽查或分类抽查方法，测试样本和训练样本占样本总量的比例需高与 1%，测试样本和训练样本的比例可为 3:7，测试样本和训练样本无交集。
   b) 模型构建：采用标注后的训练样本建立标注模型；采用标注后的测试样本测试标注模型。评估模型性能时，可采用召回率、精确率指标进行评估性能。
   c) 批量标注：使用标注模型批量执行标注任务。
2) 人工标注：应按照试标注、批量标注顺序执行。
   a) 试标注：抽取试标注样本，从需要标注的样本中，可使用随机抽查或分类抽查的方法，抽取比例需高于待标注样本总量的 1%；
   b) 批量标注：标注人员批量执行标注任务。

6.4 样本标注结果收集

1) 为防止文件外泄，由统一的人员进行样本标注结果的回收和存放；
2) 为防止文件遗漏，标注结果的（包括任务名称、任务类型、任务开始时间、任务结束时间、任务描述）等相关信息，应由对应的收集人员进行检查。
3) 由标注结果收集人员进行分类保存至电力人工智能平台中，应按照样本类型（图像视频，语音和文字）和标注方式（图像标注包括图像分类、图像目标检测、图像分割；文本标注包括文本分类、文本标注；音频标注包括音频分类、音频标注）。

6.5 样本标注结果检查

应按照样本标注要求对收集的样本标注结果进行检查，可根据样本标注和业务需求的数量，进行全量检查或抽样检查。

1) 对指定范围内的所有样本进行逐条检查为全量检查。
2) 可用随机抽查和分类抽查的方式为抽样检查。可按照随机抽样和分类抽样。随机抽样指针对不同的业务类型样本数据进行随机检查。分类抽样指针对同一个业务类型的样本数据，根据类型进行分类检查。

附 录 A

表A.1 图像视频样本描述文件内容要求

条目	内容要求	示例
样本所属项目	说明本批次样本收集工作所属的项目情况，若无项目则填无项目依托	样本所属项目：属于 xxx 项目/无项目依托；
样本来源	说明本批次样本采集来源的地区及业务领域	样本来源：来源于甘肃地区输电线路巡检业务中均压环、绝缘子设备；
创建日期	指本批次样本收集完成的日期	创建日期：2021-8-29；
样本上传单位及联系人	说明本批次样本上传的单位以及联系人	样本上传单位及联系人：中国电科院-张三；电话：139xxxx0000；邮箱：xxx@xxx.com.cn；
样本量	说明本批次样本数量	样本量：图像样本：300 张；视频样本：20 个，总时长 5h20min18s；
样本用途	说明本批次样本的用途，包括目前已经用于的业务以及将来可能用于的业务	样本用途：可用于输电线路巡检业务；

表A.2　多种多类词标注规则

词性类别	标注规则	实例
名词 n - 量词 q 多类	数词 + n-q + n，应为量词 q 类	一/m 车/q 煤/n
	汉语中部分名词临时做量词且只能前接数词“一”，应标为量词 q 类	做/v 了/ u 一/m 菜/q
	“这” “那”“每”等指示代词+ n-q +n，应标为量词 q 类	这/r 床/q 被子/n
	其他情况，应标为名词 n 类	上/v 车/n
动词 v - 名词 n 多类	该词表示一种动作时，后面带真宾语，应标为动词 v 类	编辑/v 科技/n 文献/n
	该词直接作主语或谓宾动词的宾语，应标为动词 v 类	我们/n 来/v 的/u 目的/ n 就是/v 考察/v
	该词指称人或物时，应标为名词 n 类	忘/v 了/u 买/v 一/m 把/q 锁/n
	该词作特殊动词“有”的宾语，应标为名词 n 类	领导/n 对/p 这/r 件/q 事/n 有/v 考虑/n
	该词充当了形式动词或其他准谓宾动词的准谓词性宾语，应标为名词 n 类	进行/v 一/m 次/q 深入/a 的/u 考察/n
	该词直接充当体词性短语的中心语，应标为名词 n 类	加以/v 整理/n
	该词不加助词“的”，直接充当体词性短语的修饰语，应标为名词 n 类	这个/r 研究/n 思路/n 很/d 新颖/a
动词 v - 副词 d 多类	单独做谓语，应标为动词 v 类	他/r 讽刺/d 说/v
	该词后加“地”作状语，应标为动词 v 类	主任/n 强调/v 地/u 指出/v
介词 p - 动词 v 多类	单独做谓语，应标为动词 v 类	你/r 在/v 不/d 在/v 家/n ?/w
	状语或补语，应标为介词 p 类	从/p 东/F 到/p 西/f 共/d 长/a 30/m 米/q

表 A.2（续）

词性类别	标注规则	实例
介词 p – 连词 c 多类	该词前后成分不能互换位置或者在该词的前面可以加修饰成分，应标为介词 p 类	你/r 别/d 跟/p 他/r 跑/v
	该词前后成分可以互换位置且在该词的前面不能有修饰成分，应标为连词 c 类	我/r 跟/c 他/r 都/d 是/v 大学生/n
连词 c – 副词 d 多类	该词在句子中修饰形容词、动词，应标注为副词 d 类	但/d 见/v 门上/s 贴/v 着/u 一/m 副/q 对联/n 。/w
	该词主要连接句子和子句，表示子句之间转折、让步等语义组合关系，应标注为连词 c 类	我/r 受/v 了/u 点/q 伤/Ng ，/w 不过/c 不/d 要紧/a
形容词 a – 动词 v 多类	该词在句子中带了真宾语，应标为动词 v 类	他/r 跟/p 她/r 没/d 红/v 过/u 脸/n
	该词受“很”一类程度副词修饰，应标为形容词 a 类	这/r 花/n 很/d 红/a
	该词修饰名词作定语，应标为形容词 a 类	繁荣/a 的/u 景象/n
	该词作动词的补语，应标为形容词 a 类	放/v 明白/a 一些/m
形容词 a – 名词 n 多类	该词作了“有”的宾语，应标为名词 n 类	这里/s 有/v 奥妙/n
	该词充当了准谓宾动词的准谓词性宾语，应标为名词 n 类	维护/v 环境/n 的/u 整洁/n
	该词直接充当体词性短语的中心语，应标为名词 n 类	交通/n 安全/n 是/v 第一/m 要/v 注意/v 的/u
	该词直接作主语或谓宾动词的宾语，应标为形容词 a 类	需要/v 进一步/d 努力/a
形容词 a – 副词 d 多类	该词直接作状语时，应标为副词 d 类	深入/d 研究/v 语法/n 有利/a 于/p 自然/a 语言/n 处理/n
	该词后接“地”作状语时，应标为形容词 a 类	我们/r 应当/v 深入/a 地/u 研究/v 语法/n
区别词 b – 副词 d 多类	该词作状语，应标注为副词 d 类	我们/r 会/v 共同/d 进步/v
	该词作定语或与"的“组成”的"字结构，应标注为区别词 b 类	共同/b 目标/n 是/v 完成/v 这/r 项/q 任务/n