团体标准

发 布

中国电机工程学会

20XX—XX—XX实施

20XX—XX—XX发布

电力人工智能非结构化样本脱敏规范

Specification for desensitization of unstructured samples of electric power artificial intelligence

（征求意见稿）

T/CSEE XXXX—YYYY

代替 T/XXXX

ICS 19.020

CCS K85

目 次

[前 言 4](#_Toc11222)

[1 范围 5](#_Toc2360)

[2 规范性引用文件 5](#_Toc18934)

[3 术语和定义 5](#_Toc13375)

[4 数据脱敏原则 6](#_Toc2632)

[4.1 有效性 6](#_Toc14896)

[4.2 可用性 6](#_Toc607)

[4.3 高效性 6](#_Toc6220)

[4.4 稳定性 6](#_Toc26764)

[4.5 可配置性 7](#_Toc26984)

[5 电力样本可用性定级 7](#_Toc23275)

[6 电力样本保密性定级 7](#_Toc3915)

[7 电力人工智能敏感样本定义与分类 7](#_Toc26591)

[7.1 敏感样本分类 7](#_Toc3001)

[7.2 敏感样本定义 7](#_Toc11103)

[8 脱敏场景 8](#_Toc12034)

[8.1 共享场景 8](#_Toc12229)

[8.2 开放场景 8](#_Toc12325)

[8.3 数据分析场景 8](#_Toc30207)

[8.4 开发测试场景 8](#_Toc8881)

[9 电力图像样本脱敏处理 8](#_Toc26701)

[9.1 功能要求 8](#_Toc8464)

[9.2 数据格式要求 8](#_Toc21494)

[9.3 性能要求 8](#_Toc30184)

[9.4 脱敏方法 9](#_Toc10646)

[9.5 脱敏效果评价 9](#_Toc9644)

[10 电力文本样本脱敏处理 10](#_Toc247)

[10.1 功能要求 10](#_Toc411)

[10.2 数据格式要求 10](#_Toc1170)

[10.3 性能要求 11](#_Toc30646)

[10.4 脱敏方法 11](#_Toc18981)

[10.5 脱敏效果评价 12](#_Toc12152)

[11 电力语音样本脱敏处理 12](#_Toc2314)

[11.1 功能要求 12](#_Toc16736)

[11.2 数据格式要求 12](#_Toc25648)

[11.3 性能要求 12](#_Toc1912)

[11.4 脱敏方法 12](#_Toc27874)

[11.5 脱敏效果评价 13](#_Toc14222)

[附　录　A 敏感样本定义（资料性） 14](#_Toc15285)

前 言

本文件按照《中国电机工程学会标准化管理办法》、《中国电机工程学会标准化管理办法实施细则》的要求，依据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专利的责任。

本文件由国家电网有限公司大数据中心提出并解释。

本文件由中国电机工程学会标准工作委员会\*\*\*\*工作组归口。

本文件起草单位：

本文件主要起草人：

本文件首次发布。

电力人工智能非结构化样本脱敏规范

1. 范围

本文件规定了图像、文本、和语音等电力样本脱敏应用相关的原则、脱敏场景、功能要求、性能要求、脱敏方法和脱敏效果评价。

本文件适用于图像、文本、和语音等电力样本脱敏处理，其他领域数据脱敏处理可参照使用。

1. 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。

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 20945—2013 信息安全技术 信息系统安全审计产品技术要求和测试评价方法

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 35678—2017 公共安全 人脸识别应用图像技术要求

GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型

GB/T 39477—2020 信息安全技术 政务信息共享 数据安全技术要求

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

DB 37/T 3523.2—2019 公共数据开放 第2部分：数据脱敏指南

DB 52/T 1126—2016 政府数据 数据脱敏工作指南

DA/T 77—2019 纸质档案数字复制件光学字符识别（OCR）工作规范

1. 术语和定义

GB/T 39477-2020、GB/T 5271.28—2001和DA/T 77-2019,3.3界定的以及下列术语和定义适用于本文件。

非结构化样本 Nstructured sample

没有固定结构的样本数据，可以分为文本类、音频类和图像类三种类型，每种类型都有自己的格式和规范。

样本脱敏 Sample desensitization

对电力样本中某些敏感信息按照一定规则进行数据变形，实现敏感信息样本的可靠保护。

敏感区域 Sensitive Region

电力样本中包含地点、设备名称、人脸、身份证等敏感信息的区域。

交并比 Intersection-over-Union

产生的候选框与原标记框的交叠率，即它们的交集与并集的比值，完全重叠时，比值为1。

敏感数据 sensitive data

由权威机构确定的受保护的信息数据。

注：敏感信息数据的泄露、修改、破坏或丢失会对人或事产生可预知的损害。

[来源：GB/T 39477-2020]

文本脱敏 Data desensitization

按照一定规则对电力文本原始数据进行处理，达到屏蔽敏感信息的一种数据保护方法。

人工智能 artificial intelligence

一门交叉学科，通常视为计算机科学的分支，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。

[来源：GB/T 5271.28—2001，28.01.01]

光学字符识别 optical character recognition（OCR）

将图片、照片上的文字内容直接转换为可编辑文本的一种技术。

[来源：DA/T 77-2019，3.3]

脱敏成功率 desensitization success rate

电力人工智能样本中成功完成脱敏信息占所需脱敏信息的比率。

1. 数据脱敏原则
	1. 有效性

电力人工智能样本脱敏宜确保脱敏工作有效性。电力样本在脱敏处理之后，样本中包含的敏感信息已被移除，第三方无法通过处理后的电力样本得到敏感信息，并防止利用推理和重建等方式恢复电力样本中敏感信息。

* 1. 可用性

电力人工智能样本脱敏宜确保脱敏工作可用性。脱敏后的电力样本应保持原有的特征，应尽可能保留原始样本中的有用信息，满足后续电力人工智能相关业务应用需求。

* 1. 高效性

电力人工智能样本脱敏宜确保脱敏工作高效性。电力样本脱敏过程中，宜通过自动化、可重复方式实现，在不影响有效性的前提下，平衡脱敏的力度和代价，控制时间和经济成本。

* 1. 稳定性

电力人工智能样本脱敏宜确保脱敏工作稳定性。在输入条件一致的条件下，对同一电力样本，无论脱敏多少次，应保证电力样本脱敏结果是稳定相同的。

* 1. 可配置性

电力人工智能样本脱敏宜确保脱敏工作可配置性。电力样本可根据不同的输入条件，生成不同的脱敏结果，即可按不同的使用场景需求，为用户提供不同的脱敏样本。

1.
2. 电力样本可用性定级

电力样本可用性，即执行电力样本脱敏的同时尽可能保留样本的价值，最小化样本信息损失，满足后续电力人工智能相关业务需求。针对电力样本应用的不同场景，可将样本可用性定级分为“高”、“低”两级:

a)电力样本可用性“高”，即脱敏后电力样本满足样本应用要求且样本信息损失程度较低；

b)电力样本可用性“低”，即脱敏后电力样本满足数据应用要求且样本信息损失程度较高。

1. 电力样本保密性定级

针对电力人工智能样本敏感信息泄露造成的社会危害性严重程度将电力样本保密性定级分为“高”、“低”两级：

a)电力样本中包含少量敏感信息或敏感信息价值不高，该类样本泄露，无法造成实际电力设备或人员的损害或威胁，则此类样本的样本保密性定级为“低”。

b)电力样本中包含大量敏感信息或敏感信息价值高，该类样本泄露，会造成实际电力设备或人员的损害或威胁，则此类样本的样本保密性定级为“高”。

1. 电力人工智能敏感样本定义与分类
	1. 敏感样本分类

电力人工智能敏感样本主要来自发电域、输电域、供电域和用电域，涉及图像、文本和语音多种类型，具体分类见表1。

表1 敏感样本分类

|  |  |
| --- | --- |
| **样本类型** | **范围** |
| 电力图像样本 | 输电无人机巡检影像、变电巡检机器人影像、安监现场管控影像、营销设备影像和运维工作单影像、发电设备影像、人员证件照、人脸影像、电力营销工作单、电网运维工作单、发电设备类运维工作单、电力企业财务报销单、智能电表等 |
| 电力文本样本 | 名称类、电力企业及子公司地址类、电力工作人员和客户联系类、电力工作人员证件类、电力企业资产类、电力相关金融类、电力数据类（包含电力行业发电、配电、输电、变电、用电和调度等环节的电力数据）等 |
| 电力语音样本 | 电力智能客服通话录音文件、电力企业会议录音文件、语音指令、检修录入语音文件等 |

* 1. 敏感样本定义

电力样本敏感信息包括杆号牌、设备铭牌、人脸、用户信息、地址、电话等，涉及输电巡检、变电巡检、安监、营销、智能客服等多个领域，具体定义见附录 A 。

1. 脱敏场景
	1. 共享场景

共享电力样本脱敏是指在电力系统内各单位之间进行脱敏后样本共享。

在电力样本共享场景下，需根据电力样本的保密等级来进行不同程度脱敏处理后共享。

* 1. 开放场景

开放电力样本脱敏是指电力行业企业面向公民、法人和其他组织以非排他形式有条件开放部分电力样本数据内容，如脱敏后开放。

在样本开放场景下，需根据电力样本的保密等级来进行不同程度脱敏处理后开放。

* 1. 数据分析场景

电力人工智能模型训练场景是指利用脱敏后的电力样本进行电力人工智能模型训练，电力人工智能模型开发人员获取电力样本时进行脱敏处理，脱敏处理要服从有效性和可用性。

* 1. 开发测试场景

电力人工智能模型测试场景是指利用脱敏后的电力样本进行电力人工智能模型测试，电力人工智能模型测试人员或者电力人工智能模型使用方获取电力样本时进行脱敏处理，脱敏处理要服从有效性和可用性。

1. 电力图像样本脱敏处理
	1. 功能要求

电力图像样本脱敏处理功能应具备对电力图像中特定区域进行识别、定位、标识功能，并具备对特定的敏感信息进行擦除的能力，确保数据使用人员无法利用脱敏后的电力图像识别人员、设备、线路等敏感数据。

电力图像样本脱敏应具备图像样本静态脱敏和动态脱敏的功能。对电力图像进行静态脱敏，保障电力图像脱敏样本可以多次使用。对电力图像进行动态脱敏，具备用户直接访问查看图像样本数据的能力，便于电力图像样本的运维管理等。对于电力OCR样本脱敏，应支持对电力OCR数据中包含的内容，明确其中敏感信息，同时，脱敏后保持原有数据特征，支持多次脱敏之间的数据一致性，以及保持脱敏数据的业务关联性，保证在开发、测试以及数据分析类业务不会受到脱敏的影响。

* 1. 数据格式要求

电力图像样本格式应为JPG、JPEG、JPEG2000、BMP、PNG常用格式。

* 1. 性能要求

电力图像样本脱敏应具备以下性能：

a)擦除区域和实际敏感数据区域的交并比应满足75%~90%；

b)电力图像样本脱敏速度应满足4张/秒；

c)采用人工定性评价的脱敏成功率不低于95%；

d)利用测试工具评估脱敏后数据对OCR组件系统识别功能无影响，性能指标相对于脱敏前下降不超过10%，评价方法参见9.5节。

* 1. 脱敏方法

电力图像样本脱敏宜根据技术成熟度和脱敏原则等要求选择脱敏处理方法，涉及的脱敏方法见表2。

表2 电力图像样本脱敏方法

|  |  |  |  |
| --- | --- | --- | --- |
| 图像脱敏方法类别 | 脱敏方法子类 | 描述 | 应用场景 |
| 加密技术 | 混淆 | 使用排序，循环移位，Arnold变换，幻方变换等方法，打乱电力图像样本中像素值原来所在的位置。 | 适用于电力图像样本各设备、各单位间传输 |
| 扩散 | 使用异或运算方法，在不改变像素点位置的情况下，将任一图像明文像素点的信息隐藏在尽可能多的密文像素点中。 | 适用于电力图像样本各设备、各单位间传输 |
| 灵活编码 | 用固定字母和固定位数的数字替代需要脱敏的文本。 | 合同识别中的编号识别 |
| 掩码屏蔽 | 掩码屏蔽是针对账户类数据的部分信息采用字符串代替的方式让敏感数据脱敏。 | 增值税发票识别中购买方地址及电话等 |
| 置乱 | 对敏感数据列的值进行重新随机分布，混淆原有值和其他字段的联系。 | 获奖证书识别中颁发单位等信息 |
| 替换技术 | 统一色块替换 | 使用纯色块如纯黑色来替换原图敏感区域像素级别数据，确保擦除后的数据信息不可逆和不可复原。 | 适用于电力视觉模型训练、样本查看展示 |
| 无效化 | 对字段数据值进行截断、加密、隐藏等方式让敏感数据脱敏。 | PDF转word中涉密字段信息 |
| 近似 | 采用相同含义的数据替换原有的敏感数据，例如身份证信息脱敏后仍然为有效的身份证信息。 | 身份证关键字段脱敏 |
| 生成技术 | 添加噪声 | 对电力图像样本添加随机值到敏感区域或整幅图像，同时尽可能保持脱敏后图像保持原始特性。 | 适用于图像样本统计分析 |
| 微聚集 | 用算法计算出图像样本敏感区域的平均值，并替代整个敏感区域的去标识化技术。 | 适用于电力视觉模型训练、样本查看展示 |
| 模糊技术 | 模糊、虚化、马赛克 | 对电力图像样本敏感区域进行像素级处理，比如模糊、虚化、打马赛克、均值模糊、高斯模糊等算法。 | 适用于电力视觉模型训练、样本查看展示 |

* 1. 脱敏效果评价

对电力图像样本（包括ocr）脱敏效果的评价，宜随机抽取不少于10%或100条电力图像脱敏样本作为测试集进行脱敏，执行人工定性评价。无损脱敏需进一步执行程序量化评价：

1. 人工定性评价：由评价人员对电力图像脱敏样本是否隐藏了敏感隐私信息进行定性评价。人工评价包括两部分。

一是从脱敏后的样本中随机抽取10%的电力图像样本，查看脱敏失败的样本占抽取样本的比例：

R1=A1/B1

其中，A1为脱敏失败的样本量，B1为随机抽取的样本量。

二是计算脱敏完成后的数据区域占敏感区域的比例，如下图所示：

R2= C2/B2

其中，B2为电力图像中敏感区域面积，C2为脱敏完成的区域。



b)程序量化评价：面向电力视觉模型训练和检测等样本应用任务，采用基于深度学习的评价程序，分别对脱敏前、脱敏后电力图像样本执行对应任务，获得任务的公共评价指标计算得到的结果差值，即为程序量化评价结果。

C)测试工具评价：针对ocr图像类样本，在执行脱敏工作基础上，利用测试工具评估脱敏后数据对OCR组件系统功能、性能等方面的影响，并根据验证情况不断优化脱敏规划。

1. 电力文本样本脱敏处理
	1. 功能要求

电力文本脱敏应具备对各类电力文档中的敏感文本内容按照脱敏规则进行数据变形的能力，以确保电力文本敏感隐私数据得到可靠保护，具体应满足以下要求：

a）应支持对包括但不限于UTF-8、GBK、GB2312等主流编码的文本进行脱敏；

b）应支持对包括但不限于中文、英文等电力文档主要语言文本进行脱敏；

c）应提供无损、有损的电力文本脱敏方法。对于无损脱敏，应确保电力文本脱敏后保留原始语言特征分布，以实现文本数据标注信息对脱敏前后的文本都有效。选项至少包括脱敏后文本长度相等、脱敏后文本词性相同、脱敏后文本语言相同；

d）应支持敏感文本的定义与自动识别，至少提供基于关键词匹配的敏感文本识别、基于正则表达式匹配的敏感文本识别；

e）应支持指定敏感文本的基本识别单位。至少提供基于分词的识别以及基于字符的识别。其中，基于分词的脱敏要求内置分词功能，以识别的分词为单位进行脱敏。基于字符的脱敏要求能对文本中连续出现的字符串进行脱敏。

* 1. 数据格式要求

电力文本数据格式应满足以下要求：

a）编码格式：包括但不限于UTF-8、GBK、GB2312等主流编码的文本；

b）语言：包括但不限于中文、英文等电力文档主要语言文本；

c）存储方式：包括但不限于TXT、CSV、xls、xlsx、xml等常见格式。

* 1. 性能要求

电力文本脱敏的性能应满足以下要求：

a）采用人工定性评价的脱敏成功率不低于97%。评价方法参见10.6节。

b）对于无损脱敏，需对脱敏后电力文本原始语言特征分布保留情况进行度量。对脱敏后的文本执行脱敏效果评价，性能指标相对于脱敏前下降不超过10%。评价方法参见10.6节。

* 1. 脱敏方法

电力文本脱敏主要技术手段包括：掩码技术、扰乱技术、有损技术等。其中掩码是指通过隐藏数据中部分信息的方式来对原始数据的值进行转换，又称为隐藏技术。扰乱是指通过加入噪声的方式对原始数据进行干扰，以实现对原始数据的扭曲、改变，扰乱后的数据仍保留着原始数据的分布特征。有损是指通过损失部分数据的方式来保护整个敏感数据集，适用于数据集的全部数据汇总后才构成敏感信息的场景。

宜根据电力文本数据的安全敏感性、使用风险、使用场景选择对应的脱敏方法。具体电力文本脱敏方法参见表3 。

表3 电力文本样本脱敏方法

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 脱敏方法 | 方法描述 | 示例 | 主要使用场景 |
| 1 | 掩码 | 用通用字符替换原始数据中的部分信息， 掩码后的数据长度与原始数据一样。 | 将用电地址“北京市丰台区太平桥街道丽泽路20号院1号楼”经过掩码得到“北京市丰台区太平桥街道丽泽路\*\*\*号院\*\*\*号楼”。 | 开放、共享等 |
| 2 | 规整 | 将数据按照大小规整到预定义的多个档位。 | 将用户用电量按照规模分为高、中、低三个级别，将用户用电量数据用这三个级别代替。 | 数据分析、开发测试等 |
| 3 | 扰乱 | 利用加密、重排等方式对原 始数据进行修改。保留数据 原始特征，并能经过业务校 验。 | 将用户姓名“张三”变为“ZS”。 | 开发测试等 |
| 4 | 数值随机 | 用随机生成某个范围的一个值去替换原始值的一种脱敏算法。 | 将用户电费金额1352乱序为5123。 | 共享、开放、开发测试等 |
| 5 | 均化 | 针对数值性的敏感数据，在保证脱敏后数据集总值或平均值与原数据集相同的情况下，改变数值的原始值。 | 将用电量数据7、4、9、6均化为9、3、8、7。 | 数据分析、开发测试等 |
| 6 | 数据截断 | 直接舍弃业务不需要的信息，仅保留部分关键信息。 | 将用户手机号码 13500010001截断为 135。 | 开发测试等 |

* 1. 脱敏效果评价

对电力文本的脱敏效果评价，宜随机抽样不少于10%或100条文本内容，执行人工定性评价。无损脱敏需进一步执行程序量化评价：

1. 人工定性评价：由评价人员依据有效性、可用性原则对抽样的电力文本样本的脱敏结果进行评价。在有效性方面，重点对电力样本在脱敏处理之后是否已经移除所包含的敏感信息进行评价，超过99%文本实现脱敏结果的正确移除则评价结果为有效，否则评价结果为无效；在可用性方面，重点对脱敏后的电力样本是否保持原有的特征进行评价，超过90%文本保留了原始样本中的有用信息则评价结果为可用，否则评价结果为不可用。采用以下公式进行计算：

R3=A3/B3

其中，A3为脱敏失败的文本样本内容条数，B1为随机抽取的文本样本条数。

b）程序量化评价：面向自然语言处理的信息抽取、文本分类等标准任务，采用基于深度学习及预训练语言模型技术路线的评价程序，分别对脱敏前、脱敏后电力文本执行对应任务，获得任务的公共评价指标计算得到的结果差值，即为程序量化评价结果。

1. 电力语音样本脱敏处理
	1. 功能要求

电力语音脱敏应具备对各类电力语音中的敏感内容按照脱敏规则进行数据变形的能力，以确保电力语音敏感隐私数据得到可靠保护，具体应满足以下要求：

应支持对包括但不限于WAV、AMR、M4A等主流格式的音频文件进行脱敏。

* 1. 数据格式要求

电力语音数据格式应满足以下要求：

应为音频文件格式：包括但不限于WAV、AMR、M4A等主流格式。

* 1. 性能要求

电力语音脱敏的性能应满足以下要求：采用人工定性评价的脱敏成功率不低于97%。

* 1. 脱敏方法

电力语音样本脱敏宜根据技术成熟度和脱敏原则等要求选择脱敏处理方法，涉及的脱敏方法见表4。

表4 电力语音样本脱敏方法

|  |  |  |  |
| --- | --- | --- | --- |
| 语音脱敏方法类别 | 脱敏方法子类 | 描述 | 应用场景 |
| 替换技术 | 统一语音替换 | 对电力语音敏感音频用统一音频进行替换，确保替换后的数据信息不可逆和不可复原。 | 电力智能客服场景 |
| 生成技术 | 添加噪声 | 对电力语音样本添加噪声到敏感区域，同时尽可能保持脱敏后语音保持原始特性。 | 电力会议录音转写场景 |

* 1. 脱敏效果评价

对电力语音的脱敏效果评价，宜随机抽样不少于10%或100条语音内容，执行人工定性评价。

a)人工定性评价：由评价人员对电力语音脱敏结果是否隐藏了敏感隐私信息进行定性评价，计算脱敏失败的样本占抽取样本的比例，可采用以下公式进行计算：

R4=A4/B4

其中，A4为脱敏失败的语音样本内容条数，B4为随机抽取的语音样本条数。

b)程序量化评价：面向语音识别、语音转写等标准任务，采用基于深度学习技术路线的评价程序，分别对脱敏前、脱敏后电力语音执行对应任务，获得任务的公共评价指标计算得到的结果差值，即为程序量化评价结果。

1. （资料性）

表A.1 敏感样本定义

|  |  |  |
| --- | --- | --- |
| 敏感样本分类 | 数据类型 | 分类 |
| 电力图像样本 | 输电巡检图像类 | 包含杆号牌、相序牌、巡检工单等敏感信息样本。例如：“110kV珞洪一回 09# ”中的地点线路信息。 |
| 变电巡检图像类 | 包含设备铭牌、箱体铭牌、巡检工单等敏感信息类样本。例如：“220kV #10 主变断路器 #682断路器 A相”中的设备和相序信息。 |
| 配电领域图像类 | 包含设备铭牌、箱体铭牌、杆号牌等敏感信息类样本。例如：“低压成套开关设备 额定电压 AV380V 防护等级 IP30”中的防护等级信息。 |
| 安监领域图像类 | 包含人脸、车牌、证件照、工作照、工作单等敏感信息类样本。例如：作业现场布控球拍摄的作业人员人脸数据。 |
| 营销领域图像类 | 包含电表信息、用户信息等敏感信息类样本。例如：电表中用户用电数值。 |
| 财务图像类 | 包含混合票据、火车票、报销单据、发票等票据敏感信息类样本。例如：报销单据中的客户名称、开户银行及账号。 |
| 电力文本样本 | 名称类 | 包含客户编号、客户名称、网站账户、联系人姓名等敏感信息类样本。例如：“\*\*市\*\*区\*\*小区\*\*栋\*\*室，业主名\*\*，用电量\*\*，联系电话\*\*”中的用户名。 |
| 地址类 | 包含用电地址、增值税注册地址、联系人地址、法人地址、等敏感信息类样本。例如：“\*\*市\*\*区\*\*小区\*\*栋\*\*室，业主名\*\*，用电量\*\*，联系电话\*\*”中的用户地址。 |
| 联系类 | 包含联系电话、电子邮箱、微信账号、QQ号等敏感信息类样本。例如：“\*\*市\*\*区\*\*小区\*\*栋\*\*室，业主名\*\*，用电量\*\*，联系电话\*\*”中的联系电话。 |
| 证件类 | 包含居民身份证号、军人证号、护照号、台胞证号、驾驶证号等敏感信息类样本。例如：“用户身份证号：\*\*\*\*\*\*”中的用户身份证号码。 |
| 资产类 | 包含车牌号、车架号 等敏感信息类样本。例如：员工私人车辆在系统里的登记车牌号。 |
| 金融类 | 包含银行卡号、存折账号、增值税税号、增值税账号等敏感信息类样本。例如：用电客户银行卡号“12345678901234567890”。 |
| 数据类 | 包含电力行业发电、配电、输电、变电、用电和调度等环节的发电数据、运行控制数据、电力设备参数、用户用电量等敏感信息类样本。例如：电力设备产品型号S11-xx/xx，额定容量50，损耗90（空载），损耗600（负载），外形尺寸955\*450\*1010等。 |
| 电力语音样本 | 智能客服通话录音类 | 包含用户名、用户号、身份证号、手机号等敏感信息类样本。例如：“查询户号为0306573749的电费余额”中的用户号信息。 |
| 会议录音类 | 包含人名、项目金额、身份证号、手机号等敏感信息类样本。例如：“明年项目预算金额为2000万元”中的金额信息。 |