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

电力人工智能样本存储技术要求

Technical requirements for sample storage of power artificial intelligence

（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　　次

前　言

本文件按照 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。

本文件由中国电工技术学会提出。

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：国家电网有限公司大数据中心、国网信息通信产业集团有限公司、中国电力科学研究院有限公司、国网智能电网研究院有限公司、安徽继远软件有限公司、国网福建省电力有限公司。

本文件主要起草人：李强、赵峰、邱镇、陈振宇、李博、刘识、王晓辉、李炳森、黄晓光、王晓东、秦余、张琳瑜、张国梁、白景坡、张晓航、崔冬梅、刘璟、靳敏、郭鹏天、李道兴、余江斌、郭庆、浦正国、薛濛、黄旭东、聂文萍、刘晓飞、刘健、李扬笛、林爽、杨彦。

本文件为首次发布。

电力人工智能样本存储技术要求

1 范围

本文件规定了电力行业人工智能样本包含图像、文本、音频电力样本处理技术中样本存储技术总体架构、基本要求和各项技术指标。

本文件适用于电力行业人工智能平台样本存储的建设、管理和使用。

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 41867-2022 信息技术 人工智能 术语

GB/T 25000.51-2017 软件工程 软件产品质量要求和评价体系(SQuaRE) 质量测量框架

Q/GDW 12118—2021 人工智能平台架构及技术要求

T/CES 129-2022 电力人工智能平台样本规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

样本存储系统 Sample storage system

指实现样本存储技术的软硬件系统，包括样本数据库、样本文件系统、样本元数据管理系统等。

3.2

样本存储效率 Sample storage efficiency

指样本存储系统在存储和访问样本数据时所消耗的时间、空间和资源的指标。

3.3

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

3.4

文件格式 file format

存储介质对存储信息制定的编码方式，用于识别内部储存的资料。

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。

4.1 符号

无

4.2 代号

无

4.3 缩略语

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

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

5 电力人工智能样本存储总体架构

电力人工智能样本存储技术总体架构包括：

a)样本数据，指用于电力人工智能训练和应用的原始数据，包括结构化数据和非结构化数据。非结构化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。样本数据需要被存储在一个高性能、高可靠、高可用的样本文件系统中，以便于快速地读取和处理。

b)样本元数据，指对样本数据的描述性信息，包括但不限于编号、类型、来源、标签、质量、时间、位置等内容。每个样本数据都应该有一个唯一的编号，用于标识和关联。样本元数据需要被存储在一个高效、安全、灵活的样本数据库中，以便于进行增删改查等操作。

c)样本元数据管理系统，指用于实现样本数据库和样本文件系统之间的关联和协同的系统，主要有三个功能：通过样本编号建立一一对应的关系，通过样本元数据进行检索和定位，通过样本元数据进行同步和更新。这个系统需要保证数据的一致性、完整性和可用性。

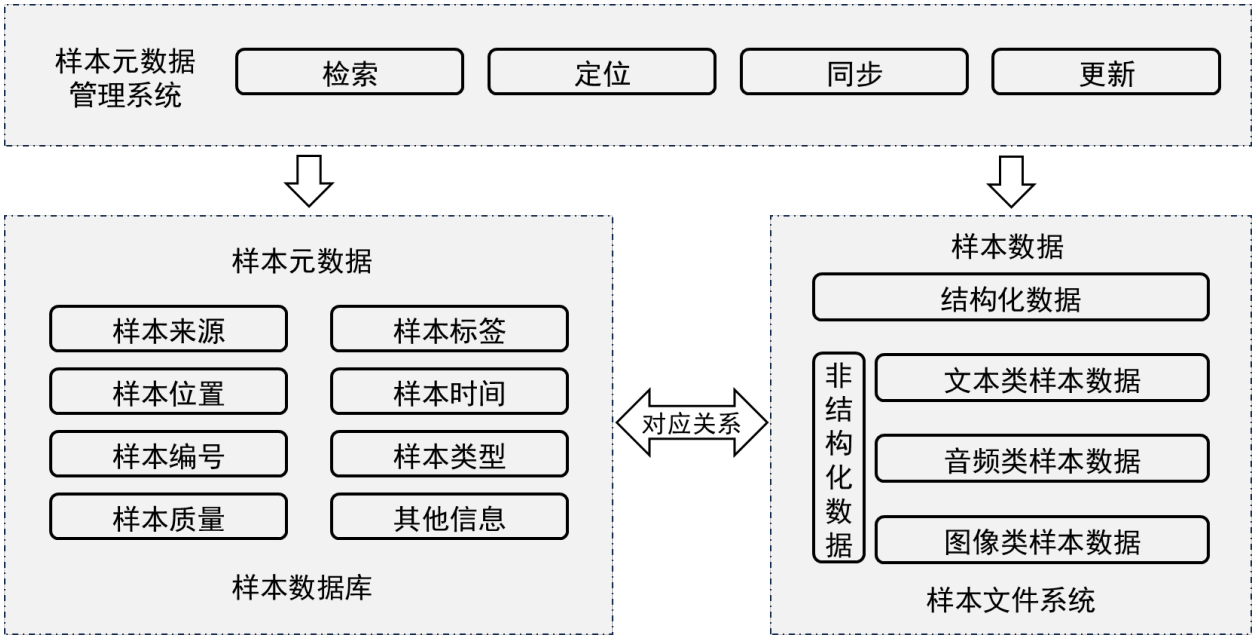

图1　电力人工智能样本存储技术总体架构图

6　样本存储技术基本要求

6.1　样本数据格式

样本数据应采用统一的文件格式进行存储，以便于后续的处理和分析。原则上同批次样本文件中，单个文件最大不得超过该批次样本文件平均大小的 200%，最小不得小于平均大小的 50%。各类型样本数据具体格式如下：

a) 文本类样本数据应采用 TXT、JSON、XML、CSV 等常见的文本文件格式进行存储，每个文件应包含一段或多段文本。

b) 音频类样本数据应采用 WAV、MP3、WMA、WAV、APE、FLAC、OGG、AAC 等常见的音频文件格式进行存储，每个文件应包含一段或多段音频。

c) 图像类样本数据应采用 JPEG、PNG、BMP、SVG、WEBP、EPS 等常见的图像文件格式进行存储，每个文件应包含一个图像。

d) 视频类样本数据应采用 MP4、M4V、WEBM、MOV、AVI、DIV 等常见的视频文件格式进行存储，每个文件应包含一段视频。

6.2　样本元数据

样本元数据是对样本数据的描述性信息，包括但不限于以下内容：

a) 样本编号：唯一标识每个样本数据的编码。

b) 样本类型：表示样本数据属于图像、音频或文本类。

c) 样本来源：表示样本数据的获取方式和来源渠道。

d） 样本标签：表示样本数据所属的类别或属性。
e） 样本质量：表示样本数据的清晰度、完整度、有效性等质量属性。
f） 样本时间：表示样本数据的采集或生成时间。
g） 样本位置：表示样本数据与电力系统中的设备或场景的关联位置。
h） 其他信息：根据不同的应用场景，可以增加其他与样本数据相关的信息。

6.3 样本数据库

样本数据库是用于存储和管理样本元数据的数据库系统，应具备以下功能：
a） 支持对样本元数据进行增、删、改、查等基本操作。
b） 支持对样本元数据进行分类、分组、排序、筛选等高级操作。
c） 支持对样本元数据进行备份、恢复、迁移等维护操作。
d） 支持对样本元数据进行安全、权限、审计等管理操作。

6.4 样本文件系统

样本文件系统是用于存储和管理样本数据的文件系统，应具备以下功能：
a） 支持对样本数据进行存储、读取、删除等基本操作。
b） 支持对样本数据进行压缩、加密、解密等高级操作。
c） 支持对样本数据进行备份、恢复、迁移等维护操作。
d） 支持对样本数据进行安全、权限、审计等管理操作。
e） 支持对样本数据进行格式转换操作。

6.5 样本元数据管理系统

样本元数据管理系统是用于实现样本数据库和样本文件系统之间的关联和协同的系统，应具备以下功能：
a） 支持通过样本编号在样本数据库和样本文件系统之间建立一一对应的关系。
b） 支持通过样本元数据在样本数据库和样本文件系统之间进行检索和定位。
c） 支持通过样本元数据在样本数据库和样本文件系统之间进行同步和更新。

7 样本存储技术技术指标

7.1 样本存储容量

指样本存储系统能够存储的最大样本数据量，单位为 GB 或 TB；原则上应大于现有样本量，同时为满足后续使用，应按实际情况预备适宜富余量。

7.2 样本存储速度

指样本存储系统在存储或读取样本数据时的平均速度，单位为 MB/s 或 GB/s；原则上平均读写速度下限为 100MB/s，不设上限。

7.3 样本存储可靠性

指样本存储系统在正常运行条件下，能够保证样本数据不丢失、不损坏、不篡改的概率，单位为%；原则上应为 100%。

7.4 样本存储可用性

指样本存储系统在正常运行条件下，能够正常响应用户请求的概率，单位为%；原则上应大于 80%，尽可能达到 100%。

7.5 样本存储安全性

指样本存储系统在正常运行条件下，能够保证样本数据不被非法获取和篡改的概率，单位为%；原则上应为 100%。

7.6 样本存储时效性

对于部分存在时效性限制的样本，应依据实际场景和使用要求制定合适的时效限制，在选取样本时应选取在规定可用时效限制内的样本，从而保证所选取的样本的时效性以及实际训练任务的顺利推进。

参 考 文 献

[1] GB/T 41867-2022 信息技术 人工智能 术语
[2] GB/T 25000.51-2017 软件工程 软件产品质量要求和评价体系(SQuaRE) 质量测量框架
[3] Q/GDW 12118—2021 人工智能平台架构及技术要求
[4] T/CES 129-2022 电力人工智能平台样本规范