ICS 号
中国标准文献分类号

CES XXX

团 体 标 准

T/CES XXX-XXXX

电力系统的大语言模型微调数据准备规范

Data Preparation Requirements and Standards for Large Language

（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：国网信息通信产业集团有限公司、国家电网有限公司大数据中心、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、北京国网信通埃森哲信息技术有限公司、四川中电启明星信息技术有限公司、国网福建省电力有限公司

本文件主要起草人：李强、赵峰、赵永生、邱镇、陈振宇、李博、刘识、李炳森、黄晓光、秦余、王晓东、张琳瑜、张国梁、刘园园、崔迎宝、王兴涛、卢大玮、吴迪、赵如意、宋卫平、杨帆、高攀、王红蕾、董梅、李欢欢、徐小云、叶林峰、赵林林、王誉博、李扬笛、杨彦、林晨翔

本文件为首次发布。

1 范围

本文件适用于电力系统的大语言模型的数据准备，包括数据的采集、清理、标注、注释和整理等所有环节。此标准的目的在于规范在大语言模型训练中的数据处理过程，以保证数据的可用性、一致性和可追溯性。本文件规定了人工智能大语言模型在电力系统中的微调数据准备规范，本文件共分为数据准备要求、数据准备规范、数据准备流程等。

本文件适用于各单位使用大语言模型技术解决相关业务需求，适用于电力系统人工智能大语言模型的应用开发等业务场景，帮助业务人员以及开发人员完成模型训练、模型微调等相关工作。

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

GB/T 41867-2022 信息技术人工智能术语

3 术语和定义

下列术语和定义仅适用于本文件。

3.1 人工智能 Artificial Intelligence

人工智能是一门交叉学科，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。

3.2 大语言模型 Large Language Model

大语言模型也称大型语言模型，是一种人工智能模型，旨在理解和生成人类语言。在大规模文本语料上训练、包含百亿级别（或更多）参数的语言模型。

3.3 指令微调 Instruction Tuning

指令微调是指可以帮助大语言模型实现人类语言指令遵循的能力，在零样本设置中泛化到未见任务上的学习方法。

3.4 数据准备 Data Preparation

指的是将原始数据进行清洗、转换、标记和结构化以适用于大语言模型的过程。

3.5 数据源 Data Sources

数据源指的是用于训练和应用大语言模型的原始数据，包括但不限于文本、图像、报告和传感器数据。

3.6 **数据预处理** Data Preprocessing

数据预处理指的是在得到原始数据之后对数据进行预处理，包括数据清洗、去重、去噪以及数据标准化等步骤。

4 缩略语

下列缩略语适用于本文件。

Json：JS对象简谱(JavaScript Object Notation)

BOM：字节顺序标记(Byte Order Mark)

5 总则

本文件规定了人工智能大语言模型在电力系统中的微调数据准备规范，本文件共分为数据准备要求、数据准备规范、数据处理流程等。其中数据准备主要用于规范电力系统的大模型在微调训练中的数据收集、数据格式以及数据隐私与安全等，数据准备规范主要用于规范电力系统大模型微调训练中的数据预处理、数据转换、数据标签与注释、数据及划分、数据格式化、数据集质量评估、数据增强以及数据更新与维护等，数据处理流程主要用于规范数据预处理的一般步骤和中文数据预处理的步骤等。具体内容组织框架见图1:

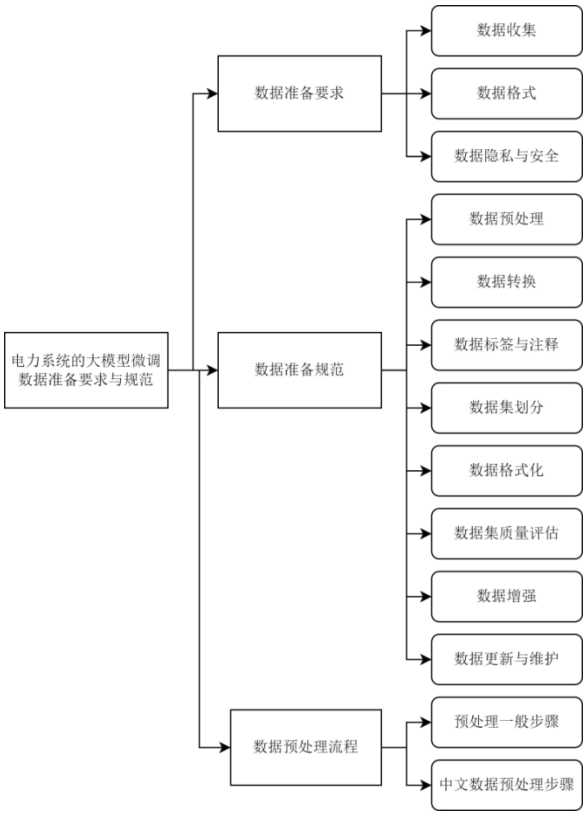

图 1 组织框架

6 数据准备规范

6.1 数据收集

本文件主要从数据来源、数据多样性与数据质量三个方面对数据收集过程进行相关的规范性要求，确保大语言模型微调技术应用过程中训练数据符合要求。

（1）数据来源

电力系统的数据要求是通过传感器、智能设备、视频监控设备、音频通信设备、移动终端等进行数据采集，收集海量结构化、半结构化、非结构化的业务数据集合。在电力系统中，大语言模型的微调与训练所使用的数据应来自可靠和权威的电力系统数据源，包括电力公司、政府部门和独立研究机构等。

（2）数据多样性

在大语言模型训练过程中所使用的数据要求应涵盖电力系统各个方面，包括发电、输电、配电、设备状态、市场数据和电力负荷等，需要根据具体的业务场景及需求，保证数据的多样性和丰富性。

（3）数据质量

大语言模型训练中应过滤低质量数据，保证数据的准确性和一致性，可分为两类方法：基于分类器的方法和基于启发式的方法。

6.2 数据格式

大语言模型微调训练中应对多样化的原始数据集进行对齐，本文件主要从数据结构化和数据标注两个方面进行要求。

（1）数据结构化

数据应以适当格式进行结构化，便于模型的理解和分析。电力系统中的数据来源复杂多样，应对收集到的数据进行结构化处理，使用统一的格式标准对数据进行结构化处理，使得大语言模型微调过程中能够更好的训练，保证模型训练的效果。

（2） 数据标注

在电力系统大模型训练中应对数据中的重要信息进行标记和注释，帮助模型理解数据的语境和含义。可采用的方法有众包、半监督、主动学习以及弱监督等，其中众包是人工标注，半监督方法指利用部分标注数据训练一个分类器等辅助标注更多的数据，主动学习方法指先从每次选出模型任务最难的样本中进行人工标注再接着训练，然后进行多次迭代，弱监督方法是设计一种标签函数，通常基于启发式。

6.3 数据预处理

大语言模型的数据清洗应包括：去除噪声数据、去除重复数据、统一标号、缺失值处理、语言检测和文本语言标准化等。数据清洗的具体步骤和技巧根据具体项目和业务需求导致数据的要求而有所不同。电力系统的大模型微调训练中要求去除不需要的数据，修复数据集中的缺失值或错误，处理异常数据和噪声，将数据转换为统一的格式和单位，保证数据质量，避免对模型的干扰，提高模型训练的效率。在清洗数据时，应进行反复测试和验证。

6.4 数据转换

数据转换应将电力系统的数据转换成统一的、适合模型使用的形式，保证数据的一致性和可用性。应包括数据编码和数据归一化，数据编码是将数据进行编码，数据归一化是将数据进行归一化处理。

6.5 数据标签与注释

电力系统的大语言模型微调中应为数据添加标签和注释，包括人工标注法和自动标注技术。其中自动标注技术可通过机器学习算法自动给数据添加标签，常用的有实体识别、事件标注等。实体识别是标记电力系统中的关键实体，如设备、线路、电力站等，事件标注是标记电力系统中的关键事件，如故障、维护、市场活动等。

6.6 数据集划分

对数据数据集进行划分要求如下：

（1）将数据集划分为训练集，验证集和测试集三个数据集；

（2）使用交叉验证来评估模型的性能

（3）通过分层抽样保证每个类别的数据在三个测试集中具有代表性，避免数据偏差。

（4）训练数据和验证数据集由输入和输出实例组成，这些实例表示模型如何执行。使用的训练和验证数据必须采用JSON（JSONL）文档格式，其中每一行代表一个{prompt-completion}对。

训练数据的格式实例：

{"prompt": "<prompt text>", "completion": "<ideal generated text>"}

{"prompt": "<prompt text>", "completion": "<ideal generated text>"}

{"prompt": "<prompt text>", "completion": "<ideal generated text>"}

除 JSONL 格式外，训练和验证数据文件必须以 UTF-8 编码并包含字节顺序标记(BOM)，并且文件大小必须小于 200 MB。

6.7 数据格式化

数据格式化应根据所选择的模型，将数据格式化为适合电力系统大语言模型接受的标准输入格式，包括文本编码、图像的张量化等。

6.8 数据集质量评估

在电力系统的大语言模型微调训练中应评估数据集的质量，确保数据的准确性、一致性和完整性。应对数据集进行及时更新与维护，保证数据集中的数据具有良好的时效性，便于不同版本的记录。

6.9 数据增强

在准备电力系统大语言模型微调数据中应通过增加数据样本和多样性来提高模型性能，解决数据不平衡问题，增加小类别的样本，可采用的方法包括Mixup、AutoAugment以及利用GAN生成新样本等等。

6.10 数据更新与维护

电力系统中大语言模型微调中应及时对数据进行更新和维护，确保数据的时效性和可靠性。在数据准备过程中应确定数据更新频率，规定数据来源（包括实时监测、传感器、数据库等），建立数据监控系统，确保及时处理数据异常变化情况，需要记录数据更新与维护的所有操作，建立清晰的数据维护历史。

6.11 数据隐私和安全

（1）隐私保护

电力数据可能包含大量的敏感信息，如用户的用电量、电费等，这些信息需要得到严格的保护。因此对大语言微调数据准备过程中涉及到的隐私数据要求如下：

a) 应对敏感信息存储和记录，对所记录的数据进行访问时要设定权限严格管控，以达到防止对数据进行未经授权的访问和数据泄露等安全问题的产生。

b) 应大语言模型设置数据安全标准等级规定，保障电力系统中大量的实时监测和控制系统的稳定运转。

（2）安全性

结合电力数据的有关特性，对数据安全性方面作出如下要求：

a) 应数据应存储和传输于安全的环境中，防止数据泄露和滥用。

b) 大语言模型构建过程中应告知使用时收集用户数据的范围，提供数据信息撤销、清除等操作。

c) 针对电力系统中数据的特性，要求大语言模型规范数据安全等级。

d) 在对大模型进行微调训练过程中，要求对数据进行全面的安全性评估，以确保大语言模型在电力领域的应用满足相关法规和安全要求。

7 大语言模型微调数据准备流程规范

本文件规定了在电力系统中通用的大语言模型微调数据准备流程规范，在应用过程中应该结合具体的业务要求及应用场景适时调整。具体的流程见图2：

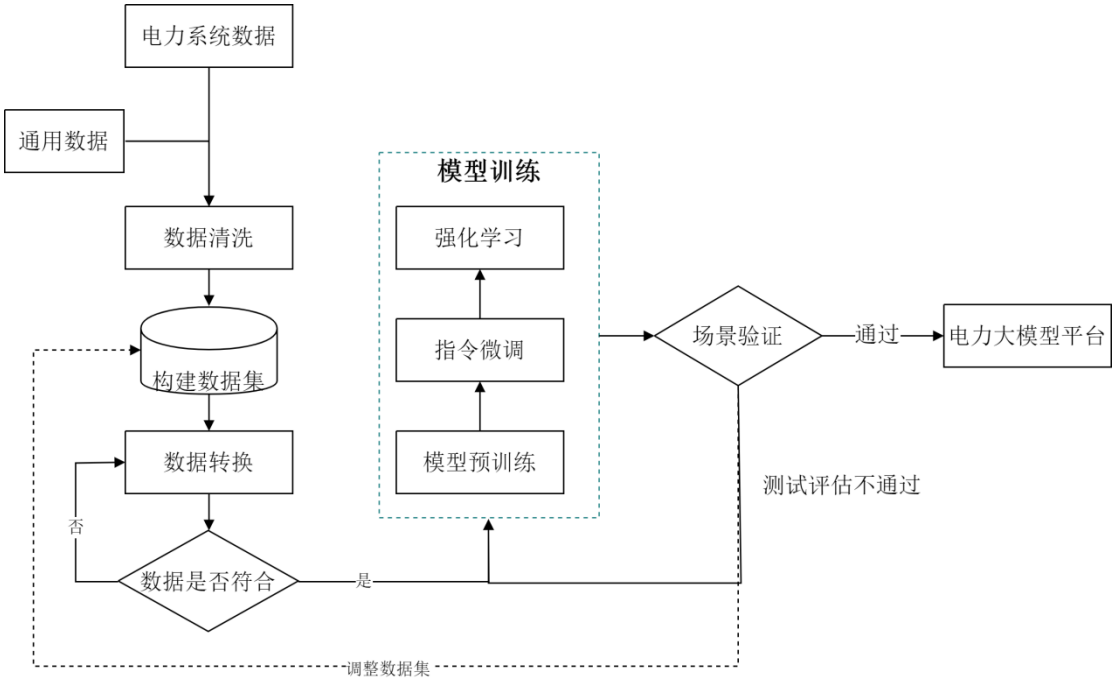

图 2 电力系统中通用的大语言模型微调数据准备流程

（1）数据收集

电力系统大语言微调数据的具体要求收集电力系统的原始数据，包括但不限于发电、输电、变电、配电、用电和调度等各环节的数据，以及能源数据、天气数据等多类型数据。

（2）数据清洗与数据集构建

大语言模型对训练数据的具体要求包括：

a）对已收集的电力系统原始数据进行筛选、标注和整理；

b）开展数据预处理，如去除无关、重复、错误、低质量的数据等，有效减少训练数据中的噪声和偏差，提高大语言模型关于电力系统场景数据的学习能力和泛化能力；

c）根据具体的应用场景和实际需求，构建微调指令集以及强化学习训练集，提高数据的质量和一致性，完成高质量的电力系统数据集构建。

（3）数据转换

对微调数据需要进行数据转换，应将清洗后的电力系统数据集转换为机器能够识别的、适合输入到大语言模型训练加载的数据格式，如向量等，使数据与模型的结构和需求相匹配，以提高模型训练的效率与性能。

（4）模型训练

在电力系统中应用大语言模型进行微调训练，应使用不同规模的训练数据来调整大语言模型的参数，使其能够完成特定的任务。模型训练可分为预训练、指令微调和强化学习等三个阶段。

（5）场景验证

应该对大语言模型在电力系统的实际应用场景中的表现进行测试和评估，应检验其是否满足具体的业务场景及实际需求的预期目标和标准。场景验证包括功能测试、性能测试、安全测试、可用性测试等方面。

（6）构建电力系统大语言模型平台

建立电力系统专用的电力系统大语言模型平台，应其随着电力系统数据及产品更新换代提供更好的实时私有化平台，并随时进行平台权限管理。