ICS 27.100
CCS F29

团　　体　　标　　准

T/CES XXX-XXXX

面向电力行业的预训练模型的通用要求

General requirements for pre-trained models for the power industry

2024-12-01 发布　　2025-01-01 实施

中国电工技术学会　发布

目　　次

前　言

本文件按照 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、安徽继远软件有限公司、国网黑龙江省电力有限公司、福建亿榕信息技术有限公司、国网福建省电力有限公司电力科学研究院、国网山西省电力公司、北京工业大学、国网河北省电力有限公司、国网青海省电力公司超高压公司、国网辽宁省电力有限公司。

本文件主要起草人：李强、陶俊、郭庆、梁翀、喻成琛、浦正国、杨彬彬、薛濛、吴小华、周伟、郭力旋、张琳瑜、王晓东、周逸平、王强、李盼盼、刘洁、李净雅、王秋琳、赵峰、余江斌、程琳、张天奇、王雷、刘晓飞、黄晓光、李小宁、李扬笛、袁永科、于涛、彭锐、田广、杨洋、张明理。

本文件为首次发布。

面向电力行业的预训练模型的通用要求

1 范围

本文件定义制备或使用面向电力行业的预训练模型的参考架构，描述了相关方及其活动，并规定了面向电力行业的预训练模型的通用技术要求。

本文件适用于面向电力行业的预训练模型的研究、制备、开发、部署和应用。

2 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。

GB/T 2900.1 电工术语 基本术语

GB/T 5271.34 信息技术 词汇 第 34 部分：人工智能 神经网络

GB/T 41867-2022 信息技术 人工智能 术语

GB/T 42755-2023 人工智能 面向机器学习的数据标注规程

3 术语和定义

GB/T 2900.1、GB/T 5271.34、GB/T 41867-2022 和 GB/T 42755-2023 界定的以及下列术语和定义适用于本文件。

3.1

预训练模型 pre-trained model

一种在广泛领域数据集上训练得到的，供以专门领域数据微调，来满足场景任务需求的深度学习模型。

注：按训练数据模态，预训练模型一般相应体现出对文本、图像、音频或视频等模态任务的处理能力及泛化性。

3.2

预训练模型服务 pre-trained model service

通过应用预训练模型为用户提供价值的方法。

注 1：服务一般满足用户获得特定输出的要求。

注 2：预训练模型服务一般含有：推理服务、微调服务、大模型小型化服务。

3.3

作业 job

一个可被测试系统执行的基本测试单元。

3.4

任务 task

被调度的训练或推理对象。

注：任务用于完成一个相对独立的业务功能。一个任务属于且仅属于一个作业。

3.5

微调 fine-tuning

为提升人工智能模型的预测精确度，一种先以大型广泛领域数据集训练，再以专门领域数据集继续训练的附加训练技术。

注 1：专门领域数据一般指下游任务数据。

注 2：常用的微调方法包括提示学习微调、全参微调、高效参数微调等。

3.6

提示语 prompt

使用预训练模型进行微调或下游任务处理时，插入到输入样本中的指令或信息对象。

3.7

提示学习 prompt learning

在不修改预训练模型结构和参数的情况下，通过向模型提供含特定任务指示性关键词的提示语，引导预训练模型在特定任务上应用其已有知识达到更好性能表现。

3.8

人工智能加速处理器 artificial intelligence accelerating processor

具备适配人工智能算法的运算微架构，能够完成人工智能应用加速运算处理的集成电路器件。

3.9

大模型 large model

是指具有大量参数和复杂结构的机器学习模型，能够处理海量数据、完成各种复杂的任务，如自然语言处理、计算机视觉、语音识别等。

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。

AI：人工智能（artificial intelligence）。

API:应用程序编程接口（application programming interface）。

FPGA：现场可编程逻辑门阵列 （field programmable gate array）。

GPU：图形处理器（graphic processing unit）。

IOPS：单位时间内系统能处理的I/O请求数量（input/output operations per second）。

LACP：链路聚合控制协议（link aggregation control protocol）。

NPU：神经网络处理器（neural network processing unit）。

PCIe：外设部件互联高速通道（peripheral component interconnect express）。

SDK：软件开发工具包（software development kit）。

TPU：张量处理器（tensor processing unit）。

UML：统一建模语言（unified modeling language）。

5 系统架构

支撑预训练模型的生态包括功能视角下的参考架构和用户视角下各相关方的技术活动。

功能视角下的预训练模型参考架构见图1，包括资源池、工具、数据资源、模型、行业应用和服务平台等。其中：

——资源池包括计算、存储、网络、资源虚拟化及调度等；

——工具包括数据工具、模型工具；

——数据资源包括通用数据、行业数据、私有数据；

——模型包括预训练模型、定制化模型。其中预训练模型包括单模态和多模态两种类型的模型，定制化模型是依据用户需求对预训练模型进行微调定制生产环境所需的模型；

——行业应用为为各行业场景用户提供预训练模型下游任务匹配服务；

——服务平台/组件贯穿各层次提供支持大规模预训练模型和相关服务的编排、部署、模型推理、运维和管理。

图 1 功能视角下的预训练模型参考架构

用户视角下的预训练模型相关方见图 2，包括基础设施提供者、数据提供者、模型提供者、应用服务者、应用消费者和管理者。其中：

——基础设施提供者包括硬件资源提供者和软件资源及工具提供者。硬件资源提供者的活动包括提供计算、存储、网络等支撑硬件服务活动。软件资源及工具提供者的活动包括提供数据处理、计算加速、模型训练、模型优化、模型验证等支撑软件服务活动；

——数据提供者进行数据采集、数据准备、数据管理等数据相关服务活动；

——模型提供者负责模型设计开发、模型预训练、模型验证、模型优化、模型部署等预训练模型相关服务活动；

——应用服务者支持平台服务、模型定制、模型推理、模型运维和管理等应用服务活动；

——应用消费者的活动包括使用模型和相关服务以及提供评估反馈；

——管理者对预训练模型在生态链各环节的安全与合规性进行管理，包括监管、审计、测试评估等活动。

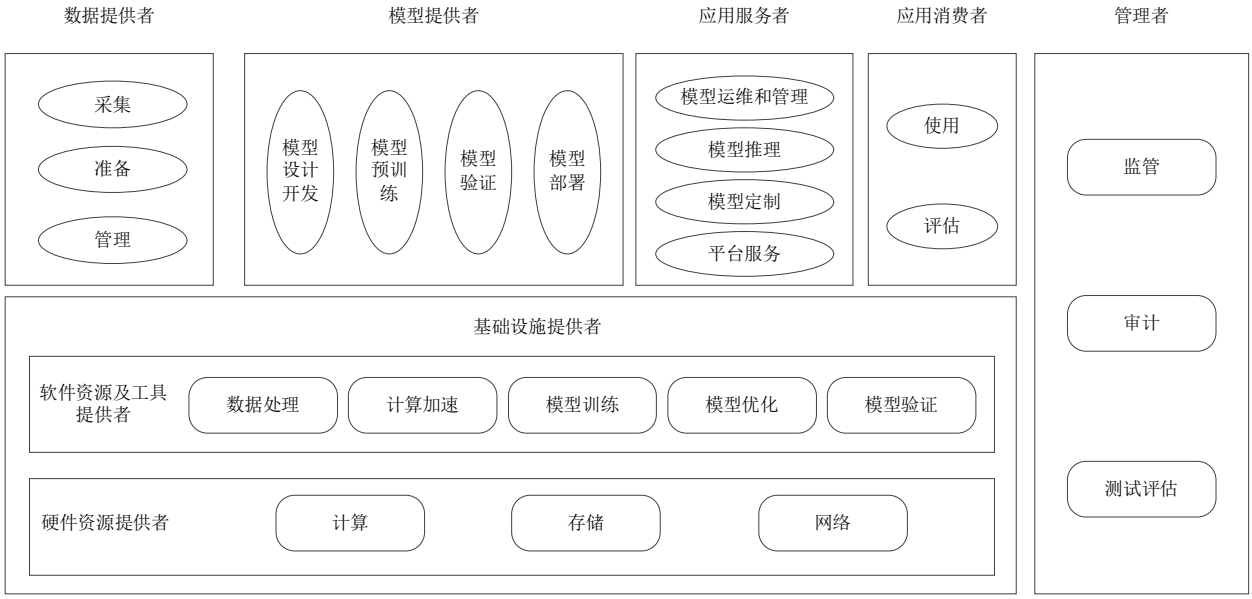

图 2 预训练模型的利益相关方及其活动

6 技术要求

6.1 资源池

6.1.1 计算资源

为模型训练和推理提供计算和数据处理等能力的实体设备（如 CPU、GPU，FPGA，NPU，TPU）或逻辑设备。计算资源应符合以下要求：

a）应能执行至少 1 种模态（如文本、图像、语音、视频）的模型的训练或推理；

b）应支持硬件加速的人工智能计算，配备分布式训练和推理计算加速库；

1）训练服务器：

——应支持不小于 4 个 100GE 网口；

——应支持电源模块、风扇模块的热插拔和备份（如 2+2 冗余，N+1 冗余等）；

2）推理服务器：

——内存总带宽应不小于 800GB/s；

——应支持不小于 2 个 PCIe 扩展槽位；

——应支持电源模块、风扇模块的热插拔和备份（如 1+1 冗余，N+1 冗余等）；

c）服务器集群单位（如机柜）宜配备不小于 128 个人工智能处理器；

d）宜支持基于硬件加速的预处理（如图像、视频编解码）；

e）应支持键值对缓存。

6.1.2 存储资源

适用于大模型训练和推理的存储资源，包含存储服务器等。存储资源用于提供数据存储和模型存储，应符合以下要求：

a）应支持数据集的分布式存储与访问，并实现冗余备份机制；

b）宜支持分布式模型训练及推理；

c）存储带宽宜不小于 200GB/s，IOPS 宜不小于 200 万；

d）宜支持内存计算；

e）宜能以存储服务器或硬磁盘为单元创建存储池，存储池宜能识别、管理固态盘、硬磁盘等不同类型存储媒体。

6.1.3 网络资源

适用于大模型训练和推理的网络资源，包含集群内交换机和路由器。网络资源应符合以下要求：
a）应支持高速网络通信协议（如 100G RoCE 等）
b）应具备模型自动切分（如基于模型结构）；
c）转发包率宜不小于 4000Mp/s；
d）应支持负载均衡；
e）应支持可靠性组网方案，如 LACP 链路聚合，M-LAG 双活等；
f）宜支持服务器集群内 40GE/100GE/200GE/400GE 全联接网络；
g）宜支持物理交换机与逻辑交换机之间的映射，实现链路备份，单台物理交换机故障不影响训练、推理任务执行。

6.1.4 资源虚拟化及调度

虚拟计算资源应符合以下要求：
a）应具备虚拟化的 CPU。
b）应具备一种以上虚拟化的人工智能加速处理器，如 NPU、GPU 等。
c）应能实时监控资源状态。
d）宜能以基于容器的方式，管理异构资源。
e）宜支持资源池内 CPU 和人工智能加速处理器间的不同配比。
f）宜支持基于角色的权限访问控制。
g）宜能自动发现和维护计算资源。
h）宜支持服务器系统与其运行应用的绑定。
i）宜提供资源故障告警、检测和还原的功能。
j）宜能使用硬件节能功能，包括资源回收、关闭和休眠。
k）宜支持资源注册、使用、配额管理和操作审计。
l）宜支持虚拟 CPU 和虚拟人工智能加速处理器的互操作。
m）应支持至少一种深度学习或机器学习框架。
n）应能执行以下至少一种场景模型的推理和训练，包含但不限于：
1）面向电力行业的视觉处理（图像处理、视频处理）；
2）面向电力行业的自然语言处理；
3）面向电力行业的声音处理（包括电力设备、人、动物、雷电、风、雪、雨、冰雹等的声音）
o）宜能训练和推理 n）中的全部类型场景的模型。

6.2 工具

6.2.1 数据工具

6.2.1.1 数据采集工具

数据工具提供数据采集功能，应符合以下要求：

a)应确定数据采集的需求、数量、渠道、所采集数据的类别（如文本，语音、图片和视频等）和范围（如话题、内容等）；

b)应能采集原始数据的类型，包括但不限于文本、视频、图像、音频等；

c)应支持从不同格式的原始数据（如 TXT、JPEG/JPG、MP4、AVI、WMV 等）中提取出模型训练所需的数据；

d)应能记录采集数据的来源、时间和采集方式；

e)应支持结构化、半结构化、非结构化的数据接入；

f)应支持多组数据或多个数据集的并行导入；

g)宜支持数据质量检测和初步清洗能力，如数据格式标准化等。

6.2.1.2 数据准备工具

数据工具提供数据准备功能，应符合以下要求：

a)数据标注流程应符合 GB/T 42755-2023 中第 6 章和第 7 章的要求；

b) 应支持数据清洗，包括文本数据的敏感词与特殊符号过滤、图像数据重建与去模糊、视频与音频数据的特定片段截取等；

c)应支持数据重组、数据标签格式转换；

d)应支持数据检索、分析等功能；

e)应支持数据增强及扩充（如添加扰动产生新数据）；

f)应支持数据质量检验。

6.2.1.3 数据存储工具

数据工具提供使用存储资源的功能，应符合以下要求：

a)应支持分布式并行存储；

b)应支持在线弹性扩展，满足容量需求和性能的线性增长；

c)应支持通过控制台、API、SDK、命令行方式操作存储资源，能按需求切换；

d)应支持标准文件系统接口，如 POSIX；

e)应支持向量库储存。

6.2.1.4 数据管理工具

数据工具提供数据管理功能，应符合以下要求：

a)应支持数据集管理的要素，包含数据集名称、版本、标注类型、标注标签、数据量、数据来源、特征版本、创建时间等；

b)应支持数据集的创建、查询、修改、删除、导入、导出、发布等；

c)应支持数据集状态信息查询，包含数据集名称、版本、标注类型、数据量、导入状态、已标注状态和版本；

d)宜支持数据可视化分析和版本管理。

6.2.2 模型工具

6.2.2.1 模型设计工具

模型设计工具，应符合以下要求：

a)应支持可视化图形界面，允许用户通过拖放、连接元素来创建模型；

b)应支持多种类型的模型设计，例如流程图、UML（统一建模语言）图、概念图等；

c)应提供预定义的模型元素和模板，使用户能够快速构建模型；

d)应支持对模型性能进行模拟和分析，以评估其行为和性能；

e)宜支持导出模型的多维度信息，如说明文档，模型代码等。

6.2.2.2 模型训练工具

模型训练工具，应符合以下要求：

a)应支持数据并行，模型并行，混合并行等分布式训练技术；

b)分布式协同训练集群在训练过程中出现节点故障（如宕机）时，应支持从断点继续并完成训练任务；

c)应能至少使用 2 种数据源或知识库，对训练任务实施集成和迁移；

d)应支持或可通过插件方式支持数据可视化、训练可视化及模型评估可视化；

e)应支持基于训练数据的整体或部分特征，构建预训练任务；

f)应支持模型历史版本和微调迭代过程中的信息记录和查询，信息包含日志，准确率、损失、参数等；

g)应支持预训练模型训练过程及应用日志的留存及获取；

h)宜提供多种并行策略，包括算子切分、算子自动并行、自定义通信算子等。

6.2.2.3 模型优化工具

模型优化工具，应符合以下要求：

a)应支持模型压缩（如剪枝、量化、知识蒸馏等），云服务实现时宜提供调用接口；

b)支持模型微调，包括：

1)应支持的数据类型包含如文本、语音、图像、视频等；

2)应支持任务类型包含单模态、多模态融合等；

3)应提供评价指标体系，包含如准确率、清晰度等；

4)宜支持基于用户反馈的微调（如基于用户反馈的强化学习）。

c)应支持参数有效性学习、混合精度训练（自动精度混合、手动精度混合）等优化训练方法，使用的精度如半精度浮点，四分之一精度整型或单精度浮点等；

d)宜支持检索增强生成功能。

6.2.2.4 模型验证工具

模型验证工具，应符合以下要求：

a)应支持预训练模型的功能（如自然语言处理、图像处理、多模态等）有效性评估；

b)应提供自动化测试功能；

c)应允许用户根据需要自定义测试参数和场景；

d)应能在测试过程中自动检测运行异常情况并提供诊断信息；

e)宜支持模型性能实时监测和日志记录。

6.2.2.5 模型部署与推理工具

模型部署和推理工具，应符合以下要求：

a)应支持的部署方式包含在线部署、批量部署、离线部署等；

b)应支持本地服务器部署，云端部署，宜支持边缘侧和移动端的模型部署；

c)应提供实现机制，支持在满足一定吞吐量条件下的低延时推理；

d)应支持模型推理过程的监控和日志记录；

e)宜支持在至少 1 种推理加速框架上部署模型；

f)宜提供工具链，基于自然语言处理模型、视觉模型、多模态模型、科学计算模型，构建下游任务。

6.3 数据资源

6.3.1 通用数据

通用数据应具有来源多样性、高质量、覆盖面广、完整性和真实性，宜尽量覆盖各类应用场景，确保大模型的训练数据具有高质量和多样性。

6.3.2 行业数据

行业数据应具备行业特征，宜尽量覆盖行业中的使用场景。宜提供定制用数据库，包含开源领域数据，具有专业性标注且在本行业具有多样性和覆盖性。

6.3.3 私有数据

私有数据应符合隐私保护法规，确保数据安全性。数据所有者应对数据使用具备控制权，包括访问权限管理和使用审计。数据需具备高质量和完整性，避免缺失值和异常值，确保数据的准确性和可靠性。

6.4 模型

6.4.1 预训练模型

6.4.1.1 一般要求

预训练模型，应符合以下要求：

a)宜支持单模态和多模态等训练方式；

b)宜支持多种模态特征提取的方法（如单塔方法、双塔方法等）；

c)宜支持的数据类型包含文本、语音、图像、视频等；

d)宜能提供相应模态的处理接口（如文本生成、图像理解等）；

e)宜支持的交互模式和协议，包含同步、异步、批量、流式、事件驱动等；

6.4.1.2 单模态

单模态预训练模型，应符合以下要求：

a）应提供单模态数据的特征提取；

b）应支持模态补全、模态掩码、模态增广、模态扩展等任务；

c）应具备至少 1 种单模态理解功能。

d） 宜支持至少 1 种单模态生成功能。

6.4.1.3 多模态

多模态预训练模型，应符合以下要求：

a)应具备至少 1 种多模态理解功能，如图文检索、视觉定位、图音检索、文音检索等；

b)应具备至少 1 种多模态预训练模型基础架构，如单塔、多塔架构等；

c)宜能提供至少 1 种多模态生成功能，如文本生成图片、图片生成文本、图片生成视频、图片生成语音、文本生成视频等；

d)宜支持对大语言模型的桥接。

6.4.2 定制化模型

基于预训练模型，定制具体电力行业细分领域所需模型，应符合以下要求：

a)应支持定制模型的参数量大小、存储容量、计算资源、网络资源、性能评价指标等；

b)应支持多种预训练模型微调方法；

c)应提供模型版本管理功能，包含模型发布、版本回退等；

d)应提供并运维预训练模型库，实现用户上传、微调和使用模型；

e)宜支持面向任务推荐定制化方法（如面向发电、输电、变电、配电、安监等场景）；

f)宜支持基于用户数据和微调数据库数据混合的模型定制。

6.5 细分领域应用

对每种预训练模型（自然语言处理，计算机视觉，多模态等），宜至少匹配 1 个下游任务。

6.6 服务平台/组件

预训练模型服务平台/组件，应符合以下要求：

a)应支持预训练模型插件开发，并提供开发协议以规定插件的规则和接口，如模型接口、输入输出数据格式、插件元数据和插件运行状态码等要求；

b)应支持部署服务升级、回滚；

c)应支持根据业务负载情况，对计算资源进行弹性伸缩；

d)宜支持预训练模型灰度发布、A/B 测试、模型版本管理；

e)预训练模型组件宜能够自动检测和修复问题，减少人工干预；

f)宜支持插件运行监控和日志记录。

_____以下无正文