|  |  |
| --- | --- |
| ICS | 35.240 |
| CCS | |  | | --- | | D:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T.pngD:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T后面的反斜杠.png GXIE |   L 70 |

广西电子学会团体标准

T/XXX XXXX—XXXX

人工智能 预训练大模型训练与应用规范

Artificial intelligence—Specifications for the training and application of pre-trained large models

XXXX - XX - XX发布

XXXX - XX - XX实施

广西电子学会  发布

1. 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。

本文件由广西电子学会提出、归口并宣贯。

本文件起草单位：广西大学计算机与电子信息学院、广西电子学会人工智能专委会、广西医科大学第一附属医院、XXX。

本文件主要起草人：白琳、张振荣、张茜、郑嘉利、XXX、兰伟。

人工智能 预训练大模型训练与应用规范

* 1. 范围

本文件规定人工智能预训练大模型在开发、训练、应用等活动中，对人工智能预训练大模型进行开发、训练、应用应遵循的基本要求。

本文件适用于预训练大模型的全生命周期管理，包括研发、训练、部署及应用环节，覆盖数据处理、算法设计、系统安全、伦理审查等技术与管理要求。

* 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。

GB/T 35273-2020 信息安全技术 个人信息安全规范

GB/T 41867 信息技术 人工智能 术语

GB/T 45288 人工智能　大模型

* 1. 术语和定义

GB/T 41867、GB/T 45288界定的以及下列术语和定义适用于本文件。

预训练大模型 **pre-trained large model**

指在大量数据上进行预训练的神经网络模型，通常包含超大规模参数。

* 1. 数据管理要求
     1. 数据收集与治理
     2. 数据传输与存储
     3. 数据使用管理
        1. 数据使用授权

所有用于预训练的数据须具备清晰、合法、可验证的授权协议，明确数据用途、使用范围、使用期限，并严格按照授权条款使用。

* + - 1. 数据使用记录

建立数据使用登记机制，记录数据使用时间、用途、对象、范围，确保使用过程可审计、可回溯。

* + - 1. 数据生命周期管理

制定数据生命周期管理规范，明确数据采集、使用、存储、销毁各阶段的管理要求，防止数据滥用、遗留风险。

* + 1. 数据伦理与社会责任
  1. 预训练大模型研发与训练规范
     1. 算法可靠性
     2. 伦理与安全约束
  2. 应用部署管理
     1. 伦理审查

部署前需提交伦理影响评估报告（含社会、法律、心理三维度分析）。

高风险场景（如金融决策、司法辅助）需设立人类监督委员会。

* + 1. 可持续发展

模型推理能耗需符合《绿色数据中心能效标准》（PUE≤1.4）。

建立模型退役机制（生命周期≥5年需重新评估技术适用性）。

2. （规范性）  
   预训练大模型伦理审查清单
   1. 数据治理类

（1）‌数据来源合法性‌

指标：数据采集是否取得主体授权或符合《数据安全法》豁免条款。

评分细则：发现未授权敏感数据（如生物特征、医疗记录）直接判定为高风险。

（2）‌数据偏见检测‌

指标：训练数据中性别、种族、地域分布的统计学偏差值（偏差率≤5%）

评分细则：每超出1%扣3分，超过15%直接禁止模型部署

（3）‌数据脱敏完整性‌

指标：身份标识字段（姓名、身份证号等）的匿名化覆盖率（要求100%）

评分细则：每发现1条可追溯原始数据的记录扣5分

* 1. 算法设计类

（1）价值对齐验证‌

指标：模型输出与社会主义核心价值观的契合度（需≥98%）

评分细则：采用VAST测试框架，得分低于90分需重新训练

‌（2）可解释性要求‌

指标：核心算法模块需提供SHAP/LIME可视化分析报告

评分细则：缺失报告扣20分，关键参数解释模糊扣5-10分

‌（3）自我演化抑制‌

指标：禁止算法架构包含自主修改权重的无监督模块

评分细则：发现此类设计直接终止项目

* 1. 应用场景类

‌（1）高风险领域控制‌

指标：医疗诊断类模型需设置人工复核强制介入机制

评分细则：未配置双人复核流程扣15分

‌（2）误导性内容过滤‌

指标：虚假信息识别准确率（阈值≥99.5%）

评分细则：每低于0.1%扣2分，低于98%暂停服务

* 1. 安全防护类

（1）对抗攻击防御

指标：对抗样本检测模块覆盖率（需达100%）

评分细则：每缺失1类攻击防护扣5分

‌（2）隐私泄露防护‌

指标：模型推理过程需实施内存隔离与差分隐私（ε≤2）

评分细则：未采用联邦学习技术扣10分

参考文献

[1] ISO/IEC 23053-2022 人工智能系统框架

