ICS 35.240.20
L60

团　　体　　标　　准

T/ISC 0005—2020

针对内容安全的人工智能数据标注指南

Guidelines for AI data annotation in content security

2020－09－24 发布　　2020－12－01 实施

中　国　互　联　网　协　会　发　布

目　　次

前　言

本文件按照 GB/T 1.1-20200《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。

本文件由中国互联网协会标准工作委员会提出并归口。

本文件起草单位：深圳市腾讯计算机系统有限公司、北京奇艺世纪科技有限公司、北京深度搜索科技有限公司、工业和信息化部电子第五研究所。

本文件主要起草人：杨晓光、鞠奇、马臣、王骏、詹博、邓理英、陈永智、刘翠香、董奕、符妍、黄佳、李久龙、周循道、黄林轶。

引　言

随着《中华人民共和国国家安全法》、《中华人民共和国网络安全法》、《互联网信息服务管理办法》、《网络信息内容生态治理规定》等法律规章制度的发布，网络运营者有责任营造清朗的网络空间、建设良好的网络生态目标，开展弘扬正能量、处置违法和不良信息。使用人工智能技术能够帮助网络运营者及时地发现和处置网络上的违法、不良信息。数据是人工智能技术的“原料”，数据标注则是将“原料”转化为机器可识别的信息的过程。本标准的作用是给数据标注过程提供指南，为机器提供优质的数据“原料”，提高机器识别违法、不良的信息的准确性。本文件也可以为其他应用领域，如智慧城市、自动驾驶、语音识别等的人工智能技术做参考。

针对内容安全的人工智能数据标注指南

1 范围

本文件规定了针对内容安全的人工智能数据标注主要过程，以及过程中的相关活动。

本文件适用于因业务需要使用人工智能技术进行内容安全审核，提供第三方数据标注服务，以及设计开发数据标注服务平台的组织等。

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 35273-2020 信息安全技术 个人信息安全规范

GB/T 37964-2019 信息安全技术 个人信息去标识化指南

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据标注 data annotation

对文本、图像、语音、视频、3D 点云等原始数据进行归类、整理、纠错、转录、翻译和添加标签等操作，以生成满足机器学习训练要求的、机器可识别的数据编码。

3.2

标签 label

标识数据的特征、类别和属性等，可用于建立数据及机器学习训练要求所定义的机器可读数据编码间的联系。标签是数据标注的结果，是机器学习训练所需的输入之一。

3.3

标注过程 annotation process

按照数据标注规范对指定数据集进行标注的过程。

3.4

标注工具 annotation tool

数据标注员完成标注任务产生标注结果时所需的工具和软件。

注1：标注工具可生成标签并提供参考模板。

注2：不同的数据类型和标注任务需要不同的标注工具。标注工具按自动化程度可分为手动、半自动、自动三种。

3.5

标注平台 annotation platform

开展标注任务的系统化框架。

注1：标注平台在包含标注工具全部功能的基础上将所有标注环节工具化，可有效地对标注任务进行全局管理和跟踪。

3.6

数据预处理 data preprocessing

为提升数据标注的效率、质量、降低人力参与强度，对原始数据进行预先处理，其中包括：数据筛选、数据切分、机器半自动预标等过程。

3.7

训练样本 training sample

数据标注后提交给需求方做模型训练样本的数据。

3.8

数据标注员/团队 data labeler/team

对文本、图像、音频、视频、3D点云等原始数据进行归类、整理、纠错、转录、翻译、编辑和添加标签等操作的工作人员或团队。

4 概述

本文件给出了针对内容安全的人工智能数据标注的主要过程，其中包括: 需求承接、标注准备、正式标注、验收交付、训练模型、上线准备等。数据标注流程架构见图 1：

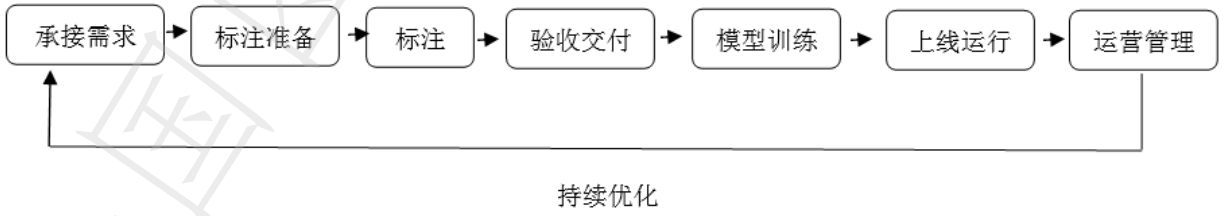

图 1 数据标注流程架构

5 需求承接

5.1 需求接收

标注团队与需求方应明确标注规模、标注形式、标注方法、标签标准、数据安全要求、标注复杂度、标注数据格式、工期约定、准确率要求、数据交付格式、说明文档以及培训细节。

5.2 需求评估

标注团队应对承接的需求进行评估，形成评估结论。

评估结论应包括现有资源能否承接该需求，需求实现路径，以及评估新增需求对现有需求的影响范围，需求承接方案等内容。

5.3 需求确认

标注团队应与需求方协商一致，将最终确认的需求形成文档并留存。

6 标注准备

6.1 数据获取

标注团队应根据需求内容，识别可获取的标注数据源渠道，评估数据源渠道的可行性，确认完成标注需求所需标注数据源构成。

数据获取过程中个人信息保护，应满足GB/T 35273-2020。

数据去标识化处理的方法，应满足GB/T 37964-2019。

6.2 数据预处理

标注团队应根据标注需求以及标注数据的特性，通过数据聚类、组合排列、数据杂质去除等方法，提高标注数据的有效性、标注效率、标注质量。数据预处理方法参见表1：

表1 数据预处理方法

维度	方法	详细内容
通用数据预处理流程	数据去重	MD5 特征值去重，相似度去重
	模型预处理	针对初步具备识别能力的模型，通过模型预测结果进行筛选，进行样本标注
	数据分类	共性无效样本分类识别
	数据聚类	基于相似度的聚类处理
	主动学习	针对初步具备识别能力的模型，通过模型标注，人工修正的方式，进行样本标注
专项数据预处理流程	针对特殊业务形式，数据类型进行专项数据预处理流程研究	多模态技术叠加，多个数据预处理流程叠加

6.3 操作规程

标注团队应：

a) 根据已确认的标注需求，形成标准化的操作规程；

b) 确保执行数据标注任务的相关人员了解操作规程。

6.4 质检方案

标注团队应：

a) 制定质检方案，确保标注结果质量。方案内容包括但不限于：

——质量责任人；

——抽样理论依据，如置信度和误差是否在可接受的范围；
——抽样方式，如随机抽样、分层抽样等；
——抽样量级，如确定整体抽样量级、阶段性抽样量级等；
——抽样频次，如按时间周期抽样、阶段性抽样等；
——反馈机制，如按时间周期反馈、阶段性反馈等；
——指标/阈值的计算方法。

b) 保留质检方案的相关成文信息。

6.5 工具/平台

标注团队应根据需求准备相应的标注工具/平台，如线下工具、平台复用、平台优化、平台新建等方式。

标注工具/平台应具备以下能力，具备包括但不限于如下能力：

a) 对文本、图像、视频、音频、3D 点云数据等各类数据进行标注；
b) 权限管理，包括：创建账号、授权管理、权限审批、角色配置；
c) 人员管理，包括：角色配置、绩效管理；
d) 流程管理，可以根据标注需求进行流程调整；
e) 版本管理，对标注内容和结果进行版本管理和控制。

6.6 人员能力

标注团队应：

a) 确定数据标注员和质检人员所需具备的能力，这些人员从事的工作影响标注的质量和有效性；
b) 基于适当的教育、培训和经验（知识库），确保这些人员是胜任的；
c) 跟踪培训的效果，并评价其有效性；
d) 保留适当的成文信息，作为人员能力的证据。

6.7 试标注

标注团队应：

a) 在正式标注前，小范围抽取数据标注员进行试标注、质检团队试质检，试运行标注的全过程；
b) 对试标注的数据量的大小、百分比等因素进行限定；
c) 保留试标注以及因试标注引起的对标注需求、标注操作规程、质检方案变更相关的成文信息。

6.8 制定标注方案

标注团队应针对特定需求制定相应的标注方案，包括但不限于：

a) 资源规划，如数据源、标注工具/平台；
b) 人力资源规划；
c) 项目进度规划；
d) 项目质量规划；
e) 风险控制措施；
f) 应急预案等。

7 标注

7.1 实施标注

按照已定标注方案，协调安排标注人员进行正式标注活动。其中包括：

a） 数据导入；

b） 任务安排；

c） 人工标注。

7.2 进度管理

标注团队应在标注过程中实时监控、管理标注的实际进度，并根据实际进度分析、预警风险，制定相应方案。

7.3 质量控制

标注团队应按照已定的质检方案进行质量控制。质量控制方法包括但不限于如下方式，见表2：

表2 质量控制方法

质量控制方法	详细描述
多人验证	多人做同一个子任务，通过标注工具的功能自动或人工辅助选择出最优、最正确的标注结果。
埋题验证	在任务进行期间，除了常规标注子任务外，在任务中混进若干已知结果的测试题，以此验证一线操作标注人员的标注水平。
标注人员状态验证	通过一定方法对标注人员的操作规范性、实时注意力状态、标注准确率等方面进行检查与监测，及时发现操作违规问题，保证数据质量。
机器验证	在任务进行期间使用机器学习方法，得到数据准确率，一旦发现离群点或明显的降低趋势，及时对标注人员预警和警告。

7.4 交付、验收

标注团队应按照事先确认的最终需求进行交付。

需求方应按照要求进行验收，如验收数据质量未达到预期，数据需求方可要求标注团队对标注数据进行修正。

双方应保留相关成文信息。

8 模型训练

8.1 模型训练

数据使用方通过运用人工标注结果数据，利用卷积神经网络、循环神经网络等算法模型学习标注后的数据特征，实现对目标样本具有一定的预测能力。

8.2 模型验证

数据使用方应对数据预测效果对模型能力进行分析评估，评价模型效果。

9 上线运行

9.1 模型测试

数据使用方应将训练后的模型接入模拟环境，验证模型效果。

9.2 运营监控

数据使用方应在上线后对模型的应用效果进行持续性监控，以此来保证该模型在线上运营的稳定性。

9.3 持续改进

数据使用方应：

a) 通过对模型上线后的效果分析，推动模型的更新迭代；
b) 与标注团队确认和选择改进机会，并采取必要措施进行改进；
c) 保留改进措施和方案相关成文信息，并作为改进标注方案的输入。

参 考 文 献

[1] GB/T 19001-2016 质量管理体系 要求
[2] T/CESA 1040-2019 信息技术 人工智能 面向机器学习的数据标注规程