ICS 35.240
CCS L 60

中华人民共和国国家标准

GB/T 42755—2023

人工智能　面向机器学习的数据标注规程

Artificial intelligence—Code of practice for data labeling of machine learning

2023-05-23 发布　　2023-12-01 实施

国家市场监督管理总局
国家标准化管理委员会　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。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：北京航空航天大学、中国电子技术标准化研究院、北京百度网讯科技有限公司、浪潮软件科技有限公司、山东省人工智能研究院、美的集团(上海)有限公司、北京智谱华章科技有限公司、北京爱数智慧科技有限公司、腾讯云计算(北京)有限责任公司、北京航天自动控制研究所、郑州中业科技股份有限公司、东软集团股份有限公司、北京海天瑞声科技股份有限公司、云从科技集团股份有限公司、深圳云天励飞技术股份有限公司、中国科学院软件研究所、上海依图网络科技有限公司、中国医学科学院生物医学工程研究所、平安科技(深圳)有限公司、上海商汤智能科技有限公司、上海人工智能实验室、上海计算机软件技术开发中心、中国航空综合技术研究所、中国科学院新疆理化技术研究所、中国质量认证中心、中汽数据(天津)有限公司、北京眼神科技有限公司、上海人工智能研究院有限公司、浙江大华技术股份有限公司、杭州趣链科技有限公司、常州微亿智造科技有限公司、长春博立电子科技有限公司、罗克佳华科技集团股份有限公司、上海交通大学、上海计算机软件技术开发中心。

本文件主要起草人：吴文峻、董建、马珊珊、刘祥龙、徐洋、贾一君、孟令中、任健、陈斌、赵豪杰、刘海涛、陈尚义、脱立恒、左家平、王丽娜、徐颂、王健宗、张楠、蔡亚森、王功明、陈敏刚、赵赫、金铸、郝玉峰、刘永辉、李玮、赵春昊、黄志龙、杨春林、王潇蔓、施佳樑、舒明雷、王英龙、匡立中、陈晓丰、吴庚、蒋慧、蒲江波、马元巍、邢警、乔宇、何聪辉、杨雅婷、马博、陶剑、胡进伟、楚思思、李军、宋海涛、沈灏、程淼、郑忠斌、李爽。

人工智能　面向机器学习的数据标注规程

1　范围

本文件规定了人工智能领域面向机器学习的数据标注框架流程。

本文件适用于指导人工智能领域面向机器学习的数据标注以及与之相关的研究、开发和应用等。

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 35274—2017　信息安全技术　大数据服务安全能力要求

GB/T 37973—2019　信息安全技术　大数据安全管理指南

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 labeling

给数据样本指定目标变量和赋值的过程。

3.2

标注任务　labeling task

按照数据标注说明对数据进行标注的活动。

3.3

数据标注方　data labeler

承担数据标注任务的人员或机构。

3.4

数据需求方　data user

提出数据标注需求的人员或机构。

3.5

标注管理方　data labeling administrator

管理数据标注任务评估、分发、交付、验收以及质量把控的人员或机构。

3.6

标注工具　labeling tool

数据标注方执行数据标注时使用的工具，标注管理方管理数据标注时使用的工具，数据需求方验收数据标注时使用的工具等所有流程相关的工具。

3.7

标注任务说明　labeling task description

数据需求方用于向标注管理方以及数据标注方明确标注任务的书面表达。

注：标注任务说明通常包含对要执行的标注任务的描述、标注方法、正反示例、验收方法与验收指标等内容。

4 数据标注流程

数据标注涉及数据需求方、标注管理方及数据标注方三方人员，主要流程包括标注任务前期准备、标注任务执行、标注结果输出三个阶段。数据标注流程见图 1。

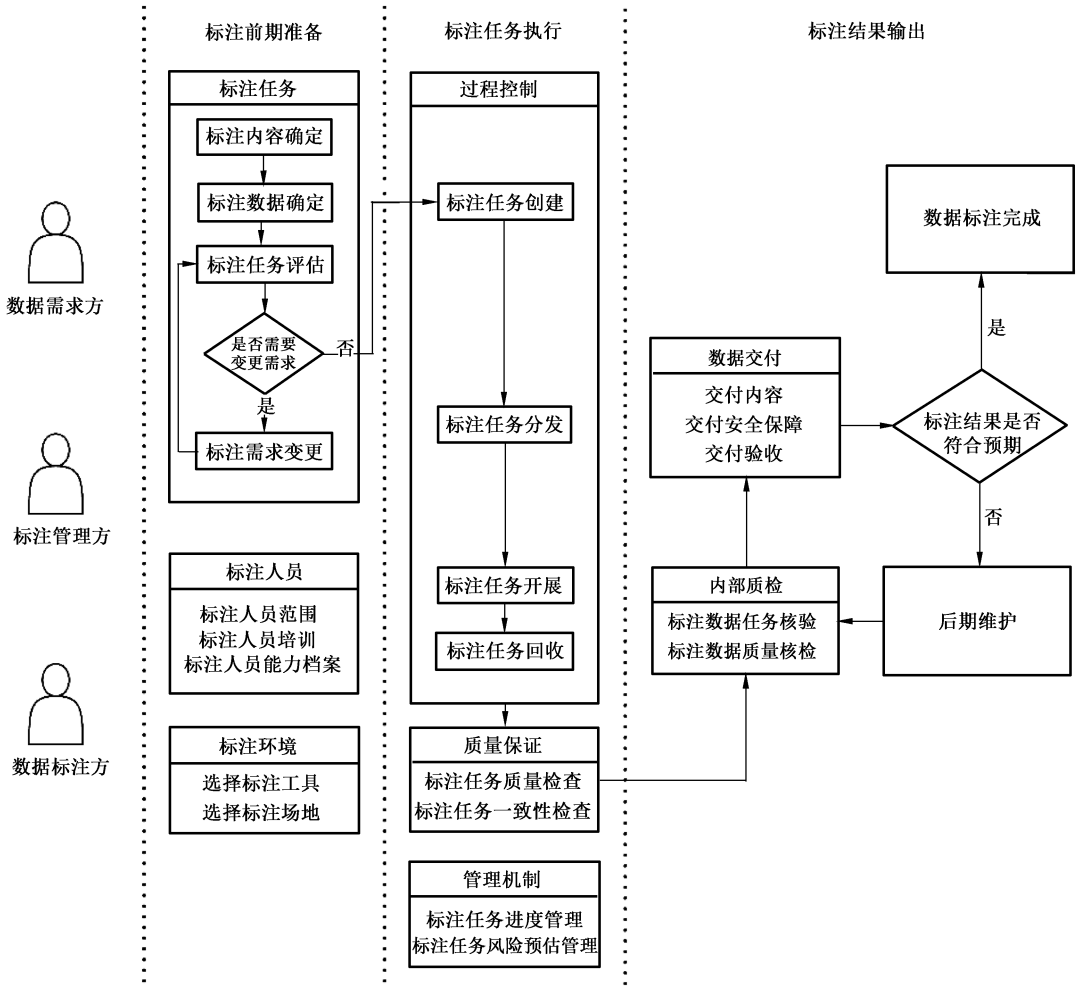

图 1 数据标注流程框架

在标注前期准备阶段，数据需求方和标注管理方应确定标注任务，完成标注内容和标注数据的确定。标注管理方评估标注任务，向数据需求方反馈是否需要变更需求，若需要则变更标注需求，并重新评估标注任务。标注前期准备阶段还应根据标注人员的要求确定数据标注方，同时确定标注环境，选择合适的标注工具和场景。在标注任务执行阶段，数据需求方、标注管理方及数据标注方三方人员应遵循标注流程的过程控制，完成标注任务的创建、分发、开展及回收。同时应保证标注任务的质量，严格遵守管理机制。在标注结果输出阶段，数据标注方应对数据标注方标注后的数据进行内部质检，质检合格后将标注后的数据交付给数据需求方。若标注后的数据符合预期，则数据标注完成；否则进行后期维护环节，数据标注方应对数据进行修正，并重启内部质检流程。

5 标注任务前期准备

5.1 标注任务

5.1.1 标注任务确定

标注内容由标注需求方在标注任务说明中提供，标注任务说明一经确认，不可修改，如需修改则进入需求变更环节，标注任务应包括但不限于：

a) 版本信息：明确当前版本编号、发布日期、发布人、发布说明（发布原因或迭代原因）；
b) 历史迭代信息（历代版本编号、发布日期、发布人、发布说明等）；
c) 项目背景：明确数据标注需求产生的原因，以及数据标注结果的应用场景；
d) 任务描述：明确数据标注任务，包括数据形式、数据规模、标注规则、相关术语、标注样例、质量要求、指标计算方式、验收流程、交付时间等；
e) 主客观描述：明确说明数据标签是根据个人专业领域知识进行标注，还是客观认识进行的标注；
f) 标注人员资质：约定标注任务参与人员的资质要求；
g) 标注结果：明确数据标注结果的交付形式；
h) 知识产权：明确数据的知识产权归属。

5.1.2 标注数据确定

5.1.2.1 待标注数据分析

数据标注前，数据需求方应对待标注数据进行分析，核对标注任务，包括：

a) 数据核查：检查待标注数据是否与标注任务说明书中的数据定义相符，核查结果及时同步给数据需求方；
b) 数据整理：建立完善的数据追踪机制，实现数据整理，以及最小粒度的数据追踪；
c) 数据处理：根据标注任务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，提高标注质量。

5.1.2.2 数据安全等级确定

根据标注任务中的数据安全描述，数据需求方应根据 GB/T 37973—2019 及 GB/T 35274—2017 相关要求，确定标注数据的安全等级。

5.1.3 标注任务评估

数据标注前，标注管理方应对标注任务进行评估，包括：

a) 根据标注任务说明，评估标注任务可行性、标注规则合理性；
b) 在数据需求方提供的小规模样本上进行预标注，将标注结果提交给数据需求方验收。在获得数据需求方确认后，再正式启动数据标注任务。

注：及时记录数据预标注流程中标注规则与数据相悖、覆盖不全或规则之间相悖的情况，并向数据需求方反馈完善标注规则。

5.1.4 标注需求变更

标注需求方需求变更时，应在标注管理方评审同意后更新标注任务说明，重新进入标注任务评估阶段。

5.2 标注人员

5.2.1 标注人员范围

数据标注方应根据标注任务内容中规定的标注人员资质需求，确定符合要求的人员进入标注人员培训环节。

5.2.2 标注人员培训

数据标注方应根据标注任务说明，对标注人员进行岗前能力培训。标注能力考试合格者，方能参与标注任务。

5.2.3 标注人员能力档案

数据标注方应建立标注人员能力档案，记录标注人员承担标注任务的相关内容，用于进行标注人员能力评估与标注质量追踪。

5.3 标注环境

5.3.1 选择标注工具

数据标注方应根据标注任务难度、数据处理规模及数据属性特征、数据安全控制层级与方式，合理选择标注工具，完成数据标注任务。

5.3.2 选择标注场地

数据标注方应根据标注任务中必要的数据安全要求，搭建数据标注场地。

6 标注任务执行

6.1 过程控制

6.1.1 标注任务创建

标注任务创建主要包括以下内容：

a) 任务包创建：数据需求方应以适应标注环境分发、标注工具读取为目标，将需要标注的数据合理分组，保证数据标注质量以及后续的任务分配；

b) 任务账户创建：数据需求方应以标注人员数量为依据，综合标注、质量分析等任务需求，根据标注环境或者标注工具，创建标注过程中所需要使用的用户账户，并分配相应的权限或账户使用规则；

c) 任务创建保障：标注管理方应制定标注任务创建和数据上传相关制度，明确数据上传分类分级、数据安全风险评估和安全监控制度，监管上传数据的合法合规性。

6.1.2 标注任务分发

标注任务分发主要包括以下内容：

a) 标注任务分发类型：标注管理方应根据标注任务要求的标注环境、标注工具，结合标注质量管理以及标注速度管理，在保证标注质量的前提下，选择主动领取或系统自动分发等任务分发类型，优化标注任务分发策略；

b) 标注任务分发保障：标注管理方应制定标注任务分发安全管理制度，明确标注任务分发日志内容，监控标注任务分发安全。

6.1.3 标注任务开展

标注任务开展主要包括以下内容：

a) 标注任务分配：数据标注方应安排数据标注人员使用分配的标注账户，对分配到的任务进行标注；

b) 标注过程反馈：数据标注方应建立标注过程反馈机制，将与标注要求不符、标注要求中未涵盖的数据等情况，及时反馈给标注管理方和数据需求方，确保标注规则与数据的匹配度；

c) 标注任务开展保障：数据标注方应明确数据脱敏和个人信息安全影响评估制度，在标注前对个人信息进行数据脱敏处理，保障数据标注流程合法合规性，并对数据标注结果进行分级制度，适当提高数据安全等级。

6.1.4 标注任务回收

在标注任务完成后，数据标注方的标注团队负责人应检查标注数量，确保所有任务包均被回收，对未能及时完成的任务包，要建立适当的沟通和回收再发放的管理机制，以确保标注任务按期完成，保证任务进度。

6.2 质量保证

6.2.1 标注任务质量检查

在标注过程中，应采用多种检查方法对标注任务质量进行检测，对不满足标注任务要求的，及时预警反馈，并查明问题原因。根据项目特性，标注任务质量检查方法可归纳为以下三种。

a) 机器验证：在任务进行期间，安排超过一名人员做同一个子任务，选择出最优、最正确的标注结果。结果选择可通过下列方式。

   1) 标注工具自动选择：通过与标注工具匹配的模型推理，或拟合若干个标注结果，选择其中置信度最高的标注结果，作为最终结果；

   2) 人工辅助选择：人工对多个标注结果进行对比，从而挑选出置信度最高的标注结果作为最终结果。对于需要特定专业知识标注的领域，进行人工辅助标注时应以多个专家的共同商议结果作为最终结果。

b) 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，以此验证标注质量。在此操作的过程中注意以下事项。

   1) 针对数据特征专题专用：对于埋题验证，应保障测试题在真题中间处于混淆的状态。因此，在出题的过程中，应针对数据的自身特征（数据的类别、场景、内容等），准备相应的题目，避免题目暴露于操作者，失去验证的效果。

   2) 限制题目的使用次数：为避免题目多次出现，引起被测者的注意，从而失去验证效果，应限制题目的使用次数。尤其是拥有容易记忆的特征点的题目（如特定脸部特征、特定文字、特定场景等），应严格限制出现的次数。

c) 标注人员状态验证：通过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；在发现操作违规问题、数据质量有下降时，应根据时间段等特征，对标注人员在这一状态内操作的标注数据进行检查或者返工等操作。

6.2.2 标注任务一致性检查

在标注任务进行期间应使用统计规则或模型验证等方法，得到标注任务一致性水平，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。

6.3 管理机制

6.3.1 标注任务进度管理

数据标注方应定期与标注管理方同步数据标注任务工作进度。

6.3.2 标注任务风险预估

在标注任务进行过程中，数据标注方应对标注人员是否能够如期达到对应的执行进度进行预估和检测，并针对可能存在的标注进度延误风险，对数据需求方、标注管理方进行适当的提示。任务风险预估和提醒的方法可分为以下两类。

a) 收集和更新：在任务进行期间，对不同的时间节点，对标注人员任务完成时间的推测和预估进行定期收集、更新，汇总于系统上，并对其中有风险的完成时间节点进行显著的提醒。这种预估和提醒的方式，应遵守以下规则。

   1) 收集、更新任务完成的推测时间的时间节点：为了在标注的工作全流程中得到尽量准确的推测时间预估，同时也避免频繁收集推测时间造成的效率损失，应在任务未开始标注、开始分配并启动标注时和距离标注结束较为接近时的节点，对推测的任务完成时间进行收集、更新。

   2) 判定任务完成风险：在上述若干类时间节点收集到推测完成标注任务的时间节点之后，应对收集到的标注人员上报的时间节点进行判定，从而推断出标注任务当前的执行是否存在逾期的风险。对于是否有逾期风险的判定，基本的判定规则是依据标注管理方扣除预估充裕的验收、返工时间后给出的截止时间，将任务完成的推测时间与之进行对比。如果推测时间晚于该截止时间，则任务存在风险，应进行风险的提醒。

b) 效率推测：在任务进行期间，根据任务当前的完成进度，以及投入的标注人员的人力、效率，进行标注效率的推测。效率的推测过程应遵守以下规范。

   1) 应以天为单位，进行时间尺度上的效率推算。标注人员的工作时间并非全天候，在全日时间内的分布并不均匀，主要集中在规定的上班时间，因此，对于标注效率的推算，宜以天为单位，能够在若干天的项目周期内，得到更准确的产能预估。

   2) 应以标注人员整组的按天效率为粒度，进行工作能力上的推算。以完成任务的整一组标注人员为整体，进行整体任务效率的预估，有效屏蔽标注人员个体在效率上的差距，得到标注任务在推进上的总体效率，更有效地反映任务的推进速度。

   3) 宜在每日结束的时间点，以天为单位对执行风险进行平均推算和提示。在每日结束的时间点，对当天的标注人员整组的按天效率进行计算，可通过历史若干天的平均效率，以及当前剩余的数据量，推算出标注任务剩余的预估工作日数。若该日数已经超过目前预计的截止时间，则认为任务有可能存在逾期的风险，此时应进行风险的及时暴露和提醒。

6.3.3 标注任务风险提醒

在预估到标注任务可能存在风险的情况下，数据标注方应对风险进行及时的暴露和提醒，从而使得数据需求方、标注管理方能及时对该逾期风险进行处理。因此风险的暴露和提醒应足够清晰，应保障消息能够触达数据需求方、标注管理方。风险预估的消息提醒应包含如下信息，以助于数据需求方、标注管理方进行恰当的情势判断：

a) 标注任务的基本概况信息：包含足以识别面临风险的标注任务的信息，包括任务的名称、需求方、标注要求、总任务量、剩余未完成的任务量等；

b) 执行任务的标注人员：包含足以识别面临风险的标注人员或团队的信息；

c) 目前预估的完成时间；
d) 完成时间的预估途径：通过何种途径预估得到完成时间，包括且不限于上述两种途径（收集和刷新/效率推测）；
e) 原本预计的截止时间。

为了使得消息能够及时传递到数据需求方、标注管理方，标注任务可能有风险的消息内容应通过各类手段进行触达，包括但不限于：

a) 电子邮件；
b) 告警短信；
c) 告警电话；
d) 应用消息推送。

7 标注结果输出

7.1 内部质检

7.1.1 内部质检要求

数据标注方应在完成数据标注，由内部质检验收合格后，提交给标注管理方。根据数据标注任务说明，合格的数据标注结果应满足：

a) 标注数据核验满足数据标注任务说明中的要求；
b) 数据质量满足数据标注任务说明中的要求。

7.1.2 标注数据任务核验

根据数据标注任务说明，数据标注方应对标注数据格式、内容进行合理性和正确性核验，以确定其满足标注要求。

7.1.3 标注数据质量检查

标注数据质量检查能够确保数据标注结果有价值，符合数据需求方的特定应用目的。根据项目特性，质量检查方法可以归纳为以下几种，标注项目负责人应根据场景需求及项目特点进行选择。

a) 逐条检查：即对整个标注项目所包含的所有标注子任务逐一核查并确认。适用于项目量级不大、人力资源充沛、时间节点不紧张、对标注数据结果的准确率要求极高的标注项目。这种方法覆盖的质量检查范围最全，同时也适用于任何形式的数据标注场景。该方法可确保标注数据输出的最高质量，尤其对于数据格式主观成分较多、应用场景较复杂的任务更有效。
b) 按比例抽查：即从全部标注数据中科学地抽取样本，对样本中的数据逐条检查，以此评判全部标注数据的质量。样本量的选择应符合统计学基本原理，足以代表全部标注数据，例如在逐包分配进行标注的同时，可以确保每包均按一定比例进行抽查，以确保抽样足够均匀，足以代表总体结果。抽查审核时，项目负责人应指定审核员完成，审核员应明确标注的详细执行要求，从而确保交付质量。
c) 抽样检验：即从整个标注项目中随机抽取少量标注子任务进行检验，据此判断该标注项目是否合格。抽样检验可分为简单抽样、系统抽样和分层抽样三种方式。
d) 机器验证：通过机器学习，包括使用已训练模型进行检查或使用迁移学习、在线学习等方法对人工标注的数据做质量检查，实现全自动或辅助人工质量检查方式。机器学习方法输出的准确率不能完全代表数据集的准确率，但能在一定程度上反映数据集的质量。
e) 第三方验证：医学等专业领域，如需对标注结果进行第三方验证的，应由有资质的第三方邀请有资质和从业经验的专家进行验证，从而确保标注结果的质量。

7.1.4 标注数据质量检查设定

在质量检查过程中，为了防止一次性不合格数据积压过多而导致延误交付，同时防止检查过于碎片化、零散导致检查效率低下、检查切换时间开销过大，对于不同任务检查的时间点，应进行如下设定，避免此类情况发生。

a) 设定质量检查间隔：通过设定质量检查间隔，使得抽样更均匀，更能有效反映出整体的质量情况；同时使得需要被返工的数据可以被以一定的时间间隔向前面的环节返工，避免大量的返工数据堆积的情况发生。

b) 设定开始检查的完成比率：在标注任务的完成比率还没有到达一定的数值之前，此时由于被完成的数据量太少，介入检查容易造成检查过于碎片化，对于检查或者返工的流转都会造成时间的损耗，应设置任务在完成一定比率时，介入进行质量的检查。该完成比率可以根据任务的总数据量进行灵活的设置。

c) 设定检查任务队列：按一定的规则对待检查任务进行排序，在有多个任务需要被同时检查时，对于任务进度更接近完成的，以及任务未检查数据占总完成数据量比重更高的任务，这些任务是离交付更接近、检查任务更重的，应被优先检查，此类任务应被排序于检查任务队列的前端。

7.2 数据交付

7.2.1 交付内容

数据交付时，数据标注方应对最终提交的数据内容进行说明。交付的内容包括：

a) 标注结果；

b) 说明文档；

c) 标注规范；

d) 原始数据。

7.2.2 交付安全保障

标注管理方和数据标注方应按照事先协商约定好的安全的递交方式递交标注结果，约束的内容包括但不限于递交数据的介质、递交数据的途径、工作数据的保存与删除原则、数据安全责任的物理或时间起始点原则等。

7.2.3 交付验收

数据标注方按流程完成标注任务后，应将成果物交付标注管理方。

a) 标注管理方应根据双方约定的确认验收标准，对数据标注质量进行检查与评价。

b) 标注管理方应及时向数据标注方反馈数据标注质量的相关结果，确定是否通过数据标注质量验收。

注：标注管理方直接或聘请第三方专业团队对数据标注结果进行验收。

7.3 后期维护

根据数据标注任务说明中后期维护的要求，在交付验收后，数据标注方应提供相关的服务。若数据质量未达到预期值，标注管理方应要求数据标注方对数据进行修正。