

黑 龙 江 省 地 方 标 准

DB 23/T XXXX—2024

人工智能数据标注总体框架规范

（征求意见稿）

起草单位：黑龙江省网络空间研究中心

联系人：白瑞

联系电话：18845165426

联系邮箱：guines@126.com

2024 - XX - XX 发布

2024 - XX - XX 实施

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这些专利的责任。

本文件中共黑龙江省委网络安全和信息化委员会办公室提出并归口。

本文件起草单位：黑龙江省网络空间研究中心

本文件主要起草人：

人工智能数据标注总体框架规范

1 范围

本文件规定了人工智能数据标注的要求，为人工智能数据标注的生产过程提供建议。

本文件适用于面向人工智能学习、研发或应用等需要实施数据标注的企业、高校、研究机构、政府机关。其他行业和机构可参照执行。

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 35295 信息技术 大数据 术语

GB/T 25069 信息安全技术 术语

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

对准备使用人工智能研究的文本、图像、音频和视频等数据进行特征标注以满足正常可用的过程。

3.2

数据标签

用于标识出数据特征并建立标注数据与人工智能可用数据之间联系的信息。

3.3

数据标注工具

能够进行数据标注及在数据标注过程中进行管理、控制等过程中发挥作用的软件的统称。

3.4

数据标注技术

对数据进行转写、切割、提取、选择、富集标注、点标注、线标注、框标注、区域标注等操作的技术。

3.5

数据标注合格率

某项标注任务中，符合标注规则要求的已标注数据总量与所有已标注数据总量的比值。

4 基本要求

4.1 标注人员

参与数据标注的人员应具有相应的资质，对于确定符合要求的人员培训环节，应做到：

a) 根据标注任务说明，对标注人员进行岗前能力培训。标注能力合格者，参与标注任务；

- b) 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。

4.2 明确职责

规定参与人工智能数据标注的所有角色的职能,应做到:

- a) 设立人工智能数据管理岗位。该岗位要求对业务、法律法规比较熟悉,应根据业务实际需要确定承担数据管理工作的部门或人员的职责;
- b) 明确各环节角色的职责。应明确数据标注人员、数据标注培训人员、数据标注质量控制员以及参与人工智能数据标注相关的其他角色的职责。

4.3 合法合规

应出台规章制度保障人工智能数据标注活动的合法合规性,并做到:

- a) 学习并严格执行与数据保护、数据安全相关的法律法规、制度等;
- b) 正确的对个人隐私和敏感数据进行处理,确保标注过程合法合规;
- c) 建立跨部门、跨组织数据标注、传输的保护制度。

4.4 保障质量

应确保人工智能数据标注的质量,并做到:

- a) 确保人工智能数据标注的准确性、可用性、完整性;
- b) 建立质量保障制度,提高数据标注合格率;
- c) 建立定期抽查,不定期检测的质量控制制度。

4.5 标注范围最小化

应确保数据标注相关角色对数据掌握范围的最小化,做到:

- a) 数据标注任务开始前,明确数据接触范围及使用范围;
- b) 提供技术或者建立制度保证数据标注过程中,数据范围不扩散;
- c) 数据标注任务完成后,及时回收数据操作权限。

4.6 数据安全

应从以下几方面做好数据保密工作,确保数据安全:

- a) 数据分发,使用必要的安全方式确保人工智能数据分发过程中的安全性要求;
- b) 数据存储,建立访问控制制度以及加密机制确保人工智能数据存在保密性要求;
- c) 加密数据的标注,使用加密算法对加密人工智能数据进行运算标注,如同态加密算法等;
- d) 数据汇总,使用数据隔离等方式确保人工智能数据汇总时满足保密性要求;
- e) 密码密钥的安全,建立人工智能数据密码密钥管理系统。

4.7 数据完整

为确保标注过程数据完整性,应做到:

- a) 接收人工智能数据可验证,应保证接收的数据已通过认证;
- b) 人工智能数据传输过程完整性,应保证标注活动数据完整性;
- c) 数据标注可靠性,确保标准过程只执行了批准的范围;
- d) 数据储存完整性,确保数据标注存储及备份数据完整行。

4.8 数据可审计

- 对数据标注各环节建立审计机制或制度，确保数据标注全过程能够对数据进行审计，应做到：
- a) 对数据标注过程信息进行记录，并保证记录过程真实可靠；
 - b) 利用合理的技术方案确保数据标注的所有过程和行为可溯源。

4.9 其他要求

- 数据标注除了考虑以上要求外，还应针对人工智能数据的特点从数据标注任务的其他方面分析，包括但不限于：
- a) 数据标注需不影响社会和公共安全；
 - b) 数据标注需考虑跨部门、跨组织的数据开放共享；
 - c) 跨境数据流动；
 - d) 知识产权保护及数据价值保护。

5 总体框架

人工智能数据标注框架应包括任务规划、任务执行、任务评审、交付验收和总结、后期维护六过程。具体过程可根据数据标注任务规模简化或裁剪，人工智能数据标注总体框架见图1。

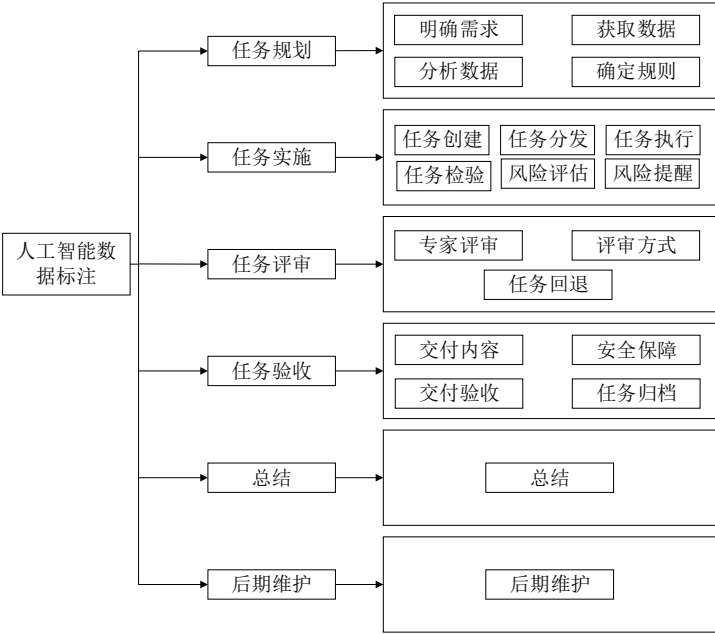


图1 人工智能数据标注总体框架

6 任务规划

6.1 明确需求

- 数据标注开始前，数据标注任务负责人应及时获取并明确数据标注的需求，包括但不限于以下内容：
- a) 数据标注需求完成的时间节点；
 - b) 数据标注的内容；

- c) 数据标注的方法；
- d) 数据验收规则；
- e) 约定标注任务参与人员的资质要求；
- f) 数据标注结果的交付形式；
- g) 数据的知识产权归属。

6.2 获取数据

数据标注项目负责人应及时获取待标注数据，数据获取方式包括：

- a) 平台接口对接的方式；
- b) 存储介质复制的方式。

6.3 分析数据

数据标注前，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。

6.4 确定规则

数据标注项目负责人应按照标注规划进行标注规则验证与细化，并完成以下内容：

- a) 确定数据试标数量。应遵循数据标注内容覆盖全面性及数据标注成本最小化要求，抽取数量作为试标数量；
- b) 数据试标。应选择中等以上技能的数据标注员，按照标注规则对抽样数据进行标注，以验证标注规则的合理性；
- c) 细化数据标注规则。应遵循易理解、易操作的原则，保持与数据标注需求一致的原则，对数据试标中发现的不合理的规则进行修改。

7 任务实施

7.1 任务创建

数据标注任务负责人应利用标注工具创建数据标注项目相关内容，包括：

- a) 任务包创建。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。应以标注人员数量为依据，综合标注数据规模、质量分析等需求，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。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。

7.2 任务分发

数据标注任务负责人应利用标注工具将数据标注任务分派给数据标注人员或数据标注团队，标注任务分发主要包括以下内容：

- a) 标注任务分发类型。应根据标注任务要求，结合标注质量管理以及标注时间管理，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。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。

7.3 任务执行

数据标注员或数据标注团队应使用数据标注工具完成指派的数据标注任务，标注执行主要包括以下内容：

- a) 标注任务分配。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。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，及时反馈标注管理，确保标注规则与数据的匹配度；
- c) 标注任务开展保障。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。

7.4 任务检验

7.4.1 质量检验

数据标注质量控制员在任务进行过程中按照标注规则进行质量检验和质量纠偏。

7.4.2 进度检验

数据标注任务负责人在任务进行中应使用数据标注工具，对数据标注进度和数据标注计划偏差情况进行检验，并及时调整资源和标准进度要求。

7.4.3 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，一旦发现明显的偏离或出现一致性降低的趋势，及时对标注人员预警。

7.5 风险预估

在标注任务进行过程中，应对标注任务是否能够达到预期进行检测和评估，并针对可能存在的风险进行提示。风险评估应包括以下几方面：

- a) 任务时间风险。包括标注任务可能在预期时间内未能完成或者标注任务所需的时间可能被低估，导致任务无法按时完成的风险；
- b) 任务进度风险。参与标注任务人员不足风险及标注人员之间沟通、协调与管理的风险，导致进度延误；
- c) 任务质量风险。数据标注不准确，或者不同标注人员主观意识和个人偏好出现标注偏差，导致标注数据的质量下降的风险；
- d) 数据隐私风险：在数据标注过程中，可能涉及到个人敏感信息的泄露，如姓名、地址等泄密的风险。

7.6 风险提醒

7.6.1 在预估到标注任务可能存在风险的情况下，应对风险进行及时的暴露和提醒，从而使得风险进行处理。风险预估的消息提醒应包含如下信息：

- 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、风险描述、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。

7.6.2 为了使得消息能够及时传递到相关任务关系人,标注任务可能有风险的消息内容应通过各类方式发送通知,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。

8 任务评审

8.1 成立评审管理委员会

数据标注完成后,数据标注任务负责人应及时成立评审管理委员会,评审管理委员会应具备以下条件:

- a) 业务范围能够满足数据标注生产工作评审需求;
- b) 专家库的专家数量具备一定规模并且专业结构分别合理;
- c) 专业技术能力满足需求并且已建立质量保障制度;
- d) 评审业绩良好,并且3年内无违法违纪行为;
- e) 组织规定的其他条件。

8.2 专家评审

评审管理委员会组织专家开展评审,评审专家应具备以下条件:

- a) 具有良好的科学道德、较强的沟通和协调能力,能够独立、客观、公正、实事求是地提出评审意见;
- b) 从事所属领域或行业专业技术工作满5年,并具有副高级及以上专业技术职称或者具有同等专业技术水平;
- c) 熟悉数据标注评审基本业务流程,掌握数据标注生产工作的基本原理、方法;
- d) 部门或组织规定的其他条件。

8.3 评审方式

评审应采取以下几种方式之一或其组合:

- a) 网络评审。组织评审专家通过网络评审系统,查阅数据标注材料并对数据标注任务完成情况进行评审;
- b) 会议评审(含视频答辩)。组织评审专家采用现场会议(网络视频)等形式,通过听取数据标注情况的汇报、质询、讨论等程序,对数据标注工作进行评审。

8.4 任务回退

评审不合格的任务,数据标注负责人应进行回收,并重新分派进行标注修改后重新评审。

9 任务验收

9.1 交付材料

评审通过后,进行数据交付,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。任务总结报告、数据标注结果等;
- b) 说明文档。数据标注情况说明,包括标签或类别等;
- c) 标注规范。数据标注方案、标注规则等;
- d) 原始数据。获取的原始待标注数据。

9.2 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递交方式递交标注结果,约束的内容包括但不限于:

- a) 递交数据的介质;
- b) 递交数据的途径;
- c) 工作数据的保存与删除原则;
- d) 数据安全责任的物理或时间起始点原则。

9.3 交付验收

按流程完成标注任务后,应将成果物交付完成任务验收,交付验证主要包括:

- a) 应根据确认的验收标准,对数据标注质量进行检查与评价;
- b) 应及时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。

9.4 任务归档

确认数据标注结果验收通过后,对生产过程文档进行分类、整理、归档。

10 总结

数据标注负责人应组织各工作角色参与项目的总结工作,编制总结报告,内容包括:

- a) 数据标注项目基本信息。描述本次数据标注项目的合作背景、数据标注内容及类型;
- b) 数据标注过程控制措施。描述本次数据标注项目在执行过程中采取的执行方法,还原标注过程;
- c) 数据标注方法总结。描述本次数据标注项目遇到的困难及采取的有效措施,总结可行的方法做经验复用;
- d) 数据标注措施改进。描述本次数据标注项目控制过程的不足,总结可以改进的方法和措施,应对下次同类项目;
- e) 其他需要总结的内容。

11 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。