

团体标准

T/CESA XXXX—2024

人工智能 数据集质量评估要求

Artificial intelligence-Data set quality evaluation requirements

征求意见稿

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

已授权的专利证明材料为专利证书复印件或扉页，已公开但尚未授权的专利申请证明材料为专利公开通知书复印件或扉页，未公开的专利申请的证明材料为专利申请号和申请日期。

202X-XX- XX 发布

202X-XX- XX 实施

中国电子工业标准化技术协会 发布



版权保护文件

版权所有归属于该标准的发布机构，除非有其他规定，否则未经许可，此发行物及其章节不得以其他形式或任何手段进行复制、再版或使用，包括电子版，影印件，或发布在互联网及内部网络等。使用许可可于发布机构获取。

目 次

前 言..... III

人工智能 数据集质量评估要求..... 1

1 范围..... 1

2 规范性引用文件..... 1

3 术语和定义..... 1

4 数据集质量要求..... 2

 4.1 通用要求..... 2

 4.2 训练数据集扩展要求..... 5

 4.3 测试数据集扩展要求..... 7

5 数据集质量评估流程..... 9

 5.1 概述..... 9

 5.2 评估准备..... 9

 5.3 评估执行..... 10

 5.4 结果确认..... 11

参考文献..... 12

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。

本文件由中国电子技术标准化研究院提出。

本文件由中国电子技术标准化研究院、中国电子工业标准化技术协会归口。

本文件起草单位： 。

本文件主要起草人： 。

人工智能 数据集质量评估要求

1 范围

本文件规定了人工智能领域数据集的质量评估要求，包括通用要求、训练数据集扩展要求、测试数据集扩展要求，给出了人工智能领域数据集的通用评估流程。

本文件适用于指导人工智能领域的开发方、用户方以及第三方等相关组织对人工智能领域数据集质量开展评估工作。

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 25000.24-2017 系统与软件工程 系统与软件质量要求与评价 第24部分：数据质量测量

GB/T 36344-2018 信息技术 数据质量评价指标

GB/T 41867-2022 信息技术 人工智能 术语

ISO/IEC DIS 5259-2 人工智能 用于分析和机器学习的数据质量 第2部分：数据质量测量 (Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2:Data quality measures)

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据质量 data quality

在指定条件下使用时，数据的特性满足明确的和隐含的要求的程度。

[来源：GB/T 25000.24—2017，4.11]

3.2

数据集 dataset

具有一定主题，可以标识并可以被计算机化处理的数据集合。

[来源：GB/T 36344-2018，2.6]

3.3

数据记录 data record

一个单元化的相关数据项的集合。

[来源：GB/T 25000.24—2017，4.15]

3.4

数据质量度量元 data quality measure

作为数据质量特征测量的结果赋值的变量。

3.5

数据质量模型 data quality model

已定义的特性集合，提供一个框架用于说明数据质量需求和评价数据质量。

[来源：GB/T 25000.12—2017，4.6]

3.6

元数据 metadata

关于数据或数据元素的数据（可能包括其数据描述），以及关于数据拥有权、存取路径、访问权和数据易变性的数据。

[来源：GB/T 36344—2018，2.2]

3.7

完整性 integrity

给定环境中，与一个实体相关联的主题数据具有所有预期属性的值和与其相关的值的程度。

3.8

准确性 accuracy

在给定环境中，数据具有正确地表示一个概念或事件的相关属性真实值的属性的程度。

4 数据集质量要求

4.1 通用要求

4.1.1 概述

通用要求是指评估数据集质量的必选指标，包含了训练数据集和测试数据集的共有指标。从通用角度看，数据集常见的质量问题包括数据不一致、数据不完整、数据不均衡、数据重复、数据缺失、格式不一致等，对于数据集质量度量的评估维度包括完整性、准确性、现时性、无偏性、相关性、安全性、数据集共享等。

4.1.2 完整性

完整性^[2]是指数据信息的缺失程度，包括变量的缺失和变量值的缺失，对于不同研究，数据的缺失程度、缺失分布、缺失原因和变量值的缺失机制不尽相同，应该予以详尽描述。当特定研究的数据缺失比例明显超过同类研究的比例时，会加大研究结论的不确定性，此时需要慎重考虑该数据能否作为支持产生真实世界证据的数据。对缺失原因的详细分析有助于对数据可靠性的综合判断。如果涉及缺失数据的填补问题，应根据缺失机制的合理假设采用恰当的填补方法。

数据集完整性的评估要求见表 1。

表 1 数据集完整性评估要求

评价指标	指标定义	计算公式
数据元素完整性	按照业务规则要求，数据集中应被赋值的数据元素的赋值程度。数据元素即数据本身，如图像、文本和视频等。	$X=A/B$ A=被赋值的数据集中元素的个数 B=预期被赋值的数据集中元素的个数
数据记录完整性	按照业务规则要求，数据集中应被赋值的数据记录的赋值程度。数据记录包括数据本身及其相关信息，如图像数据和其对应标注数据为一条记录。	$X=A/B$ A=被赋值的数据集中记录的个数 B=预期被赋值的数据集中记录的个数
元数据完整性	数据集中的数据信息的完整性。	$X=A/B$ A=数据集中数据有元数据信息的数据量 B=数据集的总数据量
空值率	空值表示数据集中缺少的信息，计算数据集中为空的字段数。	$X=A/B$ A=数据集缺少的元素数 B=预期的数据集中的元素数
数据采集类别的完整性	获取原始数据时，需要采集到各类别的数据。	$X=A/B$ A=采集数据的类别数 B=数据集预期的数据类别数

4.1.3 准确性

准确性^[1]是指数据与其描述的客观特征是否一致，包括源数据是否准确、数据值域是否在合理范围、结局变量随时间变化趋势是否合理、编码映射关系是否对应且唯一等。

数据集准确性的评估要求见表 2。

表 2 数据集准确性评估要求

评价指标	指标定义	计算公式
数据内容正确性	数据内容是否是预期数据。	$X=A/B$ A=满足数据正确性要求的数据集中元素的个数 B=被评价的数据集中元素个数 注：对于数据量很大的数据集，可以使用抽样的方法来计算其正确性。
数据重复率	特定字段、记录、文件或数据集意外重复的度量。	$X=A/B$ A=重复数据集中元素的个数 B=被评价的数据集中元素个数
数据唯一性	特定字段、记录、文件或数据集唯一性的度量。	$X=A/B$ A=满足唯一性要求的数据集中元素的个数

		B=被评价的数据集中元素个数
数据有效性	数据集中的元素符合数据集设计的需求。	$X=A/B$ A=满足数据集需求设计的元素个数 B=被评价的数据集中元素个数
标签准确率	数据集中的元素标注的准确率。	$X=A/B$ A=数据集中的元素标注正确的元素个数 B=数据集中的元素总个数

4.1.4 现时性

数据集现时性是指在给定环境中数据具有表征其正确寿命的属性的程度。
数据集现时性^[1]的评估要求见表 3。

表 3 数据集现时性评估要求

评价指标	指标定义	计算公式
基于时间段的正确性	基于日期范围内的记录数或频率分布符合业务需求的程度。	$X=A/B$ A=满足有效性要求的数据集中元素的个数 B=被评价的数据集中元素个数
基于时间点的及时性	基于时间戳的记录数、频率分布或延时时间符合业务需求的程度。	$X=A/B$ A=满足及时性要求的数据集中元素个数 B=被评价的数据集中元素个数
时序性	数据集中同一实体的数据元素之间的相对时序关系。	$X=A/B$ A=满足时序性要求的数据集中元素的个数 B=被评价的数据集中元素个数

4.1.5 无偏性

数据集的无偏性是数据集的数据分布偏差和数据内容存在偏见的程度。
数据集无偏性的评估要求见表 4。

表 4 数据集相关性评估要求

评价指标	指标定义	计算公式
分布无偏	数据集中各属性类别的整体分类的偏差程度。	$X=A/B$ A=各类别数量差异程度 B=数据集的总数据量
历史偏见	在历史过程中存在的对某些群体或事件的不公正看法或对待方式，导致历史	$X=A/B$ A=存在历史偏见的数据量

	记录中的信息出现倾向性或不客观的现象。	B=数据集的总数据量
主观偏见	数据集在收集和处理过程中受到主观因素的影响,从而导致数据反映了一定的倾向或偏见。	$X=A/B$ A=存在主观偏见的数据量 B=数据集的总数据量

4.1.6 相关性

数据集的相关性是指在给定环境中,数据集(假设其准确、完整、一致、最新等)适用于给定上下文的程度。相关性意味着其所选特征及其数据值是目标变量的良好预测因子。

数据集相关性的评估要求见表 5。

表 5 数据集相关性评估要求

评价指标	指标定义	计算公式
特征相关性	数据集中与给定上下文相关的特征的比率。	$X=A/B$ A=数据集中被认为与数据使用相关的特征数量 B=数据集的特征总数
记录相关性	数据集中与给定上下文相关的记录的比率。	$X=A/B$ A=数据集中被认为与数据使用相关的记录数 B=数据集的记录总数

4.1.7 安全性

数据集安全性的相关要求应符合GB/T 37988-2019《信息安全技术 数据安全能力成熟度模型》、GB/T 41819-2022《信息安全技术 人脸识别数据安全要求》、GB/T 41807-2022《信息安全技术 声纹识别数据安全要求》、GB/T 42016-2022《信息安全技术 网络音视频服务数据安全要求》等国家标准文件的相关规定。

4.1.8 数据集共享

数据集共享的评估要求见表 6。

表 6 数据集共享的评估要求

评价指标	指标定义	计算公式
数据共享度量	在共享数据时,对敏感数据进行消隐等操作。	$X=A/B$ A=消除敏感数据的数据量 B=需要消除敏感数据的总数据量

4.2 训练数据集扩展要求

4.2.1 多样性

数据集的多样性是指样本之间在目标数据方面的差异。在用于人工智能模型的数据集中，样本之间的充分差异很重要。如果数据集中的所有或大多数数据记录都相似，从该数据集训练的人工智能模型可能存在过度拟合的风险，因此不太具有可推广性。数据集的多样性代表了数据集如何包含各个值域、标签、聚类 and 单个数据之间的分布。生成式模型的数据增强可以提高数据的多样性，但如果原始数据集的多样性有限，这些方法可能会失败。多样性与代表性和均衡性密切相关。多样性是一种数据质量特征，可以用来评估数据集的保真度。

数据集多样性的评估要求见表 7。

表 7 数据集多样性评估要求

评价指标	指标定义	计算公式
标签丰富性	数据集中不同标签的数量。	A A=数据集中不同标签的数量
相对标签丰度	数据集中具有相同标签的单个数据（即项目、记录、帧）数量的一部分。	$X=A/B$ A=目标标签所在的单个数据的数量 B=数据集中单个数据的数量
类别大小多样性	分类样本数低于阈值的类别比率。	$X=A/B$ A=分类样本数低于阈值的类别数 B=类别总数

4.2.2 代表性

代表性定义为样本反映所研究的目标群体的程度。对于监督学习，训练数据集可以被视为样本，生产数据可以被认为是研究中的群体。当训练数据不能充分代表生产数据时，人工智能模型的训练可能无法按要求执行。

代表性数据质量特征与相关性数据质量特征相关，因为不代表所研究人群的数据集不太可能为目标变量提供良好的预测因子。

数据集代表性的评估要求见表 8。

表 8 数据集代表性评估要求

评价指标	指标定义	计算公式
代表性比率	在样本中发现的目标属性与总体数据集中发现的属性的比率。	$X=A/B$ A=样本中目标属性的数量 B=总体中属性的数量

4.2.3 可审核性

可审核性是指数据集的特征，即数据集的全部或部分经过了审核，或者相关利益相关者可以使用这些数据进行审核。对用于分析的数据集进行审核有助于提高数据的可信度，并且可能是符合要求所必需的。

例如：图像数据集用于图像识别，并由第三方承包商标记。为了确保图像正确标记，组织使用单独的第三方来审核标记图像的子集。

数据集可审核性的评估要求见表 9。

表 9 数据集可审核性评估要求

评价指标	指标定义	计算公式
经审核的记录	数据集中已审核的记录的比率。	$X=A/B$ 式中： A=数据集中已审核的记录数 B=数据集中的记录总数
可审核的记录	数据集中可供审核的记录的比率。	$X=A/B$ 式中： A=数据集中可供审核的记录数 B=数据集中的记录总数

4.3 测试数据集扩展要求

4.3.1 均衡性

当遇到数据集不均衡的情况时，优先建议标注更多的少数类样本，进而使各个类别的样本达到平衡。如果客观条件下难以标注更多的少数类样本，则建议通过在数据层面的数据重采样技术或算法层面的代价敏感算法等措施予以改进。

数据集均衡性的评估要求见表 10。

表 10 数据集均衡性评估要求

评价指标	指标定义	计算公式
群体稳定性	指标准数据集中每个变量在训练集、验证集和测试集中的分布稳定性。	$PSI = \sum_{i=1}^B (\hat{p}_i - \hat{q}_i) \times \ln \frac{\hat{p}_i}{\hat{q}_i}$ 式中，i 为数据的第 i 个分组；B 为总分组数；$\hat{p}_i$ 为验证或测试第 i 个分组占比；$\hat{q}_i$ 为训练第 i 个分组占比。
均值稳定性	指标准数据集中每个变量在训练集，验证集和测试集中的均值稳定性。	$U = \frac{\sum_{i=1}^n X_i}{n}$ 式中，X_i 为第 i 个样本；n 为样本数目。
方差稳定性	标准数据集中每个变量在训练集，验证集和测试集中的方差稳定性。	$\sigma^2 = \frac{\sum_{i=1}^n (x_i - u)^2}{n - 1}$ 式中，X_i 为第 i 个样本；n 为样本数目；U 为均值。
偏度稳定性	标准数据集中每个变量在训练集，验证集和测试集中的偏度稳定性。	$S = \frac{E (X_i - u)^3}{\sigma^3}$ 式中，X_i 为第 i 个样本；σ^3 为三阶标准差；U 为均值。
峰度稳定性	标准数据集中每个变量在训练集，验证集和测试集中的峰度稳定性。	$K = \frac{E (X_i - u)^4}{\sigma^4}$ 式中，X_i 为第 i 个样本；σ^4 为四阶标准差；U 为均值。

4.3.2 有效性

数据集的有效性表明数据集是否满足在特定任务中使用的要求。测试数据集与训练数据集应该是不同的，并且它们的特征分布应该相同。

例如：对于计算机视觉系统，数据集有效性可以是亮度或分辨率低于所需阈值的图像数量相对于数据集中所有图像或视频的最低可接受比率。对于图像分类系统，数据集有效性可以指属于一个类别的图像数量与数据集中图像总数的最低可接受比率。对于目标检测系统，数据集有效性可以是指在数据集中的所有图像或视频中，边界框的面积数量低于所需阈值的图像的数量的最低可接受比率。

数据集有效性的评估要求见表 11。

表 11 数据集有效性评估要求

评价指标	指标定义	计算公式
特征有效性	数据集中具有可接受特征的样本的比率。	$X=A/B$ 式中： A=具有可接受特征的样本数量 B=数据集中所有样本的数量
类别大小有效性	分类样本数低于阈值的类别比率。	$X=A/B$ 式中： A=分类样本的数量低于阈值的类别的数量 B=类别总数
标签有效性	数据集中具有可接受标签的样本的比率。	$X=A/B$ 式中： A=带有可接受标签的样品数量 B=数据集中所有样本的数量

4.3.3 易用性

在模型研发阶段，对于测试过程中所使用的数据集应满足数据易用性要求，能够基于深度学习框架实现相关的数据集操作。关于数据易用性要求如下：

1、可读取性

数据集应能支持深度学习框架相关接口完成读取功能，从而支持便捷地加载和处理多种类型的数据集，确保数据操作的灵活性和便捷性。

2、可融合性

数据集应能支持基于深度学习框架的融合功能，从而增强处理复杂数据的能力，深度学习框架应支持数据集融合功能，允许将多个数据集合并为一个，以满足数据整合化的需求。

3、可分割性

数据集应能支持基于深度学习框架的分割功能。深度学习框架应提供数据集分割功能，允许根据需求将原始数据集分成多个小组，例如用于训练、验证和测试的数据集，以支持不同的训练策略。

4、可采样性

数据集应能支持基于深度学习框架的采样功能。为满足不同采样策略和需求，深度学习框架应提供多种采样器接口，包括顺序采样、随机采样和按权重随机采样等，提供灵活的数据处理选择。

5、可批处理性

数据集应能支持基于深度学习框架的批处理功能。深度学习框架应支持数据集批处理机制，允许定义批量采样器，实现高效的数据处理和训练。

6、可加载优化性

数据集应能支持基于深度学习框架的加载优化功能。深度学习框架应支持智能的数据加载优化机制，如支持多进程数据读取器，并根据实际情况智能调整数据加载方式，以提高数据读取效率和整体性能。

7、可链路操作性

数据集应能支持基于深度学习框架的链路操作功能。深度学习框架应支持数据集样本的级联操作，允许进行复杂的数据处理和增强操作。

8、可子集操作性

数据集应能支持基于深度学习框架的子集操作功能。深度学习框架应支持数据集子集操作和管理功能，允许从原始数据集中选择子集进行分析和处理，提升数据处理的灵活性和精准性。

5 数据集质量评估流程

5.1 概述

数据集质量评估流程包括评估准备、评估执行、结果确认三个环节，如图1所示。

评价对象的质量应当由多个质量维度的评测来反映，单个数据质量测量是不能充分、客观评价由某一数据质量范围所限定的信息的质量状况，也不能为数据集的所有可能的应用提供全面的参考，多个质量维度的组合能提供更加丰富的信息。

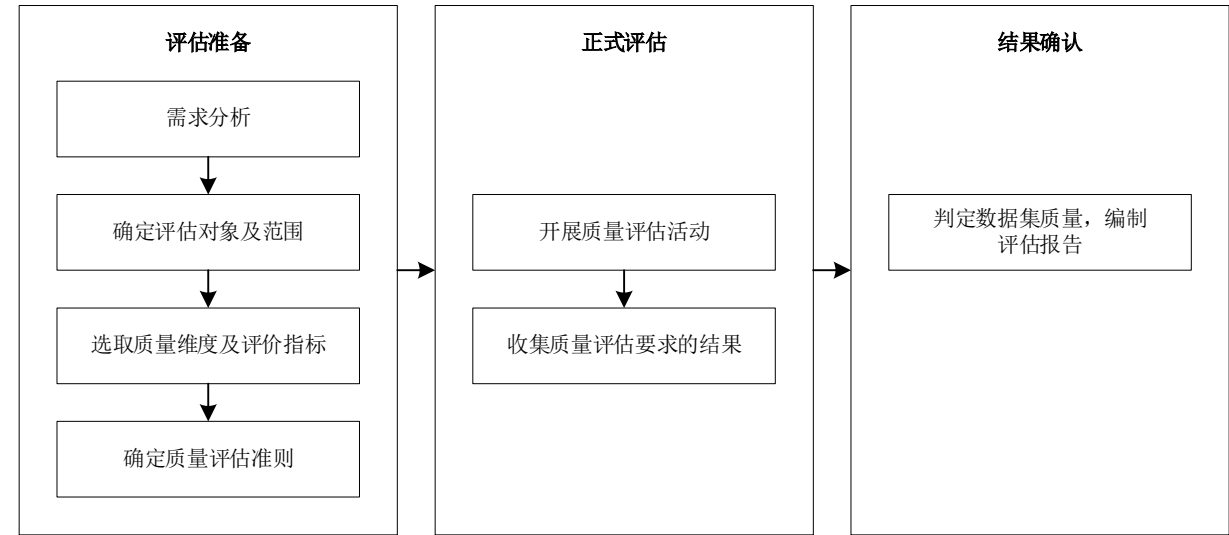


图 1 数据集质量评估流程图

5.2 评估准备

在评估准备阶段，制定本次评估的计划。评估计划包括需求分析、确定评估对象及范围、选取质量维度及评价指标、确定质量评估准则等四个方面。

5.2.1 需求分析

对具体业务数据的数据集质量评价是以业务需求为中心进行的。数据集资源不同于实体产品，具有用途个性化、多样化、不稳定等特点。因此，必须首先了解具体业务针对特定数据资源的需求特征才能建立针对性的评价指标体系。

5.2.2 确定评价对象及范围

确定当前评估工作应用的数据集的范围和边界，明确数据集在属性、数量、时间等维度的具体界限。需要说明的是，评价对象既可以是数据项也可以是数据集，但一定是一个确定的静态的集合。

5.2.3 选取质量维度及评价指标

选取质量评估要求的准则包括：

- a) 选择所有的评估项及对应的评估子项；
- b) 通用要求为必选的评估项，训练数据集扩展要求、测试数据集扩展要求、数据集安全性、数据集易用性及数据集分享为可选评估项。

注：可选的，针对数据集在人工智能产品的应用特性可以由数据集供应方和评估方等根据实际数据集情况或参考相似数据集在人工智能产品的质量评估案例协商或独立确定部分质量评估要求需要达到更高阈值要求。

5.2.4 确定质量评估准则

数据集质量评估应遵循以下准则：

- a) 选取的评估子项逐一达到评估阈值或评估范围；
- b) 选取的评估项逐一达到评估阈值或评估范围。

5.3 评估执行

在评估执行阶段，主要包括开展质量评估活动、收集质量评估要求的结果两个步骤。

可使用数据集质量评估工具开展质量评估活动及收集质量评估要求的结果，数据集质量评估工具应满足以下要求：

- a) 数据集质量评估工具应包括质量评估模型、质量分析评估、结果生成等功能模块，构建多模态、多维度的数据集质量评估模型，支持对图像、视频、语音、文本等模态的数据集进行评估。
- b) 数据集质量评估工具宜提供数据集质量评估自动化测试，用户可以根据需要设置好评估参数后实现一键化评估，并直观地了解质量评估的进度。
- c) 数据集质量评估工具宜提供可视化的评估结果，用户可以直观了解到数据集的完整性、准确性等各个要求的评分情况，支持评估报告的生成和下载。

5.3.1 开展质量评估活动

开展质量评估活动包括数据收集准备和质量评估开展。

- a) 数据收集准备的内容包括现场确认、工具有效性验证、不同质量评估要求的评估方式的确认、数据收集策略等。
- b) 质量评估开展的内容包括现场访谈、听取陈述、评审文档等。

5.3.2 收集质量评估要求的结果

收集质量评估要求的结果包括数据采集和数据确认。

a) 在数据采集阶段，质量评估要求的数据可以分为四类进行采集，即通过工具（数据集质量评估工具等）采集的数据、从介绍和演示中采集的数据、从面谈中获得的数据、以及从相关文档的评审中获得的数据。

b) 在数据确认阶段，根据质量评估要求采集的数据至少采用一项以下证实规则进行确认，即数据至少来自两个独立的来源、数据至少是在两个不同的采集期间获得的、数据至少被一个能反映确已做了某项规则的数据来源确认。

5.4 结果确认

评估方根据评估项和评估子项，最终判定被评估方的数据集质量并编制评估报告。

评估报告应包括评估总体结论和各评估项详细评估结果。各评估项的详细评估结果应包括每个评估项及其所有评估子项的评估结果。评估报告应满足评估机构及评估管理机构的编制要求。

参 考 文 献

- [1] GB/T 36344-2018 信息技术 数据质量评价指标
 - [2] ISO 8000-8:2015, IDT Data Quality——Part 8: Information and data quality: Concepts and measuring
 - [3] GB / T 25000.12-2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第12部分：数据质量模型
-