ICS 35.240
CCS L 70

中华人民共和国国家标准化指导性技术文件

GB/Z 237—2026

人工智能　可信赖数据集

Artificial intelligence—Trustworthy datasets

2026-08-27 发布

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 …… Ⅲ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　数据集可信赖要求 …… 3
5　数据集测试方法 …… 5
参考文献 …… 9

前　言

本文件为规范类指导性技术文件。

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：中国电子技术标准化研究院、中国科学院软件研究所、上海商汤智能科技有限公司、北京百度网讯科技有限公司、中科南京软件技术研究院、北京大学武汉人工智能研究院、蚂蚁科技集团股份有限公司、科大讯飞股份有限公司、上海翰纬信息科技有限公司、中移(杭州)信息技术有限公司、浙江大华技术股份有限公司、海信集团控股股份有限公司、杭州高新区(滨江)区块链与数据安全研究院、浪潮通用软件有限公司、国投智能信息科技股份有限公司、中移雄安信息通信科技有限公司、成都理工大学、江苏数兑科技有限公司、中电信数智科技有限公司、上海文鳐信息科技有限公司、贵州电网有限责任公司数智运营中心、深圳思谋信息科技有限公司、中国科学院工业人工智能研究所、浙江省质量科学研究院、北京航空航天大学、西北工业大学、讯飞医疗科技股份有限公司、湖南中医药大学、网思科技股份有限公司、超越科技股份有限公司、江苏锐创软件技术有限公司、电装智能科技(上海)有限公司、重庆通用人工智能研究院。

本文件主要起草人：马骋昊、杨光、叶珩、孟令中、吴庚、李斌斌、高卉、杨二龙、夏文华、薛云志、林冠辰、宋明、孟彦飞、咸静、孔维生、杨佳丽、蔺向楠、张驰、杜静漪、胡驰、王世海、武斌、辜凌云、周彩红、王静、杨坤、张文倩、阙锦龙、程晗蕾、左天祖、王维强、章海锋、林锋、杜忠田、余音、仲凯韬、潘宏博、鲍晨漪、马艳洁、张云畅、沈小勇、王鹏、王冀、王敏、贺志阳、梁昊、章古月、孟一丁、丁一凡、于治楼、支阿龙、刘天元、田浩、朱健、董彬超。

人工智能　可信赖数据集

1　范围

本文件规定了人工智能领域数据集的可信赖技术要求并描述了对应的测试方法。

本文件适用于人工智能领域的开发方、使用方以及第三方测评机构等相关组织对人工智能领域数据集的可信赖性开展规划、建设和评估工作。

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 25069—2022　信息安全技术　术语

GB/T 35273—2020　信息安全技术　个人信息安全规范

GB/T 41867—2022　信息技术　人工智能　术语

3　术语和定义

GB/T 25069—2022、GB/T 35273—2020、GB/T 41867—2022 界定的以及下列术语和定义适用于本文件。

3.1

数据集　dataset

具有一定主题，能标识并被计算机化处理的数据集合。

注：本文件中涉及的数据集主要指在其生存周期中，处于采集、存储、处理、分析、模型训练或应用等特定阶段，并服务于本文件所规范活动的数据集合。

[来源：GB/T 36344—2018，2.6，有修改]

3.2

可信赖　trustworthiness

满足利益相关方期望并可验证的能力。

注：可信赖的特征包括合规性、隐私保护、公平性、可追溯性、时效性、偏见缓解、安全性、相关性、一致性、可审核。

[来源：GB/T 41867—2022，3.4.2，有修改]

3.3

安全性　security

数据集在存储、传输及使用过程中，对信息的保密性、完整性和可用性的特性。

3.4

隐私保护　privacy protection

在数据集的采集、存储、处理、共享与销毁等全生存周期环节中，采取适当的技术和管理措施，防止个人信息被非法访问、泄露、滥用或通过重新识别等手段关联到特定自然人的过程。

注：隐私保护的目的是在数据处理过程中保障个人信息主体的合法权益，包括对其个人信息的知情权、决定权（控

制权)以及保密权等。

3.5

公平性　fairness

数据集的内容构成和标注方式对不同群体不存在系统性偏见的特性。

3.6

可追溯　traceability

数据集在其生命周期内,从产生、处理、存储、传输到销毁的每一个环节,都能被追踪和记录,确保数据的来源、历史和用途清晰可查的特性。

3.7

元数据　metadata

用于描述、解释、定位或以其他方式管理信息资源的结构化信息。

3.8

偏见　bias

在数据采集、处理、分析等过程中,由于主观或客观因素导致的系统性偏差,使数据无法准确、全面地反映真实世界。

注:标准数据偏见的成因包括数据源分布不均、采集手段有限、处理规则主观等。

3.9

偏见缓解　bias mitigation

通过系统性方法识别、评估、控制和持续减轻数据集中存在的偏见,以将其负面影响降低到特定应用场景可接受水平。

3.10

完整性　integrity

在数据集全生存周期内,其内容、结构、关联关系保持齐全无缺失、未被未授权篡改、未被意外破坏的属性。

3.11

毒数据　poisoned data

在数据处理过程中,因人为刻意投放、随机错误或系统偏差等原因产生的,可能导致 AI 模型失误的数据。

注:毒数据包括但不限于恶意篡改数据、虚假标注数据、逻辑冲突数据等。

3.12

时效性　timeliness

数据集在时间维度上与其应用场景和业务需求保持合理匹配的程度。

3.13

可审核　auditability

在符合法律法规、知识产权和隐私保护要求的前提下,能对数据集的来源、处理过程、标注结果或使用情况进行检查、验证或复核。

3.14

匿名化　anonymization

通过对个人信息的技术处理,使得个人信息主体无法被识别或者关联,且处理后的信息不能被复原的过程。

注:个人信息经匿名化处理后所得的信息不属于个人信息。

[来源:GB/T 35273—2020,3.14]

3.15

去标识化　de-identification

通过对个人信息的技术处理,使其在不借助额外信息的情况下,无法识别或者关联个人信息主体的

过程。

注：去标识化建立在个体基础之上，保留了个体颗粒度，采用假名、加密、哈希函数等技术手段替代对个人信息的标识。

[来源：GB/T 35273—2020，3.15]

4 数据集可信赖要求

4.1 合规性

数据集满足以下合规性要求：

a) 数据采集需遵循“知情同意”“数据最小化”原则；

b) 跨境数据合规：跨境数据集需提供数据出境安全评估报告、标准合同（如适用）、接收方安全保障承诺等文件；对跨境传输的数据集实施加密传输、访问日志留存（至少 6 个月）、定期跨境合规审计等措施。

4.2 隐私保护

数据集满足以下隐私保护要求：

a) 确保不涉及个人隐私、商业秘密和国家秘密；

b) 对于涉及敏感领域的数据，建立严格的准入和授权使用机制；

c) 根据敏感程度进行分级，并采取相应的安全保护措施；

d) 在数据流通前进行有效的匿名化或去标识化处理，以降低隐私泄露风险；

e) 采用加密机制，保护敏感数据在传输和静态存储过程中的安全，防止非授权访问；

f) 建立并实施严格的权限管理机制和身份鉴别机制，确保数据集的访问、操作（如查询、下载、修改）仅限于完成既定任务所必需的授权主体，并留存操作日志；

g) 在处理或分析高风险应用场景中，采用差分隐私、同态加密或联邦学习等隐私增强技术，进一步降低逆向破解隐私信息的风险；

h) 建立基于最小权限原则的访问控制机制，明确授权用户、使用目的及访问日志记录。

4.3 公平性

数据集满足以下公平性要求。

a) 确保数据集在采集阶段具有样本分布的代表性；涉及性别、年龄、地域、族裔等敏感属性时，其分布应符合预设的均衡要求；对于占总体比例较低的少数类群体，其在数据集中的样本占比与真实占比保持合理偏差。

b) 公平性要求应覆盖群体公平和个体公平等不同层面，并贯穿于数据生存周期全过程。

c) 在数据预处理和特征选择阶段实施偏见识别与消除措施；涉及数据清洗和特征提取时，应评估并阻断潜在偏见的传递或放大。

d) 通过加权采样、合成数据等手段，修正因采样不均衡导致的数据集系统性分布偏差。

e) 识别并评估数据中反映历史时期的特定社会偏见，并在预处理阶段采取去关联化或标签校准等措施，防止历史偏见在模型中延续。

f) 对标注任务进行去主观化设计，通过多员标注、盲审、一致性检验等方式，减少标注人员个人偏见对数据真实性的影响。

g) 建立数据集偏见动态监控机制，定期对数据集及基于其训练的模型进行偏见检测，若发现偏见超标，应及时启动偏见修正流程。

4.4 可追溯

数据集满足以下可追溯的要求。

a) 来源可追溯：能够清晰记录数据的原始来源，包括采集的时间、方式等。

b) 标注可追溯：记录标注任务的元数据，包括但不限于标注规则、标注工具版本、标注人员（或自动标注算法）标识、审核记录以及标注指令的历史变更。

c) 版本可追溯：能够唯一标识并记录数据集的不同版本，包括版本号、创建/更新时间、变更内容、变更原因及相关责任人。

d) a)～c)所要求的内容使用元数据进行记录，需满足以下要求：

   1) 元数据文件以确保其完整性与可用性的方式（如使用防篡改、抗抵赖技术）进行保存；

   2) 保存期限至少不短于数据集本身的使用与保存期限，并满足相关法律法规及合同约定的要求；

   3) 根据其敏感程度采取相应的访问控制与保密措施。

注：对于具有特殊追溯需求的领域（如电力、金融、医疗等），基于上述通用方法，结合行业监管要求与数据特点，增补特定的追溯点。

4.5 时效性

对于动态更新类数据集，数据集满足以下时效性要求：

a) 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；

b) 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；

c) 时序性：同一实体的多条记录保持合理的时序逻辑，避免时间倒置或跳跃异常；

d) 更新时效性：数据集建立定期或不间断的动态更新机制，更新周期与业务场景相匹配；

e) 多源时间对齐：对于由多个独立数据源汇聚形成的动态数据集，不同来源数据之间具备可协调的时间基准与采集节奏。

4.6 安全性

数据集满足以下安全性要求。

a) 保密性：数据集不被未授权的个人或实体访问或泄露。这包括使用加密技术来保护数据传输和存储过程中的安全，以及实施严格的访问控制措施（如身份鉴别、权限最小化分配、访问日志审计等）。

b) 完整性：采取校验、校验码、数字签名或加密哈希算法等技术手段，防止数据在存储或传输过程中被未授权篡改。

c) 可用性：数据集在需要时被授权用户访问和使用；建立数据备份与灾备机制，确保授权用户在需要时可正常访问。

d) 建立毒数据防御机制：

   1) 建立毒数据检测机制（如异常检测、来源验证、一致性校验等）；

   2) 对检测出的毒数据，实施隔离、标记或清除；

   3) 定期评估毒数据检测与排除机制的有效性（如毒数据排除率≥95%）。

e) 备份与恢复：建立定期备份机制，备份数据存储在独立安全环境，恢复测试频率宜每季度不低于一次。

f) 针对跨境使用的数据集，明确数据出境安全评估、加密传输标准、境外存储安全等要求，契合数据跨境监管需求。

g) 应急响应：建立数据安全事件应急响应机制。

4.7 相关性

数据集满足以下相关性要求：

a) 特征相关性：数据集所包含的特征与其目标应用领域或任务上下文高度相关；

b) 记录相关性：数据集中多数记录与其目标应用领域或任务上下文相关；

c) 领域相关性：数据集的整体构成（包括其特征与样本）与特定的应用领域或 AI 任务相适应。

4.8 一致性

数据集满足以下一致性要求：

a) 格式一致性：数据集中具有相同含义的数据项遵循统一的表示格式；

b) 逻辑一致性：同一数据集内部，数据记录之间或同一记录内的不同字段之间符合预定义的业务逻辑关系；

c) 相同数据一致性：当数据集由多个数据源集成时，来自不同源但对同一实体的描述信息保持一致；

d) 时序一致性：数据集中与时间序列相关的数据，其采集频率、时间戳格式和统计口径保持一致，具备时间维度上的可比性。

4.9 可审核

数据集满足以下可审核要求：

a) 制定并文档化数据集的审核程序，明确规定审核的频次、职责、方法（如自动化检查、人工抽样、第三方审计等）、合格标准以及不合格记录的处理流程；

b) 明确数据集中可供审核的数据范围，包括可审核的数据记录、元数据或处理过程；

c) 所有审核活动生成可追溯至数据集版本的记录，包括但不限于审核计划、检查项清单、抽样记录、审核发现、结论报告及批准记录，追溯保存期限不低于数据集有效期；

d) 对数据集执行覆盖范围、质量及合规性等方面的审核，记录审核结论，最终获准使用的数据集中，已审核记录的占比符合预定的质量目标；

e) 对于审核中发现的不合格记录、系统性偏差或潜在风险（如偏见、毒数据等），启动处置程序，处置措施可包括修正、剔除、标注或重新采集数据，并记录处置原因、措施和结果验证；

f) 涉及高风险或关键任务的应用场景（如金融、医疗、自动驾驶），数据集为独立的第三方审计提供必要的数据访问、元数据和过程文档支持，以验证其全生存周期的合规性与质量。

5 数据集测试方法

5.1 测试条件

测试条件包括但不限于：

a) 测试用例规模充分，覆盖采集、传输、存储、处理、共享到最终销毁的每一个阶段，而非仅集中在某一两个环节；

b) 针对所有可能接触数据集的角色类型，根据每种角色的授权边界内外行为设计用例；

c) 基于常见攻击模式反向推演数据集的薄弱点，针对每条潜在攻击路径构建验证用例；

d) 系统性地构造异常输入、边界值、格式畸变数据、并发冲突等异常场景，检验数据集的容错与隔离能力；

e) 针对时间因素对数据的影响，设计用例验证数据在不同生命周期状态之间迁移时，系统行为是否符合预期且日志记录完整；

f) 分析数据集中是否存在与敏感属性高相关的代理变量，针对这些代理变量的使用和组合查询设计测试用例；
g) 测试工具可靠、透明且可被审计，避免因工具缺陷导致漏报或误报，确保测试结果可复现、可追溯。

5.2 合规性

合规性的测试方法包括但不限于如下内容。
a) 检查数据集的采集活动是否遵循了“知情同意”与“数据最小化”原则。检查内容包括但不限于：获取个人信息的授权同意记录、隐私政策的公开情况，以及所收集数据的类型、范围、频度是否为实现处理目的所必需。
b) 检查涉及数据出境的数据集是否履行了法定的跨境提供合规义务。检查内容包括但不限于：明确涉及出境的数据范围，是否依法完成了数据出境安全评估，实际传输的数据范围、目的、接收方是否与申报或约定一致。
c) 检查对跨境传输的数据集所采取的安全保障措施。检查内容包括但不限于：是否使用了符合要求的加密技术进行传输，是否完整留存了访问日志且保存期限不少于六个月，以及是否建立了定期的跨境传输合规审计机制并保存审计记录。

5.3 隐私保护

隐私保护的测试方法包括但不限于：
a) 检查数据集中是否存在涉及个人隐私、商业秘密和国家秘密的内容；
b) 检查数据集是否建立严格的准入和授权使用机制；
c) 检查是否根据敏感程度进行分级，并采取相应的安全保护措施：
d) 检查在数据流通前是否进行有效的匿名化或去标识化处理；
e) 检查是否采用加密机制；
f) 检查是否建立并实施严格的权限管理机制和身份鉴别机制；
g) 检查是否采用隐私增强技术；
h) 检查是否建立基于最小权限原则的访问控制机制。

5.4 公平性

公平性的测试方法包括但不限于：
a) 对比数据集在不同群体（如性别、种族、年龄等）中的模型表现，检查是否存在对特定群体的系统性歧视；
b) 评估数据集中不同群体（如性别、种族、年龄等）的分布对其预期目标任务的代表性：当目标任务面向广泛人群时，应分析数据分布是否与目标应用场景的真实世界分布存在显著差异，对于定制化任务，应论证其分布设计的合理性；
c) 显性偏见筛查：通过统计分析或算法检测，识别数据集中可能存在的异常值、噪声以及明显带有偏见或歧视性内容的数据样本；
d) 代表性偏见评估：分析数据集中不同敏感属性群体（如性别、年龄组、地域）的样本分布，评估该分布是否可能因严重失衡而导致对某些群体的表征不足，进而引发模型偏见；
e) 历史偏见评估：对比不同时间段数据的分布差异、分析历史场景（如政策、技术条件）对数据的系统性影响；
f) 主观偏见评估：通过比较不同人对同一数据的标注结果，来检查是否存在个人主观偏见；
g) 特征-标签关联偏见检测：检查数据集中敏感特征或与敏感特征高度相关的特征，其取值在不

同目标标签类别或不同群体下的条件分布是否存在显著差异；

h) 缓解措施有效性验证：若已实施偏见缓解处理（如重采样、重新加权、特征调整等），通过比较处理前后数据集在 b）和 c）测试中的结果变化，验证缓解措施的有效性。

5.5 可追溯

可追溯的测试方法包括但不限于：

a) 检查数据集集成阶段是否具备并保存数据源、数据提供方、标注任务承担方及发布地址等信息的记录文件；

b) 检查数据集标注阶段是否保存了可支持追溯的关键元数据记录（如标注规则、标注人员/工具/环境标识、审核记录及标注规范的变更历史等）；

c) 检查数据集使用阶段是否有数据集版本管理记录文件，包括版本号、更新时间、更新内容、更新人员、变更依据等；

d) 若数据集包含通过模型合成或蒸馏技术产生的数据，检查是否记录其生成模型标识、源数据标识及关键生成参数等谱系信息；

e) 检查数据集注销阶段是否有数据集销毁记录，并评估所采用销毁技术方案的有效性与合规性。

5.6 时效性

时效性的测试方法包括但不限于：

a) 更新频率符合性检查：检查数据集更新的频率是否符合其目标应用场景（如实时、近实时、批量）的要求；

b) 时间信息完备性检查：检查数据集是否具有必要的时间信息的记录及时效说明，如数据生成时间戳、版本更新时间戳、数据有效期或失效条件；

c) 有效期有效性检查：检查数据集或其关键组成部分是否处于声明的有效期内；

d) 时序逻辑一致性检查：检查数据中与时间相关的逻辑关系是否正确；

e) 端到端延迟测试：在适用时，测试从数据产生到在数据集中可用（即可被查询、检索或用于训练）的端到端延迟，并验证该延迟是否符合业务规定或服务等级协议的阈值要求。

5.7 安全性

安全性的测试方法包括但不限于：

a) 通过模拟攻击（包括但不限于注入攻击、数据泄露漏洞攻击、恶意代码攻击等）、漏洞扫描等方式，检查数据集的存储和处理系统在存储、传输与访问接口等方面是否存在安全漏洞；

b) 通过完整性校验（如哈希值对比）或版本比对，检查数据是否发生了非授权的变更或损坏；

c) 检查数据集是否具备并启用了有效的访问控制机制，如身份鉴别、授权管理与访问日志记录等；

d) 检查数据集在传输和静态存储状态下是否采取了有效的安全防护措施，如加密或独立的网络/物理存储空间隔离；

e) 检查数据集是否具备有效的毒数据检测机制（如异常检测、来源验证）与明确的毒数据排除措施（如清洗、隔离、标记）；

f) 通过对比毒数据处理前后的相关评估结果（如模型鲁棒性、数据污染率），验证毒数据缓解措施的有效性；

g) 检查数据集是否具备数据备份、恢复及容灾能力，并定期验证其有效性。

5.8 相关性

相关性的测试方法包括但不限于：

a) 特征相关性测试：检查数据集所包含的每个特征字段，评估其与宣称的目标应用领域或任务上下文是否具有明确、合理的关联性；
b) 记录相关性测试：通过抽样或统计分析，检查数据集中记录的内容、标签或模式，评估其与目标应用领域或任务上下文相关的记录所占比例是否符合预期；
c) 领域相关性测试：综合评估数据集的整体构成（特征与记录），判断其是否与特定的应用领域或AI任务相适应，能否支撑有效的模型学习与预测。

5.9 一致性

一致性的测试方法包括但不限于：

a) 格式一致性测试：通过编写正则表达式或使用数据剖析工具，检查特定字段（如身份证号、电话号码、日期）的值是否符合预定义的格式规范；
b) 逻辑一致性测试：通过编写业务规则脚本或使用数据质量工具，检查数据记录内部及各记录间的逻辑关系；
c) 跨源一致性测试：通过数据比对工具或关联查询，对比来自不同数据源的同一实体的关键属性值，识别并统计差异；
d) 时序一致性测试：检查时间序列数据的时间戳间隔是否均匀，时间戳格式是否统一，并分析不同时间序列的数据在采集点上的对齐情况。

5.10 可审核

可审核的测试方法包括但不限于：

a) 过程与方法审核：检查数据集的采集、清洗、预处理（如归一化、标准化）等关键过程的方法是否被合理记录，且方法本身科学、可复现；
b) 内在一致性审核：检查数据集在时间顺序、空间关联、特征逻辑等不同维度上是否存在矛盾或冲突，确保数据自身逻辑自洽；
c) 标注质量审核：检查数据标注活动是否具备清晰、完整的规则文档，以及有效的质量控制和审核记录（如交叉验证记录、审核报告），以确保标注结果的可信度。

参 考 文 献

[1] 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

[2] GB/T 25000.24—2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第24部分:数据质量测量

[3] GB/T 35273—2020 信息安全技术 个人信息安全规范

[4] GB/T 36344—2018 信息技术 数据质量评价指标

[5] GB/T 45288.2—2025 人工智能 大模型 第2部分:评测指标与方法

[6] ISO/IEC 5259-2:2024 Artificial intelligence—Data quality for analytics and machine learning (ML)—Part 2: Data quality measures

[7] ISO 8000-8:2015 Data Quality—Part 8:Information and data quality:Concepts and measuring