ICS 35.030
CCS L 80

中华人民共和国国家标准

GB/T 45652—2025

网络安全技术　生成式人工智能预训练和优化训练数据安全规范

Cybersecurity technology—Security specification for generative artificial intelligence pre-training and fine-tuning data

2025-04-25 发布　　2025-11-01 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。

本文件由全国网络安全标准化技术委员会(SAC/TC 260)提出并归口。

本文件起草单位：北京中关村实验室、国家计算机网络应急技术处理协调中心、中国电子技术标准化研究院、北京大学、北京天融信网络安全技术有限公司、北京快手科技有限公司、阿里巴巴(北京)软件服务有限公司、北京百度网讯科技有限公司、清华大学、北京瑞莱智慧科技有限公司、天翼安全科技有限公司、中国移动通信集团有限公司、小米科技有限责任公司、阿里云计算有限公司、北京面壁智能科技有限责任公司、杭州萤石软件有限公司、北京理工大学、北京零一万物科技有限公司、中国科学院自动化研究所、联想(北京)有限公司、北京奇虎科技有限公司、科大讯飞股份有限公司、华为云计算技术有限公司、北京数安行科技有限公司、公安部第三研究所、蚂蚁科技集团股份有限公司、北京启明星辰信息安全技术有限公司、中国科学院计算技术研究所。

本文件主要起草人：徐恪、姚龙、张震、刘勇、谭知行、李琦、谢安明、许晓耕、杨光、崔天宇、郝春亮、张妍婷、薛智慧、郭建领、谷晨、姜文、叶晓俊、田天、梁伟、江为强、李家锟、彭骏涛、汪华东、郑鸿咚、洪延青、王海棠、朱贵波、孟遥、张向征、刘俊华、李峰风、刘玉红、刘楠、林冠辰、王龑、落红卫、谭映水、张峰、孙旭东、杜金浩、徐世真、安鹏、于阳、孙勇、郭洁昕、吴建亮、王霞、王金桥、高博雅、管铭、王士进、赵丽丽、王文宇、丁治国、蒋发群、盛强、吴博文。

引　言

预训练和优化训练数据是生成式人工智能的基础，直接决定了生成内容的质量和安全水平，但由于预训练和优化训练数据在收集、预处理、使用等处理活动中存在安全风险，亟需标准规范用于提高预训练和优化训练数据的安全水平。

网络安全技术　生成式人工智能预训练和优化训练数据安全规范

1　范围

本文件规定了生成式人工智能预训练和优化训练数据及其处理活动的安全要求，描述了相应的评价方法。

本文件适用于生成式人工智能服务提供者开展预训练和优化训练数据处理活动以及安全自评估，也适用于第三方机构对预训练和优化训练数据进行安全性评估。

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　信息安全技术　个人信息安全规范

GB/T 41479—2022　信息安全技术　网络数据处理安全要求

3　术语和定义

下列术语和定义适用于本文件。

3.1

生成式人工智能服务　generative artificial intelligence service

利用生成式人工智能技术向公众提供生成文本、图片、音频、视频等内容的服务。

[来源：GB/T 45654—2025，3.1]

3.2

服务提供者　service provider

以交互界面、可编程接口等形式提供生成式人工智能服务的组织或个人。

3.3

服务使用者　service user

使用生成式人工智能服务的组织或个人。

3.4

预训练　pre-training

使用大规模数据使生成式人工智能模型获得通用知识的训练过程。

3.5

优化训练　fine-tuning

在预训练基础上，使用特定领域数据使生成式人工智能模型获得面向领域服务能力的训练过程。

注：特定领域不限于某一个专业领域，通常覆盖多个领域。

3.6

预训练数据　pre-training data

用于生成式人工智能预训练的数据。

3.7

优化训练数据　fine-tuning data

所有用于生成式人工智能优化训练的数据。

3.8

元数据　metadata

定义和描述其他数据的数据。

[来源:GB/T 18391.1—2009,3.2.16]

3.9

统一资源定位符　unified resource location

用于标识互联网上资源位置的字符串。

注:通常包含协议类型(如 HTTP、FTP)、主机名、路径和查询参数等部分,用户通过其定位并访问资源。

4　通用安全要求

对服务提供者的要求如下。

a) 应制定人工智能预训练和优化训练数据的安全管理策略,包含对预训练数据和优化训练数据的保护组织、分类分级规则、数据处理活动安全、数据安全事件应急响应等。

b) 数据存储时应建立冗余备份等安全防护措施。

c) 数据传输过程中,应采取数据加密等安全防护措施,防范数据在传输过程中被窃取。

d) 训练阶段的数据应按照每批次进行安全隔离并在批次间建立数据标识,保证训练数据内容的可追溯性。

e) 开展预训练和优化训练数据处理活动时,应符合 GB/T 41479—2022 中第 5 章的相关要求。

f) 对预训练和优化训练数据中涉及个人信息的,其处理应满足 GB/T 35273 的相关要求;宜采用匿名化或去标识化技术,防止发生个人信息安全事件。

g) 应采用合理的安全保护措施及工具对训练和优化训练数据进行安全保护。

h) 对开展训练和优化训练数据处理活动的系统或平台宜至少满足等级保护三级要求。

i) 应建立并执行数据删除策略与规范,明确删除对象,经过审批并记录日志后,按数据主体请求在规定时间内删除其信息。

j) 宜采取措施确保删除的数据不能被恢复,例如重复覆写、多次格式化、物理销毁等。

k) 宜建立预训练和优化训练数据安全管理团队及监督职能部门,明确数据安全岗位和用户角色职责。

l) 应定期对预训练和优化训练数据开展安全评估,及时响应和处置预训练和优化训练数据安全事件,对涉及预训练和优化训练数据处理的关键岗位进行定期培训和考核。

m) 涉及行业数据的,应按照行业相关规定及行业标准要求采取相应的保护措施。

n) 宜对预训练和优化训练数据进行安全检测,修复或过滤被投毒数据,包括但不限于以下情况:

　1) 攻击者以降低算法模型整体表现为目的,置入大量标注错误或与设计开发目的无关的投毒数据;

　2) 攻击者以使算法模型对特定数据给出错误输出为目的,置入部分具备特定特征的投毒数据。

o) 应对预训练和优化训练数据进行真实性评估。

5 预训练数据处理活动的安全要求

5.1 数据收集

对服务提供者的要求如下。

a) 数据收集时，应对数据进行评估和记录，数据所包含的违法不良信息不应超过5%。

注1：本文件关注的违法不良信息主要是指包含GB/T 45654—2025中A.1～A.4中29种安全风险的信息。

b) 对自行收集的预训练数据，不应采集他人已明确不可采集的数据。

c) 收集开源数据集时，应遵循该数据集的开源许可协议或取得使用授权文件。

d) 对从外部数据源收集的预训练数据，应记录数据收集所涉及的数据来源：

1) 数据来源为互联网网站的，记录网站的统一资源定位符；

2) 数据来源为外部组织或个人的，记录数据集名称、来源组织，保存具备法律效力的交易合同、合作协议、许可协议或相关授权文件等；

3) 数据来源为服务使用者的，具有服务使用者的授权记录，并记录服务名称、服务使用者的标识。

e) 同类型的数据应具有多个不同的数据来源：

1) 不同的数据来源包含多个数据提供主体，包括但不限于互联网网站、其他组织或个人、服务使用者等；

2) 同类型数据中，每个数据来源的比例不低于1%。

注2：此处类型包括但不限于代码、图像、音频、视频及相同语言的文本等。

f) 所采集数据涉及个人信息的，应取得对应个人的同意或符合法律、行政法规规定的其他情形；所采集数据涉及敏感个人信息的，应取得对应个人的单独同意或符合法律、行政法规规定的其他情形。

g) 通过交易或合作等方式从其他组织或个人收集数据时，应对交易方或合作方所提供的数据、承诺以及相关证明材料进行审核。

h) 涉及数据跨境收集时，应符合相关数据跨境安全法规和标准要求。

5.2 数据预处理

对服务提供者的要求如下。

a) 应对数据进行抽样安全核验，经核验数据内容中含违法不良信息情况超过5%的，不应使用该来源数据进行训练。

b) 应确保预处理环境的安全性，进行数据处理的平台和工具安全性应与数据等级对应。

c) 宜定期对数据处理活动中涉及的平台和工具进行安全测试，防范安全风险和漏洞。

d) 应采用安全的传输、存储技术保障数据预处理活动安全。

e) 对预处理活动中涉及的个人信息，宜采取去标识化处理，并将可用于恢复识别个人的信息与去标识化后的信息隔离存储并加强访问和使用的权限管理。

f) 为数据样本添加元数据内容时，要求如下：

1) 数据样本已具有数据来源信息的，元数据内容应包括该信息；

2) 数据样本来源于互联网网站的，元数据内容应包括该样本自身或所在网页的统一资源定位符；

3) 数据样本来源于外部组织或个人数据集的，元数据内容应包括数据集名称、组织名称等信息；

4） 数据样本来源于服务使用者的，元数据内容应包括服务名称、服务使用者的标识等信息。

注 1：根据数据样本的特点加入其他相关元数据信息。

g） 对于每一种类型的训练数据，如文本、图片、音频、视频等，在将数据用于训练前，应对全部训练数据进行过滤，过滤方法包括但不限于关键词、分类模型、人工抽检等，去除数据中的违法不良信息，并记录处理情况。

h） 对训练数据中的主要知识产权侵权风险进行识别和管理，具体要求如下：

1） 应制定和完善训练数据知识产权管理策略和规则，并明确负责人；

2） 数据用于训练前，发现存在知识产权侵权等问题的，不应使用相关数据进行训练；

注 2：训练数据中包含文学、艺术、科学领域著作权作品的，重点识别训练数据及生成内容中著作权侵权问题。

3） 应建立针对知识产权问题的投诉举报渠道，并及时根据国家政策以及第三方投诉情况更新知识产权相关策略；

4） 应在用户服务协议中，向使用者告知使用生成内容的知识产权相关风险，并与使用者约定相关责任与义务。

i） 对不同模态训练数据预处理时，具体要求如下：

1） 对于文本数据，应设立关键词库、歧义关系库等敏感词库，建立文本分类模型识别隐私、敏感信息；

2） 对于图像数据，应设立敏感图像识别、分类模型或相关机制，过滤涉及违法不良信息、知识产权的图像数据；

3） 对于音频数据，应设立音频处理、识别、过滤模型或相关机制，过滤涉及违法不良信息、侵害知识产权的音频数据；

4） 对于视频数据，应结合图像、音频数据处理要求，并设立视频数据处理方法，过滤涉及违法不良信息、侵害知识产权的视频数据。

j） 对不同语言的训练数据预处理时，具体要求如下：

1） 应对涉及的不同语言数据充分评估，分别构建关键词库、分类模型等工具评估其安全风险；

2） 应对涉及不同语言翻译、转换的数据场景进行评估，保持转换前后的语义安全一致性。

k） 宜结合不同语种的语法、语境等语言逻辑设立具体的处理、过滤方法。

5.3 数据使用

对服务提供者的要求如下。

a） 应采取措施降低生成式人工智能被诱导生成安全风险内容的可能性，包括但不限于充分过滤已识别含有安全风险内容的数据样本等。

b） 如使用境外来源训练数据，在同一批次训练时应搭配使用合理比例的境内来源训练数据。

c） 涉及数据展示场景时，应对展示的必要性和安全性进行评估，不应展示含有违法不良信息的数据。

6 优化训练数据处理活动的安全要求

6.1 数据收集

对服务提供者的要求如下。

a） 优化训练数据的数据收集应符合 5.1 的要求。

b） 收集生成式人工智能服务生成的数据时，应记录所使用生成式人工智能服务的提供商、版本、获取时间、数据标识等信息。

c) 优化训练数据收集应：
   1) 对优化训练数据与优化目标的一致性进行检查，重点审查数据合规、领域适应等优化问题；
   2) 对优化训练数据错误知识、非恰当表达等问题进行检查；
   3) 对多轮迭代时的优化训练数据记录并公开历史版本的优化时间、优化内容等信息。
d) 收集垂直领域的优化训练数据时，应符合垂直领域的行业规定。
e) 收集垂直领域优化训练数据时，应建立该垂直领域数据检查机制，并对优化训练数据中存在的错误表述、隐含误导性、杜撰或篡改内容进行识别和处置。
f) 收集垂直领域优化训练数据时，宜收集获得相关行业认证或第三方公认权威认证的数据。

6.2 数据预处理

对服务提供者的要求如下。

a) 优化训练数据的数据预处理应符合 5.2 的要求。
b) 对优化训练数据进行过滤时：
   1) 应按照已建立的数据过滤机制进行数据过滤；
   2) 数据过滤的结果应使优化训练数据不包含已知的安全风险内容，未过滤的个人信息数据应当获得个人信息主体授权；
   3) 数据过滤过程中产生的中间数据或临时数据应及时删除。
c) 宜针对优化训练数据建立价值对齐检查机制，识别并处置不符合人类价值观或伦理道德的数据，相关的优化训练数据包括但不限于提示词、标注数据、知识蒸馏数据等。

6.3 数据使用

对服务提供者的要求如下。

a) 数据来源应符合 5.3 的要求。
b) 使用生成式人工智能生成内容作为训练数据时，应建立幻觉风险评估机制，识别并处置误导模型的错误知识。
c) 应以优化生成式人工智能模型为目标，选择使用优化训练数据：
   1) 在微调时，使用符合微调目标或垂直领域专业性的数据；
   2) 在对齐时，使用符合人类价值观或伦理道德的数据。
d) 宜建立优化训练数据质量的评价机制。

7 评价方法

7.1 通用安全评价方法

通用安全要求的评价方法、预期结果和结果判定如下。

a) 评价方法：
   1) 查看人工智能服务提供者是否具备人工智能预训练和优化训练数据的安全管理策略相关文档，查看文档中是否包含对预训练和优化训练数据的保护组织、分类分级规则、数据处理活动安全、数据安全事件应急响应等内容；
   2) 查看系统的运行日志和相关设备，检查是否对存储的预训练和优化训练数据进行了备份；
   3) 查看系统的运行日志和相关设备，检查在预训练和优化训练数据传输过程中是否采取了数据加密等安全防护措施；
   4) 查看系统的运行日志和相关设备，检查训练阶段的数据是否按照批次进行了安全隔离，批

次间是否建立了数据标识；

5） 检查服务提供者的数据处理活动是否符合 GB/T 41479—2022 中第 5 章的相关要求；

6） 检查服务提供者对预训练和优化训练数据中个人信息的处理是否符合 GB/T 35273 的相关要求；

7） 查看系统的设计文档、运行日志和相关设备，检查是否对预训练和优化训练数据采取了安全防护措施；

8） 检查进行预训练和优化训练数据处理活动的系统或平台是否满足等级保护三级的要求；

9） 查看服务提供者是否建立了数据删除策略与规范，查看系统运行日志和相关设备，检查是否按照相关要求或约定删除了预训练和优化训练数据；

10） 查看设计文档、运行日志和相关设备，检查服务提供者是否对删除数据采取了重复覆写、多次格式化、物理销毁等操作防止删除数据被恢复；

11） 查看组织架构文件，访谈相关人员，检查服务提供者是否建立了预训练和优化训练数据安全管理团队及监督职能部门，明确数据安全岗位和用户角色职责；

12） 查看安全评估报告、安全事件报告等，检查是否定期对预训练和优化训练数据进行了安全评估，并对发现的数据安全事件进行了响应和处置，查看培训记录、考核记录等，检查是否对涉及预训练和优化训练数据处理的关键岗位进行了定期培训和考核；

13） 检查服务提供者是否按照行业相关规定及行业标准要求，对行业数据采取了相应的保护措施；

14） 查看系统日志，检查是否对预训练和优化训练数据中的投毒数据进行识别，并对识别出的投毒数据进行修复或过滤；

15） 查看设计文档、运行日志和相关设备，检查是否对预训练和优化训练数据进行了真实性评估。

b） 预期结果：

1） 具备预训练和优化训练数据的安全管理策略，安全管理策略中包含对预训练和优化训练数据的保护组织、分类分级规则、数据处理活动安全、数据安全事件应急响应等内容；

2） 对存储的预训练和优化训练数据进行了备份；

3） 在预训练和优化训练数据的传输过程中采取了数据加密等安全防护措施；

4） 训练阶段的数据按照批次进行了安全隔离，批次间建立了数据标识；

5） 数据处理活动符合 GB/T 41479—2022 中第 5 章的相关要求；

6） 对预训练和优化训练数据中个人信息的处理符合 GB/T 35273 的相关要求；

7） 采取了身份鉴别、访问控制、加密、备份等技术措施，对预训练和优化训练数据进行了安全防护；

8） 进行预训练和优化训练数据处理活动的相关系统满足等级保护三级的要求；

9） 建立了数据删除策略与规范，并按约定删除了预训练和优化训练数据；

10） 对删除数据进行了重复覆写、多次格式化、物理销毁等操作；

11） 建立了预训练和优化训练数据安全管理团队及监督职能部门，明确了数据安全岗位和用户角色职责；

12） 定期对预训练和优化训练数据进行安全评估，并对发现的数据安全事件进行了响应和处置，对涉及预训练和优化训练数据处理的关键岗位进行了定期培训和考核；

13） 按照行业相关规定及行业标准要求，对行业数据采取了相应的保护措施；

14） 能识别对被投毒的预训练和优化训练数据，并进行修复或过滤；

15） 对预训练和优化训练数据进行了真实性评估。

c） 结果判定：

上述预期结果1)～7)、9)、12)、13)、15)均满足判定为符合，否则判定为不符合。预期结果8)、10)、11)、14)为可选评估项。

7.2 预训练数据处理活动评价方法

7.2.1 数据收集

预训练数据收集安全要求的评价方法、预期结果和结果判定如下。

a) 评价方法。

1) 查看采集时的数据来源安全评估记录，采用人工抽检的方式从中随机抽取不少于10%的记录，核查每条记录中的违法不良信息占比，以及违法不良信息占比超过5%时的处置记录。

2) 查看预训练数据收集记录，检查自行采集数据中是否包含他人已明确不可采集的数据。

3) 查看预训练数据收集记录，检查开源数据采集时是否遵循开源许可协议或具备相关授权文件。

4) 对从外部数据源收集的预训练数据，查看数据采集记录：

——数据来源为互联网网站的，检查记录中是否包含网站的统一资源定位符；

——数据来源为外部组织或个人的，检查记录中是否包含数据集名称、来源组织，检查是否保存有法律效力的交易合同、合作协议、许可协议或相关授权文件；

——数据来源为服务使用者的，检查记录中是否具有服务使用者的授权信息，检查记录中是否包含服务名称、服务使用者的标识。

5) 查看数据收集记录，检查每一种语言、模态的数据是否具有多个不同的数据来源：

——检查同一类型数据是否包含多个数据提供主体，包括但不限于互联网网站、其他组织或个人、服务使用者等；

——检查同一类型数据不同数据来源的比例。

6) 所收集数据中包含个人信息时，检查是否具有个人信息主体授权记录；所采集数据中包含敏感个人信息时，检查是否具有个人信息主体单独授权记录。

7) 以交易或合作等方式从其他组织或个人收集数据时，检查是否具备对交易方或合作方所提供训练数据、承诺以及相关材料进行审核的记录。

8) 所收集数据包含跨境数据时，核查其处理活动是否符合相关数据跨境安全法规和标准的要求。

b) 预期结果。

1) 数据来源安全评估记录中，违法不良信息占比超过5%的数据来源，未进行采集。

2) 自行采集数据中不包含他人已明确不可采集的数据。

3) 开源数据采集遵循了开源许可协议或具备相关授权文件。

4) 预训练数据收集记录中：

——数据来源为互联网网站的，记录中包含网站的统一资源定位符记录；

——数据来源为外部组织或个人的，记录中包含数据集名称、来源组织记录，保存了有法律效力的交易合同、合作协议、许可协议或相关授权文件；

——数据来源为服务使用者的，记录中保存了服务使用者的授权信息，包含了服务名称、服务使用者的标识。

5) 收集开源数据集中：

——同一类型数据具有不少于2个数据来源，包括但不限于互联网网站、其他组织或个人、服务使用者等；

——每种数据来源的比例不低于1%。

6） 所采集数据中包含个人信息时，具有相应的个人信息主体授权记录；所采集数据中包含敏感个人信息时，具有相应的个人信息主体单独授权记录；

7） 以交易或合作等方式从其他组织或个人收集数据时，记录中包含具备法律效力的交易合同、合作协议，交易方或合作方提供的数据来源、质量、安全等方面的承诺以及相关证明材料，以及服务提供者对交易方或合作方所提供训练数据、承诺、材料进行审核的记录；

8） 涉及数据跨境收集时，服务提供者的处理活动符合相关数据跨境安全法规和标准的要求。

c） 结果判定：

上述预期结果均满足判定为符合，否则判定为不符合。

7.2.2 数据预处理

预训练数据预处理安全要求的评价方法、预期结果和结果判定如下。

a） 评价方法。

1） 查看数据来源安全评估记录，采用人工抽检的方式从中随机抽取不少于10%的记录，核查每条记录中的违法不良信息占比，以及违法不良信息占比超过5%时的处置记录。

2） 检查数据在平台和工具中被确定的级别，查验组织内部对相应数据级别的安全要求，核查数据处理平台和工具的安全保障水平。

3） 检查数据处理活动中涉及的平台和工具安全测试记录及漏洞处理记录，对数据处理的平台和工具进行漏洞扫描测试。

4） 检查数据预处理活动中的传输、存储策略及对应的安全技术措施。

5） 检查数据预处理中的个人信息处理方式，是否采用去标识化技术或其他技术进行保护。

6） 随机抽样服务提供者预处理后的数据，对于每一种类型数据来源抽样数量不少于1 000个样本，检查抽样样本的来源及元数据内容：

——检查元数据内容是否包含数据样本原本所具有的数据来源信息；

——如数据样本来源于互联网网站，检查元数据内容是否包括该样本自身或所在网页的统一资源定位符；

——如数据样本来源于外部组织或个人数据集，元数据内容是否包括数据集名称、组织名称等信息；

——如数据样本来源于服务使用者，元数据内容是否包括服务名称、服务使用者的标识等信息。

7） 查看数据内容过滤记录，检查记录中是否覆盖每一种类型的训练数据，检查数据内容过滤记录的数据名称、过滤方法、过滤结果以及处理情况等信息；随机抽样服务提供者预处理后的数据，抽样数量不少于1 000个样本，检查样本是否具有违法不良信息；如有，检查是否有处理记录。

8） 检查知识产权侵权风险管理情况：

——检查是否制定和完善训练数据知识产权管理策略和规则，访谈知识产权负责人；

——随机抽样服务提供者预处理后的数据，抽样数量不少于1 000个样本，检查样本是否具有主要知识产权侵权风险识别记录及处置情况；

——检查是否建立针对知识产权问题的投诉举报渠道，通过人工测试等方式验证投诉举报渠道是否有效；查看知识产权相关策略更新日志，检查是否根据国家政策以及第三方投诉情况更新知识产权相关策略；

——查看用户服务协议，检查是否明确向使用者告知使用生成内容的知识产权相关风险，并与使用者约定相关责任与义务。

9） 检查是否对每种模态预训练数据建立对应的处理机制，以及机制的有效性：

——对于文本数据，检查是否设立关键词库、歧义关系库等敏感词库，是否建立文本分类模型识别隐私、敏感信息；

——对于图像数据，检查是否设立敏感图像识别、分类模型或相关机制，过滤涉及违法不良信息、侵害知识产权的图像数据；

——对于音频数据，检查是否设立音频处理、识别、过滤模型或相关机制，过滤涉及违法不良信息、侵害知识产权的音频数据；

——对于视频数据，检查是否应结合图像、音频数据处理要求，设立了视频数据测试数据集及相关测试方法，过滤涉及违法不良信息、侵害知识产权的视频数据。

10） 检查多语言数据预处理机制：

——检查是否对每种语言进行评估，并建立关键词库、分类模型等工具评估其安全风险；

——涉及不同语言翻译、转换的数据场景时，检查转换前后的语义安全一致性评估记录。

11） 检查是否结合不同语种的语法、语境等语言逻辑设立具体的处理、过滤方法。

b） 预期结果。

1） 数据来源安全评估记录中，违法不良信息占比超过5%的数据来源，未用作训练。

2） 数据预处理活动中平台和工具所处理的数据分级要求与组织内相关制度对应，平台和工具的安全水平与所处理的数据等级及安全要求适配。

3） 具备数据处理活动中涉及的平台和工具安全的定期测试记录及漏洞处理记录，平台和工具经过漏洞扫描测试，安全漏洞处理时效满足组织内制度要求。

4） 组织具备传输、存储策略，采用了安全的传输、存储技术，包括但不限于HTTPS、加密存储等技术。

5） 对个人信息采用了去标识化技术或其他技术进行保护，包括但不限于假名化、K匿名等。

6） 抽样样本元数据检查预期结果：

——元数据内容包含了数据样本原本所具有的数据来源信息；

——样本涉及互联网网站来源的，元数据内容中包含样本或样本所在网页的统一资源定位符；

——样本涉及其他组织或个人数据集来源的，元数据内容中包含数据集名称、来源名称的信息；

——样本涉及服务使用者来源的，元数据内容包含服务名称及服务使用者的身份标识号码记录。

7） 记录覆盖了每一种类型的训练数据，且正确、合理地记录了数据名称、过滤方法、过滤结果以及处理情况等信息；抽样样本均不含违法不良信息，或含有违法不良信息的已记录且未用于训练。

8） 知识产权侵权风险检查预期结果：

——服务提供者制定了训练数据知识产权管理策略和规则，并定期进行修订和完善，设置了知识产权负责人，并有效开展了工作；

——抽样样本不存在涉及主要知识产权侵权问题，或存在有主要知识产权侵权风险的，都已进行合法、有效的处置；

——服务提供者建立了针对知识产权问题的投诉举报渠道且渠道有效运作，能及时根据国家政策以及第三方投诉情况更新知识产权相关策略；

——用户服务协议中明确向使用者告知了使用生成内容的知识产权相关风险，并与使用者约定了相关责任与义务。

9） 服务提供者对每种模态预训练数据建立了对应的处理机制且机制有效：

——对文本数据，设立了关键词库、歧义关系库等敏感词库及文本分类模型，可识别隐私、敏感信息；

——对图像数据，设立敏感图像识别、分类模型，具备对违法不良信息、侵害知识产权的图像数据的过滤能力；

——对音频数据，设立音频处理、识别、过滤模型或相关机制，具备对违法不良信息、侵害知识产权的音频数据的过滤能力；

——对视频数据，结合图像、音频数据处理要求设立了视频数据测试数据集及相关测试方法，具备对违法不良信息、侵害知识产权的视频数据的过滤能力。

10） 具备多语言数据预处理机制：

——能对涉及的每种语言进行评估，并分别构建关键词库、分类模型等工具以进行安全风险评估；

——涉及不同语言翻译、转换的数据场景时，具备对转换前后的语义安全开展一致性评估的能力。

11） 结合不同语种的语法、语境等语言逻辑，为每种语言设置了具体的处理、过滤方法。

c） 结果判定：

上述预期结果1）、2）、4）、6）～10）均满足判定为符合，否则判定为不符合。预期结果3）、5）、11）为可选评估项。

7.2.3 数据使用

预训练数据使用安全要求的评价方法、预期结果和结果判定如下。

a） 评价方法：

1） 采用人工抽检方式从全部训练数据中随机抽取不少于4 000条数据，检查安全风险内容比例；采用关键词、分类模型等技术抽检方式从全部训练数据中抽取不少于总量10%的数据，检查安全风险内容比例；

2） 使用境外来源训练数据时，检查在同一批次训练情况下，搭配使用境内来源训练数据的记录；

3） 涉及数据展示场景时，检查服务提供者是否具有对展示的必要性和安全性的评估记录，展示的数据不包含违法不良信息。

b） 预期结果：

1） 人工抽检的抽样数据中，不含安全风险内容的数量占总抽样数量的比例不低于96%；技术抽检的抽样数据中，不含安全风险内容样本数量占总抽样数量的比例不低于98%；

2） 使用境外来源训练数据时，在同一批次训练情况下，具有合理搭配使用境内来源训练数据的记录；

3） 数据展示场景中，服务提供者对展示的必要性和安全性进行了评估，且留存相关评估记录；展示数据中均不含违法不良信息。

c） 结果判定：

实际评价结果与预期结果一致则判定符合，其他情况判定不符合。

7.3 优化训练数据处理活动评价方法

7.3.1 数据收集

优化训练数据收集安全要求的评价方法、预期结果和结果判定如下。

a） 评价方法。

1） 按照7.2.1a)规定的评价方法评价服务提供者优化训练数据的数据收集情况。
2） 查看训练数据收集记录，检查收集生成式人工智能服务生成的数据时是否记录了提供商、版本、获取时间、数据标识等信息。
3） 查看优化训练数据收集记录，从中随机抽取不少于10%的记录：
——检查收集过程中是否审查优化训练数据与优化目标一致性的记录信息，检查其中是否包括了数据合规、领域适应等方面的审查结论；
——检查收集过程中是否包括了错误知识、非恰当表达等方面的审查结论；
——对多轮迭代的优化数据，检查是否记录历次版本的优化时间、优化内容等信息。
4） 查看垂直领域优化训练数据的收集记录，随机抽取10%或不少于1 000条记录，检查是否包括垂直领域行业规定合规性检查的审查结论。
5） 查看管理策略或建设文档，检查在垂直领域数据的收集过程中是否建立了针对性的垂直领域数据检查机制；查看垂直领域优化训练数据的收集记录，随机抽取不少于10%或1 000条记录，检查是否包括数据错误表述、隐含误导性、杜撰或篡改内容的识别结果和处置结论。
6） 查看垂直领域优化训练数据的收集记录，随机抽取不少于10%或1 000条记录，检查是否包括数据源的行业认证或第三方公认权威认证的审查结论。

b） 预期结果。
1） 符合7.2.1b)的预期结果。
2） 收集生成式人工智能服务生成的数据时，优化训练数据收集记录中包含了所使用人工智能服务的提供商、版本、获取时间、数据标识等信息。
3） 抽取的优化训练数据收集记录中：
——均具有优化训练数据与优化目标一致性的记录信息，且其中包括了数据合规、领域适应等方面的审查结论；
——均包括了错误知识、非恰当表达等方面的审查结论；
——对多轮迭代的优化数据，记录中均包括了历次版本的优化时间、优化内容等信息。
4） 垂直领域优化训练数据抽取的收集记录中，均包括了垂直领域行业规定合规性检查的审查结论。
5） 管理策略或建设文档中体现了在垂直领域数据的收集过程中建立了针对性的检查机制；对垂直领域优化训练数据抽取的收集记录中，均包括了数据错误表述、隐含误导性、杜撰或篡改内容的识别结果和处置结论。
6） 对垂直领域优化训练数据抽取的收集记录中，均包括了数据源的行业认证或第三方公认权威认证的审查结论。

c） 结果判定：
上述预期结果1)～5)均满足判定为符合，否则判定为不符合。预期结果6)为可选评估项。

7.3.2 数据预处理

优化训练数据预处理安全要求的评价方法、预期结果和结果判定如下。

a） 评价方法。
1） 按照7.2.2a)规定的评价方法评价服务提供者优化训练数据的数据预处理情况。
2） 对优化训练数据进行过滤时：
——检查是否建立优化训练数据过滤机制，数据过滤时是否遵循该机制；
——从过滤后的数据中随机抽取不少于10%或1 000条数据，检查是否包含已知的安全风险内容，检查过滤后剩下的数据是否已获得个人信息主体授权；

——查看数据过滤过程中产生的中间数据或临时数据是否被及时删除。

3） 检查是否针对提示词、标注数据、知识蒸馏数据等优化训练数据建立价值对齐检查机制；查看价值对齐检查记录，检查是否识别并处置不符合人类价值观或伦理道德的数据。

b） 预期结果。

1） 符合 7.2.2b)的预期结果。

2） 检查优化训练数据过滤安全管理策略文档或管理制度：

——制定了优化训练数据过滤机制；优化训练数据过滤记录符合所制定的过滤机制；

——所抽取数据中，不包含已知的安全风险，过滤后剩下的数据已获得个人信息主体授权；

——已及时删除数据过滤过程中产生的中间数据或临时数据。

3） 在安全管理策略文档或管理制度中，针对提示词、标注数据、知识蒸馏数据等优化训练数据建立价值对齐检查机制；价值对齐检查记录中，包括了不符合人类价值观或伦理道德数据的识别结果和处置结论。

c） 结果判定：

上述预期结果 1)、2)均满足判定为符合，否则判定为不符合。预期结果 3)为可选评估项。

7.3.3 数据使用

优化训练数据使用安全要求的评价方法、预期结果和结果判定如下。

a） 评价方法。

1） 按照 7.2.3a)规定的评价方法评价服务提供者优化训练数据的数据使用情况。

2） 查看安全管理策略文档和管理制度，检查在使用人工智能生成内容作为训练数据时，是否建立幻觉风险评估机制；查看幻觉风险评估记录，检查是否包含幻觉风险识别结果和处置结论。

3） 针对优化训练数据的选择使用，查看安全管理策略文档和管理制度、优化训练数据使用记录：

——检查安全管理策略文档和管理制度中是否明确了以优化生成式人工智能模型为目标、使用符合微调目标或垂直领域专业性的数据的条款；检查优化训练数据使用记录是否包括了微调数据选择使用的相关说明；

——检查安全管理策略文档和管理制度中是否明确了以优化生成式人工智能模型为目标、使用符合人类价值观或伦理道德的数据的条款；检查优化训练数据使用记录是否包括了对齐数据选择使用的相关说明。

4） 查看安全管理策略文档和管理制度，检查是否建立优化训练数据质量的评价机制。

b） 预期结果。

1） 符合 7.2.3b)的预期结果。

2） 安全管理策略文档或管理制度中，在关于使用人工智能生成内容作为训练数据的相关内容中，明确要求了建立幻觉风险评估机制；查看幻觉风险评估记录，包含了幻觉风险识别结果和处置结论。

3） 针对优化训练数据选择使用：

——安全管理策略文档或管理制度中，明确了以优化生成式人工智能模型为目标、使用符合微调目标或垂直领域专业性的数据的要求；优化训练数据使用记录中，包括了微调数据选择使用的相关说明；

——安全管理策略文档或管理制度中，明确了以优化生成式人工智能模型为目标、使用符合人类价值观或伦理道德的数据的要求；优化训练数据使用记录中，包括了对齐数据

选择使用的相关说明。

4) 安全管理策略文档或管理制度中,明确了要建立优化训练数据质量的评价机制。

c) 结果判定:

上述预期结果 1)～3)均满足判定为符合,否则判定为不符合。预期结果 4)为可选评估项。

参 考 文 献

[1] GB/T 18391.1—2009 信息技术 元数据注册系统(MDR) 第1部分:框架
[2] GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求