ICS 35.030
CCS L 80

中华人民共和国国家标准

GB/T XXXXX—XXXX

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

（征求意见稿）

（本稿完成日期：2024 年 3 月 28 日）

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。

XXXX－XX－XX 发布　　XXXX－XX－XX 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。

本文件由全国网络安全标准化技术委员会（SAC/TC260）提出并归口。

本文件起草单位：（名单根据实际情况决定）

本文件主要起草人：（名单根据实际情况决定）

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，描述了对应的评价方法。

本文件适用于指导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开展与训练预训练和优化训练数据安全自评价，也可为监管评估提供参考。

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 AAAAA 网络安全技术 生成式人工智能数据安全标注规范

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 generative artificial intelligence

具有文本、图片、音频、视频等内容生成能力的人工智能系统。

3.2

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

利用生成式人工智能技术提供生成文本、图片、音频、视频等内容的服务。

3.3

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和个人。

3.4

服务使用者 service user

使用生成式人工智能服务的组织或个人。

3.5

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。

3.6

优化训练 fine-tuning

使用专门领域数据使生成式人工智能模型获得一定面向领域服务能力的训练过程。

3.7

预训练数据 pre-training data

所有用于生成式人工智能预训练的各类数据。

3.8

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各类数据。

4 概述

4.1 安全风险

生成式人工智能预训练和优化训练数据的安全性涉及数据自身的安全性以及生成式人工智能服务的安全性两方面。生成式人工智能的预训练和优化训练数据面临的安全风险有：

a) 数据泄露、数据窃取等风险；
b) 数据投毒风险；
c) 其他因训练数据影响生成式人工智能安全性的风险。

4.2 安全框架

生成式人工智能预训练和优化训练数据安全框架包括数据通用安全以及数据处理活动安全。数据通用安全主要包括分类分级、安全防护、安全检测、审计追溯、应急响应等。数据处理活动安全主要包括数据收集、数据预处理、数据使用等活动的安全。

生成式人工智能预训练和优化训练数据安全框架如图1所示。

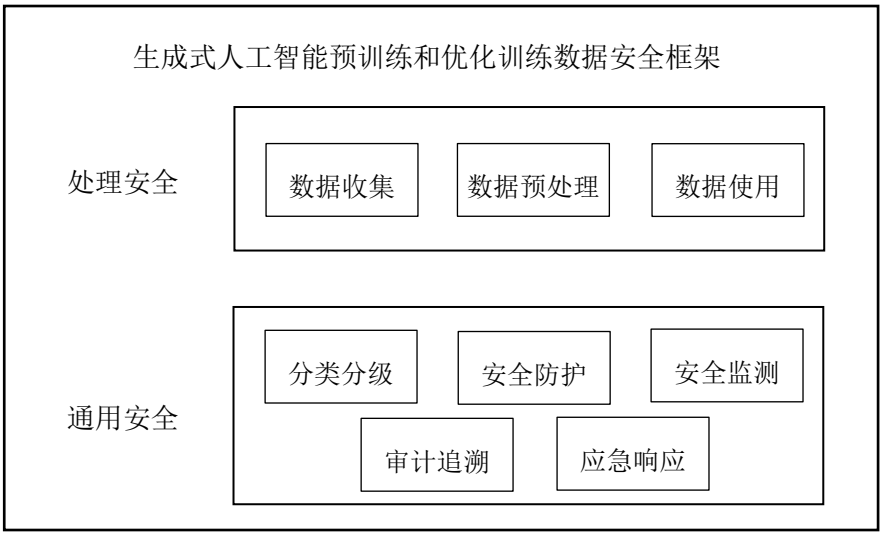

图1 生成式人工智能预训练和优化训练数据安全框架

5 通用安全要求

对服务提供者的要求如下。

a) 应对预训练和优化训练数据进行分类分级管理。
b) 应采取技术措施对预训练和优化训练数据进行安全监测，发现数据安全缺陷、漏洞等风险时及时告警并采取相应的处置措施。
c) 应采取身份鉴别、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安全防护。

d) 应建立针对预训练和优化训练数据安全事件的应急响应机制，及时有效处置发生的数据安全事件，不影响或能够尽快恢复业务的运营。
e) 应对预训练和优化训练数据的数据收集、数据预处理、数据使用等的数据处理活动进行记录，确保预训练和优化训练数据处理活动的关键操作可审计、可追溯。

6 预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) 应记录数据收集所涉及的数据来源，保存相关信息：
   1) 数据来源为互联网网站，记录网站的统一资源定位符；
   2) 数据来源为其他组织或个人，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，记录服务名称、服务使用者的身份标识号码，保存服务使用者的授权记录。
b) 同类型的数据应具有多个不同的数据来源。
   注：代码、图像、音频、视频及相同语言的文本等视为同类型的数据。
c) 通过互联网网站收集数据时，应记录所收集数据或数据所在网页的统一资源定位符。
d) 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、材料进行审核。

6.2 数据预处理

对服务提供者的要求如下。

a) 应为数据中所有数据样本添加元数据内容：
   1) 数据样本已具有数据来源信息的，元数据内容为该信息；
   2) 数据样本来源于互联网网站的，元数据内容为该样本自身或所在网页的统一资源定位符；
   3) 数据样本来源于其他组织或个人数据集的，元数据内容为数据集名称、组织名称等信息；
   4) 数据样本来源于服务使用者的，元数据内容为服务名称、服务使用者的身份标识号码等信息。
b) 应采取关键词、分类模型、人工抽检查等方式对数据含有安全风险内容情况进行识别，并记录识别情况。
   注：安全风险内容见附录A中定义的29类；关键词、分类模型要求见附录B。
c) 应对数据中的主要知识产权侵权风险进行识别并记录，例如数据中包含文学、艺术、科学作品的，重点识别数据的著作权侵权问题。

6.3 数据使用

对服务提供者的要求如下。

a) 使用包含个人信息的数据时，应取得对应个人同意或符合法律、行政法规规定的其他情形。
b) 使用包含敏感个人信息的数据前，应取得对应个人单独同意或符合法律、行政法规规定的其他情形。
c) 不应使用存在知识产权侵权问题的数据。
d) 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。

7 优化训练数据处理活动的安全要求

7.1 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6.1的要求。
b) 收集生成式人工智能生成内容等数据时，应记录所使用生成式人工智能模型或服务的版本、获取时间等信息。

7.2 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6.2的要求。
b) 生成式人工智能生成内容构成的数据样本，应添加所使用生成式人工智能模型或服务的版本、获取时间等元数据内容。
c) 优化训练数据的数据标注活动应符合GB/T AAAAA的安全要求。
d) 来源于生成式人工智能的生成数据，应重点识别数据内容是否存在安全风险内容并记录识别情况。

7.3 数据使用

对服务提供者的要求如下。

a) 优化训练所使用数据的数据来源应符合6.3的要求。
b) 使用生成式人工智能生成内容等数据时，应过滤掉存在安全风险内容的数据。

8 评价方法

8.1 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) 检查服务提供者对预训练和优化训练数据的操作过程记录和管理文档；
    2) 检查服务提供者预训练和优化训练数据所在系统和网络的设计文档、运行日志，检查相关设备的实际运行情况；
    3) 检查服务提供者预训练和优化训练数据的安全防护技术措施；
    4) 检查服务提供者是否具备应急响应小组，以及是否制定了针对预训练和优化训练数据安全事件的应急响应预案，检查安全事件的应急处置记录；
    5) 检查服务提供者是否具有记录预训练和优化训练数据的数据收集及准备阶段处理活动的日志的完整性、有效性。
b) 预期结果：
    1) 服务提供者对预训练和优化训练数据进行了分类分级操作和管理；
    2) 服务提供者已采取技术措施对预训练和优化训练数据进行安全监测，发现数据安全缺陷、漏洞等风险时及时告警并采取了相应的处置措施；
    3) 服务提供者已采取身份鉴别、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；

4) 服务提供者已具备应急响应小组，建立了针对预训练和优化训练数据安全事件的应急响应机制，并在发生安全事件时，及时有效进行了实施；
5) 服务提供者具有预训练和优化训练数据的数据收集及准备阶段关键活动日志，基于日志可对关键操作进行审计和追溯。

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。

8.2 预训练数据处理活动评价方法

8.2.1 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) 检查服务提供者是否具有数据来源记录；核验数据来源记录格式的正确性；抽样服务提供者所收集的数据，核验数据来源记录的完整性；
2) 检查服务提供者数据来源记录中同类型数据所使用数据来源的数量；
3) 检查服务提供者是否从互联网网站收集数据；抽样服务提供者收集的互联网网站数据，核查抽样样本与所记录统一资源定位符的一致性；
4) 检查服务提供者是否通过交易或合作等方式从其他组织或个人收集数据；抽样检查服务提供者对交易方或合作方提供数据、承诺、材料的审核材料。

b) 预期结果：
1) 服务提供者具有数据来源记录；数据来源涉及互联网网站的，具有统一资源定位符记录；数据来源涉及其他组织或个人的，具有数据集名称、来源组织记录，交易合同、合作协议记录、许可协议或相关授权文件等有效；数据来源涉及服务使用者的，具有服务名称、服务使用者的身份标识号码记录，服务使用者的授权记录有效；数据来源记录覆盖完整；
2) 服务提供者同种类型数据的具有多个来源数量；
3) 服务提供者未从互联网网站收集数据或所有抽样样本与所记录统一资源定位符相一致；
4) 服务提供者未通过交易或合作等方式从其他组织或个人收集数据，或具有对交易方或合作方提供数据、承诺、材料的审核材料。

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。

8.2.2 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) 随机抽样服务提供者预处理后的数据，对于每类数据来源抽样数量不少于100个样本，检查样本元数据内容的正确性；
2) 随机抽样服务提供者预处理后的数据，抽样数量不少于100个样本，检查样本是否具有安全风险内容识别记录；
3) 随机抽样服务提供者预处理后的数据，抽样数量不少于100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。

b) 预期结果：
1) 抽样样本全部具有元数据内容；样本涉及其他组织或个人数据集来源的，具有数据集名称、组织名称记录；样本涉及互联网网站来源的，具有样本或样本所在网页的统一资源定位符；样本涉及服务使用者来源的，具有服务名称及服务使用者的身份标识号码记录；
2) 抽样样本全部具有安全风险内容情况记录；
3) 抽样样本涉及知识产权侵权风险的，全部具有知识产权侵权风险记录。

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。

8.2.3 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) 检查服务提供者是否使用包含个人信息的数据；检查服务提供者是否具有个人同意记录，或是否符合法律、行政法规规定的情况；
   2) 检查服务提供者是否使用包含个人敏感信息的数据；检查服务提供者是否具有个人单独同意记录，或是否符合法律、行政法规规定的情况；
   3) 采用人工抽检方式从全部数据中随机抽取不少于4000个样本，核查服务提供者使用数据的知识产权侵权风险识别记录的准确性；
   4) 采用人工抽检方式从全部数据中随机抽取不少于4000个样本，采用关键词、分类模型等技术抽检法从全部数据中抽取不少于总量10%的数据。
b) 预期结果：
   1) 服务提供者未使用个人信息数据，或具有个人同意记录，或使用个人信息数据符合法律、行政法规规定的情形；
   2) 服务提供者未使用个人敏感信息数据，或具有个人单独同意记录，或使用个人敏感信息数据符合法律、行政法规规定的情形；
   3) 抽样样本不涉及知识产权侵权风险，或抽样样本无知识产权风险并与知识产权侵权风险识别记录一致；
   4) 人工抽检的抽样数据样本中不含安全风险内容样本数量占总抽样数量的比值不低于96%，技术抽检的抽样数据样本中不含安全风险内容样本数量占总抽样数量的比值不低于98%。
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。

8.3 优化训练数据处理活动评价方法

8.3.1 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8.2.1a)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) 检查服务提供者是否收集生成式人工智能的生成内容；检查服务提供者收集的生成式人工智能生成内容是否具有所使用生成式人工智能模型或服务的版本、获取时间等信息记录。
b) 预期结果：
   1) 符合8.2.1b)的预期结果；
   2) 服务提供者没有收集生成式人工智能的生成内容，或具有所使用生成式人工智能模型或服务的版本、获取时间等信息的记录。
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。

8.3.2 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8.2.2a)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) 随机抽样服务提供者预处理后的优化训练数据，抽样数量不少于100个样本，检查样本元数据内容的正确性；
3) 检查优化训练的标注数据是否符合GB/T AAAAA的安全要求；
4) 随机抽样服务提供者预处理后的优化训练数据，抽样数量不少于100个样本，检查样本是否具有生成式人工智能生成数据的安全风险内容识别情况。

b) 预期结果：
1) 符合8.2.2b)的预期结果；
2) 抽样样本涉及生成式人工智能生成内容的，样本的元数据内容包括生成式人工智能模型或服务的版本、获取时间等信息。
3) 优化训练标注数据符合GB/T AAAAA的安全要求；
4) 抽样样本为生成式人工智能生成内容的，具有生成式人工智能生成数据的安全风险内容识别记录。

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。

8.3.3 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8.2.3a)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) 采用人工抽检方式从全部数据中随机抽取不少于4000个样本，核查服务提供者使用数据的生成式人工智能生成数据的安全风险内容识别记录的准确性；

b) 预期结果：
1) 符合8.2.3b)的预期结果；
2) 抽样样本不涉生成式人工智能生成内容，或抽样样本安全风险内容并与知识生成式人工智能生成数据的安全风险内容识别记录一致；

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。

附录 A
（资料性）
预训练和优化训练数据的主要安全风险内容

A.1 包含违反社会主义核心价值观的内容

包含以下内容：

a）煽动颠覆国家政权、推翻社会主义制度；
b）危害国家安全和利益、损害国家形象；
c）煽动分裂国家、破坏国家统一和社会稳定；
d）宣扬恐怖主义、极端主义；
e）宣扬民族仇恨；
f）宣扬暴力、淫秽色情；
g）传播虚假有害信息；
h）其他法律、行政法规禁止的内容。

A.2 包含歧视性内容

包含以下内容：

a）民族歧视内容；
b）信仰歧视内容；
c）国别歧视内容；
d）地域歧视内容；
e）性别歧视内容；
f）年龄歧视内容；
g）职业歧视内容；
h）健康歧视内容；
i）其他方面歧视内容。

A.3 商业违法违规

主要风险包括：

a）侵犯他人知识产权；
b）违反商业道德；
c）泄露他人商业秘密；
d）利用算法、数据、平台等优势，实施垄断和不正当竞争行为；
e）其他商业违法违规行为。

A.4 侵犯他人合法权益

主要风险包括：

a）危害他人身心健康；
b）侵害他人肖像权；
c）侵害他人名誉权；
d）侵害他人荣誉权；
e）侵害他人隐私权；

f）侵害他人个人信息权益；
g）侵犯他人其他合法权益。

附录 B
（规范性）
关键词库和分类模型要求

B.1 关键词库

要求如下。

a）关键词库应具有全面性，总规模不宜少于10000个。

b）关键词库应具有代表性，应至少覆盖本文件附录A.1以及A.2中17种安全风险内容，附录A.1中每一种安全风险内容的关键词均不宜少于200个，附录A.2中每一种安全风险内容的关键词均不宜少于100个。

c）关键词库应按照网络安全实际需要及时更新，每周宜至少更新一次。

B.2 分类模型

分类模型用应完整覆盖本文件附录A中全部29种安全风险。

参 考 文 献

[1] TC260-PG-20233A 网络安全标准实践指南—生成式人工智能服务内容标识方法

[2] TC260-003 生成式人工智能服务安全基本要求