ICS 35.020
CCS L 80

DB21

辽　宁　省　地　方　标　准

DB21/TXXXX—XXXX

生成式人工智能安全要求

（征求意见稿）

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

辽宁省市场监督管理局　发布

目　　次

前　言

本文件按照GB/T1.1-2020《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。

本文件由辽宁省工业和信息化厅提出并归口。

本文件起草单位：沈阳华睿博信息技术有限公司等。

本文件主要起草人：邵华等。

本文件发布实施后，任何单位和个人如有问题和意见建议，均可以通过来电和来函等方式进行反馈，我们将及时答复并认真处理，根据实际情况依法进行评估及复审。

归口管理部门通信地址：辽宁省沈阳市皇姑区北陵大街45-2号。

归口管理部门联系电话：024-86913384。

标准起草单位通讯地址：辽宁省沈阳市和平区青年大街386号华阳国际大厦2396。

标准起草单位联系电话：18698849086。

生成式人工智能安全要求

1 范围

本文件规定了生成式人工智能的数据安全要求、模型安全要求、安全措施要求以及安全评估要求等方面的要求。

本文件适用于开展评估人工智能服务的安全水平等工作。

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 35274-2023 数据安全技术 大数据服务安全能力要求

GB/T 41479-2022 信息安全技术 网络数据处理安全要求

GB/T 42755-2023 人工智能 面向机器学习的数据标注规程

3 术语和定义

3.1

生成式人工智能 generative artificial intelligence

基于数据、算法、模型、规则，能够根据使用者提示生成文本、图片、音频、视频等内容的人工智能。

3.2

模型训练 model training

使用特定的数据集对模型进行学习的过程。

3.3

数据标注 data annotation

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。

4 基本要求

生成式人工智能安全的基本要求至少应包括：

a） 保密性。采用技术和管理手段保证数据的保密性，并记录数据处理日志及确保日志的保密性；

b） 完整性。采用完整性校验的技术和管理手段，并对人工智能数据的完整性进行监测；

c） 隐私性。采用技术和流程以保护个人信息主体的个人信息，涉及个人信息的符合相关要求；

d） 合规性。数据处理活动符合法规、政策文件、标准规范相关要求；

e） 分类分级：对生成式人工智能数据进行分类和分级，并对不同类别不同级别的数据建立相应的全流程数据安全保护措施；

f） 伦理安全：生成式人工智能数据全生命周期保障伦理安全，促进公平、公正、和谐，避免偏见、歧视、隐私和信息泄露等问题。

5 数据安全

5.1 数据采集与使用安全

生成式人工智能数据采集与使用安全的要求至少应包括：

a） 数据采集安全。在数据采集过程中，需要建立数据采集安全合规管理规范，明确数据采集的目的、用途、方式、范围等，并进行风险评估；
b） 数据存储安全。数据存储阶段需要保护数据免受未经授权访问、数据泄露、篡改或丢失的风险。采取数据加密、备份和恢复、访问控制等措施；
c） 数据传输安全。在数据传输过程中，确保数据的机密性、完整性和不可篡改性。使用加密传输、流量监控和完整性校验等技术手段；
d） 数据使用安全。在数据被访问和利用的过程中，保护数据不被滥用、泄露或未经授权访问，同时确保数据的合规性和合法使用。实施隐私保护、访问控制、数据加密等措施；
e） 数据删除安全。在数据被销毁或删除阶段，确保数据彻底删除并无法恢复，以保护数据不被非法恢复或滥用。使用数据擦除技术、删除确认和审计日志等措施；
f） 数据安全事件管理。定义安全事件、选择负责处理事件的人员、进行安全审核、制定沟通计划以及创建数据恢复计划等，以确保数据安全事件的有效管理和快速恢复。

5.2 数据内容安全

数据内容安全要求至少应包括：
a） 训练数据经过过滤，去除违法不良信息，并确保数据内容的准确性和可靠性；
b） 对数据中的知识产权侵权风险进行识别和处理，确保使用合法的数据进行训练；
c） 涉及个人信息的数据使用需获得同意或符合法律规定，确保个人隐私安全。

5.3 数据标注安全

5.3.1 标注人员

参与数据标注的人员应具备相应的资质，对确定符合要求的人员培训，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。培训合格者,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。

5.3.2 明确职责

应规定参与人工智能数据标注的所有角色的职能，并做到：
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，能够根据业务实际需要确定承担数据管理工作的部门或人员；
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、数据标注质量控制人员以及与人工智能数据标注相关的其他角色的职责。

5.3.3 合法合规

应出台规章制度保障人工智能数据标注任务的合法合规性，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。

5.3.4 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。

5.3.5 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，并做到：
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。

5.3.6 数据安全

应从以下几方面做好数据保密，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的安全性要求；

b） 数据存储，建立访问控制制度和加密机制确保人工智能数据存储保密性要求；

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。

5.3.7 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，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，保证标注活动数据完整性；

c） 数据标注可靠性，确保标注过程只执行批准的范围；

d） 数据储存完整性，确保数据标注存储及备份完整性。

5.3.8 数据可审计

应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可审计，并做到：

a） 对数据标注过程信息记录，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。

6 模型安全

6.1 语料安全

语料安全要求是确保人工智能模型训练和应用过程中数据合法性、合规性的重要环节，其要求至少应包括：

a） 使用合法来源的语料，并对语料内容质量提出量化标准；

b） 建立知识产权管理策略、识别知识产权侵权风险、完善投诉举报渠道、公开摘要信息等；

c） 建立语料来源黑名单，不使用黑名单来源的数据进行训练；

d） 对各来源语料进行安全评估，若单一来源语料中含违法不良信息超过5%，则将该来源加入黑名单；

e） 服务提供者确保个人信息处理行为具有合法性基础，即取得对应个人信息主体的同意或符合法律、行政法规规定的其他情形。

6.2 模型训练安全

模型训练安全涉及模型训练环境安全、模型训练过程安全以及其他安全等，具体要求至少应包括：

a） 确保训练所用的硬件设备安全可靠，避免设备故障或损坏导致的数据丢失或模型损坏。同时，对硬件设备进行定期维护和保养，确保其正常运行；

b） 使用正版软件，避免使用盗版软件导致的安全问题，对软件进行定期更新和升级，以修复可能存在的安全漏洞；

c） 确保训练环境的网络安全性，避免网络攻击或数据泄露，使用安全的网络连接方式，如加密传输等；

d） 在训练过程中，使用监控设备或软件，随时掌握训练进度和模型性能，对异常情况进行及时报警和处理，确保训练过程的顺利进行；

e） 模型训练过程中，对用户数据的严格保护，防止数据泄露和滥用；

f） 在模型更新或升级前，进行充分的安全评估，确保更新或升级后的模型安全性；

g） 制定安全管理策略，对模型进行定期的安全审计和漏洞修复。

6.3 模型生成内容安全

模型生成内容安全要求至少应包括：

a） 在训练过程中，将生成内容安全性作为评价生成结果优劣的主要考虑指标之一；

b） 在每次对话中，对使用者输入信息进行安全性检测，引导模型生成积极正向内容；

c） 建立常态化监测测评手段，对监测测评发现的提供服务过程中的安全问题，及时处置并通过针对性的指令微调、强化学习等方式优化模型；

d） 采取技术措施提高生成内容响使用者输入意图的能力，提高生成内容中数据及表述与科学常识及主流认知的符合程度，减少其中的错误内容；

e） 采取技术措施提高生成内容格式框架的合理性以及有效内容的含量，提高生成内容对使用者的帮助作用。

7 安全评估要求

对生成式人工智能的安全进行评估，具体要求至少应包括：

a） 生成式人工智能服务提供者应在服务上线前以及重大变更时开展安全评估，可以自行开展或委托第三方评估机构进行；

b） 语料安全评估。采用人工抽检方式，从全部语料中随机抽取不少于4000条语料，合格率不应低于96%；

c） 生成内容评估。采用人工抽检方式，从生成内容测试题库中随机抽取不少于1000条测试题，模型生成内容的抽样合格率不应低于90%；

d） 问题拒答评估。从应拒答测试题库中随机抽取不少于300条测试题，模型的拒答率不应低于95%；从非拒答测试题库中随机抽取不少于300条测试题，模型的拒答率不应高于5%；

e） 规范模型更新升级过程中的安全评估举措，需建立安全管理策略，在关键更新升级后重新进行内部安全评估；

f） 安全评估过程必须包括语料安全、模型安全、安全措施和基础设施相关条款，覆盖违反社会主义核心价值观、歧视性内容、商业违法违规、侵犯他人合法权益、无法满足特定服务类型等主要安全风险；

g） 体现坚守生成内容安全底线的基本原则，编制过程中组织一线工作者对各类评估内容指标进行细化和反复论证；

h） 对于具体的评估内容，对所需的关键词库、生成内容测试题库、拒答测试题库和分类模型等安全评估基础设施提出具体清晰的建设规范，形成了清晰、具体、可操作的安全评估标准；

i） 评估结论应明确为“符合”、“不符合”或“不适用”。对于结论为“符合”的，提供充分的证明材料；结论为“不符合”的，说明原因，并提供替代措施的有效性证明。

8 安全措施要求

8.1 用户信息保护

生成式人工智能的安全措施要求涵盖了用户信息保护，具体要求至少应包括：

a） 收集与使用。仅在提供生成式人工智能服务所必需的范围内收集个人信息，并明确告知用户信息的收集、使用目的和范围，不得非法留存能够推断出用户身份的输入信息和使用记录，避免用户隐私泄露；

b） 数据安全。采取加密技术和其他必要的安全措施，确保用户信息在存储和传输过程中的安全性，定期对存储的用户信息进行安全检查和备份，以防止数据丢失或损坏；

c） 用户权利。允许用户随时查询、更正、删除自己的个人信息。在收到用户关于个人信息处理的请求时，应在规定时间内及时响应并处理。

8.2 服务透明度与可控性

生成式人工智能的安全措施要求涵盖了服务透明度与可控性，具体要求至少应包括：

a） 明确并公开服务的适用人群、场合、用途等信息，以便用户了解并合理使用服务；

b） 在交互界面或说明文档中公开服务的局限性、所使用的模型、算法等方面的概要信息；

c） 提供用户控制选项，允许用户根据需要调整服务的使用方式，如设置隐私权限、选择服务内容等；

d） 允许用户随时停止使用服务，并保障用户在停止服务后的个人信息和数据的安全。

8.3 投诉举报与应急响应

生成式人工智能的安全措施要求涵盖了投诉举报与应急响应，具体要求至少应包括：

a） 建立用户投诉接收处理机制，设置便捷的投诉、举报入口，公布处理流程和反馈时限，及时受理、处理公众投诉举报并反馈处理结果，确保用户权益得到保障；

b） 制定应急预案，对可能发生的安全事件进行预防和应对，在发生安全事件时，及时采取措施消除安全隐患，并向相关主管部门报告；

c） 定期对服务进行安全评估和合规性检查，接受政府和相关主管部门的监管和指导，确保服务的安全性和合规性。