ICS 03.100.01
CCS A 00

团 体 标 准

T/CECC 027-2024

生成式人工智能数据应用合规指南

Compliance Guidelines for Data Application of Generative Artificial Intelligence

2024-4-16 发布 2024-5-1 实施

中国电子商会 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。

本文件由国家工业信息安全发展研究中心牵头，由北京之合网络科技有限公司负责组织，由中国电子商会归口。

本文件起草单位：国家工业信息安全发展研究中心、天翼云科技有限公司、海尔集团公司法律事务部、蔚来控股有限公司、解放号网络科技有限公司、三七互娱网络科技集团股份有限公司、北京市盈科律师事务所、北京市康达律师事务所、联想（北京）有限公司、广州华多网络科技有限公司、中移数智科技有限公司、九度数字科技（苏州）有限公司、上海邦信阳律师事务所、上海拉扎斯信息科技有限公司、上海秘塔网络科技有限公司、上海健交科技服务有限责任公司、上海宽娱数码科技有限公司、上海得帆信息技术有限公司、上海商汤科技开发有限公司、上海澄明则正律师事务所、上海中联律师事务所、上海之合网络科技有限公司、上海之爱智能科技有限公司、上海律行教育科技有限公司、上海爱奇求思教育科技有限公司、天津东方律师事务所、天津律云律师事务所、日日顺新能源科技有限公司、中国汽车工程研究院股份有限公司、中国科学院计算机网络信息中心、中电金信软件有限公司、中电信数智科技有限公司、中译语通科技股份有限公司、中国电子商会数据要素发展工作委员会、中国中小企业协会企业合规专业委员会、中国电子商会人工智能委员会、中国科学技术法学会人工智能法专业委员会、中瑞世联资产评估集团有限公司、平安科技（深圳）有限公司、北京大学武汉人工智能研究院、北京世宁律师事务所、北京市中伦文德律师事务所、北京市铭基律师事务所、北京远景视点科技有限公司、北京信工博特智能科技有限公司、北京桓润律师事务所、北京之合网络科技有限公司、北京泮顿子敬信息技术有限公司、丝芙兰Sephora、西安电子科技大学、江西火眼智能科技有限公司、江西电信信息产业有限公司、江苏品川律师事务所、江苏智伦数字技术研究有限公司、江苏数智碳链科技有限公司、阿里巴巴（北京）软件服务有限公司、武汉光谷知识产权研究院有限公司、青岛海尔生物医疗股份有限公司、英矽智能科技（上海）有限公司、杭州小影创新科技股份有限公司、枣庄市网络社会组织联合会、金杜律师事务所、郑州郑大信息技术有限公司、陕西丰瑞律师事务所、蚂蚁科技集团股份有限公司、香港浸会大学深圳研究院、美年大健康产业控股股份有限公司、浙江天册律师事务所、浙江深服人工智能科技有限公司、清图数据科技（南京）有限公司、深圳市木愚科技有限公司、厦门立马耀网络科技有限公司、新汽有限公司、睿珀智能科技有限公司。

本文件主要起草人：张穹、张平、方懿、邱惠君、李卫、刘巍、杨柳、冯立鹦、陈立彤、李丹一、朱倩倩、洪绍泉、洪祖运、洪钧、刘启铭、蔡江天、常国珍、陈府申、陈晗、陈华平、陈焕、陈杰、陈良斌、陈梦园、陈乾、陈向娟、陈怡、陈宇峰、戴学良、戴亦斌、邓超麟、邓志福、邓梓珊、刁成路、丁丁、丁亮、董格玛、董皓、董潇、杜娟、杜歆、杜雨、冯超、冯斐斐、冯祥宸、傅临黎、高辉、葛昌金、宫蕾、龚琳、郭嘉琦、郭蛟、郭璐璐、韩剑、韩琳、韩笑、郝成金、何念寒、何源泉、何媛、何昭敏、侯广、侯小菊、胡峰、胡家昊、胡俊勇、胡若玫、胡校溟、胡岩、黄凯、黄元忠、纪海良、贾颖、江海、姜婷、姜欣、蒋薇、蒋潇君、晋银涛、孔真琦、郎婷、李华、李辉、李健青、李珂、李林育、李谦、李嵘辉、李新华、李阳、李音瑶、李永锋、李悦、李泽芳、李长青、李哲、连艳、梁智刚、廖怀学、林安雯、刘朝、刘诚诚、刘丰、刘格言、刘骥、刘剑锋、刘敬霞、刘鹏、刘泊辰、刘瑞、刘欣、刘兴、刘艳阳、刘永和、龙怀春、卢丁、陆瑾、陆雨辰、罗洁、吕仁平、吕亚妹、马海曼、马晓艳、孟戈弋、聂正军、欧阳昆泼、潘永建、潘云龙、彭天基、彭晓燕、齐斌、祁彦谕、乔佳平、谯青青、邱梦赟、邱媛春、曲峰、屈文静、任洁、阮芳洋、沙俊、沙沫、单思杨、时萧楠、宋冰心、宋皓、宋俊、宋天一、孙晨荻、孙雪菲、谭洁、汤子欧、唐简捷、唐淑萍、陶毓、田莉、田茂君、朱炤沁、汪汉鸿、王斌、王彩琴、王崇、王芳、王菲、王斐、王涵、王皓、王剑锋、王捷、王菁煜、王君、王立群、王丽娜、王龙海、王淼、王祺、王瑞揆、王溪、王小敏、王轩、王艺蓉、王岳、王悦、王志林、王智滢、韦征、魏丰、翁振洋、吴刚、吴万凯、吴志强、武婕、夏海波、夏庆仁、夏文华、肖飒、谢国辉、谢尚誓、谢甜甜、熊钱富、徐婧、徐岚、徐强、徐瑞、徐云飞、许力先、许立昕、闫洋、燕雪松、杨博、杨海滨、杨海强、杨瑾煜舟、杨劲、杨军、杨思敏、杨天歌、杨晓雷、杨晓莉、杨鑫、杨旸、杨宇宙、杨忠勤、姚晶鑫、

姚雪飞、叶娟、叶俊希、尹立、于洪方、于谨源、余俊峰、俞霞、袁韶浦、袁新忠、曾玥、张翠美、张杜超、张广运、张豪、张继焕、张建民、张静、张隽、张凌、张明强、张汭、张森森、张松艳、张彤、张显显、张笑怡、张鑫、张雪芳、张延来、张逸瑞、张源、张岳、张云昊、张孜铭、张祖勤、赵梦晗、赵玉刚、赵云虎、郑珂威、郑鑫焱、钟云斌、周力思、周霖、周阳、周宇、朱彬、朱莎、朱晓薇、朱岳峰、朱政、朱中辉、邹丹莉、林戈、张怡、赵琪彦、陈骁萌、翟艺、毛姗姗、吴剑霞、金辰、聂佳彤、陈绮敏、张敏、朱彩云、董宇洲、邓欢、王怡冉、龙衍孙、王笑晗、张圆捷、陈天宇。

引　言

为应对生成式人工智能带来的安全挑战，促进生成式人工智能产业高质量健康发展，确保数据应用的各个环节符合合规性要求，确保数据全生命周期的安全运行，根据《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、《中华人民共和国著作权法》、《中华人民共和国反不正当竞争法》等相关法律，《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》、《生成式人工智能服务管理暂行办法》、《科技伦理审查办法（试行）》、《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》等相关部门规章，结合我国生成式人工智能技术和产业发展的实际，制定本文件。

生成式人工智能数据应用合规指南

1 范围

本文件规定了生成式人工智能服务在数据采集、数据标注、训练数据预处理、模型训练与测试、内容生成服务等各个数据应用环节中应遵循的数据应用合规原则与合规要求，以及可供借鉴参考的具体合规手段与合规方法。

本文件适用于指导生成式人工智能服务提供者向中华人民共和国境内公众提供生成式人工智能内容生成服务过程中所开展的数据应用合规工作。

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 29490-2023 企业知识产权合规管理体系 要求
GB/T 35273-2020 信息安全技术 个人信息安全规范
GB/T 35770-2022 合规管理体系 要求及使用指南
GB/T 41867-2022 信息技术 人工智能 术语
GB/T 42574-2023 信息安全技术 个人信息处理中告知和同意的实施指南
GB/T 42755-2023 人工智能 面向机器学习的数据标注规程
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
TC260-003 生成式人工智能服务安全基本要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能 artificial intelligence；AI

〈学科〉人工智能系统（3.2）相关机制和应用的研究和开发。
［来源：GB/T 41867-2022，定义 3.1.2］

3.2

人工智能系统 artificial intelligence system

指针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统。
［来源：GB/T 41867-2022，定义 3.1.8］

3.3

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模型及相关技术。

3.4

提供者 provider

以交互界面、可编程接口等形式面向我国境内公众提供生成式人工智能服务的组织或个人。

3.5

个人信息 personal information

以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息。

［来源：GB/T 42574-2023，定义 3.1］

3.6

敏感个人信息 sensitive personal information

敏感个人信息是一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息。

注：敏感个人信息包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息，以及不满14周岁未成年人的个人信息。

［来源：GB/T 42574-2023，定义 3.2］

3.7

数据标注 data labelling

给数据样本指定目标变量和赋值的过程。

［来源：GB/T 41867-2022，定义 3.2.29］

3.8

训练数据 training data

用于训练机器学习模型的输入数据子集。

［来源：GB/T 41867-2022，定义 3.2.34］

3.9

合成数据 synthetic data

基于计算机模拟、使用人工智能模型或基于机器学习方法、深度学习方法等技术生成的模仿现实世界观察的虚拟数据。

3.10

验证数据 validation data

用于评估单个或多个候选机器学习模型性能的数据样本。

［来源：GB/T 41867-2022，定义 3.2.35］

注：验证数据与测试数据是不重复的，通常也与训练数据不重复。但是，在没有足够的数据进行三种方式的训练，验证和测试集拆分的情况下，数据只被分为两个集——一个测试集和一个训练或验证集。

3.11

测试数据 test data

用于评估最终机器学习模型性能的数据。

［来源：GB/T 41867-2022，定义 3.2.3］

注：测试数据与训练数据无交集。

3.12

模型训练 model training

利用训练数据，基于机器学习算法，确定或改进机器学习模型参数的过程。

［来源：GB/T 41867-2022，定义 3.2.18］

3.13

模型优化 model optimization

提升模型执行速度，泛化能力，或改善利益相关方所关心的其他特性的方法。
［来源：GB/T 41867-2022，定义 3.2.19］

3.14

伦理 ethic

开展人工智能技术基础研究和应用实践时遵循的道德规范或准则。
［来源：GB/T 41867-2022，定义 3.4.8］

3.15

幻觉 hallucination

指生成式人工智能模型输出的内容基本符合逻辑与语法，但却与客观事实不符甚至完全虚构，或者不具备可理解性的现象。

3.16

偏见 bias

〈人工智能可信赖〉对待特定对象、人员或群体时，相较于其他实体出现系统性差别的特性。
［来源：GB/T 41867-2022，定义 3.4.10］
注：对待指任何一种行动，包括感知、观察、表征、预测或决定。

3.17

数据投毒 data poisoning

在训练数据中故意添加虚假、有害或恶意数据，以干扰模型训练、影响模型的输出结果。

4 合规原则

生成式人工智能数据应用应符合以下合规原则：

a) 科技伦理原则：在生成式人工智能数据应用的各个环节中，需注意遵循增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险、保持公开透明的科技伦理原则；

b) 内容安全原则：在利用生成式人工智能技术进行内容生成时，应采取有效措施避免生成违背社会主义核心价值观的内容，避免生成具有歧视性的内容，避免生成虚假有害信息等法律、行政法规禁止的内容；

c) 人格保护原则：在生成式人工智能数据应用的各个环节中，应注重保护自然人的人格利益，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益等；

d) 商业利益原则：在模型开发、服务提供等数据应用环节中，提供者应尊重他人的知识产权、数据权益等，避免实施垄断、不正当竞争等侵犯其他商业主体合法权利的行为；

e) 技术发展原则：提供者在服务提供过程中应注意及时收集反馈，提高生成内容的准确度与可靠性，不断促进人工智能技术的优化与发展；

f) 体系合规原则：提供者应搭建完善的合规管理体系，就生成式人工智能数据应用的各个环节，制定合规管理制度，采用有效的技术方法和其他治理措施，实现数据应用合规管理目标。

5 数据采集合规要求

5.1 合规性审查

对用于模型训练的数据，提供者应根据获取数据的不同方式以及数据自身的不同类别，建立数据来源和内容合法性的审查机制。

用于模型训练的数据如果同时符合5.2与5.3中的多种情形时，应同时满足各情形的收集合规要求。

5.2 采集方式

5.2.1 直接获取数据

提供者可直接从个人信息主体处获取个人信息，或在自身日常生产经营中创造生产新数据、以原始数据为基础加工生产新数据。

提供者直接从个人信息主体处获取个人信息的，应符合 5.3.2 的合规要求。

5.2.2 间接获取数据

在事先评估合法性基础的前提下，除直接获取数据外，提供者可从其他主体处间接获取数据，即通过数据交易、数据共享、公共数据授权运营等途径获取数据。

提供者应同相对方签订相应的法律协议，谨慎审核相对方的数据来源合法性以及数据可交易性，并要求相对方作出来源合法性、可交易性和可使用性承诺，或出示相关证明等。

鼓励提供者通过数据交易所等公开平台获取数据，以提升数据来源的合法合规性。

5.3 特定数据

5.3.1 公开数据获取

提供者可通过人工采集或自动爬取等手段从公共互联网获取公开数据，但应注意获取手段的合法合规，不得侵犯他人合法权益。

采用自动爬取方式的，应遵守目标网站的网络爬虫排除协议（Robots协议）等声明文件要求，避免采用破解密码、伪造用户代理（User Agent）、设置代理网际协议地址（IP地址）等技术手段进行违规爬取。应控制数据爬取的流量与频率，避免因爬取行为影响目标网站的正常运行。爬取移动互联网应用程序（App）、小程序等所依赖的网络服务应用程序接口（API）中的数据，应当遵守API的服务鉴权声明。

公开数据附有数据使用许可条件或使用限制的，提供者获取该公开数据后，应遵守相关约定。

5.3.2 个人信息收集

如提供者采集的数据类型中包含个人信息，应符合相应的法律法规和GB/T 35273-2020第5章中有关个人信息收集的合规要求，包括但不限于：

a) 在直接收集个人信息前，应依法向个人明确告知个人信息处理者的名称或者姓名和联系方式，个人信息的处理目的、处理方式，处理的个人信息种类、保存期限，个人行使法定权利的方式和程序等；
b) 如将直接获取的个人信息用于模型训练等目的，应符合 GB/T 42574-2023 第 7-9 章的规定告知并取得个人同意，或者具备其他合法性基础；
c) 对于个人自行公开或者其他已经合法公开的个人信息，如个人未明确拒绝用于模型训练等目的，处理行为未显著违背个人公开目的且相关处理不会对个人权益造成重大影响的，可视为在合理范围内进行处理；
d) 如需采集敏感个人信息用于模型训练的，应事前进行个人信息保护影响评估，在采取严格保护措施并取得个人单独同意的前提下方可使用；
e) 如处理不满十四周岁未成年人个人信息，除上款内容外，还需取得未成年人父母或其他监护人的同意，并制定专门的个人信息处理规则；
f) 间接获取的数据如包含个人信息的，应要求个人信息提供方说明个人信息来源，并确保就信息共享已履行法定的告知义务并取得个人单独同意，或者具备其他的合法性基础；
g) 根据模型训练的特定目的，遵循个人信息处理的必要性原则，在限于实现处理目的的最小范围内收集和处理个人信息；
h) 除非确有必要，否则用于模型训练的个人信息应进行去标识化处理后再进行使用。

5.3.3 知识产权保护

获取数据用于模型训练的，应采取以下手段防止对他人知识产权的侵害：

a) 对于已超过著作权保护期限进入公有领域的作品，提供者可以采集相关数据投入模型训练，但应避免在生成内容中侵犯著作权人的署名权、修改权与保护作品完整权等著作人身权；
b) 对仍在著作权保护期限内的作品，提供者应主动采取措施获取著作权人的授权，明确其作品可用于生成式人工智能的模型训练；
c) 建议提供者通过著作权集体管理组织获取著作权人的授权；
d) 对于商标权、专利权、商业秘密等其他类型的知识产权，建议提供者根据数据类型和数据来源进行必要甄别，如发现有侵权可能的，应避免采集或取得权利人的授权；
e) 提供者可依据 GB/T 29490-2023 第 4-10 章建立企业知识产权合规管理体系，并依据其附录 B 所列的“专利、商标、著作权、商业秘密典型禁止性行为”进行风险排查。

6 数据标注合规要求

6.1 标注规则的制定

为模型训练的目的需要进行数据标注的，应按法律法规以及数据需求方的要求，依据以下规定制定标注规则：

a) 标注规则应根据数据需求方对模型训练的具体要求制定；
b) 标注规则应清晰、具体、全面、细化，对标注人员具有实际操作性；
c) 标注规则的确定应有利于提高训练数据的准确性，标注过程中如发现冗余数据、错误数据、异常数据等情况应进行及时处理；
d) 标注规则的确定应有利于保持训练数据的客观性，避免因规则设计的主观性导致标注结果发生同客观情况的偏离；
e) 标注规则应进行定期审查和更新，以适应新的法律法规、技术发展和业务需求的变化。

6.2 数据标注质量评估

数据标注的全流程实施过程中应包含质量评估的环节，具体操作可依据GB/T 42755-2023第6.2和第7.1条规定的流程与方法进行实践。

质量评估可采用抽样核验、机器验证、第三方验证等方式进行，根据场景需求及项目特点，建议选择两种以上方式进行数据标注准确度和一致性检查，并根据检查结果及时进行反馈校正。

6.3 标注人员的资质、培训、考核及管理

数据标注应对标注人员提出有针对性的资质要求，并进行以下培训、考核与管理：

a) 根据标注任务和数据类型的区别，区分标注任务所需的专业知识和专业技能需求，区分图像标注、语音标注、文本标注、视频标注等不同数据标注的类型，对标注人员的基础资质提出要求；
b) 根据每次标注任务的具体要求，选择符合资质要求的人员参加岗前培训；
c) 培训应设置相应的考试环节，通过考试的人员才能进入标注任务的实操环节；
d) 除实操技术外，培训内容中必须含有必要的法律合规指南，标注人员应了解数据标注的原则与意义，未按照标注规则进行操作的责任与后果，以及在接触数据过程中所应履行的个人信息保护、数据安全合规与保密义务；
e) 就标注人员的实际工作表现建立相应的能力档案，以便在后续持续性工作中，对标注工作的人员进行筛选、抽样测试，持续提升标注工作质量。

7 训练数据的预处理合规要求

7.1 提高训练数据质量

提供者应采取有效措施提高训练数据质量，并从真实性、准确性、客观性、多样性、安全性等角度以防止训练数提升数据质量。当各方面要求不能同时满足或可能存在冲突时，提供者应进行谨慎考量，以防止训练数据的不当选择影响生成内容的质量。

7.1.1 训练数据的真实性

提供者应从数量和质量上判断所获取的数据是否具有可靠的来源，是否能够反映真实世界的情况，并通过人工或模型等方式就数据内容的真实性进行核验。

7.1.2 训练数据的准确性

提供者可采用数据去重、去除异常值、纠正错误等数据清洗方法，以提高数据集的准确性和一致性，排除噪声和偏差。

7.1.3 训练数据的客观性

训练数据宜尽可能中立和无偏见，在数据采集与后续处理环节中均应避免人为干扰、选择偏见和其他主观因素的介入。

7.1.4 训练数据的多样性

为提高模型的性能和泛化能力，应充分考虑数据来源、数据类型及样本特征分布的均衡和多样化。为防止生成存在偏见或歧视的内容，应进行充分多样化和具有代表性的数据选择，确保其包含各个民族、信仰、国别、地域、性别、年龄、职业和健康等的充分信息。

7.1.5 训练数据的安全性

为确保训练数据的安全性，应按照TC260-003中5.1列项的第一项规定对训练数据的来源进行安全评估和核验。

7.2 通过合成技术进行数据增强

提供者可在合理范围内创建并使用合成数据，按照以下原则进行数据增强训练：

a) 创建合成数据应当有真实、客观且达到一定数量的数据作为样本；
b) 合成数据原则上应保留真实数据的统计属性，但为提高数据集的多样性、补充罕见场景等目的而使用合成数据的可以不受此限；
c) 避免样本数据自身存在的偏见，该种偏见可能在生成合成数据时进一步传播；
d) 使用算法创建合成数据后，应进行适当的评估和验证，保证合成数据的质量与效用，避免出现偏差过大的问题；
e) 观测真实世界数据的变化，及时更新和维护合成数据，以保持其相关性和有效性；
f) 创建合成数据应符合相应的法律法规和伦理准则，不得创建可能造成侵权或有违伦理的数据，不得滥用合成数据；
g) 明确记录创建合成数据的具体算法与技术手段，做到可查询、可溯源。

8 模型训练与测试合规要求

8.1 模型训练

8.1.1 训练步骤

模型训练应至少包括预训练与优化训练等两重以上的训练环节。

8.1.2 预训练

预训练应选择具有合法来源的基础模型，基础模型应经过可靠性、安全性、合法性以及价值观等方面的测评，才可在此基础上进行后续训练。

8.1.3 优化训练

经过预训练后形成的算法模型，还应通过优化训练进一步使用已标注的数据进行后续流程，来优化模型训练的最终结果。

8.1.4 模型验证

在模型训练的不同环节中，均可使用验证数据对模型的参数与设置进行持续优化。验证数据可与训练数据来源于同样的数据集，但在训练过程中应保持相对独立。

8.2 模型测试

在正式为公众提供内容生成服务之前，为保证模型生成的效果，应按照以下要求进行模型测试：

a) 制定全面完整严格的测试指标体系，以减少幻觉、有害偏见和违法内容的生成；
b) 引入人工方式或其他模型进行对抗测试，根据结果反馈实现对模型性能的改进优化；
c) 建立动态调整的指标体系与测试方案，定期评估和调整指标体系，确保测试结果的有效性；
d) 测试数据的来源应独立于训练数据与验证数据，且应按照同样标准进行预处理；
e) 确保模型在经过严格测试并核验完成之后才对公众提供内容生成服务；
f) 模型评价依据、测试指标体系、测试与核验办法及采用的技术手段等，均应明确记录，做到可查询、可溯源。

9 内容生成服务合规要求

9.1 使用者尽责义务的告知

提供者应当与注册使用其服务的使用者（下称“使用者”）签订服务协议，在服务协议中明确告知使用者如下事项：

a) 生成式人工智能服务的基本特点与可能风险；
b) 使用者使用生成式人工智能服务的基本规范，包括不得利用生成式人工智能服务特性，有意识地获取违反法律法规、违反社会公德或伦理道德的内容；
c) 使用者负有审慎、尽责使用生成式人工智能服务的义务，在生成内容含有违反法律法规、违反社会公德或伦理道德的内容时，不应将此生成内容对外传播；
d) 明确告知使用者与生成内容相关的具体使用场景，例如明确生成内容是否可使用于科研、商用或自用等目的，以及其他使用限制条件；
e) 对于生成内容在特定行业的应用，尤其是对内容准确性有较高要求的如法律、医疗等领域，应向使用者重点提示风险。

9.2 生成内容的审核

提供者应建立生成内容审核机制，通过技术手段或人工审核的方式，对生成式人工智能生成的内容在对外提供前进行检测，识别并过滤其中的个人隐私信息、虚假有害信息、违法违规信息等不宜对外提供的内容。

9.3 生成内容的标识

提供者利用生成式人工智能技术向使用者提供文本、图片、音频、视频等生成内容时，需依据TC260-PG-20233A第3章的规定，通过水印等方式对生成内容进行明确标识，标识信息至少应包含“由人工智能生成”或“由AI生成”等含义。在由自然人提供服务转为由人工智能提供服务容易引起混淆时，应通过提示文字或提示语音的方式进行标识。

9.4 生成内容的异议审查机制

应建立使用者对生成内容提出异议的通知-受理机制、举报-受理机制，当使用者或举报者对生成内容合法合规性有异议，向提供者通知、举报时，提供者应按如下机制来处理：

a) 及时向使用者或举报者反馈，告知其已进入生成内容异议审核阶段；
b) 及时判断被异议的生成内容是否违反法律法规、违反社会公德或伦理道德；
c) 一旦确认被异议的生成内容违反法律法规、违反社会公德或伦理道德的，应及时采取停止生成、停止传输、消除等处置措施，并采取模型优化训练等措施进行整改；
d) 向使用者或举报者告知生成内容的异议处理情况，并视具体情况向有关主管部门报告。

9.5 使用者信息保护

提供者对使用者的个人信息、输入信息和使用记录应依法履行如下保护义务：

a) 根据必要性原则，仅收集与提供服务目的直接相关的个人信息；
b) 不得非法留存能够识别使用者身份的输入信息和使用记录；
c) 不得非法向他人提供使用者的输入信息和使用记录，除非获得使用者同意，或具有其他合法性基础；
d) 未进行明确告知并取得使用者同意的，提供者不得擅自将使用者的输入信息用于后续模型训练，除非具备其他合法性基础。

9.6 被侵权人维权支持

为应对因使用者不当使用人工智能生成内容造成他人权益损害的问题，提供者应建立被侵权人维权支持机制。在确认侵权事实属实的前提下，就被侵权人在法律框架内维护其合法权益提供合理配合，并采取必要措施防止侵害结果的扩大。

10 其他数据应用合规要求

10.1 数据安全保护

为防止数据泄露、数据投毒等安全事件的发生，提供者应按照法律法规的规定，履行如下义务：
a) 提供者应当全面加强数据处理系统、数据传输网络、数据存储环境等的安全防护，制定并落实全流程数据安全管理制度，并建议参照 GB/T 35770-2022 第 4-10 章构建相应的合规管理体系；
b) 按照网络安全等级保护的要求，提供者处理重要数据或处理一百万人以上个人信息的系统原则上应当满足三级以上网络安全等级保护要求；
c) 提供者应建立训练数据安全保障措施，宜采用经第三方认证的技术方法及软硬件工具，通过来源验证、数据清洗、访问控制、定期检查等多元机制，确保训练数据免受投毒威胁，防止模型生成内容发生偏差；
d) 提供者应当建立数据安全应急处置机制，发生数据安全事件时及时启动应急响应，防止危害扩大，消除安全隐患；
e) 数据安全事件对个人、组织可能造成危害的，建议提供者在三个工作日内将安全事件的相关情况通知利害关系人，如确实无法通知的可以采取公告方式告知。

10.2 数据分类分级

就模型训练或服务提供过程中处理的数据，提供者应依据数据的来源、敏感程度以及数据泄露是否可能危害国家安全、社会公共利益等进行分类分级，并按照数据的不同类别和级别采取不同的保护措施，及时更新满足有关法律、法规、国标、行标等规制的最新要求。

数据分类分级工作应贯穿数据应用的各个环节，并在各个环节适用统一的识别规则并采取相应的技术措施。

10.3 数据删除

因个人撤回同意等原因导致用于模型训练的个人信息需进行删除的，提供者应从数据集中将个人信息删除或进行匿名化处理，不得再用于模型训练。

已投入模型训练的相关信息如无法从模型中删除或删除成本过大的，可采用屏蔽结果等方式停止输出涉及相关信息的内容。提供者也可在数据采集时通过协议就是否需要删除模型内信息进行明确约定。

其他各类因授权到期或未经授权的训练数据删除可参照10.3条进行操作。

10.3条规定不影响提供者就其侵权行为所应承担的法律责任。

10.4 数据跨境

提供者如使用境外来源的基础模型，或从境外来源进行数据采集，在我国境内进行模型训练和内容生成的，宜关注境内外法律合规环境的差异，自查是否符合我国的法律法规要求。

提供者基于模型训练、内容生成等目的，将在中国境内采集的数据传输至境外的，应按照国家有关规定开展数据跨境合规工作。

10.5 算法备案与安全评估

如提供的生成式人工智能服务具有舆论属性或者社会动员能力，应按照国家有关规定以及本文件要求开展合规自评估工作，履行相应的算法备案及安全评估手续。提供者应在相应程序中向主管部门如实上报其数据应用合规的制度建设、落实情况与自评估结果。

参 考 文 献

[1] 中华人民共和国网络安全法(2016 年 11 月 7 日第十二届全国人民代表大会常务委员会第二十四次会议通过）

[2] 中华人民共和国数据安全法（2021 年 6 月 10 日第十三届全国人民代表大会常务委员会第二十九次会议通过）

[3] 中华人民共和国个人信息保护法（2021 年 8 月 20 日第十三届全国人民代表大会常务委员会第三十次会议通过）

[4] 中华人民共和国著作权法（1990 年 9 月 7 日第七届全国人民代表大会常务委员会第十五次会议通过）

[5] 中华人民共和国反不正当竞争法（1993 年 9 月 2 日第八届全国人民代表大会常务委员会第三次会议通过）

[6] 互联网信息服务算法推荐管理规定（2021 年 12 月 31 日国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局公布）

[7] 互联网信息服务深度合成管理规定（2022 年 11 月 25 日国家互联网信息办公室、工业和信息化部、公安部公布）

[8] 生成式人工智能服务管理暂行办法（2023 年 7 月 10 日国家互联网信息办公室、国家发展和改革委员会、教育部、科学技术部、工业和信息化部、公安部、国家广播电视总局公布）

[9] 科技伦理审查办法（试行）（2023 年 9 月 7 日科学技术部、教育部、工业和信息化部、农业农村部、国家卫生健康委员会、中国科学院、中国工程院、中国科学技术协会、中国社会科学院、中央军委科学技术委员会印发）

[10] 具有舆论属性或社会动员能力的互联网信息服务安全评估规定（2018 年 11 月 15 日国家互联网信息办公室、公安部发布）

[11] 网络信息内容生态治理规定（2019 年 12 月 15 日国家互联网信息办公室令第 5 号公布）

[12] 新一代人工智能治理原则——发展负责任的人工智能（2021 年 6 月 17 日国家新一代人工智能治理专业委员会发布）

[13] 新一代人工智能伦理规范（2021 年 9 月 25 日国家新一代人工智能治理专业委员会发布）

[14] GB/T 39335-2020 信息安全技术 个人信息安全影响评估指南

[15] GB/T 41391-2022 信息安全技术 移动互联网应用程序（App）收集个人信息基本要求

[16] Blueprint for an AI Bill of Rights, OSTP, 2022

[17] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, White House, 2023

[18] Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST, 2023

[19] EU General Data Protection Regulation，2016

[20] Ethics Guidelines for Trustworthy AI, European Commission, 2019

[21] WHITE PAPER On Artificial Intelligence – A European approach to excellence and trust, European Commission, 2020