TC260-PG-2024NA

网络安全标准实践指南

——生成式人工智能服务安全应急响应指南

征求意见稿 v1.0-202412

全国网络安全标准化技术委员会秘书处

2024 年 12 月

本文档可从以下网址获得：
www.tc260.org.cn/

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

前　言

《网络安全标准实践指南》（以下简称《实践指南》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织制定和发布的标准相关技术文件，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

本文件起草单位：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、新华融合媒体科技发展（北京）有限公司、清华大学、中国科学院计算技术研究所、中国科学技术大学、北京市公安局人工智能安全研究中心、公安部第三研究所、中国科学院信息工程研究所、浙江工业大学、阿里云计算有限公司、蚂蚁科技集团股份有限公司、科大讯飞股份有限公司、华为云计算技术有限公司、奇安信科技集团股份有限公司、北京天融信网络安全技术有限公司、中国移动通信集团有限公司、中国联合通信有限公司、国家计算机网络应急技术处理协调中心安徽分中心。

本文件主要起草人：赵芸伟、罗毅、孙锦平、闵京华、王鲁华、郭鸿飞、王建民、山世光、俞能海、许敏强、李凤华、陈悦、张妍婷、蔡津津、宣琦、彭骏涛、白晓媛、辛铮、孙勇、冯运波、王俊、王寒生、丁治国、王龑、邱勤、高枫、蒋天翔。

声　明

本《实践指南》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《实践指南》的任何部分。凡转载或引用本《实践指南》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。

摘　要

为贯彻落实《生成式人工智能服务管理暂行办法》要求，指导生成式人工智能服务提供者等有关单位做好安全应急响应工作，本文件围绕生成式人工智能服务安全事件给出了安全事件的分类和分级建议，并给出了生成式人工智能服务安全事件应急响应过程，包括应急准备、监测预警、应急处置、总结改进阶段的管理措施和技术方法，可用于指导生成式人工智能服务提供者提高安全应急响应能力。

目　录

1. 范围

本文件给出了生成式人工智能服务安全事件的分类、分级建议和生成式人工智能服务安全应急响应过程的管理措施和技术方法等内容。

本文件适用于生成式人工智能服务提供者开展安全应急响应活动。

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 20985.1-2017 信息技术 安全技术 信息安全事件管理 第1部分：事件管理原理

GB/T 20986-2023 网络安全事件分类分级指南

3. 术语和定义

下列术语和定义适用于本实践指南

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提供生成文本、图片、音频、视频等内容的服务。

3.2 生成式人工智能服务提供者 generative artificial intelligence 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 违法不良信息 illegal and unhealthy information

《网络信息内容生态治理规定》中指出的 11 类违法信息以及 9 类不良信息的统称。

注：本文件关注的违法不良信息主要是指《生成式人工智能服务安全基本要求》技术文件附录 A 中 31 种安全风险的信息。

3.4 生成式人工智能服务安全事件 generative artificial intelligence service security incident

由于人为原因、恶意攻击、模型的内在漏洞或缺陷等因素引起的，对生成式人工智能服务业务应用及其数据造成损害，对国家、社会、经济造成负面影响的事件。

4. 概述

随着生成式人工智能技术的迅猛发展，文本生成、图像生成、语

音合成、视频生成、代码生成等在各个领域的应用变得越来越广泛。然而，由于生成式人工智能的复杂性和特有风险，其安全性问题日益凸显。例如，生成内容的真实性和准确性问题可能导致严重的误导和虚假信息传播；模型篡改和对抗性攻击可能改变生成内容的性质，造成恶意使用；生成内容中的偏见和不公平性问题可能引发社会和法律纠纷。现有的安全应急响应机制在面对生成式人工智能服务安全事件时难以有效应对。生成内容的不可预测性、模型训练和推理过程的复杂性，使得及时、有效、精准的应急响应变得尤为重要。因此，制定专门针对生成式人工智能服务的安全应急响应标准，指导生成式人工智能服务提供者迅速、有效地响应安全事件已成为当前的迫切需求。

本指南描述了生成式人工智能服务安全事件的分类和分级方法。结合生成式人工智能服务的特点和复杂性，给出生成式人工智能服务安全事件在应急响应过程各个阶段的管理措施、技术方法以及协同建议。

5. 生成式人工智能服务安全事件分类

5.1 分类方法

综合考虑生成式人工智能服务安全事件的起因、威胁、攻击方式、损害后果等因素，对生成式人工智能服务安全事件宜按照 GB/T 20986-2023 5.1 分类方法进行分类，分为信息内容安全事件、数据安全事件、网络攻击事件等 10 类。

5.2 常见生成式人工智能服务安全事件

按照 GB/T 20986-2023 5.1 分类方法，结合生成式人工智能服务特点，本文件列出以下常见生成式人工智能服务安全事件。

5.2.1 信息内容安全事件

信息内容安全事件指生成式人工智能服务生成的信息内容危害国家安全、社会稳定、公共安全和利益的安全事件。按照 GB/TXXXX-XXXX《生成式人工智能服务安全基本要求》附录 A 中 31 种安全风险的信息(如违反社会主义核心价值观的内容、歧视性内容、商业违法违规内容、侵犯他人合法权益内容等)作为进一步分类方法参考。常见的生成式人工智能服务信息内容安全事件包括:

a) 违法信息生成事件

生成式人工智能服务生成违反国家法律法规的内容，可能涉及颠覆国家政权、煽动分裂国家、泄露国家机密、暴力恐怖主义、淫秽色情、引导犯罪行为（例如生成信息内容涉及网络攻击、黑客技术传播或数据窃取等犯罪技术或方法、生成信息内容涉及违法交易、走私、诈骗等犯罪行为引导）等违法信息。

b) 虚假信息生成事件

生成式人工智能服务生成虚假信息扰乱经济秩序和社会秩序，造成负面影响。此类事件包括例如虚假新闻信息（生成虚构或误导性新闻，可能引发社会恐慌），虚假财经信息（生成虚假市场信息或经济数据，扰乱金融市场，可能导致股市波动、价格操纵、投资欺诈等不

良后果），虚假医疗健康信息（生成关于疾病治疗、药物功效的虚假信息，误导公众健康选择，造成公共卫生隐患）等。

c）煽动教唆信息生成事件

生成式人工智能服务生成煽动或教唆他人从事有害行为的信息，危害社会秩序与人身安全。例如教唆自残或自杀（生成涉及教唆他人自残或自杀的信息，对个人产生严重的心理和身体伤害）。

d）权益侵害信息生成事件

生成式人工智能服务生成的信息内容侵害了社会组织或公民的合法权益。例如生成信息内容中涉及隐私侵犯（生成未经授权披露个人或组织的敏感信息）、知识产权侵犯（生成内容涉及侵犯专利、商标等受版权保护的知识产权）、名誉损害、商业秘密泄露等。

e）歧视性信息生成事件

生成式人工智能服务生成的信息内容涉及民族歧视、信仰歧视、地域歧视、性别歧视、职业歧视等。

5.2.2 数据安全事件

数据安全事件指通过技术或其他手段对生成式人工智能服务数据实施篡改、假冒、泄露、窃取等导致业务损失或造成社会危害的安全事件。常见的生成式人工智能服务数据安全事件包括:

a）数据泄露事件

无意或恶意通过技术手段导致数据泄露，包括生成式人工智能服务使用的训练数据、模型参数和用户个人信息等。例如未经授权访问

并下载了生成式人工智能服务的训练数据集，导致敏感个人信息泄露。

b）数据篡改事件

未经授权修改数据，影响生成式人工智能服务的训练和推理等。例如篡改生成式人工智能训练数据集、篡改生成式人工智能模型参数等导致生成内容异常。

c）数据投毒事件

干预深度学习训练数据集，在训练数据中加入精心构造的异常数据，破坏原有训练数据的概率分布，导致生成式人工智能模型在某些特定条件下生成异常内容。

5.2.3网络攻击事件

网络攻击事件指通过技术手段对网络实施攻击而导致生成式人工智能服务业务损失或造成社会危害的安全事件。生成式人工智能服务常见的网络安全事件包括：

a）模型篡改事件

对生成式人工智能模型的底层代码、算法逻辑或执行过程进行未经授权的恶意修改，导致模型行为异常或产生错误输出。这类篡改不仅可能导致生成内容异常，还可能通过植入后门、恶意代码等方式，使攻击者获得对系统的控制权，进一步造成数据泄露、服务中断或更大范围的网络攻击。

b）拒绝服务事件

通过非正常使用网络资源影响或破坏网络可用性，例如 DDoS 攻击导致生成式人工智能服务无法响应用户请求。

c) 漏洞利用事件

通过挖掘并利用网络配置缺陷、通信协议缺陷或应用程序缺陷等漏洞对网络实施攻击。例如攻击者通过生成式人工智能服务开放的 API 接口中存在身份验证漏洞绕过认证，非法访问、获取敏感数据或影响生成内容。

d) 社会工程事件

通过提示词话术设计、恶意轮询等手段诱导生成式人工智能服务泄露数据或执行行动。例如诱导生成个人隐私信息、诱导生成商业秘密、诱导生成恶意代码、诱导生成违法不良信息等。

5.2.4 其他安全事件

按照 GB/T 20986-2023 5.1 分类方法，其他与生成式人工智能服务相关的安全事件。

6. 生成式人工智能服务安全事件分级

生成式人工智能服务安全事件按照事件影响对象（生成式人工智能服务业务应用及数据）的重要程度、业务损失的严重程度和社会危害的严重程度三个要素进行分级。

事件影响对象主要包括生成式人工智能服务业务应用及数据，重要程度根据国家安全、社会秩序、经济建设和公众利益对事件影响对

象的依赖程度进行评估，按照 GB/T 20986-2023 6.1.2 对事件影响对象的重要程度进行分级(该要素分为特别重要、重要、一般 3 个级别)。

业务损失的严重程度取决于恢复生成式人工智能服务正常运行和消除安全事件负面影响所需付出的代价，按照 GB/T 20986-2023 6.1.3 对业务损失的严重程度进行分级(该要素分为特别严重、严重、较大和较小 4 个级别)。

社会危害的严重程度根据对国家安全、社会秩序、经济建设和公众利益等方面的危害程度进行评估，按照 GB/T 20986-2023 6.1.4 对社会危害的严重程度进行分级(该要素分为特别重大、重大、较大和一般 4 个级别)。

综合上述三个要素的级别，生成式人工智能服务安全事件按照 GB/T 20986-2023 6.2 将安全事件分为 4 个级别：特别重大事件、重大事件、较大事件和一般事件，由高到低分别为一级、二级、三级和四级。事件分级流程按照 GB/T 20986-2023 6.3 确定，表 1 描述了业务损失的严重程度与安全事件级别的关系、表 2 描述了社会危害的严重程度与安全事件级别的关系。其中，业务损失的严重程度与社会危害的严重程度同时存在时，两者中取高者确定安全事件级别，表 3 给出了生成式人工智能服务安全事件级别划分的规则描述与示例。

表 1 生成式人工智能安全事件级别与业务损失严重程度的关系

事件影响对象（生成式人工智能服务业务应用及数据）	业务损失严重程度			
	特别严重	严重	较大	较小

的重要程度				
特别重要	一级	二级	三级	三级
重要	二级	三级	三级	四级
一般	三级	三级	三级	四级

表 2 生成式人工智能安全事件级别与社会危害严重程度的关系

事件影响对象（生成式人工智能服务业务应用及数据）的重要程度	社会危害严重程度			
	特别重大	重大	较大	一般
特别重要	一级	二级	三级	—
重要	—	二级	三级	四级
一般	—	—	三级	四级

表 3 生成式人工智能服务安全事件级别划分

事件级别	描述	示例
一级（特别重大事件）	特别重大事件发生在特别重要的事件影响对象上，并且: 1） 导致特别严重的业务损失，或 2） 造成特别重大的社会危害	示例1: 由于黑客攻击，某特别重要生成式人工智能服务的核心模型参数被恶意篡改，导致整个平台无法正常生成任何内容，服务完全瘫痪。 示例2: 某文本类生成服务被诱导生成大量违反社会主义核心价值观的内容，如散布分裂国家、破坏民族团结的言论，这些内容被广泛传播，对社会稳定和国家安全造成了特别严重的影响。
二级（重大事件）	重大事件发生在特别重要或重要的事件影响对象上，并且: 1） 导致特别重要的事件影响对象遭受严重的业务损失或导致重	示例1: 某重要生成式人工智能服务遭受DDoS攻击，导致长时间服务中断。 示例2: 某重要文本类生成服务生成内容泄露大量用户

	要的事件影响对象遭受特别严重的业务损失，或 2）造成重大的社会危害	的敏感个人数据。
三级（较大事件）	较大事件发生在特别重要或重要或一般的事件影响对象上，并且: 1）导致特别重要的事件影响对象遭受较大或较小的业务损失，或重要的事件影响对象遭受严重或较大的业务损失，或导致一般的事件影响对象遭受较大(含)以上级别的业务损失，或 2）造成较大的社会危害	示例1: 某重要生成式人工智能服务由于软件错误等原因造成的短暂中断。 示例2: 某一般生成式人工智能服务生成内容包含误导性的财经信息，并通过传播影响市场稳定。
四级（一般事件）	一般事件发生在重要或一般的事件影响对象上，并且: 1）导致较小的业务损失，或 2）造成一般的社会危害	示例1: 某一般生成式人工智能服务在生成代码的任务中，生成的SQL查询代码可能被恶意利用进行SQL注入攻击。尽管这些问题在代码审查阶段可能会被发现并修正，但如果未能及时检测到，可能会导致这些漏洞进入生产环境，从而对系统安全构成一定威胁。 示例2: 某一般生成式人工智能服务在自然灾害预测系统中，生成式人工智能模型的基础数据遭到部分篡改，导致生成的气象预警报告出现轻微错误。例如，低估了一场即将到来的暴雨的强度。虽然不会导致直接的灾害，但可能

		影响到当地居民的提前准备工作，对公众安全构成一定威胁。

7. 生成式人工智能服务安全应急响应过程

7.1 应急准备

7.1.1 管理措施

应急准备阶段的详细管理措施宜按照 GB/T 20985.1-2017 5.2 实施，主要管理措施包括以下内容。。

a）应急策略制定与最高管理者承诺：制定专门针对生成式人工智能服务的安全事件应急策略，并确保最高管理者的支持与承诺。这包括识别生成式人工智能的独特风险，如生成内容的真实性、准确性和公正性问题等；

b）安全事件管理计划：制定详细的生成式人工智能服务安全事件管理计划，包括监测预警、应急处置方面的实施机制，并对其进行定期测试和评估，以确保及时发现并有效处置生成式人工智能服务安全事件；

c）安全事件应急响应预案：依据应急策略、安全事件管理计划，宜结合第 5 章、第 6 章制定不同类别（如生成内容异常、模型篡改、数据泄露等）、不同级别（一级—四级）的安全事件应急响应预案，确保各种类型、各种级别的生成式人工智能服务安全事件发生时快速有效地响应；

d）事件升级策略和程序：建立、审议安全事件升级的策略和程序；

e）上下线管理审查程序：建立、审议生成式人工智能服务上下线管理审查程序；

f）事件响应小组（IRT）：建立专门的生成式人工智能安全事件响应小组，确保其具备必要的技术和管理技能，并定期更新其知识和技能以应对生成式人工智能领域的快速发展；

g）培训和技术支持：为 IRT 成员提供定期的培训，特别针对生成内容的质量控制、数据安全、模型安全方面提供必要的技术支持。

7.1.2 技术方法

应急准备阶段，应建立并维护一个全面且定期更新的关键词库和测试题库，为生成式人工智能服务安全应急响应提供全面性、动态化的安全风险知识储备能力。

a）全面性：关键词库和测试题库应覆盖生成内容的主要安全风险，具体要求参考《生成式人工智能服务安全基本要求》8.1、8.2；

b）即时性：关键词库和测试题库应根据最新的安全威胁和安全事件进行实时更新。每当有重点新闻事件或安全威胁出现时，关键词库和测试题库应在 24 小时内更新；

c）针对性：对于可能引发安全风险的重点或热点事件，关键词库中不宜少于 20 个相关的关键词，确保能够捕捉和响应这些事件的具体内容和相关话题。应拒答测试题库不宜少于 50 题。

7.1.3 外部协同

应急准备阶段的外部协同包括:

a) 建立与产品或服务供应商、第三方安全机构、行业主管部门等利益相关方的协同机制;

b) 定期与利益相关方或相关专家进行沟通，获取关于生成式人工智能服务安全应急响应的最新动态和最佳实践，更新应急策略。

7.2 监测预警

7.2.1管理措施

监测预警阶段的详细管理措施宜按照 GB/T 20985.1-2017 5.3 实施，主要管理措施包括以下内容。

a) 制定监测策略（宜结合第 5 章安全事件分类，设置有针对性的监测策略），包括监测的频率、关键指标和阈值;

b) 按照监测策略执行监测任务，定期组织监测效果评估，依据评估意见调整和改进监测策略;

c) 建立预警机制，监测到生成式人工智能服务中的潜在威胁或服务异常时自动通知相关人员;

d) 建立快速响应机制，发现服务异常后迅速报告并启动初步调查和评估。

7.2.2 技术方法

监测预警阶段的技术方法包括:

a) 实时监测: 利用自动化监测工具与人工审查结合的方式，实

时监测生成式人工智能服务的模型行为和数据活动，及时发现异常、可疑或恶意活动。包括但不限于:

1）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;

2）对模型输入内容持续监测，防范恶意输入攻击，例如恶意轮询（监测用户请求的频率和模式）、DDoS、XSS、注入攻击等;

3）对模型输出内容持续监测，对于生成内容异常事件设置实时内容校验和异常检测机制，对于可能引发安全风险的重点或热点事件，基于应拒答测试题库，模型的拒答率不低于 95%;

4）对数据泄露、数据篡改风险重点监测数据访问行为;

5）对模型篡改风险重点监测模型的参数变化;

6）对服务中断风险重点监测网络流量异常、系统资源过载等指标;

7）定期更新漏洞数据库，使用自动化工具结合人工审查的方法定期扫描服务的安全漏洞。

b）数据分析：利用大数据分析和机器学习技术，分析监测用户输入行为、生成内容、模型参数等数据，识别生成式人工智能服务异常;

c）安全预警：对关键指标设置阈值和触发条件自动预警，对热点事件和重要舆情及时更新预警策略。

1）依据关键指标及阈值自动触发，例如异常请求率、服务系统负载等，一旦超过阈值自动触发预警；

2）跟踪热点事件或重要舆情，识别潜在的安全威胁，给出生成内容风险提示和预警，有针对性地更新关键词、测试题库。

7.2.3 外部协同

监测预警阶段的外部协同包括：

a）生成式人工智能服务提供者宜收集来自各利益相关方提供的服务异常监测线索；

b）各利益相关方共享热点事件或重要舆情的威胁情报、风险提示信息和预警信息。

7.3 应急处置

7.3.1管理措施

应急处置阶段的详细管理措施宜按照 GB/T 20985.1-2017 5.4、5.5 实施，主要管理措施包括以下内容。

a）评估与决策：对监测预警阶段的服务异常进行评估与决策，判断是否将服务异常归为安全事件并进行分类分级。记录所有评估与决策活动，以便后续分析和改进。引入独立的第三方专家进行评估，以确保评估与决策的权威性和准确性；

b）启动应急响应预案：对确定的安全事件，根据事件类型和级别，启动安全事件应急响应预案，必要时启动下线管理审查程序。

c）应急调度：按照安全事件应急响应预案，对需要应急响应的

安全事件开应急调度并协调所需的相关人员、资金和技术工具等;

d) 排查与诊断：IRT 对生成式人工智能服务安全事件原因进行排查与诊断，包括模型安全排查与诊断、数据安全排查与诊断、生成内容异常排查与诊断等。将排查与诊断的过程与结果信息进行整理与归档;

e) 处理与恢复：基于安全事件应急响应预案采取相应的应对措施进行安全事件处理和服务恢复;

f) 服务测试评估：处理与恢复后，定期评估生成式人工智能服务效果，包括模型性能、数据安全和生成内容质量等方面的测试和评估;

g) 事件关闭：包括生成式人工智能服务安全事件关闭的申请、核实、调查取证及关闭通报等;

h) 服务上线管理审查：因安全事件导致生成式人工智能服务在 b）条款启动下线管理审查程序的，需要启动服务上线管理审查程序。

7.3.2 技术方法

应急处置阶段的技术方法包括:

a) 安全事件分级处置：按照第 6 章不同事件级别采取不同的处置措施。一级事件应下线服务；二级事件应实施全面的功能限制，必要时下线服务；三级事件应实施部分功能限制，必要时下线服务；四级事件应启通过局部调整或漏洞修复等措施保持服务连续性，必要时实施部分功能限制。

b) 安全事件报送

1) 即时报送主管部门研判

对于三级及以上级别事件，如涉及严重违反社会主义核心价值观、大规模数据泄露等可能对公众造成广泛影响的事件，当 24 小时内累计发生 5 次，影响用户数量众多（用户数大于 1 万），可能造成不良社会影响的，应立即报告主管部门；

2) 自行处置定期报送

对于一般事件，服务提供者可以在内部完成处置，包括问题诊断、快速修复、和恢复服务。这些事件通常不会对服务的整体安全性或用户数据造成长期影响。完成处置后，应将安全事件情况、处理措施及结果等详细记录并定期报送给主管部门备案，以便进行后续的监督和审查；

c) 安全事件分类处置技术措施：在确认安全事件后，立即采取适当的安全技术措施，包括快速隔离（视安全事件级别和事件具体情况启动下线审查程序）、修复漏洞、对生成式人工智能模型进行调整或重新训练。例如，对于生成异常事件，通过实时调优模型参数和更新训练数据来纠正错误输出；对模型篡改事件，利用备份模型和对抗性训练技术快速恢复模型的正常功能；对数据泄露事件，采用数据加密和访问控制技术保护敏感数据，并及时更新和修复相关安全策略。

d) 服务恢复测试与评估：在安全事件处置完毕后，对生成式人工智能服务进行全面的安全测试与评估。特别关注生成式人工智能服

务的生成内容质量，采用自动化测试工具和人工评审相结合的方法，基于测试题库对生成内容的质量、准确性和一致性进行严格测试。

7.3.3 外部协同

应急处置阶段的外部协同包括:

a) 生成式人工智能服务提供者向其他利益相关方通报安全事件的类别、级别、安全故障原因等安全事件信息;

b) IRT 进行故障排查和诊断，必要时可寻求其他利益相关方以现场或远程方式提供技术支持。

7.4 总结改进

7.4.1管理措施

总结改进阶段的详细管理措施宜按照 GB/T 20985.1-2017 5.6 实施，主要管理措施包括以下内容。

a) 应急响应工作总结：生成式人工智能服务提供者应定期对应急响应工作进行分析和回顾，总结经验教训，并采取适当的后续措施，对应急响应工作的分析和回顾应形成总结报告;

b) 应急响应工作审核：生成式人工智能服务提供者应定期组织应急响应工作审核，确保应急响应过程和方法符合预定的策略和要求。评审应至少每年进行一次，或在发生重大变更、发生应急事件后进行;

c) 应急响应工作改进：应急响应工作总结、应急响应工作审核的结果应作为应急响应各项工作的改进要素。宜依据总结报告中给出

的建议项和审核结果，结合生成式人工智能服务安全应急响应的最新动态和最佳实践，不断优化应急策略和应急响应预案。

7.4.2技术方法

总结改进阶段的技术方法包括：

a）经验反馈机制：在安全事件处置完成后，建立一套反馈机制，用于收集和分析安全事件相关数据，以评估应急响应管理和技术措施的有效性，并识别潜在的改进点；

b）模拟测试：定期对安全事件应急响应过程进行模拟测试，以检验应急响应预案的实用性和有效性。测试结果可用于优化预案和提高响应能力；

c）技术审计：利用数据分析和机器学习技术，分析生成式人工智能服务安全事件日志并挖掘安全漏洞，归纳典型安全事件行为模式，以改进应急响应相关管理或技术措施；

d）知识管理和更新：将安全事件应急响应经验整理成案例库，定期更新安全操作手册和培训材料，确保 IRT 了解典型的安全事件和响应经验。

7.4.3 外部协同

总结改进阶段的外部协同包括：

a）用管联动，与行业主管部门保持密切的沟通和协调，确保应急响应预案符合最新的法规要求，在发生重大安全事件时能够迅速报送并得到主管部门的支持和指导；

b）知识协同：将生成式人工智能服务安全事件应急响应经验总结转化为培训材料，与其他利益相关方进行知识分享，提升全体安全意识和应急响应能力。

附录 A

（资料性）

生成式人工智能服务安全应急响应典型案例

A.1 信息内容安全事件—歧视性信息生成事件安全应急响应案例

A.1.1 案例概述

某生成式人工智能服务在用户交互过程中，因模型训练数据偏差，生成了一段含有轻微性别歧视的文本内容。该文本涉及刻板印象，并对特定性别进行了不恰当的评价，用户在使用该生成内容时感到不适，并向平台提交了投诉。虽然事件影响较小，但该歧视性内容引发了一些负面反馈。

A.1.2 应急准备

该案例中，应急准备工作主要包括:

生成式人工智能服务提供者制定经内部评审、高层审批过的涉及生成式人工智能服务安全事件的应急策略、管理计划、应急预案等制度文件。制度中明确上报及升级流程、IRT 团队、歧视性信息定义等内容。提供者针对歧视性信息生成事件制定了专门的应急响应预案，定期组织 IRT 团队安全培训和应急演练。

生成式人工智能服务提供者建立涉生成歧视性信息的关键词库和测试题库，定期更新涉生成歧视性信息相关的关键词库和测试题

库，涵盖性别、种族、宗教等敏感话题的词汇，并结合语义分析模型，对生成内容进行自动筛查和警报触发。

生成式人工智能服务提供者具备与利益相关方（包括但不限于供应链上下游、用户、主管部门等），建立对歧视性信息的同步、更新、处置机制，与利益相关方建立定期沟通机制和事件处置沟通机制。

A.1.3 监测预警

该案例中，歧视性信息通过用户投诉和系统监控被发现：

a）用户投诉触发事件：某用户在与生成式人工智能服务的交互过程中，生成了一段含有性别歧视的内容。生成文本中包含了关于某特定性别的刻板印象，例如：“女性通常不擅长逻辑推理，这类任务更适合男性。”该用户感到内容不适，随即向平台提交了投诉。

b）系统监控发现问题：平台的监控系统也检测到了这段内容，触发了预设的关键词警报，如“女性”、“不擅长”等敏感词，并通过语义分析确认了文本中的性别歧视倾向。

A.1.4 应急处置

该案例中，IRT 团队成员接到监测预警后，对生成信息内容进行了人工审查。

a）确认生成内容存在轻微的性别歧视倾向，虽然影响较小，但仍可能影响用户体验。按照生成式人工智能服务安全事件分类分级规则定为歧视性信息生成事件，事件级别定为四级（定级说明：考虑某

生成式服务本身的重要程度为一般，社会危害的严重程度为一般，按照定级规则定为一般事件，即四级）；

b）IRT 团队立即对该生成模型进行了调整，确保问题内容不再被重复生成。技术团队进一步优化了涉及敏感话题的生成规则，防止歧视性内容再次出现。IRT 团队通过邮件与提交投诉的用户进行了沟通，解释问题的来源并向用户澄清和致歉，确保用户了解平台已修复问题并优化了服务；

c）IRT 团队针对该生成模型进行了进一步的检查，确保类似的歧视性内容不会再次生成。同时，团队加强了内容筛查的力度，对其他可能触发歧视性内容的关键词进行了详细的分析和改进，IRT 迅速完成了问题修复并恢复服务。

A.1.5 总结改进

事件处理完毕后，IRT 团队对该事件进行了复盘，并提出了改进建议，以防止类似事件再次发生。

a）IRT 团队详细分析了此次事件的根本原因，确认生成模型在处理性别相关话题时存在偏差。团队通过复盘，决定加强对涉及性别、种族、宗教等敏感话题的内容生成的监控和规则优化；

b）IRT 团队进一步扩充了性别平等和敏感话题的关键词库，特别加强了对性别歧视、刻板印象相关词汇的监控力度，并优化了语义分析算法，以更早发现潜在的歧视性内容；

c) IRT 团队配合相关技术团队对模型的生成算法进行了调整，加强了对内容生成中的公平性要求，确保生成内容在语言表达上更加中立，不带有歧视性表述；

d) 组织 IRT 团队成员关于歧视性内容培训，提升员工识别和处理歧视性信息的能力。此外，平台还加强了与第三方机构的合作，定期评估生成内容中的潜在歧视问题，确保平台服务符合社会公平和道德标准。

A.2 信息内容安全事件—虚假信息生成事件安全应急响应案例

A.2.1 案例概述

某生成式人工智能服务因数据处理错误，生成了一则关于某知名制药公司药物存在致命副作用的虚假消息。该消息迅速在社交媒体和新闻平台上传播，引发公众恐慌，大量患者停止使用该药物，制药公司和医院因此蒙受经济损失。

A.2.2 应急准备

该案例中，应急准备工作主要包括：

a) 建立了专门的生成式人工智能安全事件响应小组(IRT)：成员包括生成式人工智能业务应急接口人、合规应急接口人、法务应急接口人、公关应急接口人等；

b）建立了生成式人工智能事件响应策略和程序，遵循快速响应、就高不就低、止损优先等原则，制定事前、事中、事后的不同流程；

c）制定了生成式人工智能事件应急预案，包括依据被指南对生成式人工智能安全事件的分类分级、具体风险场景、涉及的具体业务或产品、可能发生的生成式人工智能安全事件的类别和级别、应急流程中每个环节涉及的操作人、具体执行动作（包含但不限于涉及的工具、系统、模板、口径）、操作时效、输出结果、信息传递方式等；

d）建立了生成式人工智能事件信息感知渠道，包括但不限于内部运维感知、主管部门下发信息、批量客诉、舆情、其他应急响应组织的沟通协调等。

A.2.3 监测预警

该案例中，据客诉反馈和其他相关方的投诉反馈经人工审查监测发现生成式人工智能服务生成“最新药物 X 已经在临床试验中导致多名患者死亡，副作用严重，建议停止使用此药物。”的虚假信息。服务提供者通过实时监控机制也检测到相关虚假信息。

a）用户投诉触发事件：某用户在与生成式人工智能服务的交互过程中，生成了一段含有虚假信息的内容。生成文本中包含了关于制药公司药物存在致命副作用的虚假消息，相关用户随即向服务提供者提交了投诉。

b）系统监控发现虚假信息：系统检测到信息中出现了诸如“药物X”、“临床试验”、“死亡”、“副作用严重”等关键词。这些词语与医疗

领域的高风险敏感词库匹配，触发了自动预警机制。该机制结合语义分析技术，对生成内容的逻辑和潜在影响进行了初步分析，该信息可能会误导公众。

A.2.4 应急处置

该案例中，IRT 团队成员接收客诉反馈和监测预警提示后立即介入生成信息内容人工审查。

a）评估与决策：通过核对生成信息内容与实际医疗信息的不一致性，审查团队评估并确认该信息为虚假信息，并且潜在风险较大，可能造成虚假信息传播并引发社会恐慌。按照生成式人工智能服务安全事件分类分级规则定为虚假信息生成事件，该虚假信息事件涉及公众健康和医疗行业信任，潜在传播风险高且社会危害程度较重，事件级别定为二级（定级说明：考虑某生成式服务本身的重要程度为重要，社会危害的严重程度重大，按照定级规则定为重大事件，即二级）

b）传播控制：启动应急响应预案，IRT 立即通知技术团队对生成式人工智能服务进行临时下线，停止该模型的进一步内容生成。同时，提供者与主要社交媒体和新闻平台协同，检查已经发布的虚假信息内容并删除相关虚假信息，防止进一步扩散。对于已经产生传播影响的，发布澄清声明，澄清虚假信息的内容，提醒公众切勿轻信未经证实的信息。

c）影响评估与应对：对该事件可能对自身业务或其他相关方造

成的影响进行评估，分析该虚假信息生成事件对自身或其他相关方的经济影响、信誉影响等，联合业务、法务、合规、公关和客服团队对潜在可能产生的客诉风险、负面声誉风险等提前进行防护和准备，遏制事件影响范围继续扩大。同时 IRT 向服务提供者和其他相关方报告该事件的处置详情；

d) 处置与恢复：技术团队根据虚假信息生成的原因，修正了模型的参数设置，并对相关的数据输入和生成规则进行了优化，防止类似的虚假信息再次生成。

e) 服务测试与上线：处置与恢复后，通过针对性的关键词库和测试题库的多轮测试评估修复后的生成式人工智能服务效果，对临时下线的服务申请上线审查，通过上线审查管理程序后恢复上线服务。

A.2.5 总结改进

IRT 团队对事件进行复盘并形成书面报告，定期对应急响应工作进行分析和回顾，总结经验教训，并改进应急流程与措施。

a) 关键词库与监控规则优化：平台更新了关键词库，并对虚假信息的监控规则进行优化，增加了更灵敏的语义分析算法，确保今后能更早识别类似高风险内容；

b) 改进模型训练和数据处理机制：IRT 团队配合技术团队加强模型的训练数据管理，确保模型训练时的数据来源更为准确，同时对数据清理和标注环节进行优化，以减少生成虚假信息内容的概率；

c) 员工培训与应急预案演练：平台对 IRT 团队进行了针对性的

培训，总结本次事件的经验教训，同时定期开展应急预案演练，确保团队在未来遇到类似事件时能更迅速、精准地做出响应。