ICS 35.030
CCS L 80

T/CSAS

团 体 标 准

T/CSAS 0029—2025

人工智能数据安全要求

Requirements for security of artificial intelligence data

2025-11-28 发布 2025-12-29 实施

四川省网络空间安全协会 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。

本文件由四川省网络空间安全协会提出并归口。

本文件起草单位：成都信息工程大学、成都理工大学、全域数据信息安全重点联合实验室西南实验室、成都久信信息技术股份有限公司、泰和泰律师事务所、成都四方数安信息技术有限公司。

本文件主要起草人：白杨、刘珏麟、张露、李冬芬、多滨、李瑞、柯钦文、欧阳志强、罗荣、尹晓东、陈福中、黄恒、卢志超（排名不分先后）。

引　言

为防范人工智能生命周期各阶段面临的隐私泄露、算法歧视、模型滥用等数据安全威胁，加强人工智能技术全生命周期数据安全管理，特制定人工智能数据安全要求，旨在为人工智能研发机构、服务提供方及监管部门提供全链条规范性指导。

人工智能数据安全要求

1 范围

本文件规定了人工智能生命周期中数据收集、数据预处理、模型训练、模型评估、模型部署、模型应用、模型更新各环节的数据处理活动的安全要求，重点约束存在数据安全风险的人工智能应用场景。本文件所提的人工智能数据包括但不限于训练数据、评估数据、模型数据。

本文件适用于人工智能领域的数据处理者规范数据处理活动，监管部门、第三方评估机构对人工智能数据处理活动进行监督、管理、评估参照使用。

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 信息安全技术 术语
GB/T 35273 信息安全技术 个人信息安全规范
GB/T 35274 数据安全技术 大数据服务安全能力要求 术语
GB/T 37988 信息安全技术 数据安全能力成熟度模型
GB/T 39335 信息安全技术 个人信息安全影响评估指南
GB/T 41479 信息安全技术 网络数据处理安全要求
GB/T 43697 数据安全技术 数据分类分级规则
GB/T 45574 数据安全技术 敏感个人信息处理安全要求

3 术语和定义

GB/T 25069—2022和GB/T 35274—2023界定的以及下列术语和定义适用于本文件。

3.1

训练数据 training data

用于人工智能模型的输入数据样本子集。

[来源：GB/T 41867—2022，3.2.34，有修改]

3.2

评估数据 evaluation data

用于评估最终人工智能模型性能的数据。

[来源：GB/T 41867—2022，3.2.3，有修改]

3.3

投毒数据 poisoning data

在训练数据中被有意或恶意注入带有虚假、误导性或有害信息，使人工智能模型在后续推理或决策中产生错误、偏见或恶意输出的数据。

3.4

模型训练 model training

利用训练数据，基于人工智能算法，确定或改进人工智能模型参数的过程。

[来源：GB/T 41867—2022，3.2.18，有修改]

3.5

基座模型 base model

通过大规模无监督预训练形成的通用人工智能模型，具备广泛的语言理解与生成能力，可作为后续任务的基础架构。

注：典型示例包括Qwen系列、GPT系列、LLaMA系列等通用预训练模型。

3.6

模型混淆　model confusion

通过一系列技术手段，如网络结构加扰、参数权重加扰、算子权重加扰等，对模型的结构、输入、输出进行改造，使得模型的原始架构和参数变得不透明，难以被轻易理解或复制的一种方法。

注：常见应用包括在模型交付或部署前对权重文件进行加密混淆，或通过引入自定义算子、非标准层结构提升模型防逆向分析能力。

3.7

生成合成信息　generative synthetic information

利用人工智能技术对文本、图像、音频、视频、场景模型等进行生成或者编辑所得到的信息。

［来源：GB/T 42888—2023，3.7，有修改］

3.8

数据漂移　data drift

模型输入数据的统计特性随时间发生变化，这种变化可能是由于数据的分布、范围或频率的变化导致的。

3.9

概念漂移　concept drift

模型的输入数据与目标变量之间的关系发生的变化。

3.10

样本失效　sample failure

因时效性超期、数据异常、合规性不符、技术故障或关联数据失效等原因，导致样本不再适用于当前处理目的，需重新获取授权或停止使用的状态。

3.11

对抗样本检出率　adversarial sample detection rate

在数据集中识别出对抗样本的比例。

注：对抗样本是指通过在数据中故意添加细微的扰动生成的一种输入样本，能导致神经网络模型输出错误的预测结果。

3.12

数据污染检出率　data contamination detection rate

在数据集中识别出被污染样本的比例。污染样本如篡改、注入噪声或恶意数据。

注：污染样本是指在原始样本中经过篡改、注入噪声或恶意数据得到的样本。

3.13

模型水印　model watermark

嵌入到人工智能模型中，用于标识模型来源、确认模型权属或验证模型完整性的特征信息。

3.14

数据处理者　data processor

开展数据处理活动的个人或组织。

3.15

人工智能开发者　artificial intelligence developer

在人工智能系统的生命周期中，从事算法模型构建、数据处理、程序实现或系统集成等活动的组织或个人。

3.16

云服务提供商　cloud service provider

提供云服务的第三方公司。

［来源：GB/T 32400—2015，3.2.15，有修改］

4　概述

4.1　人工智能生命周期组成

如图1所示，人工智能生命周期由数据收集、数据预处理、模型训练、模型评估、模型部署、模型应用和模型更新等核心环节构成，这些环节涵盖了从原始数据到智能应用落地的完整流程。

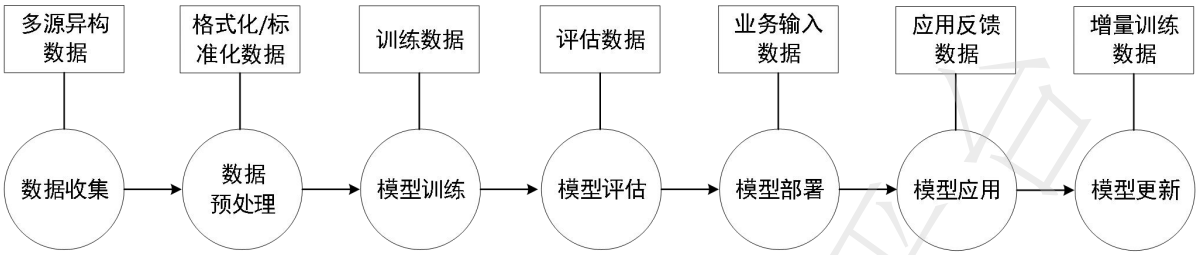

图 1　人工智能生命周期组成及数据类型

数据收集聚焦于多源异构数据的获取；数据预处理通过清洗、标注与特征工程等预处理操作形成结构化数据；模型训练通过算法从训练数据中学习规律；模型评估用于验证效果并持续优化；模型部署将训练成果转化为实际服务以满足业务需求；模型应用通过程序接口或交互界面将模型推理能力转化为实际业务功能；模型更新通过参数微调或全量重训练等方式实现模型迭代，确保系统持续适应动态环境。结合人工智能的开发范式，其生命周期由上述7个递进且迭代的模块化阶段组成。各阶段的具体活动定义如下：

a) 数据收集：组织或个人为支持人工智能开发，根据特定业务需求，系统性地采集、获取相关数据的过程；
b) 数据预处理：依据既定分析目标，对原始数据进行清洗、转换、集成和优化，以提高数据质量和可用性的活动；
c) 模型训练：利用算法，基于预处理后的数据集进行模型参数优化，使模型能够从数据中学习并建立预测或分类能力的活动；
d) 模型评估：基于测试数据对人工智能模型的性能指标进行系统性量化分析的活动；
e) 模型部署：将训练完成并通过验证的模型集成到生产环境，使其能够为实际业务提供预测或决策支持的活动；
f) 模型应用：将机器学习或人工智能模型部署到实际应用场景中，利用提示词或其他方式，如模型接口调用，以解决特定问题或提供特定功能的过程；
g) 模型更新：针对模型应用阶段发现的数据漂移、概念漂移或性能衰减等现象，结合增量训练数据和应用反馈数据，通过参数微调、结构优化或全量重训练等方式对模型进行迭代升级的活动。

4.2　人工智能数据活动主体及典型数据活动列举

人工智能数据活动涉及的数据处理角色包括数据主体、数据处理者、人工智能开发者、云服务提供商、数据标注提供商、数据经纪人等。

数据处理角色及典型数据活动列举：

a) 数据主体：数据生成与提供、数据使用授权等；
b) 数据处理者：数据预处理，如数据清洗、数据转换、特征工程等；
c) 人工智能开发者：数据加工和使用，如数据增强、模型训练、模型微调等；
d) 云服务提供商：模型部署支持、数据安全托管、数据存储与管理等；
e) 数据标注提供商：数据标注，如功能性数据标注、生成式人工智能数据标注、安全性数据标注、微调训练数据标注、偏好数据标注等；
f) 数据经纪人：数据安全风险管理、数据合规审核等。

4.3　常见安全风险

人工智能数据处理活动中常见的安全风险主要包括数据泄露风险、数据篡改风险、数据偏见歧视等，常见安全风险及典型威胁见附录A。

5　基本安全要求

对人工智能数据处理者的基本安全要求如下：

a） 应符合GB/T 37988、GB/T 41479、GB/T 35273 规定的要求；
b） 应按照GB/T 43697 的相关规定，对人工智能数据进行分类分级管理；
c） 开展人工智能数据处理活动，如涉及个人信息，应按照GB/T 39335 的规定开展个人信息安全影响评估，并形成评估报告；
d） 应采取措施确保数据主体权利，包括但不限于保障知情同意、获取人工智能数据使用情况、撤回授权、投诉、获得及时响应等；
e） 应针对不同类型、不同阶段的人工智能数据处理活动制定数据安全保护计划，并向相关数据主体公开保护计划；

注：数据安全保护计划的公开范围应包括与数据处理活动直接相关的数据处理者、人工智能开发者、云服务提供商、数据标注提供商、数据经纪人以及受监管机构监督的特定场景下的监管主体等。公开形式应根据相关方需求和合规要求，通过内部通告、协议附件或公开声明等渠道进行适度披露，在保障透明度与可追溯性的前提下实现精准覆盖。

f） 应明确人工智能数据保护负责人，负责人工智能数据保护工作；
g） 应制定数据安全评估制度，定期（如每年）对人工智能技术应用的必要性、安全措施的有效性、数据使用目的的授权情况等进行评估，并根据评估结果完善保护计划；
h） 应设立数据安全监控指标，在发生数据安全事件或确定存在发生数据安全事件的可能性时，及时评估损失情况，采取补救措施；
i） 在中华人民共和国境内收集或产生的人工智能数据应在境内存储。因业务需要确需出境的，应按照国家有关规定执行；
j） 凡涉及采用密码技术解决保密性、完整性、真实性、不可否认性需求的，应遵循密码相关国家标准和行业标准。

6 数据收集

对人工智能数据处理者的基本安全要求如下：
a） 应明确人工智能数据采集的训练目标与应用场景，规范数据的获取来源、规模及更新频度，确保符合数据安全法律法规及算法伦理要求；
b） 获取个人信息时，应征得数据主体的同意，并确保获取数据过程中数据不被泄露或篡改；
c） 收集个人敏感信息应满足GB/T 45574—2025 的要求，采取严格的数据安全保护技术（如匿名化处理等），确保敏感信息不被泄露；
d） 应对数据收集和获取环境中的设施和技术进行安全管控，确保数据的完整性、一致性和真实性；
e） 在数据收集阶段，应实施身份鉴权机制，通过数字证书、生物特征等多因子认证方式核验数据提供方的身份合法性；
f） 对于公开数据的采集，应审查数据的完整性与可靠性，确保数据未被投毒或篡改；
g） 记录数据收集和获取过程，应对数据收集和获取操作过程全程监管，确保数据来源可追溯；
h） 已授权的数据主体的样本失效后，再次收集数据前应重新取得数据主体的单独同意；
i） 应遵循最小必要原则，只采集满足业务所需的最少数据，如采集字段不超过业务目标直接关联的必要数据项。

7 数据预处理

对人工智能数据处理者的基本安全要求如下：
a） 应对数据预处理所涉及的平台、组织或机构进行合法性及正当性审查，确保其符合相关法律法规要求，并具备安全可靠的数据处理能力；
b） 应对训练数据进行审查，防止数据出现偏见、歧视等内容，并针对不同的规模、批次、可溯源性等投毒数据，进行分类处理；
c） 对于隐私数据，应采用隐私保护技术进行处理，宜采用联邦学习、差分隐私等技术，保障数据的机密性和模型的可用性。在差分隐私的应用中，应根据数据敏感度、任务需求及模型性能要求合理设置隐私预算，并建立相应的审计与评估机制，对隐私预算的分配、累计及使用过程进行记录和核查，防止形式化或过度使用造成的隐私泄露风险；

d) 使用数据标注处理数据，标注方应依据数据分类分级标准，采取差异化安全防护措施，包括但不限于数据脱敏、访问控制及加密传输等技术手段，确保标注数据全流程安全可控；
e) 对于预处理后的数据，应采用访问控制技术，如基于角色的访问控制、多因素认证等，限定数据访问权限，防止数据遭受非授权的访问、篡改、加工等操作；
f) 处理人工智能生成合成信息，应按照有关规定在元数据中添加隐式标识或显式标识；

注：显式标识是指在生成合成内容或者交互场景界面中添加的，以文字、声音、图形等方式呈现并可以被用户明显感知到的标识。隐式标识是指采取技术措施在生成合成内容文件数据中添加的，不易被用户明显感知到的标识。

g) 应建立规范化的数据预处理流程，确保各环节符合法律法规要求，并采取数字签名、加密传输等技术确保数据的完整性和机密性；
h) 应对预处理过程进行评估与监控，如发现潜在风险或异常行为，及时预警并采取相应处置措施。

8 模型训练

对人工智能数据处理者的基本安全要求如下：

a) 训练数据的使用应针对组织内部建立安全防护机制，如采用身份鉴权、访问控制等技术防止数据泄露、未授权使用等风险；
b) 对于分布式或多方合作的训练场景，应采用联邦学习等隐私保护技术，防止数据泄露；
c) 委托第三方机构开展模型训练的，应当签订书面协议，明确数据安全责任，确保受托方具备相应技术能力与合规资质，并建立全流程监督机制；
d) 应对训练数据的使用过程进行记录、审计，确保数据使用可追溯。记录内容应包括数据来源、使用时间、用途、访问主体、数据处理方式及授权信息等关键字段。相关日志应加密存储并保留不少于三年，至少每季度开展一次数据使用审计；
e) 基座模型获取应确保来源合法合规，通过正规授权渠道获取，并保留完整的授权证明文件；
f) 应对基座模型实施模块化安全评估，重点防范参数篡改、梯度扰动等攻击行为，确保模型各组件安全可靠；
g) 模型训练数据的使用应遵循国家相关法律法规，防止数据滥用；
h) 在涉及多方数据协同建模或敏感数据参与训练的场景中，宜采用同态加密或安全多方计算等隐私计算技术。采用上述技术时，应综合考虑计算性能和通信开销，确保隐私保护与模型效能的平衡。

9 模型评估

对人工智能数据处理者的基本安全要求如下：

a) 模型评估数据集应实施严格的质量控制与安全审查，重点检测数据完整性、代表性及潜在偏见风险，防范通过数据投毒实施的模型操纵攻击；
b) 使用日志记录模型评估的全过程，应确保评估活动的可追溯性与可验证性。日志内容应涵盖评估时间、评估指标、数据来源、模型版本及评估结果等关键要素。日志文件应加密存储并保留不少于三年，至少每季度开展日志审计与一致性核查，确保记录完整、真实与可追溯；
c) 采用密码技术、校验技术、隐私保护等技术为模型评估提供完整性和机密性保障；
d) 应建立模型输出内容审查机制，形成自动化检测与人工复核相结合的协同审查模式。自动化检测用于实现对生成内容的快速筛查与初步识别，人工复核用于对高风险或疑似违规内容进行精确判断。审查机制应确保覆盖全面、响应及时、处置有序，对识别出的违规内容实施分级管理与处置；
e) 委托第三方机构开展评估时，应确保评估机构具备相应资质，并签订保密协议明确数据安全责任，同时建立评估过程监督机制；
f) 设立模型数据安全指标，如对抗样本检出率、数据污染检出率等，应确保评估结果的可信度与合规性；
g) 在模型评估阶段，宜采用同态加密或安全多方计算等隐私计算技术，支持在不暴露原始数据和模型参数的前提下开展联合评测。应根据评估任务的敏感程度与资源约束，合理选择技术方案，确保评估过程中的数据安全及结果可验证性。

10 模型部署

对人工智能数据处理者的基本安全要求如下：

a) 使用开源框架部署模型，应使用漏洞扫描和代码审计等技术，对所采用的开源框架进行全面的安全评估，确保框架的安全性符合生产环境要求；
b) 在涉及敏感信息的应用场景中，应优先采用本地化部署方案，通过物理隔离和网络边界防护等措施，确保敏感数据资产与隐私信息的安全可控；
c) 应采用访问控制和身份认证等技术，限制部署阶段数据处理人员的权限，防止未经授权的数据操作或内部恶意篡改；
d) 应采用模型混淆、权重加密等技术，防范模型逆向工程及未授权滥用风险；
e) 应建立完备的日志审计机制，对模型调用行为进行全链路监控，针对异常访问、越权操作等安全事件实施实时告警与自动化响应处置；
f) 在部署过程中，对于集中存储的数据，应采用数据加密、访问控制等技术手段，防止数据泄露；
g) 建立持续监控和动态更新机制，部署后应对模型运行状态、数据输入分布及响应结果进行全链路监控。当检测到性能下降、数据漂移等情况时，自动触发模型更新流程。更新前需进行样本筛选与增量训练数据验证，更新后通过离线评测与在线验证双重环节，确保模型性能与鲁棒性持续优化。

11 模型应用

对人工智能数据处理者的基本安全要求如下：

a) 应遵循《促进和规范数据跨境流动规定》，并结合具体业务情况和相关法律要求，选择申报数据出境安全评估、与境外接收方签订标准合同、实施个人信息保护认证等合规路径，确保数据跨境流动的安全与合法合规；
b) 处理模型使用者的输入信息和使用记录等个人信息时，应当明确并向使用者告知处理目的、处理方式及保存期限等，在必要的范围内基于明确、合理的目的，以对使用者权益影响最小的方式、期限进行个人信息处理及保存，不得过度收集使用者的个人信息；
c) 应在隐私政策中明确告知模型使用者将会收集其输入数据用以训练模型、优化服务、改进产品等，并取得其同意；
d) 应对模型应用过程可能涉及的个人信息进行系统性梳理，设置并公示个人信息主体权利响应机制，及时受理和处理个人信息主体的查阅、复制、更正、补充、删除等要求；
e) 应建立外部模型使用管控机制，对使用外部模型作出明确限制，如禁止未经许可将内部数据上传至外部模型；
f) 应提示避免输入敏感数据，可以通过用户协议、隐私政策或其他形式提示模型使用者在使用模型时避免输入敏感数据，在输入第三方数据时应当取得第三方的有效授权；
g) 应通过限制模型调用次数或访问速率，并对输出结果进行必要的模糊化或扰动处理，以防止使用者通过大规模查询或重复交互推测模型内部结构。上述措施可有效抵御模型反演攻击、属性推断等典型安全风险，提升模型的防护强度与安全韧性；
h) 模型推理过程应设置硬性约束，如设置单次请求最大文本长度和递归层数限制，强制终止超限进程，防止模型遭受资源消耗型攻击；
i) 应检测模型使用者的输入和生成内容，识别是否存在涉黄、暴恐、违禁等风险；
j) 应构建数据安全管控体系，加强数据全生命周期的安全防护能力，定期审查和更新安全策略，确保安全措施的有效性和时效性；
k) 应制定数据安全应急预案，加强风险监测，在发生数据安全事件时应当立即采取补救措施并向有关主管部门报告；
l) 宜采用模型水印技术，对模型嵌入可验证标识信息，用于模型溯源、版权保护及侵权检测。

12 模型更新

对人工智能数据处理者的基本安全要求如下：

a) 应动态更新语料库，及时补充时效性数据，避免因数据过时导致模型性能下降；
b) 人工智能服务提供者应为模型使用者提供拒绝或关闭其输入数据用于训练的方式，如为模型使用者提供选项或其他控制指令，且拒绝或关闭方式应方便快捷；
c) 应建立安全的模型更新机制，确保只有经过验证和授权的模型更新才能被部署；
d) 应制定在模型更新、升级时的安全管理策略，在模型完成重要更新或升级后，需再次自行组织模型评估；
e) 对第三方数据或技术供应商进行安全评估，应确保其数据处理符合规范，如在联合开发中采用隐私保护技术，如联邦学习、安全多方计算等，避免数据外泄；
f) 应建立模型版本管理制度，记录每次更新的修改内容，并对新版本进行严格验证，防止因更新引入安全漏洞。

附 录 A
（规范性）
人工智能数据常见安全风险

本附录给出了人工智能数据常见安全风险类别，如数据泄露风险、数据篡改风险、数据偏见歧视等，如表A.1所示，并提供了人工智能安全风险及典型威胁及具体含义的说明，如表A.2和A.3所示。

表A.1 人工智能数据安全风险类别

序号	风险类别	描述
1	数据泄露风险	人工智能研发应用过程中，因数据处理不当、非授权访问、恶意攻击、诱导交互等问题，可能导致数据和个人信息泄露
2	数据篡改风险	参数、结构、功能等算法核心信息，面临被逆向攻击窃取、修改，甚至嵌入后门的风险，可导致知识产权被侵犯、商业机密泄露，推理过程不可信、决策输出错误，甚至运行故障
3	违法违规利用数据	违反法律、行政法规等有关规定，非法或违规使用、加工、委托处理数据的风险
4	数据滥用风险	由于缺乏授权访问控制、权限管控等有效的安全管控措施、人员有意或无意操作等，导致数据被未授权或超出授权范围使用、加工的风险
5	数据伪造风险	由于数据源欺骗、深度伪造等安全威胁，或者缺乏有效的安全措施、人员有意或无意操作等，导致数据或数据源被伪造、数据主体被仿冒等安全风险
6	输出不可靠	人工智能可能产生“幻觉”，即生成看似合理，实则不符常理的内容，造成知识偏见与误导
7	数据投毒风险	训练数据被攻击者篡改、注入错误、误导数据的“投毒”风险，“污染”模型的概率分布，进而造成准确性、可信度下降
8	数据偏见歧视	人工智能算法设计及训练过程中，个人偏见被有意、无意引入，或者因训练数据集质量问题，导致算法设计目的、输出结果存在偏见或歧视，甚至输出存在民族、宗教、国别、地域等歧视性内容
9	数据违规采集	人工智能训练数据的获取，以及提供服务与用户交互过程中，存在未经同意收集、不当使用数据和个人信息的安全风险
10	数据标注不规范	训练数据标注过程中，存在标注规则不完备、标注人员能力不够、标注错误等问题，不仅会影响模型算法准确度、可靠性、有效性，还可能导致训练偏差、偏见歧视放大、泛化能力不足或输出错误的风险
11	模型窃取风险	通过向黑盒模型进行查询获取相应结果，获取相近的功能，或者模拟目标模型决策边界，从而窃取模型信息
12	模型反演风险	通过特殊设计的算法，重建目标模型的私有训练样本，进而造成敏感信息泄露的风险

表A.2 人工智能安全风险

风险分类	具体风险类别	具体内容
人工智能技术内生安全风险	模型算法安全风险	可解释性不足 偏见、歧视 鲁棒性不强 输出决策不可靠 外部对抗攻击 模型缺陷扩散
	数据安全风险	违规收集使用数据 训练数据内容不当 训练数据标注不规范 数据和个人信息泄露
人工智能技术应用安全风险	网络系统安全风险	组件和算力安全 网络暴露面扩大 供应链安全 网络攻击滥用
	信息内容安全风险	输出违法有害信息 混淆事实、误导用户 污染网络内容生态

表A. 3　人工智能典型威胁及具体含义

典型威胁	具体含义
投毒攻击	通过在训练数据中植入恶意样本或修改数据以欺骗机器学习模型的方法
成员推理攻击	针对一个预训练人工智能模型，攻击者尝试推断某输入样本数据是否属于该模型的训练集合（即训练集成员）
属性推理攻击	使用模型参数作为先验知识，通过训练攻击模型推测模型训练集的全局属性统计信息
模型反演攻击	又称模型逆向攻击，是一种针对机器学习模型的攻击手段，其核心目的在于通过利用模型的输出结果推断其训练数据或其他敏感信息
数据重建攻击	攻击者利用从模型更新中提取的信息来重构参与训练的原始数据
大模型提示词注入	绕过过滤器或使用精心制作的提示操作大语言模型，使模型忽略先前的指令或执行非计划的操作
逃逸攻击	通过构造对抗样本，即在原始输入中添加人类难以察觉的微小扰动（如噪声、纹理修改），使机器学习模型在测试阶段产生错误分类或检测失效，从而绕过模型的决策逻辑
对抗攻击	对目标机器学习模型的原输入施加轻微扰动以生成对抗样本来欺骗目标模型的过程
后门攻击	攻击者在模型的训练过程中，以某种方式在训练数据中嵌入特定的触发模式，使得模型在接收到这种触发模式时表现出特定的错误行为，而在其他情况下，模型仍然正常工作
模型窃取攻击	一类窃取模型信息的恶意行为，攻击者通过向黑盒模型进行查询获取相应结果及相近的功能，或者模拟目标模型决策边界
大模型越狱	针对大语言模型的攻击手段，其核心目的是通过精心设计的输入或提示，绕过模型的安全限制，诱导模型生成不符合伦理、有害或违法的内容
深度伪造	“生成式对抗网络”的机器学习模型将图片或视频合并叠加到源图片或视频上，借助神经网络技术进行大样本学习，将个人的声音、面部表情及身体动作拼接合成虚假内容的人工智能技术
人工智能驱动的网络攻击	利用人工智能技术自动化执行网络攻击，能够提高攻击效率、降低成本，并且使攻击更加隐蔽和高效
大模型幻觉	大语言模型在生成文本时，可能会产生与事实不符、不合逻辑或完全虚构的内容
供应链攻击	攻击者通过破坏供应链中的某个环节，如开发工具、代码库、更新服务器或第三方服务，来向最终用户传播恶意代码或篡改模型
开源漏洞攻击	开发者在模型部署过程中，使用了包含漏洞的开源库框架，导致系统暴露于安全风险中，模型运行时，攻击者利用漏洞执行未授权代码，窃取敏感数据或破坏模型功能
大模型偏见	大语言模型在语言生成或决策中，因训练数据中固有的性别、种族、文化、社会等因素的不平衡而表现出的系统性偏见
智能体后门攻击	攻击者通过特定手段在基于大语言模型的智能体中植入后门的隐蔽恶意机制。该机制使得智能体在正常输入下表现符合预期，但一旦输入满足预设的触发规则，其多步中间推理过程将被操控，导致智能体生成攻击者设定的恶意输出，甚至导致进一步的衍生攻击

附 录 B
（规范性）
语料及生成内容的主要安全风险

B.1 包含违反社会主义核心价值观的内容

主要风险包括：
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。

B.2 包含歧视性内容

主要风险包括：
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。

B.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。

B.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。

B.5 无法满足特定服务类型的安全需求

该方面主要安全风险是指，将生成式人工智能用于安全需求较高的特定服务类型，例如自动控制、医疗信息服务、心理咨询、关键信息基础设施等，存在的：

a） 内容不准确，严重不符合科学常识或主流认知；

b） 内容不可靠，虽然不包含严重错误的内容，但无法对使用者形成帮助。

参 考 文 献

[1] GB/T 32400—2015 信息技术 云计算 概览与词汇

[2] GB/T 41867—2022 信息技术 人工智能 术语

[3] GB/T 42888—2023 信息安全技术 机器学习算法安全评估规范

[4] 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

[5] DB11/T 2251—2024 信息安全 人工智能数据安全通用要求

[6] ISO/IEC 23053:2022 Information technology Security techniques Guidelines for the implementation of the ISO/IEC 27002 code of practice

[7] 《中华人民共和国数据安全法》（2021年6月10日中华人民共和国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）

[8] 《人工智能生成合成内容标识办法》（2025年3月7日中华人民共和国国家互联网信息办公室印发）

[9] 《人工智能数据安全白皮书（2019年）》（2019年8月9日中国信息通信研究院安全研究所在中国人工智能高峰论坛“人工智能与大数据安全分论坛”发布）

[10] 《人工智能数据安全风险与治理》（2019年8月30日赛博研究院和上海观安信息技术股份有限公司在世界人工智能大会上共同发布）

[11] 《大模型治理蓝皮报告——从规则走向实践（2023年）》（2023年11月24日中国信息通信研究院（CAICT）、中国科学院计算技术研究所在2023江西国际移动物联网博览会发布）

[12] 《人工智能安全标准化白皮书（2023版）》（2023年5月29日全国信息安全标准化技术委员会（TC260）大数据安全标准特别工作组在全国信安标委2023年第一次标准周“人工智能安全与标准研讨会”发布）

[13] 《人工智能安全治理框架2.0版》（2025年9月15日在2025年国家网络安全宣传周主论坛上发布）

[14] 《人工智能安全标准体系1.0版（征求意见稿）》（2025年2月14日全国网络安全标准化技术委员会通过全国网安标委官网及合作媒体公开征求意见）

[15] 《促进和规范数据跨境流动规定》（2024年3月22日国家互联网信息办公室2023年第26次室务会议审议通过）

四 川 省 网 络 空 间 安 全 协 会

团 体 标 准

人工智能数据安全要求

T/CSAS 0029—2025

*

中国轻工业出版社出版

地址：北京鲁谷东街 5 号

邮政编码：100040

发行电话：(010)85119832

网址：http://www.chlip.com.cn

Email：club@chlip.com.cn

*

书号：155019·7160

印数：1—200 册　　定价：48.00 元