ICS 03.100.01
CCS A 02

团　　体　　标　　准

T/ISC 0080—2025

人工智能通用大模型合规管理体系指南

Guidelines for compliance management for general-purpose foundational model of artificial intelligence

（发布稿）

2025-7-18 发布　　2025-8-18 实施

中　国　互　联　网　协　会　　发　布

目　　次

前　言

本标准按照GB/T 1.1-2020给出的规则起草。

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这些专利的责任。

本标准由中国互联网协会归口。

本标准主要起草单位：中国信息通信研究院、蚂蚁科技集团股份有限公司、中国移动通信有限公司研究院、中国移动通信集团设计院有限公司、度小满科技（北京）有限公司、北京蜜莱坞网络科技有限公司、北京新氧科技有限公司、上海倍孜网络技术有限公司、北京美数信息科技有限公司、中国软件评测中心（工业和信息化部软件与集成电路促进中心）、内蒙古电子文件自主可控技术协会。

本标准主要起草人：张夕夜、陈慧、郭家书、程晓蕾、杨妮、施克游、赵金生、郝克勤、葛思凡、姜丽丽、邢璟、郝乙入、李鹏、夏晓晖、浦洋、金星、应晓冬、庞成军、张紫阳、聂子尧、江志鹏、郭麒、白海建、徐忠旺、高宏玲、翟腾、艾文思、刘亚洲、郭敏、吕律。

引　言

2022 年末，以 ChatGPT 为代表的大规模预训练语言模型引发各界广泛关注，成为新一轮人工智能技术应用爆发的催化剂，并由此带动人工智能通用大模型的算法创新及关键技术研究进入加速期。

通用大模型作为人工智能应用发展的核心引擎，凭借其优秀的通用性、泛化性及技术赋能特性，正渐进成为人工智能行业的新型基础设施，为经济发展与产业转型注入新动能。然而，随着通用大模型驱动的人工智能对社会结构、产业生态以及人类生活方式的影响逐步扩大，其在数据安全、可解释性、隐私保护、知识产权归属、内容安全、责任归属等方面的治理风险也在逐渐显露。

通用大模型规范发展是人工智能技术广泛应用的重要前提，从数据、算法、应用等层面加强人工智能通用大模型的合规管理是发展数字经济的应有之义。为促进组织以负责任的方式开发、提供或应用人工智能通用大模型，助力组织提高大模型合规管理水平，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》、《中华人民共和国科学技术进步法》、《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》、《人工智能生成合成内容标识办法》、GB/T 35770-2022《合规管理体系 要求及使用指南》、GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》等法律、行政法规、标准要求，制定本文件。

本文件规定了在组织内开展、支持和持续改进人工智能通用大模型合规管理工作的要求并提供指导，组织应将其合规管理的重点放在通用大模型的某些特征之上，例如数据质量、隐私保护、模型可解释性和可扩展性等。如果通用大模型与传统的人工智能技术相比引发了额外的安全风险，组织可在本文件的基础之上采取不同的保护措施。

人工智能通用大模型合规管理体系 指南

1 范围

本文件规定了在组织内开展、支持和持续改进人工智能通用大模型合规管理工作的要求并提供指导。

本文件适用于下列主体：

a) 利用通用大模型技术向公众提供生成文本、图片、音频、视频等产品或服务；

b) 正在开发或委托第三方开发通用大模型。

本文件适用于下列范围：

a) 训练/部署大模型的云平台的安全；

b) 数据处理全生命周期管理；

c) 模型安全（含备案、评估、内容安全）；

d) 服务规范。

本文件适用于下列情形：

a) 计划提升通用大模型合规管理水平；

b) 寻求外部组织对其通用大模型的合规性进行评价。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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 35770-2022/ISO 37301:2021 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 43782-2024 人工智能 机器学习系统技术要求

T/ISC 0023-2023 信息通信及互联网行业企业合规管理体系 指南

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

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, 3.1.2

3.2

人工智能系统 artificial intelligence system

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, 3.1.8]

3.3

通用大模型 general-purpose foundational model

在大规模数据集上进行训练，可以适应广泛下游任务的人工智能模型。

3.4

合规管理 compliance management

以有效防控合规风险为目的，开展包括体系构建、制度制定、风险识别、合规审查、风险应对、责任追究、考核评价、合规培训等有组织、有计划的管理活动。

[来源：T/ISC 0023-2023, 3.3]

3.5

通用大模型合规 general-purpose foundational model compliance

履行组织全部的人工智能通用大模型合规义务(3.6)。

3.6

通用大模型合规义务 compliance obligations

组织必须强制性遵守的涉及人工智能通用大模型的要求，以及组织自愿选择遵守的涉及人工智能通用大模型的要求。

3.7

通用大模型合规风险 compliance risk

因未遵守人工智能通用大模型合规义务而发生不合规事件的可能性及其后果。

3.8

训练数据 training data

用于训练机器学习模型的输入数据样本子集。

[来源：GB/T 41867-2022, 3.2.34]

3.9

数据标注 data labeling

给数据样本指定目标变量和赋值的过程。

[来源：GB/T 41867-2022, 3.2.29]

3.10

最高管理者 top management

在最高层指挥和控制组织的一个人或一组人。

[来源：GB/T 35770-2022/ISO 37301:2021, 3.3]

3.11

治理机构 governing body

对组织的活动、治理、方针负有最终职责和权限的一个人或一组人，最高管理者向其报告并对其负责。

[来源：GB/T 35770-2022/ISO 37301:2021, 3.21]

4 组织环境

4.1 通用要求

组织应确定对其开发、提供或应用的人工智能通用大模型进行合规管理的预期目标。

组织应了解与其通用大模型合规管理预期目标相关并影响其合规管理能力的内外部环境。

4.2 内部环境

与内部环境相关的考虑因素包括：

a) 组织内部与通用大模型合规管理相关的组织结构、规章制度、程序流程、技术水平、组织文化等；

b) 组织进行通用大模型合规管理的预期目标；

c) 合同义务。

4.3 外部环境

与外部环境相关的考虑因素包括：

a) 适用的法律、行政法规、部门规章及其他规范性文件的要求；

b) 监管机构对人工智能通用大模型开发和应用的司法解释或对其有影响的政策、指南和决定；

c) 与人工智能通用大模型开发和应用相关的行业惯例、技术伦理、道德规范等；

d) 人工智能通用大模型产品和服务的市场竞争格局和趋势。

4.4 合规管理工作的范围

组织应确定人工智能通用大模型合规管理工作的边界和适用性，以确定其范围。在确定该范围时，组织应考虑：

a) 4.1、4.2、4.3中提及的外部和内部环境；

b) 人工智能通用大模型合规管理工作的地理和/或组织边界；

c) 第三方对组织人工智能通用大模型合规性的要求。

该范围应以文件化信息的形式确定。

人工智能通用大模型合规管理工作的范围应确定组织根据本文件对合规管理的要求进行的职责划分、计划、支持、改进。

5 机构职责

5.1 治理机构和最高管理者

组织治理机构和最高管理者应通过以下方式展示对人工智能通用大模型合规管理工作的领导力和承诺：

a) 支持管理者制定人工智能通用大模型合规管理规章制度和合规管理预期目标，确保规章制度、预期目标与组织的人工智能技术发展战略方向相一致；

b) 支持管理者将通用大模型合规管理要求整合到组织的管理体系及业务流程中；

c) 确保通用大模型合规管理工作所需的资源充足；

d) 充分传达通用大模型合规和符合技术伦理要求的重要性；

e) 对持续改进通用大模型合规管理工作的承诺。

5.2 管理者

组织管理者应向治理机构和最高管理者汇报人工智能通用大模型合规工作计划、进展及风险，具体承办治理机构和最高管理者的合规职责，并督促合规团队的工作：

a) 制定适合组织的合规管理目标及工作计划；

b) 提出遵守通用大模型合规要求的规章制度；

c) 确立嵌入合规义务的业务流程；

d) 明确不履行合规义务的后果；

e) 促进通用大模型合规管理工作的持续改进。

该规章制度应以文件化信息的形式提供。在制定过程中可参考其他组织通用大模型合规管理规章制度相关内容，在组织内部就规章制度内容进行沟通，可酌情向外部相关方提供。

5.3 合规团队

组织宜设置合规团队，负责配合组织内部相关部门，参与组织通用大模型合规计划的设计、执行和监督工作。合规团队主要负责：

a) 具体起草本组织人工智能通用大模型合规工作计划和规章制度；
b) 组织开展或者参与风险评估、风险处理、合规检查与培训等相关工作，及时发现合规管理薄弱环节，督促违规整改和持续改进；
c) 落实本组织通用大模型合规宣传计划，定期或不定期组织或协助人事部门、业务部门开展合规培训、宣传等工作，为各业务单位提供合规咨询和支持；
d) 代表组织对接履行人工智能监管职责的部门。

5.4 业务部门

业务部门负责本领域的日常通用大模型合规管理工作。业务部门主要负责：

a) 按照通用大模型合规要求完善业务管理制度和流程，履行相应的合规义务；
b) 组织或配合合规专员进行合规风险识别和隐患排查，及时向人工智能管理部门通报风险事项，妥善应对合规风险事件；
c) 组织或配合违规调查及整改工作。

5.5 其他人员

组织所有人员应：

a) 遵守组织人工智能通用大模型合规管理的规章制度、规划及流程；
b) 报告合规疑虑、问题和漏洞；
c) 根据要求参加合规培训。

6 合规计划

6.1 概述

组织在制定人工智能通用大模型合规计划时，应考虑4.1中提到的预期目标和4.2中提到的工作范围，并确定需要履行的合规义务和可能面临的合规风险：

——保证组织通用大模型合规管理能够实现其预期目标；
——避免或减少因不合规事件而产生的不良影响；
——实现持续改进。

组织应建立并维护通用大模型风险标准，以支持：

——区分可接受和不可接受的风险；
——进行通用大模型风险评估；
——进行通用大模型风险处理；
——评价通用大模型风险处理对策。

组织应计划：

——履行义务和应对风险的措施；
——如何将这些措施整合并嵌入其人工智能管理的流程中；

——如何评估措施的有效性。

6.2 义务识别

6.2.1 明确合规义务来源

明确合规义务来源是组织开展、支持、改进合规管理工作的基础。通常，人工智能通用大模型合规义务来源于两个方面，包括组织必须要遵守的要求和自愿选择遵守的承诺。

组织必须要遵守的人工智能通用大模型合规要求包括：

——法律法规；

——监管机构发布的命令、条例或者指南；

——行政决定；

——法院判决；

——强制性标准；

——条约、公约和协议。

组织自愿选择遵守的人工智能通用大模型合规承诺包括：

——与社会团体等非政府组织签订的协议；

——与客户和公共权力机构签订的协议；

——相关产业的标准；

——组织内部制度、公开承诺等。

6.2.2 合规义务梳理及更新

组织宜系统梳理来源于其活动、产品和服务的人工智能通用大模型合规义务。

组织宜定期识别新增及变更的人工智能通用大模型合规义务，确保持续合规。

组织宜评价变更的人工智能通用大模型合规义务对合规管理工作产生的影响，并对合规管理工作计划进行必要地调整。

组织人工智能通用大模型合规义务清单应符合附录A的规定。

6.3 风险评估

组织应确定并建立人工智能通用大模型的风险评估周期与流程，该流程应：

a) 遵循组织通用大模型合规管理规章制度和预期目标并与之保持一致；

b) 评估通用大模型对组织、个人和社会可能造成的风险，如果已识别的风险成为现实，将会产生什么结果；

c) 确定风险级别，判断属于可接受还是不可接受的风险；

d) 对评估的风险进行优先级排序，优先处理风险级别较高的风险；

e) 多次重复的通用大模型风险评估宜生成客观有效、可比较的结果；

f) 保留有关通用大模型风险评估过程的文件化信息。

6.4 风险处理

基于风险评估的结果，组织应确定并建立风险处理流程，该流程应：

a) 组织应制定通用大模型风险处理计划并验证其有效性，该计划应明确不同风险级别下，组织应采取的通用大模型风险处理方案，同时制定风险应急处置方案。

b) 确定实施所选择的风险处理方案需要的所有控制措施，形成相应的操作文档，便于处理者验证没有遗漏任何必要的控制措施；

c) 当风险处理计划定义的通用大模型风险处理方案无效时，组织应按照风险处理流程对处理方案进行审查和重新验证，并更新风险处理方案。
d) 在风险事件对个人、组织造成实质性危害的情况下，宜及时以口头及书面方式向所涉主体告知事件情况、危害后果、已采取的补救措施等信息。无法逐一告知的，可采取公告方式告知；
e) 保留有关通用大模型风险处理过程的文件化信息。

6.5 计划变更

当组织因内外部环境、合规义务或风险评估结果改变，而确定需要对通用大模型合规计划进行调整时，宜通过正式的变更申请程序来进行调整。

7 支持

7.1 资源

组织应确定并提供开展、评价和持续改进通用大模型合规管理工作所需的资源。

7.2 能力

组织应：
a) 确定在其控制下从事通用大模型开发、部署、应用工作的人员具有必要的专业能力及合规意识；
b) 确保这些人员在拥有适当教育、培训或经验的基础上能够恰当地履行通用大模型合规义务；
c) 适当的文件化信息应作为能力证明予以保留。

7.3 合规培训

组织宜定期对有关人员进行培训，培训应：
a) 适合于人员的岗位及其面临的通用大模型合规风险；
b) 确保培训内容、培训方式、培训对象、培训频率等满足岗位人员履行合规义务的需要；
c) 定期举行；
d) 培训记录应作为文件化信息予以保留。

7.4 沟通及协作

组织宜建立并畅通与通用大模型合规管理工作相关的沟通渠道，制定合规规章制度时，可成立临时的跨部门规章制度制定小组，由不同部门的领导或代表组成，以确保合规规章制度的全面性和适用性；对于复杂或专业性强且存在重大合规风险的事项，可以向合规专员咨询，必要时及时向管理层汇报；面对合规专员的合规检查或调查，各部门宜积极沟通并予以配合。

7.5 合规咨询

组织宜建立通用大模型合规咨询机制，管理层和各部门员工在工作中可以向合规专员咨询通用大模型合规问题。合规专员应当不断学习、提升通用大模型合规管理水平，也可以同外部专业机构开展通用大模型合规咨询合作。

8 保障措施

8.1 采取安全技术措施

组织应当采取与所处理参数规模、类型相适应的安全技术措施，安全技术措施的有效性可以参考以下因素：

a） 是否采取数据安全与隐私保护措施，例如：将训练环境与推理环境隔离、对数据访问进行权限控制、使用匿名化和加密等技术保护用户隐私等；

b） 是否采取内容安全、准确、可靠措施，例如对使用者输入信息进行安全性检测、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等；

c） 是否定期对所使用的开发框架、代码等进行安全审计，识别和修复潜在的安全漏洞；

d） 是否采取对抗性攻击防护、模型安全审计等安全技术措施，增强抗系统性风险能力；

e） 是否建立数据、模型、框架、工具等的备份机制以及恢复策略，重点确保业务连续性。

8.2 影响评估

组织宜按计划的时间节点、时间间隔或在内外部环境发生重大变化、拟发生重大规章制度变更时进行通用大模型影响评估：

a） 影响评估应确定通用大模型的部署、预期用途和可预见地滥用对组织、个人或社会的潜在影响；

b） 影响评估应考虑通用大模型部署的具体技术和社会背景以及适用的司法管辖区；

c） 组织宜在风险评估中考虑通用大模型影响评估的结果；

d） 人工智能系统影响评估的结果应记录在案。在适当的情况下，影响评估的结果可以提供给组织定义的利益相关方。

8.3 保留文件化信息

组织宜以适当的形式和载体记录通用大模型合规管理工作产生的文件化信息，包括合规风险评估、处理措施、调查过程、审核记录，可作为证据。文件化信息应当以清晰、易读和易检索的方式保存，并采取必要措施防止泄密、不当使用或完整性受损。

9 改进

9.1 持续改进

组织宜持续改进通用大模型合规管理的适宜性、充分性和有效性。

9.2 不合规和纠正措施

当发生不合规事件时，组织宜：

a） 及时对不合规事件做出反应，并采取行动加以控制和纠正；

b） 评估是否需要采取额外行动消除不合规的原因，以使其不会在其他地方再次发生或发生：评估内容宜包括：

——不符合项的审查；

——确定不合规事件的原因；

——确定是否存在或可能发生类似的事件；

c） 审查所采取的任何纠正措施的有效性；

d） 如有必要，对组织的合规计划进行更改。

以下证据应形成文件化信息：

a） 不合格的性质以及随后采取的任何措施；

b） 任何纠正措施的结果。

附 录 A

（规范性）

人工智能通用大模型合规义务清单

表A.1规定了人工智能通用大模型合规义务清单，组织宜根据产品及服务类型等识别自身合规义务。

表A.1 人工智能通用大模型合规义务

序号	合规义务（一级）	合规义务（二级）		是否适用	备注
1	系统安全	确保系统建设和安全保障建设同步规划、同步建设、同步使用			
2		定期开展信息安全风险评估			
3		定期开展网络安全等级保护测评			
4		对于人工智能系统采用的芯片、软件、工具、算力和数据资源，应高度关注供应链安全，跟踪软硬件产品漏洞、缺陷信息并及时采取修补加固措施			
5		在规定时间内完成网络安全事件应急响应			
6	云平台安全	通过云平台进行大模型训练及部署时，选择安全可靠的云平台，避免云平台引发的网络安全风险及数据安全风险			
7		在组织内配备发现云平台安全风险的管理和技术手段			
8		通过合作协议等方式明确云平台提供方的安全合规义务与责任			
9	模型安全	依据《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》等相关规定开展通用大模型安全评估			
10		按照国家有关规定开展安全评估并按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续。			
11		提升算法可解释性、可预测性，为人工智能系统内部构造、推理逻辑、技术接口、输出结果提供明确说明			
12		在设计、研发、部署、维护过程中建立并实施安全开发规范，尽可能消除模型算法存在的安全缺陷、歧视性倾向，提高鲁棒性			
13		生成内容安全	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		
14			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		
15			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		
16		生成内容准确	采取技术借施提高生成内容响应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		
17		生成内容可靠	采取技术借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		
18	数据处理全生命周期管理	数据来源合法	以合法、正当的方式取得训练数据		
19			无损数据权益人的权益		
20			不侵害他人依法享有的知识产权		
21			处理个人信息时取得了个人信息主体同意或者获得其他合法性基础		
22			利用爬虫技术获得训练数据，爬取符合行业惯例，包括尊重Robots协议、控制爬取频率与资源占用等		
23			对个人信息采用必要的数据脱敏措施		
24		提高数据质量	通过交叉验证、第三方验证等方式，验证数据的真实性和准确性		
25			删除或纠正数据中的错误和异常值，确保数据的一致性和准确性		
26			定期更新训练数据集，以反映最新的变化和趋势，确保数据的时效性和客观性		
27			整合多个数据源，提高数据的完整性和多样性		

28			记录数据的来源、处理历史等信息，以便追踪数据质量问题，并确保数据的可靠性和可追溯性		
29		合法使用数据	未经个人同意或法律另有规定，不应使用其个人信息开展算法训练等相关活动		
30			数据用于训练前，对数据中的主要知识产权侵权风险进行识别		
31			进行个人信息保护影响评估		
32			向个人告知了接收方的名称或者姓名、联系方式、处理目的、处理方式和个人信息的种类		
33			采取措施防止未经授权的数据访问、篡改、替换或破坏		
34		规范数据标注	对于标注人员进行安全合规培训，培训内容应包括标注任务规则、标注工具使用方法、标注内容合规核验方法、标注数据合规管理要求等		
35			制定清晰、具体、可操作的标注规则，至少包括标注目标、数据格式、标注方法、质量指标等内容		
36			开展数据标注质量评估，抽样核验标注内容的准确性。对每一批标注数据进行人工抽检，发现内容不准确的，应重新标注；发现内容中包含违法不良信息高于5%的，该批次标注数据应作废		
37			宜对安全性标注数据进行隔离存储		
38	安全评估	建设关键词库、分类模型，对语料安全情况进行评估，TC260-003 生成式人工智能服务安全基本要求中给出了进一步的说明			
39		建设生成内容测试题库，对生成内容安全情况进行评估，TC260-003 生成式人工智能服务安全基本要求中给出了进一步的说明			
40		建设拒答测试题库，对问题拒答情况进行评估，TC260-003 生成式人工智能服务安全基本要求中给出了进一步的说明			
41	服务规范	明确并公开其服务的适用人群、场合、用途，指导使用者科学理性认识和依法使用通用大模型			

42		不收集非必要个人信息，不非法留存能够识别使用者身份的输入信息和使用记录，不非法向他人提供使用者的输入信息和使用记录		
43		依法及时受理和处理个人关于查阅、复制、更正、补充、删除其个人信息等的请求		
44		依据《人工智能生成合成内容标识办法》等相关规定对人工智能生成内容进行显性或隐形标识		
45		建立健全投诉、举报机制，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		
46		服务如果适用于未成年人的，应设置未成年人保护措施，积极展示有益未成年人身心健康的内容		
47		发现违法内容时，及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告		
48		发现使用者利用通用大模型从事违法活动的，依法依约采取警示、限制功能、暂停或者终止向其提供服务等处置措施，保存有关记录，并向有关主管部门报告		
49	配合监督	积极配合有关主管部门对通用大模型开展的监督检查，按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明，并提供必要的技术、数据等支持和协助		

参 考 文 献

[1] GB/T 35770-2022 合规管理体系 要求及使用指南
[2] GB/T 45288.1-2025 人工智能 大模型 第1部分：通用要求
[3] GB/T 45288.2-2025 人工智能 大模型 第2部分：评测指标与方法
[4] GB/T 45288.3-2025 人工智能大规模 第3部分：服务能力成熟度评估
[5] GB 45438-2025 网络安全技术 人工智能生成合成内容标识方法
[6] TC260-003 生成式人工智能服务安全基本要求
[7] 中华人民共和国网络安全法
[8] 中华人民共和国数据安全法
[9] 中华人民共和国个人信息保护法
[10] 中华人民共和国科学技术进步法
[11] 互联网信息服务算法推荐管理规定
[12] 互联网信息服务深度合成管理规定
[13] 生成式人工智能服务管理暂行办法
[14] 人工智能生成合成内容标识办法
[15] 人工智能安全治理框架