ICS 35.240.99
CCS L70

团　　体　　标　　准

T/ISC 0104—2026

安全应急大模型

Safety emergency large-scale model

（发布稿）

2026－05－11 发布　　2026－06－11 实施

中 国 互 联 网 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。

本文件由中国互联网协会提出。

本文件由中国互联网协会归口。

本文件起草单位：北京广监云科技有限公司、中国信息通信研究院、北京市科学技术研究院、船舶信息研究中心（中国船舶集团有限公司第七一四研究所）、北京中应安赫科技有限公司、北京华夏安科信息技术有限公司、联通数字科技有限公司、大家保险集团有限责任公司、防灾科技学院、华易数安科技（吉林省）有限公司、南方电网互联网服务有限公司、奇安信科技集团股份有限公司、亚信安全科技股份有限公司、南京赛宁信息技术有限公司、北京国科云计算技术有限公司、常州工业职业技术学院、华能信息技术有限公司。

本文件主要起草人：侯卓林、蒋阿芳、马英轩、陈丽娜、王亚飞、秦绪坤、张英香、高竞秀、丛磊、崔鹏飞、罗华伟、谷春野、吴燕雄、王冉、夏武、张源、刘岩、杨婷、颜科、邹立刚、郝亚平、潘中英、张涛。

引　言

随着数字化转型的加速推进，安全应急领域面临着日益复杂多变的挑战。传统的安全防护手段已难以应对新型安全威胁，而大模型技术的出现为提升安全应急能力带来了新的机遇。然而，当前通用大模型在安全应急领域的应用存在"通而不专"的局限性，难以满足专业场景的需求。为解决这一问题，构建《安全应急大模型》标准显得尤为必要。

本标准旨在规范安全应急领域大模型的研发、部署及应用，推动其专业化、规范化发展。通过明确技术框架、强化核心能力、规范应用场景及建立严格的评估体系，本标准致力于提升模型在安全应急场景中的精准性与适配性，有效防范数据泄露、模型滥用等安全风险，促进产业协同与融合发展，为各行业提供可信赖的智能化安全保障。标准期望解决的核心问题包括：提升通用大模型在安全应急领域的适配性，应对数据安全与隐私保护的挑战，规范模型的应用边界与伦理准则，以及通过建立统一的评测标准，解决当前技术参差不齐、缺乏统一验证的问题，促进产业健康发展。

通过本标准的制定与实施，旨在推动安全应急大模型技术的高质量发展，为应对数字化转型中的安全挑战提供有力保障，助力公共安全与应急管理的智能化升级。

安全应急大模型

1 范围

本文件规定了安全应急大模型的技术框架、安全能力要求、应用场景及评估评测体系，适用于安全应急领域大模型的研发、部署及应用。

本文件适用于从事安全应急大模型研发、部署、运营的企业和机构，以及网络安全、数据保护、应急处置等安全应急领域的相关业务场景。

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 35273-2020 信息安全技术 个人信息安全规范

GB/T 41867-2022 信息技术 人工智能 术语

GB/T 45288.1-2025 人工智能 大模型 第1部分：通用要求

3 术语和定义

GB/T 25069-2022、GB/T 41867-2022、GB/T 45288.1-2025和GB/T 35273-2020界定的以及下列术语和定义适用于本文件。

3.1

安全应急大模型 safety emergency large-scale model

在通用大模型的基础上针对安全应急领域优化训练的人工智能模型。

注：覆盖监测预警、风险评估、事件研判、态势分析等典型安全应急场景，为政府、企业等用户提供可信赖的技术支撑，提升公共安全与应急管理的智能化水平。

3.2

人工智能 artificial intelligence

针对从人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出工程系统而进行的研究和开发。

3.3

大模型 large-scale model

基于大量数据训练得到，具有复杂计算架构，能处理复杂任务，且具备一定泛化性的深度学习模型。

注：大模型的参数量由其功能和模态决定，一般不低于 1 亿。大模型训练使用的数据总量受参数量

的影响，达到收敛的大模型的参数量的对数与其训练数据总量的对数成正比。

4 基本要求

4.1 技术框架

模型采用分层架构设计，包括基础设施层、通用安全能力层和安全原子能力层，技术框架示意图见图1：

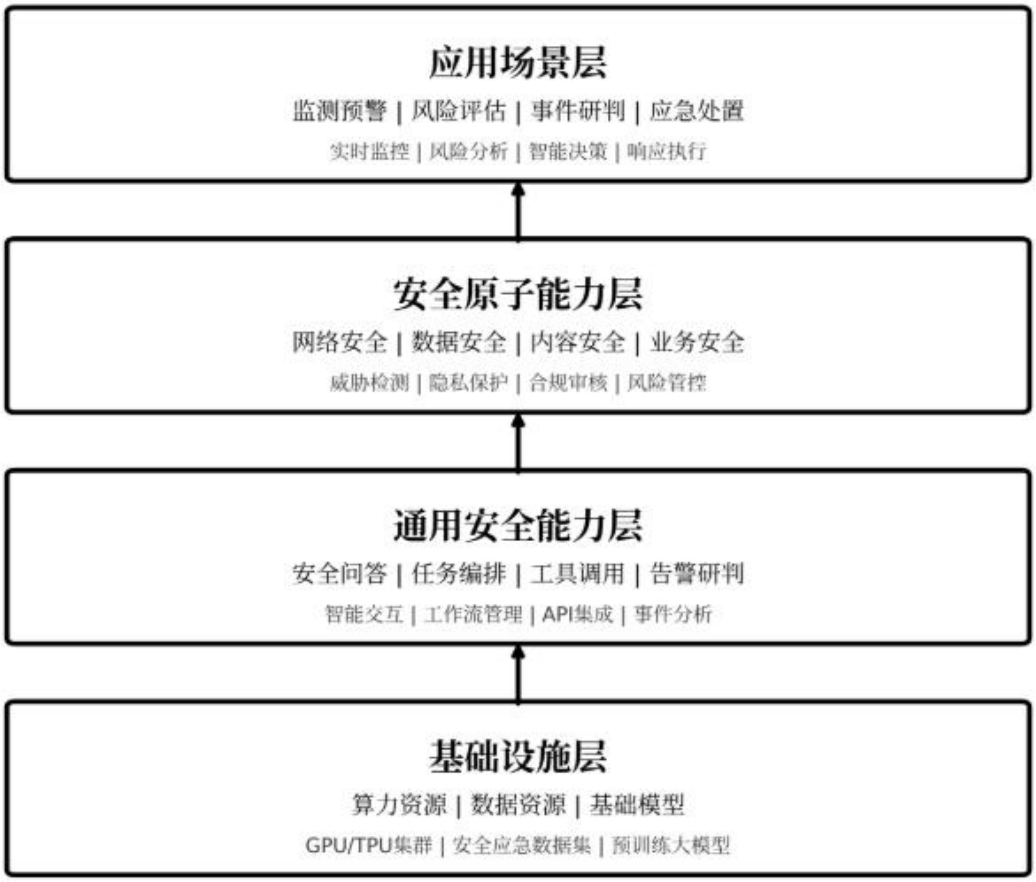

图1 安全应急大模型技术架构示意图

a) 模型采用分层架构设计，包括基础设施层、通用安全能力层和安全原子能力层。
b) 基础设施层包含算力、数据、模型等。
c) 通用安全能力层包含安全问答、任务编排、工具调用、告警研判等。
d) 安全原子能力层包含网络安全、数据安全、内容安全、业务安全等专项能力等。

4.2 场景和功能

场景和功能包含：

a）模型应用场景覆盖监测预警、风险评估、事件研判、应急处置等核心场景。

b）模型能够从视频监控、传感器等多源数据中自动识别安全隐患（如危化企业人员聚集、城市内涝风险），辅助生成应急处置方案，并支持安全报告自动生成与跨系统协同联动。

c）针对矿山、城市等专项领域，能开发专项功能模块。

4.3 信息安全

信息安全的内容包含但不限于：

a）基础网络安全：模型应具备恶意流量检测、入侵识别、漏洞挖掘、威胁狩猎等原子能力。

b）数据安全：规范数据采集、存储、处理的安全标准，确保敏感数据的分级保护与合规应用。

4.4 稳定可靠

确保可靠性应开展：

a）应建立科学的评估指标，包括模型准确性、响应时效性、数据合规性等。

b）模型在部署前通过第三方安全测试，确保其在应急响应中的可靠性与稳定性。

5 数据要求

5.1 数据收集

5.1.1 合法性

从公开渠道获取数据时，应确认数据发布方拥有合法的授权与发布权限。

收集个人信息应依据《通用数据保护条例》（GDPR）以及《个人信息保护法》等法律法规，通过清晰、易懂的告知形式，获得个人明确且自愿的授权。

5.1.2 代表性和完整性

收集的数据应具有代表性与完整性，全面反映各类自然灾害应急场景，为模型提供准确且丰富的信息输入。

注：以自然灾害应急为例，地震数据不仅要涵盖震级、震源深度等基础信息，还需包含地震发生区域的地质构造信息，因为不同地质条件下地震造成的破坏程度差异显著。洪水数据则要囊括流域面积、河流水文特征、历史洪水水位及淹没范围等内容。对于台风，除了台风强度、路径信息外，登陆地区的地形地貌（如是否有山脉阻挡）、沿海岸线特征等数据同样不可或缺。

5.2 数据存储

5.2.1 加密存储

应采取以下措施进行加密：

a）采取加密存储的手段防止数据在存储过程中被窃取或篡改。

b）如采用高级加密标准（AES）算法，可根据数据的敏感程度选择128位、192位和256位等不同的密钥长度。

c）对于涉及国家安全或重大安全应急决策的核心数据，应采用256位密钥长度的AES加密算法，以提供更高的加密强度。

d）加密存储不仅要对数据本身加密，还应对存储的元数据（如数据名称、创建时间、访问权限等）进行加密处理，确保整个数据存储体系的保密性。

5.2.2 备份和恢复

备份和恢复措施包括：

a）每周进行一次全量数据备份，能够完整保留当前数据状态，为可能出现的数据灾难提供最全面的恢复基础。

b）每天进行增量备份，则可记录自上次全量备份或增量备份以来数据的变化情况，大大减少备份数据量和备份时间。

c）备份数据应存储在异地安全的数据中心。

d）制定详细的数据恢复计划，明确在不同数据丢失或损坏场景下的恢复流程与责任人。

e）若发生硬件故障导致数据丢失，应在规定时间内（如 4 小时）启动从异地备份数据中心恢复数据的操作，并通过预演确保恢复过程顺畅，数据完整性不受影响。

5.3 数据使用

5.3.1 设置数据访问权限

数据访问权限包含：

a）应设置严格的数据访问权限。

b）根据用户在模型使用过程中的角色和职责，进行细致的数据访问级别划分。可设置模型训练人员和应急响应操作人员：

——模型训练人员可访问大量的训练数据，包括各类历史安全应急事件数据、相关环境数据以及模型训练所需的基础数据等。

——应急响应操作人员在执行具体应急任务时，仅需获取与当前应急任务直接相关的数据。

注：比如，在一场火灾应急响应中，消防指挥人员只需获取火灾发生地点周边的建筑布局、消防设施分布以及实时火势等数据，避免因访问过多无关数据而增加数据泄露风险。

5.3.2 记录数据使用情况

记录数据使用情况包含：

a）详细记录数据使用情况用于审计和追溯。

b）应精确记录每次数据访问的时间，精确到秒级。

c）记录内容包含不限于：

——访问用户的身份信息，包括用户名、所属部门及角色等，明确数据操作主体。

——访问目的，如模型训练、应急决策支持等，便于分析数据使用的合理性。

——数据操作内容，例如读取数据、修改数据、删除数据等操作类型，以及具体的数据操作对象和范围。

d）对未经授权的数据访问尝试或异常频繁的数据访问行为，采取相应的防范措施。

6 算法要求

6.1 公平性

公平性包含：

a）在模型训练过程中，应进行数据预处理。

b）需通过数据清洗技术，去除与公平性无关但可能导致偏见的数据特征。

c）采用数据增强技术，对数据量较少的地区数据进行合理扩充，使模型在训练过程中能够平等对待不同地区的数据特征。

d）宜采用科学的公平性评估指标对算法进行验证：

——差异影响比（DI）通过比较不同群体在模型输出结果上的接受率或通过率，衡量算法对不同群体的差异影响程度。

——平均机会差异（AOD）则侧重于衡量不同群体在模型输出的有利结果概率上的差异。

注：设定合理的公平性阈值（如 DI 在 0.8 - 1.2 之间，AOD 在可接受的极小范围内），确保算法在不同群体上的表现差异处于可接受水平，保障安全应急决策在不同群体间的公平性。

6.2 可解释性

可解释性包含：

a）模型输出的结果应能追溯的数据特征和算法逻辑。

b）提供多样化的可视化工具和详细的解释性文档。

注：通过图形化界面展示模型的决策树结构，使应急管理人员能够直观看到不同数据特征在决策过程中的分支走向和权重分配。

c）编写详细的解释性文档。对模型的整体架构、算法原理、数据处理流程以及关键参数设置等进行说明，帮助用户全面深入地理解模型，提高模型的可解释性和透明度。

6.3 鲁棒性

鲁棒性包含：

a）模型应能够通过滤波算法或异常值检测与处理技术，识别并纠正这些噪声数据，保持对事件（如地震、火灾等）真实情况的准确判断。

b）对于缺失值数据，模型应具备合理的填充策略，如基于历史数据的统计特征进行填充或采用机器学习算法（如 K 近邻算法）根据相似数据点进行填充。

c）面对恶意攻击，模型应通过安全防护机制（如数据加密传输、访问权限验证等）及时发现并阻止攻击，或者在遭受攻击后仍能维持一定的性能水平，避免因攻击导致错误的应急决策

d）向模型输入对抗样本，测试模型的鲁棒性。根据测试结果，对算法进行针对性优化，如改进模型结构（采用更复杂的鲁棒神经网络架构）、调整参数设置（增强模型对异常数据的敏感度阈值）等。可采用以下方法进行鲁棒性测试：

——使用 FGSM（快速梯度符号法）等攻击方法生成对抗样本，测试模型能否准确识别并应对这些恶意输入。

——数据噪声注入测试则在正常输入数据中随机添加不同类型和程度的噪声，评估模型在噪声环境下的性能变化。

7 应用要求

7.1 模型部署

模型部署包含：

a）配置服务器应及时更新服务器操作系统的安全补丁，关闭不必要的网络端口，防止黑客利用系统漏洞进行入侵。

b）配置网络设备应合理设置防火墙规则，只允许授权的网络流量进出模型部署网络。

c）安装入侵检测系统（IDS），实时监测网络流量，一旦发现异常流量（如大量的端口扫描行为），立即发出警报并采取阻断措施。

d）对模型文件进行完整性校验和数字签名，防止模型被篡改。可使用哈希算法（如 SHA - 256）对模型文件进行校验，采用数字证书对模型进行签名认证。

7.2 模型运行

模型运行包含：

a）建立全面的模型运行监控机制，实时掌握模型状态。性能指标监测包括模型的推理延迟、准确率、召回率等。

b）详细记录和深入分析模型运行日志，发现潜在安全问题和性能瓶颈，记录内容包含但不限于：

——记录模型每次推理的输入数据，包括数据的来源、格式以及关键特征值，便于在出现问题时追溯输入数据是否存在异常；

——记录输出结果，包括预测的类别、风险等级等，以及与预期结果的对比情况；

——记录运行时间，精确到毫秒级，用于分析模型推理的效率。

c）应能通过对运行日志的定期分析，发现潜在的安全隐患（如异常的输入数据模式可能暗示着恶意攻击尝试），找出性能瓶颈（如某些特定类型的输入数据导致模型推理时间过长），优化模型算法或调整数据处理流程。

7.3 更新与维护

7.3.1 模型更新

模型更新包含：

a）模型更新应经过严格的测试和验证，测试内容包含单元测试、集成测试和性能测试。

b）单元测试是对模型更新涉及的各个功能模块进行单独测试，确保每个模块在新的代码逻辑下能够正常运行。

c）集成测试是将更新后的各个模块集成在一起，测试模型整体的功能完整性和稳定性。

d）性能测试是对比新模型与旧模型在准确率、召回率、推理延迟等性能指标上的差异，确保新模型性能不低于旧模型。

e）数据的更新应遵循数据安全标准，对更新的数据进行严格的合法性和安全性审查。

f）算法更新后应进行全面测试，防止因算法调整引入新的安全漏洞或性能问题。

g）在更新过程应实施版本管理，记录每次模型更新的版本号、更新内容、更新时间以及负责人等信息。一旦新模型在应用中出现问题，能够快速回滚到上一个稳定版本，保障安全应急业务的持续运行。

7.3.2 模型维护

模型维护包含：

a）定期对模型进行维护，包括数据更新和算法优化等，同时确保维护过程的安全性。

b）维护过程中，对模型的访问权限进行重新评估和调整，确保只有授权人员能够进行模型维护操作，防止未经授权的人员对模型进行恶意修改或破坏。

8 其他要求

8.1 核心功能验证

涉及人身安全的关键场景（如矿山灾害预警、消防应急处置），模型的风险识别准确率与响应时间按照相关标准设定，应通过权威机构的第三方认证。

8.2 应急响应合规

模型在生成处置方案时，应急符合应急管理规范，确保建议的可操作性与合法性，禁止提供违反安全规程的决策支持。

8.3 算法透明

模型开发方应公开核心算法的技术原理与训练数据来源，接受监管部门的审查，防止算法偏见或黑箱操作带来的风险。