



中华人民共和国金融行业标准

JR/T XXXXX—XXXX

证券期货业智能模型安全 基本要求

Basic security requirements
for AI models in securities and futures industry

（征求意见稿）

XXXX—XX—XX 发布

XXXX—XX—XX 实施

中国证券监督管理委员会

发布

目 次

前言 II

引言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 总体安全要求 2

 4.1 总体目标 2

5 数据存储与取用安全要求 2

 5.1 数据存储安全 2

 5.2 数据取用安全 3

6 模型有效性要求 3

7 模型安全性要求 3

8 模型鲁棒性要求 4

9 模型隐私性要求 4

10 模型公平性要求 4

11 模型可解释性要求 4

12 模型持续维护能力要求 5

13 模型安全监测与响应机制要求 5

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。

本文件由全国金融标准化技术委员会证券分技术委员会（SAC/TC 180/SC4）提出。

本文件由全国金融标准化技术委员会（SAC/TC 180）归口。

本文件起草单位：北京航空航天大学、申万宏源证券有限公司、华金证券股份有限公司、金圆统一证券股份有限公司、格林基金管理有限公司。

本文件主要起草人：吴俊杰、韩潇、左源、朱天宇、部慧、王怀思、王叶、张煜、邓科夫、马文杰。

引 言

随着人工智能技术和数据驱动模型在证券期货业的深入应用，以及大语言模型技术推动智能模型在金融诸多业务的广泛应用，例如程序化投资、智能推荐、智能客服、财富管理、风险评估与风险管理等，智能模型能提升效率、改进效果的同时也会引发模型风险问题。国际上，模型风险管理（MRM）起源于银行业的风险管理要求，以美联储 SR 11-7 监管信为标志，但现有国际实践多以原则性指引为主，缺乏系统性的模型安全规范。当前，模型风险管理已从传统的操作风险分支发展为必须单独应对的风险管理领域，尤其在金融行业特别是证券期货业，亟需建立覆盖模型全生命周期的模型安全基本要求，明确安全要求的维度体系，以确保模型在设计、开发、部署、运行及退役各环节的稳定性、可控性与合规性，从源头构建模型安全防线。为此，制定证券期货业智能模型安全基本要求的金融行业标准，旨在弥补现有原则性指引在实操层面的不足，引导行业机构系统化提升模型风险防护能力，从而有效维护金融行业的稳健运行。

当前，证券期货业模型体系多样、算法复杂、应用场景广泛。不同类型的智能模型在数据依赖、算法结构、应用目标等方面差异显著，易产生模型偏差、算法歧视、数据泄露等问题并由此引发系列风险。建立统一、明确的模型安全基本要求，有助于规范模型全生命周期的模型风险管理，为模型开发方、使用方和管理机构提供清晰的安全基准，是提升模型内在安全性、公平性、可解释性与可靠性的根本依据。明确的模型安全基本要求是智能模型治理体系的基石，是保障模型科学性、公正性与安全性的前提。通过制定并遵循本文件提出的模型安全基本要求，相关机构能够在模型生命周期的各阶段落实必要的安全控制措施，构建模型风险主动防御体系，在推动技术驱动型业务创新的同时，有效防范模型失效、决策被模型误导、甚至引发合规风险等情形，从而维护证券期货市场的安全与稳定。

本文件旨在为证券期货业智能模型的全生命周期安全管理提供一套完整的模型安全要求框架。本文件结合行业监管要求和技术发展趋势，提出了涵盖数据、算法、模型、系统及治理等多个维度的具体模型安全基本要求，明确了各类模型在安全性、有效性、鲁棒性、隐私性、公平性和可解释性等方面应达到的基本水平，为金融机构、模型提供方和相关机构在智能模型的设计、开发、部署、运维与治理过程中提供明确的指引。此外，本文件亦可为第三方评估、行业监督及相关标准制定提供技术依据。

证券期货业智能模型安全基本要求

1 范围

本文件规定了证券期货业智能模型在设计、开发、部署、运行及退役全生命周期中的安全基本要求，包括总体安全要求、数据存储与取用安全、模型有效性、安全性、鲁棒性、隐私性、公平性、可解释性、持续维护能力以及安全监测与响应机制要求。

本文件适用于证券期货业各类智能模型的研发、应用与治理活动，供行业机构、模型提供方落实智能模型安全要求时使用，也可作为第三方评估机构和行业监管部门提供依据。

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 35273—2020 信息安全技术 个人信息安全规范
- GB/T 37964—2019 信息安全技术 个人信息去标识化指南
- GB/T 41479—2022 信息安全技术 网络数据处理安全要求
- GB/T 42888—2023 信息安全技术 机器学习算法安全评估规范
- JR/T 0060—2021 证券期货业网络安全等级保护基本要求
- JR/T 0223—2021 金融数据安全 数据生命周期安全规范

3 术语和定义

3.1

智能模型 intelligent model

采用机器学习等方法建立的输入与目标输出联系的计算模型，主要包含算法、超参数、参数、模型输入规范、模型输出规范五大要素。

注：大语言模型、多模态大模型等均属智能模型的子类，它们同样由这五大要素构成，是智能模型在大规模预训练范式下的具体实现。

3.2

白盒模型 white box model

使用者对其内部结构和参数的情况拥有完全信息的模型。

注：如逻辑回归、决策树、规则系统等天然具备可解释性的模型。

3.3

模型生命周期 model lifecycle

智能模型从需求分析、设计开发、验证测试、部署运行、监控维护到最终退役下线的全过程。

3.4

模型风险 model risk

因智能模型在设计、数据、算法、部署或管理过程中的缺陷，导致其输出结果偏离业务目标，或引发决策错误、财务损失、合规违规等负面影响的可能性。

3.5

模型合规性 model compliance

智能模型在其全生命周期中，遵循相关法律法规、监管要求、行业标准、技术规范及内部治理政策的能力与状态。

3.6

模型安全要求 model security requirements

为保障智能模型在其生命周期中各阶段的安全性而提出的技术、管理、运行与治理等方面的规范性要求。

3.7

去标识化 de-identification

通过对个人信息的技术处理，使其在不借助额外信息的情况下，无法识别个人信息主体的过程。

[GB/T 35273-2020, 定义3.14]

注：去除标识符与个人信息主体之间的关联性

4 总体安全要求

4.1 总体目标

本章规定了智能模型安全应达到的总体目标。第三方检测机构应基于本章要求，评估受测模型的安全治理框架、风险管控机制及体系完备性。智能模型在其全生命周期中应满足以下总体安全要求，以构建系统化、可管理、可持续的安全防御体系：

- a) 应建立权责清晰的治理框架，明确模型安全负责人及相关部门职责，制定覆盖全生命周期的安全策略与管理制度。
- b) 应构建覆盖技术、管理与运行的全流程纵深防御体系，确保从模型设计、开发到部署、运维、退役各环节均落实相应安全措施。
- c) 应实施基于风险的分级管控，依据业务重要性、数据敏感性与影响范围等方面对模型进行风险分级。
- d) 应遵循透明、可信、公平的原则，确保模型决策过程可追溯、行为可解释，且在不同群体间保持公平合理，避免不公正歧视。
- e) 应建立持续的风险评估与监控机制，在模型生命周期各阶段系统化识别、分析、处置与跟踪安全风险，并对模型性能与安全状态进行常态化监测。
- f) 应具备敏捷的应急响应与恢复能力，建立安全事件应急预案并定期演练，确保事件发生时能快速隔离、处置、恢复并按规定报告。
- g) 应确保全生命周期活动合规，遵守相关法律法规、监管要求及内部规范，并建立完备的审计与证据留存机制以支持核查。

5 数据存储与取用安全要求

5.1 数据存储安全

第三方检测机构应验证智能模型在数据存储过程中是否满足以下安全要求。

本节所指“数据”涵盖智能模型生命周期各阶段所涉及的核心数据资产，主要包括：模型训练数据、模型权重与参数数据、模型配置文件、推理输入输出数据、用户交互与反馈数据、模型运行日志与监控数据、数据标注信息、特征工程中间数据、模型评估报告数据、版本记录等。上述数据是智能模型构建、

运行、迭代与治理的基础支撑：训练数据决定模型的知识与能力边界，权重与参数数据是模型核心知识形态的存储载体，推理与交互数据反映模型运行状态与用户行为，而日志与监控数据则为模型审计、回溯与优化提供依据。根据不同业务场景，数据类型可包括表格、文本、音频、图片等。数据与模型是构成智能系统的两大核心要素，数据安全是模型安全的基础前提与关键保障。

- a) 应确保训练数据、验证数据、测试数据及运行日志的完整性、保密性和可用性。
- b) 模型运行所依托的存储系统应经过安全加固、渗透测试与漏洞扫描。
- c) 存储介质应实行物理安全与逻辑安全双重防控，访问应仅限授权人员。
- d) 对含有敏感信息的数据，应采用静态加密与分级密钥管理，密钥应独立存储并定期轮换。
- e) 应建立多点数据备份机制，对证券期货核心业务数据，宜建立异地容灾系统。
- f) 应定期对数据存储系统进行审计，验证加密、权限控制、日志记录等机制的有效性。

5.2 数据取用安全

第三方检测机构应验证智能模型相关数据在取用过程中是否满足以下要求。

- a) 数据采集与使用应遵守合法、正当、必要原则，严格执行数据最小化与目的限定原则。
- b) 应建立数据访问审批流程，对数据取用请求实施分级审核与记录留痕。
- c) 应实施去标识化处理，确保任何单一数据样本不可直接关联至特定个人。
- d) 应设定差异化的访问控制策略，对训练集、测试集、生产数据集等分别配置不同权限。
- e) 应保存数据访问的全链路日志，并建立自动化审计系统检测异常取用行为。
- f) 数据接口应启用身份认证、请求限流、签名验证等安全防护机制。

6 模型有效性要求

模型有效性是衡量智能模型达成其预设业务目标的核心能力，具体包括性能、泛化能力和业务适应性，第三方检测机构应验证智能模型有效性是否满足以下要求。

- a) 模型发布前应系统评估其在测试环境与真实业务场景中的性能表现。
- b) 应建立有效性评估指标体系，设定关键性能指标的阈值范围，并对性能衰减建立监控机制。
- c) 应验证模型输出是否满足实际业务需求，如风险识别覆盖率、客户分群合理性等。
- d) 应通过交叉验证等方法，检验模型在未知数据上的稳定性与泛化能力。
- e) 应在模拟的市场波动或异常业务情况下，验证模型决策的合理性。
- f) 检测过程中应使用具有代表性的业务数据集，覆盖主流场景与边缘案例。
- g) 对于生成式智能模型，在保证生成内容符合通用法律法规及公序良俗以外，还应严格约束输出内容不得涉及非法荐股、市场操纵及误导性投资建议，确保不干扰市场秩序、不侵害投资者权益。

7 模型安全性要求

模型安全性关注于保护智能模型资产本身免受破坏、窃取或滥用，第三方检测机构应验证智能模型安全性是否满足以下要求。

- a) 模型文件应加密存储，对模型关键算法应实施防逆向措施。
- b) 模型仓库应建立细粒度权限控制机制，仅授权人员可访问相应版本。
- c) 应建立模型版本管理体系，对模型更新、回滚、下线进行全过程记录。
- d) 模型在部署与传输环节应使用安全通道进行加密传输，并在部署前进行数字签名验证。
- e) 对于涉及高敏感业务的模型，应在可信执行环境或硬件加密容器等受保护环境中运行。

- f) 模型的应用程序编程接口（API）应具备身份认证、访问日志记录、调用速率限制与异常检测机制。
- g) 应建立模型备份与恢复机制，确保模型在损坏后可快速恢复至稳定版本。

8 模型鲁棒性要求

模型鲁棒性是指智能模型在面对异常输入、对抗性干扰或环境变化时，维持其性能与决策可靠性的能力，第三方检测机构应验证智能模型鲁棒性是否满足以下要求。

- a) 应能在面对带噪数据、对抗样本数据、分布偏移数据以及极端场景数据等异常输入时维持其性能稳定性。
- b) 应具备一定的容错机制和冗余设计，以应对输入扰动或市场剧烈波动带来的影响。
- c) 宜采用对抗训练、输入验证与过滤、模型正则化、集成学习等增强技术提升模型鲁棒性。
- d) 对于部署在生产环境的模型，应建立实时输入监控与异常检测机制。
- e) 应通过模拟测试各种极端市场环境，验证模型在这些条件下的稳定性和决策准确性。

9 模型隐私性要求

模型隐私性旨在防止在智能模型数据处理、训练及服务过程中泄露用户或企业的敏感信息，第三方检测机构应验证智能模型隐私性是否满足以下要求。

- a) 应确保在数据预处理、特征工程、模型训练及推理过程中，不会直接或间接泄露用户个人信息、敏感属性或商业秘密。
- b) 应采用差分隐私、联邦学习、同态加密、安全多方计算等技术增强模型隐私保护能力。
- c) 应验证数据处理环节中的加密和去标识化措施的有效性。
- d) 应具备抵御成员推理攻击、模型反演攻击、属性推理攻击及模型窃取攻击的能力。
- e) 应符合相关隐私法规并拥有详细的合规记录。

10 模型公平性要求

模型公平性要求智能模型的决策结果在不同群体或个体间保持公正、无不当歧视，第三方检测机构应验证智能模型公平性是否满足以下要求。

- a) 应确保模型决策在不同地域、年龄、性别、种族等受保护特征群体间具有公平性与一致性。
- b) 应通过多种公平性指标测试，定量评估模型是否存在系统性偏差。
- c) 应检测训练数据的多样性和均衡性，避免数据偏差影响模型表现。
- d) 应建立模型实时调整和偏差纠正的机制，确保模型在运行过程中持续满足公平性要求。
- e) 宜通过数据重采样、特征调整、算法优化或后处理校正等措施增强模型公平性。

11 模型可解释性要求

模型可解释性是指智能模型的决策逻辑、输入输出关系能够被相关方理解、追溯与信任的程度，第三方检测机构应验证智能模型可解释性是否满足以下要求。

- a) 白盒模型的输出结果应附带可理解的解释，其关键决策过程应能被业务方和监管机构追踪和理解。
- b) 白盒模型应提供特征重要性分析、决策规则或决策路径可视化等解释手段。

- c) 解释结果应与领域知识和业务逻辑保持一致，不存在矛盾或违反常识的决策路径。
- d) 所使用的解释方法应科学、稳定，并适用于当前模型类型与具体业务场景。
- e) 应具备模型文档化和版本控制，确保模型开发和训练过程的完整记录。
- f) 对于生成式智能模型，应强制对其生成内容进行显式标识（如水印或文字标注）。

12 模型持续维护能力要求

模型持续维护能力是确保智能模型在部署后能够长期稳定、有效运行的关键保障，第三方检测机构应验证智能模型持续维护能力是否满足以下要求。

- a) 应具备自动化的模型性能、数据分布及业务指标监测和评估机制。
- b) 应能通过对比不同时段的模型运行性能与基线，及时发现性能偏移与概念漂移。
- c) 应建立性能偏移预警与分级告警机制，当关键指标变化超过预设阈值时应自动触发告警并通知相关负责人。
- d) 应具备明确的模型迭代与更新策略，规定触发再训练或优化的条件、评估标准及标准操作程序。
- e) 应建立完善的模型版本管理、发布与回滚机制，确保在出现问题时能快速、可控地恢复服务。
- f) 应建立从数据输入、模型推理到业务反馈的完整监控闭环，确保能及时识别输入数据质量变化并调整模型。

13 模型安全监测与响应机制要求

模型安全监测与响应是针对智能模型面临的恶意攻击与安全事件，进行主动防御、及时发现、快速处置的关键防线，第三方检测机构应验证智能模型安全监测与响应是否满足以下要求。

- a) 应建立模型攻击监测与响应机制，实时监控异常访问、参数泄露和推理行为。
- b) 模型部署环境应内置攻击检测模块，识别异常查询模式、参数提取企图和对抗样本输入。
- c) 应建立模型安全事件响应流程，包括告警、隔离、模型回滚、证据留存、风险通报与复测等环节。
- d) 应对智能模型的每一次推理请求建立完整的链路追踪记录，覆盖输入提示、模型推理、工具调用、输出生成等环节，确保所有决策过程可观测。
- e) 宜建立智能模型的实时监控体系，对性能衰减、异常行为、模型漂移、安全风险等设置告警阈值和联动响应机制。
- f) 当发生严重安全事件时，应在规定时间内向上级主管部门报告并采取临时封禁或下线措施。