



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 检测要求

Security testing and evaluation 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 III

引言 IV

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 总体要求 2

 4.1 检测对象与范围 2

 4.2 检测框架 2

 4.3 检测过程 2

 4.4 检测方法 2

 4.5 检测结果与报告 3

5 数据存储与取用安全检测 3

 5.1 数据存储安全检测 3

 5.2 数据取用安全检测 4

6 模型有效性检测 4

 6.1 检测目的 5

 6.2 检测流程 5

 6.3 通过要求 5

7 模型安全性检测 5

 7.1 检测目的 5

 7.2 检测流程 5

 7.3 通过要求 6

8 模型鲁棒性检测 6

 8.1 检测目的 6

 8.2 检测流程 6

 8.3 通过要求 7

9 模型隐私性检测 7

 9.1 检测目的 7

 9.2 检测流程 7

 9.3 通过要求 7

10 模型公平性检测 7

 10.1 检测目的 7

 10.2 检测流程 8

 10.3 通过要求 8

11 模型可解释性检测 8

 11.1 检测目的 8

 11.2 检测流程 8

- 11.3 通过要求 8
- 12 模型持续维护能力检测 9
 - 12.1 检测目的 9
 - 12.2 检测流程 9
 - 12.3 通过要求 9
- 13 模型安全监测与响应机制检测 9
 - 13.1 检测目的 9
 - 13.2 检测流程 9
 - 13.3 通过要求 10

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、王叶、张煜、邓科夫、王悦青。

引 言

尽管模型风险管理在国内外受到高度重视，但普遍面临“重原则、轻检测”的困境，模型上线应用前缺少统一、可量化的安全检测方法与标准。在证券期货业，数据驱动乃至大语言模型支持的风控模型、程序化投资模型、智能推荐算法等在多业务场景已广泛部署使用。金融业务复杂性导致智能模型体系多样、算法复杂。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明显，易产生模型偏差、算法歧视、内在缺陷、对抗攻击、黑箱决策等问题，这些问题可能引发难以预见的模型风险；若模型风险防范不足，也可能会引发金融系统性风险。

国内外以模型风险原则指引的金融行业实践，令金融机构对于模型风险方法、模型安全评估存在较大技术方法差异性。金融行业特别是证券期货业制定规范化、标准化模型安全检测规范，可以将模型安全管理原则和模型安全基本要求转化为有效且较为统一的模型安全评估手段。该标准为判断模型是否满足已有的模型安全基本要求提供可用的评估方法，是验证模型风险防控措施有效性、确保模型透明度与可解释性达标的关键手段，以促进智能模型在安全、稳健的前提下服务证券期货行业的业务创新。

模型安全检测是智能模型治理体系中不可或缺的验证环节，是保障模型科学性、公正性与安全性的重要实践。通过执行本文件规定的检测流程与方法，机构及相关评估方能够在模型上线前客观评估其模型安全状况和模型风险情况，在模型运行过程中实施动态监测和持续验证，以确保模型持续符合模型安全基本要求，有效防范模型失效、决策被误导、引发合规风险等情形，从而维护证券期货市场安全与稳健运行。

本文件旨在为证券期货业智能模型安全检测工作提供具体的操作框架与评估方法。本文件以《证券期货业智能模型安全基本要求》为依据，结合行业监管要求和技术发展趋势，详细规定了涉及数据安全、模型有效性、安全性、鲁棒性、隐私性、公平性、可解释性、持续维护能力和安全监测与响应等维度的检测目标、流程、方法和通过要求，供金融机构、第三方检测机构及相关管理单位在智能模型的开发验证、上线前检测、运行期复测等过程中参考使用。本文件是落实模型安全基本要求、实现模型风险可控的关键技术支撑。

证券期货业智能模型安全测评要求

1 范围

本文件适用于第三方检测机构、证券期货经营机构、模型提供方等开展的智能模型安全检测与自评估活动。

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 17859—1999 计算机信息系统 安全保护等级划分准则
GB/T 25069—2010 信息安全技术 术语
GB/T 42888—2023 信息安全技术 机器学习算法安全评估规范
GB/T 35273—2020 信息安全技术 个人信息安全规范
JR/T 0060—2021 证券期货业网络安全等级保护基本要求
JR/T 0133—2015 证券期货业信息系统托管基本要求
JR/T 0202—2020 基于大数据的支付风险智能防控技术规范
JR/T 0206—2021 证券期货业大数据平台性能测试指引

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据 data

信息的可再解释的形式化表示，以适用于通信、解释或处理

注：可以通过人工或自动手段处理数据。

[来源：GB/T 5271.1—2000. 定义01.01.02]

3.2

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。

3.3

证券期货行业智能模型 intelligent model of securities and futures industry

证券期货行业智能模型是深度融合人工智能技术与证券期货行业业务逻辑，构建的能够进行数据驱动的预测、决策与风险控制的算法模型。

3.4

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。

注1：模型合规性涵盖数据来源与使用的合法性、算法决策的公平性与透明性、输出结果的可审计性，以及模型应用不触犯金融监管红线等要求。

注2：模型合规性是一个动态过程，需随法规政策和市场环境的变化进行持续评估与调整。

3.5

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。

注：模型风险贯穿模型全生命周期，需通过系统性评估、检测与治理进行识别与防控。

3.6

黑盒场景 black box scenario

检测者对待测模型的内部结构和参数的情况一无所知的检测环境。

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 总体要求

4.1 检测对象与范围

检测机构应将处于设计开发后、验证测试、部署上线或运行维护等任一阶段的证券期货业智能模型作为检测对象。检测范围应覆盖模型对外表现、接口行为、及可获取的相关文档与日志。

4.2 检测框架

安全检测应基于《证券期货业智能模型安全要求规范》中规定的九个安全维度展开，构建覆盖模型全生命周期的系统性检测框架。检测活动应包括但不限于：文档审查、配置核查、工具扫描、模拟攻击、性能压测、模型数据的生成、标注、存储、取用、合成质量、语义一致性、模态对齐等关键环节等。

4.3 检测过程

检测过程包括以下阶段：

- 准备阶段：明确检测目标、范围，获取模型及相关文档，搭建隔离的测试环境。
- 实施阶段：依据本文件第5至13章，综合运用多种检测方法执行具体检测项。
- 分析阶段：整理检测数据与证据，分析安全状况，评估风险等级。
- 报告阶段：编制检测报告，清晰描述检测过程、发现的问题、风险分析与整改建议。

4.4 检测方法

检测方法包括：

- 文档审查：审查安全设计文档、开发规范、配置清单、管理日志等。
- 配置核查：人工核查系统配置、权限设置、加密策略、日志策略等。
- 工具扫描：使用专用安全工具对模型文件、代码、接口进行自动化漏洞扫描。

- d) 模拟测试：构造测试用例，模拟正常、异常及恶意场景，验证模型行为。
- e) 渗透测试：在授权范围内，模拟攻击者行为，尝试发现潜在安全漏洞。

注：本规范默认检测在黑盒场景下进行。”文档审查”仅限于受检方主动提供的、非涉密的公开或授权文档；“配置核查”仅限于通过授权接口可查询的配置信息；”工具扫描”与”模拟测试”应针对模型对外接口与表现进行。

4.5 检测结果与报告

检测结果满足以下要求的，判定通过：

检测结果应以报告形式呈现，内容至少应包括：

- a) 检测概述（目标、范围、时间、人员）；
- b) 检测依据与方法；
- c) 详细检测过程与发现；
- d) 安全问题与风险等级评估；
- e) 整改建议与复测说明；
- f) 综合结论。

5 数据存储与取用安全检测

5.1 数据存储安全检测

5.1.1 检测目的

验证模型相关数据在存储过程中的机密性、完整性和可用性是否得到保障，存储系统、介质及备份恢复机制是否符合安全要求，确保静态数据安全。

5.1.2 检测流程

检测过程包括以下阶段：

- a) 审查数据安全策略，检查训练数据、验证数据、测试数据及运行日志的存储配置，验证是否实施了完整性校验（如哈希）、加密存储及访问控制，并通过模拟中断测试可用性。
- b) 审查存储系统近期的渗透测试报告、漏洞扫描报告及安全加固方案，核查中高风险漏洞的修复记录与闭环情况。
- c) 检查存储介质所在的物理环境访问控制记录与监控日志，并测试非授权账号或互联网协议（IP）地址对存储资源的访问是否被拒绝。
- d) 审查含有敏感信息数据的加密策略文档，检查静态加密是否启用，并核查密钥管理系统的独立性、分级管理策略及密钥轮换记录。
- e) 审查数据备份策略文档、备份日志及历史恢复演练报告，确认备份机制（如全量、增量）与周期是否符合要求。
- f) 审查存储系统安全审计计划与近期审计报告，核查审计内容是否覆盖加密、权限、日志等机制，并验证审计发现问题的整改情况。
- g) 审查敏感数据加密及密钥管理所采用的密码算法是否符合国家密码管理部门要求的商用密码算法（如 SM2/SM3/SM4）；对涉及关键信息基础设施的模型系统，查验商用密码应用安全性评估报告及评估中所发现问题的整改情况。

5.1.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 完整性校验机制健全，存储数据全部加密，访问控制有效，且系统具备高可用性。
- b) 近一年内经过渗透测试与漏洞扫描，且所有中高风险漏洞已修复或有效缓解。
- c) 物理访问受控，逻辑访问遵循最小权限原则，未授权访问均被阻断并记录。
- d) 敏感数据已启用符合行业要求的静态加密，密钥由独立系统分级管理并定期轮换。
- e) 建立了有效的多点备份机制，并定期进行恢复演练。
- f) 定期开展存储系统安全审计，且审计报告显示加密、权限、日志等机制运行有效，发现问题已全部整改。
- g) 敏感数据加密与密钥管理采用符合国家密码管理部门要求的商用密码算法；涉及关键信息基础设施的模型系统已开展商用密码应用安全性评估且评估结论符合要求。

5.2 数据取用安全检测

5.2.1 检测目的

验证数据采集、访问、使用、传输及销毁等环节是否符合合法、正当、必要原则，访问控制与审计机制是否健全。

5.2.2 检测流程

检测过程包括以下阶段：

- a) 审查数据采集的授权同意文件、隐私政策及业务需求文档，核对数据采集范围、使用目的与业务必要性的关联关系，评估是否符合“合法、正当、必要”及目的限定原则。
- b) 审查数据访问审批流程制度，检查近期的数据取用审批记录，验证流程是否包含申请、审核、授权、记录全环节，并测试提交申请后流程的执行情况。
- c) 审查去标识化处理的技术方案，抽取处理后样本，尝试使用背景知识或关联数据实施重识别攻击，评估其残留风险。
- d) 审查数据访问控制策略配置，检查是否针对训练集、测试集、生产数据集等设置了差异化的用户或角色权限，并使用测试账号验证权限隔离的有效性。
- e) 审查数据访问日志系统，验证日志是否完整记录访问者、时间、操作对象、操作类型及访问结果；模拟异常访问模式（如非工作时间高频访问、访问非常见数据），测试自动化审计系统是否能检测并告警。
- f) 对数据服务接口进行安全测试，包括：身份认证测试（无凭证、无效凭证访问）；请求限流测试（高频并发请求）；签名验证测试（篡改签名或重放请求）。

5.2.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 数据采集具有合法授权，使用目的明确且限于必要范围，未发现超范围采集或使用。
- b) 建立了成文的数据访问审批流程，且审批记录完整、可追溯，流程执行无例外。
- c) 采用的去标识化技术合理，样本重识别攻击成功率低于行业或机构预设的风险阈值。
- d) 实现了差异化的数据集访问控制，测试中未出现越权访问情况。
- e) 数据访问日志要素齐全、保存完整，自动化审计系统能有效识别并预警预设的异常行为模式。
- f) 数据接口具备强身份认证、有效的请求限流与防重放机制，安全测试中未发现中高风险漏洞。

6 模型有效性检测

6.1 检测目的

验证模型在预设业务场景下的性能表现是否达到预期目标，是否具备良好的泛化能力和业务适用性。

6.2 检测流程

检测过程包括以下阶段：

- a) 审查模型上线前的评估报告，包括在独立测试环境和真实业务场景中的性能测试记录，验证评估是否系统、全面。
- b) 审查模型有效性评估指标体系文档，检查是否定义了关键性能指标及其阈值。
- c) 分析模型输出与业务需求的验证报告，例如针对风险识别模型，核实其覆盖率和误报率等是否达到业务设定目标。
- d) 复现交叉验证、金融时间序列外推等测试过程，或抽样验证测试报告中的数据与结论，以检验模型在未知数据上的稳定性与泛化能力。
- e) 查验在模拟市场波动或异常业务情况下的压力测试或仿真测试报告，评估测试场景的合理性。
- f) 分析模型测试阶段所使用的数据集构成文档，评估其是否覆盖了业务主流场景、边缘案例及历史风险事件。
- g) 对于生成式智能模型，需检验模型生成内容与行业事实的对齐程度。
- h) 对于处理多模态数据的智能模型，应进行模态对齐检测。

6.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 模型发布前具备系统性的测试环境与真实场景性能评估报告，且结果达标。
- b) 已建立成文的有效性评估指标体系，设定了明确的设定的明确的关键绩效指标（key performance indicator, KPI）阈值，并部署了性能监控与衰减告警机制。
- c) 模型输出经验证满足实际业务需求，模型性能指标达到预设标准（例如交易欺诈检测、账户异常检测等场景下 F1 值（F1-score，精确率与召回率的调和平均值） ≥ 0.8 ）。
- d) 交叉验证等方法显示模型具有较好的稳定性和泛化能力，未出现严重过拟合或欠拟合。
- e) 在模拟的异常情况下，模型决策逻辑合理，未出现系统性误判或失效。
- f) 测试数据集具有代表性，覆盖了必要的业务场景，测试结论可信。
- g) 对于生成式智能模型，配有事实对齐流程拦截幻觉内容，保证生成内容基于有效事实。
- h) 对于处理多模态数据的智能模型，模型在不同模态的相同语义输入下输出的决策一致。

7 模型安全性检测

7.1 检测目的

验证模型资产的防窃取、防篡改、防逆向能力，以及模型在存储、传输、部署、运行及接口服务各环节的安全防护措施是否有效。

7.2 检测流程

检测过程包括以下阶段：

- a) 检查模型文件存储状态，验证是否加密；使用反编译或逆向工具尝试分析关键算法，评估其抗逆向能力。
- b) 查验模型仓库的权限配置与操作审计日志，验证权限粒度及日志完整性。

- c) 调阅模型版本管理系统的操作记录，检查模型更新、回滚、下线的全过程是否可追溯。
- d) 验证模型部署包在传输过程中是否使用使用传输层安全协议（TLS）/安全套接层协议（SSL）等安全通道，核查传输通道加密与数字签名所采用的密码算法是否符合国家密码管理部门要求的商用密码算法（如 SM2/SM3/SM4），并核对部署前的数字签名有效性。
- e) 对于高敏感业务模型，查验其部署文档，并通过系统命令或专用工具验证其是否实际运行于可信执行环境（TEE）或硬件加密容器内。
- f) 对模型的应用程序编程接口（API）实施渗透测试，包括身份认证绕过、参数注入、访问控制测试、异常检测触发等。
- g) 查验模型备份策略与备份文件，并执行恢复演练，验证恢复流程与时效。
- h) 对于生成式智能模型，需检验模型生成内容合规性。（例如检验模型能否拒绝诱导性违规操作请求，能否确保信息披露准确完整，拒绝配合隐瞒负面信息或做出保本保收益承诺，能否识别并拒绝生成欺诈性内容，以及能否保护用户隐私，不输出或滥用敏感信息等。）
- i) 对于第三方供应的智能模型，需对模型服务商、数据供应商、数据服务提供商等上游机构进行调研、准入审计。
- j) 对涉及关键信息基础设施的模型系统，查验商用密码应用安全性评估报告，核查评估结论及所发现问题的整改落实情况。

7.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 模型仓库访问控制有效，未授权访问被拒绝。
- b) 模型具备一定的防逆向工程能力。
- c) 模型在传输和部署环节有完整性校验机制，传输通道加密与数字签名所采用的密码算法符合商用密码算法相关要求。
- d) 模型 API 接口通过基本的安全测试。
- e) 具备有效的模型备份与快速恢复能力。
- f) 高敏感模型运行环境符合安全要求。
- g) 对于生成式智能模型，配有合规审核流程拦截违规内容，保证生成内容合规。
- h) 对于第三方供应的智能模型，对上游供应机构准入审计通过。
- i) 涉及关键信息基础设施的模型系统已开展商用密码应用安全性评估，评估结论符合要求且所发现问题已整改。

8 模型鲁棒性检测

8.1 检测目的

评估模型在面对噪声、对抗样本、数据分布变化等异常或恶意输入时的稳定性和可靠性。

8.2 检测流程

检测过程包括以下阶段：

- a) 执行噪声注入测试、对抗样本攻击测试、数据分布偏移测试及极端场景数据测试，记录模型核心性能指标的变化。
- b) 分析模型设计文档，检查是否包含输入校验、容错处理及冗余设计，并通过注入扰动数据测试其效果。

- c) 如宣称采用，则查验对抗训练、输入过滤、正则化、集成学习等技术的实施方案与记录。
- d) 查验生产环境模型的实时输入监控与异常检测系统的配置，并分析其历史告警日志。
- e) 查验针对极端市场环境的模拟测试报告。
- f) 对于生成式智能模型，需对其进行对抗性提示词测试及越狱测试。

8.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 在预设强度的噪声、对抗攻击及分布偏移下，模型核心性能衰减不超过可接受阈值（例如交易欺诈检测、账户异常检测等场景下性能衰减不超过 10%）。
- b) 模型具备有效的输入校验与容错机制，在输入扰动下不产生灾难性失效。
- c) 若采用增强技术，应有证据表明其被实施并有助于提升鲁棒性。
- d) 生产环境具备实时输入监控与异常检测能力，并能有效告警。
- e) 在极端市场环境模拟测试中，模型表现出必要的稳定性与决策合理性。
- f) 对于生成式智能模型，模型应拒绝回复与业务无关或有害的指令或生成不可接受的内容。

9 模型隐私性检测

9.1 检测目的

评估模型在数据处理、训练及服务过程中是否存在泄露用户或企业敏感信息的风险，验证隐私保护技术措施的有效性及其合规性。

9.2 检测流程

检测过程包括以下阶段：

- a) 分析数据预处理、特征工程、训练及推理流水线的安全设计，检查中间数据是否加密或脱敏，并通过渗透测试尝试窃取敏感信息以验证防护效果。
- b) 如宣称采用，查验差分隐私、联邦学习、同态加密或安全多方计算的技术方案与实施记录。
- c) 查验数据处理环节的去标识化方案，并通过抽样测试验证其有效性。
- d) 执行黑盒场景下的成员推理攻击、模型反演攻击、属性推理攻击与模型窃取攻击在内的隐私性相关攻击，记录攻击成功率。
- e) 查验隐私影响评估报告、合规性声明及相关法规遵循性证据。

9.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 全流程未发现可导致敏感信息直接或间接泄露的设计缺陷或漏洞。
- b) 若采用隐私增强技术，其参数配置合理，并有证据表明技术已落地。
- c) 去标识化措施有效，重识别风险低于预设阈值。
- d) 针对各类隐私攻击的模拟攻击成功率低于可接受风险阈值。
- e) 具备完整的隐私合规记录，符合相关法规要求。

10 模型公平性检测

10.1 检测目的

识别模型决策中是否存在对不同群体的不合理偏见或歧视。

10.2 检测流程

检测过程包括以下阶段：

- a) 选择受保护特征，划分群体，计算模型在各群体上的关键性能指标及业务结果的差异。
- b) 使用多种公平性指标对模型进行定量评估。
- c) 分析训练数据集的多样性、均衡性分析报告，检查各敏感属性组的样本分布与标注质量。
- d) 查验模型上线后的公平性监控日志、偏差告警记录及任何调整或纠正措施的文档。
- e) 如宣称采用，查验数据重采样、特征调整、算法去偏或后处理校正等公平性增强措施的实施依据及效果。

10.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 模型在不同受保护特征群体间的性能差异与业务结果差异不超过预设的公平性阈值。
- b) 公平性量化指标显示模型不存在不可接受的系统性偏差。
- c) 训练数据在可能影响公平性的维度上具有合理的代表性，无明显系统性偏差。
- d) 建立了公平性监控与纠正机制，且运行记录可查。
- e) 若采用增强措施，应有证据表明其对提升模型公平性有积极作用。

11 模型可解释性检测

11.1 检测目的

评估模型的决策过程是否透明、可理解，提供的解释是否准确、一致、有用，能否支持业务理解、合规审查与信任建立。

11.2 检测流程

检测过程包括以下阶段：

- a) 评估模型输出的解释信息，并可组织业务或合规人员评估其可理解性。
- b) 验证模型是否提供至少一种解释工具或输出，如全局/局部特征重要性排序、决策规则列表、决策路径图或反事实解释。
- c) 检查解释结果是否存在与领域知识或业务逻辑明显矛盾的情况。
- d) 评估所选解释方法的科学性及其对当前模型类型的适用性。
- e) 查验模型技术文档，包括业务目标、数据谱系、特征定义、算法选择、训练参数及版本历史。

11.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 模型输出附带解释，且目标用户能基本理解其决策依据。
- b) 提供了一种以上有效的解释手段。
- c) 解释结果与业务常识一致，无重大逻辑矛盾。
- d) 所选解释方法适用于该模型，且解释结果具有稳定性。
- e) 具备完整、清晰的模型开发与训练过程文档。

12 模型持续维护能力检测

12.1 检测目的

验证模型上线后是否具备有效的性能监控、退化发现、迭代更新和应急恢复能力，确保模型长期稳定运行。

12.2 检测流程

检测过程包括以下阶段：

- a) 查验自动化监控系统的配置，确认其持续监控模型性能指标、输入数据分布及关键业务指标，并分析历史监控报告。
- b) 分析性能偏移分析报告，检查系统是否定期对比当前性能与历史基线，并识别概念漂移。
- c) 测试性能告警机制：修改监控阈值或模拟性能下降，验证告警能否自动触发并及时通知到预设负责人。
- d) 查验模型迭代与再训练策略文档，明确触发条件、评估标准及标准操作程序。
- e) 查验版本管理系统记录，并测试回滚流程：在测试环境中部署新版本后模拟故障，执行回滚操作，验证其是否快速、可控。
- f) 查验数据质量监控日志与业务反馈闭环记录，验证系统能否识别输入数据变化并触发相应调整。

12.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 建立了自动化、多维度（性能、数据、业务）的模型监控体系。
- b) 能够定期检测并报告性能偏移与概念漂移。
- c) 性能告警机制有效，告警触发与通知及时。
- d) 具备成文的模型迭代与再训练策略，且标准明确。
- e) 版本管理严格，具备经过验证的快速回滚能力。
- f) 建立了从数据输入到业务反馈的完整监控与调整闭环。

13 模型安全监测与响应机制检测

13.1 检测目的

验证模型运行环境是否具备实时安全威胁监测、攻击识别、事件应急响应及审计追溯能力，确保能有效应对针对模型的安全事件。

13.2 检测流程

检测过程包括以下阶段：

- a) 查验模型攻击监测系统的规则配置与实时告警日志，验证其是否监控异常访问模式、可疑的参数查询及异常推理行为。
- b) 查验模型部署环境中内置或附加的攻击检测模块的配置与运行状态。
- c) 查验安全事件应急响应预案，并通过桌面推演或模拟小型事件，验证告警、分析、隔离、处置、报告、恢复、复盘等环节的衔接与有效性。
- d) 查验模型安全日志系统，检查日志内容是否完整包含模型调用、参数变更、访问行为等关键

信息，并验证日志保存期限是否不少于两年及防篡改措施。

- e) 查验安全事件报告流程文档，并核查历史异常事件的报告记录，验证其及时性与规范性。
- f) 如宣称采用，查验模型安全态势感知平台的接入情况、威胁情报整合能力及风险可视化功能。
- g) 查验发生数据安全事件时向用户告知的流程与记录，核对告知内容是否包含：信息种类、原因及可能造成的危害；补救措施及个人可采取的减轻危害措施；以及有效的联系方式。

13.3 通过要求

检测结果满足以下要求的，判定通过：

- a) 建立了针对模型攻击的实时监测机制，并能有效识别常见攻击模式。
 - b) 部署环境具备攻击检测能力，相关模块正常运行。
 - c) 制定了明确的安全事件响应流程，且相关人员熟悉职责，推演过程顺畅。
 - d) 安全日志记录完整、保存期限合规，并具备防篡改与审计检索能力。
 - e) 建立了向主管部门报告重要安全事件的明确流程与时限要求。
 - f) 若建立态势感知平台，应能实现安全数据的整合分析与动态预警。
 - g) 发生或可能发生个人信息泄露、篡改、丢失时，已及时采取补救措施并告知用户和履行个人信息保护职责的部门，告知内容包含上述三项要素。
-