第一创业证券股份有限公司企业标准

Q/FCSC 007-2024

大模型安全应用规范

Specification for the Secure Application of Big Model

2024-10 发布　　2024-10 实施

第一创业证券股份有限公司

目　　次

前　　言 II
1 范围 1
2 引用文件 1
3 定义和术语 1
4 伦理规范 1
5 数据规范 2
6 算法规范 3
7 安全规范 4
8 应用开发规范 5

前　言

人工智能（Artificial Intelligence，简称 AI）大模型技术飞速发展，其在各行各业中的应用也日益广泛。为了确保大模型技术在金融领域应用的安全性、可靠性和合规性，制定一系列规范、规程和标准势在必行。本文将大模型在公司内部安全应用规范进行描述，为公司的 AI 应用提供参考。

本规范旨在为各类 AI 应用开发工作提供一个规范化的操作框架和参考指南，确保 AI 大模型应用过程中的各个环节都能达到合规、安全、风险可控的要求。

本规范按照 GB/T1.1-2020 《标准化工作导则 第一部分：标准化文件的结构和起草规则》的规定起草。

本规范由第一创业证券股份有限公司提出。

本规范由第一创业证券股份有限公司归口。

本规范主要起草部门：第一创业证券股份有限公司信息技术中心。

本规范主要起草人： 任兴涛、龙传泽、陈志浩、侯心宇。

大模型安全应用规范

1 范围

本文件规定了第一创业证券股份有限公司的大模型安全应用的要求。

本文件适用于用于公司所有 AI 大模型相关的数据采集、模型训练、算法优化、应用开发各个环节的软件开发活动。

2 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，仅所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。

JR/T 0250-2022 证券期货业数据安全管理与保护指引

JR/T 0293-2023 期货公司监管数据采集规范

JR/T 0221—2021 人工智能算法金融应用评价规范

JR/T 0287-2023 人工智能算法金融应用信息披露指南

JR/T 0171—2020 个人金融信息保护技术规范

3 定义和术语

人工智能 Artificial Intelligence，简称AI

人工智能是智能学科重要的组成部分，它企图了解智能的实质，并生产出一种新的能以与人类智能相似的方式做出反应的智能机器。人工智能是十分广泛的科学，包括机器人、语言识别、图像识别、自然语言处理、专家系统、机器学习，计算机视觉等。

大模型 Artificial Intelligence Big Model

本文专指人工智能大模型，是指拥有超大规模参数（通常在十亿个以上）、复杂计算结构的机器学习模型。它通常能够处理海量数据，完成各种复杂任务，包括自然语言处理、图像识别等。如，大语言模型（Large Language Model，简称 LLM）。

ETL 是英文 Extract-Transform-Load 的缩写

ETL 是将业务系统的数据经过抽取、清洗转换之后加载到数据仓库的过程，目的是将企业中的分散、零乱、标准不统一的数据整合到一起，为企业的决策提供分析依据。

差分隐私 differential privacy

一种隐私保护的数据共享手段，通过添加噪声的方式（如 Laplace 噪声、指数噪声等），实现仅分享可以描述数据库的一些统计特征、而减少识别其单个个体数据记录的机会。

4 伦理规范

4.1 透明度和问责制

使用 AI 大模型的系统应向用户和相关方明确表明其决策和推荐的依据。如需人工进一步审核决策的，应进行风险提示并确认责任，并在出错或引发争议时能够解释。

4.2 隐私保护

在进行 AI 大模型应用时，应保护用户个人信息的隐私和数据的安全，不应滥用或非法使用。

4.3 公平性和非歧视性

使用 AI 大模型决策和结果时，应避免基于肤色、性别、国籍、年龄、健康、宗教等因素进行歧视，除非存在合理正当理由。

5 数据规范

5.1 数据收集和源头监管

数据的收集应遵循法律法规，从合法、可靠的来源获取，应获得相关数据所有者或授权机构的明确许可，避免借助盗取、欺诈等不当手段。

应防止数据存储、使用等环节被投放药饵数据。

5.2 数据质量和去偏差

开发人员应采取有效措施提升训练数据质量，增强数据集的真实性、准确性、客观性和多样性，数据应具备高质量、低偏差的特点，避免因数据本身偏差导致的歧视或不公平；

训练数据分布应合理：设计文档或管理文档明确训练数据分布的具体说明，防止药饵攻击数据点混入，造成模型倾斜等错误状况。

发现训练数据中存在法律、行政法规禁止的信息，应当采取必要措施，避免使用。

5.3 数据隐私和保护

在使用个人数据时，必须严格遵循相关隐私保护法规，并明确告知用户数据使用的目的和范围。

应尊重并支持数据主体对其个人信息的管理权和控制权，包括查看、更正、删除、撤销同意等权利。当数据主体要求行使这些权利时，应及时响应并配合执行。同时，积极向数据主体提供有关数据安全状况和风险的信息，促进他们更好地参与到个人信息保护中来。

在进行数据收集和使用过程中，应只收集和使用必要的数据，不过度收集数据，同时定期清理不再需要的数据以减少潜在的隐私风险和数据泄漏的可能性

应保障训练数据隐私，避免用户敏感信息泄露：设计文档需制定安全防护措施，如差分隐私加噪等。系统采取与设计文档一致的安全防护措施，保障训练数据安全，防止用户敏感信息泄露。训练数据的安全与隐私性符合 JR/T 0171—2020 的相关要求

5.4 数据窃取攻击防范

应防止训练数据在传输、存储环节被窃取或篡改：管理文档对训练数据传输、存储环节制定明确的安全管理要求。系统在数据传输、存储方面采取的安全措施与管理文档要求一致。

应防止模型在传输、存储环节被窃取或篡改：管理文档中有对模型传输和存储环节的安全管理说明。系统中算法模型在传输和存储环节不会被窃取或篡改。

6 算法规范

6.1 可解释性

大模型算法宜保证可解释性，能够透明展示运算逻辑和决策依据，避免出现“黑箱”算法，建立审查和评估机制来消除算法黑盒问题。优先选择使用可解释性强的模型和算法，以便于理解模型决策过程和结果，同时降低黑箱模型带来的潜在风险。

a) 特征定义可解释：金融应用特征定义应满足相关业务逻辑和规则、特征定义应在系统中有明确记录、特征定义应有可查、详细的 ETL 指标加工过程的记录。

b) 特征分布可解释：应符合日常生活场景和业务规则、应确保特征分布合理，并具备对缺失值、异常值检验的能力、应能展示特征的相关统计指标。

c) 特征衍生可解释：特征衍生应合理、基于业务的特征衍生，应在系统中有明确记录、基于算法的特征衍生，应在系统中明确记录和展示。

d) 特征选择可解释：特征选择过程应有一定的量化统计指标作为决策依据、特征选择过程不应有歧视性、特征选择的业务逻辑和算法依据应在系统中明确记录。

e) 建模过程算法可解释：包括回归算法、决策树算法、图算法、集成学习和深度学习算法，以及其他统计学算法等。可对使用算法进行验证和合理性说明。以回归算法为例，可系统通过统计指标对算法进行可解释性说明，包括但不限于线性回归方程的回归系数、t 检验值、F 检验值。

f) 参数可解释：应对算法参数和超参数应有明确定义，基于业务的参数定义应符合业务逻辑要求；应明确调参所依据的具体指标；应具备对不同调参方式的支持能力。

6.2 公正性

大模型算法应在设计和应用阶段遵循公正性原则，避免因算法本身的偏见导致不公平对待。

6.3 稳定性和可靠性

大模型算法应具备稳定性和容错性，确保在不同环境和数据中均能获得合理、可靠的结果。

a) 闪避攻击防范：算法应有效防御对抗样本攻击，应将对抗样本加入训练数据集。

b) 物理对抗攻击防范：算法应有效防范物理对抗攻击，应在算法使用阶段添加对抗样本检测模块。

c) 模仿攻击防范：算法应有效防范模仿攻击，AI 算法性能最佳的条件下，当测试数据集中加入10%的模仿样本进行攻击时，攻击后与原算法差异低于阈值。

d) 逆向攻击防范：应防止信息过度反馈，设计文档制定了反馈信息的要求，在满足用户需求的前提下，遵循最小够用的原则，避免反馈信息过多，造成逆向攻击。

e） 应限制探测频率：设计文档对 AI 算法探测频率进行要求和设计，避免恶意探测获取 AI 算法信息，逆向构建恶意模型。

6.4 可持续性

大模型算法应具备可持续发展的特点，能够根据实际应用需求进行升级、优化和改进。

6.5 算法依赖

基于开源模型、第三方模型做出的模型，应进行安全性风险评价。

在开源模型、第三方模型基础上做迁移调优时，应进行安全性加固。

应对开源大模型框架以及依赖库的安全性进行评估，应定期开展算法所用开源框架以及依赖库的内部审计。

6.6 可追溯性

a） 训练数据可追溯性：应记录训练数据获取时间、训练数据来源、训练数据量、存储介质的标识、训练数据应有完整签名或校验码、采样方法

b） 建模过程可追溯性：应保存建模过程中的建模脚本、记录建模过程中的软硬件环境、记录建模过程中的操作者、记录建模的起止时间戳和迭代次数、保存建模过程中参数迭代的相关记录。

c） 算法部署可追溯性：应记录 AI 算法部署的操作者、标识部署时间及相关结果、保存 AI 算法部署过程的相关脚本、记录部署的软硬件环境配置信息。

6.7 算法内控

应建立算法上线应用前的内部评审机制。

应建立算法日常监测体系。

应建立算法退出处置机制。

应全面记录算法开发至退出的全过程。

7 安全规范

7.1 数据安全和防护

AI 应用需加强数据安全保护，包括数据加密、访问控制、风险评估等措施，防范数据泄露和滥用。

7.2 模型安全和防护

对 AI 模型进行安全性评估和防护，防止恶意攻击和破坏，确保 AI 应用的可靠性和稳定性。

应保障 AI 模型算法训练步骤安全：规划方案和实施方案中制定了 AI 算法训练过程安全保障措施。系统具有相应 AI 算法训练过程安全防护的配置，并有审计日志，保障训练步骤安全，防范训练过程被窃取，避免训练信息泄露。

7.3 用户隐私保护

加强对用户个人信息的隐私保护，确保用户自主控制权限和知情权。

7.4 恶意行为识别和过滤

针对恶意行为和信息进行及时识别、过滤和阻断，维护网络生态的健康和秩序，确保其能够抵御常见的软件漏洞攻击。

8 应用开发规范

8.1 业务应用合规要求

人工智能大模型应用当确保提供的产品和服务符合相关法律、行政法规的规定，确保产品和服务的功能安全、透明、稳定、可持续。

8.2 模型仓库管理

对模型进行管理，提供对模型的可回溯方面的功能、对模型版本进行管理，提供对版本可回溯方面的功能。

8.3 系统完整性与稳定性

评估模型系统的架构设计、代码质量和部署环境的安全性，并在高负载情况下保持稳定运行。

提供算法上线的相关监控的功能，包括但不限于功能的使用、性能情况等。

8.4 账号和日志管理

应用提供账号管理访问权限控制的功能、提供日志管理和审计的功能。确认只有经过授权的用户和服务才能访问模型资源，并且具备有效的权限管理和日志记录系统。

8.5 风险控制

应具备突发情况下，向用户充分提示 AI 算法的固有缺陷和使用风险况应急处理机制。应建立风险赔偿机制，具备 AI 算法道德风险防范机制。

8.6 开发安全规范

涉及金融业务的大模型场景，原则上禁止使用外部 API 调用的大模型服务，所有 AI 大模型应用必须进行本地部署，确保数据和模型的安全性。本地部署的大模型必须符合公司内部的安全规范，确保数据不外泄，模型安全性不受威胁。

大模型应用开发过程需同时符合公司开发安全管理，以及其他相关的编码标准和规范、安全部署等规范。