主题：大模型在证券行业的应用与探索

交易技术前沿

2025 年 第 2 期 总第 62 期

ITRDC 证券信息技术研究发展中心（上海）

上海證券交易所 SHANGHAI STOCK EXCHANGE

内部资料 免费交流
准印证号（K)0671

联系电话：021-68607130
021-68607129
投稿邮箱：ftt.editor@sse.com.cn

ITRDC证券信息技术研究发展中心（上海）

中国上海市杨高南路388号
邮编：200127
公众咨询服务热线：4008888400
网址：http://www.sse.com.cn

内部资料 免费交流

本资料仅为内部交流使用，本期印200册，编印单位为上海证券交易所，面向证券期货行业发送，印刷时间为2025年6月，印刷单位为上海华顿书刊印刷有限公司。

交易技术前沿

2025 年 第 2 期 总第 62 期

运营：
证券信息技术研究发展中心（上海）

主管、主办：
上海证券交易所

刊首语

今年以来，随着以 DeepSeek 为代表的开源人工智能大模型取得重大突破，证券基金行业的人工智能应用正以前所未有的速度不断涌现、演进与迭代，行业机构因地制宜积极探索适合自身发展的大模型建设与应用模式，已探索应用的场景不仅覆盖知识库问答、客户服务、代码助手等通用领域，也广泛覆盖资产管理、投研投顾、交易策略等垂直场景。毫无疑问，人工智能大模型技术已成为驱动行业机构未来数年数字化转型深入发展，促进行业金融科技创新最重要的引擎之一。

与此同时，大模型作为一种新兴技术，在为证券基金行业描绘了广阔的智能化前景的同时，其应用之路伴随着不容忽视的风险与挑战，这既包括数据安全与隐私保护，幻觉、不可解释等模型局限性，也涉及监管合规的复杂性，以及高昂的投入成本等。本期《交易技术前沿》以“大模型在证券行业的应用与探索”为主题，精选行业机构应用大模型技术的实践案例与经验分享，以期为行业机构安全合理有效应用大模型提供有益参考。其中：

中信建投证券《智能语言生成系统助力证券经纪展业的策略与实践》将大模型应用于证券经纪业务，构建财富管理智能客服助手，提供了实时、个性化的业务服务。

汇添富基金《推理大语言模型驱动资产管理智能化变革》针对以 DeepSeek 为代表的大模型构建了四维系统架构，打造相关智能应用，服务投研、市场及投顾等核心业务领域。

申万宏源证券《FinQueryGen：探索金融资讯数据的 Text-to-SQL 应用》利用大模型和 RAG 技术实现了从自然语言到 SQL 查询的高效转换，提高金融资讯数据处理效率和准确性。

国金证券《证券行业测评垂直领域大模型的体系构建与实践》构建证券行业问答数据集和一套多维度评价指标，并构建专有的测评体系供证券行业评估垂直领域的大模型性能。

上交所技术公司的《金融大模型安全风险与应对》设计了涵盖数据安全、模型安全、环境安全、攻防安全和安全管理等层面的金融大模型安全框架，以促进大模型在金融证券领域的生态体系构建。

证券信息技术研究发展中心（上海）

2025 年 6 月 16 日

目 录

01 本期热点

02 前沿技术应用

03 实践探索

04 监管科技全球追踪

01 本期热点

智能语言生成系统助力证券经纪展业的策略与实践

潘建东，马张晖，訾顺遥，尹序鑫，梁彬，王赵鹏，刘国杨，孙冰
中信建投证券股份有限公司｜ E-mail： mazhanghui@csc.com.cn

摘　要： 本文聚焦于证券经纪展业中自然语言生成（NLG）技术的创新应用。通过深入调研证券经纪展业的现状与发展动向，我们系统梳理了其主要业务场景。接着，通过引入 NLG 技术构建高效、精准的证券经纪展业自然语言生成系统，实现与常用工作平台无缝对接，提供实时、个性化业务服务。解决传统服务模式信息过载、效率低下等问题，推动了行业智能化升级，提升客户满意度。

关键字： 证券经纪展业、自然语言生成（NLG）、人工智能技术应用

一、背景

随着金融科技的不断演进，证券经纪业务正处在一个转型与升级的关键时期。传统的服务模式已难以满足日益增长的客户需求和市场变化，数字化转型和智能化升级成为行业发展的必然趋势。自然语言生成（NLG）技术，特别是大语言模型技术的快速发展，为证券经纪展业带来了新的发展机遇和挑战。大语言模型具有强大的语言处理能力和生成能力，能够根据用户需求灵活生成高质量的文本内容。这使得 NLG 系统能够更加精准地理解用户需求，提供更加个性化的信息服务。同时，大语言模型技术还能够实现多轮对话和语境理解，进一步提升用户交互体验。

然而，尽管大语言模型技术在多个领域取得了显著进展，但在证券经纪等特定企业垂直领域的系统性应用仍处于初级阶段，面临着专业领域数据稀缺、数据质量难以保证、知识幻觉、隐私安全等多重挑战。目前，大多数尝试仅限于细微场景的应用，远远无法满足证券经纪展业对于全面、精准信息服务的迫切需求。因此，本文研究旨在填补这一空白，通过构建创新的混合自然语言生成系统，推动大语言模型技术在证券经纪展业领域的广泛应用，以满足行业对于智能化、个性化服务的升级需求。

二、经纪展业 NLG 应用场景梳理

从应用场景的角度来看，NLG 技术在证券经纪业务中的应用主要可以分为面向客户和面向内部运营两大类。

1）面向客户的应用场景

智能开户与身份验证：NLG 技术可自动生成开户确认书等文档，简化流程，提升客户体验。结合大语言模型，还能高效完成身份验证与风险管理。

投资咨询与建议：依据客户个性化需求和市场动态，NLG 能生成易懂的投资建议报告，借助大语言模型的分析预测，为客户提供精准、有价值的投资建议。

交易助手与客服查询：NLG 可提供实时交易助手服务，如交易确认、账户变动提醒等。搭配大语言模型的智能问答系统，能快速、准确地解答客户咨询。

个性化投资分析与教育：结合 NLG 和大语言模型，证券公司可依据客户风险偏好、投资目标等，提供定制化的投资建议和市场解读，助力客户提升投资能力。

2）面向内部运营的应用场景

营销内容生成：NLG 可基于大语言模型的市场分析和客户画像，自动生成个性化营销话术与宣传材料，提高获客效率。

合规管理与风险预警：NLG 能实时监控业务，预警潜在风险，结合大语言模型的数据分析，实现精准风险识别与合规保障。

知识管理与业务流程自动化：NLG 可自动化更新知识库，结合大语言模型的流程优化能力，实现业务流程的自动化和智能化升级，提升员工工作效率。

市场调研与分析：NLG 可辅助证券公司进行市场调研，自动生成报告和分析结果，基于大语言模型的数据挖掘和预测，为战略决策提供支持。

资产配置与组合管理：结合 NLG 和大语言模型，证券公司可依据客户投资目标和风险偏好，提供定制化的资产组合建议和调整策略，实现精准的资产配置服务。

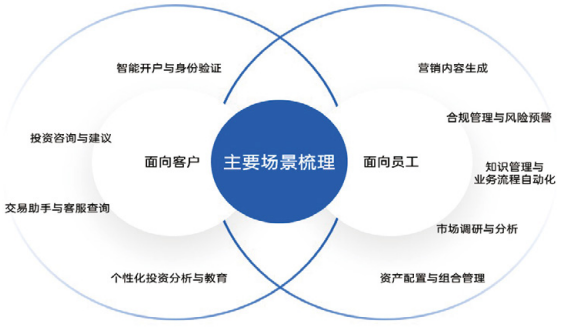


图 1 NLG 技术应用主要场景分类及特点梳理

三、面向经纪展业的 NLG 系统设计

面向经纪展业的 NLG 系统旨在构建高效、智能且用户友好的平台，为证券经纪业务关键环节提供支撑，赋能展业场景，推动业务智能化升级。我们提出创新的混合 NLG 系统构想，通过整合大语言模型与小模型、融入专家知识以及实现结构化任务调度等策略，有效解决证券经纪展业领域大语言模型面临的知识幻觉等挑战。这将为证券经纪服务带来前所未有的智能化提升，为客户提供更加卓越且风险可控的服务体验。

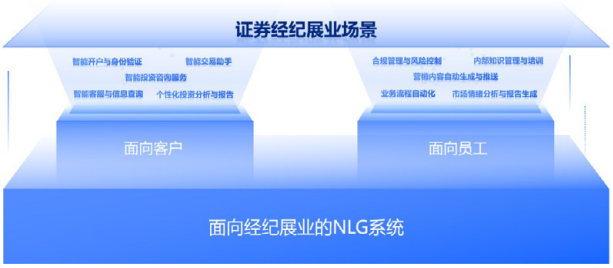


图 2 NLG 支撑证券经纪展业示意图

1）总体架构设计

面向经纪展业的 NLG 系统是综合性、智能化的自然语言生成平台。它以全面准确的数据资源为基础，借助大语言模型基础设施，深度学习海量数据并灵活应用。在模型层采用双轨并行策略，融合传统模型精准度和大语言模型泛化能力，满足经纪业务多场景文本生成需求。

该架构提供丰富 NLG 应用开发平台和工具，支持 AI Agent 快速开发部署，可自动化完成市场分析、投资建议、客户回复等任务，提升经纪业务效率和客户满意度。同时注重风险控制与安全机制，通过模型测试、审查、监控等确保文本生成合规准确，采用加密技术和访问控制保护数据安全隐私。

架构实现文本生成高效精准，具备良好的可扩展性和灵活性。模块化设计便于功能扩展升级，统一的大语言模型访问 API 层实现多模型统一访问切换，为经纪业务竞争提供有力支持。

2）NLG 系统流程

在证券经纪展业等垂直领域，NLG 技术尤其是大语言模型的落地面临重大挑战，其中最大的难题是知识幻觉。这主要是由于专业领域数据的稀缺性和数据质量的难以保证，加上模型对深入理解领域知识和高计算资源的需求，

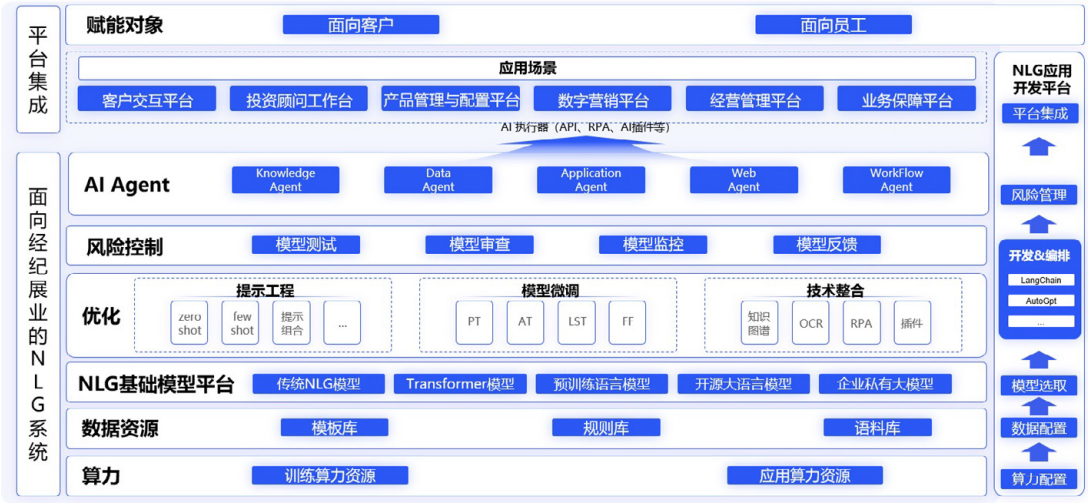


图 3 面向经纪展业的 NLG 系统总体架构示意图

使得问题更加复杂。隐私和安全性问题，以及多轮对话中的语境冲突，也增加了实际应用的难度。

为了应对这些挑战，我们可以构建一个创新的混合 NLG 系统。这个系统采用灵活的组合方式，将大语言模型和小模型相结合，基于 ReAct 框架和 CoT 思维链进行 Multi-Agent 编排调度。通过整合多个专业子 Agent，我们可以提高系统的整体性能和准确性。这些 Agent 的核心在于大型语言模型（LLM）与各种工具的协同工作，大型语言模型通过理解用户任务来推理需要调用的工具，并根据调用结果向用户提供反馈。在完成任务的过程中，Agent 可能会与用户进行多轮交互，以优化结果。

此外，为了强化基础模型的能力并提升推理准确性，融入专家知识。通过结合财富管理领域的专业知识，我们可以形成一个更精准的智能化推理分析引擎，从而提高系统的泛化能力和生成准确交互信息的能力。

同时，结构化任务调度和借鉴成熟方案也是确保模型

在实际应用中高效和灵活的关键。通过结构化的任务调度器，我们可以对交互过程中生成的结论进行结构化还原，并将其转化为可用于小模型引擎和专家规则引擎的任务。借鉴证券经纪展业领域的成熟方案，我们可以更快地实现规则引擎的构建和优化。

通过构建一个创新的混合 NLG 系统，整合多个专业子 Agent，强化基础模型能力并融入专家知识，以及实现结构化任务调度和借鉴成熟方案，我们可以有效缓解证券经纪展业领域大语言模型的幻觉问题。这将为客户提供卓越且风险可控的证券经纪服务，推动 NLG 技术在该领域的广泛应用。

3）模块组成

NLG 系统由多个核心组件构成，它们相互协作，共同实现对话式智能化服务。以下是这些核心组件的简要介绍：

大语言模型：作为框架的核心，此模型基于通识大语言模型进行微调，具备强大的自然语言处理和理解能力。它能够解析用户的意图和需求，生成自然流畅的回复，并与用户进行持续的文字或语音对话。

子 Agent 集合：为了更好地处理特定场景和任务，系统中包含多个专业的子 Agent。这些子 Agent 分别负责处理不同类型的任务，如投资咨询、账户管理、交易执行、风险评估等。每个子 Agent 都具备特定领域的知识和技能，能够独立完成相应任务，并将结果反馈给大语言模型和用户。

任务调度器：此组件负责根据用户的意图和需求，将任务分配给最合适的子 Agent 或规则引擎进行处理。它能够根据子 Agent 的专长、负载情况和历史表现等因素进行智能调度，确保任务能够高效、准确地完成。同时，任务

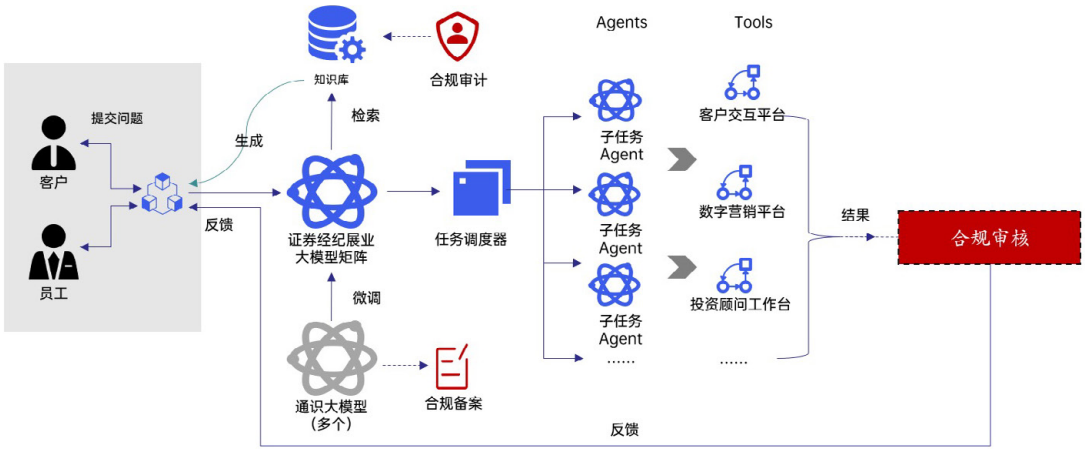


图 4 面向经纪展业的混合 NLG 系统流程示意图

调度器还支持动态添加或删除子 Agent，以适应不断变化的应用场景。

知识库：知识库存储了大量的相关知识和经验，包括市场动态、投资策略、产品信息以及历史案例和解决方案等。它为子 Agent 和大语言模型提供数据支持，帮助它们更好地理解和处理用户需求。

用户交互界面：此界面支持自然语言输入和输出，使用户可以通过语音、文字等方式轻松与系统进行沟通。

这些组件共同协作，为用户提供智能化、高效、便捷的客服服务体验。

4）合规审计

构建混合 NLG 系统时，合规性至关重要，关乎法律遵循、道德标准、用户信任和企业声誉。构建过程中需注重合规应用，具体措施如下：

选取通用大语言模型时，首要任务是检查其是否已进行合规备案。合规备案表明模型经监管机构审查，符合数据隐私、知识产权、伦理道德等法律法规要求，可降低系统运营中的合规风险。

混合 NLG 系统需结合企业知识库生成专业文本，故对知识库进行合规审计很重要。审计要关注数据来源合法性、内容敏感性及侵权风险、是否违反法律法规等，确保知识库内容合法合规，为系统运行提供基础。

即使通用大语言模型备案、知识库审计通过，系统生成文本仍需合规审核。因模型输出可能受输入影响而不可预测，含不合规内容。审核可采用自动化工具与人工审核结合方式，提高效率和准确性。

四、应用及实践—财富管理智能客服助手

在财富管理领域，为了提供更加高效、精准和个性化的客户服务，我们提出了一个基于经纪展业 NLG 系统的多场景财富管理智能助手。借鉴 GPT 的集成学习思想，通过

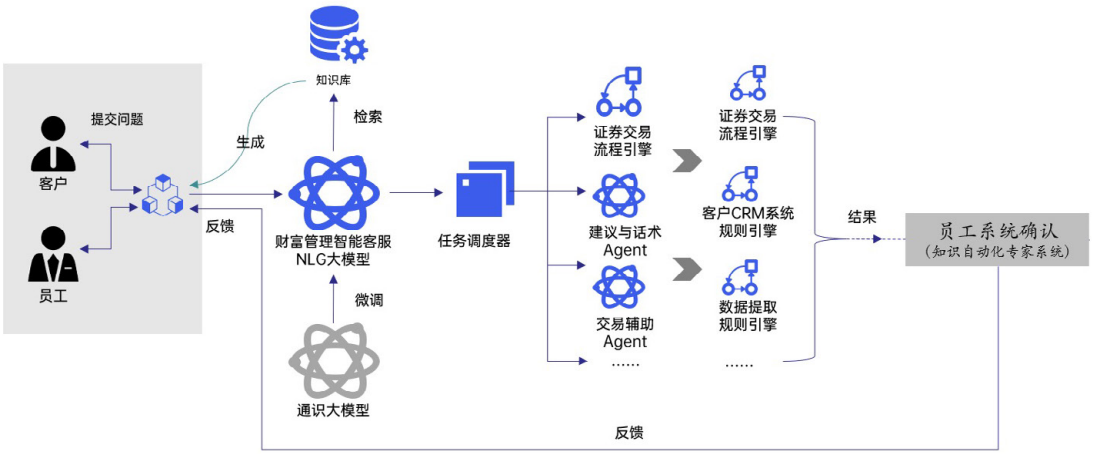


图 5 财富管理智能客服助手示意图

整合多个专业的子 Agent，构建了一个强大的混合专家系统，实现了多个智能体 Agent 的协作与动态编排调度，并赋予了计划、记忆、反思与推理等能力。

为了构建财富管理智能助手，并实现上述功能，设计以下独立的 AI Agent 组合来达到更好的效果：

身份验证 Agent：负责验证客户的身份，确保客户具备交易权限。通过多因素身份验证（如密码、生物识别、安全问题等）来确认客户身份。

数据提取与分析 Agent：从后端系统中提取客户的投资偏好、历史交易数据、市场情况等信息。对提取的数据进行预处理和分析，以识别客户的投资风格、风险承受能力等。

建议与话术 Agent：基于数据分析结果，生成个性化的投资建议和沟通话术。结合客户的投资目标和当前市场状况，定制适合客户的投资策略建议。

产品检索与解释 Agent：针对客户咨询的产品或服务，检索相关信息（如产品详情、费用、风险等级等）。提供易于理解的解释和比较，帮助客户了解产品的特点和适合性。

交易辅助 Agent：根据客户的投资目标和风险偏好，提供交易策略建议。帮助客户评估潜在收益与风险，并提供决策支持工具（如模拟交易、风险评估问卷等）。

交易处理 Agent：接收并处理客户的交易请求（如买入、卖出、转账等）。验证请求的合法性和客户的交易权限，然后将请求发送至后端系统进行执行。

结果反馈 Agent：接收后端系统的交易结果，包括成交状态、交易明细等。以直观、易于理解的方式将交易结

图 6 财富管理智能助手示意图

果反馈给客户，并提供必要的后续操作指导。

这些 AI Agent 可以独立开发、部署和管理，通过协同工作来实现财富管理智能助手的整体功能。每个 Agent 专注于处理特定类型的任务和数据，提高了系统的模块化、可维护性和扩展性。

示例场景：

张先生通过财富管理智能助手的电话客服系统，成功地对自己的投资组合进行了调整。在通话中，他首先完成了身份验证，随后财富管理智能助手的 Agent 根据他的投资历史和风险承受能力，提出了分散投资风险的建议。当张先生表示对黄金投资感兴趣时，Agent 立即为他提供了详尽的黄金投资产品信息。在 Agent 的协助下，张先生顺利完成了黄金 ETF 的购买，进一步优化了自己的投资组合。整个电话客服过程中，财富管理智能助手的 Agent 们紧密协作，为张先生提供了既全面又专业的投资服务。

应用关键技术创新如下：

基于 ReAct 框架的任务动态编排调度：财富管理智能客服 NLG 大语言模型基础模型负责解析客户意图，并根据意图将任务分配给相应的 Agent 进行处理。在财富管理智能客服框架中，我们可以定义多个数据源 Agent，分别负

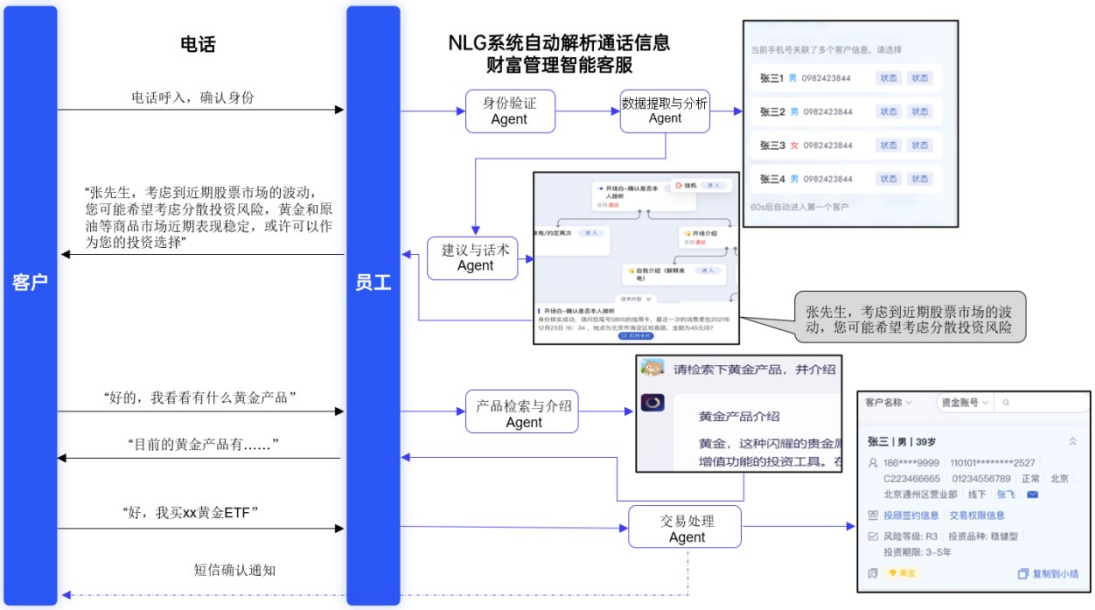


图 7 场景应用示意图

责处理不同类型的客户数据（如交易记录、投资偏好、市场动态等）。通过多 Agent 的协同工作，系统能够综合考虑多个数据源的信息，对客户的需求进行更全面地理解和分析。这种动态编排的方式使得系统能够灵活应对各种复杂的客户场景，提高服务效率和质量。

基于 RAG 检索增强生成的需求推理与交互：在财富管理领域，客户的需求往往涉及大量的专业知识和市场动态。为了更准确地理解客户需求并提供满意的解决方案，我们可以利用 RAG 检索增强技术来辅助大语言模型进行推理。通过构建包含专家经验、历史案例和市场动态等知识的知识库，并将这些知识转化为向量形式存储在向量数据库中，我们可以使用大语言模型对客户的需求进行检索和匹配。系统就能够快速找到与客户需求相关的专业知识和解决方案，提高响应速度和准确性。

沿着客户全生命周期管理，提供多种服务能力：财富管理智能客服助手具备处理客户全生命周期内各种需求的能力。这包括但不限于投资咨询、账户管理、交易执行、风险评估等。通过将各种服务能力集成到各种 Agent 中，并使用自然语言交互方式与客户进行沟通，可以为客户提供更加便捷和高效的服务体验。同时，这也降低了客服人员的工作负担，提高了工作效率。

将 NLG 技术引入财富管理智能助手领域，可以解决传统财富管理智能助手的一些痛点问题，并具有以下优势：

首先，随着财富管理行业的快速发展，各种金融产品和投资策略层出不穷，很难有一个团队能够精通全部领域的知识和技术细节。而 NLG 技术具备强大的学习能力，可以不断吸收和更新财富管理领域的知识，通过设计多个专家 Agent 的方式进行编排调度，实现对不同金融产品和投资策略的全面覆盖。同时，NLG 技术可以将多源、异构的数据转化为统一的语言形式，进行深度分析和诊断，提升

财富管理决策的效率和准确性。

其次，传统的财富管理智能助手算法往往针对特定场景和问题进行设计，且需要大量标注数据进行训练。而大语言模型基于检索增强的方式，可以充分利用无标注或少量标注的数据进行训练，降低了数据标注的成本和周期，同时提高了模型的泛化能力。这使得财富管理智能助手可以更加灵活地应对各种场景和问题，快速适应市场的变化。

此外，传统的财富管理智能助手在遇到新客户、新业务、新数据等情况时，通常需要进行大量的定制和重新训练。而大语言模型具备强大的泛化能力和自监督学习能力，可以通过简单的微调或增量学习来适应新客户和新业务的需求。同时，大语言模型的交互形式使得理财规划师和客户可以实现一定程度的松耦合，降低了人工维护客户的成本。

图 8 财富管理智能助手截图

问题	优化前	优化后
如何提取我在证券账户中的资金？	请按照以下步骤操作： 登录到您的证券经纪账户。 导航至“资金管理”或“我的资产”部分。 选择“提取资金”或类似的选项。 输入您希望提取的金额，并选择提取方式（如银行转账）。 确认提取请求，并等待资金到账。	若您希望从证券账户中提取资金，请遵循以下指导步骤以确保操作的顺畅： 步骤一：登录与验证 使用您的证券账户凭据（用户名/密码或生物识别）登录到证券经纪应用或网站。 完成双重身份验证，确保账户安全。 步骤二：导航至资金管理 在应用的主界面或侧边栏中，找到并点击“资金管理”或“我的资产”板块。 这里会显示您的总资产、证券持仓、现金余额等详细信息。 步骤三：发起提取请求 在资金管理页面中，找到并点击“提取资金”或“提现”按钮。 输入您希望提取的具体金额。请确保该金额不超过您的可用现金余额。 步骤四：选择提取方式与账户 从下拉菜单中选择您的提取方式，例如银行转账、电子钱包等。 如果您之前未绑定过提取账户，您需要先进行绑定操作。 对于银行转账，请确保您提供的银行账户信息准确无误。 步骤五：确认与提交 仔细核对提取金额、提取方式及账户信息，确保一切正确无误。 阅读并同意相关的服务条款和费用（如有）。 点击“确认”或“提交”按钮，完成提取请求。 步骤六：等待与确认 提取请求提交后，资金通常会在1-3个工作日内到达您的指定账户，具体时间取决于银行处理速度和节假日等因素。 您可以在“资金管理”或“交易历史”中查看提取状态和详细信息。 请注意，某些证券经纪公司可能会对资金提取设置限制或收取费用。在进行提取操作之前，建议您先查阅相关的帮助文档、FAQ或联系客服以获取更多信息。

图 9 财富管理智能助手优化提升效果

五、总结与展望

本文致力于构建一个高效、精准的 NLG 系统，以深度融合大语言模型技术，满足证券经纪业务在智能化服务升级过程中的迫切需求。通过引入自然语言生成技术，为证券经纪业务提供了全新的服务模式和手段，有助于打破传统服务中的信息不对称和壁垒，使得更多投资者能够享受到专业化、个性化的投资咨询服务。展望未来，随着人工智能技术的不断进步和应用场景的不断拓展，面向经纪展业的自然语言生成系统将迎来更加广阔的发展空间和应用前景。具体而言，可从以下几个方面展开：

1）模型优化与升级

持续关注大语言模型技术发展，优化模型结构和参数，提升生成质量和效率。探索跨领域知识融合，增强系统通用性和适应性，为投资者提供更精准、个性化服务。

2）多模态信息生成

结合多模态技术（如视频生成、虚拟现实），生成个性化投资演示视频或提供沉浸式投资体验，帮助投资者直观了解投资信息，提升决策信心。推动业务向智能化、高效化发展。

3）应用合规与数据安全

目前系统主要用于内部赋能，已建立完善的数据安全和合规审查机制。未来若涉及对外服务或引入外部数据，需严格遵守模型备案和数据合规要求，确保业务稳健发展。

推理大语言模型驱动资产管理智能化变革

周建军，庄明光，余哲，涂鹏
汇添富基金管理股份有限公司 | E-mail : yuzhe@htffund.com

摘　要： 随着推理大语言模型（如 DeepSeek-R1）在性能上的突破，资产管理行业的 AI 应用正在从局部试点迈向全业务链赋能。汇添富基金构建包含资源层、模型基座层、AI 平台能力层及应用场景层的四维系统架构，加速推理大语言模型智能应用在投研、市场及投顾等核心业务领域的规模化落地与应用。本文提出的智能研报洞察智能体实现了信息获取效率与投资决策质量的协同提升。智能投顾智能体基于基金数据、用户画像与市场动态提供了研究、策略与顾问一体化的智能投顾解决方案。随着对全业务链路的深入赋能，推理大语言模型有望驱动资产管理行业的智能化变革，为投资者创造更大价值。

关键字： 资产管理；推理大语言模型；智能体；研报解析；基金投顾

一、引言

2023 年大语言模型 ChatGPT 发布，其强大的语义理解与文本生成能力在全球掀起了大语言模型浪潮。在证券行业，大语言模型已得到试点应用，典型应用场景有代码辅助、客服助手、业务流程办理、投顾策略等 [1-4]。资产管理行业也有迫切大语言模型应用需求，但当前应用多限于局部试点，未能引起行业深度变革。主要原因在于大语言模型能力不完善，存在生成质量不稳定、幻觉突出、缺乏逻辑推理能力、专业领域适配不足等问题。以 OpenAI 的 o1 及 o3 模型为代表的推理大语言模型性能强，但资产管理行业对数据有极高的安全要求，核心业务无法调用外部大模型的 API 服务。开源模型能力有限，缺乏逻辑推理与数理处理等能力。

2025 年 1 月，深度求索公司开源了大语言模型 DeepSeek-V3 和 DeepSeek-R1，在模型性能与推理成本控制上实现了双重突破。该系列大语言模型采用了混合专家系统架构设计，在参数量达到 671B 的同时，算力需求相比传统稠密模型降低 90%[5-6]。其中推理模型 DeepSeek-R1 达到了世界领先水平，能够通过思维链多步推理处理复杂金融数据任务。其开源属性使金融机构能够实现超大规模模型部署的自主可控性，满足资产管理行业对敏感数据保护及系统安全的严格要求。推理大语言模型在资产管理行业的应用有望从技术探索和局部试点迈向全场景业务的规模化赋能。

二、推理大语言模型在资产管理行业应用

随着模型能力的持续提升，推理大语言模型的技术应用正经历从基础效率优化向业务价值深度挖掘的战略转型。在资产管理行业专有数据资源的支撑下，推理大语言模型技术预期将实现全业务链路的深度渗透，可覆盖投研、市场及投顾等核心领域，部分典型应用场景如图 1 所示。

图 1 推理大语言模型在资产管理行业典型应用场景

在投研领域，全生命周期覆盖事前研究、投资决策、交易执行及风险管理等核心环节。在研究环节中，大语言模型可赋能研究信息处理效率提升、智能检索问答、低频因子挖掘；在投资决策环节，模型可辅助组合优化管理、投资约束条件监测、交易行为模式解析、业绩归因模型构建、风险实时预警；交易执行环节，大语言模型可提升询价报价效率，完成交易序列分析与异常波动监测；风险管理及事后分析等方面，大语言模型能够辅助流动性与信用风险分析、舆情监控、合同审核、报告生成、指标问答等。在资产管理市场运营领域，大语言模型展现出多维应用价值：（1）业务流程方面，依托大语言模型深度语义理解能力，实现金融资讯智能生产、服务流程自动化设计及潜在销售机会挖掘；（2）客户服务智能化方面，通过整合客户渠道特征数据与需求预测数据，生成个性化资产配置方案，并依托全天候智能咨询系统实现投资全生命周期服务覆盖；（3）运营管理优化层面，模型可融合实时市场数据与历史回测数据，自动化输出市场趋势研判报告、竞争产品结构分析及系统性风险预警内容，支持客户投资策略的动态优化。推理大语言模型核心价值体现在通过数据

驱动智能运营体系，显著提升运营效率、服务精准度及客户满意度，推动资产管理行业向客户服务智能化、营销运营精细化及管理分析价值化方向转型。

在投顾领域，推理大语言模型同样可以赋能多方面应用。例如，（1）基金研究端：通过智能问答提供基金筛选与指标可视化研究；（2）策略生成端：通过实时市场数据流与历史策略知识库的深度整合，构建动态产业链图谱，提供多因子基金投资建议，建立市场异动监测预警系统，并生成策略优化方案；（3）客户服务端：基于客户画像系统与行为预测模型，实现资产配置方案的动态匹配和优化，通过智能问答看板提供市场需求结构分析、竞品解析及营销策略回测支持等功能。

三、汇添富基金推理大语言模型应用实践

汇添富基金近年来持续加大人工智能（Artificial Intelligence, AI）技术研发投入。在 ChatGPT 模型发布后，公司迅速成立 AI 项目委员会，试点推进大语言模型在资产管理行业的应用。随着大语言模型的迭代升级，特别是推理大语言模型 DeepSeek-R1 发布，资产管理行业内大语言模型的应用正从局部试点向全业务链渗透演进。汇添富也建立 AI 敏捷组织架构及工作机制，组建以互联网金融“DeepFund”团队为代表的 AI 组织。2025 年 2 月率先完成 DeepSeek-R1 的企业级私有化部署，并系统性推进投资研究、产品销售、风控合规及客户服务等核心业务的智能化升级。

3.1 系统架构

为促进以 DeepSeek-R1 为代表的推理型大语言模型在资产管理行业的实践应用，驱动核心业务效能优化与创新潜能挖掘，汇添富基金设计并构建了一套层次化、模块化的 AI 系统架构体系。如图 2 所示，依次为硬件资源层、模型基座层、AI 平台能力层及应用场景层，最终支撑覆盖多业务场景的智能化解决方案。

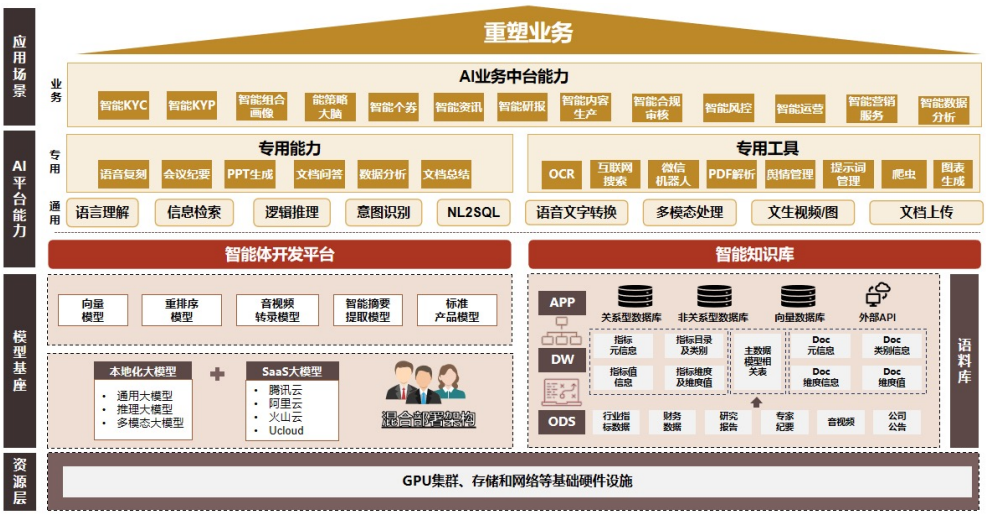


图 2 汇添富 AI 系统架构

资源层作为 AI 应用体系的资源底座，提供 GPU 集群、存储和网络等基础硬件设施。通过建设高性能算力中心，不仅确保 AI 服务的高效稳定运行，也能够同时满足资产管理行业对数据安全、隐私保护及合规性的严格要求。

第二层为模型基座层，包括 AI 模型与语料库两个方面。在 AI 模型选型与部署层面，通过构建包含通用问题及资产管理领域专业问题的评测数据集，系统性评估大语言模型在通用认知与专业领域的性能表现。针对超大参数量的推理大语言模型，我们在并发性能、响应延迟、上下文窗口长度等方面进行压力测试并优化，选择最合理的部署方式，实现与业务场景的高度适配。对于特定业务需求场景，包括意图识别、结构化信息抽取、元数据提取等应用，采用小参数量大语言模型微调方案，在确保任务精度的前提下实现计算资源的优化配置。语料库为 AI 应用提供高质量生产资料支撑。尽管推理大语言模型在通用领域展现出卓越的知识覆盖与推理能力，但受限于训练数据的时效性与领域专业性，在资产管理等垂直领域应用中存在显著局限性。我们通过制定数据治理体系，整合多源异构数据，补充与业务场景适配的动态知识，有效提升推理大语言模型在资产管理领域应答的准确性，释放其实战价值。语料

库的架构上，提供多样数据采集方式，获取公司内部各系统数据，确保数据的准确性、全面性和及时性。对采集到的原始数据进行清洗处理，转换成推理大语言模型可精准理解并处理的格式，包括文本结构化、数据结构标准化等。最终通过关系型数据库、非关系型数据库、向量数据库、外部 API 等方式为 AI 平台能力层提供全域数据视图。

AI 平台能力层为上层应用场景提供智能化支撑，如表 1 为其典型 AI 基础能力要求。智能知识库应用场景中，非结构化文本解析能力是基础，可以采用 OCR 和多模态大语言模型进行信息提取和预处理。检索召回通过文本搜索、向量检索、知识图谱检索和数据查询等方式提升信息检索的完整性与准确性。在跨系统交互方面，通过标准化 API 接口、工具调用、模型上下文协议（MCP），实现跨系统协同与数据互通。在智能体开发方面：提供流程编排开发方式，业务人员通过可视化界面配置对话机器人、流程自动化等智能体，有效降低 AI 技术应用门槛，加速智能体从局部试点向规模化生产转型；对于复杂智能体需求，提供深度开发的方案；智能体协议（Agent to Agent Protocol）支持多智能体间的连接与交互。

应用场景层是 AI 在资产管理业务中的实践落地，AI 智能体应用对多样金融数据进行分析和挖掘，生成涵盖多个关键业务领域的高质量数据，提高生产内容效率与质量，满足各业务多样化的内容需求。通过接入研究报告、自选仓位、实时行情等数据，依托产业链图谱，智能体能够评估投资风险、挖掘投资机会。接入网络热点数据，结合用户画像和产品特点，生成智能营销策略和个性化推荐方案。基金投顾方面，基于 KYC（Know Your Customer）、KYP（Know Your Products）数据，结合市场行情，利用推理大语言模型进行资产配置和基金推荐，提供智能投资组合策略。

表 1　AI 基础能力表

AI 能力大类	能力点	工程要点与难点
大语言模型	提示词工程	简洁、准确
	小参数量大语言模型	训练、微调、评测、部署与应用
	大参数量大语言模型	评测、部署与应用
多模态	语音转文字	专用短语识别
	文字转语音	音色克隆、情感语气模拟
	图片、视频识别与生成	生成稳定性与内容可靠性
智能知识库	非结构化文本提取	格式识别、版面识别、文本切块
	检索召回	文本检索、向量检索、知识图谱检索
跨系统交互	API 调用	权限与数据管控
	工具调用	清晰的任务分解
	模型上下文协议（MCP）	安全管理、上下文管理、决策管理
	自然语言转 SQL	准确率与校验方法
智能体	低代码智能体平台	易用性、兼容性与可扩展性
	深度开发方案	实现复杂任务的规划与多模态协同
	智能体协议（A2A）	决策管理与可信度管理

3.2 应用案例

汇添富基金在大语言模型推动业务变革方面持续投入，已上线多项创新应用。投资业务场景中，存在债券台账及存款报价结构化解析需求，传统基于正则表达式的解析方式识别准确率低。通过对 1.5B 参数规模的大语言模型进行微调，我们将存款报价解析准确率从 60% 提升至 77%，债券台账解析准确率从 92.7% 提升至 96.6%，2024 年上线以来平均每日解析任务超 400 条。同期上线的代码助手，平均每日调用 1325 次，已在代码开发、注释生成、代码解释、代码检查等环节形成有效辅助能力。近期，随着推理大语言模型能力的提升，汇添富已开发出多个基于 DeepSeek-R1 的智能体应用，有效赋能公司业务。直销平台“现金宝”App 上线了基于 DeepSeek-R1 模型的智能体服务，提供基金报告智能总结、对比分析与趣味化改写功能，将专业内容转化为投资者友好型解读，显著降低投资者信息获取门槛，并通过“原声朗读”增强投资者与基金经理的互动体验。基于智能研报解析的投资洞察智能体，及投研、策略与顾问一体化的智能投顾架构详解如下。

3.2.1 基于智能研报解析的投资洞察智能体

在投资业务中，基金经理日均接收研究报告邮件超过

200 封，涵盖个股研报、行业分析、宏观策略、财经资讯等多源异构数据，信息过载问题严重。传统人工处理方式下，基金经理难以精确获取高价值信息，致使投资决策效率受到显著制约。针对该业务痛点，本文提出基于智能研报解析的投资洞察方案，通过大语言模型解析研报邮件，精准提取核心观点与关键数据，识别潜在风险与机会，并按照预设逻辑生成结构化报告和开放性投资机会洞察，为基金经理提供决策参考。

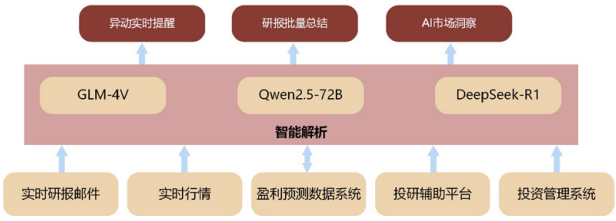


图 3 基于智能研报解析的投资洞察智能体架构图

智能体架构如图 3 所示，系统从多渠道获取源数据。研报邮件为解析源文件；实时行情系统用于个券实体及行情获取；盈利预测数据系统获取历史研报盈利预测数据并回流最新预测数据；投研辅助平台及投资管理系统获取自选订阅股票池，配置订阅规则等。

智能解析模块由文本解析及多模态解析两大功能单元构成。图 4 展示了基于多模态技术解析图片中盈利预测数据的流程图，整合了传统光学字符识别（Optical Character Recognition, OCR）预处理与多模态大语言模型的技术优势。首先获取研报邮件所有附件，包括盈利预测图片、网页格式图片、券商标识图等。随后，所有图片经过多模态嵌入模型，得到图片的向量值，通过向量相似度匹配识别盈利预测图片，有效排除其他图片的干扰。在图像处理阶段，先进行二值化处理，再通过联通区域检测算法，确认表格边界，完成图像区域分割。分割后的图片经过提示词工程输入多模态大语言模型，实现结构化盈利预测数据的高效抽取。实验对比发现，不经过图片预处理，GLM-4V 多模态大语言模型盈利预测数据提取准确率为 56%。经过图片预处理后，外文研报数据提取准确率提升至 95%，中文研报数据提取准确率达到 98%。Qwen2.5-72B 大语言模型在解析速度和准确率上表现均衡，承担实体和文本关键信息抽取任务；DeepSeek-R1 则用于研报翻译、研报总结与 AI 投资机会洞察。智能解析模块，具有模型中立性，能够轻松切换大语言模型，可以在准确率、效率、资源利用率等方面权衡，组合使用不同的模型。当前智能解析仍需预处理步骤，随着多模态模型能力的演进，端到端解析将成为趋势。

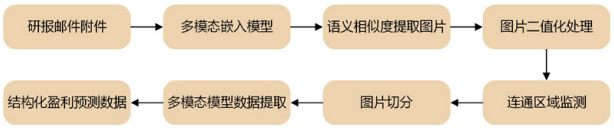


图 4 盈利预测数据智能解析流程图

经过智能解析，智能体最终输出三部分重要结果：异动实时提醒、研报智能摘要和 AI 市场洞察。实时异动提醒动态追踪盈利预测数据变动及关联市场行情波动，实现面向基金经理的风险及机遇的即时提醒；研究报告智能摘要集成多源异构研究报告的重点内容，为基金经理提供结构化研报全景视图，助力其把握市场整体态势与研究重点；AI 市场洞察，推理大语言模型自主推演市场动态与挖掘投资机会。

自上线以来，智能体运行稳定高效，日均处理研报邮件达 217 封，覆盖国内外主流券商的研究报告。智能体可以从 26% 的研报邮件中自动抽取个券盈利预测指标，包括评级、每股收益、营业收入、净利润等核心数据。对于所有研究报告，智能体自动生成摘要，压缩冗余信息并保留核心观点。推理大语言模型结合当日研报数据与市场动态，提供投资建议。经过结构化处理的数据可以实时或定时推送至订阅服务的基金经理，显著提升了研报信息获取效率，帮助用户节省 2 小时 / 日的研报阅读时间，支撑投资决策效率与质量的双重提升。

3.2.2 投研、策略与顾问一体化的智能投顾方案

基于资产管理行业专业金融数据体系，涵盖基金指标、市场行情和客户标签等多维数据，在推理大语言模型的加持下，我们提出“智能投研 - 策略中枢 - 情感化顾问”三位一体智能投顾解决方案。系统架构如图 5 所示，自下而上实现数据驱动与智能投顾的深度融合。

基础数据层可分为四个大类：基金基础数据，包括公司与全市场基金的元数据及指标数据；组合数据，包括组合成分与绩效归因指标等；资讯数据，包括热点资讯、行情数据等；客户特征数据，包括风险偏好、交易行为、投资偏好等。经过 AI 平台赋能，智能体可提供四维专业服务。首先是基金研究：基于公司内 KYP 指标搭建基金研究助手，便捷查询和维护公司基金产品数据；接入全市场基金的元数据与指标数据，实现投顾基金池深度研究，包括相似基金筛选与多维指标比对等深度分析等。第二部分是组合与策略研究：依托原有基金投顾组合分析体系，搭建了投顾组合研究助手，实现基础问答与组合偏离度计算等智能服务；依托基金指标、行情与资讯数据源，实现智能策略生成，包括历史策略优化匹配方案及推理大语言模型自主创新方案，助力投顾策略从“标准化”转向“个性化”，实现千人千面的策略覆盖。在顾问服务方面：提供市场热点解读、

专业投顾报告生成、投资观点萃取与营销文件生成等智能服务。在事后数据分析方面：智能体提供组合数据分析，包括成分、收益、波动、贡献等指标，并支持智能生成组合分析报告，助力投顾策略迭代升级；智能体同样支持营销数据分析，提供保有量、资产规模、客户分布、资金进出等销售主题数据分析，支持投顾组合维度的结构分析、产品定位与风险识别。

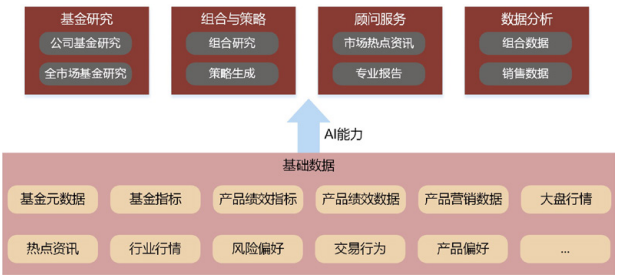


图 5 投研、策略与顾问一体化的智能投顾

智能投顾服务可提供文本、图表、音视频等多样化的内容生成，上线以来，取得了良好的业务效果。以基金研究服务为例，用户可通过自然语言交互生成基金多维可视化报告，其可溯源特性确保推理大语言模型生成内容的可信性，图 6 为其内容生成示例。相比传统的数据提取与分析流程，基金研究智能服务提供了新的基金研究范式，在研究效率、用户体验、报告生成等方面具有明显优势。

图 6 智能投顾内容生成示例

四、总结与展望

推理大语言模型正推动资产管理行业向高效化、个性化和智能化方向转型。汇添富基金已建立 AI 应用系统架构，加速推理大语言模型在投研、市场、投顾等业务领域的深入应用，目前已落地多个智能应用。智能研报洞察智能体显著提升了研报信息获取效率，并提供投资建议，辅助基金经理投资决策。投研、策略与顾问一体化的基金投顾智能应用，驱动基金、客户及市场数据与投顾策略的深度融合，有望实现机会捕捉、风险控制与客户适配的三维平衡。推理大语言模型正在驱动资产管理行业从信息化、数字化，迈向智能化，有望推动资产管理的范式重构，为投资者创造更大价值。

未来，推理大语言模型在资产管理行业的应用仍面临诸多挑战。推理大语言模型幻觉问题较非推理模型更为严重 [7]，资产管理行业需要尽量避免幻觉输出。金融数据指标口径复杂多样，目前推理大语言模型在多表关联查询场景中的准确率仍有待提高。

参考文献：

[1] 王洪涛 . 证券行业大语言模型优化方法与应用示范 [J]. 交易技术前沿，2024，56：02-10.

[2] 邓维，易卫东 . 人工智能大模型在证券行业应用路径与实践 [J]. 交易技术前沿，2024，56：21-25.

[3] 陶剑峰，张蕾，贾振兴 . 数字金融 - 中信建投证券投顾大模型推动业务创新 [J]. 交易技术前沿，2024，58：02-07.

[4] 赵岩，杨彬，夏杨铭 . 基于大模型的证券业务办理助手探索与实践 [J]. 交易技术前沿，2024，58：22-30.

[5]DeepSeek-AI, et al. DeepSeek-V3 Technical Report[R]. arXiv:2412.19437, 2025 February.

[6]DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[R]. arXiv:2501.12948, 2025 January.

[7]Vectara. DeepSeek-R1 hallucinates more than DeepSeek-V3[ER/OL]. https://www.vectara.com/blog/deep seek-r1-hallucinates-more-than-deepseek-v3. Vectara, 2025 January.

FinQueryGen：探索金融资讯数据的 Text-to-SQL 应用

褚丽恒，李楠，陈力，张进，东晓亮
申万宏源证券有限公司｜ E-mail： chuliheng@swhysc.com

摘　要： 在金融领域数据量呈爆炸式增长的背景下，传统数据处理方式在效率与准确性方面面临严峻挑战。本研究聚焦于大语言模型（LLM）与检索增强生成（RAG）技术相结合驱动的 Text-to-SQL 技术在金融资讯数据中的应用。本文详细阐述了 LLM、RAG 技术及 Text-to-SQL 的概念与发展历程等相关技术原理，通过构建金融资讯知识库，运用 RAG 技术结合提示工程提出了 FinQueryGen 模型，实现了自然语言到 SQL 查询的转换。FinQueryGen 模型与 Q-sql * 平台深度融合，协同增效显著，有效探索了 Text-to-SQL 任务在金融资讯数据场景中的应用潜力，为金融数据处理提供了创新途径。

关键字： LLM（大语言模型）；RAG；Text-to-SQL；金融资讯数据

一、引言

金融领域一直以来都面临着处理大量复杂数据的挑战，金融机构为实现精准的风险控制、合理的资产配置以及严格的合规管理，对金融数据处理的要求达到了极高的程度。传统的金融数据处理特别是复杂的查询分析往往依赖人工，需要大量专家经验，这限制了决策分析的效率，还容易出现错误。金融科技公司的崛起和传统金融机构的数字化转型加剧了金融行业的竞争，金融从业人员对高效工具的需求愈发迫切，如何实现降本增效成为行业的共同课题。随着人工智能技术的不断发展，技术的壁垒正在逐步消解，使用自然语言进行数据分析、文本理解、文本生成等工作从探索向应用迈进。

近年来，LLM 与 RAG 技术取得了突破性进展，为金融数据处理带来新机遇。LLM 如 ChatGPT 等展现出强大的自然语言处理能力，基于大规模语料库预训练能理解和生成复杂的文本内容。RAG 技术通过检索增强生成可以有效弥补 LLM 在知识更新及时性和解释性方面的不足，为 LLM 提供了实时、专业的知识支持，有效减少了模型 “幻觉” 现象，提升生成结果的准确性和专业性。在金融领域，这些技术的结合有望实现从自然语言到 SQL 查询的高效转换，从而大大提高金融数据处理的效率和准确性。

二、研究目标

本文旨在深入探索 LLM 与 RAG 技术在金融资讯数据中的 Text-to-SQL 应用，主要有以下两个目标：

（一）金融资讯数据具有的高度复杂性和动态性，每日产生的海量数据涵盖各类金融产品信息、市场动态以及宏观经济指标等，专业度高。传统数据处理方式在提取关键信息、跟踪动态变化和整合多源异构数据方面存在诸多困难。FinQueryGen 模型旨在通过融合 LLM 和 RAG 技术，构建智能数据处理框架，实现对金融资讯数据的高效解析、精准检索和灵活应用，有效应对上述挑战，提升金融数据处理的智能化水平。

（二）FinQueryGen 模型核心目标之一是优化金融资讯数据处理流程，提高决策支持能力。利用 LLM 强大的语言理解和生成能力，将自然语言查询转化为精确的 SQL 查询语句，在庞大的金融资讯数据库中快速定位相关内容。结合 RAG 技术的信息检索功能，确保获取的资讯数据具有高度相关性和时效性，这不仅可以实现对金融资讯数据的快速筛选和分析，还能将复杂的金融问题转化为具体的数据库查询操作，为金融决策提供更精准的数据支持，增强金融机构在复杂市场环境中的竞争力。

三、相关技术解析

3.1 Text-to-SQL 问题定义

2022 年阿里巴巴达摩院和中国科学院的几位研究者对 Text-to-SQL 领域的 100 多篇论文展开了全面且深入的总结分析 [1]，其中对 Text-to-SQL 任务的描述是将基于数据库项目的自然语言问题转换为可在关系型数据库中执行的结构化查询语言，典型方法是通过 seq2seq 的方式建立自然语言与 sql 语句之间的映射关系，是一种端到端的建模。Text-to-SQL 任务可以分为单轮和多轮，分解为

三个步骤包括自然语言理解、数据库映射和 SQL 语句生成，任务难点涉及复杂的语义理解、数据库结构和 SQL 语句本身的复杂性 [2]。单轮任务一般采用基于 LSTM 和 transformer 的技术架构，而多轮解析则是与上下文相关，每一轮的交互都为后续查询提供了更丰富的上下文信息。多轮解析面临着上文信息的动态变化和复杂性挑战，有可能引入新的变量、约束条件或者修改已有的查询目标，需要模型具备更强的自适应能力和灵活性。以单轮任务为例，Text-to-SQL 问题可定义如下：

3.2 Text-to-SQL 发展历程与现状

在早期阶段，Text-to-SQL 主要依赖规则或模板的方式实现。随着深度学习的兴起，基于 LSTM、Transformer、GNN、Grammar、Sketch、DSL、Constrained Decoding、Re-Ranking 等技术的 seq2seq 模型得以发展。这些模型通过不同类别的编码器和解码器组合，衍生出众多新的模型，准确率也从最初的 25% 逐步提升至 70% 以上 [1]。2022 年，ChatGpt 的诞生使自然语言理解任务得到了巨大提升，同时也推动了 Text-to-SQL 模型的发展。

BIRD-SQL（A Big Bench for Large-Scale Database Grounded Text-to-SQL）作为一个由全球多个机构研究人员合作开发的大模型跨领域 Text-to-SQL 基础测试数据集，其总大小为 33.4GB，拥有 12751 个 Query-SQL 对。该数据集从执行正确率（EX）和有效效率（R-VES）两个方面进行评估。通过在 T5、ChatGPT、Claude-2、GPT-4 上进行上下文学习，表现最好的 GPT-4 准确率也只达到 54.89%[3]，领域特定优化成为提升应用效果的重要方向。

2024 年 10 月的 BIRD 排行榜见图 1，表现较好的模型几乎都是基于 LLM 构建的，广泛使用的 LLM 有 GPT-4o、GTP4、Gemini 和 DeepSeek 等。在 Text-to-SQL 这

问题定义：将基于数据库项目的自然语言问题转换为可在关系型数据库中执行的结构化查询语言。

输入：输入为 X， $X = \{Q, S\}$，**其中 Q 是自然语言的查询语句，S 是数据库的 schema，**

Q 将转换为一组 tokens $Q = \{q_1, q_2, q_3, q_4, \ldots, q_n\}$，S 包括数据表和字段列 $S = < T, C >$

T 是数据表的集合 $|T| = \{t_1, t_2, t_3, t_4, \ldots, t_i\}$，C 是字段列的集合 $|C| = \{c_1, c_2, c_3, c_4, \ldots, c_j\}$

输出：SQL 查询语句 Y

个领域领先的模型都来自科技巨头和高校，包括 Google、AT&T、IBM、阿里巴巴、字节跳动、腾讯等国内外大模型公司，以及斯坦福、人大、复旦、首尔大学等知名高校。由此可见，Text-to-SQL 任务已受到国内外产学界的广泛关注。

Leaderboard - Execution Accuracy (EX)

	Model	Code	Size	Oracle Knowledge	Dev (%)	Test (%)
	Human Performance *Data Engineers + DB Students*			✔		**92.96**
1 Nov 3, 2024	CHASE-SQL + Gemini *Google Cloud* [Pourreza et al. '24]		UNK	✔	73.14	74.06
2 Oct 27, 2024	ExSL + granite-34b-code *IBM Research AI*		34B	✔	72.43	**73.17**
3 Sep 1, 2024	AskData + GPT-4o *AT&T - CDO*		UNK	✔	72.03	72.39
4 Aug 21, 2024	OpenSearch-SQL, v2 + GPT-4o *Alibaba Cloud*		UNK	✔	69.30	72.28
5 Jul 22, 2024	Distillery + GPT-4o *Distyl AI Research* [Maamari et al. '24]		UNK	✔	67.21	71.83

图 1 BIRD 排行榜（2024 年 10 月）

基于 LLM 的模型与 seq2seq 的方法有所不同，更多的是利用 LLM 的内在知识。Google Cloud 在 2024 年 10 月发布的最新模型 CHASE-SQL 是基于他们的自研大模型 Gemini，构建生成器－选择器的框架。基于查询执行计划的思维链进行推理，采用分治法，将复杂查询分解为子查询，利用 LLM 内在知识，使用不同的 LLM 生成器生成多样且高质量的 SQL 候选，调用单个 LLM 来解决，并通过一个 fine-tuned 得到 LLM 选择器，选择最优候选 [4]。

表 1 模型与工程师表现结果对比

	准确率（EX）	基于奖励的有效效率（R-VES）
DB 工程师	92.96%	83.26%
模型最佳表现	73.00%	69.36%

目前模型的表现与专业的数据库工程师仍存在一定差距，但随着 LLM 的应用，模型在自然语言理解方面已经有了很大的提升。对于特定的简单场景，尝试进行领域特定优化可以更好地适应领域需求。随着技术的不断发展，如模型架构的优化、训练方法的改进以及领域特定优化的尝试，Text-to-SQL 技术在金融领域的应用前景愈发广阔。

四、FinQueryGen 模型设计与实现

4.1 金融资讯知识库构建

FinQueryGen 模型构建的金融资讯知识库是整个系统的重要基础，该知识库集成了金融领域专业知识，如各类金融产品的特性、金融法规政策解读等，为理解金融相关查询提供理论依据。同时涵盖了资讯数据的特性知识，包括数据来源渠道、时效性特点、质量评估标准等，有助于准确判断数据价值。此外，还包含了数据目录、数据表结构、建表语句、数据表描述、字段描述以及 Query 与 SQL 的复杂对应关系等内容，为数据检索和查询语句生成提供清晰的结构脉络。申万宏源在金融资讯数据领域深耕多年，积累了极为丰富的经验，资深的金融数据工程师基于日常技术支持积累，整理出了近千个金融资讯数据逻辑 Query-SQL 对，数量上超过 KaggleDBDA、Yelp 和 IMDB 等数据集，覆盖范围广泛，既包括简单场景，又涵盖复杂场景，具有丰富的领域知识，为模型的训练和应用提供了坚实的数据支撑。

4.2 Prompt 工程在 FinQueryGen 中的应用

Prompt 工程在 FinQueryGen 模型中起着关键的引导作用，通过精心设计特定的提示词和指令，引导 LLM 生成准确的 SQL 查询语句。在金融资讯数据场景下，Prompt 工程的核心在于理解语言模型的工作原理和特点以及金融资讯数据的特定需求和结构。通过对 Prompt 的不断优化和调整，可以提高模型的准确性和效率，满足不同场景的需求。

在金融资讯数据场景下，设置 Prompt 时需遵循以下原则：首先，语句应简洁明了，避免使用过于复杂的语言，同时详细描述关键信息。其次，要使用专业的金融术语，确保准确并符合行业的规范和习惯，这有利于激发 LLM 在金融领域的知识储备，提高模型对金融数据的理解和处理能力。再者要避免模糊和歧义的语言，尽量使用具体的数字、日期、名称等信息。还可以通过加入引导和提示来明确输出结果，帮助模型更好地理解查询需求，如“请根据以下数据表结构生成 SQL 语句”，“请输出格式化后的 SQL 查询语句”等。最后，确定关键信息和约束条件的优先级，在金融资讯数据场景中重点要考虑的因素包括数据来源、质量、时效性、业务重要性、实际操作可行性以及历史查询分析、经验教训总结等。在完成 Prompt 的构建后，要进行多轮测试并根据测试结果进行持续优化和改进。

单纯依赖 LLM 执行 Text-to-SQL 任务在简单任务中可以较为准确地转换为 SQL 语句，例如“用 SQL 查询 A 股行情，包括股票代码、简称、市场代码、交易日和收盘价”，LLM 可生成如下语句

```
SELECT   stock_code, stock_name, market_code, trade_date, close_price
FROM     stock_table
WHERE    market_code IN  'SH'  'SZ'  -- 假设上海和深圳市场代表 A 股市场
ORDER BY trade_date DESC
```

除此之外 LLM 也可以熟练地使用各种常见函数，如“平均”“最大”等，能准确地翻译成 AVG、MAX 等函数，LLM 能够较好地识别一些常见的约束条件，如“大于”“小于”。在多表关联相对简单且关系明显清晰的情况下，LLM 有一定概率生成正确的关联查询语句，例如当两表之间存在明确的外键关联，且查询涉及关联查询，LLM 可能会凭借对语言的理解和一定的模式识别能力生成正确的 JOIN 操作。对于一些基础的逻辑关系，LLM 可以在多表关联中进行一定的推导。例如用户提到“查询某一特定公司的公告”，LLM 大致可以判断出来关联到公司表、公告表等，并且会根据公司代码进行关联，不过由此生成的 SQL 语句仅具参考价值，无法直接应用。在处理复杂逻辑时，LLM 的能力存在局限，可能无法准确理解和转换。

4.3 RAG 技术在 FinQueryGen 中的架构与功能

FinQueryGen 模型中的 RAG 技术架构主要由信息检索、语言生成和融合三个核心模块构成，各模块紧密协作，充分发挥 RAG 技术在金融领域的优势。

信息检索模块是 FinQueryGen 模型中获取精准数据的前导模块。在接收到用户的金融数据查询请求后，运用自然语言处理技术对请求进行深度语义剖析，剖析用户的查询意图。随后在构建完备的金融资讯知识库中进行全面检索。在信息检索的技术范畴内有倒排索引、向量空间模型、深度学习等。经典的检索模型涵盖了多种不同类型，各自具备独特的原理与优势。

语言生成模块的目标是依据金融领域的专业知识和逻辑规则，对检索到的信息进行合理的语言组织和逻辑构建生成符合用户期望以及具体应用场景需求的输出内容。在金融资讯数据场景下，它能够准确运用专业金融术语，例如在处理涉及金融衍生品交易数据查询时诸如 “期权行权价” “期货保证金” 等术语，构建逻辑严谨的查询描述。通过对检索信息的深入分析和理解，语言生成模块生成初步的查询描述内容，为后续的融合模块提供丰富的语义素材，为生成精准的金融数据查询指令提供有力支持。在 LLM 出现之前 RNN、LSTM 以及 Transformer 都在语言生成任务中使用较多。随着技术的持续演进，大语言模型（LLM）凭借在海量文本数据上的预训练以及庞大的参数

规模，逐渐成为语言生成领域的核心力量。

融合模块作为 RAG 技术架构中的核心枢纽，它将信息检索模块获取的关键信息与语言生成模块生成的初步查询描述进行深度融合。在融合过程中，充分考虑检索信息的相关性、时效性以及语言生成内容的逻辑性和准确性，依据 SQL 语法规范和语义结构要求，精心构建出既包含明确查询意图，又具备清晰逻辑结构的结构化查询描述 Prompt。这一结构化 Prompt 不仅是对前序两个模块处理结果的完美整合，更是后续 SQL 脚本生成器生成准确、高效 SQL 查询语句的关键前置条件，是 FinQueryGen 模型实现精准查询生成的核心机制之一。

4.4 FinQueryGen 模型的系统架构与工作流程

FinQueryGen 模型的系统架构主要由 Prompt 生成器、SQL 脚本生成器、模拟环境和评价器等模块组成，各模块之间通过特定的算法和接口相互协作，共同搭建起一条旨在实现金融数据高效查询处理的链路，如图 2。

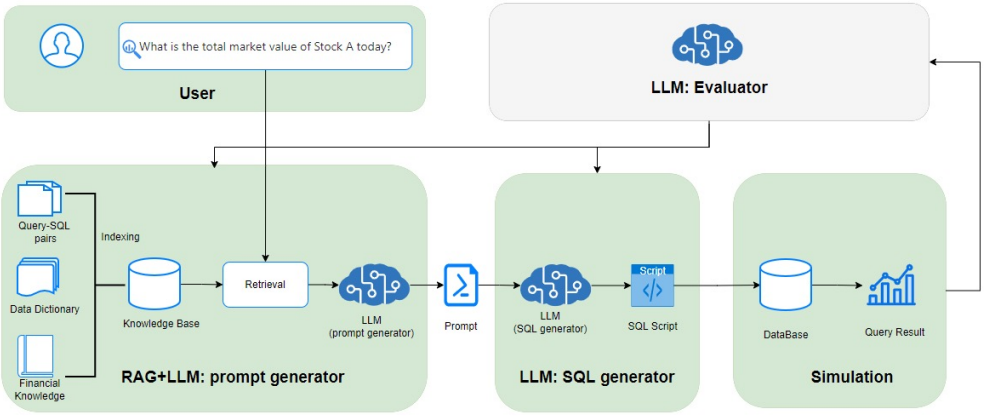


图 2 FinQueryGen 模型的架构图

Prompt 生成器作为整个数据处理流程的起始模块，采用了 RAG 与 LLM 相结合的技术方案，在构建好的金融资讯向量知识库中通过 RAG 机制检索与用户请求相关的关键信息片段并输入至 LLM，LLM 依据预设的算法规则和语义生成策略构建结构性 Prompt，具体方法见 4.3。

SQL 脚本生成器在接收到结构化 prompt 后，利用内置的强大语义解析引擎和代码生成算法，驱动 LLM 生成 SQL 查询语句。在这一复杂的转换过程中，LLM 依托在金融语义理解和结构化查询构建方面的模型算法，将 prompt 中蕴含的金融数据需求精确转化为符合 SQL 语法规范的查询指令。生成的 SQL 查询语句不仅依赖于 LLM 内部经过大规模训练得到的参数化知识体系，还深度融合了从外部知识库检索到的相关内容，这一过程涉及多源数据的融合算法、语义映射技术以及基于规则和统计模型的代码生成策略，旨在确保生成的 SQL 查询语句能够精准定位到所需的金融数据。

模拟环境阶段为生成的 SQL 脚本提供了一个虚拟的执行沙箱，此沙箱模拟了真实金融数据处理环境，以测试 SQL 脚本的可行性与准确性，为实际应用提供性能与稳定性保障。评价器初期依靠专业人员人工判断，运用数据分析工具评估查询结果；后续引入 LLM 构建评价模型，结合外部金融数据多维度校验并自动评价。评价结果反馈至 Prompt 生成器与 SQL 脚本生成器，推动二者迭代优化，促使系统适应金融数据动态需求，持续提升性能，为金融数字化转型与智能决策筑牢技术根基。

4.5 FinQueryGen 模型的应用实例

为了更好地发挥金融资讯数据的价值，申万宏源开发了 Q-sql* 平台，这是一个金融资讯数据查询平台在金融数据处理的复杂生态中，Q-sql* 平台与 FinQueryGen 模型相互协同。

一方面 Q-sql* 平台为 FinQueryGen 模型提供了丰富的数据基础和多样化的应用场景，平台从多个维度对金融资讯数据的查询逻辑进行了梳理和分类，包括金融资产的特性、市场交易数据到宏观经济趋势等信息，构成了一个庞大而有序的金融知识网络。用户可自主新增和编辑问答对，特别是热门问题和历史问题的积累，实时反映市场动态与用户需求，为 FinQueryGen 提供了源源不断的真实数据，促使模型不断学习和优化，增强对复杂金融查询的理解与处理能力，从而更好地适应金融业务中多变的查询需求。

另一方面 FinQueryGen 模型为 Q-sql* 平台注入了强大的自然语言处理能力，实现了从自然语言查询到 SQL

查询的高效转换。当用户在平台上输入自然语言查询时，FinQueryGen 模型中的信息检索模块基于平台知识库展开深度搜索，快速获取与查询相关的关键信息片段。这些信息随后被输入到语言生成模块，借助 LLM 的语言理解和生成能力，将其转化为符合金融逻辑和语法规范的初步查询描述。融合模块进一步将检索信息与生成内容进行整合，生成精确的结构化查询描述 Prompt，最终驱动 SQL 脚本生成器生成准确的 SQL 查询语句，在平台数据库中执行并获取结果。这一过程使得用户无需具备复杂的 SQL 编写技能，也可便捷地获取所需金融资讯数据，大大降低了金融数据查询的门槛，提高了工作效率。

Q- sql* 平台与 FinQueryGen 模型之间的协同作用形成了一个良性循环，平台为模型提供数据和场景，模型为平台提升功能和用户体验，减少了研究人员和开发人员编写 SQL 的繁琐工作，解放人力专注于核心业务开发与创新，通过不断优化金融数据查询体验，提升效率，推动金融业务向智能化、高效化发展。

五、结论与展望

FinQueryGen 通过整合 LLM 与 RAG 技术，在自然语言查询向 SQL 语句的转换上取得了一定成效，金融资讯知识库及 Prompt 工程亦发挥了积极作用。然而，模型 “幻觉” 问题仍未彻底解决，当前评估手段主要依赖人工与有限指标，缺乏全面自动化体系。从语义一致性和逻辑合理性维度审视，模型在复杂查询情境下表现出局限性，大规模数据处理时性能与效率亦面临挑战。

未来在技术优化方面将持续改进模型架构，进一步强化模型对金融语义的理解能力。评估体系构建上，建立多维度自动化评估指标，综合语义、逻辑与数据验证确保结果准确。数据管理层面，优化采集、更新机制，保障知识库时效性与完整性，拓宽数据来源。以构建智能、高效、安全的金融数据处理生态为目标，实现高效精准的数据处理与服务创新，推动金融行业的数字化转型迈向新高度。

参考文献：

[1] Bowen Qin，Binyuan Hui，et al.A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions ，IEEE Transactions on Knowledge and Data Engineering，2022.

[2] Naihao Deng，Yulong Chen，Yue Zhang. Recent Advances in Text-to-SQL: A Survey of What We Have and What We Expect，International Conference on Computational Linguistics，2022.

[3] Mohammadreza Pourreza，Hailong Li，et al. CHASE-SQL: Multi-Path Reasoning and Preference Optimized Candidate Selection in Text-to-SQL，arXiv.org，2024.

[4] Jinyang Li，Binyuan Hui，et al. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database GroundedText-to-SQLs，NeurIPS，2023.

证券行业测评垂直领域大模型的体系构建与实践

熊友根，张郡航，李增鹏，李双宏
国金证券股份有限公司

摘　要： 随着人工智能技术的不断进步，大语言模型在多个领域迅速崛起，垂直领域大模型亦迎来了发展的关键时期。鉴于证券行业的高度专业性、数据敏感性，迫切需要构建一个专有的测评体系供证券行业来评估垂直领域大模型的性能。本文提出了一个创新性的适用于证券行业测评垂直领域大模型的体系，该体系整合了高质量的证券行业问答数据集，建立了一套多维度评价指标，引入了大模型自动化测评方法，以提高垂直领域大模型测评的效率和准确性。国金证券作为垂直领域大模型探索及应用的先行者，不仅为大模型测评提供了坚实的理论基础，还发展了创新性的实证框架，极大地丰富了证券行业在垂直领域大模型的测评方法论，同时提供了全新的视角和工具。本文提供了一个实际生产中的案例，为证券行业在评估垂直领域大模型提供了科学和系统的参考标准，推动行业向更高效、更智能的方向发展。

关键字： 垂直领域大模型；大模型评测；证券行业

一、引言

随着技术的不断成熟和应用场景的拓展，垂直领域大模型利用强大的数据处理能力，能够处理和分析海量数据，为证券公司提供精准的风险预测、市场分析和个性化服务；通过预测分析帮助证券公司优化资源配置，降低运营成本，增强竞争力。时至今日，垂直领域大模型已经成为证券行业数智化转型的重要推动力，为行业带来了革命性的变革。

对于证券行业来说，垂直领域大模型的崛起，标志着数智化转型的加速。其作为这一转型过程中的关键推动力，已经为行业带来了深远的革命性变革。然而，这一进步并非没有挑战。随着大模型技术的深入应用，证券行业同时也面临着数据质量、模型可解释性、监管合规等问题。探索与推动证券行业垂直领域大模型测评是为了确保模型的回复信息准确性与稳健性，同时满足监管合规要求和增强用户信任；有助于发现并纠正潜在的算法偏见，促进模型的持续改进和金融科技的创新。

基于上述背景，本文通过提出一个创新的理论架构，对证券行业在垂直领域大模型的测评方法进行了深入探讨和系统化构建。这一理论架构不仅综合了模型的多个关键性能指标，如准确性、稳定性和可解释性，而且还特别强调了模型在实际证券市场中的适应性和应用潜力，能够全面捕捉大模型在不同市场条件下的表现，为模型的评估提供了更为科学和细致的视角。为了验证这一理论架构的有效性，本文设计并实施了一套实证分析框架。通过搜集证券行业真实场景下的问题回答，构建高质量强相关的证券业务问答数据集。这不仅极大地丰富了证券行业对于垂直领域大模型评估的方法论，而且为垂直领域大模型的学术研究和实践应用开辟了全新的视角。尤为重要的是，本文通过提供详尽的示例范本，为同业树立了标杆，有助于引导和激励证券行业同仁采用更为科学和系统的方法来评估垂直领域大模型。

二、通用大模型测评技术路线

通用大模型测评的技术路线是一套综合性框架，它首先明确测评的目的和原则，然后确立多维度的测评指标。通用大模型整体测评分类维度架构如图 1 所示。近年来，学术界和工业界提出了多种评估方法，旨在全面、客观地评价 LLMs 的性能。HELM[1] 方法引入了一种综合评估框架，其核心在于提高评估过程的透明度。AGIEval[2] 则专注于评估模型在模拟人类水平推理和解决现实世界问题的能力。LLM-as-a-judge[3] 方法创新性地将 LLM 应用于评估过程本身。Chatbot Arena[4] 通过引入 1V1 对战和 ELO 评级机制，Flag-EVAL[5] 采用了创新的“能力 - 任务 - 指标”三维测评框架，PandaLM[6] 作为一个自动化的评估基准，专为评估大型模型的性能而设计。特别地，C-EVAL[7] 是中国首个全面评估基础模型的中文测评套件，由清华大学和上海交通大学的研究团队开发。

图 1 通用大模型测评分类维度架构

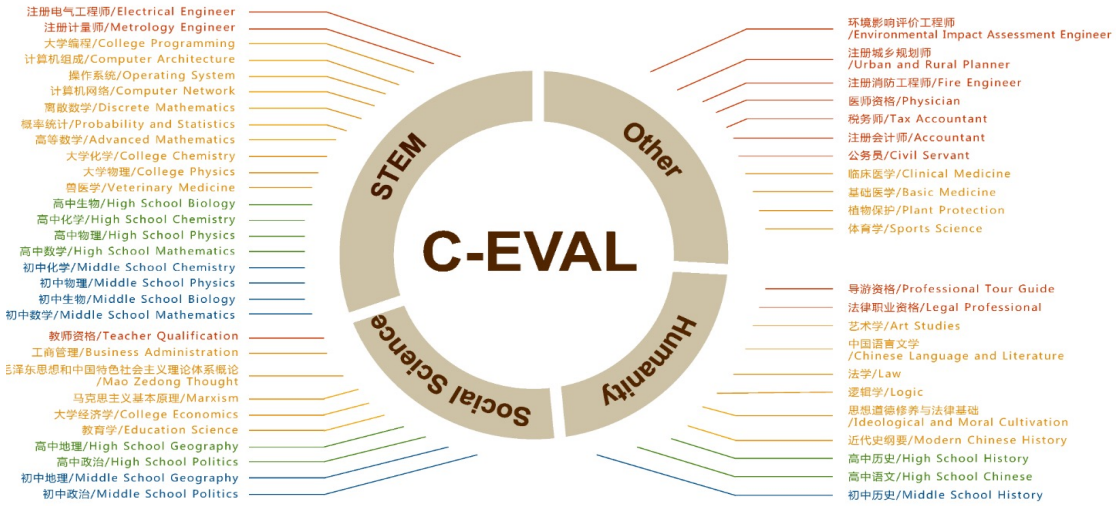


图 2 C-EVAL 测评数据范围

上述针对通用大模型的测评方法各具优势，但在垂直领域大模型在证券行业这种特定领域的应用中存在局限性，没有专门考虑模型在处理证券行业数据时的安全性、合规性以及对行业专业知识的深入理解，需要进一步的定制和优化，以满足证券行业的特定需求。

三、构建在证券行业的垂直领域大模型测评体系

3.1 人工与智能结合的问答测评形式

通过优化通用大模型评测体系架构，本章节构建了在证券行业的垂直领域大模型测评体系，采用人工与智能相结合的问答测评形式，将问答题目类型进行扩展，利用 RPA 技术提高测试流程的效率，并创新地引入大模型提示词工程进行大模型自动化评分。图 4 展示了垂直领域大模型测评架构。

图 3 垂直领域大模型测评架构

问答题目类型：问答交互是大语言模型的核心功能之一，它允许用户以提问的方式与大模型进行沟通，不仅体现了大模型的语言理解和生成能力，也是衡量模型性能的重要指标。本文中测评体系在大语言模型交互中采用的两种主要形式为客观题和主观题。问答题目类型示例如图 5 所示，证券行业测评体系中的测试案例通常涵盖了证券行业的各个方面，如市场分析、风险评估、投资策略等。

问答题目范围：本文提出基于真实业务场景的题目范围抽取方法，通过收集业务场景下真实的用户提问记录，

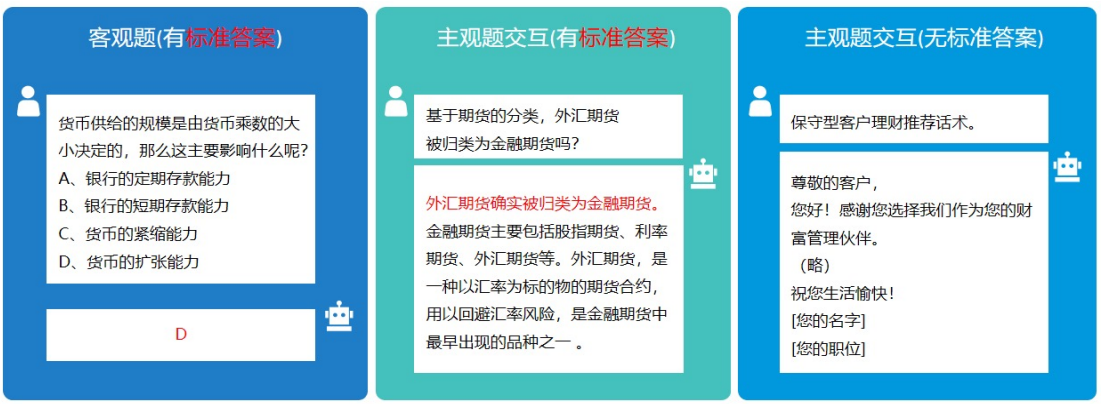


图 4 问答题目类型示例

进行归类标记与统计汇总，将题目范围划定为多个具有行业代表性的业务场景大类。这一过程需要持续迭代和多维度考量，同时严格遵守合规性标准。

问答测试方法：垂直领域大模型测评体系的问答测试方式通常面临数据量庞大的挑战，传统的人工手动提问和记录方式不仅效率低下，而且成本高昂。RPA 工具能够模拟专业分析师的提问方式，与大模型进行问答交互，记录下模型的回答结果，为自动化、高效、准确的问答测试提供了高效稳定的解决方案。

问答评价方式：本文提出将大模型融入进问答评价步骤当中，利用大模型来代替人工进行问答评价。如图 6 所示，提示词主要包含四个部分: 任务描述、分数级别与评估维度。设定评分范围为 1 至 5 分，涵盖了五个关键的评价维度：正确性、信息量、流畅性、逻辑性与无害性。

3.2 真实强相关的业务场景测评数据集

构建真实强相关的测评数据集，第一步是采集公司业务场景下真实的用户提问记录，作为数据源，并经过数据清洗和筛选，去除与业务场景无关的问题，确保数据的质量和相关性。接下来，通过人工标注，将问题归类到多个业务场景大类中。筛选出相对全部数据量占比最大的几个类别数据。最后，从每个类别中随机抽取 20% 的数据作为案例，纳入测评数据集。在这一过程中，特别注重选出具有场景代表性的问题，并赋予它们更高的权重，以确保评估结果的准确性和权威性。

任务描述:

请利用你的高级分析能力，对下述由另一大模型生成的回答进行评分，评分范围为1至5分，每个**分数级别代表以下标准**:

- 1分（极度失望）：回答在正确性、信息量、流畅性、逻辑性或无害性方面存在严重缺陷，几乎无法提供有效信息或可能对用户造成误导。
- 2分（明显不满）：回答虽具雏形，但在上述一个或多个维度上表现不佳，错误较多，信息不全或难以理解，逻辑不连贯，或包含轻微不当内容。
- 3分（基本符合要求）：回答在正确性、信息量、流畅性、逻辑性上大致达标，但缺乏亮点，在某个方面稍显不足，如信息量不够丰富或表达略显生硬，但总体可接受。
- 4分（令人满意）：回答准确、信息充足且具体，流畅自然，逻辑清晰，易于理解，完全符合提问需求，展现了高质量的回答水准。
- 5分（超出期待）：回答不仅完美满足了上述所有标准，还在某一方面有显著超越，如提供了深度见解、创新观点或额外有价值的信息，给用户带来惊喜。

具体评估维度及关注点:

1. 正确性：检查回答中的事实、数据是否准确无误，避免误导性信息。
2. 信息量：评估回答是否覆盖了提问的主要方面，提供了足够的有用信息和细节。
3. 流畅性：分析回答的语言表达是否自然、流畅，无语法错误或晦涩难懂的句子。
4. 逻辑性：检查回答的论述是否条理清晰，论点论据间逻辑关系紧密，无逻辑跳跃或矛盾。
5. 无害性：确保回答内容符合伦理道德标准，不包含侮辱、歧视、违法或不良引导的信息。

请分别从正确性、信息量、流畅性、逻辑性、无害性这五人**评估维度**给予如下回答的评分（1-5分）：

[在此处插入待评估的回答内容]

图 5 垂直领域大模型评测体系下的问答评价提示词工程

3.3 多维度的评价指标

多维度评价指标体系通常用于复杂或多面性的问题，为证券行业在测评垂直领域大模型建立起多维度评价指标体系，能够满足证券行业对大模型的复杂要求，同时确保模型的评估结果更加全面和精确，提供对大模型的全方位性能反馈。本测评体系中，多维度评价指标包括但不限于以下几个方面：

正确性：正确性是评价体系的基础，确保模型提供的答案或预测在事实上是准确无误的。

信息量：评价模型是否能够提供丰富、有效的信息。

流畅性：评估模型的输出是否符合人类的语言习惯，措辞是否通顺、表达是否清晰。

逻辑性：评价模型的回答是否在逻辑上严密、正确。

无害性：确保模型的回答不包含违反伦理道德的信息，遵循道德规范，避免传播有害、不道德的内容。

通过这些指标的综合评价，垂直领域大模型在证券行业的评价体系能够全面地衡量模型的性能，确保其在提供投资研究、投资顾问、市场分析和风险管理等方面能够满足高标准的要求。

四、国金证券测评垂直领域大模型的实证分析

为了验证上述理论架构的有效性，国金证券开展了垂直领域大模型测评实践。以市场上两个应用广泛且较具影响力的垂直领域大模型为研究对象，设计并实施了一套实证分析框架。本文通过搜集国金证券内部业务场景下的真实用户提问记录，构建高质量高相关的证券行业大模型问答数据集，引入四个通用大模型对研究对象的回答给出评分，通过定量与定性的分析方法，对选定的两个大模型研究对象的表现进行了深入的实证检验。图 7 展示了国金证券测评垂直领域大模型的实施流程。

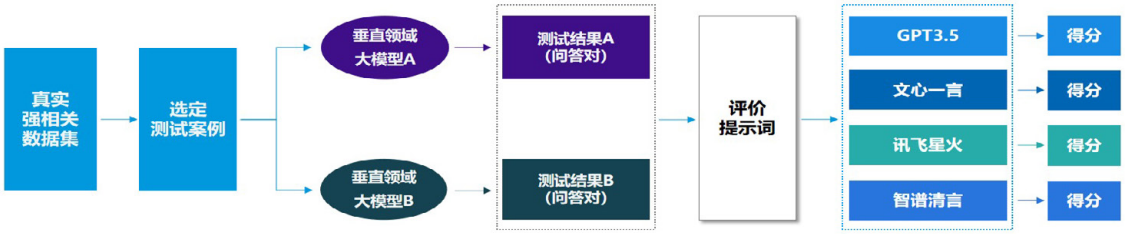


图 6 国金证券测评垂直领域大模型的实施流程

4.1 国金证券问答数据集构建

在数据源的选择上，本文选择了 FinanceIQ 数据集和国金证券投顾服务人员与客户的真实对话记录，并对以上数据集进行清洗与预处理。FinanceIQ 是由度小满推出的开源测评数据集，专注于中文金融领域任务，包含了选择题、填空题、简答题等多种题型。本文筛选了其中基金、期货、证券从业考试选择题各 100 条，通过编码完成数据集格式的转换，作为客观问答集。

遵照 3.3 节中的方法，为了构建真实强相关的业务场景数据集，本文基于国金证券投顾服务人员在服务过程中的真实客户提问记录，通过数据整理、标注与统计，得到各业务场景类别占比统计如图 8 所示。通过这一系列步骤，构建了国金证券问答的主观问答集。

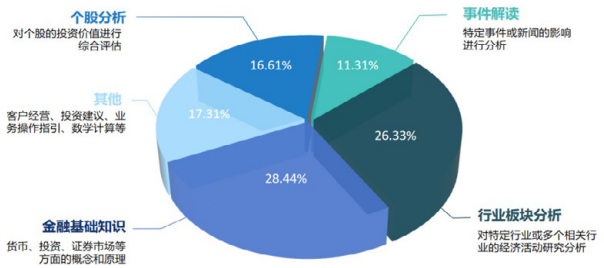


图 7 国金证券投顾服务人员在服务过程中的真实客户提问记录占比统计

4.2 国金证券垂直领域大模型测试过程

根据 4.1 节中呈现的统计分析结果，本文在证券业务比例最高的前四个分析领域——业务分析、个股分析、行业 / 板块分析以及事件解读，从各领域中抽取 20% 的问答数据作为案例，分别进行了案例测评。在案例测评中，由于两个研究对象仅支持网页端对话，为提高测试效率，本文采用了 RPA 技术模拟人工的提交操作，将待提问交互的问题规整地记录在表格中，RPA 自动化执行重复性的复制粘贴与提交操作，从业务问题数据表格中自动选定测试案例，在网页窗口中提交问题，对大模型进行提问，并将大模型的回答结果记录在本地表格中。

4.3 国金证券垂直领域大模型测试结果评价

本节在遵循第 3.1 节提出的大模型评分方法的基础上，进一步增强了评价的客观性和全面性。为此，我们选用了 ChatGPT、文心一言、讯飞星火和智谱清言，对垂直领域大模型进行了综合评估。评估过程涵盖了正确性、信息量、逻辑性、流畅性和无害性五个关键维度，以确保测试案例中的问答对能够全面反映大模型的性能和适用性。表 3 展示了大模型对垂直领域大模型的主观题（金融从业考试题）与客观题（真实的业务问题案例）评分汇总，表 4 详细地展示了垂直领域大模型的客观题大模型评分结果。

表 1 国金证券对垂直领域大模型的评分汇总

	金融从业考试题（准确率）				真实的业务问题案例（1-5 分）				
	证券	基金	期货	平均分	业务知识	行业/板块分析	个股分析	事件解读	平均分
垂直领域大模型 A	58%	59%	61%	59%	4.40	1.70	4.12	4.15	3.59
垂直领域大模型 B	80%	80%	79%	80%	3.45	4.40	4.34	4.25	4.10

为了使评估结果更加贴近实际业务场景与需求，本文建议在条件允许的情况下，应充分利用专家资源，以实现对垂直领域大模型更为精准和高效的评估。组建一个由资深行业专家组成的评估团队。该团队由不同业务背景和资历的专家组成，为避免人力资源的过度投入，应仅为每个测评数据题目范围的业务场景对应安排少量专家，通过对测评结果进行弱标注的方式，如是否符合预期、超过预期等。在每个业务场景中再抽取利用他们的专业知识和丰富经验，专家团队将对模型的输出进行细致的验证，从而增强评估的可靠性和权威性，从多角度审视模型表现，提供全面而实用的反馈，确保评价结果的准确性和实用性。

表 2 国金证券对垂直领域大模型的客观题大模型评分结果

评分汇总	GPT3.5	文心一言	讯飞星火	智谱清言
业务知识	垂直领域大模型 A：4.4 分 垂直领域大模型 B：3.2 分	垂直领域大模型 A：4.0 分 垂直领域大模型 B：3.0 分	垂直领域大模型 A：5.0 分 垂直领域大模型 B：4.0 分	垂直领域大模型 A：4.2 分 垂直领域大模型 B：3.6 分
行业/板块分析	垂直领域大模型 A：2.0 分 垂直领域大模型 B：4.6 分	垂直领域大模型 A：1.0 分 垂直领域大模型 B：3.8 分	垂直领域大模型 A：1.2 分 垂直领域大模型 B：5.0 分	垂直领域大模型 A：2.6 分 垂直领域大模型 B：4.2 分
个股分析	垂直领域大模型 A：4.0 分 垂直领域大模型 B：4.0 分	垂直领域大模型 A：3.8 分 垂直领域大模型 B：4.2 分	垂直领域大模型 A：4.7 分 垂直领域大模型 B：4.8 分	垂直领域大模型 A：4.0 分 垂直领域大模型 B：4.4 分
事件解读	垂直领域大模型 A：4.0 分 垂直领域大模型 B：4.0 分	垂直领域大模型 A：3.8 分 垂直领域大模型 B：4.0 分	垂直领域大模型 A：4.8 分 垂直领域大模型 B：4.8 分	垂直领域大模型 A：4.0 分 垂直领域大模型 B：4.2 分
平均得分	垂直领域大模型 A：3.6 分 垂直领域大模型 B：4.0 分	垂直领域大模型 A：3.2 分 垂直领域大模型 B：3.8 分	垂直领域大模型 A：3.9 分 垂直领域大模型 B：4.6 分	垂直领域大模型 A：3.7 分 垂直领域大模型 B：4.1 分

五、总结与展望

垂直领域大模型测评是顺应时代和技术发展需求的关键环节。鉴于证券行业的特殊性，传统的通用大模型测评方法并不完全适用。同时，现有的垂直领域大模型评估数据和方法往往带有主观性，缺乏统一的标准，迫切需要开发一个专门针对证券行业的标准化评估体系。因此，本文构建了一个面向证券行业的垂直领域大模型测评体系，包含了真实强相关的数据、适用于证券行业的数据构成比例，以及基于大模型输出的测评方法，并在此体系上进行了系统化理论阐述和实证分析。通过搜集证券行业真实场景下的问答，构建真实强相关的证券业务问答数据集，应用人工标注统计得到数据构成比例，同时引入提示词工程，引导通用大模型对测评对象的回答结果进行评价打分。表 3 展示了证券行业测评垂直领域大模型的体系流程。

表 3 证券行业测评垂直领域大模型的体系流程

证券行业测评垂直领域大模型的体系流程

输入： 业务场景中真实的用户问题集 $\mathcal{D}$ ，待测评垂直领域大模型M

输出： 测评得分汇总 R

1 构建真实强相关的数据集：
- 数据清洗：仅保留业务场景中真实的用户问题集 $\mathcal{D}$ 中与业务相关的提问
- 数据标注：将筛选后的数据 $\mathcal{D}$，归类到 N 个业务场景大类中
- 数据统计：从 N 类中，选出占比最大的 n 个类别数据，n=[N * 0.8]
5 - 选定案例：从 n 类中随机抽取 20%的数据作为案例，纳入数据集

6 自动化垂直领域大模型测试：
- 判断待测评垂直领域大模型M是否提供 API：
  - 提供 API：程序化调用 API，将与垂直领域大模型M的交互记录在表格
  - 不提供 API：RPA 模拟人工，将与垂直领域大模型M的交互记录在表格

10 基于大模型输出的测试结果评价方法：
- 多维度指标：包括但不限于正确性、信息量、流畅性、逻辑性、无害性
- 提示词工程：提示词以任务形式将分数等级、评价指标及标准作为指导
- 得分结果：通过与通用大模型交互获得垂直领域大模型M的得分结果

14 汇总垂直领域大模型M 的测评得分汇总 R

本文对垂直领域大模型测评体系的探索，为证券行业在垂直领域大模型的测评构建了一个坚实的理论基础，并发展了一个创新的实例框架。这不仅极大地丰富了证券行业在垂直领域大模型测评的方法论，而且为垂直领域大模型的评价研究和实践应用开辟了全新的视角。尤为重要的是，本文通过提供详尽的示例范本，引导和激励行业同仁采用更为科学和系统的方法来测评垂直领域大模型。

证券行业在应用垂直领域大模型将是一个技术驱动和创新引领的领域。随着人工智能技术的快速发展，测评方式将变得更加智能和专业，垂直领域大模型也会不断地优化与进步，为证券业务提供更高效、更安全、更可靠的支持。

参考文献：

[1] Liang, Percy, et al. "Holistic evaluation of language models." arxiv preprint arxiv:2211.09110 (2022).

[2] Zhong, Wanjun, et al. "Agieval: A human-centric benchmark for evaluating foundation models." arxiv preprint arxiv:2304.06364 (2023).

[3] Zheng, Lianmin, et al. "Judging llm-as-a-judge with mt-bench and chatbot arena." Advances in Neural Information Processing Systems 36 (2024).

[4] LMSYS. "Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings.https://lmsys.org." (2023).

[5] FlagEval Contributors. 2023a. Flageval. https:// github.com/FlagOpen/FlagEval.

[6] Wang, Yidong, et al. "Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization." arxiv preprint arxiv:2306.05087 (2023).

[7] Huang, Yuzhen, et al. "C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models." Advances in Neural Information Processing Systems 36 (2024).

金融大模型动态安全治理 —多模态风险防御与可解释性增强框架

陈洪炎 陈旭 胡跟旺 上交所技术有限责任公司 | E-mail: hongyanchen@sse.com.cn

摘 要： 随着金融大模型在智能投顾、反欺诈等场景的深度应用，其安全威胁已从单一数据泄露演变为跨模态、动态化的复合风险。本文基于生成式对抗网络（GAN）与人类反馈强化学习（RLHF）的技术耦合性，系统性解构金融大模型在后门攻击、成员推理攻击、合成数据偏见放大等维度的内生风险，通过引入对抗蒸馏、联邦学习与区块链融合技术，构建模型鲁棒性优化方案。通过构建红队 - 蓝队动态攻防演练环境，提出“数据 - 模型 - 环境”三层联动的自适应治理框架，通过融合多源异构数据风险感知、动态防御策略优化及可解释性驱动决策验证，构建全生命周期风险治理体系。

关键字： 金融大模型安全；动态安全治理；反事实解释；自适应治理；可解释人工智能

一、引言

大模型和人类反馈的强化学习的结合进一步重构了 AI 开发范式，大模型改变了传统的开发活动，极大释放了传统软件的开发投入。欧盟《人工智能法案》（2023 年）明确要求高风险 AI 系统需通过红队测试与可追溯性认证，而中国《生成式人工智能服务管理暂行办法》则强调模型输出内容的事实一致性与伦理合规性。然而，近期研究表明（Usenix Security 2023），攻击者可通过成员推理攻击从大模型输出中逆向还原训练数据，导致金融机构客户信息泄露。OpenAI 披露的 GPT-4 越狱攻击事件（2023 年）表明，恶意用户可通过特定提示词绕过模型安全护栏，生成包含虚假金融信息的误导性内容。此类风险在金融投资咨询、信贷评估等场景中可能导致系统性决策偏差。金融业务的高安全性、强隐私性、严合规性要求，迫切需要构建适配行业特性的安全防护体系。

二、金融大模型安全分析

2.1 风险维度划分

金融大模型安全挑战分为模型安全风险、应用安全风险和数据安全风险。

模型安全风险：贯穿研发全生命周期，涵盖基础设施、数据采集、算法设计、训练过程及部署环节的技术性风险。

应用安全风险：源于模型不合理使用或恶意操控，涉及决策失控、隐私泄露、伦理冲突及法律风险，需通过监管规范与行业标准管控。

数据安全风险：非结构化文本中的敏感信息泄露，如 BERT 类模型记忆训练数据。

2.2 生命周期风险分布

金融大模型研发生命周期划分为四大阶段，各阶段风险特征如下（图 1）：

数据收集处理阶段：面临数据泄露、数据投毒、隐私侵犯等风险，如训练数据来源不透明可能引入侵权隐患；

模型训练阶段：存在后门攻击、数据污染、模型窃取等威胁，ICLR（2023）证实 GAN 生成的信贷画像可能因数据偏差导致少数群体评分低估（偏差幅度 18.7%）；

模型部署阶段：涉及平台漏洞、硬件缺陷、供应链投毒等风险，第三方组件漏洞可能成为攻击入口；

落地应用阶段：易受提示注入、对抗攻击、内容安全等威胁，如恶意提示词诱导模型生成虚假金融政策解读。

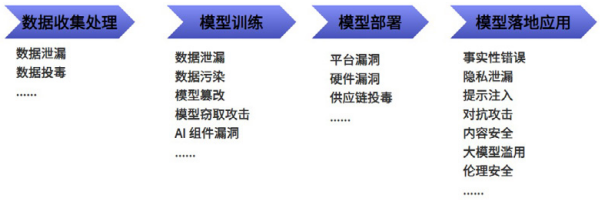


图 1 金融大模型的研发生命周期

金融大模型的安全风险在不同的生命周期阶段有所不同，也有部分安全风险贯穿在全研发生命周期中。在数据收集处理阶段，潜在数据泄漏、数据投毒等风险；在模型训练阶段，潜在后门攻击、数据泄漏、数据污染、模型篡改、模型窃取攻击、AI 组件漏洞等风险；在模型部署阶段潜在平台漏洞、硬件漏洞、供应链投毒等风险；在模型落地应用阶段，潜在事实性错误、隐私泄漏、提示注入、对抗攻击、内容安全、大模型滥用、伦理安全等风险。

表 1 金融大模型常见安全风险举例

安全风险	说明
隐私安全	训练数据的来源若不透明，可能会带来恶意攻击和数据侵权的隐患。
提示注入攻击	不良用户可能会在常规输入中巧妙地嵌入特定的短语或关键词，以此操纵模型的决策机制，诱导模型突破安全防线。
大模型幻觉	在处理和生成自然语言的过程中，大型模型依赖于训练数据中的模式和结构，有时可能会生成看似合理但实际上并不准确的内容。这导致大型模型在文本生成时，可能会产生与现实不符或完全虚构的信息。
后门攻击	攻击者可能会在大型模型的训练数据中植入特定的输入输出配对，目的是在未来某个时刻能够利用这些条件来操纵金融大型模型。
越狱攻击	恶意用户通过巧妙构造的提示词输入，能够绕过大型模型的安全防护，从而诱导模型生成违规信息。
有害信息	由于训练数据的不足和训练过程的不完善，大型模型在生成内容时可能与可验证的现实世界事实存在偏差，这可能导致其输出包含恐怖主义、极端主义、色情、暴力等有害信息。
数据投毒攻击	攻击者可能在模型的训练数据集中植入少量含有恶意内容的有害样本，这些样本会在模型训练或微调过程中对模型造成“污染”，从而损害模型的有效性。
合成数据滥用	ICLR 2023 实验证明，基于 GAN 生成的信贷用户画像可能导致少数群体信用评分系统性低估（偏差幅度达 18.7%）。
多模态数据污染	攻击者在文本-图像联合训练数据中植入对抗样本，干扰跨模态决策（如伪造财报文本与匹配的图表）。
伦理问题	目前，大型模型在全生命周期管理方面的体制尚未完善，这可能导致偏见和歧视、隐私泄露、错误信息传播以及模型决策的不可解释性等问题的出现。
侵犯知识产权	由于大模型在训练数据的采集和使用过程中可能缺乏明确的法律规范，存在被恶意利用的风险，这可能导致数据滥用和侵犯知识产权等问题。
数据泄露	数据泄露是指用户在与大型模型交互时，其敏感信息（如财务资料、交易记录和个人识别信息等）可能在发送请求或接收响应时被非法获取。这类信息泄露一般发生在模型训练、分析推理或与外部系统的数据交换过程中。
数据窃取	攻击者通过 API 查询重构模型参数（如梯度泄露），威胁知识产权（引用 IEEE S&P 2023）。
RAG 投毒	外挂知识库中的恶意文档可诱导模型生成虚假金融政策解读（案例：某券商智能客服因知识库污染误导用户操作）。
量子计算威胁	Shor 算法对 RSA 加密的破解风险（引用 Nature 2023 量子安全白皮书）。

三、金融大模型安全评估机制

为有效释放金融大模型的技术价值并防控系统性风险，需建立全生命周期安全评估机制。随着监管框架的完善，安全评估已深度融入模型研发、部署及监管环节，形成覆盖技术风险、伦理隐患与合规要求的三维防控体系。基于金融领域风险的内生性特征，本评估体系聚焦以下核心要素：（1）多模态安全评测数据集建设；（2）智能评估方法体系创新；（3）智能攻防测试环境部署。

3.1 多模态安全评测数据集建设

构建结构化安全评测数据集是实现精准风险识别的基础。该数据集需涵盖技术漏洞、伦理冲突、法律违规和社会影响四个维度，重点解析模型在虚假信息传播、有害内容生成、隐私泄露及版权侵犯等场景中的风险机理。具体包含：

(1) 公平性评估数据集：针对性别、种族、年龄等敏感维度，建立模型决策偏差量化分析基准；

(2) 毒性内容识别数据集: 构建网络有害言论特征库，强化负面信息过滤能力；

(3) 隐私合规数据集：模拟数据脱敏失效、特征泄露等场景，验证隐私保护机制有效性；

(4) 对抗韧性数据集：系统测试模型对提示注入、越狱攻击、对抗样本等威胁的防御能力；

(5) 综合评估数据集：集成多维风险因子，构建模型安全性能全景评价矩阵。

3.2 智能评估方法体系创新

3.2.1 跨学科协同评估机制

建立由技术专家、伦理学者、法律从业者组成的联合评估团队，开发统一的风险语义体系。通过“风险转译器”实现技术领域的算法缺陷与非技术维度的伦理法律要求之间的映射转换。

3.2.2 双模评估架构

人工评估层：依托专家知识库进行深度案例分析，重点处理文化适配性、伦理模糊性等复杂场景，建立多级复核机制控制评估偏差；

智能评估层：采用规则引擎 + 深度学习的混合架构，基于 MITRE ATLAS ™框架构建自动化检测系统，实现实时风险分类（准确率≥ 95%）与攻击模式识别。

3.2.3 动态攻防验证平台

部署红蓝对抗测试环境（图 2），其中：

红队模拟：实施提示工程攻击、模型越狱、数据投毒等 7 类典型攻击向量；

蓝队防护：运用注意力熵检测（F1 ≥ 89%）、上下文一致性校验等防御技术；

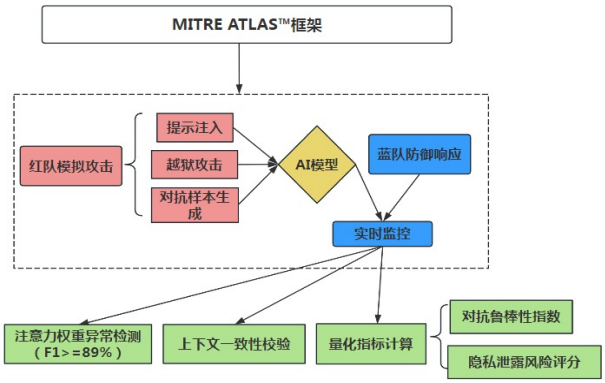


图 2 红蓝对抗测试环境

量化指标：对抗鲁棒指数（成功防御率 ×100%）、隐私泄露熵值（Σ 敏感字段泄露概率 × 权重系数）。

3.2.4 风险优先级管理系统

建立三级风险响应机制：

Ⅰ级（法律红线）：立即熔断机制；

Ⅱ级（隐私泄露）：72 小时应急响应；

Ⅲ级（技术缺陷）：版本迭代修复。

3.3 智能攻防测试环境部署

基于 MITRE ATLAS ™框架构建动态验证体系，关键组件包括：

- **攻击仿真模块：**集成对抗样本生成器、越狱攻击工具包、数据投毒模拟器；
- **防御监测模块：**部署多模态传感器，实现模型输出异常检测（响应延迟 <200ms）；
- **战术知识库：**标准化攻击模式（如特权提升、数据篡改）与防御策略的映射关系；

该体系通过持续对抗训练提升模型安全性能，经实测可使关键风险阻断效率提升 40%，误报率降低至 5% 以下。

金融大型模型面临复杂的安全威胁，包括模型固有的安全隐患，如数据泄露和隐私保护问题，以及来自外部的威胁，如恶意攻击和模型滥用等。为了增强金融大型模型的安全性，需要采取一系列策略，包括但不限于提高数据质量、加强模型训练、减少错误输出、进行安全评估和加强硬件安全等，以确保数据的安全性、算法的可解释性、模型决策的可靠性、应用的合法性和环境的安全性。至于外部风险，则主要依靠国家法律、法规和行业标准来进行管理和规范。

四、金融大模型安全治理

4.1 金融大模型安全框架

金融大模型安全框架的核心目标是构建可信、可控、可解释、可溯源的 AI 系统，通过全局视角覆盖模型全生命周期（训练、生产、应用）的安全防护。其本质是在金融业务场景下，平衡大模型的性能与安全风险，确保符合监管要求并抵御复杂攻击。金融大模型安全治理分为治理层、基础层、核心层、运维层和对抗层。金融大模型安全治理实施关键点：

合规驱动设计：将监管要求（如银保监会 AI 风控指南）拆解为技术指标（如偏见检测 SDK 集成）。

动态防御体系：建立“监测 - 分析 - 处置”闭环，例如通过 AI 安全态势感知平台联动攻防。

人机协同机制：在关键决策（如大额交易）中设置人工干预熔断点，平衡效率与安全。

供应链安全：对第三方组件（如预训练模型、开源库）进行许可证审查和漏洞扫描。

4.1.1 治理层（大模型安全管理）

法律法规：需对接《数据安全法》《个人信息保护法》及金融行业规范（如央行 AI 应用合规要求），建立模型开发与应用的合规边界。

安全评测：通过固定数据集测评（如 CERT 基准测试）和红队对抗测试（模拟攻击者视角），验证模型鲁棒性。

监管监控：部署实时监测工具（如日志审计系统），对模型输出内容进行合规性检查（如反欺诈、反歧视）。

技术结合点：需建立“合规知识库”，将法规要求转化为可量化的模型评估指标（如偏见检测阈值）。

4.1.2 基础层（数据安全）

数据清洗：采用 NLP 技术（如敏感词库、实体识别）过滤金融数据中的隐私信息（如账户号、身份证号），结合差分隐私对训练数据加噪。

数据防泄漏：基于数据分类分级（如 GDRP 等级划分），对敏感数据实施动态脱敏（如掩码、泛化），并通过区块链记录数据流转路径。

隐私计算：联邦学习（跨机构训练）与同态加密（密文推理）结合，实现“数据不动模型动”的安全协作。

挑战：金融数据高敏感性与大模型训练数据需求的冲突，需通过多方安全计算缓解。

4.1.3 核心层（模型安全）

模型结构防御：添加伪节点或冗余参数混淆模型架构。

查询控制：限制 API 调用频率，对敏感查询返回模糊结果（如故意降低置信度）。

防篡改：使用区块链技术固化模型参数哈希值，结合 TEE（可信执行环境）防止运行时篡改。

可解释性：集成 SHAP、LIME 等工具生成决策依据，满足监管审计需求。

可溯源性：为模型输出嵌入水印（如隐藏标记）或元数据（如数据来源 ID），支持追溯至原始训练样本。

减少幻觉：在金融场景中，通过知识图谱增强（如将市场规则编码为约束条件）提升事实一致性。

金融特性：需优先保障可解释性（如信贷拒因说明）和溯源性（如交易异常追踪）。

4.1.4 运维层（环境管理）

隔离与检测：部署微服务架构，将模型推理、数据存储、用户接口分离，通过零信任网络控制访问；集成 SOAR（安全编排与自动化响应）平台，实时分析日志（如 K8s 容器日志）并触发告警。

冗余与熔断：多模型投票机制（如 Ensemble Learning）降低单点故障风险；设置错误率阈值（如股价预测误差 >5% 时触发人工复核）。

漏洞挖掘：利用模糊测试（Fuzzing）模拟异常输入（如恶意构造的金融表单），结合形式化验证（Formal Verification）分析模型逻辑缺陷。

投毒检测：在数据预处理阶段加入对抗样本检测模块（如检测异常数值分布），对开源组件（如 TensorFlow 依赖库）进行供应链安全扫描。

4.1.5 对抗层（攻防安全）

1）防后门攻击

模型剪枝：识别并移除冗余参数（如 Neural Cleanse 算法）。

输入预处理：对用户输入进行规范化（如限制交易金额格式）。

2）防模型窃取

差分隐私：在模型输出中添加噪声（如 Laplace 机制）。

水印技术：通过触发特定输入（如隐藏密钥）使模型输出特征指纹。

3）防注入攻击

提示注入防御：对用户输入进行上下文校验（如限制问答范围）。

对抗训练：在训练数据中加入对抗样本（如文本扰动）。

4）金融场景强化

需重点防范针对交易系统的逻辑攻击（如操纵模型预测股价），可通过沙箱环境模拟攻击场景。

4.2 多模态风险防御与可解释性增强框架

通过动态风险感知、自适应防御与可解释性增强三个核心模块的协同作用，实现对金融大模型全生命周期的安全治理与合规支持。针对金融大模型在复杂应用场景中的安全与透明性需求设计安全治理框架（图 3）。该框架支持实时监测并量化金融大模型输入、处理和输出环节的多模态风险，包括数据投毒、对抗攻击、跨模态逻辑冲突等威胁。

4.2.1 框架协同工作流程

1）输入阶段

多模态数据（如财报 PDF、实时交易流、路演视频）

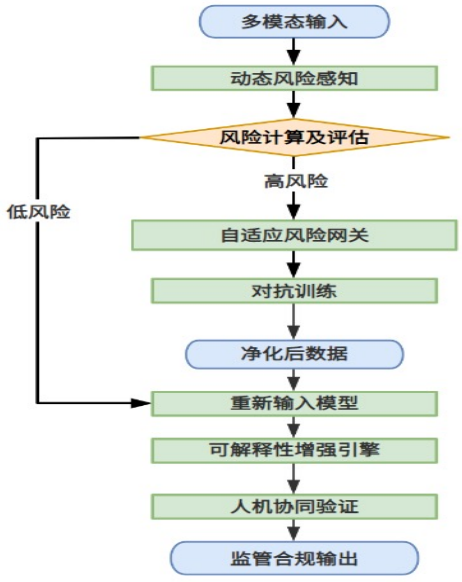


图 3 多模态风险防御与可解释性增强框架

进入动态风险感知层，经跨模态对齐与风险熵计算后分类为低风险（直接进入可解释引擎）或高风险（触发防御网关）。

2）防御阶段

高风险数据依次经过对抗样本过滤、差分隐私处理与模型微调，净化后的数据重新输入模型。

3）决策与解释阶段

可解释引擎实时生成决策逻辑链，并通过人机协同界面（如监管仪表盘）提供交互式解释。若检测到监管规则冲突，系统自动冻结交易并推送警示。

4.2.2 多模态动态风险感知

实时监测并量化金融大模型输入、处理和输出环节的多模态风险，包括数据投毒、对抗攻击、跨模态逻辑冲突等威胁。

异构数据对齐：采用跨模态注意力机制（Cross-modal Attention）融合文本、图表、时序数据。

实时风险量化：基于改进的 CVAE（Conditional Variational Autoencoder）构建风险概率图，动态更新威胁等级。

4.2.3 自适应防御网关

针对识别的高风险输入或模型中间状态，实施动态防御策略，阻断恶意攻击并修复数据 / 模型偏差。

对抗鲁棒性增强：引入 Meta-Attack 模拟器，生成多模态对抗样本进行防御预训练。

数据层防护：设计差分隐私（Differential Privacy）噪声注入策略，平衡数据效用与隐私保护。

4.2.4 可解释性增强引擎

为模型的决策过程与输出结果提供实时可解释性支持，满足监管审查与业务问责需求。

过程可解释：在 Transformer 层嵌入可解释神经元（ProtoPNet），实时可视化特征重要性。

结果可审计：生成符合 FINRA 标准的自然语言报告，自动标注风险决策依据。

冗余熔断机制：采用多模型投票（Ensemble Learning）降低单点故障风险，设置错误率阈值触发人工干预。

五、总结与展望

本文提出金融大模型动态安全治理框架，通过多模态风险防御机制和可解释性增强机制，能够有效地提升大模型的安全性和可解释性，提出从“被动合规”转向“主动免疫”的治理范式创新。未来的研究可以进一步优化框架的性能，构建联邦学习架构下的跨境风险联防体系，提高其在不同场景下的适应性及安全性。可以探索更先进的多模态数据融合技术，提高模型对复杂数据的处理能力。研究更有效的风险识别与评估方法，提高风险检测的准确性和及时性。还可以加强可解释性技术的研究，使大模型的决策过程更加透明和可理解，为大模型的广泛应用提供更坚实的安全保障。

参考文献：

[1] Goodfellow I, et al. Explaining and Harnessing Adversarial Examples. ICLR 2015.

[2] 中国人民银行 . 金融领域人工智能应用安全指南 . 2023.

02 前沿技术应用

基于大模型技术的证券金融知识库智能问答系统

潘建东，梁彬，刘国杨，王赵鹏，孙冰，马张晖，尹序鑫，訾顺遥
中信建投证券股份有限公司 | E-mail：mazhanghui@csc.com.cn

摘　要： “十四五”数字经济发展规划给金融行业数字化转型提出了新要求，金融知识的供给形式有待进一步创新。本文首先以提高金融知识供给效率为需求导向，参考金融市场基础知识问答、证券市场法律法规知识导航等实际应用场景，分析大模型的业界发展态势，形成知识实体识别和知识库问答两种大模型下游任务；然后将行业基础知识划分专题，并以“机器 + 人工”标注的方法构建小样本高质量训练语料；最后通过根据 4 种开源大模型在下游任务中的性能表现，选择性能最优者作为基础模型，按照“优质语料 + 预训练大模型 + 微调”的研究思路，构建证券金融知识库智能问答系统。

关键字： 大模型；参数微调；知识库问答；证券金融

一、引言

传统的金融证券行业在知识管理与获取的过程中存在供需机制不完善: 首先，证券从业人员需要在行业趋势预测、投资策略探讨及专业知识巩固等方面花费大量时间和精力，其决策过程往往需要收集多维度行业知识；其次，分散在企业各部门和机构的知识和经验难以沉淀成组织智慧，知识共享的效率和可靠性难以保证。现代信息技术的飞速发展促进了金融行业迈入数字化转型新篇章，如何提升企业内部的知识运用效率和管理效能、降低一线员工学习成本，成为金融行业落实“开源节流、降本增效”这一宏观战略的首要议题。

作为通用人工智能（Artificial General Intelligence，AGI）的代表，大语言模型（Large Language Model，LLM）的参数规模已达到万亿级别。LLM 的“智慧涌现”，为自然语言处理领域中的两大难题——命名实体识别（Named Entity Recognition, NER）和知识问答提供了新的方法，进而推动拟人化、智能化的知识库问答成为可能。LLM 在大规模数据训练过程中可以自动学习一些高级复杂的功能，拥有更准确的逻辑推理学习能力，在很多方面都拥有了接近人类认知的表现 [1]。聊天生成式预训练变换器 （Chat Generative Pre-trained Transformer，ChatGPT） 的出现将大语言模型技术推向了爆发阶段。互联网公司与算法科技企业深耕大模型领域，ChatGPT4.0、文心一言、Baichuan-13B 等大模型产品如雨后春笋般涌现。

本研究在把握金融行业知识供需机制不完善的痛点、跟踪人工智能领域技术前沿的基础上，以证券法律法规相关知识为例，通过利用预训练大语言模型，按照“优质语料 + 预训练大模型 + 微调”的大模型技术研究思路，训练具有语义分析、上下文关联和生成能力，能够适应知识对象实体识别、关系抽取、知识问答等多种下游任务的深度神经网络模型，构建基于大模型技术的证券金融知识库智能问答系统，为金融证券从业人员提供高精度、高效率的智能知识供给渠道，助力金融从业人员解决在行业趋势预测、投资策略探讨及专业知识巩固等方面遇到的实际问题。

二、知识库智能问答技术相关研究

构建知识库智能问答系统，其技术路径对应着自然语言处理领域中的命名实体识别和知识库问答。在自然语言处理领域，命名实体识别（Named Entity Recognition，NER）是信息抽取的第一个关键环节，旨在从大量非结构化的文本中识别出命名实体并将其分类为预定义的类型，为知识库问答中的自然语言处理任务提供基础支持；知识库问答（Knowledge Base Question Answering，KBQA）借助命名实体识别生成的精度高、关联性强的结构化知识，为给定的复杂事实型问句提供精确的语义理解或解析，并在知识库中查询推理来得到准确、简短的答案。

2.1 命名实体识别

命名实体识别是信息提取的关键子任务，它能将自然语言文本中的专有名称分为个人、地点、组织名称等类别，其准确性对诸多 NLP 应用极为重要。目前 NER 技术主要有 3 类方法：首先是基于规则和词典的方法，早期常用，依靠领域专家手动开发规则、词典，无需标注数据，能简单处理文本实体，但依赖专家知识，人工成本高，且难以

面向新领域、新实体类型或新数据集迁移和扩展。其次是基于统计的方法学习，涵盖有监督和无监督学习。近年来，基于特征的有监督学习是主流，将 NER 当作多类分类问题，用特定特征集训练模型。无监督学习则因依赖数据本身，需大量高质量数据提升性能，且缺乏领域专业知识，实体识别准确率难以保证。最后是基于深度学习的方法，随着互联网文本数据增长和深度学习技术进步而大量涌现。它无需专家特征工程，以端到端方式从原始输入学习特征表示，多层神经网络能提取复杂特征，经非线性激活函数预测实现任务，在 NER 任务中优势显著，可有效提高识别准确率和效率。

2.2 知识库问答

目前主要有两类知识库问答方法：一类基于语义解析（semantic parsing）[2]，另一类基于信息检索（information retrieval）[3]，基于语义解析的方法以符号化表征表示问句的语义，基于信息检索的方法以稠密向量表示问句的语义。相比基于信息检索的方法，语义解析的方法能够应对更多类型的问句，例如含有实体约束、类别约束、数值比较、数值排序等的问句。

近年来，得益于深度学习与自然语言处理技术的进步，知识库问答场景可借助中大型神经网络作为预训练模型，根据下游任务进行微调，以实现迁移学习。这些模型的参数规模巨大，网络结构十分复杂，在设计的预训练任务下从大规模无标注文本中学习自然语言上下文相关的意义和结构，可以捕捉到更丰富的语言特征，从而能够更好地应对各种语义解析这一自然语言处理任务。基于大模型技术的知识库问答，不仅有效规避了模型算法在“实时性”与“事实性”等方面存在缺陷，而且能借助结构化知识提升问答精度，是目前构建需要精准回答的领域知识问答场景的最佳技术方案。

三、证券金融知识智能问答大模型构建

本研究基于金融证券行业在知识管理和获取过程中存在的供需机制不完善等一系列痛点，参考证券市场法律法规知识问答、证券市场法律法规知识导航等实际应用场景，形成知识对象识别和知识库问答两种大模型下游任务，然后获取相关知识语料并构建小样本高质量标注语料，训练

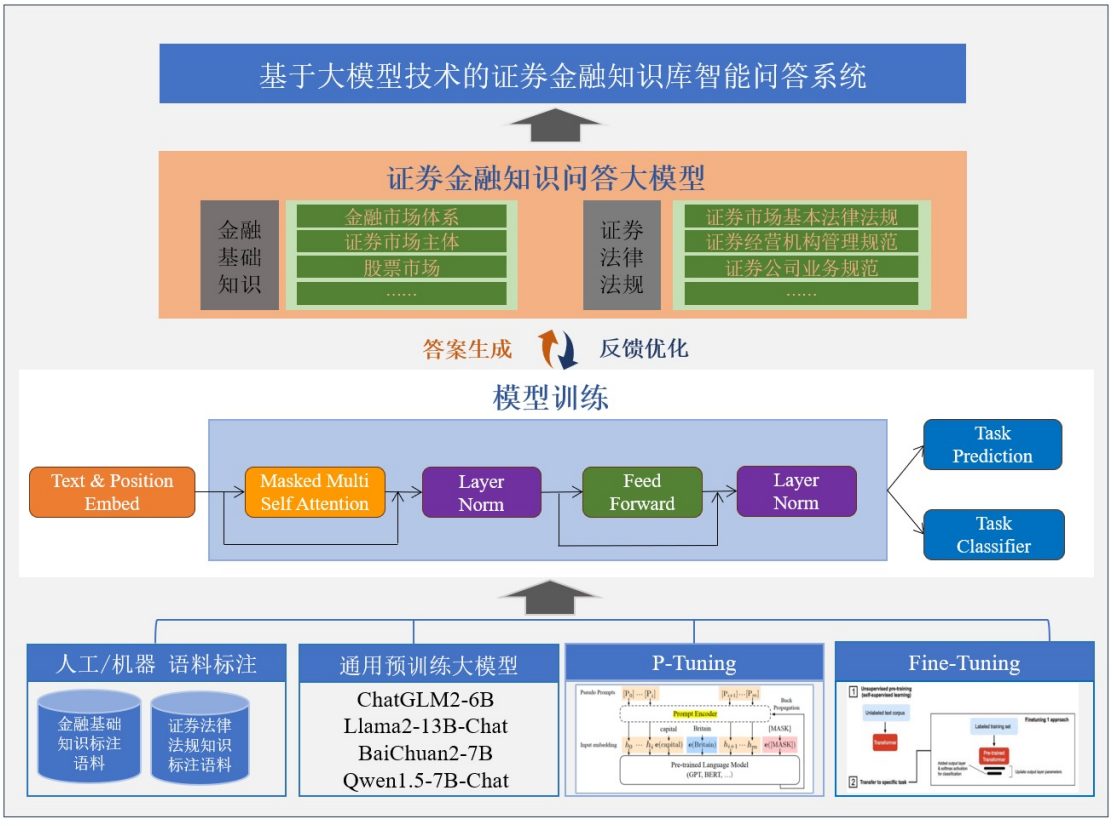


图 1 证券金融知识库智能问答系统总体框架图

金融证券领域知识大模型，构建证券金融知识库智能问答系统，总体框架如图 1 所示。

3.1 获取证券金融知识语料

本研究首先基于证券人员从业资格考试基础教材，划分出金融市场基础知识、证券市场法律法规两个专题的知识对象。然后，依据两个专题的知识对象确立实体类型。例如，证券市场法律法规专题的知识对象包含“证券市场基本法律法规”“证券经营机构管理规范”“证券公司业务规范”等；而知识对象 “证券经营机构管理规范”的实体类型包含“证券公司治理的基本要求” “证券公司与股东间关系规定”“证券公司对客户诚信义务”等，关键词为内部治理、股东关系、对客义务等。

面向证券金融知识的两个专题，依据证券人员从业资格考试基础教材中相关知识，经过文本清洗、内容重编译等数据预处理，转化为 1135 条证券金融知识语料。

3.2 小样本高质量标注语料集

分别针对证券金融知识对象识别和证券金融知识库问答两种下游任务，形成预训练大模型微调需要的小样本高质量标注语料。

1）标注知识对象实体。利用机器自动标注和人工标注相结合的方式对 1135 条知识语料进行标注。首先对每种知识对象实体类型随机抽取 20 条知识语料，使用标注工具 doccano[4] 进行人工标注。然后使用标注工具 DoTAT，对其他语料自动标注，再通过人工校对形成大模型标注语料。其中，数据属性包含知识实体、知识实体类型及其位置。

2）构建训练集和测试集。通过相关业务专家人工构建的方式，在转化的 1135 条证券金融知识语料中挑选出最契合实际业务场景的标注语料，转换出 220 条问答对。其中，70% 的标注语料和问答被用作大模型微调，30% 作为测试集检验模型性能。

3.3 构建证券金融知识问答大模型

证券金融知识问答大模型的构建过程分为模型选取、预训练模型微调、模型优化三个步骤。

1）基础模型选取。本研究基于大模型对中文语言的逻辑推理能力和是否开源可商用两方面，综合考虑基础模型的选择范围。在调研过 SuperCLUE 针对中文通用大模型的逻辑推理、 知识百科、语言理解、生成创作、对话等各种能力的排名后，结合大模型是否开源和公司实际算力配置，选择 ChatGLM2-6B （ChatGLM）、 BaiChuan2-7B（Baichuan）、Llama2-13B-Chat （Llama2） 和 Qwen1.5-7B-Chat（Qwen）作为基础模型。

2）预训练大模型微调。在大模型的 fine-tuning（微调）过程中，一种常见且高效的方法是使用 LoRA（Low-Rank Adaptation）[5]。LoRA 的核心实现方法是在模型结构中添加一个旁路分支。在训练阶段，主模型的参数被冻结，只更新旁路分支的 AB 矩阵参数。预测时，则将主模型的输出和旁路分支的预测结果进行合并。具体来说，LoRA 主要对 Transformer 结构中的多头注意力（multi-head attention）部分进行微调。

本文还采用了 Prompt-tuning 微调 [6]，该微调方法允许创建特定任务的提示，能够灵活适应各种下游任务，对于超过 10 亿参数量的模型来说，在不需要修改模型参数的低成本前提下，小样本甚至是零样本的微调性能也能够极大地被激发出来。

3）大模型的各领域实践研究中，一般将模型生成的文本存在不遵循原文或者不符合事实的现象称之为大模型幻觉 [7]。本文针对存在的大模型幻觉，采用数据清洗和检索增强两种方法进行大模型的文本生成优化。一方面，通过对数据集进行标注语料去重和人工筛选，清洗掉影响模型文本生成的脏数据；另一方面，采用检索增强生成（Retrieval Augmented Generation，RAG）方法 [8]，增强模型对领域知识的理解和生成能力。

四、结果与分析

命名实体识别和知识库问答作为构建知识库智能问答系统的关键技术路径， ChatGLM、Llama2、Baichuan 和 Qwen 等大模型在这两种下游任务中的性能表现，是判断其能否作为知识库智能问答系统基础模型的重要参考依据。

4.1 知识实体识别结果分析

在知识实体识别任务中，本文主要使用 LoRA 微调方法提升模型实体识别能力，并采取精确率（Precision，Pre）和召回率 （Recall）作为评价知识实体识别性能的指标 [9]。

4.1.1 精确率分析

精确率（Pre）是指预测为正例的样本中预测正确的比例，用以衡量检测知识实体识别的整体有效性。本文首先对比了微调前后 ChatGLM、Llama2、Baichuan 和 Qwen 大模型在金融市场基础知识、证券市场法律法规两个知识专题下的精确率，表 1 为模型在微调前（before，简称 B）与微调后（After，简称 A）的精确率变化。

表 1 大模型预训练初期精确率（Pre）对比

知识专题	ChatGLM		Llama2		Baichuan		Qwen	
	B/%	A/%	B/%	A/%	B/%	A/%	B/%	A/%
金融市场基础知识	70.7	82.5	66.2	78.3	62.7	75.9	70.5	74.1
证券市场法律法规	76.2	87.9	70.5	77.8	73.5	80.3	74.7	88.4

根据上表大模型预训练初期精确率对比不难看出：

1）微调效果因模型而异。ChatGLM 与 Qwen 在微调后，在两个知识专题上的精确率平均提升幅度为 10%~15%。而 Llama2 和 Baichuan 精确率平均提升幅度均小于 10%。

2）模型对金融知识语料的训练表现不同。无论在微调前还是微调后，ChatGLM 对证券金融知识语料的训练表现相对是最好的，在金融市场基础知识、证券市场法律法规两个知识专题下的精确率处于领先位置。

3）知识对象实体类型的数量对精确率的影响不同。定义了 6 种实体类型的“金融市场基础知识”专题的精确率较低，而定义了 4 种实体类型的“证券市场法律法规”专题的精确率更高。

4.1.2 召回率分析

召回率（Recall）的含义是指在实际为正的样本中被预测为正样本的概率。除了对比了微调前后 ChatGLM、Llama2、Baichuan 和 Qwen 大模型精确率表现，本文还分析了微调前后各大模型的召回率表现，表 2 为模型在微调前（before，简称 B）与微调后（After，简称 A）的召回率变化。

表 2 大模型预训练初期召回率（Recall）对比

知识专题	ChatGLM		Llama2		Baichuan		Qwen	
	B/%	A/%	B/%	A/%	B/%	A/%	B/%	A/%
金融市场基础知识	69.7	70.4	58.1	66.3	62.7	56.1	51.5	56.1
证券市场法律法规	70.2	72.1	67.5	73.8	70.5	66.3	70.7	62.4

根据上表大模型预训练初期的召回率对比，可以看出：

1）不同模型在微调前后的召回率提升各异。其中 ChatGLM、Llama2 大模型在微调后召回率有一定提升，而 Baichuan 和 Qwen 大模型在微调后召回率反而下降了。

2）不同模型对金融知识语料的召回率各异。其中 ChatGLM 表现最佳，在两个知识专题下的微调前后召回率均在 70% 左右。Baichuan 和 Qwen 大模型微调后召回率反而下降。

3）知识对象实体类型的数量对召回率的影响不同。定义了 6 种实体类型的“金融市场基础知识”主题的召回率较低，而定义了 4 种实体类型的“证券市场法律法规”主题的召回率更高。

4）召回率和精确率提高幅度不一致。由于模型参数的微调，在命名实体中正确识别的比例（精确率）越高，识别的实体数量（召回率）相对会有所减少。本文中精确率提高 10%~15% 的大模型，召回率提高幅度在 5% 以下，部分大模型召回率甚至较微调前降低。

4.1.3 小结

本文所研究的命名实体识别任务中，大模型实体识别表现和模型架构、标注的语料种类、实体类型数量等因素有密不可分的关系。其中 ChatGLM 大模型在预训练初期的微调前后的实体识别均有良好表现。

4.2. 知识库问答结果分析

知识库问答作为知识输出的重要环节，借助命名实体识别生成的精度高、关联性强的结构化知识，为给定的复杂事实型问句提供精确的语义理解或解析，并在知识库中查询推理来得到准确、简短的答案。在知识库问答任务中，本文使用 promp-tuning 微调和 RAG 技术优化以强化模型的生成能力，并采用幻觉率（Hallucination Rate，HR）和语义相似度（Semantic Similarity，SS）作为评价大模型答案生成质量的标准。

4.2.1 幻觉率分析

大模型生成结果中，存在幻觉现象的答案占所提供问题答案的比例（幻觉率），是评判大模型性能的重要指标。幻觉率越低，大模型生成答案的质量就越高。针对金融市场基础知识和证券市场法律法规两个知识专题，以 ChatGLM、Llama2、Baichuan 和 Qwen 大模型在微调前后生成的答案的幻觉率，衡量大模型知识库问答的输出的“事实性”。如图 2~3 所示，蓝色进度条代表微调前的幻觉率，橙色代表微调后的幻觉率。

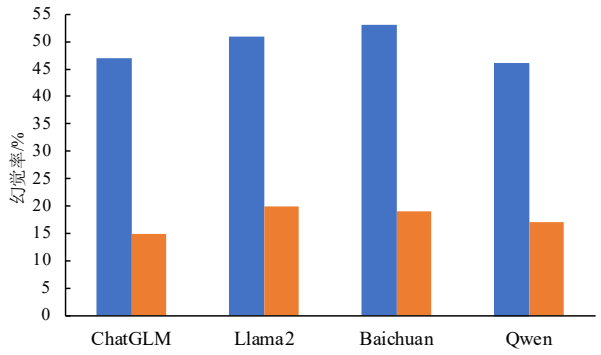


图 2 金融市场基础知识专题微调前后幻觉率

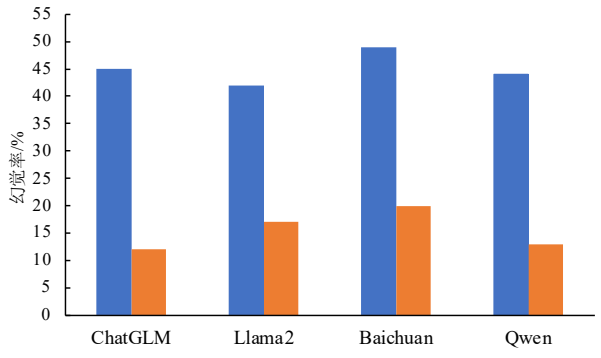


图 3 证券市场法律法规专题微调前后幻觉率

根据上图可知：不同的知识专题的幻觉率不同，“金融市场基础知识”专题的幻觉率较高，而“证券市场法律法规”专题的幻觉率较低；微调后，幻觉率的现象得到缓解，但是不同大模型的幻觉的缓解程度有所差异；微调后 ChatGLM 和 Qwen 的幻觉率在 12%~15%，比其他模型的幻觉率更低。

4.2.2 语义相似度分析

语义相似度（Semantic Similarity）是一种用于衡量两个文本之间相似性的方法，具体的方法为：针对一个知识问答的模型生成文本和原始标注语料，进行分词、去除停用词；使用词频 - 逆文本频率指数方法计算两条答案的词频向量；计算两个词频向量的余弦相似度，值越大表示模型生成的文本答案和标注语料的相似度越大。由于金融市场基础知识专题的语料数据量最高，本文以该专题知识，对比四种大模型生成文本和原始标注语料的语义相似度。

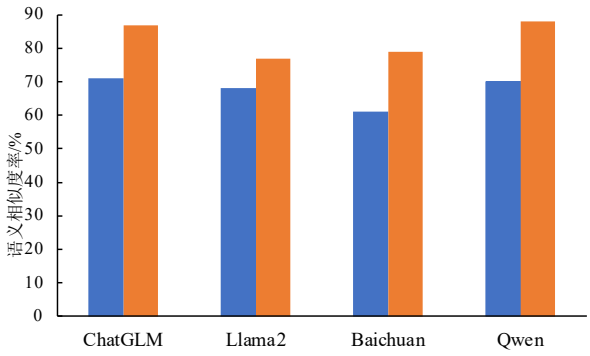


图 4 微调前后的大模型语义相似度

由图 4 可知，微调后所有大模型的语义相似度都得到了不同程度的提高，其中 ChatGLM 和 Qwen 的表现最好，微调后的语义相似度接近 90%，效果出色。

4.2.3 小结

本文在知识库问答任务中采用了 prompt-tuning 微调以及 RAG 技术，使大模型能够深入理解证券金融领域知识，并根据答案生成的质量不断调整模型的输入提示，迭代优化模型性能， 从而灵活适应知识库问答任务。其中 ChatGLM 和 Qwen 大模型在微调前后的知识库问答任务中均有良好表现。

4.3 构建证券金融知识库智能问答系统

基于 4 个模型在知识实体识别任务和知识库问答任务中的表现，本研究选择 Chat GLM 作为构建证券金融知识库智能问答系统的基础预训练模型。证券金融知识库智能问答系统的功能包括以下两个部分：

1）智能知识库问答。针对用户在相关领域知识提出的问题，由大模型生成相应答案。

2）问答信息源查看。在系统给出问题答案的同时，会把模型生成答案参考的信息源提供给用户，用户可以通过点击信息源链接获取信息源的详细信息。

图 5 证券金融知识库智能问答系统问答界面

五、结论

在传统金融证券行业知识供需机制不完善的背景下，本研究在分析大模型业界发展态势的基础上，参考证券市场法律法规知识问答、证券市场法律法规知识导航等实际应用场景，选择 ChatGLM2-6B、 BaiChuan2-7B、Llama2-13B-Chat 和 Qwen1.5-7B-Chat 这 4 种开源预训练大模型，形成知识对象识别和知识库问答两种大模型下游任务，进而构建证券金融知识库智能问答系统。在知识实体识别任务中，通过精确率和召回率对比分析了 4 种大模型的性能表现，可以得出：

1）在证券金融领域的知识对象识别和知识库问答两种大模型下游任务中，ChatGLM 大模型均有良好的表现。

2）针对预训练大模型下游任务的微调和基于 RAG 技术的模型优化可以显著提升大模型的性能，能够有效缓解大模型幻觉问题，提高模型的输出质量， 且微调效果因大模型而异。

参考文献：

[1]ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[EB/OL]. arXiv: 2303.18223, 2023.

[2]YIH W T, CHANG M W, HE X D, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015: 1321-1331.

[3] MILLER A, FISCH A, DODGE J, et al. Key-value memory networks for directly reading documents[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016: 1400-1409.

[4] DAUDERT T. A web-based collaborative annotation and consolidation tool[J]. International conference on language resources and evaluation, 2020: 7053-7059.

[5] DING N, QIN Y J, YANG G A, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature machine intelligence, 2023, 5(3): 220-235.

[6] LIU X A, JI K X, FU Y C, et al. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 61-68.

[7] BANG Y J, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[EB/OL]. arXiv: 2302.04023, 2023.

[8] PENG B L, GALLEY M, HE P C, et al. Check your facts and try again: Improving large language models with ex - ternal knowledge and automated feedback[EB/OL]. arXiv: 2302.12813, 2023.

[9] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[EB/OL]. arXiv: 2307.03109, 2023.

数据地图：大模型助力探索证券数据资产管理新路径

梁钥，聂亚妮，侯立莎，刘敏慧，褚丽恒，王延钰 ，东晓亮，王瑜
申万宏源证券有限公司 | E-mail：liangyue@swhysc.com

摘　要： 本文探讨了数据地图在证券行业数据资产管理中的应用，旨在解决金融市场数字化转型下数据规模增长、来源异构、关联复杂等问题。研究基于大模型和知识图谱技术，构建了字段级数据地图及溯源分析体系，通过大模型解析 SQL 脚本、构建知识图谱和图算法识别关键节点，实现了数据分布、流向和关联的可视化呈现。实验结果表明，该方法有效提升了数据共享与协同能力，降低了数据管理难度，优化了数据治理效率。未来，数据地图与大模型的融合将为证券行业数字化转型提供更智能化、高效化的解决方案。

关键字： 数据地图 ; 大模型 ; 知识图谱 ; 数据资产管理

一、引言

在金融市场持续深化发展与数字化转型加速的浪潮下，证券行业的数据规模和复杂程度呈现出指数级的增长。客户资料、产品信息、行情资讯、交易明细以及宏观经济数据等，以多样形式分散存储于公司内部的各类业务系统、数据库和文件中。以我司为例，目前申万宏源的数据仓库的表生成任务有近 2 万个，涉及表有 23000 余张，字段数量是表的数十倍，并且随着金融科技的发展，数据资产体量迅速增长，数据资产的管理难度激增。除此之外，数据来源广泛且异构，数据标准参差不齐，存在格式不兼容、编码差异等诸多问题。数据在不同系统间的流向复杂，呈现多路径、递归等特性。数据实体之间的关联关系更是丰富多样，包括一对一、一对多、多对多，甚至出现复杂的层次和网状结构。这些问题不仅在技术层面给数据的准确性、完整性与一致性的保障带来巨大挑战，影响数据查询和应用的效率与效果，也在证券行业关键业务领域引发连锁反应。在风险评估与管理环节，数据口径不一，将干扰风险的精准衡量与及时应对；在合规监管方面，数据溯源的复杂性以及与合规标准匹配的困难，增加了合规风险与成本；在客户营销场景中，作为精准营销基石的客户画像依赖于多系统的数据整合分析，无法清晰绘制客户画像将使营销效果大打折扣。

在这种情况下，数据地图作为一种有效的数据资产管理工具价值日益凸显。尤其是在大模型技术迅猛发展与广泛应用的当下，数据地图被赋予了更强大的功能与应用空间，为解决证券行业的数据管理难题提供了新的思路。要实现数据的共享和协同，提升数据理解与查询效率，就需整合数据资源，打破数据孤岛，以可视化可交互的方式呈现数据的分布、流向、关联等信息。随着技术的持续进步与应用的不断探索，数据地图与大模型的深度融合有望助力证券行业迈向更加智能化、高效化的发展阶段。

二、解决方案

数据地图建设的总体目标是构建全覆盖、实时、准确的数据全链路关联关系，基于数据地图打造全场景应用，提高对于全公司数据全景的掌握和重视程度。为此，我们基于申万宏源大数据平台和人工智能平台，以大模型和知识图谱为关键技术，打造了一套字段级的数据地图及溯源分析体系。总体技术架构如图 1 所示：

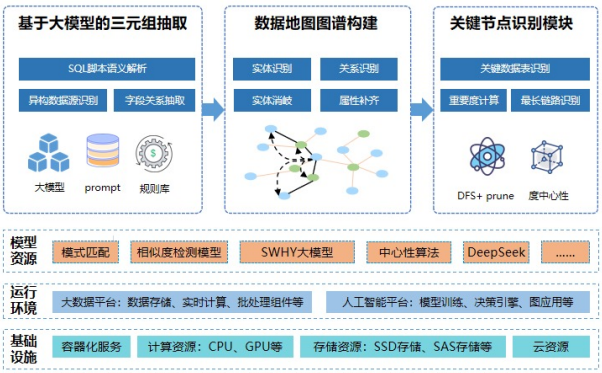


图 1 数据地图总体技术架构图

数据地图主要包括三个关键模块：大模型构建异构数据库的字段级三元组、数据地图知识图谱建设以及基于图算法的关键节点识别模块。首先，采集任务脚本，利用大模型结合元数据进行 SQL 脚本解析和执行日志解析，完成 < 字段、上 / 下游、字段 > 三元组的抽取；其次，将三元组中节点和关系分别进行实体消岐和融合，完成数据地图知识图谱的构建及存储在图数据库中；然后，从业务应用

维度配置初始阈值，通过在数据地图图谱中进行中心性算法的应用，计算每个字段和表的重要值，作为属性更新到数据图谱中。

2.1 关键技术—大模型解析 SQL

在数据资产管理领域，数据地图的智能化构建正经历革命性变革。基于大语言模型的 SQL 解析技术，特别是如 DeepSeek 等先进模型的私有化部署方案，成功突破了传统 ETL 解析工具的三大瓶颈：复杂逻辑处理能力不足、跨系统语法兼容性差、人工维护成本高昂。该技术通过深度语义理解实现了对 SQL 脚本的 " 全量解析 "，可以自动从 SQL 任务脚本中提取出源表、源字段与目标表、目标字段之间的映射关系，从而高效地构建数据地图。

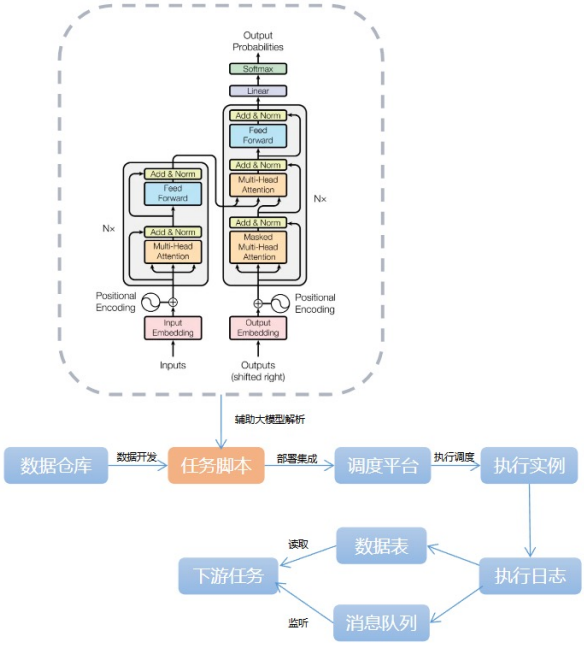


图 2 大模型解析 SQL 技术架构图

SQL 脚本解析任务图 2 所示，在整个过程中，大模型解析 SQL 的技术核心作用于辅助解析脚本环节，其通过动态解析调度状态和变量赋值的上下文信息，实现了数据血缘关系的精准捕获和智能推导。

具体来说，大模型通过自然语言处理（NLP）技术，解析 SQL 查询语句时，能够识别出关键的表名、字段名以及它们之间的关联关系。例如，模型可以识别 JOIN、WHERE、SELECT 等 SQL 语句中的逻辑结构，进一步推导出不同表和字段之间的映射规则。为了提高模型在数据映射中的精度，需要进行相关的调优：首先，进行数据预处理，确保 SQL 脚本格式统一，字段命名规范；其次，采用特定语料库进行训练，使得模型能够更好地理解数据库特有的语法和结构；最后，通过反向工程和验证，结合人工检查和模型反馈机制，对生成的映射关系进行校正和优化，确保最终结果的准确性。

在 SQL 语句的语义解构层面，大模型采用三层解析架构：

词法层： 运用自适应分词算法处理不同数据库方言（如 Hive SQL 与 Spark SQL 的语法差异），准确识别保留字、用户定义对象和嵌套别名。

语法层： 构建动态语法树（AST）时引入权重注意力机制，重点解析 JOIN 条件（占比 32% 的血缘关系）、子查询（27% 的复杂关联）和 CTE 表达式（18% 的递归逻辑）。

语义层： 通过上下文感知模型识别隐式关联，如 WHERE 子句中的跨表条件约束、窗口函数中的排序依赖等。

以典型场景为例，当解析包含 3 层嵌套的 MERGE 语句时，模型可自动标注源系统 -> 临时表 -> 维度表 -> 事实表的完整链路，并识别出多个字段的转换规则（包括类型转换、条件清洗等）。通过这种方式，大模型解决了传统单纯依赖 SQL 语法解析方式进行字段规则抽取准确率低的痛点，提升了智能化水平，降低了人工干预成本。

2.2 关键技术—基于知识图谱构建数据地图

知识图谱作为中国“人工智能 2030”的重点发展方向之一，拥有强大的信息检索能力、语义分析能力和知识表示及推理能力等，是实现认知智能的主要途径 [1]。知识图谱是重要的知识表达方式，其主要内容包括客观世界中的实体及其之间关系，是由“实体－关系－实体”三元组，以及实体及其相关属性－值对的形式组成网状知识结构 [2]。

针对庞大的数据资产难以管理的痛点，本项目使用知识图谱来构建数据仓库的表和字段级的血缘关系，以数据地图作为数据资产图谱的知识表达方式，数据地图如同一张详细的导航图，以直观可视化的方式，呈现数据资产的来源、流转路径以及最终的去向等关键信息。

本项目采用自顶向下的模式构建知识图谱，如图 3 所示，采用数据采集、知识融合、知识加工、知识应用四个步骤构建知识图谱。

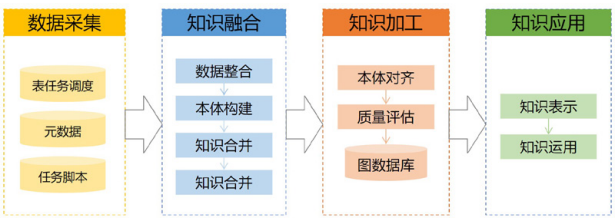


图 3 知识图谱构建流程图

2.2.1 数据采集

本项目所依赖的初始数据源有数据仓库 ETL 任务依赖表、元数据表、以及形成各个表的脚本任务，其中任务依赖表和元数据表均为结构化数据，任务脚本为 SQL 编写，利用大模型技术转化为记录字段级依赖关系的结构化数据库表。

2.2.2 知识融合

经采集得到的数据存在数据重复、数据质量参差不齐、缺少默认值等问题，经过数据整合，能够提升数据质量和数据计算效率，降低数据存储的成本。

数据整合后，进行本体构建。本项目设计的知识图谱覆盖的是金融数据资产领域，图谱 Schema 设计如图 4 所示，模型包含表节点和字段节点两类节点，其中一个表节点代表数据仓库或业务系统的一张表，一个字段节点代表某张表的一个字段，不同表中的相同字段使用不同节点表示。模型中设计了“下游表”、“下游字段”、“表包含字段”三种边类型。“下游表”边用于连接两张表节点，映射表级血缘关系；“下游字段”边用于连接两个字段节点，映射字段级血缘关系；“表包含字段”边用于连接表节点和字段节点，映射字段从属于哪张表。表节点具有负责人、重要度等属性信息、字段节点具有名称、重要度等属性信息。

当任务依赖表中存在 T2 表相关任务依赖 T1 表任务，

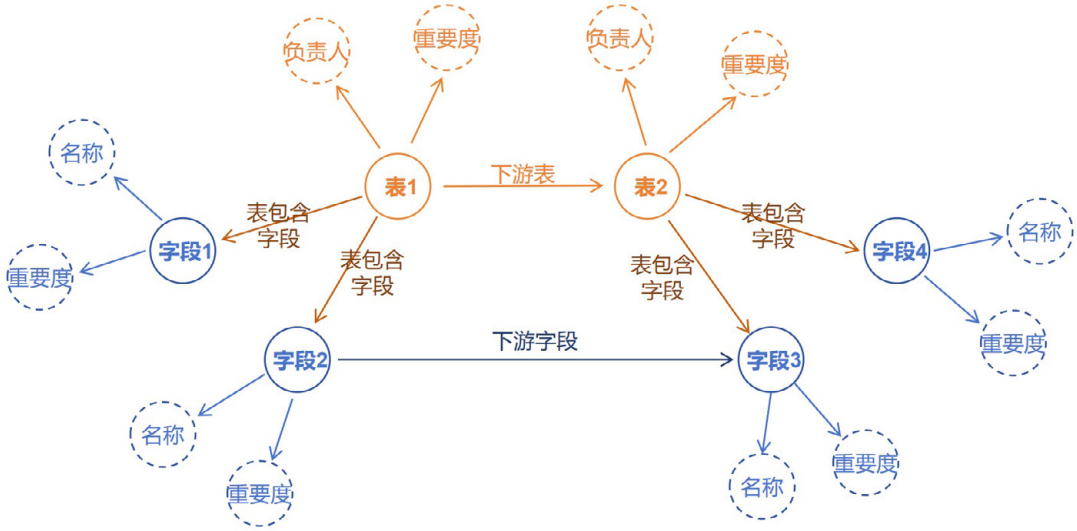


图 4 数据地图知识图谱 Schema 设计图

则图谱中存在实体下游表关系 <T1,DST,T2>，当字段依赖表中存在 C2 字段加工自 C1 字段时，则图谱中存在实体下游字段关系 <C1,DSC,C2>, 表结构信息表中存在 T1 表包含了 C1 字段，则图谱中存在 T1 表实体和 C1 字段实体的包含关系 <T1,TCC,C2>。

2.2.3 知识加工

本体对齐是指在知识图谱构建当中处理本体间的异构性，本文采用基于结构的方法来实现本体对齐，即基于本体层次关系来进行对齐。例如，可能出现不同字段节点的名称和含义一致，如两节点属于相同表节点，那么它们是对齐的，反则反之。

质量评估阶段检查图谱的一致性、时效性、完整性和准确性。预先定义好一系列规则来检查图谱质量，例如设定“如果字段 a 的下游是字段 b，且 a 属于表 A，b 属于表 B，那么表 A 的下游是表 B”，用公式表示为：

若 <a,DSC,b>&<a,TCC,A>&<b,TCC,B> , 则 <A,DST,B>，遍历图谱看是否有违背此规则的情况。

知识加工的结果将存入图数据库中，并通过持续收集和分析新的表信息来动态更新图谱，保证图谱的及时、准确。

2.2.4 知识应用

通过数据地图，我们能够清晰地看到每一项数据资产从产生源头开始，如何在不同的业务系统、数据库以及处理流程之间穿梭流动。数据地图不仅可以展示数据的线性流动，还能揭示数据在不同环节中的加工与转换过程，从而更准确地评估数据的质量和可靠性。

2.3 关键技术—图算法实现关键节点识别

在科学研究领域，复杂网络的研究已成为一个关键领域，其中对复杂网络进行深度挖掘和分析尤为重要。复杂网络中的重要节点，是指那些能够显著影响网络结构和功能的节点。破坏这些节点可能会对复杂网络造成巨大的影响，而信息在这些节点上的传播也非常迅速。因此，识别复杂网络中的重要节点，在理论和现实中都具有重要意义。

在复杂数据网络分析中，识别与分析数据指标的重要

性是理解数据网络结构和功能的关键任务之一。度中心性算法作为一种简单而有效的网络分析工具，为这一任务提供了独特的视角和方法。通过度中心性算法，可以快速识别出数据网络中与最多其他节点相连的节点，这些节点在信息传播、资源分配以及系统稳定性中往往扮演着关键角色 [3]。因此，度中心性算法在数据指标重要性识别与分析中具有广泛的应用场景，能够为复杂系统的深入研究和实际应用提供有力支持。

度中心性算法是图计算领域中一种重要的分析工具，用于衡量图谱中节点的重要性。它通过计算节点的度（即与该节点直接相连的边的数量）来评估节点在网络中的中心性。度中心性越高，表明该节点与其他节点的连接越紧密，通常在信息传播、资源分配和网络稳定性中扮演关键角色 [4]。以下是关于度中心性的计算原理：

度中心性（Degree Centrality）：度中心性测量图谱中一个节点与所有其它节点相联系的程度。一个节点的度越大，其度中心性越高，表示该节点在网络中与更多其他节点相连。在无向图（Undirected Graph）中，度中心性测量网络中一个节点与所有其它节点直接相连的程度。对于一个拥有 g 个节点的无向图，节点 i 的度中心性是 i 与其它 g-1 个节点的直接联系总数（如果是有向图，则需要考虑的出度和入度的问题）

$$C_D(N_i)=\sum_{j=1}^{g}x_{ij}(i\neq j) \quad (01)$$

由上述公式可见，图谱规模越大，度中心性的最大可能值就越高。为了消除图谱规模变化对度中心性的影响，需要进行标准化：

$$C'_D(N_i)=\frac{C_D(N_i)}{g-1} \quad (02)$$

在这个标准化度中心性测量公式中，使用节点 i 的度中心性值除以其它 g-1 个节点最大可能的连接数，得到与节点 i 有直接联系的网络节点的比例。这个比例越大，则表明其中心性越高。

利用度中心性算法，我们可以识别出数据地图中每个节点与其他节点相关联的程度，以此计算结果作为衡量数据地图中节点重要性分析的一个重要参考因素，再结合实际业务使用情况，来最终衡量数据地图中节点的重要性指标程度。

三、应用场景及效果

3.1 数据地图

数据地图是一种用于表示数据源、数据流、数据关系和分析的数据资产管理工具。它可以帮助数据工程师、分析师和其他使用数据的人更好地理解数据结构、关系和加工逻辑，从而提高数据处理和分析的效率。

数据地图是企业数据资产管理中的一种图形化工具，集中展示企业内各类数据的位置，帮助用户快速查找、理解和管理数据。它基于元数据管理，解决数据流转不清晰、查找困难、管理低效等问题，提升数据可供应性和用户体验。企业数据分散在大量数据库和表中，数据地图通过简化复杂的物理层结构，使业务用户能直接获取和理解所需数据，减少对 IT 人员的依赖，从而充分发挥数据的价值。

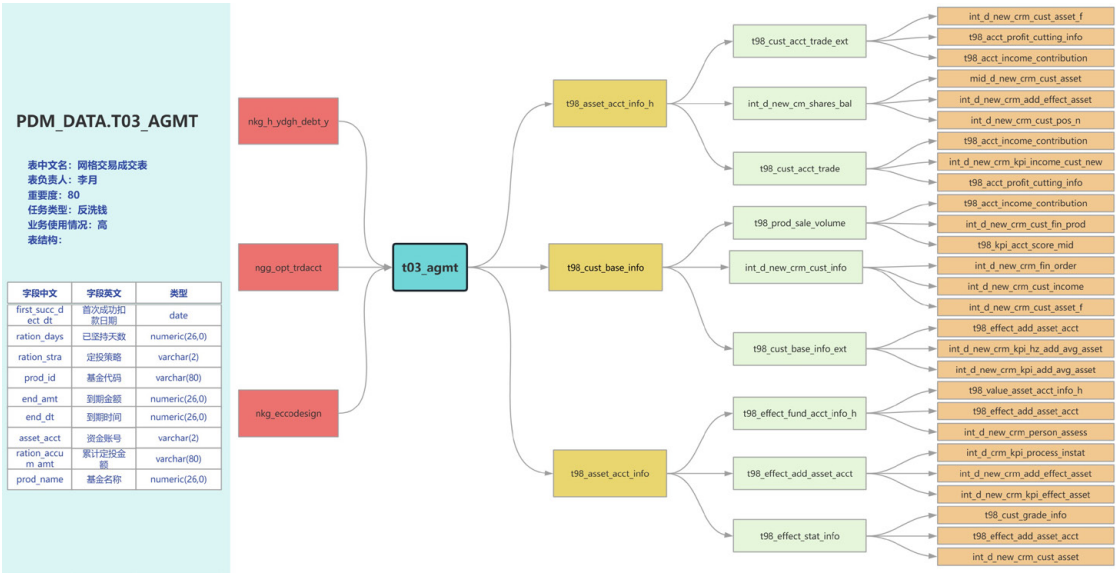


图 5 数据地图展示场景图

对于数据业务人员，可以直接申请服务调用权限并通过资产目录快速查找数据，支持通过关键字、组合条件或搜索历史进行资产检索，检索结果按资产分布类型、匹配顺序和关键属性等多维度呈现；同时，可通过表结构了解字段的业务属性（如含义、使用说明、计算口径）和技术属性（如分区概览、建表语句、样例数据）。对于数据管理人员，可直接编辑资产详情信息，支持组合条件筛选并导出为 Excel 格式，进行挂载主题、打业务标签或批量编辑操作；还可灵活创建和管理目录层级结构、挂载资产，并从数据存储、业务主题、数据质量等多视角进行全域资产分析。

3.2 数据开发助手

3.2.1 数据变更影响分析

在面对数量庞大且链路较长的数据资产时，若其中的某个任务、某张表或字段结构发生变更，很可能因难以察觉而引发事故。此时，数据地图便能发挥重要作用，它可

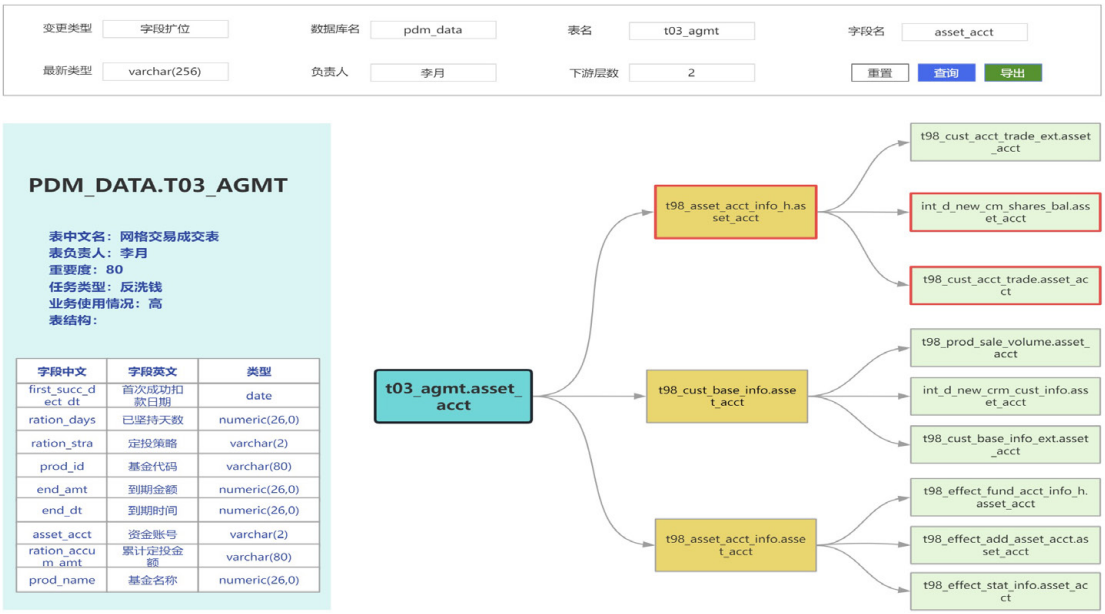


图 6 字段扩位影响分析展示图

以协助开发人员迅速定位变更的具体位置及相应的上下游链路，并且会对受到影响的部分进行高亮显示，以此提醒开发人员完成相应的变更操作。

3.2.2 二次开发查看字段口径

数据地图能够展示字段级别的血缘关系，呈现字段加工逻辑，涵盖了在数据处理过程中所运用的各种算法、规则以及转换操作，包括但不限于数据的筛选、聚合、映射、转换等操作，让开发人员清晰地了解数据在每一个环节是如何被处理和修改的。开发人员在二次开发时，可轻松掌握字段生命周期，了解哪些字段是和当前开发任务密切相关的，实现准确定位。

3.2.3 排查问题准确定位

在开发和运维人员排查问题过程中，数据地图可帮助其准确定位到问题的关键所在，无需在庞大的数据体系中盲目寻找，大大节省了时间和精力。这种精确的定位能力，使得开发人员能够更加聚焦于解决实际问题，而不是陷入对数据细节的无尽排查中，大大提升开发及运维效率。

3.3 数据重要性识别与分析

在数仓建设中，数据链路的复杂性是一个常见的问题，数据地图实现了数仓中复杂链路的可视化呈现。在企业级数据仓库建设过程中，指标体系的建设是核心环节之一。如何对体系中直接进行重要性分类，成为数据仓库成熟建设的一个关键环节，基于数据地图的图计算技术，在企业级数据仓库建设过程中，可以有效地识别和筛选出重要的指标，从而优化数据仓库的指标体系，提高数据治理的效率和质量。本项目在数据地图建设的基础上，运用度中心性算法计算出数据地图中节点的重要性权重，作为衡量指标重要性的基础参数，再结合业务方面反映的业务参数，经过整合加工后计算出指标重要性观测值，进而对其进行归类和可视化展示，作为数据指标优化的依据，对重要性最低的指标进行评估下线，释放平台资源和人力维护成本，同时将资源向重要性较高的指标倾斜，可以有效地提高数据治理的效率和质量。

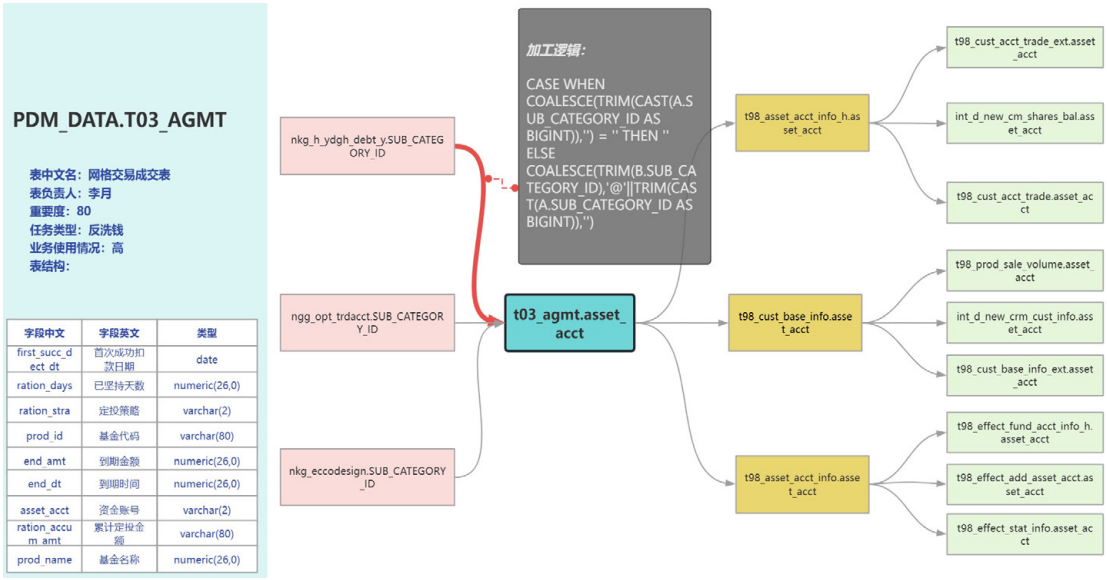


图 7 字段加工逻辑展示图

图 8 数据重要性识别应用场景图

四、结论与展望

本文深入探讨了数据地图在证券行业中的应用及其实现机制，针对当前金融市场持续深化发展以及数字化转型加速所导致的证券行业数据管理难题，提出了基于大模型和知识图谱技术的解决方案。数据地图能够以可视化的方式展示数据的分布、流向及关联，打破数据孤岛，提高数据的共享与协同能力。

随着数据地图与大模型技术的深度融合，证券行业的数据资产管理将进入一个更加智能化、高效化的新阶段。未来，我们计划从以下几个方面进一步推动相关工作，以实现更广泛的应用和更深层次的技术创新：

（1）优化大模型解析能力

提升 SQL 语义理解精度，针对复杂 SQL 脚本（如嵌套查询、递归逻辑等），进一步优化大模型的语义理解能力，确保字段级血缘关系的精准提取；支持多数据库方言，扩展大模型对不同数据库方言（如 Hive SQL、Spark SQL、PostgreSQL 等）的支持能力，提升跨系统数据整合的效率；增强上下文感知能力，通过引入更先进的上下文建模技术（如 Transformer 架构），提升模型对复杂关联关系（如 WHERE 子句中的跨表约束）的理解能力。

（2）完善数据地图知识图谱

开发自动化机制，实时跟踪数据仓库的变化（如新增表、字段变更等），并动态更新知识图谱，确保数据地图的实时性和准确性；为图谱节点增加更多属性（如数据质量评分、敏感性标记等），进一步丰富数据地图的功能；将知识图谱的应用范围扩展到更多业务系统（如交易系统、清算系统等），构建企业级数据资产全景图。

未来，随着技术的不断进步和应用的深入，数据地图与大模型的融合将为证券行业带来更智能化、高效化的发展前景。这不仅有助于提升证券公司的竞争力，还将推动整个行业的数字化转型，为金融市场的发展注入新的动力。

参考文献：

[1] 刘峤 , 李杨 , 段宏 , 等 . 知识图谱构建技术综述 [J]. 计算机研究与发展 ,[2]2016, 53(3): 582-600.

[2] 邱嘉伟 ; 基于网络拓扑的节点中心性度量方法研究 [D]; 东南大学 ;2021 年 .

[3] 陈志 ; 基于结构信息中心性的复杂网络重要节点识别 [D]; 广州大学 ;2024 年 .

[4] 杨蓉蓉 ; 王勤颖 ; 刘凤鸣 ; 基于 PageRank 算法的网络关键节点查找 [J]; 电脑知识与技术 ;2017 年 04 期 .

基于机器学习的券商公募基金精准营销研究

葛菊平，杨映紫，靳朝，潘金锐，乔天国
东吴证券 | E-mail：panjr@dwzq.com.cn

摘　要： 本文探讨了公募基金销售对证券公司业务发展的重要性，并结合当前市场竞争格局及技术趋势，提出通过机器学习模型优化基金营销策略的研究思路。文章首先回顾了券商在基金销售领域的优势和挑战，阐述了精准营销在提升客户转化率和忠诚度方面的潜力。随后，文章设计了以XGBoost为核心的客户购买行为预测模型，通过特征工程、滑动时间窗口采样及不平衡样本处理，解决了大数据背景下的建模难题。研究采用特定时间段的数据进行训练和验证，并通过 ROC-AUC、精准率和召回率等指标评估模型效果，结果显示该模型在分类性能和实践适用性方面表现优异，为证券公司在财富管理业务中的精准化、智能化营销提供了有力支持。

关键字： 大数据；机器学习；精准营销

一、引言

在当今金融市场格局中，公募基金销售业务已成为券商业务的重要组成部分。随着居民财富的持续增长以及资本市场的不断发展，公募基金市场规模逐步扩大，为券商提供了新的业务增长点和盈利机会。据《2021 中国私人财富报告》数据显示，2020 年我国个人持有可投资资产突破 240 万亿元，过往 10 年间年均复合增长率达到 13%，高净值人群数量超过 260 万人，年均复合增长率达到 15%，高净值人群可投资资产总规模达到 84 万亿元，人均财富达到 3209 万元。这庞大的可投资资产为公募基金市场的发展奠定了坚实基础，也凸显了券商在公募基金销售业务领域的广阔发展空间。

在公募基金市场中，券商扮演着重要角色。根据《证券日报》2024 年 9 月 13 日晚间公布的 2024 年上半年公募基金销售保有规模百强榜单，券商在新增的“股票型指数基金保有规模”维度表现突出，保有规模合计达 7757 亿元，市占率为 57.02%，远超银行、独立基金销售机构、保险及代理机构等竞争对手。在“权益基金保有规模”方面，虽整体有所下降，但部分头部券商仍占据重要地位，如中信证券保有规模达 1376 亿元，位居券商榜首。同时，中小券商也呈现出不同程度的增长态势，如华宝证券、山西证券等。这表明券商在公募基金销售领域具有一定优势，但也面临着市场竞争和业务转型的挑战。

随着金融与科技的紧密结合和互联网金融的兴起与快速发展，公募基金的营销环境已经发生了翻天覆地的变化。用户的消费偏好和行为习惯也随之发生了改变。券商借助机器学习和统计模型对客户数据进行深度挖掘，以构建全面而精准的客户画像。通过收集和分析客户的基本信息（如年龄、性别、职业等）、投资行为数据（交易频率、交易金额、投资偏好等）以及风险承受能力评估数据，运用聚类分析、决策树等算法，将客户划分为不同的细分群体。基于这些客户画像，券商能够实现精准营销，针对不同群体推送符合其需求和风险承受能力的公募基金产品，从而提高营销效果和客户满意度。精准营销不仅可以提升客户的购买转化率，还能增强客户对券商的信任度和忠诚度，促进长期合作关系的建立。

用户精准营销离不开购买行为预测，用户购买行为预测也一直是金融领域研究的热点之一。过往研究在传统统计分析方法的基础上，逐渐引入机器学习算法，取得了显著进展。早期研究主要依赖回归分析、时间序列分析等传统方法，但随着大数据时代的到来，机器学习模型因其强大的数据处理和模式识别能力，在用户购买行为预测方面展现出独特优势。如祝歆等人利用 Logistic 回归、支持向量机等机器学习算法构建预测模型，通过实证研究证明融合算法模型在预测效果上优于单一模型；程成等人运用主成分分析法挖掘消费者购买理财产品数据中潜在的、有价值的信息；然后使用 PCA 对网络进行降维，将主成分模拟到 BP 神经网络中建立购买互联网理财产品预测模型，预测模型训练样本结果准确率达 83.61%。

综上所述，公募基金销售业务对券商的重要性不言而喻，用户购买行为预测的研究成果和机器学习和统计模型为券商提供了理论与实践的指引，而当前券商在公募基金销售市场中的表现既显示出其竞争力，也暗示着改进与创新的空间。因此，本文将深入探讨券商如何利用机器学习模型提升公募基金营销推广效果，助力其在财富管理业务领域实现更高效、精准的发展。

二、模型设计

2.1 整体思路

在基金营销领域，多种模型被广泛应用以提升营销效果和效率。例如客户细分模型，运用聚类分析等方法将客户划分为不同群体，针对不同客户群体的特征推荐相应的产品；预测表现模型利用时间序列分析、神经网络等对基金业绩、市场走势等表现进行预测，为客户提供合理投资建议，同时也有助于制定产品推广计划；关系营销模型强调与客户建立长期关系，通过客户满意度调查、忠诚度计划等方式，提高客户留存率和复购率，借助客户口碑传播吸引新客户。在不同情形下需要根据问题的类型和已有资源决定使用的模型。

本文选择将基金营销推荐问题量化为在特定时间点对客户是否购买基金行为的概率的预测模型，基于预测出的概率值设定阈值后即为机器学习领域常见的分类模型。通过对技术问题和业务的拆解，本文的研究重点集中在这几个部分：①与实际业务结合，尽可能选择足够解释模型的变量，或变量群；②在训练模型的过程中选择合适的机器学习模型，以及最适合的参数组合；③选择合适的评价指标，评价模型的有效性；④应用模型指导业务工作，从实践结果再反馈模型优化过程。

2.2 特征工程

在机器学习的实践模型中，特征工程的重要性不容忽视。选择合适的变量可以提升模型性能，增加模型输入的信息量，使模型结果更具有可解释性。同时，变量选择在实践过程中仍然需要结合现有数据资源，做适当的转化，将已有的数据模式转化为更能传递信息的、适合机器学习模型输入的数据形态。这既要求对基金销售各个环节影响因素的充分了解，又需要对数据存储、提取和加工方式和技术的纯熟。因此，特征工程构建高效机器学习系统的关键环节，对确保模型有效性起着决定性作用。

根据基金购买过程的几个关键环节，本文主要选取 6 个大类的特征：1. 基础账户信息特征：包括客户基础身份信息（性别、年龄、职业、学历等），和客户资金账户的基础信息（账户类型、数目及开户时间）；2. 持仓信息特征：包括客户资产、盈亏、股票持仓、基金持仓及其他产品持仓信息；3. 基金交易信息特征：包括客户对场内外基金各种操作的金额及次数；4. 内外部交易属性特征：包括客户在东吴证券和其他机构购买各类产品的记录次数；5. 客户行为属性特征：包括客户在东吴证券 app 上的操作记录；6. 客户投资属性特征：客户的盈利和投资偏好信息。

在此基础上构建训练集、验证集和测试集。如下图所示：

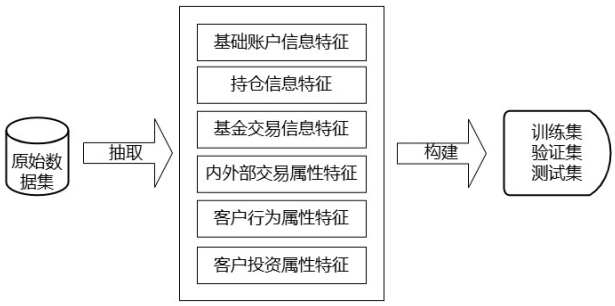


图 1 构建样本流程

由于本文的研究问题和数据集都具有时间属性，因此在建模时需要选取特定时间段的数据训练。训练时提取六个月的数据做特征，三个月的数据做标签。标签设计为三个月内有购买基金为 1，没有购买为 0。

截取数据时选择滑动时间窗口。假设当前时间节点为 t，截取逻辑如下图所示：

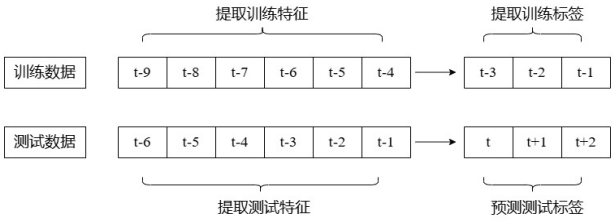


图 2 时间窗口示意

使用滑动时间窗口方式选取数据，有两方面主要原因：一是交易行为带来的时间属性，在预测时应当保证模型与测试数据的时间差不会过长，否则会出现过去的模型与当前时间点的特征不适配的情况；二是生成变量和模型本身都涉及大量的运算，在保证模型有效性的前提下应该尽量减少冗余数据和计算。

2.3 抽样思路

经过滑动时间窗口选取，并去除无效数据后的样本量大约为 120 万条左右，其中标签值为 1 的仅有 4 万左右。正负样本比例接近 1:30，属于样本极度不平衡的情况。如果不进行任何处理，会导致模型分类效果下降：如果按照正常情况选择阈值，模型往往会倾向于将大多数样本预测

为占比大的类别（负类），这会使模型评价指标不能准确反映模型在少数类（正类）上的表现。

通常情况下，样本不平衡会选择过采样、欠采样、混合采样等方式处理。在我们的模型中正样本包含了有购买基金记录的客户信息，负样本为从未有过购买记录的客户信息。对模型来说，正样本包含的信息量更大，权重也应该更大，应该充分学习正例样本；多数负样本都是仅在东吴证券开户后很少操作的客户，这些样本包含的信息量较少，应适当放弃负样本信息。

在训练集中本文选择保留全部正样本，随机抽取三倍于正样本数量的负样本的方式处理。验证集和测试集保持原有比例不变，这样可以模拟出真实情况下的模型预测结果。经过实验验证，这种欠采样方法处理后的模型结果具有较好的稳健性 。

2.4 模型原理

XGBoost 全称为“Extreme Gradient Boosting”，是脱胎于梯度提升决策树 (Gradient Boosting) 的一种机器学习算法。XGBoost 的优势众多，其正则化项可防止过拟合，提高模型的泛化能力；近似算法使它能处理大规模数据，减少计算量；能够自动识别并有效处理稀疏数据，可适应多种实际数据情况；在众多机器学习竞赛和实际应用场景中都展现出卓越的性能，在学术界和工业界等领域都被广泛应用，是一种性能非常强大的机器学习算法。其基本原理如下：

2.4.1 目标函数

给定数据集$D = (x_i, y_i)(|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$，其中 X 为特征项，Y 为标签项，n 为样本数，m 为特征数。对模型输入 x_i，输出预测$\hat{y_i}$，样本标签 y_i。预测过程可以表示为：

$$\hat{y_i} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \qquad (1)$$

其中 k 代表决策树个数，$f_k(x_i)$ 是第 k 个决策树输出的预测值。模型训练的目标是尽可能缩小预测值和真实值的差距，可以被表达为最小化目标函数：

$$L(\phi) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k) \qquad (2)$$

这里 l 是衡量预测值和实际值差距的损失函数，$\Omega(f_k)$ 是函数的正则项。

2.4.2 模型集成

XGBoost 采用加法训练的方式建构模型。它从一个初始值开始，逐步添加新的决策树来改进模型的预测。假设模型由 K 棵树组成，预测值$\hat{y_i}$的可以表示为：

$$\hat{y_i}^{(0)} = initial\ prediction \qquad (3)$$
$$\hat{y_i}^{(1)} = f_1(x_i) = \hat{y_i}^{(0)} + f_1(x_i) \qquad (4)$$
$$\hat{y_i}^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y_i}^{(1)} + f_2(x_i) \qquad (5)$$
$$\cdots$$
$$\hat{y_i}^{(t)} = \sum_{k=1}^{k} f_k(x_i) = \hat{y_i}^{(t-1)} + f_t(x_i) \qquad (6)$$

其中 x_i 是第 i 个样本的特征向量，t 是迭代次数，$f_k(x_i)$ 是第 k 棵树对第 i 个样本的预测值。当添加树时需要尽可能最小化目标函数：

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \qquad (7)$$

在这里选择 MSE（均方差）作为损失函数带入目标函数可得：

$$L^{(t)} = \sum_{i=1}^{n} (y_i - \left(\hat{y_i}^{(t-1)} + f_t(x_i)\right))^2 + \Omega(f_t) \qquad (8)$$

使用泰勒展开式来近似损失函数，可以得到：

$$\tilde{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \qquad (9)$$

这里$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$，以及$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$，分别为损失函数的一阶和二阶梯度统计量。移除常数项后目标函数被简化为：

$$\tilde{L}^{(t)} \simeq \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)] \qquad (10)$$

对于正则项 Ω(f) 可以表示为$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2$。用$I_j = i|q(x_i) = j$表示叶子 j 的样本集，目标函数扩展为：

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$
$$= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda)\omega_j^2] + \gamma T \qquad (11)$$

可以计算出最优权重ω_j^*：

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \qquad (12)$$

对应的最优目标函数为：

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \qquad (13)$$

带入实例值即可得出目标函数的具体值。

2.5 模型评价

对于分类模型的结果进行评价，有很多指标可供选择。基于数据特征和实践意义指向，本文选择了精确率（Precision）和召回率（Recall），和 ROC-AUC 三个指标。

表 1 混淆矩阵

	预测正例	预测负例
真实正例	真正例（TP）	假负例（FN）
真实负例	假正例（FP）	真负例（TN）

精确率（Precision）是指在所有被预测为正例的样本中，真正的正例所占的比例。其计算公式为：

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (14)$$

召回率（Recall）是指在所有实际为正例的样本中，被正确预测为正例的样本所占的比例。其计算公式为：

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (15)$$

精确率表现了对正类样本的预测准确性，召回率体现了对正类样本的覆盖度。在分类模型里，精确率和召回率是互相制约的：如果想要提高召回率，可以降低阈值，此时会有更多的负例被错误地分为正例从而降低准确率；类似的，如果想要提高精确率而提高阈值，此时严格的标准会使得一些正例被错误地分为负例而降低召回率。

在这样的状况下，我们需要根据实际问题的需求来确定阈值，从而达到我们的目标。在证券基金业务的营销问题下，提高阈值意味着营销名单中可能购买的比例增加，但同时有潜在客户被排除在外；如果降低阈值，则尽可能地击中所有潜在客户，又会导致营销任务加重、成本提高。此时我们需要在效率和成本上尽可能地做平衡：如果使用客户经理推广的营销方式，可以根据预算适当提高阈值，提高准确率，控制营销任务总数量；如果使用 APP 推送或短信的方式，则可以降低阈值，提高召回率，确保更多的潜在客户被覆盖到。

ROC-AUC 指标是用来衡量模型整体分类效果的指标。ROC 曲线是以假正率（$FPR = \frac{FP}{FP+TN}$）作为横轴，真正率（$TPR = \frac{TP}{TP+FN}$）作为纵轴的曲线。AUC(Area Under the Curve) 值是 ROC 曲线下的面积，AUC 值越大，模型整体预测的准确率越高，性能越好。

三、实证结果

我们选择 2023 年 12 月 1 日至 2024 年 5 月 31 日的特征数据提取模型训练信息，2024 年 6 月 1 日至 2024 年 8 月 31 日的购买行为提取模型标签训练信息。去除无效客户后随机抽取 10% 的数据按照 3:7 的比例做验证集和测试集，共 135465 条。剩余的 90% 数据经过欠采样处理后，得到模型训练集数据 140036 条。

本文主要使用 Python 平台的 XGBoost 包和 Scikit-Learn 包，选择 XGBClassifier 模型训练数据。采取格子搜索 (Grid Search) 的方式确定若干参数的值，包括估计器数量 (n_estimators)、最大深度 (max_depth)、学习率 (learning_rate)、Gamma 值、Lambda 值等。经过训练后的模型 ROC-AUC 值为 0.858，整体分类效果较好。由于样本正负比例极不平衡，我们会关注分类别的评价指标表现。在尽量保证正例召回率的选择下，模型预测的正负类的精确率和召回率如下：

表 2 预测结果评估

类别	精确率	召回率
正例	0. 341	0. 760
负例	0. 993	0. 993

四、应用结果分析

实践应用方面，我们把模型输出的潜在客户名单与业务专家挑选的可能购买客户名单一起下发给客户经理，由客户经理选择营销对象进行推广。人工和模型结果混合下发的目的一是可以保证名单的有效性，帮助客户经理完成销售任务；二是可以分析销售数据来评价模型结果，并且与人工结果进行比较。

模型上线后，前后共参与了三次基金营销活动。在分析销售数据时，我们将购买基金的客户来源进行分类，共分为四类：第一类客户既在模型输出的潜在客户名单内，又在业务专家筛选的营销名单内；第二类客户只出现在模型潜在客户名单内；第三类只出现在专家筛选名单内；第四类是没有出现在任何预测或预筛选名单内但购买基金的客户。根据该标准的占比统计如下表：

表 3 购买客户来源占比

次数	共同命中	模型预测	人工筛选	未命中
1	42. 71%	15. 95%	18. 75%	23. 58%
2	31. 47%	17. 75%	18. 16%	32. 63%
3	51. 13%	10. 65%	11. 72%	26. 50%

在四类购买客户类型中，占比最高的是人工筛选和模型共同命中的客户，其比例有时甚至可超过一半。这充分表明模型具备极高的稳定性与可靠性，模型的预测结果与实际业务逻辑高度契合，和业务人员的挑选结果也保持一致。除共同命中的客户外，模型单独预测出的购买人数比例与人工筛选出的购买客户比例大致相当，基本处于 10% 至 20% 这一区间范围内。最后一部分是未被预测到的购买客户，占比大致在 20% 至 40% 之间。

五、贡献和不足

本研究提出一种用 XGBoost 模型预测客户购买基金概率的方法。在建模过程中，依照实际问题需要进行特征工程挑选合适的变量，调节样本比例，使用交叉验证的方式筛选最优参数组合，经过调参后的模型正例预测可以达到 76% 的召回率。本文将机器学习思维和模型引入到公募基金推广营销的应用场景，构建出基金购买概率的机器学习预测模型，解决不平衡数据集下对少数类的分类识别问题。模型也同时为东吴证券的财富管理业务提供指导。基于模型的预测结果，可以为营销推广业务做决策参考，有助于业务部门和客户经理提高完成业绩目标的效率，减少较低可能性的推销过程，从而让财富管理业务领域实现更高效、精准的发展。

同时，本研究还存在许多不足之处。例如，我们仅从某一时间段客户行为特征角度获取数据进行研究，忽视了当下金融市场整体表现也是影响客户购买行为的非常重要的因素。随着市场波动，金融消费者信心也会随之发生变化，这种变化可能会在非常短时间内发生。而我们的模型选取过去六个月的特征数据预测未来三个月的购买行为，显然跟不上这样的变化，导致影响模型和实践的稳定性和持续性。后续完善过程应当添加变量以表现金融市场走势和客户信心，提高模型稳定性和适应性。

尽管本研究采用了 XGBoost 预测方法，但该方法在处理复杂的多层次关系时仍存在一些不足之处。例如，单纯使用客户交易相关信息难以有效捕捉产品本身信息对交易行为的影响，可能导致预测结果过于粗糙，没有目标性。未来研究可以探索采用知识图谱的方式改进模型，加入基金产品信息，包括基金类型、基金公司、基金经理过往业绩等变量，与客户信息相关联，可以针对某一具体产品预测出潜在客户名单，进一步提高营销效率，同时可以作为财管业务的日常维护手段，实现公司和客户的双赢。

参考文献：

[1] 慕庆宇，於勇成，杜浩斌 . 公募基金投顾助力中小券商财富管理转型 [J]. 清华金融评论 , (06):87–90, 2022.

[2] 周尚，于宏 . 券商股票型指数基金保有规模合计超 7700 亿元 [R]. 证券日报 , 2024-09-18.

[3] 王莞 . 数字化金融时代下 h 基金的销售策略研究 [D]. 中南财经政法大学 , 2020.

[4] 祝歆，刘潇蔓，陈树广，李静，张天宇 . 基于机器学习融合算法的网络购买行为预测研究 [J]. 统计与信息论坛 , 32(12):94–100, 2017.

[5] 程成，赵华，陶伟 . 数据驱动下消费者购买互联网理财产品意向预测方法 [J]. 软件导刊 , 16(01):108–111, 2017.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system[A]. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

[7] 顾汉升 . 基于机器学习的基金认购行为预测研究 [D]. 上海财经大学 , 2022.

基于图像识别的智能巡检平台研究与实践

李志龙，池烨，洪伟，石晓楠
兴业证券股份有限公司 | E-mail：shixiaonan@xyzq.com.cn

摘　要： 证券行业面临的运行风险日益严峻，统一运行监控难度大，特别是行业中大量存在的可视化界面程序的监控方面，传统巡检难以有效覆盖。人工巡检效率低、响应慢，难以应对行业运行管理的要求。兴业证券建设智能巡检平台，利用自动化技术与图像检测算法，解决巡检中的痛点问题，大幅减少人工操作，显著提升巡检的效率与准确性，实时监控、快速检测异常。巡检平台有效提升了巡检工作的标准化和自动化水平，为业务连续性的保障提供有力支撑。

关键字： 智能巡检；运维平台；图像检测；自动化

一、引言

证券行业面临严峻的运行风险管理形势，对业务连续性的要求高且标准严格。同时，该行业普遍存在系统数量较多、架构复杂及外购比例较高等问题，尤其是在可视化界面程序方面，监控统一性较差，且程序稳定性不足，频繁出现弹窗、崩溃和数据错误等现象。这些问题通过常规监控手段难以及时发现，只能依赖巡检的方式进行检测。传统的巡检方式采用人工逐台检查，效率低下且响应缓慢，难以满足现代行业对快速反应的高标准。因此，及时发现并处理异常情况变得格外关键。为应对这些挑战，证券机构正逐步向巡检智能化转型。兴业证券研发的智能巡检平台通过集中管理与图像算法技术实时检测潜在异常，并自动生成检查报告，显著提高了巡检的速度与准确性。

1.1 传统方案：人工巡检，困难重重

在传统的巡检模式中，证券公司依赖人工定时定点检查程序的运行状态。每个交易日，运维值班人员需要逐一核查各类系统程序，确保所有系统相关的可视化界面正常启动并稳定运行。这种方法不仅消耗大量人力和时间，还容易因人工疲劳或疏忽而导致错误或遗漏。此外，由于应用程序的复杂性和多样性，单个运维人员难以高效识别不同程序的异常点，进一步增加了潜在风险事件发生的可能性。

1.2 智能方案：平台巡检，高效便捷

通过引入智能化巡检平台，该平台实时从各系统主机应用程序获取可视化界面数据，并支持自定义截图范围以进行信息采集。在集中管理的巡检信息环境下，运维值班人员能够获得统一的管理视图，实时监控系统的运行状态。平台采集的截图数据还支持可视化标注，从而促进运维巡检的标准化。此外，平台利用模板匹配算法智能检测程序异常，提升了系统效率，减少了对人工干预的依赖，降低了工作量和人为错误的风险，从而进一步提高了巡检效率。

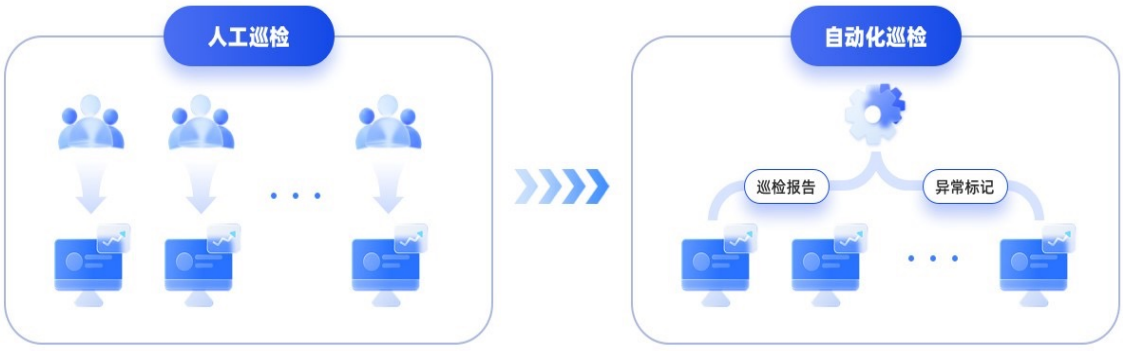


图 1 过去与现在

二、应用智能巡检平台介绍

2.1 平台架构

应用智能巡检平台的目的是提升证券行业系统运行的稳定性和效率。该平台主要包含系统管理、事务管理、任务配置、巡检管理和巡检监控五个功能模块，其系统功能如图 2 所示。

（1）系统管理：系统管理模块是巡检平台的基础，涵盖用户管理、权限管理、角色管理以及日志采集功能。系统管理员通过该模块对系统的核心操作和访问权限进行管理和配置，确保只有经过授权的人员可以执行关键任务。日志采集功能负责记录系统运行过程中发生的关键事件和异常情况，为后续的故障排查和安全审计提供重要数据支持，从而增强巡检平台的安全性。

（2）事务管理：事务管理模块负责巡检任务的综合管理，包括巡检工单的派发、事务查看和统计。在应用系统的日常运行中，值班管理员通过该模块有效地分配巡检任务，并查看任务的完成统计，确保各项任务按时完成并记录过程。事务统计功能则汇总和分析巡检任务的执行数据，帮助运维人员评估任务执行的效率和质量，从而为优化运维策略提供数据支持。

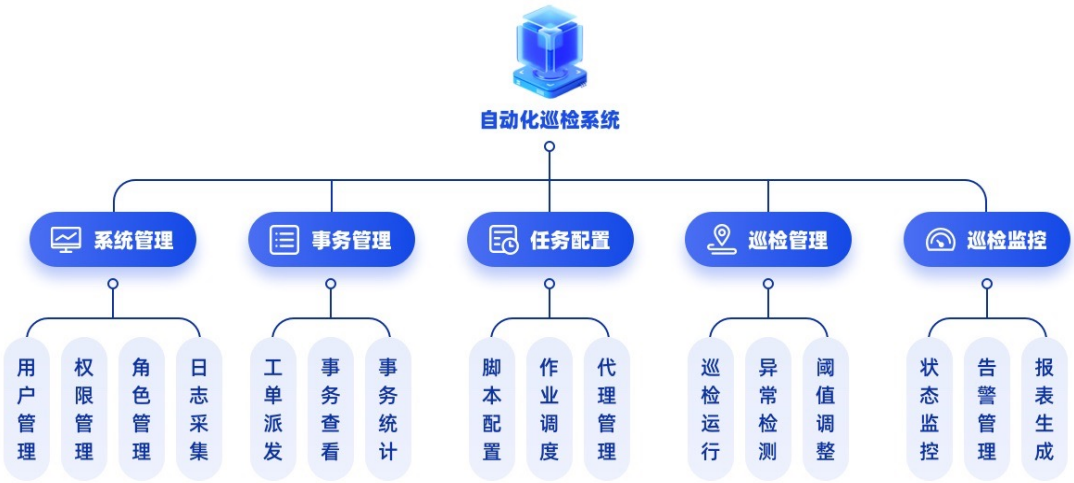


图 2 智能巡检平台功能图

（3）任务配置：任务配置模块专注于巡检任务的具体设置及主机管理，包括脚本配置、作业调度和主机代理管理。在应用系统的巡检过程中，通过合理配置巡检脚本和任务调度，可以提升巡检的多样性和全面性。

（4）巡检管理：巡检管理模块是智能巡检平台的核心组成部分，涵盖巡检运行、异常检测和阈值调整功能。巡检运行模块负责自动执行应用系统的巡检任务，确保在预定时间内进行全面的巡检，并保障巡检任务的正常运行。异常检测功能通过实时监测和分析采集到的巡检数据，自动识别并标记异常情况，同时生成告警通知以提醒运维值班人员，以便他们迅速采取措施处理。阈值调整功能则根据实时数据动态调整异常检测的阈值和参数，从而优化系统的检测精度与效率。

（5）巡检监控：巡检监控模块负责实时监控巡检任务的运行状态、告警管理和报表生成。该模块使运维值班人员能够实时跟踪应用系统巡检任务的运行状态，从而及时发现并解决潜在问题。告警管理功能处理系统生成的告警信息，帮助运维人员快速响应并解决异常情况。报表生成功能则自动生成详细的巡检结果报告和可视化分析，为值班人员提供系统运行趋势分析。

应用智能巡检平台通过集成系统管理、事务管理、任务配置、巡检管理和巡检监控五大模块，全面提升了巡检的效率与准确性。该平台不仅简化了巡检流程，减少了人为错误，还通过实时监控和数据分析，增强了系统的自适应能力和故障响应速度。此外，平台还通过详细的日志记录和报表生成，为运维管理提供了有力的数据支持，有助于优化运维策略并提高决策质量。

2.2 算法介绍

智能巡检平台通过采集各应用系统程序的信息，并经过数据加工、存储和预处理，为自动化巡检算法提供基础输入。在获取图像数据后，采用图像模板匹配算法智能识别异常信息，并将这些信息与标准模板结果整合，以判断

图 3 智能巡检平台流程图

系统是否异常。平台能够及时发出告警，通知值班人员，并生成全面的巡检报告。此外，该平台还集成了异常信息管理、工单下发、值班管理和历史审计等功能，确保整个巡检过程的透明度和可追溯性。

在应用程序运行过程中，其可视化界面可能会遇到多种异常情况，这些异常包括显而易见的弹窗提示和程序白屏现象，以及那些容易被忽视的启动错误信息。平台引入了一种基于差异感知的模板匹配图像异常识别算法，该算法能够有效识别并处理这些异常情况。该算法首先计算近一周内所有正常状态截图的平均图像，通过逐像素分析每张图像的均值，从而生成能够准确反映正常状态的一系列模板。全局模板匹配能够快速识别大范围的异常情况，例如弹窗和程序白屏，并迅速发出告警信息。对于全局匹配得分正常的图像，会进一步进行局部异常匹配，利用差分图像技术对局部细节进行精准检测。这种结构化的监测方式有效提升了异常检测的效率和准确性，确保了应用程序的稳定运行。

全局模板匹配算法在整个程序截图范围内进行分析，以检测程序截图与基准模板之间的整体相似度。这是巡检过程的第一步，旨在判断当前程序截图是否存在大范围的异常。平台采用归一化互相关（NCC）作为相似度度量方法，NCC 有效地衡量两幅图像的整体相似性，从而为快速准确地识别异常提供技术支持。

$$NCC(I_{current}, T) = \frac{\sum_{x,y}(I_{current}(x,y)-I_{current})\cdot(T(x,y)-T)}{\sqrt{\sum_{x,y}(I_{current}(x,y)-I_{current})^2}\cdot\sqrt{\sum_{x,y}(T(x,y)-T)^2}}$$

$I_{current}(x,y)$ 和 $T(x,y)$ 分别为当前程序截图和模板图像在位置 (x,y) 的像素值，$I_{current}$ 和 T 分别为它们的平均像素值。算法对 $I_{current}$ 和模板图像 T 计算 NCC 值，得到全局相似度得分 S_{global}。若 S_{global} 接近 1，则认为当前程序截图与模板非常相似，该程序无大范围全局性异常；若 S_{global} 远小于 1，则极有可能存在全局性的异常。通过设定阈值 θ_s，若 $S_{global}<\theta_s$，则判定当前截图可能存在异常，立即发送异常警告。若接近 1，继续进行局部模板匹配进一步检测。

局部模板匹配是指在图像的特定区域内进行模板匹配，以检测局部区域是否与模板图像存在显著差异。这种方法特别适用于捕捉应用程序截图中局部小范围的异常，如程序未正常运行，启动按钮未变成灰色等可以发现的异常。计算当前截图 $I_{current}$ 与模板图像 T 的差分图像 D，即：

$$D(x,y) = |I_{current}(x,y) - T(x,y)|$$

其中， D(x,y) 表示位置 (x,y) 处的差异值。差分图像 D 可以突出显示程序截图与模板之间的不同之处。对差分图像 D 进行阈值处理，标记出差异较大的区域为高差异区域。通过设定一个差异阈值 θ_D，将所有差异值大于 θ_D 的像素点标记为异常点。在标记出的高差异区域内执行模板匹配。这里同样使用归一化互相关（NCC）进行局部相似度计算，但仅在这些高差异区域内进行。计算得到局部区域的相似度得分 S_{local}。相比全局模板匹配，这一步骤更注重程序截图中局部的异常变化，可以关注到与模板细微的

差异。综合全局相似度 S_{global}、局部相似度 S_{local_final}，计算最终的异常评分 A：

$$A = \alpha \cdot S_{global} + \beta \cdot S_{local\_final}$$

α，β 是权重参数，根据实际交易程序的场景调整以平衡全局与局部。设定最终的异常评分阈值 θ_A，若 A 小于 θ_A，则当前截图被判定为异常，发送异常告警。算法通过全局和局部模板匹配相结合，能够捕捉交易程序中大范围的截图异常，又能检测出局部的小范围变化。全局匹配提供了整体的健康状态评估，能够快速检测出异常。而局部匹配则在特定区域进行更深入的检测，检查出一些细节位置的异常部分。

2.3 平台建设目标

智能巡检平台的建设目标是通过智能化手段，提高证券公司可视化界面程序的稳定性和可靠性，以确保其在高负荷和复杂环境下的高效运行。该平台利用差异感知的模板匹配算法，对可视化程序图像进行异常检测，及时识别并处理异常，从而提升处理效率，减少可能引发的事件。同时，通过减少人工工作量，降低运营成本，并有效提升工作效率。

该平台整合了一系列管理工具，实现了从数据采集到自动报表生成的巡检全流程自动化。通过这种系统化的自动化方法，显著提升了系统的稳定性、可靠性和运行效率，从而确保能够及时发现并处理潜在问题。此外，平台配备的多种工具支持运维团队深入分析系统运行状况，促进运维系统的持续优化与稳定运行，以实现降低成本和提高效率的双重目标。

三、应用智能巡检平台实践

3.1 模板生成

如图 5 所示，左侧为一系列正常图像，右侧为生成的模板图片，巡检平台能够基于某系统的正常巡检图像，自动生成对应的模板图像。平台首先采集一周内多张正常状态下的巡检图像，随后采用逐像素分析的方法，计算出各像素点的平均值，最终生成能够代表系统正常状态的模板图像。

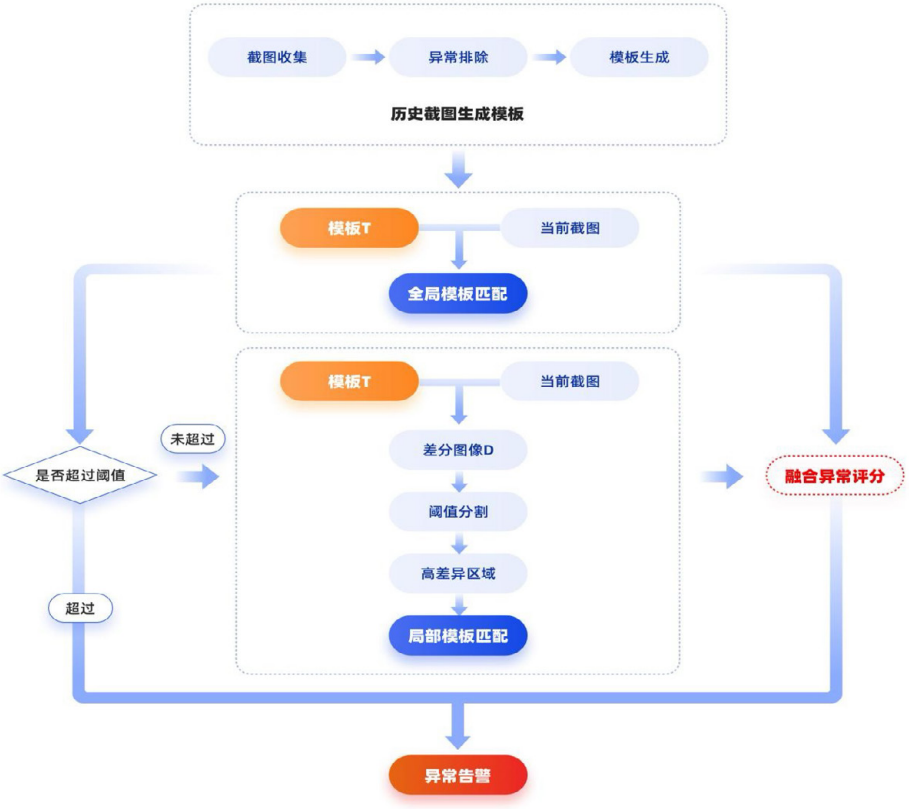


图 4 算法流程图

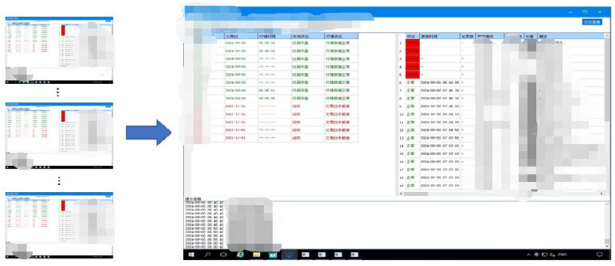


图 5 模板生成过程

3.2 异常检测

该平台不仅具备全局快速检测的功能，更注重细节处的程序异常，能够针对不同类型的异常进行精准检测。图 6 展示的是通过全局异常检测技术，系统能够快速识别出显著的程序异常并及时发出告警的常见异常情况。这种方法的优点在于它能够迅速响应明显的异常情况，确保问题能够尽快被处理。

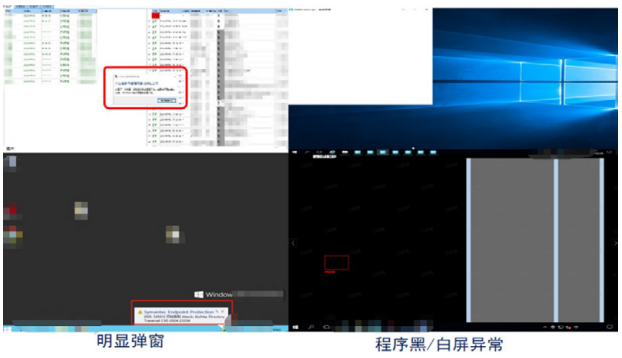


图 6 全局异常检测案例

而图 7 则呈现了局部细节检测的实例，显示了系统对一些较为细微的程序异常的有效识别。局部细节检测确保了在异常发生时不会遗漏任何细节，从而使得异常检测更加全面和精细。总体而言，这些算法设计展现了极大的灵活性，能够根据不同的异常类型采取适当的检测策略，既能够快速响应明显的异常，又能深入挖掘潜在的细节问题，从而提升系统的稳定性和可靠性。

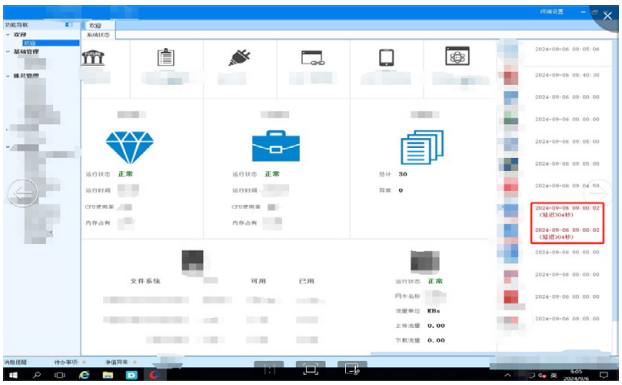


图 7 局部异常检测案例

四、总结与展望

运维管理方面，兴业证券沿着金融科技规划的蓝图方向，推进落实一系列创新举措，为业务高效稳定运行提供重要保障。本文针对运维巡检工作中的痛点难点问题，提出智能化、自动化和平台化的创新方案，有效解决行业共性问题。首先，确立了标准化的巡检操作规范，确保流程的一致性和准时性，显著提高了整体效率。同时，通过简化流程、减少非标准化操作环节，切实减轻了运维人员的工作负担，提高了整体工作效能。此外，兴业证券引入了智能巡检平台，支持集中化检查和多人同时访问，增强了巡检的全面性，并优化了资源分配。智能巡检模式的应用，打破了传统逐台巡检的方式，借助智能化手段提升了巡检的准确性与效率。这些创新措施共同推动了运维管理向高效、智能化的方向发展，为系统的稳定运行和业务的稳步增长提供了坚实基础。

展望未来工作，兴业证券将在智能化、自动化和平台化方向持续发力，充分发挥智能算法在运维管理领域的应用潜力，进一步提升智能监控和故障预测能力。同时，通过多个运维工具平台之间的数据和功能集成，将构建统一的运维管理视图，进一步提升运维效率。规范化与标准化的巡检流程推广及统一巡检平台的使用，将推动巡检工作更加自动化、智能化，减少人为错误，提高问题响应速度。这些策略的实施将推动运维管理迈向更高效、更智能的方向，确保企业信息系统的稳定运行，支撑业务的持续创新与发展。

参考文献：

[1] 张哲 . 数字化转型浪潮下基层央行 IT 运维建设 [J]. 金融科技时代 , 2021, 29(9):3.

[2] 耿辉 , 张乾尊 , 谢广斌 , 等 . 面向业务的 IT 基础运维能力建设实践 [J]. 中国金融电脑 , 2021(9):3.

[3] 王炜锋 .IT 运维价值服务转型实践 [J]. 金融科技时代 , 2023, 31(3):50-5.

03 实践探索

基于大商所 L1 行情数据：
HLS 与 RTL 混合设计在 FPGA 极速行情系统中的优化研究

刘垚，陈士阳，张旭东，张航，杨郭龙，万锟
中信建投证券股份有限公司 | E-mail：liuyaodcq@csc.com.cn

摘　要： 针对期权期货做市业务对低延迟行情的迫切需求，本文提出一种基于高层次综合（HLS）与寄存器传输级（RTL）混合设计的 FPGA 加速方案，并以大商所 Level 1 行情解析为案例展开验证。传统 RTL 开发周期长，而纯 HLS 方案难以满足性能极限需求。本研究通过任务解耦，将非关键路径（如 DMQP 解码）交由 HLS 实现以缩短开发周期，同时在延迟敏感模块（如 DMDP 的 Vint 解码与订单簿更新）采用 RTL 优化以保障吞吐量。实验表明，混合方案的行情解码延迟为 72 ns，吞吐量达 5555 万快照输出 / 秒，性能接近纯 RTL 实现，但开发周期缩短 50%。本研究为金融 FPGA 系统的敏捷开发提供了有参考价值的新思路。

关键字： 高层次综合、行情、订单簿、延迟、吞吐量

一、引言

期权期货做市业务是指做市商在期权期货市场中，通过不断提供买卖报价，为市场参与者提供流动性的一种业务模式。当市场价格发生波动时，做市商需要在亚微秒级调整其报价策略，否则可能会错失交易机会或面临不利的市场冲击。优化系统的延迟性能对于提高做市业务的执行效率和盈利能力至关重要。

FPGA（Field-Programmable Gate Array, 现场可编程门阵列）因其并行计算能力与可定制化硬件逻辑的特性，成为实现亚微秒级低延迟系统的首选方案。然而，传统 FPGA 开发依赖于寄存器传输级（Register Transfer Level, RTL）设计，需从最底层开始设计功能和时序细节，开发周期长达 3-12 个月。这一局限性在金融市场快速迭代的需求下尤为突出，导致 FPGA 技术难以广泛落地。

高层次综合（High-Level Synthesis, HLS）技术通过将 C/C++ 等高级语言自动转换为硬件描述代码，显著提升了开发效率。AMD（原 Xilinx）的实测数据显示，HLS 可将 FPGA 功能模块的开发周期缩短至传统 RTL 的 30%。但 HLS 生成的电路常因编译器优化保守性导致关键路径延迟增加，难以满足高频交易对极致性能的要求。这一矛盾使得 HLS 技术在高频场景中的应用受限。

针对上述问题，本研究提出一种 HLS 与 RTL 混合设计范式，以大商所（Dalian Commodity Exchange, DCE）Level 1 行情解析为验证场景，旨在兼顾开发效率与性能极限。通过任务解耦，将非关键路径交由 HLS 实现以加速开发，同时对延迟敏感模块采用 RTL 手动优化。实验表明，该混合开发方案在保障 72 ns 解码延迟的同时，将开发周期缩短 50%。

二、行情解析任务

大商所行情发送平台采用二进制协议，提供定时行情发送服务和历史行情查询服务。定时行情发送服务推送内容包括“L1 行情、L2 行情、市场状态、交易状态、盘后行情通知等”，使用 DMDP（DCE Market Data Dissemination Protocol）协议与用户交互。传输层采用 UDP 协议，支持双组播通道互备。历史行情查询服务提供查询的内容包括“行情快照查询、历史行情查询、合约基本信息查询”等，使用 DMQP（DCE Market Data Query Protocol）协议与用户交互。传输层使用 TCP 协议。

用户行情客户端同时接收 DMDP 和 DMQP 行情，由 DMQP 行情获取合约基本信息，由 DMDP 行情获取合约变动，用于更新订单簿，然后拼接成完整的快照输出。

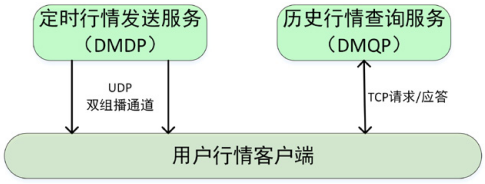


图 1 大商所行情服务

大商所行情服务使用了两种压缩技术发送行情：其一是变动与不变信息分离传输，即：当日基本信息开盘前一次性输出，交易时段仅发送行情变动；其二是使用 VINT 编码来压缩变动信息。大商所采用的这种行情发布方式既可降低交易时段网络传输带宽，又不会对客户端解码造成压力，对低延迟系统非常友好。上期所和广期所也采用了同原理的行情发布方式。

图 2 行情报文结构示意图

DMDP 和 DMQP 行情报文的组成结构相同，如图 2 所示。

客户端在做报文拆分时，需要考虑以下情形：1). 一个报文中包含 1 或多个消息；2）. 一个消息中包含 1 或多个域；3). 一个消息可以拆分到多个报文发送；4). 单个报文可以完整装下的消息，不会被拆分；5). 一个域不会拆分到不同报文中。

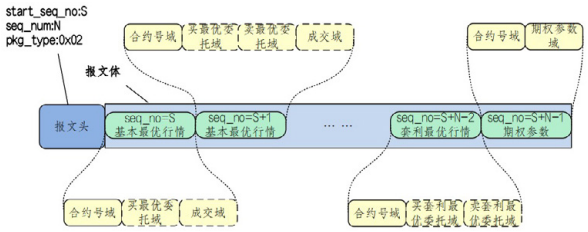


图 3 L1 定时行情报文示意图

本研究关注的 DMDP 行情是 L1 定时行情报文。L1 定时行情报文结构如图 3 所示。每个 L1 定时行情报文可能包含 N 个行情消息。报文中每个行情消息的定时行情编号是唯一的，可以通过报文头的 start_seq_no + seq_num（消息在报文中的位置）计算得出。客户端通过使用该编号实现双路互备和双路优选功能。每个行情消息包含多个域，其中【合约号域】是必填域，是每个行情消息的起点。【买最优委托域】、【卖最优委托域】和【成交域】等为选填域。定时行情发送时，只会发送在本切片内发生变化的域。

图 4 合约基本信息表

本研究关注的 DMQP 报文是合约基本信息查询应答报文，我们通过该报文获取合约号（contract_id）、合约编码基准价（codec_price）、最小变动价位（tick）、合约类型（contract_type）和初始持仓量（init_open_interest）等信息，并与合约号索引（contract_no）建立映射关系。开盘前客户端完成 DMQP 行情接收后，可以在本地构建如图 4 所示的 4 个映射表，按照合约号索引填入每个合约对应的信息。限于篇幅此处不再具体介绍 DMQP 报文格式，具体请参考官方文档《DCE 交易 7.0 组播行情发送平台开放协议规范说明》。当客户端接收到 DMDP 行情时，通过合约号索引作为地址来读取合约基本信息表，用于计算全量价格和生成完整的快照结构。

表 1 为本研究输出快照的目标数据结构及其字段说明。

表 1 输出快照的目标数据结构及字段说明

目标字段	提取来源	内容
contract_id	合约基本信息查询应答报文	合约号，Char[20]
contract_ no	L1 定时行情的合约号域	合约号索引，Int32
send_time	L1 定时行情的报文头	行情报文发送时间，Int64
seq_no	L1 定时行情的报文头	消息编号（start_seq_no+seq_num），Int32
last_price	L1 定时行情的成交域	最新成交价（codec_price+tick*变动），Int32
total_qty	L1 定时行情的成交域	总成交量，Int32
turnover	L1 定时行情的成交域	总成交额，Int64
open_interest	L1 定时行情的成交域	持仓量，Int32
bid_price	L1 定时行情的买最优委托域	最优买价（codec_price+tick*变动），Int32
bid_qty	L1 定时行情的买最优委托域	申买量，Int32
ask_price	L1 定时行情的卖最优委托域	最优卖价（codec_price+tick*变动），Int32
ask_qty	L1 定时行情的卖最优委托域	申卖量，Int32

三、整体结构设计

行情解析过程本质上是不断地拆分数据包，提取关注内容并暂存到订单簿中，待每接收一个完整合约的信息后，触发输出快照的过程。

本方案大商所 L1 行情解析的整体结构设计如图五所示。

由于 TCP 行情和双路 UDP 组播行情都是通过一个物理网口接入 FPGA 的，我们网络模块只实例化了 1 个物理底层（PMA+PCS+MAC），在 IP 层以上分别实例化 1 路 TCP 镜像模块和 2 路 UDP 镜像模块，分别接 1 路 DMQP 行情数据和 2 路 DMDP 行情数据。

DMQP 报文解码过程在当日开盘前进行，对延迟要求不高。DMQP 解码提取合约基本信息后直接保存在 OrderBookEntries 模块的合约基本信息表 (BasicOrderBook) 中对应的位置。DMQPHeaderDec 模块负责 DMQP 拆分报文并提取报文头信息，DMQPFieldDec 模块负责提取关注的基本信息字段。

DMDP 报文解码过程通过 DMDPHeaderDec 模块拆分报文并提取报文头信息，2WayOpt 模块利用报文头的定时行情编号实现双路互备和优选，然后由 DMDPFieldDec 模块提取关注的域字段，再由 VintProc 模块解码 Vint 数据并更新到 IncrOrderBook 中，同时通过判断完整接收一个合约的行情后，通知 FinishQ 模块。注意：IncrOrderBook 保存的是原始增量信息，而非使用合约基本信息计算后的结果。这样做的原因是：单拍内实现乘法计算时，主频难以提高，Decoder 的时序收敛困难。我们以 OrderBookEntries 模块内部的 SRAM 为界做跨时钟域处理，写入 SRAM 前的解码功能主频实现得更高些，在读取 SRAM 基本信息和增量信息之后，伴随快照输出同时进行乘法计算，主频与 PCIe 的 250 MHz 保持一致。这样结构更加合理，而且整体读 SRAM 次数减少一次，延迟也会少 2 拍。

FetchCalcOut 模块一旦检测到 FinishQ 非空，即启动对应合约的信息读取和计算，然后按固定格式计算并输出快照。快照结果通过 PCIe DMA 推送给主机 CPU Cache，供策略软件使用。

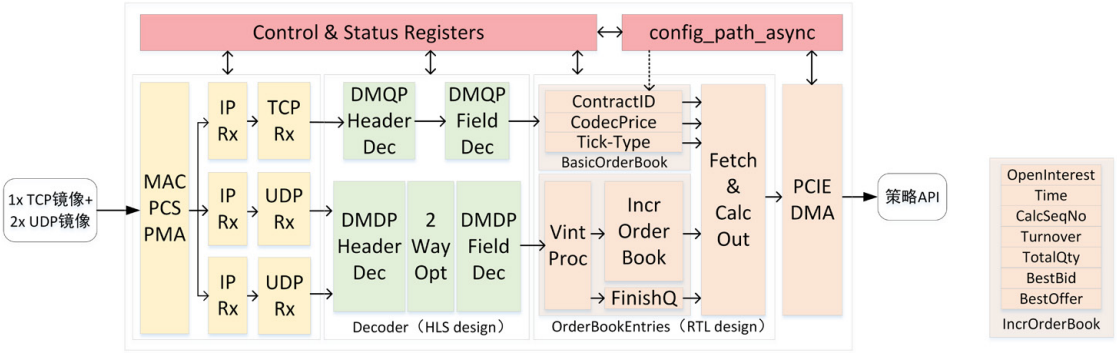


图 5 大商所 L1 行情解析的整体逻辑结构图

四、混合开发与实现

4.1 HLS 功能划分

FPGA 实现的第一个问题是决定哪部分功能由 HLS 实现。

本研究的探索目标是既不影响整体性能，又能缩短研发周期，降低设计难度。首先我们确定了不宜采用 HLS 设计的大类情况及理由如下：

1. **现有成熟 IP：**（如以太网和 PCIe DMA IP 核）都是 RTL 开发，仍保持现有设计；
2. **延迟关键路径：**性能优先，仍采用 RTL 设计；
3. **时序清晰和功能简单的需求：**RTL 设计同样快速，仍采用 RTL 设计。

适宜交给 HLS 软件实现的几个大类情况也相对清晰，情况及理由如下：

1. **非延迟关键路径的功能：**性能不需要优化到极致；
2. **时序情况较复杂的功能：**可能耗费大量调试时间；
3. **待前期探索的复杂 IP：**希望快速出结果。

针对大商所 L1 行情解码加速项目，由于 DMQP 解码仅需在盘前完成，对交易时段的性能没有影响，属于非延迟关键路径，可完全通过 HLS 方法实现。此外，在本文第 2 部分我们讲到，客户端在拆分 DMDP 时需要考虑多种情况，RTL 设计时还要考虑配合流水线节拍，极容易出现隐藏问题。由于行情数据包较大，一旦出现隐藏错误，板级调试中定位问题的等待时间较长。在高级编程语言环境可

以实现分钟级的完整数据 C 仿真，这样隐患问题在开发前期就可以解决，板级测试能够更快实现收敛。

OrderBookEntries 模块主要是由 SRAM 组成的订单簿存储空间，其内部的逻辑功能主要是配合订单簿实现快速读写 SRAM，功能清晰简单，可使用 RTL 代码实现。

4.2 HLS 设计思路与技巧

DMQP 和 DMDP 两种行情处理的第一步均需要先进行报文拆分处理。

如果用软件思维实现报文拆分处理，有一种既快速又可靠的实现方式，即：先申请一个缓存空间，将一段时间内所有输入报文连续拼接在一起放入缓存。处理时先确认缓冲中待处理的报文是一个完整的合约报文，再按照固定格式提取关注字段。这种方式虽然简化了问题，但是延迟较大。即使在非关键路径上我们也不推荐，因为这种设计在实际生成硬件时必然占用较大的 SRAM 资源，这会影响最终硬件整体可实现的主频。

在 HLS 开发中，需采用硬件导向的设计范式，通过约束驱动（如 #pragma HLS pipeline）确保流水线并行度，同时规避软件式缓存机制导致的资源浪费。

在整体设计上，为了保证流水线前后级相互独立，函数间适宜采用 FIFO（Stream）接口。FIFO 的实现方式可以根据实现的深度确定使用 BRAM、LUTRAM 或移位寄存器等资源。函数间尽量不涉及 FIFO 外的其他交互信号，这样可以减少函数间依赖关系。使用 #pragma hls dataflow 编译指令让函数间的流水线并行起来。我们建议增加使用 disable_start_propagation 选项，这个选项可以降低函数间启动信号的传递延迟。

在函数内部设计上，为了降低延迟，设计者应该编写尽可能精确的代码来增加对生成电路的确定性掌控。在此项目中，我们用到的精确编码方式包括：

1. 状态机：基于数据包的协议处理过程天然自带状态。我们采用了状态机的设计思路，用 switch-case 语句帮助软件推断出符合设计预期的状态机电路。其实，使用 if-else 语句结合计数器也可以实现数据包的处理，vitis 软件会自行判断并生成等价的电路。这样设计虽可减轻设计者的负担，但同时也减弱了设计者的掌控力。

2. 位操作：ap_uint 类型数据支持对数据的比特级操作，支持任意的位截取和位拼接操作。HLS 提供了与 RTL 编码同等的位操作能力。如果想要 FPGA 每拍吞吐一次数据，使用灵活的位操作至关重要。

3. 变量的复位控制：为保证整体功能稳定，一些寄存器电路在复位时必须能够恢复初值。精确控制可复位的变量，明确使用 static 关键字，同时将综合选项 rtl.reset 设置成 state。

4. 编译指令：函数内常用的编译指令包括：#pragma hls pipeline II = 1 和 #pragma hls latency max = 3 min = 1。Pipeline 指令可以让设计完全流水线化。Latency 指令可以相对精确地约束输出的延迟周期。

5. 调试信息：由于 HLS 输出的 RTL 代码的可读性较差，设计者应该考虑将可记录的状态和 Corner 情况全部引出到顶层，以方便调试。

4.3 RTL 设计

DMDP 解码数据进入 IncrOrderBook 存储前，需要先将 Vint 字段解码转换成整数。根据表 1 的快照结构，我们需要支持的 L1 定时行情报文的域包括：合约号域、成交域、买最优委托域和卖最优委托域。这些域的 Vint 数据分布如图 6 所示。

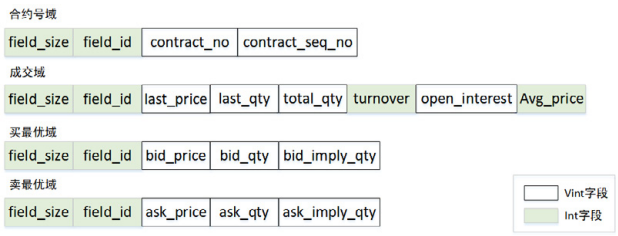


图 6 Vint 数据在不同域中的分布

Vint 编码为 Varint 编码和 ZigZag 编码的结合，指的是使用动态变化的字节数来表示整数，可将 64 位二进制编码的有符号整型编码用 1 到 10 字节表示。Varint 编码是一种将 64 位二进制编码的无符号整型根据其大小用不同长度字节进行编码的编码方式。其编码特点为数字越小所占用的字节数越少。ZigZag 编码是配合 Varint 来使用的一种为有符号整型数定义的编码。ZigZag 编码将有符号整型映射到无符号整型，而无符号整型也可以用此原则进行映射。 其算法为：

$$EncodeZigZag(n) = (n<<1)^(n>>63)$$

由于 Vint 编码的字段是变长的，且不同域的分布方式不一样，从逻辑实现上看，FPGA 流水线只有识别前一个 Vint 字段结束位，才能确定下一个字段的起始位。因此，每输入一个域，可能无法一拍内吐出所有关注字段。但是由于 Vint 编码主要由位操作组成，FPGA 实现每拍只吐出 1 个字段是比较容易的。HLS 和 RTL 设计方法都可以实现此目标。我们采用了 RTL 实现方式，因为后期若延迟优化进展到逼近极限时，要做到每输入一个域吐出所有关注字段，则 RTL 实现更加方便，且不需要改动现有结构。

RTL 设计部分还有一个会影响整体延迟的问题，即：触发快照输出的时机。为了降低延迟，我们的逻辑应最快确定一个合约信息的结束位置。我们的触发条件有两个：一是解析到有新的合约号域，且上一个合约信息有关注字段的变动，则触发上一个合约输出快照；二是解析到报文结束，且最后的合约有关注字段变动，则触发最后一个合约的快照输出。

五、结果分析

5.1 硬件资源

本研究选用的FPGA板卡为AMD的Alveo X3522PV。FPGA芯片内含1029600个CLB LUT、2112个BRAM、352个URAM和1320个DSP。板上资源包括2路DSFP28网络接口，8-lane PCI Express接口和8GB DDR4存储颗粒。本研究选用的测试服务器的型号是Dell R740，配备Xeon Gold 6226R处理器，主频2.9 GHz，睿频加速可达3.9 GHz，内含22 MB L3 Cache。

HLS开发环境选用的是Vitis 2024.1.1。Vitis软件将C++代码转换RTL后，以RTL源码的方式导入到Vivado 2024.1.1工程中。

FPGA整体工程的资源占用情况如表2所示。其中，逻辑资源占用非常低。BRAM资源占用相对高的原因是我们的订单簿容量最大支持16384只合约。我们预留出了比当前市场合约数目多出1倍的空间，以备未来扩展。

5.2 延迟

我们实现的以太网协议解析主频是312.5 MHz，L1行情解码逻辑的（包括Vint处理和Orderbook输入侧）主频是333 MHz，订单簿输出侧的计算输出和PCIe DMA主频是250 MHz。

在上述主频条件下，我们测出的关键路径上各模块的穿透延迟如图7所示。其中，PCIe DMA的延迟是64字节数据从PCIe输入到CPU L3 Cache读取的统计平均值，PCS/PMA的延迟是按照从发送端到接收端的穿透延迟的1/2计算得出。PCIe与网络功能属于技术底座，指标仅供参考。本研究主要关注行情数据处理的延迟，图7中展示了各模块有效数据的最大head-to-head延迟。从延迟指标看，行情输入到快照输出总延迟为72 ns，HLS设计部分延迟为24 ns，这已经接近RTL可实现的极限水平。

表2 FPGA资源占用情况

资源	LUT	LUTRAM	FF	BRAM	URAM	DSP
数量	41283	5181	49558	421.5	4	9
占比	4%	2%	2%	20%	1%	1%

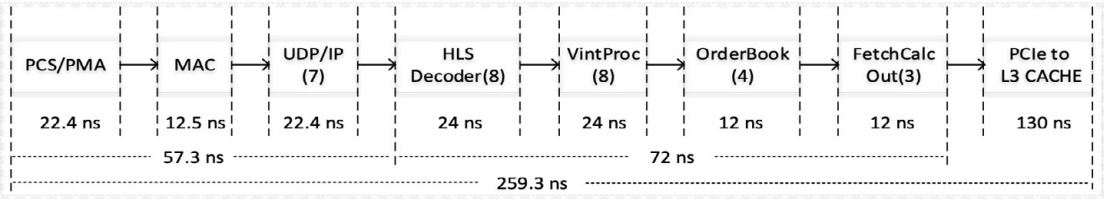


图7 关键路径上各模块的延迟性能

5.3 吞吐量

本方案中，网络处理从MAC输入到UDP输出的极限带宽为1.25GB/秒。网络输入带宽远小于此带宽，不会出现瓶颈。PCIe DMA的极限带宽为8GB/秒，快照输出有效数据72字节，使用256比特对齐传输后共占用96字节。由此算出，PCIe DMA的快照吞吐量为8333万输出/秒。

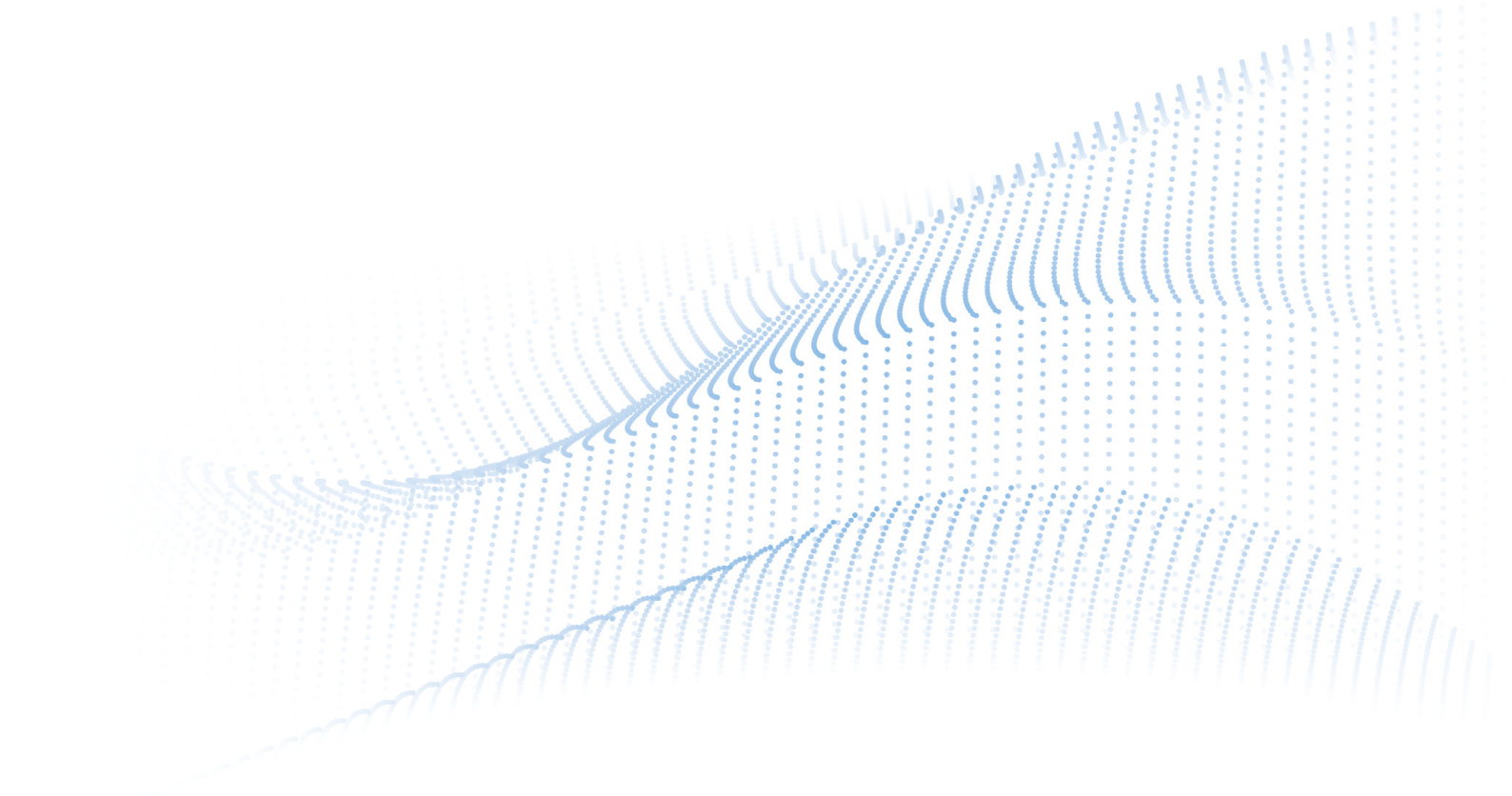
整个设计中，HLS解码和订单簿输出的设计是完全pipeline的，没有造成前级反压的情况。VintProc模块是唯一会反压的模块。VintProc模块处理最慢的域是成交域，状态机每吐出一个完整成交域需要等待6拍，由此算得的成交域的吞吐量是5555万输出/秒。完整成交域的吞吐量也即是整个设计快照的极限吞吐量。因此，本设计的极限吞吐量是每秒5555万快照输出。

5.4 开发周期

在我们的实践中，HLS 开发必须经历一定的学习和探索周期。在熟悉 HLS 开发流程并形成开发范式后，我们实际的 HLS 部分开发时间从编码到板级测试收敛大约为 2 周。根据以往 RTL 开发经验，这部分从编码到板级测试收敛耗时不少于 1 个月。如果遇到拆分情况特殊的报文，调试时间还会拉长。我们评估行情解码部分的开发周期大约可缩短 50%。开发周期缩短的主要贡献来源于以下两点：一是高级语言抽象层次较高，开发者能够从时序细节中解脱出来；二是 C 仿真运行速度快，能够让设计者在定位问题和迭代优化时，达到与软件开发相当的速度。

六、结束语

本文验证了 HLS 与 RTL 混合设计在 FPGA 极速行情系统中的可行性。实验表明，通过合理的任务解耦和混合开发，可在保障百纳秒级延迟的同时，将开发周期缩短至 RTL 方案的 50%。我们未来工作将聚焦于：1）探索更多金融加速场景，如 AI 模型和因子计算，丰富自有 IP 库；2）开发混合开发的自动化工具；3）进一步简化验证流程。



证券期货业开源软件治理的探索与实践

樊芳，沙明，李佶，房慧丽，俞小虎
上交所技术有限责任公司 | E-mail：ffan@sse.com.cn

摘　要： 当前，开源软件发展呈现出强劲的增长势头，市场规模持续扩大，技术创新不断涌现。随着金融科技的快速发展，开源软件在证券期货行业软件开发过程中的应用日益广泛，但同时也带来了安全、合规和风险管理等方面的挑战。因此，如何进行开源软件安全合规治理成为当务之急。本文调研分析开源软件治理的现状与痛点，结合上交所技术有限责任公司（以下简称技术公司）工作实际，积极探索证券期货行业开源治理实践路径，通过建立开源软件治理组织架构、发布治理制度规范、完善开源软件全生命周期治理机制、推进开源治理常态化长效化等措施，以保障开源软件的安全合规使用，进一步降低软件供应链安全风险，提升业务运行效率和开源软件安全管理水平。

关键字： 开源软件；软件开发；开源治理；软件供应链安全

一、背景

1.1 政策与监管

近年来，国家政策、地方政府、金融监管机构纷纷针对开源软件治理领域，制定并出台了一系列相关政策措施，推动开源软件产业的健康发展，促进数字化转型和创新发展，进一步保障信息安全和促进自主可控。

在国家政策层面，2021 年 3 月，《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》首次历史性地将开源纳入国民经济和社会发展的五年规划纲要之中，着重强调了对数字技术开源社区等创新联合体的支持，致力于构建更加完善的开源知识产权与法律框架，并大力倡导企业开放其软件源代码、硬件设计及应用服务，以此推动技术创新与共享。

在地方政府层面，为推动开源软件的发展和治理，已经出台了一系列相关措施，旨在通过实施多项工程来推动开源软件技术的发展和应用。例如，2023 年 5 月，由北京市经信局联合北京市科委中关村管委会、北京市发改委共同启动并发布《北京市通用人工智能产业创新伙伴计划》，该计划强调开源模式是培育软件开发新范式的重要环节，提出以开源聚合创新，构建大模型开源社区，吸引科研院所、代码托管平台、开发者及团队形成开放、包容、活跃的创新氛围。

在金融监管要求方面，2021 年 9 月 28 日，人民银行办公厅、中央网信办秘书局、工业和信息化部办公厅、银保监会办公厅、证监会办公厅联合发布《关于规范金融业开源技术应用与发展的意见》，提出了金融机构在应用开源技术时应遵循“安全可控、合规使用、问题导向、开放创新”四项基本原则，旨在规范金融机构合理、合规、安全地应用开源技术，提高应用水平和自主可控能力，促进开源技术健康可持续发展。

1.2 内部规划

为深入贯彻落实《上海证券交易所“十四五”科技战略规划》关于“全面统筹协调推进交易所开源软件治理体系建设”的要求，技术公司结合国家标准 GB/T 43698-2024《网络安全技术 软件供应链安全要求》、金融行业标准（JRT 0289-2024《开源技术规范》、JRT 0290-2024《金融业开源软件应用 管理指南》、JRT 0291-2024《金融业开源软件应用 评估规范》）及《证券期货业开源技术应用与风险管理指南》等课题研究成果，致力于构建一个安全、高效、合规的开源软件治理体系。

二、开源软件治理的现状与痛点

2.1 开源治理现状

在全球开源发展浪潮中，中国开源继续展现出强劲的发展势头和独特的创新活力。据中国开源软件推进联盟 2024 中国开源发展现状统计显示，中国开源最大的活力开发者增长速度令人瞩目，中国开发者（含香港）在 GitHub 全球用户数量规模排名第三，2023 年为 1100 万，预计 2026 年 GitHub 中国开发者规模 1800 万至 2000 万，活跃率居全球第一。2023 年 GitHub 日志数据为 14 亿，相比 2022 年增长了约 10.32%，继续凸显出开源科技的发展在全球数字化转型中的活跃与关键地位。同时，开源在

重点领域的技术与应用，也正逐渐强化和深化，已成为中国发展操作系统产业、开源数据库、大模型、汽车软件生态的重要途径，调研显示中国操作系统发行版数量显著增长，云和服务器操作系统 openEuler、OpenAnolis 和 OpenCloudOS 作为核心基础设施，持续领跑市场。

随着开源项目或技术的广泛应用，其面临的安全、供应链、合规等风险也日益加剧。据新思（Synopsys）公司《2024 年开源安全和风险分析报告》统计，在 2023 年审计的 1067 个流行代码库中，96% 的代码库包含开源代码，84% 的包含至少一个漏洞，74% 的包含高风险漏洞，对系统的稳定性和安全性构成了严重威胁。然而，当前企业开源治理能力相对薄弱，亟需战略性规划和制度指导，建立完善的开源治理体系成为当务之急，以提高企业科技创新和自主可控能力。

2.2 开源治理痛点

随着开源在各行各业的广泛应用，开源软件的安全问题日益凸显，开源软件治理的痛点和难点主要体现在以下几个方面：

一是安全性问题。开源软件安全漏洞的持续增长、漏洞层级传播和修复滞后等问题，可能导致数据泄露、系统瘫痪等严重后果；利用开源组件的已公开漏洞、后门、木马等发起软件供应链攻击。

二是合规性风险。开源许可证复杂多样，各有要求，容易引起知识产权纠纷，可能会带来法律风险，如知识产权侵权、合同违约等。开源软件可能涉及用户隐私数据的收集、处理和使用。

三是运维和管理难度大。开源软件通常依赖于其他开源库和组件，这增加了管理的复杂性。版本控制问题也是一个挑战，因为不同版本的开源软件可能具有不同的功能和安全性，亟需建立有效的依赖管理和版本控制策略，以确保所使用的开源软件的稳定性和兼容性。

三、开源软件治理体系建设

3.1 差距分析

为全面掌握开源软件治理工作现状、识别差距并制定提升路径，技术公司以《证券期货业开源技术应用与风险管理指南》为基准，对标行业开源治理能力成熟度先进级标准，于 2023 年启动开展开源软件治理能力预评估工作，旨在系统性完善开源治理体系。

技术公司严格对标《证券期货业开源风险管理能力成熟度模型》（以下简称《模型》）先进级 54 项评估指标（基础级 10 项、增强级 26 项、先进级 18 项），全面开展现状评估与差距分析。评估结果显示，开源治理工作存在以下主要问题：组织机制待完善，治理职责与分工需进一步明确；制度体系待健全，管理制度覆盖不足，流程规范性有待提升；流程管控待强化，开源软件引入评估机制缺失，第三方软件管理不足；基础设施待建设，开源组件仓库尚未建立，影响统一管控能力。针对上述问题，技术公司制定系统性改进方案，重点推进以下工作：成立专项治理组织，明确职责分工；修订完善开源治理制度体系；建立开源软件引入评估与审批机制；构建开源组件仓库，形成统一台账；完善全生命周期管理流程。通过以上措施，全面提升开源软件治理能力，实现开源治理的规范化、标准化管理。

3.2 组织架构

为确保开源软件治理工作有效落地，技术公司成立开源软件治理专项工作小组，统筹推进治理任务实施。工作小组由安全团队、技术治理团队、开发团队、测试团队、运维团队及法务团队组成，职责分工如下：安全团队主要负责制定开源软件治理制度及流程规范，开展安全漏洞及许可证风险检测，建立并维护开源软件管理台账；技术治理团队主要负责搭建并维护统一开源软件私服仓库，将开源软件管控要求嵌入软件研发电子化流程；开发团队主要负责落实开源软件引入、使用、持续管理及退出全生命周期要求，参与开源软件治理评审，确保合规使用；测试团队主要负责开源软件功能及兼容性测试，验证安全补丁有效性；运维团队主要负责监控开源软件运行状态，配合漏洞修复及版本升级；法务团队主要负责审核开源许可证合规性，提供法律风险防控支持。

3.3 制度规范

为规范开源软件全生命周期管理，技术公司发布《开源软件管理制度》，明确以下管理要求。管理架构方面，清晰界定决策层级、牵头部门及各团队职责分工，建立跨部门协作机制，确保治理要求有效落地。开源软件治理全生命周期管理方面，制定覆盖引入、使用、持续评估、退出各阶段的详细策略，建立标准化流程机制，实现闭环管理。第三方软件管理方面，要求供应商提供完整物料清单（SBOM），包含开源组件清单、已知漏洞信息、许可证合规证明等信息，严格执行安全检测要求，确保第三方软件合规上线。开源软件引入评估机制方面，建立开源软件引入评估模型（详见表 1），设置强制评估项（标 * 项），严控引入风险。

表 1 开源软件引入评估模型表

评估维度	一级指标
技术适用性	功能满足度*
	二次开发支持度
	性能满足度
安全合规性	已知安全漏洞情况*
	最新版本漏洞修复情况
	漏洞修复效率
	漏洞修复说明
	许可证合规性
产品生命力	产品成熟度
	版本更新
	问题解答率
	行业认可度
信创适配性	适配信创服务器
	适配信创操作系统
	信创替代产品
服务支持度	技术支持人*
	文档完整性*
	商业支持

3.4 技术支撑

为有效落实开源软件全生命周期管理要求，技术公司构建了三位一体的技术支撑体系，确保从引入、使用、持续评估到退出的闭环管控。

一是建立统一的开源软件私服仓库。基于现有软件制品托管平台，构建企业级开源组件私服仓库，实现开源组件集中存储、版本控制和统一分发；自动化对接全生命周期管理流程，支持组件引入审核、安全检测、使用追踪、版本查询、组件退出等核心功能。二是建立开源软件生命周期治理流程机制，深度整合现有软件研发效能平台，在设计、开发、测试、发布各阶段嵌入开源组件使用审批卡点、安全合规自动检查、许可证风险预警，实现研发流程与开源治理的有效衔接。三是完善自动化安全检测体系。部署 SCA（软件成分分析）工具，对开源软件资产进行定期扫描及发布前的扫描，自动化开展开源组件安全风险动态评估，识别组件依赖关系，实时更新开源组件资产清单，动态维护开源组件使用台账。

3.5 开源软件全生命周期治理路径

基于已建立的治理组织架构和技术支撑体系，技术公司将开源软件治理要求深度嵌入软件研发全流程。通过标准化研发流程管控，实现从架构设计、代码集成、功能测试到部署交付各阶段的开源组件闭环管理，其全生命周期治理流程如图 1 所示。

在引入阶段，技术公司通过落实开源软件需求分析与规划、风险评估与评审、引入管控等机制，确保开源软件引入安全合规。首先，根据技术系统实际需求，研判是否需要引入开源软件，开展候选组件技术调研与需求匹配度分析。其次，根据开源软件引入评估模型，对开源软件进行引入自评估，包括技术适用性、安全合规性、产品生命力、信创适配性、服务支持度等方面的评估，利用软件成分分析工具进行自动化的安全风险评估。最后，根据“谁使用谁引入”的原则，由使用方在软件研发效能平台发起引入申请流程，申请内容覆盖开源软件引入评估模型相关指标，并由开源软件治理专项工作小组开展评审，评审通过后的开源软件自动进阶到开源软件私服仓库，并对引入过程的相关材料记录存档。通过标准化评审流程与自动化工具赋

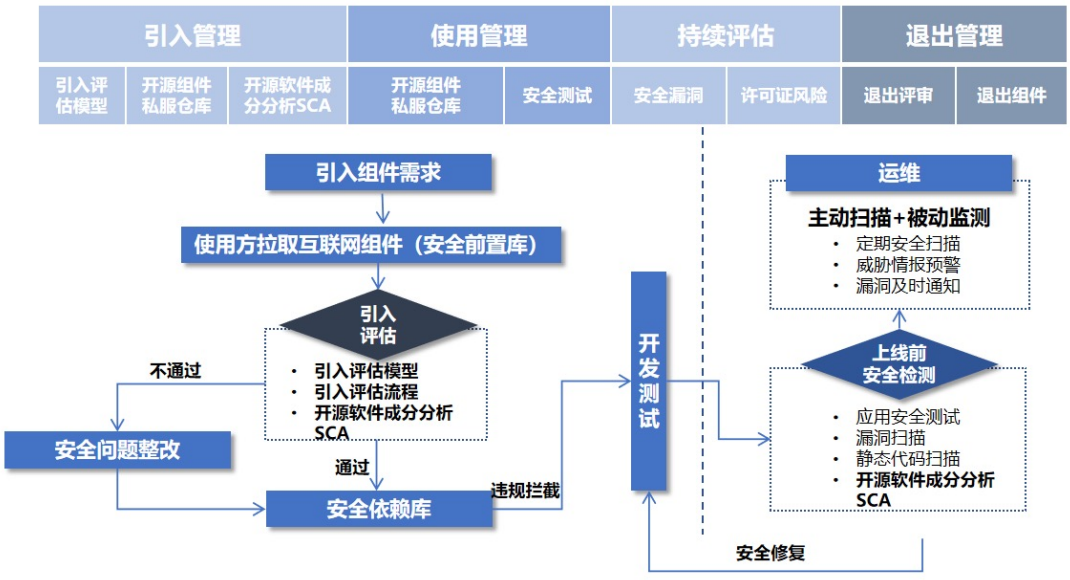


图 1 开源软件全生命周期治理路径概览图

能，确保所有引入组件经过技术验证、安全检测及合规审查，从源头降低软件供应链风险。

在使用阶段，重点通过以下措施确保开源组件使用安全可控。一是加强组件使用强制管控。分阶段推进技术系统 100% 采用私服仓库组件进行研发，严禁直接引用未经评审的开源组件。在编译环节设置强制校验卡点，若检测到未通过评审的组件，立即触发编译阻断机制，禁止未经评审的开源组件版本流入测试阶段。二是常态化开展上线前的开源组件安全检测。利用软件成分分析工具自动识别组件依赖关系、已知安全漏洞、许可证冲突风险，执行组件漏洞修复跟踪机制。通过技术卡点与自动化工具联动，从流程机制上杜绝未经评审组件的使用，形成“检测 - 修复 - 验证”闭环，显著降低开源组件使用风险。

在持续评估阶段，一是对自研技术系统使用开源组件的情况定期进行全面风险排查与安全性检测，自动化形成自研技术系统开源软件使用台账。二是及时接收外部威胁情报预警，开展开源组件风险隐患排查和处置。

在退出阶段，针对有停止服务、闭源及上级监管单位要求整治的组件，需从源头上加强开源组件风险管控，设置开源软件私服仓库黑名单，并对相关组件发起退出申请和评审，评审通过后可自动化删除相关组件。

四、开源软件治理的应用与实践

4.1 自研或合作研发技术系统的开源软件安全管控

技术公司重点通过环境隔离、流程管控、工具赋能三重机制，构建软件研发级开源组件安全防线。一是构建研发环境安全基线。自研或合作研发技术系统的开发测试环境位于公司内部独立网络，并对开发环境使用的研发集成工具进行安全管控，保障研发技术系统的环境和工具安全、合规、可控。二是加强开源软件全生命周期流程风险管控。准入管控方面，开源组件入库执行“申请 - 评审 - 检测 - 归档”四步流程，特殊场景下允许启用灰名单机制，如：允许临时使用存在已知风险的组件，要求开发者签署《网络安全风险接受单》并明确加固措施和修复期限。编译管控方面，研发编译服务器强制对接组件私服仓库，自动拦截未入库组件，触发“组件黑名单阻断”安全告警。上线前的安全检测方面，集成 SCA 工具进行软件成分分析和漏洞扫描，自动化识别分析组件依赖关系、许可证冲突风险、漏洞等级、漏洞利用难度等信息，当发现可被利用风险漏洞时或接收到外部威胁情报预警时，可快速、准确定位受影响组件清单、系统级影响面。

4.2 第三方软件的开源软件安全管控

通过执行“合同约束、技术验证、闭环管理”三重机制，构建第三方软件的开源风险防控体系。合同约束性管理方面，在项目合同协议中明确供应商的开源治理责任，要求供应商在软件交付及版本升级时提供 SBOM（软件物料清单）、安全合规评测报告，确保其使用的开源软件符合公司内部安全检测上线要求。交付物技术验证方面，公司内部安全团队对交付物的正确性、完备性、可靠性进行复核验证，要求供应商提供检测包，使用漏洞扫描工具和软件成分分析工具对第三方软件进行安全检测和分析，建立第三方软件的开源软件管理台账，加强第三方软件的安全风险管控。风险闭环处置方面，如发现第三方软件存在需要整改的风险隐患，协调供应商提供修复补丁并完成验证，同步启动备品替换预案（如存在供应链断供风险）。

五、总结与展望

未来，随着开源软件在软件开发中的广泛应用，涵盖金融服务、金融科技、大数据、AI、健康科技和生命科学等多个领域，同时开源软件的数量庞大且增长迅速，开源软件安全态势呈现出日趋严峻的趋势，如漏洞修复时间长、利用开源软件漏洞发起的供应链攻击频繁等，开源软件的安全治理刻不容缓。技术公司将紧密结合自身工作实际持续优化开源软件治理体系，以组织架构作支撑、以制度规范作引领，完善开源软件全生命周期治理路径机制，推进开源软件治理常态化长效化。

开源软件治理是软件供应链安全管理的关键环节，对于采购软硬件产品、集成建设、云服务及数据服务的运营单位而言，须与供应链的上下游协同实施安全管控措施。这些单位应积极探索建立跨行业、跨领域的开源软件治理合作机制，以共同应对开源软件所带来的挑战与风险。同时，运营单位需提升对开源软件治理的认知与重视，加强开源软件治理领域的人才培养，提升从业人员的专业技能与素质水平，也是确保开源软件治理有效性的重要措施。

分布式数字身份技术在证券行业的应用研究

夏鼎[1]，黄韦[1]，黎峰[2]，徐鑫[3]，吴鑫涛[1]、周玉勰[1]

[1] 国泰海通证券股份有限公司 | [3] 证通股份有限公司 | [4] 上海立信会计金融学院

E-mail：xiading@gtht.com

摘　要： 本文探讨了分布式数字身份（DID）技术在证券行业数字化转型中的战略价值。针对行业长期存在的数据孤岛、流程冗余、监管审计难等痛点，DID 技术通过重构数字信任体系，实现用户数据主权自主掌控、跨机构高效互信及隐私合规的有机统一。文章结合国泰海通证券的探索案例，揭示了 DID 和区块链技术如何通过统一身份标识与可验证凭证，破解证券服务割裂化、流程冗余化、数据流通低效化等难题，实现跨机构身份互认，打造证券行业数字身份信任基石，助力证券行业优化业务流程、创新业务模式，促进行业的数字化转型和高质量发展。

关键字： 分布式数字身份（DID）；可验证凭证（VC）；区块链技术

一、引言

在国家战略驱动下，分布式数字身份（DID）技术正成为我国数字经济基础设施建设的关键支柱。2024 年，国家发展改革委等三部门联合印发《国家数据基础设施建设指引》，重点提及构建“全国一体化分布式数字身份体系”，规范身份标识生成、认证等机制，为数据要素流通奠定信任基础。这一技术通过重构数字信任范式，推动三大核心转变：从机构集中管理到用户自主掌控数据主权，从单一中心化信任到多元节点协同验证，从封闭生态到开放互联的跨域协作网络。用户以加密方式持有身份凭证，仅在必要时授权验证，既落实了《个人信息保护法》查阅复制权（可携带权），又通过跨平台互信机制减少重复数据采集，在强化隐私保护的同时提升协作效能，为数据要素流通构建可信、开放的数字化底座。

在证券行业数字化进程中，传统身份认证体系存在结构性矛盾：其一，行业服务呈现“竖井式”割裂状态，用户数据分散于各机构系统中，缺乏跨机构互信互认机制，制约市场活力；其二，用户需在券商、银行等多机构重复进行身份验证，反复证明“我是我”，导致业务办理低效，既增加机构运营成本，又降低了用户体验；其三，跨机构用户数据流转路径不透明，缺乏溯源通道，使监管审计面临严峻压力。这些矛盾凸显了构建统一、可信数字身份基础设施的迫切性——通过打破数据孤岛与信任壁垒、提升数据流通效能、强化风险管控与合规能力，DID 技术为证券行业提供了破解信任断层、激活数据价值、筑牢合规根基的系统性解决方案。

国际层面，欧盟 EBSI 项目已验证 DID 在跨境协作中的价值；国内实践中，区块链服务网络（BSN）联合 CTID 数字身份链推出的实名 DID 服务，在跨境金融场景中实现“匿名注册与合规 KYC 并行”，既满足隐私保护要求，又提升数据流通效率。在证券行业，DID 可通过分布式身份标识与跨机构互信机制，解决数据确权、流转溯源与监管穿透难题，为行业业务场景提供可信基础设施，推动行业向高效、安全、合规的数字化生态转型。

本文以国泰海通证券行业分布式数字身份系统建设为探索案例，系统梳理了 DID 技术在证券行业的落地路径。基于 W3C-DID 标准与联盟链架构，本研究构建了包含身份注册、凭证核验、跨机构授权等模块的技术原型，并通过可行性分析验证技术方案在典型证券业务场景中的实施潜力。本研究突破传统中心化身份管理体系局限，提出可迁移的分布式数字身份应用框架与分布式数据可信交换协议，验证了 DID 技术在金融基础设施中实现数据确权闭环、监管穿透强化及信任机制重构的可行性，为行业数字化转型提供理论支撑与实践样板。

二、分布式数字身份技术原理

分布式数字身份体系以密码学和分布式账本技术为核心，重构了传统身份管理范式。其核心组件包括：

1）分布式标识符（Decentralized Identifier, DID）：“数字身份证”，由自己掌控，无需依赖中心机构管理。

2）可验证凭证（Verifiable Credential, VC）：权威机构签发的电子证明，可随时验证真伪。

3）可验证表述（Verifiable Presentation, VP）：用户自主组合自己的电子证明，并按需选择性展示给他人验证。

三者共同构建了自主可控、跨链互信的身份信任网络，为数据主权与隐私保护提供了技术基础，其关系如图 1 所示。

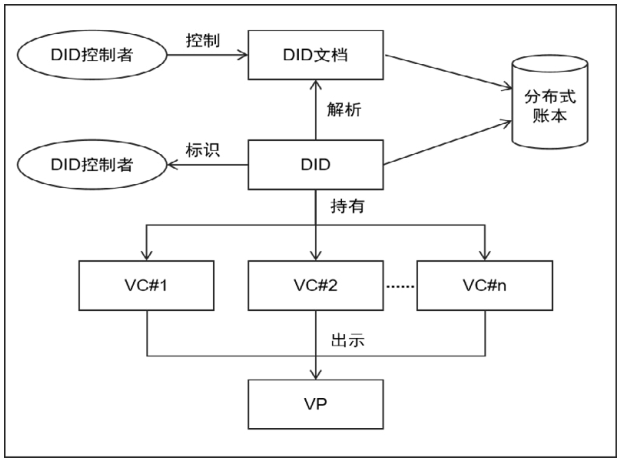


图 1 DID 基本组件关系

2.1DID：分布式数字身份应用的基石

分布式数字身份标识符（DID）是分布式数字身份的基础设施，通过分布式账本技术实现身份标识的自主控制与跨域互认。DID 一般采用国际 W3C-DID 标准，采用三段式语法结构，如 did：< 方法标识符 >：< 唯一标识符 >，来表示实体的唯一身份。

DID 的生命周期管理包括创建、解析、更新和撤销四个环节。创建时，用户通过智能合约注册 DID 并生成文档上链存证；解析时，验证者查询 DID 文档以确认身份；更新时，用户修改 DID 文档并重新上链存证；撤销时，用户或监管者通过智能合约标记 DID 失效，避免身份滥用。

DID 文档是 DID 的核心技术载体，采用标准化 JSON-LD 格式存储身份信息，包含验证方法（定义公钥及加密算法并关联控制者 DID）、控制者（明确 DID 的授权管理者身份）和证明（通过数字签名确保数据完整性，记录签名时间、公钥及签名值）等核心要素，确保身份标识的可信性与可控性。

2.2 VC: 可信数据流通的标准化容器

可验证凭证（VC）是权威机构签发的数字化“电子证明”，用于声明实体属性（如投资者资质、信用等级），其核心设计包含密码学验证、标准化数据结构及隐私保护三大特性。凭证内容通过结构化模板定义，明确主体属性、颁发者与持有者的 DID 标识符，并嵌入数字签名保障内容不可篡改，同时设置有效期限以防止长期滥用。

VC 的签发与验证流程如下：签发时，签发者根据实体属性生成 VC，使用私钥签名并上链存证，实体通过 DID 授权接收后存储至本地；验证时，解析并查询签发者 DID 文档获取公钥，验证签名有效性、凭证状态及有效期。

2.3 VP：隐私保护下的数据交互协议

可验证表述（VP）是用户向验证者提交凭证时采用的隐私保护协议，支持聚合多个 VC 并通过持有者签名确保数据可信。其核心包含持有者 DID 标识符、验证者生成的随机数、待展示的 VC 列表，以及持有者的数字签名，同时设置有效期限以限制VP的使用范围，防范长期滥用风险。

VP 的签发与验证流程如下：签发时，验证者生成随机数发送给持有者，持有者选取所需 VC 构造 VP，用 DID 私钥签名后提交给验证者；验证时，验证者解析并查询 DID 文档验证签名，检查随机数防止重放攻击，同时追溯 VP 中各 VC 的颁发者签名与状态以确保凭证真实有效。

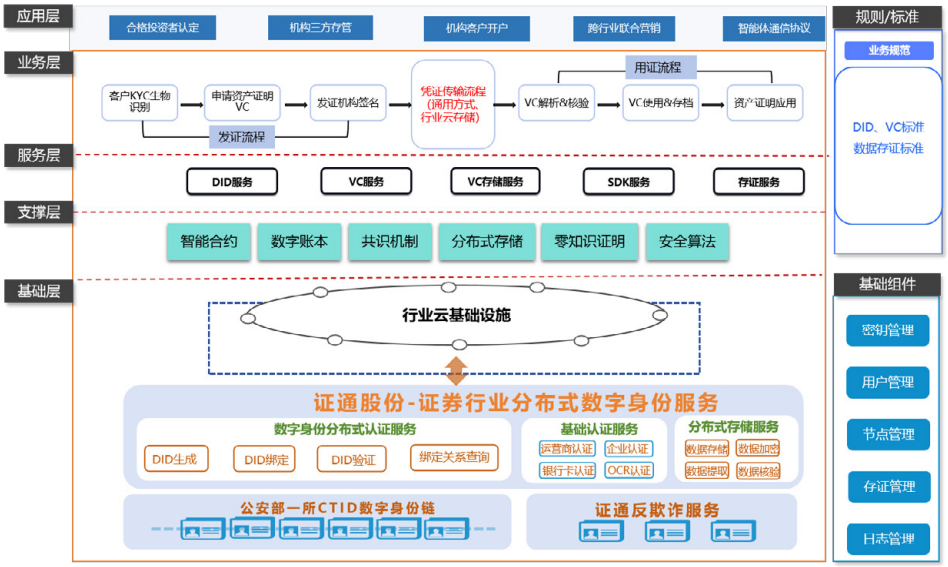


图 2 证券行业分布式数字身份应用架构

三、证券行业分布式数字身份架构设计

本文将以国泰海通证券行业分布式数字身份应用的建设情况为背景，从整体架构设计以及可验证凭证（VC）的生命周期管理两个方面进行详细介绍。

3.1 整体架构概述

证券行业分布式数字身份系统采用分层模型设计，由用户层、应用层、服务层、支撑层和基础层五个层次组成，各层协同运作，确保系统高效运行。

应用层：面向持证方、发证方、用证方和监管方，覆盖实际使用场景，包括合格投资者认定、机构三方存管、机构客户开户等应用场景。

业务层：基于 DID 和 VC 技术，实现发证与用证流程，满足身份认证和数据共享需求。

服务层：提供 DID 管理、VC 生成解析、存储服务及 SDK 支持，保障上层应用灵活调用底层能力。

支撑层：以区块链为核心，提供智能合约、共识机制和分布式存储等功能，支撑上层服务。

基础层：依托分布式身份联盟链，实现 DID 注册解析、VC 签发验证，并支持密钥管理、节点管理和存证服务。同时，通过对接证通股份反欺诈服务及公安部一所 CTID 数字身份链，确保数字身份与实体身份一致性；中心化存储模块在用户授权后代管凭证，提升用户体验。

3.2 可验证凭证（VC）生命周期管理

本架构通过分布式数字身份与区块链技术，构建以用户为中心的可验证凭证（VC）全生命周期管理体系，涵盖创建、存储、传输、使用及撤销等关键环节。以下是 VC 生命周期的具体阶段：

3.2.1 VC 创建与签发

权威机构作为发证方，基于实体属性（如投资者资质、企业信用等级）生成 VC，并使用机构 DID 私钥对 VC 进行签名，确保内容不可篡改。同时，VC 的哈希摘要记录于行业联盟链中，为后续验证提供可信依据。

3.2.2 VC 存储与存证

用户支持本地设备或云端存储 VC 文件，为增强安全性，经用户授权后，VC 的哈希摘要被记录于区块链，确保数据存储行为可追溯且不可篡改。

3.2.3 VC 传输与授权管理

用户主导 VC 的传输过程，通过分布式数据托管传输机制实现安全流转。如图 3 所示，具体包括：

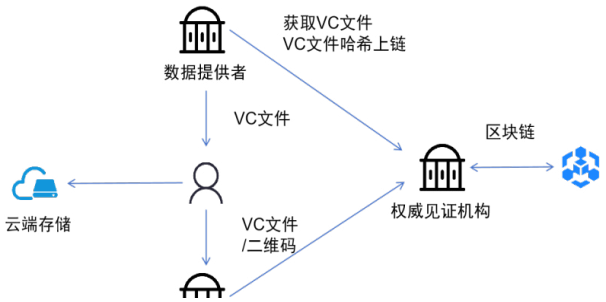


图 3 分布式数据可信托管传输协议

用户主动将封装好的 VC 文件传输至目标接收方，接收方收到文件后可通过区块链核验哈希摘要及 VC 内容，确认数据来源合规且未被篡改。用户也可选择通过二维码分享云端文件地址，由接收方自行下载并核验，进一步简化传输流程。

3.2.4 VC 使用与验证

接收方在获取 VC 后，通过解析 VC 查询颁发者的 DID 文档，提取公钥验证签名有效性，同时检查 VC 的过期时间和状态。

3.2.5 VC 撤销与失效管理

当 VC 因有效期届满或特定原因需失效时，监管方或发证方可通过智能合约标记 VC 为“已撤销”状态，同时通过区块链广播通知该 VC 失效，并将撤销记录上链存证。用户及接收方可通过区块链实时查询 VC 状态，避免无效凭证的误用。

四、分布式数字身份在证券行业的应用场景

4.1 合格投资者认定

在证券合格投资者认定中，客户需满足适当性管理要求并提供资产证明等材料，但因数据分散于不同机构，传统流程存在跨机构互认效率低、隐私风险高、券商运营成本大的问题。引入分布式数字身份（DID）技术后，投资者通过注册 DID 并签发可验证资产证明凭证（VC），帮助个人实现数字身份自主管理、个人数据自主授权；帮助企业更加合法合规的使用个人数据，促进多方数据资产流通。可信凭证由客户自主携带并传输至第三方，打破数据由数据源直接提给数据使用方的传统思路，解决数据泄露、数据滥用、数据授权等长期存在的痛点问题。该技术基于跨机构互信机制将认证时间从数天压缩至分钟级，显著降低运营成本，同时遵循《个人信息保护法》的数据最小化披露原则及个人数据携带权，实现效率与安全的双重提升。

4.2 机构三方存管无纸化

在机构客户开通三方存管协议场景中，机构客户需完成券商与银行间的三方存管协议签署。传统流程机构客户需在券商营业网点签署三方存管协议后携带至银行网点再次签署，存在协作复杂、运营成本高、客户体验差等问题。通过分布式数字身份技术，机构客户可基于 DID 进行协议签署，进而实现三方存管协议的无纸化。银行与券商通过区块链跨链验证 DID 及签署文件的真实性，简化签约流程，从而降低协作成本、提升签约效率，同时保障客户隐私安全。

4.3 机构客户开户

在证券行业，机构客户完成母公司开户后，若需在子公司（如期货公司）开立账户，通常需重复提交法人资料、营业执照等核心材料，导致流程冗余、数据重复验证、隐私泄漏风险上升。通过分布式数字身份技术，子公司可向机构签发授权的 VC，将该凭证提交母公司验证，母公司通过 DID 核验 VC 真实性后，直接将已验证材料传输至子公司，全程无需重复提交材料。这一方式以数字信任替代传统纸质流程，既降低跨机构开户的重复操作成本，又提升了机构客户的业务体验，推动行业从“材料驱动”向“信任驱动”的服务模式升级。

4.4 跨行业联合营销

在跨机构联合营销场景中，客户在合作机构办理业务时可享受特定权益（如优惠或优先服务），但传统模式需双方交换客户名单进行核验，存在隐私泄露和合规风险。利用分布式数字身份技术，客户可向机构申请身份证明 VC 文件。合作机构可通过验证客户的 VC 快速确认客户身份并授予相应权益。同时，基于相同的身份系统，还可以开展数字藏品等联合营销活动。分布式数字身份技术简化了跨行业身份核验流程，有效保护客户隐私，同时提升业务效率与用户体验，为跨机构联合营销提供安全、高效的数字化支持。

4.5 智能体通信协议

在未来证券行业 AI 智能体跨机构通信场景中，智能体需调用外部数据源与服务以支持业务决策（如查询跨境交易数据或敏感客户信息），但现有 MCP 协议依赖中心化 OAuth 授权服务器，存在单点故障风险及权限管理灵活性不足的问题。基于分布式数字身份技术，可以为 MCP 客户端与服务器注册唯一身份标识，并关联可验证凭证，实现分布式的身份管理和动态权限控制。该方案有效解决了中心化授权的局限性，支持监管机构即时冻结高风险 AI 智能体的权限，保障系统安全运行，满足数据隐私不出域的合规要求，为反洗钱协作和跨市场数据共享等复杂场景提供高效、可信的技术支撑。

五、总结与展望

数字化转型的核心在于构建可信的数据要素流通体系，而分布式数字身份（DID）技术通过重构数字信任范式，有效解决了证券行业传统身份认证体系中的数据孤岛、流程冗余及监管穿透等痛点，为行业建立了自主可控、跨机构互信的数字化底座。基于可验证凭证（VC）与可验证表述（VP），DID 实现用户数据主权自主掌控，显著降低跨机构协作成本，并推动行业从“材料驱动”转向“信任驱动”模式。区块链技术支撑的分布式架构通过智能合约，既保障数据确权与流转透明性，又为监管穿透提供技术保障，实现效率与安全的有机统一。

尽管分布式数字身份技术展现出巨大潜力，但在实际应用中仍面临诸多挑战：

1）技术标准化：目前 DID 方法尚未形成统一标准，导致不同系统间的互操作性不足。未来需推动国际与国内标准组织合作，制定兼容性强的技术规范。

2）生态建设与协作：分布式数字身份的广泛应用依赖于多方协作，包括权威机构（如公安部）、金融机构（如券商、银行）以及监管机构的共同参与。当前，行业生态建设尚处于初期阶段，需进一步加强联盟链部署、数据共享协议设计等方面的协同努力。

分布式数字身份不仅是技术层面的革新，更是金融行业迈向智能化、可信化的重要基石。未来，我们期待 DID 技术能够在证券行业中更广泛地赋能金融创新，为投资者提供更加便捷、安全的服务体验，同时助力行业构建开放、高效、透明的数字化生态。

参考文献：

[1] Decentralized Identifiers (DIDs) v1.0. https://www.w3.org/TR/did-core/ 2022. 7.

[2] Verifiable Credentials Data Model v1.1. https://www.w3.org/TR/vc-data-model/ 2022. 3.

[3] 全国金融标准化技术委员会 . 区域性股权市场分布式数字身份技术规范 : JRT 0325—2024 [S]. 北京 : 中国金融出版社 , 2024.

[4] 王妮娜 , 杨帆 , 桑杰 , 许雪姣 , . 国内外分布式数字身份建设研究 [J]. 信息安全研究 , 2023, 9(10): 993-.

[5] 焦志伟 , 吴正豪 , 徐亦佳 , 魏凡星 . 基于隐私保护的分布式数字身份认证技术研究及实践探索 [J]. 信息通信技术与政策 , 2024, 50(1): 59-66.

运维数据湖平台在数智化实践中的探索与落地

毛梦非，王东，姜婷婷，王厦，刘博，刘志，刘青竹
国泰海通证券股份有限公司 | E-mail：xiading@gtht.com

摘　要： 国泰海通证券依据《金融科技发展规划（2022—2025 年）》，构建运维数据湖，旨在提升数据中心运维的智能化、自动化水平，强化数据治理，实现数据的高效利用。

文中涵盖统一集中管理、高效数据处理、数据治理能力及智能算法服务。湖仓一体架构和数据编织与数据网格技术的引入，增强了数据的整合与自治性，提升数据管理的灵活性与扩展性。实践探索场景包括指标异常检测、指标趋势分析、告警智能标签等。

运维数据湖的未来发展具备在智能化、全面化及创新化方向上的潜力和推动证券行业数智化转型的核心价值。通过本文研究，国泰海通证券不仅提升了运维管理效能，更为金融行业数智化转型提供了实践范例与理论支撑。

关键字： 运维数据湖 ; 运维数据治理 ; 湖仓一体 ; 数据网格

一、引言

随着金融科技的飞速发展，金融机构的数据中心运维已成为其提供差异化服务能力和提升客户体验的关键支撑。金融科技正逐步成为推动金融领域深刻变革的重要引擎。在这一背景下，中国人民银行印发《金融科技发展规划（2022—2025 年）》，旨在加速金融机构的数字化转型，并强化金融科技审慎监管。规划中不仅强调“数据”在金融科技中的核心地位，还设定了具体的发展目标，以期实现整体水平与核心竞争力的跨越式提升。

作为数据中心数智化转型的数据支撑，运维数据湖打破传统数据仓库的局限性，提高了数据分析的效率和准确性，使得跨领域、跨平台、跨环境的数据分析变得简单可行。采用湖仓一体的架构更有效地统一和整合日益增多的数据资源，将运维场景中多源异构的数据集中纳管，实现运维数据治理，通过数据模型的自动化映射和基于业务需求的数据查询与服务发布，为下游场景提供高质量数据服务，实现数智化实践探索和落地，为运维管理提供有力支持。

二、运维数据中台发展现状

2.1 企业数据中台的现状与挑战

在当前大数据环境下，企业面临着多源异构数据的爆炸式增长，导致一系列数据管理难题，如数据沼泽和信息孤岛。传统的架构不仅储存了大量的冗余和陈旧数据，而且严重影响了数据资产的有效利用。为了克服这些限制，多数企业通常采取了一种双层架构策略，即数据湖、数据仓库和关系型数据库等多种数据架构并存。这样做的目的是充分利用数据湖和数据仓库各自的优点，数据湖在存储多样性和数据归集方面表现出色，而数据仓库则在事务管理和数据分析方面有优势，典型架构如图 1 所示：

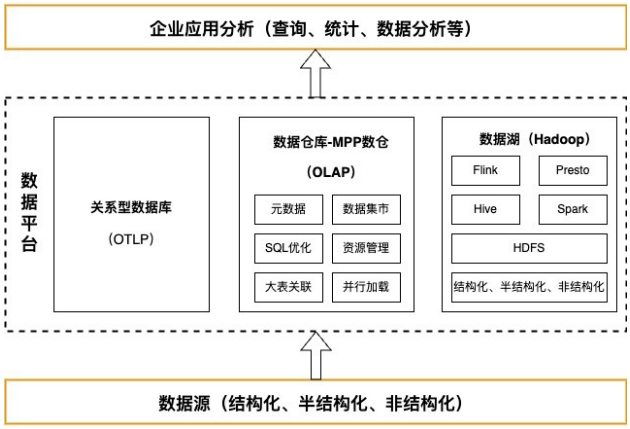


图 1

湖仓一体架构不仅解决了数据孤岛问题，还通过实时数据流和批处理相结合，大大提高了数据处理效率。与传统的数据湖和数据仓库相比，湖仓一体化具有更强的灵活性和扩展性，能够更好地满足现代企业对数据实时性、一致性和可用性的需求。这一转变标志着湖仓一体化从一个理论概念发展为一个实用的解决方案，已经在多个行业和应用场景中得到了广泛的应用和验证。

2.2 目标能力

整体目标是通过平台建设为国泰海通数据中心运维体系增强如下能力：

统一集中管理能力。实现多源、多域、异构、云上云下、多环境、多样化运维数据的统一采集和管理，努力做到应收尽收，运维数据全方位覆盖的集中、纳管和汇聚，形成体系化的运维数据管理体系和对象模型构建。

高效的数据处理能力。实现基于流处理引擎和批量任务一体的大规模数据处理架构，覆盖运维数据进行实时清洗、解析、处理、分析的复杂性和时效性要求以及满足海量、多并发的特殊需求，包括实时性要求不高的入库后异步 T+1 模式的统计分析。

运维数据治理能力。参照行业团标的运维数据管理规范对数据进行分层分域分类建模管理、数据质量管理、数据生命周期管理及数据安全管理，实现运维数据管理的一体化、规范化和标准化，为上层智能应用场景提供数据支撑。

智能算法服务能力。能够根据不同场景、针对不同的业务逻辑需求提供从理论到工程化的算法建模，并对算法模型进行调优、调参、发布和管理；算法模型的实践落地，不仅适用于交易和市场分析，更能广泛应用于日常运维管理，为运维人员提供分析和决策能力，提升运维效率和降本增效。

三、运维数据湖建设

3.1 平台架构设计

3.1.1 逻辑架构图

建设国泰海通运维数据湖和上层数据集市，梳理运维数据治理规范，支撑云上云下、流程操作、性能日志等各类数据接入；建设数据集市，实现数据服务标准、敏捷，打通数据生产和数据消费，为信息化运行场景的智能化演进提供数据和算力支撑。

原始数据：通过消息队列（kafka）、API 接口、数据库 JDBC、文件的方式融合多源异构数据，具备增量接入与全量接入能力；

核心数据：通过数据的清洗和转换，按照设定的数据模型生成标准数据，并加载到数据仓库中；平台基于湖仓一体的框架，数据湖作为一个汇聚多样原始数据的底层存储，数据仓库层则负责数据的精细加工、转换和聚合，以支持更高级别的查询和业务分析。

标准数据：可通过规则计算或智能算法服务进行进一步的数据加工，形成数据集市。

数据集市：通过统一的消费接口为消费场景提供数据服务。

3.1.2 技术架构图

技术架构中各个分层详情如图 3 所示：

(1) 数据接入

采集层基于国泰海通自研采集模块和开源采集器结合的模式，收集分散在服务器上的日志、指标数据，同时支持通过 Kafka 实现集中式的数据采集，便于兼容现有的监控数据，增加了数据源的多样性和保障数据的时效性。

(2) 数据计算

计算层使用华为 MRS 大数据计算平台，在设计上采用了 Lambda 架构和 Kappa 架构，以满足不同的数据处理需求。这种双架构的设计不仅提供了极大的灵活性，也确保了平台在不同应用场景下都能提供高效、准确的数据处理能力。

(3) 数据开发

支持可视化数据开发和基于 DAG 的作业编排，以缩短数据开发的投产周期和降低开发门槛。提供开箱即用的算子用于数据清洗、处理和计算，同时支持在线调试和查看算子处理结果。也支持自定义 Flink 和 Flink SQL 作业。

图 2 运维湖平台逻辑架构图

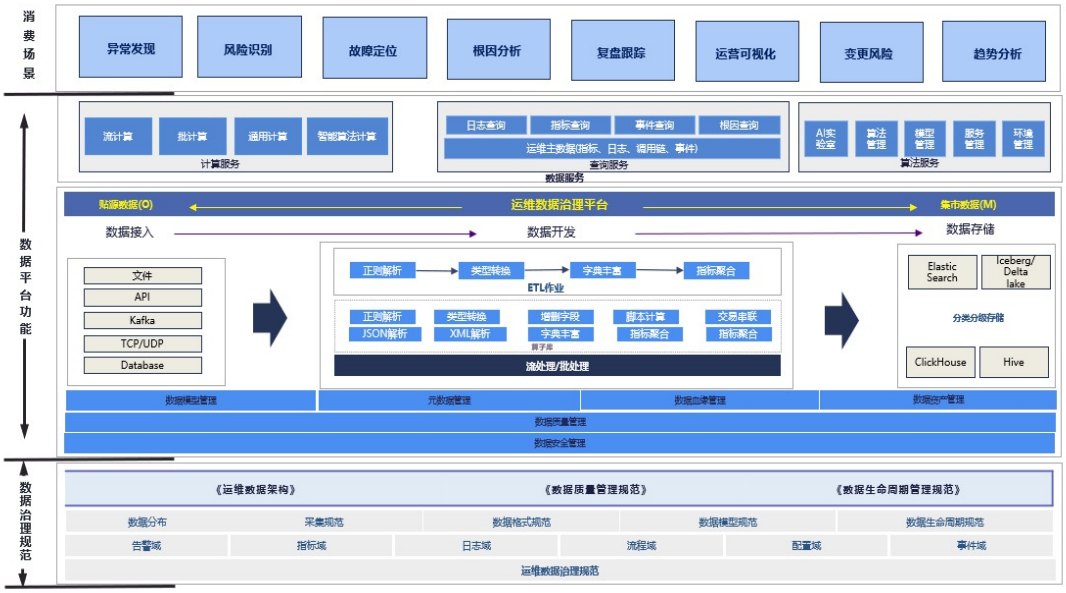


图 3 运维湖平台技术架构图

(4) 数据存储

存储层针对运维数据的特性采用了多种数据库（Clickhouse、Elasticsearch、HDFS 等）设计，以满足不同类型和规模数据的存储需求。提供更高的存储和计算效率，还能实现资源的统一管理和优化，从而降低总体运营成本。

(5) 数据服务

在数据服务方面，平台采用了统一服务网关的技术方案。这一方案能够统一管理所有的服务请求和数据流，实现服务的高可用和负载均衡。同时，统一服务网关还能提供安全控制和流量监控等高级功能，确保数据在传输和使用过程中的安全性和可靠性。

3.2 统一数据接入

基于湖仓一体架构进行大数据平台建设之后，国泰海通证券对其运维监控体系的存储进行了全面升级。这不仅优化了数据存储和查询性能，还大大提高了数据管理的灵活性和扩展性。全面接入了业务、应用、数据库、中间件、系统、硬件和网络等维度监控（如表 1 所示），同时也接入了大量的非结构化日志数据。

目前日志接入业务系统数量超 200 套，包括应用日志、系统日志、中间件日志和数据库日志，覆盖 Linux、Windows、Aix 和信创环境的 Kylin、欧拉等。支持冷热历史日志数据的实时查询与关键字的统计分析。日增日志条数达 45 亿条，日增存储量为 3T。热数据周期为 60 天。在热数据的存储周期之后（1 年周期），冷数据查询频次下降，通过 Airflow 工作流程被迁移到 HDFS 中作为存档数据存储，降低存储成本。

3.3 高效数据处理

基于湖仓一体架构的大数据平台，能够实现对运维数据的全面接入、标准化处理与汇聚，以及提供高效的数据

表 1 指标接入统计表

数据类型	指标数量	日接入条数	日存储量
主机基础数据	158 个	2.6 亿条/天	27G/天
中间件数据	540 个	6100 万条/天	5.8G/天
数据库数据	138 个	2000 万条/天	5.6G/天
硬件数据	240 个	1.2 亿条/天	17G/天
网络数据	35 个	3000 万条/天	4.1G/天
业务性能数据	610 个	6200 万条/天	6.4G/天

处理服务、数据计算服务和数据存储服务。以下是针对海量运维数据进行高效数据处理的详细策略。

首先，明确数据处理的功能定位，确保平台能够针对明确的数据处理规则需求进行自动化加工汇聚，并向上层场景提供所需数据。对于数据处理规则未定或随业务需求变化的数据，平台提供灵活的数据处理服务，支持上层场景根据业务需求进行数据处理。

其次，高效的数据处理架构构建实现运维数据的集中管控、分级处理和高效计算。通过实时数据流处理引擎，满足运维数据处理实时、海量、多并发的特殊需求。针对运维数据流的状态转化和存储模式进行专项优化，确保数据处理的速率和效率。

在数据处理过程中，引入任务调度引擎，实现对总量指标和日志数据任务的多模式优雅分级调度。通过同步双写和异步入库统计模式，确保业务优先级和实时性要求并存的任务得到高效处理，同时保证资源的合理分配和高效执行。

最后，为了满足智能化的运维场景分析需求，构建可灵活编排的批量算法计算框架至关重要。这一框架能够将不同场景的算法分析能力整合，解决复杂分析场景难以关联计算的难题。并通过图形编排方式，支持在线快速配置实现数据开发需求，包括实时数据聚合、关联以及任务资源配额的设置等。

数据开发模式支持可视化拖拽式与 SQL 式两种模式，同时兼容 Jar 包运行模式，可视化拖拽的模式也叫作管线作业模式，管线作业同时支持流处理和批处理，通过将算子连接成管线（Pipeline）可描述一个数据处理任务。数据处理任务由大量丰富开箱即用的算子组成，处理过程涵盖数据输入、数据处理、数据计算、数据输出等环节，可自定义编排组合，如下图 5 所示：

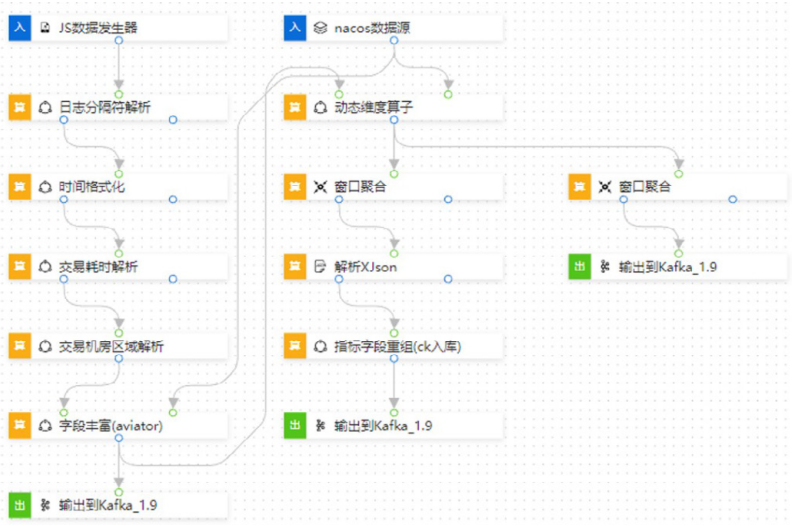


图 4 运维湖平台数据处理任务图

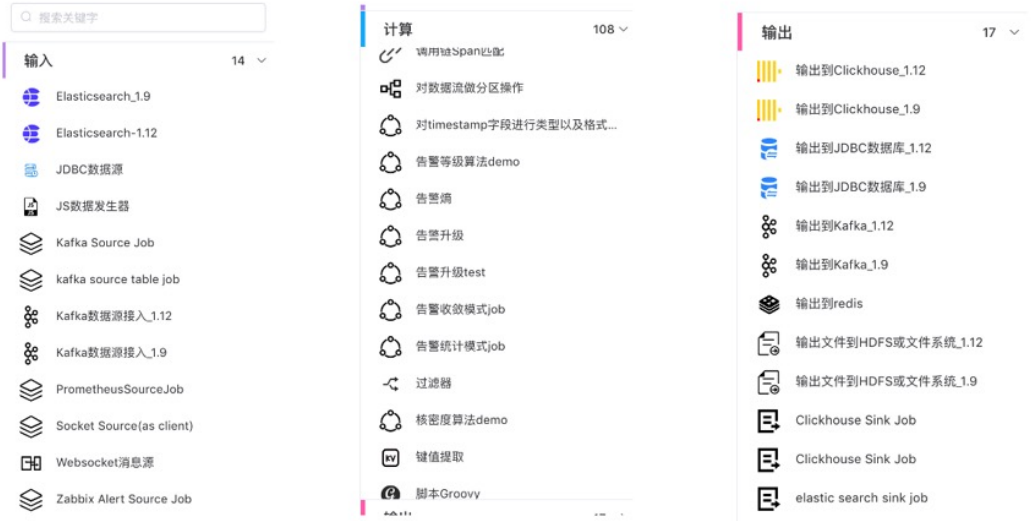


图 5 运维湖平台数据处理任务算子列表

3.4 运维管理体系

3.4.1 对象模型管理

基础设施指标对应 CMDB 运维数据模型对象的分类，数据分为设备层、操作系统、网络管理、网络层、系统层、中间件、数据库、物理服务器、业务应用层。具体列表如下:

3.4.2 指标体系构建

平台建设中，通过对象建模、维度建模，结合 CMDB 数据完成指标体系构建。纳管指标对象实例分层分类分域，结合运维数据治理理念进行合理存放，统计各个分层的指标数量、实例数量、对象接入数量，以及各系统的设备数及其他各类指标数，以便从宏观层面观察指标的动态变化，支撑后续指标消费应用场景。

表 2 运维湖平台模型统计表

对象分类名称	中文名	英文名	对象模型表名称	对象类型标识
国泰海通-数据库	国泰海通-达梦	DAMENG	db_dameng	DAMENG
国泰海通-数据库	国泰海通-DB2	DB2	db_db2	DB2
国泰海通-数据库	国泰海通-MONGODB	MONGODB	db_mongodb	MONGODB
国泰海通-数据库	国泰海通-MYSQL	MYSQL	db_mysql	MYSQL
国泰海通-中间件	国泰海通-Elasticsearch	Elasticsearch	mid_elasticsearch	Elasticsearch
国泰海通-中间件	国泰海通-Hbase	ht-Hbase	mid_hbase	Hbase
国泰海通-中间件	国泰海通-Hive	Hive	mid_hive	Hive
国泰海通-中间件	国泰海通-Kafka	Kafka	mid_kafka	Kafka
国泰海通-中间件	国泰海通-Nginx	Nginx	mid_nginx	Nginx
国泰海通-中间件	国泰海通-RabbitMq	RabbitMq	mid_rabbitmq	RabbitMq
国泰海通-中间件	国泰海通-Zookeeper	Zookeeper	mid_zookeeper	Zookeeper
国泰海通-中间件	国泰海通-宝兰德 BES	宝兰德 BES	mid_bes	bes
国泰海通-操作系统	国泰海通-Linux	Linux	os_linux	Linux
国泰海通-操作系统	国泰海通-Windows	Windows	os_windows	Windows
国泰海通-操作系统	国泰海通-AIX	AIX	os_aix	AIX
国泰海通-物理服务器	国泰海通-ARM 服务器	ARM server	phy_arm	arm
国泰海通-物理服务器	国泰海通-x86 服务器	x86 server	phy_x86	x86
国泰海通-物理服务器	国泰海通-网络设备	network_machine	phy_network_machine	network_machine
国泰海通-业务应用层	国泰海通-应用	sys_info	sys_info	sysInfo
国泰海通-业务应用层	国泰海通-业务	sys_info2	sys_business	sysBusiness
……	……	……	……	……

图 6 运维湖平台指标体系模型图

图 7 运维湖平台系统指标统计图

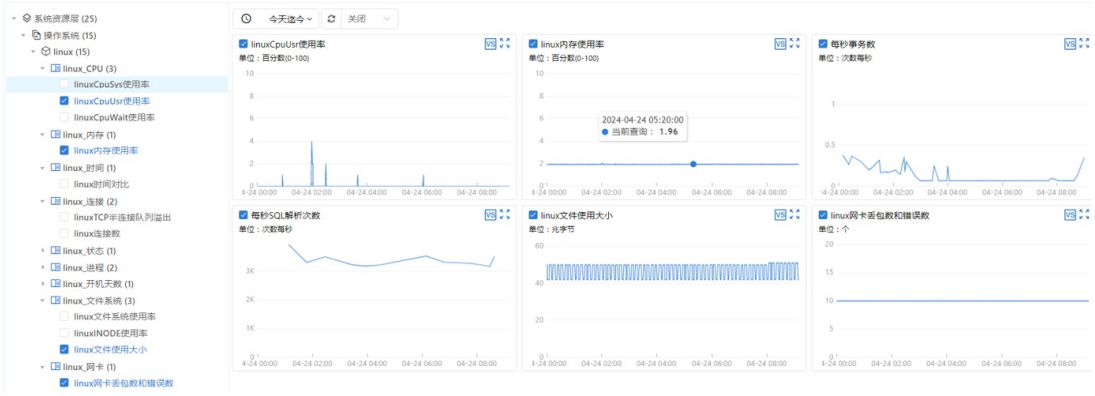


图 8 运维湖平台指标实例图

指标管理根据指标类型分层，以指标维度查询实例指标，以设备维度查询指标，展示设备信息，多指标平铺展示，单指标查询支持多实例比对、展示昨日、上周、上一交易日的指标曲线以便纵向比对。

体系建设实现效果可针对不同维度指标快速查询，不同层级、不同对象的初步统计和报表分析。

四、基于湖仓一体的场景化运维

在湖仓一体架构落地之后，运维数据服务能力得到了显著提升，新架构不仅解决了数据存储和查询的多样性和复杂性问题，还极大地提升了数据管理和应用的灵活性和高效性。这为多种运维场景提供了强有力的数据支持。

4.1 基于运维数据湖的指标异常检测场景

4.1.1 场景描述

为了弥补静态阈值无法考虑数据波动性或季节性变化而导致在正常波动范围内的误报或漏报，指标异常检测场景引入自适应性更强的动态阈值检测技术。根据运维数据湖中采集指标的周期性变化和递增递减趋势，实时识别出该指标的异常变化，在生产问题发生之前，先于业务人员感知异常。自动感知和自适应节假日等非交易时间，采用无监督或者半监督学习的方式自动校正基线范围，并且以特定的规则进行告警。

4.1.2 建设成效

完成运行实时检测任务创建共计 169 个，支持交易量、成功率、延时指标、系统连接数、应用日志条数等 200+ 指标的实时监测，涉及业务和机器性能指标分类共计十余种，AI 算法生成异常检测基带，与昨日、上周、上交易日指标线同框展示（图 9）。同时，触发异常合并告警后对接告警平台进行后续通知和跟踪。

以 Linux 连接数指标效果统计进行分析，数据特征受交易日历影响，非交易日的数据量明显下降，与总体指标情况一致，呈现一定趋势，如（图 10）所示：

数据分析结果为 linux 连接数的上述实例每日存在轻微变化，并且该指标在 2024-05-27 检测出偏离基带（如上图黄点），没有存在连续偏离的情况，不触发告警。

运维湖指标体系接入异常检测指标数量 200+，实例数 20 万 +，实时指标数据流检测耗时延迟为 10ms 内，经

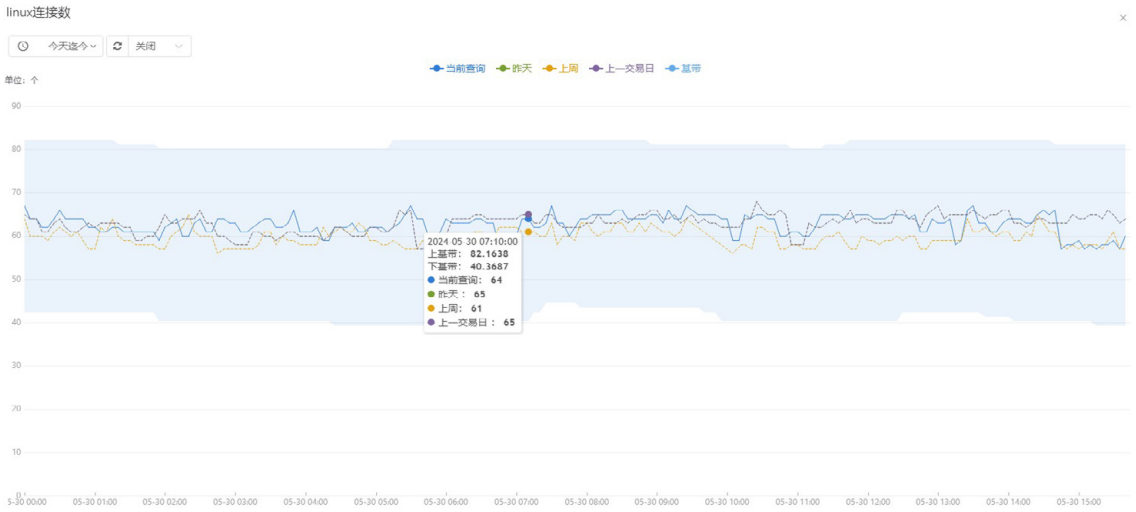


图 9 linux 连接数指标同框展示曲线图

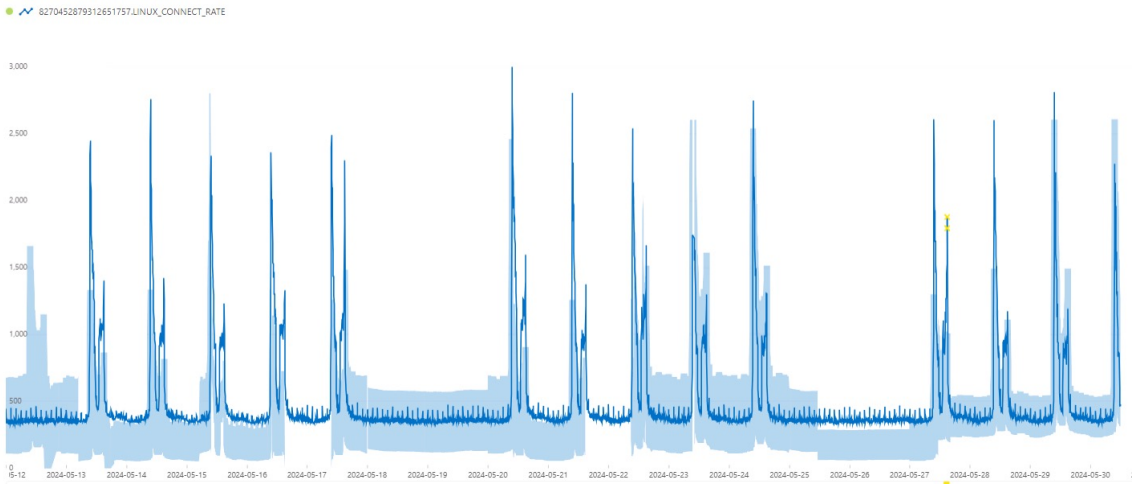


图 10 Linux 连接数指标曲线基带检测图

过调参去除毛刺及完善日常波动区间，提取连续周期和多样波动性特征等操作之后，数据异常检测准确度为 95% 以上。

4.2 基于运维数据湖的指标趋势分析场景

4.2.1 场景描述

为解决实际运维生产中表空间、文件系统使用率无法预测资源瓶颈到达的未来时间点以及指标存在持续缓慢异常增长未及时发现的痛点，文中引入单指标预测类算法，针对天周期、周周期或其他的季节性趋势较强，有历史趋势变化，且对于数据中蕴含非线性增长趋势有自然极限或饱和状态的资源类指标进行预测，并在故障产生前提前发出告警通知。

算法原理

将时间序列方法与机器学习结合的预测算法，它综合了时序数据通常具备的特性：周期性、趋势性、季节性，以及在生产生活中常见的节假日效应，并在趋势项中考虑突发情况存在的变点，以此结合机器学习方法进行分段线性拟合或分段逻辑回归，从而较好地适应数据中易出现波动变化的趋势项，最终做出对未来的评估。

4.2.2 建设成效

文中通过趋势特征提取的 AI 算法，抓取缓慢劣变递增趋势的指标实例，如下所列：

√ windows 内存使用率
√ windows 虚拟内存使用率
√ tomcat jvm 使用率
√ linux 连接数
√ 系统磁盘使用率
√ 数据库表空间

图 11 指标实例趋势劣变统计图

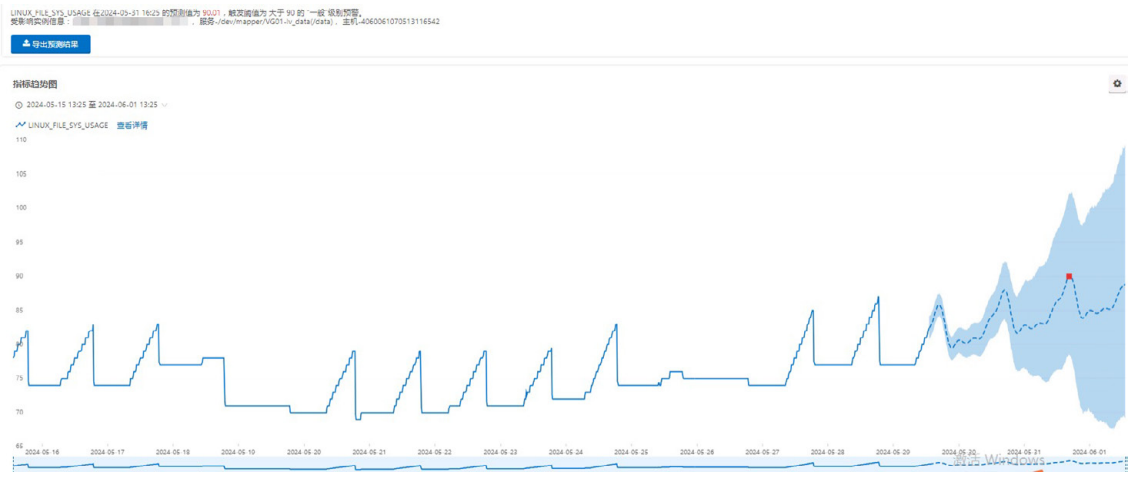


图 12 文件系统使用率实例趋势预测图

文中涉及的文件系统使用率指标纳管共计 100469 个实例，每日定时任务运行进行实时拟合，性能瓶颈触及前预警并对接告警中心进行通知。如下图（图 12）所示，在 05 月 31 日 16:25 磁盘使用率达到 90%，触发告警。

4.3 基于运维数据湖的告警智能标签场景

4.3.1 场景描述

由于 IT 业务繁多和监控工具告警策略复杂，国泰海通生产环境运维日增严重告警数量 2000+，需要通过规则结合 AI 智能算法进行告警上下游影响判断、影响范围和区分告警的有效性，更需要机制基于告警产生规律进行轻重缓急分类处理，对于周期性或者频繁出现的噪声告警识别过滤，达到快速从海量告警中甄别真正有价值信息的目标，提高有限人力资源投入的效率最大化。

4.3.2 建设成效

通过 AI 算法对历史数据进行总结归纳，对当前告警进行标签化。让业务人员对不同标签告警进行不同的关注，主要包括：

√ 对近期内从未出现过告警标记为新奇告警，提示业务人员关注；

√ 对短时间内出现大量告警标记为激增告警，提示业务人员关注；

√ 对近期出现频率较高的告警标记为高频告警，对告警降噪；

√ 对周期性出现的告警标记为周期告警，对告警进行降噪；

√ 对偶然出现的告警标记为偶发告警，提升业务人员引起重视。

从海量告警中快速识别出真正的生产问题，为运行部提供分析判断的依据，极大提高数据中心整体的运维能力。一方面标签数据接入 E 海智维，面向所有用户，展示告警标签以及告警产生的算法标签特征，用于数据的进一步统计以及推动告警治理工作；另一方面统计严重和紧急的告警，根据算法标签以及统计的出现次数最多 topN 的告警分类，在告警分析报告中予以重点展示，主要面向各部门、条线的负责人。

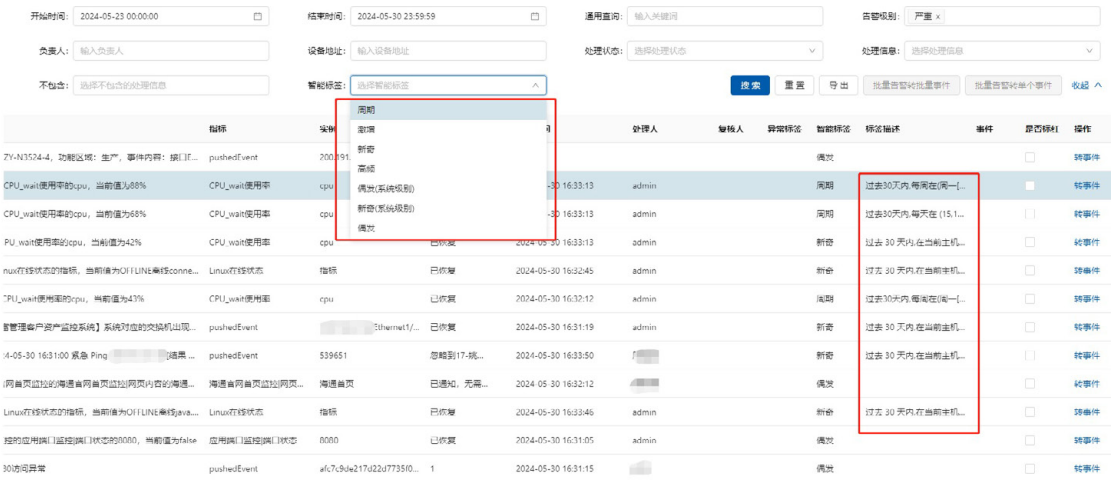


图 13 告警算法标签特征展示

通过文本相似度算法，对告警数据进行标签提取，针对结果复杂告警，人为调整匹配模板，保证基础告警标签准确度。基于 IP+ 指标名称 + 实例名称 + 告警等级组合维度生成标签，并按照如下方式进行告警标签准确性校验，结论是基础告警准确率可达 95% 以上：

（1）查询近 3 个月数据，基于告警量排序，取 topN 验证准确性；

（2）查询近 3 个月数据，随机抽验某一天的告警；

4.4 基于运维数据湖的告警治理报告场景

4.4.1 场景描述

众多的 IT 基础设施监控每天产生成千上万的告警信息，需要从运维数据湖中提取一定时间范围内的告警事件，包括但并不限于告警量、告警等级、告警部门、业务系统、负责人员、主机 IP 和处理时长等关键信息、分析从不同角度对告警数据进行统计分析并生成告警治理报告。

4.4.2 建设成效

基于运营视角，根据业务方对于报告统计维度需求，告警治理报告从不同角度不同方式进行统计，得到重点关注对象，如下所示：

》告警数量

》告警等级分布

》告警部门分布

》业务系统分布

》负责人员分布

》主机 ip 分布

》处理时长分布

文章实践落地，基于运维数据湖中纳管的告警历史和实时数据，从系统侧对告警进行细分统计，着重关注新增、周期、激增告警和超时未处理告警。

图 14 告警治理报告统计维度展示

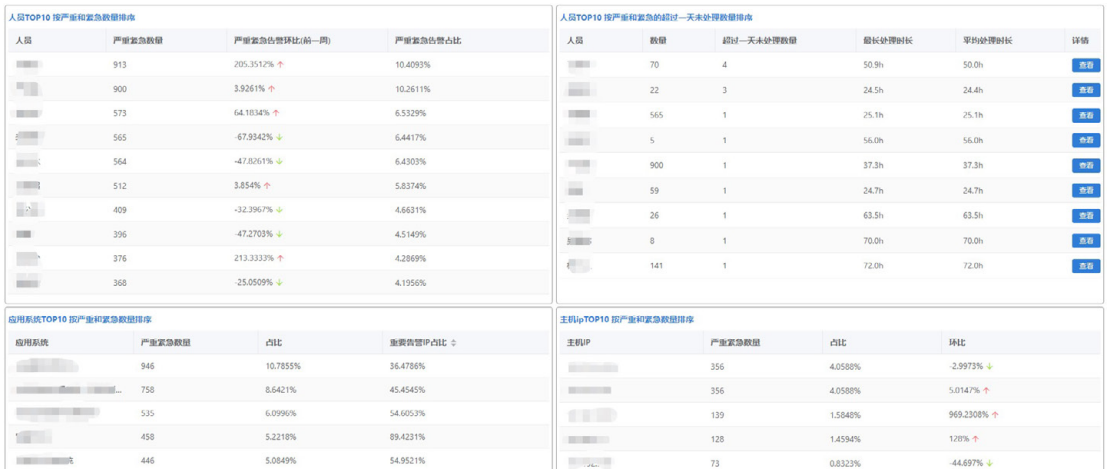


图 15 告警治理报告重要系统分析 topN 排名

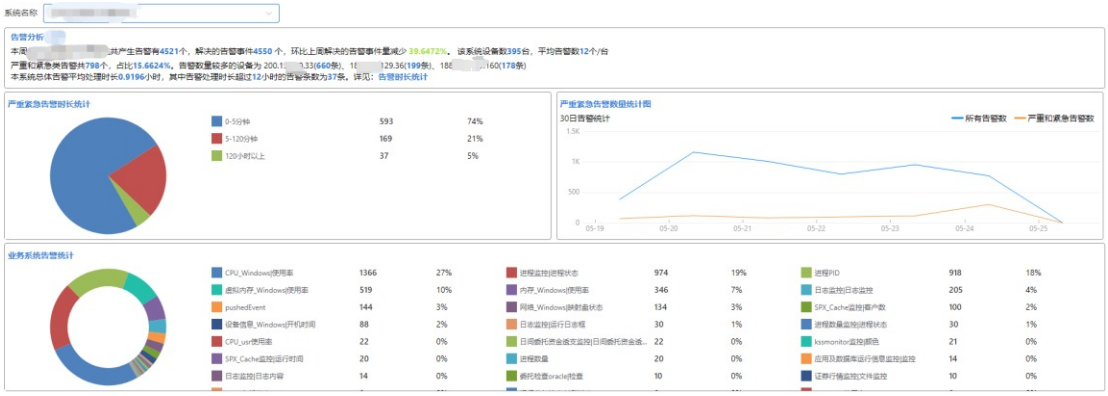


图 16 告警治理报告重要系统告警统计

图 17 告警治理报告重要系统同环比分析

五、总结与展望

在证券行业的快速发展和数字化浪潮的推动下，国泰海通证券针对运维数据湖的构建与维护进行研究和落地实践，旨在解决当前生产环境中数据管理的复杂性、统一性和相关性问题。通过全领域数据接入、标准化数据服务体系以及基于自研查询引擎的统一数据查询机制，不仅确保数据的实时有效，更为证券行业运维数据管理提供了参考范本。

在场景应用层面，本文章探讨了指标异常检测、告警智能标签等多个核心运维场景，通过数据纳管、分析和智能算法实现了故障的快速响应和风险的有效预警，展现了运维数据湖在提升证券公司业务效率和风险管理水平方面的巨大潜力。

展望未来，运维数据湖的发展将呈现以下趋势：一是智能化，通过结合 AI 算法、机器学习、深度学习、LLM 运维大模型等先进技术，实现运维的自动化、智能化，进一步赋能运维场景，提升运维分析和处置效率；二是全面化，将运维数据湖的应用范围从传统的 IT 运维领域扩展到业务运营、风险管理等多个领域，实现数据的全面整合和深度分析；三是创新化，不断引入新框架、新模式，推动运维数据湖的前瞻发展，奠定坚实的数据基础和扩展能力，帮助运维人员提升整体的执行和开发效率，助力实现运维自治提供强大动力。

综上所述，本文章提出的基于治理理念的运维数据湖构建与维护方案具有较高的理论价值和实践意义，对于支撑证券 IT 行业的持续发展及提升运维效能带来切实的帮助。将来，国泰海通证券将继续关注运维数据湖的前沿动态，不断探索和创新，为证券行业数智化转型和发展贡献更多力量。

参考文献：

[1] JANSSEN N E. The evolution of data storage architectures：examining the value of the data Lakehouse. Enschede：University of Twente，2022。

[2] Data Lakehouse，Databricks，网址：https://www.databricks.com/glossary/data-lakehouse，最后访问日期：2023 年 8 月 30 日。

[3] ARMBRUST M，GHODSI A，XIN R，et al. Lakehouse：a new generation of open platforms that unify data warehousing and advanced analytics，Stanford，网址：https://cs.stanford.edu/people/matei/papers/2021/cidr_lakehouse.pdf，最后访问日期：2023 年 8 月 30 日。

[4] 中国证券业协会印发《证券公司网络和信息安全三年提升计划（2023-2025）》，中证协网站，时间：2023-06-09，网址：https://www.sac.net.cn/ljxh/xhgzdt/202306/t20230609_60404.html，最后访问日期：2023 年 8 月 30 日。

[5] 金融科技政策重视程度提升，证券业 IT 市场规模不断扩大，搜狐财经，时间：2023-07-27，网址：https://www.sohu.com/a/706690082_121353797，最后访问日期：2023 年 8 月 31 日。

[6] Data Lake vs Warehouse vs Data Lakehouse | Know the Difference，Xenonstack，时间：2022-10-31，网址：https://www.xenonstack.com/insights/data-lake-vs-warehouse-vs-data-lakehouse，最后访问日期：2023 年 9 月 2 日。

[7] 首批！中国信通院“可信大数据”湖仓一体数据平台建设成熟度专项测试正式启动，大数据技术标准推进委员会，时间 2023-09-04，网址：https://mp.weixin.qq.com/s/_1ZuG7LGFnhG2cCCnuTHnw，最后访问日期：2023 年 9 月 5 日。

[8] 湖仓一体技术与产业研究报告（2023 年），百度文库，网址：https://wenku.baidu.com/view/3d768c0959cfa1c7aa00b52acfc789eb172d9efe.htm，最后访问日期：2023 年 9 月 5 日。

一种估计并行双模型召回率的新统计学方法

何峰，陈俊

大连飞创信息技术有限公司 | E-mail：hefeng@dce.com.cn

摘　要： 召回率，即模型识别出的正样本与所有正样本的比例，是衡量模型性能的重要指标。在人工智能技术的应用中，一些特殊场景对模型召回率有着极高的要求，如重大疾病检测、恐怖活动预警、金融欺诈检测等。针对该问题，集成模型因其子模型存在多重检测机制而被广泛采纳。随着人工智能的发展，算法复杂度及参数规模均呈指数级增长，单一模型的算力需求已不容小觑，集成模型则更为可观，所以在实验前准确估计算法召回率就显得尤为重要。鉴于此，基于超几何分布，结合子模型先验召回率，本文提出一种可精准估计并行双模型召回率的新统计学方法。基于传统概率模型的对比实验结果显示，该方法在精度、准度及可靠性三方面均表现优异。不仅如此，本文还将该方法推广至并行多模型召回率的估计，实现方法的深度泛化。此外，基于多模态大语言并行双模型，结合大连商品交易所市场操纵行为认定，本文理论上设计一套市场操纵行为辅助预警系统，为期货交易所的监查系统建设及数字化转型提供有价值参考。

关键字： 召回率；人工智能；超几何分布；并行双模型；市场操纵行为辅助预警系统

一、引言

如今，随着人工智能技术逐渐落地于生活中的诸多场景，人们在享受人工智能技术带来便利的同时，也能够接受使用过程中的些许瑕疵，如电商平台推荐的产品未包含用户心怡的品牌；使用文心一言时，生成的内容还需要人工校验修改；“萝卜快跑”无人车会车时出现互相等待的短暂死循环等等。然而，一些特殊场景却对人工智能模型有着极高的要求，如重大疾病检测、恐怖活动预警、金融违规操作检测等。这些场景的共同点是允许多重检测，如心脑血管类病患是否存在发病风险的专家会诊、恐怖活动预警后的预警判定及威胁评估、金融违规操作案件定性前的多重审批等。但不能放过任何一个可疑案例，杜绝漏网之鱼的出现。

此类特殊场景具备对真阳、真阴、假阳样本友好，但排斥假阴样本的特性，相关定义如表 1 混淆矩阵所示。其中，真实正、负样本表示样本的客观属性。模型预测正、负样本即为模型在样本的客观属性未知前提下，对样本正负的判定结果。

当样本为真实正、负样本，同时模型也预测其为正、负样本，即为表 1 中的真阳、真阴样本，表示模型预测结果与客观事实一致。当样本为真实负、正样本，模型预测其为正、负样本，模型预测结果与客观事实相悖，则会出现表 1 中的假阳、假阴样本，即统计学中的第一、二类错误。得益于特殊场景的多重检测机制，假阳样本会在检测中排除。然而假阴样本虽为真实正样本却未被模型识别，致使漏网之鱼的出现，在上述特殊场景中可能导致不可估量的损失。

综上，在特殊场景中选定的模型指标，应能充分体现模型在假阴样本识别方面的能力。召回率 Recall，即真阳样本与所有真实正样本的比例，在机器学习和统计学中，也被称为灵敏度（Sensitivity），是评估分类模型性能的一个重要指标，定义如公式（1）所示。在特殊场景的模型应用过程中，召回率被广泛选取为衡量模型性能的指标。

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

目前针对上述特殊场景，现有研究已构建出基于单一模型或集成模型的解决方案。例如在心脏病发作风险预测领域，基于 Kaggle 相关数据集，本文对比验证

表 1 混淆矩阵

	真实正样本	**真实负样本**
模型预测正样本	真阳样本（TP）	假阳样本（FP）
模型预测负样本	**假阴样本（FN）**	真阴样本（TN）

了如下算法。首先是 Nissa 团队 [1] 基于支持向量回归（Support Vector Regression， SVR）的单一模型，以及基于 Adaboost 的集成模型解决方案，召回率分别为 0.7933 与 0.8584。然后是 Rahman 团队 [2] 基于 Transformer 的算法，召回率为 0.8611。最后是 Jumphoo 团队 [3] 基于复杂模型 CNN（Convolutional Neural Network）及 Transformer 的集成算法 Conv-DeiT，召回率为 0.9298。

由以上结果可知，因为集成模型的各子模型分别对样本进行判定，形成多重检验，所以适合假阴样本的甄别，能够有效提高模型的召回率，其效果优于单一模型。如上述实验中，Adaboost 结果强于 SVR，Conv-DeiT 结果好于 Transformer。同时，通过对比 Transformer 与 Adaboost 的实验结果，可得当模型复杂度足够高，也可优于集成模型，但效果不显著。

此外，本文还调研了操纵股票价格行为检测领域。基于中国证券监督管理委员会（下称证监会）2014 年 4 月 1 日至 2022 年 7 月 31 日的操纵股票价格行为的处罚公示 [4]，Liu 团队 [5] 完成样本的收集与整理，以波动率、成交额、收益率、异常收益率为特征，分别采用决策树（Decision Tree）、随机森林（Random Forest）、多 RNN（Recurrent Neural Network）并行集成、多 LSTM（Long Short-Term Memory）并行集成，对该违规行为进行识别，召回率分别为 0.5877、0.564、0.7962、0.8152。

虽然集成模型可获得较好结果，但也存在如下痛点。第一，随着人工智能的发展，特别是进入大模型时代以来，单一模型的复杂度及参数规模与日俱增，目前已十分可观，模型运行存在算力需求高、周期长的痛点。第二，模型效果与其训练时的数据集样本量及熵值成正比。然而部分特殊场景无法提供足够规模且分布均匀的样本集，如恐怖活动预警、金融违规行为检测等。第三，实验具有偶然性，有限次的实验结果不足以体现模型真实效果。因此在实验前，对集成模型的召回率进行精准估计就显得尤为必要。

鉴于此，基于统计学理论，结合子模型先验召回率，本文提出一种可精准估计并行双模型召回率的新统计学方法，其适用于所有的非生成式机器学习模型。基于传统概率模型的对比实验结果显示，该方法在准度、精度及可靠性三方面均表现优异。不仅如此，本文还将该方法推广至并行多模型召回率的估计，实现方法的深度泛化。此外，以多模态大语言并行双模型为基础，结合大连商品交易所市场操纵行为认定 [6]，理论上本文设计了一套市场操纵行为辅助预警系统，为期货交易所的监查系统建设及数字化转型提供有价值参考。

二、方法

该部分包含两点主要内容。其中 2.1 对本文中使用的数学符号及基本算子，如组合计算等进行说明。2.2-2.4 详细阐述了本文的核心算法，即并行双模型的基于传统概率模型的召回率估计、基于超几何分布的召回率估计，以及两种召回率估计方法在并行多模型上的深度泛化。

2.1 数学符号及指标定义

对本文中用到的数学符号及相应指标定义如下。N 表示样本集的总样本量。e_i 表示模型 i 基于给定样本集的假阴样本量。r_i 表示模型 i 的先验召回率，如公式（2）所示。m 表示用到的模型数量。C_n^k 表示组合计算，即从 n 个样本中有放回抽取 k 个，共有多少种可能场景，如公式 (3) 所示。

$$r_i = 1 - e_i / N \quad (2)$$

$$C_n^k = \frac{n!}{k!(n-k)!} \quad (3)$$

在本文中，当且仅当两个子模型对同一样本均作出假阴判定时，并行双模型假阴判定成立。参考现有研究，结合场景特殊性，本文亦选用召回率作为模型性能指标，定义如公式（1）所示。

2.2 基于传统概率模型的召回率估计

基于上述设定，给出基于传统概率模型的并行双模型召回率估计，如公式（4）所示。其中，在已知两个子模型先验召回率情况下，首先计算两个子模型同时出现假阴样本的概率，然后推出并行双模型的召回率估计值 $Recall_{td}$ 。

$$Recall_{td} = 1 - (1 - r_1) \times (1 - r_2) \quad (4)$$

2.3 基于超几何分布的召回率估计

在统计学中，超几何分布（Hypergeometric distribution）是一种离散概率分布，在 19 世纪初期由拉普拉斯系统总结后正式提出 [7]，用来描述从有限集合中成功无放回抽出指定种类样本次数的概率。随着统计学和概率论的发展，超几何分布在质量控制、人口统计、生物学以及经济学等领域的应用变得更加广泛。下面将通过具体示例介绍超几何分布的计算过程。

如图 1 所示，其中包含了若干水果。做以下假设，假设图 1 中有 200 个水果，其中圆框内的青枣有 20 个。现需从 200 个水果中随机抽取 20 个，其中至少包含一个青

枣的概率是多少？计算过程如下所示。

图 1 超几何分布计算流程举例

$$P=\sum_{i=1}^{20}\frac{C_{20}^{i}\times C_{180}^{20-i}}{C_{200}^{20}}=0.160929$$

其中分母表示从 200 个水果中无放回抽取 20 个的可能场景数量。i 表示抽取的 20 个水果中包含青枣的数量，C_{20}^{i} 则表示从 20 个青枣中抽取 i 个的可能场景数量。C_{180}^{20-i} 表示从不包含青枣的 180 个水果中抽取 20-i 个的可能场景数量。以上就是超几何分布的一个具体应用案例。

综上，基于特殊场景相关设定，结合超几何分布，推导并行双模型的召回率估计如公式（5-6）所示。

$$Pvalue_{hg}=\sum_{i=1}^{\min(e_1,e_2)}\frac{C_{e_1}^{i}\times C_{N-e_1}^{e_2-i}}{C_{N}^{e_2}} \quad (5)$$

$$Recall_{hg}=1-P\_value_{hg} \quad (6)$$

类似上面案例，在样本量为 N 的集合中，公式（5）计算了子模型 1 假阴的e_1个样本中至少存在一个与子模型 2 的e_2个假阴样本相同，即并行双模型出现假阴样本的概率。同基于传统概率模型召回率估计的方法类似，公式（6）可得并行双模型的召回率估值$Recall_{hg}$。

2.4 方法泛化

基于已推导的并行双模型召回率估值，结合统计学中相关理论，本文对并行多模型的召回率估值进行推导如下。同时，并行多模型的假阴判定同并行双模型类似，即当且仅当所有子模型对同一样本判定假阴时，假阴成立。

首先，本文泛化推导了基于传统概率模型的召回率估值，如公式（7）所示。

$$Recall_{td\_g}=1-\prod_{i=1}^{m}(1-r_i) \quad (7)$$

然后，对并行多模型的基于超几何分布的召回率估计进行推导，如公式（8-9）所示。

$$Pvalue_{hg\_g}=\sum_{i=1}^{\min(e_1,e_2,\cdots,e_m)}C_{e_1}^{i}\times\prod_{j=2}^{m}\frac{C_{N-e_1}^{e_j-i}}{C_{N}^{e_j}} \quad (8)$$

$$Recall_{hg\_g}=1-P\_value_{hg\_g} \quad (9)$$

公式（8）的核心思想同并行双模型的召回率估值推导一致。其结果为多个子模型基于同一样本集的实验，所产生的各假阴样本集至少包含一个相同样本的概率。公式（9）同公式（6）相似，计算真阳性样本在所有正样本中的比例。

最后，本文还考虑了一种特殊情况，即当 m 个子模型的先验召回率相同时，两种方法的并行多模型召回率估计如公式（10-12）所示。其中$Recall_{td_gs}$表示基于传统概率模型的召回率估值，而$Recall_{hg_gs}$则表示基于超几何分布的召回率估值。

$$Recall_{td\_gs}=1-(1-r)^{m} \quad (10)$$

$$Pvalue_{hg\_gs}=\sum_{i=1}^{e}C_{e}^{i}\times\left(\frac{C_{N-e}^{e-i}}{C_{N}^{e}}\right)^{m-1} \quad (11)$$

$$Recall_{hg\_gs}=1-P\_value_{hg\_gs} \quad (12)$$

三、结果分析

该部分包含两点主要内容。3.1 描述了两种并行双模型召回率估值方法的实验结果，并详细解析了相较于传统模型，基于超几何分布的召回率估计的三点优势。3.2 则详细介绍了本文设计的市场操纵行为辅助预警系统。

3.1 召回率估值结果分析

基于上述两种并行双模型召回率的估值方法，本文以子模型先验召回率$r_1=r_2=0.9$为例。首先代入公式（3）可得基于传统概率模型的召回率估值$Recall_{td}=0.99$。然后代入公式（5-6）可得基于超几何分布的召回率估值结果如表 2 所示。

表 2 中每一行前三列表示基于对应样本量的测试集，各子模型的假阴样本量。后两列为并行双模型的相应 P_value 值及召回率。这里进行两点说明。第一，因在基于每行相应样本量测试中，子模型召回率应不低于 0.9。所以每行的子模型可能不尽相同。第二，当样本量在 5000 及以上时，相应的 P_value 值不等于 0，而是一个极小的

表 2 基于超几何分布的并行双模型召回率估值

N	e_1	e_2	P_value	$Recall_{hg}$
100	10	10	0.738471	0.281529
200	20	20	0.378211	0.621789
500	50	50	0.02782	0.97218
1000	100	100	0.000198	0.999802
2000	200	200	5.5778e-9	0.999999
5000	500	500	0.0	1.0
10000	1000	1000	0.0	1.0

小数。但由于现有算力的计算位数有限，已无法显示结果中非 0 的位数，所以记为 0，相应召回率记为 1。

由此可得如下三点结论：

第一，基于超几何分布的召回率估计在准度上表现优异。基于传统概率模型的召回率估计在定义上计算的是两个子模型同时出现假阴样本的概率。而在定义上，基于超几何分布的召回率估计计算的是两个子模型同时至少在一个相同样本上出现假阴判定的概率，所以其召回率估值相较于传统模型更为准确，二者间关系如公式（13-15）所示。

$$Recall_{td} \le Recall_{hg} \quad (13)$$

$$Recall_{td\_g} \le Recall_{hg\_g} \quad (14)$$

$$Recall_{td\_gs} \le Recall_{hg\_gs} \quad (15)$$

第二，基于超几何分布的召回率估计在精度上具备一定优势。基于传统概率模型的召回率估计在计算过程中，仅考虑子模型是否同时出现假阴判定这两种情况。而基于超几何分布的召回率估计在计算过程中，不仅考虑了子模型是否同时在相同样本上出现假阴，且进一步量化了假阴样本的数量，即公式（5）中的变量 i。所以其相较于传统模型更为精确。如表 2 所示，当测试集样本量达 2000 时，估值精度可达 $1e^{-9}$。

第三，基于超几何分布的召回率估计充分考虑了子模型在不同样本量测试集上的表现，使得估值更为合理可靠。如表 2 所示，当在样本量 200 的测试集上达到召回率 0.9 的子模型，其在样本量 5000 的测试集上未必能够取得相同效果。但基于传统概率模型的召回率估计未对该方面进行约束，所以其估值存在一定的不可靠性。

3.2 市场操纵行为辅助预警系统

本文不仅实现并行双模型召回率的精准估计，还结合多模态大语言模型（Multimodal Large Language Models，MLLM），以及大连商品交易所市场操纵行为认定，理论上设计了一套期货市场操纵行为辅助预警系统，如图 2 所示。

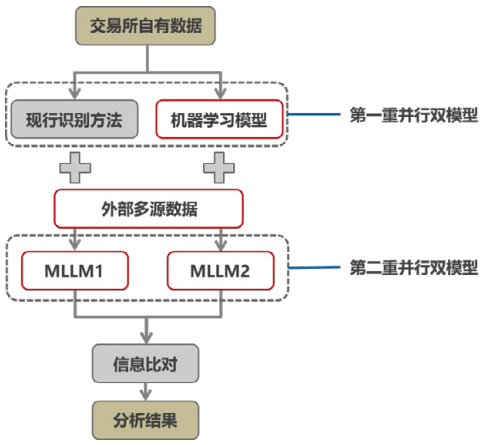


图 2 基于多模态大语言模型的市场操纵行为辅助预警系统

首先，图 2 中的市场操纵行为辅助预警系统保留了现有的识别方法。在输入同为交易所自有数据前提下，为其适配一个结构相对简单的机器学习识别模型，构成第一重并行双模型。该结构在保证运行效率的同时，有效避免假

阴样本的出现。

然后，本文选取外部多源数据，如客户在其他金融领域的投资、征信情况，以及万得、彭博等机构对客户投资品种的相关音视频解析及研报。该多源数据经加工后，结合第一重并行双模型识别结果，完成第二重并行双模型输入向量的构建。

最后，本文采用两个异构的多模态大语言模型 MLLM1、MLLM2 构建第二重并行双模型。结合上述多维输入向量，各模型完成判定，并通过信息比对得出最终的分析结果，为市场操纵行为的预警及后续判定提供有效辅助。

本文提出的市场操纵行为预警辅助系统具备如下三点优势：第一，该系统在保留现行识别方法基础上，采用了两重并行双模型，充分发挥了并行模型在规避假阴样本方面的优势。第二，该系统不仅使用了期货交易所自有数据，还同时采用了外部多源数据。各维度数据间可形成有效互补，能够更为全面地刻画用户的交易行为，为深度提升模型召回率奠定基础。第三，多模态大语言模型的应用，不仅能够实现多源数据的有效整合，还可实现数据间逻辑的全面深入推理，在模型端为假阴样本的规避提供强有力的保障。

四、总结

本文针对召回率要求极高的特殊场景，结合超几何分布，提出了一种新的估值方法，可实现并行双模型召回率的精准估计。同基于传统概率模型的对比实验结果显示，新方法在准度、精度及可靠性三方面均表现优异。不仅如此，本文还推导出并行多模型的召回率估计，使得新方法得以深度泛化。此外，基于多模态大语言并行双模型，结合大连商品交易所市场操纵行为认定，本文还设计了一套市场操纵行为辅助预警系统，为期货交易所的监查系统建设及数字化转型提供助力。

同时，本文也存在诸多不足，如本文提出的新方法并未在实践中得以充分检验。在未来工作中，本文成果会逐步应用于真实案例，其中不足也会被加以修正。再者，本文在召回率估值结果分析部分，仅展示子模型召回率均为 0.9 条件下的并行双模型估值，后续会增加更多实验结果佐证方法在准度、精度及可靠性方面的优势。最后，本文设计的市场操纵行为辅助预警系统目前尚处于理论设计阶段，如果后续落地需根据实际情况逐步改进。在接下来的工作中，会陆续对相关工作予以补充、完善及深化。

参考文献：

[1] N. Nissa, S. Jamwal, and M. Neshat, "A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques," Computation, vol. 12, no. 1, p. 15, 2024. doi:10.3390/computation12010015.

[2] A. U. Rahman, Y. Alsenani, A. Zafar, K. Ullah, K. Rabie, and T. Shongwe, "Enhancing heart disease prediction using a self-attention-based transformer model," Sci Rep, vol. 14, no. 1, 2024. doi:10.1038/s41598-024-51184-7.

[3] T. Jumphoo, K. Phapatanaburi, W. Pathonsuwan, P. Anchuen, M. Uthansakul, and P. Uthansakul, "Exploiting Data-Efficient Image Transformer-Based Transfer Learning for Valvular Heart Diseases Detection," IEEE Access, vol. 12, pp. 15845-15855, 2024. doi:10.1109/ACCESS.2024.3357946.

[4] 中国证券监督管理委员会 , " 行政处罚决定 ." http://www.csrc.gov.cn/csrc/c101971/zfxxgk_zdgk.shtml?channelid=17d5ff2fe43e488dba825807ae40d63f

[5] C. Liu, S. Li, and L. Shi, "A stock price manipulation detecting model with ensemble learning," Expert Syst. Appl., vol. 248, p. 123479, 2024.

[6] 大连商品交易所 , " 合规交易相关问题 ." http://www.dce.com.cn/dalianshangpin/ywfw/ywzy/jcyfkywzy/505264/6270063/index.html

[7] 王幼军 ." 拉普拉斯概率理论的历史研究 ," 博士 , 物理学 (物理学史), 上海交通大学 , 2003.

04 监管科技全球追踪

监管科技全球追踪

2024 年 12 月，瑞士金融市场监管局 FINMA 发布人工智能风险管理指南，指导金融机构更好管控人工智能应用风险，要求其在开发、部署和使用人工智能系统时，遵循一系列风险管理原则和实践，确保系统安全、可靠和合规运行。

2024 年 12 月，国家数据局、中央网信办、工信部、公安部、国资委印发《关于促进企业数据资源开发利用的意见》，提出针对新技术应用和新模式新业态，探索建立“沙盒监管”机制，构建鼓励创新、弹性包容的治理环境。

2024 年 12 月，国家金融监管总局发布《银行保险机构数据安全管理办法》，从强化数据治理顶层设计、落实分类分级管理、健全数据安全管理体系、加强个人信息保护、完善风险监测处置机制等方面提出要求，规范数据处理活动、保障数据安全。

2025 年 1 月，香港金融管理局发布 2025 年工作重点，涵盖应对信贷风险环境变动、打击诈骗等，同时将聚焦监管科技和金融科技推广，包括生成式人工智能沙盒、分布式分类帐技术监管孵化器，实施以数据、科技主导的监管框架，并进行科技成熟度评估。

2025 年 2 月，中国证监会发布《关于资本市场做好金融“五篇大文章”的实施意见》，提出加快推进数字化、智能化赋能资本市场，加强行业机构金融“五篇大文章”服务能力等 18 条政策举措。

2025 年 2 月，美国证券交易委员会（SEC）正式成立网络和新兴技术部门（CETU），核心工作聚焦调查借助人工智能、机器学习等新兴技术实施的欺诈活动，打击针对零售经纪账户的黑客攻击，以及揭露利用社交媒体、暗网和欺诈网站进行的欺骗行径。

2025 年 2 月，欧洲央行清算支付系统 TARGET2 宕机超 7 小时。该系统承载着欧元区各国央行间的实时全额自动清算服务，每日处理超 3 万亿欧元的支付和金融交易。官方公告称该事件由“基础设施硬件失效”所引发。

2025 年 3 月，国际证监会组织（IOSCO）发布关于资本市场人工智能的咨询报告，指出金融机构日益借助人工智能辅助算法交易、投资研究、情绪分析等应用，风险挑战则包括人工智能的恶意使用、模型和数据考量、集中化与外包及第三方依赖、人机系统交互等方面。

2025 年 3 月，美国证券交易委员会（SEC）扭转加密货币执法方向，至少驳回或叫停 8 起针对大型加密货币公司的执法案件，包括涉及 Coinbase、币安等行业巨头的诉讼，预示数字资产的监管方式可能迎来重新评估。

2025 年 3 月，国家公共数据资源登记平台上线试运行。根据登记管理要求，登记机构按照行政层级和属地原则提供规范化、标准化、便利化登记服务。直接持有或管理公共数据资源的党政机关和事业单位，应登记纳入授权运营的公共数据资源。

2025 年 3 月，纳斯达克宣布，在获得监管批准和行业协调的前提下，自 2026 年下半年起提供每周五天的 24 小时交易，这标志着继纽交所后又一交易所加入延长交易时段的行列。

2025 年 4 月，英国金融行为监管局（FCA）计划推出新的人工智能实时测试服务，旨在为开发面向消费者或市场的人工智能模型的公司，提供监管支持与协作测试环境，使其能在产品推向市场前完成验证。

2025 年 4 月，瑞士国家银行、国际清算银行和世界银行成功开发概念验证（PoC）平台，以分布式账本技术 (DLT) 替代传统纸质本票票据，自动执行相关业务流程，降低成本。

2025 年 4 月，长桥集团（LongBridge）发布证券行业首个券商 MCP 协议，通过标准化接口使投资者直接用自然语言指令让 AI 执行投资分析、股票交易及风险管理等操作，同时配套“LLMs Text”工具支持量化投资者以自然语言编写代码。

2025 年 4 月，美商品期货交易委员会（CFTC）就 7X24 小时交易的有关事项征询意见，征询重点包括：CFTC 监管的衍生品市场扩展至全天候交易对交易清算、风险管理体系的差异化影响，24/7 交易模式可能引发的市场诚信、客户保护及零售交易等领域风险，配套清算系统在连续运作环境下的挑战等。

2025 年 4 月，英国央行发布政策文件，指出人工智能在算法交易中的应用可能加剧市场波动性并放大金融体系的不稳定性，随着越来越多金融机构采用人工智能进行投资决策，市场将面临“算法趋同”风险。

2025 年 4 月，中国人民银行等六部门联合印发《促进和规范金融业数据跨境流动合规指南》，促进中外资金融机构金融业数据跨境流动更加高效、规范，进一步明确数据出境的具体情形以及可跨境流动的数据项清单，便利数据跨境流动。

2025 年 5 月，越南胡志明市证券交易所（HOSE）计划上线启用新的交易系统——KRX 系统。该系统自 2012 年开始建设，由韩国证券交易所研发，预期提供更高效、稳定的交易结算环境。

2025 年 5 月，国际标准化组织（ISO）正式发布《老龄化社会 老龄化包容性数字经济通用要求与指南》（ISO 25556:2025），该标准由我国牵头制定，是 ISO 首个人口老龄化视角下的数字经济标准。

《交易技术前沿》征稿启事

《交易技术前沿》由上海证券交易所主管、主办，主要面向全国证券、期货等相关金融行业的信息技术管理、开发、运维以及科研人员。近期重点征稿主题如下：

一、云计算

（一）云计算架构

主要包含但不限于：云架构剖析探索，云平台建设经验分享，云计算性能优化研究。

（二）云计算应用

主要包含但不限于：云行业格局与市场发展趋势分析，国内外云应用热点探析，金融行业云应用场景与实践案例。

（三）云计算安全

主要包含但不限于：云系统下的用户隐私、数据安全探索，云安全防护规划、云安全实践，云标准的建设、思考与研究。

二、人工智能及大模型技术

（一）应用技术研究

主要包含但不限于：大语言模型 /AIGC 的数据处理和治理、可解释的人工智能及大语言模型、用于大语言模型 /AIGC 的神经网络架构、训练和推理算法、多模态 AI 等。

（二）应用场景研究

主要包含但不限于：基于人工智能或大语言模型的智能客服、语音图像文本等数据挖掘、柜员业务辅助等。

主要包含但不限于：金融预测、反欺诈、授信、辅助决策、金融产品定价、智能投资顾问等。

主要包含但不限于：金融知识库、风险控制等。

主要包含但不限于：机房巡检机器人、金融网点服务机器人等。

三、数据中心

（一）数据中心的迁移

主要包含但不限于：展示数据中心的接入模式和网络规划方案；评估数据中心技术合规性认证的必要性；分析数据中心迁移过程中的影响和业务连续性；探讨数据中心迁移的实施策略和步骤。

（二）数据中心的运营

主要包含但不限于：注重服务，实行垂直拓展模式；注重客户流量，实行水平整合模式；探寻数据中心运营过程中降低成本和提高服务质量的途径。

四、分布式账本技术（DLT）

（一）主流分布式账本技术的对比

主要包含但不限于：技术架构、数据架构、应用架构和业务架构等。

《交易技术前沿》征稿启事

（二）技术实现方式

主要包含但不限于：云计算 + 分布式账本技术、大数据 + 分布式账本技术、人工智能 + 分布式账本技术、物联网 + 分布式账本技术等。

（三）应用场景和案例

主要包含但不限于：结算区块链、信用证区块链、票据区块链等。

（四）安全要求和性能提升

主要探索国密码算法在分布式账本中的应用，以及定制化的硬件对分布式账本技术性能提升的作用等。

五、信息安全与 IT 治理

（一）网络安全

主要包括但不限于：网络边界安全的防护、APT 攻击的检测防护、云安全生态的构建、云平台的架构及网络安全管理等。

（二）移动安全

主要包括但不限于：移动安全管理、移动互联网接入的安全风险、防护措施等。

（三）数据安全

主要包括但不限于: 数据的分类分级建议、敏感数据的管控、数据共享的风险把控、数据访问授权的思考等。

（四）IT 治理与风险管理

主要包括但不限于：安全技术联动机制、自主的风险管理体系、贯穿开发全生命周期的安全管控、安全审计的流程优化等。

六、交易与结算相关

（一）交易和结算机制

主要包含但不限于: 交易公平机制、交易撮合机制、量化交易、高频交易、高效结算、国外典型交易机制等。

（二）交易和结算系统

主要包含但不限于: 撮合交易算法、内存撮合、双活系统、内存状态机、系统架构、基于新技术的结算系统等。

投稿说明：

1、本刊采用电子投稿方式，投稿采用 Word 文件格式（格式详见附件），请通过投稿信箱 ftt.editor@sse.com.cn 进行投稿，收到稿件后我们将邮箱回复确认函。

2、稿件字数以 4000-6000 字左右为宜，务求论点明确、数据可靠、图表标注清晰。

3、不设固定截稿日期，常年对外收稿。收齐一定数量的稿件后将尽快组织专家评审。

4、投稿联系方式 021-68607130, 021-68607129 欢迎金融行业的监管人员、科研人员及技术工作者投稿。稿件一经录用发表，将酌致稿酬。

附件：投稿格式（可通过电子邮件索要电子模版）

标题（黑体 二号 加粗）
作者信息（姓名、工作单位、邮箱）（仿宋 GB2312 小四）
摘要：（仿宋 GB2312 小三 加粗）
关键字：（仿宋 GB2312 小三 加粗）
一、概述（仿宋 GB2312 小三 加粗）
二、一级标题（仿宋 GB2312 小三 加粗）
（一）二级标题（仿宋 GB2312 四号 加粗）
1、三级标题（仿宋 GB2312 小四 加粗）
（1）四级标题 （仿宋 GB2312 小四）
正文内容（仿宋 GB2312 小四）
图：（标注图 X. 仿宋 GB2312 小四）
正文内容（仿宋 GB2312 小四）
表：（标注表 X. 仿宋 GB2312 小四）
正文内容（仿宋 GB2312 小四）
三、结论 / 总结（仿宋 GB2312 小三 加粗）
四、参考文献（仿宋 GB2312 小四）

电子平台：

欢迎访问我们的电子平台 http://www.sse.com.cn/services/tradingtech/transaction/。我们的电子平台不仅同步更新当期的文章，同时还提供往期所有历史发表文章的浏览与查阅，欢迎关注！