# 证券公司高质量数据集标准研究

**【摘要】**随着人工智能与大数据技术在证券行业的深度融合，数据已成为驱动业务创新与风险管理的关键要素。然而，在智能化转型过程中，行业普遍面临数据质量参差不齐、多模态数据处理困难、数据标准不统一、流通机制不健全等突出问题，严重制约了智能化业务模型的准确性与泛化能力。建立覆盖数据全生命周期的高质量数据集标准，推动数据资源的规范化、可信化与要素化，已成为行业数字化转型的迫切需求。本课题在全面调研证券公司高质量数据集建设现状与瓶颈的基础上，借鉴《高质量数据集建设指南》的顶层框架，对数据采集、预处理、分析建模与应用服务等关键环节进行了系统性梳理，按照分层建设、迭代优化的原则，对多源数据接入、质量控制、智能标注、评估认证等核心过程进行了标准化设计，并在此基础上构建了覆盖数据源、处理流程、资产管理与应用场景的高质量数据集建设步骤，形成了适用于全行业推广的高质量数据集标准草案，为证券行业数据要素化与智能化升级提供了重要支撑。

**关键词：**多模态数据；数据标注；难例标注；数据质量；

# 正　文

## 一、引言

### （一）课题研究背景及意义

2024 年 1 月，国家数据局等 17 部门联合发布《"数据要素×"三年行动计划（2024—2026 年）》，明确提出在金融服务等重点领域实施"数据要素×"行动，要求"提升数据要素在金融领域的配置效率，推动数据在风险定价、投资决策等场景的深度应用"。这一重要文件为证券行业的数据要素化发展指明了具体方向和实施路径。与此同时，中国信息通信研究院发布的《高质量数据集建设指南》为行业提供了具体技术规范，系统阐述了高质量数据集在数据采集、标注、评估等环节的技术标准，特别强调了训练数据质量对人工智能应用效果的决定性影响。在证券行业数字化转型加速推进的背景下，证监会与国家标准委联合印发的《关于加强证券期货业标准化工作的指导意见》进一步明确要求"建立健全行业数据标准体系，以标准化支撑行业数字化转型"。随着深度学习、大模型等人工智能技术在证券行业的快速应用，投资研究、交易执行、风险管理等核心业务环节对高质量训练数据的需求呈现爆发式增长。然而，行业当前面临的数据标注不规范、质量评估体系缺失、多源数据融合困难等问题，严重制约了人工智能技术在证券业务的深度应用效果。在此背景下，开展高质量数据集标准研究不仅是落实国家数据要素战略的

重要举措，更是推动证券行业人工智能技术规模化应用的基础工程，具有重要的战略意义和实践价值。

对证券公司而言，高质量数据集建设是推动人工智能技术实现突破性发展的关键基石。在模型训练层面，高质量数据集直接决定了机器学习算法的准确性和泛化能力。以量化投资为例，基于经过严格质量控制的市场行情数据、公司基本面数据和另类数据训练的交易模型，能够显著提升策略的稳定性和超额收益水平。特别是在当前大模型技术快速发展的背景下，证券公司需要构建大规模的高质量金融语料库，这些语料数据需要经过专业的清洗、标注和校验，才能支撑大模型在金融领域的有效训练。在算法优化层面，高质量数据集能够有效提升人工智能系统的推理能力。例如，在智能投研场景中，经过精准标注的上市公司财报数据、行业研报数据和宏观经济数据，能够帮助自然语言处理模型更准确地理解金融文本语义，提升信息提取和知识推理的准确性。在智能风控领域，完整且准确标注的交易行为数据、账户流水数据和市场数据，能够支撑更复杂的异常检测算法，显著提升风险识别的精准度。此外，高质量数据集还是推动人工智能技术持续迭代的重要保障。通过建立标准化的数据采集、标注和评估流程，证券公司能够构建持续自我完善的数据飞轮，为机器学习模型的持续优化提供稳定可靠的数据供给。特别是在深度学习模型训练中，高质量标注数据能够有效防

止模型过拟合，提升在实际业务场景中的泛化能力，这对于证券业务这种高动态性、高复杂性场景尤为重要。

证券公司高质量数据集建设作为企业知识体系数字化转型的核心基础，通过与知识库建设的深度融合，为企业知识的系统化积累与智能化应用提供了全新范式。在数据输入层面，高质量数据集通过标准化的采集流程和严格的质量校验，将原本分散在各个业务系统的非结构化文档、研究报告、客户交流记录等知识载体，转化为具有统一标准和规范格式的结构化知识单元。这些经过标注、清洗和分类的知识数据进入企业知识中台后，依托本体建模、知识图谱等技术，建立起概念清晰、关联紧密的知识网络体系，有效解决了传统知识管理中信息孤岛、检索困难、关联性弱等痛点。在知识应用层面，基于高质量数据集构建的知识中台能够为业务人员提供精准的知识检索、智能问答和决策支持服务，同时通过持续学习机制不断吸收新的知识输入，形成知识积累的良性循环。这种以高质量数据集为驱动、以知识中台中台为载体的新型知识管理模式，不仅显著提升了企业知识的利用效率和价值转化能力，更为企业构建持续创新的知识生态系统奠定了坚实基础。

### （二）课题研究目标及主要内容

1. 研究目标

遵循证券期货业标准体系建设思路，参考国内外高质

量数据集建设框架与人工智能数据集建设思路，构建证券行业高质量数据集标准体系，指导证券公司开展数据采集、标注、评估和应用的规范化工作。通过建立覆盖多模态数据的全流程质量标准，推动训练数据的规范化管理和质量控制，为行业人工智能技术的规模化应用提供可靠数据支撑。借助高质量数据集标准的推广实施，助力证券公司智能化业务场景的快速落地、算法模型的精准优化，以及基于数据驱动的业务创新与风险管控能力提升。

2.研究主要内容

本课题《证券期货行业高质量数据集标准化研究》聚焦行业智能化转型中的数据基础瓶颈，通过系统调研与分析，已完成高质量数据集建设核心框架的初步构建。研究围绕数据采集、预处理、分析建模与应用服务四大阶段，梳理出包含多源数据接入、质量控制、智能标注、评估认证等16个关键环节的标准化流程，并针对多模态数据特征制定了涵盖基础质量、安全评估、内容评估和应用评估的四维指标体系。课题组通过专家访谈与业务-技术联动机制，深入识别了数据采集格式混乱、标注规范缺失、流通机制不健全等行业核心瓶颈，并提出以标准化推动数据要素化流通的解决方案。研究已形成高质量数据集建设的方法框架，完成行业统一的高质量数据集标准草案研制、试点完成了数据集在证券业务场景的场景化应用。

## 二、研究方法

（一）研究方法

由于本课题无市场直接参考内容，课题组首先通过对实地深度访谈多名人工智能专家、科技部门专家、业务部门专家等，对于人工智能及生态发展中使用过程中存在的问题难点及成因进行了解析，获得了高质量数据集的适用的场景和需要面对的问题。其次利用中电研究院现有数据能力，在众多必要步骤和要点上进行了深入实践，覆盖高质量数据集采集环节、质量控制环节、智能标注环节、数据评估环节、发布订阅等，对各个环节进行了评估和设计,并结合知识库及知识中台的建设,进行了深入研究。最后课题组针对相关成果进行整理分析，对高质量数据集建设过程方法、步骤、评估等内容进行总结。

对高质量数据集标准化的关键工作环节进行系统化梳理与深入探索，重点聚焦于多态采集、多态数据质量、多态数据标注及多态数据发布等核心流程。在多态采集环节，研究涵盖结构化数据、非结构化文本、图像、音频、视频等多源异构数据的标准化采集规范。针对多态数据质量，研究制定多维度的质量评估体系。在多态数据标注领域，研究构建标注管理的全流程标准体系，加强对难例样本专项管理。在多态数据发布环节，研究数据集版本管理、权限控制、溯源追踪和合规审查的标准化发布流程。研究过程要求对各环节标准名称、技术定义和质量要求的精细化定义，研究形成完

整的工作流程说明与方法论体系，为证券行业高质量数据集建设提供了可操作、可落地的标准化解决方案。并结合证券期货行业数据平台建设状况，进行高质量数据集在应用场景的抽象展望。

（二）研究步骤

本课题的整体研究工作分为现状调研和材料整理、人工智能模型设计和标准研制、课题总结三个阶段：

1.现状调研和材料整理阶段

本阶段主要工作：

一是收集本课题组参与单位高质量数据集材料、技术研究成果；

二是开展证券公司高质量数据集建设现状调研工作，调研工作采用以线上调研为主、资料调研和专家访谈为辅的调研方式，选择科技公司和重点证券公司，开展高质量数据集全流程调研，全面了解不同公司高质量数据集发展、人工智能落地执行、智能模型建设以及高质量数据集应用建设等情况。

2.高质量数据集设计和标准研制阶段

本阶段是课题研究内容的核心环节，主要工作包括：

一是经典数据治理建设流程，并结合人工智能工作要点和应用场景要求，进行高质量数据集建设流程梳理，确定关键环节和步骤；

二是针对中电研究高质量工作成果，从环节定义、说明、技术、评估和验证等多个角度进行补充，确保本次标准设计成果是适合各公司实际，可落地应用推广；

三是研究市场中已经呈现的高质量数据集成果，包括评估成果、试点成果、行业异同等，为高质量数据集标准设计的合理性和规范性奠定基础；

四是研究科技公司和证券公司已有的应用实践和高质量数据集的设计成果，尤其是知识库或知识中台的构思及实战方案；

五是开展高质量数据标准设计，对前期梳理的框架成果、验证成果进行整合，确保成果覆盖的高质量数据集的全链路，后期对相关成果验证与评审支撑有效论证；

六是开展工作验证工作，对设计完成的高质量数据集标准进行合理性、有效性的验证。

3.课题总结阶段

完成高质量数据集标准草案和研究报告撰写，并组织专家论证和征求意见，完善标准草案内容。

## 三、研究结果

课题组严格按照任务书要求，遵循集中规划、统一标准、分工协作、分步实施工作原则，制定了课题研究计划、详细的工作方案，建立了参与单位联动机制，保质保量顺利完成各项研究任务，并形成了证券公司高质量数据集标准研究报

告和标准草案。课题研究的关键内容总结如下：

（一） 证券公司高质量数据集现状调研

采用市场调研、资料调研和专家访谈等相结合的调研方法，完成科技公司与证券公司高质量数据集现状调研工作，摸清高质量数据集当前发展及建设现状、证券公司高质量数据集应用情况，厘清了引起人工智能发展不准确问题的重要因素，形成了对证券公司高质量数据标准研究课题调研初步总结，为高质量数据集通用标准的构建奠定基础。

（1） 科研机构在高质量数据集的推进工作调研

通过调研 5 家国家级科研机构在金融高质量数据集领域的推进情况，发现他们主要聚焦于前沿方法论探索、基准数据集构建以及关键瓶颈问题的攻坚上。它们的工作虽不直接面向业务场景，却为金融人工智能的理论突破和技术标准化提供了不可或缺的基石。其核心贡献在于，通过构建高纯度、高复杂度的实验级数据集，定义了诸多金融 AI 任务的性能天花板，并揭示了数据质量本身对模型泛化能力与公平性的决定性影响。

在基准数据集构建与算法评测方面，致力于创建旨在解决特定金融分析难题的标准化数据集。例如，针对金融文本信息抽取这一核心需求，一些顶尖大学的研究团队会构建并发布经过精细标注的上市公司年报、券商研报或财

经新闻数据集。这些数据集的标注工作远超简单的实体识别，可能深入到标注文本中的财务指标变化、管理层情感倾向、业务风险提示等深层语义关系。以金融关系抽取数据集为例，其构建过程通常包括：由金融专家设计详尽的标注 schema，明确各类实体，如公司、产品、人物等内容，并标准和关系，如竞争、供应、投资等的定义；然后由具备金融背景的标注人员进行多轮标注与交叉校验，确保标注的准确性与一致性；最后，数据集会划分为公开的训练集和隐藏的测试集，用于举办国际学术竞赛。它们的工作直接证明，一个定义清晰、标注精准、经过严格质量控制的基准数据集，是衡量和提升金融自然语言处理技术水平的先决条件。

在探索数据驱动的新兴研究范式上，科研机构扮演着先锋角色。一个典型的领域是基于另类数据的宏观经济预测或企业行为分析。研究人员会系统性地收集和处理海量的、非传统的数字化痕迹。构建此类数据集的工作极具挑战性：首先需要从杂乱无章的原始信号中提取出可量化的特征；其次，需要将这些另类数据与传统的宏观经济学指标或公司财务报表等标准数据进行精确的时间对齐和关联分析，以验证其预测效力。这类研究虽然处于探索阶段，但其成功与否高度依赖于另类数据本身的质量、清洗方法的科学性以及与传统金融数据融合的准确性。它从另一个

维度警示，即便数据源再新颖，如果缺乏一套对数据获取、预处理和融合质量的通用评估标准，研究结论的可靠性与可复现性将难以保证。

在数据伦理与模型可解释性研究方面，科研机构的工作则直面金融高质量数据集的“暗面”。他们通过构建带有已知偏差属性的测试数据集，系统地研究算法公平性问题。或者通过构建包含对抗性样本的数据集，测试模型的鲁棒性。他们研究深刻地揭示，金融数据集中存在的历史偏差、样本不平衡或标注主观性，会被模型学习并放大，导致决策不公或潜在风险。科研机构的这些前沿探索不仅推动了高质量内涵的深化，从单纯的准确性扩展到公平性、鲁棒性与可解释性，也为制定金融高质量数据集的标准提供了至关重要的伦理维度和安全性考量，论证了在标准中纳入偏见检测与消减规范的必要性。

（2） 大型企业在高质量数据集的推进工作调研

大型科技企业在高质量数据集领域的实践，体现了一种规模化、体系化与价值导向的工业级执行力。它们的核心工作是将数据视为核心生产原料，立足自身数据中台体系，构建高质量数据集支撑的技术基础设施和严密的组织流程，确保数据从原始状态到赋能业务的高质量转化，并最终将这种能力产品化，服务于广泛的行业生态。

在数据生产的基础设施与流程建设上，以阿里巴巴的

Data Works、腾讯的 TBDS 等平台为代表，它们实现了从数据接入、清洗、标注、加工到服务化的流水线管理。针对自身业务产生的海量原生数据，它们建立了强制性的数据规范与接入标准，确保数据在源头处的格式统一与血缘清晰。在数据清洗环节，部署了自动化的数据质量探测规则，实时监控数据的完整性、准确性和一致性，并对不符合质量要求的数据进行隔离与告警。对于非结构化数据，如用于训练内容审核模型的图片和视频，它们建立了规模庞大的人机协同标注体系。这套体系包括标准化的标注任务设计、专业的标注人员培训、多层次的质检流程，如抽样校验、交叉验证等以及基于标注一致性的质量评估模型。

在数据的内部消费与价值闭环构建上，高质量数据集是科技公司核心业务算法的生命线。在推荐系统、广告精准投放、智能风控等关键场景中，模型的迭代升级完全依赖于高质量特征数据集和标注数据的持续供给。任何一方的数据噪声都会直接污染模型，导致业务不准确。大型企业业务团队会与数据团队紧密协作，针对特定场景定义专属的数据质量指标，并将其纳入模型的测试评估体系，形成“数据质量监控->模型训练->线上效果评估->反馈驱动数据质量优化”的持续迭代闭环。

在高质量数据集建设过程中，互联网企业正通过系统

化方法提升数据质量与价值。一方面，企业大规模构建覆盖多领域的问答对数据集，通过人机协同标注与多轮质检机制，确保问答内容的准确性与逻辑一致性。另一方面，针对模型易错场景开展难例挖掘与标准制定，组织专家团队对边界案例进行精细化标注，形成具有教学价值的挑战集，持续推动模型优化迭代。同时，企业着力建设结构化知识库，将非结构化信息转化为实体-关系-属性组成的知识图谱，为模型训练提供可靠的事实依据。这些举措共同构成了数据质量提升的闭环：知识库为基础标注提供权威参考，难例标准驱动数据标注方向，问答对则成为模型优化的直接燃料，三者协同显著增强了数据集的知识密度与学习效用。

（3） 证券企业在高质量数据集的推进工作调研

证券公司在业务数字化与智能化的内生驱动下，正全力推进高质量数据集的建设和应用工作。其重心从过去分散的数据采购，转向以价值发现和风险防控为目标的企业级数据治理与平台化整合。然而，证券公司传统的数据治理和质量提升工作往往集中于结构化数据，对于大模型所需要的多模态数据往往关注不高，但目前多模态数据集的数据质量的多维度挑战已成为制约其人工智能应用效能提升的最主要障碍，也从反面论证了推进数据集标准化建设的紧迫性。

证券公司在高质量数据集方面的核心工作，首先体现在对投资研究相关业务上。头部券商早已不满足于直接使用原始的市场行情和公司财报数据。它们构建了复杂的数据预处理与分析流水线。对于采购的基础金融数据会进行严格的数据校验与清洗，处理数据异常、计算错误，字段缺失等问题。

其次，在内部数据资产的治理与平台化服务方面，证券公司正努力通过建设数据中台来提升数据的可用性与质量。这项工作涉及对核心业务系统产生的数据进行标准化和资产化。具体工作包括：制定非结构化数据管理规范，明确大模型知识入库流程；制定企业统一的客户、产品、账户等主数据标准；建立跨业务条线的数据质量监控体系，对关键业务指标的准确性、及时性进行常态化核查与督办；通过数据仓库或数据湖技术，将经过整合和清洗的数据，以主题域的形式提供给业务部门。它从实践层面证明，建立并强制执行企业内部的统一数据标准和质量管控流程，是确保公司级数据分析与 AI 应用结果可信的基石，能有效避免因数据源不一致、口径冲突导致的决策失误或监管合规风险。

证券公司在实践中暴露出的数据质量痛点，深刻揭示了行业级标准的缺失所带来的普遍困境。第一，数据来源的多样性与标准不一问题突出。外部数据供应商各有其数

据格式和定义，内部系统数据模型老旧且不统一，导致数据整合成本高企。第二，数据标注的专业门槛与成本极高。在智能投研、合规审查、监管知识库等场景中，需要对复杂的金融非结构化进行信息抽取，这要求标注人员具备深厚的金融知识，导致标注工作难以规模化，且质量波动大，直接影响模型性能上限。第三，数据合规要求限制了数据的完整性与融合度。对数据的融合与应用需格外谨慎，可能导致用于构建训练模型的数据集存在特征缺失，影响模型的全面性与准确性。这些普遍存在的挑战清晰地表明，券商各自为战、内部标准不一，无法从根本上解决数据质量的系统性难题。因此，当前券商在数据治理和中台建设上的所有努力，都在共同呼唤一套行业公认的、涵盖数据采集、预处理、质量评估、标注规范乃至合规红线的高质量数据集通用标准。这套标准能大幅降低券商的数据整合与治理成本，提升 AI 应用的基线水平，为整个行业数字化转型的深化提供关键支撑。

（4） 中电金信在高质量数据集的推进工作调研

中电金信构建了高质量数据集的四级分层管理体系，数据来源广泛，涵盖外部公开数据、内部业务积累数据及仿真构建数据，具备全面、丰富、高质量、时效性强的特点。L1 为合法公开金融信息，通过日常收集、官网爬取及自动生成获取；L2 是经脱敏的行业共性数据，由各业务部

门提交并结合专家生成数据对；L3 为企业核心商业秘密与业务方法论，来自部门归纳整理；L4 是受严格法规约束的客户信息，采用脱敏或联邦学习技术，各层级按特征与范围规范管理。

中电金信高质量数据集标注覆盖 SFT、CoT、Agent 等多种类型，以系统化提升数据集及模型性能。其中预训练数据占比 70%，是构建知识与基础推理能力的基石；SFT 数据占 20%，助力模型适应特定领域并对齐人类价值观；CoT 与 RL 强化数据各占 2.5%，分别提升模型复杂推理与动态问题解决能力。业务覆盖银行大部分子领域，如营销管理、风险管理等，但在金融市场、投资银行等领域存在空缺，计划与银行合作完善。同时投入 1.02 亿元，由二十余个部门、三千余人专业团队及多领域专家协作，保障建设推进。

中电金信金融高质量数据集建设成果显著，1.0 版本拥有 1TB 预训练数据与 98 万条微调数据，涵盖银行、信贷、保险等领域，有效支撑金融客服问答、信贷尽调报告生成等场景的模型训练。2.0 版本基于市场需求与业务积累，参考 890 个全域场景，精选 152 个场景构建，数据集规模达 2TB，包含 30 万条高质量数据指令与 200 万条指令数据总数。细分领域中，客户管理、营销支持、信贷业务等均有对应场景数量与高质量数据指令数，数据来源

涵盖四级分层数据，经业务专家梳理与手工标注，部分结合仿真造数，进一步提升数据集实用性与专业性。

中电金信高质量数据集采用数据寻源、数据采集、知识工程、数据标注、数据对齐反馈等 5 步进行，通过建设数据资源地图组织、数据清晰过滤、智能标注、难例标注、数据对齐、数据反馈等关键步骤组织进行。

当前高质量数据集建设已在理论、产业、业务应用等层面形成多主体协同推进格局，虽在标准统一、部分领域覆盖等方面存在待解问题，但已为行业发展及通用标准构建积累丰富实践经验。

### （二） 证券公司高质量数据集框架

完成高质量数据核心建设框架五大步骤，包括高质量数据寻源接入、高质量数据预处理与转换阶段、高质量数据标注建模阶段、高质量数据集应用服务阶段、高质量数据集评估管理阶段等。

（1）高质量数据集寻源接入阶段

在高质量数据集建设的初始阶段，准确界定采集寻源范围是确保后续工作成效的基础。这一阶段需要系统性地规划数据资产的边界，明确需要采集的数据类型与内容形式。当前工作主要涵盖结构化与非结构化两大类别：结构化数据包括数据字典、业务指标、用户标签及各类表格数据，这些数据具有明确的字段定义和关系模型，是量化分析与模型训练

的核心原料；非结构化数据则包括长文本、图片、音频、视频等，这些数据蕴含丰富的语义信息和场景特征，对提升模型的理解与泛化能力至关重要。通过对这些数据类型的全面梳理，形成了完整的数据资源地图，为后续的精准采集与标准制定提供了清晰的目标范围。

在明确采集范围后，需要建立相应的采集标准体系来保障数据质量。这项工作正针对不同数据对象的特性，从多个维度展开标准化建设。名称标准确保同一数据对象在不同系统中的标识一致性；定义标准明确每个数据对象的业务含义和范畴边界；口径标准规定数据的计算逻辑和统计方法，确保数据可比性；结构标准定义数据的组织格式和关系模式；参数标准明确技术参数的有效范围和取值规范；编码标准统一各类代码值的表示方式。这些标准相互关联、彼此支撑，共同构成了完整的数据采集规范框架，确保从采集源头就建立起统一的质量基准。

在实际标准实施过程中，需要根据不同数据对象的特性采取差异化的标准制定策略。对于结构化数据，重点规范字段命名、数据类型、值域范围、约束条件等要素，建立严格的数据模型管理机制。对于文本类数据，需要制定内容格式、字符编码、语言规范、质量要求等标准，特别对长文本还需规定分段规则、标注方法等。对于多媒体数据，图片需规范格式、分辨率、色彩空间等参数；音频需明确采样率、位深

度、声道数等技术指标；视频需规定编码格式、帧率、分辨率等参数要求。这种分类别、分层级的标准实施方式，既确保了标准要求的针对性，又保持了整体框架的一致性，使得各类数据在采集阶段就能满足后续加工处理的质最要求。

（2）高质量数据集预处理与转换阶段

预处理与转换是高质量数据集建设的核心环节，其目的在于将原始、粗糙的源数据转化为洁净、规整、适于模型消费的优质数据。这一阶段首先需要建立一套系统化的机制框架，提供顶层指导。数据清洗机制是这一流程的起点，目的是系统性处理数据中的缺失值、异常值、格式不一致及逻辑冲突等问题，恢复数据的完整性与准确性。为进一步提升数据信噪比，数据降噪机制专注于识别并剔除数据中无意义的干扰信息。为防止数据集虚胖及过拟合，数据防注水机制通过技术手段杜绝无效、重复或高度相似数据的混入，保证数据集的纯粹性与有效性。最后，数据抽样机制为大规模数据集的处理与评估提供了可行性方案，它确保无论是用于模型训练还是效果验证，所抽取的样本都能无偏地代表整体数据分布。这四大机制共同构成了一个环环相扣的质量控制闭环，为生产高标准数据集奠定了坚实的流程基础。

在标准机制框架下，针对不同预处理步骤的技术研究是提升数据质量的关键驱动力。在数据清洗与集成环节，精确与模糊匹配研究至关重要。它既能够进行严格的唯一标识符

匹配，以整合来自不同源头的同一实体数据，更需要发展高效的模糊匹配算法，以处理名称不完全一致、存在错别字或别名的情况，确保数据关联的准确性。面对海量多媒体数据，图像与文本去重研究成为数据防注水的核心技术。它通过感知哈希、特征向量相似度计算及嵌入模型比对等先进方法，精准识别并剔除内容重复或高度近似的图片与文本，极大提升数据集的多样性和有效性。此外，规则过滤与数据增强研究构成了辩证统一的两个方面。一方面，基于业务逻辑与安全合规要求，通过规则过滤主动剔除敏感、低质或带有偏见的数据；另一方面，为了弥补数据分布的不平衡或样本量的不足，通过数据增强技术，如图像的旋转裁剪、文本的同义词替换与句式改写，在保证语义不变的前提下，合法地扩充高质量训练样本，增强模型的鲁棒性。

为确保预处理与转换工作的科学性、一致性与可复现性，必须对其核心内容进行标准化，并持续推进自动化能力建设。标准化工作主要涵盖三个层面：算法标准规定了在特定场景下应优先选用的处理算法及其参数配置。规则标准则将业务知识、合规要求固化为可执行的判断逻辑。知识标准则更进一步，它构建了支撑整个预处理过程的领域知识体系。在上述标准的基础上，数据质量自动增强研究致力于将分散的处理流程整合成智能化的流水线，利用机器学习模型自动检测数据异常、推荐清洗策略甚至执行增强操作，从而将人力从

繁复的劳动中解放出来，实现数据处理效率与质量稳定性的跨越式提升。

预处理与转换工作是一套可管理、可度量、可持续优化的工业化生产标准，为构建真正高质量的数据集提供核心保障。

（3）高质量数据集标注建模阶段

在高质量数据集的标注建模阶段，数据标注管理流程演进为一个以技术为驱动、人机协同的精细化过程。通过构建一个闭环、高效、质量可控的标注生产体系。其起点是任务定义与规范制定，即根据模型训练目标，明确标注对象、标注维度及详细的标注指南，实现自动化标注与辅助标注工具的深度集成。自动化标注利用成熟的预训练模型或规则引擎，对原始数据进行批量预处理和初步标注，极大提升了基础标注任务的效率，将人力资源集中于更具价值的复杂任务。辅助标注工具则通过交互式界面、智能预填充、实时一致性检查等功能，为标注员提供强力支持，有效降低主观误差、提升标注速度与一致性。

建立系统的难例标准并开展专业化标注成为提升数据集质量的关键。证券行业作为高专业门槛领域，问答对的构建要求标注者具备领域知识，判断答案的准确性与完整性。指标原子化标注要求将复杂的业务指标拆解为不可再分的原子要素并进行独立标注，这些为模型的组合推理能力提供

了训练基础。标签精细化分类标注构建层次化、多维度的标签体系，以刻画数据的细微差别。针对这些复杂任务，可建设专门的辅助标注工具，例如集成领域知识图谱的智能提示系统、用于指标拆解的可视化配置界面等，以降低标注难度，保证专业性要求的落地。

为确保分析建模产出的数据集具备高置信度，需要在多个关键节点建立严格的质量修正模型。预训练筛选标准旨在为自动化标注或主动学习筛选出最具价值的训练样本，定义数据多样性、代表性及难例比例的量化指标。数据置信筛选标准关注评估和过滤自动化标注或众包标注结果的可靠性，通过模型自置信度、多个模型预测的一致性等方式设定过滤阈值。后处理标准规定标注结果所需的规范化处理。校验迭代标准建立了人机协同的质检流程，明确抽检比例、异议仲裁机制以及基于质检反馈的标注规范迭代优化机制。数据增强标准则在需要扩充数据时，为确保增强后数据的真实性与语义一致性提供准则，防止引入虚假模式。这些研究构成了一个从数据输入、处理到产出的全链路质量保障模型体系。

（4）高质量数据集应用服务阶段

在数据应用与服务阶段，高质量数据集的合成与验证将来自不同源头、经过预处理和标注的多个子数据集，按照统一的规范和逻辑进行整合，可以形成面对不同各行业、不同场景、不同量级的的复合数据集。解决数据源之间的语义对

齐、时空对齐和尺度对齐等复杂问题。同时，进行数据集验证及性能检测工作。通过一套多维度、量化的评估指标体系，对合成后的数据集进行校验。内容包括数据整体的完整性、一致性、准确性和时效性；数据分布的均衡性与无偏性；通过抽样人工核查对关键数据项进行精度验证。性能检测侧重于数据集在机器学习任务上的表现，通过基准测试评估其对于模型训练效果的提升程度。合成与验证共同构成了数据集交付应用前的质量闭环，是构建可信数据服务的首要环节。

在数据集有效赋业务方和开发者，需定制灵活、高效的数据集服务模式。第一，提供原始的标准化数据集包下载，满足用户对数据进行深度定制和本地化处理的需求。第二，构建在线数据 API 服务，通过接口实时调用特定数据子集或特征，对实时性要求高的模型生产环境。第三，提供嵌入工作流的特征平台服务，将高质量数据集加工后的特征直接推送到模型训练和推理流水线中。服务模式确认后，同步进行数据集应用场景设计，定义高质量数据集从资产到价值的转化路径。

为确保数据应用与服务过程的规范性、可重复性和结果的可信度，建立全链路的标准化体系至关重要。这一体系覆盖从数据集生成到最终消亡的全生命周期。数据集标准化合成规范数据合并、连接、聚合等操作的具体流程、算法和输出格式。冲突检测与消解标准则提供了当不同来源的数据出

现矛盾时，进行识别、判断和解决的决策规则与优先级策略。数据集版本标准管理规定了版本的命名规则、变更日志的记录规范、回溯机制以及不同版本间的兼容性声明。Benchmark检验标准定义了用于衡量数据集性能的基准任务、评估指标、测试环境以及性能基线，确保评估结果的客观可比性。最后，应用场景发布标准明确了在特定场景下使用数据集时应遵循的伦理指南、合规性要求和性能预期。这些标准共同构成了数据服务质量的坚实保障。

确保高质量数据集在应用服务阶段能够被一致地管理、可靠地交付和合规地使用，最终实现其业务价值的最大化。

（五）、高质量数据集评估反馈管理阶段

在高质量数据集建设的分析评估阶段，评估反馈管理构成确保数据集持续优化的核心闭环。

该阶段系统性地从四个关键维度展开评估管理工作：

1、基础质量维度聚焦数据内在属性。

✓ 准确性：包括标注正确率、数据错误率等；

✓ 一致性：包括如逻辑冲突检测、跨源数据比对等；

✓ 完整性：包括如字段填充率、覆盖率统计等；

✓ 时效性：包括如数据更新周期、有效时间戳比例等；

2、安全评估维度重点关注合规与风险控制，涉及数据脱敏有效性、隐私泄露风险评估、敏感信息识别准确率及合规性审计，确保数据集符合相关法律法规与伦理要求；

3、内容评估维度深入数据语义层面，考察信息密度、内容多样性、偏见指数及知识准确性，防止数据内容出现系统性偏差或知识谬误；

4、应用质量维度则直接验证数据集在真实场景中的价值，通过模型训练耗时、资源消耗、收敛速度等验证效能性。模型通用性和表征对齐性等。

高质量数据集建设需经历寻源接入、预处理转换、分析建模、应用服务及评估反馈五大核心步骤，形成从数据采集到价值验证的闭环管理体系。这一过程中，标准化工作具有重要意义：它通过统一数据定义、处理流程、质量指标和应用规范，确保数据集在一致性、可靠性及合规性方面达到可验证的高标准。标准化不仅提升了数据生产和使用的效率，降低了协作成本，更为数据要素的可信流通与规模化应用奠定了坚实基础，是释放数据价值、驱动人工智能产业高质量发展的关键支撑。

### （三） 高质量数据集建设标准

#### （1） 高质量数据集寻源阶段数据标准

若要将寻源接入工作全面标准化，需要建立包含以下核心信息项的标准规范体系：首先是源数据标识信息，包括数据源名称、编码、版本、提供方等基础元数据；其次是技术规范信息，涵盖数据格式、编码方式、接口协议、

更新频率等技术要求；第三是质量基准信息，明确完整性、准确性、一致性、时效性等质量指标的阈值要求；第四是合规安全信息，规定数据分类分级、隐私保护、权限控制等安全管理要求；第五是管理流程信息，包括采集计划、验收标准、问题处理等流程规范。这些标准信息项共同构成了寻源接入阶段的标准化框架，通过明确的规范要求和可操作的执行细则，确保数据从源头就具备高质量特性，为后续的数据治理和应用奠定坚实基础。

（2） 高质量数据集预处理和转化阶段

若要将预处理与转换阶段全面标准化，这套规范至少应包含以下核心内容：

1、流程控制标准信息项：明确每个处理步骤，如清洗、去重、增强的先后顺序、触发条件、输入与输出数据的规格定义，以及各环节的质量检查点与验收标准。

2、技术与算法标准信息项：详细记录各步骤所采用的特定算法、模型或工具的名称、版本号及其关键参数配置。同时，需规范算法效果的评价指标，如去重准确率与召回率、数据增强的保真度，及其目标值。

3、规则与知识标准信息项：对于基于规则的处理，需明文规定所有规则的逻辑条件、判定阈值和执行动作。对于依赖知识的处理，需指定所引用的知识库或标准术语集的名称、版本及其使用方法。

4、质量度量与元数据标准信息项：定义预处理后数据所需达到的质量维度，包括完整性、准确性、一致性、唯一性、时效性及其量化的度量方法和目标值。同时，强制要求记录完整的处理元数据，包括数据血缘、处理日志、参数变更记录等，确保处理过程的全链路可追溯与可审计。

（3）高质量数据集标注建模阶段

标注建模工作全面标准化，也是一组覆盖流程、工具、质量和管理的标准信息项。具体而言，关键标准信息项包括：

1、流程管理标准信息项：明确标注任务的生命周期管理规范，包括任务创建、标注员资质要求、任务分配机制、进度监控指标、以及成果交付与验收流程的标准定义。

2、工具与接口标准信息项：规定辅助标注工具和自动化标注服务必须支持的核心功能、性能指标、输入输出数据格式以及人机交互接口的通用规范，确保工具间的互操作性和标注体验的一致性。

3、质量度量与置信度标准信息项：定义各环节的质量评估指标、计算方法、目标阈值以及标注一致性的量化度量方式。同时，规范置信度分数的计算方法和基于置信度的分级处理规则。

4、难例与领域知识标准信息项：针对不同行业，规范难例的定义、分类体系、标注方法论以及所引用的证券知识库的版本与使用规范。

5、元数据与溯源标准信息项：强制记录每次标注任务的完整元数据，包括标注工具版本、标注员信息、所用标准版本、质检记录、迭代历史等，实现从原始数据到最终标注结果的全链路溯源与审计。通过规范上述信息项，可使分析建模阶段从高度依赖个人经验，转变为可复制、可评估、可持续优化的工业化生产过程。

（4）高质量数据集阶应用服务阶段

数据应用与服务阶段全面标准化，需对关键活动定义明确、可操作的标准信息项。标准应至少包含以下核心内容：

1、合成与溯源信息项：必须规范记录数据集合成所依据的蓝图或配方，包括所有输入数据集的版本标识、使用的合成算法与关键参数、以及合成过程中执行的所有转换操作的完整日志，确保合成结果的可复现性和全链路溯源。

2、质量与验证信息项：需明确定义数据集出厂质量报告必须包含的量化指标、内部质量检验的方法与结果、以及第三方验证报告的链接或索引。规范性能基准测试的详细环境和结果数据。

3、服务与版本信息项：必须标准化数据集服务的元数据，包括服务模式（API/包下载）、访问接口规范、数据格式说明、

使用许可协议和服务级别协议。在版本管理上，严格规定版本的唯一标识符、版本变更说明的模板、以及版本的生命周期状态。

4、场景与合规信息项：对于推荐的应用场景，规范其场景描述、适用的业务问题、预期的价值收益、已知的局限性以及在该场景下使用的具体配置建议或最佳实践。必须包含数据安全与合规性声明，明确数据分类等级、隐私保护措施和相关的法规政策遵从性说明。

（5）高质量数据集评估反馈阶段

完成高质量数据集评估细化，参考信通院部分成果，设计评估项 4 大项 12 小项。从基础质量维度、安全评估维度、内容评估维、应用质量维度进行高质量数据评估管理工作。其中设计的标准内容包括准确性、一致性、完整性、时效性、效能性、模型通用性、表征对齐性等方面内容及具体指标设计。例如模型训练效能评估指标包括训练耗时、收敛速度、泛化误差进行综合评价。

为将此评估体系标准化，需规范以下关键信息项：

一是质量度量标准信息项，明确定义各维度指标的计算公式、测量方法、评估频率及合格阈值；

二是评估流程标准信息项，规范评估环境配置、数据抽样方法、基准模型选择及对比实验设计；

三是元数据标准信息项，要求完整记录每次评估的版本

信息、环境参数、原始结果及分析结论；

四是反馈改进标准信息项，建立评估结果与数据迭代的联动机制，明确问题分类、优先级判定及整改验证要求。通过这些标准化建设。

## 四、研究结论与建议

### （一）课题总结

一是完成高质量数据集发展现状调研。通过市场调研、资料分析与专家访谈相结合的方法，系统梳理了科技公司、证券公司及科研机构在高质量数据集建设与应用方面的现状，输入了中电金信研究现状，明确了当前高质量数据集在金融领域的发展水平与应用瓶颈。研究发现，科研机构聚焦于前沿数据方法与基准数据集的构建，大型科技企业依托规模化数据基础设施实现数据生产与价值闭环，证券公司则从业务需求出发推进数据治理与平台化整合，但在高数据标准统一、专业标注能力及合规应用等方面仍面临普遍挑战。在此基础上，需要从数据寻源、预处理、标注建模、应用服务与评估反馈的高质量数据集建设五阶段框架，强调了全流程标准化在提升数据一致性、可靠性、安全性及业务价值方面的重要作用。

二是完成高质量数据集标准设计。构建了一套贯穿高质量数据集全生命周期的综合性标准体系，通过系统性地定义

数据寻源、预处理、标注建模、应用服务及评估反馈各环节的核心标准信息项，确立了覆盖数据标识、技术规范、质量基准、流程控制、算法参数、工具接口、合规安全与溯源元数据等关键要素的统一规范。强调以量化指标和可操作细则确保数据从采集到应用的全流程质量可控、过程可溯与结果可信，推动数据生产从依赖个人经验向标准化、工业化模式的根本转变，为构建一致、可靠、高效的高质量数据集提供了完整的理论框架与实践指南，为行业数据治理与人工智能产业化发展奠定了坚实基础。

三是编制高质量数据集标准研究报告。

基于高质量数据集标准现状调研和设计成果，在证标委WG22工作组指导下，根据《标准化工作导则第1部分：标准化文件的结构和起草规则》（GB/T 1.1—2020）以及金融行业标准化有关要求，基于课题研究成果，编制完成《证券公司高质量数据集标准草案》。

（二）课题展望

证券行业高质量数据集标准化工作任重道远，面对科技和业务的不断演变，课题组将持续投入和改进，推动证券公司高质量数据集标准化工作不断发展，主要有以下四个方面：

一是持续优化，将本课题研究成果进一步凝练优化，提供给行业机构参考使用，如对高质量数据集建设流程梳理方法进行总结提升、标准成果的完善补充等；

二是成果细化，高质量数据集的建设需要成熟的运营指标、质量指标、安全指标、工具指标等全方面的标准指引，需要在建设过程中结合各大行业机构的探索实践不断细化和丰富，逐步构建起完整的行业高质量数据集标准体系，指导高质量数据集成果落地；

三是成果推广，推动将高质量数据集及标准相关成果在其他证券公司试点使用，促进全行业人工智能能力升级；

四是标准落地，建议对证券公司高质量数据集建设进行增补扩展，建立高质量数据集标准，工作计划如下：

（1）完善标准草案，形成工作组讨论稿。组建标准起草工作组，对高质量数据集标准草案进行研讨，优化完善标准草案内容，同步向证标委提出行业标准立项建议（时间计划：2026 年 1 月至 6 月）；

（2）标准征求意见。邀请行业机构开展标准征求意见工作，对高质量数据集标准进行修订完善（时间计划：2026 年 7 月至 12 月）；

（3）完成标准制订各阶段工作。配合证标委秘书处完成标准送审、报批及发布阶段的各项工作，形成证券公司高质量数据集标准发布版（时间计划：2027 年 1 月至 2027 年 12 月）。

| | | | |
|---|---|---|---|
| 课题负责人 | 杜啸争 | 中电金信研究院、商业分析事业部 | 副院长、总经理 |
| 课题成员 | 杜啸争 | 中电金信研究院、商业分析事业部 | 副院长、总经理 |
| | 李楠 | 中电金信研究院数据工程实验室 | 主任 |
| | 李娜 | 中电金信商业分析事业部 | 首席咨询专家 |
| | 革远平 | 中电金信软件有限公司 | 高级副总裁&证券行业部总经理 |
| | 马野 | 中电金信商业分析事业部交付部 | 部门经理 |
| | 罗嗣汉 | 中电金信商业分析事业部新兴交付部 | 部门经理 |
| | 戴永恒 | 中电金信研究院数据工程实验室 | 数据资产平台产品负责人 |
| | 李静 | 招商证券综合业务开发部 | 总经理 |
| | 王超 | 招商证券综合业务开发部 | 大数据团队负责人 |
| | 左银康 | 国信证券技术管理部 | 数据治理组组长 |