主题：“资本市场注册制下信息披露审核与监管关键技术研究”项目增刊

交易技术前沿

2025 年 第 4 期 总第 64 期

ITRDC 证券信息技术研究发展中心（上海）

国家重点研发计划“资本市场注册制下信息披露审核与监管关键技术研究”项目

所属专项：社会治理与智慧社会科技支撑（平安中国）

专项主责单位：国家发展和改革委员会

项目管理专业机构：中国21世纪议程管理中心

推荐单位：中国证券监督管理委员会

项目牵头承担单位：上交所技术有限责任公司

项目参与单位：北京邮电大学、华东师范大学、复旦大学、同济大学、上证所信息网络有限公司、华泰证券股份有限公司、海通证券股份有限公司（现国泰海通证券股份有限公司）、汇添富基金管理股份有限公司、北京明朝万达科技股份有限公司

执行期限：2021年12月 至 2024年11月

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 4 期 总第 64 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

序言 1

资本市场在经济和金融运行中具有牵一发而动全身的作用。党的十八大以来，以习近平同志为核心的党中央高度重视资本市场发展，就注册制改革、提高上市公司质量、加强资本市场风险防控等作出一系列重大决策部署，为新时代新征程中国资本市场改革发展事业指明了前进方向，提供了根本遵循。

与此同时，新一轮科技革命和产业变革加速演进。党的二十大报告明确提出“加快实施创新驱动发展战略”，这一战略部署不仅为经济社会发展指明了方向，更深刻揭示了科技创新在破解发展难题、塑造发展新动能中的核心地位。我国资本市场诞生三十多年来，信息技术有效促进了金融普惠发展、提高了信息传播效率、创新监管与服务模式，对我国资本市场在短时间内实现跨越式发展，实现投资者账户数量世界第一、市场规模世界第二，起到了无法替代的关键作用。另一方面，近年来随着市场快速演变与金融科技快速迭代，金融风险的隐蔽性、交叉性、传染性显著增强，资本市场传统的监管模式与风险防范手段越来越难适应复杂多变的形势，迫切需要以科技筑牢金融安全防线，用创新赋能监管升级，为市场安全平稳运行和金融强国建设保驾护航。

国家重点研发计划作为我国当前最高层次的科技计划体系之一，是国家立足改革发展全局，聚焦“四个面向”，整合产学研优势资源，加强科技创新的重大战略性部署，为各行业、各领域借助科技手段破解发展难题提供了国家级的科研平台支撑。其中，“社会治理与智慧社会科技支撑（平安中国）”重点专项面向社会安全、智慧司法、金融监管等领域重大需求，构建社会治理与智慧社会理论体系，加强关键共性技术和专用装备研发，促进社会治理体系和治理能力现代化，服务更高水平“平安中国”建设。

2021 年，上交所技术公司牵头承担了“社会治理与智慧社会科技支撑（平安中国）”金融监管领域首批“揭榜挂帅”项目“资本市场注册制下信息披露审核与监管关键技术研究”。作为“社会治理与智慧社会科技支撑（平安中国）”的总体专家组成员，我全程见证了其立项、实施到结项的整个过程。注册制以信息披露为核心，对监管与审核的时效性、专业性和精准性提出更高要求，项目主要针对监管实践中碰到的监管信息难获取、风险问题难识别、信披审核难问准“三难”问题进行攻关，旨在以科技突破发展瓶颈，有效助力重大改革落地，项目立项时机可谓恰到好处。项目实施紧扣证券交易所的“上市审核”和“公司监管”业务场景，研发并应用自然语言处理、大数据分析、知识图谱等多项先进技术，形成信息披露文档智能核验、舆情风险动态识别、市场异常交易和异常波动实时预警、企业科创属性可信评价等重要成果，在交易所和证券、基金公司的实际业务中得到应用和验证，取得良好成效。总体而言，该项目响应了中央金融工作会议、新“国九条”有关要求，为资本市场防风险、强监管、促发展提供了重要的科技支撑，体现了国家重点研发计划为国选题、为国立项、为国攻关的战略导向。

未来，全球竞争与博弈日趋激烈，资本市场深化改革也进入深水区。以数字化、智能化推动资本市场高质量发展将是一个长期过程，需要产学研用各方携手并进，发挥合力协同创新。期望本期《交易技术前沿》专刊能成为一个窗口，让更多行业同仁了解项目成果，共同探讨技术赋能信息披露审核与市场监管的未来路径，为建设与金融强国相适应的高质量资本市场更多智慧与力量。

柴洪峰 *

2025 年 7 月 10 日

* 中国工程院院士、复旦大学金融科技研究院院长，国家重点研发计划“社会治理与智慧社会科技支撑（平安中国）”重点专项总体组专家。

序言 2

党的十八大以来，以习近平同志为核心的党中央始终高度重视科技创新，强调企业发挥科技创新中的关键作用。习近平总书记在全国科技大会上发表重要讲话指出要强化企业科技创新主体地位，支持企业牵头或参与国家重大科技项目。“十四五”以来，在证监会《证券期货业科技发展“十四五”规划》有力指引下，证券期货行业机构积极发挥行业科研主力军的作用，通过产学研合作，积极承接国家重大科技计划，开展资本市场金融科技创新试点，多措并举提升金融科技水平，促进资本市场高质量发展。

为把握数字化发展机遇，赋能资本市场强监管、防风险、促高质量发展目标，突破资本市场改革发展面临的难点堵点问题，在证监会科技司的大力推荐和指导下，2021 年，上交所技术公司牵头高等院校、证券基金公司和金融科技企业，承接国家重点研发计划“社会治理与智慧社会科技支撑（平安中国）”重点专项“资本市场注册制下信息披露审核与监管关键技术研究”项目。项目重点围绕注册制下信息披露审核与监管实践中的“监管信息难获取、风险问题难识别、信披审核难问准”三难业务痛点，提炼科学问题、攻关关键技术、形成解决办法。通过研发成果的应用示范与场景验证，项目实现预期建设目标，在提升信息披露审核与市场监管质效方面取得显著成效。

一是智能辅助核验提升上市公司信息披露质量。注册制以信息披露为核心，招股书、年报等信息披露文档篇幅长、内容密集，规范性和专业性强。实际工作中上市公司信息披露低级错误时有发生，包括文字编辑错误、数据填报错误、内容疏忽遗漏等，容易引发负面舆情、动摇投资者信心，影响资本市场整体形象。同时，伴随着注册制改革的不断深化，证券交易所的信息披露监管审核工作面临量增、质升双重挑战，传统人工校对的局限性日益凸显。

项目通过研发富格式非结构化金融文档智能理解、基于大模型的信披内容合规性检测等关键技术，并融合金融领域知识，打造信息披露文档智能审核系统，破解人工审核校验的效率瓶颈和差错风险。项目实现审核要素自动识别准确率 98.64%、召回率 98.07%，达到行业领先水平。在项目应用示范期间，系统日均抽取能力达 2300 篇，形成 152 类超 250 万条结构化信披数据，向证监会监管大数据仓库、上海大数据中心和 8 家证券基金经营机构免费共享，降低行业用数成本。系统同时支持智能检测披露内容缺失、数据勾稽偏差、表述不一致、错别字等低级错误，将传统数百页的文档审核从数小时压缩至约 6 分钟，误报率低于 3%，有效提升了监管审核的效率和准确性。相关功能已向沪市 2000 余家上市公司提供服务，在 2025 年 4 月下旬年报披露高峰期间共审核 3.8 万份文件，审核结果 19.6 万个，推动信息披露从“事后抽查”向“事前预防、事中干预”转型，逐步形成“科技驱动——市场自律——精准监管”的新型信披治理生态。

二是语义分析助力精准识别市场舆情风险事件。金融舆情具有敏感度高、传播速度快等特点，一条突发舆情可能在几分钟内引发市场大幅波动。国家网信办、证监会等多部门联合部署强化金融信息管理、打击股市“小作文”乱象，凸显了新形势下舆情监测在维护市场秩序、保护投资者权益中的重要性。然而，舆情事件具有强时效性和动态演化特征，新事件类型、传播模式及风险因子不断涌现，传统依赖大量静态标注数据的事件检测方法在金融监管场景下面临准确率和鲁棒性欠佳的不足。

项目通过改进小样本事件检测模型并提出一种新的零样本检测技术，构建资本市场舆情风险识别预警系统，解决模型训练样本少、数据呈长尾分布等行业共性难题。系统日均采集分析金融市场舆情超 6 万条、推送舆情预警超 1.9 万条，大幅提高复杂场景下金融舆情事件识别的准确性和泛化性，对比国际同类算法，取得 10% 以上的整体性能提升，针对未见过的事件类型，提升幅度更达 15%-20%。同时，系统改变了传统舆情预警的研发范式，支持业务人员以自助方式配置化实现新舆情事件类型的识别需求，在华泰证券应用示范过程中，新舆情预警需求实现耗时从 1-2 月缩短至 5-10 分钟完成，节省开发测试人力超 600 人月。

三是公司画像赋能上市公司监管与企业科创属性评价。上市公司监管与服务工作含非标业务多、人力投入大、社会关注度高，深度应用科技手段是提升监管效能与服务质量的必然选择。项目围绕财务异常风险识别和企业科创属性可信评价等日常业务中的典型难点问题，结合上交所公司画像系统建设，为监管与服务工作带来多维度的效能提升与模式优化。

财务造假严重扰乱资本市场秩序，通常具备手段隐蔽、模式复杂、动机多样、层级多、链条长等特征。传统财务异常测算方法存在未能充分整合舆情等非财务信息、难以区分财务舞弊行为的异质性等缺陷，且专家循证成本高，线索发现效率也有待提升。项目提出基于同群效应和集成学习算法的全新财务舞弊识别框架，对已披露的财务欺诈案例具有 98.2% 的高召回率，并通过计算“净检测值”验证了相关方法对比国际同类技术具有更高的经济和实用价值，相关成果被管理学国际顶刊《Management Science》录用。在上交所对科创板上市公司 2023 年年报的审核工作中，系统对合计 114 封问询函涉及的上市公司共提示 603 条预警线索，总体采纳率约 45%，有效提升审核问询效率，支持监管“长牙带刺”、有棱有角。

科创板坚持“硬科技”定位，但监管审核人员大多不具备深厚技术背景，对于细分领域的企业科创属性较难快速作出专业判断。国际上常用的企业评级方法忽视中国特色创新生态，与国家战略和社会价值导向偏离。针对上述问题，项目利用基于预训练大模型的信息检索等关键技术，融合科创板监管审核经验，从政策契合度、知识产权、产业链地位等多维度构建了 266 项科创属性指标的计算方法和评分逻辑，形成国内首个针对科创板的“硬科技”评价体系。同时，项目通过构建企业关系、产业链和供应链知识图谱，实现关联探寻、关系脑图、风险图谱、股权穿透、集团系谱、关系查询等重要功能，辅助审核人员对拟上市企业“问清楚”，中介机构“核清楚”，把好“入口关”。

下一步，上交所拟以本次国家重点研发计划项目工作为契机，持续做好研究成果的转化应用与行业推广。一是拓展应用场景，将公司画像等成果向债券、基金、REITs 等更多金融产品赋能，并向更多行业机构推广试用。二是技术迭代优化，探索数字人等多模态大模型在监管审核与服务中的创新应用。三是深化产学研合作，与行业机构、科研院所和科技企业等建立常态化的研究合作与交流机制，加强前沿技术的研究与应用。同时，上交所也将继续强化科技赋能，不断探索人工智能、大数据等前沿技术在资本市场监管与服务领域的深度应用，贯彻落实资本市场新“国九条”，扎实做好金融“五篇大文章”，着力打造科技领先的世界一流交易所，为提升服务实体经济质效、助力金融强国建设作出更大贡献。

王泊 *

2025 年 7 月 10 日

* 上海证券交易所党委委员、副总经理，国家重点研发计划“社会治理与智慧社会科技支撑（平安中国）”重点专项“资本市场注册制下信息披露审核与监管关键技术研究”项目负责人。

刊首语

为贯彻落实中央有关决策部署，推动资本市场核心机构主导的产学研融通创新，在科技部、发改委和 21 世纪中心的大力支持下，在证监会科技监管司的推荐和指导下，上交所技术公司联合北京邮电大学、华东师范大学、复旦大学、同济大学、上证信息公司、华泰证券、海通证券（现国泰海通）、汇添富基金和北京明朝万达共 10 家单位，承接了“十四五”国家重点研发计划“社会治理与智慧社会科技支撑（平安中国）”重点专项揭榜挂帅项目“资本市场注册制下信息披露审核与监管关键技术研究”。

该项目由上交所王泊副总经理担任项目负责人，实施周期三年，自 2021 年 12 月起至 2024 年 11 月止，并于 2025 年 4 月顺利通过国家重点研发计划相关管理机构组织的专家结项评审。项目重点围绕注册制下证券交易所开展信息披露审核与监管实践中的“监管信息难获取、风险问题难识别、信披审核难问准”三难业务痛点，提炼科学问题、攻关关键技术、形成解决办法。通过研发成果的应用示范与场景验证，实现预期建设目标，在提升信息披露审核与市场监管质效方面取得显著成效。

本期《交易技术前沿》作为“资本市场注册制下信息披露审核与监管关键技术研究”项目专刊，收录了项目组部分核心成果论文或报告，以期推动成果交流共享，并为相关领域技术人员提供参考借鉴。其中：

“全景纵览”阐述了项目助力资本市场防风险、强监管、促发展总体情况，从全局视角概括了项目的核心成果及其应用成效，深入剖析了科技创新和数智化对资本市场及证券交易所“十五五”高质量发展的重要作用，归纳项目助力上交所做好金融五篇大文章的阶段性成果。

“前沿攻坚”聚焦项目的理论技术研究与创新，分别介绍了结合新闻大数据和同群效应的财务舞弊识别框架，非结构化文档智能理解与要素抽取，基于检索增强和大模型的财报分析，以及企业科创属性指标计算等关键技术。

“工程实践”重点关注研究成果如何转化为实际应用，涵盖金融文档智能理解能力测评，金融舆情风险识别与预警系统实现，融合金融舆情的证券市场态势分析，基于大模型技术的问询函自动生成，投资者视角下的企业科创属性评价技术应用，以及基于微服务与隐私计算技术的数据安全共享服务平台。

证券信息技术研究发展中心（上海）

2025 年 7 月 10 日

目录

01 全景纵览

02 前沿攻坚

03 工程实践

04 项目大事记

01 全景纵览

科技赋能防风险 创新聚力强监管 —“资本市场注册制下信息披露审核与监管关键技术研究”成果集萃

王泊、唐忆、黄越、徐广斌、陆伟
上海证券交易所

摘　要： 本文介绍了“资本市场注册制下信息披露审核与监管关键技术研究”项目取得的主要创新成果。研发信息披露文档智能审核系统，通过分层嵌套实体识别模型等技术，将招股说明书要素识别准确率提升至 98.64%，并在上海证券交易所部署，大幅降低人工修订率。构建资本市场舆情风险识别与处置系统，提出小样本和零样本检测技术，对比国际同类先进算法 F1 值显著提升，落地多家机构，实现高效舆情预警。创新企业科创属性和财务经营综合评价体系，科创属性指标数量增至 266 项，改进科创行业分类，研发财务异常识别方法，集成到公司画像系统，辅助监管审核与投资决策。基于大模型技术，研发动态词表生成技术，提升问询函生成质量与效率。项目成果在上交所等多家机构应用示范，提升监管审核质效，获得多方用户肯定。

关键字： 信息披露审核；舆情风险识别；财务异常识别；科创属性评价；监管系统

一、引言

“资本市场注册制下信息披露审核与监管关键技术研究”项目聚焦我国资本市场“防风险、强监管、促高质量发展”目标，以科技手段提升注册制下的企业上市审核与上市后持续监管的工作质效，更好助力新时代新征程的资本市场改革与发展。

上市审核方面，注册制以信息披露为核心，要求证券发行人真实、准确、完整地披露公司信息。信息披露文档是非结构化数据，无法被计算机直接分析处理，面对动辄数百页甚至上千页的招股说明书等信息披露文档，人工审核效率低、耗时长。项目通过打造信息披露文档智能审核系统，突破富格式金融长文档智能理解关键技术，大幅提升信息披露文档的审核要素识别与抽取准确率，基于抽取的结构化信披数据，自动完成基础的信披文档内容完备性、一致性审核。同时，科创板聚焦支持“硬科技”企业，而企业上市审核人员大多不具备技术背景，对于细分领域的企业科创属性较难作出专业判断。项目构建科创企业评价系统和产业链、供应链、企业关联关系图谱，多维度构建具有中国特色的科创属性评价指标，辅助科创板上市审核工作。

持续监管方面，注册制强调“放管结合”，强化股市风险研判，增强资本市场内在稳定性，同时在出口端严格退市制度，严厉打击各类违法违规行为。进入大数据时代，舆情对资本市场的影响日渐加大，金融舆情事件具有传播快、多样化、动态演变等特点，且突发的舆情事件经常是前所未见或在以往历史中只发生了少量几次，传统检测算法难以有效识别。项目通过研究小样本和零样本舆情事件检测技术，及时准确识别舆情中的风险事件和其涉及的企业实体，同时利用大语言模型技术自动生成问询函初稿，辅助监管人员及时开展监管处置，回应市场关切。其次，财务造假严重扰乱资本市场秩序、破坏市场生态，项目突破传统单一的财务异常测算方法难以区分财务舞弊行为的异质性、不能充分利用舆情等非财务数据等局限，利用集成学习、人机协同方式优化财务舞弊识别模型，准确提示财务异常线索。另一方面，企业上市后的交易监管也是监管机构的工作重心之一，随着投资者数量和交易量的快速增长，对于拉抬打压、虚假申报等异常交易的实时预警计算要求也越来越高。项目通过建设完善安全高效、自主可控的交易监管系统，及时识别预警异常交易线索，并结合信息技术和金融学理论研究程序化交易等实践中遇到的新问题、新风险。

最后，项目研究成果在用户单位上海证券交易所（以下简称上交所）有机集成，融入上交所已有科技监管体系，并在上市审核、公司监管、交易监管等部门开展应用示范，同时相关成果还分别在华泰证券、海通证券、汇添富基金公司等证券行业机构开展应用示范，取得良好成效。未来，项目成果还将持续向更多市场机构辐射推广，赋能各类市场参与者各司其职，统筹推进资本市场防风险、强监管、促高质量发展重大任务。

二、信息披露文档要素智能抽取与审核

我国资本市场大部分信息披露文档仍为 PDF 格式的

非结构化数据，不便于监管部门或投资者从中快速获取有效信息。同时，信息披露文档是富格式金融长文档，具有文本、表格、图表等多种要素的，同时还大量使用实体共指、嵌套关系、长程修饰等复杂语法结构，导致通用领域的文本理解与抽取模型效果不佳。

已有提高实体识别准确率的方法主要集中在基于序列标注的模型上，包括通过多标签标注层来识别嵌套实体，引入分层模型以改进效果等。然而，这些方法仍然面临标签不平衡和错误传播的问题，同时忽视了嵌套实体在高维度语义信息聚合中的复杂性，尤其是在金融领域，长文本中的复杂嵌套实体及其同义替代导致的数据稀疏问题，严重影响了识别准确度。针对上述问题，项目在传统模式匹配算法基础上，设计了分层嵌套实体识别模型，迭代聚合不同维度语义信息，低层识别单个实体，高层识别嵌套实体，显著提升了对复杂嵌套结构的识别效果。相比于传统的基于序列标注的多标签或动态叠加识别层方法，项目成果有效解决了错误传播和数据稀疏问题。

另一方面，针对关系触发词多样和冗余信息干扰导致的关系抽取准确率不高问题，现有的异构图神经网络方法主要依赖于元路径聚合等手段处理图中的多样性与复杂性，但未能充分结合依存句法解析等语言学信息。本项目提出的基于依存句法解析的异构图卷积神经网络，融合了实体类型、语义相似度、句法结构等多维信息，并结合对比学习策略，显著提高了关系抽取的准确性。

另外，针对由于市场环境变化快速，导致历史模型面临性能衰退的挑战，项目提出了一套基于领域适配的向量模型编码技术以及结合增强语义编码的聚类与分类策略，不仅有效解决传统模型在快速变化的金融市场中的适应性不足问题，而且在提高迭代效率的同时，保证了模型的稳定性与性能。最后，项目在工程上构建了人工标注、机器学习到模型抽取的信披要素抽取平台，以人在回路方式实现业务人员快速通过在线可视化的方法完成对披露文档的要素标注并迭代模型训练。

综合利用以上技术，项目实现对 545 个招股说明书要

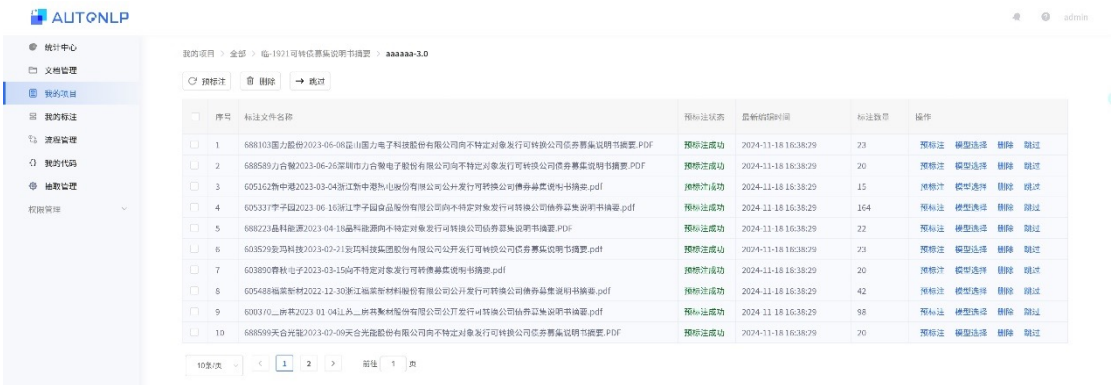


图 1 关键要素抽取功能

图 2 合规性审核分析功能

素字段识别准确率从立项前约 85% 提升到 98.64%，同时平均召回率 98.07%，达到国内外同行业先进水平。目前，信息披露文档智能审核系统已在用户单位上海证券交易所集成部署，系统年处理信息披露文档约 20 万份，日均产生数据记录 1.5 万条，人工修订率从 43% 降至 16%，大幅减少人工参与度，提升监管审核效率。

在合规性审核过程中，系统对勾稽关系、关联交易、出资瑕疵、股份质押、司法风险五类主要问题进行检测，同时根据各类子问题的检测结果返回对应的预警信息。在应用示范过程中，系统发现多起信息披露不完整问题。例如，在针对发行人 A 的招股说明书（申报稿）的关联交易问题审核中，检测到发行人与 B 公司的交易内容不完整，缺少对流程合规性（关联交易是否经由股东大会通过）的内容披露，经系统判断可能存在关联交易不实问题，提出关联交易问题预警（如图 3 所示）。在针对发行人 C 的招股说明书（申报稿）的出资瑕疵问题审核中，检测到发行人在出资缴纳声明（无法判断资金是否实缴）、股东背景声明（部分股东缺失）、股份转让声明（部分证明材料缺失）等内容的披露不完整，经系统判断可能存在出资核验、股权变更信息披露等子类问题，提出出资瑕疵问题预警（如图 4 所示）。

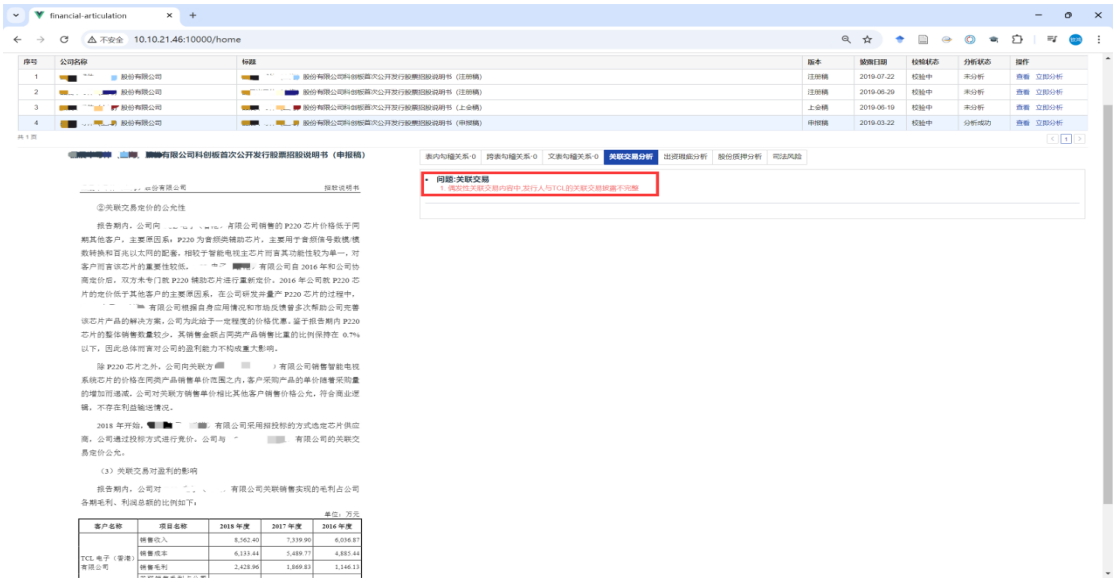


图 3 关联交易问题预警

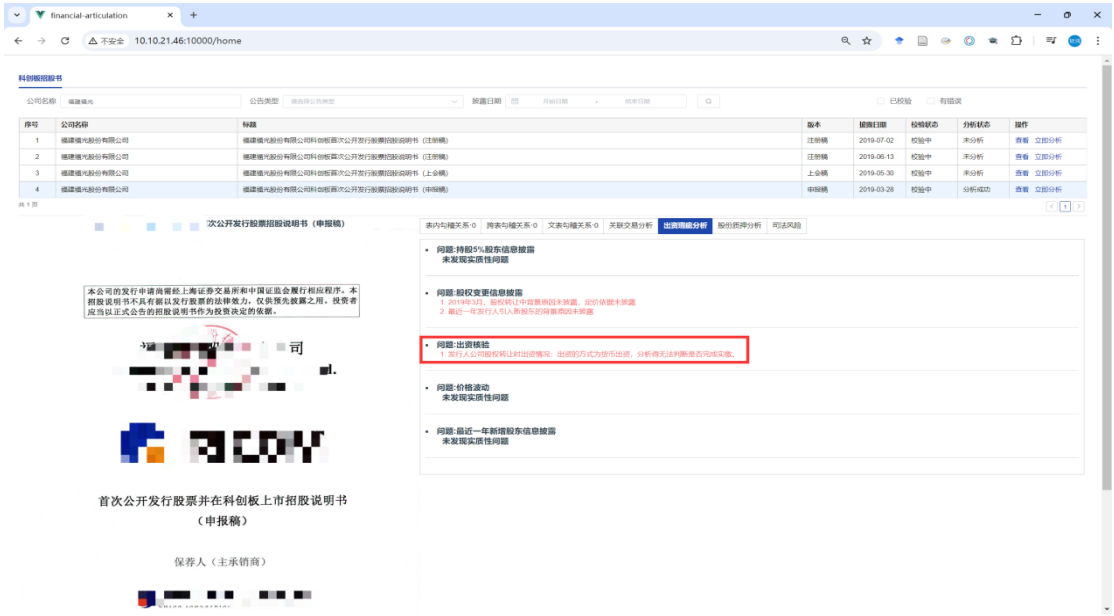


图 4 出资瑕疵问题预警

此外，针对已开源的金融文档评测集中，文本长度一般都小于 1000 字，无法满足对于金融长文档的评测需求，缺少测试基线等问题，项目组基于相关研究成果开源面向金融场景的长文档测评集 FinLongEval，设计了针对 43 篇金融长文档的 12 大类共 347 道问题，从相关性、有用性、流畅性、连贯性、一致性和忠实度 6 个维度对人工智能大模型生成的答案进行评估，填补了该垂直领域的测试基线空白。

图 5 项目开源的金融长文档理解测评集

三、开放域低资源市场舆情风险识别

在金融领域，市场的波动常受突发舆情事件影响，罕见或全新的事件对金融风险检测和管理构成了重大挑战。例如，突发的政策变化、公司丑闻等事件，都可能对市场稳定性产生严重负面作用。然而，由于这些事件发生的频率较低，历史数据往往不足，传统依赖大量标注数据的检测方法在相关场景下表现不佳，难以训练出能够泛化到新事件的模型。这种数据稀缺性进一步加剧了问题的复杂性，使得在有限样本条件下准确及时地识别舆情风险事件变得尤为困难。解决金融舆情中小样本和零样本场景下的问题能够改进罕见事件的检测能力，帮助金融机构和监管机构更好地动态管理风险，减少突发市场动荡的可能性及其影响，增强整体市场稳定性，促进金融生态系统的韧性。项目改进了现有的小样本事件检测技术，同时提出了一种新的零样本检测技术，大幅提高模型在复杂金融事件场景下的效果。

小样本检测技术方面，通过在原型网络算法中引入标签增强和对比学习来提高模型的鲁棒性和稳定性。具体而言，为了充分利用预训练模型捕获上下文中字符语义特征的能力，项目构造了人工模板，将标签转换成一个语义完整的句子和样本进行衔接，从而帮助模型捕获相应特征。通过利用标签增强拉近同类样本的表示，弱化模型对于数据的依赖性，同时引入对比学习在高维空间上优化样本表示，聚合同类数据，拉远不同类数据，降低数据敏感性提高模型鲁棒性。相比于国际上已有的小样本事件检测基线系统（原型网络 prototype network, 孪生网络 Siamese network, 元学习模型 SNAIL），本项目所采用的小样本检测模型在标准评测数据集 FewEvent，MAVEN 上 F1 值分别提高了 4.7% 与 9.2%。另外在数据中混有 40% 噪声的情况下，模型整体具有 10% 的性能增益。通过性能比较，本项目所提出的技术同时增强了小样本检测模型的鲁棒性和稳定性。

零样本检测技术方面，提出了基于有序对比学习的端到端零样本事件检测技术。该技术利用对比学习的自监督学习范式，消除了模型训练对未知事件类型标注样本的依赖。同时利用有序对比学习通过不同方法构造出与原始样本高低各异相似度的对比样本，包括 Dropout 样本、重写样本、同质样本以及异质样本四类对比样本，通过计算有序对比学习损失函数，以维护这些对比样本与原样本在语言模型隐空间中距离的有序关系，帮助语言模型充分学习同一事件类型内样本之间的共同特征，同时区别不同事件类型的样本。相比于当前最先进的无监督事件聚类算法 SCCL，半监督事件检测算法 SS-VQ-VAE，本项目成果在标准数据集 ACE-2005，FewShotED 上取得了 10% 以上的 F1 值提升。值得一提的是，在未见过的事件类型上，F1 提升幅度到达 15% 至 20%。实验证明，本项目所提出的零样本事件识别技术取得了已公开的最优识别性能与识别效率。

综合以上技术，项目构建的资本市场舆情风险识别与处置系统落地上海证券交易所、华泰证券、江苏省股交中心等机构，用户总人数超过 3000 人，主要包括交易所的上市审核人员、公司监管人员、市场服务人员，证券公司的投资顾问、投资研究员、风险控制研究员等。用户可通过 web、轻应用方式查看利用自然语言算法处理后的舆情数据，也可以通过 API 接口调用、kafka 消息订阅、页面嵌入等方式或者混合方式进行服务接入应用。

相比行业内其他舆情系统，本系统基于小样本和零样本事件检测技术，结合强化学习和主动学习技术构建了人机协同的监管预警回路，可通过配置化的方式实现对新舆情事件类型识别的自助式满足，从而将原先需要 1~2 个月的舆情预警需求开发测试上线流程缩短至 5~10 分钟完成，由业务人员自行少量标注后即可快速上线并持续增强模型效果，支持个性化舆情订阅，具有完整的事件标签体系，支持用户实时定义、抽取新出现事件。

目前，系统日均采集金融市场舆情达 6 万条，配置的 601 条预警日推送舆情预警消息超过 1.9 万条，对上市企业的重点舆情事件识别准确率超过 90%，为用户及时预警各类问题风险。例如 A 公司在上市审核期间收到《行政处

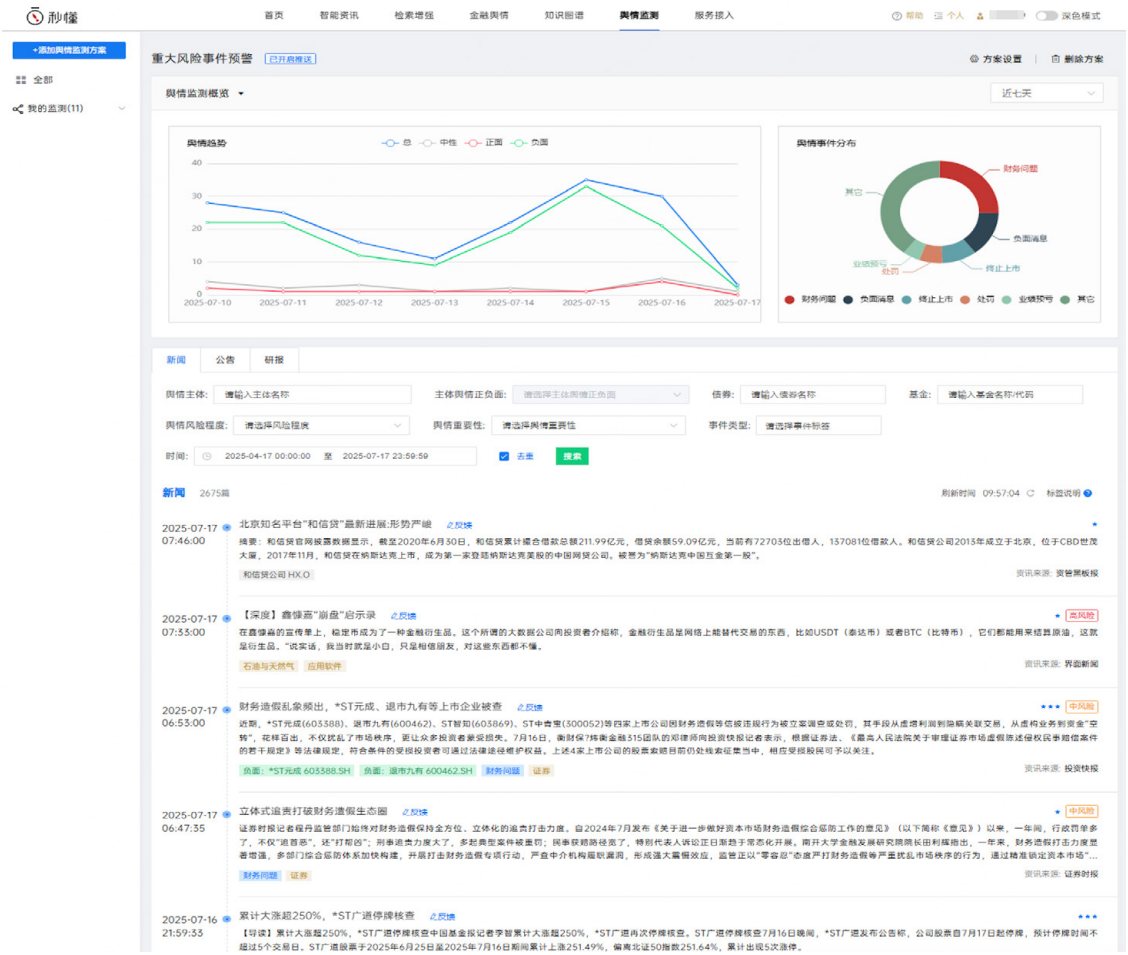


图 6 舆情风险事件识别

图 7 实时舆情订阅

罚决定书》，不久 A 公司通过上市委审议并披露招股书（注册稿），但该处罚事项未予披露。上交所审核人员通过系统舆情提醒发现后，督促保荐机构核查，并对发行人、保荐机构、保代予以口头警示。在华泰证券应用示范过程中，系统捕捉到上市公司 B 收到监管问询函形成负面舆情事件，系统自动生成预警信号推送给相关研究员（如图 8 所示），研究员收到信号后，结合公司基本面、市场价格等综合评估公司资质，决定是否调整公司评级。又如，系统捕捉到某境外上市的公司 C 盘前股价大幅下跌舆情，系统自动匹配到多个客户以该标的为质押券，可能产生履约保障比不足的信用风险，及时推送给对口监控风险经理，分析舆情影响，按规及时处置提示，避免了可能造成的损失。

图 8 系统推送的舆情预警示例

四、企业科创属性和财务经营综合评价体系

公司画像对监管部门的监管审核工作和投资者的投资决策均起到重要参考作用。但国内外现有产品主要聚焦公开信息展示，缺乏有效研判企业科创属性，分析产业链、供应链“卡脖子”环节的产品工具，财务指标与非财务信息的联动运用较少。同时，科创板公司的发展阶段、经营特征、市场环境与传统行业存在较大差异。已有的企业评价分析系统往往采用较为宽泛的行业划分标准选取同行业可比公司，可比性较弱，进而影响同行业对比指标的有效性。

本项目以信息披露、舆情、专利、标准、科技成果、研发投入、核心团队情况等数据为基础，通过领域知识、特征分箱和权重证据（WOE）转换技术进行特征筛选，利用机器学习算法回归分析，实现对企业科创属性的精准、综合评分定级，多维度构建科创属性指标的计算方法和评分逻辑，指标数量从 100 项提升至 266 项。相较之前同类型的高新技术企业评价指标，该指标体系紧紧贴合资本市场发行审核监管和投资业务需求，同时融入了国家科技创新和产业发展政策要求。特别地，项目改进了传统的科创行业分类方法，利用企业专利数据形成专利关键词向量，基于词向量相似度细化划分科创行业，更准确开展同行业比较和相关指标计算。政策契合度评估方面，计算招股书中产品词和政策文件中政策词之间的相关性，利用基于预训练大模型的双塔式信息检索算法（一塔是招股书的向量编码，一塔是政策文件的向量编码），计算企业主营产品和业务与政策的契合程度，相关功能为行业首创。

其次，优化企业关系、产业链和供应链知识图谱，实现关联探寻、关系脑图、风险图谱、股权穿透、集团系谱、关系查询等重要应用功能，辅助上市审核和公司监管业务人员对企业各类关系脉络、产业链的上下游等快速查询分析。从股权、任职、投资、电话邮箱地址等维度构建关联方挖掘模型，智能挖掘企业与企业、企业与自然人之间现有和历史的潜在关联关系，辅助识别出注销或转让子公司、隐匿关联关系、与关联方之间通过多个非关联方转换为非关联业务等不当手段。结合审核问询、现场督导检查等方式，交叉验证是否存在向关联方输送利益、关联交易不公允、虚增销售收入等行为。

另一方面，财务造假严重扰乱资本市场秩序、动摇投资者信心、侵犯投资者权益，其通常具备手段隐蔽、模式复杂、动机多样、层级多、链条长等特征。传统财务异常测算方法存在无法充分利用未被揭露的财务异常，难以区分财务舞弊行为的异质性等缺陷，且专家循证成本高、识别工作效率有待提升。项目研发财务数据异常识别模块，改进提出 5S 方法、FFPI 方法、CFFI 方法等三种财务数据异常测度方法，同时基于同群效应和集成学习算法构建全新的财务舞弊识别框架，利用前沿自然语言处理技术结合上市公司动态业务范畴，突破传统行业分类下财务舞弊行为同群程度低的桎梏，搭建财务舞弊聚类分析下的指标算法。创新地实现财务数据异常因子和模型的归因，提供高可信和高解释度的模型辨识效果，落实开展业务层面的财务可信度分析和财务健康度分析。

上述功能模块已集成到上交所新一代公司画像系统、海通证券智能风险预警中心，辅助监管与审核人员判断企业“硬科技”属性，及时发现企业财务问题线索，智能提示财务异常归因。相关功能界面如图 9、10 所示。

上交所相关业务部门利用系统开展 2023 年度年报审核工作时，对系统提示的问题线索认真拆解对比，总体判断相关预警线索对监管发函问询具备参考价值，为人工审核提供积极补位，一定程度减少了人工审核的盲区。在海通证券应用示范过程中，根据历年案例统计，财务风险提前预警平均天数为 1123 天，问题企业被系统预测出中风险以上比例为 72.78%，预测局部风险以上比例为 97.28%。基于相关应用成果发表的论文《做好资本市场科技金融服务大文章——基于证券交易所的视角》被人大复印报刊资料转载。

图 9 公司画像系统主页

五、基于大模型的舆情分析及自动问询函生成

在项目实施过程中，大语言模型技术迅速兴起并发展为人工智能、自然语言理解等领域的主流技术，在多种任务中取得远超中小模型的 SOTA 效果。项目密切关注大模型等前沿技术发展趋势，及时吸收并优化技术方案。一方面，在 BERT 模型基础上（十亿参数规模），扩展到六十亿至一百三十亿规模的 GPT 基座模型，通过指令微调统一处理多个任务，在事件识别、事件抽取、黑嘴识别等场景领先或持平多个单任务模型，各项任务 F1 指标平均提升超过 2%。

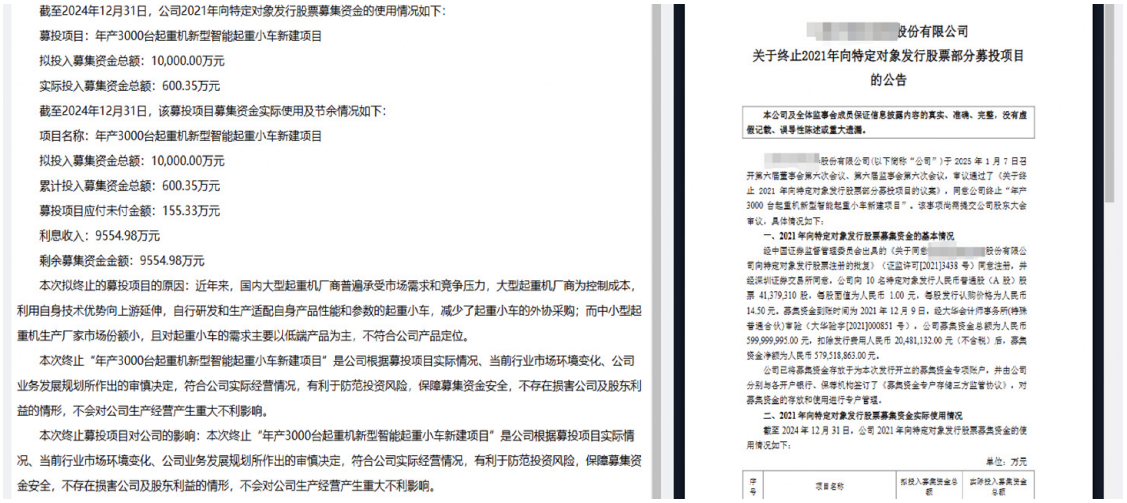
另一方面，针对现有大语言模型的词表有限且固定，而资本市场问询函语义丰富、问询内容长且复杂的问题，基于十亿至百亿规模的大模型研发动态词表生成技术，自适应地根据风险事项调整大模型词表，高效生成专业的问询函初稿。动态词表技术改变了传统大模型的固定生成方式，实现降本增效，并且能突破大模型的不可解释与幻觉问题，对生成文本进行引证，使生成内容可解释、高置信。相比于当前采用标准静态词表技术的语言模型（例如 GPT-2, TinyLlama），经测试利用本项目所提出的动态词表扩张技术后，MAUVE 值（用来判断生成的文本与人类写作相似程度的一种方法）提升 25%，同时有效提高了模型推理速度，延迟降低 20%。特别地，在金融、法律等垂直领域中，本所项目所提出的技术显著优于微调后的通用模型（20% 的 MAUVE 值提升），证明了动态词表在高效领域迁移与高效推理上的技术潜力。相关成果《Generation with Dynamic Vocabulary》在行业顶会 EMNLP2024 发表。

六、成果应用示范

项目主要研究成果集成落地到上交所，辅助监管审核人员开展日常工作。在上交所应用示范过程中，系统平均月活 337 人，月平均登录次数 4291 人次，月平均访问功能次数 12977 次，系统日均抽取处理信息披露文档 2300 篇，产生数据记录 1.5 万条，日均采集市场舆情信息 6 万条，推送舆情预警 1.95 万条，日均异常交易预警运算超 2 万次，实时交易数据处理瞬时峰值达 22.6 万笔 / 秒。此外，通过数据安全共享平台，在上交所和北京邮电大学等 10 家项目参与单位间安全对接和传输信披文档、舆情、财务等数据，累计上传数据 19761 份，共享 / 下载数据 43605 份。同时系统识别处理后的 152 类超 150 万条结构化信息披露数据向证监会监管大数据仓库、上海大数据中心和东方证券、富国基金等 8 家证券基金公司共享，降低行业取数用数成本。

总体上看，在上交所的应用有效提升了一线监管与审核工作质效，在实际业务中提供了诸多有价值的预警或线索，符合预期效果。例如，在科创板再融资审核中，审核人员通过科技评价、产业链分析功能判断某公司的再融资募投项目与公司主营业务相关性较低，且公司在该领域技术储备薄弱，最终发行人撤回申请。在上市审核业务中，系统帮助上交所审核人员发现了某发行人在上市审核期间收到行政处罚但未予披露上报等多个违规问题线索。在交易监管业务中，项目研发的多维度异常交易预警及时向监管人员提示拉台打压、操纵市场等违法违规行为，部分案件被公开披露或报道，有效彰显金融监管“长牙带刺”、有棱有角。

同时，项目成果积极向证券行业辐射，分别在华泰证券、海通证券、汇添富基金公司等机构开展应用示范。在华泰证券应用过程中，舆情风险识别相关成果在企业运营、市场动态监测及投资策略调整等方面发挥了积极作用。例如，系统捕捉到某证券股价大幅下跌舆情，系统自动匹配到多个客户以该标的为质押券，可能产生履约保障比不足


截至2024年12月31日，公司2021年向特定对象发行股票募集资金的使用情况如下：

募投项目：年产3000台起重机新型智能起重小车新建项目

拟投入募集资金总额：10,000.00万元

实际投入募集资金总额：600.35万元

截至2024年12月31日，该募投项目募集资金实际使用及节余情况如下：

项目名称：年产3000台起重机新型智能起重小车新建项目

拟投入募集资金总额：10,000.00万元

累计投入募集资金总额：600.35万元

募投项目应付未付金额：155.33万元

利息收入：9554.98万元

剩余募集资金金额：9554.98万元

本次拟终止的募投项目的原因：近年来，国内大型起重机厂商普遍承受市场需求和竞争压力，大型起重机厂商为控制成本，利用自身技术优势向上游延伸，自行研发和生产适配自身产品性能和参数的起重小车，减少了起重小车的外协采购；而中小型起重机生产厂家市场份额小，且对起重小车的需求主要以低端产品为主，不符合公司产品定位。

本次终止“年产3000台起重机新型智能起重小车新建项目”是公司根据募投项目实际情况、当前行业市场环境变化、公司业务发展规划所作出的审慎决定，符合公司实际经营情况，有利于防范投资风险，保障募集资金安全，不存在损害公司及股东利益的情形，不会对公司生产经营产生重大不利影响。

本次终止募投项目对公司的影响：本次终止“年产3000台起重机新型智能起重小车新建项目”是公司根据募投项目实际情况、当前行业市场环境变化、公司业务发展规划所作出的审慎决定，符合公司实际经营情况，有利于防范投资风险，保障募集资金安全，不存在损害公司及股东利益的情形，不会对公司生产经营产生重大不利影响。

[redacted]股份有限公司

关于终止2021年向特定对象发行股票部分募投项目的公告

本公司及全体监事会成员保证信息披露内容的真实、准确、完整，没有虚假记载、误导性陈述或重大遗漏。

[redacted]股份有限公司（以下简称“公司”）于 2025 年 1 月 7 日召开第六届董事会第六次会议、第六届监事会第六次会议，审议通过了《关于终止 2021 年向特定对象发行股票部分募投项目的议案》，同意公司终止“年产 3000 台起重机新型智能起重小车新建项目”，该事项尚需提交公司股东大会审议，具体情况如下：

一、2021 年向特定对象发行股票募集资金的基本情况

经中国证券监督管理委员会出具的《关于同意[redacted]股份有限公司向特定对象发行股票注册的批复》（证监许可[2021]3438 号）同意注册，并经深圳证券交易所同意，公司向 10 名特定对象发行人民币普通股（A 股）股票 41,379,310 股，每股面值为人民币 1.00 元，每股发行认购价格为人民币 14.50 元，募集资金到账时间为 2021 年 12 月 9 日，经大华会计师事务所(特殊普通合伙)审验（大华验字[2021]000851 号），公司募集资金总额为人民币 599,999,995.00 元，扣除发行费用人民币 20,481,132.00 元（不含税）后，募集资金净额为人民币 579,518,863.00 元。

公司已将募集资金存放于为本次发行开立的募集资金专项账户，并由公司分别与各开户银行、保荐机构签订了《募集资金专户存储三方监管协议》，对募集资金的存放和使用进行专户管理。

二、2021 年向特定对象发行股票募集资金实际使用情况

截至 2024 年 12 月 31 日，公司 2021 年向特定对象发行股票募集资金的使用情况如下：

单位：万元

序号	项目名称	拟投入募集资金总额	实际投入募集资金总额

图 10 问询函自动生成功能

的信用风险，推送给对口监控风险经理，分析舆情影响，按规进行处置提示。

在海通证券（现国泰海通）应用过程中，系统提供的企业财报可信度异常识别功能较为准确地识别企业财报造假动机、手段、科目及行为，为投行条线在企业上市服务、企业存续期管理提供重要参考。2023 年以来，证监会和地方证监局处罚的 64 家公司中，系统对 51 家公司预先给出中高风险提示，剩余的 13 家也给出了局部风险预警，有效避免了财富管理、自营投资等业务可能的损失。同时，相关成果也同步对接服务上海证监局、上海市国资委等机构。

在汇添富基金公司应用示范过程中，企业评价功能在机构的投资实践中展现显著成效。该系统已覆盖 130 只权益类基金产品，管理资金规模超 590 亿元，其中包括 3 只关注科创板的公募基金，管理金额达 63 亿元。系统的实施优化了企业创新能力评分和估值分析过程，显著减少人工干预、提升相关业务效率，同时降低了人为错误发生的概率，为投资决策的科学性和准确性提供保障。

项目成果亦获得行业高度认可。包含异常交易实时预警计算等项目成果的上交所新一代交易监管系统获得中国人民银行 2022 年度金融科技发展奖一等奖。包含财务异常分析等项目成果的海通证券智能风险预警中心获得中国人民银行 2022 年度金融科技发展奖二等奖。项目数据安全共享平台建设运用的隐私保护及可信传输技术获得中国通信学会科学技术奖二等奖。

数智化助力证券交易所高质量发展

王泊，徐广斌
上海证券交易所

摘　要： 本文围绕我国证券交易所“十五五”数智化发展，系统分析了数智化在提升资本市场服务实体经济能级、强化核心竞争力、夯实金融基础设施安全、优化科技监管效能等方面的关键作用。数智化不仅是交易所应对复杂形势的必然选择，更是构建行业数字生态、助力经济高质量发展的核心驱动力，为我国建设世界一流交易所提供了理论支撑和实践框架。面对全球科技革命与金融强国建设的双重机遇，证券交易所需健全网络信息安全体系、升级交易系统架构、深化数字化转型、布局数字基础设施及加强科技治理，推动实现安全高效、开放包容、国际领先的数智化转型目标。

关键字： 证券交易所；高质量发展；金融科技；监管科技；金融基础设施

一、“十五五”时期证券交易所科技发展面临的机遇和挑战

现代资本市场的运转高度依赖信息科技，我国证券交易所自开业伊始即采用领先的电子化交易模式，经过三十多年的发展，已建构起完备的市场基础设施体系，有效支撑交易、监管和服务等核心功能的高效安全运行，助力我国资本市场在短时期内实现了跨越式发展。“十五五”时期，数智化成为证券交易所发展的关键路径和核心驱动力，助力交易所新阶段下的高质量发展。

“十五五”时期，我国资本市场进入进一步全面深化改革和加快迈向高质量发展的关键阶段，在新“国九条”和证监会“1+N”等政策指导和支持下，交易所以加快打造安全、规范、透明、开放、有活力、有韧性的资本市场为目标，以防风险、强监管、促高质量发展为主线，以做好金融“五篇大文章”为动力，努力开创高质量发展新局面，加快建设世界一流交易所。另一方面，以数据和科技为核心驱动的全球新一轮科技革命和产业变革加速演进，以人工智能、大数据和云计算为代表的新兴技术与传统金融业务深度融合，数智化成为世界一流交易所重要的发展战略。

同时，随着资本市场快速发展和数字化转型迈入新阶段，交易所的科技工作面临着前所未有的变化与挑战。一是面对外部形势的复杂变化和金融强国的战略要求，防风险能力需要进一步巩固，安全运行保障水平、网络安全防护能力与关键技术的自主安全可控水平仍需进一步强化。二是业务模式加速创新，服务范围不断扩大，投资者结构快速演变，“大数据 +AI”监管需进一步加强。三是数字化转型深入推进，在安全治理的基础上，需要加强数据开发、流通与共享，扩大赋能成效。四是科技治理需要提升，强化规范建设与资源配置，加大对人才和创新的支持。

二、深刻认识数智化对证券交易所“十五五”高质量发展的重要意义

一、是有助于顺应数字化新阶段，提升服务实体经济能级

数智化的网络化、自动化和智能化特征，有利于高效匹配超大规模市场的投融资双方需求，更好发挥资本市场的枢纽作用。利用大数据、人工智能等手段，可对实体经济企业进行精准评估和画像，将资金引导至新质生产力企业，满足不同类型、不同发展阶段实体经济企业的融资需求。数智化与绿色化融合，可更好地识别和评估绿色项目，丰富绿色金融产品体系，助力实体经济实现绿色低碳转型和可持续发展。数智化转型降低金融服务的门槛和成本，为普惠金融客户提供更个性化的金融产品和服务，让更多中小企业、老年和低收入群体能够享受便捷的金融服务，更好践行人民性。

二、是有助于提升核心竞争力，促进高质量发展

交易系统是证券交易所的核心竞争力之一，利用先进技术研发交易系统，可使交易系统在具备低时延海量订单处理能力的同时，获得灵活扩展能力，满足各类证券交易和国际竞争需要。利用高可配置化与模块化设计，按需搭建系统功能并联通外部集成，形成具有普适竞争力的平台，输出到各类交易场所。利用互联网、云、区块链打造新型交易结算系统，可突破时空限制，实现快速展业。推动发行、上市、交易、结算和服务等全流程实现数字化，可全面提升市场的运行效率，提高透明度和规范性并降低交易成本。

三、是有助于夯实自主可控安全高效的金融基础设施，建设金融强国

夯实自主可控安全高效的金融基础设施，是推动高质量发展和建设金融强国的应有之义。在日益严峻的网络安全态势下，安全自主可控刻不容缓，利用智能运维、网络攻防等领域的技术突破，可更好落实“三法一条例”等网络安全重要法律法规，提供更全面、精准和高效的解决方案。通过加密技术、下一代防火墙、入侵检测系统等手段打造立体化网络安全防护能力，能够有效保护数据，并在攻击发生时快速响应，有效抵御外部冲击。基于分布式容灾机制，可确保交易系统的不间断运行，并在发生自然灾害等极端状况下迅速恢复交易系统。

四、是有助于提升科技监管质效，加强风险防范能力

随着科技对金融的渗透融合加深加快，新业态新模式新情况不断涌现，传统监管模式面临的工作点多线长、量大面广与监管人力不足的矛盾愈发突出。利用大数据和人工智能技术，可以对网络舆情、交易行为和信息披露等进行全要素、全链条、全覆盖的实时监管，让监管更加智能高效。通过大数据基础设施，可以更好收集、挖掘和共享海量数据，实现监管信息的共享和监管协同，促进监管模式转型。监管科技（RegTech）通过将监管法规转化为机器可读代码，实现自动化报告生成和合规性检查，可减少人工操作的误差、提高效率，并及早发现和处置各类风险，有效防范化解重大风险。

五、是有助于提升治理水平和能力，提高业务运行效能

借助数智化推动治理水平和能力的提升，是数字时代推进治理体系和能力现代化的必然选择。数智化治理可提供更客观全面的信息支持和更精准的分析，支持决策的科学化和精准化，提升管理水平和运行效能。通过构建统一数字化平台，替代“烟囱式”系统，可实现跨部门业务协同、跨领域资源调度、跨场景决策联动和跨层次上下联动，推动“碎片化治理”向“系统化治理”。通过构建智能中枢，可消除数据孤岛，实现业务流程的集成贯通和自动化，并在智能分析优化流程的基础上，借助智能助理、数字人、AI 智能体等提升工作效率。

六、是有助于金融基础设施互联互通，支持高水平对外开放

通过新技术加强资本市场基础设施的互联互通，有利于促进系统互联和信息互通，提升证券交易、结算和支付等跨境金融服务的便捷性和高效性。构建安全、高效的数据共享平台，可利用多方安全计算、联邦学习等技术在保障数据隐私下实现跨机构数据的安全共享及“可用不可见”。稳步加强“数字金融”境内境外“双循环”，可促进金融资源的国际优化配置，吸引全球流动性，提升我国定价权。通过参与全球数字金融标准等国际规则的制定和推广，可增强国际竞争力和规则影响力，助力人民币国际化、“一带一路”发展等战略落地。

七、是有助于带动行业协同发展，构建行业数字新生态

“十五五”期间，科技创新驱动成为证券行业重要的增长引擎，数智化有利于交易所更好发挥行业科技发展排头兵作用，引领打造全局协同的治理体系和能力聚合的数字发展生态。加强行业信息技术研究平台建设，有利于发动行业力量对行业科技发展的热点共性问题攻关，汇聚行业优势力量打通关键领域的堵点和瓶颈。依托资本市场金融科技创新试点等创新项目，可探索新技术、新模式和新应用，促进行业守正创新。加强产学研用协同创新，有助于突破重大共性关键技术，促进科学创新、人才培养及知识产权建设。

三、上海证券交易所“十五五”数智化发展路径思考

“十五五”时期，上海证券交易坚持以习近平新时代中国特色社会主义思想为指导，围绕打造科技领先的数字智能型交易所目标，积极运用信息科技提升效率、竞争力和服务实体经济的能力，助力实现全面、协调、可持续高质量发展，为加快实现以科技为核心竞争力的世界一流交易所奠定坚实基础。

一、是健全网络和信息安全一体化管理，筑牢安全底线

树牢安全底线思维，推动业务与技术运行安全、网络安全和数据安全“四安全”一体化管理，建立“事前预防型”安全运行体系。健全软件安全开发管理制度，将安全理念和安全工具深植软件研发建设全过程。建立高效可靠的业务全流程数字化系统，推行“直通式”处理，实现业务技术操作运行“大闭环”。优化应急案例场景、流程机制和响应效率，持续开展业技联动的实战化应急演练。打造数

字化智能化运维，建构集中智能运维平台，完善监测体系，构建多模态知识库，借助 AIOps 优化运维流程，提升运维效率和质量。

建设态势感知平台“一个大脑”和边界防护、主机防护、终端防护、网络防护“四种能力”，增强动态防御和主动防御能力。实现关键信息系统全面自主可控，发挥行业信息技术应用创新基地功能，促进行业标准、指南和案例库等建设。健全数据安全管理体系，完善数据分类分级保护和数据安全评估，优化数据权限管理和安全审计机制。利用多活数据中心、异地备份和行业灾备中心等措施，确保关键业务的高可用性和容灾能力。构建信息技术风险数字化管理体系，完善风险库、风险因子和风险对策，强化对重大建设、关键系统和新技术应用的风险防控。

二、是实现交易系统升级换代，支持核心业务发展创新

全面建成低时延分布式开放架构的第四代核心交易系统（G4 系统），满足完全自主可控，实现系统在可用性、可维护性及技术指标等方面达到国际领先水平。丰富系统功能，满足多样化、个性化交易需求，稳步推进 G4 系统在股、债、基、衍“四大市场”全面部署。持续优化在用交易系统，运用高性能网络、内存数据库、大并发接入等先进技术减少时延、提升容量、提高扩展和吞吐能力，加固安全，确保高效、稳定、安全的交易环境。适时启动第五代交易系统预研，支撑证券产品 7×24 小时交易，支持交易系统上云和集成 AI，夯实与一流交易所相匹配的证券交易基础设施。

稳步推进交易系统国际互联互通，支持与国际金融业务网络、金融信息交换网络及交易订单路由网络互联互通。利用移动互联网、区块链等新技术构建创新型交易综合服务平台，实现业务敏捷化部署，在支持传统交易业务的同时，支持场外业务落地能力和泛金融产品交易能力。探索交易系统产品化国际化，开展境内外市场技术合作，为参与交易、行情、数据和清算等业务的相关方提供国际化、标准化接口。根据国际交易场所业务场景、监管环境和技术条件裁剪定制交易系统，提高交易系统竞争力和国际影响力。

三、是深化数字化转型，提升科技监管与服务效能

推进数据创新应用，充分发挥数据的要素价值和乘数效应，赋能业务创新、风险管理和投资者服务，提升一线监管的精准性有效性。建构“集中统一、数据丰富、安全可靠、智能高效”的数据基础设施，实现数据统一集中、管理和服务。健全数据要素资源体系，按需拓展数据资源，建设数据资产管理服务平台。完善数据标准与制度体系，从资源目录、供需对接、质量管理、服务管理等全方面提升数据治理能力，实现数据盘、整、规、治、用规范化标准化。提高多维度多形式的数据服务能力，丰富数据工具和产品，探索敏捷式开发和矩阵式管理。建立数据流动分类管理、安全传输和审查机制，在确保数据安全的前提下，有序促进数据的开放共享和流通交易。

以重点数字工程为抓手，深化推进转型。建设一体化综合科技监管平台，对接画像、数据和 AI 等通用能力，覆盖上市审核、公司监管和债券监管等一线监管全链条。深入推进画像系统等监管科技建设，形成对市场参与者和产品等主体的全景式穿透分析，重点加强信息披露监管、财务舞弊分析和异常交易识别功能。深化“一网通办”建设，打造以用户为中心的精准推送、服务直达和便捷办理，从“一网通办”向“一网好办”升级。持续建设“一网统管”，打造财务、人力和办公等数字管理体系，做到数据“只报一次”。推进人工智能建设，围绕重点需求创新应用，配套推进算力、算法和场景建设，为内外部用户提供易用的人工智能服务和功能调用。

四、是适度超前布局数字基础设施建设，助力资本市场高质量发展

大力推进“上证云”建设，发挥云计算弹性、兼容和资源整合优势，多模式扩展云基础设施，扩大服务的规模、范围和对象，建构安全灵活高效、国际先进的云平台。坚持“应云尽云”发展方针，实现业务系统全面上云，稳步探索核心交易系统云化建设，构建新型交易基础设施。围绕市场发展和业务创新需要，推出层次丰富、种类多样的云应用、云服务和云产品，为行业提供更安全易用的平台化服务和定制化部署。为行业机构信息系统建设提供多样化的云解决方案，支持敏捷部署、集约管理，降低中小机构建设门槛和成本。依托行业云聚集多方合作，结合大数据、人工智能、区块链等新技术，赋能上证链、上证 e 服务和星企航等产品升级。探索基于云基础设施的科技监管和金融科技创新，助力资本市场数智化转型。

推进建设先进管理、国际标准、绿色智能的一流数据中心。构建数据中心数智化运维，采用智能运维、机器人巡检、物联网等技术提升运维和安全保障能力。适应智算平台等创新发展，强化数据中心机房布局、通信电力和功能配置等能力建设。综合运用 AI 驱动节能、优化冷却系统、利用可再生能源等方法，进一步减少能耗和碳排放，通过绿电、绿证和碳普惠交易等碳中和措施，打造零碳数据中心。结合数据中心布局，优化建设高可靠冗余网络架构，综合

运用 5G、IPv6 和卫星互联网等技术打造固移融合的空天地一体化网络。积极参建行业云、网、库、链、智慧监管等重点工程，助力监管效能提升和行业运营成本降低。

五、是加强科技治理，培育科技发展“新质生产力”

健全集团化科技治理体系，加强对科技资源、科技力量和科技工作的统筹，构建适配数智化转型的组织架构。强化科学规划，建立覆盖规划设计、实施运行、考核评测和改进完善的循环控制机制，完善需求论证、统筹与跟踪机制，提升信息系统应用成效。借鉴先进标准和最佳实践，提升科技制度体系的科学性，规范相关科技活动，防范化解潜在风险。打通技术研发工具链，加强 AI 在研发中的应用，推广人机协同开发和零代码、低代码开发模式。推进 IT 管理数智化、规范化和平台化，实现信息化建设全生命周期覆盖，提升科技治理信息共享。强化科技部门与财务、审计等内控部门联动监督，完善采购廉政风险长效化防控机制，优化外包管理的流程和机制。健全信息技术和数据服务机构管理体系，加强对供应商的日常管理，确保提供安全可靠的信息技术产品和服务。

充分发挥交易所的创新主体作用，加强人工智能、大数据、云计算等关键技术的应用研究和标准供给，积极参与国际技术交流与标准建设，增强前瞻性技术储备。以“人工智能 +”、“数据要素 ×”等金融科技创新试点项目为抓手，发挥科技创新示范引领作用。立足证券信息技术研究发展中心（上海）平台优势，持续推动技术共研、成果共享，深化协同创新机制，做优做大科技生态圈。利用国家级科研平台推进科技创新 , 对难点共性问题开展攻关，为关键领域发展提供支撑和引领。打造与数智化发展相适应的科技人才队伍，培养人员的创新意识和创新能力，健全人才评价与人才激励机制，营造有利才能发挥的干事氛围和鼓励创新的文化环境。加强宣传引导，做好科技成果推广，营造良好的科技发展舆论环境，打造科技宣传品牌。

02 前沿攻坚

基于新闻大数据的财务报表舞弊识别技术 *

范剑青 [1]、刘庆富 [2]、王泊 [3]、郑凯鑫 [4]

[1] 普林斯顿大学 | [2] 复旦大学 | [3] 上海证券交易所 | [4] 上海交通大学

摘　要： 本研究提出 PeerMeta 财务舞弊识别框架，基于 2001–2022 年中国经济新闻库构建连续型舞弊倾向指标 FSFP，结合财务、治理、市场及同群传染与同群比较因子，采用 19 种分类器堆叠自适应元学习集成。实证表明：召回率达 0.982，新增 FSFP 与同群因子显著提升模型性能，经济价值评估显示更高边际收益，多项稳健性检验证实方法有效，为舞弊预警与监管政策设计提供新思路。

关键字： 财务舞弊；新闻舆情；同群效应；元学习

一、引言

近年来，伴随资本市场的快速发展，财务报表舞弊事件屡见不鲜，典型案例包括康得新、康美药业、獐子岛、乐视网等。此类舞弊行为不仅严重侵蚀投资者利益，还会对市场信心造成持久损害，甚至引发系统性风险。如何及早、准确地识别潜在舞弊行为，已成为学术界与实务界关注的焦点。现有文献在挖掘影响舞弊的内部治理机制、外部审计质量及市场环境等方面取得了丰硕成果，但往往在样本划分与舞弊标签测度上依赖于已确认的事件，如财务重述、监管处罚及审计师非标准意见等断点型指标（Firth, 2011; Armour et al., 2017; Karpoff et al., 2017），易导致样本选择偏差，并难以捕捉潜在但尚未披露的舞弊倾向。

为填补上述空白，本文基于新闻媒体作为监管信息的重要补充，提出了一种连续型舞弊倾向测度 FSFP，并在此基础上构建了丰富的特征集合与集成算法模型，旨在更早期、更准确地识别舞弊风险。具体而言，本文在以下三个方面做出主要贡献：

（1）创新标签测度：首次基于中国经济新闻数据库中 2001–2022 年逾 30 万篇和财务舞弊相关的财经新闻报道，构建同时涵盖“财务”与“舞弊”术语并伴随负面情感的报道段落加权频次指标，形成连续型 FSFP，通过区分潜在与显性舞弊行为，提高早期预警灵敏度。

（2）丰富特征构建：在传统财务比率、公司治理与市场估值因子的基础上，利用年度报告“公司业务概要”文本，采用 Doc2Vec 提取公司业务向量，基于余弦相似度构建行业无向网络，并设计同群传染因子与同群差异因子，揭示同群竞争压力与社会学习对舞弊决策的影响。

（3）多算法元学习：集成 19 种主流分类算法，采用堆叠泛化与自适应分类两种元学习策略，通过多折交叉验证优化超参数，实现模型性能的显著提升，召回率达到 0.982，F1 值、AUC 均优于基准模型。

* 本文是项目下设课题“异常波动传导与异常交易风险识别预警”（课题编号：2021YFC3340703）的研究成果，课题负责人：刘庆富（复旦大学）；本文部分相关成果已被《Management Science》录用。

二、文献回顾

传统舞弊检测研究多依赖于断点型二值指标，如企业财务重述、监管处罚及审计师非标准意见等，用以标记舞弊公司。然而，重述与处罚往往滞后于舞弊行为发生数年，且仅覆盖已被查实案例，易导致样本选择偏差，降低检测模型的泛化能力。此外，少数学者尝试采用司法文件或内部举报信息，但数据来源不具普遍性。

近年来，随着大数据与文本挖掘技术的发展，研究者开始利用新闻舆情、社交媒体及分析师评论等非财务信息构建舞弊预警模型。Dong et al. (2018) 利用美国股票财经平台帖子情感构建新闻情绪指数，对财务舞弊风险进行预测；Li et al. (2022) 结合微博文本与财务指标，实现了对中国上市公司舞弊行为的早期预警。然而，现有研究多停留于基于整体新闻情感的静态衡量，缺乏针对舞弊主题的高精度术语集构建与连续变量测度。

行为决策理论与社会学习理论指出，企业行为不仅受到自身治理结构与市场环境的影响，也会受到同群行为的示范与传染效应（Yu et al., 2015; 陆蓉和常维，2018；Foroughi et al., 2022）。部分学者基于行业分类构建同群平均财务比率，发现高同群平均杠杆率会增加企业风险偏好（Chen et al., 2022）。但这些研究通常依赖于静态行业划分，忽略企业间业务多样性与动态竞争关系。Doc2Vec 等文本嵌入技术为构建基于业务相似度的动态行业网络提供了可能，但在舞弊检测领域尚未得到充分应用。

元学习方法因其抗噪能力与性能稳定性，广泛应用于金融风险预测（Abbasi et al., 2012；Fan et al., 2023; 张学勇和施懿，2023）。Stacking、Boosting 与 Bagging 等算法已被引入违约

预测与股价波动研究中。但针对舞弊检测场景的元学习研究相对较少，多数仅采用单一算法或简单集成，未能充分利用算法多样性与特征异构性。本文通过引入适应性学习框架，实现了不同分类器权重的动态调整，进一步提升了模型的灵活性与鲁棒性。

三、基于新闻舆情的财务舞弊倾向测度

3.1 数据来源与预处理

本文利用中国经济新闻数据库中 2001–2022 年间共计 30 万余篇财经新闻，筛选包含中文关键词“财务”“舞弊”的初始文章。通过 TF-IDF 与 Word2Vec 技术，构建财务与舞弊两组术语集，包含多个高相似度关键词，并由研究团队进行人工校验。最后，对筛选后的文章按照年度与公司维度进行标注，剔除与舞弊无关的噪声报道。表 1 展示了最终构建的财务舞弊术语集。

表 1 财务舞弊术语集

术语类型		术语集
财务舞弊		财务造假，财务欺诈，财务舞弊，虚增利润，虚增收入，披露不实，虚假记载，虚假陈述
财务+舞弊	财务	财务，会计，审计，税款，账目，资产，负债，收入，利润，业绩，偿付，现金流量，预算，开支，经营，债务，资金，信贷，费用，融资，投资，报表/财报，披露
	舞弊	造假/作假，涉嫌，指控，失实，爆出，虚假，弄虚作假，捏造，炒作，不实，隐瞒，利益输送，爆料，起底，查出，不属实，篡改，黑幕，舞弊/欺诈，疑点，违反，虚报，揭露，纠纷，举报，偷工减料，虚增，虚减

3.2 FSFP 指标构建

在同一公司年度报告披露期内，统计新闻段落中同时包含财务与舞弊术语且 LSTM 情感分析得分小于 0 的文本数量，以加权词频（ATF-IIF）进行度量，并除以该期总新闻段落数，形成连续型 FSFP 指标。该指标取值范围为 [0,1]，数值越高表示该公司当期潜在舞弊倾向越大。

$$FSFP_{it} = \frac{1}{N_{it}} \sum_{j=1}^{M_{it}} (ATF \times IIF)_{itj}$$

其中，$(ATF \times IIF)_{itj}$ 是公司在时期 t 第 j 篇与财务舞弊相关的新闻报道中的财务舞弊术语的 ATF-IIF，M_{it} 是公司 i 在时期 t 财务舞弊相关新闻报道的总数，N_{it} 是公司 i 在时期 t 所有新闻报道的数量，无论是否与财务舞弊相关。ATF 是文章中术语集 K 中所有术语的聚合频率，即 $ATF = \sum_{k \in K} TF_k$ ，它反映了新闻报道对财务舞弊的重视程度。另一方面，IIF 是新闻报道中提及的个别公司数量除以当年公司总数的倒数，即 $IIF=1/(\#firm_{tj}/\#firm_t)$，因此讨论过多公司的新闻报道权重较低。引入 N_{it} 是为了减少公司规模或技术发展等内生因素对新闻报道数量的影响。

表 2 展示了 FSFP 与传统断点指标在样本覆盖率与描述性统计上的对比。Penalty、Restate 和 NSAudit 分别表示证监会行政处罚、财务重述和非标准审计意见。FSFP 在 77.5% 的公司样本中为非零值，均值为 0.389，标准差为 1.385；而重述与处罚指标仅覆盖约 30% 的样本，且分布高度偏态。

表 2 财务舞弊测度指标的描述性统计

测度指标	观察值	非零值	公司覆盖率	均值	标准差
FSFP	56213	21204	0.7752	0.389	1.385
$FSFP_{non\text{-}zero}$	21204	21204	0.7752	0.708	1.895
$FSFP_{winsorized}$	56213	21204	0.7752	0.361	0.826
Penalty	56213	1349	0.1553	0.024	0.154
Restate	56213	7139	0.4165	0.127	0.333
NSAudit	56213	1855	0.1426	0.033	0.179
测度指标	最小值	25%分位数	中位数	75%分位数	最大值
FSFP	0	0	0	0.415	46.092
$FSFP_{non\text{-}zero}$	0.12	0.183	0.348	0.882	46.092
$FSFP_{winsorized}$	0	0	0	0.415	5.943
Penalty	0	0	0	0	1
Restate	0	0	0	0	1
NSAudit	0	0	0	0	1

四、元学习模型架构

检测模型的输入变量包括特征集和标签。特征集由 605 个财务舞弊风险因子构成（39 个基础因子、4 个同群传染因子、39 个同群比较因子、39 个时间趋势因子、484 个季度因子）。其中，基础因子的选择参考了 Song et al. (2014) 和 Lin et al. (2015) 的研究，包括 15 个财务因子、13 个治理因子、8 个市场因子和 3 个情绪因子。定义同群传染因子为同一社区内其余公司 FSFP 的加权平均值，衡量同群中舞弊倾向的集群效应；定义同群比较因子为公司自身 FSFP 与社区平均 FSFP 之差，用以捕捉相对偏离程度。群体的识别以年度报告“业务概览”段落为文本输入，应用 Doc2Vec 模型提取每家公司的 300 维向量表示，基于余弦相似度构建无向加权网络，并采用社区发现算法确定。同群传染因子与比较因子共同反映行业竞争与差异化决策对舞弊风险的影响。

为缓解高维数据的多重共线性问题，采用 Zhou et al.（2024）提出的 FarmPredict 模型，将这 605 个因子分解为 k 个潜在因子与 605 个特异性因子，并以此构建最终的 k + 605 维特征集。样本标签为基于新闻覆盖的 FSFP 值划分的三分类变量：

- 非舞弊（label = 0）：FSFP = 0，约占样本的一半以上；
- 低风险舞弊（label = 1）：FSFP 在非零样本中低于中位数，约占 25%；
- 高风险舞弊（label = 2）：FSFP 在非零样本中高于中位数，约占 25%。

本研究的检测模型称为 SG-AL，包含两层结构：（1）底层由 19 种常用分类器构成（见表 3），涵盖线性 / 二次判别分析（LDA、QDA）、Logit/Probit、朴素贝叶斯、

决策树（ID3、C4.5、CART）、RUSBoost、前馈神经网络、支持向量机、多项式 / 径向基核 SVM，以及多种度量学习与最近邻方法；（2）顶层结合了堆叠泛化（Stacked Generalization）与自适应学习（Adaptive Learning）两种元学习算法，分别用于整合底层分类器预测结果并反馈测试集最确定性预测，以持续优化模型。为了减少过拟合，促进模型更平衡，使其更好地泛化到样本外数据，本文在元学习算法中引入了 L2 正则化。

财务舞弊预测采用动态方式（滚动窗口预测），对每个预测年份 t，取前 5 年（即 t − 5 到 t − 1）的样本数据作为训练集，确保模型能学习最近五年的财务与新闻特征；图 2 总结了本文的检测框架流程图。通过这一滚动预测方式，论文能够检验模型在不同年份的稳定性与前瞻性，为中国 A 股财务舞弊监测提供具有实务价值的年度预警指标。

五、实证结果与性能评估

模型将样本分为“非舞弊”“低舞弊风险”“高舞弊风险”三类，分别计算各组的精确率（Precision）和召回率（Recall），并以整体准确率（Accuracy）衡量综合预测效果。采用当年样本为测试集，滞后 1 至 5 年样本为训练集，测试期自 2006 年至 2022 年。结果显示，高舞弊风险组平均召回率达 0.927，意味着 1000 起舞弊案例能识别 927 起；非舞弊组平均精确率为 0.948，表示预测为无舞弊的 1000 家公司中，仅有 52 起误判；整体平均准确率 0.823，表明模型稳定可靠。此外，低舞弊风险组的精确率和召回率分别为 0.696 和 0.86，进一步验证了模型在不同风险等级下的预测能力。这些指标共同说明，模型在识别舞弊行为方面具有较高的准确性和稳定性。

进一步通过逐一移除风险因子和算法组件，比较性能变化，并以热力图呈现：蓝色代表移除后性能下降，红色则相反。所有组件对准确率均有正向贡献，其中季度因子和自适应学习算法效果最显著；因子类对召回率的提升整体优于算法类，部分算法因过度追求精确率反而降低召回。

金融因子虽能提升准确率与精确率，却显著削弱召回率，原因在于违规者对财务数据的巧妙操控使模型难以捕捉舞弊痕迹。相较之下，同业传染因子创新性强：移除后高、低舞弊组召回率均下降约 10%，尤其对轻度舞弊影响更大，说明行业内的传染效应有助于发现潜在或隐蔽舞弊。此外，同业传染因子与同业比较因子互为补充：前者提升低风险舞弊样本识别，后者对高风险样本更敏感，两者结合增强

表 1 底层分类器说明

大类名称	子类名称	子类符号	说明
判别分析	线性判别分析	LDA	多元正态分布的均值不同但协方差矩阵相同
	二次判别分析	QDA	多元正态分布的均值和协方差矩阵均不同
定性响应回归	Logit 回归	Logit	因变量服从逻辑分布函数
	Probit 回归	Probit	因变量服从累积正态分布函数
贝叶斯分类	朴素贝叶斯	NaiveBayes	基于贝叶斯定理
决策树	ID3 决策树算法	Tree-ID3	基于绝对信息增益计算分类结果不确定性
	C4.5 决策树算法	Tree-C4.5	基于相对信息增益计算分类结果不确定性
	CART 决策树算法	Tree-CART	基于 GINI 指数计算分类结果不确定性
人工神经网络	单层人工神经网络	ANN1	包含一个隐藏层
	双层人工神经网络	ANN2	包含两个隐藏层
	三层人工神经网络	ANN3	包含三个隐藏层
支持向量机	线性支持向量机	SVM-Lin	核函数为线性函数
	多项支持向量机	SVM-Poly	核函数为多项式函数
	RBF 支持向量机	SVM-RBF	核函数为 RBF 型函数
距离度量学习	K 最邻近算法	KNN	用欧式距离度量两点距离
	近邻成分分析	NCA	随机选择邻近点度量两点距离
	信息距离度量学习	ITML	基于信息熵计算两点距离
	大边界最近邻算法	LMNN	用马氏距离度量两点距离

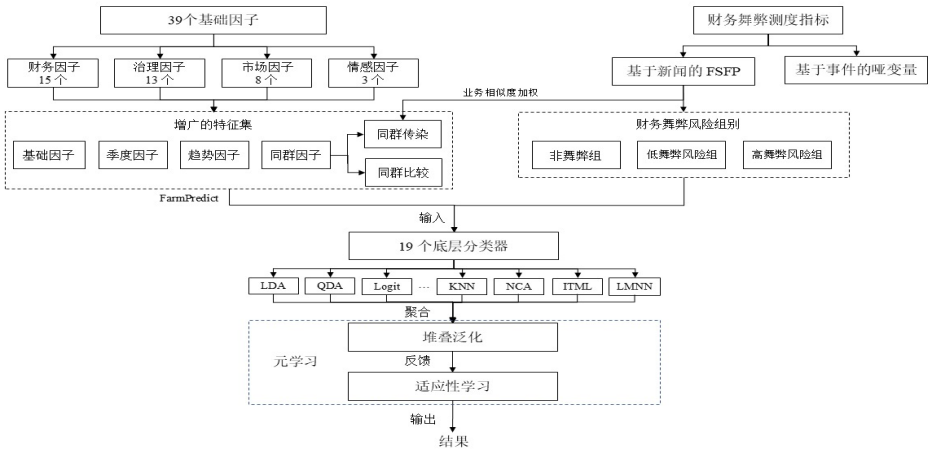


图 1 财务舞弊识别流程图

表 3 样本外识别结果

年度	非舞弊		低舞弊风险		高舞弊风险		总样本
	Prec	Rec	Prec	Rec	Prec	Rec	Acc
2006	0.934	0.713	0.684	0.898	0.769	0.9	0.813
2007	0.921	0.709	0.711	0.896	0.729	0.882	0.806
2008	0.946	0.747	0.674	0.821	0.762	0.923	0.817
2009	0.933	0.751	0.687	0.873	0.844	0.949	0.839
2010	0.976	0.781	0.733	0.843	0.753	0.947	0.846
2011	0.976	0.748	0.699	0.881	0.745	0.908	0.828
2012	0.964	0.711	0.725	0.905	0.729	0.939	0.823
2013	0.939	0.692	0.723	0.884	0.719	0.944	0.81
2014	0.925	0.744	0.718	0.823	0.745	0.934	0.819
2015	0.924	0.782	0.681	0.808	0.835	0.94	0.836
2016	0.973	0.712	0.68	0.886	0.744	0.927	0.816
2017	0.954	0.712	0.638	0.826	0.73	0.892	0.793
2018	0.96	0.731	0.714	0.844	0.732	0.956	0.823
2019	0.949	0.763	0.724	0.889	0.758	0.89	0.834
2020	0.915	0.728	0.679	0.857	0.817	0.944	0.821
2021	0.973	0.747	0.671	0.856	0.798	0.955	0.834
2022	0.946	0.77	0.696	0.838	0.789	0.93	0.834
平均值	0.948	0.738	0.696	0.86	0.764	0.927	0.823

了模型的全面检测能力。

总体而言，该框架通过多维度因子与多种算法的有机融合，实现了高水平的准确率、精确率和召回率平衡，为财务舞弊识别提供了切实可行的技术方案。未来可在保持因子层面挖掘深度的前提下，优化算法配置，并强化同业网络效应，以进一步提升真实舞弊行为的发现率。

六、结论与政策建议

本研究提出基于新闻报道的连续舞弊倾向指标 FSFP，并结合文本嵌入构建同群因子，采用多算法元学习，实现了对财务报表舞弊的高效检测。实证结果表明，新框架在召回率与经济价值方面均取得显著提升，为学术界与实务界提供了新的测度思路与实用工具。

对企业治理而言， FSFP 指标可作为董事会、审计委员会的预警工具，帮助企业识别内部控制薄弱环节；同群因子可为治理层提供行业对标分析，优化风险管理流程。对监管政策而言，监管机构可将新闻覆盖型连续指标纳入信息披露监管体系，鼓励媒体加强对疑似舞弊行为的舆论监督；可考虑构建基于新闻监测的动态监管白名单与高风险名单。

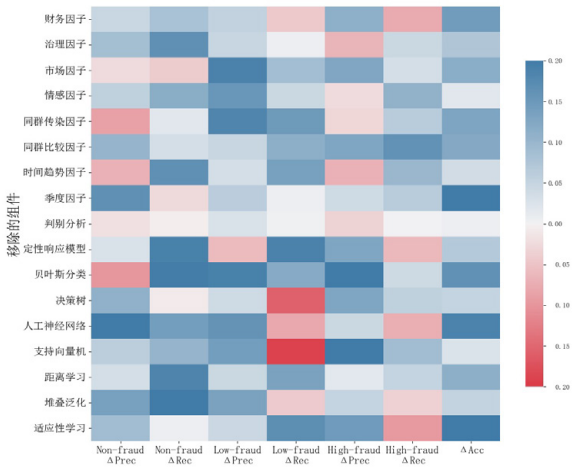


图 2 识别框架各组件对模型性能的贡献度

参考文献：

[1] Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. MIS Quarterly, 1293-1327.

[2] Armour, J., Mayer, C., & Polo, A. (2017). Regulatory sanctions and reputational damage in financial markets. Journal of Financial and Quantitative Analysis, 52(4), 1429-1448.

[3] Chen, Q. (2022). Relationship between Financial Asset Allocation, Leverage Ratio, and Risk - Taking of Small - and Medium - Sized Enterprises in China: Taking Environment - Related Industries as an Example. Journal of Environmental and Public Health, 2022(1), 2431428.

[4] Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. Journal of Management Information Systems, 35(2), 461-487.

[5] Fan, J., Liu, Q., Wang, B., & Zheng, K. (2023). Unearthing financial statement fraud: Insights from news coverage analysis. Available at SSRN 4338277.

[6] Firth, M., Rui, O. M., & Wu, W. (2011). Cooking the books: Recipes and costs of falsified financial statements in China. Journal of Corporate Finance, 17(2), 371-390.

[7] Foroughi, P., Marcus, A. J., Nguyen, V., & Tehranian, H. (2022). Peer effects in corporate governance practices: Evidence from universal demand laws. The Review of Financial Studies, 35(1), 132-167.

[8] Karpoff, J. M., Koester, A., Lee, D. S., & Martin, G. S. (2017). Proxies and databases in financial misconduct research. The Accounting Review, 92(6), 129-163.

[9] Li, J., Yu, L., Mei, X., & Feng, X. (2022). Do social media constrain or promote company violations?. Accounting & Finance, 62(1), 31-70.

[10] Lin, C. C., Chiu, A. A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. Knowledge-Based Systems, 89, 459-470.

[11] Song, X. P., Hu, Z. H., Du, J. G., & Sheng, Z. H. (2014). Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China. Journal of Forecasting, 33(8), 611-626.

[12] Yu, X., Zhang, P., & Zheng, Y. (2015). Corporate governance, political connections, and intra - industry effects: Evidence from corporate scandals in China. Financial Management, 44(1), 49-80.

[13] Zhou, Y., Fan, J., & Xue, L. (2024). How much can machines learn finance from Chinese text data?. Management Science, 70(12), 8962-8987.

[14] 陆蓉 , 常维 . 近墨者黑 : 上市公司违规行为的“同群效应”[J]. 金融研究 , 2018, (8): 172-189.

[15] 张学勇 , 施懿 . 基于元学习的财务舞弊识别研究 [J]. 管理科学学报 , 2023, 26(10):95-113.

基于检索增强的智能研报生成技术 *

梁佳艺 [1]，岑黎彬 [2]，王晓玲 [1]，吴苑斌 [1]

[1] 华东师范大学 | [2] 维沃移动通信（深圳）有限公司

摘　要： 根据公司的财报、股市信息和舆情信息撰写个股分析研报是金融研究员重点关注的问题。基于大模型的研报生成技术难以解决从长篇幅公司财报中检索关键信息的问题，导致研报撰写的质量较差。因此，本文提出基于检索增强的金融研报生成技术。具体来说，使用检索增强技术从财报中提取有价值的信息，使用经过微调的大语言模型进行研报大纲的撰写，根据大纲再次检索进行信息检索，分段撰写个股分析研报。在真实数据上的实验表明，本文提出的方法能够提升研报的可读性和可解释性，能够显著降低研报撰写的人工成本，提升研报撰写效率。

关键字： 金融研报撰写；检索增强技术；大语言模型

一、引言

中国金融市场具有巨大的潜力。在投资需求不断增长的环境下，如何依托高精度、实时更新的市场行情信息，实现自动化个股分析报告的生成，已成为金融领域的重要研究课题。在海量异构金融数据（如财报、舆情新闻、交易数据）快速更新的背景下，如何基于高精度、实时化的数据实现个股分析研报的自动化生成，已成为金融科技研究中的重要课题。如何从长篇幅的公司财报中提取关键信息，并从海量市场数据中筛选出重要内容。

研报生成算法面临异构信息的高效检索与融合和研报结构的多样性与生成内容的专业性两个挑战。首先，个股分析所依赖的信息来源包括结构化的公司财务报表和非结构化的舆情新闻，二者在数据格式、语义结构和关注重点上存在显著差异。如何从长篇幅、冗余度高的文本中高效检索出与股价波动密切相关的片段，是研报自动生成的基础难题。其次，研报结构的多样性与生成内容的专业性：个股分析研报通常包括公司概况、行业趋势、财务分析和未来展望等模块，内容组织形式复杂、风格高度专业，缺乏统一模板。这对自动文本生成系统在篇章结构规划、信息组织及语言风格控制方面提出了巨大挑战。

为了解决上述挑战，本文提出了一种基于检索增强的研报生成算法，该方法通过将大语言模型与检索增强模型灵活应用，分别应对文本检索和研报撰写过程中的关键问题。具体而言，首先将网页上的新闻和财报数据标准化为统一的文本格式，并利用嵌入模型构建向量数据库。随后，通过研报信息对财报和新闻内容进行检索增强训练，使得模型能够基于关键词检索出相关的财报和新闻片段，从而提高信息获取的准确性。研报撰写模块将研报生成拆分为研报大纲生成和分段落撰写两个任务，首先根据经过微调的大语言模型生成的研报大纲，在存储财报信息的向量数据库中再次进行检索，再根据检索的内容进行分段落研报撰写。最终，通过格式整理与内容整合，生成完整的个股分析研报。

本文的主要贡献如下：

1. 面向财报和新闻的检索增强策略：构建向量数据库，以实现对财务报告和新闻文本的高效检索。通过关键字匹配与向量相似度计算，精准提取与目标主题相关的内容，提高数据利用效率。

2. 实现结构引导的研报分阶段生成算法：以生成大纲为引导，结合分节检索与微调生成，实现研报结构清晰、逻辑连贯的内容组织方式。

* 本文是项目下设课题“开放域舆情风险识别、预警与处置”（课题编号：2021YFC3340702）的研究成果，课题负责人：王晓玲（华东师范大学）。

二、检索增强技术

向大语言模型添加检索增强模块，可有效缓解其在文本生成过程中的幻觉现象 [1]。根据输入内容检索文档，再基于检索到的文档生成完整回答的检索增强方法在文档生成和事实性回答 [2] 等方面都取得了优秀的表现。在进行长文本的检索增强时，大模型需要从文档中提取更多的知识，现有的方法通常将一个问题拆分为多个子问题分别进行检索 [3]。为了主动决定检索的时间以及检索内容，前瞻性主动检索增强生成模型 FLARE 定义了基于置信度的查询策略 [4]，在大语言模型生成低概率标记时进行检索。在金融文本生成领域，为了让模型掌握金融知识并具备金融推理能力，TFR 使用基于金融知识图谱的检索增强框架 [5]，采用两阶段检索方式，第一阶段利用金融知识图谱进行初

始信息选择，第二阶段基于所选信息进行推理，生成金融市场分析报告。

三、基于检索增强的研报生成技术

3.1 问题定义

研报生成任务的定义如下：$\mathcal{D}\text{report} = (X_i, Y_i)_{i=1}^{N}$ 表示研报生成的数据集，其中 X_i 为输入信息集，Y_i 为对应的目标研报文本。$X_i = \{F_i, N_i, C_i\}$，分别表示第 i 个样本的输入数据：F_i 表示公司财报文本；N_i 为相关新闻文本；C_i 为公司基本信息，如行业、市值、成立时间等；$Y_i = S_i^1, S_i^2, \ldots, S_i^K$ 表示目标研报文本，由多个子段落组成，其中每个 S_i^K 表示研报的一个组成部分，如公司概况、财务分析、舆情解读等。

研报生成模型的目标是学习一个条件生成函数：

$$\mathcal{G}_\theta : X_i \longmapsto Y_i$$

其中 G_θ 是参数为 θ 的生成模型，用于从多源输入 X_i 中建模文本结构与语义关系，并生成完整的研报 Y_i。

3.2 方法

本文提出一种基于检索增强的研报生成算法，架构如图 1 所示。该架构主要由两个模块组成：（1）检索增强模块首先将多模态的输入信息转换为统一的文本格式，并构建向量数据库，通过检索增强技术检索与个股研报具有强相关性的舆情和财报片段；（2）研报撰写模块结合大语言模型获得的知识与检索到的财报信息，生成个股分析研报的大纲并分阶段撰写。

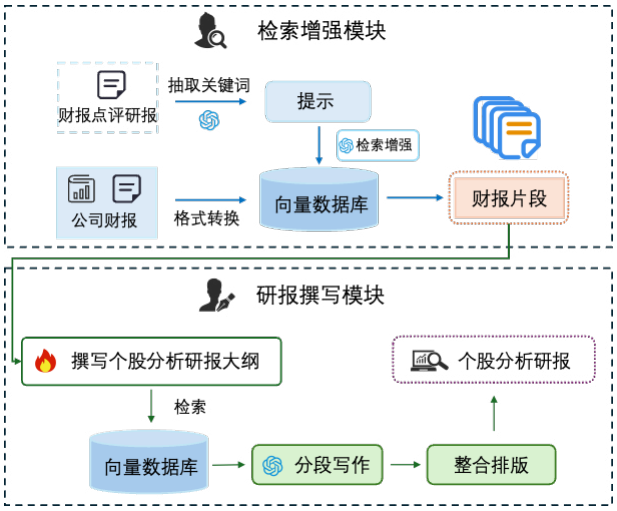


图 1 基于检索增强的研报生成算法结构图

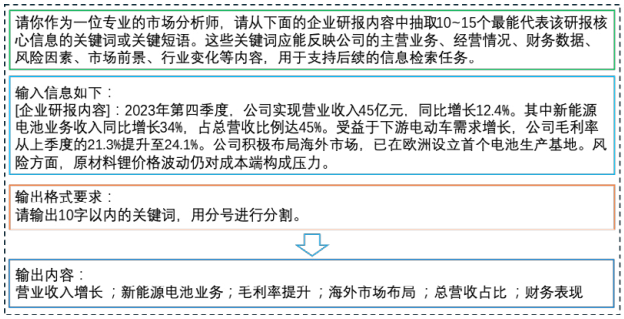


图 2 小样本的提示词示例

3.2.1 检索增强模块

检索增强智能体实现的主要任务包括将多种形式的输入内容转换为文本格式，将文本信息向量化存储进向量数据库，根据小样本的个股分析研报构造向量检索的提示词，对新闻文本和公司财报进行检索。

对企业财报的解析和分析首先需要解决不同模态数据的解析问题。财报中大量数据来自表格，虽然可以利用 PDF 解析工具从中将表格解析成结构化文本，但是现有开源的小参数模型（7 - 14 B），对解析后的纯表格文本十分不敏感，容易在计算表格中的数据时出现错误。为了解决这个问题，本文在 Qwen2.5 模型 上对用自然语言描述表格的任务进行微调训练，让模型具备理解和描述表格的能力。此外，企业财报数据的篇幅很大，无法直接作为大语言模型的输入，因此需要考虑长文本分割的方法。本文使用 Faiss 向量数据库架构对分割后的文本构建向量索引。

在检索增强阶段，本文提出了一种小样本的提示词优化生成策略，如图 2，以实现对公司财报和舆情信息的高效内容检索。具体而言，考虑到不同公司个股分析研报数量有限，直接依赖大模型进行关键词提取易受到内容分布偏差的影响，本文设计了一种小样本微调机制，使模型能够从有限的研报样本中学习关键词构建策略。

我们将每篇研报样本表示为 $\mathcal{D} = \{(x_i, k_i)\}_{i=1}^{N}$，其中 x_i 表示第 i 篇研报的正文内容，k_i 表示与该研报对应的目标关键词集合，用于检索相关财报段落。在训练阶段，我们使用一个基于指令微调的文本生成模型 M_θ，优化目标如下：

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_{\text{CE}}(\mathcal{M}_\theta(x_i), k_i)$$

其中 L_{CE} 为交叉熵损失，衡量模型输出关键词与真实关键词集合之间的差异。

在生成阶段，模型根据当前任务目标生成提示词（prompt）形式如图 2。这些关键词将被进一步嵌入向量数据库查询接口，用于检索结构化财报数据或舆情段落，以实现信息增强的目标。

此外，为提升关键词的语义表达能力，本文引入了词向量监督机制，即将模型生成的关键词k_i'与目标片段中实际出现的关键词向量 v_j 进行相似度对齐，辅助损失函数如下：

$$\mathcal{L}_{sim} = \sum_{i=1}^{N} \sum_{k \in K_i'} \max_j (1 - \cos(k, v_j))$$

[1] https://github.com/QwenLM/Qwen2.5

最终的训练目标为联合优化：

$$\mathcal{L}_{total} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{sim}$$

其中 λ 为权重系数，控制语义对齐损失的参与程度。

我们将通过微调模型生成的关键词集合作为查询输入，分别用于对公司财报文本和舆情新闻片段进行相关性检索。为实现高效的语义级匹配，本文采用向量化表示与相似度计算的方法，将关键词与待检索文本统一映射到同一语义空间中。具体而言，设生成的关键词集合为$\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$，其中每个关键词 q_i 被编码为向量$e_{q_i} \in R^d$，同时构建包含所有待检索片段的向量数据库$\mathcal{C} = \{c_j\}_{j=1}^{M}$，其中每个$c_j \in R^d$表示一段财报或新闻文本的向量表示。

在实际检索过程中，我们采用如下余弦相似度公式计算关键词向量与片段向量之间的相似度分数：

$$\text{sim}(q_i, c_j) = \frac{e_{q_i} \cdot c_j}{|e_{q_i}| \cdot |c_j|}$$

然后对每类文本（如财报和新闻）分别进行相似度排序，选取得分最高的前 K_1 财报片段和前 K_2 个新闻片段，分别作为对应的检索结果输出，用于后续的股价预测与研报撰写模块。最终的检索结果可表示为：

$$\mathcal{R}_{\text{rpr}} = \text{TopK}_{K_1}\left(\text{sim}(\mathcal{Q}, \mathcal{C}_{\text{rpr}})\right), \quad \mathcal{R}_{\text{nw}} = \text{TopK}_{K_2}\left(\text{sim}(\mathcal{Q}, \mathcal{C}_{\text{nw}})\right)$$

其中 C_{rpr} 和 C_{nw} 分别表示财报和新闻构建的向量集合。通过这种检索增强机制，我们能够从冗余的大规模原始文本中筛选出与分析目标高度相关的信息，为下游推理和生成提供更加精准且上下文相关的知识支持。

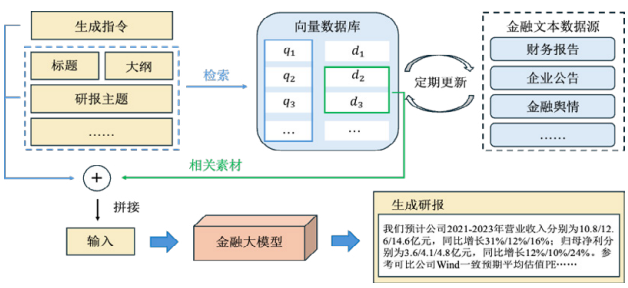


图 3 研报撰写具体实现图

3.2.2 研报撰写模块

在研报撰写阶段，本文使用两阶段撰写的思路，首先撰写个股分析研报的大纲，接着根据大纲的内容再次检索后分段创作，如图 3 所示。

第一阶段为结构性规划阶段，根据少量标注数据对基座模型进行微调，提升模型对研报篇章结构的理解。根据检索增强智能体从向量数据库中检索到的信息构建研报的大纲结构，包括但不限于“公司概况”“行业环境”“利好利空因素分析”与“未来股价预测”四个板块。该大纲作为后续内容撰写的结构性提示，有助于提升生成内容的条理性与逻辑一致性。

第二阶段为内容生成阶段，系统根据每一节大纲中的子主题，利用其对应的提示在向量数据库中进行进一步的定向语义检索，从而获得与当前小节高度相关的新闻与财报信息（记为 $\mathcal{R}_i = \{r_{i1}, r_{i2}, \ldots, r_{iK}\}$）。模型基于这些检索片段 R_i 及其对应的大纲主题 T_i，构造增强型提示词：

$$\text{Prompt}_i = \text{“请基于以下信息撰写关于}T_i\text{的小节内容：”} + \mathcal{R}_i$$

随后，大语言模型（如 LLaMA 或 FinGPT）在微调后的权重基础上，对每一小节内容进行生成，过程记为：

$$\text{Paragraph}_i = \text{LLM}_{\text{finetuned}}(\text{Prompt}_i)$$

所有生成的段落 {{Paragraph}_1,{Paragraph}_2,...,{Paragraph}_n} 通过排版与整合模块进行统一格式化，最终形成具备专 Chinese LLaMA - 7B Chinese LLaMA - 7B 业水准的个股分析研报文档。

四、实验与分析

4.1 数据来源

本文使用东方财富网上沪深 300 的个股对应的企业财报、舆论数据以及点评研报，作为研报生成的来源数据。在微调检索增强的提示词阶段，本文使用了 100 条人工标注的研报，对大语言模型进行了微调。在研报大纲撰写阶段，文本使用了继续使用 100 篇人工标注过的研报进行大语言模型的微调。在研报的分段落撰写阶段，文本针对研报段落撰写和小标题撰写两个子任务，分别使用了 400 条和 600 条标注过的内容进行微调。

4.2 基准模型与评估指标

为了确保实验的准确性，本文引入 4 个基准模型用于对比股价波动预测任务，引入两个大语言模型作为基准模型用于对比研报生成任务的文本质量。基线大模型的介绍

如下：

• GPT-3.5 Turbo-16k 模型：使用思维链提示，在模型上进行零样本的测试，通过调用 API 的方式调用该模型。此外，使用上下文学习的方式，提供财报片段的示例，调用模型执行研报大纲生成和研报分段撰写的任务。

• FinGPT 模型：该模型是一个端到端的金融大语言模型，在研报大纲生成阶段和研报分段生成阶段，对模型进行微调训练。

本文在对撰写的个股分析研报进行评估时使用文本质量自动评估的困惑度指标（PPLscore）对文本质量进行客观评估。

4.3 个股分析研报评估

本文使用困惑度指标对生成文本的语言质量进行了评估。困惑度指标是衡量语言模型生成文本自然性和流畅性的常用指标，其值越低，表示模型生成的语言序列越符合自然语言分布，即文本的连贯性与语法合理性越强。表 3 展示了 FIRA 与两个基线模型在困惑度指标上的表现结果。图 4 为研报撰写智能体生成的完整的个股分析研报示例。

表 1 FIRA 与基线模型生成个股分析研报的文本质量评估

模型	GPT-3.5	FinGPT	本文方法
PPL 分数	96.2	87.6	85.3

金山办公(688111 CH): LLM重要落地场景

LLM持续投入，产品智能化体验不断提升

2022年公司收入38.85亿元，同增18.4%，归母净利11.18亿元，同增7.3%，扣非净利9.39亿元，同增11.7%；22Q4收入10.90亿元，同增20.0%，归母净利3.04亿元，同增57.4%，扣非归母净利2.76亿元，同增100.3%。22年公司利润增速慢于收入主要由于研发投入加大，22年研发投入13.31亿元，同增23.1%，占收入比34.0%，研发人员2922人，占员工总数68.0%。22年公司LLM围绕AIGC及LLM领域持续投入，产品智能化体验不断提升。

2C业务：用户基础持续扩大，增值服务持续丰富

2022年，公司订阅业务收入27.4亿元，同增43.5%，占总收入比71%，同增12pct。截至2022年12月31日，公司主要产品月活设备数5.73亿，同增5.3%，其中WPS Office PC版月活2.42亿，同增10.5%，WPS Office移动版月活3.28亿，同增2.2%。截至2022年12月31日，公司累计年度付费个人用户数达2,997万。2022年，公司针对用户需求研发并上线新功能，WPS会员增值功能活跃度持续增长，用户口碑不断提升。同时新推出了图片处理服务、票据服务、多端音频转写、视频编辑压缩处理、文档对象批量处理等多项会员特权，备受用户欢迎。2022年，公司针对团队用户新增了70多项服务应用，涵盖协同办公、生产供应链、人事行政、客户管理、店铺运营等场景，上线低代码平台，为小型团队的业务全流程管理提供轻应用服务。

2B业务高速增长，机构订阅及服务收入同增55.06%

2022年，公司2B业务高速增长，国内机构订阅及服务收入6.92亿元，同增55.06%，国内机构授权收入8.36亿元，同减13.18%。2022年，公司数字办公平台收入同增57%，数字办公产品新增政企客户3990家，带动国内机构订阅及服务高速增长。2022年，公司公有云SaaS在期付费企业数同增51%，付费企业续约率超70%，金额续费率超100%，带动公有云领域相关收入同增超100%。2022年，公司密切关注信创领域政策变化及客户需求，提前布局各地下沉市场及行业信创业务，机构客户因正版化需求对公司产品采购持续增加，抵消了信创订单收缩的部分影响。

AI技术持续迭代，产品智能化程度不断提升

2022年，金山办公利用AI能力帮助用户校对总字数达3,340亿个，全年OCR处理图片数量达146亿份，智能美化功能月活跃用户数量达237万。2022年9月，公司在第三届CSIG图像图形技术挑战赛中获得单项赛道及总决赛双冠军。在文档校对方向，实现了金山办公和黑马校对的双引擎整合，成为中文校对领域的佼佼者。报告期内，公司不断完善算法能力，优化文档数据搜索功能，持续推进产品智能化进程。同时，积极探索结合AIGC、LLM技术的下一代人机交互体验，并在文档翻译、听读、中英文校对、语音及音频转写、智能辅助写作及排版、表格数据智能分析、PPT一键生成及美化等场景进行融合，为用户提供便捷优质的智能化服务体验。

图 4 使用基于检索增强的研报生成技术生成的研报示例

通过分析实验结果，我们发现研报撰写智能体在 PPL 指标上取得了最低分（85.3），优于 GPT-3.5（96.2）和 FinGPT（87.6），说明其生成的个股分析研报文本在语言流畅性、逻辑结构和语义连贯性方面更为自然。该结果验证了 FIRA 提出的 “两阶段研报生成策略”的有效性：首先通过结构性规划明确研报提纲，再结合检索增强结果进行逐段撰写，有效降低了生成内容的冗余与不一致性。此外，通过将大语言模型面向金融场景的微调，研报撰写智能体更好地适应了个股分析研报中专业术语与因果结构丰富的语言风格。

五、总结与未来展望

为了完成基于异构信息的个股分析研报撰写任务，文本提出了一种基于检索增强的研报撰写算法。该算法利用检索增强技术筛选舆情信息和公司财报，同时，通过大语言模型撰写研报的大纲，并进行财报信息的二次检索，再分段完成研报的撰写。

本文的方法虽然使用了多次检索增强从很长的财报中检索需要的财报片段，但是在基于微调的方法在训练数据不平衡时可能会导致数据集中占比较低的行业研报生成效果不佳。未来的研究可以针对小样本的研报生成进行数据增强工作。

参考文献：

[1] Fan W, Ding Y, Ning L, et al. A survey on rag meeting llms: Towards retrieval-augmented large language models[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024: 6491-6501.

[2] Guu K, Lee K, Tung Z, et al. Retrieval augmented language model pre-training[C]// International conference on machine learning. PMLR, 2020: 3929-3938.

[3] Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models[C]// International Conference on Learning Representations (ICLR). 2023.

[4] Jiang Z, Xu F F, Gao L, et al. Active retrieval augmented generation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 7969-7992.

[5] Chen Y, Wu F, Wang J, et al. Knowledge-augmented Financial Market Analysis and Report Generation[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. 2024: 1207-1217.

金融文本中的信息抽取 *

李小明 [1]，陈艺文 [1]，常思维 [1]，何豪杰 [2]，孙晓飞 [2]

[1] 上证所信息网络有限公司 | [2] 北京香侬慧语科技有限责任公司

摘　要： 随着金融科技的快速发展，金融文档中蕴含着大量的结构化和非结构化信息。有效地从金融文本中抽取关键信息对于风险评估、投资决策和监管合规具有重要意义。本文提出了一个综合性的金融文本信息抽取框架，包含三个核心组件：基于深度学习的文本信息抽取、基于二维注意力机制的表格信息抽取以及智能表头识别系统。通过在真实金融数据集上的实验验证，本文方法在信息抽取准确性和效率方面均取得了显著提升，为金融文档智能化处理提供了有效解决方案。

关键字： 信息抽取，深度学习，注意力机制，金融文本，表格处理

一、引言

1.1 研究背景

金融行业每天产生海量的文档和报告，包括年报、招股说明书、审计报告、研究报告等。这些文档中包含了关键的财务数据、风险信息和业务指标，是投资决策、风险管理和监管监督的重要数据源。传统的人工信息抽取方式效率低下且容易出错，难以满足现代金融业务的实时性要求。

随着自然语言处理和计算机视觉技术的发展，自动化信息抽取技术为解决这一问题提供了新的途径。然而，金融文档具有专业性强、格式复杂、表格密集等特点，对信息抽取技术提出了更高的要求。

1.2 研究挑战

金融文本信息抽取面临以下主要挑战：

1. 文本复杂性：金融文本语言专业、术语繁多、句式复杂，传统 NLP 方法难以准确理解语义。

2. 表格多样性：金融文档中的表格格式多样，包含复杂的表头结构、合并单元格、多层嵌套等，传统的表格处理方法难以适应。

3. 上下文依赖：表格中的数据往往依赖于文本说明和表头信息，需要综合考虑多模态信息。

4. 实时性要求：金融应用场景对信息抽取的速度和准确性都有很高的要求。

1.3 研究贡献

本研究贡献如下：

1. 文本 - 表格双流处理架构：提出了统一的文本 + 表格信息抽取处理架构；

2. 二维空间注意力模型：设计行列分离的位置编码方案与曼哈顿距离衰减函数，使复杂表格的单元关系预测准确率提升至 89.7%

3. 动态表头解析系统：开发基于样式特征与语义聚类的层次识别算法，能实现对复合表头的结构的准确还原。

* 本文是项目下设课题“智能信披审核和监管数据安全共享关键技术研究”（课题编号：2021YFC3340701）的研究成果，课题负责人：周琳娜（北京邮电大学）。

二、相关工作

2.1 文本信息抽取

文本信息抽取主要包括命名实体识别（NER）、关系抽取（RE）和事件抽取等任务。早期方法主要基于规则和统计机器学习，如条件随机场（CRF）等。随着深度学习的发展，基于神经网络的方法逐渐成为主流。

LSTM-based 方法： 长短时记忆网络（LSTM）由于其能够处理长序列依赖关系的特性，在文本序列标注任务中表现出色。Huang 等人提出的 BiLSTM-CRF 模型在 NER 任务上取得了显著效果。

BERT-based 方法： BERT（Bidirectional Encoder Representations from Transformers）通过双向编码器和预训练机制，在多项 NLP 任务上刷新了纪录。Devlin 等人的工作证明了 BERT 在信息抽取任务上的强大能力。

金融领域应用：近年来，研究者们开始将深度学习方法应用于金融文本处理。FinBERT、SecBERT 等专门针对

金融领域预训练的模型相继出现，为金融文本理解提供了更好的基础。

2.2 表格信息抽取

表格信息抽取作为文档理解的重要分支，近年来受到了广泛关注。早期的方法主要基于规则和模板匹配，如Embley等人提出的基于启发式规则的表格结构识别方法。随着机器学习技术的发展，研究者开始采用统计学习方法进行表格理解，如支持向量机、条件随机场等。

深度学习时代的到来为表格信息抽取带来了新的机遇。Chen 等人提出了基于卷积神经网络的表格结构识别方法，通过将表格转换为图像进行处理。Qasemi 等人设计了基于循环神经网络的表格内容理解模型。然而，这些方法往往将表格视为一维序列或二维图像，未能充分利用表格的结构化特征。

三、文本信息抽取

3.1 命名实体识别

命名实体识别又称作专名识别，是信息抽取的重要子任务，其目的是识别文本中具有特定意义的实体，如人名、地名、组织机构名称、日期、专有名词等。

命名实体识别是一项基础性的关键任务，主要包括两部分：实体边界识别、实体类别识别。英语中的命名实体具有比较明显的形式标志，例如实体这种的每个词首字母大写，因此实体边界识别相对容易。而汉语作为象形文字，没有大小写的区分，相比英文等拼音文字来说，针对中文的命名实体识别任务往往更有挑战性。主要包括：

• 中文没有空格这种天然的词语界限标志，命名实体识别的准确率会受到分词结果准确率的影响。例如“南京市长江大桥”，若分词结果为“南京 市长 江大桥”那么分词错误将会传播至命名实体识别。

• 在中文文本中，命名实体常出现嵌套现象，例如“北京大学口腔医院”这一组织机构名中还嵌套着同样作为同样是组织机构名的“北京大学”。

• 现代汉语文本中，常出现中英文交替使用的情况，例如“ARPU 值”中文是“每用户平均收入”。这时汉语命名实体识别任务还要识别英文命名实体。

对于命名实体识别信息抽取系统，项目组使用了基于多轮问答的阅读理解系统，根据预先生成好的问题模板，从文章中抽取指定信息。其中阅读理解系统的输入为文章和问题，输出为从文章中选取的区域，以此作为对问题的回答。课题组使用的阅读理解模型基于预训练模型 BERT，如下图所示。

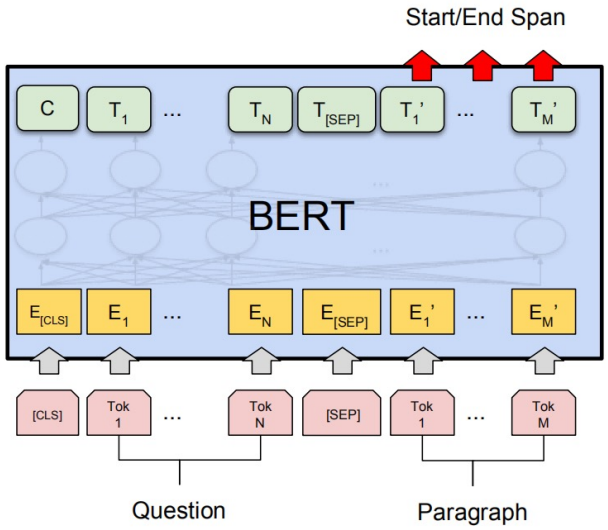


图 1 BERT 模型结构

首先在海量语料上使用 Masked-Language Model 训练得到 BERT，随后在机器阅读理解数据集上微调，预测每个字属于问题回答的起始或结尾，并取概率最高的区域作为最终的回答。

3.2 关系抽取

关系抽取（Relation Extraction, RE）的主要目的是针对抽取得到的命名实体进行语义关系抽取，在两个实体之间建立关系连接，形成关系三元组。相比命名实体识别，关系抽取是一项更有挑战性的任务：

• 关系抽取依赖命名实体识别的输出，如果识别到的实体是错误的，则会将错误传播至关系抽取。

• 关系的语义更复杂，表达两种同一种关系时，往往有多种说法，例如“公司 2018 年产煤量 1000 万吨”和“公司 2018 年生产煤炭 1000 万吨”所抽取出的关系三元组是一样的。

• 关系的自由度更高，在两个实体间往往存在多种关系，例如库克和苹果公司，库克是苹果公司的 CEO，同时库克也持有苹果公司的股票，因此这两个实体间有“CEO”和“持股”两种关系。

• 关系抽取所需要关注的语义片段往往更长，语法结构更复杂。一般来说命名实体识别可能只需要一句话即可对内部的实体进行识别，而关系抽取涉及两个实体，不仅要识别一句话内的关系，还会涉及多句间的关系，往往涉及指代消解（Coreference Resolution, CR），即对代指词找到其所代指的实体。

项目组结合金融中的关系存在语义复杂且种类多变的特点，使用基于深度学习的监督和半监督关系抽取器，在海量

文中进行关系抽取。具体方法描述如下。

基于监督的方法

首先对所需抽取的命名实体和它们之间的关系进行定义，将这个问题建模为一个分类问题，并训练两种分类器，第一个分类器是“是 / 否”二分类器，即判断两个实体是否存在关系。如果存在关系，则再送第二个分类器，即给实体间的关系判断类别。使用 CNN、LSTM，结合注意力机制（Attention）进行关系的分类和抽取。

基于半监督的方法：使用监督的方法，可以得到准确率很高的抽取器，然而其覆盖较为有限。因此项目组通过 Bootstrap、远程监督的方法进行关系抽取。基于 Bootstrap 的方法，利用少量实例作为种子三元组集合，通过不断地迭代从非结构化文本中抽取关系，然后从新学到的实例中学习新的模式集合，寻找和发现新的潜在关系三元组。基于远程监督的方法，是结合金融信息知识图谱的大量知识库信息，构建大量有噪声的训练数据，使用基于 CNN 和注意力机制的方法，并结合强化学习（Reinforcement Learning, RL），从文本中自动进行关系抽取。

机器阅读理解

机器阅读理解（Machine Reading Comprehension, MRC）是 NLP 中的核心通用任务之一。因为机器阅读理解极强的通用性，上述任务均可以视作机器阅读理解的一个特例子任务，因此它也是最为复杂的一个任务。

机器阅读理解可以抽象为以下任务：给定一段文章，并给出一个问题，在机器阅读了文章内容后，针对这个问题，在文章中寻找需要抽取的正确答案。它能够让计算机帮助人类在海量文本中找到想要的信息，从而极大地减轻人们获取信息的成本。

针对金融领域信息抽取中主要为事实类短答案的特点，首先做表示学习，将问题和文章都转化成向量，再利用注意力（Attention）与自注意力（Self-Attention）机制，通过深度学习网络学习到与问题相关的回答的语义向量的模式，预测候选答案的开始和结束位置。并结合问题和答案的基础语义相关性，在多个候选答案中进行选择。既可针对一个信息点只有唯一答案，也可扩展到一个信息点有多个答案的情况。

3.3 针对金融实体的特殊优化

金融新闻、行业报告、公告等金融领域的文本具有其语法、用词上的独特结构，因此我们针对该领域做了特殊优化。具体而言，该领域的命名实体识别任务仍然面临着以下问题：1. 目前市面上的命名实体识别工具只能识别粗粒度通用命名实体。粗粒度通用命名实体包括人名，组织机构名，地点。由于金融数据有大量的行业专有名词，粗粒度通用命名实体并不能满足金融系统的需求。2. 目前大部分的命名实体识别工具是在新闻语料如《人民日报》上进行评估。但是在金融数据上表现不佳，因此不能支撑系统中上游模块的使用。3. 目前没有公开大规模金融数据的命名实体识别语料。

我们具体结合系统的实际需求和金融行业知识，定制研发了系统中的命名实体识别工具。1. 定义了细粒度的命名实体。例如，将粗粒度的组织机构分成细粒度的公司名，股票交易市场，教育机构，公共设施等有更加准确语义信息的命名实体。2. 结合金融行业专家积累构建了金融实体词典。金融实体词典能够进一步帮助提高命名实体识别工具的效果。3. 构建了 TB 级高质量的金融语料库。大规模金融语料库能够帮助提高模型的表现能力。4. 构建了新词发现系统。通过使用深度神经网络（DNN）对每天更新的新闻中词语的互信息、信息熵以及上下文词向量进行融合，自动识别有价值信息的领域新词。具体模型的结构可见下图：

定制化细粒度命名实体识别模块能够满足舆情系统中对金融命名实体识别的实际需求，并且可以作为特征提高其他模块模型预测的准确率。

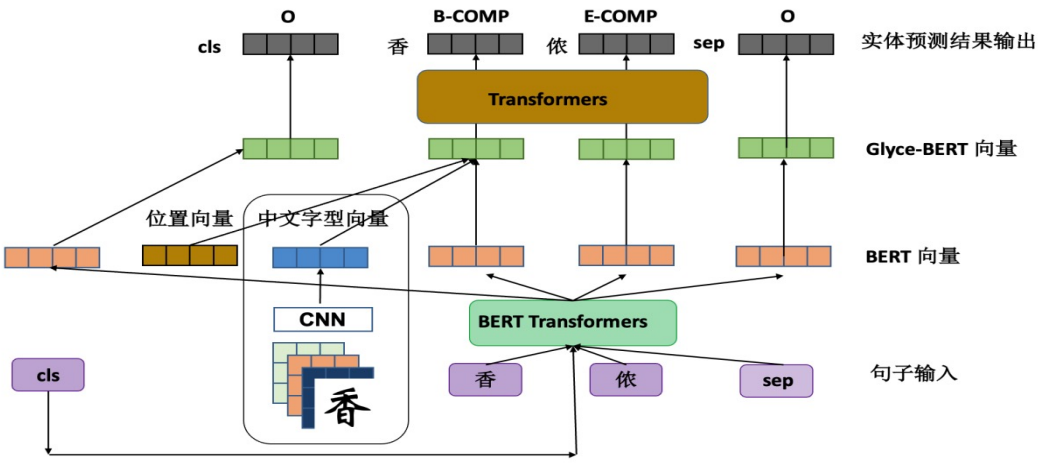


图 2 命名实体识别模型结构

四、表格信息抽取

4.1 问题定义

给定一个表格T，包含m行n列的单元格矩阵 $T = \{t\_\{I,j\} \mid 1 \le i \le m, 1 \le j \le n\}$，其中 $t_{i,j}$ 表示第i行第j列的单元格内容。表格信息抽取任务目标是：抽取表格中的关键信息实体 $E = \{e_l \mid l = 1,2,3,\ldots,L\}$。

4.2 文本抽取在表格抽取上遇到的问题

文本信息抽取中，我们使用了 transformers 模型来进行信息抽取。然而， 文本抽取中的 Self Attention 机制使用一维位置编码来表示序列中元素的位置信息。这种一维的机制在处理二维表格时会遇到 以下问题：

二维空间关系缺失：表格中的单元格具有明确的行 - 列坐标关系（如 (row=2, col=3)），而传统的一维位置编码（如绝对位置或相对位置编码）只能表示线性序列中的顺序信息（如 pos=5），无法建模单元格之间的二维空间关系。

长距离依赖与稀疏注意力冲突：表格中跨行或跨列的语义关联（如跨多行的表头与数据对齐）需要长距离依赖建模，但一维位置编码的全局注意力机制会引入大量无关的远距离计算。

为了解决上述问题，我们提出了二维注意力机制，具体来说：

1. 使用二维位置编码，充分刻画行和列的位置信息
2. 引入距离函数对二维空间相关性进行建模

4.3 二维表格注意力机制

4.3.1 二维位置编码

对于表格数据，我们需要设计二维位置编码来表示单元格在表格中的二维坐标。

给定单元格 t_(i,j)，其二维位置编码由行位置编码 $PE_{row(i)}$ 和列位置编码 $PE_{col(j)}$ 组成：

$$PE_{2D(i,j)} = PE_{row(i)} \oplus PE_{col(j)}$$

其中⊕表示向量拼接操作。行位置编码和列位置编码采用正弦余弦函数：

$$PE_{row(i,2k)} = \sin\left(\frac{i}{10000^{\frac{2k}{d}}}\right)$$

$$PE_{row(i,2k+1)} = cos\left(\frac{i}{10000^{\frac{2k}{d}}}\right)$$

$$PE_{col(i,2k)} = \sin\left(\frac{j}{10000^{\frac{2k}{d}}}\right)$$

$$PE_{col(i,2k+1)} = cos\left(\frac{j}{10000^{\frac{2k}{d}}}\right)$$

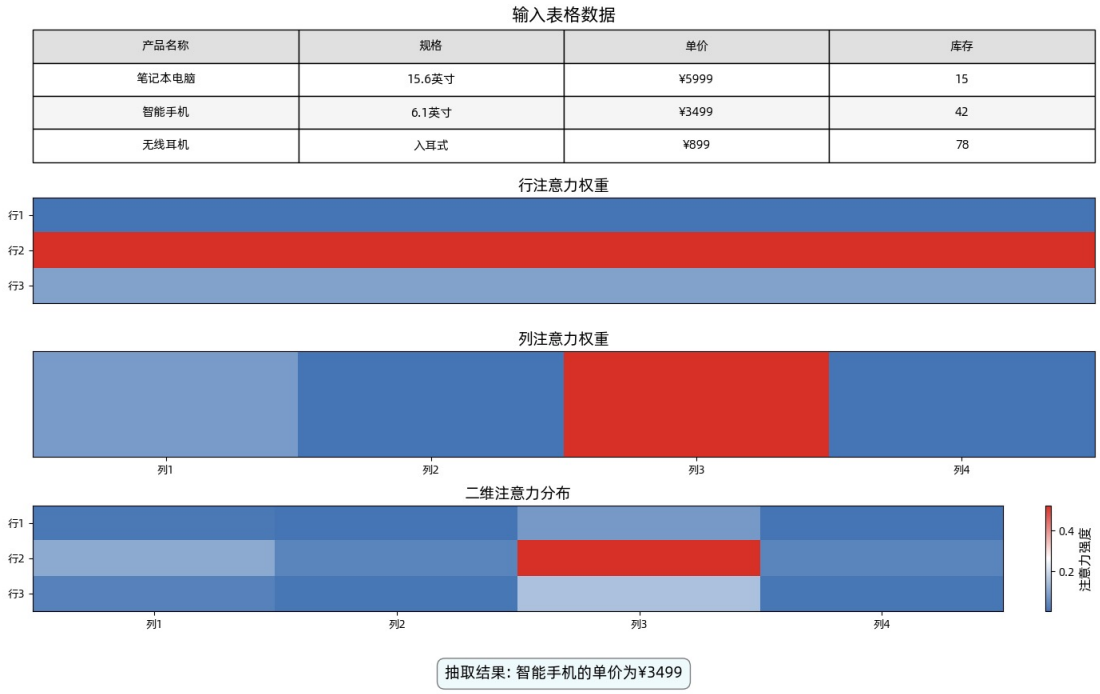


图 3 表格二维注意力示例

4.2.2 二维空间相关性建模

表格单元格之间的相关性不仅取决于其语义内容，还与其在二维空间中的相对位置密切相关。我们设计了空间距离感知的注意力权重计算方法：

$$\alpha_{(i,j),(p,q)} = \text{softmax}(\frac{Q_{i,j} \cdot K_{p,q}^T}{\sqrt{d_k}} + \beta \cdot f_{spatial}(i,j,p,q))$$

其中 $f_{spatial}(i,j,p,q)$ 是空间距离函数：

$$f_{spatial}(i,j,p,q) = -\gamma \cdot (|i-p| + |j-q|)$$

参数 β 和 γ 控制空间距离对注意力权重的影响程度。

五、表头识别

5.1 问题定义

表头识别是表格理解的关键步骤之一。传统方法主要基于表格的视觉特征，如字体、位置、边框等进行识别。复杂表头识别面临更大挑战。多行表头涉及表头的层次结构理解，跨单元格表头需要识别单元格的合并关系。一个带有复杂表头的表格样例如下图所示：

（二）公司主要客户销售情况

报告期内，公司向前五大客户销售的情况如下：

序号	客户名称	销售的具体内容	获得业务的形式	销售金额（万元）	占当期销售总额比重
		2020 年 1-6 月			
1	中国联合网络通信有限公司	安防视频监控产品销售：摄像机、设备箱、主机箱、机柜、防雷器、支架、存储卡、配电柜等	招投标	2,713.78	23.48%
2	中国人民解放军92330部队政治工作部	社会安全系统解决方案：崂山区某单位安全防范基础设施完善配套工程	招投标	2,046.45	17.71%
3	中共酒泉市金塔县委政法委员会	社会安全系统解决方案：酒泉市雪亮工程（金塔县）	招投标	1,550.71	13.42%
4	厦门市公安局	社会安全系统解决方案：重点区域防控系统建设	招投标	851.16	7.36%
5	南靖县公安局交通警察大队	社会安全系统解决方案：南靖县城区智能交通管控系统建设项目	招投标	810.55	7.01%
		合计		7,972.64	68.98%
		2019 年度			
1	重庆市江津区公安局	社会安全系统解决方案：江津区社会公共安全视频监控建设联网应用工程	招投标	9,933.42	19.58%
2	中国电信股份有限公司	社会安全系统解决方案：天竺山景区视频安防监控系统项目； 安防视频监控产品销售：前端监控设备、平台存储网络设备、边界及安全设备等； 维保服务：海沧区重大安保视频监控及信息化系统	招投标、竞争性谈判	9,211.33	18.16%

图 4 带表头的表格样例

5.2 总体流程

我们将表头分为以下几种类型：

- 简单表头：单行单列表头
- 多行表头：跨越多行的层次化表头
- 跨列表头：跨越多列的类别表头
- 复合表头：同时跨行跨列的复杂表头
- 表格内表头：出现在表格内的表头，例如不同每 N 行出现一次的跨全列的表头，如下图：

序号	客户名称	销售金额（万元）	占年度销售额比例
	重庆比亚迪锂电池有限公司	25.66	0.17%
	深圳市比亚迪锂电池有限公司坑梓分公司	6.41	0.04%
	小 计	**8,570.14**	**56.48%**
2	宁德时代新能源科技股份有限公司	975.37	6.43%
	江苏时代新能源科技有限公司	420.00	2.77%
	时代上汽动力电池有限公司	385.25	2.54%
	小 计	**1,780.62**	**11.74%**
3	青山控股集团有限公司	1,238.94	8.17%
4	万向一二三股份公司	796.46	5.25%
5	中材锂膜有限公司	434.48	2.86%
合 计		**12,820.63**	**84.49%**
2020 年度			
1	深圳市赢合科技股份有限公司	4,604.59	42.11%
2	西安众迪锂电池有限公司	2,858.27	26.14%
	重庆比亚迪锂电池有限公司	405.98	3.71%
	商洛比亚迪实业有限公司	127.59	1.17%
	青海弗迪电池有限公司	70.80	0.65%
	深圳市比亚迪供应链管理有限公司	7.08	0.06%
	小 计	**3,469.72**	**31.73%**
3	上海卡耐新能源有限公司	769.23	7.03%
4	哈尔滨万鑫石墨谷科技有限公司	463.58	4.24%
5	华鼎国联四川动力电池有限公司	349.14	3.19%
合 计		**9,656.26**	**88.31%**

图 5 表格内表头示例

针对每个候选表头区域，提取以下特征：

- 文本特征：词汇、语法、语义特征
- 位置特征：在表格中的绝对和相对位置
- 视觉特征：字体、对齐方式、边框等
- 结构特征：与相邻单元格的关系，覆盖的行数、列数
- 合并单元格特征：识别单元格的范围以及对应的行列

由于表头在绝大多数情况下都是按行出现的，即，每一行要么全部是表头要么全部不是，因此，我们利用这条前置信息，设计了如下分类方案：

1. 使用前述的特征提取，提取所有单元格特征；
2. 对于每一行，确认每一个单元格是否应当与下一行进行合并，如果合并，则该行一定不是表头；
3. 否则，进入真正的分类模型；

5.3 分类模型

5.3.1 模型结构与输入格式

我们采用预训练的 BERT 模型作为基础架构，利用其强大的上下文理解能力捕捉表头的语义和结构特征。模型结构分为以下部分：

1. 输入嵌入层：将文本、位置和视觉特征融合为统一表示

2.12 层 Transformer 编码器：提取跨模态的深层特征

3. 分类头：包含两层 FFN+ReLU 的二元分类器（表头 / 非表头）

特别地，我们在 BERT 的原始输入中新增了两种特殊 token：

• [CELL]：标记单元格起始位置

• [ROW]：标记行分隔位置

每个表格样本被构造为如下序列格式：

[CLS] [ROW] [CELL] 文本特征 [SEP] 位置特征 [SEP] 视觉特征 [ROW] [CELL] ...

具体特征编码方式：

1. 文本特征：

• 原始文本经过 WordPiece 分词

• 添加 POS 标签和 NER 标签作为附加特征（如北京 / LOC）

2. 位置特征：

• 绝对位置：(行号 , 列号) → 线性映射为 256 维向量

• 相对位置：与相邻单元格的行列偏移量

3. 视觉特征：

• 字体特征：加粗 =1/ 斜体 =2/ 下划线 =3 → 嵌入矩阵

• 对齐方式：左 =0/ 中 =1/ 右 =2 → one-hot 编码

• 边框类型：实线 =0/ 虚线 =1/ 双线 =2 → 嵌入矩阵

4. 结构特征：

• 跨行 / 列数通过可学习标量量化

• 合并单元格关系用指针网络标记

示例输入（数值已简化）：

[CLS] [ROW] [CELL]　销　量 [SEP] (1,1) [SEP] 1_0_0 [ROW]

[CELL] 手机 [SEP] (2,1) [SEP] 0_1_0 [CELL] Q1 [SEP] (2,2) [SEP] 0_2_1

5.3.2 训练数据生成

我们采用半自动化的数据构建流程：

1. 原始数据来源：数据来源包括研报表格、公告以及根据表头模板自动合成的数据；

2. 标注方法：我们首先通过使用规则引擎生成初步标签 实现自动预标注，再通过标注平台进行人工二次审核，能在保持准确性的同时大大提升标注速度；

3. 数据增强: 利用语言模型根据表格模板进行重新生成，叠加基于当前表格内容进行改写，从而完成数据的扩增；

4. 类别平衡：令简单表头 vs 复杂表头 = 3:7；令正负样本比例通过动态采样保持 1:1.5。

5.3.3 模型训练

在模型训练过程中，我们采用了两阶段训练方法。

第一阶段为预训练任务，包括 Masked Cell Modeling（MCM） 和 Header Relationship Prediction（HRP）。MCM 任务随机 mask 15% 的单元格内容，并预测被 mask 的文本和格式特征；HRP 任务通过二分类判断两个单元格是否属于同一表头。

第二阶段为微调任务，主任务为行级表头分类，采用 Focal Loss 作为损失函数（γ=2, α=0.25），同时设置两个辅助任务：单元格合并预测使用交叉熵损失，表头类型分类采用多标签 softmax。

六、总结

本文提出了一套面向金融文本信息抽取的综合解决方案，通过深度学习与注意力机制的结合，有效解决了金融文档处理中的关键挑战。提出的文本 - 表格双流处理架构、二维空间注意力模型和动态表头解析系统。该框架已在年报分析、招股书解析等实际场景中成功应用，为金融文档智能化处理提供了可靠的技术支撑。

展望未来，该研究可进一步扩展至跨文档关联分析、动态自适应系统等方向，以应对金融领域不断变化的需求。随着技术的持续优化，本研究成果有望推动金融信息处理向全面智能化方向发展，为风险评估、投资决策和监管合规提供更强大的技术支持。该框架的通用性设计也为其他领域复杂文档的信息抽取提供了可借鉴的解决方案。

高新技术企业科创属性评价研究 *

黄越，俞喆华，余勇，谢金浩，王忠
上交所技术有限责任公司

摘　要： 文章主要设计了一种针对国家高新技术企业的科创属性评价指标体系，包含政策契合度、知识产权能力、团队竞争力、企业运营成熟度和产业链地位等五大维度。该评价指标体系首次对高新技术企业的科创属性进行了系统性评价，提出将企业战略发展方向与政府政策契合程度结合到一起，再综合融入知识产权能力、团队竞争力和企业产业链地位等多方面因素，并结合工业指标进行分析，使得该评价指标体系更加科学性、系统性和可操作性。

关键字： 科创属性；评价体系；政策契合度；产业链地位

一、引言

习近平总书记在党的二十大报告中指出：“教育、科技、人才是全面建设社会主义现代化国家的基础性、战略性支撑。必须坚持科技是第一生产力、人才是第一资源、创新是第一动力，深入实施科教兴国战略、人才强国战略、创新驱动发展战略，开辟发展新领域新赛道，不断塑造发展新动能新优势。”

党的十八大以来，党中央提出《国家创新驱动发展战略纲要》，强调科技创新是提高社会生产力和综合国力的战略支撑，必须摆在国家发展全局的核心位置。2021 年 3 月，国务院发布《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》，特别提到“提升企业技术创新能力，强化企业创新主体地位”，要求全力贯彻实施创新驱动发展战略，在产业、企业、区域、重大工程和人才队伍建设等方面着力发挥作用 [1]。在国家层面，从党中央、国务院到各大部委、地方政府，陆续出台各项政策，引导企业在高科技领域大力投入、钻研探索，大力提倡高新技术企业，重视科技能力的建设，为我国社会的高质量发展提供更多的科技源头供给，推动经济的高速增长，进一步提高我国在国际上的竞争力；在高新技术企业层面，科技能力建设必将提高企业的核心竞争力，推动企业的高速发展，在日益激烈的市场竞争和利润获取中占据有利之地。

2018 年 11 月 5 日，习近平总书记在首届中国国际进口博览会开幕式上宣布，在上海证券交易所设立科创板，坚持面向世界科技前沿、面向经济主战场、面向国家重大需求，鼓励和服务于符合国家战略、突破关键核心技术、市场认可度高的科技创新企业，重点支持新一代信息技术、高端装备、新材料、新能源、节能环保以及生物医药等高新技术产业和战略性新兴产业，推动互联网、大数据、云计算、人工智能和制造业深度融合，引领中高端消费，推动质量变革、效率变革、动力变革。

科创板旨在帮助科技创新企业上市融资，其服务对象是符合国家发展战略、掌握关键技术的高新技术公司。如何客观、高效、准确地评价企业的科创属性，对于判定企业属于何种竞争力类型，评估企业是否符合国家创新驱动发展战略，有着至关重要的作用。2007 年，国家科技部发布《科技企业孵化器评价指标体系 (试行)》[2]，主要目标是解决国家级的高新技术小企业的评价，更好地服务科技创业企业和地方经济发展。2020 年，中国证监会发布《科创属性评价指引（试行）》[3]，明确企业科创板上市的“三项指标”和“五种情形”，并于 2021 年进一步修订，形成了“四项指标”和“五种情形”的科创属性评价指引标准。

从总体上来看，现有的研究存在明显缺陷，主要表现在全面性和时效性两个方面。从涵盖范围上看，现有工作主要聚焦于企业技术创新资金投入、创新成果数量及其经济效益，忽略了企业技术和产品与国家发展战略的契合程度的判定；从研究时效上看，现有的评价方法大多发表于上海证券交易所设立科创板之前，对现阶段高新技术企业的科创属性评价指导意义较为有限。文章旨在提出一种多元多维科创属性评价体系，创新性地将“政策契合度”融入企业科创属性评价，以期与时俱进地对高新技术企业的科创属性进行科学性、系统性、全面性的评价。

* 本文是项目下设课题“科创企业评价与行业综合应用示范”（课题编号：2021YFC3340704）的研究成果，课题负责人：唐忆（上海证券交易所）；本文部分相关成果已在《中国市场》发表；文章所研究内容仅在学术范围内探讨，不具任何行政审核指导倾向性。

二、文献综述

近年来，科研工作者在企业科创属性的评价方面取得了不少进展，主要可分为定性研究 [4,5,6] 和定量研究 [7,8,9,10,11,12,13,14]。

定性的研究主要关注科创属性评价的定义、内容以及必要性。刘健等 [4] 总结了科技型中小企业评价工作的政策背景和重要意义。阳雨潇等 [5] 选取创新能力较强的四个行业，利用财务数据分析技术，发现企业创新投入的持续增长对企业绩效和社会责任均有正向作用。根据姚玉明等 [6] 的研究，技术创新能力分为六大维度，创新资源投入能力、创新决策与管理能力、创新倾向、研究开发能力、产品制造能力和市场营销能力。

定量研究则主要利用建模分析来从数据中挖掘潜在的指标。张筱辰等 [7] 提出将快速聚类等数据分析算法引入企业创新能力评估，从产学研深度融合视角预测企业创新能力发展的趋势；孙妍娜等 [8] 借用期权股价模型 , 探讨高新技术企业价值的评估方法，为评估高科技企业提供了一种全新视角；余伟中 [9] 等、张艳春等 [10] 利用实物期权定价模型对高新技术企业价值进行评估，为企业价值评估提供了一种新的思维方式；陈收等 [11] 利用回归方程模型分析高科技上市公司面板数据，发现在企业成长期和成熟期，研发投入与企业绩效呈正相关性和滞后性；许志晋等 [12] 首次将模糊综合评判法引入企业技术创新能力的评价之中，对技术创新能力评价的一般程序展开定量地探讨；赵林海等 [13] 认为知识型企业的技术创新能力是与其获取和开发资源的能力严格相关，提出了一套基于资源的评价指标体系，从外部网络资源、人力资源、企业家资源和经济资源四个方面对知识型中小企业的技术创新能力进行评价；曹萍等 [14] 构建了企业技术创新能力评价指标体系，运用 AHP 方法和多级模糊数学评价模型对企业技术创新能力进行分析，通过专家打分和数学模型的量化分析以及对指标体系的有效控制，在一定程度上避免了数据主观性的不足，客观地反映出企业技术创新能力的情况。

三、科创属性评价体系的设计

3.1 科创属性

科创属性，为评价一家企业的科技创新的含金量，其目的是落实国家部署的企业科创定位，支持和鼓励“硬科技”企业发展，加速科技成果向现实生产力转化，促进经济发展向创新驱动转型。

目前中国证监会发布《科创属性评价指引（试行）》(以下简称《指引》)[15]，其中明确科创属性的评价侧重点为“4+5”，主要包括：一是最近三年研发投入占营业收入比例 5% 以上，或最近三年研发投入金额累计在 6000 万元以上；二是研发人员占当年员工总数的比例不低于 10%；三是形成主营业务收入的发明专利 5 项以上；四是最近三年营业收入复合增长率达到 20%，或最近一年营业收入金额达到 3 亿元；五是拥有的核心技术经国家主管部门认定具有国际领先、引领作用或者对于国家战略具有重大意义；六是作为主要参与单位或者核心技术人员作为主要参与人员，获得国家自然科学奖、国家科技进步奖、国家技术发明奖，并将相关技术运用于主营业务；七是独立或者牵头承担与主营业务和核心技术相关的国家重大科技专项项目；八是依靠核心技术形成的主要产品（服务），属于国家鼓励、支持和推动的关键设备、关键产品、关键零部件、关键材料等，并实现了进口替代；九是形成核心技术和主营业务收入相关的发明专利（含国防专利）合计 50 项以上。

作为当前我国政府正部级行政机构颁布的关于科创属性评价的政策文件，《指引》对现阶段中国关键核心技术的创新迭代能力和科技强国战略贯彻实施具有引领作用。文章基于《指引》的主要关切，结合创新驱动发展战略的指导思想，创新性地将“政策契合度”融入企业科创属性评价，提出一种多元多维科创属性评价体系，以对高新技术企业的科创属性进行科学性、系统性、全面性的评价。

3.2 多元多维企业科创属性评价指标体系设计

文章将高新技术企业的科创属性评价体系划分为五大主题，分别为政策契合度、知识产权能力、团队竞争力、企业运营成熟度和产业链地位。上述主题由于其实际操作过程中的可行性和人为认知性，可分为显性量化指标和隐性非量化指标。

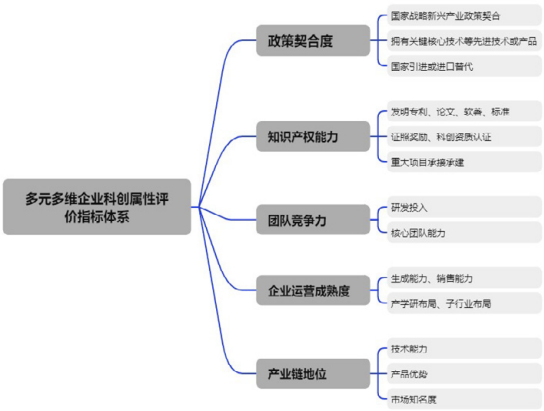


图 1　高新技术企业科创属性评价体系

3.2.1 政策契合度

企业需优先在方向上符合全人类和国家创新发展战略，客观体现企业创新与其社会责任的深度耦合，与社会创新整体同向而行，尤其是高新技术企业，在兼顾短、中、长期利益的同时，还需要考虑国家利益和社会意义，甚至是人类未来发展。因此高新技术企业在其实际业务、产品和技术发展方向上应遵循国家对于战略性新兴产业、“硬科技”方向和“卡脖子”技术的扶持和鼓励预期，切实为国家和人民探索研发核心技术，推动国家鼓励和支持的关键设备、关键产品、关键零部件、关键材料等关键产品和技术的研发生产，实现一定程度上的国际引领或进口替代。

从科创板六大领域（新一代信息技术、高端装备、新材料、新能源、节能环保和生物医药）出发，目前学者梳理了党中央、国务院、发改委、财政部、工业和信息化部、科技部和中央网信办等单位下发的 400 余份政策文件 和 760 余家科创板上市公司的招股书。政策契合度评估通过大语言模型对招股书进行实体识别，解析企业的主营业务、产品和技术情况等信息；借助文心一言向量和相似度算法匹配相关领域的政策文件。通过挖掘公司主营业务与国家政策支持或鼓励的方向的契合程度，探索该公司主营产品或业务是否遵循国家对战略性新兴产业、“硬科技”方向、“卡脖子”或进口替代产品的扶持和鼓励预期；获知该公司的核心技术是否为国内领先甚至国际引领，判定企业产品是否为适应国家和市场重大需求、关系人民生命健康或关系产业链安全的产品。

[1] https://kcb.sse.com.cn/kczl/ty/.

3.2.2 知识产权能力

科创属性的根源在于合理地评价企业科技创新的含金量。企业科技创新中重视实际产品和与之相关的知识产权等无形资产的建设，其中包括但不限于专利、论文和标准等。

专利是企业技术知识产权的最核心体现，发明专利是企业商业价值的核心输出，合理地申请并使用专利权，也是高新技术企业无形资产的有效保护，是企业发明创造成果的重要屏障。因此，专利，尤其是发明专利，可以高效地评估企业科创属性。发明专利的规模、被引用数、核心性、重要性、与主营业务相关性、增长趋势和所有权稳定性等，均需要被纳入评估体系中。当然，对于有失效的发明专利、存在重大权属纠纷的发明专利、共同持有的发明专利，需要进行负面考量。

论文是企业科学知识产权的代表性产出，与专利并称为企业科学与技术的双重核心节点。论文是企业对自然界客观事实规律的探索和新知识新方法的发现创新，也是技术革新的重要基石，两者通常也是相辅相成、关联依赖，互相作用、互相促进，双螺旋式发展的。论文，尤其是高水平会议和核心期刊且被高频次引用的论文，可以体现高新技术企业的科学创新竞争力。这类企业在创新能力的强端，他们优先布局科学创新，在积累到一定知识程度后，逐步向技术转化，形成专利或实际科技研究成果，并进而商品化。在一定程度上，论文的主题覆盖面塑造了企业专利和技术的广度，论文的单领域主题先进性成就了企业专利和技术的深度。因此，论文，尤其是高水平论文，可以评估企业科创属性。论文的规模、被引用数、顶级会议或核心期刊发表、与主营业务相关性、增长趋势、科技成果转化程度等，也需被纳入评估体系中。

标准是总结、传播科学与技术的最佳技术和最佳实践[16]。所谓“三流的企业做产品，二流的企业做品牌，一流的企业做标准”。高水平标准的发布，是对企业重复性事物和概念所做的统一规定，它以科学、技术和实践经验的综合成果为基础，经有关方面协商一致，由主管机构批准，以特定形式发布，作为共同遵守的准则和依据。标准依据标准化对象的不同分为技术标准、管理标准和工作标准。技术标准包括基础标准、产品标准、工艺标准、检测试验方法标准及安全、卫生、环保标准等。标准依据主管机构的不同，分为国家标准、行业标准、地方标准和企业标准；依据指导方式的不同，分为强制性标准和推荐性标准。因此，技术标准的规模、增长趋势，尤其是强制性国家标准的规模，需被纳入评估体系中。

此外，针对特殊类型企业，如计算机软件和生物医药等，需要针对该类型企业的业务特征，考虑特定科创属性评价因素。

软著是计算机软件行业中一方或多方众多计算机软件从业人员的脑力劳动的成果体现。一个软件系统，从预调研、需求分析、架构设计、产品研发到测试等多个过程，耗费了大量的人力、物力，同时也需要企业巨额资金投入支持。而软著是计算机软件企业研发成果的唯一知识产权载物，承担着为企业保护核心竞争力，防复制防抄袭的使命。因此，软著的规模、重要性、增长趋势和所有权稳定性等，在评价计算机软件企业科创属性时，需要重点考虑。

药物临床试验研究，是生物医药行业中承载新药研发人员众多智慧的结晶。对药企来说，一款新药的研发过程繁冗复杂，时间周期长，研发成功率极低，能发展到药物临床试验阶段并获批的少之又少。通常情况下，药企会先进行实验室细胞模型筛选，研究后进行动物体内的药理学、毒理学等研究，然后才可进行临床试验研究过程。临床试验研究分为四期，一期二期为初步测试疗效和耐受性等，判定是否有条件批准上市；三期四期为确证性研究和大样本安全性评价，判断是否批准上市。通常进入上述四个周

期的药品，试验期越靠后离批准上市越近，离实际生产经营获利越近。因此，申报及被批准的不同临床试验期的项目规模、趋势和布局结构应纳入生物医药企业的科创属性判定因素。

除此之外，高新技术企业的科创属性评价需关注技术创新成果的转化情况。对于企业及其科技成果，在各项国际、国家、部委和地方评优评奖中获得高水平奖励和证明者，各类被认定的重点实验室、新型研发中心、企业技术中心、工业设计中心和高新技术企业等均可作为该企业在科创属性方面先进性的佐证，也应被纳入评估体系中。

3.2.3 团队竞争力

高新技术企业科技创新的源泉动力是极具竞争力的专业研发团队和成熟的高级别管理团队。考察高新技术企业的竞争力，首要关注其是否重视高精尖高素质人才和丰富相关从业经验的高级管理人员的吸纳，尤其是与企业主营业务直接相关的高新技术专业人才、有过大型竞品企业高级管理经验和行业知名企业履历的高级管理人才；其次关注其是否从公司层面敢于大力增加研发相关经费的投入，特别是与公司经营规模相适配的研发经费投入。

基于以上，文章建议在企业科创属性评价体系中纳入团队中核心成员专业技术能力或管理能力认定和研发投入状况进行评估。其中，核心成员专业技术能力或管理能力认定包括技术核心、业务核心和高级管理人员的学历知识组成、学术界认定、工业界认定和履职经历认定；研发投入包括近年来研发累计投入金额、研发增长金额、研发金额增长趋势，研发投入占营业收入比重、研发投入占营业收入比重增长率等。除此之外，企业稳定在册的研发人员规模、核心研发人员规模、研发人员占全部人员比例、核心研发团队规模、研发岗位情况、高级研发岗位情况等也是客观评价企业团队竞争力的可考虑因素。

分析时注意，研发人员需为直接与企业签订劳动合同，专为开展研发活动的人员，且短期内没有突击式、爆发式大批量增大规模；研发投入需为费用化的研发费用和资本化的开发支出，包括研发人员的薪酬、生产和研发的设备和生产线、研发的物料和人工费等。

3.2.4 企业运营成熟度

高新技术企业科技创新后，还需关注其科技成果市场情况，因此成熟的企业运营机制也是重要考量因素，运营成熟度评价可以反映企业的业务流程、运营效率和资源配置情况。企业是否重视原材料供应端和市场开发管理，尤其在实际生产中的生产原材料端和销售端上的话语权和稳定性；企业是否重视与相关高校和企业的产学研结合多样化发展，在主营业务的相关子领域的深度攻关布局，为维护产品生态而贡献力量。

通常情况下，原材料供应端和市场开发管理主要依据财务指标和实际调研得出，主要体现指标为在购货上游的话语权、购买端资金周转率、供应商集中度、下游端话语权和存货周转率、下游客户货源集中度、营业收入增长规模和营业收入复合增长率等。工业指标上，相关指标的计算通常结合利润表的权责发生制和现金流量表的收付实现而产生，具体为净现比、收现比、付现比和应付账款应收账款比。其中，净现比是经营活动产生的现金流量净额与净利润的比值，比值越大，企业盈利质量越高；收现比是销售商品和提供劳务收到的现金与营业收入的比值，比值越大，企业对下游的议价能力越强；付现比是购买商品、接受劳务支付的现金与主营业务成本的比值，比值越小，企业对上游供应商的议价能力越强。产学研布局具体分为委外研发、高校合作研发、企业合作研发和企业共建研发等；依照合作形式上分为团队攻关、人才培养和技术转移等；合作时长分为长期合作、短期合作、不定期合作等；一般情况下，与知名高校、龙头企业联系更加密切，合作关系更加长久体现该企业运营更成熟。

3.2.6 产业链地位

产业链是现代社会分工引起的，在交易机制的作用下不断依据产业价值进行组织深化得到。产业链地位是企业在产业层次、产业关联、资源加工深度和满足需求程度等四个方面价值的综合体现。高新技术企业应重视产业链中市场地位建设，通过技术能力、产品优势、市场推广和企业品牌内涵等方面的持续创新不断提高产业竞争力；同时与相关领域的高校、企业进行产学研的共建联合，进行合作创新和集成创新，提高市场知名度、市场占有率、行业排名、国际排名等关键指标，凸显企业在产业链上的关键节点作用 [17，18]。

从过程论，生产效率、交易成本和企业间关系很大程度上决定了企业在产业链上的可创造价值；所处产业链的分工精细度、市场交易活跃度和产业链发展完善程度综合体现了企业在产业链上的增值价值。从结果论，企业在产业链中的议价能力，包括定价权和谈判能力，反映企业在市场竞争中的市场定价能力和谈判能力。应付账款代表公司对上游供应商的占款，应收账款代表了下游客户对公司的占款。应付账款 / 应收账款的背后，是公司对上下游之间议价权的体现，该比率越大，说明公司的产业链地位越高。

四、结论

文章对高新技术企业对科创属性进行了定义，并依据国家相关政策要求设计了一套企业科创属性评价体系，从政策契合度、知识产权能力、团队竞争力、企业运营成熟度和产业链地位等五大维度视角对科创属性进行了定性分析。该评价体系对高新技术企业的科创属性进行了系统性

地、全面地、科学地评价，创造性地融入了政府政策契合程度和企业产业链地位等多方面因素，并结合工业指标进行分析，使得该评价指标体系具有可落地性。

参考文献：

[1] 赵竹青 . 科技如何支撑高质量发展 [J]. 科技传播 ,2022,14(11):2.

[2] 中华人民共和国科学技术部 . 关于印发《科技企业孵化器评价指标体系（试行）》的通知 [EB/OL].(2007-12-20)[2022-12-29].https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2010before/200802/t20080227_143711.html.

[3] 中国证券监督管理委员会 .【第 21 号公告】《科创属性评价指引（试行）》[EB/OL].(2020-03-20)[2022-12-29].http://www.csrc.gov.cn/csrc/c101950/c1048009/content.shtml.

[4] 刘健 , 黄丽娟 . 科技型中小企业评价与发展政策研究 [J]. 财务与金融 ,2022(2):33-37.

[5] 阳雨潇 . 创新投入、企业社会责任与企业绩效 [J]. 中国商论 ,2022(20):126-128.

[6] 姚玉明 . 谈谈企业的技术创新能力 [EB/OL]. [2022-12-26].http://www.weighment.com/newsletter/year2007/m2/892.htm.

[7] 张筱辰 , 汤玲玲 . 产学研深度融合与企业科技创新能力提升研究 [J]. 技术与市场 ,2022,29(8):38-40,43.

[8] 孙妍娜 , 邓恩 . 高新技术企业评估方法探讨 -- 期权估价法 [J]. 湖南广播电视大学学报 ,2004(1):53-55.

[9] 余伟中 . 运用实物期权方法给高新技术企业评估——以某生物医药企业为例 [D]. 上海 : 上海财经大学 ,2002.

[10] 张艳春 , 谭家彬 . 实物期权方法在高新技术企业价值评估中的运用 [J]. 商场现代化 ,2009(26):24-25.

[11] 陈收 , 邹增明 , 刘端 . 技术创新能力生命周期与研发投入对企业绩效的影响 [J]. 科技进步与对策 ,2015,32(12):72-78.

[12] 许志晋 , 凌奕杰 , 宋凤珍 . 企业技术创新能力的模糊综合评判 [J]. 科学研究 ,1997(1).

[13] 赵林海 . 知识型中小企业技术创新能力的评价 [J]. 统计与决策 ,2007(1):67-68.

[14] 曹萍 , 张剑 . 企业技术创新能力的评价 [J]. 中国管理信息化 ,2009,12(2):89-92.

[15] 关于修改《科创属性评价指引（试行）》的决定 [EB/OL].(2021-04/17).http://www.gov.cn/zhengce/zhengceku/2021-04/17/ content_5600280.htm.

[16] NARAYANAN V K,CHEN T.Research on technology standards:accomplishment and challenges[J].Research policy,2012,41(8):1375-1406.

[17] 曾琼 , 舒巧 . 重庆市产学研融合现状与制约因素及对策建议 [J]. 科技资讯 ,2021,19(17):78-82+87.

[18] 张士运 , 肖雯 , 谢海涛 . 科技自主创新背景下产学研深度融合发展研究与建议 [J]. 科技智囊 ,2020(12):38-43.

03 工程实践

大语言模型面向金融长文档智能理解的能力评测系统 *

俞定甫、张宇豪、胡斯涵、杨忠良、周琳娜
北京邮电大学

摘　要： 在金融应用场景中，随着大语言模型的快速发展，其对长文档的理解与推理能力已成为智能投研、量化投资与合规审计等任务的关键支撑。然而，现有金融评测体系主要聚焦于通用知识问答，缺乏对模型在处理金融长文档中的综合性测试，难以满足实际业务中对高结构性文本的理解需求，尤其是在如招股说明书等合规性审阅场景下。为此，本文构建了一个面向金融场景的大语言模型长文档性能评测系统。该系统涵盖券商研究报告、公司公告、政策文件等八类典型金融文档，设计包括信息提取、关键数据提取、事件分析、逻辑推理等在内的十二类评测任务，全面覆盖金融语境下的核心应用需求。系统采用六维度评价指标体系，从相关性、流畅性、连贯性、有用性、一致性与忠实度对模型表现进行评估，并引入三种主流大语言模型开展实验对比。同时，针对金融长文本中涉及的数据计算与陷阱问题，设计差异化评估机制，以更准确反映模型在高复杂度任务中的表现。评测系统具备良好的可扩展性，为推动大模型在金融合规与智能审阅等高要求场景中的落地应用提供了有效支撑。

关键字： 大语言模型；金融合规；长文档理解；评测系统；六维度评估

一、引言

近年来，随着大语言模型（LLMs）的快速发展，其在任务理解和推理能力方面取得显著进展，已广泛应用于法律 [1]、教育 [2]、金融 [3] 等多个领域。伴随模型参数规模的持续扩大及训练数据的多样化，这些模型展现出新兴能力的同时，也引入了更多不确定性与潜在风险。在此背景下，构建科学、标准化的评估基准，以量化模型在特定领域的有效性，已成为大模型持续演进过程中的关键环节。

在金融领域，已有评估体系如 FinanceBench[4] 和 FinEval[5] 分别面向英文开放问答与中文选择题任务，评估模型在金融知识理解与问答方面的能力。然而，当前主流评测基准仍集中于标准化任务，对模型处理金融长文档等结构复杂、语义密集文档的能力覆盖不足。特别是在招股书审阅等金融合规场景中，模型面临逻辑一致性判断、关键数据提取、陷阱识别等多重挑战，现有基准难以系统反映其实际表现。

在众多金融应用场景中，合规性审核对大语言模型提出了尤其严苛的能力要求，尤其是在处理信息密度高、结构复杂的 IPO 招股说明书时，模型需具备稳定的文本理解、逻辑一致性判断与数据提取能力。

因此，本文提出面向招股书等金融长文档场景的大模型评测系统，以推动大模型在合规场景中的可信落地。该评测系统的主要贡献体现在三个方面：1）该系统覆盖 8 类典型金融文档，设计 12 类任务类型，包括信息提取、文本理解、表格解析、数据计算和陷阱问题识别等，贴合金融实务需求；2）提出六维度评估框架，从相关性、流畅性、连贯性、有用性、一致性和忠实度等方面全面评估模型输出质量，特别适用于衡量模型在招股书合规性审核等真实场景中的稳定性与可靠性；3）评测系统具有良好的可扩展性，支持后续接入多种模型与新任务场景，具备广泛应用前景。

二、相关工作

大语言模型的进步催生了各种评估系统，每个框架都有其独特的功能。本节将详细概述当前评估框架的现状。

Liang 等人 [6] 提出了 HELM，对大语言模型进行全面评估，涵盖语言理解、生成能力、连贯性、上下文敏感性、常识推理和领域知识等多个方面。其目标是全方位评估语言模型在各类任务和领域中的整体表现。HuggingFace 中的 OpenLLM 开放大语言模型排行榜，通过提供一个公开竞赛平台来比较和评估不同大语言模型在各类任务上的表现。

OpenCompass 进一步集成了多个数据集 / 任务，提供开源、高效、全面的大模型评测开放平台。Zheng 等人 [7] 提供了一个大语言模型评估平台 Chatbot Arena，用户通过匿名的 Elo 评分系统对模型的响应进行投票。FlagEval 是一个多语言、多模态的评估平台，涵盖了中文和英文的自然语言处理（NLP）和计算机视觉（CV）任务的基准测试。

An 等人 [8] 提出了一个全面的长文本大语言模型评估基准 L-Eval，从规模较小的类似公共数据集中重新标

注数据和指令，以确保质量。此外，它还优化了评估程序和基线，以获得更准确的结论。Shaham 等人 [9] 提出了 ZeroSCROLLS，这是一个专注于长文本理解（长文档类型）的零样本基准测试，它通过改编和新引入的数据集，全面评估了一些开源和封闭的大型语言模型。Bai 等人 [10] 提供了一个中英双语和多任务数据集 LongBench，具有不同长度、分布、模式、语言和领域的各种序列，用于全面评估长语境理解能力。

现有评测系统只有 L-Eval 包含少量金融长文档评测题（仅 8 篇英文文档，且只覆盖业绩会这一个场景），缺乏对于金融长文档理解、分析能力的综合全面评估，本文构建的 FinLongEval 金融长文档性能评测系统正好填补了这一空白。

* 本文是项目下设课题“智能信披审核和监管数据安全共享关键技术研究”（课题编号：2021YFC3340701）的研究成果，课题负责人：周琳娜（北京邮电大学）；本文部分相关成果已在第五届 CSIG 中国媒体取证与安全大会发表。

三、金融长文档评测集构建

金融长文档评测集的设计应紧贴实际金融业务场景，覆盖真实存在的信息处理需求，反映典型的行业挑战与合规要求。尤其在招股书等高规范监管文件的智能审核中，大语言模型需具备对长篇幅、结构复杂、信息密集文本的理解与提取能力。为此，FinLongEval 评测集整理了 8 类具有代表性的金融长文档，并结合招股书审核中常见的披露信息抽取、风险因素识别与财务数据理解等任务，设计了 12 类核心问题类型，系统化评估模型在复杂金融文本中的实用性能。

（一）金融长文档类型及描述

在金融长文档类型上，本评测集总共涵盖了 8 类一级分类文档，18 类二级分类文档，各类文档的基本情况如下所述：

- 券商研究报告：涵盖个股研报、行业研报、宏观研报、金工研报这四类常见券商研报，文本长度在 1 万字至 3 万字之间；
- 上市公司公告 / 募集书：涵盖拟上市公司招股书、债券募集书、基金募集说明书、上市公司年报、业绩预告 & 快报、股权激励公告等，文本长度大部分在 10 万字至 30 万字之间；
- 财经资讯：涵盖财经评论、主流财经媒体的财经早报等，文本长度在 3 千字至 1 万字之间；
- 会议路演：涵盖业绩交流会、策略会等会议文字，文本长度在 1 万字至 5 万字之间；
- 政策文件：涵盖国务院政策文件、政府工作报告、人民银行的货币政策报告等文件，文本长度在 1 万字至 5 万字之间；
- 学术论文：涵盖货币政策、外汇储备、疫情研究等金融学术类文章，文本长度在 1 万字至 3 万字之间；

1.https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

2.https://opencompass.org.cn/

3.https://flageval.baai.ac.cn/

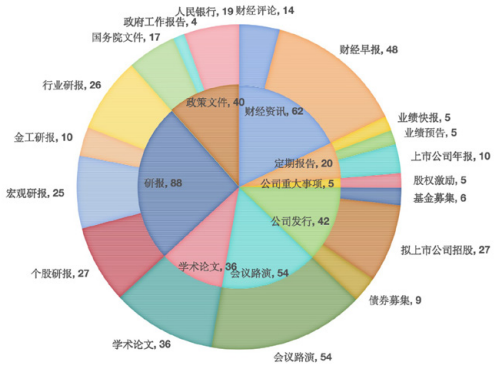


图 1 金融长文档文件类型分布图

图 1 所示为本评测集的文件类型分布比例图，券商研究报告、上市公司公告 / 募集书（定期报告、公司发行、公司重大事项）、财经资讯、会议路演文件，占比分别为 25.4%、19.3%、17.9%、15.6%。在文件类型覆盖度和覆盖比例上面我们尽量做到与实际业务场景的需要处理文件类型分布比例保持一致。

（二）问题类型及描述

为全面评估大模型在金融长文档处理中的能力，结合投研分析、文档合规审查、投教服务等实际业务场景，设计了 12 类不同类型的问题。通过这些问题，旨在从不同维度和场景对大模型进行充分的评估与测试。具体的问题类型、考查目标及其对应的业务场景如下所述：

1. 信息提取：从文档中提取特定信息片段，如公司名称、项目名称、关键人物等。在招股书中，常用于提取发行人基本信息、股东结构、募集资金用途等核心条目。

2. 表格提取：识别并解析文档中的结构化表格内容，如财务数据、预测指标等。适用于招股书中财务摘要、募集资金使用表等关键表格的结构还原与内容提取。

3. 关键数据提取：从金融文档中提取特定数值信息，如从公司业绩说明会的纪要中提取营收、毛利等关键指标，或从公司报告中提取产销量等运营指标。

4. 阅读理解：在理解上下文的基础上，提取信息并给出逻辑

合理的分析结论。用于分析招股书中的毛利率变化、风险因素解释等，需要结合数据和语义进行推理判断。

5. 事件分析：针对资讯中的事件，大模型可以进行深入推理，提供更多分析信息。常见场景包括量化投资分析、政策解读、ESG 因子挖掘等。

6. 逻辑推理：在投资研究中，分析师常要求大模型基于公司或行业基本面进行推理，辅助预测公司未来业绩或行业走势。

7. 关键词提取：快速提炼金融文档的核心关键词，如会议路演中的关键词总结，帮助识别主旨。

8. 文本摘要：快速生成金融投研文档的摘要，如会议路演或政策分析文档的简要总结，适用于机构问答分析、文本因子提取等场景。

9. 生成提纲：用于写作辅助，帮助生成简要提纲后进行深入撰写。例如，为政府报告或政策解读提供初步框架。

10. 对话人分辨：在会议路演等多人参与的场景中，准确识别各个发言者的角色，确保总结和分析内容的准确性。

11. 陷阱问题：金融行业高度依赖精准信息，大模型的“幻觉问题”直接影响其在金融场景中的可应用性和系统可靠性。

12. 数据计算：基于文档中已知数据，进行进一步的数据计算与推导，如计算衍生指标以辅助决策分析。

（三）文件的长度分布

如图 3 所示，本次评测集中的各类文件字数分布情况表明，超过 80% 的文件字数在 1 万字以上，显著超出现有金融评测集的平均文本长度。此外，超过 40% 的文件字数超过 2.5 万字，远超当前典型商用大模型的上下文窗口长度。本评测集中还包含字数超过 50 万字（约 500 页）的超长金融文档。

四、FinLongEval 评测系统

上一章节详细介绍了金融长文档评测集的构建过程及其组成，该评测集为金融领域长文本理解任务提供了可靠的基准。在此基础上，FinLongEval 评测系统得以建立，用于对不同模型在金融领域长文本处理任务中的表现进行系统化评测。基于评测集，FinLongEval 评测系统包含了评测任务选择、模型生成答案和打分模型评测等关键功能，从而实现对模型性能的全面评估。

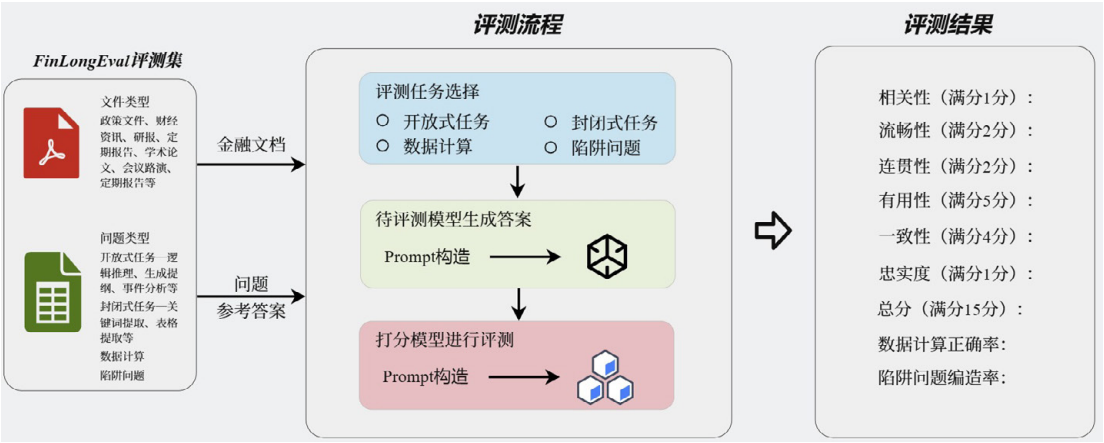


图 2　FinLongEval 评测系统评测流程

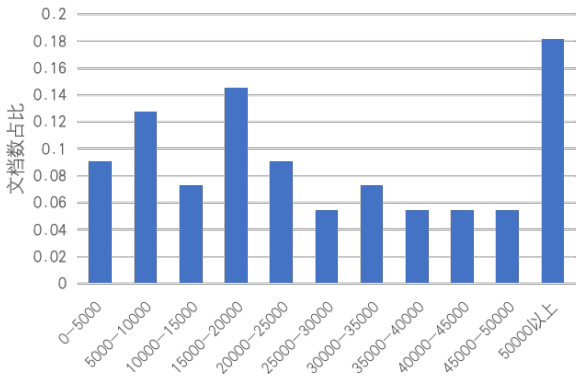


图 3　金融长文档文件长度分布图

如图 2 所示，FinLongEval 评测系统各个模块的具体结构如下所述。

（一）评测任务选择

为全面评估大模型在处理金融长文档方面的能力，同时充分考虑智能投研、量化投资等实际业务场景，设计了 12 类不同的问题类型。这些问题可根据任务特点划分如下：

· 开放式任务：无唯一正确答案，通常涉及理解、推理和生成，典型问题包括逻辑推理、文本摘要、事件分析、生成提纲、信息提取等；

· 封闭式任务：有明确正确答案范围，聚焦于精确信息的抽取和识别，如关键数据提取、表格提取和关键词提取；

· 陷阱问题：用于测试模型是否能识别并回避误导性信息，特别考察模型在处理潜在风险措辞或虚假前提时的稳健性，贴近招股书风险因素审查任务；

· 数据计算任务：要求模型根据给定文本中的数值进行计算与推导，评估其财务计算和逻辑一致性能力，在招股书财报交叉验证等场景中具有现实应用价值。

这一任务体系覆盖金融文档处理的核心维度，既支持基础信息识别，也支持高阶语义分析与合规性推理，具备良好的通用性和行业适应性，后续章节将结合具体模型测试与评估方式进行详细介绍。

（二）待评测模型答案生成

由第‘三、（三）’节可知，评测集中的大部分金融文档篇幅均在数万字以上，评测任务以长文本问答为主，这对大语言模型的上下文处理能力提出较高要求。因此，评测所选模型需满足以下条件：商业模型应支持上传完整文件进行交互；API 调用与本地部署模型需支持不低于 128K tokens 的上下文窗口，以保证任务对话在招股书等超长文档场景中的完整覆盖与语义连续性。

1、已有评测结果模型

本系统评测的模型类型涵盖商业模型、闭源 API 模型和开源本地模型三类：

商业模型：包括 Alphabox[4]、ChatDOC[5]、ChatGPT4[6]、ChatPDF[7]、Claude2[8]、Moonshot[9]、WarrenQ[10] 和 ERNIE Bot 3.5[11]，均具备文档上传能力，部分模型如 ChatDOC 和 ChatPDF 专注于表格识别与 PDF 解析，适用于招股书中的结构化内容抽取任务。答复生成采用手动上传文档并提问的方式，处理结果经人工筛选后用于评分。

闭源模型（API 调用）：如 Doubao-lite-128k 与 ERNIE-Speed-128K，支持 128K 上下文窗口，响应速度快，适合金融智能问答任务的集成部署。

开源模型（本地部署）：包括 GLM-4-9B-chat[11]、ChatGLM3-6B-128K 和 Qwen2-7B-Instruct[12]，其中部分模型通过优化配置可支持 128K 上下文窗口，适用于大规模本地化评测与灵活适配。

```
1  question_prompt = f"根据文件内容，回答问题：{question}"
2  final_prompt = '文件内容包含在<file></file>中\n<file>\n' + file_content + '\n</file>\n' + question_prompt
```

图 4 模型生成答案 Prompt，其中 file_content 为文档的内容，final_prompt 为最终输入给模型的 Prompt。

针对招股书合规性审核等任务，这些模型可分别承担如财务摘要识别、风险因素抽取、核心信息生成等子任务的生成响应，支持多文档输入、内容压缩与表格解析能力，是高结构性金融文档处理的重要基础。其具体回答构造方式见图 4 所示，核心 Prompt 由文档内容与标准问题模板组成。

2、模型调用方式支持

评测系统支持 HTTP 调用、Python SDK 与本地部署三种模型接入方式，便于适配不同模型接口与企业环境。若模型兼容 OpenAI SDK，可直接通过 API Key 完成对接。该模块为后续将评测任务集成至招股书自动审核平台提供了良好的扩展基础。

（三）打分模型评测过程

为确保评估体系能够准确反映大语言模型在金融文本处理中的实际能力，本文选用 GPT-4 Turbo[13]、Qwen2-72B-Instruct 和 Doubao-pro-32k 三款中文表现突出的模型作为打分工具，并通过官方推荐方式完成接入与部署，以保障评分过程的稳定性和一致性。

4.、Alphabox - https://www.alphabox.top/

5、ChatDOC - https://www.chatdoc.com/

6.、ChatGPT4 - https://chatgpt.com/?model=gpt-4

7、ChatPDF - https://www.chatpdf.com/

8、 Claude2 - https://claude.ai/

9、Moonshot - https://kimi.moonshot.cn/

10、 WarrenQ - https://www.warrenq.cn/

11、ERNIE Bot 3.5 - https://yiyan.baidu.com/

针对开放式与封闭式任务，FinLongEval 系统采用六个关键维度对模型生成回答进行评分，分别为：相关性、流畅性、连贯性、有用性、一致性和忠实度。该评分框架在设计上充分考虑了金融合规场景的需求，尤其适用于招股说明书等文档的结构化审阅任务。例如，在合规性审核中，模型不仅需回答准确，更需逻辑严密、表达规范、内容真实，避免幻觉或遗漏关键条款。

· 相关性（1 分）：回答内容与问题的匹配程度；

· 流畅性（2 分）：语言是否通顺、表达是否专业，适应招股书中对正式语体与书面规范的要求；

· 连贯性（2 分）：是否具备良好的逻辑组织与常识支持，避免因理解片段化而引发误读；

· 有用性（5 分）：回答是否完整覆盖关键信息，具备审阅价值，特别关注定量数据或重要业务描述是否充分支撑结论；

· 一致性（4 分）：是否准确回应问题，避免答非所问或信息断裂，尤其适用于合规条款与财务指标准确匹配的场景；

· 忠实度（1 分）：回答是否忠实于源文件内容，避免虚构信息或逻辑跃迁，对招股书审核中的“模型幻觉”具备直接警示作用。

该六维框架广泛适用于招股书合规性智能审核中的各类典型任务，包括发行人基本信息提取、风险因素归纳、财务数据分析等内容。通过标准化的 Prompt 构造（见图 5），打分模型可对生成结果进行一致、客观的自动化评估。

此外，针对涉及事实判断与数值推理的任务，如“风险表述是否存在逻辑漏洞”或“财务指标是否计算准确”，系统设立了专门的陷阱问题和数据计算任务类型。此类任务采用专门设计的 Prompt（见图 6 与图 7），聚焦于幻觉识别、错误推理与计算可靠性，进一步模拟大模型在高风险文本审核中的应用边界。

通用任务评估 Prompt:
任务描述: 根据给定的问题、参考答案和待评估的答案，按照 6 个标准对评估的答案进行评分：相关性、流畅性、连贯性、有用性、一致性和忠实度。首先给出一个整数分数，接着提供解释。
六维度评估标准:
{EVALUATION_PROMPT}
问题:
{question}
参考答案:
{reference_answer}
待评估的答案:
{generated_answer}
输出格式示例:
相关性: X 分
理由: [解释]
流畅性: X 分
理由: [解释]
...

图 5 打分模型评测 Prompt

陷阱问题评估 Prompt:
任务描述: 根据给定的问题、参考答案和生成的模型答案，判断生成的答案是否包含捏造内容，并提供理由。
问题:
{question}
参考答案:
{reference_answer}
生成的模型答案:
{generated_answer}
输出格式:
编造/未编造
理由: [解释]

图 6 陷阱问题评测 Prompt

数据计算评估 Prompt:
任务描述: 根据给定的数据计算问题、参考答案和生成的模型答案，评估生成的答案是否计算正确。最终计算结果必须与参考答案完全一致，才能被视为正确。
问题:
{question}
参考答案:
{reference_answer}
生成的答案:
{generated_answer}
输出格式:
正确/错误
理由: [解释]

图 7 数据计算评测 Prompt

（四）评估结果

本节展示了 13 个大语言模型在开放式任务和封闭式任务上的评测结果，如表 1 所示，并进一步分析了它们在陷阱问题与数据计算任务中的表现。评测结果显示，ChatGPT 4 和 Claude2 在整体性能上表现最为出色，尤其在“有用性”和“一致性”维度上优势明显。前者在开放式任务中表现均衡，后者则在封闭式任务中得分显著领先，这与它们所采用的大规模训练数据与优化策略密切相关，使其在处理复杂任务时具备更强的稳定性和忠实度。

相比之下，开源模型如 GLM-4-9B-chat 与 Qwen2-7B-

表 1 GPT-4 Turbo 对当前已有大语言模型在封闭式任务和开放式任务上的评测结果

模型名称	访问方式	相关性	流畅性	连贯性	有用性	一致性	忠实度	总分
开放式任务								
Alphabox	网页端	0.99	2	1.97	4.09	3.03	0.76	12.84
ChatDOC	网页端	0.94	2	1.76	3.21	2.27	0.53	10.71
ChatGPT 4	网页端	1	2	1.95	4.31	3.2	0.88	13.34
ChatPDF	网页端	0.85	2	1.65	2.59	1.78	0.45	9.32
Claude2	网页端	0.99	2	1.92	4.1	2.88	0.76	12.65
Moonshot	网页端	0.95	2	1.92	3.74	2.53	0.65	11.79
WarrenQ	网页端	0.89	1.99	1.54	2.42	1.7	0.4	8.94
ERNIE Bot 3.5	网页端	0.89	1.99	1.67	2.95	2.02	0.4	9.92
GLM-4-9b-chat	本地部署	0.99	2	1.86	4.1	2.9	0.65	12.5
Qwen2-7B-Instruct	本地部署	0.98	2	1.82	3.88	2.73	0.68	12.09
ChatGLM3-6B-128k	本地部署	0.97	2	1.73	3.72	2.69	0.65	11.76
Doubao-lite-128k	API 调用	0.98	1.99	1.74	3.6	2.72	0.7	11.73
ERNIE-Speed-128K	API 调用	0.98	2	1.71	3.68	2.6	0.61	11.58
封闭式任务								
Alphabox	网页端	0.97	2	1.89	3.94	2.64	0.5	11.94
ChatDOC	网页端	0.97	2	1.78	3.44	2.67	0.72	11.58
ChatGPT 4	网页端	1	2	2	4.17	3.11	0.78	13.06
ChatPDF	网页端	0.81	1.97	1.53	2.31	1.38	0.28	8.28
Claude2	网页端	1	2	2	4.6	3.6	1	14.2
Moonshot	网页端	1	2	1.96	4	2.73	0.62	12.31
WarrenQ	网页端	0.97	1.94	1.67	3.06	2.22	0.58	10.44
ERNIE Bot 3.5	网页端	0.91	2	1.74	2.37	1.37	0.34	8.73
GLM-4-9b-chat	本地部署	1	2	1.73	4	3.19	0.65	12.57
Qwen2-7B-Instruct	本地部署	1	2	1.73	4.15	2.96	0.62	12.46
ChatGLM3-6B-128k	本地部署	1	2	1.65	3.85	3	0.69	12.19
Doubao-lite-128k	API 调用	1	2	1.85	3.92	3	0.65	12.42
API 调								

表 2 已有大语言模型的陷阱问题编造率和数据计算正确率结果统计表

模型名称	陷阱问题编造率	数据计算正确率
Alphabox	65%	0
ChatDOC	70%	14%
ChatGPT 4	40%	86%
ChatPDF	28%	14%
Claude2	60%	43%
Moonshot	35%	29%
WarrenQ	75%	0
ERNIE Bot 3.5	75%	0
GLM-4-9b-chat	40%	29%
Qwen2-7B-Instruct	70%	14%
ChatGLM3-6B-128k	75%	0
Doubao-lite-128k	65%	43%
ERNIE-Speed-128K	70%	0

Instruct 在相关性与流畅性方面接近商用模型，但在一致性与忠实度上仍有一定差距，反映出当前开源模型在高精度任务中的局限性。不过，其发展潜力依然值得期待。相较之下，ChatPDF 和 ERNIE Bot 3.5

的整体表现偏弱，尤其在封闭式任务中准确率不高，说明在应对结构复杂、逻辑严谨的金融文本时仍存在显著挑战。

金融领域对大语言模型在数据处理能力上的要求极为严格，评估大语言模型在数据计算任务中的表现是衡量其能否满足金融计算需求的关键标准。如表 2 显示，除了 ChatGPT 4 之外，其他大语言模型在数据计算能力上表现普遍欠佳，正确率较低。值得注意的是，尽管 Alphabox 和 WarrenQ 是专为金融领域开发的模型，其数据计算正确率却为 0%。这表明，尽管这些模型针对金融领域进行了专门的训练和调优，但在处理精确计算任务时仍存在严重不足，未来需要重点加强这一领域的改进。

结合上述结果，可以看出当前大语言模型在处理金融长文档，尤其是招股说明书这类高结构性、高风险文本中的表现差异显著。在招股书的合规性智能审核应用中，模型不仅需具备信息提取和语义理解能力，还必须在关键数据核验、风险措辞判断等方面保持高度一致性和忠实度。因此，模型在陷阱问题和数据计算任务上的表现可直接反映其在招股书审阅场景中的可用性与可信度。整体来看，商用模型在高精度合规审查任务中仍具显著优势，而开源模型尚需在核心能力上进一步打磨，以实现更广泛的智能审阅落地。

（五）评测系统界面展示

1、评测任务选择

可供选择的评测任务如图 8 所示。

图 8 评测任务选择

2、待评测模型选择

已有评测结果的模型如图 9 所示，除此之外，还支持评测其他模型，如果要评测的模型兼容 OpenAI SDK，可以直接通过上传 API Key 和服务 endpoint 进行评测，如图 10 所示。

图 9 已有评测结果模型选择，其中“通”代表通用模型，“垂 - 金”代表垂直领域 - 金融模型

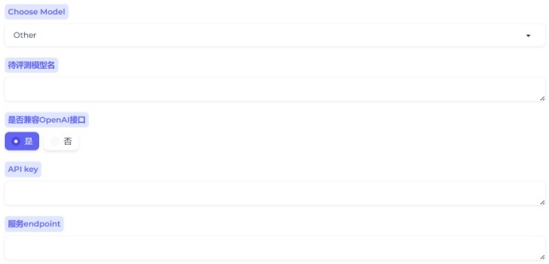


图 10 OpenAI SDK 兼容模型评测

3、打分模型选择

可供选择的打分模型如图 11 所示。

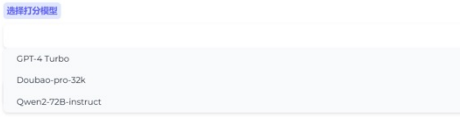


图 11 打分模型选择

4、评测结果展示

如图 12 所示，评测任务选择所有，打分模型选择 GPT-4，待评测模型选择 Alphabox，然后点击 Submit 进行评测，就可以得到该模型的开放式和封闭式任务的六维度评估结果以及数据计算正确率和陷阱问题编造率。

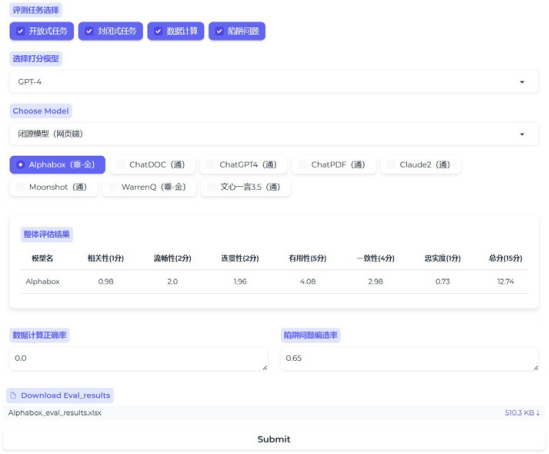


图 12 评测结果展示图

（六）面向招股书场景评测任务设计

为增强评测系统在金融监管实际场景中的代表性与适用性，本文在 FinLongEval 任务体系中引入了以招股说明书和保荐书为代表的重要金融长文档作为评测素材，构建覆盖 IPO 审核关键环节的信息抽取类任务。招股书文本结构复杂、篇幅较长，内容涉及发行人基本情况、股权结构、核心技术、财务指标、募集资金用途、风险因素、行业趋势等多个板块，是信息密集型、高规范性的代表性文本。结合“上海合晶硅材料股份有限公司”和“中

电港”等企业的招股材料，本文设计了一批高质量的任务样例，覆盖信息提取、阅读理解和表格提取三类评测任务，全面反映模型在合规性审查过程中的关键能力要求。任务示例包括对发行人基本信息、控股结构、主营业务和风险因素的提取与解读，对三年财务指标变动趋势的理解，以及对募集资金项目、证券发行基本情况等表格化信息的准确解析。这些任务不仅具备明确的参考标准和标注依据，便于多模型横向比较和自动化评分，同时贴合监管审阅实际流程，有助于验证大语言模型在智能化尽职调查与招股文件审阅中的实用能力，进一步拓展 FinLongEval 系统在监管科技场景下的应用深度。

为便于展示大模型在招股书场景中需应对的典型任务形式及其对应的评测类型，本文整理部分代表性问题如表 3 所示，这些问题均选自真实招股材料，并经标准化处理后用于评测系统任务生成与模型能力测试。

表 3 评测案例展示

场景	任务类型	模型能力要求	示例问题
招股书风险因素识别	阅读理解	多维风险解析与文本总结能力	发行人面临的主要风险有哪些？要求：对每种风险因素分别进行描述。
资金用途合理性判断	表格提取	表格+文本融合处理，项目投资结构理解	告诉我，发行人的募集资金的投资项目情况。要求表格呈现+每个项目介绍。
科创板定位合规判断	信息提取	政策条款对照能力、逻辑一致性验证	发行人符合科创板定位的依据是什么？

五、总结与展望

综上所述，本文构建了面向金融长文档场景的大语言模型评测系统 FinLongEval，并设计六维度评估框架，用以系统评估模型在金融文本理解、关键信息提取与语义一致性方面的表现。通过对 13 个主流大语言模型的实验结果分析显示，当前模型在长文档处理方面尚存在显著提升空间，尤其在数据计算准确性与陷阱问题识别等高风险任务中表现有限。在招股书等合规性审阅场景中，这些能力正是决定模型可用性和安全性的重要因素。

本评测系统以招股书智能合规审核为主要应用导向，特别关注招股说明书等高结构性、高合规要求文档所需的关键能力，如法定要素提取、风险披露判别与财务逻辑校验等。系统任务设计紧贴实际审核流程，可为招股书合规性审核中的模型选型提供可靠参考。未来工作将重点围绕三个方面展开：一是将系统从 Demo 形态部署为可远程调用的服务平台，以支持实际审阅任务；二是引入更高质量的中文打分模型，提升系统对复杂问答与逻辑一致性的评判能力；三是扩展评测样本，重点覆盖不同类型的招股书文本，进一步增强系统在金融合规智能审核场景下的实用性与扩展性。

参考文献：

[1] Fei Z, Zhang S, Shen X, et al. InternLM-Law: An Open Source Chinese Legal Large Language Model[J]. arXiv preprint arXiv:2406.14887, 2024.

[2] Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.

[3] Zhou Z, Shi J X, Song P X, et al. LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model[J]. arXiv preprint arXiv:2406.04614, 2024.

[4] Islam P, Kannappan A, Kiela D, et al. Financebench: A new benchmark for financial question answering[J]. arXiv preprint arXiv:2311.11944, 2023.

[5] Zhang L, Cai W, Liu Z, et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models[J]. arXiv preprint arXiv:2308.09975, 2023.

[6] Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models[J]. arXiv preprint arXiv:2211.09110, 2022.

[7] Zheng L, Chiang W L, Sheng Y, et al. Judging llm-as-a-judge with mt-bench and chatbot arena[J]. Advances in Neural Information Processing Systems, 2024, 36.

[8] An, C., Gong, S., Zhong, M., Zhao, X., Li, M., Zhang, J., Kong, L., Qiu, X. L-Eval: Instituting standardized evaluation for long context language models // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand, 2024: 14388-14411.

[9] Shaham U, Ivgi M, Efrat A, et al. ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 7977-7989.

[10] Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., Li, J. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand, 2024: 3119-3137.

[11] GLM T, Zeng A, Xu B, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools[J]. arXiv preprint arXiv:2406.12793, 2024.

[12] Yang A, Yang B, Hui B, et al. Qwen2 technical report[J]. arXiv preprint arXiv:2407.10671, 2024.

[13] Achiam J, Adler S, Agarwal S, et al. GPT-4 Technical Report[J]. arXiv preprint arXiv:2303.08774, 2023.

面向监管的金融舆情大模型系统及实现 *

马朝阳 [1]，王新宇 [1]，杜威 [1]，梁佳艺 [1]，吴苑斌 [1]，王晓玲 [1]，杨忠良 [2]，周琳娜 [2]
[1] 华东师范大学 | [2] 北京邮电大学

摘　要： 随着互联网的快速发展，金融与互联网的联系日益紧密，海量的金融舆情信息迅速传播并动态变化，对市场秩序产生重要影响。为了应对金融舆情信息爆炸带来的分析挑战，本文实现了一套面向监管的金融舆情大模型系统。首先采集多源金融数据，并利用大模型实现自动标注。然后基于预训练的大语言模型，结合参数高效微调方法构建多任务框架，实现金融事件检测、网络黑嘴识别和研报生成等功能。最后通过前后端技术开发，集成算法与可视化模块，构建面向金融监管的舆情分析系统。

关键字： 金融舆情；大语言模型；参数微调；系统实现

一、引言

在如今数字化以及互联网快速发展的背景下，金融领域的舆情信息膨胀十分迅速。在金融领域，舆情信息始终保持快速传播和动态变化的状态。对于某个公司有益的舆情信息，可能会促进股价上涨从而推动公司的发展，但是负面的信息可能会引起相反的结果，对公司以及中国市场的经济造成不可想象的损失。投资者间的信息交互行为会进一步对市场个体间所持有的认知、情绪和意见产生影响，这不仅推动金融网络舆情的演化，还同时影响股票市场投资者决策行为，造成股票市场剧烈震荡。由此可见金融舆情事件的涉及面广、影响深远，特别是在市场波动时期，大量的舆情信息往往能显著影响市场秩序。根据 2023 年中共中央、国务院印发的《党和国家机构改革方案》[1] 中强调，中央金融委员会的建立可以推进金融稳定发展的顶层设计等工作，推动及时发现金融领域的舆情风险。互联网金融舆情的分析以及处理属于当下金融智能领域重点研究方向之一，借助对海量舆情数据加以建模并分析，可切实识别潜在金融事件、辨别不良舆论风险，给市场监管和投资决策给予有力的数据支持。不过金融舆情文本一般有着非结构化程度高、领域术语众多以及标注数据匮乏等特性，在传统的金融舆情分析中，大多数的内容是通过人工收集、分析和汇总的方式进行处理，这不仅效率低，而且随着数据量的剧增，人工处理的方式逐渐无法满足快速发展的需求。

随着人工智能技术对金融文本进行分析的方法逐渐成熟，本文致力于构建一个面向监管的金融舆情大模型系统，系统集成人工智能技术相关技术，提供舆情处理和分析功能，迎合多样化的互联网金融舆情分析任务。系统以大语言模型为基础，结合高效的微调方法，把金融舆情处理构建成三个具体子任务，即金融事件检测、网络黑嘴识别以及研报生成。整个系统通过自动化采集、模型调用以及直观展示，构建起了一套完备的业务流程框架。

二、金融舆情处理大模型

为提升大模型在金融领域的任务表现，本文聚焦金融事件检测、网络黑嘴识别与研报生成三大任务，基于开源的大语言基座模型进行指令微调 [2]（Instruction Tuning），基于 ChatGLM2-6B 的金融舆情处理大模型的整体微调架构如图 1 所示。本文提出的金融舆情处理大模型在逻辑上采用三层架构设计，从数据到任务实现端到端的优化。

（1）数据层

由于通用大模型在应对金融领域中的信息抽取与生成等任务时往往存在理解偏差、表达不准确等问题，所以有必要通过领域特定的数据对模型进行微调，使大模型在金融术语理解和结构化输出方面表现优异。并且，金融舆情分析领域里的公开数据集大多集中于特定的子任务，其构建目标和评价体系存在着较大差异，难以覆盖本文中涉及的多任务需求以及复杂多变的金融舆情应用场景。为此本文在融合开源数据的基础上，结合数据采集技术扩展数据的来源，构建了一个面向多种金融场景的中文金融舆情数据集，支持多样化的舆情分析需求。

（2）模型层

本文挑选了拥有较强中文处理能力的 ChatGLM2-6B[3] 大语言模型，该模型是基于通用预训练模型 GLM（General Language Model）的基础上量化优化的对话模型，采用了 Transformer Decoder 结构，包含 32 层注意力网络，每层配备 32 个注意力头，隐藏层维度为 4096。

通过旋转位置编码技术，模型能够有效捕捉长距离依赖关系，支持最大 2048 个 token 的上下文窗口，非常适合处理金融文本的复杂语义和长文档分析。所以它拥有十分强大的推理能力和上下文理解能力，同时该模型经过高效压缩和优化，可在消费级显卡上部署。本文同时使用了参数高效微调技术 LoRA[5] 对 ChatGLM2-6B 进行定制化训练，在保持模型通用能力的同时显著提升了金融领域任务的性能表现。

（3）多任务应用层

基于金融舆情的业务需求，模型实现了三大核心功能。金融事件检测任务针对金融文本如新闻、评论等，识别金融事件类型以及事件中的要素信息；网络黑嘴识别任务着重分析舆情文本内容，分辨其中有无误导性、虚假性或操纵市场意图的信息，以便及时发现潜在风险；研报生成任务要求模型依据输入的关键信息点，自动撰写符合金融领域规范的完整研报段落，提高信息加工与决策支持效率。

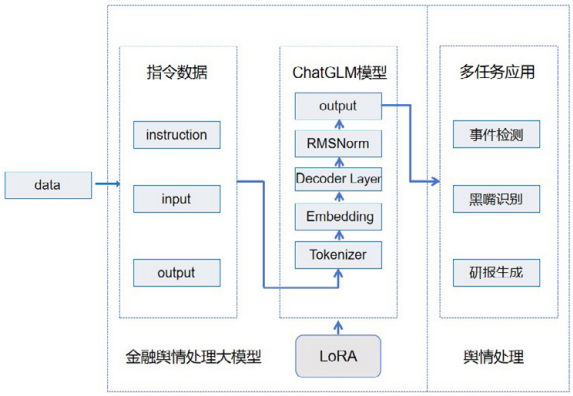


图 1 模型微调架构

* 本文是项目下设课题“开放域舆情风险识别、预警与处置”（课题编号：2021YFC3340702）的研究成果，课题负责：王晓玲（华东师范大学）。

三、实验与结果分析

3.1 训练数据集构建

本文采用的数据用于金融事件检测、网络黑嘴识别和研报生成三个任务，采集的数据主要来源于专业金融平台、贴吧等，对采集的舆情信息进行标准化处理与存储，主要包括清洗无效字段和冗余标签，去除空格、换行符、特殊字符等，统一输出格式为 JSON 文件等步骤，每条记录包括舆情标题和对应的正文内容等信息。

完成数据收集之后要进行数据的标注，此过程通过构建有“指令—输入—输出”三元组结构的数据样本，来契合大语言模型在指令微调阶段对数据格式的要求，传统的人工标注方式往往耗费时间长且成本高，很难契合大规模数据的需求。本文选用当前主流的大语言模型 ChatGPT 作为标注工具，ChatGPT 是由 OpenAI 在 2022 年发布的一款聊天机器人大模型，它能够给予在预训练阶段所见的模式和统计规律来生成回答，可以完成翻译、问答、分类、推理等各种任务，具有强大的自然语言处理能力，其后续版本 GPT4[4] 是 OpenAI 在 2023 年发布的大模型，综合性能进一步提升，在多个自然语言任务中已接近甚至达到人类水平，相比 ChatGPT，GPT4 会对事件的标识有更准确的判断。本文先使用 ChatGPT 进行标注，后使用 GPT4 对 ChatGPT 后标注的事件进行精标，以进一步提高数据集的质量。数据标注过程包括以下几个步骤：提示词设计、数据格式转化、长文本过滤和不重复均匀采样等步骤，以此针对三种不同类型的任务，构建了相应的任务数据集。

下面仅介绍标注后的金融事件检测数据集的示例。

金融事件检测数据集包含了事件类型识别数据集和要素抽取数据集两种类型，事件类型识别旨在判断一段金融新闻文本中涉及的预定义事件类型，如“控股股东变更”“减持事件”等，为事件结构化处理打下基础。事件要素抽取进一步在识别出的事件基础上，抽取对应的关键论元要素，如“交易对手方”“交易金额”“增持比例”等。本文基于金融新闻数据，定义了 18 类候选事件标签与 80 余类细粒度事件论元类型，涵盖 IPO、增持、减持、回购、股权质押、限售解禁、诉讼纠纷、财务异常、高管变动等典型金融舆情事件，确保任务具有丰富的标签语义，满足舆情建模下游任务需求。

数据集基于指令式监督学习范式进行构建，采用 <instruction, input, output> 三元组结构。如图 2 所示，展示的是事件类型识别的示例数据，每条数据包括了：

instruction：对模型提出的具体任务指令如“请检测新闻文本对应哪种事件”或者“请检测新闻文本有哪些事件论元”。

input：金融新闻原文及事件候选列表或者论元候选列表。

output：对应的标准结构化答案，使用 JSON 格式返回事件类别或者事件论元。

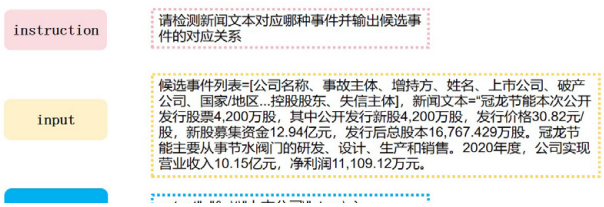


图 2 金融事件类型识别数据集

3.2 LoRA 微调

本文使用了 LoRA[5] 技术对 ChatGLM2-6B 进行参数高效的领域适配训练，LoRA 可在不改变大部分预训练模型参数的情形下，仅对新增的可训练低秩矩阵进行优化，从而降低微调成本以及减少显存占用。其具体流程包含以下几个步骤：首先按照 ChatGLM2-6B 的网络结构设计低秩矩阵，接着对模型代码进行修改，在 Transformer 的注意力层中引入低秩分解模块，代替原有的全参数更新策略，使其与原始网络协同工作。最后在微调阶段，模型仅更新嵌入的低秩权重，让其他模型权重维持冻结状态。

本文针对三类任务分别训练了独立的 LoRA 适配器模块，各个适配器都专门针对与之对应的任务来进行低秩参数更新，每个 LoRA 适配器在对应任务上单独进行训练。原始权重矩阵为 $W \in R^{d\times d}$，LoRA 的做法是引入两个低秩矩阵 $A \in R^{d\times r}$、$B \in R^{r\times d}$，其中 $r \ll d$，用来近似权重变化量，即：

$$\Delta W = AB \tag{1}$$

最终，微调后的权重表示为：

$$W' = W + \Delta W = W + AB \tag{2}$$

其中，原始权重 W 保持冻结，仅训练 A 和 B。这样一来，在主体模型权重保持冻结的状况下，可迅速适配不同任务的需求，系统在推理阶段的时候，可以依据当前任务动态加载相应的适配器权重，灵活地支持金融事件检测、网络黑嘴识别、研报生成多种任务类型，使得模型在多任务环境下的切换效率得到提高，处理并行性也有所提升。

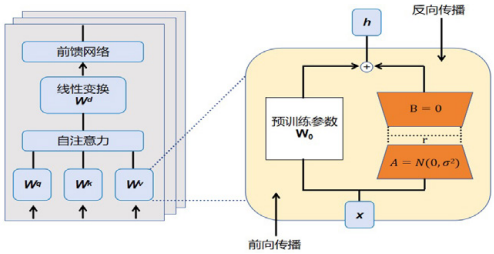


图 3 LoRA 原理

3.3 实验结果

本文对于金融事件检测和网络黑嘴识别任务采用 F1 分数的宏观平均值作为评价指标，F1 值是召回率（Recall）和精确率（Precision）这 2 项指标的调和平均。可以从表 1 中看出，与直接运用预训练模型来进行推理的基准方法相比较，经过微调处理后的模型在性能指标方面出现了明显提升，表明了模型可精准区分不同种类的金融事件，较为稳定地判定文本里是否存在某类金融事件，以及对黑嘴有一定的识别能力。

表 1 事件检测和黑嘴识别结果评估

任务类型	F1 值(LoRA)	F1 值(zero-shot)
事件类型识别	81.00	65.00
事件要素抽取	72.30	58.50
黑嘴识别	72.36	62.40

针对研报生成任务，本文借助 ROUGLE-L 和 BLEU-4 这两个指标去验证模型的效果。ROUGLE-L 用于计算生成文本和参考的财报文本之间的重叠度来评估生成文本的质量，BLEU-4 比较模型输出文本与参考文本的相似度来度量模型。可以在表 2 中看出，模型经过微调之后生成的研报与参考文本的一致性具有显著提升，并且在结构和内容覆盖方面较为完整，内容比较连贯。

表 2 研报生成结果评估

任务类型	BLEU-4	ROUGE-L
研报生成(LoRA)	27.0	43.6
研报生成(zero-shot)	18.0	35.2

四、面向监管的金融舆情大模型系统

经过对网络金融舆情以及相关系统展开调研，本文设计了一套面向监管的金融舆情大模型系统，系统后端采用 Django 框架，通过调用并集成微调后的 ChatGLM2-6B 多任务模型，前端基于 Vue.js 与 Element UI 开发交互式可视化界面。本文设计的这套系统适用于企业，又适合政府以及其他管理机构使用，该系统可识别文本当中的主体和事件，借助对事件展开分析，可帮助企业快速获取并处理网络信息，灵活应对舆情变化，政府以及相关管理机构也可借助该系统实时掌握网络热点事件，为政府与企业在舆情信息处理方面进行科学决策给予有效支持。

4.1 系统功能模块

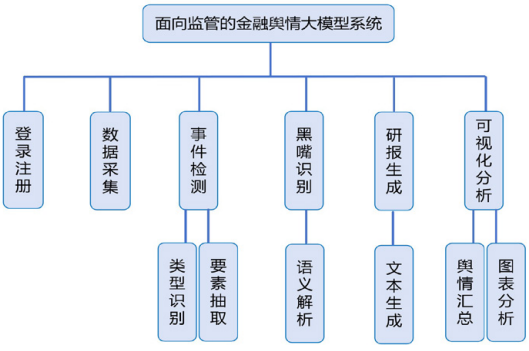


图 4 系统功能模块

互联网金融舆情监控系统的核心功能模块如图4所示。

（1）用户登录注册模块

提供用户身份验证与权限管理，支持普通用户与管理员两种角色，账户信息统一存储在数据库中。

（2）舆情数据采集模块

通过数据采集技术自动采集金融相关文本数据，采集内容标准化处理后存入数据库。

（3）金融事件检测模块

利用大模型解析舆情文本，自动识别金融事件及其关键要素，输出事件类型与论元。

（4）网络黑嘴识别模块

基于大语言模型识别可疑评论与文章，结合语义和情感分析判断是否存在虚假信息并说明原因。

（5）研报生成模块

整合舆情与事件数据，结合用户输入生成结构化金融研报，依托向量检索与微调大模型完成内容撰写。

（6）舆情可视化分析模块

以图表形式展示舆情数据，主要包括：舆情汇总、历史趋势、黑嘴分析、类型分析、研报分析等。

4.2 系统架构

本系统采用分层架构设计，从下至上划分为基础设施层、算法能力层、数据处理层、业务服务层、接口层与表现层。各层职责分明，相互解耦，增强了系统的可维护性与可扩展性，用户鉴权机制贯穿核心层级，确保系统安全与数据访问控制，系统整体架构图如图 5 所示。

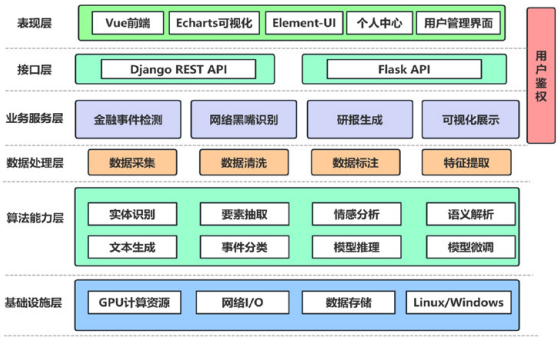


图 5 系统架构图

4.3 系统展示

这一小节将给出系统实现的部分模块示例。

图 6 系统首页展示了系统通过数据采集得到的各类平台上的数据数量，以及舆情数量趋势图和来源趋势图，还列举了部分最新舆情展示，支持点击查看详细信息，便于用户快速获取最新热点信息。

图 6 系统首页

图 7 中所示为金融舆情智能识别和处理系统的数据采集模块界面示例。该模块旨在自动化采集金融领域相关内容并存储记录，为后续舆情分析与事件检测等提供原始数据支持。

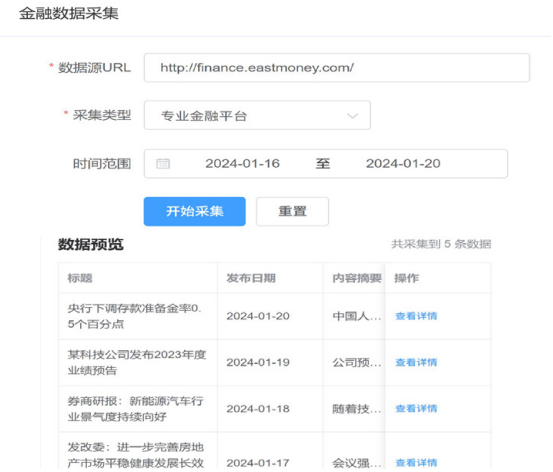


图 7 数据采集示例

金融事件检测模块基于大模型推理能力，对收集到的舆情文本数据进行深入解析，从中自动识别潜在的金融事件，并对事件中的关键要素进行结构化提取，输出金融事件类型和事件论元。图 8 给出了事件检测使用示例。

图 8 事件检测示例

网络黑嘴模块依托大语言模型，通过文本识别与分析，对网络中可疑的评论、文章进行识别。结合语义检测、情感分析等信息，判断是否存在虚假内容并给出具体原因。该模块能够对用户输入的原始金融新闻或公告文本等进行判断是否为网络黑嘴。图 9 给出了网络黑嘴识别的使用示例。

图 9 黑嘴识别示例

数据可视化页模块包含了事件统计、黑嘴分析、研报分析三个子模块，分别对三大功能的使用情况进行数据可视化分析，其中使用了雷达图、柱状图、热词云图等可视化展示了数据的分布、趋势等特征。图 10 展示了事件检测的相关可视化统计示例。

图 11 展示了事件类型分布统计和研报来源分布的可视化示例。

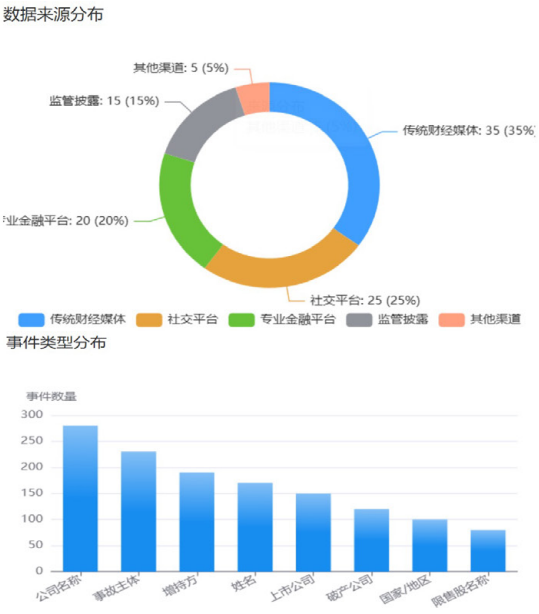


图 10 事件检测可视化示例

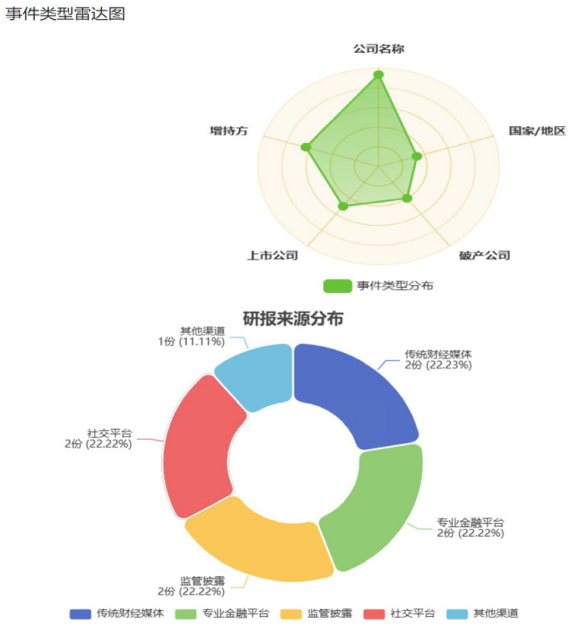


图 11 事件类型分布和研报来源分布可视化示例

五、总结与展望

本文围绕金融舆情处理的核心需求，构建了一套包括技术理论、算法模型和系统实现的完整解决方案，虽然研究成果于金融舆情处理方面有一定成效，但以下几个方向有待剖析：

其一，提高模型的泛化能力，当前模型在细分金融场景里的表现存在不足，后续要借助增加增量学习等技术，强化模型的泛化能力。

其二，进行多模态数据融合分析，现有的系统主要侧重于文本处理，未来可拓展到视频舆情、语音识别等领域，构建多模态融合的舆情分析框架。

其三，实现系统轻量化部署，针对高频舆情场景，可增添边缘计算、模型压缩等技术，提高系统响应效率和性能。

其四，完善系统功能，系统目前的功能主要是围绕三大核心功能开展，未来可增添更多个性化功能，比如实施舆情预警功能、用户画像构建、热点舆情个性化推送等功能，丰富系统的用途。

参考文献：

[1] 国务院．党和国家机构改革方案 [EB/OL]. https://china.huanqiu.com/article/4C6P9KT0GdN, 2023.

[2] Zhou M., Yang F., Wang S., et al. Instruction tuning for large language models: A survey[J]. arXiv preprint arXiv:2306.04761, 2023.

[3] Du Z., Qian Y., Yang A., et al. GLM: General language model pretraining with autoregressive blank infilling[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022: 320–335.

[4] OpeniAI. GPT-4 Techinical Report[J]. arXiv preprint arXiv:2303.08774,2023.

[5] Hu E. J., Shen Y., Wallis P., et al. LoRA: Low-rank adaptation of large language models[C]//Proceedings of the 10th International Conference on Learning Representations , 2022.

融合金融舆情的股市态势分析技术及实现 *

戴雨霖，吴苑斌，王晓玲
华东师范大学

摘　要： 本文针对当前股价预测方法普遍依赖单一历史股价序列、忽略舆情信息而导致预测准确性不足的问题，提出一种融合大小模型的多模态股价预测方法。该方法利用小模型对股价序列进行初步预测，并将预测结果与对应时间段内的舆情信息一并输入大模型，辅助其进行结果修正，从而提升预测精度。为实现时间序列与文本信息的有效融合，本文采用序列分块方式，将时间序列转换为类文本格式，与自然语言对齐后输入模型。在多个数据集及不同预测长度任务上，本方法均优于现有模型，性能提升超过 10%。此外，实验证明，大模型在时间序列预测中依然符合尺度规律，并具备良好的少样本泛化能力。

关键字： 大语言模型；多模态；时序预测；数值推理；融合预测

一、引言

在金融行业高度竞争的背景下，数据科学技术始终扮演着至关重要的角色。股票交易作为金融领域的核心组成部分，长期以来吸引着学术界与产业界的广泛关注。股票市场不仅承担着资源配置和风险对冲等重要功能，其价格走势还受到公司基本面、宏观经济、突发事件以及公众情绪等多种因素的共同影响。因此，如何借助先进技术手段对股价走势进行准确预测，已成为量化交易研究中的关键课题之一。

传统的股价预测方法多依赖于结构化历史数据和统计建模技术，这些方法在建模稳定数据时具有效果，但在面对现实股市中的非线性、高噪声和数据稀疏等问题时，往往存在泛化能力不足的缺陷。随着信息环境的日益复杂，社交媒体、新闻报道、公告等非结构化数据的激增使得投资者对信息的依赖不断向多模态发展，股价的变化越来越体现出多源信息交互影响的特征。这一趋势推动了预测范式从传统的单一模态向多模态融合的深度演进。

在此背景下，大语言模型（Large Language Model, LLM）作为自然语言处理领域的前沿成果，展现出强大的语义理解与推理能力，为金融非结构化信息的建模带来了新的可能。LLM 在少样本学习、零样本推理等方面表现出色，具备在数据稀缺情境下进行有效预测的能力，为年轻公司或低频交易标的的走势预测提供了理论基础。近年来，部分研究开始探索将 LLM 引入金融时序建模场景，尝试将文本情绪、政策公告等舆情因素与历史价格信息相结合，以期实现对股价的更深层次理解与预测。然而，LLM 本身的序列建模机制多基于自然语言的离散、平稳特性，直接应用于金融市场这一高度非平稳、连续变化的环境时仍面临诸多挑战。同时，其庞大的参数量与计算成本也限制了其在实时交易决策系统中的直接部署。

为有效应对上述问题，本文提出一种多模态融合的股价预测框架，利用轻量级模型处理结构化时序数据，同时引入 LLM 对非结构化舆情文本进行建模，通过基于分块编码、时序重构与模态对齐的机制，实现对多源信息的深度建模与融合。该方法既保留了小模型的效率优势，又充分释放了大模型在语义理解方面的潜力，为多模态金融数据的融合预测提供了一种可行路径。

* 本文是项目下设课题“开放域舆情风险识别、预警与处置”（课题编号：2021YFC3340702）的研究成果，课题负责：王晓玲（华东师范大学）。

二、多模态股价预测框架

2.1 引言

如上一章所述，股票市场的数据存在数据非平稳性强以及具有非常高的异构性等特点。传统的预测方法多基于单一模态而缺乏多模态信息互相补充。针对上述问题，在本章中提出了一种基于大小模型融合的多模态股价预测框架。该框架较好地融合大小模型各自的优势，并使用小模型的时序编码能力以及大模型的文本分析能力对股价作出了详细的预测。

本章采用的方法主要由以下几部分组成：小模型预测输出、输入分块策略、时序重编码、舆情文本前缀编码以及模态融合输出，整体模型架构可以见图 1。对于一个多

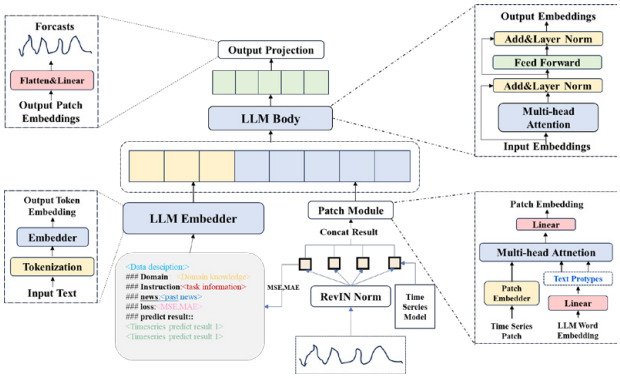


图 1 模型架构图

变量股价时间序列 $X \in R^{k\times d}$，其中 k 为时间窗口的长度，d 为变量维度，该序列首先经过 RevIN 归一化得到归一化数据，随后使用不同的小模型对时序数据进行预测。在得到预测结果后，对预测结果进行分块嵌入处理来实现模态的统一。随后利用 LLM 的分析能力，使用大模型收集各个小模型的预测结果，并分析各个小模型的损失，同时根据过去时间段内的新闻文本，经过大模型分析后对时序信息进行补充，从而得到最终结果。

值得注意的是，整个框架只有小模型的参数、时序重编码部分以及模态融合输出部分的参数需要经过计算以及反向传播进行更新，而主干大语言模型部分的参数则进行冻结处理。相较于需要经过完全对齐的多模态数据来更新所有参数的视觉领域大模型或者其他类型的多模态大模型，本节提出的方法由于不需要更新大量参数，因此非常易于优化，只需要少量时序与文本对齐的数据以及训练步数即可得到强大的预测效果。

2.2 问题描述

考虑到股价预测本质上时间序列分析问题，因此可以用如下问题来建模：考虑一段具有多个数值特征的股价时间序列 $P_{target} \in R^{k\times d}$，其中 k 表示时间窗口的长度，d 为变量维度例如股票开盘价和收盘价等；同时还有与时间序列对齐的舆情文本数据，表示为 $\mathcal{M} = \{M_1, M_2, \dots, M_k\} \in R^{k\times l\times w}$，其中 l 表示在一个时间步数内的新闻篇数，w 则表示一篇新闻内最大字符数量。对于新闻字符数不足 w 或者篇数不足 l 的舆情文本，将不进行补全或者补零处理。

本文将任务表述为一个多变量预测单变量的回归问题，其目标是预测目标股票在第 t 天的股票收盘价 p_i^t，将本框架的各个参数记录为 θ，整个过程可以用公式表述为：

$$\mathcal{X} \in R^{N\times T\times F} \rightarrow^{\theta} p \in R^{N\times 1} \qquad (1)$$

2.3 模型架构

2.3.1 小模型预测输出

每条输入数据首先通过 RevIN 归一化进行归一化处理，使其具有零均值和单位标准差。随后，为了充分利用不同小模型捕捉时序特征的能力，考虑使用多个架构差别较大的模型来实现对股价时序数据的分析预测，从而互补模型之间的优势并弥补单个模型的劣势。将归一化的时间序列输入各个小模型得到预测结果并将各个结果进行拼接，得到向量 $r \in R^{m\times n\times 1}$，其中 m 为小模型的个数，随后将 r 经过 RevIN 反归一化得到最终结果。同时记录各小模型的 MSE、MAE，以作为下一步 LLM 分析的参考与依据。

2.3.2 输入分块策略

对于各个小模型输出并拼接后的预测结果，考虑到时间序列模态是连续的模态，而自然语言的形式往往是离散的，因此可以考虑对连续的时间序列进行分块，把连续的序列转换为离散的时间块，每个块包含一个时间窗口内的部分信息，这些时间块就可以以类似于自然语言的形式来进行数据分析。考虑到自然语言的上下文连贯性，类似性质也应该在时间块之间得到体现，因此每个时间块之间的数据可以有部分的重叠来实现不同时间块的数据共享。

时间块的数量可以由以下过程计算得到。给定一个长度为 L 的时间序列，每个时间块的长度为 P，每个块之间不重叠部分的长度为 S，则该时间序列生成的时间块总个数为

$$N = \lfloor \frac{(L-P)}{S} \rfloor - 2 \qquad (2)$$

最终将得到时间块的序列形状为 $x_p \in R^{p\times N}$。

通过分块的策略，预测得到的时间序列将被转换为与自然语言相似的离散序列并用于进一步输入 LLM 进行分析。除了上述优势以外，分块策略也可以大幅提升模型计算效率。对于原有长度为 L 的时间序列，可以认为时间序列中的每个值都是一个 token，直接输入模型 f 会产生 O(f(L)) 的开销；而分块后则可以将 token 数量降低 $\frac{L}{S}$，这也意味着通过控制超参数 S 即可控制模型内存使用量与计算复杂度，从而降低模型计算开销。

2.3.3 时序重编码模块

虽然将时序数据进行了分块来实现了时序数据的离散化，但直接将时间块与自然语言的 token 序列一起分析仍存在问题，这是因为两者尚未对齐到同一语义空间。为了解决这一问题，本文使用了时序重编码模块，通过将时间序列映射到自然语言模态的形式来激活 LLM 对时序数据的分析和推理能力。不过自然语言为文本形式而时间序列为数字形式，直接用自然语言对时间序列进行无损地表示是无法实现的。因此在不进行微调的情况下使用 LLM 对时间

序列进行分析存在一定的困难。

考虑到上述问题，本文使用经过 LLM 预训练完成的词嵌入 $E \in R^{V \times D}$来对分块完成的时间块序列进行重编码，其中 V 为词表大小。考虑到直接使用原始的词嵌入其维度过大并包含了大量的冗余信息，并且金融领域各个词汇之间的相关性与普通语境下不同需要模型进行重新学习，因此考虑将词表内各个单词对应的向量进行线性映射来维护一个参数量较小的映射层 $\mathrm{E} \in R^{V' \times D}$，其中 V' ≪ V，具体过程可见图 2，可以看到图中 Stock 和 Price 这两个单词被分为了一类向量，并运用于接下来与时间块进行重编码。这个方法非常便捷有效并且能够让时间块学到更多的语义信息，从而使以后的 LLM 对时间序列进行更有效的修正。为了实现时序重编码，本文使用了一层多头注意力层来实现模态的融合。对于每个注意力头 k={1,⋯,k}，定义 Query 矩阵 $Q_k^{(i)} = \hat{X}_P^{(i)} W_k^Q$，Key 矩阵 $K_k^i = E' W_k^K$，以及 Value 矩阵 $V_k^{(i)} = E' W_k^V$，其中 $W_k^Q \in R^{d_m \times d}, W_k^K, W_k^V \in R^{D \times d}$，D 表示 LLM 在隐藏层的维度，$d = \lfloor \frac{d_m}{K} \rfloor$。完整重编码时间序列的过程可以用如下公式表示：

$$Z_k^{(i)} = ATTENTION\left(Q_k^{(i)}, k_k^{(i)}, V_k^{(i)}\right) = SOFTMAX\left(\frac{Q_k{}^{(i)} K_k^{(i)}}{\sqrt{d_k}}\right) V_k^{(i)} \quad (3)$$

通过对每个注意力头的 $Z_k^{(i)}$ 进行聚合来得到该模块的输出结果向量 $Z^{(i)} \in R^{P \times d_m}$，并通过一个线性映射来将 $Z^{(i)}$ 映射到大模型隐藏层的维度。

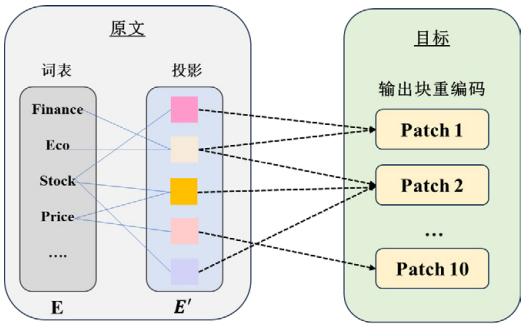


图 2 时序重编码模块

2.3.4 舆情文本前缀编码

在 LLM 的工作过程中，提示词起到了一个重要的指引性作用，合理的提示词可以激活 LLM 在不同任务的表现能力。如之前所述，直接将时间序列翻译成自然语言会导致信息上的丢失，同时在该模块中，考虑到时间序列单一模态的问题，需要舆情文本对该模态进行进一步补充。近几年的研究表明，其他数据模态 (例如图像) 可以无缝地作为提示词的前缀来进行整合，从而有效地基于这些输入来进行推理。因此，可以考虑：舆情文本信息是否也能作为提示词的一部分，并作为前缀来补充时间序列上下文的信息，从而激发 LLM 在时间序列领域的表现能力。

针对该想法，本文提出舆情文本前缀编码模块，并指引 LLM 进行预测。对于一个时间段的舆情文本数据，本文采用将这些文本拼接的方式输入给 LLM，不同日期间的舆情数据用 <sep> 分割。其中一个提示词样例可以见图 3

```
Description:CMIN-CN数据集由来自沪深300的300支股票组成
----------
<|start_prompt|>
### Domain: 股价受舆情文本影响剧烈，并且历史数据容易出现波动
### Instruction: 这是<M>个时间序列模型对前<T>步的股价预测结果, 结合舆情信息判断其预测正确性并预测接下来<L>步股价数据
###Company: 公司名字
### News: 过去时间段的新闻信息
### Loss: MSE,MAE
<|end_prompt|>
```

图 3 提示词样例

数据集描述为 LLM 提供了有关输入时间序列的基本背景信息，同时补充了舆情文本信息来让 LLM 对小模型的时序预测结果得到更好的判断。

2.3.5 模态融合输出

如图 1 所示，在完成上述所有操作之后，将拼接完成的向量输入给 LLM 进行分析，得到 LLM 的输出向量 T。舍弃 T 的文本前缀部分，只保留时间块部分，将该部分进行线性映射来得到最终的预测结果 Y。

三、实验设计与结果分析

3.1 实验设置

3.1.1 数据集与基线模型设置

在本节中将对模型性能进行一个详尽的评估。通过各种实验表明，本文提出的方法在各方面领先于基线模型，尤其是在少样本和零样本学习的背景下。在本次实验中主要使用以下三个数据集在固定时间窗口长度以及不同预测长度下进行评估：

ACL18 由美国股市中的 87 只股票组成，涵盖 9 个行业。它还包含了两种类型的数据：来自 Twitter 的推文和来自 Yahoo Finance 的历史股票价格。

CMIN-CN、CMIN-US[1] 数据集。其中 CMIN-US 包括美国按市值排名前 110 的股票；CMIN-CN 则包含中国主要股票市场沪深 300 指数的全部 300 只成分股。CMIN-US 和 CMIN-CN 均包含金融文本以及历史股票价格对齐的股票数据。这两个数据集中的历史价格数据均来自 Yahoo Finance。

本文将提出的方法与大量的 SOTA(State of the Art)

方法进行了比较。对于这些方法，本文均将其在上述三个数据集上进行了复现。本文采用的基线模型包含了一系列不同架构的模型，包括 Autoformer[2]、Non-Stationary Transformer[3]、iTransformer[4]、Informer[5]。对于其他架构的模型，本文也对其中表现出色的模型进行了复现比较，包括 DLinear[6]、LightTS[7]。并将最新的 SOTA 模型 Time-LLM[8] 与本方法进行了对比，Time-LLM 是基于 LLM 以及时序分块的模型，是目前时序预测领域表现最好的模型之一。

股价走势预测也是股市分析的一个重要组成部分，这是一个二分类任务，其目标是预测未来股价的上涨或下跌，在本节中进一步探索了本文提出的方法在股价走势预测方面的能力。在股价走势预测领域也提出了许多预测的方法，包括 ALSTM[9]，Adv-LSTM[10]， CMIN。其中 CMIN 模型也是多模态融合模型，集中了舆情文本信息与股价时序信息进行预测。本文均对比了上述模型的表现效果。

3.1.2 实验指标

对于股价预测这类回归任务，本文使用 MSE、MAE 这两类指标来进行评估，其计算公式如下：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 \quad (4)$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y_i}| \quad (5)$$

而对于股价走势预测这类分类任务，如之前的工作一样，选用了准确率 (Accuracy, Acc) 和马修斯相关系数 (Matthews Correlation Coefficient，MCC) 两个经典的指标来进行评估：

$$ACC = \frac{tp+tn}{tp+tn+fp+gn} \quad (6)$$

$$MCC = \frac{tp*tn-fp*tn}{\sqrt{(tp+fp)(fn+tp)(fn+tn)(fp+tn)}} \quad (7)$$

3.2 预测结果分析

3.2.1 全样本预测结果

全样本预测结果表明，所提出的方法在多个数据集和不同预测时间跨度下均展现出显著的性能优势。在 CMIN-US 数据集上，无论是短期还是长期预测，该方法均取得了最小的 MSE 和 MAE，显示出强大的鲁棒性和泛化能力。特别是在短期预测中，模型展现出对噪声的敏感性控制能力，而在长期预测中则体现了良好的动态建模能力。在 CMIN-CN 数据集上，该方法同样表现出色，即使在数据更为复杂和非平稳性更强的情况下，依然能够取得优于其他先进方法的结果，反映出模型在处理复杂市场环境中的适应性。在 ACL-18 数据集上，该方法在所有预测时间段均再次取得最优表现，随着预测步长增加，模型误差的增长速度显著小于其他方法，进一步说明了其在面对长期趋势

表 1 全样本预测结果（回归任务）

Methods		Ours		Stationary		iTransformer		Time-LLM		LightTS		Autoformer		DLinear		Informer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
CMIN-US	12	**0.143**	**0.191**	0.185	0.281	0.176	0.274	0.181	0.279	1.358	0.760	0.291	0.371	0.509	0.470	15.124	2.521
	24	**0.160**	**0.168**	0.357	0.397	0.342	0.386	0.343	0.391	12.763	2.414	0.489	0.490	1.458	0.802	21.063	3.317
	36	**0.329**	**0.371**	0.475	0.471	0.475	0.467	0.486	0.476	21.909	3.199	0.616	0.553	2.090	0.971	22.283	3.416
	48	0.707	0.565	0.791	0.603	**0.586**	**0.523**	0.602	0.530	28.701	3.922	0.702	0.584	2.704	1.112	21.638	2.618
CMIN-CN	12	**0.612**	**0.393**	2.561	0.627	0.849	0.449	0.665	0.413	2.352	0.745	2.303	0.656	0.708	0.414	18.097	1.923
	24	1.360	0.556	3.540	0.866	1.501	0.608	**1.320**	**0.588**	8.251	1.396	2.838	0.746	1.568	0.613	21.437	2.274
	36	**1.873**	**0.700**	5.955	1.084	1.949	0.706	2.062	0.731	12.129	1.797	4.218	0.964	2.522	0.777	18.097	1.923
	48	**2.437**	**0.792**	7.017	1.193	2.717	0.883	2.710	0.849	15.741	2.122	5.344	1.110	3.767	0.931	22.968	2.387
ACL-18	12	0.102	0.208	0.101	0.209	**0.086**	**0.193**	0.103	0.210	0.199	0.313	0.181	0.271	0.105	0.220	0.965	0.561
	24	**0.142**	**0.243**	0.167	0.268	0.169	0.274	0.153	0.256	0.690	0.591	0.332	0.368	0.226	0.327	1.645	0.856
	36	**0.252**	**0.332**	0.337	0.380	0.261	0.339	0.264	0.336	0.969	0.686	0.448	0.431	0.379	0.426	2.209	1.048
	48	0.349	0.388	1.278	0.742	**0.345**	**0.392**	0.357	0.398	1.097	0.736	0.685	0.533	0.512	0.487	3.108	1.252
1st count		8		0		2		2		0		0		0		0	

表 2 全样本预测结果（分类任务）

Methods	Ours		ALSTM		Adv-LSTM		CMIN		StockNet		DTML	
	ACC.	MCC	ACC.	MCC	ACC.	MCC	ACC.	MCC	ACC.	MCC	ACC.	MCC
CMIN-US	**67.14**	**0.251**	51.81	0.032	52.75	0.052	62.69	0.209	58.23	0.081	57.44	0.191
CMIN-CN	**57.79**	**0.059**	51.64	0.006	51.73	0.012	53.43	0.046	52.46	0.022	52.06	0.031
ACL-18	**59.37**	**0.161**	53.35	0.023	53.49	0.025	55.28	0.111	54.53	0.045	54.42	0.083

表 3 少样本预测结果

Methods		Ours		Stationary		iTransformer		Time-LLM		LightTS		Autoformer		DLinear		Informer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
CMIN-US	12	**4.354**	**0.999**	4.257	0.956	5.373	1.126	5.542	1.146	294.11	11.17	14.007	1.933	111.76	6.385	641.00	16.42
	24	**9.478**	**1.490**	15.595	2.036	10.548	1.605	11.073	1.810	481.46	13.98	17.754	2.285	138.30	7.155	630.13	16.33
	36	**10.783**	**1.594**	21.916	2.427	13.283	1.888	13.810	1.924	760.81	17.74	19.734	2.463	109.85	6.327	659.91	17.25
	48	**11.207**	**1.715**	26.427	2.806	16.386	2.173	16.319	2.137	1022.3	21.02	24.080	2.662	77.141	5.174	665.78	17.43
CMIN-CN	12	**6.654**	**1.249**	12.353	1.561	8.905	1.410	8.316	1.342	502.96	11.63	23.274	2.440	265.77	8.445	642.00	16.25
	24	**10.278**	**1.320**	16.137	2.045	16.744	1.990	15.694	1.916	739.76	14.27	30.091	2.896	361.43	9.756	631.13	16.43
	36	**15.562**	**1.944**	17.142	4.937	23.730	2.376	22.391	2.296	1014.1	16.41	38.654	3.116	275.17	8.491	661.91	17.54
	48	**22.317**	**2.526**	26.284	2.119	29.364	2.659	28.043	2.616	1124.1	17.92	46.167	3.528	146.01	5.876	662.78	17.46
ACL-18	12	0.774	0.221	1.222	0.731	0.809	0.578	**0.643**	**0.518**	168.10	8.10	2.742	1.367	15.949	2.795	195.34	10.10
	24	1.603	0.495	1.874	0.898	1.451	0.777	**1.321**	**0.753**	279.58	10.84	4.289	1.597	15.776	2.770	163.21	9.56
	36	**2.012**	**0.552**	2.667	1.082	2.327	0.973	2.962	1.124	277.74	10.83	6.977	2.773	14.267	2.614	201.53	11.04
	48	**2.489**	**0.712**	8.489	1.950	3.217	1.150	5.164	1.568	257.24	10.23	10.415	2.845	11.672	2.353	178.41	10.39
1st count		**10**		**0**		**0**		**2**		**0**		**0**		**0**		**0**	

和波动性增强时的稳定性和鲁棒性。本文采用的方法通过融合多模态信息，充分发挥了大小模型各自的优势，有效提升了股价预测的准确性和泛化能力。

本文进一步探索了模型在股价趋势预测方面的能力，其结果如表 2 所示，红色加粗表示在当前实验设置下效果最好的指标，而蓝色下划线则表示效果其次的指标。可以看出，本文所提出的方法在股价走势预测方面具有显著的性能优势。具体而言，该方法在 CMIN-US、CMIN-CN 和 ACL-18 三个数据集上均取得了最高的准确率（ACC）和马修斯相关系数（MCC）。这表明，该方法不仅能够准确预测股价的上涨或下跌，还能在不同市场环境下保持较高的预测一致性。与其他模型相比，该方法通过简单修改 LLM 的指令提示词和输出层结构，即可快速适配至股价走势的二分类预测任务，无需对模型本体进行复杂重构，体现了其良好的迁移学习能力。此外，该方法在 CMIN-CN 数据集上取得了 57.79% 的准确率，相较于其他模型如 CMIN 的 53.43%，显示出其在处理复杂市场数据时的优越性。这些结果进一步证明了该方法在多模态信息融合和股价预测任务中的有效性和泛化能力。

3.2.2 少样本预测结果与零样本预测结果

LLM 在少样本学习领域表现出强大的能力，即使只有少量的样本也能表现出良好的效果。因此在本节中，将探讨经过时序重编码后的 LLM 是否在预测时序数据时是否保留了相关的能力。在本实验中将仅仅使用少量样本（例如

仅有 10% 的训练数据) 对各个模型进行训练，并使用相同的训练集和验证集来进行评估。本节分别设置了各模型在 20% 训练数据下的表现，即取每个公司前 20% 天数的股价数据用于训练，如表 3 所示。由表分析可知，当训练样本较少的情况下，本文采用的方法所取得的效果要优于其他模型，这其中最终的原因可能就是我们同时利用大小模型各自的能力，也就是小模型的时序分析能力以及大模型文本分析和内在消化的能力。同时也可以看到，对比其余仅使用小模型的方法，Time-LLM 与本文采用的方法均没有出现大幅下滑，其中 Time-LLM 下滑了约 35% 左右的性能，而本文采用的方法下降了约 25% 左右的性能，其他小模型性能则出现了极大的下滑，这也说明了大语言模型在时序分析方面的潜力。

除了少样本学习以外，LLM 在零样本学习方面也有着巨大的潜力。因此在本节中，将探讨本文采用的方法在跨领域的零样本时序预测方面的能力。具体而言，本节中要实现的任务为：当一个时序模型在数据集 A 上进行训练后，评估其在数据集 B 上的预测效果，在预测前该时序模型不会对数据集 B 上的任何数据进行训练。各模型的表现效果如表 4 所示，箭头左侧为训练数据集，箭头右侧为测试数据的训练样本，时间窗口长度为 96，预测长度为 48。可以看到本文采用的方法所得到的效果也依然比众多模型要出色，相比于第二名 Time-LLM，其提升效果大约为 10%。对比上一节少样本学习部分可以发现，当现有数据量较少的情况下，本文采用的方法要优于其他模型，对比于 Time-LLM，虽然 Time-LLM 也使用了 LLM 来进行时序预测，但是其表现效果也不如本文采取的方法，这也是由于本文结合了大小模型的优势，通过小模型的结果以及舆情文本数据进一步激发了 LLM 在股价预测领域的能力。

表 4 零样本预测结果

Methods	Ours		Stationary		iTransformer		Time-LLM		LightTS		Autoformer	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
CMINCN→CMINUS	**0.408**	**0.428**	1.854	0.861	1.557	0.749	0.604	0.531	12.563	2.413	1.591	0.822
CMINCN→ACL18	**0.663**	**0.425**	1.637	0.545	1.732	0.578	0.758	0.376	1.456	0.642	1.389	0.612
CMINUS→CMINCN	**1.478**	**0.720**	9.235	2.645	6.261	2.078	2.112	0.762	23.763	5.828	9.091	1.844
CMIN-US→ACL18	**0.723**	**0.497**	1.737	0.641	1.001	0.478	0.758	0.376	2.406	0.732	1.289	0.597
ACL18→CMINUS	**0.568**	**0.438**	2.104	0.861	1.057	0.639	0.604	0.531	13.558	2.414	1.691	0.922
ACL18→CMINCN	**1.878**	**0.748**	10.562	2.745	5.261	1.378	2.112	0.762	22.862	5.621	9.091	1.844

四、总结

综上所述，本文围绕股价预测中模态信息缺失的问题，提出了一种融合小模型与大语言模型（LLM）的多模态预测方法，利用时间序列与舆情信息的协同建模，有效提升了预测的鲁棒性与准确性。通过构建跨模态对齐机制，使得自然语言信息能够辅助时间序列预测，弥补了传统方法在信息利用方面的不足。实验验证表明，该方法在多个预测任务中表现优异，且具备较强的泛化能力。

尽管本研究已取得一定成果，但仍存在改进空间。例如，当前 LLM 在处理长文本舆情信息时存在推理效率低与信息截断的问题，未来可探索更高效的长文本压缩或筛选机制；此外，本文主要关注单一公司数据，未充分考虑公司之间的关联性，未来可将企业之间的竞争或协同关系纳入模型建模范畴，从而进一步提升模型对现实市场的拟合能力。后续工作将继续拓展多模态融合方法的深度与广度，以实现更强泛化性与更高精度的股价趋势预测。

参考文献：

[1] Luo D, Liao W, Li S, et al. Causality-guided multi-memory interaction network for multivariate stock price movement prediction[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). 2023: 11207–11222.

[2] Wu H, Xu J, Wang J, et al. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting[C]//Advances in Neural Information Processing Systems. 2021, 34: 22419–22430.

[3] Liu Y, Wu H, Wang J, et al. Non-stationary transformers: Exploring the stationarity in time

series forecasting[J]. Advances in neural information processing systems, 2022,35:9881-9893.

[4] Liu Y, Hu T, Zhang H, et al. itransformer: Inverted transformers are effective for time series forecasting[J]. arXiv preprint arXiv:2310.06625, 2023.

[5] Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(12): 11106–11115.

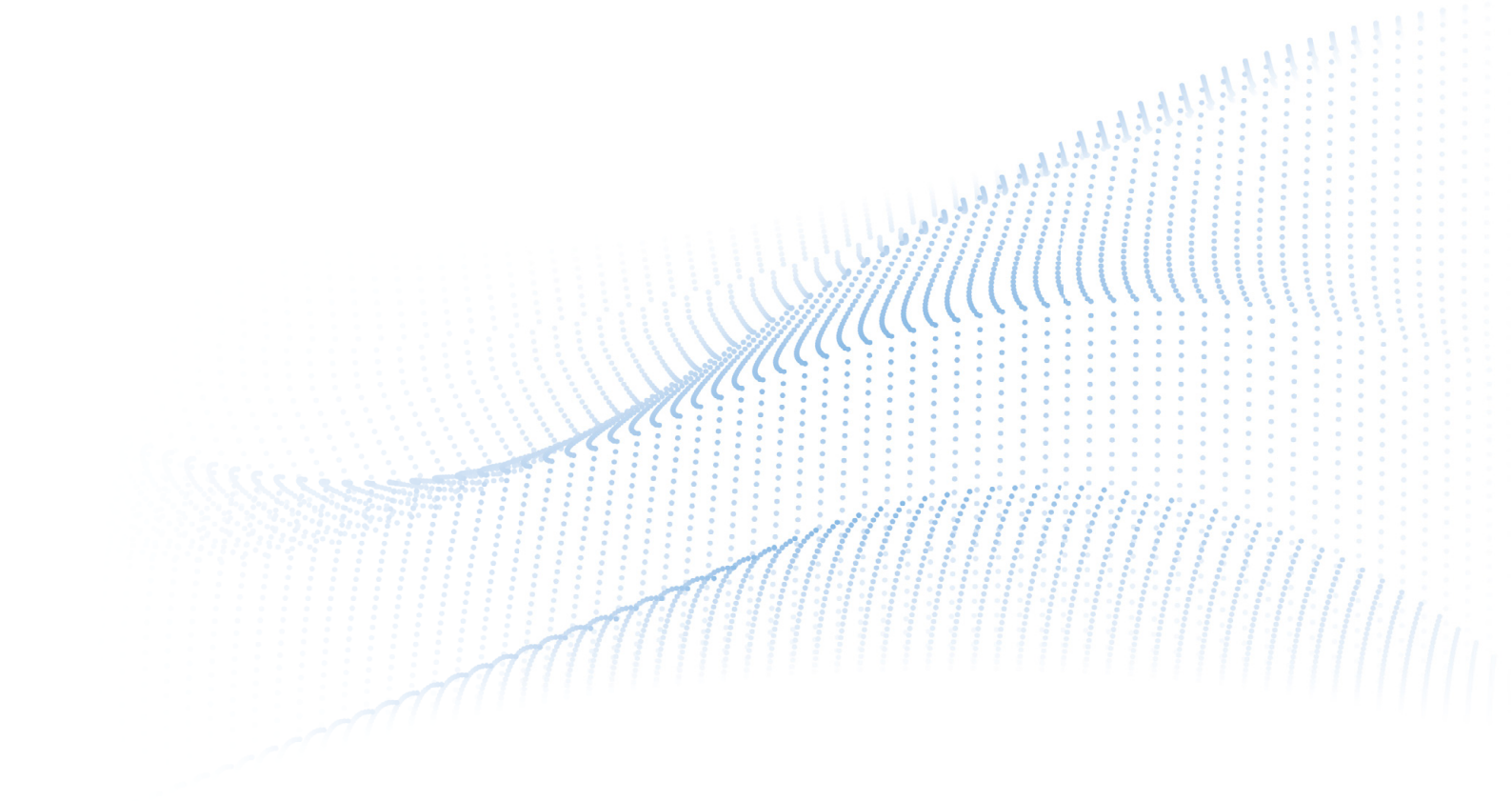
[6] Wen Q, Ma Z, Zhang J, et al. DLinear: A simple disentangled linear forecasting model[EB/OL]. arXiv:2211.06624, 2022[2025-06-08]. https://arxiv.org/abs/2211.06624.

[7] Campos D, Zhang M, Yang B, et al. LightTS: Lightweight time series classification with adaptive ensemble distillation[J]. Proceedings of the ACM on Management of Data, 2023,1(2):1-27.

[8] Jin M, Wang S, Ma L, et al. Time-llm: Time series forecasting by reprogramming large language models[J]. arXiv preprint arXiv:2310.01728, 2023.

[9] Wang Q, Hao Y. ALSTM: An attention-based long short-term memory framework for knowledge base reasoning[J]. Neurocomputing, 2020,399:342-351.

[10] Tao F, Liu G. Advanced LSTM: A study about better time dependency modeling in emotion recognition[C], 2018. IEEE, 2018.



基于大模型技术的监管问询函生成 *

吴苑斌 [1]，谢欣余 [1]，刘燕婷 [1]，杜威 [1]，王晓玲 [1]，潘明慧 [2]，王玲 [2]
[1] 华东师范大学 | [2] 华泰证券股份有限公司

摘　要： 随着金融监管场景的数字化升级，人工撰写监管问询函面临效率低下、专业性依赖强等问题，难以适应高频复杂的监管需求。本文提出基于大模型技术的问询函生成方案，首先构建了包含问询函原文、关联新闻及财报数据的多源数据集。然后采用参数高效微调技术和动态词表技术构建生成模型。最后给出了实验结果、生成示例和未来优化方向。

关键字： 大语言模型；问询函生成；参数微调；动态词表

一、引言

在商业活动与监管实践中，问询函作为信息获取、风险核查与问题追踪的重要工具，广泛应用于证券监管、企业合规审查、商务调查等场景。从证券交易所针对上市公司财务异常、重大交易的问询，到企业内部对业务部门的合规性质询，问询函的准确性与专业性直接影响调查效率和决策质量。然而，传统问询函生成主要依赖人工，这一过程不仅耗费大量人力与时间，而且容易受到审核人员主观因素的影响，导致问询函的质量和一致性难以保证。同时，随着金融创新的不断推进，新的业务模式、风险因素不断涌现，人工生成问询函在应对复杂多变的场景时，很难保证全面和精准。随着数字化转型加速，各行业产生的数据量呈爆炸式增长，问询函的需求也愈发高频和复杂，人工撰写模式已难以满足实时、高效的业务需求。

为了提升问询函生成的效率与质量，一些基于简单规则或传统机器学习的方法曾被尝试应用 [1]，但这些方法存在明显的局限性。它们难以有效捕捉大量信息中复杂的语义关系和深层的风险信息，在处理新型业务和复杂事务状况时，生成的问询函往往无法切中要害，无法满足监管的严格要求。

近年来，大语言模型在自然语言处理领域取得了突破性进展。以 ChatGPT 为代表的大模型，凭借其强大的文本理解、生成和推理能力，在众多领域展现出巨大的应用潜力。大模型通过对海量文本数据的预训练，能够学习到丰富的语言知识和语义表征，这使得它们在处理复杂文本任务时具有独特的优势。其知识覆盖广，能跨领域整合信息，生成逻辑严谨、风格多样的内容，同时凭借强大的推理能力，能处理文本蕴含的复杂逻辑关系，还可快速响应需求、批量定制文本，为复杂文本生成提供了高效解决方案。

然而大模型虽然有以上优势，但由于问询函的专业性较强，在面对问询函生成这一应用场景时通用大模型仍然有缺陷，其输出结果往往存在针对性不足、合规性偏差等问题。因此，通过对大模型进行微调、添加动态词表以适配特定任务，成为提升问询函生成质量的关键。基于大模型技术的问询函生成研究，能够利用模型强大的语义理解与文本生成能力，实现问询函从数据提取、逻辑构建到文本生成的全流程自动化，显著提升工作效率；还能通过专业数据微调以及加入动态词表，使模型学习问询函的行业规范、语言风格及逻辑框架，确保生成内容的合规性与专业性。这一研究对于降低人工成本、优化监管流程、提升风险防控能力具有重要的理论价值与现实意义，有望推动问询函生成从传统人工模式向智能化、精准化模式的变革。

* 本文是项目下设课题“开放域舆情风险识别、预警与处置”（课题编号：2021YFC3340702）的研究成果，课题负责：王晓玲（华东师范大学）；本文部分相关成果已 EMNLP2024 会议发表。

二、问询函

2.1 问询函的含义

问询函是指具有监督、管理或审查职能的机构、组织或个人，在对特定事项进行调查、核实或审查过程中，以书面形式向相关主体发出的信息索取与质疑文件。从功能属性来看，问询函兼具信息获取与风险预警双重作用：一方面，通过结构化或开放式的问题设计，要求被问询方对财务数据异常、业务经营合规性、重大事项披露完整性等内容作出详细解释与补充说明；另一方面，借助质询过程揭示潜在风险点，督促相关主体规范运营、提升信息透明度。

在应用场景上，问询函广泛存在于资本市场监管、企业内部审计、政府行政监督等领域。例如，证券交易所向上市公司发出的问询函，聚焦定期报告、重大资产重组等公告中的疑点，保障投资者知情权；企业内部审计部门通过问询函核查业务流程合规性，防范经营风险；政府部门则利用问询函收集政策执行反馈，优化公共管理决策。问询函的文本特征表现为严谨的逻辑架构、规范的专业术语和明确的问题导向，其内容需围绕特定事实依据展开，措辞需兼顾权威性与客观性，以确保质询的有效性与法律效力。问询函具体样例如图 1 所示。

关于XXXX公司相关媒体报道的问询函

XXXX公司：

我部关注到，有媒体报道称，X月X日X时X分许，公司控股子公司气化车间发生高压气体喷出事故，导致现场多名在高处作业的工作人员被喷射坠落。

根据《上海证券交易所股票上市规则》第13.1.1条等有关规定，现请你公司核实相关事故的具体情况，对公司生产经营的影响，并及时履行信息披露义务，妥善做好事故相关处置工作

请公司收到本问询函后立即披露，并在次一交易日内书面回复我部，并履行相关信息披露义务。

上海证券交易所上市公司管理二部

X年X月X日

图 1 问询函样例

2.2 困难与挑战

问询函作为监管机构、企业内部审查等场景下获取信息的重要工具，其生成质量直接影响后续决策与监管效能。然而，当前问询函的自动生成面临诸多挑战。问询函需精准匹配复杂多样的业务场景与监管要求，传统基于规则和模式匹配的生成方法，难以应对不断变化的问询需求与新型业务模式，这要求生成系统必须持续迭代升级，以适应动态的问询场景；问询函内容需准确传达意图，同时保持专业、严谨的语言风格，但其中涉及的专业术语、逻辑推理及对复杂事实的表述，使语义理解和内容生成极具难度，需要模型具备深度语义理解、逻辑构建和专业语言表达能力；在实际应用中，可用于训练问询函生成模型的高质量标注数据极为有限，重复、简单的问询函样本易导致模型过度拟合，难以满足复杂场景下的问询需求。

上述难题给问询函自动化生成带来阻碍，而大模型的出现为解决该问题提供了新方向。大模型凭借庞大的知识储备和强大的信息处理能力，能更高效地理解业务场景和监管要求，辅助生成问询函。

虽然通用大模型凭借庞大的知识储备和强大的信息处理能力，为问询函生成提供了新的解决思路，但在实际应用中仍存在明显欠缺。通用大模型的训练数据涵盖广泛领域，缺乏对问询函生成场景的针对性优化，导致其生成的问询函在专业术语准确性、监管要求契合度等方面难以满足实际需求。此外，通用大模型生成内容的逻辑性和严谨性不足，容易出现问询要点遗漏、问题表述模糊等情况。因此，要实现高质量的问询函自动生成，本文对通用大模型进行针对性优化，训练提升模型在问询函生成任务中的专业性、准确性和逻辑性，从而推动问询函生成技术的发展。

三、数据准备

为了满足微调和构建动态词表的需求，需要准备问询函数据和相关的新闻、公告、财报、法律条文数据，数据准备的整体流程如图 2 所示。

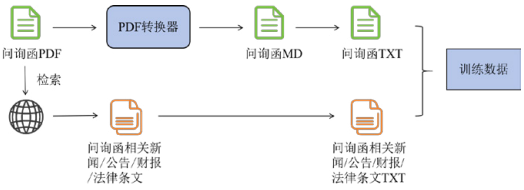


图 2 数据准备

3.1 数据收集

本研究的数据采集围绕问询函及相关支撑性文本展开，主要数据包括两大类：

1）问询函 PDF 文件；

2）问询函对应的新闻报道、公司财报、重大事项公告、法律条文等文本数据。

本研究采用人工搜集模式确保数据质量。数据来源主要包括权威监管机构、新闻资讯网站和企业信息披露平台三大类，具体采集流程如下：

首先访问证券交易所披露网站。在检索过程中，通过设定 “问询函” “媒体报道” “年报问询” 等关键词，结合企业名称、发布时间等筛选条件，人工定位并下载上市公司问询函 PDF 文件。

其次在新闻资讯类平台，通过组合 “[问询函核心事件]+ 时间” 等关键词进行人工筛查。在检索结果页面，逐篇浏览新闻报道、专家解读文章，重点筛选与问询函事件存在强关联的内容，包括事件背景、舆论观点、市场反应等。最后保证筛选到的新闻内容尽量专业且避免提及问询函。

接着检索第三方财经网站等，如东方财富网，针对问询函相关的企业，根据问询函时间及内容，获取对应时间财报以及与问询函内容相关的公告。

最后收集问询函相关的法律条文，并将所有数据按问询函分组进行整理。

数据采集完成后，对数据质量进行检验。主要关注以下内容：一是剔除二次问询，二次问询函本质上是对已有问询结果的补充和延伸，其生成逻辑与首次问询存在差异，

为了避免其干扰模型训练核心目标，需要将二次问询移出数据集；二是内容比对，确保每个问询函有相关新闻媒体报道，同时核查问询函与相关支撑性文本的相关性；三是格式审查，确保 PDF 文件正确完整、新闻报道网址可正常打开；四是逻辑校验，检查财报数据勾稽关系、新闻报道时间线是否合理。经审核剔除重复文件、无效链接及内容缺失文件后，形成高质量的原始数据集，为后续研究奠定基础。

3.2 数据格式转换

为适配语言模型的训练与微调需求，需将原始文件转换为文本格式。使用 PDF 转换工具，将问询函 PDF 转换为 markdown 格式，再由 markdown 格式转换为 txt 格式。对于新闻报道、财报公告和法律条文数据，直接保存为 txt 格式。

3.3 构建训练数据

为了适配后续微调的需求，需要对收集的数据进一步处理，构建训练数据。在之前的处理中，已剔除缺乏对应首次问询的二次问询函数据，确保训练数据集聚焦于首次问询场景下的完整逻辑链条。对留存的问询函、相关新闻、公告及财报等文本，将其转化为“系统提示 - 用户输入 - 助手回答” 的三元组格式：

1）系统提示部分明确任务目标，在本研究中使用“你现在是一名交易所监管人员，请你根据下面的新闻内容生成相应的问询函。”作为指令；

2）用户输入部分整合新闻报道、财报关键数据等原始信息；

3）助手回答部分为人工撰写的标准问询函文本。通过严格的人工审核机制，确保数据标注一致性，避免因数据偏差影响模型训练效果。

四、问询函生成

4.1 基于大模型微调的问询函生成

完成数据准备之后，下一步需要对模型整体框架进行搭建。本研究以大模型为基座，设计和训练了问询函生成模型，模型训练的整体流程如图 3 所示。尽管大模型具备强大的自然语言处理能力，但问询函生成属于专业性较强的应用场景，其原生训练数据难以满足对问询函逻辑构建、专业术语运用、合规性表述等特定需求。因此，本研究采用参数高效微调方法对模型进行微调，引导模型学习问询函生成的逻辑框架与语言规范，使其能够准确、高效地完成问询函生成任务。本节将详细阐述以 ChatGLM3 为基座，运用 P-tuning v2 方法进行模型微调具体过程。

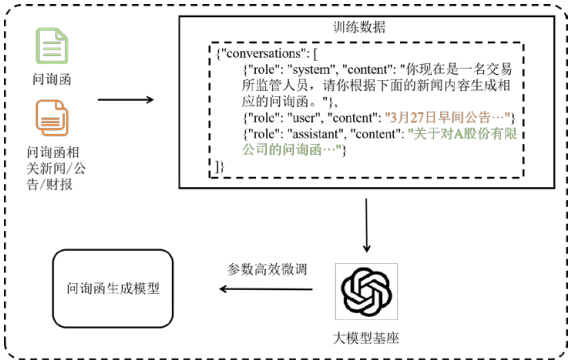


图 3 基于大模型微调的问询函生成

4.1.1 基座模型

ChatGLM3 是由智谱 AI 和清华大学 KEG 实验室联合发布的新一代开源双语对话语言模型 [2]。它基于 Transformer 结构，采用了自注意力机制，能够更好地理解上下文信息，从而提高对话的准确性和流畅性。此外，ChatGLM3 还支持多种功能，包括工具调用、代码执行、知识图谱搜索与推理等复杂场景。

本研究选用 ChatGLM3 作为核心基座模型，主要基于以下考量：其一，在模型架构层面，ChatGLM3 的多层 Transformer 结构使其在处理长文本时表现优异。问询函生成任务通常需要整合新闻报道、财报数据等多源长文本信息，该模型能够有效捕捉不同信息片段之间的语义关联，构建连贯的逻辑框架。其二，从语言理解能力来看，ChatGLM3 在大规模中文语料上进行预训练，对中文语言模式、专业术语有深入理解。问询函作为中文专业文档，涉及大量金融、法律领域的专业术语和规范表述，ChatGLM3 的预训练基础能够为准确生成此类文本提供有力支持。其三，在知识推理方面，ChatGLM3 具备较强的逻辑推理能力，能够基于给定信息挖掘潜在问题和风险点。这与问询函生成任务的核心需求高度契合，即需要从复杂的业务信息中识别异常点并提出针对性问题。最后，ChatGLM3 提供了开源版本和商业授权选项，具有良好的可扩展性和社区支持，便于进行定制化开发和技术迭代。

4.1.2 模型微调

确定基座模型后，为了提升大模型在问询函生成任务上的性能，需要对其进行微调操作，使其获取问询函生成的能力。由于使用的基座模型 ChatGLM3 参数规模达十亿级，难以对其进行全量参数微调，因此本文采用参数高效微调的方法对其进行微调，具体来说选用 P-tuning v2 微调技术 [3]。

相较于早期连续 prompt 技术，P-tuning v2 针对中小规模模型优化了参数更新机制，在参数量小于 10B 的模型微调中，能够以更少的可训练参数实现更高的任务适配精度。本研究采用的 ChatGLM3-6B 模型，其参数规模恰好处于 P-tuning v2 的最佳适用区间，因而选择该技术开展问询函生成任务的优化。

P-tuning v2 的核心机制在于通过构建多层级连续提示向量实现对模型输出的精准引导。该方法将任务指令编码为特殊的嵌入向量，在训练过程中冻结模型原生参数，仅针对这些提示向量进行迭代更新。具体而言，P-tuning v2 在 Transformer 架构的不同层次引入独立的前缀 token，使模型在不同语义处理阶段能够动态聚焦关键信息。这种分层优化策略不仅大幅降低了计算复杂度——可训练参数仅占原模型的 0.1%-3%，还能有效避免因大规模参数调整引发的过拟合风险，在保持模型泛化能力的同时显著提升特定任务表现。

在问询函生成任务中，用 P-tuning v2 对 ChatGLM3-6B 进行微调。设置前缀序列长度为 128，最大序列长度为 16000，使其适用于处理新闻报道等长文本。利用 P-tuning v2 将训练数据中的任务指令部分从离散的形式转化成连续向量的形式，用来提示基座模型根据新闻内容生成问询函。P-tuning v2 既能够使基座模型学习到问询函生成的能力，又能减少微调参数量。微调后，我们得到基于 ChatGLM3-6B 的问询函生成模型。

4.2 基于动态词表的问询函生成

除了使用大模型微调进行问询函生成外，本研究还使用动态词表来进行问询函生成。问询函生成可视为是条件文本生成任务之一。条件文本生成可以分为两种：抽取式和生成式。抽取式方法通过计算句子或词的重要性分数，将抽取的重要信息作为摘要。生成式方法则主要利用序列到序列结构来生成文本。抽取式方法生成结果会和原文相关性高，但摘要内容局限，缺乏灵活性和创新性，而生成式的方法灵活，但生成的内容可能会存在语法错误、幻觉等问题。因此该任务考虑将二者结合，从相关文本中划分重要短语，加入到预训练语言模型的词表中，构成动态词表，在维护创新性的同时，保持模型的生成内容与提供的原文尽可能相关。

具体来说，问询函生成的动态词表由三部分组成：舆情事件词表，问询函模板词表以及语言模型的原始词表。舆情事件词表主要是通过将与舆情相关的新闻、财报和公告等文本等进行划分，从中获取与舆情的相关短语，加入模型词表。由于问询函具有相对统一的模板，包含通用话术，因此考虑统计问询函相对固定、常用的短语作为问询函模板词表，帮助模型获取问询函的固定模板信息。同时，为了能使模型生成连贯、通顺、可读的文本，模型仍然保留语言模型的原始词表。

模型的具体运行流程包括以下几个步骤：

1）从构建的问询函、舆情数据集中检索与当前样本 S_i 最为接近的 K 条数据 $\{d_1,d_2...d_k\}$，按照前向最大匹配算法，对当前样本进行短语划分 $\{p_1,p_2...p_n\}$。

2）将 GPT-2 作为短语编码器 Phrase-Encoder，预训练 GPT-2 作为解码器进行文本生成。

3）将检索的 K 条新闻、财报、公告等数据 $\{d_1,d_2...d_k\}$ 对应的短语 $\{p_1,p_2...p_n\}$ 分别送入短语编码器，它首先利用原始静态词表的分词器对短语 p 进行分词，得到 $p=\omega_1,\omega_2,\text{-},\omega_s$，然后通过因果编码器以及多层感知机（MLP），得到短语词元的隐层表示，将短语 p 最后一个词元对应的隐层向量作为该短语 p 的向量表示。

4）将当前样本 S_i 作为前缀送入 GPT-2 让模型继续生成，模型在生成时，会根据最后位置的隐层表示与 GPT-2 的原始词表（token_embeddings）以及短语（phrase_embeddings）的相似度选择最为接近的某个词或者短语作为生成的下一个词，不断重复该步骤，直至生成结束或达到模型可接受的最大长度。

$$last\_hidden\_state=GPT2(S_i) \quad (1)$$

$$next\_token=argmax(last\_hidden\_state*[toke\_embeddings,phrase\_embeddings]) \quad (2)$$

图 4 展示了基于动态词表的问询函生成流程示例，在问询函生成过程中，模型可以从三部分词表中选择短语或者词进行生成，当模型需要舆情相关信息时，可以从舆情事件划分的短语中选择合适的短语作为下一个生成的结果，当模型需要问询函模板信息时，则从绿色的问询函模板词表中选择合适的词。基于动态词表的问询函生成不仅解决了以往将舆情信息直接作为条件文本信息送入模型，受到语言模型最大输入长度限制的问题，同时也考虑了舆情和模板的相关信息，使得文本生成更有依据。

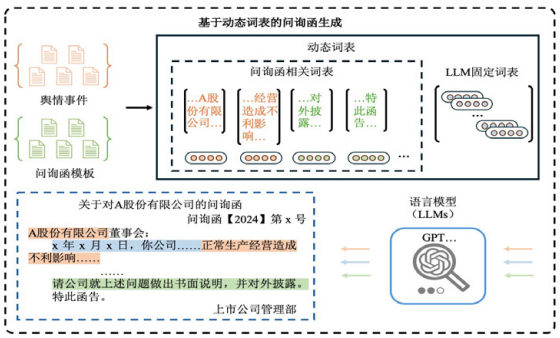


图 4 基于动态词表的问询函生成

五、实验

5.1 数据集

对于测试数据集，采用与第三节相似的收集方法，主要收集新闻报道，以及同期的公告、财报和相关的法律条款等内容。构建“指令 - 内容 - 公司”三元组格式的测试数据：

1）指令部分用于引导模型生成问询函，在本研究中使用“你现在是一名交易所监管人员，请你根据下面的新闻内容生成相应的问询函。”作为指令。

2）内容部分即为新闻报道、公告、财报等用于生成问询函的内容。

3）公司部分即生成的问询函应该问询的对象。

同时，在验证动态词表方面，实验采用 Wikitext-103 数据集，Wikitext-103 是一个大型文本数据集，全部从维基百科的 Good 与 Featured 文章中提炼出来，主要用于自然语言处理（NLP）任务，特别是在预训练模型和生成模型的研究中。

5.2 结果

在测试时，将任务指令和相关文本输入到问询函生成模型中，模型可以自动生成符合要求的问询函文本，图 5 和图 6 展示了模型生成问询函的两个具体示例，展示了其在实际应用中的有效性与准确性。

分析生成结果可以发现，生成的问询函能够遵循监管规则及公文格式要求，涵盖问题定性、事实引用、问询要点及回复期限等核心要素。同时能够做到逻辑连贯性，从舆情事实陈述到问询要点展开，逻辑链条清晰，问题层层递进，便于被问询方针对性回应。并且采用监管问询函的

input

"instruction": "你现在是一名交易所监管人员，请你根据下面的新闻内容生成相应的问询函。",
"content":"认为手机游戏《烈火遮天》抄袭《热血传奇》，盛大公司以不正当竞争为由将奇天乐地公司、苹果公司、卓易讯畅公司和华为公司诉至法院，要求停止侵权、赔礼道歉并赔偿经济损失1000万元。
盛大公司诉称，盛大公司是推动中国互动娱乐产业发展的领军企业，其商标“盛大”和“SHANDA”被认定为驰名商标。《热血传奇》业内简称“传奇”，是一款由韩国公司制作的顶级网络游戏，盛大公司获得该游戏在中国大陆地区的独家代理权，并于2001年9月开始正式运行。
2014年9月，盛大公司发现奇天乐地公司开发了一款与《热血传奇》极其近似的手机游戏《烈火遮天》，经进一步取证、鉴定，发现《烈火遮天》完全抄袭了《热血传奇》，在作品的角色形象名称、地图场景、技能、动画特效等多个方面与《热血传奇》构成实质性相似，使得众多玩家对《烈火遮天》与《热血传奇》及改编自《热血传奇》的系列手机游戏发生混淆。
另外，奇天乐地公司在推广《烈火遮天》的过程中，假冒盛大公司的知名商品、服务名称，盗用盛大公司享有著作权的知名网络游戏形象图片，使得众多网友对《烈火遮天》及其所在网站与《热血传奇》等系列游戏发生混淆。
近来，盛大公司发现网络苹果应用商店为社会公众提供《烈火遮天》的下载服务，公众可以直接通过iphone、ipad、itouch等产品进入网络苹果应用商店，将《烈火遮天》下载到苹果设备中运行并获取经济利益。卓易讯畅公司和华为公司均在其网站上向公众提供《烈火遮天》的下载服务，侵害了盛大公司的著作权。
盛大公司认为，各被告借助《热血传奇》的知名度，擅自侵犯自己的著作权并进行不正当竞争，已经对社会公众产生严重误导，抢占了盛大公司的市场份额并赚取了巨额利润，给盛大公司造成了巨大损失。"

output

上 海 证 券 交 易 所
上证公函【2015】0806号
关于对盛大网络股份有限公司相关媒体报道事项的问询函
盛大网络股份有限公司：
2015年 8月 21日，公司于 2014年 9月发现奇天乐地公司开发了一款与《热血传奇》极其近似的手机游戏《烈火遮天》，并在 2015年 7月 2日向相关公司发出《停止侵权、赔礼道歉并赔偿经济损失 1000万元》的诉讼函。近日，有媒体报道称，公司因《烈火遮天》涉嫌抄袭《热血传奇》，被相关公司诉至法院。根据本所《股票上市规则》第 17.1条规定，请公司就以下事项进行进一步核实和补充说明。
一、根据前期诉讼函，公司称《烈火遮天》完全抄袭了《热血传奇》，在多个方面与《热血传奇》构成实质性相似，请公司补充说明《烈火遮天》与《热血传奇》各自主营业务、产品设计、程序编写、运营管理等方面的差异，以及《烈火遮天》是否构成对《热血传奇》的侵权。
二、公司称，奇天乐地公司在推广《烈火遮天》的过程中，假冒盛大公司的知名商品、服务名称，盗用盛大公司享有著作权的知名网络游戏形象图片，请公司补充说明：（1）奇天乐地公司是否存在上述侵权行为；（2）公司前期诉讼函是否具有法律约束力，是否能够实现公司获得赔偿的目的；（3）结合目前案件进展，分析公司后续诉讼请求的可行性，并提示相关风险。
三、请公司就《停止侵权、赔礼道歉并赔偿经济损失 1000万元》的诉讼请求进行调整，降低赔偿金额，并说明调整原因。
四、请公司就本次诉讼事项进行风险提示，说明可能存在的诉讼\延滞、判决结果不公等风险，并充分提示相关风险。
请你公司于 2015年 8月 26日之前，以书面形式回复我部。
证券交易部
二〇一五年八月

图 5 模型生成具体示例 1

input

"instruction": "你现在是一名交易所监管人员，请你根据下面的新闻内容生成相应的问询函。",
"content":"建设粤港澳大湾区，如何用好港口资源？13日，广州市人大代表、广州港集团董事长蔡锦龙在接受记者采访时表示，粤港澳大湾区拥有世界级港口群，其中包括了广州港等3个全球排名前列的港口。广州正制定新一轮国际航运中心三年行动计划，到2020年初步目标拟为港口吞吐量6.5亿吨，集装箱吞吐量2500万标箱，国际航线达到120条。与此同时，广州港拟通过股权转让等形式，整合省内其他七市港口资源。
…在蔡锦龙看来，广州在粤港澳大湾区建设中，发挥核心增长极作用。他提及，具体到港口，下一步加强港口资源整合是大势所趋。目前江苏、浙江等地已纷纷实现了港口资源整合。我省推进港口资源也将有助于避免大湾区内部港口重复建设、过度竞争及各自为政，从而推动港口群的互联互通、共建共享共发展。
…事实上，广州港集团也积极走向全球，在美国、欧洲设立了办事处，邀请当地航运知名人士担任办事处代表，同时通过国际会议、论坛等多种形式，广州港面向船公司、物流公司、制造企业等营销推介进一步加强，进口和出口双双实现提升。目前国际航线已经达到91条，其中欧洲航线13条，美洲航线9条。港口国际货物占比也从过去占比不足一成提升到四成。
与此同时，国家正在探索建设自由贸易港，这也将给广州国际航运中心建设带来新的机遇。蔡锦龙表示，自由贸易港关键就是要实现人员、商品、资金等要素自由进出，真正实现“境内关外”。对于港口来讲，国际中转货物就可以不用办通关手续，有助于节约成本、节约时间，能吸引更多船公司来开展业务，企业也可以在港口自由拆装货物、加工产品后再出口，加上公司注册便利化，将有助于集聚金融、物流、法律等全球高端要素，推动临港工业及现代服务业发展。
"

output

上 海 证 券 交 易 所
上证公函【2018】2536 号
关于对广州港集团股份有限公司近期媒体报道相关事项的监管问询函
广州港集团股份有限公司：
近期，公司对外表示，将整合省内其他七市港口资源，通过股权转让等形式实现。同时，公司表示，已制定新一轮国际航运中心三年行动计划，拟实现港口吞吐量 6.5 亿吨、集装箱吞吐量 2500 万标箱、国际航线达到 120 条。请公司核实并补充披露以下事项：
一、媒体报道称，下一步将加强港口资源整合，通过股权转让等市场手段实现。请公司补充说明，目前相关整合的具体情况及现阶段进展，尚需履行的审批程序，并进行风险提示。
二、媒体报道称，广州港集团已积极走向全球，通过国际会议、论坛等多种形式，加强对外营销推介。请公司补充说明，目前国际航线、港口国际货物占有率等国际航运业务的具体情况。
三、媒体报道称，国家正在探索建设自由贸易港，这也将给广州国际航运中心建设带来新的机遇。请公司补充说明，目前广州港集团在自由贸易港方面的具体举措及现阶段进展，尚需履行的审批程序，并进行风险提示。
四、请公司结合上述问题，对公司的国际航运中心建设进行自查，并补充披露目前进展情况、尚需履行的审批程序。
五、请公司全体董事、监事、高级管理人员勤勉尽责，保护中小投资者利益，认真核查说明上述问题，并就公司国际航运中心建设的发展方向、重大事项进行审批。
请你公司于 2018 年 12 月 18 日之前，以书面形式回复我部。
上海证券交易部
二〇一八年十二月

图 6 模型生成具体示例 2

表 1 动态词表实验结果 [4]

模型	MAUVE	Rep-2	Rep-3	Rep-4	Diversity	Latency(s)	PPL
Transformer	20.47	41.96	36.82	33.74	24.30	1.10	**3.60**
RETRO	19.59	43.78	38.58	35.35	22.33	4.43	3.96
KMM-LM	19.92	43.79	38.76	35.69	22.13	10.36	3.48
CoG	21.61	34.77	30.67	28.35	32.41	1.04	7.89
MWT	24.74	33.78	26.72	22.76	37.48	1.13	5.58
Ours	**25.69**	**27.77**	**20.80**	**17.08**	**47.44**	**0.99**	8.03

正式书面语体，表述严谨、准确，符合公文语言规范。

为了验证动态词表模型的有效性，实验在 Wikitext-103 数据集上进行开放文本生成。在该实验中，模型输入是长度为 32 个词元的句子前缀，由模型生成后续的 128 个词元。同时，为了针对不同的前缀构造不同的动态词表，该方法利用测试样本输入前缀从语料库中检索相关文档，并从检索的文档中划分短语作为新增词元加入语言模型词表当中。

本实验中从文本流畅度 MAUVE，文本重复度 Rep-n，文本多样性 Diversity，困惑度 PPL 以及时间开销 Latency 等多个方面将动态词表模型与以往的文本生成模型进行对比，Baselines 包含 Transformer，KNN-LM，Retro，COG，MWT 等。具体实验结果如表 1 所示。

从实验结果可以看出：从生成质量上看，增加动态词表的语言模型在流畅度上（MAUVE）上比标准语言模型要高5.22%。从生成效率上看，动态词表拥有最优的生成速度。这是因为动态增加的每一个短语词元包含多个固定词元，因此在解码相同长度的文本时需要更少的解码步。

六、总结与展望

本研究聚焦问询函生成的智能化转型，针对传统人工生成模式效率低、一致性差及通用大模型专业性不足等核心问题，提出基于大模型微调和基于动态词表的问询函生成方法。实验表明，优化后的模型可以生成逻辑连贯、格式规范、符合监管要求的问询函，提升了生成效率与内容专业性，为问询函生成的智能化转型提供了有效解决方案。

尽管本研究在问询函生成领域取得了初步成果，但仍存在进一步优化的空间。未来的研究可以从以下几个方面展开：

1）目前的问询函生成主要通过调用模型直接生成，虽然能够快速输出结果，但在复杂场景下可能无法充分优化生成内容。未来可以考虑采用多步生成策略，进一步提高生成质量。

2）问询函生成需要充分考虑监管人员和被问询方的需求与反馈，未来可以设计用户交互界面，允许用户在生成过程中实时调整生成参数或提供反馈信息，模型根据用户反馈动态调整生成内容，从而实现更加个性化和精准化的问询函生成服务。

3）可以进一步拓展到多模态数据融合的场景。例如，结合财务报表的可视化图表、新闻报道的图片或视频等多模态信息，帮助模型更全面地理解业务背景和风险点，从而生成更加丰富和准确的问询函内容。

参考文献：

[1] 深圳证券信息有限公司 . 一种问询函生成方法、系统及其装置 :CN202111277389.0[P]. 2024-11-05.

[2] GLM T, Zeng A, Xu B, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools[J]. arXiv preprint arXiv:2406.12793, 2024.

[3] Liu X, Ji K, Fu Y, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint arXiv:2110.07602, 2021.

[4] LIU Y, JI T, SUN C, et al. Generation with dynamic vocabulary[C/OL] //AL-ONAIZAN Y, BANSAL M, CHEN Y N. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics, 2024: 18931-18948.

买方视角下的企业科创能力量化评价体系 *

马振民，庄明光，李媛
汇添富基金管理股份有限公司

摘　要： 本文介绍了买方视角下的企业科创能力评价体系构建项目，旨在为资本市场的投资者提供科学的决策依据。项目聚焦于六大科技创新领域，包括新一代信息技术、高端装备、新材料、新能源、节能环保及生物医药，选取了各领域内具有典型性的企业作为分析样本。评价体系涵盖行业状况、业务状况、产品与技术、管理能力、财务控制及外部支持六大关键维度，采用随机森林算法作为核心评价模型，确保了模型的泛化能力和稳定性。项目成果被整合至相关平台后，开发了涵盖行业信息、公司深度资料、估值分析及创新能力量化对比的综合功能模块。通过多维度数据整合与可视化展示，该平台可为投资者提供更全面的企业分析视角，帮助其精准判断企业的成长潜力与技术竞争力。

关键字： 科创能力、买方视角、量化评价体系

一、项目背景

本研究从买方机构视角出发，将投资者对上市公司成长性和价值的需求作为核心考量。我们深知投资者期望在资本市场中获得稳健且可观的回报，因此需要一套科学合理的评价标准来筛选出真正具有科创潜力和成长价值的企业。通过制定这一体系，我们旨在为投资者提供更具参考价值的决策依据，助力其在复杂多变的市场环境中把握投资机会，实现资产的稳健增值。同时，这也有助于促进社会主义市场经济的良性发展，推动科技与资本、产业之间的高水平循环，为完善资本市场基础制度贡献一份力量。

二、项目目标

本项目的目标是基于买方视角，针对六大科创领域（新一代信息技术、生物医药、新能源、节能环保、高端设备、新材料）的科创属性进行评价，设计一套综合考量企业的核心技术、科技创新能力和市场认可度等关键要素的评价指标体系。通过引入科创属性评价指标，可以更好地衡量企业在科技创新方面的实力，提高对科创板企业的筛选能力，这将有助于优化市场资源配置，推动科技创新的发展，并为投资者提供更明晰的投资标的选择，促进资本市场的稳定和健康发展。

* 本文是项目下设课题“科创企业评价与行业综合应用示范”（课题编号：2021YFC3340704）的研究成果，课题负责：唐忆（上海证券交易所）。

三、研究范围与方法

3.1 研究范围

项目的研究范围聚焦于六大科技创新领域，包括新一代信息技术、高端装备、新材料、新能源、节能环保以及生物医药。这些领域不仅是国家战略支持的关键产业，也是全球资本市场的关注热点。为确保研究的代表性，项目通过严格的筛选标准选取了各领域内具有典型性的企业，作为分析样本和评价指标的主要来源。数据主要来自公开数据库、政府官方网站及企业发布的财务文件和业务说明书，确保了数据的可靠性和准确性。

3.2 研究方法

3.2.1 筛选机制

为了能够在众多企业特征中快速筛选出最有效的科创属性，我们通过自下而上的调研方法，深入到二级子行业中，精准识别那些在技术创新、市场份额和品牌知名度方面表现卓越的行业龙头企业。这些企业不仅在自身领域中占据主导地位，还能反映行业的整体发展趋势与未来走向。因此，它们在科创能力上的表现为我们提供了宝贵的研究依据和参考。筛选机制主要包括以下几个步骤：

1）行业龙头地位：行业龙头地位是首要筛选标，这些企业通常是行业的技术引领者或市场领导者，能够在更大程度上反映出行业的创新能力和未来发展潜力。

2）财务数据完整性：公司必须拥有三年以上的完整财务数据，以确保对企业科创能力的评估具有充分的数据

支撑。财务数据是分析企业盈利能力、财务健康状况以及成长趋势的关键指标，也是客观评估其科创能力的重要依据。

3）科创板上市企业优先：科创板是中国专门为科技创新型企业设立的交易板块，能够在该板块上市的企业，通常已经经过了严格的审查和筛选，因此它们在创新能力上具有较高的市场认可度。优先选择这些企业，能够确保评价模型在应用时具有较强的参考性和可信度。

通过上述筛选条件，成功筛选出了一批具有代表性的科创企业，这些企业在技术创新、财务表现、市场竞争力等方面均表现优异，为行业发展提供了典型案例，也为构建细分的科创能力评价指标提供了重要依据。

3.2.2 指标构建

在前期对企业进行筛选后，我们提取了不同领域代表性企业在多个维度（行业、经营、财务、产品与技术、投融资、外部支持）上的特征和属性，以体现其科创能力。通过这一过程，共总结得到了 223 个指标，这些指标涵盖了企业在各个方面的表现，以体现其科创能力。

为了进一步优化评价体系，我们将这一系列指标按照数据的可得性进行排序，从高可得性指标排到低可得性指标，根据量化模型的需求，选择了 24 个具备可量化和操作性的核心评定指标放入评价标准中（见表 1）。这套模型不仅考虑了指标的重要性，还赋予了每个指标相应的权重和得分，以确保综合评价的准确性和客观性。对于未上市或新加入科创板的公司，可以将其纳入该量化模型进行科创能力的评价和打分。通过量化评价模型，可以对这些公司的科创能力进行量化的评估，得出一个综合的评分结果。这一评分结果可以为投资者、风险投资机构和其他利益相关者提供重要参考，帮助他们更好地了解和判断企业的科技创新实力。

表 1 科创能力量化评价体系指标表

一级指标	二级指标
业务状况-企业资质	实控人属性
	科创属性
	公司成立时长
	战略性新兴产业分类
	是否专精特新企业
业务状况-营销能力	销售人员人数占比
产品与技术-知识产权	授权发明专利
	专利总数
	发明专利/专利总数
产品与技术-研发实力	研发支出/同比增长
	研发支出占营业收入比例
	研发人员占比
	研发投入
	研发人员数占比
	博士人数占比
	高管学历
财务特点-盈利能力	营业收入
	营业收入同比增长
	毛利润
	销售毛利率
外部支持-吸引资金能力	是否香港/海外上市
	是否科技公司入股
	阳光私募持股比例
	政府补助

四、评价体系构建

4.1 评价体系框架

项目建立的科创能力评价体系涵盖了六大关键维度：行业状况、业务状况、产品与技术、管理能力、财务控制及外部支持。这一全面的评价框架能够从多角度捕捉企业的创新能力。鉴于各维度的多样性和非线性特征，项目选用随机森林算法作为核心评价模型。随机森林具备强大的多维特征处理能力和非线性关系建模能力，能够有效整合科创企业的各类创新数据。通过多棵决策树的综合判断，模型保证了良好的泛化能力和稳定性，并通过定期更新以适应市场变化，从而实现对企业创新能力的实时反馈。

4.2 模型介绍及结果

4.2.1 算法思路

采用 bootstrap 法从原始样本集中随机选择一部分数据来构建多棵决策树，并通过集成决策树的预测结果来进行最终的预测。

4.2.2 算法优势

通过对随机生成的决策树进行集成可减少单个决策树的不确定性，降低过拟合风险，提高模型的鲁棒性和准确性；由于每个决策树都是独立构建的，因此算法具有较好的并行性，能够高效地处理大规模数据集。最关键的是，随机森林算法可以评估特征的重要性，在此项目中可以帮助我们筛选出哪些特征对于科创能力的评价更为关键。虽然机器学习算法自带一定的不透明性，但随机森林算法在训练

中可通过基尼指数的高低以对特征重要性进行判断，从而提供了近似于模型但透明度更高的赋权方式。

4.2.3 模型特征

1）开放性：该算法具有开放性，不限制于特定目标或问题的定义，可以根据用户选择的数据和指标来处理多种类型的问题。模型的灵活性和适应性使其能够适应不同领域和需求的数据分析任务。

2）适应性：该算法具有适应性，它能够通过从用户的多次反馈中学习和改进自身，不断优化其性能。它拥有强化学习的能力，能够根据环境和反馈信息来调整自身的行为，以提供更准确和有效的结果。

3）可扩展性：该算法具有可扩展性，可与其他系统或服务进行集成，与现有的技术和平台进行无缝连接。同时，模型可迭代和更新，使其能够不断适应新的需求和技术变化，为用户提供更广泛的功能和服务。

4.2.4 模型步骤及结果

1）定义学习对象：明确需要评估的企业科创能力的具体指标。

2）推荐指标：根据前期研究和数据分析，推荐适合的评价指标。

3）机器学习：利用随机森林算法进行模型训练。

4）客户偏好指标：通过模型训练，确定哪些指标对于科创能力的评价更为重要。

5）特征重要性排序：对推荐的指标进行重要性排序。

6）特征重要性判断：根据基尼指数判断特征的重要性。

7）重要 - 指标方向判断：确定哪些指标对科创能力有正面或负面影响。

8）不重要指标剔除：剔除对科创能力影响较小的指标。

9）确认指标权重：根据特征重要性确定每个指标的权重。

10）构建模型：最终构建完整的科创能力评价模型。

五、项目成果与应用

5.1 项目成果

项目成果被整合至相关平台后，开发了涵盖行业信息、公司深度资料、估值分析及创新能力量化对比的综合功能模块。行业信息模块提供不同科创领域的市场概况，用户能够了解行业整体发展趋势及竞争格局。公司深度资料展示模块则为用户提供了详细的公司背景、核心技术、财务表现等信息，帮助用户全方位分析目标企业的创新潜力和市场竞争力。估值与创新能力得分模块通过实时数据跟踪，帮助用户更准确地判断企业的投资价值与科技创新潜力。

5.2 应用情况

5.2.1 公司生命树嵌入模块

公司生命树嵌入模块通过科学的科创标签分类，将企业按创新能力分为 A 至 D 四类，并以不同颜色标示，直观展示企业在科创板中的相对位置和创新表现。模块支持快速分层识别，帮助投资者高效筛选出创新能力强的优质标的，同时通过深度分析入口提供企业的创新能力评分、专利情况等详细信息，为投资决策提供数据支持。此外，模块通过动态更新实时调整分层标签，助力投资者及时识别潜在风险企业，优化策略并提升决策效率。

5.2.2 行业综合信息展示页

行业信息综合屏是科创能力评价系统的核心模块之一。该模块为买方机构提供三方面支持：行业整体概况分析、企业创新能力评分分布图谱，以及重点公司创新特征详解。该模块通过实时更新的动态评分和分组功能，帮助投资者快速识别高潜力企业或创新能力下滑的公司，规避潜在风险。行业分析和波动预警功能支持对行业创新表现进行长期跟踪，提示环境恶化或竞争加剧的信号，助力资产配置优化。在市场波动时，模块提供的分层信息便于筛选抗风险能力强、创新突出的企业，支持防御性策略部署。

5.2.3 公司深度资料展示页

公司情况页面是科创能力评价系统中的核心模块，通过整合企业的创新能力评分、行业排名、专利状况、财务数据及产业链地位等多维度信息，为投资者提供详尽的分析支持。该模块实时跟踪企业创新能力和财务表现，帮助投资者识别成长潜力与潜在风险。通过动态更新，投资者能够掌握目标企业科创能力的最新变化。页面还提供产业链地位监控，评估企业在上下游中的角色变化对其未来表现的潜在影响。此外，专利更新追踪和技术与财务数据整合功能支持评估企业技术积累、创新持续性及商业化效率，帮助投资者提前识别技术创新的竞争优势或劣势。

5.2.4 企业估值与科创能力得分对比展示页

企业估值和科创能力得分对比展示页面通过结合企业的估值数据（如市盈率、市净率）与创新能力得分，直观展示了企业在资本市场表现与科技创新能力之间的关系。该页面支持用户动态对比不同时期的估值与创新能力变化趋势，为评估企业的市场价值、创新潜力及估值合理性提供了重要依据。在持续跟踪阶段，该页面为投资团队提供了企业投资价值和创新能力的综合视图，通过对估值合理性及历史趋势的动态分析，帮助判断企业的长期增长潜力，监控市场对科技创新企业的预期偏差，支持制定更科学的

投资策略，实现收益与风险的最佳平衡。

六、项目意义与展望

本项目通过建立基于随机森林算法的科创企业评价模型，为投资者识别和评估创新型企业提供了有力工具。在汇添富投资管理公司和汇添富基金管理股份有限公司的实践应用中，该模型依托综合评估框架，通过整合多维度数据，对企业创新能力进行了合理量化评估，为投资者有效识别长期增长潜力的优质企业与潜在高风险企业提供了科学支持。应用期间，该模型显著优化了企业创新能力评分和估值分析的过程，避免了传统人工审核中可能存在的漏判和延误，大幅缩短了决策周期和风险响应时间，提升了整体业务效率。

我们的研究成果将辅助市场参与者，来识别和评估符合国家科技创新战略的优秀企业，提升投资者对科创企业的识别和判断能力。基于科创企业评价系统的多模块功能的协同应用，买方机构将能够更有效地评估企业创新能力与市场表现，掌握行业动态变化，从而做出更加前瞻性的投资决策。通过提供科学的数据支持和智能化解决方案，将助力机构投资者在资本市场中实现价值发现，提升投资回报，同时推动科创企业的健康发展与资本市场的长期稳定，有效促进资金向具有强大科技创新能力的企业倾斜，为资本市场支持实体经济高质量发展提供有力保障。通过该模型，我们希望为国家的科技创新和产业发展贡献力量，同时助力投资者在不断变化的市场环境中做出明智的决策。

参考文献：

[1] 上海证券交易所 . 上海证券交易所科创板企业上市推荐指引 .20190304.

[2] 张鹏 . 上交所细化科创板第五套上市标准 [J]. 法人 ,2022(07):67-70.

[3] 李硕容 . 科创板首次公开募股规则审核研究——基于对科创板上市公司的数据 [J]. 中国农业会计 ,2023,33(13):93-95.DOI:10.13575/j.cnki.319.2023.13.030.

[4] 曹崇延 , 王淮学 . 企业技术创新能力评价指标体系研究 [J]. 预测 ,1998(02):67-69.

[5] 肖锎 , 吴彦 . 企业科技创新能力评价研究综述 [J]. 办公室业务 ,2022(11):47-50.

基于微服务与隐私计算技术的数据安全共享服务平台 *

安鹏，张卓晖，喻波
北京明朝万达科技股份有限公司

摘　要： 本文针对多源数据融合共享应用的需求，应用基于隐私计算技术，在可信受控存储的基础上，结合密态计算和协同计算，提供机构间融合数据共享服务模式，并构建合规的数据安全共享平台应用平台，提供微服务架构及 API 网关技术支撑的服务注册、发布、订阅、调用、注销等全生命周期的数据服务管理 . 此外，针对多源数据融合共享应用的需求，采用基于微服务的数据安全共享架构，通过访问控制策略管理、敏感数据流转监测、异常行为监测与管控等技术，实现内外部数据的跨域安全共享 . 构建海量多维数据的融合共享服务平台，实现跨平台、跨主体、跨部门的多方数据安全计算，解决不可流通数据的协同应用问题。

关键字： 数据安全；微服务；隐私计算；安全共享；多方安全计算；共享服务平台

数字经济时代，数据要素已成为关键生产要素，其重要性日益凸显 . 在数字化转型的过程中，数据的流通和共享成为必需，与此同时敏感数据泄露的风险也随之加大 . 随着《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规的研究制定，我国数据保护法律法规体系将更为清晰、严谨 . 对数据的有效监管实现了有法可依，填补了数据安全保护立法的空白，完善了网络空间安全治理的法律体系 . 在强监管趋势下，粗放型数据交易模式上升为触犯法律红线的行为，目前业务仍处于此类灰色地带的企业将遭受重创，须积极探索符合合规要求的业务路线 . 保障数据安全，数据的合规合法使用成为数据流通和共享的前提 [1]. 在数据安全领域，隐私计算因保护数据安全、打破数据孤岛等优势，其优秀落地场景与案例越来越多，随着“数据安全”体系的不断完善，隐私计算 [2-3] 将实现数据价值最大化，成为数据流通和共享必需的基础设施 .

本文从数据安全共享服务平台的设计出发，对平台架构和系统组成进行详细描述，采用微服务架构实现数据资源服务总线，通过任务驱动的协同机制实现基于隐私计算的安全计算系统 . 最终在平台内部构建数据安全监测和数据集中管控系统，保证系统运行的稳定和安全。

* 本文是项目下设课题“智能信披审核和监管数据安全共享关键技术研究”（课题编号：2021YFC3340701）的研究成果，课题负责人：周琳娜（北京邮电大学）；本文部分相关成果已在《信息安全研究》发表。

一、数据安全共享服务平台架构设计

数据安全共享服务平台由 1 套平台（数据安全共享服务平台）、2 个系统（数据安全管控系统、数据安全计算系统）、3 种终端（数据安全共享 SDK、数据安全共享网关、数据安全共享节点）组成，如图 1 所示。

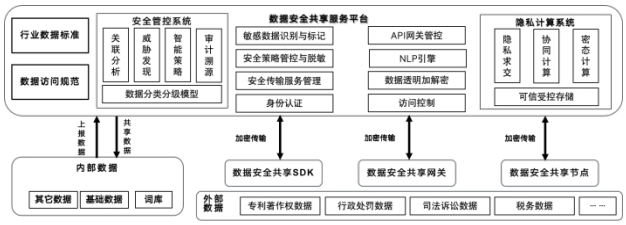


图 1 数据安全共享服务平台

平台通过支持多种数据安全接入传输的终端网关与数据安全接入服务实现外部数据大批量、高并发的安全接入与传输；并通过基于微服务架构的数据资源服务总线进行安全共享，满足符合访问授权规范要求的数据共享与访问操作 . 内部数据的共享也通过本平台完成，相关数据的访问操作必须符合访问授权规范要求 . 其他数据的共享也可以选择通过本平台实现发布使用与访问控制 [4]。

针对部分所需数据不能被直接获取，甚至部分数据不能被访问的情况，平台提供了基于隐私保护的数据安全共享节点，采用在可信受控存储环境下的多方协同分析与多方安全计算 [5-6] 实现对受限数据的“可用不可见”.

平台融入了数据安全管控系统 [7]，实现了智能匹配相应的安全管控策略，对所有数据操作进行全程安全管控和全生命周期的审计溯源，采用基于人工智能的用户异常行为分析，实现异常行为、安全风险的自动感知与处置 .

1.1 数据资源服务总线

相对于传统的资源服务总线，基于微服务架构的数据资源服务总线具有可弹性扩展、分布式、自维护、轻量级、松耦合等特点，也叫微服务 API 网关 . 采用面向服务的体系结构实现数据资源应用间的数据共享和使用，主要解决数据资源的封装问题 [8]. 如图 2 所示，包括：

数据使用方 . 需要通过总线获取数据服务的请求程序 .

数据提供方 . 在总线上提供数据服务的服务程序 .

数据服务注册 . 数据提供方将自己的数据服务和服务规约发布到服务注册中心，以便数据使用方可以发现和访问该服务 .

数据服务管理 . 总线为了发现数据服务请求、提供过程中存在的问题，记录数据服务请求、提供的内容，了解数据服务的状况、性能，从而对数据服务进行控制 .

数据服务内容 . 包括数据服务请求内容和数据服务提供内容 .

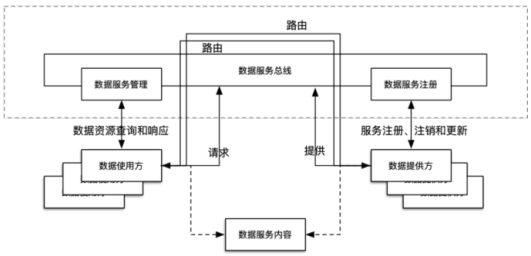


图 2 数据服务总线

数据服务总线主要功能包括对接服务、级联服务、网关服务及跨网授权、权限控制、服务注册、访问审计、日志同步等，以满足不同业务场景下的技术要求 .

1）服务注册 .

服务提供者依据服务资源注册信息格式要求，将自己的数据资源服务和服务规约发布到服务注册中心，由服务总线统一管理服务目录，以便提供调用 .

2）服务请求 .

服务请求者依据服务文档的请求调用报文格式，构造服务请求报文并发送至总线 . 服务调度主要通过代理访问模式实现，即将服务请求发往服务接口所挂接的移动服务总线，由服务总线通过路由代理访问服务接口，并返回结果；服务请求也支持直接访问模式，即由认证及授权服务根据服务请求方的权限信息，向服务请求方授予访问令牌，服务请求方拿到令牌后再向服务接口方发送请求 . 服务提供方需检查访问令牌，通过后直接向服务请求方提供服务 [9-10].

3）异步服务请求 .

服务总线支持服务请求者的异步服务请求 . 服务总线缓存服务请求的返回结果，当服务请求者获取异步请求返回结果时，再将返回结果发回给服务请求者 .

4）服务路由 .

服务路由是服务总线基于服务请求进行路由匹配的核心功能 . 服务总线接收到服务发起方通过权限校验后提交的服务请求后，开始进行路由匹配，匹配成功后就开始处理该请求，并将服务响应结果传输给服务提供方 .

相对于传统的服务总线，服务总线需要满足海量的应用访问请求，需要支持分布式的扩展 . 服务总线的路由支持 1 个 API 多个后端节点模式（即集群模式）；后端支持 IP 地址注册，也支持服务名称注册；使用服务名称注册时，移动服务总线必须提供一种可靠的服务注册发现机制，确保后端节点地址的动态变化 .

5）服务编排 .

服务编排指将多个服务进行编排形成新的服务 . 对于服务调用方，只关心想要的结果，并不关心调用的复杂过程 . 支持直观方式定义的新组合服务流程 (工作流或代码级编排)，通过少量的可视化定制化开发，即可实现服务的编排功能 .

6）访问控制 .

对接入服务总线的服务请求方和服务接口进行身份合法性验证 . 对服务请求方发出的请求进行权限检查，对于越权访问予以拒绝 . 访问控制可以是对应用层的权限审查，也可以支持对访问发起方的权限检查 . 服务总线即支持客户端身份和用户身份的访问控制，也支持同时对两种身份的访问控制 .

7）流量控制 .

流量控制可以用于管控 API 的被访问频率、应用的请求频率、用户的请求频率等 . 流量控制的时间单位可以是分钟、小时、天 . 同时支持流控例外，允许设置特殊的应用或者用户 .

8）服务管理 .

① 服务监控：实现对服务接口等相关资源的运行状态监控、性能监控、负载监控以及异常自动告警；从服务接口的在线率、访问量、访问成功率、响应速度等方面对服务质量进行评价和排名；基于监控日志，从地区、应用、时间、频度等多个维度，对服务资源运行情况进行统计分析，并采用业务视角展现服务资源的实战成果 .

② 调用链跟踪：服务总线支持识别请求方发送过来的跟踪信息，从而形成 1 条调用链并保存到日志中，后续可以通过直观的方式看到一个请求从客户端发起，经过网关路由，再到后端节点，甚至数据库的调用链，也支持查看每个环节的消耗时间、错误状态和采集到的关联日志 .

③ 异常处理：服务总线在接收到服务请求到服务结束期间发生一切异常都有完整处理 . 一方面要让服务请求方

知道服务调用失败，即异常反馈；一方面网关需要知道异常情况，即记录异常日志 .

9）安全防护 .

支持多种认证方式，支持 HMAC(SHA-1，SHA-256) 算法签名 . 支持 HTTPS 协议，支持 SSL 加密 . 防攻击、防注入、请求防重放、请求防篡改 [11].

10）安全审计 .

主要通过日志采集、分析和处理实现对服务行为进行安全审计 . 行为日志主要包括服务资源注册、授权和访问等 3 种类型 . 采集的数据项目应符合相关要求；同时通过采集汇总服务总线节点和信息资源服务资源的状态和日志信息，以此为基础提供日志查询、统计分析功能，为服务总线的运行维护提供数据支持 .

11）访问协议 .

服务总线对外包括接口支持 HTTP 和 HTTPS 协议，支持 HTTPS 证书管理；服务总线调用后端接口支持 HTTP 和 HTTPS 协议，同时支持把后端节点的 HTTP 协议转换为对外暴露 HTTPS 协议 .

12）访问鉴权 .

服务总线对外提供统一安全控制策略 . 应用访问总线时必须经过鉴权，鉴权通过后方允许访问，否则予以拦截 . 访问鉴权的模式有以下 4 种：

① 应用鉴权 . 访问服务接口的使用者必须是已注册服务的用户 . 该部分由使用方进行申请，总线完成注册，同时为服务使用者分配可使用服务的权限 .

② IP 地址鉴权 . 基于 IP 地址对服务使用者进行身份认证 . 对于不在服务使用者所申请的 IP 地址范围内产生的服务调用，总线给予拦截和告警 . 对于通过多重路由或映射导致不能获得实际 IP 地址时，应采用访问令牌方式进行替代 .

③ 用户身份认证 . 通过 OAuth2，JWT 等标准实现对所注册的接口和用户进行身份认证和权限控制 .

④ 请求校验 . 支持参数类型、参数值（范围、枚举、正则、JSON Schema）校验，无效校验直接会被 API 网关拒绝，减少无效请求对后端造成的资源浪费 .

1.2 数据安全接入服务

1.2.1 数据安全接入网关

数据安全接入服务基于代理技术开发，使用 TLS 连接提供安全服务，通过重写链接和端口来处理远程用户对内网的访问请求，采用国密加密算法，进行链路数据加密 [12]. 主动采集系统自身运行状态信息、客户端访问流量信息，确保做到过程可信、结果准确、证据可查，有效实现了“主动 / 被动安全防御”的结合，保护内部网络不被攻击，内部资源不被窃取 . 具有维护简单、移动性强、访问控制能力强等特点 .

1）数据加密传输 .

采用 TLS 协议保证通讯双方的信息安全，依赖可靠的 TCP 传输层来传输和接收数据；支持国产加密算法，进行链路数据加密；采取特有应答纠错机制，包括确定应答与重发、记录重组等机制，保证数据包有序、完整到达安全接入网关 TLS 会话模块 .

2）日志监控审计 .

提供配置远程客户端和服务发布的可视化管理平台；实时监控内网资源访问情况，自动记录相关日志；随时查看每个在线客户端的情况，可以随时中断可疑会话 .

3）资源服务发布 .

支持内网资源在安全接入平台以服务的形式发布，已发布的服务，可被客户端通过安全接入网关访问；支持对发布的服务生成唯一签名；支持服务端发布多个内网资源服务，每个服务可进行独立的认证、配置与管理；所有内网资源服务的 IP、端口对客户端不暴露 [13].

4）远程接入管控 .

支持在安全接入平台管理远程接入的客户端，客户端需在安全接入平台配置并认证，方可远程访问内部资源服务；支持配置认证多个客户端，每个客户端可进行独立的认证、配置与管理；支持同一客户端同时访问多个内网资源服务；客户端使用服务端已发布服务的签名和证书与服务端具体服务进行 TLS 握手认证 .

集成弹性伸缩、身份认证、通道管理、流量监控、服务管控等，支持跨区容灾和就近路由，规避单可用区可能存在的不可抗力风险，提高服务的高可用性和容灾能力 [14]. 线性扩展，包括本身的扩展性及业务的扩展性具有最灵活的安全接入方式，支持 Web 代理、文件共享、端口转发、网络扩展、支持 IPv6 网络 . 动态检测接入条件最优网关，智能优选接入链路，确保良好业务和数据应用体验 .

1.2.2 API 服务接口规范

所有服务的接口均基于 HTTP/HTTPS 协议，符合 Swagger 2.0 接口描述规范 . 服务提供方和服务使用方必须同时使用同一种类型的技术来进行开发和调用，调用的服务通过 HTTP URL 中特定属性进行标识，具体详见接口协议 .

服务的接口数据包含业务所有的业务数据，数据采用 JSON 格式表示，并且符合相应的 JSON Schema. 服务提供方和服务使用方必须同时使用同一种格式进行数据交互 .

1 个服务应该只实现 1 个业务功能 . 服务应是无状态的，2 次请求之间无须状态和会话的保持，并可以采用轮询的方式在负载均衡器上进行注册 .

服务请求和返回的报文应符合 JSON Schema 格式 . 服

务请求方和提供方应采用通用的 JSON 解析器来构造和解析数据，JSON 不同含义的段落应定义明确含义的字段名称，相同内容的数据应采用数组来进行描述，双方可根据 JSON 名称和路径进行精确定位，不应根据字段的顺序来获取字段值，字段值不受字段顺序调整的影响 . 报文统一采用 UTF-8 进行编码 .

为提高数据查询类服务的通用性和性能，查询类服务在入参中定义返回字段列表，服务提供方根据入参中指定的字段返回信息 .

服务提供方应对请求报文格式和关键信息进行合规性和业务校验，防止非法访问和入侵 .

服务调用方和服务提供方通常采用同步调用的方式进行请求，如需要使用异步调用可采用消息队列或服务调用方定义异步通知接口来实现 .

服务接口采用微服务架构进行开发和部署 [15]. 微服务是指开发一个单个小型的具备业务功能的服务，每个服务都有自己的处理和轻量通讯机制，可以部署在单个或多个服务器上 . 微服务架构指一种松耦合的、有一定上下文的面向服务架构 . 相对于单体架构和 SOA，微服务架构的主要特点是组件化、松耦合、自治和去中心化 .

1.3 数据安全计算系统

数据安全计算是平台基于隐私计算技术对外输出的数据计算服务能力，如图 3 所示：

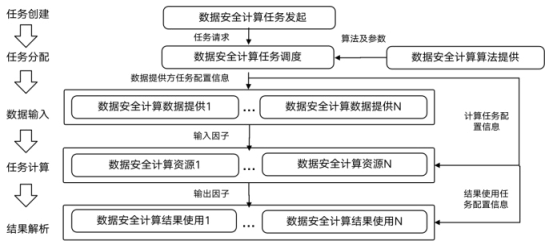


图 3 数据安全计算系统

跨信任域多参与方的数据安全计算以及联邦学习 [16-17] 任务，主要包括任务创建、任务分配、数据输入、任务计算、结果解析等步骤：

1）任务创建 .

任务发起方配置、核实数据安全计算任务所需资源，发起计算任务 . 数据提供方对所有的数据使用进行授权，任务发起方和数据提供方为同一实体的情况除外 . 数据提供方可委托调度方对数据进行使用授权，也可在任务创建前对数据进行预授权 . 数据使用授权和后续任务分配阶段可合并执行 .

2）任务分配 .

调度方验证任务请求信息的合法性，包括身份验证和数据授权的合法性 . 验证通过后生成任务配置信息，发送给数据提供方、计算方和结果使用方 . 数据提供方、计算方和结果使用方收到任务配置信息后进行验证 . 各参与方保存收发的任务配置信息 .

3）数据输入 .

数据提供方从数据源读取数据并生成输入因子，通过安全通道发送给指定计算方 . 数据提供方保存任务配置信息，并对发送的输入因子进行存证 .

4）任务计算 .

计算节点接收各数据提供方的输入因子，按照数据安全计算协议进行协同计算生成输出因子 . 将输出因子发送至结果使用方 .

5）结果解析 .

结果使用方对输出因子进行解析得到计算结果 . 并对结果进行存证 .

1.4 数据安全管控系统

数据安全管控系统是一个针对数据安全监测和数据集中管控的系统，能够收集各接入系统上报的安全事件和业务运行信息，对信息进行存储、分析、展示和响应控制，达到安全运行集中监测和管理的目的，如图 4 所示 . 同时可以帮助管理人员进行线上业务实时监控、业务异常原因定位、应用的数据统计分析、安全数据的分析和审计 . 还可以对出现的安全事件进行及时的响应控制，实现对终端的接入管控，主动断开存在安全威胁终端的连接 [18]，对内部的数据和应用服务进行保护 .

图 4 数据安全管控系统

1）数据采集 .

数据采集是系统安全监控的基础，主要负责采集基础信息和运行数据，采集的过程主要采用探针技术，在不同模块中安装探针，实现获取数据的目标 . 数据采集探针用来探测终端、网络、应用、数据基础信息外，还负责探测安全事件和业务运行数据，并将探测结果定时上报至监控中心 .

2）数据分析 .

平台对采集的信息进行分析，并做分类处理 . 采集的信息一般将被分为 2 大类：统计信息和安全事件 . 统计信息包括设备信息和流量信息等 . 相应的，安全事件则是违背监测策略项的内容 . 通过对采集到的监测信息进行分类、分析，在平台内进行可视化展示 .

3）数据展示 .

主要对采集到的信息，通过引入安全框架进行分析，对得到的结果通过大屏进行可视化展示 . 主要包括：整体安全信息分析展示用户信息分析展示、网络信息分析展示、终端信息分析展示、应用信息分析展示、数据信息分析展示、安全事件分析展示构成 .

4）响应控制 .

对发生的安全事件，提供具体管控能力 . 主要包括，告警提示、终端控制、应用控制以及数据控制 . 可以在管控平台的策略模块根据安全事件的严重程度，针对用户、网络、终端、应用、数据配置不同的管控策略 .

5）平台管理 .

平台管理负责对安全保护环境中的计算节点、安全区域边界、安全通信网络实施集中管理和维护，包括用户身份管理、终端信息管理、接入设备管理、权限管理、应急处理等，为平台的安全提供基础性保障 . 平台管理符合国家相关安全规定和标准，监测内容标准化、采集数据格式标准化、设备接口标准化、违规信息处理标准化 .

二、总结

数据安全共享服务平台综合应用微服务、隐私计算等技术，满足了跨行业、跨区域的多源数据安全对接、传输与共享需求 . 采用 API 网关进行安全共享，满足符合访问授权规范要求的数据共享与访问操作，在保证数据安全前提下提供数据共享服务能力 . 针对内外部数据不能被直接获取，甚至部分数据不能被访问的情况，平台基于隐私计算技术提供了数据安全计算系统与数据安全共享节点，采用任务驱动的系统模式，利用可信受控存储环境下的数据分析与多方安全计算实现对受限数据的“可用不可见”.

参考文献：

[1] 蒋凯元 . 多方安全计算研究综述 [J]. 信息安全研究 , 2021,7(12): 1161-1165

[2] 范江波 . 以个人数据权益保护为核心的大数据权益保护研究 [J]. 信息安全研究 , 2021,7(12): 1166-1177

[3] 国家工业信息安全发展研究中心 . 中国隐私计算产业发展报告 （2020—2021）[R]. 北京：国家工业信息安全发展研究中心， 2021

[4] 李凤华 , 李晖 , 贾焰 , 等 . 隐私计算研究范畴及发展趋势 [J]. 通信学报 , 2016, 37(4):1-11

[5] 丁毅 , 沈薇 , 李海生 , 等 . 面向 CNN 的区块链可信隐私服务计算模型 [J]. 电子学报 , 2022, 50(6):1399-1409

[6] 张勇 , 李丹丹 , 韩璐 , 等 . 隐私保护的群体感知数据交易算法 [J]. 通信学报 , 2022,43(5):1-13

[7] 熊金波 , 周永洁 , 毕仁万 , 等 . 边缘协同的轻量级隐私保护分类框架 [J]. 通信学报 , 2022, 43(1):1-11

[8] 张贺 , 王忠杰 , 陈连平 , 等 . 面向持续软件工程的微服务架构技术专题前言 [J]. 软件学报 , 2021, 32(5):1229-1230

[9] 彭鑫，赵文耘，吴毅坚，等 . 一个基于服务请求语言的统一 Web 服务框架 [J]. 计算机科学 , 2006, 33(1):86-90

[10] Zhang J， Jiang Z L， Li P , et al. Privacy-Preserving Multikey Computing Framework for Encrypted Data in the Cloud[J]. Information Sciences, 2021, 575(3):217-230

[11] Li C , Lv Q , Li N , et al. A novel deep framework for dynamic malware detection based on API sequence intrinsic features[J]. Computers & Security, 2022, 116:102686

[12] Xue K， Liu Z， Zhu H , et al. Advances in privacy-preserving computing[J]. Peer-to-Peer Networking and Applications, 2021, 14(3):1348-1352

[13] Fazal R， Shah M A， Khattak H A , et al. Achieving data privacy for decision support systems in times of massive data sharing[J]. Cluster Computing, 2022(18):1-13

[14] 钱文君 , 沈晴霓 , 吴鹏飞 , 等 . 大数据计算环境下的隐私保护技术研究进展 [J]. 计算机学报 , 2022, 45(4):669-701

[15] Du W， Atallah M J . Atallah, Secure multi-party computation problems and their applications: A review and open problems[C] //Proc of Workshop New Secur. Paradigms, 2001: 13–22

[16] 张泽辉 , 富瑶 , 高铁杠 . 支持数据隐私保护的联邦深度神经网络模型研究 [J]. 自动化学报 , 2022, 48(5): 1273-1284

[17] 刘艺璇 , 陈红 , 刘宇涵 , 等 . 联邦学习中的隐私保护技术 [J]. 软件学报 , 2022, 33(3):36

[18] Shukla S， Patel S J . A novel ECC-based provably secure and privacy-preserving multi-factor authentication protocol for cloud computing[J]. Computing, 2022, 104(5):1173-1202

FinBERT2：弥合 LLM 在金融领域部署差距的双向编码器

徐璇 [1]，温富方 [2]，储贝林 [1]，付志兵 [2]，林钦鸿 [1]，刘佳琪 [2]，费斌杰 [2]，李渔 [2]，杨忠良 [1]，周琳娜 [1]

[1] 北京邮电大学 网络空间安全学院 北京 102206

[2] 北京熵简科技有限公司 北京 100026

E-mail：sh22xuxuan@bupt.edu.cn

摘　要： 在自然语言处理中，重点已经从像 BERT 这样的仅编码器小型语言模型转向像 GPT-3 这样的仅解码器大语言模型。然而，LLMs 在金融领域的实际应用存在了三个局限性：（1）LLMs 在判别任务上的表现往往比微调的 BERT 更差，尽管计算资源成本更高，例如金融报告中的市场情绪分析；（2）生成任务的应用严重依赖检索增强生成方法来提供实时和专业的信息，而通用检索器在领域特定检索任务上表现不佳；（3）在其他基于特征的场景中存在额外的不足，例如主题建模。我们介绍了 FinBERT2，这是一个在高质量、金融特定的 32B token 语料库上预训练的专业双向编码器。作为更好的骨干模型，FinBERT2 可以通过以下成就弥合 LLMs 在金融特定部署中的差距：（1）判别性微调模型（Fin-Labelers）在五个金融分类任务上平均优于其他（Fin）BERT 变体。（2）对比学习微调模型（Fin-Retrievers）在五个金融检索任务上优于开源和专有嵌入模型；（3）基于 FinBERT2 变体，我们构建了 Fin-TopicModel，为研报标题实现更好的聚类和主题表示。

关键字： FinBERT；语言模型；密集检索器；主题建模

一、引言

早期 LLMs 主要依赖具有掩码语言建模（MLM）的仅编码器架构，如 BERT[1]、RoBERTa[2] 和 XLM[3]。然而，在 2018 年至 2021 年期间，该领域从单任务微调转向大规模多任务学习。GPT-3 证明了扩展可以显著缩小自回归架构与其他架构之间的性能差距。此外，自回归模型提供了更大的任务适应性、统一建模范式和降低工程复杂性等优势。因此，仅解码器模型已成为 LLMs 开发的主导范式。

同样，金融领域见证了从早期 FinBERT[4] 到参数范围从数十亿到数万亿的大规模 FinLLMs 的转变，如 FinGPT、BloombergGPT 和 FinLlama。这些模型利用领域特定数据并在金融文本语料库上进行后训练，增强了它们在金融应用中的理解和生成能力。然而，包括金融适应版本在内的 LLMs 并不能完全替代 BERT。它们在现实世界部署中仍然面临局限性。

首先，虽然 LLMs 作为大规模多任务模型表现出强大的泛化能力和鲁棒性，但它们对于特定 NLU 任务并不总是最优的。在某些单任务场景中，微调的 BERT-base 模型往往优于它们。例如，有研究 [5] 在 25 个分析性 NLP 任务上评估 ChatGPT 时发现，与最先进（SOTA）方法相比，其零样本和少样本性能平均下降了约 25%。对于复杂任务，下降更为明显。同样，在假新闻检测 [6] 中，GPT-3.5 的表现不如 BERT 等专门的较小模型。此外，LLMs 对于数据密集型任务（如标记金融报告）成本高且速度慢，而较小的 BERT 类模型（0.1B 参数）更高效。

其次，在需要外部金融知识的生成场景中，如金融报告的实时问答，LLMs 依赖基于嵌入的检索来确保准确性和及时性。这种方法被称为检索增强生成（RAG），需要高效的离线检索。因此，基于双编码器 BERT 架构构建的密集检索器（DRs），如 M3E[7]、BGE[8] 和 BCE[9]，已成为主流。这些 DRs 通过在大规模弱监督句子对上进行批内负学习，以及使用挖掘的硬负样本进行对比微调，在通用基准上实现了高检索精度。然而，尽管进行了广泛训练，领域外泛化仍然有限。这些 DRs 甚至在没有对标记数据集进行进一步微调的情况下无法达到 BM25 水平。

第三，除了检索之外，LLMs 在基于特征的任务中效果较差或不够成熟，如基于聚类的主题建模、衡量盈利惊喜和市场反应，以及提取股价预测因子。生成模型通常缺乏这些应用的成熟用例，因为它们需要紧凑高效的特征编

码和灵活的任务特定适应微调。例如，主题建模优先考虑行业特定特征，而股票回报预测依赖情绪特征。然而，仅解码器LLMs由于其固有的架构约束而难以满足这些要求。

上述挑战阻碍了中小型金融企业的应用，这些企业的场景通常是专门化和多样化的。为了解决这个问题，我们重新评估了轻量级、本地化和可定制的 FinBERT 模型的价值，并提出了一个混合架构，将 FinBERTs（作为缓解这些局限性的领域专家）与 LLMs（作为具有上下文学习能力的通用生成模型）集成。

具体而言，我们预训练了 FinBERT2，这是其前身 FinBERT1[4] 的增强版本。FinBERT2 在精心策划的包含 32B token 的中文金融语料库上训练，并进一步针对标记、检索和主题建模 [10] 等下游任务进行优化。在这些任务中，FinBERT2 可以有效地替代、协助或补充 LLMs，为金融应用提供更高效和可部署的 NLP 系统。我们的贡献可以总结如下：

1）我们在 32B token 中文金融语料库上预训练了 FinBERT2 以注入领域知识。据我们所知，这是中文金融）领域 BERT 类语言模型最大的预训练语料库，并使用金融定制分词器进行训练。

2）作为标记的更高效和高性能替代方案，Fin-Labelers 在五个金融分类任务上平均优于其他（Fin）BERT 变体 0.4%–3.3%，优于领先的 LLMs（如 GPT-4-turbo、Claude 3.5 Sonnet、Qwen2）9.7%–12.3%。

3）作为增强的 RAG 助手，Fin-Retrievers 超越了开源和专有嵌入模型。它们在五个金融检索任务上比 BGE-base-zh 平均提升 +6.8%，比 OpenAI 的 text-embedding-3-large 平均提升 +4.2%。

4）在主题建模等基于特征的应用中，基于 FinBERT2 变体构建的 Fin-TopicModel 为金融标题实现了更好的聚类和主题表示。

二、方法

2.1 FinBERT2 概述

我们的工作从由 Fin-Corpus 和 Fin-downstream 数据集组成的数据层开始，通过产生 Fin-Tokenizer 和 FinBERT2-base/large 的基础层，到具有三个主要组件的下游应用层：（1）Fin-Labeler，针对五个下游金融任务进行微调，包括市场情绪分类、行业分类和命名实体识别（NER）；（2）Fin-Retriever，通过对比学习为五个金融检索任务训练；（3）Fin-TopicModel，集成了从其他 FinBERT2 变体派生的关键组件和改进。

2.2 FinBERT2 的预训练

2.2.1 预训练的 Fin-Corpus

与 FinBERT1 相比，FinBERT2 模型的预训练语料库有了显著扩展，总 token 数增加到 32B（99G）。

1) 分析师报告语料库（16B tokens，53G）：我们编制了 260 万份金融分析报告，涵盖股票 / 期货研究、行业分析和机构评论等二十多种类型的报告。数据集跨越过去 15 年，并经过了详细的数据清洗。

2) 公司公告语料库（6.4B tokens，19G）：来源于国内上市公司官方网站的网络爬取公告，该数据集包括来自各个行业的广泛企业披露，如财务报告、重大事件声明、股东大会通知、股票回购计划和高管变更。它跨越 20 年，格式已标准化以与报告数据格式保持一致。

3）度小满开源新闻 FinCorpus（9.6B tokens，27G）：包括从多个来源聚合的文章和信息，包括主要金融新闻网站和社交媒体，该数据集提供了金融新闻和见解的综合集合。它跨越 20 年，格式已标准化以与报告数据

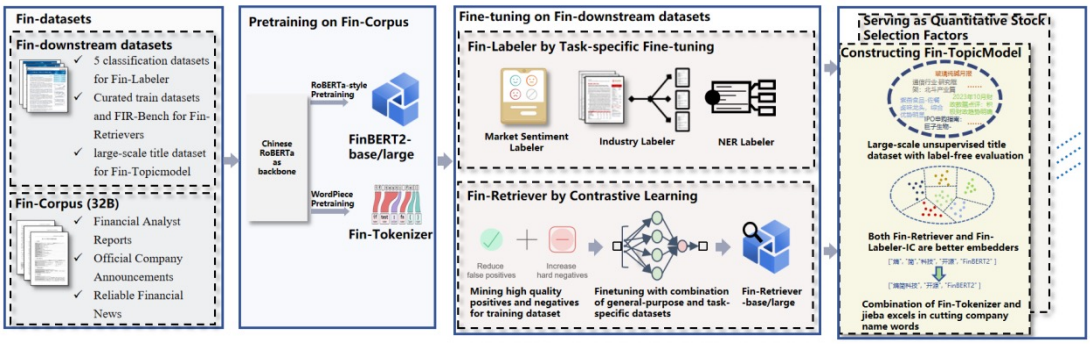


图 1 工作概览

格式保持一致。

2.2.2 过滤低质量 Fin-Corpus

我们的 32B 金融语料库包含噪声数据，包括 URL、冗长内容以及不连贯或重复的文本。由于使用 LLMs 进行全数据集过滤成本高昂，我们将 LLM 判断质量的能力提炼到轻量级 BERT 中。

我们使用 Qwen2.5-72B-Instruct 对 100K token 子集进行 1-10 分评分，将 8 分以上的数据标记为高质量，4 分以下的标记为低质量。这产生了 4K 实例训练集（每类 2K，90%–10% 训练 - 测试分割）。

在数据集上微调 RoBERTa-wwm-Chinese 产生了一个准确率超过 99% 的分类器，我们用它来过滤整个语料库，移除 15% 的低质量数据。

2.2.3 Fin-Tokenizer 的扩展词汇表

使用 WordPiece 算法，我们从 32B token 金融语料库和 "Colossal, Cleaned, Common Crawl (C4)" 数据集中提取领域特定词汇表，后者是 Chinese-RoBERTa 原始预训练数据的一部分。这个过程将模型的词汇表扩展了 14,000 个单词，包含了大量高频金融术语和公司名称（如 BYD）。基于 Fin-Tokenizer，我们在金融语料库上进行后预训练，产生 FinBERT2，一个更好地适应领域特定任务的模型。

2.3 Fin-Labelers 的任务特定微调

2.3.1 Fin-Labelers 的五个下游数据集

由于现有公共中文金融数据集与真实世界商业实践之间的差异，我们通过直接从金融终端系统提取和注释数据构建了五个金融分类数据集。该数据集涵盖三个金融应用场景，包括：

1) 报告相关行业分类（IC）：根据中国国际信托投资公司（CITIC）一级行业分类对报告相关段落进行分类，涵盖 28 个行业类别。

2) 市场情绪分类（MSC）：该任务旨在对与金融事件或资产相关的文本评论进行情绪分类，促进市场情绪分析和股票相关性研究。第一种情绪分类应用于报告，包括代表不同情绪极性和强度的四个类别。第二种专注于新闻情绪分类，包括两个类别：正面和负面。

3) 金融报告中的命名实体识别（NER）：识别和提取金融报告中出现的实体（如公司或个人名称）。

2.3.2 任务特定微调细节

BERT 中的微调通常优化所有参数以最大化分类概率。对于像 MSC 和 IC 这样的序列分类任务，[CLS] token 的嵌入被输入到全连接层以预测类别标签概率。对于像 NER 这样的 token 分类任务，每个 token 的对应向量用于预测。

2.4 Fin-Retriever 的对比学习

我们在 64,000 个金融样本和 150,000 个通用样本上使用对比学习微调 FinBERT2，这些样本经过精心策划，具有平衡的正负样本以获得优化性能。领域检索性能的评估基于金融信息检索基准（FIR-Bench），其包含五个金融数据集，使用召回率进行评估。

2.4.1 构建 Fin-Retriever 的训练数据集

为了增强 Fin-Retriever 的金融和通用检索能力，我们在通用和金融特定文本对的组合上微调预训练的 FinBERT2。训练数据包括 64,000 个金融 QA 数据、100,000 个 T2Retrieval 数据、40,000 个 MMarcoRetrieval 数据和 10,000 个 DuRetrieval 数据。此外，我们优化了设置并增强了训练数据集中正负样本的质量。

1) 挖掘足够的硬负样本：由于收集的训练数据集缺乏足够的负样本，我们利用 FinBERT2-base 为每个查询挖掘最多 50 个硬负样本。此外，金融数据被集成到样本池中以增强数据集的多样性，从而提高模型的领域特定能力。

2) 平衡正负样本比例：实现正负样本之间的平衡对于最佳性能至关重要。经过大量实验，我们发现 2 个正样本到 8 个负样本的比例产生最佳结果。将负样本数量增加到 15 个保持可比性能，而减少到 5 个导致性能下降。对于比 Fin-Retriever-base 容量更大的 Fin-Retriever-large，每个查询的负样本数量从 10 个增加到 15 个以更好地利用其容量。

3) 过滤假正样本和假负样本：在 T2Retrieval、MMarcoRetrieval 和 DuRetrieval 数据集中，正样本可能并不总是直接回答查询。为了解决这个问题，我们利用 Qwen2.5-72b-Instruct 进行过滤，确保正样本包含足够的信息来有效响应查询，从而提高其质量。同样，为了确保负样本不包含可能回答查询的信息，我们使用 Qwen2.5-72b-Instruct 进行进一步过滤。通过组合跨数据集的高质量正负样本，我们构建了一个强大有效的训练数据集。

2.4.2 金融信息检索基准（FIR-Bench）

为了全面评估其领域特定能力，我们策划了五个金融检索测试数据集（FIR-Bench），这些数据集来源于我们的业务数据，确保对其在金融领域性能的全面评估。

1) 单文档金融问答测试数据集（Sin-Doc FinQA）：该数据集由链接到正负文档样本的查询组成。对于每个查询，候选文档来自同一篇文章，正文档数量从最少 1 个到最多 10 个不等。平均而言，每个查询与 8.4 个文档相关联，其中 2.6 个是正样本。

2) 多文档金融问答测试数据集（Multi-Docs FinQA）：与 Sin-Doc FinQA 不同，Multi-Docs FinQA 的

文档样本来自不同文章；因此，它包含更多的正负文档。正文档的最大数量限制为 50 个。在语料库中，每个问题平均与 9,384 个文档相关联，其中 14 个平均被标记为正样本。

3) 三种来源的金融检索数据集：这包括三个数据集——金融研究报告、指标和公告。我们使用余弦相似度度量来计算每个数据集中正样本的召回率。

2.5 构建 Fin-TopicModel 的管道

我们在 Fin-Retriever 和其他 FinBERT2 相关组件的基础上开发了 Fin-TopicModel。它使用 HDBSCAN 算法对 Fin-Retriever 嵌入进行无监督聚类以获得多个聚类（主题）。对于每个聚类，它使用 c-TF-IDF（基于类别的术语频率-逆文档频率）来衡量聚类内单词的重要性。通过分析每个聚类中的高频词汇，它自动生成主题描述。

2.5.1 大规模无监督标题数据集与无标签评估

我们构建了一个包含从 2022-2024 年发布的 59,014 篇文章中提取的 56,540 个报告标题的数据集，平均标题长度为 27 个字符。该数据集在没有手动注释的情况下进行了无标签主题建模。综合评估由使用 LLMs 的主观评分（如连贯性、简洁性和信息性）、聚类指标如轮廓系数和 Calinski-Harabasz 指数，以及包括主题多样性和异常值率的额外指标组成，这些指标提供了进一步的见解。这个框架在缺乏标记数据的情况下促进了主题建模的全面和无监督探索。

2.5.2 从 FinBERT2 变体编码

在传统实践中，检索模型通常用于从文本中提取嵌入以支持主题建模任务。为此，我们选择了 Fin-Retriever 来编码文档。同时，我们相信 FinLabeler-IC（微调的行业分类模型）允许 BERT 在训练期间进一步学习与特定行业分类任务相关的语义信息。这些微调的嵌入捕获了识别行业的详细语义特征，使其非常适合主题建模任务。后续实验证实，两个选择的模型在主题建模中表现出不同的优势。

2.5.3 精确词汇切分

我们使用 Fin-Labeler-NER 从标题中提取 3,290 个公司相关命名实体，这些实体被添加到 Jieba 的自定义词典中以增强分词。为了比较 Fin-Tokenizer 的金融词汇与 Jieba 的默认分词，我们在 Fin-Tokenizer 的 13,804 个自定义金融术语上测试了 Jieba，揭示了 1,724 个分词不一致。分析显示 545 个是双字符术语（主要是辅助词或副词），1,179 个是四字符术语（主要是公司名称）。

Fin-Labeler-NER 依赖神经方法，而 Fin-Tokenizer 使用统计方法进行子词分词。为了结合神经和统计方法的优势，我们设计了一个合并分词器。使用贪婪策略，它选择具有最长覆盖范围的分词切分，从而增强基于 NER 分词的准确性和覆盖范围。

三、评估和分析

3.1 Fin-Labelers

我们对 FinBERT2 与通用领域中文 BERTs（如原始 BERT、MacBERT 和 RoBERTa）和金融领域中文 BERTs（如 FinBERT1 和 Mengzi-Fin）进行了基准测试。我们还对领先的 LLMs 进行了基准测试，包括 Qwen2-72b-Instruct、

表 1 FinBERT 2 和其他基线模型在分类任务中的表现

	Backbone	IC	MSC(4 labels)	MSC(2 labels)	NER(person)	NER(company)	Avg
LLMs	Qwen2-72b-Instruct	0.9250	0.4880	0.8850	0.9669	**0.8995**	0.8329
	GPT-4-turbo	0.8600	0.4750	0.8880	0.9315	0.8787	0.8066
	Claude-3.5-Sonnet	0.9030	0.5230	0.8650	**0.9957**	0.8683	0.8310
General BERTs	BERT-base-chinese	0.9166	0.8676	0.8840	0.9901	0.8269	0.8970
	Chinese-MacBERT-base	0.9128	0.8616	0.9422	0.9854	0.8324	0.9069
	Chinese-RoBERTa-wwm-ext	0.9196	0.8841	0.9424	0.9901	0.8158	0.9104
FinBERTs	FinBERT1-base	0.9294	0.9147	0.9453	0.9901	0.8481	0.9255
	Mengzi-BERT-base-fin	0.9083	0.8657	0.9498	0.9902	0.8324	0.9093
	FinBERT2-base (ours)	0.9398	**0.9249**	0.9546	0.9901	0.8378	0.9295
	FinBERT2–large (ours)	**0.9432**	0.9131	**0.9573**	0.9804	0.8514	**0.9291**

GPT-4-turbo 和 Claude-3.5-Sonnet。

3.1.1 与 LLMs 和其他 BERTs 比较

Qwen2-72b-Instruct（0.8329）、GPT-4-turbo（0.8066）和 Claude-3.5-Sonnet（0.8310）的平均 F1 分数显著低于 FinBERT2-base（0.9295）和 FinBERT2-large（0.9291）。在具有挑战性的市场情绪分类任务上，LLMs 得分低于 0.523，远远落后于 FinBERT2-base（0.9249）。对于 NER 任务，LLMs 表现出优势——例如，Claude-3.5-Sonnet 在公司名称方面表现出色（0.868）。通用 BERTs 在五个任务上平均得分为 0.8970–0.9104，而 FinBERT2-base 和 FinBERT2-large 以 0.9295 和 0.9291 的分数优于它们。对于复杂的四类市场情绪分类（MSC）任务，FinBERT2-base 达到 0.9249，而通用模型为 0.8841，突出了其捕获领域特定细微差别的能力。

3.1.2 不同任务和架构的分析

在简单 NER 任务的背景下，如人名识别，一些 LLMs 可能表现出实体遗漏问题，导致性能不如微调的 BERT 模型。然而，在复杂的 NER 任务中，如公司名称识别，LLMs 能够更好地展示更强的泛化能力。这表明 LLMs 和基于 BERT 的模型的性能根据 NER 任务的复杂性而显著变化。此外，FinBERT2-large 在复杂 NER 任务上优于 FinBERT2-base。这表明增加模型规模有助于改善复杂 NER 任务的泛化。

MSC（4 标签）由于依赖行业特定注释标准而对 LLMs 提出了独特挑战。注释标准来自专业分析师，表现出某些行业特定特征。例如，在传统行业中，10% 的增长可能被认为是高度积极的，而在另一个行业中，它可能只被视为中等积极。LLMs 的上下文学习能力通常局限于表面语义和数值。这种限制限制了它们对涉及细微、行业特定标准的任务的适应性，这些标准通常难以通过提示表达。

3.2 Fin-Retriever

我们将 Fin-Retriever 与类似参数大小的流行开源模型（BAAI 通用嵌入（BGE）、双向对比嵌入（BCE））以及 OpenAI 的专有模型进行比较。Fin-Retriever-base 和 Fin-Retriever-large 在 FIR-Bench 上都优于通用检索器，展示了它们强大的领域专业化。Fin-Retriever-large 始终实现最高召回分数，平均 R@k 为 0.746，而 Fin-Retriever-base 也超过大多数基线，平均达到 0.730。与表现最佳的通用模型 text-embedding-ada-002-large（0.723 平均）相比，两个 Fin-Retriever 变体都表现出优越的金融检索能力。特别是在研究报告和公告中，Fin-Retriever 模型表现出显著优势。

表 2 Fin-Retriever 与其他密集检索器在 FIR-Bench 上的性能表现

DR Model	Sin-Doc-FinQA			Multi-Docs-FinQA		Research Reports		Announcements		Indicators	
	R@1	R@3	R@5	R@20	R@50	R@10	R@20	R@10	R@20	R@5	R@10
BGE-base-zh	0.479	0.815	0.906	0.238	0.318	0.921	0.960	0.387	0.482	0.910	0.930
BCE-embedding-base	0.513	0.824	0.902	0.227	0.309	0.967	0.978	0.318	0.421	0.803	0.915
text-embedding-3-small	0.511	0.823	0.906	0.197	0.234	0.864	0.872	0.473	0.509	0.863	0.928
text-embedding-3-large	**0.560**	0.845	0.920	0.215	0.257	0.951	0.960	0.492	0.526	0.940	0.965
Fin-Retriever-base (Ours)	0.520	0.846	0.916	0.307	0.398	**0.987**	**0.991**	0.566	0.642	0.950	**0.975**
Fin-Retriever-large (Ours)	0.554	**0.867**	**0.937**	**0.315**	**0.402**	0.983	0.987	**0.571**	**0.664**	**0.960**	0.970

表 3 主题建模任务中不同嵌入模型的表现

	Avg LLM-score (0-3)			Cluster Qulity Metrics		
Embedder in pipeline	Coherence	Conciseness	Informativity	Silhouette Coefficient	Calinski Harabasz Score	Davies Bouldin Score
BGE-base-zh	1.765	1.550	1.865	0.141	11.394	1.254
BCE-embedding-base	1.790	1.460	1.870	0.171	12.483	1.158
text-embedding-3-small	1.744	1.533	1.809	0.106	11.204	1.274
text-embedding-3-large	1.795	1.445	1.825	0.182	12.934	1.155
Fin-Labeler-IC	1.830	1.500	1.845	0.170	11.108	1.108
Fin-Retriever-base	**1.835**	1.515	**1.905**	**0.192**	**13.296**	1.077
Fin-Retriever-large	1.760	**1.570**	1.860	0.174	12.690	**1.070**

3.3 Fin-TopicModel

3.3.1 指标

评估主题模型是一个复杂且不断发展的挑战。我们为 Fin-TopicModel 引入了一套无监督评估指标，从而避免了标记数据的需要。指标分为三大类。首先，主观评分利用 LLMs 评估主题描述，如连贯性、简洁性和信息性。其次，采用聚类质量指标，包括轮廓系数和 Calinski-Harabasz 指数。第三，纳入主题多样性和异常值率等补充指标以提供更细致的理解。

3.3.2 基于 Fin-Retrievers 的 Fin-TopicModel 分析

基于 Fin-Retriever-base 的 Fin-TopicModel 在主观评分（连贯性、简洁性和信息性）和聚类质量指标方面都表现出全面的优越性能。它实现了最高连贯性分数（1.835），并在聚类紧凑性和分离性方面表现出色，这反映在最高 Calinski-Harabasz 指数（13.296）和低 Davies-Bouldin 分数（1.077）上。这些结果突出了 Fin-Retriever-base 在主题建模任务中的更好能力。此外，它实现了高主题多样性分数（0.218），同时保持最低异常值率（0.222）。相比之下，Fin-Retriever-large 也表现良好，但略逊于 Fin-Retriever-base。这种性能差距可能归因于增加的窗口长度，导致短文本表示的劣势。

3.3.3 基于 Fin-Labeler-IC 的 TopicModel 分析

基于 Fin-Labeler-IC 的 Fin-TopicModel 在主题建模中展示了独特的优势和权衡。它生成最高数量的主题（13,417），显著多于其他模型，这表明其捕获数据中细微差别的潜力。此外，它实现了竞争性的语义质量和聚类质量，连贯性（1.830）和信息性（1.845）分数。然而，这种粒度以较低的主题多样性（0.196）为代价，表明主题的潜在冗余或过度分割。尽管其微调方法更简单，Fin-Labeler-IC 通过利用其行业特定训练目标，本质上与主题相关性一致，因此适合主题建模。这种权衡突出了任务特定嵌入对应用的价值。

四、结论

在这项工作中，我们介绍了 FinBERT2，一个在已知最大的中文金融语料库（32B tokens）上预训练的专业双向编码器。我们的结果表明，仅编码器模型在金融 NLP 中仍然发挥着关键作用，补充了仅解码器 LLMs 的优势。具体而言，FinBERT2 实现了（1）在判别性金融分类任务上的优越性能，优于现有（Fin）BERT 变体和领先的 LLMs；（2）通过 Fin-Retrievers 增强检索能力，超越开源和专有嵌入模型；（3）通过 Fin-TopicModel 改进主题建模，产生更好的聚类和主题表示。这些发现突出了基于编码器的架构在金融 AI 中的持续相关性，特别是在需要高精度和领域特定理解的场景中。

参考文献：

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [C]. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[2] Yinhan Liu, Myle Ott, Naman Goyal, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [J]. ArXiv, 2019.

[3] Guillaume Lample. Cross-lingual language model pretraining [J]. arXiv preprint arXiv:1901.07291, 2019.

[4] Panpan Hou, Mengchao Zhang, Zhibing Fu, et al. FinBERT [EB/OL]. 2020. https://github.com/valuesimplex/FinBERT.

[5] Jan Koco ´n, Igor Cichecki, Oliwier Kaszyca, et al. ChatGPT: Jack of All Trades, Master of None [J]. Information Fusion, 2023, 99: 101861.

[6] Beizhe Hu, Qiang Sheng, Juan Cao, et al. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(20): 22105-22113.

[7] Yuxin Wang, Qingxuan Sun, Sicheng He. M3E: Moka Massive Mixed Embedding Model [EB/OL]. 2023.

[8] Shitao Xiao, Zheng Liu, Peitian Zhang, et al. C-Pack: Packaged Resources To Advance General Chinese Embedding [EB/OL]. 2023. https://arxiv.org/abs/2309.07597.

[9] NetEase Youdao Inc. BCEmbedding: Bilingual and Crosslingual Embedding for RAG [EB/OL]. 2023. https://github.com/netease-youdao/BCEmbedding.

[10] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [J]. arXiv:2203.05794 [cs.CL], 2022. https://arxiv.org/abs/2203.05794.

04 项目大事记

项目大事记

2021 年 5 月	“社会治理与智慧社会科技支撑”重点专项 2021 年度“揭榜挂帅”榜单发布。
2021 年 9 月	提交项目申报书及相关材料。
2021 年 12 月	项目获批立项。
2022 年 7 月	召开项目启动会，项目实施方案通过专家论证。
2022 年 12 月	项目单位复旦大学召开“第七届金融科技与智能监管国际峰会”，组织研讨项目阶段性成果与技术方案。
2023 年 4 月	项目中期考核指标测试方案通过专家论证。
2023 年 5 月	项目完成原型系统建设。
2023 年 6 月	项目应用示范方案通过用户单位和专家论证。
2023 年 6 月	召开第一次用户单位沟通交流会，宣传推介项目成果。
2023 年 8 月	通过项目中期检查暨第一次里程碑考核。
2023 年 12 月	完成项目成果集成，在上交所、华泰证券、海通证券、汇添富基金等用户单位开展应用示范。
2024 年 1 月	参与“社会治理与智慧社会科技支撑”重点专项金融监管项目群研讨交流。
2024 年 1 月	代表性成果分别由上海证券交易所和海通证券申报并获得中国人民银行 2022 年度金融科技发展奖一等奖、二等奖。
2024 年 4 月	通过项目第二次里程碑考核。
2024 年 9 月	项目结项考核指标测试方案通过专家论证。
2024 年 11 月	完成所有既定任务，项目实施结束。
2024 年 11 月	代表性成果由北京邮电大学申报并获得中国通信学会科技技术奖二等奖。
2025 年 1 月	项目组完成课题结题验收与绩效评价。
2025 年 4 月	通过项目结项验收与综合绩效评价。

《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、数字货币及稳定币在交易结算场景的应用等。

（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！