THE FORELAND OF
TRADING TECHNOLOGY

内部资料 免费交流
《准印证》编号沪（K）0671

交易技术前沿

2023年 第三期 总第54期

本期主题

大语言模型专刊

No.3

内部资料 2023 年第三期（总第 54 期）
准印证号：沪（K）0671

NO.3

主管：上海证券交易所
主办：上海证券交易所
总编：邱勇、蔡建春
副总编：王泊
执行总编：唐忆
责任编辑：徐广斌、徐丹、陆伟、王昕、黄淦

上海市杨高南路 388 号
邮编：200127
电话：021-68607129，021-68602496
传真：021-68813188
投稿邮箱：ftt.editor@sse.com.cn

篇首语

2023 年 5 月，习近平总书记主持召开二十届中央财经委员会第一次会议时强调，新一代信息技术等战略性新兴产业是引导未来经济社会发展的重要力量，要把握人工智能等新科技革命浪潮。GPT-3.5 和 ChatGPT 的发布标志着新一轮以大语言模型（LLM）技术领衔的人工智能发展再起高潮。面对大语言模型强劲发展的技术趋势，百度、科大讯飞、360 等国内互联网巨头紧跟人工智能技术热点，先后推出旗下的大语言模型产品，在中文理解能力、多模态、内容生成等方向上百花齐放。证券期货业作为数字技术应用的重点领域，也在积极探索大模型技术的应用场景，推进机构的数字化、智能化转型，促进业务效能提升和行业高质量发展。本期《交易技术前沿》专刊的主题是“大语言模型”，本刊编辑部在前期征稿、邀稿的基础上，精选并收录行业关于大语言模型和通用人工智能技术在数字员工、合规建设、金融文本识别等领域探索与应用的优秀文章，为行业推广大模型应用提供借鉴经验。

国金证券的《证券行业大语言模型的探索与实践》探讨了证券行业大模型与现有系统融合模式的可行性，提出了将 AIGC 与 RPA 结合，实现现有软件系统的重构方法，并针对存在的技术问题提供了相应的解决方案。

海通证券的《海通证券虚拟数智人的应用实践与探索》基于元宇宙理念研发推出虚拟数字员工，集成人物渲染、智能对话引擎、智能语音识别、智能语义理解、多模态智能交互等多项 AI 技术，实现科技赋能，丰富金融服务模式。

华东师范大学等的《基于大模型技术的网络黑嘴识别》响应国家网络黑嘴专项整治号召，针对自媒体违规发布财经新闻、歪曲解读经济政策等行为，提出了一套基于大模型技术的网络黑嘴识别系统框架，探索以 ChatGLM2、Baichuan2 等中文大模型作为基座，并采用不同的方法微调的“网络黑嘴”识别模型。

上海金仕达软件的《金融合规大模型的研究与实践》聚焦于业务合规管理，介绍了基于法律法规库的智能合规咨询服务、智能合规监测系统等金融合规大模型的实践案例。

北京邮电大学的《基于大模型的金融招股书智能审批系统设计与实践》从证券公司信披文档审核的痛点问题出发，将招股书智能审批问题分解为勾稽关系校验和合规性审核两个任务，并利用元素提取、语义分割、自回归大模型等技术探索解决方案。

《交易技术前沿》编辑部

2023 年 12 月 14 日

目录 Contents

大模型与人工智能整体战略

大模型与人工智能运营探索

大模型驱动业务创新

基础运算技术研究

信息资讯采撷

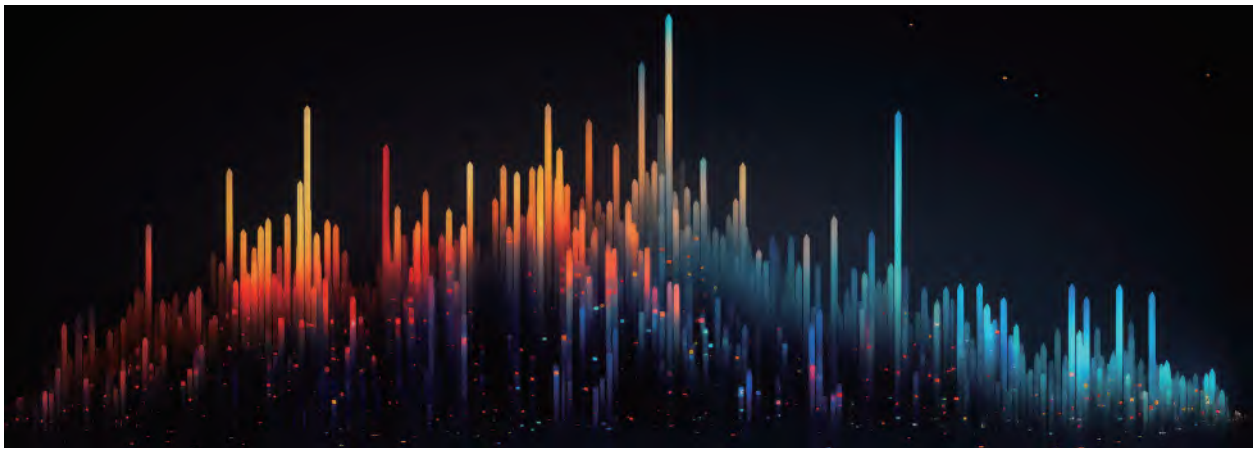
大模型与人工智能整体战略

1 证券行业大语言模型的探索与实践

2 上证信息智能运营体系的规划与建设

证券行业大语言模型的探索与实践

王洪涛[1]、熊友根[2]、周轶珅[3]、李双宏[3]、李增鹏[3]/
[1]国金证券股份有限公司 首席信息官 上海 201204
[2]国金证券股份有限公司 科技研发部 上海 201206
[3]国金证券股份有限公司 信息技术部 上海 201204
Email：lishuanghong@gjzq.com.cn



在证券行业中，IT 系统的稳定性与前沿技术的快速应用之间存在着天然的矛盾。为了平衡稳态和敏态之间的差异问题，本文探讨了证券行业采用大模型与现有系统结合的新型模式，包括 AI 外挂式、AI 内嵌式和 AI 原生式三种结合模式。提出了通过将 AIGC 与 RPA 结合，实现现有软件系统的重构方法。本文探索了基于开源预训练大语言模型结合本地数据进行本地化训练微调的技术路线，并针对存在的技术问题提供了相应的解决方案，为证券公司的大模型本地化提供了实践与参考。国金证券以 AI 助手为切入点，探索大语言模型提升工作效率赋能业务发展，积极推动大模型与现有系统分级耦合并成为 AI 中台建设的突破口，为证券行业人工智能建设提供启示和借鉴。

关键词：证券行业；大语言模型；AIGC；AI 中台

1 引言

国务院《新一代人工智能发展规划》指出抢抓人工智能发展带来的重大战略机遇，打造我国人工智能发展的先发优势，加速推进建设创新型国家的进程。在数据不断累积、算法不断突破、算力不断提高的背景下，人工智能生成内容（Artificial Intelligence Generated Content，AIGC）技术迎来迅猛发展，正在催生全新的产业体系，将为这一目标的实现提供强有力的支撑。现阶段

AIGC 的发展属于深度融合阶段。在这个阶段，AIGC 技术和其他技术和领域深度融合，发挥出 1+1 大于 2 的价值。这种融合可以使 AIGC 技术更加全面和智能，实现更加复杂和高级的内容生成。

作为 AIGC 在文本生成领域的里程碑式应用，OpenAI 公司发布的对话式语言大模型 ChatGPT 能够模拟人类的语言行为与用户进行自然、流畅的交互，执行语义理解、内容生成、情感分析等多种任务[1]。大模型作为凝练海量数据内在精华的“隐式知识库”，是一种学习能力强、泛化能力强、能够解决传统难处理的复杂任务的人工智能应用载体，结合了大数据、高算力和强算法的优势。快速发展的大模型为 AIGC 赋能各行各业的发展提供了新动力。

证券行业有多元化的业务场景和广泛的数字化转型升级需求，是大模型极佳的垂直落地场景。通过快速、准确地进行信息整合及自动化任务处理，大模型能够进一步推动金融行业数字化转型。然而，证券行业对技术可靠性和稳定性的要求非常严格，这是因为证券交易涉及大量资金和客户敏感信息，任何系统故障或数据泄露都可能导致严重的后果。AIGC 技术的不成熟导致证券行业在应用 AIGC 技术时，面临与行业特点相关的风险和问题。本文着重解决证券 IT 系统的稳定性与前沿技术的快速发展之间的矛盾，提出了一种大模型与证券 IT 系统分级耦合的技术理念，并探索利用 AIGC+RPA（Robotic Process Automation）的技术路线来改造现有软件系统。结合现阶段开源大模型技术框架，提出一种适合证券公司自研大模型的技术路线，并进行国金证券大模型探索与实践的案例展示，为同行提供技术路线和推广模式的借鉴。

2 大模型在证券行业的应用

证券行业大模型建设作为新兴事物，没有太多垂直行业的经验可参考，我们提出证券行业大模型建设思路如（图 1）。

大模型与现有系统的结合模式。证券 IT 系统必须高度稳定可靠，而大模型的发展是日新月异的，二者对频繁变化的容忍程度存在巨大差异。因此，如图 1 所示，大模型与现有系统的集成应遵循**“由松至紧逐步耦合”**的原则，采用外挂式、嵌入式和原生大模型相结合的方式来进行大模型的应用。外挂式是指用大模型的交互接口，把任务分发给相应的 IT 系统进行数据查询、结果反馈，再利用大模型进行内容的整合；内嵌式是指在 IT 系统嵌入对话式接口来调用大模型，IT 系统的功

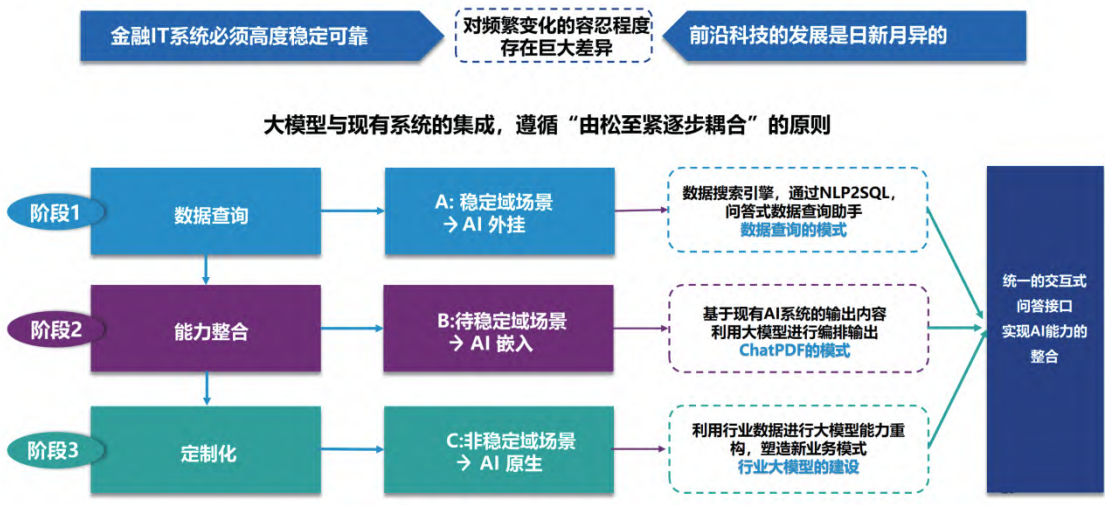


图 1：大模型与现有系统的结合模式

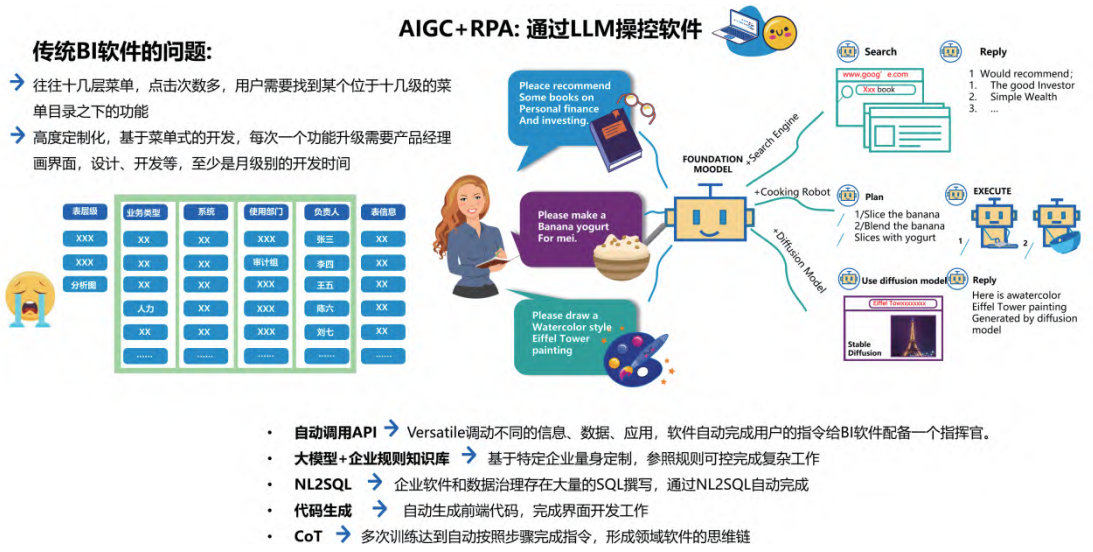


图 2：大模型重构软件系统

能可通过大模型交互式问答的方式调用，让对该系统不熟悉的人员也可以快速高效的使用；AI 原生是指采用行业数据和信息，进行大模型的微调，使大模型具有该行业的属性。

AIGC+RPA 重构软件系统。RPA 像强有力的机械手臂，可以自动执行任务，提供高效的工作辅助。AIGC 是机械大脑，具备分析、整合、创造的能力，可利用智能算法和模型生成全新的内容。如图 2 所示，将 RPA 和 AIGC 结合，让机械手臂的动作与机械大脑的思考相互配合，可以自动化处理重复和繁琐的任务，这有助于证券公司提供更智能的服务。

可兼容信创 GPU 的弹性算力池化。为充分利用现有 GPU 算力支持大模型推理，如图 3 所示，将 GPU 资源池化能力扩展到整个数据中心，解耦 AI 应用和 GPU 服务器，实现弹性调度 vGPU 资源。该方案兼容英伟达 CUDA 生态和国产 GPU 体系 NeuWare，ROCm，能建设异构的 GPU 算力集群。基于国产 GPU 生态体系，可以构建信创大模型和算力资源池。

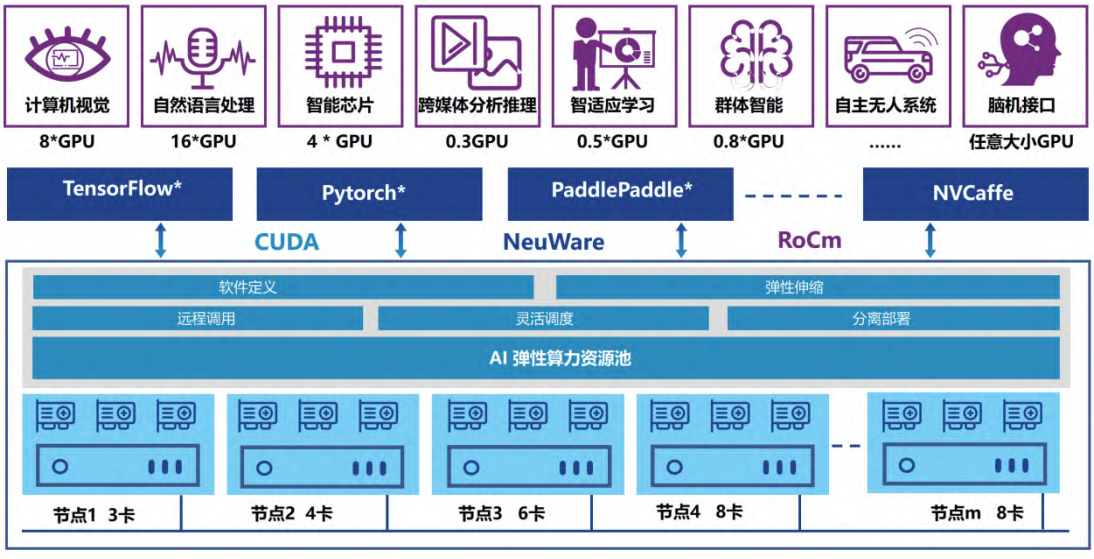


图 3：GPU 算力池化方案

3 证券公司大模型建设技术方案探索

证券公司的大模型规划应以业务人员和客户的需求为导向，用大模型赋能证券业务的发展，着重拓展以大语言模型为核心的交互式问答场景挖掘。

3.1 技术路线 1：预训练通用大模型

利用大规模数据集和语料库训练出的垂直领域的大模型具有丰富的专业领域知识和强大的文本生成能力。例如，彭博社基于自主构建的 FinPile 数据集训练的 BloombergGPT 大模型已经被证明能够比通用大模型更好地处理金融领域的数据和任务。

由于大模型的参数量通常介于几十亿到上千亿之间，其预训练需要大量的 GPU 算力资源作为支撑（ChatGPT 每次训练的成本超百万美元），这对于大部分证券公司而言难以负担。

3.2 技术路线 2：基于通用大模型进行参数微调

参数高效微调是性价比相对较高的强化大模型行业领域知识的手段，主流方法包括 P-Tuning、LoRA 等。参数高效微调的主要思想是在保持原有大模型全部或绝大部分参数不变的基础上，通过增加或改变少量参数的方式改善模型在特定任务上的性能，受影响的参数量通常仅为全量参数的 0.1% 左右 [2]。

大模型参数微调的训练成本适中，在行业场景对话效果良好。然而，该方法可能面临灾难性遗忘的问题，即模型在学习新任务或适应新环境后会忘记甚至是丧失以前习得的知识，造成模型在原有任务或环境下性能大幅下降，一定程度上限制了大模型在证券行业的应用。

3.3 技术路线 3：基于外部知识库和提示工程对通用大模型调优

提示工程通过构建和优化输入来引导大语言模型生成更加精确、可靠和符合预期的输出文本，能够有效激发通用模型的迁移学习能力，从而提升在目标领域的应用效果，为用户提供更佳体验。

针对证券行业的特点，选择基于外部知识库和提示工程对通用大模型调优技术方案最合适。首先可以构建包含专业术语、产品信息、股票行

图 4：预训练垂类大模型

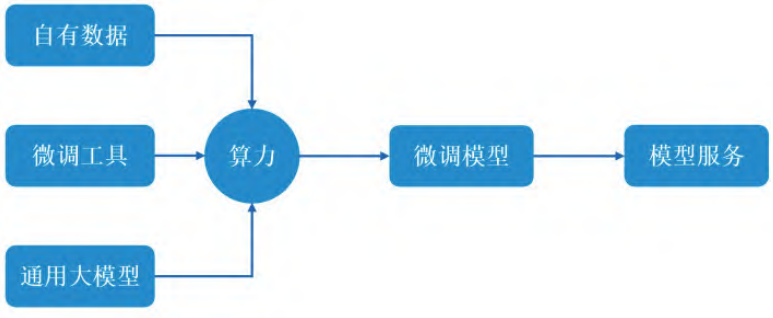


图 5：通用大模型进行参数微调

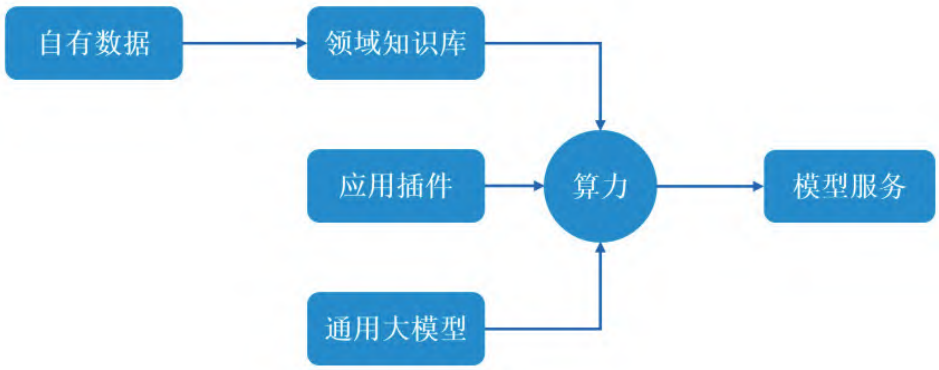


图 6：基于外部知识库和提示工程对通用大模型调优

情等结构化信息的外部知识库，然后通过知识库检索以添加融合了证券知识的提示对通用大模型进行调优，使大模型学习证券领域的语言表达方式以及业务逻辑知识，从而展现出更好的解决证券问题的能力。该方法在保留良好对话效果的同时，其训练成本非常低，能够有效避免模型微调后的灾难性遗忘问题以及减少模型出现事实错误（幻觉）的情况。

3.4 通用大模型本地化训练与调优

大模型的基础能力维度包括查询、整合和推理，如表 1 所示。为探索何种技术路线更有助于证券大模型的快速落地，对多个代表性大模型进行了本地化部署以探索其适用场景及能力边界。

Llama2 是 Meta 公司发布的大模型，相较其上一代训练数据增加了 40%，在包括推理、编码和知识测试在内的各基准测试中性能优越。美中不足的是 Llama2 的训练数据集英文语料占 89.7%，而中文语料仅占 0.13%，导致其中文能力短板明显，难以完成流畅、有深度的中文对话，这与本地化部署结果一致。

GitHub 社区基于大规模中文数据集和 LoRA 方法对 Llama2 进行微调，得到了中文对话能力增强的 Llama2-Chinese 大模型。然而，本地化部署结果显示，LLaMA2-Chinese 出现了明显的灾难性遗忘现象，存在答非所问、中英文夹杂甚至是乱码的情况。

InternLM，即书生 · 浦语大模型，由商汤科技和上海 AI 实验室发布，其预训练集针对中英文构建，来源包括网页、书籍、百科、学术论文、代码等。经过本地化部署及测试，InternLM 的综合性能优异，但是缺乏对金融领域的深入了解，难以准确捕捉金融领域的语义信息。

ChatGLM2 是智谱 AI 及清华 KEG 实验室发布的大模型，在保留初代模型对话流畅、部署门槛较低等特性的基础之上，性能更强大，推理速度提升 42%。ChatGLM2 专门针对中文问答和对话进行了优化，测试结果表明其解决通用问题的能力良好，但在金融领域的表现仍有待提高。

ChatLaw 是北大和兔展 AIGC 联合实验室发布的中文法律大模型，能够为证券从业人员提供普惠的法律服务以满足合规要求。ChatLaw 基于判例文书、法律法规和地方政策，并辅以资深律师的人工标注构建了大规模法律数据集以进行预训练。本地化部署结果显示其在法律任务上的表现不尽如人意，并未大幅优于通用大模型。出现这一现象可能的原因是：其基座为 Llama，对中文的支持一般；因模型微调而导致的灾难性遗忘问题；没有基于法律知识库进行法条检索并生成提示词以提高模型性能。

LangChain 是一款帮助开发者利用语言模型构建端到端应用的框架。它提供了一套能简化由大型语言模型和聊天模型提供支持的应用程序的创建过程的工具、组件和接口，可以轻松管理与

语言模型的交互、连结多个组件、整合 API 和向量数据库等额外资源。

3.5 基于知识库微调预训练大模型技术方案的问题及解决方案

针对证券行业技术的特点，我们选择 LangChain+ 通用大模型结合的技术路线进行模型的微调和领域知识库的训练，但是该技术方案仍存在一些问题：

文档内部的表格和非结构化数据的解析，极大地影响知识库的效率。当前方案使用 OCR 解析 PDF 文件，对表格、图标等非结构化数据的解析能力不足，后续可将包含行列结构的 Markdown 文档提供给大模型以实现更精确的解析。

Embedding 模型影响知识库的构建。Embedding 模型的作用是用连续的、低维的向量表示离散的符号或词汇。Embedding 模型提供文档的切词、向量计算等功能，如果该模型的能力不够精细，将极大影响大模型检索问题答案的能力。目前，LangChain 框架使用的名为 Text2Vec 的 Embedding 模型暂无法满足高精度的问答需求。

多重复内容的文档构建的知识库对大模型起到反作用。多文档间相互有影响，可能导致特定领域知识被稀释，出现训练越多，精度越差的情况，因此高质量、无重复、无矛盾的文档集非常重要。为了解决该问题，可以对文档分类构建不同的专业知识库，并限制用户提问的细分领域。

全链路大模型成本高且推理能力有限。未来会是大小模型并存，大模型部署的资源需求高，面向资源不足的实际场景就需要提升小模型。此外，现有实验也表明专业领域小模型在特定任务上可以比大模型的相关模块精度更高。

我们认为建设证券大模型时，外部知识库和提示工程的技术路线不仅可行且有更好的性价比。自研基模型预训练成本高昂，数据集构建难度大，大部分企业难以负担。模型微调可能导致灾难性遗忘，企业内部专业数据总体体量有限，且算力要求不容忽视。相较之下，利用外部知识库和应用插件的算力要求低，能使大模型对证券领域的语义理解更为准确，减轻其幻觉问题并有效提高对话能力。

为了解决上述问题，如图 7 和图 8 所示，我们提出以下两种技术改进方案以供证券公司在未来应用大模型加快数字化转型进程：

1、大模型内部功能模块替换为高精度的商用模块，并挂载相关业务知识库，形成**业务专用大模型**；

2、串流式替换，加强文档的解析能力，利

表 1：大模型基础能力维度

维度	能力	示例
查询	知识问答	什么是证券公司次级债务?
整合	内容生成	撰写一份国金指数的广告文案
	语义理解	某人夸奖对方办事能力强，对方回答“哪里”。这里的“哪里”是什么意思?
	代码编程	写一个从 User 表提取 Age 字段大于 20 的所有数据的 SQL 语句
推理	逻辑推理	小明的老婆诞下双胞胎。以下哪个推论是正确的? A、小明家有两个孩子。B. 小明家既有男孩也有女孩。
	数学能力	鸡兔同笼问题

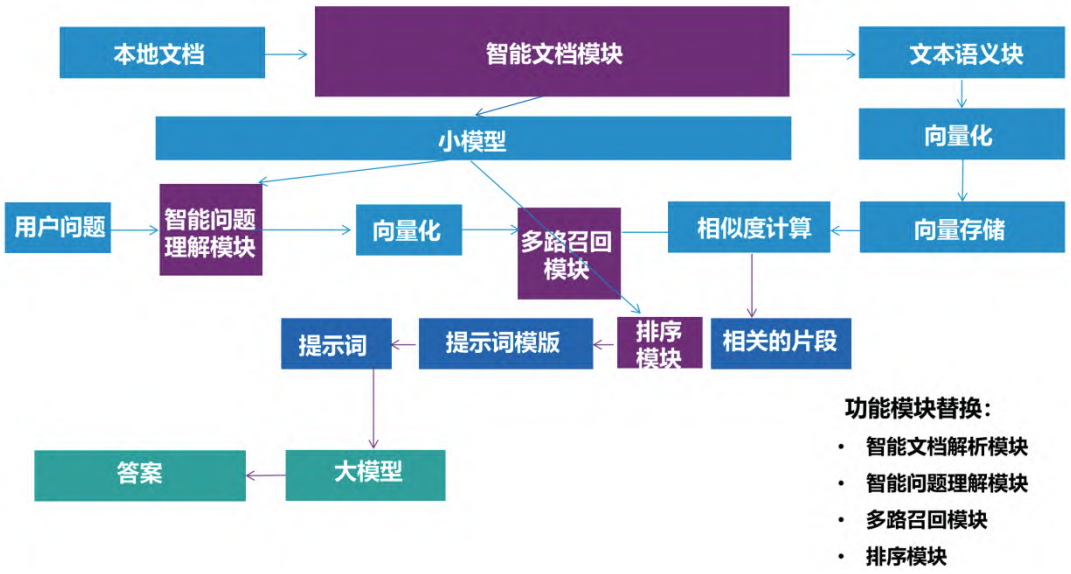


图 7：业务专用大模型技术改进方案

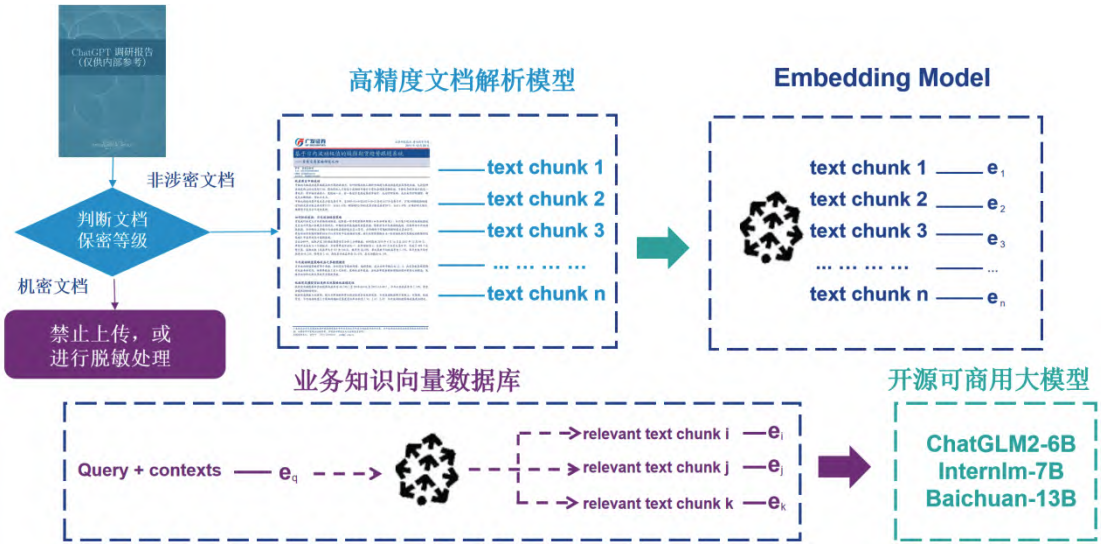


图 8：文档对话大模型技术改进方案

用大模型进行编排整合，形成**文档对话式大模型**。

同时，可以结合 NLP2SQL 等技术路线，整合数据库查询功能，提升大模型的问答和数据图表绘制的能力。

4 国金证券大模型建设的探索与实践

4.1 建设思路

国金证券紧扣新五年科技战略发展规划“融合业务，平台赋能，打造一流券商科技组织作为科技愿景和目标”的定位，积极探索大模型建设。国金证券的 AI 的规划会以业务需求和客户需求为导向，用 AI 赋能证券业务的发展。主要着重拓展两方面的 AIGC 应用：一方面是以 AIGC、大语言模型为核心的交互式问答场景挖掘，另一方面是证券交易风控领域的 AIGC 算法探索，通过外挂式和嵌入式的模式，与交互式大语言模型接口进行整合。

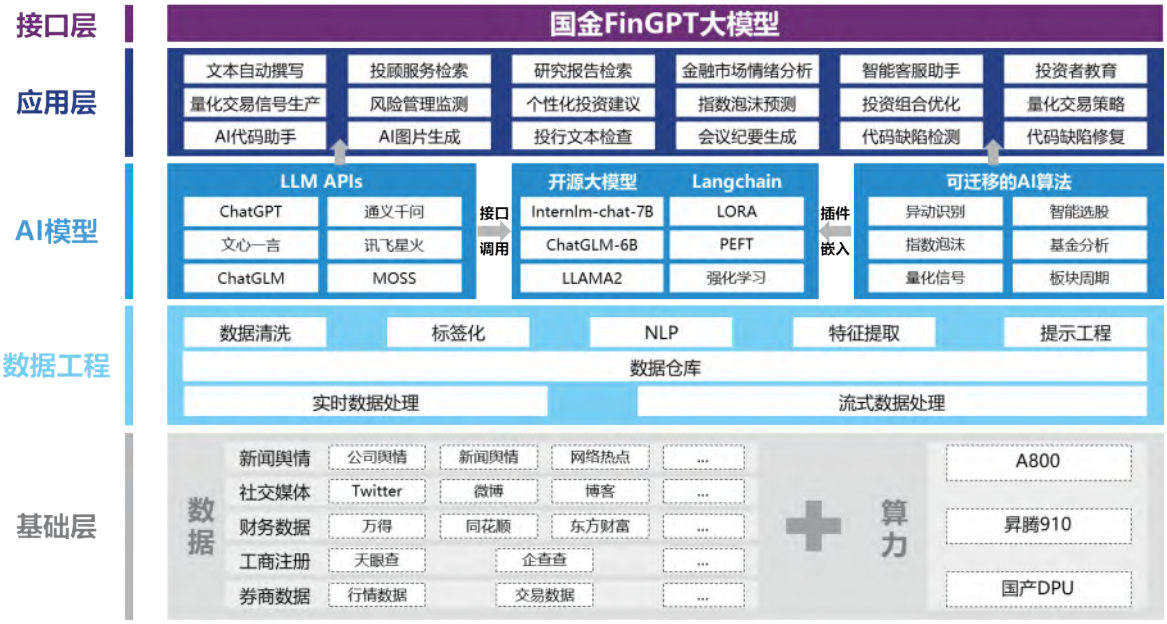


图 9：国金证券大模型架构规划

图 9 展示了国金证券大模型的架构规划。在 AI 中台，大模型的定位是智能化业务辅助应用的入口，形成统一的 AI 大模型入口，来赋能不同的公司人员。在证券业务应用系统中基于大模型带来的 AI 能力，将形成各式各样的智能化应用入口对 AI 中台做向下赋能。AI 中台将更聚焦于应用智能化实现，做向上业务的支撑。同时，采用外部大模型 API 与内部本地化大模型的相互配合，并进行相关数据权限的内部审核和分级管控，从而实现工作效率的提升。

4.2 国金证券大模型实践成果

基于对不同大模型和技术路线的探索，我们选用 LangChain 和 ChatGLM2 构建了国金证券大

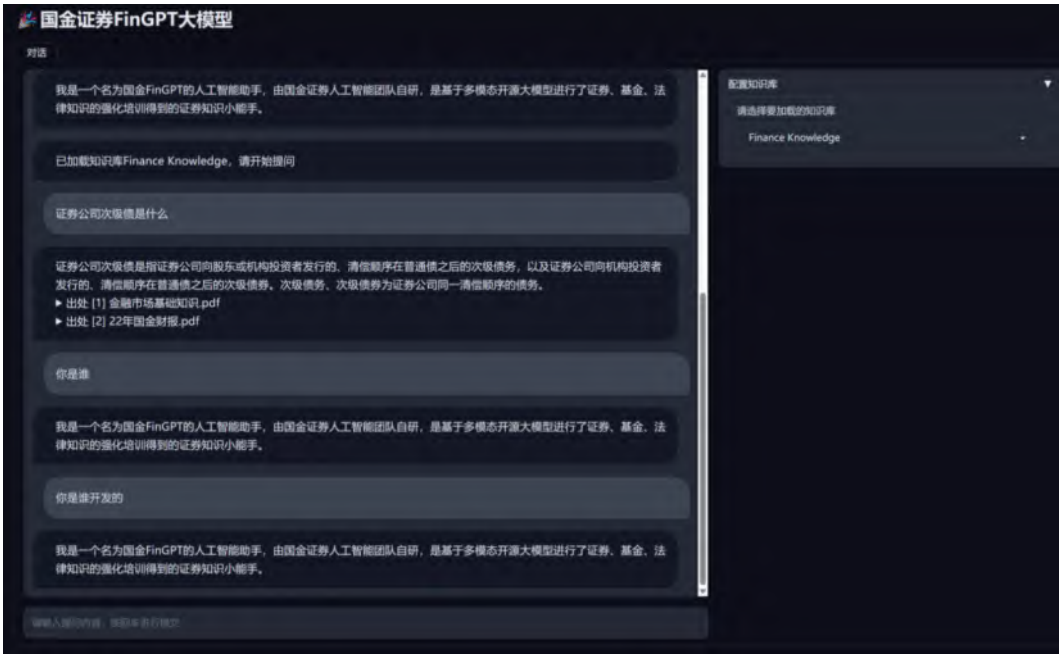


图 10：国金证券大模型对话界面

模型，图 10 展示了系统对话界面，可以看出通过知识库检索大模型能够准确回答证券相关的专业问题。通过挂载不同的知识库，可赋能不同的业务场景。

LangChain 是一种集成自动提示工程的框架，支持多种大模型的接入。LangChain 的实现原理及文本处理过程如图 11-13 所示，主要步骤如下：

加载本地知识文件。通过非结构化加载器读取文本，目前支持加载 pdf、txt、md、docx 等格式的文件。

文本切分。根据预定义规则或语义，将大型文本拆分成较小的文本块。例如，可以根据中文文章常见的中止符号进行文本切分。

构建知识向量库。基于 Embedding 模型将切分好的文本块向量化，并存储在数据库中以供知识检索。

用户请求向量化及检索匹配。通过检索知识库，找出与用户请求向量最相似的前 k 个文本向量，然后并添加上下文，通过模板自动生成提示。

大语言模型响应。将用户请求和提示一起作为输入提交给大模型生成回答。基座大模型应具有优秀的泛化能力和适当地参数量，此处我们选用 ChatGLM2。

4.3 AI 助手作为大模型的抓手

基于现阶段相关技术成熟度，在证券行业中 AIGC 和大模型的探索场景仍应主要集中在企业内部或企业机构客户，而非零售或大众客户。这是因为这些企业通常拥有更多的数据资源和更复杂的业务需求，需要更高效、准确的技术支持来提升其业务水平和竞争力。同时，企业内部用户全本地化大模型可以很好地控制使用权限，防止内部数据隐私泄露。虽然 ToC 市场也在逐渐增长，但由于涉及个人隐私和数据安全等问题，ToC 市场的 AI 应用面临很多挑战和限制。因此，目前大多数 AIGC 公司在探索应用场景时会以 ToB 市场为主要目标客户群体。

国金证券将内部办公场景作为切入点是比

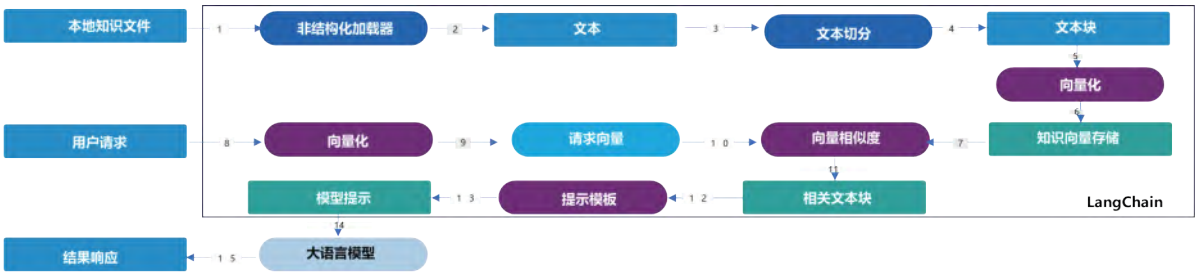


图 11：知识库与大模型耦合基本原理

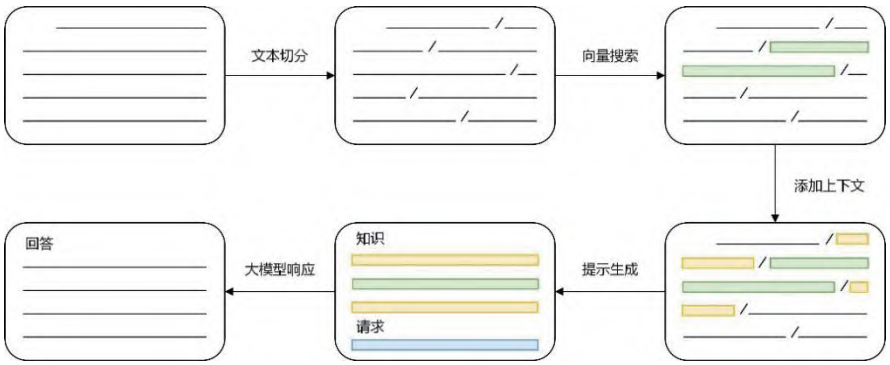


图 12：大模型 + 知识库文本处理过程

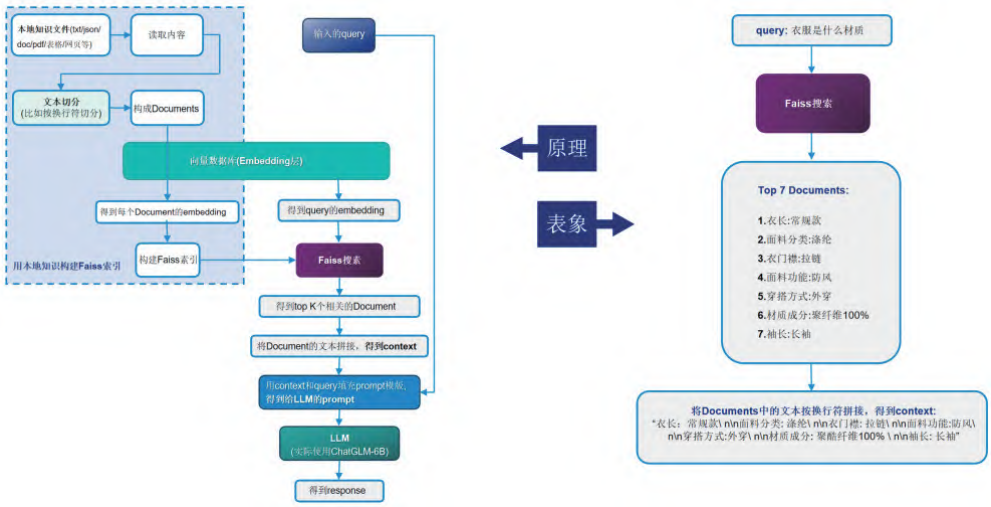


图 13：基于 Query 和 Context 让大模型得到高质量回答

较合适的选择。通过办公场景逐步沉淀 AI 能力，逐步拓展应用场景，同时也可利用 AI 办公助手来逐步收集业务部门的需求，为后续大规模应用 AIGC 做技术准备和 AI 赋能业务的铺垫。如图 14 所示，国金证券建设的 AI 办公助手采用外部 API 和内部自研大模型相结合的模式，通用问题可以采用外部 API 问答，针对内部数据采用本地化服务，可以提升公司员工的办公效率。

该助手具有以下特点：

角色定制：AI 办公助手可以根据员工的角色进行定制，例如，对于销售部门的助手，可以定制销售数据分析辅助，客户沟通辅助等相关功能。

模型选择：AI 办公助手可以根据不同的任

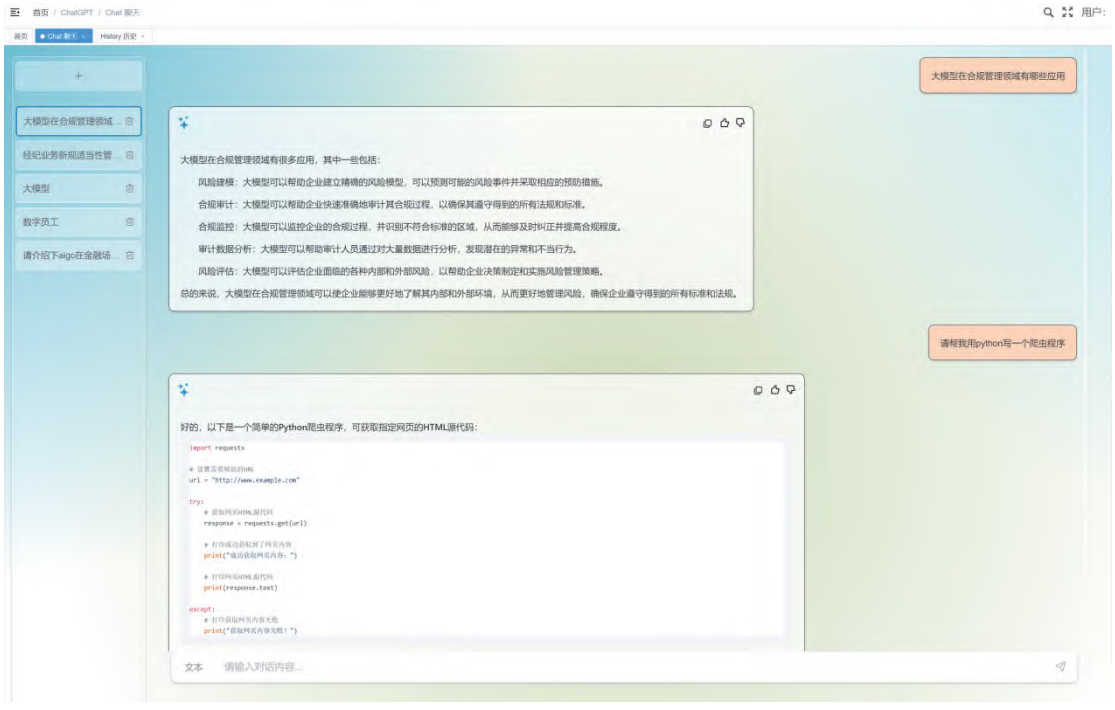


图 14：AI 助手对话界面

务和场景选择不同的模型进行调整。

敏感过滤：AI 办公助手可以过滤掉敏感信息，保护公司的数据和隐私。

权限管理：AI 办公助手可以进行权限管理，确保每个员工只能访问其需要的数据和功能。例如，在数据查询中，助手可以根据员工的角色和职务进行数据权限控制，避免数据泄露。

AI 助手作为 AI 中台的核心组成部分，为公司内外提供全方位的 AI 技术支持和服务。用户可以轻松访问和使用各种 AI 算法和大语言模型，实现智能化的办公处理和管理，极大地提高公司的办公效率。

外挂式 AI 算法的推广将推动 AI 技术的广泛应用与落地。外挂式算法可根据具体业务需求进行定制化开发。这种灵活性使得外挂式算法能够更好地适应不同业务的特点，实现更精准和高效的解决方案。

AI 需求汇总的媒介将成为公司员工获取 AI 技术和知识的重要途径。通过该媒介，员工能及时了解最新的 AI 研究成果和应用案例，并反馈自己的使用心得、建议，帮助开发者进行 AIGC 的场景探索和系统优化，同时员工也可以提升自己的 AI 应用水平，逐步培养 AI 思维和数字化思维。这有助于培养公司内部的数字化思维和创新氛围，推动企业数字化转型和发展。

培育 AI 文化和建设 AI 生态系统是推动公司数字化转型的重要举措。建设完善的 AI 生态系统可以为公司提供更多的业务模式创新，促进各公司各业务条线的协同创新和共同发展。

5 总结与展望

证券行业对 IT 的稳定性的要求与前沿技术的日新月异发展之间，存在对 IT 系统频繁调整的天然矛盾。本文探索了一种证券行业采用大模型的结合模式，分为 AI 外挂式、AI 内嵌式和 AI 原生式三种结合模式；并提出将 AIGC 与 RPA 结合，重构现有软件系统的技术设想。本文探索了基于开源可商用大模型结合本地数据进行模型训练的技术路线，存在的问题以及解决方案，为证券公司的大模型本地化进行了探索与实践。国金证券将以大模型在公司内部员工助手中的应用为突破口，将大模型建设作为 AI 中台建设的切入点，积极推动人工智能在提升内部工作效率和新业务模式拓展方面的作用，力争在创新与发展中实现突破。

参考文献：

[1] 桑基韬 , 于剑 . 从 ChatGPT 看 AI 未来趋势和挑战 [J]. 计算机研究与发展 ,2023,60(06):1191-1201.

[2] 天翼智库 . 迎接大模型时代：大模型发展简史及攻略 [J]. 互联网天地 ,2023(05):8-15.

上证信息智能运营体系的规划与建设

王波、张晓军、孙志峰、何帅兵、田秋实、马强 / 上证所信息网络有限公司 基础架构部 上海 200120
E-mail：bwang@sse.com.cn

随着新技术的不断应用，特别是人工智能技术的不断发展，加速了数字化时代的演进过程，也加速了金融行业的数字化转型。在数字化转型过程中，智能化运维有望得以落地，进一步提升运维能力，成为当前探索和实践的方向。

上证所信息网络有限公司（以下简称“上证信息”），在交易所数字化转型过程中，通过智能化运维的解决方案，打造了一个覆盖基础设施到业务系统的 IT 智能运营体系，统一纳管运营数据资源以实现数据资产标准化，实现了高频场景的批量处理以满足自动化的要求，构建运营类数据采集、计算和存储的数据一体化框架，实现异常检测、告警收敛智能化场景，建立起适用于当下金融科技发展模式的数字化产品运营体系和治理结构。

关键词：人工智能；数字化转型；智能化运维；智能运营体系

1 背景及意义

上证信息以 " 创造卓越、为证券市场提供一流信息服务 " 为宗旨，肩负着打造信息产业链和提供市场基础设施的重任。目前上证信息负责运营 Level-2 商业行情、上证云、上交所官网及 APP、上证路演中心等多个明星产品，上证云是其中运营规模最大的一个产品。

1.1 上证云简介

上证云是面向证券、基金、监管机构、核心机构等金融机构推出的云计算服务，依托全国 20 个数据中心节点，通过先进的混合云架构以及行业场景导向的产品设计，具备完善的用户服务体系及丰富的安全运营管理经验，严格遵循国家相关部门监管政策，打造云链网一体化基础设施，为金融机构提供技术领先、稳定可靠、安全合规

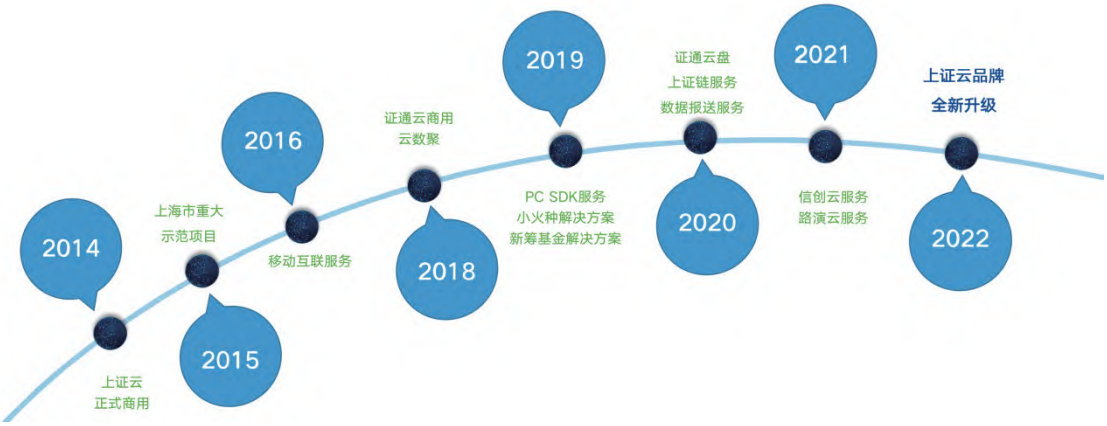


图 1：上证云发展历程

的云计算服务。

1.2 挑战与难题

随着公司业务的发展，IT 运维团队面临了越来越多的挑战，主要体现在：

1) 从技术系统的规模上来看，有 24 个技术系统，100 多个子系统，600 多个模块。

2) 从基础设施的规模上来看，有 20 个机房，6000 多台主机，800 多台网络设备

3) 全年发生的上线发布、版本升级和配置变更的台次超过 4 万次。

4) 面对监管压力、市场舆论和投资者投诉，零容忍的安全运行压力。

面对这样的运维规模和监管压力，带来了一些急需解决的难题，比如：

1) 如何尽早的从用户或者业务角度发现故障，在基础监控的基础上，提升业务运行监控。

2) 如何更快的定位故障问题，出现故障后，更快的确定这个故障是网络问题，还是系统问题，还是中间件、应用的问题

3) 如何更好的管理硬件资源，准确无误的摸清家底，有效的对分配的资源进行回收、再利用，或者进行替换

4) 如何更快的交付资源和交付应用，满足 SLA 的承诺，满足用户的期望。

1.3 智能化运维

2016 年，Gartner 首次提出了智能运维概念。在 Gartner 的《Market Guide for AIOps Platforms》报告中对智能运维（AIOps）做出了如下定义，AIOps 平台是结合大数据、人工智能（AI）或机器学习功能的软件系统，对数字化转型过程中 IT 系统不断产生的数据量、数据类型，进行采集和分析，用以增强和部分取代广泛应用的现有 IT 运维流程和事务，包括可用性和性能监控、事件关联和分析、IT 服务管理以及运维自动化。

从 Gartner 的定义来看，智能运维是传统运维体系和大数据、人工智能技术结合的产物。AIOps 将人工智能应用于运维领域，基于已有的运维数据，通过机器学习的方式来进一步解决自动化运维没办法解决的问题。如果要确保业务的可用性、敏捷性，必须以 IT 环境的全量数据汇聚能力为基础，整合每个单独系统所提供的数据服务能力，降低多个系统的运维复杂程度。

AIOps 围绕质量保障、成本管理和效率提升的基本运维场景，逐步构建智能化运维场景。在质量保障方面，保障现网稳定运行细分为异常检测、故障诊断、故障预测、故障自愈等基本场景；在成本管理方面，细分为指标监控，异常检测，资源优化，容量规划，性能优化等基本场景；在效率方面，分为智能预测，智能变更、智能问答，

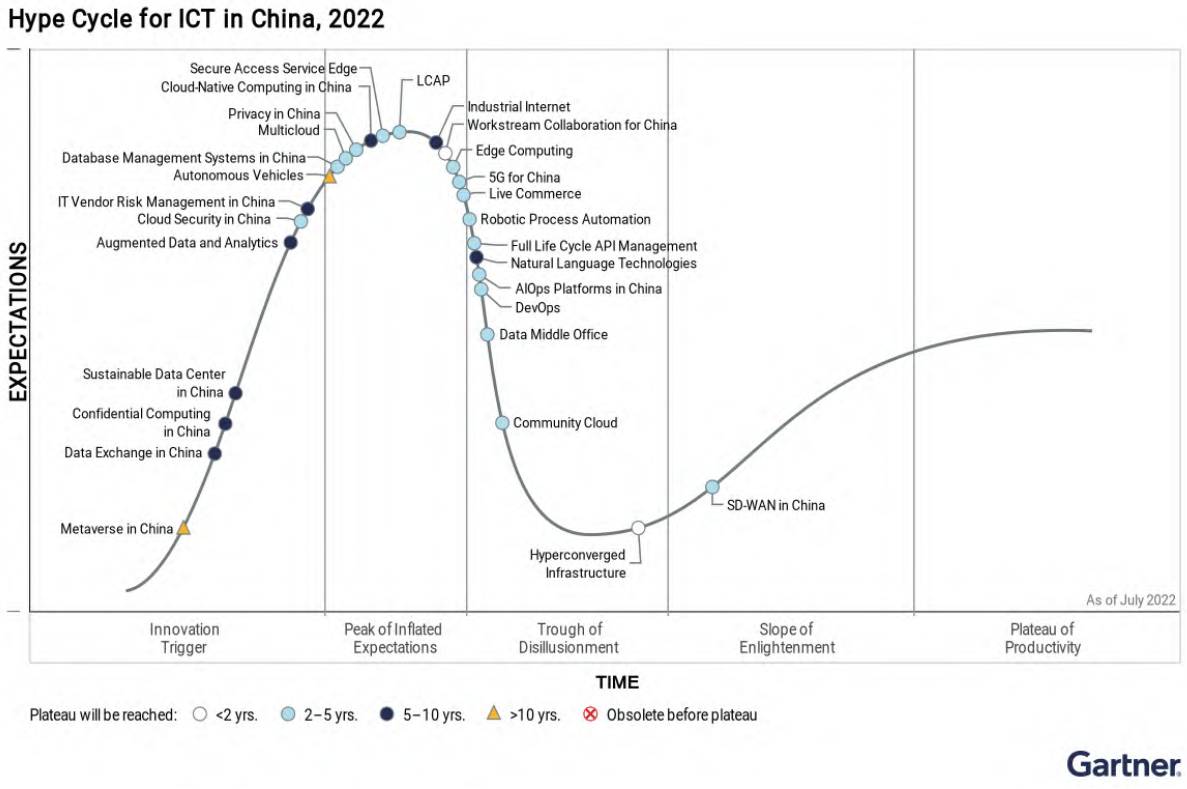


图 2：2022 年 ICT 成熟度曲线

智能决策等基本场景。

借鉴 AIOps 的理念，为了应对公司运营的挑战和难点，上证信息提出了从以下 4 个方面进行能力的提升：

1）提升变更效率和质量。通过自动化发布平台，提升自动化发布能力、应急回滚能力、机器灰度能力

2）增强异常监测和告警。推动基础设施层、应用层、业务层监控能力的建设；推动集中监测展示的能力建设

3）增强应急事件的处理。结合事件，及时更新应急预案，并定期进行演练；提高应急操作能力，提升应急操作效率。

4）提升故障的根因定位。结合应用 CMDB 中配置信息，确定应用的拓扑关系，落定 AIOps 应用场景，形成智能化根因分析能力

图 3：典型的智能化场景

2 智能运营体系的整体规划

“规划先行、谋定后动”，2018 年，上证信息启动了智能运营体系的建设规划，按照“标准化”、“自动化”、“一体化”和“智能化”的建设路线逐步实施，先从传统运维向技术运营转型，最终实现智能运营。

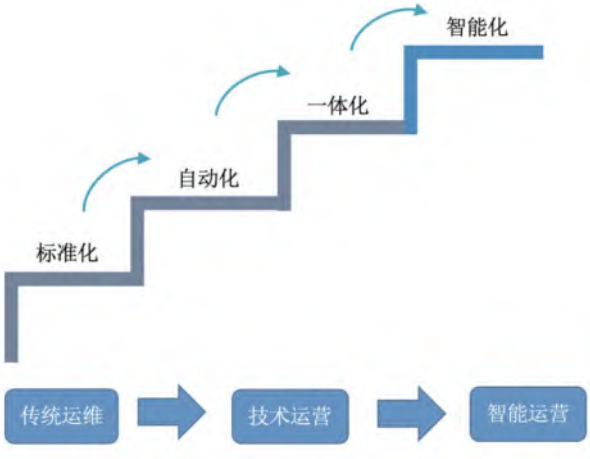


图 4：智能运营体系的规划

依托 AIOps 的相关理念，结合当前金融科技中最前沿的大数据、人工智能、容器化技术、分布式消息队列、微服务架构等相关技术，并结合证券行业的特点和公司的现状，设计了智能运营体系的整体架构。使用 CMDB 统一纳管运营数据资源以实现数据资产标准化，实现了高频场景的批量处理以满足自动化的要求，构建运营类数据采集、计算和存储的数据一体化框架，并通过 AI 大脑实现异常检测、告警收敛智能化场景，实现集成监控告警、运维流程、可视化大屏等功能的一体化运营门户。

3 相关的建设方案

3.1 标准化

标准化是自动化的基础，全方位建立标准化，有助于快速实现场景的自动化，尽可能少地处理各种个性化的情景，减少自动化难度。

在智能运营体系的架构中，标准化主要围绕两部分进行建设，一个是制定和发布了一系列的标准和规范，比如：监控告警规范、日志规范、部署规范等，另外一个是配置管理的建设，明确规定了 CMDB 作为唯一权威的配置数据源，将运维的配置信息按照标准的模型存入 CMDB。

配置管理是 IT 服务管理的核心流程，为各

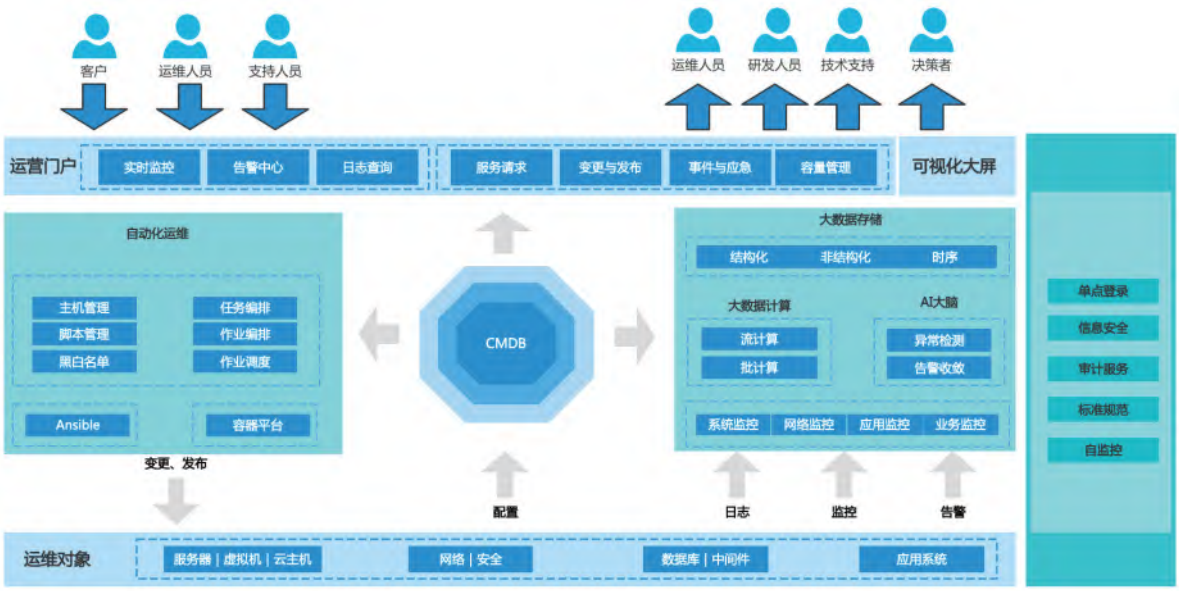


图 5：智能运营体系的整体架构

项运维工作提供准确、一致、规范的配置数据，具有重要意义，能够提升整体运维 / 运营的效率。CMDB 是配置管理的主要支撑工具，用于存储和管理配置数据，，为各项运维工作提供准确、一致、规范的配置数据，是实现 IT 基础设施“对象数字化”的基础，也是支撑 IT 服务流程和运维活动的“行为数字化”的前置条件。

CMDB 的整个建设过程中，始终以消费场景驱动，为 IT 服务流程、IT 运维管理的工作提供交付。首先，从技术需求和管理需求方面挖掘和收集价值场景，进一步分析各类场景所需的数据，以及结合各类场景针对数据的使用条件和依赖关系，形成模型设计的核心四要素，即：“模型对象”、“模型分组”、“模型属性”以及“模型关联”，并使用人工流程和自动采集的方式，作为提升配置数据质量的保障手段。

3.2 自动化

自动化的目标是，将日常操作频率高且具备通用性的人工操作流程，通过建设自动化运维平台，实现此部分流程的自动化和界面化，提升变更效率的同时，有效降低人为风险和人力成本。自动化运维过程中，还将数据标准化的工作纳入自动化范畴，配置数据的变更主要依靠自动化触发，少量人工变更的部分依靠流程的强执行力保证。

基于开源工具 ansible，自主研发了一套界面友好、操作简单、功能丰富、易扩展、易于脚本编制和生成的运维操作平台，能够满足现有技术系统的运维需求。通过自动化运维平台，使网络运维、系统运维、应用运维等角色可以在其权限

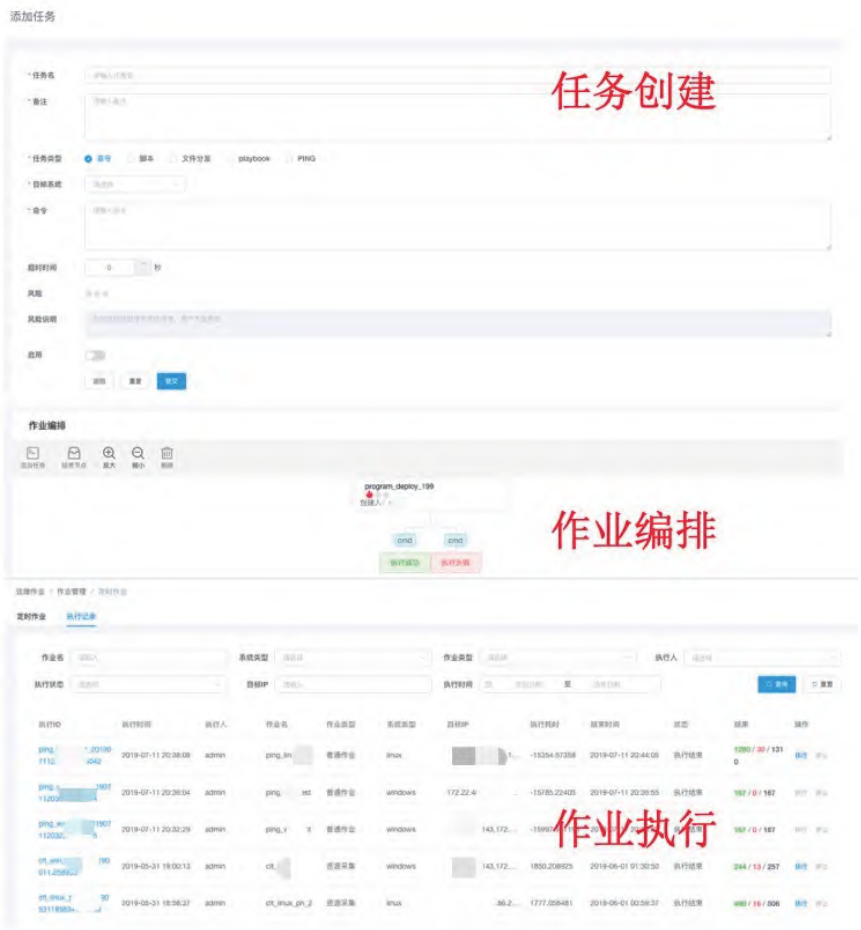


图 7：自动化运维操作流程

下完成日常运维工作中对应的具体操作，如应用运维人员完成应用部署或版本升级工作、系统管理人员完成主机的磁盘清理和网络管理人员完成 NAT 映射或开关白名单工作。

从上图可知，运维人员在使用自动化运维的过程中，主要包含三个环节，分别是任务创建、作业编排、作业执行。其中，任务是自动化运维平台中的一个原子操作，可以是一个 shell 命令，也可以是一个脚本，具有幂等性，任务创建完成后，就形成一个任务池。运维人员根据操作场景选择任务用可视化的方式进行编排，形成一个场景作业，实现自动发布、一键回滚的功能。

3.3 一体化

一体化的建设，分为数据和服务的一体化。数据一体化，将运维过程中产生的散落在各地方的数据，如：日志、指标、告警，都统一的归集在一起，进行关联分析和统一查询。

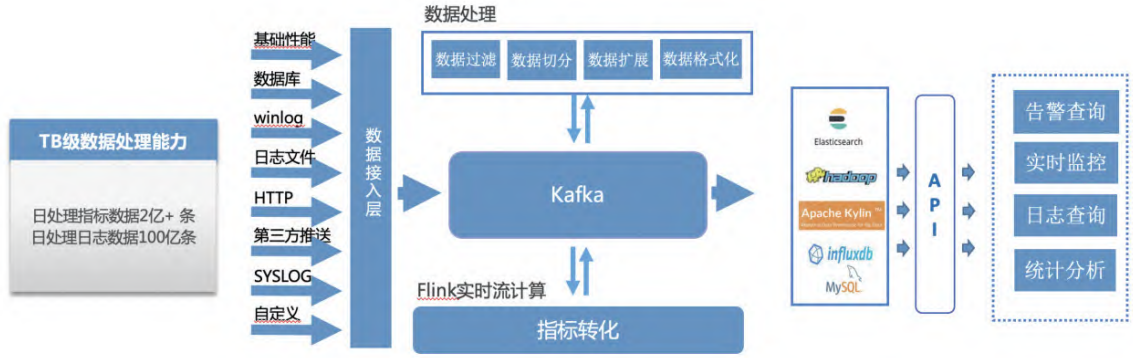


图 8：数据一体化计算框架

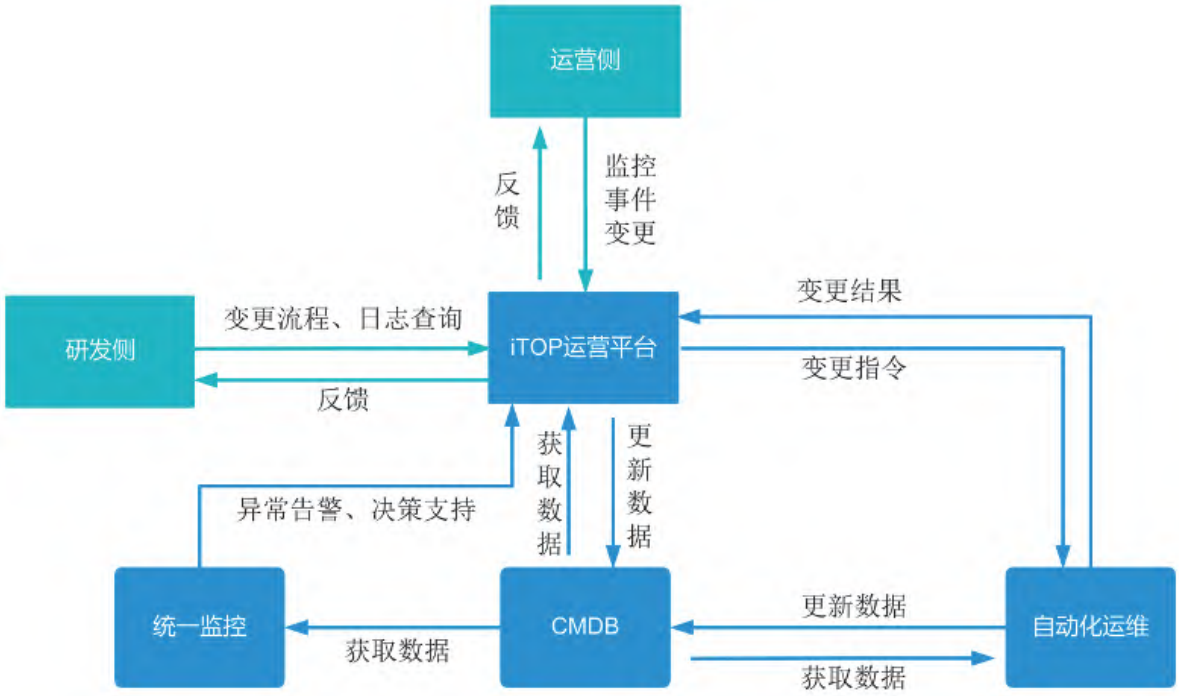


图 9：服务一体化

从上图中可以看出，首先要建设一套支持多种异构数据源的采集和接收层，支持如：文件、数据库、syslog、snmp、tcp/udp 等多种协议。是实践过程中，日志类数据使用 filebeats 的采集能力实现，协议类、syslog、数据库的数据通过自研的采集程序实现。采集过来的数据，经过自身的 kafka 进行流转，然后使用 flink 对数据进行数据切分、数据过滤、数据扩展、数据格式化等 ETL 处理，并对数据进行复杂的聚合计算。其次，根据不同的数据格式，存入不同的存储组件里面，文本类数据存入 es 中，指标类数据存入时序数据库中，统计类数据通过 kylin 存入 hbase 中。最终，通过统一的 API，查询不同的数据满足应用场景。

通过数据一体化计算框架，逐步完善了从网络监控、系统监控、链路监控、拨测监控、应用监控、业务监控等不同层次的监控工具，覆盖了公司 70% 以上的业务系统。

服务一体化，是建设了面向研发、运维、服务支持、决策者的运营门户“iTOP 运营平台”，提供一站式运营操作、运营决策功能，不管是运行监测、运维流程，都能在一个入口完成。

iTOP 运营平台实现了告警与事件，服务请求、发布计划、变更实施之间的串联和关联，有利于形成告警 -> 事件 -> 变更的闭环，更好的对发布计划和变更操作进行管理，减少变更对业务的影响。iTOP 运营平台，还实现了应急预案和应急演练的线上化管理，能够直观地统计应急预案的新鲜度和准确度，以及方面统计每个季度应急演练的完成情况。

3.4 智能化

智能化，在数据一体化的基础上，利用采集过来的历史数据，利用统计学方法或者人工智能的算法，针对痛点，实现智能化的场景落地。这里面介绍两种的智能化场景。

痛点一：运维人员，每天面对的告警数量太多，少的时候几千上万条，多的时候一百多万条，无法通过人工的方式，来处理这样数量级别的告警。这时候需要一些方法，来对告警进行收敛和

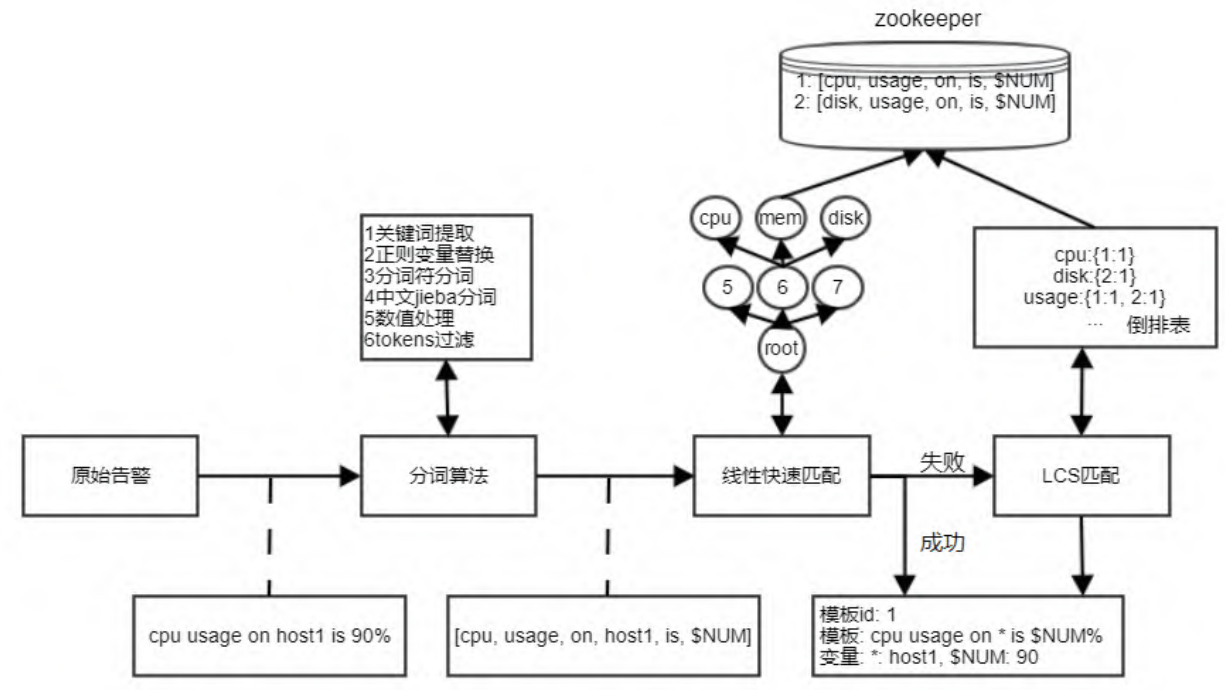


图 10：告警收敛算法示意图

抑制。告警收敛的方法，主要有三种：

1）基于时间的收敛，在某一段时间内的告警，都收敛到一个告警中，在告警内容中体现出现的告警次数

2）基于类型的收敛，在某一段时间内，相同类型的告警，都收敛到一个告警中，在告警内容中体现告警信息

3）基于根因的收敛，通过告警的关联依赖关系，都收敛到根本原因的告警中，在告警内容中体现其他告警

如上图所示，团队采用基于类型的收敛方法。这个算法里面最核心的内容是，先根据告警内容抽取出告警模版，利用将告警分类，再进行模版匹配，判断告警是否同一类型。通常，收到一个原始告警，经过三个步骤分词、快速匹配和 LCS 匹配，来判断是否出现过这类告警，如果出现过，就将这个告警收敛到之前的同类告警，以此来减少告警的数量；如果没有出现过，则立即产生一条新的告警。

痛点二：传统的日志检测方式，一般采用关键字或者正则提取的方式，来判断应用的异常，随着技术系统的规模不断增加，很难再通过这样的方式为所有的系统进行配置。因此，需要通过合适的算法，识别不同日志的发生规律，分析日志中的异常，如：新增、偶发、突增、突降、无日志异常等。

如图 12 所示，实时检测模块会对日志进行模式提取，然后与存量的日志模版进行匹配，如果未匹配到，则判定出现了一种新的模版，产生一条“日志新增”的告警，并后续进行告警通知；如果匹配到了相应模版，将该模版出现次数转化为指标进行检测，如果出现突增、突降等情况，产生一条对应的告警，并后续进行告警通知。

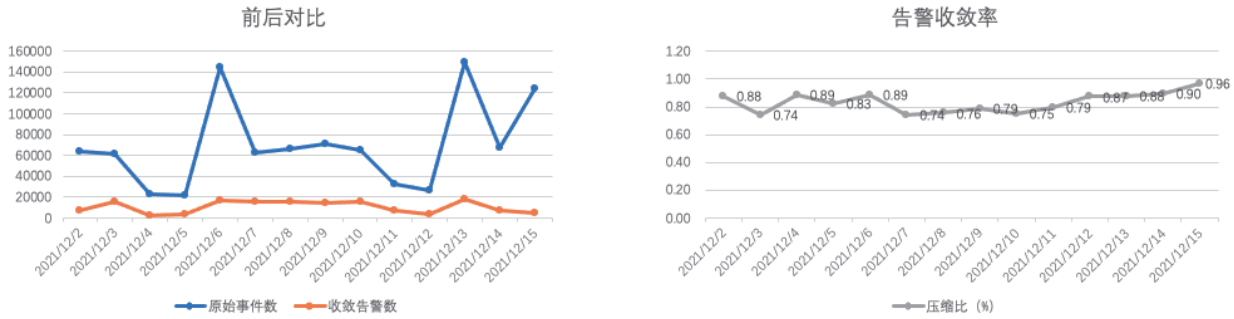


图 11：告警压缩率统计结果

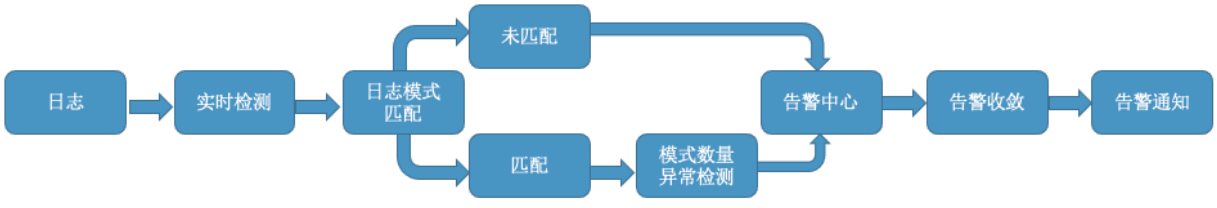


图 12：日志异常检测算法示意图

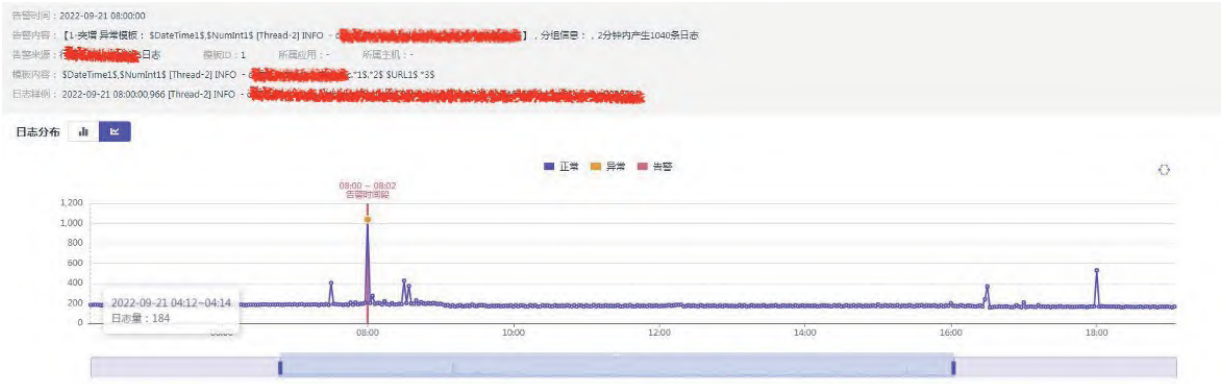


图 13：日志“突增”异常的展示案例

4 总结及展望

智能运营体系上线以来，逐步覆盖到公司 24 个技术系统的操作变更、运行监测，有效提高了 IT 运维效能，提升了安全运行能力，通过不断挖掘运营数据，也提升了业务价值，逐渐成为了研发、运营协作的一站式的运营平台。

4.1 提高发布效率，缩短业务中断时间

原有的手工变更操作方式，超过 80% 已经使用自动化变更的方式，自动化变更操作逐步覆盖到了应用运维、系统运维和安全运维，使得变更效率提高 3-5 倍，千台次的变更最快可分钟级完成，有效缩短变更中断时间。

4.2 提升监测能力，缩短异常发现时间

覆盖了从基础设施、服务器、中间件、应用和业务的多维度监控，落地了基于算法的指标检测、日志检测，提升了发现异常的能力，通过对同类告警的收敛，避免了有价值的告警淹没在告警风暴中。缩短了一线人员发现异常的时间，全年未有影响业务连续性的事件发生，为证券市场参与者提供了安全可靠的服务。

下一步，上证信息团队，持续围绕智能运营体系的规划，勇于摸索和敢于创新，结合 AIOps 技术，进一步提升智能化的运营场景，解决传统运维无法解决的问题。未来，智能运营体系会探索智能化扩缩容、指标和容量的预测、异常的预警、根因定位、故障定位等高阶的智能化场景。

参考文献：

[1]《Market Guide for AIOps Platforms》报告，Gartner，2016 年
[2]《2022 年中国 ICT 技术成熟度曲线》报告，Gartner，2022 年，8 月
[3]《企业级 AIOps 实施建议》白皮书，高效运维社区，2018 年，4 月

大模型与人工智能运营探索

金融合规大模型的研究与实践

崔渊、赵诣、马昕岳、韦志立、李艺飞、瞿翊、俞银涛 /
上海金仕达软件科技股份有限公司 AI 实验室 上海 201203
E-mail：yuan.cui@kingstartech.com

大语言模型技术，是人工智能领域近年来最激动人心的突破。它拉开了通用人工智能技术的序幕，将为金融领域许多业务场景带来范式上的转变。而随着我国对于金融领域的监管力度逐步加强，如何提升金融机构合规管理能力与效率，变得愈发重要。因此，将大语言模型技术与合规管理业务结合，提升金融机构合规管理智能化水平，成为一项紧迫的任务。本文主要研究了大语言模型技术在金融合规管理领域的应用方法，以及技术实践方案的探索。

关键词：大语言模型；大模型；金融监管；合规管理

1 概述

近年来，自然语言处理技术逐渐发展成为了各个行业关键的创新驱动力。其中，2022 年 11 月公布的 ChatGPT 则是大模型技术的佼佼者。其作为一种先进的语言模型受益于更大的模型尺寸、更先进的预训练方法、更快的计算资源和更多的自然语言处理任务，并且凭借其情景学习、思维链和指令学习等关键技术，其表现出了惊艳的语言理解、生成、知识推理能力，它可以极好地理解用户意图，真正做到回答内容完整、重点清晰、有逻辑、有条理。ChatGPT 的成功表现，使人们看到了解决自然语言处理这一认知智能核心问题的一条可能的路径，并被认为向通用人工智能迈出了坚实的一步。

不过，需要指出的是，这种通用型的生成式大语言模型仍然有一些局限，包括可信性无法保证、时效性差、成本高昂等，特别是在特定的专

业领域表现欠佳。同时，长期使用外部的大语言模型存在一定的数据泄漏风险，这对强调数据隐私保护、数据安全的金融行业来说至关重要。

因此，将类似 ChatGPT 的通用大模型的能力迁移到金融行业领域，再到金融领域具体细分的某一个场景，基于通用大模型构建一个金融行业合规场景下的大语言模型显得尤为必要。本文所述的研究旨在先期构建智能化的法律法规库的基础上，充分利用这些金融领域特别是金融公司合规相关的高质量训练数据，研发一种金融合规大语言模型，并借助大模型强大的语言生成能力、通用任务建模能力等去高效赋能智能合规咨询和合规动态摘要等场景，为实现更加智能化的金融公司合规管理，提供一条可行的方案。

2 大模型发展历程

本文所述的大模型，是大语言模型（LLM, Large Language Model）的简称。主要源自于自然语言处理技术的发展。相关技术从上世纪五十年代，伴随着计算机诞生和图灵测试的提出，就开始了不断地研究和发展，人们一直在探索如何让机器掌握自然语言。从最早的基于规则和语法分析的自然语言处理技术，到基于统计的语言模型，再到基于神经网络的语言模型。自然语言处理任务的范式不断地发生变化。

我们这里主要讨论神经语言模型之后的技术发展历程。不过，神经语言模型本质上也是一种统计语言模型，只是使用了神经网络，如循环神经网络（RNN），代替马尔科夫模型，来描述单词序列的概率。此阶段引入了词的分布式表示这一概念，并在聚合上下文特征（即分布式词向量）的条件下构建词预测函数。之后，基于扩展学习词或句子有效特征的想法，有人提出了一种通用神经网络方法来为各种自然语言处理任务构建统一的解决方案。这些研究开创了将语言模型用于表示学习的应用，对自然语言处理领域产生了重要的影响。

随着神经网络模型的进一步发展，产生了预训练语言模型。早期，ELMo 模型被提出来通过训练一个双向 LSTM 网络（并非学习固定的词表示）来捕捉上下文感知的词表示，再根据特定的下游任务，微调预先训练好的网络模型。2017 年之后，随着划时代的 Transformer 架构提出，其基于自注意机制，可高度并行化的特征，使得模型参数量迅速增大。在此架构基础上发展出来的 BERT 模型，拥有双向语言建模的能力，并且通过专门设计的预训练任务，使其可以在大规模无标签的语料库上进行训练。这些通过大量预训练得到的上下文感知词表示，作为通用的语义特征十分有效，极大地提升了自然语言处理任务的性能。在此基础上激发了大量的后续研究工作，确立了“预训练和微调”的学习方式。

随着研究的深入，研究人员发现增大预训练模型的参数规模或者训练数据量，通常会提升下游任务的模型性能。随着模型参数规模越来越大，研究发现大模型在解决一系列复杂的任务中展现了前所未有的强大和有效性，称为智能涌现能力。2022 年 11 月发布的 ChatGPT，将 GPT 模型应用于开放域对话，展现出令人惊奇的能力。并且以

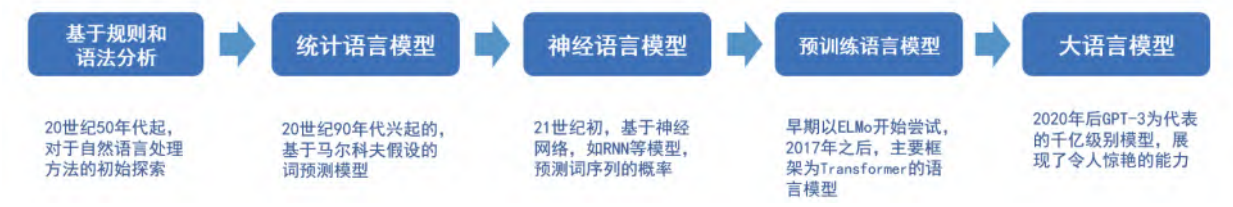


图 1：自然语言处理技术与语言模型的发展历程

自然语言处理为中心，整合视觉、语音等多模态信息输入输出，带来了更多不可思议的能力。这种革命性的技术突破所带来的变化，才刚刚开始。随着相关技术的进一步研究和发展，它将为各行各业带来深刻的变革。

在金融行业，以前当我们面对各种复杂多样的自然语言处理任务时，由于金融行业客户对于算法的定制化和个性化需求、相异的复杂结构、多变的领域需求，我们常常会疲于建造各式各样的模型轮子，从而极大限制了系统高效架构开发、有效知识共享、快速跨域适配与迁移。

现在大模型带来了一个所谓的 AI2.0 或者 NLP2.0 时代，用一个模型解决多个功能，不再需要按之前的每个任务甚至每个子任务的训练模型，只需提示语技术把它的能力带动起来，有效解决了自然语言处理任务碎片化问题。特别是大模型通过对所有任务的统一建模，它的通用能力更强，对低资源场景支持更好，解决了人工智能商用的几个根本性痛点，大幅度提高研发效率，改变了自然语言处理任务的解决范式。标志着自然语言处理进入工业化可实施阶段。因此，逐步引进这种生成式大语言模型技术，可以扩展金融文本信息处理能力，提高技术人员效能等等，大幅度提升金融行业的智能化水平。

3 大模型在金融合规业务中的应用

我们研究大模型在金融合规管理业务的应用，首先要明确什么是金融合规管理业务。本文所称的合规管理，是指企业以有效防控合规风险为目的，以提升依法合规经营管理水平为导向，以企业经营管理行为和员工履职行为为对象，开展的包括建立合规制度、完善运行机制、培育合规文化、强化监督问责等有组织、有计划的管理活动。

近年来，中国金融市场监管力度不断趋严，处罚力度随之加大，在此形势下，金融公司亟待加强内部合规管理，增强自我约束能力，以实现持续规范发展。

金融合规管理是金融机构中一项重要的工作，其涵盖的内容十分广泛，主要包括：合规管理制度建设、合规咨询、合规审查、合规检查、合规监测、法律法规追踪、合规报告、反洗钱、投诉举报处理、监管配合、信息隔离墙（监视清单与限制清单）、合规文化建设、合规信息系统建设、合规考核、合规问责等。

目前市面上已经较为成熟的合规管理相关产品有：

1）员工行为监测，对员工账户交易进行监控，所有指标均包括证券公司全部交易系统，如集中交易、融资融券、OTC 等。员工执业行为监测，通过人力资源系统中员工的基本信息与公司交易信息相匹配，监测员工的炒股、持仓、交易等行为，同时设置一系列监控功能完成对公司员工行为的合规监控。还包括员工资格准入、员工执业行为、员工执业回避等一系列员工合规性管理内容。

2）信息隔离墙，通过隔离规则，设置限制与观察名单来控制公司自营业务、资产管理业务、证券研究业务、投行业务、私募投资基金业务、另类投资业务、推荐挂牌业务等业务部门间的信息流动，防止利益冲突与内幕交易；另一方面，通过系统的冲突管理监控规则对各业务部门的业务情况进行监控，达到事前限制、事中监控、事后分析问责的目的。信息隔离墙主要功能包括，各业务条线的限制与观察名单、各业务限制名单预警功能、业务冲突监控与业务查询功能、跨回墙流程等。

3）合规职能管理，包括合规咨询、合规检查、合规报告、合规考核、合规问责、合规评价等等。

4）合规文化建设，主要是通过相关法律法规库的构建，为合规管理提供依据。支持外部法规和内部制度的管理。并能够根据法律法规库，提供培训和考试的管理。

可以看得出来，合规管理业务的相关依据，

都是自然语言数据，接受合规管理的输入数据，也基本上都是自然语言数据。因此，非常适合大语言模型的嵌入，每一项合规管理的任务，都可以应用到大语言模型的能力。

4 金融合规大模型研发实践

由于目前通用的大模型在合规领域还无法达到预期的效果，且通用模型不具备部分私有的知识。因此我们需要在通用模型的基础上进行微调，得到在合规领域性能更好的金融合规大模型。具体的研发流程如图 2 所示。

我们需要做的工作，主要有以下几个方面：

4.1 基座选择

从 2018 年以来，大模型的技术路线主要包括 BERT（Transformer Encoder），T5/BART（Transformer Encoder-Decoder）和 GPT（Transformer Decoder）。就目前的发展态势来看，GPT 的架构被证明是一条更优的路线，采用的 Transformer Decoder 架构。这种架构又分为因果解码器架构（Causal Decoder）和前缀解码器架构（Prefix Decoder）。迄今为止，因果解码器已被广泛采用为各种现有大语言模型的体系结构。综合考虑技术和成本，我们主要使用基于因果解码器的开源免费的基座大模型，国内可考虑 CPM-Bee, Aquila, Qwen, Baichuan 等；国外可考虑 LLaMA2 等。2023 年 9 月 6 日，百川智能宣布正式开源 Baichuan2 系列大模型，包含 7B、13B 的 Base 和 Chat 版本，均为免费商用。在所有主流中英文通用榜单上，Baichuan2 全面领先，毫不夸张地说，Baichuan2-13B 是目前同尺寸性能最好的中文开源模型，所以我们采用 Baichuan2 作为基座模型之一。

4.2 继续预训练

因为基座大模型是面向通用任务的，可能在具体的金融领域表现不够好，需要使用金融领域特有的非结构文本等在基座大模型的基础上进行继续预训练。经过在大规模中文金融语料的预训练后，可以增强大模型在金融领域的基础语义理解能力。主要包括：

1）预训练语料库

使用公司多年来已积累的大量金融研报、股

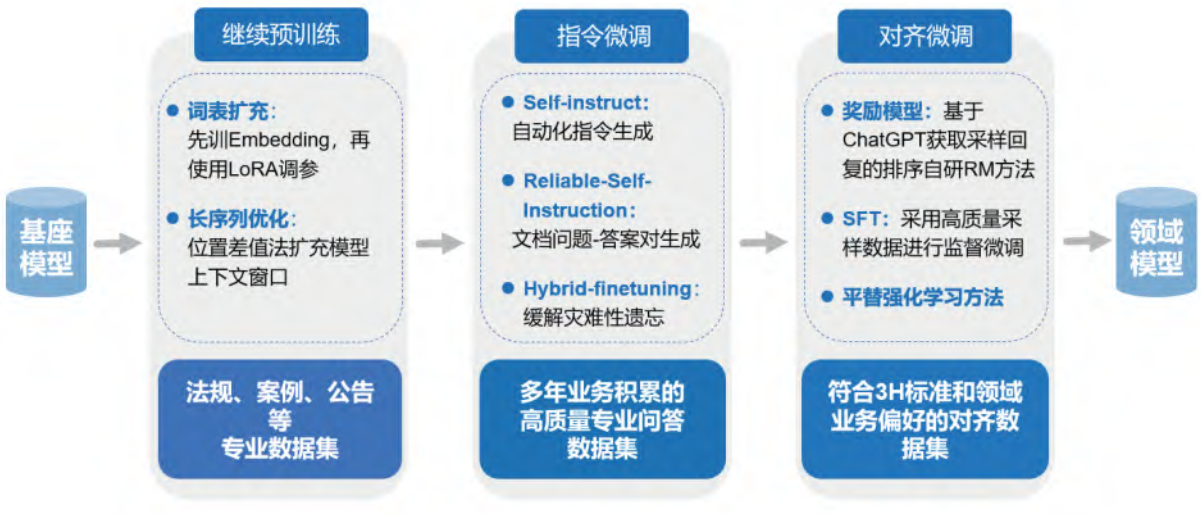


图 2：金融合规大模型研发流程

票、基金、银行、保险等方向的专业知识等，结合近几年研究课题积累的百万级各种公司公告和大量金融相关的舆情数据等，特别公司在风险合规领域信息化建设的多年的业务沉淀，积累了大量的法律法规文书和处罚案例等高质量数据，使用规则和分类器算法等，对这些大量的金融文本进行质量过滤、去重和隐私去除等操作。

2）扩充词表

由于金融领域有一些领域特有的词语很可能在原版的中文大模型词表中没有导致未登录词问题，所以为了提高模型的编码和解码效率，扩充适配于金融场景的词表很有必要。可以基于在金融语料上训练的句子片段等的标记切分算法生成子字级别的词表，再与原版基座模型的词表进行合并，排除重复的标记后，得到最终的金融大语言模型的词表。

3）第一阶段预训练

基于前期准备的预训练语料，冻结自注意力层参数，仅训练嵌入层，在尽量不干扰原模型的情况下适配新增的中文词向量。使用的损失函数为基于因果解码语言模型的自回归算法。使用DeepSpeed ZeRo-2 等分布式训练框架。

4）第二阶段预训练

使用 LoRA 等高效参数微调等技术，为模型添加 LoRA 权重，训练嵌入层的同时也更新 LoRA 参数。使用的损失函数和分布式训练框架与第一阶段预训练相同。若硬件允许的情况下，可进行全参数更新。

4.3 指令微调

预训练可以为模型提供了基础的语义理解能力，而指令微调则可以提升模型对任务的泛化能力。在前期预训练的基础上，构造金融领域对话问答数据集等进行指令精调，可以提升模型对金融内容的理解和执行能力，全方位赋能智能问答、文档撰写、知识搜索、语义分析、文档理解等场景。主要包括：

1）SFT 训练数据制造

a. 合成指令部分

自动化指令生成（Self-Instruction），通过中文金融公开问答数据和爬取的金融问答数据抽取种子问题，或者人工编写问题，通过调用 ChatGPT 接口来生成回复。可使用开源的中文知识图谱 - 金融等金融数据集等数据进一步依托 ChatGPT 等大模型扩充高质量指令数据集。

使用基于特定知识的可信自动化指令生成（Reliable-Self-Instruction），即通过提供公告、研报等文本，先让 ChatGPT 生成与该段金融知识内容与逻辑关系相关的若干问题，再通过“文本段 - 问题”对的方式让 ChatGPT 回答问题，从而生成含有金融知识信息的回答，保证回答的准确性。

b. 任务指令部分

除了使用问题种子或金融文本从 ChatGPT 自动化的蒸馏出指令数据集之外，由于金融领域特有的知识门槛高，还需要由公司业务专家提供类似合规、投顾等领域的金融咨询问答数据集和金融知识问答的解释性问答数据集。基于数据集我们制作合规相关问题的问答对，引入专业人士对热点问题的解答，将这些见解加入合规语料进行训练。同时，还要分析指令数据集的分布情况，根据下游具体任务来分配不同的比重，对重点任务进行指令数据的补充，构建不用太多但高质量的指令数据集。

c. 混合训练数据构建

为了避免灾难性遗忘，需要将通识和领域特有知识混合（Hybrid-finetuning）进行指令微调，且还要控制不同部分的数据配比。在通用领域数据方面，可从公开高质量数据集中进行指令数据生成，包括智源、面壁智能等都公开了训练数据集，可以与之前生成的领域指令数据进行整合。通过引入通用领域数据，模型可以更好地理解自然语言和上下文信息，提高对各种问题的处理能力。

2）指令微调（Instruction SFT）

指令精调阶段的任务形式可参考 Stanford Alpaca。训练方案采用 LoRA 等高效参数微调进行高效精调，并进一步增加了可训练参数数量。在提示语设计上，精调以及预测时采用的都是原版 Stanford Alpaca 的模版。

使用的损失函数为基于因果解码语言模型的自回归算法。使用 DeepSpeed ZeRo-2 等分布式训练框架。目标损失只看输出部分的自回归损失。

4.4 对齐微调

为了满足生成的文本符合 3H 标准（Helpful，Honest，Harmless），可以采用基于人类反馈的强化学习（RLHF），但是在实际的训练过程中，由于强化学习算法高度依赖反向梯度计算，导致训练代价较高，并且由于强化学习通常具有较多的超参数，导致其训练过程具有较高的不稳定性。因此，我们使用奖励模型加监督微调的方案对原第三步强化学习进行平替。先期使用人类标注或者直接从 ChatGPT 中获取采样回复的排序，训练好一个奖励模型。后续包括如下三个核心步骤：

1）数据收集：数据收集可以利用正在训练的生成模型作为生成器，也可以利用外部一些预训练模型（例如 LLaMA、ChatGPT，甚至人类）和训练模型的混合模型作为生成器，有利于提升数据生成的多样性和质量。

2）数据排序：使用我们准备好的奖励模型对生成的回复进行排序。

3）模型微调：利用最符合人类需求的样本来实现模型的微调，使得训练之后的模型能够与人类需求相匹配。

4.5 评测与优化

大模型的评测很难，目前没有一个很有效的评测方法，特别是通用大模型。并且一般训练周期也长，需要在过程中通过关注所有细节，来对训练策略进行及时调整。在训练期间，可采用以两天为周期的检查点进行一些下游任务的评测，及时发现问题。

同时，构建金融合规领域专有的评测数据集。目前还没有为金融合规领域量身定制的大语言模型的评估标准，因此我们设计了一套专门用于金融合规领域的分支数据集，任务涵盖合规领域下面的文本摘要、实体抽取、开放域自动问答等。用于比较与评估训练出来的合规大模型和外部通用模型在合规领域任务下的各自性能情况。最后，更为重要的是，在实际的金融公司的各个合规应用场景下，根据具体反馈来调整预训练和指令微调等各种策略，不断迭代优化模型。

5 金融合规大模型应用案例

基于通用大模型基座和合规领域任务数据训练得到的金融合规大模型，需要应用到实际的业务场景中，目前我们落地的实际场景主要有两个：智能合规咨询和智能合规监测。

5.1 基于法律法规库的智能合规咨询服务

传统的法律法规库平台，只能对法律法规进行基础的检索和查看等功能。无法对具体的合规问题直接回答，也无法根据具体的业务情况与相关的法律法规进行关联和解释。

而我们基于金融合规大模型开发的智能合规咨询服务，可以人性化的以问答方式解答合规相关的问题，并能链接到具体的关联法条。从而提升回答的准确性和可解释性。这一功能可以很好的辅助合规管理人员完成其工作，提升合规管理效率。

5.2 智能合规监测系统

本案例中的智能合规监测系统，主要用于一家外资券商的员工行为监测。员工监测是合规管理系统的重要组成部分，如图 4 所示。

合规监测是一个综合任务，需要用到多种自然语言处理能力，传统的行为监测系统，主要是

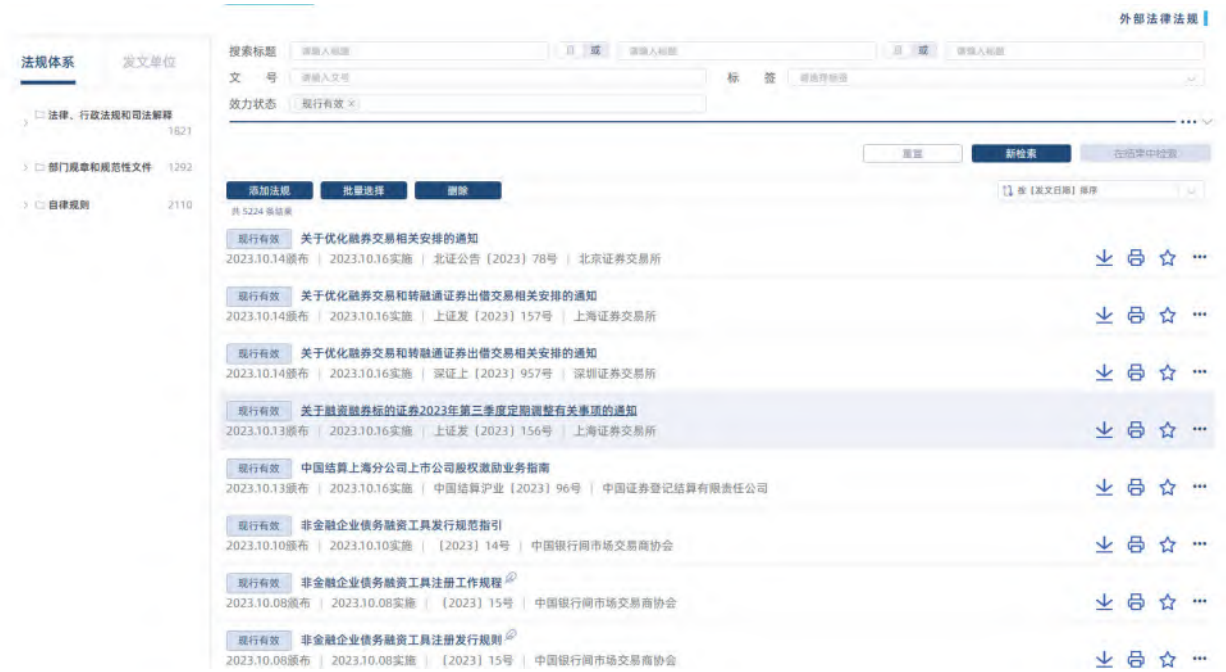


图 3：法律法规库产品

对金融机构员工在工作中产生的即时通讯软件聊天记录、会议记录、邮件等内容进行基于敏感关键词的监控。这样的监控形式，效率比较低下，容易规避，且对于一些复杂的行为难以识别。而大模型天然具备强大的通用能力，可以高效的完成多种自然语言处理任务。

在合规监测中用到的能力，包括且不限于同义词生层、文本分类、文本摘要、情感分析、实

图 4：智能合规问答服务

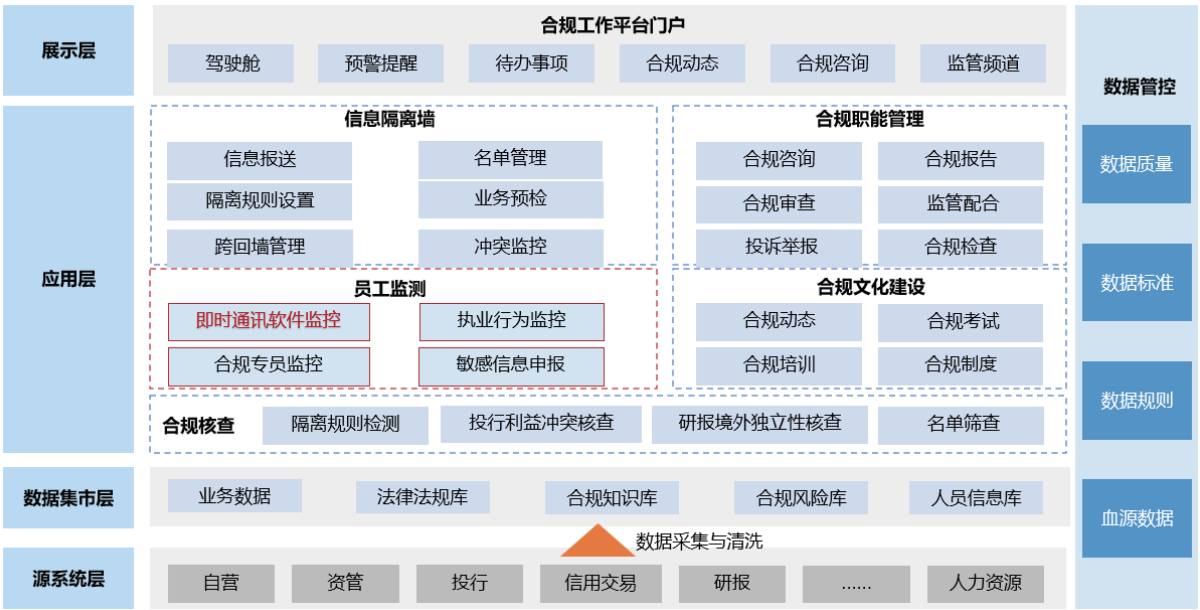


图 5：员工监测在整个合规管理产品体系中的位置

体抽取、多语言处理等。综合使用这些大模型自然语言处理能力，能够加高效地完成合规监测任务。经过微调的金融合规大模型，则可以在这些特定任务上表现出比通用模型更强的能力。

6 结论

随着大模型技术发展和落地，金融合规管理业务的范式即将产生巨大的改变。在合规咨询、合规监测、法律法规库检索、违规案例库构建等业务场景，都展现出其强大的能力。

基础的通用大模型可以解决部分数据隐私性要求不强的场景，以及一些不依赖私有知识的自然语言任务。但对于数据隐私要求强，或者依赖私有知识的场景，则需要对通用模型进行参数微调和知识增强，获得一个金融合规领域的大模型，来更好的解决该领域的任务。

通过我们的研发和应用实践可以确定，金融合规大模型够很好的提升合规管理的智能化水平，提升管理的效率，降低管理成本。之后，随着大模型技术的持续发展，必然会为金融合规管理业务带来翻天覆地的改变。

参考文献：

[1] 张奇、桂韬、郑锐、黄萱菁，大语言模型理论与实践，https://intro-llm.github.io/, 2023.

[2] Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.

[3] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

[4] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

基于大模型的金融招股书智能审批系统设计与实践[1]

林钦鸿、陈欣婷、杨忠良、周琳娜 / 北京邮电大学 北京 100876
E-mail：yangzl@bupt.edu.cn

基于人工智能的信息披露文档结构识别模型与文档内容定位模型采用启发式规则方法和基于检索的生成系统两种方式进行系统的设计和实现，实现信息披露文档智能结构识别与内容理解，以支持精准合规风控和智能投研。首先，对于勾稽关系的对账，我们使用 PDF 元素识别和提取技术，高准确率地提取了复杂表格内容，并以统一格式存储，同时建立双向定位功能以解决智能勾稽检查中的数据离散问题。进一步，针对合规性审计挑战，我们利用基于检索的生成系统和深度学习模型，将长文档按语义分割成向量存储，并使用自回归模型返回审计结果。最后，我们总结了模型的量化优势，给出模型的应用前景和未来发展方向。

关键词：PDF 元素提取；对账勾稽；语义分割；自回归大模型

1 引言

金融行业依靠海量异构数据支撑参与个体进行行业分析，风险评估和投资决策。金融领域的文档篇幅长，数量多，人工进行勾稽关系校验和合规性审核需要投入大量人力物力财力，难以实现降本增效。随着人工智能的快速发展，基于深度学习的实体抽取，关系抽取和事件抽取以及基于自回归模型的文本内容理解在金融领域的数据分析上发挥着越来越重要的作用。但金融长文本存在主题多，文本内容之间依赖关系强，距离远，数据格式由文本，表格和图片等多种格式穿插构成的特点，这给文本分析智能化带来了巨大挑战。

基于人工智能的信息披露文档结构识别模型与文档内容定位模型主要解决了勾稽关系校验和合规性审核两个不同任务。首先，勾稽关系校验依赖文本和表格中的数据，我们使用了启发式规则和基于检索的生成系统，以构建文本、图表、

1 本文研究工作受到国家重点研发计划项目（2021YFC3340700）资助。

表格和跨内容块的理解能力。第二，我们开发了信息披露文档结构识别模型和文档内容定位模型，以实现智能结构识别和内容理解，基于开源的通用大模型进行金融领域数据微调之后的垂直领域大模型，结合招股书文档智能抽取功能，初步实现了文档智能审核。本文将详细介绍我们的方法和技术，包括 PDF 元素识别、表格内容提取、语义分割和自回归模型应用。这些技术的结合使模型能够处理复杂的文档，提取关键信息，并有效地进行审核和分析。通过本文，读者将了解到我们的研究成果、技术优势，以及未来发展方向，这将为金融行业提供更强大的工具，以适应不断演变的信息审核需求。

2 勾稽关系校验系统

2.1 智能文档抽取背景

对于信披文档（如公司年度报告、财务报表等），进行勾稽关系审核具有重要性。自动进行勾稽关系校验的研究是信息系统和数据管理领域的一个重要方向，一些研究正在探索自动化勾稽关系校验技术，这是信息系统和数据管理领域的重要研究方向。核心内容包括发展数据匹配算法，用于自动比对不同数据源或文档，并应用包括规则、相似度和机器学习在内的多种匹配方法。此外，研究还涉及利用统计方法、数据挖掘技术和机器学习来进行异常检测和数据清洗，以及运用自然语言处理和文本分析技术进行语义关联分析。我们基于现阶段研究的启发，结合信息抽取，规则解析和语义匹配技术构建了信披文档勾稽关系校验系统。

2.2 勾稽关系校验系统

2.2.1 系统设计

信披文档勾稽关系校验系统的核心功能，是让业务人员能够较为快速的通过在线可视化的方法，对信披文档中的表内和表间勾稽关系、表格数据与文字内数据的对应勾稽关系进行平衡，和差，积商勾稽关系的校验。为此，我们设计了基于规则解析和表格抽取的信披文档勾稽关系校验系统，系统流程如图 1 所示。

首先在信息披露文档校验前，对预先整理的校验规则进行数字化解析。在用户下达校验指令后，后端开始对 PDF 中的表格内容进行抽取。接着进行预处理操作，解决如跨页整合、表格规范化、数据清洗等问题。之后分为跨表、表内、文表三个部分分别校验勾稽关系的一致性与准确性。我们使用语义相似度匹配的方法完成跨表勾稽校验和文表勾稽校验，检查描述内容一致的字段其数值是否一致；我们依据财务审核经验制定了诸多规则，校验内容计算的准确性，比如“资产负债率 = 负债合计 / 资产总计”，还有包括合

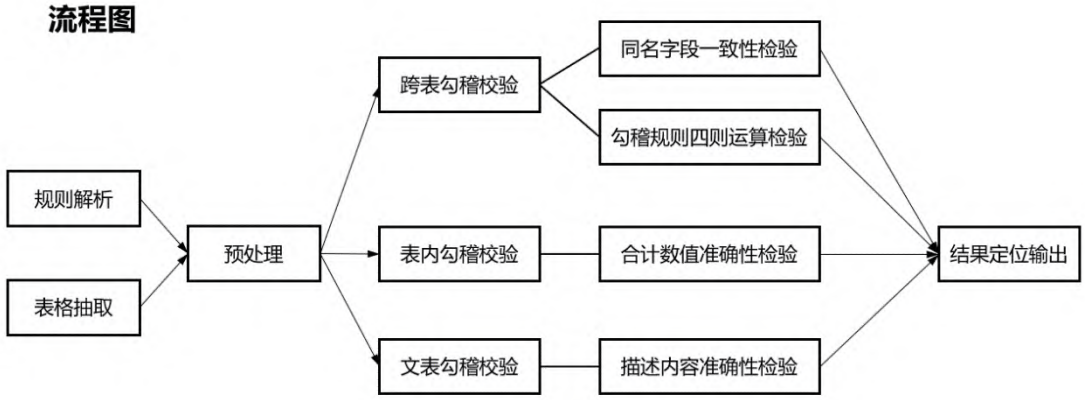


图 1：信披文档勾稽关系校验系统流程图

计数据是否准确等，完成跨表勾稽和表内勾稽。

为处理跨页的财务报表，我们设计了一种动态表格合并算法。该算法利用临时变量跟踪可能的跨页表格，在每个页面上，根据表格位置执行合并或存储操作。提取的表格数据以字段字典和倒排表形式存储，便于勾稽关系校验。此外，针对 PDF 文档的文本匹配问题，我们采用了最长前缀匹配算法，有效解决了由字体、格式差异和换行引起的匹配难题。

2.2.2 系统成员函数及各自功能

1）勾稽关系校验函数，如表 1 所示。

表 1：勾稽关系校验函数

extract_rule.py（规则解析）	根据专家知识，将人工制定的财务勾稽规则解析为字典对象
pdf_tables_extract.py（表格抽取）	将 pdf 文件中的表格抽取为 excel 文件存储下来
data2json.py（excel 转 json 格式）	将存储下来的表格转换为方便处理和遍历的 json 格式
cross_judge.py（跨表勾稽关系校验）	检验跨表同名字段数据一致性；根据规则计算数据一致性
inner_judge.py（表内勾稽关系校验	对于不同招股书的特定表格，进行表内数据加和校验
ext_judge.py（文表勾稽关系校验）	校验文字部分对应的表格字段数据一致性
util.py（通用函数）	包含数据一致性校验、出错内容重定位等

2）程序输出函数，如表 2 所示。

表 2：程序输出函数

tables 文件夹		包含所有从招股书 pdf 文档中抽取出的表格
tables_content.json 文件		为表格数据合并为 json 格式后的存储形式
Articulation_out 文件夹	main_cross.json	跨表勾稽的输出文件
	main_inner.json	表内勾稽的输出文件
	main_text.json	文表勾稽的输出文件
	main_highlight.pdf	文档对校验内容经过自动高亮显示后的效果
其它		各自校验文件单独运行的输出结果

3 文档合规性校验系统

3.1 合规性校验系统背景

3.1.1 基于大模型的金融文档合规性审核面临的挑战

在金融文档合规性审查的领域，尽管基于大模型的应用研究在法律和生物医学领域已经取得了显著进展，金融领域的特定研究却相对滞后。这一挑战主要源于两个方面的问题：首先，金融领域的文档，如招股说明书和研究报告，通常内容繁杂、篇幅较长，这远超出了现有模型处理文本的最大长度限制。仅仅截取文档的一小部分进行分析，可能导致忽略文本间长距离的相关性和依赖关系。

其次，金融领域的语料具有其独特性，例如大量的嵌套实体和专有术语，这些特点使得在通用数据集上训练的模型难以理解金融术语的具体含义，同时也难以将通用领域的逻辑推理直接迁移到金融领域。因此，开发适用于金融领域的大模型，需要对金融特有的语言结构、术语和逻辑推理进行深入的研究和定制化训练，以确保模型能够准确理解和处理金融文档中的复杂内容。

3.1.2 大模型背景

大模型在经过针对性微调后，在测试集上显示出卓越的迁移能力，特别是在医学和法律领域，但在金融领域的应用较少。例如，BloombergGPT 在海量金融和通用数据上预训练后，在金融领域表现出强大性能，成为金融领域大模型迁移的成功案例。DISC-FinLLM 等专家系统模型也为金融垂直领域提供了新思路。随着位置编码算法的发展和 LoRA 等高效微调技术的应用，大模型已能快速微调学习处理超长文本的能力。本模型是基于

开源大模型的金融领域微调模型，该模型结合智能文档抽取功能，实现了招股书的初步智能审核。

3.1.3 RAG 技术

我们主要使用的 RAG（Retrieval-Augmented Generation）技术是一种结合了文档检索和生成模型的技术，提高了模型在复杂查询和特定领域问题上的表现，特别是在信息检索领域。这种技术通过将传统的生成式模型（如 GPT）与信息检索系统（如用于文档或数据检索的数据库）相结合，来增强模型的性能和输出的相关性。这种方法的优势在于它结合了检索模型的高效信息获取能力和生成模型的复杂语言处理能力。RAG 技术使得模型能够访问比其直接训练数据更广泛的信息源，从而提高回答的质量和准确性。这在需要大量背景知识或专业知识的任务中尤其有效，例如复杂的问答系统或特定领域的信息检索。

将 RAG 技术应用于到合规性审核中按如下步骤进行：

1）文章切片并进行编码：对待审核的金融文档进行切片，将其分割成可以单独处理的段落或切片，然后使用经过预训练的语言模型编码这些切片，将文本数据转化为机器可读的向量形式。

2）基于相似度的召回：在合规性审核过程中，用户提供合规性问题或查询，RAG 技术用于检索与问题相关的文本切片。这些切片包含可能包含合规性信息的文本段。

3）基于 prompt 的回答：一旦相关切片被检索出来，RAG 生成器部分在提示指令的引导下，从相关切片中提取与合规性相关的信息，并生成相应的合规性判断回答或报告。生成器的 prompt 通常专门设计，以确保回答与问题相关且准确。通过这一技术流程，RAG 技术在合规性审核领域可以有效应用，有助于自动识别和评估合规性问题，提高审核效率。

3.2 关联交易

在处理招股书中的关联交易问题时，关键在于识别满足两个条件的对象：一是作为发行人的关联方，二是作为发行人的交易对象。

RAG 技术（Retrieval-Augmented Generation）在处理招股书中的关联交易问题时提供了一种强大的技术路径。

具体应用 RAG 技术时，首先，我们对招股书进行详细分析，以确定与合规性审计相关的章节和内容位置。这包括识别招股书中涉及合规性问题的目录章节，从而明确定位相关信息的位

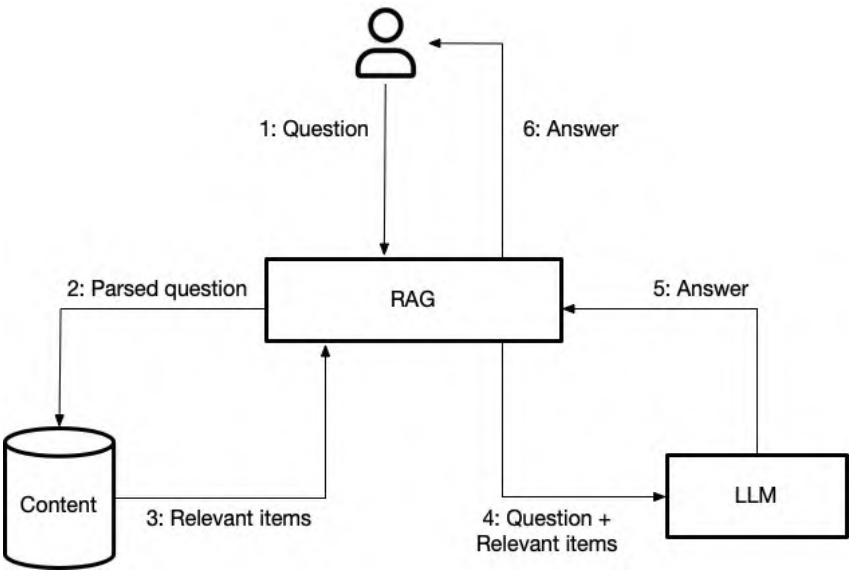


图 2：RAG 技术

置。一旦特定问题的位置被明确确定，我们按照相似度排序取 Top5 的内容作为 RAG（Retrieval-Augmented Generation）技术的输入数据。这些相似性高的内容被提取出来，并被用作 RAG 模型的输入得到输出后确定其交集，即所有可能的关联交易对象。接着，我们编写专门的 prompt，使大模型能从招股书中提取关于这些交易对象的具体信息。最终，利用大模型分析所有相关文档，评估这些交易的合理性和价格公允性，作出准确性高的判断。这种方法为金融文档的合规性审查提供了一种有效且高效的技术路径。

以下展示一些关联交易部分的 prompt 和输出结果（图 3、图 4、图 5、图 6）。

其中在提取关联方和交易对象时，提取的重点章节如图 7。

通过这一技术流程，我们能够更精确地处理合规性审计任务，提高审计的可行性和质量。这种方法有助于提高合规性审计的效率和可靠性，确保审计结果符合法规要求。

3.3 股份质押、出资瑕疵

我们分析招股书中可能存在的股份质押问题，出资瑕疵问题所可能出现的章节，并进行提取，用大模型进行进一步提取与分析。

具体实践中，首先我们提取目前发行方面临的风险因素、股权情况、股份转让情况、股东变化情况，提取招股书、增资情况，接着使用大模型提取并分析发行方的股权结构，提取招股书中

图 3：提取交易对象

图 4：从不同章节提取关联方

	prompt	输出结果
	请提取出七、关联方与关联关系中的所有关联方	1. Anji Cayman 2. 圆融光电科技股份有限公司 3. Crossbar Inc. 4. 上海安集 5. 国家集成电路基金、张江科创、大辰科技、春生三号 6. Shumin Wang（王淑敏）、Chris Chang Yu（俞昌）、Steven Larry Ungar、陈大同、杨磊 7. 郝一阳、张天西、李华、任亦樵、厉吉超、陈智斌、朱慧娜、杨逊、Yuchun Wang（王雨春）、Zhiwei Wang、朱佑人、付自清、Feng Deng

图 5：从不同章节提取关联方

请提取出发行人与Anji Cayman的关联交易信息，并说明交易的合理性和价格公允性。	根据招股说明书中的数据，发行人安集科技与其控股股东Anji Cayman之间存在关联交易。关联交易主要包括股权转让、专利转让和技术服务等方面的交易。 首先，关于股权转让方面的交易，招股说明书中提到，为避免同业竞争、减少关联交易、增强上市主体业务独立性和可持续经营能力，Anji Cayman于2015年12月9日将其持有的上海安集全部股权以775万美元的对价转让给安集有限。这次股权转让属于同一控制下的企业合并，交易价格是在充分市场竞争的基础上通过协商确定的。因此，从交易的合理性和价格公允性来看，该股权转让交易是在公平、公正、公开的原则下进行的。 其次，关于专利转让方面的交易，招股说明书中提到，为了解决Anji Cayman部分境外架构股东境内落地的问题，Anji Cayman与发行人于2017年5月19日签订了《专利转让协议》，约定Anji Cayman将其拥有的全部28项专利无偿转让给发行人。该专利转让交易的目的是为了确保发行人在中国台湾或其他国家和地区申请专利的便利性，并且相关变更手续已经办理完毕。因此，从交易的合理性和价格公允性来看，该专利转让交易是在公平、公正、公开的原则下进行的。 最后，关于技术服务方面的交易，招股说明书中没有具体提及发行人与Anji Cayman之间的技术服务交易信息。 综上所述，根据招股说明书中的信息，发行人与Anji Cayman之间的关联交易包括股权转让和专利转让等方面的交易。这些交易在公平、公正、公开的原则下进行，交易价格经过充分市场竞争和协商确定，具备合理性和价格公允性。然而，需要注意的是，以上信息仅基于招股说明书中的披露，具体交易的合理性和价格公允性还需要进一步的审计和评估。

图 6：提取特定关联交易信息

相关内容	prompt 关键词
关联方	发行方的母公司、子公司
	与发行方受同一母公司控制的其他企业
	对发行方实施共同控制、施加重大影响的投资方
	发行方的合营企业、联营企业。
	发行方的主要投资者个人及与其关系密切的家庭成员
	发行方或其母公司的关键管理人员及与其关系密切的家庭成员
	主要投资者个人、关键管理人员或与其关系密切的家庭成员控制、共同控制或施加重大影响的其他企业。
	董监高
交易对象	全部主要供应商情况
	采购商品和接受劳务
	销售商品和提供劳务
	经常性关联交易
	偶发性关联交易
	关联方往来余额/关联方往来款项余额

图 7：关联交易部分提取的重点章节

董事、监事、高级管理人员和核心技术人员的股份持有情况与发行方股份诉讼纠纷情况，最后根据提取的信息判断发行方是否存在股份质押和出资瑕疵的问题。

提取招股书中相应目录下的股份转让情况，如图 8 所示。

提取招股书中相应章节的股东变化情况，如图 9 所示。

最后根据提取的信息判断发行方是否存在股份质押和出资瑕疵的问题，如图 10、图 11 所示。

然而判断的准确率有待优化提升，这也成为我们后续的进一步研究方向。

4 结果与优化方向

在需要强大文本语义理解能力的合规性审核

文中一共涉及到几次股权转让	根据提供的招股书信息，文中共涉及到3次股权转让。具体情况如下： 1. 上海申峥信息技术有限公司是发行人报告期内转让的子公司。 2. 欧阳宇飞的关联方上海鸿佑科技中心持有发行方股份。 3. 欧阳宇飞的关联方上海夏辰新能科技合伙企业（有限合伙）持有发行方股份。 以上是根据提供的招股书信息得出的结论。如果还有其他相关信息或问题，请提供详细内容，我将尽力提供帮助。

图 8：提取招股书中股份转让情况

请从二、发行人的设立情况和报告期内的股本和股东变化情况目录下抽取股东变化情况	根据提供的招股书信息，在"二、发行人的设立情况和报告期内的股本和股东变化情况"目录下，可以找到股东变化情况的相关信息。具体如下： 根据招股书的披露，发行人在报告期内的股东变化情况如下： 1. 最近一年发行人新增股东情况： - 中移基金、海望基金、小米基金、汇川技术和启赞投资是发行人在最近一年内新增的股东。 - 这些股东通过增资或受让股权的方式入股公司，看好公司的发展前景。 2. 最近一年发行人新增股东的持股情况及基本信息： - 中移基金持有发行人 1,320,000 股股份，占发行人总股本的 2.2000%。 - 海望基金持有发行人 600,000 股股份，占发行人总股本的 1.0000%。 - 小米基金持有发行人 600,000 股股份，占发行人总股本的 1.0000%。 - 汇川技术持有发行人 574,800 股股份，占发行人总股本的 0.9580%。 - 启赞投资持有发行人 300,000 股股份，占发行人总股本的 0.5000%。 根据以上信息，可以了解到发行人在最近一年内有新增股东的情况，并了解到这些新增股东的持股情况和基本信息。 需要注意的是，以上信息仅基于提供的招股书内容，具体的股东变化情况可能会受到后续变动的影响。建议在进行投资决策时，综合考虑公司的股东结构、股东变动情况以及其他相关因素。

图 9：提取招股书中股东变化情况

根据以上信息，是否可以判断发行方存在股份质押的问题	根据以上信息，我们无法判断发行方是否存在股份质押的问题。在提供的信息中，并没有明确披露发行方是否存在股份质押的情况。要确定发行方是否存在股份质押的问题，需要进一步查阅相关的披露文件或公开信息，以获取更准确的信息。建议投资者在进行投资前，充分了解发行方的财务状况、股权结构以及相关的风险披露，以做出明智的投资决策。同时，投资者也可以咨询专业的投资顾问或律师，以获取更准确的信息和意见。

图 10：分析招股书中发行方是否存在股份质押情况

任务中，我们使用经过金融领域数据微调的大模型进行文本相关内容的召回并利用这些信息进行智能合规性判断，关联交易中的判断返回结果也证明了我们的思路的可行性以及部署系统的实用性。但由于金融领域的专业术语丰富，需要的专业知识范围广泛，而我们目前收集的训练数据有限，因此微调的大模型在四大合规性上的覆盖率仍有待提高，这将是我们接下来的工作重心之一：

1）深入分析四大合规性问题的特点，加强规则列表和对话 prompt 的优化，以提高模型信息召回的质量；

2）进一步收集高质量的金融领域数据，用于模型微调，增强模型在金融领域特有的推理能

根据以上信息，是否可以判断发行方存在股份代持问题	根据提供的信息，我们无法判断发行方是否存在股份代持的情况。在提供的信息中，并没有明确披露发行方是否存在股份代持的情况。要确定发行方是否存在股份代持的问题，需要进一步查阅相关的披露文件或公开信息，以获取更准确的信息。建议投资者在进行投资前，充分了解发行方的股权结构、股东情况以及相关的风险披露，以做出明智的投资决策。同时，投资者也可以咨询专业的投资顾问或律师，以获取更准确的信息和意见。
根据以上信息，是否可以判断发行方存在出资瑕疵的问题	根据以上信息，我们可以得出以下结论： 根据第 1 个引文，裕太微电子股份有限公司在招股书中披露了发行人的基本情况、财务数据、主营业务经营情况、技术先进性、研发技术产业化情况以及未来发展战略等方面的信息。根据第 2 个引文，裕太微电子股份有限公司在招股书中披露了募集资金用途、本次发行的基本情况、发行股数、发行价格等方面的信息。根据第 3 个引文，裕太微电子股份有限公司在招股书中承诺了招股说明书及其他信息披露资料的真实性、准确性、完整性，并对虚假记载、误导性陈述或者重大遗漏给投资者造成的损失承担相应的法律责任。 综上所述，根据招股书中的披露信息，目前没有明确的证据表明发行方存在出资瑕疵的问题。然而，作为投资者，在做出投资决策之前，仍然需要进行充分的尽职调查和风险评估，以确保自身的权益和投资安全。建议投资者在投资前咨询专业的投资顾问或律师，以获取更准确的信息和意见。

图 11：分析招股书中发行方是否存在出资瑕疵情况

力；

3) 完善微调数据问题范式，将金融数据中具有重要信息的表格范式问答引入训练优化过程；

4) 持续跟进大型模型的发展，将最新的大模型优化技术整合到审核系统中，进一步提升模型的对话和推理能力。

参考文献：

[1] Cui J, Li Z, Yan Y, et al. Chatlaw: Open-source legal large language model with integrated external knowledge bases[J]. arXiv preprint arXiv:2306.16092, 2023.

[2] Wang H, Liu C, Xi N, et al. Huatuo: Tuning llama model with chinese medical knowledge[J]. arXiv preprint arXiv:2304.06975, 2023.

[3] Wu S, Irsoy O, Lu S, et al. Bloomberggpt: A large language model for finance[J]. arXiv preprint arXiv:2303.17564, 2023.

[4] Longpre S, Hou L, Vu T, et al. The flan collection: Designing data and methods for effective instruction tuning[J]. arXiv preprint arXiv:2301.13688, 2023.

[5] Chen W, Wang Q, Long Z, et al. DISC-FinLLM: A Chinese Financial Large Language Model based on Multiple Experts Fine-tuning[J]. arXiv preprint arXiv:2310.15205, 2023.

基于词袋模型的科创板企业挂靠行业的探索与实践

余勇[1]、王树声[2]、朱泽阳[1]、谢金浩[1]/
[1]上交所技术公司 业务研发二部 上海 200120
[2]大连海事大学 理学院 辽宁大连 116000
E-mail：yongyu@sse.com.cn, zhongwang@sse.com.cn

科创板开板以来，大力倡导具有“硬科技”和“卡脖子”技术的科创企业申报上市。抛弃传统人工判断、利用机器学习将企业自动识别挂靠到某一行业是一个新挑战。首先，本文通过对 5 种分词工具的分词效果进行了比较，发现在公开数据集上分词工具 PKUseg 具有最好的分词效果。然后基于企业招股说明书，本文通过构建词袋模型，使用 5 种不同中文分词工具、4 种向量构建方法，并采用 5 折交叉验证作为实验方案，进行了企业挂靠科创板行业算法的探索与研究。最后通过实验验证“PKUSeg 词性标注 + 方法二”组合是该挂靠算法的最优解，在前 1 到 3 个行业标签下一级挂靠准确度分别达 88.11%，90.30%，90.30%。

关键词：科创板；企业挂靠；词袋模型

1 背景

自科创板正式开板以来，作为我国金融供给侧结构性改革的重要举措，科创板不仅为充满生机的科技创新企业提供直接融资支持，更成为一片“试验田”。科创板聚焦于战略新兴产业领域，摒弃了传统 IPO 制度的盈利门槛要求，强化科创属性为导向，更有利于创新型、成长型企业的 IPO 制度安排的包容性，更加适合新技术、新产业、新业态、新模式的科技创新企业上市。

强调科创板上市企业的科创属性，是科创板立足“硬科技”本位的基本体现，为企业“硬科技”水平和科创成色的界定提供了基准锚，为科创板企业上市申报指明了方向。通常情况下，需要对企业的主营业务进行深入了解，明确其主营业务面对某一特定行业和技术，再深入了解企业的科创属性高低。此时，通过大量人工的学习、识别、标注，纵然可以将企业挂到对应的产业链上，但人力成本和知识储备要求很大。如何抛弃传统人工方法、利用机器学习将企业自动归属到某一行业是一个新挑战。基于此，本文尝试应用词袋模型，利用企业招股说明书，在实现企业挂靠科创板行业方面进行算法研究和实践。最后通过在已有企业招股说明书上进行实验，验证算法的可行性。

本文的文章结构如下：第二章介绍词袋模型与分词工具的技术现状；第三章详细阐述了一级行业挂靠算法原理；第四章进行分词工具的比较和行业挂靠算法实现的交叉验证实验以及结果分析；第五章为思考和展望。

2 技术现状

2.1 词袋模型

自然语言处理（Natural Language Processing，NLP）面临的数据通常是非结构化的杂乱文本，必须进行处理将文本数据转化为数字（比如向量）才可以用于机器学习算法。词袋模型就是一种流行且实用的文本表示方法。

词袋模型（Bag of Words Model）[1] 将文本中的每个单词都视为独立的特征，忽略它们之间的顺序和语法。在词袋模型中，文本被表示为一个固定大小的向量，其中向量的每个维度对应于一个词，并且该维度的值表示该词在文本中出现的次数。如果一个单词在文本中没有出现，则对应的权值为 0。

词袋模型最直接的应用体现在，将两篇文本通过词袋模型变为向量模型，通过计算向量的余弦距离来计算两个文本间的相似度。不妨设通过词袋模型得到向量 A 和 B，记为 A = [a1，a2，a3，a4，a5，a6，a7]，B = [b1，b2，b3，b4，b5，b6，b7]，则余弦距离表示为：

$$cos(A,B)=\frac{A\cdot B}{\|A\|_2\cdot\|B\|_2}=\frac{\sum_{i=1}^{7}(a_i\cdot b_i)}{\sqrt{\sum_{i=1}^{7}(a_i\cdot a_i)}\cdot\sqrt{\sum_{i=1}^{7}(b_i\cdot b_i)}},\quad(1)$$

由此可得以上三个句子的相似度为 cos(A,B) = 0.6708，cos(A,C)=0.8660，cos(B,C)=0.7746。 通常来说，相似度越高，文本内容越相近。因此可以认为句子 1 和句子 3 最为相似。

在行业挂靠任务中，对不同行业语料分词生成行业词袋，用于招股书内容匹配对应行业。根据实际应用的需要，频数可以替换为 one-hot 编码、TF-IDF 等，而计算相似度的方法也可使用欧氏距离、皮尔逊相关系数 [2] 等。但是直接进行分词会存在大量重复且无关的词汇，比如助词、介词等，极大的影响到文本向量的生成和相似度的计算。所以我们需要使用分词工具，进行词法分析（词性标注或命名体识别）[3]，根据词性或词语类别配合正则表达式和停用词库筛选所需的关键词，以保证构建行业词袋的质量和匹配结果的精准。

2.2 分词工具

在人机自然语言交互中，成熟的中文分词算法能够达到更好的自然语言处理效果，帮助计算机理解复杂的中文语言。因为我们的任务需求，词性标注和命名体识别是非常重要的分词任务，需要借助成熟的分词工具实现。下表所列是近几年成熟且知名的开源分词工具及其相关信息。

3 基于词袋模型的行业挂靠实现

在我们的研究中，拟上市企业招股书挂靠行业实际上是文本相似度的计算问题。换言之，我

表 1：开源分词工具及相关信息表

开源分词工具	分词	词性标注	命名体识别	自定义词典	自训练模型	特点
jieba	√	√	×	√	×	速度快，支持精确模式、全模式、搜索引擎模式三种分词模式。
PKUSeg	√	√	×	√	√	提供多领域的个性化的预训练模型，比如新闻、医药、旅游等。
THULAC	√	√	×	√	√	使用目前世界上规模最大的人工分词和词性标注中文语料库。
HanLP	√	√	√	√	√	1）面向生产环境的多语种自然语言处理工具包； 2）功能完善、性能高效、架构清晰、语料时新； 3）但是单次请求的语料数量有限制。
LTP	√	√	√	√	√	1）一整套中文语言处理系统，包括词法等6项中文处理核心技术； 2）但是命名体识别仅能识别人名、地名、机构名。
PaddleNLP	√	√	√	√	√	1）可实现包括但不仅限于中文分词、词性分析等最全的中文任务。 2）覆盖自然语言理解与自然语言生成两大核心应用。 3）快速模式=词性标注，精确模式=命名体识别。

们想通过计算新招股说明书与以前已有各行业文本内容的相似度来确定该企业所属行业，算法总体流程图如图 1。接下来详细介绍词袋构建和计算的具体细节。

3.1 词袋构建

初始词袋的构建是行业挂靠任务的关键。基于《战略性新兴产业分类（2018）》文件，并对标科创板六大行业，我们将初始语料库分为新一代信息技术、高端装备制造、新材料、新能源、生物医药、节能环保六大行业，其中新能源汽车归属于节能环保。构建初始词袋的语料来源于：1）《战略性新兴产业分类（2018）》文件中一级产业下的二级产业名称、三级产业名称、四级产业名称、国民经济行业名称、重点产品和服务名称；2）公开披露已上市并有《战略性新兴产业分类（2018）》标注的科创板企业招股说明书中的相关文字描述。招股说明书中的相关文本主要包括："第二节 概述"中的"发行人主营业务经营情况"；"第六节 业务与技术"中的"发行人主营业务及主要产品和服务情况"、"公司核心技术情况"、"发行人销售情况和主要客户"、"发行人原材料采购

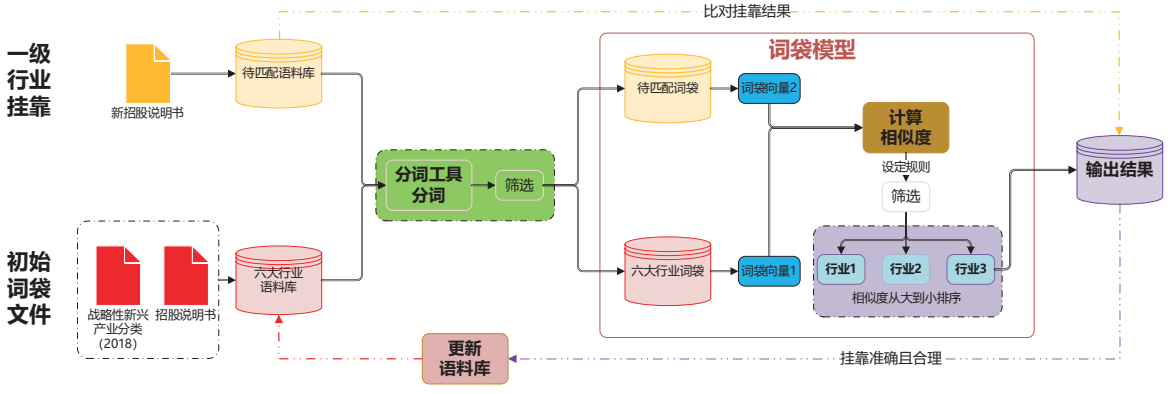


图 1：一级行业挂靠算法流程图

和主要供应商”等。待匹配词袋的语料库同样取自招股说明书的该部分文本。

接下来使用相同的分词和筛选规则处理语料库，得到基于已有文件的初始行业词袋和根据新文件生成的待匹配词袋。我们主要使用的 6 种分词工具如表 3 中所列，不同的分词工具的标注类型有略微的不同，但总体上对分词结果的筛选依照以下的规则：

1）使用正则表达式去除所有的标点符号和纯数字内容。有些像 5G、3D 等数字与中英文组合的词语是行业专有名词需要放入词袋，但纯数字和符号就没有任何区分行业的意义。

2）保留名词和动词词性的相关词汇，去除所有无意义的助词、语气词、连词、介词等。

3）使用命名体识别时除了以上俩条，还可以对名词进行进一步的筛选。以表 2 为例，还应该去除所有的人物类、地点类、场所类、文化类、饮食类、组织机构类等名词类别，保留词汇用语、场景事件、信息资料、个性特征、作品类、术语类、物体类、生物类、药物类、疾病损伤类。

4）使用停用词库。停用词（Stop Words）库是人为设定的一个字典，如果分词得到词汇存在于停用词库中，则其不进入词袋。我们所用停用词库主体是中文停用词表、哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词库等众多通用停用词库的并集，并在多次测试后，我们在此并集的基础上增添了若干词汇。

总而言之，我们选词的出发点就在减少噪声的基础上尽可能保留行业相关词汇，生成具有行业代表性且词汇丰富的行业词袋。这既保证了不同词袋之间的区分度较大，通过相似度计算可以轻松挂靠行业；又保证了在招股说明书的内容较少、描述较独特、用词较特殊时，也可以有足够的词参与匹配。

3.2 计算相似度

在构建好初始词袋和待匹配词袋后，我们需要将词袋转化为向量才可以进行文本相似度的计算。本文第二节技术现状中介绍了在某一词袋内句子向量的生成和相似度的计算，但是该方法并不能用于词袋与词袋之间相似度的计算。因此，基于该种方法思路，我们设计了构建词袋向量的方法，用该向量进行词袋与词袋之间相似度的计算。

假设某一行业的初始词袋 WordBag1 包含 k 个词汇 $\{word_{11}, word_{12}, \ldots, word_{1k}\}$，每个词汇的词频为 $\{num_{11}, num_{12}, \ldots, num_{1k}\}$，词频的数字位数极大多为十位到千位。现对某一新招股说明书分词筛选处理后，得到待匹配词袋 $WordBag_0$，该词袋包含 m 个词汇 $\{word_{21}, word_{22}, \ldots, word_{2m}\}$，每个词汇的词频为 $\{num_{21}, num_{22}, \ldots, num_{2m}\}$，词频的数字位数为个位或十位。我们设计的方法为：

方法一：不妨将 $WordBag_1$ 的词频表示为词袋向量 $Vecbag_1 = [num_{11}, num_{12}, \ldots, num_{1k}]$。若 $WordBag_1$ 中的词汇存在于 $WordBag_0$ 中，则用 WordBag0 的词频替换该词汇在词袋向量 $Vecbag_1$ 中的值；若 $WordBag_1$ 中的词汇不存在于 $WordBag_0$ 中，则令该词汇在词袋向量 $Vecbag_1$ 中的值为 0（此时即是词频为 0）。由此得到 $WordBag_0$ 的词袋向量 $Vecbag_0 = [num_{21}, num_{22}, \ldots, num_{2k}]$，实质上是由 $Vecbag_1$ 对应位置词汇的词频替换得到。

方法二：先将 $WordBag_1$ 与 $WordBag_0$ 合并后去重，得到两个词袋的并集 $WordBag_{all}$。不妨设这个并集包含 n 个词汇 $\{word_{31}, word_{32}, \ldots, word_{3n}\}$，其中 $n \geqslant k$ 且 $n \geqslant m$，$WordBag_{all}$ 的向量表示为 $Vec_{all} = [word_{31}, word_{32}, \ldots, word_{3n}]$。$WordBag_1$ 和 $WordBag_0$ 作为这个并集的两个子集，各包含有该并集中的部分词汇，故在 Vec_{all} 中使用词频填充替换对应位置的词汇，得到 n 维词袋向量 $Vecbag_1 = [num_{11}, \ldots, 0, \ldots, num_{1k}, \ldots, 0, \ldots]$、$Vecbag_0 = [\ldots, num_{21}, \ldots, 0, \ldots, num_{2m}, \ldots, 0]$。

为了使以上方法得到的结果更加直观，下面进行举例说明。假设 $WordBag_0$ 包含词汇 $word_1$、$word_2$、$word_3$、$word_4$，词频为 num_{11}、num_{12}、num_{13}、num_{14}，$WordBag_2$ 包含词汇 $word_1$、$word_2$、

$word_3$、$word_5$，词频为 num_{21}、num_{22}、num_{23}、num_{25}。其中 $word_1$、$word_2$、$word_3$ 同时存在于两个词袋中，而 $WordBag_1$ 中的 $word_4$ 不存在于 $WordBag_0$ 中，$WordBag_0$ 中的 $word_5$ 不存在于 $WordBag_1$ 中。根据方法一，词袋向量表示为 $Vecbag_1 = [num_{11}, num_{12}, num_{13}, num_{14}]$、$Vecbag_0 = [num_{21}, num_{22}, num_{23}, 0]$；根据方法二，$WordBag_{all}$ 词袋向量表示为 $Vec_{all} = [word_1, word_2, word_3, word_4, word_5]$，$WordBag_1$ 和 $WordBag_0$ 向量表示为 $Vecbag_1 = [num_{11}, num_{12}, num_{13}, num_{14}, 0]$、$Vecbag_0 = [num_{21}, num_{22}, num_{23}, 0, num_{25}]$。根据公式 (1)，这两种词袋向量所求余弦相似度分别为：

$$cos_1(Vec_{bag1}, Vec_{bag2}) = \frac{num_{11} \cdot num_{21} + num_{12} \cdot num_{22} + num_{13} \cdot num_{23}}{\sqrt{num_{11}^2 + num_{12}^2 + num_{13}^2 + num_{14}^2} \cdot \sqrt{num_{21}^2 + num_{22}^2 + num_{23}^2}}, \quad (2)$$

$$cos_2(Vec_{bag1}, Vec_{bag2}) = \frac{num_{11} \cdot num_{21} + num_{12} \cdot num_{22} + num_{13} \cdot num_{23}}{\sqrt{num_{11}^2 + num_{12}^2 + num_{13}^2 + num_{14}^2} \cdot \sqrt{num_{21}^2 + num_{22}^2 + num_{23}^2 + num_{25}^2}}。 \quad (3)$$

很明显，这两种方式计算得到的余弦相似度并不完全相同，公式 (3) 的结果小于公式 (2)。实际上，相比于方法一，方法二实际考虑了待匹配词袋种所有词对相似度的影响，但是会因为分词筛选得到大量噪音词从而降低了相似度。另外，由于待匹配词袋来源于全新的招股说明书，其中词汇出现的频率不可控，这会极大地影响相似度结果。因此我们参考 one-hot 编码，在方法一和方法二的基础上，仅仅使用 1 与 0 构成待匹配词袋向量：

方法三：此时我们不考虑待匹配词袋 WordBag0 中词频的影响，只关注是否出现某些关键词汇。不妨将 $WordBag_1$ 的词频表示为词袋向量 $Vecbag_1 = [num_{11}, num_{12}, \dots, num_{1k}]$。若 $WordBag_1$ 中的词汇不存在于 $WordBag_0$ 中，则令该词汇在词袋向量 $Vecbag_1$ 中的值为 1；若 $WordBag_1$ 中的词汇不存在于 $WordBag_0$ 中，则令该词汇在词袋向量 $Vecbag_1$ 中的值为 0。由此得到 $WordBag_0$ 的 k 维词袋向量 $Vecbag_0 = [\dots, 1, \dots, 0, \dots, 1, \dots]$。

方法四：同样避免待匹配词袋 $WordBag_0$ 中词频的影响。先将 $WordBag_1$ 与 $WordBag_0$ 合并后去重，得到两个词袋的并集 $WordBag_{all}$。不妨设这个并集包含 n 个词汇 $\{word_{31}, word_{32}, \dots, word_{3n}\}$，其中 $n \geqslant k$ 且 $n \geqslant m$，并集词袋向量表示为 $Vec_{all} = [word_{31}, word_{32}, \dots, word_{3n}]$。$WordBag_1$ 和 $WordBag_0$ 作为这个并集的两个子集，各包含有该并集中的部分词汇。对于 $WordBag_1$，在 Vec_{all} 中使用的词频填充替换对应位置的词汇，得到 n 维词袋向量 $Vecbag_1 = [num_{11}, \dots, 0, \dots, num_{1k}, \dots, 0, \dots]$；对于 $WordBag_0$，在 Vec_{all} 中使用 1 填充替换对应位置的词汇，在 Vec_{all} 中不包含于 $WordBag_0$ 的词汇则用 0 填充替换，得到 n 维词袋向量 $Vecbag_0 = [\dots, 1, \dots, 0, \dots, 1, \dots, 0]$。

同样的，以方法一和方法二所举例子为例。根据方法三，词袋向量表示为 $Vecbag_1 = [num_{11}, num_{12}, num_{13}, num_{14}]$、$Vecbag_0 = [1, 1, 1, 0]$；根据方法四，$WordBag_1$ 和 $WordBag_0$ 的词袋向量表示为 $Vecbag_1 = [num_{11}, num_{12}, num_{13}, num_{14}, 0]$、$Vecbag_0 = [1, 1, 1, 0, 1]$。根据公式 (1)，利用这两种词袋向量所求余弦相似度分别为：

$$cos_3(Vec_{bag1}, Vec_{bag2}) = \frac{num_{11} + num_{12} + num_{13}}{\sqrt{num_{11}^2 + num_{12}^2 + num_{13}^2 + num_{14}^2} \cdot \sqrt{3}}, \quad (4)$$

$$cos_4(Vec_{bag1}, Vec_{bag2}) = \frac{num_{11} + num_{12} + num_{13}}{\sqrt{num_{11}^2 + num_{12}^2 + num_{13}^2 + num_{14}^2} \cdot \sqrt{2}}。 \quad (5)$$

综上所述，以上提出的四种方法的计算结果不尽相同而且各有优劣，如何选择要综合考虑词频、词汇等多方面因素。所以在我们的实践中，如何分词和筛选生成高质量词袋仍是最关键的步骤，相似度的计算和结果的筛选都只是可替换的过程。

3.3 行业挂靠的实现

利用上一小节提出的使用词袋向量计算余弦相似度的方法，我们将得到待匹配词袋与六大行业初始词袋的相似度结果。出于对一个公司存在多个方面的业务和产品的考量（比如某公司可能既生产芯片，又生产制造芯片的设备，可能既属于新一代信息技术又属于高端装备制造），有必

要依据相似度对六大行业进行降序排序，使用一定的规则进行筛选输出多个挂靠结果。

不妨设降序排序后的相似度为 cos_1、cos_2、cos_3、cos_4、cos_5、cos_6，对应于行业 Lab_1、行业 Lab_2、行业 Lab_3、行业 Lab_4、行业 Lab_5、行业 Lab_6。首先，将相似度最高的行业，即 cos_1 对应的 Lab_1，作为该公司的行业标签 $label_1$。其次，对于非相似度最高的行业，若它们跟 Lab_1 的相似度差值不大，则可以作为该公司的行业标签 $label_2$ 或行业标签 $label_3$。在我们的设计中，最多只能记录到 $label_3$，而相似度差值不大则需满足以下公式：

$$\frac{cos_1 - cos_i}{cos_1} \le a \ , \ i = 2,3,4,5,6 \ 。 \quad (6)$$

其中 a 是超参数，它是用来筛选标签二和标签三的阈值，在后续的实验中，统一设置 a=0.1。在经过以上计算、筛选的流程后，得到的一个或多个行业标签，即是我们所需的拟上市企业的行业挂靠结果。

4 工作成果

4.1 实验设置

前面章节详细介绍了现有的技术和本文提出的行业挂靠算法。对于行业挂靠算法的验证实验以及结果分析是本章节的核心内容。另外，在行业挂靠算法中，合理且准确的挂靠结果依赖于分词获取的高质量词袋。所以本章节主要进行两组对比实验：1）在公开标准数据集上各分词工具分词实验；2）基于招股说明书的行业挂靠算法验证实验。下面将详细介绍实验数据、算法设置、评价指标等详细实验设置。

4.1.1 实验数据来源

对于第一组实验，我们将采用 SIGHAN 2005 国际中文自动分词评测的 PKU 和 MSR 数据集进行分词的测试。该数据集由中国微软研究所、北京大学、香港城市大学、台湾中央研究院联合发布，用以进行中文分词模型的训练与评测，其中 PKU、MSR 是简体中文数据集。

对于第二组实验，我们提取了 500 份带有一级行业标签的上市企业招股说明书文件。这些带标签的文件将被用于构建初始词袋和通过交叉验证测试挂靠准确性。

4.1.2 算法设置

第一组实验主要对比表 3 中 jieba、PKUSeg、THULAC、LTP、PaddleNLP 这五种开源分词工具的分词效果。由于 HanLP 单次请求的语料数量有限制，在实验和实际应用中有着明显的桎梏，故不考虑将该工具加入本文的对比试验中。

第二组实验则基于第一组实验的五种分词工具，使用词性标注或命名体识别筛选分词结果，利用 3.2 小节中提出的四种词袋向量生成方法，并采用 5 折交叉验证的方式将招股书文件划分为训练集和测试集来设置实验，进行挂靠算法的验证。除了单独使用词性标注或命名体识别进行筛选外，还设置了串行应用于 LTP 和并行应用于 PaddleNLP。此外，PKUseg 工具中有专门针对医疗领域的分词模型（我们将其记为 PKUseg-m），也一并进行了测试。

4.1.3 评价指标

在机器学习中，基于二分类问题的混淆矩阵（见图 2），模型的评价指标 [5] 主要包括但不仅限于准确率（Accuracy），精准率（Precision），召回率（Recall），F1 值（F1-Score）：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+TN}, \ Precision = \frac{TP}{TP+NP}, \ Recall = \frac{TP}{TP+FN}, \ F1-Score = \frac{2*Precision*Recall}{Precision+Recall}. \quad (7)$$

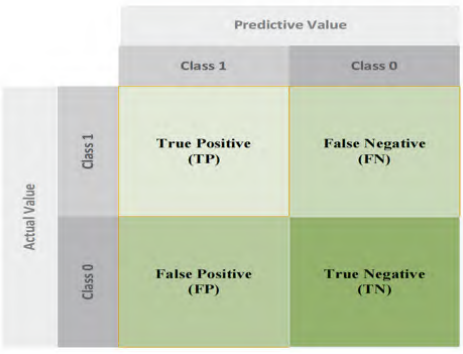


图 2：混淆矩阵

其中 F1 值实际上是精准率和召回率的加权平均。本文采用精准率，召回率和 F1 值来作为实验的评价指标。

必须说明的是，第一组实验作为自然语言处理中的分词测评，所用的分词性能评价指标实质上与分类问题有所区别。假设对于长为 n 的字符串，分词结果是一系列单词。设每个单词按照其在文中的起止位置可记作区间 [i, j]，其中 0 ≤ i ≤ j ≤ n。则某句子分词的标准答案为区间集合 A，某次分词结果所有词语构成的区间集合为 B，此时精准率和召回率应该表述为：

$$Precision = \frac{|A \cap B|}{|B|}, \quad Recall = \frac{|A \cap B|}{|A|}。 \tag{8}$$

这里的取模代表的是集合内区间的个数。同时，完成分词所需时间也应该作为工具分词效率的考察指标。

另外，第二组实验可以认为是多分类问题，我们统一使用准确率进行评价，并取每组 5 折交叉验证的平均值作为实验结果。最后出于企业的业务产品可能存在多个行业交叉情况的考量，我们记录了每组实验前 1 到 3 个行业标签的匹配准确率，分别用 $label_1$，$label_2$，$label_3$ 分组区别。

4.2 结果与分析

4.2.1 在公开标准数据集上各分词工具分词效果

如表 4 所示是在 PKU 和 MSR 两种公开数据集下五种分词工具的分词结果。明显可见，在分词效率方面，jieba 的分词效率最高，PKUSeg 的分词效率与其差距较小。至于分词性能方面的考量，在 PKU 数据集上 PKUSeg 具有最高的精准率和 F1 值，THULAC 具有最高的召回率，但是 THULAC 的分词效率要远远低于 PKUSeg。在 MSR 数据集上 PaddleNLP 具有最高的精准率和 F1 值，PKUSeg 具有最高的召回率和第二高的精准率和 F1 值，而且 PKUSeg 的分词效率要远高于 PaddleNLP。综上所述，PKUSeg 在保证了分词性能的前提下，仍具有非常高的分词效率，说明 PKUSeg 的综合分词效果最好，在实际应用中可以优先考虑使用 PKUSeg 分词。在不考虑所需时间的情况下，THULAC 和 LTP 也不失为较好的分词工具。

4.2.2 基于招股说明书的行业挂靠算法的验证结果

如表 5 所示是利用不同分词工具、不同词袋向量计算方法在 5 折交叉验证的设置下进行一级行业挂靠测试的结果。纵向来看，分词工具 PKUSeg 的第一个标签 label1 的准确率最高，分词工具 THULAC 的前两个和前三个标签的准确率最高。横向来看，使用 2 个和 3 个行业标签时，方法四的挂靠准确率最高，分别在 9 组实

表 2：在公开标准数据集上各分词工具分词结果表

数据集	评价指标	jieba	PKUSeg	THULAC	LTP	PaddleNLP
PKU	精准率	0.8597	**0.9489**	0.9250	0.9379	0.8693
	召回率	0.7997	0.9202	**0.9285**	0.9209	0.7973
	*F*1值	0.8286	**0.9343**	0.9268	0.9293	0.8318
	所需时间（秒）	**6.2387**	8.4312	227.2746	106.3537	39.5753
MSR	精准率	0.8199	0.8623	0.8326	0.8301	**0.8955**
	召回率	0.8212	**0.8815**	0.8793	0.8592	0.8702
	*F*1值	0.8205	0.8718	0.8553	0.8444	**0.8827**
	所需时间（秒）	**4.9331**	9.4591	243.5654	119.3604	46.3125

表 3：一级行业挂靠测试的结果表

方法 \ 分词工具		Jieba 词性标注			PKUSeg 词性标注			PKUSeg-m 词性标注		
		$label_1$	$label_2$	$label_3$	$label_1$	$label_2$	$label_3$	$label_1$	$label_2$	$label_3$
方法一	准确率	0.7706	0.8216	0.8234	0.8014	0.8400	0.8439	0.7728	0.8252	0.8269
方法二	准确率	**0.8239**	0.8615	0.8652	**0.8343**	0.8618	0.8657	**0.8106**	0.8520	0.8574
方法三	准确率	0.7686	0.8527	0.8691	0.7563	0.8553	0.8716	0.7500	0.8474	0.8674
方法四	准确率	0.8053	**0.8917**	**0.8992**	0.7965	**0.8818**	**0.9016**	0.7949	**0.8884**	**0.9082**
方法 \ 分词工具		THULAC 词性标注			LTP 词性标注			LTP 串行		
		$label_1$	$label_2$	$label_3$	$label_1$	$label_2$	$label_3$	$label_1$	$label_2$	$label_3$
方法一	准确率	0.6847	0.7767	0.8112	0.7738	0.8259	0.8278	0.7738	0.8259	0.8278
方法二	准确率	**0.8113**	0.8905	0.9142	**0.8190**	0.8563	0.8599	**0.8190**	0.8563	0.8599
方法三	准确率	0.7094	0.8467	0.8872	0.7683	0.8496	0.8730	0.7683	0.8496	0.8730
方法四	准确率	0.7833	**0.9063**	**0.9278**	0.8058	**0.8758**	**0.8957**	0.8058	**0.8758**	**0.8957**
方法 \ 分词工具		PaddleNLP 词性标注			PaddleNLP 命名体识别			PaddleNLP 并行		
		$label_1$	$label_2$	$label_3$	$label_1$	$label_2$	$label_3$	$label_1$	$label_2$	$label_3$
方法一	准确率	0.7547	0.8106	0.8145	0.7238	0.7862	0.8027	0.7445	0.7860	0.7899
方法二	准确率	**0.8003**	0.8449	0.8523	**0.8051**	0.8486	0.8542	**0.7785**	0.8311	0.8350
方法三	准确率	0.7321	0.8368	0.8588	0.7402	0.8426	0.8534	0.7373	0.8172	0.8317
方法四	准确率	0.7725	**0.8582**	**0.8747**	0.7768	**0.8524**	**0.8684**	0.7656	**0.8411**	**0.8501**

验中取到了 9 次和 9 次最大值；仅仅考察行业标签 label1 时，方法二的挂靠准确率最高，在 9 组实验中取到了 9 次最大值。综合来看，其中使用“PKUSeg 词性标注 + 方法二”的组合得到了最高的挂靠准确率，且 PKUSeg 的分词效率和分词性能都比较优秀，故该组合可以作为目前一级行业挂靠算法的最优解。

值得一提的是，LTP 的串行并未取得更好的结果；PaddleNLP 的并行甚至得到了更差的实验结果。除此之外，结合图 3 分析，实际上前两个

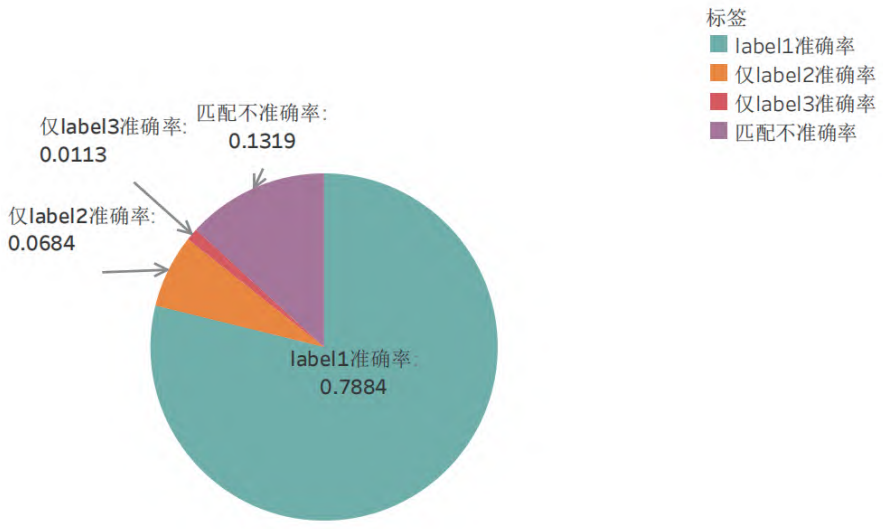


图 3：行业标签 1 到 3 的平均准确率占比

行业标签的匹配准确度明显比仅仅第一个行业标签准确有明显提升，这正符合我们前期观察数据得到的结论，即某一企业的具体业务、产品的行业界限并不明显，可能同属多个行业。而前三个行业标签比前两个行业标签的匹配准确度提升并不显著，可能的原因在于:1）同属三个行业业务、产品的企业很少；2）实际上有些行业两两之间差距并不大，但是三三进行比较时就会突出其中某一个与其他的差别。

5 思考与展望

本文阐述了科创属性评价体系建立过程中对于企业挂靠行业算法的需求，详细介绍了词袋模型和分词工具，基于词袋模型提出了企业挂靠行业算法。通过对现有开源分词工具的对比试验和挂靠行业算法的验证实验，得到了该算法的最优方法组合“PKUSeg 词性标注 + 方法二”，该组合在前 1 到 3 个行业标签下一级挂靠准确度分别达 83.43%，88.18%，90.16%。但作为探索性的项目，本研究仍存在不足和可提升的角度。

一是数据层面。本文提出的算法所使用数据中归属不同行业的招股书文件数量有明显的区别，据此生成的词袋大小也有区别。专门进行招股书文件的标注以提升数据质量、通过词语相似度的方法扩充行业词袋以达到数量平衡等，这些都是后续可以探索和研究的工作。

二是算法层面。今年人工智能领域热门话题是通用大模型，利用通用大模型训练出针对行业挂靠问题的专用模型是一种可以尝试的思路，如何将大模型行业挂靠任务结合也值得思考和研究。

参考文献：

[1] 黄春梅，王松磊 . 基于词袋模型和 TF-IDF 的短文本分类研究 [J]. 软件工程 ,2020,23(03):1-3.

[2] Fei Gao, Weikai He, Wenhao Bi. Ensemble extended belief rule-based systems with different similarity measures for classification problems[J]. International Journal of Approximate Reasoning,2023,163:109054.

[3] 王腾阳，赵小丹，胡林 . 基于词性标注规则的马铃薯文献信息抽取方法 [J]. 科学技术与工程 ,2023,23(27):11562-11569.

[4] 袁里驰 . 基于 BiLSTM-CRF 的中文分词和词性标注联合方法 [J]. 中南大学学报（自然科学版),2023,54(08):3145-3153.

[5] Nawaf Almaskati. Machine learning in finance: Major applications, issues, metrics, and future trends[J]. International Journal of Financial Engineering,2022,09(03):2250010.

大模型驱动业务创新

基于大模型技术的网络黑嘴识别

杜威[1]、刘燕婷[1]、吴昊伦[1]、王新宇[1]、纪焘[2]、吴苑斌[1]、王晓玲[1]、王玲[3]/
[1] 华东师范大学 上海 200062
[2] 复旦大学 上海 200433
[3] 华泰证券股份有限公司 南京 210019
E-mail：52265901025@stu.ecnu.edu.cn，ybwu@cs.ecnu.edu.cn

随着数字化、网络化发展，金融行业也进入了数据浪潮之中。网络黑嘴因其泄露信息、传播虚假消息等特性对金融系统的稳定运行构成危害。本文基于大模型技术提出了一套识别网络黑嘴的系统框架，首先系统将网络黑嘴分为三类，通过公开的大模型 API 对收集到的黑嘴文本进行自动化标注，获取训练数据。然后挑选已开源的大模型作为基座，用训练数据进行微调从而让系统拥有识别网络黑嘴的能力。最后我们给出当前系统的性能与后续的优化方向。

关键词：网络黑嘴；大模型技术；自然语言处理；参数微调

1 引言

随着移动互联网和新媒体的快速发展，抖音、微信、微博、快手等舆情平台及其“自媒体”账号屡屡发生违规发布财经新闻、歪曲解读经济政策、唱衰唱空金融市场、充当“黑嘴”博人眼球、造谣传谣、敲诈勒索等违法违规行为，严重扰乱网络传播秩序，对金融体系的稳定和用户信任构成威胁。为推动经济社会持续健康发展营造良好网上舆论环境，国家网信办会同国家发展改革委、财政部、中国人民银行、证监会、银保监会等财经主管部门，多次督促指导舆情平台进行黑嘴专

项整治[1]。在2021年的“清朗·商业网站平台和自媒体违规采编发布财经类信息专项整治”行动中，处置违规自媒体账号达2929个，清理有害信息47153条，反映出网络“黑嘴”肆虐的现状与问题的严重性。相较于以往金融领域的任务，如虚假新闻，敏感言论等，网络黑嘴往往在真实的舆情的基础上添油加醋，通过诱导、断章取义、恶意解读等方式扭曲事实，这类隐式的黑嘴文本，通过以往处理虚假新闻、敏感言论等的技术难以正确识别。然而，基于人工审核的“黑嘴”识别难以应对实时产生的海量舆情数据，因此“黑嘴”的自动化、智能化识别研究意义重大。

金融网络平台上舆情信息的语义多样和动态变化特点使得传统的黑嘴识别方法难以对抗不断演化的黑嘴形式。另外，传统识别方法虽然能识别黑嘴类别，但是无法自然地解释判断依据、输出识别原因。近年来，大语言模型的发展突飞猛进，尤其是OpenAI现象级产品ChatGPT的出现，标志着大模型成为实现通用人工智能的最佳技术。大模型目前在文本理解、文本分析、文本推理等自然语言处理任务中表现出了强大的性能，展现出了其在解决金融领域网络黑嘴问题上的巨大潜力。通过学习海量金融相关数据，深度理解行业术语、市场动态和用户行为，大模型能够更准确地识别和过滤潜在的网络黑嘴信息。其强大的处理和情感分析能力使得大模型能够识别那些伪装在正常金融交流中的恶意信息，为金融从业者和用户提供更为安全的网络环境。

1.1 网络黑嘴简介

在金融领域，网络黑嘴是指那些通过网络渠道，以恶意的、虚假的、欺诈性的手段，试图获取金融机构或个人用户敏感信息、资金或其他财产的行为。国家网信办针对财经类“自媒体”账号、主要公众账号平台、主要商业网站平台财经版块、主要财经资讯平台这四类网上传播主体定义了八种重点打击违规问题[1]。网络黑嘴活动在金融领域呈现多样化和复杂性，涉及诸多形式这些网络黑嘴活动给金融领域带来了严峻的安全挑战，要求金融机构和相关当事人采用先进的网络安全技术和措施，以保护用户的资金和敏感信息。大模型等先进的人工智能技术被广泛应用于网络安全领域，助力金融行业更有效地应对这些网络黑嘴威胁。

1.2 大语言模型技术简介

随着大模型技术的不断完善，目前大模型已应用于医药、制造、能源、电力、化工、交通等多行业多领域，生成式AI赋能千行百业，将为企业创造前所未有的商业价值。以DeepMind开发的制药领域的大模型AlphaFold2为例，其以强大的分析能力在蛋白质结构预测任务上取得了惊人的进展，大大减少了医药研究开发成本。而大模型的优秀的分析推理能力非常适合用于当前海量金融舆情的文本处理，当前的大模型技术可以监测社交媒体、新闻、论坛等各类平台上的金融相关舆情。通过对大规模文本数据的分析，大模型能够识别并理解关键信息，从而迅速捕捉市场上涌现的网络黑嘴通过对这些黑嘴信息的初步分析总结，能够一定程度上地减少金融从业人员，监管人员等的工作量，提高金融行业的整体运行效率。我们将利用OpenAI发布的GPT API来辅助我们获取训练数据，之后利用开源的大模型作为基座模型训练网络黑嘴识别系统。

接下来将具体介绍网络黑嘴在本系统中的定义与分类情况。

2 网络黑嘴

2.1 网络黑嘴分类

在本系统中，基于网信办对网络黑嘴的八种定义，挑选其中三类影响恶劣的黑嘴类型进行识别：

类别一：胡评妄议、歪曲解读我财经方针政

策、宏观经济数据，恶意唱空我金融市场、唱衰中国经济等；

类别二：充当金融“黑嘴”，恶意唱空或哄抬个股价格，炒作区域楼市波动，扰乱正常市场秩序；

类别三：炒作社会恶性事件、负面极端事件，煽动悲情、焦虑、恐慌等情绪，借以推销所谓“财商课”、各类保险产品等。

系统在训练大模型识别网络黑嘴时也会根据上述分类来进行。目前系统能够对网络黑嘴进行两种粒度的分类，一是简单地判断一段文本是否属于网络黑嘴，即一个“是 / 不是”的二分类任务；二是在区分给定文本是否为网络黑嘴的基础上，对属于网络黑嘴的文本进行上述三个类别的划分。具体样例如图 1 所示。

2.2 困难与挑战

“黑嘴”歪曲解读金融政策，恶意唱空我国经济，散布恐慌情绪，企图动摇国际社会对华投资信心，妄图引发我国内金融动荡。这些都给新形势下维护金融安全带来了新挑战。

网络黑嘴采用多种巧妙的手法，包括变种、混淆和伪装等，使得传统的规则和模式匹配方法难以捕捉到其特征，这要求识别系统需要不断升级和演进，以适应新型攻击；网络黑嘴往往试图隐藏其真实意图，采用模糊语言、暗示性词汇或讽刺性表达，使得识别系统在解析语义和情感时面临挑战，这需要模型具备深层次的语义理解和上下文感知能力；大多数时候，正常的网络交流远远多于恶意的网络黑嘴行为，这导致了数据不平衡，使得模型更容易倾向于过度拟合正常样本而忽略潜在的恶意行为。

上述几点都是传统模型难以解决的问题，因此大模型的入场使得这一问题见到了曙光。大模型知识库丰富，分析、处理信息能力强的优点较好地适配了当前金融领域的这一需求，通过对金融领域舆情信息分析，大模型可以自动且全面得发现“网络黑嘴”，从而及时预警，辅助决策，减少不良影响，对维护金融安全有积极作用。

然而目前黑嘴数据集较少并且现有的黑嘴数据集几乎不满足 2.1 中三类金融领域相关的网络黑嘴定义。因此为了解决这个问题，本系统考虑将公众号等作为金融领域数据来源，并借助 ChatGPT 来标注满足定义的黑嘴数据。

3 数据准备

为了解决网络黑嘴数据缺失的问题，我们调用了 OpenAI 开放的大模型 ChatGPT 和 GPT4 的 API 接口从海量无结构文本中，生成符合系统中黑嘴识别任务输入格式的网络黑嘴训练数据。

3.1 ChatGPT 和 GPT4

ChatGPT 是由 OpenAI 发布于 2022 年 11 月

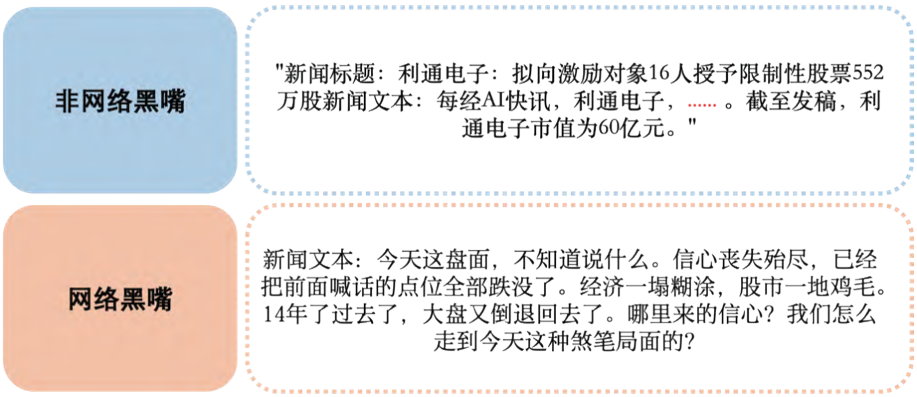


图 1：网络黑嘴样例

30 日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，因此将它用于网络黑嘴事件的识别标注。它的原理主要基于 GPT 模型结构、无监督预训练技术、微调技术、奖励模型、人类反馈的强化学习模型等技术，具体的技术细节如图 2 所示。

而 GPT4[2] 则是 OpenAI 在 2023 年 3 月 14 日发布的大模型，也是目前综合性能最为强大的语言大模型，在多个自然语言任务中接近或达到了人类水平。因此相比 ChatGPT，GPT4[2] 会对黑嘴事件的标识有更准确的判断，但是考虑到 GPT4[2] 的 API 高昂的调用成本，GPT4[2] 在本任务中仅负责对 ChatGPT 判别为黑嘴的事件进行精标，以进一步提高数据集的质量。

3.2 数据收集

我们把在微信公众号和股吧上分别采集到的 60326 条和 9983 条原始数据按新闻标题和新闻文本的格式存储为 json 文件，在去除了多余的空格缩进以及 html 标记后就得到较为结构化的文本数据集。

3.3 数据标注

完成了数据的采集工作后就需要对数据进行标注，但是对大量的数据进行人工标注的人力成本和时间成本是很高的，因此选择调用 ChatGPT 以及 GPT4[2] 等大模型的 API 接口来自动化地完成数据预标注流程。具体流程如图 3 所示，包含提示词设计、数据格式转换、长文本过滤和不重复均匀采样等步骤。

1）提示词设计。为了尽可能地提升数据标注的质量，首先需要设计相关的提示词来让大模型的输出更符合我们的预期。提示词设计的指导原则之一就是具体，即必须要明确地定义任务并说明期望的输出形式。因此在黑嘴识别任务中，需要明确地交代网络黑嘴的类别以及相对应的分类标准并提前告知模型针对每种类别的期望回复格式。另外，为了增强分类结果的可信度和可解释性，还要求模型对分类结果进行原因分析。

2）数据格式转换。在明确了输入模版之后，

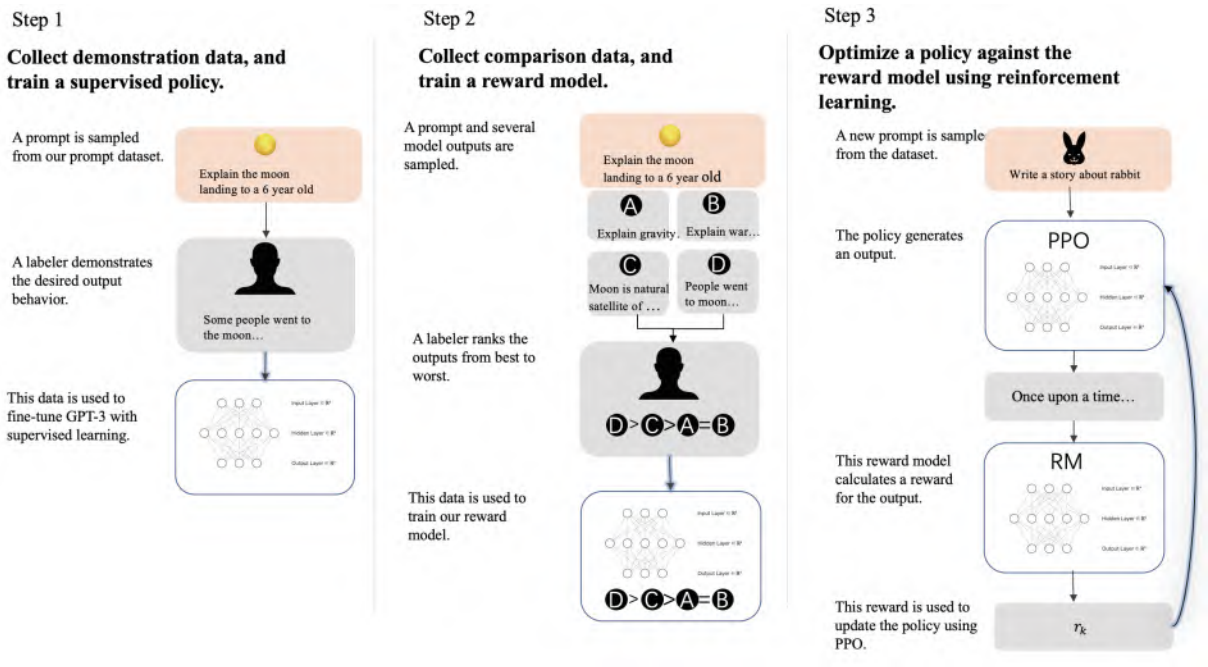


图 2：GPT 技术细节

需要对原始的数据格式进行转换以符合设计的模版，也就是要将每条 json 数据中的“新闻标题”和“新闻文本”标签下存储的文本进行拼接，然后作为判别语料填入提示词模版中。

3）长文本过滤。所有大模型可接受的输入 token 的总长度是有限的，并且不同模型可接受的最大 token 长度都不同。ChatGPT-3.5-turbo 的最大可接受 token 长度是 4K，而 GPT4[2] 的最大可接受 token 长度是 8K。因此需要对输入请求的长度进行判断，过滤掉过长的请求。

4）不重复均匀采样。完成了上述的数据处理流程后，为了采集到尽可能多的不同长度的数据，需要在不重复的基础上对不同长度的数据进行均匀采样。

5）数据标注。将预处理、采样后的数据及提示词等信息拼接成 ChatGPT 等所需的格式请求获取标注的结果。

4 模型设计

在完成了数据的标注之后下一步就要选择对应的基座模型。在本系统中，我们以大模型为基座设计、训练的”网络黑嘴”识别模型，其整体流程如图 4 所示。虽然大模型基座拥有良好的问答能力，但其金融相关的训练数据较少，不能很好地回答金融领域相关问题，更不能有效识别金融”网络黑嘴”。我们利用 SFT（Supervised Fine-Tuning，监督微调）的方法对模型进行微调，使其拥有识别”网络黑嘴”的能力。本节采用了 ChatGLM2[6]、Baichuan2[5] 等中文大模型作为基座，采用不同的方法对模型进行微调获得”网络黑嘴”识别模型。

4.1 基于 Baichuan2-7B-chat 的模型

Baichuan2[2] 是百川智能继 Baichuan 系列后推出的新一代开源大语言模型。在多个权威的多语言通用领域 benchmark 上取得了相同模型尺寸下的最佳效果。其中，Baichuan2-192K 已开启内测，以 API 调用的方式开放给百川智能的核心合作伙伴，已经与财经类媒体及律师事务所等机构达成了合作，将 Baichuan2-192K 全球领先的长上下文能力应用到了传媒、金融、法律等具体场

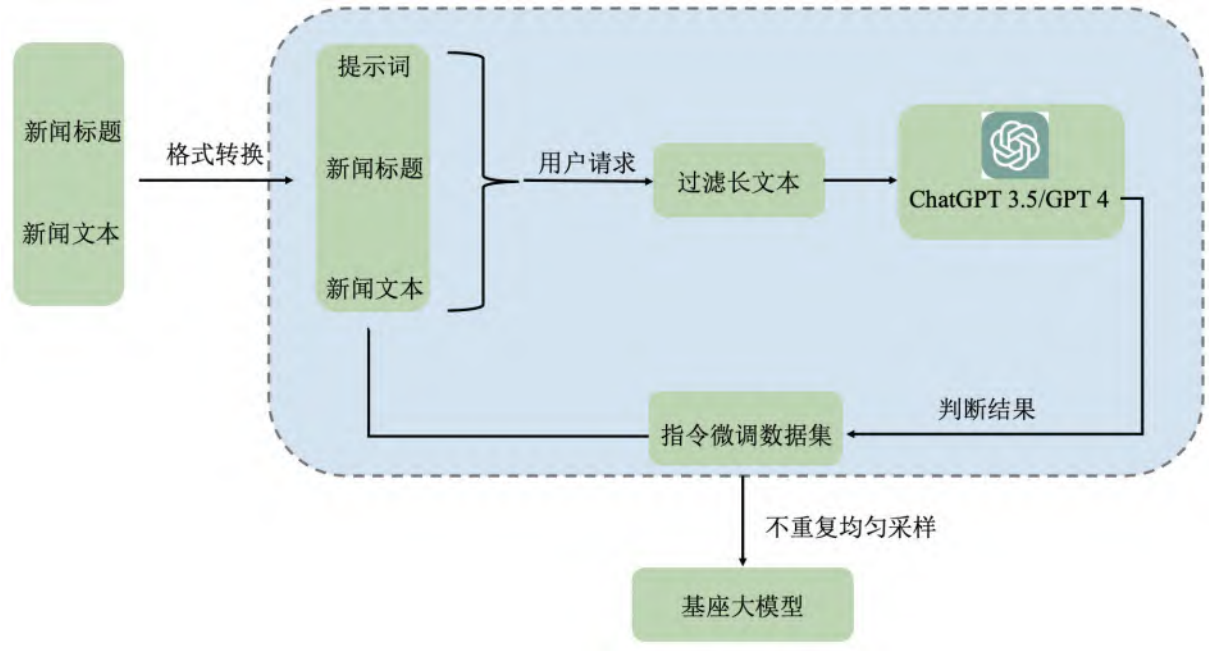


图 3：数据标注流程

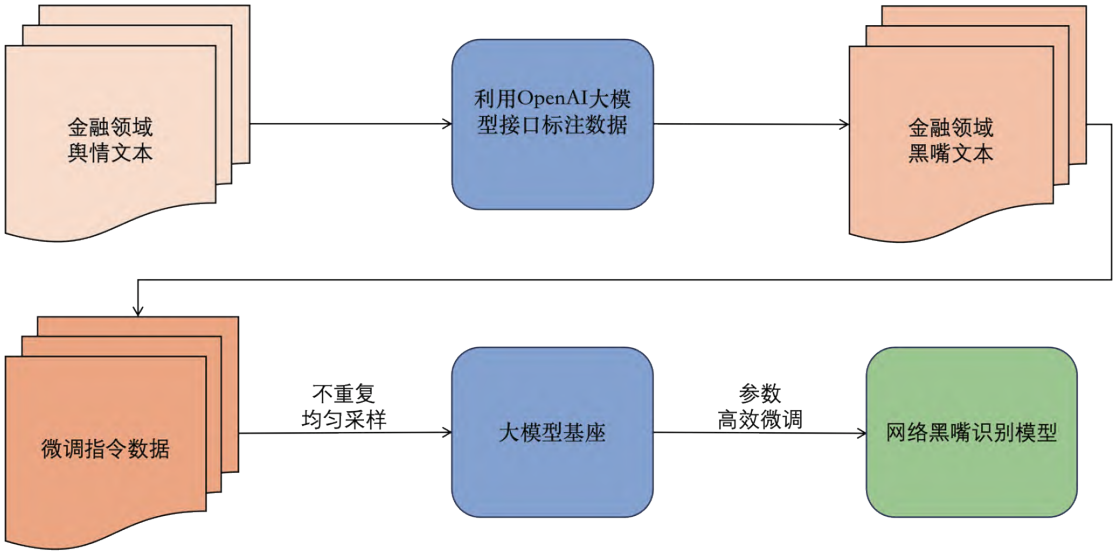


图 4：模型训练的整体流程

景当中。因此，我们认为 Baichuan 系列模型是一个在金融文本分析领域非常有潜力的基座模型，因而选取其作为网络黑嘴识别任务的其中一个基座模型，考虑到训练和存储开销，选用 Baichuan2-7b-chat[2] 进行指令微调。

4.2 基于 ChatGLM2-6B 的模型

ChatGLM2[6] 是一个开源的、支持中英双语的对话语言模型，基于 General Language Model (GLM)[7] 架构。中英双语进行训练，对中文的支持。目前 ChatGLM2 在中文的各项学术测试集的性能都位居开源数据集的榜首，甚至部分任务的性能超过了 GPT4，因此 ChatGLM2 也被认为是当下最适合用于处理中文文本的开源大模型之一，且部署简单，因此挑选 ChatGLM2 作为基座模型是非常合适的。并且 ChatGLM 系列模型的更新频率极高，后续模型更新后也能够及时迭代系统中的 ChatGLM2 基座模型，提升模型性能。考虑到存储和开销，本系统的选择 ChatGLM-6B 作为基座模型之一。

5 模型训练

完成数据标注和模型设计后，下一步利用 SFT 对大模型基座进行微调作，使其学习到金融“网络黑嘴”的相关知识，进而获取判别“网络黑嘴”的能力。然而，作为基座的大模型如 Baichuan2-7B-chat, ChatGLM2-6B 等参数规模达十亿级，在低资源硬件上对其行全量参数微调不再可行。因此本文采用参数高效的方法进行微调，在保证微调效果的基础上，降低资源开销。

5.1 Lora 技术

LoRA[3] 是常见的参数高效微调方法。通过 LoRA[3] 技术，可以将原参数矩阵分解为低秩矩阵的乘积，从而大大降低需要微调的参数量。在训练时，为了进一步降低存储开销并提高训练效率，仅将 LoRA[3] 应用到模型一部分参数矩阵的更新中而保持其余大部分参数矩阵不改变。

LoRA[3] 基本原理如图 5 所示，冻结预训练好的模型权重参数，在冻结原模型参数的情况下，通过往模型中加入额外的网络层，并只训练这些新增的网络层参数。由于这些新增参数数量较少，这样不仅微调的成本显著下降，还能获得和全模型微调相近的效果。

相关研究表明高维参数矩阵的信息可以通过学习它的一个低维嵌入来近似表示，因此在 LoRA 的训练过程中可以将需要微调的参数矩阵

的秩设定为一个较小的值，从而大大减少了需要微调的参数量。另外，LoRA 相比较其他的微调技术还具有模块化、可迁移的优点，也就是说通过更换在不同下游任务上微调得到的 LoRA 矩阵，就可以将模型应用到不同的下游任务中，这就便于不同下游任务之间的切换，及时对于同一个下游任务也能很方便地进行模型迭代。基于以上特点，我们选择使用 LoRA 对 Baichuan2-7b 模型进行微调。具体来说，在黑嘴识别系统中，我们将 LoRA 矩阵的秩设定为 8，用于计算 Baichuan2 模型结构的全连接网络参数以及 attention 矩阵参数的。训练完成后，通过合并 LoRA 微调的矩阵和冻结的原模型参数，就得到了微调后的 Baichuan2-7b 模型。

5.2 P-tuning 技术

基于连续 prompt 的技术也是常用的参数高效微调方法。相较于以往的连续 prompt 技术，如 P-tuning[4] 方法，P-tuning v2 在模型参数量小于 10B 的情况下微调效果更好。我们的基座模型均小于 10B，因此我们选用 P-tuning v2 进行微调。

如图 6 所示，P-tuning v2 方法是一种连续 prompt 方法，即将 prompt 作为特殊的标记并转换为 embedding 加入到向量中，在训练时冻结模型原本的参数，只训练这些 prompt。P-tuning v2 利用多层提示，如同前缀优化，将不同层中的提示作为前缀 token 加入到输入序列中，并独立于其他层间。由于 P-tuning v2 只训练前缀 prompt，因此其参数量只有原模型参数量的 0.1%-3%，且效果与全量微调相当。

在本系统中，我们利用 P-tuning v2 对 ChatGLM2-6B[6] 进行微调。我们将利用自动化标注获取的数据整理成 SFT 的形式，包含 instruction,

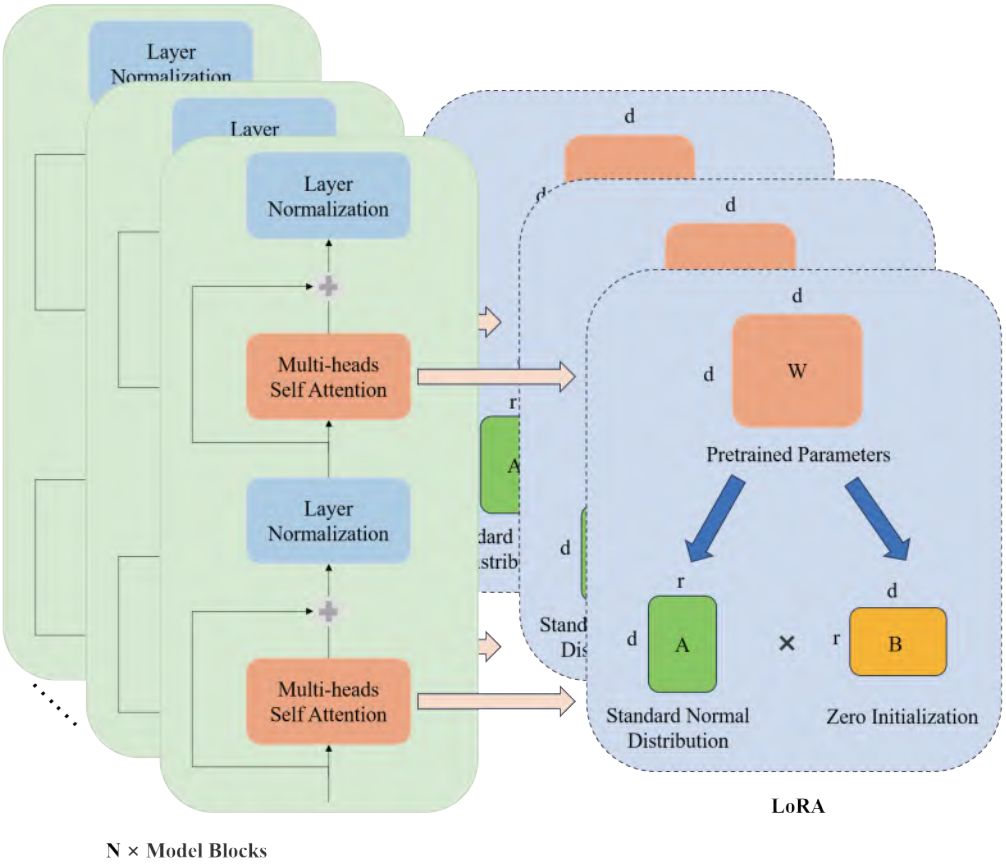


图 5：LoRA 原理图

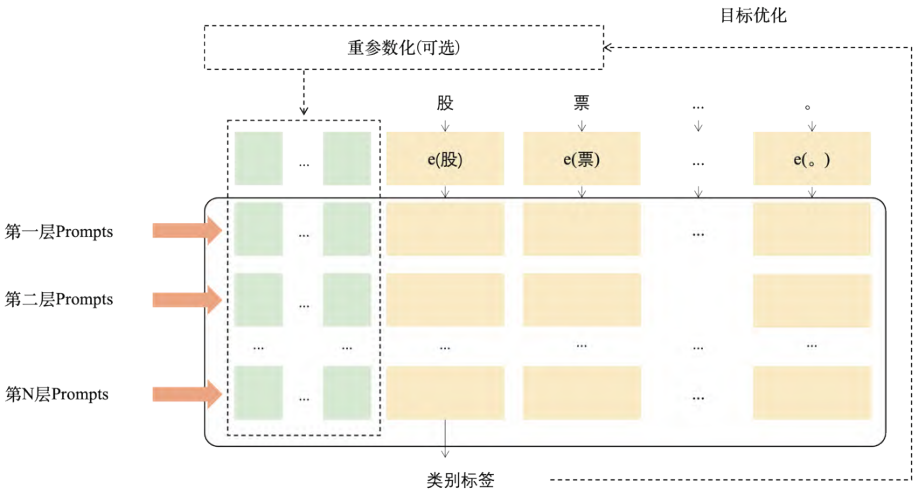


图 6：P-tuning v2

input, output 三个部分。利用 P-tunning v2 将 instruction, 即 prompt 从离散的形式转化成连续向量的形式，用来提示基座模型有关 " 网络黑嘴 " 的定义以及要求识别 " 网络黑嘴 " 的指令。P-tuning v2 既能够使基座模型学习到 " 网络黑嘴 " 相关的概念，又能减少微调参数量。微调后，我们得到基于 ChatGLM2-6B[6] 的 “网络黑嘴” 识别模型。

6 实验结果

我们使用 ChatGPT 标注的 1000 条” 网络黑嘴” 数据用于完成指令微调的 ChatGLM-6B 模型的测试。测试结果如表 1 所示，ChatGPT 识别为“黑嘴” 的共 185 条，本模型识别为 “黑嘴” 的共条 205 条，ChatGPT 识别为 “黑嘴” 的数据与本模型识别为 “黑嘴” 的数据重叠 141 条，模型准确率为 68.78%，召回率为 76.21%；在专家筛选的 15 条” 黑嘴” 数据中，我们对比了两个版本 v1 和 v2 的数据标注的测试结果，他们的区别是所使用的 prompt 不同。模型最好可以召回 13 条，召回率达 86.67%，结果如表 2 所示，类别正确表示模型在具体的三分类黑嘴中也预测正确，召回率只考虑模型是否识别出了黑嘴。图 7 给出了一个具体的识别输出实例。

表 2：专家标注数据测试结果

模型	预测结果
Baichuan for data_v1	召回：11；类别正确：3
Baichuan for data_v2	召回：10；类别正确：0
Chatglm2 for data_v1	召回：13；类别正确：5
Chatglm2 for data_v2	召回：12；类别正确：7

表 1：ChatGLM 模型识别结果

准确率		召回率			
68.78%		76.21%			
预测结果	非黑嘴	黑嘴			
		黑嘴类别一	黑嘴类别二	识别为黑嘴类别三	总和
	795	137	67	1	205

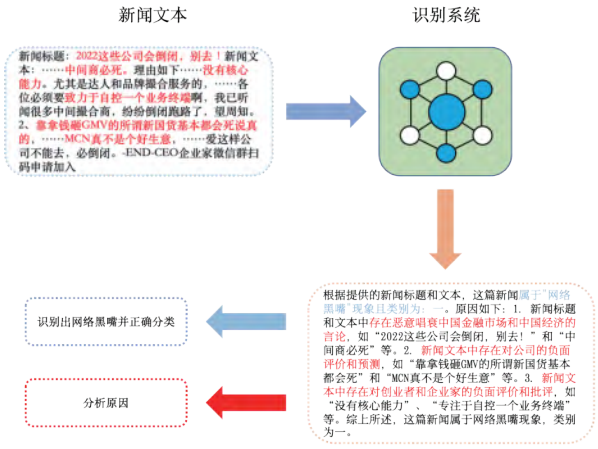


图 7：识别系统输出实例

7 后续优化思路

1）我们将调研类似情境下的 GPT 开发案例，尝试优化利用 GPT 自动标注的 prompt，提升其标注准确率。根据过往的实践经历来看，不同的 prompt 对于 GPT 生成的效果影响较大，因此寻求一个更优秀的 prompt 或许可以成为一个提升标注质量的突破口。

2）邀请专业人士对标注数据进行筛选，提高数据质量。目前看来网络黑嘴识别这一任务的数据标注还是高度依赖于专家知识，目前已有的优质标注过少，模型可学习到的领域知识还非常有限。如果可以通过专家来标注部分数据，那对我们的数据收集工作将有巨大帮助。

3）尝试更强大的基座模型，如 ChatGLM v3 等。当前正处于大模型井喷的时段，几乎每天都会有新的大模型问世，因此适当地跟上潮流，尝试一些新的开源大模型，或许也能够取得不错的效果。

参考文献：

[1] 国家网信办启动清朗·商业网站平台和"自媒体"违规采编发布财经类信息专项整治 [EB/OL]. http://www.cac.gov.cn/2021-08/27/c_1631652531513374.htm

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

[3] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

[4] Liu X, Ji K, Fu Y, et al. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022: 61-68.

[5] Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.

[6] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

海通证券虚拟数智人的应用实践与探索

任荣、蔚赵春、应原、姚振、李鑫芸 / 海通证券股份有限公司
E-mail：lxy14619@haitong.com

元宇宙作为与现实共生的虚拟世界，能够给客户带来沉浸、真实、刺激的消费体验。金融行业作为数字化转型先行者，元宇宙的应用将再次推动数字经济和产业数字化转型的快速发展，其所具备的沉浸式体验和如临其境的交互感能为金融服务带来全新的服务模式。虚拟数字人作为元宇宙的核心组成部分，凭借其虚实结合、沉浸式消费、人机协同等优势，在金融领域亦有非常广阔的应用前景。海通证券积极拥抱这一发展趋势，寻求元宇宙技术嵌入证券业金融应用场景的可能，基于元宇宙理念率先研发推出公司虚拟数字员工，实现科技赋能、降本增效，丰富金融服务模式，拓宽金融应用场景。

1 背景

科技的快速进步不断改变着我们的生活方式，尤其在金融领域，人工智能（AI）、大数据和云计算等前沿科技正在逐步改变着传统的业务模式，推动着金融行业向更高效、更智能的方向发展。虚拟现实（含增强现实、混合现实）是新一代信息技术的重要前沿方向，亦是数字经济的重大前瞻领域。“十四五规划”中明确指出将“虚拟现实和增强现实”列入数字经济重点产业，提出以数字化转型整体驱动生产方式、生活方式和治理方式变革，催生新产业新业态新模式，壮大经济发展新引擎。

证券市场是金融行业不可或缺的一部分，它的繁荣与活跃对于资本市场乃至整个经济都具有重大意义。然而，证券市场中的交易数据量极大、更新速度极快，对于数据处理和决策分析的能力提出了严苛的要求。传统的人工操作和分析方法在处理这种大数据环境下的任务时往往显得力不从心。面对这一挑战，人工智能技术，尤其是深度学习和大数据分析技术，展现出了它们的强大潜力，利用人工智能技术

能够在短时间内处理和分析大量的数据，提取有价值的信息，为投资决策提供科学、精确的支持。

鉴于此，海通证券深刻认识到了人工智能技术在未来证券交易中的重要性和必要性，以相关政策为指引，发挥头部券商的引领作用，大胆闯、率先试，积极布局元宇宙新赛道，基于互联网、区块链、5G、人工智能等技术打造全新的公司虚拟数字人品牌形象，引入重塑传统商业模式的虚拟员工。从智慧内容播报、智能品宣、智能服务、智慧网点等各类业务场景、用户交互到金融消费，虚拟数字人可提供全新的、沉浸式的服务模式，全方位提升用户交互体验。通过虚拟数智人项目的实施，引领证券行业走向智能化的未来，构建起一个高效智能、具有颠覆性的证券金融服务新范式。

2 项目介绍

数智人集成了人物渲染、智能对话引擎、智能语音识别、智能语义理解、多模态智能交互等各项先进 AI 技术，实现交互式数智人与用户侧的智能交互对话。数智人主要通过动作捕捉、二维 / 三维建模、语音合成等技术高度还原真实人类。由人工智能所驱动的数智人，拥有近似真人的形象以及逼真的表情动作，唇形动作能与声音实时同步，且具备表达情感和沟通交流的能力。打造出的高度拟人化虚拟数字形象，能像真人般与人互动沟通，带来全新的感官体验。数智人平台提供标准化对接方式供业务方对接至自己的终端，在终端接收数智人的音视频信息。数智人通过业务终端获取信息，通过语音、视觉等多项识别技术，完成信息的转化；再分发至自然语言处理模块，通过自然语言理解技术，完成任务 / 问答 / 闲聊式对话内容反馈；最终，通过语音及音视频影响合成技术，合成音视频流 / 文件，供业务终端获取呈现。

2.1 技术原理

2.1.1 多模态虚拟数字人技术

多模态数字人集成了人物渲染、智能对话引擎、智能语音识别、智能语义理解、多模态智能交互等各项先进 AI 技术，实现交互式数智人与用户侧的智能交互对话。由人工智能所驱动的数智人，拥有近似真人的形象以及逼真的表情动作，唇形动作能与声音实时同步，且具备表达情感和沟通交流的能力，带来全新的交互体验。数智人平台提供标准化对接方式供业务方对接至自己的终端，在终端接收数智人的音视频信息。数智人通过业务终端获取信息，

图 1：数智员工平台虚拟数智人模块应用架构

通过语音、视觉等多项识别技术，完成信息的转化；再分发至自然语言处理模块，通过自然语言理解技术，完成任务 / 问答 / 闲聊式对话内容反馈;最终，通过语音及音视频影响合成技术，合成音视频流 / 文件，供业务终端获取呈现。如图 1。

2.1.2 大模型技术

大模型（LARGE MODEL）技术是人工智能领域中的一种重要技术，其基本原理是通过分布式训练来构建一个巨大的神经网络模型，该模型可以在大规模数据集上进行训练，从而获得强大的预测和泛化能力。具体来说，大模型技术通过分布式计算框架，将大量低成本的小模型(即前向传递的神经网络)进行并行训练，以期获得更好的预测和泛化能力。这些小模型可以是基于卷积神经网络（CNN）、循环神经网络（RNN）、长短时记忆网络（LSTM）等不同类型的神经网络。在大模型训练过程中，首先需要准备一个包含大量数据的数据集，然后将这些数据集分成多个小块并分配给不同的小模型进行训练。每个小模型独立地对数据集进行前向传递，并将其结果相加以获得输出。然后，这些小模型的输出被合并起来，以获得最终的预测结果。如图 2。

基于人工智能及大模型技术的应用，对于大模型精调、提示工程、知识增强、检索增强、人类反馈的强化学习（RLHF）等大模型相关新技术进行了探索应用，结合数智人形象实现更加智能化的交互方式。AIGC 技术能够帮助企业自动化处理大量数据和复杂问题。利用大模型技术和虚拟人形象结合智能终端构建智能交互系统，能够为用户提供更为智能化和自然的交互体验，实现人机智能交互。这种基于 AIGC 和大模型技术的人机智能交互方式，将为未来的智能系统带来更加智能、高效和自然的交互体验。

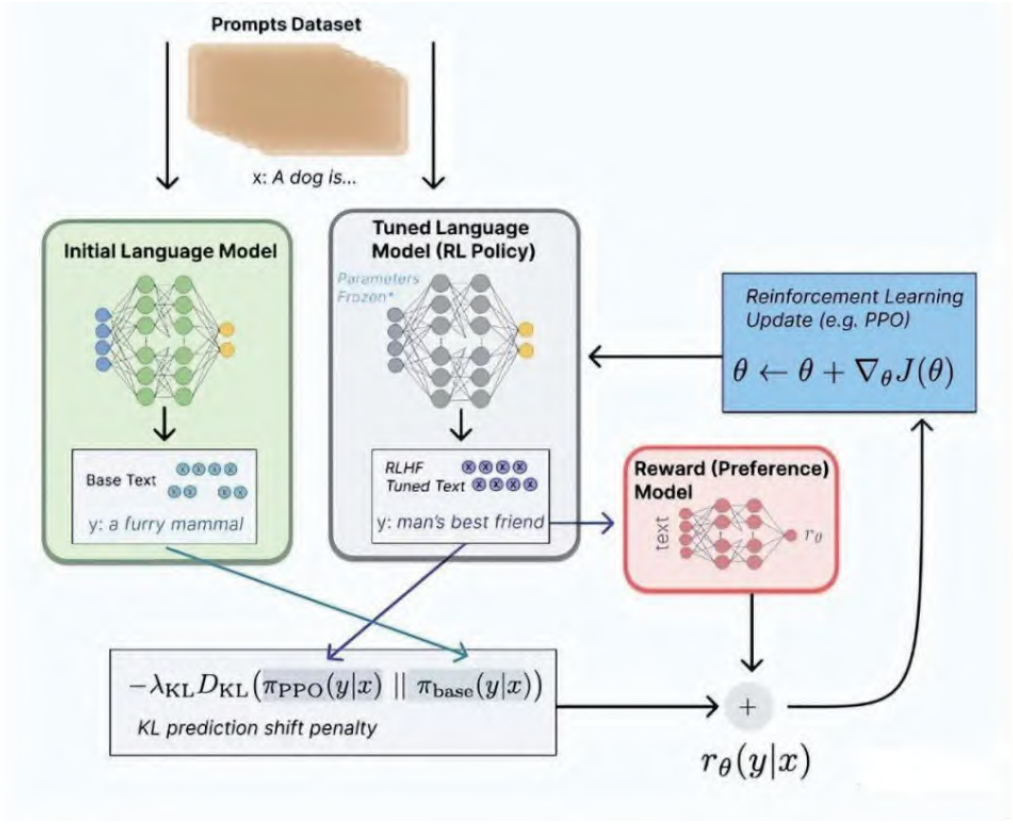


图 2：大模型技术原理图

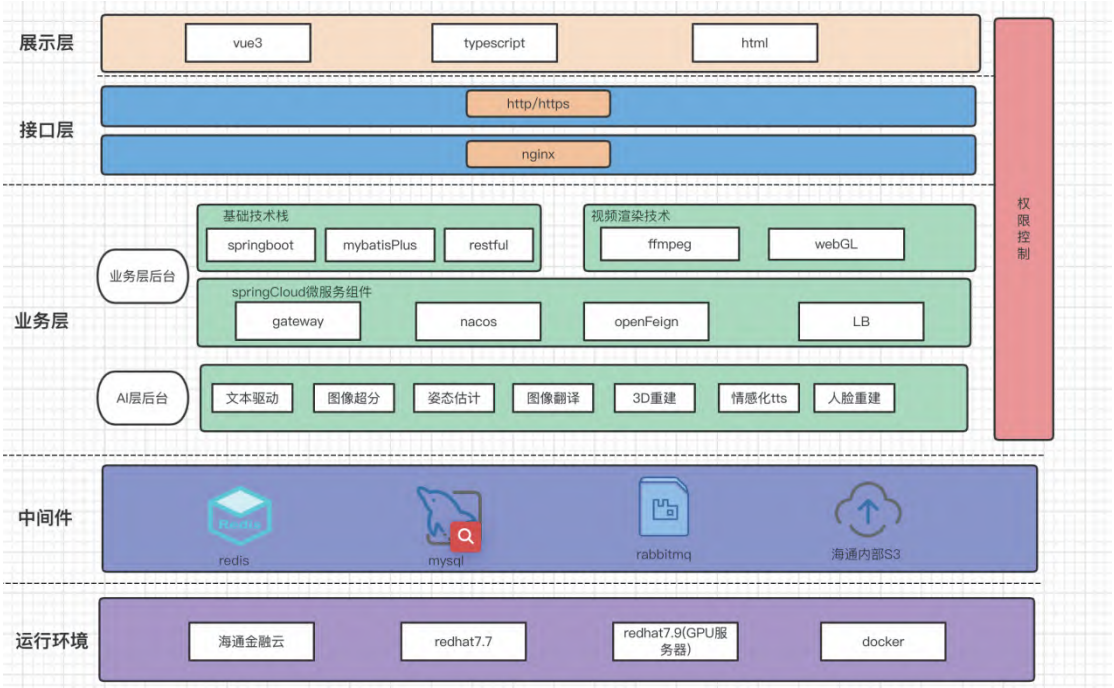


图 3：数智员工平台视频制作模块技术栈

2.2 平台整体架构

数智人基础平台主要包括人物模型渲染、动作驱动、表情驱动、智能交互、图像处理等相关的底层能力，文本驱动、语音驱动、获取视频流、SDK 等各终端对接能力，以及系统管理、内容运营等后台运营能力。如图 3。

目前数智人技术平台自下而上主要分为人工智能层、对接能力层、终端展示层、应用层。其中人工智能层是数智人平台最核心的模块，这一代虚拟数字人有三个最核心的内容——形貌表情系统、骨骼行为系统、灵魂认知系统，这三个很大程度上都要基于人工智能，并且未来随着 AI 能力的发展会逐步实现实时驱动。按照功能模块可以主要分为 AI 核心底层模块、数

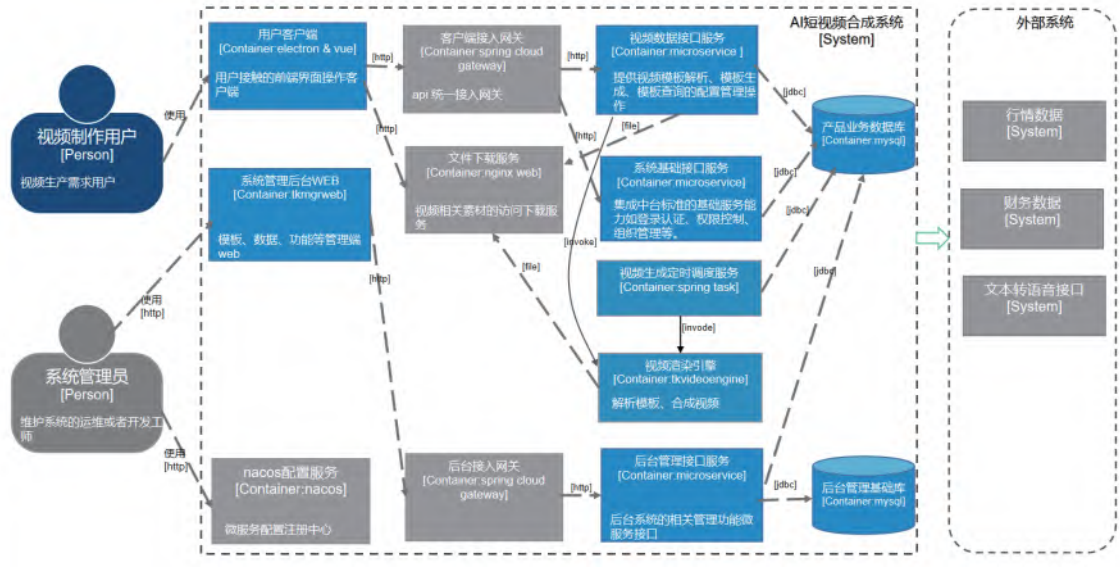


图 4：数智员工平台视频制作模块交互原理

字人形象管理模块、视频制作管理模块、人机交互模块，对外提供内容播报能力和人机交互能力。

数智人视频生产平台可以满足数智人线上投教基地建设、数智人行情资讯解读、数智人营销理财推荐等需求，可以提升客户黏度，将有价值的信息及时准确的触达投资客户，为客户提供更好的服务体验；通过建设数智人培训系统，可以实现快速制作数智人视频，且可以根据需求及时迭代课件视频，节省大量的人力，同时保持培训的灵活性和先进性。如图 4、图 5。

数智人交互能力可以借助智能终端完成实时语音对话，用户可以通过语音咨询天气、股票行情、人文百科等各种信息。同时，数智人交互大屏还支持对话 QA 定制，用户可以根据自己的需求定制化配置问题和答案，让大屏更加贴合实际使用场景，提高用户体验。不仅如此，数智人交互大屏还支持多轮对话，用户可以和大屏进行更加深入和复杂的交互，获取更加精准和满意的答案，可以很好的助力智慧网点的发展。如图 6。

图 5：数智员工平台形象运营中心

图 6：数智人交互大屏

2.3 项目创新性

1）基于大数据和人工智能算法，打造智能化的审核决策中心

数智人融合了公司多种科技能力，包含人工智能平台、流程自动化机器人、大数据平台等多项金融科技技术以及 OCR、ASR、TTS、NLP 等多类基础能力，打造了新一代内容生成、人机交互的新基建，打造更为智能化的审核决策中心。

2）行业首创 3D 可交互数智人，全面助力服务升级

建设完成了证券行业首个 3D 多模可交互，同时支持投教、投研、客服、品宣等业务场景的金融科技创新应用。科技拉近了人与人的距离，有了数智人的陪伴，海通为用户提供了多元化、人性化、智能化的新体验，创造证券业务数字化新模式在投资产品介绍、投资者教育、客户智能服务、VTM 智能引导等对客场景中显著提升客户的体验。

3）灵活的视频编辑平台，智慧运营更高质量

建设完成了运营后台，提供自主业务场景配置，支持视频素材管理、快速生成数智人播报视频，支持基于图片、文本、话术、视频、音频等素材进行配置，支持画中画等功能，为快速、便捷生成高质量新媒体内容提供新模式。

4）AI 大模型与金融数据应用结合，助力数智人服务提质增效

海通大模型目前具备了文本理解、文本生成、多模态理解、多模态生成、场景理解五项基本能力。依托国内智源悟道 2.0 人工智能服务体系、以及自有算力、算法、数据、人才四位一体的基础能力，重点着眼于大模型在金融领域的知识理解能力、内容生成能力以及安全问答能力，对于大模型精调、提示工程、知识增强、检索增强、人类反馈的强化学习（RLHF）等大模型相关新技术进行了探索应用。该版本大模型拥有百亿级参数，初步具备自由闲聊、合规知识问答、内容摘要等多类型任务的服务能力，通过研发支持服务面向 RPA 智能工单支持服务以及虚拟人问答生成、为后续大模型的深入应用迈出了坚实的一步。沉淀 AI 大模型模型与金融数据资产，探索 AI 大模型的多模态服务、可信应用、知识增强、高性能训练推理等方面的技术创新，推进智能人和 AI 大模型结合在营销、风控、客服等领域的探索和试点应用。

5）“数字员工”升级为“数智员工”

随着 AI 技术的成熟发展，数智人形象逐步会发展为企业数字员工的最理想的载体和呈现方式，通过传统数字员工技术与虚拟数智人的结合，将“数字员工”逐渐升级为“数智员工”，让更多的“数字员工”从后台走向前台，逐渐走向用户的视野，进一步提升用户体验，让员工在工作中感受到科技带来的幸福感。

3 应用实践

证券行业存在众多数智人应用场景，在客户服务与支持、投资顾问、市场分析与预测、风险管理、教育和培训等领域，利用虚拟数智人帮客户解决问题，从而改善客户体验，助力公司打造智能化、形象化、可交互的数智人形象，同时可以起到对外展现公司科技能力，进一步提升行业影响力。

3.1 应用场景

3.1.1 智慧内容生产

通过智慧内容生产可以提升券商服务质量，更好的满足客户投资选股等需要，提升客户黏度；利用科技手段赋能资讯视频服务过程，将有价值的信息及时准确的触达投资客户，为客户提供更好的资讯服务体验；满足证券行业客户适老化支持，助力公司品牌建设及业务数字化拓展。

图 7：智慧内容生产

3.1.2 智慧服务

通过数智人智慧服务可以完成数智人开户引导、数智人客户服务、虚拟坐席等需求，使用交互型数字人服务，对接语音识别技术、文本知识库技术、以及音视频平台，完成交互式数字人客服与相关 AI 技术的能力串联，为用户侧提供数字人音视频交互沟通能力，满足客户数字人开户引导、数字人客户服务、虚拟坐席等需求。

通过数字人与用户实时智能互动的方式，提升业务办理成功率和用户体验；同时通过数字人对外统一服务形象，加强品牌效应，提升服务温度与科技创新力。

通过在用户端融入虚拟数字人形象，实现数字人实时智能交互与可视化服务，推动证券公司金融服务升级、服务模式创新，使金融服务更加智能化、有温度。

3.1.3 培训考核

图 8：智慧服务

图 9：培训考核

证券公司总部需要对各分支机构进行产品和服务培训，真人培训老师通过线下或者线上方式组织培训，占用大量人力，且时间不够灵活，当产品和服务更新速度较快时，组织真人老师进行培训的更新速度无法满足需求；再者，一线客户经理流动性较大，新人入职需要进行常规产品和服务培训，真人培训和答疑能解决的问题有限。

通过建设数字人培训系统，生成数字人培训课件，可以实现快速制作数字人培训视频，且可以根据需求及时迭代课件视频，保持培训的灵活性和先进性；通过使用数字人培训和考核讲师，可以实现专业知识培训考核，提升客户经理对客能力。

3.1.4 智慧主持

自数智人推出以来，“小海”还作为虚拟主持人参与多次各种线上或线下的活动，通过互动问答的方式为嘉宾带来了新颖、优质的智能互动体验。虚拟数智主持人突破了地域和时间的限制，可以随时为用户提供在线服务，同时具备人类无法比拟的信息处理能力，可以确保提供更准确、更专业的信息。

3.1.5 数字员工商城

随着数智员工在公司的广泛应用，出现一些高重合度的需求场景，如何充分利用已有的数智员工而非重复开发去满足这些需求也是公司急需解决的问题。我们将可共享的数智员工基础能力发布到数智员工在线商城，同时结合 AIGC、GPT 等创新技术提供在线功能体验，通过云端向各终端的商城推送数智员工；用户也无需单独安装众多应用程序，通过商城即可实现数智员工的使用及运行过程中的状态监控。目前已积极探索研制了各种类型的数智员工：图生图、文生图、错别字审核、文本比对数智员工等，提升用户体验。

3.2 项目成果

本项目综合研究数智人形象建模、人物渲染、人机交互以及视频渲染等相关技术，提出了一套完整的数智人平台建设和应用实践解决方案，搭建高可用的数智人基础平台，进行统一管理，以支持快速业务流程的开发和投产。

通过数智人视频制作平台，将可以充实企业自身运营视频制作的管理，利用丰富的模板可以

进一步提升视频制作效率，从而减少了在视频制作的投入精力。“小海”上岗后，化身智慧播报员、智能主持、智能培训官、智能客服等多重角色，在投资者教育、资讯播报、大会主持、开户引导、培训等 10 余个业务领域充分发挥潜能，赋能数金、运营、财富、投行、财务、人力等多个业务条线。“小海”也在各类行业峰会活动中亮相主持，为嘉宾带来了新颖、优质的智能互动体验。“小海”的加入使得业务更加数字化、智慧化、专业化，持续赋能业务数智化，加速推动公司数字化转型，助力公司高质量发展。同时，通过数智人形象打造公司自己的品牌宣传大使，既展示了公司的科技实力，又传达了公司的服务理念。数智人的出现，不仅可以吸引客户的注意力，提升公司的知名度，还可以增强客户对公司的认同感，帮助公司建立独特的科技品牌形象。

4 探索与展望

在数智人未来的发展中，我们看到了巨大的潜力和无限的可能。未来，海通将持续深化和扩展人工智能在金融领域的应用，将更加智能化的投资顾问、更精确的风险预测、更人性化的客户服务以及更广泛的应用领域无缝结合，以形成一个全面的、创新的金融服务体系。

1）深入挖掘虚拟数智人应用场景

证券公司线下投教基地建设是根据政策要求和证券公司对客投教的关键路径之一，为提升展厅的投教宣传作用，数字人虚拟展厅应运而生。作为当前元宇宙建设的具体形式，数字人虚拟展厅运用 AI 驱动的 3D 虚拟数字人技术，实现用户、虚拟人、虚拟场景、虚拟物件的多元化互动。同时还可利用数智人进行投教直播，通过真人声音或摄像头驱动，数字人投教直播能够呈现更高表现力，同时节省人力成本。此外，还可利用数智人建设智慧网点，结合公司客服知识库和大语言模型的支持，提供智能化的交互对话服务，在营业网点为客户提供友好的用户体验。相信随着人工智能技术的不断发展，数智人将会在更多的场景得到应用，为金融服务带来更多便利和智慧。

2）对接金融大语言模型，助力数智人服务提质增效

大模型的引入将赋予虚拟数智人更强大的能力，包括决策支持、自然语言处理、个性化服务、情感交互和知识图谱构建等。大模型处理海量数据可提供高层次决策支持，帮助投资者做出智能投资决策。自然语言处理方面的优势提升了交互体验，个性化服务预测用户投资偏好并提供建议。通过大模型增强情感识别和交互能力，虚拟数智人更好地理解回应用户情绪，提供人性化服务，构建金融知识图谱增强虚拟数智人的知识储备与理解能力。虚拟数智人与大模型的结合将极大地推动金融行业的数字化和智能化转型，带来更高效、更智能、更人性化的服务，为我们开启金融科技的新篇章。

相信未来的虚拟数智人将成为金融科技领域的一颗璀璨明星，其强大的智能化能力将极大推动金融行业的数字化和智能化转型，引领金融科技的未来发展。

基于AI技术的全场景数智化服务平台的实践应用

潘建东、马张晖、梁彬、尹序鑫、孙冰、王赵鹏、刘国杨 /
中信建投证券股份有限公司 信息技术部 北京 100010
E-mail：mazhanghui@csc.com.cn

在居民财富管理需求日益强烈以及金融业数字化转型的大背景下，人工智能技术赋能财富管理与提升服务品质成为各大券商经纪业务的工作重点。本文主要阐述了中信建投证券基于人工智能技术并结合实际业务需求，对全场景数智化服务平台的探索与建设，并总结了客户服务平台数智化建设的实践经验。

关键词：金融服务；数字转型；客户服务平台；人工智能

1 概述

进入 21 世纪以来，国内居民可支配收入增长显著，对于金融机构提供高质量财富管理服务的需求日益强烈。与此同时，移动互联网快速普及，社交媒体主导交互新时代，客户的咨询需求更加多样 [1]，证券公司等金融机构在进行财富管理业务时，普遍面临着庞大的客户群体与综合服务能力不匹配的问题，如何高效地提供对客服务是金融机构必须应对的难题。

在数字化转型的大背景下，以 5G 通信、人工智能（AI）、区块链等技术为代表的新一代数字化技术的普及应用，为解决传统金融服务难题提供了新应对途径 [2]。为提升客户体验及服务效能，中信建投证券于 2019 年开始积极探索前沿 AI 技术，推动感知智能、数据智能融合的全场景客户服务平台落地，在合规风控、智能营销、客户服务等领域提供了“AI+”数字化转型新思路。

2 全场景数智化服务平台的探索与建设

传统金融服务由于存在着人工依赖程度高、渠道规划性弱以及业务效率低下等缺点，在面临愈发复杂的市场环境时显得愈发力不从心[3]。为此，中信建投证券积极开展了数字化转型的战略部署，构建全场景数智化客户服务平台。该平台通过基于人机协作的系统构建，覆盖客服全流程的业务模式，能够最大程度实现人工和 AI 的优势互补。

2.1 设计目标

（1）打造证券领域专业领先的智能知识库

全场景数智化服务平台项目，旨在建设统一、标准、高效的知识管理平台，实现知识的集约化、精细化及智能化管理，提高知识管理和共享能力，进而打造证券领域专业领先的知识库。首先，通过将分散在企业各部门和机构的知识点汇聚并有机整合，提高企业内部的知识运用效率和管理效能；然后以人工智能等技术手段对知识进行分类、标记、模型化，形成完整的知识管理和业务分析框架，实现对知识资源的全面掌控和分析；最后通过建设知识库，达成数据共享和知识交流，提升知识共享的高效性和可靠性，为业务发展提供更加均衡和有效的支撑，推动行业创新发展。

（2）以信息科技高效赋能证券从业人员

证券从业人员需要在行业趋势预测、投资策略探讨及专业知识巩固等方面花费大量时间和精力，其决策往往需要顾及多方面因素，且经验差异会导致人员专业能力差异较大。全场景数智化服务平台项目，旨在将行业知识储备、服务经验和科技能力等整合起来，为员工提供工作效率提升、决策质量优化及智能服务供给等各项服务。本平台基于知识图谱、人工智能及音视频技术，建立高效的知识管理和服务共享机制，同时向投顾服务过程提供坐席全流程的智能辅助，从而赋能员工、助力财富管理业务高效开展。

（3）实现证券财富管理服务降本增效

随着财富管理的服务规模和业务范围逐年扩大，券商面临日益增长的运营成本压力和客户需求，全场景数智化平台项目的目标之一便是实现财富管理服务的降本增效，实现经营效益和客户满意度的双重提升。其目标包含引入多模态技术实现客户管理精细化、服务流程智能化及合规管控实时化，旨在利用海量用户行为数据实现客户行为识别和画像生成，助力营销服务升级；同时，计划提供智能外呼等融合文图音频等多模态赋能的营销方式，并且具备基于自然语言处理的实时质检能力保障营销合规，从而大幅降低人力资源成本，提高财富管理服务流程效率，实现企业的长远发展和客户的满意度提升。

2.2 逻辑架构

基于 AI 技术构建的全场景数智化服务平台，已经实现了与各大核心业务系统的互联互通，涵盖了公司服务“服务前”、“服务中”与“服务后”的全生命周期，主要功能应用根据客户服务的不同阶段，分为事前智“慧”（服务前），事中智“助”（服务中）和事后智“学”（服务后）。

（1）事前智“慧”（服务前）

事前智“慧”，即在客户服务前期智慧构建用户画像，精准洞察客户财富管理需求。

首先，构建知识中台。即基于人工智能技术的知识中台从跨业务条线的多源异构数据中构建问答对、知识图谱、全文检索等结构化知识，并将结构化数据与客户标签体系融合，形成以客户为中心的知识表示。知识中台的持续生产能力保障各服务系统持续跟踪并精准洞察客户需求，同时对可配置投资标的大量信息进行加工生产，将投资标的信息与客户需求信息进行匹配，助力各客服模块提供针对性的服务方案。

其次，客户画像分析。根据客户的过往通话记录、业务办理情况、风险匹配情况进行分析，

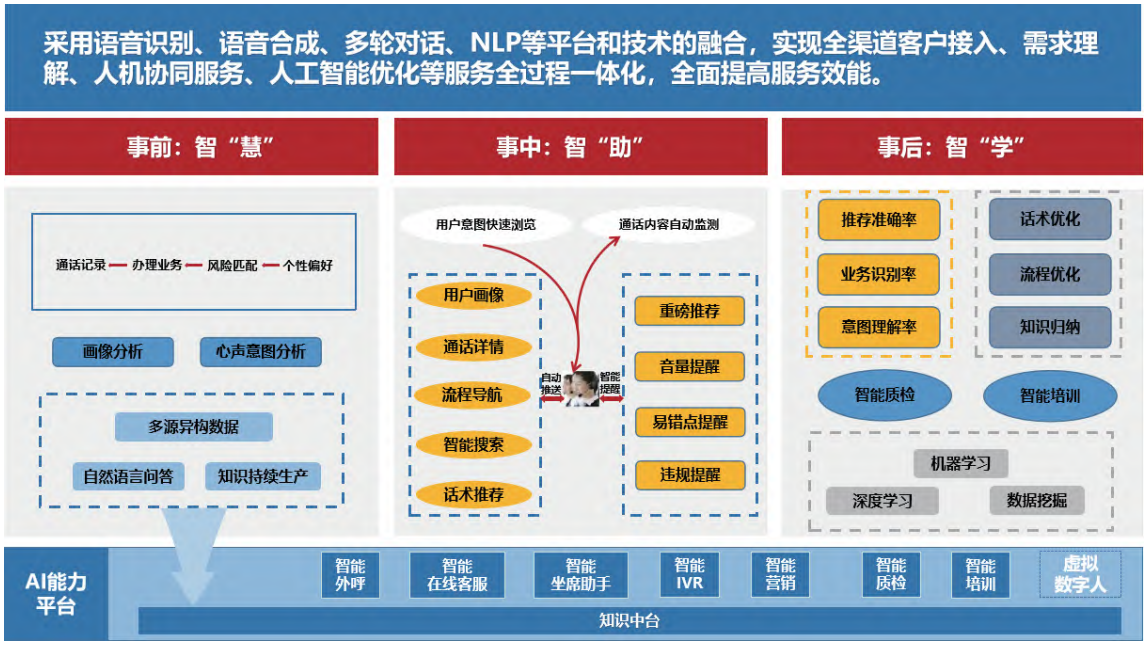


图 1：全场景数智化服务平台的逻辑架构

将客户自身使用习惯与特点进行个性化分类，为客户选择自身偏好的服务渠道、个性化的服务模式提供数据抓手。最终实现客户千人千面的财富管理和服务需求整合，针对不同客户提供一站式个性化定制服务。

（2）事中智“助”（服务中）

事中智“助”，即在客户服务期间通过 AI 技术辅助工作，提升客户服务质量，显著实现提效降本。

智能机器人革新传统电话按键导航，结合大模型搭建 AI 客服，通过 TTS、NLP 等多项技术提供一站式语音问答交互。智能外呼系统结合用户大数据画像，针对不同财富管理偏好的客户进行自动外呼，达成客户经理邀约、投顾产品签约回访、新股中签通知等财富管理和综合服务。智能助手系统通过实时监测通话中客户提及敏感词、服务禁忌语、语速等，侦测互动双方情绪，实时统计、检索与分析通话内容，捕捉投诉问题，将业务知识与沟通经验通过实时弹屏辅助坐席解答。同时实时评估业务类型与话术流程的匹配度，提醒一线员工及时修正，辅助人工坐席优化综合服务质量。

（3）事后智“学”（服务后）

事后智“学”，即在客户服务后期对全服务流程进行回归分析，学习最优的营销和服务策略，不断优化客户财富管理和综合服务。

智能质检系统利用 ASR 技术，对全量、全员通话录音自动化质检。精准锁定问题录音，并推送给人工审核即时分析，极大增强了客服系统应对突发情况、异常指标的快速响应能力。系统针对全流程服务记录，从业务属性、客户属性、员工属性等多个维度进行分类、排序，通过对比发现差异，根据差异改造完善各个服务节点及服务流程。以提升服务完成率为导向，利用人工智能和大数据技术分析构建各业务服务路由模型，构建 AB 测试服务流程探索机制，为客户提供针对性的个性化财富管理方案。

2.3 技术架构

中信建投证券通过对全场景数智化服务平台的建设，首先实现了基于人工智能技术的知识中台模块，支持多源异构数据的加工处理、自动持续的知识生产、知识经验便捷沉淀及共享等功能，为构建高级分析工具和构建组织统一的认知及决

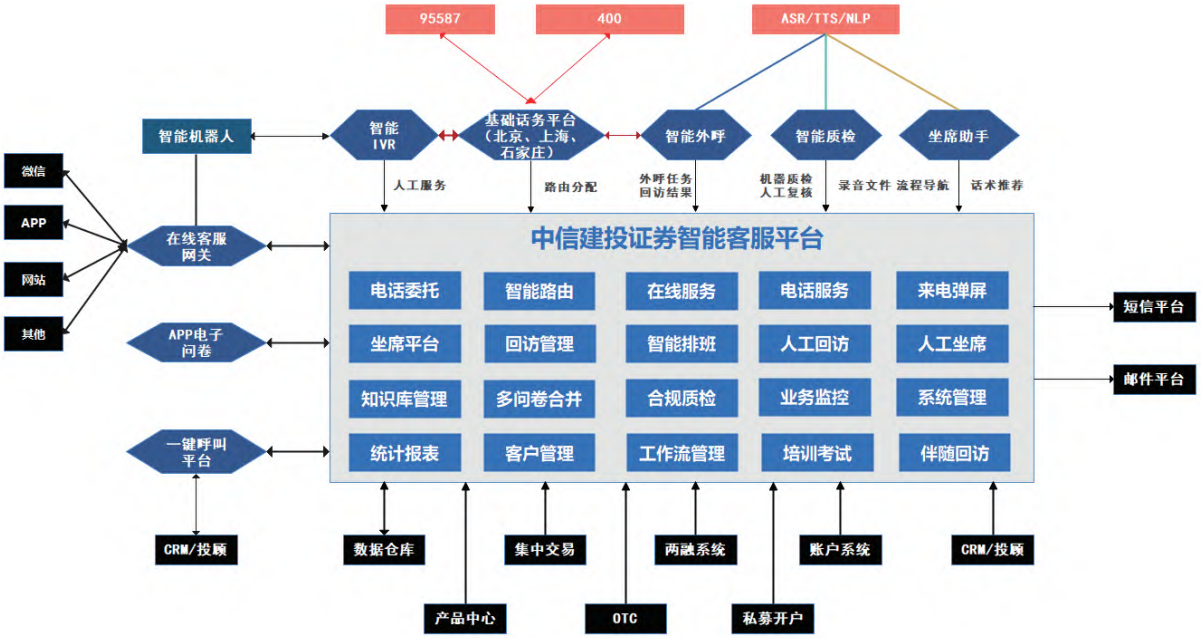


图 2：全场景数智化服务平台架构

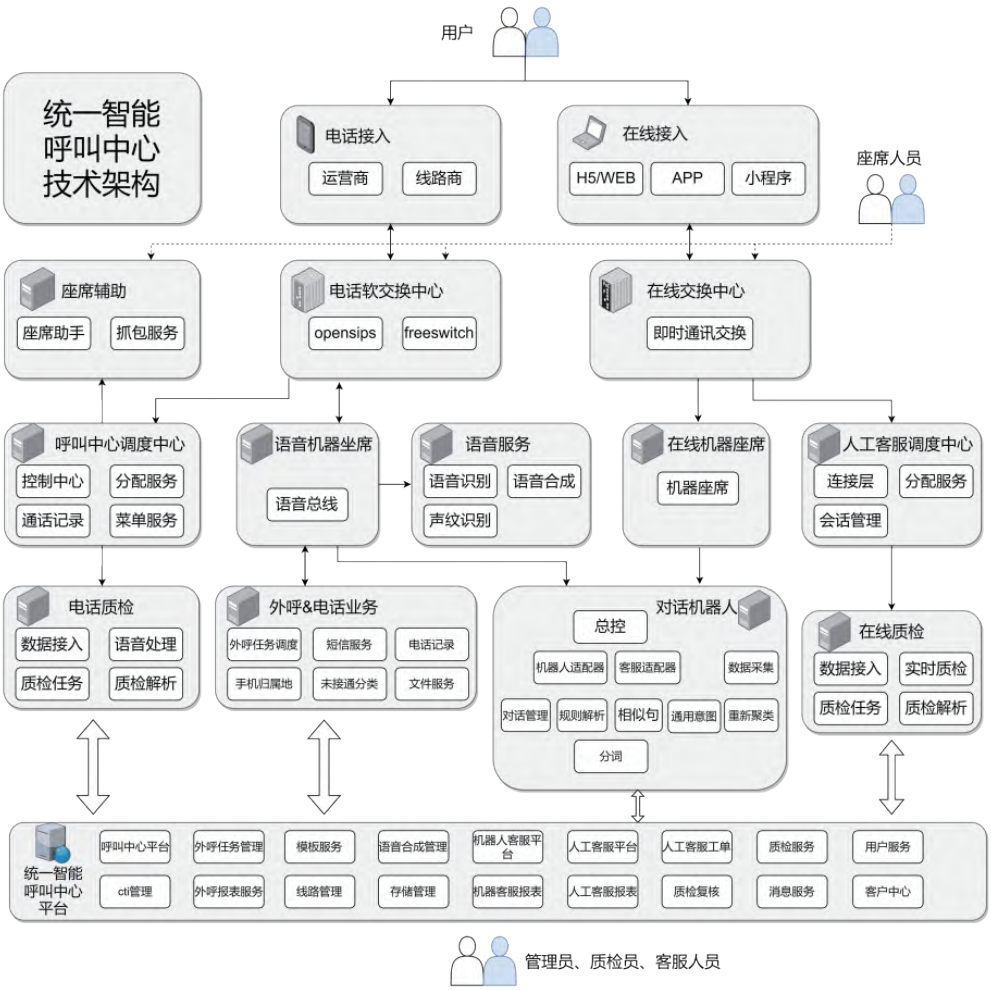


图 3：智能化呼叫平台系统技术架构

策视野提供知识基础。然后，建设基于人工智能技术的坐席辅助模块，利用语音识别、自然语言处理、检索等技术，实现对话交互过程中的实时话术推荐、实时合规质检、实时业务分析、客户画像分析等功能，协助坐席为客户提供合规高质的电话服务。最后打造开放的专业领域的 AI 赋能基座，基于组织数据建设语音识别、语音合成、自然语言处理、图像识别等 AI 组件，探索语言大模型在金融垂直领域的应用，为各项智能化应用提供基础 AI 能力。

以电话呼叫业务、员工坐席两大业务场景为例，针对全场景数智化平台中的智能化呼叫平台、员工赋能平台两大系统的技术架构进行阐述。

（1）智能化呼叫平台系统

中信建投证券智能客服团队及运营管理部分别构建了智能外呼、智能电话客服系统，并于 2021 年将智能外呼、智能电话客服统一为智能化呼叫平台系统，完成“从单个产品的解决方案，到整体智能化呼叫中心解决方案”的调整。

图 3 为智能化呼叫平台系统技术架构，系统包含多种智能功能模块。例如，坐席辅助模块可实时监听人工坐席和用户的语音通话并转写，通过 nlp 触发相关提醒，获取到机器人的答案给坐席参考；语音机器人与对话机器人模块，提供机器人坐席服务，支持和客户进行实时语音与文字交互，并生成人机对话交互记录；在线质检模块则从人工客服获取坐席和用户的文本对话记录，进行质检处理等等。

（2）员工赋能平台系统

为配合财富管理转型战略规划顺利实施，财富管理部提出打造员工赋能平台，为员工搭建分享和获取经验的知识平台，并在平台之上不断叠加智能辅助、协同工具，帮助员工提高综合服务能力。

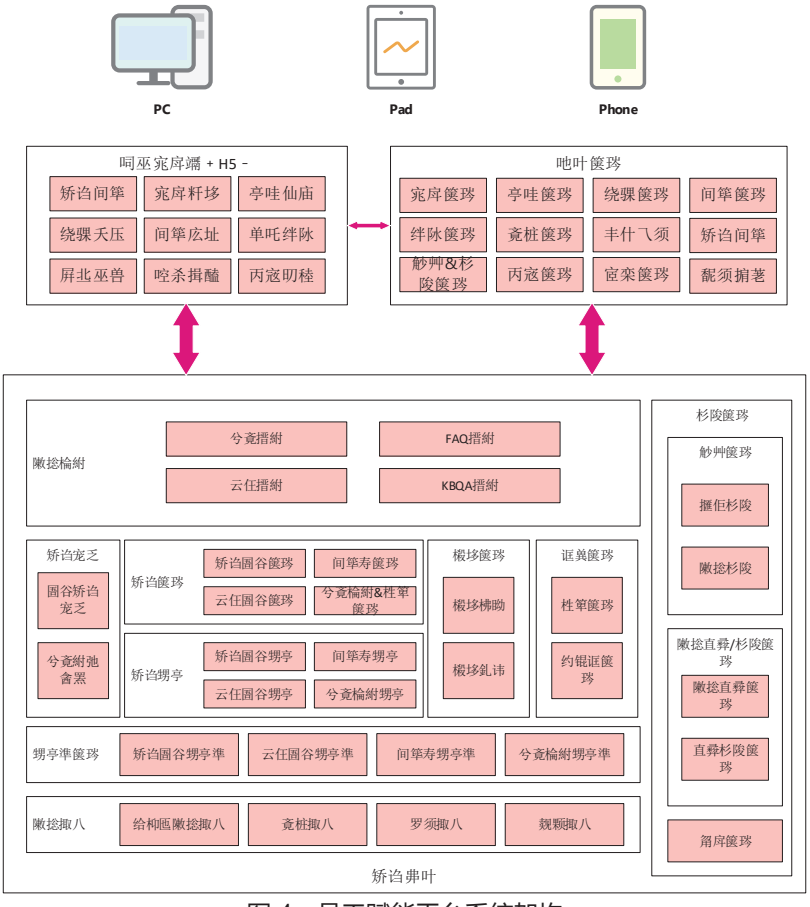


图 4：员工赋能平台系统架构

员工赋能平台的底层是基于 NLP、ASR、知识图谱等智能 AI 技术构建的知识中台。通过该平台，可以实现多源异构知识的采集、汇聚；多维知识的持续自动生产和分发。基于平台的功能，进一步为员工提供知识分享创作、快速搜索、专家查找、业务协同组队等功能。

3 全场景数智化服务平台建设成效

中信建投证券秉承科技赋能的核心理念，用数字化思维赋能金融综合服务，通过建设全场景数智化客户综合服务平台，为公司客服业务全流程提供有力支持。

3.1 经济效益

（1）数智化赋能业务

目前全场景数智化客户服务平台已经赋能多个业务部门，在前期不断探索和积累的基础上，人工客户服务和智能客户服务共同使用形成合力，更好地为客户服务。智能客户服务不受工作日和工作时间限制，可根据客户的碎片化时间需求，解决以往重复性较高的问题，人工客户服务则根据客户实际情况提供个性化优质服务。将二者有效结合，能够真正为企业节约成本，提高工作效率。其中，智能电话客服机器人系统语音转写准确率 90%+，问题回答准确率 85%，整通对话完成率达到 70%，7*24 小时的全天候服务在疫情封控期间发挥巨大的作用。智能外呼系统平均每年可完成约 110 万通电话，平均每月 9.1 万通，按人工每人每工作日处理 160 件回访任务计算，平均每工作日可节省 26 个人工，提效降本效果显著。

（2）有效提高合规审核效率

通过人与智能的协同配合，让智能更好地赋能合规人员。通过智能合规质检系统，实现了合规问题找得准、找得全、找得快，有效提高了合规审核人员的审核效率，有效降低了漏检、误检所带来的额外工作负担，目前质检系统每周可全量质检 6 万通会话数据，替代 80% 的质检工作，获得了合规审核人员的认可。

（3）数据分析赋能

通过对全流程客服数据的持续分析，针对客户、渠道、业务特点，建设智能培训、智能潜客筛选、智能服务路由等模块，不断优化服务节点和服务流程，构建服务闭环，提升服务触达效率，提升服务质量。2023 年全场景数智化客户综合服务平台服务触达成功率同比提升 15%，服务满意度同比提升 10%。

3.2 敏捷价值

（1）快速响应客户需求：全场景数智化客户服务平台采用智能语音识别与自然语言处理技术、聊天机器人等工具，可以快速理解客户问题和需求，并提供快速响应和解决方案，大大缩短了客服响应时间。这有助于提高客户满意度和忠诚度。

（2）灵活定制服务：全场景数智化客户服务平台可以根据不同企业的需求和情况进行定制开发和部署，以满足不同的客户服务需求。例如，一些企业可能需要实时监测并管理社交媒体上的客户反馈，而另一些企业则需要对移动应用程序进行深度集成，此时全场景数智化客户服务平台的灵活性将发挥重要作用。

（3）数据驱动决策：全场景数智化客户服务平台通过数据分析和挖掘技术，可以收集和分析大量客户数据，以便更好地了解客户需求和行为，并提供个性化服务。数据驱动的决策也可以帮助企业更好地预测和规划客户服务需求，从而提高客户满意度和企业效益。

（4）持续优化服务：全场景数智化客户服务平台可以实时监测客户服务质量和效率，并通过数据分析和挖掘技术进行持续优化。例如，通过分析客户服务热点问题和处理时间等指标，可以及时调整机器人回答问题的算法和策略，提高解决问题的准确性和速度。

3.3 创新优势

（1）智能化助力业务

项目通过自动分类和归纳用户的需求，可以减轻人工坐席的负担，提高服务效率。通过构建知识图谱，企业可以更好地管理和利用自身的知识和信息资源，利用自然语言处理技术进行语义搜索和推荐，提高用户体验和满意度。此外，项目可以通过数据挖掘和分析技术，对用户需求和行为进行分析，从而为企业提供更加精细化的服务。在客户服务过程中，客户会向坐席提出相关的业务问题，系统可智能判断客户意图，自动推送出客户问题对应的业务知识，以帮助坐席对客户进行服务，无需跨系统手动查找业务知识。

（2）提升客户画像描绘能力

项目利用自然语言处理等技术根据客户的对话内容，自动提取客户意图画像，记录客户全方位的关键信息，便于下次了解客户详细情况，并具备 API 接口，可将提取的客户动态画像反馈至客户管理系统。

（3）扎根证券垂直领域，实现实时违规检测

项目利用自然语言处理技术实时监测坐席说话内容，对违规内容进行实时提醒，以醒目颜色在辅助前端 SDK 上进行提示标记，以颜色区分违规等级。

（4）领域知识助力专业化服务

知识库内容包括服务话术、证券业务知识、业务办理流程等，知识形态支持文本、图片、网址链接（自动解析网页内容），辅助前端 SDK 具备方便的搜索入口，坐席可在辅助前端 SDK 检索相关知识内容，支持关键词检索，保证坐席能快速地检索对应知识；支持模糊搜索，可自动对搜索内容进行智能分词，最大化地将检索词和分词内容的知识检索出来；支持拼音搜索，误输入拼音的情况下，无需重新输入，支持对拼音进行检索，将相关的知识检索出来。

4 总结

中信建投证券响应“十四五”数字规划，积极开展数字化转型的战略部署，针对一线业务痛点，构建全场景数智化客户服务平台。该平台整合了行业知识、服务经验和科技能力，建立高效的知识管理和服务共享机制并覆盖客服全流程的业务模式，能够最大程度实现人工和 AI 的优势互补。全场景数智化平台的建设，为公司带来了客户管理精细化、服务流程智能化，财富管理成本的降低等一系列提升。当前，生成式人工智能正在引发新一轮智能化浪潮，中信建投证券时刻把握 AI 技术前沿，探索其在智能投研、智能投顾、智能风控等业务领域中的应用，不断推动 AI 大模型在金融服务领域的落地。

参考文献：

[1] 谢海斌 . B 公司全渠道智能云客服平台业务战略研究 [D]. 华南理工大学，2020.

[2] 余丽霞，李政翰 . 金融科技对商业银行盈利能力和经营风险的影响研究——基于文本挖掘的实证检验 [J]. 金融监管研究，2023(4)：62-79.

[3] 肖钢，潘建东，徐政钧等 . 金融咨询服务平台智能化建设与应用 [J]. 金融科技时代，2022，30(04)：16-19.

[4] 孟凡玥 . 基于微服务的客户服务平台的设计与实现 [D]. 北京交通大学，2020.

[5] Sage launches Europe's first peer-to-peer customer service platform for SMEs; As the UK starts to un-lock, businesses across the country join forces to share insights and best practice[J]. M2 Presswire,2021.

基础运算技术研究

郑商所五档组播行情FPGA解码加速器设计与实现

张旭东、万锟、刘垚、陈士阳、马龙 / 中信建投证券股份有限公司 北京 100010
E-mail：zhangxudongbj@csc.com.cn

郑商所五档组播行情 FPGA 解码加速器是通过 FPGA 来实现对郑商所五档组播行情的解码加速功能。FPGA 收到组播行情后进行解码，并在内部实时维护一个合约索引表和订单簿，经过数据规整化后，通过 PCIe DMA 将最新五档行情推送到策略服务器。由于采用定制化模板，策略服务器接收到的行情数据可以直接使用，从而实现了最优的行情解码加速。

关键词：郑商所；FPGA；行情解码；加速

1 背景

近些年中国金融衍生品市场进入了快速发展阶段，郑商所期货期权交易业务以其丰富的风险管理功能受到越来越多投资者的青睐，也对商品市场的平稳运行起到了至关重要的作用。中信建投证券作为郑商所的主做市商之一，始终致力于提升自身期货期权做市业务水平。

做市业务属于一种典型的高频交易业务，系统的运行速度直接决定了业务的盈利能力。行情解码作为做市系统的关键路径之一，对速度有着极致的要求，软件方案通常使用低延迟网卡接收网络行情，再通过 CPU 在 Cache 中做行情解码。这种方案在行情解码方面虽然也有不错的表现，但是当合约数量增加、突发数据变多时性能会出现瓶颈，导致行情获取的延迟和抖动增加。本文提出一种基于 FPGA 技术的郑商所五档行情解码加速方案，与软件方案相比，该方案能够做到更

低的落地延迟和零抖动，有效提升郑商所期货期权做市业务的竞争力。

2 整体结构设计

2.1 报文结构

郑商所行情数据传输协议为二进制协议，采用网络字节序的大端存储方式，以组播的形式发送，每一个 UDP 数据包包含若干个报文，报文结构如图 1 所示。每个报文头部固定占用 4 个字节，对报文类型、消息个数和正文长度进行声明，正文部分存放多个相邻的消息，其消息总个数和总长度均在报文头部结构中标识。每个消息头部固定占用 2 个字节，对消息长度进行声明，后续部分为消息正文。

2.2 整体结构

郑商所行情解码加速器通过信息提取与对齐模块（ExtractWithAlignment）接收 UDP 协议栈解析后的报文数据，根据报文长度、消息个数、消息长度等信息进行数据提取和对齐，解码模块（Decoder）根据报文类型对数据进行解码和分类存储，在内部实时维护一个合约索引表（IndexTable）和订单簿（OrderBook），数据重组模块（DataRestruct）根据业务需求提取订单簿内容并转化为定制化的精简数据结构，最终通过 PCIe DMA 上传。ConfigSpace 模块用于寄存器配置和状态信息反馈。整体结构如图 2 所示。

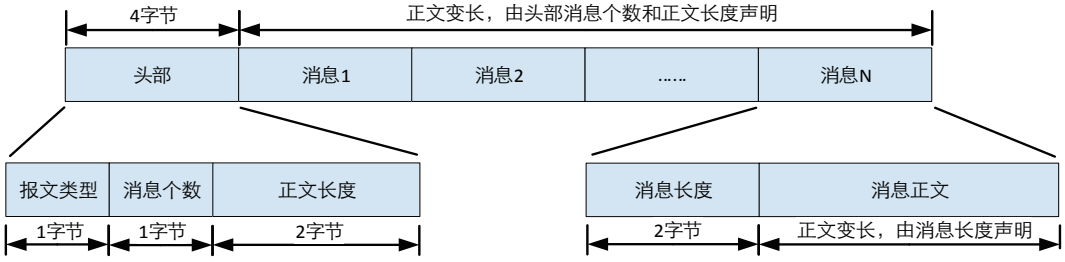


图 1：报文结构

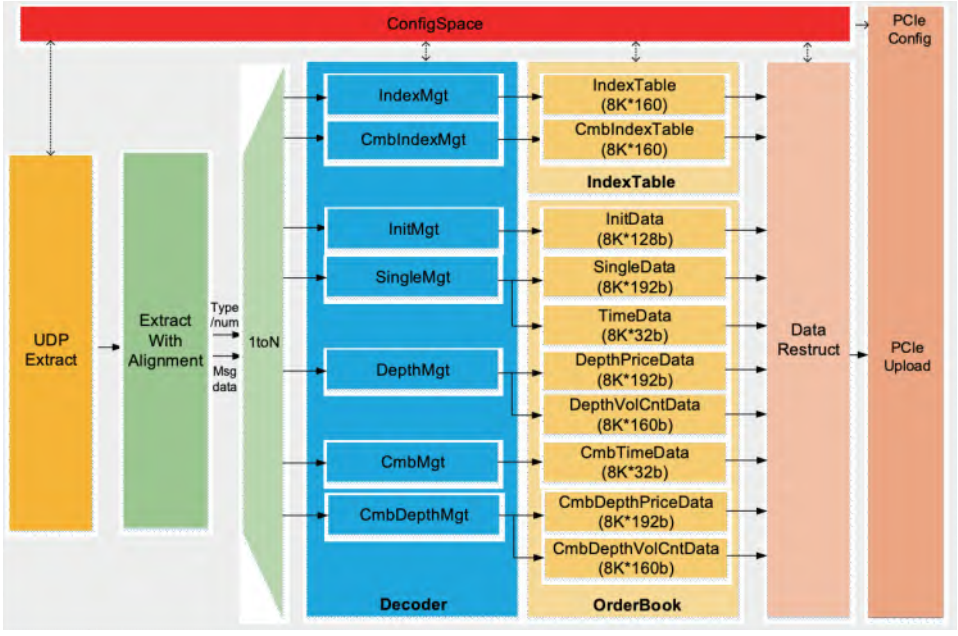


图 2：郑商所行情解码加速器整体结构

3 FPGA 实现

3.1 信息提取与对齐

信息提取与对齐模块首先将报文头部剥离，提取报文类型、消息个数和报文长度字段，然后逐个消息提取消息长度和消息体数据。由于可能存在多个报文以及多个消息混合在一起传输，不能保证行情数据包是 64 位对齐的，为了方便后级逻辑识别和处理，模块将以消息为单位对有效数据进行提取和 64 位对齐处理。处理过程的状态机跳转如图 3 所示。

信息提取处理过程的核心操作是对报文结构的解析，状态机首先需要对报文尾部进行检测，当报文尾部出现时，对报文结构进行分类解析。报文尾部出现时可能存在以下几种典型状态：

1）数据仅包含当前报文尾部：该状态在下一拍时直接处理新的报文即可；

2）数据包含当前报文尾部以及下一报文的部分报文头部：该状态寄存下一报文的部分报文头部，在下一拍时解析其余报文头部、报文正文；

3）数据包含当前报文尾部以及下一报文的报文头部：该状态寄存下一报文的报文头部，在下一拍时解析报文正文；

4）数据包含当前报文尾部以及下一报文的报文头部、报文正文：该状态寄存下一报文的报文头部，部分报文正文，在下一拍时解析剩余报文正文。

如果未检测到报文尾部，则需要对消息尾部进行检测，当消息尾部出现时，对消息结构进行分类解析。消息尾部出现是可能存在以下几种典型状态：

1）数据仅包含当前消息尾部：该状态在下一拍时直接处理新的消息即可；

2）数据包含当前消息尾部以及下一消息的部分消息头部：该状态寄存下一消息的部分消息头部，在下一拍时解析其余消息头部、消息正文；

3）数据包含当前消息尾部以及下一消息的消息头部：该状态寄存下一消息的消息头部，在下一拍时解析消息正文；

4）数据包含当前消息尾部以及下一消息的

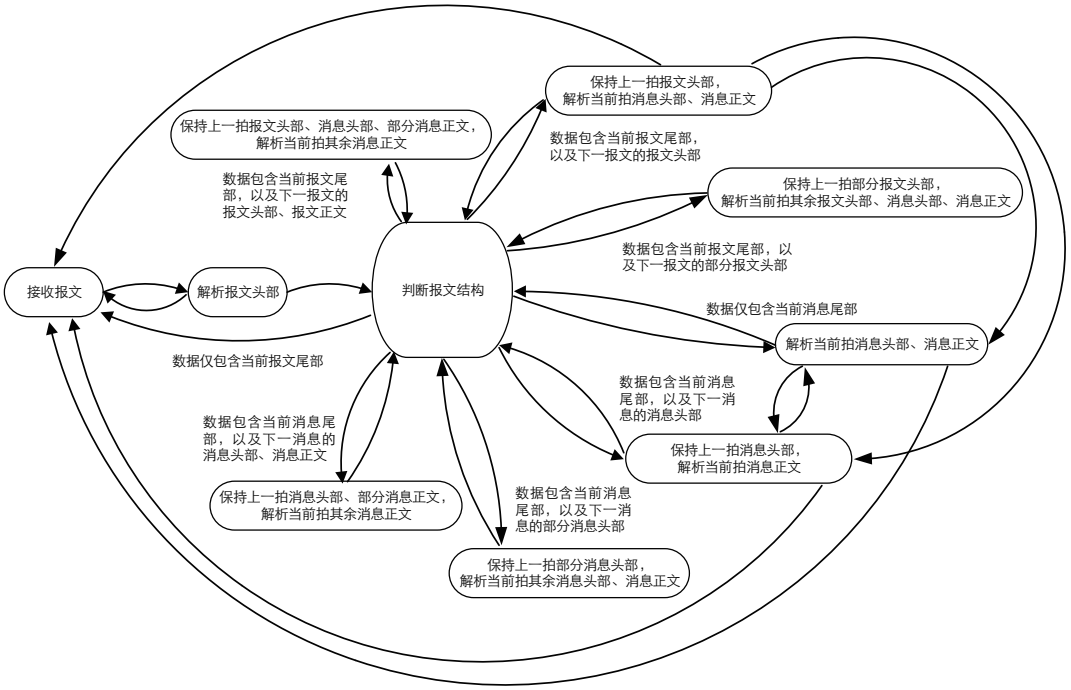


图 3：信息提取处理过程状态机跳转

消息头部、消息正文：该状态寄存下一消息的消息头部，部分消息正文，在下一拍时解析剩余消息正文。

在完成有效信息提取后进行 64 位对齐处理即可进行解码操作。

3.2 行情解码

行情解码模块主要依据报文类型和数量字段对行情数据进行区分，再对不同类型数据分别进行解码。做市业务主要关注如表 1 所示的合约索引报文、初始行情报文、单腿行情报文、组合行情报文、单腿深度行情报文和组合深度行情报文几种类型。

表 1：报文类型对应关系

值	类型	描述
0x05	PACKAGE_INSTRUMENT_IDX	合约索引报文
0x06	PACKAGE_INSTRUMENT_INIT	初始行情报文
0x10	PACKAGE_INSTRUMENT	单腿行情报文
0x11	PACKAGE_CMBTYPE	组合行情报文
0x20	PACKAGE_INSTRUMENT_DEPTH	单腿深度行情报文
0x21	PACKAGE_CMBTYPE_DEPTH	组合深度行情报文

3.2.1 合约索引信息管理

FPGA 内部需要存储所有合约索引（Index）对应的合约编码（InstrumentId）。根据目前交易所的合约数量要求，我们预留最大 8192 个合约编码的存储空间。

合约索引信息报文从系统启动开始至系统关闭，期间每 50 毫秒推送一个 UDP 数据包，包内包含一个报文，报文内包含多个消息，消息的结构如表 2 所示。合约索引的分配是从 0 自增的连续数据，因此设计合约索引表时，直接将合约索引当做存储地址，通过存储地址访问合约索引表内的合约编码。

硬件启动后，索引信息管理模块自动将合约索引表全部清 0。接收行情后根据合约类型区分出是单腿合约还是组合合约，再解析出 Index 和 InstrumentId，然后用 Index 做地址，把 InstrumentId 分别写入对应的合约索引表的 SRAM 中。

合约索引行情主要区分首包和其他包，如图 4 所示，状态机在解析完首包的包头和包体之后，继续解析其他包的包头和包体，直到全部行情解析完成。

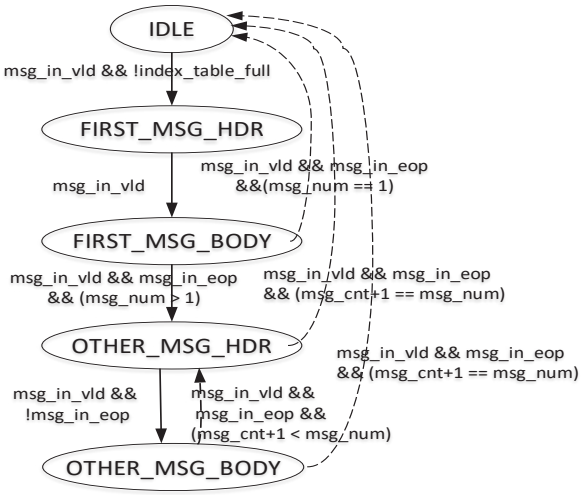


图 4：合约索引信息解析状态机

3.2.2 初始行情管理

初始行情报文从系统启动开始至系统关闭，期间每 50 毫秒推送一个 UDP 数据包，每个 UDP 数据包内包含一个报文，报文内包含多个消息。每个消息由多个字段组成，每个字段长度为 4 字节，最多可以包含 7 个字段。初始行情报文结构如图 5 所示。

表 2：合约索引信息报文结构

字段类型	数据类型	字段长度	备注
交易日	uint32	4	仅在每个报文中的第一个消息中有该字段，如：20210222
合约类型	uint8	1	0-单腿合约；1-组合合约
合约索引	uint16	2	单腿合约与组合合约索引彼此独立，均从 0 开始
合约编码	string	变长	如：期货为 SR909、期权为 SR009C4900、组合为：SPD-AP005/AP007

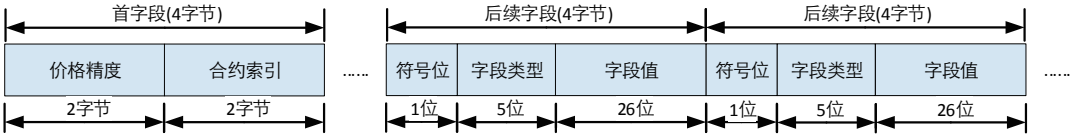


图 5：初始行情报文结构

本设计主要关注并解析如表 3 所示的字段类型。在对初始行情进行解析时，主要区分首字段和其他字段，状态机在处理完首字段后，继续处理其他字段。当处理完所有字段后，将行情中解析出的 4 种类型字段值组合成，通过 Index 进行索引，对订单簿中的初始行情 RAM 进行更新。

表 3：初始行情报文字段类型表

值	类型	描述
1	LastClose	昨收盘
2	LastClear	昨结算
4	LimitUp	涨停价
5	LimitDown	跌停价

3.2.3 单腿行情和组合行情管理

单腿行情和组合合约的消息结构和图 5 初始行情报文类似，由多个长度为 4 字节的字段组成，最多包含 23 个字段。第一个字段包含价格精度和合约索引，第二个字段开始需要按位进行解析。本设计主要关注和解析如表 4 所示字段类型。

表 4：单腿行情和组合行情字段类型

单腿行情		
值	类型	描述
4	LastPrice	最新价
9	Volume	成交量
10	OpenInterest	持仓量
16	UpdateTime	更新时间
18	UpdateTimeUsec	更新时间（微秒部分）
19	TradeTurnover1	总成交金额（part1）
20	TradeTurnover2	总成交金额（part2）
组合行情		
值	类型	描述
7	UpdateTime	更新时间
8	UpdateTimeUsec	更新时间（微秒部分）

在对单腿行情进行解析时，主要区分首字段和其他字段，状态机在处理完首字段后，继续处理其他字段。更新时间由秒和微秒两部分组成，通过公式（1）计算得到单位为毫秒的更新时间。总成交金额 TradeTurnover 用两个字段表示，读取时先以整数形式将 part1 和 part2 读出，然后使用公式（2）将两个整数拼接起来，最后根据价格精度将其转换为精确的总成交金额。

$$UpdateTimeStamp = UpdateTime \times 1000 + UpdateTimeUsec \div 1000 \quad (1)$$

$$TradeTurnover = (TradeTurnover1 << 26) \mid TradeTurnover2 \quad (2)$$

当处理完所有字段后，将解析出的最新价、成交量、持仓量、总成交金额、精度等字段组合，通过 Index 进行索引，对订单簿中的单腿行情 RAM 进行更新，同时将更新时间供单腿深度行情管理模块使用。

组合行情的处理方式和单腿行情处理方式类似，但组合行情仅解析更新时间供组合深度行情管理模块使用即可。

3.2.4 单腿深度行情和组合深度行情管理

单腿合约的深度行情由单腿行情报文（0x10）和单腿深度行情报文（0x20）组成。组合合约的深度行情由组合行情报文（0x11）和组合深度行情报文（0x21）组成。深度行情 UDP 数据包结构如图 6 所示。

深度行情报文由多个字段组成，最多可以包含 11 个字段。第一个字段包含价格精度和合约索引，第二个字段开始需要按位进行解析，每个字段长度为 8 字节，前 4 字节解析规则与初始行情报文相同，包含符号位、字段类型、价格，后 4 字节包含该价位的委托量和订单个数，深度行情结构如图 7 所示。

本设计解析如表 5 所示字段类型的深度行情

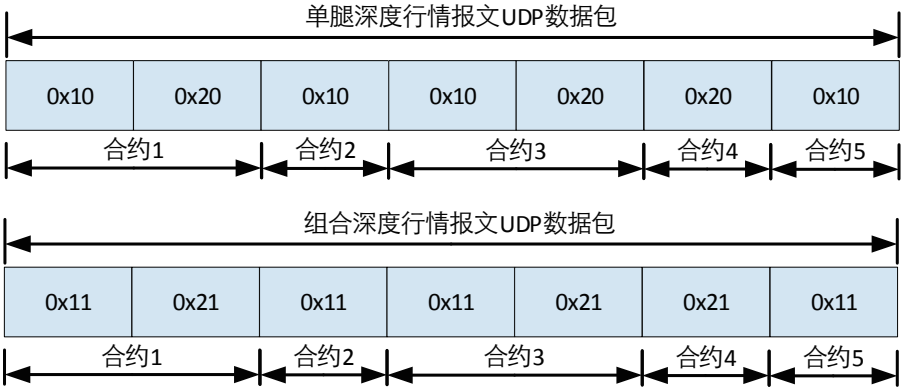


图 6：深度行情 UDP 数据包结构

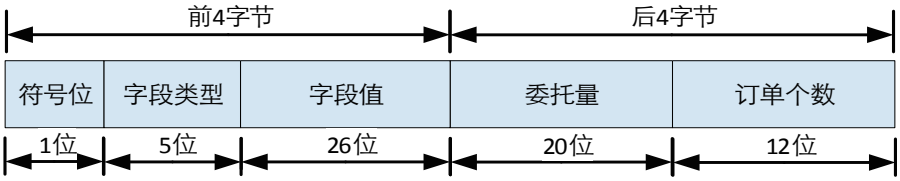


图 7：深度行情消息结构

数据信息，采用面积换速度的设计思路，将深度行情解析为五档买价、五档卖价、五档买量、五档卖量四部分数据，分别通过一拍存储到订单簿中对应存储的 RAM 中，减少了因为数据量大导致的 RAM 存取时间开销。

表 5：深度行情字段类型

值	类型	描述
1	BidDepth1	买 1 价、量、订单数
2	AskDepth1	卖 1 价、量、订单数
3	BidDepth2	买 2 价、量、订单数
4	AskDepth2	卖 2 价、量、订单数
5	BidDepth3	买 3 价、量、订单数
6	AskDepth3	卖 3 价、量、订单数
7	BidDepth4	买 4 价、量、订单数
8	AskDepth4	卖 4 价、量、订单数
9	BidDepth5	买 5 价、量、订单数
10	AskDepth5	卖 5 价、量、订单数

如图 8 所示，状态机在处理完最后一个字段后，需要获取当前报文的更新时间，再通过 Index 索引到上一次推送相同合约行情时的时间来进行对比。如果时间相同，则不进行更新，如果时间不同，则向后级推送该行情，并将更新时间通过 Index 索引对 Orderbook 中的时间戳管理 RAM 进行更新。深度行情管理模块在每处理完一个深度行情时产生一个 FinishQ 和 Finish_Data 信号，通知数据规整化模块从订单簿中获取并上传最新的行情数据。

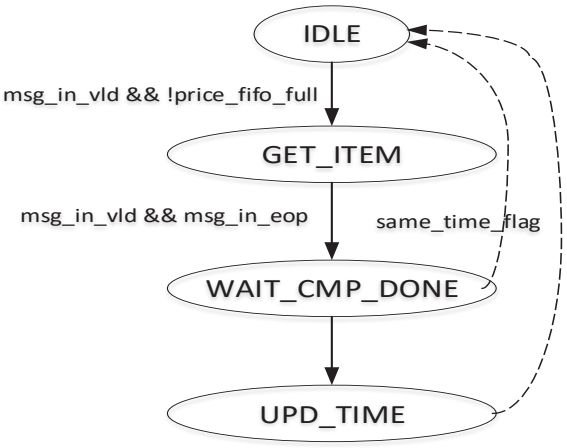


图 8：深度行情解析状态机

3.3 合约索引表和订单簿

合约索引表和订单簿均通过 Index 值作为地址进行索引。

合约索引表用来维护合约编码，分为单腿合约索引表和组合合约索引表两部分，合约编码最大位宽是 20 字节，均存储在位宽 20 字节、深度 8192 的双口 SRAM 中。合约索引信息管理模块每收到一个合约索引消息，该模块都会按合约索引作为存储地址将合约编码写入 SRAM；数据规整化模块每完成一个深度行情的更新，该模块都会通过合约索引在对应的存储地址读取一份合约编码。

订单簿用来维护合约索引、初始行情、单腿行情、组合行情、单腿深度行情、组合深度行情等信息，其他行情都会被 FPGA 自动过滤。为了提高并行度，订单簿将不同类型的数据分别存储于不同的 SRAM 中，再通过 Index 索引到对应的 SRAM。根据策略需要，FPGA 在订单簿内部维护单腿行情、组合行情的最新数据，再通过单腿深度行情、组合深度行情的 FinishQ 信号触发所有行情信息的合并及上传动作。

3.4 数据规整化

数据规整化（DataRestruct）模块分为单腿深度规整化模块和组合深度规整化模块两部分，主要用于读取两类深度行情的订单簿数据并进行精简规整化输出。如图 9 所示，上半部分为单腿深度规整化模块的处理流程，下半部分为组合深度规整化模块的处理流程，两个模块的行情输出结果通过 Mux2to1 模块进行 Round Robin 轮询调度后输出给 PCIe DMA 接口。

数据规整化模块首先检查对应的 FinishQ 信号，每当产生 FinishQ 信号时，根据 Finish_Data 存储的合约索引判断待读取的 SRAM 地址，单腿深度行情从订单簿中读取并整合 UpdateTimeStamp、InstrumentId、InitData、SingleData、DepthPriceData、DepthVolCntData 数据存入对应的 FIFO 中，再根据 FormOut 模块的状态机填充精简数据结构并输出到后级 MUX 轮询，组合深度行情整合 CMBUpdate-TimeStamp、CMBInstrumentId、CMBDepthPriceData、CMBDepthVolCntData 数据存入对应的 FIFO 中，再根据 CMBFormOut 模块的状态机填充精简数据结构并给到后级 MUX 轮询输出。

本设计中用于 PCIe DMA 上传的精简规整化

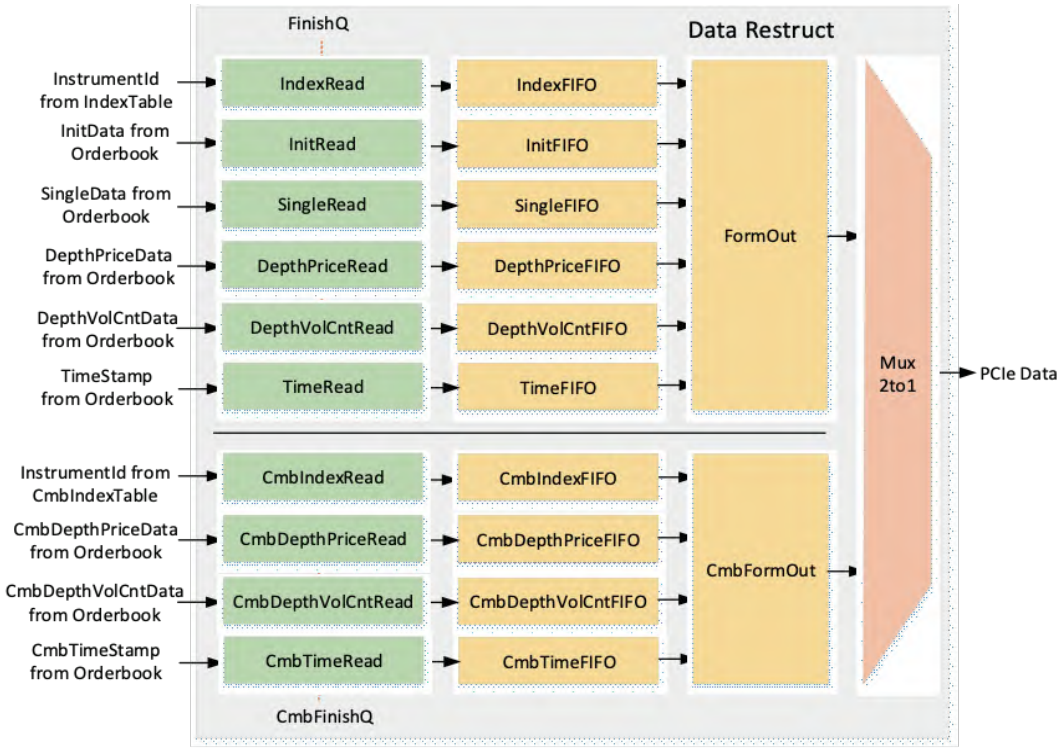


图 9：数据规整化模块处理流程

数据结构如表 6 所示，填充时每一拍（pat）向 PCIe DMA 接口输出 256bit 位宽的行情数据，总共 6 拍完成一个深度行情模板的输出。模板通过 Tag 字段区分两种类型的行情，通过 Decimal 字段维护五档行情的精度，数据结构尽可能根据软件需求进行字段填充，将同类型的数据聚集在一起方便策略服务器进行 Cache 读取。模板内预留了 Rsv 位用于填充未来策略算法可能需要的行情字段，并在专门设计预留了 8 字节的 rsv 位用于填充 pcie header 报文头部信息，最大限度保证策略服务器内存的 Cache 对齐，从而减小处理器读取数据时的延迟开销。

表 6：数据规整化后的精简数据结构

PCIe 分组序号		高 32 位	低 32 位
Pat5	PCIe[255:192]	pcie header	
	PCIe[191:128]	{Askvolume5, AskCount5}	{Bidvolume5, BidCount5}
	PCIe[127:64]	{Askvolume4, AskCount4}	{Bidvolume4, BidCount4}
	PCIe[63:0]	{Askvolume3, AskCount3}	{Bidvolume3, BidCount3}
Pat4	PCIe[255:192]	{Askvolume2, AskCount2}	{Bidvolume2, BidCount2}
	PCIe[191:128]	{Askvolume1, AskCount1}	{Bidvolume1, BidCount1}
	PCIe[127:64]	AskPrice5	BidPrice5
	PCIe[63:0]	AskPrice4	BidPrice4
Pat3	PCIe[255:192]	AskPrice3	BidPrice3
	PCIe[191:128]	AskPrice2	BidPrice2
	PCIe[127:64]	AskPrice1	BidPrice1
	PCIe[63:0]	func(TradeTurnover1, TradeTurnover2)	
Pat2	PCIe[255:192]	Rsv	Rsv
	PCIe[191:128]	Rsv	Rsv
	PCIe[127:64]	LastPrice	Depth_Decimal
	PCIe[63:0]	Rsv	Tag
Pat1	PCIe[255:192]	LimitDownPrice	LimitUpPrice
	PCIe[191:128]	LastClearPrice	LastClosePrice
	PCIe[127:64]	Rsv	Rsv
	PCIe[63:0]	Rsv	Rsv
Pat0	PCIe[255:192]	OpenInterest	Volume
	PCIe[191:128]	InstrumentID[159:96]	
	PCIe[127:64]	InstrumentID[95:32]	
	PCIe[63:0]	InstrumentID[31:0]	UpdateTimeStamp

4 性能分析

郑商所行情解码加速器性能方面主要关注穿透延迟、吞吐量和抖动。板级测试方法如图 10 所示。郑商所行情源回放服务器产生的行情通过一层交换机分为 TCP 行情和 TCP 镜像行情，策略服务器安装一块 SolarFlare 网卡和一块 FPGA，网卡连接 TCP 行情用于软件的网络行情接入，FPGA 连接 TCP 镜像行情用于硬件的网络行情接入。FPGA 解码完成后通过 PCIe 接口将行情存储在策略服务器的内存中，策略软件在解码完成后将结果和内存中 FPGA 的输出结果通过结果比较函数进行对比，计算出平均时延和胜率。

策略服务器选用 Dell R760 服务器，CPU 型号为 Intel Xeon Gold 6226R，主频 2.90GHz，32 核。网卡型号为 Solarflare SFC9220 10/40G。行情解码加速器选用 Xilinx U50 系列 FPGA 板卡，集成自研通用 MAC、UDP 协议栈和 PCIe DMA 模块，FPGA 板卡工作频率可以达到 380MHz，支持 10G 网络接入。测试时分别在行情解码加速器的入口和出口处打点，其中 M1 为行情在 UDP 协议栈出口处时间戳打点，M2 为行情在 PCIe DMA 入口处时间戳打点，M2 和 M1 的时间差值便是行情解码加速器的硬件穿透延迟。根据实际测试结果，郑商所行情解码加速器的硬件穿透延迟约为 79ns。

吞吐量受限于整条通路的瓶颈位置，抖动取决于吞吐出现瓶颈时待处理行情的平均排队时间。堵塞与否是由网络行情大规模输入时，解码模块处理速度和报文信息提取对齐模块处理速度是否匹配决定的。本设计解码模块的处理速度能够满足信息提取对齐模块的处理速度，实现流水处理，因此不会造成反压和阻塞的情况。根据实际测试结果，郑商所行情解码加速器在 10 倍速行情回放压力测试中，无数据反压和丢包现象，输出零抖动。

使用郑商所生产仿真测试环境进行性能和正确性比对测试，延迟和胜率测试结果如表 7 所示，在输出约 376 万个深度行情的情况下进行对比，FPGA 解码方案和生产环境所使用的低延迟网卡 + 软件解码方案输出结果完全一致，并且具备较高的胜率和一定的延迟优势。

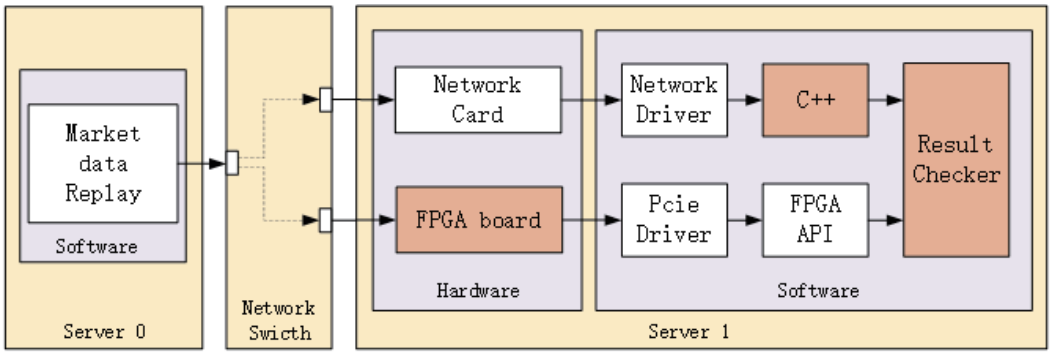


图 10：软硬件性能比对测试框图

表 7：FPGA 硬件解码方案与网卡 + 软件解码方案性能比对结果

输出行情类型	输出行情数量	FPGA 延迟优势	FPGA 胜率
单腿深度行情	3606298	1052ns	96.6%
组合深度行情	162642	582ns	92.4%

5 结束语

本文针对郑商所期货期权做市业务需求，提出并设计实现了一种基于 FPGA 的行情解码加速器，并且在实现行情解码的基础上实现了数据规整化等定制化功能，使得策略服务器接收到的数据可以直接进行使用，不需要进行二次处理，优化了行情解码的延迟水平。通过和目前生产环境使用的 SolarFlare 网卡和软件解码方案进行性能比较，FPGA 硬件方案具备较高的胜率和一定的延迟优势，对提高做市业务水平具有一定的价值。

基于隐私计算的风险监测算法研究

袁梦泽、颜挺进、李乔、陈林博 / 中国证券登记结算有限责任公司上海分公司 技术开发一部 上海 200127
E-mail：mzyuan@chinaclear.com.cn

数字化转型依赖于全局性、多视角的数据价值融合，在此过程中，保护各市场参与主体权益与挖掘数据价值间的矛盾日益凸显。为了在数据有序共享、综合应用的同时确保信息安全与合规，本文探索隐私计算技术，研究内容如下：

针对金融宏观监测中信息割裂及监测平台难以复用的问题，提出基于联邦学习融合同态加密的隐私计算算法原型，以隐私计算中联合建模技术为主体，通过各项支撑技术实现跨领域信息融合，探索隐私计算在释放数据要素、进行风险监测的可行性。

关键词：隐私计算；风险监测；同态加密；联邦学习

1 绪论

1.1 研究背景与意义

在数字经济时代，数据价值进一步凸显，成为主导经济发展的关键生产要素。从数据规模和量级看，我国构建全球领先的超大规模数据市场各项条件已经具备。《金融科技发展规划（2022-2025 年）》中指出：“把握数字经济发展新趋势，发挥数据要素倍增作用，将数字元素注入金融服务全流程，将数字思维贯穿于业务运营全链条。”党中央、国务院高度重视数据要素市场的培育发展，近年来出台了一系列政策措施积极完善基础制度建设，对繁荣数据要素市场提出了明确的要求。

在此背景下，各金融机构积极响应政策指示，激活数据要素潜能，加快推进数字化转型。数字化转型依赖于全局性、多视角的数据价值融合，在此过程中，维护资本市场秩序、保护各市场参

与主体权益与挖掘数据价值间的矛盾日益凸显，如何在发挥数据融合价值的同时、强化数据安全，已成为数字化转型的关键环节。

作为提升数据安全及隐私保护水平的重要技术手段，隐私计算技术可分离“数据可见的信息部分”和“无需看见就可计算的使用价值”，实现“数据可用不可见”，基于此，数据流通主体可以不再是明文数据本身，而是数据特定使用价值。

综上，为释放数据要素红利、共同搭建数据流通基础设施，本文研究隐私计算技术，针对金融宏观监测中信息割裂及监测平台难以复用的问题，提出基于联邦学习融合同态加密的隐私计算平台，以隐私计算中联合建模技术为主体，融合各项跨领域信息。

1.2 研究现状

1.2.1 大数据计算场景应用现状

在数据输入阶段，学术界提出了根据数据的敏感性分离存储和计算数据的解决思路，典型代表方案为联邦学习。经典的联邦学习开源框架包括 OpenMind 提出的 PySyft，微众银行提出的 FATE，百度提出的 Paddle FL，牛津大学提出的 Flower 等。

在数据计算过程中，为了保证数据的机密性和计算隐私性，目前行之有效的手段是对传输数据进行加密，并结合安全多方计算、硬件增强或者访问模式隐藏等主流方法实现隐私计算。针对基于混淆电路的安全多方计算，Goyal 等人提出了一种针对隐蔽敌手的安全两方计算方案，采用剪切和选择技术减少两方之间传输的数据量和通信步骤数；Huang 等人引入流水线电路执行的思想，通过并行处理部分电路的混淆步骤和估值步骤，实现了更快速的安全两方计算协议。

1.2.2 金融业应用现状

当前国内外金融业已有的隐私计算技术应用主要有两种：一是通过联邦学习实现的联合建模；二是使用多方安全计算技术直接实现隐私查询、联合建模及联合统计。

金融业的联邦学习试点应用在模型训练环节将交换的梯度或参数进行加密，以保证各建模参与方的数据隐私性。平安科技利用自身的用户行为数据，和银行或保险机构的客户金融数据进行联邦学习建模，建立更精准的保险产品定价；江苏银行基于联邦学习技术对智能化经营进行了联合开发和方案部署，共同进行金融风控模型训练，提升模型效果；美国金融科技公司 Consilient 与 Intel 合作建立联邦学习反洗钱平台，各参与银行在其可信执行环境内明文训练本地模型，并将中间梯度或参数以加密形式传输至中心计算服务器的可信执行环境内进行聚合并计算。

金融业的多方安全计算可服务隐私查询、联合统计、数据交易等计算场景。中国农业银行启动了两个试点项目：一是为跨行跨境的信贷场景提供匿踪查询服务、多方统计服务；二是研究解决电信诈骗黑名单、反欺诈、反洗钱等隐私计算场景。英国金融行为监管局在 2019 年举办了全球反洗钱和反欺诈技术竞赛，获胜小组应用基于多方安全计算的技术方案，解决了欺诈行为识别、客户身份识别、交易关联分析等场景的数据壁垒痛点。交通银行在普惠金融、联合风控、精准营销等一系列方面都应用了隐私计算技术，包括：基于多方安全知识图谱计算的中小微企业融资服务，通过多方安全图计算等技术作为开户真实性意愿审核的辅助手段等。

2 相关技术简介

2.1 隐私计算概述

2016 年，李凤华等学者提出：隐私计算是面向隐私信息全生命周期的计算理论和方法，是在隐私信息的所有权、管理权和使用权分离的前提下面向隐私度量、隐私保护与隐私分析复杂性的可计算模型与公理化系统。抛开学术定义，隐私计算是指在保护数据本身不对外泄露的前提

下，实现数据分析计算的一类技术集合。

Gartner 公司认为，按照实现的功能，隐私计算技术可分为以下三类：提供可信的环境来执行处理或分析；在处理或分析之前转换数据和 / 或算法；在不公开数据的情况下执行数据本地处理或分析。

2.2 联邦学习概述

联邦学习是一种多个参与方在数据不出本地的前提下共同完成某项机器学习任务的活动。通过联邦学习，不同的数据拥有方可以在不交换彼此数据的情况下，建立一个虚拟的共有模型，其效果等同于各方把数据聚合在一起建立的模型。为避免交换中间因子导致原始数据泄露和反推，联邦学习多与多方安全计算、同态加密、差分隐私等密码学技术结合，对交互的中间因子进行加密和保护。

根据各参与方数据特点的不同，联邦学习可分为横向联邦学习、纵向联邦学习和联邦迁移学习。横向联邦学习适用于特征重合较多、样本重合较少的数据集间计算，比如不同地区的银行间，他们的业务相似（特征相似），但用户不同（样本不同）。以样本维度（即横向）对数据集进行拆分，以特征相同而样本不完全相同的数据部分为对象进行训练，如图 1 所示。

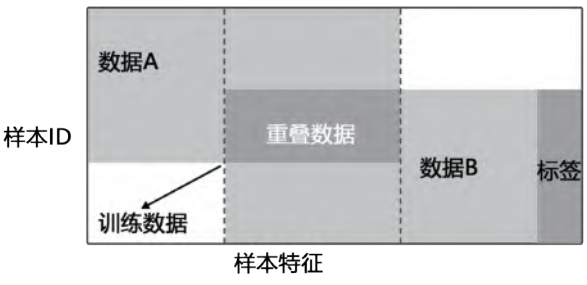


图 1：横向联邦学习

纵向联邦学习适用于样本重合较多、而特征重合较少的数据集间联合计算的场景，比如同一地区的商超和银行，其用户都为该地区的居民（样本相同），但业务不同（特征不同）。以特征维度（即纵向）对数据集进行拆分，以样本相同而特征不完全相同的数据部分为对象进行训练，如图所示。

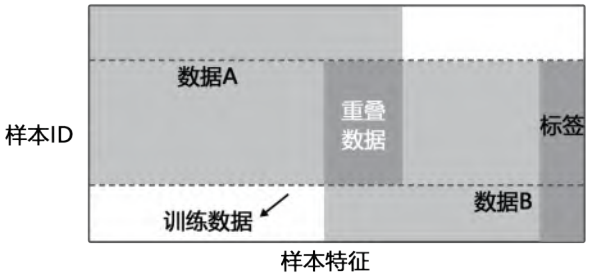


图 2：纵向联邦学习

联邦迁移学习则适用于数据集间样本和特征重合均较少的场景，如不同地区的银行和商超间的联合，主要适用于以深度神经网络为基模型的场景。在这样的场景中，不再对数据进行切分，而是利用迁移学习来弥补数据或标签的不足，如图所示。

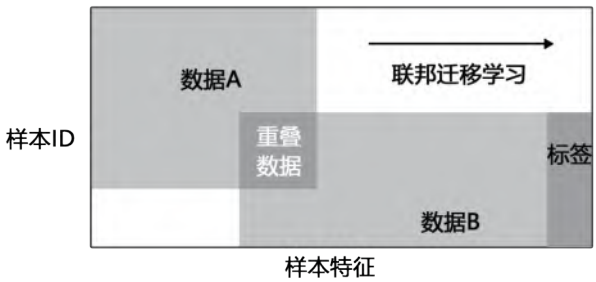


图 3：联邦迁移学习

3 基于联邦学习融合同态加密的风险监测算法

3.1 风险监测目标

金融科技蓬勃发展，在为金融发展注入新动能的同时，也使金融风险更易跨机构、跨业态、跨市场外溢，传播速度明显加快，给传统金融监测体系和监测手段带来新的挑战。传统风险监测面临信息割裂的困境。一是金融机构只将信息上报给相关监测主体，单个监测主体难以掌握金融风险全貌；二是建立新的统一监测平台，可能造成监管资源难以复用。

可以看出，需要探索无需归集各业务数据、而以各项支撑技术实现跨领域信息融合的方案，实现数据的及时性、完整性监测，同时节约监测

基础设施建设成本。

3.2 算法设计

对于非实时性风险监测平台，通常 T+1 日核对 T 日发生的交易数据即可满足需求，涉及交易金额、交易人身份证号等信息，属于隐私数据，且数据体量大，对建模准确性有较高的要求。综合考虑采用联邦学习算法，同时为了增加对协调方和参与者的信任，该方案还将结合同态加密，防止交换梯度造成原始数据的泄露。

3.2.1 算法描述

在典型的联邦学习场景中，记参与方 A 拥有特征 X_1 、X_2，无法独立建模；参与方 B 拥有特征 X_3 、Y。二者建立联合模型，效果会超过参与方 B 单独建模。

以逻辑回归（Logistic Regression）为例，其损失函数和梯度公式如下所示：

$$J_S(\theta)=\frac{1}{n}\sum_{i\in S}\log\left(1+e^{-y_i\theta^T x_i}\right) \quad (4)$$

$$\frac{dJ}{d\theta}=-\frac{1}{n}\sum_{i\in S}\left[y_i\log\left(\frac{1}{1+e^{-\theta^T x_i}}\right)+(1-y_i)\log\left(1-\frac{1}{1+e^{-\theta^T x_i}}\right)\right] \quad (5)$$

其中，各参与方需要交换 x_i 、y_i 以计算各自的梯度和损失，为解决信息明文传输带来的隐私泄露风险，引入同态加密算法，即：对密文处理的结果、与明文处理后再加密结果相同，本文基于 Paillier 算法，该算法属于部分同态加密，支持加法与常数乘法运算，根据泰勒展开可得：

$$\log(1+e^{-z})=log2-\frac{1}{2}z+\frac{1}{8}z^2-\frac{1}{192}z^4+O(z^6) \quad (6)$$

代入上式，得：

$$\frac{dJ}{d\theta}=\begin{bmatrix}X_A^T\left(\frac{1}{4}X_A\theta_A^T+\frac{1}{4}X_B\theta_B^T-y+0.5\right)+\lambda\theta_A\\ X_B^T\left(\frac{1}{4}X_A\theta_A^T+\frac{1}{4}X_B\theta_B^T-y+0.5\right)+\lambda\theta_B\end{bmatrix}\times\frac{1}{n} \quad (7)$$

希望参与方 A、B 能够尽量地进行单独的计算，再通过加密信息交互获得各自的梯度计算结果，因此需要对计算的任务进行一定划分，在每一轮参数更新中，各参与方需要按序进行如下的计算和交互：

1. 参与方 A 和 B 各自初始化自己的参数，协调方 C 生成秘钥对并分发公钥给 A 和 B。

2. 参与方 A 计算$\frac{1}{4}X_A\theta_A^T$，使用公钥加密后发送给 B。

3. 参与方 B 计算$\frac{1}{4}X_B\theta_B^T-y+0.5$，使用公钥加密后发送给 A。

4. 此时 A 和 B 能各自计算

$\left[\left[X_A^T\left(\frac{1}{4}X_A\theta_A^T+\frac{1}{4}X_B\theta_B^T-y+0.5\right)+\lambda\theta_A\right]\right]$和

$\left[\left[X_B^T\left(\frac{1}{4}X_A\theta_A^T+\frac{1}{4}X_B\theta_B^T-y+0.5\right)+\lambda\theta_B\right]\right]$，其中 [[]] 表示同态加密形式。

5.A 和 B 需要加密的梯度发送给 C 来进行解密，但为了避免 C 直接获得梯度信息，A 和 B 可以将梯度加上一个随机数 R_A 和 R_B，再发送给 C。C 获得加密梯度进行后进行解密再返还 A 和 B。

A 和 B 只需要再减去之间加的随机数就能获得真实的梯度，更新其参数。

算法流程如下表所示：

	Party A	Party B	Party C
Step0	初始化θ_A	初始化θ_B	
Step1			生成密钥对，并分发公钥
Step2	$\left[\left[\frac{1}{4}X_A\theta_A^T\right]\right]\rightarrow B$	$\left[\left[\frac{1}{4}X_B\theta_B^T-y+0.5\right]\right]\rightarrow A$	
Step3	计算加密的梯度	计算加密的梯度	
Step4	$[[\frac{dJ}{d\theta_A}+R_A]]\rightarrow C$	$[[\frac{dJ}{d\theta_B}+R_B]]\rightarrow C$	
Step5			$\frac{dJ}{d\theta_A}+R_A\rightarrow A$ $\frac{dJ}{d\theta_B}+R_B\rightarrow B$
Step6	更新参数	更新参数	

3.2.2 算法具体流程

风险监测算法具体流程如图 4 所示，包括协调模块、计算模块以及若干参与方。首先，协调模块发起计算请求，触发联邦学习实例，通过作业调度服务将计算任务同步给计算模块，计算模块向参与方生成并分发公钥；其次，参与方之间分别计算和自己相关的特征中间结果，运用 Paillier 算法加密后互相交互，用来求得各自梯度和损失；随后，参与方分别将加密后的梯度数据添加掩码，转发给计算模块；最后，计算模块解密梯度和损失后回传给参与方，参与方去除掩码，并更新模型。

3.3 原型验证

搭建了基于 Paillier 算法的联邦学习原型，包括服务端和客户端。数据集来自 Kaggle 上一个用于研究金融诈骗的数据集，基于非洲某国家真实的金融交易记录，通过 PaySim 模拟器模拟生成。输入矩阵包括交易时间、交易类型、交易金额、交易双方余额等，输出矩阵为欺诈交易的标志（即“是”或“否”）。

参与联邦学习的五家机构各自拥有一部分数据，每家机构的特征相同，数据内容不同，采用横向联邦学习的训练方法，该算法可保证：（1）每个机构的原始数据和加密数据均不离开本地；（2）机构间仅共享加密后的梯度数据；（3）协调模块不能根据计算过程推断交易数据及数据所属的机构。

Paillier 算法支持加法与常数乘法运算，为验证算法性能，本文在单数加密、单数解密、加密数相加、加密数与非加密数相乘等场景进行测试用例的设计和执行。运行服务器为 24 核，250G 内存。

图 5 展示了密钥大小分别为 1024 bit、2048 bit、4096 bit、8192 bit 时，上述测试场景的耗时情况，可以看出，算法主要耗时出现在加密与解密模块，耗时情况近秒级，随着密钥大小增长为原先的 4 倍时，加解密过程耗时呈指数增长，其

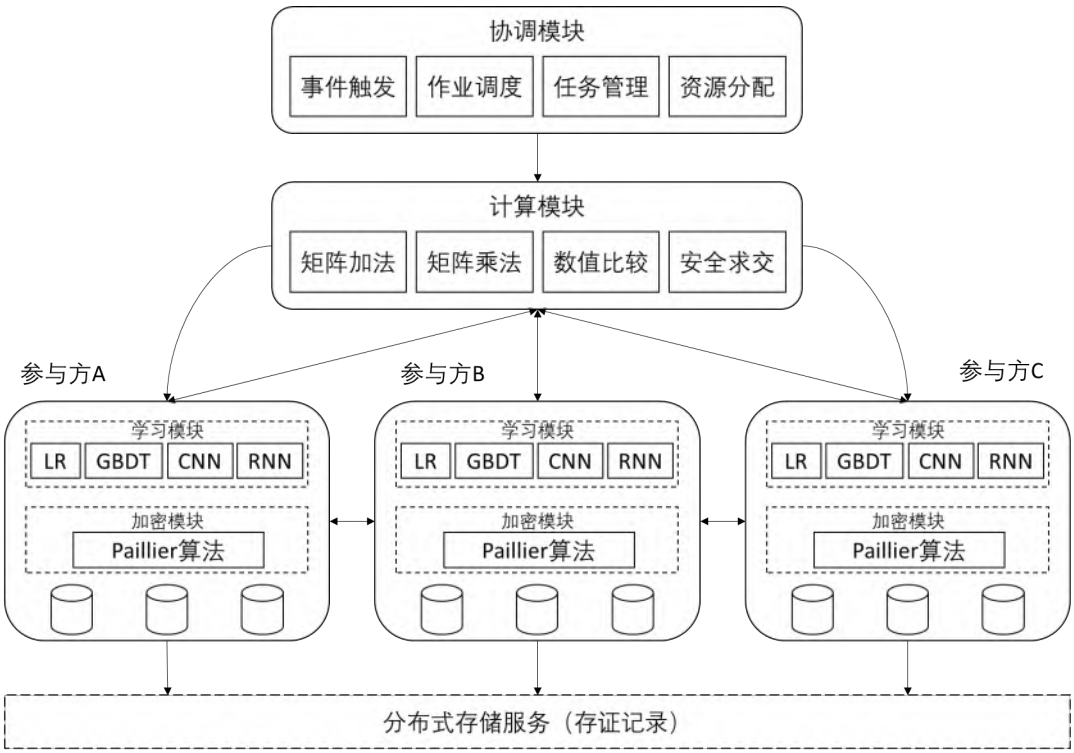


图 4：风险监测算法具体流程

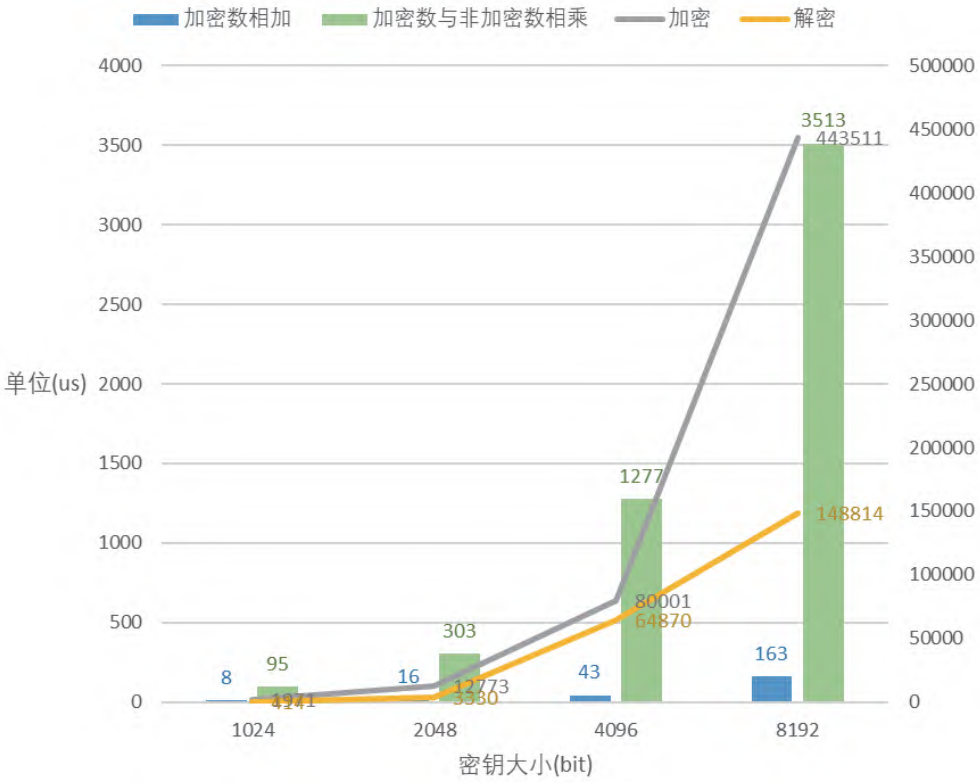


图 5：Paillier 算法在不同密钥大小的耗时情况

中加密过程的耗时增长更明显；另一方面，加密数相加、加密数与非加密数相乘运算耗时情况较为理想，与密钥大小大致呈线性关系。

为了进一步研究算法加解密的性能情况，本章分析了加密的内存占用情况，如图 6 所示，可以看出，加密数字量从 10000 增加到 90000 时，内存占用量呈线性增长；对每个数字来说，加密过程所耗的平均内存基本稳定，最大物理内存平均值为 1.0425 KB。

使用线性回归作为神经网络主体。训练结果表明：相比于本地学习，本章搭建的联邦学习可在数据不出本地、梯度加密的同时，提升模型的

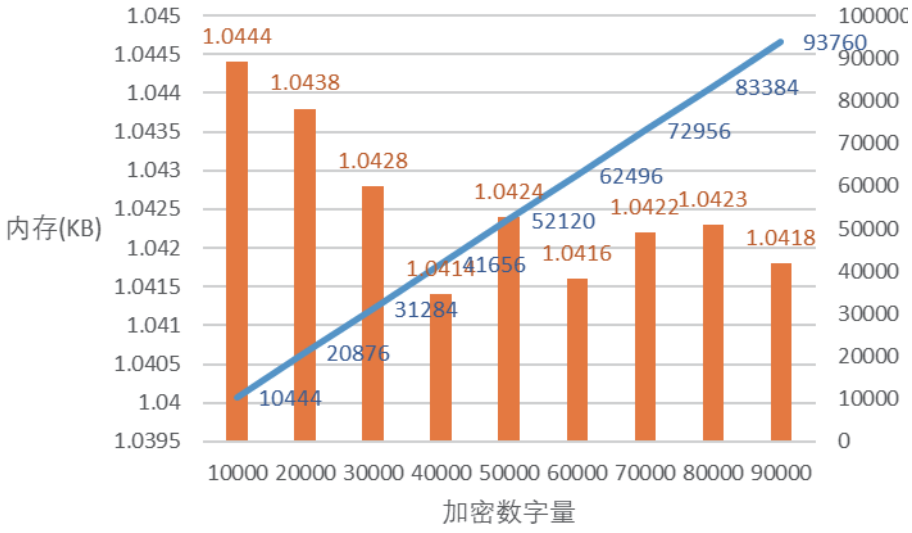


图 6：Paillier 算法在不同密钥大小的内存情况

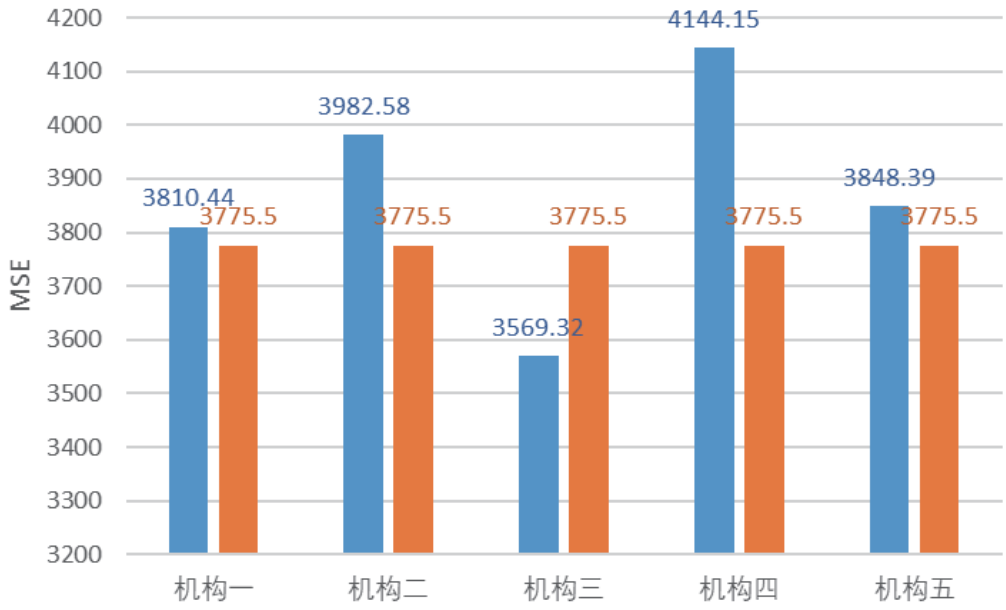


图 7：联邦学习与本地学习均方差对比

准确性。如图 7 所示，均方差（MSE）提高 2.48%，说明联邦学习下模型准确度相对于本地学习更高。

4 总结与展望

数据作为一种新型生产要素正在颠覆全球社会的发展模式，深入数据要素使用、加快数据要素市场培育等工作已提升至国家战略。在数字化转型过程中，各金融机构积极探索、把握机遇，为国家数据要素市场和数据要素新生态建设贡献力量。

隐私计算等新兴技术可以在“数据不出域”的前提下，与参与方共享挖掘数据要素潜在价值，打破了传统技术局限，为深入挖掘数据要素价值提供了必要条件。本文在隐私计算领域成熟的技术路径基础上，探索可行的隐私管理与应用方案，为金融风险监测等场景提供安全可靠的数据保护和计算分析，旨在为数据价值融合、释放数据红利摸索一条可行的路线，助力数字化转型。

需要指出的是，当前隐私计算技术成熟度仍有待提高。虽然隐私计算在金融、保险、互联网等领域已经开展了部分应用，但距离实际应用仍面临技术通用、提升性能、加深业务融合、完善生态等挑战。

隐私计算技术蓬勃发展的同时，仍需在建设数据基础设施的前提下，通过优化算法和协议设计、软硬件协同设计等方式提升计算交互效率，形成数据应用标准和适应数据要素特性的监测模式，使技术真正造福于社会。同时，我们也将继续跟进和探索隐私计算技术的最新发展，致力于在保证数据安全的前提下，最大限度地发挥数据融合价值，打造具备数字化能力的金融技术系统。

参考文献：

[1] Ryffel T, Trask A, Dahl M, et al. A generic framework for privacy preserving deep learning[J]. arXiv preprint arXiv:1811.04017, 2018.

[2] Liu Y, Fan T, Chen T, et al. FATE: An Industrial Grade Platform for Collaborative Learning With Data Protection[J]. J. Mach. Learn. Res., 2021, 22(226): 1−6.

[3] Beutel D J, Topal T, Mathur A, et al. Flower: A friendly federated learning research framework[J]. arXiv preprint arXiv:2007.14390, 2020.

[4] Shiffman G，Zarate J，Deshpande N，Yeluri R，Peiravi P. 2020. Federated Learning through Rev−olutionary Technology[R/OL].https：//consilient.com/white− paper/fed−erated−learning−through−revolu−tionary−technology/.

[5] 苏建明，叶红，吕博良，程佩哲．联邦学习在商业银行反欺诈领域的应用 [J]. 中国金融电脑 ,2021(02):39−42.

[6] 龚光庆．隐私计算技术在银行业的应用探索 [J]. 中国金融电脑，2021.

信息资讯采撷

监管科技全球追踪

监管科技全球追踪

8 月 21 日，全国信息安全标准化技术委员会发布国家标准《信息安全技术 数据安全风险评估方法》征求意见稿。明确评估内容在信息调研基础上围绕数据安全管理、数据处理活动安全、数据安全技术、个人信息保护等方面开展。

8 月 29 日，为维护银行、保险和养老金行业的稳定性，澳大利亚审慎监管局（APRA）公布了其最新计划。该计划概述了 APRA 未来四年的优先事项以及如何应对影响全球金融体系的新型风险和潜在风险。

9 月 4 日，德国联邦金融监督管理局（BaFin）宣布，自 9 月 1 日以来，其官方网站“bafin.de”遭受了分布式拒绝服务（DDoS）攻击，导致其网站无法打开，其余系统未受到影响。

9 月 5 日，人民银行召开加强支付受理终端及相关业务管理工作会。会议指出，人民银行高度重视收单市场管理和发展，深化支付供给侧结构性改革，不断夯实制度框架，净化支付市场环境。

9 月 5 日，金融稳定委员会（FSB）主席在致二十国集团领导人的文件中论述了 2020 年二十国集团加强跨境支付路线图第一阶段工作所取得的成就。

9 月 12 日，美国网络安全和基础设施安全局（CISA）发布开源软件安全图谱，阐述如何在联邦政府内实现开源软件的安全应用，并支持健康、安全和可持续的全球开源软件生态系统。

9 月 13 日，2023 年“科技产业金融一体化”专项路演在雄安新区举办。工信部党组成员、副部长辛国斌指出，举办专项路演，既是落实习近平总书记重要讲话精神的具体实践，也是引导金融资源助力雄安新区实体经济高质量发展的重要举措。

9 月 19 日，英国议会审议通过了《在线安全法案》（Online Safety Bill）草案。该《法案》指出，无论设立在英国本土或海外，只要服务于英国用户，数字和社交媒体服务商都受其制约。《法案》还特别强调了对未成年用户的保护。

9 月 19 日，Visa 和 Swift 宣布合作加强双方资金流动的全球网络连接，从而简化 B2B 国际支付。本次合作旨在为金融机构及其客户在跨境汇款时提供更多选择，同时提高端到端交易的速度和透明度。

9 月 22 日，根据《系统重要性银行评估办法》的要求，中国人民银行、国家金融监督管理总局开展了 2023 年度我国系统重要性银行评估，认定 20 家国内系统重要性银行，其中国有商业银行 6 家，股份制商业银行 9 家，城市商业银行 5 家。

9 月 28 日，中国人民银行、国家金融监督管理总局、中国证券监督管理委员会、国家外汇管理局、香港金融管理局、香港证券及期货事务监察委员会、澳门金融管理局决定进一步优化粤港澳大湾区“跨境理财通”业务试点。旨在优化投资者准入条件、扩大参与机构范围和“南向通”“北向通”合格投资产品范围、适当提高个人投资者额度、优化宣传销售安排等。

9 月 28 日，欧洲央行称正在探索使用类似 ChatGPT 的大语言模型（LLM）进行文档分析和软件测试。欧洲央行对人工智能的使用持“谨慎”态度，并考虑数据隐私、法律约束和道德考虑等问题。

10 月 4 日，国际清算银行（BIS）和多国中央银行合作启动 Mandala 项目，该项目探讨将特定司法管辖区的政策和监管要求编码为跨境使用案例（如外国直接投资、借贷和支付）的通用协议的可行性。

10 月 11 日，国务院发布关于推进普惠金融高质量发展的实施意见，推进普惠金融高质量发展。意见指出，未来五年，基础金融服务更加普及，经营主体融资将更加便利，金融支持乡村振兴更加有力，金融消费者教育和保护机制更加健全，金融风险防控将更加有效，普惠金融配套机制将更加完善。

10月19日与12月5日，伦敦证券交易所（LSE）在两个月内两次发生交易暂停事故，数百支小盘股受到波及，但未影响富时 100 指数、富时 250 指数和国际订单簿的交易。据官方公告，两次事故均与其交易系统第 3 撮合分区的故障有关。

10 月 27 日，人民银行召开提升境外来华人员支付服务水平工作动员部署会。会议要求，各有关单位要全面总结重大赛事支付服务保障工作经验，按照“大额刷卡、小额扫码、现金兜底”的工作思路，以丰富支付服务供给、深化支付场景建设为切入点，加快落实各项工作措施。

10 月 18 日，欧洲数据保护委员会（EDPB）和欧洲数据保护监督机构（EDPS）就数字欧元作为中央银行数字货币的拟议法规发表了联合意见。意见提出，将数字欧元作为现金之外的另一种支付手段，为个人提供在线和离线电子支付的可能性。

11 月 8 日，美国消费者金融保护局（CFPB）提出，建议对提供数字钱包和支付程序等服务的大型非银行公司进行监管，并要求非银行金融公司，尤其是每年处理超过 5 万笔交易的大型公司，与大型银行、信用合作社和其他已经由 CFPB 监管的金融机构遵守相同的规则。

11 月 10 日，工商银行在美全资子公司工银金融服务有限责任公司（ICBCFS）在官网发布声明称，美东时间 11 月 8 日，ICBCFS 遭受了勒索软件攻击，导致部分系统中断。

11 月 20 日，中国人民银行、科技部、国家金融监管总局、中国证监会联合召开科技金融工作交流推进会。会议强调，金融管理部门、科技部门和各金融机构要深入贯彻落实党中央关于加快建设科技强国、金融强国的重要部署，进一步健全国家重大科技任务和科技型中小企业两个重点领域的金融支持政策体系，组织开展科技金融服务能力提升专项行动。

2023年四季度《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，以季度为单位发刊，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。2023 年二季度征稿主题如下：

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、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、多模态大语言模型等。

（二）应用场景研究

主要包含但不限于：基于大语言模型的智能客服、语音数据挖掘、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。

（二）技术实现方式

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

（三）应用场景和案例

主要包含但不限于：结算区块链、信用证区块链、票据区块链等。

（四）安全要求和性能提升

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

五、信息安全与 IT 治理

（一）网络安全

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

（二）移动安全

主要包括但不限于：移动安全管理、移动互联网接入的安全风险、防护措施等。

（三）数据安全

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

（四）IT 治理与风险管理

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

六、交易与结算相关

（一）交易和结算机制

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

（二）交易和结算系统

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

投稿说明

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

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

3、本期投稿截止日期：2024 年 1 月 15 日。

4、投稿联系方式 021-68607129，021-68602496 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。

《交易技术前沿》编辑部

证券信息技术研究发展中心（上海）

附件：投稿格式（可通过电子邮件索要电子模板）

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！

联系电话：021-68607129
021-68602496
投稿邮箱：ftt.editor@sse.com.cn

ITRDC
证券信息技术研究发展中心（上海）
上海證券交易所 SHANGHAI STOCK EXCHANGE
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