金融行业大模型应用探索与实践

上海期货信息技术有限公司 范宏婷 薛利 赵博

一、引言

金融作为国民经济的血脉，其发展历程始终与技术创新紧密相连。近年来，深度学习技术的飞速发展，特别是大模型的广泛应用，为金融行业带来了一场由数据驱动、智能决策引领的变革。

大模型的兴起是人工智能技术发展史上的一个重要里程碑。这些模型通过海量的数据训练，获得了强大的文本处理、语义理解和生成能力，能够像人类一样理解并处理复杂的语言信息。在金融领域，利用专业数据训练后的大模型，可以使得金融数据的处理与分析不再局限于传统的统计方法和简单规则，并且能够深入到数据的本质，挖掘出隐藏其中的价值信息和潜在规律。

对于金融数据的深度挖掘而言，大模型的出现极大地提高了数据处理的效率和准确性。传统上，金融数据分析往往依赖于专家经验和复杂模型，而大模型则能够自动学习数据中的特征，发现隐藏的关联和模式，为金融机构提供更加全面、深入的市场洞察。这种能力不仅有助于金融机构更好地理解市场动态，还能够为风险管理提供更加科学、高效的工具。

此外，大模型还具备实时监测市场动态的能力。通过实时分析新闻、社交媒体等渠道的信息，大模型能够及时发现并预警潜在的市场风险或机遇，为金融机构提供及时有效的决策支持。这种能力在金融市场波动频繁、信息瞬息万变的今天尤为重要，有助于金融机构更好地把握市场脉搏，做出更加明智的风险管理决策。

本文旨在探索大模型在金融行业的应用潜力与实践路径，一方面分析当前大模型的发展概况和技术原理，探讨大模型应用中主要技术攻关方向，形成可复用的方法论，丰富大模型在金融行业的应用理论体系，为后续研究与实践提供理论支撑；另一方面，讨论当前金融行业大模型应用实践及痛点问题，提出探索利用大模型赋能金融行业科技监管的实践路径，促进金融创新，提升金融市场监管效能，增强风险管理能力，推动金融行业高质量发展。下文将从技术理论基础、技术攻关方向、应用探索实践、痛点问题、总结与展望这五个方面进行阐述。

二、技术理论基础

近年来，大模型经历了从萌芽初探到蓬勃兴起的飞跃。本节将从大模型发展历程、技术原理等方面对大模型技术理论基础进行论述。

（一）大模型的发展概况

大模型作为“大数据+大算力+强算法”结合的产物，以其强大的数据处理和分析能力，在自然语言处理、计算机视觉、语音识别等领域展现出巨大潜力。本节将从国外和国内两个维度，综述大模型的发展历程及发展趋势。

1. 国外大模型发展概况

（1）发展历程

国外大模型的发展始于深度学习技术的突破，尤其是以Transformer架构为基础的预训练模型的出现。自2018年OpenAI发布GPT（Generative Pre-Trained）-1以来，大模型技术经历了从预训练模型到大规模预训练模型，再到超大规模预训练模型的演进过程。其中，GPT系列模型以其卓越的自然语言处理能力引领了行业潮流。

（2）发展趋势

多模态发展：国外大模型正朝着多模态方向发展，旨在实现文本、图像、语音等多种模态数据的综合理解和生成。例如，GPT-4已经具备多模态理解与多类型内容生成能力，Sora具备文生视频的能力，未来更多跨模态大模型将不断涌现。

参数量增加：计算能力的提升和数据量的增长为大模型参数量的不断攀升提供了必要的条件。从GPT-1的1.17亿参数到GPT-4的数千亿参数，大模型的性能显著提升，能够更好地捕捉数据中的复杂模式和规律。

2. 国内大模型发展概况

（1）发展历程

国内大模型的发展起步较晚，但近年来呈现出加速追赶的态势。自2022年OpenAI发布ChatGPT以来，国内科技巨头和科研机构纷纷加大投入，推出了一系列具有竞争力的大模型产品。国内大模型在过去几年中取得了实质性突破：以百度文心一言、阿里通义千问、科大讯飞星火、清华智谱ChatGLM4、月之暗面KIMI等为代表的第一梯队大模型，整体能力已逼近GPT-4。尤其是在中文领域，部分模型的表现已可比肩GPT-4。

（2）发展趋势

长文本处理：国内大模型在长文本处理方面取得了显著进步。今年3月18日，月之暗面KIMI支持200万长文本输入；3月22日阿里宣布通义千问将向所有用户免费开放1000万字的长文档处理能力；3月23日，360智脑宣布正式内测500万字长文本处理能力。这些模型在理解、生成长文本方面表现出色，能够应对复杂对话和文档处理任务。

多模态探索：国内大模型也在积极探索多模态应用。如商汤科技的“日日新SenseNova5.0”大模型在图文感知能力上达到全球领先水平；昆仑万维发布的文生音乐大模型天工Skymusic，展现了多模态生成能力；智谱清言的清影智能体可实现文生视频、图生视频；月之暗面KIMI可作为PPT助手，辅助办公。

（二）大模型的技术原理

1. 大模型技术架构

大模型通常指的是包含超大规模参数（通常在十亿个以上）的神经网络模型，其核心特点包括庞大的规模、涌现能力、更好的性能和泛化能力、多任务学习、大数据训练、强大的计算资源需求、迁移学习和预训练能力、自监督学习、领域知识融合等。大模型通过在大规模数据集上进行预训练，然后在特定任务上进行微调，以提高模型在新任务上的性能。大模型的分类主要基于输入数据类型和应用领域。按照输入数据类型，可以分为语言大模型、视觉大模型和多模态大模型；而根据应用领域的

不同，大模型又可分为通用大模型L0、行业大模型L1和垂直大模型L2。

主流的大模型的网络架构一直沿用自然语言处理（NLP）领域最热门最有效的架构——Transformer结构。相比于传统的循环神经网络（RNN）和长短时记忆网络（LSTM），Transformer具有独特的注意力机制（Attention），这相当于给模型加强理解力，对更重要的词能给予更多关注，同时该机制具有更好的并行性和扩展性，能够处理更长的序列，成为NLP领域具有奠基性能力的模型，在各类文本相关的序列任务中取得不错的效果（图1）。

根据网络架构的变形，主流的框架可分为Encoder-Decoder、Encoder-Only和Decoder-Only，具体如下：

（1）Encoder-Only

仅包含编码器部分，主要适用于不需要生成序列的任务，只需要对输入进行编码和处理的单向任务场景，如文本分类、情感分析等。这类代表是BERT相关的模型，例如BERT（Bidirectional Encoder Representations from Transformers），RoBERTa（Robustly Optimized BERT），ALBERT（A Lite BERT）等。

（2）Encoder-Decoder

既包含编码器也包含解码器，通常用于序

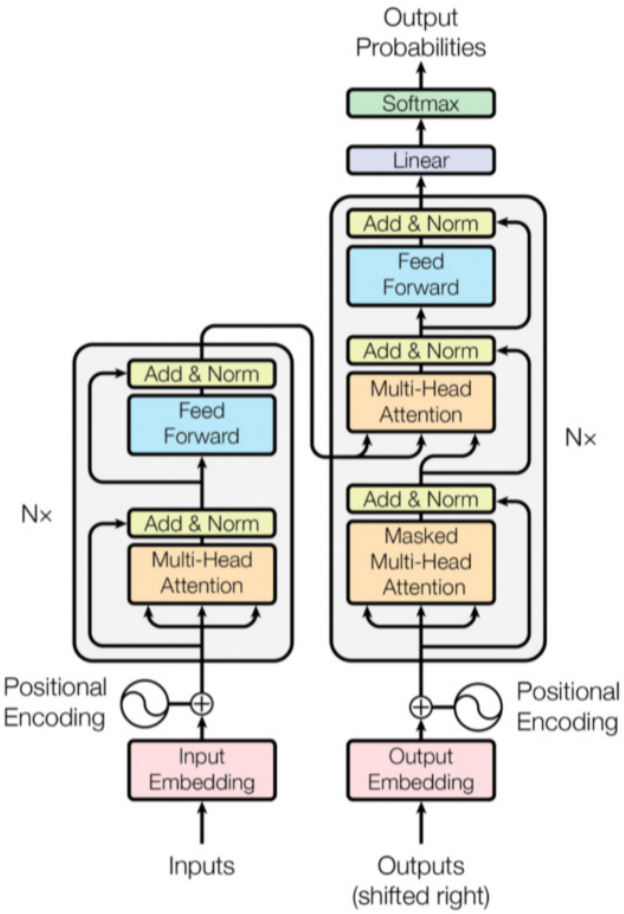

图 1：Transformer 模型架构图

列到序列（Seq2Seq）任务，如机器翻译、对话生成等。这类代表是以Google训练出来T5为代表相关大模型。

（3）Decoder-Only

仅包含解码器部分，通常用于序列生成任务，如文本生成、机器翻译等。这类结构的模型适用于需要生成序列的任务，可以从输入的编码中生成相应的序列。同时还有一个重要特点是可以进行无监督预训练。在预训练阶段，模型通过大量的无标注数据学习语言的统计模式和语义信息。这种方法可以使得模型具备广泛的语言知识和理解能力。在预训练之后，模型可以进行有监督微调，用于特定的下游任务（如机器翻译、文本生成等）。这类结构以GPT模型为代表，所有该家族的网络结构都是基于Decoder-Only的形式来逐步演化。

2. 大模型的微调

常用的大模型微调技术包括Fine-tuning、P-Tuning和LoRA（Low-Rank Adaptation），具体如下：

（1）Fine-tuning

Fine-tuning是通过对预训练模型在特定任务的数据集上进行进一步训练，以调整模型参数来适应新任务的过程。这通常涉及在预训练模型的顶部添加一个分类器层或任务特定层，并在目标任务的训练数据上对整个模型（或部分模型）进行微调。Fine-tuning能够充分利用大规模预训练模型学习到的通用知识，并通过特定任务的数据来优化模型性能。特点一是全面调整。Fine-tuning会调整模型中的所有参数，以适应新任务的数据分布。二是高效迁移。Fine-tuning能够将预训练模型在相似任务上学到的知识迁移到新任务上，加速模型收敛。三是易过拟合。如果新任务的数据量较少，Fine-tuning可能会导致模型过拟合。

（2）P-Tuning

P-Tuning的核心思想是在特定位置插入可训练的Token，从而让模型更好地理解下游任务的需求。这种方法特别适用于需要捕捉复杂上下文信息的任务，如生成式任务和问答系统等。P-Tuning通过在输入序列中插入额外的Prompt Token，形成新的输入序列，并通过学习这些Token来捕捉任务相关信息。

P-Tuning技术有两种版本，P-Tuning v1和P-Tuning v2。其中P-Tuning v1是对Prompt部分进行进一步编码计算，使用LSTM或多层感知机（MLP）对Prompt Embedding进行一层处理，以加速收敛。而P-Tuning v2则是对P-Tuning v1的改进，它在每一层Transformer中都加入了Prompt Tokens作为输入，从而使得深层结构中的Prompt对模型预测产生更直接的影响。

P-Tuning v2相比于P-Tuning v1，移除了重参数化的编码器，并且引入了多任务学习，通过在多任务的Prompt上进行预训练，然后再适配下游任务，以提高模型的泛化能力。此外，P-Tuning v2回归了传统的分类标签范式，采用随机初始化的分类头应用于Tokens之上，增强了通用性。

P-Tuning的特点一是参数高效。P-Tuning只调整少量的提示参数，而保持预训练模型的参数不变，从而大大减少了微调所需的参数数量。二是适应性强：由于提示词可针对特定任务进行优化，因此P-Tuning在少样本或零样本学习场景下表现出色。三是表达丰富。P-Tuning v2进一步改进了P-Tuning v1，使用

连续的提示词代替离散的Token，使提示的表示更加灵活和丰富。

（3）LoRA

LoRA是一种高效的大模型微调技术，其核心原理是在预训练的大模型中引入低秩矩阵来近似权重的更新，从而实现参数的高效更新。与传统的全参数微调相比，LoRA通过仅更新模型中一小部分参数来减少计算和存储成本，同时保持了模型在特定任务上的性能。

LoRA的特点一是低秩调整。LoRA基于预训练模型内在的低秩特性，通过调整一个低秩矩阵来模拟全参数微调的效果，从而减少对预训练模型的破坏。二是计算成本低。由于只微调了额外的线性层，LoRA的计算成本远低于全参数微调。三是易集成。LoRA已经被集成在多个深度学习框架中，如HuggingFace的PEFT代码库，方便开发者使用。

三、技术攻关方向

在金融行业内，除了头部银行和保险机构外，关于大模型技术的资源配置与人才储备普遍显得较为薄弱，鲜有金融机构能够独立承担并成功实施大模型的全面训练任务。鉴于此，金融机构纷纷调整技术策略，将技术攻关的焦点聚焦于大模型提示词工程、知识工程构建与优化、大模型微调训练技术以及运营管理平台的建设与完善。

（一）提示词工程

大模型凭借人类输入的提示词能够生成目标内容。然而，这一过程的效果显著依赖于提示词的质量。提示词工程作为一种优化策略，其框架的构建对于显著提升大模型的性能与能力至关重要，它不仅是提升模型响应准确性的关键，也是促进模型在多样化应用场景中表现优越性的基础。一般情况下，提示词框架包括以下几个方面的内容：

1.角色定义

在训练阶段，大模型使用的数据种类非常丰富，但质量参差不齐。默认情况下，其生成高质量数据和低质量数据的概率相近。当给大模型指定角色后，生成的概率会更偏向定义的高质量解决方案上，即让大模型对自身定位更加清晰。

2.目标

设置大模型生成的预期结果。

3.生成限制

给大模型加上约束，能够使大模型的生成更加合理接近实际目标，避免大模型生成如凭空捏造、不符合输出预期要求/格式等情况。

4.已知内容

大模型针对舆情、研报、政策等知识进行提炼解读，则可以将这些舆情、研报、政策等内容当作已知内容供大模型理解。

5.指令

人工输入的指令信息，以便纠正或补充信息，让大模型更好地生成结果。

6.输出规则

指定大模型生成的最终要求，比如生成markdown格式、json格式、python代码格式以及具体包含的内容等。

7.输出样例

样例涉及两个概念：zero-shot和few-shot。简单来说就是没有指定特定任务的训练样本和提供少量特定任务的训练样本，以帮助模型更好地理解任务。

当同一组提示词在应用于不同大模型时，所激发的能力提升效果呈现出差异性。这一现

象揭示了不同大模型架构对于提示词信息的敏感度与解析能力的差异。此外，当任务细节在提示词中被更加详尽且清晰地描述时，大模型生成的内容质量会显著提升，表现为更高的准确性、相关性和逻辑性。高质量的提示词不仅应当准确传达任务的核心意图，还需包含足够的上下文信息和具体细节，以引导模型生成更为精准和符合期望的输出。

因此，为了最大化大模型的效能，大模型在应用时需将更多精力投入提示词工程的优化上，包括设计更加精准、全面且富有引导性的提示词，以及探索不同模型对提示词反应的敏感性和偏好，从而制定出更加适配特定任务和模型的提示词策略。

（二）知识工程

在使用大模型时，知识工程作为预先集成的已知内容库，能够显著增强大模型回答问题的准确性、全面性和专业性，扮演着至关重要的角色。然而，知识数据的质量参差不齐，存在诸如表述口语化、上下文连贯性缺失、标点符号使用不规范等问题。这些问题若未经处理直接纳入知识库，将不可避免地削弱大模型的输出效能与可靠性，从而影响用户体验。

例如，图2所示的语料如果直接作为知识库的内容输入，当向大模型提问“第一章第二条的具体内容”时，由于语料未经结构化处理且缺乏明确的上下文标识，大模型存在极大的误答风险，这种局限性凸显了知识工程在提升模型性能方面的关键作用。

上海期货交易所交易规则

第一章 总则

第一条 为了规范期货交易行为，保护期货交易各方的合法权益和社会公共利益，根据国家有关法律、法规、规章和《上海期货交易所章程》，制定本规则。

第二条 上海期货交易所（以下简称交易所）根据公开、公平、公正和诚实信用的原则，组织经中国证券监督管理委员会（以下简称中国证监会）批准的期货交易。

图 2：原始语料示例

通过应用知识工程的方法，可以对原始语料进行深入地梳理与重构。这一过程涉及对信息的结构化提取、逻辑关系的明确界定以及必要时的语义增强。通过这一精细化的处理，能够构建出一个更为精确、易于模型理解的知识库。当再次面对相同的查询时，经过知识工程优化后的知识库能够显著提升大模型的响应准确率，有效避免误答现象，从而增强大模型在特定知识领域内的应用效能。

具体而言，图2的语料可大致切分为：文章名|章节|条目|内容。

（1）上海期货交易所交易规则|第一章 总则|第一条|为了规范期货交易行为，保护期货交易各方的合法权益和社会公共利益，根据国家有关法律、法规、规章和《上海期货交易所章程》，制定本规则。

（2）上海期货交易所交易规则|第一章总则|第二条|上海期货交易所（以下简称交易所）根据公开、公平、公正和诚实信用的原则，组织经中国证券监督管理委员会（以下简称中国证监会）批准的期货交易。

知识工程并非一项短期或一次性任务，而是具有长远影响的战略性工作。鉴于当前大模型技术的迅猛发展态势，高质量的知识工程成果不仅能够即时提升现有模型的推理与理解能力，更能在未来技术迭代与模型更替中保持其价值与效用，实现“一次投入，长期受益”的效果。即便面对不同架构或算法的大模型更新，优质的知识基础依然能够确保模型性能的持续优化与稳定输出。构建并持续优化知识工程体系对于推动大模型技术的发展与应用具有不可估量的价值。它不仅是提升大模型性能的关键路径，也是确保大模型在复杂多变的应用场景中保持高效、准确与专业的基石。因此，加大对知识工程领域的投入与研究力度，是当前及未来大模型技术发展中不可或缺的一环。

（三）微调训练

微调训练作为增强大模型在特定垂直领域适应性的有效手段，在特定应用场景中展现出不可或缺的价值。然而，这一过程往往伴随着高昂的实施成本，包括对数据集的大量需求、对高性能计算资源的依赖，以及显著的人力、时间和经济投入。

鉴于当前大模型技术日新月异的发展态势，新一代大模型在综合能力上常常显著超越其前身，甚至超越了经过微调的前代大模型。这一现象直接导致了微调训练相对于大模型自然迭代的效益比降低，即在资源投入与性能提升之间出现了不成比例的情况。

进一步地，鉴于当前硬件资源的稀缺性与人力资源的紧张状态，除非是对于具备“专业”“精准”“特定”或“创新”等特点的场景，否则从成本效益分析的角度来看，大规模地实施模型微调训练并不具备高性价比。因此，在资源有限的情况下，合理评估并优化微调训练的应用场景，以最大化资源利用效率，成了研究与实践的重要课题。

（四）运管平台

大模型技术正以前所未有的态势引领着技术革新的浪潮。在此过程中，各类大模型凭借其在特定应用领域的独特优势，展现出多元化的应用场景与价值。然而，这一技术繁荣的背后，也伴随着实际应用中的一大挑战：频繁地在不同模型间进行切换，不仅过程繁琐，还显著降低了工作效率与用户体验。

鉴于此，构建一个统一、全面的大模型运营管理平台显得尤为重要且迫切。通过优化模型切换机制，简化操作流程，从而实现对大模型的高效管理与协调。具体而言，该平台需具备以下核心能力：

1.统一接口与标准化管理

通过定义统一的API接口和标准化管理规范，消除不同模型间的兼容性问题，确保用户可以无缝切换并使用各类大模型。

2.智能调度与优化配置

利用先进的调度算法和资源配置策略，根据实际应用需求智能选择最优模型，并动态调整资源分配，以达到性能与成本的最佳平衡。

3.可视化监控与性能分析

提供直观的可视化界面，实时展示模型运行状态、性能指标及资源消耗情况，帮助用户快速定位问题并进行优化调整。

4.灵活扩展与定制开发

支持快速集成新模型，同时提供灵活的定制开发接口，满足用户多样化的业务需求，推动技术创新与应用深化。

构建一个全面、高效的大模型运营管理平台，不仅能够显著提升大模型在实际应用中的便捷性与效率，还将进一步推动该领域技术的普及与发展，为金融行业的数字化转型提供强有力的技术支撑。

四、应用探索实践

相较于传统判别式AI在金融领域的广泛应用，大模型具备更强的通用泛化能力，能够高效应对复杂多变的信息理解、内容生成及多轮对话等任务需求，在金融领域内存在巨大的价值创造潜力。具体而言，在金融机构层面，根据英伟达发布的针对近400家金融机构的调研，43%的金融机构已经在使用大模型，另有55%金融机构正在研究并考虑应用大模型；在金融从业者层面，根据麦肯锡2023年调研数据，金融行业从业者反馈“在工作中使用大模型”、“在生活中使用大模型”和“在工作和生活中均使用大模型”的数量占比达42%，而这一比例在麦肯锡2024年调研数据中上升至48%。由此可见，国内外各类金融机构正在积极探索大模型落地场景，金融从业者对大模型工具的需求和使用也与日俱增。

本节将通过梳理目前证券行业、银行行业在大模型应用的相关实践案例，分析大模型在提升业务效率，增强风险管理能力方面的基本情况。此外，本节还将重点分享上海期货交易所（以下简称上期所）在运用大模型方面的探索实践情况。

（一）证券行业

在证券行业中，投资研究与投资顾问作为两大核心业务领域，其效率与质量的提升对于证券公司的整体竞争力具有至关重要的作用。大模型的引入，为这两个领域带来了革命性的变革，有效应对了信息过载、个性化服务不足、风险管理复杂及合规性监控挑战等长期存在的业务痛点。本节将从生产力提升与降本增效、投顾场景下的应用实践、投研场景下的应用实践三个方面进行探索研究。

1. 生产力提升与降本增效

在短期内，证券行业的大模型应用聚焦于投研与投顾场景中的日常事务处理，通过自动化与智能化手段，可以显著提升工作效率并降低成本。具体而言，大模型能够替代重复性高的劳动任务，如数据整理、初步分析等，使投研人员与投顾人员得以将更多精力集中于高价值的策略制定与个性化服务上。多家头部证券公司，如中信建投、申万宏源、国泰君安及海通证券，已率先探索大模型在各自业务中的应用潜力，旨在实现降本增效与用户体验优化。

2. 投研场景下的应用实践

（1）中信建投智能投研平台

中信建投利用大模型构建了智能投研平台，通过整合数百万条投研报告与金融资讯，以及大量音视频数据进行深度训练，实现了投研工作的智能化升级。该平台不仅提升了投研效率，还促进了投资知识的普惠传播与资产管理能力的提升。

（2）申万宏源智能研报降维服务

针对研报内容复杂、信息提取成本高的痛点，申万宏源采用传统算法与大模型相结合的方式，实现了研报的自动化智能降维解读。该

服务已在关键事件管理系统、公司详情、行业详情、经济解读、债券解读、配置策略、热门板块等多个业务场景中上线，显著提高了研报信息的利用效率与价值发挥。

3. 投顾场景下的探索与创新

（1）国泰君安君弘灵犀大模型

国泰君安推出的君弘灵犀大模型，凭借其千亿参数级多模态能力，构建了涵盖多模态处理、智能选股、综合诊断、热点资讯等十大功能的智能服务体系。该模型不仅解决了传统模型在语义理解、多轮对话等方面的不足，还以全新的交互模式为海量用户提供高质量的智能化服务。

（2）中信建投分阶段构建投顾大模型

中信建投分阶段构建投顾大模型，从智能运营助手到投顾专家助手，再到智能专家投顾，逐步实现了从被动响应到主动服务的转变，显著提升了投顾服务的个性化与专业化水平。

（3）西南证券投顾AI创作台系统

该系统利用大模型处理每日海量舆情与资讯，通过智能整合与定制化生成，为投资顾问提供精选资讯内容，有效减轻了投顾人员的信息筛选负担，使其能更专注于为客户提供精准的投资建议。

（二）银行行业

截至2024年9月底，42家A股上市银行2024年中期报告已披露完毕。在2024年中期报告中，19家上市银行提到了在大模型技术应用方面的最新进展，展现了金融行业在人工智能领域的深度探索与实践。本节将从大模型技术应用在银行行业的分布特征、应用进展、业务价值三个方面进行深入分析。

1. 分布特征

国有银行引领潮流。在6家上市国有银行中，有5家（中国工商银行、中国邮储银行、中国农业银行、中国建设银行、中国银行）明确提及了大模型技术的应用，彰显了其在金融科技领域的领先地位和前瞻布局。

股份制银行紧随其后。9家上市股份制银行中，8家（招商银行、平安银行、中信银行、兴业银行、浦发银行、中国民生银行、中国光大银行、浙商银行）亦积极投身于大模型技术的探索与应用，体现了其在金融创新方面的活跃度和执行力。

城商行与农商行差异显著。相比之下，17家上市城商行中仅有6家（北京银行、江苏银行、上海银行、南京银行、长沙银行、重庆银行）提及大模型应用，而10家上市农商行则无一涉及[1]，这反映出不同规模银行在技术应用能力和战略选择上的差异。

2. 应用进展

以中国工商银行、平安银行为代表的头部银行，已在大模型技术的多场景应用中展现出显著的引领作用。中国工商银行不仅深化了千亿级大模型技术的建设与赋能，还率先实现了企业级金融大模型的全栈自主可控训练和推理部署，覆盖了金融市场、信贷风控、网络金融等多个领域。平安银行则通过自主研发的大模型开放平台，为营销支持、内部运营、风险管控、办公辅助等多个业务领域提供了通用能力模型和一站式场景定制服务，进一步提升了业务效率和服务质量。

其他银行在大模型技术的应用上则显得更

[1] 数据来源：沙丘智库。

为谨慎和保守。中国银行等部分银行仅在年报中披露了部分试点场景，如代码辅助等，而多数中小银行，特别是农商行，则仍处于观望阶段。

3. 业务价值

根据沙丘智库发布的报告，当前银行大模型技术落地的Top5场景分别为员工办公助手、编码助手、智能客服、知识助手和信贷审批[2]。这些应用场景的广泛覆盖，不仅提升了银行内部运营的效率和质量，也为客户提供了更加便捷和个性化的服务体验。例如在辅助员工提高作业效率的场景中，大模型技术已展现出明显的业务价值。又例如中信银行通过结合AI大模型迭代智能机器人，实现了客户服务效能的显著提升；兴业银行则利用大模型技术打造研报摘要助手，有效减轻了人工负担并提高了信息处理效率。

（三）上期所

近期，上期所作为国家重要金融基础设施，分别在基于文本数据的画像标签抽取及报告辅助生成等典型应用场景进行了初步探索实践，用以验证大模型的知识抽取能力及知识信息生成能力。画像标签抽取及基于画像的分析报告辅助生成两个典型应用场景可为后续知识库问答、代码辅助生成等类似应用场景提供实践借鉴。

1. 画像标签抽取

（1）问题与挑战

现有画像体系通常依赖预先定义结构和内容的知识图谱，通过结构化表示对数据进行组织与管理。这一过程首先需要将原始的文本数据通过一系列的预处理转化为结构化数据，包括但不限于文本的分词、实体识别、关系抽取等，以确保数据的可读性和可操作性。随后，基于结构化数据生成标签，实现对个体或实体的特征描述和分类。然而，这一传统的流程存在一定的局限性，主要体现在对文本数据的直接利用程度较低，难以挖掘其中的潜在意图、观点信息，且内容和结构受知识图谱定义的限制等问题。

大模型技术的飞速发展为文本数据的分析处理提供了强大的助力，提升了从文本数据中挖掘潜在意图、观点信息的能力，从而丰富画像标签维度。然而，在处理长文本时，大模型面临着资源消耗升高与上下文理解能力下降的双重挑战。一方面，庞大的计算资源需求可能限制了模型的实用性与效率；另一方面，长文本中的信息密度高，上下文关联复杂，模型在理解与处理时可能会出现断链或误解的情况，影响最终的输出质量。

传统的策略是将长文本分割为多个片段，这些片段可以作为知识库的一部分，用于后续的查询与信息检索。通过这种方式，能够针对特定问题召回与之最相关的一组文本片段，从而为大模型提供局部的、针对性的信息支持。然而，这种处理方式不可避免地会导致信息的损失。一方面文本片段的分割可能破坏了原始文本中原本连续且重要的上下文关系，另一方面受召回数量的限制，大模型获取的有效信息不完整。

对于画像体系的构建而言，全面、完整的信息是至关重要的，缺失信息的处理方法显然不符合构建画像体系的需求，因为它无法提供足够丰富、连贯的信息，从而限制了画像的准

[2] 数据来源：《2024 年金融业生成式 AI 技术应用跟踪报告》。

确性和价值。鉴于以上问题，上期所在实践中提出了一种基于大模型的叶子到根部的画像标签抽取方法。

（2）画像标签抽取方法

画像标签抽取方法的步骤包括（图3）：

第一步：获取文本数据。

第二步：根据文本数据的标题，基于规则、关键词匹配对文本数据进行筛选、分类，完成初步处理。

第三步：利用BERT模型对处理后的文本进行二次筛选，保留和画像实体相关的文本数据。

第四步：根据文本格式，如\n等，对文本数据进行切片，获得多个文本片段。如文本切片字符数少于50，则计算当前文本片段与上一文本片段的向量相似度r1，当前文本片段与下一文本片段的向量相似度r2，若r1>r2，则将当前文本片段与上一文本片段合并，否则，与下一文本片段合并。

第五步：利用大模型总结归纳每一个文本片段的内容，并根据每一个总结归纳的文本片段生成一个或多个推荐的画像标签。

第六步：利用大模型将生成的推荐标签进行聚类，总结归纳，并生成上一级画像标签。

第七步：重复第五步和第六步，直到画像标签无法再聚类。

第八步：将由大模型自下而上生成的画像标签进行结构化处理，利用知识图谱，实现画像展示。

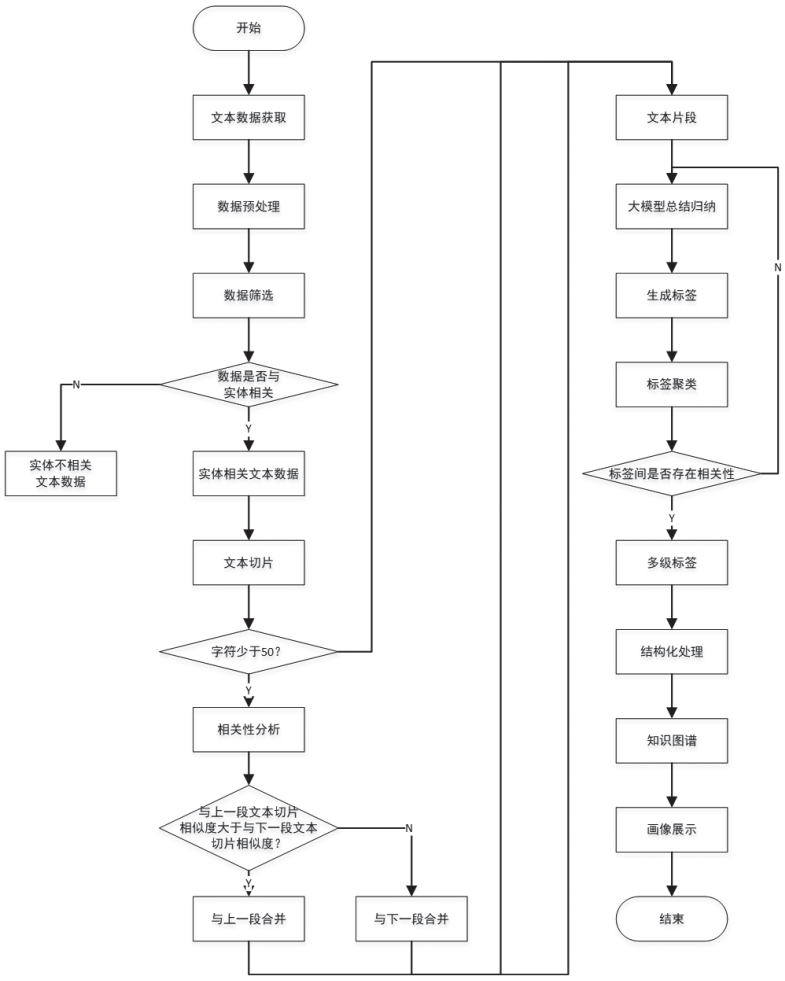

图 3：画像标签抽取方法流程

（3）结果展示

1）基于舆情数据的画像标签抽取（图4）

利用本地部署的大模型对当日原油舆情数据进行分析，抽取画像标签。画像标签包括原油期货行情概述、原因分析、影响和未来走势分析。

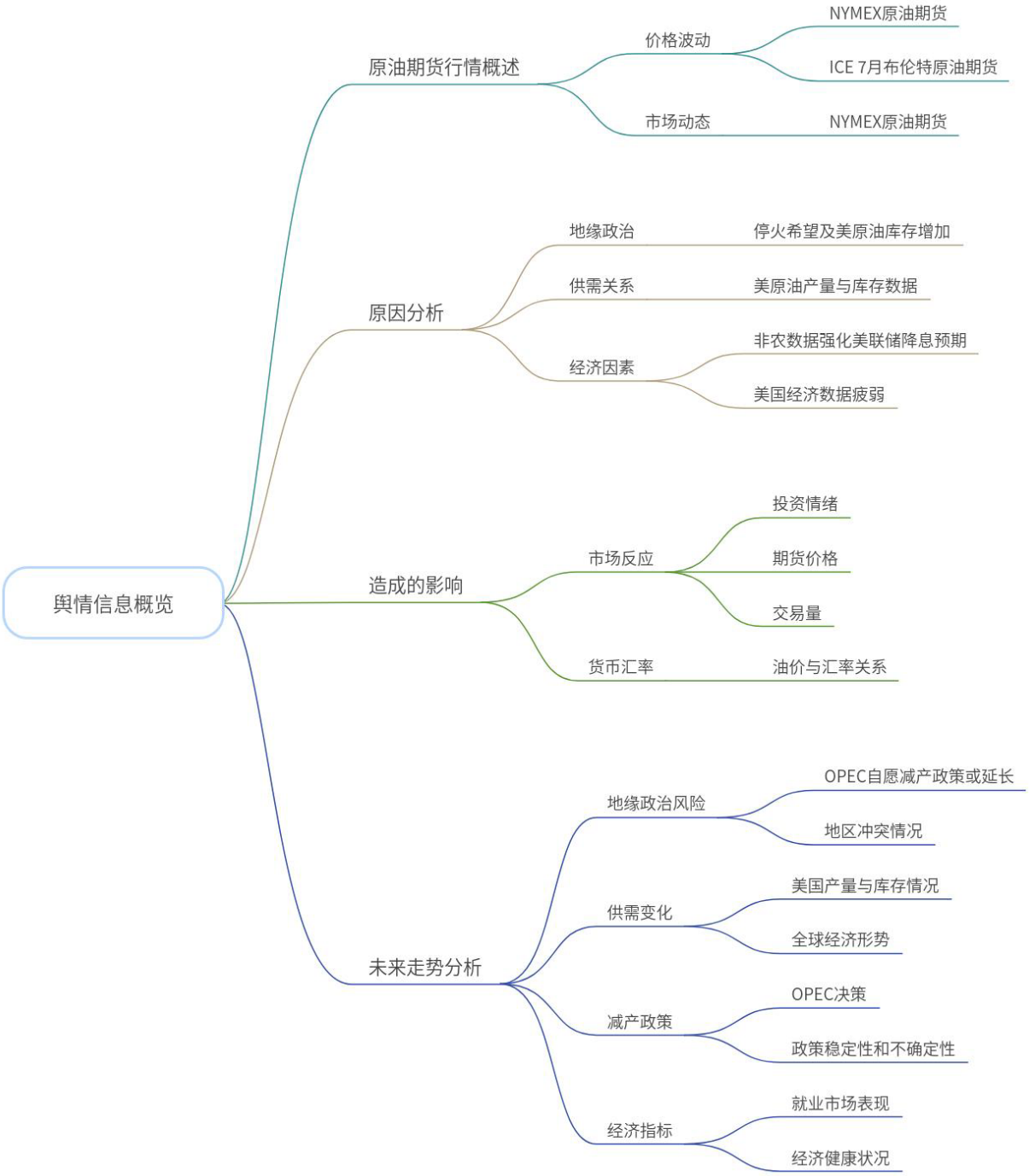

图 4：基于舆情数据的画像标签抽取

2）基于研报数据的画像标签抽取（图5）

利用本地部署的大模型对原油期货行业的研报数据进行分析，抽取画像标签。画像标签包括事件概述、事件原因、事件影响、各方观点、应对措施和未来走势预测。

图 5：基于研报数据的画像标签抽取

3）基于政策数据的画像标签抽取（图6）

利用本地部署的大模型对原油期货行业的政策数据进行分析，抽取画像标签，可实现对品种、合约、政策调整（手续费、保证金、涨跌幅限制）、调整时间等维度的标签抽取。

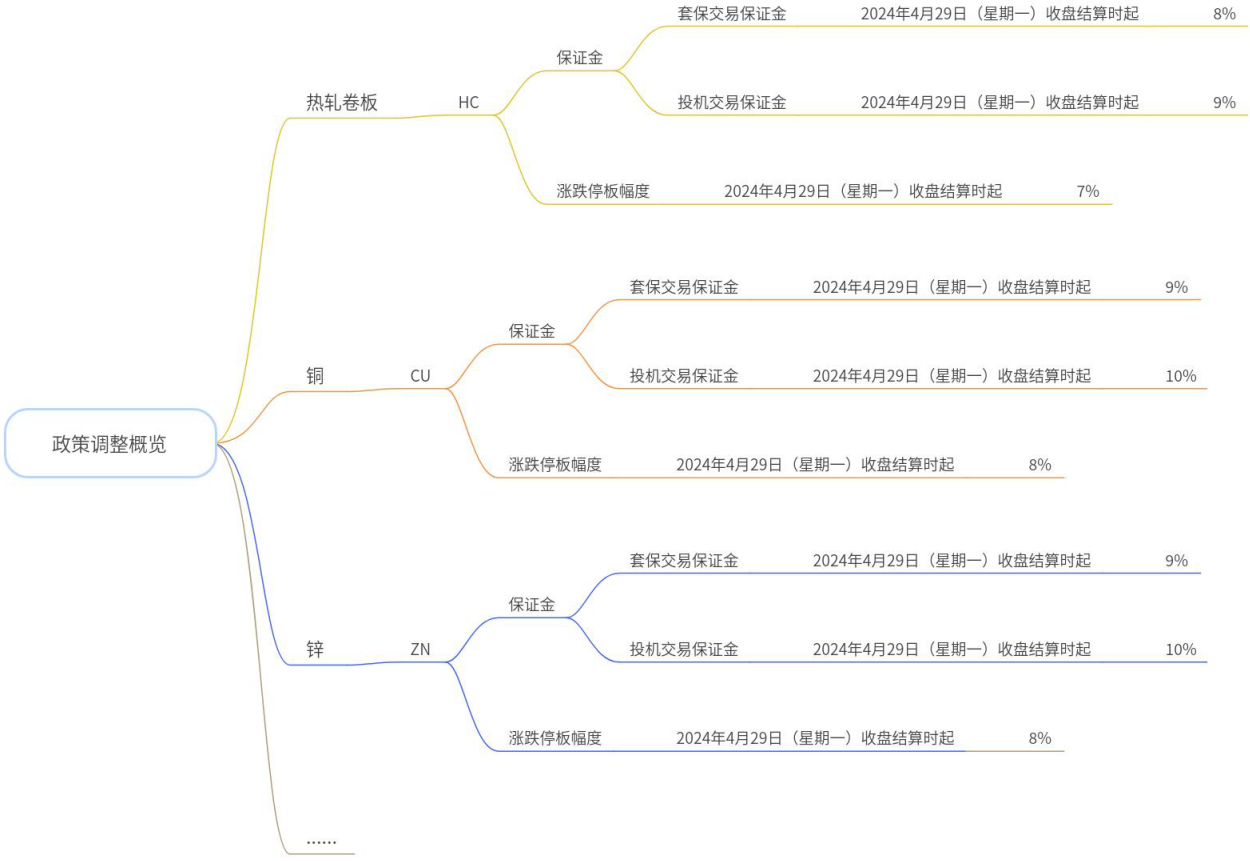

图 6：基于政策数据的画像标签抽取

2. 报告辅助生成

（1）问题与挑战

证券期货行业作为信息敏感型领域，因突发事件派生出的舆情事件时有发生，及时有效的舆情分析手段可以帮助从业者理解和预测事件的发展趋势，也为决策提供有力的支持。尤其在自媒体广泛普及的背景下，各类行业机构均对突发事件分析报告的自动生成技术有强烈需求。然而，传统的分析报告生成方式往往依赖于人工，这不仅需要耗费大量的人力开展信息采集、观点解析、内容生成等工作，而且可能受到个人主观因素的影响。因此，亟需寻找一种更高效、客观的分析报告生成方法。

大模型具有强大的语言理解能力和生成能力，可以在短时间内处理大量的文本数据，自动提取关键信息，生成结构清晰、逻辑严密的分析报告，为分析报告生成提供了全新的解决方案。例如在观点解析时，大模型可通过Prompt，将同类观点融合，不同类观点分类，并给出支撑观点的事实依据；在内容撰写时，可根据大纲内容，结合提供的知识库或采集的信息，自动编写相关内容；在摘要抽取时，可根据分析报告的内容，抽取关键信息，形成摘要。

（2）报告辅助生成方法

对于大模型而言，用户可以通过自然语言的方式对其进行提问以及提出任务要求。但直接让大模型生成一份文档报告，得到的生成反馈往往效果不佳。这是因为大模型的自由推理使得生成报告的整体结构及内容难以符合预期，而且大模型也存在惰性问题，让其一次性生成全部内容，得到的报告内容通常存在不具体、不全面的问题。

为了充分发挥大模型的能力以及写出一份较为完整可靠的报告，在使用一定规模上下文长度的大模型同时，可以将其输出Token最大值进行调整，并且使用不同的提示词模板构造提示链。流程图如下（图7）：

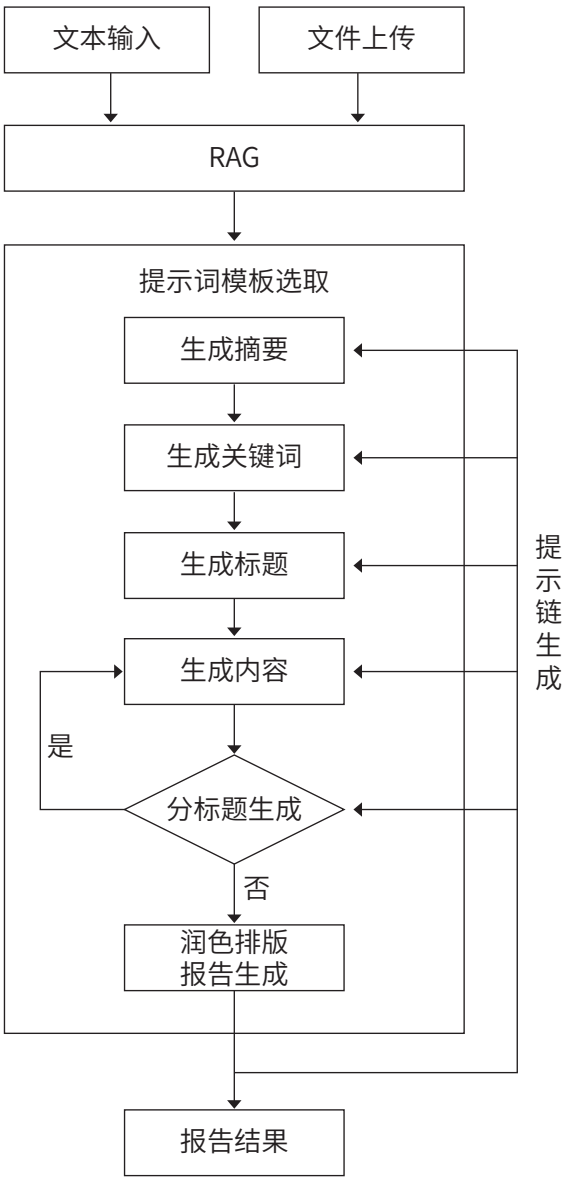

图 7：基于大模型的报告生成流程

（3）结果展示

1）极端行情分析报告

利用大模型收集舆情、研报数据，对行情进行概述，分析行情异动原因，实现极端行情分析报告辅助生成（图8）。

2）风险事件分析报告

利用大模型收集舆情数据，对风险事件进行概述，分析事件原因及影响，实现风险事件分析报告辅助生成。内容包括摘要、引言、事件概述、原因分析、影响、应对策略、未来走势分析、结论等方面（图9）。

(一) 单日涨、跌幅达到一定程度(3%)

1.4月18日原油期货价格大跌3.15%

（1）事件概述

4月18日，原油期货价格经历了显著的下跌。当日，上海国际能源交易中心的原油期货主力合约收盘时下跌了3.15%，具体而言，主力2406合约收报642.4元，较前一交易日下跌了20.9元。与此同时，纽约商品交易所的WTI原油期货价格也遭遇了类似的命运，5月交割的WTI原油期货价格下跌了2.67美元，收于每桶82.69美元，跌幅为3.13%。

（2）原因分析

地缘政治因素：尽管中东地区的地缘政治局势紧张，特别是伊朗和以色列之间的冲突，但市场普遍认为……

图 8：极端行情分析报告

关于商品期货量化交易被暂停传闻事件的分析

摘要：本报告基于大模型技术，对近期关于商品期货量化交易被暂停传闻事件进行分析。近期，关于商品期货市场量化交易被暂停的传闻在金融市场中引起广泛关注。中国证监会澄清此为不实消息，并表示各商品期货交易所将对手续费减收政策进行优化调整。尽管商品期货量化交易并未被暂停，但此次事件仍可能对期货公司和量化高频交易机构产生影响。监管部门可能会加强对期货市场的监管，以维护市场的公平和稳定。期货公司和量化交易机构需要密切关注政策动态，及时调整业务策略，以应对可能的挑战。

一、引言

近期，一则关于“商品期货市场量化交易被暂停”的传闻(以下简称传闻事件)在金融市场中引起了广泛关注。这一传闻迅速传播，引发了市场的恐慌和担忧。然而，随后中国证监会相关部门负责人对此进行了澄清，表示这一传闻为不实消息……

图 9：风险事件分析报告

3）舆情分析报告

利用大模型对每日舆情数据进行分析，对具体期货品种的行情进行概述，分析原因及影响，并展望未来走势，实现舆情分析报告辅助生成（图10）。

三、5月6日关键舆情分析

1.原油期货行情概述

国际油价回调：假期期间，国际油价出现大幅回调。外盘Brent原油期货主力合约跌破83美元/桶，WTI原油期货主力合约跌破78美元/桶。

国内原油期货下跌：节后国内原油期货主力合约2406大幅度跳空下跌超4%，报614.5元/桶，跌5.01%。

2.原因分析

地缘政治因素：

…巴以冲突出现缓和可能，国际油价回吐前期地缘溢价。

…中东地缘局势有所缓和，哈马斯对埃及提出的与以色列停火倡议作出回应，派代表团前往埃及首都开罗……

图 10：舆情分析报告

五、痛点问题

尽管金融行业已在大模型领域具备了一定的应用探索经验，但仍面临着不少痛点问题，如数据质量低、隐私数据易泄露等数据问题，硬件资源采购困难、利用率低、难以满足安全可控要求等算力问题，定制化需求复杂、小型化需求迫切等模型问题，以及模型构建和应用中的安全、合规等监管问题，在一定程度上限制了大模型在行业中的快速应用落地。本节将从数据、算力、模型以及安全监管四个方面进行分析。

（一）数据

大模型训练面临着数据规模和质量不足的问题。大模型的有效性依赖于海量且高质量的数据，而目前大模型的训练语料主要依赖互联网公开的内容，其中低质量的数据容易造成大模型泛化能力偏低、出现歧视偏见以及幻觉问题，难以满足金融行业对专业性、严谨性方面的要求。此外在数据隐私保护方面，金融行业的核心业务涉及大量私有数据，如内部文档等。大模型在训练和应用过程中需要处理这些数据，如何确保这些数据在处理过程中的隐私保护成为关键问题，目前仍缺乏安全的数据共享机制。

（二）算力

1. 硬件资源采购困难

金融行业在推进大模型应用时，首先面临的是硬件资源不足的问题。GPU和NPU作为大

模型训练与推理的核心硬件，其采购面临诸多限制。一方面，受限于供应链稳定性、国际政治经济环境等因素，GPU和NPU的供应可能出现波动；另一方面，金融行业对于高性能硬件的需求量大，但采购资金、审批流程复杂，导致采购周期较长，难以满足快速发展的需求。

2. 硬件资源利用率低

金融行业，特别是证券期货行业，其核心业务时间通常较为集中，导致硬件资源在非核心业务时间大量闲置。如何有效调度和管理这些硬件资源，提高其在非核心业务时间的利用率，成为亟待解决的问题。当前，缺乏高效的资源调度机制和优化算法，使得硬件资源的利用率难以达到理想状态。

3. 难以满足安全可控要求

当前，大模型的生态系统主要依赖于国外相关企业，如N卡产品，国产硬件和软件的适配需要投入大量的经济、人力和时间成本，且可能面临技术壁垒和兼容性问题。因此，如何在保障性能的同时，实现大模型生态的自主可控，成为金融行业面临的重要挑战。

（三）模型

1. 定制化需求复杂

金融行业对大模型的定制化需求极高，不同业务场景需要不同的大模型支持。然而，大模型的定制化开发涉及复杂的算法设计、数据标注和模型训练过程，需要投入大量的人力和时间成本。此外，不同金融机构之间的业务需求差异较大，如何实现大模型的快速定制和灵活部署，成为亟待解决的问题。

2. 小型化需求迫切

在金融行业中，尽管大模型展现出了强大的潜力和价值，但大多数金融机构面临着一个不容忽视的现实：其硬件资源相对有限，难以直接支撑起千亿级大模型的训练与推理需求。这种资源限制迫使金融行业对模型的小型化提出了迫切需求。小型化不仅意味着模型在参数规模上的精简，更要求模型在保持一定性能的前提下，能够更有效地利用有限的计算资源，从而适应更多金融机构的实际运行环境。

（四）安全监管

金融行业对安全、合规有着极高的要求，大模型的应用必须严格遵守相关法律法规和监管要求。然而，大模型在训练和应用过程中可能涉及数据泄露、算法歧视等安全风险，如何确保大模型的应用符合安全、合规要求，成为金融行业关注的焦点。此外，在大模型技术不断发展的过程中，监管政策也在不断调整和完善，如何及时跟进监管政策变化，确保大模型的应用始终符合安全、合规要求，成为金融行业面临的长期挑战。

六、总结与展望

（一）总结

1. 探索实践多样，“颠覆性”应用尚未出现

总体而言，大模型在金融行业的应用仍处于初期探索尝试阶段。目前，金融机构大部分将大模型应用于业务流程中较为基础的非决策性环节，但在涉及高度专业性、精准性的金融投资建议及核心分析决策等业务场景，仍鲜有成功案例。例如大模型在支付、信贷、保险、财富管理等多个金融细分领域主要集中在优化客户服务体验、提升数据挖掘及分析能力、辅助业务运营等方面；在需要高度专业判断、承担核心决策职责的情境中，如资本管理、风险管理和监管科技等，大模型尚难以达到专业人

员分析决策水平，更多扮演辅助角色。因此，当前大模型技术在金融行业尚未出现“颠覆性”的变革应用。

2. 痛点问题具有行业共性，落地难题亟待解决

金融机构普遍面临着数据管理与算力支撑等方面的痛点问题。

首先，大模型在金融领域的应用受限于数据质量，低质量数据导致模型泛化能力不足、偏见及幻觉问题频发。因此，金融机构亟需构建一套强大的数据管理体系，涵盖从数据收集、存储、清洗到标注的全链条，确保数据的准确性、完整性、合规性和安全性。这要求不仅要有高效的数据采集与存储平台，还需推动数据标准化，促进多源数据在模型训练中的无缝融合。

其次，随着模型复杂度提升，算力需求激增，尤其对于需实时处理海量数据的金融机构而言，高性能算力基础设施的建设迫在眉睫。鉴于中小机构难以独立承担高昂成本，政府及大型金融机构可携手合作，共建随需调度的算力资源池，以降低行业整体使用成本。

最后，鉴于基础设施建设的高投入特性和长尾特征的使用场景，金融机构应积极寻求跨行业合作，与科技、能源、政府等多方力量联合，共同推进数据与算力共享平台的建设，并探索分布式能源网络等创新方式，以降低成本、提升效率，共同推动金融行业基础设施的全面发展。

（二）展望思考

1. 强化数据治理，提升数据质量

数据既是金融行业的核心资产，也是决定大模型应用有效性的前提。相对于其他行业而言，金融行业的数据质量虽然相对较高，但相较于国外相关领域的数据质量仍有不小差距。金融机构应将数据治理作为一项长期性、基础性的工作，通过系统化的数据清洗、管理和加工，提升数据质量，为大模型的应用提供坚实的数据基础。同时，金融机构应主动承担起数据治理的责任，虽然可以借助第三方专业机构的力量，但核心工作仍需亲力亲为，确保数据安全与合规。

2. 构建行业生态圈，促进大模型资源共享

大模型的开发与维护需要巨大的资金投入和专业人才支持，这对于中小金融机构而言是一大挑战。因此，金融机构应摒弃“单打独斗”的思维模式，积极寻求多方合作，共同推动大模型的建设与应用。通过构建行业生态圈，实现资源共享、优势互补和成本节约的合作模式。同时，金融机构在与第三方机构合作过程中应注意风险防控，确保合作关系的稳定性和可持续性。

3. 重视垂直领域大模型在金融行业的应用

当前大多数金融机构尚不具备部署并维护千亿级通用大模型的能力。鉴于这一现状，金融机构应转而聚焦于垂直领域大模型。因为这类大模型具有专业性强、参数量小、资源需求不高的特点，不仅满足金融机构的业务需求，还可实现成本效益的最优化。

4. 坚持审慎监管原则，平衡创新与风险

金融行业的创新发展离不开监管的支持与引导。在推动大模型等新技术应用的过程中，给予市场充分的探索空间和时间的同时，监管部门应密切关注新技术应用可能带来的风险和挑战，如消费者权益保护、道德伦理冲突等，并根据实践情况适时调整监管政策。在制定监

管政策时，监管部门应采取包容审慎的态度，既要保障市场的健康发展又要鼓励创新。通过平衡创新与风险的关系，推动金融行业与大模型技术的深度融合与发展。

5. 坚持需求导向，注重大模型的实用性

对于中小金融机构而言，大模型的应用应紧密围绕市场需求，确保技术的实用性和商业价值。只有当大模型能够切实帮助金融机构降本增效、提升风险管理水平并优化客户体验时，其应用才具备持久的商业可持续性。因此，金融机构在引入大模型时，应充分考虑市场需求，注重实际效果评估，确保技术投资与业务发展的紧密结合。

（责任编辑：支文纲）

作者简介：

范宏婷、薛利、赵博，均任职于上海期货信息技术有限公司，研究方向为大数据及人工智能。