上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

蚂蚁集团 ANT GROUP

BCTC 国家金融科技测评中心 National FinTech Evaluation Center 银行卡检测中心 Bank Card Test Center

大模型在金融领域的应用技术与安全白皮书

摘要

大模型技术带来了AI的新一轮技术变革和产业应用。构建大模型在金融领域完善的开发框架和应用框架，可助力现有金融业务进行数字化转型。但其应用也面临着诸多风险，需要进行进一步防控。除了针对通用的大模型幻觉风险的防护围栏，还需要针对金融领域的应用进行隐私风险防控、大模型攻击防御、可解释性增强、可溯源性增强以及有害内容防控，从而更好的助力传统金融业务。除此之外，金融领域大模型治理框架的搭建、评测集的构建和人才体系的培养则有利于促进大模型在金融领域的生态体系构建。

基于此，本白皮书主要围绕大模型在金融领域的应用技术及安全防控研究，延伸至大模型在金融领域的评测框架及人才培养体系进行分析。应用技术方面，主要基于大模型的开发框架和应用框架及应用实践进行了探讨。在应用风险防控方面，主要聚焦在大模型金融领域的风险及安全防控手段，同时借鉴了国外的人工智能应用风险治理框架；而大模型评测主要聚焦在大模型的评测框架以及大模型在金融领域的评测；最后大模型人才培养体系构建则强调了人才需求、人才教育体系、跨界合作及人才评估认证。

本篇白皮书为本系列的第一本，主要围绕大模型的应用技术及风险防控技术进行撰写，分为五个章节。第一章节主要为大模型的概述，第二章节主要聚焦于大模型的技术分析，第三章节主要聚焦于大模型的风险与防控，第四章节给出了大模型的评测方式，第五章节则衍生到大模型发展中的人才培养。

目录

概述 01

1.1 大语言模型技术发展概述

语言建模（Language Model）可分为四个发展阶段，分别为统计语言模型、神经语言模型、预训练语言模型、大模型语言模型。

其中最早的统计语言模型基于统计学习来预测单词，而后演进成为神经语言模型基于神经网络方法预测单词。在神经网络语言模型中，通过使用神经网络，将单词映射为向量作为网络模型的输入来估计单词序列的概率。随着注意力机制被引入，注意力层（Attention Layers）在文本中建立了词之间的相关性，使得模型在生成下一个单词时，考虑到整体语句的意思，从而建立了 Transformer 架构，提升了模型理解和生成语言的能力。

但随着参数的增加，需要大量人力来标注数据，因此 OpenAI 提出了预训练语言模型（Generative Pre-Trained Transformer），通过无监督学习在大规模无标签语料库上进行预训练任务，在预训练中模型学会了基于前一个单词预测后一个单词。除此之外，模型还可以针对特定的任务基于更小的数据集进行微调，提升在特定领域的性能。基于此，通过不断叠加数据增加模型参数规模以及优化模型的提示工程，不仅可以解决更复杂的任务，同时也拥有了更强大的文本涌现能力[1]，从而演进成为大模型语言模型（以下简称“大模型”）。

大模型浪潮爆发后，国内各企业纷纷推出自研大模型，大模型应用迎来了蓬勃发展的阶段。据测算，我国 2030 年基于大模型的生成式人工智能市场规模有望突破千亿元人民币。

与此同时，国内垂直行业领域的大模型也成为各个行业头部企业未来的发展趋势之一，其中前沿的垂类大模型涉及领域包括媒体影视、电商、广告营销、游戏、医疗、教育

[1] Zhao et al,《A Survey of Large Language Models》

及金融行业。比如在金融领域，大型科技企业如华为推出了盘古金融大模型，而蚂蚁集团则在外滩大会发布了金融大模型“AntFinGLM”并应用于蚂蚁集团内部产品“支小宝”和“支小助”。

金融行业大模型在所有行业垂直大模型中落地速度相对较快。金融领域拥有天然的大量数据积淀，从而为大模型应用提供了良好的数据基础。同时金融领域大模型的应用场景较多，基于这些不同的场景，大模型有助于从不同角度提升原有从业人员及机构的工作效率。比如大模型情绪分析的功能可帮助从业者基于投资者情绪状态预测股票的价格；大模型精确度的提升可帮助从业者预测市场走势，大模型可基于过去大量的金融数据学习预测未来市场趋势帮助投资者和金融机构做出更合理的决策；而复杂任务的处理可协助从业者将大模型用于交易策略上，通过分析大量交易信息，大模型或可识别交易中的风险参数并给出风险防控策略。

1.2 大模型引领中国金融领域科技的国际化发展

因此，通过提升金融服务的效率和质量，大模型可提升我国金融机构的核心竞争力。首先大模型的自然语言理解与内容生成能力可以与用户进行多轮问答对话，提升金融客服的服务效率。其次，通过大模型进行智能数据挖掘处理，金融机构能够更快速准确地获取市场趋势的洞察，做出更明智的决策。同时，大模型可以迅速了解各国的法律、监管规定和市场动态，为金融机构提供国际化的业务洞察和决策支持，帮助中国从业者更好地理解和适应国际市场的业务需求和规则。

海外金融科技公司已经在积极探索和持续深化大模型在金融服务领域的应用。Bloomberg 已推出 BloombergGPT，一个基于 500 亿参数训练的应用于金融领域自然语言处理的大模型。据研究，当前此大模型在金融任务包括金融资讯分类任务（FPB），预测特定领域的金融新闻及话题（FiQA SA），股指推理（ConFinQA）等特定任务上的表现大幅领先于现有的近似规模的开放模型[2]。BloombergGPT 的推出说明海外在大模型金融科技应用方面已经取得了一定的成果。除此之外，一些传统金融机构也通过基

[2] Wu et al, 《Bloomberg GPT: A Large Language Model for Finance》

础大模型的应用提升业务竞争力，大型国际投行 Morgan Stanley 已将 GPT-4 应用在财富管理领域打造内部智能助手从而辅助其财富管理顾问快速搜索所需资讯，高效地为客户提供服务。与此同时头部对冲基金 Citadel 也拟在全公司各条业务线中应用 ChatGPT，提升业务运作效率。

而我国大模型和数字金融已有较好的产业发展基础，宜抓住此轮大模型科技变革机遇，进一步提升我国数字金融国际竞争力。2023 年中央金融工作会议提出将数字金融上升到国家战略部署的新高度，而大模型等新技术将进一步扩展金融科技的发展空间。根据《金融科技发展规划(2022-2025 年)》，目前应要抓住全球人工智能发展新机遇，深化人工智能技术在金融领域的应用。因此，我们应把握大模型技术浪潮，提升金融科技全球竞争力。

大模型应用技术分析 02

2.1 大模型在金融领域的应用挑战

由于金融行业的专业性、严谨性、合规性等特点，在把大模型技术应用到金融领域时，需要解决下述挑战，如图 2-1 所示。

通用大模型的金融专业性不足	金融领域具有高度的专业性，涵盖了复杂的金融理论、模型和实践，有着独特的术语内涵和表达方式。这些内容在常规的大数据训练集中往往表现不足，使得通用大模型在理解复杂的金融概念和操作上显得力不从心。
通用大模型的金融情境理解能力不足	金融市场高度情境敏感，同一事件在不同的情境下可能释放出不同的信号。例如，某一公司发布的财务报告如果不符合市场预期，对于该公司而言可能是负面的，但对于寻求低估值入市的投资者而言却可能是一个机会。通用大模型很难精准把握这种情境下的语义差异和心理预期，这就要求模型能够更加敏感地对待金融语境和事件，需要对这些模型进行金融情境的深度训练和优化。
通用大模型难以完成较复杂的金融指令	金融领域在交易过程中存在大量较复杂的工具指令，如限价单、止损单等，都需要精确的表达和执行。这些指令往往与特定的金融逻辑紧密相关，通用大模型如果不能准确执行这些复杂的金融指令，就很难在金融领域中得到有效应用。
通用大模型难以满足金融场景的定制化需求	金融领域具有高度的多样性，不同的机构和场景可能有着截然不同的需求。例如，投研场景会关注实时热点分析，投顾场景需关注投资者安抚等。通用大模型无法满足这些多样化和定制化的需求，从实践来看在落地过程中还涉及到具体的定制化调优。
通用大模型难以满足金融领域应用的合规要求	金融市场受到严格的法规制约，包括反洗钱（AML）、客户了解程序（KYC）、数据保护法规、适当性义务等。这些法规要求金融机构在处理客户数据和执行交易时必须遵循特定的规则和程序。通用大模型可能在设计时没有充分考虑这些合规性问题，因而在应用时可能无法确保机构的业务操作符合监管要求。

图 2-1 大模型应用到金融领域时需解决的挑战

面对上述挑战，金融机构在应用大模型到金融业务场景的过程中，一般需要经过两个主要步骤：一是从通用大模型进一步训练调优出专业的大模型；二是以大模型为核心，结合金融专业知识库、金融专业工具库、智能体、安全合规组件等构成一个可满足金

融领域安全应用要求的应用系统，来支撑在金融应用各场景中的应用，如下图所示。

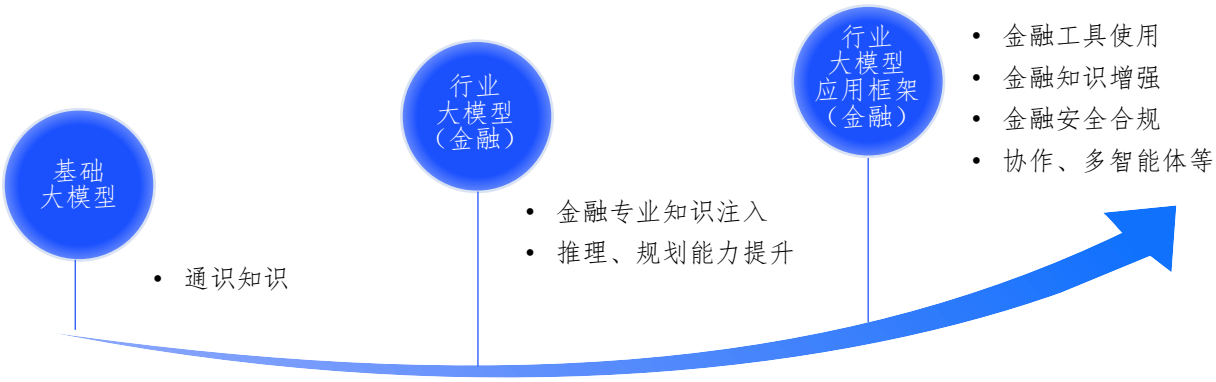

图 2-2 大模型在金融领域落地应用路线图示意

2.2 金融领域的行业大模型开发技术

2.2.1 开发技术框架

一个完整的大模型构建和应用流程如下图所示，包括：从数据收集和处理开始，通过领域适配训练使模型理解金融语境，然后通过性能优化确保模型的实用性和高效性，接着处理幻觉问题以提高事实性，最终实现复杂推理的能力。

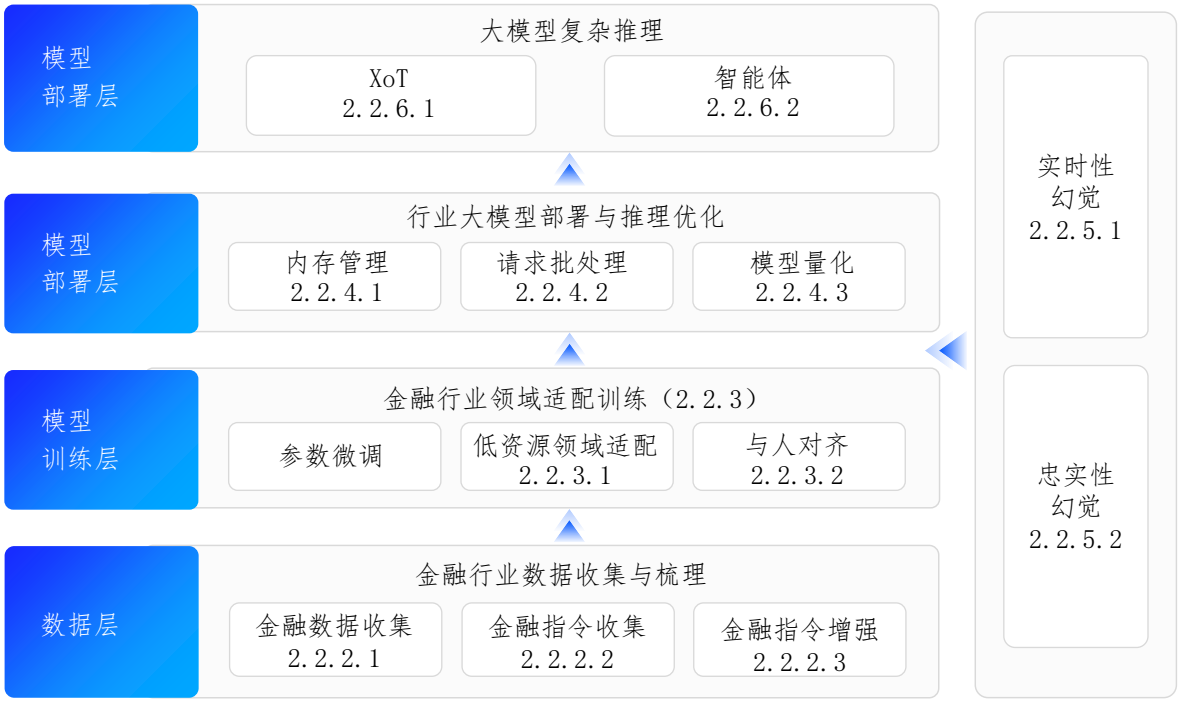

图 2-3 大模型开发技术框架

框架中各层主要关注的问题如下：

- **数据层**：构建大模型的第一步是数据收集和处理，这涉及搜集金融领域的大量数据集，包括公司公告、金融新闻、投资研报等。此外，为了使大模型具备处理下游各类金融任务的能力，还需要收集多样的、高质量的金融指令数据。

- **模型训练**：此处主要关注大模型领域适配训练，通常包括有监督的参数微调和对齐技术，以调整模型对金融术语、概念和上下文的理解，使其更好地适应金融行业需求，并符合人类价值观。此外，还需要考虑到低资源条件下领域适配技术，以满足实际应用中成本和条件的要求。

- **模型部署**：金融应用中模型的快速响应至关重要。需要考虑在特定的硬件资源下，如何提高模型的推理效率，从而改善用户体验和决策支持的实时性。

- **复杂推理**：金融场景的复杂推理能力是大模型的高级功能，允许模型进行多步推理和决策支持，这通常涉及到构建复杂的推理链、使用情景模拟和智能体决策技术等。

- **幻觉降低**：金融领域的高准确率和事实性要求，需要大模型能够有效处理幻觉问题以降低误导性决策风险，这包括开发和应用技术来识别和纠正模型在生成预测或解释时可能产生的忠实性幻觉和事实性幻觉等。

2.2.2 金融数据收集与梳理

2.2.2.1 金融数据集收集

金融数据集的构建是一项综合性工程，涉及预训练数据、指令数据和安全数据这三种主要类别（如表 2-1 所示），每一类别的数据都对大型金融语言模型的训练起到不可或缺的作用。

数据类别	描述	主要数据来源	具体描述
预训练数据	负责为模型输送必要的语境认知、语言结构理解以及广泛的知识背景。在金融领域的大型模型预训练过程中，引入专业金融数据是至关重要的，它确保了模型能够准确把握金融行业特有的知识和表达风格，与通用大模型不同，金融语料往往存在获取困难，数据非结构化等特点	企业财务报告	包括但不限于财务报表、盈利预测和负债情况等。这些数据主要来源于公司的年度和季度报告，可通过上市公司的公告、证券交易平台以及金融数据服务供应商获得。使用这些数据需对表格、图表等进行转换，以便模型能够解析和理解其结构化的数据格式
		金融领域学术论文与书籍	这些文献深入探讨金融理论的基础知识，包含专业教材、投资指南、个人理财策略、经济学原理等内容。这些资源可以通过学术数据库或图书馆访问
		行业分析报告及市场研究	这类报告提供关于特定行业或市场的深入分析和洞见。源自金融咨询公司和市场研究机构的报告往往需要通过商业采购来获取
		金融产品说明	诸如基金投资策略、保险条款等介绍性资料，这些信息多由券商、基金公司以及保险产品供应商提供
指令数据	构建金融指令集的目的是使人工智能模型适应金融领域的专业性和复杂性，增强对金融术语、计算、规范的理解与应用能力。这为用户提供精准、合规的专业建议和决策支持，同时满足特定金融角色的需求，推动金融多样化服务	金融知识指令	覆盖金融、投资、经济、会计等基础理论，和针对保险、基金、证券等具体金融产品和服务的行业应用知识，金融知识指令有助于提高模型在处理专业金融问题时的准确性和专业表达
		金融计算指令	包括财务分析和复杂计算公式的操作，金融计算指令不仅要求大模型具有数值计算能力，并且需要有将金融问题转化为计算问题的理解能力，相关指令可以使模型具备执行精确计算的能力，帮助用户做出更好的财务决策

		金融遵循指令	金融行业受到严格的监管和合规要求，具有高度专业与严谨的特性。金融遵循指令确保输出内容符合金融行业规范和写作标准
		金融角色指令	大模型的应用受众包含专业的投资研究员以及非金融专业用户，通过构建不同的金融角色，如投资顾问、分析师，基金经理等，在构建具体应用时可以使模型更好地服务于特定的用户群体。
安全数据	大模型在提升知识与表达能力的同时，需要具备安全底线，不能表达不符合金融、人道价值观的问题，也不能出现频繁拒答的情况，从而误导用户，这一部分的数据构建往往需要具备专业金融知识的专家协助	拒答数据集	此数据集确保在大模型遇到敏感议题、潜在的隐私泄露风险、法律合规约束，以及可能导致误解的金融咨询请求时，能够恰当地选择不予回答。构建此数据集的挑战在于准确定义拒答的边界，确保模型在遵循合规性的同时，依然能够提供有价值的信息。该数据集需定期更新，以确保其内容与最新的监管政策和行业规范同步
		金融价值观	该数据集涵盖了与金融行业伦理标准和法律规定相契合的案例、规章及导则，旨在训练大模型在提供咨询服务时，确保输出内容符合行业的合规性标准 例如，模型在未持牌的情况下，应避免提供具体的投资建议、预测市场走势或对板块、市场、股指未来点位进行预判，同时不得对国内市场进行不当描述

表 2-1 金融数据集类别

2.2.2.2 金融指令数据集构建与增强

高质量金融指令数据集的构建对大模型在金融领域的应用效果提升非常重要。大模型在特定场景中应用时，其核心能力之一是对人类指令的准确响应，以提供与人类意

图和价值观一致的反馈。这一能力依赖于有监督微调，即使用成对的（指令，响应）数据对模型进行进一步训练。这种训练方法以“遵循用户指令”为目标，约束模型输出，以确保其在处理请求和查询时的行为符合预期。在金融领域，准确和专业的数据对于风险评估和决策至关重要，当前金融数据非标准化和碎片化问题如数据类型和格式的混杂、知识来源的分散，制约了大模型的应用效果。

金融指令数据集构建主要面对数据质量不一和高质量数据稀缺的挑战。指令微调数据集的发展历程如图 2-4 所示。当前技术解决方案主要在两个方向寻求突破：一是指令生成技术的创新，通过设计预期形式和自动化方法（如自动化的指令生成器）来批量生成高质量数据；二是指令处理技术的改进，旨在优化数据筛选和构建过程，确保即便在低质量数据的情况下也能有效微调。通过上述策略，大模型能够更准确、有效地处理复杂金融场景中的指令，提升其在实际金融应用中的可靠性和专业性。

图 2-4 指令微调数据集的发展历程

自动化指令生成技术正成为当前解决数据分布不平衡和质量参差不齐等问题的关键。如图 2-5 所示，主要包括自指令方法、进化指令和指令适应等技术。这些发展展示了自动化金融指令数据生成技术在提高模型在复杂任务中表现、降低人工成本、以及提升数据生成多样性和质量方面的重要作用。随着这些技术的不断进步，可以预见大模型可以更好解决在金融应用中的数据稀缺挑战。

图 2-5 自动化指令生成技术进展

2.2.3 金融领域适配与参数微调

在大模型的适配应用中，微调技术扮演重要角色。通过微调，大模型不仅保留了模型在预训练期间获得的广泛知识，还能够细致地适应金融领域的具体需求。金融领域对模型的能力要求尤其严格，不仅要求模型理解复杂的金融术语和原则，还要求在日益复杂的监管环境中做出合规的决策。通过微调，大模型在学习了通用数据的基础上，进一步吸收了特定金融任务的细节。这种精确调整模型参数的技术确保模型的输出不仅精确，而且符合金融行业的高标准和法规要求，这对增强金融机构的信任度、降低运营风险以及提高决策效率至关重要。

本节主要关注高效参数微调和与人对齐的微调技术。这些微调技术的应用，确保了大

模型在在有限的算力资源下，专业性、精确性、伦理性和实用性方面都能达到更高的标准，为金融行业的发展提供强有力的技术支持。

2.2.3.1 高效参数微调

在金融行业中，尤其是在资源有限或对计算成本敏感的环境下，高效参数微调（Parameter-efficient fine-tuning，PEFT）技术允许即使是小型机构也能利用先进的大型预训练模型来强化其数据分析和决策过程。通过优化计算资源的使用，高效参数微调降低了大模型进入门槛，使得大模型能够在不牺牲性能的前提下快速适应金融特定任务。这使得缺乏大规模计算能力的用户也能从大模型中受益。PEFT 技术中三种常见方法如下图的简要介绍。

图 2-6 PEFT 常见方法

未来，PEFT 技术的发展可能集中在提升重参数化方法的泛化能力和表达能力，以及探索基于多层 Transformer 的自适应微调方法，以进一步提高模型在特定领域如金融的准确性和效率。

2.2.3.2 与人对齐技术

与人对齐的微调则专注于提升模型的道德和社会意识，确保其输出不仅在技术上先进，而且在伦理和价值观上与人类社会的期望保持一致。在金融领域，这意味着模型生成的预测或决策不仅要准确、可靠，还要公正、透明，并且符合行业规范。随着人工智能决策在经济和社会层面的影响日益增大，确保模型行为符合人类价值观变得更为重要。与人对齐的微调可以减少偏见、提高模型的普遍接受度，建立金融服务中更强的信任和可靠性。通过对齐，大模型能更好地服务于人类，提高决策质量，降低风险，增强客户信任。

- **基于强化学习和人类反馈训练的对齐技术：**RLHF (Reinforcement Learning from Human Feedback)是一种结合了监督学习和强化学习的技术，目的是根据人类反馈优化模型的行为。该技术被 OpenAI 用于 ChatGPT 的与人对齐，是最广为人知的对齐技术之一。这一过程涉及结合监督微调和强化学习来训练模型。监督微调使用人类注释的数据来教导模型期望的行为。然后，强化学习根据人类反馈细化这些行为，鼓励模型生成更符合人类偏好和指令的响应。RLHF 使用了 PPO（Proximal Policy Optimization）作为强化学习算法，用于将奖励模型的分数作为反馈来调整模型的行为。RLHF 的关键在于它将人类的直观判断和反馈直接融入模型的训练过程中，使模型能够更好地理解并遵循人类的价值观和意图。

- **对强化学习的化简：**基于 PPO 的 RLHF 存在代价高、训练困难等问题。因此，后续的方法关注如何改进 PPO 策略，以获得代价更低、更稳定的结果。RAFT (Reward Aligned Fine Tuning)通过使用奖励函数排名的样本来替代 PPO，这种方法计算效率更高，避免了标准强化学习算法所需的繁重梯度计算。RAFT 在平衡奖励与生成质量方面表现出色。DPO（Direct Preference Optimization）同样简化了复杂且不稳定的 PPO 过程，直接使用基于人类偏好的二元交叉熵目标来优化语言模型策略。这种方法消除了对显式奖励建模和强化学习的需求，使其更稳定、性能更好且计算效率更高。CoH(Chain of Hindsight)简化了奖励函数和强化学习，将所有反馈转化为句子并对模型进行微调来学习。这种方法让模型能从正面和负面的反馈中学习，提高了模型识别和纠正错误的能力。

总体来说，这些方法都旨在通过不同方式确保大模型在决策支持、风险评估和预测等方面能够反映人类的价值观和伦理原则，从而提高模型的社会接受度和信任度。

2.2.4 大模型推理

大模型推理是指使用训练好的模型对新输入数据进行理解、总结、生成及预测的过程。由于金融领域的行业特殊性，大模型推理往往对速度及吞吐量有较高的要求。

首先，金融行业具有时效性和实时决策性。金融市场的动态变化迅速，股票价格的波动、市场新闻的发布、政策变动等都可能影响最终决策，而传统人工需要花费大量精力做到实时响应,但大模型则能够快速地进行推理,以便在关键时刻提供准确的结论。

其次，优质的用户体验是金融服务成功的关键因素。广义上的用户不仅包含使用金融终端应用的普通用户，也包括研究员、基金经理等广大从业人员。大量高频的请求也使得大模型推理服务需要具备较大的吞吐量,从而处理尽可能多的数据来提升用户体验。本节主要从内存管理、请求批处理、模型量化这三个角度阐述推理优化技术。

2.2.4.1 内存管理

在大型语言模型，特别是基于 Transformer 架构的模型中，内存管理技术能有效提高推理效率和降低资源消耗。Transformer 的 Attention 机制虽然能精确捕捉上下文关系，却在推理过程中消耗大量的时间和空间资源。因此，内存管理技术主要解决在如何高效管理 GPU 内存空间的问题，特别是 Attention 操作的内存需求。

内存优化基本思路。内存管理的基本策略是利用现代 GPU 的内存层次结构，包括 SRAM 和 HBM，来优化大模型的推理服务。不同类型的内存有其特定的优缺点，例如 SRAM 虽内存小但速度快，而 HBM 则内存大但速度较慢。有效的内存管理策略旨在平衡这些内存类型的特性，优化数据存取效率。

基于 SRAM 的内存优化	基于 HBM 的内存优化
FlashAttention 是一种典型的基于 SRAM 的优化技术。SRAM 作为片上内存，虽内存容量有限（如 A40 的 20 MB），但运算速度极快（可以达到 19 TB/s）。FlashAttention 通过将 Attention 操作的计算从 HBM 转移到 SRAM，实现分块计算和算子融合。这样不仅减少了对 HBM 的 IO 读写需求，还有效克服了内存交换带来的计算瓶颈，节省内存的同时提高了推理速度。	PagedAttention 是一种典型的基于 HBM 的优化技术。HBM 作为传统显存，容量较大（如 40 GB）但运算速度较慢（如 1.5 TB/s）。在大模型推理过程中，KV cache 保存当前上下文以预测下一个单词，但其大小高度可变且不可预测。PagedAttention 通过在非连续内存空间中存储连续的 key 和 value，解决了由于内存分配不当（如碎片化）导致的 HBM 内存浪费问题。这使得相同的 HBM 空间可以支持更大的批处理大小，极大提高了系统的吞吐量。

图 2-7　内存优化方法

2.2.4.2 请求批处理

传统批处理采用静态批处理（Static batching）方式，批大小在推理完成之前保持不变。因此在之前的请求没有处理完毕时，当前的请求必须一直等待。这种处理方式的吞吐量较低。为了解决这一问题，动态批处理和和连续批处理技术被提出。

动态批处理	连续批处理
与静态批处理相比，动态批处理（Dynamic batching）会在新的请求到来时将模型的输入填充到相同的长度并加入到批次中，以提高 GPU 的利用率。然而，批请求的输入中填充长度影响了内存利用率。	连续批处理（Continuous batching）通过细粒度的调度和内存优化技术有效地解决了静态批处理和动态批处理中存在的问题。这类技术方案允许请求根据需要加入或离开批次，从而消除了对请求的输入进行填充的需求。vLLM 实现了连续批处理。在 vLLM 中，新到来的请求会抢占现有请求的生成过程以进行提示词处理（prompt processing）。DeepSpeed-FastGen 同样实现了连续批处理，不同的是 DeepSpeed-FastGen 会将长的提示词分成小块，并在多个前向传播中进行调度，在最后的前向传播时执行生成过程。

图 2-8　动态批处理和连续批处理方法

2.2.4.3 模型量化

模型量化是一种高效的网络参数压缩方法，它通过将神经网络的参数和状态从 32 位或 16 位浮点数转换为更低的精度（例如 8 位或 4 位），来提升推理速度并减少显存占用。量化降低了单位数据的位数，从而减少了计算过程中的 IO 通信量，使得通过增

加批大小的方式进一步提高模型推理的吞吐量。量化方法根据实施时机的不同，可分为训练中量化和训练后量化。

训练中量化	训练后量化
训练中量化（Quantization-Aware Training），也被称为在线量化，需要在量化的同时结合反向传播对模型权重进行调整。这种方法利用额外的训练数据以保留预训练数据的分布特征。例如，LLM-QAT 是一种用于大模型压缩的先进技术，它通过利用预训练语言模型本身生成的数据，更好地保存了数据的分布特征。与基于 RTN（round-to-nearest）的方法相比，LLM-QAT 在多个指标上显示了显著的优势。	训练后训练（Post-Training Quantization），也称为离线量化，是在已训练的模型上进行的，通常使用少量或不使用额外数据进行模型量化过程的校准。GPTQ 是 PTQ 中的一项著名技术方案，它基于早期的 OBD（Optimal Brain Damage）方法开发。OBD 是一种剪枝方法，通过去除影响目标最小的参数来减少模型参数。GPTQ 在此基础上应用了懒批次更新和乔列斯基分解，克服了通信带宽的瓶颈并减少了计算量。这使得 GPTQ 能够在短时间内将大模型高效、精确地压缩到低精度。

图 2-9　模型量化技术

2.2.5 幻觉问题与缓解策略

在金融领域应用中，大型语言模型面临的一个重要挑战是幻觉问题，尤其是内容的非忠实性（Faithfulness）和非事实性（Factualness）。这些幻觉影响模型输出的可靠性，对基于这些输出的决策产生负面影响。因此，有效缓解幻觉对于确保金融领域的精准实施与严谨推理至关重要。

幻觉的定义：一般可分为事实性幻觉和忠实性幻觉两类：

- **事实性幻觉：** 指生成内容与可验证的现实世界事实之间存在差异，如事实不一致或捏造。
- **忠实性幻觉：** 指生成回答与用户意图不一致，如指令不一致和上下文不一致。

幻觉的产生源自大模型开发的多个流程，如下图所示。

预训练阶段	智能体将根据问题的性质和需求，设计出一个解决问题的框架或计划。这个框架或计划可以来自于专家的经验，也可以通过搜索类似问题的解决方法来生成。

图 2-10　幻觉的产生原因

2.2.5.1 事实性幻觉的缓解策略

针对大型语言模型在金融领域应用中遇到的事实性幻觉问题，以下是一些有效的缓解策略：

- **高质量数据集的使用**：通过使用高质量、专业领域的数据集，如维基百科和"textbook-like"数据源，可以提高模型在事实方面的准确度。还可以向上采样事实性强的数据，提升数据集中准确信息的比例，以增强大模型的事实性。

- **诚实导向的微调（Honesty-oriented SFT）**：在训练数据中加入模型无法回答问题的实例（如“Sorry, I don't know”），培养模型自我边界认知能力。旨在减少模型在不确定情况下的过度自信，但需注意避免过度拒识的风险。

- **强化学习（RLHF）**：通过设计针对幻觉的奖励分数，在 RLHF 阶段优化模型。能有效减轻幻觉，但也可能使模型过于保守，削减其能力。

- **对比解码（Contrastive Decoding, CD）**：利用更强大模型和较弱大模型在单词预测概率上的差异作为关键决策依据。优先选择预测概率差异较大的单词，生成流畅、词汇丰富且内容连贯的文本。

- **对比层解码（DoLa）**：通过对比不同变换器层的输出来提高语言模型的事实性。该方法利用了一个观点：事实知识在语言模型的较高层中更为突出。通过比较高层和低层的输出，并强调高层的知识，DoLa 减少了幻觉，提高了生成内容的真实性。

这些策略涵盖了从数据质量改进到微调方法创新，以及解码策略优化等多个方面，旨在全面提升大模型的事实性。特别是在数据集选择、训练策略设计以及推理过程优化方面，这些方法可以有效减少幻觉，增强模型输出的可靠性和准确性。

2.2.5.2 忠实性幻觉的缓解策略

忠实性幻觉影响着模型的可靠性和准确性。以下是几种有效的缓解策略：

- **思维链（Chain-of-Thought，CoT）**：通过引导大型语言模型展开详细的推理过程，思维链技术提高了模型在复杂问题上的逻辑性和连贯性。这种方法特别适用于大规模模型，能有效提升推理的准确性。

- **上下文预训练和检索增强**：上下文预训练通过优化训练数据的组织方式，增强了模型对上下文的理解能力。检索增强（RAG）则通过结合外部知识源，增强了模型的信息检索和整合能力，从而提升了其在复杂任务中的表现。

这些策略从不同方面缓解了忠实性幻觉问题，提高模型输出的忠实度和可靠性，进而增强在金融领域等专业应用中的实用性。

2.2.6 金融领域复杂推理

2.2.6.1 思维链增强方法

思维链被认为是一种开创性且最具影响力的提示工程技术，它指引大模型提供中间多步推理过程来获得最终结果。但是，这种常规的线性链式结构一定程度限制了对金融领域的复杂任务上的推理能力，于是需要进一步采用思维链增强方法来提高大模型在

金融领域的推理能力。

方法类别	具体描述
思维链结构变体方法	常规的线性链式结构一定程度限制了对金融领域的复杂任务上的推理能力，于是可采用程序语言或算法（Algorithm-of-Thought）代替自然语言，利用程序算法作为推理链条；为进一步拓展思维链探索广度，构造思维树结构（Tree-of-Thought），使用树搜索算法对不同推理路径进行探索；对于更复杂的金融任务，引入图拓扑结构（Graph-of-Thought），进行信息聚合和多路径推理，以获得更通用、更全局的推理视角。
思维链推理结果验证方法	一方面，对思维链每一个金融分析和推理步骤进行细粒度校验，通过演绎推理检验前后推理的一致性，即前向推理验证。另一方面，根据金融问题和模型的预测结果来反向推理其发生条件，通过比较推测出的条件与真实条件的一致性来判断推理的正确性，即反向推理验证。Google 提出的 Self-Consistency 方法生成多个答案候选，并在其中寻找一致性，最终选择最一致的答案，可有效提高大模型在金融知识问答和文本补全等任务上的性能。
思维链推理过程验证方法	与推理结果验证方法相对，该方法专注于推理链中每一个单独的推理步骤的效验。例如，Self-Check 方法通过对推理过程的每一步进行验证来确保逻辑的严密性；GRACE 方法则进一步优化这种验证，通过引入额外的校验机制提高推理的可信度。
思维链问题分解方法	对于复杂金融推理任务，可采用自顶向下的问题分解策略，将一个复杂问题分解成若干个子问题，然后逐一解决从而得到最终答案。另一种常用方法是采用一种迭代分解策略，每次迭代分解出一个子问题并对其进行推理解答，以递推方式进行后续问题分解和回答。
外部知识增强方法	从金融知识库、金融知识图谱、以及金融相关的百科和词典等，引入外部金融知识，从其中获取结构化知识进行知识指导下的思维链推理，同时根据结构化知识对推理的真实性和可信性来进行验证。

表 2-2

2.2.6.2 智能体推理

金融市场的高度复杂性和快速变化对分析方法有了更高的要求。传统分析方法通常依赖于固定模型和有限的数据处理能力，因而难以适应这种动态性。而智能体（Agent）可以通过持续学习和自我调整，更有效地理解和适应市场变化。它们具有处理大量多样化信息的能力和实时反应机制，能够解决传统方法难以应对的复杂金融问题。

智能体是通过在特定环境中感知、思考和行动来实现特定的目标的计算实体，具备自主性、反应性、社会性、主动性等特征。在金融领域，智能体通过计划、记忆和行动等三个模块的紧密配合来实现目标。计划模块制定和优化策略，记忆模块存储经验和知识，而行动模块将这些策略和知识转化为具体行动。这种协同作用使得智能体能够有效地处理复杂金融任务，并持续学习和适应变化，以提高其在金融环境中的性能和效率。

在计划模块方面，智能体借鉴了人类处理复杂任务时将其解构为更简单的子任务来完成，根据执行过程中的环境反馈结果，迭代进行计划修正，其中主要包括了任务分解和模型自我反思两个关键过程。

任务分解是将复杂任务分解为更易于管理的子任务，并为每个子任务制定合理计划。一类常用方法包括解决简单金融问题的以思维链（CoT）为代表的逐步规划和执行方法、解决较复杂金融问题的以思维树（ToT）为代表的多路规划并择优路径选择方法、以及解决多因子耦合复杂关系金融问题的思维图（GoT）为代表的更复杂操作规划策略等。该类方法本质上是通过精心设计提示，激发和调动大模型中潜藏着的更擅长规划的认知部分。另一类方法则是借助外部金融领域专用问题规划器，进行整体系统性逻辑规划，例如，利用大语言模型首先将问题翻译成问题规划域定义语言（PDDL）描述，然后利用外部专用规划器搜寻最佳计划，然后生成计划规划语言，最后再利用大语言模型将该计划规划语言翻译成以自然语言表达的计划，以驱动行动模块执行任务。

智能体自我反思（Self-reflection）是对以前制定的计划进行回顾性思考，以纠正

之前错误认知并完善行动决策来不断改进计划效果。这类自我反思主要来源于智能体内部反馈机制、与人类互动获得的反馈以及从环境中获取的反馈三个方面。

基于内部反馈机制的反思	通过智能体内部机制强调学习过程中的自我调整和持续改进。例如，Reflexion 框架通过自我反思和语言反馈提升智能体的推理能力。该框架在标准强化学习环境中加入语言元素，学习避免重复错误的经验，通过内部记忆映射适应环境。在金融领域，智能体每次执行任务后，通过启发式函数评估当前效果，并决定是否重置所处的环境，以更好地应对快速变化的金融市场的挑战。
基于人类互动反馈的反思	通过与人类直接互动获得反馈，有效确保智能体与人类的价值观和偏好一致，同时有助于缓解幻觉问题，对于金融领域强监管、强规范的要求下这点尤为重要。例如，在 ChatGPT 的训练中采用的基于人类反馈的强化学习 RLHF 方法。
基于环境反馈的反思	智能体利用客观世界或虚拟环境的反馈进行反思。例如，ReAct 将推理和行动结合起来应用到大模型上，其中推理轨迹有助于模型归纳、跟踪、更新行动计划，并辅助进行异常处理；而行动则通过与知识库、维基百科 API、环境等外部信息源交互收集必要反馈信息。金融市场环境瞬息万变，如何实时地对环境反馈做出快速反思和应对，又能够兼顾短期、中期和长期的市场趋势是对金融 Agent 提出更高自我反思要求。

图 2-11

在记忆模块方面，智能体需要特定的记忆机制来确保熟练处理一系列连续任务，其中记忆模块负责存储从环境中感知到的信息，并利用这些记忆促进未来的行动。这种机制有助于智能体积累经验、自我进化，以更加一致、合理、有效的方式行动。智能体涵盖多种记忆类型，包括感知记忆、短期记忆和长期记忆。

感知记忆	能够在原始刺激结束后保持对感官信息的印象，包括图像记忆（视觉）、回声记忆（听觉）和触摸记忆（触感）等，可作为金融领域相关数值、文本、图像和视频等多种模态的智能体原始输入。
短期记忆	存储智能体所知信息，以及执行复杂的学习和推理等认知任务所需要的信息，如包括提示工程的上下文学习等。该类型记忆时间较短且影响范围有限，受到智能体网络框架 Transformer 的上下文窗口长度的限制。所以，为了增强智能体的记忆能力，尤其记忆垂直领域的上下文信息（如金融领域的行业规范、任务要求以及当前金融市场情况等），可通过增加 Transformer 的输入长度来实现。例如 LONGMEM 通过解耦模型的记忆与知识，将上下文长度扩展至 65K，提升了智能体对丰富的提示示例的支持能力。

长期记忆	将信息存储较长时间，理论上可实现永久存储无限多的数据。例如，智能体在推理过程中需要查询外部的各类金融报告、金融数据库和知识库等，实现快速检索和访问数据。常用的实现方法是利用向量数据库，基于人工智能中的嵌入技术将金融文本、图像、音视频等非结构化数据压缩为多维向量。利用这种向量化数据管理方式构建结构化向量数据库，智能体可在其中进行快速、高效的数据存储和检索，从而赋予了智能体更为强大的长期记忆能力。

图 2-12

在行动模块方面，负责采取合适的行动将决策转化为具体结果。智能体的行动包括文本输出、工具使用和具身行动等三种主要类型，在金融领域，目前前两种类型应用更广泛，而后者正处于探索和发展阶段。

文本输出	使用工具	具身行动
大模型赋予智能体的原生能力，允许其通过自然语言与用户和环境进行交流。金融领域的智能客服、报告生成、文档审核、舆情分析等都属于该种类型。	工具可以显著拓展智能体的行动空间，使得它可以在推理和规划阶段利用各种外部资源，在行动阶段使用专业工具执行特定领域的任务。例如，HuggingGPT 将大模型作为中心控制器，通过自然语言与各种 AI 模型协同工作，共同解决复杂任务。金融领域已经存在大量的金融方法和工具，如投资策略、风险检测、收入预测、金融衍生品定价、资产组合优化和量化交易等。这些工具都可通过金融智能体进行系统性的规划、调度和调用，以完成那些专业化的金融任务。	智能体主动感知和理解物理环境，根据大模型内部知识做出决策从而产生近似人类行为的方式与世界互动，更全面地理解和适应环境。这类具有具身行动的智能机器人目前还处于快速发展中，目前仅在室内引导、交互式顾客服务等少数金融场景下初步应用。

图 2-13

2.3 行业大模型在金融领域的应用框架

2.3.1 应用框架

如前所述，在开发出具有应用到金融领域的行业大模型后，还需要以大模型为核心，结合金融专业知识库、金融专业工具库、智能体、安全合规组件等，进一步构成一个可满足金融领域安全应用要求的应用系统，如图 2-14 所示。

图 2-14 大模型应用框架

应用框架中各模块的主要功能介绍如下：

- **应用请求方：**在金融应用各场景中，向大模型系统发起服务请求的请求方。根据具体应用场景不同，可以是通过用户交互界面直接请求大模型的客户，也可以是需要调用大模型服务的其他金融应用。

- **输入内容安全组件：**对于应用请求方提出服务请求内容（Prompt）进行分析，并判断服务请求是否存在安全合规风险，如存在安全风险，可以对请求进行拦截。

- **大模型：**应用系统中的核心模块，对用户的输入内容分析，并判断是否需要调用金融知识库或金融工具库获取金融专业知识或者金融逻辑处理结果，并综合处理后得到返回给请求方的响应内容。

- **智能体：**可与大模型交互，自主的对复杂金融任务进行分解、规划、执行，并可通过学习和经验不断总结优化的一类工具。

- **金融知识库：**可以提供高时效、专业、可信和丰富的金融专业知识，来补足大模型在金融专业性上的不足。

- **金融工具库：**通过 API 接口对外提供金融专业工具服务能力的工具集合。

- **输出内容安全组件：**对于大模型生成的待返回给请求方的内容进行分析，并判断待输出内容是否存在安全合规风险，如存在安全风险，可以对输入内容进行安全改写，或者进行拦截。

2.3.2 金融知识库

大模型通过集成检索增强生成技术可显著提升其性能。检索增强技术，即 Retrieval-Augmented Generation（简称 RAG），结合了信息检索和答案生成两个步骤，通过从一个专门构建的知识库中检索相关信息来辅助生成更加准确和具有根据的回答。为此，首先需要创建一个全面的金融知识库，该库应包括历史金融数据、最新市场动态、研究报告、市场分析等内容。接着，通过将这些信息转换为高维向量表示，以便高效地进行相似性搜索。当用户给出提问时，可以采用 FAISS 等先进的向量搜索算法，以实现从知识库中迅速而准确地检索相关信息。通过将检索后的信息结合问题以 Prompt 形式输入语言模型即可获得经过检索增强后的回答。

金融知识库需及时更新以降低大模型生成误导性回答的风险。一般而言，金融领域知识库可包括行情类（如新闻资讯、热带事件）、投教百科知识类、专业内容（如研报）、董监高事实类（如基金经理、董监高等）等知识，在经过知识加工（如拆条、标题生成、实体识别、时效判别、向量表达等）后更新到知识库中。当大模型调用时，根据请求的查询（Query）词，经过预处理后（如意图识别、时效识别、关键词识别等），检索召回到最新的相关知识向量条目，并进行融合处理后返回给大模型相关的知识答案。这些答案可以帮助大模型降低产生错误或虚构的信息（即所谓的“幻觉”）的概率。实时或定期刷新知识库中的向量表示可以确保模型能够检索到最新的信息，从而减少依赖过时数据而产生误导性回答的风险。此外，上下文敏感的检索机制可以进一步确保生成的回答不仅基于客观事实，而且与用户查询的具体上下文紧密相关。

例如，当用户提问“巴菲特为什么减持比亚迪”，在响应用户的请求查询后，大模型

首先识别任务需求，判断是否需要调用金融知识库来检索相关新闻资讯。如需调用，大模型会去知识库检索多篇相关最新资讯，并获取到金融库检索召回到最为相关的知识答案。最后，大模型会将所有信息进行捏合并作为输入的提示词 Prompt，并通过自身的逻辑和表达能力生成最后的答案。

2.3.3 金融工具库

当前大模型在处理逻辑推理和高度专业化的复杂金融指令时仍有不足，可通过金融工具库进行补充。在当前的金融技术领域，尽管先进的大模型已能够执行一定的复杂任务，但它们在执行数值计算及处理高度专业化的复杂指令时仍会遇到较大挑战，主要是由于模型在逻辑推理和信息即时更新方面还存在局限性。因此，为了提升大模型的准确性并扩展其应用范畴，可以利用专门的金融工具库来补充其功能。金融工具库通常包括金融计算器、实时股票和基金查询系统、基金经理分析工具以及投资组合诊断工具等。通过这些工具的辅助，大模型能够更加精确地处理用户指令，尤其是那些涉及到专业金融知识和数据处理的任务，比如实时股票信息查询或进行复杂财务计算等。

大模型需要“学会”调用工具来提供最佳答案。对于输入的用户请求 Prompt，大模型需要在经过意图识别、实体抽取后进行需求分析(判断需求是否超出模型自身边界)，以决定是否需要调用外部工具。如果判定为需要，模型将进入决策阶段，选取恰当的工具，并构造适当的调用格式来访问对应金融工具库的 API。待被调用的金融工具库执行完任务并给大模型返回计算结果后，模型会结合原始用户指令、工具输入以及输出结果来生成综合性的回答。针对那些需要多个工具联合使用来解决的复杂问题，模型可以通过多轮工具 API 调用并汇总结果后来得到最终返回内容。为了使大型语言模型具备调用金融工具库的能力，可采用提示工程策略或对模型进行专门训练，教授其如何正确判断是否需要使用以及如何使用工具。

例如，对于用户查询当日股票价格的需求，由于大模型自身无法生成实时更新的数据，它就可以调用股票查询工具以获取最新的价格信息。此外，当用户需要计算净利润时，

模型也可以利用金融计算器的功能来辅助完成这项数值密集型的计算任务。

2.3.4 安全围栏工具

大模型在金融应用的前提是能够保障安全合规。当前大模型安全问题已成为产业关注热点，这些问题存在于从数据到算法到模型应用的全周期关键节点，除了前文中提到的幻觉问题，更包含隐私风险、模型攻击、缺乏可解释性、缺乏可溯源性、以及有害内容生成等。大模型安全问题在合规要求更为严格的金融领域则显得更为突出。例如在没有相应牌照的情况下，模型不得进行基金推荐服务，不得采用明显的销售推广话术，也不得传播有悖金融价值观的信息。

在提升模型原生安全能力基础上，再结合安全围栏工具，是当前保障大模型应用系统整体输入输出安全合规的可行方案。提升模型原生安全的措施包括：一是通过对安全指令的训练，模型能够更精确地识别和响应复杂的人类指令，从而增强其对复杂道德问题和金融合规要求的理解，并减少误解的风险；二是通过强化学习等机制，借助人类反馈调整自身偏好，从而更好地理解并遵循安全和道德规范。不过，当前仅依赖大模型自身想确保大模型的安全合规内容生成仍存在较大压力。安全围栏工具相当于在大模型外围又加上了一个“防护盾”，通过智能化风控技术，可以帮助大模型挡住外界的恶意提问，同时对生成的回答内容进行风险过滤，保障大模型上线后从用户输入到生成输出的整体安全防御。具体而言，输入内容安全组件对服务请求进行分析，筛选敏感或不合规内容，并采用模糊匹配和深度模型深入理解上下文，以识别安全风险（例如非金融相关查询）并在必要时拦截请求；输出内容安全组件负责监控和审查模型的响应，通过实时监测，在线风控大模型部署或正则策略以及离线安全改写机制，确保输出内容的金融合规性。

2.3.5 多智能体协同

在金融行业中，智能体的概念已成为提高决策质量的关键要素。尽管金融智能体通常配备了丰富的金融知识和较强的逻辑推理能力，但面对高度复杂和不断变化的金

融市场，单一智能体仍存在局限。因此，构建一个协同工作的多智能体系统（MAS）成为提升整体性能和效率的方式。为了有效地完成复杂的金融任务，多智能体系统需要解决以下主要问题：

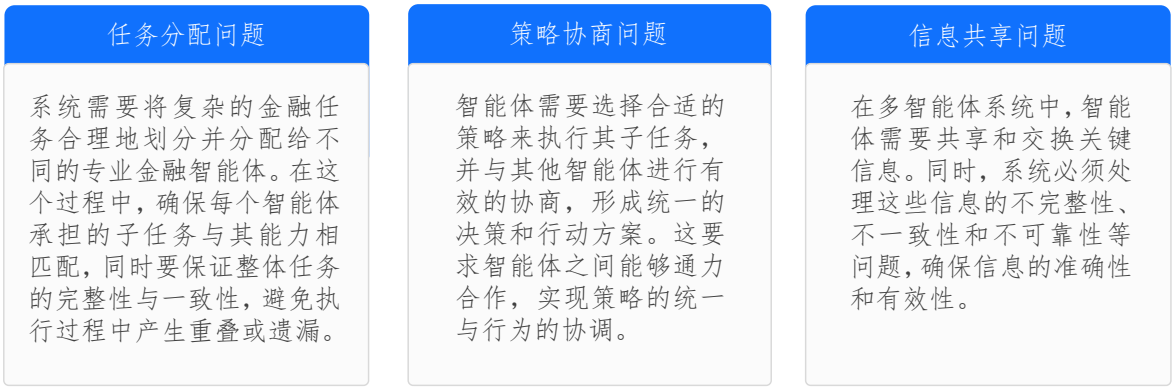

图 2-15

框架的设计：为解决上述问题，金融智能体系统的框架可按照人类专家组解决问题的方式设计拆分，即：策划（Engineering）、执行（Executing）、表达（Expressing）、评价（Evaluating）4E 范式。这种范式将问题的解决过程分为四个阶段，以实现复杂任务的逐步拆解、细化为可解决的单一任务、最终完成整体目标。

图 2-16

框架的价值：金融智能体框架的价值在于其能够将复杂问题的解决过程标准化、系统化。通过这一框架，智能体在确保解决方案的有效性和准确性的同时，保证了通用性，使得其可应用于解读金融市场热点、债券舆情分析、政策解读等各类问题。此外，该框架为人类大脑工作原理的理解和模拟提供了新的思路，为生成式模型在金融行业的落地发展打开了新的视角。

2.3.6 案例分析：巴菲特减持比亚迪股份

背景：2023 年 1 月，金融市场上出现了值得关注的大事件：巴菲特透过港交所权益披露信息，显示其半年内累计减持比亚迪股份超过 7000 万股。此举涉及的资金高达 150 亿港元，引起了市场和投资者的广泛猜测和讨论。这一决策背后的原因成为了分析的焦点。

应用 4E 框架解读：

- **策划（Engineering）：**首先，策划节点通过针对问题的理解以及大模型的原生知识可以针对上述的事件进行解读框架的拆解，例如：巴菲特对于投资的理念和原则是什么？他注重什么样的投资机会？比亚迪的业务状况和财务表现如何？公司的内在价值是如何评估的？巴菲特为什么选择在2008年金融危机后买入比亚迪股票？他持有比亚迪股票的原因是什么？

- **执行（Executing）：**在执行阶段，执行节点会根据策划节点的问题分析框架去执行相关的金融知识库检索，使用相关的金融工具库查询相关数据以及进行简单的逻辑推理和归纳，形成一个初步的答案，例如：针对巴菲特的投资理念和原则，执行节点会去搜索相关金融知识库关于巴菲特的咨询新闻，并通过大模型的理解生成能力总结出一个初步答案。针对比亚迪的财务状况，则可以通过调用专业金融工具库查询企业 2008 年后的相关营收，利润等具体财务咨询，并通过大模型的分析能力进行总结。

- **表达（Expressing）：**表达节点会将各个执行节点的答案捏合成一份详尽的报告，其中可能包含巴菲特减持比亚迪的潜在原因：比亚迪股价与内在价值的关系变化、比亚迪与其他新能源竞争对手的竞争力比较、以及巴菲特可能的资产配置调整逻辑。

- **评价（Evaluating）：**在评价阶段，智能体会对报告中的每项分析提出批判性的评估。它会检查所得出结论的合理性、准确性，以及是否全面覆盖了影响巴菲特

投资决策的所有潜在因素。如果最终结论没有回答原问题，或回答本身有逻辑性问题，则会提出修改意见或进行改写。

结论：通过 4E 框架的应用，可以将大模型基座，金融知识库，金融工具库串联起来，针对基座本身无法回答的实时复杂问题进行拆解并结合实时资讯，金融数据进行专业性回答。虽然无法完全揭晓巴菲特的真实动机，但通过框架的系统化分析，可以提供一个全面、合理的理论解释。

2.4 大模型的应用实践

2.4.1 投研场景

（1）应用背景

及时准确地获取金融信息、高效的金融分析工具，是影响投研水平的关键因素之一。随着财富管理行业的快速增长和普惠化，投研所需覆盖的资产和市场大幅扩展，原先金工定量+专家定性的人工模式，在效率效果上都难以满足发展诉求，新趋势也带来了新的挑战。

（2）应用方案或者产品介绍

蚂蚁集团支小助通过自动化采集，将研报、新闻、分析师音视频素材输入大模型，借助大模型的多模态理解能力，通过观点归纳和数据结构化，协助工作人员完成市场的高效解读。

（3）应用效果

支小助投研版的实测数据表明，其每日可辅助每位投研分析师高质量地完成超过100+篇研报和资讯的金融逻辑和观点提取，完成 50+金融事件的推理和归因，并将典型的量化分析任务的效率从天级别提升到小时级别，带来了明显的生产力提升。

2.4.2 保险场景

(1) 应用背景

“蚂蚁保”是服务千万在保用户、普惠性的互联网保险售卖平台，具有用户体量大（上亿在保用户）、投保性价比高等普惠特点，需要严格控制理赔运营成本。由于报案所需医疗凭证种类繁多，专业性强，这给用户的报案材料提交和理赔审查带来了困难，导致用户补充材料率较高，同时人工审查时间也变得更长，进而影响了结案周期。

(2) 应用方案或者产品介绍

“蚂蚁保” 通过搭建智能化理赔平台，建设了高精度的“自动化信息提取”和“自动化核赔”双智能引擎。自动化信息提取通过融合文档的图像、版面以及文字信息，构建高精度的自动化信息提取平台，实现材料分类、材料去重、凭证归档、凭证 KV 提取、票据表格识别等功能模块。自动化核赔通过将借助十万级典型理赔案件提取信息和结论，构造了高精度核赔决策模型。进行自动化核赔时，核赔决策模型首先针对用户上传的理赔材料，利用自然语言处理技术，进行关键信息（时间、诊断、手术、既往症、医院等）的实体识别、关系抽取和并按医疗事件进行组装，从而形成结构化的理赔案件。通过大模型的 CoT 逻辑思维链能力，该系统能够快速准确地判断理赔申请的有效性，避免人工审核中可能出现的主观性和误判。此外，与传统的基于分类的黑箱模型不同的是，本系统不仅能够给出核赔结论，在需要拒赔时还能够给出具体的拒赔原因，提升了用户体验。

(3) 应用效果

“保险理赔凭证识别和保险医学 NLP 引擎”可以作为健康险两核、保顾、健康服务等多个场景辅助甚至部分高发常规案例辅助医学背景业务专家高效诊断。

2.4.3 个人金融智能助理

(1) 应用背景

智能理财助理是旨在协助个人更有效地管理和配置资产。尽管智能理财助理在智能客

服，风险管理等方面已取得显著进展，而大模型在其中的应用则聚焦在非持牌的金融资讯推荐和投教知识上，但智能理财助理要完全替代人工金融专家仍面临一系列挑战。这些挑战包括金融信息过载、复杂金融任务拆解、专业术语晦涩，缺乏个性化投资建议等问题。

(2) 应用方案或者产品介绍

针对通用大模型专业金融知识缺失的问题，应用在智能理财助理中的大模型引入了可信、多元、实时的泛金融内容和知识，构建起百亿级别 Token 级别的通用+蚂蚁金融语料并通过模型知识注入与信息检索赋予智能理财助理兼具广度和深度的“知识力”。

金融行业的复杂性与用户期望的简明性之间存在着巨大的差距。为了弥合这一鸿沟，支小宝智能理财助理应用通过扩展上下文窗口至 32K，以深入理解用户意图，实现更连贯的多轮对话；通过构建对话仿真工具，蚂蚁内部训练了对话仿真工具，模拟专业理财专家与用户的对话，提升其理财领域语言能力；

针对通用大模型在金融领域应用面临的安全性及合规性问题，蚂蚁聘请超过 100 名金融专家对生成内容在隐私保护、合规表达、内容安全、上下文关联等多个维度评估，使用基于人类反馈的 RLHF 让大模型对齐金融业务的合规需求，并通过后置校验的方式保障安全底线及输出内容的合规性，在数据，模型，输出层面建起了“安全防护围栏”。

(3) 应用效果

通过大模型的范式，支小宝 2.0 有了兼具广度和深度的金融知识，专业金融工具调用能力，个性化的表达能力，以及安全可信的围栏能力。

2.4.4 零样本金融合同要素提取

（1）应用背景

在合同合规性审查领域，合同要素提取起着至关重要的作用。这个过程使审查人员能

够全面了解合同的内容和条款，识别潜在的风险和违规行为。确保合同的合规性、有效性和可执行性是任何组织合同管理工作的核心。通过有效的合同要素提取，审查的效率和准确性可以显著提高，为组织提供强有力的合同管理支持。

（2）应用方案或产品介绍

合同要素提取的一个重要挑战是，不同合同的抽取字段各不相同，且某些字段的训练样本稀少甚至完全缺失。为应对这一挑战，上财课题组提出了零样本要素提取的概念。这一创新目标旨在使模型具备对任意字段的抽取能力，即使对于那些之前未见过的字段。

为了提高要素提取的准确率，上财课题组基于合作公司提供的标注数据，训练了一款支持零样本要素提取的先进的大语言模型。此外，为了增加模型对于表格型数据的理解能力，增加了训练数据中表格内容的字段比例，提高了训练数据的质量。这一调整使得大模型在测试数据集上的综合准确率进一步提升。

（3）应用效果

要素提取大模型在测试数据集上的综合准确率达到了 85%，相较于 ChatGPT 3.5 的 53% 准确率，有了显著提升。对于金融和合同管理领域的组织而言，这意味着模型将提供更高效和可靠的合同合规性审查支持，从而降低潜在的法律风险和合同纠纷的发生。

大模型的应用安全 03

大模型在金融领域的实践需要考虑多方因素，除了大模型技术框架对现有金融业务的效率提升以外，金融业务的专业性、严谨性及合规要求对大模型在金融领域的应用实践也提出了更加严格的风险防控措施要求。

3.1 大模型应用在金融业务领域的风险分析及防控措施

图 3-1 大模型开发框架中的风险防控[3]

大模型在金融相关业务应用中有几大类风险维度及相应防控措施，其中包括针对全流程的隐私风险防控以及模型攻击防控；针对数据收集处理、适配与参数微调以及推理过程的可解释性增强；针对推理过程和生成内容的可溯源性增强及针对生成内容的有害内容防控。

3.1.1 大模型的隐私风险防控

由于金融业务所涉及的数据敏感，从模型开发到模型应用的过程中均有可能涉及用户隐私信息，而这些隐私信息不仅包含敏感的个人信息，更包括某些用户的资产信息。

[3] 图 3-1：预训练模型开发框架不在本框架图内

这些用户隐私的过度使用及间接泄露，可能会成为金融犯罪活动的导火索。

3.1.1.1 隐私泄露种类

隐私风险泄露根据攻击的方法分为基于记忆的隐私风险泄露和基于推断的隐私风险泄露。

基于记忆的隐私风险泄露是指大模型在学习中会形成对训练数据的记忆。这一方面可能导致敏感训练数据的泄露，另一方面可能导致数据在上下文中的误用。例如大模型可能在回复针对某用户的查询时泄露其它用户的电子邮箱。

而基于推理的隐私泄露是指大模型利用自身推理能力产生的隐私泄露问题。例如模型可能基于公共论坛或社交网络帖子自动推断出个人作者的各种属性。这极大地降低了侵犯隐私的成本，使得攻击者能在更大的范围内进行攻击[4]。

攻击类别	攻击方法	具体描述
基于记忆的隐私泄露	成员推断攻击	攻击者可以利用训练好的模型预测一个特定示例是否被用于训练该模型。方法可分为三类，分别是基于分类器的方法、基于度量的方法和差分比较方法。基于分类器的方法代表是影子训练(shadow training)，即在知道目标模型结构和训练算法的情况下，构建多个影子模型模拟目标模型行为，并利用影子模型的训练数据集构建成员推断数据集来训练攻击模型；基于度量的方法通常利用模型倾向于对存在于训练数据中的样本赋予更高的置信度这一观察来定义度量指标。而差分比较方法(differential comparison)首先构建非成员数据集，然后以迭代的方式将目标数据集中的样本移动到非成员集中。样本移动后集合距离的变化决定该样本是否为成员。成员推断攻击可能导致严重的隐私问题，例如针对金融信贷模型进行成员身份攻击可能会泄露训练集成员的信贷状况。

[4] Staab et al, 《Beyond Memorization: Violating Privacy via Inference with Large Language Models》.

	训练数据提取攻击	攻击旨在从模型中恢复训练数据。狭义上，它的目标是逐字逐句重构完整的训练样本，而广义上，它也可以指推断出和训练样本语义相似的数据。在黑盒设置下，狭义的训练数据提取攻击通常分为根据输入的提示进行解码和利用成员推断攻击对生成的结果进行过滤两个阶段。在 GPT-2 上 ，该攻击方式能成功恢复一个人的全名、地址和电话号码。此外，该攻击的有效性和模型大小、训练数据重复次数之间存在对数线性关系。狭义的训练数据提取攻击可以通过设计新型解码算法进行规避，例如 MEMFREE 解码 ，其在生成的每一步中避免选择会创建训练集中存在的 n-gram 的标记。然而这些方法依然无法规避从模型中推断出语义相似训练数据的问题。
基于推理的隐私泄露	自由文本推断攻击	通过人工构建提示从公开文本中推断出个人作者的隐私属性，例如住址，性别和年龄等
	对抗性交互攻击	模型以某种方式引导用户的对话，使他们产生的文本能够让模型推断出潜在敏感的信息

表 3-1 隐私攻击种类

3.1.1.2 隐私防控方法

针对上述隐私攻击，基于模型开发与应用流程，可分别应用数据治理、模型训练和模型后处理阶段的隐私防控手段。

隐私风险防控阶段	具体描述
数据收集与处理阶段	在数据收集和处理阶段可进行数据治理，清除训练数据中的敏感信息。数据治理是隐私防御中最直接的方式。PII（个人身份信息）清除是针对个人身份信息泄露的一种数据治理方法，用于从文本中删除个人身份信息，可能包括姓名、地址、电话号码、身份证号码等可用于识别特定个人的敏感数据。