



烽火通信

金融大模型智能问答解决方案

技术白皮书

版本：01

烽火通信科技股份有限公司

二零二四年八月

目 录

目 录.....	2
1 概述.....	4
1.1 大语言模型技术发展概述.....	4
1.2 大模型引领中国金融领域科技的国际化发展.....	4
2 大模型应用技术分析.....	6
2.1 大模型在金融领域的应用挑战.....	6
2.2 金融领域的行业大模型开发技术.....	7
2.2.1 开发技术框架.....	7
2.2.2 金融数据收集与梳理.....	8
2.2.3 金融领域适配与参数微调.....	11
2.2.4 大模型推理.....	12
2.2.5 幻觉问题与缓解策略.....	14
2.2.6 金融领域复杂推理.....	15
3 金融大模型智能问答解决方案设计.....	17
3.1 行业金融大模型方案架构.....	17
3.2 金融大模型智能问答方案功能模块介绍.....	18
3.3 金融大模型智能问答方案创新点.....	19
3.4 金融大模型智能问答方案应用实践.....	19
3.4.1 方案核心业务功能介绍.....	19
3.4.2 方案主要应用场景.....	20
3.4.3 方案应用成效.....	20

4 大模型评测	21
4.1 通用大模型评测框架	21
4.2 大模型在金融领域的评测概述	24
4.2.1 业务合规性	26
4.2.2 事实准确性	26
4.2.3 推理正确性	26
4.2.4 事件实时性	26
4.2.5 评估覆盖广度和深度	26
4.2.6 业务实践性	28
4.2.7 中文金融测评集	28
4.3 大模型在金融领域的评测实践	29
4.3.1 上财 FinEval 金融数据集	29
4.3.2 蚂蚁大模型评测集 Fin-Eva	30

1 概述

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,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 and 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(FIQASA)，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础大模型的应用提升业务竞争力，大型国际投行 MorganStanley 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。

2 大模型应用技术分析

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。

 <p>通用大模型难以完成较复杂的金融指令</p>	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。
 <p>通用大模型难以满足金融场景的定制化需求</p>	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。
 <p>通用大模型难以满足金融领域应用的合规要求</p>	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。
 <p>通用大模型的金融专业性不足</p>	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。
 <p>通用大模型的金融情境理解能力不足</p>	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，需要对模型进行金融情境的深度训练和优化。

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。

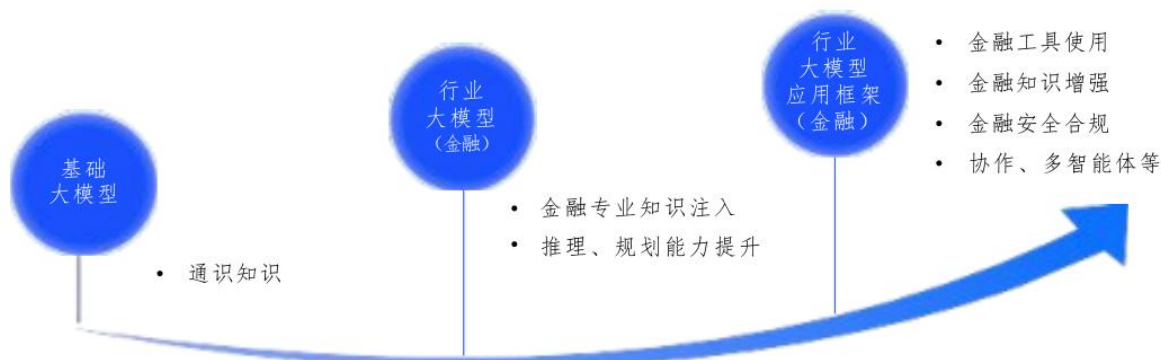


图 2-2 大模型在金融领域落地应用路线图示意

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

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透过商业采购来获取
		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

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。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。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，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。

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。

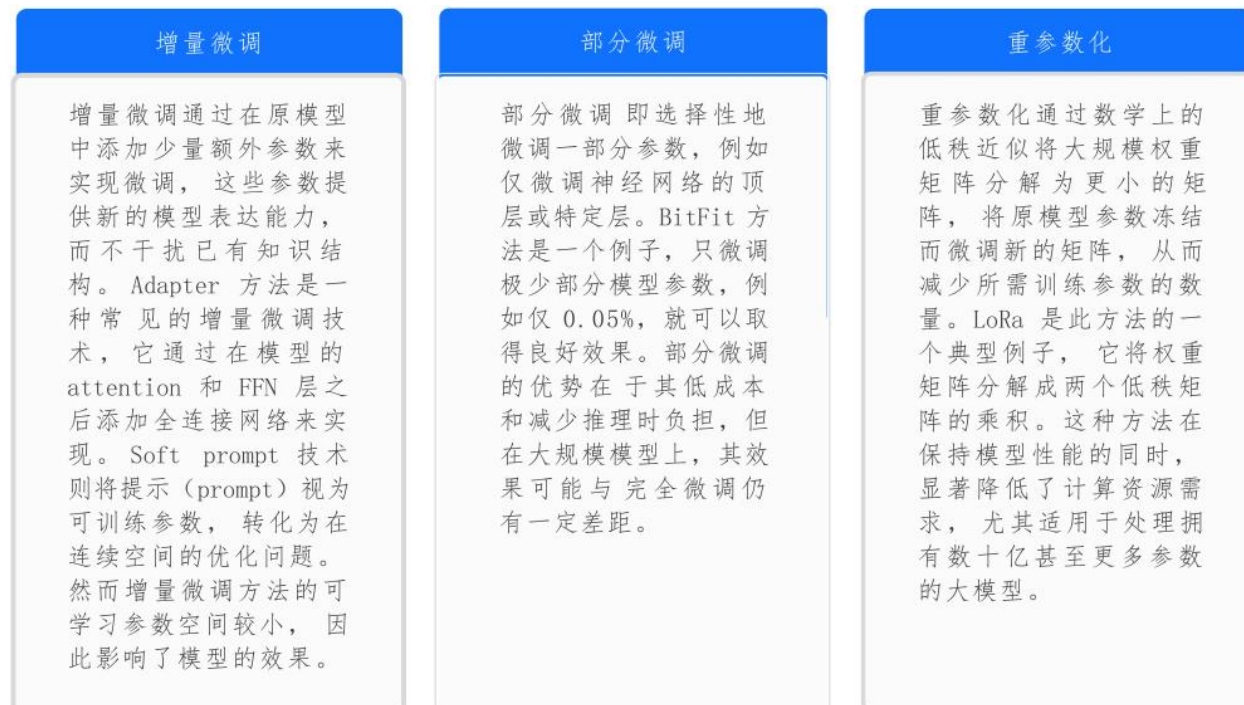


图 2-4PEFT 常见方法

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 and 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。DP0(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。

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，动态批处理和连续批处理技术被提出。



图 2-5 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 I/O 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。

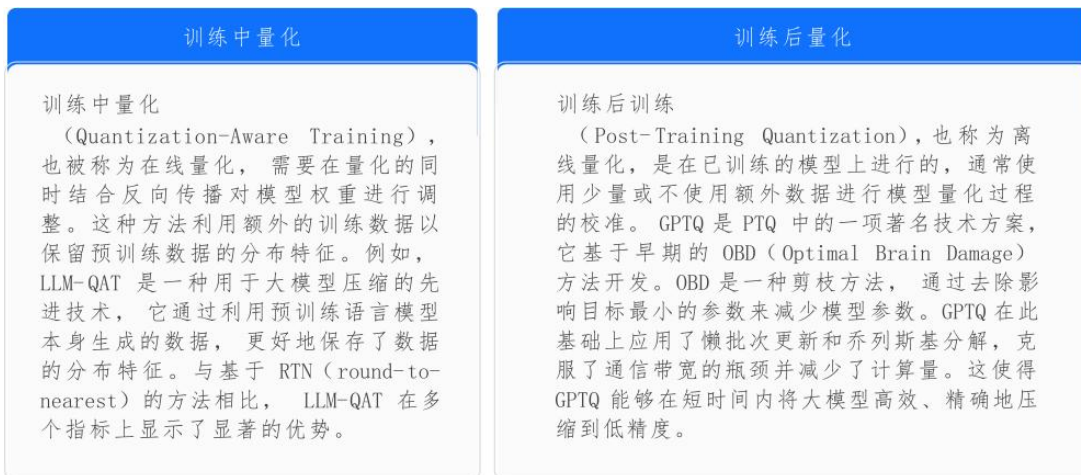


图 2-6 模型量化技术

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。

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。

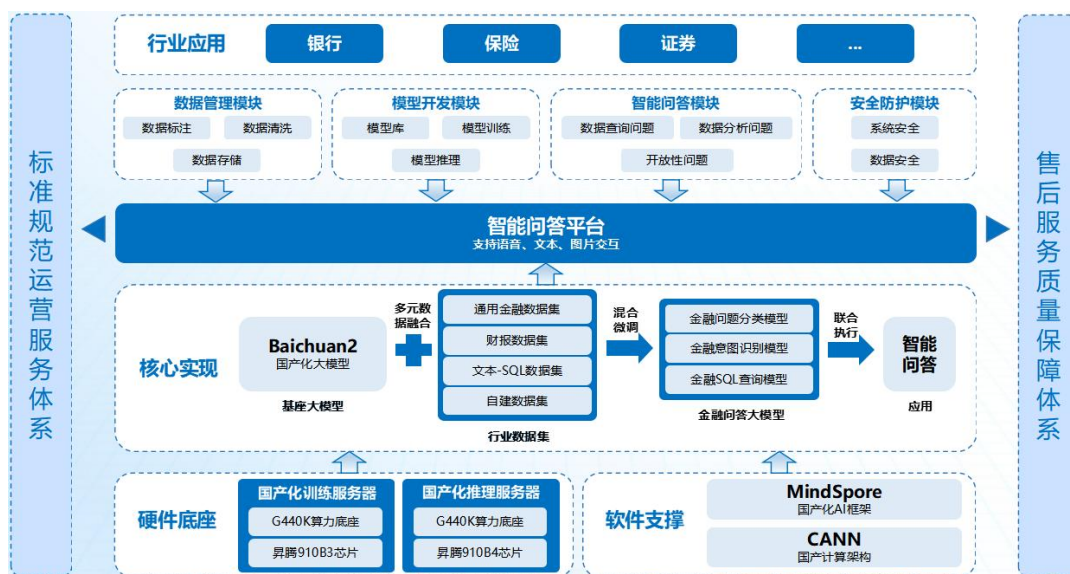
2.2.6.2 智能体推理

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。

3 金融大模型智能问答解决方案设计

3.1 行业金融大模型方案架构

金融大模型智能问答解决方案采用全栈全国产化技术，构建了一个金融问答大模型解决方案，可以极大地提升金融办公场景的智能化和效率，具体方案图如下图所示：



- ◆ 算力底座使用自研的国产化 AI 服务器，其技术完全自主可控，提供强大的算力资源，这使得智能问答系统能够处理大量的数据，执行复杂的算法，以及支持高并发的用户请求；
- ◆ 软件支撑采用国产化计算架构 CANN 和国产化 AI 框架 MindSpore。其中，CANN 是针对 AI 场景推出的异构计算架构，能充分释放昇腾 AI 处理器的计算效率。MindSpore 是面向“端-边-云”全场景设计的 AI 框架，提供动静统一的编程体验以提升算法的开发效率。两者结合可极大加速金融问答大模型的开发流程和部署效率；
- ◆ 核心实现思路是以优秀中文能力表现的 ChatGLM3 为基座模型，利用多种类金融行业数据集进行模型微调，开发出金融问答大模型簇（金融问题分类模型、金融意图识别模型、金融 SQL 查询模型），并通过这些大模型串流并行实现智能金融问答应用；
- ◆ 前端展示界面使用了一套智能问答平台系统，企业用户能通过该系统，进行数据管理、模型开发、智能问答以及安全防护等，其中问答交互支持语音、文本、图片等不同类型的输入。

此外，本方案还提供完善的标准规范运营服务体系 and 售后服务质量保障体系，可确保项目的高效运转与客户体验的持续优化。通过集成化的管理流程、个性化的服务设计、前沿技术的应用、全面的风险控制以及持续的创新驱动，实现了服务的高效响应和质量的稳定保障。

3.2 金融大模型智能问答方案功能模块介绍

数据管理模块负责处理金融行业数据。主要功能包括：

- ◆ 数据清洗：识别并删除金融数据中的错误、异常和重复的数据；
- ◆ 数据标注：对金融数据进行半自动化或手动化标注；
- ◆ 数据存储：将清洗和标注后的数据保存到数据库中，以便开发使用。

模型开发模块提供构建、训练和部署金融问答大模型所需的工具。主要功能包括：

- ◆ 模型库：模型库是一个预先构建的模型集合，包括常用的大模型。用户可以直接使用这些模型，也可以在其基础上进行定制和改进。
- ◆ 模型训练：用户仅需选择合适的损失函数、优化器以及训练策略，即可使用选好的模型和数据来开发金融大模型。
- ◆ 模型推理：用户将训练好的模型应用于新数据进行能力测试。满足应用条件的模型可用于替换智能问答模块的运行模型。

智能问答模块利用金融问答大模型来理解和回答用户的问题。主要功能包括：

- ◆ 回答数据查询问题：理解用户的数据查询请求，并从数据中检索相关信息，并生成目标结果。
- ◆ 回答数据分析问题：理解用户的数据分析请求，并从数据中统计相关信息，并生成分析结果。
- ◆ 回答开放性问题：理解用户问题意图，生成流程连贯的合理答案。

安全防护模块用于确保使用的安全性和合规性。主要功能包括：

- ◆ 系统安全：采用网络安全、身份验证、授权、访问控制和安全审计等措施，以防止未授权访问和系统攻击。
- ◆ 数据安全防护：使用数据加密、数据脱敏、备份和恢复策略，还包括对敏感数据的合规性管理和监控。

3.3 金融大模型智能问答方案创新点

(1) **全栈国产化方案**: 针对先进算力和大模型技术的封锁问题, 提出一种可定制化的国产化智能问答平台架构, 以自研的训练和推理服务器为算力底座, 结合先进异构计算架构 CANN 和优秀 AI 框架 MindSpore, 实现了金融问答大模型的高效开发和敏捷部署, 缩短了 10% 的开发时间, 减少了使用国外技术可能带来潜在安全隐患风险, 保障了系统的安全性和可靠性。

(2) **多源数据融合技术**: 针对通用大模型行业能力弱的难题, 提出一种多源数据训练的策略, 通过混合通用金融数据集、企业年报数据、文本-SQL 数据以及自建数据集的数据表征, 加强对通用知识处理、企业信息理解以及数据检索的能力, 提高了大模型在金融领域的知识覆盖率和准确性, 实现了问题回答精度近 30% 的提升。

(3) **混合微调模型方法**: 针对金融问答中大模型泛化性差、能力不均衡的问题, 提出一种基于 Pturning-v2 的混合微调策略, 采用任务划分的模型微调技术和模型联合响应技术, 实现了 6 大类金融问题的精准问答, 且所有类别问题的回答准确率均在 90% 以上。

(4) **多维度交互技术**: 提出一种多维度交互方式, 支持语音、文本、图像等输入形式, 实现了高度自由化问答应用。

3.4 金融大模型智能问答方案应用实践

基于国产化的智能问答金融大模型解决方案可在金融行业中得到广泛应用, 为银行、保险和证券等不同领域的机构带来实际效益。

3.4.1 方案核心业务功能介绍

- ◆ **智能问答功能**: 该解决方案通过精准的问题分类和意图识别, 能够快速准确地理解客户需求, 并提供个性化的服务。同时, 它还支持多轮对话和上下文理解, 能够与客户进行流畅的交流和沟通, 提供快速准确的问答服务, 以解决客户咨询和常见问题。
- ◆ **智能投顾服务**: 该解决方案能够根据金融市场的变化和趋势, 为客户提供个性化的投资建议和策略。它能够帮助客户抓住投资机会并规避风险, 实现资产的保值增值。
- ◆ **风险管理助手**: 该解决方案能够利用大数据分析和智能算法, 对客户公司的信用风险、市场风险等进行全面评估和预警。它能够帮助金融机构及时发现潜在风险并采取相应措施, 确保业务的稳健发展。

3.4.2 方案主要应用场景

- ◆ 银行业务场景：在银行领域，智能问答模块可以提供客户咨询服务，如账户管理、贷款咨询、理财产品推荐等。数据管理模块和模型开发模块支持风险评估和信用评分，帮助银行进行信贷决策。
- ◆ 保险业务场景：在保险行业，解决方案通过智能问答模块协助客户了解保险产品、理赔流程和保险条款。同时，数据管理模块帮助保险公司进行风险管理和客户数据分析。
- ◆ 证券业务场景：证券行业可以利用智能问答模块为投资者提供市场分析、投资建议和交易咨询。模型开发模块支持证券市场趋势预测和投资策略的制定。

3.4.3 方案应用成效

- ◆ 智能问答：在银行、保险和证券行业，智能问答模块实现了对客户咨询的快速响应，提升了服务体验，减少了人工客服依赖，人力成本节省 40%。
- ◆ 风险管理：智能问答大模型帮助金融机构进行风险评估和预警，提高风险管理效率，整体效率可提升 30%。
- ◆ 数字员工：在工商银行、中国银联等机构中，金融大模型充当数字员工处理业务工单，工单处理效率整体提升 50%。
- ◆ 智能投顾：金融问答大模型可快速准确分析公司财报，财报分析准确率可达 95%，显著提升投资决策水平。
- ◆ 办公辅助与培训助手：金融问答大模型可充当培训助手，简化培训流程，降低时间成本，提高培训效率。

4 大模型评测

4.1 通用大模型评测框架

随着人工智能技术的不断成熟，大语言模型在金融领域的应用蓬勃发展，大模型评测工作的重要意义得以凸显出来。测评基准对于通用大模型至关重要。通用大模型的评测框架可分为评估领域、能力维度、数据分类、题型分类、构建方式及评测方法六个层面：

评测层面	具体描述
评估领域	决定了评测框架的应用范围。不同的领域，如金融、医疗或法律，拥有不同的数据特性、任务需求和合规要求。因此，评测框架必须能够适应特定领域的唯一需求和挑战。一个优秀的领域专用评测框架，对于牵引相关领域模型迭代优化，起着至关重要的作用。
能力维度	<p>知识能力：衡量模型对广泛知识的了解程度，关注于模型在不同教育层次的学科知识掌握情况，从义务教育、高等教育以及职业教育等角度，通过对各级学科进行分类，构建完整的学科能力评测方案。</p> <p>推理能力：该维度针对模型的逻辑分析与问题解决能力。模型的推理能力不仅涉及数学计算和逻辑判断，还包括对复杂问题的因果推断和程序代码的生成与 debug 能力。</p> <p>理解能力：评估模型对于给定文本的深入理解程度，包括对文字含义的把握、主旨的抽取、语境的解读以及创意写作。评测时可以通过提供文章摘要、批判性阅读分析、以及围绕特定主题的创意写作任务来衡量模型的理解能力。</p> <p>语言能力：关注模型在理解和生成语言方面的能力，包括对字词的精准辨识、分类、含义解释以及新词创造；对语句、段落的语法结构进行解析和错误纠正；以及将一种语言翻译为另一种语言的能力。通过设计细致的语</p>

	<p>言测试，包括同义词辨析、句子改写、多语种翻译等任务，来全面评估模型的语言处理能力。</p> <p>安全能力： 确保模型的输出不仅符合技术要求，还要符合社会和道德规范，这是避免潜在风险和不当使用的关键。通过设置与实际应用场景相符的测试用例和模拟情境，可以检验模型在各种复杂环境下的安全表现。</p>
<p>数据分类</p>	<p>大模型评测集的数据来源包含网络爬虫、教科书、业务数据，以及问答社区和知识库等渠道，旨在全面衡量模型的知识掌握和实际应用能力。</p> <p>网络爬虫数据为模型提供了丰富的语言环境和现实世界的情境，包括最新的新闻、流行话题和公众讨论。教科书是权威的学术资源，它们给模型带来了正式的学科知识和概念性学习材料。</p> <p>业务数据则来源于特定行业或企业，这些数据集中于具体领域的专业知识和场景，对于评估模型在专业环境中的适用性至关重要。</p> <p>问答社区和知识库如知乎提供了用户生成的问题与答案，这些实际的交互数据可以检验模型的应答质量和问题解决能力。</p> <p>综合这些来源，评测集能够更精确地揭示模型在理解和生成语言、处理知识信息，以及为用户交互方面的实际表现。</p>
<p>题型分类</p>	<p>文本补全测试： 评估模型预测和插入缺失文本片段的能力，要求模型展现对上下文的理解并准确推断出合适的内容。</p> <p>多项选择测试： 旨在衡量模型能否在多个可能答案中选择最合适的一项，从而考验模型的知识储备，阅读理解和分析判断能力。</p> <p>文本摘要生成： 检验模型提取关键信息并有效压缩长篇文章为简洁摘要的能力，这对于评估模型的信息处理和概括能力至关重要。</p> <p>代码生成： 此类测试专注于模型理解编程语言规范并根据功能需求生成准确代码的能力，是衡量其技术应用潜力的关键。</p>

	<p>工具调用：测试模型能否正确使用特定工具或服务来完成任务，如查询数据库、调用 API 接口等，这反映了模型的实用性和交互能力。</p>
构建方式	<p>对于 PDF 格式的数据，可以采用 OCR 技术进行电子化处理，随后通过人工干预进行清洗和校正，以确保构建出高质量的评测题目。相比于可能被模型预训练过的网页文本格式试题，PDF 格式的数据更能保障评估结果的客观性，以避免数据穿越的潜在影响。</p> <p>对于未标注的教科书或专业资料，可以利用语言模型的转换功能，将这些内容转化为填空题、问答题以及选择题等形式。鉴于生成式大模型可能产生的幻觉问题，直接利用大模型生成题目可能无法确保其有效性。因此，利用教科书和专业资料作为基础，借助模型本身的语言理解能力构建评测题目是一种非常有效的方案。</p> <p>此外，专家构造的评估集也是评测工具箱中的重要组成部分。这类评估集能够有效避免数据泄露问题，并且人类专家能够创造众多独特而富有挑战性的评估数据。然而，专家构造评估集也面临规模有限、创建和更新成本高昂的局限性。</p> <p>针对业务数据的题目构建，可以通过精心设计的提示（prompt）和规则将业务数据转化成具体的评测题型，以此评估模型在实际业务环境中的适用能力。</p>
评测方法	<p>客观评估：客观评估通过量化指标来衡量模型在特定任务上的表现，是小模型时代主流的评估方法，常用的评估指标包括：准确率、F1 分数、ROUGE 指标、METEOR 分数以及 pass@k 指标等。</p> <p>主观评估：在实施大型语言模型的评估时，纯粹依赖于客观指标并不能完全捕捉到模型的语言能力及其在安全性方面的细微差别。因此，采用以人类评价者的主观感知为基础的评估方法，能够更全面地衡量模型的综合性能。主观评估则依靠人类专家根据经验和判断来进行，它涉及对模型性能的个人感知评价和比较，旨在识别模型的优势和潜在的改进空间。主观评估常考察内容的流畅度、逻辑一致性和符合标准性等因素，提供更全面和深入的评估视角，弥补了客观评估可能存在的不足，尽管如此，主观评估过程往往耗时且人力成本较高。</p>

	<p>对于人工评估，使用 GPT-4 进行评估可作为其替代方法（团队内部评估过，GPT-4 作为人工评估替代方案，与专业达标人员评估相关度高，且效率大大提升）。除了使用 GPT4 辅助评估，业界还曾以众包方式让不同的大模型进行匿名随机的对抗测评。这种评级基于国际象棋等竞技游戏中广泛使用的 Elo 评分系统（Elo 是一种计算玩家相对技能水平的方法，通过两名玩家之间的评分差异可以预测比赛的结果），在 ChatbotArena 评测基准和著名的中文通用大模型综合性评测基准 SuperCLUE 中都应用了这种评估方法。</p> <p>上下文学习与思维链：针对大型模型特有的新场景适应能力和逻辑推理能力，研究领域已发展出具有代表性的评估方法如“上下文学习”（In-Context Learning）和“思维链推理”（Chain of Thought, CoT）等。</p> <p>在 Zero-shot learning 能力的评估中，关键在于考察模型在未经特定任务训练的情况下的表现能力。模型被要求依赖于其在预训练阶段习得的知识与推理技巧，直接应对新颖任务的挑战。该评估手段突出了模型的普适性，以及其对未知场景的适应与处理能力。</p> <p>对于 Few-shot learning，评估聚焦于模型在接触有限的任务相关样例（通常 1-5 个）后的表现。此评估方法测试了模型在极少量信息支持下对新任务的快速学习与适应性，反映了模型在小样本学习环境中的预测效率。</p> <p>而 Chain of Thought (CoT) 推理的评估框架，则是要求模型在输出最终答案前，展示其一系列的中间推理步骤。这种方法不仅衡量了答案的正确性，而且深入评价了模型处理问题的逻辑和推理路径。CoT 推理尤其适用于那些需要复杂多步骤推理的问题，为评估模型的逻辑能力提供了有效途径。</p>
--	---

表 4-1 通用大模型评测框架

4.2 大模型在金融领域的评测概述

大模型在金融领域应用的评测与通用大模型评测之间存在一种深刻的关联性。首先，金融能力的评测建立在通用能力的基础之上。这些基础能力涵盖语言理解、指令执行、逻辑推理、数学计算以及内容生成等多个方面。在金融评测进行之前，模型必须要在通用评测中证明其在这些领域中的能力，确保有足够的底层支持来执行更为复杂的金融任务。在此基础之上，金融大模型评测要求模型不仅要拥有通用能力，还要具备专业的金融知识和技能。这一层次的能力扩展需要模型在理解广泛的金融概念和进行专业化推理上有所增强，类似于在通识教育基础上发展出专门的职业技能

能。而就金融领域的大模型而言，其评测意义在于能够基于完整性、针对性及区分度，全面地评估模型在处理复杂金融数据和情景时的能力。

首先，完整性是金融领域大模型评测的一个重要方面，目前的评测框架主要是针对知识广度的评估，因此大多集中在通用知识的研究上。这种评测对象过于分做的方法可能无法全面反映出参与者在面对复杂金融任务时的真实能力。除了通用能力评测外，完整的评测框架应当还包含金融通用能力评测和金融场景能力评测两大模块。对于金融通用而言，对其在金融、经济、会计和资格证书等领域表现进行测评是一种科学的框架构建方式。金融通识掌握和应用的宽度和深度，决定了能够多大程度上客观的反馈大模型的总体综合能力。以金融通识掌握深度为例，蚂蚁集团定义了“L1 识记级-15 自省级”的不同深度，用以评测大模型对金融通识的掌握应用能力。对于金融场景而言，需要评测模型能否适应不同场景下金融业务的需求，如针对股票、期货、基金、保险、证券和信托等业务的表现能力。蚂蚁集团开发了相应的评测框架，其包含认知，生成、金融知识，金融逻辑和安全合规五大金融场景模块，能够对大模型在金融场景领域的表现能力进行科学完备的评估。

其次，针对性也是金融领域大模型评测需要关注的重要方面。金融领域的特殊性包含了业务合规性、事实准确性、推理正确性、事件实时性等方面。业务合规性涉及业务适当性、数据隐私安全等多个方面，其难点在于法律、规章等官方文件高度分散，系统性梳理和构建评测集进行有效评测具有挑战性。在大模型中，事实准确性比通用大模型更为重要，因为金融业务强调严谨性，基础数据和事件的事实性对复杂推理和金融计算的可靠性至关重要。除此之外，金融领域大模型在推理正确性方面的要求极高，因为金融系统与居民、企业、机构等密切相关，推理或计算错误可能导致严重影响。对于金融业务，能够准确且实时地解读和推理计算相关政策、事件和宏观金融数据非常关键。

最后，在金融领域大模型与通用大模型的评测对比中，区分度显得尤为重要。金融场景任务评测的代表性要求评测集能覆盖并代表实际发生的金融场景任务，这包括使用真实业务数据(经处理确保合规)和经过实际金融场景任务检验的评测任务及数据。金融领域评测的专业性包括领域宽度和深度。领域宽度包括金融任务和职业资格类认证在内的广泛领域，需符合 MECE 原则并经过人类专家评审;领域深度则涵盖实际应用的深度，其主要考虑因素包括可解释性、可靠复现性和符合 MECE 原则。最后，区分度涉及与人类专家、通识基座和金融基座的比较，旨在区分金融模型与通识模型的能力差异，并为大模型提供提升指引。这包括比较金融模型与人类专家的能力差异，以及不同金融模型之间在某些维度或深度上的能力差异。

因此，对于金融大模型评测而言，一个具备完整性的评测体系应涵盖从基础知识到高级金融理论、从通用应用到特定领域应用的全方位评估。除此外，特殊性任务的设计和评估也应该得到足够的重视，以确保评估结果能够真实反映出参与者在金融特殊性任务方面的能力。总之，金融大模型的评测是一个多维度、全方位的过程，它要求模型不仅要具备强大的技术能力，还要能在实际的金融环境中安全、高效地运作。

基于上述大模型在金融领域的评测概述，与通用大模型相比，大模型在金融领域的评测也应考虑结合金融行业特征的维度包括业务合规性、事实准确性、推理正确性、事件实时性、评估覆盖广度及深度。

4.2.1 业务合规性

金融领域的业务合规性涉及广泛的法律和监管要求，包括但不限于业务适当性和数据隐私安全。业务适当性要求确保金融产品的发行方、销售者以及服务提供者遵循一系列与客户利益相关的义务。鉴于法规和条例的多样性和不断演变，构建一个能系统性整合这些复杂信息的模型，并通过合规性评估框架证实其有效性，是实现大模型合规性的关键挑战。

4.2.2 事实准确性

大模型的准确性直接决定了其推理结果的可信度。金融行业对于数据和事实的精确性有着非常高的标准，因此大模型在处理和验证基础数据及事件的真实性方面需要特别谨慎。确保模型的事实准确性不仅是对模型质量的基本要求，也是防止错误推理和决策的关键。

4.2.3 推理正确性

金融系统对模型的逻辑推理和计算的正确性有着严格的要求。大模型必须具备高度可靠的推理能力以避免给个人、企业和机构带来不利影响。因此，在评估推理正确性时大模型必须证明其在处理复杂金融问题时的准确性和稳健性。

4.2.4 事件实时性

政策/事件/宏观金融数据等时刻都在变化，对于金融业务无论个人业务或机构业务而言，能够准确的实时的对相关事件/进行解读研判，如何评估大模型应用的实时有效性，对引导大模型落地实践应用起着关键的作用。

4.2.5 评估覆盖广度和深度

金融知识的掌握程度，包括知识的广度和深度，是评估大模型能力的另一关键维度大模型应在从 L1 识记级到 15 自省级的不同层次上展现其深入理解和应用金融知识的能力。这种评估不仅揭示了模型对金融概念的掌握程度，也反映了模型在实际场景中的应用潜力。

难度	定义	内容	举例
L1	识记	概念点，答案为名词、术语的解释	<p>风险溢价是什么？</p> <p>什么是趸交？</p>
L2	关联	知识关联，知道知识点正确的子集，并能识别判定错误的部分	<p>将抵押贷款组合打包成可在市场上交易的资本市场国内工具过程被称为（）</p> <p>A. 证券化； B. 金融深化； C. 市场一体化； D. 分散化</p>
L3	掌握	推理分析，通常涉及比较/递进/简单的分析/简单计算题	<p>直接融资和间接融资分别指什么，它们的区别是什么？</p> <p>某人购买了 10 万元的终身寿险。在保险期间，不幸被一辆汽车撞死。按照有关法律规定，肇事司机应该赔偿其家属 5 万元。事后该被保险人的丈夫持单向保险公司索赔，保险公司对该案件的处理方式是（）。</p> <p>A. 赔偿 10 万元， B. 先赔偿 10 万元，然后再向肇事司机追偿 5 万元赔款， C. 赔偿 5 万元， D. 不赔，因为不属于保险责任</p>
L4	应用	推理计算，多知识点/应用计算解决问题	<p>某公司预期未来三年股利收益分别为 5 元/股，7 元/股，6 元/股，当前资本成本率 5%，股价 20 元/股，股价被高估还是低估？</p> <p>王某，男 35 岁，现投保 5 年期定期寿险一份，保险金额为 10 万元，假设死亡给付发生在期末，利率为 2%，35 岁那年的死亡率为 0.001，则王某 35 岁那年的自然保费是（）。</p> <p>A. 96 元， B. 98 元， C. 100 元， D. 102 元</p>
L5	创造	论述题，解释现有现象或综合应用解决问题	<p>你认为中国现在的股票市场是有效的吗？请用相关理论进行分析和论述。</p>

表 4-2：知识深度分级

4.2.6 业务实践性

与通用评测关注模型通用能力不同，金融领域相关评测通用需要考虑大模型在落地中的业务实践性。评测集应当依托于真实的金融业务流程，使用在实际业务生产中产生的数据(在遵守法律法规并进行必要的清洗及脱敏处理后)来构建评测案例。相比之下，仅从互联网上公开获取的数据往往缺乏必要的真实性和有效性，无法全面反映模型在实际业务中的表现。

4.2.7 中文金融测评集

下面列举国内院校以及工业界开源出的几个中文金融测评集。作为金融行业大模型评测的第一批构建单位，从不同角度切入及不断完善该领域的评测体系。

金融评测集	领域	简介	发行方	语言	评估题型	评分方式
PIXIU	金融	包括 5 类任务、9 个数据集。任务包括金融情感分析、新闻标题分类、NER、QA、股价走势预测。	武汉大学、中山大学、云南大学、NYU、四川大学、西安交大、佛罗里达大学	英文	综合	综合白盒
FinEval	金融	是一个包含高质量多项选择题的集合，涵盖金融、经济、会计和证书等领域。它包括 4,661 个问题，涵盖了 34 个不同的学术科目。	上海财经大学	中文	多选	Acc白盒

表 4-3 中文金融测评集

4.3 大模型在金融领域的评测实践

下面通过两个具体的案例，来展示评测的具体步骤。第一个是上海财经大学的 FinEval 金融评测集，展示学术界是如何构建金融评测的维度，第二个是蚂蚁集团的 Fin-Eva 金融评测集，展示工业界如何对金融业务进行评测以及评测的工业框架。

4.3.1 上财 FinEval 金融数据集

为了辅助开发者更好的研发中文大模型，财大团队耗时三个月的时间，构造一个中文的，有足够区分度的，多学科的评测基准，命名为 FinEval。FinEval 是一个高质量的多项选择题的集合，涵盖金融，经济，会计和证书等四大领域。它包括 4661 个问题，涵盖了 34 个不同的学科。从选题的角度来看，金融领域和会计领域分别包括 10 个不同的科目，经济领域和证书领域分布包括 7 个科目。在数据集分割方面，开发集、验证集、测试集和总集各包含 34 个主题，分别是由 170、1151、3340 和 4661 个问题组成。数据源主要基于相关领域权威性考试各类真题和模拟题对知识大纲的要求，由上海财经大学统计与管理学院张立文副教授课题组牵头，金融学院闵敏教授及其他各学院老师协助完成，所有数据均为原创，这保证了数据源的准确性和权威性。就评估方法而言，FinEval 采用了一系列提示类型，包括 zero-shot 和 few-shot，以及仅回答和思维链提示，这确保模型性能评估的专业性和先进性。该团队向外界公开评测数据和评测代码，未来将持续进行迭代更新，并提供开放性的平台化评测服务，旨在为行业提供综合评估解决方案。

项目地址:<https://github.com/SUFE-AIFLM-Lab/FinEval>

四大领域	详细介绍
金融领域	为专业人员提供了做出明智金融决策和导航全球金融环境所需的关键技能。
经济领域	着重于理解国家和全球经济系统，使个人能够分析经济趋势，并有效地为该领域作出贡献。
会计领域	提供全面的财务管理和合规知识，塑造专业人员在财务决策制定和风险管理方面的专业技能。
证书领域	包含精算、会计和金融等领域的证书考试，用于验证专业人员的知识和技能，增强职业前景和行业认可。

表 4-4 FinEval 评估的四大领域

4.3.2 蚂蚁大模型评测集 Fin-Eva

蚂蚁集团为大模型评测专门设计了 Fin-Eva 金融评测集，其设计目的不仅是一个金融评估数据集，更能帮助加速领域的发展，拓展大模型应用的边界。Fin-Eva 涵盖金融认知、领域知识、金融逻辑、内容生成以及安全合规五大类能力 33 个子维度共 8446 个测评题，题目类型为单选题。Fin-Eva 覆盖财富管理、保险、投资研究等多个金融领域，数据来源包括蚂蚁各业务领域、开源数据、模型蒸馏，经过数据脱敏、文本聚类语料精筛等处理过程后，结合金融领域专家的深度加工最终构建而成。目前对外开放评测数据及评测代码，未来持续迭代并开放平台化评测托管服务，为行业提供一站式评估。项目地址:<https://github.com/SUFE-AIFLM-Lab/FinEval>



五大能力	评估维度
金融认知类	考察模型金融文本的理解和提取能力
领域知识类	考察模型是否具备全面的金融领域知识，以及能否通过专业能力考试
金融逻辑类	考察模型是否具备完成复杂金融任务的推理和计算能力
内容生成类	考察模型总结和生成专业金融文本的能力
安全合规类	考察模型能否辨别金融领域的安全和合规问题

表 4-5 Fin-Eva 评估的五大能力

4.4 金融大模型测试结果

烽火通信金融大模型智能问答系统具体使用测试结果如下：

烽火通信金融大模型问答系统

历史问题列表

- 我想知道2023年烽火通信营业外支出是多少元?
- 请提供高德红外公司2023年的企业研发经费与营业收入比值
- 请提供九州通公司2023年的企业研发经费与营业收入比值
- 请提供九州通公司2019年的企业研发经费与营业收入比值

智能问答

欢迎使用金融大模型

请输入问题...

提交

问题所属分类

A 公司基本信息类问题：包含股票代码简称，公司名称，外文名称，法定代表人，注册地址，办公地址，公司网址网站，电子信箱等。

B 公司员工信息类问题：包含员工人数，员工专业，员工类别，员工教育程度等。

C 财务报表类问题：包含资产负债表，现金流量表，利润表汇总存在的字段，包含费用，资产，金额，收入等。

D 计算类问题：无法从年报中直接获得，需要根据计算公式获得，包括增长率、比率、比重、占比等。

结果注释表

字母编号	分类子选项	归属问题类别
A	公司基本信息类问题	数据查询问题
B	公司员工信息类问题	
C	财务报表类问题	数据分析问题
D	计算类问题	
E	统计类问题	
F	开放性问题	
		开放性问题

烽火通信金融大模型问答系统

问题:

2023年烽火通信营业外支出是多少元?

"问题"属于以下哪性问题类别:

A 公司基本信息类问题：包含股票代码简称，公司名称，外文名称，法定代表人，注册地址，办公地址，公司网址网站，电子信箱等。

B 公司员工信息类问题：包含员工人数，员工专业，员工类别，员工教育程度等。

C 财务报表类问题：包含资产负债表，现金流量表，利润表汇总存在的字段，包含费用，资产，金额，收入等。

D 计算类问题：无法从年报中直接获得，需要根据计算公式获得，包括增长率、比率、比重、占比等。

E 统计类问题：需要在数据表中查找数据，过滤、排序后得出结论。

F 开放性问题：包括但不限于开放式问答、分析推理、计算等。

答案:

烽火通信2023年营业外支出为184741.52元。烽火通信2023年营业外支出占营业收入的0.04%。根据2023年年报披露，烽火通信2023年营业收入为459494.22元。

数据分析问题结果

烽火通信金融大模型问答系统

问题:

什么条件下会有研发?

"问题"属于以下哪性问题类别:

A 公司基本信息类问题：包含股票代码简称，公司名称，外文名称，法定代表人，注册地址，办公地址，公司网址网站，电子信箱等。

B 公司员工信息类问题：包含员工人数，员工专业，员工类别，员工教育程度等。

C 财务报表类问题：包含资产负债表，现金流量表，利润表汇总存在的字段，包含费用，资产，金额，收入等。

D 计算类问题：无法从年报中直接获得，需要根据计算公式获得，包括增长率、比率、比重、占比等。

E 统计类问题：需要在数据表中查找数据，过滤、排序后得出结论。

F 开放性问题：包括但不限于开放式问答、分析推理、计算等。

答案:

研发是指企业为了获得新的产品或服务，而进行的技术创新和知识积累的过程。研发活动通常涉及大量的资金投入、人力资源投入、设备投入、材料投入等。研发活动的目的是为了提高企业的核心竞争力，实现可持续发展。

开放性问题结果

烽火通信金融大模型问答系统

问题:

烽火通信2023年营业外支出占比多少元?

"问题"属于以下哪性问题类别:

A 公司基本信息类问题：包含股票代码简称，公司名称，外文名称，法定代表人，注册地址，办公地址，公司网址网站，电子信箱等。

B 公司员工信息类问题：包含员工人数，员工专业，员工类别，员工教育程度等。

C 财务报表类问题：包含资产负债表，现金流量表，利润表汇总存在的字段，包含费用，资产，金额，收入等。

D 计算类问题：无法从年报中直接获得，需要根据计算公式获得，包括增长率、比率、比重、占比等。

E 统计类问题：需要在数据表中查找数据，过滤、排序后得出结论。

F 开放性问题：包括但不限于开放式问答、分析推理、计算等。

答案:

2023年烽火通信营业外支出为184741.52元。

数据查询问题结果