

大语言模型与经济金融文本分析： 基本原理、应用场景与研究展望

林建浩, 孙乐轩

(中山大学岭南学院, 广州 510275)

摘要 大语言模型 (LLMs) 拥有强大的自然语言处理能力, 本文系统梳理了相关领域的前沿文献, 并详细论述其为经济金融领域的文本分析带来的新研究机会. 首先, 本文介绍了 GPT 和 BERT 这两类最具代表性的 LLMs 及一系列经济金融领域专用的 LLMs, 并阐述应用 LLMs 分析文本数据的基本原理. 其次, 本文从两个方面归纳了 LLMs 在经济金融文本分析研究中的应用场景: 一是 LLMs 在计算文本相似度、提取文本向量进行预测、文本数据识别与分类、构建特定领域词典、主题建模与分析、文本情绪分析等传统文本分析场景中所呈现的显著优势; 二是 LLMs 拥有强大的人类对齐能力, 由此开创了全新的应用场景, 即作为经济代理人, 模拟人类生成对于文本的信念或预期并进行经济决策. 最后, 本文总结了 LLMs 在开拓经济金融文本分析研究的新范式过程中面临的局限性以及现有研究不足, 并针对这些问题可能催生的新研究议题进行展望.

关键词 大语言模型; 自然语言处理; 文本分析; 经济与金融

Text Analysis in Economics and Finance with Large Language Models: Fundamentals, Applications, and Future Prospects

LIN Jianhao, SUN Lexuan

(Lingnan College, Sun Yat-sen University, Guangzhou 510275, China)

Abstract Large language models (LLMs) have powerful natural language processing capabilities. In this paper, we systematically review the recent literature in this field and highlight the new research opportunities that LLMs bring to text analysis

收稿日期: 2024-07-25

基金项目: 国家社会科学基金 (22AZD121); 国家自然科学基金 (71991474, 72073148, 72273156, 72303258)

Supported by National Social Science Foundation of China (22AZD121); National Natural Science Foundation of China (71991474, 72073148, 72273156, 72303258)

作者简介: 林建浩, 博士, 中山大学岭南学院教授, 博士生导师, 研究方向: 宏观金融, 计量经济学, E-mail: linjh3@mail.sysu.edu.cn; 通信作者: 孙乐轩, 博士研究生, 研究方向: 宏观金融, 人工智能经济学, 文本数据计量, E-mail: sunlx7@mail2.sysu.edu.cn.

in economics and finance. First, we introduce GPT and BERT, the two most representative LLMs, as well as a number of LLMs developed specifically for economic and financial applications. Additionally, we also elaborate on the fundamental principles behind applying LLMs for text data analysis. Second, we summarize the applications of LLMs in economic and financial text analysis from two perspectives. On the one hand, we highlight the significant advantages of LLMs in traditional text analysis scenarios, such as calculating text similarity, extracting text vectors for prediction, text data identification and classification, building domain-specific dictionaries, topic modeling and analysis, and text sentiment analysis. On the other hand, LLMs have strong human alignment capabilities, thus opening up entirely new application scenarios, i.e., acting as economic agents that simulate humans in generating beliefs or expectations about texts and making economic decisions. Finally, we summarize the limitations and existing research gaps that LLMs face in pioneering new paradigms of economic and financial text analysis research, and discuss potential new research topics that may arise from these issues.

Keywords large language models; natural language processing; text analysis; economics and finance

1 引言

得益于近年来文本分析技术的快速发展和文本数据可得性的提高, 经济学和金融学中的数据分析已广泛地从传统的量化数据拓展到了文本等非结构化数据, 为相关研究提供了全新的视角和工具来探索经济金融现象、预测市场动态以及评估政策影响 (沈艳等, 2019; 姚加权等, 2020; 洪永淼和汪寿阳, 2021). 如图 1 所示, 目前文本分析技术已经历了长达七十余年共三个发展阶段并逐渐走向成熟 (车万翔等, 2021; Zhao et al., 2023): 1) 早期探索阶段 (1950s–1990s): 该阶段主要使用基于理性主义的规则方法, 利用专家总结的符号逻辑知识来处理小规模的自然语言现象. 但这类方法主观性较强、错误率较高且仅能用于小规模文本分析. 2) 统计语言模型阶段 (1990s–2000s): n-gram 模型 (Brown et al., 1992)、概率潜在语义分析 (probabilistic latent semantic analysis, PLSA) 模型 (Hofmann, 2000) 和潜在狄利克雷分配 (latent dirichlet allocation, LDA) 模型 (Blei et al., 2003) 等在该阶段得到快速发展和应用. 例如, Larsen and Thorsrud (2019) 利用 LDA 模型提取报纸文本数据中蕴含的新闻主题, 并研究它们在预测和解释经济波动方面的作用. 范小云等 (2022) 基于 n-gram 模型的思想开发了一种新的中文金融文本情绪词典. 这类模型虽能有效避免主观性, 但所面临的“维数灾难” (curse of dimensionality) 问题严重限制了其对文本的处理规模与速度, 且仅能学习词汇间的共现关系而忽略了语序信息. 3) 神经语言模型阶段 (2000s–2017): 在这一阶段, Word2Vec 等词嵌入 (word embedding) 技术广泛应用 (Mikolov et al., 2013), 通过将文本表示为低维稠密向量从而避免了维数灾难. 例如, 姜富伟等 (2021) 借助 Word2Vec 算法扩充构建了一个全新的中文金融情感词典. 何勇等 (2023) 基于因子增广回归与深度神经网络模型, 实现了从财经新闻文本等另类数据中提取交易信号. 但这类技术对上下文信息的建模能力仍然较弱, 难以捕捉语序信息与文本间的长距离依赖关系. 2017 年谷歌提出的 Transformer 架

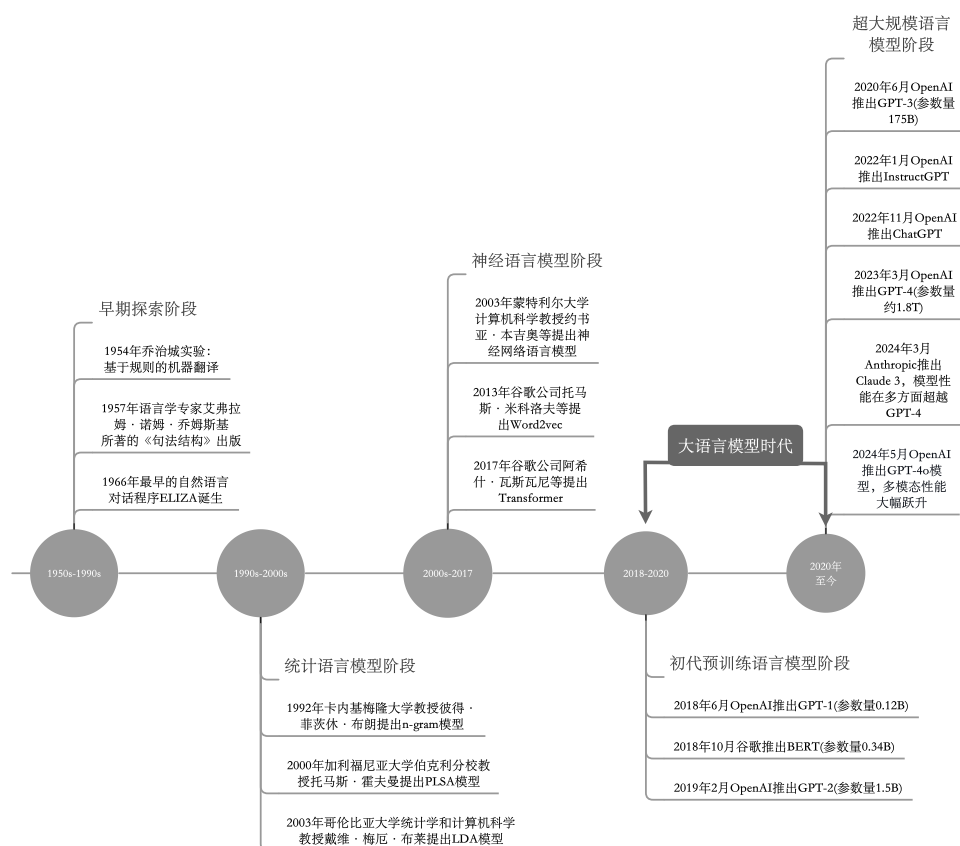


图1 文本分析技术已进入大语言模型时代

构 (Vaswani et al., 2017) 虽能有效解决这些问题, 但由于模型参数规模和训练数据规模较小, 故仍难以在下游文本分析任务中取得良好效果。

目前, 经济和金融文本数据逐渐多元化和海量, 常见数据包括企业的年报、招聘广告、专利文本、IPO 招股说明书, 经济、金融领域重要人物演讲, 政府机构定期和不定期发布的各类政策和信息, 以及个人或机构在推特、微博、论坛帖子、购物平台、微信公众号等社交媒体发布的海量信息。这些文本具有规模庞大、语义复杂和形式多变等特点, 对文本分析技术的有效性提出了新挑战。前述三个发展阶段中的文本分析方法虽已在经济和金融研究中得到广泛应用, 但由于维数灾难、忽略语序及上下文信息和模型规模较小等局限性, 这些传统方法也逐渐难以满足该领域对于文本数据处理的需求。我们亟需一种技术能够克服传统方法的不足, 同时能够尽量模拟人类对文本进行阅读、处理、甚至形成自己的判断, 进而大幅提升文本分析的效果, 挖掘文本数据中蕴含的更多信息。大语言模型 (large language models, LLMs) 技术正是在这一背景下应运而生的。

2018 年 6 月, Open AI 公司的科研技术团队发表了一篇标题为《通过生成式预训练提高语言理解能力》¹ 的论文 (Radford et al., 2018), 提出了一种基于多层 Transformer 和大

¹英文题目为: Improving Language Understanding by Generative Pre-Training.

规模无标注文本训练得到的生成式预训练 (generative pre-trained transformer, GPT) 模型, 用于提升自然语言理解任务的效果, 正式将文本分析技术带入第四个发展阶段 —— 大语言模型时代 (如图 1 所示). 大语言模型时代标志着文本分析正在利用更大规模的文本数据以及更深层的神经网络模型学习更丰富的文本语义表示, 开启了基于“通用文本预训练 + 特定任务微调 (fine-tuning)”² 的自然语言处理新范式 (车万翔等, 2021; 赵鑫等, 2024). 随后, 一系列结构与功能类似的大语言模型层出不穷, 其中最具代表性且应用最为广泛的两类模型是 2018 年 10 月由 Google 公司研究人员推出的 BERT (bidirectional encoder representations from transformers) 模型 (Devlin et al., 2018) 和由 OpenAI 公司开发的 GPT 系列模型及相关的衍生产品³. 特别地, OpenAI 公司在 2020 年发布的 GPT-3 模型实现了模型参数规模的飞跃式提升, 是 GPT-2 的 100 余倍 (达到了 175B 的规模) (Floridi and Chiriatti, 2020), 从而进一步展现出一些初代预训练语言模型所不具备的“涌现能力” (emergent ability)⁴, 标志着“超大规模语言模型”的诞生. 借助 Transformer 架构的关键技术和大规模数据的预训练及特定任务微调, LLMs 不仅克服了传统方法的局限性, 还拥有了一些传统方法所无法具备的优势. 例如, LLMs 能够充分结合输入文本的上下文信息, 能够非常灵活地处理各种类型的文本, 且能够模仿人类阅读文本并生成相应的信念、预期或决策.

由于其前所未有的优势和出色的分析效果, LLMs 已逐渐应用于各类文本数据的处理及分析中. 特别地, 在经济学和金融学领域内, 近期涌现出了大量利用以 GPT 和 BERT 为代表的 LLMs 进行文本数据分析的前沿研究 (Eisfeldt and Schubert, 2024; 姜富伟等, 2024). 目前已有综述论文总结了 LLMs 在经济或金融领域的应用, 但主要关注应用案例或除文本分析外其他方面的应用. 例如, Li et al. (2023b) 更关注如何通过技术手段训练得到适用于处理

²通用文本预训练是指在大规模未标注的通用文本数据集上训练一个高容量的语言模型, 使其学习更加丰富的语法、语义和上下文信息; 特定任务微调是指在预训练好的模型基础上, 通过在下游特定任务上进行有监督的微调, 以使模型更好地适应该任务. 在该过程中, 模型通过有标注数据学习在特定任务上进行准确地预测或分类. “通用文本预训练 + 特定任务微调”的模式可类比于学生在小学、初中和高中阶段学习语文、数学、英语、物理、化学和历史等基础学科知识, 夯实基础并构建初步的知识体系 (预训练阶段). 而当学生进入大学后, 则将根据选择的专业 (目标任务) 学习某领域更深层次、更专业化的知识 (微调阶段), 并在之后应用于未来的工作中 (处理下游任务).

³2018 年, OpenAI 公司发布了首款 GPT 模型, 即 GPT-1, 其采用“通用文本预训练 + 特定任务微调”的方法解决各类下游任务. 之后, GPT-2 和 GPT-3 通过增加预训练数据和参数规模, 大幅提升了模型性能. 建立在 GPT-3 基础上, OpenAI 又通过各类技术进一步提升了模型性能, 推出了功能强大的 GPT-3.5 模型. 2022 年 11 月, ChatGPT 震撼问世, 其能够通过与人进行交互的形式解决多种任务, 并通过网络 API 功能为用户和开发者提供语言模型的便捷体验. 2023 年 3 月, OpenAI 发布了具有里程碑意义的 GPT-4 模型 (Achiam et al., 2024), 将该类模型性能推向新高度, 并进一步引入了多模态功能, 衍生得到 GPT-4V 模型以及实现实时文本语音视频交互的 GPT-4o 模型.

⁴涌现能力 (emergent ability): 如果一个系统的属性不是其任何基本元素的属性, 那么它就是涌现的 (Wei et al., 2023a). 例如蚂蚁汇聚成群时会展现出一种不可思议的“智能” (如自动发现从蚁群到达食物的最短路径), 而这种能力是每只蚂蚁所无法具备的. 而本文的涌现能力是指当模型扩展到一定规模时, 模型的特定任务性能突然出现显著跃升的趋势, 远超过随机水平. 大语言模型的三种典型涌现能力有: 上下文学习 (in-context learning, ICL) 能力、指令遵循 (Instruction Following) 能力以及逐步推理 (step-by-step reasoning) 能力 (赵鑫等, 2024).

金融任务的 LLMs, 而不是如何利用 LLMs 分析金融文本数据. Korinek (2023) 则重点关注数十个与 GPT-4 的对话案例, 从构思和反馈、写作、背景研究、数据分析、写代码和数学推导等六个角度展现了 LLMs 如何帮助经济学家改进研究工作, 其中粗略涉及对文本数据进行分析, 但缺乏对研究文献的系统归纳. Nie et al. (2023) 重点介绍了 LLMs 在金融领域应用的最新进展, 包括语言任务、情绪分析、金融时间序列分析、金融推理和大模型智能体系统等, 但对与文本分析相关的研究仅简略提及, 并未对其不同应用场景下的研究文献进行详细总结. 李春涛等 (2024) 则侧重介绍 Stata 集成命令 chatgpt 的具体应用案例, 缺乏对各类经济金融大模型进行梳理和归纳.

综上, 目前尚缺少文献对 LLMs 在经济金融文本分析中的应用研究进行系统性地归纳与评述, 也很少有文献在研究综述的基础上探讨这一重要研究方向的争议与局限性以及未来研究方向的展望. 因此, 本文试图从以下三个方面来填补现有文献的空白. 首先, 本文对这一研究领域中亟待梳理的知识体系和现有前沿研究进行了系统地归纳总结, 更加侧重介绍经济金融领域的专用大模型以及各类文本分析场景. 具体而言, 本文简要介绍了两类最具代表性的 LLMs — GPT 和 BERT, 并整理归纳了一系列经济金融领域专用的 LLMs. 同时, 本文通过对 LLMs 分析文本数据的基本原理进行简要介绍, 厘清了 LLMs 的基本架构与重要技术, 并阐述了将 LLMs 应用于经济金融领域文本分析的原因. 之后, 本文通过梳理现有前沿研究, 从两个方面归纳了 LLMs 在经济金融文本分析中的应用场景: 第一个方面是 LLMs 在传统文本分析场景中的应用与优势, 包括六个应用场景: 1) 利用 LLMs 计算文本相似度; 2) 利用 LLMs 提取文本向量进行预测; 3) 利用 LLMs 进行文本数据识别与分类; 4) 利用 LLMs 构建特定领域词典; 5) 利用 LLMs 进行主题建模与分析; 6) 利用 LLMs 分析文本情绪, 如分析新闻媒体情绪以预测股票收益、分析管理层语调、量化央行政策沟通等. 在这六个场景中, LLMs 通常表现出显著优于传统方法的分析效果, 展现出强大的应用潜力. 第二个方面是 LLMs 作为经济代理人所开创的全新场景, 即凭借 LLMs 强大的人类对齐 (Human Alignment) 能力, 模拟人类生成对于文本的信念或预期 (generating beliefs with LLMs).

其次, 本文进一步对该领域中还未形成广泛共识的问题及争议进行深入探讨, 以指明未来研究所面临的风险与挑战. 一些研究对 LLMs 在经济金融文本分析中的应用持积极态度, 认为经济金融系统是一个超高维动态复杂系统, 大语言模型能帮助领域内研究者以更精细的分辨率来分析系统中的文本数据, 进而更精确地捕捉经济金融系统的动态和规律 (Korinek, 2023; 汪寿阳等, 2023). 然而, 另有一些研究对 LLMs 的文本分析能力表示质疑, 并对其在经济金融领域的应用表示担忧. LLMs 饱受质疑的一点就是其可复现性较差, 输出结果很大程度依赖于人类输入的提示 (Prompts), 且即便输入的提示完全相同, LLMs 不同次运行的输出结果也可能有差异, 这一局限性将极大地挑战经济学实证研究的可信性, 并阻碍其在经济金融领域的广泛应用 (Ash and Hansen, 2023). 此外, LLMs 的“黑箱”问题所导致的难以解释性、“幻觉 (Hallucination)”问题 (Ji et al., 2022; Yao et al., 2024) 所产生的错误结果、“欺骗能力” (Deception Abilities) 所导致的误导性 (Hagendorff, 2024)、性别刻板印象等信念偏见 (Chen et al., 2024) 以及其他一系列局限性都可能会对 LLMs 的文本分析效果产生影响. 本文将在对文献梳理的基础上进一步讨论这些关于 LLMs 局限性的争议.

最后, 本文针对这一极具开创性、想象空间和发展前景的领域中存在的诸多争议和挑战提出可能的解决策略, 并在梳理现有研究不足的基础上进一步明晰未来可能的研究方向. 作为最先进的自然语言处理工具, LLMs 将对教育、科研工作者以及学生对待数据科学的方式产生重大影响 (Tu et al., 2024; 陈晓红等, 2024). 同时, 其不仅会改变经济主体行为与经济运行方式, 还将对经济学未来的研究范式产生深远影响 (洪永淼和汪寿阳, 2023, 2024; 程兵, 2023). 一方面, 本文将针对关于 LLMs 局限性的争议来探讨未来研究应如何改进或规避这些问题. 另一方面, 本文将从现有前沿研究的不足出发探讨未来研究应关注哪些重要问题与方向, 并归纳潜在的研究机会与策略.

2 模型及基本原理

目前主流的 LLMs 几乎都是基于 Transformer 模型框架设计的. Transformer 是由多层多头自注意力 (Multi-head Self-attention) 模块堆叠而成的深度神经网络模型 (Vaswani et al., 2017). 初始的 Transformer 模型包括编码器和解码器两部分, 但这两部分均可以独立使用, 例如基于解码器架构的 GPT 模型和基于编码器架构的 BERT 模型. LLMs 的特点是使用了更长的向量维度和更深的模型层次, 因此具有规模更加庞大的模型参数, 但并没有对 Transformer 原本的结构与配置做出较大改变 (车万翔等, 2021; 赵鑫等, 2024). 在此基础上, 研究者开发了各类 LLMs, 如 GPT、BERT、BLOOM 和 LLaMA 等, 并通过进一步的模型特化得到了一系列专用于经济金融领域的 LLMs.

2.1 GPT 和 BERT

初始的 GPT 模型是一个基于 12 层 Transformer 解码器架构的单向语言模型, 即从左至右按顺序对输入的文本序列进行建模, 并使用无监督学习技术在大量文本数据上进行预训练. GPT 的预训练任务是预测下一个词, 需要将当前词之后的词全部遮盖 (Mask), 只能根据当前词及其之前的词来对下一个词进行预测 (Radford et al., 2018). 这意味着 GPT 更适合生成下一个单词或句子的任务 (生成式任务), 如生成自然语言对话. OpenAI 公司为用户提供了两种途径访问 GPT 模型, 一种是网页版的 ChatGPT, 另一种是 OpenAI 提供的一系列应用程序接口 (API)⁵. 前者主要用于日常聊天与实时交互, 而后者则常用于搭建应用程序、大规模数据处理和科学研究.

在使用 GPT 这类生成式 LLMs 时, 人类用户需要向其发出自然语言指令, 这些指令被称为提示 (Prompts). 这些指令的格式和内容可能会影响模型的输出结果. 为了获得理想的输出结果, 用户通常需要尝试不同的指令方式, 由此产生了一种被称为“提示工程 (Prompt Engineering)”的关键技术. 该技术通过设计精良的提示来引导 LLMs 生成所需的输出结果, 其目的是提高特定任务的输出质量, 这在优化 LLMs 的性能方面至关重要. 赵鑫等 (2024) 进一步总结了提示设计的四项基本原则, 以便研究者设计更为有效的任务提示: 1) 清晰地表达任务目标; 2) 分解为简单且详细的子任务; 3) 提供少样本示例; 4) 采用模型友好的提示格式.

⁵读者可查阅该网址以获取 OpenAI API 的详细介绍与教程:
<https://platform.openai.com/docs/api-reference>.

BERT 模型是一个基于多层 Transformer 编码器的双向语言模型, 它并不像 GPT 一样从左至右按顺序对输入文本进行单向建模, 而是实现了对文本的双向建模, 因此可以更好地捕捉句子中的上下文信息 (Devlin et al., 2018). BERT 模型通过不同的预训练任务在无标记数据上对模型进行训练. 该过程具体包括两个基本预训练任务: 掩码语言模型 (masked language model, MLM) 和下一句预测 (next sentence prediction, NSP).

MLM 任务类似于完形填空任务, 它会随机地遮盖文本中的一些词, 并要求模型预测这些被遮盖的词. 通过反复运行该步骤进行训练, 模型可以学习得到词语的上下文特征、语法结构及句法特征等信息. NSP 任务会将两个句子 A 和 B 输入模型, 并要求模型判断句子 B 是否是句子 A 的下一句, 通过反复的训练, 模型就可以学习得到句子之间的语义关系. 这意味着 BERT 更擅长处理需要理解整个句子或段落并对其进行分类或标注的任务 (判别式任务).

综上所述, 本文将 GPT 模型和 BERT 模型的关键特点与主要优势总结归纳于表 1 中.

表 1 GPT 与 BERT 比较

关键点/ 主要优势	GPT	BERT
架构	基于 12 层 Transformer 解码器的单向语言模型	基于多层 Transformer 编码器的双向语言模型
预训练任务	预测下一个词	掩码语言模型 (MLM) 和下一句预测 (NSP)
上下文处理	单向建模, 只能从左到右按顺序处理文本	双向建模, 能够捕捉句子的上下文信息
提示工程	通过精心设计的提示优化输出结果	通常不依赖提示工程, 更多依赖上下文信息
访问方式	OpenAI 网页版 ChatGPT 或 API	通常通过 API 或部署在本地使用
主要优势	更适合生成式任务 (如文本生成、自然语言对话、写作辅助等)	更擅长判别式任务 (如文本分类、标注、命名实体识别等)

2.2 经济金融领域专用的 LLMs

尽管 GPT 和 BERT 等 LLMs 可以获得解决各种文本分析任务的通用能力, 但越来越多的研究表明 LLMs 的能力可以根据特定目标进一步微调, 且经过微调的 LLMs 在特定领域内文本语料上的表现更好 (Li et al., 2023b; Nie et al., 2023). 如表 2 所示, 目前已有大量研究利用经济和金融领域的标注数据对 GPT 和 BERT 等各类 LLMs 进行训练或微调⁶, 得到了适用于经济和金融领域特定任务的 LLMs.

首先是以 OpenAI 的 GPT 模型为基准开发得到的 LLMs. 这一系列的研究并不多, 且主要在 2024 年之后才集中出现, 主要原因可能是 GPT-3.5 及之后的系列模型均未开源且价格较为昂贵, 不利于下游应用和开发. Ploutos 是一个根据 GPT-4 模型微调得到的 LLM 框架, 能够融合文本和数字信息来预测股票走势, 在预测准确性和可解释性方面均优于现有方法 (Tong et al., 2024). FinSentGPT 模型是 Ardekani et al. (2024) 在 ADA GPT-3 模型的基础上利用美国经济新闻文本数据和欧洲中央银行货币政策公告的多语言数据集微调得到

⁶以下网址分别介绍了如何进行 GPT 模型和 BERT 模型的微调:
<https://platform.openai.com/docs/guides/fine-tuning> (GPT 微调),
https://www.tensorflow.org/tfmodels/nlp/fine_tune_bert (BERT 微调).

表 2 经济金融领域专用的 LLMs 概览

发布时间	GPT	BERT	BLOOM	LLaMA	其他
2024	FinSentGPT, Ploutos	BERT-Jiang (中文)		LLaMA-2-Econ, GreedLlama, FinLlama	SilverSight (中 文), FinTral
2023		EconBERTa, Cen- tralBankRoBERTa, FinBERT-23	Xuan Yuan 2.0 (中文), BloombergGPT	InvestLM, Instruct-FinGPT, FinMA, FIN-LLAMA, FinGPT	DISC-FinLLM (中文), CFGPT (中文), FinVis-GPT
2022		ESGBERT			
2021		Mengzi-BERTbase- fin (中文), FinBERT-21			
2020		FinBERT-20			
2019		FinBERT-19			

注: 表格中标注“(中文)”表示相应的模型是使用中文语料库训练得到的中文金融大语言模型, 其余大模型均以英文语料库为主进行训练.

的, 其在情绪分类的准确性上超越了传统的机器学习模型, 并且与 FiGAS 模型 (Barbaglia et al., 2023) 性能相当, 且能够实现跨语言情绪分析.

由于 BERT 模型良好的开源性和优异的自然语言处理性能, 以 BERT 为基准开发经济金融 LLMs 的研究非常丰富. 例如, 一些研究基于各类不同金融文本数据集并采用不同微调策略得到了一系列 FinBERT 模型 (Araci, 2019; Yang et al., 2020; Liu et al., 2021; Huang et al., 2023a), 这些 FinBERT 模型在金融文本情绪分析、金融信息检索、命名实体识别、金融文本分类和问题解答等金融下游任务中均表现出色. Zhang et al. (2021) 以 RoBERTa 为基准, 利用由金融新闻、公告和金融研究报告组成的 20GB 中文金融语料库训练得到了一个轻量级的中文金融大语言模型 Mengzi-BERTbase-fin. 在 ESG 和公司金融方面, Mehra et al. (2022) 利用一个 ESG 特定文本语料库对原始 BERT 模型进行微调, 得到了一个专用于 ESG 文本分类任务的模型 ESGBERT. 在量化央行政策沟通方面, Pfeifer and Marohl (2023) 使用三个最大的中央银行演讲稿英语数据库中的人工标注数据训练 RoBERTa, 从而得到了一个可以对五种常见的经济主体和两种央行沟通情绪进行分类的模型 CentralBankRoBERTa. 在经济学命名实体识别 (named entity recognition, NER) 方面, Lasri et al. (2024) 使用由 150 万篇经济学论文组成的语料库训练 BERT 模型, 得到的 EconBERTa 模型能够在下游经济学领域的 NER 任务中表现出良好的性能. 在金融市场与资产定价方面, 姜富伟等 (2024) 结合中国特色金融市场的独特特征, 在 BERT 模型的基础上融合了结构化金融市场数据和非结构化金融文本大数据, 训练得到了一个适用于我国金融领域的中文金融大模型 (在表 2 中记为“BERT-Jiang”), 并优化了金融市场情绪测度与资产价格风险预测.

除了 GPT 和 BERT 模型, 还有一些研究基于其他种类的 LLMs 开发适用于下游经济金融任务的 LLMs. 2022 年, Scao et al. (2023) 发布了一个具有 1760 亿参数的多语言大语言模型 BLOOM, 因其多样性和可获取性, 一些研究将其应用于经济金融领域. 例如, Wu et al. (2023) 基于 BLOOM 架构并利用彭博社提供的大规模金融数据训练得到 BloombergGPT 模型, 该模型在金融特定任务上的表现优于 OPT、BLOOM 等通用开源 LLMs, 并且在通用任务上能达到与这些通用 LLMs 相近的性能. Zhang and Yang (2023) 则利用中文金融新闻、市场报告和社交媒体讨论等文本对 BLOOM 模型进行训练, 得到了首个千亿级参数的中文金融大语言模型 Xuan Yuan 2.0.

2023 年 2 月, Touvron et al. (2023a) 推出了 LLaMA 模型, 参数范围从 7B 到 65B 不等, 其训练数据的透明度和模型结构的开源程度均较高, 且性能超越了 GPT-3 等大模型, 这使得该模型一经发布就得到了广泛应用. 在经济金融领域也出现了一些基于 LLaMA 开发的专用模型, 例如 FIN-LLAMA (Todt and Babaei, 2023)、FinMA (Xie et al., 2023)、Instruct-FinGPT (Zhang et al., 2023) 和 InvestLM (Yang et al., 2023b), 这些 LLMs 在金融特定任务中均表现出专业性能. 其中, 通过指令微调方法得到的 Instruct-FinGPT 显著提升了金融情绪分析的性能, 尤其是在处理数值敏感性和上下文理解方面. 而 InvestLM 则在 LLaMA-65B 模型的基础上利用金融投资相关的指令数据集进行微调, 从而具有了卓越的理解金融文本和提供前瞻性投资建议的能力. 2023 年 7 月发布的 LLaMA 2 模型是对初代 LLaMA 模型的加强 (Touvron et al., 2023b), FinGPT (Yang et al., 2023a)、FinLlama (Konstantinidis et al., 2024)、GreedLlama (Yu et al., 2024) 和 LLaMA-2-Econ (Keles and Bayraklı, 2024) 等均是在其基础上开发的经济金融 LLMs.

除了上述模型外, 还有一些基于其他模型框架开发的适用于经济金融领域的 LLMs. 例如基于 LLaVA 架构开发的、能够对金融图表进行分析的多模态大模型 FinVIS-GPT (Wang et al., 2023b), 以及基于 Mistral-7B 模型构建的 FinTral (Bhatia et al., 2003). 同时, 另有一些研究利用中文金融文本数据进行训练, 得到了一系列中文金融大语言模型, 例如基于 InternLM-7B 模型微调得到的 CFGPT (Li et al., 2024), 基于 Baichuan-13B 模型构建的 DISC-FinLLM (Chen et al., 2023b) 以及基于 Qwen 1.5-7B 模型开发的 SilverSight (Zhou et al., 2024) 等模型.

2.3 LLMs 分析文本数据的基本原理

生成词嵌入是 LLMs 进行后续文本分析任务的第一步. 与 Word2Vec 等传统词嵌入技术不同 (具体不同之处的对比见表 3), GPT 和 BERT 等基于 Transformer 架构的 LLMs 首先通过位置编码 (Position Embedding) 技术将输入文本序列的位置信息引入初始词嵌入, 实现了对文本词序信息的直接建模, 之后通过多头自注意力机制聚合不同位置的文本信息, 将文本中各个位置之间的语义依赖关系引入初始词嵌入, 再经过后续多层的处理及反复迭代增强, 生成最终蕴含词序信息与丰富上下文语义关系的词嵌入表示. 为此, OpenAI 公司提供了一系列基于 GPT 的专用模型用于生成这种包含词序与上下文信息的词嵌入, 如 text-embedding-ada-002 (以下简称 ada-002)、text-embedding-3-small 和 text-embedding-

表 3 基于 Transformer 架构的 LLMs 与传统词嵌入技术的不同之处

对比维度	基于 Transformer 架构的 LLMs	Word2Vec 等传统词嵌入技术
位置信息引入	通过位置编码引入词序信息, 即使词语完全相同但位于不同位置, 词嵌入也不同	无法直接引入位置信息, 无法区分不同位置的相同词语
上下文信息捕捉	通过多头自注意力机制, 全局聚合来自不同位置的词语语义, 能够处理长距离依赖关系	只能基于固定窗口内的局部上下文, 无法处理较长距离的语义关系
语义依赖处理	多头自注意力机制可直接建模不同距离的词元之间的复杂依赖关系	无法直接处理词元之间的复杂语义依赖
词嵌入包含信息	词嵌入同时包含词序信息和上下文语义, 经过了多层迭代增强	仅基于词语共现信息生成词嵌入, 缺乏上下文语义信息

3-large⁷. 而 BERT 的词嵌入则可通过 Python 的 sentence_transformers 库中提供的模型和方法生成⁸.

具体而言, 在基于 Transformer 架构的 LLMs 中, 输入的词元 (token)⁹ 序列首先会经过输入嵌入模块 (Input Embedding Module) 并被映射为一个具有固定维度 D 的词向量 $v_t \in \mathbb{R}^D$, 但这些向量并不包含位置信息, 无法被 LLMs 识别. 因此, 可以根据输入的词元在序列中的绝对位置生成唯一的位置嵌入向量 $p_t \in \mathbb{R}^D$, 并与之前的词向量 v_t 相加以实现位置编码, 生成初始的输入嵌入序列 $X = x[x_1, \dots, x_t]$, 其中 $x_t = v_t + p_t$. 这一序列包含了文本的位置信息, 即使初始的词向量相同, 但如果它们在输入序列中所处的位置不同, 其最终的词嵌入序列表示也会不相同. 而 Word2Vec 等传统词嵌入技术仅通过学习词语的共现信息来生成词嵌入, 即便词语相同但输入序列不同, 也可能会生成相同的词嵌入. 因此, 位置编码技术使得 LLMs 能够额外地捕捉到文本数据的词序信息. 需要注意的是, 虽然绝对位置编码能够引入位置信息, 但它仅适用于训练样本中出现过的位置, 而无法涵盖从未出现过的位置, 这可能会严重制约其处理长文本的能力, 所以 LLMs 通常会对上下文窗口¹⁰ 的长度做出限制.

通过位置编码技术得到的初始词嵌入虽然能解决对语序的建模问题, 但仍无法捕捉文本

⁷读者可查阅 OpenAI 的词嵌入指南以获取更多相关信息:

<https://platform.openai.com/docs/guides/embeddings>.

⁸读者可查阅以下网址获得关于 sentence_transformers 库的具体介绍: <https://www.sbert.net>, 同时可查阅以下网址获得 BERT 词嵌入模型的具体信息: <https://huggingface.co/sentence-transformers>.

⁹需要注意的是, 词元 (token) 并不等同于单词. 词元可以包括单词、标点符号、特殊字符、换行符, 甚至单词片段. 根据经验, 对于普通英文文本而言, 100 个标记大约相当于 75 个单词. 读者可以查阅该网址以直观地感受单词与词元的不同之处: <https://platform.openai.com/tokenizer>.

¹⁰上下文窗口是 LLMs 在生成响应时可以分析和处理的最大信息量, 是提示 (输入) 和响应 (输出) 的文本总长度, 通常以词元的数量来衡量. 例如, GPT-4 模型的上下文窗口长度为 128K, 但需要注意的是, 这并不意味着用户可以向 OpenAI API 提供 128K 词元的文本输入. 这是因为提示 (输入) 和响应 (输出) 是共享这个窗口的, 如果提示的文本长度达到了上下文窗口长度的限制, 那么模型就没有空间输出响应了. 因此, 准确估计一段文本的词元数量对于使用 LLMs 进行文本分析是非常重要的. 具体而言, 用户可以使用 OpenAI 提供的 Python 软件包 tiktoken (<https://github.com/openai/tiktoken>) 来计算一段文本的词元数量.

中不同位置词元之间的语义依赖关系。因此, Transformer 模型创造性地引入了多头自注意力机制, 它能够直接对文本序列中间隔任意距离的词元之间的交互关系进行建模, 聚合来自于文本中不同位置词元的语义信息。多头自注意力机制可视作多组结构相同但映射参数不同的自注意力模块的集成。在每组自注意力模块中, 输入的初始词嵌入序列会被映射为对应的查询 (Query)、键 (Key) 和值 (Value) 三个矩阵。每组查询、键和值的映射构成一个“头”, 并独立计算其自注意力的输出。最后, 将不同头的输出拼接在一起, 并通过一个权重矩阵进行映射, 生成最终的输出。这一输出中蕴含了丰富的上下文信息, 模型会进一步将该输出与初始词嵌入序列 X 进行相加和层归一化以实现上下文信息的引入, 并经过前馈神经网络 (feed forward network) 处理以捕捉更复杂的语义交互关系, 最后多次迭代上述步骤以增强最终的词嵌入表示。

多头自注意力机制对文本信息的处理过程可以类比于参加一个具有多个分论坛的大型学术会议, 每个分论坛就像多头自注意力机制中的每个“头”, 均专注于讨论一个特定的话题, 并从不同角度分析这一话题。同时参加多个分论坛就可以获得不同的见解, 将各类信息组合在一起就可以形成更加全面和深刻的理解。因此, 多头自注意力机制可以使模型从不同角度同时关注输入序列的所有位置, 从而实现全局上下文的信息整合和语义提取。而 Word2Vec 等传统词嵌入技术只能基于极短固定窗口内的词语进行处理, 无法捕捉到窗口外的词语信息, 因而缺乏对较长距离的文本间依赖关系的处理能力, 只具有局部上下文信息的提取能力。

接下来, 在后续进行下游的文本分类或情绪分析等文本分析任务时, LLMs 通常会在生成词嵌入的结构基础上再连接一个多层神经网络, 这个神经网络有两个关键结构: 首先是池化层 (Pooling), 其主要目的是聚合上述生成的词嵌入以得到文本整体的语义信息并实现数据降维; 其次是全连接层 (fully connected layers), 其主要目的是根据聚合后的文本向量预测生成其对应的类别标签或情绪极性, 相当于一个分类器。在进行分类或情绪分析任务前, 首先需要对 LLMs 进行有监督下游任务微调, 即根据带有类别或情绪极性标注的文本数据训练这个多层神经网络, 从而获得该神经网络的具体拟合参数。之后, 再将待分析的无标签文本输入该模型以预测得到其对应的类别标签或情绪极性。

2.4 在经济金融领域中应用 LLMs 进行文本分析的原因

上述内容已介绍了 GPT 和 BERT 等 LLMs 的基本架构和进行文本分析的基本原理。接下来, 一个自然而然的问题就是, 我们为什么要利用 LLMs 分析经济金融领域的文本? 也就是说, LLMs 在经济金融领域的文本分析中具有哪些优势? 为什么会具有这些优势? 本文将 LLMs 的优势总结为以下三点, 并相应解释了 LLMs 具有这些优势的原因。

首先, LLMs 具有全局上下文理解能力, 能够充分结合输入文本的上下文信息。LLMs 的全局上下文理解能力主要得益于 Transformer 模型中的位置编码和多头自注意力机制这两个重要技术, 指的是 LLMs 在处理自然语言时, 能够综合考虑整个文本中的所有词语及其相互关系, 以便更准确地理解和生成具有一致性和连贯性的语言。这种能力有助于捕捉复杂的语义关系、长距离依赖性以及整体语境, 是使得 LLMs 文本分析效果显著提升的关键原因之一, 对于处理和分析一些语义模糊不清的企业年报、财经新闻及央行沟通等经济金融文本至关重要。而传统基于关键词的文本分析方法和 Word2Vec 等传统词嵌入技术更关注单个词语

而不是文章整体,忽略了文本的语序及上下文信息对语义的影响.换句话说,如果两段文本的词语完全相同但语序有巨大区别,这些传统方法通常会认为两段文本相同.此外,传统方法需要对文本进行繁琐的预处理后才能分析,这一过程也会损失掉或扭曲部分文本信息.相比之下,LLMs 更加关注文章整体的语义,能够捕捉到文本数据的复杂性和细微之处,同时也不需要经过复杂的预处理步骤.

其次,LLMs 具有较强的迁移学习能力,能够非常灵活地处理各种类型的文本. LLMs 通常是在一个非常庞大的文本语料库中进行预训练的,该语料库包含各种主题和语言的文本,这使得 LLMs 具备了零样本(或少样本)学习(zero-shot (or few-shot) learning)能力(Brown et al., 2023; Nie et al., 2023),因此能够在不提供任何样本或提供少量示例样本的情况下直接对经济金融领域内的特定文本数据进行分析,可以较好地理解经济金融领域内的一些术语和特定表达,具有较强的文本分析泛化能力.而传统文本分析方法对于不同类型文本数据进行分析的泛化能力较差.例如,在分析经济金融领域的文本情绪时,传统方法需要构建经济金融领域的特定词典,如 LM 词典(Loughran and McDonald, 2011),若直接使用通用情绪词典或其他领域的词典,分析效果可能会大打折扣.又例如,经济金融领域中经常会涉及跨语言文本(例如多个国家央行的政策沟通文件)的分析,在处理时,传统方法对不同语言的文本进行预处理的方式会有较大差别,比如英文单词之间天然有空格隔开,非常便于分词,但通常需要进行词形还原等操作;而中文分词通常较为复杂,有时甚至需要自定义分词词典以防止某些词组被错误分词.相比之下,LLMs 不用考虑这些问题,能够更好地进行跨语言的文本分析(Winata et al., 2021).

最后,LLMs 具有较好的人类对齐能力,能够模仿人类阅读文本并生成相应的信念、预期或经济决策. LLMs 的人类对齐能力指的是这些模型能够理解和遵循人类的指令及伦理价值观,模拟人类生成接近或符合人类期望和社会行为规范的文本内容(Wang et al., 2023a).在文本分析方面,LLMs 的文本阅读理解能力及文本交互能力非常接近人类,能够很好地模仿人类在阅读某段文本之后的反应(产生观点或决策),因此可以自行对某个具体事件、问题或变量(如未来的失业率将上升还是下降?)进行判断,这代表了某种类似于人类的预期或信念.然而,其他传统文本分析方法通常仅能基于关键词进行一般化的情绪分析,计算一个总体的情绪指数,难以生成针对某一特定经济变量或事件的观点.但需要说明的是,这一点通常只是生成式 LLMs 所独有的能力,而 BERT 等精于判别式任务的 LLMs 难以实现(未经过任何微调训练之前的初始 BERT 模型理论上无法完成问答任务).

3 大语言模型在经济金融文本分析中的应用场景

目前 LLMs 已被广泛应用于经济金融领域的文本分析任务,本文分别从两个方面对现有前沿研究中的应用场景(如图 2 所示)进行梳理和归纳.一是 LLMs 在传统文本分析场景中的应用与优势,包括六个应用场景:计算文本相似度、提取文本向量进行预测、文本数据识别与分类、构建特定领域词典、主题建模与分析、文本情绪分析.二是 LLMs 开创的全新文本分析场景,即 LLMs 能够作为经济代理人模拟人类生成对于文本的信念或预期.

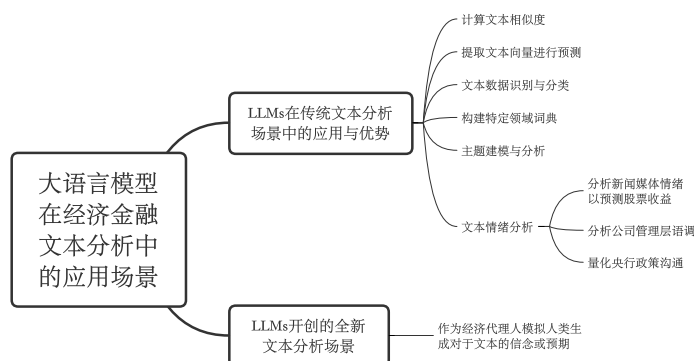


图2 大语言模型在经济金融文本分析中的应用场景

3.1 LLMs 在传统文本分析场景中的应用与优势

LLMs 具有许多相较于传统文本分析方法的优势, 因此能够取得显著优于传统方法的分析效果, 从而优化传统文本分析任务. 常见的传统文本分析任务有六大应用场景, 在这些场景中, LLMs 通常会表现出相对更优的性能.

3.1.1 利用 LLMs 计算文本相似度

文本相似度是衡量文档之间在用词或语义上的相似程度的指标. 最常见的一种算法是将文本向量化之后, 计算两个文档所对应的向量之间的余弦相似度 (Gentzkow et al., 2019). 传统词袋模型通常将文本表示成文档-词条矩阵 (document-term matrix), 通过计算矩阵中任意两个行向量之间的余弦距离即可得到相应两个文档之间的相似度. 而 Word2Vec 等词嵌入技术会将文档中的每个词转换为其对应的词嵌入向量, 之后通过对这些词嵌入向量取平均值或求和等方法将其聚合为一个向量, 再进一步通过余弦相似度计算两个文本向量之间的相似度. 由于这些传统方法生成的文本向量忽略了文本的语序信息, 所以由此计算得到的文本相似度更侧重于两个文档在用词上的相似程度. 相比之下, 基于 LLMs 生成的词嵌入向量计算的文本相似度更能体现两个文档在整体语义上的相似程度.

具体而言, Vamvourellis et al. (2023b) 基于 SBERT、PaLM 和 GPT 等 LLMs 生成的能够代表公司业务特征的词嵌入计算公司描述之间的余弦相似度以识别相似公司, 进一步发现这些相似公司之间的股票收益相关性显著高于行业内平均水平. 利用类似的方法, Breitung and Müller (2023) 提出了一种全新的全球商业网络衡量方法, 旨在评估企业之间的经济关联性, 从而充分展示了该方法的强大性能. 该方法汲取了 Hoberg and Phillips (2010, 2016) 的灵感, 但与传统方法的不同之处在于, 作者利用了 ada-002、Luminous 和 T5-XXL 等 LLMs 生成的词嵌入来计算公司描述之间的余弦相似度. 这种新的衡量方法在识别客户、供应商和竞争对手网络等关系方面表现优异, 相较于传统的行业分类方法具有更大的优势. 胡增玺和马述忠 (2023) 利用 Sentence-BERT (SBERT) 提供的 paraphrase-multilingual-MiniLM-L12-v2 模型生成具有语义意义的定长句子的词嵌入向量, 用于计算专利摘要文本和每个数字经济核心产业类别描述之间的文本相似度, 从而实现数字创新的识别.

3.1.2 利用 LLMs 提取文本向量进行预测

文本数据中蕴含着丰富的预测信息,但作为非结构化数据,若不经处理则很难直接用于经济预测.已有研究利用传统的词袋模型 (Kalamara et al., 2022) 和 Word2Vec 等词嵌入技术 (Yang et al., 2022) 提取文本的向量表示作为预测因子进行经济预测,并取得了良好的预测效果.近年来,一些研究逐渐开始尝试利用 LLMs 生成的词嵌入进行预测,与传统方法进行对比,LLMs 提取的文本向量能够进一步提升预测效果.

具体而言,Barbaglia et al. (2021) 利用 BERT 等 LLMs 和 Word2Vec、GloVe 等传统文本分析技术提取新闻文本的词嵌入,并将其作为特征变量输入一个深度神经网络模型用于预测 S&P 500 指数的已实现波动率,结果显示 BERT 模型提取的词嵌入在预测效果上表现更好.Chen et al. (2023c) 利用 LLMs 综合分析了来自 16 个国家、用 13 种不同语言撰写的新闻文章,从中提取文本的词嵌入向量作为特征变量以预测股票市场收益,同时将其与传统的 Word2Vec、SESTM (Ke et al., 2019) 等方法对比,展现了 LLMs 在提取新闻文本数据信息以预测市场收益率方面的卓越能力.类似地,Tan et al. (2023) 通过相同的思路和方法验证了 LLMs 提取的中文新闻文本向量也能有效预测中国股票市场的收益率.Yang (2023) 利用 ada-002 模型提取专利文本的词嵌入,并将其作为预测因子输入深度机器学习模型预测专利价值和申请的接受与否,结果发现这种包含文本整体语义信息的词嵌入能够提高预测的准确性.

3.1.3 利用 LLMs 对文本数据进行识别与分类

文本的识别与分类是指根据文本的某些特征或标准为输入数据分配预定义的标签¹¹.传统文本分析技术在处理分类任务时通常有两类方法,一类是无监督学习方法,如词典法等;另一类则是有监督学习方法,如支持向量机和神经网络算法等 (沈艳等, 2019). 近期一些研究通过使用 LLMs 对文本数据进行识别与分类并取得了相比于传统方法更好的效果.

具体而言,Korinek (2023) 通过一个与 GPT-4 的对话实例展示了新一代 LLMs 在文本数据分类方面的能力.在对话中,作者根据 Wei et al. (2022b) 提出的思维链提示原则,通过提示要求 GPT-4 执行分类任务并进行理由说明,以引导 LLMs 在执行实际分类之前对其响应进行推理.类似地,Eloundou et al. (2023) 也采用这一方式利用 GPT-4 系统地估算了 LLMs 对美国劳动力市场的潜在影响.Chen et al. (2024) 通过类似方式利用 GPT-3.5 研究了劳动力市场的不匹配现象,对具有特定特征的求职者是否适合特定工作进行判断,从而恢复了分类变量中被忽略的信息.这一过程一般通过使用 Open AI 提供的 API 进行提示以获取 GPT 模型的输出,同时为了最大程度地确定模型输出,这些研究通常会通过设置模型的

¹¹需要说明的是,虽然从自然语言处理的技术角度来看,情绪分析的操作流程在本质上与文本分类任务并无差异,但本文并不是从技术角度出发进行梳理,而是从 LLMs 的各类应用场景角度出发对文献进行归纳的.由于目前有大量应用研究使用 LLMs 进行情绪分析,且情绪分析是文本分析的一个重要应用场景,因此本文在 3.1.6 节对情绪分析的应用研究进行集中讨论.这一梳理思路与 Dong et al. (2024)、Nie et al. (2023) 及沈艳等 (2019) 等文献一致.

temperature 参数¹² 来尽量控制输出结果的可重复性。

此外, 大量研究使用 BERT 模型及其变体进行文本数据分类。Hansen et al. (2023) 研究了五个英语国家的 2.5 亿多个职位空缺信息, 利用基于 DistilBERT (Sanh et al., 2019) 改进的 LLMs 对这些招聘信息是否属于远程工作进行识别与分类, 实现了 99% 的准确率, 远远超过了传统的词典法, 也超过了其他传统的机器学习方法。为了通过专利摘要语义判断该专利技术是否属于劳动节约型技术, 刘青和肖柏高 (2023) 利用 5000 条人工标注数据训练 BERT 模型以用于大规模的专利分类, 结果发现, BERT 模型在样本内的预测准确度接近 100%, 样本外的预测准确度也高达 90%。类似地, 诸竹君等 (2023) 采用中国专利数据和 BERT 模型对微观层面与服务型制造相关的专利进行识别与分类, 探索性地构造了制造业企业服务型制造技术创新代理变量。金星晔等 (2024) 首先对 2006 年至 2020 年上市公司年报中的句子进行人工标注, 随后利用这些标注结果对多种机器学习模型 (包括大语言模型) 进行训练和微调。通过比较不同模型的分类效果, 选择了效果最佳的 ERNIE 大模型作为句子分类模型, 用于预测整个文本中句子的标签, 最终构建了企业数字化转型的新指标, 结果显示该新指标优于传统词典法所构建的指标。

3.1.4 利用 LLMs 构建特定领域词典

词典法是一种传统的文本分析方法, 其核心思想是基于预先设定的词典, 通过统计文本数据中不同类别词语的出现次数, 并结合不同的加权方法来提取文本信息。在使用词典法时, 关键的一步是选择或构建合适的词典 (沈艳等, 2019)。经济金融领域的研究中常见的英文特定词典包括 Harvard IV-4 词典、Henry 词典、Diction 词典和 LM 词典。但在一些特定领域的研究中, 通常需要研究者根据文本特征和研究目的自行构建领域内的词典。

一些研究利用 LLMs 为构建特定领域词典提供了一套具有可操作性的方案。具体而言, Chava et al. (2021a) 利用 RoBERTa (robustly optimized BERT approach) 模型 (Liu et al., 2021) 构建了一个环境与社会 (E&S) 领域的词典, 并使用该词典分析了 2007–2019 年期间美国上市公司的盈利电话会议, 结果发现环境话题的讨论与更高的污染减排和更多的未来绿色专利相关。Chava et al. (2021b) 则利用 BERT 模型从盈利电话文本中构建了一个新兴技术短语词典, 发现基于 BERT 构建的新兴技术短语词典能有效区分市场讨论的可信度, 并研究分析了股票市场对公司讨论新兴技术的直接正向反应。为了解决 LM 词典更新困难的问题, Gupta et al. (2023) 采用 BERT 模型从美国证券交易委员会文件 (10-K 和 8-K 报告) 中自

¹²temperature 是一个控制 ChatGPT 生成文本多样性的参数。该参数的值范围在 0 到 2 之间。当 temperature 的值接近 0 时, 模型生成的文本将更加确定和一致, 有可能重复性更强, 也就是说模型将更倾向于选择概率最高的输出。这会使得生成的内容较为固定和单一。当 temperature 的值接近 2 时, 模型生成的文本将会更加多样和有序, 也就是说模型在选择下一个词或短语时会考虑更多可能性, 这会使生成的内容更具多样性和创造性, 但也可能导致模型的输出变得不可预测或者不连贯。为保证结果的可复现性, 现有大多数研究均将该参数设置为 0, 因为较低的 temperature 有助于提高一致性, 而不会降低准确性 (Gilardi et al., 2023)。类似地, OpenAI 近期引入了一个新参数 Seed, 目的是使输出更具确定性。通常设置该参数为一个整数, 作用类似于随机数生成器的种子, 但并不保证输出结果完全确定。使用相同的 seed 设置和完全相同的其他参数很可能在不同的请求中产生相同的输出, 具体可参阅: https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter。

动创建词典,并比较了该自动创建的词典与 LM 词典对财务业绩的预测效果.结果发现相较于传统的 LM 词典,使用 BERT 模型自动生成的词典能够更有效地提取文本特征,并能捕捉到传统情绪词典之外的信息.

3.1.5 利用 LLMs 进行主题建模与分析

主题模型近年来在经济和金融领域逐渐得到广泛运用.其代表模型是由 Blei et al. (2003) 提出的 LDA 模型.该模型主要有四个方面的局限性:首先,主题数量的选择较为主观.其通常需要人为地给出一个主题数量,而主题数量的选择对主题的生成和文档的归类产生较大影响(沈艳等, 2019).其次,忽略了单词上下文的语义关系.该模型的输入通常是基于词袋法表示的词频矩阵,从而忽略了词与词之间的语义关系(Grootendorst, 2022).再次,在短文本的主题分析上通常表现不佳. LDA 模型通常需要庞大的文本数据集作为训练数据,因为其需要足够的词语共现信息才能识别主题(Grootendorst, 2022).最后, LDA 模型计算效率较低,在处理一些大规模文本数据时通常非常耗时.

随着 LLMs 相关技术的迅速发展,一些研究开始探索利用 LLMs 进行主题建模与分析,试图解决传统 LDA 模型所面临的问题. Grootendorst (2022) 提出了一个名为“BERTopic”的主题模型¹³,该模型主要通过三个步骤识别主题:首先,使用预训练语言模型 BERT 将每段文本转换词嵌入向量.然后,根据语义相似性对这些词嵌入向量进行聚类:先对高维词嵌入向量进行降维处理,接着对降维后的向量进行聚类操作,将语义相似的文本聚类到同一个语义簇(即主题)中.最后,使用基于类的 TF-IDF (class-based TF-IDF) 算法,计算每个语义簇中各词语的重要性得分(即 TF-IDF 值),进而提取每个簇中最重要的若干词作为该语义簇的关键词,实现对主题的识别.识别主题之后,可进一步计算文本到各个主题中心点的距离,以获取该文本属于各个主题的概率.这种创新方法克服了传统 LDA 主题模型的局限性:首先, BERTopic 能够根据算法自动确定主题数量,不需人为预先设定,避免了主观因素所带来的偏误.其次, BERTopic 基于 BERT 模型生成文本的词嵌入,从而充分捕捉了文本的整体语义信息,能够更加准确地表示文档内容.再次, BERTopic 非常适用于对短文本进行主题提取,由于它并不是文本集合的生成概率模型,因此不需要提供大量的词语共现信息训练模型.最后,与 LDA 等传统主题模型相比, BERTopic 可以调用 GPU 进行数据处理,从而大大提高了计算效率.

目前已有一些文献开始探索如何利用 BERTopic 模型进行实证研究.具体而言, Born et al. (2023) 利用 BERTopic 模型从 Twitter 文本中提取与通货膨胀相关的主题用于构建高频通货膨胀指数.郑晓龙等 (2023) 基于 BERTopic 对东方财富网股吧平台的用户发帖和评论文本的话题进行分析,发现在低估信息事件(如利空谣言)传播过程中,大量用户选择相信未经证实的利空政策并呼吁出手该股,而少部分用户持相反态度,呼吁抄底持股并怀疑存在恶意做空行为. Hajek and Henriques (2024) 利用 BERTopic 对关于特定公司的新闻文章进行主题的提取与分析,发现从新闻文章中提取的主题特征显著提高了并购目标预测和投资的超额回报预测的准确性. Tang et al. (2024) 使用 BERTopic 分析了关于公司债券发行者的新闻标题的主题信息,结果表明,该主题指标可以作为预警信号以提高对 3 个月范围内公司债券

¹³作者提供了一个 Python 软件包 bertopic 用于算法实现: <https://github.com/MaartenGr/BERTopic>.

违约的预测性能, 进一步证明了 BERTopic 在新闻标题等短文本分析中的优越性。

近期, 也有一些研究尝试利用 GPT 进行主题建模与分析, 提出了 “TopicGPT” (Pham et al., 2023) 和 “GPTopic” (Reuter et al., 2024) 等方法, 并证明了相比于 BERTopic 模型, 这种基于 GPT 的主题模型所识别的主题与人类注释的真实主题更加接近。但目前该类方法还尚未在经济金融领域内得到应用。

3.1.6 利用 LLMs 进行文本情绪分析

度量情绪是文本大数据在经济金融领域内的一大应用 (沈艳等, 2019)。在这类分析中, 一段文字会被归类为具有积极、消极或中性的语调, 并通过赋予数值来区分语调的强度。目前利用 LLMs 对文本进行情绪分析的研究文献较为丰富, 主要集中在分析新闻媒体情绪以预测股票收益、分析公司管理层语调、量化央行政策沟通等多个应用方向。一些研究将 LLMs 方法与传统的文本情绪分析方法 (如人工赋值法、情绪词典法和某些机器学习方法) 进行对比, 展现了 LLMs 方法的优越性。

1) 分析新闻媒体情绪以预测股票收益

新闻媒体情绪反映了媒体报道文本中所包含的乐观和悲观情绪, 这类情绪通常会对市场中的经济主体产生影响, 进而影响资产价格或投资决策。目前有大量文献利用 LLMs 分析新闻媒体情绪以预测股票收益, 这是该领域当前的研究热点, 相关研究层出不穷。例如, 许雪晨和田侃 (2021) 将 BERT 模型应用到财经新闻文本的情绪分析中, 并将文本的情绪特征与股市行情交易数据相结合, 有效提高了股指趋势预测的准确率。Chen et al. (2023a) 利用 ChatGPT 从 1996 年至 2022 年《华尔街日报》的新闻标题和警示中提取股市的利好和利空消息, 研究发现高比例的利好消息与随后的月度市场回报之间存在显著关联。此外, ChatGPT 识别的文本情绪信息更可能与宏观经济状况相关。相比于小型 LLMs (如 BERT 或 RoBERTa) 和传统文本分析方法, 大型 LLMs 表现出更强的能力。基于 GPT-3.5 训练期截止点之后¹⁴的新闻标题数据集, Lopez-Lira and Tang (2023) 研究了使用 ChatGPT 等 LLMs 对新闻标题进行情绪分析以预测股市收益的潜力。结果发现, 相比于传统情绪分析方法, ChatGPT 分析的情绪具有更强的收益预测能力。何勇等 (2024) 利用 BERT 和 ChatGPT 处理财经新闻文本数据及传统金融数据, 从中获取交易信号, 构建了适用于中国股票市场的投资交易模型。

2) 分析公司管理层语调

上市公司年报的管理层讨论与分析 (MD&A)、盈利电话会议和其他公开信息披露文件等文本中通常包含上市公司管理层对当前经营状况的分析和未来发展方向的讨论, 因此往往能反映管理层的决策和意图 (沈艳等, 2019)。新近研究表明, LLMs 在提取和分析管理层语调方面具有不可忽视的潜力。例如, Cao et al. (2023) 使用 FinBERT 模型对 10-K 文件中的 MD&A 部分进行情绪分析, 并将其与人类读者所感知到的情绪进行对比, 考察了机器技术引

¹⁴这一做法用于避免可能存在的前瞻性偏误问题。LLMs 的前瞻性偏误问题是指: 由于 LLMs 的训练期和该研究所抽取的样本期有重叠, 且 LLMs 的训练样本通常十分庞大和复杂, 所以由 LLMs 生成的情绪分析结果可能并不是其对于当前文本所传递情绪的直接判断, 而只是其对于训练样本的记忆和复述, 从而导致了前瞻性偏误。因此, 这一做法可以确保作者的分析结果是基于模型训练数据中不存在的信息, 从而允许对其情绪分析能力进行更准确的评估。

入所带来的影响。Kim et al. (2024) 利用 GPT-3.5 模型对 MD&A 和盈利电话会议的文本内容进行凝练总结,发现凝练后的文本能够有效捕捉并放大原文本的情绪,且对股票异常回报率的预测能力更强,即凝练后文本的信息含量更高。Beckmann et al. (2024) 利用最新的 GPT-4-Turbo 模型对 S&P 500 公司的盈利电话会议的文本进行分析,从 25 个维度对异常财务沟通进行识别与分类。结果表明,非常规财务沟通普遍存在于企业高管和分析师的交流互动中,且会引发显著的负面市场反应,尤其是当高管提供的信息模糊不清、分析师的问题缺乏深度挖掘时,市场的反应会格外明显。

3) 量化央行政策沟通

央行通过定期发布政策报告、召开新闻发布会和发表讲话等方式向市场传递货币政策信息。研究人员通常从这些政策性文本数据中提取信息,以探究央行政策沟通对金融市场的影响。目前测度央行沟通的文献中常见的传统方法主要有:主观赋值法、关键词计数法、基于通用词典的情感指数法以及机器学习方法(林建浩等, 2021)。一些研究开始尝试使用 LLMs 对央行政策沟通进行测度与分析。例如, Hansen and Kazinnik (2023) 研究了 GPT 模型对美联储讲话 (Fedspeak) 的理解和阐释能力,证明了 GPT 模型在识别联邦公开市场委员会 (FOMC) 公告中的政策立场方面超越了其他流行的测度方法。Gorodnichenko et al. (2023) 在研究中利用 BERT 模型量化了 FOMC 沟通文本的情绪,并使用其微调模型 RoBERTa 和 FinBERT 验证了测度结果的稳健性。Peskov et al. (2023) 使用 GPT-4 来量化 FOMC 在通胀问题上的不同意见。研究发现,文字记录和会议纪要反映了委员们对经济前景的不同观点,但这些观点在最终的公开声明中几乎完全被忽略,因此仅凭声明来测度 FOMC 的情绪并不能充分反映鹰派和鸽派之间的分歧。

3.2 LLMs 开创的全新文本分析场景

上述六个场景均可使用传统文本分析方法实现, LLMs 则凭借全局上下文理解能力和较强的迁移学习能力这两个优势超越了传统方法,进而为传统文本分析任务提供了一种全新且性能更优的备选方案。在本小节中,本文将进一步关注 LLMs 如何凭借其第三个优势(即人类对齐能力)开创一类新的文本分析场景——作为经济代理人 (economic agents) 模拟人类生成对于文本的信念或预期,而这是所有传统方法所无法实现的应用场景。LLMs 拥有大量的训练数据和庞大的参数空间,表现出了能够模仿人类行为的涌现能力 (Brown et al., 2023; Wei et al., 2023a; Bubeck et al., 2023; Zhao et al., 2023), 能够实现较好的人类对齐效果,因此可以被视作经济社会中的一个代表性主体 (representative agent), 标志着“AI 行为科学”的出现 (Meng, 2024)。

现有研究的实验证据也支持了上述观点,即 LLMs 确实能够模拟人类生成与人类接近的信念或偏见,甚至能够模仿人类行为进行决策。例如, Aher et al. (2023) 进行了一系列“图灵实验”,旨在测试 LLMs 是否能在各种环境下模仿人类,并发现了积极的结果。Horton (2023) 同样发现,在一系列经济学实验中,通过给 LLMs 提供禀赋、信息、偏好等设定,然后通过模拟在不同场景中探索它们的行为,发现 LLMs 的行为与人类非常相似,可以作为经济代理人。类似地, Chen et al. (2023d) 发现 LLMs 在多个领域的决策基本都是具有经济理性的,其理性得分甚至高于人类受试者,具有做出正确决策的潜力。此外, Argyle et al. (2023) 在政治

学研究中利用 GPT 生成了模拟人类的调查样本, 也发现了类似的结果. 进一步地, Leng and Yuan (2024) 发现 LLMs Agents 表现出一系列类似人类的社会行为, 并能够在实验中替代人类受试者. Li et al. (2024) 发现 LLMs 在经济环境中表现出类似人类的决策和适应能力, 并具备感知、反思和决策能力, 进而提出了 EconAgent 的概念. 对宏观经济活动的模拟实验表明, 与基于规则或学习的 Agents 相比, EconAgent 可以做出更接近人类的工作和消费决策, 并模拟出更合理的宏观经济现象.

基于上述文献的理论与实验基础, 一些经济学者探索利用 LLMs 模拟人类阅读文本并据此生成相应的信念或预期, 同时验证了这些生成式 LLMs 生成的信念或预期与大规模经济调查的指标较为接近, 具有替代人类受访者生成模拟调查样本的潜力. 具体而言, Bybee (2023) 利用 GPT-3.5 阅读《华尔街日报》的历史新闻文章样本, 并通过提示询问其对各种金融和宏观经济变量走势的预期, 由此构造了一个由 GPT 生成的预期指标. 研究进一步发现该指标与现有的大规模预期调查指标具有高度相关性, 具有替代或扩展现有调查数据时间序列的潜力. 类似地, Jha et al. (2024) 利用 GPT-3.5 分析了盈利电话会议记录文本, 令其模仿金融专家对公司未来的资本支出生成预期, 由此构造了 ChatGPT 投资得分. 结果发现, 该投资得分与大规模 CFO 调查的投资指标具有显著的相关性, 这表明 GPT 能够从大量文本中提取相关信息, 生成接近于公司管理层的信念, 具有补充大规模高管调查的潜力. 进一步地, Zarifhonarvar (2024) 探究了不同知识源的文本数据对 LLMs 通货膨胀预期形成的影响. 研究发现, 文本数据的信息类型会对 LLMs 所生成的预期的方差产生显著影响. 此外, 当 LLMs 接收到关于抵押贷款利率的信息时, 它们会根据这些新信息调整自己生成的通货膨胀预期, 这一点与人类受访者在调查中的行为是类似的.

在近期, LLM AI Agent (译作“大模型智能体”) 的相关研究与市场化产品层出不穷. 大模型智能体是对上述研究与相关概念和技术的进一步拓展, 是能够感知环境、做出决策并独立执行行动的智能实体 (赵鑫等, 2024; Gao et al., 2019). 这一概念和技术正逐渐在经济金融领域展开应用, 其中最主要且与业界实务联系最为紧密的应用就是基于大模型智能体的智能投顾平台. 这些智能体能够处理海量数据并以高精度度提供投资建议或执行交易, 其依托 LLMs 的自然语言处理能力综合并解读财经新闻、市场报告或公司财务报告, 从而显著提升市场预测、交易策略和投资建议的准确性 (Nie et al., 2023; Yang et al., 2023). 例如, 基于 LLMs 的交易与投资智能体可以将多模态数据 (包括文本、图片和视频) 进行整合, 由此构建新的智能体 FinAgent (Zhang et al., 2023). FinAgent 可以进行股票和加密货币市场的量化和高频交易. 多元记忆检索系统和工具增强特性使 FinAgent 能够与各种数据源和工具进行互动, 提升了在动态交易环境中的适应性和性能. 此外, 这些交易与投资智能体在持续学习和环境适应方面同样表现出色. 例如, Yu et al. (2024) 提出的 FinMem 智能体引入了分层记忆和角色设计, 增强了代理处理海量多源金融数据并转化为交易决策的能力. 受到人类认知过程的启发, FinMem 的记忆模块包括工作记忆和分层长期记忆组件. 这允许 FinMem 按照不同信息的相关性和时效性进行分类和排序, 在保留更长时间历史信息的同时, 对新的投资信息做出敏捷反应. FinMem 还可以通过现实世界的测试, 不断演化其交易策略, 展示出自动改进的灵活决策机制.

综上可知, 这一应用场景与之前六个应用场景有本质区别: 之前的场景不涉及利用

LLMs 模拟人类生成对于某一特定事件或指标的看法以及作出针对性的决策和行动, 因此利用传统的文本分析技术也可实现, 而该全新场景展示了 LLMs 所独有的人类对齐能力在经济金融文本分析中的应用.

4 未来研究展望

综合上述各应用场景中的前沿文献, 本部分首先将 LLMs 与传统文本分析方法进行对比, 总结归纳了利用 LLMs 进行经济金融文本分析的局限性及争议, 进而针对这些局限性、争议及现有前沿研究中的不足对未来 LLMs 助力经济金融文本分析研究的发展方向和相关策略进行展望.

4.1 利用 LLMs 进行经济金融文本分析的局限性及争议

虽然 LLMs 在经济学和金融学研究中的应用逐渐广泛, 且能够帮助研究者大幅提高工作效率, 但仍然存在诸多局限性及争议有待未来研究解决. 结构复杂的 LLMs 如何适用于经济金融文本数据的挖掘有待研究者进一步尝试、探索与规范化 (Ash and Hansen, 2023). 相比于传统文本分析方法, LLMs 主要存在以下四个方面的局限性及争议:

第一, 文本分析结果的可复现性和可解释性有待改善, 且可信性和客观性遭受质疑. 首先, LLMs 的输出结果通常较为随机, 且输出结果通常对人类提示的特定措辞非常敏感. 尽管一些 LLMs 如 OpenAI 的 GPT 模型提供了用于控制输出结果一致性和确定性的参数, 但仍然无法确保结果能够百分之百的复现. LLMs 输出结果的难以复现性和较差的稳健性会使经济学的实证研究面临可信性威胁.

其次, 目前自然语言处理的发展方向是开发结构越来越复杂的模型, 而这些模型的透明度和可解释性越来越低¹⁵ (Zhao et al., 2024; Luo and Specia, 2024), 这一“黑箱”问题比传统文本分析方法更加突出, 将成为其应用于经济学和金融学研究中的一大阻碍 (Ash and Hansen, 2023; 洪永淼和汪寿阳, 2024). 以文本情绪分析为例, LLMs 通过复杂的非线性运算处理海量数据, 尽管其情绪分析可能更准确, 但研究者无法解释模型的决策过程和权重分配等信息, 难以追溯到具体的词语或语法结构, 其内部机制是模糊不清的. 相比之下, 情绪词典等传统方法虽然灵活性较差, 但具备可解释性, 研究者可以清晰看到哪些词语触发了情绪判断, 如“增长”代表积极情绪, “下滑”代表负面情绪. 因此, 在使用 LLMs 时, 虽然其分析结果可能比传统方法更接近人类, 但当研究者试图解释为什么模型在某一特定文本上做出了某种判断时, 遇到的挑战要远大于传统方法. “黑箱”问题会使 LLMs 在涉及监管、透明性要求、学术验证时可能面临更多质疑, 特别是在经济政策制定或金融市场预测等领域, 需要更高的解释性和可控性.

再次, 一些研究发现 LLMs 有时会产生“幻觉 (Hallucination)”而输出一些错误的分析结果 (Ji et al., 2022; Yao et al., 2024), 甚至还具有“欺骗能力”, 会对用户产生误导 (Hagendorff, 2024), 这将严重影响 LLMs 文本分析结果的可信性. 例如, 通过 Prompt 询问 GPT-4o 模型: “请你列举出 2024 年发表在经济学顶级期刊 American Economic Review 的两篇论文标题”, GPT-4o 的回答如图 3 所示. 经过人工核对可知, 第一篇论文标题完全不存在, 第二篇论文虽

¹⁵截至本文成稿之时, OpenAI 尚未披露 GPT-4 的具体模型参数和训练数据.

请你列举出2024年发表在经济学顶级期刊American Economic Review的两篇论文标题

已搜索 3 个网站

Here are two papers published in 2024 in the *American Economic Review*:

1. **"A Theory of Bank Runs and Deposit Contracts"** by M. Kathleen Ngangoue & Andrew Schotter.
2. **"When Losses Turn into Loans: The Cost of Weak Banks"** by Laura Blattner, Luisa Farinha, & Francisca Rebelo

These papers explore issues related to banking stability, examining the causes and consequences of weak financial institutions.

图3 GPT-4o 产生“幻觉”的对话实例

然发表于 *American Economic Review*, 但发表时间是 2023 年 6 月。因此, GPT-4o 产生了“幻觉”, 生成了错误的信息。由于目前 LLMs 被广泛应用于金融分析、报告撰写、经济预测等场景, 因此“幻觉”问题可能会导致不同信息接收者之间的信息不对称现象加剧, 进而有可能造成导致市场价格的不合理波动甚至经济政策的不确定性与风险。

最后, 一些研究探讨了 LLMs 的输出结果是否具有客观性, 发现 LLMs 具有性别刻板印象 (Chen et al., 2024) 和意识形态偏见 (Agiza et al., 2024) 等信念偏见, 这也可能会影响文本分析结果的准确性。

第二, 可能存在前瞻性偏误问题。在某些预测或测度任务中, 由于 LLMs 的预训练窗口固定且样本期通常较长, 待分析的文本数据的样本期可能会与 LLMs 的训练样本期有重叠, 从而可能引致前瞻性偏误问题。目前文献中的主流解决方法是, 分别报告 LLMs 预训练截止点之前的样本内分析结果和截止点之后的样本外分析结果, 如果二者没有明显差异, 则认为不存在前瞻性偏误问题。例如, Lopez-Lira and Tang (2023)、Bybee (2023) 以及 Jha et al. (2024) 均采用这种方法证明了其结果的稳健性。但这种方法通常会面临样本外数据量不足的限制, 因为一些 LLMs 的预训练截止点在距今两三年甚至几个月之前, 某些季度指标可能难以收集足够的时间序列样本用于样本外预测。而传统文本分析方法可以通过滚动 (rolling) 或递归 (recursive) 的方式进行训练和样本外预测, 从而有效缓解前瞻性偏误问题。

第三, 对于经济金融领域中长文本的分析能力有待检验。GPT 和 BERT 等 LLMs 通常会限制单次输入文本的长度, 即会设置上下文窗口的长度限制, 从而使其难以对企业年报或新闻文章等长文本进行分析。现有研究通常有两种处理思路, 如 Vamvourellis et al. (2023b) 和 Breitung and Müller (2023) 先将长文本拆分成文本块或句段输入模型, 最后再将输出结果在文章整体层面加总或平均; 而 Chen et al. (2023c) 则直接将长文本正向截断¹⁶后输入模型。但这两种处理方法都有可能造成长文本信息的扭曲或损失。此外, 对于一些性能优越且常被使用的闭源大模型 (如 GPT-4) 而言, 过长或规模较大的文本会产生极高的处理费用¹⁷。

¹⁶ 正向截断处理指最多只保留一段文本的前 n 个词元, 其中 n 为上下文窗口的最大长度。

¹⁷ 例如, 向 OpenAI 的 API 发出的每个请求都会按词元数量计费, 因此处理规模庞大的文本数据时可能会产生高昂的费用。

但需要说明的是,目前 LLMs 的长上下文处理领域发展十分迅速,近期已涌现出大量研究及产品,使得 LLMs 的长文本处理能力不断提高 (Dong et al., 2023; Huang et al., 2024; Pawar et al., 2024),且处理成本也不断降低。例如 OpenAI 在 2024 年 4 月推出的 GPT-4-Turbo 模型宣称将上下文窗口拓展至 128K,而在同年 5 月推出的 GPT-4o 模型将 API 价格降低至 GPT-4-Turbo 的 50%。因此,可以预见,LLMs 的这一局限性在不久的将来有望得到改善,但其对于经济金融领域长文本的分析能力仍有待进一步实证检验。

第四,可能存在隐私泄露问题。经济金融领域涉及大量敏感的客户数据和交易信息,如果这些数据在大模型训练或推理过程中未得到适当保护,可能导致客户隐私泄露,从而引发信任危机、法律风险和经济损失。LLMs 通常需要大量数据进行训练,在集中式训练中,数据会被汇总到单一服务器,一旦服务器受到攻击或模型存在漏洞,敏感信息就可能被恶意获取,增加了隐私泄露的风险。而在推理阶段,用户输入的文本可能包含敏感信息,模型的生成结果可能会无意间引用与特定用户相关的具体数据。例如,2023 年 11 月,OpenAI 公司的某位用户发现,如果反复要求 ChatGPT 重复某个单词,可能会使其在生成内容中泄露训练数据中的个人隐私信息 (赵鑫等, 2024)。虽然这一漏洞现已修复,但也同时反映出大模型在数据安全性和隐私性方面的局限性。

4.2 未来研究方向和相关策略

4.2.1 针对局限性及争议

针对上述 LLMs 所面临的局限性及争议,本文逐一提出了未来的研究重点及可能的解决策略。

首先,LLMs 输出的文本分析结果依赖于人类输入的提示,为了提升 LLMs 输出结果的可复现性和稳健性,必须确立经济学实证研究中稳健提示设计所依据的一般原则。因此,未来的一个研究重点是:如何根据不同类型的经济金融文本分析任务设计具有针对性的提示框架或策略,使得 LLMs 在相应任务上取得更好的分析效果与更稳健的输出结果。例如,Pham and Cunningham (2024) 发现使用“未来叙事”提示策略¹⁸可以显著提高 GPT-4 模型在预测未来事件时的准确性。同时,由于 LLMs 文本分析的结果具有随机性,即便输入相同的提示也可能得到不同的输出,其中一些输出可能存在严重的误差。因此,对于经济金融的实证研究而言,仅依赖于 LLMs 某一次的输出结果进行分析很可能会使实证结果产生偏误。未来研究可以考虑进行大量重复实验,获得 LLMs 的多次输出结果,并以此为样本进行估计,构建 LLMs 文本分析结果的置信区间。

此外,在可解释性方面,未来研究应关注如何提高 LLMs 生成的词嵌入的经济学可解释性。例如,可借助 t-SNE、UMAP 等降维技术将高维词嵌入向量可视化,通过二维或三维图表展示词嵌入的分布,直观地观察词与词之间的聚类情况。同时也可以尝试借助或改进一些可解释 AI (XAI) 技术如 SHAP (Shapley additive explanations) (Lundberg and Lee, 2017) 等,用于分析词嵌入向量的特征重要性,了解不同维度在词语意义表达中的贡献,并解释词

¹⁸“未来叙事”提示策略是一种通过让生成式 LLMs 讲述设定在未来的虚构故事来间接预测或描述过去已发生事件的方法。这种策略利用了生成式 LLMs 的创造力和叙事构建能力,以故事形式嵌入预测信息,从而可能提高预测的准确性。

嵌入向量的组成部分以及它们在不同任务中的影响。在未来, 研究者也需建立 LLMs 输出结果可解释性的评估指标。例如, Tong et al. (2024) 构建了诚实度 (Faithfulness) 和信息度 (Informativeness) 这两个可量化的指标来评估 Ploutos 模型所生成内容的可解释性。

在可信性方面, 为了尽量确保 LLMs 文本分析结果的准确性和真实性, 未来研究可以尝试构建对 LLMs 的响应内容进行真实性核查的工具。Krishna et al. (2019) 开发的 GenAudit 是一个初步尝试, 其可以协助检查 LLMs 对文本分析任务的响应内容是否准确, 通过修改或删除没有事实依据的生成内容来编辑 LLMs 的响应, 并为有事实依据的生成内容提供证据支持。但 GenAudit 对于 LLMs 在经济金融领域文本中分析的纠错能力并未得到有效检验, 未来研究应考虑在此基础上构建适用于经济金融领域文本的 LLMs 文本分析核查工具。

其次, 为了缓解 LLMs 在文本分析任务中的前瞻性偏误问题, 未来的研究可以考虑如何实现 LLMs 预训练窗口的滚动或递归。当前大多数的 LLMs 均采用固定的预训练窗口, 这种做法虽能够节省训练时间和成本, 但在利用 LLMs 对某一特定时段的文本进行分析时, 可能会将预训练窗口中未来时段的信息引入分析进而产生偏误。在未来, 随着算力不断提高, LLMs 的预训练时间和成本将不断降低, 因此可以尝试将 LLMs 预训练窗口进行滚动或递归 (Sarkar and Vafa, 2024)。同样地, 近期 Drinkall et al. (2024) 提出的 TimeMachineGPT (TiMaGPT) 也为 LLMs 的前瞻性偏误问题提供了解决方案。TiMaGPT 模型是由一系列特定时间点的 LLMs 组成的, 每个 LLM 都是在特定截止日期之前发布的数据上预训练的, 以确保每个特定时点的 LLMs 不会包含这一时点之后的任何信息, 进而能够实现 LLMs 的样本外预测, 避免了前瞻性偏误与样本外数据不足的问题。这种做法类似于对 LLMs 的预训练窗口进行递归。

再次, 未来研究应考虑如何在实证研究中检验 LLMs 对经济金融长文本的分析能力。一方面, 未来研究应尝试构建专门用于评估 LLMs 对经济金融文本的分析效果的指标。另一方面, 未来研究应在实证研究中对利用 LLMs 对长文本整体直接进行分析的效果与将长文本拆分或截断后再输入 LLMs 进行分析的效果有何不同, 评估哪种处理方式具有更好的分析效果, 探究不同处理方式分析效果的优劣是否会因分析任务和文本的不同而有所差别, 进而归纳出利用 LLMs 进行长文本分析的适用任务和特定文本。

最后, 为了有效缓解 LLMs 的隐私泄露问题, 未来研究可从两个角度思考解决策略。首先, 保证预训练文本数据的纯净性。一种直接且有效的方法是利用启发式方法来检测和剔除这些个人信息 (Laureçon et al., 2022), 如关键词识别。这类方法采用基于规则的方法来过滤预训练数据集中的隐私信息。在文本收集过程中, 一旦检测到这些预标注的隐私信息, 就会进行替换或直接删除整个文档。其次, 保证预训练数据和训练过程的本地化。联邦学习 (federated learning) 是一种分布式学习 (distributed learning or training) 方法, 它允许模型在多个参与者的本地数据上进行训练, 而不需要将数据集中到一个地方。这种方法减少了对中央数据存储的依赖, 从而降低了隐私泄露的风险 (Kuang et al., 2024)。参与者在本地训练模型并只发送模型参数更新到中央服务器, 而不用上传原始数据。这种做法有助于保护数据隐私, 并降低数据泄露的风险。未来研究可以聚焦如何优化联邦学习算法以解决大模型隐私泄露问题。

4.2.2 针对现有研究不足

根据现有前沿研究中的一些不足之处, 本文将分别从三个方面针对性地提出未来可能的研究方向及相应的研究策略.

1) LLMs 在传统文本分析场景中的研究机会

在传统文本分析场景中, 利用 LLMs 提取文本向量进行预测以及主题建模与分析这两个场景中的应用研究并不丰富, 具有进一步延伸的空间和潜在研究机会.

首先, 未来研究可尝试拓展基于 LLMs 提取的文本向量在宏观经济预测和资产定价中的应用, 并加强经济学解释. 现有研究主要将 LLMs 提取的文本向量作为特征变量进行预测, 但大多集中在金融或微观层面的预测, 尚未有研究探索其在宏观经济预测或资产定价方面的效果, 同时缺乏对文本向量预测能力来源的经济学解释. LLMs 提取的文本向量可以理解为对文本叙事注意力的提取, 其中蕴含着丰富的预测信息. 在未来, 研究者可以探索这些文本向量是否可以作为宏观经济预测问题中的预测变量或资产定价问题中的定价因子; 此外, 研究者需要进一步探究这些文本向量的经济学含义与可解释性, 以便其更好地探究其预测能力的来源.

其次, 未来研究可探索 LLMs 主题模型如何优化主题分析的效果. 目前该应用场景的技术研究刚刚起步, 因此在经济金融领域内的应用研究也相对较少. 现有研究主要基于 BERTopic 模型进行文本主题分析, 还尚未出现基于 GPT 主题模型进行分析的经济和金融研究, 这是未来可以考虑尝试的应用方向. 其次, 研究者可以从 LLMs 主题模型的优势出发思考未来的研究改进. 例如, 由于传统 LDA 模型的计算速度较慢, 在处理大规模文本时往往需要消耗很长的时间, 因此在一些预测问题中, 研究者通常不会滚动或递归地使用主题模型, 而仅使用训练期起始点之前的文本数据训练 LDA, 之后再将训练好的模型用于样本外推断 (Larsen and Thorsrud, 2019; Ellingsen et al., 2022). 这一操作可能会损失掉大量文本信息. 相比于 LDA 等传统主题模型, LLMs 主题模型的计算速度有大幅提升, 因此可以考虑在一些预测任务中实现主题模型的滚动或递归, 以解决潜在的前瞻性偏误问题.

同时, 通过文献梳理, 我们发现大语言模型在我国经济学和金融学领域的应用仍处于初级阶段, 大多应用研究只是利用大语言模型执行一些简单的文本分析任务, 如计算文本相似度和文本分类等. 中国具有不同于西方世界的经济、政治、文化以及社会制度, 其中蕴含着具有独特价值的中国问题, 需要中国学者运用国际前沿的技术、方法和工具, 借鉴国际同行所使用的研究范式和方法, 讲述能够使国际同行理解的中国故事 (刘景江等, 2023). 因此, 未来中国学者需要重点关注大语言模型如何赋能具有中国特色的经济学与金融学问题研究.

首先, 研究者可以尝试在更多研究方向和更多样化的中文文本中利用 LLMs 计算的文本相似度来构建具有中国问题背景的新指标. 例如, 通过计算地方政府工作报告文本的相似度来测度地方政府的改革创新意识、计算国企披露文本的相似度来衡量国企改革力度.

其次, 研究者可以探索 LLMs 在中文文本中各种分类任务上的适用性, 例如基于中国各级政府采购合同文本对不同采购项目进行分类、根据“专精特新”企业的年报文本对其所属行业进行分类等.

此外, 研究者可以根据中国问题背景或特定研究领域构建更多的专业领域词典. 利用 LLMs 构建特定领域词典的现有研究相对较少, 但不同研究领域对于特定词典的需求较大,

尤其在中国问题背景下, 这一应用场景具有较大的研究空间。例如, 研究者可以尝试利用 LLMs 基于政府工作报告、企业年报等文本构建关于“新质生产力”和“企业数字化转型”等具有中国现实问题背景的专用词典。

最后, 文本情绪分析通常与语言特性紧密相关, 应该拓展 LLMs 对中文文本情绪分析的新思路。一方面, 可以探索利用 LLMs 分析投资者情绪以研究中国股票市场。作为全球第二大市值的股票市场, 中国股票市场的重要性与日俱增, 但与 LLMs 分析文本情绪相关的研究却很少关注中国股票市场。中国特有的政治、经济、文化和制度等因素使得对中国股票市场的研究具有较大挑战性。在这一独特背景下, 未来可以探索如何借助 LLMs 分析中文股吧、微博和中文财经新闻等文本以测度投资者情绪进而分析中国股票市场。另一方面, 可以尝试利用 LLMs 量化中国央行政策沟通。现有国外研究主要基于美联储和欧央行的英文文本训练 LLMs 以量化其政策沟通, 而中国人民银行的沟通文本具有其独特性和复杂性, 如果直接使用基于英文文本训练的模型, 可能难以取得较好的分析效果。因此, 如何训练或微调得到能够精准量化中国央行政策沟通的 LLMs 是一个重要的研究话题。

2) LLMs 开创的全新文本分析场景的研究前景

LLMs 在这一应用场景中所展现的人类对齐能力是其最核心且独有的能力, 与传统文本分析技术有着本质区别。这一场景下的研究正在飞速发展, 具有非常广阔的研究前景。在未来, 我们可以从行为科学和经济学等跨学科视角探索如何构建和利用人工智能经济人 (homo silicus, 另译作“硅基人”), 进而开启“AI 行为科学”领域的研究。人工智能经济人的使用方式与经济学家使用经济人 (homo economicus) 的方式相同: 因为大语言模型能够对各种文本输入做出高度接近人类的反应, 因此它们可以被赋予禀赋, 或被置于不同的情景中, 进而便于研究者探索其行为。不过在使用人工智能经济人的情况下, 这一过程是通过计算机模拟实现的, 而不是传统的数学推导 (Horton, 2023)。

LLMs 具备模仿人类并替代自然人参与调查或实验的潜力, 因此可以被视作人工智能经济人。未来的研究可以尝试训练 LLMs 表现出特定的行为特征, 从而作为智能体扮演不同背景的个体 (如不同收入水平的人群)。通过阅读网络媒体中的大规模文本内容, LLMs 能够生成具有时间、空间和个体多维度异质性的信念或预期。这一能力使其能够捕捉和探索特定时期、特定地区或特定群体的高频预期或情绪信息, 特别是在某些难以获取数据的情境下具有巨大应用空间。

此外, 在未来的研究中, 人工智能经济人还有可能会被用于模拟不同政策和经济环境变化对经济行为的影响。例如, 研究者可以通过调整输入文本中的政策信息, 观察人工智能经济人的不同反应, 从而预测实际政策实施后的潜在效果。这种模拟方法不仅速度快、成本低, 而且能够在实际政策实施前提供有价值的决策参考, 减少政策试错的风险。

3) 进一步开发适用于中文经济金融文本的 LLMs

正如前文所述, 已有研究表明, 使用特定领域 (如经济和金融领域) 的高质量文本数据对 LLMs 进行微调可以提高模型在该领域进行文本分析的性能。未来研究可考虑进一步开发和应用适用于中文经济金融文本的 LLMs, 这不仅是科学问题, 更是开拓细分市场的潜在增长点。

目前, 在华尔街投行中, AIGC 的应用非常热门。高盛集团、花旗集团等公司纷纷开发类

似 ChatGPT 的产品, 这些产品不仅能够帮助客户提高投资回报率, 还能够帮助员工节省大量时间, 提升工作效率. 例如摩根大通正式发布的 IndexGPT 就是基于 GPT-4 进行微调后的产品, 其利用特定主题关键词创建投资篮子主题索引, 然后通过 LLMs 进行深度分析, 扫描新闻等内容以生成投资指数, 展示公司的整体投资价值.

然而, 从现有研究来看, 主流的经济金融大语言模型均是使用英语为主的经济金融文本数据集训练得到的, 因为这些英文数据集已得到学界和业界的广泛认可, 且 FinBERT 等模型已广泛应用于英文的经济金融文本分析的学术研究中. 近期虽有一些中国学者尝试利用中文的经济金融文本数据训练得到了一些适用于中文的经济金融大语言模型, 但这些模型仍停留在技术开发阶段, 尚未在中国经济学和金融学的研究中得到广泛应用, 其能否理解一些中国特有的经济学和金融学问题有待进一步的研究考察.

由于经济和金融领域的数据可能涉及隐私问题, 目前公开可用的数据资源相对较少, 尤其缺乏大规模的预训练数据集 (赵鑫等, 2024) 和高质量的中文数据集, 因此通过训练或微调得到适用于该领域且能够理解中国经济金融问题的 LLMs 具有更高的挑战性. 因此, 在未来我们需要加快建设和补充高质量的中文经济金融文本数据集, 尤其是蕴含中式价值观和中国自主的经济学知识体系内涵的语料数据库, 以便开发得到真正适用于中国经济金融问题的 LLMs. 这是兼具学术创新机会和市场开拓机会的领域, 不仅有助于提升中国经济金融领域的基础研究能力, 还有利于推动“人工智能+”落地具体应用场景, 赋能产业升级与企业数字化转型.

5 结论

本文主要关注现有研究如何利用大语言模型对经济和金融领域的文本数据进行分析. 首先, 本文简要介绍了两类最具代表性的大语言模型 GPT 和 BERT 及一系列经济金融领域专用的 LLMs, 同时阐述了 LLMs 分析文本数据的基本原理, 并解释了使用 LLMs 分析经济金融文本的原因. 然后本文归纳总结了该领域内近期的前沿研究文献, 从两个方面全面介绍了 LLMs 在经济金融文本分析中的应用场景, 分别是: 计算文本相似度、提取文本向量进行预测、文本数据识别与分类、构建特定领域词典、主题建模与分析、文本情绪分析和模拟人类生成对于文本的信念或预期. 最后, 本文分别针对利用 LLMs 进行经济金融文本分析的局限性和争议展开讨论, 并根据现有前沿研究的不足之处对未来研究的发展方向和主要策略进行了前瞻性展望.

在大语言模型时代, 文本分析技术正在经历前所未有的变革, 而文本大数据在经济学和金融学领域的应用方兴未艾 (沈艳等, 2019), 因此大语言模型的迅猛发展势必会对经济学和金融学领域中的文本分析实证研究产生巨大的冲击. 本文的写作回应了当前相关研究领域的关注焦点与迫切需求. 通过本文, 我们期望能够帮助读者厘清相关概念与基本原理, 同时更加系统全面地把握现有前沿研究的脉络, 从而为研究者进一步探究大语言模型在经济金融文本分析中的潜力与局限性提供更加清晰地研究方向与策略建议, 最终推动大语言模型在经济学和金融学领域广泛应用, 促进相关学术研究的创新与进步.

参 考 文 献

- 车万翔, 郭江, 崔一鸣, (2021). 自然语言处理: 基于预训练模型的方法 [M]. 北京: 电子工业出版社.
- Che W X, Guo J, Cui Y M. Natural Language Processing: A Pretrained Model Approach[M]. Beijing: Publishing House of Electronics Industry.
- 陈晓红, 杨柠屹, 周艳菊, 曹文治, (2024). 数字经济时代 AIGC 技术影响教育与就业市场的研究综述 —— 以 ChatGPT 为例 [J]. 系统工程理论与实践, 44(1): 260–271.
- Chen X H, Yang N Y, Zhou Y J, Cao W Z, (2024). AIGC Affecting Education and Employment in Era of Digital Economy — Take ChatGPT as an Example[J]. Systems Engineering — Theory & Practice, 44(1): 260–271.
- 程兵, (2023). 以 ChatGPT 为代表的大语言模型打开了经济学和其他社会科学研究范式的巨大新空间 [J]. 计量经济学报, 3(3): 589–614.
- Cheng B, (2023). Artificial Intelligence Generative Content (AIGC) Including ChatGPT Opens a New Big Paradigm Space of Economics and Social Science Research[J]. China Journal of Econometrics, 3(3): 589–614.
- 范小云, 王业东, 王道平, 郭文璇, 胡焱翊, (2022). 不同来源金融文本信息含量的异质性分析 —— 基于混合式文本情绪测度方法 [J]. 管理世界, 38(10): 78–101.
- Fan X Y, Wang Y D, Wang D P, Guo W X, Hu X Y, (2022). Heterogeneity Analysis of Information Content for Financial Text from Different Sources: A Hybrid Text Sentiment Measurement Method[J]. Management World, 38(10): 78–101.
- 何勇, 焦丽, 杨艺, 祝怡菲, (2024). AI 大模型赋能金融市场量化投资? 基于另类数据与传统金融数据的研究 [J]. 计量经济学报, 4(3): 761–783.
- He Y, Jiao L, Yang Y, Zhu Y F, (2024). AI Large Language Model Empowers Quantitative Investment in Financial Markets? Research Based on Alternative Data and Traditional Financial Data[J]. China Journal of Econometrics, 4(3): 761–783.
- 何勇, 李琪琪, 焦丽, 黄文萱, (2023). 另类数据在中国股票市场投资中有用吗? —— 基于财经短视频、图像、文本数据的探究 [J]. 计量经济学报, 3(4): 1008–1031.
- He Y, Li Q Q, Jiao L, Huang W X, (2023). Is Alternative Data Useful in China Share Market Investment? — Empirical Study Based on Financial Short Video, Image and Text Data[J]. China Journal of Econometrics, 3(4): 1008–1031.
- 洪永淼, 汪寿阳, (2024). ChatGPT 与大模型将对经济学研究范式产生什么影响?[J]. 计量经济学报, 4(1): 1–25.
- Hong Y M, Wang S Y, (2024). How Will ChatGPT and Large Models Influence the Research Paradigm in Economics?[J]. China Journal of Econometrics, 4(1): 1–25.
- 洪永淼, 汪寿阳, (2021). 大数据如何改变经济学研究范式?[J]. 管理世界, 37(10): 40–55.
- Hong Y M, Wang S Y, (2021). How is Big Data Changing Economic Research Paradigms?[J]. Management World, 37(10): 40–55.
- 洪永淼, 汪寿阳, (2023). 人工智能新近发展及其对经济学研究范式的影响 [J]. 中国科学院院刊, 38(3): 353–357.
- Hong Y M, Wang S Y, (2023). Impacts of Cutting-edge Artificial Intelligence on Economic Research Paradigm[J]. Bulletin of the Chinese Academy of Sciences, 38(3): 353–357.
- 胡增玺, 马述忠, (2023). 市场一体化对企业数字创新的影响 —— 兼论数字创新衡量方法 [J]. 经济研究, 58(6): 155–172.
- Hu Z X, Ma S Z, (2023). The Impact of Market Integration on Firm Digital Innovation: An Analysis on Digital Innovation Measurement Methods[J]. Economic Research Journal, 58(6): 155–172.

- 姜富伟, 刘雨旻, 孟令超, (2024). 大语言模型、文本情绪与金融市场 [J]. 管理世界, 40(8): 42–64.
- Jiang F W, Liu Y M, Meng L C, (2024). Large Language Model and Textual Sentiment Analysis in Chinese Stock Markets[J]. Management World, 40(8): 42–64.
- 姜富伟, 孟令超, 唐国豪, (2021). 媒体文本情绪与股票回报预测 [J]. 经济学 (季刊), 21(4): 1323–1344.
- Jiang F W, Meng L C, Tang G H, (2021). Media Textual Sentiment and Chinese Stock Return Predictability[J]. China Economic Quarterly, 21(4): 1323–1344.
- 金星晔, 左从江, 方明月, 李涛, 聂辉华, (2024). 企业数字化转型的测度难题: 基于大语言模型的新方法与新发现 [J]. 经济研究, 59(3): 34–53.
- Jin X Y, Zuo C J, Fang M Y, Li T, Nie H H, (2024). Measurement Problem of Enterprise Digital Transformation: New Methods and Findings Based on Large Language Models[J]. Economic Research Journal, 59(3): 34–53.
- 李春涛, 闫续文, 张学人, (2024). GPT 在文本分析中的应用: 一个基于 Stata 的集成命令用法介绍 [J]. 数量经济技术经济研究, 41(5): 197–216.
- Li C T, Yan X W, Zhang X R, (2024). Application of GPT in Textual Analysis: An Introduction to a Community Developed Command within Stata[J]. The Journal of Quantitative & Technical Economics, 41(5): 197–216.
- 林建浩, 陈良源, 罗子豪, 张一帆, (2021). 央行沟通有助于改善宏观经济预测吗? —— 基于文本数据的高维稀疏建模 [J]. 经济研究, 56(3): 48–64.
- Lin J H, Chen L Y, Luo Z H, Zhang Y F, (2021). Does Central Bank Communication Improve Macroeconomic Forecasting? — High-dimensional Sparse Modeling Based on Text Data[J]. Economic Research Journal, 56(3): 48–64.
- 刘景江, 郑畅然, 洪永淼, (2023). 机器学习如何赋能管理学研究? —— 国内外前沿综述和未来展望 [J]. 管理世界, 39(9): 191–216.
- Liu J J, Zheng C R, Hong Y M, (2023). How Can Machine Learning Empower Management Research? A Domestic-Foreign Frontier Review and Future Prospects[J]. Management World, 39(9): 191–216.
- 刘青, 肖柏高, (2023). 劳动力成本与劳动节约型技术创新 —— 来自 AI 语言模型和专利文本的证据 [J]. 经济研究, 58(2): 74–90.
- Liu Q, Xiao B G, (2023). Labor Costs and Labor-saving Innovation: Evidence from AI Language Model and Patent Texts[J]. Economic Research Journal, 58(2): 74–90.
- 沈艳, 陈赞, 黄卓, (2019). 文本大数据分析在经济学和金融学中的应用: 一个文献综述 [J]. 经济学 (季刊), 18(4): 1153–1186.
- Shen Y, Chen Y, Huang Z, (2019). A Literature Review of Textual Analysis in Economic and Financial Research[J]. China Economic Quarterly, 18(4): 1153–1186.
- 汪寿阳, 李明琛, 杨昆, 林文灿, 姜尚荣, 等, (2023). ChatGPT+ 金融: 八个值得关注的研究方向与问题 [J]. 管理评论, 35(4): 3–11.
- Wang S Y, Li M C, Yang K, Lin W C, Jiang S R, et al. (2023). Chatgpt+Finance: Eight Noteworthy Research Directions and Issues[J]. Management Review, 35(4): 3–11.
- 许雪晨, 田侃, (2021). 一种基于金融文本情感分析的股票指数预测新方法 [J]. 数量经济技术经济研究, 38(12): 124–145.
- Xu X C, Tian K, (2021). A Novel Financial Text Sentiment Analysis-based Approach for Stock Index Prediction[J]. The Journal of Quantitative & Technical Economics, 38(12): 124–145.
- 姚加权, 张银澎, 罗平, (2020). 金融学文本大数据挖掘方法与研究进展 [J]. 经济学动态, (4): 143–158.
- Yao J Q, Zhang K P, Luo P, (2020). Text Mining in Financial Big Data and Its Research Progress[J]. Economic Perspectives, (4): 143–158.
- 赵鑫, 李军毅, 周昆, 唐天一, 文继荣, (2024). 大语言模型 [EB/OL]. [2024-07-25]. <https://llmbook-zh>.

- github.io.
- Zhao X, Li J Y, Zhou K, Tang T Y, Wen J R, (2024). The Chinese Book for Large Language Models[EB/OL]. [2024-07-25]. <https://llmbook-zh.github.io>.
- 郑晓龙, 白松冉, 曾大军, (2023). 面向复杂决策场景的认知图谱构建与分析 [J]. 管理世界, 39(5): 188–204.
- Zheng X L, Bai S R, Zeng D J, (2023). Construction and Analysis of Cognitive Graphs for Complex Decision-making Scenarios[J]. Management World, 39(5): 188–204.
- 诸竹君, 谢然成, 郭志芳, 余晓, (2023). 服务型制造技术创新与企业劳动要素市场势力 —— 基于 BERT 语言模型的微观证据 [J]. 中国工业经济, (12): 135–152.
- Zhu Z J, Xie R C, Guo Z F, Yu X, (2023). Service-oriented Manufacturing Technological Innovation and Firms' Labor Market Power: Micro-level Evidence Based on BERT Language Model[J]. China Industrial Economics, (12): 135–152.
- Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. (2023). GPT-4 Technical Report[J]. arXiv Preprint arXiv: 2303.08774.
- Agiza A, Mostagir M, Reda S, Agarwal S, Ahmad L, (2024). Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in LLMs[J]. arXiv Preprint arXiv: 2404.08699.
- Aher G V, Arriaga R I, Kalai A T, (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies[C]// International Conference on Machine Learning. PMLR: 337–371.
- Araci D, (2019). Finbert: Financial Sentiment Analysis with Pre-trained Language Models[J]. arXiv Preprint arXiv: 1908.10063.
- Ardekani A M, Bertz J, Bryce C, Dowling M, Long S C, (2024). Finsentgpt: A Universal Financial Sentiment Engine?[J]. International Review of Financial Analysis: 103291.
- Argyle L P, Busby E C, Fulda N, Gubler J R, Rytting C, (2023). Out of One, Many: Using Language Models to Simulate Human Samples[J]. Political Analysis, 31(3): 337–351.
- Ash E, Hansen S, (2023). Text Algorithms in Economics[J]. Annual Review of Economics, 15: 659–688.
- Barbaglia L, Consoli S, Manzan S, (2023). Forecasting with Economic News[J]. Journal of Business & Economic Statistics, 41(3): 708–719.
- Barbaglia L, Consoli S, Wang S, (2021). Financial Forecasting with Word Embeddings Extracted from News: A Preliminary Analysis[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing: 179–188.
- Beckmann L, Beckmeyer H, Filippou I, Menze S, Zhou G, (2024). Unusual Financial Communication-evidence from ChatGPT, Earnings Calls, and the Stock Market[J]. Available at SSRN: 4699231.
- Bhatia G, Nagoudi E M B, Cavusoglu H, et al. (2024). Fintral: A Family of GPT-4 Level Multimodal Financial Large Language Models[J]. arXiv Preprint arXiv: 2402.10986.
- Blei D M, Ng A Y, Jordan M I, (2003). Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 3(1): 993–1022.
- Born B, Dalal H, Lamersdorf N, Steffen S, (2023). Monetary Policy in the Age of Social Media: A Twitterbased Inflation Analysis[R]. Available at SSRN: 4567305.
- Breitung C, Müller S, (2023). Global Business Similarity Networks[J]. Available at SSRN 4395079.
- Brown P F, Della Pietra V J, Desouza P V, Lai J C, Mercer R L, (1992). Class-based N-gram Models of Natural Language[J]. Computational linguistics, 18(4): 467–480.
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan J D, et al. (2020). Language Models are Few-shot Learners[J]. Advances in Neural Information Processing Systems, 33: 1877–1901.
- Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, et al. (2023). Sparks of Artificial General Intelligence: Early Experiments with gpt-4[J]. arXiv Preprint arXiv: 2303.12712.

- Bybee J L, (2023). The Ghost in the Machine: Generating Beliefs with Large Language Models[J]. arXiv Preprint arXiv: 2305.02823.
- Cao S, Jiang W, Yang B, Zhang A L, (2023). How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI[J]. *The Review of Financial Studies*, 36(9): 3603–3642.
- Chava S, Du W, Malakar B, (2021a). Do Managers Walk the Talk on Environmental and Social Issues?[R]. Georgia Tech Scheller College of Business Research Paper: 3900814.
- Chava S, Du W, Paradkar N, (2021b). More than Buzzwords? Firms' Discussions of Emerging Technologies in Earnings Conference Calls[J]. Available at SSRN 3862645.
- Chen J, Tang G, Zhou G, Zhu W, (2023a). Chatgpt, Stock Market Predictability and Links to the Macroeconomy[J]. Available at SSRN: 4660148.
- Chen W, Wang Q, Long Z, Zhang X, Lu Z, et al., (2023b). Disc-finllm: A Chinese Financial Large Language Model Based on Multiple Experts Fine-tuning[J]. arXiv Preprint arXiv: 2310.15205.
- Chen Y, Fang H, Zhao Y, Zhao Z, (2024). Recovering Overlooked Information in Categorical Variables with LLMs: An Application to Labor Market Mismatch[R]. National Bureau of Economic Research: w32327.
- Chen Y, Kelly B T, Xiu D, (2023c). Expected Returns and Large Language Models[R]. Available at SSRN: 4416687.
- Chen Y, Liu T X, Shan Y, Zhong S F, (2023d). The Emergence of Economic Rationality of GPT[J]. *Proceedings of the National Academy of Sciences*, 120(51): E2316205120.
- Devlin J, Chang M W, Lee K, Toutanova K, (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv Preprint arXiv: 1810.04805.
- Dong M M, Stratopoulos T C, Wang V X, (2024). A Scoping Review of ChatGPT Research in Accounting and Finance[J]. *International Journal of Accounting Information Systems*, 55: 100715.
- Dong Z, Tang T, Li L, Zhao W X, (2023). A Survey on Long Text Modeling with Transformers[J]. arXiv Preprint arXiv: 2302.14502
- Drinkall F, Rahimikia E, Pierrehumbert J B, Zohren S, (2024). Time Machine GPT[J]. arXiv Preprint arXiv: 2404.18543.
- Eisfeldt A L, Schubert G, (2024). AI and Finance[R]. National Bureau of Economic Research, No. w33076.
- Ellingsen J, Larsen V H, Thorsrud L A, (2022). News Media Versus FRED-MD for Macroeconomic Forecasting[J]. *Journal of Applied Econometrics*, 37(1): 63–81.
- Eloundou T, Manning S, Mishkin P, Rock D, (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models[J]. arXiv Preprint arXiv: 2303.10130.
- Floridi L, Chiriatti M, (2020). GPT-3: Its Nature, Scope, Limits, and Consequences[J]. *Minds and Machines*, 30: 681–694.
- Gao C, Lan X, Li N, Yuan Y, Ding J, et al. (2024). Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives[J]. *Humanities and Social Sciences Communications*, 11(1): 1–24.
- Gentzkow M, Kelly B, Taddy M, (2019). Text as Data[J]. *Journal of Economic Literature*, 57(3): 535–574.
- Gilardi F, Alizadeh M, Kubli M, (2023). Chatgpt Outperforms Crowd Workers for Text-annotation Tasks[J]. *Proceedings of the National Academy of Sciences*, 120(30): E2305016120.
- Gorodnichenko Y, Pham T, Talavera O, (2023). The Voice of Monetary Policy[J]. *American Economic Review*, 113(2): 548–584.
- Grootendorst M, (2022). Bertopic: Neural Topic Modeling with a Class-based TF-IDF Procedure[J]. arXiv Preprint arXiv: 2203.05794.
- Gupta A, Rawte V, Zaki M J, (2023). Predicting Firm Financial Performance from SEC Filing Changes Using Automatically Generated Dictionary[J]. *Computational Economics*, 64: 1–28.

- Hagendorff T, (2024). Deception Abilities Emerged in Large Language Models[J]. Proceedings of the National Academy of Sciences, 121(24): E2317967121.
- Hajek P, Henriques R, (2024). Predicting M&A Targets Using News Sentiment and Topic Detection[J]. Technological Forecasting and Social Change, 201: 123270.
- Hansen A L, Kazinnik S, (2023). Can Chatgpt Decipher Fedspeak?[J]. Available at SSRN 4399406.
- Hansen S, Lambert P J, Bloom N, Davis S J, Sadun R, (2023). Remote Work across Jobs, Companies, and Space[R]. National Bureau of Economic Research: w31007.
- Hoberg G, Phillips G, (2010). Product Market Synergies and Competition in Mergers and Acquisitions: A Text-based Analysis[J]. The Review of Financial Studies, 23(10): 3773–3811.
- Hoberg G, Phillips G, (2016). Text-based Network Industries and Endogenous Product Differentiation[J]. Journal of Political Economy, 124(5): 1423–1465.
- Hofmann T, (2000). Learning the Similarity of Documents: An Information-geometric Approach to Document Retrieval and Categorization[J]. Advances in Neural Information Processing Systems, 12: 914–920.
- Horton J J, (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?[R]. National Bureau of Economic Research: w31122.
- Huang A H, Wang H, Yang Y, (2023a). Finbert: A Large Language Model for Extracting Information from Financial Text[J]. Contemporary Accounting Research, 40(2): 806–841.
- Huang Y, Xu J, Lai J, Jiang Z, Chen T, et al. (2023b). Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey[J]. arXiv Preprint arXiv: 2311.12351.
- Jha M, Qian J, Weber M, Yang B, (2024). Chatgpt and Corporate Policies[R]. National Bureau of Economic Research: w32161.
- Ji Z, Lee N, Frieske R, Yu T, Su D, et al. (2023). Survey of Hallucination in Natural Language Generation[J]. ACM Computing Surveys, 55(12): 1–38.
- Kalamara E, Turrell A, Redl C, Kapetanios G, Kapadia S, (2022). Making Text Count: Economic Forecasting Using Newspaper Text[J]. Journal of Applied Econometrics, 37(5): 896–919.
- Ke Z T, Kelly B T, Xiu D, (2019). Predicting Returns with Text Data[R]. National Bureau of Economic Research, No.w26186.
- Keles O, Bayraklı O T, (2024). LLaMA-2-Econ: Enhancing Title Generation, Abstract Classification, and Academic Q&A in Economic Research[C]// Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing@ LREC-COLING: 212–218.
- Kim A, Muhn M, Nikolaev V V, (2024). Bloated Disclosures: Can Chatgpt Help Investors Process Information?[J]. Chicago Booth Research Paper, (23–07): 2023–2059.
- Konstantinidis T, Iacovides G, Xu M, Constantinides T G, Mandic D, (2024). Finllama: Financial Sentiment Classification for Algorithmic Trading Applications[J]. arXiv Preprint arXiv: 2403.12285.
- Korinek A, (2023). Generative AI for Economic Research: Use Cases and Implications for Economists[J]. Journal of Economic Literature, 61(4): 1281–1317.
- Krishna K, Ramprasad S, Gupta P, Wallace B C, Lipton Z C et al. (2024). Genaudit: Fixing Factual Errors in Language Model Outputs with Evidence[J]. arXiv Preprint arXiv: 2402.12566.
- Kuang W, Qian B, Li Z, Chen D, Gao D, et al. (2024). Federatedscope-llm: A Comprehensive Package for Fine-tuning Large Language Models in Federated Learning[C]// Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining: 5260–5271.

- Larsen V H, Thorsrud L A, (2019). The Value of News for Economic Developments[J]. *Journal of Econometrics*, 210(1): 203–218.
- Lasri K, de Castro P V Q, Schirmer M, San Martin L E, Wang L, et al. (2023). Econberta: Towards Robust Extraction of Named Entities in Economics[C]// *Findings of the Association for Computational Linguistics*: 11557–11577.
- Laurençon H, Saulnier L, Wang T, Akiki C, Villanova del Moral A, et al. (2022). The Bigscience Roots Corpus: A 1.6TB Composite Multilingual Dataset[J]. *Advances in Neural Information Processing Systems*, 35: 31809–31826.
- Leng Y, Yuan Y, (2024). Do LLM Agents Exhibit Social Behavior?[J]. *arXiv Preprint arXiv*: 2312.15198.
- Li J, Bian Y, Wang G, Lei Y, Cheng D, et al. (2023a). Cfgpt: Chinese Financial Assistant with Large Language Model[J]. *arXiv Preprint arXiv*: 2309.10654.
- Li N, Gao C, Li Y, Liao Q, (2024). Econagent: Large Language Model-empowered Agents for Simulating Macroeconomic Activities[J]. *arXiv Preprint arXiv*: 2310.10436v3.
- Li Y, Wang S, Ding H, Chen H, (2023b). Large Language Models in Finance: A Survey[C]// *Proceedings of the Fourth ACM International Conference on AI in Finance*: 374–382.
- Liu Y, Ott M, Goyal N, Du J, Joshi M, et al. (2019). Roberta: A Robustly Optimized BERT Pretraining Approach[J]. *arXiv Preprint arXiv*: 1907.11692.
- Liu Z, Huang D, Huang K, Li Z, Zhao J, (2021). Finbert: A Pre-trained Financial Language Representation Model for Financial Text Mining[C]// *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*: 4513–4519.
- Lopez-Lira A, Tang Y, (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models[J]. *arXiv Preprint arXiv*: 2304.07619.
- Loughran T, McDonald B, (2011). When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks[J]. *The Journal of Finance*, 66(1): 35–65.
- Lundberg S M, Lee S I, (2017). A Unified Approach to Interpreting Model Predictions[J]. *arXiv Preprint arXiv*: 1705.07874.
- Luo H, Specia L, (2024). From Understanding to Utilization: A Survey on Explainability for Large Language Models[J]. *arXiv Preprint arXiv*: 2401.12874.
- Mehra S, Louka R, Zhang Y, (2022). ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices[J]. *arXiv Preprint arXiv*: 2203.16788, 2022.
- Meng J, (2024). AI Emerges as the Frontier in Behavioral Science[J]. *Proceedings of the National Academy of Sciences*, 121(10): E2401336121.
- Mikolov T, Sutskever I, Chen K, Corrado G S, Dean J, (2013). Distributed Representations of Words and Phrases and Their Compositionality[J]. *Advances in Neural Information Processing Systems*, 26.
- Nie Y, Kong Y, Dong X, Mulvey J M, Poor H V, et al. (2024). A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges[J]. *arXiv Preprint arXiv*: 2406.11903.
- Pawar S, Tonmoy S M, Zaman S M, Jain V, Chadha A, et al. (2024). The What, Why, and How of Context Length Extension Techniques in Large Language Models — A Detailed Survey[J]. *arXiv Preprint arXiv*: 2401.07872.
- Peskoff D, Visokay A, Schulhoff S, Wachspress B, Blinder A, et al. (2023). GPT Deciphering FedSpeak: Quantifying Dissent Among Hawks and Doves[C]// *Findings of the Association for Computational Linguistics*: 6529–6539.
- Pfeifer M, Marohl V P, (2023). Centralbankroberta: A Fine-tuned Large Language Model for Central Bank Communications[J]. *The Journal of Finance and Data Science*, 9: 100114.

- Pham C M, Hoyle A, Sun S, Resnik P, Iyyer M, (2023). Topicgpt: A Prompt-based Topic Modeling Framework[J]. arXiv Preprint arXiv: 2311.01449.
- Pham V, Cunningham S, (2024). Chatgpt Can Predict the Future When it Tells Stories Set in the Future About the Past[J]. arXiv Preprint arXiv: 2404.07396.
- Radford A, Narasimhan K, Salimans T, Sutskever I, (2018). Improving Language Understanding by Generative Pre-training[R/OL]. [2024-07-25]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- Reuter A, Thielmann A, Weisser C, Fischer S, Säfken B, (2024). Gptopic: Dynamic and Interactive Topic Representations[J]. arXiv Preprint arXiv: 2403.03628.
- Sanh V, Debut L, Chaumond J, Wolf T, (2019). Distilbert, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter[J]. arXiv Preprint arXiv: 1910.01108.
- Sarkar S K, Vafa K, (2024). Lookahead Bias in Pretrained Language Models[J]. Available at SSRN 4754678.
- Scao T L, Fan A, Akiki C, Pavlick E, Ilić S, et al. (2022). BLOOM: A 176B-Parameter Open-access Multilingual Language Model[J]. arXiv Preprint arXiv: 2211.05100.
- Tan L, Wu H, Zhang X, (2023). Large Language Models and Return Prediction in China[J]. Available at SSRN 4712248.
- Tang W, Bu H, Zuo Y, Wu J, (2024). Unlocking the Power of the Topic Content in News Headlines: BERTopic for Predicting Chinese Corporate Bond Defaults[J]. Finance Research Letters: 105062.
- Todt P B W, Babaei R, (2023). FIN-LLAMA: Efficient Finetuning of Quantized LLMs for Finance[EB/OL]. [2024-07-24]. <https://github.com/Bavest/fin-llama>.
- Tong H, Li J, Wu N, Gong M, Zhang D, et al. (2024). Plutos: Towards Interpretable Stock Movement Prediction with Financial Large Language Model[J]. arXiv Preprint arXiv: 2403.00782.
- Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, et al. (2023a). Llama: Open and Efficient Foundation Language Models[J]. arXiv Preprint arXiv: 2302.13971.
- Touvron H, Martin L, Stone K, Albert P, Almahairi A, et al. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models[J]. arXiv Preprint arXiv: 2307.09288.
- Tu X, Zou J, Su W, Zhang L, (2024). What Should Data Science Education Do With Large Language Models?[J]. Harvard Data Science Review, 6(1).
- Vamvourellis D, Toth M, Bhagat S, Desai D, Mehta D, et al. (2023). Company Similarity Using Large Language Models[J]. arXiv Preprint arXiv: 2308.08031.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. (2017). Attention is All You Need[J]. Advances in Neural Information Processing Systems, 30.
- Wang Y, Zhong W, Li L, Mi F, Zeng X, et al. (2023a). Aligning Large Language Models with Human: A Survey[J]. arXiv Preprint arXiv: 2307.12966.
- Wang Z, Li Y, Wu J, Soon J, Zhang X, (2023b). Finvis-gpt: A Multimodal Large Language Model for Financial Chart Analysis[J]. arXiv Preprint arXiv: 2308.01430.
- Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, et al. (2022a). Emergent Abilities of Large Language Models[J]. arXiv Preprint arXiv: 2206.07682.
- Wei J, Wang X, Schuurmans D, Bosma M, Xia F, et al. (2022b). Chain-of-thought Prompting Elicits Reasoning in Large Language Models[J]. Advances in Neural Information Processing Systems, 35: 24824–24837.
- Winata G I, Madotto A, Lin Z, Liu R, Yosinski J, et al. (2021). Language Models are Few-shot Multilingual Learners[J]. arXiv Preprint arXiv: 2109.07684.
- Wu S, Irsoy O, Lu S, Dabrowski V, Dredze M, et al. (2023). Bloomberggpt: A Large Language Model for Finance[J]. arXiv Preprint arXiv: 2303.17564.

- Xie Q, Han W, Zhang X, Lai Y, Peng M, et al. (2023). Pixiu: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance[J]. arXiv Preprint arXiv: 2306.05443.
- Yang H, Liu X Y, Wang C D, (2023a). Fingpt: Open-source Financial Large Language Models[J]. arXiv Preprint arXiv: 2306.06031.
- Yang H, Zhang B, Wang N, Guo C, Zhang X, et al. (2024). Finrobot: An Open-Source AI Agent Platform for Financial Applications Using Large Language Models[J]. arXiv Preprint arXiv: 2405.14767.
- Yang S, (2023). Predictive Patentomics: Forecasting Innovation Success and Valuation with ChatGPT[J]. arXiv Preprint arXiv: 2307.01202.
- Yang Y, Tang Y, Tam K Y, (2023b). Investlm: A Large Language Model for Investment Using Financial Domain Instruction Tuning[J]. arXiv Preprint arXiv: 2309.13064.
- Yang Y, Uy M C S, Huang A, (2020). Finbert: A Pretrained Language Model for Financial Communications[J]. arXiv Preprint arXiv: 2006.08097.
- Yang Y, Zhang K, Fan Y, (2022). Analyzing Firm Reports for Volatility Prediction: A Knowledge-driven Text-embedding Approach[J]. INFORMS Journal on Computing, 34(1): 522–540.
- Yao J Y, Ning K P, Liu Z H, Ning M N, Liu Y Y, et al. (2023). LLM Lies: Hallucinations are Not Bugs, but Features as Adversarial Examples[J]. arXiv Preprint arXiv: 2310.01469.
- Yu J, Huber M, Tang K, (2024). Greedllama: Performance of Financial Value-Aligned Large Language Models in Moral Reasoning[J]. arXiv Preprint arXiv: 2404.02934.
- Yu Y, Li H, Chen Z, Jiang Y, Li Y, et al. (2024). Finnem: A Performance-enhanced LLM Trading Agent with Layered Memory and Character Design[C]// Proceedings of the AAAI Symposium Series, 3(1): 595–597.
- Zarifhonarvar A, (2024). Experimental Evidence on Large Language Models[J]. Available at SSRN 4825076.
- Zhang B, Yang H, Liu X Y, (2023). Instruct-fingpt: Financial Sentiment Analysis by Instruction Tuning of General-purpose Large Language Models[J]. arXiv Preprint arXiv: 2306.12659.
- Zhang W, Zhao L, Xia H, Sun S, Sun J, et al. (2024). A Multimodal Foundation Agent for Financial Trading: Toolaugmented, Diversified, and Generalist[C]// Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining: 4314–4325.
- Zhang X, Yang Q, (2023). Xuanyuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters[C]// Proceedings of the 32nd ACM International Conference on Information and Knowledge Management: 4435–4439.
- Zhang Z, Zhang H, Chen K, Guo Y, Hua J, et al. (2021). Mengzi: Towards Lightweight Yet Ingenious Pre-trained Models for Chinese[J]. arXiv Preprint arXiv: 2110.06696.
- Zhao H, Chen H, Yang F, Liu N, Deng H, et al. (2024). Explainability for Large Language Models: A Survey[J]. ACM Transactions on Intelligent Systems and Technology, 15(2): 1–38.
- Zhao W X, Zhou K, Li J, Tang T, Wang X, et al. (2023). A Survey of Large Language Models[J]. arXiv Preprint arXiv: 2303.18223.
- Zhou Y, Li Z, Tian S, Ni Y, Liu S, et al. (2024). Silversight: A Multi-Task Chinese Financial Large Language Model Based on Adaptive Semantic Space Learning[J]. arXiv Preprint arXiv: 2404.04949.