大语言模型和人工智能代理在量化投资与交易中的应用[1,2,3]

周鸿松

香港科技大学金融数学实践教授、
晨晗咨询（香港）有限公司董事

1. 简介

在所有被视为人工智能（AI）早期采用者的行业中，金融行业是最早的一批。量化投资与交易，作为金融的子领域，在硬件系统（例如用于衍生品定价的 GPU）和软件工具（例如用于价格预测的神经网络）方面，甚至在它们成为主流热点之前就开始进行探索。量化金融领域如此早地采用 AI 的原因有二：其一，量化金融研究人员和从业者积累的数据通常结构良好（例如高频订单簿数据），因此相对而言，将新开发的 AI 模型应用于解决实际问题更为直接；其二，量化投资者和交易者始终在寻找创新方式以开展收益预测和风险管理实践，因此他们始终有动力尝试将新的量化模型纳入其投资策略中。然而，许多量化投资者和交易者也是 AI 的务实使用者，如果对 AI 研究和基础设施建设的投资不能相对迅速地得到证明，他们可能会对基于 AI 的模型和基础设施持批判态度。

自 2022 年底 OpenAI 推出 ChatGPT 产品以来，随着生成式 AI 的进步，许多量化投资者和交易者对这些新工具的潜力感到兴奋，它们有望帮助实现更好的成果，例如更准确的价格预测模型和更强大的研究与交易基础设施。这种最新的生成式 AI 发展，实际上是近年来对量化投资和交易实践产生影响的第二波 AI 浪潮，第一波是基于机器学习和深度学习的函数包的研究和交易基础设施的采用，但值得思考的是，尽管发展仍处于早期阶段，这些基础设施通常用 Python 等 AI 友好型编程语言实现，这第二波以大语言模型（LLMs）和人工智能代理（AI Agents）为代表的 AI 潮流是如何成型的，以及其可能如何进一步影响量化投资和交易实践。

[1] 本文基于作者于 2025 年 4 月 24 日至 25 日在香港举办的芝加哥大学“量化金融中的人工智能研讨会”上所作的演讲。

[2] 本文将“量化交易” —— 在广义上 —— 视为一种通常具有中频或高频性质的量化投资策略。因此，本文中“量化投资”和“量化交易”这两个术语可以互换使用。

[3] 免责声明：本文中表达的观点仅代表作者个人观点，不一定与 ECC Info 的观点一致；本文的任何内容均不能被视为作者和 / 或 ECC Info 提供的投资建议或建议；所有计算机代码仅用于说明目的。

2. 大语言模型与财务预测

大语言模型（LLMs）是基于深度学习领域内序列模型的新一代自然语言处理（NLP）模型，许多 AI 技术在该领域内开发出来以处理和理解人类语言。虽然这通常作为行业机密，但据估计，行业级 NLP 模型（例如智能手机、车载导航系统等中的语音识别工具）拥有的参数数量通常在数亿到数十亿之间。相比之下，据估计 OpenAI 的生成式 AI LLM ChatGPT 可能拥有高达 1.7 万亿个参数，而 Meta 开源的 LLM 平台 Llama - 2 可能拥有高达 700 亿个参数。对 ChatGPT 等 LLM 成功至关重要的一个技术发展是 Transformer 模型，该模型最初由谷歌开发，并已得到极大扩展和改进，能够处理比原始谷歌论文描述的更复杂的自然语言处理问题。Transformer 技术发展的关键结果之一是 LLMs 逐渐形成的“学习”或“推理”能力。DeepSeek R1 模型就是这种“推理”能力的最佳例证，用户可以观察到该模型如何推理并形成对用户提出的问题（或一系列问题）的回答。

量化投资与交易的一个基本假设可以描述为“历史会重演”：在量化投资者 / 交易者能够充分探索的信息集中，如果一个“模式”在该信息集中以高统计置信度出现（例如两个变量之间存在统计上显著的关系，并且其中至少一个变量是可投资的），投资者 / 交易者通常会假设这种模式将在未来数据中再次出现，并可被用来构建用于实际投资和交易活动的预测模型。许多财务数据属于时间序列类型，表明数据具有序列性且存在时间因果关系。这种因果效应也出现在带有时间因果关系的人类推理过程中。因此，自然会产生这样的想法：LLM 是否能够帮助识别这种“模式”，就像量化分析师一直使用的许多统计推断工具一样。

与其他统计工具一样，LLMs 确实可以帮助量化分析师从历史数据中寻找此类“模式”。然而，由于 LLMs 最擅长处理带有上下文的人类语言，因此构建实际预测模型的最简单方法是使用人类语言序列排列的历史数据来训练 LLM。图 1 展示了一种所谓的“少量样本学习”方法，用于训练 LLM 从历史数据中学习，然后对新的提示给出预测答案。该示例表明，当以人类语言的问题和答案序列形式呈现时，LLM 可以使用历史时间序列数据来检测模式，并利用该模式产生预测结果。

虽然上述例子是否是有效利用高度结构化数据来预测未来价格的方式还存在争议，但如果将用于训练统计模型的相同时间序列数据输入到 LLM 中， 看看 LLM 是否也能产生与其他统计模型相比，质量相似或更好的预测结果，这非常有趣 。鉴于 LLM 通常被认为与统计模型（例如线性回归）具有大相径庭的结构，量化分析师可能会对这种新方法感兴趣，特别是当 LLM 方法与传统统计方法在某种程度上 “正交”（orthogonal）时。

图 1: 使用“少量样本学习”方法训练 LLM 以利用人类语言提示预测股票收益的示例。

```
#Example 3: leading-lagging effects on Nvidia
def fewShotLearningExample3(client):
    model = "gpt-4o-mini"
    messages = [
        {
         "role": "user",
         "content": "AMZN went up 1%;"
                    "TSLA went up 4%;"
                    "Did Nvidia go up or down? By how much?"
        },
        {
         "role": "assistant",
         "content":  "Nvidia went up 2%."
        },
        {
         "role": "user",
         "content": "AAPL went up 4%;"
                    "MSFT went up 2%;"
                    "Did Nvidia go up or down? By how much?"
        },
        {
         "role": "assistant",
         "content":  "Nvidia went up 0.5%."
        },
          ⋮
        {
         "role": "user",
         "content": "GOOGL went down 1%;"
                    "AMZN went up 4.5%;"
                    "Did Nvidia go up or down? By how much?"
        },
        ]

    response = client.chat.completions.create(
        model=model,
        messages=messages,
        temperature=0.0,
    )

    return response
```

3. 大语言模型与因子投资

因子投资是组合投资的常见方法，已有数十年历史。该方法基于有效市场假说（EMH）。EMH 认为，资产价格反映了市场参与者所拥有的信息内容，参与者对资产的定价不能超出他们的信息内容；因此，如果市场拥有的信息内容与每个单独参与者相同（即市场在资产信息方面完全透明），则没有参与者能够战胜市场。在现实世界中，市场参与者只能近似地拥有市场所拥有的信息量；从数学上讲，这意味着市场参与者可以基于资产对市场所展示的各种“因子”的“敞口”来对资产价格进行建模，如下面的方程（1）所示。

$$\mathbb{E}(R_i) = a_i + \Sigma_{k=1}^{K} \beta_{i,k} \mathbb{E}(F_k) + \epsilon_i \quad (1)$$

其中，$\mathbb{E}(R_i)$ 是第 i 个资产的预期收益，$\mathbb{E}(F_k)$ 是第 k 个因子的期望值。$\beta_{i,k}$ 通常被称为第 i 个资产对第 k 个因子的敞口，而 a_i 可以被视为第 i 个资产的回归截距，ϵ_i 则是第 i 个资产回归的噪声项。请注意，因子通常由研究市场的量化分析师识别。量化投资者和交易者常用的常见因子通常来自基本面因子、技术因子、流动性因子、风险因子、情绪因子等。一旦确定了这些因子的值，就可以使用公式 (1) 通过在资产的下一期回报数据与资产的当前和 / 或历史因子值之间建立线性回归模型来预测资产回报。

随着机器学习（ML）和深度学习（DL）工具的更广泛应用，量化分析师除了构建线性回归模型外，还常常构建 ML 或 DL 模型来进行收益预测。图 2 提供了样本因子值数据和股票未来收益数据，两者均为时间序列数据。因子值通常被用作“特征”，而未来收益通常被用作“标签”。如今，将特征输入到 ML/DL 模型中，并优化模型以获得预定义指标的最小值（该指标衡量预测收益与标签之间的距离）是一种常见做法。该过程通常涉及许多技术细节，如训练、验证、测试、特征选择与工程、消融测试等。

随着 LLMs 的出现，自然会考虑 LLMs 如何像现有的 ML 和 DL 平台一样帮助进行收益预测任务。鉴于 LLMs 被训练为理解人类语言以及用户与 LLM 之间的对话上下文，因此将图 2 中所示的结构化数据转换为“问题和答案”序列，并使用所谓的“少量样本学习”方法来训练 LLM 进行收益预测是有意义的。在许多方面，这种“因子 - Q&A”转换类似于图 1 中所示的示例，唯一的区别在于因子模型案例涉及对投资组合中所有股票的同时预测。借助 OpenAI 的 gpt - 4o（一种 LLM），我们能够测试使用 LLM 进行此类预测的方法，并将其预测结果与其他模型进行比较。结果显示，使用相对简单的 LLM 回测安排，提示的 LLM 为股票月度收益预测生成了 0.0141 的信息系数。相比之下，使用相同因子数据和回测设置的线性回归给出了 0.0135 的信息系数。我们相信，随着对 gpt - 4o 模型的预测参数进行更多微调（由于我们使用的 gpt - 4o 模型的令牌数量限制），LLM 方法具有进一步训练以获得更好结果的潜力。尽管如此，将因子模型数据应用于 LLMs 进行回报预测的初步结果还是令我们感到鼓舞。

值得一提的是，已有文章讨论了 LLM 如何用于进行资产价格收益预测，例如基于 LLM 分析新闻数据的情绪交易信号等。对于传统的基于因子的投资策略，情绪通常被视为一种可以与其他因子结合构建预测模型的因子。上述具有初步结果的方法表明，LLM 可以与情绪以外的其他因子一起使用。

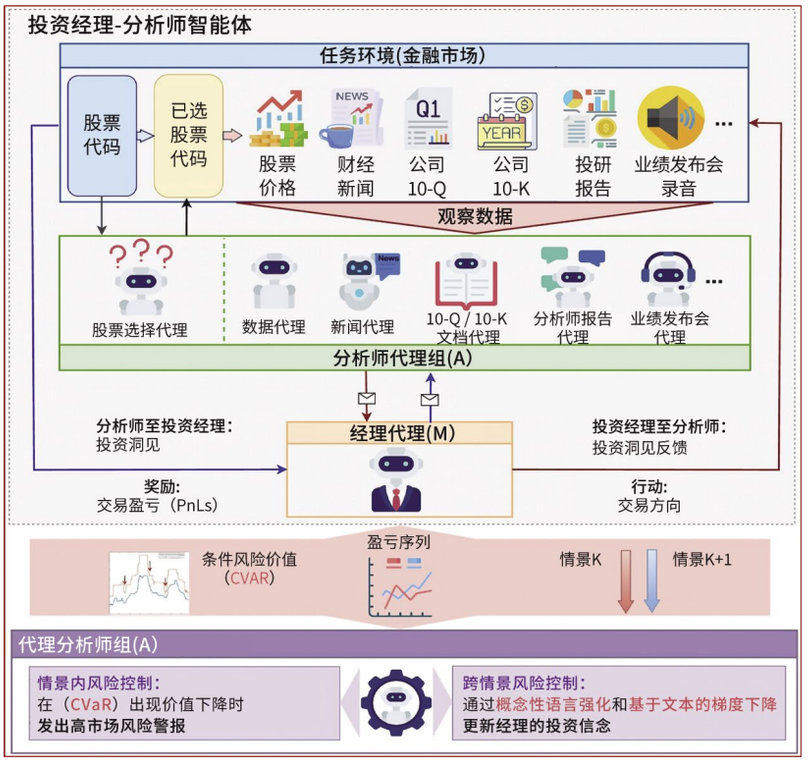
4. 大语言模型和人工智能代理对量化投资与交易的影响

尽管 LLMs 对许多量化投资者和交易者来说仍然相对较新，但已经有关于它们如何帮助改进量化投资与交易实践的讨论。除了上述使用 LLMs 和因子数据进行收益预测的例子外，LLM 驱动

图 2：样本因子值数据。

Date	Stock	return	factor0	factor1	factor2	factor3	factor4	factor5	factor6
1/17/2016	s030178	-0.073296	-0.870595	0.595826	-0.547213	0.578947	0.500000	0.333333	-26.831818
1/17/2016	s031103	-0.077591	-0.961591	-0.840146	-0.475808	0.368421	0.444444	0.277778	-83.330313
1/17/2016	s032349	0.040172	-0.872872	-0.685289	-0.603744	0.105263	0.388889	0.888889	-62.500000
1/17/2016	s032417	-0.142062	-0.544862	-0.486238	0.444499	0.947368	0.611111	0.555556	-70.010005
1/17/2016	s032595	-0.078917	-0.945611	-0.762440	-0.390260	0.263158	0.277778	0.388889	-59.994309
1/17/2016	s033025	-0.140298	-0.704588	-0.544040	-0.454021	0.421053	0.666667	0.611111	-45.650357
1/17/2016	s036348	-0.038072	-0.798549	-0.700183	-0.531638	0.052632	0.722222	0.666667	-92.105633
1/17/2016	s037954	-0.038463	-0.038391	-0.085787	-0.169622	0.526316	0.555556	0.444444	-85.713481
1/17/2016	s044851	-0.019024	-0.772756	-0.433683	-0.410546	0.631579	0.055556	0.111111	-87.156609
1/17/2016	s045428	-0.064403	-0.985007	-0.457822	-0.310600	1.000000	0.888889	1.000000	-76.922706
1/17/2016	s047438	0.019364	-0.866620	-0.651674	-0.358468	0.789474	0.333333	0.222222	-71.428571
1/17/2016	s047593	-0.079305	-0.628971	-0.704191	-0.531438	0.157895	0.944444	0.944444	-78.947903
1/17/2016	s082343	-0.131399	-0.966012	-0.846198	-0.674174	0.894737	0.166667	0.055556	0.000000
1/17/2016	s088124	-0.104753	0.838525	-0.151860	-0.296260	0.736842	0.111111	0.166667	-30.008352
1/17/2016	s000815	0.080729	-0.701756	-0.698688	-0.196433	0.473684	1.000000	0.833333	-5.263585
1/17/2016	s000982	-0.030818	-0.102490	-0.534912	-0.301932	0.842105	0.833333	0.777778	-58.334681
1/17/2016	s001239	-0.093252	-0.512272	-0.433202	-0.557644	0.210526	0.777778	0.722222	-58.623020
1/17/2016	s000467	-0.163342	-0.500122	-0.653268	-0.649318	0.315789	0.222222	0.500000	-36.364622
2/15/2016	s030178	-0.104855	0.000000	0.065906	0.000000	0.684211	0.473684	0.000000	-33.333333
2/15/2016	s031103	0.001341	-0.763763	-0.495871	-0.480528	0.736842	0.368421	0.277778	-46.161565
2/15/2016	s032349	-0.114484	-0.647798	-0.437074	-0.599899	0.421053	0.210526	0.222222	-50.003818
2/15/2016	s032417	0.007955	-0.455321	-0.134210	-0.477922	0.210526	0.315789	0.777778	-52.779322
2/15/2016	s032595	-0.004425	-0.737261	-0.559029	-0.535527	0.526316	0.526316	0.444444	-37.495553
2/15/2016	s033025	-0.031828	-0.423896	-0.224683	-0.519168	0.263158	0.421053	0.333333	-45.711147
2/15/2016	s036348	-0.198769	0.000000	-0.465938	-0.478242	0.842105	0.631579	0.555556	-58.822913
2/15/2016	s037954	-0.067201	-1.000000	-0.655516	-0.648005	0.947368	0.684211	0.666667	-45.457151
2/15/2016	s044851	-0.099952	-0.662266	-0.537875	-0.319304	0.315789	0.894737	0.722222	-33.335995
2/15/2016	s045428	-0.025117	-0.793182	-0.487534	-0.363760	0.578947	0.842105	0.500000	-15.789647
2/15/2016	s047438	0.057492	0.000000	-0.337171	-0.334099	1.000000	0.947368	0.944444	-80.048810

图 3：“经理 - 分析师”AI 代理系统设计模式（见 Yu 等人；arXiv:2407.06567v3 [cs.CL] 2024 年 11 月 7 日）。



的 AI 代理被认为有助于简化金融机构的运营。AI 代理可以被视为一个自我维持的 AI 系统，它能够反思其环境，基于预定义的目标执行特定任务，在必要时与其他 AI 代理交互，并适应新环境和更新的目标。在许多情况下，AI 代理需要拥有一个或多个 LLM 作为其骨干智能平台，以完成由人类或另一个 AI 代理分配的任务。凭借这样的 AI 代理，量化投资者或交易者可以构建一个模仿当前投资公司或交易台运营的 AI 系统。例如，Yu 等人（2024 年；见图 3）讨论的“基金经理 - 分析师”代理组设计是一种有趣模式，投资公司可以利用它来构建这样的运营平台，许多量化分析师可以在其中执行各种任务，如数据收集和清理、股票和行业分析、投资相关任务（如收益预测和投资组合构建）等。有趣且令潜在用户放心的是，“经理”AI 代理可以被训练为执行管理任务，如规划、将任务分配给各个“分析师”AI 代理、管理各个“分析师”AI 代理的工作进展、形成最终投资决策等。

我们相信，这种基于 LLM 的 AI 代理系统可以大大提高量化投资公司和 / 或交易台的生产力，特别是中小投资公司。正如 Kelly 等人（2024 年）所讨论的，只要因子的测量误差得到控制，包含在收益预测模型中的特征越多，从基于 ML/DL 的模型中获得的收益就越多。由于财务资源有限，中小投资公司缺乏研究大量投资因子的能力。因此，这些公司倾向于专注于对其投资策略重要的有限数量的因子。相比之下，大公司通常能够负担得起研究和在其投资策略中纳入更多因子的费用，从而可能获得更好的投资业绩。借助 LLM，中小公司可以在其上构建基于 LLM 的 AI 代理系统，许多自动化的“分析师”AI 代理可以帮助收集数据、进行因子分析和回测，并推荐表现最佳的新因子纳入投资策略。当然，这也意味着即使中小投资公司仍然需要投资于 LLM 和 AI 代理系统，这些系统通常伴随着陡峭的学习曲线和潜在的高维护成本。

5. 人工智能驱动的量化投资与交易的“圣杯”

值得注意的是，基于 LLM 的量化投资策略可能潜在地遭受“前瞻性偏差”，因为用于这些分析的 LLMs 仅是特定日期发布的特定版本模型的“快照”。换句话说，LLMs 在回测分析中考虑的历史日期并不存在，因此，回测结果可能潜在地暴露于前瞻性偏差，因为用于回测的 LLMs 可能使用了比回测包含的许多日期“更新”的数据进行训练。

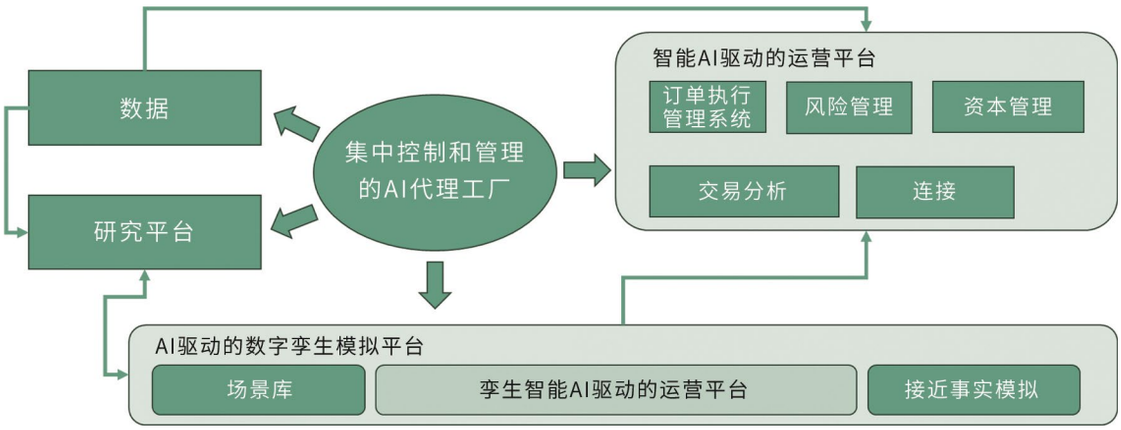
对于许多量化投资公司和交易台而言，解决此类“前瞻性偏差”问题的唯一方法是通过“情景分析”：对于 LLM 发布日期之后的日期，可以创建具有不同经济后果的“市场情景”。换句话说，而不是假设“历史会重演”用于未来的投资决策，量化分析师可以创建“接近现实”的市场条件集合，例如牛市和熊市、高通胀和低通胀环境、自然灾害的出现等。通过这种方式，量化分析师可以通过模拟尽可能多地穷尽各种结果，其中他的 / 她的投资策略可以得到测试。

我们将这种基于集合的“接近现实”模拟支持环境称为人工智能驱动的量化投资与交易平台设计与开发的“圣杯”。在实现这一“圣杯”方面，我们认为可以考虑“数字孪生”概念。例如，图 4 提供了这种量化投资与交易平台“圣杯”设计的高层说明。在该设计的中心是一个集中控制和管理的 AI 代理工厂，在其中可以制造、维护、重新训练和重新分配能够执行不同任务的 AI 代理，以确保其性能状态符合任务要求。

6. 结论

如同 Python 改变了量化投资与交易的基础设施一样，LLMs/AI Agent 也将改变量化投资与交易的基础设施。然而，ML/DL/Python 的影响与 LLMs/AI Agents 的潜在影响之间的区别可能在于，后者可能有助于简化量化投资公

图 4：AI 赋能量化投资与交易平台的“圣杯”设计。



司的运营，从而进一步“民主化”量化投资业务的许多方面（例如通过基于 LLM 的 AI Agent 进行低成本的 alpha 研究）。尽管如此，我们相信所有量化基金可能仍然希望投资于基于人类的特定领域研究和测试能力，尽管 AI 在此过程中将变得越来越有用。在整个过程中，中小公司可能会在短期内从 LLM/AI 代理范式中获益更多，特别是从研究生产力的角度来看；然而，这也意味着在获得实际收益之前需要经历相对陡峭的学习曲线。持续学习并不断接受培训！

参考文献

Yu Yangyang, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie, *FinCon: A Synthesized LLM Multi - Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making*, arXiv:2407.06567v3 [cs.CL] 7 Nov 2024

Kelly, Brian, Semyon Malamud and Kangying Zhou, *The Virtue of Complexity in Return Prediction*, Journal of Finance, Vol. LXXIX, No.1, Feb 2024