



Financial Market Applications of LLMs

20.APR.2024 . 7 MIN READ

The AI revolution drove frenzied investment in both private and public companies and captured the public's imagination in 2023. Transformational consumer products like ChatGPT are powered by Large Language Models (LLMs) that excel at modeling sequences of tokens that represent words or parts of

Richard Dewey

Co-founder and CEO of Proven. Previously worked in financial markets at PIMCO,

words [2]. Amazingly, structural understanding emerges from learning next-token prediction, and agents are able to complete tasks such as translation, question answering and generating human-like prose from simple user prompts.

Not surprisingly, quantitative traders have asked: can we turn these models into the next price or trade prediction [1,9,10]? That is, rather than modeling sequences of words, can we model sequences of prices or trades. This turns out to be an interesting line of inquiry that reveals much about both generative AI and financial time series modeling. Be warned this will get wonky.

LLMs are known as autoregressive learners -- those using previous tokens or elements in a sequence to predict the next element or token. In quantitative trading, for example in strategies like statistical arbitrage in stocks, most research is concerned with identifying autoregressive structure. That means finding sequences of news or orders or fundamental changes that best predict future prices.

Where things break down is in the quantity and information content of available data to train the models. At the 2023 NeurIPS conference, Hudson River Trading, a high frequency trading firm, presented a comparison of the number of input tokens used to train GPT-3 with the amount of trainable tokens available in the stock market data per year HRT estimated that, with 3,000 tradable stocks, 10 data points per stock per day, 252 trading days per year, and 23400 seconds in a trading day, there are 177 billion stock market tokens per year available as

Elm Partners, Tudor
and Royal Bridge.

Ciamac Moallemi

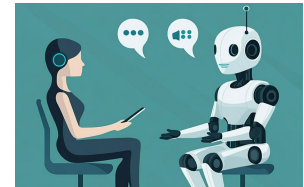
Ciamac Moallemi is
William von Mueffling
Professor of Business
at Columbia
University, where he is
also Director of the
Briger Family Digital
Finance Lab.

| RECENT POSTS

1.

What's Missing From
LLM Chatbots: A
Sense of Purpose

09.SEP.2024



2.

We Need Positive
Visions for AI
Grounded in
Wellbeing

03.AUG.2024



3.

market data. GPT-3 was trained on 500 billion tokens, so not far off [6].

		Number of Stocks	3K
GPT-3 parameters	175B	× 10 years of data	30K
GPT-3 tokens	500B	× 252 trading days per year	7.6M
		× 23,400 seconds per trading day	177B

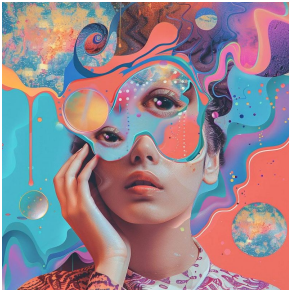
numbers courtesy of HRT 2023 NeuRIPS presentation

But, in the trading context the tokens will be prices or returns or trades rather than syllables or words; the former is much more difficult to predict. Language has an underlying linguistic structure (e.g., grammar) [7]. It’s not hard to imagine a human predicting the next word in a sentence, however that same human would find it extremely challenging to predict the next return given a sequence of previous trades, hence the lack of billionaire day traders. The challenge is that there are very smart people competing away any signal in the market, making it *almost* efficient (“efficiently inefficient”, in the words of economist Lasse Pedersen) and hence unpredictable. No adversary actively tries to make sentences more difficult to predict — if anything, authors usually seek to make their sentences easy to understand and hence *more* predictable.

Looked at from another angle, there is much more noise than signal in financial data. Individuals and institutions are trading for reasons that might not be rational or tied to any fundamental change in a business. The GameStop episode in 2021 is one such example. Financial time series are also constantly changing with new fundamental information, regulatory changes, and occasional large macroeconomic shifts

A Brief Overview of
Gender Bias in AI

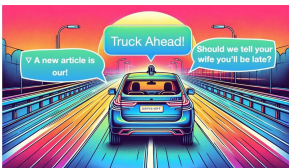
08.APR.2024



4.
Mamba Explained
27.MAR.2024



5.
Car-GPT: Could LLMs
finally make self-
driving cars happen?
08.MAR.2024



| TAGS

- Art
- Conference
- Deep Learning
- Ethics
- Explainability

such as currency devaluations. Language evolves at a much slower pace and over longer time horizons.

On the other hand, there are reasons to believe that ideas from AI will work well in financial markets. One emerging area of AI research with promising applications to finance is multimodal learning [5], which aims to use different modalities of data, for example both images and textual inputs to build a unified model. With OpenAI's DALL-E 2 model, a user can enter text and the model will generate an image. In finance, multi-modal efforts could be useful to combine information classical sources such as technical time series data (prices, trades, volumes, etc.) with alternative data in different modes like sentiment or graphical interactions on twitter, natural language news articles and corporate reports, or the satellite images of shipping activity in a commodity centric port. Here, leveraging multi-modal AI, one could potentially incorporate all these types of non-price information to predict well.

Another strategy called 'residualization' holds prominence in both finance and AI, though it assumes different roles in the two domains. In finance, structural 'factor' models break down the contemporaneous observations of returns across different assets into a shared component (the market return, or more generally returns of common, market-wide factors) and an idiosyncratic component unique to each underlying asset. Market and factor returns are difficult to predict and create interdependence, so it is often helpful to remove the common element when making predictions at the individual asset level

Generative
Models

Graphs

Healthcare

History

Human
factors

Impacts

Interpretability

Language

LLM

Machine
Learning

NLP

Overviews

Perspectives

Podcast

Policy

Quantum ML

Reinforcement
Learning

Science

and to maximize the number of independent observations in the data.

In residual network architectures such as transformers, there's a similar idea that we want to learn a function $h(X)$ of an input X , but it might be easier to learn the residual of $h(X)$ to the identity map, i.e., $h(X) - X$. Here, if the function $h(X)$ is close to identity, its residual will be close to zero, and hence there will be less to learn and learning can be done more efficiently. In both cases the goal is to exploit structure to refine predictions: in the finance case, the idea is to focus on predicting innovations beyond what is implied by the overall market, for residual networks the focus is on predicting innovations to the identity map.

A key ingredient for the impressive performance of LLMs work is their ability to discern affinities or strengths between tokens over long horizons known as context windows. In financial markets, the ability to focus attention across long horizons enables analysis of multi-scale phenomena, with some aspects of market changes explained across very different time horizons. For example, at one extreme, fundamental information (e.g., earnings) may be incorporated into prices over months, technical phenomena (e.g., momentum) might be realized over days, and, at the other extreme, microstructure phenomena (e.g., order book imbalance) might have a time horizon of seconds to minutes.

Capturing all of these phenomena involves analysis of multiple time horizons across the context window. However, in finance,

Speech
Recognition

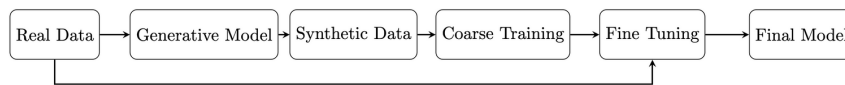
Trends

Vision

XAI

prediction over multiple *future* time horizons is also important. For example, a quantitative system may seek to trade to profit from multiple different anomalies that are realized over multiple time horizons (e.g., simultaneously betting on a microstructure event and an earnings event). This requires predicting not just the next period return of the stock, but the entire term structure or trajectory of expected returns, while current transformer-style predictive models only look one period in the future.

Another financial market application of LLMs might be synthetic data creation [4,8]. This could take a few directions. Simulated stock price trajectories can be generated that mimic characteristics observed in the market and can be extremely beneficial given that financial market data is scarce relative to other sources as highlighted above in the number of tokens available. Artificial data could open the door for meta-learning techniques which have successfully been applied, for example, in robotics. In the robotic setting controllers are first trained using cheap but not necessarily accurate physics simulators, before being better calibrated using expensive real world experiments with robots. In finance the simulators could be used to coarsely train and optimize trading strategies. The model would learn high level concepts like risk aversion and diversification and tactical concepts such as trading slowly to minimize the price impact of a trade. Then precious real market data could be employed to fine-tune the predictions and determine precisely the optimal speed to trade.



Financial market practitioners are often interested in extreme events, the times when trading strategies are more likely to experience significant gains or losses. Generative models where it's possible to sample from extreme scenarios could find use. However extreme events by definition occur rarely and hence determining the right parameters and sampling data from the corresponding distribution is fraught.

Despite the skepticism that LLMs will find use in quantitative trading, they might boost fundamental analysis. As AI models improve, it's easy to imagine them helping analysts refine an investment thesis, uncover inconsistencies in management commentary or find latent relationships between tangential industries and businesses [3]. Essentially these models could provide a Charlie Munger for every investor.

The surprising thing about the current generative AI revolution is that it's taken almost everyone – academic researchers, cutting edge technology firms and long-time observers – by surprise. The idea that building bigger and bigger models would lead to emergent capabilities like we see today was totally unexpected and still not fully understood.

The success of these AI models has supercharged the flow of human and financial capital into AI, which should in turn lead to even better and more capable models. So while the case for GPT-4 like models taking over quantitative trading is currently

unlikely, we advocate keeping an open mind. Expecting the unexpected has been a profitable theme in the AI business.

References

1. “Applying Deep Neural Networks to Financial Time Series Forecasting” Allison Koenecke. 2022
2. “Attention is all you need.” A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones... Advances in Neural Information Processing Systems, 2017
3. “Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models” . Lopez-Lira, Alejandro and Tang, Yuehua, (April 6, 2023) Available at SSRN
4. “Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls.” SA Assefa, D Dervovic, M Mahfouz, RE Tillman... - Proceedings of the First ACM International Conference ..., 2020
5. “GPT-4V(ision) System Card.” OpenAI. September 2023
6. “Language models are few-shot learners.” T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan... - Advances in Neural Information Processing Systems, 2020
7. “Sequence to Sequence Learning with Neural Networks.” I.Sutskever,O.Vinyals,and Q.V.Le in Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.
8. “Synthetic Data Generation for Economists”. A Koenecke, H Varian - arXiv preprint arXiv:2011.01374, 2020
9. C. C. Moallemi, M. Wang. A reinforcement learning approach to optimal execution. Quantitative Finance, 22(6):1051–1069, March 2022.

10. C. Maglaras, C. C. Moallemi, M. Wang. A deep learning approach to estimating fill probabilities in a limit order book. Quantitative Finance, 22(11):1989–2003, October 2022.

Citation

For attribution in academic contexts or books, please cite this work as

Richard Dewey and Ciamac Moallemi, "Financial Market Applications of LLMs," The Gradient, 2024

```
@article{dewey2024financial,  
  author = {Richard Dewey and Ciamac Moallemi},  
  title = {Financial Market Applications of LLMs},  
  journal = {The Gradient},  
  year = {2024},  
  howpublished = {\url{https://thegradient.pub/financial-market-applications-of-llms}},  
}
```

You Might Be Interested In

| DEEP LEARNING

Modern AI is Domestification

27.MAY.2023 / TED XIAO

| DEEP LEARNING

In-Context Learning, In Context

29.APR.2023 / DANIEL BASHIR

| HISTORY

How Machine Learning Can Help Unlock the World of Ancient Japan

17.NOV.2019 / ALEX LAMB

Tags

Art

Conference

Deep Learning

Ethics

Explainability

Generative Models

Graphs

Healthcare

History

Human Factors

Impacts

Interpretability

Language

LLM

Machine Learning

NLP

Overviews

Perspectives

Podcast

Policy

Quantum ML

Reinforcement Learning

Science

Speech Recognition

Trends

Vision

XAI