Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models

University of Florida
Author: Alejandro Lopez-Lira and Yuehua Tang
Reporter: Yanrui Zhou

May 13, 2024

Outline

Research Motivation

- In financial economics, using **LLMs** remains relatively **uncharted territory**, especially concerning their ability to predict stock market returns;
- The performance of **LLMs** in predicting financial market movements is an **open question**:
  - pos: these models are capable of understanding the context of natural language;
  - neg: these models are not explicitly trained for stock return prediction.
- This paper aims to fill this gap by studying the potential of LLMs in extracting the context from news headlines to predict stock returns.

Research Questions

Q: Whether LLMs which are not trained in predicting returns are able to forecast stock market returns while using news headlines data?

Q1: Does the approach based on LLMs outperform **existing methods**?

- Traditional way: sentiment score provided by a leading data vendor.

Q2: Among **various kinds of LLMs**, which one performs well?

- GPT-1, GPT-2, BERT, ChatGPT-3.5 and ChatGPT-4.

Q3: Whether the predictability will change among **different sample**?

- e.g., small and large-cap stocks; positive and negative-news stocks.

Q4: Which reasoning related **concept** can help LLMs predict better?

- stock purchases by insiders, earning guidance, dividends, et al.

Overall Contributions

1. Demonstrate the value of ChatGPT could be employed in **financial industry** and inspire further research;
2. Help **regulators and policymakers** understand the potential benefits and risks associated with the adoption of LLMs;
3. Benefit **asset managers and investors** by providing empirical evidence on the efficacy of LLMs in predicting stock market returns;
4. Contribute to the broader **academic discourse** on artificial intelligence applications in finance.

Academic Contributions

1. Contribute to recent papers that use ChatGPT in the context of economics.
   - Prior studies[1]:
     - pos: ChatGPT can decode Fedspeak, help economics teaching and enhance writing jobs.
     - neu: ChatGPT is no better than simple models when using numerical data.
   - Expand: This paper is among the first to study the potential of LLMs in **financial markets** and the investment decision-making process.
2. Contribute to papers that uses textual analysis and ML to study finance questions.
   - Prior studies[2]: quantify document tone, extract features from textual material, et al.
   - Expand: This paper is the first to evaluate the capabilities of ChatGPT in forecasting stock returns.

[1] e.g., Hansen and Kazinnik (2023), Cowen and Tabarrok (2023), et al.

[2] e.g., Jegadeesh and Wu (2013), Chin and Fan (2023), et al.

Academic Contributions

3. Contribute to recent papers that uses linguistic analyses of news articles to extract sentiment and predict stock returns.
   - Prior studies[3]:
     - studies media sentiment and **aggregate stock returns**.
     - uses the sentiment of firm news to predict **future individual stock returns**.
   - Expand: This paper studies whether LLMs can extract **additional information** that predicts stock market reactions.
4. Contribute to recent papers about employment exposures and vulnerability to AI-related technology.
   - Prior studies[4]: examined the extent of job exposure and vulnerability to AI technology.
   - Expand: In the finance domain, LLMs can help participants process information well.

[3] e.g., Tetlock (2007), Jiang, Li, and Wang (2021), et al.
[4] e.g., W. Jiang et al. (2022), and Noy and Zhang (2023), et al.

Data

- Datasets:
  - CRSP daily returns: NYSE, NASDAQ, AMEX.
  - News headlines: news agencies, financial news websites, and media platforms;
  - RavenPack: a prominent news sentiment analysis data provider[5].
- Sample selection:
  - Period: 2021/10 - 2022/12, allow for a more accurate "out-of-sample" assessment;
  - Match: 67,586 headlines of 4,138 unique companies.
  - Filter: consider "relevance score", select complete articles, eliminate duplicate headlines.

[5] We do not use the RavenPack enhance headlines that potentially contain more information.

Methods: Prompt

- Prompts are essential for enabling ChatGPT to perform a wide range of language tasks, this study uses the following prompt:
  - Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer "YES" if good news, "NO" if bad news, or "UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of _company_name_ in the _term_ term?
    Headline: _headline_

Methods: Prompt

There is an example:

- Headline: Rimini Street Fined $630,000 in Case Against Oracle
- ChatGPT's answer:
  - YES
  - The fine against Rimini Street could potentially boost investor confidence in Oracle's ability to protect its intellectual property and increase demand for its products and services.
- However, software analytics tool gives a negative sentiment score of -0.52.

Methods: Empirical Design

- ChatGPT score: +1(YES), 0(UNKNOWN) and -1(NO);
- Match the headlines to the next trading period return:
  1. before 6 a.m. : the return of the same trading day;
  2. 6 a.m. -> 4 p.m. :
     traded at the same day's close and sold at the close of the next trading day;
  3. after 4 p.m. : the return of the next trading day
- Run linear regressions of the next day's stock returns on the ChatGPT score and other scores[6].

[6] For the more basic LLMs, we employ a different strategy.

Methods: Empirical Design

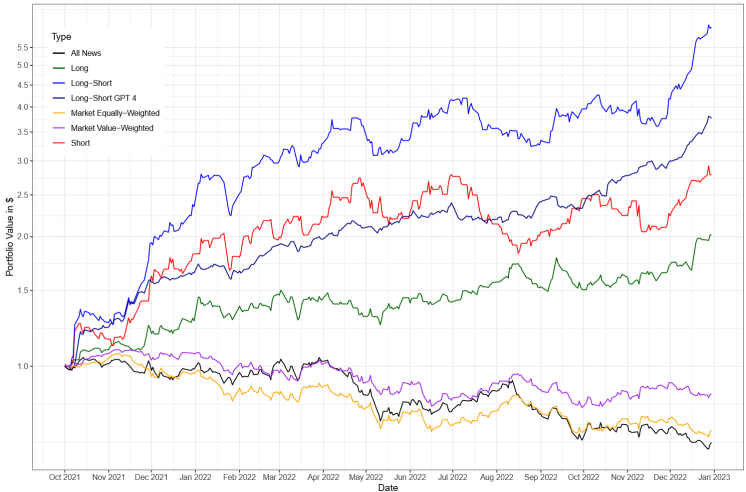
- Estimate the following regression specification:

$$r_{i,t+1} = a_i + b_t + \gamma^{'} x_{i,t} + \varepsilon_{i,t+1} \tag{1}$$

  - $r_{i,t+1}$ is stock i's return over a subsequent trading period;
  - a_i, b_t are fixed effects;
  - $x_{i,t}$ is vector of scores;
  - Standard errors are double clustered by date and firm.

Result: Long-Short Strategies based on ChatGPT Scores

- Form zero-cost portfolios:
  - buy stocks with a positive ChatGPT score;
  - sell stocks with a negative ChatGPT score;
  - rebalanced daily.
- We find strong evidence of the power of ChatGPT scores in predicting stock returns.



Result: Long-Short Strategies based on ChatGPT Scores

- The following table shows the descriptive statistics of various portfolios:

This table reports the following statistics of the different trading strategies as specified in Figure 1: Sharpe ratio, mean daily returns, standard deviation of daily returns, and maximum drawdown. The strategies include (i) the long and short legs of the strategy based on ChatGPT 3.5, (ii) the long-short strategy based on ChatGPT 3.5, (iii) the long-short strategy based on ChatGPT 4, (iv) equal-weight and value-weight market portfolios, and (v) an equal-weight portfolio in all stocks with news the day before (regardless of news direction).

	Long (L)	Short (S)	L-S ChatGPT	L-S GPT-4	Market EW	Market VW	All News EW
Sharpe Ratio	1.72	1.86	3.09	3.80	-0.99	-0.39	-0.98
Daily Mean (%)	0.25	0.38	0.63	0.44	-0.10	-0.04	-0.11
Daily Std. Dev. (%)	2.32	3.26	3.25	1.84	1.55	1.49	1.83
Max Drawdown (%)	-16.94	-34.39	-22.79	-10.40	-36.12	-26.68	-38.70

- We find that although L-S ChatGPT's Daily Mean outperforms L-S GPT-4, L-S GPT-4 has higher Sharpe Ratio and lower Daily Std.

Result: Results from Predictive Regressions

- Study whether LLMs outperforms traditional method:

$$r_{i,t+1} = a_i + b_t + \gamma^{'} x_{i,t} + \varepsilon_{i,t+1}$$

	(1)	(2)	(3)	(4)	(5)	(6)
GPT-3.5-score	0.259***	0.243***				
	(5.259)	(4.980)				
event-sentiment-score		0.058		0.038	0.118*	
		(1.122)		(0.683)	(2.272)	
GPT-4-score			0.176***	0.167***		
			(5.382)	(4.768)		
bart-large-score						0.142***
						(4.653)
Num.Obs.	60 755	60 755	60 755	60 755	60 755	60 176
R2	0.184	0.184	0.184	0.184	0.184	0.185
R2 Adj.	0.121	0.121	0.121	0.121	0.121	0.121

- LLMs can significantly forecast future stock return;
- When GPT-3.5/4 score are controlled, traditional method is insignificant.

Result: Results from Predictive Regressions

- Compare the performance of various language models:

$$r_{i,t+1} = a_i + b_t + \gamma^{'} x_{i,t} + \varepsilon_{i,t+1}$$

	(1)	(2)	(3)	(4)	(5)	(6)
distilbart-mnli-12-1-score	0.150***					
	(4.919)					
GPT-2-large-score		0.035				
		(1.051)				
GPT-2-score			0.001			
			(0.025)			
GPT-1-score				0.034		
				(1.304)		
bert-score					−0.226	
					(−3.703)	
bert-large-score						0.001
						(0.020)
Num.Obs.	60 755	60 176	60 176	60 755	60 176	60 176
R2	0.184	0.185	0.185	0.184	0.185	0.185
R2 Adj.	0.121	0.121	0.121	0.121	0.121	0.121

- BART models show predictability but weaker than GPT-3.5/4;
- More basic LLMs like GPT-1/2 and BERT models don't show any predictability.

Result: Results from Predictive Regressions

Small-cap vs Large-cap firms:

- This article further shows that the predictability of the ChatGPT scores:
  - is present among both small and large-cap stocks;
  - is more pronounced among smaller stocks.

Portfolio Performance of all models:

Model	Pos	Neut	Neg	LS	t LS	α_M	t α_M	$R2_M$	α_{FF5}	t α_{FF5}	R^2_{FF5}
Gpt-4	0.09	-0.18	-0.35	0.44	4.24	0.45	4.31	1.14	0.41	4.01	5.20
Gpt-3.5	0.25	-0.21	-0.38	0.63	3.44	0.63	3.41	0.47	0.60	3.28	4.15
Gpt-1	-0.10	0.01	-0.19	0.09	0.69	0.09	0.71	0.24	0.09	0.67	0.41
Gpt-2	-0.03	-0.20	0.20	-0.23	-1.38	-0.23	-1.37	0.03	-0.24	-1.39	1.83
Gpt-2-Large	-0.06	-0.02	-0.16	0.10	0.92	0.10	0.95	0.40	0.10	0.94	0.64
Bart-Large	-0.01	0.08	-0.16	0.15	1.40	0.15	1.41	0.04	0.13	1.25	1.91
Distilbart-Mnli-12-1	-0.04	0.12	-0.28	0.24	2.12	0.24	2.13	0.08	0.22	1.91	3.77
Bert	-0.23	-0	-0.08	-0.14	-1.16	-0.12	-1.05	12.60	-0.09	-0.79	17.99
Bert-Large	-0.06	-0.06	-0.11	0.04	0.23	0.05	0.25	0.36	0.04	0.19	4.24
Event-Sentiment	-0.04	-0.11	-0.32	0.29	1.94	0.28	1.90	0.47	0.25	1.70	3.04

- Gpt-4 and Gpt-3.5 show significant alpha, and the prior one is more significant.

Interpretability: Evaluating ChatGPT's Reasoning Capabilities

- Traditional **ML** models:
  - focus on prediction, often lacking interpretability.
- **LLMs** like ChatGPT:
  - offer predictions associated with qualitative explanations.
- Postulate that these explanations contain information that can enhance the accuracy of financial forecasts:
  - Refining ChatGPT-4's textual explanations to distill the core reasoning[7];
  - Translate ChatGPT-4's explanations into a more quantifiable format;
    - use TF-IDF to get quantitative data.
  - Employ regularized logistic regression models.
    - y: whether the stock price moved in the direction predicted by ChatGPT.
    - train: distinct models for positive and negative explanations.

[7] removing explicit sentiment indicators.

Interpretability: Evaluating ChatGPT's Reasoning Capabilities

- After regression of individual words:
  - extract the terms with the highest and lowest coefficients;
    - Positive: enhance the prediction accuracy;
    - Negative: reduce the prediction accuracy.
  - but the research above only focuses on individual words.
- Further study the contextual environment of these influential words:
  - identify the words that frequently accompany a given target word;
  - of course, less significant words are filtered out.

Interpretability: Results of Positive Explanations

- The reasoning is more accurate when it's related to:
  - purchases by insiders;
  - earnings guidance;
  - earnings per share;
  - dividends.
- The reasoning is less accurate when it's related to:
  - partnerships or new developments.

Panel A: Positive Influence

Influential Word	Coefficient	Top Accompanying Words
purchase	0.61	future, shows, significant, number, demonstrates
guidance	0.50	indicate, revenue, stability, earnings, likely
share	0.39	earnings, market, indicate, typically, lead
dividends	0.37	generally, seen, sign, generating, profits
higher	0.35	lead, typically, attracts, indicate, sales
returns	0.31	shareholder, stability, attract, indicate, value
generating	0.28	profits, sign, generally, seen, sharing
number	0.26	significant, future, acquisition, shows, purchase
sharing	0.26	generating, profits, sign, generally, seen
insider	0.23	future, positively, significant, number, indicate

Panel B: Negative Influence

Influential Word	Coefficient	Top Accompanying Words
development	-0.54	progress, positively, new, investor, lead
profits	-0.45	generating, sign, generally, seen, sharing
stability	-0.28	sign, generally, seen, indicating, commitment
profitability	-0.27	announcement, shareholder, typically, quarterly...
sales	-0.24	indicate, likely, higher, boost, positively
commitment	-0.21	sign, generally, seen, shareholder, stability
declaring	-0.20	sign, seen, generally, quarterly, indicating
demand	-0.19	lead, increase, partnership, positively, likely
partnership	-0.18	likely, lead, revenue, boost, increase
lead	-0.16	revenue, typically, partnership, higher, collab...

Interpretability: Results of Negative Explanations

- The reasoning is more accurate when it's related to:
  - risk of downgrade;
  - risk related to credit;
  - factors that impacted earnings or revenue negatively;
  - fraud or reputational damages.

Influential Word	Coefficient	Top Accompanying Words
significant	0.92	indicate, lack, selling, number, chairman
indicate	0.78	lack, significant, selling, number, future
risk	0.64	downgrade, credit, investor, higher, outlook
headline	0.48	suggests, likely, earnings, issues, sales
impacted	0.46	likely, earnings, revenue, reduced, drop
director	0.45	indicate, lack, number, sale, future
issues	0.43	headline, sales, impacting, revenue, reduced
number	0.39	lack, indicate, significant, selling, future
fraud	0.36	securities, reputational, investor, loss, headline
reputational	0.33	securities, fraud, losses, headline, lead

Influential Word	Coefficient	Top Accompanying Words
prospects	-0.90	lack, significant
credit	-0.63	downgrade, outlook, future, risk, investor
chairman	-0.62	indicate, lack, selling, significant, number
lack	-0.52	indicate, selling, number, significant, future
outlook	-0.47	downgrade, future, credit, investor, lowered
sale	-0.47	lack, indicate, number, future, director
revenue	-0.44	lower, likely, decreased, expectations, profit
earnings	-0.40	likely, impacted, lower, sales, decline
losses	-0.38	reputational, decreased, securities, impacts, lead
sales	-0.32	lower, decreased, indicate, decline, earnings

Conclusion

Findings:

1. ChatGPT's scores of news headlines can predict subsequent stock returns;
2. More basic LLMs such as GPT-1, GPT-2 cannot accurately forecast returns;
3. The predictability of ChatGPT scores presents in both small and large stocks;
4. Predictability is stronger among smaller stocks and stocks with bad news;
5. We propose a new method to evaluate and understand the models' reasoning capabilities.

Ideas

- If possible, use the content of the news may increase the accuracy of prediction;
- Use ChatGPT to analysis firm's earning calls or other materials and research the reasoning capability.
- Research explanations deeply.
- et al.