Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis

Paul Glasserman[1], Caden Lin[2]
[1]Columbia Business School, Decision, Risk, and Operations Division
[2]Columbia University

Introduction

Large language models (LLMs) like ChatGPT have taken the world by storm. Recent research indicates that LLMs can extract profitable trading signals through sentiment analysis of financial news. However, backtesting such strategies is difficult as LLMs are trained on years of data, and backtesting periods overlap. This bias can take two forms:

- Look-ahead bias: the LLM may have specific knowledge of the returns that follow an event
- Distraction: the LLM's general company knowledge interferes with its sentiment analysis

We investigate these biases by analyzing the performance of LLM-advised trading strategies. In particular, we compare trading performance based on the original headlines with de-biased strategies where we remove the relevant company's identifiers from the text.

Figure 1a: Potential impact of look-ahead bias or distraction on model performance

Figure 1b: Eliminating look-ahead bias and distraction through entity anonymization

Trading Strategy and Performance

We use two sets of news headlines: a Thomson-Reuters (TR) archive covering S&P500 companies and a web-scraped dataset covering 6,000+ publicly traded firms. We follow Lopez-Lira and Tang (2023) to implement an algorithm that executes trades based on GPT's sentiment analysis of news headlines. All portfolios are equally-weighted and rebalanced daily at market close.

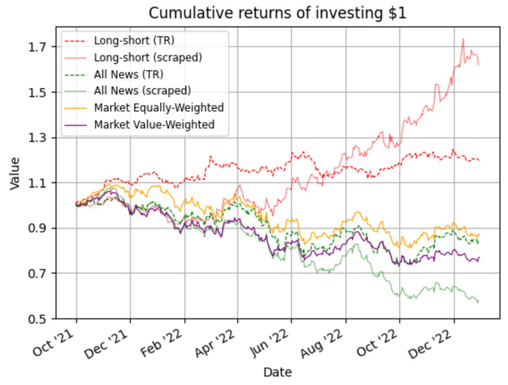

Figure 2: Out-of-sample trading performance using the original headlines

Results

We run the trading algorithm on both the original and replaced headlines, in-sample and out-of-sample. Surprisingly, we observe that the replaced headlines perform better in-sample with statistically significant differences, indicating that the distraction effect outweighs any positive look-ahead bias, by encouraging the LLM to make incorrect, overconfident decisions. We also find that anonymizing the headlines increases the predictive value of the scores, particularly in-sample, which is consistent with the theory of look-ahead bias and distraction.

		Scraped		Thomson Reuters	
		Original	Replaced	Original	Replaced
In-sample	No. of obs.	1699	1699	1695	1695
	Mean	25.08	30.97	10.74	13.84
	Std. dev.	183.87	219.32	77.27	81.64
	t-stat	-1.84		-2.38	
	p-value	0.066		0.017	
Out-of-sample	No. of obs.	314	314	314	314
	Mean	16.32	11.09	6.07	12.23
	Std. dev.	141.52	148.26	85.25	98.21
	t-stat	1.20		-1.86	
	p-value	0.23		0.064	

Table 1: Significance tests comparing original and replaced portfolios

	Scraped		Thomson-Reuters	
	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
orig_score	45.9***	23.5***	15.7***	13.4***
std. error	10.06	7.54	2.15	5.03
t-stat	(4.556)	(3.116)	(7.328)	(2.661)
rep_score	49.3***	23.5***	20.9***	22.36***
std. error	10.83	7.60	2.64	6.15
t-stat	(4.55)	(3.086)	(7.881)	(2.867)
Wald test	1.587	0.000391	24.239	3.315
p-value	0.2077	0.9842	0.0000	0.0686
No. of obs.	199650	59212	314376	49440

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 2: Stacked regressions to compare original and replaced GPT scores

We also find that companies traded by GPT are significantly larger than average in-sample and out-of-sample, pointing to the impact of a distraction effect.

		long	short	other	long-other	short-other
in-sample	Mean	35.37	42.37	17.91	17.46***	24.46***
	Std. dev.	129.17	139.27	93.22		
	No. of obs.	27454	9473	62898		
out-of-sample	Mean	46.30	70.32	32.18	14.12***	38.15***
	Std. dev.	143.74	217.04	123.47		
	No. of obs.	9602	4452	15552		

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 3: Comparison of average market cap of GPT-recommend companies

We also see that portfolios driven by replaced headlines are correlated with the market, which further indicates biased predictions. In theory, our strategies should be uncorrelated to the market, as they are based on short-term trading signals.

	Scraped				Thomson-Reuters			
	In-Sample		Out-of-Sample		In-Sample		Out-of-Sample	
	Original	Replaced	Original	Replaced	Original	Replaced	Original	Replaced
const	12.77***	14.74***	18.78***	14.13***	9.09***	11.44***	3.77	3.54
std err.	3.917	4.389	6.623	6.384	1.783	2.007	4.117	4.979
t-stat	3.26	3.358	2.837	2.213	5.102	5.701	0.917	0.712
rm-rf	0.429***	0.348***	0.273***	0.0850	0.317***	0.0677	0.220***	2.21
std err.	0.0420	0.215	0.0532	0.0694	0.0192	0.0488	0.0331	0.0677
t-stat	10.217	3.032	5.131	1.226	16.571	1.389	6.654	0.326
No. of obs.	1699	1699	314	314	1695	1695	314	314
R	0.058	0.033	0.078	0.009	0.14	0.005	0.124	0.001

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 4: Regression results of long-short portfolio returns on the excess return of the market

Conclusion

Extensive training data allows LLMs to make short-term stock predictions without specific training. However, such data limits opportunities to design and backtest LLM-based strategies. We find that entity anonymization improves LLM performance in-sample, suggesting that a distraction effect—where general knowledge of companies interferes with sentiment analysis capabilities—hurts overall LLM performance. Our findings are robust across two different news sources.

References

Lopez-Lira, Alejandro and Yuehua Tang (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv: 2304.07619 [q-fin.ST].