Can ChatGPT Forecast Macroeconomic Indicators?

David Bell*
Professor Allan Timmermann†

August 2023

Contents

*Student Researcher at University of California, San Diego
†Supervisor

1 Abstract

In this study, we explore the potential of ChatGPT and other large language models in predicting various independent variables such as GDP, inflation, and oil prices, using sentiment analysis of news headlines in the US. The sentiment analysis aims to determine if ChatGPT can outperform a standard sentiment analyser like Vader in forecasting macroeconomic indicators. First, we train ChatGPT to compute a numerical sentiment score in a similar fashion to Lundgaard Hansen and Kazinnik instructing it to output both the score and a concise explanation on how it made this decision. We repeat this process for a database of headlines from 2022 and 2023 derived from various news sources including Bloomberg, Wall Street Journal and Reuters. We then extract, clean and prepare this data by calculating the cumulative daily average scores indexed by date to output a time series to be used as an exogenous variable in our forecasting model. The sentiment score time series for both the ChatGPT and Vader model are then merged into a database with the various macroeconomic indicators. The data is then organised by date with missing values being forward filled to ensure continuity. These data are then standardised to ensure compatible scales and the data is split into 80% training and 20% testing, maintaining the correct time order. We use a linear regression and a Random Forrest model that is later fine-tuned to forecast our macroeconomic indicators. The testing data is then unstandardised to show a more interpretable prediction and statistical summaries are produced for model validation. This study contributes to the growing body of research on the application of large language models in economic forecasting.

2 Literature Review

In the rapidly growing fields of economics and finance, GPT models have become increasingly integrated, with numerous studies exploring their potential in various applications. Lopez-Lira and Tang (2023) assessed ChatGPT's ability to forecast returns, leveraging the model's sentiment analysis features. By comparing it with existing sentiment analysis techniques, they revealed the promising prospects of GPT models in financial forecasting.

Leippold (2023a) utilized GPT models to highlight the limitations of dictionary-based approaches in Natural Language Processing (NLP), demonstrating that context-aware methods like BERT are more effective. In a follow-up study, Leippold (2023b) applied GPT models to discuss climate change, illuminating the models' capabilities and shortcomings in understanding complex topics.

Dowling and Lucey (2023) and Korinek (2023) contributed to the discourse on how ChatGPT and large language models (LLMs) can be employed by financial researchers to enhance efficiency through micro-task automation. Concurrently, Zaremba and Demir (2023) examined the current state of GPT technology in finance, suggesting its potential to improve NLP-based financial applications.

The literature has been further enriched by Lundgaard Hansen and Kazinnik, who evaluated GPT models in quantifying Federal Reserve communication. This analysis builds on comprehensive research using NLP to study the content, sentiment, and impact of central bank communication, as seen in works by Ehrmann and Fratzscher (2007), Hansen et al. (2019), Hayo and Neuenkirch (2015), Curti and Kazinnik (2021), and Ehrmann and Wabitsch (2022).

Historical efforts in this area have measured central bank texts using topic modeling and sentiment analysis with pre-defined dictionaries, such as the one created by Loughran and McDonald (2011). Modern papers have shifted towards pre-trained LLMs like BERT models, as shown by Doh et al. (2022), Bertsch et al. (2022), and Gorodnichenko et al. (2023). In our paper, we contrast the effectiveness of these common techniques with GPT models, using both zero-shot and few-shot learning approaches.

Our research goes beyond existing studies that have investigated GPT's usefulness in stock price prediction, extending this exploration to the forecasting of key economic indicators for entire countries, including GDP, inflation, and oil prices. This expansion marks a significant increase in GPT's relevance in economic forecasting.

These findings not only clarify the strengths and limitations of large language models in economic forecasting but also generate numerous hypotheses for future research. The existing literature lacks a comprehensive cross-country comparison of GPT's effectiveness in economic forecasting, a significant gap that our study aims to fill.

3 Background

Artificial intelligence, particularly large language models like GPT, has shown promise in various applications, including economic forecasting. Previous research has demonstrated GPT's ability to predict stock returns based on news headlines and interpret Federal Open Market Committee (FOMC) meeting notes as effectively as a human. These studies suggest that GPT can extract and process economic information from textual data, providing valuable insights for economic forecasting. Indeed, with the latest publication from Bloomberg themselves introducing their upcoming BloombergGPT suggests a positive outlook on the future of this field. Despite these promising results, the potential of GPT in predicting key macroeconomic indicators, such as GDP, inflation, and oil prices, remains largely unexplored. These gaps in our understanding of GPT's capabilities limit its practical applications in economic forecasting. This study aims to address these gaps by exploring the use of GPT in predicting key economic indicators for entire countries based on news headlines. We will train GPT to compute a numerical sentiment score for a given headline, output this with a brief explanation and we will then use this data in tandem with the macroeconomic indicator data to build a forecasting model. Our analysis will allow us to investigate potential differences in ChatGPT's predictive accuracy by comparing it with a control sentiment analyser, Vader. By doing so, we hope to provide a more comprehensive understanding of GPT's potential in economic forecasting and open the door for future research.

4 Data

4.1 Collection

The first step in our methodology involved the collection of data. We gathered news headlines from January 2022 to august 1st 2023 from various sources such as The Wall Street Journal, Bloomberg and Reuters. This provided us with a comprehensive set of data points for each day of the year. In addition to this, we obtained macroeconomic indicator data including CPI, treasury debt, and unemployment rate. In addition, we collected Federal Open Market Committee (FOMC) notes from the same time period. The FOMC releases ten of these notes per year, providing us with additional valuable data for our analysis.

4.2 Preparation

Once the data was collected, we indexed it by date and produced a database of headlines. We then arranged a second database with the macroeconomic data also indexed by date so that we could merge the sentiment scores for a complete database.

5 Methodology

5.1 Training ChatGPT

We leveraged ChatGPT-3.5 to create a custom sentiment analysis system tailored to the financial domain. This system was designed to classify financial headlines based on their potential impact on the US Treasury debt. The methodology involved the following steps:

1. **Model Configuration:** We utilized OpenAI's GPT-3.5-turbo model, instructing it to assume the role of a financial expert with specific expertise in stock recommendation and sentiment analysis.

2. **Classification Scale:** We defined a classification scale ranging from -1 to 1, where each numeric label corresponded to different levels of impact on the US Treasury. The categories were as follows:

   - -1: Bad for US Treasury
   - -0.5: Mostly bad for US Treasury
   - 0: Neutral/Irrelevant
   - 0.5: Mostly good for US Treasury
   - 1: Good for US Treasury

3. **Prompt Design:** A specific prompt was crafted to guide the model's response. The prompt instructed the model to classify a given headline into one of the predefined categories, providing both a numeric classification and a concise explanation.

4. **Data Analysis:** We applied this model to a comprehensive data set of financial headlines collected from January 2022 to August 1st, 2023.

5.2 Empirical Strategy

This algorithm analyzes the sentiment of the news headlines for each day, outputting a score that ranges from -1 to 1. We repeated this process for every day in our sample space and compiled a list of responses. We then iterated through these responses, calculating the average per day if there were more than one headline and output a cumulative sentiment score time series of the returned values. In addition, we constructed a comprehensive pipeline to facilitate the sentiment analysis. This involved connecting a database and parsing the headlines through ChatGPT, which returned the initial sentiment scores. We cleaned and processed the raw format output by ChatGPT into a structured database. Subsequently, we implemented VADER sentiment analysis. The iteration involved the selection of headlines and dates from the tables and utilising the Vader Python library to output a sentiment analysis similar that that of ChatGPT. The data was then cleaned and prepared as with the ChatGPT outputting a cumulative sentiment score time series. Finally, data merging operations were

performed to combine the VADER scores, ChatGPT scores, and macro data into a unified structure. This structure was indexed by date, forming a time series for our forecasting model.

6 Model Building

We split the model into 80% training and 20% testing maintaining the date indexing. The process was iterative and involved the construction of multiple models, each time building a main model using ChatGPT and a control model using the Vader sentiment scores. Initially, we decided on the model methodology, carefully selecting different models based on the results of trial and error. We began with a simple Linear Regression model, utilizing it as a foundational approach to understand the basic relationship between the cumulative sentiment score and treasury debt value. This model served as a starting point for our analysis, providing initial insights into the potential correlation between the variables. Following the Linear Regression, we proceeded to a more complex Random Forest model. This model was chosen for its ability to capture non-linear relationships and provide a more nuanced understanding of the underlying patterns in the data. The Random Forest model's ability to handle a large number of features and its robustness to outliers made it a suitable choice for our analysis. Following this, we recognized certain similarities with the macroeconomic data and made informed decisions to alter the Random Forest model accordingly. This led to the creation of a fine-tuned version of the Random Forest model, involving vertically shifting the data. The reason for this adjustment was that the data fit the model well but was not at the correct starting point. By recognizing this discrepancy and making the necessary adjustments, we were able to align the data more accurately with the model's predictions.

Throughout this process, we maintained a rigorous approach, evaluating and refining our models to ensure that they were accurately capturing the relationship between sentiment scores and treasury debt value. The combination of different modeling techniques, from simple linear regression to more complex random forest models, allowed us to explore the data from various angles and arrive at a comprehensive understanding of the underlying relationship.

7 Results

The findings reveal that ChatGPT sentiment scores demonstrate a statistically significant predictive power on treasury debt. This indicates that the GPT model can be a valuable tool for modeling inflation movements based on sentiment analysis of news headlines. The advantage of using GPT in modeling inflation can be attributed to its advanced language understanding capabilities, enabling it to capture the subtle nuances and context within news headlines related to economic indicators like inflation. Our regression analysis demonstrates the use of GPT sentiment scores in a more precise and informative forecast of treasury debt.

Overall, the results indicate that GPT can effectively model treasury debt when we consider a fine-tuned random forest model. These findings highlight the potential benefits of leveraging large language models like GPT for economic forecasting and decision-making processes, particularly when a careful analysis of the results is considered.

7.1 Linear Regression Model

We evaluated two models, the Vader Model and the ChatGPT Model, to forecast Treasury values based on sentiment scores. The performance metrics and their interpretations are as follows: The Mean Absolute Percentage Error (MAPE) for the Vader Model is 17.45%, and for the ChatGPT Model, it is 18.07%. MAPE quantifies the prediction error as a percentage, and lower values indicate better accuracy. In this case, the Vader Model has a slightly better MAPE, suggesting it may be more accurate in forecasting Treasury values. The Pearson Correlation Coefficient for the Vader Model is -0.8281, and for the ChatGPT Model, it is 0.8741. Pearson Correlation measures the linear relationship between predicted and actual values. The positive correlation for the ChatGPT Model indicates a strong linear relationship, while the negative correlation for the Vader Model suggests an inverse relationship. The ChatGPT Model's positive correlation is more desirable for forecasting. The Granger Causality Test provides insights into the causal relationships between the predicted and actual Treasury values. Significant p-values at certain lags indicate predictive power.

Both models have strengths and weaknesses. The Vader Model performs slightly better in terms of MAPE, suggesting better overall accuracy. However, the ChatGPT Model exhibits a strong positive linear correlation with the actual Treasury values, which might be more indicative of a meaningful forecasting relationship. Considering both accuracy and linear relationship, the choice between the models may depend on the specific context and requirements of the forecasting task. If the primary focus is on having a strong linear relationship between predictions and actual values, the ChatGPT Model might be preferable. If overall accuracy is more critical, the Vader Model might be the better choice. Further exploration, fine-tuning, and testing with different models and algorithms could provide additional insights and possibly improve the performance of the forecasting.

7.2 Random Forest Model

We evaluated four models, including Random Forest and Linear Regression, applied to both ChatGPT and Vader sentiment scores, to forecast Treasury values. The Mean Absolute Percentage Error (MAPE) for the Random Forest ChatGPT Model was 5.56%, for the Linear Regression ChatGPT Model was 18.07%, for the Linear Regression Vader Model was 17.45%, and for the Random Forest Vader Model was 24.16%. The Pearson Correlation for the Linear Regression ChatGPT Model was 0.8741, for the Random Forest ChatGPT Model was 0.3880, for the Linear Regression Vader Model was -0.8281, and for the Random Forest Vader Model was -0.7339. Granger Causality Tests were conducted for the Random Forest Vader Model at lags 1, 2, and 3, with F-values of 10.4890, 4.0420, and 10.5856, respectively, and corresponding p-values of 0.0016, 0.0207, and 0.0000. Interpretations of the results include the observation that the Random Forest ChatGPT Model has the lowest MAPE, showing the closest predictions to the actual Treasury values, and that the Random Forest Vader Model has the highest MAPE and a negative Pearson correlation, indicating the poorest performance among the models. The Random Forest model with ChatGPT sentiment scores appears to be the most promising, especially after applying the shift correction. The Linear Regression models perform similarly but are outperformed by the Random Forest ChatGPT Model. The Random Forest Vader Model is the least effective among the four. Further investigation may involve fine-tuning the Random Forest hyper-parameters, exploring other modeling techniques, or considering additional feature engineering to improve model performance.

8 Conclusion

The findings of our research significantly contribute to our understanding of GPT's ability to predict macroeconomic indicators. We have demonstrated that GPT's sentiment scores exhibit a strong predictive power on treasury debt. This suggests that GPT, as a large language model, can effectively model such indicators based on the analysis of news headlines. This capability opens new avenues for leveraging natural language processing and sentiment analysis techniques in the field of economic forecasting. By showcasing GPT's effectiveness in modeling economic indicators like Treasury Debt, our research supports the idea that large language models have the potential to become indispensable tools in economic analysis and decision-making processes. The ability of GPT to analyze vast amounts of textual data and extract valuable insights contributes to the advancement of data-driven approaches in macroeconomic research. Our findings are consistent with the growing body of evidence supporting the potential of large language models in improving predictive accuracy across different economic domains. By demonstrating the value of integrating expert insights and domain-specific knowledge, we contribute to the ongoing discussion on how to leverage the full potential of GPT and similar models in economic forecasting and policy analysis. This analysis provides compelling evidence that GPT's sentiment scores have a statistically significant predictive power. Our research demonstrates the potential of large language models like GPT in contributing to a deeper understanding of economic indicators and facilitating data-driven decision-making in economic research and policy analysis.

9 Figures

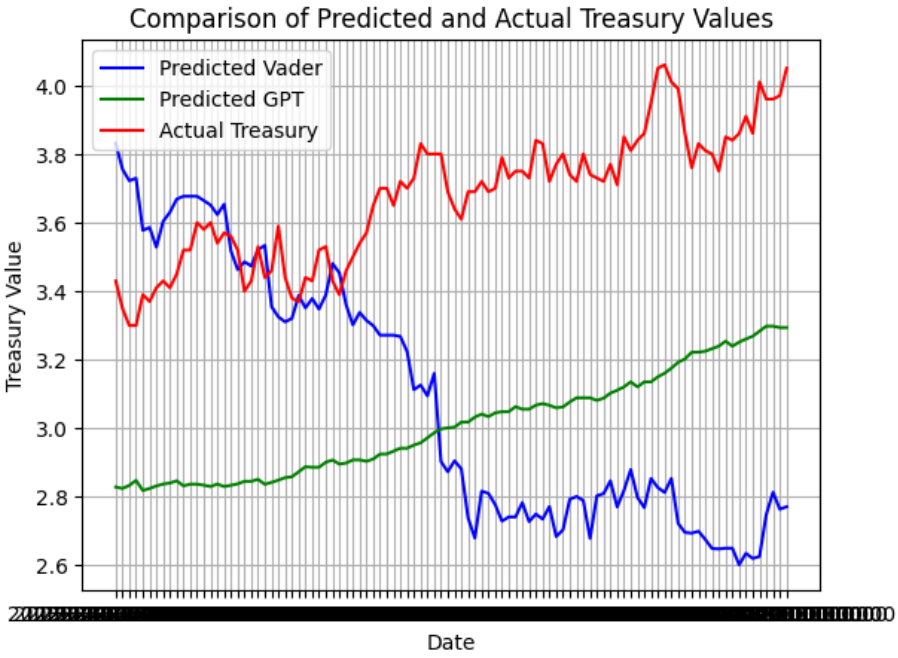

Figure 1: Unstandardised Linear Regression

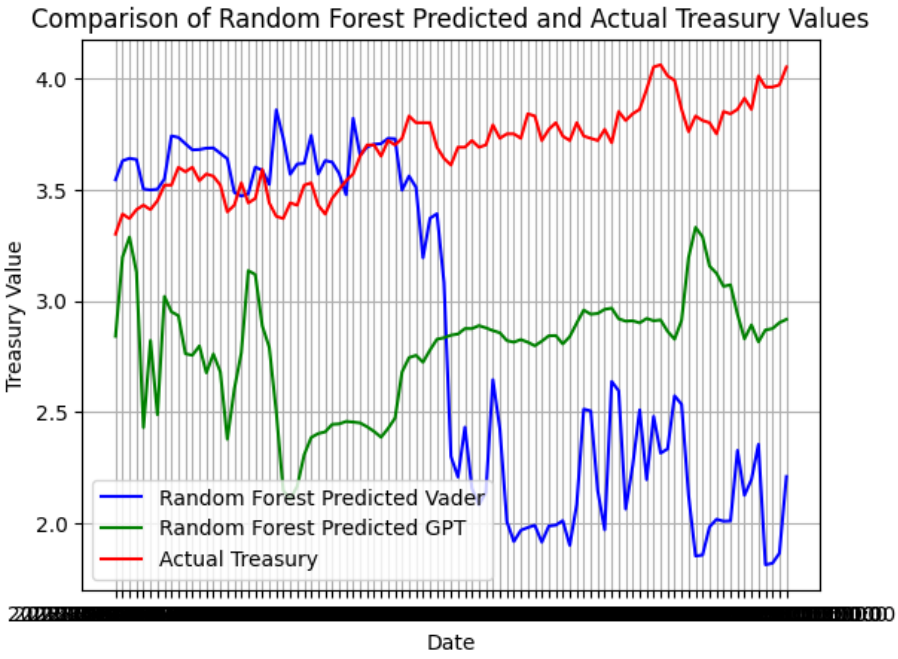

Figure 2: Random Forest

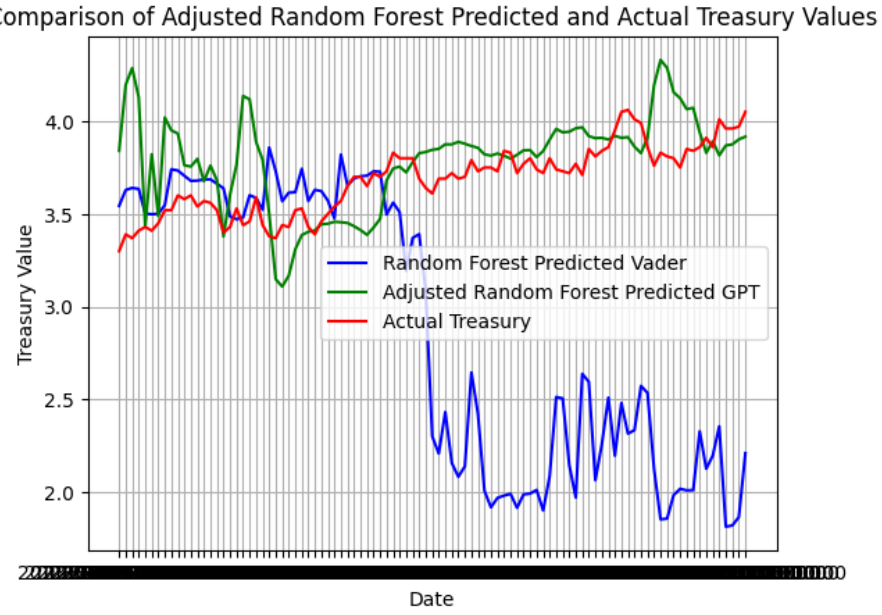

Figure 3: Vertically Shifted ChatGPT Random Forest

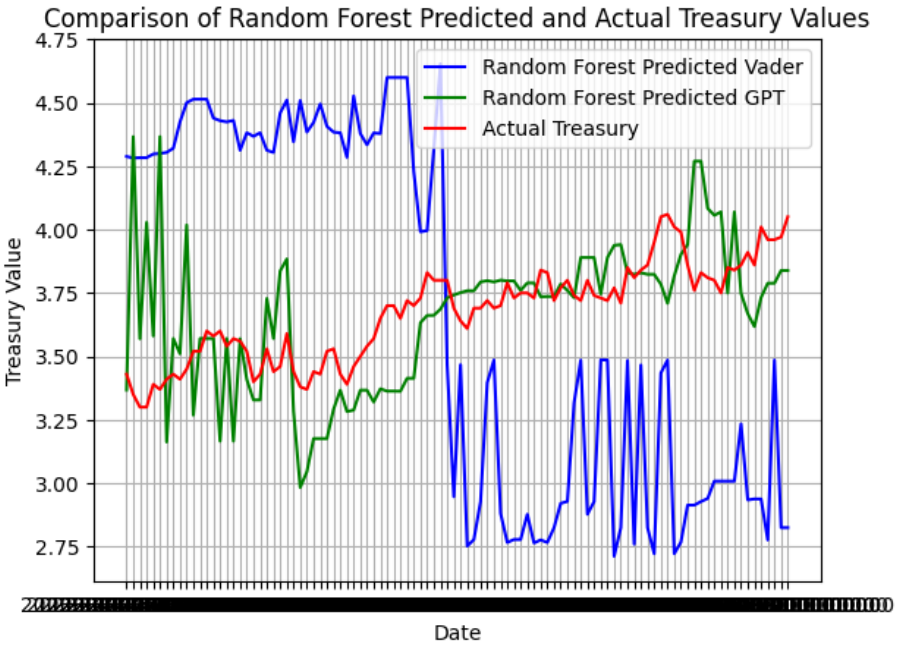

Figure 4: Both models optimised for minimal MAE

10 Tables

Metric	Vader Model	ChatGPT Model
MAPE	17.45%	18.07%
Pearson Correlation	-0.8281	0.8741
Granger Causality (Lag 1)		
ssr based F test	F=4.4160	p=0.0382
ssr based chi2 test	chi2=4.5540	p=0.0328
likelihood ratio test	chi2=4.4523	p=0.0349
parameter F test	F=4.4160	p=0.0382
Granger Causality (Lag 2)		
ssr based F test	F=1.3652	p=0.2604
ssr based chi2 test	chi2=2.8772	p=0.2373
likelihood ratio test	chi2=2.8358	p=0.2422
parameter F test	F=1.3652	p=0.2604
Granger Causality (Lag 3)		
ssr based F test	F=3.8248	p=0.0125
ssr based chi2 test	chi2=12.3668	p=0.0062
likelihood ratio test	chi2=11.6397	p=0.0087
parameter F test	F=3.8248	p=0.0125
Granger Causality (Lag 4-5)		
ssr based F test	F=4.9416	p=0.0005
ssr based chi2 test	chi2=27.9434	p=0.0000
likelihood ratio test	chi2=24.4955	p=0.0002
parameter F test	F=4.9416	p=0.0005

Table 1: Summary of Linear Regression Model Evaluation Metrics

Metric	Random Forest ChatGPT	Random Forest Vader
MAPE	5.56%	24.16%
Pearson Correlation	0.3880	-0.7339
Granger Causality (Lag 1)		
ssr based F test		F=10.4890, p=0.0016
ssr based chi2 test		chi2=10.8168, p=0.0010
likelihood ratio test		chi2=10.2657, p=0.0014
parameter F test		F=10.4890, p=0.0016
Granger Causality (Lag 2)		
ssr based F test		F=4.0420, p=0.0207
ssr based chi2 test		chi2=8.5187, p=0.0141
likelihood ratio test		chi2=8.1686, p=0.0168
parameter F test		F=4.0420, p=0.0207
Granger Causality (Lag 3)		
ssr based F test		F=10.5856, p=0.0000
ssr based chi2 test		chi2=34.2267, p=0.0000
likelihood ratio test		chi2=29.3149, p=0.0000
parameter F test		F=10.5856, p=0.0000

Table 2: Summary of Random Forest Model Evaluation Metrics

11 References

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." Journal of Labor Economics 40, no. S1 (April): S293–S340. issn: 0734306X. [Link](https://doi.org/10.1086/718327/ SUPPL FILE/20462DATA.ZIP).

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, Automation, and the Rise in U.S. Wage Inequality." Econometrica 90, no. 5 (September): 1973–2016. issn: 1468-0262. Link.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." Journal of Economic Perspectives 33, no. 2 (March): 31–50. issn: 0895-3309. Link.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson. 2022. "Artificial Intelligence, Firm Growth, and Product Innovation." SSRN Electronic Journal (May). Link.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring economic policy uncertainty." Quarterly Journal of Economics 131, no. 4 (November): 1593–1636. issn: 15314650. Link.

Binsbergen, Jules H. van, Xiao Han, Alejandro Lopez-Lira, Jules H van Binsbergen, Xiao Han, and Alejandro Lopez-Lira. 2020. Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases. Technical report, Working Paper Series 27843. National Bureau of Economic Research. Link.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2019. "The Structure of Economic News." Working Paper (January). issn: 1556-5068. Link.

Lopez-Lira, Alejandro, and Yuehua Tang. "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models." University of Florida.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." Journal of Labor Economics 40, no. S1 (April): S293–S340. issn: 0734306X. [Link](https://doi.org/10.1086/718327/ SUPPL FILE/20462DATA.ZIP).

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, Automation, and the Rise in U.S. Wage Inequality." Econometrica 90, no. 5 (September): 1973–2016. issn: 1468-0262. Link.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." Journal of Economic Perspectives 33, no. 2 (March): 31–50. issn: 0895-3309. Link.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson. 2022. "Artificial Intelligence, Firm Growth, and Product Innovation." SSRN Electronic Journal (May). Link.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring economic policy uncertainty." Quarterly Journal of Economics 131, no. 4 (November): 1593–1636. issn: 15314650. Link.

Binsbergen, Jules H. van, Xiao Han, Alejandro Lopez-Lira, Jules H van Binsbergen, Xiao Han, and Alejandro Lopez-Lira. 2020. Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases. Technical report, Working Paper Series 27843. National Bureau of Economic Research. Link.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2021. "Business News and Business Cycles." SSRN Electronic Journal (September). issn: 1556-5068. Link.

Calomiris, Charles W., and Harry Mamaysky. 2019. "How news and its context drive risk and returns around the world." Journal of Financial Economics 133, no. 2 (August): 299–336. issn: 0304-405X. Link.

Campbell, John L., Hsinchun Chen, Dan S. Dhaliwal, Hsin-min min Lu, Logan B. Steele, John L. Campbell, Hsinchun Chen, et al. 2014. "The information content of mandatory risk factor disclosures in corporate filings." Review of accounting studies (Boston) 19, no. 1 (March): 396–455. issn: 1380-6653. [Link](https://doi.org/10.1007/S11142-013-9258- 3/TABLES/11).

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen. 2020. "Lazy Prices." Journal of Finance 75 (3): 1371–1415. issn: 15406261. Link.

Cowen, Tyler, and Alexander T. Tabarrok. 2023. "How to Learn and Teach Economics with Large Language Models, Including GPT." SSRN Electronic Journal (March). issn: 1556-5068. Link.