

Quantifying uncertainty: a new era of measurement through large language models

Francesco Audrino, Jessica Gentner, Simon Stalder

SNB Working Papers

12/2024



EDITORIAL BOARD SNB WORKING PAPER SERIES

Marc-Antoine Ramelet
Enzo Rossi
Rina Rosenblatt-Wisch
Pascal Towbin
Lukas Frei

DISCLAIMER

The views expressed in this paper are those of the author(s) and do not necessarily represent those of the Swiss National Bank. Working Papers describe research in progress. Their aim is to elicit comments and to further debate.

COPYRIGHT©

The Swiss National Bank (SNB) respects all third-party rights, in particular rights relating to works protected by copyright (information or data, wordings and depictions, to the extent that these are of an individual character).

SNB publications containing a reference to a copyright (© Swiss National Bank/SNB, Zurich/year, or similar) may, under copyright law, only be used (reproduced, used via the internet, etc.) for non-commercial purposes and provided that the source is mentioned. Their use for commercial purposes is only permitted with the prior express consent of the SNB.

General information and data published without reference to a copyright may be used without mentioning the source. To the extent that the information and data clearly derive from outside sources, the users of such information and data are obliged to respect any existing copyrights and to obtain the right of use from the relevant outside source themselves.

LIMITATION OF LIABILITY

The SNB accepts no responsibility for any information it provides. Under no circumstances will it accept any liability for losses or damage which may result from the use of such information. This limitation of liability applies, in particular, to the topicality, accuracy, validity and availability of the information.

ISSN 1660-7716 (printed version)
ISSN 1660-7724 (online version)

© 2024 by Swiss National Bank, Börsenstrasse 15,
P.O. Box, CH-8022 Zurich

Quantifying Uncertainty: A New Era of Measurement through Large Language Models

Francesco Audrino*, Jessica Gentner[†] and Simon Stalder[‡]

October 16, 2024

Abstract

This paper presents an innovative method for measuring uncertainty via large language models (LLMs), which offer greater precision and contextual sensitivity than the conventional methods used to construct prominent uncertainty indices. By analysing newspaper texts with state-of-the-art LLMs, our approach captures nuances often missed by conventional methods. We develop indices for various types of uncertainty, including geopolitical risk, economic policy, monetary policy, and financial market uncertainty. Our findings show that shocks to these LLM-based indices exhibit stronger associations with macroeconomic variables, shifts in investor behaviour, and asset return variations than conventional indices, underscoring their potential for more accurately reflecting uncertainty.

JEL Classification: C45, C55, E44, G12

Keywords: Uncertainty measurement, Large language models, Economic policy, Geopolitical risk, Monetary policy, Financial markets.

*University of St. Gallen and Swiss Finance Institute (SFI); francesco.audrino@unisg.ch

[†]University of St. Gallen and Swiss National Bank (SNB); jessica.gentner@snb.ch

[‡]University of Lugano (USI) and Swiss National Bank (SNB); simon.stalder@snb.ch

We thank Daniele Ballinari, Fabian Fink, Francesco Franzoni, Lukas Frei, Thomas Maag, Harry Mamaysky, Erik Senn, Alexander Wehrli and the entire Technology and Data Science team of the Money Market and Foreign Exchange division at the Swiss National Bank (SNB) as well as the seminar participants at the SNB and the University of St. Gallen and the anonymous referee for helpful comments and discussions. The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of the SNB. The SNB takes no responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper.

1 Introduction

The global relevance of economic policy uncertainty is undeniably profound, as evidenced by significant events such as the Brexit referendum in June 2016. The uncertainty surrounding Britain’s exit from the European Union not only created fluctuations in the British markets but also reverberated across global financial markets. Despite its paramount importance, the multifaceted nature of economic policy uncertainty makes it an elusive and complex phenomenon to quantify and understand. Traditional methods, such as counting the occurrence of specific terms used to construct the economic policy uncertainty index (EPU) developed by Baker, Bloom, and Davis (2016), have been vital in shaping our existing understanding. However, as we transition into an era marked by rapid technological advancements, more advanced tools, such as large language models (LLMs), have become available. The mere reliance on word frequency without scrutinising contextual nuances creates a discernible gap in our ability to fully grasp and quantify uncertainty.

This paper addresses two main research questions: First, can LLMs enhance the measurement of uncertainty by capturing nuances that conventional methods overlook? Second, can advanced LLM-based uncertainty indices effectively predict macroeconomic indicators as well as asset returns and shifts in financial market behaviours? We hypothesise that LLMs, by leveraging their superior language comprehension capabilities, will provide a more nuanced and contextually aware quantification of uncertainty than traditional methods and that these measures will correlate strongly with macroeconomic indicators and market movements, particularly during periods of high uncertainty.

To explore these questions, we employ state-of-the-art LLMs, including GPT-4 and open-source models such as LLaMa-2-7B-Chat and Zephyr-7B- β , to analyse and classify textual data from a comprehensive dataset of Wall Street Journal (WSJ) articles covering significant economic events from 1998 to 2023. To keep the costs low, the open-source models are fine-tuned via a novel dataset created by GPT-4, with a focus on their ability to detect different types of uncertainty: economic policy uncertainty, geopolitical risk, monetary policy uncertainty and financial market volatility. The best-performing model, the fine-tuned Zephyr-7B- β model, is used to construct various uncertainty indices. Our findings confirm that our LLM-based indices surpass traditional indices in capturing the complexities of uncertainty and exhibit significant correlations with macroeconomic indicators, asset returns, and mutual fund flows.

The implications of this research are profound. By integrating advanced computational models with traditional economic analysis, this paper not only broadens the scope of uncertainty measurement but also enhances the predictive accuracy of macroeconomic indicators and financial market

responses to economic shocks. This presents a compelling case for policymakers, financial analysts, and economic researchers to adopt LLMs in their analytical frameworks, offering a more dynamic and timely approach to managing economic risk and reacting to uncertainty.

Measuring uncertainty in newspaper articles is highly relevant because the media plays a critical role in shaping public perception and investor sentiment. Newspapers such as the WSJ provide timely and comprehensive coverage of economic and political events, making them valuable sources of information for gauging market sentiment and expectations. The language and tone used in these articles can reflect both the factual state of the economy and the prevailing mood among market participants, offering a real-time snapshot of uncertainty. By examining the language and context of news articles, our approach reveals how uncertainty, as measured by newspaper articles, can affect the economy and financial markets.

Our first contribution is to demonstrate that LLMs can be fine-tuned effectively via a relatively small and inexpensive dataset created by GPT-4 Turbo, making this approach both practical and scalable. We select two open-source models, LLaMa-2-7B-Chat and Zephyr-7B- β and fine-tune them via a dataset of WSJ articles classified by GPT-4 Turbo. By performing instruction tuning and utilising a parameter-efficient fine-tuning technique known as QLoRA, we achieve cost-effective fine-tuning on a single GPU, ensuring ease of replication and broad accessibility.

We benchmark the fine-tuned models against the state-of-the-art GPT-4 Turbo, using accuracy and F1 scores as our primary metrics. The fine-tuned Zephyr-7B- β achieves high accuracy and F1 scores across all uncertainty categories, demonstrating that LLMs can be adapted for specific tasks with limited data and without the time-consuming human labelling of the dataset. This finding underscores the potential for widespread application of LLMs in economic and financial contexts, where cost and computational resources are often constrained.

Furthermore, we show that LLMs outperform traditional methods such as the widespread bag-of-words (BoW) method in identifying uncertainty from news articles by capturing textual nuances that BoW methods overlook. To evaluate the accuracy of the labelling results, we conduct a human judgement study where the authors manually classify a subset of articles according to their perceived uncertainty. This provides a benchmark to assess how well each method captures various types of uncertainty. Our results show that the GPT-4-generated labelling aligns more closely with human judgment than the traditional methods do, demonstrating superior precision in identifying nuanced expressions of uncertainty.

This highlights the ability of LLMs to understand and interpret context in ways that traditional methods cannot, providing a more accurate and reliable tool for measuring uncertainty from textual

data. Leveraging the contextual understanding capabilities of LLMs allows us to detect nuances that BoW methods overlook, thereby demonstrating the superior precision and reliability of LLMs in capturing uncertainty. Figure 1 illustrates the development of LLM-based and BoW-based indices over time, supporting our claim that the LLM-based index captures elements of uncertainty missed by BoW methods.¹ The shaded areas indicate periods when the discrepancies between the LLM-based and BoW-based indices are particularly pronounced. For example, significant events such as 9/11 and the Iraq invasion caused more substantial spikes in the LLM-based economic policy uncertainty index than did the BoW-based index. Conversely, the monetary policy index showed minimal change. During the 2008-2009 global financial crisis, the LLM-based indices spiked earlier and remained elevated longer than the BoW-based indices did. These examples indicate that our LLM-based indices better capture and contextualise uncertainty, providing a more nuanced understanding than BoW-based indices.

Finally, we investigate whether the LLM-based uncertainty indices are associated with changes in macroeconomic variables, shifts in financial market behaviour as measured by mutual fund flows, and asset returns. We employ different econometric analyses to assess the predictive power of the uncertainty indices and compare the results with those obtained from traditional uncertainty measures.

Our results show that LLM-based uncertainty indices are more strongly associated with macroeconomic variables, mutual fund flows, and asset returns than traditional indices are. Using a vector autoregression (VAR) model, we examine how uncertainty shocks impact macroeconomic variables such as industrial production and employment. The impulse response functions demonstrate that uncertainty shocks, as measured by our LLM-based indices, exhibit clearer, more pronounced, and sustained effects on industrial production and employment than traditional indices do. Additionally, we find that increased uncertainty, as captured by LLMs, correlates with shifts in mutual fund holdings from riskier to safer assets, reflecting investors' risk-averse behaviour during periods of heightened uncertainty. Specifically, a one standard deviation increase in economic policy uncertainty leads to an immediate outflow in equity funds of 0.38% (approximately 100 million USD) and inflows in government bond funds of 0.25% (approximately 123 million USD). Moreover, higher levels of different types of uncertainty are linked to both contemporaneous and future returns across different asset classes, with increased uncertainty correlating with lower equity returns and higher returns on safe-haven assets for up to three months. These findings suggest that LLM-generated uncertainty indices not only enhance our understanding of market sentiments but also provide valuable tools for forecasting macroeconomic variables and financial market trends, offering practical insights for investors and

¹BoW-based indices have different means and standard deviations than our LLM-based indices do. For comparability, we standardise the BoW-based indices to match the means and standard deviations of the LLM-based indices. Our findings remain consistent without standardisation, and unstandardised results are available upon request.

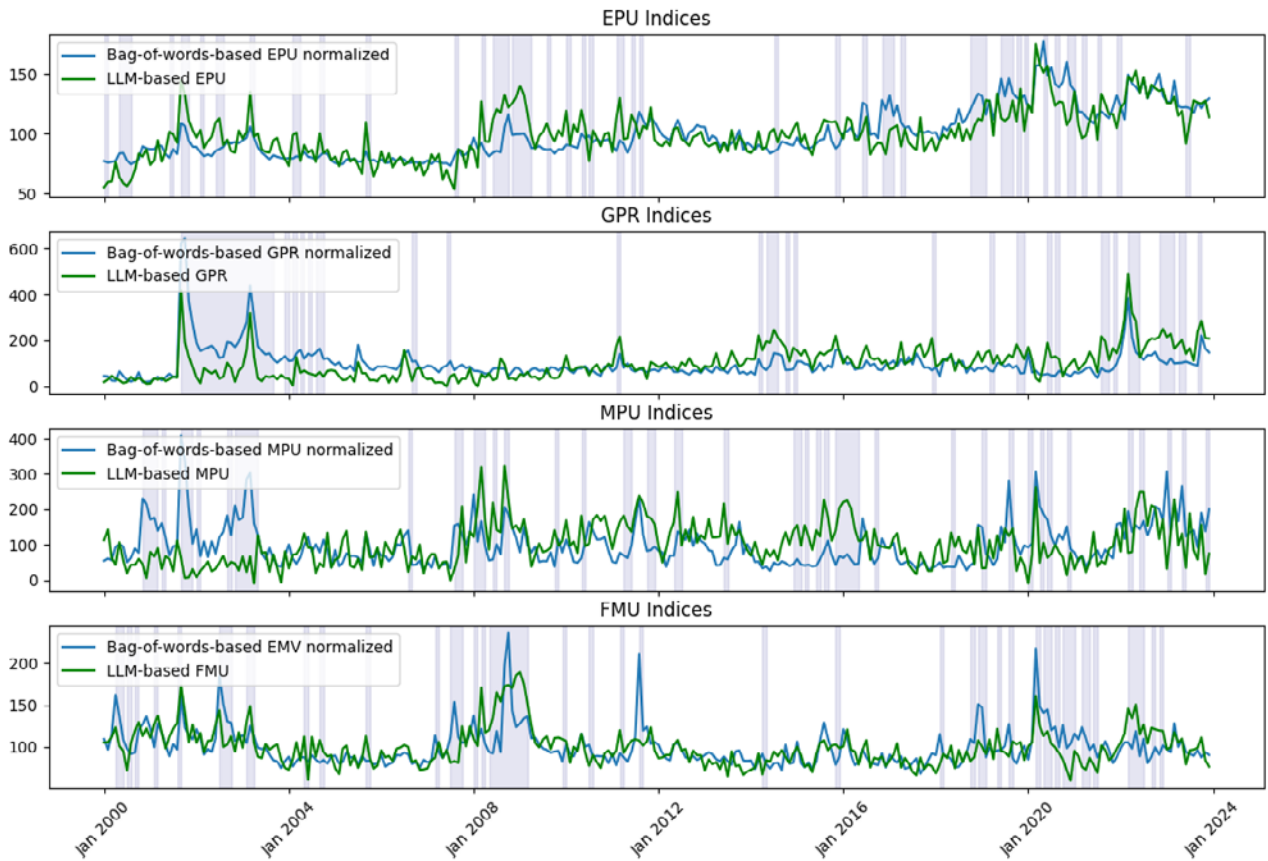


Figure 1: This figure shows the development of the BoW-based and LLM-based uncertainty indices over time. EPU stands for economic policy uncertainty (Baker, Bloom, & Davis, 2016), GPR stands for geopolitical risk (Caldara & Iacoviello, 2022), MPU stands for monetary policy uncertainty (Baker, Bloom, & Davis, 2016) and FMU stands for financial market uncertainty (Baker, Bloom, Davis, & Kost, 2019). The BoW based indices are normalised such that they have the same mean of 100 and the same standard deviation as the LLM-based indices. The shaded areas represent times in which the difference between the WSJ BoW-based indices and the LLM-based indices is above the 80th percentile.

policymakers.

We make significant contributions to multiple strands of literature. First, our work advances the methodologies for measuring uncertainty via textual data. Traditional approaches, such as the EPU index by Baker et al., 2016, rely on BoW techniques that are limited by predefined vocabularies and lack contextual understanding. Our research shows that the use of LLMs can overcome these limitations, providing a more nuanced and comprehensive analysis of textual data.

Additionally, we contribute to the literature on the impact of uncertainty on economic activity by showing that our LLM-based indices exhibit meaningful correlations with various macroeconomic variables. This improved measurement provides deeper insights into how uncertainty influences economic performance. Specifically, we observe a stronger and more pronounced relationship between economic policy uncertainty, as captured by the LLM-based indices, and economic activity compared with the BoW-based indices.

Furthermore, we advance the asset pricing literature by exploring the relationship between our refined LLM-based uncertainty indices and asset prices. Our indices offer improved precision in predicting market behaviour and asset prices, providing valuable insights for financial modelling and investment strategies. As our indices capture uncertainty in a more nuanced way, they serve as a valuable tool for investors, enabling more informed portfolio adjustments, improving risk management, and helping to capitalise on market opportunities.

Finally, our study contributes to the growing body of literature on the application of LLMs in economics and finance. We illustrate the advanced analytical capabilities of LLMs and their potential for economic forecasting and market analysis, thus highlighting the intersection of machine learning and economic research and opening new avenues for interdisciplinary studies.

The remainder of the paper is organised as follows. Section 2 provides a comprehensive review of the existing literature on uncertainty. Section 3 describes the dataset used and the methodology for constructing and evaluating the LLM-based uncertainty indices. In Section 4, we assess the effectiveness of these indices in capturing uncertainty, examining their correlation with economic activity, mutual fund flows and asset returns. Finally, Section 5 concludes.

2 Related Literature

We contribute to several strands of the existing literature. First, we contribute to the literature on methods for measuring uncertainty via text data. Significant research has been conducted on measuring uncertainty from textual data, primarily news articles. A prominent example is the EPU index introduced by Baker et al., 2016. This index analyses terms in newspapers to gauge policy-related uncertainty. To be more precise, the EPU index is calculated as the proportion of newspaper articles that include at least one term from each of three groups of words that cover "economy", "policy" and "uncertainty".² The EPU index has inspired the development of various uncertainty indices targeting different domains. For example, Baker et al. (2016) extended the EPU index to capture uncertainty surrounding monetary policy, whereas Baker, Bloom, Davis, and Kost (2019) built an index that measures equity market volatility (EMV). Caldara and Iacoviello (2022) built an index that covers geopolitical risk (GPR) and that is constructed based on the share of articles mentioning specific words related to geopolitical risks. Similar approaches have been applied to data from alternative sources, such as Twitter (Baker, Bloom, Davis, & Renault, 2021), "The Economist"

²The "economy" group contains the words "economic", "economy", "economics", and "economies"; the "policy" group contains "congress", "deficit", "Federal Reserve", "the Fed", "legislation", "regulation", "regulatory", "regulator", "regulators", and "White House" and the "uncertainty" group includes "uncertain", "uncertainty", and "uncertainties".

intelligence reports (Ahir, Bloom, & Furceri, 2022), and infectious disease reports (Baker, Bloom, Davis, Kost, Sammon, & Viratyosin, 2020). Additionally, country-specific indices based on this BoW methodology have been developed.³

However, the limitation of the BoW method lies in its reliance on predefined terms, which ignores contextual nuances. For instance, an article on "On-Line Banking"⁴ may contain words on economic policy uncertainty despite not being related to the topic at all. This paper addresses this limitation by using LLMs for measuring different types of uncertainty, enabling a more nuanced understanding of the textual context and surpassing the limitations of traditional methods.

Alternative measures of uncertainty, employing methodologies beyond simple word counts, have been proposed in the literature. Julio and Yook (2012) explored political uncertainty via election year dummy variables in statistical models. Jurado, Ludvigson, and Ng (2015) measured macroeconomic uncertainty through principal component analysis of various indicators. Furthermore, some prominent uncertainty indices in the literature are based on internet search volumes. Manela and Moreira (2017) utilised support vector regression to model the relationship between the internet search volume of specific keywords and the VIX, creating the NVIX. Da, Engelberg, and Gao (2015) used internet search volumes for terms such as "recession" to quantify uncertainty. Glasserman, Mamaysky, and Qin (2023) introduced an entropy-based measure of news novelty that outperforms standard measures in forecasting stock returns. While their approach uses a recurrent neural network to capture shifts in the distribution of news text over time, our method employs LLMs to capture deeper contextual relationships and semantic nuances in news articles, offering an advanced understanding of uncertainty and its market implications.

Additionally, we contribute to the literature on the impact of uncertainty on economic activity by showing that our LLM-based indices exhibit meaningful correlations with various macroeconomic variables. Similar types of analyses have been performed by Baker et al. (2016), who showed that higher economic policy uncertainty is associated with negative effects on economic activity. Bybee, Kelly, Manela, and Xiu, 2023 measured the state of the economy via textual analysis of business news and showed that news plays a large role in forecasting macroeconomic dynamics. van Binsbergen, Bryzgalova, Mukhopadhyay, and Sharma (2024) developed a text-based measure that predicts GDP, consumption, and employment growth.

Moreover, we contribute to the asset pricing literature by carefully examining how our nuanced LLM-based uncertainty indices relate to asset prices. A substantial body of past research has investi-

³For an overview of the many different indices, refer to <https://www.policyuncertainty.com>.

⁴The original article can be found here: <https://www.wsj.com/articles/SB898796563903106000>.

gated the connection between uncertainty indices and financial market returns and volatility. Many indices, including those mentioned above, have proven instrumental in explaining these links, particularly the impact of events such as elections or crises. Baker, Bloom, Davis, and Sammon (2021), for example, found that policy-related news triggers a significant portion of market jumps, further supported by similar patterns observed with Financial and Economic Attitudes Revealed by Search (FEARS) and NVIX (Da et al., 2015; Manela & Moreira, 2017). These findings highlight the pervasive effects of uncertainty on market dynamics, extending beyond stocks. Positive correlations have been observed between sentiment indices and safe haven assets such as gold, the Swiss franc and the Japanese yen (Balcilar, Bonato, Demirer, & Gupta, 2017; Nasr, Bonato, Demirer, & Gupta, 2019).

Specific events such as the COVID-19 pandemic further illustrate how adverse news can impact market returns. Mamaysky (2023) observed depressed risky asset returns and elevated volatility during the pandemic. Similarly, the appearance of unusual news predicts an increase in stock volatility (Glasserman & Mamaysky, 2019). Delving deeper into the role of the media, a study using daily content from a popular WSJ column revealed that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals. Furthermore, periods of unusually high or low pessimism are associated with high market trading volume, underscoring the nuanced influence of media sentiment on market dynamics (Tetlock, 2007).

Finally, we contribute to the rapidly growing literature on the use of LLMs in economics and finance. Recent advancements have paved the way for the innovative application of LLMs in financial prediction. Jiang, Kelly, and Xiu (2023) employed LLMs to predict stock returns, revealing their consistent superiority over traditional word-based models. Furthermore, Lopez-Lira and Tang (2023) utilised models such as ChatGPT to predict stock market returns through sentiment analysis of news headlines, thus exemplifying the potential and value of employing sophisticated tools such as ChatGPT in the domain of financial market forecasting. Bybee (2023) examined the ability of ChatGPT to form expectations from news data. In addition to ChatGPT, many other capable LLMs have been introduced, many of which can be used in an open-source setting. Notable examples include Meta’s LLaMa (Touvron et al., 2023) and Zephyr (Tunstall et al., 2023). Additionally, many publications have shown how to leverage these models at relatively low costs to build effective LLMs. For example, Dettmers, Pagnoni, Holtzman, and Zettlemoyer (2023) demonstrated how complex multibillion parameter models can be fine-tuned on a single GPU. Zhang, Yang, and Liu (2023) and Zhang et al. (2023) demonstrated the strong performance of fine-tuned open-source models. Recent advances in the research on LLMs have allowed us to create our own effective and transparent model at a low cost. In contrast to previous applications of LLMs in finance and economics, this is, to the best of

our knowledge, the first study to use open-source LLMs to classify uncertainty based on newspaper articles. Our insights demonstrate the effectiveness of these new methods for quantifying uncertainty.

3 Construction of Uncertainty Indices

In this section, we outline the data and methodology used to construct uncertainty indices from textual data with an LLM. We aim to create robust, reproducible indices that capture various dimensions of uncertainty over significant period of time.

3.1 Data

We use newspaper articles from the WSJ, a prominent U.S. newspaper renowned for its comprehensive business and financial news coverage. The WSJ is distinguished by its commitment to journalistic excellence, exemplified by rigorous editing processes and high standards.⁵ Our dataset includes all the articles published from 1998 until 2023, thus covering 26 years. This period encompasses several major economic events and policy shifts, providing valuable insights for our analysis.

Our dataset includes both article titles and full texts, allowing us to leverage the capabilities of an LLM based on the entire newspaper text. We exclude articles with fewer than ten words to filter out items such as crossword puzzles or photo galleries. Additionally, to reduce the computational effort for our LLMs, we remove articles that contain more than 4,096 tokens.⁶ Articles exceeding 4,096 tokens account for less than 0.3% of the total data. Consequently, our dataset comprises 1,179,967 articles containing a total of 771,174,808 words. Table 1 provides a detailed overview of the data.

	Mean	Standard Deviation	Min	Max	Sum
Articles per Month	4,311.8	1,080.9	1,615	6,784	1,179,967
Words per Article	653.6	387.9	10	2,609	771,174,808

Table 1: Summary statistics of the WSJ text dataset used in this study, covering the period from 1998 until 2023.

Figure 2 shows the number of monthly articles and words over time, along with the average number of words per article in recent years. Notably, the WSJ maintained a relatively stable number

⁵For example, the WSJ is recognised as a newspaper of record; for details, see <https://libguides.mcmaster.ca/news/record>. Additionally, the newspaper has received numerous awards for good journalism, including 39 Pulitzer Prizes.

⁶LLMs process text data in the form of tokens. In English, a token typically consists of approximately four characters. On average, 4,096 tokens are approximately 3,000 words. For more details, see <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>. The performance of LLMs in processing texts with a large number of tokens varies depending on the location of the relevant information in the text (Liu et al., 2024). For WSJ articles, the most important information is usually placed at the beginning. Liu et al. (2024) show that LLMs perform best when the relevant information is at the beginning or end of a text.

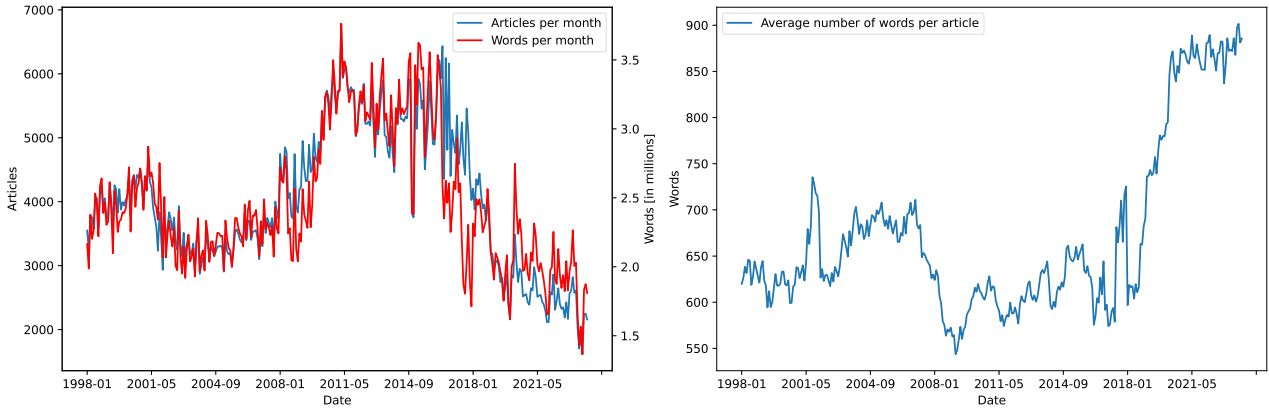


Figure 2: In this figure, the left-hand side panel displays the monthly number of articles and the monthly number of words in the WSJ dataset over time. The right-hand side shows the average word count per article over time.

of equally long articles over time. However, since approximately 2019, there has been a trend toward fewer but longer articles. The total number of articles published in the WSJ has decreased slightly, and despite the longer average article length, the overall word count has slightly decreased. One explanation is that after 2019, the WSJ stopped publishing short articles, such as company announcements containing only 50 words or less.⁷ Nevertheless, the remaining articles tend to be consistently longer after 2019, as shown in Figure A.1 in the Appendix.

Unlike other applications of WSJ text data⁸, our dataset includes any article exceeding ten words without excluding content based on the publication date or other criteria. This approach results in, to the best of our knowledge, the most complete and up-to-date dataset for WSJ text, covering the past 26 years.

3.2 Construction of LLM-based Uncertainty Indices

To explore the capabilities of contemporary LLMs in detecting uncertainty in newspaper texts, we choose two open-source models: LLaMa-2-7B-Chat and Zephyr-7B- β . These models are fine-tuned via a dataset constructed with GPT-4 Turbo. The models’ performance is then benchmarked against GPT-4 Turbo, and the best-performing open-source LLM is used for index construction. Our selection criteria consider model size, accessibility, and performance in existing benchmarks (see Section B.1 in the Appendix for more details on the model selection).

LLMs such as LLaMa-2 (Touvron et al., 2023) and Zephyr (Tunstall et al., 2023) demonstrate

⁷See, for example, <https://www.wsj.com/articles/funding-snapshot-german-lender-smava-receives-65-million-1515513785>.

⁸For example, Bybee et al. (2023) and Bybee (2023) utilised the same WSJ text data spanning from 1984 to 2017 and 2021, respectively. However, they excluded articles published on weekends, articles with fewer than 100 words, and articles not covering economic content. Additionally, they used other criteria to remove articles, such as excluding those without page-citation tags.

remarkable proficiency across a broad spectrum of tasks. However, their true potential lies in their adaptability to specialised objectives. An important consideration in our study is the commitment to keeping our model easily reproducible, promoting transparency and accessibility. We prioritise efficiency, avoid unnecessary demands on computational resources, and focus on models that are openly available. This approach aligns with best practices for reproducibility and facilitates wider adoption and understanding of our work within the research community.

Why do we not use OpenAI’s GPT-4 Turbo for index construction? It is widely regarded as one of the industry’s leading models and is often used as a benchmark (Zhang et al., 2023; Zhou et al., 2023). However, it is proprietary, and querying it via API incurs costs, making it impractical for our emphasis on transparency and replicability. Nonetheless, to fine-tune and compare the performance of LLaMa-2-7B-Chat and Zephyr-7B- β , we query a subset of the WSJ data through the API of GPT-4 Turbo. It would have been too costly for a human to read and label 20,000 full-text articles, as this is time-consuming, and querying the articles via API only costs approximately 300 USD at the time of the query. The GPT-4 Turbo generated data are then used to fine-tune and benchmark the other models. For more details on the fine-tuning dataset, see Section B.2 in the Appendix.

For classification, we construct a prompt to instruct the LLMs on definitions of uncertainty, assessing whether an article implies a rise, fall or an unchanged level of uncertainty or whether it influences it at all. We also ask for the magnitude of uncertainty, which is quantified on a scale from 0 to 1, and the model’s confidence level in its response. Additionally, the prompt categorises the news, identifies the global or regional implications of the event, and specifies the predominantly affected asset. This comprehensive set of inquiries aims to extract nuanced and detailed information regarding the impact of news articles on uncertainty across various dimensions. We experimented with different variations of the prompt and focused on the one that yielded results in line with our human assessment. The resulting prompt is used to create the instruction fine-tuning datasets via GPT-4 Turbo and for the subsequent classification that is necessary for the index generation. The exact prompt is detailed in Figure B.2 in the Appendix.

After creating the fine-tuning dataset, we compare the results from our GPT-4 dataset, as described in Table B.1 in the Appendix, to those of the traditional BoW method pioneered by Baker et al. (2016) and a human evaluation of the articles to ensure that the GPT-4 Turbo-generated labels align closely with human judgement. We focus on the EPU index, as it is the most common index used in the literature and by practitioners, and assume that the results are similar for the other types of uncertainty. Table A.1 in the Appendix displays the results of the comparison between the classification by GPT-4 Turbo and the BoW method. In most cases, the GPT-based EPU and the EPU, as

in Baker et al. (2016) (BoW-based EPU), deliver the same result. A total of 61.9% of all articles are identified by both methods to affect economic policy uncertainty similarly (true/true or false/false). However, the results show that in many cases, an article is classified by GPT-4 as affecting economic policy uncertainty (38.8%), whereas the BoW-based EPU only identifies very few articles (1.3%). As a result, we have a small set of articles (0.3%) where, compared with GPT’s classification, the traditional method delivers a ”false positive” and a larger set of articles (37.8%) of ”false negatives” (see Figure A.2 in the Appendix).

To investigate which method aligns more closely with human judgment in capturing economic policy uncertainty, we conduct a manual classification of 110 articles. Our focus is on potential misclassifications; thus, we focus on articles where the two models yield different outcomes. We select all 55 articles that are classified as ”true” by the BoW method, whereas GPT-4 classifies them as ”false”. We then randomly sample an additional 55 articles from the pool of 7,271 where GPT-4 classifies them as ”true” while the BoW does not.⁹ For the human evaluation, we then read all the articles and classify whether they affect economic policy uncertainty according to their definition.¹⁰ We then compare the results of the human evaluation with the model predictions to evaluate which model has a better fit in line with the human perception of uncertainty. Table 2 shows the results. GPT-4 does a much better job of identifying economic policy uncertainty in terms of accuracy. On average, more than 80% of the investigated articles are classified in line with the human judgement by GPT-4, whereas the traditional method is only in line with the human judgement in less than 20% of the articles. Interestingly, GPT-4 performs particularly well in the second sample of articles, which consists of the articles that are classified as ”true” by GPT-4 but not captured by the BoW method. This is an indication that the original BoW method can be too restrictive in identifying all articles that affect economic policy uncertainty. Overall, the results support our hypothesis that an LLM-based approach can deliver improved results for identifying uncertainty based on newspaper articles.

The traditional BoW method has several shortcomings in accurately capturing economic policy uncertainty. First, as highlighted by Baker et al. (2016), approximately 5% of newspaper articles classified by this method reflect decreasing rather than increasing uncertainty. Thus, for example, an article

⁹To ensure representativeness, we ensure in the sampling process that we select at least one article per month over the total period.

¹⁰We use the same definition as in Baker et al. (2016) and the LLM prompt in this paper. EPU is defined as follows: ”Uncertainty over who makes or will make policy decisions that have economic consequences. Current and past uncertainty over what economic policy actions will be undertaken. Uncertainty over the economic effects of policy actions in the past, present or future. Economic uncertainty induced by policy inaction or related to policy developments or motivated by non-economic considerations.”

Period	BoW-based EPU	GPT-based EPU	N
Sample 1	0.27	0.73	55
Sample 2	0.09	0.91	55
Combined	0.18	0.82	110

Table 2: This table depicts the accuracy of the prediction of the BoW-based EPU and the GPT-based EPU with respect to economic policy uncertainty compared with a human evaluation of 110 articles. Sample 1 consists of all articles that are classified as "false" by the GPT-based EPU and as "true" by the BoW-based EPU in Table A.1 in the Appendix. Sample 2 includes 55 randomly drawn articles from a subset of articles that are classified positively by GPT-4 and narrowly negatively by the BoW method. Combined includes both samples.

that explicitly mentions lower uncertainty following an election outcome¹¹ is misclassified as describing increasing economic policy uncertainty because of the presence of keywords such as "economics," "policy," and "uncertainty." Second, the BoW method struggles with context-based misclassifications. For example, articles may include all relevant keywords from the three economic policy uncertainty categories but in an unrelated context, such as discussing hypothetical events or using negations. Examples include articles about weather forecasting or the introduction of "online banking" in 1998¹² and historical analyses, such as those arguing Franklin D. Roosevelt was the greatest leader of the 20th century¹³. Third, the BoW method is overly restrictive in its classification criteria. For example, an article must contain one of the terms "uncertain," "uncertainty," or "uncertainties" to be classified as relevant. However, many articles discuss economic policy uncertainty without explicitly using these terms. In our sample, out of the 7,271 articles classified as "true" by GPT-4 but "false" by the BoW method, approximately one-third (1,911) include words from two of the three required categories but lack one keyword, often "uncertainty." A large portion of these articles (1,591) do not explicitly mention "uncertainty" but still discuss inherently uncertain topics. For instance, an article¹⁴ on the U.S. government's response to the mortgage crisis highlights the ambiguity in future policy actions but does not use the term "uncertainty." Given these points, it is clear that LLMs offer greater precision and contextual awareness. They are not limited by the strict keyword constraints of BoW and better capture the nuances of uncertainty, as humans interpret it from newspaper articles.

After creating and evaluating the fine-tuning dataset, we use it to fine-tune the two open-source LLMs: LLaMa-2-7B-Chat and Zephyr-7B- β . Previous studies (Zhang et al., 2023; Zhou et al., 2023) demonstrated the effectiveness of fine-tuning open-source LLMs for various tasks. We perform an instruction fine-tuning of the two open-source LLMs for classifying uncertainty based on newspaper

¹¹See <https://www.wsj.com/articles/SB943700739275961154>.

¹²See <https://www.wsj.com/articles/SB894256527525075500> and <https://www.wsj.com/articles/SB898796563903106000>.

¹³See <https://www.wsj.com/articles/SB946504042719485841>.

¹⁴See <https://www.wsj.com/articles/SB122641622440217445>.

articles, considering the evolving nature of financial language over the last 25 years. Initial fine-tuning uses randomly selected data from 1998 and 1999. We subsequently use this fine-tuned LLM to construct uncertainty indices from 2000 to 2009, followed by further fine-tuning with randomly selected data from 2008 and 2009 to construct uncertainty indices until the end of 2023. This approach ensures that the LLMs are fine-tuned to our specific task while reflecting some capabilities of GPT-4 Turbo. During fine-tuning, the models learn not only the appropriate classification but also the instructions, ensuring that they adopt the required formatting for their responses (see Section B.3 in the Appendix).

To assess the performance of our fine-tuned LLMs, we compare the fine-tuned LLaMa-2-7B-Chat and Zephyr-7B- β models to their non-fine-tuned base models using a subset of data classified by GPT-4 Turbo. The main evaluation metrics are accuracy and the F1 score, which are common measures in the literature. Accuracy indicates the percentage of predictions that are correct overall, whereas the F1 score is the harmonic mean of precision and recall, which jointly minimises the risk of Type I and Type II errors. As a result, a high F1 score is desirable, and typically, an F1 score greater than 0.7 is considered strong (Powers, 2011).

GPT-4 Turbo, with its extensive parameter count and cutting-edge performance, serves as an ideal benchmark for fine-tuned open-source models. Comparing the selected models against GPT-4 allows us to not only measure their relative performance but also to understand how these open-source models stand against one of the highest-performing LLMs. This comparison is crucial for assessing the viability and potential of open-source LLMs in performing complex language tasks. Assuming that the accuracy of GPT-4 Turbo is very high, as the model is currently considered one of the best-performing LLMs available, which our human evaluation study confirmed, we start the performance evaluation by benchmarking the fine-tuned and non-fine-tuned LLaMa-2-7B-Chat and Zephyr-7B- β models against the classification performed by GPT-4 Turbo. The results are shown in Table 3. Notably, the fine-tuned Zephyr-7B- β model consistently outperforms the other models across all uncertainty categories, achieving the highest scores in terms of both accuracy and the F1 score. This is evident in almost all uncertainty categories, achieving high accuracy and F1 scores of over 0.80 in most categories. In contrast, the LLaMa-2 model shows moderate to low performance. Fine-tuning significantly improves its performance, although it remains inferior to the fine-tuned Zephyr-7B- β . Zephyr-7B- β in its original form, demonstrates varied performance. However, its fine-tuned version shows remarkable performance, outshining the other models in almost all categories.

The results indicate that the fine-tuned Zephyr-7B- β model is the best-performing model in this study. The model’s fine-tuning process plays a crucial role in achieving superior accuracy and F1 scores across diverse uncertainty scenarios. Given these results, we use the fine-tuned Zephyr-7B- β for

Type of Uncertainty	Model	1998/1999		2008/2009	
		Accuracy	F1 Score	Accuracy	F1 Score
EPU	LLaMa-2-7B-Chat	0.30	0.47	0.44	0.61
	Fine-tuned LLaMa-2-7B-Chat	0.42	0.51	0.66	0.69
	Zephyr-7B- β	0.42	0.50	0.60	0.65
	Fine-tuned Zephyr-7B- β	0.70	0.74	0.83	0.84
GPR	LLaMa-2-7B-Chat	0.70	0.70	0.64	0.63
	Fine-tuned LLaMa-2-7B-Chat	0.94	0.96	0.95	0.96
	Zephyr-7B- β	0.35	0.20	0.34	0.19
	Fine-tuned Zephyr-7B- β	0.97	0.98	0.98	0.98
MPU	LLaMa-2-7B-Chat	0.09	0.18	0.13	0.23
	Fine-tuned LLaMa-2-7B-Chat	0.84	0.91	0.84	0.87
	Zephyr-7B- β	0.17	0.15	0.21	0.16
	Fine-tuned Zephyr-7B- β	0.89	0.92	0.94	0.95
FMU	LLaMa-2-7B-Chat	0.29	0.36	0.36	0.43
	Fine-tuned LLaMa-2-7B-Chat	0.53	0.59	0.66	0.69
	Zephyr-7B- β	0.52	0.59	0.54	0.59
	Fine-tuned Zephyr-7B- β	0.76	0.78	0.81	0.82

Table 3: Comparison of Accuracy and F1 Score Benchmarked Against GPT-4 Turbo. EPU stands for economic policy uncertainty (Baker, Bloom, & Davis, 2016), GPR stands for geopolitical risk (Caldara & Iacoviello, 2022), MPU stands for monetary policy uncertainty (Baker, Bloom, & Davis, 2016) and FMU stands for financial market uncertainty (Baker, Bloom, Davis, & Kost, 2019). Note that the prompt used for the inference with the non-fine-tuned LLaMa-2 model is slightly different than the prompt for the other models, as the non-fine-tuned LLaMa-3 model otherwise creates answers that are in a random format, making it very difficult to extract the relevant numbers.

index construction.

To decrease the computational resources needed, we randomly select 10% of the WSJ articles for the index creation and classify them.¹⁵ We then construct our different indices of uncertainty every month in a similar way as in Baker et al. (2016). However, as our model classifies the data with greater granularity, we have different options to calculate the index. Specifically, we calculate the indices in three different ways:

1. Define X_t as the monthly count of articles where the selected uncertainty index is "increasing".
2. Define $X_t = \sum_{i=1}^n \sqrt{Y_i Z_i}$, where Y_i is the estimate of the magnitude and Z_i is the estimate of

¹⁵For a robustness check, we selected an additional 10% of the WSJ articles from the period 2000 to 2009 and compared the indices constructed using this subset with those based on 20% of the data. The comparison reveals that the indices derived from the larger dataset exhibit only marginal differences. Specifically, the correlation between the two sets of indices ranges from 0.91 to 0.97, and the average absolute difference between them is less than one-third of one standard deviation of the respective indices.

the confidence of article i for every article $i = 1, 2, \dots, n$ in month t where the uncertainty index is "increasing".

3. Define $X_t = U_t - D_t$, where U_t is the X_t defined under 2. and where D_t is the same definition but for articles where the uncertainty index is "decreasing".

We then scale X_t by dividing it by the total number of articles per month. Finally, we compute the mean M of the series and normalise it by multiplying X_t with $(100/M)$.

3.3 Evaluation of LLM-based Uncertainty Indices

We start the evaluation of the resulting uncertainty indices by simply analysing the resulting indices visually. The indices are plotted in Figure 3. All index computation methods lead to similar results for all uncertainty categories. The shaded areas represent significant periods in which all index computation methods result in an index value above the 90th percentile. We now proceed to discuss the development of each type of uncertainty index over time.

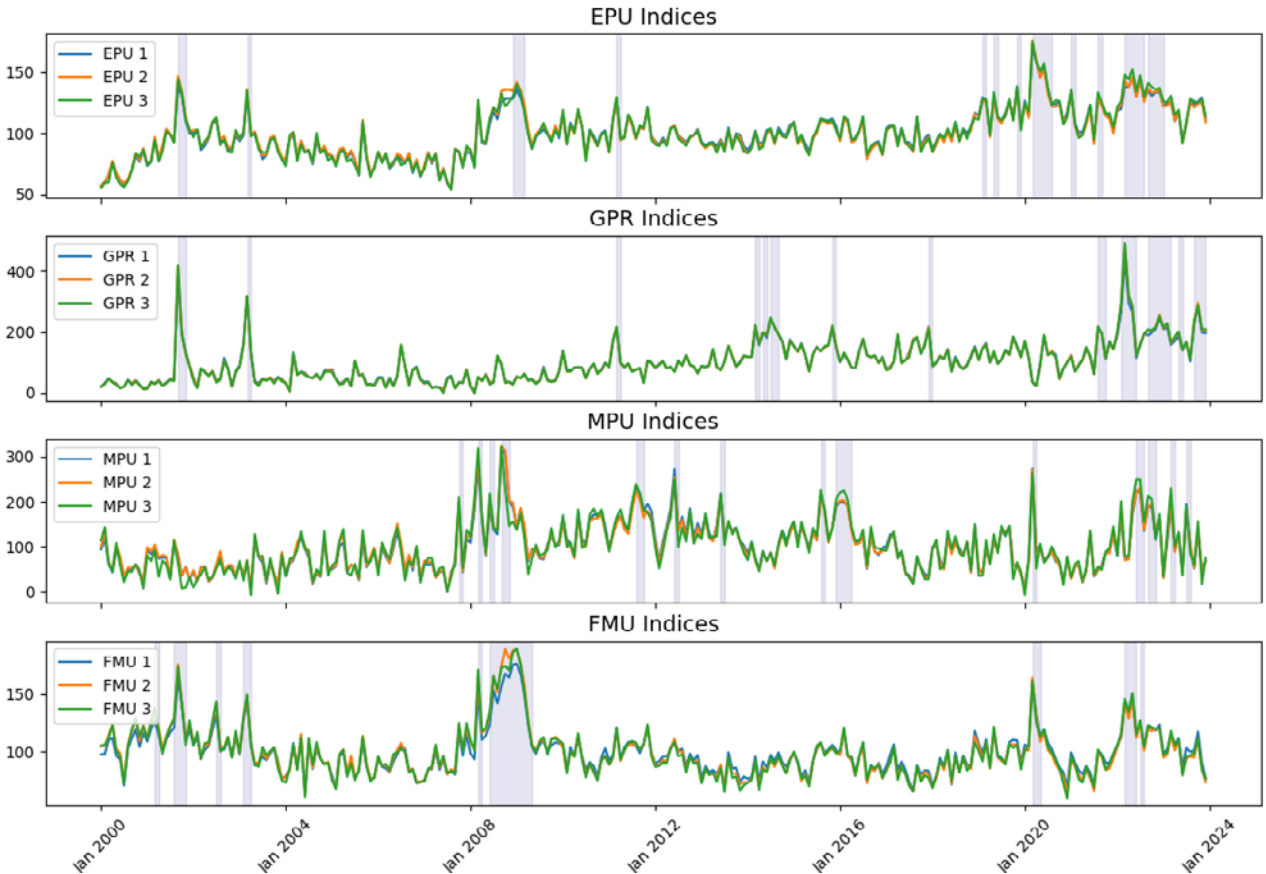


Figure 3: This figure shows the development of the different uncertainty indices computed by slightly different methods over time. 1, 2 and 3 correspond to the methods described in Section 3.2. The shaded areas represent times in which all index computation methods result in index values above the 90th percentile.

Economic Policy Uncertainty (EPU): The shaded spikes in the EPU indices over the years coincide with major global events. In September 2001, the EPU indices surged due to the terrorist attacks on the World Trade Center and the Pentagon. In March 2003, a notable rise in uncertainty was again observed, aligning with the onset of the Iraq War. The spike that started in December 2008 can be directly linked to the global financial crisis. In March 2011, the EPU indices peaked, likely in response to the Arab Spring. The beginning of 2019 saw a series of increases in uncertainty, with spikes starting in February, May, and November, which may be related to escalating trade tensions between the U.S. and China. This pattern of volatility continued into the subsequent years with the spread of COVID-19 in March 2020. In January 2021, the indices again reflected significant economic policy uncertainty amid the political transition in the U.S. and the ongoing pandemic. By August 2021, the focus shifted to the economic implications of the U.S. military withdrawal from Afghanistan and the surge of the COVID-19 delta variant. The March 2022 spike corresponded with the invasion of Ukraine by Russia, whereas the September 2022 increase in the EPU indices was possibly a result of persistent global economic challenges, including inflation and regional geopolitical conflicts.

Geopolitical Risk (GPR): The shaded regions in the GPR indices coincide with moments of heightened geopolitical tensions. The September 2001 rise was tied to the 9/11 terrorist attacks, which had profound and far-reaching geopolitical consequences. In March 2003, the escalation was due to the U.S.-led invasion of Iraq, initiating a long-term military engagement with broad geopolitical implications. The Arab Spring, beginning in late 2010 and continuing into March 2011, brought widespread social and political upheaval across the Arab world, contributing to a noticeable spike. The annexation of Crimea by Russia in March 2014 and the subsequent conflict in Eastern Ukraine during mid-2014 significantly increased geopolitical risks. November 2015 saw a peak corresponding to the Paris terrorist attacks, which intensified global concerns about terrorism. The situation in Afghanistan in August 2021, particularly the takeover by the Taliban, sharply increased global geopolitical risks. The invasion of Ukraine by Russia in February 2022 further heightened these risks.

Monetary Policy Uncertainty (MPU): The shading in the MPU indices suggests periods of significant monetary policy uncertainty. From October 2007 to September 2008, peaks coincide with the onset and escalation of the global financial crisis. Another notable spike occurred in August 2011, indicating renewed concerns about economic stability, likely influenced by events such as the European sovereign debt crisis. The period from June 2012 to June 2013 shows spikes reflecting uncertainties surrounding the eurozone crisis and the tapering of quantitative easing by major central banks, highlighting market volatility and anxieties about policy effectiveness. The peaks from August 2015 to December 2015 suggest increased uncertainty surrounding central bank policy decisions amid

concerns about slowing global growth and efforts to normalise monetary policy. The spike in March 2020 coincides with the onset of the COVID-19 pandemic and the subsequent economic downturn. From June 2022 to July 2023, spikes likely indicate ongoing uncertainty surrounding central bank policy responses to evolving economic conditions, including concerns about inflationary pressures and the pace of interest rate hikes.

Financial Market Uncertainty (FMU): In March 2001, the bursting of the dot-com bubble triggered heightened volatility, prompting a spike in the FMU indices. Another notable spike occurred in August 2001, likely reflecting continued concerns following the dot-com crash and increased uncertainty after the September 11 terrorist attacks. July 2002 saw another increase in volatility, possibly linked to economic uncertainties and corporate scandals such as the Enron scandal. The spike in February 2003 coincides with rising tensions surrounding the Iraq War and broader geopolitical uncertainties, contributing to market uncertainty. March 2008 and June 2008 witnessed spikes marking the onset of the global financial crisis. In March 2020, the outbreak of the COVID-19 pandemic caused a significant market crash, driving a spike in uncertainty. March 2022 saw another spike, likely due to concerns about inflation, rising interest rates, and geopolitical tensions, which contributed to market instability.

All three index computation methods lead to similar results, but the third method, as described in Section 3.2, is the most concise method reflecting increasing and decreasing uncertainty as well as confidence and magnitude, which is why we focus on the indices computed via the third method in the following.

We further evaluate our indices by comparing them to the indices generated via the BoW method as pioneered by Baker et al. (2016). First, we graphically compare the LLM-based indices to the BoW-based indices. Figure 1 shows the development of the LLM-based and BoW-based indices over time. Where the spikes in the indices are aligned, it suggests a consensus in recognising significant uncertainty events. However, the shaded areas represent times in which the difference in the indices is particularly high. For example, during the 9/11 terrorist attacks that occurred in September 2001 as well as the U.S.-led invasion of Iraq in March 2003, the LLM-based EPU spiked more than the BoW-based EPU did, whereas the LLM-based GPR spiked marginally less than the BoW-based GPR did, and the LLM-based MPU almost did not spike at all. The developments of the LLM-based indices seem intuitive, as the events that occurred are less connected to monetary policy uncertainty and more connected to geopolitical risk and economic policy uncertainty. Increased economic policy uncertainty stems from the events that occurred, which is why it makes sense that not only the GPR index indices but also the EPU index spiked during these events. Furthermore, during the global financial crisis

in 2008-2009, the LLM-based EPU, MPU and FMU spiked earlier and remained high for a longer period than their BoW-based counterparts did, indicating that the LLM-based indices captured the uncertainty during the global financial crisis more effectively. This underscores the LLM’s superior ability to contextualise and parse uncertainties, potentially offering an enhanced metric over the BoW model.

After comparing the indices graphically, we now examine their correlations. The correlations between our LLM-based indices and the BoW indices based on the same WSJ dataset are moderate: EPU stands at 0.35, GPR at 0.44, EMV/FMU at 0.42, and MPU at 0.24. These numbers are in line with our projections, considering the differing methodologies employed to quantify uncertainty. The correlations highlight both commonalities and differences in the detection of pivotal uncertainty events by each approach. The moderate correlations suggest that the LLM’s nuanced interpretation of uncertainty diverges from that of conventional models. This discrepancy may be attributed to the LLM’s capacity for deep linguistic analysis, which allows it to pick up on subtleties that the BoW method may overlook. The correlations between the LLM-based indices and the normalised BoW indices based on the original dataset used by Baker et al. (2016) are greater (except for the MPU): EPU stands at 0.75, GPR at 0.49, EMV/FMU at 0.58, and MPU at 0.14, as presented in Figure 4. The higher correlation suggests that, despite focusing on only 10% of the articles in the WSJ dataset, the LLM-based indices capture nuances of uncertainty more similarly to the BoW indices derived from the full dataset of 10 newspapers than to the BoW indices based on 10% of the WSJ articles.

Figure 4 additionally shows the correlation between the LLM-based uncertainty indices and other publicly available uncertainty measures, such as the VIX, MOVE index, Citi Economic Surprise Index (CESI), and Michigan Consumer Sentiment Index. The figure highlights that our LLM-based indices capture a distinct and nuanced portrayal of uncertainty, as the correlations are rarely above 0.5. Furthermore, the correlation coefficient between the LLM-based FMU and the VIX is 0.68, and with the MOVE, it is 0.62, which is drastically greater than those for the other LLM-based uncertainty indices. This makes intuitive sense, as the VIX and the MOVE measure expectations of stock and bond market volatility. The FMU is a text-based index capturing financial market uncertainty, whereas the definition of the uncertainty captured in the other indices is not directly linked to financial markets.

4 Uncertainty, Economic Activity and Financial Markets

We present three applications to examine the associations between uncertainty, as measured by our LLM-based indices, and changes in macroeconomic variables, investor behaviour, and asset return

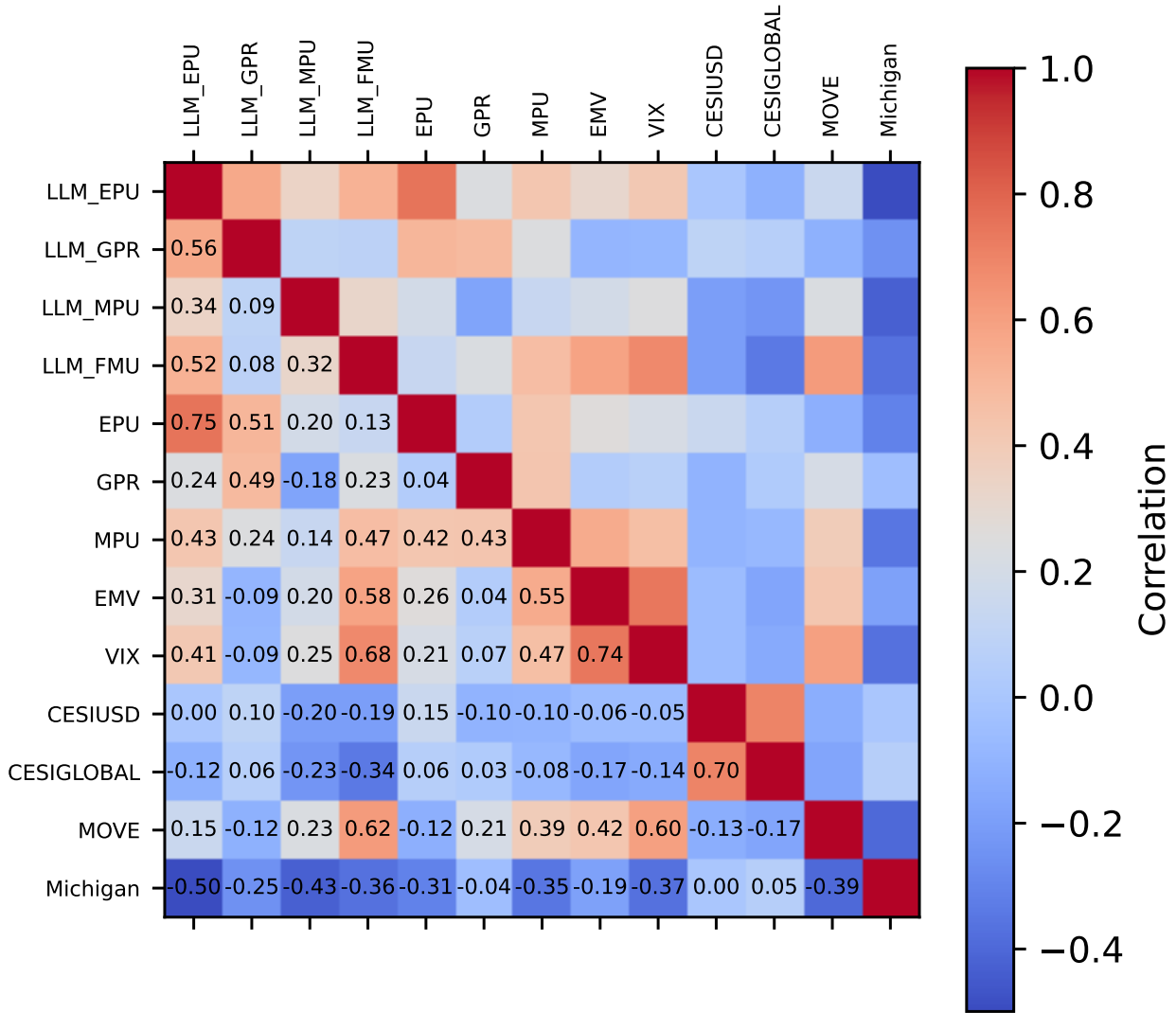


Figure 4: This figure displays the correlations among the various indices from 2000 to 2023, encompassing established measures such as economic policy uncertainty (EPU), geopolitical risk (GPR), monetary policy uncertainty (MPU), and equity market volatility (EMV). Additionally, it includes our LLM-based versions of these indices denoted with "LLM_". The figure also features the VIX, the Citi Economic Surprise Index (CESI) for both the U.S. and global economies, the MOVE index, and the Michigan Consumer Sentiment index.

variations. First, we investigate whether uncertainty shocks correlate with weaker macroeconomic performance, controlling for standard macroeconomic and policy variables. Second, we explore the impact of different types of uncertainty on investor behaviour. Specifically, we analyse whether greater uncertainty, as indicated by our indices, leads to mutual fund flows into more or less risky assets, and we assess the presence of a "flight-to-safety" effect during periods of heightened uncertainty. Finally, we examine the influence of various types of uncertainty on future returns, investigating whether increased uncertainty is predictive of the future performance of specific asset classes. In each of these applications, our LLM-based indices capture effects that significantly differ from those identified via traditional BoW indices. The LLM-based indices are associated with more pronounced effects, highlighting the added value of our novel approach.

4.1 Economic Activity

To explore the dynamic relationships between uncertainty shocks and economic activity, we employ a VAR model akin to the framework proposed by Baker et al. (2016). VAR models are well suited for analysing the interactions among multiple time series variables over time without imposing a predefined structural form. By using the same model as Baker et al. (2016), we facilitate direct comparisons with their findings and demonstrate the added value of our approach. However, VARs pose challenges in establishing causality owing to the need for assumptions regarding variable interactions. Nonetheless, they offer insights into how uncertainty levels and macroeconomic variables correlate with future economic activity.

Our baseline VAR specification uses monthly data on our LLM-based economic policy index, the S&P 500, effective federal fund rates, U.S. industrial production (IP) and U.S. employment (EMP) from 2000 until 2023. All the data, except for our policy index, are sourced from Bloomberg. We use three lags and take the natural logarithm of S&P500, IP and EMP. To identify the shocks to uncertainty, we employ a Cholesky decomposition. This ordering assumes that uncertainty shocks occur independently and precede changes in economic variables. Specifically, we order the variables in the baseline model as EPU, S&P500, the fed funds rate, EMP, and IP.

Following estimation, we conduct impulse response analysis to explore how a shock to the LLM-based EPU index influences economic activity over time. To facilitate comparison with Baker et al. (2016), we simulate a shock equivalent to the index's change from its average before the 2008 financial crisis (2005 to 2006) to its average during the subsequent high-uncertainty period (2011 to 2012).¹⁶ Figure 5 displays the impulse responses for industrial production and employment.

¹⁶This is equal to a positive innovation of 23.3 points, which is almost equal to one standard deviation (20.7 points).

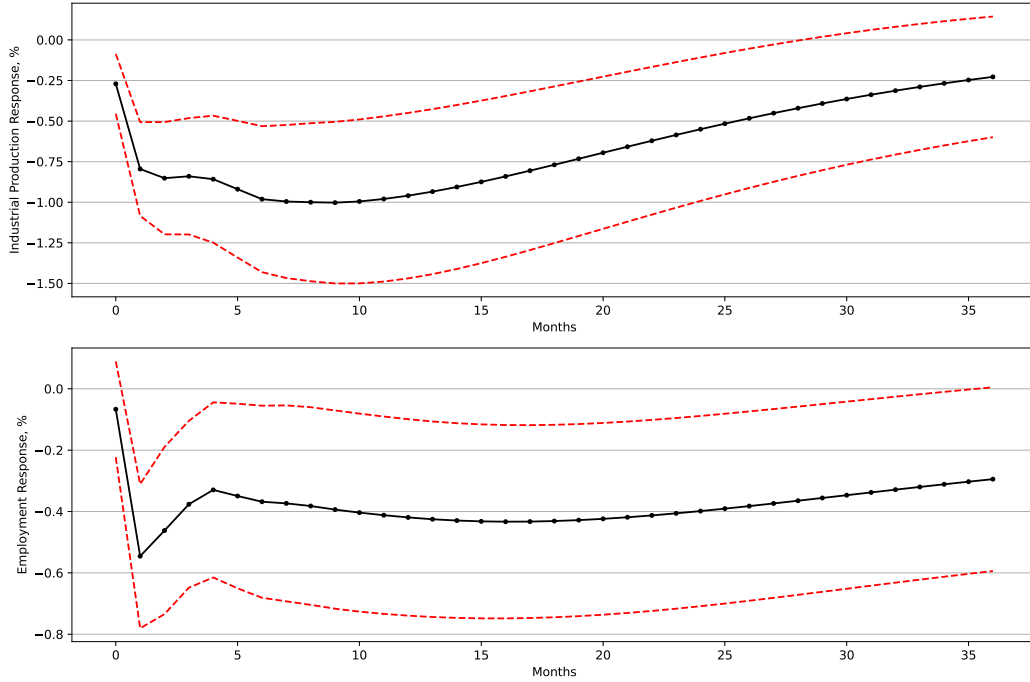


Figure 5: Impulse response functions for changes to U.S. industrial production and U.S. employment after a positive shock to our LLM-based EPU index. The impulse response functions are generated from a VAR that is estimated on monthly data with three lags for the period of 2000 until 2023. Identification is based on a Cholesky decomposition with the following ordering: EPU index, $\log(\text{S\&P 500})$, fed funds rate, $\log(\text{employment})$ and $\log(\text{industrial production})$. The solid lines represent the impulse response functions for up to 36 months ahead, and the dashed lines are 90% confidence bounds.

Our LLM-based economic policy uncertainty (EPU) index is significantly linked with subsequent adverse impacts on economic activity, specifically industrial production and employment. A positive shock to the EPU index induces an approximately 1% decline in industrial production and a 0.4% decrease in employment.

These effects align with the findings of Baker et al. (2016) for the period from 1985 to 2016. However, when their index is applied to the same VAR model for the period from 2000 to 2023, we find that their index is associated with much weaker and mostly insignificant effects on economic activity. Figure A.2 in the Appendix compares the impulse responses from two models: one using the LLM-based EPU index and the other using the original BoW-based index. Additionally, the model with the LLM-based index has a better in-sample fit, improving the predictions of EP and IP by 1% and 1.5%, respectively, as measured by the mean-squared error. We also conducted the same analysis via BoW-based models with our WSJ data, which did not yield significant results.¹⁷ These findings support our argument that LLM-based indices offer a significant advantage over traditional models used in the literature.

We also conduct several robustness checks, estimating the VAR with various specifications.

¹⁷The results are available upon request.

Specifically, we vary the lag lengths and the orderings of the variables in the Cholesky decomposition. Additionally, we include other variables, such as the VIX and the Michigan Consumer Sentiment index, to control for potentially overlooked types of uncertainty. We also add for comparison the traditional BoW-based index as control. Figure A.3 in the Appendix demonstrates that our LLM-based EPU index remains robust across these alternative model specifications. Shocks to the index level are consistently associated with significant changes in economic activity, with effects persisting for up to three years. As an additional validation, we integrate other uncertainty indices, as depicted in Figure 4 and described in Section 3.3. Despite improving the in-sample fit, as measured by the mean-squared error, the estimated effect of economic policy uncertainty on economic activity remains consistent in magnitude. These results underscore the enhanced utility of our index in capturing and predicting economic dynamics.

Using the same VAR framework, we extend our analysis to investigate the impacts of various types of uncertainty on economic activity. Our findings indicate that economic policy uncertainty has the most pronounced effects on U.S. industrial production and employment. In contrast, geopolitical risk, as measured by our LLM-based approach, has no significant effect on economic activity. Monetary policy uncertainty and financial market uncertainty, however, exhibit effects that are approximately half the magnitude of that of economic policy uncertainty on industrial production and employment. Figures A.4, A.5 and A.6 in the Appendix display the impulse response functions for shocks to the other indices.

4.2 Flight-to-Safety

To assess how our LLM-based uncertainty indicators are related to investor behaviour, we examine the effect of changing levels of uncertainty on mutual fund flows. Da et al. (2015) argued that most investors in mutual funds are individual investors who are most likely sentiment traders. Moreover, Sammon and Shim (2024) showed that the opposite side of mutual fund flow is taken by mainly firms, short sellers and insiders of specific stocks. Thus, we use mutual fund flows to investigate whether our indices can capture the "flight-to-safety" effect of individual investors. "Flight-to-safety" means that in cases of high uncertainty, investors have a higher demand for assets that are perceived as safe. Therefore, they shift their investments from riskier assets, such as equities, to safer assets, such as government bonds. As a result, we hypothesise that higher uncertainty leads to outflows in equity mutual funds and inflows in mutual funds that invest in government bonds.

To investigate mutual fund flows, we select all available U.S. mutual funds from Morningstar that are listed in the categories "Equities" and "U.S. Intermediate Government". We then query monthly

data on returns and net assets for each fund from the CRSP Survivorship Bias Free Mutual Fund Database. We compute the mutual fund flows as follows:

$$flow_{i,t} = \frac{TNA_{i,t} - TNA_{i,t-1}}{TNA_{i,t-1}} - r_{i,t}, \quad (1)$$

where $TNA_{i,t}$ denotes fund i 's total net assets in month t and r denotes fund i 's return in month t . Following Coval and Stafford (2007), we exclude funds that have an extreme change in the total net assets over one month, i.e., we impose that $-0.5 < \frac{TNA_{i,t} - TNA_{i,t-1}}{TNA_{i,t-1}} < 2.0$. Table 4 displays the summary statistics of the mutual fund data. Our data cover many large funds that invest in either equities or intermediate government bonds.¹⁸ We see that over the whole sample period of 24 years, equity funds, on average, attracted monthly inflows of 0.86%, whereas government bonds on average attracted monthly outflows of 0.11%.

		Equities	Government Bonds
Monthly observations		17,771	9,685
Number of funds		112	39
Net assets per fund (in mio. USD)	Min	0.10	0.99
	Max	4,759.14	12,016.14
	Median	28.03	474.94
	Average	216.28	1,261.57
	Standard deviation	575.02	2,212.07
Return (%)	Average	0.69	0.26
	Standard deviation	5.36	1.14
Fund flow (%)	Average	0.86	-0.11
	Standard deviation	11.89	6.82

Table 4: This table displays summary statistics for funds selected from the CRSP Survivorship Bias Free Mutual Fund Database over the time period of 2000 to 2023. The monthly fund flows are calculated as described in Equation 1.

To assess the existence of a "flight-to-safety" phenomenon in mutual fund flows, we estimate the following regression:

$$flow_{i,t+k} = \beta_0 + \beta_1 LLM_{l,t} + \beta_2 VIX_t + \beta_3 MOVE_t + \gamma_j \sum_{j=1}^{12} r_{i,t-j} + \delta_j \sum_{j=1}^{12} flow_{i,t-j} + \epsilon_{i,t+k}, \quad (2)$$

¹⁸Equity funds predominantly invest in stocks. Intermediate government bond funds are defined as funds that "have at least 90% of their bond holdings in bonds backed by the U.S. government or by government-linked agencies. [...] These portfolios have durations between 3.5 and six years (or, if the duration is unavailable, average effective maturities between four and 10 years)". See https://awgmain.morningstar.com/webhelp/glossary_definitions/mutual_fund/glossary_mf_ce_Morningstar_Category.html for more information.

where $flow_{i,t+k}$ is the fund flow as defined in Equation 1 for mutual fund i in month $t+k$. $LLM_{l,t}$ is our LLM-based index for the type of uncertainty $l \in (EPU, GPR, MPU, FMU)$ in month t , VIX is the VIX at time t and $MOVE$ the MOVE at time t , which represent the implied volatility of option prices for equities and government bonds.¹⁹ We include twelve months of lagged mutual fund flows and fund returns, as Coval and Stafford (2007) showed that flows in and out of mutual funds are strongly influenced by past performance and past in- and outflows. They proposed including up to twelve months of lagged past returns and fund flows to forecast future fund flows. The regression results are displayed in Table 5.

The results show that our indices capture a flight-to-safety effect. Our measures of economic policy uncertainty and geopolitical risk predict persistent outflows from equity mutual funds for several months, whereas monetary policy uncertainty also predicts equity outflows, although with lower statistical significance. Notably, economic policy uncertainty predicts significant government bond inflows for up to two months, whereas geopolitical risk and monetary policy uncertainty do not. This is intuitive since geopolitical risk may lead to preferences for assets other than U.S. government bonds, and monetary policy uncertainty increases interest rate unpredictability, making bonds a less ideal hedge. Interestingly, our financial market uncertainty index predicts significant bond inflows for up to six months, as U.S. government bonds are seen as safe during financial market turmoil.

Our coefficients are economically significant. For example, a one standard deviation²⁰ increase in the LLM-based EPU index results in an immediate equity outflow of 0.38%, equating to almost 100 million USD per month, with outflows persisting for several months. Conversely, a one standard deviation increase in economic policy uncertainty leads to a 0.25% inflow into government bond funds, approximately 123 million USD.²¹ These findings support the hypothesis that during periods of heightened uncertainty, as measured by our LLM-based indices, investors shift from riskier equity funds to safer government bonds. While this effect is clear for economic policy uncertainty, it is less pronounced for other uncertainties. Geopolitical risk mainly causes equity outflows without significant bond inflows, and monetary policy uncertainty triggers some equity outflows but no significant bond inflows, likely owing to its direct impact on interest rates. Interestingly, higher financial market uncertainty increases demand for government bonds without causing significant equity fund outflows, suggesting that investors might be shifting from other asset classes to government bonds in such times. Future research should explore these dynamics further.

¹⁹As a robustness check, we also exclude the VIX and MOVE from our regressions, with results remaining consistent.

²⁰The standard deviations of our indices from 2000 until 2023 are: EPU 20.7, GPR 68.9, MPU 59.2, FMU: 22.3

²¹Our sample comprises 112 equity funds and 39 intermediate government bond funds, with average assets totalling USD 216 million and USD 1,262 million per fund, respectively. A monthly inflow of 0.38% and an outflow of 0.25% translate to approximately USD 100 million and USD 123 million per fund type, respectively.

k	Equities				Government Bonds				Equities				Government Bonds			
	β_1	SE	R^2	N	β_1	SE	R^2	N	β_1	SE	R^2	N	β_1	SE	R^2	N
Equities (LLM-based)																
0	-0.0184***	0.0044	0.11	17771	0.0123**	0.0061	0.10	9685	-0.0162***	0.0037	0.11	17771	0.0092*	0.0050	0.10	9685
1	-0.0162***	0.0043	0.10	17754	0.0108**	0.0047	0.06	9685	-0.0158***	0.0035	0.10	17754	0.0012	0.0037	0.06	9685
2	-0.0201***	0.0050	0.08	17737	0.0103**	0.0042	0.05	9685	-0.0156***	0.0039	0.08	17737	0.0033	0.0031	0.05	9685
3	-0.0265***	0.0053	0.07	17718	0.0086*	0.0049	0.04	9685	-0.0212***	0.0046	0.07	17718	0.0029	0.0036	0.04	9685
4	-0.0164***	0.0044	0.06	17611	0.0109**	0.0054	0.04	9648	-0.0178***	0.0047	0.06	17611	0.0001	0.0040	0.04	9648
5	-0.0190***	0.0045	0.06	17504	0.0031	0.0049	0.04	9611	-0.0223***	0.0043	0.06	17504	0.0026	0.0061	0.04	9611
6	-0.0188***	0.0046	0.05	17397	0.0081	0.0061	0.03	9575	-0.0248***	0.0053	0.05	17397	0.0008	0.0041	0.03	9575
GPR (LLM-based)																
0	-0.0035***	0.0010	0.11	17771	0.0012	0.0017	0.10	9685	0.0016*	0.0009	0.11	17771	0.0016	0.0020	0.10	9685
1	-0.0030***	0.0011	0.10	17754	0.0001	0.0012	0.06	9685	0.0024**	0.0010	0.10	17754	0.0003	0.0010	0.06	9685
2	-0.0026**	0.0011	0.08	17737	-0.0002	0.0011	0.05	9685	0.0025**	0.0011	0.08	17737	0.0001	0.0010	0.05	9685
3	-0.0028**	0.0011	0.07	17718	0.0009	0.0009	0.04	9685	0.0036***	0.0010	0.07	17718	0.0009	0.0015	0.04	9685
4	-0.0038***	0.0012	0.06	17611	0.0019	0.0015	0.04	9648	0.0027**	0.0012	0.06	17611	0.0014	0.0020	0.04	9648
5	-0.0050***	0.0014	0.06	17504	0.0003	0.0014	0.04	9611	0.0021	0.0013	0.06	17504	-0.0006	0.0015	0.04	9611
6	-0.0034***	0.0011	0.05	17397	0.0002	0.0009	0.03	9575	0.0032**	0.0015	0.05	17397	-0.0017**	0.0008	0.03	9575
MPU (LLM-based)																
0	-0.0015	0.0011	0.11	17771	0.0018	0.0012	0.10	9685	-0.0023**	0.0009	0.11	17771	0.0038**	0.0019	0.10	9685
1	-0.0014	0.0014	0.10	17754	0.0012	0.0012	0.06	9685	-0.0006	0.0012	0.10	17754	0.0025*	0.0013	0.06	9685
2	-0.0024*	0.0014	0.08	17737	0.0002	0.0015	0.05	9685	0.0007	0.0012	0.08	17737	0.0026*	0.0014	0.05	9685
3	-0.0043***	0.0016	0.07	17718	-0.0024**	0.0011	0.04	9685	-0.0007	0.0013	0.07	17718	0.0005	0.0014	0.04	9685
4	-0.0009	0.0014	0.06	17611	-0.0034**	0.0014	0.04	9648	0.0018	0.0011	0.06	17611	0.0008	0.0016	0.04	9648
5	-0.0034**	0.0014	0.06	17504	-0.0020	0.0013	0.04	9611	-0.0003	0.0014	0.06	17504	-0.0008	0.0017	0.04	9611
6	-0.0018	0.0015	0.05	17397	0.0009	0.0018	0.03	9575	-0.0013	0.0016	0.05	17397	-0.0001	0.0010	0.03	9575
FMU (LLM-based)																
0	0.0061	0.0045	0.11	17771	0.0184**	0.0073	0.10	9685	-0.0081*	0.0048	0.11	17771	0.0206***	0.0072	0.10	9685
1	0.0057	0.0041	0.10	17754	0.0250***	0.0066	0.06	9685	0.0081*	0.0043	0.10	17754	0.0095	0.0067	0.06	9685
2	0.0005	0.0052	0.08	17737	0.0248***	0.0080	0.05	9685	0.0012	0.0053	0.08	17737	0.0177***	0.0053	0.05	9685
3	-0.0020	0.0056	0.07	17718	0.0220***	0.0082	0.05	9685	-0.0037	0.0057	0.07	17718	0.0105*	0.0057	0.04	9685
4	0.0177***	0.0053	0.06	17611	0.0237***	0.0075	0.04	9648	0.0014	0.0043	0.06	17611	0.0111**	0.0046	0.04	9648
5	0.0044	0.0051	0.06	17504	0.0223***	0.0075	0.04	9611	-0.0073*	0.0042	0.06	17504	0.0176***	0.0054	0.04	9611
6	0.0123***	0.0047	0.05	17397	0.0255***	0.0073	0.04	9575	-0.0003	0.0036	0.05	17397	0.0198***	0.0060	0.03	9575

Table 5: The regression results of estimating Equation 2 using our own LLM-based uncertainty indices and using the uncertainty methods produced by the BoW methods. The BoW indices are those reported by Baker, Bloom, and Davis (2016) and Caldara and Iacoviello (2022) (see policyuncertainty.org). The regression is estimated on monthly data for different models. We include the VIX, the MOVE and twelve months of lagged returns and lagged fund flows as control variables in all specifications. We estimate one model for bond flows and for equity flows per index. We set $k = 0, 1, 2, 3, 4, 5, 6$ to study the persistence of the effect on future fund flows. The coefficient β_1 shows whether our uncertainty indicators affect mutual fund flows at time $t + k$. The standard errors are clustered at the mutual fund level. *, ** and *** denote that the coefficient estimates are significant at the 10%, 5%, and 1% significance levels, respectively.

Next, we compare our LLM-based indices with the traditional BoW methods to investigate whether the LLM-based indices excel at measuring uncertainty. To do so, we estimate the same regression as in Equation 2 but with different BoW-based indices to explain fund flows. First, we estimate the regression for mutual fund flows for the popular indices published by Baker et al. (2016) and Caldara and Iacoviello (2022). Next, we construct the BoW-based indices based on the same data that have been used for the LLM-based construction of the index. We then estimate the same regressions again based on these indices. In this way, we can compare the performance of the LLM-based indices with that of traditional BoW-based indices in different dimensions. We demonstrate that our indices excel at predicting mutual fund flows compared with the published indices that are based on multiple newspapers. We also show that the LLM-based indices outperform an index built on the basis of much fewer data using the BoW methods. The results based on the original indices are displayed in Table 5. The results for the BoW-based indices that use the same WSJ data as the LLM-based indices are reported in Table A.2 in the Appendix.

The results show that the BoW methods are less precise in identifying mutual fund flows than our LLM-based method is. If we compare the results for the EPU, MPU and EMV generated based on the same WSJ data used for our LLM-based indices (see Table A.2) with the results from our LLM-based indices (see Table 5), we find that the LLM-based indices are associated with stronger and more significant mutual fund flows. Using the same data from the WSJ with the LLM and BoW methods, we show that the BoW method can predict only some inflows to government bonds in times of high geopolitical risk. For all other measures and flows, our LLM-based method shows a more pronounced association with mutual fund flows. The results from the BoW-based indices only slightly support the hypothesis that in times of higher uncertainty, investors switch from riskier equity funds to safer bond funds. However, the LLM-based counterparts perform better in identifying the size and significance of the effects. If we run the same regressions for the indices that are reported by Baker et al. (2016) and Caldara and Iacoviello (2022) and are based on the original dataset used (10 newspapers), we also find no improvement (see Table 5). They show results similar to those of the BoW method based on the WSJ data, with one notable exception: GPR is positively associated with equity inflows, a result that defies intuition. In summary, the study suggests that the BoW-based indices, at best, perform comparably to our LLM-based indices in some categories. However, in most categories, the LLM-based indices have more power to predict mutual fund flows and identify a potential flight-to-safety effect; this is a remarkable result, considering that the LLM-based indices draw from a much smaller dataset than the traditional indices do.

To assess the significance of the differences between LLM- and BoW-based indices in predicting

mutual fund flows, we also combine both indices into one single regression equation. The results, presented in Table A.3 in the Appendix, further demonstrate the additional predictive power of the LLM-based index over traditional BoW-based methods. The results demonstrate that LLM-based indices offer significant added value in measuring uncertainty. These indices capture different types of uncertainty more accurately and directionally correct; specifically, when LLM-based indices are higher than BoW-based indices are, they effectively capture a notable shift from equity funds to safer bond funds. The impact of the LLM-based indices is often as substantial as the effects of the BoW-based indices; this underscores our earlier findings that LLM-based indices excel at measuring uncertainty linked to the flight-to-safety phenomenon.

4.3 Asset Returns

Uncertainty is not only associated with the previously discussed flight-to-safety effect but also can directly influence asset returns. For example, this can be the case for a change in asset demand through investors pulling out of riskier assets (such as the equity market) and investing in assets that are perceived as safe (e.g., government bonds). We study whether uncertainty, as measured by our LLM-based indices, can predict future asset returns. To do so, we estimate the following regression:

$$r_{i,t+k} = \beta_0 + \beta_1 LLM_{l,t} + \beta_2 VIX_t + \beta_3 MOVE_t + \gamma_m \sum_{j=1}^{12} r_{i,t-j} + \epsilon_{i,t+k}, \quad (3)$$

where $LLM_{l,t}$ is our LLM-based index for the type of uncertainty $l \in (EPU, GPR, MPU, FMU)$ in month t , VIX is the VIX at time t and $MOVE$ the MOVE at time t , which represent the implied volatility of option prices for equities and government bonds.²² $r_{i,t+k}$ is the monthly return of asset i at time t . We use twelve months of lagged returns and run the regression for different asset classes. We select the S&P 500, Eurostoxx 50 and the MSCI World indices to represent equity markets and the 2-year U.S. treasury bond to represent the government bond market. We further include the Swiss franc (CHF) and Japanese yen (JPY) as assets that are considered to be "safe" assets in the market (Ranaldo & Söderlind, 2010). We also include the price of gold in USD (XAU), as gold is often referred to as a safe asset (Baur & McDermott, 2010). Table 6 presents the regression results, including only those assets and indices that are statistically significant.

We find that our measures of uncertainty are associated with future asset returns. Economic policy uncertainty is correlated with future returns in the CHF for up to three months. An increase in economic policy uncertainty at time t leads to significantly increased returns in the CHF nominal

²²As a robustness check, we also exclude the VIX and MOVE from our regressions, with results remaining consistent.

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	$r_{i,t}$	$r_{i,t+1}$	$r_{i,t+2}$	$r_{i,t+3}$	$r_{i,t}$	$r_{i,t+1}$	$r_{i,t+2}$	$r_{i,t+3}$
CHFNEER				JPYNEER				
EPU	0.0091*** (0.0033)	0.0071** (0.0034)	0.0059* (0.0034)	0.0052* (0.0028)	-0.0032 (0.0068)	-0.0092 (0.0069)	0.0025 (0.0080)	0.0022 (0.0091)
R^2	0.09	0.07	0.06	0.05	0.18	0.12	0.08	0.07
N	282	282	282	282	282	282	282	282
MPU	0.0034*** (0.0011)	0.0025* (0.0014)	-0.0000 (0.0012)	0.0008 (0.0011)	0.0054** (0.0024)	0.0059* (0.0031)	0.0017 (0.0027)	0.0022 (0.0030)
R^2	0.10	0.07	0.05	0.05	0.20	0.13	0.08	0.07
N	282	282	282	282	282	282	282	282
UST2Y				XAU				
GPR	0.0380* (0.0199)	0.0244* (0.0130)	0.0052 (0.0182)	-0.0008 (0.0147)	-0.0034 (0.0039)	-0.0083** (0.0034)	-0.0056* (0.0031)	-0.0032 (0.0034)
R^2	0.16	0.15	0.13	0.07	0.07	0.07	0.08	0.06
N	282	282	282	282	282	282	282	282
SPX				EUROSTOXX50				
FMU	-0.0669*** (0.0171)	-0.0558** (0.0222)	-0.0405** (0.0198)	-0.0467** (0.0235)	-0.0574*** (0.0201)	-0.0409** (0.0195)	-0.0464*** (0.0178)	-0.0462* (0.0249)
R^2	0.35	0.12	0.10	0.08	0.26	0.05	0.06	0.04
N	288	288	288	288	288	288	288	288
MSCIWORLD								
FMU	-0.0711*** (0.0173)	-0.0622*** (0.0230)	-0.0438* (0.0232)	-0.0477* (0.0252)				
R^2	0.36	0.11	0.09	0.06				
N	288	288	288	288				

Table 6: Results of the regressions described by Equation 3. We estimate the effects of past returns and our LLM-based indices on contemporaneous and future returns. All the coefficients are displayed as percentages. The control variables include the VIX, the MOVE and twelve months of lagged returns. The table displays the coefficient β_1 for the LLM-based index. The standard errors are shown in brackets below. We display the results for different types of uncertainty on four different types of assets. CHFNEER and JPYNEER stand for the Swiss franc and Japanese yen nominal effective exchange rates, respectively. UST2Y stands for the 2-year U.S. treasury bond yield, XAU for the gold price in USD, SPX for the S&P 500, EUROSTOXX50 for the Eurostoxx 50 and MSCIWORLD for the MSCI World equity indices. $r_{i,t}$ is the contemporaneous return in the same month as the uncertainty index is computed. $r_{i,t+k}$ represents the monthly value in month $t+k$, where $k=0,1,2,3$. We use Newey and West (1987) heteroscedasticity- and autocorrelation-consistent standard errors with twelve lags. *, ** and *** denote that the coefficient estimates are significant at the 10%, 5%, and 1% significance levels, respectively.

exchange rate. An increase in geopolitical risk increases the 2-year U.S. treasury bond yields for up to three months, and higher monetary policy uncertainty is associated with a higher nominal effective exchange rate of the JPY. The effects for the currencies are consistent with the explanation that increased uncertainty leads to a higher demand for assets that are perceived as safe; thus, these assets exhibit higher returns. However, for the U.S. treasury yields and gold prices, we expect that higher geopolitical risk leads to an increased demand for U.S. treasuries and gold and, thus, to lower yields and higher gold prices. Geopolitical risk might affect asset prices differently than the other types of uncertainty that we explore. One potential explanation could also be that geopolitical risk differs depending on its originating region. Increased geopolitical risk in a major economy might have stronger effects on investor sentiment. This could also explain why we do not see increased inflows into government bond funds (see Table 5). The effect of geopolitical risk on asset prices is an interesting topic for future research. For the equity market as measured by the S&P 500, the Eurostoxx 50 and the MSCI World index, we find that increased financial market uncertainty leads to persistently lower equity returns. The effect is strongly significant for up to three months. This is to be expected given that greater uncertainty related to financial markets leads to a decline in equity prices, similar to the known effect that the VIX has on equity prices (Whaley, 2009). All effects are economically meaningful. A one standard deviation change in our LLM-based indices²³ leads to a significant change in returns of 0.15% for the CHF, 1.68% for the 2-year U.S. treasury bond yield, 0.57% for gold, 0.35% for the JPY, 1.24% for the S&P 500, 0.91% for the Eurostoxx 50 and 1.39% for the MSCI World in the following month; this is in contrast to average monthly returns of 0.19% for the CHF, 1.55% for the 2-year U.S. treasury bond yield, 0.85% for gold, -0.05% for the JPY, 0.6% for the S&P 500, 0.32% for the Eurostoxx 50 and 0.46% for the MSCI World within our sample.

We also run the same regressions for the traditional BoW indices calculated on the basis of the same WSJ data used in our LLM-based indices and for the indices reported by Baker et al. (2016) and Caldara and Iacoviello (2022). Tables A.4 and A.5 in the Appendix display the results. Similar to the mutual fund flows, we find that the LLM-based indices perform much better than the BoW-based methods do. If the BoW-based methods can even identify significant effects, they are less strong or in the opposite direction, as intuition would suggest. These findings additionally support our hypothesis that the LLM-based methods provide an additional benefit in measuring uncertainty.

²³Recall that the standard deviations of our indices over the full sample are as follows: EPU 20.7, GPR 68.9, MPU 59.2, and FMU 22.3

4.4 Discussion of the Results

In conducting empirical analyses using pretrained LLMs, there is a potential risk of look-ahead bias. This bias occurs when an LLM’s training data include information about future events, potentially contaminating an analysis intended to rely solely on contemporaneous or historical data. Given that the training periods of models such as Zephyr-7B- β and GPT-4 Turbo overlap with our sample timeframe used in the empirical analysis in the previous section, it is important to assess whether their classifications are influenced by memorisation of training data or data leakage.

One approach to address this issue is to conduct an empirical analysis using only out-of-sample data. Several studies have demonstrated that empirical analyses using LLMs hold both in-sample and out-of-sample, mitigating concerns regarding look-ahead bias. For example, Jha, Qian, Weber, and Yang (2024) demonstrated that an LLM-based investment score continues to predict capital expenditures beyond the model’s training period, confirming the model’s predictive ability without data leakage. The analysis also includes tests where firm, people and product identities are masked, and the results remain robust despite reduced readability, further reducing the possibility of look-ahead bias. Similarly, Chen, Kelly, and Xiu (2022) and Lopez-Lira and Tang (2023) presented evidence of asset price predictability by LLMs in an out-of-sample setting. Furthermore, Bybee (2023) shows that LLM-generated economic expectations remain significantly correlated with existing survey measures, even when tested with data outside the model’s training window. This result reinforces the assumption that the LLM was generalising from its training data rather than memorising it, confirming its ability to perform out-of-sample. However, an out-of-sample empirical analysis presents a challenge, as the limited number of available out-of-sample data points reduces the reliability of the results. This limitation is particularly significant in our case because we are working with monthly data, and Zephyr-7B- β is built on the Mistral-7B model, which was only released in September 2023. As a result, the small amount of monthly out-of-sample data is insufficient for drawing robust conclusions. Furthermore, relying on older models is not a viable alternative, as they tend to exhibit significantly reduced performance compared with more recent models.

Another approach to mitigating potential look-ahead bias is to concentrate on periods where such a bias is more likely to occur. One illustrative example is the global financial crisis. Articles containing the term "subprime" may have increasingly been labelled as signalling rising uncertainty even before the onset of the crisis. To evaluate this, we compare the EPU labels generated by Zephyr-7B- β with those produced by human coders. The human coders, the authors, are aware that the articles are written prior to the global financial crisis, providing a benchmark to assess whether the model

exhibits look-ahead bias. Our sample includes all 29 articles published before the global financial crisis that contain the word "subprime." Zephyr-7B- β labelled 15 of these articles as indicating increased EPU, and 11 of these are similarly labelled by the human coders. Therefore, in more than 73% of the cases, the model's labels for increasing EPU align with the human coders' assessment. This result is consistent with the findings presented in Section 3.2. We repeat this procedure by randomly selecting 29 articles that were published not before but during the global financial crisis. We again find that 11 out of 15 positively labelled articles match the evaluation of the human coders. Consequently, this example, along with evidence from past studies, suggests that look-ahead bias is unlikely to significantly affect our results. Zephyr-7B- β does not disproportionately label articles with specific terms differently in the lead-up to significant events, such as the global financial crisis; this provides anecdotal evidence that the model can label articles accurately even before events associated with high uncertainty and highlights the robustness and reliability of the LLM-based predictions.

Look-ahead bias could particularly impact our equity price regressions, given that stock return predictability remains a contested topic in the academic literature. We acknowledge that previous studies typically report lower predictive power for stock returns at monthly frequencies (see, for instance, Welch and Goyal, 2008 and Campbell and Thompson, 2008). However, based on earlier findings indicating minimal look-ahead bias in LLM-driven price predictions (Lopez-Lira & Tang, 2023), along with our illustrative examples, we argue that look-ahead bias is not the primary driver of our results. With respect to the theoretical question of why heightened uncertainty might negatively predict future stock returns, we recognise the viewpoint that increased uncertainty could increase the risk premium, lowering prices today but signalling higher future returns. However, this view may overlook the potential persistence of uncertainty and its deeper impact on market expectations and corporate behaviour. When uncertainty is not just a short-term spike but reflects deeper structural concerns, such as political instability, economic policy changes, or shifts in global trade dynamics, it can depress future earnings potential and discourage long-term investment. As both Baker et al. (2016) and the present study show, heightened uncertainty negatively impacts economic activity. We follow the argument of Baker et al. (2016) that increased uncertainty can prompt businesses to delay or cancel investments while also increasing the cost of debt and equity financing, further deterring investment. Additionally, households may adopt more cautious spending habits in response, reducing overall consumption. Consequently, higher uncertainty today can signal lower stock prices in the future, as reflected in our results.

Despite the previously outlined challenges, we argue that our results offer valuable insights and contribute meaningfully to the ongoing academic discourse. We have grounded our analysis in a

robust body of literature and provided extensive references to support our methodology and findings. Additionally, we have incorporated illustrative evidence where possible to help rule out potential concerns of bias.

5 Conclusion

This paper introduces an innovative method for quantifying uncertainty through the use of LLMs, enhancing the precision and contextual sensitivity of uncertainty measurement compared with the methods used in traditional indices such as the EPU index by Baker et al. (2016), which is based on a BoW method. By leveraging the capabilities of LLMs, our study addresses the limitations of traditional BoW methods, which often overlook contextual nuances in textual data. Our approach enables a more nuanced understanding of different types of uncertainty, including economic policy uncertainty, geopolitical risk, monetary policy uncertainty, and financial market volatility.

Our methodology involves fine-tuning open-source LLMs, specifically LLaMa-2-7B-Chat and Zephyr-7B- β , using a dataset of WSJ articles classified by GPT-4 Turbo. Compared with traditional methods, GPT-4 Turbo demonstrates better alignment with the human judgment in classifying uncertainty, as shown in our human evaluations. We further show that LLMs can be fine-tuned at low costs, making this approach practical and scalable. Fine-tuned models such as Zephyr-7B- β , which are evaluated against GPT-4 Turbo in terms of accuracy and F1 scores, achieve high performance across all uncertainty categories, highlighting the adaptability and efficiency of LLMs for specific tasks with limited data. Using this fine-tuned open-source LLM, we create various LLM-based uncertainty indices.

The LLM-based indices reveal significant insights into the relationships between uncertainty and macroeconomic indicators as well as financial market dynamics. Using a VAR model, the study investigates the impact of uncertainty shocks on macroeconomic variables such as industrial production and employment. The impulse response functions indicate more pronounced effects than traditional indices do, with LLM-based indices demonstrating clearer and more sustained impacts. Moreover, the findings of several estimated regressions suggest a "flight-to-safety" effect, where increased uncertainty leads to outflows from equity mutual funds and inflows into government bond funds. Additionally, our indices are predictors of future asset returns, with higher uncertainty associated with higher returns for safe-haven assets such as the CHF and JPY and lower returns for the S&P 500.

The use of these indices in macroeconomic and financial market analysis highlights the practical utility of LLMs in managing economic risk and uncertainty. The high performance of the Zephyr-7B- β

model highlights the potential for LLMs to revolutionise the analysis and quantification of uncertainty. Thus, this paper not only validates the effectiveness of quantifying uncertainty but also sets the stage for their application among financial market professionals and policymakers, enhancing decision-making processes and enriching economic analyses. The open-source nature of our models promotes transparency and replicability, allowing further research to build on our work and explore new applications of LLMs in economics and finance.

References

- Ahir, H., Bloom, N., & Furceri, D. (2022). The World Uncertainty Index. *National Bureau of Economic Research*.
- Baker, S. R., Bloom, N., Davis, S., & Renault, T. (2021). Twitter-derived measures of economic uncertainty. *Working paper*.
- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4), 1593–1636.
- Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M., & Viratyosin, T. (2020). The unprecedented stock market reaction to COVID-19. *Review of Asset Pricing Studies*, 10(4), 742–758.
- Baker, S. R., Bloom, N., Davis, S. J., & Kost, K. J. (2019). Policy news and stock market volatility. *National Bureau of Economic Research*.
- Baker, S. R., Bloom, N., Davis, S. J., & Sammon, M. C. (2021). What triggers stock market jumps? *National Bureau of Economic Research*.
- Balcilar, M., Bonato, M., Demirer, R., & Gupta, R. (2017). The effect of investor sentiment on gold market return dynamics: Evidence from a nonparametric causality-in-quantiles approach. *Resources Policy*, 51, 77–84.
- Baur, D. G., & McDermott, T. K. (2010). Is gold a safe haven? International evidence. *Journal of Banking & Finance*, 34(8), 1886–1898.
- Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Bybee, L. (2023). Surveying generative AI’s economic expectations. *arXiv preprint arXiv:2305.02823*.
- Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2023). Business news and business cycles. *Journal of Finance*, forthcoming.
- Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, 112(4), 1194–1225.
- Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4), 1509–1531.
- Chen, Y., Kelly, B. T., & Xiu, D. (2022). Expected returns and large language models. *Available at SSRN 4416687*.
- Coval, J., & Stafford, E. (2007). Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics*, 86(2), 479–512.

- Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies*, 28(1), 1–32.
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.
- Glasserman, P., & Mamaysky, H. (2019). Does unusual news forecast market stress? *Journal of Financial and Quantitative Analysis*, 54(5), 1937–1974.
- Glasserman, P., Mamaysky, H., & Qin, J. (2023). New news is bad news. *arXiv preprint arXiv:2309.05560*.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Jha, M., Qian, J., Weber, M., & Yang, B. (2024). *Chatgpt and corporate policies* (tech. rep.). National Bureau of Economic Research.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Jiang, J., Kelly, B. T., & Xiu, D. (2023). Expected returns and large language models. *Available at SSRN*. <https://ssrn.com/abstract=4416687>
- Julio, B., & Yook, Y. (2012). Political uncertainty and corporate investment cycles. *Journal of Finance*, 67(1), 45–83.
- Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3), 1177–1216.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.
- Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.
- Mamaysky, H. (2023). News and markets in the time of covid-19. *Available at SSRN*. <https://ssrn.com/abstract=3565597>
- Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1), 137–162.
- Nasr, A. B., Bonato, M., Demirer, R., & Gupta, R. (2019). Investor Sentiment and Crash Risk in Safe Havens. *Journal of Economics and Behavioral Studies*, 10(6A(J)), 97–108.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703.

- OpenAI. (2023a). ChatGPT [Large language model]. <https://chat.openai.com/chat>
- OpenAI. (2023b). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.
- Ranaldo, A., & Söderlind, P. (2010). Safe haven currencies. *Review of finance*, 14(3), 385–407.
- Sammon, M., & Shim, J. J. (2024). Who clears the market when passive investors trade? *Available at SSRN*. <https://ssrn.com/abstract=4777585>
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. (2023). Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.
- van Binsbergen, J. H., Bryzgalova, S., Mukhopadhyay, M., & Sharma, V. (2024). (almost) 200 years of news-based economic sentiment. *National Bureau of Economic Research*.
- Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4), 1455–1508.
- Whaley, R. E. (2009). Understanding the vix. *Journal of Portfolio Management*, 35(3), 98–105.
- Zhang, B., Yang, H., & Liu, X.-Y. (2023). Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. (2023). Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Additional Figures and Tables

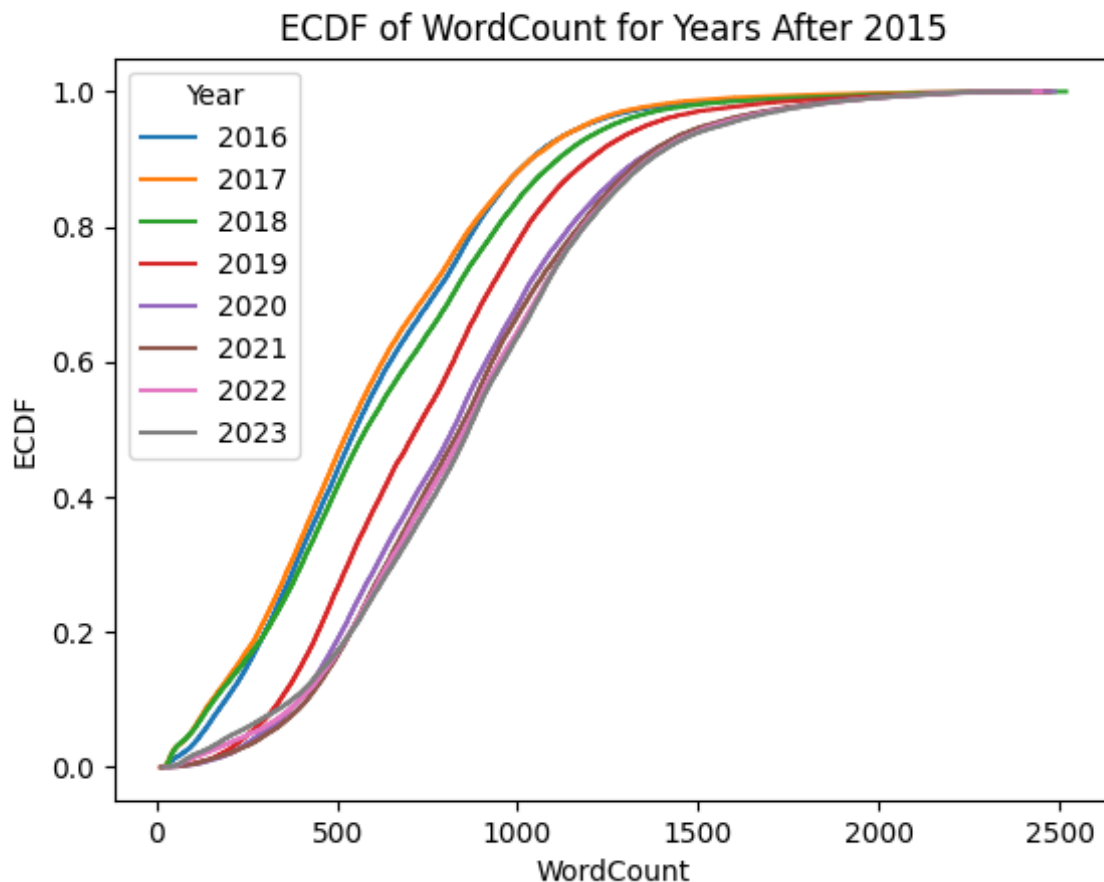


Figure A.1: This figure shows the cumulative distribution functions based on the word count of all the articles since 2015. A trend toward consistently longer articles emerges after 2019.

		BoW-based EPU		
		True	False	N
GPT-based EPU	True	198 (1.0%)	7,271 (37.8%)	7,496 (38.8%)
	False	55 (0.3%)	11,731 (60.9%)	11,786 (61.2%)
	N	253 (1.3%)	19,002 (98.7%)	19,255 (100%)

Table A.1: This table shows the classification of our subset of articles in 1998, 1999, 2008 and 2009 by the BoW method proposed by Baker, Bloom, and Davis (2016) (BoW-based EPU) and by GPT-4 Turbo (GPT-based EPU). True means that the article affects economic policy uncertainty as defined by the respective method. For the BoW-based EPU this means that an article must contain at least one word in each of the groups "economic", "policy" and "uncertainty". For the GPT-based EPU the LLM must identify the article as having an "increasing" or "decreasing" effect on economic policy uncertainty.

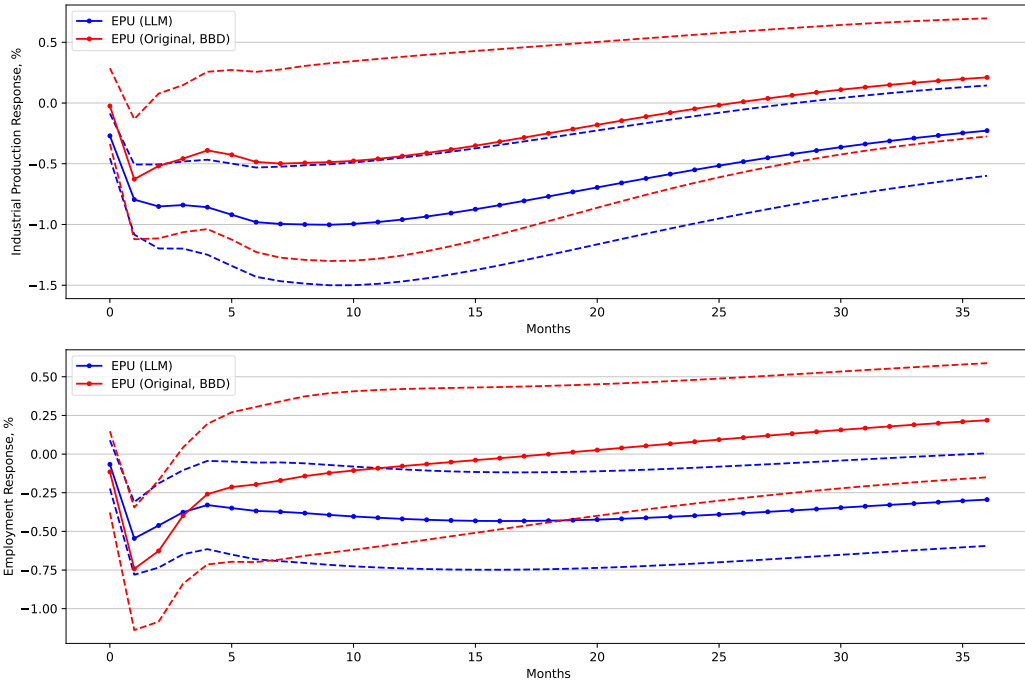


Figure A.2: This figure displays the same impulse response functions as described in Figure 5 (in red). Additionally, it displays (in blue) the impulse responses for a model using the EPU index for 2000 until 2023 published online by Baker, Bloom, and Davis (2016).

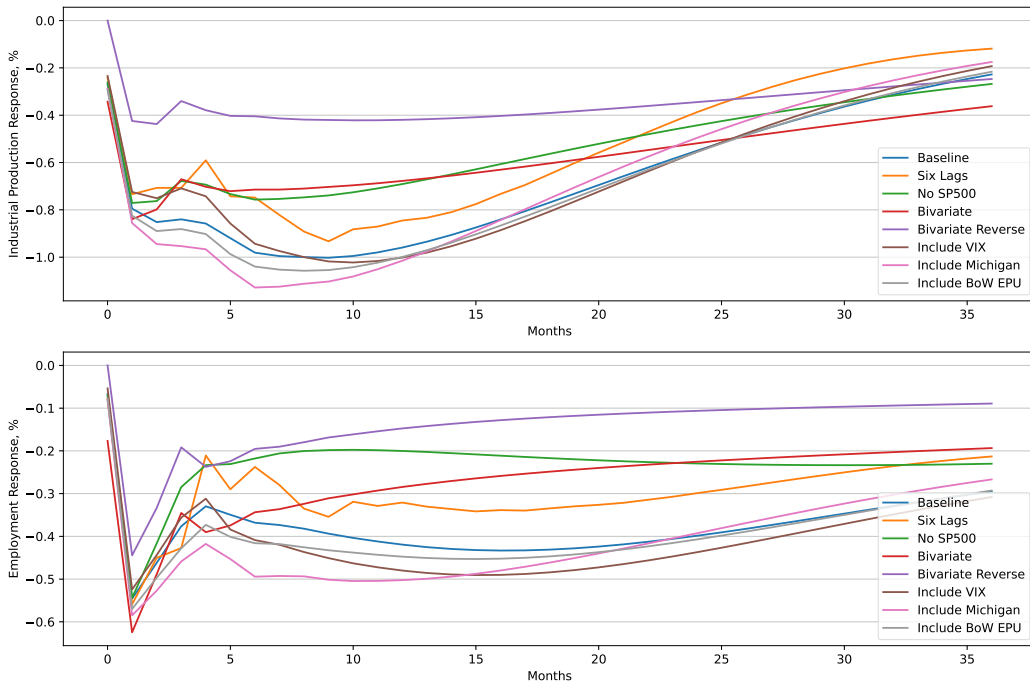


Figure A.3: This figure displays the impulse response functions for various model specifications. The baseline model is shown in Figure 5. The additional models include variations with an increased lag length, a bivariate model containing only our index and the specified economic variable, a bivariate model with reversed ordering, and models with different combinations of included variables compared with the baseline.

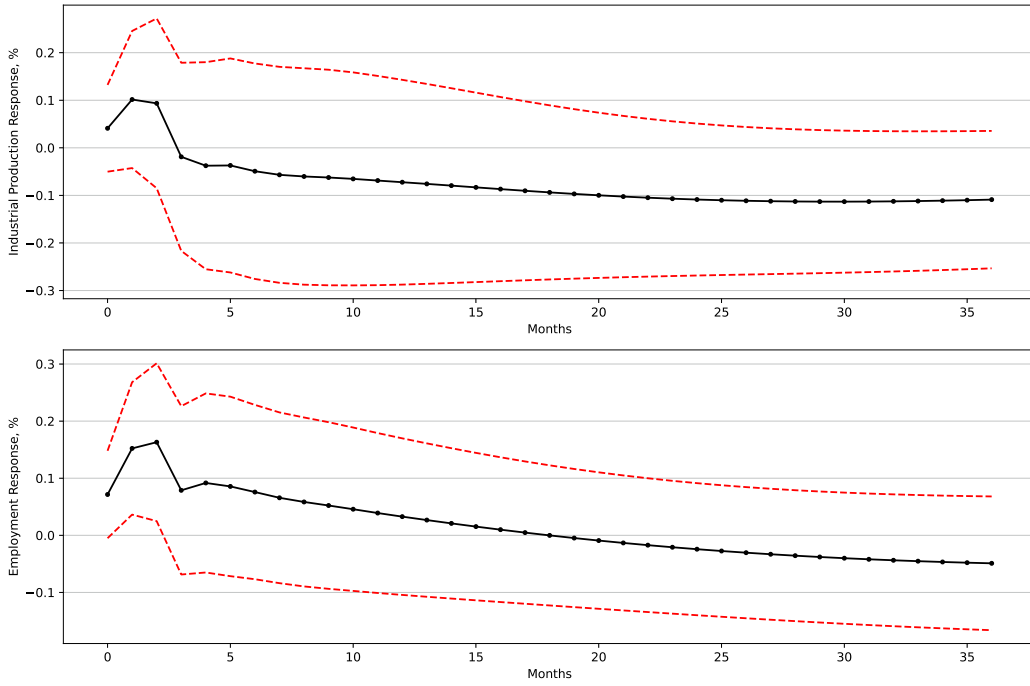


Figure A.4: This figure displays the same impulse response functions as described in Figure 5 including a 90% confidence bound for a model that includes our LLM-based geopolitical risk index.

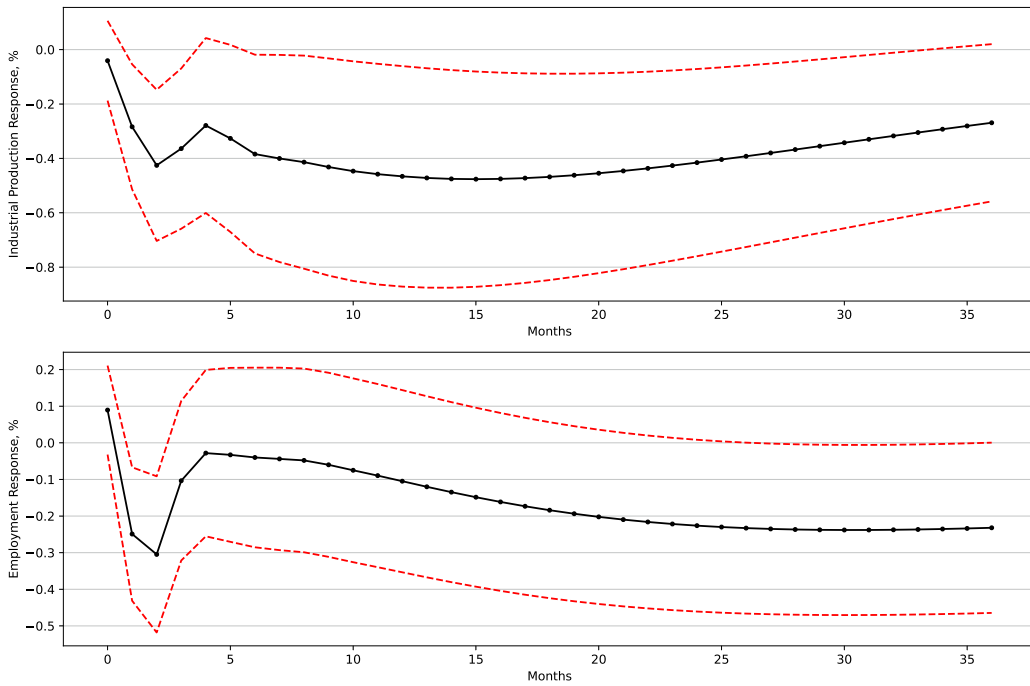


Figure A.5: This figure displays the same impulse response functions as described in Figure 5 including a 90% confidence bound for a model that includes our LLM-based monetary policy uncertainty index.

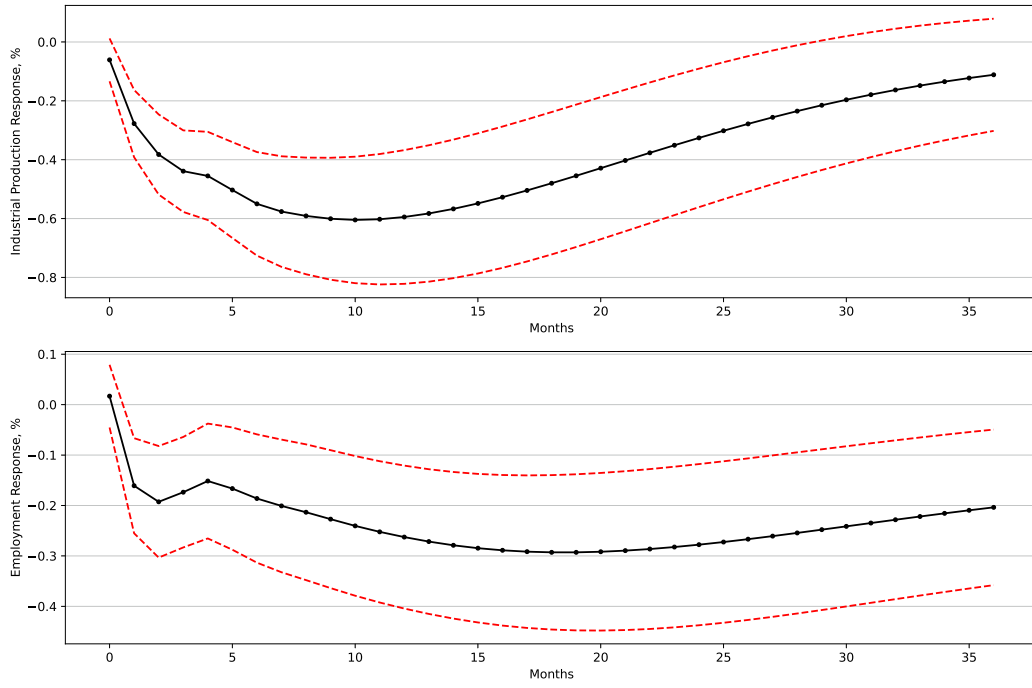


Figure A.6: This figure displays the same impulse response functions as described in Figure 5 including a 90% confidence bound for a model that includes our LLM-based financial market uncertainty index.

Table A.2

EPU		Equities			Government Bonds			
k	β_1	Std. Error	R^2	N	β_1	Std. Error	R^2	N
0	-0.0008	0.0035	0.11	17771	0.0033	0.0042	0.10	9685
1	-0.0066*	0.0036	0.10	17754	0.0061**	0.0031	0.06	9685
2	-0.0024	0.0036	0.08	17737	-0.0014	0.0022	0.05	9685
3	-0.0095**	0.0037	0.07	17718	0.0028	0.0029	0.04	9685
4	0.0004	0.0024	0.06	17611	0.0048	0.0034	0.04	9648
5	-0.0063	0.0038	0.06	17504	0.0105**	0.0047	0.04	9611
6	-0.0010	0.0040	0.05	17397	0.0010	0.0040	0.03	9575
GPR		Equities			Government Bonds			
k	β_1	Std. Error	R^2	N	β_1	Std. Error	R^2	N
0	-0.0013*	0.0007	0.11	17771	0.0009	0.0011	0.10	9685
1	0.0007	0.0008	0.10	17754	0.0009	0.0009	0.06	9685
2	-0.0017**	0.0008	0.08	17737	0.0023**	0.0011	0.05	9685
3	0.0010	0.0010	0.07	17718	0.0029***	0.0010	0.04	9685
4	-0.0003	0.0008	0.06	17611	0.0014	0.0012	0.04	9648
5	-0.0023***	0.0008	0.06	17504	0.0013	0.0011	0.04	9611
6	-0.0017*	0.0009	0.05	17397	0.0023**	0.0010	0.03	9575
MPU		Equities			Government Bonds			
k	β_1	Std. Error	R^2	N	β_1	Std. Error	R^2	N
0	0.0006	0.0012	0.11	17771	0.0015	0.0015	0.10	9685
1	-0.0013	0.0013	0.10	17754	0.0009	0.0012	0.06	9685
2	0.0003	0.0012	0.08	17737	-0.0019*	0.0010	0.05	9685
3	-0.0012	0.0013	0.07	17718	0.0001	0.0009	0.04	9685
4	0.0022**	0.0011	0.06	17611	-0.0007	0.0009	0.04	9648
5	-0.0001	0.0011	0.06	17504	0.0022	0.0015	0.04	9611
6	0.0006	0.0015	0.05	17397	-0.0003	0.0011	0.03	9575

(continues on next page)

Table A.2 (*continued*)

FMU k	Equities				Government Bonds			
	β_1	Std. Error	R^2	N	β_1	Std. Error	R^2	N
0	0.0063*	0.0037	0.11	17771	0.0114***	0.0026	0.10	9685
1	0.0086**	0.0036	0.10	17754	0.0085*	0.0045	0.06	9685
2	0.0129***	0.0046	0.08	17737	0.0028	0.0041	0.05	9685
3	0.0037	0.0041	0.07	17718	-0.0007	0.0042	0.04	9685
4	0.0126**	0.0050	0.06	17611	0.0026	0.0032	0.04	9648
5	0.0062	0.0039	0.06	17504	0.0066*	0.0038	0.04	9611
6	0.0078**	0.0039	0.05	17397	0.0142***	0.0037	0.03	9575

Table A.2: The regression results of estimating Equation 2 via the uncertainty methods produced by the BoW methods. The BoW indices have been constructed based on the methods in Baker et al. (2016) and Caldara and Iacoviello (2022) **using the same WSJ data that were used for constructing the LLM-based indices. Furthermore, the indices have been standardised to have the same mean and standard deviation as their LLM-based counterparts depicted in Table 5.** The regression is estimated on monthly data for different models. We include the VIX, the MOVE and twelve months of lagged returns and lagged fund flows as control variables in all specifications. We estimate one model for bond flows and for equity flows per index. We set $k = 0, 1, 2, 3, 4, 5, 6$ to study the persistence of the effect on future fund flows. The coefficient β_1 shows whether our uncertainty indicators affect mutual fund flows at time $t + k$. The standard errors are clustered at the mutual fund level. *, ** and *** denote that the coefficient estimates are significant at the 10%, 5%, and 1% significance levels, respectively.

$k = 0$	Equities				Government Bonds			
	(1) EPU	(2) GPR	(3) MPU	(4) FMU	(5) EPU	(6) GPR	(7) MPU	(8) FMU
BoW	-0.0166*** (0.0050)	-0.0034*** (0.0010)	-0.0008 (0.0012)	0.0130*** (0.0049)	0.0122* (0.0067)	0.0016 (0.0018)	0.0027 (0.0018)	0.0220*** (0.0065)
ΔLLM	-0.0203*** (0.0045)	-0.0037*** (0.0012)	-0.0012 (0.0012)	0.0067 (0.0044)	0.0123* (0.0065)	0.0010 (0.0019)	0.0014 (0.0012)	0.0133** (0.0066)
R^2	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10
N	17771	17771	17771	17771	9685	9685	9685	9685
$k = 1$								
BoW	-0.0189*** (0.0052)	-0.0016 (0.0012)	-0.0022 (0.0017)	0.0148*** (0.0051)	0.0127** (0.0053)	0.0007 (0.0013)	0.0017 (0.0012)	0.0251*** (0.0088)
ΔLLM	-0.0156*** (0.0041)	-0.0042*** (0.0011)	-0.0006 (0.0013)	0.0061 (0.0040)	0.0094** (0.0045)	-0.0003 (0.0010)	0.0010 (0.0014)	0.0200*** (0.0063)
R^2	0.10	0.10	0.10	0.10	0.06	0.06	0.06	0.06
N	17754	17754	17754	17754	9685	9685	9685	9685
$k = 2$								
BoW	-0.0191*** (0.0057)	-0.0030** (0.0012)	-0.0018 (0.0016)	0.0137** (0.0054)	0.0051 (0.0038)	0.0012 (0.0010)	-0.0015 (0.0015)	0.0251*** (0.0095)
ΔLLM	-0.0216*** (0.0050)	-0.0023* (0.0013)	-0.0020 (0.0014)	0.0002 (0.0051)	0.0112** (0.0049)	-0.0018 (0.0014)	0.0010 (0.0015)	0.0244*** (0.0077)
R^2	0.08	0.08	0.08	0.08	0.05	0.05	0.05	0.05
N	17737	17737	17737	17737	9685	9685	9685	9685
$k = 3$								
BoW	-0.0278*** (0.0059)	-0.0013 (0.0011)	-0.0044** (0.0019)	0.0025 (0.0061)	0.0051 (0.0039)	0.0023* (0.0012)	-0.0023* (0.0012)	0.0228*** (0.0082)
ΔLLM	-0.0254*** (0.0051)	-0.0042*** (0.0014)	-0.0040*** (0.0015)	-0.0016 (0.0056)	0.0069 (0.0054)	-0.0011 (0.0008)	-0.0019 (0.0012)	0.0246*** (0.0087)
R^2	0.07	0.07	0.07	0.07	0.04	0.04	0.04	0.05
N	17718	17718	17718	17718	9685	9685	9685	9685

Table A.3: Results of the regressions described by $flow_{i,t+k} = \beta_0 + \beta_1 BoW_{i,t} + \beta_2 \Delta LLM_{i,t} + \beta_3 VIX_t + \gamma_j \sum_{j=1}^{12} r_{i,t-j} + \delta_j \sum_{j=1}^{12} flow_{i,t-j} + \epsilon_{i,t+k}$, where $\Delta LLM = LLM_{i,t} - BoW_{i,t}$. The traditional BoW indices are calculated on the same WSJ data as those used for the LLM indices. We estimate the regression for four different types of indices and for mutual fund flows into a bond and an equity fund. All indices have been standardised to have the same mean and standard deviation. The coefficient β_1 is represented by the row BoW, and the coefficient β_2 is represented by the ΔLLM row. The standard errors are included in brackets below the coefficients and are clustered at the mutual fund level. *, ** and *** denote that the coefficient estimates are significant at the 10%, 5%, and 1% significance levels, respectively. We compare our financial market uncertainty (FMU) index with the equity market volatility (EMV) index from Baker, Bloom, Davis, and Kost (2019).

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	$r_{i,t}$	$r_{i,t+1}$	$r_{i,t+2}$	$r_{i,t+3}$	$r_{i,t}$	$r_{i,t+1}$	$r_{i,t+2}$	$r_{i,t+3}$
CHFNEER				JPYNEER				
EPU	0.0058*	0.0063**	0.0010	-0.0001	0.0094*	0.0004	0.0011	-0.0007
	(0.0034)	(0.0027)	(0.0031)	(0.0038)	(0.0055)	(0.0043)	(0.0051)	(0.0061)
R^2	0.08	0.07	0.05	0.05	0.19	0.11	0.08	0.07
N	282	282	282	282	282	282	282	282
MPU	0.0022	0.0007	-0.0006	-0.0011	0.0030	-0.0006	-0.0005	-0.0019
	(0.0014)	(0.0011)	(0.0012)	(0.0013)	(0.0022)	(0.0018)	(0.0021)	(0.0025)
R^2	0.08	0.06	0.05	0.05	0.19	0.11	0.08	0.07
N	282	282	282	282	282	282	282	282
UST2Y				XAU				
GPR	0.0386	-0.0220	-0.0137	0.0126	-0.0017	0.0002	0.0018	-0.0064*
	(0.0248)	(0.0177)	(0.0200)	(0.0198)	(0.0038)	(0.0036)	(0.0042)	(0.0038)
R^2	0.16	0.15	0.14	0.08	0.07	0.05	0.07	0.06
N	282	282	282	282	282	282	282	282
SPX				EUROSTOXX50				
EMV	0.0019	-0.0102	0.0064	-0.0148	0.0101	0.0087	0.0105	0.0001
	(0.0129)	(0.0121)	(0.0150)	(0.0125)	(0.0145)	(0.0145)	(0.0195)	(0.0154)
R^2	0.31	0.09	0.09	0.06	0.24	0.04	0.04	0.02
N	288	288	288	288	288	288	288	288
MSCIWORLD								
EMV	0.0043	-0.0055	0.0132	-0.0102				
	(0.0148)	(0.0133)	(0.0143)	(0.0128)				
R^2	0.32	0.07	0.08	0.04				
N	288	288	288	288				

Table A.4: Results of the regressions described by Equation 3 for the same specifications as displayed in Table 6. We estimate the effect of past returns and **the traditional BoW indices calculated on the same WSJ data as those used for the LLM indices** on contemporaneous and future returns. All the coefficients are displayed as percentages. The control variables include the VIX, the MOVE and twelve months of lagged returns. The table displays the coefficient β_1 for the LLM-based index. Standard errors are shown in brackets below. We display the results for different types of uncertainty on four different types of assets. CHFNEER and JPYNEER stand for the Swiss franc and Japanese yen nominal effective exchange rates. UST2Y stands for the 2-year U.S. treasury bond yield, XAU for the gold price in USD, SPX for the S&P 500, EUROSTOXX50 for the Eurostoxx 50 and MSCIWORLD for the MSCI World equity indices. $r_{i,t}$ is the contemporaneous return in the same month as the uncertainty index is computed. $r_{i,t+k}$ represents the monthly value in month $t+k$, where $k = 0, 1, 2, 3$. We use Newey and West (1987) heteroscedasticity- and autocorrelation-consistent standard errors with twelve lags. *, ** and *** denote that the coefficient estimates are significant at the 10%, 5%, and 1% significance levels, respectively.

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	$r_{i,t}$	$r_{i,t+1}$	$r_{i,t+2}$	$r_{i,t+3}$	$r_{i,t}$	$r_{i,t+1}$	$r_{i,t+2}$	$r_{i,t+3}$
CHFNEER				JPYNEER				
EPU	0.0075** (0.0030)	0.0029 (0.0034)	0.0009 (0.0032)	-0.0003 (0.0037)	-0.0002 (0.0068)	-0.0077 (0.0055)	-0.0029 (0.0063)	-0.0110* (0.0065)
R^2	0.09	0.06	0.05	0.05	0.18	0.11	0.08	0.08
N	282	282	282	282	282	282	282	282
MPU	0.0040*** (0.0015)	-0.0006 (0.0011)	-0.0013 (0.0010)	-0.0026* (0.0014)	0.0013 (0.0019)	-0.0075*** (0.0018)	-0.0044** (0.0019)	-0.0085*** (0.0019)
R^2	0.10	0.06	0.05	0.06	0.18	0.14	0.09	0.11
N	282	282	282	282	282	282	282	282
UST2Y				XAU				
GPR	0.0097 (0.0140)	0.0155** (0.0076)	0.0069 (0.0167)	0.0046 (0.0129)	-0.0006 (0.0029)	-0.0054** (0.0023)	-0.0016 (0.0024)	0.0017 (0.0025)
R^2	0.14	0.14	0.13	0.08	0.07	0.06	0.07	0.06
N	282	282	282	282	282	282	282	282
SPX				EUROSTOXX50				
EMV	-0.0093 (0.0175)	-0.0364 (0.0246)	0.0006 (0.0206)	-0.0151 (0.0159)	-0.0315 (0.0211)	-0.0308 (0.0241)	-0.0142 (0.0218)	0.0092 (0.0209)
R^2	0.31	0.10	0.08	0.06	0.25	0.05	0.04	0.02
N	288	288	288	288	288	288	288	288
MSCIWORLD								
EMV	-0.0228 (0.0206)	-0.0384 (0.0273)	0.0069 (0.0195)	-0.0137 (0.0149)				
R^2	0.32	0.09	0.07	0.04				
N	288	288	288	288				

Table A.5: Results of the regressions described by Equation 3 for the same specifications as displayed in Table 6. We estimate the effect of past returns and **the traditional BoW indices provided by Baker, Bloom, and Davis (2016) and Caldara and Iacoviello (2022)** on contemporaneous and future returns. For comparison, the indices are standardised to have the same mean and standard deviation as our LLM-based indices. All the coefficients are displayed as percentages. The control variables include the VIX, the MOVE and twelve months of lagged returns. The table displays the coefficient β_1 for the LLM-based index. The standard errors are shown in brackets below. We display the results for different types of uncertainty on four different types of assets. CHFNEER and JPYNEER stand for the Swiss franc and Japanese yen nominal effective exchange rates. UST2Y stands for the 2-year U.S. treasury bond yield, XAU for the gold price in USD, SPX for the S&P 500, EUROSTOXX50 for the Eurostoxx 50 and MSCIWORLD for the MSCI World equity indices. $r_{i,t}$ is the contemporaneous return in the same month as the uncertainty index is computed. $r_{i,t+k}$ represents the monthly value in month $t+k$, where $k=0,1,2,3$. We use Newey and West (1987) heteroscedasticity- and autocorrelation-consistent standard errors with twelve lags. *, ** and *** denote that the coefficient estimates are significant at the 10%, 5%, and 1% significance levels, respectively.

B LLM Selection and Fine-Tuning

This section describes the technical implementation of our LLM-based index. It includes details on the LLM selection, construction of the fine-tuning dataset and the parameter-efficient fine-tuning with QLoRa.

B.1 LLM Selection

LLaMa-2-7B-Chat: As an offshoot of Meta’s ambitious LLaMa-2 project²⁴, the LLaMa-2-7B-Chat model is based on the standard transformer architecture and offers a unique combination of sophisticated language modelling in a compact package, swiftly gaining popularity and industry recognition. Its swift rise to prominence is a testament to its advanced capabilities and the industry’s acknowledgement of its potential. Despite its relatively small size of 7 billion parameters, this model incorporates advanced features of its larger counterparts. This balance between size and complexity is pivotal, especially given that our model is easily reproducible and, therefore, does not want to access unnecessarily large resources; this aligns with our objective of classifying different types of uncertainty in newspaper articles, a task that requires both linguistic precision and contextual awareness.

Zephyr-7B- β : Developed by HuggingFace, Zephyr-7B- β is notable for its exemplary performance in open-source benchmarks, as highlighted in Figure B.1. Zephyr-7B- β is a fine-tuned version of Mistral 7B (Jiang, Sablayrolles, et al., 2023). The Mistral 7B model stands out because it uses Sliding Window Attention (SWA), a technique from Beltagy, Peters, and Cohan (2020). SWA makes the model more efficient in terms of memory and computing. The selection of Zephyr-7B- β is motivated by its proven efficiency in language processing and generation, despite its small number of parameters. Zephyr-7B- β not only matches but also occasionally surpasses the performance of larger models such as the 70 billion parameter version of the previously mentioned LLaMa-2 Model. The model shows that using larger teacher models to create synthetic datasets is very effective for fine-tuning. Zephyr was fine-tuned via the Ultrachat dataset from ChatGPT, which was further aligned with the GPT-4-generated UltraFeedback through direct preference optimization (DPO). This process allows Zephyr to learn from the teacher models, indirectly capturing the human preferences they were trained on through RLHF (Tunstall et al., 2023).²⁵

GPT-4 Turbo: GPT-4 Turbo was developed by OpenAI and is widely considered one of the best-performing models on the market.²⁶ Notably, the model and its parameters are not openly

²⁴See <https://llama.meta.com/llama2> for more details on the project.

²⁵See <https://huggingface.co/collections/HuggingFaceH4/zephyr-7b-6538c6d6d5ddd1cbb1744a66> to access the datasets used to fine-tune Zephyr-7B- β .

²⁶Refer to Zhang et al., 2023 or Tunstall et al., 2023 for a comparative performance analysis of different LLMs. For

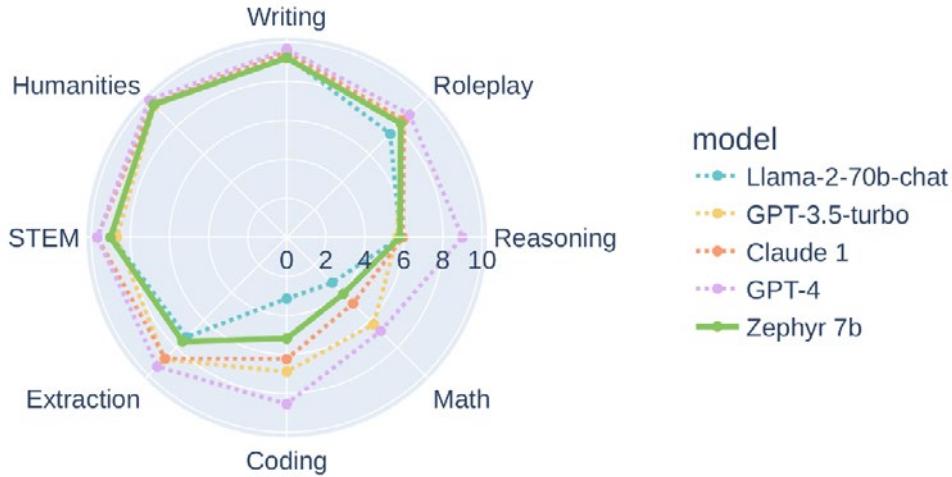


Figure B.1: This figure displays the performance of different LLMs, including Zephyr-7B- β , on MT-Bench, a benchmark designed to measure the ability of LLMs to engage in various types of conversations. Source: Tunstall et al., 2023.

accessible to users; only paying customers can engage with it via chat or API. Consequently, it would be infeasible and too costly to create the whole uncertainty indices via GPT-4 Turbo. Despite this limitation, we use GPT-4 Turbo as a benchmark for the open-source LLMs and for the creation of the fine-tuning dataset, as it would be too costly to create the fine-tuning dataset through human labelling. Thus, we classify a random set of articles through the LLM’s API. These articles are then used for fine-tuning and evaluating the open-source LLMs.

The selection of LLaMa-2-7B-Chat and Zephyr-7B- β is grounded in a strategic approach to evaluating high-performing, yet computationally feasible, models against one of the industry’s best. This comparison aims to shed light on the evolving landscape of LLMs, offering insights into the trade-offs between model size, computational demands, and performance efficiency.

B.2 Construction of the Fine-Tuning Dataset

The usage of GPT-4 Turbo comes at a marginal cost, but Zhou et al., 2023 and Zhang et al., 2023 have shown that fine-tuning instances of open-source LLMs with datasets that include between 1,000 and 10,000 entries can lead to high performance. Zhou et al., 2023 even highlighted that LLMs can become highly efficient at their tasks after fine-tuning with even a relatively small dataset of 1,000 entries. In our case, we select 10% of our newspaper articles from 1998 until 1999 and 2008 until 2009 for fine-tuning. The articles are selected at random every month to avoid adding certain biases to the data. With this procedure, we obtain approximately 9,000 and 10,000 articles per sample period. We then query the GPT-4 Turbo API with this article and a suitable prompt. At prices

additional details on GPT-4 Turbo, see <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>.

available when accessing the API,²⁷ OpenAI charges approximately 300 USD for those queries.²⁸ Furthermore, we set the so-called temperature of the model to the lowest value of zero to ensure consistency in the results. A lower temperature allows for fewer different responses (OpenAI, 2023b) and thus increases the consistency of the responses in our classification dataset. A consistent dataset is important because we do not want to introduce additional randomness over time. For example, we queried different articles with the same prompt ten times each and obtained almost identical results. The only difference was a slight variation in the level of confidence (within 0.1) and minor changes in the wording of the explanation.

For the classification, we construct a prompt aimed at instructing the different LLMs on the different definitions of uncertainty. The prompt instructs the model to assess whether an article implies a rise, fall or unchanged level of uncertainty or whether the article influences uncertainty at all. Additionally, we ask for the magnitude of uncertainty, which is quantified on a scale from 0 to 1, and the model's confidence level in its response, which is also measured on a scale from 0 to 1. Moreover, we prompt the model to categorise the news, identify whether the event has global or regional implications, and specify which asset is predominantly affected by the reported news. This comprehensive set of inquiries aims to extract nuanced and detailed information regarding the impact of news articles on uncertainty across various dimensions.

Figure B.2 presents the prompt inquiring about the influence of a newspaper article on different types of uncertainty. We use the same definitions for the EPU, GPR, MPU and EMV in our LLM-based indices as in the original papers by Baker et al. (2016), Caldara and Iacoviello (2022) and Baker et al. (2019). The only difference is that we define the EMV not only for equity markets but also for financial markets. Thus, we call it financial market uncertainty (FMU).²⁹

Table B.1 shows the responses from classifying randomly selected articles with GPT-4 Turbo using the prompt outlined before. The responses among the two sample periods are in the same area of magnitude. The model indicates a substantial rise in economic policy uncertainty and financial market uncertainty. The confidence in these predictions is relatively high, emphasising the model's ability to detect heightened uncertainty. On the other hand, instances of decreasing uncertainty are notably low, with average magnitudes and moderate confidence levels. This makes intuitive sense, as newspapers typically report on new events that lead to uncertainty. If these events then dissipate again or the uncertainty decreases, this is often not a newsworthy story. Many articles are classified as

²⁷See <https://openai.com/pricing#language-models>.

²⁸The total cost of generating the GPT-4 data was slightly less than 600 USD.

²⁹We did run all the results in this paper for both the EMV and FMU definitions and did not find significantly different results. The results are available at request by the authors.

You are evaluating if newspaper articles affect different types of uncertainty.

Economic policy uncertainty (EPU) is defined as:

Uncertainty over who makes or will make policy decisions that have economic consequences. Current and past uncertainty over what economic policy actions will be undertaken. Uncertainty over the economic effects of policy actions in the past, present or future. Economic uncertainty induced by policy inaction or related to policy developments or motivated by non-economic considerations - e.g., national security concerns.

Geopolitical risk (GPR) is defined as:

The threat, realization, and escalation of adverse events associated with wars, terrorism, and any tensions among states and political actors that affect the peaceful course of international relations.

Monetary policy uncertainty (MPU) is defined as:

Policy uncertainty as defined in the EPU but specifically associated with monetary policy.

Financial market uncertainty (FMU) is defined as:

Uncertainty that specifically concerns financial markets, is characterized by unpredictable fluctuations and volatility are primarily impacting financial markets.

Here is a newspaper article:

Title: "{row.Title}"

Text: "{row.Text}"

What is the association between this news and EPU?

What is the association between this news and GPR?

What is the association between this news and MPU?

What is the association between this news and FMU?

Under which news category would you categorize this article?

What asset will be most affected by this news?

What region is the origin of the potential uncertainty in the news article?

You must give 23 answers. Write your answer as:

"EPU": "increasing/decreasing/unchanged/not affected",

"EPU_Confidence": "float (0-1)",

"EPU_Magnitude": "float (0-1)",

"EPU_Explanation": "text (less than 25 words)",

"GPR": "increasing/decreasing/unchanged/not affected",

"GPR_Confidence": "float (0-1)",

"GPR_Magnitude": "float (0-1)",

"GPR_Explanation": "text (less than 25 words)",

"MPU": "increasing/decreasing/unchanged/not affected",

"MPU_Confidence": "float (0-1)",

"MPU_Magnitude": "float (0-1)",

"MPU_Explanation": "text (less than 25 words)",

"FMU": "increasing/decreasing/unchanged/not affected",

"FMU_Confidence": "float (0-1)",

"FMU_Magnitude": "float (0-1)",

"FMU_Explanation": "text (less than 25 words)",

"news_category": "Financial Markets, Economy, Politics or Other",

"most_affected_asset": "S&P500, EuroStoxx50, Gold, 2yearUST, 10yearUST, JPY, CHF or Other",

"origin_region": "Global, US, Europe, Asia or Other"

Figure B.2: The complete prompt inquiring about the influence of a newspaper article on different types of uncertainty.

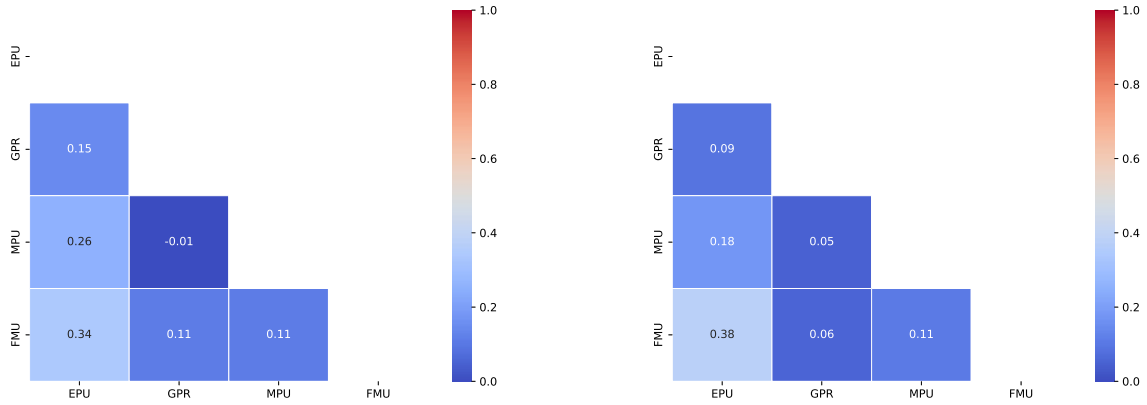
affecting uncertainty, yet the model categorises them with unchanged levels of uncertainty. However, the magnitude and confidence of these answers are relatively low, indicating that the model might have some difficulties in classifying these articles. Finally, a substantial percentage of all the articles are classified as not affecting uncertainty with low magnitude and high confidence. The model seems to be very confident about which articles do not affect specific types of uncertainty.

		1998/1999				2008/2009			
		EPU	MPU	GPR	FMU	EPU	MPU	GPR	FMU
increasing	Percentage	32.94	5.69	3.69	33.47	41.54	5.45	4.51	33.01
	Avg. Magnitude	0.61	0.64	0.68	0.56	0.61	0.63	0.66	0.57
	Std. Magnitude	0.06	0.10	0.12	0.13	0.06	0.10	0.11	0.12
	Avg. Confidence	0.74	0.78	0.79	0.74	0.74	0.77	0.80	0.73
	Std. Confidence	0.06	0.08	0.11	0.11	0.05	0.09	0.09	0.10
decreasing	Percentage	1.83	1.48	0.22	4.93	0.88	0.53	0.35	1.65
	Avg. Magnitude	0.62	0.53	0.59	0.47	0.62	0.52	0.55	0.53
	Std. Magnitude	0.04	0.17	0.15	0.18	0.06	0.17	0.15	0.14
	Avg. Confidence	0.74	0.80	0.81	0.78	0.74	0.79	0.78	0.77
	Std. Confidence	0.05	0.08	0.06	0.09	0.05	0.06	0.09	0.09
unchanged	Percentage	30.38	7.68	0.60	38.14	30.16	6.37	1.20	33.84
	Avg. Magnitude	0.22	0.27	0.12	0.26	0.22	0.29	0.23	0.26
	Std. Magnitude	0.07	0.16	0.16	0.14	0.09	0.18	0.19	0.14
	Avg. Confidence	0.70	0.67	0.56	0.65	0.69	0.64	0.59	0.62
	Std. Confidence	0.06	0.11	0.11	0.10	0.07	0.11	0.11	0.09
not affected	Percentage	34.84	85.15	95.49	23.46	27.38	87.62	93.91	31.45
	Avg. Magnitude	0.05	0.01	0.01	0.01	0.03	0.01	0.01	0.01
	Std. Magnitude	0.05	0.04	0.03	0.04	0.05	0.03	0.02	0.03
	Avg. Confidence	0.94	0.98	0.99	0.97	0.95	0.99	0.99	0.97
	Std. Confidence	0.05	0.04	0.03	0.07	0.08	0.05	0.04	0.08
Total		9,164	9,164	9,164	9,164	10,095	10,095	10,095	10,095

Table B.1: Results from classifying randomly selected articles with GPT-4 Turbo. For the two subsamples that we classified with GPT-4 Turbo (using randomly selected data from 1998/1999 and 2008/2009), we show how often the model classified an article as "increasing", "decreasing", "unchanged" or "not affected" per uncertainty type. We also report the averages and standard deviations of the magnitude and confidence of the effects as estimated by GPT-4 Turbo.

The EPU and FMU measures seem to be affected equally often in similar directions, whereas the MPU and GPR uncertainties are relatively rarely affected. However, the question arises of how the different types of uncertainty are correlated and whether GPT-4 Turbo can effectively distinguish the different types. Figure B.3 shows the correlations of the classifications of the model in both samples. The correlations are low, which indicates that the different indices capture different types of uncertainty. Additionally, there are no noticeable significant autocorrelations observed in the classification of the different types of uncertainty.³⁰

³⁰The significance of the autocorrelations has been tested via the Ljung-Box test on various lags. The results are available from the authors.



(a) Correlations for 1998/1999

(b) Correlations for 2008/2009

Figure B.3: The correlations between the different types of uncertainty as classified by GPT-4 Turbo. The model classifies a newspaper article per category as "increasing", "decreasing", "unchanged" or "not affected". The correlation among these values is then calculated and depicted in these charts. A value close to 1 indicates that the model considers that the newspaper articles affect uncertainty in almost the same way.

B.3 Parameter-efficient Fine-Tuning with QLoRA

In the evolving landscape of artificial intelligence, the parameter-efficient fine-tuning (PEFT) technique has revolutionised the customisation of large-scale LLMs for specific-domain tasks. PEFT serves as a sophisticated means of model refinement that judiciously tunes a selected set of parameters, thereby maximising efficiency and effectiveness while minimising the need for extensive computational resources. Techniques such as low-rank adaptation (LoRA) and its quantised counterpart, QLoRA, as detailed in Hu et al., 2021 and Dettmers et al., 2023, exemplify the success of PEFT by demonstrating how strategic, minimal adjustments can lead to significant improvements in model performance.

PEFT shines in scenarios where models need to be fine-tuned for a specific objective with comparably small datasets. In such cases, LoRA under the PEFT umbrella leverages a model adapter to infuse flexibility and adaptability into the LLM, paving the way for precise text response generation. Quantisation techniques refine this process further, compressing model data for more efficient GPU usage.

LoRA, a flagship technique within PEFT, minimises fine-tuning demands by inserting trainable low-rank matrices into the transformer architecture's layers. This innovative step conserves the foundational weights of the model, slashing the number of parameters that require adjustment for task-specific learning. This method not only conserves computational resources but also retains the model's high performance.

Instruction-based prompt engineering complements these technical innovations, serving as guides

for input processing to fine-tune the model’s response generation. As GPT-4 Turbo (OpenAI, 2023b) is currently considered one of the most advanced LLMs available, it serves as our benchmark and is used to generate the instruction tuning dataset by classifying newspaper articles from the WSJ. We use GPT-4 Turbo for our instruction tuning dataset because of its high performance (see Section 3.2) and the cost-effectiveness of using an LLM compared with human classification. Human classification of a total of 19,000 articles over the two periods demands substantial time and resources, making it practically unfeasible.

We start the training of the LLaMa-2-7B-Chat and Zephyr-7B- β models with similar training parameters as outlined in Zhang et al., 2023. However, we reduce the training batch size and evaluation batch size to 1 while setting the gradient accumulation steps and evaluation accumulation steps to 32, leading to a global training and evaluation batch size of 32 to reduce the GPU memory requirements. We set the maximum token length to 4,096, which is the maximum token limit given in both models. Furthermore, we refrain from using DeepSpeed (a deep learning optimisation library), as we use a single GPU for training and use the previously described QLoRa method instead. We use 4-bit quantisation to train the models on a single GPU. Furthermore, we use the bf16 computation datatype, NF4 data type and double quantisation, as Dettmers et al., 2023 have shown that this approach is most effective and does not degrade performance. Like in Dettmers et al., 2023, we apply the paged AdamW optimiser to avoid memory spikes during gradient checkpointing. Moreover, we set the LoRa r parameter to 8, as a higher r increases the computational resources needed, and as Hu et al., 2021 revealed that LoRa already shows high performance with a low r . Finally, we set the LoRa α to 16, as it is common to set it to twice the value of the LoRa r , and we set the dropout rate to 0.1. Furthermore, we train the model on all linear layers of the base model (Dettmers et al., 2023) rather than focusing only on the attention modules. The training parameters are shown in Table B.2.

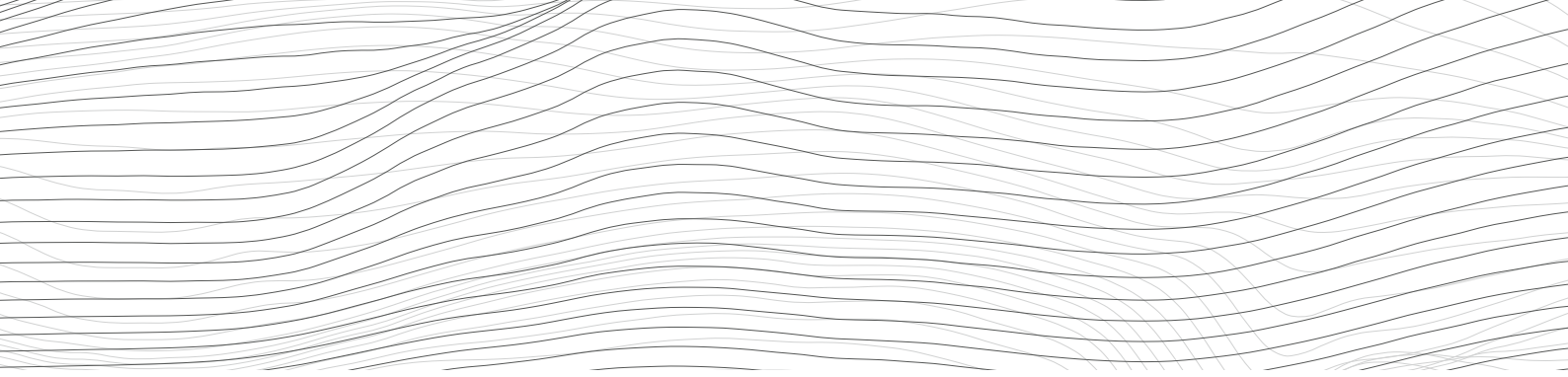
For training, we use Hugging Face’s SFTTrainer, which is a light wrapper around the transformer trainer, allowing the training of LLMs on a custom dataset in a simple way. The GPT-4 classified WSJ dataset, which contains approximately 9,000 to 10,000 randomly selected articles per period, is split into training, validation and test datasets. The instruction tuning is performed on 80% of the GPT-4 classified WSJ dataset, while we also perform an evaluation during training using 10% of the dataset and include an early stopping parameter to avoid overfitting on the data used for training. The training is stopped when the evaluation loss worsens for 10 evaluation calls, whereby the evaluation is performed every 10 steps. The remaining 10% of the dataset is used for model tests, which are presented in Section 3.2.

Parameter	Value
Epochs	3
Training batch size	1
Gradient accumulation steps	32
Evaluation batch size	1
Evaluation accumulation steps	32
Optimizer	Paged AdamW 32-bit
Learning rate	1e-5
LR scheduler	CosineAnnealing
Warmup ratio	0
Weight decay	0.1
Lora r	8
Lora α	16
Lora dropout	0.1
Max token length	4096
GPU	A100 (80GB)

Table B.2: Training parameters used to train the open-source LLMs.

Recent SNB Working Papers

- 2024-12 Francesco Audrino, Jessica Gentner, Simon Stalder:
Quantifying uncertainty: a new era of measurement through large language models
- 2024-11 Marc-Antoine Ramelet, Anna Zeitz:
Oil price shocks and household heterogeneity: the income side
- 2024-10 Jayson Danton, Terhi Jokipii:
A decade of low interest rates: impact on Swiss bank profitability
- 2024-09 Anders Brownworth, Jon Durfee, Michael Junho Lee, Antoine Martin:
Regulating decentralized systems: evidence from sanctions on Tornado Cash
- 2024-08 Valentin Grob, Gabriel Züllig:
Corporate leverage and the effects of monetary policy on investment: a reconciliation of micro and macro elasticities
- 2024-07 Thomas Nitschka:
Evidence on the international financial spillovers of the New York Bankers' Panic of 1907
- 2024-06 Milen Arro-Cannarsa, Rolf Scheufele:
Nowcasting GDP: what are the gains from machine learning algorithms?
- 2024-05 Jessica Gentner:
The role of hedge funds in the Swiss franc foreign exchange market
- 2024-04 Tobias Cwik, Christoph Winter:
FX interventions as a form of unconventional monetary policy
- 2024-03 Lukas Voellmy:
Decomposing liquidity risk in banking models
- 2024-02 Elizabeth Steiner:
The impact of exchange rate fluctuations on markups – firm-level evidence for Switzerland
- 2024-01 Matthias Burgert, Johannes Eugster, Victoria Otten:
The interest rate sensitivity of house prices: international evidence on its state dependence
- 2023-08 Martin Brown, Laura Felber, Christoph Meyer:
Consumer adoption and use of financial technology: “tap and go” payments
- 2023-07 Marie-Catherine Bieri:
Assessing economic sentiment with newspaper text indices: evidence from Switzerland
- 2023-06 Martin Brown, Yves Nacht, Thomas Nellen, Helmut Stix:
Cashless payments and consumer spending
- 2023-05 Romain Baeriswyl, Alex Oktay, Marc-Antoine Ramelet:
Exchange rate shocks and equity prices: the role of currency denomination
- 2023-04 Jonas M. Bruhin, Rolf Scheufele, Yannic Stucki:
The economic impact of Russia's invasion of Ukraine on European countries – a SVAR approach
- 2023-03 Dirk Niepelt:
Payments and prices
- 2023-02 Andreas M. Fischer, Pinar Yeşin:
The kindness of strangers: Brexit and bilateral financial linkages
- 2023-01 Laura Felber, Simon Beyeler:
Nowcasting economic activity using transaction payments data
- 2022-14 Johannes Eugster, Giovanni Donato:
The exchange rate elasticity of the Swiss current account
- 2022-13 Richard Schmidt, Pinar Yeşin:
The growing importance of investment funds in capital flows



SCHWEIZERISCHE NATIONALBANK
BANQUE NATIONALE SUISSE
BANCA NAZIONALE SVIZZERA
BANCA NAZIUNALA SVIZRA
SWISS NATIONAL BANK

