Automation of Text-Based Economic Indicator Construction: A Pilot Exploration on Economic Policy Uncertainty Index

ABSTRACT

The growing popularity of text-as-data in various domain-specific applications and research has often relied on manually selected keywords or annotations. Although labor-intensive, expensive and time-consuming, the effectiveness of these efforts is not always guaranteed, especially in the early stages of research. This predicament raises the question of the extent to which large language models (LLMs) can aid in verifying the potential of a nascent research idea. This paper seeks to explore the reliability of LLM-suggested keywords in the automatic construction of the Economic Policy Uncertainty (EPU) index, a representative keyword-based economic indicator utilized across numerous countries. Our findings affirm the utility of LLMs in facilitating the automation of text-based EPU index construction. Furthermore, we delve into the potential of LLMs in enhancing the indicator construction process.

CCS CONCEPTS

- **Applied computing → Economics**.

KEYWORDS

Automation, Economic indicator construction, Large language model

1 INTRODUCTION

The use of text as data has gained popularity in various fields, including finance [6, 13] and economics [2, 17], where researchers predominantly employ two approaches: relying on manually-selected keywords or utilizing human-annotated data. Both techniques demand considerable human labor; for instance, Baker et al. [5] reviewed 12,000 news articles to identify those related to economic policy uncertainty, while Chen et al. [7] annotated over 47,000 news articles to train models. This manual effort is undertaken without assurance of the viability of the research concept or an ability to estimate success beforehand. However, the advent of Large Language Models (LLMs) presents an opportunity to assist in the evaluation and validation of early-stage research ideas. This paper simulates the process of constructing text-based economic indicators and aims to ascertain the extent to which LLMs can aid in assessing ideas related to this process.

The Economic Policy Uncertainty (EPU) index, a prominent text-based economic indicator developed through manually selected keyword sets and quantification of relevant news article frequencies, serves as a case study [5]. This methodology has been adopted by over 29 countries, each employing unique keyword sets. However, if economists want to study global economic policy uncertainty and considering there are over 200 countries, it's almost impossible to manually construct keywords for each country, and that's where LLMs can assist and a reliable automation procedure becomes vital. This study contemplates the feasibility of automating the EPU index construction versus the necessity of extensive manual annotation as conducted by Baker et al. [5]. Furthermore, we evaluate the potential of LLM-suggested keywords in EPU index development and the discrepancies between these and indexes based on expert-curated keyword sets.

Moreover, building on Chen et al. [7]'s observation that even expertly selected keywords yield 40% irrelevant content, this research explores the self-refinement capability of LLMs in the context of EPU index construction. Specifically, it examines LLMs' efficiency in excluding news articles that, despite containing EPU keywords, are deemed irrelevant to economic policy uncertainty. The findings suggest that employing a limited number of annotated instances (few-shot learning) outperforms both zero-shot learning and fine-tuning approaches.

In summary, this work presents three main contributions:

1. We highlight the potential of LLMs in constructing the EPU index and discuss the trade-offs among various methods.
2. We show that for identifying and removing unrelated news articles, using a tailor-made supervised model remains superior to the LLM-based approach in terms of accuracy, explainability, and predictability.
3. We demonstrate that an EPU index, constructed using LLM-suggested keywords, can achieve significant explainability and predictability.

2 RELATED WORK

Thanks to advances in NLP, LLMs are leading researchers towards reducing the significant effort required in data annotation. Exploring the potential of LLMs in zero-shot or few-shot contexts has emerged as a crucial topic. The most prevalent method to directly employ LLMs in research involves treating them as classifiers or predictive models. Jha et al. [12] utilizes GPT-3.5 to analyze corporate earnings conference calls, predicting changes in capital spending policy for the subsequent year. Smales [16] employs both GPT-3.5 and GPT-4 [14] as classifiers to categorize sentiments and attitudes—positive, negative, hawkish, dovish—regarding the monetary policy decisions made by the Reserve Bank of Australia. Yang [20] leverages Ada-002 OpenAI text embeddings to extract qualitative features from patent application documents to predict the acceptance rate and value of patents. Moreover, the potential of LLMs to assist researchers is an area of promising direction. This paper offers an early exploration of the automation of text-based economic indicator construction. We focus on the EPU index and provide an

CIKM'24, October 21–25, 2024, Boise, Idaho, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-XXXX-XXXXX.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Results of keyword recommendation with different task description.

Model	Task Description	Precision	Recall	F1
GPT-3.5	Simple	10.43%	30.75%	11.77%
	Definition	11.81%	38.67%	13.72%
GPT-4	Simple	9.01%	28.81%	10.31%
	Definition	10.38%	29.55%	11.44%
Claude 3 Sonnet	Simple	9.71%	30.01%	11.09%
	Definition	12.14%	41.42%	14.25%

analysis of each step in the index's construction. Although our findings reveal the potential of LLMs in constructing the EPU index, we also highlight the gap between supervised methods and the value of human effort in domain-specific index construction.

3 PRELIMINARY

The EPU Index was initially proposed by Baker et al. [5]. Constructing the EPU index involves several steps. Initially, researchers categorize three keyword sets relevant to economics, policy, and uncertainty, necessitating substantial human effort. For instance, Baker et al. [5] reviewed 12,000 news articles with considerable student assistance and frequent discussions to refine these keyword sets. The keyword set for *economics* includes "economic" and "economy"; for *uncertainty*, it comprises "uncertain" and "uncertainty"; and for *policy*, it encompasses "Congress", "deficit", "Federal Reserve", "legislation", "regulation", and "White House". After deciding on the keyword sets, economists count the number of articles containing at least one keyword in each set to construct the EPU index. Specifically, to construct the EPU index, prior research tallies the monthly count of news articles (N_{EPU}) that include at least one term from each predefined set. Consequently, the EPU index (EPU) is derived using Equation 1.

$$EPU = \frac{N_{\text{EPU}}}{N_{\text{All}}} \tag{1}$$

where N_{All} denotes the aggregate number of news articles disseminated within a given month. Following the construction of the EPU index, researchers examine its explanatory power and predictive capacity concerning various macroeconomic and microeconomic indicators.

Subsequently, Chen et al. [7] identified that 40% of the articles filtered using these expert-selected keyword sets were not directly relevant to economic policy uncertainty. To address this, they introduced a denoising step, annotating 47,000 news articles to train supervised models. Their results suggest that removing the noisy articles can improve the explanatory power and predictive capacity. This paper focuses on exploring the potential of leveraging LLMs in the initial steps (keyword recommendation and denoising) of constructing the EPU index and evaluating the index's utility in terms of explainability and predictability.

4 METHOD

In order to understand the capability of LLMs in keyword recommendation, we examine several well-performing LLMs, including GPT-3.5, GPT-4, and Claude 3 Sonnet [3]. Considering that the EPU index is country-specific, we adjust for *nationality* to assess if LLMs can encapsulate the unique characteristics of a given country. Keywords for each country are generated in the primary language spoken there. We investigate results under five nationality settings: U.S. (English), China (Simplified Chinese), Taiwan (Traditional Chinese), Japan (Japanese), and South Korea (Korean). Furthermore, inspired by studies indicating that role-playing may enhance LLMs' performance [10], we also explore the effect of *role* on performance in keyword recommendation. Specifically, we assign LLMs roles such as a newspaper editor, an economist, the Central Bank Governor, or the Minister of Economic Affairs. Additionally, we assess the impact of the *description* of the task on outcomes, i.e., whether providing a general versus a detailed task description, i.e. the definitions in [5], influences the results.[1]

Table 2: F1 scores of keyword recommendation with different roles. The bolded results indicate the best performance among different roles, and "All" represents the average scores across the five countries explored.

Country	Model	Editor	Economist	Minister	Governor
All	GPT-3.5	13.63%	13.38%	13.88%	**13.99%**
	GPT-4	11.23%	**11.99%**	11.53%	11.00%
	Claude 3 Sonnet	13.71%	13.97%	14.43%	**14.91%**
Taiwan	GPT-3.5	14.72%	15.19%	**15.54%**	14.74%
	GPT-4	14.16%	**14.55%**	14.13%	12.03%
	Claude 3 Sonnet	11.25%	12.39%	11.44%	**13.96%**

Table 3: Performance of GPT-3.5 on the denoise task.

Approach		Micro-F1	Macro-F1
Zero-Shot	w/o CoT	0.415	0.572
	w/ CoT	0.401	0.580
Few-Shot	w/o CoT	0.540	0.585
	w/ CoT	**0.672**	**0.674**
Fine-Tuned	w/o CoT	0.372	0.588
	w/ CoT	0.417	0.579
Supervised Model [7]		0.907	0.905

For the denoising task, we utilize the annotated dataset from Chen et al. [7] to evaluate LLMs' effectiveness in zero-shot, few-shot, and fine-tuning scenarios. We also compare the explainability and predictability of the EPU index, constructed under various experimental conditions, employing the following time series model from prior research [15]:

$$y_t = \beta EPU_t + \sum_{i=0}^{2} \alpha_i y_{t-i} + \alpha + \epsilon_t\,, \tag{2}$$

where y represents the target macroeconomic variable, and t denotes the period t. We further examine predictability with the equation:

$$y_{t+1} = \beta EPU_t + \sum_{i=0}^{2} \alpha_i y_{t-i} + \alpha + \epsilon_{t+1}. \tag{3}$$

[1] For detailed prompts, please check out our GitHub repo: https://anonymous.4open.science/r/auto-EPU-0702/README.md

Table 4: Explainability and predictability of the EPU index, constructed based on different keyword sets from GPT-3.5 with varying settings. *, **, and * denote significance under 95%, 99%, and 99.7% confidence levels, respectively. Original denotes the original EPU index that was constructed based on expert-selected keywords.**

Type	Y	Explainability					Predictability				
		Original	Simple	Definition			Original	Simple	Definition		
			Economist	Economist	Governor	Minister		Economist	Economist	Governor	Minister
Employment	Non-Farm	*	**	**	**	*	***	***	***	***	**
Price	CPI								*	*	
Production	IPI	**					**		*	*	*
Stock	Avg. Stock	***		*	*						

5 EXPERIMENT

5.1 Keyword Recommendation

In this section, we evaluate the capability of LLMs to recommend keywords, using expert-selected keywords for various countries as the ground truth for assessment [4, 5, 8, 9, 11]. We conducted ten trials for each model under varying settings and compared the mean precision, recall, and F1 score.

Table 1 presents the results of using different task descriptions. Regardless of the model used, the performance with a detailed task description, i.e., the definition provided by experts in the previous work, is superior to that obtained using a simple description. These results show the importance of integrating expert knowledge in task design and providing the necessary definitions. Conversely, Table 2 displays the results of using various role settings. Adopting specific roles, such as economist, central bank governor, or minister of economic affairs, leads to higher performance compared to the role of a newspaper editor. As we aim to compare these findings with the EPU index constructed for Taiwan, the results from Taiwan are also reported. The keywords recommended by GPT-3.5 are much closer to those suggested by experts, thereby justifying the use of GPT-3.5 in the subsequent discussions.

5.2 Denoise

In this section, we assess the denoising effectiveness of GPT-3.5 across zero-shot, few-shot, and fine-tuning scenarios. Further, we explore the impact of Chain-of-Thought (CoT) reasoning, as described by Wei et al. [19], on performance outcomes. Our methodology aligns with that of Chen et al. [7], employing micro-F1 and macro-F1 scores as evaluative measures. Results are detailed in Table 3. Primarily, models utilizing CoT consistently outperform those that do not. Secondly, within these experiments, the few-shot approach surpasses both zero-shot and fine-tuned models in effectiveness, suggesting that minimal annotations may suffice for removing noises that occur in constructing text-based economic indicators. Thirdly, despite few-shot learning achieving the highest performance among LLM applications, a significant disparity remains when compared to specialized supervised models [7]. This disparity underscores the trade-off between employing LLMs with minimal instance requirements and the comprehensive data annotation needed for custom-tailored models.

5.3 Explainability and Predictability

To compare the utility of the constructed indicators, economists routinely evaluate their explainability and predictability. This paper assesses the EPU indices against four macroeconomic variables: non-farm employment rate, Consumer Price Index (CPI), Industrial Production Index (IPI), and average stock market price (Avg. Stock), which serve as dependent variables. We utilize the same newspaper sources as in previous research [7] spanning from 2003 to 2020.

Table 5: Explainability and predictability of the denoised EPU index. ● denotes significance under 90% confidence level.

Y	Explainability		Predictability	
	Supervised	GPT-3.5	Supervised	GPT-3.5
Non-Farm	***	●	***	***
CPI	*		*	
IPI	*	●		
Avg. Stock	*	**	●	

Table 4 presents the statistics. First, the EPU index, enriched with keywords suggested by LLMs, exhibits improved explainability and predictability in the employment ratio compared to the original EPU index. This highlights the potential of LLMs in automating the construction of text-based economic indicators. Secondly, in the CPI analysis, only the keyword set under a specific definition setting yielded significant predictability for CPI and IPI. It underscores the importance of precise and detailed definitions in tasks. Additionally, role simulations as an economist or Central Bank Governor yielded better outcomes than as the Minister of Economic Affairs, indicating that role design influences results. Thirdly, while none of the EPU indices could predict the growth rate of the market price, the original EPU index showed better explainability for the stock market. This result points to a discrepancy between model-selected and expert-selected keyword sets.

We further compare the performance of a tailor-made supervised model [7] and GPT-3.5 (few-shot) in denoising from both explainability and predictability perspectives, as shown in Table 5. Interestingly, while the supervised model outperformed in three of the four economic indicators from an explainability standpoint, GPT-3.5 demonstrated good performance in three indicators and even surpassed the supervised model in explaining the stock market growth rate. From a predictability aspect, the model from prior work outperformed GPT-3.5.

In summary, this section illustrates the potential of GPT-3.5 in the automated construction of EPU indices and emphasizes the significance of task design and role-play. Compared with expert-selected keyword sets, GPT-3.5 achieved comparable performance in some cases and inferior in others. However, in the denoising

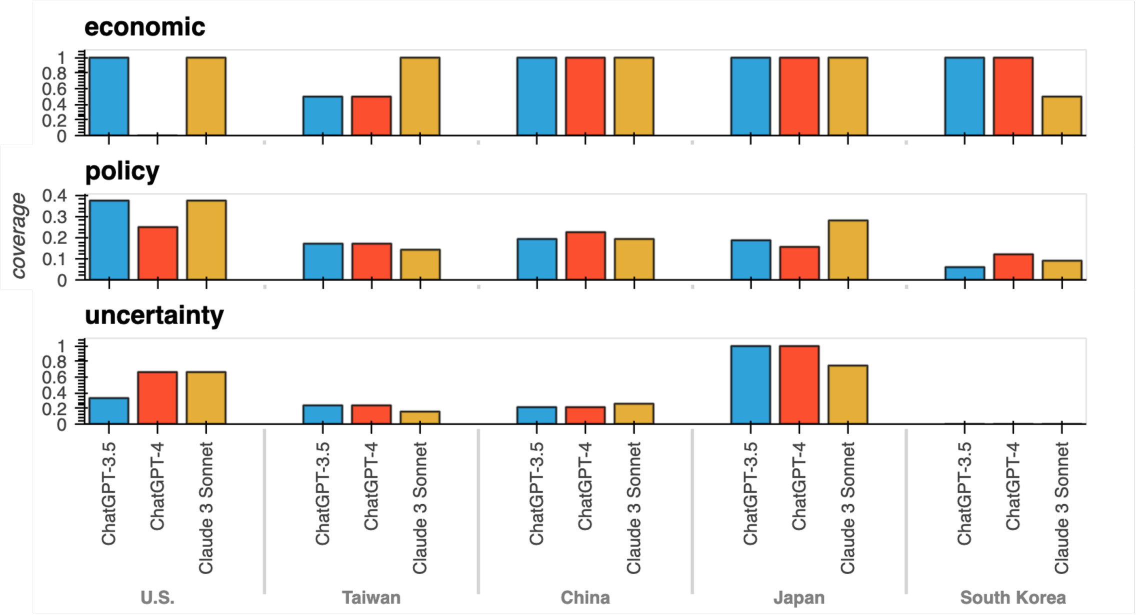

Figure 1: Coverage in keyword recommendation.

phase, the challenges of using GPT-3.5 to eliminate noise caused by keyword-based article selection persist. With a tailor-made model and extensive annotated data, the utility of the constructed EPU index would surpass that achieved using GPT-3.5 for denoising.

6 DISCUSSION

6.1 Recommended Keyword Analysis

Given that the outcomes of using the strategy with definition and playing as an economist in Table 4 are the best from both explainability and predictability aspects, we explore the same strategy for all countries in the keyword recommendation task. Figure 1 shows the coverage of the keyword recommendation results of different models for various countries. Here, coverage refers to whether the expert-selected keywords are included in the model-recommended sets. We observe that models perform well in recommending economic-related keywords and are satisfactory for uncertainty-related keywords, particularly for the U.S. and Japan. However, the policy-related keywords are more country-oriented. LLMs perform relatively worse compared with other cases, and regardless of the LLMs used, the performance for the U.S. is consistently the best. This result aligns with the findings of previous studies [1, 18], indicating that models are primarily trained on the cultures, behaviors, and solutions of English-speaking users. It also offers insight for future studies interested in constructing country-oriented economic indexes, highlighting the need to consider the models' ability to understand country-specific topics.

6.2 Limitation

Although we explore the keyword suggestion task across various countries, we limit our investigation of the denoising task and the examination of explainability and predictability to Taiwan's data due to the availability of news articles and the reproducibility of previous studies. Future studies could extend our insights to different countries, particularly benefiting researchers with limited budgets for annotations. Additionally, we did not test the initial step of constructing an index, namely designing and defining the index; instead, we adopted the definition from a prior study. This approach restricts our focus to keyword recommendation and denoising tasks and precludes discussions on the creativity of LLMs. On the other hand, our primary aim is to explore how LLMs can facilitate the automation of text-based economic indicator construction, not focusing on model development or training. Future research could delve deeper into constructing an economic indicator with robust explainability and predictability from scratch.

7 CONCLUSION

This paper explores using LLMs to automate the creation of text-based economic indicators like the EPU index. LLMs can streamline keyword selection and article denoising tasks, traditionally requiring significant human labor. The findings indicate that an LLM-designed EPU index, with properly refined prompts, aligns well with conventional indexes. Particularly, the few-shot learning method excels over other methods in denoising, showing that minimal targeted data can substantially boost LLM effectiveness. Nonetheless, there's still a performance gap compared to manually created indexes. It indicates the need for domain expertise in LLMs. Despite this, LLMs present a valuable tool in economic research, especially where quick, scalable methods are essential, though with certain limitations. Future research could focus on refining the integration of LLMs in economic research, potentially through hybrid models that combine human expertise with the scalability and efficiency of LLMs.

REFERENCES

[1] Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language We Prompt Them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 6330–6340. https://aclanthology.org/2024.lrec-main.560

[2] Hites Ahir, Nicholas Bloom, and Davide Furceri. 2022. The World Uncertainty Index. https://doi.org/10.3386/w29763 arXiv:29763

[3] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

[4] Elif C Arbatli, Steven J Davis, Arata Ito, and Naoko Miake. 2017. *Policy uncertainty in Japan.* Technical Report. National Bureau of Economic Research.

[5] Scott R. Baker, Nicholas Bloom, and Steven J. Davis. 2016. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics* 131, 4 (Nov. 2016), 1593–1636. https://doi.org/10.1093/qje/qjw024

[6] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter Mood Predicts the Stock Market. *Journal of Computational Science* 2, 1 (March 2011), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[7] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, and Hsin-Hsi Chen. 2021. Constructing Noise Free Economic Policy Uncertainty Index. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* 2915–2919.

[8] Dooyeon Cho and Husang Kim. 2023. Macroeconomic effects of uncertainty shocks: Evidence from Korea. *Journal of Asian Economics* 84 (2023), 101571.

[9] SJ Davis, D Liu, and SX Sheng. 2019. Economic policy uncertainty and trade policy uncertainty for China. *Working Paper. University of Chicago Booth School of Business* (2019).

[10] Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. LEGO: A Multi-agent Collaborative Framework with Role-playing and Iterative Feedback for Causality Explanation Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9142–9163. https://doi.org/10.18653/v1/2023.findings-emnlp.613

[11] Yu-Lieh Huang, Jin-Huei Yeh, and Chung-Chi Chen. 2019. Economic Policy Uncertainty Index for Taiwan. *Taiwan Economic Review* (2019).

[12] Manish Jha, Jialin Qian, Michael Weber, and Baozhong Yang. 2024. ChatGPT and Corporate Policies. https://doi.org/10.3386/w32161 arXiv:32161

[13] Katherine Keith and Amanda Stent. 2019. Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 493–503. https://doi.org/10.18653/v1/P19-1047

[14] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[15] Shibley Sadique, Francis In, Madhu Veeraraghavan, and Paul Wachtel. 2013. Soft information and economic activity: Evidence from the Beige Book. *Journal of Macroeconomics* 37 (2013), 81–92.

[16] Lee A. Smales. 2023. Classification of RBA Monetary Policy Announcements Using ChatGPT. *Finance Research Letters* 58 (Dec. 2023), 104514. https://doi.org/10.1016/j.frl.2023.104514

[17] Minchae Song and Kyung-shik Shin. 2019. Forecasting Economic Indicators Using a Consumer Sentiment Index: Survey-based versus Text-Based Data. *Journal of Forecasting* 38, 6 (2019), 504–518. https://doi.org/10.1002/for.2584

[18] Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning. *arXiv preprint arXiv:2309.04766* (2023).

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[20] Stephen Yang. 2023. Predictive Patentomics: Forecasting Innovation Success and Valuation with ChatGPT. https://doi.org/10.48550/arXiv.2307.01202 arXiv:2307.01202 [cs]