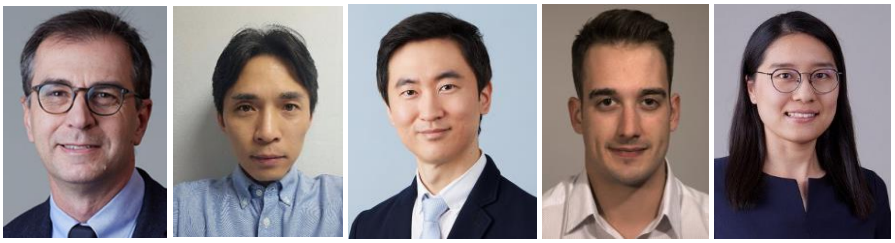
SUERF
The European Money and Finance Forum

SUERF Policy Brief
No 1054, December 2024

Enhancing Central Bank Communication: Domain-Specific Language Models



Leonardo Gambacorta, Byeungchun Kwon, Taejin Park, Pietro Patelli, and Sonya Zhu
Bank for International Settlements (BIS)

Keywords: large language models, gen AI, central banks, monetary policy analysis

JEL codes: E58, C55, C63, G17

Abstract

Central bank communication is a critical tool for managing public expectations regarding monetary policy decisions. Recent advances in economic research increasingly leverage Natural Language Processing (NLP) to quantify the information conveyed through such communication. This policy brief introduces central bank language models (CB-LMs), specifically tailored for applications in the central banking domain. We show that CB-LMs outperform foundational models in predicting masked words within central bank-specific idioms and excel in classifying monetary policy stances from FOMC statements, surpassing even state-of-the-art generative Large Language Models (LLMs) in the latter task. Furthermore, while the leading generative LLMs show exceptional performance in complex tasks, such as analysing long news articles with limited training data, they face notable challenges related to confidentiality, transparency, replicability and cost-efficiency.

Disclaimer: This policy brief is based on "CB-LM: language models for central banking", BIS Working Paper, No 1215/2024. The views expressed are those of the authors and do not necessarily reflect those of the Bank for International Settlements.

Introduction

Economic literature increasingly applies natural language processing (NLP) techniques to analyse monetary policy communications (Acosta and Meade, 2015; Ehrmann and Talmi, 2020). While these studies offer valuable insights, they often rely on language models trained on general text corpora. This limitation may restrict the models' ability to fully capture the nuances specific to central banking and monetary economics. Recent research suggests that retraining language models on domain-specific datasets can significantly enhance their performance in specialised NLP tasks.

To address the need for domain-specific NLP tools in monetary economics and central banking, we introduce central bank language models (CB-LMs), which are trained on a large-scale central banking corpus. Using encoder-based models like BERT and RoBERTa, we retrain these models with central bank speeches and policy papers. CB-LMs demonstrate superior capabilities in understanding central bank semantics, outperforming foundational models in predicting masked words in central bank idioms and classifying monetary policy stance. We also compare CB-LMs with state-of-the-art generative Large Language Models (LLMs), which despite their extensive pretraining on diverse datasets and minimal retraining requirements, could face challenges in achieving domain-specific precision.

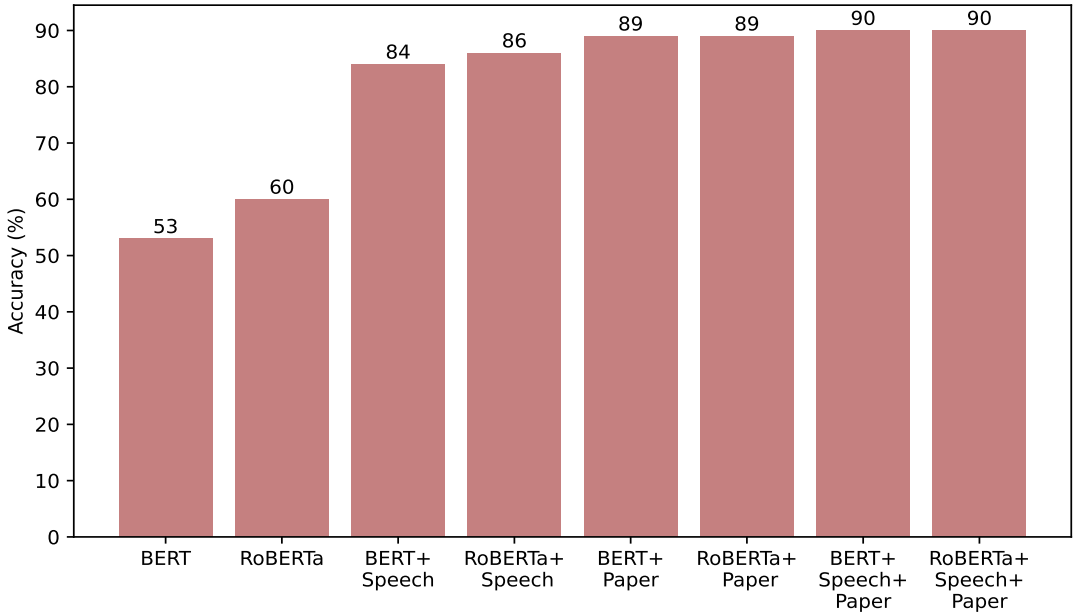
Our goal is to develop domain-specific models that enable more accurate analysis of monetary policy. Additionally, we explore the adaptability of different LLMs in central banking and assess their performance across different training methods and tasks. Our findings aim to guide central bankers in selecting language models best suited to their specific needs.

A two-step approach: domain adaptation and fine-tuning

The development and application of CB-LMs involves two key phases: domain adaptation and fine-tuning. In the first phase, the model is trained using unsupervised learning on an extensive corpus of central banking texts, comprising 37,037 research papers and 18,345 speeches. This phase enables the model learn linguistic elements such as grammar, idioms, semantics, and structural patterns unique to central banking. In the second phase, the model undergoes supervised learning on task-specific datasets, a process known as fine-tuning. This step refines the model’s capabilities for specialised tasks, such as sentiment classification within the central banking context. We chose foundational language models like BERT and RoBERTa due to their widespread acceptance in the NLP community and their relatively manageable computational requirements. These models were then customised specifically for the central banking domain.

For domain-adaptation, we used Masked Language Modelling (MLM) to enhance the models' bidirectional understanding of central banking terminology. This approach involves randomly masking words in sentences and retraining BERT and RoBERTa to predict the masked words. Figure 1 illustrates the performance of our six CB-LMs compared to the foundational models in predicting masked words within idioms specific to central banking. All CB-LMs significantly outperform their foundational models, with the top-performing models accurately predicting 90 out of 100 masked words. In comparison, RoBERTa and BERT predict only 60 and 53, respectively. These results demonstrate the successful adaptation of CB-LMs to the central banking domain. Moreover, we find that performance improvements are correlated with the size of the training datasets, with models trained on a combined dataset of research papers and speeches achieving the greatest enhancements.

Figure 1. Performance of the masked word test



Notes: This graph compares the performance of foundation models and our six CB-LMs in the masked word test. Y-axis represents the percentage of correct predictions from each model.

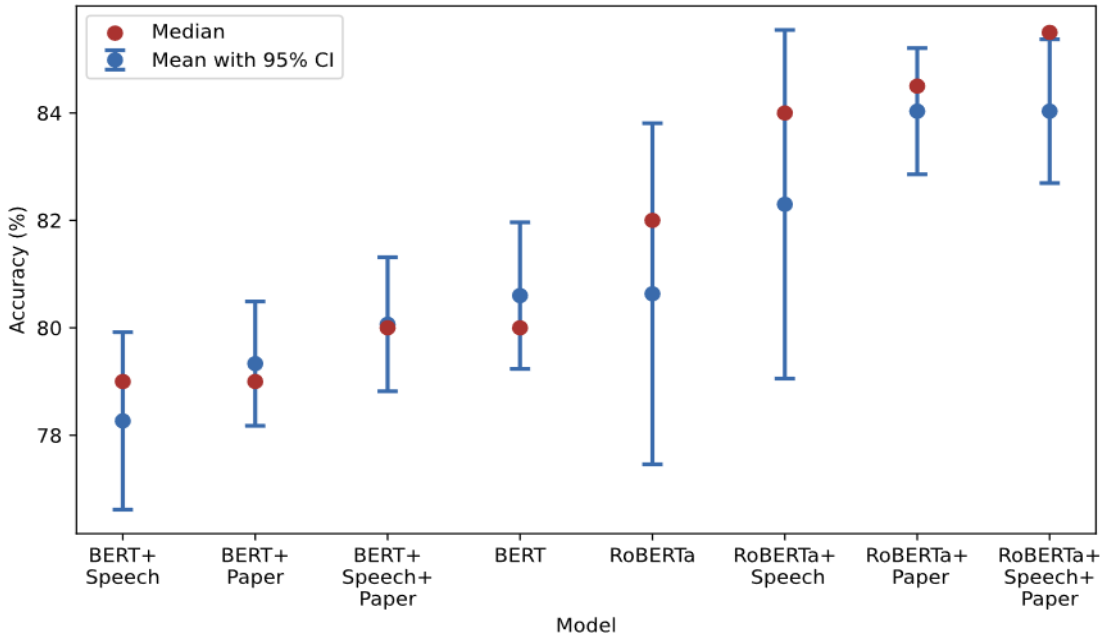
Monetary policy sentiment analysis

Building on the successful domain adaptation phase, we fine-tune CB-LMs to enhance their ability to perform specialised tasks, specifically classifying the monetary policy stance of Federal Open Market Committee (FOMC) statements as dovish, hawkish or neutral. Excelling in this task could significantly aid central bankers in formulating and executing effective monetary policy communication strategies.

For this application, we train CB-LMs using a dataset of 1,243 sentences from historical FOMC statements (1997–2010), each manually labelled by Gorodnichenko et al. (2023). We randomly allocate 80% of the sentences for training and the remaining 20% for testing. To ensure the robustness of our findings, we repeat the evaluation process 30 times, measuring the out-of-sample performance based on the percentage of correctly classified sentences.

The results show that RoBERTa-based CB-LMs consistently outperform their foundational counterparts (Figure 2). For instance, the top-performing CB-LM, extensively trained on central bank papers and speeches, achieves a mean accuracy of 84%, compared to 81% for the foundational RoBERTa model. In contrast, BERT-based CB-LMs do not clearly exhibit improved performance. While domain adaptation generally aims to enhance the LLM performance, these findings suggest that such adaptations do not guarantee improvements across all scenarios.

Figure 2. Classifying monetary policy stance



Notes: This figure reports the performance of CB-LMs alongside two foundation models in classifying the stance of FOMC statements. Sentences from the FOMC statements are manually labelled as Dovish, Hawkish or Neutral. The models are fine-tuned with 80% of these sentences and their corresponding manual labels. Then, we task the language models with predicting the monetary policy stance for the rest 20% of sentences. The prediction from language models is considered as "correct" when it is consistent with the expert's manual label.

A comparison with generative LLMs

We extend our analysis by evaluating state-of-the-art generative LLMs, including ChatGPT, Llama and Mixtral, in the context of central bank communication. To this end, we replicate the monetary policy sentiment analysis using these generative models, with two key methodological differences. First, we forgo domain adaptation, given these models' extensive pretraining on diverse datasets likely containing central bank-related text. Second, we employ various task-specific training strategies, including fine-tuning and in-context learning methods. For fine-tuning, we leverage supervised fine-tuning (Ziegler et al, 2019), direct preference optimization (Rafailov et al, 2023) and proprietary techniques provided by OpenAI. For in-context learning, we apply few-shot learning with examples that are either randomly sampled or retrieved as the most similar to the task.

Our findings indicate that fine-tuning significantly enhances generative LLM performance. For example, ChatGPT-3.5 achieves up to 88% accuracy post fine-tuning, compared to 56% without it. Model size emerges as a critical factor, with larger models like Llama-3 70B consistently outperforming smaller versions such as Llama-3 8B. Additionally, retrieval-based in-context learning with task-specific few-shot examples further improves performance, boosting ChatGPT-4 Turbo's accuracy to 81%, compared to 73% with random sampling methods.

However, despite these improvements, generative LLMs often underperform encoder-based models like RoBERTa in simple classification tasks, underscoring fundamental differences in architecture and training objectives. These results suggest that while generative LLMs excel in complex, multi-dimensional tasks, smaller encoder-based models could remain more efficient and suitable for straightforward classification applications, particularly given their lower computational requirements.

To evaluate the applicability of language models in more challenging scenarios, we examine the performance of CB-LMs and generative LLMs in analysing monetary policy sentiment in US monetary policy news articles, manually classified as hawkish, dovish or neutral. This task requires long-text analysis and relies on limited training data.

In this test, generative LLMs, such as ChatGPT-4 and Llama-3 70B achieve higher accuracy (80–81%) without additional training, benefiting from extensive pretraining and superior ability to handle complex texts. In contrast, the best-performing RoBERTa-based CB-LM achieves an average accuracy of 65% after fine-tuning, only marginally surpassing ChatGPT-3.5's zero-shot predictions. These findings underscore the advantage of leading generative LLMs in managing long-range dependencies and navigating contextual variability in extended texts, offering a clear edge over CB-LMs for complex application.

Despite these advantages, large generative LLMs present also significant risks. The main challenges for central banks in using these models include: (1) concerns about confidentiality and privacy when handling sensitive data, (2) a lack of transparency and replicability due to proprietary architectures, (3) high computational and financial costs associated with their deployment and fine-tuning, and (4) the risk of variability in outputs, which may undermine consistency in critical applications like policy analysis. These challenges necessitate careful evaluation before adopting such models for central banking tasks.

Conclusions

This policy brief describes the introduction of CB-LMs – language models retrained on a large-scale collection of central banking texts. By using prominent models like BERT and RoBERTa and adding texts tailored to central banking – including speeches, policy notes and research papers – CB-LMs effectively capture domain-specific semantics, terminologies and contextual nuances. Our primary goal is to develop and publicly release CB-LMs to advance NLP analysis in monetary economics and central banking.[1] Additionally, our comprehensive assessment of different LLMs across various training settings provides insights into model selection tailored to central bankers' specific tasks and technical requirements.

We show that CB-LMs outperform their foundational models in predicting masked words within central bank idioms. Some CB-LMs surpass not only their original models but also state-of-the-art generative LLMs in classifying monetary policy stances from FOMC statements. CB-LMs excel at processing nuanced expressions of monetary policy, which could make them valuable tools for central banks in real-time analysis and decision-making. Nonetheless, in more complex scenarios – such as those involved limited data for fine-tuning and processing longer text inputs – the largest LLMs, like ChatGPT-4 and Llama-3 70B, exhibit superior performance. However, despite their advantages, deploying these LLMs presents significant challenges for central banks, including concerns about confidentiality, transparency, replicability and cost-efficiency.

[1] The full documentation of CB-LMs can be downloaded here: https://www.bis.org/publ/work1215.htm.

References

Acosta, M., and E. Meade (2015), "Hanging on Every Word: Semantic Analysis of the FOMC's Post-meeting Statement", FEDS Notes 2015-09-30, Board of Governors of the Federal Reserve System.

Ehrmann, M. and J. Talmi (2020), "Starting from a Blank Page? Semantic Similarity in Central Bank Communication and Market Volatility", Journal of Monetary Economics, 111: 48–62.

Gorodnichenko, Y., T. Pham and O. Talavera (2023), "The Voice of Monetary Policy". American Economic Review, 113(2), 548–584.

Rafailov, R., A Sharma, E Mitchell, C.D. Manning, S. Ermon, and C. Finn (2023), "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model", Advances in Neural Information Processing Systems, 36.

Ziegler, D. M., N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving (2019), "Fine-Tuning Language Models from Human Preferences", arXiv preprint, arXiv:1909.08593.

About the author(s)

Leonardo Gambacorta is the Head of the Innovation and the Digital Economy unit at the Bank for International Settlements. His main interests include the monetary transmission mechanisms, the effectiveness of macroprudential policies on systemic risk, and the effects of technological innovation on financial intermediation.

Byeungchun Kwon is a Senior Financial Market Analyst at the Bank for International Settlements, with research interests in financial market simulation using deep reinforcement learning.

Taejin Park is the Head of Financial Markets Research Support at the Bank for International Settlements, specialising in AI technologies as well as energy and climate change economics.

Pietro Patelli is a Senior Data Scientist at the Bank for International Settlements and the Financial Stability Board. His research interests include machine learning, financial markets, and international economics.

Sonya Zhu is an Economist at the Monetary and Economic Department of Bank for International Settlements (BIS). Her research interests are empirical asset pricing, market microstructure and information economics, with a particular focus on the interplay between monetary policy and financial markets. Sonya holds a PhD in Finance from the Stockholm School of Economics.

SUERF Policy Briefs and Notes disseminate SUERF Members' economic research, policy-oriented analyses, and views. They analyze relevant developments, address challenges and propose solutions to current monetary, financial and macroeconomic themes. The style is analytical yet non-technical, facilitating interaction and the exchange of ideas between researchers, policy makers and financial practitioners.

SUERF Policy Briefs and Notes are accessible to the public free of charge at https://www.suerf.org/publications/suerf-policy-notes-and-briefs/.

The views expressed are those of the authors and not necessarily those of the institutions the authors are affiliated with.

Editorial Board: Ernest Gnan, David T. Llewellyn, Donato Masciandaro, Natacha Valla

Designed by the Information Management and Services Division of the Oesterreichische Nationalbank (OeNB)

SUERF Secretariat
c/o OeNB, Otto-Wagner-Platz 3A-1090 Vienna, Austria
Phone: +43 1 40 420 7206
E-Mail: suerf@oenb.at
Website: https://www.suerf.org/