BIS Working Papers
No 1215

CB-LMs: language models for central banking

by Leonardo Gambacorta, Byeungchun Kwon, Taejin Park, Pietro Patelli, Sonya Zhu

Monetary and Economic Department

October 2024

JEL classification: E58, C55, C63, G17

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

This publication is available on the BIS website (www.bis.org).

ISSN 1020-0959 (print)
ISSN 1682-7678 (online)

CB-LMs: language models for central banking

Leonardo Gambacorta[+]*, Byeungchun Kwon[+], Taejin Park[+], Pietro Patelli[+], Sonya Zhu[+]

[+] Bank for International Settlements (BIS), * CEPR

Abstract

We introduce central bank language models (CB-LMs) — specialised encoder-only language models retrained on a comprehensive corpus of central bank speeches, policy documents and research papers. We show that CB-LMs outperform their foundational models in predicting masked words in central bank idioms. Some CB-LMs not only outperform their foundational models, but also surpass state-of-the-art generative Large Language Models (LLMs) in classifying monetary policy stance from Federal Open Market Committee (FOMC) statements. In more complex scenarios, requiring sentiment classification of extensive news related to the US monetary policy, we find that the largest LLMs outperform the domain-adapted encoder-only models. However, deploying such large LLMs presents substantial challenges for central banks in terms of confidentiality, transparency, replicability and cost-efficiency.

JEL Classification: E58, C55, C63, G17.

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

We thank Douglas Araujo, Flavio Bazzana, Claudia Biancotti, Carolina Camassa, Chris Cox, Fernando Perez-Cruz, and seminar participants at the BIS and Generative AI in Finance Paper Development Workshop at TU Dresden for useful comments and suggestions. The views expressed in this article are those of the authors and do not necessarily reflect those of the BIS.

1. Introduction

Communication is becoming an increasingly important tool for central banks to manage public expectation. Leveraging the power of language models, there is a growing body of economic literature applying Natural Language Processing (NLP) techniques to decipher central bank communication. Words, vocal tones and body languages have all been found as effective channels for central banks to manage public expectations.

While these studies have made valuable contributions, most language models used in economic literature have been trained on general text corpora, which inherently limits their ability to fully capture the intricacies and nuances specific to central bank communication. Notably, Gorodnichenko et al (2023) and Hansen and Kazinnik (2023) employ BERT (Bidirectional Encoder Representations from Transformer) and GPT (Generative Pre-trained Transformers) for their analysis, both of which are general-purpose language models. To overcome this limitation, recent NLP literature (Lee et al (2019), Huang et al (2023)) suggests that retraining language models on targeted and more comprehensive domain-specific corpora can significantly enhance the performance of NLP analysis.

To address the need for domain-specific NLP analysis in monetary economics and central banking research, we develop central bank language models, referred to as CB-LMs, which are specifically trained on a large-scale central banking corpus. In developing these models, we leverage prominent encoder-only language models, including BERT and RoBERTa (Robustly Optimized BERT Pretraining Approach), and retrain them with a corpus that includes central bank speeches and policy and research papers.

We find that CB-LMs excel in understanding the specific semantics, terminologies, and contextual nuances within the central bank domain. In particular, they outperform their foundational encoder-only models in two key areas: 1) predicting masked words in central bank idioms and 2) classifying stance in official monetary policy decision statements. Furthermore, we compare CB-LMs with state-of-the-art generative Large Language Models (LLMs). The latter require less retraining on central banking corpora due to their extensive pretraining on vast and diverse data sets.

The main aim of this paper is to develop and disseminate high-performing CB-LMs that have the potential to level the playing field for NLP analysis in monetary economics and central banking. By providing researchers and practitioners with access to domain-specific language models, our research open ups new possibilities for more accurate and insightful analysis of monetary policy and related topics. Additionally, in this paper, we explore the adaptability of different LLMs within the context of central banking, examining their performance across diverse training approaches and various downstream task scenarios. Our comprehensive assessment of these LLMs across different training settings offers central bankers deeper insights, enabling them to make more informed decisions when selecting models tailored to the specific requirements of their tasks and technical environments.

The rest of the paper is organized as follows. Section 2 provides a literature review, outlining the current landscape of NLP research in monetary economics and finance. Section 3 describes our approach, detailing the data collection process and the technical steps involved in developing CB-LMs. Section 4 employs a masked word test to evaluate the performance of foundational models and CB-LMs in identifying idioms commonly utilised by central bankers. Section 5 presents the applications of CB-LMs in classifying monetary policy sentiment. Section 6 expands the performance analysis to state-of-the-art generative LLMs. Section 7 introduces a more demanding task to establish further benchmarks among LLMs. Section 8 addresses key considerations for deploying generative LLMs within central banking contexts. The last section concludes.

2. Literature review

Communication tools, including the release of meeting transcripts/minutes and speeches by officials, have been increasingly used by central banks in the past two decades. A number of studies evaluate the effectiveness of central bank communications by leveraging different techniques from computational linguistics.

Traditional approaches typically involve a bag-of-word technique, in which only word frequency matters. For instance, Acosta and Meade (2015) and Ehrmann and Talmi (2020) count sentiment-embedding words in monetary policy statements. By using sentiment analysis, they are able to gauge the emotional tone and subjective aspects of the text, providing additional insights into the policy stance. In contrast, Boukus and Rosenberg (2006) consider the distribution of semantics and extract economic themes from Federal Open Market Committee (FOMC) minutes using Latent Semantic Analysis. They demonstrate that the release of FOMC minutes moves Treasury yields, and the particular reaction depends on the specific themes identified. This heterogeneity in reaction is very relevant because it highlights the differential impact of various economic themes on market behaviour, underscoring the importance of nuanced communication in monetary policy.

More advanced textual analysis approaches have been adopted in the literature to filter relevant information. Hansen et al (2017) use a probabilistic topic modelling algorithm — Latent Dirichlet Allocation (LDA) (see Blei et al (2003)) — to decompose FOMC transcripts in terms of the fraction of time spent covering a variety of topics. They find that the publication of meeting transcripts after 1994 reduces the deliberations of FOMC committee members, implying that transparency in central bank communication can influence the internal dynamics of policy discussions.

Recently, a few studies started to apply deep learning models to monetary economic research. Curti and Kazinnik (2023) apply convolutional neural network (CNN) to FOMC press conference videos, a method particularly relevant for analysing visual cues and non-verbal communication, adding another dimension to the understanding of central bank communication. Gorodnichenko et al (2023) apply BERT to FOMC press conference audios, focusing also on the tone of the reading to extract more nuanced information. Interestingly, both Curti and Kazinnik (2023) and Gorodnichenko et al (2023) find that financial markets respond to non-verbal communications of Federal Reserve chairs.[1]

Our CB-LMs differ from existing models by being specifically designed and trained on a large-scale central banking corpus, capturing the specific semantics, terminologies, and

[1] Complementing the NLP-focused research, Aruoba et al (2021) utilise a large-scale macroeconomic model to evaluate the impact of the Federal Reserve's new monetary policy framework. Their approach provides a comprehensive analysis of the macroeconomic implications of monetary policy decisions, offering valuable insights into the broader economic context. However, their methodology relies on traditional macroeconomic modelling techniques and does not incorporate the nuanced textual analysis offered by NLP techniques.

contextual nuances within the domain. In this paper, we introduce several domain-specific language models, offering new possibilities for more accurate and insightful analysis of monetary policy and related topics.

In a similar vein, Aruoba and Drechsel (2023) use deep learning techniques to develop a sentiment analysis approach in the spirit of Hassan et al (2020), capturing the sentiment surrounding economic concepts within a 10-word window in a document. They used a dictionary of positive and negative terms, which they modified to better suit the language used in Fed documents. Each positive or negative word influenced the sentiment score of the concept, providing a nuanced understanding of the sentiment surrounding each economic concept. This approach, while different in its application, shares with our CB-LMs the aim of capturing the specific semantics, terminologies, and contextual nuances within the domain of central banking and monetary economics.

Besides monetary policy, there is also a rapidly growing literature that adopt textual analysis to measure a vast array of economic variables, counting the frequencies of topic words in text files. For instance, textual analysis has been applied to measure economic policy uncertainty (Baker et al (2016), Husted et al (2020)), partisan conflict (Azzimonti (2018)), geopolitical risk (Caldara and Iacoviello (2022)), and product similarity (Hoberg and Phillips (2010), Hoberg and Phillips (2016)). Our CB-LMs' main advantage over a word-counting approach is that the transformer-based language models capture "context" of the text much better than simple frequency counting. As a result, CB-LMs are sufficiently flexible and can be easily applied for the measurement of economic variables besides monetary policy stances, providing a comprehensive tool for economic analysis.

Our study is closely related to the work of Pfeifer and Marohl (2023), which investigates the types of economic agents (government, financial intermediaries, households, firms) involved in central bank communication and the sentiment surrounding them. They developed a fine-tuned language model for sentiment classification, utilizing a general-purpose language model, RoBERTa, and a corpus of manually labelled central bank speeches. Our study contributes to the literature by introducing domain-adapted models that achieve better performance for downstream NLP tasks in central banking.

3. Methodology

In this section, we describe the steps involved in developing CB-LMs. These models are essentially refined adaptations of foundational language models tailored for the nuanced domain of central banking and monetary policy.

The development of CB-LMs involves two fundamental phases: domain adaptation and fine-tuning. In the domain adaptation phase, the model undergoes unsupervised learning on an extensive text corpus. This process imparts a foundational understanding of linguistic elements, encompassing grammar, idioms, semantics, and structural patterns. Through this phase, the model develops a comprehensive grasp of language. Subsequent to the domain adaptation phase, the model undergoes fine-tuning, adapting to specific tasks through supervised learning on a more focused, task-oriented dataset. This process refines the model's parameters, enhancing its performance in specialised tasks like text classification and question-answering within the central banking context.

Figure 1 presents the domain adaptation process of our language models. We start by assembling a corpus of central banking, including speeches and research papers curated by the BIS through its Central Bank Hub.[2] For the purpose of this paper, our dataset incorporates 37,037 research papers (2.7 Gigabytes) and 18,345 speeches (0.34 Gigabytes). We then pre-process the text data and generate three sets of encoded tokens from them: one based on the speech text, one based on the research paper text, and a combined set encompassing both the speech and paper text.

In the third step, we choose foundational language models to be utilised in the central banking domain. These models are typically trained on general text such as Wikipedia and BookCorpus. Our selection of foundation models is guided by two key criteria. First, the models should have a broad acceptance and usage within the NLP community. Second, their computational requirements, typically associated with model size, must align with our computation capabilities, especially the graphics processing unit (GPU) infrastructure. Taking these factors into account, we select the base BERT model (Devlin et al (2019))

[2] The Central Bank Research Hub contains central bank publications that are featured on the Research Papers in Economics (RePEc) website.

developed by Google, and the base RoBERTa model (Liu et al (2019)) developed by Meta. We customise selected foundation models to the central banking domain. By adapting the two foundation models to these three datasets, we end up with six unique central bank language models.

To improve the models' bidirectional understanding of central banking terminology, we employ Masked Language Modelling (MLM). More specifically, we randomly mask a token in a sentence in the central bank datasets, and retrain BERT and RoBERTa to predict these masked tokens from their surrounding tokens.

To validate the effectiveness of CB-LMs in downstream NLP tasks related to central banking, we fine-tune them for specific applications related to monetary policy communication, benchmarking their performance against the original foundation models.

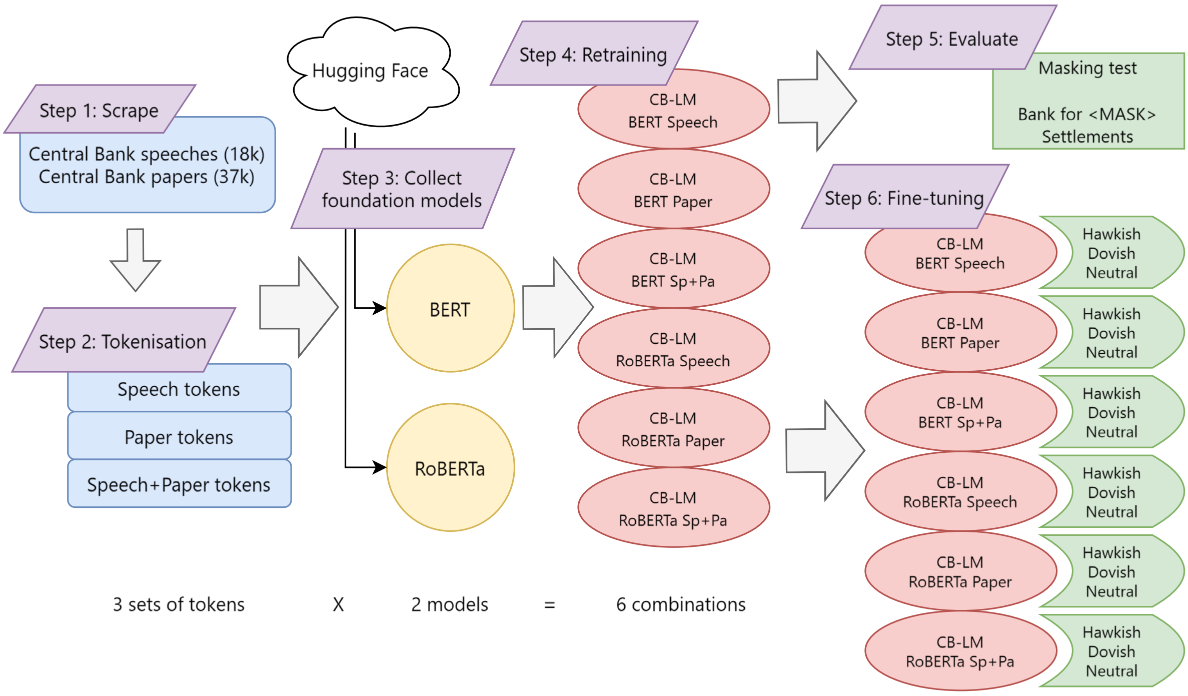

Figure 1. **Overview of CB-LMs development process.** This graph shows the steps for developing the CB-LMs. Step 1-5 are steps for re-training the foundational models with central bank texts. And Step 6 is to fine-tune the models for downstream tasks.

4. Evaluating domain adaptation: predicting central banking idioms

In this section, we follow the NLP literature and use a standard masked word test to evaluate the performance of CB-LMs. The test uses a manually curated dataset comprising

100 idioms commonly used by central banks (see Table 1). The idioms prevalent in the domain of central banking may not be immediately comprehensible to language models that are less versed in this specialised field. Each of these idioms consists of at least three words, with one of their middle words randomly masked. This approach enables us to evaluate language models' bidirectional comprehension of terminology specific to central banking. Superior performance in this assessment signifies the effectiveness of the retraining aimed at domain adaptation.

Masked idioms	Correct answers
Accommodative <mask> policy	monetary
Asset <mask> program	purchase
Balance <mask> payments	of
Bank <mask> International Settlements	for
Basel <mask> on Banking Supervision	Committee
Bretton <mask> system	Woods
Capital <mask> ratio	adequacy
⋮	⋮

Table 1. **Example of idioms used in the masked word test.** For the full list of the test dataset, see Appendix 1.

Figure 2 shows the performance of our six CB-LMs and the two foundation models in predicting masked words. Notably, all six CB-LMs outperform the foundation models by a considerable margin. Specifically, CB-LMs successfully predict 90 out of the 100 masked words. In contrast, the foundational models RoBERTa and BERT correctly predict only 60 and 53 words, respectively. These outcomes suggest the successful adaptation of CB-LMs into the central banking domain. It is noteworthy that performance improvements are directly proportional to the size of the training datasets. Models trained with a combined dataset of papers and speeches exhibit the greatest performance enhancements, followed by those trained solely on papers or speeches.

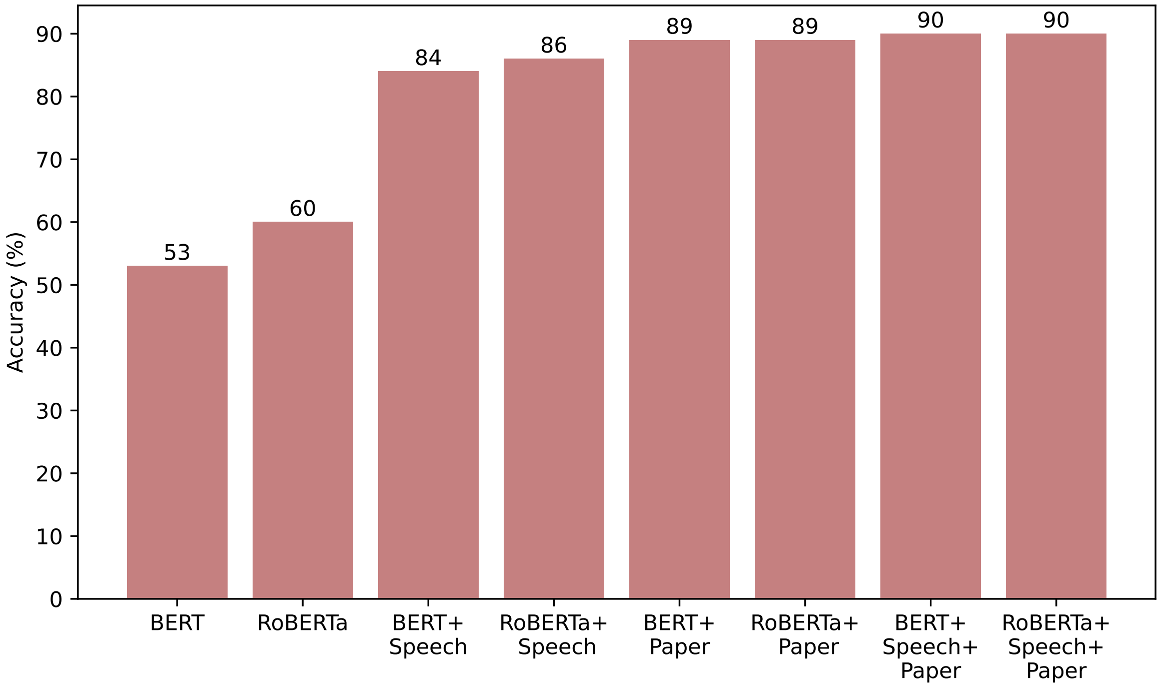

Figure 2. **Performance of the masked word test.** This graph compares the performance of foundation models and our six CB-LMs in the masked word test. Y-axis represents the percentage of correct predictions from each model.

5. Monetary Policy Sentiment Analysis

The outperformance of the CB-LMs in the masked word test suggests that these models are adeptly optimised to yield more accurate and context-aware predictions in the central bank domain. To exemplify this capability, we fine-tune CB-LMs for a specific application pertinent to monetary policy sentiment analysis. This fine-tuning process is tailored to refine the CB-LMs' parameters to enhance their performance in recognising and adapting to particular nuances, vocabulary, styles, or objectives associated with monetary policy communications.

We consider an application to evaluate the performance of CB-LMs for the classification of monetary policy stance within each sentence of FOMC statements. Achieving excellent performance in this task would greatly benefit central bankers to formulate and execute monetary policy communication strategies.

Monetary policy shocks, including those embedded in central bank communications, impact the real economy (Nakamura and Steinsson (2018)). As a result, central bank governors always need to consider, ex ante, the potential implications of their public communications on financial stability. However, this ex-ante evaluation is subtle and often cumbersome for central banks, largely because evaluating the sentiment from textual information typically involves subjective judgements from several central bank officials with relevant expertise and experience. For example, it can be difficult, even for an expert, to evaluate the sentiment from a sentence like “the policy path remains data-dependent”, if not considered in the right context. Against this background, our CB-LMs provide a novel tool for understanding the stance of a central bank statement using a scientific and systematic approach. They can offer quantitative measures that can greatly complement the qualitative expertise of central bank officials. Recognising the importance of this task, we fine-tune CB-LMs to classify the monetary policy stance of historical FOMC statements and test their performance against general-purpose foundation language models, such as BERT and RoBERTa.

For this exercise, we use the dataset from Gorodnichenko et al (2023) where each sentence is already manually labelled by several domain experts to ensure sufficient consistency. The dataset comprises historical FOMC statements spanning the period from 1997 to 2010. These statements are further split into 1,243 sentences, and then manually categorised into three groups: dovish, hawkish, or neutral.

For the experiment, we randomly select 80% of these sentences to fine-tune each CB-LM. We then assess the models' ability to predict the stances of the remaining 20% of the sentences, referred to as the test data. To maintain the representativeness of the full dataset, we ensure that the distribution of labels (hawkish, neutral and dovish) deviated by less than 5 percentage points between the full dataset and the test data. Among them, we identify the top 30 datasets that best preserved these label distributions, applying the training and test datasets consistently across all models used in this application.

We evaluate the out-of-sample performance of our CB-LMs based on the percentage of sentences correctly classified in the test data. We repeat this process using the selected 30 datasets to ensure robustness and reliability in our findings.

Figure 3 illustrates the average and median accuracy of various language models in classifying monetary policy stance. In this application, results for the RoBERTa- and BERT-based CB-LMs distinctly diverge. All RoBERTa-based models exhibit enhanced performance compared to their foundational model, with the models

RoBERTa+Paper+Speech [3] and RoBERTa+Paper showing statistically significant improvements. The top-performing CB-LM achieved a mean accuracy of approximately 84%, while their foundational model, RoBERTa, has a mean accuracy of 81%. As in the previous section, we observe a direct correlation between the amount of retraining data for domain adaptation and performance for RoBERTa-based CB-LMs. Those retrained with both the Paper and Speech datasets achieve the best results, followed by those retrained solely with the Paper, and finally those with only the Speech dataset.

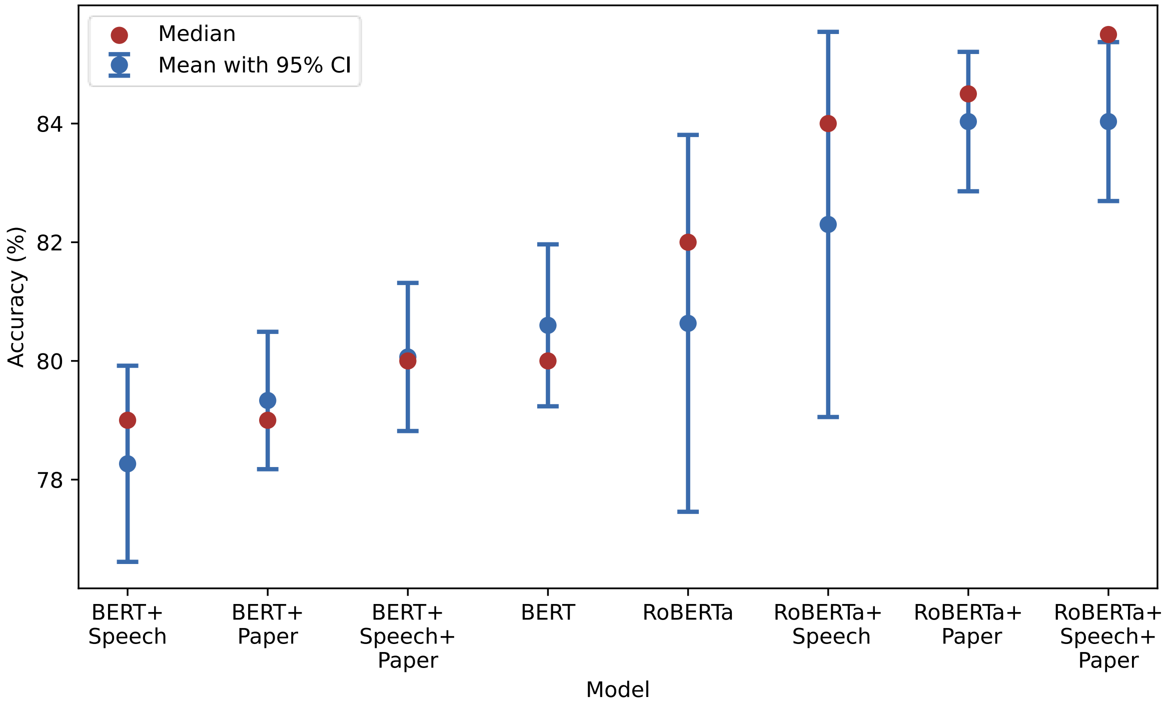

Figure 3. **Classifying monetary policy stance**. This figure reports the performance of CB-LMs alongside two foundation models in classifying the stance of FOMC statements. Sentences from the FOMC statements are manually labelled as Dovish, Hawkish or Neutral. The models are fine-tuned with 80% of these sentences and their corresponding manual labels. Then, we task the language models with predicting the monetary policy stance for the rest 20% sentences. The prediction from language models is considered as "correct" when it is consistent with the expert's manual label.

By contrast, the domain-adapted BERT models do not clearly exhibit improved performance relative to their foundation model. While domain adaptations generally intend to enhance the performance of LLMs, the results suggest that they do not guarantee performance improvements in every scenario. LLMs operate on probabilistic algorithms, meaning their

[3] When evaluating the monetary policy stance using RoBERTa+Paper+Speech, the sample median exceeds the upper bound of the confidence interval for the sample mean. This occurs because the distribution is highly negatively skewed, with a concentration of higher performance values.

performance can vary depending on the training data, as well as the model's specific parameters and architecture. This variability underscores that the same domain adaptation strategy might yield unexpected outcomes in certain models or contexts.[4]

Pinpointing the exact reason for this performance deterioration is difficult due to the complexity and often opaque nature of machine learning algorithms; nonetheless, overfitting is frequently identified as a contributing factor. Overfitting occurs when a model learns from the noise in the training data rather than the underlying signal, compromising its ability to generalise effectively to new data. This issue is particularly pronounced when the training data is smaller and lacks diversity, failing to represent the full spectrum of possible inputs (Kohavi and Sommerfield (1995), Dos Santos et al (2009), Charilaou and Battat (2022)). Such a scenario might explain the observed underperformance when integrating BERT, trained on a comparatively limited corpus. It is noteworthy that RoBERTa - with its additional 15 million parameters and training on a more extensive range of data - may be less susceptible to overfitting under similar conditions, due to its enhanced ability to generalise (Liu et al (2019)).

6. Generative LLMs vs CB-LMs

In this section, we evaluate the performance of state-of-the-art generative LLMs such as OpenAI’s ChatGPT (OpenAI (2023)), Meta’s Llama 3, Mistral (Jian et al (2023a)) and Mixtral (Jian et al (2023b)) in the context of central bank communication. For this purpose, we repeat the application in the previous section with these generative models that are known to excel in generating coherent and contextually appropriate text due to their sophisticated design and extensive pre-training on diverse text corpora. The generative LLMs are primarily decoder-only models designed to predict subsequent words in a

[4] The variability in training outcomes can even be influenced by different random seeds, which affect aspects such as weight initialization and the order in which training data is presented, introducing variance in the learning process (Dodge et al (2020)). This variability contributes to observed disparities in model performance, such as nonuniform improvement with increasing model size (Zhong et al (2021)), and a nonlinear relationship in the performance between upstream tasks, which involve general model training, and downstream tasks, which are specific applications of the trained model (Abnar et al (2021)).

sequence, thereby enabling them to generate text that closely mimics human behaviours. This capability renders them exceptionally adaptable to a variety of NLP applications.

The pre-training phase for these models involves exposure to vast datasets, incorporating potentially significant volumes of central banking-related content. The GPT architecture, which forms the foundation of ChatGPT, is known to have 175 billion parameters in its GPT-3.5 version and over 1 trillion parameters in GPT-4.[5] Similarly, Llama-3 versions are equipped with either 8 billion or 70 billion parameters, while Mistral and Mixtral have 7 billion and 47 billion parameters respectively. This broad knowledge base suggests a potential familiarity of the models with central banking context, reducing the necessity for domain-specific adaptation. However, to optimise performance for specialised tasks, targeted training is still required. We explore two approaches to this training: fine-tuning and in-context learning techniques.

For **fine-tuning**, we examine several methods such as Supervised Fine-Tuning (SFT; Ziegler et al (2019)), Direct Preference Optimization (DPO; Rafailov et al (2023)), and a proprietary technique developed by OpenAI for ChatGPT. SFT involves adjusting the model using a labelled dataset that is pertinent to the desired task. This process entails feeding the model a curated dataset where the input-output pairs are explicitly annotated, allowing the model to learn the specific patterns and nuances required for the task. By repeatedly updating the model's parameters based on this dataset, SFT aims to enhance the model's accuracy and reliability in generating relevant outputs.[6]

DPO, on the other hand, seeks to enhance model performance through reinforcement learning strategies directly targeted at the task. DPO uses a reward-based system to iteratively refine the model. The model generates outputs for given inputs, and these outputs are evaluated based on a predefined reward function that measures how well the outputs align with the desired outcomes. Positive feedback (rewards) is given for desirable outputs, and negative

[5] The specific details of GPT models such as GPT-3.5 and GPT-4 are proprietary and not fully disclosed by OpenAI.

[6] In our analysis, we adopt low-rank adaptation (LoRA; Yu et al (2023)), a fine-tuning method to improve training efficiency. LoRA reduces the number of trainable parameters by decomposing weight matrices into low-rank forms, thus maintaining performance while significantly lowering computational costs and memory requirements. This makes it particularly well-suited for fine-tuning large models with limited computational resources.

feedback (penalties) is given for undesirable ones. Over the training process, the model learns to optimise its performance by maximising the cumulative rewards, thereby improving its ability to handle the task at hand.

Additionally, we conduct **in-context learning**, particularly few-shot learning, to adapt the models to our specific requirements without extensive training. In-context learning involves the model using context provided in the input to understand and perform the task. Specifically, in few-shot learning, we present the model with a small set of task-related examples, known as "shots," within the input prompt. These examples demonstrate the desired output given particular inputs, effectively showing the model how to handle similar tasks. One of the key advantages of this technique is that it requires no modification to the core model architecture. Instead of updating the model's parameters, few-shot learning leverages the model's pre-existing capabilities and extensive knowledge base acquired from diverse training datasets. This makes it highly efficient, as it can be executed swiftly with limited computational resources compared to fine-tuning methods (BIS (2024)).

When selecting examples for the few-shot learning, we employ two strategies: random sampling and a retrieval-based method that prioritises examples most relevant to the target task. Random sampling involves the random selection of five examples from the training dataset without specific criteria, ensuring diversity. The retrieval-based method prioritises three examples that are most relevant to the target task by leveraging a vector database constructed from our training data. Each example in the training set is transformed into a high-dimensional vector representation using a pre-trained embedding model (Sentence-Transformers, Reimers and Gurevych (2019)), and these vectors are stored in the vector database. When presented with a new task, we convert the task description into an embedding using the Sentence-Transformers and perform a similarity search against the vector database using cosine similarity as the metric. This process retrieves examples with the highest cosine similarity to the target task, thereby selecting examples that are semantically closest.

Due to high computing demands, we train and test the generative LLMs using only a single random sample. For fine-tuning, we randomly select 80% of the data to train these models and evaluate their performance on the remaining 20% of data. The same test sample is used consistently across all generative LLMs.

Table 2 shows the performance of each model on monetary policy sentiment classification as discussed in Section 5. For comparison, we also include the foundation models. This comparative analysis underscores the potential of generative LLMs to adapt effectively to domain-specific challenges and highlights the strengths and limitations of each training method.

We find that fine-tuning generative LLMs generally enhances their performance. Without fine-tuning, the performance of generative LLMs tends to be lower than that of our foundational models for the CB-LMs. The "largest" fine-tuned models, including ChatGPT-3.5 and Llama-3 70B (4-bit), exhibit superior performance in this application. Notably, the fine-tuned ChatGPT-3.5 models achieve the highest accuracy rates, exceeding 85%. Additionally, Llama-3 70B (4-bit) models achieve an accuracy of 80-85% after fine-tuning using DPO. These results indicate significant promise compared to the best-performing CB-LM (RoBERTa + Paper & Speech), which averaged 83% accuracy.

Our findings align with the existing literature, which suggests that the performance of LLMs correlates with model size (Kaplan et al (2020)) and that performance improvements can emerge unexpectedly as size increases (Wei et al (2022)). For instance, models based on Llama-3 70B (4-bit) exhibit robust performance in our specific application, whereas their compact version of Llama-3 8B models perform significantly worse, achieving only 56% accuracy before fine-tuning and 63% afterwards. Interestingly, despite their larger size compared to the BERT or RoBERTa models discussed in the previous section, most generative LLMs exhibit considerably lower performance in accurately classifying monetary policy sentiments, even after fine-tuning.

Our analysis implies that generative LLMs may underperform in simple classification tasks compared to smaller, encoder-only models like BERT or RoBERTa. When examining only foundation models, all generative models underperform the encoder-only models, with accuracy ranging from 50% to 70% depending on the model. Even the ChatGPT-4 Turbo model, which is considered the best-performing LLM to date, achieves only 71% accuracy, lower than the average 80% accuracy of the RoBERTa foundation model.

This performance gap might stem from differences in model design and training objectives. Designed for text generation, generative models prioritise broad context and coherence over precise accuracy, which is crucial for classification. In contrast, encoder-only models like BERT and RoBERTa, are optimised for transforming input text into vector representations, or embeddings, that are well-suited for classification tasks.

Model	Type	Fine-tuning	In-context learning	Accuracy
ChatGPT-3.5 Turbo	Foundation	none	none	56%
ChatGPT-3.5 Turbo	Foundation	none	random	69%
ChatGPT-3.5 Turbo	Foundation	none	retrieval-based	71%
ChatGPT-3.5 Turbo	Fine-tuned	OpenAI	none	88%
ChatGPT-3.5 Turbo	Fine-tuned	OpenAI	random	87%
ChatGPT-3.5 Turbo	Fine-tuned	OpenAI	retrieval-based	85%
ChatGPT-4 Turbo	Foundation	none	none	71%
ChatGPT-4 Turbo	Foundation	none	random	73%
ChatGPT-4 Turbo	Foundation	none	retrieval-based	81%
Llama-3 8B	Foundation	none	none	56%
Llama-3 8B	Fine-tuned	DPO	none	63%
Llama-3 70B Instruct (4-bit)	Foundation	none	none	71%
Llama-3 70B Instruct (4-bit)	Foundation	none	random	38%
Llama-3 70B Instruct (4-bit)	Fine-tuned	DPO	none	73%
Llama-3 70B Instruct (4-bit)	Fine-tuned	DPO	random	73%
Llama-3 70B Instruct (4-bit)	Fine-tuned	DPO	retrieval-based	71%
Llama-3 70B Instruct (8-bit)	Foundation	none	none	72%
Llama-3 70B Instruct (8-bit)	Foundation	none	random	68%
Llama-3 70B Instruct (8-bit)	Foundation	none	retrieval-based	76%
Llama-3 70B (4-bit)	Fine-tuned	DPO	none	83%
Llama-3 70B (4-bit)	Fine-tuned	DPO	random	80%
Llama-3 70B (4-bit)	Fine-tuned	DPO	retrieval-based	85%
Mistral 7B	Foundation	none	none	50%
Mistral 7B	Foundation	none	retrieval-based	66%
Mistral 7B	Fine-tuned	SFT	none	38%
Mistral 7B	Fine-tuned	DPO	none	68%
Mistral 7B	Fine-tuned	DPO	retrieval-based	69%
Mixtral 7X8B	Foundation	none	none	60%
Mixtral 7X8B	Foundation	none	random	65%
Mixtral 7X8B	Foundation	none	retrieval-based	75%

Table 2. **Performance Comparison of Generative LLMs in Classifying Monetary Policy Stance.** This table provides a comparative analysis of the performance of various generative LLMs in accurately classifying the stance of monetary policy, a task described in Section 5. The models are assessed based on their ability to accurately classify the monetary policy stance of sentences from FOMC statements as Dovish, Hawkish or Neutral. The terms "4-bit" and "8-bit" in the model's name refer to quantization levels used to enhance computational efficiency, albeit with some potential trade-off in performance.

Additionally, the vast parameter space of generative models requires significant data for full fine-tuning. In this test, with limited training data and minimal layer updates during fine-tuning due to the high computational demands, there is a potential for overfitting. Although the exact cause is unclear, this might explain our observation that certain fine-tuning techniques or in-context learning strategies rather degrade performance, sometimes significantly. For instance, the performance of Mistral 7B decreases from 50% to 38% with the SFT method.[7] Similarly, the performance of Llama-3 70B Instruct (4-bit) drops from 71% to 38% following the random sampling in-context learning, which is significantly lower than the outcome of its 8-bit counterpart, despite both versions using the same training samples. Despite these challenges, it remains noteworthy that the "largest" generative LLMs such as ChatGPT and Llama-3 70B exhibit robust performance, especially after fine-tuning.

Training strategies also affect the performance of generative LLMs. Notably, as shown in Table 2, fine-tuning yields better performance than in-context learning when applied to ChatGPT-3.5 Turbo. Specifically, training ChatGPT-3.5 Turbo using in-context learning increases its performance from 56% to 69% with a random sampling method, and up to 71% with a retrieval-based method. By contrast, fine-tuning elevates the performance of ChatGPT-3.5 to as high as 88%.

Regarding in-context learning, it is noteworthy that the retrieval-based approach generally proves more effective than the random sampling approach. For example, with ChatGPT-4 Turbo, using the most similar examples improves accuracy to 81%, compared to 73% when randomly selected examples are used. This pattern is consistent with other models, such as Llama-3 70B Instruct (8-bit) and Mixtral 7X8B. Additionally, our tests with Mistral 7B models suggest that DPO is more effective than SFT in our experimental setup. However, we do not find conclusive evidence that combining fine-tuning with in-context learning further enhances performance.

To conclude, this section demonstrates that without fine-tuning, generative LLMs do not outperform smaller encoder-only models like BERT and RoBERTa, in classifying monetary

[7] Performance does not improve, even when varying the numbers of epochs.

policy sentiment at the sentence level. With fine-tuning, only the largest and most advanced models, ChatGPT 3.5 and Llama-3 70B, outperform BERT and RoBERTa.

7. Testing the limits of CB-LMs

In this section, we conduct a more challenging test to compare the performance of the top-performing CB-LMs identified in Section 5, specifically RoBERTa-based models, with leading state-of-the-art generative LLMs such as ChatGPT-3.5 Turbo, ChatGPT-4 Turbo, and Llama-3 70B. We assess the capabilities of these models in handling more complex tasks, thereby highlighting their potential applicability and effectiveness in various central banking contexts.

We incorporate two additional complexities to mirror complex real-life scenarios more accurately. First, we extend the analysis to longer texts. Analysing monetary policy stance involves more than just sentence-level sentiment classification in official monetary policy statements. A comprehensive understanding of monetary policy communications requires consideration of the entire text and its context, rather than merely analysing each sentence in isolation. Typically, market reactions are influenced by the overarching message or conclusions of speeches, statements, or news articles, rather than by sentiment expressed in individual sentences. Using longer texts introduces greater variability in language and context, complicating sentiment detection for language models. As the complexity of relationships between preceding and subsequent text segments increases, maintaining accuracy and consistency across extended text sequences becomes more challenging, especially as models must process long-range dependencies.

Second, we provide fewer training datasets than in the previous application. The process of conducting LLM-based analysis often faces challenges due to the need for high-quality labelled data to fine-tune LLMs for specific downstream tasks. This often manual process requires significant expertise and effort, frequently resulting in a scarcity of sufficient training data. This setup limits the models' ability to learn from diverse examples and generalise effectively.

These factors make the task more challenging than sentence-level classification with a large training dataset, where shorter texts and ample data provide clearer and more consistent signals for the models to learn from.

With these considerations in mind, we design an application aiming at analysing the monetary policy sentiment in key U.S. monetary policy news, predominantly speeches delivered by Federal Reserve governors. Similar to the previous application, the ability to anticipate market shifts in advance of policy communications is important for central banks to fine-tune communications, aiming to sidestep unintended market turbulence.

In pursuit of this aim, we leverage our internal daily newsletter dataset, which includes a highly selective compilation of U.S. monetary policy discussions from January 2015 to June 2023. These discussions are carefully identified and summarised by analysts from the BIS, making the best effort to ensure a clean, concise, and unbiased representation of key monetary policy messages. The documents are normally intended for a specialised audience and include brief direct quotes as well as succinct summaries of comments made by policymakers, without additional colour or unnecessary context. The documents contain five sentences and 67 words on average per news. From this dataset, we select and classify 237 news items deemed to articulate hawkish, dovish, or neutral views regarding upcoming policy rate decisions. To ensure the robustness of our classification process, we cross-check our classifications against corresponding market reactions on the date when the information was released. The market's reaction is measured through daily fluctuations in the estimated probability of a rate decision at the upcoming FOMC meeting, drawn from the futures forward-curve structure from Bloomberg.[8] To ensure balance in our analysis, the number of news items is quite evenly distributed across the three stance categories, with 89 hawkish, 83 neutral, and 65 dovish cases.

Like the previous application, we randomly select 80% of the events to fine-tune the RoBERTa-based CB-LMs and use the remaining 20% for out-of-sample predictions. To ensure robustness, this process is repeated 30 times. We also employ consistent strategies to

[8] See Appendix 2 for more details.

sample the training and test datasets, specifically selecting the top 30 datasets that best preserve label distributions of the full dataset.

For generative LLMs, including ChatGPT and Llama-3 70B, we apply a "zero-shot" learning approach. These models are tasked with classifying the entire dataset without any preliminary training process. Figure 4 shows the performance of most advanced CB-LMs within this application. The best performing RoBERTa-based models achieve an average accuracy of 58%, 64%, and 65%, the latter narrowly surpassing ChatGPT-3.5's performance of 64%.

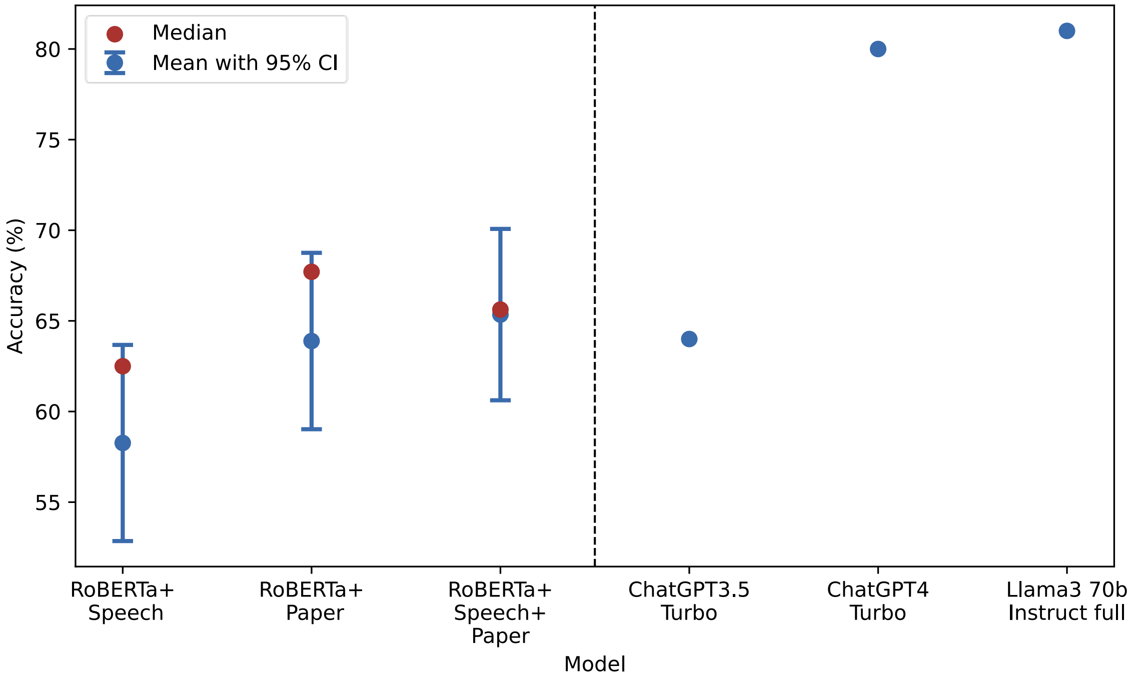

Figure 4. **Accuracy of CB-LMs in classifying the sentiment of monetary policy communications.** This figure reports the performance of our three CB-LMs alongside leading generative LLMs, in classifying the direction of expected rate decisions in response to US monetary policy discussions.

Meanwhile, ChatGPT-4 and Llama-3 70B Instruct models achieve superior performance, with an accuracy of 80% and 81% respectively, even without additional fine-tuning or in-context learning. This finding underscores that these models can inherently understand the nuances of monetary policy sentiments from their extensive pre-training, allowing them to adeptly apply this knowledge to new datasets (BIS (2024)). The advantage gained from their ability to process longer, context-rich texts, which aligns well with the general writing style of the test data, likely contributes to these outcomes. Furthermore, the extended context window of these models offers significant benefits in understanding and responding to complex and long text datasets.

8. Key Considerations for Using Generative LLMs for Central Banking

8.1 Adopting proprietary models

The test in Section 7 highlights the exceptional performance of the "largest" generative LLMs in complex NLP tasks. This illustrates how central banks can immediately begin to utilise such technologies in certain monetary policy contexts with minimal or no model training, significantly enhancing the usability of LLMs for central bankers. Particularly, the use of ChatGPT, which is often recognised as the best performing LLM, offers considerable convenience. This model requires no specialised hardware, and its services are highly accessible, even to those without advanced technical skills.

However, the use of such proprietary models in the context of central banking presents several challenges that have to be properly taken into account. The first major concern is confidentiality and privacy. When central banks send sensitive or confidential monetary policy-related information to external servers for training or application of these models, there is a risk that this data could be leaked or misused. Despite ongoing improvements in how proprietary models handle client data, transmitting sensitive information to an external server could violate internal information management policies or breach the terms of agreements with data providers, such as news media.

Recently, some cloud service providers began to offer services that allow proprietary LLMs to be hosted within a company's secure cloud environments. This enables the use of confidential or sensitive data with the models, potentially alleviating privacy and security concerns. However, this convenience comes with substantial cost implications. Integrating such a cloud environment requires significant investment.

Furthermore, whether operating LLMs within internal cloud environments or external servers, these models usually incur charges for both data input and output. Central banking tasks often require specific knowledge and in turn fine-tuning of the model, a process that can be data-intensive and, therefore, costly. Moreover, even after fine-tuning, accessing the custom model requires additional expenses. The cost can be considerable, particularly for tasks involving large amounts of input data or repetitive processes. This can make the use of a proprietary LLM less cost-effective in the long run.

Another critical issue is the lack of transparency and replicability. Proprietary models, hosted on external servers and controlled by a private company, create challenges in ensuring consistency and reliability. LLMs, by their nature, generate outputs based on probabilistic algorithms, often yielding varied results for identical requests. This characteristic, unless explicitly addressed with features for replicability as seen in models like BERT and RoBERTa, raises concerns about the models' dependability. Such variability is particularly problematic when these models are employed in critical domains like policy analysis and decision-making.

In a similar vein, the opacity of proprietary models is another limitation because users lack direct access to the model's architecture and parameters. Furthermore, any changes made by the model's provider can force users to adapt their processes, potentially leading to backward compatibility issues. In addition, if technical or performance issues arise on the external server, users have no control or means to directly address these problems.

To sum up, while the benefits of high-performance, proprietary LLMs are clear, central banks need to carefully consider the potential risks and costs. Issues related to operational security, scientific integrity, and financial impact are critical considerations in this assessment.

8.2 Adopting open-source models

As an alternative to proprietary LLMs, high-performance open-source LLMs may be considered for central banking applications. As shown in the previous two application cases, Llama-3 70B, an open-source LLM, has shown performance comparable to that of the proprietary ChatGPT-4. However, deploying such open-source models introduces a different set of challenges.

First, our findings indicate that only the largest models consistently demonstrate robust performance in the central banking applications. There exists a notable performance gap both across different models and among various sizes of the same model, as summarised in Table 2. Utilising these large models, such as Llama-3 70B, necessitates substantial computational resources, particularly in terms of GPU capabilities. Additionally, providing high-capacity GPUs to end-users within an organisation requires further resources. This includes building an internal cloud infrastructure that is strong and stable enough to handle the intensive computational demands of these models.

To circumvent significant upfront investments, some may opt for conducting LLM training and processing on a public cloud computing service. However, this approach can introduce issues similar to those encountered with proprietary LLMs, particularly concerning confidentiality and privacy. Such an arrangement potentially exposes sensitive data to the security risks inherent in public cloud environments.

Another challenge is the ongoing management and customisation of open-source LLMs. Unlike proprietary models that often come with vendor support and regular updates, open-source models require in-house expertise for maintenance, updates, and customisation to specific central banking needs. For example, fine-tuning open-source LLMs for specific downstream tasks requires highly technical staff and may require continuous training and development programs.

Additionally, even when an open-source LLM is developed by a reputable large technology company, the operation and maintenance of the model often depend on solutions and updates provided by a diverse and global community of developers. This reliance can present challenges in maintaining consistency with internal IT compliance standards, which are critical in the highly secure environment of a central bank. The variability in community support and the need for frequent adaptation can complicate the integration of open-source LLMs into existing IT ecosystems, potentially leading to gaps in compliance and increased operational risk.

These aspects underscore the necessity for central banks to carefully evaluate not only the initial capabilities of open-source LLMs but also the long-term implications of their integration into their operational environment. Adequate planning for the management and customisation of these models is essential to ensure they meet their operational requirements.

9. Conclusions

This paper evaluates the efficacy of domain-adapted language models across various applications within the central banking domain. We introduce CB-LMs by adapting foundational encoder-only language models, such as BERT and RoBERTa, into domain-specific language models tailored for central banking tasks. This process involves extensive retraining of the foundational models with a rich corpus of central banking text datasets—

speeches, papers, and a combination of speeches and papers—enabling the CB-LMs to outperform their foundational counterparts in downstream tasks such as masked word prediction and monetary policy sentiment analysis.

We find that CB-LMs, particularly those based on the RoBERTa model, consistently deliver superior performance compared to foundational language models. This evidence implies that CB-LMs understand nuanced expressions of monetary policy, which is critical for the real-time analysis and decision-making processes of central banks.

By contrast, the state-of-the-art generative LLMs such as ChatGPT, Llama-3, Mistral and Mixtral models do not necessarily guarantee superior performance. Our experiments indicate that many of them fail to exhibit clear outperformance relative to smaller encoder-only models like BERT and RoBERTa, in specific applications like sentence-level monetary policy sentiment classification. However, when faced with more challenging scenarios — such as more limited data for fine-tuning and longer texts to analyse — the most advanced generative LLMs, like ChatGPT-4 and Llama-3 70B, may outperform the smaller models.

These findings underscore the need for strategic consideration in the application of language models within central banking. Smaller models may be preferable when training data is ample and task complexity is lower, while generative models excel in more complex contexts. Important factors in selecting language models for central banking also include considerations of confidentiality, privacy, transparency, replicability, cost-efficiency, and the infrastructure and skills required for development and maintenance.

In conclusion, the results of this study advocate for the strategic integration of CB-LMs into central banking analytical frameworks, providing a significant enhancement over traditional NLP models. While benefits are clear, the fast-evolving technologies and environments require continuous attention and investments, making it difficult for a single institution to keep up with the pace. In this context, central banks can significantly benefit by sharing development and application experiences, models, best practices and AI tools to foster a "community of practice" (BIS (2024)). The continued refinements and knowledge sharing promise to revolutionise the way central banks utilise language models to inform and guide monetary policy decisions.

References

Abnar, Samira, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. 2021. "Exploring the Limits of Large Scale Pre-Training." arXiv preprint arXiv:2110.02095.

Acosta, Miguel, and Ellen E. Meade. 2015. "Hanging on Every Word: Semantic Analysis of the FOMC’s Postmeeting Statement." FEDS Notes 2015-09-30, Board of Governors of the Federal Reserve System (U.S.).

Aruoba, Borağan S., Pablo Cuba-Borda, Mauricio Ulate, Frank Schorfheide, and Sergio Villalvazo. 2021. "The Macroeconomic Impact of the Fed's New Monetary Policy Framework." Journal of Monetary Economics 124: 1–26.

Aruoba, Borağan S., Pablo Cuba-Borda, and Frank Schorfheide. 2023. "Reading the Fed’s Mind: Detecting Policy Decisions from the Minutes of the FOMC." Journal of Monetary Economics.

Aruoba, Borağan S., and Thomas Drechsel. 2023. “Identifying monetary policy shocks: A natural language approach”. No. w32417. National Bureau of Economic Research.

Azzimonti, Marina. 2018. "Partisan Conflict and Private Investment." Journal of Monetary Economics 93: 114–131.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty." The Quarterly Journal of Economics 131, no. 4: 1593–1636.

Bank for International Settlements (BIS). 2024. "Artificial Intelligence and the Economy: Implications for Central Banks." BIS Annual Economic Report, Ch. III.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." J. Mach. Learn. Res. 3: 993–1022.

Boukus, Ellyn, and Joshua V Rosenberg. 2006. "The Information Content of FOMC Minutes." Available at SSRN 922312.

Caldara, Dario, and Matteo Iacoviello. 2022. "Measuring Geopolitical Risk." American Economic Review 112, no. 4: 1194–1225.

Charilaou, Paris, and Robert Battat. 2022. "Machine Learning Models and Over-Fitting Considerations." World Journal of Gastroenterology 28, no. 5: 605–607.

Curti, Filippo, and Sophia Kazinnik. 2023. "Let’s Face It: Quantifying the Impact of Nonverbal Communication in FOMC Press Conferences." Journal of Monetary Economics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Dodge, Jesse, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping." arXiv preprint arXiv:2002.06305.

Dos Santos, Eulanda M., Robert Sabourin, and Patrick Maupin. 2009. "Overfitting Cautious Selection of Classifier Ensembles with Genetic Algorithms." Information Fusion 10: 150–162.

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

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera. 2023. "The Voice of Monetary Policy." American Economic Review 113, no. 2: 548–584.

Hansen, Anne Lundgaard, and Sophia Kazinnik. 2023. "Can ChatGPT Decipher Fed Speak?" Available at SSRN: https://ssrn.com/abstract=4399406.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2017. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." The Quarterly Journal of Economics 133, no. 2: 801–870.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2020. "The Global Impact of Brexit Uncertainty." Tech. rep., National Bureau of Economic Research.

Hoberg, Gerard, and Gordon Phillips. 2010. "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis." The Review of Financial Studies 23, no. 10: 3773–3811.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-Based Network Industries and Endogenous Product Differentiation." Journal of Political Economy 124, no. 5: 1423–1465.

Huang, Allen H., Hui Wang, and Yi Yang. 2023. "Finbert: A Large Language Model for Extracting Information from Financial Text." Contemporary Accounting Research 40, no. 2: 806–841.

Husted, Lucas, John Rogers, and Bo Sun. 2020. "Monetary Policy Uncertainty." Journal of Monetary Economics 115: 20–36.

Jiang, A.Q., A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, L Renard Lavaud, M-A Lachaux, P Stock, T Le Scao, T Lavril, T Wang, T Lacroix, and W El Sayed. 2023. "Mistral 7B." arXiv:2310.06825 [cs.CL].

Jiang, A.Q., A Sablayrolles, A Roux, A Mensch, B Savary, C Bamford, D Singh Chaplot, D de las Casas, E Bou Hanna, F Bressand, G Lengyel, G Bour, G Lample, L Renard Lavaud, L Saulnier, M-A Lachaux, P Stock, S Subramanian, S Yang, S Antoniak, T Le Scao, T Gervet, T Lavril, T Wang, T Lacroix, and W El Sayed. 2024. "Mixtral of Experts." arXiv:2401.04088 [cs.LG].

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language." arXiv preprint arXiv:2001.08361.

Kohavi, Ron, and Dan Sommerfield. 1995. "Feature Subset Selection Using the Wrapper Model: Overfitting and Dynamic Search Space Topology." In Proceedings of the First International Conference on Knowledge Discovery and Data Mining, KDD'95, 192–197.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." Bioinformatics 36, no. 4: 1234–1240.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692

Nakamura, Emi, and Jón Steinsson. 2018. "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect." The Quarterly Journal of Economics 133, no. 3: 1283-1330.

OpenAI. 2023. "Gpt-4 Technical Report." arXiv preprint arXiv:2303.08774.

Pfeifer, Moritz, and Vincent P. Marohl. 2023. "Central Bank RoBERTa: A Fine-Tuned Large Language Model for Central Bank Communications." The Journal of Finance and Data Science 9: 100114.

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

Reimers, N., and I. Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." arXiv:1908.10084 [cs.CL].

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. 2022. "Emergent Abilities of Large Language Models." Transactions on Machine Learning Research (TMLR), arXiv:2206.07682 [cs.CL].

Yu, Y., C-H Huck Yang, J. Kolehmainen, P. G. Shivakumar, Y. Gu, S. Ryu, R. Ren, Q. Luo, A. Gourav, I-F Chen, Y-C Liu, T. Dinh, A. Gandhe, D. Filimonov, S. Ghosh, A. Stolcke, A. Rastow, and I. Bulyko. 2023. "Low-Rank Adaptation of Large Language Model Rescoring for Parameter-Efficient Speech Recognition." arXiv preprint, arXiv:2309.15223 [cs.CL].

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 [cs.CL].

Zhong, Ruiqi, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. 2021. "Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level." arXiv preprint arXiv:2105.06020.

Appendix 1: Masked idiom dataset

Masked idioms	Correct answers
Accommodative <mask> policy	monetary
Asset <mask> program	purchase
Balance <mask> payments	of
Bank <mask> International Settlements	for
Basel <mask> on Banking Supervision	Committee
Bretton <mask> system	Woods
Capital <mask> ratio	adequacy
Central <mask> (CCP)	counterparties
Central <mask> balance sheet	bank
Central <mask> digital currency	bank
Collateralized <mask> obligation	debt
Committee <mask> Payments and Market Infrastructure	on
Commodity <mask> index	price
Contractionary <mask> policy	monetary
Core <mask> price index	consumer
Countercyclical <mask> buffer	capital
Credit <mask> swap	default
Cross-Currency <mask> Swaps	Basis
Currency <mask> of reserves	composition
Decentralized <mask> (DeFi)	finance
Distributed <mask> technology	ledger
Domestic <mask> important bank	systemically
Effective <mask> bound	lower
Effective <mask> funds rate	federal
Effective <mask> rate	exchange
Efficient <mask> hypothesis	market
Emerging <mask> and developing economies	market
Emerging <mask> economies	market
European <mask> Bank	Central
Exchange <mask> pass-through	rate
Exchange <mask> regime	rate
Expansionary <mask> policy	monetary
Financial <mask> board	stability
Fixed <mask> rate	exchange
Floating <mask> rate	exchange
Foreign <mask> intervention	exchange
Foreign <mask> investment	direct
Foreign <mask> reserves	exchange
Global <mask> important banks	systemically
Gross <mask> debt	external
Gross <mask> product	domestic
Interbank <mask> rate	offered
Interest <mask> on deposit facility	rate
Interest <mask> on excess reserves	rate
Interest <mask> parity	rate
Interest <mask> risk	rate
Interest <mask> swap	rate
Interest <mask> targeting	rate
International <mask> Fund	Monetary
International <mask> of Insurance Supervisors	Association
International <mask> position	investment
Inverted <mask> curve	yield
Labor <mask> participation rate	force
Lender <mask> last resort	of
Liquidity <mask> test	stress
London <mask> Offered Rate	Interbank

Long-term <mask> operation	refinancing
Long-term <mask> rates	interest
Macroprudential <mask> measures	policy
Main <mask> operation	refinancing
Marginal <mask> facility	lending
Monetary <mask> committee	policy
Monetary <mask> framework	policy
Monetary <mask> stance	policy
Monetary <mask> transmission	policy
Natural <mask> of interest	rate
Natural <mask> of unemployment	rate
Negative <mask> rates	interest
Net <mask> debt	external
Neutral <mask> policy	monetary
Nominal <mask> rate	exchange
Non-accelerating <mask> rate of unemployment	inflation
Non-bank <mask> institution	financial
Non-bank <mask> intermediation	credit
Open <mask> operations	market
Overnight <mask> facility	deposit
Overnight <mask> swap	index
Pegged <mask> rate	exchange
Producer <mask> index	price
Purchasing <mask> parity	power
Real <mask> exchange rate	effective
Real <mask> rate	exchange
Safe <mask> assets	haven
Secured <mask> financing rate	overnight
Short-term <mask> rates	interest
Sovereign <mask> crisis	debt
Sovereign <mask> fund	wealth
Special <mask> Rights	Drawing
Special <mask> vehicle	purpose
Systemically <mask> financial institution	important
Targeted <mask> refinancing operations	longer-term
Terms <mask> trade	of
Tier <mask> capital	1
Tight <mask> policy	monetary
Too <mask> to fail	big
Unconventional <mask> policy	monetary
Value <mask> Risk	at
Velocity <mask> money	of
Yield <mask> control	curve
Zero <mask> bound	lower

Appendix 2: Description of data used to measure market reactions to US monetary policy news

Market reactions are measured through daily changes in the estimated probability of a rate decision at the upcoming FOMC meeting, drawn from the futures forward-curve structure from Bloomberg. This is done by taking the difference of futures-implied interest rates before and after an event day with central bank news and rescale it to the standard size of a central bank rate hike or cut (25 bps for the Federal Reserve).

In this way, we can determine the percentage of a hike or cut 'priced in' to markets following central bank news. A simple example is the following. Let us imagine that the current overnight interest rate is 1.00%. The overnight rate expected after the central bank meeting is 1.125%. The assumed rate move size of the central bank is 25 basis points. Here, we would subtract the current overnight rate from the rate expected after the meeting: 1.125%-1.00% = 12.5 basis points. We would divide this difference by the assumed move size: 12.5/25= 0.50. We could then report that 50% of a hike was priced in for that particular event.

Over the period analysed by this study, this indicator ranges from -376 to 368 (see Figure A1). Hawkish (dovish) news is defined as that news raising (lowering) the index by more than 1 percentage point. Neutral news is defined as those causing minimal changes to the index (ie below 1 percentage point).

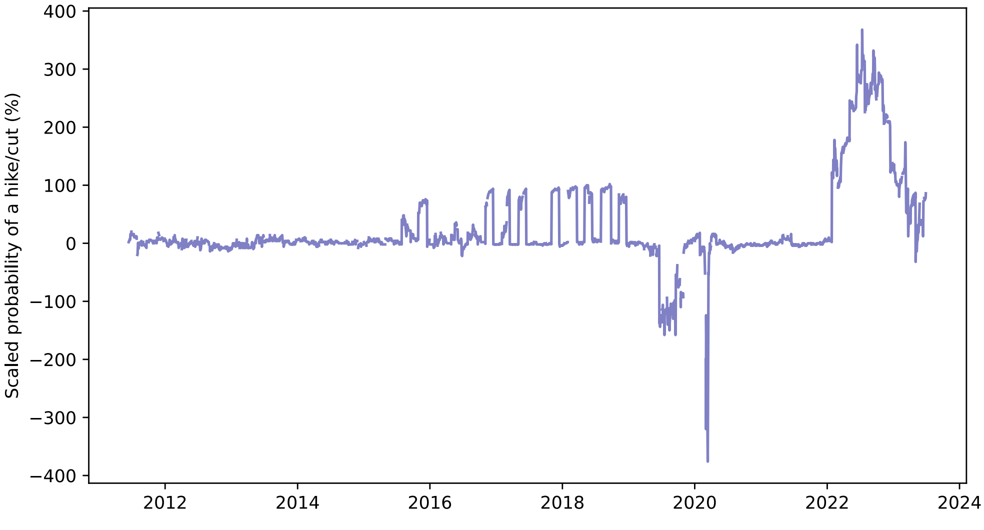

Figure A1. **Expected rate decisions at the upcoming FOMC meeting.** This figure shows the level of rate changes implied by Fed Funds Futures for the upcoming FOMC meeting. A level of 50% implies that markets expect a rate hike of 12.5 bps (ie 25 bps with a 50% of probability). Similarly, a level of 300% implies that markets expect a rate hike of 75 bps.

Previous volumes in this series

1214 September 2024	The impact of financial crises on industrial growth: lessons from the last 40 years	Carlos Madeira
1213 September 2024	Chinese Banks and their EMDE Borrowers: Have Their Relationships Changed in Times of Geoeconomic Fragmentation?	Catherine Casanova, Eugenio Cerutti, and Swapan-Kumar Pradhan
1212 September 2024	House price responses to monetary policy surprises: evidence from US listings data	Denis Gorea, Oleksiy Kryvtsov and Marianna Kudlyak
1211 September 2024	Non-bank lending and the transmission of monetary policy	Dominic Cucic and Denis Gorea
1210 September 2024	Which exchange rate matters to global investors?	Kristy Jansen, Hyun Song Shin and Goetz von Peter
1209 September 2024	Latin America’s non-linear response to pandemic inflation	Rafael Guerra, Steven Kamin, John Kearns, Christian Upper and Aatman Vakil
1208 September 2024	Generative AI and labour productivity: a field experiment on coding	Leonardo Gambacorta, Han Qiu, Shuo Shan and Daniel M Rees
1207 September 2024	The rise of generative AI: modelling exposure, substitution and inequality effects on the US labour market	Raphael Auer, David Köpfer, Josef Švéda
1206 August 2024	Covered interest parity: a forecasting approach to estimate the neutral band	Juan R. Hernández
1205 August 2024	The Measure Matters: Differences in the Passthrough of Inflation Expectations in Colombia	Andres Sanchez-Jabba and Erick Villabon-Hinestroza
1204 August 2024	Climate Policies, Labor Markets, and Macroeconomic Outcomes in Emerging Economies	Alan Finkelstein Shapiro and Victoria Nuguer
1203 August 2024	Strike while the Iron is Hot: Optimal Monetary Policy with a Nonlinear Phillips Curve	Peter Karadi, Anton Nakov, Galo Nuno, Ernesto Pasten, and Dominik Thaler
1202 August 2024	Are low interest rates firing back? Interest rate risk in the banking book and bank lending in a rising interest rate environment	Lara Coulier, Cosimo Pancaro and Alessio Reghezza
1201 July 2024	Crypto Exchange Tokens	Rodney Garratt, Maarten R.C. van Oordt

All volumes are available on our website www.bis.org.