DEUTSCHE BUNDESBANK EUROSYSTEM

Technical Paper

Monetary-Intelligent Language Agent (MILA)

01/2025

Felix Geiger
Dimitrios Kanelis
Philipp Lieberknecht
Diana Sola

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

Non-technical summary

Research Question:
Monetary policy communication has gained significant relevance during the last decades and is by now part of the established instrument kit used by central banks to steer expectations of market participants and the private sector. The difference to other monetary policy instruments like policy rates or asset purchase programs is that it conveys qualitative messages in natural language with explicit or implicit information about the intended monetary policy stance, future policy rates and the central bank's view of the macroeconomic outlook. This introduces new challenges for economic analysis and policy advice, as the textual nature of this communication requires a different set of methods and instruments for both quantitative and qualitative analysis.

Contribution:
We present the Monetary-Intelligent Language Agent (MILA), an innovative AI tool that uses advanced prompt engineering techniques and is based on large language models (LLMS). We design MILA to analyse monetary policy statements by the Governing Council of the European Central Bank (ECB) and speeches by members of its Executive Board. It decomposes these messages into smaller, understandable segments, classifies them, and aggregates the results to derive indicators. MILA offers three main advantages: First, it increases transparency by allowing users to attribute the classification results to specific text segments. Second, it simplifies the integration of expert knowledge and provides essential traceability of results under varying assumptions, making MILA highly adaptable. Third, it ensures a rigorous calculation of the indicators and avoids typical errors in linguistically generated numerical values. An extensive evaluation shows the advantages of our approach in terms of accuracy and consistency.

Results:
According to MILA, the ECB Governing Council's 2011-2024 communication was broadly in line with the macroeconomic environment in the euro area. Monetary policy communication was predominantly dovish, especially at the peak of the coronavirus pandemic in 2020. In 2021, the Governing Council balanced its inflation narrative, but communication on the interest rate path remained dovish. For the period of monetary policy tightening between 2022 and 2023, MILA is seeing a marked shift towards hawkish communication. In 2024, communication became more balanced.

Nichttechnische Zusammenfassung

Fragestellung:
Geldpolitische Kommunikation hat in den letzten Jahrzehnten zunehmend an Bedeutung gewonnen und dient den Zentralbanken als wichtiges Instrument, um die Erwartungen der Marktteilnehmer und des Privatsektors zu beeinflussen. Im Gegensatz zu traditionellen geldpolitischen Instrumenten wie Leitzinsen oder Anleihekaufprogrammen verwendet dieses Instrument natürliche Sprache, um qualitative Botschaften über den geldpolitischen Kurs, künftige Leitzinsen und wirtschaftliche Aussichten zu vermitteln. Dies bringt neue Herausforderungen für die wirtschaftliche Analyse und die Politikberatung mit sich, da der qualitative Charakter von Kommunikation andere Methoden und Instrumente für die quantitative und qualitative Analyse erfordert.

Beitrag:
Wir präsentieren den Monetary-Intelligent Language Agent (MILA), ein innovatives KI-Tool, das fortschrittliche Techniken des Prompt-Engineering nutzt und auf großen Sprachmodellen (LLMs) basiert. Wir konzipieren MILA, um geldpolitische Statements des Rates der Europäischen Zentralbank (EZB) und Reden ihrer Direktoriumsmitglieder zu analysieren. Es zerlegt diese Mitteilungen in kleinere, verständliche Segmente, klassifiziert sie und aggregiert die Ergebnisse, um Indikatoren zu erstellen. MILA bietet drei Hauptvorteile: Erstens, es erhöht die Transparenz, indem es den Nutzern ermöglicht, die Klassifikationsergebnisse auf spezifische Textsegmente zurückzuführen. Zweitens, es vereinfacht die Integration von Expertenwissen und bietet Nachvollziehbarkeit unter verschiedenen Annahmen, was es anpassungsfähig für unterschiedliche Analysen macht. Drittens, es gewährleistet eine rigorose Berechnung der Indikatoren und vermeidet typische Fehler sprachlich generierter numerischer Werte. Eine umfangreiche Evaluation zeigt die Vorteile unseres Ansatzes in Bezug auf Genauigkeit und Konsistenz.

Ergebnisse:
Laut MILA entwickelte sich die Kommunikation des EZB-Rats von 2011-2024 weitgehend im Einklang mit dem makroökonomischen Umfeld im Euroraum. Die geldpolitische Kommunikation war überwiegend „taubenhaft“, besonders zum Höhepunkt der Corona-Pandemie 2020. 2021 wurde das Inflationsnarrativ des EZB-Rats zunehmend balancierter, die Leitzinskommunikation blieb aber zunächst taubenhaft. Für die geldpolitische Straffungsphase zwischen 2022 und 2023 verzeichnet MILA eine deutliche Wende hin zu „falkenhaften“ Pressekonferenzen und Reden. 2024 wurde die Kommunikation wieder balancierter.

Monetary-Intelligent Language Agent (MILA)

Felix Geiger[1], Dimitrios Kanelis[1], Philipp Lieberknecht[1], Diana Sola[1]

April 2025

Abstract

Central bank communication has become a crucial tool for steering the monetary policy stance and shaping the outlook of market participants. Traditionally, analyzing central bank communication required substantial human effort, expertise, and resources, making the process time-consuming. The recent introduction of artificial intelligence (AI) methods has streamlined and enhanced this analysis. While fine-tuned language models show promise, their reliance on large annotated datasets is a limitation that the use of large language models (LLMs) combined with prompt engineering overcomes. This paper introduces the Monetary-Intelligent Language Agent (MILA), a novel framework that leverages advanced prompt engineering techniques and LLMs to analyze and measure different semantic dimensions of monetary policy communication. MILA performs granular classifications of central bank statements conditional on the macroeconomic context. This approach enhances transparency, integrates expert knowledge, and ensures rigorous statistical calculations. For illustration, we apply MILA to the European Central Bank's (ECB) monetary policy statements to derive sentiment and hawkometer indicators. Our findings reveal changes in the ECB's communication tone over time, reflecting economic conditions and policy adaptions, and demonstrate MILA's effectiveness in providing nuanced insights into central bank communication. A model evaluation of MILA shows high accuracy, flexibility, and strong consistency of the results despite the stochastic nature of language models.

JEL Classification: C45, E31, E44, E52, E58

Keywords: Central bank communication, monetary policy, sentiment analysis, artificial intelligence, large language models

[1]Deutsche Bundesbank, Wilhelm-Epstein-Str. 14, 60431 Frankfurt am Main, Germany

[‡]This paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

1. Introduction

Monetary policy communication has gained significant relevance during the last decades and is by now part of the established instrument kit used by central banks to steer expectations of market participants and the private sector. The difference to other monetary policy instruments like interest rates or asset purchase programs is that it conveys qualitative messages in natural language with explicit or implicit information about the intended monetary policy stance, future policy rates and the central bank's view of the macroeconomic outlook. This introduces new challenges for economic analysis and policy advice, as the textual nature of this communication requires a different set of methods and instruments for both quantitative and qualitative analysis.

Traditionally, the evaluation of textual data has relied on human expertise, where various features of a text are identified through manual analysis (Ehrmann and Fratzscher, 2007). Recent studies, such as Pfeifer and Marohl (2023) and Gambacorta et al. (2024), have demonstrated the use of fine-tuned language models to analyze central bank communication. However, fine-tuning requires the creation of extensive datasets for training, a time-consuming task that also carries the risk of error-prone human labeling. Additionally, as we demonstrate, using available off-the-shelf data can lead to misleading results. The emergence of powerful large language models (LLMs) has driven a paradigm shift from fine-tuning to prompt engineering. Against this background, we introduce our novel Monetary-Intelligent Language Agent (MILA), which uses advanced prompt engineering techniques and combines them with insights from the comprehensive literature on central bank communication. In Deutsche Bundesbank (2025), we use MILA as our novel AI tool to perform a detailed analysis on the monetary policy statements (MPS) and speeches of the Executive Board of the European Central Bank (ECB) to measure various semantic dimensions of euro area monetary policy communication.[1]

We develop MILA as a framework based on LLMs that performs granular classifications, which are then aggregated via statistical methods to derive a comprehensive indicator. For the granular analysis, we break down central bank communication into text segments or categories that are easily understood by humans. We then classify these text segments using an advanced prompting strategy, while considering the macroeconomic background and document context, which ensures that the overall context is retained during the granular analysis. Subsequently, we aggregate the individual results into a document-level indicator using mathematical calculations.

Our approach has three major advantages. First, analyzing communication at a granular level increases the transparency of the classification results. Breaking texts into smaller, interpretable units allows users to trace decisions back to specific inputs and clarify how each text segment contributes to the indicator score. Furthermore, our approach leverages the LLM not only for classifications, but also to generate explanations for its decisions. While this does not align

[1] Until the strategic review in 2021, the official name of the MPS was "Introductory Statement." During the review, the name was changed, and several adjustments were made to the document. For simplicity, we will refer to the document as MPS, while acknowledging these changes.

with the traditional definition of Explainable AI, it effectively enables users to verify results and conduct plausibility checks. Second, a granular analysis simplifies the complementary use of expert knowledge and provides essential traceability of results under varying assumptions, making MILA highly adaptable. This also supports counterfactual analyses, helping to assess how communication may be perceived under different macroeconomic assumptions. Third, aggregating granular results into a document-level indicator through statistical and mathematical calculations ensures that the indicator is rigorously calculated rather than derived from the next-word prediction probabilities of LLMs. This approach avoids the pitfalls of linguistic generation, such as sensitivity to input variations – especially in longer texts – and the risk of language model hallucinations.

The framework that we introduce is independent of the employed LLM such that MILA can execute the same analysis with different models, which enables a comparison of different LLM performances. The results presented in this paper are derived by using Llama 3.1 70B (Grattafiori et al., 2024). We focus on the MPS since November 2011 and derive a variety of indicators about the tone of the statement based on a Positive/Negative metric (*Sentiment*) and a Hawkish/Dovish metric (*Hawk-O-Meter*) showcasing MILA's capabilities. These indicators measure the amount of positive and negative information that the statement contains on the economic outlook as well as explicit and implicit information on the future stance of monetary policy.

We assess MILA's classification performance in an evaluation. A general gap in the literature is the lack of an extensive annotated dataset on the communication of euro area central bank officials that researchers and practitioners can use as a reliable reference for model evaluation. Therefore, we rely on available off-the-shelf datasets and demonstrate the flexibility and accuracy of MILA in out-of-sample evaluations. During the evaluation, we emphasize some weaknesses of existing approaches and recommend the use of LLMs to assist in the future creation of annotated datasets for fine-tuning smaller models or model evaluation. Finally, we show the superiority of our granular analysis approach in terms of replicability compared to direct full document analyses. Despite the stochastic nature of language models, MILA provides high consistency of results, making it a reliable and promising tool for monetary policy analysis and advice.

The remainder of this paper is structured as follows: In Section 2, we provide an overview of the methodology and training process of the language models that MILA is built on, along with our prompt engineering approach. In Section 3, we introduce a novel methodology to derive sentiment indicators from the communications of ECB officials, measuring the economic optimism and pessimism in their messages and expectations. In Section 4, we present a new methodology to derive the Hawk-O-Meter, which measures the hawkishness or dovishness of monetary policy announcements regarding the interest rate path and the interpretation and narrative of economic developments. In Section 5, we demonstrate the replicability of our results and evaluate MILA using labeled datasets from the literature, while emphasizing several methodological weaknesses of established approaches. The final section concludes.

2. Methodology: The Artificial Intelligence of MILA

MILA is an LLM-based framework that conducts granular classifications, which are subsequently aggregated through statistical methods to generate a comprehensive indicator. It studies two different dimensions: Sentiment Analysis and Hawk-Dove-Analysis of the Monetary Policy Statement, which we discuss in more detail in Sections 3 and 4, respectively. Within the Hawk-Dove-Analysis, we distinguish between the Decision and the Narrative Hawk-O-Meter.

The framework operates at different levels of granularity – sentence, paragraph, or category level – depending on the analysis performed. At the sentence level, for instance, classifications are made on a sentence-by-sentence basis and subsequently aggregated into a comprehensive indicator for the document. To enhance the accuracy of granular classifications, MILA incorporates various contextual elements, such as preceding text segments or the inflation context, tailored to the specific type of analysis conducted. Beyond its granularity-based approach, MILA integrates additional features to enhance transparency. Depending on the analysis type, it either lists all relevant text segments contributing to a classification decision or generates a concise one-sentence explanation to justify the classification outcome.

At the core of our classification methodology is a multi-layered prompting strategy, incorporating three key principles from the literature on prompt engineering: role-based prompting, prompt chaining, and few-shot prompting. First, we implement role-based prompting to guide the LLM's response generation by assigning it a specific identity: *"You are an economist specialized in monetary policy analysis at a central bank."* This role description is placed at the beginning of every system prompt to establish a consistent expert perspective. Second, we employ a prompt chaining strategy to distribute different steps of the classification process across multiple prompts, breaking down complex tasks into individual steps. For example, in the first step, a language model determines whether communication is hawkish or dovish. In the subsequent step, it evaluates whether the statement is strong or moderate. This approach reduces the risk of hallucinations by avoiding the pitfalls of requiring the model to generate the entire output in a single request. Our prompt chaining method is inspired by Wu et al. (2022) and Sun et al. (2024), who demonstrate that decomposing tasks into a sequence of steps improves accuracy, increases transparency, and facilitates the identification of mistakes or inefficiencies. Third, we apply few-shot prompting, providing three to four carefully curated examples for each step within each prompt chain for every indicator. This approach enhances the model's ability to deal with more challenging examples and understand our preferred output format.

MILA is designed to be independent of the specific LLM employed, allowing it to perform the same analysis with different underlying models. However, different LLMs may need adjusted prompts for the best results. The results presented in this paper are based on Llama 3.1 70B, which is one of the most advanced language models currently available.[2] The training process of Llama consists of two phases. In the first phase, pre-training, the model is trained using next-

[2]We use Llama 3.1 70B on-premise in a private cloud operated by the Deutsche Bundesbank.

token prediction to learn the underlying structure of textual data. For this purpose, an extremely large text corpus is utilized. The goal of this step is to enable the model to predict tokens that appropriately continue a given text. At this stage, however, the model is not yet capable of handling complex instructions or specific tasks. In the second phase of Llama's training process, post-training, the model is fine-tuned to understand and execute instructions accurately. This process involves multiple rounds of fine-tuning. The first step is Supervised Fine-Tuning (SFT), where the model is trained on curated datasets containing input-output pairs provided by human experts. This allows the model to learn how to respond appropriately to specific queries and tasks. Building on this, Direct Preference Optimization (DPO) is applied, which is a simpler alternative to Reinforcement Learning from Human Feedback (RLHF) for aligning language models to human preferences. During this step, the model is trained to prioritize responses that human evaluators have rated as high-quality. This approach enhances the relevance and usability of the model's outputs, making it exceptionally effective at handling complex and diverse tasks.[3]

3. Sentiment Analysis of the Monetary Policy Statement

We create an indicator to gauge the level of positivity (or optimism) within the MPS. MILA is tasked with classifying the MPS of interest in an interval between +1 (*very positive*) and -1 (*very negative*). Additionally, the classification should treat announcements and decisions related to monetary policy as neutral to avoid any possible bias, as central banks invariably present new decisions as the right choice given the prevailing macroeconomic circumstance.[4]

Before conducting the sentiment classification, it is crucial to first extract information on the current inflation dynamics and outlook to ensure a state-dependent analysis. Policymakers generally view a decline in economic growth or a rise in unemployment as negative. However, the classification of price level changes depends on the prevailing inflation gap. When the inflation rate is below the central bank's target, increases towards the target are considered positive. Conversely, increases that exceed the inflation target in the medium term are viewed as negative developments. MILA extracts the relevant inflation context by examining the document's date and retrieving the corresponding ECB/Eurosystem staff forecast information from the associated MPS.[5] The information extraction is performed via an LLM request, which generates a standardized summary of the current and expected inflation trends used in subsequent steps

[3]Grattafiori et al. (2024) provide a detailed description of Llama 3.

[4]A central bank inherently possesses a degree of subjective bias when communicating monetary policy decisions. From a communicative standpoint, it is inconceivable to present changes in monetary policy by highlighting expectations of negligible or adverse effects. Consequently, monetary policy decisions are either presented without additional framing or are cast in a positive light. Alternatively, implicit positive or negative connotations can be inferred directly from the decisions themselves. Nevertheless, we consider this aspect as a potential enhancement to our baseline model.

[5]In our baseline model, MILA derives inflation expectations from the ECB/Eurosystem staff forecast. However, this could be replaced with different forecasts, including those with higher frequency, such as inflation expectations from option prices. This would allow for the analysis of the tone of the statement to be conditional on varying inflation expectations.

of our prompt chain. By incorporating this inflation context, the model is able to understand current and expected inflation dynamics in relation to the 2%-inflation target. In addition to inflation, we inform the model that signs of economic growth, low or decreasing unemployment, as well as improvements in financial and banking stability are generally positive developments. Conversely, signs of economic downturn, a weak labor market or rising unemployment, as well as financial and banking instability are considered negative.

For the sentiment analysis of the MPS, we derive an indicator from the most transparent textual level by conducting a granular analysis at the sentence level. Dividing the text into individual sentences for the indicator derivation enhances transparency, as each sentence represents a meaningful unit that a human reader can easily interpret and evaluate (Maibaum et al., 2024). This applies especially to central bank communication (Kanelis and Siklos, 2025) where even individual sentences can have a strong impact. Moreover, assigning ground truth labels to individual sentences allows for better interpretability of classification results compared to analyzing the entire document at once.

As contextual elements, we use the inflation context available at the release date, extracted as previously described, along with a predetermined number of preceding sentences to consider the broader document context. In the further analysis, we set this number to one, meaning MILA always considers the preceding sentence $k-1$ when classifying sentence $k \in \{1, ..., K\}$, where K is the total number of sentences in the MPS of date t.[6] For additional transparency, we require the model to generate a one-sentence explanation for each sentence classification before making a decision. This approach also helps to identify potential hallucinations and increases the likelihood of correct classification by requiring the model to justify its decision first.[7]

MILA classifies each sentence as either *Positive*, *Neutral*, or *Negative*. For example, MILA classifies the sentence *"After decreasing by 3.6%, quarter on quarter, in the first quarter of 2020, euro area real GDP is expected to have contracted even further overall in the second quarter, broadly in line with the June 2020 Eurosystem staff macroeconomic projections."* from the MPS in July 2020 as *Negative* with *"The sentence is classified as negative because it indicates a contraction in euro area real GDP, reflecting worsening economic conditions."* as the corresponding classification reasoning. As context for the sentiment classification of this sentence, MILA uses the inflation context *"The expected medium-term inflation rate for the immediate years ahead, starting with the specified year, is below the ECB's target. Consequently, increases in inflation are viewed as positive."* and the previous sentence in the document *"At the same time, economic indicators remain well below the levels recorded before the pandemic, and the recovery is in its early stages and remains uneven across sectors and jurisdictions."*

Having obtained a detailed classification at the sentence level, MILA can apply different calcula-

[6]For the first sentence of a document MILA does not consider the preceding sentences.

[7]This results from the next-token prediction logic of language models, which operate by predicting the probability distribution of the next word in a sequence given the preceding context.

tions or statistical analyses without the necessity to further utilize language modeling.[8] Following the literature on central bank communication (Kanelis and Siklos, 2025), we use equation (1) to calculate the net sentiment of the MPS of the press conference at t:[9]

$$Sentiment_t = \frac{\#PosSent_t - \#NegSent_t}{\#PosSent_t + \#NegSent_t} \tag{1}$$

We calculate the sentiment of the MPS at t by subtracting the number of negative sentences from the number of positive ones and then dividing by the sum of positive and negative sentences. This continuous indicator takes a value between -1 and 1. The value is 1 (-1) if all sentences with a non-neutral sentiment are positive (negative), and 0 if the text features an equal amount of positive and negative sentences.[10] Figure (1) displays the sentiment analysis for all MPS since November 2011[11], operating under the assumption that each sentence is accorded equal importance.[12]

According to the artifical intelligence assessment, the sentiment in ECB press conferences has been predominantly positive since 2011. From the reduction of the deposit rate to 0% in July 2012 until the end of 2013, MILA records a noticeably negative tone. This primarily reflects a pessimistic narrative about economic dynamics in the context of the European sovereign debt crisis. From 2014 until the end of Mario Draghi's tenure as ECB President in November 2019, the sentiment was largely optimistic. The subsequent COVID-19 pandemic was also reflected in ECB press conferences: while the lockdowns in spring 2020 and winter 2020/2021 were accompanied by negative sentiment, the increasing distribution of vaccines and the reopening of the economy in 2021 led to optimism. MILA recorded another downturn in sentiment around the time of the Russian invasion of Ukraine and the sharp rise in inflation in early 2022. Since the last interest rate hike in September 2023, the sentiment has gradually become more positive and can be considered broadly balanced at the end of 2024.[13]

[8] Given that language models are primarily designed for text generation, it is prudent to refrain from using them for mathematical computations.

[9] Once all sentences are classified, we can apply various statistical techniques for evaluation. It would, for example, be possible to assign varying weights to sentences based on their importance. For illustration purposes, however, we will focus on a single aggregation technique that possesses linear properties and is intuitive.

[10] We use this formula to calculate the sentiment even though it disregards the document length, due to the similar lengths of the MPS. This approach contrasts, for example, with speeches.

[11] In November 2011, Mario Draghi assumed the role of President of the ECB.

[12] Alternatively, we can focus on sentences related to specific topics or assign weights based on particular heuristics or their importance.

[13] For a more in-depth analysis, see Deutsche Bundesbank (2025).

4. Hawk-Dove-Analysis of the Monetary Policy Statement

In this section, we explain how MILA assesses the hawkishness of monetary policy communication through both explicit and implicit signals regarding interest rates. *Explicit signals* encompass communications and commitments that provide market participants with clear information about the interest rate trajectory, such as forward guidance. *Implicit signals* involve descriptions and interpretations of economic developments and inflation, offering market participants indirect insights into likely monetary policy actions. We investigate the hawkishness of monetary policy communication using our Hawk-O-Meter, which is based on Hawk-Dove metrics from the literature (Apel and Blix Grimaldi, 2014) and evaluates different communication characteristics compared to the Positive-Negative metric (Picault and Renault (2017); Pfeifer and Marohl (2023)).[14] Within the Hawk-Dove-analysis, we differentiate between the Decision Hawk-O-Meter and the Narrative Hawk-O-Meter, which we describe in detail below.

4.1. Introducing the Decision Hawk-O-Meter

The initial paragraphs of the MPS contain the monetary policy decisions, including communication about the future trajectory of interest rates. This section of the statement thereby provides insights into the ECB's expectations and intentions regarding the development of interest rates and typically concludes with transitions such as *"Let me now explain our assessment in greater detail"* or *"I will now outline in more detail..."*. MILA's Decision Hawk-O-Meter quantifies the ECB's monetary policy stance by analyzing the interest rate communication in these initial paragraphs, classifying the decisions on a scale from -1 (*Very Dovish*) to +1 (*Very Hawkish*).

The Decision Hawk-O-Meter derives its indicator through a granular analysis at both the paragraph and the category level. Unlike other analyses in this work, focusing on individual sentences for the Decision Hawk-O-Meter may be misleading, as signaling interest rate changes can depend on altering just a few key words. Consequently, instead of deriving the indicator at the sentence level – the most granular approach – we shift to the paragraph level to assess the communicative content.[15] Deriving a quantitative value on a scale from -1 (*Very Dovish*) to +1 (*Very Hawkish*) is challenging, even for a language model, given the inherently qualitative nature of central bank communication. Therefore, we define four categories and design a score-based approach to characterize the monetary policy decisions as communicated by the president:

- **Interest Rate Decision Score:** Are the key interest rates lowered, raised, or maintained, and if changed, by how many basis points? (**Score:** +0.3, +0.15, 0, -0.15, -0.3)

[14]The Hawk-Dove metric is most suitable for central bank communication on monetary policy and the economy, while its application to central bank communication with a different thematic focuses can lead to measurement bias.

[15]It is possible to remain at the sentence level by providing high weights to sentences containing information on monetary policy decisions or forward guidance. However, the paragraph level of the document is more suitable if specific sentences are dominating.

- **Interest Rate Outlook Score:** Are signals for future reductions or increases communicated? Is the language cautious, neutral, data-dependent, or clearly directive, or is it subject to directional distortion? (**Score:** +0.3, +0.15, 0, -0.15, -0.3)
- **Inflation Score:** Are the decisions justified by a direct reference to inflation developments, and is there an indication of upward or downward risks? (**Score:** +0.2, +0.1, 0, -0.1, -0.2)
- **General Tone Score:** What is the tone of the language, and what other announcements are made with implications for the interest rate path? Is the language particularly cautious, supportive, expansionary, or markedly restrictive? (**Score:** +0.2, +0.1, 0, -0.1, -0.2)

Each category is analyzed individually by an LLM request to ensure a comprehensive understanding of the final score. Unlike the previously introduced vertical prompt chain – where the output of one LLM serves as input for the next – we now employ a horizontal prompt chain, where an LLM evaluates each category independently.

To further enhance transparency, MILA identifies and lists all relevant sentences within the paragraphs that contribute to each category score. This process effectively allows MILA to *find the needle in the haystack*, leveraging a capability refined in Llama 3.1 70B during post-training (Grattafiori et al., 2024). Once relevant sentences are identified, MILA awards a score for each category based on specific criteria.[16] To obtain the final Decision Hawk-O-Meter score measuring the degree of hawkishness of the announced monetary policy decisions, we sum the individual category scores.

In figure (2), we visualize our Decision Hawk-O-Meter for the presidency of Christine Lagarde, that is, from December 2019 until December 2024.[17] According to MILA, the ECB's communication regarding its monetary policy decisions was highly dovish throughout the first 1.5 years of the COVID-19 pandemic. Although it became gradually less dovish over the course of 2021, the communication still emphasized the need for a prolonged accommodative monetary policy stance as of December 2021. At the beginning of 2022, with incoming inflation data beginning to cast doubt on the narrative of temporary inflationary pressures, the communication regarding monetary policy decisions shifted towards a more balanced tone. The Russian invasion of Ukraine, which caused substantial upside inflation risks through supply bottlenecks and rising energy prices, led to an abrupt shift towards a more hawkish stance, particularly at the onset of the tightening cycle with the first policy rate hike in over a decade and the exit from negative interest rate territory in July 2022. The indicator records the most hawkish tone in late 2022 and early 2023, when the ECB raised policy rates by 50 basis points or more at each meeting. Since then, communication has become less hawkish over the course of 2023. However, it continued to emphasize the need for restrictive policy rates for approximately a year following the

[16] We provide the criteria for each category in Appendix A.

[17] We did not create an interest rate indicator for Mario Draghi's presidency because short-term interest rates were largely at the effective lower bound, coupled with a clear downward bias.

final interest rate hike in September 2023. Since September 2024, MILA has observed a more balanced and, most recently, a dovish tone.[18]

4.2. Introducing the Narrative Hawk-O-Meter

In the previous section, we derived a Hawk-O-Meter solely from the monetary policy decisions and forward guidance communicated by the ECB president in the first part of the MPS. MILA's Narrative Hawk-O-Meter focuses on the second part of the statement, where the ECB president elaborates on the analysis conducted by the ECB governing council and provides insights into the Eurosystem's expectations regarding inflation and economic development, based on both real analysis and the monetary and financial analysis pillars. Typically, the second part of the statement begins with phrases such as *"Let me now explain our assessment in greater detail"* or *"I will now outline in more detail..."*. Focusing on the central bank's communication and narrative regarding inflation and the general economy, rather than the actual decisions, is a well-established approach in the literature on central bank communication (Apel et al., 2022). This is because interest rate decisions are ultimately an endogenous outcome of the macroeconomy. In other words, from the perspective of a Taylor rule (Taylor, 1993) this indicator emphasizes communication about macroeconomic parameters rather than the interest rate variable itself.

For the Narrative Hawk-O-Meter, we return to a granular analysis at the sentence level, classifying each sentence into one of five categories: *Hawkish, Moderate Hawkish, Neutral, Moderate Dovish, Dovish.* The sentence classification follows a two-step approach. In the first step of the classification chain, MILA categorizes each sentence as *Hawkish, Neutral* or *Dovish.* We instruct MILA to classify sentences as *Hawkish* if they provide information that suggests an increased likelihood of monetary tightening or implicitly imply such a stance. This includes statements on increasing inflation or inflation pressure, increases in purchasing power, adverse supply shocks, low unemployment and higher wage growth, accelerating economic activity and a strong or even unsustainable money or credit growth. Furthermore, we prompt MILA to classify critical statements on public debt or support for inflation-reducing economic policies as hawkish.[19] Reverse communication, such as references to falling inflation rates, rising unemployment, or weak economic activity, are correspondingly classified as dovish. Additionally, calls for fiscal policy stimulus are also considered dovish. Neutral statements are those without economic reference or without implications for monetary policy.

Once a sentence is classified as *Hawkish* (*Dovish*), MILA proceeds to the second step in the prompt chain to determine whether the statement should be categorized as *Moderate Hawkish (Moderate Dovish)* or *Hawkish (Dovish).* A key criterion in this refinement step is whether the statement emphasizes upward (downward) risks to inflation or provides clear justifications

[18]For a more in-depth analysis, see Deutsche Bundesbank (2025).

[19]This instruction on the evaluation of fiscal policy is based on the metric of fiscal hawkishness or dovishness. According to this metric, a communicated preference for increased debt-financed government spending is considered dovish, while a reduction in the deficit or avoidance of public debt is considered hawkish. The resulting effect on monetary policy, however, is not necessarily clear.

for monetary policy decisions. This prompt-chaining approach facilitates the more challenging qualitative distinction between moderate and strong classifications.

MILA leverages different contextual information for the Narrative Hawk-O-Meter. Similar to the Sentiment Analysis described in Section 3, it incorporates the inflation context derived from ECB/Eurosystem staff projections and uses the previous sentence as document context. In addition, the Narrative Hawk-O-Meter uses the monetary policy decisions announced during the same press conferences as context. For enhanced transparency, MILA generates a one-sentence explanation for the sentence classifications.

After classifying all sentences into one of five categories, we derive the overall narrative Hawk-O-Meter for the MPS using equation (2), which is similar to the sentiment indicator but modified to incorporate the moderate categories.

$$Hawkometer_t = \frac{\#HawkSent_t + 0.5 * (\#ModHawkSent_t - \#ModDoveSent_t) - \#DoveSent_t}{\#HawkSent_t + \#ModHawkSent_t + \#ModDoveSent_t + \#DoveSent_t} \quad (2)$$

Furthermore, we refine our analysis to exclusively consider sentences that 1) relate to inflation, termed as *Inflation Hawk-O-Meter*, or 2) relate to economic growth and employment, referred to as *Real economy Hawk-O-Meter*. In this case, we use equation (2), but include also the neutral sentences in the denominator to calculate sentiment indicators for both inflation and the real economy.[20]

The upper panel of figure (3) displays the Hawk-O-Meter analysis of the entire MPS since November 2011, and the lower panel illustrates topic-based Hawk-O-Meter indicators (constructed similarly to the topic-based sentiment indicators). According to artificial intelligence, the economic narrative of the ECB governing council has been predominantly dovish since the end of 2011. Significant ECB governing council meetings during Mario Draghi's presidency are assessed as particularly dovish, such as in August 2012 (following the *"Whatever it takes"* statement), June 2014 (reduction of the deposit facility rate into negative territory), January 2015 (start of the Asset Purchase Program - APP), and March 2016 (increase of the APP purchase volume from 60 to 80 billion per month). During Christine Lagarde's presidency, the ECB governing council press conferences at the beginning of the COVID-19 pandemic in early 2020 (start of the Pandemic Emergency Purchase Program - PEPP) are also rated as highly dovish by MILA. The lower panel of figure (3) shows that for the period from 2011 to 2021, it is particularly notable that the Hawk-O-Meter for inflation and the real economy show a strong correlation over large parts. This is consistent with a narrative of prolonged weakness in aggregate demand, associated downward pressure on inflation dynamics, and the need for highly accommodative

[20]To identify sentences related to a specific topic, we use automatic keyword searching. We create lists of keywords pertaining to inflation or the real side of the economy, and annotate the sentences if at least one keyword appears.

monetary policy.

The monetary tightening phase between 2022 and 2023 was associated with a hawkish economic narrative. According to MILA, the inflation narrative became less dovish over the course of 2021 and, in early 2022, in light of unexpectedly strong rising inflation rates, turned hawkish for the first time in around ten years. In the following months, the press conferences became increasingly hawkish about inflation. The peak of the Hawk-O-Meter was at the end of 2022, when the inflation rate in the Eurozone exceeded 10%. In 2023 and 2024, the inflation narrative gradually became less hawkish. Most recently, it is assessed by MILA as balanced. In contrast, the communication of the ECB regarding real economic developments remained predominantly dovish or at most balanced even after the end of the COVID-19 pandemic. Such a divergence in the narrative between inflation and the real economy indicates that the ECB governing council, following the Russian invasion of Ukraine in February 2022, primarily emphasized supply-side disruptions – with inflationary but growth-damaging effects. The rather dovish communication about the real economy since mid-2023 also reflects the prolonged weakness in growth in the euro area.[21]

5. Evaluation, Replicability and Advantages of MILA

In this section, we evaluate MILA's performance in classifying textual data from the sphere of central banking using annotated datasets and demonstrate the advantages of MILA in comparison to established methods and approaches in the literature. The main challenge is the current lack of high-quality annotated data on the communication of the Eurosystem, which researchers can use to train and evaluate models. During our evaluation with existing classifications related to the euro area, we noticed some limitations in the datasets, which we report in addition to the evaluation results. In future work, we would like to explore the use of large-scale artificial intelligence models like MILA to create annotated data and evaluate central bank communication classifiers.

To assess MILA's performance, we consider several metrics: accuracy, precision, recall, and F1-Score. Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. Precision indicates the proportion of true positive results among all positive predictions made by the model, while recall shows the proportion of true positive results among all actual positives. The F1-Score, as the harmonic mean of precision and recall, balances both metrics. While accuracy provides a general sense of model performance, the F1-Score is useful in scenarios with imbalanced datasets, ensuring both precision and recall are considered.

5.1. Evaluating Sentiment Analysis

First, we analyze the performance of MILA in classifying statements from central bankers regarding economic optimism or pessimism. For this exercise, we use the annotated dataset on

[21] For a more in-depth analysis, see Deutsche Bundesbank (2025).

sentences from ECB speeches from Pfeifer and Marohl (2023). Pfeifer and Marohl (2023) manually labeled sentences from speeches of the Federal Open Market Committee (FOMC) members into the categories *Positive* and *Negative* and trained, evaluated, and compared several machine learning models, with RoBERTa (Liu et al., 2019) emerging as the most successful.[22] Note that for the fine-tuning of the models, the authors used *pseudo-labels* from the ECB and BIS datasets. Pseudo-labeling is a technique in which a model trained on labeled data is used to generate labels for unlabeled data. The pseudo-labeled data that exceeds a confidence threshold, meaning the model is highly confident in its classification, is then added to the training set.

A limitation of the dataset is the absence of a *Neutral* category, which is crucial to avoid misclassifying the typically balanced and diplomatic communication of central bankers. Additionally, due to the pseudo-labeling, the authors only collect sentences that were automatically classified with high confidence. As a result, the dataset primarily consists of sentences that are easier to understand and classify, and contain few difficult nuances.[23] Despite these limitations, we use the pseudo-labeled ECB sentences for evaluation since these sentences and the approved labels by the authors are unknown to MILA. To adjust MILA to this dataset, we constrain it to conduct binary classification without considering any inflation context or further information beyond the sentence.

For comparison, we use the dictionary from Loughran and McDonald (2011) (LM (2011)),[24] which is widely used to measure central bank communication (e.g., Arouba and Drechsel, 2024), even though the word lists are designed for companies and are therefore susceptible to measurement error and further well known limitations (Picault and Renault, 2017).[25] A natural candidate for comparison would be the model introduced by Pfeifer and Marohl (2023), but its application reveals issues of in-sample bias and overfitting.[26] Since MILA and the Loughran and McDonald (2011) dictionary are not trained or designed on these sentences and corresponding labels, we can apply both methods to the full sample without introducing any in-sample bias. Table (2) shows the performance of both approaches using standard classification metrics.

---

[22]For the FOMC sentences, Pfeifer and Marohl (2023) also labeled the most likely audiences. Classifying communication based on the intended audience is an important contribution to the literature, as it extends the perspective of the analysis.

[23]The main reason for such a strategy is the significant amount of time required and the necessity of having annotators with sufficient domain knowledge to ensure accuracy. Additionally, there is a risk of subjective bias or labeling inconsistency over time when conducted manually.

[24]We use the most current version of the dictionary from February 2024.

[25]The advantage of lexicographic methods lies in their low computational costs and the transparency of classification decisions, as they primarily involve word counting.

[26]We used the replication code to reproduce the training and test data for the labeled ECB sentences as described by Pfeifer and Marohl (2023). In our reproduction, their fine-tuned RoBERTa model achieved an accuracy of 100% for both training and test data, which is indicative of overfitting. The overfitting likely occurred because the authors used pseudo-labels from RoBERTa to fine-tune the same RoBERTa model. Additionally, Pfeifer and Marohl (2023) did not report the classification results for the FED, ECB, and BIS individually, but only for the entire sample, potentially overlooking this issue.

Metric	LM (2011)	MILA
Accuracy	0.76	0.96
Precision	0.61	0.90
Recall	0.97	0.99
F1-Score	0.75	0.94

Table 1: Comparison of Classification Metrics for Sentiment Classification: We compare the dictionary from Loughran and McDonald (2011) with MILA using the pseudo-annotated data from Pfeifer and Marohl (2023).

The analysis demonstrates a considerable performance of MILA, showing strict superiority over the bag-of-words approach, even when applying a very constrained version of MILA. MILA demonstrates balanced results with reliable positive classifications and very little false negative. In contrast, the bag-of-words approach is more prone to error, and the gap between Precision and Recall indicates that the dictionary is biased towards positive classification in the context of central banking communication. This results in communication being measured as more positive than it actually is.

5.2. Evaluating the Hawk-O-Meter

After evaluating MILA's performance in sentiment analysis, we now move on to assess its sentence classification capabilities for the narrative Hawk-O-Meter. Once again, the primary challenge is the lack of high-quality annotated data from senior central bank officials of the Eurosystem.[27] For the evaluation, we use the manually annotated dataset from Nitoi et al. (2023), which consists of 1,998 randomly drawn sentences from the minutes of the Polish, Hungarian, Romanian, and Czech central banks. These sentences are categorized as either *Hawkish*, *Dovish*, or *Neutral*. An obvious limitation for our analysis is that this textual data comes from countries that are not members of the euro area. However, their central banks are part of the European System of Central Banks (ESCB), making them more aligned than the FED.

Once again, we constrain MILA to classify without considering information beyond the sentences, as such information is not provided.[28] For comparison, we use the bag-of-words approach from Apel and Blix Grimaldi (2014) (AB (2014)), which the authors applied to the minutes of the Swedish Riksbank to measure the hawkishness of the deliberations. Nitoi et al. (2023) provide the results of their model evaluation in their paper.

The evaluation reveals a weak performance of the bag-of-words approach, indicating a high number of incorrect classifications. Although the dictionary performs better than randomly selecting one of the three categories, a precise and detailed analysis beyond basic document

[27]We do not evaluate the decision Hawk-O-Meter, as establishing an objective benchmark for it goes beyond the scope of this paper.

[28]Since the dataset only provides three categories, we classify *Moderate Dovish* as *Dovish* and *Moderate Hawkish* as *Hawkish*.

Metric	AB (2014)	MILA
Accuracy	0.43	0.84
Precision	0.67	0.84
Recall	0.43	0.84
F1-Score	0.35	0.83

Table 2: Comparison of Classification Metrics for the Hawk-Dove Metric: We compare the dictionary from Apel and Blix Grimaldi (2014) with MILA using the annotated data from Nitoi et al. (2023).

classification is likely to be biased. In contrast, MILA performs well on unseen data, achieving an accuracy of 84%. However, Nitoi et al. (2023) report achieving an accuracy of 88% by fine-tuning BERT on their data.[29] Our next step is to identify the reasons for the lower performance of MILA by examining the dataset and the classifications.

We employ an AI to analyze and comprehend the differences in classification between MILA and the annotations by Nitoi et al. (2023). The AI is informed that the sentences originate from the minutes of the four central banks that are part of the ESCB but not the euro area, whereas MILA is specifically designed for euro area central bank communication. We provide the AI with the sentence, the human classification, and the classification by MILA, and request that the AI determines whether the disagreement results from a mistake, institutional differences, or a lack of necessary additional context that the sentence does not provide.

Using AI and manual verification, we identified 64 (out of a total of 328) instances where disagreements can be attributed to institutional differences between these central banks and 67 instances where disagreements stem from a lack of additional contextual information that the sentence does not provide. We assume that Nitoi et al. (2023) have accurately classified all these cases, leveraging their additional knowledge of eastern European central banks, which their analysis focuses on. However, the same sentence would require a different classification if it originated from a euro area central bank.[30] If we remove these sentences from the evaluation, MILA achieves an accuracy of around 90%.[31]

[29]Nitoi et al. (2023) conducted the evaluation on a test set, while we evaluated MILA and the dictionary on the entire available dataset.

[30]The central banks of Hungary, Romania, and Poland have different inflation targets compared to the Eurosystem and are accustomed to higher growth rates due to their economic convergence process. In these countries, discussing an inflation target above two percent does not necessarily indicate the need for monetary tightening, as it would in the euro area. For example, the sentence "*Looking at the recent developments in the dynamics of consumer prices, Board members showed that the annual inflation rate had remained flat at 4.10 percent in May.*" is classified as *Neutral* by the authors, while MILA considers it as *Hawkish*. Similar discrepancies arise in discussions on the exchange rate and financial stability.

[31]Using AI and manual verification, we identified an additional 73 cases where a clear classification seems to depend on more information or where the annotation strategy of Nitoi et al. (2023) seems problematic. While many sentences are borderline cases for which introducing moderate categories would be appropriate, other annotations are puzzling. For example, Nitoi et al. (2023) classify the sentences "*In the context of the assessment of external developments, it was said that the higher economic growth combined with the lower expected inflation*

Overall, our observations indicate that datasets on other central banks should be used with caution when evaluating Eurosystem communication and should not be used to train models for analyzing Eurosystem communication without appropriate adjustments.[32] This also underscores the challenges involved when classifications necessitate an annotated data set for model training, which is central to the fine-tuning paradigm in machine learning. Nontheless, using the dataset with caution and manually verifying the classifications, we demonstrate that MILA's Hawk-O-Meter achieves accurate results.

5.3. Replicability of MILA's Results

A distinctive feature of generative language models is their stochastic nature, enabling them to produce varied outputs for the same prompt. This *artificial creativity* offers advantages and sets these models apart from deterministic and mechanical lexicographic methods or traditional machine learning techniques. However, for analyzing monetary policy communication, the results must be reliable, which would not be the case with highly volatile outputs if used repeatedly.[33] One of the core theses of this paper is that context-dependent analysis of individual sentences to rigorously derive indicators at the document level is significantly more beneficial in several respects than directly analyzing the entire document. This approach not only enhances the transparency of the results but also provides greater flexibility in analysis, avoids linguistically generated values, and ensures replicability. It is much easier for a language model to accurately assess a single sentence or text fragment than to evaluate an entire longer text, where it may not be clear which part of the text the model has prioritized.[34]

To demonstrate the advantage of granular analysis in terms of replicability, we created the narrative Hawk-O-Meter and sentiment indicator for the presidencies of Draghi and Lagarde three times. For comparison, we modified the MILA prompts to directly analyze the individual statements to immediately generate a final numerical score, instead of calculating the final score from individual sentences. We also generated triple time series from the overall document approach. Next, we calculated the cross-correlations for the individual versions of the indicator aggregated from sentences and the indicator generated directly by the language model, and presented the results in Table (5.3).

The analysis reveals a strong correlation, close to one, for the aggregated sentence-based indica-

---

perhaps suggested growth in potential output without the need to tighten monetary policy." and "*The opinion was expressed that rapid economic growth should not be a reason for tightening monetary policy.*" as hawkish.

[32] Nitoi et al. (2023) do not analyze the communication of the Eurosystem.

[33] One way to ensure de facto determinism in generative language models is to set the temperature parameter to 0. However, this results in a significant decline in the performance of these models. During our experiments, we observed a notable increase in hallucinations and misclassifications. These findings were consistent across both Llama 3.1 70B and GPT-4, the latter tested via the Azure Cloud.

[34] Requesting reasonings prior to classification enhances the model's consistency and offers the user some level of transparency. However, this does not equate to true explainability of AI. LLMs can fabricate virtually any explanation and optimize it linguistically, enabling them to justify any numerical value for classification. As the text length increases, manual verification becomes significantly more challenging.

	Narrative Sentiment						Narrative Hawk-O-Meter					
	MILA			Full Document			MILA			Full Document		
	First	Second	Third	First	Second	Third	First	Second	Third	First	Second	Third
Mario Draghi												
First	1.000			1.000			1.000			1.000		
Second	0.982	1.000		0.595	1.000		0.980	1.000		0.666	1.000	
Third	0.979	0.980	1.000	0.705	0.827	1.000	0.982	0.983	1.000	0.633	0.543	1.000
Christine Lagarde												
First	1.000			1.000			1.000			1.000		
Second	0.985	1.000		0.778	1.000		0.994	1.000		0.906	1.000	
Third	0.984	0.988	1.000	0.780	0.892	1.000	0.993	0.993	1.000	0.976	0.918	1.000

Table 3: Comparison of the Replicability of Narrative Sentiment and Hawk-O-Meter by Aggregating Individual Sentences (MILA) Versus the Whole Document Approach. We generate each indicator three times and calculate the correlations.

tors generated multiple times for both presidencies. This high correlation indicates replicability and reliability. We view the minor differences that still appear as a feature rather than a flaw. The varying classification of the same sentences with each request serves as an identifier for ambiguous communication, which policymakers can address to enhance clarity.[35] The full document approach demonstrates greater replicability for the Hawk-O-Meter during Lagarde's presidency, potentially due to clearer language compared to Mario Draghi's statements. However, reliability is significantly lower compared to our aggregated sentence-based indicators. In summary, the analysis supports our core thesis that generating numerical classification values directly with a LLM based on full documents should be avoided.[36]

6. Conclusion

In this paper, we introduce the MILA, an innovative framework that leverages advanced prompt-engineering techniques and LLMs to analyze and measure different semantic dimensions of monetary policy communication in the euro area. MILA's granular classification approach, combined with statistical aggregation, enhances transparency, integrates expert knowledge, and ensures rigorous calculations. Our application of MILA to the ECB's monetary policy statements since November 2011 reveals significant changes in the tone of communication, reflecting economic conditions and monetary policy shifts. The sentiment and Hawk-O-Meter indicators MILA derived from the MPS provide nuanced insights into the ECB's communication strategies and

[35] A correlation analysis does not provide information about the levels of the replicated indicators. However, for our aggregated sentence-based approach, the levels are also nearly identical, as validated through visual analysis and by comparing the mean values.

[36] We also avoid this approach for the Decision-Hawk-O-Meter, as the individual scores are transformed from qualitative statements based on various communicative aspects of monetary policy announcements (see Appendix A).

demonstrate the framework's effectiveness in analyzing central bank communication.

Furthermore, our evaluation highlights MILA's high accuracy, flexibility and consistency of results. During the evaluation, we identified several weaknesses in existing datasets and evaluations in the literature, such as the absence of classification categories or in-sample bias. The shift from fine-tuned models to prompt engineering provides a flexible solution to overcome the lack of high-quality training datasets. In the future, models like MILA could be used to facilitate the creation of training and evaluation datasets. Overall, this study underscores that advanced AI techniques can enhance the understanding and evaluation of central bank communication.

References

Apel, M., Blix Grimaldi, M., 2014. *How Informative Are Central bank Minutes?* Review of Economics 65.

Apel, M., Blix Grimaldi, M., Hull, I., 2022. *How Much Information Do Monetary Policy Committees Disclose? Evidence from the FOMC's Minutes and Transcripts.* Journal of Money, Credit, and Banking 54(5), 1460–1489.

Deutsche Bundesbank, 2025. *Monetary Policy Communication According to AI.* Monthly Report March.

Ehrmann, M., Fratzscher, M., 2007. *Communication by Central Bank Committee Members: Different Strategies, Same Effectiveness?* Journal of Money, Credit, and Banking 39, 509–541.

Gambacorta, L., Kwon, B., Park, T., Patelli, P., Zhu, S., 2024. *CB-LMs: Language Models for Central Banking.* BIS Working Papers .

Grattafiori, A., et al., 2024. *The Llama 3 Herd of Models.* Technical Report. arXiv .

Kanelis, D., Siklos, P.L., 2025. *The ECB Press Conference Statement: Deriving a New Sentiment Indicator for the Euro Area.* International Journal of Finance & Economics , 652–664.

Loughran, T., McDonald, B., 2011. *When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks.* Journal of Finance 66(1), 35–65.

Maibaum, F., Kriebel, J., Foege, J.N., 2024. *Selecting Textual Analysis Tools to Classify Sustainability Information in Corporate Reporting.* Decision Support System 183.

Nitoi, M., Pochea, M.M., Radu, S.C., 2023. *Unveiling the sentiment behind central bank narratives: A novel deep learning index.* Journal of Behavioral and Experimental Finance 38.

Pfeifer, M., Marohl, V.P., 2023. *CentralBankRoBERTa: A fine-tuned large language model for central bank communications.* The Journal of Finance and Data Science 9, 2405–9188.

Picault, M., Renault, T., 2017. *Words are not all Created Equal: A New Measure of ECB Communication.* Journal of International Money and Finance , 136–156.

Sun, S., Yuan, R., Cao, Z., Li, W., Liu, P., 2024. *Prompt Chaining or Stepwise Prompt? Refinement in Text Summarization.* Accepted to Findings of ACL 2024 .

Taylor, J.B., 1993. *Discretion versus Policy Rules in Practice.* Carnegie-Rochester Conference Series on Public Policy 193, 195–214.

Wu, T., Terry, M., Cai, C.J., 2022. *AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts.* CHI'22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems , 1–22.

Figures

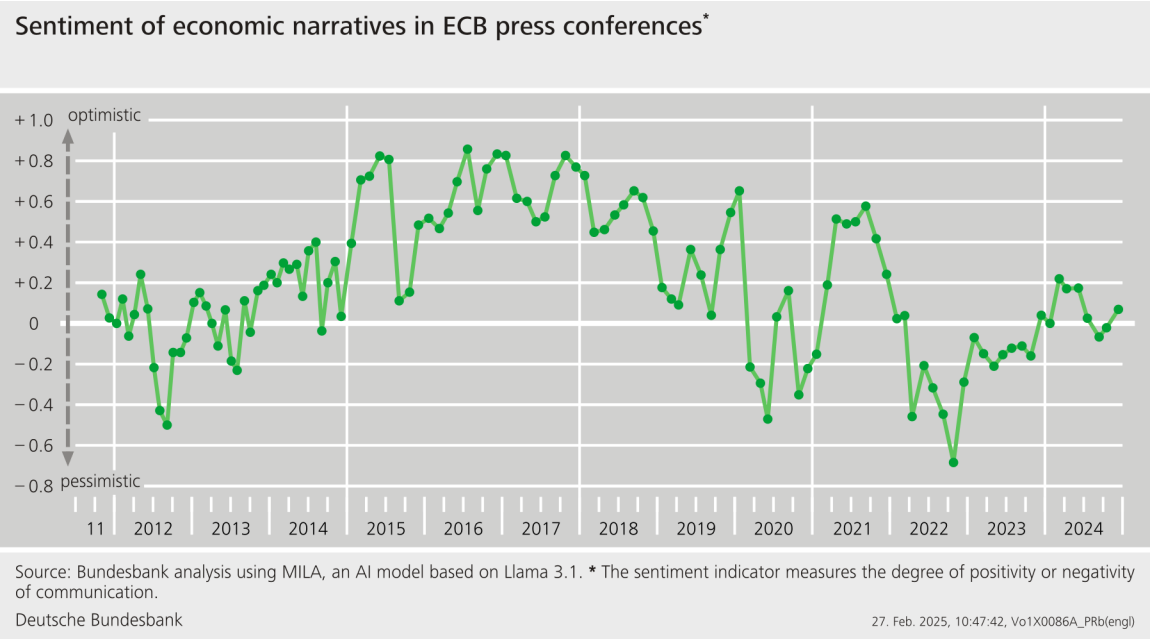

Figure 1: Narrative Sentiment Indicator for the Monetary Policy Statement

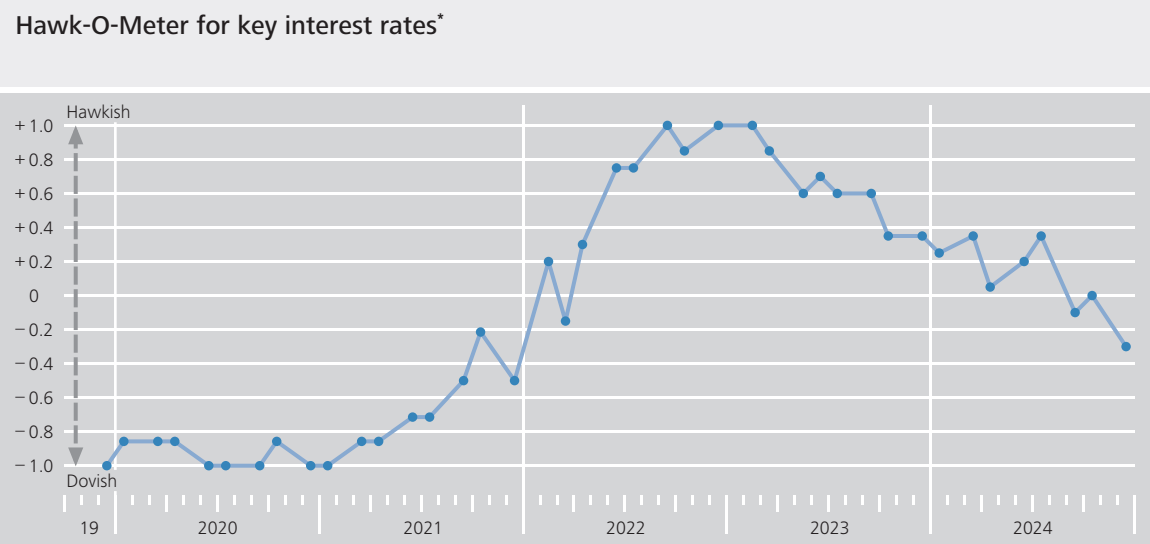

Figure 2: Decision Hawk-O-Meter for the Monetary Policy Statement

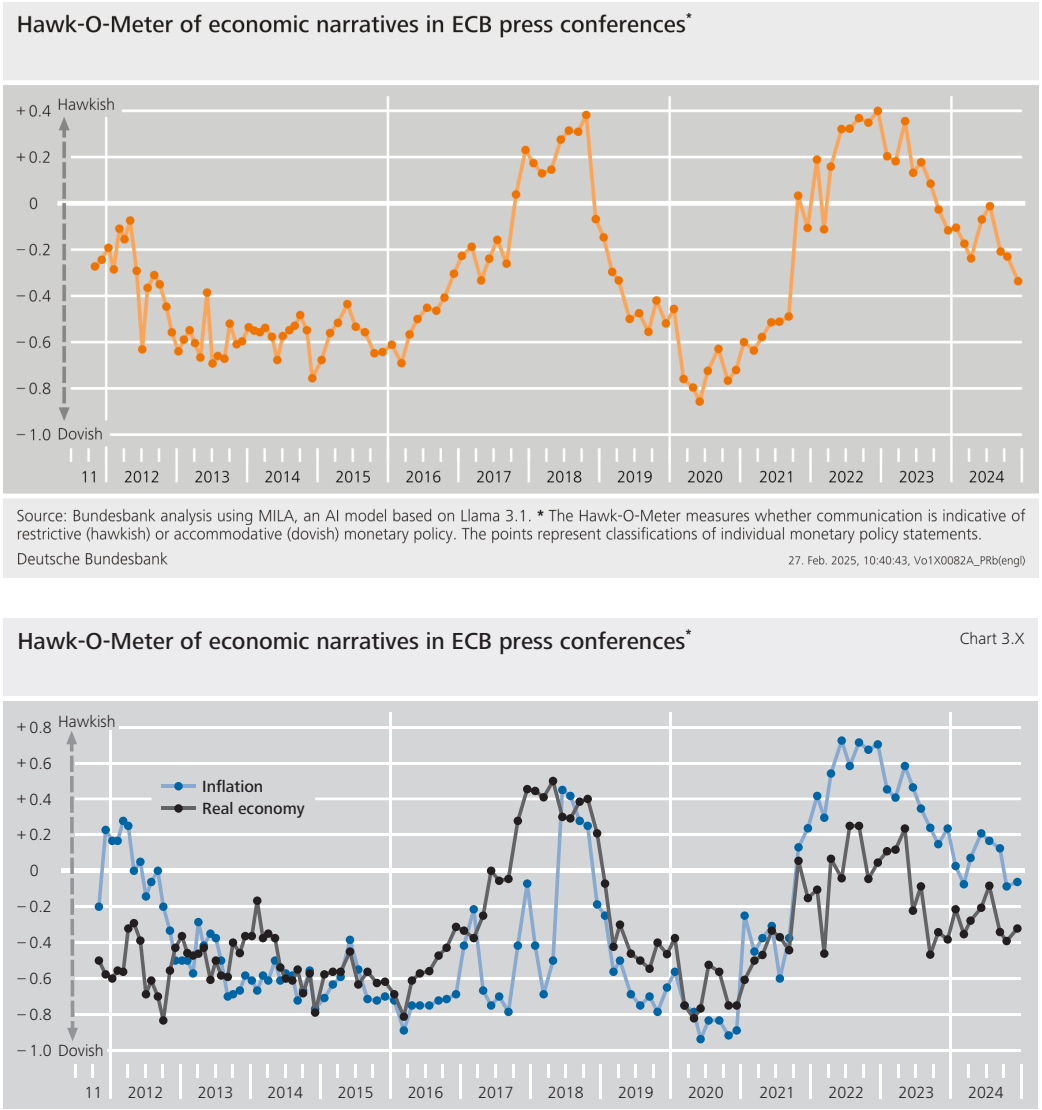

Figure 3: Narrative Hawk-O-Meter for the Monetary Policy Statement

Appendix A. Conditions and Scoring for the Decision Hawk-O-Meter

- **Interest Rate Score:** Will key interest rates be lowered, raised, or maintained, and if changed, by how many basis points?

Interest Rate Score	
Condition	**Score**
Increase by 50 basis points or more:	0.30
Increase by 25 basis points:	0.15
No change:	0
Decrease by 25 basis points:	-0.15
Decrease by 50 basis points or more:	-0.30

- **Interest Rate Outlook Score:** Are signals for future reductions or increases communicated? Is the language cautious, neutral, data-dependent, or clearly directive, or is it subject to directional distortion?

Interest Rate Outlook Score	
Condition	**Score**
Clear commitment to future rate cuts or monetary easing or explicit announcements of a long-lasting expansionary policy stance:	0.30
Subtle hints toward possible easing measures:	0.15
Data-dependent, neutral stance (no clear directional guidance):	0
Subtle hints toward possible tightening measures or explicit announcements to maintain currently restrictive policy levels for a longer time:	-0.15
Clear commitment to future rate hikes or monetary tightening:	-0.30

- **Inflation Score:** Are the decisions justified by direct reference to inflation developments, and is there an indication of upward or downward risk?

Inflation Score	
Condition	**Score**
Inflation clearly above target, strong emphasis on the need for immediate tightening actions:	0.20
Inflation above target and suggestion that further measures to contain it may be needed:	0.10
Inflation around target, neutral stance or no information provided:	0.00
Inflation somewhat below target indicating concern that current policy may not be sufficient to reach the target:	-0.10
Inflation clearly below target, emphasis on the need for more stimulus:	-0.20

- **General Tone Score:** What is the tone of the language, and what other announcements are made with implications for the interest rate path? Is the language particularly cautious, supportive, expansionary, or markedly restrictive?

General Tone Score	
Condition	**Score**
Clearly restrictive, firm stance on tightening and inflation control:	0.20
Slightly restrictive tone, more focus on controlling inflation:	0.10
Balanced and neutral tone:	0.00
Cautious but generally supportive tone:	-0.10
Strongly expansionary, supportive tone, focusing on easing measures (especially beyond key interest rates):	-0.20