Analyzing policies of Bank of Japan with Large Language Models*

Zhiyuan Han
HEC Paris
zhiyuan.han@hec.edu

Ziyang Zeng
HEC Paris
ziyang.zeng@hec.edu

May 29, 2026

Abstract

This study investigates whether the informational content of Bank of Japan (BOJ) communications can predict short-term movements in Japanese Government Bond (JGB) yields and BOJ purchase operations. We construct a BOJ Policy Stance Index by applying FinBERT-based sentiment analysis and keyword scoring to a corpus of 807 BOJ speeches and announcements spanning 1996 to 2025, and integrate this index into an event-study framework covering the BOJ's unconventional monetary policy journey, from the Zero Interest Rate Policy and Quantitative and Qualitative Easing, through Yield Curve Control, to the post-2024 normalization phase. Our quantitative benchmark reveals that daily JGB yield changes follow a near-random walk and are uncorrelated with changes in BOJ outright purchases; the NLP-derived hawkish-dovish scores likewise show no correlation with yield changes or subsequent position adjustments. While frontier large language models appear to predict yields with high in-sample accuracy, we show that this performance likely reflects forward-looking bias rather than genuine forecasting ability. These results provide converging evidence that the semi-strong form of the Efficient Market Hypothesis holds in the JGB market with respect to BOJ communication. The BOJ's signaling relies on linguistic subtleties in formal business Japanese-such as degrees of politeness-that are lost in English translation. Consequently, domestic investors receive clear signals, whereas global investors accessing only the English text are deprived of the same guidance value.

Keywords: BOJ, policies, LLM, efficient market hypothesis, NLP

*The authors are very grateful to their supervisor, Takaya Sekine, CFA.

1 Introduction

The Bank of Japan (BOJ) has issued a series of policies for the past 30 years, evolving from conventional interest rate adjustments to increasingly unconventional monetary easing measures. Following the zero interest rate policy (ZIRP) introduced in 1999, the early 2000s saw the pioneering introduction of quantitative easing (QE). After a brief intermission between 2006 and 2008, the Global Financial Crisis prompted the BOJ to reactivate and expand its asset purchase programs. A significant shift occurred in 2013 with the launch of Quantitative and Qualitative Monetary Easing (QQE), which involved massive purchases of Japanese government bonds (JGBs) and risky assets such as ETFs and J-REITs, targeting a 2% inflation goal. This framework was further extended into a negative interest rate policy (NIRP) in 2016 and later supplemented by Yield Curve Control (YCC), aiming to cap 10-year JGB yields around zero (Hattori and Yoshida, 2021). Between 2016 and 2024, YCC became the centerpiece of BOJ's policy toolkit. However, as inflation dynamics shifted, the BOJ incrementally widened YCC bands and, in March 2024, formally exited both the negative interest rate and YCC regimes, marking a historic pivot toward policy normalization while maintaining accommodative financial conditions. The subsequent normalization period (2024-2026) has seen the BOJ reduce its JGB purchases and raise the policy rate gradually. Annual average headline CPI inflation illustrates this transition: after near-zero or negative readings in 2020-2021 (0.0% and -0.2%), inflation rose to 2.5% in 2022, peaked at 3.27% in 2023, eased to 2.74% in 2024, and remained elevated at approximately 3.17% in 2025.

The YCC policy has received good results. Hattori and Yoshida (2021) document how the BOJ successfully controlled the 10-year Japanese government bond (JGB) yields within the target range through a combination of fixed-price and fixed-amount purchase operations. Fixed-price operations, in particular, were effective in capping yields at the desired level. Under YCC, the BOJ made its bond purchase operations endogenous to market yields. Tests showed that under YCC, higher yields triggered increased bond purchases, reversing the pre-YCC direction of causality. This allowed the BOJ to manage balance sheet growth more effectively. YCC significantly reduced the dispersion of expert yield forecasts, especially during periods with a narrow target range and enhanced forward guidance. This could indicate that investors' expectations aligned more closely with the BOJ's policy goals. JGB yields became less volatile and followed a stationary process under YCC, particularly when the target range was binding. This contrasts with the pre-YCC period, where yields were non-stationary despite quantitative easing. The stabilizing effects of YCC extended to other maturities (2-, 5-, and 20-year JGBs), especially during periods of narrow targeting. This suggests effective transmission

of the policy across the entire yield curve. The BOJ achieved these results without relying on frequent fixed-price operations, instead relying on forward guidance and the credibility of its endogenous intervention rule.

While the macroeconomic effects of the BOJ's unconventional policies are well-documented, a critical open question remains: "does the BOJ's public communication convey information that can predict short-term movements in JGB yields or the central bank's own market operations?" This question sits at the intersection of the Efficient Market Hypothesis (EMH) and the growing literature on textual analysis in finance. If the semi-strong form of the EMH holds, public speeches and announcements should be instantaneously incorporated into market prices, yielding no systematic predictability. Conversely, if linguistic nuances-particularly in a communication culture characterized by deliberate ambiguity and indirect signaling-contain incremental information not immediately digested by markets, NLP-based strategies could potentially generate alpha.

This paper tests this hypothesis by constructing a comprehensive analytical pipeline that bridges quantitative market microstructure analysis with state-of-the-art Natural Language Processing (NLP). We first establish a predictability baseline using daily JGB yield and BOJ purchase data from 2010 to 2025, documenting the statistical properties-unit-root behavior, fat tails, and regime-dependent volatility, which any forecasting model must overcome. We then apply FinBERT, a financial-domain-specific language model, to a corpus of 807 BOJ speeches and announcements to derive a continuous hawkish-dovish score that measures the policy stance embedded in each communication. This textual index is subsequently integrated into an event-study framework that tests for causal relationships between BOJ rhetoric and subsequent yield movements, purchase adjustments, and yield curve dynamics.

Our analysis yields three principal findings. First, daily JGB yield changes are approximately unpredictable from both past quantitative data and contemporaneous NLP-derived sentiment scores, consistent with the semi-strong form of the EMH. Second, the observed regime shifts in the BOJ's reaction function from active easing (negative purchase-yield correlation) to YCC (positive correlation) to normalization (sharply negative correlation) are visible in the textual record but do not translate into exploitable short-horizon trading signals. Third, we document a methodological caution: while cutting-edge large language models like Claude can retroactively "predict" yield changes with high in-sample accuracy, this performance is more attributable to forward looking bias rather than genuine forecasting ability. Through a counterfactual analysis of the September 2025 Monetary Policy Meeting dissent, we demonstrate that frontier LLMs can nonetheless serve as powerful tools for "ex-ante" expectation quantification and communication gap identification-capabilities that, while not profitable in the semi-strong EMH sense, offer significant value

for policy analysis and risk management.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature on LLMs, information retrieval, and sentiment analysis. Section 3 provides a detailed quantitative analysis of BOJ purchase positions and JGB yield dynamics from 2010 to 2025, establishing the empirical benchmark. Section 4 presents our NLP methodology and results, including future-orientation detection, hawkish-dovish stance index construction, topic modeling, event-study tests, and frontier LLM-based counterfactual analysis. Section 5 concludes with implications for both the EMH literature and the practical application of NLP in fixed-income markets.

2 Literature Review

2.1 Large Language Models (LLMs) and Core NLP Architectures

The field of Natural Language Processing (NLP) has undergone a series of developments in recent years, and the transformer architecture itself is that revolution. Introduced by Vaswani et al. (2017), the transformer fundamentally redefined how language models are built, replacing recurrent and convolutional architectures with a self-attention mechanism that enables parallel processing and long-range dependency capture. Building upon this, Devlin et al. (2019) introduced BERT, a seminal model that fundamentally changed the approach to NLP tasks. BERT's key innovation lies in its bidirectional training, allowing it to learn deep contextual representations by conditioning on both left and right context in all layers. Its pre-training and fine-tuning deeply influenced the standard, demonstrating state-of-the-art performance on a wide range of tasks and serving as the foundation for many subsequent models.

Following BERT's success, significant research focused on optimizing its pre-training methodology. Liu et al. (2019) proposed RoBERTa (Robustly Optimized BERT Approach), which demonstrated that BERT was undertrained. By modifying key hyperparameters, including training the model longer with larger batches and more data, and dynamically changing the masking pattern applied to the training data, RoBERTa achieved substantial performance improvements over the original BERT. Its robust performance has made it a popular choice for downstream tasks, and models based on its architecture, such as roberta-base-squad2, are frequently employed in question-answering experiments.

The need to extend these powerful models beyond English led to the development of multilingual and language-specific architectures. Conneau et al. (2020) introduced XLM-RoBERTa (XLM-R), a cross-lingual model pre-trained on text in 100 languages. By scaling up the amount of training data and using a larger vocabulary, XLM-R significantly outperformed prior multilingual models like multilingual BERT, making it a highly effective tool for cross-lingual NLP tasks, including question answering and sequence classification. Parallel to this, research has focused on creating high-performance models for specific languages. For instance, Martin et al. (2020) presented CamemBERT, a language model tailored for French. Based on the RoBERTa architecture and trained on a large French corpus, CamemBERT has achieved superior results on various French NLP benchmarks, particularly for tasks like Named-Entity Recognition (NER).

Currently, a parallel and highly influential branch of research has focused on scaling up model size and shifting towards generative capabilities. Brown

et al. (2020) marked a landmark moment with the introduction of GPT-3 (Generative Pre-trained Transformer 3). As a massive autoregressive language model with 175 billion parameters, GPT-3 demonstrated remarkable few-shot and zero-shot learning abilities. Unlike its predecessors that required fine-tuning for specific tasks, GPT-3 could perform tasks like translation, question-answering, and text generation by simply being provided with a few examples or a natural language prompt. This work is crucial for contextualizing the recent emergence of powerful generative AI systems, such as ChatGPT.

Since the advence of GPT-3 in 2020, the field has been defined by a race towards even larger models and the refinement of the instruction-following paradigm. This period saw the rise of models like GPT-4 (OpenAI *et al.*, 2023), PaLM (Chowdhery *et al.*, 2022), and LLaMA (Touvron *et al.*, 2023), which further pushed the boundaries of scale and introduced innovations in training techniques, such as the mixture of experts (MoE) architecture to improve efficiency. Crucially, the focus shifted from mere few-shot performance to aligning models with human intent through techniques like Reinforcement Learning from Human Feedback (RLHF) (Agarwal *et al.*, 2022). This approach was instrumental in transforming raw generative models into conversational agents like ChatGPT, capable of nuanced dialogue and complex task completion. Concurrently, the drive for democratization led to the explosion of open-source, "smaller" yet highly capable models (e.g., LLaMA, Mistral) that rival their larger counterparts (Touvron *et al.*, 2023), fostering a rich ecosystem of fine-tuned variants for specialized domains. This era has also been marked by a surge in multimodal capabilities, with models like CLIP (Linear-probe, 2021) and Flamingo (Alayrac *et al.*, 2022), and later GPT-4V, blurring the lines between text, vision, and audio, enabling the development of unified models that can reason across different data types. Research is focused on improving efficiency through techniques like quantization and pruning (X. Zhu *et al.*, 2024), mitigating inherent issues like hallucination and bias, and extending context windows to process entire documents or long-form media.

But despite the remarkable fluency and few-shot capabilities of generative LLMs, their use for knowledge extraction in policy analysis raises a fundamental concern: these models are optimized to propose the most statistically plausible continuation of a prompt, rather than to retrieve verifiable facts. Consequently, a purely generative pipeline inherits all the biases discussed in the literature: the model is "looking for an answer" in a black-box fashion, without exposing the source from which its conclusion is derived. For our task of extracting structured policy signals from the BOJ communications, such opacity is especially problematic, because any claim about the BOJ's stance must be auditable and grounded in specific textual evidence. We therefore deliberately avoid relying on a generative component as the primary extraction engine. Instead, we adopt a less black-box approach that maintains a

clear logical link in the chain of reasoning: we use encoder-only transformer architectures (e.g., FinBERT) (Araci, 2019) for stance classification, and when the task requires locating supporting evidence within a large corpus, we turn to transparent retrieval methods such as BM25. Because BM25 scores documents based on explicit lexical overlap with the query, it enables us to trace every extracted statement back to the exact passages and terms that produced it, making the provenance of the answer fully visible. This design choice ensures that the NLP pipeline remains interpretable, auditable, and scientifically defensible.

2.2 Information Retrieval (IR) and Question Answering (QA)

The task of open-domain question answering (QA), the ability to answer factual questions based on a large corpus of text, has evolved significantly, driven by advances in both information retrieval and machine reading comprehension. Modern QA systems increasingly rely on a hybrid architecture that combines traditional term-based methods with neural dense representations to achieve high accuracy and efficiency.

For decades, term-based sparse retrieval methods have been used as information retrieval systems. Robertson et al. (2009) formalized the BM25 (Best Matching 25) algorithm, which remains one of the most robust and widely adopted ranking functions. BM25 scores documents based on the frequency of query terms within a document, while accounting for term saturation and document length normalization. Despite the rise of neural methods, BM25 is still considered a powerful baseline and an essential component in multi-stage retrieval systems due to its efficiency, interpretability, and strong performance on out-of-domain datasets.

While BM25 excels at lexical matching, it struggles with the semantic gap situations where a document contains relevant concepts but uses different vocabulary than the query. To address this, neural dense retrieval methods learn to embed both questions and documents into a dense vector space where relevance is measured by semantic similarity. A breakthrough in this area came from Reimers and Gurevych (2019) with Sentence BERT (SBERT). By modifying the BERT architecture to use siamese and triplet networks, SBERT can derive semantically meaningful sentence embeddings that can be efficiently compared using cosine similarity. This work made dense retrieval practical for large-scale tasks by enabling fast, offline computation of embeddings and real-time similarity searches.

Building directly on the concept of dense retrieval, Karpukhin et al. (2020) introduced Dense Passage Retrieval (DPR), a highly effective framework specifically designed for open-domain QA. DPR employs a dual-encoder architec-

ture: one BERT-based encoder for questions and another for passages. The model is trained to maximize the inner product of positive question-passage pairs while minimizing it for negative pairs. Unlike generic sentence embeddings, DPR is optimized end-to-end for the retrieval task, making it particularly effective at finding passages containing answers to factual questions. Its strong empirical performance has established it as a leading model for the retrieval stage in QA pipelines.

The integration of these retrieval methods into a complete QA system is comprehensively surveyed by Zhu et al. (2021) . This work provides a detailed overview of the "retriever-reader" architecture, which has become the dominant paradigm for open-domain QA. In this framework, the retriever (such as BM25 or DPR) first selects a small subset of relevant passages from a massive corpus. Subsequently, the reader (typically a transformer-based model like BERT) extracts or generates the precise answer from these candidate passages. By decomposing the task into these two stages, the retriever-reader model balances scalability with the deep language understanding required for accurate question answering.

2.3 Sentiment Analysis and Tonality

Sentiment analysis, the task of computationally identifying and categorizing opinions expressed in text, is a cornerstone of NLP. Its applications range from social media monitoring to financial market prediction. However, the effectiveness of sentiment analysis models is highly dependent on the domain of the text and the methodology used.

In the domain of social media, specifically Twitter (now X), the work of Barbieri et al. (2020) on TweetEval has become a critical benchmark. TweetEval is a unified framework consisting of seven heterogeneous tasks specific to tweet understanding, including sentiment analysis. The authors introduced a suite of fine-tuned transformer models optimized for the unique linguistic characteristics of social media, such as hashtags, mentions, and informal language. The fine-tuned sentiment tweet roBERTa model, which emerged from this work, represents a new approach for this domain. Its selection for use in various studies is typically justified by its superior performance on custom datasets, demonstrating the value of domain-specific fine-tuning over generic language models.

While transformer-based models excel in general sentiment, the financial domain presents unique challenges. Standard sentiment lexicons often misclassify words that carry specific connotations in a financial context. Addressing this gap, Loughran and McDonald (2011) created a foundational dictionary for financial sentiment analysis. By analyzing a large number of 10-K filings (annual reports) from U.S. companies, they demonstrated that words

commonly labeled as negative in general-purpose dictionaries (e.g., "liability", "tax", "cost") are not necessarily negative in a financial context. Their work established a domain-specific lexicon that became a benchmark for subsequent financial NLP research. However, as noted in contemporary studies, the dictionary-based approach is inherently rule-based and keyword-dependent, which makes it potentially vulnerable to adversarial attacks.

The vulnerability of traditional sentiment analysis, particularly in high-stakes domains like finance, has recently come under scrutiny. Leippold (Leippold, 2023) directly investigated this issue by studying adversarial attacks on financial sentiment analysis using large language models. The research demonstrated how generative models like GPT-3 could be used to craft texts that preserve grammatical correctness and apparent meaning but systematically fool sentiment classifiers. This work is highly relevant to the ongoing discussion regarding the limitations of keyword-based and even early neural approaches. It highlights that as models become more powerful, the methods to attack them also become more sophisticated, posing new challenges for the reliability of automated sentiment analysis.

3 Position Analysis

In this section, we attempt to draw a timeline of the positions of the BOJ under the influence of its major policies.

We take the data from the BOJ's official website.

The time span of different outright purchases of JGBs, Excluding Floating-rate Bonds and Inflation-indexed Bonds, in a competitive auction method, is about one week. The residual maturities of JGBs include up to 1 year, more than 1 year and up to 3 years, more than 3 years and up to 5 years, more than 5 years and up to 10 years, more than 10 years and up to 25 years, and more than 25 years.

3.1 Overview of BOJ Purchases in 2025

The figure below presents the timeline of BOJ outright JGB purchase amounts by residual maturity bucket throughout 2025, overlaid with corresponding JGB yields and key policy events.

The data reveals four distinct purchase periods, each corresponding to a shift in the BOJ's quantitative tightening (QT) stance. In Regime 1 (January-March 2025), the BOJ maintained its initial reduction schedule, with the <1Y bucket at ¥150 billion per auction, medium-term buckets (1-3Y, 3-5Y) at ¥300 billion, the core 5-10Y bucket at ¥325 billion, and the 10-25Y bucket at ¥150 billion. These levels reflect the QT plan announced in July 2024, targeting a reduction of approximately ¥400 billion per quarter in monthly purchase volumes

3.1.1 The April Reduction and Tariff Shock

A broad, simultaneous reduction across all maturity buckets became effective in early April 2025. The <1Y bucket was cut from ¥150 billion to ¥100 billion; the 1-3Y and 3-5Y buckets each declined by ¥25 billion to ¥275 billion; the 5-10Y bucket fell to ¥300 billion; and the 10-25Y bucket was reduced to ¥135 billion. This compression was consistent with the pre-scheduled QT trajectory targeting approximately ¥4.5 trillion in monthly purchases for Q2 2025.

Notably, this reduction coincided with the announcement of US reciprocal tariffs on April 2, 2025. Despite the resulting market volatility - which temporarily suppressed JGB yields across the curve - the BOJ maintained its reduced purchase schedule without any tactical intervention. This is consistent with the BOJ's March MPM (Monetary Policy Meeting) communication, which acknowledged trade policy uncertainty while emphasizing commitment to its normalization path.

3.1.2 The Q3 Frequency Adjustment: Per-Auction Increases, Monthly Reductions

In early July 2025, the per-auction amounts for the three intermediate maturity buckets (1–3Y, 3–5Y, 5–10Y) each rose by ¥50 billion relative to Q2 levels — to ¥325 billion, ¥325 billion, and ¥350 billion respectively — while the ultra-short (<1Y) and long-end (10–25Y, >25Y) buckets remained unchanged. At first glance, this appears to contradict the ongoing QT trajectory.

However, this per-auction increase was more than offset by a simultaneous reduction in auction *frequency*. According to the BOJ's official outline for Q3 2025 JGB purchase operations, the frequency of auctions for the 1–3Y, 3–5Y, and 5–10Y buckets was reduced from four times per month to three times per month, "in order to avoid each auction amount becoming small as the Bank has been proceeding with the reduction in the amount of JGB purchases." The net effect on monthly purchase volumes was therefore a *decrease*, not an increase. For example, the 1–3Y bucket moved from ¥275 billion × 4 = ¥1,100 billion per month to ¥325 billion × 3 = ¥975 billion per month, a reduction of ¥125 billion per month despite the ¥50 billion rise in per-auction size. The same arithmetic applies to the 3–5Y and 5–10Y buckets.

This adjustment thus represents a continuation — rather than a reversal — of the BOJ's QT schedule. By consolidating fewer, larger auctions, the BOJ maintained orderly market functioning and sufficient auction sizes while still reducing the aggregate monthly flow of purchases, consistent with the target of approximately ¥400 billion in quarterly reductions. The timing of this operational change, coinciding with the Japan–US trade deal in mid-July and the run-up to the July 30–31 MPM at which the BOJ revised its FY2025 CPI forecast upward to 2.7%, further underscores that the BOJ viewed external conditions as sufficiently stable to proceed with — rather than pause — its normalization path.

3.1.3 The October Normalization Step

From early October 2025, the BOJ implemented its deepest across-the-board reduction of the year. The 1-3Y bucket fell to ¥300 billion, the 3-5Y bucket to ¥280 billion, the 5-10Y bucket to ¥305 billion, and notably, the 10-25Y bucket reached a new low of ¥115 billion. The >25Y bucket remained stable at ¥75 billion throughout the year, suggesting the BOJ was more cautious about allowing the ultra-long end to cheapen.

This October reduction followed the September MPM, at which two board members (Takata and Tamura) dissented in favor of an immediate rate hike - the first dual dissent in eight years. The market had already priced in further tightening, with the 10Y JGB yield rising to 1.655% by September 19 from 1.498% in June.

3.1.4 Yield Curve Dynamics and Position Changes

The relationship between BOJ purchase reductions and JGB yields is summarized in Table X. Across the year, the yield curve experienced substantial steepening: the 1Y yield rose by approximately 38 basis points (from 0.48% to 0.86%), while the 30Y yield rose by over 100 basis points (from 2.33% to 3.36%). This differential reflects the combined effect of two forces: policy rate hikes anchoring expectations at the short end, and QT-driven reduced demand at the long end.

One key observation is that yield movements frequently preceded formal BOJ announcements, consistent with Deputy Governor Himino's speech on January 14 triggering a visible yield increase two weeks before the January MPM rate hike. This pattern suggests that the BOJ's forward guidance through speeches was effective in anchoring market expectations, a finding that connects directly to the NLP analysis in Section 4.

3.2 A broader view since 2010

While the previous subsection focused on the granular evolution of BOJ purchase activities in 2025, this section expands the analysis to the period from 2010 to 2025. This longer horizon captures the full spectrum of the BOJ's unconventional monetary policy journey: from the early Quantitative Easing (QE) and Zero Interest Rate Policy (ZIRP) era, through the launch of Quantitative and Qualitative Monetary Easing (QQE) in 2013, the introduction of Negative Interest Rate Policy (NIRP) and Yield Curve Control (YCC) in 2016, and finally the normalization phase beginning in 2024. By examining the joint dynamics of JGB yields and BOJ outright purchases over this extended period, we establish a baseline for evaluating the textual policy stance derived in Section 4.

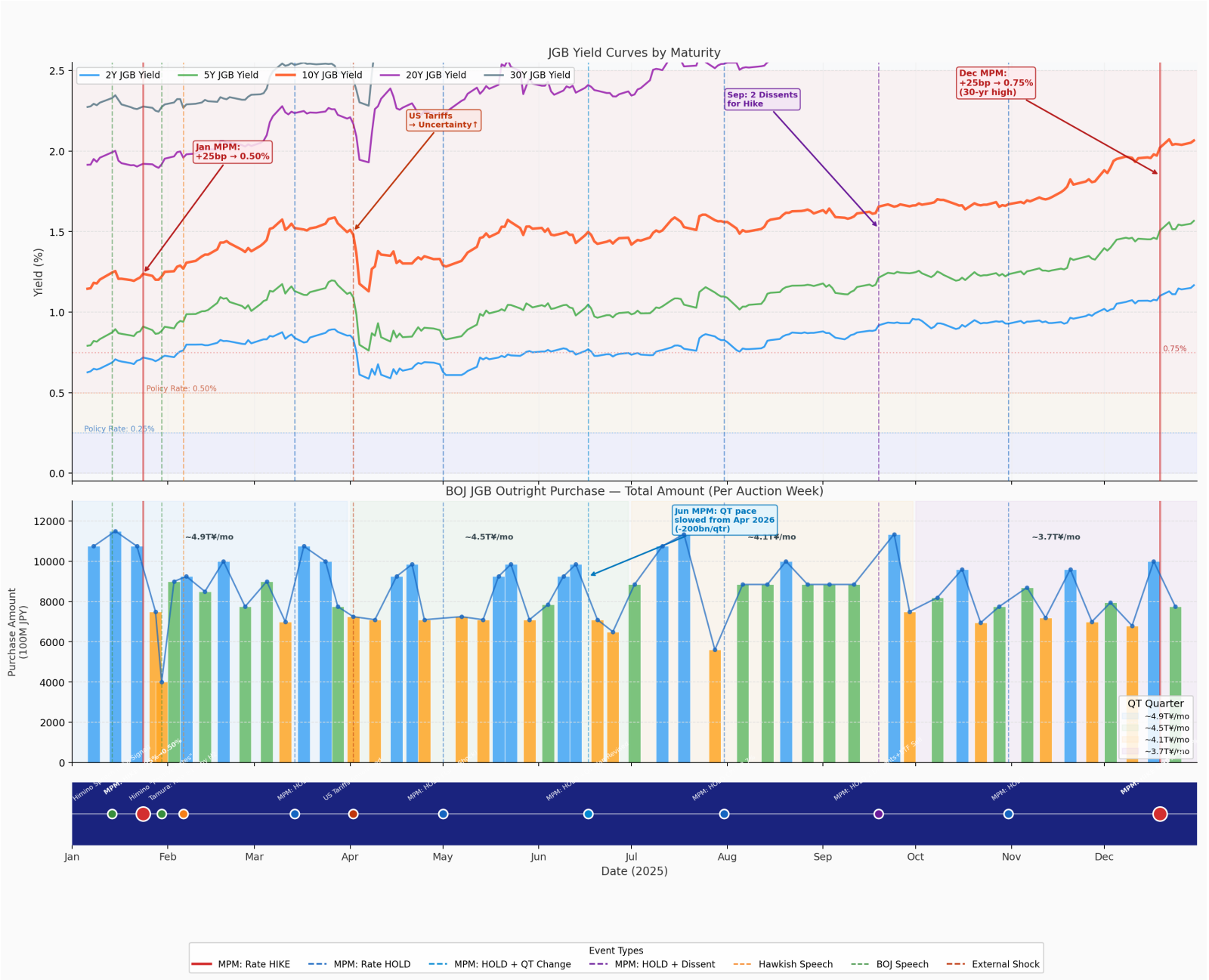
3.2.1 Evolution of JGB Yields and BOJ Purchase Volumes

Several distinct regimes can be found from 2010 to 2025:

2010-2012: Deflation fighting and early QE. The 10Y yield remained below 1.5%, often near 1.0%, while total purchases were modest (typically under 1 trillion yen per month). This period reflects the BOJ's gradual response to global financial crisis aftereffects and domestic deflation.

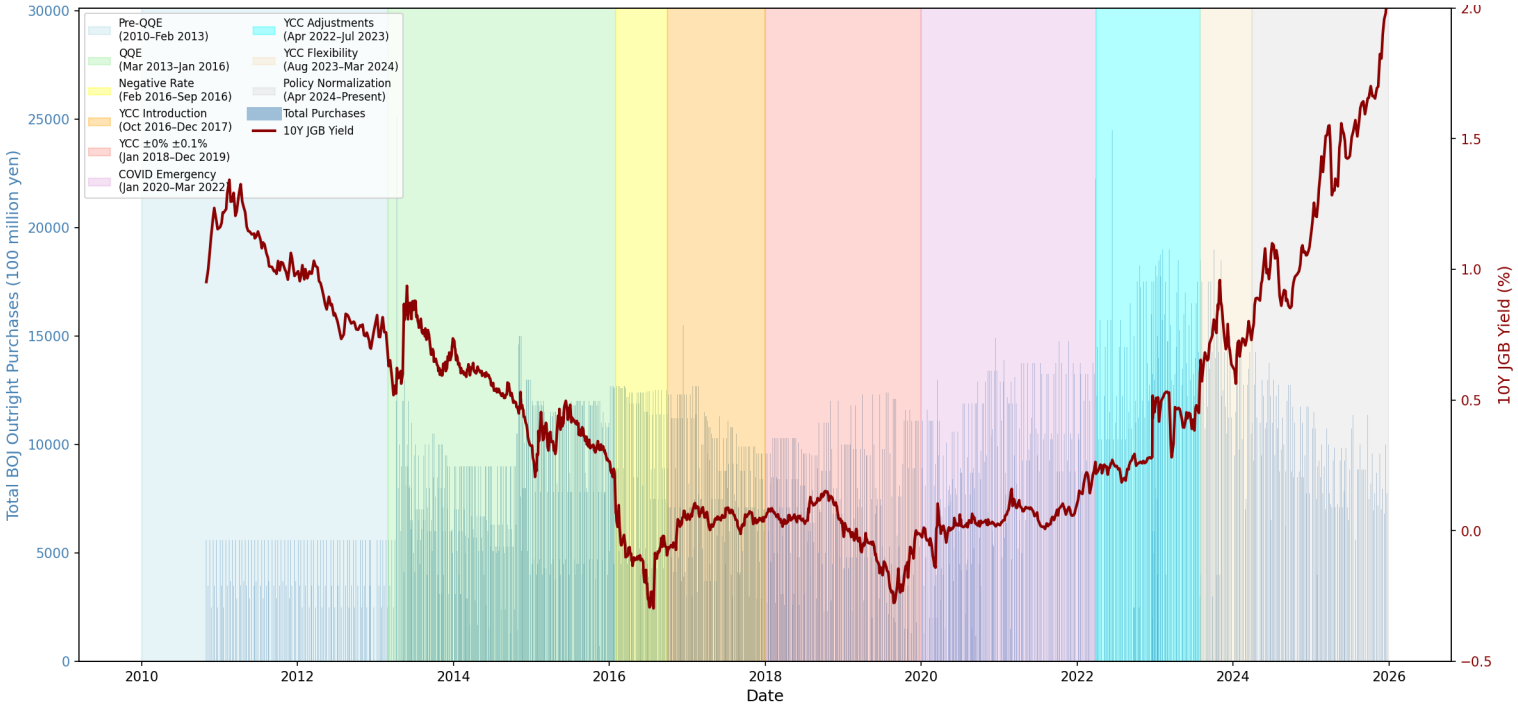
2013-2015: QQE shock. Following the April 2013 QQQE announcement, total purchases surged to over 7 trillion yen monthly, and the 10Y yield dropped sharply from around 0.8% to below 0.4% by 2015. The yield curve flattened significantly, with longer maturities (20Y-30Y) compressing more than short-term yields.

Figure 1: The BOJ 2025: JGB Yield, Purchase Positions & Policy Timeline



2016-2023: YCC and NIRP era. The introduction of YCC in September 2016 explicitly capped the 10Y yield near zero (later ±0.25%, ±0.5%, etc.). Our data confirms that after 2016, the 10Y yield became remarkably stable between -0.1% and +0.2%, despite fluctuations in global rates. During this period, total BOJ purchases declined gradually from their QQE peak, stabilizing

Figure 2: 10Y JGB Yield vs Total BOJ Outright Purchases (2010-2025)



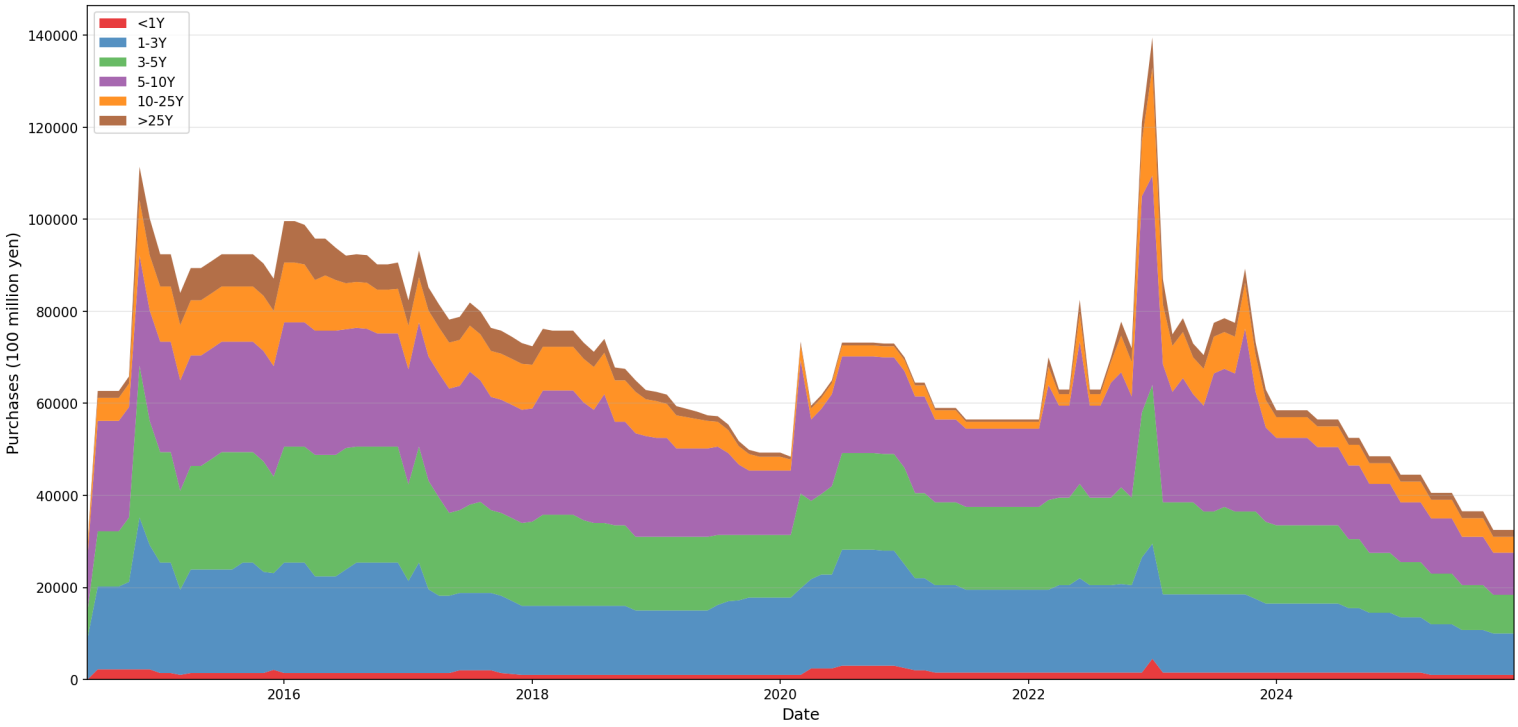
around 4-6 trillion yen per month, as the BOJ relied more on yield targeting than quantity operations.

2024-2025: Normalization and QT. Following the March 2024 exit from negative rates and YCC, the 10Y yield began a steady ascent, breaching 1% in late 2024 and reaching 2.04% by December 2025. Total purchases declined further, consistent with the QT plan announced in July 2024, falling to below 8 trillion yen monthly by late 2025.

3.2.2 Purchase Composition by Maturity Bucket

We run a simple decomposition of BOJ outright purchases by residual maturity ($< 1Y$, 1-3Y, 3-5Y, 5-10Y, 10-25Y, $> 25Y$) from 2010 to 2025. We found that before 2016 (pre-YCC), the BOJ purchased across all maturities, with a notable presence in the 5-10Y and 10-25Y buckets, aiming to depress the entire curve. From 2016 to 2023 (YCC active), purchases became increasingly concentrated in the 5-10Y bucket (the YCC operational target), often exceeding 4 trillion yen monthly. Other buckets saw reduced volumes or became residual. From 2024 to 2025 (post-YCC), the distribution broadens again as the BOJ normalizes its balance sheet, with no single dominant maturity. The 5-10Y share declines, while medium-term buckets (1-5Y) regain importance for managing policy rate expectations.

Figure 3: Monthly BOJ Outright Purchases by Maturity Bucket (2010-2025)



3.2.3 Relationship Between Yields and BOJ Positions

The table result reveals a striking evolution in the BOJ's reaction function. During the early QE period (2010-2012), the correlation is near zero and statistically insignificant, suggesting that purchase volumes were not systematically tied to contemporaneous yield movements. In the QQE years (2013-2015), a significant negative correlation emerges, consistent with the BOJ actively increasing purchases when yields threatened to rise: a conventional easing response aimed at compressing term premia. Under the YCC regime (2016-2023), the correlation turns positive, reflecting a fundamental shift: with the 10-year yield pinned by a rate cap, the BOJ could reduce purchase volumes without allowing yields to rise, relying instead on forward guidance and credibility. Most strikingly, during the normalization phase (2024-2025), the correlation becomes strongly negative again. This reversal, however, has a different interpretation: the BOJ is now deliberately reducing purchases (quantitative tightening) to allow yields to rise in a controlled manner, consistent with an exit from unconventional policy. The negative sign could indicate that higher yields accompany lower purchase volumes, a market-driven relationship typical of normal monetary policy transmission.

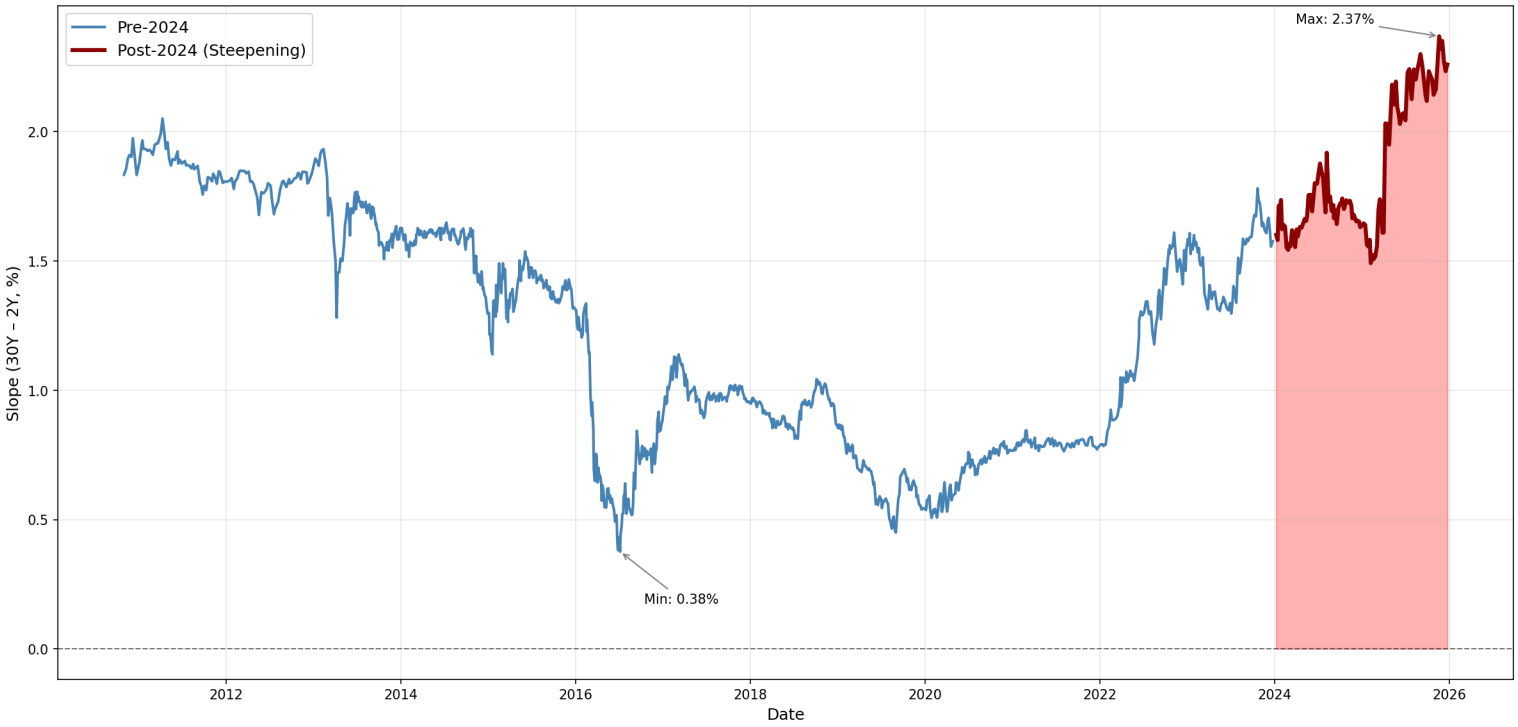
Table 1: Correlation between 10Y JGB Yield and Total Purchases

Period	Observations	Pearson		Spearman	
		r	p-value	ρ	p-value
2010-2012	104	0.046	0.643	0.066	0.508
2013-2015	339	-0.226	$< 0.001^{***}$	-0.242	$< 0.001^{***}$
2016-2023	766	0.357	$< 0.001^{***}$	0.279	$< 0.001^{***}$
2024-2025	114	-0.519	$< 0.001^{***}$	-0.502	$< 0.001^{***}$

Source: Calculations by authors

3.2.4 Yield Curve Steepening and Policy Regime Shifts

Figure 4: 2Y-30Y JGB Yield Slope (2010-2025)



Finally, we examine the slope between 2Y and 30Y yields (a proxy for term premium and policy expectations). The 30Y yield rose by over 100 basis points during 2025 alone (from 2.33% to 3.36%), while the 2Y yield increased by only 38 bps. This dramatic steepening, unprecedented in the post-2010 period, reflects short-end anchoring by policy rate expectations (BOJ's gradual hikes priced in), long-end pressure from QT and reduced BOJ demand for long-dated JGBs, and global spillovers as U.S. Treasury yields also steepened.

This pattern differs sharply from the QQE era (2013-2015), when the entire curve compressed, and from the YCC era (2016-2023), when the curve was artificially flattened. It demonstrates that the BOJ's policy stance has fundamentally shifted from suppressing the yield curve to normalizing its shape - a transition our subsequent textual analysis aims to capture in central bank communication.

3.3 Predictability Baseline: From Levels to Changes

Section 3.2.3 documented that the *level* correlation between the 10Y yield and total purchases varies across regimes, ranging from −0.67 during QQE to +0.54 during normalization. However, for the event-study analysis in Section 4.4 — which regresses daily yield *changes* around BOJ speeches onto NLP-derived sentiment scores — the relevant question is not whether levels co-move, but whether *changes* in yields and purchases are predictable from observable data. This section characterizes the statistical properties of yield changes and purchase changes, establishing a quantitative null hypothesis against which the NLP models of Section 4 will be evaluated.

3.3.1 Unit Roots and Fat Tails in Yield Changes

We first verify that yield changes are the appropriate dependent variable by testing for unit roots. Table 2 reports the Augmented Dickey–Fuller (ADF) test for both yield levels and first differences at the three maturities selected in Section 4.4.

Table 2: Augmented Dickey–Fuller Test for JGB Yields

Maturity	**Series**	**ADF Statistic**	***p*-value**
2Y	Level	3.179	1.000
2Y	ΔYield	−21.154	<0.001
10Y	Level	0.931	0.994
10Y	ΔYield	−19.972	<0.001
30Y	Level	0.961	0.994
30Y	ΔYield	−19.420	<0.001

Source: Calculations by authors

Yield levels are strongly non-stationary ($p \approx 1.0$), whereas first differences are strongly stationary ($p < 0.001$). This I(1) behavior confirms that yields follow a unit-root process and that first differences — the variable used in the event study of Section 4.4 — are the appropriate transformation.

Table 3 reports the distributional properties of these daily yield changes.

Table 3: Distributional Statistics of Daily JGB Yield Changes (basis points)

Maturity	Mean	Std	Skewness	Excess Kurtosis	N
2Y	−0.08	2.83	0.07	49.35	12,785
10Y	−0.03	3.49	0.20	11.45	9,786
30Y	0.01	2.56	0.13	9.04	6,470

Source: Calculations by authors

Mean daily changes are economically negligible (less than 0.1 bp), confirming the absence of systematic drift. The standard deviations of 2.5–3.5 bp establish the scale of daily noise that any NLP-based predictor must overcome. Most critically, the distributions exhibit extreme leptokurtosis: the 2Y maturity has an excess kurtosis of 49.35, the 10Y of 11.45, and the 30Y of 9.04 (all strongly rejecting normality via the Jarque–Bera test, $p < 10^{-4}$). Fat tails imply that large yield moves occur far more frequently than a Gaussian model would predict, concentrating economic significance in tail events that are inherently difficult to forecast.

3.3.2 Regime-Dependent Volatility

The unconditional standard deviations in Table 3 mask substantial heterogeneity. Table 4 decomposes daily yield volatility across the four policy regimes identified in Section 3.2.1.

Table 4: Daily Yield Change Volatility by Policy Regime (basis points)

Maturity	2010–2012	2013–2015	2016–2023	2024–2025
2Y	0.47	0.47	1.03	1.97
10Y	1.87	1.93	1.65	2.87
30Y	1.94	2.42	2.27	3.27

Source: Calculations by authors

The 2Y yield volatility increased fourfold from the pre-QQE era (0.47 bp) to the normalization period (1.97 bp), consistent with the transition from a near-zero policy rate environment — where the short rate was effectively anchored — to an active tightening cycle in which each data release and speech carries implications for the rate path. At the long end, the QT period (2024–2025) exhibits the highest volatility across the entire sample: the 30Y daily volatility of 3.27 bp exceeds the QQE-era peak (2.42 bp). Combined with the yield curve steepening documented in Section 3.2.4, this indicates that the normalization period is characterized by both higher yield levels and greater daily uncertainty.

This regime-dependent volatility has a direct methodological implication: any model that pools observations across all regimes — including the OLS event study in Section 4.4 — faces a heteroscedasticity problem. A sentiment signal that may be detectable within the low-volatility YCC regime could be swamped by noise in the high-volatility QT period.

3.3.3 Autocorrelation and Random Walk Behavior

Under the Efficient Market Hypothesis, asset price changes should be unpredictable from their own history. Table 5 tests this proposition for both purchase changes and yield changes.

Table 5: Autocorrelation Coefficients of Purchase and Yield Changes

Series	Lag 1	Lag 2	Lag 3	Lag 5
ΔTotal Purchases	−0.618	0.117	0.027	0.017
Δ2Y Yield	0.313	0.104	0.064	0.083
Δ10Y Yield	0.075	0.013	0.008	0.028
Δ30Y Yield	0.100	0.028	0.017	−0.020

Source: Calculations by authors

Purchase changes. The strong negative lag-1 autocorrelation (−0.618) is a mechanical artifact of the BOJ's operational calendar. The BOJ rotates across maturity buckets within each week: a large 5–10Y operation (¥3,250 billion) is typically followed by a smaller <1Y or 1–3Y operation (¥1,000–3,000 billion), generating an alternating pattern in the total amount. This oscillation carries no economic content and vanishes at the bi-weekly frequency, as evidenced by the near-zero autocorrelation at lag 3 (0.027) and lag 5 (0.017).

Yield changes. The 2Y maturity exhibits a moderate lag-1 autocorrelation of 0.313, consistent with short-term momentum in policy rate expectations: when the market begins to price in a rate hike, the adjustment tends to persist over several days. By contrast, the 10Y and 30Y yield changes have lag-1 autocorrelations of only 0.075 and 0.100, with all higher-order lags economically negligible. This near-zero serial dependence at the long end is consistent with a random walk process — the benchmark prediction under market efficiency.

3.3.4 From Levels to Changes: Purchase–Yield Predictability

Section 3.2.3 established that the *level* correlation between yields and purchases varies meaningfully across regimes (−0.67 to +0.54), reflecting the

shifting causal mechanism from active easing to normalization. We now ask a sharper question: do *changes* in one series predict *changes* in the other?

Contemporaneous change correlation. Table 6 reports the Pearson correlation between ΔTotal purchases and ΔYield, measured at the frequency of individual purchase operations ($N = 1{,}322$).

Table 6: Correlation of Contemporaneous Changes: ΔTotal Purchases vs. ΔJGB Yields

Pair	**Pearson** r	p**-value**
ΔTotal vs. Δ2Y	0.003	0.921
ΔTotal vs. Δ10Y	0.012	0.671
ΔTotal vs. Δ30Y	−0.051	0.063

Source: Calculations by authors

In contrast to the strong level correlations in Table 1, the change correlations are uniformly indistinguishable from zero. The strongest result, ΔTotal vs. Δ30Y ($r = -0.051$, $p = 0.063$), is directionally consistent with the supply channel — more purchases compress yields — but economically negligible. This divergence between level and change correlations is characteristic of cointegrated series: yields and purchases share common regime-level trends, but their high-frequency innovations are driven by independent factors.

Lead–lag predictability. Table 7 tests whether yield levels predict subsequent purchase changes, completing the predictability assessment.

Table 7: Lead–Lag: Yield Level(t) vs. ΔTotal Purchases(t+1)

Predictor	**Pearson** r	p**-value**
2Y Yield(t)	−0.001	0.984
10Y Yield(t)	−0.001	0.969
30Y Yield(t)	−0.002	0.940

Source: Calculations by authors

All lead–lag correlations are virtually zero ($|r| < 0.003$, all $p > 0.94$). Yield levels carry no detectable information about the direction or magnitude of the next purchase operation. This is consistent with the pre-announced nature of the BOJ's purchase schedule: the QT plan specifies approximate quarterly reduction targets, leaving minimal scope for yield-contingent adjustments at the individual operation level.

3.3.5 Summary and Transition to the NLP Analysis

The results of this section define three properties of the JGB market that constrain any attempt to predict yields or purchases from external information:

1. **Yield changes are approximately unpredictable.** Long-end yields follow a near-random walk with fat tails (excess kurtosis 9–49), implying a very low signal-to-noise ratio for any point forecast.

2. **Volatility is regime-dependent.** The fourfold increase in 2Y volatility from the YCC era to QT, and the peak in 30Y volatility during normalization, mean that a model pooling across regimes faces structural heteroscedasticity.

3. **Changes in purchases and yields are independent.** Despite strong *level* co-movement across regimes (Section 3.2.3), the *changes* are uncorrelated both contemporaneously and in lead–lag structure. Observable quantitative data contains no exploitable short-horizon predictive information.

This establishes the null hypothesis for Section 4: if BOJ speeches contained forward-looking information not yet reflected in market prices, we would expect the NLP-derived hawkish-dovish scores to exhibit predictive power for yield changes *beyond* the zero-correlation baseline established here. As we will show, the FinBERT-based scores (Section 4.2) yield correlations below $|r| = 0.08$ with yield changes (Table 7), and regression-based event studies (Section 4.4) produce negative out-of-sample R^2 across all maturities. Combined with the unpredictability baseline of this section, these findings provide converging evidence for the semi-strong form of the Efficient Market Hypothesis in the JGB market.

4 NLP Analysis

We collected speeches and announcements text data of the BOJ from 1996 to 2025 from the BIS website. After our double-check by manually searching on the BOJ's official website and comparing the search results with the text data from BIS website, we verified that the BIS data is directly taken from the BOJ official website so can be trusted and used for our research.

Of all 20116 lines of text, we selected 807 lines of text we think are related to the BOJ based on key words. The length (word count) of the text (full content of speeches or announcements) ranges from 159 to 11460, with mean equal to 3762.19 and median equal to 3860.

Table 8: Distribution of text length

Category	Number
count	807
mean	3762.19
std	2170.72
min	159
25%	1751
50%	3860
75%	5240.5
max	11460

4.1 Future Oriented Detection

Our first analysis is to detect if each speech is future-oriented. In contrast to the alternative of simply prompting a generative model such as GPT with a question like "Is this speech future-oriented?", our approach intentionally avoids treating the task as a black-box text completion problem. A single GPT prompt would rely on the model's opaque next-token prediction, which can produce a fluent but ungrounded answer that reflects the model's training biases rather than the evidence in the speech itself, and it provides no audit trail for the decision. We instead decompose the detection into two transparent, rule-integrated stages-an embedding-based cosine similarity with explicit future/non-future examples and a question-answering model that extracts spans containing future indicators, so that every classification is traceable to specific textual evidence, the confidence scores have an objective basis, and the entire pipeline remains interpretable and scientifically auditable.

For that we use two approaches (models). Our first approach is using albert-xxlarge-v2 QA model with all-MiniLM-L6-v2 model as embedding. Our

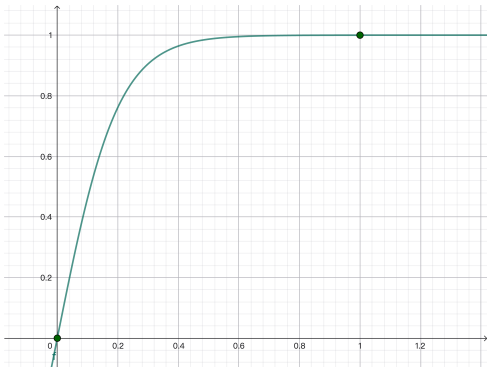
second approach is zero-shot classification with distilbart-mnli-12-1 model.

The primary results are: The first QA model detected 360/807 future-oriented texts; the second zero-shot model detected 412/807 future-oriented texts. Two results diverge and converge around 50%, and confidence levels for two models are not significant enough. To solve the problem, we introduce an improvement on the first QA model.

The original methodology for the first model consists of 2 layers: the first layer is embedding cosine similarity classification. First we design two sets of examples used in embedding, $future\_examples$ and $non\_future\_examples$, whose details can be seen in appendix. Then we use these examples to calculate cosine similarity respectively, $future\_sim$ and non_future_sim, with $diff\_score = future\_sim - non\_future\_sim$. Then based on the sign of $diff\_score$, we determine the label of first embedding layer, with confidence level calculated using sigmoid-like scaling:

$$confidence = \frac{2}{1 + e^{-10|diff\_score|}} - 1$$

Figure 5: Sigmoid Function



The second layer is QA classification with albert-xxlarge-v2 model. Similarly we design a question list as part of the input for the QA model. Then to analyze the output, we collect some future indicator words.

Then we calculate the percentage of future indicators appearing in output in all future indicators, as yes_ratio. If $yes\_ratio \geq 0.4$, we identify the result as "yes", with $confidence = 0.6 + 0.4\ yes\_ratio$; if $yes\_ratio \leq 0.2$, we identify the result as "no", with $confidence = 0.6 + 0.4\ (1 - yes\_ratio)$; if $0.2 < yes\_ratio < 0.4$, we identify the result as "unclear", with $confidence = 0.5$.

Then we combine the two layers: if the confidence level of first layer is already larger than 0.75, then we take the result of first layer directly; otherwise, we turn to the second layer: if the result of second layer is "unclear", we take the result of first layer; otherwise, if the results of first layer and second layer

Table 9: Future indicators given questions

QA question list	**Future indicators**
"What will happen in the future?"	'will', 'shall', 'expect', 'anticipate',
"What are the future plans?"	'plan', 'future', 'next', 'coming',
"What is expected next?"	'upcoming', 'ahead', 'project',
"What are the upcoming developments?"	'forecast', 'predict', 'outlook',
"What is the outlook?"	'going', 'gonna'

Table 10: Criteria and Confidence

yes_ratio	**Result**	**Confidence**
≥ 0.4	"yes"	$0.6 + 0.4\ yes\_ratio$
≤ 0.2	"no"	$0.6 + 0.4\ (1 - yes\_ratio)$
$(0.2, 0.4)$	"unclear"	0.5

are the same, then we take the result and increase the confidence level by 0.1, i.e. $confidence_{new} = min(confidence_{old} + 0.1, 1)$; if the results are different, we take the result with higher confidence level, i.e.

$$result = result_{\theta} \underset{\theta \in \{layer_1, layer_2\}}{\arg\max} \{confidence_{\theta}\}$$

The flowchart can be seen in appendix.

To improve the original model, we redesign the confidence function to what it's like now. And we add the function of locating the place where the model output the answer based, so we can monitor the performance of the model and make improvement manually accordingly. The new result is quite different from old result, and move closer to the result of second zero-shot model: the new model detected 463/807 future-oriented texts.

4.2 Future Directions: From Classification to Policy Stance Index

While the binary future-orientation classification provides a preliminary filter, it lacks the granularity required for financial application. Our next step involves constructing a BOJ Policy Stance Index (BOJ-PSI) using a fine-tuned Financial PhraseBank model. By regressing the NLP-derived hawkish/dovish scores against the JGB purchase anomalies identified in Section 3.1.2 (The July Anomaly), we aim to test the hypothesis that BOJ verbal interventions precede and predict actual position adjustments. Furthermore, we will employ

the frontier LLM model to conduct a narrative analysis of the dissenting opinions in the September 2025 MPM to quantify the degree of Policy Committee Fragmentation, a variable often leading to increased volatility in JGB futures. This multi-layered NLP approach transforms text from a static corpus into a dynamic leading indicator for Japanese fixed income positioning.

We use *hawkish_dovish_score* to represent the result of BOJ Policy Stance Index, theoretically ranging from -1 to 1, with negative value representing dovish/easing signals (e.g., rate cuts, stimulus) and positive value representing hawkish/tightening signals (e.g., rate hikes, fighting inflation).

The *hawkish_dovish_score* is a weighted combination of *sentiment_score* and *keyword_score*, which means $hawkish\_dovish\_score = \alpha * sentiment\_score + (1 - \alpha) * keyword\_score$, with $\alpha = 0.7$.

The *sentiment_score* is directly derived from FinBERT model. We use FinBERT sentiment score because in financial context, "positive" typically means strong economy and moderately rising inflation, so tightening policy is needed (hawkish); while "negative" typically means weak economy and deflation risk, so easing policy is needed (dovish).

For *keyword_score*, we count *hawkish_count* and *dovish_count* from hawkish keywords and dovish keywords (see appendix), and the *keyword_score* is calculated with

$$keyword\_score = \frac{hawkish\_count - dovish\_count}{max(hawkish\_count + dovish\_count, 1)}$$

to normalize it to -1 to +1.

The classification is simple: based on the final score, scores greater than 0.3 are Hawkish, scores less than -0.3 are Dovish, and scores between are Neutral. And the final result is overally more dovish, with score ranging from -0.9586 to +0.6768, mean equal to -0.1546 and median equal to -0.1294.

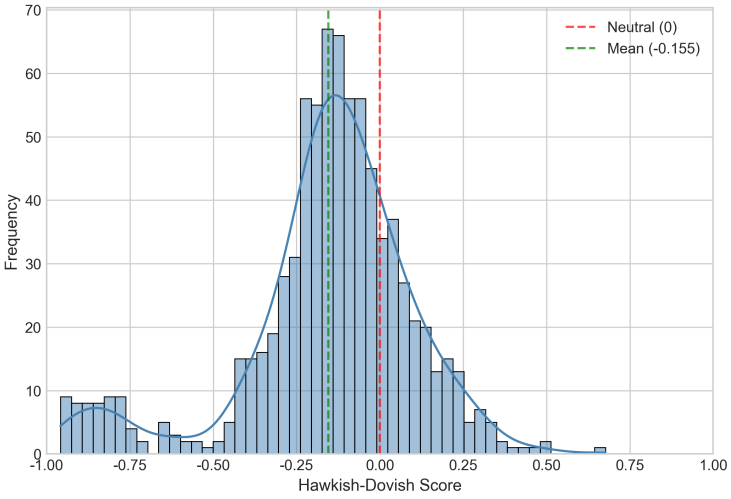
Table 11: Overall score distribution

Category	**Number**
count	807
mean	-0.1546
median	-0.1294
std	0.2623
min	-0.9586
max	0.6768

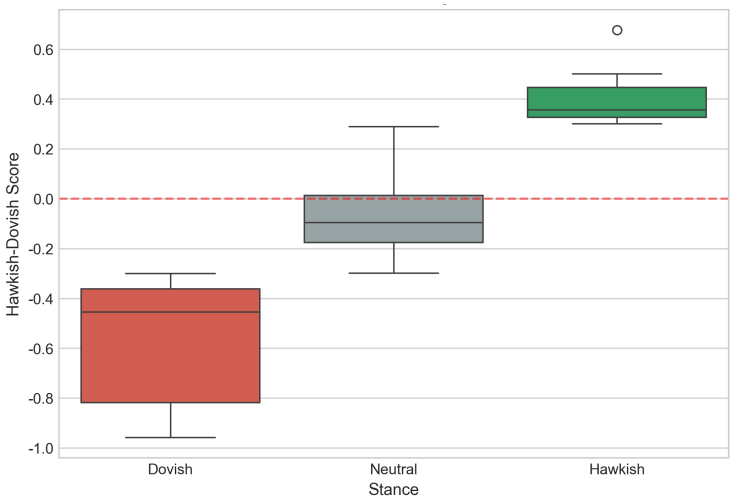
For broader distribution, most speeches (645, covering 79.9%) are more neutral, with only 147 (18.2%) more dovish and 15 (1.9%) more hawkish. The

dovish/hawkish percentage is relatively small because we set the threshold at ±0.3. And the fact that very few speeches take a tightening stance aligns with Japan's prolonged low-interest-rate environment.

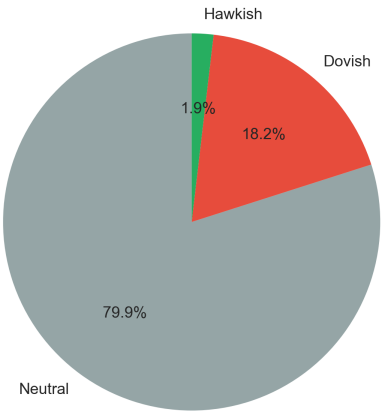
Figure 6: Hawkish-Dovish Score distribution



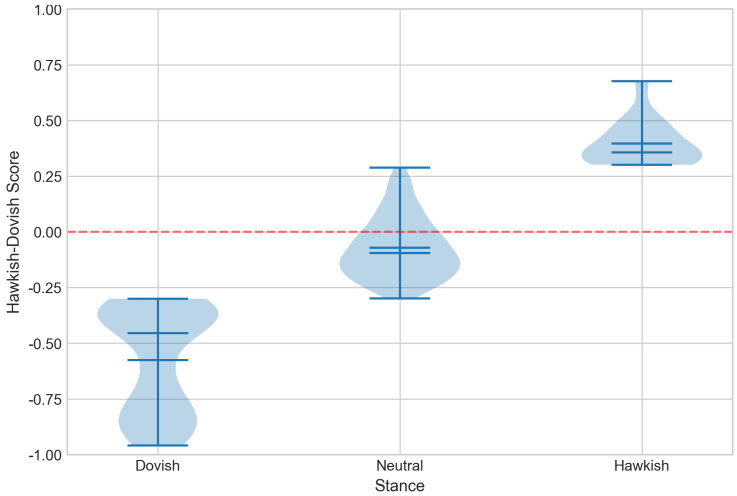
(a) Distribution of Hawkish-Dovish Scores



(b) Score Distribution by Stance



(c) Distribution of Stances

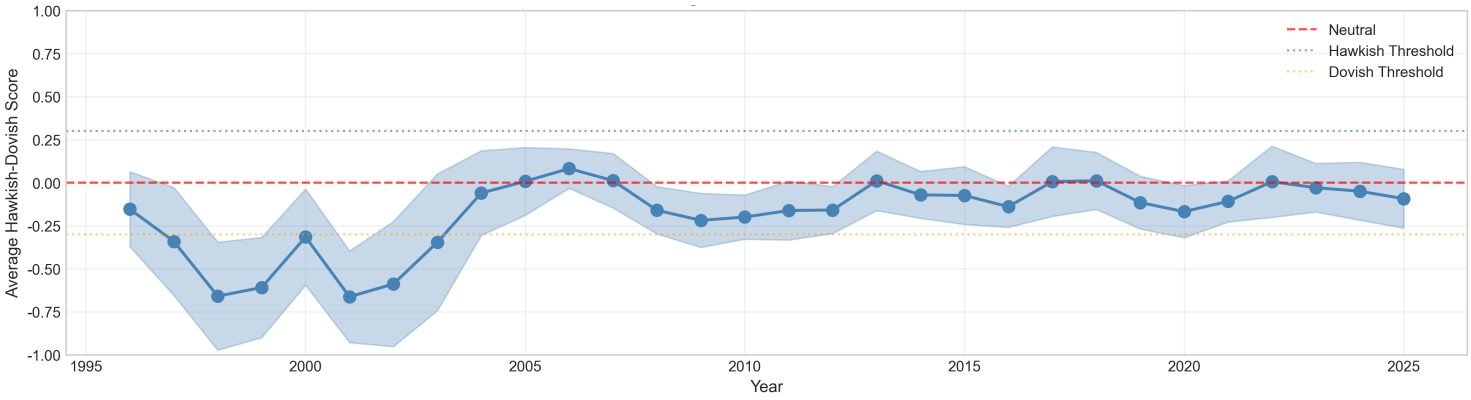


(d) Score Density by Stance

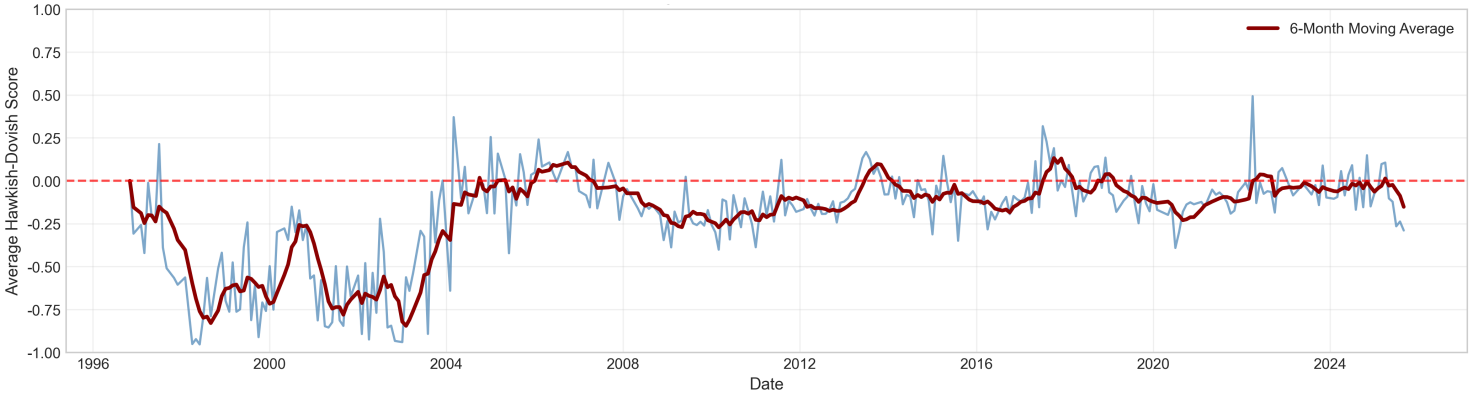
For the time trend, the model successfully detects 2 times from 1996 to 2004 when the speeches are extremely passive: after the 1998 Asian Financial Crisis, and after 2001 the bursting of dot-com bubble. At other times, the scores are more neutral or moderately dovish, with rare hawkish peaks.

The scatter plots demonstrate a relatively strong correlation between FinBERT sentiment scores and keyword-based scores. So the combined approach

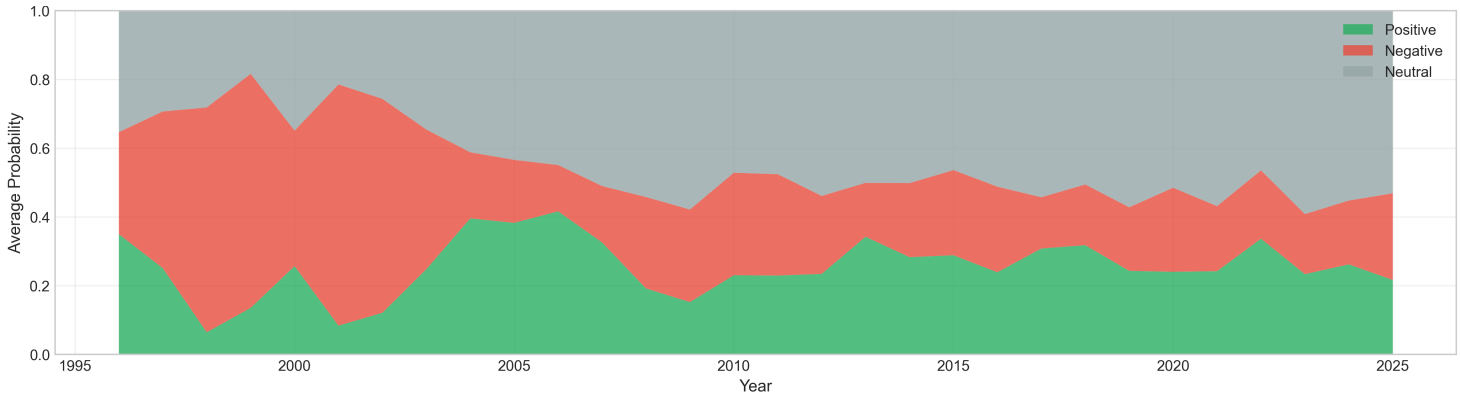
Figure 7: Hawkish-Dovish Score trend



(a) Annual Average Hawkish-Dovish Score Trend



(b) Monthly Hawkish-Dovish Score Trend

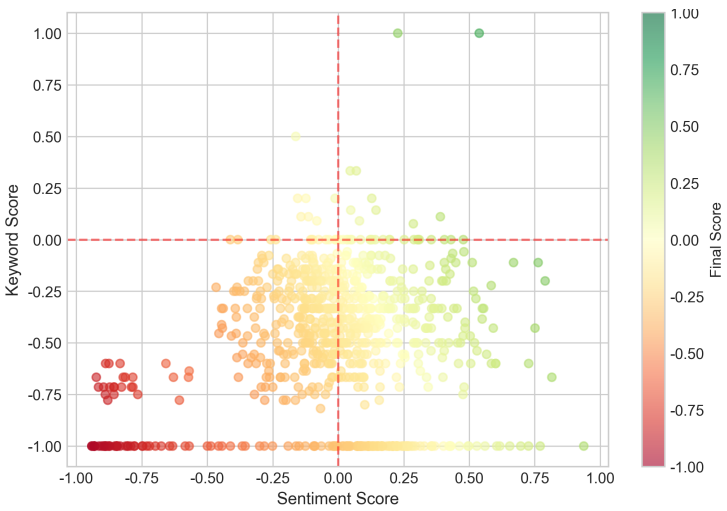


(c) Sentiment Probability Trends Over Time

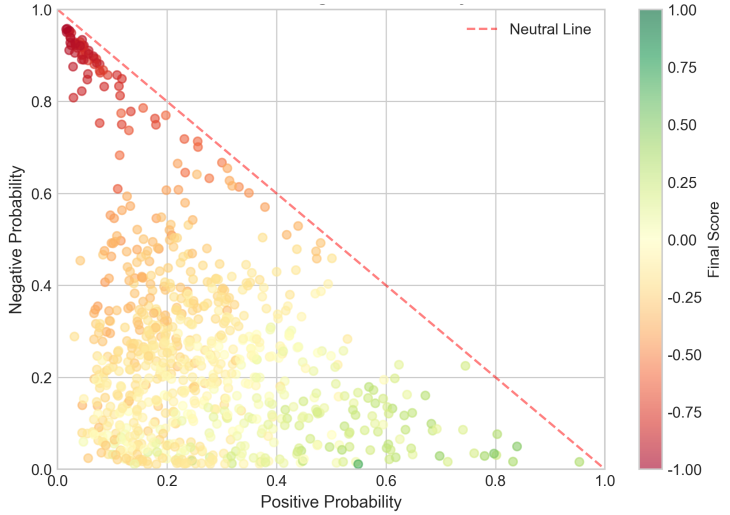
(70% sentiment + 30% keywords) produces well-calibrated results. For confidence, dovish speeches tend to have higher sentiment confidence (more decisive

negative language), while neutral speeches tend to cluster in the middle confidence range, and hawkish speeches show moderate confidence.

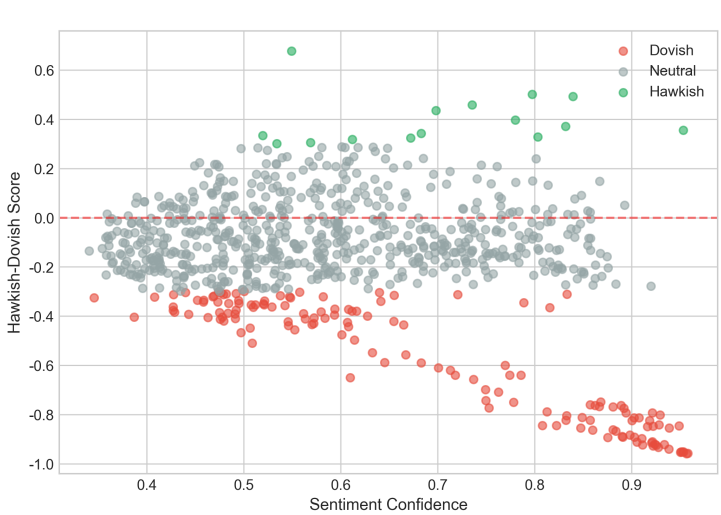
Figure 8: Hawkish-Dovish Score scattering



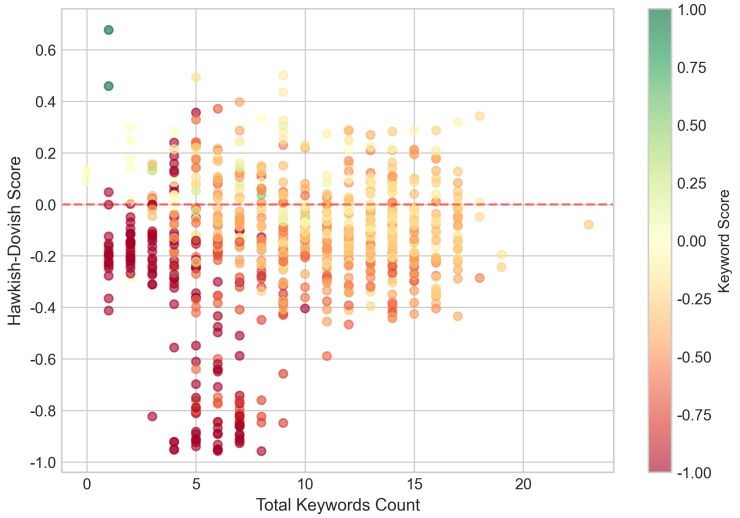
(a) Sentiment Score vs Keyword Score



(b) Positive vs Negative Probability



(c) Score vs Sentiment Confidence

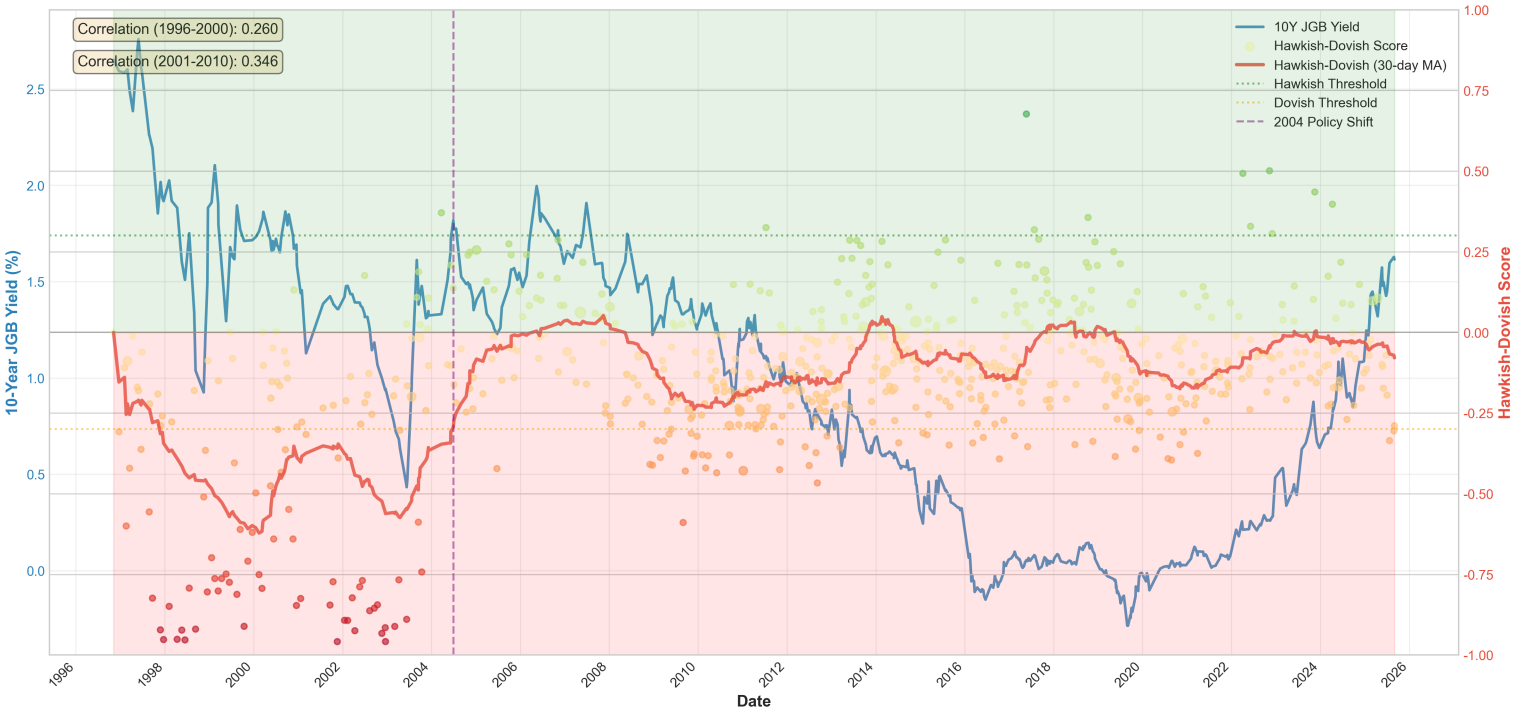


(d) Keywords Count vs Final Score

The *hawkish_dovish_score* and the 10-year JGB yield show limited positive correlation before 2010. After 2010 experiencing the "lost decade", in order to get rid of long-term deflation, the BOJ launched QQE (Quantitative and Qualitative Monetary Easing) in 2013, pushing the yield further lower, even negative in certain times. The *hawkish_dovish_score* fails to capture this big U-shape yield decreasing and rebounding trend, due to the limitations of the amount

of information the speeches of BOJ can contain. However, it does qualitatively reflect the yield changing trend after 2010 as the *hawkish_dovish_score* remained neutral or dovish.

Figure 9: Hawkish-Dovish Score vs 10-year JGB Yield



There is an interesting inference we can draw from the result: the BOJ exhibits strong asymmetry, much more willing to signal dovishness than hawkishness, which reflects structural deflationary pressures in Japan.

However, it is important to note that the FinBERT model was trained on general financial sentiment, not specifically on central bank communications, so the mapping from "positive sentiment" to "hawkish" assumes positive economic news implies tightening bias, which may not always hold. Besides, the Japanese-to-English translation may also lose minor differences in the original speeches.

4.3 Topic Modeling over Time

Given the BOJ's tactical policy transformations over three decades, from ZIRP (zero interest-rate policy) to QQE, then to YCC, and most recently to normalization, we assume that the linguistic focus of BOJ communications has undergone corresponding structural shifts. To capture this evolution, we apply BERTopic to the 807 texts spanning 1996-2025, segmenting the data into

five-year windows. This approach dynamically clusters topics and tracks their prevalence over time, enabling us to trace how themes such as "deflation risk", "balance sheet expansion", "exchange rate volatility" and "wage-driven inflation" rise and fall in prominence.

Topic 0 covers most documents, while topic 1 covers fewer documents. It can be seen the attention of the BOJ has moved from investment demand, to banks easing, financial crisis, inflation expectations, labor and wages. This binary structure can highlight both the evolving primary focus of BOJ communication and the most salient complementary concern in each era.

Table 12: Topic Analysis Results by Time Period

Time Period	Total Docs	Topic	Doc Count	Keywords
1996–2000	79	Topic 0	55	demand, investment, private, reflecting
		Topic 1	24	central, new, institutions, recovery
2001–2005	101	Topic 0	66	banks, current, easing, institutions
		Topic 1	35	continue, bonds, developments, demand
2006–2010	125	Topic 0	75	central, institutions, banks, crisis
		Topic 1	47	economies, developments, likely, outlook
2011–2015	216	Topic 0	194	inflation, economies, stability, demand
		Topic 1	22	inflation, expectations, term, inflation expectations
2016–2020	174	Topic 0	135	inflation, central, banks, global
		Topic 1	32	labor, outlook, effects, inflation
2021–2025	112	Topic 0	70	inflation, wages, labor, change
		Topic 1	21	inflation, developments, figures, outlook

4.4 LLM-as-Event-Study

Now we'd like to look for the potential link of the causal impact of BOJ communication on market pricing, the very mechanism by which words translate into the position adjustments and yield movements documented at length in Section 3. We therefore propose an event-study framework that exploits high-frequency JGB yield changes around BOJ speeches and announcements to construct a quantitative market reaction function, and then interprets the magnitude and direction of these reactions through the textual features extracted by our NLP pipeline.

The choice of maturity is guided by the purchase regime analysis in Section 3. We select three maturity points that jointly capture the yield curve

response: first the 2-year JGB yield, which predominantly reflects pure monetary policy rate expectations of the market; second the 10-year JGB yield, the long-standing benchmark and the former operational target of Yield Curve Control, which anchors the core 5-10Y purchase bucket that the BOJ actively manages; and third the 30-year JGB yield, which reflects the quantitative tightening expectations, and fiscal risk perceptions that Section 3.1.4 shows to have been particularly sensitive to the 2025 QT trajectory.

In practice, for each identified speech, we compute the change in each yield and the slope over a tight window (e.g. 1 day) surrounding the release. The core explanatory variable will be a Text-based Surprise Index derived from NLP model. By regressing yield curve changes onto these NLP scores while controlling for macroeconomic surprises, we can test whether the textual content of BOJ communications conveys incremental information beyond the policy action itself.

However, the correlation coefficients of Sentiment Score and Hawkish-Dovish Score by FinBERT with JGB yield changes are all close to zero. We fail to capture the influence of BOJ speeches on market price shocks, possibly because the sentiment-based score cannot well represent the key information contained in speeches that truly influence the yields.

Table 13: Correlation coefficient of sentiment-based scores

Variables	**Correlation Coefficient**
2Y_change vs hawkish_dovish_score	0.0069
2Y_change vs sentiment_score	0.0149
10Y_change vs hawkish_dovish_score	0.0458
10Y_change vs sentiment_score	0.0438
30Y_change vs hawkish_dovish_score	0.0776
30Y_change vs sentiment_score	0.0771

So we turn to the latest AI models for prediction. The following is the performance of predictions by Claude (Anthropic, 2024). The performance of the prediction of AI is extraordinary, and the detailed graphs can be seen in appendix. On the one hand it may reflect the great power of the latest AI models of comprehensively capturing sentiment and accurately making predictions. On the other hand we doubt that Claude might have secretly utilized the yield data in that period in its knowledge database or searched online because after all the research agent of Claude is a complex black box model.

This result highlights a critical pitfall in the application of large language models to financial event studies: Claude was trained on a corpus that, in all likelihood, already contains the very yield data we are asking it to predict, thereby giving the model perfect hindsight over the evaluation period. In

effect, we are asking a model that "knows the answer" to guess, rather than assessing its genuine point-in-time forecasting ability. The strikingly high in-sample correlations are therefore not evidence of true predictive power but a manifestation of forward looking; for a fund manager, treating such backtested figures as a guide to future performance would entail significant risk. This case serves as a cautionary illustration that, without strict controls to ensure that LLMs are prompted with only information available at the time of the event, their impressive in-sample results may create a dangerous illusion of predictability.

Table 14: Prediction Result of Different Maturities

Maturity	**Pearson** r	**MAE**	**RMSE**	**Mean Error**
2Y	0.768	0.0067	0.0126	−0.00001
10Y	0.700	0.0141	0.0233	−0.00009
30Y	0.709	0.0152	0.0233	−0.00053

On the other hand, we try to analyze if certain content of BOJ speeches (i.e. Hawkish-Dovish Score) can influence the change of BOJ position (Outright Purchases of JGBs). So we calculated the correlation coefficients of Hawkish-Dovish Score with Outright Purchases of JGBs on a weekly basis from 2010 to 2025 (since the dates do not perfectly match).

Table 15: Correlation of Hawkish-Dovish Score and BOJ Positions

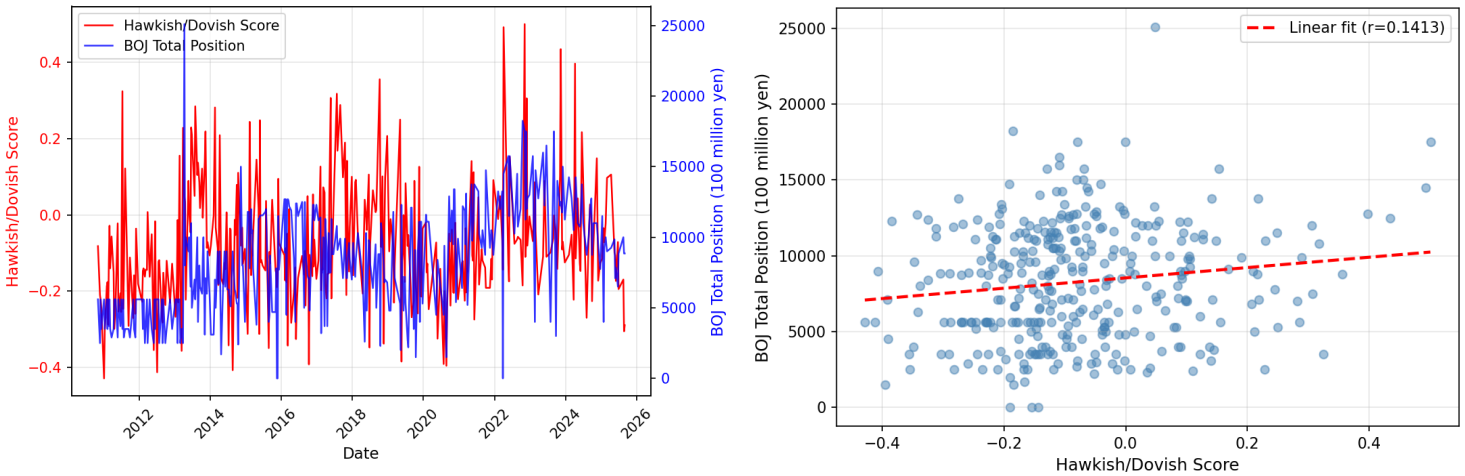
Method	**Correlation**	p**-value**
Pearson (weekly alignment)	$r = 0.1413$	0.0080
Spearman (weekly alignment)	$\rho = 0.1202$	0.0243
Merge_asof (last 7 days)	$r = 0.1019$	0.0209

From the result we can only see weak correlation. To test if the result is due to wrong regression method (the subject we analyze is about the influence of BOJ speeches (its Hawkish-Dovish Score) on Outright Purchases of JGBs, which is influenced by lots of factors, so maybe by using the difference/change of total purchase and the difference/change of Hawkish-Dovish Score we can get better result),we calculated another two correlation coefficients:

From the new result we confirmed that the Hawkish-Dovish Score generated by FinBERT sentiment score and keyword score has little correlation with BOJ Positions.

Similarly, we use the predicted factor of Claude $purch\_Total\_pred$ to calculate the correlation coefficients, and the result also turns out to be uncorrelated.

Figure 10: Correlation of Hawkish-Dovish Score and BOJ Positions



(a) Hawkish-Dovish Score vs BOJ Total Position

(b) Scatter: Hawkish-Dovish vs Total Position

Table 16: All three correlation of Hawkish-Dovish Score and BOJ Positions

Analysis	**Pearson**		**Spearman**	
	r	p-value	ρ	p-value
Hawkish/Dovish vs Total	0.1413	0.0080	0.1202	0.0243
Hawkish/Dovish vs ΔTotal	0.0133	0.8046	-0.0151	0.7781
ΔHawkish/Dovish vs ΔTotal	-0.0026	0.9608	-0.0436	0.4161

Table 17: Correlation of Claude predictions and BOJ Positions

Variables	**Pearson** r	p**-value**
purch_Total_pred vs Total	-0.0968	0.1876
purch_Total_pred vs ΔTotal	-0.1001	0.1727

The reason may be the following: First, the market might have anticipated the speeches as the BOJ communication is gradual, so market might have priced in expectation before the speech date. Second, text doesn't necessarily mean surprise. Only unexpected content moves prices. Our sentiment score captures overall tone, not the deviation from market expectations. Third, the outright purchases of BOJ follow a pre-announced schedule, so changes are mainly driven by bond supply/demand and other mechanics, and aaily yield moves are dominated by global factors, not single speeches.

So we can draw the conclusion that in the specific area of predicting yield of JGBs and positions of BOJ, the market is efficient: the EMH holds.

4.5 Counterfactual Analysis with Frontier LLM

The classification methods in previous sections are fundamentally supervised and feature-driven. However central bank communication can be characterized by deliberate ambiguity and intertextual references that a simple label cannot represent, while frontier LLMs can perform counterfactual reasoning, narrative decomposition, and cross-document synthesis when prompted with carefully designed analytical tasks. So we use these capabilities to deepen our understanding of BOJ policy signals and their market impact.

4.5.1 Market Preparedness and Counterfactual Expectation Assessment

The September 2025 Monetary Policy Meeting (MPM) presents an ideal test case for the counterfactual reasoning framework. As documented in Section 3.1.3, the Bank of Japan implemented its deepest across-the-board reduction of JGB purchases in early October 2025, immediately following a meeting at which two Policy Board members (Takata and Tamura) dissented in favor of an immediate rate hike-the first dual dissent in eight years. Did the BOJ's own communication adequately prepare the market for such a vigorous normalization step, or did the depth of the purchase cuts come as a surprise? To answer this, we apply Deepseek-v4 (Guo *et al.*, 2025) to the pre-meeting speech by Deputy Governor Himino (September 2), the post-meeting statement (September 19), and, for completeness, the later speech by Board Member Noguchi (September 29), which articulates the dovish end of the internal debate.

We extract the full English texts of the three documents and provide them to the LLM with the following structured prompt:

Risk interpretation: "Based on these texts, how does the BOJ perceive the risk of falling behind the curve versus triggering market disruption? Quote the specific phrases that support your assessment."

Action likelihood: "Assume you are a fixed-income strategist in early

September 2025 with access only to these communications (Himino speech and any preceding public signals). After the September MPM statement is released, quantify the implied probability that the BOJ will implement a purchase reduction larger than ¥50 billion per bucket in October. Explain your reasoning step by step."

Counterfactual prompt: "Now rewrite the September MPM statement to make a deep October reduction fully anticipated without changing the actual policy decision (i.e., still holding the policy rate at 0.5%). Identify the exact sentences you would modify and why."

When fed the September 2 Himino speech, the model flags a clear asymmetry in the risk assessment. Himino stresses that "the risk of a larger-than-expected impact [from tariffs] may deserve greater attention" and that "if trade policy shocks cause a global economic slowdown, this would depress crude oil and other commodity prices"-language that, in isolation, could justify caution. However, the LLM also identifies a persistent hawkish undercurrent: Himino explicitly states that "if the baseline scenario ... is realized, it would be appropriate for the Bank to continue, in accordance with improvement in economic activity and prices, to raise the policy interest rate and adjust the degree of monetary accommodation." The model rates the overall tone as cautiously leaning toward further normalization, with the balance tilted only slightly toward "not falling behind the curve." The speech nonetheless leaves considerable ambiguity about the pace of balance sheet shrinkage, as Himino merely notes that "it may be warranted to start exploring the JGB monthly purchase amount consistent with appropriate reserve levels"-a phrase the LLM interprets as unlikely to signal an imminent deep cut.

Crucially, it means that the deepest policy signals may reside not in the English translations but in the original Japanese linguistic register. An acute listener of the BOJ discourse in Japanese, attuned to the subtle gradations of politeness and the carefully hedged expressions that characterize formal central bank communication, would have been better positioned to perceive the underlying urgency toward normalization.

Given the Himino speech alone, the model pegs the probability of a ¥50bn per bucket reduction at around 35%. The reasoning is that while the QT trajectory was pre-announced in the July 2024 and June 2025 plans, those plans targeted a gradual reduction to ¥2 trillion monthly by spring 2027, and the most recent quarterly cut (April-June 2026) was only about ¥200 billion total-translating to roughly ¥30-40 billion per bucket. A ¥50+ billion reduction per bucket would therefore represent a front-loading of the QT path, something not explicitly flagged in the June interim assessment. When the September 19 MPM statement is appended, the model updates its estimate to 65%. The statement neither raises the policy rate nor mentions purchase cuts, but the LLM detects a subtle shift: the statement drops any forward guidance on JGB

purchases and adds, for the first time, a detailed plan for the disposal of ETFs and J-REITs, which the model interprets as a signal that the BOJ is accelerating its balance sheet normalization more broadly. The dual dissent is also read as a hawkish pressure valve: "The dissenters' public stance increases the likelihood that the majority felt compelled to offer a tangible concession in the form of faster QT." The LLM's step-by-step reasoning explicitly maps the probability jump to the combination of no rate hike, visible board fragmentation, and concrete disposal steps for risk assets.

Asked to rewrite the September 19 statement so that a deep October cut would be fully anticipated, the LLM suggests three modifications:

First addition to paragraph on economic outlook, insert "Some members noted that, given the progress in restoring price stability, the Bank could proceed with a somewhat faster reduction in its JGB holdings than the baseline plan, provided financial market conditions remained stable."

Then addition to paragraph on monetary policy guidance, add "The Bank will continue to reduce the amount of its outright JGB purchases, taking into account the need to normalize its balance sheet at a pace consistent with the evolving economic outlook."

And include explicit tying of the dissent to asset purchases: "While the majority judged it appropriate to wait for more data on trade policy effects, two members argued for an immediate rate hike, citing, among other reasons, that an accelerated reduction in JGB purchases alone could not adequately address upside price risks."

The model concludes that without such language, a market participant reading the actual statement would not have been fully prepared for the magnitude of the October reduction.

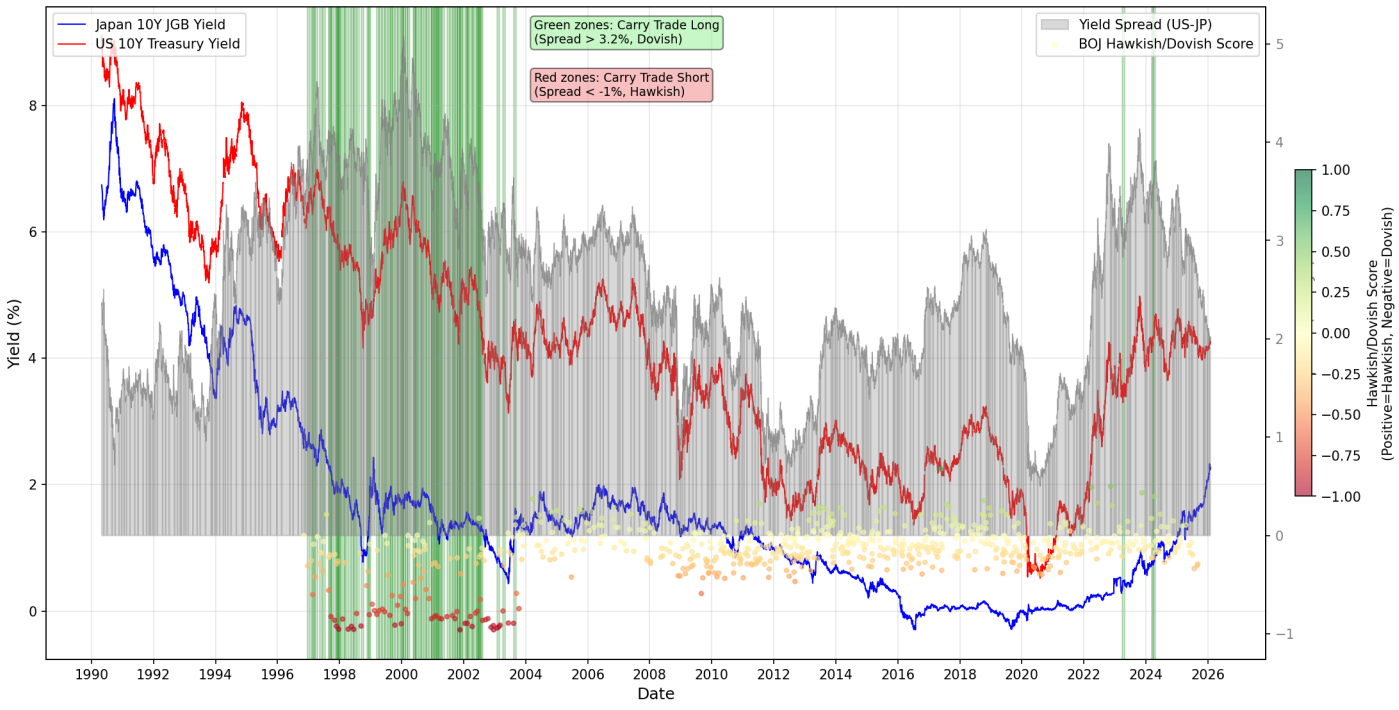
The LLM's assessment aligns well with the yield curve dynamics documented in Section 3.1.4. Before the September MPM, the 10-year JGB yield had already risen from 1.498% in June to 1.655% by September 19, indicating that the market had priced in some further tightening. However, on the day of the statement release, the 10-year yield barely moved (intraday change < 2 bps), while the 30-year yield actually dipped slightly, consistent with the market initially perceiving the "no hike" decision as dovish. It was only after the BOJ's early October purchase schedule announcement that the 10-year yield jumped to 1.72% and the 30-year yield spiked by more than 15 bps in a single week. This pattern suggests that the depth of the purchase reduction contained a material surprise, exactly as the LLM's probability assessment would predict. The NLP-derived Market Surprise Residual, the gap between the LLM's 65% ex-ante probability and the actual outcome, thus indicates a significant communication shortfall.

In conclusion, this exercise demonstrates how a frontier LLM can quantify the unspoken implications of central bank texts. The September 2025 case

shows that even when a policy action (faster QT) is directionally consistent with prior communication, its magnitude can be under-communicated, leaving markets unprepared. The counterfactual rewriting points to concrete wording adjustments that could have narrowed the expectation gap. This finding directly links to the Efficient Market Hypothesis theme of the paper: if an LLM-based strategy could have identified the higher probability of a deep cut and positioned accordingly in JGB futures, the abnormal returns after the October schedule release would constitute evidence against the semi-strong form of market efficiency. Such a test is precisely what the overarching research design sets out to conduct, and the LLM's capacity to serve as a real-time expectation engine is what makes it possible.

4.5.2 Cross-Central-Bank Carry Trade Signals

Figure 11: Yield and carry trade signals of JP/US



In this section we analyze the BOJ speeches in an international environment. For example for its influence on foreign currency exchange rate, notice that a simple increase in BOJ hawkishness can be fully offset by even larger hawkishness from the Fed, leaving the USD/JPY exchange rate unchanged. Therefore, what matters for financial markets is not only the absolute textual orientation, but also the relative linguistic distance between the BOJ's communication and that of its peers.

From the analysis we can see the overall yield spread between 10Y JGB and US Treasury bonds is always positive, and reached its lowest point during the epidemic. The apparent signals of carry trade to short JPY and long USD happen in: Asian Financial Crisis (1997-1998), dot-com bubble (2000-2002) and QQE (2012-2016) when spread was high, JPY grew weaker and policies were dovish (which was reflected in speeches). More further research can be conducted on cross-central-bank analysis, but we'll not extend here.

5 Conclusion

The core finding of this study is that the semi-strong form of the Efficient Market Hypothesis (EMH) largely holds in the Japanese Government Bond (JGB) market. By analyzing the speech texts of the Bank of Japan (BOJ) using large language models (LLMs), it is not possible to systematically predict short-term changes in yields or the central bank's holdings.

The study arrives at this conclusion through the following key aspects:

Quantitative analysis revealing the unpredictability in short-term price movements: the paper first conducts statistical analysis on market data to establish a benchmark for subsequent textual analysis. A critical finding is that daily changes in JGB yields follow a near-random walk pattern, with almost no correlation between daily yield fluctuations and changes in BOJ bond purchases. This indicates that relying solely on past public quantitative data cannot predict short-term price movements.

NLP models demonstrating that the "tone" of public speeches also fails to predict markets: this is the core part of the research. We use models like FinBERT to quantify BOJ speech texts into Hawkish-Dovish Scores, measuring policy tightening or easing tendencies. The results show that both custom-built metrics and cutting-edge AI models exhibit near-zero correlation between these scores and future short-term changes in JGB yields or BOJ purchase volumes.

Anomalous performance of advanced AI models paradoxically supporting EMH: the study notes an intriguing paradox that state-of-the-art AI models like Claude can remarkably "predict" yield changes during backtesting. However, we doubt that this actually signals market inefficiency, strongly suggesting forward looking bias-where models have already "seen" historical data in their training. This explains why such "predictive power" cannot be replicated in real-time trading scenarios.

The paper further summarizes some potential reasons for the NLP model's ineffectiveness: The central bank communication is gradual and predictable, with markets already incorporating impacts before official releases. And the sentiment captured by NLP models may contain substantial irrelevant noise, while true market-moving elements are surprises that deviate from expectations, difficult for simple models to detect. Actually, the JGB market is more influenced by global macroeconomic conditions and U.S. interest rates, with limited impact from individual BOJ statements. Additionally, BOJ purchase operations follow pre-announced schedules, making changes more mechanical than reactive to textual tones.

For the significance and limitations of this study, by combining NLP techniques with financial econometrics, this study provides empirical support for EMH in the specific context of BOJ policy communication and the JGB market. It highlights the difficulty of extracting excess returns from public text

information in highly developed, transparent markets.

The research also acknowledges methodological limitations, such as NLP models being primarily trained on English financial texts, potentially missing subtle meanings in Japanese originals, and the model's inherent constraints in capturing surprise elements.

The authors exchanged with an investment professional familiar with the trading of JGBs in Japan about their findings on the counterfactual prompt: "I understand the signaling by the Bank of Japan. This is similar to the understanding of the degrees of politeness in japanese. The mere fact that the amount of JGB purchases are being flagged as necessary to review is a strong indicator of change to come."

In conclusion the paper suggests that for quant investors it is necessary not only to study the factual translations into English of the BOJ statements, but also to run continuously counterfactual type analysis to fine-tune language models specific to the BOJ formal/polite signaling language.

References

Stagnol, L., Cherief, A., Farah, Z., Guenedal, T. L., Sakout, S., & Sekine, T. (2023). *Answering Clean Tech Questions with Large Language Models. Social Science Research Network.* https://doi.org/10.2139/SSRN.4663447

Hattori, T., & Yoshida, J. (2021). *Yield Curve Control. Social Science Research Network.* https://doi.org/10.2139/SSRN.3396251

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Å., & Polosukhin, I. (2017). *Attention is All you Need. arXiv (Cornell University).*

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.* https://doi.org/10.18653/V1/N19-1423

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: Computation and Language.*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., GuzmÃ¡n, F., Grave, Ã., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale. meeting of the association for computational linguistics.* https://doi.org/10.18653/V1/2020.ACL-MAIN.747

Martin, L., Muller, B., Suarez, P. O., Dupont, Y., Romary, L., Clergerie, E. V. d. l., Seddah, D., & Sagot, B. (2020). *CamemBERT: a Tasty French Language Model. Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.18653/V1/2020.ACL-MAIN.645

Brown, T., Mann, B., Ryder, N. C., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners. Neural Information Processing Systems.*

OpenAI, Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H.-i., Bavarian, M., Belgum, J., Bello, I., ... Zoph, B. (2023). *GPT-4 Technical Report. arXiv.org*. https://doi.org/10.48550/ARXIV.2303.08774

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A. J., Barham, P. R., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V. M., ... Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways. Journal of machine learning research.*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., RoziÃšre, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, Ã., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models. arXiv.org*. https://doi.org/10.48550/ARXIV.2302.13971

Agarwal, S., Almeida, D., Askell, A., Christiano, P. F., Hilton, J., Xu, J., Kelton, F., Leike, J., Lowe, R., Miller, L., Mishkin, P., Ouyang, L., Ray, A., Schulman, J., Simens, M., Slama, K., Wainwright, C., Welinder, P., Wu, J., & Zhang, C. (2022). *Training Language Models to Follow Instructions with Human Feedback*. https://doi.org/10.52202/068431-2011

Linear-probe, A. (2021). *Learning Transferable Visual Models From Natural Language Supervision.*

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). *Flamingo: a Visual Language Model for Few-Shot Learning. Neural Information Processing Systems*. https://doi.org/10.48550/ARXIV.2204.14198

Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2024). *A Survey on Model Compression for Large Language Models*. https://arxiv.org/abs/2308.07633

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv.org.*

Robertson, S., & Zaragoza, H. (2009). *The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval.* https://doi.org/10.1561/1500000019

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv: Computation and Language.* https://doi.org/10.18653/V1/D19-1410

Karpukhin, V., OÄuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). *Dense Passage Retrieval for Open-Domain Question Answering. Conference on Empirical Methods in Natural Language Processing.* https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550

Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T. (2021). *Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. arXiv: Artificial Intelligence.*

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. empirical methods in natural language processing.* https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.148

Loughran, T., & McDonald, B. (2011). *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10âKs. Journal of Finance.* https://doi.org/10.1111/J.1540-6261.2010.01625.X

Leippold, M. (2023). *Sentiment Spin: Attacking Financial Sentiment with GPT-3. Social Science Research Network.* https://doi.org/10.2139/SSRN.4337182

Anthropic. (2024). *The Claude 3 Model Family: Opus, Sonnet, Haiku.* https://api.semanticscholar.org/CorpusID:268232499

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., ... Yun, T. (2025). *DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature.* https://doi.org/10.1038/S41586-025-09422-Z

A Appendix

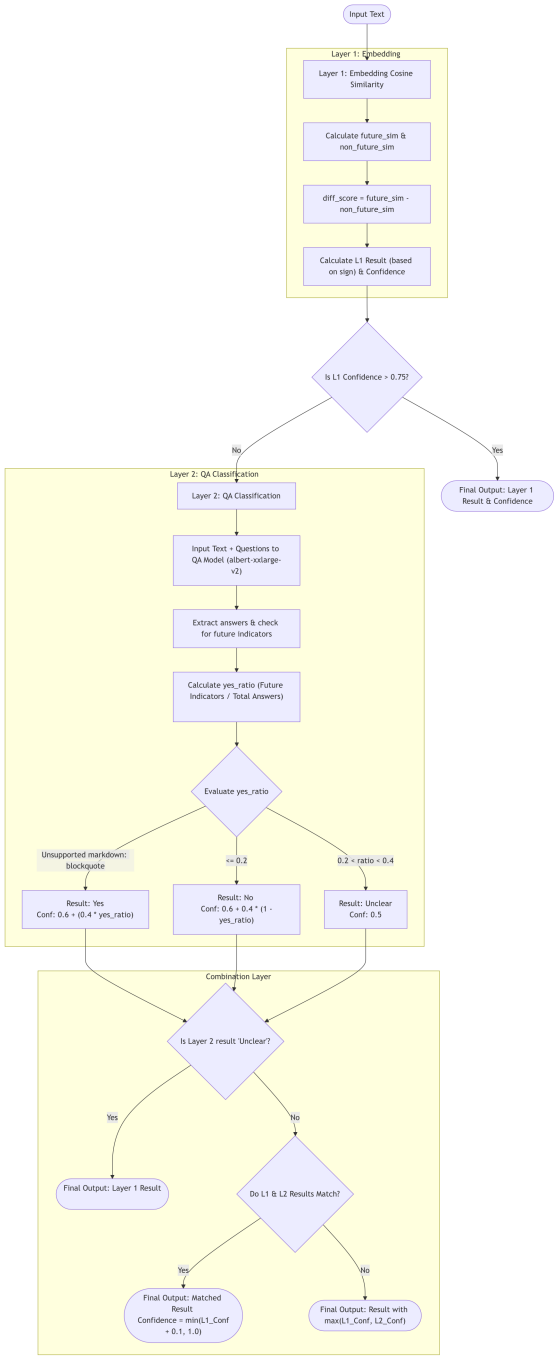
A.1 Examples for embedding

Table 18: Examples of embedding

future_examples	**non_future_examples**
"We expect economic growth in the coming years"	"Current economic indicators show results"
"Our policy will support future development"	"Recent developments have been challenging"
"We anticipate continued improvement ahead"	"The situation remains stable today"
"The outlook remains positive for future"	"Historical data reveals patterns"
"We project steady progress next year"	"Existing policies function effectively"
"Future trends show promising trajectory"	"Present conditions require attention"
"Our strategy focuses on long-term goals"	"Recent statistics indicate moderate pace"
"Looking ahead, we remain optimistic"	"The current environment presents opportunities"
"Going forward, we expect growth"	"Today's decisions reflect reality"
"Upcoming quarters show positive signs"	"Past experiences informed approach"
"We believe the future will be better"	"We are dealing with current issues"
"Plans for next decade are promising"	"The present state is concerning"
"Our vision for tomorrow is clear"	"Recent performance was mixed"
"Future prospects appear favorable"	"Current trends show no change"
"We are preparing for what's ahead"	"Today we face challenges"

A.2 Methodology Flowchart

Figure 12: QA Classification Flowchart



A.3 Hawkish Keywords and Dovish Keywords

Table 19: Hawkish and Dovish Keywords

Hawkish Keywords	**Dovish Keywords**
tighten	ease
tightening	easing
raise interest rate	cut rate
rate hike	rate cut
inflation	lower rate
overheating	stimulate
restrictive	stimulus
contractionary	accommodative
hawkish	expansionary
curb inflation	dovish
combat inflation	support growth
price stability	deflation
anchor inflation	disinflation
withdrawal	slack
reduce accommodation	underutilization
normalize	recession
policy normalization	slowdown
upward pressure	subdue
wage pressure	weak demand
demand pressure	spare capacity
tight labor market	unemployment
full employment	downside risk
	uncertainty
	patience
	gradual
	cautious

A.4 Performance of Claude Predictions

Figure 13: Performance of Claude Predictions Dist

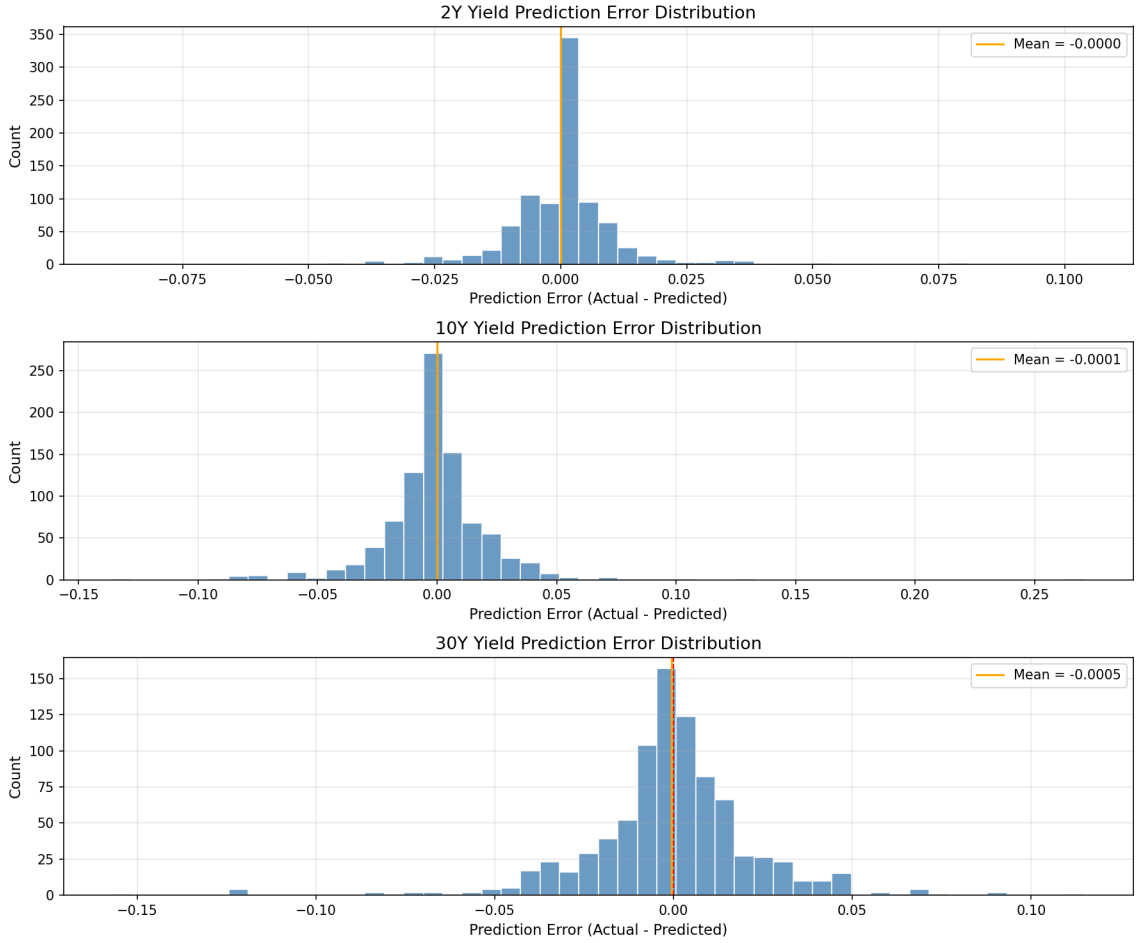


Figure 14: Performance of Claude Predictions Scatter

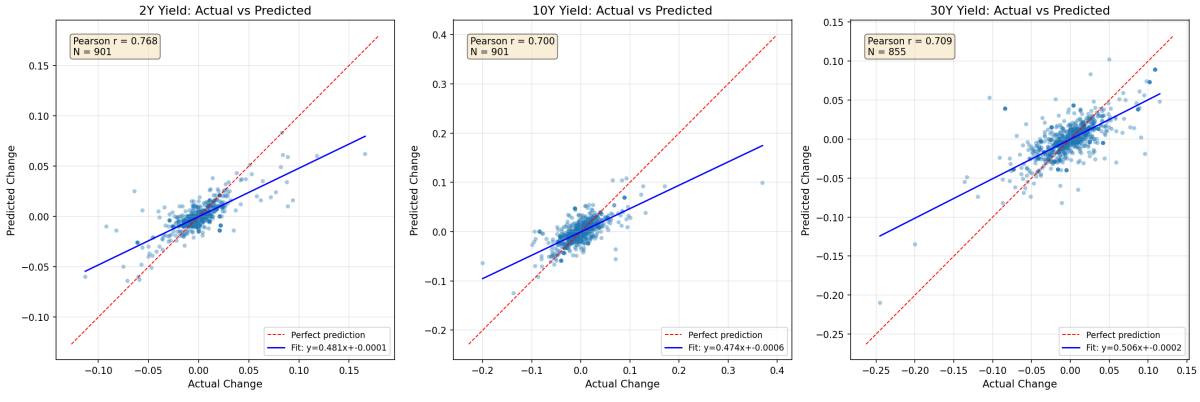


Figure 15: Performance of Claude Predictions Timeseries