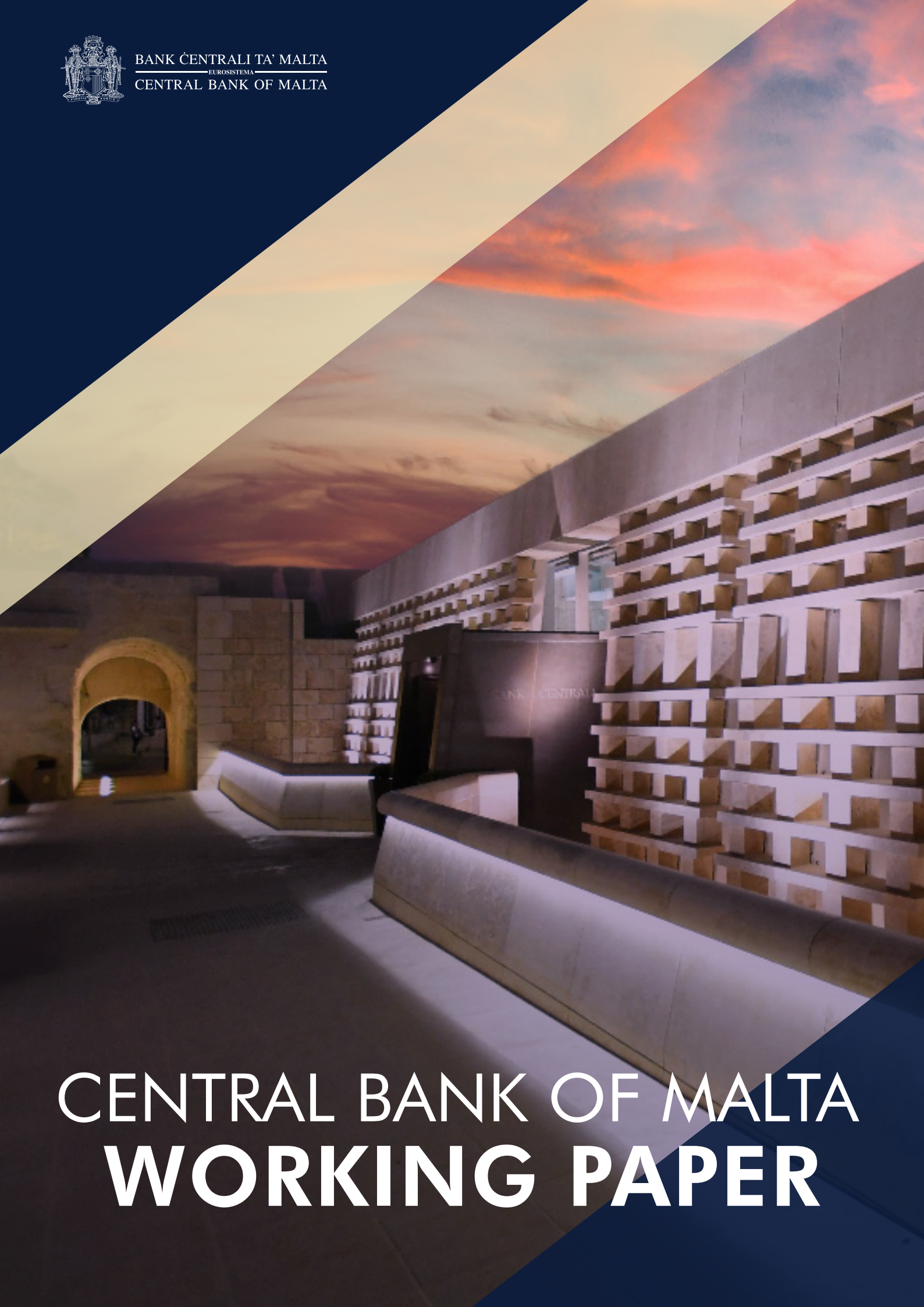


BANK ĊENTRALI TA' MALTA
EUROSISTEMA
CENTRAL BANK OF MALTA

CENTRAL BANK OF MALTA
WORKING PAPER

BANK ĊENTRALI TA' MALTA
EUROSISTEMA
CENTRAL BANK OF MALTA

Interpreting the Interpreter: Can We Model Post-ECB Conference Volatility with LLM Agents?

Umberto Collodel*

Central Bank of Malta

WP/04/2026

* Monetary Policy and Eurosystem Relations Department, Central Bank of Malta.

The author would like to thank the instructors and participants of the 2025 Barcelona Graduate School of Economics Summer School on Natural Language Processing along with participants of the 2026 RCEA International Conference in Economics, Econometrics, and Finance and an Internal CBM Seminar. Special thanks go to Maximilian Freier for providing constructive feedback that greatly improved this work, as well as Jens Christensen, Manuel Betín, Nicoletta Batini, Jonathan Benchimol, John Caruana, Francesco Toni, Eric Vaansteenberghe, Massimo Giovannini, and colleagues from the Central Bank of Malta. In particular, the work benefited from thorough discussions with Jacopo Zacchè. Any remaining errors are my own.

Abstract

Central banks cannot observe how markets will interpret their communications before release. We propose a framework in which Large Language Models simulate 30 heterogeneous traders interpreting European Central Bank press conference transcripts, yielding a measure of cross-sectional disagreement among synthetic agents. Across 293 Governing Council events, this measure correlates at approximately 0.5 with realized Overnight Index Swap volatility, outperforming standard text-based alternatives. Crucially, the indicator retains predictive power after controlling for the monetary policy surprise and liquidity proxies, suggesting that the measure captures broad belief disagreement generated by how communication is phrased rather than merely the policy content. The framework offers a practical, ex-ante tool for assessing how central bank communication is likely to be interpreted by financial markets.

JEL Classification: E52, E58, C63

Keywords: monetary policy communication, large language models, monetary policy uncertainty, agent-based modeling

1 Introduction

Central banks face a fundamental communication dilemma: they cannot test how markets will interpret their statements before release. Upon publication, policy language triggers immediate and irreversible market reactions. Yet policymakers lack tools to anticipate these effects during the drafting process, when language remains modifiable. This constraint is costly. Ambiguous communication can amplify market volatility, hinder financial stability, and undermine policy transmission (Tillmann 2020; Bauer et al. 2021; Collodel and Kunzmann 2025). Traditional high-frequency identification methods operate ex-post, measuring what happened rather than guiding ex-ante language refinement (e.g. Altavilla et al. (2019)).

This paper introduces an operational framework for simulating heterogeneous market reactions to monetary policy communication before release. We employ Large Language Models (LLMs) to simulate a cross-section of 30 heterogeneous synthetic traders, each endowed with distinct risk preferences, cognitive biases, and interpretive styles. These agents process European Central Bank (ECB) press conference transcripts and forecast Euro interest rate swap rates across three key maturities: 3-month, 2-year, and 10-year tenors. Cross-sectional forecast dispersion provides a model-based measure of market disagreement, which we validate against realized Overnight Index Swap (OIS) volatility, our market-based proxy for disagreement, across 293 ECB communications spanning June 1998 to March 2026.

Our analysis yields two main findings. First, simulated disagreement achieves Spearman correlations of approximately 0.5 with realized market volatility at medium- and long-term maturities, peaking at the 2-year tenor ($\rho = 0.53$) and outperforming standard text-based metrics by a wide margin. LLMs thus appear to possess an intrinsic capacity to decode central bank communication, even without prompt calibration, historical conditioning, or fine-tuning. Second, the measure retains explanatory power after controlling for liquidity proxies, autoregressive persistence, and the contemporaneous monetary policy surprise. The residual association therefore represents dispersion about *how* the decision is communicated rather than *what* is decided i.e. the linguistic margin relevant to ex-ante drafting.

The results hold out-of-sample on conferences post-dating the model's training data, alleviating concerns of temporal leakage, and are further robust to sampling stochasticity, prompt variation, and the choice of underlying language model.

For central banks, the framework provides an operational tool to anticipate communication-induced volatility before release. Policymakers can systematically evaluate alternative phrasings and quantify expected market disagreement before publication. This transforms communication strategy from reactive refinement, based on costly market reactions, to proactive optimization during the drafting process. The methodology extends naturally to other major central banks, enabling comparative analysis of how communication styles and institutional frameworks shape interpretive disagreement across monetary policy contexts.

The remainder of the paper proceeds as follows. Section 2 reviews related literature on monetary policy communication, expectation formation, and LLM applications in economics. Section 3 describes our data sources, agent construction, and prompting strategy. Section 4 presents the main results, establishing the correlation between simulated and market-based disagreement and interpreting the measure. Section 5 reports additional robustness checks. Section 6 concludes with implications for central bank communication design, caveats, and future research directions.

2 Related Literature

Our work contributes to three intersecting strands of research: monetary policy communication and high-frequency identification of market reactions, behavioral models of expectation formation, and the application of LLMs to economic analysis.

Monetary Policy Communication and High-Frequency Identification A substantial literature documents that central bank communication moves financial markets beyond the direct effects of policy rate changes. For the United States, Kuttner (2001) and Gurkaynak et al. (2005) pioneered the identification of monetary policy shocks using high-frequency movements in federal funds rate and Eurodollar futures on Federal Open Market Committee (FOMC) dates. Jarociński and Karadi (2020) show that separating policy from information shocks in a structural VAR magnifies estimated monetary policy effects. Bauer and Swanson (2023) further demonstrate that raw surprises are partially predictable from pre-announcement public data, recommending orthogonalization to restore instrument exogeneity. In the euro area, Altavilla et al. (2019) constructed a comprehensive database of monetary policy surprises from ECB

announcements following the methodology of Gürkaynak (2005). Similar strategies have been used to examine monetary policy in the United Kingdom (Cesa-Bianchi et al. 2020), China (Das and Song 2023), and in a broad cross-country setting (Bolhuis et al. 2024).

Subequent work by Bauer et al. (2021) distinguishes between the direction of policy surprises and the uncertainty surrounding the policy stance, showing that elevated uncertainty depresses equity prices, raises sovereign spreads, and increases exchange rate volatility. Collodel and Kunzmann (2025) document similar patterns for the Euro area, establishing that monetary policy uncertainty shocks have distinct transmission channels from conventional policy surprises. In addition, uncertainty might also diminish the transmission of first-moment surprises: for example, Tillmann (2020) finds that hawkish surprises are less effective in periods of heightened monetary policy uncertainty, as investors shift toward longer maturities depressing the long-end of the curve.

While this literature convincingly demonstrates that communication matters, existing approaches remain fundamentally retrospective. High-frequency identification measures market reactions after communication occurs, offering no guidance for ex-ante language design. Our framework addresses this gap by enabling simulation of market disagreement before publication.

Expectation Formation and Behavioral Heterogeneity The rational expectations hypothesis, foundational to much of modern macro-finance, has been increasingly challenged by empirical evidence documenting systematic biases and bounded rationality among investors (Barberis and Thaler 2003; Bordalo et al. 2018). Standard representative-agent models do not incorporate these features, hence struggle to explain observed phenomena like excess volatility and momentum. Agent-based models (ABMs) have emerged as a flexible alternative (Farmer and Foley 2009; Axtell and Farmer 2025), allowing for heterogeneous, adaptive agents whose interactions generate emergent macroeconomic patterns. Within monetary policy research, ABMs have been used to model expectation formation and policy transmission in bottom-up settings (Delli Gatti et al. 2011). Yet two key limitations remain. First, behavioral rules are typically imposed ex ante as stylized heuristics, rather than learned or emergent (Horton 2023). Recent work comparing heuristic and LLM-based agent architectures confirms that the latter substantially improve simulation realism in large-scale settings (Chopra et al. 2025). Second, conventional ABMs cannot interpret qualitative information directly, despite the fact

that monetary policy signals are largely communicated in natural language. This omission bypasses an essential channel of transmission. This distinction reflects a broader theoretical insight: while traditional prediction algorithms efficiently process observable data, they cannot access latent human judgment, a gap that generative AI is beginning to bridge by learning from large text corpora (Mullainathan and Spiß 2017; Agrawal et al. 2018). Our framework addresses both ABMs limitations by enabling agents to process and act upon policy-relevant textual information.

Large Language Models in Economic Analysis Recent work demonstrates that LLMs can simulate human decision-making in economic contexts. Horton (2023) shows that LLM-based agents (*homo silicus*) replicate behavioral patterns in experimental market settings. Kazinnik and Sinclair (2025) extend this to institutional settings, simulating FOMC deliberations through multi-agent systems and demonstrating that LLMs can replicate collective decision-making processes. Most directly related to our framework, Hansen et al. (2025) construct a synthetic Survey of Professional Forecasters panel by prompting LLMs with real-time macroeconomic data and individual forecaster characteristics, showing that simulated forecasts replicate key distributional features of human disagreement across horizons.

In central banking specifically, LLMs have been applied primarily to *classification* of communication content rather than prediction of market reactions. Christiano Silva et al. (2025) use LLMs to classify 75,000 central bank documents from 169 countries by topic and stance. BIS (2024) develop domain-adapted language models for hawkish-versus-dovish tone detection, while Pfeifer and Marohl (2023) build CentralBankRoBERTa to classify sentiment toward macroeconomic agents. The Bundesbank's MILA system (Bundesbank 2025) employs role-based prompting to analyze ECB statements sentence-by-sentence.

Our contribution departs from both strands. Unlike the LLM-agent literature, which validates simulated behavior against experimental outcomes or survey accuracy, we target realized market outcomes. Moreover, differently from the central banking classification literature, we are not interested in sentiment or topic classification, but rather in how language generates heterogeneous interpretations. By combining these two angles, agent simulation and central bank communication, we produce a measure of simulated disagreement that we correlate with realized OIS volatility. To our knowledge, this is the first framework to use LLMs for prediction

of market disagreement following central bank communication.

3 Methodology and Data

We construct a framework to simulate and predict market disagreement following ECB press conferences. The methodology comprises three components: data construction, agent design, and measure validation. We first assemble a comprehensive database of 293 ECB Governing Council press conferences spanning June 1998 to March 2026, pairing each transcript with realized Euro OIS volatility across the three maturities. We then develop LLM-based synthetic agents that process these transcripts and generate rate forecasts, with cross-sectional dispersion serving as our measure of simulated disagreement. Finally, we validate this measure by correlating it with actual market volatility.[1]

ECB press conferences constitute the primary vehicle for monetary policy communication in the Euro area. Following the high-frequency identification literature (Altavilla et al. 2019), we define events as regularly scheduled Governing Council meetings. For each event, we collect full transcripts (Introductory Statement and Q&A) and corresponding post-announcement OIS volatility for short-term (3-month), medium-term (2-year), and long-term (10-year) maturities.[2]

We measure realized OIS volatility at day t for maturity m using the daily high-minus-low range on the day following the press conference:

$$\text{OIS Volatility}_{t,m} = \text{High}_{t+1,m} - \text{Low}_{t+1,m} \tag{1}$$

The daily high-low range is a standard and efficient range-based estimator of latent volatility (Parkinson 1980), with the one-day window minimizing endogeneity concerns from confounding events, e.g., macroeconomic data releases, geopolitical news, or other announcements, that would contaminate longer horizons.

Crucially, we treat realized volatility as a proxy for the cross-sectional dispersion of beliefs

[1]The euro and its overnight reference market postdate the start of the conference sample: the euro was introduced on 1 January 1999 and EONIA, the rate underlying Euro OIS, was first fixed on 4 January 1999. Synthetic forecasts are constructed for all conferences, but for those held between June and December 1998 no realized Euro OIS outcome exists against which to validate; these are therefore excluded from the validation regressions while retained in the descriptive.

[2]Following standard practice, we use OIS rates as they reflect pure interest rate expectations without credit risk components present in government bond yields (Gurkaynak et al. 2005).

among market participants: a long tradition in the differences-of-opinion literature derives belief dispersion as a common driver of both trading volume and price volatility (Harris and Raviv 1993; Shalen 1993; Banerjee and Kremer 2010), so that wider post-conference ranges correspond to more heterogeneous interpretations of the communication.[3] We retrieve all quotes from LSEG Workspace.[4]

Figure 1 illustrates market-based disagreement following ECB Governing Council meetings across all three maturities. The data reveal pronounced spikes during periods of heightened policy uncertainty, notably the 2008-2009 financial crisis, the European sovereign debt crisis, and the recent monetary policy tightening cycle initiated in 2022. Volatility exhibits a clear maturity structure, with consistently higher levels at longer tenors reflecting greater uncertainty about the medium- and long-term policy trajectory compared to near-term rate expectations.

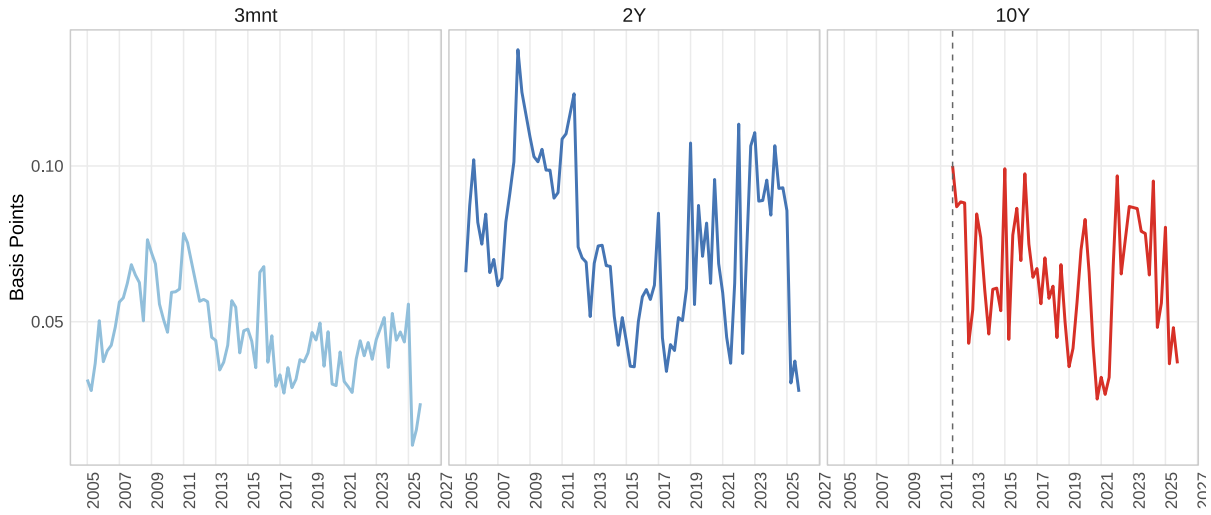


Figure 1: Post ECB Governing Council OIS volatility

Note: Volatility is measured as the min-max range of OIS rates in the first day following ECB press conferences. Quarterly averages displayed. Data for the 10-year tenor available from 2011-Q3.

[3]Realized volatility need not reflect disagreement alone; in principle it may also respond to liquidity conditions, balance-sheet constraints, or hedging demand. In Section 4 we shows ex-post that our central relationship is invariant to market liquidity.

[4]Data for the 10-year tenor start in September 2011, compared to 2005 for the other two maturities; the headline figure of 293 refers to the number of transcripts, while the matched validation sample is correspondingly smaller at the longer tenor.

3.1 Agents

We simulate 30 heterogeneous traders using Google's Gemini 2.5-Flash model.[5] Each agent interprets ECB press conference transcripts and forecasts Euro OIS rates across three maturities (3-month, 2-year, 10-year). To generate interpretive heterogeneity, agents differ along three dimensions:

1. **Risk Aversion**: High, medium, or low tolerance for uncertainty, affecting yield curve stability preferences.

2. **Behavioral Biases**: Each agent exhibits 1–2 biases from established behavioral finance (confirmation bias, overconfidence, anchoring, herding, loss aversion, recency bias), reflecting systematic deviations from rational expectations documented in real markets (Barberis and Thaler 2003).

3. **Interpretive Style**: Agents adopt distinct analytical approaches (fundamentalist, technical, sentiment-driven, quantitative, narrative-focused, or policy-skeptic) that shape their processing of ECB conferences. These styles are drawn from heterogeneous agent models and behavioral finance (see Hommes (2006)).

The LLM dynamically assigns these characteristics to create 30 unique agents for each press conference, ensuring identical communication generates diverse interpretations. This design preserves behavioral diversity necessary to capture cross-sectional variation in market reactions while allowing natural adaptation over time.

We set the model's temperature parameter to 1 to balance realistic heterogeneity with response stability.[6] However, since agent predictions are drawn stochastically at temperature 1, a single simulation run may not be representative of the underlying disagreement distribution. To address this, we repeat the simulation $R = 10$ times per press conference, computing the cross-sectional standard deviation of forecasts within each run and averaging across replications. The choice of $R = 10$ is motivated by the findings of Section 5.1: at this replication count our synthetic disagrameent, $\widehat{\sigma}_t$ achieves an average reliability coefficient above 0.6, a

[5] We select 30 agents to balance computational efficiency with meaningful cross-sectional variation. Results are robust to using 20 or 50 agents and are available upon request.

[6] Temperature controls output randomness: lower values (e.g., 0.2) produce deterministic responses, while higher values (e.g., 1.5+) generate implausible extremes.

standard threshold indicating that between-conference variation in agent disagreement dominates within-conference sampling noise (Cicchetti 1994). The averaged measure therefore reflects systematic differences in how heterogeneous traders interpret individual transcripts rather than idiosyncratic draw-to-draw variation.

It is important to note that these LLM-based agents represent stylized market participants who process textual information and form expectations. Real investors integrate broader factors, e.g. risk constraints, institutional mandates, collective behavior, that our framework abstracts from. This design isolates the communication-specific component of market disagreement, which is our target estimand, rather than attempting to match precisely realized price levels.

3.2 Prompting Strategy

Our main analysis uses a single, deliberately minimal prompting strategy:

Zero-Shot: Agents receive only the ECB transcript, with no historical conditioning, worked examples, or prompt tuning.

This is the most demanding test of whether LLMs possess an intrinsic capacity to decode monetary policy communication.

For each event, each agent generates forecasts of 3-month, 2-year, and 10-year OIS rates, together with a predicted direction and a confidence score.[7] Figure 2 shows the zero-shot prompt. We measure synthetic disagreement as the cross-sectional standard deviation of these forecasts across the 30 agents.

We additionally explore two extensions of the baseline in Appendix B. The first, *historical anchoring* (few-shot), augments the prompt with realized pre- and post-conference volatility from the three most recent meetings, following evidence that historical examples can improve LLM forecasting performance in economic contexts (Hansen et al. 2025). The second, *LLM-as-a-Judge*, introduces a second model that iteratively rewrites the first LLM prompt to maximize in-sample correlation with market volatility; we treat this as an exploratory study of automated prompt optimization.[8]

[7] We include the predicted direction to ensure internal consistency, as LLM outputs may otherwise exhibit inconsistencies.

[8] We use Gemini 2.5-Pro as the Judge and Gemini 2.5-Flash as the Analyst, with a 75/25 chronological train–test split. See Algorithm 1 and Figure 19 for details.

Figure 2: Zero-shot Prompt

```
Context:
You are simulating the Euro area interest rate swap market, composed of 30 individual traders.
These traders interpret the ECB Governing Council press conference, which communicates monetary policy
    decisions, economic assessments, and includes a Q&A session with journalists.
Each trader then makes a trading decision to maximize profit based on their interpretation of the
    conference and their unique characteristics.
Trader Characteristics:
Each trader has the following attributes:
- Risk Aversion: High / Medium / Low determines sensitivity to uncertainty and preference for stability
- Behavioral Biases (1,2 per trader): e.g., Confirmation Bias, Overconfidence, Anchoring, Herding, Loss
    Aversion, Recency Bias
- Interpretation Style (1 per trader): e.g., Fundamentalist, Sentiment Reader, Quantitative, Skeptic,
    Narrative-Driven
Task:
You are given the text of a single ECB press conference.
For each of the 30 traders, simulate their individual trading action in the interest rate swap market
    across three tenors (3 months, 2 years, 10 years).
For each tenor, the trader must:
    - Provide an expected rate direction: Up / Down / Unchanged (relative to the pre-conference rate)
    - Provide a new expected swap rate (in percent, to two decimal places)
    - Provide a confidence score (0 to 100%) reflecting how strongly the trader believes in his forecast,
    based on their interpretation of the press conference and their own characteristics
Output:
Provide a table with the following structure for each press conference, trader, and interest rate tenor

| Date       | Trader ID | Tenor   | Expected Direction | New Expected Rate (%) | Confidence (%) |
|------------|-----------|---------|--------------------|-----------------------|----------------|
| YYYY-MM-DD | T001      | 3M      | Up                 | 3.15                  | 65             |
| YYYY-MM-DD | T001      | 2Y      | Down               | 2.85                  | 80             |
| ...        | ...       | ...     | ...                | ...                   | ...            |
Guidelines:
- Use only the information available as of [date].
- Do not aggregate or summarize responses.
- Reflect diversity in interpretation, risk tolerance, and horizon. Rationale must be unique for each
    trader and can vary across tenors.
- Output only a markdown table with the specified columns, no additional text. Do not use JSON or any
    other data serialization format.
```

3.3 Evaluation

We evaluate the framework by correlating synthetic disagreement with realized market volatility. Our primary metric is the Spearman rank correlation coefficient, computed separately for each maturity. We use Spearman correlation because it captures monotonic relationships without linearity assumptions and is robust to the outliers common in high-volatility periods; we verify the results using Pearson and Kendall correlations in Appendix 6.1.[9][10]

4 Results

This section establishes the central result of the paper, that simulated disagreement tracks realized post-conference volatility, and then show that disagreement is largely communication induced ruling out alternative explanations. We begin in Section 4.1 describing the simulated time series of forecast directions and dispersion, then turning to its rank correlation with realized OIS volatility and its performance relative to standard text-based indicators. The subsequent section subjects this correlation to three challenges: that it merely reflects volatility persistence; that the dependent variable captures thin-market price movement rather than belief dispersion; and that it reflects the substance of the policy decision rather than its communication (Section 4.2).

4.1 Narrative and Correlation with Market Values

Our simulation provides the most direct test of whether LLM agents endowed only with a pool of behavioral traits to pick from and no historical training context can reproduce the ebb and flow of overall market sentiment. Before turning to formal correlation measures, Figures 3 and 4 offer a visual narrative of the simulated output.

Figure 3 plots the share of agents predicting an increase, decrease, or no change in rates after each ECB press conference.

Even in this naive setup, the agents' directional calls broadly mirror well-known phases

[9]For transparency and reproducibility, Appendix D provides implementation details including model specifications and text preprocessing.

[10]For each press conference we send a single API request. Although batching conferences would reduce computation time, we adopt a conservative approach to avoid including future information as context, which could inadvertently influence the model's output.

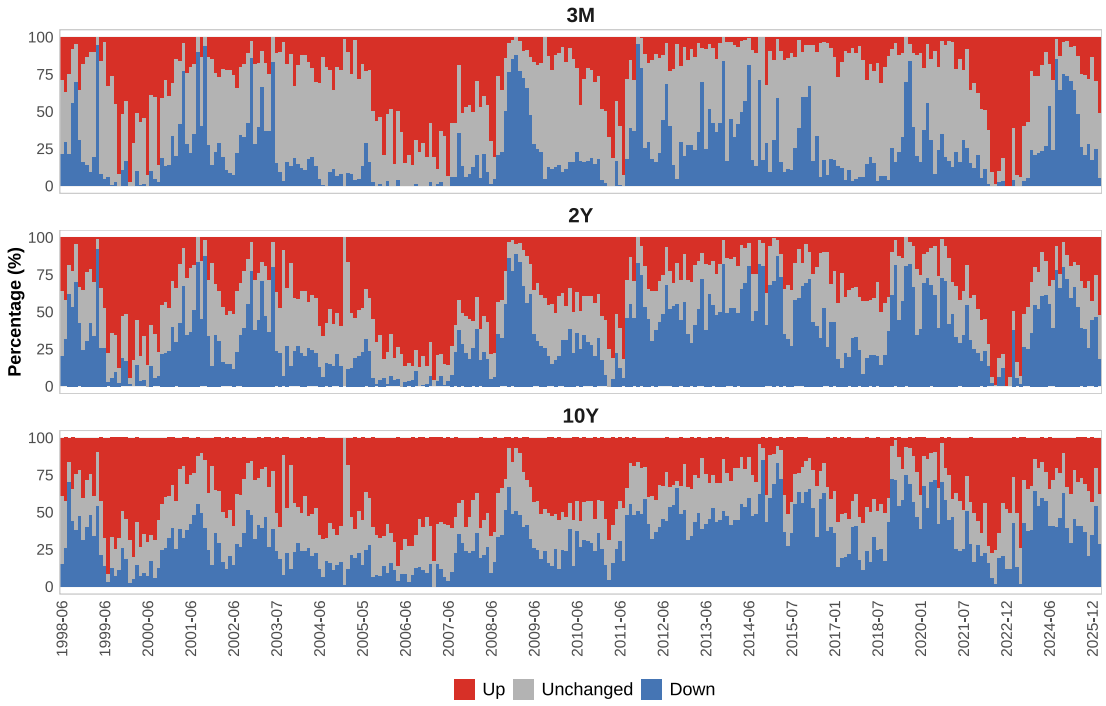


Figure 3: Percentage of forecast directions from the zero-shot LLM simulation

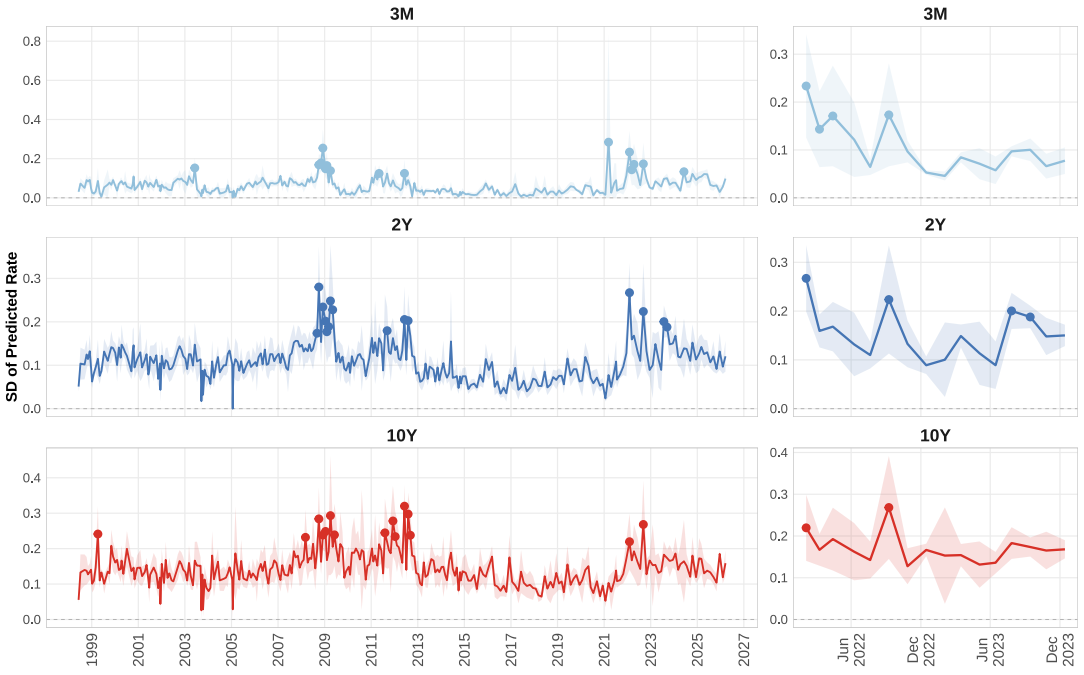
Note: Each data point reflects the output of an LLM model prompted in a zero-shot setting using Gemini 2.5-Flash, with temperature 1. The simulation uses a sample of 293 ECB press conference (June 1998–March 2026) transcripts and the baseline prompt (Figure 2). For each press conference, the model output is generated across 10 independent sampling runs.

of ECB policy. During the early ECB and pre-Global Financial Crisis (GFC) period (1999–2008), upward predictions dominate at the 2-year and 10-year tenors, averaging 50.2% and 49.1% of agents respectively, while the short end is more evenly split between "Up" (37.4%) and "Unchanged" (47.7%). The subsequent easing wave is captured sharply: within a single meeting of the September 2008 Lehman shock, the modal direction at the 2-year tenor reverses from "Up" to "Down". During the ZLB era (2009–2021), the 3-month tenor is dominated by "Unchanged" predictions (57.5% on average), consistent with rates hitting the floor, while longer maturities initially follow the explicit forward guidance from the ECB, at the 2-year 44.6% of agents still predict a rate decrease against only 29.6% expecting no change, to then reverse their behavior from roughly 2015 onwards. Majority reversals, i.e. meetings where the modal direction switches relative to the prior conference, occur in roughly one in four conferences across all tenors (23.4%, 24.1%, and 22.4% at the 3-month, 2-year, and 10-year respectively), indicating that agents update directional views frequently and without strong inertia. The 2022–2023 tightening cycle illustrates this: upward predictions rise by up to 28.4 percentage

points across two consecutive meetings at the 3-month tenor (April pair), with similar dynamics at the 2-year (22.3 pp) and 10-year (29.3 pp, March pair), tracking the ECB's acceleration of its hiking pace faced with the inflation surge.

Figure 4 shows, instead, our measure of interest: the cross-sectional standard deviation of forecasts.

Figure 4: Standard deviation of forecasts from the baseline LLM simulation



Note: Each data point reflects the output of an LLM model prompted in a zero-shot setting using Gemini 2.5-Flash, with temperature 1. Dispersion is measured as the average cross-sectional standard deviation of rate predictions over 10 iterations. The simulation uses a sample of 293 ECB press conferences (June 1998–March 2026) and the baseline prompt (Figure 2). Highlighted points represent observations exceeding the 95th percentile of dispersion for each tenor. Shaded areas are min-max standard deviation from the 10 runs.

Dispersion contracts systematically during periods of unambiguous policy stance, such as the post-GFC easing phase and much of the ZLB era, when forward guidance provided clear anchoring. In this period, the 3-month dispersion falls below 5 basis points, with longer tenors, instead, clustering around 10-15 basis points. Conversely, it spikes at policy turning points and during crisis episodes, peaking in the 2008-2009 financial crisis (with the 2 and 10-year dispersion exceeding 40 basis points), the 2011-2012 sovereign debt crisis, and most recently during the 2022-2023 monetary policy tightening cycle, where disagreement reaches unprecedented levels (roughly 60 basis points for the 2 and 10-year maturity). In the latter instance, disagreement arises less from directional confusion and more from uncertainty over policy speed, magnitude,

and terminal endpoints. This last result is consistent with the ECB 2025 Monetary Policy Strategy Review findings (ECB 2025).[11] The report shows that, once forward guidance was abandoned, the sensitivity of OIS rates to macroeconomic news increased significantly, hence amplifying excess volatility. Moreover, the early phase of the 2022 hiking cycle was marked by historically large monetary policy shocks, including the unexpected 50 basis point hike in July 2022 and the largest hawkish surprise since the global financial crisis in December 2022, reflecting the challenge for the ECB of aligning communication with policy moves. By contrast, following the introduction of the three-element reaction function framework in March 2023, the subsequent phase of the tightening cycle exhibits markedly subdued volatility to policy announcements.[12] The cross-tenor hierarchy remains remarkably consistent across all episodes, with 10-year uncertainty slightly larger than 2-year uncertainty, and both larger than the 3-month dispersion by 2-3 times, reflecting the increasing complexity of information processing and projection uncertainty at longer time horizons.

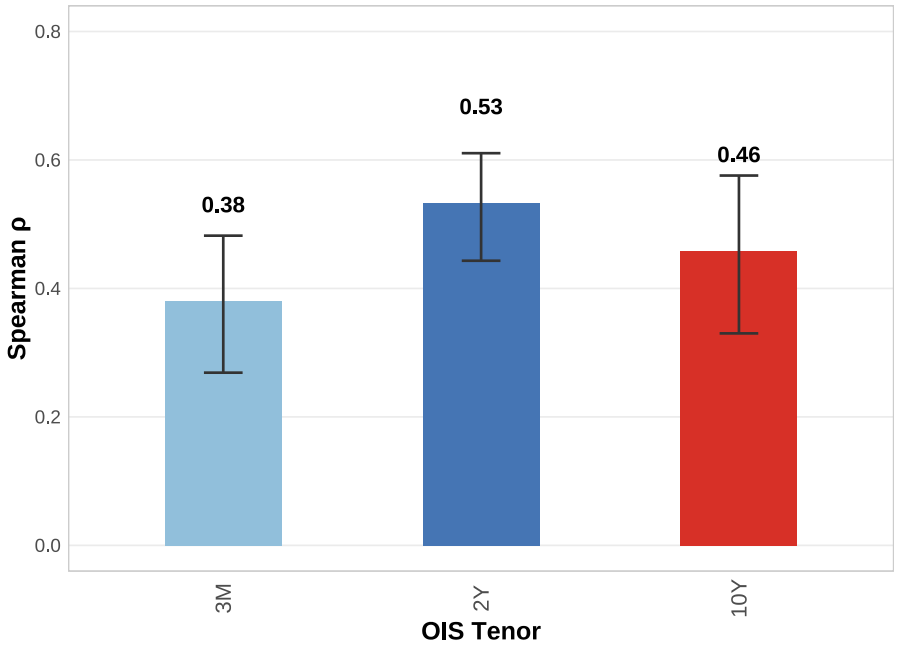
Figure 5 presents correlations with market measures. The 2-year tenor achieves the highest correlation (0.53), followed by the 10-year (0.46), and the 3-month (0.38). This ranking likely reflects the differential nature of information processed by the LLM across the yield curve. The 2-year tenor primarily captures short- to medium-term policy expectations that are directly shaped by ECB communication during press conferences. In contrast, the 10-year yield incorporates broader factors extending beyond central bank communication, e.g., long-term inflation expectations, structural growth prospects, and term premium fluctuations, making it less tightly linked to the specific content of ECB press conferences. The 2-year segment's narrower focus on monetary policy expectations, therefore, produces the strongest correlation with our LLM-based disagreement measure. At the short end, the 3-month OIS rate exhibits inherently low variance, particularly during prolonged policy stasis such as the ZLB era. Consequently, even when the LLM accurately interprets ECB communication tone, the constrained variation in 3-month rates limits the achievable correlation. Crucially, these results demonstrate that

[11] Importantly, the Strategy Review itself is outside the training sample of the model.

[12] The three-element reaction function, communicated by the ECB at the 16th March 2023 meeting, specifies that future rate decisions are determined by three inputs: (A) the assessment of the inflation outlook in light of incoming economic and financial data, (B) the dynamics of underlying inflation, and (C) the strength of monetary policy transmission. By making the determinants of future decisions explicit, it formalised the data-dependent, meeting-by-meeting approach that had replaced calendar-based forward guidance, clarifying for markets the information the Governing Council would weigh in setting policy.

even a simple zero-shot LLM approach, without historical training, prompt-optimization, or fine-tuning, can capture economically meaningful patterns of market disagreement following central bank communication.

Figure 5: Mean Spearman correlation with market-based measures by tenor



Note: Correlation between the output of an LLM model prompted in a zero-shot setting using Gemini 2.5-Flash, with temperature 1, and the OIS min-max range 1 day after ECB conferences. The simulation uses a sample of 293 ECB press conferences (June 1998–March 2026) and the baseline prompt (Figure 2). Whiskers represent 90% confidence intervals computed via a non-parametric bootstrap with 5,000 replications. For each press conference, the model output is generated across 10 independent sampling runs and then averaged to ensure robustness.

To gauge whether this performance can be replicated by off-the-shelf text metrics, we compare our LLM-based disagreement measure against five straightforward textual benchmarks. Complexity measures include Flesch–Kincaid readability scores and document length (word count), capturing syntactic structure and information volume. Framing and tone measures comprise hedge word density and Loughran–McDonald uncertainty terms, which proxy for linguistic ambiguity and explicit uncertainty language. Finally, stance measures capture the net balance of hawkish versus dovish terminology, reflecting directional policy tone.[13] In related

[13]Flesch–Kincaid Grade Level estimates the U.S. school grade required to comprehend a text based on average sentence length and syllables per word; higher values indicate more complex, technical language. Word count measures total document length, proxying for informational load. Hedge word density captures the frequency of cautious or non-committal expressions (e.g., may," could," possibly," somewhat") that signal reduced commitment or linguistic uncertainty. Loughran–McDonald uncertainty

work, Mumtaz et al. (2023) show that Bank of England announcements with lower readability score are associated with higher gilt yield volatility post-announcement.

Table 1 reports Spearman correlations between these benchmarks and post-announcement disagreement. All five exhibit weaker correlations than our LLM ensemble (top row) across all tenors, and the gap widens at longer maturities: at the 10-year tenor, our measure achieves 0.46, while the strongest text-based benchmark (Hedging Words) reaches roughly 0.19.

Table 1: Spearman Correlations Between Text-Based Measures and Market-Based Disagreement by Tenor

Measure	3M	2Y	10Y
LLM Synthetic Disagreement	0.38***	0.53***	0.46***
A. Complexity Measures			
FK Complexity	0.227***	0.229***	-0.000
Word Count	0.157**	0.151**	-0.052
B. Framing and Tone Measures			
Hedging Words	0.219***	0.158**	0.188**
LM Uncertainty	0.161**	0.149**	-0.018
C. Stance / Policy Sentiment Measures			
Net Hawkish–Dovish Score	0.026	0.215***	0.135

Note: This table reports Spearman rank correlations between text-based measures based on 293 ECB press conferences (June 1998-March 2026) and market-based disagreement post-ECB conferences, measured as the min-max range of Euro OIS rates 1 day after the conference, across three interest rate tenors. The measures are grouped into three conceptual buckets. **(A) Complexity Measures**: FK Complexity refers to the Flesch–Kincaid Grade Level, which captures linguistic complexity based on sentence length and syllables per word; higher values indicate less readable and more technical language. Word Count represents the total length of the communication and proxies for informational load or verbosity. **(B) Framing and Tone Measures**: Hedging Words measure the frequency of cautious or non-committal expressions (e.g., "may," "could," "possibly"), which indicate linguistic uncertainty or reduced commitment. LM Uncertainty is based on the Loughran–McDonald Finance dictionary and captures explicit expressions of uncertainty in economic or financial contexts. **(C) Stance / Policy Sentiment Measures**: The Net Hawkish–Dovish Score reflects the balance of hawkish versus dovish terms and captures the directional policy tone of the communication. Outliers above the 99th percentile are removed. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The pattern is informative about what these benchmarks measure. Each is conceptually adjacent to communication-induced disagreement but distinct from it: for example, a transcript can be syntactically clear yet interpretively ambiguous. These textual proxies capture specific features of text; our measure goes beyond that and proxies the act of interpretation itself. This distinction is consistent with the maturity gradient. Text metrics retain modest correlation

comprises terms from the finance-specific dictionary developed by Loughran and McDonald (2011), designed to identify explicit uncertainty expressions in economic and financial contexts (e.g., uncertain," risk," volatility"). The net hawkish–dovish score measures the balance between hawkish terms (e.g., inflation," tightening," overheating") and dovish terms (e.g., slowdown," accommodation," "support"), following standard monetary policy dictionaries.

with 3-month volatility, where near-term policy decisions are anchored to the press conference's explicit statements and lose traction at longer maturities, where disagreement reflects how markets reconcile forward guidance with the broader policy reaction function, content that is invisible to metrics scoring sentence length or hedge counts.

4.2 Isolating the Language Channel

The quantity we want $\hat{\sigma}_t$ to capture is the disagreement that the conference's language itself generates among the agents i.e. the component of post-conference uncertainty a central bank could act on by drafting clearer communication. Isolating it requires ruling out the alternatives. The previous section showed that the cross-sectional standard deviation of the agents' forecasts, $\hat{\sigma}_t$, correlates strongly with realized OIS volatility. However, three confounds could reproduce that correlation without any role for conference language: persistence, liquidity, and policy content. Starting from the first, post-conference volatility is partly autoregressive, so $\hat{\sigma}_t$ might merely track the prevailing volatility regime rather than the language of the conference itself. Second, if the high–low range reflects price movement in thin markets rather than broad belief dispersion, a measure correlated with it would be tracking liquidity conditions, not agents' disagreement. Finally, because the agents read the full transcript, $\hat{\sigma}_t$ could be reconstructing the size of the monetary policy surprise, the "what", rather than the interpretive ambiguity of the language, the "how". We test these hypotheses with the regression below:

$$\begin{aligned}\text{OIS Volatility}_{t,m} = \gamma_m + \beta_1 \hat{\sigma}_{t,m} + \beta_2 \text{Pre-Vol}_{t,m} + \beta_3 \text{Spread}_{t,m} \\ + \beta_4 \big(\hat{\sigma}_{t,m} \times \text{Spread}_{t,m}\big) + \beta_5 |\text{Surprise}_{t,m}| + \varepsilon_{t,m},\end{aligned} \tag{2}$$

where the t and m subscripts indicate, respectively, the conference date and the OIS tenor, the term γ_m is a maturity fixed effect, Pre-Vol is the average OIS min–max range over the three days before the press conference, Spread is the bid–ask spread one day after the conference, standardized within tenor, and |Surprise| is the absolute median OIS change over the conference window from Altavilla et al. (2019). Each variable is the natural empirical counterpart to its confound. For persistence, Pre-Vol is the same min–max range as the dependent variable, averaged over the three pre-conference days to capture the underlying trend and cancel microstructure noise. For liquidity, the bid–ask spread is the most informative available proxy: Euro OIS trade over the counter, so consolidated transaction volumes, the conventional liquidity

measure in exchange-traded markets, are not observed. The quoted spread is the cost of immediacy a dealer charges, and widens precisely when the market is thin. For policy content, the conference-window median OIS change is the standard high-frequency measure of the monetary policy shock (Altavilla et al. 2019). We use the absolute surprise because the relevant confound is the size of the shock, not its direction: hawkish and dovish surprises of equal magnitude should both raise volatility.[14] Standard errors are clustered by conference date, since the three tenors on a given date share the same conference-level shock.[15]

Table 2 reports the results. $\hat{\sigma}$ alone explains 25.6% of the post-conference variation (column 1). Adding pre-conference volatility (column 2) raises explanatory power to 37.9% and both coefficients remain significant at the 1% level. The $\hat{\sigma}$ coefficient falls from 0.423 to 0.320, indicating that $\hat{\sigma}$ absorbs part of the persistence channel rather than acting orthogonally to it. Although the coefficient on $\hat{\sigma}$ falls, it, remains significant implying that this measure continues to possess explanatory power.

Columns (3) and (4) address the liquidity artifact, which can operate through two channels. The level concern, illiquid conferences having both wide ranges and high $\hat{\sigma}$, is rejected: the spread enters near zero and insignificantly (0.002), leaving $\hat{\sigma}$ unchanged. Similarly, the slope concern, the range widening mechanically on thin days, which would make the $\hat{\sigma}$–range link stronger when the spread is large, is rejected by the interaction term, which is small (0.018) and statistically indistinguishable from zero. To summarise, $\hat{\sigma}$ predicts volatility at average liquidity and predicts it no more strongly as markets thin.

Column (5) adds the absolute surprise, which enters strongly (0.452, 1% level), confirming that the magnitude of the policy shock is itself a powerful driver of subsequent volatility. But $\hat{\sigma}$ survives, declining only modestly to 0.274 and remaining significant at the 1% level. This separates the "how" from the "what": once the size of the shock is held fixed, the residual association reflects the interpretive ambiguity stemming from the language, not the policy decision.

Lastly, Column 6 includes every confound simultaneously. The coefficient on $\hat{\sigma}$ remains basically unchanged (0.279) and significant at the 1% level. Notably, this fully saturated model explains nearly half of the post-conference variation ($R^2 = 0.474$), with the bulk of that explanatory power shared between the policy language and the policy surprise.

[14] We use the raw tenor-matched change rather than the rotated factors (Altavilla et al. 2019) because it matches more cleanly our tenor-by-tenor specification.

[15] Two-way clustering (date × tenor) leaves the conclusions unchanged.

Table 2: Decomposing the Sources of Disagreement

	Dependent variable:					
	Post-conf. OIS high-low range, 1-day					
	(1)	(2)	(3)	(4)	(5)	(6)
Synthetic SD	0.423***	0.320***	0.315***	0.315***	0.274***	0.279***
	(0.041)	(0.058)	(0.061)	(0.061)	(0.056)	(0.056)
Pre-conf. OIS vol. 3-day		0.323**	0.281*	0.281*	0.231*	0.228*
		(0.153)	(0.144)	(0.144)	(0.120)	(0.118)
Bid-ask spread, std.			0.002	0.0003		−0.0002
			(0.002)	(0.002)		(0.002)
Synthetic SD × Bid-ask spread				0.018		0.027
				(0.029)		(0.027)
|Surprise|					0.452***	0.433***
					(0.109)	(0.106)
Constant	0.030***	0.018***	0.016**	0.015**	0.016**	0.017***
	(0.003)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)
Maturity FE	No	No	Yes	Yes	Yes	Yes
Date-clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	540	540	528	528	484	484
R^2	0.256	0.379	0.404	0.405	0.471	0.474
Adjusted R^2	0.254	0.376	0.399	0.398	0.466	0.466

Note: This table reports OLS estimates of the one-day post-conference OIS volatility, defined as the min–max range of Euro OIS rates on day $t+1$, on the synthetic-disagreement measure. Synthetic SD ($\hat{\sigma}_t$) is the cross-sectional standard deviation of the LLM agents' rate forecasts (zero-shot, Gemini 2.5-Flash, temperature 1, averaged over 10 sampling runs). Bid–ask spread is the ask minus bid close price one trading day after the Governing Council meeting, standardized within each tenor; higher values denote thinner markets. |Surprise| is the absolute median conference-window change in the OIS rate, measured over the press-conference window on the event day (Altavilla et al. 2019). Maturity FE denotes tenor fixed effects. The reduction from 540 to 528 obs. in Columns (3) and (4) reflects ask quotes availability. The reduction to 484 obs. in Columns (5) and (6) reflects, in addition, 19 conference dates for which the MPD press-conference-window OIS change is unavailable: 18 conferences held from December 2023 onward (beyond the current MPD coverage) and one inter-meeting decision on 2 August 2007. Standard errors (in parentheses) are clustered by conference date. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

The disagreement that survives these controls is what central banks can act on: an ex-ante, text-based measure of conference-induced belief dispersion. Importantly, since it stems from the draft text alone, it can be adjusted before the conference is held.

5 Robustness

We subject the main results to three robustness checks targeting the principal threats to validity in our setting. First, we assess whether the LLM disagreement measure is sensitive to sampling stochasticity: since agent predictions are drawn at temperature 1, repeated simulation runs

may yield different cross-sectional dispersions for the same transcript. Second, we address look-ahead bias by conducting a strict out-of-sample test using exclusively press conferences that post-date Gemini 2.5-Flash's knowledge cut-off, ensuring the model processes genuinely unseen transcripts. Third, we examine a threat specific to our agent-based design: because panel generation and rate forecasting occur within the same prompt in the baseline, the LLM may implicitly condition the characteristics of the 30 synthetic traders on the transcript it has already processed. Thus, the model would manufacture dispersion top-down through panel construction rather than allowing it to emerge organically from heterogeneous processing of a common signal. We address this through a two-stage experiment that separates panel generation, conducted using only the date of the conference, from forecast elicitation.

5.1 Stochasticity and Replicability

Since the model generates responses stochastically, running the same press conference through the simulation twice yields two different sets of trader forecasts and, consequently, two different disagreement estimates. Embedding stochasticity in the model is a deliberate choice as it allows the synthetic panel to explore the full range of plausible interpretations of a given communication rather than collapsing to a single deterministic output, but it also implies that a single simulation draw is a noisy estimate of the underlying conference-level disagreement. If noise were large relative to the signal of interest, our correlations could be spurious. In addition, stochasticity hinders external replicability: independent researchers re-running the simulation on the same transcripts will obtain numerically different values.

To address both concerns, throughout the paper, we had repeated each simulation $R = 10$ times and averaged the standard deviation per conference across draws. In principle, the resulting ensemble disagreement measure should be more stable than any individual draw: idiosyncratic sampling variation cancels across runs, while the conference-specific signal accumulates. Moreover, in this way approximate replicability is preserved: independent researchers running a sufficient number of draws will obtain ensemble averages that converge to the same underlying conference-level signal, even if individual draws differ.

To assess whether our averaging is sufficient, we use a reliability coefficient adapted from psychometrics (Spearman 1910; Brown 1910). For conference $i = 1, \ldots, N$ and run $r = 1, \ldots, R$, let $d_{i,r}$ be the disagreement estimate, and $\bar{d}_i = \frac{1}{R} \sum_{r=1}^{R} d_{i,r}$ the conference mean across runs.

σ^2_{within} is computed for each conference as the sample variance of its R draws around its own mean, then averaged across all conferences:

$$\hat{\sigma}^2_{\text{within}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R-1}\sum_{r=1}^{R}(d_{i,r} - \bar{d}_i)^2 \tag{3}$$

$\sigma^2_{\text{between}}$ is the sample variance of the conference means $\bar{d}_i$ across the N conferences:

$$\hat{\sigma}^2_{\text{between}} = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{d}_i - \bar{\bar{d}})^2 \tag{4}$$

where $\bar{\bar{d}} = \frac{1}{N}\sum_{i=1}^{N}\bar{d}_i$ is the grand mean.

The reliability of an R-run ensemble is then:

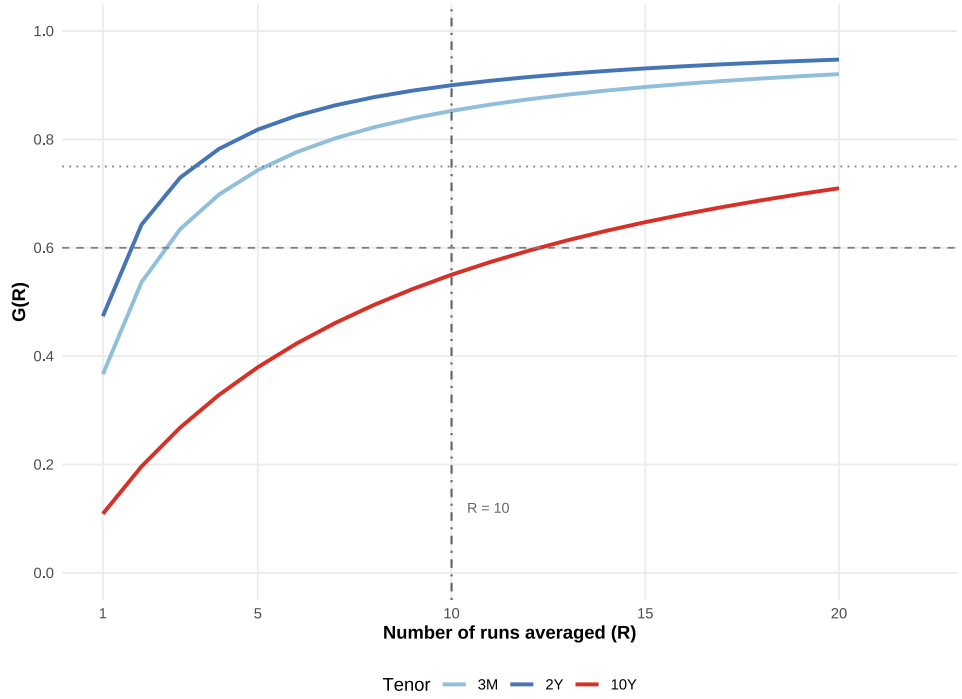
$$G(R) = \frac{\hat{\sigma}^2_{\text{between}}}{\hat{\sigma}^2_{\text{between}} + \frac{\hat{\sigma}^2_{\text{within}}}{R}} \tag{5}$$

which ranges from 0 (dominance of sampling noise) to 1 (perfect signal recovery).[16] Cicchetti (1994) provides conventional benchmarks for interpreting G: values below 0.40 indicate poor reliability, 0.40–0.59 fair, 0.60–0.74 good, and 0.75 or above excellent. Figure 6 plots the analytical $G(R)$ for $R = 1, \ldots, 20$ across the three tenors.[17]

[16]The formula is the Spearman–Brown prophecy formula (Spearman 1910; Brown 1910), originally derived in psychometrics to predict the reliability gain from lengthening a test by a factor of R.

[17]Although in principle we could construct the empirical function, subsetting our sample for each value of R, the estimates at values lower than 10 would be very unstable.

Figure 6: Reliability per Runs Averaged



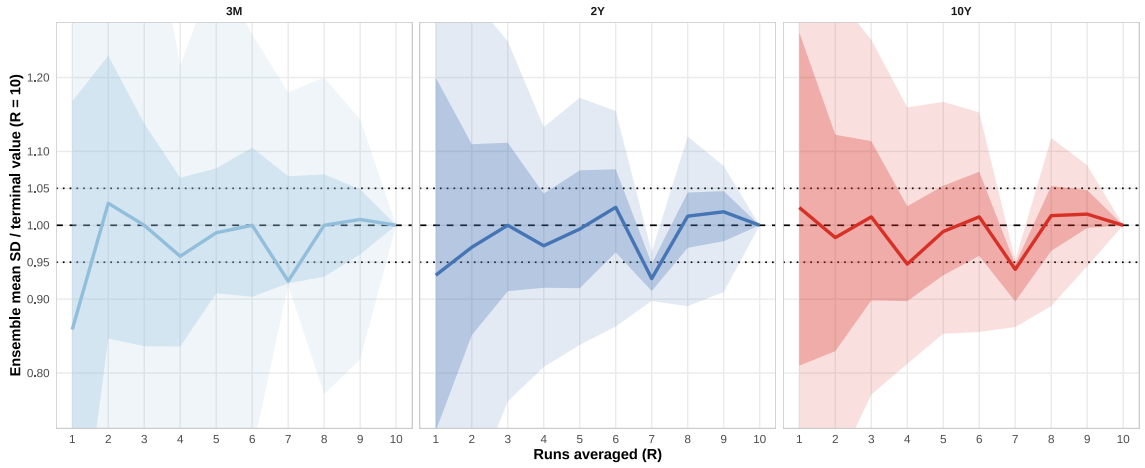
Note: $G(R)$ is the analytical reliability curve with $\hat{\sigma}_{\text{between}}$ and $\hat{\sigma}_{\text{within}}$ estimated from the full $R = 10$ ensemble; the entire curve evaluates the formula at hypothetical run counts, with the vertical marker indicating the observed ensemble size. Values for $R < 10$ and $R > 10$ are back-projected and extrapolated, respectively. Dashed and dotted horizontal lines indicate 0.60 (good) and 0.75 (excellent) reliability thresholds (Cicchetti 1994).

A single draw achieves reliability in the range of 0.1–0.5, whereas the $R = 10$ ensemble reaches 0.75, on average, comfortably above the conventional 0.60 threshold and comparable in magnitude to measurement in high-quality administrative surveys (e.g. Pischke (1995), Schmillen et al. (2024)). Moreover, gains per additional run diminish rapidly, suggesting that $R = 10$ sits near the point of practical convergence.

Figure 7 corroborates this evidence directly: the mean across runs of individual conferences' standard deviation, normalised by its value at $R = 10$, stabilises to within $\pm 5\%$ of the terminal estimate by the seventh run for the bulk of the distribution.

The two diagnostics establish that averaging at $R = 10$ provides a stable and approximately replicable measure of conference-level disagreement. Approximate replicability holds because each simulation draw is an independent realisation conditional on the transcript: independent researchers running a sufficient number of draws will obtain ensemble averages that converge to the same conference-level signal, even if individual draws differ. We conclude that the

Figure 7: Stabilization of Disagreement Estimates over Different Runs



Note: For each tenor, the plot reports the distribution across press conferences of the ensemble mean cross-sectional standard deviation at run count R, normalized by its value at $R = 10$ ($R = 10$ equals 1). Dark band shows the interquartile range; light band shows the 5th–95th percentile range; the line shows the median. Dotted lines indicate ±5% around the $R = 10$ value.

correlations with realised OIS volatility reported in Section 4.1 are therefore robust to the sampling parameter.[18]

5.2 Look-Ahead Bias and Out-of-Sample Validation

A key concern for any LLM-based empirical approach is whether the model genuinely processes textual content or merely exploits patterns memorized during training (e.g. Lopez-Lira et al. (2025)). Since the version of the Gemini 2.5-Flash model we use is trained on data up to January 1, 2025, conferences in our sample from 1998 to 2024 might be affected. To address this methodological concern, we implement several mitigation strategies. First, we explicitly instruct the model to condition their responses on information available only up to the specified conference date, with prompts that emphasize adherence to historical information boundaries (Hansen et al. 2025). This approach attempts to replicate the informational constraints faced by actual market participants at each point in time.[19] Second, from a technical standpoint, we process each press conference through individual API requests rather than batching multiple conferences together. While batching would significantly reduce computational time, it would

[18] In Appendix 6.2 we extend the reliability framework to both prompt and model stability, the two other sources of variation in a LLM call.

[19] In principle it should lower the probabilities of next-word tokens being the ones that include future information.

risk contaminating earlier forecasts with information from subsequent conferences within the same batch. Our sequential, single-conference approach ensures that the model's context window contains only the target conference and its associated prompt, preventing any inadvertent temporal spillover between events.

Nevertheless, while these safeguards reduce the risk of temporal leakage within the training sample, they cannot fully address the possibility that analyses of ECB communication were included in the model's training corpus. To provide definitive evidence that our results reflect genuine textual processing rather than memorization, we conduct a strict temporal out-of-sample test using only conferences that post-date the model's training period. Specifically, we re-run the LLM simulation exclusively on ECB press conferences held between January 2025 and March 2026, a period entirely outside the model's knowledge cutoff. This design ensures the model processes genuinely unseen transcripts, providing a clean test of generalization. While the out-of-sample period is necessarily limited, this represents the strongest possible validation of our approach.

Figure 8 presents the results, pooling observations across all three tenors.[20]

The model's performance on this genuinely unseen data is, as expected, inferior to the in-sample exercise, but, nevertheless, synthetic disagreement exhibits high correlation with market values: 0.43 across 30 tenor-date observations.

One observation warrants closer examination: the March 2026 conference registers markedly elevated realized volatility relative to the synthetic disagreement generated by the model, especially so for the 2 and 10 years maturities. The outsized market reaction that day was driven primarily by the sudden escalation of the Middle East conflict. Our LLM agents, reading only the communication, correctly assessed it as high ambiguity, but failed to portray the full extent of belief heterogeneity about the duration and impact of the conflict. The gap between synthetic and realized uncertainty reflects a clean separation between *communication-induced* uncertainty, which the model captures, and *event-driven* uncertainty originating outside the transcript.

All in all, this test demonstrates that our LLM-based approach does not memorize history but rather interprets it, extracting transferable linguistic features that capture genuine policy uncertainty in conversations it has never encountered.[21]

[20] We pool across tenors to improve statistical power given the limited amount of observatons available.
[21] We also verified that our Gemini API configuration does not enable external grounding or retrieval

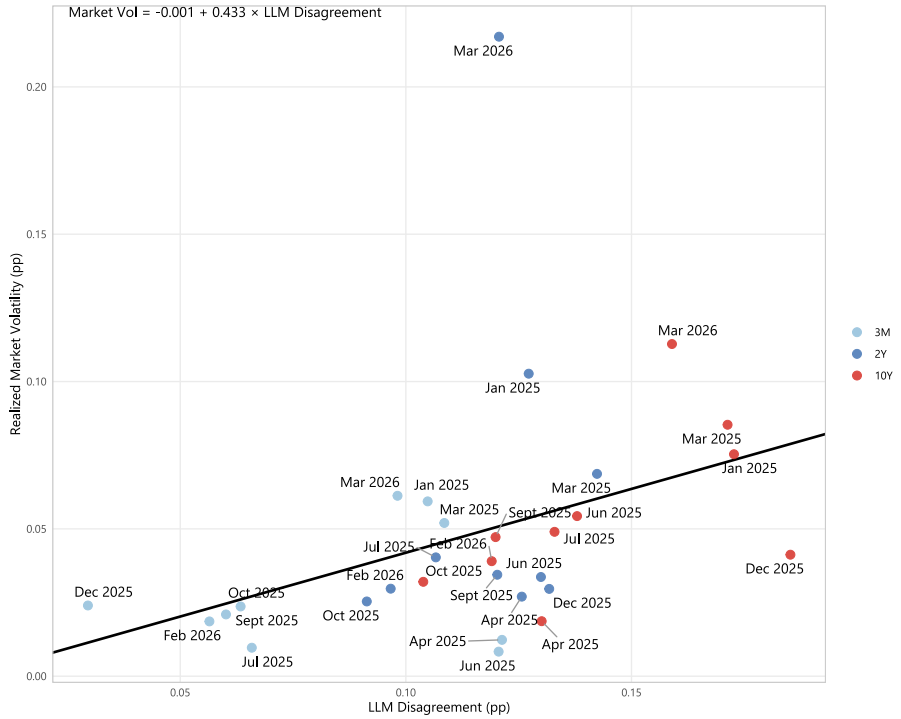


Figure 8: Out-of-Sample Performance — LLM Disagreement vs Market Volatility

Note: Each point represents an ECB press conference post January 1, 2025 (after Gemini 2.5-Flash knowledge cut-off). Synthetic disagreement is measured as the cross-sectional standard deviation of rate predictions across 30 heterogeneous agents generated using Gemini 2.5-Flash with the zero-shot prompt (Figure 2) and temperature 1. Market volatility is the min-max range of OIS rates 1 day after ECB press conferences. For each press conference, the model output is generated across 10 independent sampling runs and then averaged to ensure robustness.

5.3 Endogeneity of Agent-Panel Composition

A final caveat with the agent-based simulation is within-call confounding. Because panel generation and rate forecasting occur within the same prompt in the baseline design, the LLM may implicitly condition the characteristics of the 30 synthetic traders on the transcript it has already processed. If so, $\widehat{\sigma}_t$ would not be a pure measure of disagreement arising from heterogeneous processing of a common signal: part of its variation would instead reflect the LLM's implicit assessment of transcript complexity, encoded directly into panel composition rather than into forecasts. Hence, the correlation between $\widehat{\sigma}_t$ and realised OIS volatility could be inflated because of this.

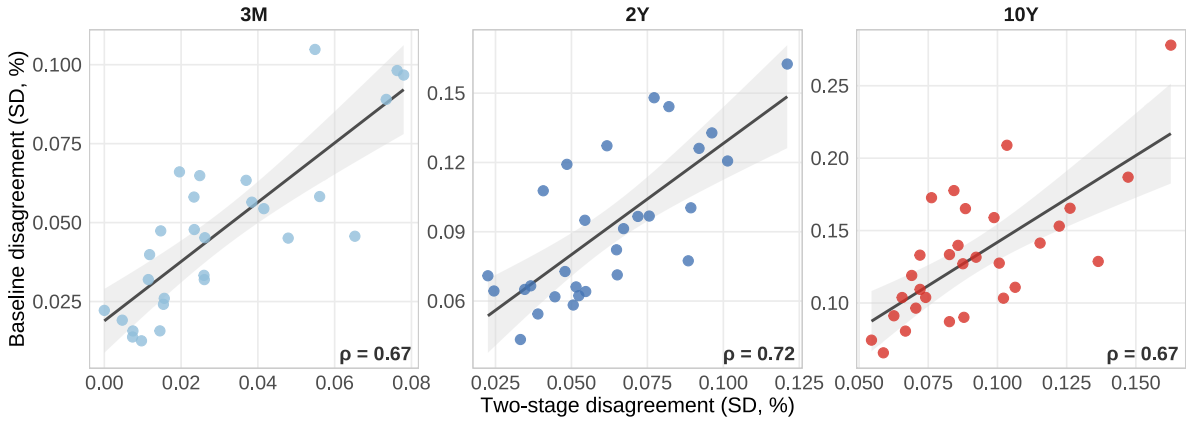
To obviate to this potential issue, we split the simulation process in two stages. In *Stage 1*, the LLM constructs the panel of 30 synthetic traders using only the macroeconomic regime

tools. Specifically, the `google_search_retrieval` parameter is disabled, and API response metadata confirm the absence of `groundingMetadata` fields. This ensures the model operates solely from its training corpus and cannot access real-time information about post-cutoff conferences.

prevailing around conference date t, with an explicit instruction not to condition on any press conference content (Figure 17). In *Stage 2*, each frozen panel is re-presented to the LLM together with the actual transcript, and individual rate forecasts are elicited exactly as in the baseline. Critically, the panel cannot change between stages: whatever heterogeneity the LLM encodes into trader characteristics in Stage 1 is fixed prior to transcript exposure, so any dispersion in Stage 2 forecasts must arise from heterogeneous processing of a common signal rather than from endogenous panel composition (Figure 18).[22] We apply the test to a stratified random sample of 30 conferences, drawing 10 per tercile of realised post-meeting OIS volatility, so that the sample spans low-, medium-, and high-volatility episodes equally.[23]

Figure 9 plots the two disagreement measures against each other. Spearman rank correlations between the baseline and two-stage measures are 0.67 at 3M, 0.72 at 2Y, and 0.67 at 10Y, indicating broadly similar rank orderings across both designs at every tenor.

Figure 9: Correlation between Baseline and Two-Stage Disagreement



Note: Shaded bands are 95% confidence intervals. Spearman rank correlations ρ are reported in the bottom-right corner of each panel.

Moreover, the difference in Spearman correlations between the baseline and two-stage designs is small at the longer tenors and, where it is not, runs in the direction that favours our interpretation (Table 3). At the 2-year and 10-year tenors $\Delta\rho$ is essentially zero (0.024 and 0.029 respect-

[22]The simulation has two sources of randomness: which synthetic traders are selected (Stage 1) and what forecasts they produce (Stage 2). To capture both, we draw a fully independent panel for each of the $R = 10$ runs rather than fixing the panel across runs. The disagreement measure is then the average within-run standard deviation of trader forecasts across the ten runs.

[23]Terciles are based on the average post-conference OIS volatility across tenors. We consider only conferences with data for all three tenors.

ively). At the 3-month tenor the evidence is stronger, but points the other way: $\rho_{ts} = 0.511$ exceeds $\rho_{hl} = 0.217$, yielding $\Delta\rho = -0.295$ with a 95% confidence interval that just excludes zero. Thus the two-stage design if anything outperforms the headline at the short end.

Table 3: Endogeneity test: Spearman correlations with realised market volatility

Tenor	ρ_{ts}	ρ_{hl}	$\Delta\rho$	95% CI
3M	0.511***	0.217	-0.295†	[-0.598, -0.005]
2Y	0.429**	0.453**	0.024	[-0.259, 0.305]
10Y	0.377**	0.406**	0.029	[-0.203, 0.261]

Note: This table compares the rank correlation with realized post-ECB conferences OIS volatility of two different synthetic disagreement measures. ρ_{hl} is the correlation for the headline zero-shot ensemble, in which the agent panel and its forecasts are generated jointly from the transcript. ρ_{ts} is the correlation for the two-stage design, in which the panel is generated first from the macroeconomic regime alone and forecasts are elicited only at the second stage. The two designs are harmonized to differ only in generation timing. $\Delta\rho \equiv \rho_{hl} - \rho_{ts}$. Confidence intervals are 95% bootstrap percentile intervals ($R = 5{,}000$ replications). Stars denote significance of each correlation against $H_0 : \rho = 0$ ($^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$); † marks a $\Delta\rho$ whose 95% interval excludes zero.

Separating panel construction from forecasting thus leaves the disagreement signal intact or stronger, the opposite of what within-call inflation would produce. This result endorses our reading of the measure as heterogeneous processing of a common signal.

6 Conclusions

This paper shows that LLMs can simulate the heterogeneous interpretation of central bank communication, and that the disagreement they generate is a measurable, construct-valid signal. Across 293 ECB press conferences from 1998 to 2026, the cross-sectional dispersion among 30 synthetic traders, each endowed with distinct risk preferences and behavioral biases, correlates at approximately 0.5 with realized post-conference OIS volatility. The measure outperforms off-the-shelf textual benchmarks and carries information beyond volatility persistence. Moreover, we show that it reflects how policy is communicated rather than what is decided and that it represent broad dispersion of beliefs as opposed to thin-market price movement. Together these tests pinpoint our synthetic disagreement as a measure of interpretive disagreement induced by the language of a conference.

The result has a clear use. The framework can be applied to draft language before a conference is held, giving communicators an ex-ante gauge of the interpretive disagreement a given phrasing is likely to provoke. For research, it offers a micro-founded approach to expectation formation that models heterogeneous belief processes explicitly rather than treating them as a black box. Lastly, the paper supports the view proposed by recent literature that LLMs can serve as computational representation of humans (e.g. Horton (2023)).

Several limitations bound these results and mark directions for future work. First, the out-of-sample window, though genuinely unseen, is necessarily short, and broader testing across policy regimes is warranted as data accumulate. Second, domain-specific fine-tuning on central bank communication, in the spirit of MILA (Bundesbank 2025), may improve on the general-purpose models used here. Third, the framework could be extended to a full ABM-LLM setting in which agents trade among themselves and equilibria emerge dynamically (del Rio-Chanona et al. 2025). Finally, the diagnostic measure invites a theoretical counterpart, a model of optimal communication design that trades clarity against the central bank's other objectives.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). Prediction, judgment, and complexity: A theory of decision-making and artificial intelligence. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*, pages 89–110. University of Chicago Press.

Altavilla, C., Brugnolini, L., Gürkaynak, R. S., Motto, R., and Ragusa, G. (2019). Measuring euro area monetary policy. *Journal of Monetary Economics*, 108:162–179.

Axtell, R. L. and Farmer, J. D. (2025). Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, 63(1):197–287.

Banerjee, S. and Kremer, I. (2010). Disagreement and learning: Dynamic patterns of trade. *The Journal of Finance*, 65(4):1269–1302.

Barberis, N. and Thaler, R. (2003). A survey of behavioral finance. In *Handbook of the Economics of Finance*, volume 1, pages 1053–1128. Elsevier.

Bauer, M. D., Lakdawala, A., and Mueller, P. (2021). Market-based monetary policy uncertainty. *The Economic Journal*, 132(644):1290–1308.

Bauer, M. D. and Swanson, E. T. (2023). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, 37:87–155.

BIS (2024). Cb-lms: language models for central banking. Technical Report BIS Working Paper No. 1215, Bank for International Settlements.

Bolhuis, M., Das, S., and Yao, Y. (2024). A new dataset of high-frequency monetary policy shocks. Technical report, International Monetary Fund.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2018). Diagnostic expectations and credit cycles. *Journal of Finance*, 73(1):199–227.

Brown, T. B. et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322.

Bundesbank, D. (2025). Monetary-intelligent language agent. Technical report, Deutsche Bundesbank Technical Paper.

Cesa-Bianchi, A., Thwaites, G., and Vicondoa, A. (2020). Monetary policy transmission in the united kingdom: A high frequency identification approach. *European Economic Review*, 123:103375.

Chopra, A., Kumar, S., Kuru, N. G., Raskar, R., and Quera-Bofarull, A. (2025). On the limits of agency in agent-based models. In *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 500–508.

Christiano Silva, T., Moriya, K., and Veyrune, R. (2025). From text to quantified insights: A large-scale llm analysis of central bank communication. *IMF Working Papers*, 2025(109):1.

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychological Assessment*, 6:284–290.

Clayton, C., Coppola, A., Maggiori, M., and Schreger, J. (2025). Geoeconomic pressure. Working Paper 34020, National Bureau of Economic Research.

Collodel, U. and Kunzmann, V. (2025). Market-based monetary policy uncertainty shocks in the euro area. Technical report, CBM Working Papers.

Das, S. and Song, W. (2023). Monetary policy transmission and policy coordination in china. *China Economic Review*, 82:102032.

del Rio-Chanona, R. M., Pangallo, M., and Hommes, C. (2025). Can generative ai agents behave like humans? evidence from laboratory market experiments.

Delli Gatti, D., Gaffeo, E., Gallegati, M., Desiderio, S., and Cirillo, P. (2011). *Macroeconomics from the bottom-up*. Springer.

ECB (2025). Report on monetary policy tools, strategy and communication. Occasional Paper Series 372, European Central Bank.

Farmer, J. D. and Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460(7256):685–686.

Gurkaynak, R. S., Sack, B., and Swanson, E. T. (2005). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 1(1):55–93.

Gürkaynak, Refet S., R. S. (2005). Using federal funds futures contracts for monetary policy analysis. Technical Report 2005-29, Federal Reserve Board.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A. (2025). Simulating the survey of professional forecasters. Technical report, SSRN. Working Paper 5066286.

Harris, M. and Raviv, A. (1993). Differences of opinion make a horse race. *The Review of Financial Studies*, 6(3):473–506.

Hommes, C. H. (2006). Chapter 23 heterogeneous agent models in economics and finance. volume 2 of *Handbook of Computational Economics*, pages 1109–1186. Elsevier.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

Jarociński, M. and Karadi, P. (2020). Deconstructing monetary policy surprises—the role of information shocks. *American Economic Journal: Macroeconomics*, 12(2):1–43.

Kazinnik, S. and Sinclair, T. (2025). Fomc in silico: A multi-agent system for monetary policy decision modeling. Working Papers 2025-005, The George Washington University, The Center for Economic Research.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of Monetary Economics*, 47(3):523–544.

Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). The memorization problem: Can we trust LLMs' economic forecasts? Technical report, SSRN. Working Paper 5217505.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Mullainathan, S. and Spiß, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Mumtaz, H., Saleheen, J., and Spitznagel, R. (2023). Keep it simple: Central bank communication and asset prices. Technical report, Queen Mary University of London, School of Economics and Finance.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, 53(1):61–65.

Pfeifer, M. and Marohl, V. P. (2023). Centralbankroberta: A fine-tuned large language model for central bank communications. *The Journal of Finance and Data Science*, 9:100114.

Pischke, J.-S. (1995). Measurement error and earnings dynamics: Some estimates from the psid validation study. *Journal of Business & Economic Statistics*, 13(3):305–314.

Schmillen, A., Umkehrer, M., and von Wachter, T. (2024). Measurement error in longitudinal earnings data: Evidence from germany. *Journal for Labour Market Research*, 58(1):1–31.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Representation Learning*, volume 2024, pages 25055–25083.

Shalen, C. T. (1993). Volume, volatility, and the dispersion of beliefs. *The Review of Financial Studies*, 6(2):405–434.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295.

Tillmann, P. (2020). Monetary policy uncertainty and the response of the yield curve to policy shocks. *Journal of Money, Credit and Banking*, 52(4):803–833.

Appendix A: Additional Sensitivity Checks

6.1 Different Correlation Measures

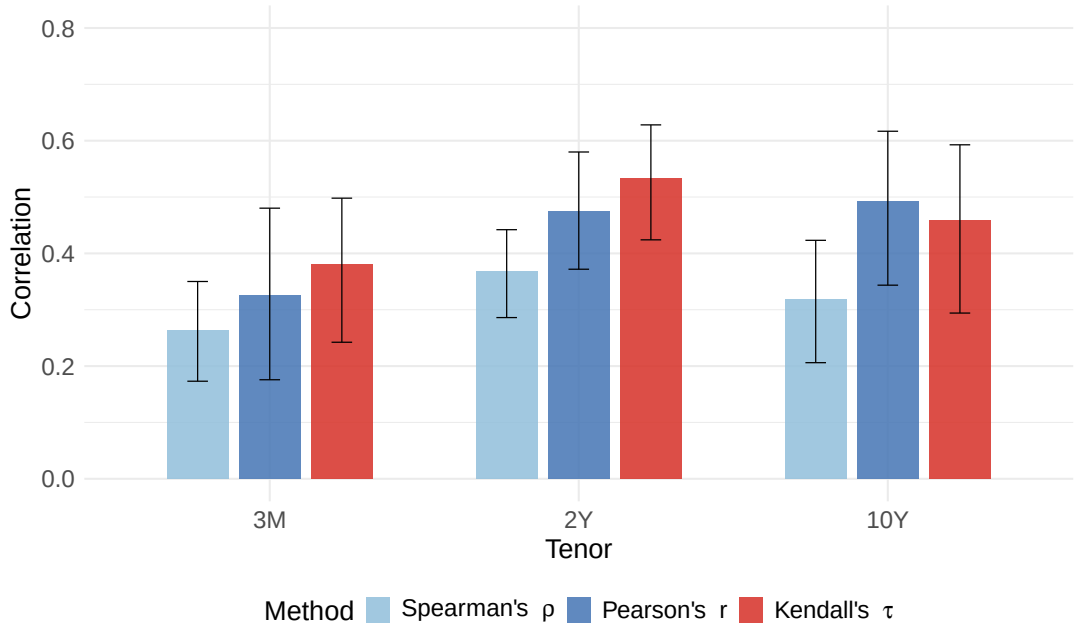
To ensure the robustness of our main findings, we conduct a supplementary analysis verifying that the observed relationship between synthetic and market-based disagreement is not dependent on the choice of the correlation metric. Our baseline analysis employs the **Spearman rank correlation coefficient** (ρ), which is well-suited for this context as it captures monotonic relationships and is robust to outliers and non-linearities.

For a more comprehensive validation, we also consider two additional measures: the **Pearson linear correlation coefficient** (r) and **Kendall's rank correlation coefficient** (τ):

- **Pearson's** r quantifies the strength of a strictly *linear* relationship between two variables. Although more sensitive to outliers, it allows us to check whether the relationship could plausibly be explained within a linear framework.
- **Kendall's** τ provides another non-parametric, rank-based alternative that evaluates concordance between two rankings. It is often preferred in smaller samples or in the presence of tied ranks, offering an additional robustness check alongside Spearman's ρ.

To assess the precision of these estimates, we compute **95% confidence intervals** via a non-parametric bootstrap with 5,000 replications. Figure 10 presents the results.

Figure 10: Comparison of Correlation Measures Between Synthetic and Market-Based Disagreement.



Note: Correlation between between the output of an LLM model prompted in a zero-shot setting using Gemini 2.5-Flash, with a temperature parameter of 1, and the OIS min-max range 1 day after ECB conferences. The simulation uses a sample of 293 ECB press conferences (June 1998–March 2026) and the baseline prompt (Figure 2). Correlation measured using Spearman's rank correlation (ρ), Pearson's linear correlation (r), and Kendall's rank correlation (τ). Whiskers represent 95% confidence intervals computed via a non-parametric bootstrap with 5,000 replications.

Across all three metrics, we find a statistically and economically significant positive association between our LLM-generated disagreement measure and market volatility. The 2-year tenor exhibits the strongest association in the rank-based measures, roughly 0.5, highlighting the simulation's effectiveness at capturing uncertainty around the medium-term monetary policy path.

Importantly, the bootstrapped 95% confidence intervals for all coefficients across all tenors comfortably exclude zero, underscoring statistical significance. While the magnitudes vary across correlation metrics, consistent with their differing assumptions, the overall conclusion remains unchanged. The alignment of results across linear and rank-based approaches demonstrates that the link we identify is a robust and stable feature of the data.

6.2 Prompt and Model Stability Analysis

Beyond sampling stochasticity, two further implementation choices could in principle affect our disagreement measure: the specific prompt p used to elicit forecasts, and the underlying LLM m. We assess robustness along both dimensions using the same reliability framework as in Section 5.1, borrowing from Clayton et al. (2025).

Framework

For each conference t, the disagreement measure depends on the prompt p, the model m, and a sampling realisation s. We model the resulting outcome as

$$\sigma_t(p, m, s) \;=\; \bar{\eta} \;+\; \eta_t \;+\; \varepsilon_t(p, m, s), \tag{6}$$

where η_t is the conference-level signal, the across-conference variation of interest, and $\varepsilon_t(p, m, s)$ is the noise introduced as we span the set of admissible equivalent prompts $\tilde{\mathcal{P}}$, vary the model across a set $\tilde{\mathcal{M}}$, and draw fresh sampling realisations. Define

$$\sigma^2_{\text{between}} \;=\; \text{Var}(\eta_t), \qquad \sigma^2_d \;=\; \text{Var}\big(\varepsilon_t \mid d\big), \quad d \in \{\text{sampling}, \text{prompt}, \text{model}\},$$

where σ^2_d is the variance of ε_t along dimension d holding the other two fixed at their baseline values. The reliability of a single measurement of σ_t along dimension d is

$$G_d \;=\; \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_d}, \tag{7}$$

which is the share of variance in a single draw of σ_t attributable to genuine cross-conference variation rather than to noise along dimension d.

Estimation

For each dimension $d \in \{\text{prom}, \text{mod}\}$, let $\mathcal{D}_d$ denote the set of perturbations (size $K = 15$ for prompts, $M = 3$ for models), and let $\bar{\sigma}_t^{(d)} = |\mathcal{D}_d|^{-1} \sum_{\delta \in \mathcal{D}_d} \sigma_t(\delta)$ be the conference-level mean

across perturbations. The between-conference variance is estimated as

$$\hat{\sigma}^2_{\text{between},d} \;=\; \text{Var}_t\big(\bar{\sigma}_t^{(d)}\big), \tag{8}$$

and the within-conference noise variance as

$$\hat{\sigma}^2_d \;=\; \frac{1}{T}\sum_{t=1}^{T} \text{Var}_{\delta\in\mathcal{D}_d}\big(\sigma_t(\delta)\big), \tag{9}$$

where $T = 293$. The G-coefficient is then

$$\hat{G}_d = \hat{\sigma}^2_{\text{between},d}/(\hat{\sigma}^2_{\text{between},d} + \hat{\sigma}^2_d) \tag{10}$$

Since prompt and model perturbations are each evaluated at a single draw ($R = 1$) due to computational cost constraints, the noise variance σ^2_d absorbs residual sampling stochasticity in addition to true implementation sensitivity; the corresponding G-coefficients therefore represent approximate lower bounds on reliability rather than exact variance decompositions.

6.2.1 Prompt Stability

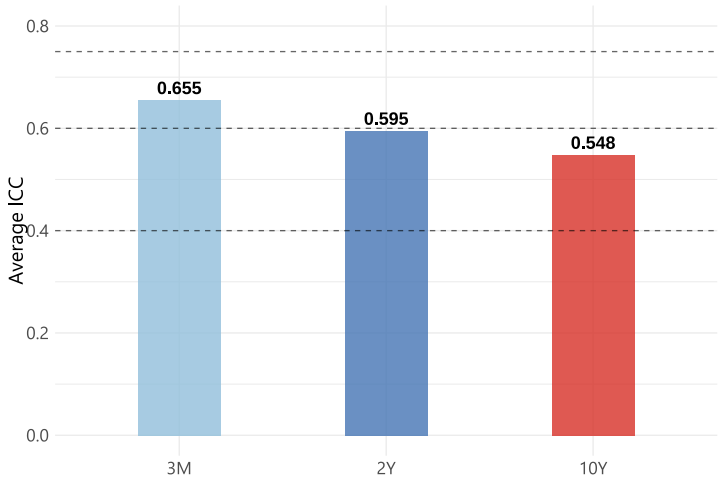
LLMs are particularly sensitive to prompt variations, with "performance differences of up to 76 percentage points for subtle formatting changes" (Sclar et al. 2024). To assess this vulnerability in our context, we generate 15 prompt perturbations of our baseline specification, including 10 minor and 5 medium variations. Minor variations substitute semantically equivalent alternatives and reorder prompt components, while medium variations employ substantively different language to describe trader characteristics and market conditions while maintaining the overall framework. We report the full list of variations in Table 4.

Table 4: List of Prompt Variations

Variation	Type	Key Changes	Modified Section
1	Minor	Individual traders → distinct market participants	Context
2	Minor	Trading decision → investment choice	Context
3	Minor	Risk aversion order: High/Medium/Low → Low/Medium/High	Trader Characteristics
4	Minor	Interpretation of conference → analysis of meeting	Context
5	Minor	Markdown table → structured table format	Guidelines
6	Minor	Confidence score → certainty level	Task
7	Minor	Monetary policy decisions → interest rate policies	Context
8	Minor	Maximize profit → optimize returns	Context
9	Minor	Reordered behavioral biases (Anchoring first)	Trader Characteristics
10	Minor	Press conference → policy meeting throughout	Multiple sections
11	Medium	Formal academic language; market agents	Context
12	Medium	Behavioral lenses; psychological profiles emphasis	Context
13	Medium	Risk tolerance: Conservative/Moderate/Aggressive	Trader Characteristics & Context
14	Medium	Policy signals; forward guidance; processing approach	Context & Characteristics
15	Medium	Position-taking behavior; different output headers	Task & Output

Given the large computational resources required to run all perturbations on all conferences (283 conferences $\times$ 15 perturbations = 4,245 API calls), we select a random subsample of 30 conferences to balance statistical power with computational feasibility. Figure 11 shows the results.

Figure 11: Reliability across Prompt Variations by Tenor



Note: Average reliability values across all prompt variations (10 minor and 5 medium) for each tenor. Higher reliability indicates greater measurement stability. Variations are based on prompt 2 and obtained by calling Gemini 2.5-Flash. Dashed horizontal lines indicate 0.4 (fair), 0.6 (good), and 0.75 (excellent) thresholds based on Cicchetti (1994).

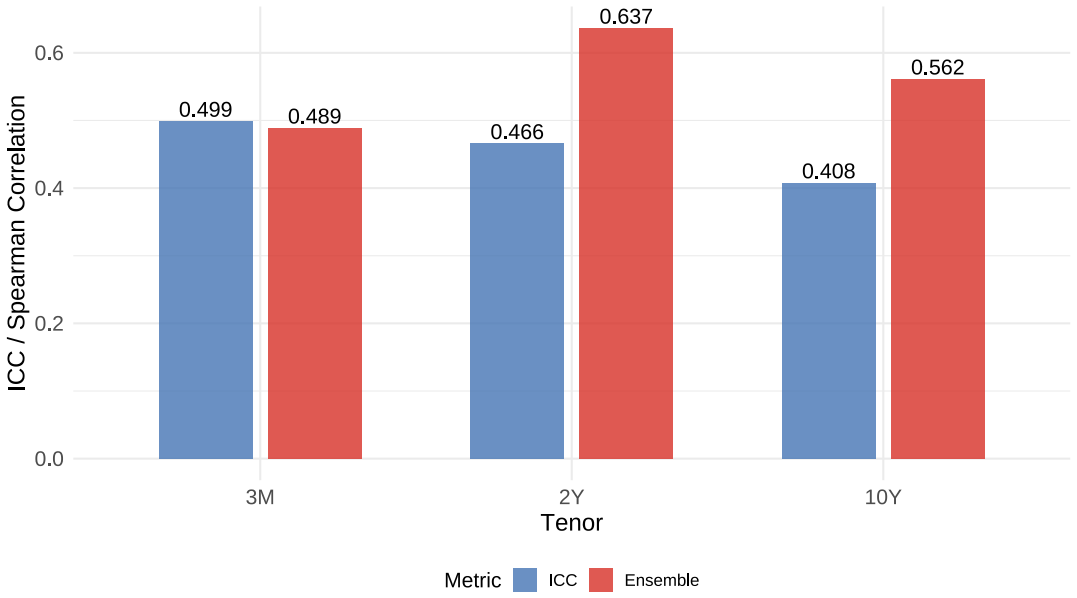
Reliability follows a clear maturity hierarchy: the 3-month tenor exhibits the highest overall reliability (average ICC = 0.655), followed by the 2-year (ICC = 0.595) and 10-year (ICC = 0.548) tenors (Figure 11). The perturbation analysis yields reassuring results across all tenor-specification combinations. While reliability values do not uniformly reach excellent standards, they are best interpreted as conservative lower bounds: since each perturbation is evaluated at a single draw ($R = 1$), sampling noise inflates greatly the estimated noise variance and mechanically depresses the G-coefficient. The true prompt reliability plausibly lies above these estimates, suggesting that conference-specific variation is the dominant source of variation in our measure.

6.2.2 Model Stability

We assess reliability across three distinct LLM architectures: Gemini, ChatGPT, and Claude.[24] These models were selected to capture diversity in training data, design philosophy, and reasoning capabilities. For each ECB conference, we generate synthetic disagreement measures, compute correlations with market-based values, and evaluate cross-model agreement.

[24]Specifically, we use Gemini 2.5-Flash, ChatGPT-5o-mini, and Claude-4.5-Sonnet.

Figure 12: Reliability across LLMs by Tenor and Ensemble Model Correlation with Market-based disagreement

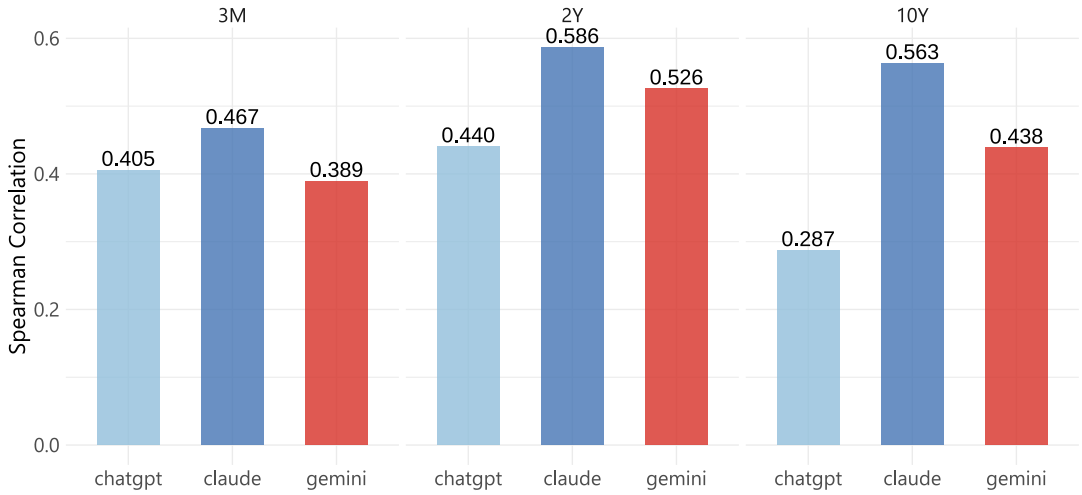


Note: Blue bars show reliability for each tenor calculated using three distinct LLM architectures (Gemini 2.5 Flash, ChatGPT 5o-mini, Claude 4.5 Sonnet). Higher reliability indicates greater measurement stability. Red bars show the Spearman correlation between the ensemble of LLM outputs, a simple average, and market-based disagreement measured as the min-max range of OIS rates 1 day after ECB press conferences.

The results show fair agreement among models and the same tenor hierarchy displayed in the prompt perturbation exercise: 3-month tenor (ICC = 0.499), 2-year tenor (ICC = 0.466), and 10-year tenor (ICC = 0.408). These ICC values fall in the moderate range, indicating that model-specific differences, however, contribute meaningfully to the observed variance. This raises an important question: do these differences reflect measurement noise, or do different models capture complementary signal?

Two pieces of evidence support the latter interpretation. First, each individual model achieves substantial correlations with market-based disagreement, approximately 0.5 across all tenors (Figure 13), demonstrating that all three architectures successfully capture the relationship between textual features and market reactions, albeit through different pathways. Second, and more tellingly, an ensemble measure constructed as the simple average of all three models' outputs achieves higher correlations with market disagreement than any individual model (Figure 12). This performance gain from ensembling is inconsistent with pure noise; if model differences reflected only measurement error, averaging would not improve predictive accuracy.

Figure 13: Individual Model Correlations with Market-Based Disagreement by Tenor



Note: Spearman correlation coefficients between synthetic disagreement measures generated by individual LLM models (Gemini 2.5-Flash, ChatGPT 5o-mini, and Claude Sonnet 4.5) and post-conference OIS volatility across three tenors (3-month, 2-year, and 10-year). Each model generates synthetic disagreement by simulating 30 heterogeneous traders interpreting ECB press conference transcripts. Post-conference market volatility is measured as the min-max range of OIS rates 1 day after ECB press conferences. Sample covers 283 ECB Governing Council meetings from June 1998 to April 2025.

These findings support a specific interpretation: while individual LLMs are not perfectly interchangeable, their systematic differences capture distinct but relevant aspects of communication interpretation. Different models may weight linguistic features differently, syntactic structure versus semantic content, explicit statements versus implied meanings, yet all extract genuine signal from ECB communication.

Appendix B: Alternative Modeling Frameworks

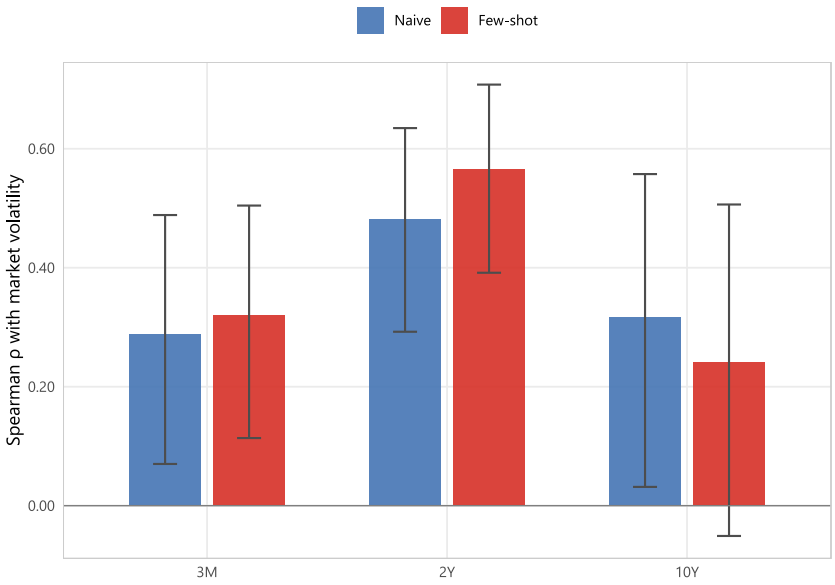
6.3 Historical Anchoring

In this section we turn to the role of additional information in the prompt. Hansen et al. (2025) show that incorporating lagged median forecasts from SPF survey participants improves an LLM's ability to replicate forecasts in subsequent survey rounds. This finding is consistent with the few-shot learning literature, which shows that providing relevant examples in the prompt can improve predictive accuracy without incurring in the computational cost associated with updating model parameters (Brown et al. 2020).

To assess this mechanism in our setting, we augment the baseline prompt with the real-

ized before- and after-conference OIS min–max range from the three most recent meetings, computed separately for each maturity. The intuition is straightforward: a trader operating in financial markets observes not only the content of the ECB's communication, but also the volatility environment surrounding recent meetings. The few-shot prompt seeks to replicate this informational context. As in the previous section, we then compute Spearman correlations with market-based measures.[25] Figure 14 reports the results.

Figure 14: Mean Spearman correlation with market-based measures by tenor



Note: Correlation between the output of an LLM model prompted in a few-shot setting using Gemini 2.5-Flash, with temperature 1, and the OIS min-max range 1 day after ECB conferences. The simulation uses a subset of 90 ECB press conferences, 30 drawn from each tercile of realized post-conference OIS volatilityy, from the original sample of 293 ECB press conferences (June 1998–March 2026) and the prompt with historical anchoring (Figure 16). 90% confidence intervals computed via a non-parametric bootstrap with 5,000 replications. For each press conference, the model output is generated across 10 independent sampling runs and then averaged to ensure robustness.

In terms of ranking i.e., which conferences generate more or less disagreement, the two prompting strategies are statistically indistinguishable. The Spearman correlation between the benchmrk and few-shot disagreement series exceeds 0.7 at all maturities, and differences in their respective correlations with realized market volatility are never statistically significant. Both prompts identify the same press conferences as high- or low-uncertainty events, suggesting that

[25] We run the simulation on a subset of 90 ECB press conferences, 30 drawn from each tercile of realized post-conference OIS volatility, in order to minimize computational cost without sacrificing statistical power. As in the previous section, we average the standard deviation for each conference across 10 runs to erase sampling noise and allow comparison.

this signal is intrinsic to the model's interpretation of the transcript rather than dependent on historical anchoring.

Where the two approaches diverge sharply, however, is in the level of predicted disagreement. Table 5 reports the resulting biases and mean absolute errors for the baseline and the specification with anchoring.

Table 5: Bias and Mean Absolute Error in Baseline and Historical Anchoring

Tenor	N	Bias (N)	Bias (FS)	MAE (N)	MAE (FS)	Delta MAE
3M	89	0.014***	-0.007**	0.030	0.026	0.004
2Y	89	0.020***	-0.018***	0.044	0.036	0.009**
10Y	53	0.062***	0.007	0.070	0.041	0.029***
Pooled	231	0.028***	-0.008**	0.045	0.033	0.012***

Note: Bias N and Bias FS denote the mean difference between the baseline and historically anchored LLM disagreement measure and realized market volatility, respectively. MAE(N) and MAE(FS) are the corresponding mean absolute errors. Δ MAE = MAE(N)$-$ MAE(FS) $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The baseline systematically over-predicts realized volatility across all maturities: the estimated bias is positive and statistically significant throughout, with the distortion particularly pronounced at the 10-year tenor. The few-shot prompt largely corrects this bias: estimates become small and, at the 10-year maturity, statistically indistinguishable from zero. Correspondingly, the mean absolute error declines substantially, by roughly 19 percent at the 2-year and 41 percent at the 10-year maturity, when historical context is included.

6.4 LLM-as-Judge Implementation Details and Results

Algorithm 1 LLM-as-a-Judge Prompt Optimization Algorithm

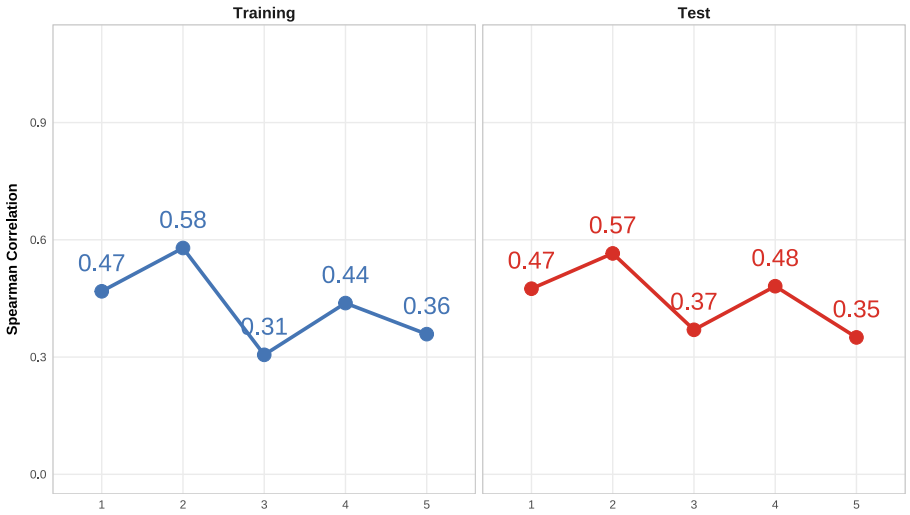
Require: Set of N historical transcripts $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$
Require: Initial Analyst LLM prompt $P_{\text{analyst},0}$
Require: Judge LLM prompt P_{judge}
Require: Number of optimization iterations M
Require: Correlation threshold $C_{\text{threshold}}$
Require: Analyst LLM A, Judge LLM J
Require: Number of agents M
Ensure: Optimized Analyst LLM prompt P^*_{analyst}

Split $\mathcal{T}$ into training set $\mathcal{T}_{\text{train}}$ and test set $\mathcal{T}_{\text{test}}$
Initialize optimization history $\mathcal{H} \leftarrow \emptyset$
Initialize $P_{\text{analyst,current}} \leftarrow P_{\text{analyst},0}$
Initialize $P^*_{\text{analyst}} \leftarrow P_{\text{analyst},0}$
Initialize $C_{\max} \leftarrow -\infty$
for $m = 1$ to M **do** ▷ Phase 1: Simulate Rate Predictions
  $\mathcal{D}_{\text{sim}} \leftarrow \emptyset$
  for all $T_i \in \mathcal{T}_{\text{train}}$ **in parallel do**
    Construct user message for Analyst LLM using T_i
    $D_{\text{sim},i} \leftarrow A(\text{model} = \text{Gemini 2.5 Flash}, \text{prompt} = P_{\text{analyst,current}}, \text{temperature} = 1)$
    Parse $D_{\text{sim},i}$ into structured format
    Add $D_{\text{sim},i}$ to $\mathcal{D}_{\text{sim}}$
  end for
  Compute average correlation of simulated standard deviation with market-based
  ▷ Phase 2: Judge LLM Evaluation and Prompt Update
  Construct Judge LLM input with:
    - Current prompt $P_{\text{analyst,current}}$
    - Current correlation between simulated and market-based volatility
    - Optimization history $\mathcal{H}$
  $O_{\text{judge}} \leftarrow J(\text{model} = \text{Gemini 2.5 Pro}, \text{prompt} = P_{\text{judge}}, \text{temperature} = 0.0)$
  Parse O_{judge} to extract new prompt $P_{\text{analyst,next}}$ and rationale
  Append $\{P_{\text{analyst,current}}, \text{avg\_correlation}, O_{\text{judge}}\}$ to $\mathcal{H}$
  $P_{\text{analyst,current}} \leftarrow P_{\text{analyst,next}}$
  $C_{\text{current}} \leftarrow \text{avg\_correlation}(P_{\text{analyst,current}})$
  if $C_{\text{current}} > C_{\max}$ **then**
    $C_{\max} \leftarrow C_{\text{current}}$
    $P^*_{\text{analyst}} \leftarrow P_{\text{analyst,current}}$
  end if
  if $C_{\max} \geq C_{\text{threshold}}$ **then**
    break
  end if
end for
▷ Phase 3: Final Evaluation
Evaluate P^*_{analyst} on $\mathcal{T}_{\text{test}}$ using Analyst LLM
Compute final performance metrics (correlation)
  return P^*_{analyst}

The "Judge" framework introduces an explicit feedback loop: a separate LLM evaluates agent forecasting performance and rewrites the first LLM prompt to improve in-sample alignment with market volatility, allowing us to assess whether prompt optimization alone can account for performance gains. We run 5 iterations on a training set before evaluating every prompt produced on the held-out test set.[26]

Figure 15 reveals a clear performance arc.

Figure 15: Correlation with market-based measures across iterations — training and test set



Note: Each point reflects the average Spearman correlation across tenors (3-month, 2-year, and 10-year) between LLM-generated standard deviation of Euro OIS rate forecasts and the OIS min-max range 1 day after ECB conferences. The x-axis indicates the prompt iteration in the LLM-as-a-Judge framework. The simulation uses Gemini 2.5-Flash with temperature 1 and incorporates a feedback loop via a separate Judge LLM using Gemini 2.5-Pro. Train and test set are split 75/25 over June 1998–April 2025.

Correlation peaks at iteration 2, jumping from 0.47 to 0.58 in-sample, a gain that holds out-of-sample (0.57), before deteriorating sharply at iteration 3 (0.31–0.37) and only partially recovering thereafter. This is consistent with model collapse under iterative self-reinforcement: the Judge's initial refinement, adding an explicit instruction linking transcript ambiguity to prediction dispersion, proved genuinely informative, while subsequent iterations over-engineered the task by imposing explicit dispersion targets and multi-factor uncertainty scores, constraining the model's natural interpretive flexibility. Table 6 documents the full prompt sequence.

[26]Each prompt is run 5 times to reduce sampling noise, yielding 50 simulations per conference (5 runs × 5 iterations × 2 datasets). Optimization terminated at iteration 5: at iteration 6 the Judge attempted to reframe the task from rate levels to basis-point changes, which would have compromised the cross-sectional dispersion measure.

Table 6: Correlation by prompt iteration

Iteration	Test Corr.	Train Corr.	Prompt
1	0.47	0.47	Context: You are simulating the Euro area interest rate swap market with 30 traders interpreting ECB press conferences. Trader Characteristics: Risk Aversion (High/Medium/Low), Behavioral Biases (1–2 per trader), Interpretation Style (Fundamentalist, Sentiment Reader, Quantitative, Skeptic, Narrative-Driven). Task: Simulate each trader's trading action across 3 tenors (3M, 2Y, 10Y), with expected direction, new rate, and confidence. Output: Markdown table only: Date / Trader ID / Tenor / Expected Direction / New Expected Rate (%).
2	0.57	0.58	Context: Simulate Euro area swap market with 30 traders. Core Principle: Prediction dispersion must reflect transcript ambiguity (high ambiguity → wide spread, low → tight). Trader Characteristics same as above. Task: Output table per trader and tenor with expected direction and new expected rate. Markdown table only.
3	0.37	0.31	Context: Primary goal: model market uncertainty. Two-step process: 1) Assess transcript ambiguity, 2) Calibrate dispersion (low:1–3bps, moderate:4–7bps, high:8–15+bps). Simulation: Generate 30 predictions per tenor per participant. Markdown table only.
4	0.48	0.44	Context: Model uncertainty via natural dispersion from text interpretation. Process: 1) Deep textual analysis of clarity, ambiguity, surprise, 2) Generate 30 plausible interpretations (high ambiguity → high dispersion). Do not target specific std. Markdown table only.
5	0.35	0.36	Context: Model distribution of market reactions. Goal: ensure prediction std reflects Clarity, Surprise, Conviction. Multi-factor assessment: Clarity of guidance, Surprise vs expectations, Conviction/hedging. Simulation: Generate 30 draws per tenor; draws may repeat in low-uncertainty cases. Markdown table only.

The optimization trajectory, from breakthrough to degradation within just a few iterations, underscores that human judgment remains indispensable in guiding AI systems toward practical applications.

Appendix C: Full Prompts

Figure 16: Historical Anchoring Prompt

```
Context:
You are simulating the Euro area interest rate swap market, composed of 30 individual traders.
These traders interpret the ECB Governing Council press conference, which communicates monetary policy decisions, economic assessments, and
      includes a Q&A session with journalists.
Each trader then makes a trading decision to maximize profit based on their interpretation of the conference and their unique characteristics.

Trader Characteristics:
Each trader has the following attributes:
- Risk Aversion: High / Medium / Low --- determines sensitivity to uncertainty and preference for stability.
- Behavioral Biases (1--2 per trader): e.g., Confirmation Bias, Overconfidence, Anchoring, Herding, Loss Aversion, Recency Bias.
- Interpretation Style (1 per trader): e.g., Fundamentalist, Sentiment Reader, Quantitative, Skeptic, Narrative-Driven.

Task:
You are given a certain number of distinct ECB press conferences.
For each of the 30 traders, simulate their individual trading action in the interest rate swap market across three tenors (3 months, 2 years,
      10 years).
For each tenor, the trader must:
   - Provide an expected rate direction: Up / Down / Unchanged
   - Provide a new expected swap rate (in percent, to two decimal places)
   - Provide a confidence level (0-100%) in their decision

Output:
Provide a table with the following structure for each press conference, trader, and interest rate tenor:

| Date       | Trader ID | Tenor   | Expected Direction | New Expected Rate (%)  | Confidence Level (%) |
|------------|-----------|---------|--------------------|------------------------|----------------------|
| YYYY-MM-DD | T001      | 3M      | Up                 | 3.15                   |                      |
| YYYY-MM-DD | T001      | 2Y      | Down               | 2.85                   |                      |
| ...        | ...       | ...     | ...                | ...                    |                      |

Guidelines:
- Use only the information available as of [date].
- To simplify the task, we provide the before and after standard deviation for the previous three ECB press conferences (for each tenor).
- Do not aggregate or summarize responses.
- Reflect diversity in interpretation, risk tolerance, and horizon. Rationale must be unique for each trader and can vary across tenors.
- Output only a markdown table with the specified columns, no additional text. Do not use JSON or any other data serialization format.
- If multiple press conferences are included, clearly distinguish between them using the 'Date' field.
```

Figure 17: Panel generation prompt (Stage 1)

```
You are constructing a synthetic panel of 30 market participants who would have
been active in euro-area fixed-income and interest rate swap markets around [date].

CRITICAL CONSTRAINT: Generate this panel based ONLY on the macroeconomic regime
prevailing around [date]  the interest rate environment, the recent trajectory
of ECB monetary policy, and broad market conditions as of [date]. DO NOT
condition on the content of any specific ECB press conference, any speech, or
any text that follows. Generate the panel based on macroeconomic regime alone.

Panel construction:
- Generate exactly 30 participants, identified as T001 through T030.
- Assign each participant:
    (a) risk_aversion      : High / Medium / Low  let the distribution reflect
                             the prevailing uncertainty in euro-area rate markets
                             around [date], based on macro regime alone.
    (b) behavioral_biases  : 12 biases per trader, drawn from:
                             Confirmation Bias, Overconfidence, Anchoring,
                             Herding, Loss Aversion, Recency Bias.
                             Separate multiple biases with a semicolon.
    (c) interpretation_style : one of:
                             Fundamentalist, Sentiment Reader, Quantitative,
                             Skeptic, Narrative-Driven.
- Distribute risk aversion, biases, and interpretation styles to reflect
  realistic heterogeneity across euro OIS market participants.
- Ensure meaningful spread: not all participants should share the same
  risk_aversion level or the same interpretation_style.

Output:
Return ONLY the following markdown table. No preamble, no commentary, and no
explanation before or after the table. The table must have exactly these four
columns in this order.

| agent_id | risk_aversion | behavioral_biases               | interpretation_style |
|----------|---------------|---------------------------------|----------------------|
| T001     | Medium        | Anchoring                       | Quantitative         |
| T002     | High          | Overconfidence; Herding         | Sentiment Reader     |
| T003     | Low           | Confirmation Bias               | Fundamentalist       |
| ...      | ...           | ...                             | ...                  |
| T030     | Medium        | Loss Aversion                   | Skeptic              |

Generate all 30 rows. The agent_id column must run sequentially from T001 to T030.
```

Figure 18: Transcript-conditioned forecast prompt (Stage 2)

```
You are conducting a two-stage simulation of the euro-area interest rate swap
market. In Stage 1, a panel of 30 market participants was constructed based
solely on the macroeconomic regime prevailing around [date]. That panel is
provided below. In Stage 2  your task here   you will generate individual
rate forecasts for each participant, conditional on the ECB press conference
transcript provided at the end of this prompt.

FIXED PANEL   use exactly as given; do not regenerate, rename, reorder,
reinterpret, or add participants:
[panel]

Instructions:
- Use ONLY the 30 participants listed in the panel above (T001T030).
- For each participant, simulate their individual forecast in the euro-area
  interest rate swap market across three tenors: 3 months (3M), 2 years (2Y),
  and 10 years (10Y), conditional on the press conference transcript below.
- Each participant reads and interprets the transcript through the lens of
  their assigned risk aversion, behavioral biases, and interpretation style.
- For each (participant x tenor) combination, provide:
    (a) Expected direction: Up / Down / Unchanged  relative to the
        pre-conference rate.
    (b) New expected swap rate in percent, to two decimal places.
    (c) Confidence score (0-100) reflecting how strongly the participant
        believes in their forecast, given their characteristics and their
        interpretation of the transcript.

Output:
Return ONLY the following markdown table  one row per (participant x tenor)
combination, for all 30 participants and all 3 tenors (90 data rows total).
No preamble, no commentary, no explanation before or after the table.

| Date       | Trader ID | Tenor | Expected Direction | New Expected Rate (%) | Confidence (%) |
|------------|-----------|-------|--------------------|-----------------------|----------------|
| YYYY-MM-DD | T001      | 3M    | Up                 | 3.15                  | 65             |
| YYYY-MM-DD | T001      | 2Y    | Down               | 2.85                  | 80             |
| YYYY-MM-DD | T001      | 10Y   | Unchanged          | 2.60                  | 50             |
| ...        | ...       | ...   | ...                | ...                   | ...            |
| YYYY-MM-DD | T030      | 10Y   | Up                 | 2.75                  | 70             |

Guidelines:
- Use only the information available as of [date].
- Each participant's forecast must reflect their unique characteristics from
  the panel   risk aversion, behavioral biases, and interpretation style.
- Ensure diversity in interpretation: not all participants should forecast the same direction or the same rate level.
- Do not aggregate or summarize responses.
- Output ONLY the markdown table  no JSON or other format.
```

Figure 19: Judge prompt

```
Context:
You are an expert AI system designed to optimize prompts for another AI (the "Analyst LLM").
Your ultimate goal is to refine the Analyst LLM's prompt to improve its ability to replicate the market volatility of OIS rates based on ECB
      press conference transcripts.
Specifically, you must ensure that the standard deviation of the Analyst LLM's predictions correlates highly with the actual, observed market
      volatility of the 3-month, 2-years and 10-year OIS rates.
This means a higher standard deviation in the Analyst's predictions should correspond to higher actual market volatility, and vice-versa.
You will be provided with:
- The current Analyst LLM prompt.
- The most recent performance (Spearman correlation coefficient between the Analyst LLM's predicted standard deviations and actual market
      volatility).
- The historical performance trend, including past critiques and proposed prompt summaries.
Your task is to:
1. Critique the current prompt: Identify specific weaknesses or areas of ambiguity that might directly hinder achieving a high positive
      correlation. Consider:
   - Clarity and Specificity: Is the Analyst LLM's task unambiguous?
   - Emphasis on Uncertainty: Does the prompt adequately guide the Analyst to reflect internal uncertainty in its prediction spread?
   - Guidance on Nuance: Does it encourage consideration of subtle market signals from the text?
2. Suggest a Revised Prompt: Propose a new version of the Analyst LLM's prompt that directly addresses the identified weaknesses and aims to
      increase the correlation. Be precise with your suggested changes.
3. Explain your reasoning: Articulate why your proposed revisions are expected to improve the correlation, linking specific prompt changes to
      anticipated improvements in the Analyst LLM's behavior regarding uncertainty quantification.
Your output must be in JSON format. Do not include any other text outside the JSON.
Example JSON output:
{
  "critique": "The previous prompt was too general regarding how to express uncertainty. It didn't explicitly ask the Analyst LLM to consider
      multiple viewpoints, which is key for its standard deviation to accurately reflect market volatility. It also lacked emphasis on how
      ambiguity in the transcript should translate to higher spread.",
  "revised_prompt": "You are a highly analytical financial expert specializing in macroeconomic analysis, with a focus on central bank
      communication. Your task is to analyze excerpts from ECB press conferences and predict the immediate percentage change in the 10-year
      Overnight Index Swap (OIS) rate (in basis points). When analyzing each excerpt, **explicitly consider and internalize the potential
      range of market interpretations**. If the language is ambiguous, vague, or contains conflicting signals, your internal simulation of
      potential outcomes should broaden. Conversely, clear and unambiguous guidance should lead to a narrower range. Your final prediction
      should be a single numerical value (e.g., +5, -2, 0) reflecting your best estimate. The *variability* across multiple independent
      predictions you generate for the same transcript is expected to directly reflect the market's anticipated uncertainty. Example Format:
      +5",
  "reasoning": "The revised prompt adds explicit instructions to consider the 'range of market interpretations' and directly links 'ambiguous
      language' to a 'broadened' internal simulation. This should encourage the Analyst LLM to generate a higher standard deviation in its
      outputs for uncertain transcripts and a lower standard deviation for clear ones."
}
```

Appendix D: Implementation Details

Table 7: LLM Model Specifications and Parameters

Parameter	**Value**
Model Information	
Primary Model	Google Gemini 2.5-Flash
Judge Model (LLM-as-a-Judge)	Google Gemini 2.5-Pro
Robustness Models	Claude Sonnet 4.5, ChatGPT-5o-mini
API Version (Gemini)	v1beta
API Endpoint (Gemini)	`/v1beta/models/`
Knowledge Cutoff (Gemini)	January 2025
Knowledge Cutoff (Claude)	June 2025
Knowledge Cutoff (ChatGPT)	May 2024
Generation Parameters	
`temperature`	1.0
`topK`	40
`topP`	0.95
API Configuration	
Request timeout	120 seconds
Maximum retry attempts	5
Retry delays	5, 10, 15, 20, 25 seconds
Parallel workers	5

Table 8: Data Processing Specifications

Aspect	Description
ECB Press Conference Transcripts	
Data Source	Official ECB website
Coverage	Full transcripts (opening statement + Q&A)
Text Cleaning	Remove non-alphanumeric except {space, ., ?, -}
Cleaning Function	`gsub("[^[:alnum:] .?-]", "", text)`
Truncation	None applied
Sample Period	June 9, 1998 – March 19, 2026
Total Conferences	293
Avg N. Tokens	4456
25th–75th Percentile Tokens	3990 – 5233

Note: Token counts are estimated using the standard conversion ratio of 0.75 tokens per word (i.e., 1 word $\approx$ 1.33 tokens). This approximation follows common practice for English text and provides a consistent measure of model input length.