Interpreting Fedspeak with Confidence: A LLM-Based Uncertainty-Aware Framework Guided by Monetary Policy Transmission Paths

Rui Yao[*1], **Qi Chai**[*1], **Jinhai Yao**[*2], **Siyuan Li**[1], **Junhao Chen**[1], **Qi Zhang**[†2], **Hao Wang**[†1]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Antai College of Economics and Management, Shanghai Jiaotong University
{ryao663, qchai315, sli974, jchen024}@connect.hkust-gz.edu.cn, {jh_Yao,zhang.qi}@sjtu.edu.cn, haowang@hkust-gz.edu.cn

Abstract

"Fedspeak", the stylized and often nuanced language used by the U.S. Federal Reserve, encodes implicit policy signals and strategic stances. The Federal Open Market Committee strategically employs Fedspeak as a communication tool to shape market expectations and influence both domestic and global economic conditions. As such, automatically parsing and interpreting Fedspeak presents a high-impact challenge, with significant implications for financial forecasting, algorithmic trading, and data-driven policy analysis. In this paper, we propose an LLM-based, uncertainty-aware framework for deciphering Fedspeak and classifying its underlying monetary policy stance. Technically, to enrich the semantic and contextual representation of Fedspeak texts, we incorporate domain-specific reasoning grounded in the monetary policy transmission mechanism. We further introduce a dynamic uncertainty decoding module to assess the confidence of model predictions, thereby enhancing both classification accuracy and model reliability. Experimental results demonstrate that our framework achieves state-of-the-art performance on the policy stance analysis task. Moreover, statistical analysis reveals a significant positive correlation between perceptual uncertainty and model error rates, validating the effectiveness of perceptual uncertainty as a diagnostic signal.

Code —
https://github.com/yuuki20001/FOMC-sentiment-path

Introduction

The Federal Open Market Committee (FOMC) is the key institution responsible for managing the U.S. economy and implementing monetary policy (Connolly and Struby 2024). Through tools such as interest rate adjustments and open market operations, monetary policy regulates the money supply to influence economic activity (Pflueger and Rinaldi 2022; Volk 2024). FOMC's statutory mandate is to promote "maximum employment, stable prices, and moderate long-term interest rates". This reflects a broader goal of maintaining a sustainable balance between economic growth, labor market performance, and inflation control. Given the global importance of the Federal Reserve, its communications have a direct and substantial influence on financial markets worldwide (Karnaukh and Vokata 2022; Suh 2025).

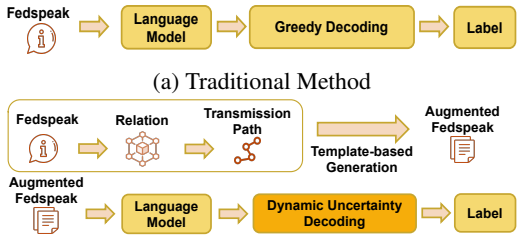
Federal Reserve communications, known as "Fedspeak", are challenging to interpret. This challenge arises from their inherent ambiguity, where the same term may imply different policy stances depending on the broader economic context. For example, a "strong" labor market might be a dovish signal (indicating no immediate rate hikes) in a weak economy, but it can become a hawkish signal (suggesting tightening) in an overheated economy. This strong dependence on context poses a major challenge for traditional sentiment analysis models.

Traditional methods in financial sentiment analysis often face a trade-off between performance and interpretability. For instance, dictionary-based approaches are simple and highly interpretable, but their performance is often limited as they struggle to understand complex contexts or incorporate external knowledge. In contrast, language models like FinBERT (Araci 2019; Liu et al. 2021), after being fine-tuned with domain-specific supervised data (Shah, Paturi, and Chava 2023), can capture contextual nuances effectively and achieve superior performance. However, the black-box nature of such model leads to a lack of transparency in their decision-making process, resulting in poor interpretability.

LLMs offer an accessible and effective solution for financial sentiment analysis, particularly for deciphering central bank policy stances. Recent studies (Peskoff et al. 2023; Hansen and Kazinnik 2024) have shown that closed-source LLMs, such as the GPT-4 series, exhibit strong zero-shot capabilities on policy stance analysis tasks. Research from central banks (Gambacorta et al. 2024; Geiger et al. 2025) demonstrates that fine-tuned LLMs can match human-level accuracy in interpreting monetary policy stance signals. However, existing work primarily focuses on performance metrics, often neglecting critical aspects of LLM behavior such as reliability, bias, and hallucinations. Improving the reliability and interpretability of LLMs, as well as understanding their limitations in policy stance analysis, is essential for reducing systemic risks and enhancing financial stability (Leitner et al. 2024).

In economics, uncertainty is often categorized as risk (known probabilities) or ambiguity (unknown probabilities).

[*]These authors contributed equally.
[†]Corresponding author.



(a) Traditional Method

(b) Our Method

Figure 1: **Comparison between (a) the traditional method and (b) our proposed method**. Our method introduces a domain-specific reasoning approach grounded in the monetary policy transmission mechanism to emulate how human experts analyze policy stances, providing the model with relevant domain knowledge. In addition, we employ a dynamic uncertainty decoding module to capture perceptual uncertainty and estimate the model's confidence in its predictions, thereby improving overall reliability.

Recent work in behavioral economics borrows the computer science concepts of epistemic and aleatoric uncertainty to describe how investors interpret uncertainty (Walters et al. 2023). Epistemic uncertainty reflects missing knowledge, while aleatoric uncertainty stems from randomness. This framing aligns closely with how LLMs operate: like investors, they make predictions under incomplete information. We adopt this perspective to quantify model uncertainty and use it to identify the confidence of predictions. We build on this analogy and treat LLMs as policy stance analysts.

Therefore, to align the model behavior with that of human analysts, we augment the input texts by extracting financial entity relations and reasoning over monetary policy transmission paths using structured templates. This augmentation is designed both to emulate expert-level economic reasoning and to reduce the model's cognitive risk by compensating for missing domain-specific knowledge. Then, we redefine the model prediction uncertainty as perceptual uncertainty (PU), decomposed into cognitive risk (CR) and environmental ambiguity (EA). Building on this formulation, we introduce a dynamic uncertainty decoding module that adapts model behavior based on the PU level. We tune PU–related hyperparameters on the validation set and use them to evaluate the model prediction confidence.

The contribution of this work is as follows:

- We incorporate a domain-specific reasoning grounded in the monetary policy transmission mechanism to emulate the analytical process of human experts and enhance economic interpretability on the policy stance analysis task.
- We introduce a dynamic uncertainty decoding module with a PU metric that helps identify potentially unreliable predictions, aiming to improve overall prediction reliability.
- Our framework achieves state-of-the-art performance while enhancing transparency, economic interpretability, and human-AI collaboration in policy stance analysis.

Related Work

AI in Finance

Recent applications of LLMs in the financial domain have led to the emergence of domain-specific models. proprietary models like BloombergGPT (Wu et al. 2023) and open-source models such as FinGPT (Liu et al. 2023) are often instruction-tuned to perform financial tasks such as sentiment analysis. Existing financial evaluation benchmarks, such as FinQA (Chen et al. 2021), FinBen (Xie et al. 2024), FinTextQA (Chen et al. 2024a), and FinDER (Choi et al. 2025), are designed to assess the performance of LLMs on a diverse range of tasks. These tasks cover numerical reasoning, retrieval-augmented generation evaluation, general understanding, and financial regulation and compliance. These benchmarks establish a standard for measuring the reliability of LLMs in various financial tasks. Another line of research enhances LLM reasoning by decomposing complex and ambiguous tasks into structured, domain-informed components, as seen in FinEntity, EFSA, and DEFINE (Tang et al. 2023; Chen et al. 2024b; Hu et al. 2024). These works highlight the value of integrating domain priors with structured task design to improve accuracy and reliability.

FOMC Analysis

Fedspeak plays a pivotal role in shaping market expectations and asset prices, offering key insights into the Federal Reserve's economic outlook and policy intentions (Ehrmann and Talmi 2020; Cieslak and Vissing-Jorgensen 2021; Mathur et al. 2022; Gómez-Cram and Grotteria 2022). Its nuanced language requires careful interpretation, as tone and sentiment in communications can significantly affect financial variables and market volatility (Gorodnichenko, Pham, and Talavera 2023; Curti and Kazinnik 2023). Approaches to analyzing Fedspeak have evolved from early dictionary-based methods (Loughran and McDonald 2011) and topic modeling techniques (Boukus and Rosenberg 2006; Edison and Carcel 2021) to Transformer-based models that better capture contextual subtleties (Liu et al. 2023; Shah, Paturi, and Chava 2023). LLM-based methods have become increasingly popular for policy stance analysis (Peskoff et al. 2023; Hansen and Kazinnik 2024). Domain-specific adaptations have shown promise in improving classification accuracy (Gambacorta et al. 2024; Geiger et al. 2025). However, most existing approaches limit LLMs to predicting stance labels, underutilizing their potential.

Methodology

Policy Stance Definition

When setting monetary policy, the Federal Reserve relies on deliberations and votes by FOMC members to reach collective decisions. These decisions reflect an aggregation of individual views and preferences. Committee members hold different policy stances: Members who prefer supporting increased output and employment and exhibit a stronger aversion to unemployment risk are considered dovish. Members who prefer controlling inflation and exhibit stronger aversion to inflation risk are considered hawkish (Bordo and

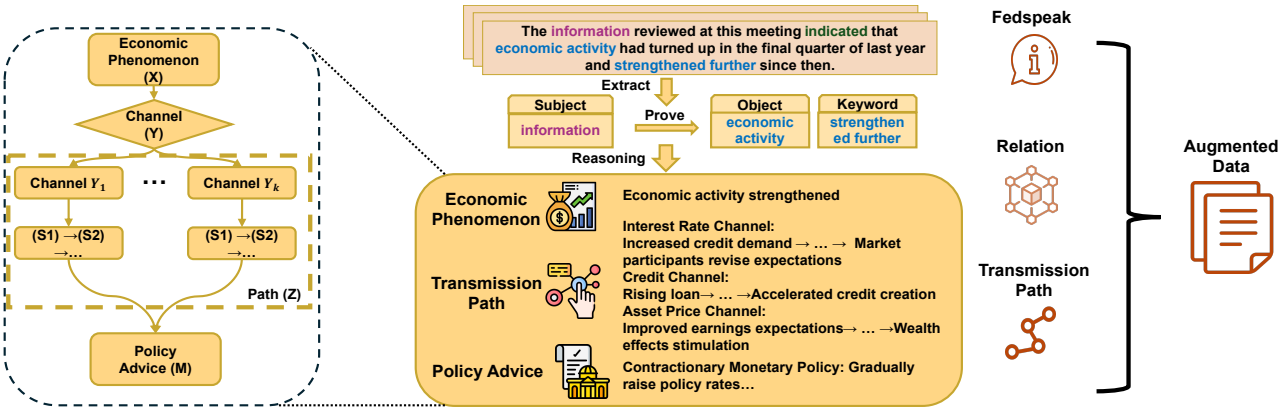


Figure 2: **The workflow of data augmentation**.We extract economic entity relations from Fedspeak, and then perform reasoning grounded in the monetary policy transmission mechanism using structured templates to derive policy advice.

Istrefi 2023). From a trade-off perspective, dovish members place more weight on maximizing employment, whereas hawkish members assign more weight to maintaining price stability (Shah, Paturi, and Chava 2023; Smales 2023). We follow prior literature in assigning stance labels based on these definitions. We define policy stances as dovish, hawkish, and neutral. The detailed definitions and explanations are illustrated in the detailed data augmentation section of the appendix.

Data Augmentation with Domain Reasoning

Financial Entity Relations The relations embedded in Fedspeak are complex and multi-layered. To enhance the reliability of interpreting Fedspeak, we decompose these financial entity relations, thereby improving our model's reasoning capability and accuracy (Li and Sanna Passino 2024). Within the financial entity relations analysis framework, atomic relations serve as the minimal logical units for semantic decomposition. We define the entity set as $\mathcal{E}$ (representing all financial and economic entities) and the atomic relation set as $\mathcal{R} = \{CAUSE, COND, EVID, PURP, ACT, COMP\}$. Atomic relations are formally defined in Equation 1:

$$r(e_i, e_j) \in \mathcal{R} \quad , \quad \forall e_i, e_j \in \mathcal{E} \tag{1}$$

where e_i is the subject entity, e_j is the object entity, and r is one of the six core relation types. These six core relation types are listed in the detailed data augmentation section of the appendix.

Given a path $T : e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} \dots \xrightarrow{r_n} e_{n+1}$, the decomposition of financial entity relations follows Equation 2:

$$T \Rightarrow \bigcup_{k=1}^{n} r_k(e_k, e_{k+1}) \quad , \quad r_k \in \mathcal{R} \tag{2}$$

In Equation 2, if a sentence contains a multi-step logical chain or causal chain (e.g., A causes B, B causes C), it must be decomposed into a sequential list of atomic relations within the same subtask. We construct an entity classification function $C : \mathcal{E} \to E$ that maps to the information sources E in Fedspeak. When extracting entity relations, the information source (e.g., official sources: Fed chairs, official documents, committee members; external sources: journalists, analysts) and its stance (official statements, data interpretation, external analysis, direct questions, rhetorical questions) must be identified.

Monetary Policy Transmission Paths Existing research on monetary policy primarily focuses on understanding how policy tools, such as interest rate adjustments, affect key economic targets like inflation and employment. These studies trace the monetary policy transmission paths through the financial system and the real economy, offering a theoretical foundation for policy design (Passos, Carrasco-Gutierrez, and Loureiro 2024; Bosshardt et al. 2024; Suh 2025; Alessandri, Òscar Jordà, and Venditti 2025).

Given the importance of monetary policy transmission paths, we construct the transmission path: Define a quadruple $\mathbf{\Gamma} = (\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{M})$, where $\mathbf{X} \in R^n$ is an n dimensional shock vector representing economic phenomena, policy implementation, policy framework shifts, or external shocks, serving as the starting point; $\mathbf{Y} = \{y_1, y_2, \dots, y_k\}$ is the set of monetary policy transmission channels, with k being the number of channels, including the credit channel, asset price channel, aggregate demand channel, etc.; $\mathbf{Z} = \{Z_1, Z_2, \dots, Z_p\}$ is the monetary policy transmission path, defined as the sequence of expectation shifts or market/economic indicator responses within a specific transmission channel. For each channel $y_i \in \mathbf{Y}$, there exists a transmission path Z_{y_i}, which consists of a series of state transitions, as shown in Equation 3:

$$Z_{y_i} = f_{\text{i,n}} \circ f_{\text{i,n-1}} \circ \cdots \circ f_{\text{i,j}} \circ \cdots \circ f_{\text{i,1}} \tag{3}$$

where $f_{\text{i,j}}$ denotes the mapping function at step j in channel y_i. For each channel y_i, the state sequence is given by Equation 4:

$$S_{\text{i},j} = \begin{cases} \phi_i(\mathbf{X}) & j = i^{(0)} \\ g_{\text{i},j}^{(j)}(S_{\text{i},j}) & j = i^{(1)}, i^{(2)}, \dots, i^{(n)} \end{cases} \tag{4}$$

In Equation 4, j denotes the number of steps in channel y_i, ϕ_i represents the function that maps the initial shock $\mathbf{X}$ to

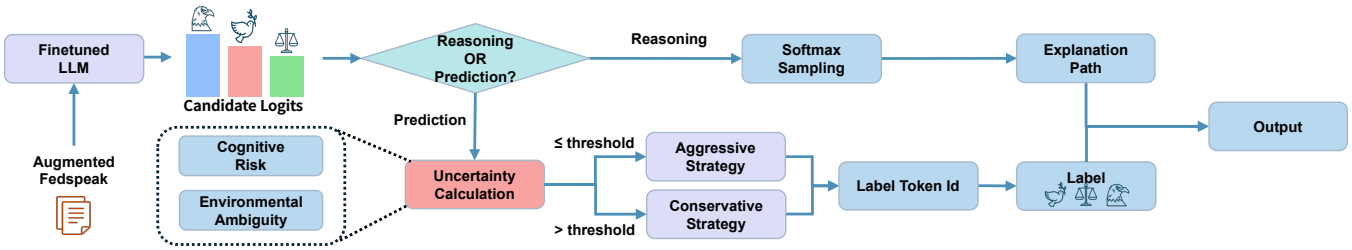


Figure 3: **Overview of Dynamic Uncertainty Decoding module**.When the LLM is about to generate a prediction token, we obtain the corresponding logits over the vocabulary. Dynamic uncertainty decoding module quantifies the model's PU via estimating CR and EA. The decoding strategy for the current token is selected based on whether the PU exceeds the threshold.

the initial state of this channel, and S_{i,n_i} is the final state of the channel. $\mathbf{M} \in \mathbf{P}$ is the policy advice, $\mathbf{P}$ denotes the policy space, and the entire transmission path of channel y_i is aggregated into $S^{(i)} = \{S_{\mathrm{i},0}, S_{\mathrm{i},1}, \ldots, S_{\mathrm{i},n_i}\}$. The policy advice $\mathbf{M}$ is a function of the aggregated state, as given by Equation 5:

$$\mathbf{M} = h\left(S^{(1)}, S^{(2)}, \ldots, S^{(i)}, \ldots, S^{(k)}\right) \tag{5}$$

where the final policy advice is generated by aggregating channel outputs and influenced by multiple channels and transmission paths. Through the above analysis, we adopt a structured framework to analyze the monetary policy transmission mechanism, with Equation 6 as the main path.

$$\mathbf{X} \rightarrow \mathbf{Z} \rightarrow \mathbf{M} \tag{6}$$

The detailed transmission path of monetary policy is illustrated in Figure 2.

Dynamic Uncertainty Decoding

LLMs may generate unreliable information. During next-token prediction, reliability can be affected by ambiguity in the input data and insufficient domain knowledge. We treat our model as a policy analyst. Policy analysts face reliability issues in decision-making due to EA and CR. EA stems from data-related uncertainty, such as distributional shifts, semantic ambiguity, and other stochastic variations in the input (Walters et al. 2023; le Roux and Bopp 2025). EA causes policy analysts to consider multiple alternative options during decision-making. CR arises from limited domain knowledge, missing information, or skill gaps (Walters et al. 2023). CR causes policy analysts to make unreliable judgments due to cognitive limitations. Higher levels of EA or CR increase the policy analyst's PU. We incorporate a measure of PU into our model framework. Let the LLM be denoted as $\mathcal{M}$, and let the prompt be tokenized into vector representations $\mathbf{q}$. We use the top-k logits from $\mathcal{M}$ to construct a Dirichlet distribution (Sensoy, Kaplan, and Kandemir 2018; Ma et al. 2025), as shown below:

$$\alpha_k = \mathcal{M}(\tau_k|\mathbf{q}, \mathbf{a_{t-1}}) \quad , \quad \alpha_0 = \sum_{k=1}^{K} \alpha_k \tag{7}$$

where τ_k presents the token associated with the k^{th} largest logit, and α_0 constitutes the total evidence parameter of the Dirichlet distribution, equaling the sum of the k largest logits. The model predicts the next token a_t based on the representations $\mathbf{q}$ and the previously generated tokens $\mathbf{a_{t-1}} = a_1 a_2 \ldots a_{t-1}$.

Our measure of EA is defined as follows:

$$EA(a_t) = -\sum_{k=1}^{K} \frac{\alpha_k}{\alpha_0}(\psi(\alpha_k + 1) - \psi(\alpha_0 + 1)) \tag{8}$$

where ψ represents the digamma function, defined as $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$. EA denotes the expected entropy of the predictive distribution. Higher entropy indicates a more uniform data distribution and greater ambiguity. Increased EA implies higher inherent uncertainty in the data, making the model uncertain about the input content. In such cases, a high level of EA indicates that the model is uncertain and confused about the output.

Our measure of CR is defined as follows:

$$CR(a_t) = \frac{K}{\sum_{k=1}^{K}(\alpha_k + 1)} \tag{9}$$

CR is inversely related to the total evidence $\sum_{k=1}^{K}(\alpha_k + 1)$. A higher CR indicates that the model has less accumulated evidence and domain knowledge, suggesting limited capability to process the input effectively. Under high CR, the model tends to make less reliable decisions.

Our measure of PU is defined as follows:

$$PU = EA \times CR \tag{10}$$

where PU is jointly determined by EA and CR. When both EA and CR values are high, the model struggles to identify clear information in the input and lacks sufficient domain knowledge for accurate judgment. This results in elevated PU and less reliable outputs. To enhance the model's understanding of policy analysis and improve its reliability, we integrate uncertainty quantification into our framework with the goal of reducing PU.

Experiment

Experiment Setup

To simulate the reasoning process of human analysts, we design a set of structured templates grounded in the mech-

Model	Meeting Minutes		Press Conference		Speeches		All Categories	
	Macro F1	Weighted F1	Macro F1	Weighted F1	Macro F1	Weighted F1	Macro F1	Weighted F1
Zero-Shot								
GLM-4-9B	0.2855	0.3058	0.3438	0.3876	0.3065	0.3807	0.3152	0.3478
HD-Dissent	0.3032	0.4219	0.3359	0.5962	0.2468	0.3288	0.3109	0.4774
Qwen3-8B	0.5144	0.5325	0.5670	0.5843	0.5748	0.6450	0.5572	0.5854
Qwen3-14B	0.5083	0.5367	0.4236	0.4650	0.5129	0.5969	0.5135	0.5514
GLM-Z1-9B	0.5998	0.5995	0.5951	0.6143	0.5761	0.6284	0.6023	0.6154
Deepseek-R1	0.6269	0.6355	0.5873	0.6094	0.5680	0.6490	0.6201	0.6385
Gemini-2.5-pro	0.5953	0.5823	0.6703	0.6620	0.6116	0.6550	0.6286	0.6275
Phi-4	0.6366	0.6385	0.6023	0.6105	0.6140	0.6736	0.6349	0.6488
GPT-4.1	0.6500	0.6456	**0.6803**	**0.6900**	0.6326	0.6980	0.6662	0.6763
AICBC	0.6701	0.6741	0.6735	0.6805	0.6175	0.6841	0.6637	0.6802
Fine-Tuned								
GLM-4-9B	0.6442	0.6514	0.5955	0.6087	0.6146	0.6832	0.6390	0.6585
Phi-4	0.6481	0.6560	0.5706	0.5927	0.6609	0.7096	0.6495	0.6683
FinBERT	0.5922	0.5958	0.6117	0.6844	0.6398	0.6513	0.6185	0.6380
GLM-Z1-9B	0.6426	0.6514	0.6366	0.6503	0.6527	0.7032	0.6573	0.6736
Qwen3-8B	0.6504	0.6536	0.5783	0.5985	0.6621	0.7242	0.6586	0.6745
Qwen3-14B	0.6227	0.6288	0.6286	0.6363	0.6052	0.6819	0.6360	0.6534
Ours	**0.7449**	**0.7394**	0.6672	0.6699	**0.7291**	**0.7718**	**0.7327**	**0.7426**

Table 1: Performance comparison of our method against zero-shot and fine-tuned baselines on the FOMC dataset. ***Zero-Shot*** denotes the baseline models; ***Fine-tuned*** applies LoRA fine-tuning on FOMC data to the ***Zero-Shot*** models; **Ours** further integrates the proposed data augmentation and uncertainty decoding module with the ***Fine-tuned*** Qwen-3-14B.

anism of monetary policy transmission. Using these templates, we augment the original Fedspeak texts through a hybrid human-AI procedure to construct a supervised fine-tuning dataset. Our dynamic decoding framework adaptively selects a strategy guided by an PU threshold. For low PU, it employs an aggressive strategy by selecting the label from the top-ranked vocabulary token; for high PU cases, it adopts a conservative strategy, sampling from the two tokens with the highest logits. To compute the PU, we first apply a ReLU activation to the token-level logits to derive their evidence. This evidence is then mapped to three canonical labels (HAWKISH, DOVISH, and NEUTRAL) based on a predefined token-to-label mapping. We construct a candidate evidence set by combining the aggregated label logits scores with the individual logits scores of all unmapped tokens. The top-K logits scores from this set are then selected, and the PU is computed over them to capture ambiguity and low evidence.

We utilized the open-source ModelScope Swift (Zhao et al. 2025) framework for all model training and inference. All base models were fine-tuned using the LoRA method. During inference, we applied greedy decoding and incorporated FlashAttention for efficiency. For dynamic uncertainty decoding, hyperparameters were selected based on validation set performance. The search space was defined as follows: **Top-K:** $\{3, 10, 15, 20, 25, 30\}$; **Threshold percentiles:** $\{1, 0.95, 0.9, 0.85, 0.8, 0.75, 0.7\}$; **Sampling temperature:** $\{0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.5, 2.0\}$. All stochastic components were controlled with a fixed random seed 42 to ensure reproducibility. All experiments were conducted on four NVIDIA A800 80GB GPUs. Full implementation details are provided in the appendix.

Benchmark and Evaluation Metrics

We train and evaluate our method on the trillion dollar Words (Shah, Paturi, and Chava 2023) FOMC dataset, which contains three distinct categories of Federal Reserve communications: meeting minutes, press conference transcripts, and speeches. The dataset spans from January 1996 to October 2022, covering multiple economic cycles, including the dot-com bubble, the 2008 financial crisis, and the COVID-19 pandemic period. We report two metrics for evaluations and comparisons: Macro-F1 and Weighted-F1. Weighted-F1 is the original metric established for the FOMC dataset, ensuring consistency with previous research and enabling fair comparison with existing baselines.

Main Results

Performance Analysis We compare our method with several widely used language models like GPT (Achiam et al. 2023), Phi (Abdin et al. 2024), GLM (Zeng et al. 2024), Qwen (Yang et al. 2025), Deepseek (Guo et al. 2025), and Gemini (Comanici et al. 2025). The comparison also includes previous approaches like HD-Dissent (Peskoff et al. 2023), AICBC (Fanta and Horvath 2024), and FinBERT (Liu et al. 2021). For open-sourced language models, we report both zero-shot and fine-tuned results respectively. The results presented in Table 1 demonstrate the effectiveness of our proposed approach across all three categories of Federal Reserve communications. Our proposed method achieves substantial improvements over all baselines, demonstrating the effectiveness of our approach for financial sentiment analysis. Specifically, our method achieves 0.7327 Macro-F1 and 0.7426 Weighted-F1 on the

combined dataset, representing significant improvements of 6.6% and 6.2% respectively over the strongest baseline (0.6662 and 0.6802). These results highlight the effectiveness of our methods.

The performance gains are particularly pronounced for meeting minutes and speeches. On meeting minutes, our method achieves 0.7449 Macro-F1 and 0.7394 Weighted-F1, substantially outperforming the best baseline by 7.4% and 6.5% respectively. For speeches, we also observe improvements, with our method reaching 0.7291 Macro-F1 and 0.7718 Weighted-F1, representing gains of 6.7% and 4.7% over the best baseline. However, in the press conference transcripts, our methods show reductions of 1.3% and 2% compared to GPT-4.1. The reason may be that press conferences involve real-time interactions where the context from prior questions and answers within the same session is critical for accurate sentiment classification. Our current approach may not fully capture these dynamic dependencies, which larger models like GPT-4.1 handle more effectively due to their enhanced contextual understanding.

Fine-tuning vs. Zero-shot An important observation from our results is the varying effectiveness of fine-tuning across different model architectures. While fine-tuning generally improves performance for most models, the magnitude of improvement varies significantly. For instance, GLM-4-9B shows substantial gains from fine-tuning (from 0.3152 to 0.6390 Macro-F1 on combined data), while larger models like Phi-4 show more modest improvements. This suggests that fine-tuning is particularly beneficial for models that may lack sufficient pre-training on financial domain data. Notably, zero-shot models like GPT-4.1 achieve competitive performance without fine-tuning, indicating strong out-of-the-box capabilities for financial text understanding. However, our approach still outperforms these strong baselines.

Ablation Analysis

To better understand the contributions of various components in our proposed approach, we performed ablation studies by removing individual components step by step and evaluating their impact on performance in all categories. The results of ablation experiments are presented in Table 2.

Model	All Categories	
	Macro F1	Weighted F1
Ours	0.7327	0.7426
w/o PU	0.7291	0.7378
w/o Transmission Path	0.6538	0.6699
w/o Entity Relationships	0.6397	0.6551
Original Qwen3	0.6360	0.6534

Table 2: Ablation study results of our methods.

Among the results, **w/o PU** refers to using greedy decoding for all samples, **w/o Transmission Path** indicates the removal of monetary policy transmission path information, **w/o Entity Relationships** means the model receives only the policy stance label guidelines, **original Qwen3** represents the baseline with all proposed modules removed. The transmission path shows the most substantial impact on model performance. Its removal results in the largest performance degradation. This significant drop demonstrates that explicit transmission paths are essential for complex financial sentiment analysis.

The transmission path likely enables the model to analyze the sentiment by breaking down the decision process into interpretable steps, leading to more accurate and consistent predictions. The entity relationships also demonstrate substantial contribution to the model. It not only serves as the foundation for the transmission path but also brings performance improvement to the model. The uncertainty component shows the most modest improvement, however, the uncertainty quantification still provides valuable contributions by helping the model better calibrate its predictions and handle ambiguous cases more effectively.

Uncertainty Analysis

To validate the effectiveness of our uncertainty quantification algorithm, we conducted an analysis to examine the relationship between PU and prediction accuracy. We hypothesize that correctly predicted samples should exhibit lower PU compared to incorrectly predicted ones, which would demonstrate the reliability of our uncertainty estimation. We evaluated our uncertainty quantification approach in various K values on the FOMC dataset. For each prediction, we calculated the PU based on the model outputs, then classified predictions into correct and incorrect groups, and computed the mean PU for each group.

To rigorously test our hypothesis, we employed three complementary statistical methods:

- **T-test**: To compare mean PU between correct and incorrect predictions
- **Mann-Whitney U test**: A non-parametric alternative to validate results without assuming normal distribution
- **Logistic regression**: To quantify the predictive power of PU for classification accuracy

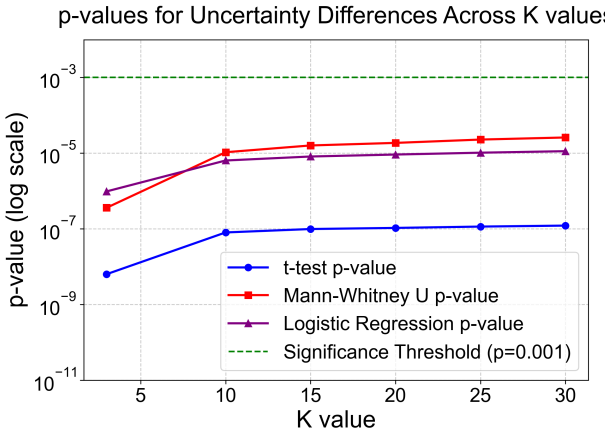


Figure 4: P-values of T-test, Mann-Whitney U test and logistic regression for different K values on FOMC dataset.

Test results are shown in Figure 4. Both statistical tests yield consistently significant results across all k-values. The blue line shows p-values of T-test ranging from 6×10^{-9} to 1.2×10^{-7}, all substantially below the significance threshold (green dashed line at p=0.001). The red line demonstrates p-values of the Mann-Whitney U test from 4×10^{-7} to 3×10^{-5}, providing non-parametric confirmation of significant differences and ensuring robustness regardless of data distribution assumptions. The purple line represents p-values of logistic regression from 1×10^{-6} to 1.2×10^{-5}, indicating that PU serves as a significant predictor of correctness.

The above results confirm a significant positive correlation between model PU and model error rate. This finding provides crucial guidance for practical applications: when the model's PU is high, the reliability of its predictions correspondingly decreases. In real-world financial sentiment analysis scenarios, the most prudent strategy is to actively seek human expert intervention or choose to abstain from providing an answer when the model detects that it cannot make accurate judgments (i.e., when the PU exceeds threshold). This strategy can effectively prevent potential losses caused by incorrect predictions, particularly in high-risk financial decision-making environments.

However, to ensure fair comparison with existing baseline methods, we did not provide models with the option to "refuse to answer" in our main experiments. Instead, all models were required to provide explicit predictions for every sample. To further validate the effectiveness of uncertainty quantification in practice, we designed additional experiments to demonstrate the accuracy differences when model predictions have PU above or below thresholds. We list the accuracy of the low PU group and the high PU group, respectively, in Table 3.

PU	All Categories	
	Macro F1	**Weighted F1**
Low	0.7791	0.7822
High	0.2473	0.4372

Table 3: Performance of predictions with low vs. high PU on the FOMC dataset.

These results indicate that predictions with low PU are highly reliable, achieving strong performance across both Macro F1 and Weighted F1 metrics. In contrast, the high PU group exhibits significantly poorer performance, indicating a higher likelihood of errors. This large performance gap highlights that high PU predictions negatively impact overall model accuracy, effectively undermining the robust results of the low PU group. These findings further reinforce the positive correlation between PU and prediction error rate, confirming PU as a reliable signal of prediction reliability.

Case Study

We conduct a case study to analyze model behavior under different sentence types. Results are shown in Table 4. The model performs well when processing explicit signals, especially those with contrastive shifts (e.g., but, however), where semantic redirection is clear. It also correctly captures transitional tones when contextual cues are clear. However, the model struggles in two cases. First, under contextual confusion, even when surface keywords are present, the sentence may reflect third-party views or rely on external background knowledge. Without such knowledge, the model misinterprets the sentence and makes overly aggressive decisions. Second, under implicit statements, where no clear keywords or contextual anchors exist, the model fails to resolve the underlying intent through economic reasoning. In this case, it adopts a conservative strategy but still produces incorrect predictions. Table 4 categorizes these cases by sentence type, prediction correctness, and decision strategy. The original sentences and detailed analysis for case study are provided in the detailed case study section of the appendix.

Sentence Type	Prediction	Strategy
Explicit + Contrastive	Correct	**Aggressive**
Contextual Confusion	Incorrect	**Aggressive**
Implicit Statement	Incorrect	**Conservative**

Table 4: Case study of different sentence types.

Conclusion

We present a domain-specific reasoning framework grounded in the monetary policy transmission mechanism to emulate expert analysis of policy stances. We propose a dynamic uncertainty decoding module that effectively identifies cases where the model lacks sufficient knowledge or exhibits contextual confusion, rendering it unable to provide viable predictions. Under high perceptual uncertainty, our approach enables the model to identify unreliable outputs, improving reliability under realistic decision-making conditions. We validate our approach through a comprehensive hyperparameter search and demonstrate that the framework achieves state-of-the-art performance on the policy stance analysis task. Beyond improved robustness, our framework also supports human analysts by highlighting predictions that are well-grounded versus those requiring caution. Our work offers a new direction for evaluating the reliability, transparency, and economic interpretability of LLM predictions in policy stance analysis.

Limitation

While incorporating domain knowledge and dynamic decoding enhances performance and reliability, our approach still relies on hand-crafted templates and has runtime and strategy constraints. The main limitation is the limited availability of fine-grained data from central banks such as the BoE and ECB. Additionally, the framework needs improvement to better handle context ambiguity and implicit statements. We leave these challenges for future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62406267) and the Guangzhou Municipal Science and Technology Project (No. 2025A04J4070). Zhang acknowledges financial support from the National Science Foundation of China (No. 72373097).

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alessandri, P.; Òscar Jordà; and Venditti, F. 2025. Decomposing the monetary policy multiplier. *Journal of Monetary Economics*, 152: 103783.

Araci, D. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Bordo, M.; and Istrefi, K. 2023. Perceived FOMC: The making of hawks, doves and swingers. *Journal of Monetary Economics*, 136: 125–143.

Bosshardt, J.; Di Maggio, M.; Kakhbod, A.; and Kermani, A. 2024. The credit supply channel of monetary policy tightening and its distributional impacts. *Journal of Financial Economics*, 160: 103914.

Boukus, E.; and Rosenberg, J. V. 2006. The information content of FOMC minutes. *Available at SSRN 922312*.

Chen, J.; Zhou, P.; Hua, Y.; Xin, L.; Chen, K.; Li, Z.; Zhu, B.; and Liang, J. 2024a. FinTextQA: A Dataset for Long-form Financial Question Answering. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6025–6047. Bangkok, Thailand: Association for Computational Linguistics.

Chen, T.; Zhang, Y.; Yu, G.; Zhang, D.; Zeng, L.; He, Q.; and Ao, X. 2024b. EFSA: Towards event-level financial sentiment analysis. *arXiv preprint arXiv:2404.08681*.

Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.-H.; Routledge, B.; et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Choi, C.; Kwon, J.; Ha, J.; Choi, H.; Kim, C.; Lee, Y.; Sohn, J.-y.; and Lopez-Lira, A. 2025. FinDER: Financial Dataset for Question Answering and Evaluating Retrieval-Augmented Generation. *arXiv preprint arXiv:2504.15800*.

Cieslak, A.; and Vissing-Jorgensen, A. 2021. The economics of the Fed put. *The Review of Financial Studies*, 34(9): 4045–4089.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Connolly, M. F.; and Struby, E. 2024. Treasury buybacks, the Federal Reserve's portfolio, and changes in local supply. *Journal of Banking & Finance*, 168: 107286.

Curti, F.; and Kazinnik, S. 2023. Let's face it: Quantifying the impact of nonverbal communication in FOMC press conferences. *Journal of Monetary Economics*, 139: 110–126.

Edison, H.; and Carcel, H. 2021. Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. *Applied Economics Letters*, 28(1): 38–42.

Ehrmann, M.; and Talmi, J. 2020. Starting from a blank page? Semantic similarity in central bank communication and market volatility. *Journal of Monetary Economics*, 111: 48–62.

Fanta, N.; and Horvath, R. 2024. Artificial intelligence and central bank communication: the case of the ECB. *Applied Economics Letters*, 0(0): 1–8.

Gambacorta, L.; Kwon, B.; Park, T.; Patelli, P.; and Zhu, S. 2024. *CB-LMs: language models for central banking*. Bank for International Settlements, Monetary and Economic Department Basel ....

Geiger, F.; Kanelis, D.; Lieberknecht, P.; and Sola, D. 2025. Monetary-Intelligent Language Agent (MILA). Technical report, Technical Paper.

Gómez-Cram, R.; and Grotteria, M. 2022. Real-time price discovery via verbal communication: Method and application to Fedspeak. *Journal of Financial Economics*, 143(3): 993–1025.

Gorodnichenko, Y.; Pham, T.; and Talavera, O. 2023. The voice of monetary policy. *American Economic Review*, 113(2): 548–584.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hansen, A. L.; and Kazinnik, S. 2024. Can chatgpt decipher fedspeak? *Available at SSRN 4399406*.

Hu, Y.; Wang, X.; Yao, W.; Lu, Y.; Zhang, D.; Foroosh, H.; Yu, D.; and Liu, F. 2024. Define: Enhancing llm decision-making with factor profiles and analogical reasoning. *arXiv preprint arXiv:2410.01772*.

Karnaukh, N.; and Vokata, P. 2022. Growth forecasts and news about monetary policy. *Journal of Financial Economics*, 146(1): 55–70.

le Roux, S.; and Bopp, F. 2025. Social learning under ambiguity—An experimental study. *Journal of Behavioral and Experimental Economics*, 114: 102323.

Leitner, G.; Singh, J.; van der Kraaij, A.; and Zsámboki, B. 2024. The rise of artificial intelligence: benefits and risks for financial stability. *Financial Stability Review*, 1.

Li, X. V.; and Sanna Passino, F. 2024. FinDKG: Dynamic Knowledge Graphs with Large Language Models for Detecting Global Trends in Financial Markets. In *Proceedings*

of the 5th ACM International Conference on AI in Finance, ICAIF '24, 573–581. New York, NY, USA: Association for Computing Machinery. ISBN 9798400710810.

Liu, X.-Y.; Wang, G.; Yang, H.; and Zha, D. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*.

Liu, Z.; Huang, D.; Huang, K.; Li, Z.; and Zhao, J. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, 4513–4519.

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.

Ma, H.; Chen, J.; Zhou, J. T.; Wang, G.; and Zhang, C. 2025. Estimating LLM Uncertainty with Evidence. *arXiv preprint arXiv:2502.00290*.

Mathur, P.; Neerkaje, A.; Chhibber, M.; Sawhney, R.; Guo, F.; Dernoncourt, F.; Dutta, S.; and Manocha, D. 2022. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2276–2285.

Passos, F. V.; Carrasco-Gutierrez, C. E.; and Loureiro, P. R. A. 2024. Monetary policy through the risk-taking channel: Evidence from an emerging market. *The Quarterly Review of Economics and Finance*, 98: 101923.

Peskoff, D.; Visokay, A.; Schulhoff, S.; Wachspress, B.; Blinder, A.; and Stewart, B. M. 2023. GPT Deciphering Fedspeak: Quantifying Dissent Among Hawks and Doves. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6529–6539.

Pflueger, C.; and Rinaldi, G. 2022. Why does the Fed move markets so much? A model of monetary policy and time-varying risk aversion. *Journal of Financial Economics*, 146(1): 71–89.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shah, A.; Paturi, S.; and Chava, S. 2023. Trillion dollar words: A new financial dataset, task & market analysis. *arXiv preprint arXiv:2305.07972*.

Smales, L. A. 2023. Classification of RBA monetary policy announcements using ChatGPT. *Finance Research Letters*, 58: 104514.

Suh, J. E. 2025. An interest rate rule following the natural rate of interest for optimal monetary policy. *Economic Modelling*, 147: 107040.

Tang, Y.; Yang, Y.; Huang, A.; Tam, A.; and Tang, J. 2023. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15465–15471.

Volk, M. 2024. The transmission of targeted monetary policy to bank credit supply. *The Quarterly Review of Economics and Finance*, 94: 104–112.

Walters, D. J.; Ülkümen, G.; Tannenbaum, D.; Erner, C.; and Fox, C. R. 2023. Investor behavior under epistemic vs. aleatory uncertainty. *Management Science*, 69(5): 2761–2777.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Xie, Q.; Han, W.; Chen, Z.; Xiang, R.; Zhang, X.; He, Y.; Xiao, M.; Li, D.; Dai, Y.; Feng, D.; et al. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37: 95716–95743.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; et al. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 29733–29735.