Evidence-Traced and Rules-Guided LLM Pipeline for Predicting China's Stock-Market Trend in Response to FOMC Press Conferences

Chen Xin

The Hong Kong University of Science and Technology, Hong Kong, China
kennyxin0524@gmail.com

Abstract

This paper studies whether language in Federal Open Market Committee (FOMC) press conferences helps predict the next trading-day trend of China's stock market. We propose an evidence-traced and rules-guided LLM pipeline for event-level prediction without model fine-tuning. In our setting, evidence-traced means that each forecast is grounded in a compact set of retrieved transcript spans, and the output includes span-level citations that make the decision auditable. The pipeline separates each press conference into an opening-statement view and a Q&A view, applies evidence retrieval, and produces a structured label in {up, down, unknown}. We use self-consistency voting and a selective threshold τ to allow abstention when evidence is weak. We also distill a lightweight rulebook from the development period and reuse it as soft guidance at test time. On a time-based split, the rules-guided opening view achieves 71.4% accuracy at 71.8% coverage on the held-out test set (AUC = 0.655), outperforming dictionary sentiment, FinBERT, TF–IDF logistic regression, and zero-shot LLM baselines. Q&A-only results stay near chance, suggesting that the prepared opening statement carries more reliable short-horizon cross-market signal.

Keywords: Central bank communication, FOMC press conference transcripts, China stock market (A-share), LLM pipeline, Selective prediction

1. Introduction

Central banks use communication as a core monetary policy instrument. Policy language can shift expectations even when the policy rate does not change. Prior work shows that communication affects asset prices through channels such as forward guidance and risk premia[1][2]. In the Federal Reserve system, each Federal Open Market Committee (FOMC) meeting is followed by a press conference. The press conference has a prepared opening statement and a journalist-driven Q&A session. These two components play different informational roles. The opening statement is more structured. The Q&A often reveals uncertainty, clarifies the reaction function, and responds to gaps raised by journalists[2]. This structure motivates a two-view analysis.

A large empirical literature shows that markets react to both policy actions and policy language around FOMC events. Studies find that statements and related communications contain information beyond the target rate move and can shift market expectations[3][4]. Researchers also document systematic timing patterns around scheduled announcements, including pre-announcement drifts, which motivates studying event-time communication content rather than only policy outcomes[5]. These effects can transmit internationally. Monetary policy surprises and communication shocks can move foreign equity and bond markets through global linkages and risk sentiment[6]. However, many existing prediction-oriented approaches still rely on coarse text features. Dictionary-based measures compress complex policy language into simple counts and may miss domain-specific meanings that drive interpretation[7]. These approaches also rarely provide span-level evidence for an event-level directional call.

This paper develops an **evidence-constrained** LLM pipeline for predicting the next trading-day direction of China's stock market following an FOMC press conference, without parameter fine-tuning. The pipeline processes the opening statement and the Q&A as separate views. It retrieves a compact set of evidence spans for each view. It then prompts the LLM to output a structured verdict in ***{up, down, unknown}*** together with quoted evidence. The system uses repeated sampling and aggregation to reduce variance and stabilize decisions[8][9]. The system also uses selective prediction with abstention. A threshold τ controls whether the model outputs up/down or returns unknown[10]. This design yields an explicit accuracy–coverage trade-off.

This work makes three contributions. First, it proposes a **zero-finetune, evidence-traced** pipeline that grounds each decision in retrieved spans, so readers can audit what text drove the prediction. Second, it introduces a **two-view evaluation protocol** (opening vs. Q&A) that tests which press-conference component carries more reliable short-horizon cross-market signal. Third, it benchmarks against representative baselines, including dictionary-based signals and domain-pretrained sentiment models such as FinBERT, to clarify what the evidence-traced workflow adds beyond traditional NLP features[7][11].

2. Related Work

2.1 Central Bank Communication and Market Effects

Central banks use communication as a policy instrument, not just a supplement to interest-rate decisions. A broad survey shows that communication shapes expectations and asset prices through multiple channels, including forward guidance and risk premia[1][2]. For the Federal Open Market Committee (FOMC), asset prices respond to both policy actions and policy language. Researchers show that statements and press communications contain information beyond the target rate move and can shift market expectations [3][4]. Markets also react around the timing of FOMC events. Evidence such as the pre-FOMC drift suggests that investors price in information systematically before scheduled announcements, which reinforces the value of studying event-time communication content[5]. These effects are not confined to U.S. markets. Monetary policy surprises and communication shocks transmit internationally and move foreign equity and bond markets through global risk sentiment and financial linkages[6]. Finally, the delivery channel can matter. The Chair's tone and spoken content can shift market outcomes even after controlling for the written statement, which motivates separating the opening statement from the Q&A in empirical analysis[12].

2.2 Computational Approaches to Financial Text

Early computational finance NLP relied on lexicon-based features to quantify tone, sentiment, and uncertainty in a transparent way. These methods are easy to audit, but they often produce coarse document-level signals and they can be fragile when they move across domains or genres. Work on central-bank text has therefore moved toward richer representations, including topic- or embedding-based decompositions that "look inside" the communication rather than treating it as a single sentiment score. For example, text-analysis pipelines have been used to extract systematic information from alternative FOMC statements and related communication artifacts, which supports the idea that careful textual modeling can recover policy-relevant structure beyond headline decisions[4]. In parallel, domain-adapted encoders such as FinBERT show that pretraining on financial language can materially improve sentiment modeling with limited labeled data, but these models still typically output scores or embeddings rather than end-to-end, auditable event decisions[11].

2.3 Prompting, Evidence, and Reliability Controls for LLM-Based Forecasting

Large language models enable a different workflow because they can be prompted to perform structured inference without task-specific fine-tuning. Chain-of-thought style prompting improves reasoning on complex tasks, and self-consistency further stabilizes outputs by sampling multiple reasoning paths and aggregating the final decision[9]. Evidence-traced designs also align well with retrieval-augmented generation, where a system retrieves relevant text and constrains generation to grounded evidence, which is especially useful when you want outputs to be auditable at the span level[13]. Empirically, early finance studies show that LLMs can extract predictive signals from text and can compete with or outperform simpler sentiment baselines in certain forecasting settings, which motivates using LLMs as inference engines rather than as sentiment scorers[14]. For deployment, however, the pipeline still needs decision-risk controls. Selective prediction with a reject option provides an explicit accuracy–coverage trade-off, and calibration tools such as temperature scaling help align confidence with empirical correctness[10][15].

3. Problem Formulation

We study whether the textual content of Federal Open Market Committee (FOMC) press conferences contains predictive information about the next-day direction of China's A-share market. For each FOMC meeting mmm, the official transcript is parsed into an opening statement O_m and a Q&A session Q_m. Let $T_m = (O_m , Q_m)$ denote the full press-conference text.

We focus on the Shanghai Composite Index (SSE Composite). Let $P_{T_0}(m)$ be the closing level of the SSE Composite on the first Chinese trading day after the press conference, and $P_{T_{-1}}(m)$ the close on the previous trading day. The next-day return is

$$r_{T_0}(m) = \frac{P_{T_0}(m) - P_{T_{-1}}(m)}{P_{T_{-1}}(m)}$$

We convert this into a binary direction label

$$y_m = 1\left[r_{T_0}(m) \geq 0\right] \in \{0,1\}$$

where $y_m = 1$ denotes an up day and $y_m = 0$ denotes down or non-positive.

The prediction task is therefore: Given the opening and Q&A (O_m , Q_m) of a press conference, produce a probabilistic forecast for the next-day SSE Composite direction y_m . Rather than training a parametric model on this small sample of meetings, we treat a large language model (LLM) as a zero-finetune event classifier that operates under strict evidence constraints. For each view $v \in$ { opening , qa } and meeting m, the LLM produces a three-way probability vector

$$p_m^{(v)} = \left(p_{up}^{(v)}, p_{down}^{(v)}, p_{unk}^{(v)} \right),\ p_{up}^{(v)} + p_{down}^{(v)} + p_{unk}^{(v)} = 1 ,$$

where "*unk*" corresponds to an explicit "unknown / abstain" option that will be used for selective prediction.

4. Methodology

4.1 Overview

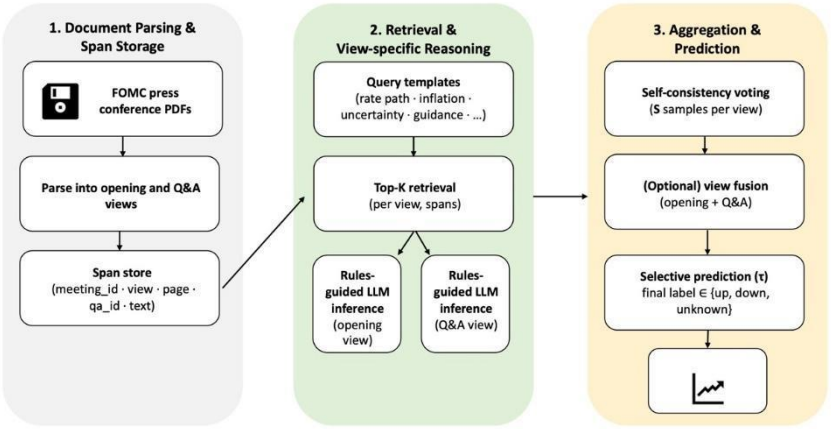

Figure 1: Overview of the proposed evidence-traced and rules-guided multi-view LLM pipeline for predicting stock-market trend from FOMC press-conference transcripts.

Figure 1 summarizes the end-to-end workflow of our evidence-traced and rules-guided multi-view LLM pipeline, from span construction and retrieval to voting and selective prediction. At a high level, the steps are:

Span store construction. Official FOMC press-conference PDFs are parsed into short text spans with metadata (meeting id, view, page, QA id). This creates a structured **span store** for each meeting.

Evidence retrieval. A small set of macro-financial query templates (rates, inflation, growth, uncertainty, global risks, etc.) is used to retrieve top-K spans from the opening and Q&A views using a standard retrieval model.

LLM classification with self-consistency. For each view, the LLM receives only the retrieved evidence spans and is prompted to output a JSON verdict in {up, down, unknown} with a confidence score and quoted evidence. We repeat sampling multiple times and aggregate via **self-consistency voting**, following the intuition of Wang et al.[9]

Rulebook-guided enhancement (proposed). On the development period, we let the LLM explain its decisions with chain-of-thought (CoT) reasoning[8] and then ask it to summarise these rationales into an explicit rulebook that links textual patterns to A-share reactions. This rulebook is then injected back into the prompts as domain knowledge at evaluation time.

Selective prediction. The final decision uses a threshold-based reject rule inspired by selective prediction / reject-option work such as SelectiveNet: the system only outputs up / down when the view-level probability and confidence exceed a threshold τ; otherwise it abstains (unknown)[10].

4.2 Evidence-Constrained Span Retrieval

The first methodological component is to construct a span-level evidence set so that the LLM is forced to base its decision on traceable passages rather than the entire transcript.

For each meeting m m and view $v \in$ { opening , qa }, we split the parsed text into spans

$s_i = (meeting_{id}, view, page, qa_{id}, text)$,

and collect the set of spans for that view as

$$S_m^v = \{ s_i : s_i belongs\ to\ meeting\ m\ and\ view\ v \}$$

We design a small set of natural-language query templates Q = { q_1, … , q_L} corresponding to macro themes such as rate path, inflation, growth, financial conditions, and uncertainty. For a query q and span s_i , a retrieval model R (q , s_i) assigns a relevance score. For each view and query, we select the top- K_v spans:

$$TopK\left(q, S_m^{(v)}, K_v\right) = arg\ arg\ R\ (q, s_i)$$

The evidence set for meeting m m and view v v aggregates top- K_v spans over all queries:

$$E_m^{(v)} = \bigcup_{q \in Q} TopK\left(q, S_m^{(v)}, K_v\right)$$

Only spans in $E_m^{(v)}$ are passed to the LLM, which makes every prediction evidence-traced.

4.3 Evidence-Constrained LLM Classifier with Self-Consistency

Given the evidence set $E_m^{(v)}$, we prompt the LLM as a **constrained classifier**:

it is told to act as a financial analyst,

it may only use the supplied spans as evidence,

it must quote 3–5 key spans (with page / qa_id),

and it must output a JSON record with fields:

verdict ∈ {up, down, unknown}, confidence ∈ [0,1], evidence[], reason.

To mitigate the stochasticity of LLM generation, we adopt **self-consistency sampling:** instead of trusting a single chain of thought, we generate multiple independent samples and aggregate them by voting, as proposed by Wang et al[9] For each pair (m,v), we call the LLM S_v times and obtain

$\{(c_j, r_j)\}_{j=1}^{S_v}$,

where $c_j \in$ {up, down, unknown} is the j-th verdict and $r_j \in [0,1]$ is the corresponding confidence.

We convert these samples into view-level probabilities through simple counts:

$n_{up} = \sum_{j=1}^{S_v} \quad 1[c_j = up], n_{down} = \sum_{j=1}^{S_v} \quad 1[c_j = down], n_{unk} = \sum_{j=1}^{S_v} \quad 1[c_j = unknown]$

$p_{up}^{(v)} = \frac{n_{up}}{S_v}$, $p_{down}^{(v)} = \frac{n_{down}}{S_v}$, $p_{unk}^{(v)} = \frac{n_{unk}}{S_v}$

$$avg\_conf^{(v)} = \frac{1}{S_v} \sum_{j=1}^{S_v} \quad r_j$$

This transforms raw LLM outputs into well-defined probabilities that can be fused across views and used by the selective-prediction rule.

4.4 Rulebook-Guided LLM

The main methodological contribution is a rulebook-guided enhancement that distills domain rules from the development set and injects them as structured prior knowledge into the LLM.

On the development period (2011–2020), we first run the evidence-constrained CoT prompting baseline in **Section 4.3**, asking the LLM to produce detailed reasoning for each meeting. Following the idea of chain-of-thought prompting, these rationales expose how the model internally links phrases such as "rates will need to stay higher for longer" or "data-dependent and cautious" to risk-on or risk-off reactions[8].

We then construct a meta-prompt that feeds the model a batch of such rationales together with the ground-truth labels $\hat{y_m}$, and ask it to synthesise a compact rulebook organised by themes:

- rate-path rules (hawkish vs dovish),
- inflation and price-stability rules,
- growth / labour-market rules,
- risk and uncertainty language,
- global and China-specific spillover rules.

The resulting rulebook is stored in a machine-readable JSON file (rule descriptions, indicative phrases, suggested direction). At evaluation time (both dev and test), we prepend this rulebook to the evidence in the prompt and require the LLM to:

identify which rules are triggered by the supplied spans $E_m^{(v)}$;

combine the activated rules into a final verdict c_j and confidence r_j;

justify the decision with a short rationale and explicit rule references.

This rulebook-guided LLM keeps the model zero-finetune, but performs a form of pseudo-supervised domain adaptation: dev-set experience is encoded in the rulebook and reused for unseen meetings, improving stability and interpretability compared with a pure CoT baseline.

4.5 Selective Prediction and Decision Rule

Finally, we translate the view-level probabilities into an operational decision with a reject option. Let a given view (or a fused

combination of views) produce p_{up}, p_{down}, $p_{unk}$and avg_conf. We define

$$top\_prob = (p_{up}, p_{down}),$$

$$c^* = \{up \ \ if \ \ p_{up} \geq p_{down};\ down\ otherwise\}$$

Given a user-chosen threshold $\tau \in [\,0\,,\,1\,]$, the selective-prediction rule is

$$\hat{y_m} = c^*\ if\ max(top\_prob, avg\_conf) \geq \tau\ and\ p_{unk} < 0.95;\ unknown\ otherwise$$

By sweeping τ on the development set we obtain an accuracy-coverage curve, analogous to selective-classification work in deep learning[10]. A lower τ yields higher coverage but lower accuracy, while a higher τ yields fewer but more reliable predictions. In the experiments, we fix τ based on this trade-off and report out-of-sample performance on the test period.

5. Experiment

5.1 Experimental setup

Data, task, and evaluation protocol. We study an event-level binary classification task. Each sample is one FOMC press conference, where we use the transcript text (parsed into opening-statement and Q&A views) as input and predict the direction (up/down) of the SSE Composite on the next China trading day. We adopt a time-based split to avoid look-ahead bias: meetings from 2011–2020 form the development set (dev) and meetings from 2021–2025 form the held-out test set (test). We use dev only to (i) distill the rulebook for rules-guided prompting and (ii) select the abstention threshold τ for selective prediction; we then freeze these choices and report all results on test. For all LLM-based variants, we use deepseek-chat through the DeepSeek API (OpenAI-compatible endpoint)[16].

Metrics. Because the proposed system supports abstention ("unknown"), we report both coverage (the fraction of meetings for which the model outputs *up/down* rather than *unknown*) and accuracy (computed only on the covered subset). We additionally report balanced accuracy and the area under the ROC curve (AUC), where AUC measures the threshold-independent ranking quality of the model scores (with 0.5 corresponding to random guessing), thereby reducing sensitivity to potential class imbalance and to any single operating threshold. For transparency, we also log τ-sweep results to characterize the accuracy–coverage trade-off and plot selective prediction curves.

Proposed method and τ selection. Our main model is Rules-Guided Prompting, where a rulebook distilled from dev is injected into the prompt as soft decision constraints. For selective prediction, we choose τ on dev using the opening-statement τ sweep. Concretely, we set τ = 0.50 (dev-selected) and apply this same value to the test split without modification. Although the test τ sweep shows that τ = 0.40 would yield slightly higher test accuracy, we do not select τ based on test, as doing so would constitute test tuning.

Baselines. We compare against four baselines spanning classical finance text heuristics, traditional text classification, domain-pretrained sentiment modelling, and unconstrained LLM prompting:

Dictionary sentiment + Logistic Regression (LM Dict + LR). We compute sentiment/uncertainty features using the Loughran–McDonald word lists and train a logistic regression classifier on dev, then evaluate on test. This is a standard finance-text baseline grounded in domain-specific lexicons[17].

TF–IDF Bag-of-Words + Logistic Regression (TF–IDF + LR). We vectorize the transcript using TF–IDF and train a linear classifier, a strong baseline family for text classification[18].

FinBERT sentiment features (FinBERT). We apply a finance-domain pretrained language model for sentiment and derive a directional signal from its sentiment outputs, then evaluate on the same dev/test split[11].

Naïve zero-shot LLM (forced decision). We prompt an LLM to directly output *up/down* on the full transcript without evidence constraints, rule guidance, or selective prediction. This represents generic zero-shot inference with large language models[19].

5.2 Main Result and Comparison

We report test-set (2021–2025) performance in **Table 1**. We report our main result at the **dev-selected threshold τ = 0.50**, and we do not tune τ on the test set. For completeness, Table 1 includes both the **opening-statement** and **Q&A** views under the same evaluation protocol. **Figure 2** plots the accuracy–coverage trade-off for the opening view as τ varies.

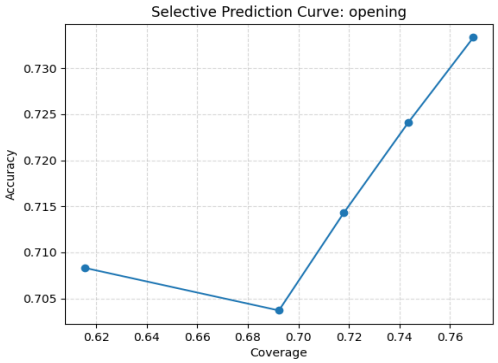

Figure 2. Selective prediction curve (Opening view): Accuracy vs. Coverage as τ varies.

Table 1. Test-set performance (2021–2025). Results are reported with a fixed τ selected on dev (τ = 0.50 for our method). Coverage is the fraction of meetings with a non-*unknown* prediction.

Method	Input	τ	Coverage	Accuracy	Balanced Acc	AUC
Rules-Guided Prompting (Ours)	Opening	**0.50**	**0.718**	**0.714**	**0.697**	**0.655**
Rules-Guided Prompting (Ours)	Q&A	0.50	0.795	0.452	0.370	0.463
TF–IDF + LR	Full text (Q&A)	–	1.000	0.564	0.500	0.567
LM Dict + LR	Full text	–	1.000	0.436	0.500	0.500
FinBERT (sentiment-based)	Truncated long text	–	1.000	0.436	0.500	0.567
Zero-shot LLM (forced)	Full transcript	–	1.000	0.487	–	–
Zero-shot LLM (+unknown)	Full transcript	–	~0.000	–	–	–

Table 1 shows that rules-guided prompting delivers the strongest out-of-sample performance under selective prediction. Using only the opening statement at τ = 0.50, our method achieves 71.4% accuracy at 71.8% coverage, with balanced accuracy 0.697 and AUC 0.655. This result exceeds all baselines in Table 1. The strongest traditional classifier, TF–IDF + LR, reaches 56.4% accuracy with full coverage and AUC 0.567. Sentiment-driven baselines remain near chance in discrimination: LM Dict + LR and FinBERT both achieve 43.6% accuracy with balanced accuracy 0.500. These comparisons show that the performance gain is not explained by "always predicting," because our method achieves higher accuracy while still allowing principled abstention.

The view split reveals a clear difference in signal quality. Under the same rules-guided framework and threshold, the Q&A view achieves 45.2% accuracy at 79.5% coverage with AUC 0.463, which is substantially weaker than the opening view. This gap suggests that the Q&A transcript introduces more topic drift and noise for short-horizon direction prediction. It also supports our choice to prioritize the opening statement as the primary decision view in the main comparison.

Naïve prompting is not competitive without the pipeline structure. When we query the LLM on the full transcript with a forced up/down answer, test accuracy is 48.7%, which is close to chance. When we allow an "unknown" option without evidence constraints or decision rules, coverage drops to near zero, because the model abstains on almost all meetings. Together, these results show that structure—retrieval-based evidence, rules guidance, and a selective decision policy—is necessary to turn general language understanding into a stable event-level predictor.

Why we report τ = 0.50 although τ = 0.40 is best on test. The test-time τ sweep indicates that τ = 0.40 would yield slightly higher test accuracy for the opening view. However, τ must be fixed before observing test outcomes to avoid leakage. We therefore set τ = 0.50 based on development selection and report the corresponding test performance as the main result. We include the full τ sweep in the appendix for transparency, but we do not use it for model selection.

6. Discussion

6.1 Sensitivity to threshold τ and view choice

Table 2 reports the sensitivity of the proposed system to the abstention threshold τ and the choice of textual view on the held-out test set (2021–2025). Because our pipeline supports abstention ("unknown"), we report both coverage (the fraction of meetings for which the model outputs *up/down* rather than *unknown*) and accuracy (computed only on the covered subset). We additionally report balanced accuracy and AUC to reduce sensitivity to potential class imbalance, and we include Brier score to reflect the quality of probabilistic predictions.

Table 2. Sensitivity to threshold τ and view choice on the held-out test set (2021–2025). Metrics include coverage, accuracy on the covered subset, balanced accuracy, AUC, and Brier score. The dev-selected operating point τ=0.50 is highlighted for reporting test performance under a no-test-tuning protocol.

View / (\tau)	Coverage	Acc.	Bal. Acc.	AUC	Brier
Opening @ 0.40	0.769	0.733	0.701	0.655	0.316
Opening @ 0.50 (dev-selected)	**0.718**	**0.714**	**0.697**	**0.655**	**0.316**
Opening @ 0.60	0.615	0.708	0.689	0.655	0.316
QA @ 0.40	0.974	0.447	0.403	0.463	0.553
QA @ 0.50	0.795	0.452	0.370	0.463	0.553
QA @ 0.60	0.692	0.481	0.342	0.463	0.553
Fused @ 0.40	0.000	—	—	0.485	0.422
Fused @ 0.50 (dev-selected)	**0.000**	**—**	**—**	**0.485**	**0.422**
Fused @ 0.60	0.000	—	—	0.485	0.422

As shown in **Table 2**, the opening statement view provides the strongest and most stable signal on the test period. At the dev-selected operating point τ=0.50, the opening view achieves 0.714 accuracy at 0.718 coverage, with AUC 0.655 and Brier 0.316. Increasing τ from 0.40 to 0.60 reduces coverage (0.769 → 0.615), consistent with a stricter abstention rule, while conditional accuracy remains relatively stable (0.733 → 0.708). This pattern suggests that, under rules-guided prompting, the opening statement is both informative and robust to moderate threshold changes, making it a reliable operating view for deployment-oriented settings where abstention is permitted.

In contrast, the Q&A view exhibits substantially weaker predictive quality despite higher coverage. At τ=0.50, Q&A covers 0.795 of meetings but attains only 0.452 accuracy, accompanied by a lower AUC (0.463) and a markedly worse Brier score (0.553). These results indicate that the Q&A transcript, as currently processed, injects considerable uncertainty and noise. A plausible interpretation is that the Q&A view contains heterogeneous content (e.g., repeated clarifications, journalist framing, topic switching, or discussion of constraints and contingencies) whose relation to next-day A-share direction is not easily extracted by a simple evidence selection and aggregation strategy. Importantly, this does not imply that Q&A lacks incremental information; rather, it suggests that Q&A is harder to operationalize and may require more structured retrieval and aggregation to isolate policy-relevant signals.

Finally, the fusion variant in this experiment collapses to complete abstention (coverage=0) across all tested τ. When fusion outputs are dominated by uncertainty (e.g., high p_{unk} or near-zero decisive mass), the selective decision rule yields "unknown" for all meetings, rendering accuracy undefined. This failure mode highlights that naïve multi-view fusion can be brittle when component views have mismatched uncertainty behaviour or poorly aligned confidence. In practice, multi-view fusion should be treated as a modelling problem in its own right—requiring reliability-aware weighting, calibration, or learning-based combination—rather than a fixed heuristic. Given the stability and strength of the opening view in **Table 2**, our main empirical conclusion in this work is that opening-only rules-guided prompting is currently the most dependable configuration under the proposed evaluation protocol.

6.2 Why structure matters: from naïve prompting to an auditable selective-prediction pipeline

Naïve prompting does not yield a reliable event-level decision in our setting. Table 1 shows that a zero-shot LLM with an "unknown" option abstains almost everywhere, which results in near-zero coverage. Table 1 also shows that forcing a binary up/down answer raises coverage to 1.0 but yields ≈48.7% test accuracy, which is close to chance[19]. In contrast, our evidence-traced pipeline achieves 71.4% accuracy at 71.8% coverage for the opening view under the dev-selected τ = 0.50. This gap indicates that *prompt structure and decision policy* matter more than a "single call" to the model.

Our pipeline makes the decision process grounded and auditable. The system first retrieves a compact set of relevant spans and then requires the model to justify its verdict with quoted evidence. This design follows the general idea of retrieval-plus-grounding, where retrieved passages provide provenance and improve controllability compared with unconstrained generation[13]. The pipeline also enforces a structured output format, which reduces ambiguity in evaluation and keeps the

prediction trace inspectable at the event level.

The pipeline also makes uncertainty measurable and actionable. Self-consistency aggregates multiple sampled reasoning paths into a more stable distribution, instead of relying on a single stochastic output[8][9]. Selective prediction then uses τ to convert that distribution into an explicit **accuracy–coverage** trade-off, rather than forcing low-evidence guesses[10]. Prior work also suggests that language models can provide meaningful self-evaluation signals when the task format elicits them properly[20]. In our results, "unknown" therefore becomes a principled fallback, not an ad-hoc refusal, and the model's uncertainty becomes a controllable part of the decision rule.

7. Limitation

This study evaluates event-level prediction on a small and regime-dependent sample. FOMC press conferences occur only a few times per year. The mapping from U.S. policy language to next-day China market direction can also shift across macro regimes, such as tightening versus easing cycles and different volatility states. These factors increase estimation variance and can make the effect of the abstention threshold τ look unstable or non-monotonic in finite samples. The empirical scope is also narrow. The paper focuses on one communication channel, one main index, and a one-day horizon, so external validity remains limited.

The outcome definition is intentionally simple, but it is also noisy. The label is the sign of the next trading-day return. Many drivers of that return are not contained in the transcript, including domestic announcements, concurrent global shocks, positioning, liquidity conditions, and intraday news flow. As a result, even an evidence-grounded interpretation of policy language may still miss the realized direction. This places a practical ceiling on accuracy for transcript-only methods.

The pipeline is less robust when inputs become noisier or when views behave differently. The Q&A view is challenging because it is multi-topic and repetitive, and it can dilute market-relevant clarifications with generic policy language. The fusion failure on the test set reinforces this point. The fused variant collapses to near-zero coverage under the current decision rule. This negative result suggests that multi-view combination is not a simple "additive" gain in this task. It can fail when the two views carry mismatched uncertainty and inconsistent signals.

8. Conclusion

This paper studies whether the textual content of Federal Reserve FOMC press conferences can provide actionable information for predicting the direction of the next trading-day return of the China A-share market. We propose an engineering-oriented, auditable LLM pipeline that operates without parameter fine-tuning: it retrieves a compact evidence set from the transcript, constrains the model to reason only over retrieved spans, and aggregates multiple stochastic generations via self-consistency voting to obtain view-level class probabilities and uncertainty.

To improve robustness across time, we distill a human-readable rulebook from the development period and inject it into the prompt as soft decision constraints during inference. The system further supports selective prediction by allowing an abstention option ("unknown") and applying a dev-selected threshold τ to trade off accuracy and coverage in a principled manner.

Empirically, on the held-out 2021–2025 test set, the rules-guided model applied to the opening statement achieves 71.4% accuracy at 71.8% coverage (balanced accuracy 0.697 and AUC 0.655) with $\tau = 0.5$, outperforming sentiment-only baselines and a naive zero-shot LLM call. In contrast, the Q&A view is substantially noisier under the current retrieval and prompting scheme, suggesting that structured policy language in the prepared opening statement is more reliable for out-of-sample directional discrimination. Overall, the results support the practical value of evidence-constrained LLM inference with abstention for high-uncertainty market prediction tasks and motivate future work on topic-aware retrieval, reliability-aware fusion, and broader cross-market evaluation.

Reference

[1] A. S. Blinder, M. Ehrmann, M. Fratzscher, J. de Haan, and D.-J. Jansen, "Central bank communication and monetary policy: A survey of theory and evidence," Journal of Economic Literature, vol. 46, no. 4, pp. 910–945, 2008, doi: 10.1257/jel.46.4.910.

[2] S. Hansen and M. McMahon, "Shocking language: Understanding the macroeconomic effects of central bank communication," Journal of International Economics, vol. 99, suppl. 1, pp. S114–S133, 2016, doi: 10.1016/j.jinteco.2015.12.008.

[3] R. S. Gürkaynak, B. Sack, and E. T. Swanson, "Do actions speak louder than words? The response of asset prices to monetary policy actions and statements," International Journal of Central Banking, vol. 1, no. 1, pp. 55–93, 2005.

[4] T. Doh, D. Song, and S.-K. Yang, "Deciphering Federal Reserve communication via text analysis of alternative FOMC statements," Federal Reserve Bank of Kansas City, Research Working Paper RWP 20-14, Oct. 2020 (updated Oct. 2025), doi: 10.18651/RWP2020-14.

[5] D. O. Lucca and E. Moench, "The pre-FOMC announcement drift," The Journal of Finance, vol. 70, no. 1, pp. 329–371, 2015, doi: 10.1111/jofi.12196.

[6] J. Hausman and J. Wongswan, "Global asset prices and FOMC announcements," Journal of International Money and

Finance, vol. 30, no. 3, pp. 547–571, 2011, doi: 10.1016/j.jimonfin.2011.01.008.

[7] T. Loughran and B. McDonald, “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,” The Journal of Finance, vol. 66, no. 1, pp. 35–65, 2011, doi: 10.1111/j.1540-6261.2010.01625.x.

[8] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” arXiv:2201.11903, 2022, doi: 10.48550/arXiv.2201.11903.

[9] X. Wang et al., “Self-consistency improves chain-of-thought reasoning in language models,” in Proc. Int. Conf. Learn. Representations (ICLR), 2023. [Online]. Available: https://openreview.net/forum?id=1PL1NIMMrw. (See also: arXiv:2203.11171.)

[10] Y. Geifman and R. El-Yaniv, “SelectiveNet: A deep neural network with an integrated reject option,” arXiv:1901.09192, 2019, doi: 10.48550/arXiv.1901.09192.

[11] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” arXiv:1908.10063, 2019, doi: 10.48550/arXiv.1908.10063.

[12] Y. Gorodnichenko, T. Pham, and O. Talavera, “The voice of monetary policy,” NBER Working Paper 28592, Mar. 2021, doi: 10.3386/w28592.

[13] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020. (See also: arXiv:2005.11401.)

[14] A. Lopez-Lira and Y. Tang, “Can ChatGPT forecast stock price movements? Return predictability and large language models,” SSRN, 2023, doi: 10.2139/ssrn.4412788.

[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 70, pp. 1321–1330, 2017.

[16] DeepSeek, “Your First API Call,” DeepSeek API Docs. Accessed: Dec. 23, 2025. [Online]. Available: https://api-docs.deepseek.com/.

[17] Software Repository for Accounting and Finance (SRAF), “10X Summaries: 1993–2024,” University of Notre Dame. Accessed: Dec. 20, 2025. [Online]. Available: https://sraf.nd.edu/sec-edgar-data/lm_10x_summaries/.

[18] S. Wang and C. D. Manning, “Baselines and bigrams: Simple, good sentiment and topic classification,” in Proc. 50th Annu. Meeting Assoc. Comput. Linguist. (ACL), vol. 2 (Short Papers), pp. 90–94, 2012. [Online]. Available: https://aclanthology.org/P12-2018/.

[19] T. B. Brown et al., “Language models are few-shot learners,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020. (See also: arXiv:2005.14165.)

[20] S. Kadavath et al., “Language models (mostly) know what they know,” arXiv:2207.05221, 2022, doi: 10.48550/arXiv.2207.05221.