FedSpeak Decoder: Does Linguistic Surprise Predict Market Surprise?

Bohong (Lucia) Liu | Stanford CS109 Probability Challenge, Summer 2026

Interactive demo: https://94007369.github.io/fedspeak-decoder/

Abstract

Federal Open Market Committee (FOMC) statements are short, repetitive, and scrutinized word by word. I built *FedSpeak Decoder*, an interactive tool that measures how surprising a statement is relative to earlier Federal Reserve language and displays market reactions after similar historical statements. Using 221 statements from 2000–2025, I test whether unusually surprising language is associated with unusually large changes in the 2-year Treasury yield. High-surprise statements have a higher observed large-move rate (31.8% versus 21.8%), but the mutual information is only 0.0092 bits and a 5,000-draw permutation test gives $p = 0.130$. The result suggests a direction, not reliable predictive evidence.

1. Motivation

A phrase such as "inflation has eased" may matter precisely because it is unusual in the Fed's historical vocabulary. The project asks:

> *When the Fed changes its language, does the information content of that change help us anticipate an unusually large move in short-term interest rates?*

The website lets a user paste or edit a statement, highlights unusual words, retrieves similar meetings, and summarizes their market reactions. Its most plausible users are event-driven macro and fixed-income traders, central-bank watchers, risk managers, and financial journalists who need a rapid first-pass reading of a release. The simple bag-of-words model is inexpensive and auditable: every score can be traced to words and historical observations. This transparency comes at the cost of word order, semantic nuance, and investor expectations, so the output is a screening signal rather than trading advice.

An important distinction motivates the model. A statement with varied vocabulary can have high *lexical entropy* without being vague. I therefore do not interpret raw entropy as "policy ambiguity." The primary measure is *historical surprisal*: how improbable the words are under a language distribution estimated from statements available before that meeting.

2. Data and Random Variables

The corpus contains 221 official FOMC statements from 2000–2025. I align them with the daily 2-year constant-maturity Treasury yield from FRED (series `DGS2`). After excluding the first statement, 220 meetings have both a chronological language score and a market observation.

For meeting t, let S_t denote average linguistic surprisal and define the absolute event-day move

$$M_t = 100\left|y_t - y_t^-\right| \quad \text{basis points}, \tag{1}$$

where y_t is the first available event-date yield and y_t^- is the previous trading-day yield. I define binary random variables

$$X_t = \mathbf{1}\{S_t \geq \text{median}(S)\}, \tag{2}$$

$$Y_t = \mathbf{1}\{M_t \geq Q_{0.75}(M)\}. \tag{3}$$

Thus $X_t = 1$ is HighSurprise (threshold 8.722 bits) and $Y_t = 1$ is LargeMove (threshold 8 basis points). Thresholding makes the test easy to explain, although it discards magnitude.

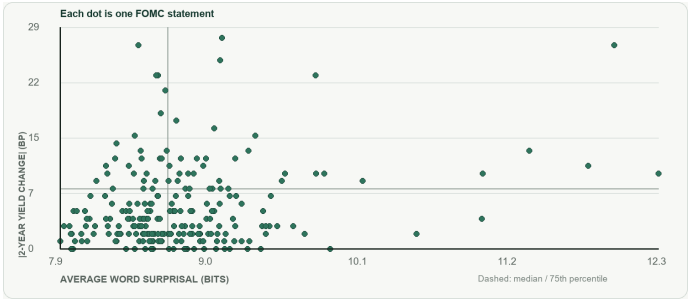

Figure 1: Average word surprisal versus the absolute event-day 2-year yield change ($n = 220$). Dashed lines mark the median surprisal and 75th-percentile move.

3. From Entropy to Surprisal

For a statement with empirical word probabilities $p(w)$ and vocabulary V, normalized lexical entropy is

$$H_{\text{norm}}(W) = \frac{-\sum_{w \in V} p(w) \log_2 p(w)}{\log_2 |V|}. \tag{4}$$

This statistic, bounded by $[0, 1]$, describes vocabulary dispersion inside one statement. It is displayed in the interface but is not treated as a market predictor.

For historical surprisal, let $C_{t-1}(w)$ be the count of word w before meeting t, N_{t-1} the total number of earlier tokens, and $|V|$ the corpus vocabulary size. Additive smoothing with $\alpha = 0.5$ prevents unseen words from receiving probability zero:

$$\widehat{P}_t(w) = \frac{C_{t-1}(w) + \alpha}{N_{t-1} + \alpha|V|}. \tag{5}$$

The information in a word and average statement surprisal are

$$I_t(w) = -\log_2 \widehat{P}_t(w), \qquad S_t = \frac{1}{n_t} \sum_{i=1}^{n_t} I_t(w_i). \tag{6}$$

Common boilerplate receives little weight; historically rare words receive more. Because each statement is scored only against prior language, future frequencies are not used to evaluate the past.

4. Interactive Bayesian Estimate

For a user-entered statement, the decoder finds $k = 8$ historical neighbors. Text is lowercased and tokenized with the regular expression `[a-z]+(?:'[a-z]+)?`; one-letter tokens and 25 common stop words are removed. No stemming or lemmatization is applied. For document d, the vector is raw term frequency, not

TF–IDF:

$$v_d(w) = c_d(w)\mathbf{1}\{I(w) > 7\}, \qquad \cos(a,b) = \frac{v_a \cdot v_b}{\|v_a\|_2 \|v_b\|_2}. \tag{7}$$

Here $I(w)$ is computed from the smoothed frequency of w in the complete 221-statement corpus. Thus "informative vocabulary" means words exceeding 7 bits of corpus surprisal; surprisal filters the coordinates but does not weight them. Cosine similarity compares term-frequency direction, reducing sensitivity to statement length. Statements with a market observation are ranked by Eq. 7, and the top eight are retained.

4.1. Preprocessing and design choices

The stop-word list is deliberately short: articles, conjunctions, auxiliary verbs, and other very common function words are removed, while economically meaningful words such as *inflation*, *employment*, *risks*, and *uncertainty* remain. Possessives such as *committee's* are preserved as tokens. Numbers and punctuation are discarded, so the current similarity model cannot distinguish, for example, a 25-basis-point change from a 50-basis-point change unless the surrounding words differ.

Raw term frequency was chosen because the separate $I(w) > 7$ rule already removes the most common boilerplate. Applying TF–IDF as well would downweight words a second time and would make the demonstration harder to explain. Likewise, stemming was avoided so that displayed words correspond exactly to what the user typed. This improves transparency but treats forms such as *increase*, *increased*, and *increases* as different coordinates. The method is therefore a sparse bag-of-words comparison: it ignores word order, syntax, and semantic equivalence.

The threshold of 7 bits has a probability interpretation. Since $I(w) = -\log_2 P(w)$, retained words have smoothed corpus probability below $2^{-7} \approx 0.0078$. It is a heuristic interface parameter rather than a fitted optimum. The neighbor count $k = 8$ is also fixed in advance: smaller k gives more local but noisier comparisons, whereas larger k gives a more stable estimate but may include less relevant meetings.

Let L be the number of these neighbors followed by a LargeMove. Reporting L/k directly would be unstable, so the tool uses a uniform Beta prior for the unknown conditional probability p:

$$p \sim \text{Beta}(1,1), \tag{8}$$

$$p \mid L \sim \text{Beta}(L+1, k-L+1), \tag{9}$$

$$\mathbb{E}[p \mid L] = \frac{L+1}{k+2}. \tag{10}$$

This posterior mean is the displayed "large-move probability." It shrinks a small neighborhood toward $1/2$ and is best interpreted as a smoothed historical conditional frequency, not a live forecast.

The full posterior also describes uncertainty:

$$\text{Var}(p \mid L) = \frac{(L+1)(k-L+1)}{(k+2)^2(k+3)}. \tag{11}$$

With only eight neighbors, that variance can be substantial. For example, if three of eight meetings produced a large move, the raw frequency is $3/8 = 0.375$, the posterior mean is $4/10 = 0.400$, and uncertainty remains wide. The point estimate should therefore be read together with the underlying neighbor outcomes rather than as precise predictive confidence.

The interface shows colored word surprisals, a historical percentile, and each neighbor's similarity and realized move. Low similarities warn that the text lies outside the model's historical support.

Table 1: Historical event-study results.

Statistic	Low S_t	High S_t
P(LargeMove | group)	21.8%	31.8%
Mean M_t	5.19 bp	5.82 bp
Mutual information		0.0092 bits
Permutation test		$p = 0.130$

4.2. From a user action to an output

An interaction proceeds in six deterministic steps. First, the text is normalized and tokenized. Second, every retained token receives the smoothed corpus surprisal $I(w)$; these values determine the highlighting and their mean determines the displayed statement score. Third, tokens with $I(w) \le 7$ are removed only for retrieval, and the remaining counts form v_a. Fourth, cosine similarity is calculated against every historical vector v_b. Fifth, the eight highest-scoring eligible statements are selected and their absolute DGS2 moves are classified using the same 8-basis-point threshold as the event study. Sixth, L successes update the Beta$(1,1)$ prior and the posterior mean is displayed.

These stages answer different questions: word highlighting identifies rarity, cosine similarity identifies lexical resemblance, and the Beta update summarizes outcomes among the retrieved neighbors. None alone estimates the causal effect of a phrase, and lexical similarity need not imply semantic equivalence.

5. Mutual Information Test

If X denotes HighSurprise and Y denotes LargeMove, mutual information measures how much observing X reduces uncertainty about Y:

$$I(X;Y) = \sum_{x,y \in \{0,1\}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}. \tag{12}$$

Under independence, $I(X;Y) = 0$. A finite sample nevertheless produces a nonnegative estimate, so I construct a null distribution by randomly permuting Y 5,000 times while holding X fixed. The Monte Carlo p-value is

$$\widehat{p} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{I_b \ge I_{\text{obs}}\}}{B+1}, \qquad B = 5000. \tag{13}$$

Permutation preserves the observed number of high-surprise statements and large moves while breaking their pairing. It therefore asks a precise probability question: if language and market moves were unrelated, how often would random reassignment produce dependence at least this large? The added ones in Eq. 13 make the simulated test valid even when no shuffled statistic exceeds the observed value.

6. Results

Table 1 and Fig. 2 summarize the experiment. Large moves occur after 31.8% of high-surprise statements versus 21.8% of low-surprise statements. Mean absolute moves differ less: 5.82 versus 5.19 basis points.

The observed mutual information is only 0.0092 bits, and 13.0% of shuffled datasets produce at least as much. Therefore, the experiment does not reject independence at the conventional 5% level. The difference in group rates is descriptive evidence of

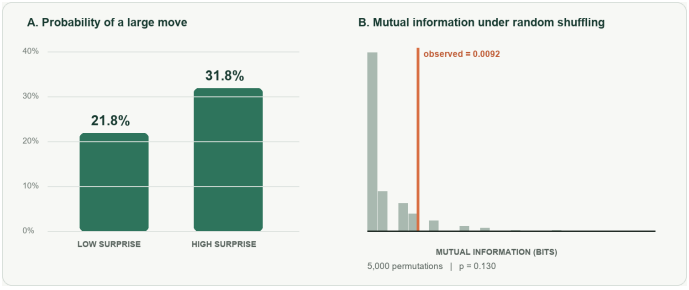

Figure 2: Observed group probabilities (left) and the permutation null distribution of mutual information (right). The orange line is $I_{\text{obs}} = 0.0092$.

a possible relationship, not strong evidence that language alone predicts market reactions.

The effect is easier to interpret in probability units. Splitting the 220 observations at the median creates 110 meetings per language group. The estimated large-move probability rises by 10.0 percentage points, from 0.218 to 0.318, while the unconditional rate is 0.268. In contrast, the difference in mean absolute moves is only 0.63 basis points. Together these summaries suggest that any signal is concentrated near the upper tail rather than shifting the entire move distribution. However, the permutation result shows that a gap of this size is not exceptionally rare in a sample of 220 meetings.

6.1. Connection to CS109

The project combines four course ideas in one pipeline. Entropy summarizes uncertainty in a discrete word distribution; surprisal converts an observed word into information measured in bits; conditional probability describes outcomes among comparable statements; and Beta–Binomial conjugacy regularizes that probability when only eight neighbors are available. Mutual information then measures dependence without assuming a linear relationship, while random permutation approximates its sampling distribution under the null. The website is therefore an interface for inspecting explicit probabilistic assumptions rather than merely a text classifier.

6.2. Sensitivity and alternative specifications

The reported hypothesis test fixes a chronological language model, the median split for surprise, the 75th percentile for a large move, and the 2-year yield as the outcome. These choices make the experiment interpretable, but alternative cutoffs could change the estimated gap. Repeatedly searching cutoffs and reporting only the strongest one would create a multiple-comparisons problem. A confirmatory extension should preregister the thresholds or replace the binary variables with a continuous model evaluated out of sample.

The retrieval system and the hypothesis test also serve different purposes. The formal mutual-information result uses chronological statement surprisal for all 220 meetings. The interactive neighbor search instead uses full-corpus word frequencies to create one stable vocabulary for arbitrary user input. Consequently, the neighbor probability is an exploratory historical analogy, not the estimator tested by the permutation experiment. Keeping this distinction explicit prevents the interactive display from being mistaken for validated forecasting performance.

Several richer representations are possible: TF–IDF could provide graded rarity, lemmatization could merge inflected forms, and sentence embeddings could capture paraphrases. Each adds assumptions and requires out-of-sample evaluation. I retain raw frequency because every coordinate and neighbor can be audited directly.

6.3. Cross-central-bank generalization

The framework can extend to the ECB, Bank of England, Bank of Japan, and People's Bank of China (PBOC), but each institution needs its own corpus and local market outcome. For China, quarterly PBOC Monetary Policy Committee releases or China Monetary Policy Reports could be paired with Chinese government-bond yields or RMB exchange-rate moves. A fair comparison would use within-bank surprisal percentiles. Differences across banks could reveal whether concise statements, vote disclosure, press conferences, or translation practices change how much information markets extract from official language. Generalization can therefore test how communication institutions shape market response.

7. Limitations

- Markets respond to text relative to investor expectations, which are not directly observed.
- Daily yields also include the rate decision, press conference, and concurrent news. Intraday data would isolate the statement window more cleanly.
- A unigram model ignores word order and negation. Comparing changed sentences with the immediately previous statement would better match how traders read FOMC releases.
- Thresholding improves interpretability but discards information. A preregistered continuous, out-of-sample test would be stronger.

There is also selection and timing uncertainty. FOMC language conventions changed across chairs and policy regimes, so an early-2000s statement may be a poor comparison for a modern one even when its bag-of-words vector is close. Daily DGS2 observations do not identify the exact announcement-window response and may absorb other macroeconomic news. A stronger event study would use intraday Treasury futures, market-implied expectations before each meeting, and rolling or regime-specific language models.

8. Conclusion

FedSpeak Decoder demonstrates how entropy, surprisal, Bayesian updating, and mutual information can turn subtle policy language into an interactive probability experiment. Historical surprisal is more defensible than equating raw word entropy with ambiguity, and the Beta-Binomial neighbor model communicates uncertainty rather than false precision. The present conclusion is deliberately modest: unusual language may accompany larger rate moves, but this sample does not establish a reliable predictive relationship.

Sources and AI-use disclosure

Federal Reserve Board, *FOMC Historical Materials*; Federal Reserve Bank of St. Louis, FRED series DGS2; People's Bank of China, *Monetary Policy Committee Meetings*; C. E. Shannon, "A Mathematical Theory of Communication," 1948. I used Codex for code scaffolding, interface styling, and figure generation. I independently collected and verified the data and sources, and I reviewed all definitions, calculations, and interpretations. The final analysis, results, and report reflect my own research decisions and work.