Monetary Policy Shocks: A New Hope

Large Language Models and Central Bank Communication

Rubén Fernández-Fuertes
Department of Finance, Bocconi University
ruben.fernandez@phd.unibocconi.it | rubenfernandezfuertes.com

Motivation

Classic Question: How does **monetary policy** transmit to the real economy and asset prices? Key challenge: separate **shocks** from endogenous responses.

Market-Based Measures

▷ Raw surprises are **contaminated** and **predictable** (Bauer & Swanson 2023)
▷ Different instruments (FF1/4, ED1/4, MP1) give **different results**
▷ **Doesn't track** the FOMC communication timeline

Narrative Approaches

▷ Human dependent (R&R 2023: "expect to be made redundant by computers")
▷ Pure narrative (R&R 1989, 2023) is **binary** (Shock/No Shock)
▷ Limited to internal documents (**5-year lag**)

This Paper: Multi-agent LLM extracts **probabilistic expectations** from Fed docs → **cleaner surprises**

Data: 269 FOMC meetings (1996–2025)

Methodology

Direct extraction of $\mathbb{E}[\Delta i_t|\mathcal{B}_t]$
not *ex post* econometric cleaning

Why Multi-Agent Systems?

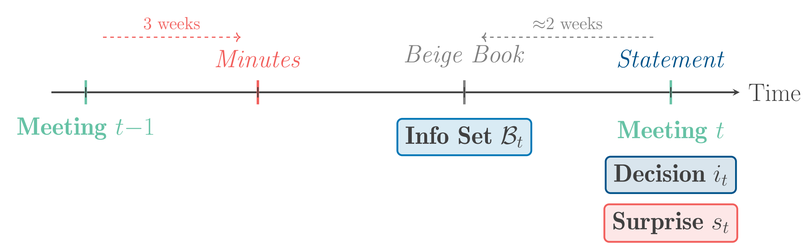
▷ **Context constraints:** Cannot process entire document sets simultaneously
▷ **Task complexity:** LLMs become "confused" with multiple instructions
▷ **Specialized expertise:** Different documents require domain-specific analysis

Key Equations:

$$\text{Surprise: } \varepsilon_t = \Delta i_t - \mathbb{E}[\Delta i_t|\mathcal{B}_t]$$
$$\text{Prior: } \mathbb{E}[\Delta i_t] = \sum_k p_k \cdot \Delta i_k$$

▷ $\mathcal{B}_t$: Beige Book + Minutes + History
▷ **Probabilistic:** full distribution, not point estimate
▷ **Real-time:** uses only pre-meeting information

FOMC Communication Timeline



Information Decomposition:

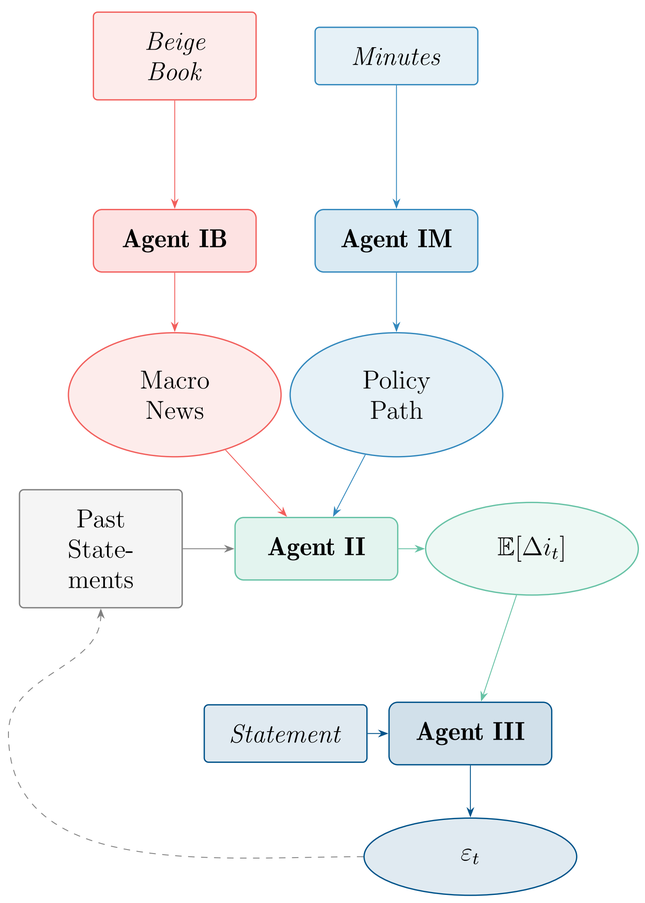
Public Info $\mathcal{B}_t$: Minutes, Beige Book, Speeches

Private Info $\mathcal{F}_t \setminus \mathcal{B}_t$: Staff forecasts, Deliberations

$$\underbrace{\Delta i_t - \mathbb{E}[\Delta i_t|\mathcal{B}_t]}_{\text{LLM Surprise } s_t} = \underbrace{\mathbb{E}[\Delta i_t|\mathcal{F}_t] - \mathbb{E}[\Delta i_t|\mathcal{B}_t]}_{\text{Fed Private Info}} + \underbrace{\Delta i_t - \mathbb{E}[\Delta i_t|\mathcal{F}_t]}_{\text{Pure Shock}}$$

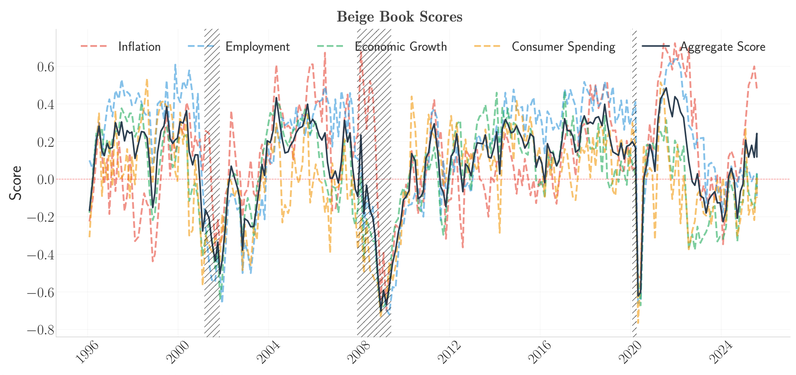
$i_t = \mathbb{E}[i_t] + s_t$. ***MZ:*** i_t *on* $\mathbb{E}[i_t] \to \alpha$=0, β=1. ***Calibration:*** Δi_t *on* $\hat{s}_t \to \beta$=1.

Multi-Agent Framework



GPT-4.1-mini | Knowledge cutoff: June 2024

Pipeline Outputs

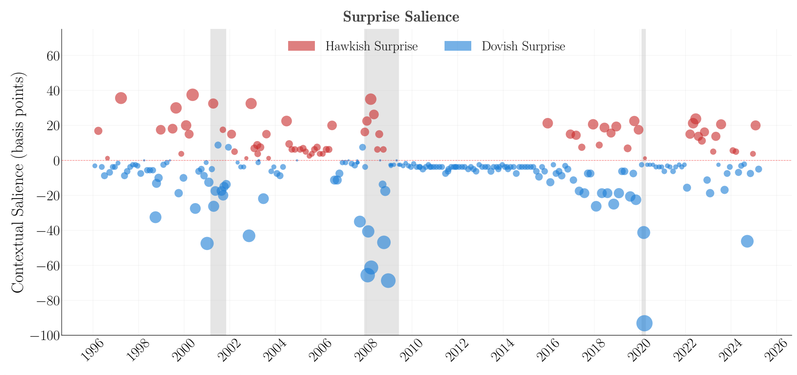


Stage 1: BB Predicts Δi_t

	(1)	(2)	(3)	(4)
i_{t-1}	-.006	-.008	-.011*	-.013**
	(.006)	(.006)	(.006)	(.006)
BB^{agg}		.327***		
		(.054)		
BB^{emp}			**.666****	**.208*****
			(.267)	(.071)
BB^{infl}			.311	.088
			(.221)	(.057)
BB^{growth}			.255	.028
			(.258)	(.087)
R^2	0.4%	12.6%	13.5%	14.5%

Stage 1 Findings:

Policy inertia alone explains virtually nothing ($R^2 = 0.4\%$). The Beige Book aggregate achieves $R^2 = 12.5\%$, capturing 94% of full model power. **Employment dominates**: a one-unit increase predicts 25bp tightening. Adding inflation barely improves fit, not because the Fed ignores it, but because inflation data (CPI/PCE) is already public. Employment discussions provide **incremental forward-looking content** (hiring intentions, labor tightness) that official statistics miss.



Stage 2: Mincer-Zarnowitz Test

	(1)	(2)	(3)	(4)	(5)	(6)
i_{t-1}	-.006	-.008	-.001	-.008	**-.014****	-.001
	(.006)	(.006)	(.004)	(.006)	(.006)	(.005)
BB^{agg}	—	**.325*****	.070	**.325*****	**.317*****	.076*
		(.054)	(.044)	(.054)	(.053)	(.044)
$E[\Delta i]$	—	—	**.990*****	—	—	**1.01*****
			(.069)			(.072)
$\sigma[\Delta i]$	—	—	—	-.032	—	.448
				(.371)		(.291)
$Skew[\Delta i]$	—	—	—	—	**-.028****	.004
					(.012)	(.010)
R^2	.4%	12.5%	**50.6%**	12.5%	14.2%	**51.0%**

Stage 2 Findings:

Agent II synthesizes Beige Book scores with Minutes-based forward guidance to form probability distributions. This **quadruples** explanatory power: R^2 rises from 12.5% to **50.6%**. Negative skewness alone predicts cuts: when Fed communications signal asymmetric downside risks, cuts follow. But skewness is subsumed by E[Δi], which passes the **Mincer-Zarnowitz test** ($\alpha \approx 0$, $\beta \approx 1$). The first moment is a **sufficient statistic** for distributional information.

Stage 3: Surprise Calibration

Surprise Validity Test

$$\Delta i_t = \alpha + \beta \hat{s}_t + u_t, \quad H_0\text{: } \alpha = 0,\ \beta = 1$$

α=0: unbiased; β=1: well-calibrated (1bp $\hat{s}_t \to$ 1bp Δi_t)

	Narrative		Market-Based				
	LLM	R&R	FF4	FF1	MP1	ED1	ED4
β	**1.02**	.78**	1.83**	2.38***	1.41	1.68*	1.32
	(.08)	(.09)	(.36)	(.49)	(.27)	(.36)	(.31)
R^2	**.52**	.43	.17	.16	.15	.16	.14
N	272	178	221	221	221	221	222

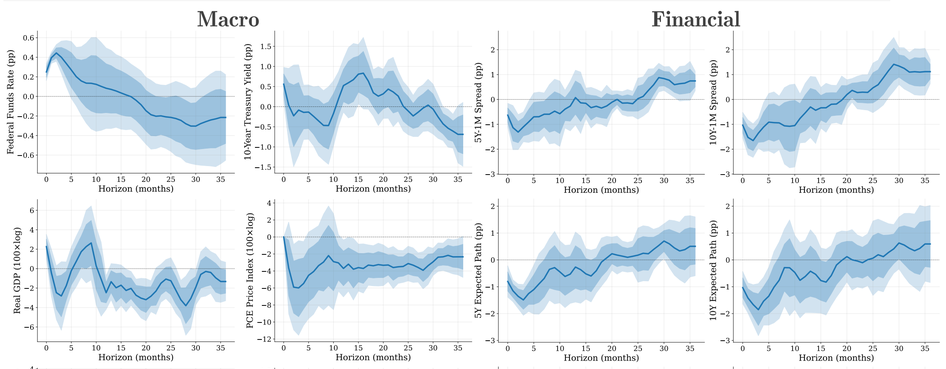
Predictability Regressions

	Narrative		Market-Based			
	R&R	LLM	FF1	FF4	ED1	ED4
Comm. Idx	-.01	.02*	-.00	.00	.00	.01**
	(.02)	(.01)	(.00)	(.00)	(.00)	(.01)
Nf. Payrolls	.01	**.02***	.00	**.00****	**.01****	**.01*****
	(.03)	(.01)	(.00)	(.00)	(.00)	(.00)
Term Spr.	**-.05*****	-.02*	-.00	**-.01*****	-.01*	**-.01****
	(.01)	(.01)	(.00)	(.00)	(.00)	(.00)
S&P 500	-.02	**.04*****	.01	**.01*****	**.01****	**.02*****
	(.02)	(.02)	(.00)	(.01)	(.01)	(.01)
R^2	.20	**.16**	.05	.15	.11	.20
N	184	223	230	230	230	231

Impulse Responses

$$y_{t+h} - y_{t-1} = \alpha_h + \beta_h\, i_t + \gamma_h X_t + u_{t+h}$$

IV: $\hat{s}_t$ instruments for i_t | 25bp shock, ZLB excluded, HAC s.e.



▷ **No price puzzle**: GDP, IP, Prices all ↓ after tightening—resolves classic identification problem.

▷ **Expectations channel**: yield curve **flattens** (5Y/10Y spreads fall); term premia rise.

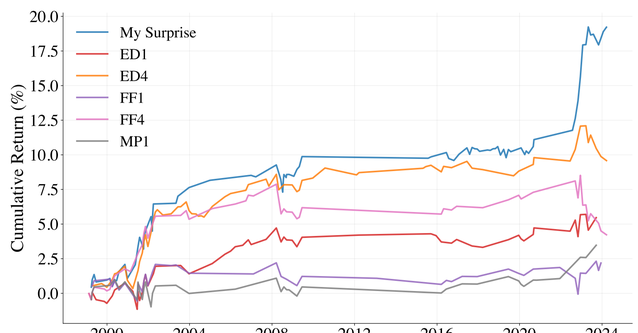
▷ **Tighter inference**: narrower CIs than market-based IV (higher first-stage F).

$$r_t^{(n)} - r_t^{(1m)} = \textstyle\sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) \mathbb{E}_t[\Delta r_{t+i}^{(1m)}] + \theta_t^{(n)} \quad \text{(Spread = Exp. Path + Term Premium)}$$

Trading Strategy

$$\text{Pos}_t = -\text{sign}(\hat{s}_t) \times \mathbb{I}[|\hat{s}_t| > q_{67}]$$

Hawkish → short 1m / long 10Y (bet on flattening)



Performance (180-day, Equal-Weight)

Signal	Ann.Ret	Sharpe	Hit	Tot.Ret
LLM	**0.69%*****	**0.68**	67.1%	**19.2%*****
ED4	0.36%**	0.41	63.4%	9.6%**
FF4	0.16%	0.18	57.1%	4.2%
ED1	0.21%	0.24	55.9%	5.5%
FF1	0.09%	0.11	50.0%	2.2%

▷ LLM **outperforms** ED4 by **92%** at 6-month horizon.

▷ Returns peak at 6–10 months = when IRFs show max spread adjustment.

▷ Duration hedging (DV01) doubles returns by isolating spread signal.

Validation

▷ **Temporal controls:** Agents process only pre-meeting docs; prompts anchored to date

▷ **Automated audit:** Scans for forbidden terms ("looking back"); rejects violations

▷ **Out-of-sample test:** If cheating (look-ahead bias!) → out-of-sample var. ≫ in-sample; gap ≈3bp (negligible)

▷ **J&K validation:** Loads on MP shocks, not central bank information

▷ **Multi-run stability:** 17 runs → 4bp std dev (vs ±25–100bp typical surprises)

Contributions

▷ **Method:** First multi-agent LLM for MP shock identification

▷ **Measurement:** Priors unbiased (MZ: α=0, β=1); surprises well-calibrated

▷ **Empirics:** No price puzzle; curve flattens via expectations channel

▷ **Practice:** Real-time from public docs (no 5-year Greenbook lag)

▷ **Trading:** Narrative surprise outperforms best market measure by **92%**