A Large-Scale LLM Analysis of Central Bank Communication

Ole Gregarek
IMF – Monetary and Capital Markets
Central Bank Operations
ogregarek@imf.org

with Thiago Christiano Silva, Kei Moriya, Romain Michel Veyrune

13th ECB Conference on Forecasting Techniques
European Central Bank
March 23-24, 2026

Motivation and Contribution

1

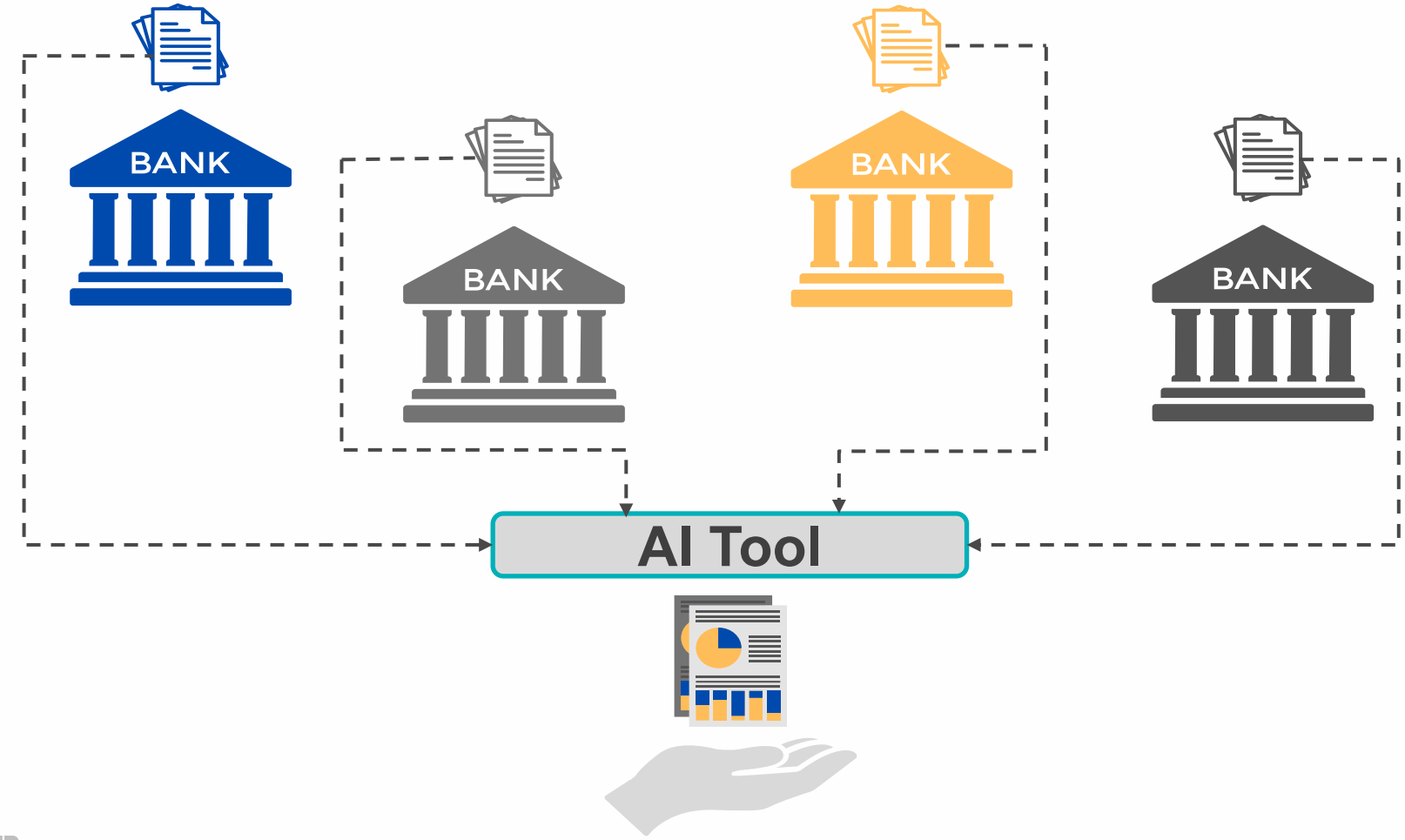
Measuring communication

WHY
Communication matters

CHALLENGE
Text is complex

WHAT
AI quantifies messages

RELEVANCE
Insight to action



Contribution

- An automated tool to classify central bank communications at the sentence level by:
  - Topic
  - Communication stance
  - Audience
  - Policy sentiment
- AI tool is applied to an enormous dataset of central bank communications with 171 central banks
  - 78k+ documents, 25M+ sentences
  - This empirical exercise offers unparalleled scope for global central bank communication analysis
- The classifier transforms qualitative central bank communications into measurable quantitative insights, enabling systematic analysis of policy signals, content, and styles with metrics

Use Cases of the AI Classification Tool

1.- Enhancing policy communications

Improve clarity and transparency of monetary policy and financial stability communications

2.- Benchmarking and peer comparison

Comparative analysis and regional trends

3.- IMF use cases

Central bank communication technical assistance missions + Central Bank Transparency Code Missions

Data and Exploratory Analysis

2

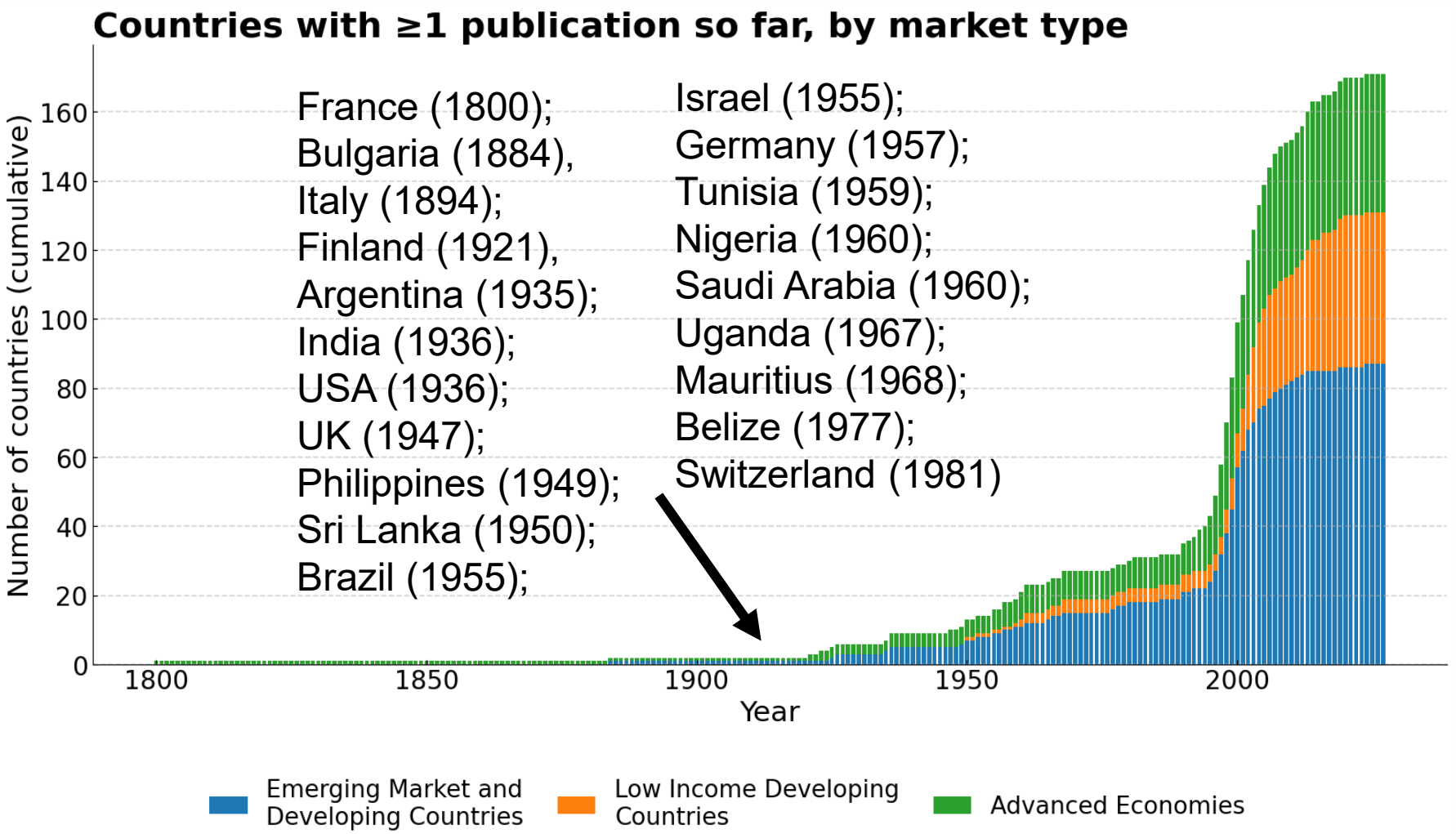
Data

1800 to 2026

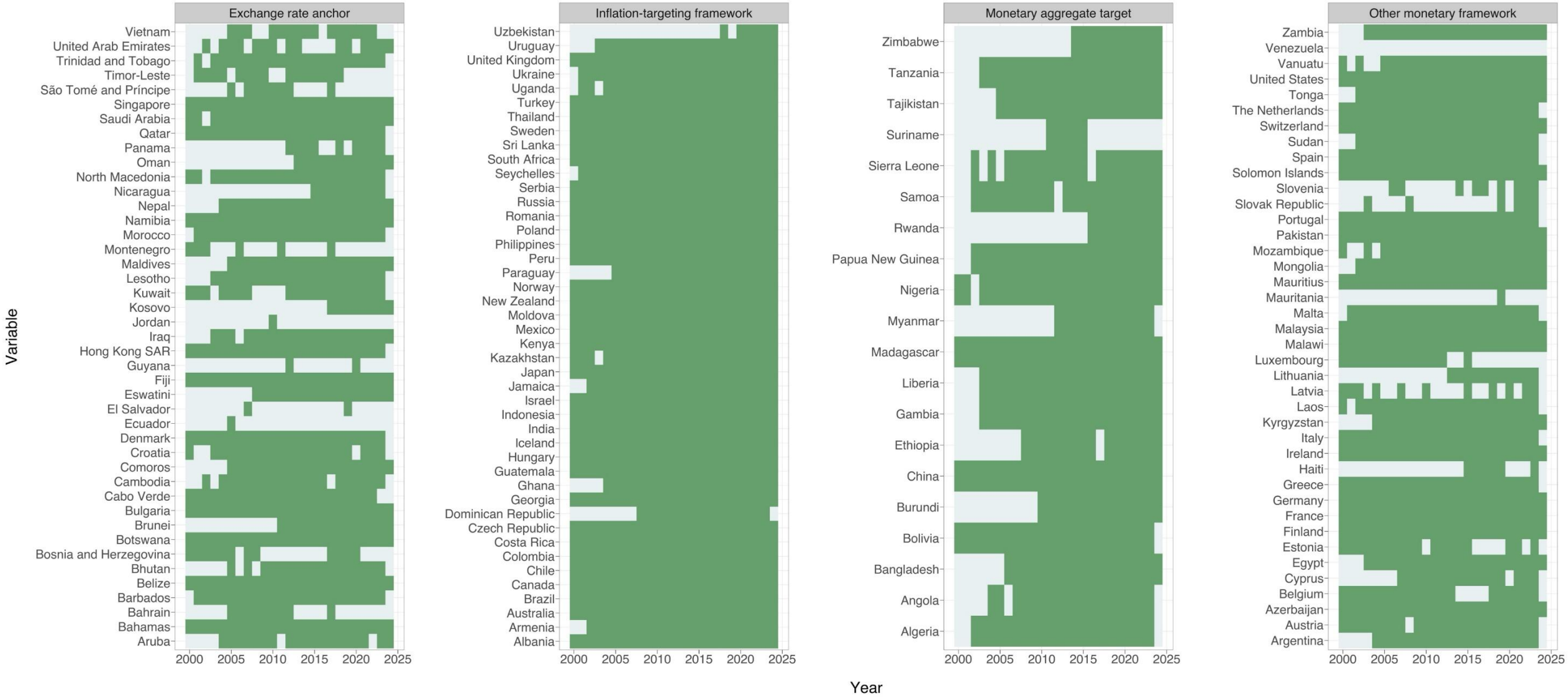
171 Central Banks

78K+ Documents
25M+ Sentences

Multilingual

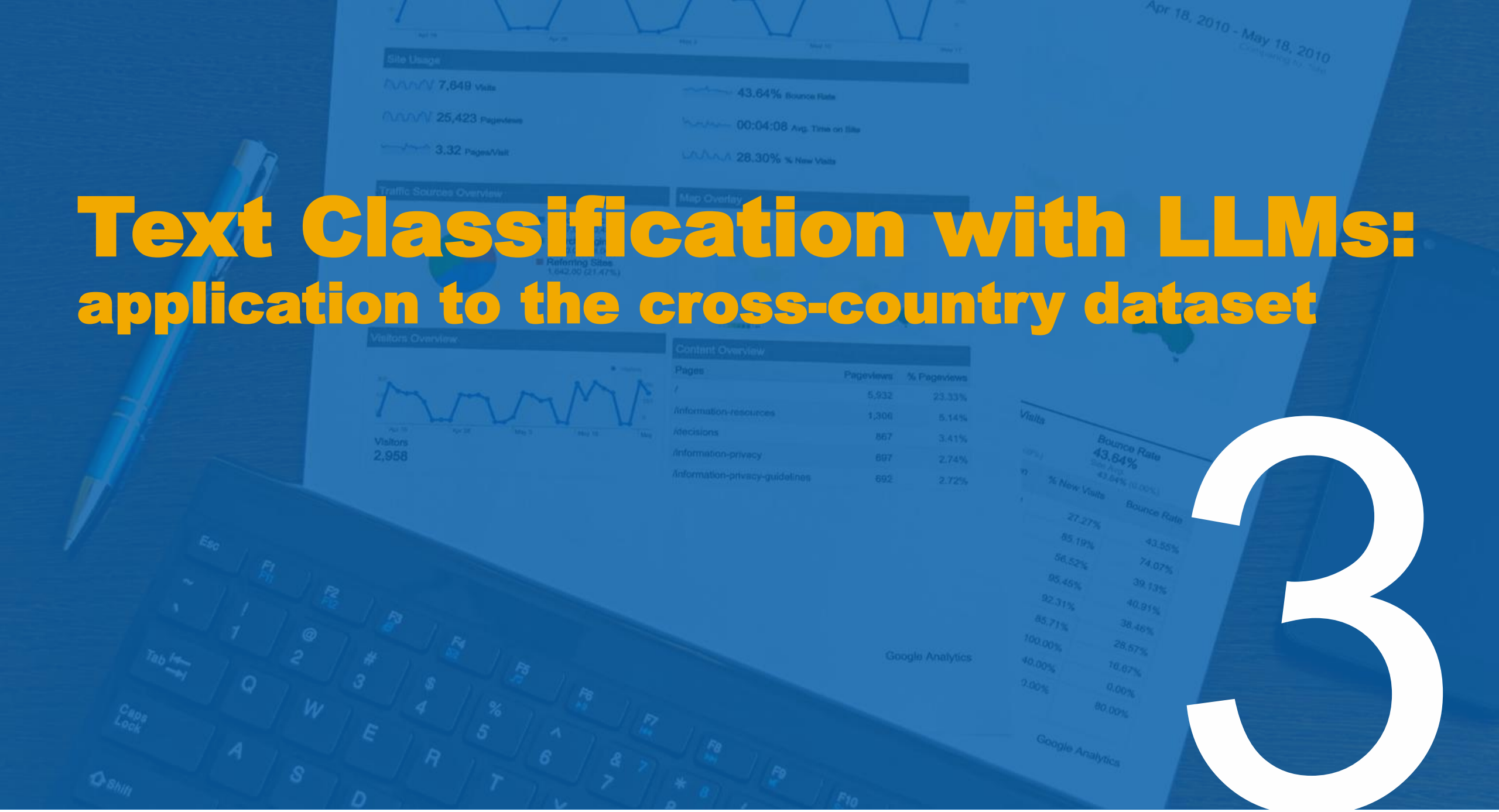


Data coverage by central bank (since 2000) - regular publications



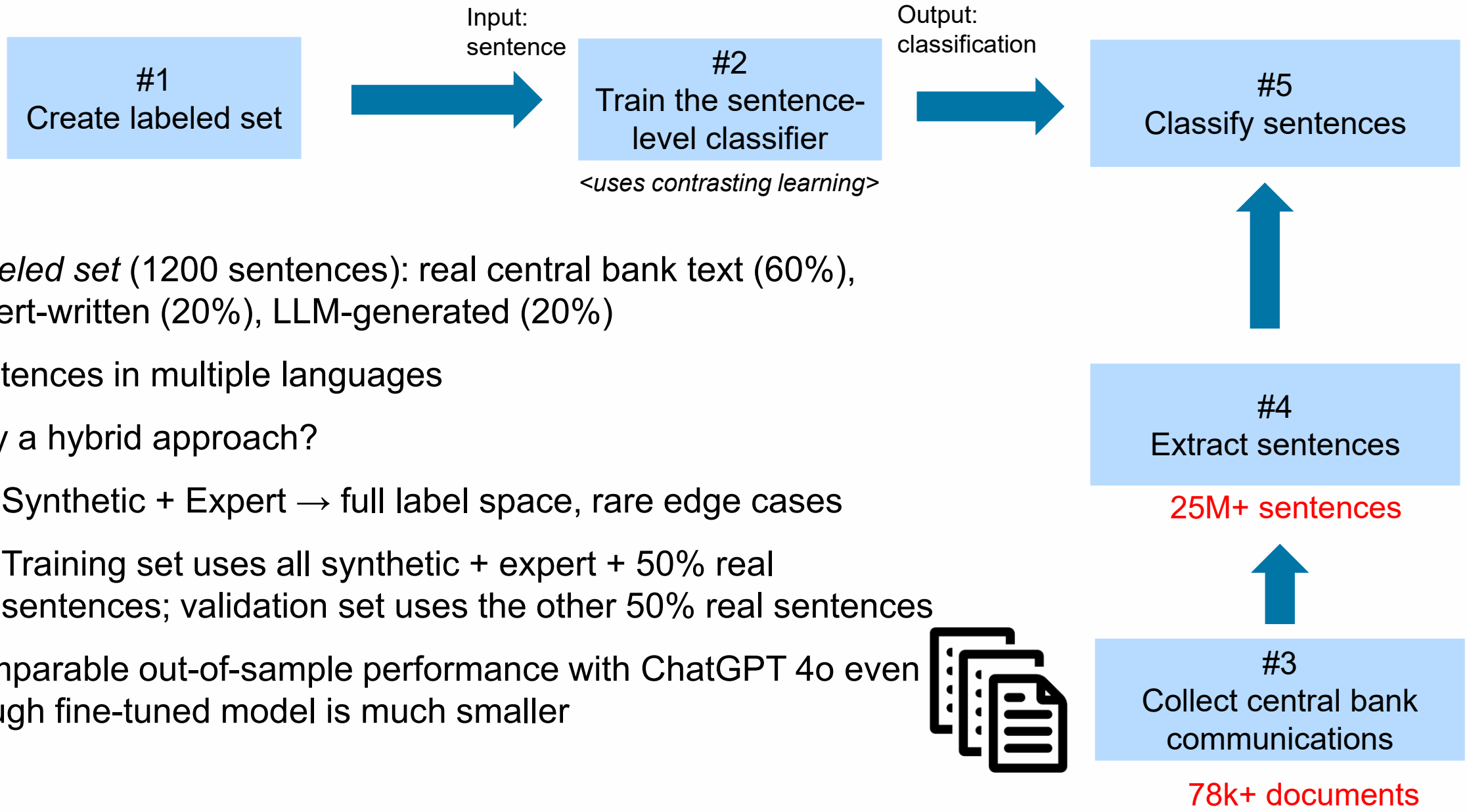
Note: The dataset includes ECB, ECCB, BEAC, and BCEAO. A cell is green if at least one publication is present.

Text Classification with LLMs:
application to the cross-country dataset



3

Schematic: Text Classification of Central Bank Communication Using LLMs



- *Labeled set* (1200 sentences): real central bank text (60%), expert-written (20%), LLM-generated (20%)
- Sentences in multiple languages
- Why a hybrid approach?
  - Synthetic + Expert → full label space, rare edge cases
  - Training set uses all synthetic + expert + 50% real sentences; validation set uses the other 50% real sentences
- Comparable out-of-sample performance with ChatGPT 4o even though fine-tuned model is much smaller

Semantic analysis: schematic of the text classifier

Lower inflation will bring the necessary predictability and stability to economic activity, and the banking system is prepared to provide financing for companies.

⋮

Due to emerging economic indicators and potential market volatility, it is crucial for the financial sector to proactively assess risks to ensure long-term financial stability.

Sentence-level classifier

Finetuned Multilingual LLM

(official translations in English if available; otherwise, native language)

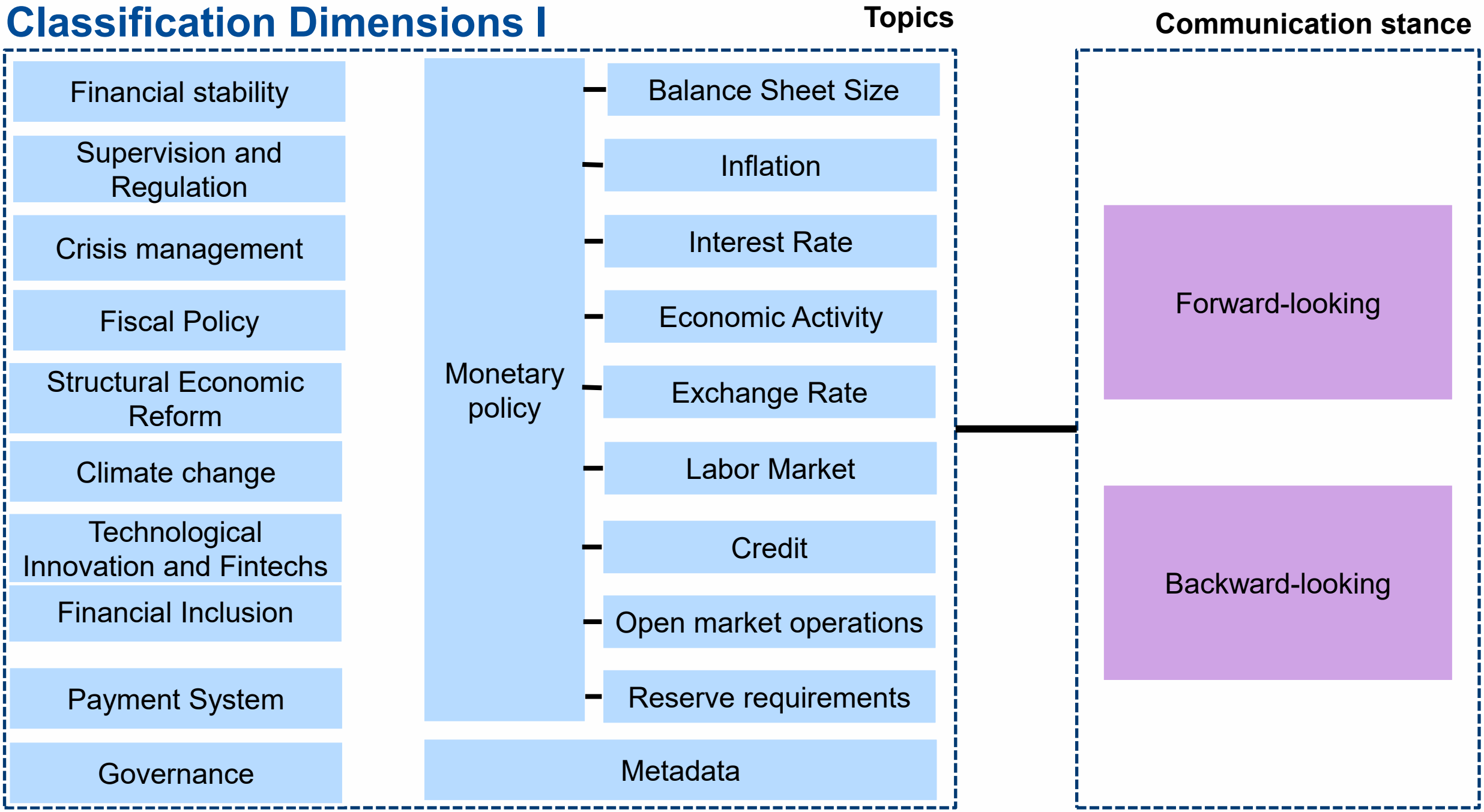
- ***Topic***: *Monetary policy – inflation*
- ***Communication stance***: *Forward-looking*
- ***Audience***: *Financial sector*
- ***Sentiment***: *Dovish*

⋮

- **Topic**: Financial stability
- **Communication stance**: Forward-looking
- **Audience**: Financial sector
- **Sentiment**: Risk-highlighting

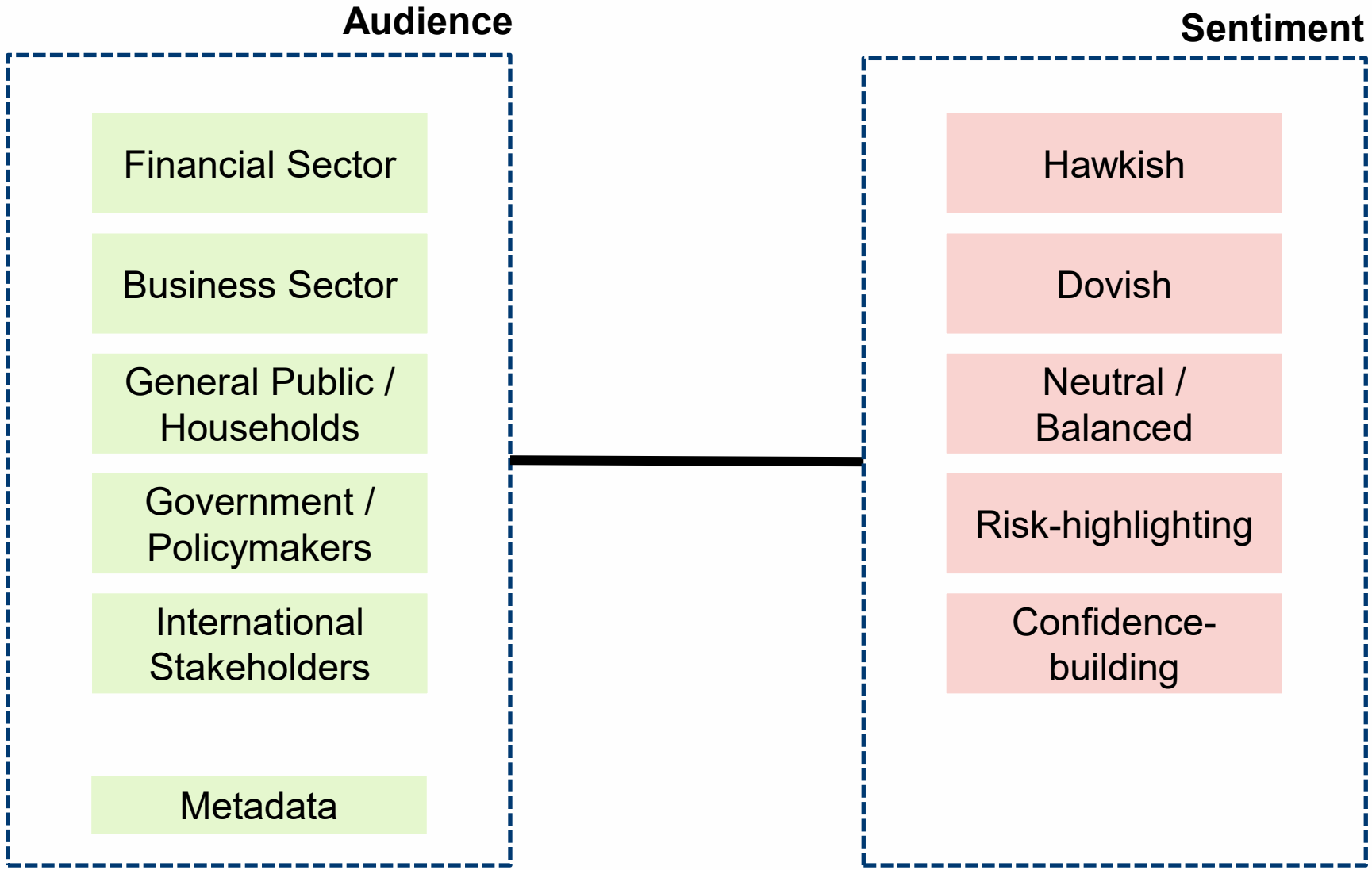
Sentences extracted from CB communications
$\approx 78k$ documents (25 million sentences)

Classification Dimensions I



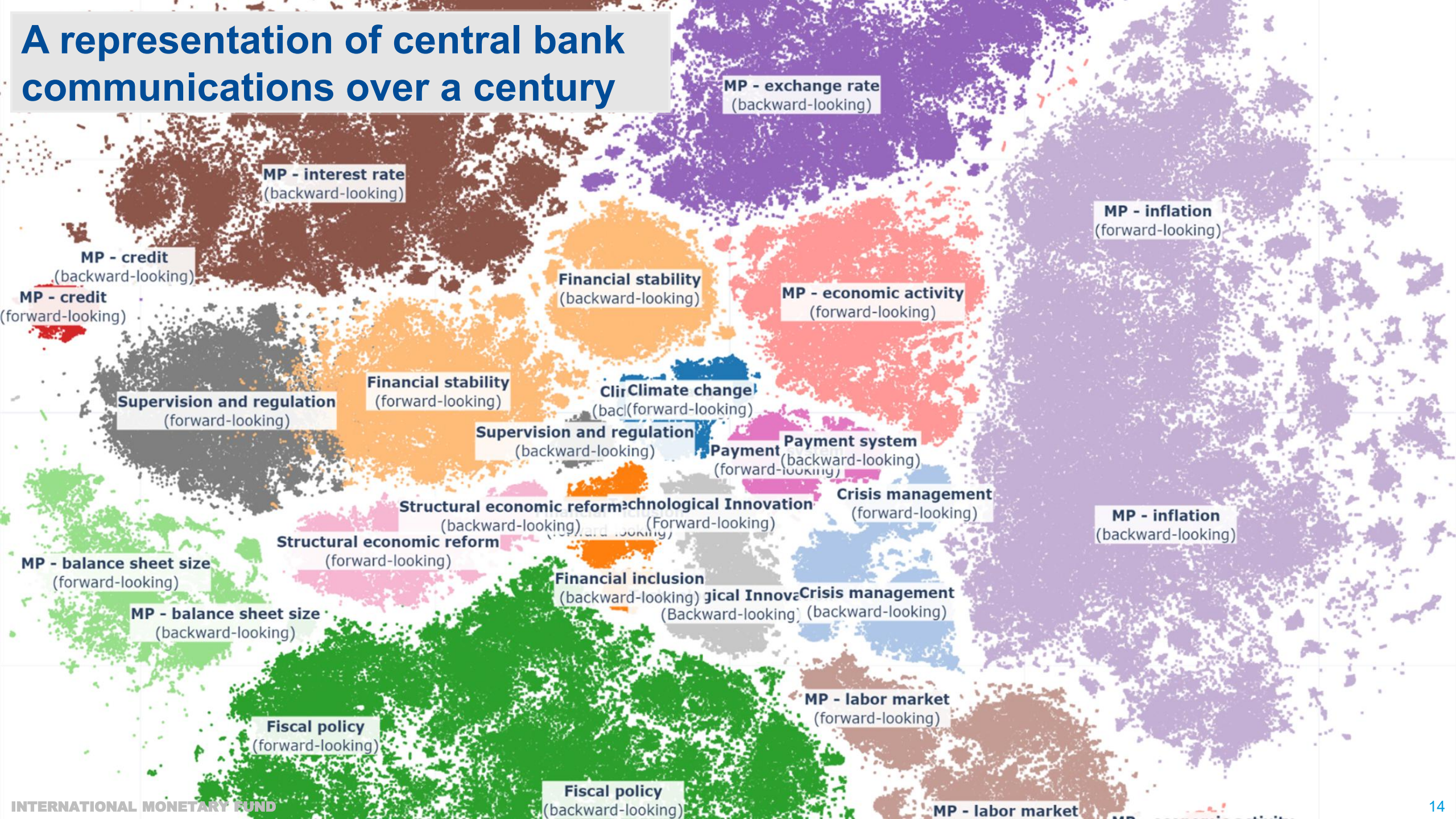
OBS: The “metadata” class contains sentences with no economic information. 6.3% of all the sentences in the dataset are classified as “metadata”.

Classification Dimensions II



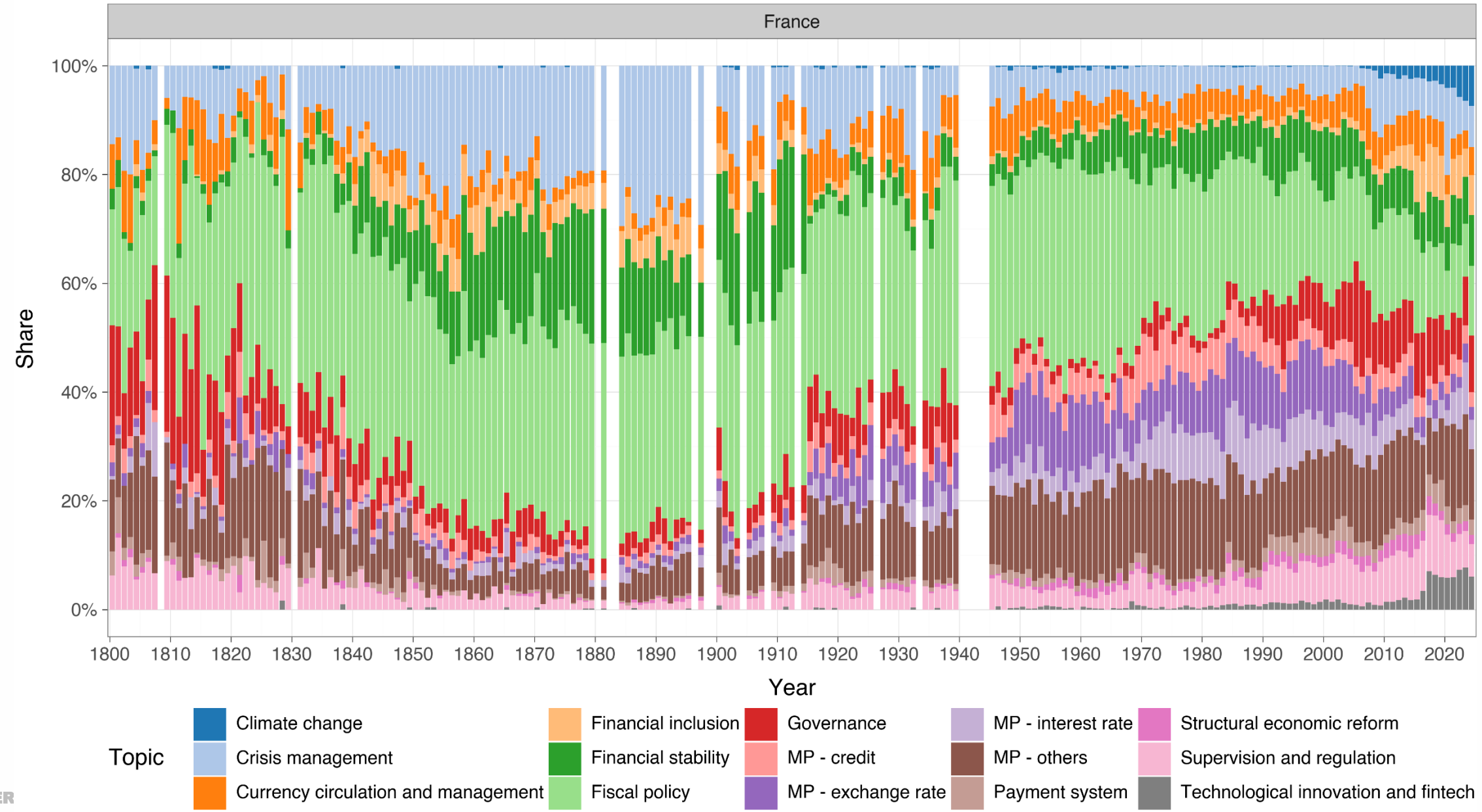
OBS: The "metadata" class contains sentences that convey no economic information. 14.1% of all the sentences in the dataset are "metadata".

A representation of central bank communications over a century

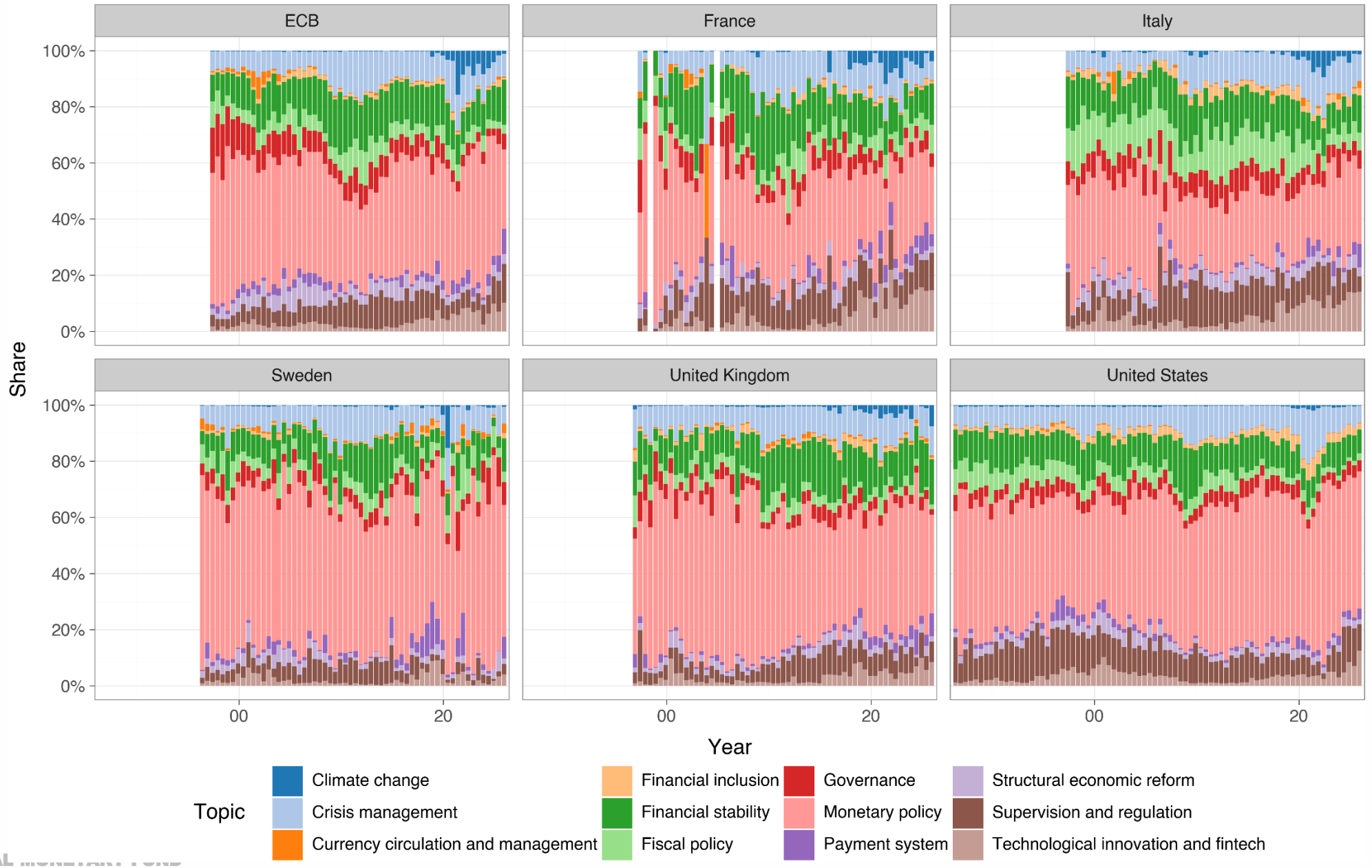


200+ years of Banque de France's communication in a standardized format

Only in French language until 2020; then in English (and French)



Speeches topics are heterogeneous across countries



Global Trends: Monetary Policy Topics

By level of development

Topics
Comm. Stance

Advanced economies

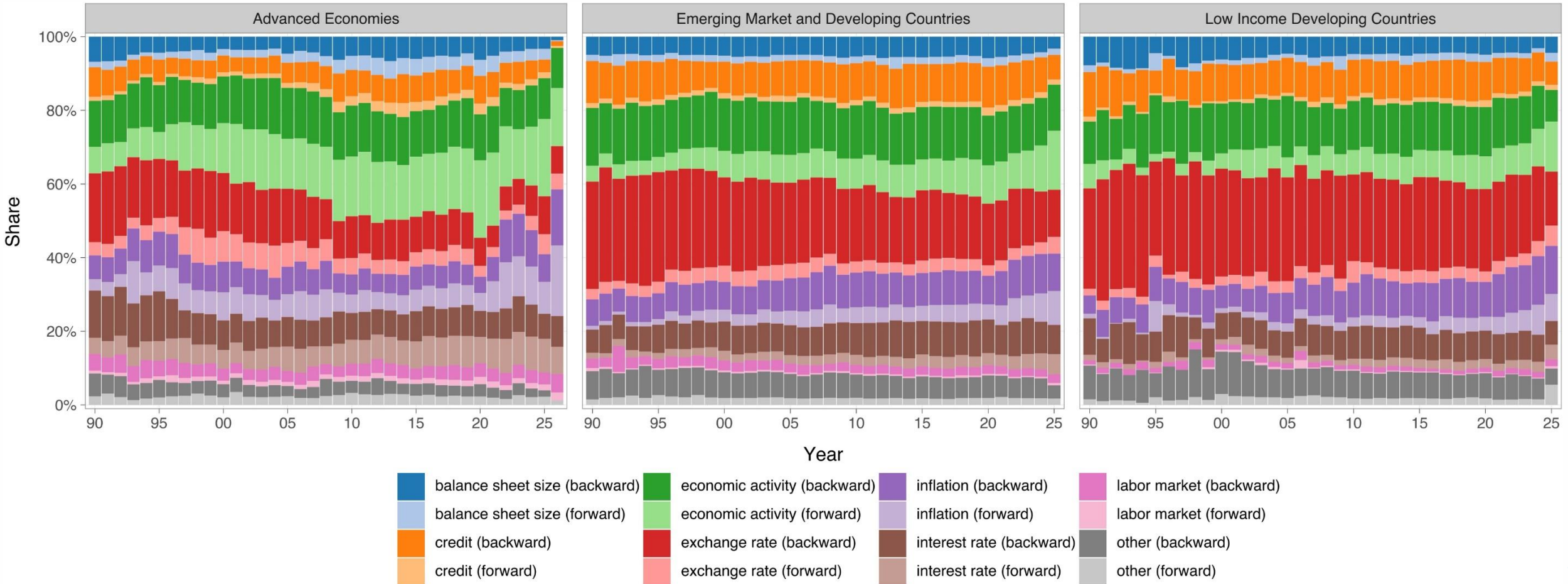
- Higher emphasis on **interest rates** than **inflation**
- More forward-looking sentences
- **Exchange rate** statements decrease

Emerging market economies

- Higher emphasis on **inflation** than **interest rates**
- **Exchange rate** statements decrease

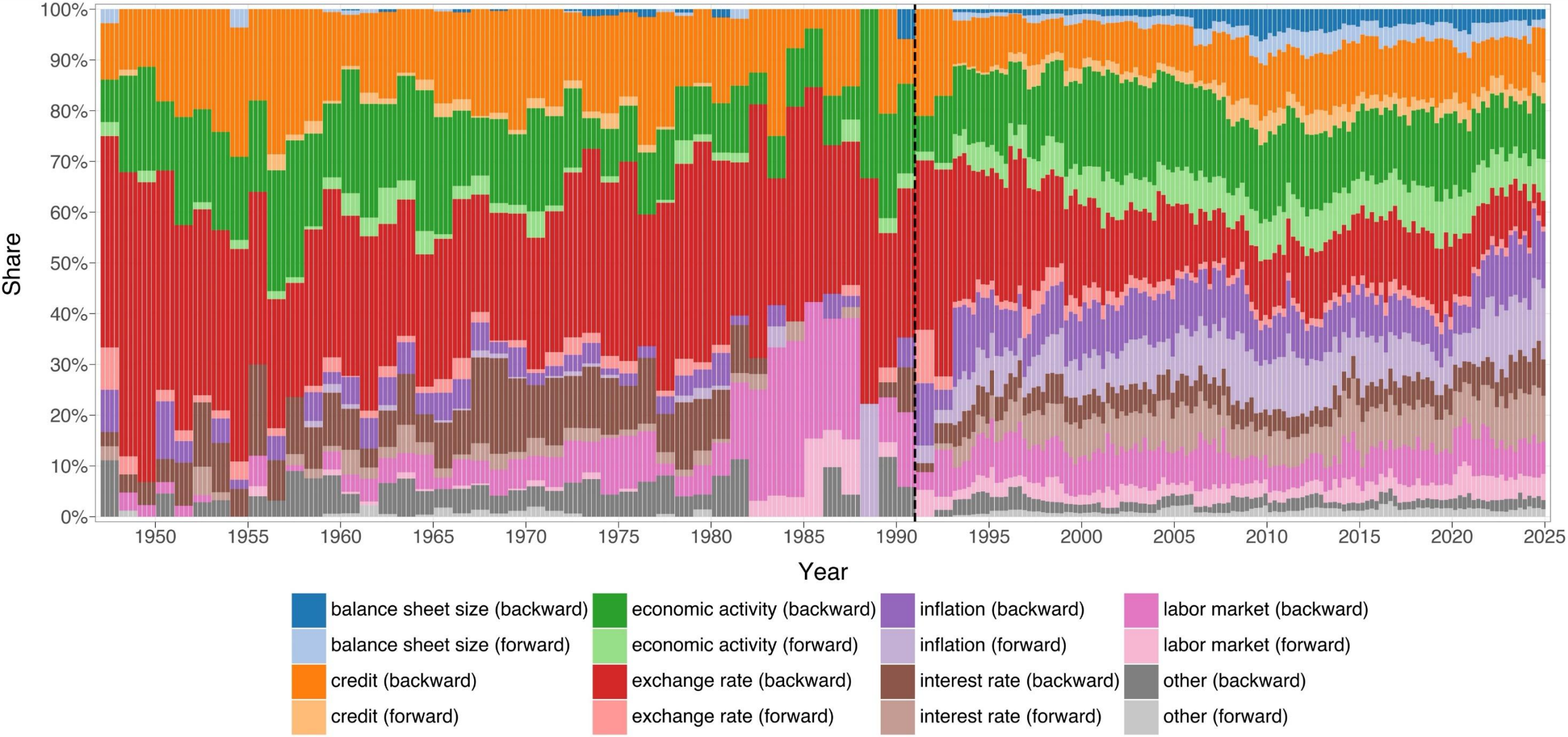
Low-income economies

- Higher emphasis on **inflation** than **interest rates**
- **Exchange rate** statements decrease but are comparatively more relevant

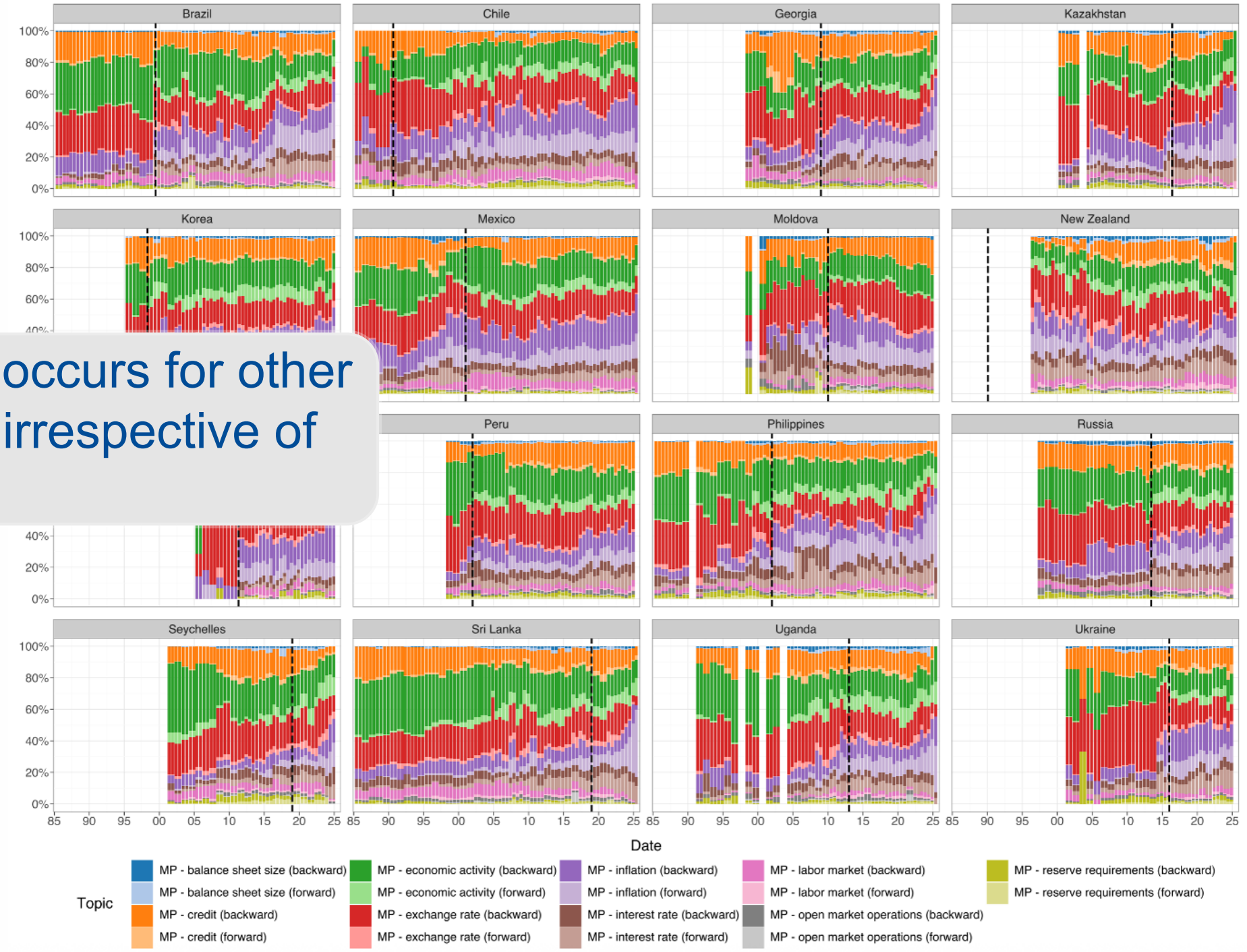


Inflation-targeting adoption deeply changes monetary policy communication

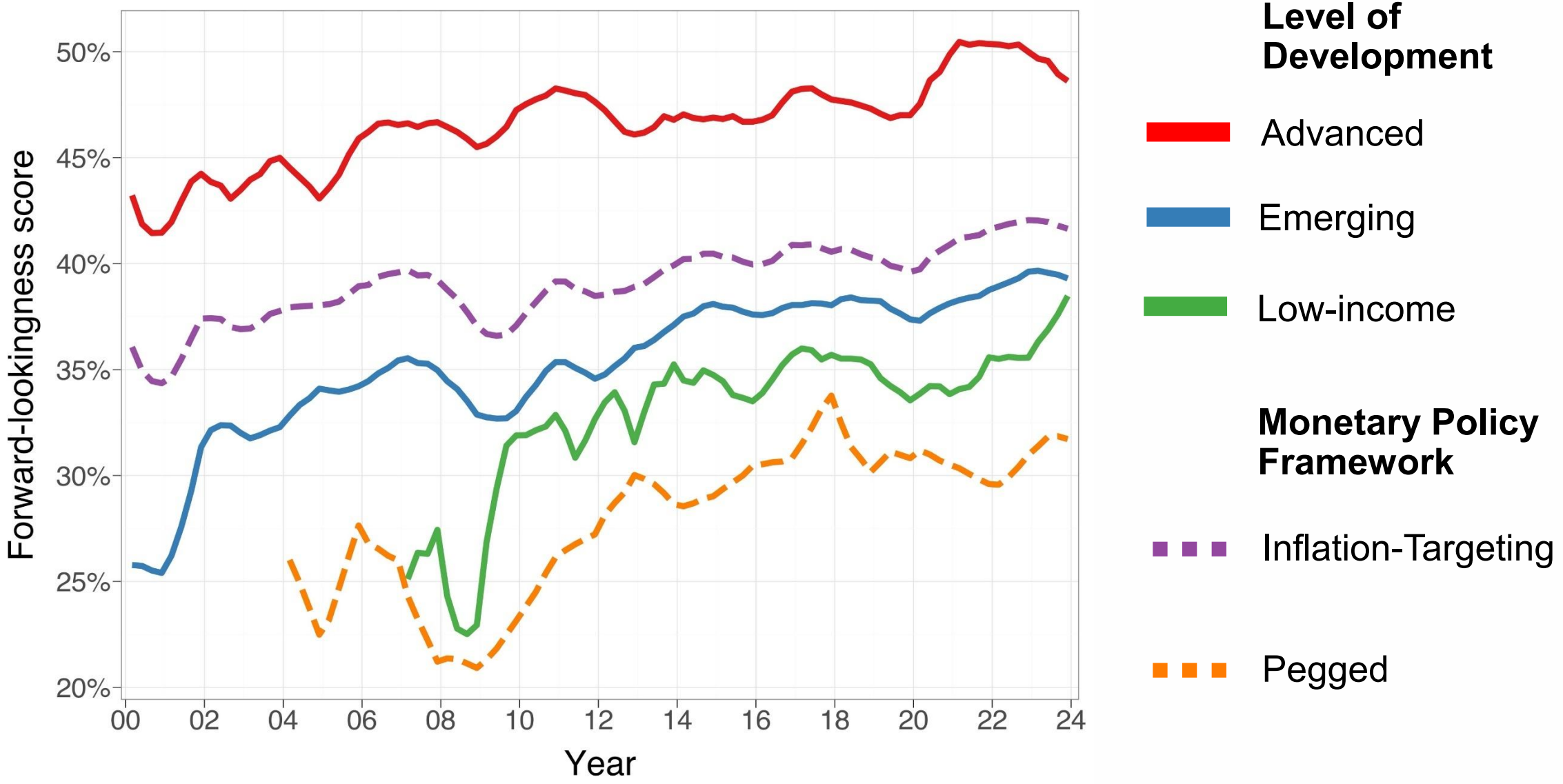
UK Monetary Policy Communication



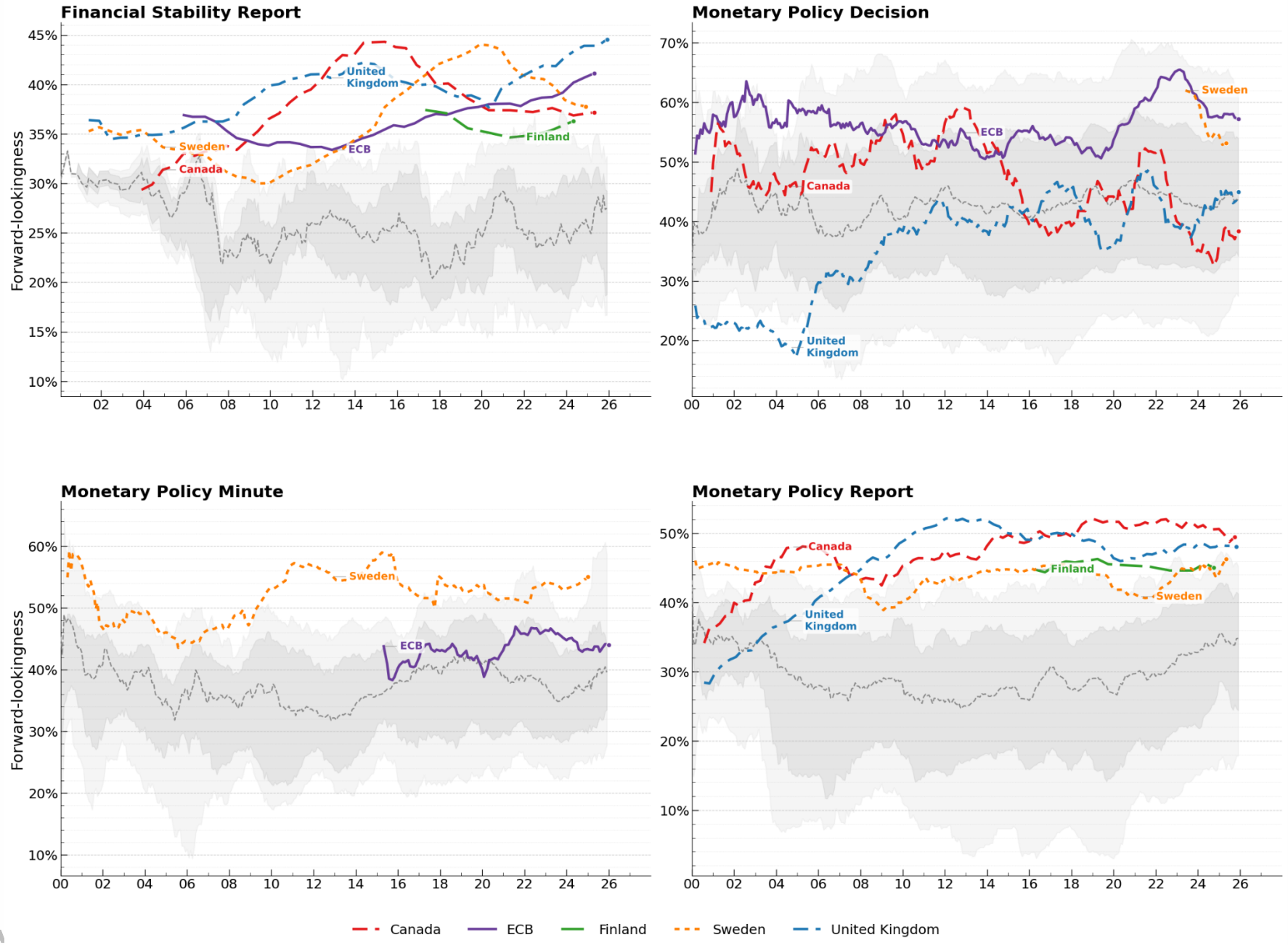
The same change occurs for other inflation-targeters, irrespective of development level



Communication is becoming more forward-looking worldwide



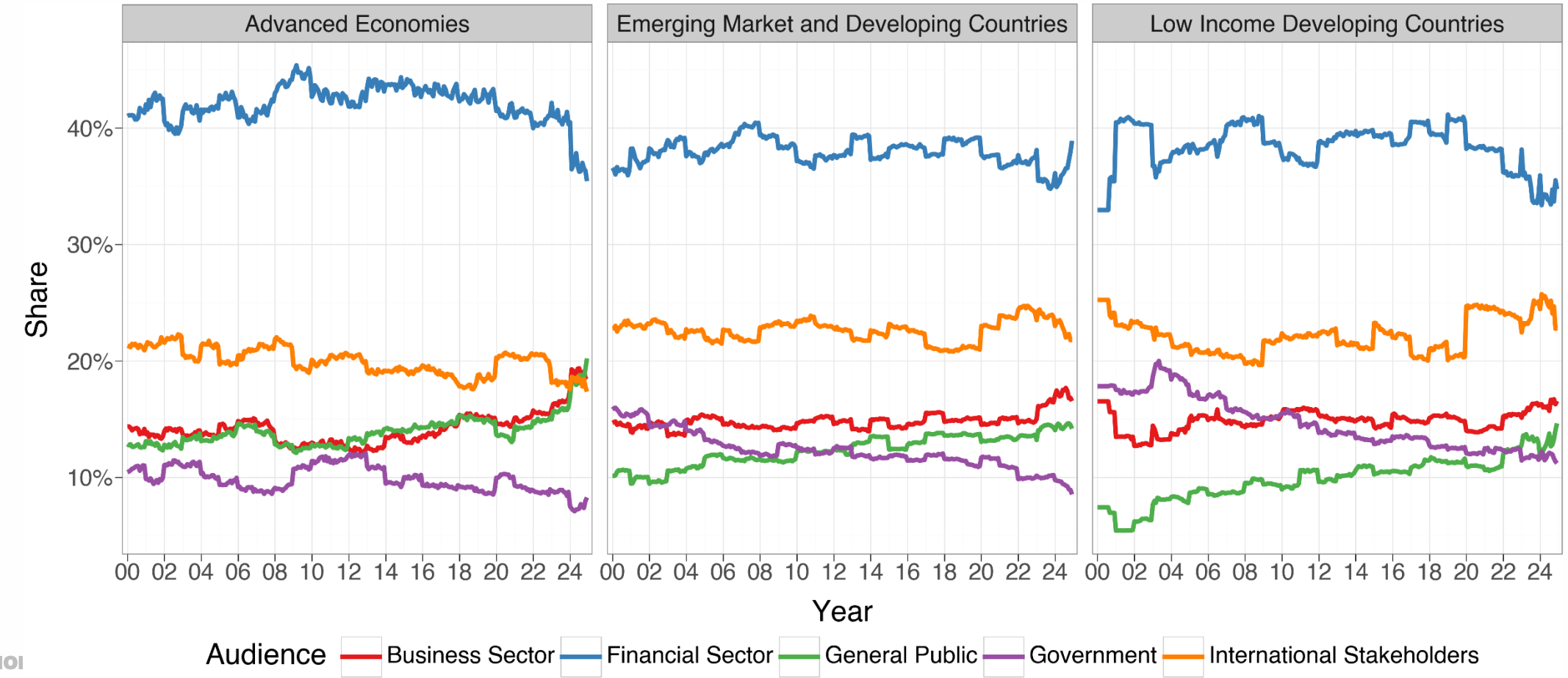
High dispersion in how forward-looking documents are



Global Trends: Audience in Central Bank Communications

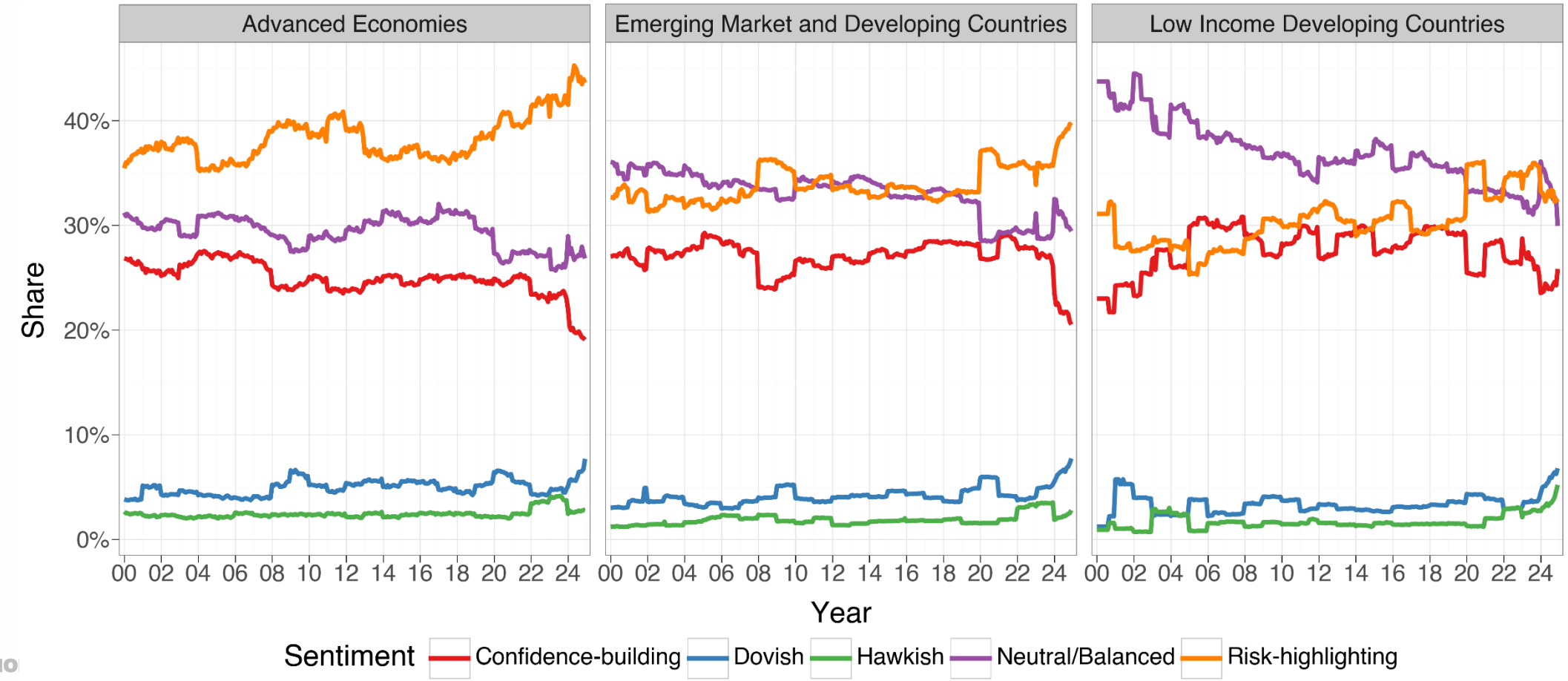
By level of development

- Though decreasing, the main recipient is the **financial sector,** especially in advanced economies
- Attention to the **government** is negatively correlated with the economy's level of development
- Increased inclusiveness of the **general public** in central bank discourse
- Attention to **general public** and **business** is similar in advanced economies, but a gap exists in others



Global Trends: Sentiment in Central Bank Communications
By level of development

- Attention to **risk highlighting** is proportional to the economy's level of development
- **Neutral/balanced** text is inversely proportional to the economy's level of development
  - Neutral text limits to fact-based statements, typically without economic rationale and explanations



Metrics from Central Bank Communication

4

Metric: net policy sentiment

$$\text{Net Policy Sentiment} = \frac{H - D}{\text{H} + \text{D}}$$

Symbol	Definition
H	Hawkish (tightening signal)
D	Dovish (easing signal)

Interpretation:

- Captures the balance between *tightening* (hawkish) and *easing* (dovish) communication signals based on **monetary policy decisions only**

Two dimensions

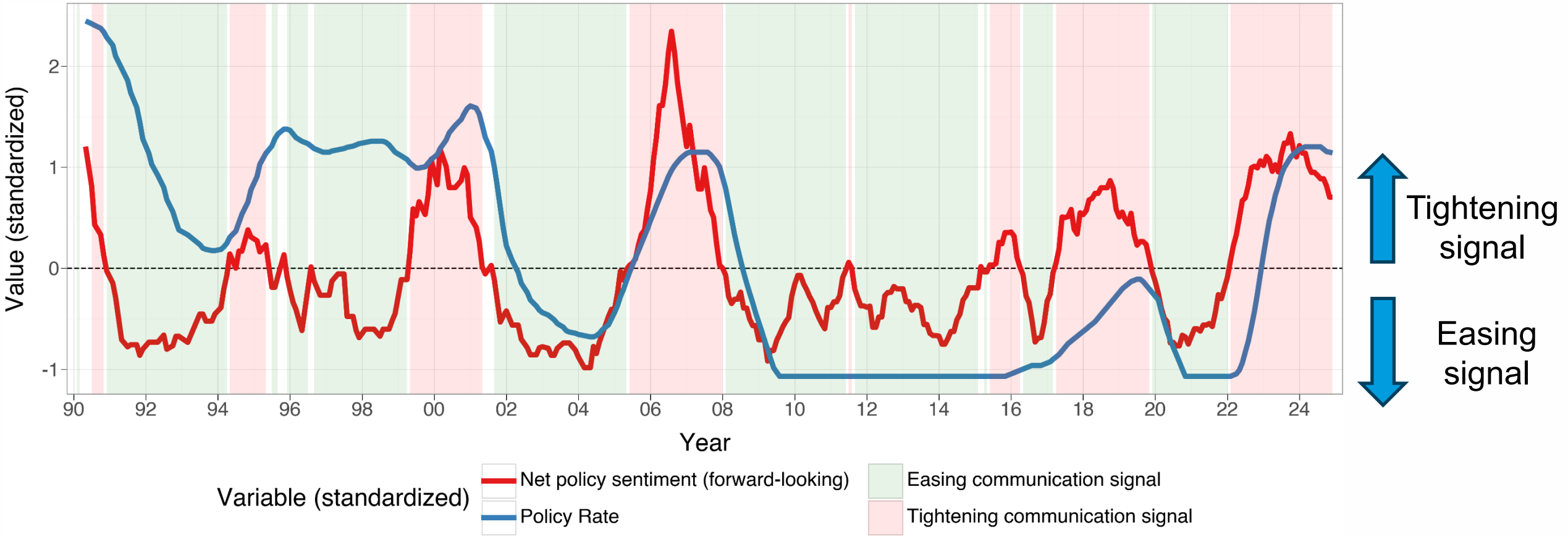
- *Forward-looking net sentiment*: Reflects future-oriented signals, including forward guidance
- *Backward-looking net sentiment*: Reflects past- or present-oriented signals, often linked to policy justification

Why is this metric relevant?

- Enables comparing the central bank's communication signals with the actual monetary policy stance

US net policy sentiment vs. policy rate

On monetary policy decisions



- Net policy sentiment (forward) leads the policy rate often
- Net policy sentiment (forward) embeds more than policy rate changes: it contains elements of forward guidance, QE (e.g., in post-GFC)

Use: check the alignment of communication signals with the intended monetary policy stance

Forward-looking net policy sentiment has informational value for future changes in policy rates and market interest rates

	Δ Policy Rate$_{i,t+1}$			Δ T-Bill Rate$_{i,t+1}$			Δ T-Bond Rate$_{i,t+1}$		
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
Net Policy Sentiment$_{i,t}$									
Total	0.101***			0.022			0.102*		
	(0.022)			(0.021)			(0.057)		
Forward-looking		0.083***	0.162***		0.055***	0.051***		0.105*	0.135*
		(0.017)	(0.033)		(0.020)	(0.012)		(0.058)	(0.070)
Backward-looking		0.062***			-0.025			0.024	
		(0.014)			(0.019)			(0.015)	
Gap (Fwd - Bwd)			-0.096***			0.038			-0.036
			(0.022)			(0.029)			(0.024)
Straightforwardness Index$_{i,t}$	-0.010	0.006	0.006	0.021	0.023	0.023	-0.022	-0.006	-0.006
	(0.011)	(0.009)	(0.009)	(0.021)	(0.018)	(0.018)	(0.028)	(0.019)	(0.019)
Explanation Index$_{i,t}$	-0.003	-0.006	-0.006	-0.010	-0.013	-0.013	-0.016	-0.024	-0.024
	(0.013)	(0.013)	(0.013)	(0.022)	(0.021)	(0.021)	(0.021)	(0.018)	(0.018)
Net Confidence Index$_{i,t}$	-0.008	-0.008	-0.008	-0.012	-0.013	-0.013	-0.012	-0.013	-0.013
	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(0.015)	(0.021)	(0.021)	(0.021)
Exchange Rate (USD/local)$_{i,t}$	-0.041	-0.041	-0.041	0.052	0.049	0.049	-0.171	-0.171	-0.171
	(0.101)	(0.106)	(0.106)	(0.064)	(0.064)	(0.064)	(0.102)	(0.106)	(0.106)
Inflation (CPI)$_{i,t}$	1.124**	1.128**	1.128**	0.063	0.066	0.066	0.336	0.333	0.333
	(0.452)	(0.451)	(0.451)	(0.183)	(0.181)	(0.181)	(0.505)	(0.508)	(0.508)
Country Fixed Effects	x	x	x	x	x	x	x	x	x
Time Fixed Effects	x	x	x	x	x	x	x	x	x
Additional Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5318	5318	5318	4456	4456	4456	3040	3040	3040
R^2	0.168	0.173	0.173	0.105	0.108	0.108	0.145	0.148	0.148

- Data: panel data with 160 economies (this is not causality)
- Bottomline: the results point to the importance of central bank communication, showing that it covaries with future market prices and financial conditions

Metric: net confidence index

$$\text{Net Confidence Index} = \frac{C - R}{\text{C} + \text{R}}$$

Symbol	Definition
C	Confidence-building
R	Risk-highlighting

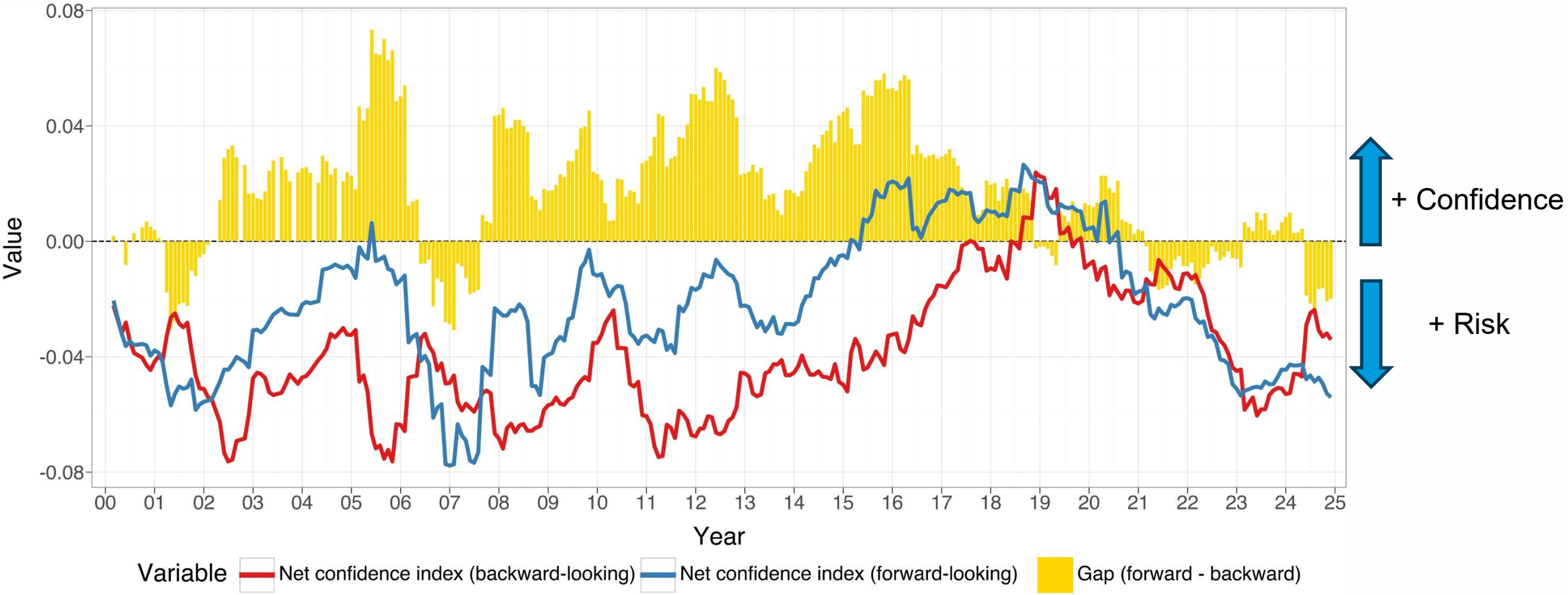
- The **net confidence index** measures central bank's risk communication (ranges from -1 to 1)
  - Higher values: optimistic tone, signaling confidence
  - Lower values: cautious tone, signaling risks and concerns

- **Forward vs. Backward-looking**
  - Forward-looking: Expectations about future economic or financial risks.
  - Backward-looking: Assessment of current/past conditions.

- **Gap (forward − backward components)**
  - A positive gap (forward > backward) signals increasing optimism about future conditions compared to the present—a possible indicator of improving outlook
  - A negative gap (forward < backward) indicates rising future concerns compared to current conditions—a potential early signal of emerging risks or instability

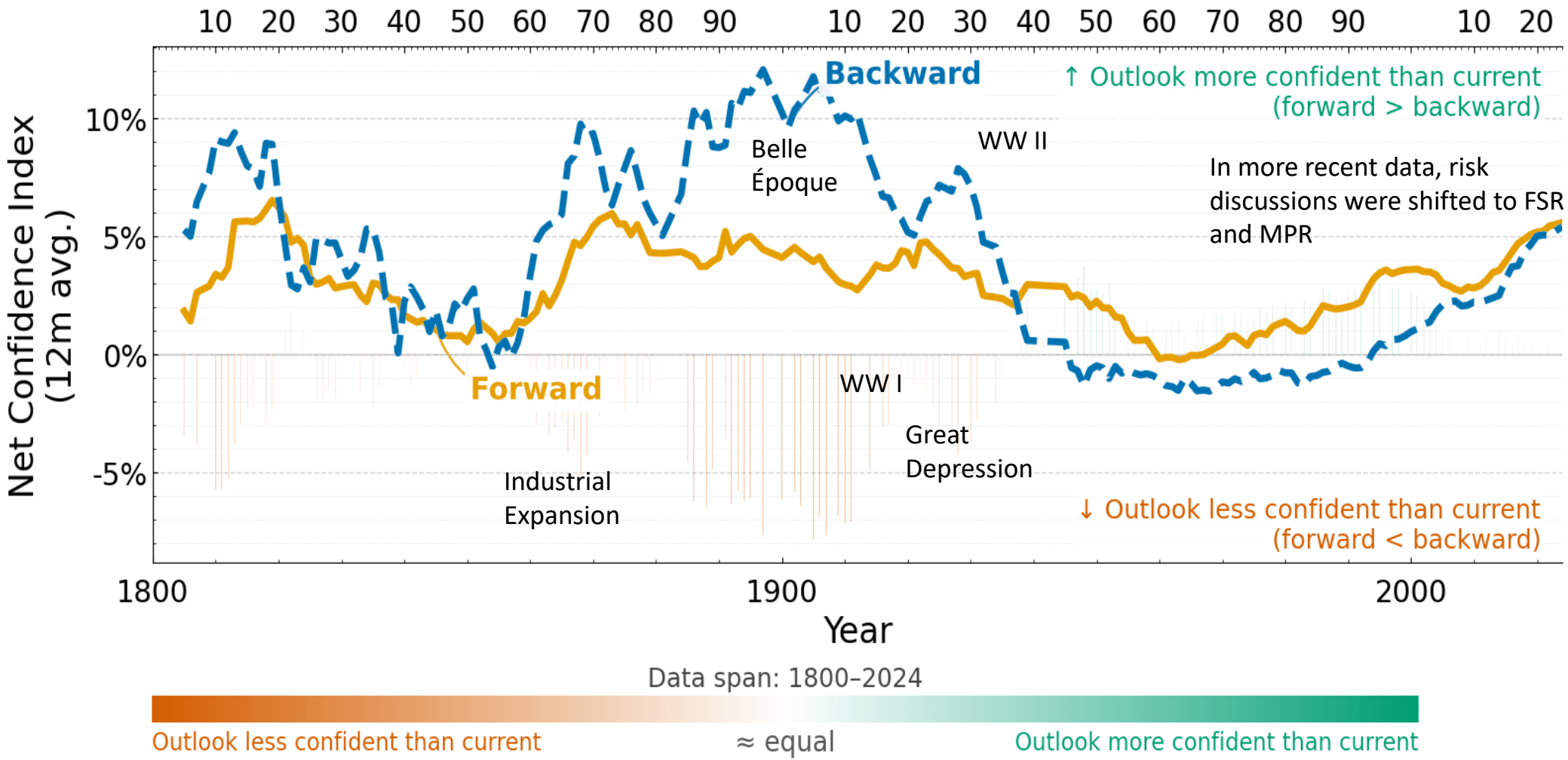
US net confidence index

Risk communication patterns (on all regular documents)



- Net confidence scenario captures the risk scenario in the outlook
- Forward-looking metric leads the backward-looking metric in many episodes
- The US risk communication strategy seems to build confidence going forward while highlighting relatively more risks when looking at the past and current conditions

Banque de France's risk communication over 200 years

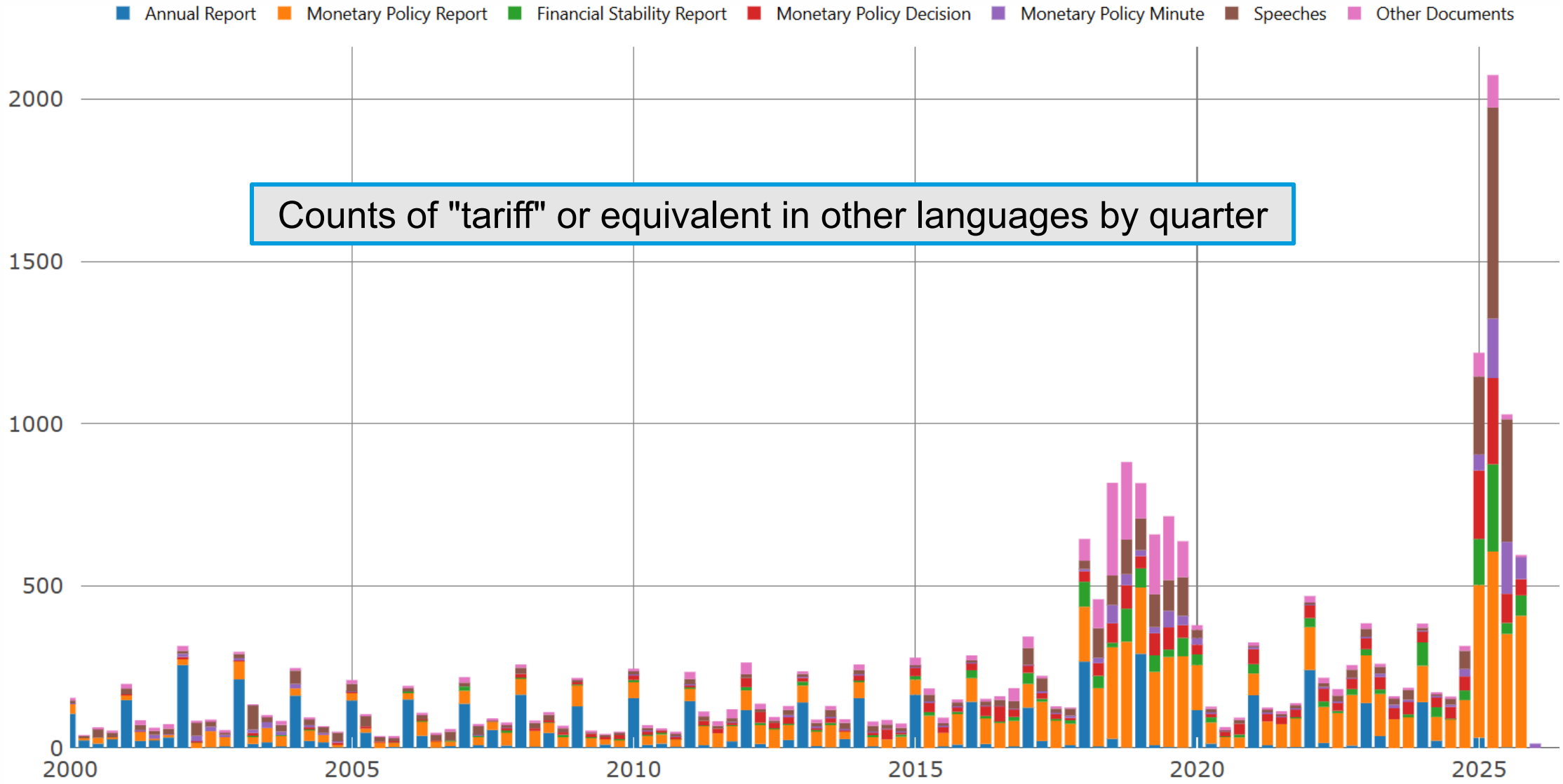


Examples of downstream tasks

5

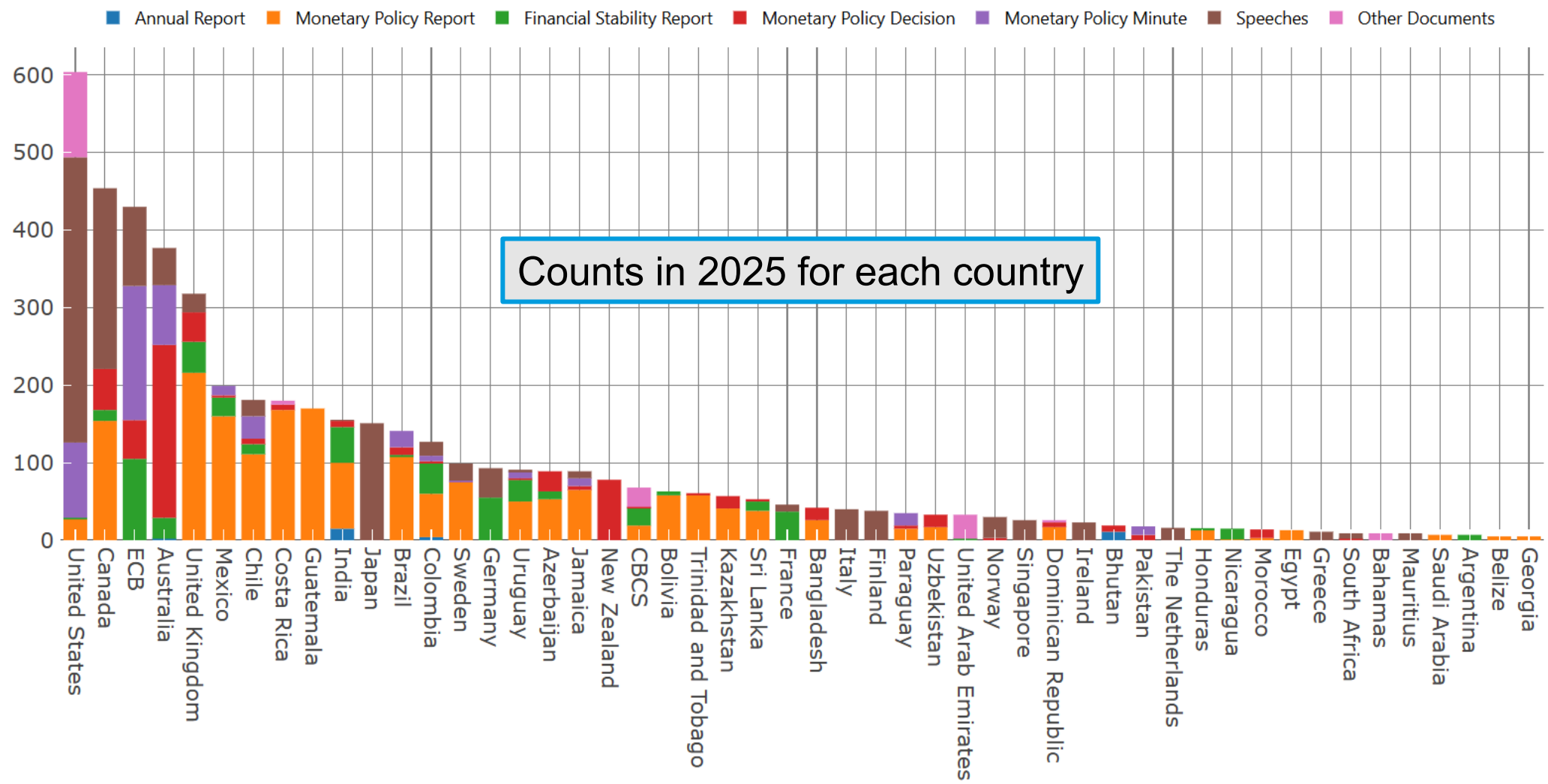
Tariffs: Who is Talking?

- Our data allows sifting through 200+ years of data for specific keywords



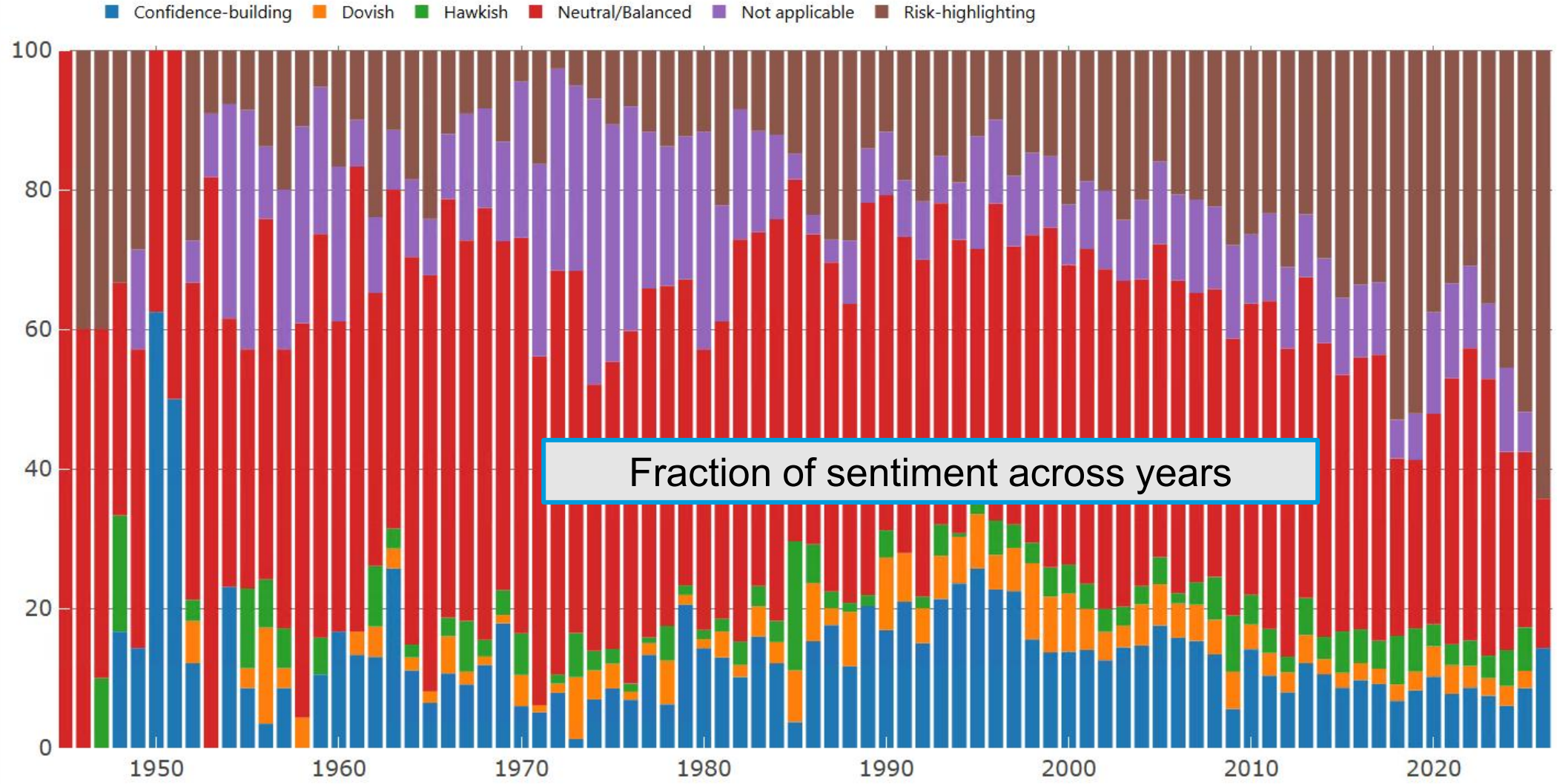
Tariffs – Which Countries, Which Medium?

- US mentions tariffs mostly in **speeches** but not in decisions
- ECB mentions in **MP Minutes** and **FSR**, Australia, New Zealand quote in **MP Decisions**



Tariffs – What is the Sentiment?

- Check how dimension "sentiment" is classified for all sentences containing "tariff" (or equivalent)
- Traditionally has been mostly neutral/balanced, **recently more risk-highlighting**



Key takeaways, uses, and opportunities

Key takeaway: No one-fits-all communication metric—each measure captures a distinct dimension and serves a different purpose

- **Same stance, different purpose:** two MPDs can be equally *hawkish/dovish* (**Net Policy Sentiment**), yet one is **forward-looking** while the other is backward-looking—markets react differently even if the stance is the same.

- **Same topic, different signal:** a rise in "financial stability" coverage can mean **risk warnings** (**Risk-highlighting**) *or* **reassurance/toolkit readiness** (**Confidence-building**)—same theme, opposite communication intent.

Current use: IMF using in (i) communication technical assistance & (ii) Central Bank Transparency Code

Provides quantitative metrics on the recipient country, typically benchmarked against group of peers

- **Web application** developed and provided to member countries in technical assistance missions

Future work: Methodology is general: applying the same idea with focus on financial stability

A Large-Scale LLM Analysis of Central Bank Communication

Ole Gregarek
IMF – Monetary and Capital Markets
Central Bank Operations
ogregarek@imf.org

with Thiago Christiano Silva, Kei Moriya, Romain Michel Veyrune

13th ECB Conference on Forecasting Techniques
European Central Bank
March 23-24, 2026

Design considerations: selection of the LLM

Decoder-only models (e.g., GPT)

- Autoregressive: optimized for generation, not classification
- Weak supervision via prompt engineering is required for label control
- Opaque label boundaries; hard to audit
- High inference cost (token-dependent pricing)

Encoder-only models (e.g., sentence transformers)

- Bidirectional: optimized for classification & similarity
- Direct sentence embeddings from fine-tuning
- Relatively more transparent (you have probabilities)
- Cost-efficient at scale (no output token costs)

Bottomline:

- Large generative LLMs (e.g., GPT) are versatile but **cost-prohibitive at scale** for classification tasks
- **Small, fine-tuned models** can match or exceed performance when applied to **well-defined, domain-specific tasks**
- We use a **compact, open-source, multilingual sentence transformer** (bge-m3) tailored for economic text classification

Design considerations: training and validation sets

Labeled Dataset Construction

Total: 1,200 sentences

Split by source and purpose:

Synthetic (240 examples, 20%)

•Generated with LLM prompts to ensure semantic coverage & diversity. Expert-verified.

Expert-constructed (240, 20%)

•Built by domain experts for precision and policy relevance.

Real central bank text (720, 60%)

•Stratified sample, multiple languages → labeled by experts to reflect authentic communication.

Train–Validation Split

•**Training Set (840)**
• Synthetic + Expert + 50% Real
•**Validation Set (360)**
• Only remaining 50% Real

Key: validation should be well-designed so that we can iterate and have directions on whether we are improving!

Why Hybrid Design?

•Purely real data = expensive, unbalanced, and incomplete.
•Synthetic + expert = controlled signal, full label coverage.
•Real = grounding in actual usage.

Best practice: Train broadly, validate realistically.

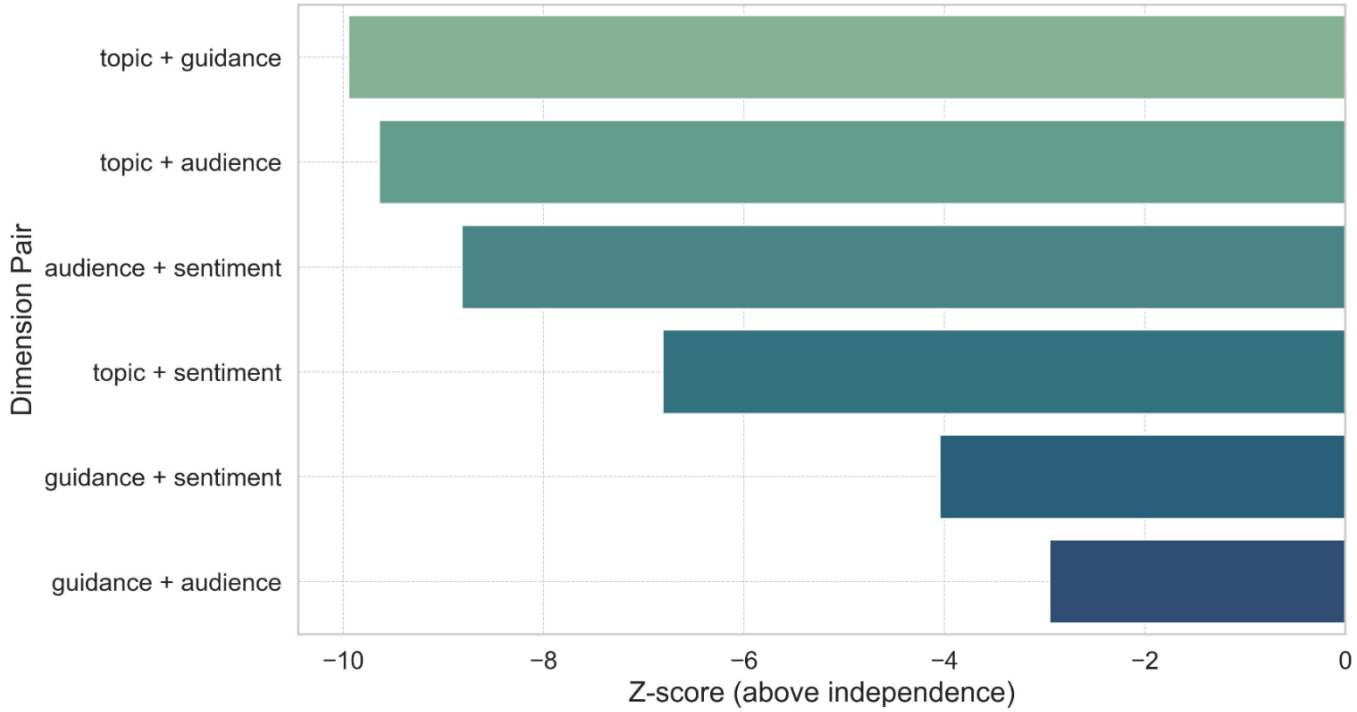
Out-of-sample performance on annotated real data

Dimension	Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)	Precision (Micro)	Recall (Micro)	F1 (Micro)	Cohen's Kappa
Topic	Our Classifier	0.689	0.699	0.645	0.650	0.689	0.689	0.689	0.666
	ChatGPT 4o	0.731	0.593	0.535	0.551	0.731	0.731	0.731	0.711
Comm. stance	Our Classifier	0.924	0.925	0.910	0.917	0.924	0.924	0.924	0.834
	ChatGPT 4o	0.828	0.591	0.551	0.570	0.828	0.828	0.828	0.658
Audience	Our Classifier	0.706	0.713	0.699	0.700	0.706	0.706	0.706	0.622
	ChatGPT 4o	0.506	0.529	0.463	0.437	0.506	0.506	0.506	0.400
Sentiment	Our Classifier	0.700	0.612	0.595	0.594	0.700	0.700	0.700	0.589
	ChatGPT 4o	0.704	0.557	0.393	0.418	0.704	0.704	0.704	0.592

Notes: *Accuracy* is the proportion of correctly classified instances across all classes. *Precision* is the proportion of true positives among predicted positives, measuring reliability of positive predictions. *Recall* is the proportion of true positives among actual positives, measuring completeness. *F1-score* is the harmonic mean of precision and recall, providing a balanced measure of accuracy. Metrics marked *Macro* compute the arithmetic mean across all classes, giving equal importance to each class regardless of size. Metrics marked *Micro* aggregate true positives, false positives, and false negatives across all classes, thus being weighted by class frequency and reflecting performance on more frequent categories. *Cohen's Kappa* corrects accuracy for chance agreement, providing robustness to class imbalance.

Why two classifiers, four dimensions?

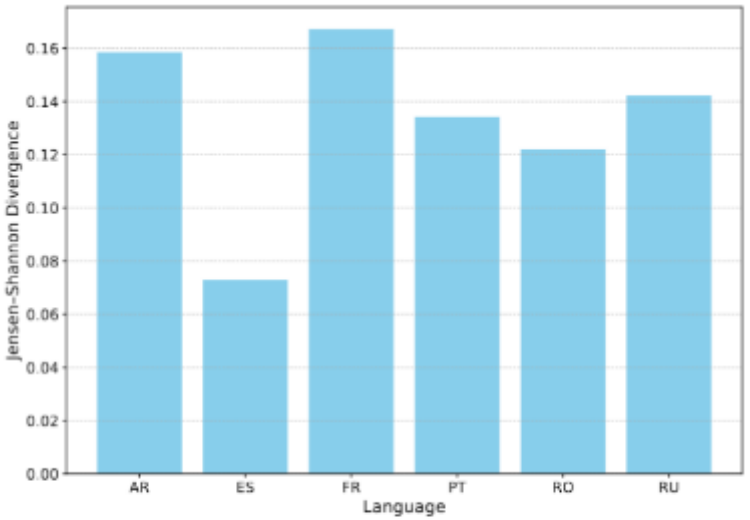
Figure 6: Empirical Pairwise Coupling Between Classification Dimensions



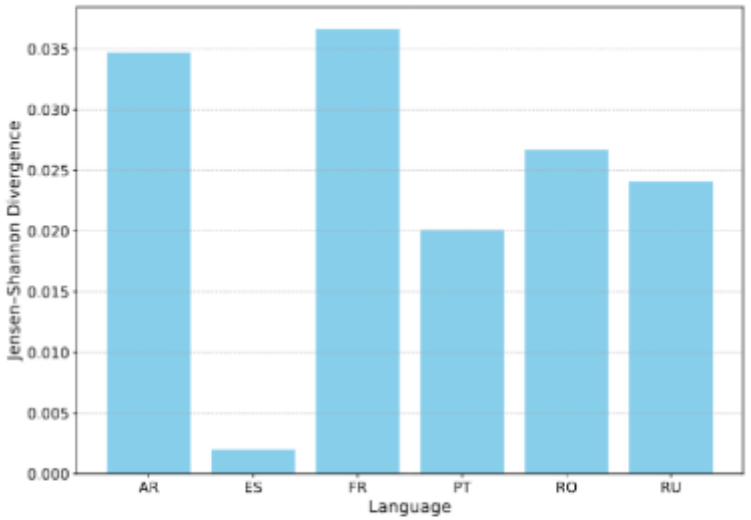
Notes: This figure reports the empirical pairwise coupling between the four classification dimensions using Z-scores. The figure ranks all pairs of sentence-level labels—topic, communication stance, audience, and sentiment, based on how strongly they co-occur relative to what would be expected under statistical independence. Each Z-score compares the number of unique co-occurrences observed in the labeled dataset against a null distribution generated by randomly permuting one label while holding the other fixed. Higher absolute Z-scores indicate greater empirical structure. The pairs *topic* + *guidance* and *audience* + *sentiment* show the strongest dependency, supporting their use as the basis for the two jointly estimated classifier modules. The null ($Z = 0$) represents the threshold for statistical independence.

Consistency across major languages

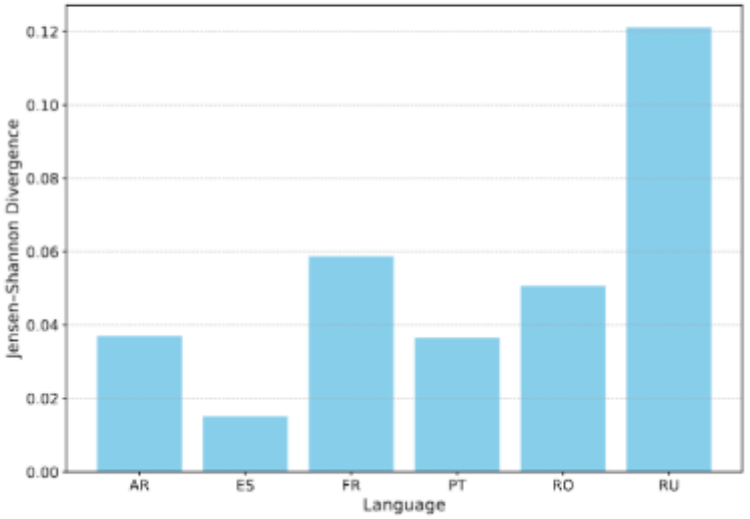
Take local language doc., translate with ChatGPT 4o, and check consistency



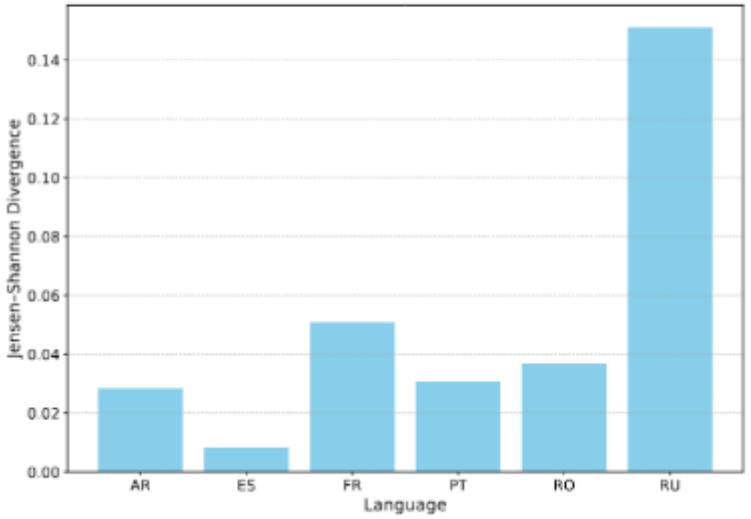
(a) Topic



(b) Communication stance



(c) Audience

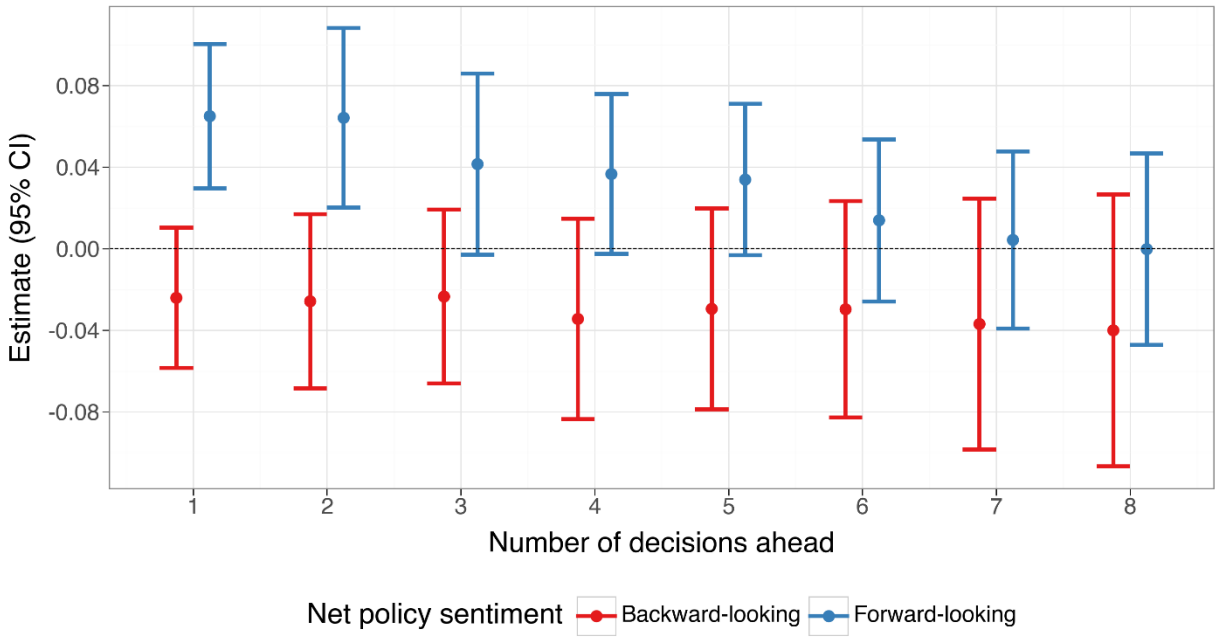


(d) Sentiment

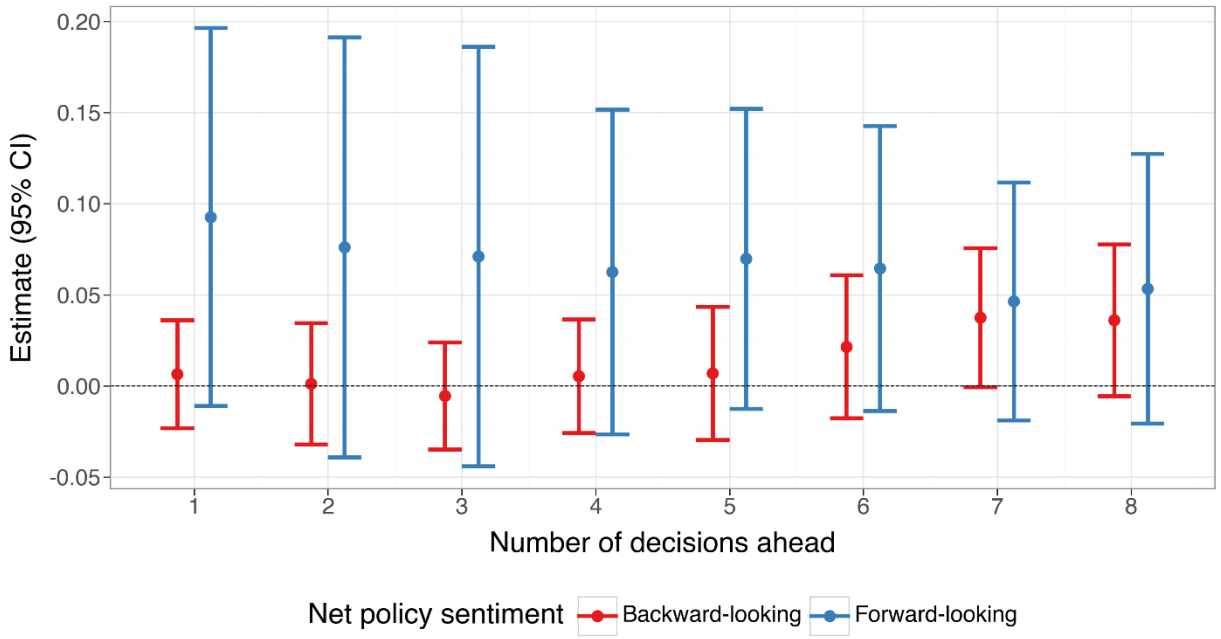
Net policy sentiment (forward) predicts future T-bills and T-bond rates

Different horizons

Short-term market rates (T-bill rates)



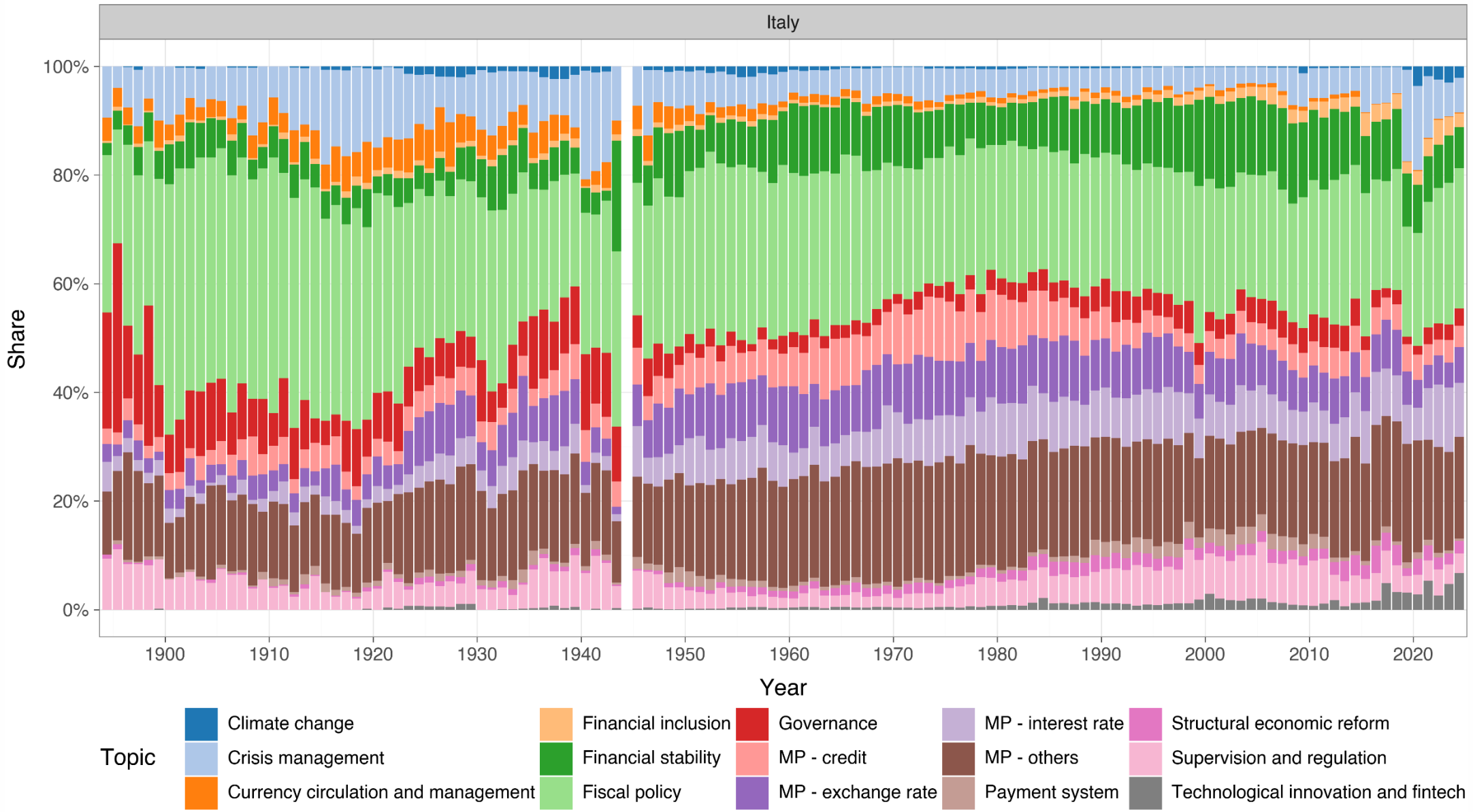
Long-term market rates (T-bond rates)



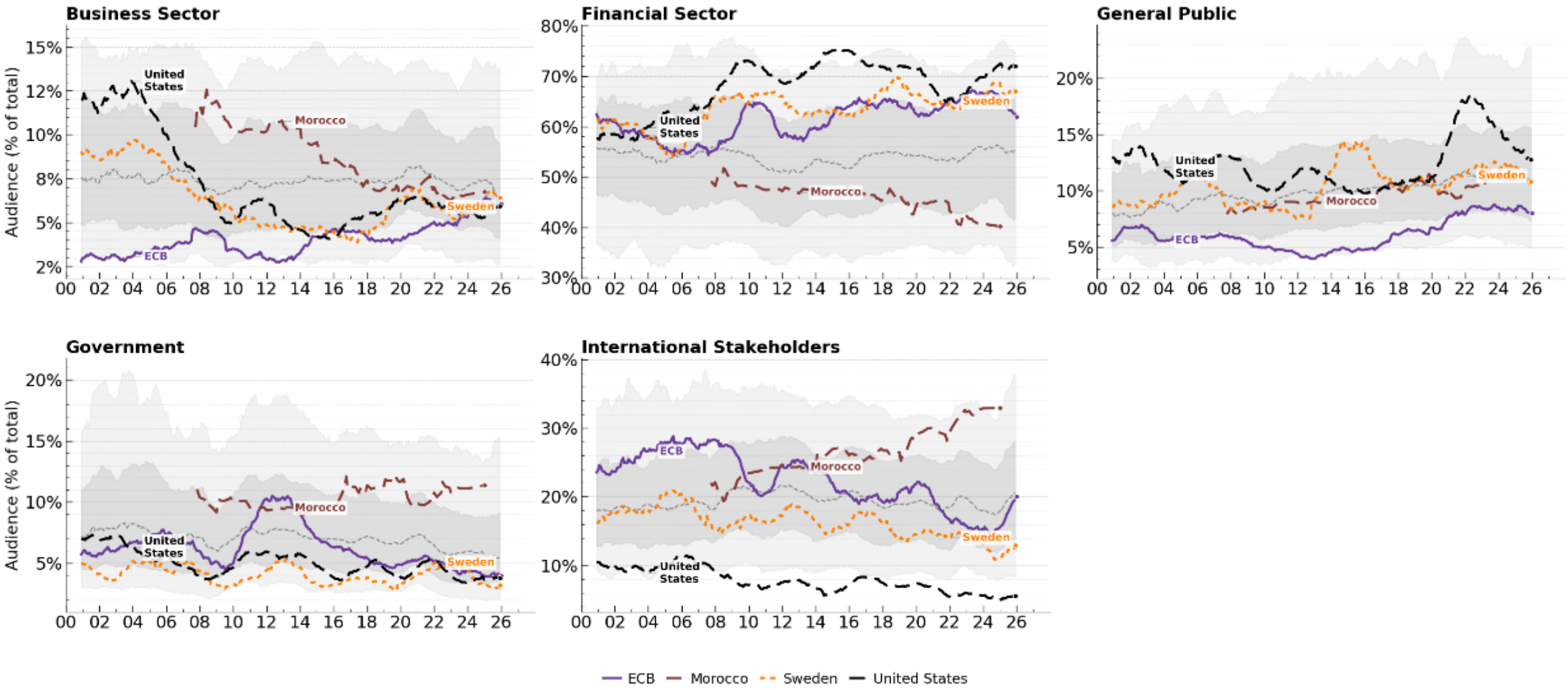
Bottomline: the results corroborate the transmission of central bank communication signals into market rates going forward

Banca d'Italia Annual Reports since 1894

Only in Italian language until 1922; then in English (and Italian)



Financial sector is the main recipient of central bank communication; attention to general public has been increasing



Forward-looking communication gains power over longer horizons:
Overnight index swaps (OIS) – limited sample (15 economies)

	OIS Rate$_{i,t+1}$ (1 month) (I)	OIS Rate$_{i,t+1}$ (3 months) (II)	OIS Rate$_{i,t+1}$ (6 months) (III)	OIS Rate$_{i,t+1}$ (12 months) (IV)
Net Policy Sentiment (Forward)$_{i,t}$	0.017**	0.024***	0.029***	0.036***
	(0.007)	(0.007)	(0.008)	(0.008)
Net Policy Sentiment (Backward)$_{i,t}$	0.017	0.032	0.034	0.036
	(0.013)	(0.020)	(0.025)	(0.028)
Straightforwardness Index$_{i,t}$	-0.014**	-0.012**	-0.008	-0.002
	(0.006)	(0.005)	(0.008)	(0.009)
Explanation Index$_{i,t}$	0.025*	0.014	0.004	-0.014
	(0.011)	(0.010)	(0.011)	(0.014)
Net Confidence Index$_{i,t}$	-0.001	-0.001	-0.007	-0.006
	(0.004)	(0.005)	(0.008)	(0.009)
Policy Rate$_{i,t}$	1.949***	1.951***	1.950***	1.838***
	(0.017)	(0.057)	(0.073)	(0.097)
Inflation (CPI)$_{i,t}$	-3.207***	-1.922***	-1.785**	-0.532
	(0.715)	(0.588)	(0.736)	(0.960)
Exchange Rate (USD/domestic)$_{i,t}$	-0.231	-0.164	-0.286	-0.165
	(0.193)	(0.199)	(0.259)	(0.192)
Country Fixed Effects	x	x	x	x
Time Fixed Effects	x	x	x	x
Observations	491	504	556	605
R^2	0.991	0.983	0.976	0.968

- Intuition: Testing the effect of forward-looking communication on OIS rates cleanly identifies monetary policy expectations from economic agents, as the OIS floating leg is tied to the overnight money market interest rate.

- Bottomline: The impact of the net policy sentiment (forward-looking) strengthens with the contract's maturity, confirming that communication plays a larger role at longer horizons through expectations.

Metric: communication straightforwardness index

$$\text{Straightforwardness index} = \frac{N + |H - D|}{\text{N} + \text{H} + \text{D}}$$

Symbol	Definition
H	Hawkish (tightening signal)
D	Dovish (easing signal)
N	Neutral

Motivation:

- Measures whether central banks deliver clear and coherent policy signals, assessing the extent to which monetary policy communications convey a unidirectional stance versus presenting multiple potential policy paths

Interpretation:

- The index measures the homogeneity of directional (hawkish or dovish) communication signals in the same monetary policy decision
- A higher straightforwardness indicates that a single communication direction is given
- A lower straightforwardness indicates that hawkish and dovish communication elements coexist within the same monetary policy decision due to internal contradictions or the deliberate discussion of multiple scenarios
  - This feature can introduce interpretive ambiguity, particularly in economies where financial markets have less experience parsing complex forward guidance

What is the net confidence index really capturing?

A formal econometric test

Granger Causality: Backward Net Confidence and VIX

Lag	F-stat	p-value	Signif?	Direction
1	1.689	0.194		Backward → VIX
2	0.281	0.755		Backward → VIX
3	0.213	0.887		Backward → VIX
4	0.573	0.682		Backward → VIX
5	0.641	0.669		Backward → VIX
6	0.366	0.900		Backward → VIX
1	0.728	0.394		VIX → Backward
2	4.736	0.009	***	VIX → Backward
3	3.493	0.016	**	VIX → Backward
4	3.126	0.015	**	VIX → Backward
5	2.82	0.016	**	VIX → Backward
6	2.3	0.034	**	VIX → Backward

Granger Causality: Forward Net Confidence and VIX

Lag	F-stat	p-value	Signif?	Direction
1	4.148	0.042	**	Forward → VIX
2	1.237	0.291		Forward → VIX
3	1.095	0.351		Forward → VIX
4	1.418	0.227		Forward → VIX
5	1.114	0.352		Forward → VIX
6	1.239	0.285		Forward → VIX
1	0.01	0.922		VIX → Forward
2	0.021	0.979		VIX → Forward
3	0.994	0.396		VIX → Forward
4	1.662	0.158		VIX → Forward
5	1.434	0.211		VIX → Forward
6	1.184	0.314		VIX → Forward

Key insight 1: VIX → Backward. Central banks justify current risk conditions when looking backward (reactive)
Key insight 2: Forward → VIX. Forward-looking statements have predictive power of future risks, a suggestive evidence of the relevance of central bank communication in shaping future market conditions

Data: almost 70 different types of central bank documents

More than 30 languages, with 87% of the documents in English

Document Type	Description	# Docs	First seen	Last seen
Annual Reports	Key institutional communication, detailing central bank governance, financial statements, economic developments, and policy implementation. Often required by legislation.	4356	1800	2025
Monetary Policy Decisions	Communication issued after interest rate decision detailing rationale behind the policy move.	15276	1936	2025
Monetary Policy Reports	Overview of the central bank's monetary policy stance, actions, and economic outlook. Typically published quarterly.	5218	1993	2025
Financial Stability Reports	Overview of the financial sector risks and vulnerabilities and outlook. Typically published semiannually.	2184	1996	2025
Speeches	Speeches by central bank decision makers. Most of the data comes from Campiglio et al. (2025)	37245	1986	2025
Other Specialized Docs	Press releases and reports on specialized topics that do not fall into the above categories	11558	1976	2025