BANCO DE ESPAÑA
Eurosistema

RASI & RAUI

SENTIMENT AND UNCERTAINTY INDICATORS USING ARTIFICIAL INTELLIGENCE

MORTEZA GHOMI AND SAMUEL HURTADO

DECEMBER 2025

RASI: Retrieval Augmented Sentiment Indicator

RAUI: Retrieval Augmented Uncertainty Indicator

They are indicators using artificial intelligence

We can calculate specific indicators for topics of interest at any time

A full draft of the paper is already online: https://ghomimorteza.github.io/index/Project6.html

The usual methodology for calculating sentiment and uncertainty indicators is based on counting words, in two stages

Selection of relevant news articles: using a dictionary of words that define a topic

Quantification of sentiment or uncertainty: using dictionaries of positive/negative terms, or of terms that capture uncertainty

An LLM model could improve both the selection and the quantification steps
Problem: if the database is very big, asking an LLM about every news piece is not feasible (too expensive, or too slow)
(Dow Jones DNA for Spain means >4 million news articles from 2000 to 2025)

But we can use AI techniques to improve in both steps (selection and quantification)

RAG (Retrieval Augmented Generation) is an AI technique usually employed for building chatbots, e.g. to make one that answers using the contents of the Bank's intranet

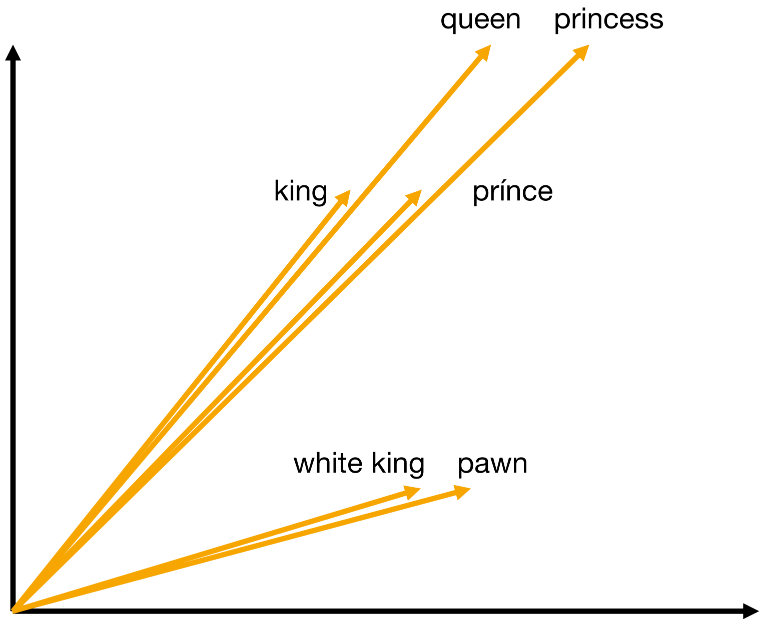
Query: how do I file for a reduction of working hours to care for a family member?
RAG:

- Searches the intranet and selects the 20 paragraphs that have content that is closest to that query
- Sends that information to an LLM with the following prompt:
  Consider the following information:
  content of first search result
  content of second search result
  etc
  Using that information, tell me: how do I file for a reduction of working hours to care for a family member?
- Return to the user the response of the LLM

It allows developers to use a generic LLM (e.g. GPT, Deepseek, Llama, Gemma, Phi, etc) to provide answers based on a specific set of information, that can be of very big size (the LLM only receives the first search results) and can be confidential (this information doesn't need to go outside the institution), without having to train the model (the information is not coded in the weights of the model: easier and with a lower chance of generating hallucinations)

The source retrieval part is usually done through semantic search, using an embeddings model

A text embeddings model is a function that transforms text into numbers
It generates a projection of a text into a multidimensional space (e.g. 4096 dimensions)



King, queen, prince and princess should be in the same area of that space
It should hold (approximately) that prince + (queen – king) = princess
Models don’t look at individual words: white king should be in a very different area, close to pawn

There are very big embedding models: it's not word embedding, but text embedding

LLMs use text embeddings as the first stage of their processing
E.g.. Qwen2-7B is an LLM with 7B parameters, of which 1.1B are used for embeddings

But there are also models specifically for embeddings, that will work better for semantic search than using the embeddings layers of a general purpose LLM

One model can be good in English and bad in Spanish, or good for computing topics and bad for economics topics: there are benchmarks, but it's not enough to just use the one with the best benchmark score, we are going to test which one works best for our task in particular

The embeddings model is used in two steps

Pre-processing of the database:

We project each news piece (or each paragraph in the BdE intranet) into the embeddings space
This takes time, but not as much as processing all that information with an LLM
(200x faster than using deepseek in ollama, 20x faster than gpt-4o-mini)

We only need to do this once (no need to repeat it for each new search) (it took ~9 days)

Search the database:

Project the query (e.g. “inflation and prices” or “international trade”) into the embeddings space,
and select the N articles that are closest to that query in that space
(the usual metric to calculate this distance is cosine similarity)

BANCO DE ESPAÑA
Eurosistema

Previous step: choose an embeddings model that works well for our task
finding news articles that talk about a specific topic in a database in Spanish

Candidates to test:

5 open source models available in ollama that seemed pomising
bge-m3, mxbai, nomic, mle5l, qwen gte-1B

2 models from openAI that are available in our private Azure AI forge space
openai2, openai3large

dictionaries
the traditional methodology that just looks for a specific list of words

gpt-4o-mini (also available in Azure; more expensive and slower than embeddings models)
just ask if a news piece belongs in a topic or not (yes/no)

Random sample of 443 news articles from Spanish newspapers, from the relevant sections (approx. the size of the database for an average day), classified by hand by a trainee

11 test queries (side task: learn which queries work, and what's the best way to define the query):

- international trade
- indicators of economic activity
- wages
- firms
- banks and financial institutions
- energy
- tourism
- inflation and prices
- employment
- debt
- financial markets

E.g. for international trade the full query is (in Spanish):

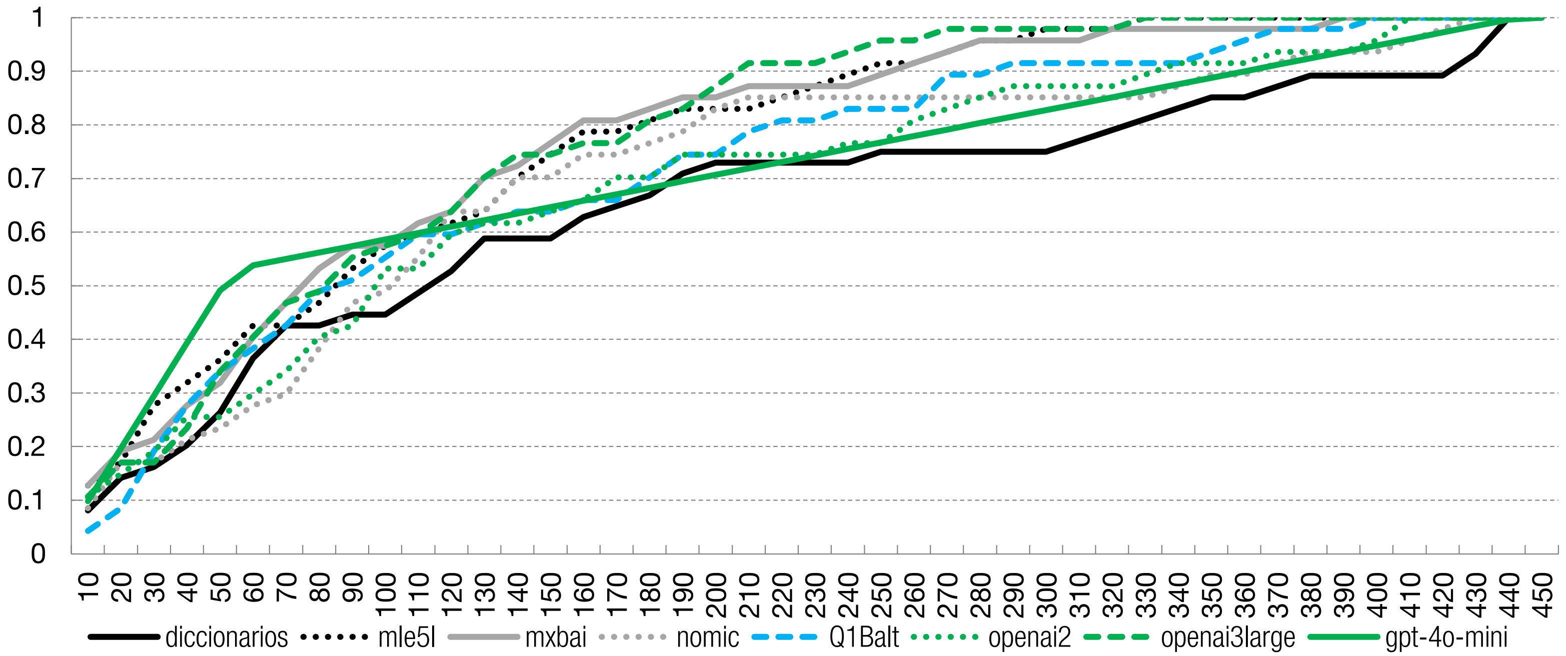
International trade, tariffs, borders, global value chains, international transport of goods, exports and imports of goods and services, tourism, world trade organization, protectionism, globalization, international logistics, foreign investment, competitiveness

(one thing we learned: this is too long and too broad)

For each model, for each query, we calculate the cosine similarity of all news articles in the random sample, and sort them from closest to farthest

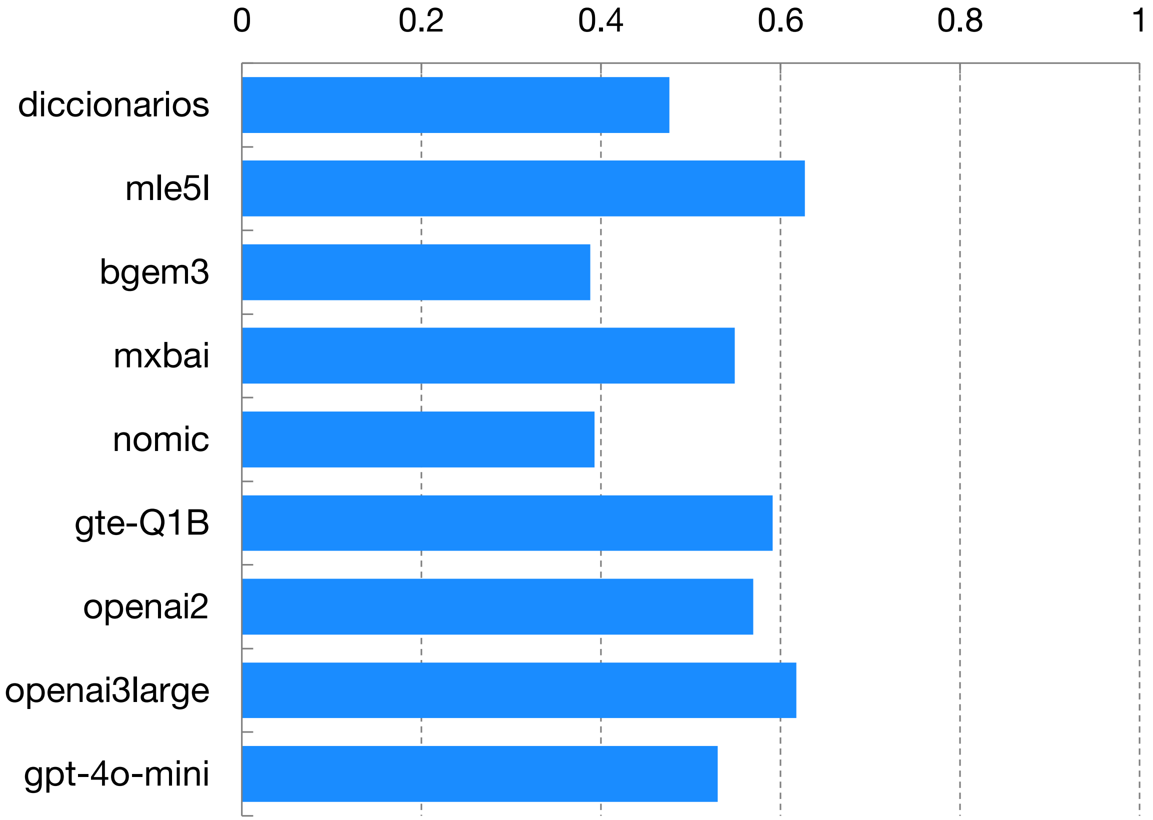
We look at the percentage of news articles that talk about a topic (according to the human) that have been found (vertical axis) by each embeddings model in its first X news (horizontal axis)

ACCURACY OF EMBEDDING MODELS, COMPARED TO HUMAN CLASSIFICATION, FOR THE TOPIC OF INTERNATIONAL TRADE (49 NEWS PIECES OUT OF 443)



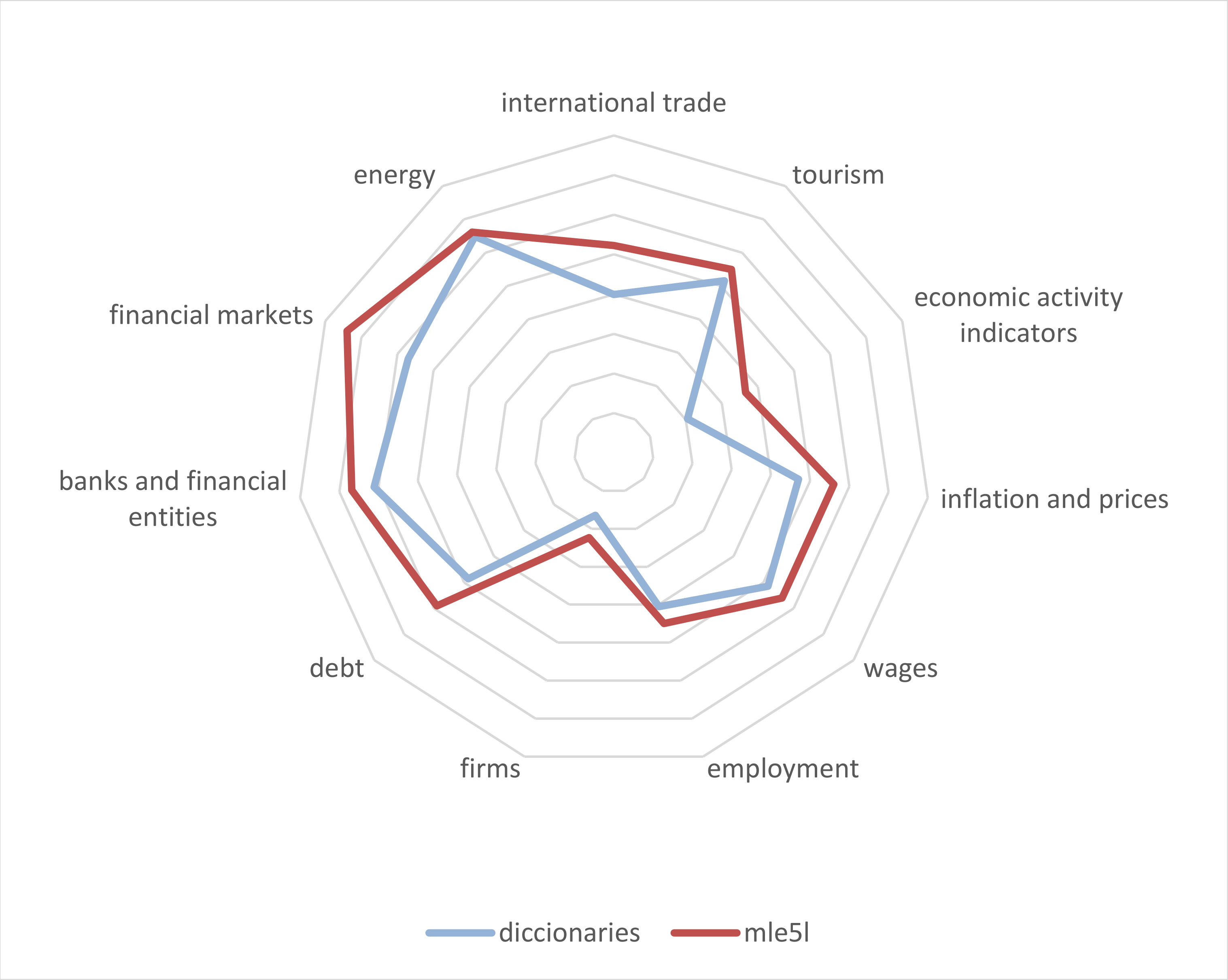
The Gini index (the area between that line and the diagonal) is a good statistic to summarize how well each model is performing: we want the first brackets to contain all the hits, and to have only misses from there onwards (high inequality, high Gini)

The average of the Gini for the 11 test queries summarizes how well each model performed:



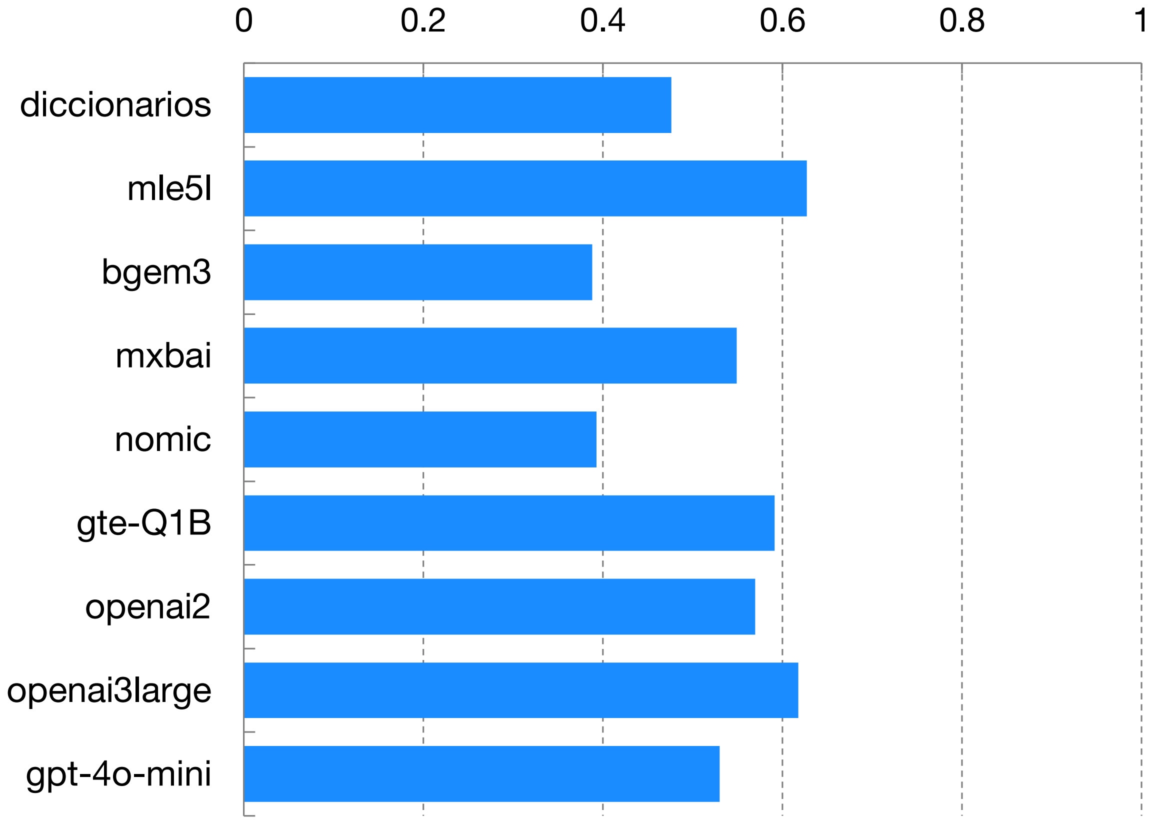
multilingual-e5-large:

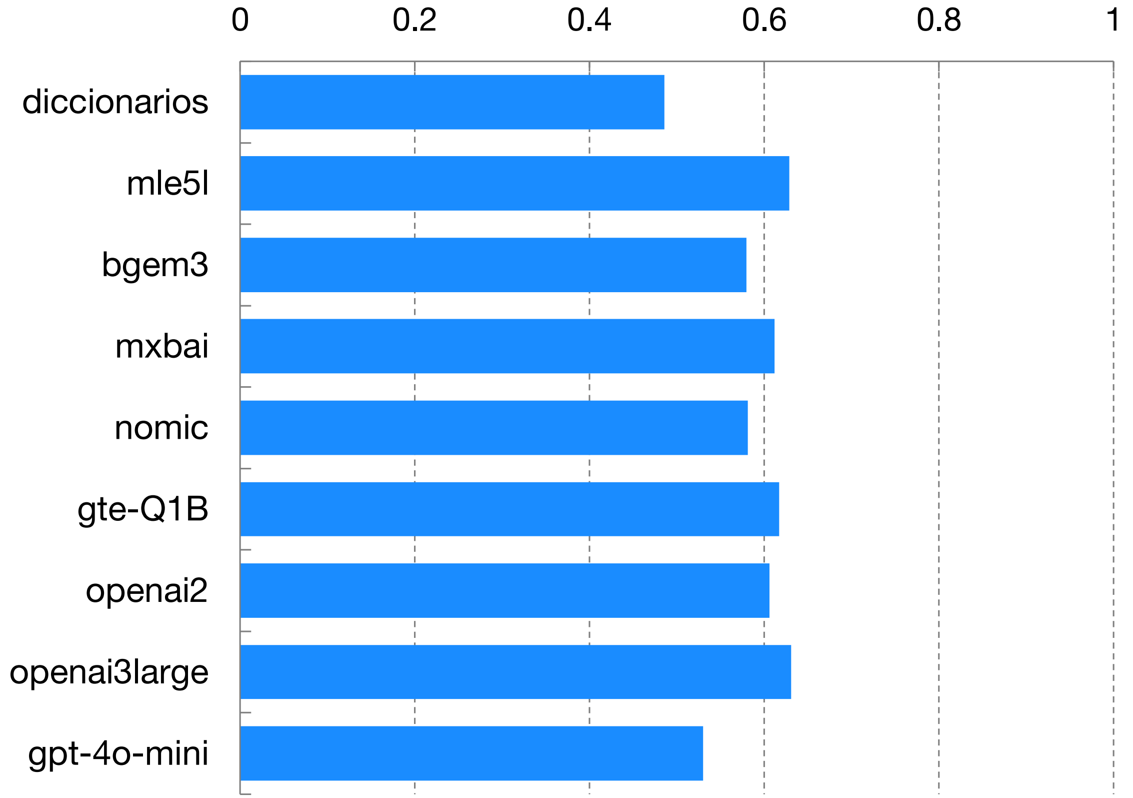
- Developed by Microsoft (Wang et al, 2024)
- Available in huggingface and in ollama
- Can be executed locally (or on private cloud instance)
- Free to use
- "Small" (560 million parameters)
- "Fast" (~9 days to process the full >4M articles of the 2000-2025 database)



BANCO DE ESPAÑA
Eurosistema

We can improve the selection using gpt to get rid of the false positives in a second stage (helps more in the case of models that didn't work particularly well in the first stage)





- Models specifically for embeddings work better than using the embeddings layers of a general purpose LLM

- No model is perfect: we will use mle5 to do a selection, and check later with the big LLM

- Text preprocessing: unneeded, and can harm the results, better use original text as is

- Test with “inflation and prices”: the short query (3w) works better than the long query (133w)

- Test with “wages” vs “employment”:
    - The embeddings model can distinguish correctly between them
    - The errors make sense.
        - *In wages the false positives are news about unemployment benefits, unemployment, layoffs, working hours, scholarhsips, public tenders, industrial policy, firm mergers, GDP growth, inflation, etc*
        - *In employment there are more false positives, about unemployment benefits, productivity, etc*

- Some topics are difficult to define, and for them the semantic search doesn’t work very well

- Embeddings model finally selected: multilingual-e5-large

- Topics to be used in our sentiment and uncertainty indicators:

    1. Geopolitical tensions
    2. International trade and tariffs
    3. Exports markets and external demand
    4. Financial markets
    5. Prices of energy and other raw materials
    6. Inflation, prices and markups
    7. Wages and collective bargaining
    8. Fiscal policy and sustainability of public finances
    9. Housing market
    10. Confidence of agents and internal demand

(we will have to repeat the model selection exercise using these topics, but it will take time, and it's not a priority right now)

Once the relevant news articles have been selected…

- 130.000 for each of the 10 topics (on average, 100 per week: 25*52*100=130.000)
- A cosine similarity threshold is calculated for the whole sample (2000-2024)
- There will be more news articles evaluated in times when a topic is talked about more

We ask GPT-4o-mini to provide a value for the level of sentiment and uncertainty in each article

- The prompt is long and detailed, and also asks for confirmation that the article talks about the topic

- The prompt contains ten varied examples (different for each topic) of headlines (real or made up) and the value of sentiment and uncertainty that we would assign to it
- These examples are very important to get consistent standards across answers from GPT (in the initial tests, without examples, the results were very noisy)

- This runs on a private instance in Azure AI Forge (to meet the terms of use of our news database) (for each topic it takes around 20 hours and costs around 40 euros)

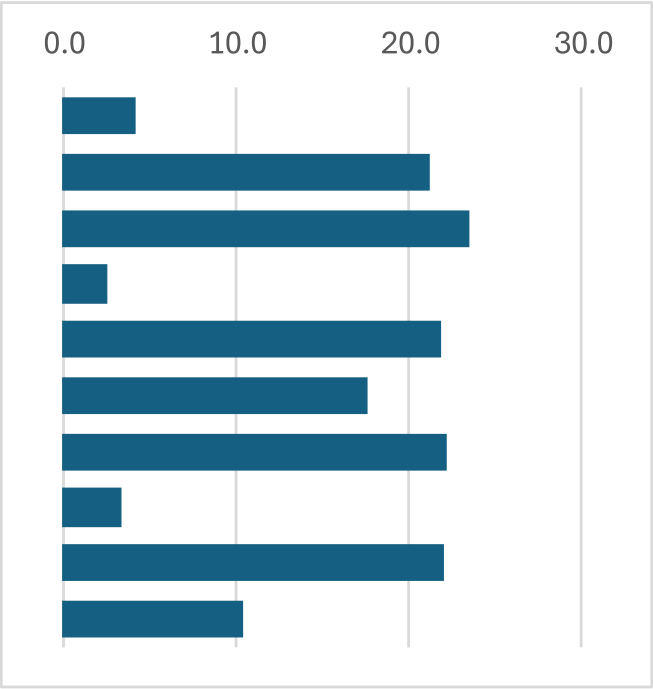
This is a news article that talks about fiscal policy and the sustainability of public finances.
Give me a numerical assessment of the sentiment (0 if very negative, 5 if neutral, 10 if very positive) and the uncertainty (0 if there is little uncertainty, 10 if there is a lot of uncertainty) that reflects how this news talks about fiscal policy, sustainability of public finances, taxes, subsidies, pensions, public spending, fiscal rules, government budgets, etc.
(if the news does not talk about any of these topics, simply respond None, None, None).
Also tell me if the news refers to Spain or another country (in case of doubt, we can assume it refers to Spain).
Give me a simple answer. If the evaluation is that the value for sentiment is Vs and the value for uncertainty is Vi, and the news refers to Spain, give the answer in the following format:
Sentiment Vs, uncertainty Vi, Spain yes.
Here are some examples of hypothetical news headlines, and what their assessment might be:
Garriga (Vox) predicts that Sánchez's rearmament plan will entail a hidden tax increase (Sentiment 4, Uncertainty 6, Spain yes).
PSOE of Calahorra denounces a 211 percent increase in debt and warns of more taxes (Sentiment 2, Uncertainty 7, Spain yes).
PP demands Sánchez to lower taxes for farmers in the face of US tariff uncertainty (Sentiment 5, Uncertainty 7, Spain yes).
The city council approves a budget of 13.2 million that prioritizes social spending, education, and the local economy without raising taxes (Sentiment 8, Uncertainty 2, Spain yes).
Trump threatens Harvard with withdrawing tax-free status after freezing 2.2 billion for not yielding to his political demands (Sentiment 3, Uncertainty 8, Spain no).
PP government does not rule out further tax cuts in Cantabria: we may have some margin (Sentiment 7, Uncertainty 6, Spain yes).
Prime Minister of Portugal presents a program with lower taxes and more public housing (Sentiment 8, Uncertainty 2, Spain no).
The State grants Navarra more tax collection powers for new taxes (Sentiment 5, Uncertainty 5, Spain yes).
The Senate urges the Government to present budgets with the abstention of Sumar, PNV, and Junts, without certainty that they can be approved (Sentiment 4, Uncertainty 9, Spain yes).
Sánchez reiterates his commitment to presenting a budget this year and works with parliamentary groups (Sentiment 7, Uncertainty 7, Spain yes).
Thousands of people demonstrate in the three Basque capitals for decent pensions (Sentiment 6, Uncertainty 7, Spain yes).
The pension reform passes the first AIReF exam, which warns of sustainability (Sentiment 7, Uncertainty 8, Spain yes).
Paula Conthe: The improvement in the risk premium is here to stay (Sentiment 7, Uncertainty 6, Spain yes).
The Government begins negotiations on the General Budgets (Sentiment 6, Uncertainty 5, Spain yes).

This is the news article to evaluate:

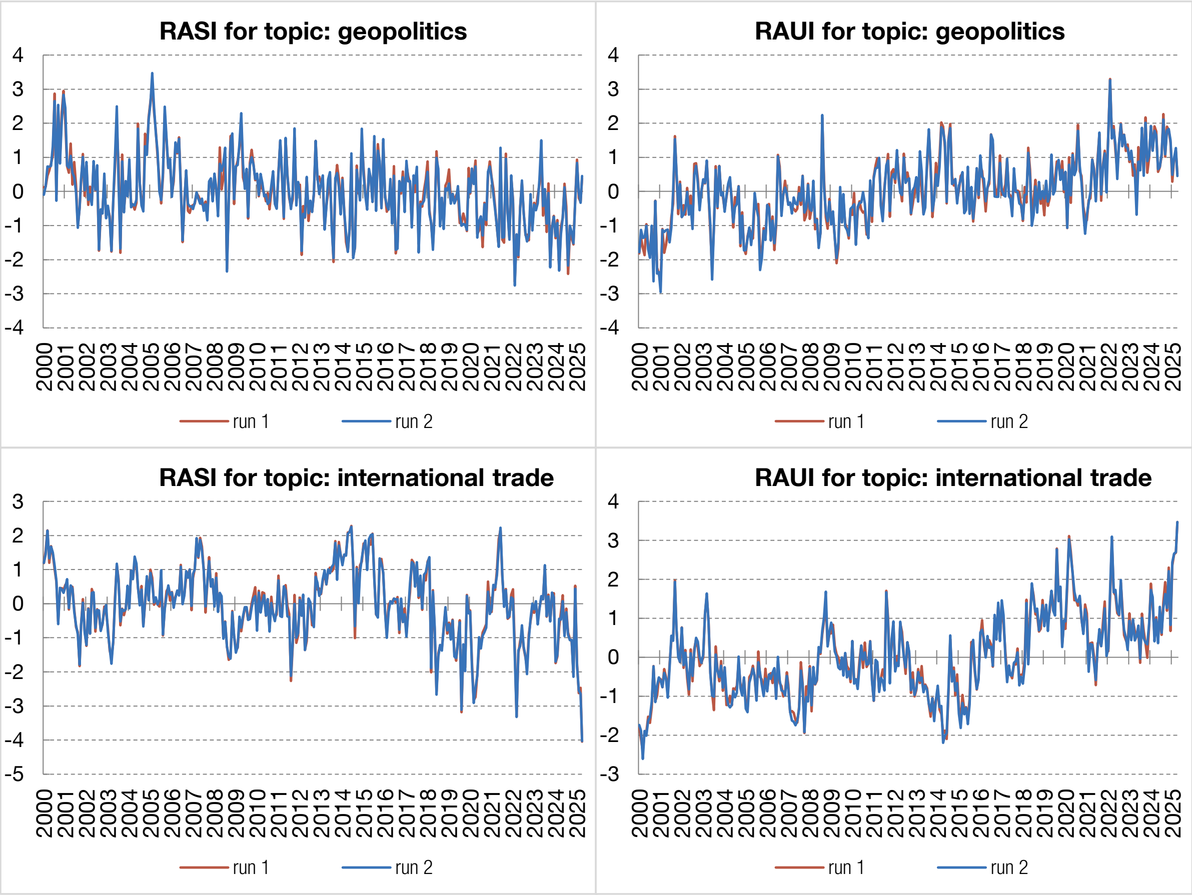
The answers from GPT confirm that in general terms the news articles selected with the embeddings model talk about the correct topics, with percentages of false positives that are around 4% for easier topics and around 20% for more difficult topics.

Percentage of false positives in each topic

Geopolitics	4.2
International trade, tariffs, global value chains	21.3
Export markets and external demand	23.5
Financial markets	2.6
Price of energy and other raw materials	21.8
Inflation, prices and markups	17.6
Wages and collective bargaining	22.1
Fiscal policy and sustainability of public finances	3.4
Housing market	22.0
Confidence of agents and internal demand	10.5



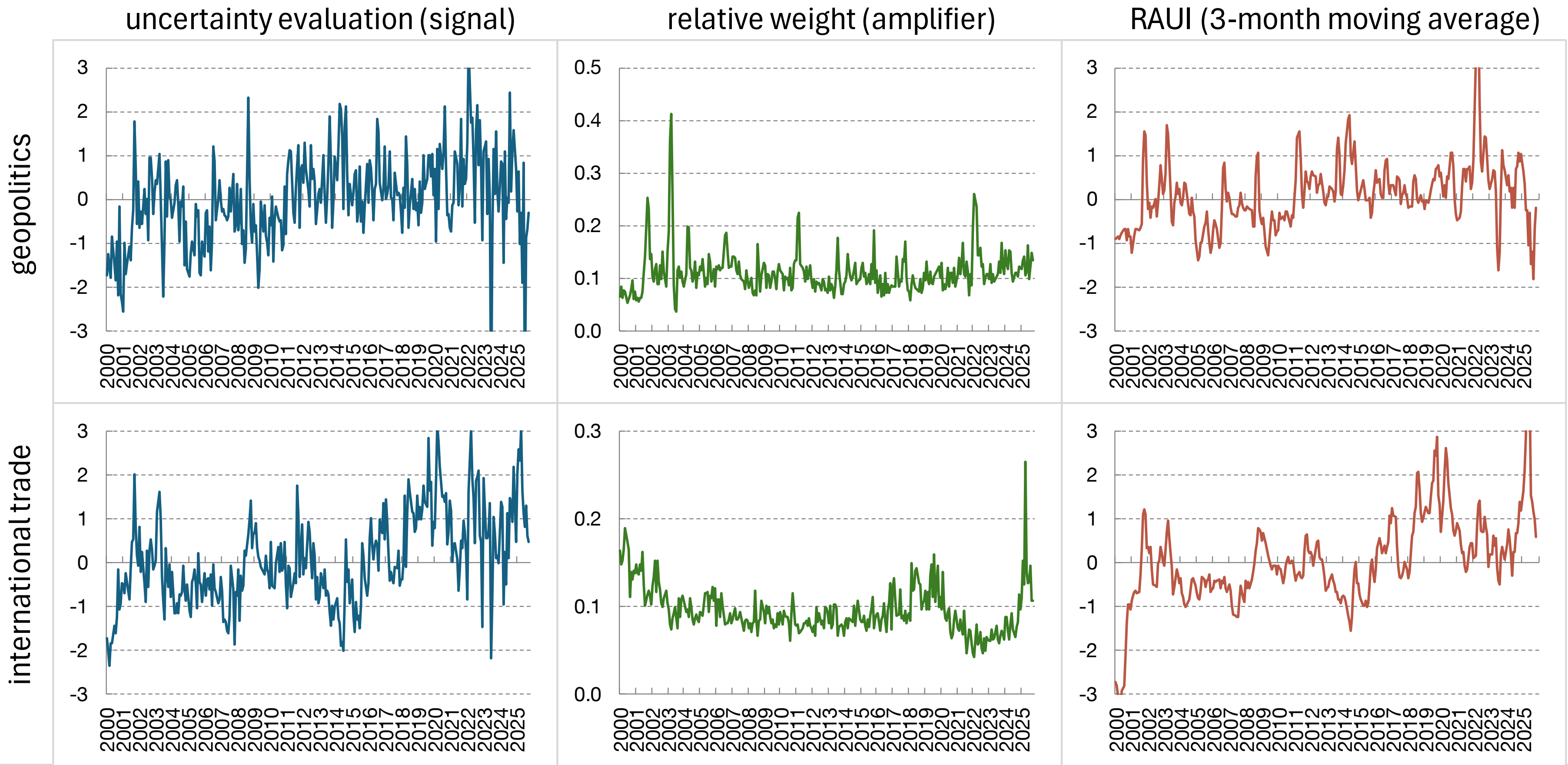
The answers from GPT are reasonably consistent: asking a second time about all the news pieces in a given topic, the correlation between answers is 0.8-0.9 if we calculate it using individual news pieces, and 0.98-0.99 if we calculate it on the monthly series



t02 comercio internacional

incertidumbre	0	1	2	3	4	5	6	7	8	9	10	rech
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.6	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
3	0.0	0.0	0.5	2.3	1.3	0.1	0.0	0.0	0.0	0.0	0.0	0.1
4	0.0	0.0	0.1	1.4	8.4	1.8	0.7	0.1	0.0	0.0	0.0	0.3
5	0.0	0.0	0.0	0.0	2.0	3.1	2.0	0.7	0.0	0.0	0.0	0.2
6	0.0	0.0	0.0	0.0	0.7	1.8	6.4	4.5	0.1	0.0	0.0	0.1
7	0.0	0.0	0.0	0.0	0.2	0.8	5.0	21.9	2.6	0.0	0.0	0.2
8	0.0	0.0	0.0	0.0	0.0	0.0	0.1	2.4	4.8	0.2	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
rechaza	0.0	0.0	0.1	0.1	0.3	0.2	0.2	0.1	0.0	0.0	0.0	20.6

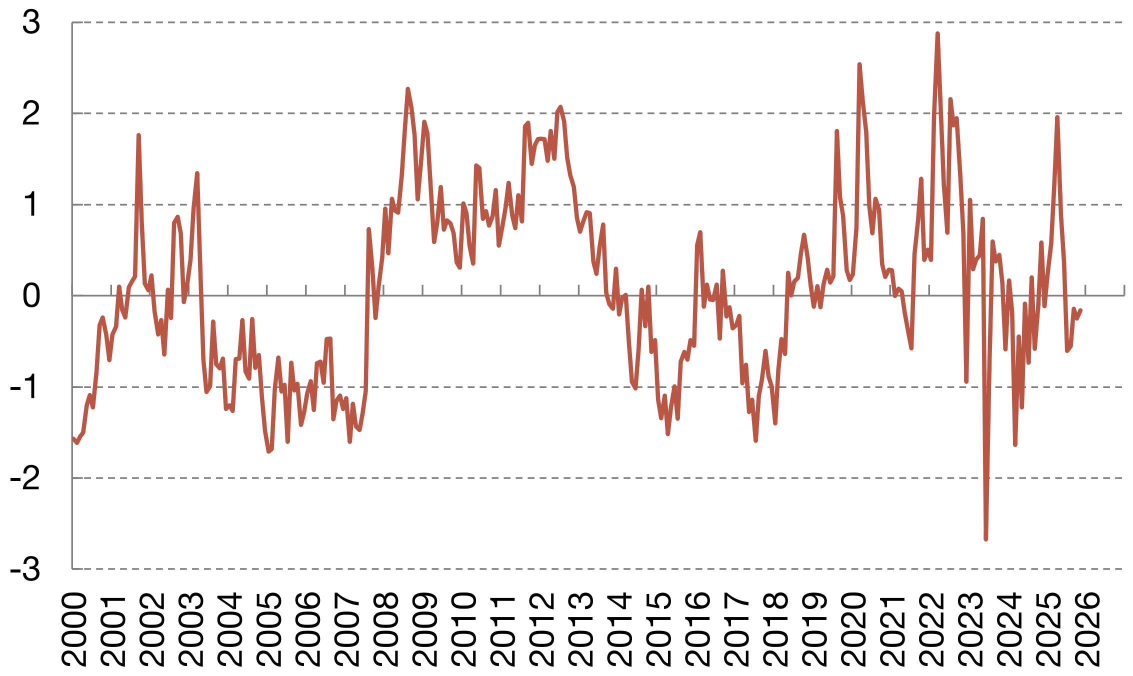
With the answers from GPT for each news article, we calculate the average of the values for sentiment and uncertainty of the articles identified and confirmed as belonging to each of the ten topics, in each month. The values are normalized subtracting the mean and dividing by the standard deviation. The relative weight of each topic in each month is used as an amplifier to generate the final RASI or RAUI for each topic.



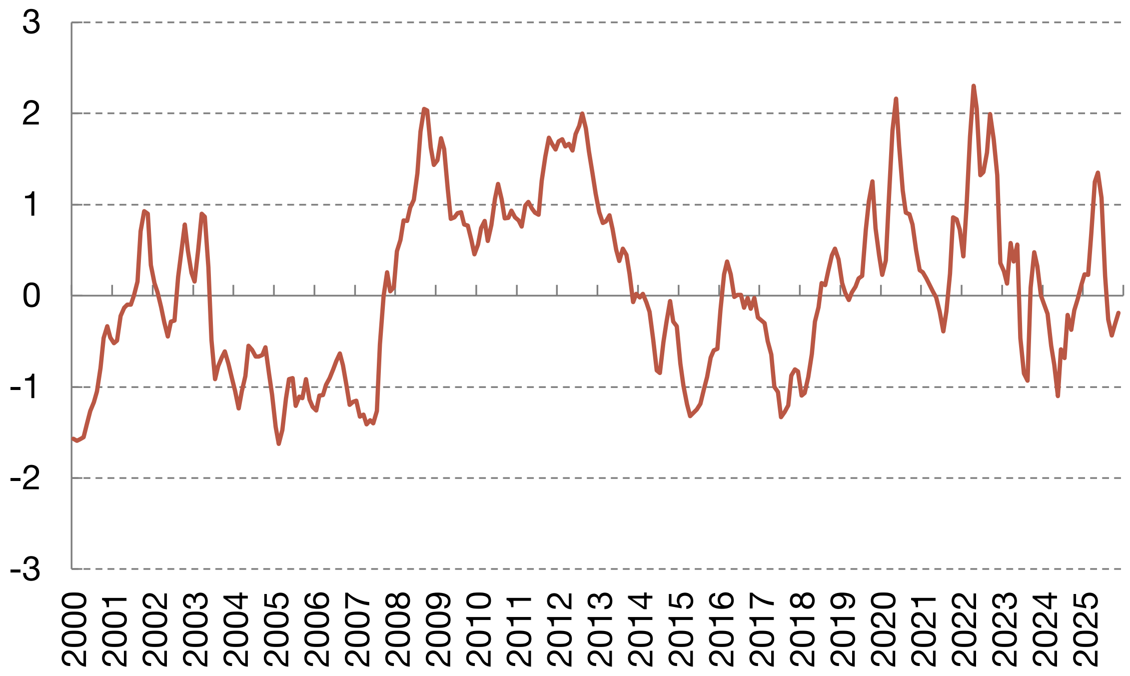
BANCO DE ESPAÑA
Eurosistema

The RASI or RAUI for the individual topics can be aggregated easily to generate an overall indicator
Since the indicators are usually noisy, we often look at a 3-month moving average

Overall RAUI (monthly series)

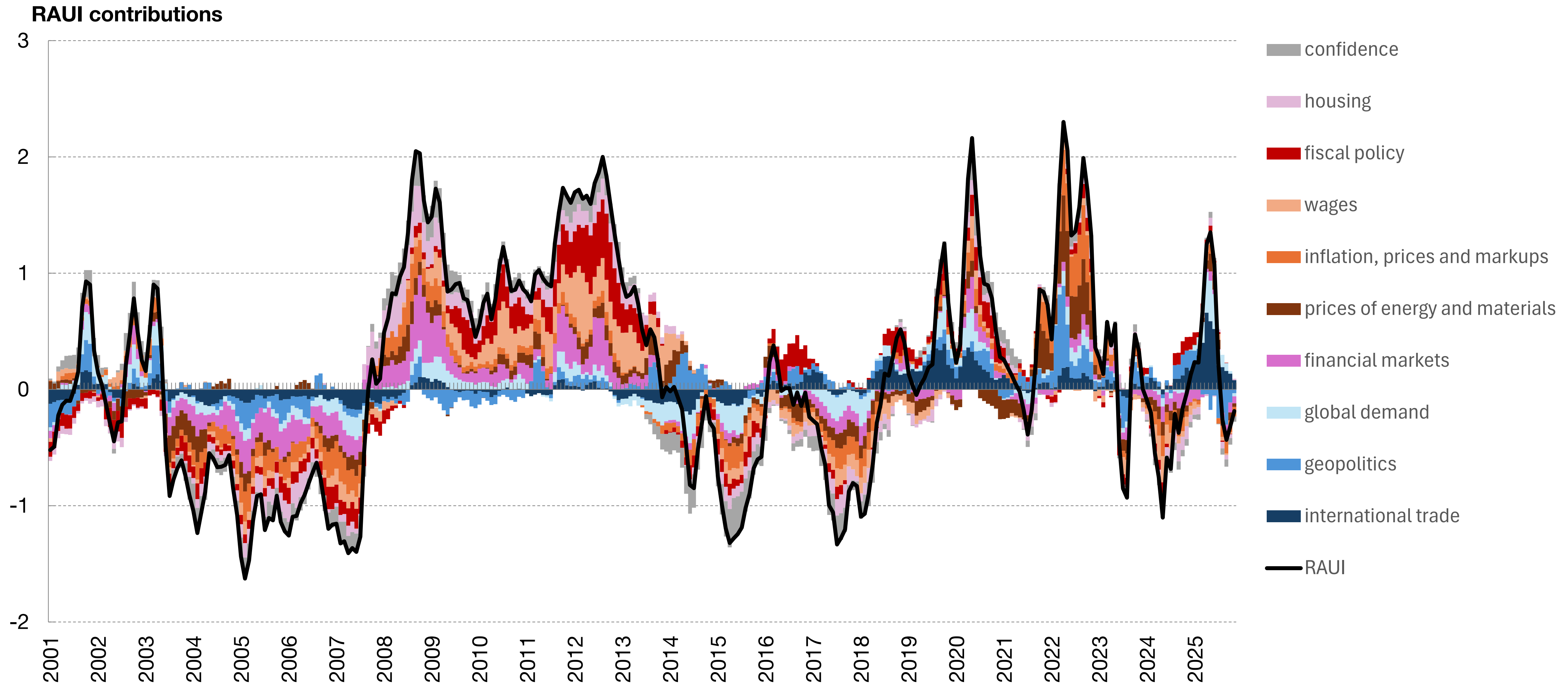


Overall RAUI (3-month moving average)



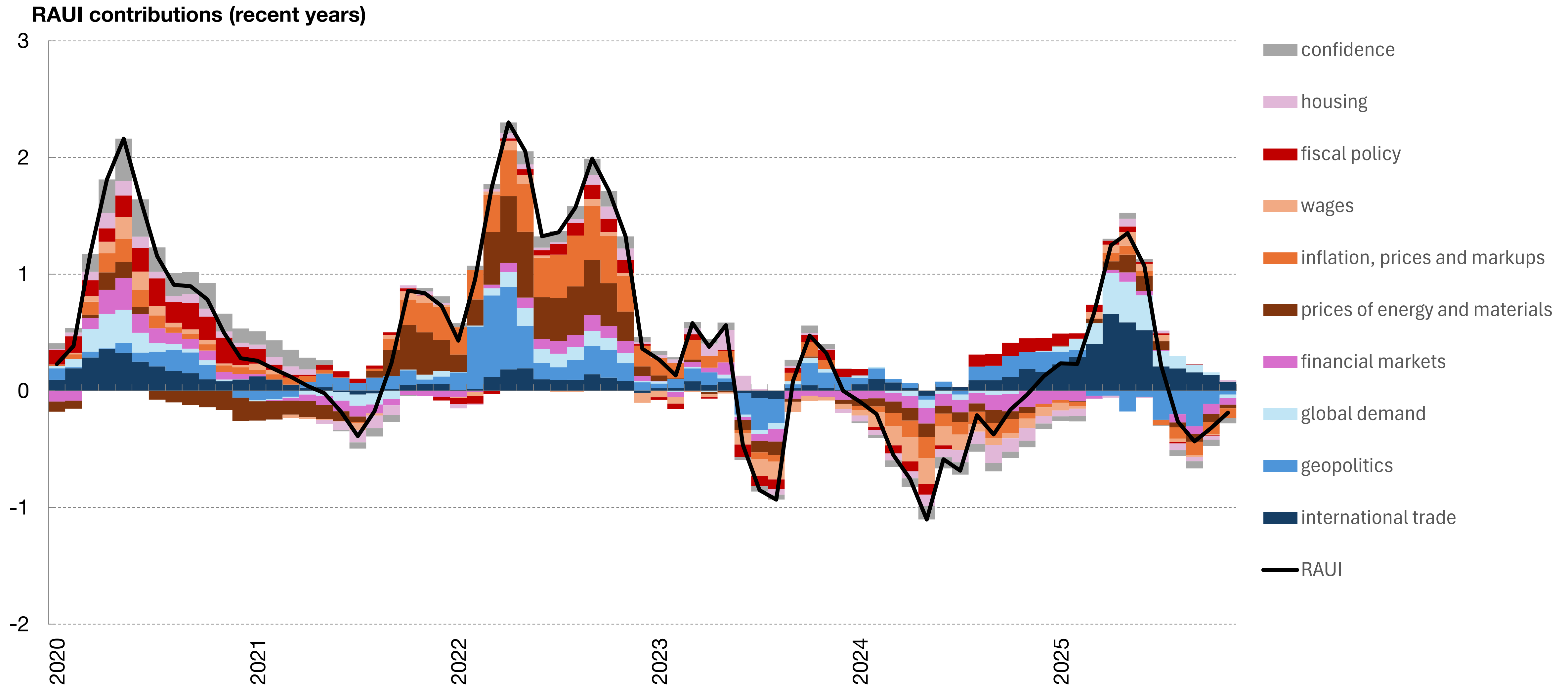
BANCO DE ESPAÑA
Eurosistema

An advantage of this methodology is that we can plot the contribution of each topic to the overall indicators

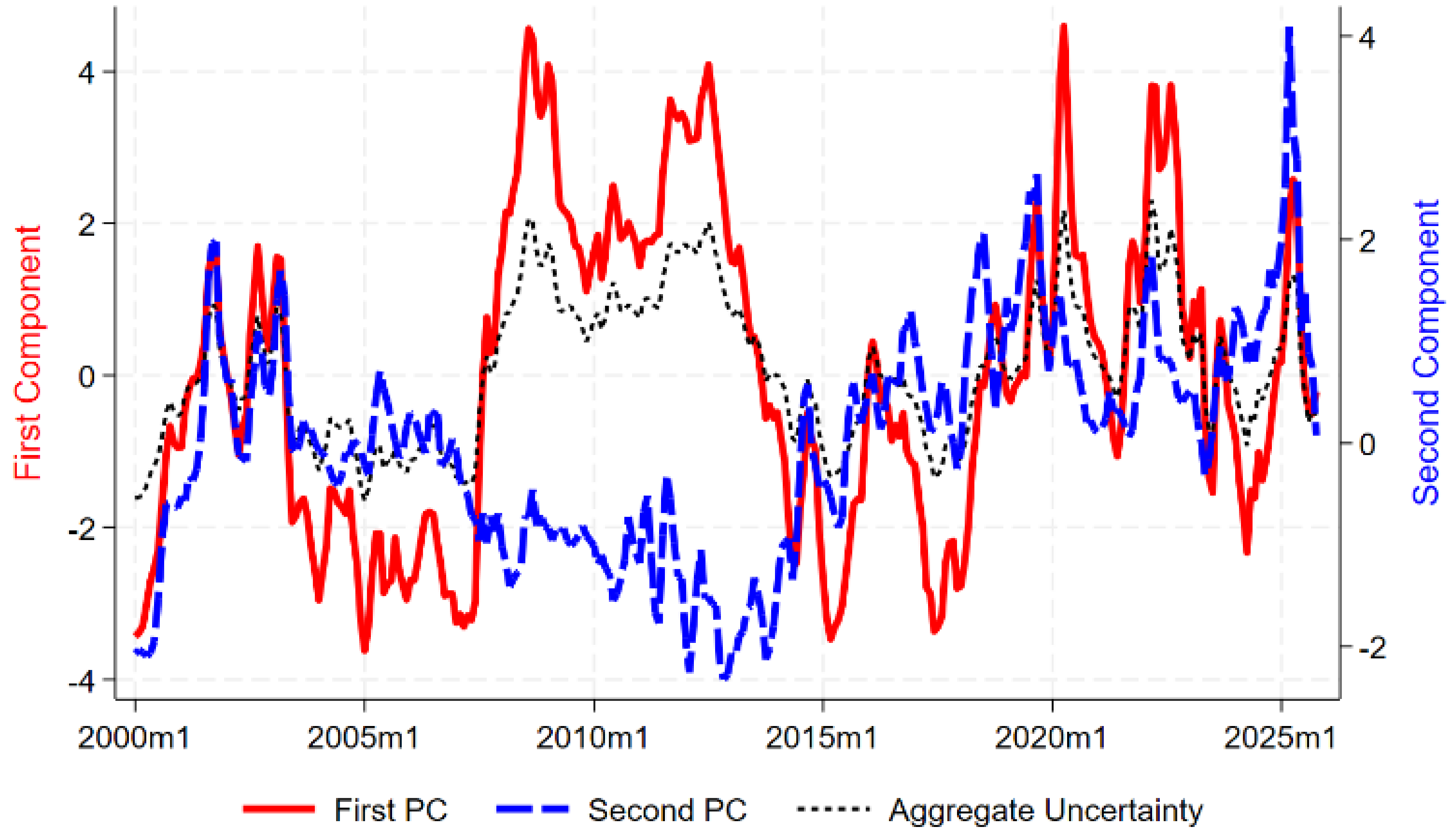


BANCO DE ESPAÑA
Eurosistema

An advantage of this methodology is that we can plot the contribution of each topic to the overall indicators

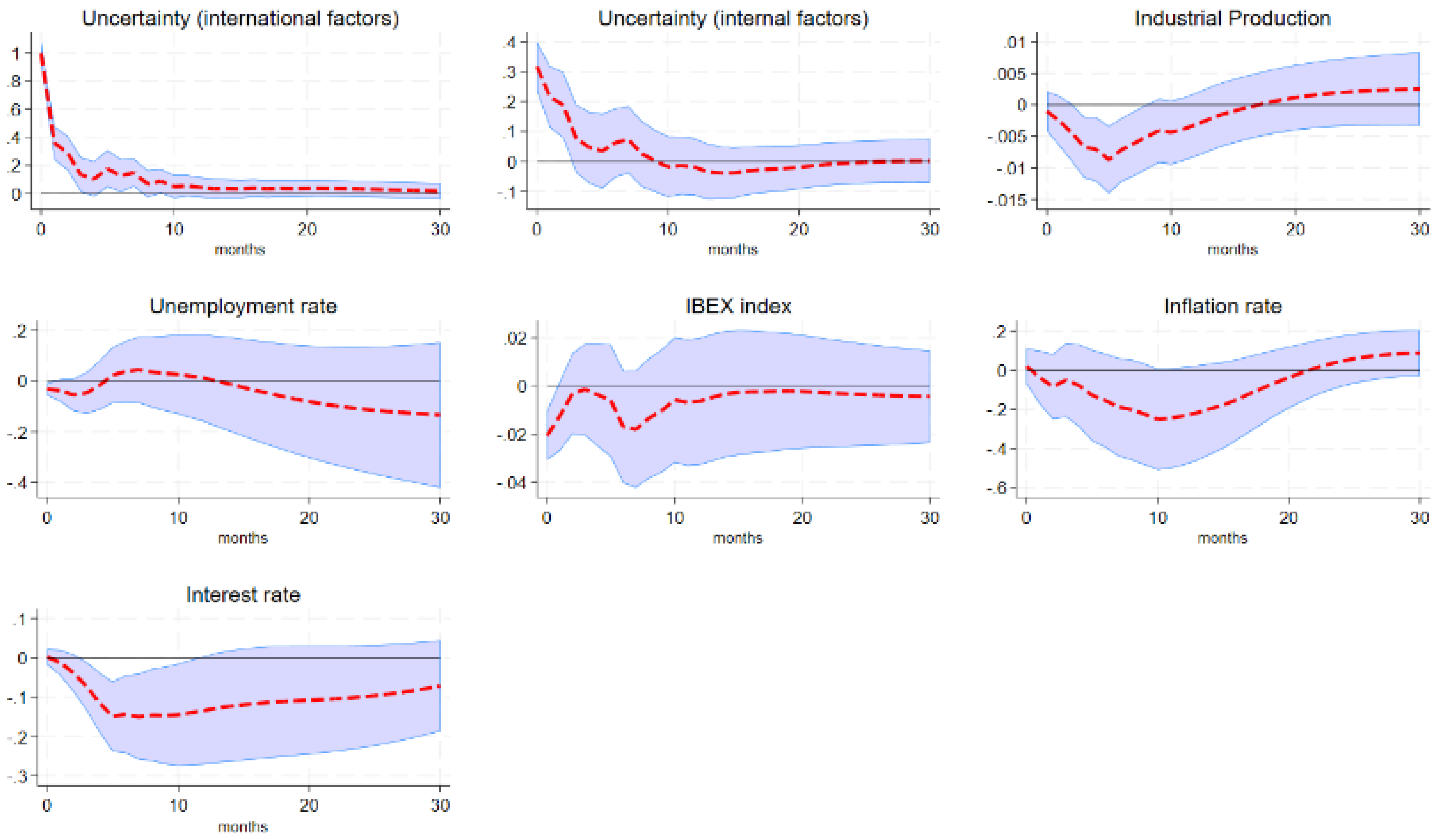


- Estimating a single big VAR with all RAUI doesn't work: too much detail
- We do PCA to summarize our 10 RAUI into two principal components: the first one captures mostly internal uncertainty, while the second one captures mostly external uncertainty

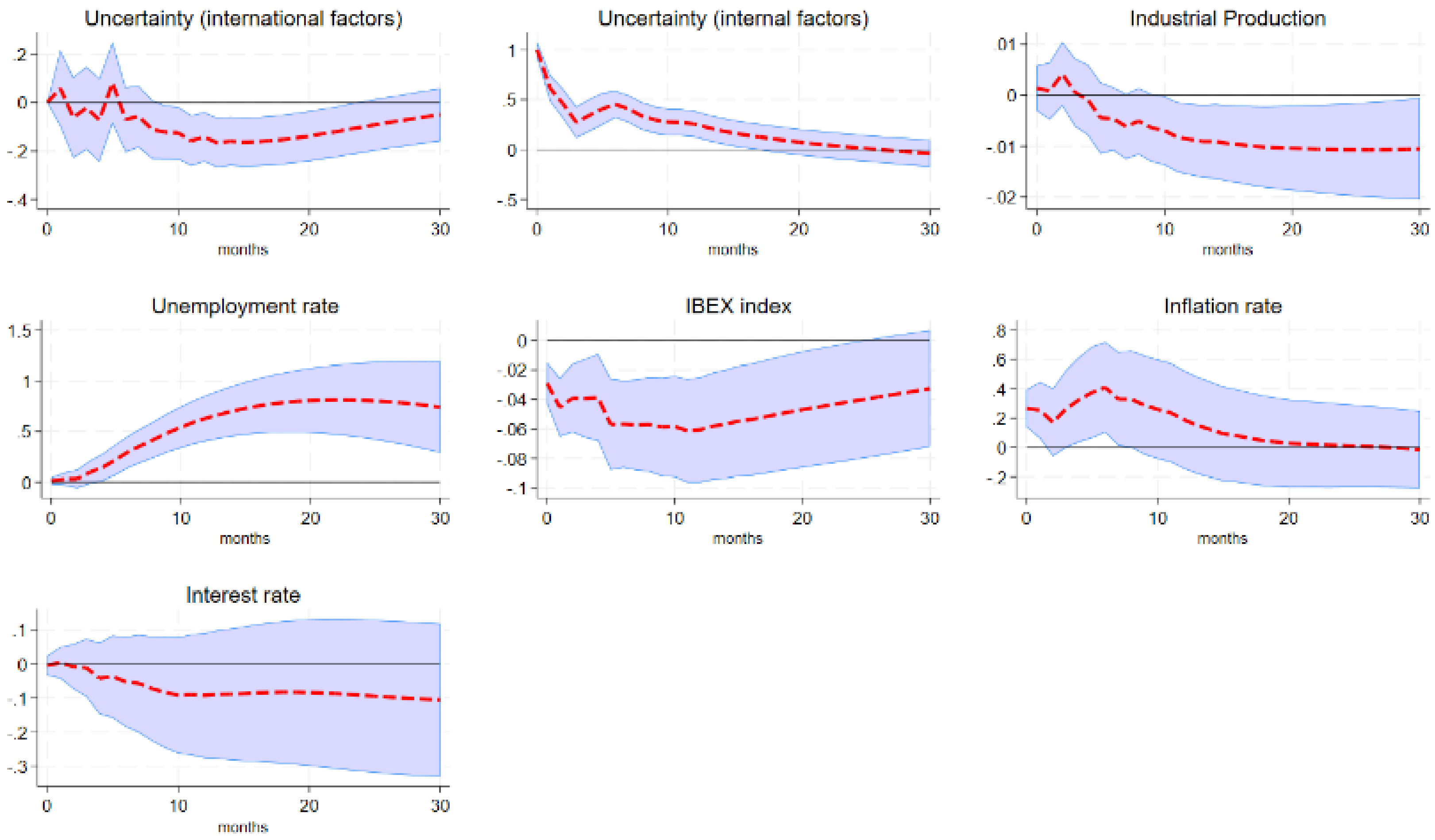


BANCO DE ESPAÑA
Eurosistema

A shock to external uncertainty generates significant responses in output, the stock market and interest rates



A shock to internal uncertainty doesn't affect interest rates, but generates significant responses in output, unemployment, the stock market and inflation



We estimate the relationship between forecast errors and the level of uncertainty when the projections were elaborated, and use it to generate a time-varying fan chart around the projections