EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Domenic Kellner, Jan Hannes Lang,
Lukas Joseph Nagy, Marek Rusnák

A SPOT in the dark: using AI to assess financial stability risks

No 3262

Abstract

Financial stability risks consist of two distinct components: vulnerabilities and possible trigger events. While there has been considerable progress regarding the measurement of vulnerabilities, the assessment of possible trigger events remains largely qualitative. To fill this gap, we employ Large Language Models to extract information about the **S**everity and **P**robability **O**f potential **T**rigger events (**SPOT**) from a large dataset of financial news articles over the period 2005 – 2026. The SPOT indicator increases ahead of major historical trigger events, correctly identifies trigger sources, and helps to improve forward looking model estimates of downside risks to the economy. The results indicate that the use of AI-based signal extraction from text can be a promising avenue to improve the monitoring of financial stability risks.

JEL codes

JEL: C55, C88, E32, E44, G01

Keywords

Financial stability; Artificial intelligence; Crisis indicators; Growth-at-risk;

Non-technical summary

Financial stability risks typically arise from the interaction between underlying macro-financial vulnerabilities and adverse trigger events that can materialise unexpectedly. While substantial progress has been made in monitoring vulnerabilities, the systematic assessment of potential trigger events remains challenging and often relies on qualitative judgement. This paper proposes a novel framework to quantify such trigger events using artificial intelligence (AI) and large-scale text analysis.

We introduce SPOT, a novel AI-based indicator to measure the **S**everity and **P**robability **O**f **T**riggers. SPOT is generated using large language models (LLMs) and a large dataset of financial news articles. The LLM assesses whether an article describes events capable of adversely affecting euro area economic activity or financial stability in the future and then performs a structured assessment along four dimensions: the probability that the event could materialise, the severity of its potential macro-financial impact, the expected time horizon of the impact, and the dominant source of the trigger. These article-level assessments are then aggregated into the SPOT indicator to capture both the frequency and intensity of potential triggers over time.

The resulting SPOT indicator tracks major episodes of financial and macroeconomic stress and provides detailed information on the nature of emerging risks. The benchmark SPOT indicator rises ahead of the global financial crisis, the euro area sovereign debt crisis, the COVID-19 pandemic and periods of heightened geopolitical tensions. The framework also allows for a decomposition of SPOT by trigger source and by attributes such as probability, severity and expected timing, thereby offering a structured and comprehensive view of evolving financial stability risk perceptions and their underlying drivers.

We further show that SPOT provides information beyond existing risk indicators, including measures of financial stress, geopolitical risk and economic policy uncertainty. While correlated with these indicators, SPOT captures broader forward-looking narratives reflected in financial news and incorporates information about the probability and potential impact of adverse events. This additional information proves economically meaningful: incorporating SPOT into a growth-at-risk framework significantly improves the assessment of downside risks to economic growth, particularly when combined with measures of financial vulnerabilities.

Overall, the results demonstrate that AI-based text analysis can enhance financial stability monitoring by systematically extracting forward-looking signals from large volumes of unstructured information.

1 Introduction

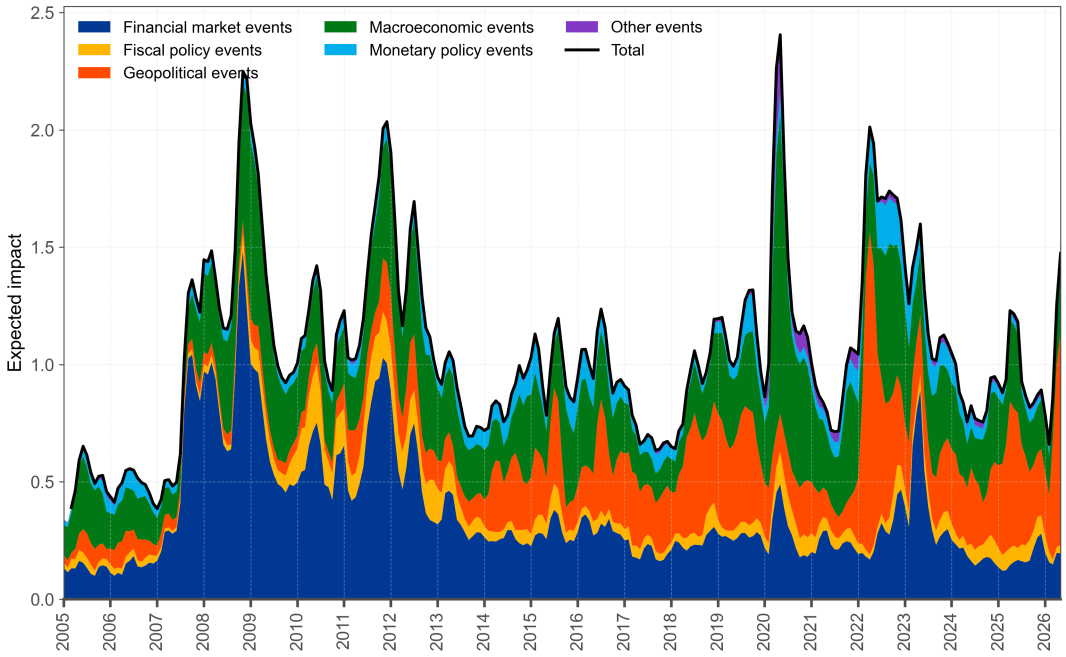
Since the global financial crisis, the analysis of financial stability risks has gained prominence at central banks, supervisory authorities, and international institutions. At a conceptual level, financial stability risks can be broken down into two components. First, there are vulnerabilities which reflect imbalances or fault lines within the financial system or within non-financial sectors that can propagate or amplify shocks, such as excessive leverage, maturity and liquidity mismatches, or asset price misalignments. Second, there are potential triggers which reflect events that could unearth vulnerabilities or cause them to unravel in a disorderly manner, such as bank failures, interest rate shocks, a loss of confidence, wars, or pandemics (Covitz et al., 2015; Fell and Schinasi, 2005; Fell et al., 2024). While there has been considerable progress in measuring vulnerabilities using individual or composite indicators (Alessi and Detken, 2011; Borio and Drehmann, 2009; Borio and Lowe, 2002; Detken et al., 2014; Lang et al., 2019; Schularick and Taylor, 2012; Schüler et al., 2020), the monitoring of potential triggers remains largely qualitative due to the complex and constantly evolving nature of such triggers. Hence, the assessment of financial stability risks and especially of potential triggers is akin to walking barefoot in a dark room filled with broken glass: you are not sure whether the next step is risky or not, due to limited visibility regarding what is lying ahead.

The aim of this paper is to generate an additional SPOT of light into the analysis of financial stability risks, by employing recent advances in Artificial Intelligence (AI) to systematically assess and quantify the potential **S**everity and **P**robability **O**f **T**riggers (**SPOT**) based on news. Specifically, we apply a structured Large Language Model (LLM) prompt to more than 1 million newspaper articles from the *Financial Times* starting in 2005 to classify whether a given article describes a potential future trigger event that could have a severe negative impact on the euro area economy or financial stability, and to assess four key properties of the potential trigger based on the information contained in the article: the probability that the trigger event could materialise, the severity of its potential macro-financial impact, the expected time horizon of the impact, and the dominant source of the trigger. These article-level LLM assessments are then aggregated into the SPOT indicator, which reflects the evolution of the potential **S**everity and **P**robability **O**f **T**riggers over time. More specifically, the SPOT indicator is calculated as the average of the trigger probability multiplied by the trigger severity across all articles at a given point in time.[1] The motivation for our proposed approach is that newspaper articles often discuss potential trigger events and LLMs can be used to extract this information systematically.

[1] An article that is classified as not covering a potential trigger event has a trigger probability and trigger severity of zero.

Our AI-based benchmark SPOT indicator, which reflects the average expected impact of potential trigger events across all articles at a given point in time[2], reaches peaks around the onset of major historical episodes of financial instability or stress, such as the global financial crisis, the euro area sovereign debt crisis, the Covid-19 pandemic, or the Russian invasion of Ukraine (See Figure 1). Moreover, the decomposition of the SPOT indicator by trigger source allows for further insights into underlying drivers and helps form a risk narrative. For instance, financial market triggers dominated during the global financial crisis, other exogenous triggers increased in importance during the COVID-19 pandemic, while geopolitical triggers spiked following Russia's invasion of Ukraine in 2022, and more recently in 2026 due to the war in the Middle East. While the LLM we employ to create the SPOT indicator is trained on data up to October 2023, we mitigate potential look-ahead bias via careful prompt design. First, the system prompt explicitly instructs the LLM not to use outside knowledge, future information or speculation. Second, the article assessment prompt explicitly instructs the LLM to only use information contained in the article as of its publication date for classifying the article and assigning the four key attributes of trigger probability, trigger severity, time horizon and trigger source.

Figure 1: Benchmark SPOT indicator with decomposition by trigger source over time



Notes: The benchmark SPOT indicator is calculated as the average expected impact (probability*severity) of trigger events across all articles at a given point in time. The probability ranges from 0 (unlikely) to 3 (high probability). Severity ranges from 0 (negligible impact) to 3 (very severe impact). Trigger sources are classified as macroeconomic events, financial market events, geopolitical events, monetary policy events, fiscal policy events or other (exogenous) trigger events.

2 The expected impact is equal to the trigger probability multiplied by the trigger severity.

We also use the growth-at-risk modelling approach pioneered by Adrian et al. (2019), i.e. quantile local projections for the lower tail of the future real GDP growth distribution, to formally evaluate the usefulness of our AI-based SPOT indicator for monitoring tail risks. The estimation results from euro area panel growth-at-risk models for 1-quarter and 1-year ahead horizons show that the SPOT indicator significantly improves model fit as measured by the tick loss function, and outperforms other commonly used indicators capturing financial stress (Kremer et al., 2012), geopolitical risk (Caldara and Iacoviello, 2022), policy uncertainty (Baker et al., 2016), uncertainty (Ludvigson et al., 2021), or volatility (Engle and Campos-Martins, 2023). The estimation results also reveal that interacting the SPOT indicator with vulnerability indicators that have been found useful within growth-at-risk models (Lang et al., 2025) further improves the explanatory power of the models. This finding shows that there are important amplification mechanisms at work between vulnerabilities and triggers in driving economic tail risks, and that it is useful to include both of them within growth-at-risk models. Overall, the results presented in this paper suggest that the AI-based SPOT indicator can measure potential trigger events and underlying trigger sources in a systematic way and provide useful information for the monitoring of financial stability risks.

Our paper makes three contributions. First, we contribute to the literature on the use of AI for extracting economically relevant signals from text data. Most papers in this field have used LLMs to extract sentiment scores (e.g. Bond et al. (2023), Lefort et al. (2024), Zhang et al. (2025), or Kwon et al. (2025)), whereas we extract information about specific types of events and their characteristics based on customised definitions that draw on financial stability concepts. Second, we contribute to the literature on constructing economic and financial risk indicators (e.g. Kremer et al. (2012), Baker et al. (2016), Correa et al. (2017), Ludvigson et al. (2021), Caldara and Iacoviello (2022), or Engle and Campos-Martins (2023)), filling a specific gap in the field of financial stability, namely the construction of an indicator to measure the potential **S**everity and **P**robability **O**f **T**riggers. Third, we contribute to the growth-at-risk literature (Adrian et al., 2022, 2019; Aikman et al., 2019, 2018; Chavleishvili et al., 2021; Falconio and Manganelli, 2020; Figueres and Jarocinski, 2020; Galán, 2020; Hartwig et al., 2021; Lang et al., 2025; Plagborg-Moller et al., 2020) by showing that triggers matter alongside vulnerabilities for measuring future downside risks to the economy, and that there are important interactions and amplification effects between vulnerabilities and triggers in driving economic tail risks.

The remainder of the paper is structured as follows. Section 2 discusses the related literature in more detail. Section 3 presents our prompting strategy, the data used, and how the SPOT

indicator is constructed. Section 4 provides an overview of empirical results for the benchmark SPOT indicator and various more granular SPOT versions that can be constructed based on the different extracted trigger attributes. In Section 5 we compare the SPOT indicator to other indicators suggested in the literature, perform the panel growth-at-risk model estimations, and show various robustness exercises. Section 6 concludes.

2 Literature review

Text data potentially contains economically relevant information that can complement traditional numeric data sources. Methods for extracting information from text typically require processing large corpora (Gentzkow et al., 2019) and have benefited substantially from recent advances in computational power and language modeling. Our paper relates to a growing literature that uses text-based methods to construct economic and financial indicators, and in particular to recent work employing Large Language Models (LLMs) for sentiment measurement, uncertainty quantification, and financial stability analysis.

A large body of earlier research constructs economic indicators from text using relatively simple Natural Language Processing (NLP) techniques such as keyword dictionaries, word counts, or bag-of-words approaches. Prominent examples include the Economic Policy Uncertainty (EPU) index by Baker et al. (2016), the Financial Stability Sentiment (FSS) index by Correa et al. (2017), and the Geopolitical Risk (GPR) index by Caldara and Iacoviello (2022). While these methods are transparent and easy to implement, they rely on predefined vocabularies and typically fail to capture semantic context, narrative structure, or evolving economic language (Loughran and McDonald, 2011). Recent work therefore seeks to improve text-based measurement by incorporating contextual language models and neural-network-based classifiers.

Several studies demonstrate the potential of LLMs for generating sentiment-based indicators from financial news. Bond et al. (2023) use ChatGPT to construct sentiment-based market indicators from daily news summaries. Lopez-Lira and Tang (2023) show that GPT-based models can extract economically meaningful signals from news headlines that predict market reactions. Similarly, Lefort et al. (2024) apply LLMs to financial news headlines to derive sentiment scores for NASDAQ index predictions. Zhang et al. (2025) propose *FinSentLLM*, which combines an ensemble of LLMs with structured financial signals and expert-driven semantic cues to improve financial sentiment forecasting. Their approach captures inter-model agreement and domain-specific information and yields improvements in predictive accuracy and long-run comovement

with major stock indices.

Beyond sentiment scores, Nyman et al. (2021) show that changes in the coherence and emotional structure of financial narratives systematically precede episodes of financial instability. Their findings support the view that news text embeds forward-looking information about systemic risk, providing a conceptual foundation for event-based monitoring frameworks. While this work provides an important conceptual foundation, monitoring of macro-financial risks often requires more compressed and easily interpretable indicators. Kwon et al. (2025) construct sentiment indices for macroeconomic growth and inflation from a large set of U.S. news articles. Their approach directly extracts economic narratives from press coverage and decomposes sentiment into interpretable drivers such as demand and supply forces. The resulting indicators closely track conventional macroeconomic measures and improve forecasting performance when incorporated into predictive models, suggesting that text-based sentiment captures information not contained in traditional data sources.

A related strand of literature applies LLMs to measure economic uncertainty and policy-related risks. Audrino et al. (2024) introduce an LLM-based framework for measuring economic uncertainty from newspaper texts that exploits contextual language understanding to classify sources of uncertainty across domains such as geopolitics, economic policy, monetary policy, and financial markets. The resulting indices exhibit stronger correlations with macroeconomic developments and financial market volatility than traditional benchmarks. Similarly, Chen et al. (2024) revisit the construction of the Economic Policy Uncertainty index by showing that keyword-matching approaches neglect semantic context and introduce noise. By replacing keyword matching with neural-network classifiers, they improve predictive performance and out-of-sample accuracy.

Our paper contributes to this literature by proposing a structured LLM-based framework to identify potential financial stability trigger events from news in a forward looking manner and to assess their probability, severity, timing, and source. While existing work primarily focuses on sentiment or uncertainty measurement using LLMs, we construct SPOT indicators that explicitly capture potential trigger events for financial stability risks and examine their interaction with financial vulnerabilities within a growth-at-risk framework.

3 SPOT methodology and data

This section describes the construction of the SPOT indicator based on a large set of *Financial Times* (FT) articles that are evaluated in a structured way using the LLM GPT-4o-mini. We first present the three-stage LLM prompting pipeline to identify articles that discuss trigger events and to extract information on the probability, severity, expected horizon, and source of the potential trigger events. We then introduce the characteristics of the underlying dataset of newspaper articles and present the results from the article-level LLM assessments. Finally, we describe the aggregation procedure to transform article-level classifications and assessments into SPOT indicators that are suitable for financial stability monitoring and empirical analysis.

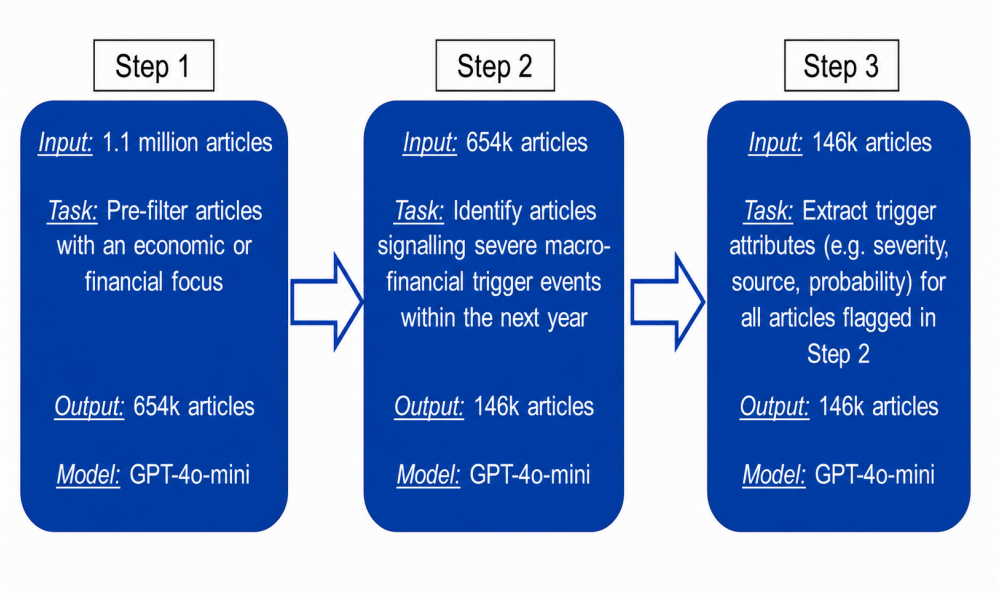
3.1 Prompting strategy

The proposed prompting framework decomposes the assessment of each newspaper article into three sequential steps (Figure 2). First, to identify articles that cover economic or financial content. Second, to detect articles that discuss forward-looking trigger events. Third, to extract key attributes of the potential trigger events. Each step represents a different prompt that we ask the LLM to evaluate via an API. This modular prompt design mitigates context window limitations, reduces cognitive load on the model, and improves the consistency and interpretability of the resulting classifications. The prompting framework relies on rigorously specified instructions that define key economic and financial stability concepts, enforce forward-looking discipline, and structure model outputs. In addition, the model is explicitly instructed to use only information contained in the article as of its publication date and to avoid using external knowledge, future information, or speculation. This restriction mitigates information contamination and helps to ensure that the resulting classifications reflect real-time information contained in the newspaper article, consistent with the risk monitoring objective. The full documentation of all three prompting steps can be found in Appendix A.1.

In the first step, all newspaper articles are pre-filtered to identify whether the primary focus is economic or financial and whether the economic or financial discussion is explicit, substantive, and central to the article. Economic and financial content is defined to include macroeconomic variables (such as inflation, GDP, interest rates, or unemployment), financial markets or instruments (including equities, bonds, spreads, banks, or non-bank financial institutions), or economic policy with a direct and concrete economic impact. Articles concerning individual firms are retained only if firm-specific developments are described as having implications for the broader economy or financial system. Articles that are strictly non-economic or non-financial, or that use economic terminology only metaphorically or without analytical substance, are ex-

cluded. When the classification is ambiguous, the model is instructed to conservatively exclude the article. This pre-filtering step removes unrelated content (e.g. articles only covering politics, art, travel, etc.) and ensures that subsequent analysis focuses on articles with relevant information.

Figure 2: Three-stage prompting pipeline



Notes: Each stage represents a different prompt that is sent to the LLM to evaluate via an API. The prompts can be found in A.1

In the second step, the pre-filtered articles are classified whether they signal current trigger events or potential future trigger events that could have a severe negative impact on economic activity or the stability of the financial system within the euro area over a 1–12 month horizon ($Class = 1$). A severe negative impact on economic activity is defined as negative real GDP growth, significant declines in industrial production, or substantial increases in unemployment, while a severe negative impact on financial stability includes banking-sector losses, significant increases in interbank lending rates, or large outflows of bank deposits. The classification explicitly adopts a forward-looking perspective, identifying statements about forecasts, expectations, warnings, or predictions concerning potential future developments affecting the euro area. Articles describing only current or past developments without credible future implications for euro area economic activity or financial stability are classified as non-trigger events. Particular emphasis is placed on euro area relevance: articles concerning non-euro area developments or firm-specific events are classified as trigger events only if plausible spillovers to euro area economic activity or financial stability within the specified horizon are indicated or strongly implied.

In the third step, articles identified as signaling potential trigger events are further characterised along several dimensions capturing the probability ($Prob$), severity (Sev), time horizon, and source of the potential trigger. To provide an additional consistency check for the second stage, the third stage explicitly allows zero-value categories, thereby giving the model an exit option even when an article has been previously classified as signaling a trigger event. This design acts as a form of double verification, allowing the model to revise earlier assessments if the evidence for a genuine macro-financial risk trigger is weak upon closer evaluation.

Probability is measured on a four-point scale ranging from 0 (unlikely, implying effectively zero probability) to 3 (high probability, indicating a very plausible outcome likely to materialise), with intermediate values representing low and medium probability. Severity is also measured on a four-point scale, where 0 indicates negligible or limited effects on output, employment, or financial stability; 1 indicates mildly severe outcomes such as contained sectoral stress or limited financial tightening; 2 indicates severe outcomes involving broad-based macroeconomic or financial disruption; and 3 indicates very severe or systemic impacts, including banking distress, sovereign funding crises, or widespread economic contraction.[3]

Lastly, the model also assigns a time horizon indicating when the impact is most likely to materialise, distinguishing between impacts that have already materialised, impacts expected within 1–3 months, 3–6 months, 6–12 months, and 12–36 months. Finally, the model identifies the dominant source of the trigger event by assigning the article to one of several predefined categories. These include macroeconomic events (such as demand disturbances, commodity price shocks, or exchange rate adjustments), financial market events (including liquidity shortages, banking-sector instability, or reassessments of asset risk), geopolitical events (such as wars, political instability, trade conflicts, or sanctions), monetary policy events, fiscal policy events, or other (exogenous) shocks such as natural disasters or pandemics. A few examples demonstrating how the LLM classifies and assesses selected articles based on the structured prompts are shown in Table 1.

Look-ahead bias is a potential concern when LLMs are used to generate historical forecasts, as the model may have been exposed during training to information about events that occurred after the forecast date (see e.g. Sarkar and Vafa, 2024; Glasserman and Lin, 2024). In our setting

[3] We experimented with alternative scaling approaches, including continuous numerical scales and coarser categorical classifications. In practice, a four-point categorical scale provided the best balance between interpretability, stability, and discriminatory power. Continuous scales produced less stable outputs, while coarser classifications compressed the distribution of assessments excessively. This finding is consistent with recent evidence suggesting that large language models perform more reliably in ordinal classification tasks than in precise numerical calibration exercises (see, e.g., Qin et al., 2024; Zhao et al., 2023).

Table 1: Example outputs from the prompting pipeline

Article title (including link)	Published date	Stage 1: Economic and financial relevance	Stage 2: Trigger-event classification	Stage 3: Trigger-event characterisation			
		filter_label	Class	Prob	Sev	Hor	Trigger Source
Ten years of social media have left us all worse off	27/12/2019	0	NA	NA	NA	NA	NA
Strong expansion in eurozone business activity boosts hiring	23/08/2021	1	0	NA	NA	NA	NA
Donald Trump's sweeping tariffs ignite $2.5tn rout on Wall Street	03/04/2025	1	1	3	3	1	Geopolitical

this concern is mitigated by the fact that the model is not instructed to forecast future outcomes directly, but to classify and assess the information contained in each article using the article text and publication date as the relevant information set. This makes the task conceptually closer to text annotation than to direct forecasting, in line with the emerging literature using LLMs for automated text classification and annotation tasks (e.g. Törnberg, 2024; Gilardi et al., 2023). To further mitigate potential information leakage, the system prompts in Stages 2 and 3 explicitly instruct the model to base its assessment solely on the information contained in the article as of its publication date and not to use outside knowledge, future information, or speculation. Such prompt constraints are commonly used in the LLM annotation literature to anchor model outputs to the provided text and reduce hallucinations or the incorporation of external information. The full prompts are reported in Appendix A.1. Robustness checks related to the LLM-choice and training cut-off dates are discussed in Section 5.3.

Overall, the three-step prompting pipeline enables scalable and systematic identification of potential trigger events from a large set of newspaper articles, while maintaining conceptual clarity. The modular structure further enhances operational flexibility, as individual steps can be updated independently in case one wants definitions or emerging trigger source categories to be adjusted. In particular, revisions to the attribute characterizations can be implemented by modifying only the final prompting stage without re-running the entire pipeline, which facilitates continuous monitoring and regular updating of risk assessments. The sequential prompting design also improves computational efficiency by restricting more demanding classification tasks to a progressively smaller subset of relevant articles.

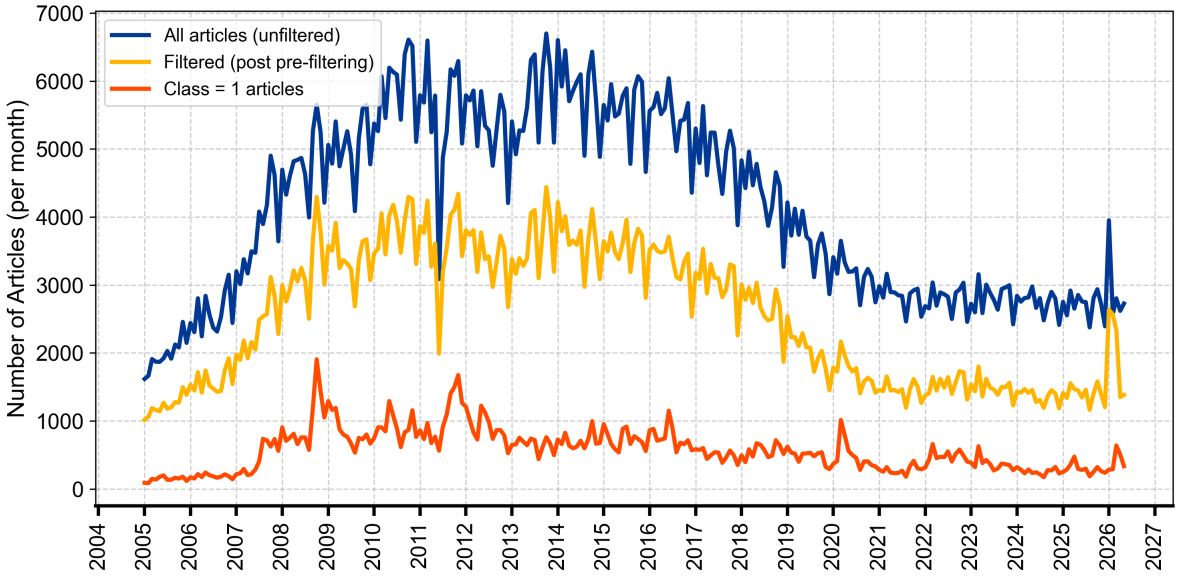
3.2 Dataset

We apply the three-stage prompting framework described above to a large dataset of newspaper articles from the *Financial Times* (FT). The FT serves as our data source due to its consistent high-quality focus on economic and financial topics in Europe, which ensures thematic

coherence and reduces the need for extensive pre-screening of unrelated content. The dataset covers articles published between January 2005 and May 2026 and provides broad coverage of macroeconomic and financial developments relevant for financial stability analysis. For each article, we extract the title, body text, and publication date, which serve as inputs to the LLM classification procedure. The publication date ensures that model assessments rely only on information available at the time of publication (as instructed in the LLM system prompt), thereby mitigating potential look-ahead bias and supporting the real-time risk monitoring purpose of the framework.

The raw dataset comprises approximately one million articles, with an average length of around 600 words and roughly 4,200 articles published per month (Figure 3). Overall data quality is high. Articles with missing body text or incomplete metadata are excluded, and duplicate entries are removed from the sample. The number of available articles varies over time, reflecting changes in archiving practices, topic coverage, and distribution rights that affect data availability. This variation underscores the importance of constructing indicators based on article content rather than relying solely on article counts, a potential limitation of many conventional dictionary-based approaches.

Figure 3: Number of Articles over time for the three prompting stages



Notes: The blue line shows the number of articles per month before the first stage i.e., the raw data. The yellow line shows the number of articles before the second stage and after pre-filtering articles for their economic relevance. The orange line shows the number of articles per month classified as $Class1$ in the second stage, that are passed to the stage three prompt.

The three-stage prompting procedure sequentially reduces the number of articles are pro-

cessed in each step (Figure 2 and Table A.1 in the Appendix). After initial data cleaning, approximately 1.065 million articles enter the first stage of the pipeline, where the pre-filter removes articles without a primary economic or financial focus (around 39% of the sample). The second stage evaluates the remaining 654 thousand articles for forward-looking trigger events relevant for euro area economic activity or financial stability and excludes a further 78% of articles that do not cover trigger events. In total, approximately 146 thousand newspaper articles are retained for the final stage, in which the potential trigger events described in the articles are characterised along several dimensions such as probability, severity, time horizon and trigger source (Figure 3).

Table 2: Marginal distribution of extracted attributes (Class 1 articles)

Attribute	Category	Total (N)	Share (%)
Trigger source	Financial market events	51,340	35.1
	Macroeconomic events	45,177	30.9
	Geopolitical events	27,002	18.5
	Fiscal policy events	10,984	7.5
	Monetary policy events	9,646	6.6
	Other events	2,001	1.4
Probability	Unlikely / effectively zero probability (0)	43	0.0
	Low probability (1)	4,280	2.9
	Medium probability (2)	95,094	65.1
	High probability (3)	46,733	32.0
Severity	Negligible / limited effects (0)	74	0.1
	Mildly severe (contained stress) (1)	10,828	7.4
	Severe (broad disruption) (2)	115,009	78.7
	Very severe / systemic (3)	20,239	13.8
Horizon	Already materialising (0)	5,441	3.7
	Within 1–3 months (1)	85,114	58.2
	Within 3–6 months (2)	27,533	18.8
	Within 6–12 months (3)	27,782	19.0
	Within 12–36 months (4)	280	0.2

Notes: Shares are computed relative to all Class 1 articles. Probability and severity are measured on a 0–3 scale and horizon on a 0–4 scale.

The marginal distributions of the extracted attributes from the third stage prompt are reported in Table 2. 65% and 79% of trigger articles are assigned a medium probability and severity respectively. High-probability and high-severity trigger events are identified in 32% and 14% of the relevant articles, reflecting the rarity of crisis level triggers in news coverage. As a robustness feature of the prompting design, in the third stage the model is allowed to assign category 0 for probability and severity even for articles previously identified as trigger events. Such cases are rare ($\leq 0.1\%$), confirming the validity of the step 2 classification. For the

horizon attribute, category 0 indicates that the adverse impact is already materialising rather than expected in the future. This occurs in only 3.7% of trigger articles and is most likely due to the fact that during crisis periods the distinction between triggers being contemporaneous rather than forward-looking becomes somewhat blurry. More than half of the potential triggers are assessed to have an impact within 1-3 months. In terms of trigger sources, around 35% and 31% of triggers relate to financial market and macroeconomic events respectively, with a further 18.5% related to geopolitical events.

Table 3: Joint distribution of attributes (row-wise shares, %, Class 1 articles)

		Probability				Severity				Horizon					Trigger source					
Attribute	Category	0	1	2	3	0	1	2	3	0	1	2	3	4	Macro	FinMkt	Geo	MP	Fiscal	Other even
Probability	0					100	0	0	0	98	0	0	0	2	74	9	0	0	0	16
	1					1	98	1	0	1	22	17	57	4	42	27	11	6	11	3
	2					0	7	93	0	0	49	25	26	0	34	31	18	7	8	1
	3					0	0	58	42	11	80	7	2	0	23	45	20	6	6	1
Severity	0	58	42	0	0					59	0	0	14	27	73	9	0	5	0	12
	1	0	39	61	0					0	35	20	43	2	43	26	12	8	10	2
	2	0	0	77	23					1	58	21	20	0	32	33	19	7	7	1
	3	0	0	2	98					20	74	4	2	0	16	53	21	2	7	1
Horizon	0	1	1	3	96	1	1	24	75						23	61	6	1	5	2
	1	0	1	55	44	0	4	78	18						26	40	18	7	8	1
	2	0	3	86	11	0	8	89	3						41	27	18	6	7	1
	3	0	9	88	4	0	17	82	1						39	23	23	6	8	2
	4	0	68	22	9	7	65	22	5						44	24	11	4	10	6
Trigger source	Macro	0	4	72	24	0	10	82	7	3	48	25	24	0						
	FinMkt	0	2	57	41	0	5	74	21	6	67	15	12	0						
	Geo	0	2	64	34	0	5	79	16	1	57	18	23	0						
	MP	0	3	70	28	0	9	87	5	1	64	18	17	0						
	Fiscal	0	4	70	26	0	9	78	12	3	59	17	21	0						
	Other even	0	7	67	26	0	11	75	14	7	54	14	24	1						

Notes: Entries are row-wise shares (in %). Within each row, reported blocks sum to 100 (the block corresponding to the row attribute is left blank to avoid redundancy). Probability is measured on a four-point scale from 0 (unlikely, effectively zero probability) to 3 (high probability). Severity is measured on a four-point scale from 0 (negligible impact) to 3 (very severe or systemic impact). The horizon indicates the expected timing of the impact: 0 = already materialising, 1 = within 1–3 months, 2 = within 3–6 months, 3 = within 6–12 months, and 4 = within 12–36 months. Trigger source categories are abbreviated as follows: Macro (macroeconomic events), FinMkt (financial market events), Geo (geopolitical events), MP (monetary policy events), Fiscal (fiscal policy events), and Other (exogenous events).

The joint distributions of extracted trigger attributes in Table 3 provide further insights. Slightly less than half of the high probability triggers are also high severity triggers. On the other hand, almost all of the high severity triggers also have a high probability of materialising. Around 90% of the high probability and high severity triggers are associated with short horizons ($\leq$ 3 months). The occurrence of high probability and high severity triggers is therefore concentrated during periods of imminent stress. Conversely, the model rarely assigns high probability or high severity to triggers with distant horizons ($>$ 3 months), which is consistent with the notion that news coverage of distant triggers will often be more speculative and vague when it comes to their potential occurrence and impact. Instead, medium probability and severity are predominant for triggers with horizons of 3 to 12 months. The distributions of severity, probability, and horizon characteristics are rather similar across the different trigger sources. However, short horizon triggers are dominated by financial market events (40%),

whereas medium to longer horizon triggers are dominated by macroeconomic events (Table 3).

3.3 Construction of SPOT indicators

The classification into trigger articles ($Class = 1$) and non-trigger articles ($Class = 0$) from prompting step 2 and the extracted attributes from prompting step 3 can be used to create various SPOT indicators. Before presenting these indicators, it is useful to define some notation. Let N_t denote the total number of economic or financial articles at a given point in time t, and let N_t^{C1} and N_t^{C0} denote the number of $Class1$ and $Class0$ articles respectively, where $N_t = N_t^{C1} + N_t^{C0}$. Whenever we sum over the entire set of articles we use the index i and whenever we sum over a subset of articles we use the index j. By default, whenever an article is classified as $Class0$ in prompting step 2, all of the attributes that are extracted in prompting step 3 are automatically set to zero. Based on these notation conventions we can easily define the following SPOT indicators.

Trigger Frequency Indicator. The share of articles signaling potential trigger events:

$$\overline{\text{Freq}}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Class_i = \frac{1}{N_t}\sum_{j=1}^{N_t^{C1}} Class_j = \frac{N_t^{C1}}{N_t} \tag{1}$$

Average Probability Indicator. The average probability of triggers across all articles:

$$\overline{\text{Prob}}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Prob_i = \frac{1}{N_t}\sum_{j=1}^{N_t^{C1}} Prob_j \tag{2}$$

Average Severity Indicator. The average severity of triggers across all articles:

$$\overline{\text{Sev}}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Sev_i = \frac{1}{N_t}\sum_{j=1}^{N_t^{C1}} Sev_j \tag{3}$$

Average Expected Impact Indicator (Benchmark SPOT indicator). This measure captures both the frequency and intensity of trigger signals:

$$\overline{\text{Impact}}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Prob_i \times Sev_i = \frac{1}{N_t}\sum_{j=1}^{N_t^{C1}} Prob_j \times Sev_j \tag{4}$$

Decompositions of Indicators. The SPOT indicators above can also be systematically decomposed by a certain attribute A (e.g. by trigger source, time horizon, probability, or severity), where N^A denotes the number of different attribute types and $N_t^{C1|a}$ denotes the number of trigger articles associated with a specific attribute type a. Such decompositions allow for a more detailed analysis of the nature of triggers and can help form a coherent risk narrative:[4]

$$\overline{\text{Measure}}_t = \sum_{a=1}^{N^A} \left(\frac{1}{N_t} \sum_{j=1}^{N_t^{C1|a}} Measure_j \right) \tag{5}$$

Sectoral SPOT Indicators. Frequency, average probability, severity, and expected impact can also be calculated for different trigger sources S, where $N_t^{C1|S}$ denotes the number of trigger articles associated with a specific trigger source:

$$\overline{\text{Measure}}_t^S = \frac{1}{N_t} \sum_{j=1}^{N_t^{C1|S}} Measure_j \tag{6}$$

Trigger Article Indicators. Average probability, severity, and expected impact can also be calculated conditioning on trigger articles only:

$$\overline{\text{Measure}}_t^{C1} = \frac{1}{N_t^{C1}} \sum_{j=1}^{N_t^{C1}} Measure_j \tag{7}$$

The default time aggregation of articles into SPOT indicators is done at a monthly frequency. To smooth short-term fluctuations and enhance interpretability, all indicators are transformed into three-month moving averages.

4 Empirical results

This section provides an overview of the empirical properties of the various SPOT indicators described in subsection 3.3.

[4] The decomposition of expected impact by probability/severity is not possible. For SPOT indicators representing the probability or severity, the decompositions can be conducted along the complementary dimension.

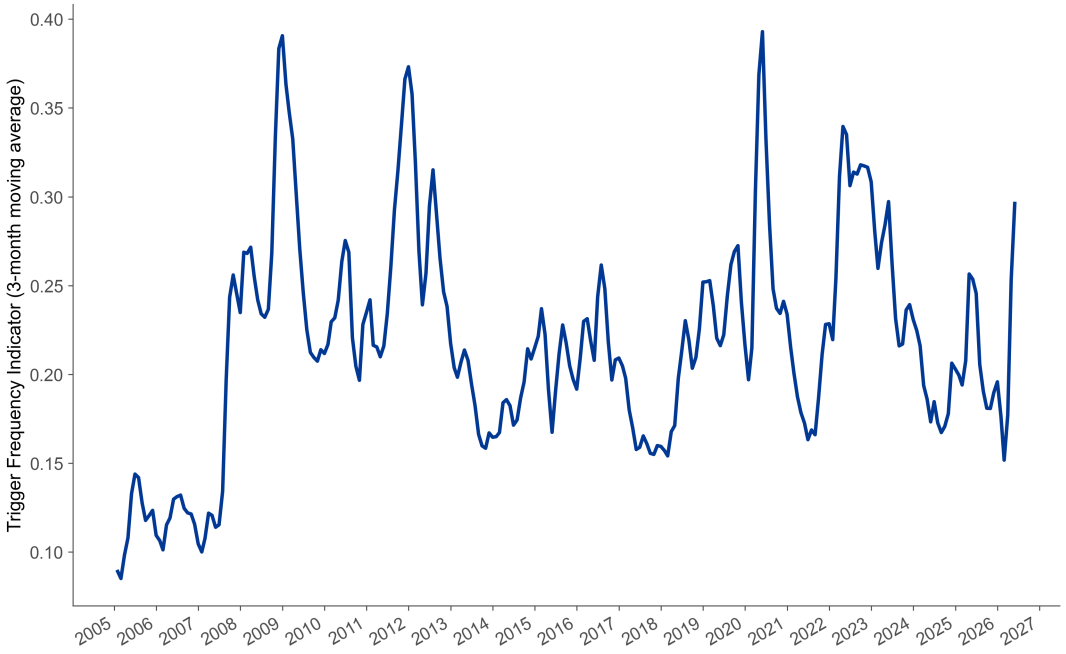
4.1 Aggregate SPOT indicators and decompositions

The first observation is that the frequency of trigger articles varies considerably over time, between 10% and close to 40% (Figure 4 panel a). The largest peaks in the frequency of trigger articles are reached during the global financial crisis in 2008, and the Covid-19 crisis in 2020. The second observation is that while the average probability and average severity of trigger articles display a high positive correlation, they also seem to contain complementary information (Figure 4 panel b). For example, the amplitude of the average probability is almost twice as large as the amplitude of the average severity. Moreover, there are time periods when these two attributes of potential triggers show different dynamics, for example during 2025 and 2026. As shown in section 5.2 below, combining information about the frequency, probability, and severity of trigger articles improves model estimates of future downside risks to the economy. Hence, our benchmark SPOT indicator (the average expected impact) combines these three dimensions.

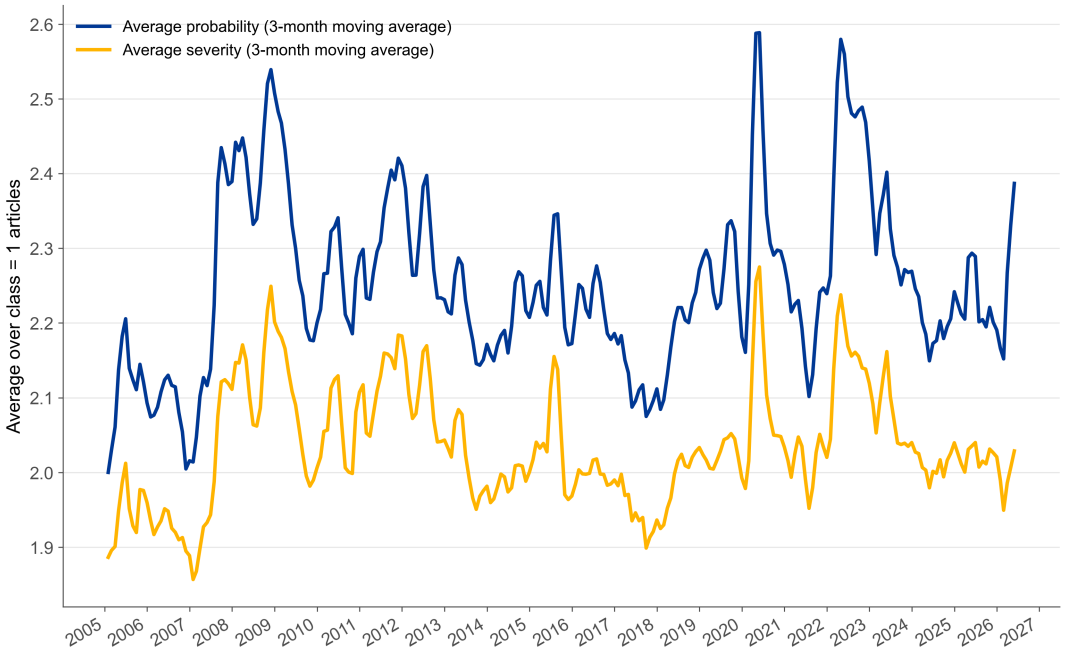
The benchmark SPOT indicator increases ahead of major historical trigger events and correctly identifies the relative importance of different trigger sources, as shown in Figure 5. For example, it reaches peaks around the onset of the global financial crisis in 2008, the euro area sovereign debt crisis in 2011, the Covid-19 pandemic in 2020, and the Russian invasion of Ukraine in 2022 (Figure 5). The decomposition of the benchmark SPOT indicator into trigger sources allows for further insights into underlying drivers and helps form a risk narrative. For instance, financial market triggers dominated before and during the global financial crisis and the euro area sovereign debt crisis, while other exogenous triggers increased in importance during the COVID-19 pandemic (Figure 5). More recently, geopolitical triggers spiked following Russia's invasion of Ukraine in 2022 and again during 2026 amidst heightened geopolitical tensions. It is important to note that the most recent spike in the SPOT indicator in 2026 happens well beyond the training cut-off date for the LLM. Overall, these patterns suggest that the proposed SPOT indicator can measure potential trigger events in a systematic way and provide meaningful information about underlying drivers.

Figure 4: Evolution of aggregate SPOT indicators over time

(a) Frequency of trigger articles

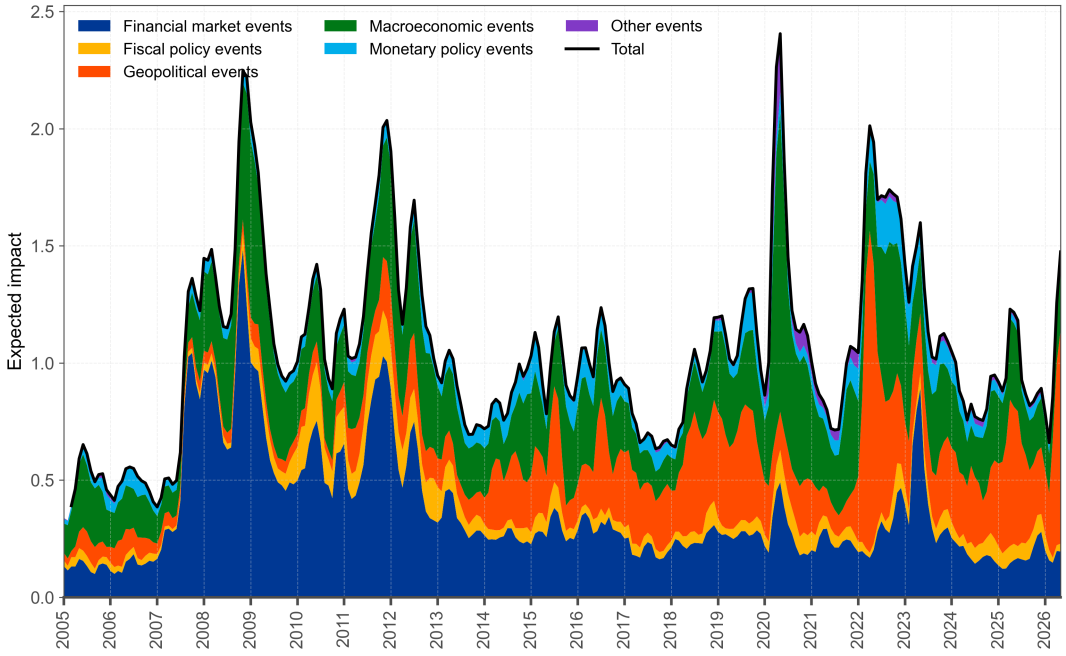


(b) Average probability and average severity of trigger articles



Notes: The trigger frequency indicator shows the share of Class1 articles over the sum of Class0 and Class1 articles. The average probability and severity measures are shown here as averages across only Class1 articles.

Figure 5: Benchmark SPOT indicator with decomposition by trigger source over time

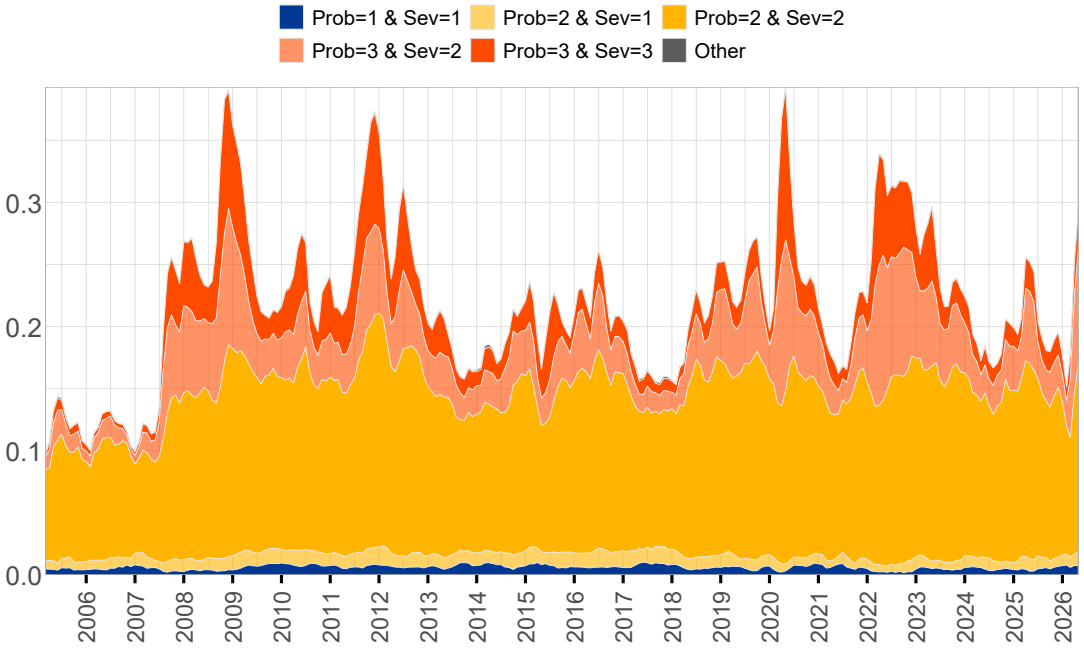


Notes: The benchmark SPOT indicator is calculated as the average expected impact (probability*severity) of trigger events across all articles at a given point in time.

The decomposition of the trigger frequency indicator by the severity and probability of the potential triggers allows for further insights (Figure 6). Across the full sample, events with a "medium probability" and a "medium severity" account for the largest share of classified trigger articles, reflecting a persistent baseline of concerns about potential trigger events in news coverage.[5] During major stress episodes like the global financial crisis, the euro area sovereign debt crisis, the COVID-19 pandemic, and the Russian invasion of Ukraine, the composition of classified trigger articles shifts considerably: the share of trigger articles with a high severity rating and/or a high probability rating jumps up, indicating that news narratives intensify. It is worth noting that the war in the Middle East in 2026 mainly led to an increase in the share of high probability trigger articles, rather than also high severity articles. These patterns suggest that monitoring the probability–severity distribution of triggers can capture variation in perceived risk beyond what the trigger frequency alone would convey.

5 Such trigger articles are assigned a probability = 2 and a severity = 2.

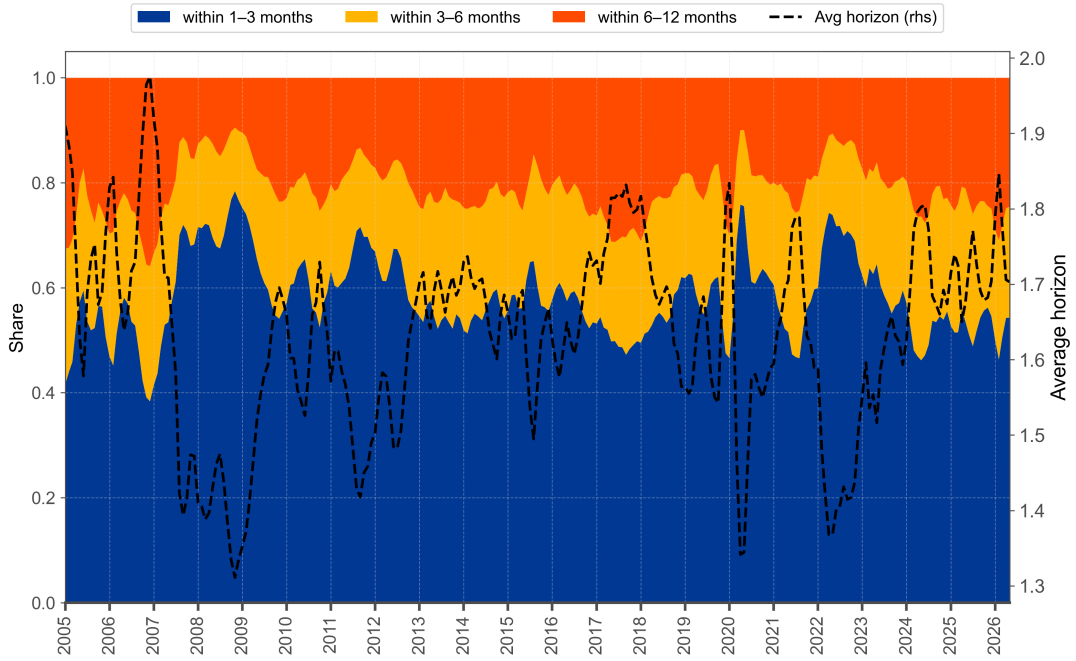
Figure 6: Decomposition of trigger frequency indicator by probability and severity



Notes: The chart shows the frequency of trigger articles by probability and severity classes. Probability is measured on a four-point scale from 0 (unlikely, effectively zero probability) to 3 (high probability). Severity is measured on a four-point scale from 0 (negligible impact) to 3 (very severe or systemic impact). Values shown are 3-month moving averages.

Beyond the frequency, probability and severity of potential trigger events, the SPOT indicator can also capture a time-horizon dimension. The decomposition of identified trigger articles in Figure 7 shows that trigger-related news coverage is concentrated at short horizons. More than half of trigger articles (58%) are assigned to the short-term bucket (one to three months), with the remainder split roughly equally between the medium-term (three to six months) and longer-term buckets (six to twelve months), as also shown in Table 2. However, the composition of trigger horizons varies markedly over time. The weighted average horizon of trigger articles tends to shorten ahead of and during major stress episodes, which is consistent with the notion that, as trigger events become more imminent, media coverage places greater emphasis on near-term adverse developments. From a surveillance perspective, the SPOT horizon provides a complementary signal on the timing of potential triggers, alongside SPOT signals about their perceived probability and severity.

Figure 7: Share of trigger articles by time horizon of potential impact



Notes: The horizon indicates the expected timing of the impact: 0 = already materialising, 1 = within 1–3 months, 2 = within 3–6 months, 3 = within 6–12 months, and 4 = within 12–36 months. Horizons 0 and 4 were dropped from the chart as they represent a small fraction of classified trigger articles and the former is also inconsistent with the definition of forward looking triggers. Average horizon is expressed in terms of the horizon category (ranging from 1 to 3) and not in terms of months.

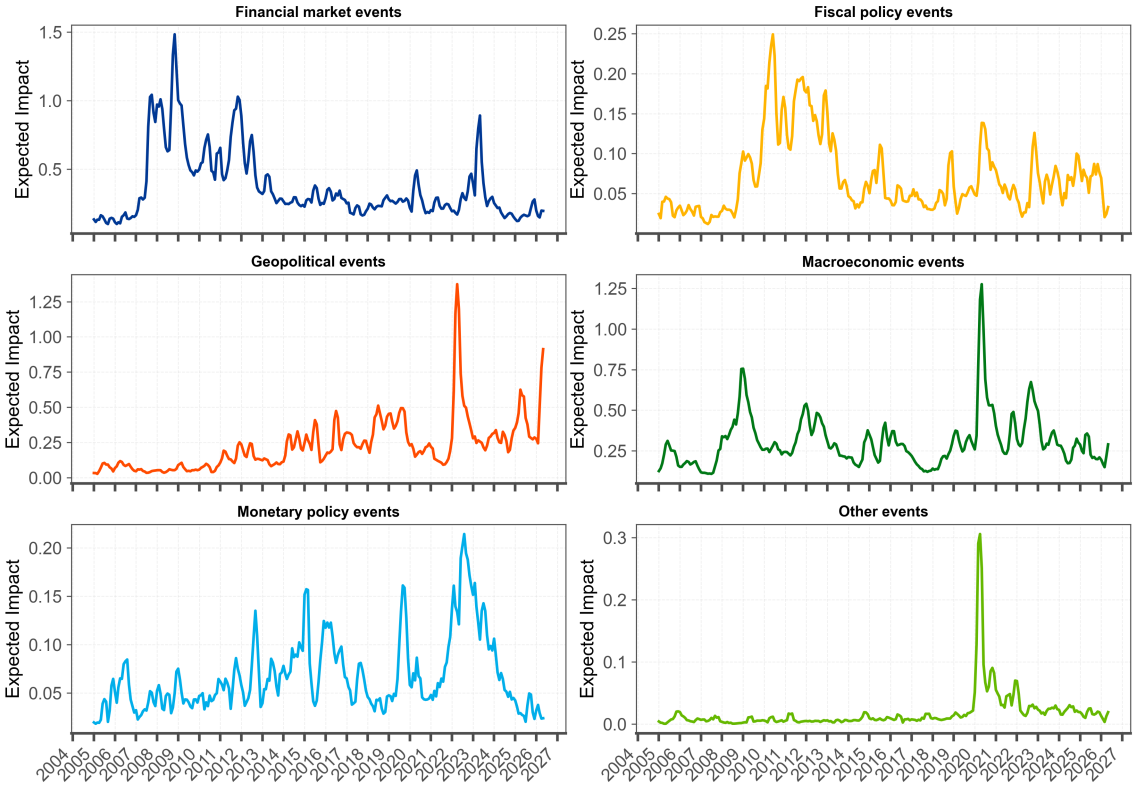
4.2 Granular SPOT indicators

Building on the aggregate SPOT decompositions, this subsection further zooms in on four granular dimensions: (i) source-specific SPOT indicators, (ii) higher-frequency SPOT indicators, (iii) horizon-specific SPOT indicators, and (iv) SPOT indicators conditional on probability and severity thresholds.

While the decomposition of the benchmark SPOT indicator into trigger sources can help identify drivers and form an overall risk narrative, sectoral SPOT indicators for individual trigger sources make the underlying drivers even more visible and can be used to complement the benchmark SPOT indicator for monitoring specific trigger sources (Figure 8). For example, the sectoral SPOT indicator for fiscal triggers shows clear peaks in 2010 and 2011, which are less visible in the aggregate decomposition chart (Figure 5). The important role of other exogenous triggers during the Covid-19 pandemic is also much more visible when looking at the sectoral SPOT indicators. The sectoral SPOT indicator for geopolitical triggers also helps to identify a clear upward trend over the past 15 years in the importance of such trigger events and shows a clear spike at the current end of the sample in 2026 (Figure 8). Hence, the sectoral SPOT

indicators can provide useful complementary information to the benchmark SPOT indicator.

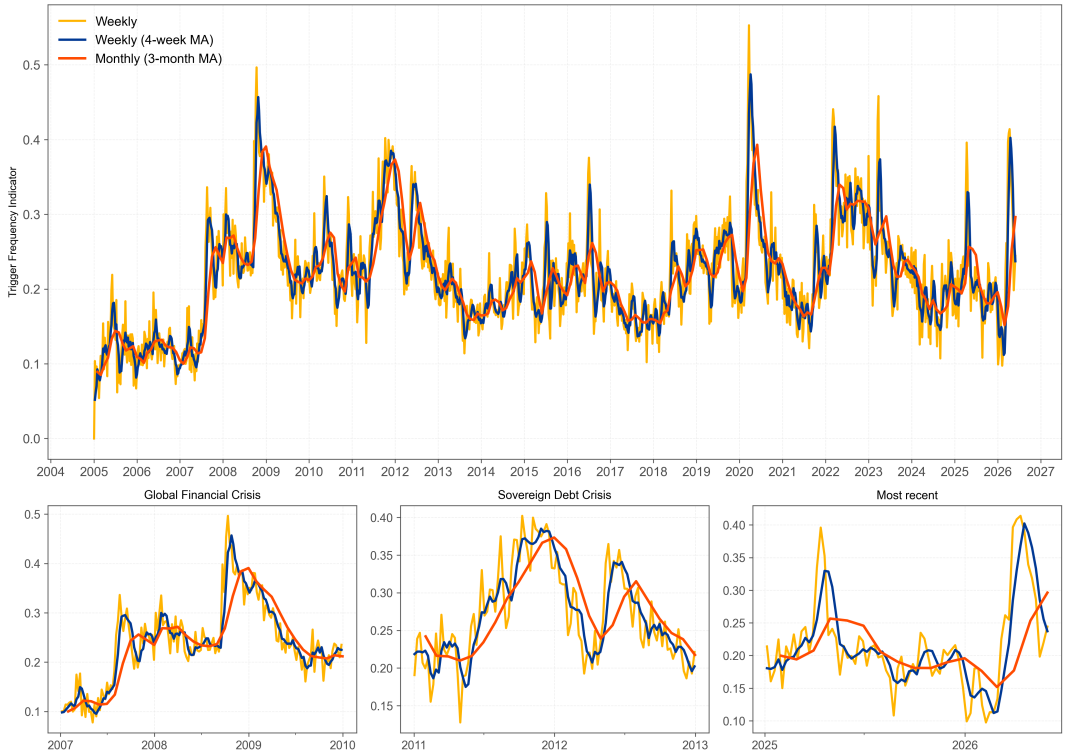
Figure 8: Expected impact indicators for different trigger sources



Notes: The expected impact is calculated across both classes as explained in subsection 3.3.

SPOT indicators can also be constructed at different time aggregations. While the benchmark indicator is calculated as a 3-month moving average of monthly values, higher-frequency measures can provide more timely signals when new triggers emerge. Weekly SPOT indicators capture short-term fluctuations in news coverage of triggers and can signal emerging financial stability risks earlier, whereas monthly and quarterly aggregations smooth temporary movements and highlight more persistent developments in potential trigger events. For example, at the onset of the global financial crisis, the euro area sovereign debt crisis, and Russia's invasion of Ukraine, higher-frequency SPOT indicators increased earlier than lower-frequency time aggregations, as shown in Figure 9. However, these more timely signals come at the cost of increased volatility, implying a trade-off between responsiveness and noise. Monitoring different time aggregations of SPOT can therefore be useful.

Figure 9: Trigger frequency indicators for different time aggregations

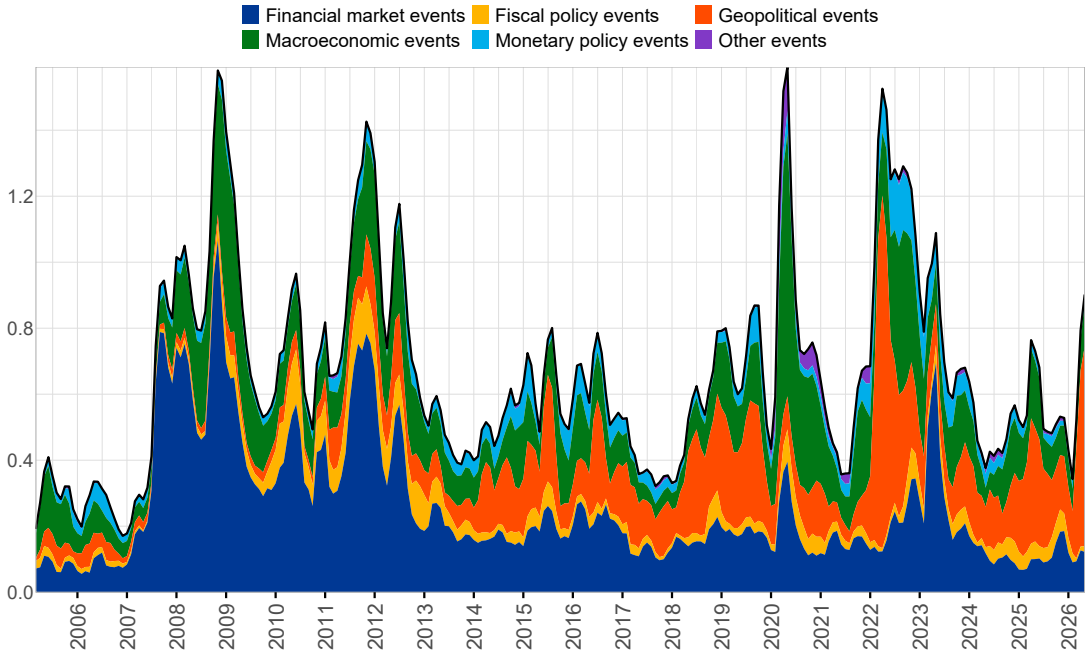


Notes: The trigger frequency indicator is calculated across both classes as explained in subsection 3.3.

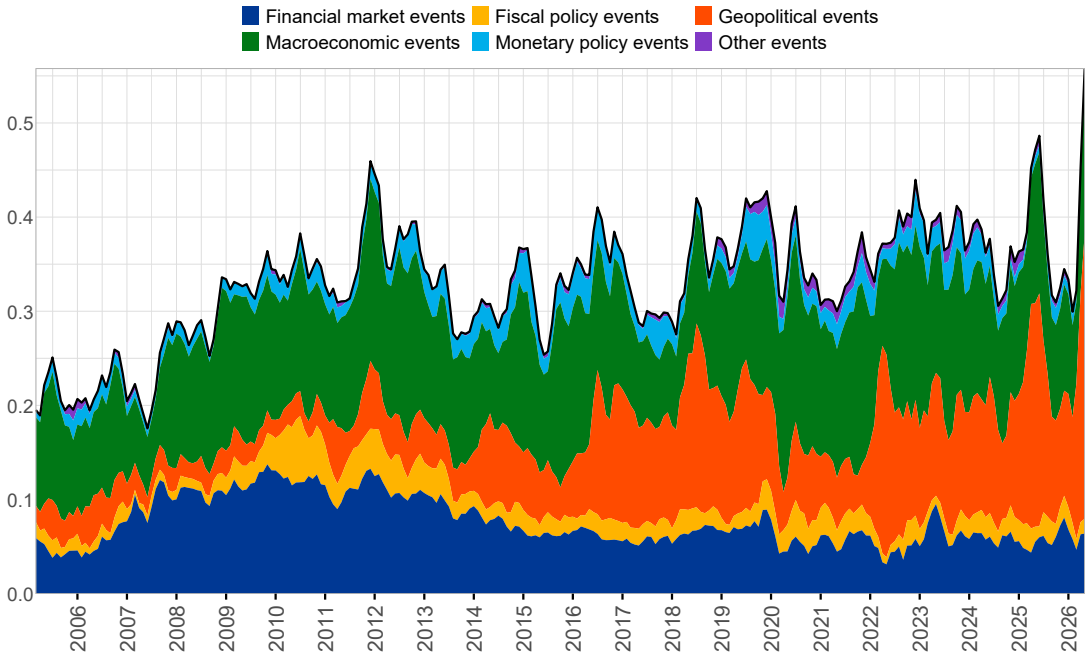
Constructing separate expected impact indicators for short horizons (one to three months) and longer horizons (beyond three months) can further help to identify the immediacy of potential triggers. As shown in Figure 10, the expected impact of potential short-term triggers varies considerably more over time than the expected impact of potential medium-term triggers. In particular, the expected impact of short-term triggers increases significantly around major past crisis episodes, with substantial time-variation in the underlying trigger sources (Figure 10, panel a): during the first half of the sample period, financial market events dominated, with some role also for macroeconomic and fiscal policy trigger events, whereas geopolitical events gained in importance during the second half of the sample, with some role also for macroeconomic and monetary policy events. In contrast, the expected impact of medium-term triggers shows a more secular upward trend over the past 20 years, driven by geopolitical events and macroeconomic events, with peaks around the euro area sovereign debt crisis, the "Liberation day" tariff announcements in 2025 and the recent geopolitical tension in the middle east in 2026 (Figure 10, panel b).

Figure 10: Expected impact indicators for different time horizons

(a) Average expected impact for short horizons (1-3 months)



(b) Average expected impact for longer horizons (more than 3 months)

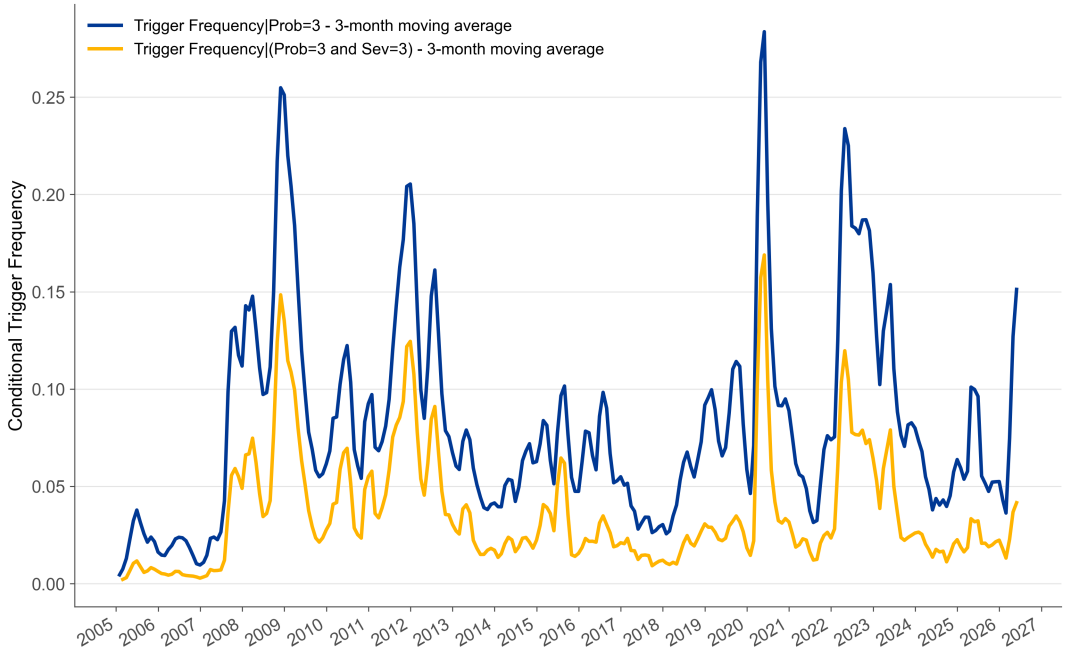


Notes: The indicators are calculated as the average expected impact (probability*severity) of trigger events for the respective horizon group across all articles at a given point in time.

Finally, granular SPOT indicators conditioning on certain probability and severity thresholds can also provide further insights. For example, the trigger frequency conditioning on

high-probability events is consistently larger than the frequency conditioning also on high-severity events, suggesting that news coverage more frequently signals an increased probability of trigger events materialising rather than also having a highly severe macro-financial impact (Figure 11). In addition, the frequency of trigger events classified as highly probable and highly severe mainly shoots up during episodes of major stress, whereas the frequency of high-probability trigger events increases more often and tends to rise somewhat in advance. These patterns suggest that increases in the frequency of triggers with a high-probability and high-severity signals more imminent and consequential threats, while increases in the frequency of high-probability events alone can signal earlier shifts in potential trigger events. Moreover, the frequency of high-probability and high-severity triggers appears less responsive to news with uncertain outcomes, as illustrated by the relatively muted response to the "Liberation day" tariff announcements in 2025, which were widely discussed but not uniformly assessed as having a very severe macro-financial impact. At the current end of the sample in 2026 the frequency of high probability trigger articles also rises much more sharply than the frequency of high probability and high severity articles.

Figure 11: Trigger frequency indicators for different probability and severity thresholds



Notes: The conditional trigger frequency indicators measure the share of Class1 articles that meet the specific trigger conditions relative to the sum of Class0 and Class1 articles (see subsection 3.3).

5 Evaluation of SPOT indicators

5.1 Comparison to other risk indicators from the literature

In this section, we compare SPOT indicators to selected indicators proposed in the literature for measuring financial stress, uncertainty, geopolitical risk or other relevant aspects that can relate to trigger events. The indicators are listed in Table 4 together with selected summary statistics and Figure 12 presents pairwise correlations among the various indicators and also with annual real GDP growth for the euro area. A number of observations are worth noting.

Table 4: Overview of indicators evaluated in the empirical exercise

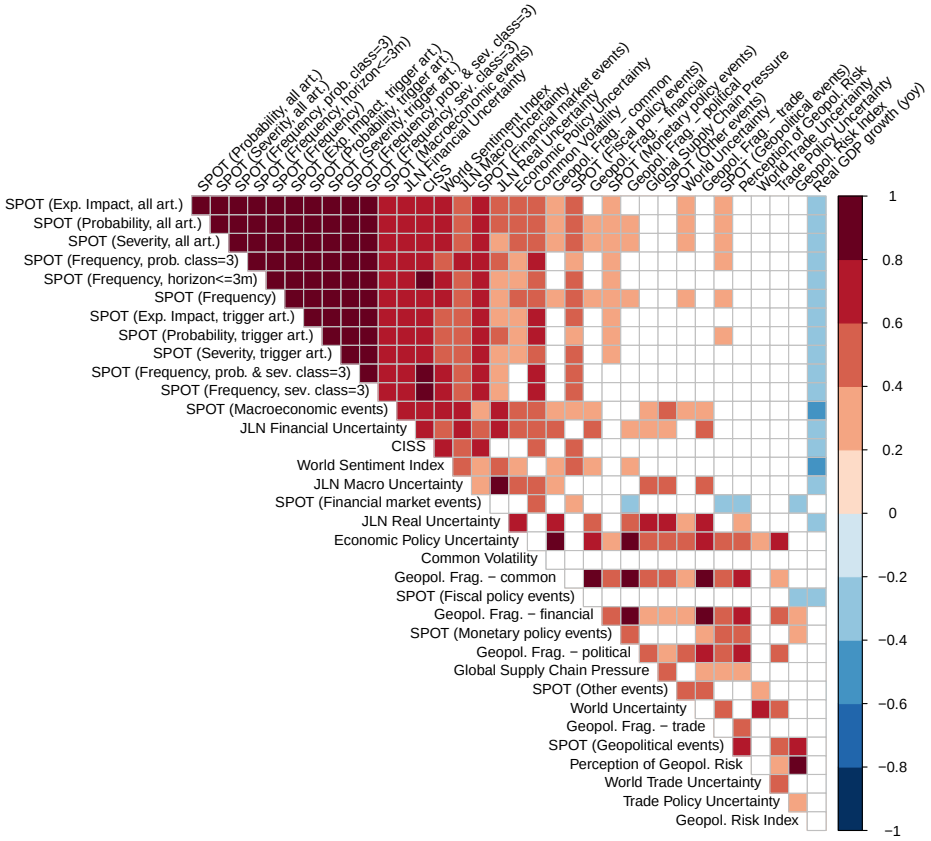
Indicator	Source	Mean	St. deviation	Median	25th perc.	75th perc.	Min	Max	Latest obs.
SPOT (Frequency)	This paper	0.22	0.06	0.22	0.18	0.24	0.10	0.39	2025 Q4
SPOT (Probability, all art.)	This paper	0.50	0.16	0.48	0.40	0.57	0.20	0.98	2025 Q4
SPOT (Probability, trigger art.)	This paper	2.25	0.12	2.24	2.18	2.30	2.02	2.52	2025 Q4
SPOT (Severity, all art.)	This paper	0.45	0.14	0.43	0.36	0.51	0.19	0.86	2025 Q4
SPOT (Severity, trigger art.)	This paper	2.04	0.08	2.02	1.99	2.09	1.89	2.21	2025 Q4
SPOT (Exp. Impact, all art.)	This paper	1.06	0.39	1.00	0.82	1.22	0.39	2.22	2025 Q4
SPOT (Exp. Impact, trigger art.)	This paper	4.72	0.42	4.65	4.45	4.98	3.90	5.71	2025 Q4
SPOT (Frequency, horizon<=3m)	This paper	0.13	0.05	0.12	0.10	0.16	0.04	0.31	2025 Q4
SPOT (Frequency, prob. class=3)	This paper	0.07	0.04	0.06	0.04	0.08	0.01	0.20	2025 Q4
SPOT (Frequency, sev. class=3)	This paper	0.03	0.02	0.02	0.02	0.04	0.00	0.10	2025 Q4
SPOT (Frequency, prob. & sev. class=3)	This paper	0.03	0.02	0.02	0.01	0.04	0.00	0.09	2025 Q4
SPOT (Financial market events)	This paper	0.37	0.25	0.28	0.22	0.47	0.10	1.24	2025 Q4
SPOT (Fiscal policy events)	This paper	0.07	0.05	0.06	0.04	0.09	0.01	0.25	2025 Q4
SPOT (Geopolitical events)	This paper	0.22	0.18	0.19	0.09	0.29	0.03	1.16	2025 Q4
SPOT (Macroeconomic events)	This paper	0.31	0.15	0.28	0.21	0.37	0.11	0.99	2025 Q4
SPOT (Monetary policy events)	This paper	0.07	0.04	0.06	0.04	0.08	0.02	0.19	2025 Q4
SPOT (Other events)	This paper	0.02	0.03	0.01	0.01	0.02	0.00	0.29	2025 Q4
CISS	Kremer et al. (2012)	0.17	0.19	0.08	0.03	0.27	0.00	0.90	2025 Q4
Common Volatility	Engle and Campos-Martins (2023)	0.57	0.16	0.55	0.47	0.66	0.19	1.10	2025 Q2
Economic Policy Uncertainty	Baker et al. (2016)	162.85	72.65	151.87	111.72	210.02	55.39	370.51	2025 Q1
Geopol. Frag. - financial	Fernandez-Villaverde et al. (2024)	-0.63	0.63	-0.78	-0.95	-0.23	-1.68	0.99	2024 Q1
Geopol. Frag. - political	Fernandez-Villaverde et al. (2024)	1.51	1.30	1.72	0.45	2.52	-0.66	3.48	2024 Q1
Geopol. Frag. - trade	Fernandez-Villaverde et al. (2024)	-0.75	0.33	-0.89	-0.97	-0.68	-1.12	0.28	2024 Q1
Geopol. Risk Index	Caldara and Iacoviello (2022)	100.25	26.04	90.85	84.92	109.19	69.69	224.60	2025 Q2
Geopol. Frag. - common	Fernandez-Villaverde et al. (2024)	-0.32	0.45	-0.44	-0.52	-0.14	-0.96	0.76	2024 Q1
Global Supply Chain Pressure	FED New York	0.16	1.06	-0.12	-0.48	0.33	-1.32	4.26	2025 Q2
JLN Financial Uncertainty	Ludvigson et al. (2021)	0.99	0.06	0.99	0.94	1.03	0.90	1.13	2024 Q3
JLN Macro Uncertainty	Ludvigson et al. (2021)	0.93	0.07	0.91	0.88	0.97	0.84	1.13	2024 Q3
JLN Real Uncertainty	Ludvigson et al. (2021)	0.89	0.05	0.86	0.86	0.91	0.83	1.10	2024 Q3
Perception of Geopol. Risk	Alonso-Alvarez et al. (2025)	99.83	48.44	92.14	66.69	105.54	55.02	358.89	2024 Q4
Trade Policy Uncertainty	Caldara et al. (2019)	70.25	104.00	34.60	26.64	61.87	20.78	782.33	2025 Q2
World Sentiment Index	Ahir et al. (2022)	0.82	0.68	0.82	0.49	1.30	-1.40	2.16	2025 Q3
World Trade Uncertainty	Ahir et al. (2022)	11.04	29.53	1.03	0.15	3.73	0.00	174.34	2025 Q1
World Uncertainty	Ahir et al. (2022)	22.06	9.31	19.65	15.10	26.46	8.24	174.34	2025 Q1

Notes: Summary statistics are computed on quarterly observations starting from 2005 Q1. Indicators available at monthly frequency are transformed to quarterly values by computing the moving average over 3 months. Sectoral SPOT indicators reflect the expected impact for that trigger source. World Uncertainty is in thousands.

First, SPOT indicators are negatively correlated with real GDP growth, which is consistent with the notion that an elevated trigger probability and severity is associated with weaker economic activity (Figure 12). Second, various SPOT indicators exhibit high correlations among each other. Nonetheless, this is mostly true for the aggregate SPOT indicators, while sectoral SPOT indicators exhibit lower correlations. Third, the benchmark SPOT displays positive but imperfect correlations with established indicators such as the Composite Indicator of Systemic Stress (CISS), the Geopolitical Risk Indicator (GPR), and the Economic Policy Uncertainty Indicator (EPU). This shows that while these measures share common dimensions, SPOT captures information beyond what is reflected in widely used financial stress and risk indicators.

Fourth, sectoral SPOT indicators cluster with their thematically relevant counterparts — SPOT for geopolitical events groups with GPR, trade policy uncertainty, and geopolitical fragmentation measures, while SPOT for financial market events groups with Common Volatility, CISS and JLN financial uncertainty. This pattern lends support to the validity of the LLM-based classification of trigger sources.

Figure 12: Correlations between SPOT indicators and other indicators

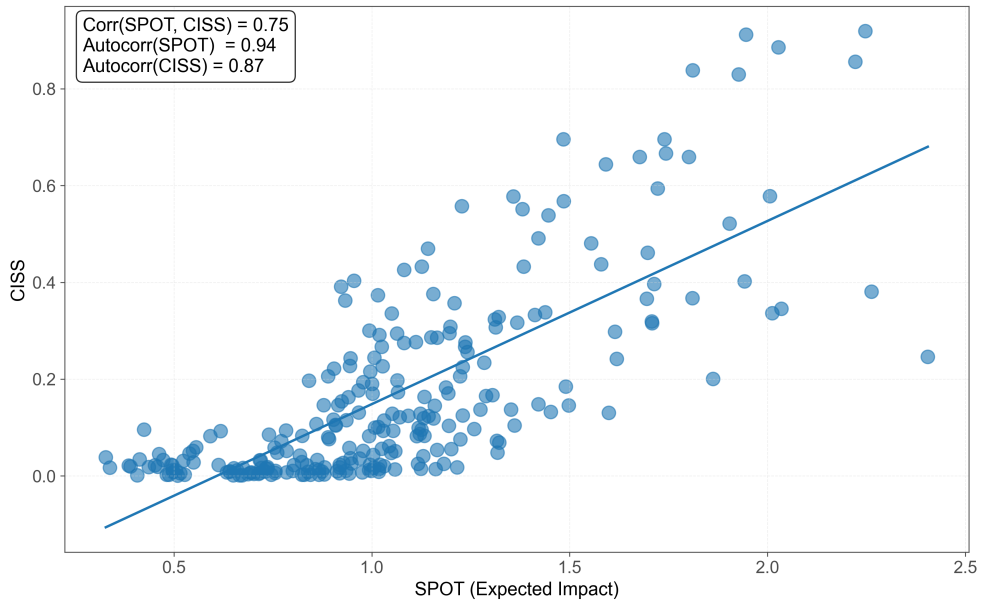


Notes: Computed on quarterly sample 2005Q1-2024Q4 using 3 month averages of monthly values. White squares denote correlations that are not significant at the 99% level. For the World Sentiment Index, negative values are used so that increases denote higher risk. Indicators are ordered based on the first principal component loadings.

Compared to the CISS (Kremer et al., 2012), a widely used financial stress indicator for the euro area, the benchmark SPOT indicator exhibits a clear positive correlation of 0.75, yet the relationship is imperfect (Figure 13). This suggests that the SPOT indicator provides complementary news-based information beyond what is captured by financial market conditions. Notably, the SPOT indicator displays greater persistence than the CISS, consistent with its narrative nature: while the CISS, which is constructed from volatility measures and cross-correlations across financial market segments, tends to respond swiftly to new information, the SPOT indicator reflects the more gradual build-up and unwinding of trigger-related narratives in financial

news. Moreover, a key advantage of the SPOT indicator over purely market-based stress measures is its better interpretability. Whereas the CISS captures how financial markets react in a given situation, through elevated volatility and tighter cross-market co-movements, it does not reveal the underlying narratives for the observed stress. In contrast, the SPOT indicator's decomposition into trigger sources provides a structured narrative of what is driving the increase in potential trigger events, adding information for financial stability surveillance purposes.

Figure 13: Visual comparison of the benchmark SPOT indicator with the CISS

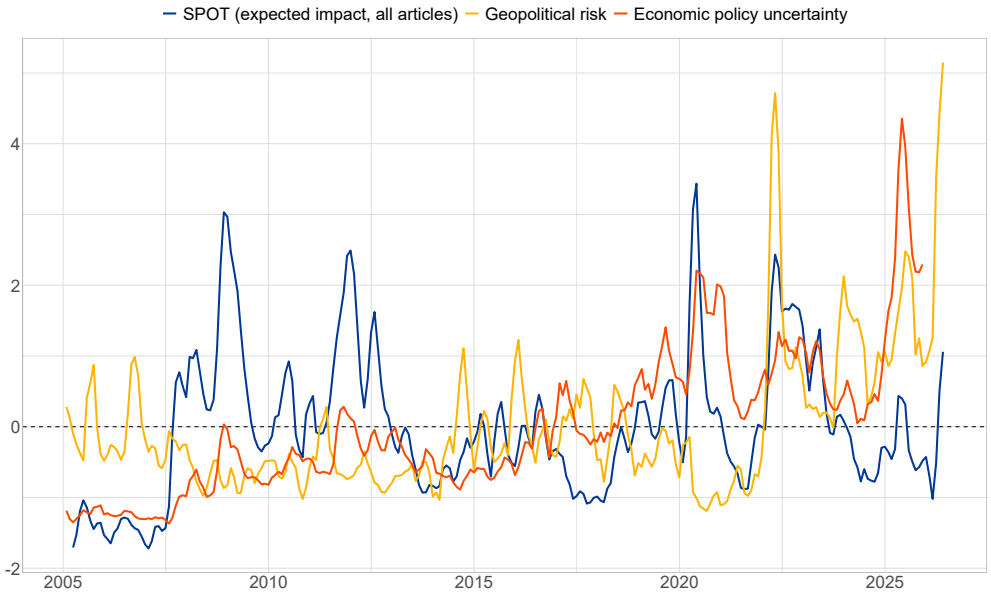


Notes: The benchmark SPOT (Expected Impact) and CISS values are monthly observations.

We also compare the benchmark SPOT indicator to two prominent text-based indicators: the GPR of Caldara and Iacoviello (2022) and the EPU of Baker et al. (2016). The comparison reveals episodes of clear co-movement (Figure 14). For instance, both the benchmark SPOT indicator and EPU rise markedly during the COVID-19 pandemic in 2020 and amid heightened trade policy tensions in 2025, while the SPOT indicator and the GPR index increase together following Russia's invasion of Ukraine in 2022 and more recently during the war in the Middle East in 2026. However, there are also notable episodes where the SPOT indicator captures potential triggers that these more specialised indicators do not. For example, during the global financial crisis and the euro area sovereign debt crisis the SPOT indicator rises substantially, whereas neither the GPR nor the EPU fully reflect the severity of these episodes, given that they were not primarily driven by geopolitical events or policy uncertainty. In addition, even in periods of co-movement, the amplitudes of the respective indicators can differ, reflecting their distinct underlying construction. One useful feature of the SPOT indicator is that it classifies a

trigger article only when it suggests a potentially meaningful adverse effect on the economy or financial system over a forward-looking horizon, rather than simply flagging the use of geopolitical or policy-uncertainty related language in news. By combining trigger event identification with an assessment of the likely macro-financial impact, SPOT can provide information that complements standard dictionary-based risk indicators.

Figure 14: Visual comparison of the benchmark SPOT indicator with the GPR and the EPU



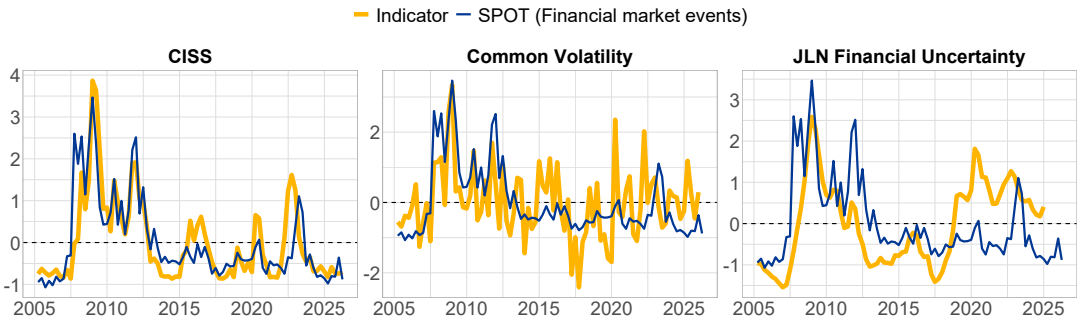
Notes: Indicators are standardised to have a mean of zero and a standard deviation of one. The latest available data for the global economic policy uncertainty index is November 2025.

Finally, several sectoral SPOT indicators are compared to established risk and uncertainty indicators (Figure 15). Three observations are worth noting. First, consistent with the correlation analysis above (Figure 12), the sectoral SPOT indicators broadly co-move with thematically close counterpart indicators: for example with the CISS, common volatility, and JLN financial uncertainty in the case of the financial market SPOT; with the GPR, EPU, and trade policy uncertainty for the geopolitical SPOT; and with the JLN macro and real uncertainty for the macroeconomic events SPOT. This thematic co-movement corroborates the consistency of the LLM-based trigger source classification. Second, the co-movement is loose rather than exact. There are episodes during which the sectoral SPOT indicators rise while the corresponding other indicators remain comparatively muted, and vice versa. Such occasional divergences suggest that the LLM-based SPOT indicators may capture aspects of trigger perceptions that may not be fully reflected in market-based or dictionary-based indicators. Third, despite these differences, the overall time-series profile of the sectoral SPOT indicators is broadly consistent with that of established measures, indicating that the SPOT indicators do not generate spurious signals

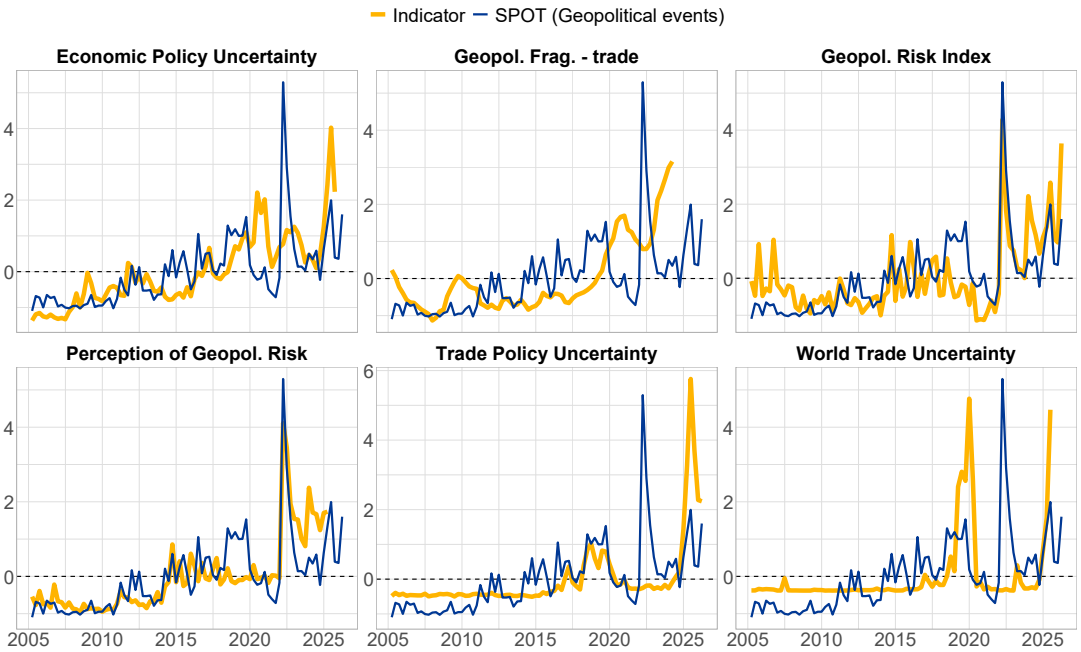
unrelated to the risk environment.

Figure 15: Visual comparison of sectoral SPOT indicators and thematically aligned indicators

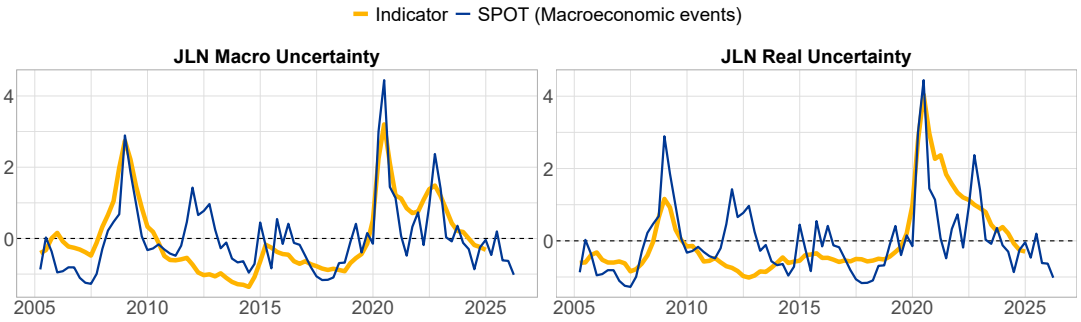
(a) Financial market events



(b) Geopolitical events



(c) Macroeconomic events



Notes: Indicators are standardised to have a mean of zero and a standard deviation of one. Sectoral SPOT indicators shown capture expected impact of all articles. Indicators are shown as 3-month moving averages at quarterly frequency and therefore look slightly different from the monthly indicators shown in Figure 8. The last shown observation is 2026 Q1.

5.2 SPOT performance evaluation in growth-at-risk models

To assess the information content of SPOT indicators compared to other risk indicators more formally, we employ the growth-at-risk model framework which can be used to estimate future downside risks to the economy (Adrian et al., 2022, 2019; Lang et al., 2025). In particular, we estimate quarterly panel quantile local projection models for various horizons h as follows:

$$Q_{y_{i,t+h,\tau}} = \alpha_i^{h,\tau} + \rho^{h,\tau} y_{i,t} + \beta^{h,\tau} \mathbf{X}'_{\mathbf{i,t}} + \varepsilon_{i,t+h,\tau}$$

where quantiles are denoted by τ, $y_{i,t+h}$ is the average real GDP growth rate in country i between time period t and $t+h$, $\mathbf{X}'_{\mathbf{i,t}}$ is a vector of explanatory variables, while $\alpha_i^{h,\tau}$ are country fixed effects, and $\varepsilon_{i,t+h,\tau}$ denotes an error term. We apply a two-step estimation approach for panel quantile regressions following Canay (2011). In the first step, we estimate the unobserved fixed effects using a within estimator, assuming that country fixed effects are constant across quantiles. In the second step, we run a standard conditional quantile regression on the dependent variable that has been adjusted by subtracting the fixed-effect estimates from the first step (see Canay, 2011, for further details).

To compare the performance of different indicators, we use the tick loss as our measure of model fit, a criterion widely used to assess the accuracy of value-at-risk models. The majority of growth-at-risk studies that adopt a formal evaluation metric make use of the tick loss function (Brownlees and Souza, 2021; Carriero et al., 2020; Figueres and Jarocinski, 2020; Giacomini and Komunjer, 2005). The tick loss is the value of the objective function that is minimized by the quantile regression. More precisely, it is computed as follows:

$$TL_{h,\tau} = (\tau - \mathbf{1}(\widehat{\varepsilon}_{h,\tau} < 0))\widehat{\varepsilon}_{h,\tau},$$

where $\mathbf{1}()$ denotes the indicator function, τ is the quantile of interest, and $\widehat{\varepsilon}_{h,\tau}$ is the residual (i.e., actual value minus predicted value).

As a benchmark model, we estimate a panel quantile regression with current annual and quarterly real GDP growth as explanatory variables as well as the systemic risk indicator (SRI) of Lang et al. (2019) to capture financial stability vulnerabilities.[6] As shown by Lang et al. (2025), the SRI outperforms other vulnerability indicators in terms of in-sample explanatory power and out-of-sample forecasting performance for medium-term growth-at-risk in euro area countries and therefore constitutes a key benchmark against which the additional information

[6] We also add the lagged SRI, and an interaction term for when the SRI is positive

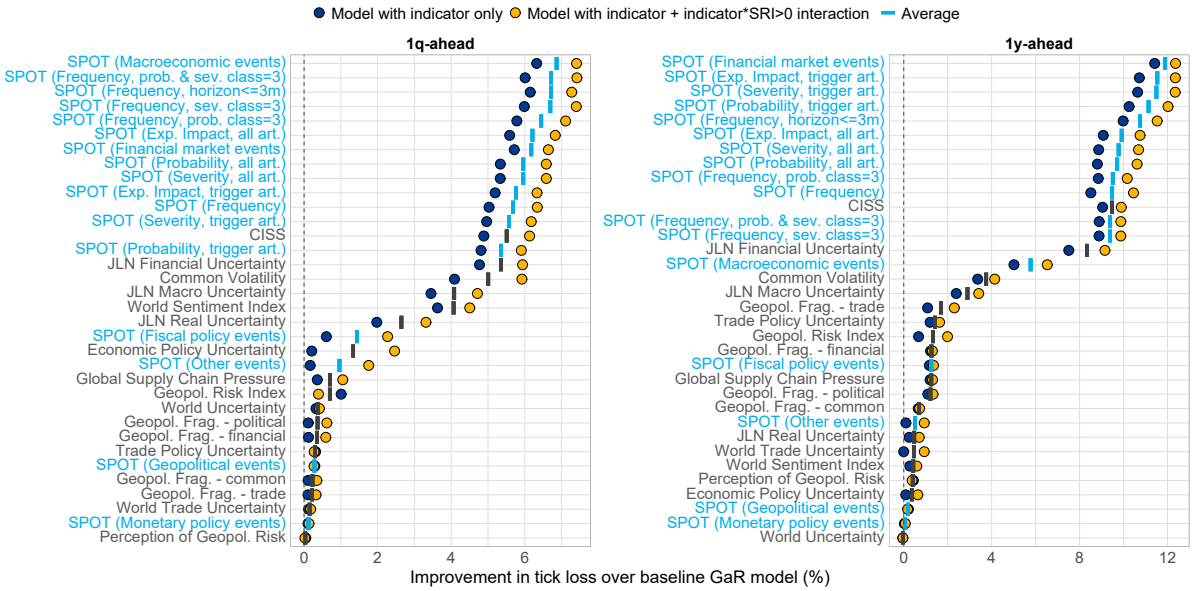
content of other indicators should be evaluated. To account for the exogenous nature of the covid pandemic shock, we add dummies for the following periods: 2020 Q1, ..., 2020 Q4, and 2021 Q2. The models are estimated on a sample of euro area countries[7] over the period 2005 Q1 to 2024 Q1.

We focus on the lower tail of the future real GDP growth distribution by estimating the 10th quantile ($\tau = 0.1$) for 1-quarter and 1-year ahead horizons ($h = 1, 4$). Each potential trigger indicator is added to the benchmark model one at a time. To make the results comparable, we balance observations across all models (i.e. the dataset is balanced within each country). We also consider specifications that interact the indicator with the SRI whenever the SRI is positive, to capture the potential amplification of triggers when vulnerabilities are elevated. For each specification, we compute the percentage improvement in the average tick loss relative to the benchmark model.

Key results from the growth-at-risk exercises are reported in Figure 16. Overall, SPOT indicators significantly improve the performance of the benchmark model and consistently outperform other standard financial stress and risk indicators. This holds for both the 1-quarter ahead and 1-year ahead horizon, with improvements compared to the benchmark model fit of more than 6% and 10% respectively. Given the high correlation between various aggregate SPOT indicators documented in Figure 12, it is not surprising that their growth-at-risk model performance is similar. Nevertheless, SPOT indicators that combine information about the frequency of triggers with information about the probability/severity of triggers tend to slightly dominate SPOT indicators that are solely based on the trigger frequency. In addition, SPOT indicators that combine information about the probability with information about the severity (expected impact) perform slightly better than indicators solely based on the probability or the severity. These findings indicate that combining different trigger dimensions in aggregate SPOT indicators seems desirable. The benchmark SPOT indicator (average expected impact) is among the top six best-performing SPOT indicators for both growth-at-risk projection horizons.

[7] In addition to euro area countries, we include DK and SE but exclude HR and BG due to data availability.

Figure 16: Performance of SPOT and other indicators in growth-at-risk models



Notes: Improvements in tick loss function are in percentage relative to the benchmark model with GDP and the SRI (including its lag and a dummy variable for cases when SRI>0). Indicators are ordered by largest average improvement in model performance (i.e. percentage decrease in the tick loss relative to the benchmark model) for the two sets of model specifications (indicator only and indicator*SRI>0 interaction models).

The best-performing SPOT indicator for the 1-quarter ahead horizon is the expected impact of macroeconomic trigger events, while for the 1-year ahead horizon the best-performing SPOT indicator is the expected impact of financial market trigger events (Figure 16). However, neither of these sectoral SPOT indicators outperform the benchmark SPOT indicator (SPOT Exp. impact, all art.) at both horizons. Other granular SPOT indicators that slightly outperform the benchmark SPOT indicator for the 1-quarter ahead horizon are the frequency of triggers with a high probability and/or high severity. However, at the 1-year ahead horizon this slight outperformance vanishes. Another granular SPOT indicator that performs well at both growth-at-risk horizons is the frequency of immediate trigger events (trigger horizon $\leq$ 3 months). Among other established risk indicators, those related to financial markets provide the largest improvements in growth-at-risk model fit (CISS, JLN Financial Uncertainty, and Common Volatility). However, their performance remains below the performance of the benchmark SPOT indicator. Most other risk indicators are significantly worse than the benchmark SPOT. The strong growth-at-risk model improvement of financial market indicators (including the sectoral SPOT indicator for financial market events) is likely related to the fact that the estimation sample includes two crises driven by financial markets (the global financial crisis and the sovereign debt crisis).

As shown in Figure 16, the growth-at-risk model performance of SPOT indicators improves

further, by around 2 percentage points, when interaction terms between the SPOT indicators and the SRI are included (see yellow vs blue dots). However, the overall growth-at-risk performance ordering of various SPOT indicators does not change much when allowing for such interaction terms. These results indicate that there are important interactions and amplification mechanisms between underlying financial vulnerabilities, as measured by the SRI, and potential trigger events captured by SPOT: while triggers act as immediate catalysts for stress, the differential impact on the lower tail of the growth distribution is driven by pre-existing macro-financial vulnerabilities. This highlights the importance of a holistic approach to monitoring financial stability risks, combining information about vulnerabilities with information about potential trigger events.

The findings regarding the growth-at-risk performance of various SPOT indicators are robust to changing the sample period or looking at different quantiles, as shown in Annex A.3.

5.3 Robustness and sensitivity analysis of SPOT

A potential concern when constructing text-based indicators with large language models is that the resulting signal may depend on the specific model used. To assess the robustness of the SPOT framework, we therefore compare the benchmark implementation based on GPT-4o-mini with several alternative frontier models, namely GPT-4o, GPT-5-mini and GPT-4.1-nano. To ensure comparability across models, we draw a fixed sample of 100 articles per month from the set of articles passing the first-stage economic and financial relevance filter. The same article sample is then evaluated using identical second- and third-stage prompts across all models.

At the article level, the models differ primarily in their trigger-classification thresholds. As shown in Table 5, GPT-4o and GPT-5-mini classify substantially fewer articles as trigger events than the GPT-4o-mini benchmark, with trigger rates of 7.8% and 10.8% respectively, compared with 24.2% for GPT-4o-mini. This suggests that the larger models apply more conservative trigger-classification criteria. The nature of the disagreement across LLMs is also informative. GPT-4o and GPT-5-mini identify only around one-third of the trigger articles detected by the benchmark (Table 5 column 2), reflecting their substantially lower trigger rates. However, their precision relative to the benchmark is very high (96% and 85% respectively), indicating that articles classified as signalling trigger events by these models are almost always also classified as signalling trigger events by the benchmark LLM (Table 5 column 4). In other words, these LLMs mainly identify a subset of the trigger articles detected by GPT-4o-mini, rather than a different set of trigger articles altogether. At the same time, agreement on non-trigger articles is very high, ranging from 97.9% to 99.6%. This suggests that the main differences arise from the

treatment of borderline trigger cases, while there is broad agreement across LLMs on both clear trigger articles and non-trigger articles. GPT-4.1-nano differs somewhat from the larger models in that it classifies a larger share of articles as trigger events than GPT-4o and GPT-5-mini, but still substantially fewer than the GPT-4o-mini benchmark. At the same time, it exhibits lower agreement with the benchmark on trigger classifications and lower precision than the larger models, suggesting a slightly noisier classification.

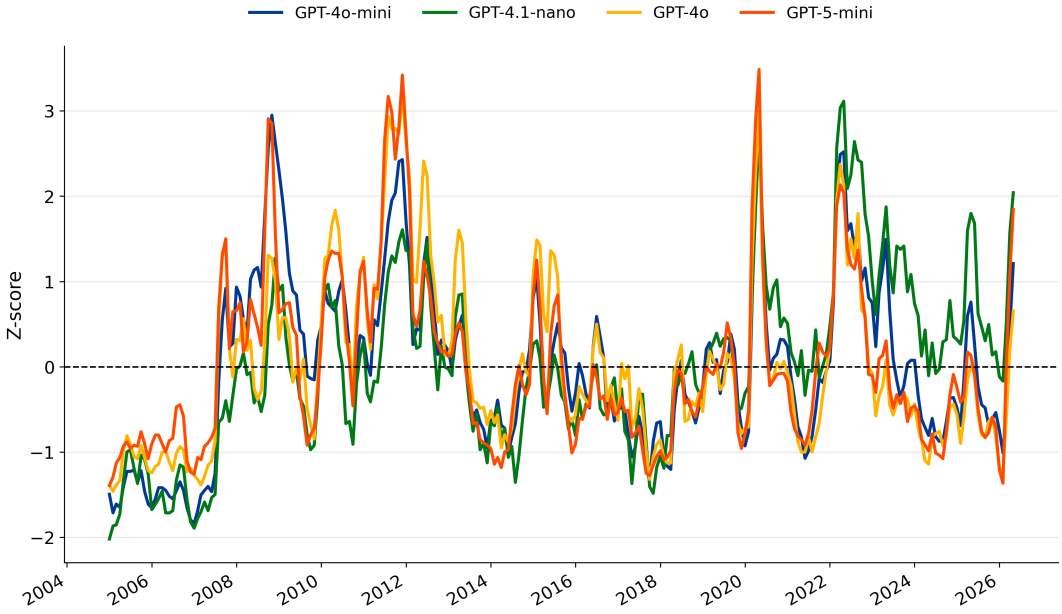
Table 5: Article-level classification agreement across different LLMs

Model	Trigger rate (%)	Benchmark 1s matched (%)	Benchmark 0s matched (%)	Precision (%)
GPT-4o-mini	24.2	100.0	100.0	100.0
GPT-4.1-nano	13.4	37.8	94.4	68.3
GPT-4o	7.8	31.0	99.6	96.1
GPT-5-mini	10.8	38.0	97.9	85.2

Notes: The trigger rate reports the share of articles classified as trigger events by each model. "Benchmark 1s matched" reports the share of benchmark trigger articles that are also classified as trigger events by the alternative model, while "Benchmark 0s matched" reports the share of benchmark non-trigger articles that are also classified as non-trigger events. Precision measures the share of articles classified as trigger events by an alternative model that are also classified as trigger events by the GPT-4o-mini benchmark.

For each model, we construct the benchmark SPOT indicator as the monthly average expected impact associated with identified potential trigger events. All indicators are transformed into three-month moving averages and standardised to z-scores to facilitate direct comparison across models. Figure 17 shows that the resulting SPOT indicators display similar dynamics across the main frontier models. In particular, all models identify the major stress episodes associated with the global financial crisis, the euro area sovereign debt crisis, the COVID-19 shock and the geopolitical and energy stress episode following Russia's invasion of Ukraine. The broad cyclical dynamics of the SPOT indicator therefore appear robust to the underlying LLM being used.

Figure 17: Z-scored SPOT indicators based on different LLMs



Notes: The figure reports z-scored benchmark SPOT indicators constructed using alternative large language models on a fixed sample of 100 articles per month. Indicators are transformed into three-month moving averages and standardised to z-scores to facilitate comparison across models.

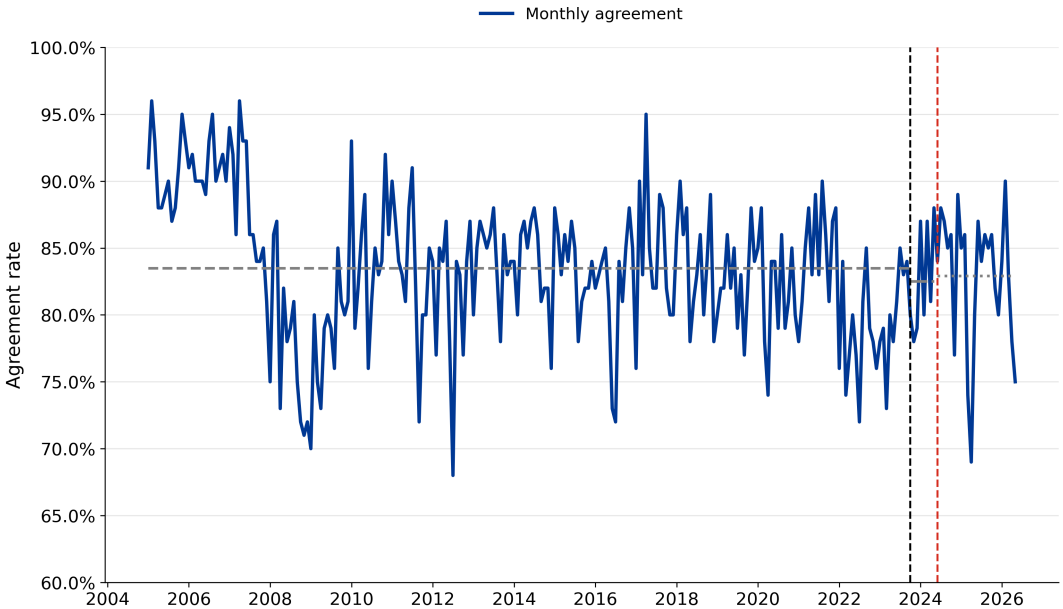
The conclusion that cyclical dynamics of the SPOT indicator are robust to using different LLMs is further supported by the high pairwise correlations between the z-scored SPOT indicators reported in Table 6. The strongest alignment is observed between GPT-4o and GPT-5-mini, with a correlation of 0.92, while GPT-4o-mini and GPT-5-mini exhibit a correlation of 0.88. More generally, the benchmark SPOT indicator generated with GPT-4o-mini displays high positive correlations of 0.78 to 0.88 with the SPOT indicators that are generated using the alternative LLMs, suggesting that the dynamics of the SPOT indicator are robust to using different frontier models. The SPOT indicator based on GPT-4.1-nano displays somewhat lower correlations with the other SPOT indicators that are based on larger LLMs, suggesting that the smallest LLM generates somewhat noisier classifications and assessments, while preserving the broader dynamics of the benchmark SPOT indicator.

Table 6: Correlation of z-scored SPOT indicators across models

	GPT-4o-mini	GPT-4.1-nano	GPT-4o	GPT-5-mini
GPT-4o-mini	1.000	0.778	0.847	0.883
GPT-4.1-nano	0.778	1.000	0.661	0.666
GPT-4o	0.847	0.661	1.000	0.922
GPT-5-mini	0.883	0.666	0.922	1.000

We additionally assess the possibility of look-ahead bias by making use of differences in model training cutoff dates. Figure 18 reports the monthly agreement rate between GPT-4o-mini and GPT-5-mini together with the respective model training cutoffs. The agreement rate is defined as the share of articles in a given month that receive the same binary trigger classification (Class = 0 or Class = 1) from both models. The average agreement rate remains highly stable across all periods, averaging 0.835 before the GPT-4o-mini cutoff, 0.825 between the GPT-4o-mini and GPT-5-mini cutoffs, and 0.829 after the GPT-5-mini cutoff. The absence of any visible structural break across these periods, particularly between the two training cutoffs, suggests that the models do not systematically rely on information contained in their training data when classifying articles. Instead, the results indicate that the classifications are primarily driven by the information contained in the articles themselves and by the definitions provided in the structured prompt, as intended. Moreover, the spike of the benchmark SPOT indicator during spring 2026 due to the war in the Middle East (Figure 5) is another indication that look-ahead bias is not driving spikes in the SPOT indicator, as the training cut-off for GPT-4o-mini is well before 2026.

Figure 18: Agreement rate between GPT-4o-mini and GPT-5-mini over time



Notes: The agreement rate is defined as the share of articles receiving the same binary trigger classification from both models. The vertical dashed lines indicate the training cutoffs of GPT-4o-mini and GPT-5-mini, respectively. Horizontal dashed lines report average agreement rates before the GPT-4o-mini cutoff, between the two model cutoffs, and after the GPT-5-mini cutoff.

Overall, the robustness exercises suggest that the SPOT indicator is relatively robust across sufficiently capable frontier models and that look-ahead bias should not be a major concern

given the careful prompt design. Differences across LLMs mainly reflect varying degrees of conservativeness in trigger classification rather than fundamentally different assessments of time-variations in triggers. The stability of the benchmark indicator across model architectures and training vintages provides additional support for the robustness and reproducibility of the SPOT indicator.

6 Conclusion

In this paper we developed SPOT, an AI-based indicator that measures the potential **S**everity and **P**robability **O**f **T**riggers (**SPOT**) for financial stability risks based on news. We showed that our benchmark SPOT indicator spikes ahead of major historical episodes of financial stress or instability, such as the global financial crisis, the euro area sovereign debt crisis, the Covid-19 pandemic, or the Russian invasion of Ukraine. Moreover, the decomposition of SPOT by trigger source allows for further insights into underlying drivers and helps form a risk narrative. Sectoral SPOT indicators for individual trigger sources, e.g. for geopolitical or fiscal triggers, make the underlying drivers even more visible and can be used to complement the benchmark SPOT indicator for monitoring specific trigger sources. Hence, the SPOT framework provides a comprehensive view of the dynamics, drivers, and characteristics of perceived triggers for financial stability risks.

We also showed that various SPOT indicators contain useful information about future downside risks to the economy and outperform other commonly used indicators capturing financial stress, geopolitical risk, policy uncertainty or volatility in euro area panel growth-at-risk models. In addition, interacting the SPOT indicator with vulnerability indicators further improves the explanatory power of growth-at-risk models, indicating important amplification mechanisms between vulnerabilities and triggers in driving economic tail risks.

Overall, the results presented in this paper suggest that AI-based signal extraction from text offers a promising avenue to enhance the monitoring of financial stability risks. Recent research shows that machine learning and LLMs can be used to extract relevant information from unstructured data to construct indicators of sentiment, uncertainty, and macro-financial risks. Building on these advances, the SPOT indicator presented in this paper demonstrates how LLMs can be applied to financial news to provide a systematic and interpretable measure of potential trigger events, thereby complementing existing vulnerability indicators for monitoring financial stability risks. Looking ahead, a promising avenue for applied financial stability research appears to be the application of similar approaches to new or previously underutilised

data sources, including financial disclosures, supervisory information, market intelligence findings, or social media posts. While challenges related to model reliability, interpretability, and data governance remain, AI-based approaches have the potential to become an increasingly important component of financial stability surveillance frameworks.

References

Adrian, Tobias, Federico Grinberg, Nellie Liang, Sheheryar Malik, and Jie Yu, "The Term Structure of Growth-at-Risk," *American Economic Journal: Macroeconomics*, July 2022, *14* (3), 283–323.

_ , Nina Boyarchenko, and Domenico Giannone, "Vulnerable Growth," *American Economic Review*, April 2019, *109* (4), 1263–1289.

Ahir, Hites, Nicholas Bloom, and Davide Furceri, "The World Uncertainty Index," NBER Working Papers 29763, National Bureau of Economic Research, Inc Feb 2022.

Aikman, David, Jonathan Bridges, Sinem Hacioglu Hoke, Cian O Neill, and Akash Raja, "Credit, capital and crises: a GDP-at-Risk approach," Bank of England working papers 824, Bank of England September 2019.

_ , _ , Stephen Burgess, Richard Galletly, Iren Levina, Cian O'Neill, and Alexandra Varadi, "Measuring risks to UK financial stability," Bank of England working papers 738, Bank of England July 2018.

Alessi, Lucia and Carsten Detken, "Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity," *European Journal of Political Economy*, September 2011, *27* (3), 520–533.

Alonso-Alvarez, Irma, Marina Diakonova, and Javier J. Perez, "Rethinking GPR: The sources of geopolitical risk," Working Papers 2522, Banco de Espana May 2025.

Audrino, Francesco, Jessica Gentner, and Simon Stalder, "Quantifying uncertainty: a new era of measurement through large language models," Working Papers 2024-12, Swiss National Bank 2024.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, "Article Navigation Journal Article Editor's Choice Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, November 2016, *131* (4), 1593–1636.

Bond, Shaun A., Hayden Klok, and Min Zhu, "Large Language Models and Financial Market Sentiment," *SSRN Online Library*, October 2023.

Borio, Claudio and Mathias Drehmann, "Assessing the risk of banking crises - revisited," *BIS Quarterly Review*, March 2009, pp. 29–46.

_ **and Philip Lowe**, "Asset prices, financial and monetary stability: exploring the nexus," BIS Working Papers 114, Bank for International Settlements July 2002.

Brownlees, Christian and André B.M. Souza, "Backtesting global Growth-at-Risk," *Journal of Monetary Economics*, 2021, *118* (C), 312–330.

Caldara, Dario and Matteo Iacoviello, "Measuring Geopolitical Risk," *American Economic Review*, April 2022, *112* (4), 1194–1225.

_ **,** _ **, Patrick Molligo, Andrea Prestipino, and Andrea Raffo**, "Does Trade Policy Uncertainty Affect Global Economic Activity?," FEDS Notes 2019-09-04, Board of Governors of the Federal Reserve System (U.S.) Sep 2019.

Canay, Ivan A., "A simple approach to quantile regression for panel data," *Econometrics Journal*, October 2011, *14* (3), 368–386.

Carriero, Andrea, Todd E. Clark, and Marcellino Massimiliano, "Nowcasting Tail Risks to Economic Activity with Many Indicators," Working Papers 20-13R2, Federal Reserve Bank of Cleveland May 2020.

Chavleishvili, Sulkhan, Robert F. Engle, Stephan Fahr, Manfred Kremer, Simone Manganelli, and Bernd Schwaab, "The risk management approach to macro-prudential policy," Working Paper Series 2565, European Central Bank June 2021.

Chen, Chung-Chi, Yu-Lieh Huang, and Fang Yang, "Semantics matter: An empirical study on economic policy uncertainty index," *International Review of Economics & Finance*, January 2024, *89* (Part A), 1286–1302.

Correa, Ricardo, Keshav Garud, Juan M. Londono, and Nathan Mislang, "Sentiment in Central Banks' Financial Stability Reports," International Finance Discussion Papers 1203, Board of Governors of the Federal Reserve System (U.S.) Mar 2017.

Covitz, Daniel, Nellie Liang, and Tobias Adrian, "Financial Stability Monitoring," *Annual Review of Financial Economics*, December 2015, *7* (1), 357–395.

Detken, Carsten, Olaf Weeken, Lucia Alessi, Diana Bonfim, Miguel M. Boucinha, Christian Castro, Sebastian Frontczak, Gaston Giordana, Julia Giese, Nadya Jahn, Jan Kakes, Benjamin Klaus, Jan Hannes Lang, Natalia Puzanova, and Peter Welz, "Operationalising the countercyclical capital buffer: indicator selection, threshold identification and calibration options," ESRB Occasional Paper Series No. 5, ESRB June 2014.

Engle, Robert F. and Susana Campos-Martins, "What are the events that shake our world? Measuring and hedging global COVOL," *Journal of Financial Economics*, None 2023, *147* (1), 221–242.

Falconio, Andrea and Simone Manganelli, "Financial conditions, business cycle fluctuations and growth at risk," Working Paper Series 2470, European Central Bank September 2020.

Fell, John and Garry Schinasi, "Assessing Financial Stability: Exploring the Boundaries of Analysis," *National Institute Economic Review*, 2005, *192*, 102–117.

_ , Sandor Gardo, Benjamin Klaus, Jonas Wendelborn, and Stefan Wredenborg, "Communication for financial crisis prevention: a tale of two decades," *ECB Financial Stability Review*, November 2024, *2*.

Fernandez-Villaverde, Jesus, Tomohide Mineyama, and Dongho Song, "Are We Fragmented Yet? Measuring Geopolitical Fragmentation and Its Causal Effect," NBER Working Papers 32638, National Bureau of Economic Research, Inc Jun 2024.

Figueres, Juan Manuel and Marek Jarocinski, "Vulnerable growth in the euro area: Measuring the financial conditions," *Economics Letters*, 2020, *191* (C).

Galán, Jorge E., "The benefits are at the tail: Uncovering the impact of macroprudential policy on growth-at-risk," *Journal of Financial Stability*, 2020, p. 100831.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, September 2019, *57* (3), 535–574.

Giacomini, Raffaella and Ivana Komunjer, "Evaluation and Combination of Conditional Quantile Forecasts," *Journal of Business & Economic Statistics*, October 2005, *23*, 416–431.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, July 2023,

120 (30), e2305016120.

Glasserman, Paul and Xinyu Lin, "Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis," SSRN Working Paper, SSRN Electronic Journal 2024.

Hartwig, Benny, Christoph Meinerding, and Yves S. Schüler, "Identifying indicators of systemic risk," *Journal of International Economics*, 2021, *132* (C).

Kremer, Manfred, Marco Lo Duca, and Daniel Hollo, "CISS - a composite indicator of systemic stress in the financial system," Working Paper Series 1426, European Central Bank Mar 2012.

Kwon, Byeungchun, Taejin Park, Phurichai Rungcharoenkitkul, and Frank Smets, "Parsing the pulse: decomposing macroeconomic sentiment with LLMs," BIS Working Papers 1294, Bank for International Settlements Oct 2025.

Lang, Jan Hannes, Cosimo Izzo, Stephan Fahr, and Josef Ruzicka, "Anticipating the bust: a new cyclical systemic risk indicator to assess the likelihood and severity of financial crises," Occasional Paper Series 219, European Central Bank February 2019.

_ , Marek Rusnák, and Moritz Greiwe, "Medium-Term Growth-at-Risk in the Euro Area," *IMF Economic Review*, 2025.

Lefort, Baptiste, Eric Benhamou, Jean-Jacques Ohana, David Saltiel, Beatrice Guez, and Damien Challet, "Can ChatGPT Compute Trustworthy Sentiment Scores from Bloomberg Market Wraps?," *SSRN Online Library, Paris Dauphine University*, Feburary 2024, pp. 1–32.

Lopez-Lira, Alejandro and Yuehua Tang, "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models," *arXiv preprint*, April 2023.

Loughran, Tim and Bill McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, January 2011, *66* (1), 35–65.

Ludvigson, Sydney C., Sai Ma, and Serena Ng, "Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response?," *American Economic Journal: Macroeconomics*, October 2021, *13* (4), 369–410.

Nyman, Rickard, Sujit Kapadia, and David Tuckett, "News and narratives in financial sys-

tems: Exploiting big data for systemic risk assessment," *Journal of Economic Dynamics & Control*, June 2021, *127*.

Plagborg-Moller, M., L. Reichlin, G. Ricco, and T. Hasenzagl, "When is Growth at Risk?," *Brookings Papers on Economic Activity*, Spring 2020.

Qin, Zhen, Jiacheng Wu, Jiaming Shen, Tianyu Liu, and Xiaojun Wang, "LAMPO: Large Language Models as Preference Machines for Few-Shot Ordinal Classification," in "Proceedings of the First Conference on Language Modeling (COLM)" 2024.

Sarkar, Ritam and Keyon Vafa, "Lookahead bias in pretrained language models," *arXiv preprint arXiv:2410.09827*, October 2024.

Schularick, Moritz and Alan M. Taylor, "Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870-2008," *American Economic Review*, April 2012, *102* (2), 1029–61.

Schüler, Yves S., Paul P. Hiebert, and Tuomas A. Peltonen, "Financial cycles: Characterisation and real-time measurement," *Journal of International Money and Finance*, 2020, *100* (C).

Törnberg, Petter, "Best practices for using large language models in computational social science," *Sociological Methods and Research*, 2024.

Zhang, Zijian, Rong Fu, Yangfan He, Xinze Shen, Yanlong Wang, Xiaojing Du, Haochen You, Jiazhao Shi, and Simon Fong, "FinSentLLM: Multi-LLM and Structured Semantic Signals for Enhanced Financial Sentiment Forecasting," *Preprint in arXiv*, Septmeber 2025.

Zhao, Yilun, Yuxuan Long, Haonan Liu, Linyong Nan, Lifu Chen, Ryo Kamoi, Yijun Liu, Xiaoman Tang, Rui Zhang, and Arman Cohan, "DocMath-Eval: Evaluating Numerical Reasoning Capabilities of LLMs in Understanding Long Documents with Tabular Data," *arXiv preprint arXiv:2311.09805*, 2023.

Appendix

The appendix provides additional methodological details, robustness exercises, and supplementary figures and tables supporting the empirical results presented in the main text. More specifically, the appendix reports the prompts used in the three-stage LLM prompting pipeline, additional descriptive statistics on the underlying dataset and classification procedure, supplementary lead-lag evidence for the SPOT indicator, and robustness checks for the Growth-at-risk exercises under alternative samples and quantile specifications. These additional materials complement the main analysis and provide further transparency regarding the construction and evaluation of the SPOT framework.

A.1 Prompting pipeline

This part of the appendix reports the prompts used in the three-stage prompting pipeline described in section 3. The pipeline is designed to decompose the classification and assessment tasks into three sequential steps. Stage 1 filters articles for economic and financial relevance. Stage 2 classifies whether a relevant article signals a current or potential trigger event for euro area economic activity or financial stability. Stage 3 characterises each trigger event by its probability, severity, expected horizon and trigger source. A key design feature of the prompts in Stages 2 and 3 is the restriction that the model should use only information contained in the article as of its publication date. The system prompt explicitly instructs the model not to use outside knowledge, future information or speculation. This is intended to mitigate look-ahead bias and to ensure that the classifications reflect the information set available at the time the article was published. While prompting alone cannot fully rule out information leakage, this restriction is important for aligning the exercise with a real-time financial stability monitoring setting. All prompts are executed using GPT-4o-mini with deterministic inference settings: `temperature` $= 0$, `top_p` $= 1$, `presence_penalty` $= 0$, and `frequency_penalty` $= 0$. These settings are chosen to maximise classification consistency and reproducibility across articles and prompting stages. In particular, setting the temperature to zero reduces randomness in model outputs and helps ensure that identical or highly similar articles receive stable classifications across repeated runs. The remaining parameters are kept at their default deterministic values to avoid introducing additional variation in token selection or response structure. This is particularly important in our setting, where the objective is structured and comparable text classification rather than creative text generation. The same deterministic inference settings are also kept for all robustness exercises in subsection 5.3.

Stage 1: Economic and financial relevance filter

```
SYSTEM PROMPT:
You are a strict binary classifier. You follow instructions exactly
and output valid JSON only.

USER PROMPT:
Classify the following Financial Times article.

Output 1 ONLY IF the article's PRIMARY focus is economic or financial
AND the economic/financial content is explicit, substantive, and central
to the article.

Economic/financial content must include at least one of:
- macroeconomic variables (e.g. inflation, GDP, interest rates, unemployment)
- financial markets or instruments (e.g. equities, bonds, spreads, banks,
  investment funds, non-bank financial institutions)
- economic policy with direct, concrete economic impact

Output 1 for articles about one specific firm ONLY if the article describes
an impact on the economy or wider financial system of the firm-specific event.

Output 0 if the article is strictly non-economic or non-financial
(e.g. technology, travel, health, culture, politics) without an economic
or financial angle or if economic terms are used metaphorically or without
analytical substance.

When in doubt, output 0.

Return ONLY valid JSON in the following format:
{"filter_label": 0}

Article text:
{text}
```

Stage 2: Trigger-event classification

```
SYSTEM PROMPT:
You are an economic and financial analyst. Return a single JSON object only
(no code fences, no markdown, no explanations) and include exactly the
requested keys. All numeric fields must be numbers, not strings.
Use only information contained in the article as of its publication date -
do not use outside knowledge, future information, or speculation.

USER PROMPT:
You are evaluating, in a forward-looking manner, whether a given newspaper
article signals current trigger events or potential future trigger events
that could have a severe negative impact on economic activity and/or the
stability of the financial system within the euro area over the next
1 to 12 months.

Definitions:
Severe negative impact on economic activity: negative real GDP growth,
```

```
significant declines in industrial production, or significant increases
in the unemployment rate.

Severe negative impact on financial stability: losses for the banking system,
a significant increase in interbank lending rates, or large outflows of bank
deposits.

Forward-looking: the article contains statements or implications about
potential future events, conditions, or outcomes for the euro area
(e.g. forecasts, expectations, warnings, guidance, or predictions) over
the next 1-12 months.

Not forward-looking: trigger events that are already leading to, or have
already led to, a severe negative impact on economic activity and/or the
stability of the financial system in the euro area.

Important guidance:
Focus on implications for the euro area. If the article concerns individual
firms or non-euro-area events, set Class=1 only if plausible spillovers to
the euro area's economic activity and/or stability of the financial system
are indicated or strongly implied within 1-12 months.

If the article only describes current/past impacts with no credible future
euro-area implications, set Class=0.

Tasks for each article:
Class (0-1 integer): Classify whether the article covers trigger events
aligned with the above forward-looking criteria for the euro area (1) or
not (0).

Here is a newspaper article:
Title: '{title}'
Text: '{text}'

Return valid JSON ONLY with exactly these keys and types:
{
  "Class": <int>
}
```

Stage 3: Trigger-event characterisation

```
SYSTEM PROMPT:
You are an economic and financial analyst. Return a single JSON object only
(no code fences, no markdown, no explanations) and include exactly the
requested keys. All numeric fields must be numbers, not strings.
'Trigger_Source' must be one string chosen from the listed categories.
Use only information contained in the article as of its publication date -
do not use outside knowledge, future information, or speculation.

USER PROMPT:
You are evaluating articles that signal current trigger events or potential
future trigger events that could have a severe negative impact on economic
activity and/or the stability of the financial system within the euro area
```

```
over the next 1 to 12 months. It is your task to assess the nature of the
trigger event, and the associated probability, severity and time horizon.

Definitions:
Severe negative impact on economic activity: negative real GDP growth,
significant declines in industrial production, or significant increases
in the unemployment rate.

Severe negative impact on financial stability: losses for the banking system,
a significant increase in interbank lending rates, or large outflows of bank
deposits.

Not forward-looking: trigger events that are already leading to, or have
already led to, a severe negative impact on economic activity and/or the
stability of the financial system in the euro area.

Tasks for each article:

Prob (0-3 integer): Assess the probability that the trigger event described
in the article could have a severe impact on euro area economic activity
or financial stability in the future:
0 = Unlikely, meaning zero probability
1 = Low probability, highly speculative, unlikely to materialise
2 = Medium probability, possible outcome, could potentially materialise
3 = High probability, very plausible outcome, likely to materialise

Sev (0-3 integer): Assess the severity of the negative impact described
in the article if it materialises:
0 = not severe, limited or negligible effect on euro area output, employment,
    or financial conditions
1 = mildly severe, noticeable but contained effect
2 = severe, broad-based or macroeconomically significant impact
3 = very severe, systemic or crisis-level impact across the euro area

Hor (0-4 integer): Assess the time horizon when the impact is most likely
to materialise:
0 = impact is already occurring/already occurred
1 = within 1-3 months
2 = within 3-6 months
3 = within 6-12 months
4 = within 12-36 months

Trigger_Source (string): Select exactly one of the following categories
that best describes the source/nature of the trigger event discussed in
the article:
- Macroeconomic events
- Financial market events
- Geopolitical events
- Monetary policy events
- Fiscal policy events
- Other exogenous events
- No shock/trigger event

Here is a newspaper article:
```

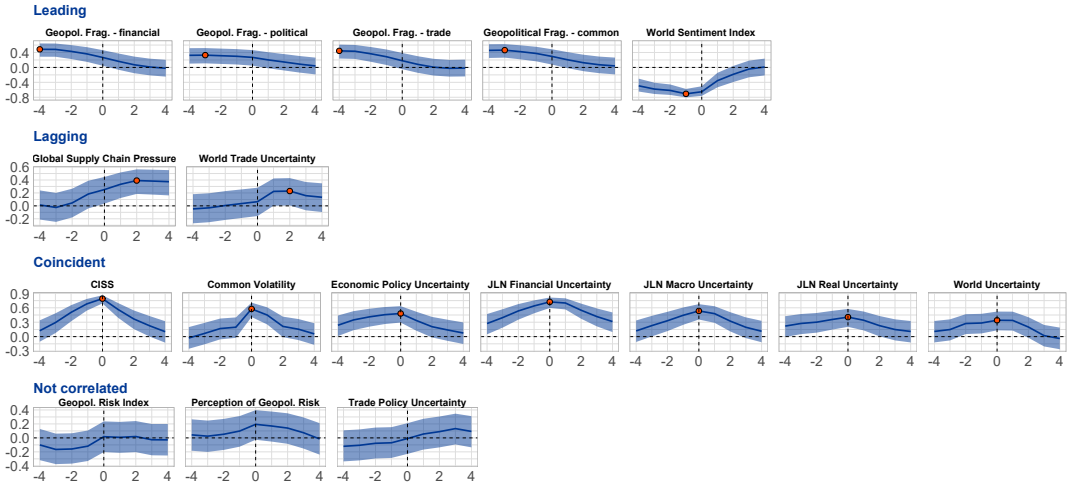
```
Title: '{title}'
Text: '{text}'

Return valid JSON ONLY with exactly these keys and types:
{
  "Prob": <int>,
  "Sev": <int>,
  "Hor": <int>,
  "Trigger_Source": "Macroeconomic events" |
                    "Financial market events" |
                    "Geopolitical events" |
                    "Monetary policy events" |
                    "Fiscal policy events" |
                    "Other exogenous events" |
                    "No shock/trigger event"
}
```

A.2 Additional data and indicator characteristics

Table A.1: Data and Classification Summary

Dataset characteristics	**Value**
Raw total number of articles	1,065,164
Sample period	Jan 2005 – May 2026
Average article length (words)	599
Average article length (characters)	4,186
Classification pipeline	
Articles entering Stage 1	1,065,164
Excluded in Stage 1 (non-economic content)	411,203 (38.6%)
Articles entering Stage 2	653,961
Excluded in Stage 2 (non-trigger events)	507,791 (77.6%)
Articles entering Stage 3	146,170

Figure A.1: Lead-lag relationship between SPOT and other indicators



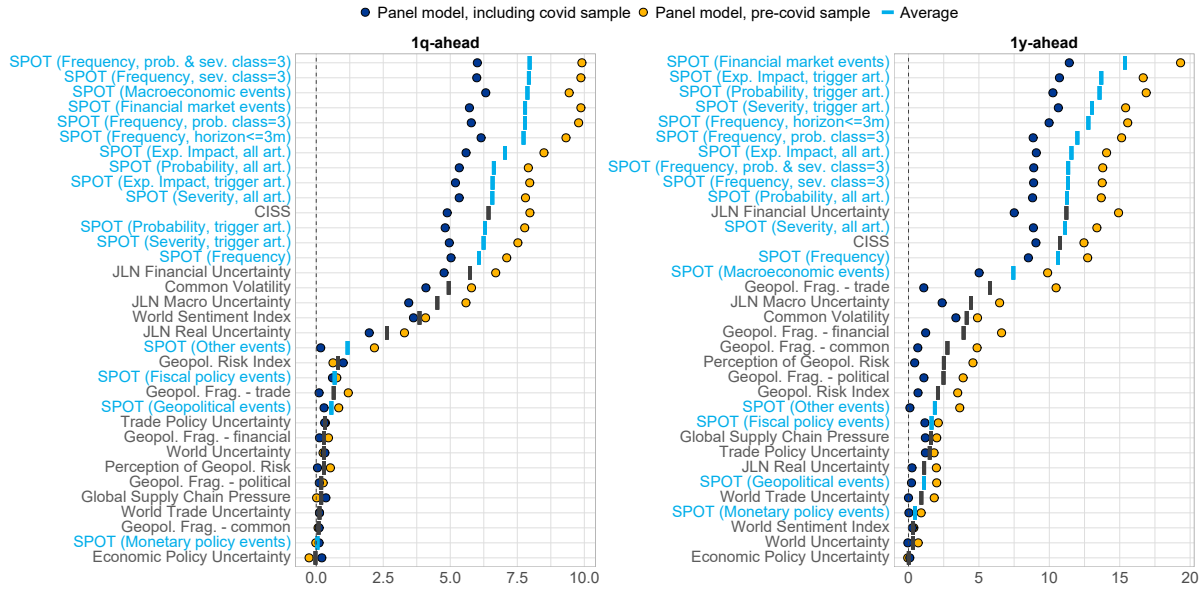
Notes: Computed on quarterly sample 2005Q1-2024Q4 using 3m averages of monthly values. Red dot denotes maximum correlation (significant at 5% level). Blue lines denote cross-correlations between SPOT - expected impact (all articles) and a given indicator at various leads (negative x-axis) and lags (positive x-axis). Light blue area denotes 95% confidence intervals.

A.3 Robustness of Growth-at-risk results

Because various modelling choices might affect the growth-at-risk results reported in the main part of the paper, we also estimate alternative specifications to assess the robustness of the results. First, we test whether the results are sensitive to the estimation sample. A key concern is that the large drops in real GDP during the COVID pandemic might disproportionately drive the results, even though the shock was exogenous. To address this, we include a set of dummy variables in our baseline specification reported in the main part of the paper. As an additional check, we re-estimate the models on a sample that ends in 2019 Q4. Figure A.2 reports these results, where the yellow dots represent results for the pre-COVID sample. The results indicate that the relative performance ordering of the indicators for the pre-Covid sample is very similar to that for the model estimated on the full sample including the COVID period, although including the COVID period in the sample reduces overall model performance somewhat.

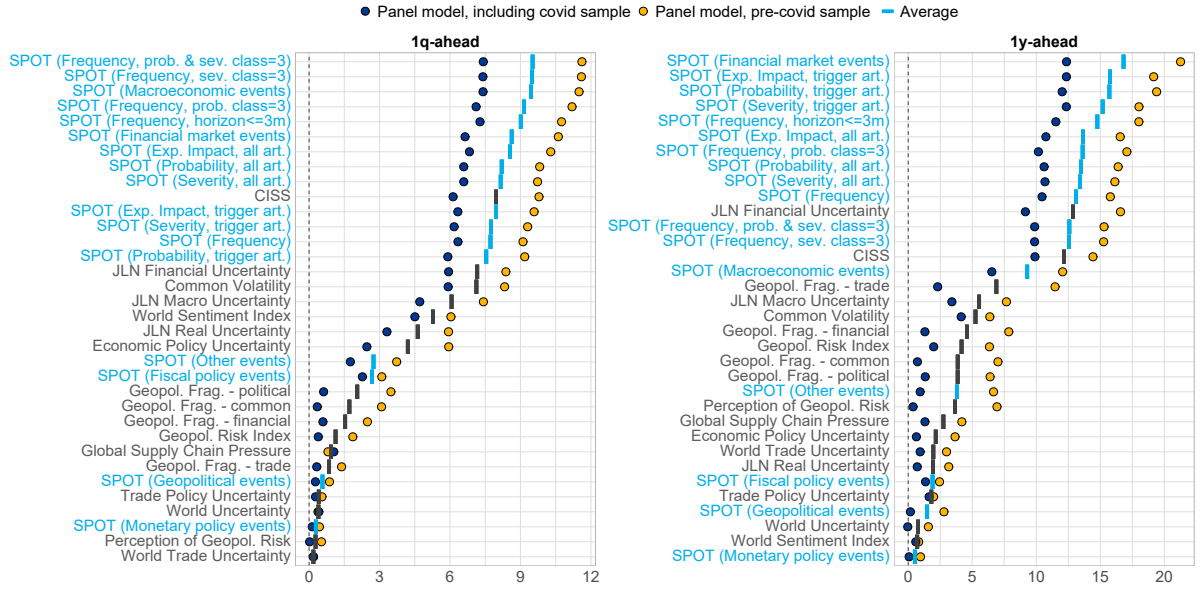
Similarly, Figure A.3 compares models with interactions estimated on the sample that includes the COVID period (dark blue dots) with those estimated on the sample that ends before the pandemic (yellow dots). As with the models without interactions, the key findings of the benchmark specification continue to hold.

Figure A.2: Performance of the indicators for various sample periods



Notes: Improvements in tick loss function are in percentage relative to the baseline model with GDP and SRI. The models are ordered by the average performance of indicators on the two samples.

Figure A.3: Performance of the indicators with interactions for various sample periods

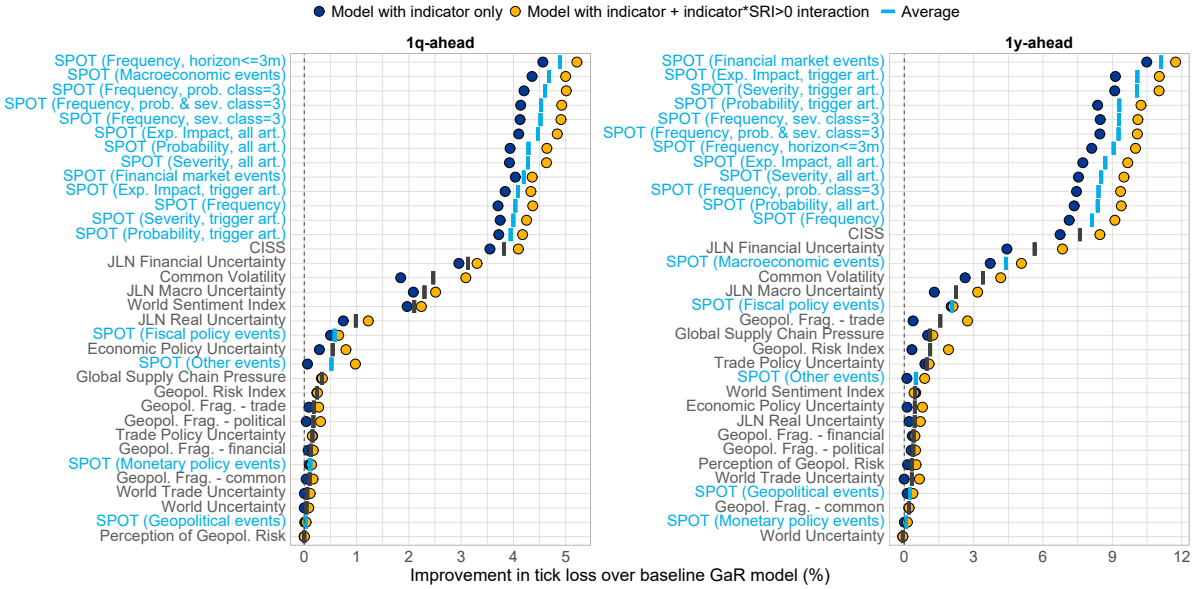


Notes: Improvements in tick loss function are in percentage relative to the baseline model with GDP and the SRI (including its lag and a dummy variable for cases when SRI>0). The models are ordered by the average performance of indicators on the two samples.

Next, we examine how sensitive the results are to the choice of quantile level. Setting $\tau = 0.1$ means that the estimates are driven by the lowest 10% of observations. While focusing on the lower tail is in line with the key goal of the study - namely, to capture rare, extreme crisis periods

that are most relevant for financial stability surveillance - the number of influential observations is rather small. For example, for quarterly data over a 20-year period, the results are driven by only the eight lowest values for each country. We therefore re-estimate the models for the 25th quantile ($\tau = 0.25$), which still reflects the lower part of the GDP growth distribution but is based on more observations. These results, reported in Figure A.4, indicate that the key messages from the benchmark specification remain largely unchanged, thereby corroborating the robustness of our findings.

Figure A.4: Performance for models evaluated at different quantile ($\tau = 0.25$)



Notes: Improvements in tick loss function are in percentage relative to the baseline model with GDP and SRI. The models are ordered by the average performance of indicators.

Acknowledgements

We would like to thank participants at the National Bank of Slovakia Financial Stability Seminar, the IMF MCM Policy Forum, the BoF/ESRB Conference on AI and Systemic Risk Analytics, and the ECB Financial Stability Seminar for helpful comments and suggestions. Large language models were used to create the SPOT indicator, as described in more detail in the paper.
The views expressed in this paper are those of the authors and do not necessarily reflect those of the European Central Bank.

Domenic Kellner
European Central Bank, Frankfurt am Main, Germany; email: domenic.kellner@ecb.europa.eu

Jan Hannes Lang
European Central Bank, Frankfurt am Main, Germany; email: jan-hannes.lang@ecb.europa.eu

Lukas Joseph Nagy
Marburg University, Marburg, Germany; email: lukas.nagy@uni-marburg.de

Marek Rusnák
European Central Bank, Frankfurt am Main, Germany; email: marek.rusnak@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

PDF ISBN 978-92-899-7991-7 ISSN 1725-2806 doi:10.2866/7426173 QB-01-26-190-EN-N