The AI-GPR Index: Measuring Geopolitical Risk using Artificial Intelligence*

Matteo Iacoviello† Jonathan Tong‡

July 13, 2026

Abstract

We introduce an improved measure of geopolitical risk, constructed by applying artificial intelligence to newspaper articles. Our approach replaces traditional keyword-based methods with semantic understanding: instead of searching for specific word combinations, we use OpenAI's GPT-4o mini model to read newspaper articles and assess their geopolitical risk intensity. The daily AI-GPR index scores around 4.6 million articles from the New York Times, Washington Post, and Chicago Tribune from 1960 through 2025. The approach reduces false positives from articles mentioning war or terrorism in non-geopolitical contexts while capturing relevant articles that lack exact or common dictionary terms. The AI-GPR index also assigns gradations of risk intensity rather than simple binary yes-or-no classifications, providing more nuanced measurement. We demonstrate the potential of our approach with several applications. The AI-GPR index produces a more precisely estimated negative effect of geopolitical risk on stock returns, and decomposing it into anticipated threats and realized acts reveals that markets price the two differently. A second LLM classification layer yields a time series of geopolitical risk-driven oil supply disruptions by region, which we use to trace the macroeconomic effects of geopolitical oil shocks. Finally, the layered architecture decomposes geopolitical risk by country and by directed country pair, and we show that the resulting bilateral index predicts a contraction in trade between countries as a result of rising bilateral geopolitical tensions.

JEL Classification: D80, E66, F51, G15.

Keywords: Geopolitical Risk, Textual Analysis, Large Language Models, Measurement Error, Content Classification.

*We thank David Sorensen, Flora Haberkorn, the Proquest TDM Studio team, and seminar participants at the Federal Reserve Board, the Financial Risks Forum in Paris, University of Naples Partenope and Cambridge University for helpful comments. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of anyone else associated with the Federal Reserve System. The data presented in this paper are available at https://www.matteoiacoviello.com/ai_gpr.html.

†Federal Reserve Board of Governors and CEPR. Email: matteo.iacoviello@frb.gov

‡University of Wisconsin. Email: jtong45@wisc.edu

1 Introduction

Geopolitical risk—encompassing wars, terrorism, and tensions between nations—is an important driver of macroeconomic fluctuations. Central to the empirical understanding of its effects is the ability to quantify latent political tensions from sparse and noisy data. Caldara and Iacoviello (2022) established a framework for this measurement with the Geopolitical Risk (GPR) Index, using dictionary-based keyword counts to measure geopolitical instability through newspaper archives.[1] Their index, measured as the share of newspaper articles discussing geopolitical risks, has since become a standard benchmark in the literature as a proxy for the risks that shape trade patterns, investor decisions, and asset prices.

While keyword-based methods have proven effective at capturing global risks and other economic phenomena, they are subject to the inherent constraints of dictionary-based matching. These methods struggle to distinguish between articles where a geopolitical event is the primary subject from those where keywords appear in only a peripheral or historical context. Furthermore, fixed dictionaries may miss evolving terminology or fail to capture the nuance of diplomatic language. This paper builds directly upon the conceptual architecture of Caldara and Iacoviello (2022), but proposes a methodological refinement: the transition from keyword matching to semantic understanding.

We introduce the AI-GPR Index, constructed by applying a Large Language Model (LLM), specifically OpenAI's GPT-4o mini, to a corpus of approximately 4.6 million articles from the *New York Times*, *Washington Post*, and *Chicago Tribune* between 1960 and 2025. Our approach maintains the definition of geopolitical risk established in the prior literature but utilizes the LLM's contextual awareness to score articles on a continuous scale. This allows for a more granular identification of risk, distinguishing the relative severity of geopolitical events as they are presented in the news. The resulting index complements the original GPR by offering a smoother and more accurate index that retains the transparency of the original framework through auditable prompts and classification criteria.

We demonstrate the value of the AI-GPR index through a number of empirical applications. First, we revisit the relationship between geopolitical risk and financial markets. We find that the AI-GPR index provides a more precisely estimated negative effect of geopolitical innovations on stock market returns, particularly when isolating the persistent component of risk. A further classification layer decomposes the index into anticipated threats and realized acts, and find that the two are priced differently: markets price the gradual and persistent buildup of threats, and respond to realized acts only insofar as they are unanticipated.

Second, we employ a layered classification architecture to identify articles specifically related

[1] The use of dictionary-based methods itself builds on the foundational work of Baker, Bloom, and Davis (2016) and Hassan et al. (2019), among others.

to geopolitical oil supply disruptions. This yields a novel time series of GPR-driven oil shocks that can be disaggregated by geographic region. We validate the "Oil GPR index" using a proxy structural VAR, in the spirit of Verduzco-Bustos and Zanetti (2026), but replacing their original GPR instrument with a novel proxy constructed from oil price surprises on days when the Oil GPR index spikes; the identified shock generates a persistent rise in oil prices and a significant contraction in economic activity.

Third, we leverage the LLM to identify the country-level actors involved in geopolitical events and their specific roles as initiators, respondents, or spillover parties. We use these labels to reconstruct the directed actions between country-level actors. Altogether, this yields country-level indices of geopolitical risk that trace geopolitical risk involving a country over time, extending the country sub-indices of Caldara and Iacoviello (2022) to two hundred nations, along with high-frequency, directed country-pair indices that capture the source and target of bilateral tensions. By mapping these directed links, we can also identify the specific country pairs that drive global risk. We validate the bilateral index estimating its effects on bilateral trade flows, finding that in a standard gravity equation, higher bilateral GPR is associated with lower bilateral trade.

We make all the measures presented in this paper available to the research community in a regularly updated companion webpage. The scope of these new measures has itself been shaped by the community; the high-frequency country-level decompositions, the oil-market measures of geopolitical risk, and the bilateral stress indicators introduced in this paper each respond to requests and suggestions that have accumulated over the years from researchers, policymakers and market participants who have relied on and popularized the original index of Caldara and Iacoviello.

Our work contributes to a rapidly emerging literature applying LLMs to economic measurement (e.g., Ito, Sato, and Ota, 2025 and Clayton, Coppola, Maggiori, and Schreger, 2025). We demonstrate that LLM-based approaches do not replace, but rather augment, the dictionary-based methods that have defined the field. As computational costs continue to decline, we believe this high-resolution approach to text analysis will become a standard template for measuring complex political and economic phenomena that are otherwise hard to quantify. Our focus on geopolitical risk also connects our work to the growing literature on geoeconomics (see Mohr and Trebesch, 2025 and Clayton, Maggiori, and Schreger, 2026) and provides an improved set of data to make sense of geopolitical risk both within and across country borders.

Section 2 describes the data and the LLM-based methodology. Section 3 presents the AI-GPR index. Section 4 assesses the index's validity and robustness, contrasting it with keyword classification, validating it against human coding, bounding measurement error, and confirming insensitivity to the choice of model. Section 5 introduces three further dimensions of measurement: threats and acts, oil, and country and bilateral risk. Section 6 presents our empirical applications, and Section 7 concludes.

2 Data and Methodology

2.1 Data Sources

Our construction of the AI-GPR index sources newspaper data from the New York Times, Washington Post, and Chicago Tribune, all of which provide comprehensive international coverage and maintain consistent editorial standards over time. The articles we score span from 1960 to 2025 and are sampled daily.[2]

We adopt a two-stage process that first filters for articles likely to discuss geopolitical events and subsequently passes them through an LLM for scoring. The advantages of this method are twofold. First, the keyword screen efficiently discards many unrelated and irrelevant articles that carry no discussion of geopolitical risk at all. Second, for articles that pass the screen, the LLM classifies geopolitical risk content and intensity more accurately than query-based alternatives.

We use the following query on the ProQuest TDM Studio database to filter for articles likely to discuss geopolitical risk:

```
SEARCH QUERY:

PUBDATE(>=19600101) AND PUBDATE(<=20251231) AND LA(English) AND (airstrike*
OR alliance OR annex* OR attack* OR blockade* OR bomb* OR cease-fire OR
combat OR confrontation OR conflict* OR coup OR crisis OR diplomat* OR
diplomacy OR embargo OR enem* OR hostage* OR hostil* OR instability OR
invade* OR invasion* OR militia OR military OR missile* OR nuclear OR
refugee* OR riot* OR sanction* OR sovereign* OR terror* OR treaty OR troop*
OR truce OR unrest OR violence OR war OR weapon*) AND (publication("New
York Times") OR publication("Washington Post") OR publication("Chicago
Tribune")) AND DTYPE(article OR commentary OR editorial OR feature OR
'front page article' OR 'front page/cover story' OR news OR report OR
review)
```

The query focuses on articles containing at least one keyword potentially related to geopolitical risk. We further restrict our attention to articles categorized as news content rather than advertisements or non-editorial material. Our keywords span three categories of geopolitical risk: direct conflict and military action (war, invasion*, attack*, terror*, conflict*, ...), political crisis and instability (crisis, instability, unrest, coup...), and strategic responses and diplomacy (sanction*, embargo, blockade*...). Our search extracts 4.6 million articles from the universe of articles pub-

[2] The headline corpus counts, summary statistics, and computational-cost figures reported in this paper refer to the 1960–2025 period covered by the search query below. For the time-series figures and the financial-market regressions of Section 6.1, we extend the index through March 2026 by applying the identical sampling, screening, and scoring pipeline to articles published through that date. The companion webpage hosts the regularly updated series.

lished across the three publications over the sample period. Of these 4.6 million articles, 1.2 million (26 percent) receive a positive GPR score from the LLM. Section 4.3 verifies that the keyword filter has a negligible false-negative rate.

For each selected article, we extract metadata (publication date, headline) and its textual content from the ProQuest database. To manage computational costs, we truncate article text to the first 2,000 characters[3].

2.2 LLM Classification: Definition, Prompt, Implementation

We use Caldara and Iacoviello (2022)'s definition of geopolitical risk, which encompasses wars, escalation of existing events, terrorist attacks, and diplomatic tensions. This definition provides clear boundaries for defining geopolitical risk while encompassing the full range of relevant event types. Importantly, it focuses on current risks and events, not historical retrospectives, fictional depictions, or metaphorical uses of geopolitical terminology.

We use the following system prompt to instruct the LLM to evaluate articles:

```
CLASSIFICATION PROMPT:
You will be given a news article.  Classify the article's assessment of
geopolitical risk based only on what the article states or strongly implies.

Geopolitical risk is defined as the threat, realization, and escalation of adverse
events associated with Wars (initiation of new conflicts); Escalation of existing
wars; Major Terrorist attacks; Tensions between nations or involving cross-border
political actors that affect international relations.

Assign a geopolitical risk score from 0.0 to 1.0 based on the following scale:
0.0--0.2:  No mention of geopolitical risks.
0.2--0.4:  Mentions of minor tensions, diplomatic disputes, or isolated incidents
with limited escalation risk.
0.4--0.6:  Discussion of significant tensions, ongoing regional conflicts, or
moderate escalation risk; some concern about war or terrorism.
0.6--0.8:  Substantial discussion of major war initiation/escalation risks, active
terrorism threats, or high likelihood of significant escalation.
0.8--1.0:  Extensive coverage of imminent or new war, major war escalation, severe
terrorism threats, or critical threat to international stability.

Movies or books about geopolitical risks, stories remembering old events,
obituaries, anniversaries of old events should be given 0 unless they explicitly
mention current risks.
Note:  Purely domestic political events (elections, protests, internal policy
debates) should score 0.0 unless they have implications for international tensions
```

[3] Due to the inverted pyramid structure of journalism where the most important information is presented first, the truncation should capture the essential geopolitical content for the vast majority of articles. As we show in Section 4.3, our baseline scores based on truncated text are nearly identical to those obtained using the full article text.

```
or cross-border conflicts.
```

As the prompt illustrates, we use zero-shot prompting where the model receives explicit criteria and a scoring scale but no labeled examples[4]. The scale from 0.0 to 1.0 allows the model to rate articles with varying intensities of geopolitical risk rather than binary classification.

We set the LLM's temperature parameter set to zero, which minimizes stochastic variation in responses to ensure consistent scoring across multiple runs. We use structured output to return model output in JSON format, making extraction and processing of scores more efficient.[5]

For each article, we construct a request consisting of the system prompt and the article's textual content (the headline and first 2,000 characters of the article). All scores fall within the specified 0.0–1.0 range.

2.3 Index Construction

We construct the AI-GPR index by aggregating individual article scores at daily frequency. For each date t, we compute:

$$\text{AI-GPR}_t = \frac{1}{\bar{S}} \times \frac{1}{A_t} \sum_{i=1}^{N_t} S_{it} \tag{1}$$

where S_{it} represents the geopolitical risk score for article i at time t, N_t is the total number of scored articles published on date t, A_t is the total number of articles published on date t across all three newspapers (not just those matching geopolitical keywords), and $\bar{S}$ is a normalization constant chosen such that the index has a mean of 100 over the 1985–2019 period. Dividing by A_t accounts for variation in newspaper output over time and is adopted from Caldara and Iacoviello (2022) for consistency.[6]

The index thus measures the intensity-weighted share of newspaper coverage devoted to geopolitical risk. The numerator sums the risk intensity of each article while the denominator scales by total newspaper output. In Appendix A.2 we show that the index is robust to the main construc-

[4] One-shot or few-shot prompting would require embedding one or more example articles in every request; at the scale of approximately 4.6 million articles, even a single example would roughly double input token costs. The zero-shot approach keeps prompt length short and cost proportional to the number of articles, while the detailed instruction set substitutes for examples by giving the model precise, auditable criteria.

[5] Structured output is particularly valuable for classification tasks with multiple output dimensions, such as the second-stage classifiers introduced in Section 5.

[6] The denominator A_t counts articles containing at least seven common words ("the," "be," "to," "of," "and," "at," and "in"), as in Caldara and Iacoviello (2022). This choice filters out non-editorial content such as photos with a brief caption, classified advertisements, table-of-contents entries, stock quotes, sports results, and brief notices, and serves as a normalization for trends in newspaper article volume over the years. The total number of articles containing all these words is about 8.5 million. Were we to count every object tagged as an article, the total would be almost twice as large.

tion choices, particularly whether article scores enter on the intensive or the extensive margin. We verify that the index is not mechanically driven by long-run drift in the scored share N_t/A_t.

Having the AI-GPR index aggregated daily allows it to capture high-frequency fluctuations in geopolitical risk. Unlike monthly or quarterly indices, the daily frequency preserves information about short-lived risk spikes and allows for better identification of the specific timing of geopolitical events. This precise temporal resolution is particularly valuable for event studies and examining the immediate market reactions to geopolitical developments.

2.4 Computational Considerations

A practical concern for LLM-based approaches is computational cost. At current pricing, processing one article through GPT-4o mini costs approximately \$0.0001 for our truncated text length. For our sample of approximately 4.6 million articles from 1960–2025, the total computational cost is approximately \$450. These costs will continue to decline as API pricing falls, which should enable larger and more widespread usage of LLMs for sentiment analysis.

We also find that the processing time is manageable. Scoring one article takes roughly one second. With parallelization across five API calls, we can process the full corpus of 4.6 million articles in approximately 250 hours. These practical considerations suggest that LLM-based text classification is increasingly feasible for large-scale economic research applications, including the construction of long historical time series.[7]

3 The AI-GPR Index

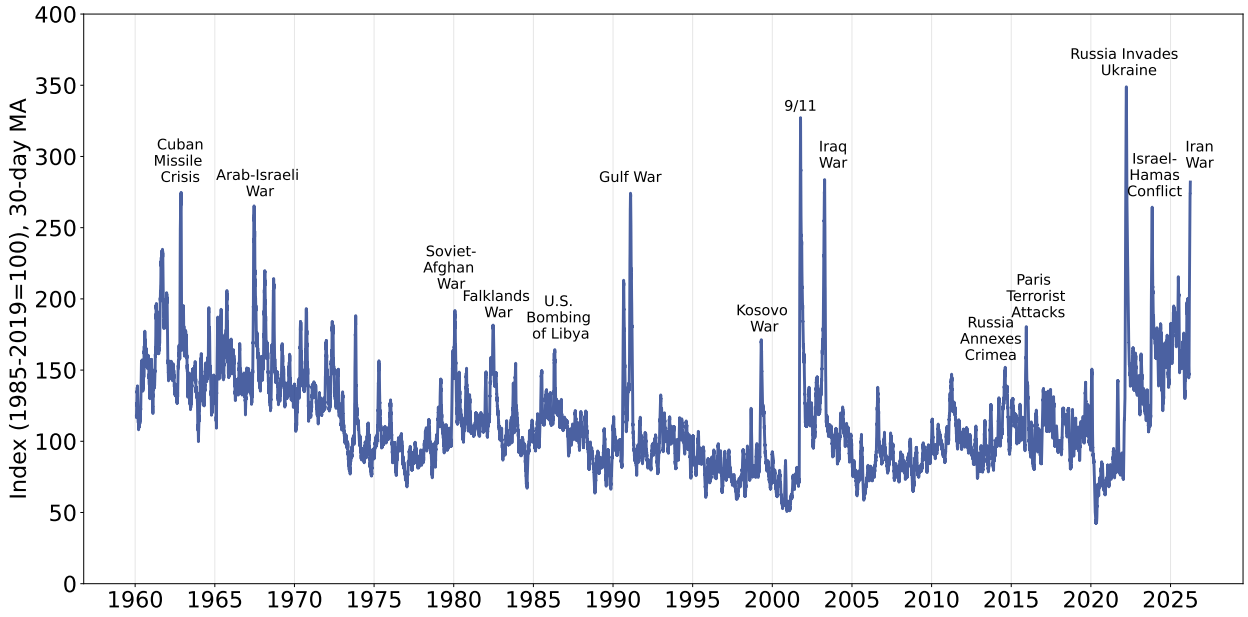
Figure 1 displays the AI-GPR index over the full sample, with labels marking the major geopolitical episodes of the past six decades.

3.1 Summary Statistics

Table 1 presents summary statistics for the AI-GPR index alongside the keyword-based GPR index over the 1960–2025 period as a benchmark. The keyword-based index applies the Caldara and Iacoviello (2022) methodology—which classifies articles using proximity searches of geopolitical and risk-related terms—to the same three newspapers used for the AI-GPR (New York Times, Washington Post, and Chicago Tribune). Caldara and Iacoviello (2022) publish two versions of the index: a historical index based on three newspapers that extends back to 1900, and a recent index based on ten newspapers that begins in 1985. Our version is more comparable with the three-newspaper version, where we sample from the same three newspapers as the AI-GPR back

[7] Using larger and more powerful models, such as GPT-5-mini, is more costly and requires longer processing times, but may of course become feasible over time as these models grow more powerful and cheaper.

Figure 1: The AI-GPR Index



Notes: The AI-GPR index plotted as a 30-day moving average, from January 1960 through March 31, 2026. The index sums LLM-assigned geopolitical risk scores (0–1) across all articles each day, normalized by total newspaper article count, and is scaled to a mean of 100 over 1985–2019. Regularly updated data are available on the companion webpage. Source: authors' calculations using ProQuest TDM Studio data.

to 1960. The full search string is reported in Appendix A.5. Both indices are normalized so that the mean equals 100 over 1985–2019.

Several features of the AI-GPR index stand out, as reported in Table 1. It is markedly smoother than the original GPR with a lower standard deviation and a thinner right tail. Its first percentile sits well above zero, with no days at all recording zero geopolitical content, consistent with the fact that even quiet days carry some low-level geopolitical discussion. The AI-GPR is also more persistent, with a higher 90-day autocorrelation, reflective of the fact that geopolitical conditions evolve smoothly rather than spiking and reverting.[8]

3.2 Comparison with the Original GPR Index

Figure 2 plots the AI-GPR alongside the original GPR index at monthly frequency. The two series co-move closely overall but exhibit notable divergences.

In the 1960s–1970s, the AI-GPR index is elevated relative to the original GPR, perhaps indicative of the LLM's ability to non-conventional language to describe geopolitical tensions that

[8] We measure persistence as the autocorrelation of each index's 90-day moving average: specifically, the correlation between the smoothed series on day t and on day $t-90$. This statistic captures slow-moving variation in geopolitical conditions rather than day-to-day volatility, and is consistent with the definition reported in Table 1.

Table 1: AI-GPR Index: Summary Statistics

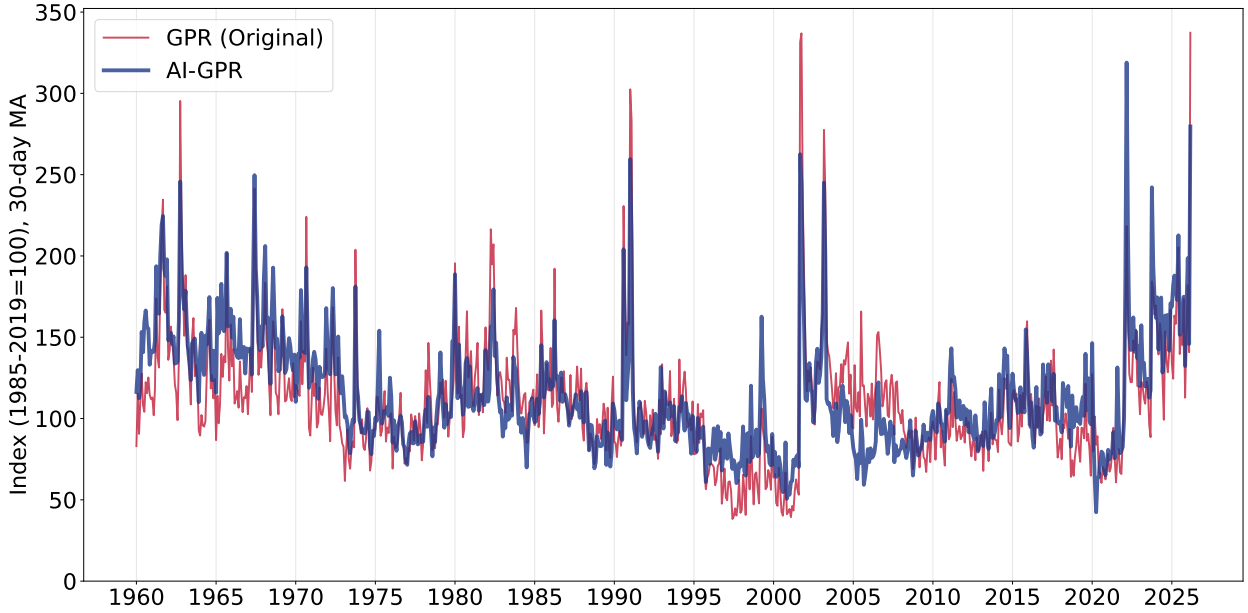
Statistic	**AI-GPR**	**GPR (Original)**
Total newspaper articles	8,514,822	8,514,822
Articles sampled	4,571,496	—
Articles scored positive (> 0)	1,166,083	262,096
Standard Deviation	48.52	60.55
Skewness	1.64	1.97
1st Percentile	41.97	18.14
Median	105.46	99.15
99th Percentile	266.95	300.16
90-day Autocorrelation	0.73	0.62
Correlation with Original GPR	0.69	1.00
Positive Articles Share (%)	15.00	3.28
Positive Among Sampled (%)	26.16	3.28

Notes: Summary statistics computed over the 1960–2025 sample. The "GPR (Original)" column is the historical GPR index of Caldara and Iacoviello (2022), computed on the same three newspapers as the AI-GPR over 1960–2025. Both indices are normalized to a mean of 100 over 1985–2019. The 90-day autocorrelation is the correlation between the 90-day moving average on date t and on date $t-90$. Positive Articles Share (%) is the share of all newspaper articles classified as containing geopolitical risk: for the AI-GPR, articles receiving a positive score; for the original GPR, articles matching the keyword proximity search. Positive Among Sampled (%) is the average daily share of positive articles among those in the queried sample: for the AI-GPR, the share of articles sent to the LLM that receive a score > 0; for the original GPR, this coincides with Positive Articles Share (%) since there is no separate sampling stage. The first three rows report totals over the full sample: total newspaper articles (the normalization denominator), articles sampled (sent to the LLM for scoring; the original GPR has no separate sampling stage), and articles scored positive (AI-GPR score > 0, or a keyword match for the original GPR).

aren't in the keyword dictionary. The Cold War-era tensions, the Vietnam War, and the Yom Kippur War all register prominently. In the 1990s, the AI-GPR index is higher during the Bosnian War (1992–1995), suggesting that the LLM captures the sustained geopolitical significance of this conflict more than keyword counting. In the 2000s, both indices spike sharply after the September 11 attacks and during the Iraq War. The AI-GPR exhibits a somewhat faster decline after the initial Iraq War spike, consistent with the LLM's ability to distinguish between articles primarily covering active conflict versus articles mentioning the Iraq war in other contexts. From 2010 onward, the indices track each other closely, though the AI-GPR is somewhat more elevated in recent years. This may reflect the LLM's ability to capture emerging geopolitical risks (e.g., cyberwarfare, economic coercion) that do not always trigger traditional keyword matches.

Consistent with the finding of Carlson and Dell (2025), the AI-GPR index finds that the share of articles discussing geopolitical risks is higher than in the original GPR tabulations. As shown in Table 1, the AI-GPR classifies 15 percent of all articles as mentioning geopolitical risks, compared

Figure 2: AI-GPR and Original GPR Index



Notes: The AI-GPR index (blue, thick) and the original GPR index (red, thin), both plotted at monthly frequency. The original GPR here is the three-newspaper version of the Caldara and Iacoviello (2022) index—their historical index, which extends back to 1900—computed on the same three newspapers as the AI-GPR; the comparison is with this version rather than their ten-newspaper index, which begins in 1985. Both indices are normalized to a mean of 100 over 1985–2019. The AI-GPR index sums LLM-assigned geopolitical risk scores across all articles each day, normalized by total newspaper article count, then averaged to monthly frequency. The original GPR counts articles matching specific keyword proximity searches. Source: authors' calculations using ProQuest TDM Studio data. Last observation: March 2026.

to just 3.3 percent under the original, keyword-based approach.[9] The two indices nevertheless remain substantially correlated and track each other closely around major events, suggesting that while the keyword approach may understate the share of geopolitically relevant newspaper articles, it still captures the same peaks and troughs that define the phenomenon.

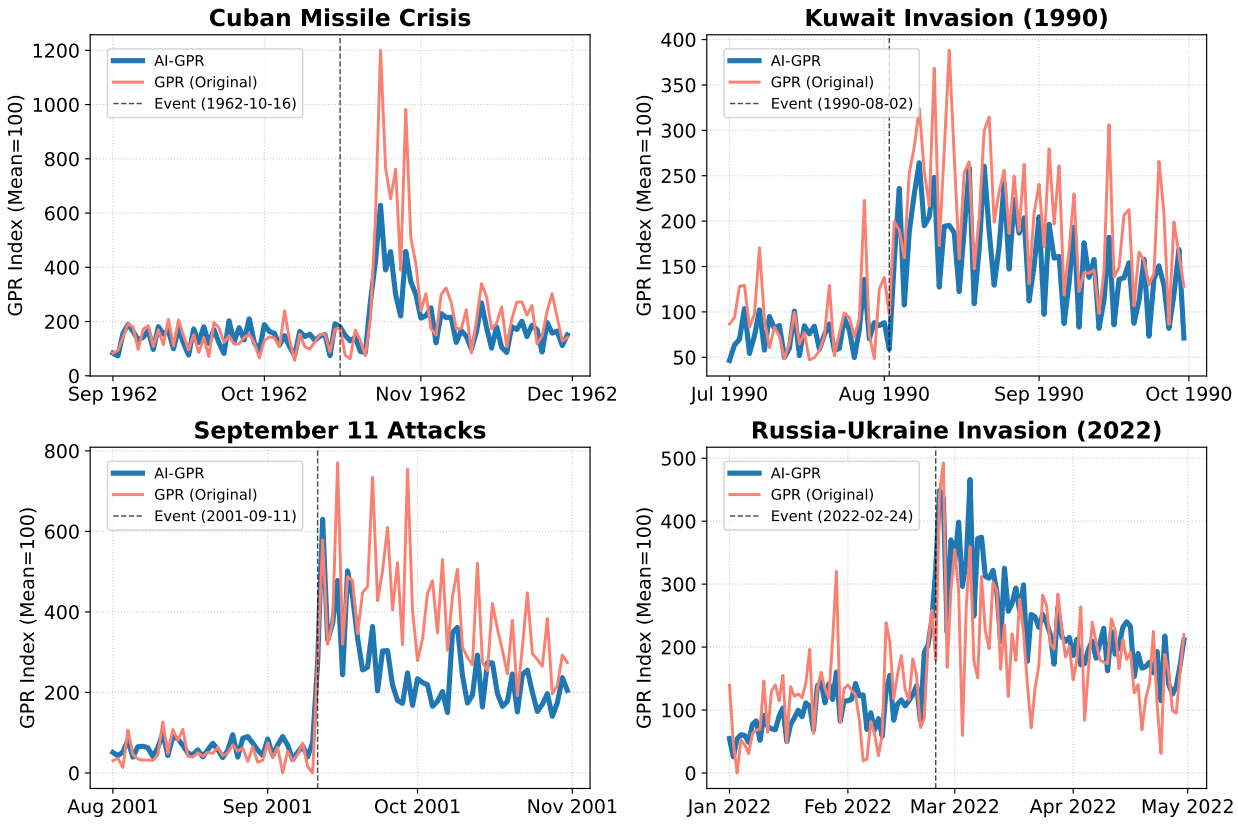
3.3 Performance on Historical Events

Figure 3 compares the AI-GPR and original GPR around four major geopolitical events: the Cuban Missile Crisis (1962), the Kuwait Invasion (1990), the September 11 Attacks (2001), and Russia's invasion of Ukraine (2022). Both indexes spike at the onset of each event, and both indexes reveal some effects driven by the weekly cycle in newspaper articles volume. This high-frequency variation disappears when using a standard 7-day moving average (not shown).

For more recent episodes like September 11 and the Russia-Ukraine invasion, the AI-GPR

[9] Among articles sent to the LLM, 26.2 percent receive a positive score.

Figure 3: AI-GPR vs. Original GPR Around Key Geopolitical Events



Notes: Each panel shows the daily AI-GPR index (blue, thick) and original GPR index (salmon, thin) around a major geopolitical event. Vertical dashed lines mark the event date. Both indices are at daily frequency (no smoothing) and normalized to a mean of 100 over 1985–2019. Source: authors' calculations.

index shows less day-to-day volatility than the original index. One possible explanation is that any journalistic language and style is interpretable by the LLM, so the AI model can parse varied writing styles more accurately than a keyword-based approach that relies on the exact appearance of specific terms. For earlier episodes such as the Cuban Missile Crisis, the AI-GPR also captures the risk elevation but with a smoother profile, perhaps because the AI approach is less sensitive to day-to-day variation in keyword frequency.

4 Validation and Robustness

Having presented the index, we now assess its validity and robustness. We first contrast the LLM classifier with keyword-based selection, then validate the scores against careful human coding, quantify the remaining sources of measurement error, and confirm that the index is insensitive to the choice of language model.

4.1 LLM versus Keyword-Based Classification

Our LLM-based approach differs from traditional keyword-based classification methods in several fundamental ways. Caldara and Iacoviello (2022) classify articles by searching for specific combinations of words from predefined categories. For example, an article is classified as relevant to "war threats" if it contains a geopolitical term ("war," "conflict," "hostilities") within a specified proximity of a risk term ("threat," "danger," "crisis"). To reduce false positives, the original GPR methodology employs exclusion words like "movie," "anniversary," and "obituary" that disqualify an article from classification even if it contains the requisite geopolitical keywords.

This exclusion strategy addresses some common sources of false positives but remains imperfect. Articles about war films may not use the word "movie" and still discuss fictional content. Historical retrospectives may avoid words like "anniversary" while still focusing on past events rather than current events. Moreover, the binary nature of their exclusion methodology (article is either included or excluded entirely) cannot handle articles that legitimately discuss both current geopolitical risks and historical context.

Our approach handles these cases more flexibly by leveraging the LLM's semantic understanding. Rather than applying binary exclusion rules, the model evaluates whether geopolitical content is central to the article and whether it pertains to current rather than historical events. An article that mentions a war film in passing while primarily analyzing current military tensions would receive an appropriate non-zero score, whereas an article primarily reviewing a war movie would correctly receive a zero score.

Beyond the false positive and false negative concerns, keyword-based approaches face additional limitations. First, keyword matching is context-insensitive: it cannot distinguish whether geopolitical content is central or peripheral to an article, so an article titled "The Price of War in Ukraine" receives the same binary classification as one mentioning "the war on inflation" in a subordinate clause. Keyword-based matching weighs every occurrence of a geopolitical term equally, so it tends to overvalue the geopolitical content of articles that happen to contain such terms. Second, the approach is terminology-dependent, requiring exhaustive keyword lists; articles discussing "armed confrontation" or "military standoff" may be missed if these exact phrases are not in the dictionary or if the proximity requirements are not satisfied. In practice, it is neither realistic nor feasible to create a query that accounts for all possible geopolitical terms used in newspapers, so some articles will always be missed with keyword-based approaches. Third, binary classification is not able to differentiate the intensity or centrality of geopolitical content. Our LLM-based approach addresses these three limitations through semantic understanding. The model evaluates whether geopolitical content is central to the article, uses context to distinguish current risks from historical references, recognizes semantically relevant content regardless of specific terminology, and provides graduated scores reflecting severity. Rather than applying rigid exclusion rules, the model exercises contextual

judgment about relevance, which we believe resolves the biggest flaws of the original approach.

The tradeoff is reduced transparency. While keyword lists and exclusion rules are fully auditable, LLM-based classification relies on complex neural networks and is effectively a black box. To introduce as much predictability and human oversight as possible into the LLM's classifications, we make deliberate choices in how we prompt and call the model. Our prompt-based instruction set provides explicit criteria, and we set the temperature of the model to 0 to minimize output randomness. The model's decisions can be examined by reviewing its classifications on validation samples, and from our manual auditing of sample articles, the model's scoring is consistent with our own judgment.

4.2 Validation Against Human Judgment

To assess whether the model systematically over- or under-scores articles relative to a human reader, we hand-scored a random sample of about 2,000 articles drawn from all three newspapers, applying the same 0-to-1 rubric given to the LLM. Each article was scored by an independent human coder and a large subset by a second, which lets us benchmark the model not only against a human reader but against how closely the human coders agree with one another.[10] Taking the consensus of the human coders as the reference, the machine scores track it closely: their correlation is 0.85, and the two agree on whether an article carries any geopolitical risk in over 90 percent of cases. This exceeds the agreement the human coders reach with one another, who correlate at 0.78 on the double-coded articles, so the model reads geopolitical risk at least as consistently as a second human would. Because the AI-GPR is a continuous score rather than a binary label, classification-error rates are only a coarse summary of this agreement, but they permit a like-for-like comparison with keyword selection: dichotomizing at any positive value, about 13 percent of the articles the model flags are ones a human coder scores as zero, against about 16 percent for the original Caldara and Iacoviello keyword query applied to the same articles—below the 21 percent they report from their own audit.[11] On a common footing, then, a smaller share of what the AI-GPR flags consists of articles a human reader judges to carry no geopolitical risk. Appendix A.3 reports the full three-way comparison; the model's and the humans' average scores are essentially identical, so the classifier introduces no aggregate bias in levels. The pattern we find—a large recall gain over keyword selection at no cost in precision—mirrors what Hartley (2025) documents for the Economic Policy Uncertainty index of Baker, Bloom, and Davis (2016), where language-model classifiers evaluated on the original human audit raise recall from 45 to 82 percent while precision

[10] Full details of the audit—the inter-rater agreement, the cross-tabulation of human and machine scores, a taxonomy of the disagreements, and illustrative cases—are reported in Appendix A.3.

[11] This rate is the share of flagged articles—those receiving a positive score—that the human scores as zero, the quantity comparable to the figure Caldara and Iacoviello report, which is likewise defined over the articles their keyword search selects. We compute it over individual article–coder ratings rather than the coders' consensus, so that an article one coder scores as zero and another as positive is not averaged into a positive benchmark.

also improves, suggesting that the advantage of semantic over dictionary classification is a general feature of news-based indexes rather than specific to geopolitical risk.

Because the two scores diverge only modestly, almost always by a small value on the scale, the disagreements are few enough to read individually, and doing so is informative about how to interpret the machine index. Where the model scores above the human, two patterns recur: it occasionally tends to read cross-border economic and financial disputes—tariffs, currency crises, frayed trade ties—and episodes of purely domestic unrest as geopolitical risk, even though the rubric reserves the concept for war, terrorism, and military tension. Its errors of timing are roughly offsetting: it sometimes scores retrospective coverage as though the events were current, and sometimes misses the fresh aftermath of an attack, so the two biases largely cancel in the aggregate. At the top of the scale the model is conservative, rarely using the highest gradations, so it modestly understates the intensity of the most extreme episodes. A further set of disagreements arise where the human credited context that an article only implies, and so reflect ambiguity in the coding rubric as much as model error. On balance, given this high correlation and near-universal agreement, the audit supports treating the AI-GPR scores as a reliable proxy for careful human coding.

Carlson and Dell (2025) propose a statistical correction for indices built from text classifiers. Their key insight is that keyword- or model-based classifiers produce biased proxies for the true concept of interest, and that this bias can be removed by combining classifier predictions with a human-labeled validation sample that covers the full span of the data. Applied to the GPR index of Caldara and Iacoviello, they find that the keyword-based series underestimates geopolitical risk once the correction is applied. A practical requirement of this approach is a validation sample large and representative enough to pin down the bias across the span of the data. Our LLM approach instead reduces this bias at the classification stage—by replacing keyword matching with semantic understanding—producing a similarly elevated series directly. The two approaches are complementary: because the LLM lowers classifier error at the source, the residual bias their framework targets is smaller to begin with, and their correction could still be applied to the AI-GPR as a further refinement. Two features of our audit would matter for such an application. Their framework requires a validation sample that is representative of the corpus (missing at random), whereas our audit panel deliberately oversamples articles the model scores as geopolitical, so only its random component—or a suitably reweighted version—would qualify. And the correction can only be as good as the human labels it is anchored to: our double-coding shows that even careful human coders disagree with one another somewhat more than the model disagrees with their consensus, a bound on annotation quality that their framework takes as given.

4.3 Measurement Error

Four potential sources of measurement error affect the AI-GPR index: model stochasticity, text truncation, attenuation from the two-stage filtering process, and hindsight knowledge embedded in the language model.

Model stochasticity. Even at temperature zero, LLMs are not deterministic and may produce slightly different outputs across calls. To quantify this, we run GPT-4o mini three times on the same subset of 1,050 articles. Scores are very nearly identical across the three runs—the vast majority of articles receive exactly the same score—confirming that classification noise from model stochasticity is negligible. Raising the temperature to 0.5 and 1.0 barely changes this: within- and cross-temperature agreement remains high and mean scores are virtually indistinguishable, so the choice of temperature has negligible impact on classification outcomes.[12] We further bound this concern through the multi-model comparison reported in Section 4.4, where scores are highly correlated across four OpenAI models and an open-weight Llama model—spanning two generations, different model sizes, and different model families—implying that classification is driven by article content rather than idiosyncratic model behavior.

Text length. Our baseline classifier truncates article text at 2,000 characters to reduce processing time and cost. To verify that this truncation does not materially affect classification, we re-score the same articles using their complete text. Truncated and full-text scores are highly correlated and identical for the large majority of articles; full-text scores are only modestly higher on average—consistent with longer articles carrying additional geopolitical content beyond the opening paragraphs—while the cross-article ranking is nearly unchanged.[13] We conclude that truncation introduces minimal classification error.

First-stage attenuation. The keyword pre-filter could introduce measurement error by missing geopolitically relevant articles (false negatives) or by including irrelevant ones (false positives). We assess the false-negative rate by applying the LLM classifier to a random sample of articles that did not match any of our search keywords. Only 0.9 percent received a positive geopolitical risk score, and those that did had uniformly low scores, confirming that the keyword filter captures the vast majority of geopolitically relevant content with a negligible false-negative rate. This compares favorably with the 2.6 percent false-negative rate reported by Caldara and Iacoviello (2022),[14] whose

[12] Across the three temperature-zero runs the pairwise correlation is 0.99, with at least 98 percent of articles scored identically and a mean absolute difference below 0.004. Cross-temperature agreement between temperature 0.0 and 1.0 is 93.2 percent, with a mean absolute difference of 0.014. Correlation and agreement against the human-annotated sample remain very high and similar across the temperature 0.5 and 1.0 runs.

[13] The correlation between truncated and full-text scores is 0.93, with 86.7 percent of articles scored identically; the mean score rises from 0.129 to 0.153 with full text.

[14] See Appendix B.6 in Caldara and Iacoviello (2022).

keyword filter permanently excludes articles without a second-stage review. False positives from our keyword filter are handled naturally by the LLM scoring step: articles that match keywords but do not discuss geopolitical risk receive a score of zero and do not contribute to the index.

Hindsight knowledge. Because GPT-4o mini is trained on text through a recent cutoff, it scores historical articles already knowing how the events they describe unfolded—that the Cuban Missile Crisis was resolved without war, or how the response to September 11 played out. Were the model to lean on this knowledge, it could systematically misstate the risk perceived by a contemporaneous reader. To probe this, we re-score the articles surrounding four major events—the Cuban Missile Crisis, the 1990 invasion of Kuwait, the September 11 attacks, and Russia's 2022 invasion of Ukraine—with a modified prompt that fixes each article's own publication date as the present ("Today is {date}"), supplies the date and headline, and explicitly instructs the model that it has no knowledge of events occurring after that date. The resulting series is nearly indistinguishable from the baseline. On average the hindsight-corrected scores run marginally higher for the Cuban Missile Crisis and September 11, and marginally lower for the Kuwait and Russia-Ukraine episodes, but the differences are very small throughout. The slight elevation for the two crises that resolved without the feared catastrophe is the direction one would expect if the baseline carried mild hindsight dampening, with the model scoring those episodes a touch less alarmingly than a contemporaneous reader would; that the other two events tilt the other way confirms the effect is neither systematic nor material. Appendix A.4 plots the comparison and reproduces the full prompt. We find the same on a representative sample: re-scoring the randomly audited articles with the date-anchored prompt changes the scores hardly at all (correlation 0.95, mean absolute difference 0.02), and the small average shift is again slightly upward, consistent with mild hindsight dampening in the baseline rather than inflation. We read this as evidence that any hindsight effect is too small to matter: stripping the model's knowledge of how events unfolded leaves the index's level and profile essentially unchanged. One caveat bears noting: instructing a model that it lacks knowledge of the future does not erase that knowledge from its parameters, so the test establishes that invoking or denying such knowledge barely moves the index, not that the knowledge is absent altogether.

4.4 Robustness to AI Model Choice

To assess whether our choice of LLM affects the resulting index, we classify a random sample of 1,050 articles using four OpenAI models—GPT-4o mini (our baseline), GPT-4o, GPT-5-mini, and GPT-5—and an open-weight model, Llama 3.1 8B Instruct. All models receive the same prompt. We set temperature to zero for the models that support it; GPT-5-mini and GPT-5 do not expose

an adjustable temperature, so they run at their default setting with reasoning effort set to high.[15] GPT-4o mini's scores are highly correlated with those of every alternative, and the share of articles classified as containing geopolitical risk is similar across all of them.[16] This agreement holds across models spanning two generations, different model sizes, and both proprietary and open-weight model families, suggesting that the AI-GPR index is not sensitive to the specific model used. Given that GPT-4o mini is substantially cheaper and produces highly correlated scores, we adopt it as our baseline for the full corpus. The agreement with an open-weight model also speaks to reproducibility: proprietary models are periodically deprecated, a concern Carlson and Dell (2025) raise for AI-derived measures generally, and the Llama results indicate that the index can be reconstructed with a model whose weights are permanently and publicly available.

5 The Anatomy of Geopolitical Risk

We further leverage the semantic understanding of LLMs to measure three additional dimensions of geopolitical risk: threats and acts, oil, and country-level and bilateral risk. The threats-and-acts measure identifies whether geopolitical risk is anticipated or realized. The oil measure captures the intensity of oil-supply disruptions driven by geopolitical events. The country and bilateral measures track which countries are involved in geopolitical events and the directed relationships between them. Each is produced by applying an additional classification layer to the underlying articles.[17]

5.1 Threats and Acts

A key innovation of Caldara and Iacoviello (2022) was to separate geopolitical risk into *threats*—adverse events that are anticipated, feared, or warned of but have not yet materialized—and *acts*—events that have actually occurred or are underway. The distinction is economically meaningful, because anticipated and realized risk need not move the economy in the same way, as our analysis of stock returns in Section 6.1 makes clear. We build an analogous decomposition of the AI-GPR index with a second LLM classification layer that reads each article and judges the temporal nature of the risk it describes.

[15] For Llama, we run meta-llama/Llama-3.1-8B-Instruct locally with 8-bit quantization, applying the model's standard instruct chat template, and decode greedily for near-deterministic output.

[16] The pairwise correlations between GPT-4o mini and the alternatives are 0.90 (GPT-4o), 0.87 (GPT-5-mini), 0.86 (GPT-5), and 0.91 (Llama); within the GPT-5 family, GPT-5-mini and GPT-5 correlate at 0.94. The positive-classification share ranges from 26 to 33 percent across models.

[17] Every subindex in this section is normalized by the same denominator as the headline index in equation (1): the relevant geopolitical risk scores are summed and divided by A_t, the total count of newspaper articles published on date t across all three papers (not just those in the relevant category). Using this common denominator throughout—rather than, for instance, the number of articles within each category—keeps all of the indexes on the same scale as the aggregate AI-GPR index and ensures their levels remain directly comparable.

We apply the classifier to every article that receives a positive AI-GPR score, passing the same text supplied to the baseline classifier to GPT-4o mini at temperature zero.[18] The model places each article in one of three categories: a *threat*, when it centers on risks, warnings, tensions, or potential escalations that have not yet come to pass; an *act*, when it reports an adverse event that has occurred or is occurring, such as an attack, invasion, or coup; or *both*, when it substantively discusses a realized event together with the risks of further escalation that surround it. Because articles in the *both* category speak to each dimension, we count them toward both subindexes, so that the threat series aggregates the articles labeled *threat* or *both* and the act series aggregates those labeled *act* or *both*.[19] Each subindex is then built exactly like the headline AI-GPR index: we sum the geopolitical risk scores of its articles, divide by total newspaper volume, and rescale the series to average 100 over 1985–2019.

Figure 4 plots the two subindexes at monthly frequency, and they trace distinct geopolitical rhythms. Acts spike sharply around realized conflicts—the 1991 Gulf War, the September 11 attacks, the 2003 Iraq War, and Russia's 2022 invasion of Ukraine—while threats run higher during the early Cold War and other spells of elevated but unrealized tension, when coverage is dominated by warnings and brinkmanship rather than open conflict. As Section 6.1 shows, this distinction is more than descriptive: threats and acts are priced quite differently in equity markets.

5.2 Oil Markets

We construct a second subindex about oil supply disruptions driven by geopolitical events using a two-step classification process similar to Section 5.1. In the first step, we restrict attention to articles receiving a GPR score above 0.5 and, to economize on classification costs, retain those mentioning at least one oil- or energy-related keyword; articles without such keywords are recorded as carrying no oil-supply risk. In the second step, we apply a second LLM prompt to the retained articles.[20] This prompt asks GPT-4o mini to determine whether the article discusses an oil or energy supply disruption caused by a geopolitical event and, if so, which geographic region(s) are affected from a predefined list: Middle East, Russia, USA, Venezuela, North Africa, West Africa, Central Asia, North Sea, Canada, Mexico, Latin America, Southeast Asia, and China.[21]

For each date t, we construct the oil disruption ratio:

$$\text{Oil GPR}_t = \frac{\sum_{i \in D_t} S_{it}}{A_t} \tag{2}$$

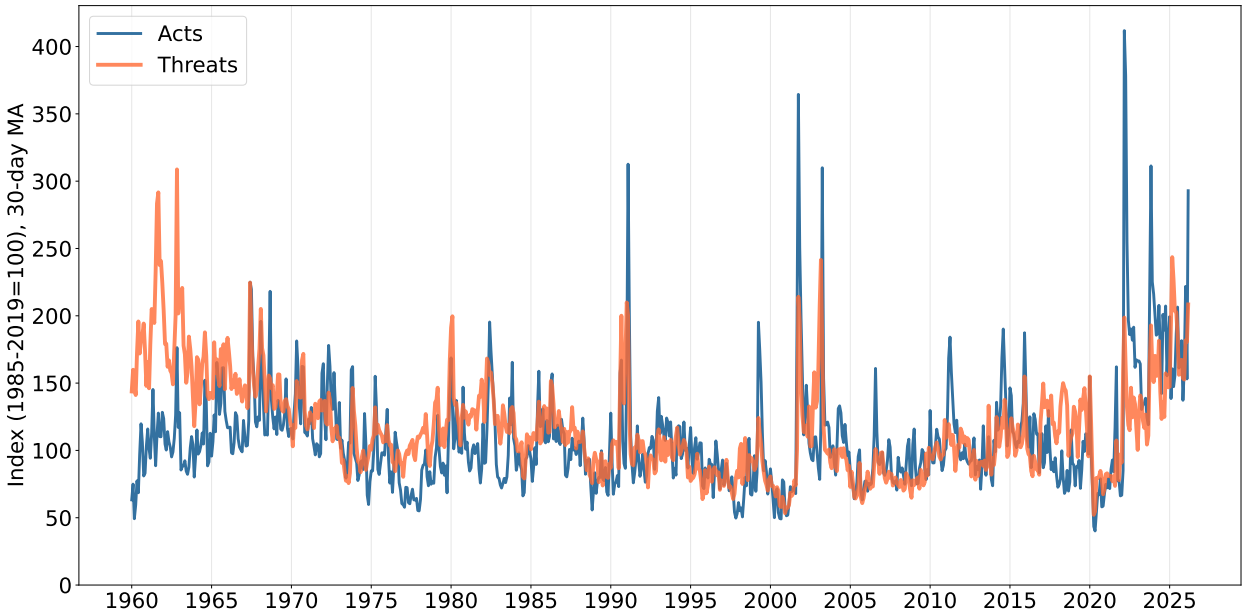
[18] As in the baseline scoring step, the model reads the first 2,000 characters of each article. The full prompt is reproduced in Appendix A.6.

[19] About 15 percent of classified articles fall in the *both* category, with roughly 48 percent labeled threats and 37 percent acts.

[20] To improve classification accuracy, the second-stage classifiers pass the first 3,000 characters of each article to GPT-4o mini, compared to 2,000 characters for the base GPR scoring step.

[21] The full classification prompt is located in Appendix A.6.

Figure 4: Geopolitical Threats and Acts (Monthly)

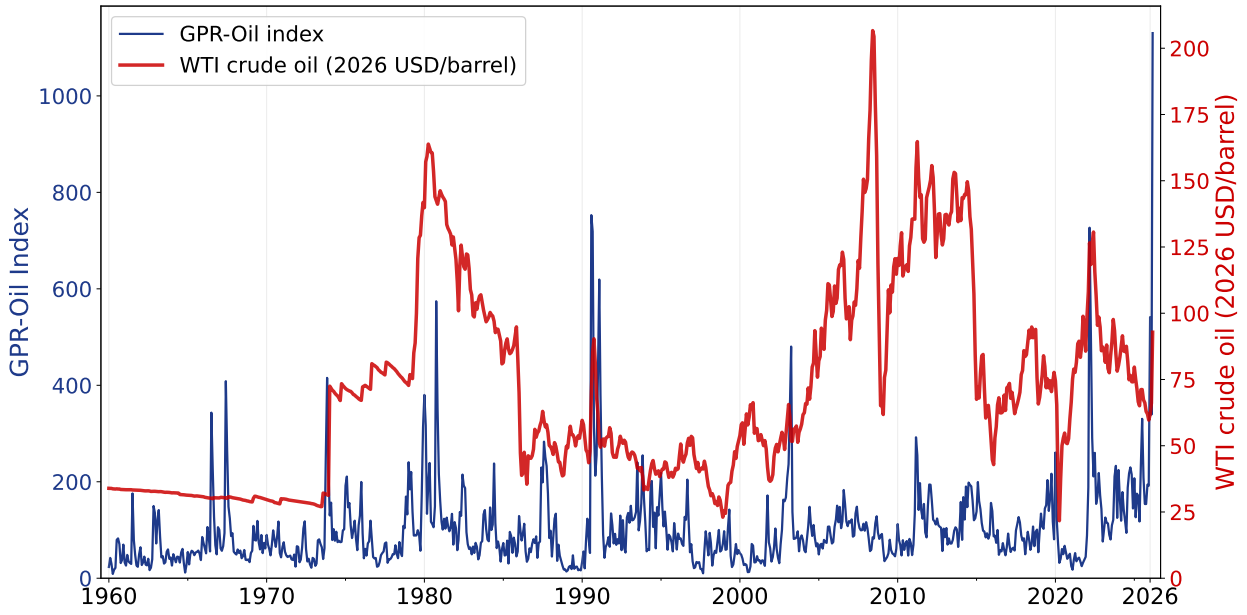


Notes: The threats subindex (orange, thick) and acts subindex (blue, thin) of the AI-GPR index, plotted at monthly frequency. A second LLM classification layer assigns each geopolitically relevant article (AI-GPR score > 0) to "threat," "act," or "both"; articles labeled "both" contribute to each series. Each series is the 30-day moving sum of the AI-GPR scores of its articles divided by the 30-day moving sum of total newspaper articles, indexed so that each series averages 100 over 1985–2019, then averaged to monthly frequency. Source: authors' calculations using ProQuest TDM Studio data. Last observation: March 2026.

where D_t is the set of articles on date t classified as discussing a geopolitical oil disruption, S_{it} is the article's GPR score (from the first classification layer), and A_t is the total number of newspaper articles published on date t. The numerator weights each disruption article by its GPR intensity, so that an article extensively covering an oil crisis (score near 1.0) contributes more than one with a passing mention (score near 0.5). The denominator normalizes by total publishing volume, analogously to the construction of the AI-GPR index. Regional oil disruption indices are constructed similarly, restricting D_t to articles classified as affecting a particular region.

Figure 5 plots the Oil GPR index (30-day moving average, monthly) alongside real WTI crude oil prices over 1960–2025. The index closely tracks the major geopolitical oil supply events: the 1973 Arab oil embargo, the 1979 Iranian Revolution, the 1990–91 Gulf War, and the Russia-Ukraine conflict and associated sanctions from 2022 onward. The co-movement between the Oil GPR index and oil prices confirms that the LLM-based classification is capturing the geopolitical episodes most associated with supply disruptions. A regional decomposition of the Oil GPR index by geography—showing the shifting dominance of the Middle East through the 1990s, Russia after 2014, and Venezuela during its political crises—is provided in Figure A.3 of the appendix.

Figure 5: GPR Oil Index and Real Oil Prices



Notes: The dark blue line (left axis) shows the Oil GPR index defined in equation (2), plotted as a 30-day moving average. The red dashed line (right axis) shows the monthly real WTI crude oil price (nominal WTI deflated by the U.S. CPI, expressed in 2019 dollars). Sources: authors' calculations; WTI prices from FRED (series WTISPLC/DCOILWTICO); CPI from FRED (series CPIAUCSL). Last observation is March 31, 2026.

Of all articles with a GPR score above 0.5, about 13 percent contain oil-related keywords and are sent to the second classification layer. Of these, roughly three-quarters are classified as discussing a geopolitical oil supply disruption. The Middle East accounts for nearly two-thirds of all disruption mentions, consistent with its dominant role in geopolitical oil risk throughout the sample period. Russia is the second most frequently mentioned region at 14 percent, with its share rising sharply after 2014.

5.3 Country Risk and Bilateral Risk

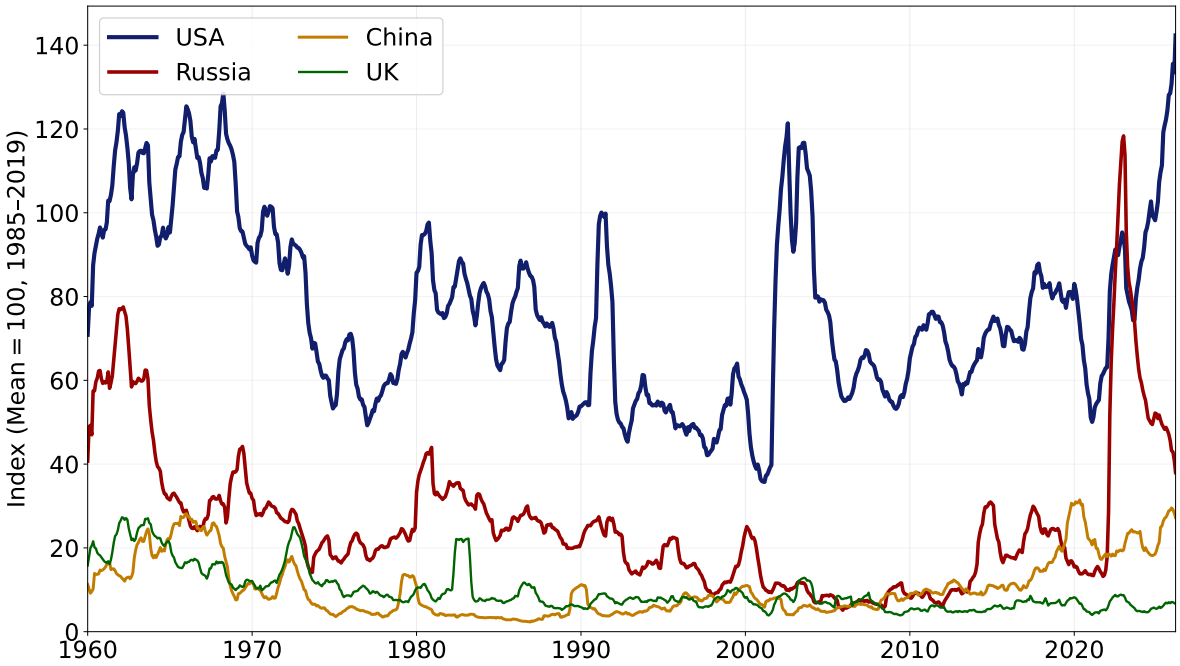
Our final set of measures focuses on country-level and bilateral geopolitical risk.

The country subindex measures the intensity of geopolitical events associated with a particular country. To construct the subindex, we parse each article through a classification layer to identify which countries are involved in the geopolitical event and their corresponding actor roles. We aggregate the AI-GPR scores for each country across all roles and scale to the overall index.[22] Figure 6 presents monthly series for the United States, Russia, China, and the United Kingdom—the four countries most often featured in newspaper coverage of geopolitical risks. The companion

[22] The full country classifier prompt is in Appendix A.6.

webpage includes time series of geopolitical risk for 200 countries.

Figure 6: AI-GPR by Country



Notes: Monthly AI-GPR contribution for the four most prominent countries across all actor roles (initiator, respondent, spillover), smoothed with a 12-month moving average. Series are scaled to the overall AI-GPR index (mean = 100, 1985–2019). Source: authors' calculations using the second-stage LLM country classifier applied to articles with positive GPR score. Last observation: March 2026.

The United States is the most persistently prominent country, reflecting its role as the global hegemon engaged in or affected by conflicts worldwide throughout the sample.[23] Russia registers two distinct elevated periods: the Cold War era through the late 1980s, and the post-2014 period following the annexation of Crimea and the full-scale invasion of Ukraine in 2022. China's GPR contribution rises steadily from the late 1990s onward, consistent with its growing global presence and the intensification of geopolitical competition in the Indo-Pacific. Country-level decompositions of this kind extend the country sub-indices of Caldara and Iacoviello (2022), who constructed keyword-based series for selected nations. The LLM approach enables a more precise construction of sub-indices across all countries mentioned in the news without requiring pre-specified keyword lists.

We build on the country subindex to create the bilateral index. This index tracks the level of

[23] Of course, the large coverage of U.S.-related events is also a byproduct of our choice to focus on U.S. newspapers. An active and promising line of research has begun constructing text-based measures of geopolitical risk using a broad array of non-U.S. news sources. See for instance Ambrocio et al. (2025), Andres-Escayola et al. (2022), and Bondarenko et al. (2024) for some recent examples.

geopolitical risk between initiator-respondent country pairs. We construct bilateral indices for the top 1,200 country pairs as the sum of AI-GPR sentiment scores in articles where the first country is classified as initiator and the second as respondent, so that the USA→Russia and Russia→USA series are distinct.[24]

This subindex represents an innovation that would not have been possible using past keyword approaches, which could identify that Russia and the United States both appeared in an article but could not distinguish whether Russia was acting on the United States, whether the United States was responding to Russian pressure, or whether both countries merely appear in adjacent paragraphs about unrelated topics. The LLM classifier extracts the directed relational structure of each event, identifying which country initiates the geopolitical action and which is the target.

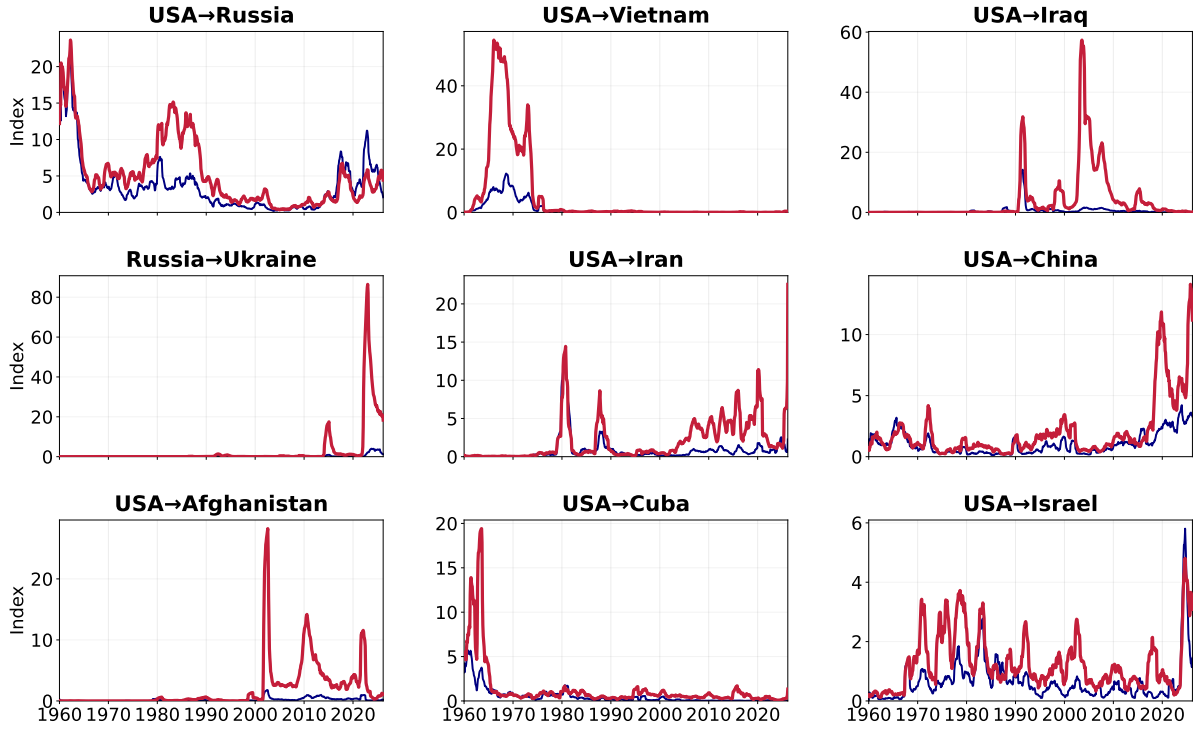
Figure 7 presents the nine country pairs with the highest average bilateral intensity, each shown in both directions: the thick red line is the more intense direction (named in the panel title) and the thin blue line is its reverse. The United States is the dominant initiator in eight of the nine pairs, reflecting both its central role in postwar geopolitics and the US-centric newspaper sources underlying the index. The pairs span the Cold War confrontation with the Soviet Union (USA→Russia, USA→Cuba), the major US military interventions of the postwar era (USA→Vietnam, USA→Iraq, USA→Afghanistan), ongoing great-power and regional tensions (USA→China, USA→Iran, USA→Israel), and Russia's 2022 invasion of Ukraine (Russia→Ukraine). The directional split is itself informative: when a large power confronts a smaller one—USA→Iraq, Russia→Ukraine, USA→Afghanistan—coverage casts one country overwhelmingly as the initiator and the reverse series barely registers; between comparably sized powers, such as the United States and Russia or the United States and China, initiation is attributed far more evenly across the two directions.

The same classification layer also labels each article by the type of geopolitical event it describes—among them military conflict, diplomatic tension, terrorism, civil war, nuclear threat, coup or regime change, and sanctions.[25] Although we do not pursue these categories as a separate application, they line up with familiar historical patterns: military conflict tracks the major wars of the postwar era, peaking around the Korean and Vietnam Wars, the Gulf War, and Russia's 2022 invasion of Ukraine; terrorism rises sharply after the September 11 attacks; and sanctions grow steadily more prominent after 2014 as economic coercion becomes a favored geopolitical instrument.

[24] Each directed-pair index is scaled analogously to the country-level indices: the raw sentiment sum for the pair is multiplied by the ratio of the overall AI-GPR index to the total coded sentiment in that month. This scaling is algebraically equivalent to dividing the pair's sentiment sum by the same total article count A_t used for the benchmark index (and applying the benchmark's rescaling constant), so the bilateral series share the aggregate index's denominator and scale rather than being renormalized by the smaller set of articles involving each pair. Because an article in which one country acts against multiple respondents contributes its full sentiment score to each initiator–respondent pair (and vice versa), the bilateral indices do not in general sum to the aggregate index.

[25] The event-type classification prompt is listed in Appendix A.6.

Figure 7: AI-GPR Bilateral Indexes: Top Country Pairs, Both Directions



Notes: Each panel shows both directed bilateral AI-GPR sub-indices for a country pair, smoothed with a 12-month moving average. The thick red line is the more intense direction—initiator → respondent, as named in the panel title—and the thin blue line is the reverse direction. Each directed sub-index aggregates AI-GPR sentiment scores from articles in which the first country is classified as initiator and the second as respondent, scaled to the overall AI-GPR index. The nine pairs have the highest average intensity in their dominant direction over the sample. Last observation: March 2026. Source: authors' calculations using the second-stage LLM country classifier applied to articles with positive GPR score.

6 Applications: Geopolitical Risk in Financial Markets, Oil Markets, and Trade

We present a series of applications relating the three dimensions of geopolitical risk we have studied thus far to financial markets, oil, and trade. First, we examine how geopolitical threats and acts are priced in the stock market using the threats and acts subindexes. We then study the macroeconomic response to oil shocks identified with a novel instrument constructed from the Oil GPR index. Lastly, we examine how bilateral geopolitical risk relates to international trade using our bilateral GPR index.

6.1 Financial Markets and Geopolitical Threats and Acts

A large literature studies whether geopolitical events depress stock markets. We revisit this question using the AI-GPR index and daily data on U.S. stock returns from 1960 to March 2026.

We merge the AI-GPR index with daily excess market returns.[26] Since stock markets are closed on weekends and holidays while the AI-GPR is computed for every calendar day, we construct a trading-day GPR measure as follows. If the market is closed for k consecutive calendar days before trading day t, we assign to day t the average of the AI-GPR over those $k+1$ days (including day t itself), so that geopolitical information accumulating over non-trading days is correctly attributed to the next trading day. We then work with the first difference of this series, standardized by its full-sample standard deviation. All GPR variables in the regressions below are standardized in this way. We estimate the regression:

$$r_t^e = \alpha + \beta \, \Delta\text{GPR}_t + \varepsilon_t \tag{3}$$

where r_t^e is the excess market return (in percent). Table 2 reports the results estimated on data collapsed to weekly frequency, where stock market returns are cumulated weekly and the GPR change is computed week over week. The relationship is negative and economically meaningful: a one-standard-deviation increase in the AI-GPR is associated with a decline of about fourteen basis points in the weekly market return.

The OLS estimate in equation (3) conflates predictable and unpredictable movements in GPR. To disentangle these, we decompose geopolitical risk into anticipated and unanticipated components. We estimate an AR(4) model for the standardized GPR change:

$$\Delta\text{GPR}_t = \phi_0 + \sum_{j=1}^{4} \phi_j \, \Delta\text{GPR}_{t-j} + u_t \tag{4}$$

and decompose ΔGPR_t into the persistent (predictable) component $\widehat{\Delta\text{GPR}}_t$ and the shock (unpredictable) component $\hat{u}_t$. We then estimate:[27]

$$r_t^e = \alpha + \beta_1 \, \widehat{\Delta\text{GPR}}_t + \beta_2 \, \hat{u}_t + \varepsilon_t \tag{5}$$

We estimate both regressions—the level specification in equation (3) and the anticipated vs.

[26] Through November 28, 2025, we use data from the Fama-French daily factors dataset. See https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. Beginning December 1, 2025, when Fama-French daily factors are not yet available, we extend the return series using the daily excess return on the S&P 500 index from Haver Analytics, computed as the daily simple return on the index minus the daily-compounding equivalent of the three-month Treasury bill rate.

[27] This approach follows the empirical tradition of Barro (1979), who decomposes macroeconomic variables into anticipated and unanticipated components to test whether agents respond differently to predictable trends versus stochastic innovations.

Table 2: Weekly Stock Returns and Geopolitical Risk

	Aggregate GPR		Threats and Acts	
	(1) OLS	(2) Decomposition	(3) OLS	(4) Decomposition
ΔGPR_t	-0.136^{**}			
	(0.057)			
$\Delta\text{GPR}_{t,\,threats}$			-0.081^{**}	
			(0.037)	
$\Delta\text{GPR}_{t,\,acts}$			-0.084	
			(0.061)	
Persistent $\widehat{\Delta\text{GPR}}_t$		-0.271^{**}		
		(0.134)		
Persistent $\widehat{\Delta\text{GPR}}_{t,\,threats}$				-0.220^{**}
				(0.103)
Persistent $\widehat{\Delta\text{GPR}}_{t,\,acts}$				-0.063
				(0.131)
Shock $\hat{u}_t$		-0.119^{*}		
		(0.062)		
Shock $\hat{u}_{t,\,threats}$				-0.052
				(0.044)
Shock $\hat{u}_{t,\,acts}$				-0.090^{**}
				(0.042)
Constant	0.137^{***}	0.138^{***}	0.136^{***}	0.138^{***}
	(0.037)	(0.037)	(0.037)	(0.037)
Observations	3,452	3,448	3,452	3,448
R^2	0.004	0.004	0.004	0.004
Sample	1960–2026	1960–2026	1960–2026	1960–2026

Notes: Weekly regressions of excess market returns (r_t^e, in percent) on standardized changes in the AI-GPR index. Columns (1)–(2) use the AI-GPR index; columns (3)–(4) use its threats and acts subindexes. Weekly stock returns are the sum of daily excess returns within the week, and the weekly GPR variable is the change in the within-week average of the trading-day GPR measure, standardized over the full sample. Columns (1) and (3) report OLS regressions of r_t^e on the GPR change(s). Columns (2) and (4) decompose each GPR change into a persistent component (fitted values from an AR(4) model) and a shock component (residuals). Excess market returns are from the Fama-French daily factors dataset through November 28, 2025, extended with Haver SP500 excess returns (SP500 simple return minus the daily-compounded 3-month T-bill rate) from December 1, 2025 onward. OLS columns report heteroskedasticity-robust standard errors in parentheses; decomposition columns report pairs-bootstrap standard errors (1,000 replications). The estimation sample runs through March 2026. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

unanticipated decomposition in equation (5)—twice: first using the aggregate AI-GPR index, and then replacing it with the threat and act subindexes entered jointly. The two pairs of regressions map onto the four columns of Table 2—the aggregate index in columns (1) and (2), and its threat-and-act refinement in columns (3) and (4).

For the aggregate index, the predictable, persistent component of geopolitical risk depresses returns far more than the unexpected component—its effect is more than twice as large—and both are statistically significant. This finding has a clear economic interpretation. While a sudden geopolitical event like the September 11 attacks generates an immediate negative return, the "long tail"

of that risk, the fact that geopolitical risk remains high in the weeks that follow, is what sustains an elevated equity risk premium. Investors do not merely react to headlines; they aggressively re-price assets as geopolitical uncertainty becomes embedded in the macroeconomic outlook.

The gap between the muted aggregate OLS estimate and the much larger effect of the persistent component suggests that a simple linear regression understates the role of geopolitical risk, attenuated by the high-frequency noise of weekly shocks, which are large in magnitude but transient in their effect on asset prices.[28] After decomposing the risk, the market is roughly twice as sensitive to the persistent component of geopolitical risk than the raw OLS estimate suggests.

Separating the index into its threat and act components sharpens this picture. In levels, it is geopolitical *threats* whose effect on returns is precisely estimated, while realized *acts*—though comparable in magnitude—are too noisy to register as statistically significant. The decomposition into anticipated and unanticipated movements reveals why the two are priced so differently. Threats bite through their predictable part: as a threat persists and works its way into the outlook, markets re-price gradually, so it is the slow-moving, anticipated component of threat risk that weighs on returns. Acts work the other way around. Because events that can be foreseen are already embedded in prices, only the *unanticipated* occurrence of an act moves markets—and it is indeed the surprise in acts, not their predictable part, that carries the effect. This same asymmetry explains why acts seem to do nothing in the simple regression: pooling a meaningful surprise effect together with a negligible anticipated one averages the act response toward zero, the very dilution that mutes the aggregate estimate. The threat-and-act results thus reinforce the central lesson of the aggregate decomposition: what matters for asset prices is not the level of geopolitical risk but whether it is anticipated or unexpected—a distinction that a single regression cannot capture.

The improved day-to-day precision of the AI-GPR index is central to these results. When we repeat the same analysis using the keyword-based GPR index, the results are qualitatively similar but economically and statistically weaker: the contemporaneous relationship with returns is roughly a third smaller and only marginally significant, and in the decomposition the persistent component barely clears conventional significance while the shock component loses it altogether. These differences illustrate how measurement imprecision in the keyword-based index may attenuate the estimated effects. Keyword-based methods misclassify articles at higher rates, so the resulting index contains more variation that is not geopolitically relevant, biasing coefficients toward zero in a standard errors-in-variables fashion. The AI-GPR's semantic scoring reduces this day-to-day noise, yielding a sharper signal that strengthens the estimated relationship between geopolitical risk and stock returns.

[28] This result echoes the findings in Gonçalves et al. (2025), who examine the pricing of geopolitical risk using stock market returns. They use Caldara and Iacoviello's decomposition of geopolitical risk into forward-looking threats (GPT) and realized acts (GPA), and find that only the threat component is consistently priced across assets and forecasts equity premia, whereas realized acts have a much weaker and less stable link to risk premia.

6.2 Oil Market Shocks and their Macroeconomic Effects

The Oil GPR index provides a natural instrument for isolating oil price movements driven by geopolitical supply disruptions, as distinct from demand-driven fluctuations or weather-related supply shocks. We borrow this idea from Verduzco-Bustos and Zanetti (2026), who use innovations in the Caldara and Iacoviello (2022) GPR index to identify geopolitically driven oil shocks; we replace their measure with the Oil GPR index, which leads to better inference by targeting exactly the subset of geopolitical news related to energy supply.[29] We estimate a four-variable proxy structural VAR (Mertens and Ravn, 2013; Stock and Watson, 2012) over the period 1975–2024 at monthly frequency. The system includes the nominal WTI oil price, global oil production, U.S. monthly GDP (the Brave-Butters-Kelley index, Brave et al. 2019), and the Oil GPR index. The VAR is estimated with 12 lags, and the external instrument—monthly WTI oil price surprises on Oil GPR event days—identifies the structural geopolitical oil price shock.[30]

Figure 8 presents the impulse responses to a geopolitical oil supply shock, normalized to generate a 33 percent increase in oil prices on impact. The Oil GPR index itself spikes on impact and then gradually dissipates as tensions ease. Oil production declines, consistent with a supply disruption. U.S. GDP falls by roughly 1 percent over one to two years, reflecting the contractionary effect of higher energy costs on economic activity as well as the accompanying disruptions caused by any conflict or tensions that directly affects the oil market. These results indicate that the Oil GPR index isolates economically meaningful supply-side disturbances, and that geopolitically driven oil shocks trace out sizable macroeconomic responses.

6.3 Bilateral Geopolitical Shocks and International Trade

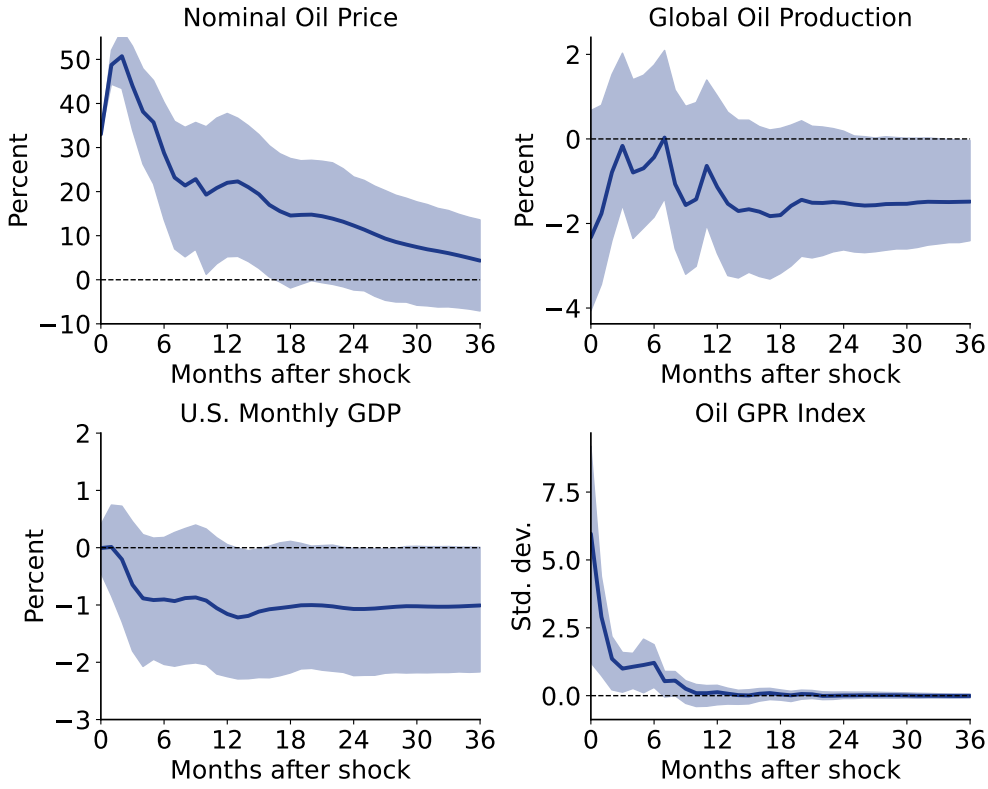
The bilateral index lets us examine whether higher bilateral geopolitical risk predicts lower bilateral trade. We merge our bilateral GPR series for up to 1,200 country pairs with the CEPII Gravity database (Conte et al., 2023), which provides annual bilateral trade flows from BACI—covering up to 200 countries—together with standard gravity controls. Our dependent variable is total two-way trade for each country pair, obtained by summing the two directed BACI flows (exports in each direction); because BACI trade data begin in 1996, the estimation sample spans 1996–2020. We include all pairs for which we construct bilateral GPR (up to 1,200) and for which matching trade data are available in BACI.[31] Both bilateral trade and bilateral GPR are standardized within country pair to remove pair-specific means and scales, and GPR enters with a one-year lag. Table 3 reports the estimates. With year fixed effects alone (column 1) and with year fixed effects plus log

[29] As an additional step, we construct the monthly instrument by summing daily WTI log-price changes on days when the Oil GPR index spikes above two standard deviations, shifted back one day to account for the newspaper reporting lag. This ensures the identified shock reflects oil price movements triggered by geopolitical news rather than pre-existing price trends.

[30] The instrument is strong: the first-stage F-statistic is 35, well above the conventional weak-instrument thresholds.

[31] Trade flows are deflated to constant 2012 dollars using the US GDP deflator.

Figure 8: Impulse Responses to a Geopolitical Oil Price Shock



Notes: Impulse responses to a geopolitical oil supply shock identified via a proxy structural VAR. The VAR includes four monthly variables (1975–2024): nominal WTI oil price, global oil production, U.S. monthly GDP (Brave-Butters-Kelley index), and the Oil GPR index, estimated with 12 lags. The external instrument is monthly WTI oil price surprises accumulated on days when the Oil GPR index spikes above two standard deviations. The shock is normalized to generate a 33 percent increase in oil prices on impact, which builds to roughly 50 percent at the peak. Shaded areas represent 80 percent confidence intervals from a wild bootstrap with 500 draws.

PPP-adjusted GDP of both countries (column 2), a one-standard-deviation increase in bilateral GPR is associated with roughly a 0.03–0.04 standard-deviation decline in bilateral trade in the following year. The association weakens under the more demanding origin-by-year and destination-by-year fixed effects (column 3), where the point estimate is about −0.02 and no longer statistically significant—indicating that much of the relationship operates through broader country-level trends absorbed by those fixed effects. Figure 9 displays the partial correlation after removing year effects and GDP controls.

Table 3: Bilateral Geopolitical Risk and Trade Flows

	(1) Year FE	(2) Year FE + GDP	(3) Country×Year FE
Bilateral GPR_{t-1} (std.)	-0.033^{***}	-0.043^{***}	-0.021
	(0.011)	(0.013)	(0.013)
ln GDP PPP, origin		0.016^{***}	
		(0.003)	
ln GDP PPP, destination		0.021^{***}	
		(0.003)	
Observations	17,135	12,905	11,547

Notes: Dependent variable is total two-way bilateral trade (BACI, sum of both directed export flows, deflated to 2012 USD), standardized within country pair. Bilateral GPR is standardized within pair and enters with a one-year lag. Column 1 includes year fixed effects. Column 2 adds log PPP-adjusted GDP (PWT) for origin and destination. Column 3 uses origin-by-year and destination-by-year fixed effects, which absorb GDP. Standard errors clustered by country pair. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Because BACI trade data begin in 1996, the sample covers 1996–2020. Source: bilateral GPR from this paper; trade and gravity controls from Conte et al. (2023).

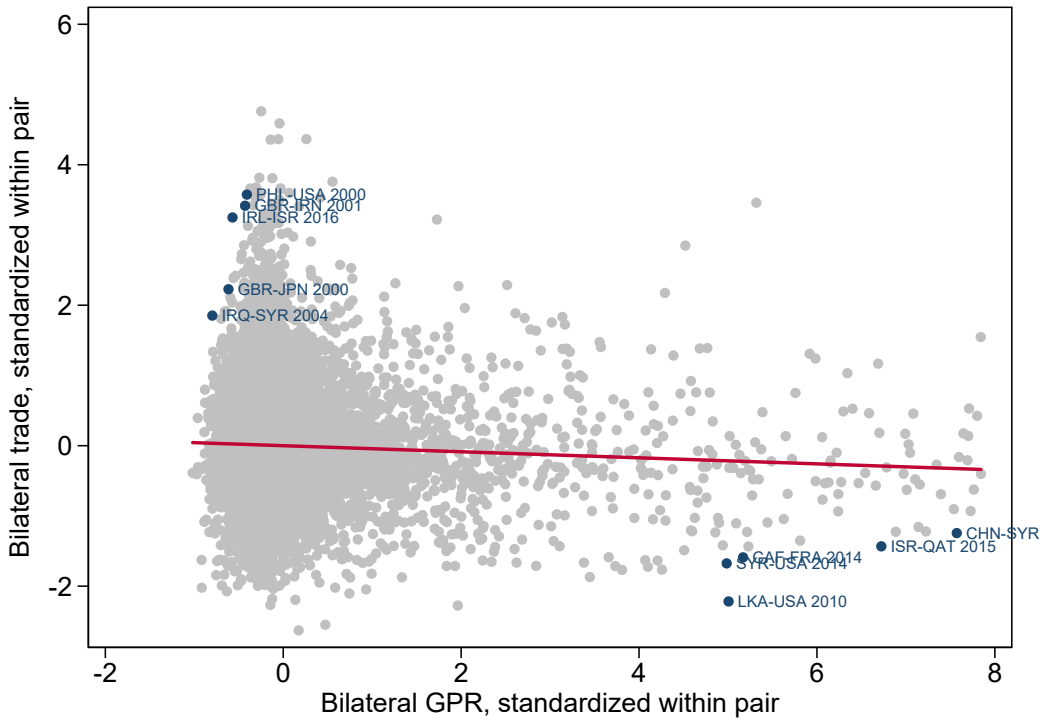
7 Conclusion

We have introduced the AI-GPR Index, an improved measure of geopolitical risk that uses large language models to evaluate newspaper content. Our approach maintains the conceptual framework established by Caldara and Iacoviello (2022) while improving measurement precision through semantic understanding rather than keyword matching.

The methodology offers several practical advantages. LLM-based classification reduces false positives from topically irrelevant keyword matches and false negatives from semantically relevant articles using alternative terminology. The approach provides graduated risk scores rather than binary classification, allowing the index to reflect intensity as well as presence of geopolitical content. Unlike rigid exclusion rules that disqualify entire articles based on specific words, our method exercises contextual judgment about whether geopolitical content is central and current. The method extends straightforwardly to non-English publications and other text sources.

We illustrated several applications of the index. First, the AI-GPR produces a more precisely estimated negative effect of geopolitical risk on stock returns and reveals that markets are especially sensitive to the persistent component of GPR innovations; decomposing the index into threats and acts shows that this persistent effect runs through anticipated threats, while the surprise component of realized acts also weighs on prices. Second, by combining the AI-GPR with a second classification layer, we constructed a novel historical time series of geopolitical oil supply disruptions by region,

Figure 9: Bilateral Geopolitical Risk and Trade Flows: Partial Correlation



Notes: Scatter plot of within-pair standardized total two-way bilateral trade against within-pair standardized bilateral GPR (one-year lag), both residualized on year fixed effects and log PPP-adjusted GDP of origin and destination. The 10 observations with the largest influence on the estimated slope are labeled with country-pair and year. The line shows the OLS fit. Source: bilateral GPR from this paper; trade from Conte et al. (2023). Sample: 1996–2020 (annual data).

revealing the shifting geography of energy-related geopolitical risk over six decades and the macroeconomic effects of geopolitically driven oil shocks. Third, by extracting the geopolitical actors in each event and their roles, we decomposed risk to the level of individual countries—extending the analogous sub-indices of Caldara and Iacoviello (2022) using richer LLM-based scores—and, in a genuinely new step, to directed country pairs, identifying which countries are acting on whom rather than merely co-occurring in coverage; we show that bilateral GPR predicts bilateral trade in a standard gravity framework, with higher bilateral GPR associated with lower bilateral trade.

Our main contribution is demonstrating that LLM-based approaches can be applied at scale to construct consistent historical time series spanning multiple decades. As computational costs continue to decline, LLM-based approaches become increasingly practical for large-scale text analysis in economics. The approach is complementary to existing keyword-based methods, offering researchers and policymakers an additional tool for tracking geopolitical risk dynamics across time

and space.

Future research could extend this methodology to construct firm-level or sector-level measures of geopolitical risk exposure, analyze how geopolitical risks transmit across countries and markets using the bilateral and country-level indices, or examine the relationship between geopolitical risks and various economic outcomes. The flexibility of prompt-based classification extends naturally to a wide range of economic and social phenomena beyond geopolitical risk, and we hope this framework offers a useful template for such work.

References

Ambrocio, G., Z. Fungáčová, J. Heikkinen, E. Kerola, I. Korhonen, and A. Norring (2025). Northern insights: Geopolitical risk from finnish news media. *Bank of Finland Research Discussion Paper* (13).

Andres-Escayola, E., C. Ghirelli, L. Molina, J. J. Pérez, and E. Vidal Muñoz (2022). Using newspapers for textual indicators: which and how many?

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics 131*(4), 1593–1636.

Barro, R. J. (1979). Unanticipated money growth and unemployment in the united states: Reply. *The American Economic Review 69*(5), 1004–1009.

Bondarenko, Y., V. Lewis, M. Rottner, and Y. Schüler (2024). Geopolitical risk perceptions. *Journal of International Economics 152*, 104005.

Brave, S. A., R. A. Butters, and D. Kelley (2019). A new "big data" index of U.S. economic activity. *Economic Perspectives 43*(1). Federal Reserve Bank of Chicago.

Caldara, D. and M. Iacoviello (2022). Measuring geopolitical risk. *American Economic Review 112*(4), 1194–1225.

Carlson, J. and M. Dell (2025). A unifying framework for robust and efficient inference with unstructured data. arXiv:2505.00282.

Clayton, C., A. Coppola, M. Maggiori, and J. Schreger (2025, June). Geoeconomic pressure. Working Paper.

Clayton, C., M. Maggiori, and J. Schreger (2026). A framework for geoeconomics. *Econometrica 94*(1), 105–136.

Conte, M., P. Cotterlaz, and T. Mayer (2023). The CEPII gravity database. Working paper, CEPII.

Gonçalves, A. S., A. Melone, and A. Ricciardi (2025). The pricing of geopolitical tensions over a century. *Fisher College of Business Working Paper*. WP 2025-03-010.

Hartley, E. (2025). Narratives to numbers: Large language models and economic policy uncertainty. arXiv:2511.17866.

Hassan, T. A., S. Hollander, L. van Lent, and A. Tahoun (2019). Firm-level political risk: Measurement and effects. *Quarterly Journal of Economics 134*(4), 2135–2202.

Ito, A., M. Sato, and R. Ota (2025). A novel content-based approach to measuring monetary policy uncertainty using fine-tuned llms. *Finance Research Letters 75*, 106832.

Mertens, K. and M. O. Ravn (2013). The dynamic effects of personal and corporate income tax changes in the United States. *American Economic Review 103*(4), 1212–1247.

Mohr, C. and C. Trebesch (2025). Geoeconomics. *Annual Review of Economics 17*.

Stock, J. H. and M. W. Watson (2012). Disentangling the channels of the 2007–2009 recession. *Brookings Papers on Economic Activity 2012*(1), 81–135.

Verduzco-Bustos, G. and F. Zanetti (2026). The effects of geopolitical oil price shocks. Working paper.

A Appendix

A.1 Alternative Versions of AI Model to Classify Articles

We classify a random sample of 1,050 articles using four OpenAI models—GPT-4o mini, GPT-4o, GPT-5-mini, and GPT-5—together with the open-weight Llama 3.1 8B Instruct, all with the same prompt. We set temperature to zero for the models that support it; the GPT-5 models do not expose an adjustable temperature and run at their default settings. Table A.1 reports the full pairwise correlation matrix.

Table A.1: Cross-Model Score Correlations

	GPT-4o mini	GPT-4o	GPT-5-mini	GPT-5	Llama
GPT-4o mini	1.00				
GPT-4o	0.90	1.00			
GPT-5-mini	0.87	0.86	1.00		
GPT-5	0.86	0.87	0.94	1.00	
Llama 3.1 8B	0.91	0.86	0.85	0.85	1.00

Notes: Pairwise Pearson correlations of geopolitical risk scores assigned by each model to a random sample of 1,050 newspaper articles. All models use the same classification prompt, at temperature zero where supported; the GPT-5 models run at their default settings. "Llama" denotes the open-weight Llama 3.1 8B Instruct model.

All pairwise correlations exceed 0.85, with the highest agreement (0.94) between GPT-5-mini and GPT-5. The mean score ranges from 0.10 (GPT-4o) to 0.17 (GPT-5-mini), reflecting modest differences in calibration that wash out after normalization. These results confirm that the AI-GPR index is robust to model choice—across two generations of OpenAI models, across model sizes, and across proprietary and open-weight model families.

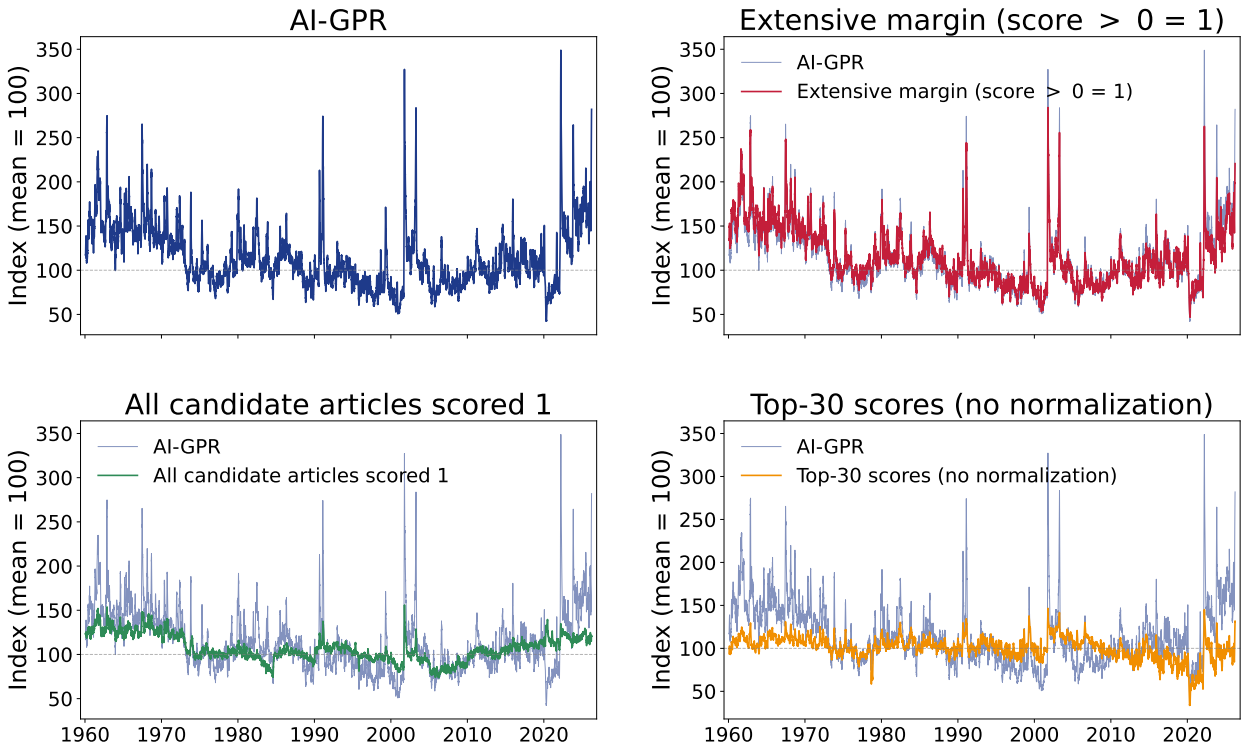
A.2 Alternative Ways of Constructing the Index

The baseline index in equation (1) makes two choices a reader might dispute: it weights each article by its graded risk score S_{it} (the intensive margin), and it normalizes the score sum by total newspaper output A_t rather than by the number of scored articles N_t. To show the role of these choices, we recompute the index under four alternative constructions, each reindexed to a mean of 100 over 1985–2019: (i) the baseline AI-GPR, $\sum_i S_{it}/A_t$; (ii) an *extensive-margin* version in which every article with $S_{it} > 0$ counts as one, $\#\{i : S_{it} > 0\}/A_t$, discarding intensity; (iii) an *all-candidate* version in which every scored article counts as one regardless of its score, equal to the scored share N_t/A_t; and (iv) an *unnormalized* version equal to the sum of the thirty highest scores each day, $\sum_{\text{top }30} S_{it}$, with no denominator.

Figure A.1 plots the four series and Table A.2 reports their pairwise correlations. The baseline and the extensive-margin version are very close (correlation 0.97): the intensity weighting is not what generates the index's movements, although it emphasizes some of the spikes. Replacing the graded score sum in the numerator with a simple count of scored articles—the all-candidate share N_t/A_t—lowers the correlation to 0.70, and removing the denominator A_t entirely (the unnormalized top-30 sum) gives 0.48 in levels; both nonetheless track the baseline closely in year-on-year changes (0.66 and 0.85, respectively). These construction choices therefore change the index's high-frequency movements little.

The concern that the index might merely reflect long-run drift in the scored share N_t/A_t—as newspaper composition and the share of topics shift over six decades—is addressed by the third construction. As Figure A.1 shows, the share of candidate articles is a smooth, slowly drifting series with none of the event spikes that define the geopolitical risk index, and among the four it least resembles the baseline's sharp, event-driven movements. In addition, the series shows a U-shape that mostly reflects the inverted U-shape in the pattern of total newspaper articles over the sample. The index's economically relevant variation—the sharp rises around the Gulf War, September 11, the Iraq War, and Russia's invasion of Ukraine—comes from the geopolitical content of the articles, not from how many articles clear the very generous keyword filter. We conclude that the AI-GPR index measures the intensity-weighted share of newspaper coverage devoted to geopolitical risk and that this object is robust to the construction choices examined here.

Figure A.1: Alternative Versions of the Index



Notes: Four alternative constructions of the AI-GPR index, each shown as a 30-day moving average and reindexed to a mean of 100 over 1985–2019. *AI-GPR* is the baseline $\sum_i S_{it}/A_t$; *Extensive margin* counts each article with $S_{it} > 0$ as one; *All candidate articles scored 1* counts every scored article as one, equal to N_t/A_t; *Top-30 scores* sums the thirty highest scores each day with no normalization. Source: authors' calculations. Last observation: March 2026.

A.3 Additional Details on the Human Audit

This appendix gives the full results of the human validation summarized in Section 4.2. We hand-scored a random sample of about 2,000 articles—drawn from the *New York Times*, the *Washington Post*, and the *Chicago Tribune* over January 1960 to December 2025—using the same rubric supplied to the model, which assigns each article a geopolitical risk score between 0 and 1. Each article was

Table A.2: Correlations Across Index Constructions

	(1)	(2)	(3)	(4)
(1) AI-GPR (baseline)	1.00			
(2) Extensive margin	0.97	1.00		
(3) All candidate (N_t/A_t)	0.70	0.75	1.00	
(4) Top-30 (unnormalized)	0.48	0.41	0.08	1.00

Notes: Pairwise Pearson correlations of the four index constructions defined in the text, computed on the daily series (reindexed to a mean of 100 over 1985–2019) over 1960–2026. A further variant not shown in the figure, normalizing by scored articles N_t instead of total articles A_t—the average score among scored articles, $\sum_i S_{it}/N_t$—yields a series correlated 0.89 with the baseline.

scored by one human coder and a large subset (about 1,700 articles) by a second; we take the consensus of the available coders as the human score and compare it with GPT-4o mini's score on the first 2,000 characters of each article, treating any entry outside the 0-to-1 scale as missing.

The human coders agree with one another imperfectly: on the double-coded articles their scores correlate at 0.78 and concur on whether an article carries any geopolitical risk in about 92 percent of cases, a reminder that even careful readers applying the same rubric reach different judgment calls, and a natural benchmark for any classifier. Against the human consensus, the machine scores correlate at 0.85—above the coders' agreement with each other—agree on the presence of any geopolitical risk in about 90 percent of cases, and have a mean absolute difference of about 0.05, roughly the spacing of a single step on the scale. Table A.3 reports the full cross-tabulation; the off-diagonal mass sits almost entirely on the first off-diagonal, so disagreements are overwhelmingly one-step judgment calls at a rubric boundary, such as whether an ongoing regional conflict constitutes "significant tensions" (0.4) or "major war escalation risks" (0.6). The model's and the humans' average scores are essentially identical (both 0.122), so the article-level disagreements are offsetting and introduce no aggregate bias in levels.

The model's disagreements with the humans are, moreover, concentrated in precisely the articles the humans themselves find ambiguous. On the double-coded articles where the two coders assign identical scores, the model matches their common score exactly in about 90 percent of cases, with a mean absolute deviation of 0.02; where the coders disagree with each other, the model's mean absolute deviation from the consensus rises to 0.14. Model error thus tracks genuine ambiguity in the articles rather than striking at random, the same pattern Hartley (2025) documents for policy-uncertainty classification, where model misclassification falls by half as human-coder certainty rises.

Scoring the same audited articles with both classifiers also allows a like-for-like comparison against the human labels, free of the differences in sample, rubric, and ground truth that complicate cross-study figures. Table A.4 reports this three-way comparison, pooling all individual article–coder ratings (about 3,700 pairs, with each double-coded article entering once per coder) and applying the original Caldara and Iacoviello (2022) keyword query—exclusion words included—to the same texts. Benchmarking each coder's rating separately, rather than the consensus mean, ensures that an article one coder scores as zero and another as positive does not count as positive by construction. Within each method's own set of flagged articles, the share a human coder scores as zero is broadly similar (a false-discovery rate of about 13 percent for the AI-GPR and 16 percent for the keyword query, the latter below the 21 percent Caldara and Iacoviello report from their own audit), confirming on a common footing the precision advantage discussed in Section 4.2. Measured against the consensus mean instead, the false-discovery rates are somewhat lower for both methods

Table A.3: Human vs. AI Score Cross-Tabulation

	AI score				
Human score	0.0	0.2	0.4	0.6	0.8
0.0	1366	60	23	3	0
0.2	43	34	47	7	0
0.4	39	60	140	100	0
0.6	3	0	21	73	5
0.8	0	1	3	15	4

Notes: Cross-tabulation of the human consensus score (rounded to the nearest 0.2) and the GPT-4o mini score for the roughly 2,000 audited articles. Rows are human scores, columns are model scores. The model never assigned a score above 0.8. Off-diagonal mass is concentrated on the first off-diagonal, that is, one-step disagreements.

(9.4 and 15.1 percent), since averaging turns some single-coder zeros into positive benchmarks. The keyword query flags far fewer articles overall; we caution against reading its high miss rate as an unconditional recall figure, however, because the audit deliberately oversamples articles the model scores as geopolitical and so is enriched relative to the full corpus.

Table A.4: Human, AI-GPR, and Keyword GPR on the Audited Panel

	Flagged as GPR (% of ratings)	False-discovery (% of flagged)	Miss rate (% of human GPR)
Human coders	30.5	—	—
AI-GPR	29.2	13.3	16.9
Keyword GPR	6.9	16.3	81.1

Notes: The sample pools all article–coder ratings in the audit: each article enters once for every human coder who scored it, about 3,700 ratings in total across the *New York Times*, *Washington Post*, and *Chicago Tribune*. An article is flagged by a method when its score is positive; the keyword row applies the original Caldara and Iacoviello (2022) query, including its exclusion words, to the same texts. The false-discovery rate is the share of a method's flagged articles that the human scores as zero. The miss rate is the share of articles the human scores as positive that the method fails to flag. The audit oversamples articles the model scores as geopolitical (a 30/70 design built for human review), so the miss rate is specific to this panel and is not a recall rate for the full corpus; the false-discovery rate, computed within each method's own flagged set, does not depend on the sampling design.

Reading the disagreeing articles and recording notes on each case, we group the discrepancies into five recurring patterns.

First, the model treats occasional economic and financial tensions between countries as geopolitical risks. Articles on a proposed U.S. tariff on Japanese goods (1971), the financial crisis in Latvia (2009), Iceland's refusal to repay British and Dutch lenders (2010), and strained Taiwan–China business ties (2000) were scored 0.4 to 0.6 by the model and 0 to 0.2 by us. The rubric defines geopolitical risk around war, terrorism, and military tensions. Trade and financial disputes between friendly nations strain "ties" in a literal sense but rarely carry risk of armed conflict, and a human coder discounts them accordingly. Many of the cases in which the model scores higher than

the human are of this type, consistent with the model anchoring on surface vocabulary ("threat," "retaliation," "frayed ties") rather than the military nature of the underlying risk.

Second, the model treats domestic unrest and domestic security as geopolitical risk, even if the original definition emphasizes the interstate nature of such risks. Internal events with security overtones—the 1991 Crown Heights riots, labor strife framed around a general strike, an article on the disciplining of unruly airline passengers before the 2021 inauguration—were scored 0.4 to 0.6 by the model. The prompt explicitly instructs that domestic political events should score zero absent international implications, but the model appears to weight words such as "violence," "riot," and "security" heavily even in domestic settings.

Third, the model makes some minor temporal-attribution errors in both directions. The rubric instructs that retrospective coverage of old events should score zero unless current risks are mentioned. In some cases the model fails to recognize aftermath coverage as backward-looking: a 2004 article on legal turmoil in a Detroit terrorism prosecution (events eight months past) received 0.6, and a 2002 piece on routine airport security received 0.4. In other cases it applies the retrospective rule too aggressively: human-interest profiles published days after the 1993 World Trade Center bombing and the 1998 embassy bombings were scored 0 by the model even though the attacks were current events signaling elevated terrorism risk. The model evaluates the 2,000-character snippet in isolation, whereas a human coder knows the article's date and the surrounding history.

Fourth, the human coder sometimes credits context the snippet only implies. A 1961 wire item on a Navy missile contract was scored 0.2 by us (arms development at the height of the Cold War) and 0 by the model, which is arguably the more faithful reading of the prompt's instruction to classify based only on what an article states or strongly implies. A brief 2003 item on the capture of Ali Hassan al-Majid scored 0.6 in our coding—it presupposes the ongoing Iraq war—and 0 from the model. Reports on the humanitarian consequences of conflict, such as a 1973 piece on war-induced famine in Cambodia, scored higher in our coding because we treated humanitarian devastation and the implied risk of escalation as risk content while the model adhered to the narrower definition. In these cases the human is more generous than the prompt strictly licenses, so the disagreement reflects the rubric as much as the model.

Fifth, the model occasionally scores the backdrop rather than the subject of an article. When a major conflict serves as background for a story about something else—a 2023 piece on how the war in Ukraine affects livelihoods in a Norwegian border town, a 2017 color piece on the mood in a Beirut neighborhood—the model seems to scores the conflict (0.6) while we score the article (0.2). The conflicts are real, but the articles convey little incremental information about war risk. These are occasional instances where a human coder weighs the subject of the article, while the model weighs the most alarming entity mentioned in it.

Finally, the model appears to compress the top of the scale: it rarely uses the highest gradations and almost never scores above 0.8, whereas human coders occasionally do. The compression is mild and affects few observations, but it implies that the machine index understates the relative intensity of the most extreme episodes. Taken together, the audit supports treating the LLM scores as a reliable proxy for careful human coding, with the caveat that the index seems to run slightly hot at the bottom of the scale, and slightly cold at the top.

A.4 Hindsight Robustness Check

This appendix accompanies the hindsight discussion in Section 4.3. To test whether the model's knowledge of how historical events unfolded contaminates its scores, we re-score the articles around four major events with a date-anchored variant of the baseline prompt. The variant fixes each article's own publication date as the present, supplies the date, headline, and publisher alongside

the article text, and instructs the model that it has no knowledge of anything that happened afterward. The full prompt is reproduced below; its geopolitical risk definition and 0.0–1.0 scoring scale are identical to the baseline prompt of Section 2.2.

```
HINDSIGHT-CHECK PROMPT (system message, prepended with the article's own date):

Today is {date}
You are a geopolitical risk analyst reading a news article.  You possess
absolutely zero knowledge of any events that occurred after today.
Classify the article's assessment of geopolitical risk based only on what the
article states or strongly implies.
[The geopolitical risk definition and the 0.0--1.0 scoring scale follow verbatim
from the baseline prompt in Section 2.2.]

User message:
Date:  {date}
Title:  {title}
Publisher:  {publisher}
{article text}
```

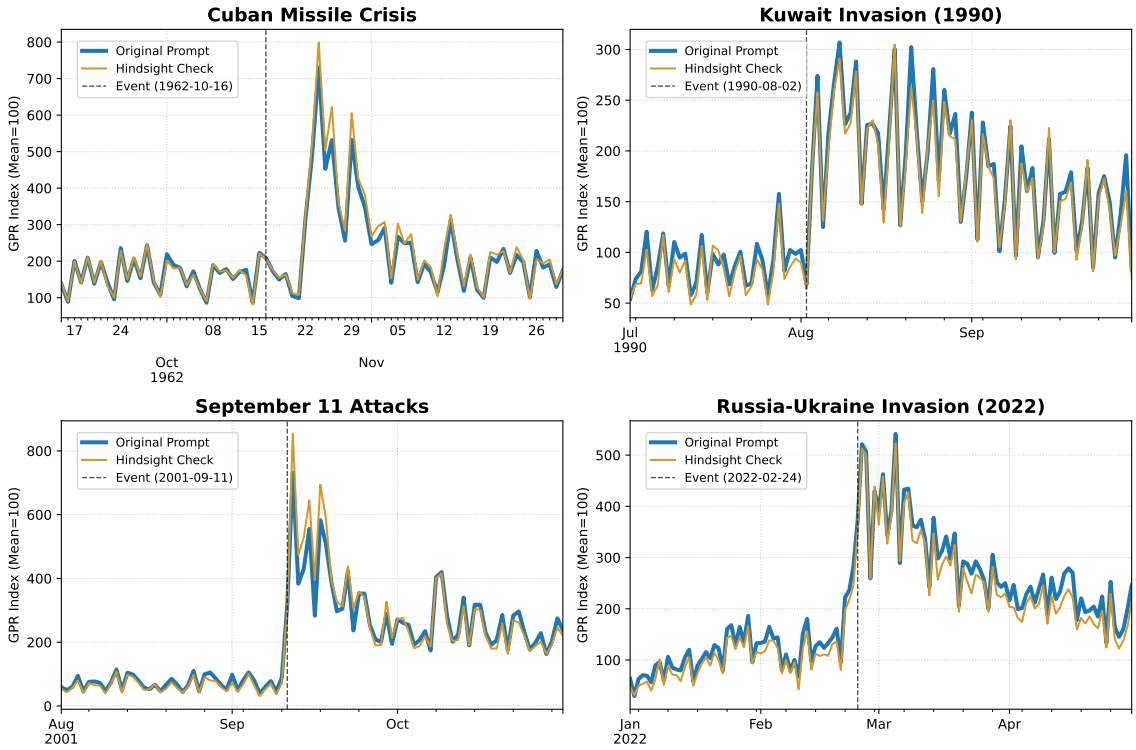
Figure A.2 plots the resulting daily series against the baseline around each event. The two are nearly indistinguishable. On average the hindsight-corrected scores are marginally higher for the Cuban Missile Crisis and September 11—the direction consistent with mild hindsight dampening in the baseline—and marginally lower for Kuwait and Russia-Ukraine, but the differences are very small throughout.

A.5 The Original Caldara and Iacoviello (2022) Keyword-Based GPR Index

As described in Section 3, we compare the AI-GPR index against the original, keyword-based GPR index developed by Caldara and Iacoviello (2022), using the same three newspapers and time period as the AI-GPR. In the original Caldara and Iacoviello (2022) formulation, the search terms are organized into eight categories, each combining a geopolitical term group with a risk/threat term group using either the AND or the N/2 (words must appear within 2 words of each other) proximity operators. The full search string is:

```
Caldara and Iacoviello (2022) KEYWORD-BASED GPR SEARCH QUERY:
(war OR conflict OR hostilities OR revolution* OR insurrection OR uprising OR
revolt OR coup OR geopolitical) N/2 (risk* OR warn* OR fear* OR danger* OR threat*
OR doubt* OR crisis OR troubl* OR disput* OR concern* OR tension* OR imminen* OR
inevitable OR footing OR menace* OR brink OR scare OR peril*)
OR (peace OR truce OR armistice OR treaty OR parley) N/2 (menace* OR reject* OR
threat* OR peril* OR boycott* OR disrupt*)
OR (military OR troops OR missile* OR ‘‘arms’’ OR weapon* OR bomb* OR warhead*)
AND (buildup* OR build-up* OR blockad* OR sanction* OR embargo OR quarantine OR
ultimatum OR mobiliz*)
OR (‘‘nuclear war’’ OR ‘‘atomic war’’ OR ‘‘nuclear missile*’’ OR ‘‘nuclear bomb*’’
```

Figure A.2: Baseline and Hindsight-Anchored AI-GPR Around Key Events



Notes: Each panel plots the daily AI-GPR index around a major geopolitical event under the baseline prompt (blue, thick) and under the hindsight-check prompt (orange, thin), which fixes the article's publication date as the present and instructs the model that it has no knowledge of subsequent events. Both series are scored on the same articles, normalized by total newspaper article count, and scaled to a mean of 100 over 1985–2019. Vertical dashed lines mark the event date. Source: authors' calculations using ProQuest TDM Studio data.

```
OR ‘‘atomic bomb*’’ OR ‘‘h-bomb*’’ OR ‘‘hydrogen bomb*’’ OR ‘‘nuclear weapon*’’)
AND (risk* OR warn* OR fear* OR danger* OR threat* OR doubt* OR crisis OR troubl*
OR disput* OR concern* OR tension* OR imminen* OR inevitable OR footing OR menace*
OR brink OR scare OR peril*)
OR (terroris* OR guerrilla* OR hostage*) N/2 (risk* OR warn* OR fear* OR danger*
OR threat* OR doubt* OR crisis OR troubl* OR disput* OR concern* OR tension* OR
imminen* OR inevitable OR footing OR menace* OR brink OR scare OR peril*)
OR (war OR conflict OR hostilities OR revolution* OR insurrection OR uprising OR
revolt OR coup OR geopolitical) N/2 (begin* OR begun OR began OR outbreak OR
‘‘broke out’’ OR breakout OR start* OR declar* OR proclamation OR launch*)
OR (allie* OR enem* OR foe* OR army OR navy OR aerial OR troops OR rebels OR
insurgen*) N/2 (drive* OR shell* OR advance* OR offensive OR invasion OR invad* OR
clash* OR attack* OR raid* OR launch* OR strike*)
```

```
OR (terroris* OR guerrilla* OR hostage*) N/2 (act OR attack OR bomb* OR kill* OR
strike* OR hijack*)
NOT (movie* OR film* OR museum* OR anniversar* OR obituar* OR memorial* OR arts OR
book OR books OR memoir* OR ‘‘price war’’ OR game OR story OR history OR veteran*
OR tribute* OR sport OR music OR racing OR cancer OR ‘‘real estate’’ OR mafia OR
trial OR tax)
```

The resulting index is the ratio of articles matching the proximity search to total articles published on each date, normalized to have a mean of 100 over 1985–2019. This is the "GPR (Original)" series reported throughout the paper.

A.6 Prompts for the AI-GPR Subindexes

We use the following prompts to classify articles for the GPR subindexes.

A.6.1 Threats and Acts GPR

The threats and acts classification prompt identifies the nature of geopolitical risk. Relevant articles are labeled based on whether they predominantly discuss anticipated risk (threats), realized risk (acts), or both.

```
THREATS AND ACTS CLASSIFICATION PROMPT:
Analyze this newspaper article about geopolitical risk.  This article has been
identified as discussing an adverse geopolitical event (Geopolitical risk).
Geopolitical risk (GPR) encompasses adverse geopolitical events and associated
risks, including:
- Wars (nuclear and conventional)
- Terrorist threats and acts
- Military buildups and tensions
- International crises and conflicts
TASK: Classify the nature of geopolitical risk discussed in this article.
The three types of geopolitical risk nature are defined as follows:

  1. threat:  Risks, tensions, warnings, or potential escalations related to
     adverse geopolitical events

  2. act:  Actual adverse geopolitical events that have occurred or are actively
     occurring (wars, attacks, terrorist acts, invasions, coups, etc.)

  3. both:  Substantive discussion of both realized events AND associated
     threats/risks

Rules:
- Choose ONLY ONE: 'threat', 'act', or 'both'
- Be precise:  only choose 'both' if the article substantially discusses BOTH
  temporal dimensions
- Brief mentions of past context while focusing on future risks = 'threat'
- Brief mentions of future risks while focusing on current/past events = 'act'
```

A.6.2 Oil GPR

The Oil GPR classification prompt first identifies whether an article talks about potential or realized oil supply disruptions. Relevant articles are then assigned regional labels based on where the oil supply disruption/threat is and which regions are affected. Multiple countries can be listed under the regional label.

```
OIL-GPR CLASSIFICATION PROMPT:
Analyze this newspaper article about geopolitical risk.
TASK 1: Determine if this article discusses events that could disrupt oil or
energy supplies. Oil/Energy disruption includes:
- Military conflicts in oil-producing regions
- Sanctions on oil-producing countries
- Attacks on oil infrastructure, pipelines, refineries, or shipping routes
- Political instability in major oil producers
- Blockades of oil shipping lanes (Strait of Hormuz, Suez Canal, etc.)
- Natural disasters affecting oil production
- Oil embargoes or export restrictions
- Terrorist attacks on energy infrastructure
Answer "yes" if the article discusses events that could affect oil/energy
production, transportation, or supply. Answer "no" if the article discusses
geopolitical events unlikely to affect oil/energy markets.

TASK 2: If "yes", identify which oil/energy producing region(s) are affected.
Select one or more from this list ONLY: Middle_East, Russia, USA, Venezuela,
North_Africa, West_Africa, Central_Asia, North_Sea, Canada, Mexico, Latin_America,
Southeast_Asia, China, Other
REGIONAL MAPPING RULES:
- Use 'Middle_East' for: Saudi Arabia, Iran, Iraq, Kuwait, UAE, Qatar, Bahrain,
  Oman, Yemen, Israel, Syria, Lebanon
- Use 'Russia' for: Russia, Soviet Union, USSR, Russian Federation
- Use 'USA' for: United States, America, USA, U.S. (as oil producer, not just
  consumer)
- Use 'Venezuela' for: Venezuela
- Use 'North_Africa' for: Libya, Algeria, Egypt, Tunisia
- Use 'West_Africa' for: Nigeria, Angola, Equatorial Guinea, Gabon, Congo
- Use 'Central_Asia' for: Kazakhstan, Azerbaijan, Turkmenistan, Uzbekistan
- Use 'North_Sea' for: Norway, UK (when referring to North Sea oil production)
- Use 'Canada' for: Canada (oil sands, Alberta, etc.)
- Use 'Mexico' for: Mexico
- Use 'Latin_America' for: Other Latin American oil producers (Brazil, Colombia,
  Ecuador, Argentina)
- Use 'Southeast_Asia' for: Indonesia, Malaysia, Brunei, Vietnam
- Use 'China' for: China (as oil producer)
- Use 'Other' for: Any other oil/energy producing region not listed above
IMPORTANT:
```

```
- Focus on regions WHERE the disruption is occurring, not where it's being
  discussed
- Include ALL regions mentioned that could experience oil/energy disruption
- If article mentions global impact or multiple regions, list all relevant ones
- If uncertain whether it affects oil/energy, err on the side of ``yes" if the
  region is a major producer
```

A.6.3 Country GPR (and Event)

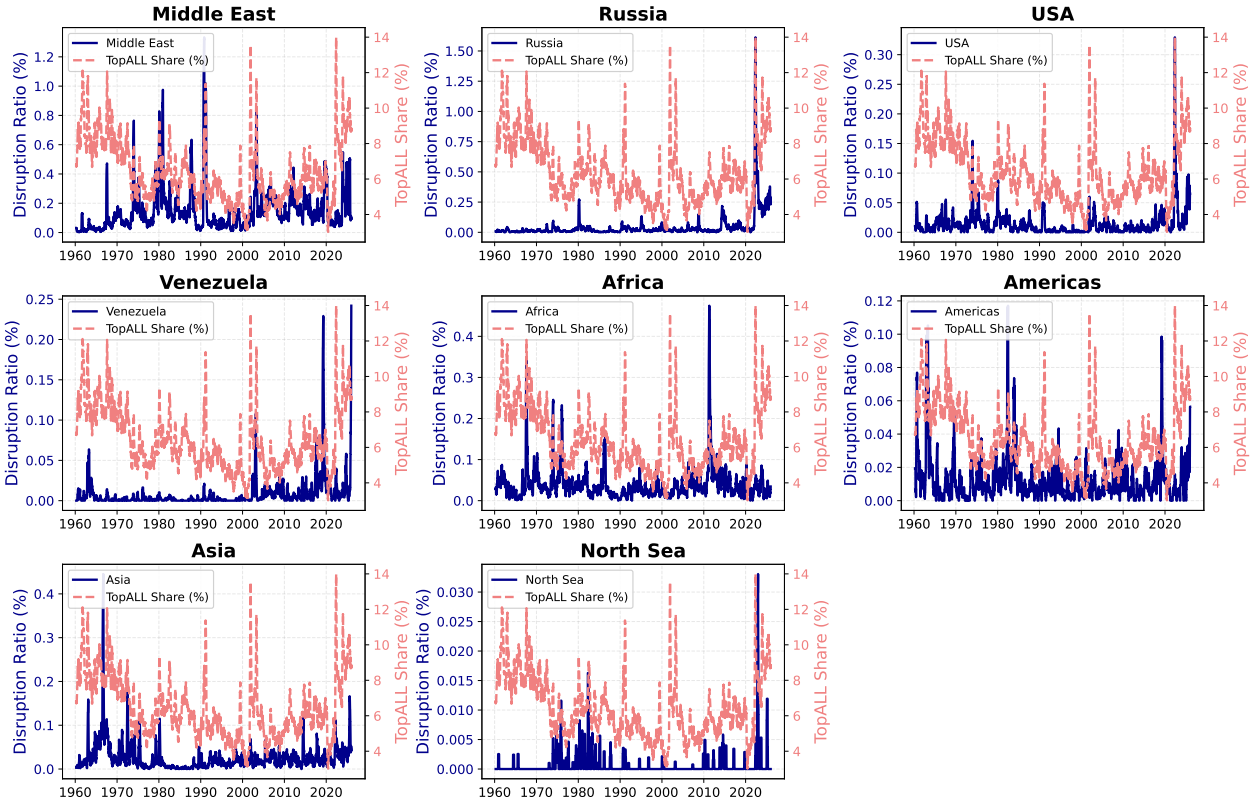
The country classifier prompt identifies the main countries involved in a geopolitical incident as portrayed in a given article. Relevant countries are assigned one of three actor labels indicating whether they are an initiator, a respondent, or a spillover party. The prompt also assigns an event-type label for the article's central geopolitical conflict.

```
COUNTRY GPR:
Analyze this newspaper article about a geopolitical risk event.
TASK: Identify the key country-level actors in this geopolitical event and assign
each a role.  Roles are defined as follows:

- initiator:  the country that launched or triggered the hostile action (military
  attack, invasion, sanctions, blockade, etc.)
- respondent:  the country that is the direct object of the hostile action
- spillover:  a country not directly involved but significantly affected (energy
  shock, refugee flows, trade disruption, etc.)

Rules:
- List up to 5 actors total; include only countries with a meaningful role in THIS
  article
- Use standard country names (e.g.  "United States", "Russia", "Saudi Arabia",
  "Israel")
- If a supra-national body (NATO, UN, EU) is the dominant actor, use its name
- Assign exactly one role per actor
- When initiation is genuinely contested or unclear (civil wars, mutual
  escalations, disputed territories), assign both main parties the role of
  initiator
- If the article is too vague to identify actors, return an empty actors list
Respond in JSON format only:
{"actors":  [
  {"country":  "<name>", "role":  "<role>"},
  ...
],
"event_type":  "<one of:  military_conflict, sanctions, terrorism,
diplomatic_tension, coup, civil_war, nuclear_threat, other>"}
```

A.7 GPR-Driven Oil Disruptions by Region

This appendix shows a decomposition of the Oil GPR index of Section 5.2 by geographic region. For each region, we apply the same two-stage LLM classification used for the aggregate index but retain only articles in which the oil or energy disruption occurs in, or affects, that region; each series is the sentiment-weighted count of such articles divided by total newspaper volume. Figure A.3 reports the eight regional panels and makes visible the shifting geography of energy-related geopolitical risk over the past six decades.

Figure A.3: GPR-Driven Oil Disruptions by Region



Notes: Each panel shows the regional Oil GPR index (90-day moving average) for a specific geographic region. The index is computed as the ratio of sentiment-weighted articles about oil disruptions in that region to total newspaper articles. The light red dashed line in each panel shows the overall AI-GPR share for reference. Regions are: Middle East, Russia, USA, Venezuela, Africa (North + West), Americas (Canada, Mexico, Latin America), Asia (Central, Southeast, China), and North Sea. Source: authors' calculations using the two-stage LLM classification described in Section 5.2. Last observation: December 2025.