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

Matteo Iacoviello† Jonathan Tong‡

March 11, 2026

Very Preliminary Draft

Abstract

We introduce an improved measure of geopolitical risk that builds on Caldara and Iacoviello (2022) and uses artificial intelligence to evaluate newspaper content. Our approach replaces keyword matching with semantic understanding: instead of searching for specific word combinations, we use one of the language models underlying ChatGPT to read newspaper articles and assess their geopolitical risk intensity. The daily AI-GPR index scores about 5 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 yes-or-no classifications, providing more nuanced measurement even at high frequencies. We demonstrate the potential of our approach with four applications. The AI-GPR index produces a stronger and more precisely estimated negative effect of geopolitical risk on stock returns. Combined with a second LLM classification layer, it yields a historical time series of geopolitical risk-driven oil supply disruptions by region. A third classification layer maps directed networks of geopolitical actors—initiators, respondents, and spillover countries—across major historical episodes. Finally, the layered architecture decomposes geopolitical risk by event type, by country, and by bilateral country pair—an approach that goes beyond what keyword-based methods can achieve.

JEL Classification: D80, E66, F51, G15.

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

*We thank Flora Haberkorn, the Proquest TDM Studio team, and seminar participants at the Federal Reserve Board 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, tensions between nations, and terrorism—is an important driver of macroeconomic fluctuations. Central to the empirical understanding of its effects is the ability to quantify latent political tension 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 track global instability through newspaper archives.[1] Their index has since become a standard benchmark in the literature, providing a transparent and computationally efficient proxy for the risks that shape trade patterns, policy uncertainty, investor sentiment, and asset prices.

While keyword-based methods have proven highly effective at capturing the broad contours of global risk and other economic phenomena, they are subject to the inherent constraints of dictionary matching. These methods may struggle to distinguish between articles where a geopolitical event is the primary subject from those where keywords appear in a peripheral or historical context. Furthermore, fixed dictionaries may miss evolving terminology or fail to capture the nuanced intensity 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 five million articles from the *New York Times*, *Washington Post*, and *Chicago Tribune* between 1960 and 2025. Our approach maintains the definitions established in the prior literature but utilizes the LLM's contextual awareness to score articles on a continuous scale (0.0 to 1.0). This allows for a more granular identification of risk, effectively separating the "signal" of central geopolitical events from the "noise" of peripheral mentions. The resulting index complements the original GPR by offering a smoother and more accurate time series that retains the transparency of the original framework through auditable, prompt-based classification criteria.

We demonstrate the value of the AI-GPR index through four 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. Second, we employ a layered classification architecture to identify articles specifically related to geopolitical oil supply disruptions. This yields a novel, 65-year time series of GPR-driven oil shocks disaggregated by geographic region.

Third, we leverage the LLM's ability to parse complex linguistic structures to extract "subject-

[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.

verb-object" relationships, identifying country-level actors and their specific roles during adverse geopolitical events—as initiators, respondents, or spillover parties. We use this data to construct directed network graphs that visualize the topology of global conflict across major historical episodes.

Fourth, we provide a bilateral decomposition of geopolitical risks that represents an important advance relative to prior methods. While keyword-based approaches are limited to identifying the co-occurrence of country names within a text, our methodology explicitly identifies directed actions between states. This allows us to construct a high-frequency, "who-on-whom" measure of geopolitical stress, providing a structural view of bilateral state relations that was previously inaccessible at this scale. By mapping these directed links, we can identify the specific nodes and edges that drive aggregate global risk.

We make all the measures presented in this paper available to the research community in a companion webpage. The scope of these new measures has itself been shaped by that 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 Caldara and Iacoviello's index.

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 many 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.

The paper proceeds as follows. Section 2 describes the data and the LLM-based methodology. Section 3 presents the AI-GPR index and compares it with the keyword-based benchmark. Section 4 reports our four empirical applications. Section 5 details the decompositions of geopolitical risk, and Section 6 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.

We adopt a two-stage process that first filters for articles likely to discuss geopolitical events and then passes them through an LLM for scoring. The advantages of this method are two-fold. For one, we are able to more efficiently screen for relevant articles. Given the broad coverage of newspapers, many articles are unlikely to have any discussion of geopolitical risks. For articles that are screened through, the LLM is more accurately able to classify the articles for geopolitical risk and intensity than alternative approaches like query-based classifications.

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 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 criteria extracts 4.6 million articles from the 8.5 million total articles published across the three publications over the sample period.[2] Of these 4.6 million articles, 1.2 million (26 percent) receive a positive GPR score from the LLM. Section 2.5 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

[2] The denominator counts articles containing at least seven common words ("the," "be," "to," "of," "and," "at," and "in"). This choice filters out non-editorial content such as photos with a brief caption, classified advertisements, table-of-contents entries, and brief notices, and serves as a normalization for trends in newspaper articles volume over the years.

the first 2,000 characters. Due to the inverted pyramid structure of journalism where the most important information is presented first, the truncation captures the essential geopolitical content for the vast majority of articles. As we show in Section 2.5, our baseline scores based on truncated text are nearly identical to those obtained using the full article text.

2.2 LLM Classification: Definition, Prompt, Implementation

We use Caldara and Iacoviello (2022)'s definition of geopolitical risk, which encompasses wars (initiation of new military conflicts between states or organized groups), escalation of existing wars (intensification of ongoing military conflicts), major terrorist attacks (violent acts intended to intimidate civilian populations or governments), and tensions among states and political actors (diplomatic disputes, threats, military buildups, and other developments that affect the peaceful course of international relations). This definition provides clear boundaries for what constitutes geopolitical risk while encompassing the full range of relevant events. Importantly, it focuses on current risks and actual 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 among states and political actors
that affect the peaceful course of 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.
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, books, anniversaries of old events, and obituaries should receive a score
of 0.0 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
or cross-border conflicts.
```

As the prompt illustrates, we use zero-shot prompting: the model receives explicit criteria and

a scoring scale but no labeled examples. One-shot or few-shot prompting would require embedding one or more example articles in every request; at the scale of approximately 4.5 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. 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 implement the model with temperature set to zero, which minimizes stochastic variation in responses and ensures consistent scoring across multiple runs. We use structured output to return model output in JSON format, making extraction and processing of scores more efficient./[3]

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). The API returns a JSON object containing the geopolitical risk score. We validate that 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.

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 researchers to better identify the specific timing of geopolitical events and their economic effects. This precise temporal resolution is particularly valuable for event studies and examining the immediate market reactions to geopolitical developments.

[3] We use structured output because it makes retrieving model output efficient and standardized. But structured output would be particularly useful for complex classification tasks with multiple dimensions to score and more complex prompts and instructions.

2.4 Comparison with Keyword-Based Approaches

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 the inclusion of a "geopolitical" word equally, so it tends to overvalue the geopolitical content of articles that contain "geopolitical" terms. Second, the approach is terminology-dependent, requiring exhaustive keyword lists; articles discussing "armed confrontation," "military standoff," or "bellicose rhetoric" may be missed if these exact phrases are not in the dictionary or if the proximity requirements are not satisfied. In practice, it isn't realistic nor feasible to create a query that accounts for all possible "geopolitical" terms that are 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 involves complex neural networks and are effectively a black-box. To introduce as much certainty and human autonomy as possible into the LLM's classifications, we make deliberate, strategic 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.

2.5 Measurement Error

Four potential sources of measurement error affect the AI-GPR index: model stochasticity, text truncation, systematic bias relative to human judgment, and attenuation from the two-stage filtering process.

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. The correlation between the first two runs is 0.99, with 98% of articles receiving identical scores and a mean absolute difference of 0.004. The correlation between the first and the third run is also 0.99, with 98.6% of articles receiving identical scores and a mean absolute difference of 0.003. This near-perfect reproducibility confirms that classification noise from model stochasticity is negligible. To further bound this concern, we re-ran the model at temperatures of 0.5 and 1.0 on the same articles. Within-run agreement remains above 91 percent even at temperature 1.0, and cross-temperature agreement between temperature 0.0 and temperature 1.0 is 93.2 percent with a mean absolute difference of 0.014. Mean scores are virtually indistinguishable across temperature settings, confirming that the choice of temperature has negligible impact on classification outcomes [4]. We further bound this concern through the multi-model comparison reported in Section A.1: pairwise correlations across four OpenAI models—spanning two generations and different model sizes—range from 0.86 to 0.94, 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

[4] Correlation and agreement across temperatures remain robust when compared to the human-annotated sample. The correlation between human annotations and model scores is 0.8 for both the temperature 0.5 and temperature 1.0 runs, and the agreement rate is 75 percent for both runs.

the same 1,050 articles using the complete article text. The correlation between truncated and full-text scores is 0.93, with 86.7 percent of articles receiving identical scores. Full-text scores are modestly higher on average (mean 0.153 versus 0.129), consistent with longer articles containing additional geopolitical content beyond the opening paragraphs, but the cross-article ranking is nearly identical. We conclude that truncation introduces minimal classification error.

Systematic bias. The LLM may systematically over- or under-score certain types of articles relative to human readers. To assess this, we independently scored 1,050 articles using the same definition and gradations provided to the LLM. The Pearson correlation between human and AI scores is 0.812, with an overall agreement rate of 91.6 percent on whether an article contains geopolitical risk. Disagreements are nearly symmetric and concentrated among articles with the lowest positive AI scores, reflecting marginal geopolitical risk content rather than clear misclassification. The AI's type-I error rate of 13.3 percent compares favorably with the 21 percent reported by Caldara and Iacoviello (2022) for keyword-based classification. The mean AI score slightly exceeds the mean human score, suggesting mild upward bias. Human-AI agreement is stable across decades, with Pearson correlations ranging from 0.78 to 0.85 decade by decade and no systematic trend, suggesting that measurement error is not concentrated in any particular historical period.

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. Fewer than 1 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),[5] whose 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.

Relationship to debiasing approaches. 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 Caldara and Iacoviello's GPR index, they find that the keyword-based series underestimates geopolitical risk once the correction is applied. A practical challenge with this approach is that the validation sample must be large enough and representative enough to

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

pin down the bias over time, which can require extensive hand-annotation. Our LLM approach addresses the underestimation problem directly at the classification stage—by replacing keyword matching with semantic understanding—and produces a similarly elevated series without requiring a post-hoc statistical correction. The stable human-AI agreement across decades documented above (correlations 0.78–0.85) suggests that residual bias in the AI-GPR is small and does not vary systematically over time, limiting the scope for the kind of time-varying underestimation that the Carlson and Dell's correction is designed to address.

2.6 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 4.5 million articles from 1960–2025, 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. With parallelization across multiple API calls, we can process the full corpus in approximately 250 hours of computation time. 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.

3 The AI-GPR Index

3.1 Summary Statistics

Table 1 presents summary statistics for the AI-GPR index alongside a keyword-based GPR index over the 1960–2025 period. 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). Note that the original Caldara and Iacoviello (2022) index uses ten newspapers and begins in 1985; our implementation extends coverage back to 1960 using three newspapers to ensure a like-for-like comparison with the AI-GPR. The full search string is reported in Appendix A.2. Both indices are normalized so that the mean equals 100 over 1985–2019.

Several features of the AI-GPR index stand out. The AI-GPR index has a lower standard deviation (48.5 vs. 60.6) and lower maximum value (718 vs. 1,200) than the original GPR, reflecting the index's greater smoothness. The AI-GPR has a higher minimum (21.3 vs. 0.0), consistent with the fact that news articles always carry some low-level geopolitical discussion even on quiet days. The higher 90-day autocorrelation of the AI-GPR (0.73 vs. 0.62) indicates greater persistence:

geopolitical conditions evolve smoothly rather than spiking and reverting.[6] The correlation between the two indices is 0.69, indicating that they capture similar dynamics while exhibiting meaningful differences.[7]

Figure 1 displays the AI-GPR index as a 90-day moving average over the full 1960–2025 sample, with labels highlighting five major geopolitical episodes.

3.2 Comparison with the Original GPR Index

Figure 2 plots the AI-GPR alongside the original GPR index as 90-day moving averages. 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, consistent with the AI model's ability to detect geopolitical risk in articles that use language conventions of the era that may not match 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 effectively 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.

As conjectured by Carlson and Dell (2025), the AI version of the GPR index features a share of articles discussing geopolitical risks that 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 to just 3.3 percent under the original, keyword-based approach.[8]. Nevertheless, the original GPR retains considerable validity as a measure of geopolitical risk: the two series share a correlation of 0.69 and co-move closely around major events, suggesting that while the keyword approach may

[6] 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.

[7] A comparison of classification rates with Caldara and Iacoviello (2022) can be reconciled as follows. Caldara and Iacoviello (2022) classify 3.3 percent of articles as GPR-relevant, but report a 21 percent Type-I error rate and a 2.6 percent false-negative rate among excluded articles. Correcting for both, their implied true GPR rate is approximately 5.1 percent ($= 3.3 \times 0.79 + 96.7 \times 0.026$). In our approach, 26 percent of sampled articles (15 percent of all articles) receive a positive score, but two-thirds of these receive scores of 0.2 or 0.4, reflecting low-intensity geopolitical risk content that keyword matching systematically misses. Only one-third—approximately 9 percent of sampled articles, or 5 percent of all articles—score above 0.5. This high-intensity subset is directly comparable to Caldara and Iacoviello's adjusted rate of 5.1 percent, suggesting both approaches identify a similar core of high-salience geopolitical articles. Small differences in these rates reflect the fact that Caldara and Iacoviello's validation statistics refer to their ten-newspaper index over 1985–2019, while ours covers three newspapers from 1960–2025.

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

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 six major geopolitical events: the Cuban Missile Crisis (1962), the Yom Kippur War (1973), the Kuwait Invasion (1990), the Gulf War (1991), the September 11 Attacks (2001), and Russia's invasion of Ukraine (2022). The daily (non-smoothed) indices reveal important differences.

For more recent episodes—September 11 and the Russia-Ukraine invasion—the AI-GPR index shows less day-to-day volatility while still capturing the spike in risk. One possible explanation is that modern journalistic language is more readily interpretable by the LLM: the AI model can parse contemporary writing style 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 and Yom Kippur War, the AI-GPR also captures the risk elevation but with a smoother profile, likely because the semantic approach is less sensitive to day-to-day variation in keyword frequency.

3.4 Model Comparison

To assess whether our choice of LLM affects the resulting index, we classify a random sample of 1,050 articles using four different OpenAI models: GPT-4o-mini (our baseline), GPT-4o, GPT-5-mini, and GPT-5. All models receive the same prompt and are run at temperature zero. The pairwise correlations between GPT-4o-mini scores and each alternative model are 0.90 (GPT-4o), 0.87 (GPT-5-mini), and 0.86 (GPT-5). Within the GPT-5 family, the correlation between GPT-5-mini and GPT-5 is 0.94. The share of articles classified as containing geopolitical risk (score > 0) is remarkably similar across models, ranging from 26 to 33 percent. These results indicate strong agreement across models spanning two generations and different model sizes, 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.

4 Applications

We illustrate the usefulness of the AI-GPR index through three applications. The first revisits the relationship between geopolitical risk and stock returns, showing that the AI-GPR provides a sharper estimate of this effect. The second uses the AI-GPR index as an input to a second LLM classification layer that identifies articles specifically about oil supply disruptions, producing a novel historical time series of GPR-driven oil shocks by region. The third applies a further LLM layer to extract the country-level actors involved in geopolitical events and map directed networks of

conflict and spillover.

4.1 Geopolitical Risk and Stock Returns

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.[9] 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 simple average of the AI-GPR over those $k+1$ days (including day t itself), so that geopolitical information accumulating over weekends and holidays 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:

$$r_t^e = \alpha + \beta\, \Delta\text{GPR}_t + \varepsilon_t \tag{2}$$

where r_t^e is the excess market return (in percent). Table 2 reports the results, estimated on data collapsed to weekly frequency, where weekly returns are cumulated and the GPR change is computed from week to week. The OLS estimate of β is -0.14: a one-standard-deviation increase in the AI-GPR innovation is associated with a 14 basis point decline in the weekly stock market return.

The OLS estimate in equation (2) 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:

$$\Delta\text{GPR}_t = \phi_0 + \sum_{j=1}^{4} \phi_j\, \Delta\text{GPR}_{t-j} + u_t \tag{3}$$

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:[10]

$$r_t^e = \alpha + \beta_1\, \widehat{\Delta\text{GPR}}_t + \beta_2\, \hat{u}_t + \varepsilon_t \tag{4}$$

[9] 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.

[10] 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 reports the results. The estimated coefficient on the persistent component ($\hat{\beta}_1 = -0.27$) is more than twice as large in absolute value as the coefficient on the shock component ($\hat{\beta}_2 = -0.12$). The persistent and shock components are statistically significant at the 5 and 10 percent level, respectively. This finding has a clear economic interpretation. While a sudden geopolitical event—such as the September 11 attacks—generates an immediate negative return, it is the "long tail" of that risk—the fact that the world remains dangerous in the weeks that follow—that truly drives the equity risk premium upward and prices downward. Investors do not merely react to headlines; they aggressively re-price assets as geopolitical uncertainty becomes embedded in the macroeconomic outlook.

The discrepancy between the aggregate OLS coefficient (-0.14) and the decomposed coefficients (-0.27 and -0.12) suggests that a simple linear regression suffers from an information dilution effect. After decomposing the risk, the market is roughly twice as sensitive to the persistent component of geopolitical risk than the raw OLS estimate suggests. The aggregate OLS estimate is attenuated by the high-frequency noise of weekly shocks, which are large in magnitude but transient in their effect on asset prices.[11]

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 OLS coefficient on the keyword-based index is -0.09, roughly one-third smaller than the AI-GPR estimate and significant only at the 10 percent level. In the decomposition, the persistent component (-0.24) is marginally significant at the 5 percent level, while the shock component (-0.07) loses significance entirely. These differences illustrate how measurement noise in the keyword-based index attenuates the estimated effects: because keyword matching misclassifies articles at higher rates on any given day, the resulting daily index contains more non-geopolitical variation, which biases 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.

4.2 GPR-Driven Oil Supply Shocks

As a second application, we combine the AI-GPR classification with a second LLM layer to construct a historical time series of oil supply disruptions driven by geopolitical events. We apply a two-step classification process. In the first step, described in Section 2.2, each article receives a GPR score from 0.0 to 1.0. In the second step, we apply a second LLM prompt to all articles in the universe of

[11] This result echoes the findings in Gonçalves et al. (2025), who examine the pricing of geopolitical risk using stock market returns. They use the Caldara and Iacoviello (2022)'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.

geopolitically relevant articles—those receiving a GPR score above 0.5.[12] This prompt asks GPT-4o-mini to determine whether the article discusses an oil or energy supply disruption caused by a geopolitical event (yes/no), 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.

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

$$\text{Oil-GPR}_t = \frac{\sum_{i \in D_t} S_{it}}{A_t} \tag{5}$$

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 0.3). 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 4 plots the aggregate Oil-GPR index as a 90-day moving average alongside the AI-GPR index over 1960–2025. The oil disruption measure captures the major geopolitical oil supply events: the 1973 Arab oil embargo, the 1979 Iranian Revolution, the 1990 Gulf War, and the sanctions-related disruptions in the 2010s–2020s.

Figure 5 decomposes the Oil-GPR index by region. The panels illustrate the shifting geography of geopolitical oil risk over six decades. The Middle East dominates through the 1990s, reflecting the Arab-Israeli conflicts, the Iran-Iraq War, and the Gulf War. Russia becomes prominent in the 2000s and especially after 2014 and 2022. Venezuela registers during the political instability of the 2000s and 2010s. North Africa shows episodic spikes during the Libyan civil wars.

Table 3 reports summary statistics for the oil disruption classification. 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. North Africa, the United States, Southeast Asia, and West Africa each account for between 5 and 11 percent of mentions. Because articles may reference multiple producing regions—for instance, an article about OPEC production cuts may mention both the Middle East and Venezuela—the regional shares sum to more than 100 percent.

[12] 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.

4.3 Geopolitical Risk Networks

As a third application, we use the AI-GPR classification as a foundation for mapping the structure of geopolitical conflict. We apply a second-stage LLM classifier to articles with GPR scores above 0.5, asking GPT-4o-mini to identify the country-level actors involved in each geopolitical event and to assign each actor one of three roles: initiator (the country that launched or triggered the hostile action), respondent (the country that is the direct target), or spillover (a country not directly involved but significantly affected, e.g., through energy shocks, refugee flows, or trade disruption).

For each article, the model identifies up to five actors and classifies the event type (military conflict, sanctions, terrorism, diplomatic tension, coup, civil war, or other). We apply this classifier to articles from four historical episodes: the Cuban Missile Crisis (1962), the Gulf War (1990–91), the September 11 attacks (2001–02), and the Russia-Ukraine conflict (2022–23). For each period, we restrict the sample to high-GPR articles whose text contains at least one event-specific keyword (e.g., "Cuba," "missile," "Kennedy" for the Cuban Missile Crisis), drawing up to 1,000 articles per period.

From the classified articles, we construct directed network graphs where nodes represent countries and edges represent co-occurrence in geopolitical events. Node color indicates the country's dominant role (initiator, respondent, or spillover), node size is proportional to the country's total GPR-weighted involvement, and edge width reflects the intensity of co-occurrence between actor pairs. Directed edges run from initiators to respondents, with softer links connecting initiators to spillover countries.

Figure 6 presents the resulting networks for the four episodes. The Cuban Missile Crisis panel shows the United States and Russia as the central initiator-respondent pair, with Cuba appearing as a respondent and countries such as China and France as spillover nodes.[13] The Gulf War panel highlights Iraq as the dominant initiator with Kuwait as the primary respondent and a broad set of spillover nations including Saudi Arabia, Israel, and the United States. The September 11 panel features the United States as the primary respondent to terrorism, with Afghanistan as initiator and a wide network of spillover countries reflecting the global reverberations of the attacks. The Russia-Ukraine panel shows Russia as the dominant initiator and Ukraine as respondent, with NATO, the United States, and European nations appearing as spillover countries reflecting the economic and security implications of the conflict.

These network visualizations illustrate how the layered LLM architecture can extract structured relational information from unstructured text, going beyond aggregate risk measurement to characterize the actors and directionality of geopolitical conflict. The approach could be extended to construct time-varying network measures of geopolitical centrality, connectedness, or contagion

[13] India was another important node during the Cuban crisis. China launched a major invasion of India's northern border on October 20, 1962, precisely when the U.S. and Soviet Union were occupied with the Caribbean standoff. The crisis forced India to fight a war while major powers were distracted.

risk.

5 Decomposing Geopolitical Risk

Beyond constructing an aggregate index, the LLM architecture enables structured decomposition of geopolitical risk in ways that keyword matching cannot easily replicate. Because each article is classified not only for overall risk intensity but also for the type of event, the countries involved, and the roles those countries play, the same scored corpus supports multiple complementary decompositions. Caldara and Iacoviello (2022) constructed event-type and country sub-indices of their keyword-based GPR index; we extend their framework and introduce a third dimension—bilateral country pairs—that requires the relational understanding that large language models uniquely provide.

5.1 GPR by Event Type

The country classifier assigns each article to one of eight event categories: military conflict, diplomatic tension, terrorism, civil war, nuclear threat, coup or regime change, sanctions, and other. We aggregate the AI-GPR sentiment scores within each category at monthly frequency and scale the resulting series to the overall AI-GPR index, so that the sub-indices are directly comparable in units. Figure 7 presents the 12-month moving averages for all eight categories.

Military conflict is the most prevalent category and closely tracks the aggregate index, with prominent peaks during the Korean and Vietnam Wars, the Gulf War, the September 11 aftermath, and Russia's invasion of Ukraine in 2022. Diplomatic tension forms a lower-level but persistent baseline that elevates during periods of superpower rivalry and subsides in detente. Terrorism rises sharply following the September 11 attacks and remains elevated through the mid-2000s before declining, reflecting both the actual trajectory of global terrorism and the evolution of media coverage. Sanctions become increasingly prominent after 2014, capturing the growing use of economic coercion as a geopolitical instrument. The decomposition broadly confirms the event-type patterns documented by Caldara and Iacoviello (2022) using keyword-based classification, while extending them with continuous intensity scores across a richer set of categories.

5.2 Geopolitical Risk for One Hundred Countries

For each article the classifier identifies which countries are involved in the geopolitical event and assigns each an actor role. We aggregate the total AI-GPR sentiment contribution for each country across all roles and scale to the overall index. Figure 8 presents monthly series for the United States, Russia, China, and the United Kingdom—the four countries most consistently featured in the geopolitical risk literature. The companion webpage includes time series of geopolitical risk for

the one hundred countries most often mentioned in the 65-year history.

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. 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 systematic ranking across all countries mentioned in the news without requiring pre-specified keyword lists for each.

5.3 Bilateral Geopolitical Risk

The bilateral index is the genuinely new contribution of this decomposition. Keyword approaches can identify that Russia and the United States both appear in an article, but they cannot distinguish whether Russia is acting on the United States, whether the United States is 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. We construct bilateral indices for each country pair as the sum of AI-GPR sentiment scores in articles where both countries appear as geopolitical actors, treating the pair symmetrically across initiator and respondent roles for the purpose of constructing time series.

Figure 9 presents the 12-month moving averages for the six highest-ranked bilateral pairs by cumulative GPR intensity over 1960–2025. The United States–Russia pair dominates the Cold War era and surges again after 2022, the two great geopolitical rivalries of the sample bookending the post-Cold War interlude. The India–Pakistan pair registers persistent elevated risk reflecting the long-running territorial and nuclear rivalry on the subcontinent. The United States–Vietnam pair rises sharply during the Vietnam War and then subsides entirely—a clean illustration of bilateral risk that is historically bounded and geographically specific. The China–United States pair shows a steady rise from the early 2000s with notable acceleration after 2017, consistent with the literature on the emergence of great-power competition between the two economies. Across all pairs, the bilateral index illustrates not merely that two countries are salient in geopolitical coverage but that they are bound together by a specific conflictual relationship—a level of relational precision that keyword-based methods simply cannot achieve.

6 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 demonstrated four 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. 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, revealing the shifting geography of energy-related geopolitical risk over six decades. Third, by applying a further classification layer to extract geopolitical actors and their roles, we mapped directed networks of conflict across major historical episodes, demonstrating how the layered architecture can recover structured relational information from unstructured text. Fourth, we decomposed geopolitical risk by event type, country, and bilateral pair. The event-type and country decompositions extend the analogous sub-indices of Caldara and Iacoviello (2022) using richer LLM-based scores; the bilateral dimension is genuinely new, identifying which countries are acting on whom rather than merely co-occurring in coverage.

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 network classification approach, or examine the relationship between geopolitical risks and various economic outcomes. The flexibility of prompt-based classification also enables researchers to measure related concepts such as domestic political risk, climate-related risks, or other forms of economic uncertainty using similar methods.

References

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.

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.

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.

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.

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

Table 1: AI-GPR Index: Summary Statistics

Statistic	**AI-GPR**	**GPR (Original)**
Standard Deviation	48.52	60.55
Skewness	1.64	1.97
1st Percentile	41.97	18.14
25th Percentile	79.61	68.22
Median	105.46	99.15
75th Percentile	138.33	139.28
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 full 1960–2025 daily sample. The AI-GPR index is computed as the sum of LLM-assigned geopolitical risk scores for all articles on each date, normalized by total newspaper article count. The original GPR is from Caldara and Iacoviello (2022). 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 average daily share of all newspaper articles classified as containing geopolitical risk: for the AI-GPR, articles receiving a score > 0; for the original GPR, articles matching the keyword proximity search. Both are expressed as a share of total newspaper articles published. 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 (i.e., matching the broad keyword filter) that receive a score > 0; for the original GPR, this coincides with Positive Articles Share (%) since there is no separate sampling stage.

Table 2: Weekly Stock Returns and Geopolitical Risk

	(1) OLS	(2) Decomposition
ΔGPR_t	-0.136^{**}	
	(0.057)	
Persistent component $\widehat{\Delta\text{GPR}}_t$		-0.271^{**}
		(0.134)
Shock component $\hat{u}_t$		-0.119^{*}
		(0.062)
Constant	0.137^{***}	0.138^{***}
	(0.037)	(0.037)
Observations	3,452	3,448
R^2	0.004	0.004

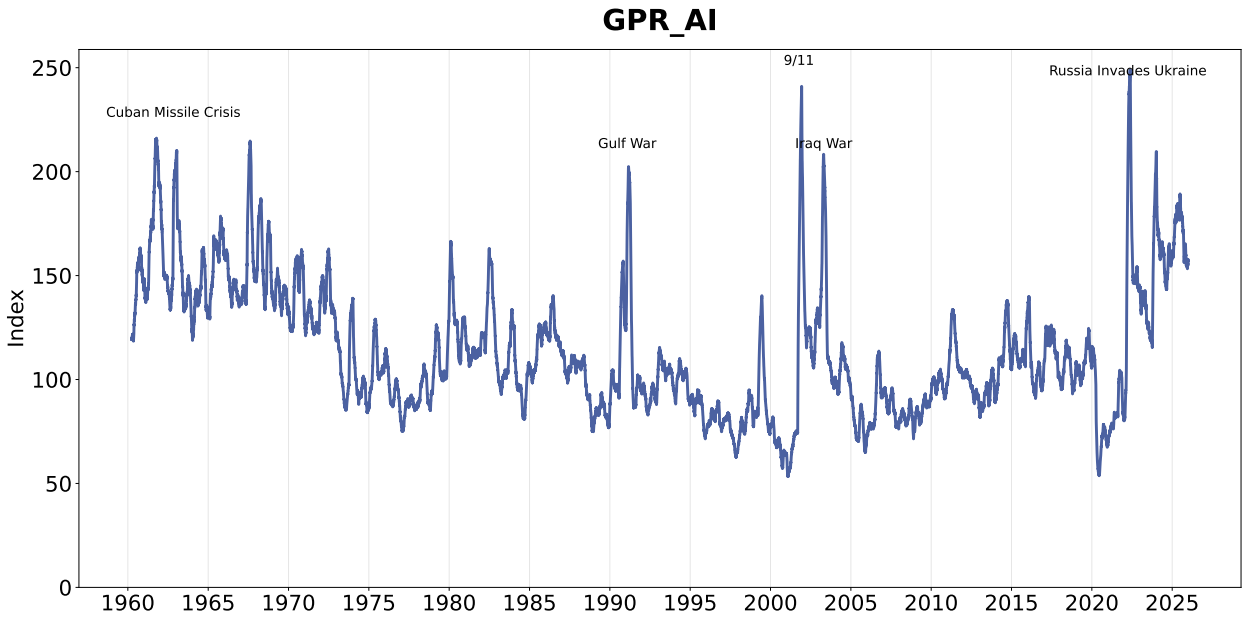
Notes: Weekly regressions of excess market returns (r_t^e, in percent) on the standardized change in the AI-GPR index. Weeks run Monday through Sunday; weekly stock returns are the sum of daily excess returns within the week, and the weekly GPR variable is the change in the within-week average of the trading-day GPR measure, standardized over the full sample. Column (1): OLS regression of r_t^e on ΔGPR_t, as defined in equation (2). Column (2): decomposition of ΔGPR_t into a persistent component $\widehat{\Delta\text{GPR}}_t$ (fitted values from an AR(4) model) and a shock component $\hat{u}_t$ (residuals), as in equation (4). The persistent component coefficient (-0.27) is roughly twice the aggregate OLS coefficient (-0.14), indicating that the OLS estimate is attenuated by transient high-frequency noise. The sample spans 1960–2026 (through March 6, 2026). 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. Column (1) reports heteroskedasticity-robust (HC1) standard errors in parentheses. Column (2) reports pairs-bootstrap standard errors (1,000 replications) to account for the generated regressors in the two-step procedure. A block bootstrap with 4-week blocks yields standard errors of 0.138 and 0.063 for the persistent and shock components, respectively ($p = 0.049$ and $p = 0.059$), confirming that both components remain significant at conventional levels under this more conservative approach. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 3: GPR-Driven Oil Disruption Classification: Summary Statistics

Statistic	**Value**
Classification rates (% of articles with GPR score > 0.5*):*	
Containing oil/energy keywords	12.6%
Classified as oil disruption	9.3%
Of keyword articles classified as disruption	73.8%
Disruptions by region (% of disruption articles):	
Middle East	62.9%
Russia	14.1%
North Africa	10.8%
USA	5.7%
Southeast Asia	5.8%
West Africa	5.7%
Latin America	5.2%
China	3.8%
Venezuela	3.0%
Other	3.8%

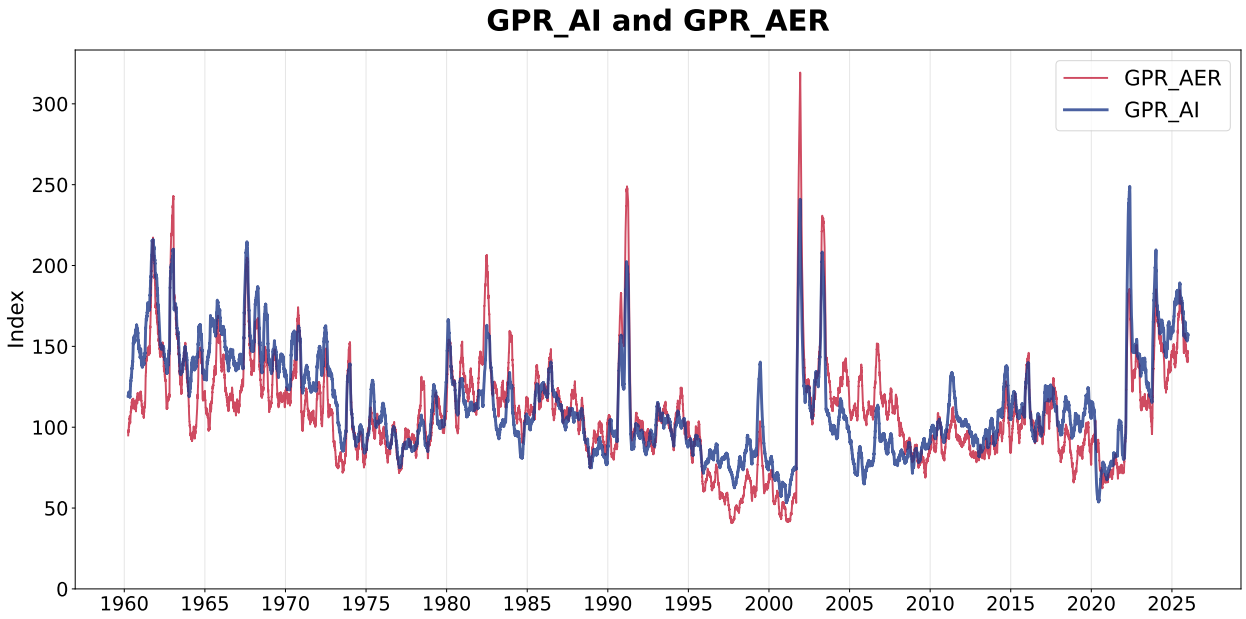
Notes: Summary statistics for the two-stage oil disruption classification. All articles with a GPR score above 0.5 are classified by a second LLM prompt to determine whether the article discusses an oil or energy supply disruption driven by a geopolitical event and, if so, which region(s) are affected. Regional percentages sum to more than 100% because articles may mention multiple regions. "Other" includes Central Asia, Mexico, North Sea, and Canada. Sample: New York Times, Washington Post, and Chicago Tribune, 1960–2025.

Figure 1: The AI-GPR Index (90-Day Moving Average)



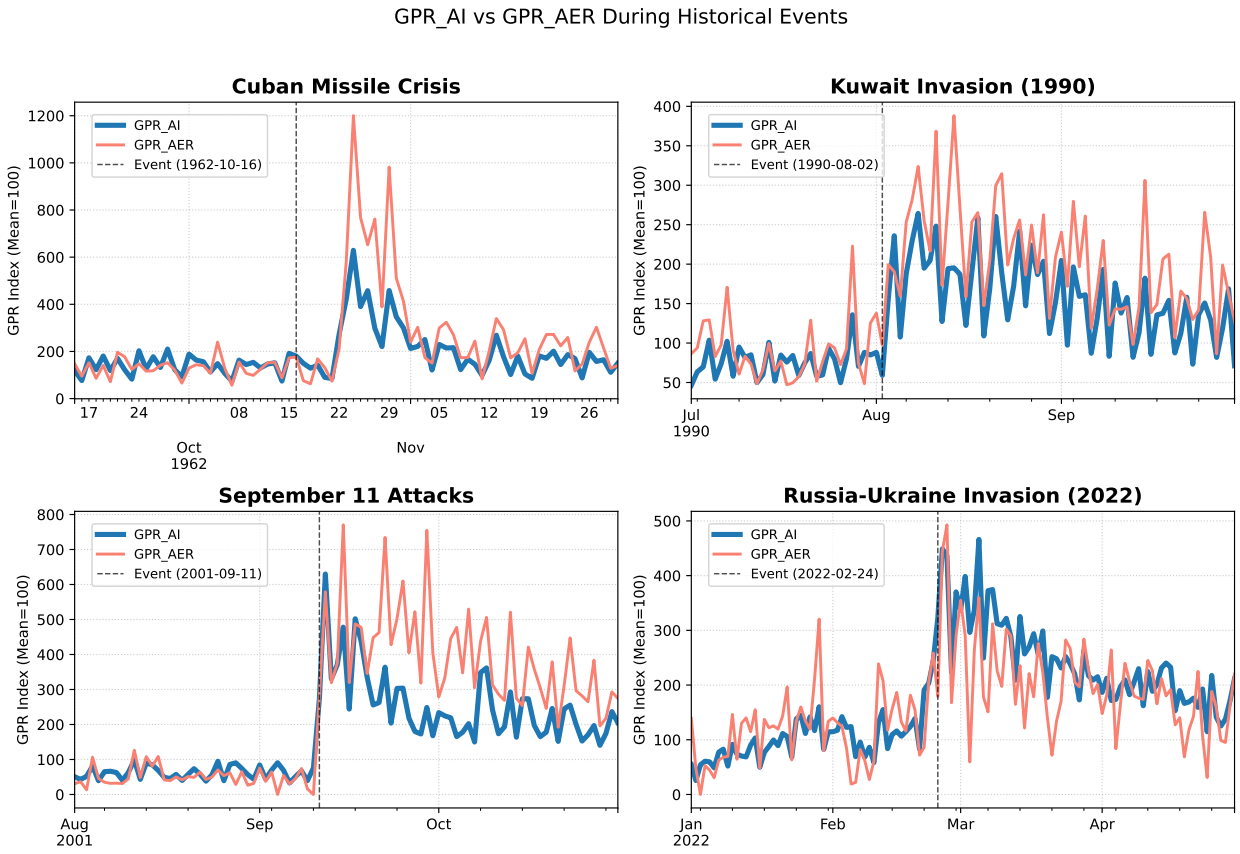
Notes: The AI-GPR index plotted as a 90-day moving average, 1960–2025. The index sums LLM-assigned geopolitical risk scores (0–1) across all articles each day, normalized by total newspaper article count. The index is normalized to a mean of 100 over 1985–2019. Labels indicate the Cuban Missile Crisis (1962), the Gulf War (1991), September 11 (2001), the Iraq War (2003), and Russia's invasion of Ukraine (2022). Source: authors' calculations using ProQuest TDM Studio data.

Figure 2: AI-GPR vs. Original GPR Index (90-Day Moving Average)



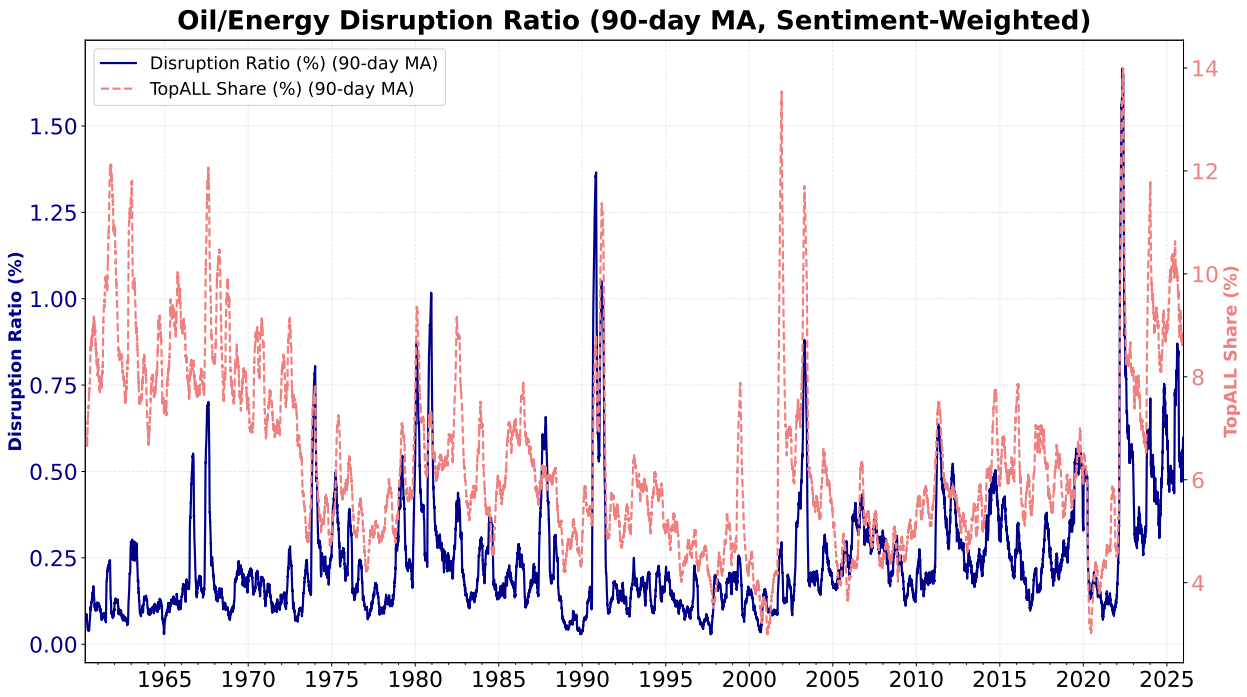
Notes: The AI-GPR index (blue, thick) and the original GPR index from Caldara and Iacoviello (2022) (red, thin), both plotted as 90-day moving averages, 1960–2025. Both indices are normalized to a mean of 100 over 1985–2019. The AI-GPR index sums LLM-assigned geopolitical risk scores across all articles each day, normalized by total newspaper article count. The original GPR counts articles matching specific keyword proximity searches. Source: authors' calculations using ProQuest TDM Studio data.

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



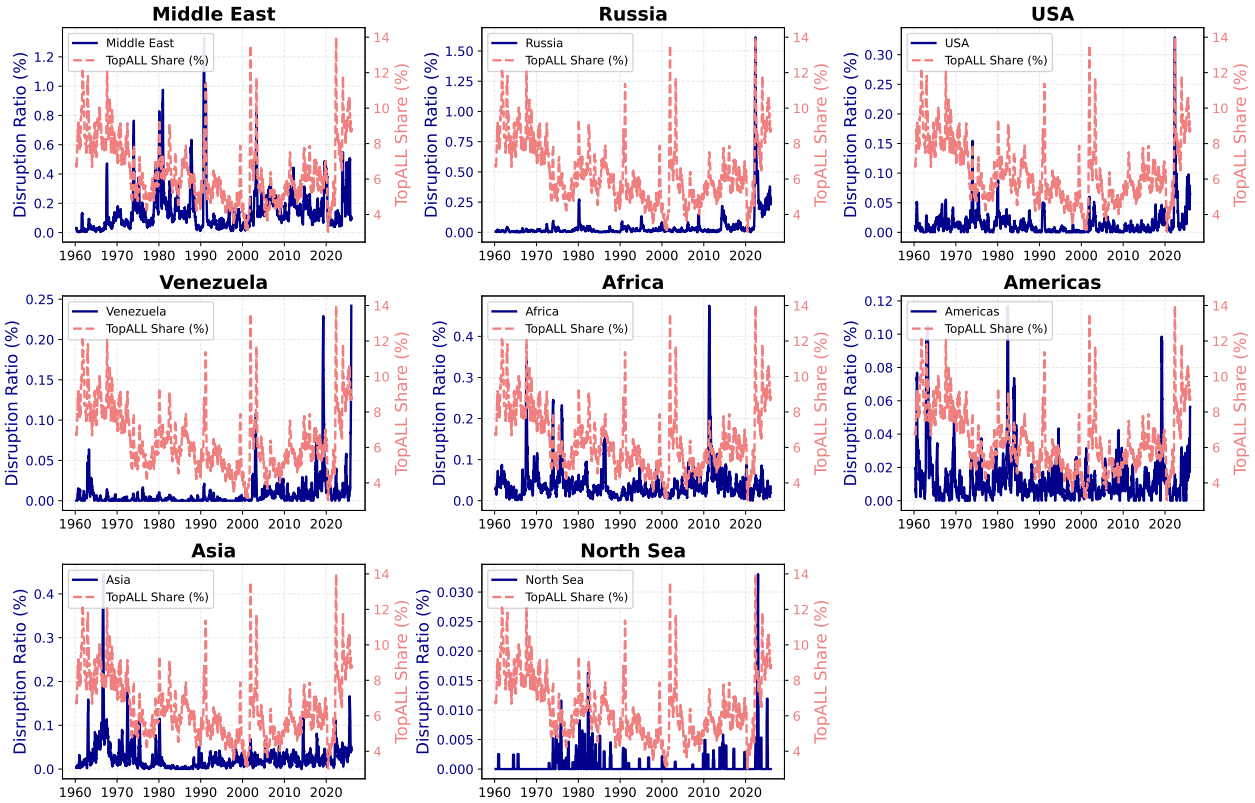
Notes: Each panel shows the daily AI-GPR index (blue, thick) and original GPR index (salmon, thin) around a major geopolitical event. Vertical dashed lines mark the event date. Top row: Cuban Missile Crisis (October 1962), Yom Kippur War (October 1973), Kuwait Invasion (August 1990). Bottom row: Gulf War (January 1991), September 11 Attacks (September 2001), Russia-Ukraine Invasion (February 2022). Both indices are at daily frequency (no smoothing) and normalized to a mean of 100 over 1985–2019. Source: authors' calculations.

Figure 4: GPR-Driven Oil Supply Disruption Index



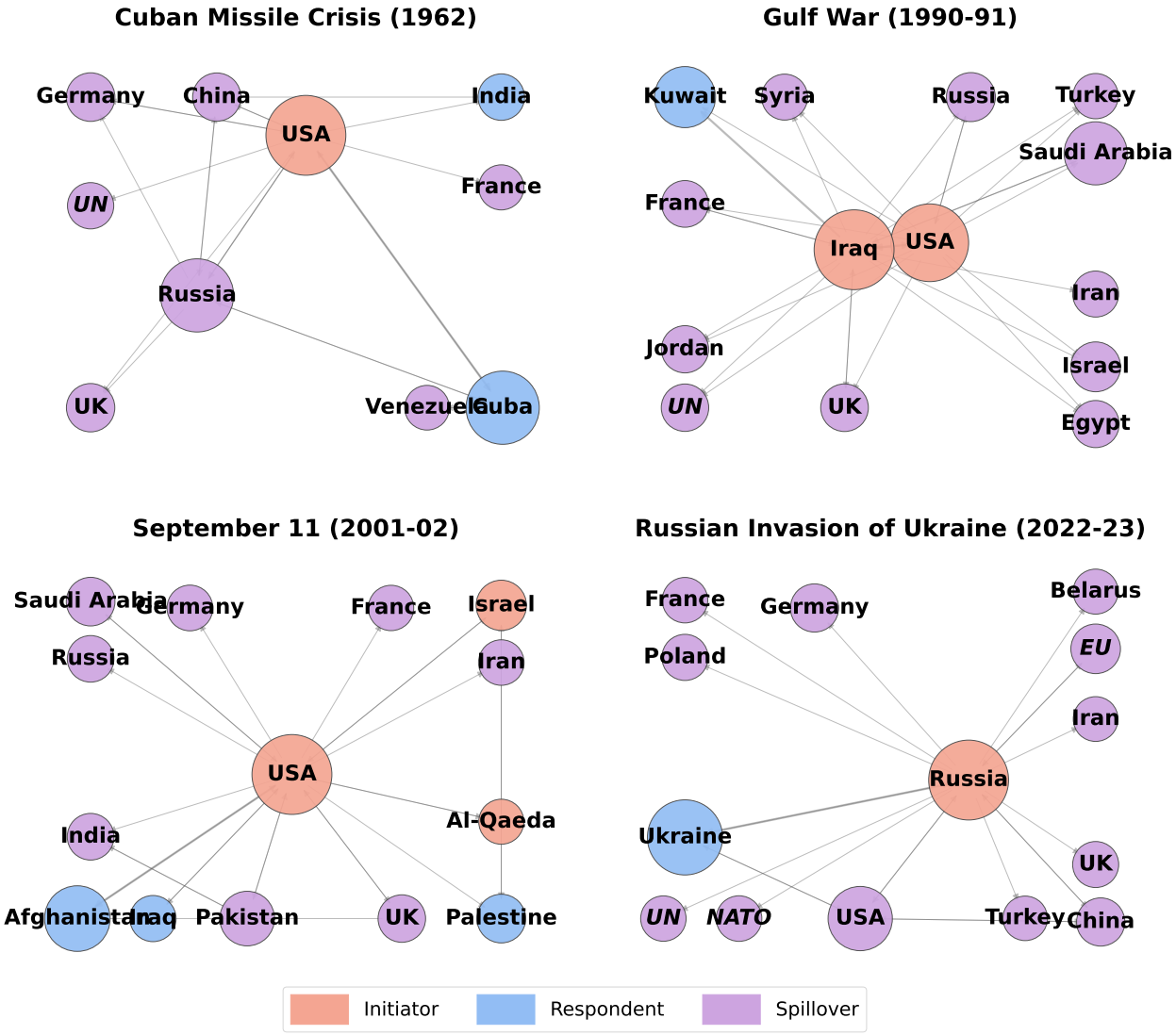
Notes: The dark blue line (left axis) shows the Oil-GPR index defined in equation (5): the daily ratio of sentiment-weighted oil disruption articles to total newspaper articles, plotted as a 90-day moving average. Each disruption article is weighted by its GPR score (0–1). The light red dashed line (right axis) shows the AI-GPR index as a share of total articles for comparison. Major peaks correspond to the 1973 Arab oil embargo, the 1979 Iranian Revolution, the 1990–91 Gulf War, and the 2022 Russia-Ukraine conflict. Source: authors' calculations.

Figure 5: GPR-Driven Oil Disruptions by Region



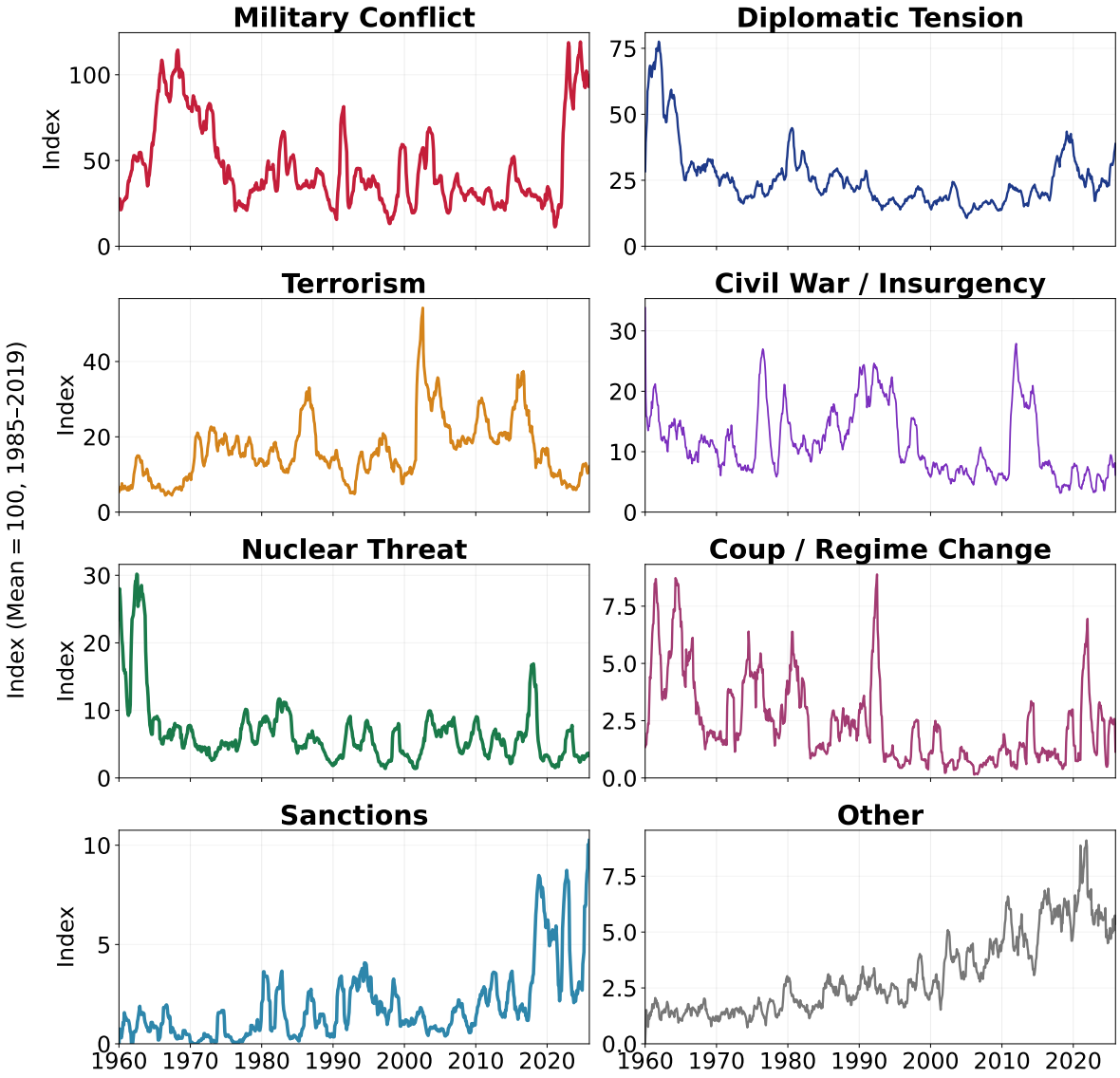
Notes: Each panel shows the regional Oil-GPR index (90-day moving average) for a specific geographic region. The index is computed as the ratio of sentiment-weighted articles about oil disruptions in that region to total newspaper articles. The light red dashed line in each panel shows the overall AI-GPR 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 4.2.

Figure 6: Geopolitical Risk Networks Around Major Events



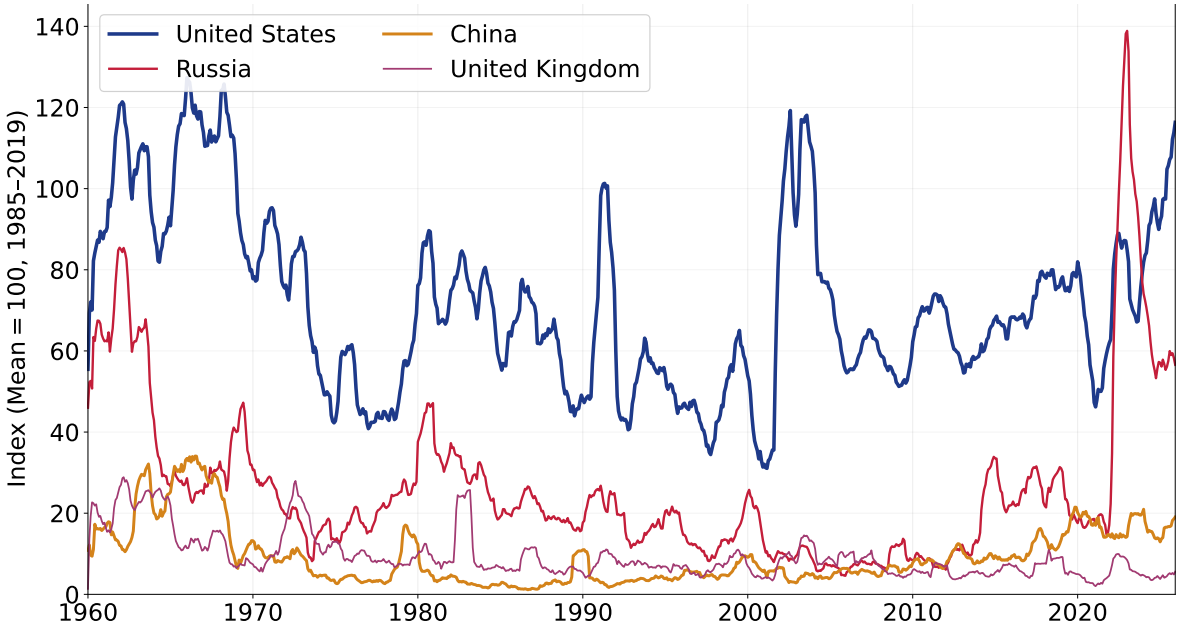
Notes: Directed network graphs of geopolitical actors around four major episodes. Each node represents a country (or supra-national body). Node color indicates the dominant role assigned to that country across articles in the episode: initiator (red), respondent (blue), or spillover (purple). Node size is proportional to GPR-weighted involvement (the sum of GPR scores across all articles mentioning the actor). Edge width reflects co-occurrence intensity between actor pairs, with directed arrows running from initiators to respondents. Up to 14 top nodes by total involvement are shown per panel. Source: authors' calculations using two-stage LLM classification on articles with GPR score > 0.5.

Figure 7: AI-GPR by Event Type (12-Month Moving Average)



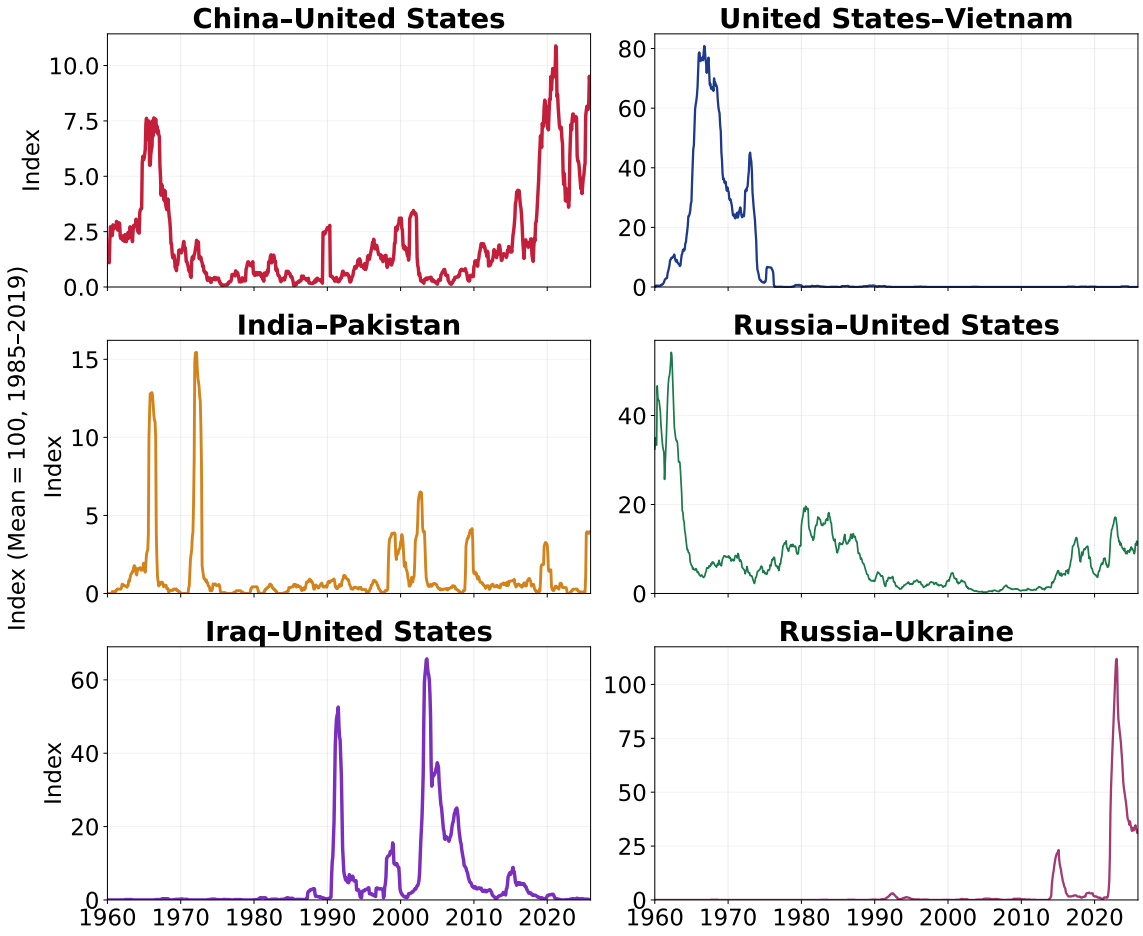
Notes: Each panel shows the monthly AI-GPR sub-index for one event category, smoothed with a 12-month moving average. Sub-indices are scaled to the overall AI-GPR index (mean = 100, 1985–2019). Event categories are assigned by the second-stage LLM classifier applied to articles with GPR score > 0.5. Source: authors' calculations.

Figure 8: AI-GPR by Country (12-Month Moving Average)



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 GPR score > 0.5.

Figure 9: AI-GPR Bilateral Index: Top Country Pairs (12-Month Moving Average)



Notes: Each panel shows the monthly bilateral AI-GPR sub-index for one country pair, smoothed with a 12-month moving average. The bilateral index aggregates AI-GPR sentiment scores from articles in which both countries appear as geopolitical actors (initiator or respondent), treating the pair symmetrically across directions. Pairs are ranked by cumulative GPR intensity over 1960–2025. 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 GPR score > 0.5.

A Technical Appendix

A.1 Alternative Model Specifications

We classify a random sample of 1,050 articles using four OpenAI models—GPT-4o-mini, GPT-4o, GPT-5-mini, and GPT-5—all with the same prompt at temperature zero. 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
GPT-4o-mini	1.00	0.90	0.87	0.86
GPT-4o		1.00	0.86	0.87
GPT-5-mini			1.00	0.94
GPT-5				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.

All pairwise correlations exceed 0.86, 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 LLMs and across model sizes within the same family.

A.2 Keyword-Based GPR Index

As described in Section 3, we construct a keyword-based GPR index using the Caldara and Iacoviello (2022) methodology on the same three newspapers and time period as the AI-GPR. The search terms are organized into eight categories, each combining a geopolitical term group with a risk/threat term group using the N/2 proximity operator (words must appear within 2 words of each other). The full search string is:

```
GPR ORIGINAL 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*’’
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.