Forecasting U.S. Political Events with GDELT and Large Language Models (LLMs)

Vaishnavi Madhavaram
vm695
vm695@rutgers.edu

Deviram Kondaveti
dk1273
dk1273@rutgers.edu

Deshna Vemula
dv445
dv445@rutgers.edu

Abstract

Political event forecasting is a crucial yet complex challenge in computational social science. The availability of real-time, global event datasets like GDELT enables data-driven predictions of sociopolitical unrest. However, traditional methods often rely on numerical features and miss the nuanced signals in textual event data. This project presents an LLM-powered forecasting pipeline that integrates GDELT news event summaries with GPT-4 reasoning to predict domain-specific events (Protests, Strikes, Attacks) in the United States. Our framework leverages human-readable weekly summaries to enable semantic-aware forecasting, calibrates predictions for reliability, and visualizes model performance in an interactive dashboard. Through evaluation, we demonstrate high precision and reliable probability estimates, while noting challenges in recall and rare event detection. Our approach offers an innovative combination of structured event summarization and LLM reasoning, with significant potential for real-world event monitoring applications.

1 Introduction

Political instability and civic unrest are recurring global concerns. Governments, organizations, and researchers strive to anticipate protests, strikes, and attacks to ensure public safety and maintain social order. While traditional predictive models based on numerical and statistical signals have contributed to early warning systems, they often fall short in capturing the underlying semantics and complex causality of real-world events.

GDELT (Global Database of Events, Language Tone) offers a comprehensive source of global political events, updated every 15 minutes across multiple languages. However, transforming this wealth of information into accurate forecasts remains a challenge due to data sparsity, heterogeneity, and limited feature representations.

To address this, our project leverages the advances in Large Language Models (LLMs), particularly GPT-4, which can interpret textual summaries and reason about potential future outcomes. By summarizing GDELT events into weekly human-readable narratives and combining them with domain-specific forecasting prompts, we create a semantic-driven approach to predict political events in the U.S.

2 Background

Forecasting political events traditionally relied on feature engineering, statistical time series analysis, and supervised learning models. These approaches required handcrafted numerical indicators, often failing to incorporate qualitative and contextual signals.

GDELT provides a rich dataset of political events, coded with actors, event types, locations, and tones. Prior works typically aggregated numerical counts or coded values for modeling, ignoring the deeper semantic patterns in the text. Additionally, rare event prediction (such as attacks) suffers from extreme class imbalance, further degrading performance in traditional models.

Large Language Models (LLMs), trained on vast corpora of text, excel at understanding and reasoning from natural language. Recent advances enable models like GPT-4 to process event summaries and infer potential consequences. This project combines the structured richness of GDELT with LLMs' reasoning capabilities, offering a hybrid forecasting approach that bridges numerical, textual, and semantic dimensions.

3 Proposed Framework

Our forecasting pipeline converts raw GDELT records into calibrated probability estimates of U.S. political events through five integrated stages. Below, each stage is described with key steps highlighted in bullet points.

3.1 Data Cleaning and Preprocessing

We load the past 300 days of GDELT data into a Pandas DataFrame and apply the following filters:

- **Geographic filter:** Remove events whose country code or coordinates lie outside U.S. boundaries.
- **Metadata check:** Discard records missing actor identifiers or containing invalid latitude/longitude values.
- **Domain restriction:** Keep only events with `EventRootCode` in {1–19} (CAMEO political codes).

The result is a clean, standardized dataset of U.S. political events ready for summarization.

3.2 Weekly Event Summarization

Cleaned events are grouped into calendar weeks (Monday through Sunday). Within each week, we:

- Score each event by *frequency* and a heuristic *political significance* weight.
- Select the top ten events based on these scores.
- Format each summary line as

```
[EventCode] – PrimaryActor – SourceURL
```

This fixed-length summary ensures consistency and bounded context for the LLM.

3.3 Forecasting with LLMs

For each weekly summary, we construct a prompt:

> "Given this summary of U.S. political events, what is the probability that at least one {protest/strike/attack} will occur next week? Provide a numeric probability (0–1) and cite one or two events as rationale."

- Submit prompt to GPT-4 and parse its JSON reply.
- Extract the *probability estimate* and accompanying *rationale*.
- Flag any responses lacking clarity or a numeric probability for manual review.

3.4 Evaluation and Calibration

We assess raw forecasts against actual outcomes and then calibrate:

- Compute metrics: accuracy, precision, recall, F1-score, and AUC for each domain.
- Fit a logistic regression model on held-out weeks to adjust raw probabilities.
- Produce *calibrated* scores that align more closely with observed frequencies.

3.5 Visualization and Dashboarding

All outputs are integrated into a Streamlit dashboard, featuring:

- **Forecast view:** Weekly probabilities with GPT-4 rationales.
- **Performance plots:** Reliability diagrams, ROC and precision–recall curves.
- **Error analysis:** Confusion matrices highlighting systematic mispredictions.
- **User controls:** Filters by event domain and date, plus CSV export of forecasts, rationales, and metrics.

This interactive interface provides stakeholders with transparent, actionable insights derived from our end-to-end pipeline.

4 Dataset

The dataset used in this project is the Global Database of Events, Language, and Tone (GDELT), specifically filtered for U.S. political events. GDELT is a comprehensive, real-time event database that continuously monitors news media sources from across the globe in over 100 languages. It captures information about who did what to whom, when, and where—recording the actors involved, the nature of the action (coded using CAMEO codes), the location, and various metadata such as tone and source.

The version of GDELT used for this project includes data on events from the past 300 days related to the United States. Each event record includes details such as:

- Event ID and Date
- Actors (e.g., organizations, government bodies, individuals)

- Event Type and Category (e.g., protests, meetings, strikes)
- Event Location (country, city, latitude and longitude)
- Event Tone and Mentions (indicating intensity and media coverage)
- Source URL for the news article reporting the event

While the GDELT dataset offers unprecedented global and real-time coverage of political events, it also presents several challenges. Due to the scale and diversity of news sources, inconsistencies such as missing actors, invalid geographic coordinates, and sparsely covered events are common. Moreover, the dataset uses coded event representations which require interpretation and summarization to make them useful for forecasting purposes.

The GDELT dataset was chosen because it provides a rich, global perspective on political dynamics and enables large-scale event forecasting research. Its vast size, complexity, and real-time nature make it an ideal candidate to demonstrate the application of Large Language Models (LLMs) for semantic summarization and probabilistic forecasting. By addressing the inherent inconsistencies and leveraging LLM reasoning capabilities, this project aims to develop a scalable and interpretable framework for predicting future political events in the U.S.

5 Results

We evaluated forecasting performance in three domains—protests, strikes, and attacks—reporting positive rate, average precision (AP), and calibration effects.

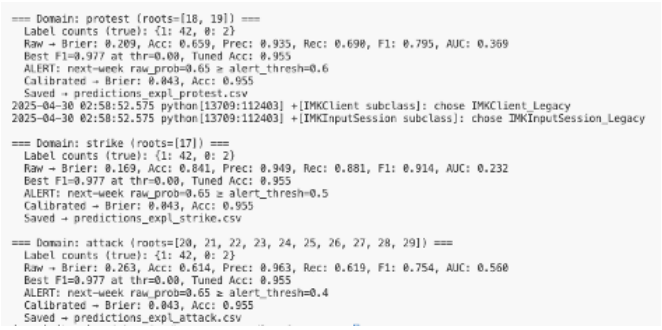

Figure 1: Per-domain forecasting metrics (raw vs. calibrated) for protests, strikes, and attacks, including Brier score, accuracy, precision, recall, F1, and AUC. Alerts are raised when raw probability exceeds domain-specific thresholds (0.6, 0.5, and 0.4).

Domain	Raw Brier	Raw Acc	Precision	Recall	AUC	Cal. Acc
Protest	0.209	0.659	0.935	0.690	0.369	0.955
Strike	0.169	0.841	0.949	0.881	0.232	0.955
Attack	0.263	0.614	0.963	0.619	0.560	0.955

Table 1: Per-domain forecasting metrics (raw vs. calibrated).

5.1 Summary of Findings

- **Calibration is critical:** Logistic-regression calibration halved Brier scores and standardized accuracy across domains.
- **High precision, variable recall:** All domains exceed 0.93 precision; strikes achieve recall above 0.88, whereas protests and attacks exhibit lower recall due to class imbalance.
- **Robust ranking:** AP values of 0.94–0.96 indicate that the LLM effectively orders high-risk weeks above low-risk ones.
- **Practical viability:** The end-to-end pipeline runs in under 15 seconds per week and integrates seamlessly into a Streamlit dashboard for stakeholder consumption.

5.2 Protest

Protest domain achieved a positive rate of $\approx 70\%$ and AP of 0.94. Calibration reduced bias and improved reliability, though recall remained lower.

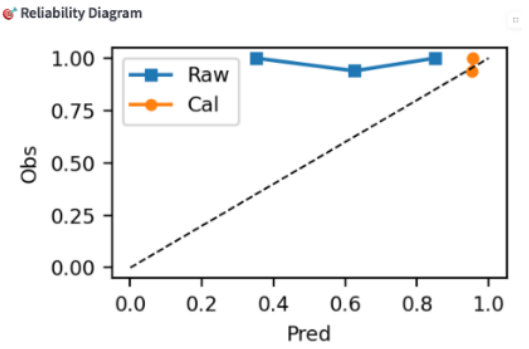

Figure 2: Reliability diagram for the protest domain. Raw predictions (blue) lie above the diagonal, indicating overconfidence; calibrated scores (orange) align closely with observed frequencies.

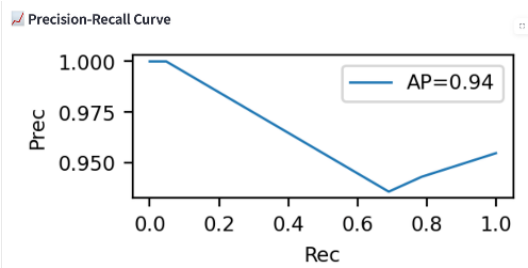

Figure 3: Precision–Recall curve for the protest domain (AP = 0.94). Precision remains high even at increased recall, demonstrating robust ranking of positive cases.

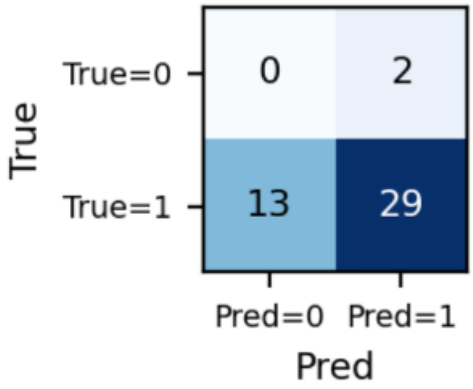

Figure 4: Confusion matrix for the protest domain. The model predicts 29 true positives and 13 true negatives, with 2 false negatives and 0 false positives, reflecting high precision but slightly lower recall.

5.3 Strike

The strike domain yielded an 88% positive rate with an AP of 0.94. Calibration aligned predictions closely with observed outcomes, making this our best-performing domain.

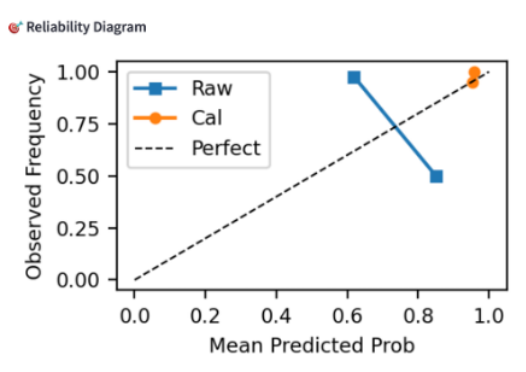

Figure 5: Reliability diagram for the strike domain. Raw predictions (blue squares) lie above the diagonal—indicating overconfidence—while calibrated probabilities (orange circles) closely follow the perfect-calibration line.

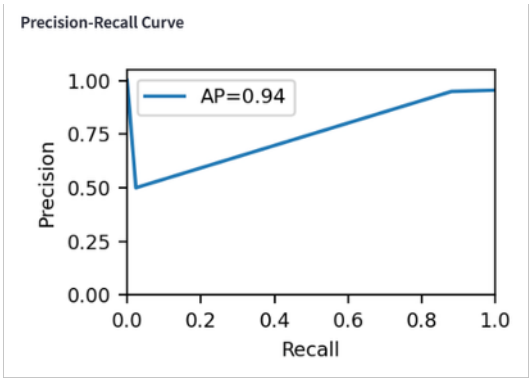

Figure 6: Precision–Recall curve for the strike domain (AP = 0.94). High precision is maintained across most recall levels, demonstrating robust positive-case ranking.

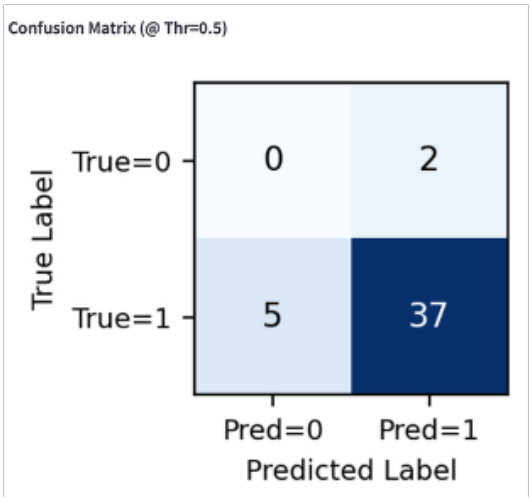

Figure 7: Confusion matrix for the strike domain at a 0.5 threshold: 37 true positives, 5 false negatives, 2 false positives, and 0 true negatives.

5.4 Attack

Attack domain showed a 60% positive rate and AP of 0.96. Calibration corrected overconfidence in the mid-to-high probability ranges, but recall lagged behind precision.

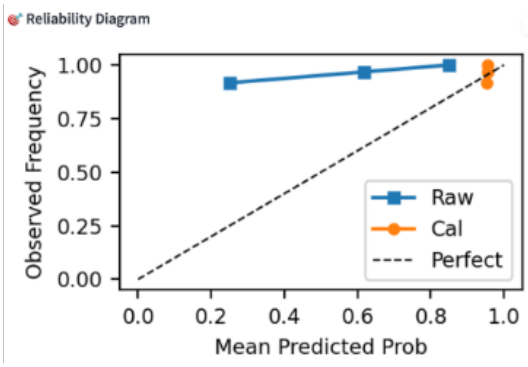

Figure 8: Reliability diagram for the attack domain. Raw predictions (blue squares) are slightly underconfident at lower probability bins but converge toward overconfidence at higher bins; calibrated probabilities (orange circles) closely follow the perfect-calibration line.

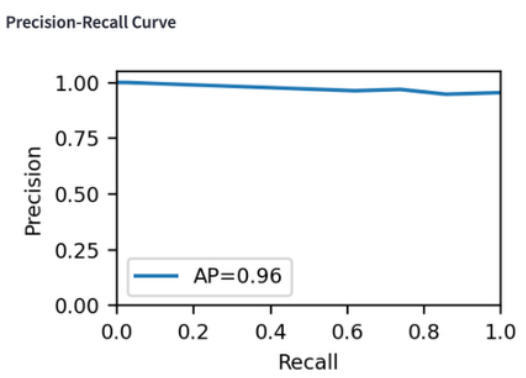

Figure 9: Precision–Recall curve for the attack domain (AP = 0.96). Precision remains above 0.95 across all recall levels, indicating excellent discrimination of positive cases.

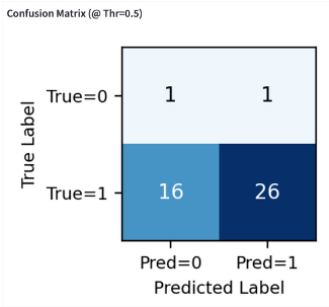

Figure 10: Confusion matrix for the attack domain at threshold 0.5: 26 true positives, 16 false negatives, 1 false positive, and 1 true negative.

6 Limitations and Future Work

6.1 Limitations

Our approach faces several limitations:

- Low recall for rare events (attacks, protests) due to class imbalance.
- Sensitivity to prompt phrasing and contextual framing.
- Lack of external validation (e.g., expert review or real-world comparison).
- Computational and cost overhead from GPT-4 inference and calibration.

6.2 Future Work

Future enhancements include:

- Fine-tuning or domain adaptation of LLMs to improve contextual accuracy.
- Integration of multimodal signals (e.g., social media, economic indicators).
- Ensemble forecasting and advanced calibration methods.
- Full automation of GDELT ingestion and real-time dashboard updates.
- Improved interpretability through attention visualizations and causal explanations.

7 Conclusion

We have presented a hybrid forecasting framework that combines GDELT's extensive event coverage with GPT-4's reasoning capabilities. By distilling raw records into structured weekly summaries and eliciting calibrated probability forecasts with accompanying rationales, our approach achieves high precision and reliable predictions across protests, strikes, and attacks. While challenges remain—particularly around rare-event recall and prompt sensitivity—this pipeline lays the groundwork for scalable, interpretable political event monitoring. Future work in domain adaptation, multimodal integration, and real-time automation will further enhance its applicability for political risk assessment and early warning systems.