GENOME: A New Geopolitical Event Methodology and Dataset using Large Language Models

Alessandro Dell'Orto* **Jesse Kommandeur***
The Hague Centre for Strategic Studies
alessandrodell'orto@hcss.nl jessekommandeur@hcss.nl

Abstract

Quantitative research in international relations relies heavily on structured event data, yet existing automated datasets lack up-to-date coverage of both conflictual and cooperative interactions. We introduce GENOME (Geopolitical Event News Observatory, Mapping, and Extraction), an automatically extracted dataset that implements PLOVER's 16 event types and extends its Actor–Recipient schema with a Third Party role to capture multilateral relations from newswire data. GENOME's pipeline comprises event extraction, ontology-based classification, entity normalization, and deduplication, leveraging GPT models with one-shot prompting and enforced structured outputs. We compare GENOME against POLECAT dataset over a five-month overlap period across event volume, temporal dynamics, and geographical coverage. Results show that while the two datasets align closely on conflict event types, GENOME captures a more balanced distribution of cooperative events, particularly verbal interactions nearly absent in POLECAT. GENOME also demonstrates improved temporal precision by attributing events to their inferred date of occurrence rather than publication date, and effective deduplication of highly covered events.

1 Introduction

Collecting, coding and analysing geopolitical events has been a pivotal challenge for quantitative studies in International Relations since the 1960s (McClelland, 1961; Yonamine, 2016). Several datasets have been compiled over the following decades, where the main trade-off has been the balance between cost and reliability of manually annotated datasets or the size and rapidity of automated solutions. Most widely diffused manually coded ones are either historic and no longer maintained, exclusively conflict-focused or limited to specific domains, regions or diplomatic processes (Raleigh et al., 2023; Olsen et al., 2024).

On the other hand, automatically extracted datasets are based on event ontologies, rulebooks that provide a schema for representing events in a structured, machine-readable way. Numerous ontologies have been developed since the late 1970s (McClelland, 1978; Azar, 1980; Bond et al., 1994; King and Lowe, 2003b), giving rise to a range of automatically coded datasets (King and Lowe, 2003a; O'Brien, 2010; Jenkins et al., 2012; Leetaru and Schrodt, 2013; Salam et al., 2020). Most notably, CAMEO (Gerner et al., 2002) served as the foundation for ICEWS, GDELT, and Phoenix, and was succeeded by the PLOVER ontology, redesigned for greater simplicity, flexibility, and compatibility with machine learning approaches (Open Event Data Alliance, 2024). In 2023, POLECAT was introduced as an automatically coded dataset with global coverage of both conflict and cooperation political events (Halterman et al., 2023a), based on PLOVER, using the NGEC coder which employs transformer-based models to extract events from multilingual news sources (Halterman et al., 2023b). Around August 2024, POLECAT has stopped receiving updates, leaving a gap for up-to-date automatically extracted events covering both conflictual and cooperative interactions. Another relevant issue is that these ontologies lack publicly available gold datasets to test Event Extraction (EE) tasks on them, forcing scholars to adapt their methods to other event structures where such gold standards exist, such as ACE05 (Halterman et al., 2023b; Gao et al., 2023; Brandt and Sianan, 2025).

In recent years, Large Language Models (LLMs) have captured significant attention for socio-political EE, though zero-shot approaches often fall short for rigorous codebook-based measurement (Cai and O'Connor, 2023; Chen et al., 2024; Halterman and Keith, 2025). Earlier work on zero-shot ranking for socio-political texts found that re-

* Equal contribution.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 83–95

trieval quality degrades as the target label becomes more general, and that declarative label formulations outperform dictionary definitions (Akdemir and Hürriyetoğlu, 2022). Key challenges include hallucination, output inconsistency, deviation from ontology schemas, and degraded performance on non-Western sources and low-resource languages (Thapa et al., 2025), as well as mismatches between the nuanced, evolving nature of real-world events and the rigid dyadic structures of current ontologies (Brandt and Sianan, 2025). Existing LLM-based EE pipelines and datasets have all been either conflict or region-focused, like the horn of Africa (Bai et al., 2025; Meher and Brandt, 2025; Semnani et al., 2025). However, tracking low-intensity, cooperative, and verbal interactions is valuable for early-warning and conflict prediction, as escalation is best understood as a sequential process in which preceding lower-intensity interactions carry meaningful predictive signal (Halkia et al., 2020; Beardsley et al., 2024).

Given the current gap left by POLECAT in automatically extracted geopolitical events from news data that tackle both conflictual and cooperative, verbal and material inter-state interactions, we introduce GENOME (Geopolitical Event News Observatory, Mapping, and Extraction). The name reflects the dataset's ambition: just as genomics maps the fundamental building blocks of biological life, GENOME aims to map the building blocks of geopolitical interaction through underlying patterns of conflict and cooperation between states. GENOME implements PLOVER's 16 event types and extends its Actor-Recipient schema to capture multilateral relations from newswire data that more closely mirror real-world geopolitical interactions. It incorporates up-to-date newswire articles and a two-phase extraction-classification pipeline leveraging GPT models, along with a series of entity normalization and deduplication techniques.

While GENOME, like most other ontology-based datasets, lacks a manually annotated gold standard, we evaluate it by comparing a 5-month overlapping sample of approximately 28,000 events against POLECAT: even with different input sources, both datasets show similar behaviour in conflict events and temporal dynamics, while differing significantly in how they track cooperation, especially verbal. We also show that GENOME better aligns events with their actual occurrence dates (rather than news publication dates), accurately resolves international entities such as the IMF and NATO, and reduces redundancy among heavily reported events. The dataset and analysis scripts are publicly released.[1] The remainder of this paper describes GENOME's design and ontology changes (Section 2), details our methodology (Section 3), presents our experimental setup and results (Sections 4–5), and discusses findings and limitations (Section 6).

2 Concept and Design

2.1 Sources

GENOME receives as input a corpus of English-language newswire articles collected from a commercial source, primarily covering international affairs and including a publication date. The corpus currently spans from January 2024 to January 2026 and contains approximately 148,000 articles from around 2,400 unique sources, of which roughly 30,000 articles fall within the February–June 2024 period used for comparison (Sections 4–5).

2.2 Ontology Changes

GENOME events are based on a simplified and slightly modified version of the PLOVER ontology. They are *simplified* because the system only focuses on the "what" of the "What–How–Why" logic from PLOVER. It does this by classifying events by their Event Type, using the same 16 types grouped into four categories based on their Conflict/Cooperation and Material/Verbal nature. Moreover, it follows the same temporal and reality logic (focusing on things that have already happened and are not just being discussed), explicit Actor logic (where Actors must be explicitly named while the Recipient role is optional), and compound logic (where events that involve multiple Actors for a single role generate a single entry).

They are *modified* because we extend PLOVER's Actor–Recipient framework by introducing the overarching concept of an Agent, defined as an individual, organization, country, or social group that is capable of taking action (speaking, moving, fighting) or having a political action directed at them. Within this framework, we identify three Agent roles:

- **Actor:** The Agent that initiates or performs the action described in the event.

[1] https://github.com/HCSS-Data-Lab/Submission-GENOME

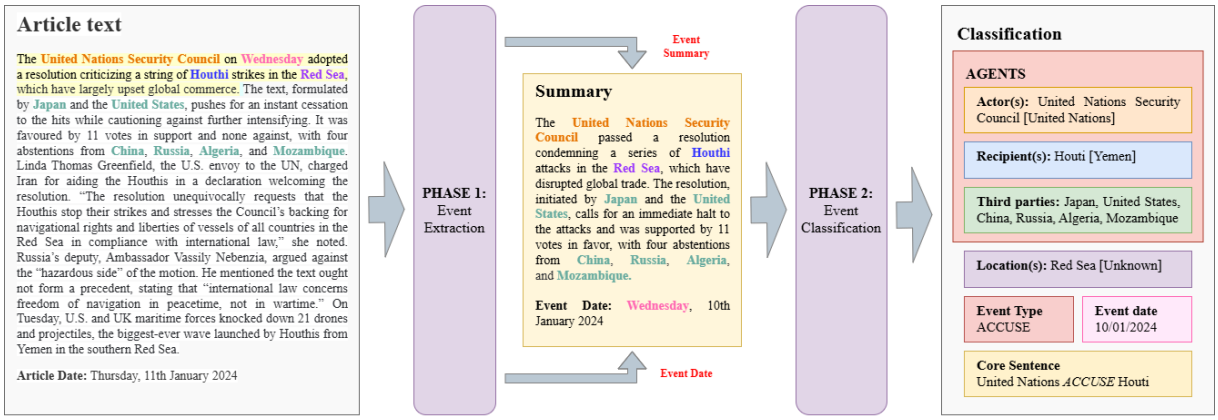

Figure 1: Simplified overview of the event extraction, classification, and normalization modules of the GENOME pipeline.

- **Recipient:** The Agent that is the direct target of the action. This role is optional, as not all events have a clearly defined recipient.
- **Third Party:** An Agent that provides contextual information relevant to the event but is neither the initiator (Actor) nor the target (Recipient). For example, in "A accused B of arming C," A is the Actor and B is the Recipient, while C is a Third Party. The implied action "B arming C" is not treated as a separate event, as it violates the reality logic.

We introduce the Third Party role for two reasons. First, in pilot experiments, we observed that LLMs tend to assign a role to every identified Agent, so it is useful to provide a designated category for context-bearing entities. Second, similar to how POLECAT further classifies Actors and Recipients, Third Parties can also be categorized based on their role (e.g., "facilitator" or "reporter") and nature (e.g., "government" or "military"). This creates a pathway from dyadic actor–recipient representations to richer multi-entity event structures (e.g., hypergraph-style representations) that more explicitly preserve temporal and relational structure (Ahrabian et al., 2025).

3 Methodology

3.1 Pipeline and Model Choice

Our pipeline consists of four modules executed sequentially: (1) extracting event summaries from articles' raw text and inferring the event date; (2) classifying events by applying our version of the PLOVER ontology; (3) normalizing ambiguous names and resolving entities; and (4) deduplicating multiple occurrences of the same event. Figure 1 visualizes a simplified version of the first three modules. For both the extraction and classification phases, our strategy relies on one-shot prompting combined with enforced structured outputs using Pydantic models. During our testing, OpenAI models proved to be the most robust choice, offering the best balance across performance, cost, and native support for both structured output and fine-tuning capabilities.

3.2 Event Extraction

The extraction task is primarily a summarization task, which does not require the advanced reasoning capabilities of frontier LLMs. We deliberately separate event extraction from classification to reduce the complexity of individual tasks, allowing us to utilize more cost-effective, lower-parameter models. Moreover, distilling the raw text into a structured, concise event summary standardizes the input for the classification task, significantly reducing input size and variability (You et al., 2025).

To reduce input tokens and isolate the core news lead, article texts are truncated at the sentence boundary nearest the 200-word mark. This approach is highly effective, but requires the input text to follow the standard newswire format, where the main news is placed at the beginning of the article. `gpt-4o-mini` was chosen as the most cost-effective option for this kind of task.

We inject the article's date in long british format and pre-cleaned text as a prompt to extract event objects. Each object comprises three components. First, an inferred event date (e.g., deriving "14th January 2024" from an article dated "Monday, 15th January 2024" that reports an event "on Sunday"); second, an event description containing all funda-

mental and contextual elements—a structure that mimics the concise event summaries of the ACLED dataset (Raleigh et al., 2023); and third, a source quote representing the verbatim text used to construct the object. The prompt explicitly requests each event description to be fully self-contained and understandable without reference to external information, to articulate the core action of the event in a concise form, and to complement it with the main relevant contextual information. Lastly, the model is tasked to extract the core event of the article, and to only report multiple events if a single "core" event cannot be unequivocally identified. This ensures that extracted events are the actual subject of the newswire article, rather than contextual or historical events mentioned in passing.

3.3 Event Classification

Event objects are expanded by a second LLM call, which classifies them based on our extension of the PLOVER ontology described in Section 2.2. This task, particularly correctly identifying the core event action and assigning each involved party to its Agent role, is more cognitively complex. However, this complexity is reduced by the concise and consistently structured summaries produced in the previous step, which yield a highly consistent format for both input and output.

To reduce cost, we leverage this repetitiveness by fine-tuning `gpt-4.1-nano` on 775 event descriptions classified with `gpt-5.1`, with an 80/20 train/test split, using OpenAI's supervised fine-tuning tool. The process yielded a training loss of 0.114 and a validation loss of 0.122, demonstrating successful model convergence and strong generalization to unseen data with minimal overfitting.

The classification prompt includes the definition of Agent and of Agent roles as they are defined in Section 2.2, along with examples to distinguish recognized entities that are agents (Actor, Recipient, Third Party) from those that are not (e.g., locations such as "the Middle East", or concepts such as "Climate Change"). To facilitate reasoning on roles, we ask the model to report the main action as "A *verb* B", where A and B are the Actor or Recipient and the verb is the event type from the PLOVER ontology (Figure 1). Finally, the model maps all entities to their corresponding country, which serves as the basis for the normalization step.

3.4 Entity Normalization

Entity normalization proceeds in two stages. First, extracted names of countries and international bodies are matched via fuzzy search against a manually curated reference list of 248 countries and 40 organizations, along with a series of synonyms. Unmatched entries are resolved manually. This process successfully resolved all inferred country names, reducing the count of unique countries and international organizations by 81.7%.

Second, agent names are grouped by their normalized country or organization. These names are converted into vector embeddings using a Sentence-Transformer model and clustered based on semantic similarity using FAISS. The most frequent entry in each cluster is selected as the canonical normalized name. Consequently, the volume of unique agent names decreased by 35.4% to 32,624 entries.

3.5 Deduplication

To consolidate duplicates extracted from multiple sources, we implement a multi-criteria deduplication pipeline. Candidate duplicate events are first filtered by temporal proximity within a 2-day window, then scored using a weighted composite of four similarity components: semantic similarity of event summary embeddings using `all-MiniLM-L6-v2` (0.45), actor and recipient name overlap via Jaccard similarity (0.25), event type match (0.20), and location overlap (0.10). Pairs exceeding a composite threshold of 0.80 are linked as duplicates. Duplicate clusters are resolved via connected components on the resulting similarity graph, merging all contributing article identifiers.

3.6 Operational Cost

Across the full two-year corpus (roughly 6,000 newswire articles ingested per month), total OpenAI API spend was approximately $55, comprising $19 for extraction with `gpt-4o-mini`, $33 for classification with the fine-tuned `gpt-4.1-nano`, and a one-off $4 fine-tuning job, all using the batch API at a 50% discount on standard rates. This corresponds to about $2.30 per month, or on the order of $0.0004 per processed article. Because fine-tuning is performed once on a silver-standard sample, its cost is amortized over all subsequent extractions and does not scale with corpus size.

4 Experimental Setup

Evaluating GENOME presents inherent challenges common to automated event extraction: no manually annotated gold standard exists for the PLOVER ontology, and direct record-level alignment with POLECAT is infeasible due to differences in input data and GENOME's extended Actor–Recipient schema introduced in Section 2.2. Consequently, the next best alternative and the most robust evaluation feasible under these constraints is to assess how similarly GENOME behaves with respect to POLECAT and known real-world geopolitical events. We therefore adopt a systematic indirect comparison against POLECAT of multiple versions of the GENOME dataset and a cross-comparison between those versions, complemented by a case study to validate the datasets against known occurrences.

4.1 Dataset Variants

We evaluate our pipeline against the publicly accessible version of POLECAT (Scarborough et al., 2023). All quantitative comparisons are restricted to the five-month overlap period between the datasets (February to June 2024). We compare POLECAT against three GENOME configurations:

- **GENOME (article_date)**: The output of our pipeline after the normalization step (Section 3.4), timestamping events using the *publication date* of the source news article.
- **GENOME (event_date)**: The same table as GENOME (article_date), but incorporating the *event date* extracted by the model during the extraction phase (Section 3.2).
- **GENOME (dedup)**: The output of the pipeline after the deduplication step (Section 3.5), timestamping events with the same extracted *event date* as GENOME (event_date).

While POLECAT includes an analogous extracted event date, we present evidence in Sections 5 and 6 suggesting that its dates more closely reflect publication timing rather than actual occurrences. Lastly, in subsequent sections, when the specific date choice is irrelevant to an evaluation metric, we refer to the non-deduplicated versions collectively as GENOME, and the deduplicated version as GENOME (dedup).

4.2 Approach and Metrics

We structure our evaluation around four complementary objectives. First, we assess event volume and balance, comparing the total number of events recorded by each dataset and their distribution across the four PLOVER quad categories (verbal/material $\times$ conflict/cooperation) and across PLOVER's 16 individual event types. This reveals whether the two systems, despite different input sources, agree on how geopolitical activity is categorized under a shared ontology.

Second, we examine temporal dynamics. Using daily event counts, we compute summary statistics including means, standard deviations, and the coefficient of variation (CV) to quantify volatility. We compare weekday and weekend means through a weekend/weekday ratio (W/D) to assess sensitivity to media publication cycles. We also analyse the distribution of the lag between article publication date and inferred event date, and we measure the Pearson correlation between daily event volume and deduplication rate (defined as the proportion of events removed on a given day) to assess whether deduplication preferentially targets high-volume days. To formally quantify cross-dataset agreement, we compute Pearson correlations between POLECAT and each GENOME variant on daily event counts, both overall and stratified by PLOVER quad category.

Third, we evaluate geographical coverage and entity structure. We compare the sets of unique country dyads across datasets. For both POLECAT and GENOME, we treat a country dyad as the set of events in which both countries appear among the Actor or Recipient Agent roles. We examine the overlap and composition of shared and exclusive dyads, particularly in terms of event types and international organization resolution. At the country level, we quantify the volume gap and identify where surpluses concentrate. We also report the average number of actors, recipients, and third-party countries per event to explore structural differences.

Fourth, we conduct dyad-level case studies against known real-world occurrences. The primary case study examines Israeli–Iranian dyadic interactions between March 25 and April 25, 2024. This period contains three major, heavily covered events: the Israeli airstrike on the Iranian embassy complex in Damascus (April 1), the Iranian retaliatory strikes on Israel (April 13), and the Israeli

attack on Iranian military sites (April 19). With the same criteria we isolate the 14 dyads with at least 100 events in both POLECAT and GENOME, for which we compute the same daily-count correlation by quad category. The first case study serves as a qualitative check on temporal precision and deduplication effects against known real-world occurrences, while the 14-dyad correlation sweep tests whether aggregate agreement persists once geographic mix is controlled for.

5 Results

5.1 Event Volume and Balance

Volume: Figure 2 presents the distribution of event types across datasets. POLECAT records substantially higher event volume than GENOME (157,328 vs. 28,530 events), with a ratio of approximately 5.5:1, despite drawing from far fewer sources (242 vs. 1,036). GENOME (dedup) reduces the number of events to 16,795 (60% of the GENOME initial volume), bringing the ratio with POLECAT above 9:1 (full counts in Table 4 in the Appendix).

Macro proportions: POLECAT is markedly more conflict-heavy (77.7% conflict vs. 22.3% cooperation), whereas GENOME presents a more balanced distribution (56.3% conflict vs. 43.7% cooperation), with the deduplicated version showing an even more balanced split. Furthermore, the composition of these interactions differs significantly: POLECAT's cooperation is primarily material (19.4%) rather than verbal (2.9%), while GENOME displays the inverse, heavily favouring verbal cooperation (34.2%) over material actions (8.5%).

Event-Type proportions: Conflict event types are distributed similarly across both datasets. Cooperation, however, diverges sharply. In verbal cooperation, POLECAT is almost entirely dominated by CONCEDE (94.1%), with no AGREE or SUPPORT events recorded, while GENOME distributes most of its verbal cooperation across CONSULT (66.6%) and AGREE (24.9%). Material cooperation also differs structurally, though both datasets agree on AID as the dominant type.

5.2 Temporal Dynamics

Daily flow: POLECAT exhibits the highest volatility (CV 40.9%), driven by a sharp weekend drop (W/D 0.321) that pulls its mean well below the median (Table 1). GENOME (dedup) shows the most stable flow (CV 30.1%) but a wider weekday–weekend gap than the non-deduplicated versions, while GENOME (event_date) is notably more volatile than GENOME (article_date).

Metric	POLECAT	GENOME (art.)	GENOME (evt.)	GENOME (dedup)
Mean	1048.85	190.20	190.20	111.97
Median	1235.0	188.0	184.5	116.0
Std	429.34	62.93	75.80	33.67
CV (%)	40.9	33.1	39.9	30.1
Wkday	1302.24	206.40	206.55	124.28
Wkend	418.33	149.88	149.51	81.33
W/D	0.321	0.726	0.724	0.654

Table 1: Events per day summary statistics (Feb–Jun 2024). Wkday = mean weekday count, Wkend = mean weekend count, W/D = weekend/weekday ratio.

Δ (days)	Count	% of Total
< 0	1,247	4.4
$= 0$	14,933	52.3
1	8,946	31.4
2	1,440	5.0
3–5	1,133	4.0
6–10	529	1.9
> 10	302	1.1

Table 2: Distribution of event date vs. article date lag (Δ) for GENOME prior to deduplication (Feb–Jun 2024). Total events: 28,530.

Event vs. Article Date: GENOME's inferred event dates follow expected behaviour: most events (52.3%) fall on the article publication date, with shares declining steadily as lag increases, and only 4.4% show a negative difference; we attribute this to hallucination, time-zone mismatches, or scheduled future events (e.g., "The meeting will start on Thursday") (Table 2). This is clear in the Israel–Iran case study (Figure 3), where GENOME (event_date) produces sharp spikes on the exact dates of the three major events, while both POLECAT and GENOME (article_date) show a visible, similar delay.

Deduplication effects: The daily volume of non-deduplicated GENOME events and the deduplication rate are significantly correlated ($r = 0.57$, $p < 0.001$), consistent with high-volume days containing more duplicate reports. The Israel–Iran case study (Figure 3) shows the same pattern.

Aggregate correlation: Looking beyond the Iran–Israel case, daily Pearson correlations between POLECAT and the three GENOME variants over the full 150-day window are $+0.40$ (article_date), $+0.30$ (event_date), and $+0.58$ (dedup),

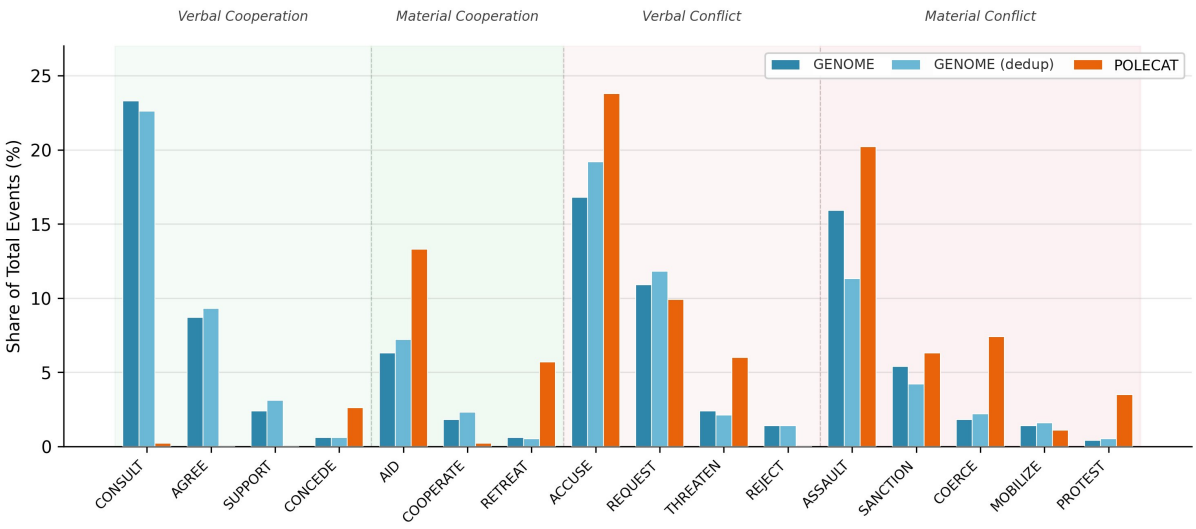

Figure 2: Distribution of event types by share of total events (%) for GENOME, GENOME (dedup), and POLECAT (Feb–Jun 2024). Background shading distinguishes the four PLOVER quad categories. Conflict event types show close alignment between datasets, while cooperation—especially verbal—diverges sharply.

Quad category	article	event	dedup
Overall	+0.40	+0.30	+0.58
Verbal coop.	+0.41	+0.37	+0.39
Material coop.	+0.25	+0.28	+0.43
Verbal conf.	+0.41	+0.33	+0.51
Material conf.	+0.09	−0.00	+0.20

Table 3: Daily-count Pearson r between POLECAT and each GENOME variant by PLOVER quad category (Feb–Jun 2024).

with deduplication strengthening agreement in all four PLOVER quads (Table 3). Despite their similar event-type shares, material conflict shows the weakest aggregate daily-count agreement across all three variants (Pearson $r = +0.09$ article_date, $r = -0.00$ event_date, $r = +0.20$ dedup).

Dyad-level correlation: Restricting to the 14 dyads with at least 100 events in both GENOME and POLECAT, material conflict correlation rises sharply: positive in 13 of 14 dyads with article_date r ranging from $+0.12$ to $+0.81$ (Iran–Israel). The sole exception, Ukraine–United States, is also the only dyad where cooperation dominates (verbal cooperation $r = +0.60$). Across all quad categories, daily correlation with POLECAT is highest under article_date in 8 of 14 dyads, confirming POLECAT's bias toward publication cycles (full per-dyad breakdown in Appendix Table 5).

Day-of-week analysis: The overall weekend/weekday ratio remains stable across GENOME date versions (Table 1), but switching to event dates shifts the composition: weekend material conflict rises by 8.2%, while verbal and cooperative events decline. Across all GENOME versions, material events are less sensitive to day-of-week effects than verbal ones, which drop visibly on weekends (Figure 4). POLECAT, by contrast, shows a uniform drop across all categories.

5.3 Geographical Coverage and Entity Structure

Dyad-level comparison: POLECAT has far more unique country dyads (4,471 vs. GENOME's 1,691), but including GENOME's third-party countries narrows this gap substantially (4,191). Only 2,384 dyads are shared, yet these cover the top dyads by volume (e.g., Israel–Iran, Russia–Ukraine, USA–China), with higher mean events per dyad in the shared group—suggesting convergence on the most salient relationships.

Exclusive dyads and entity resolution: The datasets' exclusive dyads differ structurally: nearly half of GENOME's belong to verbal cooperation (under 2% for POLECAT). GENOME also resolves specific international bodies (NATO, IMF, G7) that POLECAT aggregates under a generic "International Organization" label. Conversely, POLECAT's exclusive dyads include state-to-state pairs such as Chile–Venezuela and Colombia–India, indicating gaps in GENOME's non-Western coverage.

Country-level distribution: For most countries, POLECAT records substantially more events than GENOME: the five with the largest gap (India, Russia, Israel, Ukraine, Brazil) alone account for a dif-

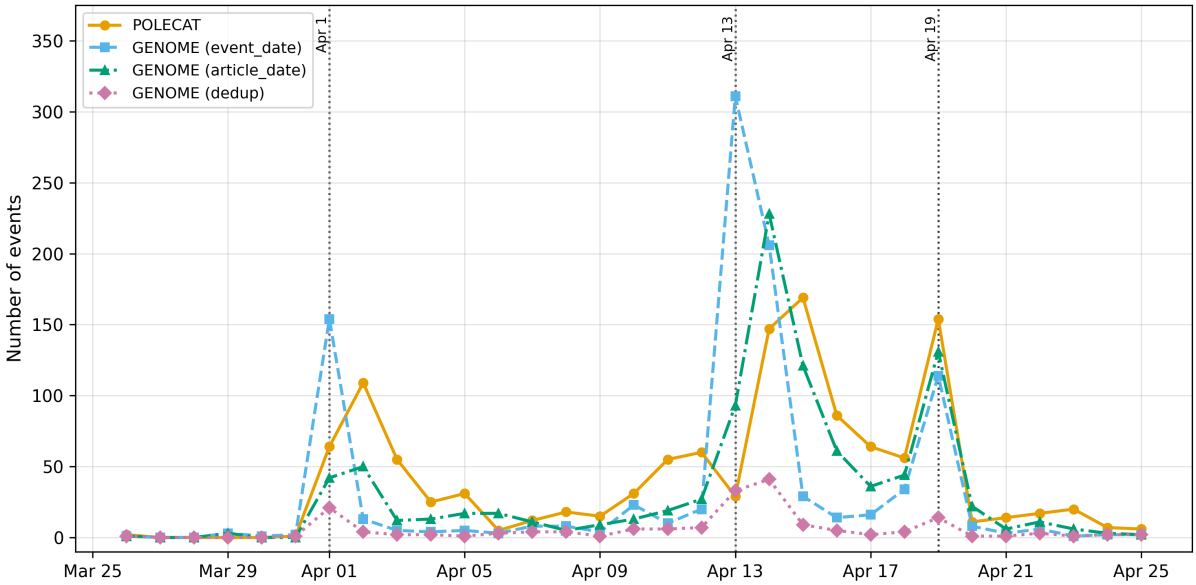

Figure 3: Daily event counts for the Israel–Iran case study (Mar 25 – Apr 25, 2024).

ference of 42,000 events. This disparity is also pronounced in Latin America, the Caucasus, and South and Southeast Asia. GENOME exceeds POLECAT only for a narrow cluster, most notably Palestine and the United Nations.

Event structure: GENOME incorporates more countries per event: 1.34 actors, 0.88 recipients, and 1.09 third parties on average, compared to POLECAT's 1.05 actors and 0.55 recipients.

6 Discussion

GENOME and POLECAT can be understood as similar products with different underlying mechanics, which is reflected in differences in event volume, event composition, temporal attribution, and entity representation. GENOME produces a substantially lower event volume, largely because it typically yields one event per article while encoding richer agent structures. In contrast, POLECAT extracts more events per article and covers a broader range of sources, including some non-Western regions. It is also likely that POLECAT introduces more duplicates of the same underlying event, either due to more redundant inputs or a higher extraction rate within individual articles.

Despite strong alignment between the two datasets on conflict event types, reflecting their shared ontology, a notable divergence emerges in cooperative events. POLECAT contains almost no CONSULT events and shows a generally low volume of verbal cooperation. This pattern suggests that routine diplomatic interactions, such as meetings and official statements, are filtered out at the input stage rather than misclassified. GENOME, by contrast, captures a more balanced distribution of cooperative events, which is particularly relevant for early warning applications where lower intensity interactions may carry predictive signal. Despite operating on different input corpora, the two datasets converge on the same major events. Daily-count correlations with POLECAT are weakest in aggregate, with material conflict in particular appearing nearly uncorrelated. However, once restricted to dyads where both systems record substantial activity, material conflict correlation turns sharply positive in nearly all cases. The aggregate gap therefore reflects differing geographic coverage rather than mismatched classifications.

GENOME also appears better able to attribute events to their actual date of occurrence rather than the publication date of the reporting source. This distinction is reflected in the increased volatility observed when moving from GENOME (article_date) to GENOME (event_date): publication dates introduce artificial smoothing due to reporting delays, whereas event dates cluster reports onto the day the event occurred, producing sharper temporal signals around known geopolitical developments. This effect is especially visible in the Israel–Iran case study (Figure 3). POLECAT, in contrast, shows a uniform decline in events over weekends, indicating stronger sensitivity to media publication cycles. This bias is further confirmed at the dyad level, where POLECAT's daily counts align most closely

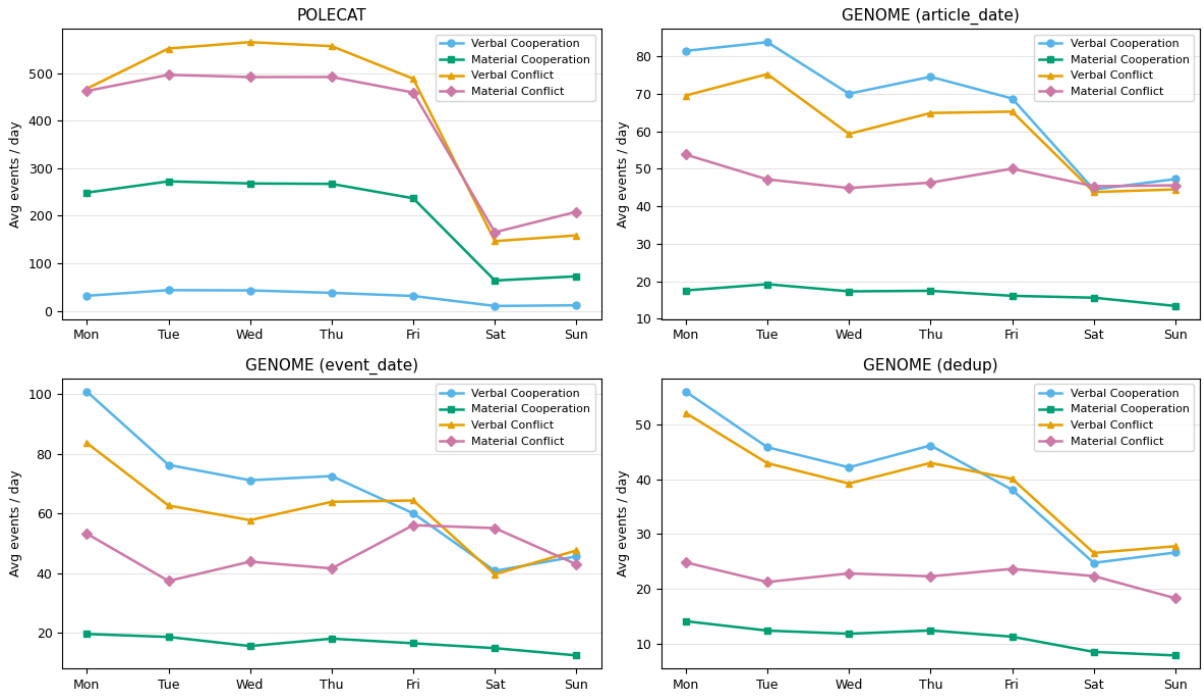

Figure 4: Day-of-week event volume by category across datasets.

with GENOME (article_date) in the majority of shared dyads. GENOME also exhibits a weekend reduction, but this is concentrated in verbal events, consistent with reduced diplomatic activity and public statements, while material events remain relatively stable. This suggests that GENOME's pipeline partially mitigates media cycle bias. Its deduplication step further reduces redundancy in highly reported events, although its accuracy still requires formal evaluation.

Finally, GENOME demonstrates more granular entity resolution. It distinguishes a range of international organisations, including NATO, IMF, WTO, BRICS, and the ICC, whereas POLECAT largely restricts such entities to major supranational bodies like the European Union and the United Nations, grouping others under a generic "International Organization" category. This allows GENOME to capture supranational interactions in greater detail. In addition, the Third Party role further supports the shift toward multi-entity event representations described in Section 2.2.

7 Conclusion

As of today, GENOME provides a continuously updated, automatically extracted dataset of geopolitical events covering both conflictual and cooperative, verbal and material inter-state interactions. Its pipeline demonstrates that LLM-based extraction can produce structured event data that aligns with established ontologies and captures dynamics consistent with real-world patterns. However, while POLECAT embodies a more granular detail level (e.g., geolocalization, Wikidata normalization, context-mode entries), GENOME currently represents a foundation to be built upon. Future work will focus on reintroducing these features, enriching the dataset with finer-grained classification of events and agent attributes, and expanding coverage through more diverse, multilingual input sources. Additionally, developing a manually annotated gold standard for the PLOVER ontology remains a priority, both for GENOME and for the broader socio-political event data community, as it would enable systematic evaluation of extraction quality across competing systems.

Limitations

Several limitations should be acknowledged. First, GENOME relies exclusively on English-language newswire sources from a single commercial provider, which introduces both structural and linguistic bias toward Western media perspectives and English-speaking regions. This likely explains the coverage gaps in Latin America, the Caucasus, and South and Southeast Asia identified in the geographical comparison with POLECAT. Relying on Western-centric, English-language media

inherently restricts the observability of localized or low-intensity geopolitical interactions in the Global South.

Second, GENOME lacks a manually annotated gold standard for evaluation, a limitation shared by the broader automated geopolitical event extraction community. While the indirect comparison with POLECAT and known real-world occurrences provides useful indicators of systemic validity, the lack of record-level ground truth prevents a formal calculation of standard precision, recall, and F1 metrics for the extraction and classification modules. Without such ground truth, it remains difficult to determine whether divergences from POLECAT, particularly in cooperative event types, reflect genuine improvements in coverage or systematic biases introduced by the LLM-based pipeline. Developing such a dataset remains a priority to enable systematic evaluation across competing systems.

Third, to minimize hallucination and standardize inputs, the extraction module is deliberately prompted to identify the single "core" event of an article, discarding secondary, historical, or contextual interactions unless strictly necessary. While this design choice successfully reduces noise and redundancy, it inherently caps the recall of the system, potentially missing valid but subordinate events embedded later in the text. Additionally, article texts are truncated at approximately 200 words under the assumption of a standard newswire inverted-pyramid structure. Articles that deviate from this format, such as feature pieces, opinion columns, or non-Western journalistic conventions, may yield incomplete or inaccurate extractions. The 4.4% of events with negative date lags further suggests that the date inference mechanism is susceptible to hallucination, time-zone ambiguities, and references to scheduled future events.

Fourth, while the multi-criteria deduplication pipeline and entity resolution steps demonstrably reduce the redundancy of highly covered events and clean the dataset's network structure, their accuracy has not been formally evaluated against human judgments. The chosen similarity thresholds (e.g., the 0.80 composite score) were optimized iteratively on pilot data but may require domain-specific tuning to prevent the over-merging of distinct but similar events occurring in close temporal proximity. Similarly, the 35.4% reduction in unique agent names through embedding-based clustering risks merging distinct entities that share similar names or descriptions.

Fifth, the current implementation utilizes a simplified version of the PLOVER ontology. It focuses on the 16 root event types and introduces a novel Third Party role, but it temporarily omits finer-grained sub-types, exact geolocalization, Wikidata normalization, and contextual-mode tagging that POLECAT provides, limiting the granularity of downstream analyses.

Finally, the pipeline depends on proprietary, closed-source LLMs (the OpenAI GPT family), including `gpt-4o-mini` for extraction and a fine-tuned `gpt-4.1-nano` for classification. While cost-effective and highly performant for structured outputs, this reliance introduces concerns regarding reproducibility over time, as underlying model weights and behaviors may be updated by the provider without notice. The fine-tuning process further relies on silver-standard labels produced by `gpt-5.1` rather than human annotations, potentially propagating systematic errors into the classification module.

References

Kaiqian Ahrabian, Ethan Boxer, and Jay Pujara. 2025. Toward better temporal structures for geopolitical events forecasting. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM 2024)*, pages 588–602. Association for Computational Linguistics.

Kiymet Akdemir and Ali Hürriyetoğlu. 2022. Zero-shot ranking socio-political texts with transformer language models to reduce close reading time. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, pages 124–132. Association for Computational Linguistics.

Edward E. Azar. 1980. The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution*, 24(1):143–152.

Rui Bai, Dong Lu, Sheng Ran, Emily M. Olson, Himank Lamba, Allison Cahill, Joel Tetreault, and Alejandro Jaimes. 2025. CEHA: A dataset of conflict events in the horn of africa. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics.

Kyle Beardsley, Patrick James, Jonathan Wilkenfeld, and David Quinn. 2024. What is escalation? Measuring crisis dynamics in international relations with human and LLM generated event data. ArXiv:2402.03340.

Doug Bond, Brian Bennett, and William Vogele. 1994. Data development and interaction events analysis using KEDS/PANDA: An interim report. Technical report, International Studies Association, Washington, DC.

Patrick T. Brandt and Marlo Sianan. 2025. Measurement of event data from text. *Frontiers in Political Science*, 6.

Erica Cai and Brendan O'Connor. 2023. A monte carlo language model pipeline for zero-shot sociopolitical event extraction. ArXiv:2305.15051.

Ruoxi Chen, Chengzhi Qin, Wenxing Jiang, and Donna Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of ChatGPT for event extraction. *Preprint*, arXiv:2303.03836.

Deborah J. Gerner, Philip A. Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. Paper presented at the International Studies Association, New Orleans.

Myrsini Halkia, Stefano Ferri, Michail Papazoglou, and Marlies Kampen Schellens. 2020. Dynamic global conflict risk index. Technical report, Publications Office of the European Union. JRC Technical Report.

Andrew Halterman, Benjamin E. Bagozzi, Andreas Beger, Philip A. Schrodt, and Grace I. Scarborough. 2023a. PLOVER and POLECAT: A new political event ontology and dataset. SocArXiv.

Andrew Halterman and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis*.

Andrew Halterman, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace I. Scarborough. 2023b. Creating custom event data without dictionaries: A bag-of-tricks. ArXiv:2304.01331.

J. Craig Jenkins, Charles L. Taylor, Marianne Abbott, Thomas V. Maher, and Lindsey Peterson. 2012. The world handbook of political indicators IV. Technical report, Mershon Center for International Security Studies, The Ohio State University.

Gary King and Will Lowe. 2003a. 10 million international dyadic events. IQSS Dataverse Network.

Gary King and Will Lowe. 2003b. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642.

Kalev Leetaru and Philip A. Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, pages 1–49.

Charles A. McClelland. 1961. The acute international crisis. *World Politics*, 14(1):182–204.

Charles A. McClelland. 1978. World event/interaction survey, 1966–1978. Technical Report 5211, ICPSR.

Shrey Meher and Patrick T. Brandt. 2025. ConflLlama: Domain-specific adaptation of large language models for conflict event classification. *Research & Politics*, 12(3).

Sean P. O'Brien. 2010. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104.

Helene Brinken Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 40–53. Association for Computational Linguistics.

Open Event Data Alliance. 2024. PLOVER: Political language ontology for verifiable event records — event, actor and data interchange specification (draft version 2.0).

Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. 2023. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10:74.

Shehzad Salam, Patrick Brandt, Vito D'Orazio, Jennifer Holmes, Javier Osorio, and Latifur Khan. 2020. An online structured political event dataset based on CAMEO ontology. SocArXiv.

Grace I. Scarborough, Benjamin E. Bagozzi, Andreas Beger, John Berrie, Andrew Halterman, Philip A. Schrodt, and Jevon Spivey. 2023. POLECAT weekly data.

Sina J. Semnani, Peng Zhang, Wenxuan Zhai, Hao Li, Nicholas Beauchamp, Tyler Billing, Koko Kishi, Michaela Li, and Monica Lam. 2025. LEMONADE: A large multilingual expert-annotated abstractive event dataset for the real world. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25813–25852. Association for Computational Linguistics.

Surendrabikram Thapa, Suresh Adhikari, Hristo Tanev, and Ali Hürriyetoğlu. 2025. Challenges and applications of automated extraction of socio-political events at the age of large language models. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025*, pages 6–19.

Jay Yonamine. 2016. A guide to event data: Past, present, and future. *All Azimuth: A Journal of Foreign Policy and Peace*.

Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2025. Event-based evaluation of abstractive news summarization. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 504–510. Association for Computational Linguistics.

Appendix

A Event Type Counts

See Table 4.

B Dyad-level Pearson correlations

See Table 5.

Table 4: Event counts and shares by type for GENOME, GENOME (dedup), and POLECAT (Feb–Jun 2024).

	GENOME		Dedup		POLECAT	
Type	**N**	**%**	**N**	**%**	**N**	**%**
Verbal cooperation						
CONSULT	6,646	23.3	3,794	22.6	267	0.2
AGREE	2,486	8.7	1,565	9.3	0	0.0
SUPPORT	686	2.4	513	3.1	0	0.0
CONCEDE	168	0.6	105	0.6	4,227	2.7
Material cooperation						
AID	1,787	6.3	1,207	7.2	21,175	13.5
COOPERATE	525	1.8	382	2.3	319	0.2
RETREAT	166	0.6	88	0.5	9,101	5.8
Verbal conflict						
ACCUSE	4,786	16.8	3,232	19.2	37,580	23.9
REQUEST	3,121	10.9	1,989	11.8	15,738	10.0
THREATEN	685	2.4	351	2.1	9,484	6.0
REJECT	387	1.4	238	1.4	29	0.0
Material conflict						
ASSAULT	4,526	15.9	1,903	11.3	30,254	19.2
SANCTION	1,554	5.4	701	4.2	10,093	6.4
COERCE	506	1.8	374	2.2	11,858	7.5
MOBILIZE	401	1.4	268	1.6	1,746	1.1
PROTEST	100	0.4	85	0.5	5,457	3.5
Total	**28,530**	**100**	**16,795**	**100**	**157,328**	**100**

Table 5: Daily-count Pearson r between POLECAT and each GENOME variant (art = article_date, evt = event_date, dd = dedup) for the 14 dyads with at least 100 events in both datasets, broken down by PLOVER quad category (Feb–Jun 2024). Dyads ordered by combined volume. "n/a" indicates that one or both sides had zero events in the corresponding quad.

	Overall			V-Coop			M-Coop			V-Conf			M-Conf		
Dyad	art	evt	dd	art	evt	dd	art	evt	dd	art	evt	dd	art	evt	dd
Russia–Ukraine	+0.24	+0.20	+0.15	+0.12	+0.29	−0.04	+0.07	+0.14	+0.15	+0.24	+0.30	+0.18	+0.21	+0.11	+0.11
Iran–Israel	+0.86	+0.46	+0.58	+0.32	+0.36	−0.02	+0.00	+0.18	+0.21	+0.91	+0.62	+0.58	+0.81	+0.42	+0.47
Israel–United States	+0.38	+0.26	+0.33	+0.03	+0.09	+0.08	+0.13	+0.11	+0.03	+0.30	+0.18	+0.25	+0.18	+0.17	+0.31
Russia–United States	+0.61	+0.56	+0.47	−0.04	−0.07	−0.07	+0.15	+0.08	+0.09	+0.46	+0.41	+0.47	+0.37	+0.31	+0.08
Israel–Syria	+0.20	+0.23	+0.11	n/a	n/a	n/a	+0.06	+0.06	+0.06	+0.06	+0.06	+0.06	+0.24	+0.26	+0.10
China–United States	+0.49	+0.50	+0.45	−0.07	−0.06	−0.07	−0.07	+0.08	+0.11	+0.31	+0.36	+0.39	+0.24	+0.17	+0.23
Israel–Lebanon	+0.43	+0.35	+0.25	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	+0.35	+0.23	+0.14	+0.40	+0.36	+0.28
Iran–United States	+0.65	+0.65	+0.55	n/a	n/a	n/a	+0.03	+0.22	+0.20	+0.56	+0.48	+0.51	+0.53	+0.61	+0.36
Ukraine–United States	+0.46	+0.29	+0.07	+0.60	+0.49	−0.06	+0.32	+0.26	+0.18	−0.05	+0.04	−0.00	−0.04	−0.04	−0.04
United States–Yemen	+0.16	+0.08	+0.32	n/a	n/a	n/a	−0.02	−0.02	−0.01	+0.10	+0.03	+0.03	+0.18	+0.10	+0.35
China–Philippines	+0.59	+0.33	+0.35	+0.28	+0.28	+0.37	−0.03	−0.03	−0.03	+0.51	+0.45	+0.38	+0.25	+0.13	+0.07
North Korea–South Korea	+0.70	+0.49	+0.56	+0.26	+0.16	−0.04	−0.04	−0.04	−0.04	+0.51	+0.36	+0.41	+0.49	+0.47	+0.36
Russia–United Kingdom	+0.64	+0.70	+0.44	−0.03	−0.03	−0.03	n/a	n/a	n/a	+0.31	+0.27	+0.29	+0.74	+0.78	+0.51
United Kingdom–Yemen	+0.11	+0.14	+0.42	n/a	n/a	n/a	−0.01	−0.01	−0.01	−0.04	−0.04	−0.04	+0.12	+0.14	+0.47