Challenges and Applications of Automated Extraction of Socio-political Events at the age of Large Language Models

Surendrabikram Thapa[1], Surabhi Adhikari[2], Hristo Tanev[3], Ali Hürriyetoğlu[4]
[1]Virginia Tech, USA, [2]Columbia University, USA,
[3]European Commission, Joint Research Centre, Italy,
[4]Wageningen Food Safety Research, Netherlands
[1]sbt@vt.edu, [2]surabhi.adhikari@columbia.edu,
[3]hristo.tanev@ec.europa.eu, [4]ali.hurriyetoglu@wur.nl

Abstract

Socio-political event extraction (SPE) enables automated identification of critical events such as protests, conflicts, and policy shifts from unstructured text. As a foundational tool for journalism, social science research, and crisis response, SPE plays a key role in understanding complex global dynamics. The emergence of large language models (LLMs) like GPT-4 and LLaMA offers new opportunities for flexible, multilingual, and zero-shot SPE. However, applying LLMs to this domain introduces significant risks, including hallucinated outputs, lack of transparency, geopolitical bias, and potential misuse in surveillance or censorship. This position paper critically examines the promises and pitfalls of LLM-driven SPE, drawing on recent datasets and benchmarks. We argue that SPE is a high-stakes application requiring rigorous ethical scrutiny, interdisciplinary collaboration, and transparent design practices. We propose a research agenda focused on reproducibility, participatory development, and building systems that align with democratic values and the rights of affected communities.

1 Introduction

Socio-political events (SPEs) are occurrences involving political or social actors that have significance for societies or governance. Protests, conflicts, elections, policy changes, and diplomatic interactions are examples of SPEs. In computational terms, an SPE can be represented as a structured record of who did what to whom, when and where, extracted from text (Cai and O'Connor, 2023). Event extraction systems seek to transform unstructured data (e.g. news articles, social media posts) into structured event representations (often as tuples like source–action–target with time and location) (Hu et al., 2024). Such structured event databases enable large-scale analysis of political dynamics and serve as inputs for monitoring conflict, tracking trends, and forecasting crises (Hu et al., 2024). In both academic research and real-world decision-making, having timely and accurate event data is crucial. Analysts use these databases to understand patterns of violence, policymakers use them for early warnings, and humanitarian organizations for situational awareness.

Automated SPE extraction has grown in importance as the volume of text data (news, social media) explodes beyond human coding capacity. Traditional rule-based or supervised systems have been used to populate global event databases (e.g. extracting 'who attacked whom') for decades. Recently, large language models (LLMs) have begun to play a transformative role in this space. LLMs like GPT-3.5 and GPT-4 can, in principle, read and interpret complex texts to identify events with minimal task-specific training. Early experiments show that advanced LLMs (e.g. GPT-4) significantly outperform previous models in zero-shot political event coding, handling nuanced distinctions better and generalizing with fewer examples (Hu et al., 2024). The success of GPT-4 in following event coding guidelines highlights the vast potential of LLMs for this task (Hu et al., 2024). At the same time, LLMs introduce new challenges (like hallucination and transparency issues, discussed later) that must be managed. This position paper takes a hybrid technical and policy-oriented view of automated socio-political event extraction in the era of LLMs, examining not only the algorithmic and data-centric hurdles but also the ethical, legal, and societal implications of these technologies.

2 Technical Challenges in SPE

Despite progress, automated SPE extraction faces numerous technical challenges.

2.1 Ambiguity and Coreference

Language describing socio-political events is often ambiguous. A single phrase can imply different event types depending on context (e.g. "sanction"

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, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-002

could mean an economic sanction or simply approval) (Cai and O'Connor, 2023; Hürriyetoğlu et al., 2022a; Danilova and Popova, 2014). Identifying whether an event actually occurred or is hypothetical (modality) also requires understanding subtle cues (did a politician promise an action or actually do it?). Moreover, the information about one real-world event may be scattered across multiple sentences or reports. Systems must perform coreference resolution to merge mentions referring to the same event. For example, in the text 'A protest broke out in CityX… The demonstration continued into the night', linking 'protest' and 'demonstration' is non-trivial. Recent efforts have been made to explicitly evaluate event coreference linking across sentences (Hürriyetoğlu et al., 2022b). However, ambiguity and cross-sentence reference remain open problems. Without resolving these, an automated system might count one event multiple times or miss it entirely.

2.2 Temporal and Spatial Grounding

Every event entry needs a when and where (Abraham et al., 2018; Westin, 2025). Extracting accurate temporal and geospatial information is challenging. News text may describe an event with relative times ('earlier today', 'last week') that require context (e.g., publication date) to resolve. Locations can be mentioned at various granularities (a city, a region, a country), and many event coders need coordinates, which requires mapping place names to a gazetteer (Hürriyetoğlu et al., 2024). Ensuring that each event is anchored to the correct date and place is vital for analysis (e.g., distinguishing two protests on different days). Temporal ordering (figuring out the sequence of events) is also difficult when texts jump around chronologically. Techniques from temporal IE and geographic entity resolution are needed as part of any robust SPE pipeline. These tasks remain hard, especially in noisy or terse text (like social media), where time/place might not be explicitly stated.

2.3 Multilinguality and Low-Resource Languages

Socio-political events occur worldwide, and being able to extract events from multiple languages is essential for global coverage. Many high-profile event extraction systems have focused on English (or a few major languages) due to data availability. However, relying only on English sources creates a biased pictures (Claro et al., 2019; Miok et al., 2024). The challenge is that NLP resources (annotated data, pretrained models) for many low-resource languages are limited. Progress is being made (Hürriyetoğlu et al., 2022b). Still, performance typically drops for truly low-resource languages (with different scripts or limited data).

2.4 Dataset Quality and Reproducibility

High-quality training and evaluation data are expensive to create (Thapa et al., 2023). Annotating event mentions in text (especially with detailed role labeling or fine-grained event types) is time-consuming and often requires expert knowledge of political contexts (Olsen et al., 2024; Cardie and Wilkerson, 2008). As a result, existing datasets may be small, sparse, or inconsistently annotated. Many academic event extraction datasets (e.g. ACE 2005, TAC KBP event tracks) focus on a limited ontology and are not perfectly aligned with the needs of socio-political analysis (Doddington et al., 2004; Mitamura et al., 2015). On the other hand, political science event datasets (like ICEWS or ACLED) contain high-level coded events but are not released with their source texts (often due to copyright), making it hard to use them for supervised learning or to reproduce results (Raleigh et al., 2010; O'brien, 2010). This raises a reproducibility challenge. A research group may train a model on proprietary news data and output a set of events, but without public text data, others cannot replicate the extraction process. Furthermore, different datasets use different schemas, making it hard to compare systems. Annotation consistency is also an issue, as complex events can suffer from low inter-annotator agreement if guidelines are vague.

2.5 Event Schema and Ontology Design

What counts as an "event" and how it is categorized can vary greatly. Designing an ontology (schema) for events is a foundational challenge that affects extraction (Danilova and Popova, 2014; Xiang and Wang, 2019). SPE extraction has been guided by schemas like CAMEO (Conflict and Mediation Event Observations) which defines a hierarchy of around 20 top-level event classes and over 200 subtypes for political interactions (from cooperative acts like appeals or meetings to conflictual ones like protests, attacks) (Parolin et al., 2019; Gerner et al., 2002). Other ontologies exist (ACE's schema for general events, custom schemas for cybersecurity events, etc.), and social science projects have proposed new ones (e.g., PLOVER, a recent political

violence ontology aligning with CAMEO) (Halterman et al., 2023). The schema design problem has two elements: (1) deciding on the categories and their granularity (balancing detail with annotator reliability), and (2) ensuring models can generalize across schema changes. A rigid ontology may become outdated as new event types emerge (for example, "COVID lockdown protest" might not fit neatly into older categories). On the other hand, very broad definitions reduce analytical usefulness.

3 Applications and Use Cases of LLMs

3.1 Conflict Early Warning and Crisis Forecasting

One of the original motivations for machine-coded event data was to feed conflict early warning systems (Hegre et al., 2019). Projects like the Integrated Crisis Early Warning System (ICEWS) have used continuous streams of coded events (protests, violence, cooperation events, etc.) to predict instability and conflict outbreaks. By analyzing trends e.g. a spike in protests or escalating repressive events, these systems aim to forecast the risk of civil war, mass atrocities, or other crises, enabling preventative action. Automated event extraction greatly speeds up the data pipeline for such systems, which need near-real-time updates from daily news. LLMs could enhance early warning by improving the recall of relevant events (catching subtle precursors in text) and by summarizing situational reports (Foisy et al., 2025; Baek et al., 2023). For example, an LLM might synthesize disparate reports into a narrative of escalating tension.

3.2 Use by Governments and International Organizations

Governments and intergovernmental organizations (IGOs) are heavy users of event data (Ngai et al., 2025). Intelligence and defense agencies use event extraction to monitor global security like identifying terror attacks, troop movements, or diplomatic gestures in open sources. The U.S. government's ICEWS program is one example where automated event data directly supports analysts. Diplomatic services might track protest movements or election-related unrest in real time to inform embassy staff. At the IGO level, organizations like the United Nations or regional bodies (African Union, EU) may utilize event data for peacekeeping and policy decisions (Nohuddin and Zainol, 2020; Amicarelli and Di Salvatore, 2021). The U.N.'s crisis mapping initiatives and the World Bank's political risk assessments rely on understanding the event landscape. Here, comprehensiveness and reliability of event extraction are key. An LLM-powered system might help by reading situation reports or local news in various languages and highlighting events of concern, thus augmenting human analysts.

3.3 NGOs and Humanitarian Monitoring

Non-governmental organizations (NGOs), especially in the human rights and conflict prevention space, have been both producers and consumers of event data (Alhelbawy et al., 2020). A notable example is ACLED (Armed Conflict Location & Event Data Project), an NGO-driven effort that manually curates conflict and protest events across the world. ACLED (Raleigh et al., 2010) and others (e.g. Crisis Group, Human Rights Watch's data teams) might use automated extraction to extend their reach, scanning local media or social platforms for reports of violence that their human coders can then verify and add. Humanitarian organizations can benefit from real-time event feeds to coordinate responses. For instance, knowing about protests turning violent could help the Red Cross prepare, or detecting displacement events could trigger UNHCR action. LLMs could assist these NGOs by quickly summarizing large volumes of community radio transcripts or Facebook posts from affected communities, pulling out events like "village attacked by armed group" or "aid convoy blocked by protesters."

3.4 Event Databases and Knowledge Graphs

In academia and policy research, curated event databases are valuable for studying patterns of conflict, cooperation, and social movements (Zhao et al., 2024). Automated extraction is used to populate and update these databases continuously (Deng et al., 2024; Gottschalk and Demidova, 2018). For example, the GDELT project has attempted to automatically ingest global news and output coded events for every day. While impressive in scale, such efforts sometimes sacrificed precision for breadth. With LLMs, there is potential to improve the quality of automated event databases. An LLM can consider subtler contexts than keyword-based systems, thereby potentially reducing false positives. Moreover, LLMs can help unify or reconcile events. If multiple news reports describe the same protest from different angles, an LLM might consolidate them into one entry with a more com-

plete description (this borders on automatic summarization of events). Knowledge graphs are another use where events can be nodes linking actors, places, and dates in a graph database. Querying such graphs can answer complex questions (e.g. "find all confrontations between government forces and tribe X in the past year"). Automated SPE extraction is what supplies the raw material for these knowledge bases. LLMs could be used to populate new types of relations in graphs, like sentiment or causal links (e.g. "protest led to policy change"). There is active research on using LLMs to enrich knowledge graphs with event information extracted from text (Deng et al., 2024).

3.5 Analytical Tools and Summarization

Finally, a growing application is the use of LLMs for higher-level analysis of event data. Rather than just populating a database, an LLM can help analysts make sense of the data (Kumar et al., 2024). For instance, given a chronology of extracted events, an LLM could produce a narrative report or timeline summary ("In June, a series of protests in X province escalated into clashes by August, prompting government crackdown in September..."). This moves into the realm of report generation and explanatory analysis. Automating such analytical tasks has policy value as busy decision-makers may not have time to read dozens of incident reports, but a well-crafted summary or even an on-demand Q&A powered by an LLM (e.g. "Has violence against civilians increased this month compared to last?") could be immensely helpful. Some prototypes in media monitoring have used LLMs to summarize global news on a topic across countries. For example, summarizing how different countries' press are reacting to a conflict. Those same capabilities can be tuned to summarizing event data. Additionally, interactive exploration via natural language questions is an exciting use case. For example, an analyst could ask the system (which has ingested an event database) questions in English and get answers or charts, without needing to write code or SQL. LLMs can serve as an interface between humans and complex event data, broadening access to insights. Caution is warranted to keep the LLM "grounded" in actual data (so it doesn't fabricate answers). Combining retrieval methods with LLMs (so the model bases answers on retrieved event records) is one technique being explored for this purpose (Arslan et al., 2024).

4 Limitations of LLMs, Multilingual and Global Considerations

4.1 Technical Limitations

Introducing LLMs into the pipeline brings its own set of technical caveats (Thapa et al., 2025). By design, generative LLMs will fill in gaps and produce plausible text even when the input is uncertain. This can lead to hallucinated events, i.e. the model might assert that an event occurred that isn't actually supported by the source (Zhang et al., 2025; Ji et al., 2023; Shiri et al., 2024; Liu et al., 2025). For example, if given a vaguely worded report, an LLM might "assume" a protest happened when in reality the text was speculating. Ensuring faithful extraction requires grounding the LLM to the source text. Relatedly, LLM outputs can be inconsistent; the same prompt might yield slightly different extractions on different runs (due to sampling variability), which is problematic for a deterministic database update. Stability and calibration of confidence in extracted facts are therefore technical issues to solve. Another limitation is interpretability as deep learning models, especially large generative ones, are often black boxes. Understanding why a model classified something as, say, an "attack" versus an "arrest" can be difficult, hindering our ability to trust and refine the system. LLMs also have practical limitations like they may struggle with very long documents (context length limits), or with remembering a long list of ontology definitions without confusion.

4.2 Non-Western Contexts and Local Nuance

Many event extraction tools and models have been developed primarily on Western news sources and in languages like English, Spanish, or French (Aliyu et al., 2024; Kulkarni and Dogra, 2024). Applying these to events in, say, rural Africa or Central Asia can pose problems. The way events are reported, the cultural context, and the actors involved may differ greatly (Hürriyetoğlu et al., 2022b). For example, a "protest" in one country might be described very differently in another country's media (or might not be reported openly at all). Local idioms or euphemisms (e.g., referring to rebel militants as "our boys" in some context) might mask what an event is about. Also, the salience of event types can differ. Events like tribal clashes, land disputes, election violence, etc., each have unique markers. An extraction system needs to be tuned into these nuances. This often requires

involving regional experts in the loop, or at least using region-specific data to fine-tune models. One promising avenue is to engage local journalists or organizations to help create training data (perhaps via annotation or feedback) for their context, creating a more inclusive global system. LLMs, with their ability to absorb vast multi-domain text, might already know some culturally specific references, but careful prompt engineering is needed to make them work for less-covered contexts.

4.3 Cross-Lingual and Low-Resource Techniques

As mentioned, multilingual capability is crucial. There are a few approaches to handle it (Jafri et al., 2024; Alghamdi et al., 2024). One is machine translation (MT), i.e., translate all foreign texts to a pivot language (e.g. English) and then run an English event extractor (Chew et al., 2025; Cabrera, 2024). This was a common strategy in earlier systems, but MT errors can lead to missed or wrong events (especially if translation alters proper names or event verbs). Another approach is using multilingual models like multilingual BERT or XLM (Pires et al., 2019; Conneau et al., 2020), which have some cross-lingual transfer ability. Such models can sometimes be trained on a high-resource language and still be applied to a related low-resource language. Few-shot learning with LLMs could shine where one could prompt an LLM in a target language with a few examples of event annotations in that language (or even in English, relying on its cross-lingual knowledge) and get results. There is early research on prompt-based cross-lingual IE which is encouraging. Additionally, active learning could be employed i.e., the system asks humans to translate or verify a few critical pieces to improve itself iteratively.

4.4 Multimodal Event Extraction

Socio-political events are not only described in text; they may be captured in images, videos, or even satellite data (Bhandari et al., 2023; Thapa et al., 2024). A protest might be live-streamed, a damage assessment might come from satellite imagery, a social media image might show evidence of an attack. Multimodal event extraction seeks to combine text with other data sources to improve event detection and validation. For instance, an automated system could corroborate a reported protest (text) with social media images geotagged in that city showing crowds. LLMs are expanding into multimodal models (e.g. vision-language models like GPT-4's multi-modality or others that can process images) (Thapa et al., 2025; Fei et al., 2024). A future SPE pipeline might take a news article and also any attached photo or video transcript, and use both to decide what happened. Multimodal analysis can improve recall (catch events that text missed but image shows) and precision (disambiguate events by seeing visuals). It also helps in contexts where text might be propagandistic and images can sometimes cut through biases (though they have their own issues of authenticity).

4.5 Bias and Representation in Global Data

Global event extraction must grapple with bias in sources (Xiang and Wang, 2019; Spiliopoulou et al., 2020; Dev et al., 2021). Many regions lack independent media, or any media coverage at all of certain event types (e.g. state repression might be hidden). As a result, automated systems might reflect state narratives or international media agendas. Being aware of these gaps is part of a global perspective. There are efforts to include non-traditional sources. For instance, using reports from NGOs or crowdsourced data to complement news. A balanced approach might merge information from local citizen reports with mainstream media, with the AI model reconciling them. Bias mitigation techniques can be applied, such as calibration (if a known bias exists, adjust the data distribution) (Garrido-Muñoz et al., 2021; Sun et al., 2019). Ultimately, a global system may need regional tuning, as what works well for event extraction in Europe might need rethinking for Central Africa. Community evaluations and workshops (like regional "data challenges") could help identify where current models fall short. Inclusivity in the development process (having NLP researchers and social scientists from diverse regions) is also vital to ensure the tools are attuned to global realities and not just Western media patterns.

5 Policy and Ethical Challenges

5.1 Surveillance and Authoritarian Misuse

A powerful SPE extraction system can turn into a double-edged sword. On one hand, it can provide transparency and early warnings about crises; on the other, it could enable authoritarian surveillance at an unprecedented scale (Yabanci, 2025; Roberts and Oosterom, 2024). Repressive regimes might use automated event detection to track dissident

activities or protests in real-time, flagging leaders and participants for reprisal. Unfortunately, this is not just hypothetical. AI-driven surveillance and policing systems are already used by authoritarian governments and have been found effective in suppressing political unrest and entrenching regimes. If an event extraction tool can scrape social media and news to pinpoint every protest or strike as it begins, authorities could quickly crack down, undermining civil liberties. Even in democratic societies, law enforcement has shown interest in such tools. This kind of proactive surveillance blurs the line between public safety and infringement of the right to assemble.

5.2 Privacy and Human Rights

Related to the above, the privacy implications of large-scale event monitoring are significant (Baldassarre et al., 2024). Socio-political events often involve individuals like protesters, activists, and even victims of violence. If an automated system is parsing social media for events, it might incidentally capture personal data like names of organizers, eyewitness accounts, etc. Even news articles can contain personal identifying information in event descriptions. Using AI to aggregate and analyze this at scale can amplify privacy risks. For instance, extracting a "protest event" from a Facebook post could reveal the poster's political participation without their consent. Furthermore, in conflict zones or authoritarian contexts, being identified in an event report (e.g., as attending a demonstration) could endanger one's safety. Human rights organizations worry that indiscriminate use of such technology could lead to abuses such as compiling watchlists of protesters or surveilling minority communities under the guise of event detection.

5.3 Misinformation and Propaganda

Automated event extraction systems could inadvertently become conduits for misinformation or propaganda if not carefully managed. These systems rely on source data which may be inaccurate or biased. For example, state-controlled media might report a fabricated event (e.g. a false "terror plot foiled") or exaggerate an incident for propaganda. If an automated pipeline naively extracts that into the event database, it lends credence to the false narrative and propagates it to any downstream users (analysts, alert systems, etc.). There is a real risk of false positives where an SPE system could report an event that never actually happened, due to either misinterpretation or malicious input. In the context of political events, such an error can have serious consequences (imagine a system that mistakenly alerts to a "coup attempt" that was just a rumor, and governments could react harshly). Systems should thus cross-validate events with multiple sources or official reports when possible.

5.4 Bias, Fairness, and Data Provenance

Automated SPE extraction inherits and can even amplify biases present in source data (Huang et al., 2024; Kumari et al., 2024). Media reporting bias is well documented. For instance, studies find that international media severely underreport violence in certain regions compared to others. If an event extraction system relies on those media, the resulting database will systematically undercount or underplay conflicts in those underreported regions. This raises fairness concerns around analyses using the data might over-focus on areas that the media highlight and neglect others. Bias can also creep in through the algorithms. If an ML model were trained mostly on, say, Western news text, it might not recognize event triggers in the rhetoric of other cultures or might misclassify events that don't fit its learned patterns. Furthermore, LLMs themselves carry biases from their training data; they might be more likely to extract events that sound "newsworthy" in a Western sense, for example.

6 Recommendations and Guidelines

6.1 Robust Dataset Creation and Sharing

The community should establish best practices for creating and sharing event data. This includes clear documentation of inclusion criteria, coding methodologies, and known limitations of any event dataset. Data collectors (whether researchers or organizations) have a responsibility to explicitly state what sources they use, what counts as an event, and what biases might result. When possible, datasets should be shared in a form that supports reproducibility. For example, reference URLs or source snippets for each coded event (within copyright constraints) should be provided. Creative solutions like releasing machine-readable summaries or embeddings of text can be explored to respect copyright while still enabling method comparison. The community could benefit from an open repository of annotated texts for events (perhaps using texts that are in the public domain or licensed for research) to serve as a benchmark. Moreover, any new event ontology

or schema should ideally be published openly, with rationales for design, to encourage standardization or at least interoperability between projects.

6.2 Integration of LLMs with Human Oversight ("Human-in-the-Loop")

To harness LLM power while safeguarding against errors, a human-in-the-loop approach is highly recommended (Amirizaniani et al., 2024; Cohn et al., 2024). LLMs can be used to draft event annotations or suggest events, but human analysts or annotators should verify critical details, especially for high-impact events. For instance, an LLM might summarize a complex report into a tentative event entry; a human can then check the source, correct any misinterpretation, and approve it. This not only prevents spurious data from entering official records but also allows humans to catch subtle biases the AI might introduce. Output validation is crucial and automated confidence scores from models can guide which events need human review (low confidence or novel event types get flagged). Additionally, employing multiple systems (e.g., an LLM and a rule-based checker) in parallel and comparing outputs where disagreements can be routed to humans can be useful. This kind of cross-validation workflow ensures that LLMs augment rather than replace expert judgment in sensitive applications.

6.3 Transparent Model Use & Explainability

Any use of LLMs or AI for SPE extraction in policy or public-facing contexts should be transparent (Foisy et al., 2025). Stakeholders (from end-users of an event dataset to citizens potentially affected by its use) deserve to know if an event was identified by a human, a classical algorithm, or an LLM, and what the reliability might be. We recommend developing explainability tools specific to event extraction. For example, if an LLM classifies something as an "armed attack" event, the system should ideally provide a rationale or highlight the evidence in text that led to this classification. Techniques such as step-by-step reasoning prompts or modular pipelines can help with interpretability. At the very least, event records generated or assisted by AI could carry a tag or confidence level. In high-stakes use (e.g. legal accountability for conflict incidents), one might decide that no event enters the official record without either two independent sources or human verification similar to journalistic standards. Transparency reports on system performance, biases found, and corrections made would also build trust in the technology.

6.4 Ethical Guidelines and "Do No Harm" Policies

It is imperative to establish and follow ethical guidelines for deploying SPE extraction, particularly in volatile and sensitive regions. Drawing on principles from humanitarian and human rights domains, developers should adopt a "Do No Harm" mentality by anticipating how the technology could cause harm and work to mitigate it. For example, if deploying a system to monitor protests in an oppressive regime, measures should be taken so the data is not easily accessible to the regime to target individuals (perhaps aggregating or anonymizing certain elements). Collaboration with ethics boards or oversight committees can provide external review of such deployments. Access control might be one guideline. For example, sensitive event data (like locations of protest organizers) might only be shared with vetted parties like NGOs, not made fully public. The community could formulate a code of conduct or ethics checklist for SPE projects, including considerations like 'have we accounted for bias?', 'are the communities being monitored aware or have a say?', 'is there a risk of misuse and how are we preventing it?' For LLM-specific issues, guidelines should stress not to over-rely on AI without verification, and to always have a human accountability in the loop for decisions made from event data. When working in conflict zones, respecting local laws and norms, and protecting sources (e.g. journalists or informants who are reporting events) is also part of ethical use.

6.5 Bias Awareness and Correction

To address fairness, we recommend that any large-scale SPE extraction effort include an explicit bias assessment phase. This might involve comparing the AI-extracted data with known baselines (perhaps human-curated datasets like ACLED in some regions) to see where discrepancies lie. If certain event types or areas are consistently under-detected, the model or pipeline should be adjusted (additional training data for those cases, or lowering thresholds). Bias correction techniques such as reweighting events from underrepresented regions can be applied to the output data. Another best practice is involving local stakeholders in evaluating the system's output, like having experts from different regions review the events detected in their region for completeness and accuracy. Not only

does this catch biases, but it also builds a more inclusive system. Data provenance, as mentioned, should be maintained. Each event record ideally links to its source material, which allows users to judge source reliability and bias. If an event comes only from a single source with a strong slant, perhaps the system can flag that (like "source is state media"). Users of the data should be educated on these provenance flags. In essence, continuous auditing for bias and an openness about the system's limits will improve fairness and trustworthiness.

6.6 Collaboration Among Stakeholders

Finally, we urge a strong collaboration between the technical developers (NLP researchers, scientists) and the policy community (political scientists, ethicists, legal experts, and practitioners on the ground). This cross-domain dialogue can ensure that the tools developed address real needs and align with norms. For example, engaging with human rights organizations might highlight the need for certain event categories (like "internet shutdown event") that technologists hadn't considered. Policymakers, on the other hand, should stay informed about the capabilities and limits of the latest tech, avoiding both unrealistic expectations and ungrounded fears. Joint workshops or working groups can produce normative guidelines that marry technical possibilities with ethical guardrails. We recommend formulating clear use policies for different scenarios, e.g., guidelines for using event extraction in election monitoring versus in conflict zones (the latter might require more restraint). By working together on scenario planning, the community can preemptively set standards for responsible use (similar to how bioethics guides biomedical innovations).

7 Future Directions

7.1 Hybrid Extraction Models

Future research will likely explore hybrid models that combine the strengths of LLMs with structured symbolic knowledge (He et al., 2025; Shaik and Doboli, 2025). For example, an LLM could be used to interpret text and draft possible events, but a symbolic reasoner or knowledge graph ensures consistency with known facts (preventing obvious contradictions or impossibilities). Integrating expert-defined rules (from event coding manuals) into LLM prompts or architectures could yield systems that are both flexible and precise. One concrete direction is leveraging existing political ontologies and knowledge bases to guide LLMs, e.g., providing a model with a library of event type definitions and historical examples to reduce ambiguity. This addresses the question posed by researchers like 'can we use expert knowledge to enhance efficiency without extensive new data?'. Progress in prompt engineering and fine-tuning will make LLM outputs more controllable, which is crucial for complex event schemas.

7.2 Adaptive and Continual Learning

Socio-political realities evolve, and so must our extraction systems. A promising avenue is continual learning (Wang et al., 2024) for LLM-based extractors, i.e., the ability to update the model as new event types emerge or new slang/terms enter the lexicon, without forgetting past knowledge. This could involve periodic fine-tuning on newly annotated events or streaming adaptation where the model's prompts are adjusted based on feedback. One challenge is avoiding "catastrophic forgetting" when adapting to new domains (Kirkpatrick et al., 2017). Research into LLMs that can plugin new information (modular learning or using external memory) will benefit SPE greatly, as it means, for example, the system that was never trained on "COVID-19 lockdown protest" could learn that category on the fly. Additionally, ontology evolution should be handled, as event schemas are revised (which happens in social science as new patterns like cyber warfare become relevant), systems need to incorporate those changes.

7.3 Multimodal and Multilingual Fusion

Building on current trends, the future will likely see fully multimodal event extraction in practice. This means models that simultaneously process text, images, video, and maybe audio to detect and validate events. A protest event, for instance, could be confirmed by both a news text and a tweet with a photo. Research into multimodal transformers and alignment techniques (like aligning image detection of violence with text reports) is burgeoning. By 2025 and beyond, we anticipate systems that can, say, take a live social media feed (text + images) and output structured events to dashboards for crisis responders. On the multilingual front, future work may achieve more universal models that work across dozens of languages via a combination of improved training data and leveraging LLM's polyglot capabilities. There is also room for transfer learning between languages and modalities.

For example, an event described in French text and an Arabic tweet might be linked as the same event through a shared embedding space.

7.4 Narrative Construction & Causal Analysis

Moving up the value chain, an exciting research frontier is automated narrative and causality extraction. It's not just about listing events, but understanding how they connect. Future LLM-driven systems could attempt to identify causal or temporal relationships. For example, protest A led to government response B, which triggered conflict C. Some early studies are looking at event chains and temporal reasoning with LLMs. If successful, this could produce draft analytical reports or help populate causal graphs of events, which are immensely useful for political analysis (like understanding escalation paths or conflict dynamics). There is also potential for what-if analysis. With generative models, one could simulate how a sequence of events might unfold under different scenarios, giving policymakers a tool to explore consequences (though this enters speculative territory and would need robust grounding in data). Additionally, as LLMs become more explainable, we might use them to interrogate event data like "Why did violence increase in region X?" and the system might highlight a series of coded events (e.g. arrests, then protests, then clashes) as an explanation. Achieving this level of reliable narrative construction will require advances in discourse understanding and knowledge integration for LLMs.

7.5 Data Responsibility and Ethics

On the policy side, a major future direction is establishing international norms or agreements on the responsible use of AI for social data analysis. Just as there are treaties and agreements on the use of certain surveillance (for instance, UN discussions on digital privacy), we may see efforts to set guidelines for technologies like event extraction, especially as they get more powerful with LLMs. Researchers and practitioners should collaborate in forums to develop a code of ethics specific to computational event monitoring. This could encompass agreements on not facilitating human rights abuses, ensuring data sharing for humanitarian purposes, and perhaps even certification of systems (an independent audit to say an event extraction system meets certain bias and transparency standards). Work in this direction will involve not just technical people, but also lawyers, ethicists, and the communities being monitored. Another aspect is education and literacy. Future efforts should include training for policymakers and journalists on how to interpret AI-generated event data, to avoid misuse or misinterpretation.

7.6 Open Research and Collaboration

Finally, a future direction that underpins all others is maintaining an open and interdisciplinary research environment. The challenges at this sociotechnical junction are complex; solving them will require insights from NLP, machine learning, political science, conflict studies, ethics, and more. We envision more joint research endeavors like political scientists formulating problems that NLP folks can help solve, and NLP advances (like new LLM capabilities) being rapidly tested on social science use cases. There is also likely to be increased benchmarking and evaluation efforts specific to SPE, creating shared tasks that evaluate not just extraction accuracy but also bias, fairness, and utility in downstream analysis. A "roadmap" paper from a multi-disciplinary team could periodically assess where we stand and recalibrate goals (for example, setting a goal to achieve a certain reliability in low-resource languages by year X). As foundation models evolve (e.g., new versions of GPT or open-source LLMs with tens of billions of parameters), continually applying them and assessing their fit for event extraction tasks will be an ongoing process. Keeping this work open (publishing results, sharing models) will ensure broad access and avoid a scenario where only a few large players dominate the technology (which could be risky if their interests don't align with public interest).

8 Conclusion

In conclusion, automated socio-political event extraction sits at a pivotal point with the rise of LLMs. The coming years will likely bring substantial improvements in capability with support for more languages, more nuanced detection, and richer outputs. At the same time, ensuring these advancements are applied responsibly and benefit the global community is a collective task for researchers, practitioners, and policymakers. By recognizing the challenges and actively working on both technical solutions and ethical safeguards, we can harness LLMs to better understand and respond to the socio-political events that shape our world.

References

Susanna Abraham, Stephan Mäs, and Lars Bernard. 2018. Extraction of spatio-temporal data about historical events from text documents. *Transactions in GIS*, 22(3):677–696.

Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. 2024. Fake news detection in low-resource languages: A novel hybrid summarization approach. *Knowledge-Based Systems*, 296:111884.

Ayman Alhelbawy, Mark Lattimer, Udo Kruschwitz, Chris Fox, and Massimo Poesio. 2020. An nlp-powered human rights monitoring platform. *Expert Systems with Applications*, 153:113365.

Yusuf Aliyu, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani BA Rahman, and Mujaheed Abdullahi. 2024. Sentiment analysis in low-resource settings: a comprehensive review of approaches, languages, and data sources. *IEEE Access*, 12:66883–66909.

Elio Amicarelli and Jessica Di Salvatore. 2021. Introducing the peacekeeping operations corpus (pkoc). *Journal of Peace Research*, 58(5):1137–1148.

Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. Developing a framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*.

Muhammad Arslan, Saba Munawar, and Christophe Cruz. 2024. Political-rag: using generative ai to extract political information from media content. *Journal of Information Technology & Politics*, pages 1–16.

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

Seungwon Baek, Do Namgoong, Jinwoo Won, and Seung H Han. 2023. Automated detection of social conflict drivers in civil infrastructure projects using natural language processing. *Applied Sciences*, 13(20):11171.

Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernández Nieto, Domenico Gigante, and Azzurra Ragone. 2024. Fostering human rights in responsible ai: A systematic review for best practices in industry. *IEEE Transactions on Artificial Intelligence*, 6(2):416–431.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1994–2003.

Luis Cabrera. 2024. Babel fish democracy? prospects for addressing democratic language barriers through machine translation and interpretation. *American Journal of Political Science*, 68(2):767–782.

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

Claire Cardie and John Wilkerson. 2008. Text annotation for political science research.

Edward Chew, Mahasweta Chakraborti, William Weisman, and Seth Frey. 2025. Machine translation for accessible multi-language text analysis. *Computational Communication Research*, 7(1):1.

Daniela Barreiro Claro, Marlo Souza, Clarissa Castellã Xavier, and Leandro Oliveira. 2019. Multilingual open information extraction: Challenges and opportunities. *Information*, 10(7):228.

Clayton Cohn, Caitlin Snyder, Justin Montenegro, and Gautam Biswas. 2024. Towards a human-in-the-loop llm approach to collaborative discourse analysis. In *International Conference on Artificial Intelligence in Education*, pages 11–19. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Vera Danilova and Svetlana Popova. 2014. Sociopolitical event extraction using a rule-based approach. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 537–546. Springer.

Songgaojun Deng, Maarten de Rijke, and Yue Ning. 2024. Advances in human event modeling: from graph neural networks to language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6459–6469.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2021. On measures of biases and harms in nlp. *arXiv preprint arXiv:2108.03362*.

Ling Ding, Xiaojun Chen, Jian Wei, and Yang Xiang. 2023. Mabert: mask-attention-based bert for chinese event extraction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(7):1–21.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Citeseer.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8.

Laurence-Olivier M Foisy, Étienne Proulx, Hubert Cadieux, Jérémy Gilbert, Jozef Rivest, Alexandre Bouillon, and Yannick Dufresne. 2025. Prompting the machine: Introducing an llm data extraction method for social scientists. *Social Science Computer Review*, page 08944393251344865.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

Deborah J Gerner, Philip A Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. 2002. The creation of cameo (conflict and mediation event observations): An event data framework for a post cold war world. In *annual meeting of the American Political Science Association*, volume 29.

Simon Gottschalk and Elena Demidova. 2018. Eventkg: A multilingual event-centric temporal knowledge graph. In *European semantic web conference*, pages 272–287. Springer.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics*.

Andrew Halterman, Benjamin E Bagozzi, Andreas Beger, Phil Schrodt, and Grace Scraborough. 2023. Plover and polecat: A new political event ontology and dataset. In *International Studies Association Conference Paper*.

Qiyuan He, Jianfei Yu, and Wenya Wang. 2025. Large language model-enhanced symbolic reasoning for knowledge base completion. *arXiv preprint arXiv:2501.01246*.

Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, et al. 2019. Views: A political violence early-warning system. *Journal of peace research*, 56(2):155–174.

Yibo Hu, Erick Skorupa Parolin, Latifur Khan, Patrick Brandt, Javier Osorio, and Vito D'Orazio. 2024. Leveraging codebook knowledge with nli and chatgpt for zero-shot political relation classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–603.

Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in opinion summarisation from pretraining to adaptation: A case study in political bias. *arXiv preprint arXiv:2402.00322*.

Ali Hürriyetoğlu, Osman Mutlu, Fatih Beyhan, Fırat Duruşan, Ali Safaya, Reyyan Yeniterzi, and Erdem Yörük. 2022a. Event coreference resolution for contentious politics events. *arXiv preprint arXiv:2203.10123*.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022b. Extended multilingual protest news detection - shared task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 223–228, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024. Chunav: Analyzing hindi hate speech and targeted groups in indian election discourse. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Amit Kulkarni and Varun Dogra. 2024. Comprehensive survey of event extraction methods in natural language processing. In *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, pages 925–929. IEEE.

Raghvendra Kumar, Ritika Sinha, Sriparna Saha, and Adam Jatowt. 2024. Extracting the full story: a multimodal approach and dataset to crisis summarization in tweets. *IEEE Transactions on Computational Social Systems*.

Gitanjali Kumari, Anubhav Sinha, Asif Ekbal, Arindam Chatterjee, and Vinutha B N. 2024. Enhancing the fairness of offensive memes detection models by mitigating unintended political bias. *Journal of Intelligent Information Systems*, 62(3):735–763.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.

Jiangwei Liu, Liangyu Min, and Xiaohong Huang. 2021. An overview of event extraction and its applications. *arXiv preprint arXiv:2111.03212*.

Wenxuan Liu, Zixuan Li, Long Bai, Yuxin Zuo, Daozhu Xu, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2025. Towards event extraction with massive types: Llm-based collaborative annotation and partitioning extraction. *arXiv preprint arXiv:2503.02628*.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. Eventplus: A temporal event understanding pipeline. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 56–65.

Charles McClelland. 1978. World event/interaction survey, 1966-1978. *WEIS Codebook ICPSR*, 5211(640):49.

Kristian Miok, Encarnación Hidalgo Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, and Marko Robnik-Šikonja. 2024. Multi-aspect multilingual and cross-lingual parliamentary speech analysis. *Intelligent Data Analysis*, 28(1):239–260.

Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2015. Overview of tac kbp 2015 event nugget track. In *TAC*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Eric WT Ngai, Ariel KH Lui, and Brian CW Kei. 2025. Natural language processing in government applications: a literature review and a case analysis. *Industrial Management & Data Systems*, 125(6):2067–2104.

Puteri N.E. Nohuddin and Zuraini Zainol. 2020. Discovering explicit knowledge using text mining techniques for peacekeeping documents. *Int. J. Bus. Inf. Syst.*, 35(2):152–166.

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 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.

Erick Skorupa Parolin, Sayeed Salam, Latifur Khan, Patrick Brandt, and Jennifer Holmes. 2019. Automated verbal-pattern extraction from political news articles using cameo event coding ontology. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE intl conference on high performance and smart computing,(HPSC) and IEEE intl conference on intelligent data and security (IDS)*, pages 258–266. IEEE.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Clionadh Raleigh, Rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of peace research*, 47(5):651–660.

Tony Roberts and Marjoke Oosterom. 2024. Digital authoritarianism: a systematic literature review. *Information Technology for Development*, pages 1–25.

Philip A Schrodt. 2001. Automated coding of international event data using sparse parsing techniques. In *annual meeting of the International Studies Association, Chicago*.

Philip A Schrodt, Shannon G Davis, and Judith L Weddle. 1994. Political science: Keds—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587.

Hashmath Shaik and Alex Doboli. 2025. Using a symbolic knowledge graph to address llm limitations in analog circuit topology generation. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00528–00533. IEEE.

Fatemeh Shiri, Farhad Moghimifar, Reza Haffari, Yuan-Fang Li, Van Nguyen, and John Yoo. 2024. Decompose, enrich, and extract! schema-aware event extraction using llms. In *2024 27th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

Evangelia Spiliopoulou, Salvador Medina Maza, Eduard Hovy, and Alexander Hauptmann. 2020. Event-related bias removal for real-time disaster events. *arXiv preprint arXiv:2011.00681*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383.

Fereshta Westin. 2025. Time, technique and text: scoping review of temporal information extraction and categorisation in documents. *Journal of Documentation*, 81(7):135–156.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Bilge Yabanci. 2025. Surveil, datafy, publicize: digital authoritarianism and migration governance in turkey. *Democratization*, 32(4):1016–1041.

Ziyao Zhang, Chong Wang, Yanlin Wang, Ensheng Shi, Yuchi Ma, Wanjun Zhong, Jiachi Chen, Mingzhi Mao, and Zibin Zheng. 2025. Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):481–503.

Bang Zhao, Yilong Zhao, and Ying Mao. 2024. A method for judicial case knowledge graph construction based on event extraction. In *Proceedings of the 2024 9th International Conference on Intelligent Information Technology*, pages 62–69.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

A.1 Related Works

Systematic political event data collection dates back to the Cold War era. In the 1960s and 70s, political scientists began manual coding of international events from news reports to enable quantitative analysis (Olsen et al., 2024). Influential early datasets like WEIS (World Events Interaction Survey) and COPDAB (Conflict and Peace Data Bank) catalogued interstate events (e.g. protests, conflicts, diplomatic acts) by human annotation of news archives (McClelland, 1978; Olsen et al., 2024; Azar, 1980). These pioneering efforts demonstrated the value of structured event data but were labor-intensive and limited in scope (covering only certain actors or regions). By the late 1980s, researchers recognized that much of this coding could be automated by text processing. The Kansas Event Data System (KEDS) in the early 1990s was a seminal rule-based system that used dictionaries and patterns to code events from newswire feeds (like Reuters) (Schrodt et al., 1994). KEDS (and its successor TABARI) could scan sentences for keywords indicating actions (e.g. 'attack', 'meet') and map them to predefined event types, initiating the era of machine-coded event databases (Schrodt, 2001). These early systems were capable of coding thousands of articles, paralleling developments in the NLP field of information extraction.

In the 1990s and 2000s, the NLP community's work on event extraction evolved in parallel. Early information extraction (IE) tasks in NLP, such as the MUC competitions and later ACE, involved identifying event "triggers" and participants in text (for example, extracting a terrorist bombing event with its perpetrator, target, date, etc.) (Grishman and Sundheim, 1996; Doddington et al., 2004). While political scientists' event databases aimed at capturing abstract real-world events (often aggregating information across sources), NLP tasks focused on text-bound events with token-level annotations (Olsen et al., 2024). This led to a divergence. Socio-political event databases prioritized what actually happened in the world (even if details were spread across multiple documents), whereas NLP event annotations captured what was explicitly mentioned in a single text. Nonetheless, by the 2010s there was convergence in methodology. Statistical and ML-based approaches emerged for event extraction. For example, supervised classifiers to detect if a sentence describes a protest,

or sequence labeling models to mark event triggers and arguments. Researchers began applying emerging deep learning techniques to event extraction, achieving improvements over brittle pattern-matchers (Olsen et al., 2024). However, these supervised models required substantial annotated data (which was scarce for fine-grained socio-political events) and often struggled to adapt when event schemas or ontologies changed.

The late 2010s and early 2020s saw the advent of large pretrained language models, culminating in today's LLMs (Thapa et al., 2025; Naveed et al., 2023). Initially, these models were used as contextual encoders in neural event extraction pipelines (Hu et al., 2024; Ma et al., 2021; Ding et al., 2023). For example, BERT-based classifiers for protest detection or relational models for 'who did what to whom' (Liu et al., 2021). More recently, prompt-based extraction and in-context learning have become feasible. Given a prompt describing event categories or a few examples, an LLM can attempt to parse new texts into structured event records without explicit retraining. This zero-shot or few-shot capacity is attractive for socio-political events, which often require flexibility to new event types or languages. Early studies are mixed but promising. For instance, one study found GPT-4 could achieve nearly the performance of a supervised classifier in coding political event types, and even exceeded some rule-based systems in recall (Hu et al., 2024). At the same time, prompting LLMs for complex, fine-grained event coding exposes issues (memory limits for long ontology descriptions, prompt sensitivity, etc.), indicating that LLMs are not a silver bullet (Thapa et al., 2025; Li et al., 2024; Ziems et al., 2024). The field has now reached a point where hybrid approaches are being explored like combining LLMs with knowledge bases, using retrieval-augmented generation (RAG) for factual grounding, and integrating human feedback for higher fidelity. This sets the stage for understanding the technical challenges that persist and the new considerations that arise in the LLM era.