Political Analysis (2025), **0**, 1–29
doi:10.1017/pan.2025.10027

PA

ARTICLE

Extractive versus Generative Language Models for Political Conflict Text Classification

Patrick T. Brandt[1], Sultan Alsarra[2], Vito D'Orazio[3], Dagmar Heintze[1], Latifur Khan[4], Shreyas Meher[5], Javier Osorio[6] and Marcus Sianan[1]

[1]School of Economic, Political and Policy Sciences, University of Texas at Dallas, USA; [2]Computer Science, King Saud University, Saudi Arabia; [3]Political Science, West Virginia University, USA; [4]Engineering and Computer Science, University of Texas at Dallas, USA; [5]Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam, Rotterdam, The Netherlands; [6]Department of Political Science, University of Arizona, USA

Corresponding authors: Patrick T. Brandt; Email: pbrandt@utdallas.edu; Sultan Alsarra; Email: salsarra@ksu.edu.sa

(Received 19 December 2024; revised 23 July 2025; accepted 24 July 2025)

Abstract

We review our recent ConfliBERT language model (Hu *et al.* 2022 [ConfliBERT: A Pre-Trained Language Model for Political Conflict and Violence]) to process political and violence-related texts. When fine-tuned, results show that ConfliBERT has superior performance in accuracy, precision, and recall over other large language models (LLMs) like Google's Gemma 2 (9B), Meta's Llama 3.1 (7B), and Alibaba's Qwen 2.5 (14B) within its relevant domains. It is also hundreds of times faster than these more generalist LLMs. These results are illustrated using texts from the BBC, re3d, and the Global Terrorism Database. We demonstrate that open, fine-tuned models can outperform the more general models in terms of accuracy, precision, and recall, and at a fraction of the cost.

Keywords: natural language processing; event data coding; event data

Edited by: Daniel J. Hopkins and Brandon M. Stewart

1. Introduction

How does one compare the latest large language models (LLMs) to prior methods for text-as-data applications where political science domain knowledge is well developed and important? Choosing the appropriate tool for the extraction and classification of relevant information from large and often unstructured corpora is a contemporary and ongoing challenge. As social (political) scientists, we possess replicable and encoded domain expertise to understand texts in our field and apply appropriate methods for our tasks (either with humans, text-as-data, natural language processing (NLP), or other methods). How then should one combine the insights of domain experts and computational scientists to evaluate which models are useful for extracting the domain information across various tasks with attention to accuracy, cost, and other metrics of interest? Should we use simpler information extraction tools or newer, generative, and more costly LLMs? To answer these questions, we compare ConfliBERT (Hu *et al.* 2022), our domain-specific, extractive encoder model, to a selection of more recent generative LLMs. We examine different political science text-as-data applications, such as event extraction, classification, and named entity recognition (NER), with comparisons in terms of accuracy, processing speed, and other performance metrics. Through these analyses, we gain an understanding of the capabilities of longer-established extractive NLP tools versus more recent generative models.

An area of focus with significant application and domain expertise is conflict event data coded from news reports. The transformation of news texts into structured "who-did-what-to-whom" event data

© The Author(s), 2025. Published by Cambridge University Press on behalf of The Society for Political Methodology.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2025.10027

is fundamental in international relations and studies of conflict and political violence. The process of gathering and preparing data for analysis in this domain is to assemble a corpus, filter for relevant information, identify target events, and annotate event attributes. This process can be costly and time-consuming: time required for data collection, structuring and filtering large amounts of text, training human annotators to apply the event ontology, and several rounds of quality controls to ultimately achieve a curated text corpus. This approach parallels the widely systematized way to process text-as-data in much of international relations and the social sciences (Croicu and Eck 2022; Grimmer, Roberts, and Stewart 2022; O'Connor, Stewart, and Smith 2013). Ultimately, the purpose is to extract relevant information on source actors (who), actions (did what), targets and actors (to whom), and other attributes for political science research.

Although computational methods have been used to analyze political texts for decades (Gerner *et al.* 1994), tools for these tasks have developed rapidly with the advent of LLMs. Models that were commonly used include *extractive* LLMs that can be trained specifically for classification, NER, and to annotate other features of the text. This is the focus of models like BERT, RoBERTa, DistilBERT, ELECTRA, and ConfliBERT, which are all variously sized (layered) encoder neural network models. More recently, this also includes *generative* LLMs that both encode the original text and provide a decoder to summarize the output features of interest, the generative output from a prompt. This includes many of the now familiar LLMs like Gemma, Llama, Qwen, ChatGPT, etc. In this research, we compare these extractive and generative types of LLMs for common text analysis tasks. The focus is not on a comparison across BERT-alike models, but on a domain-specific, fine-tuned BERT model (ConfliBERT) to more recent generative LLMs.

We make three significant contributions. *First*, compared to recent generative LLMs, ConfliBERT has superior performance based on classification metrics (AUC, F_1, etc.) applied to multiple datasets (BBC, Global Terrorism Database [GTD], and re3d) and tasks (binary and multi-class classifications and NER). Specifically, ConfliBERT outperforms Meta's Llama 3.1 (Dubey *et al.* 2024), Google's Gemma 2 (Team Gemma *et al.* 2024), and Alibaba's Qwen 2.5 (Hui *et al.* 2024) in relevant tasks. These results show that fine-tuned models used to extract political conflict information from domain-relevant texts can outperform the more general models in terms of accuracy, precision, and recall. This is consistent with prior work from Hürriyetoğlu *et al.* (2021), Kent and Krumbiegel (2021), Ollion *et al.* (2023), Wang (2024), and Croicu and von der Maase (2025). *Second*, ConfliBERT is hundreds of times faster than generative LLMs at identical tasks. This time savings is important when processing hundreds of thousands or millions of documents, as is often the case for large-scale event coding projects like Georeferenced Event Data or Militarized Interstate Dispute (Palmer *et al.* 2022; Sundberg and Melander 2013). The savings is amplified when used for active learning, iterative coding, and additional rounds of fine-tuning. *Third*, ConfliBERT models are open and extensible, so these results align with other recent and related political science work and developing standards (Barrie, Palmer, and Spirling 2024; Burnham *et al.* 2024).[1] We ran the generative LLMs locally using the Ollama backend, a framework that provides the instruction-tuned model variants that are standard for task-based research, not the raw base models. For efficiency, these models were deployed with 4-bit quantization, a common practice that explains the low memory footprint in our results while representing a typical trade-off between performance and computational cost. This setup ensures our comparison is against LLMs as they are practically applied by researchers.

We begin with a short review of ConfliBERT, a domain-specific, pre-trained, and then fine-tuned model for the analysis of conflict texts. We then discuss how this model can be compared to newer, larger LLMs that are generative and thus more costly in terms of computational time and initial resource setup. Finally, we discuss and present the relative performance of the various models.

[1]We have documented ConfliBERT in full and made it available together with a clear set of documents and repositories describing the training data and how versions for Spanish and Arabic contexts can be processed. This is documented on our project sites: https://eventdata.utdallas.edu, https://github.com/eventdata/conflibert, and https://huggingface.co/eventdata-utd.

https://doi.org/10.1017/pan.2025.10027

2. ConfliBERT as an Extractive Domain Tool

For political science applications, we want to use a tool like ConfliBERT to accomplish three key information extraction and summarization tasks that are part of "coding event data": 1) filtering politically relevant information in a corpus, 2) identifying events, and 3) encoding their attributes. The first is well-solved in multiple ways using tools such as support vector machines, topic models, or dictionary-based methods (Beieler *et al.* 2016). The second, event identification, is crucial to create valid and reliable event datasets. These form the backbone of many quantitative analyses in the field. But this identification often requires iterations and revisions, requiring speed and computational efficiency as well as accuracy. Perhaps the most challenging aspect in this text processing is the third—the detailed annotation of event attributes. This is the "who," "what," "to whom," "where," and "when" of each identified event. This requires not just NER, but also understanding the roles these entities play in the event and the relationships between them.

Transformer architectures and LLMs show considerable promise across these event coding and text analysis tasks. For example, Parolin (2021), Parolin *et al.* (2021), and Parolin *et al.* (2022) explore the use of general (non-domain specific) Transformer models for cross-lingual, multi-label, and multi-task classifications in English, Spanish, and Portuguese. Base models, such as pre-trained BERT, were incorporated, adapted, and extended for different event coding tasks by changing the attention layers and recalibrating the parameters (Parolin *et al.* 2021). These innovations led to improvements in the accuracy, precision, recall, and F_1 of the classifications over original BERT and RoBERTa models across languages (Parolin *et al.* 2021, Table III).

We address how the filtering and extraction of annotations for conflict reports can be done with ConfliBERT, a domain-specific model that we pre-trained using the BERT architecture.[2] Unlike BERT and the many other general-purpose LLMs pre-trained on all sorts of text data, ConfliBERT is pre-trained with domain-specific texts about political conflict, violence, and international relations.[3] Our curated corpus of 33.7 GB of text consists of an expert domain corpus and a mainstream media corpus. The expert domain corpus (2,293 MB) contains political conflict texts and professional sources related to diplomacy, such as the United Nations, intergovernmental organizations (INGOs), think tanks, and government agencies. The mainstream media corpus contains the (a) Mainstream Media Collection (MMC) (20 GB), a corpus collected from 35 news agencies worldwide, (b) Gigaworld corpus (8,818 MB), which includes media coverage from seven international English newswires from 1994 to 2010,[4] (c) Phoenix Real-Time (PRT) event dataset (2,425 MB), which combines data from over 400 news agencies worldwide, and (d) Wikipedia's political events articles (2,845 MB), which were extracted from the 2021 Wikipedia dump (Hu *et al.* 2022). To remove texts unrelated to our domain, documents were filtered for relevance where appropriate.

ConfliBERT has previously been shown to perform better than BERT models (cased and uncased) based on macro F_1 statistics 1) across training set sizes (Hu *et al.* 2022, Figure 2) and 2) across relevant tasks in multiple test datasets related to political conflict and violence—such as 20News, GLOCON, GTD, SATP, InsightCrime, India Police Events, CAMEO codebook examples, MUC-4, and re3d—which are used across political science, national security, and NLP comparisons. Hu et al. (2022, Table 3 and Figure 1) establish ConfliBERT's superiority against a baseline of BERT (Devlin *et al.* 2018) for these datasets. For binary classification (BC) and NER, ConfliBERT was better than BERT (based on using cased and uncased models) using F_1 and macro F_1 statistics for weighted precision and recall.[5]

[2]When we started this in 2021, the Bidirectional Encoder Representations from Transformers (BERT) LLM was among the best, open models available. More recently, we introduced a conflict-specific adaptation of Llama (Dubey *et al.* 2024), a generative AI-based model, which we termed ConfLlama and trained on conflict-specific training data (Meher and Brandt 2025). More details are provided in Section 5.

[3]Training examples are listed at https://github.com/eventdata/ConfliBERT/tree/main/pretrain-corpora and test or evaluation ones are at https://github.com/eventdata/ConfliBERT/tree/main/data.

[4]Stories that were already covered in the MMC were removed from this corpus.

[5]The cross-model comparisons include the BBC News dataset (Greene and Cunningham 2006), a sample SATP dataset https://satp.org/, the 20 Newsgroups dataset (Lang 1995), the Gun Violence dataset (Pavlick *et al.* 2016), and the Event Status

https://doi.org/10.1017/pan.2025.10027

The performance of ConfliBERT has been validated and independently established by 1) Häffner *et al.* (2023) who find it superior to dictionary-based classifiers for conflict prediction, 2) the complete fine-tuning of the ConfliBERT model by Wang (2024) for similar tasks, 3) Croicu (2024) give additional and independent evidence of the model's strong performance relative to known alternatives for different conflict texts and related tasks, and 4) Croicu and von der Maase (2025) use of the model as part of a classification pipeline for a refined version of the UCDP GED data.

ConfliBERT builds on insights from the NLP literature and introduces 1) domain-specific pre-training, 2) fine-tuning training corpora from the conflict/political science domain, and 3) specific downstream tasks, such as BC, multi-label classification, and NER. It uses a transformer language model architecture and large amounts of politically relevant news texts as training data (Devlin *et al.* 2018). Similar to BERT, the pre-trained models minimize loss on masked token prediction, next sentence prediction, or both tasks. The pre-training for these models is either continuous or from scratch. Continuous means that it uses weights from another LLM as the starting point and tunes by minimizing loss on our domain corpus. Scratch means that we do not begin with the pre-learned weights, so learning only from the domain corpus.

We re-cast the problems of the political science domain into those more commonly seen in the information and computer science domains of NLP and inferences. This trades human annotation and classification costs for computational resources, which grow more powerful and cheaper. However, we need to bridge the way social scientists think about information extraction with how computational linguists and information scientists think about information extraction. Specifically, they focus on labeling spans of text corresponding to linguistic or contextual entities. In contrast, we focus on event attributes, their modality, and characteristics (Olsen *et al.* 2024). As a domain-specific, pre-trained LLM, ConfliBERT can help identify and categorize key features of political events from text without a fully specified ontology of actors or their interactions. These ontologies are required for dictionary-based approaches (Boschee *et al.* 2015).

Similar domain-specific BERT models have been shown to outperform generic BERT models in other scientific fields, such as biomedical (SCIBERT, Beltagy, Lo, and Cohan 2019), material sciences (MatSCIBERT, Gupta *et al.* 2022), legal (LegalBERT, Chalkidis *et al.* 2020), finance (FinBERT, Araci 2019), clinical notes (ClinicalBERT, Huang, Altosaar, and Ranganath 2020), and patent texts (patentBERT, Lee and Hsiang 2019). The benefits of the domain-specific approach of ConfliBERT extend to other languages as well. Recent extensions of the English language ConfliBERT model to ConfliBERT-Spanish (Yang *et al.* 2023) and ConfliBERT-Arabic (Alsarra *et al.* 2023) address the lack of non-English trained LLMs and permit the use of ConfliBERT's classification abilities to these two additional languages. Both are political conflict domain-specific LLMs without machine translations to English. Yang *et al.* (2023) pre-train and fine-tune ConfliBERT Spanish.[6] Compared to two Spanish-based models, mBERT and BETO—in all three tasks NER, binary, and multi-class classification—ConfliBERT Spanish outperformed the generic Spanish language models (Yang *et al.* 2023, Table II). Alsarra *et al.* (2023) introduce the same approach in ConfliBERT Arabic, a language-specific LLM that outperform competing models in the majority of cases on Arabic datasets that contained political, conflict-related,

dataset (Huang *et al.* 2016) for BC tasks. For NER tasks, ConfliBERT CONT cased achieved the highest macro F_1 for the source and target labeling NER task on the CAMEO Codebook (Gerner, Jabr, and Schrodt 2002) dataset, while ConfliBERT SCR uncased showed the highest macro F_1 on both the MUC-4 (DARPA 1992) and the DSTL (2018) datasets.

[6]Like the setup of the English-based ConfliBERT, this model was pre-trained on 11.7 GB of Spanish domain-specific text from (a) Spanish-language news websites, focusing exclusively on relevant categories, (b) websites of NGOs specializing in the field of human rights, violence, crime, and politics from different Spanish-speaking countries, and (c) Spanish text from the United Nations using MultiUN and the European Union's Directorate General for Translation. The pre-training was conducted on 12 layers, 768 hidden units, 12 attention heads, and 110M parameters and took about three days for each continual model using two Nvidia A-100 GPUs with 40 GB of memory each. An Adam optimizer (Kingma and Ba 2014) with a peak learning rate of 5e-5 and linearly decay was trained on 512-token sequences to account for long paragraphs in the new data.

https://doi.org/10.1017/pan.2025.10027

and international content (Alsarra *et al.* 2023, see Tables 3 and 4).[7] On datasets that did not primarily contain these specific topics, regular BERT models performed better than ConfliBERT Arabic. Its non-English variants, ConfliBERT-Spanish outperforms BERT variants like mBERT and BETO (Yang *et al.* 2023); and, ConfliBERT-Arabic does the same relative to AraBERT (Osorio *et al.* 2024).

3. The Event Coding Problem

In conflict event data research, scholars break down texts into key attributes: actors (sources and targets), actions, locations, and dates.[8] With actor coding, there are two broad approaches that, with the exception of training data, eschew human coding: mining past data to propose new groups or categories of actors (Solaimani *et al.* 2017b) and machine learning or transformer approaches using BERT-based and other models (Alsarra *et al.* 2023; Dai, Radford, and Halterman 2022; Halterman *et al.* 2023; Hu *et al.* 2022; Parolin *et al.* 2022; Yang *et al.* 2023). Prior work codes actions using sparse parsing with human-annotated dictionaries (Osorio *et al.* 2020; Schrodt 2001), whereas newer approaches handle new ontology or action extensions through up-sampling (Halterman and Radford 2021), natural language inference (NLI) (Croicu 2024; Dai *et al.* 2022; Halterman *et al.* 2023; Hu *et al.* 2022; Lefebvre and Stoehr 2023; Parolin *et al.* 2022), or zero-shot (ZS) prompts (Hu *et al.* 2024). Geographic coding in earlier work relied on the location inferred from the actors to identify where the event occurred. Some approaches to determine location use sparse parsing (Osorio *et al.* 2020), word embedding and NER (Halterman 2017; Imani *et al.* 2017; Imani, Khan, and Thuraisingham 2019, e.g.,), and even BERT (Halterman *et al.* 2023). For date or time coding of events, researchers generally parse the byline of the news report to acquire the publication date (Osorio *et al.* 2020), but the publication and the event occurrence dates are not always the same. A recent approach is to apply BERT technology to extract date information from the news story (Halterman *et al.* 2023). All of these information extraction approaches are prone to various errors (Brandt and Sianan 2025), and the latest methods attempt to reduce them using BERT-alike language models.

ConfliBERT provides a domain-level solution to these coding tasks. For most generative and extractive tasks, an LLM needs broad pre-training. These training steps generate huge costs in terms of 1) training data and its acquisition, 2) human/expert time, and 3) computational complexity to combine and produce the relevant model. In a domain-specific application, several choices make these challenges much more feasible for a social science tool like ConfliBERT. First, creating an extractive LLM or a BERT LLM (or even, for that matter, a *simple* predictive or generative suggestion model) can be done much more rapidly and cheaply. Since there is domain knowledge and insight provided in the initial training steps, steps 1 and 2 above for training a generic LLM are greatly scaled back, resulting in a superior model in a shorter period of time. Second, ConfliBERT can then be augmented or expanded (which we demonstrate below) to focus on harder tasks, such as ontology extension (Radford 2021), actor detection and recognition (Solaimani *et al.* 2017a, b), and image processing applications (Steinert-Threlkeld 2019; Wen *et al.* 2021).

[7]The model was pre-trained on 11.5 GB of Arabic source text from 84 sources originating from 19 Arabic-speaking countries, across 84 sources from news sites, mainstream media, and government sites, such as national news agencies representing a corpus of political, conflict, and political violence-related text. All text was collected in modern standard Arabic and contained news articles from the political, international and local sections of the sources to ensure that it represented domain-specific source text. The initial data collection was filtered using relevant keywords from the CAMEO ontology to further ensure the domain-specific relevance. The models were then trained using 12 base layers, 768 hidden units, 12 attention heads, and 110M parameters, in line with the approach utilized for training ConfliBERT Spanish.

[8]For those interested in getting event data and not exploring the weeds of this process, see Halterman *et al.*'s (2023) Political Language Ontology for Verifiable Event Records (PLOVER) and the POLitical Event Classification, Attributes, and Types (POLECAT) dataset, which are a record of domestic and international political interactions described in international news reports from 2010 to the present. The news reports are in English or machine translated into English from Arabic, Chinese, French, Portuguese, Russian, or Spanish before they are coded.

https://doi.org/10.1017/pan.2025.10027

The extraction of actors, action events (verbs), and additional information from texts for political science and international relations studies of conflict are accommodated in three different NLP tasks. The three main tasks that ConfliBERT addresses are:

Classification Which texts contain relevant information about politics, conflict, and violence? We give examples of this below based on data from the BBC and re3d text corpora. These are:
1. binary classifications: yes/no questions;
2. multi-label classifications: in a series of reports about protests, which types of protest are present (labor, peaceful, violent, etc.)?

Named Entity Recognition What are the "who" and "whom" that characterize the event? These are most typically the linguistic subjects and objects of the sentences and clauses, subject to textual disambiguation and co-referencing. But making sense of them becomes a task for a political scientist to identify the source/initiator of a political event toward a target or other political actor. We give an example below using texts about terrorist attacks from the GTD. We use NER to identify both traditional entities (Persons and Locations) and event-specific roles (Victims and Perpetrator Organizations). This approach is sometimes referred to as role-aware NER or event argument extraction and shares similarities with semantic role labeling. For this study, we train a single NER model to identify all entity types. We acknowledge that a more complex approach could involve training distinct models for each event type to resolve role ambiguities (e.g., an entity as a "victim" in one event and an "accuser" in another), an avenue we leave for future work.

Masking/Coding new entities and/or events is the extension of any ontology of new kinds of events. This can include teaching a model which events are new ones, ones to be excluded, or newly emergent actors and their roles.

The first two of these tasks may be viewed as a supervised learning problem and handled with statistical or machine learning algorithms. In this setting, the model is trained to learn and predict patterns based on repeated past examples or interactions. For example, D'Orazio *et al.* (2014) use support vector machines to classify texts on international conflict. This and similar approaches rely heavily on training data and may have trouble predicting out-of-sample when new patterns and types of conflict emerge. ConfliBERT improves on previous approaches like this by using longer embedded patterns of related text and its ability to comprehend context. It is able to accomplish this in situations where 1) the events or entities to be classified are rare and there are few examples to learn from or 2) where there is a class imbalance and the event or entity is not necessarily rare, but there are few relative to more common ones.

The last task is harder, but can begin with an LLM or a BERT-like model. To determine whether an event is similar to a prior one (or a related class or actor), we can provide examples that omit the thing to be predicted (masking or hiding it) and then assess how well the model performs. This problem can apply to new actors and events and the determination of whether the model/coder is correct relies on the domain knowledge of the social scientist. Further, the identification of new actors is often a masking task for which BERT-alike models are designed.[9]

These tasks could be done with extractive LLM models like ConfliBERT or could use newer generative models. The next section explores these choices and how they can be compared across several examples.

4. ConfliBERT Examples

ConfliBERT is an engine or baseline for extracting information about political texts. It 1) sorts political violence texts from other ones (classification), 2) identifies possible political actors and entities (since it was trained to do so and does NER with this knowledge), and 3) provides for masking and question

[9]Our recent work in actor detection proposes distant supervised (DS) and ZS approaches for extracting political actors and their roles using pre-trained language models (Hu *et al.* 2024; Parolin 2021).

https://doi.org/10.1017/pan.2025.10027

and answer (QA) tasks for coding. Clearly, the process and methodology here can be adapted to use other methods beyond BERT (other options and more recent LLMs are explored below). It can then be extended in domain areas/expertise, as well as scope conditions to include new languages, etc., via masking, fine-tuning, or other extensions. In this section, we illustrate how these three main tasks can be accomplished by the model before turning to specific comparisons to other models in the next section.

An example of the first task is the BC of news articles to determine their relevance to gun violence. Using a dataset comprising BBC news articles and the 20 Newsgroups corpus, we trained ConfliBERT to discern whether a given news item pertains to gun violence incidents. This fine-tuning task is significant for both domestic and international conflict studies, since it shows how to filter rapidly large volumes of news data about something like gun violence-related events. The ability to quickly identify relevant articles from a diverse news corpus can significantly enhance researchers' capacity to track and analyze (gun-related) conflicts in real-time.

For this BC task, ConfliBERT distinguishes between *gun violence-related and non-gun violence-related incidents*. Consider these examples:

Example

Input: "Two Lashkar e Jhangvi LeJ militants Asim alias Kapri and Ishaq alias Bobby confessed to killing four Rangers in Ittehad Town of Karachi, the provincial capital of Sindh."
Output: Gun Violence Related (1)
Input: "More than a week after a woman Communist Party of India-Maoist (CPI-Maoist) cadre was killed in an encounter in the forests of Lanjigarh block in Kalahandi District, the Maoists identified her as Sangita and called a bandh (general shutdown) in two Districts in protest against the killing." Output: Gun Violence Related (1)

The second task expands on this BC to a more nuanced multi-class classification of attack types. Employing the GTD to train ConfliBERT, it can classify attacks into nine distinct categories, including bombing/explosion, armed assault, assassination, and various forms of hostage-taking. Here are examples from the South Asia Terrorism Portal (SATP) dataset:

Example

Input: "Islamic State (IS) in the latest issue of its online magazine Dabiq claimed that the five of the nine Gulshan café attackers were suicide fighters... The mujahidin held a number of hostages as they engaged in a gun battle with apostate Bengali police and succeeded in killing and injuring dozens of disbelievers before attaining shahadah."
Output: Armed Assault
Input: "The ongoing construction work of an interstate bridge on Pranhita River on Maharashtra-Telangana border was thwarted by the Naxalites [Left Wing Extremists, LWEs] who set an excavator on fire and also damaged other equipment at the construction site at Gudem in Aheri taluka (revenue unit) of Gadchiroli District on April 26."
Output: Facility/Infrastructure Attack
Input: "Three boys sustained injuries when a landmine went off in Atmar Khel area of Baizai tehsil (revenue unit) in Mohmand Agency of Federally Administered Tribal Areas (FATA) on June 18."
Output: Bombing/Explosion

The third task ConfliBERT addresses is NER, crucial for extracting structured information from unstructured text, enabling more detailed and systematic analyses of conflict actors and targets. Using event reports (from MUC-4), which contain annotations of terrorism events, we fine-tune ConfliBERT to identify and classify entities, such as Organizations, Physical Targets, Victims, and Individuals. Here is an NER classification example using text from SATP:

https://doi.org/10.1017/pan.2025.10027

Example

Input: "A senior Muttahida Qaumi Movement (MQM)[ORG] worker identified as Sohail Rasheed[PERSON], 30, was shot dead near his home in Naeemabad[LOC] in Korangi Town[LOC] of Karachi[LOC], the provincial capital of Sindh[LOC], on June 19[DATE]."
Output:
Organization: Muttahida Qaumi Movement (MQM)
Victim: Sohail Rasheed Physical Target: Not specified
Location: Naeemabad, Korangi Town, Karachi, Sindh
Date: June 19

The versatility in these tasks suggests potential applications in fields, such as international relations, security studies, and public policy. Providing a tool that can simultaneously categorize events, identify key actors and targets, and filter relevant information from large text corpora, ConfliBERT offers a powerful means of analyzing the complex landscape of modern conflicts.

ConfliBERT was pre-trained in 2021 on data that at this point is nearly four years old (Hu *et al.* 2022). So a question is, how well does it do with more contemporaneous events and data? Consider the following example that has been processed using the interface at https://eventdata.utdallas.edu/conflibert-gui/ or https://huggingface.co/spaces/eventdata-utd/ConfliBERT-Demo:[10]

Example

Input: Former President Donald Trump, the 2024 presumptive Republican presidential nominee, was escorted off the stage by Secret Service after gunshots were fired at his rally in Butler, Pennsylvania. Mr. Trump was injured from the incident, with blood appearing on the right side of his face. This occurred two days before the start of the Republican National Convention in Milwaukee. The Butler County, Pennsylvania, district attorney told the Associated Press that a shooter was dead and a rally attendee was killed.
Output:
Organization: "secret service", "republican national convention", "the associated press", "district"
Person: "former president donald trump, the 2024 presumptive republican presidential nominee", "attorney"
Temporal: "two day"
Location: "butler, pennsylvania", "milwaukee", "butler county, pennsylvania"

The outputs for each of the coding tasks are:

Binary Classification for Political Violence "Positive: The text is related to conflict, violence, or politics (Confidence: 99.85%)."

Multilabel Classification "Armed Assault (Confidence: 98.40%)/Bombing or Explosion (Confidence: 5.39%)/Kidnapping (Confidence: 0.44%)/Other (Confidence: 0.95%)."

The only notable classification question is the placement of the "district attorney" as a person or organization. Such an error can easily be corrected with additional fine-tuning about legal actors and titles. This would then affect the comparability of later downstream performance metrics.

The scalability of this approach is particularly noteworthy. Once trained on these diverse tasks, ConfliBERT can be rapidly deployed to process large volumes of new data, enabling real-time or near-real-time analysis of emerging conflicts. This capability is invaluable for researchers and policymakers who need to quickly assess and respond to evolving conflict situations.

[10]"Former President Donald Trump Removed From Stage After Shots Fired at Pennsylvania Rally", CSPAN, July 13, 2024. Accessed 2024-09-09.

https://doi.org/10.1017/pan.2025.10027

5. Evaluating ConfliBERT versus Other LLMs

The focus here is on ConfliBERT's efficacy in the two critical NLP tasks of BC and NER *compared to recent developments like generative AI LLMs*. Gauging ConfliBERT's comprehension and extraction of information from conflict-related texts can be benchmarked against more recently created baselines from much larger LLMs like Gemma 2, LLama 3.1, and Qwen 2.5. The goal is to assess the quality of an LLM like ConfliBERT and compare it to larger, more costly, and more computationally expensive alternatives.

We do this initially for two datasets that were used in the earlier comparisons of ConfliBERT to BERT: the BBC News Dataset and re3d.[11] The BBC News dataset is used for the BC task (Greene and Cunningham 2006) and consists of 2,225 news articles, with 1,490 records for training and 735 for testing. The articles cover five categories: business, entertainment, politics, sport, and technology. For the conflict classification task, the dataset articles are relabeled as either conflict-related (1) or not conflict-related (0) by expert coders who analyzed each article's content and context. The BBC News dataset provides a diverse range of news articles, thus testing ConfliBERT's performance in sorting conflict-related content across various domains. The BC task mimics real-world scenarios where analysts must quickly identify political conflict-relevant information from a stream of news articles.

Once such articles or reports are identified via BC, political actor and action classification are the next relevant NER tasks. To compare ConfliBERT and more recent alternatives on this task, the re3d dataset is used (Relationship and Entity Extraction Evaluation Dataset https://github.com/dstl/re3d/). These data are specifically designed for defense and security intelligence analysis, focused on the conflict in Syria and Iraq, and providing domain-specific content across various source and document types with differing entity densities. The entities of interest include organizations, persons, locations, and temporal expressions. Ground-truth labels were established by annotators using a hybrid process.[12] The re3d dataset is valuable to evaluate ConfliBERT's and LLMs' extraction of relevant entities from conflict-related texts.

5.1. Methodology

Across the two datasets in this section, the performance of a task is done using versions of 1) ConfliBERT, 2) Meta's Llama 3.1 (8B), 3) Google's Gemma 2 (9B), and 4) ConflLlama (8B). Note that these are the most recent versions of these LLMs in mid-2024. To run the generative LLMs, we utilized the Ollama platform, which facilitates local model inference. This framework ensures that we are using the *instruction-tuned* variants of the models by default, which is the standard and appropriate choice for task-based applications like classification and NER. For computational efficiency on standard research hardware, Ollama deploys these models with 4-bit quantization. This practical choice significantly reduces memory usage but may also impact model precision. This methodological setup allows for a multi-faceted comparison. We evaluate our domain-specific, supervised extractive model (ConfliBERT) against:

- State-of-the-art generative LLMs used in a *zero-shot* capacity (Llama 3.1 and Gemma 2), reflecting a common and accessible workflow for researchers.
- A generative LLM that has also undergone supervised, domain-specific fine-tuning (ConfLlama), helping to distinguish the effects of model architecture versus training paradigm.

The comparisons conducted within the scope of this article therefore assess ConfliBERT's performance against both readily accessible LLMs and a more tailored generative counterpart. A general comparison of these different approaches is presented in Table 1.

[11] Both datasets are available for public use and can be accessed through the ConfliBERT GitHub repository.

[12] Described at https://github.com/dstl/re3d/. A modified version of the re3d dataset, which underwent several preprocessing steps, is used: These include tokenization and minor cleaning, removal of entity labels with confidence scores below 0.5, and resolution of overlapping entities by keeping only the largest span. The dataset was then converted to CoNLL 2003 format for compatibility with standard NER evaluation tools.

https://doi.org/10.1017/pan.2025.10027

Table 1. A comparison of extractive vs. generative LLMs across settings.

Setting	Extractive models	Generative models
Supervised	Trained on labeled data to classify/predict new data	Trained to generate text based on input
Zero-shot	Uses pre-learned rules on new data	Generates based on prompts without specific tuning
Pre-trained	Encoder model trained on large data	Uses generative models like GPT
Fine-tuned	Uses pre-trained model trained on task-specific data	Model tuned for specific generative tasks (e.g., QA)

The generative LLMs utilized in this article represent some of the most recent available models. The models and approaches include:

Meta's Llama 3.1 is the latest version of the Llama series of language models (Dubey *et al.* 2024). With 7 billion parameters, it strikes a balance between computational efficiency and performance. For the purpose of this comparison, we used the base model and a ZS approach.

Google's Gemma 2 has 9 billion parameters, represents a significant advancement in the field of LLMs (Team Gemma *et al.* 2024), offering robust performance across a wide range of NLP tasks while maintaining a relatively compact size. Similar to the Llama 3.1 comparison, we used a Gemma 2 base model, and applied a ZS approach.

Alibaba's Qwen 2.5 has a large pre-training corpus focused on math and coding. Another key improvement, especially in the context that we are using the model for, is the greater accuracy in generating structure outputs (as JSON objects). Qwen 2.5 was also utilized in its base model variant, using a ZS approach.

ConfLlama based on LlamA-3 8B, was specifically fine-tuned on the GTD using QLoRA with a learning rate of 2e-4 and LoRA rank of 8. The model was trained with gradient checkpointing enabled and 4-bit quantization, achieving convergence with loss reduction from 1.95 to approximately 0.90 (Meher and Brandt 2025). We employ both $Q4_{K_M}$ and $Q8_0$ quantizations for comprehensive performance analysis. Additional details about ConflLlama's architecture, training methodology, and prompt engineering are provided in Appendix C. In contrast to the base model variants of Llama 3.1, Gemma 2, and Qwen 2.5, the ConflLlama model is a fine-tuned model.

Various performance metrics quantify how well the models classify an event or its key attributes (e.g., actors, actions, locations, and dates). These metrics essentially compare the ground truth with what the machine extracts to produce a numerical result. This *distance* between the two demonstrates the degree of congruence, and the goal for event data scientists is to achieve 100% congruence across multiple possible sources of error (Althaus, Peyton, and Shalmon 2022; Brandt and Sianan 2025). For BC, the precision, recall, and the F_1 score are reported. The focus here is the F_1 statistic, the geometric mean of the precision and recall of the classifications, combining both attributes. The NER tasks are evaluated using token-level precision, recall, and macro F_1 score, which assesses the model's ability to correctly label each token (including the "O" tag for non-entities).

5.2. Binary Classification

To evaluate its performance, ConfliBERT was fine-tuned on the training split of the BBC News dataset for this BC task.[13] Table 2 shows the BC performance for the BBC News and re3d texts. ConfliBERT has high recall for conflict-related texts, suggesting a strong ability to identify relevant content. ConfliBERT's disparity between precision and recall for the conflict class indicates that it flags more texts as conflict-

[13]Our fine-tuning process follows the methodology outlined in our public repository, which utilizes the `Simple Transformers` library. Datasets are prepared in their required format (e.g., tab-separated for classification and CoNLL for NER), and models are trained on an NVIDIA A100 GPU with standard hyperparameters until convergence on a validation set.

https://doi.org/10.1017/pan.2025.10027

Table 2. Performance metrics for binary classifications of BBC texts.

Model	Class	Precision	Recall	F_1	Support
ConfliBERT	Conflict	0.5385	0.9245	0.6806	53
	Weighted Avg	0.9096	0.8571	0.8706	322
Gemma 2 (9B)	Conflict	0.2759	0.3019	0.2883	53
	Weighted Avg	0.7637	0.7547	0.7590	322
Llama 3.1 (8B)	Conflict	0.2923	0.3585	0.3220	53
	Weighted Avg	0.7730	0.7516	0.7614	322

Table 3. Performance metrics for named entity recognition of re3d texts.

Model	Class	Precision	Recall	F_1	Support
ConfliBERT	Micro Avg	0.4706	0.1956	0.2763	450
	Weighted Avg	0.4790	0.1956	0.2659	450
Gemma 2 (9B)	Micro Avg	0.4558	0.3556	0.3995	450
	Weighted Avg	0.4802	0.3556	0.4009	450
Llama 3.1 (8B)	Micro Avg	0.3863	0.3133	0.3460	450
	Weighted Avg	0.4052	0.3133	0.3489	450

related. Meanwhile, Gemma 2 and Llama 3.1 lack the nuanced understanding required for the specific task. Their performance remains poor and they consequently have lower F_1 scores. While Llama 3.1 is marginally better at detecting conflict-related content compared to Gemma 2, it still struggles significantly with this classification task.

Gemma 2 and Llama 3.1 show a bias towards classifying texts as non-conflict compared to ConfliBERT—evident from their poor performance on the conflict class. This imbalance suggests that general LLMs may overfit the majority class (non-conflict), potentially due to class imbalance in the training data or limitations in their ability to capture the nuanced features that distinguish conflict-related texts. The performance of Gemma 2 and Llama 3.1 shows that for a basic classification task, a domain-specific model that focuses on a local context is likely superior to a larger more general model when put to the same task. We turn to the issue of further fine-tuning the Llama, Gemma, and related models below.

5.3. Named Entity Recognition Results

For this task, the fine-tuned ConfliBERT model was compared against two general-purpose LLMs, Gemma 2 and Llama 3.1, using the re3d dataset. The models' performance was evaluated using a strict entity-level F_1 score, which requires a model to identify the exact span of tokens and the correct label for an entity to be considered a true positive. This provides a meaningful measure of practical performance than token-level accuracy. To ensure a fair comparison, the LLMs were guided by an instructed prompt that defined all valid entity types and specified a structured JSON output (see Appendix B).

The overall results of the models for re3d are summarized in Table 3. They reveal a critical trade-off between the comprehensive recall of the LLMs and the precision of the specialized model. While Gemma 2 achieves the highest weighted average F_1 score (0.4009), this is largely driven by its higher recall. ConfliBERT's strength lies in its precision and reliability. It achieved the highest precision score on key, frequent entities like Person and was perfect in its Money classifications. Most importantly, ConfliBERT exhibited perfect discipline by adhering strictly to the required entity schema, producing zero invalid labels.

In contrast, both generative LLMs failed to adhere to the explicit constraints of the prompt. Despite being provided with a definite list of valid categories, Gemma 2 "hallucinated" non-existent labels

https://doi.org/10.1017/pan.2025.10027

Table 4. Performance metrics for ConfliBERT, Llama 3.1, and Gemma 2 models.

Model	Task	Execution time (s)	GPU usage (%)
ConfliBERT	Classification	3.52	95.93
	NER	1.42	95.63
Llama 3.1 (8B)	Classification	575.23	92–94
	NER	489.39	92–94
Gemma 2 (9B)	Classification	730.14	90–97
	NER	866.23	90–97

like "Event" and "PhoneNumber," while Llama 3.1 invented a "Vehicle" category. This failure to follow instructions, even with a detailed prompt, makes them unreliable for automated coding systems where data integrity is paramount. Details of these other categories are in Appendix A.

While the LLMs are capable of finding more potential entities, their lack of discipline presents a significant challenge. For specialized domains like political conflict analysis, where precision and the reliability of the output schema are critical, ConfliBERT's performance represents a much stronger and more practical showing. It proves to be a more robust tool, effectively distinguishing signal from noise without introducing a new layer of error from fabricated categories.

5.4. Computational Performance Comparison

For both the BC task using the BBC News data and the NER task with the re3d, we recorded the execution time.[14] Table 4 presents the timings for each model and task combination. The most striking difference is the execution time: for classification, ConfliBERT took only 3.52 seconds, while Llama 3.1 (Gemma 2) took 575.23 (730.14) seconds. For NER, ConfliBERT completes the task in 1.42 seconds, compared to 489 (866) seconds for Llama 3.1 (Gemma 2). The speed of ConfliBERT can be attributed to its parallel architecture for processing input data. ConfliBERT's superior performance stems from its ability to process inputs in parallel. BERT-based models, like ConfliBERT, can efficiently batch multiple inputs and process them simultaneously. Generative LLMs, like Gemma and Llama, typically process inputs sequentially so each text requires a separate request to the model, introducing additional computational overhead. While we parallelize these models by batching multiple task requests, there are context-length constraints on processing the texts that differ across the models.

6. Classifying Texts about Terrorist Attacks

Some pre-training of the generative LLMs could bring their performance up to or exceeding the performance of ConfliBERT (see Wang 2024). Fine-tuning a model like ConfliBERT involves training task-specific parameters on top of the base text representations. This is a critical process for adapting the model to new domains or extending its capabilities. This asks for a comparison of ConfliBERT with more recent generative LLMs *with pre-training on political conflict texts*. Critical is replicability and service as a baseline comparison for event feature classification across new LLMs. This example replicates a common problem: one has identified political conflict-related texts (or prior dataset to be extended) and organized them (say in a CSV, JSON, or other database) for analysis with standard NLP to extract the relevant event information. This then leaves open the choices of the LLM and the pre-training. As an illustration, consider the short texts in the GTD (LaFree and Dugan 2007).[15] GTD is a

[14]These tests were conducted on a Macbook Pro with an Apple M2 Pro processor (12-core CPU and 19-core GPU) and 16 GB of unified memory. The software environment was macOS Sequoia 15.5 (build 24F74).

[15]One consideration in selecting this example is the need for shareable and non-copyrighted texts for replication. This removes a barrier to entry for replication that would exist if data requiring a copyright, extensive downloads (scraping), or a

https://doi.org/10.1017/pan.2025.10027

good choice because it 1) is a comprehensive open-source database of terrorist events, 2) contains the conflict classification tasks (what kind of attack is in the event?), 3) provides consistent, well-structured texts for NLP tasks, and 4) is classified by experts: one knows from the codebook and the dataset who perpetrated the terrorist attacks, the nature of the attacks and the types of victims. One cannot use these texts for the BC task, but they are suitable for evaluating models' NER and event multi-label classifications.

The task here is predicting the categorization of terrorist attacks from each GTD text description, comparing the various LLMs' codings to the original (human) GTD annotations of the terrorist attack types. ConfliBERT is compared to the aforementioned Llama and Gemma varieties, a larger LLM (Qwen 2.5), and a fine-tuned variant of Llama that we denote as ConflLlama.[16] The training prompts for the generative LLMs are given in Appendix B. The selection of evaluation metrics (ROC, accuracy, precision, recall, and F_1 score) follows standard practices in conflict event classification (Schrodt and Van Brackle 2012).

For testing and evaluation, GTD data from 1970 to 2016 are used to train the LLMs and they are tested with data from 2017 to 2020. Most of the GTD events have no texts for 1970–1997, so this is mainly based on training texts from 1998 to 2016. The LLM coded texts produce sets of BCs of each of the nine GTD event types across 37,709 texts recorded in GTD using each of the six models (ConfliBERT, ConflLlama4, ConflLlama8, Gemma, Llama, and Qwen). The first three of these are ones we produce, the latter three are "off the shelf" from Ollama.

6.1. Basic Classification Results

Figure 1 shows the comparative analysis of model performance differences across the LLM architectures. Here, the left column presents receiver-operator characteristic curves (ROCs) for the conflict-trained LLMs, while the right column presents the same for the general (non-conflict data trained), generative LLMs. The results in the right column are closer to a 45° line, indicating nearly random classification performance by event-type. The area under the curve (AUC) for each event type is in the lower right.[17] Across models, the higher accuracy of the ConfliBERT is evident and generally best for events about bombings and kidnappings (the green and gold lines) across the models—the most common kinds of attacks.

One criticism of only using an accuracy to compare models is that it is inflated by predicting the dominant class for imbalanced problems like the classifications here. Figure 2 shows the models' precision–recall curves, in parallel with Figure 1. The best precision–recall curves are those that follow the top, northeastern edge of the plot.[18] ConfliBERT has the highest precision–recall combinations for similar events (i.e., for the same colored lines). Of the larger generative AI models, only the more recent and much larger Qwen model comes close to ConfliBERT and ConflLlama in precision and recall performance, but only for kidnappings and bombings.

Precision–recall curves are a function of the cutoff used to classify a prediction as a match to GTD. Choosing the wrong cutoff, one may miss the benefits of a model to detect events (and mis-state their precision and recall in Figure 2). Figure 3 presents the F_1 score for the precision–recall as a function of the chosen cutpoint for the correct classification *for each event type*. Unlike Figures 1 and 2, these are grouped by types of events, so the colors used indicate the models here. Ideally, the values should be high across the cutpoints, like those for ConfliBERT. ConfliBERT and Qwen have the best F_1 scores, followed

license is used—e.g., Linguistic Data Consortium corpora. Further, we want something that is timely and relevant to conflict scholars. The short descriptions from GTD described here fit the bill.

[16] ConflLlama is a base Llama model given the training data and the classification prompt in Appendix C.

[17] AUC scores are relevant in conflict event classification due to the balance of precision and recall (Schrodt and Van Brackle 2012).

[18] The numeric precision and recall scores commonly seen in tables are weighted averages over the appropriate axes of these plots.

https://doi.org/10.1017/pan.2025.10027

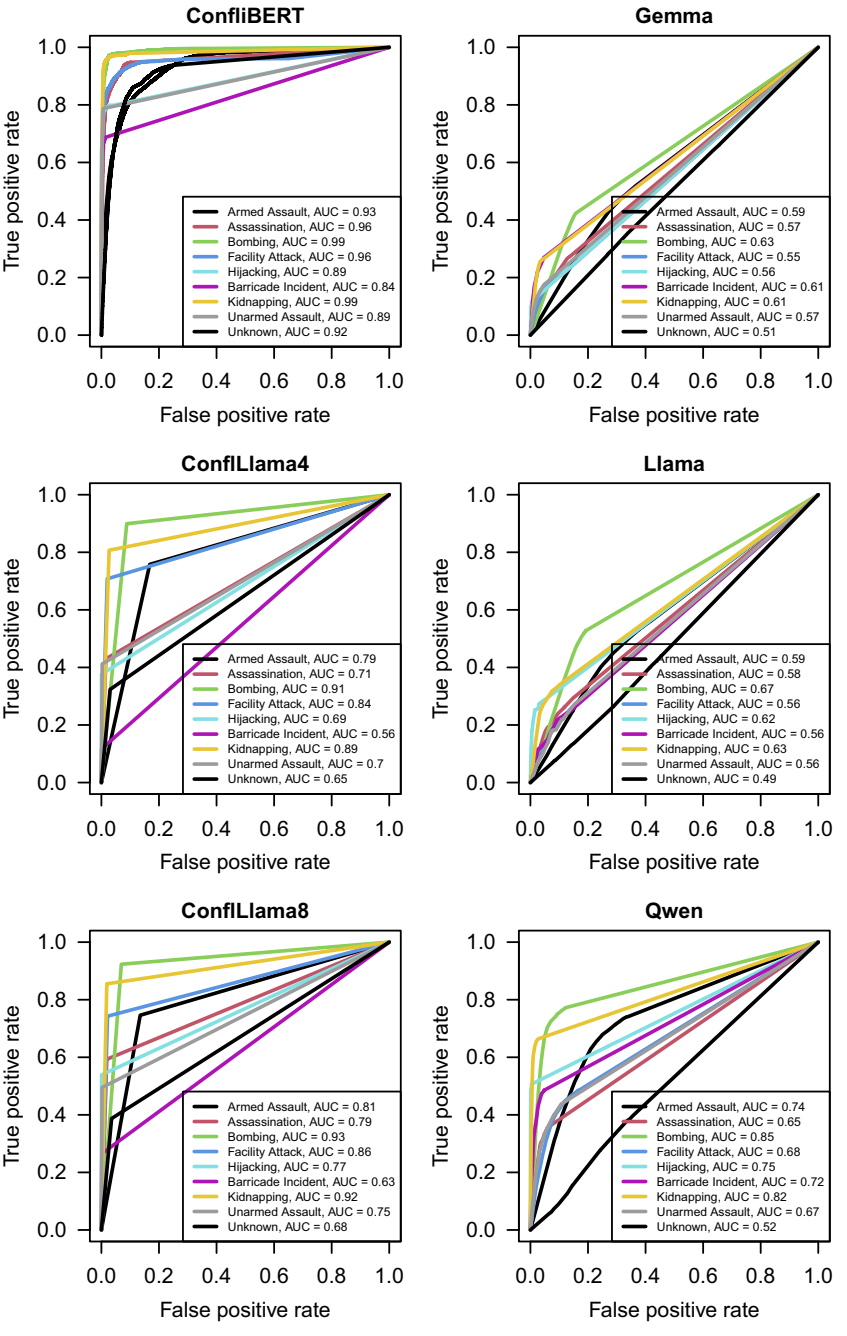

Figure 1. ROC and AUC for each LLM and event type.
Note: Curves along the northwestern edge are better.

by ConflLlama. These results align with previous findings suggesting that domain-specific fine-tuning often outperforms larger, general-purpose models (Gururangan *et al.* 2020). Like in other specialized domains, ConfliBERT's strong performance can be attributed to its training on conflict-related data.

https://doi.org/10.1017/pan.2025.10027

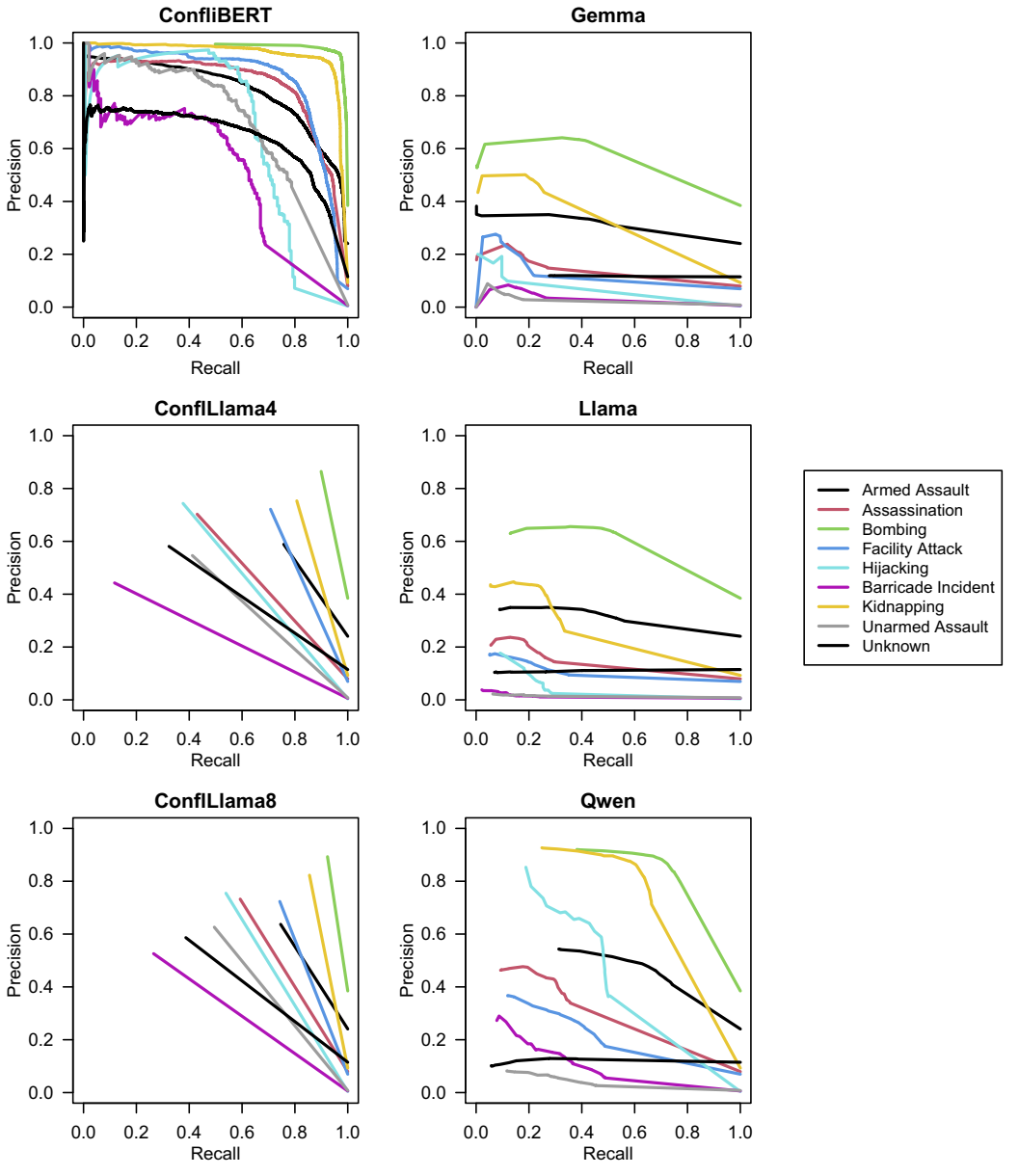

Figure 2. Precision–recall curves for each LLM and event type.
Note: Curves along the northeastern edge are better.

For the GTD conflict-related text analysis tasks, ConfliBERT outperforms the baseline competitors across all metrics as shown in aggregate in Table 5. Its considerable speed improvements over larger models also reflect broader trends in NLP research emphasizing the importance of computational efficiency (Schwartz *et al.* 2020).[19] For the fewer than 40K sentences evaluated here, this is remarkably fast, yet for a larger document processing-training problem, the more general generative LLMs like Gemma, Llama, and Qwen are likely computationally prohibitive. While general-purpose LLMs continue to improve, these results reinforce previous findings that specialized models can achieve

[19] Processing times were measured on identical hardware configurations to ensure fair comparison.

https://doi.org/10.1017/pan.2025.10027

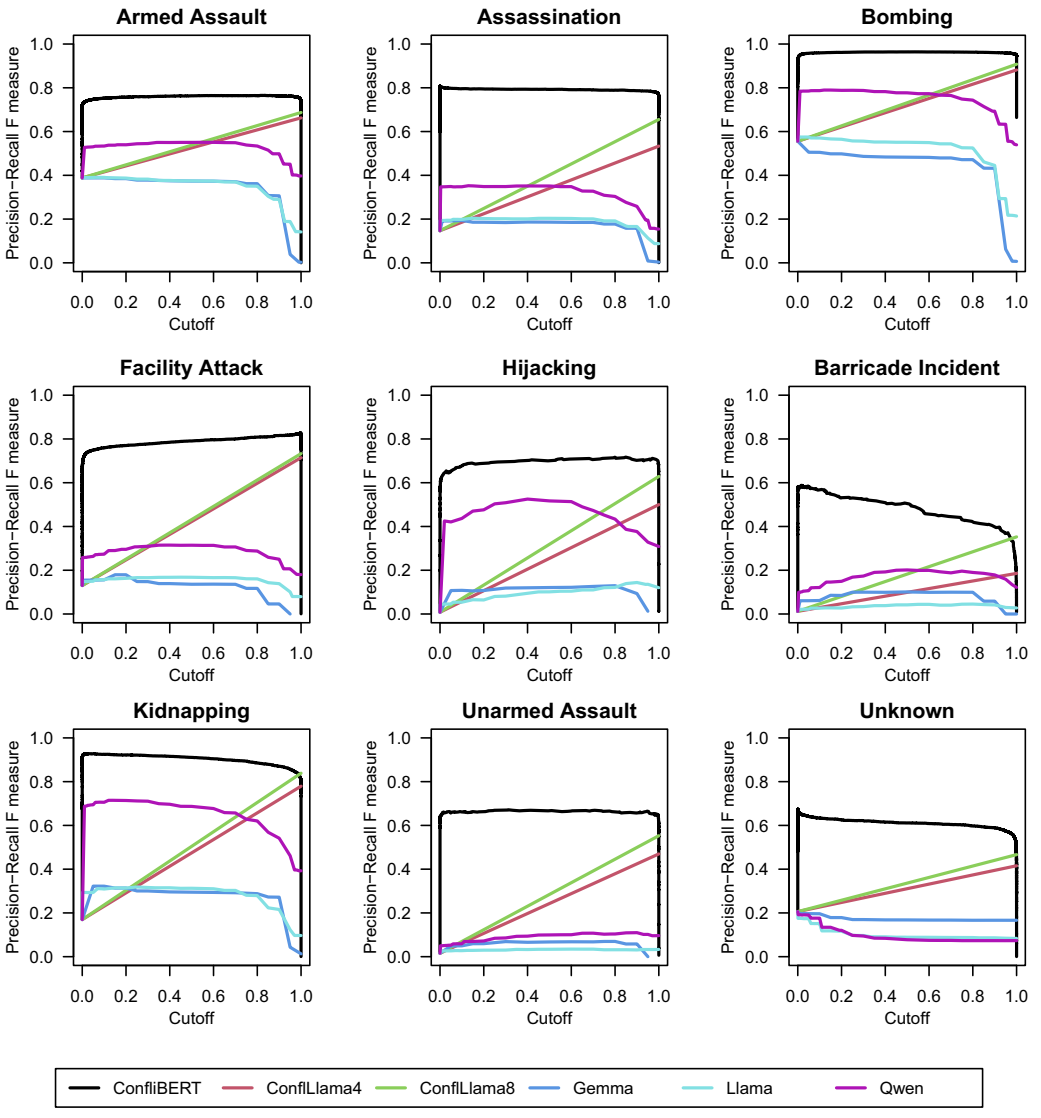

Figure 3. F_1 scores across cutoffs for each event type model.
Note: Higher curves are better.

superior performance in domain-specific tasks while maintaining significantly lower computational requirements (Strubell, Ganesh, and McCallum 2019).

6.2. *Multi-Label Classification Performance*

Incidents that involve more than one event type are documented with multi-label classifications in the GTD. This occurs say when an incident includes an armed attack or assault in the course of a kidnapping. Multi-label classification is important in conflict event coding, as real-world events often exhibit characteristics of multiple attack types (Radford 2021). Less than 10% of the post-2016 (the test period) data has multi-label events. Multi-label classification results, presented in Table 6, demonstrate ConfliBERT's ability to handle complex event categorizations. The model achieved a subset accuracy of 79.38% and the lowest Hamming loss (0.035), indicating superior performance in scenarios where events may belong to multiple categories. The close alignment between predicted label cardinality

https://doi.org/10.1017/pan.2025.10027

Table 5. Model performance comparison (macro averages).

Model	Accuracy	Precision	Recall	F_1	Total time	Time/Document	Relative speed
ConfliBERT	0.84	0.79	0.71	0.74	27.6 s	0.0016s	759.49×
ConflLlama-Q4KM* (8B)	0.73	0.66	0.54	0.57	49.9 m	0.1746 s	7.15×
ConflLlama-Q8* (8B)	0.77	0.70	0.62	0.65	52.3 m	0.1831 s	6.82×
Gemma 2 (9B)	0.33	0.26	0.21	0.21	3.1 h	0.6605 s	1.89×
Llama 3.1 (8B)	0.35	0.22	0.24	0.21	3.3 h	0.7191 s	1.74×
Qwen 2.5 (14B)	0.54	0.43	0.42	0.40	5.8 h	1.2490 s	1.00×

Note: *ConflLlama timing measurements were performed on Delta HPC resources and are not directly comparable to other models' timing metrics.

Table 6. Multi-label classification metrics.

Metric	ConfliBERT	ConflLlama-Q8	ConflLlama-Q4	Qwen	Gemma	LLaMA
Subset Accuracy (%)	79.38	72.40	68.80	50.99	30.70	32.03
Hamming Loss	0.035	0.052	0.061	0.096	0.133	0.148
Partial Match (%)	79.66	73.80	71.10	55.04	30.65	35.64
Label Cardinality						
True	0.963	0.963	0.963	0.963	0.963	0.963
Predicted	0.907	0.975	–	0.903	0.711	0.932

(0.907) and true label cardinality (0.963) suggests that the model has effectively learned to capture the multiple classification complexity of conflict events without over- or under-predicting.

The performance of ConfliBERT across all metrics suggests several important implications for conflict event classification. First, the results demonstrate that ConfliBERT with domain-specific fine-tuning can substantially outperform larger, general-purpose models, even when the latter have significantly more parameters (Gururangan *et al.* 2020). The model's strong performance on rare event types is particularly noteworthy, as it addresses a common challenge in conflict event classification. This suggests that the fine-tuning process successfully captures the nuanced characteristics of different attack types, even with limited training examples.

6.3. Validity Comparisons

Another assessment of the classification differences from the LLMs is to consider how their distributions change over the event types. At any one point in (recent) time, it may not be evident how the (mis-) classification of a given type of events affects inferences. But if there were systemic biases in LLM classification, they are more evident as more types over events are collected—an inherently time-series process for these data. This is particularly relevant in say a changepoint analysis of the drivers of trans-national terrorism like that addressed in Santifort, Sandler, and Brandt (2013, Figures 1–3), who use cumulative sums of terrorist event type classifications over time that would be severely biased upward or downward by LLM mis-classifications like those documented above.

Figure 4 shows the cumulative time series of the number of each type of GTD terrorist event from 2017 to 2020 as a dashed line. LLMs whose classifications are above this line are over-predicting/over-classifying the number of events of a given type, while those under the dashed line are the reverse. A few immediate patterns jump out: the non-conflict pre-trained LLMs under classify bombing events (uppermost-right plot)—so Gemma, Llama, and Qwen. Second, the Llama, Qwen, and Gemma models

https://doi.org/10.1017/pan.2025.10027

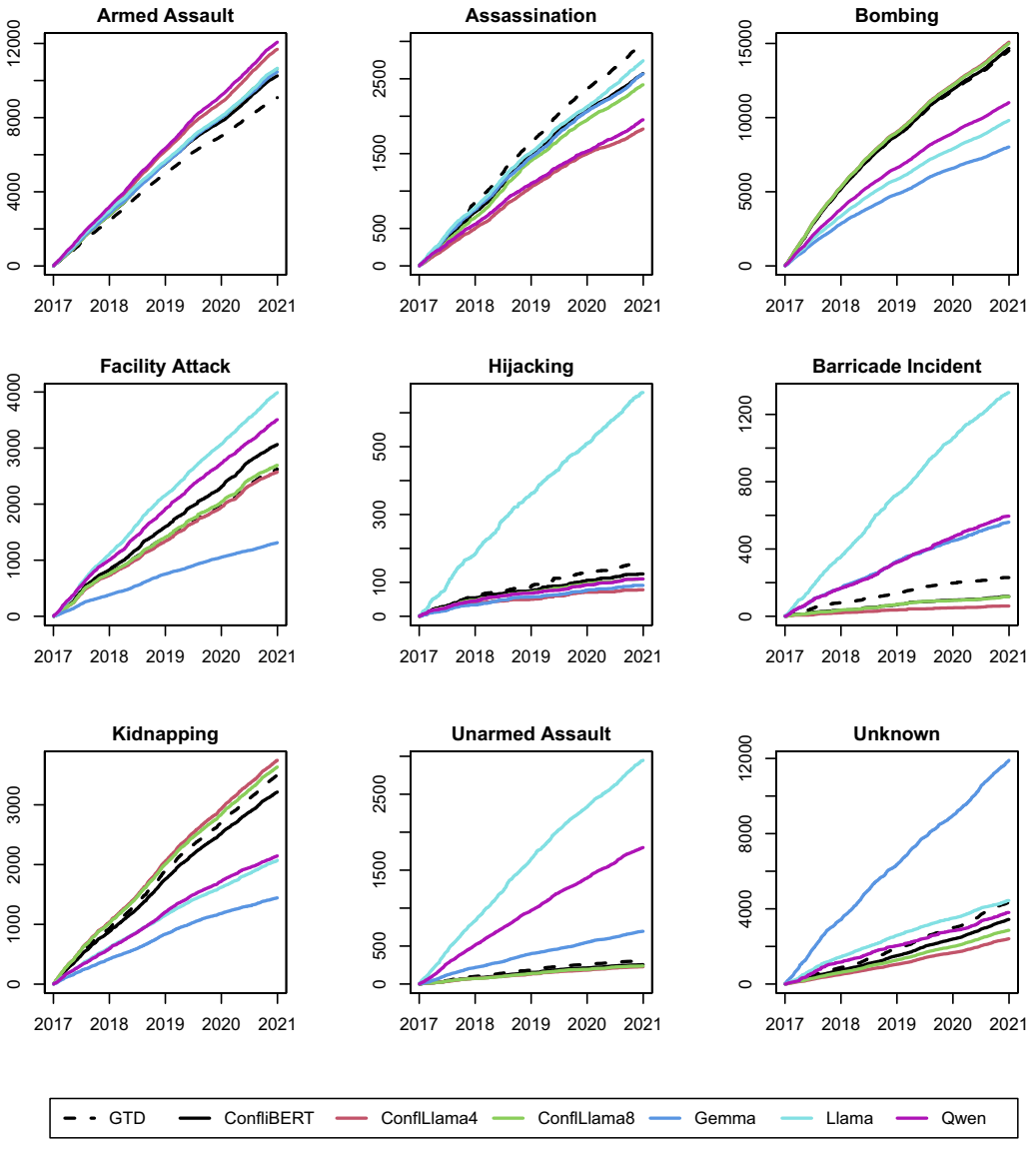

Figure 4. Cumulative number of predicted events, 2017–2021, by type and model.

generally do poorly with the rarest event types (hijackings, barricade incidents, and unarmed assaults), but their performance includes over and under predictions relative to GTD's human-coded data.

The reason such deviations and the relative performance of the LLMs matters here is that it will effect downstream time-series analyses since systematically mis-measured event types will lead to incorrect time-series dynamics and inferences that would confound those in works like Santifort *et al.* (2013). This builds on a key point since it shows that even using more data and more sophisticated methods for encoding texts, the issues of aggregation over time will still be important and affect inferences (Shellman 2004).

7. Conclusion

Conflict research and event data have a fruitful history of incorporating NLP approaches to advance methods of unstructured data processing (Beieler *et al.* 2016; Schrodt 2001, 2012). This work continues along that path. The adoption of NLP techniques like those employed by ConfliBERT improves how

https://doi.org/10.1017/pan.2025.10027

political scientists can extract and study events and political interactions. These tools offer the potential to analyze larger volumes of text data, enabling more comprehensive studies. They can reduce bias in event coding by applying consistent criteria across large datasets, identify patterns and trends that might be missed by human coders, and enable near-real-time analysis of political events as they unfold. While LLMs generally hold this potential, the approach in ConfliBERT incorporates domain-specific knowledge, resulting in superior performance and even faster data processing for text classification and summarization tasks.

There are a series of conclusions to be drawn from this analysis. The results leverage the existing infrastructure of BERT-alike LLMs and conflict researchers' expertise to advance scholarship on conflict processes and international security. The contribution is that domain-specific knowledge—the things international and civil conflict scholars know—should be part of the information extraction process used to 1) filter relevant reports (BC), 2) identify events, and 3) annotate their attributes (NER). A BERT-based model *plus* domain knowledge is able to do this in a way that is better on several metrics as documented in Section 5.

ConfliBERT has several advantages over comparable contemporaneous methods for machine coding events. First, it is easily deployed and replicable as a method since it is open source and can be deployed on conventional hardware. Second, it is significantly better on comparable, relevant quality metrics and faster than rival or even newer generative AI methods that used decoder technologies with graphical processing units (GPUs). Third, it can be rapidly deployed to detect new event data and their characteristics.[20] This means it can be tuned and adjusted as needed for new cases, data, and texts. This allows users to improve the extraction, coverage (geographically, and as we show, linguistically), across new data and training domains. Fourth, this means that additional downstream tasks, such as recoding texts, extracting additional variables or features, etc., are all much faster and easier than what has historically been the case. We show this in our examples, where differences are seen in the classifications of terrorist event types in the GTD dataset across the LLMs. ConfliBERT and domain-specific models provide much better results compared to generalist LLMs like Gemma, Llama, and Qwen. Fifth, ConfliBERT continues to maintain its superior performance when compared to the most recent encoder models, such as ModernBERT (see Appendix D).

Beyond an infrastructure outline for political scientists to engage with texts about conflict and violence, there are several other contributions of note here. First, ConfliBERT builds on a known ontology (CAMEO/PLOVER) (Schrodt 2012) for coding events and provides a set of tools for continuing to do so. This allows for additional fine-tuning of the models and a flatter development and learning curve. Unlike current large-scale general LLMs, this allows researchers to openly and quickly work in this area (the span from ConfliBERT in Hu *et al.* (2022) to the recent paper by Osorio *et al.* (2024) is less than 36 months.)

Second, the typical social science conflict researcher need not build their own ConfliBERT: one can fine-tune or extend this model since it is open and available for use via our website and Hugging Face. About 200 GB of combined training data are invested in ConfliBERT, ConfliBERT Spanish, and ConfliBERT Arabic. Additional classifications and training based on new ideas, texts, actors, etc., can be added and evaluated. We have done this in the efforts to extend beyond event coding just in English by working not just with a language and domain-specific dictionary approach (Osorio and Reyes 2017), but a general BERT-like model in Spanish (Yang *et al.* 2023) and Arabic (Alsarra *et al.* 2023). This shows how the domain-specific approaches can bring old codebooks and ontologies into the LLMs (Hu *et al.* 2024). So prior domain knowledge about regions, languages, and events can be part of how LLMs are used to encode and understand new texts and data.

Third, this approach is sometimes better than using larger LLMs. Unlike large generative LLMs, an encoder model like ConfliBERT better fits what a social scientist needs, which is data extraction, organization, and (predictive) classification. Section 5 shows this in terms of performance metrics like accuracy, precision, F_1, etc. It is also faster to use ConfliBERT. While the initial LLM training for

[20] We thank a reviewer/commentators for emphasis on this point.

https://doi.org/10.1017/pan.2025.10027

ConfliBERT and its language variants took thousands of GPU hours, the work in Section 5.4 only takes hours of computing time on current laptops. Deploying this on a real data problem is scalable and feasible: it is 300–400 times faster than using a proprietary LLM for NER and 150–200 times faster for BC.

Finally, there are problems that can be addressed, such as learning about and connecting events and actors. One area of interest is extending ontologies and NER to recognize and learn about new events and actors—who is the next leader, insurgent, or what are they doing? This is related to a literature on continual learning and catastrophic forgetting in LLMs. There is work in this area that can be applied and used to aid models like ConfliBERT as well (e.g., Li *et al.* 2022). This would also be useful for extending text-as-data methods across networks of texts, languages, etc.

Appendix A. NER Performance by Entity Type for re3d

Table A1. Full per-class performance metrics for named entity recognition models for re3d.

Model	Entity class	Precision	Recall	F_1	Support
ConfliBERT					
	DocumentReference	0.0000	0.0000	0.0000	5
	Location	0.4800	0.2105	0.2927	114
	MilitaryPlatform	0.1429	0.2500	0.1818	4
	Money	1.0000	1.0000	1.0000	2
	Nationality	0.0000	0.0000	0.0000	6
	Organisation	0.4394	0.1503	0.2239	193
	Person	0.6471	0.1692	0.2683	65
	Quantity	0.4444	0.4706	0.4571	17
	Temporal	0.5833	0.2059	0.3043	34
	Weapon	0.4000	0.6000	0.4800	10
	Micro Avg	*0.4706*	*0.1956*	*0.2763*	*450*
	Macro Avg	*0.4137*	*0.3056*	*0.3208*	*450*
	Weighted Avg	*0.4790*	*0.1956*	*0.2659*	*450*
Gemma 2 (9B)					
	DocumentReference	0.0000	0.0000	0.0000	5
	Event	0.0000	0.0000	0.0000	0
	Location	0.6122	0.5263	0.5660	114
	MilitaryPlatform	0.3333	0.2500	0.2857	4
	Money	0.0000	0.0000	0.0000	2
	Nationality	0.1200	0.5000	0.1935	6
	Organisation	0.3008	0.1917	0.2342	193
	Person	0.6389	0.3538	0.4554	65
	PhoneNumber	0.0000	0.0000	0.0000	0
	Quantity	0.3333	0.4706	0.3902	17

(continued)

https://doi.org/10.1017/pan.2025.10027

Table A1. Continued.

Model	Entity class	Precision	Recall	F_1	Support
	Temporal	0.9259	0.7353	0.8197	34
	Weapon	0.7500	0.3000	0.4286	10
	Micro Avg	*0.4558*	*0.3556*	*0.3995*	*450*
	Macro Avg	*0.3345*	*0.2773*	*0.2811*	*450*
	Weighted Avg	*0.4802*	*0.3556*	*0.4009*	*450*
Llama 3.1 (8B)					
	DocumentReference	0.0909	0.2000	0.1250	5
	Location	0.5644	0.5000	0.5302	114
	MilitaryPlatform	0.0625	0.2500	0.1000	4
	Money	0.0000	0.0000	0.0000	2
	Nationality	0.0000	0.0000	0.0000	6
	Organisation	0.3070	0.1813	0.2280	193
	Person	0.3158	0.2769	0.2951	65
	Quantity	0.2308	0.1765	0.2000	17
	Temporal	0.8750	0.6176	0.7241	34
	Vehicle	0.0000	0.0000	0.0000	0
	Weapon	0.3846	0.5000	0.4348	10
	Micro Avg	*0.3863*	*0.3133*	*0.3460*	*450*
	Macro Avg	*0.2574*	*0.2457*	*0.2397*	*450*
	Weighted Avg	*0.4052*	*0.3133*	*0.3489*	*450*

Appendix B. LLM Prompts

ConfliBERT and ConflLlama are fine-tuned specifically for terrorist event classification, without explicit prompting for output classifications given input event texts. For the general-purpose LLMs (e.g., Gemma, Qwen, and Llama), the following structured prompt is used:

Listing 1. Prompt for multi-label event classification.

```
Classify each of the following events into up to three of these categories,
    ↪ providing probabilities for each:
Assassination, Armed Assault, Bombing/Explosion, Hijacking,
Hostage Taking (Barricade Incident), Hostage Taking (Kidnapping),
Facility/Infrastructure Attack, Unarmed Assault, Unknown

For each event, return only a JSON object with category names as keys and
    ↪ probabilities as values.

Example format:
{"Armed Assault": 0.7, "Bombing/Explosion": 0.2, "Unknown": 0.1}

Events:
```

We follow key principles on effective LLM prompting (Liu, Zhang, and Gulla 2023; Wei *et al.* 2022). Its structured format with explicit probability requirements builds on research showing that quantitative outputs improve model classification tasks (Brown *et al.* 2020). The multi-label approach, allowing up

https://doi.org/10.1017/pan.2025.10027

to three categories, reflects the complex classification task and the original GTD structure—allowing direct comparisons. The JSON output format facilitates consistent parsing and evaluation, addressing challenges in systematic event coding. This standardization enables direct comparison with both human annotations and across models, while maintaining interpretability.

For the NER task, we employed a similar structured approach:

Listing 2. Prompt for named entity recognition (NER).

```
You are an expert in Named Entity Recognition (NER) for analyzing texts about
    ↪ political conflict and events. Your task is to identify and extract all
    ↪ named entities from the user's text according to the provided entity
    ↪ definitions.

Entity Definitions:
- Organisation: A formal group of people (e.g., "United Nations").
- Person: A specific individual's name (e.g., "Carter").
- Location: A geographical place (e.g., "Geneva", "Iraq").
- Weapon: A specific type of weapon (e.g., "Javelin missile").
- Nationality: An adjective describing origin (e.g., "Ukrainian").
- Temporal: A phrase indicating a time or date (e.g., "next week").
- DocumentReference: A reference to a specific document (e.g., "Resolution
    ↪ 242").
- Money: A specific monetary value (e.g., "$10 million").
- Quantity: A number and a unit that is not money (e.g., "50 kilograms").
- MilitaryPlatform: A major military asset (e.g., "HMS Ocean").

Output Instructions:
Return a single JSON object with a key "entities" containing a list of objects
    ↪ . Each object must have keys "entity_text" (the exact text) and "
    ↪ entity_label" (the corresponding label, without B- or I- prefixes).
Example:
Text: "The Taliban attacked Kabul with rockets last Tuesday."
Output:
{
  "entities": [
    {"entity_text": "The Taliban", "entity_label": "Organisation"},
    {"entity_text": "Kabul", "entity_label": "Location"},
    {"entity_text": "rockets", "entity_label": "Weapon"},
    {"entity_text": "last Tuesday", "entity_label": "Temporal"}
  ]
}
User Text to Analyze:
"{text}"
```

For BC, we used this simplified format:

Listing 3. Prompt for binary conflict classification.

```
You are an expert text classifier. Your task is to classify the following text
    ↪  as either 'Conflict' or 'Not Conflict'.

'Conflict' refers to texts about war, violence, political unrest, or
    ↪ significant social tensions.

Your response MUST be a single JSON object and nothing else. Do not add
    ↪ explanations or markdown formatting.

Example format:
{"classification": "Conflict"}

Text to classify:
"{text}"
```

https://doi.org/10.1017/pan.2025.10027

Appendix C. ConflLlama: Implementation Details

C.1. Fine-Tuning Approach

ConflLlama employs a supervised fine-tuning approach on GTD data, but importantly, it uses a generative text completion methodology rather than a classification head. Unlike ConfliBERT's encoder-based classification framework, ConflLlama was fine-tuned to generate attack type labels directly as text using a next-token prediction objective—this is distinct from both additional pretraining and instruction fine-tuning approaches.

The model was fine-tuned using parameter-efficient fine-tuning (PEFT), specifically through quantized low-rank adaptation (QLoRA). In this approach, the entire base Llama-3 8B model remains frozen and quantized to 4-bit precision, with only the low-rank adaptation matrices being trainable. These adaptation matrices were applied to specific components of the transformer architecture: query, key, and value projections in the attention mechanism (q_proj, k_proj, v_proj, and o_proj), as well as gate projections and feed-forward components (gate_proj, up_proj, and down_proj). The LoRA adaptation used a rank of 8 and an alpha scaling factor of 16, with no dropout applied during training. This configuration resulted in training only approximately 0.5% of the model's parameters (roughly 41.9 million parameters), substantially reducing computational requirements while allowing effective domain adaptation.

The fine-tuning objective utilized standard language modeling cross-entropy loss over the output tokens (next-token prediction), focusing on the tokens in the "Attack Types:" section of our template. We did not implement any custom classification-specific loss functions or add classification heads to the model. Training progress was monitored through the language modeling loss, which showed convergence from an initial value of approximately 1.95–0.90 over the course of training. Unlike classifier models, where metrics such as accuracy or F_1 score might be tracked during training, our generative approach meant that the primary training signal was the language modeling loss itself, with classification metrics calculated only during evaluation phases.

Our training implementation used an AdamW optimizer with 8-bit quantization, a learning rate of 2e-4 with linear decay, and gradient accumulation steps of 8 for an effective batch size of 8. Memory efficiency was further enhanced through gradient checkpointing, allowing the model to fit within the constraints of a single A100 40 GB GPU while maintaining performance. The model was trained for 1,000 steps, which was sufficient for convergence on the GTD dataset.

C.2. Prediction Methodology

Unlike traditional classification models that output probability distributions over fixed classes, ConflLlama generates the attack type labels as actual text strings. The prediction process works through a structured format where the event description is provided within a prompt template, after which the model generates text to complete the prompt. The generated text is then parsed to extract the predicted attack type labels, which are compared with the ground truth for evaluation purposes.

This generative approach offers significant advantages for multi-label classification tasks. By producing text rather than class probabilities, ConflLlama naturally accommodates cases where multiple attack types apply to a single event. Furthermore, this methodology potentially allows the model to adapt to new classification schemes through additional fine-tuning, as it does not rely on a fixed classification architecture with predetermined output classes.

C.3. Prompt Templates

C.3.1. Training and Evaluation Prompt

For consistent training and evaluation, we employed a structured prompt template that clearly delineates between the input event description and the expected output classification. The template takes the following form:

https://doi.org/10.1017/pan.2025.10027

```
Below describes details about terrorist events.
>>> Event Details:
{summary}
>>> Attack Types:
{combined_attacks}
```

During the training phase, both the event details and attack types were provided to the model, allowing it to learn the association between descriptions and classifications. During evaluation, only the event details were provided, and the model generated the attack types based on its fine-tuned knowledge.

C.3.2. Prompt Engineering

The selection of an appropriate prompt template is crucial for model performance. We examined multiple prompt variations to identify the most effective format. Alternative formulations included a direct classification request: "Classify the following terrorist event into its attack type(s): Event: {summary} Attack Type(s):" as well as a more elaborate expert-framed request: "You are an expert in terrorism analysis. Based on the following event description, identify all applicable attack types from the GTD schema: {summary}."

Through empirical testing, we observed only marginal performance differences between these prompts, with variances of approximately ±1.5% in F_1 score. Notably, after fine-tuning was complete, the exact prompt wording had substantially less impact than would be expected in ZS models, as the language model had already adapted to the fundamental task structure during training.

Appendix D. ConfliBERT versus ModernBERT

To evaluate whether newer-generation BERT architectures offer performance advantages for terrorism event classification, we fine-tuned ModernBERT on the same GTD used for ConfliBERT training, creating[21] Confli-mBERT.

ModernBERT offers several advantages over other BERT architectures: a larger pre-training corpus (2 trillion tokens), modern architectural improvements (Rotary Positional Embeddings and Local-Global Alternating Attention), enhanced long-context understanding (8,192 tokens native window),[22] and more efficient inference through Flash Attention. Despite these theoretical advantages, ConfliBERT consistently outperformed Confli-mBERT across most metrics:

Table D1. Overall performance metrics.

Metric	ConfliBERT	Confli-mBERT	Difference
Overall Accuracy	84.04%	79.66%	+4.38%
Average F_1 (all types)	0.7441	0.6001	+0.1439
Average AUC (all types)	0.9304	0.7777	+0.1527

The performance gap between models showed a strong negative correlation with class prevalence ($r = -0.83$), with ConfliBERT demonstrating significantly better handling of rare attack types:

[21] The model can be accessed through Hugging Face.

[22] ModernBERT can process documents with a larger context window of 8,192 tokens compared to BERT's 512 tokens, previous research by Pappagari *et al.* (2019), Park, Vyas, and Shah (2022), and Osorio *et al.* (2025) have applied a chunking strategy for longer documents. After splitting the document into segments of 512 tokens each, these authors then independently classified each segment and applied majority voting to derive final labels. Confidence scores for each of the processed segments were then averaged across all document segments, thereby permitting the processing tasks on documents that exceed the 512 token window of ConfliBERT.

https://doi.org/10.1017/pan.2025.10027

Table D2. Performance on rare vs. common attack types (F_1 score).

Attack type	Prevalence	ConfliBERT	Confli-mBERT	Difference
Hijacking	0.4%	0.7000	0.3653	+0.3347
Hostage taking (barricade)	0.6%	0.4971	0.1516	+0.3455
Unarmed assault	0.8%	0.6667	0.3137	+0.3529
Facility/infrastructure attack	7.0%	0.7901	0.7666	+0.0235
Assassination	7.9%	0.7924	0.6552	+0.1373
Hostage taking (kidnapping)	9.3%	0.9111	0.9067	+0.0045
Unknown	11.5%	0.6125	0.5889	+0.0236
Armed assault	24.1%	0.7630	0.7150	+0.0481
Bombing/explosion	38.5%	0.9637	0.9383	+0.0254

This analysis demonstrates that for specialized domain classification with significant class imbalance, domain-specific pre-training is more valuable than general language understanding. ConfliBERT's architecture is better suited for handling rare event classes, and newer language model architectures do not automatically translate to better performance on specialized classification tasks. Class imbalance handling appears more important than model size or architectural sophistication. These findings challenge the assumption that newer, larger language models inherently perform better across all tasks, emphasizing the continued importance of domain-specific models for specialized applications.

Acknowledgments. Previous versions have been presented at the virtual Methods in Event Detection Colloquium (September 2024), TexMeth 2025 at the University of Houston (February 2025), the 66th Annual Convention of the International Studies Association, Chicago, Illinois (March 2025), and the APSA Virtual Research Group on Advancing the Use of Computational Tools in Political Science (April 2025). Thanks for the suggestions and feedback go to Scott Althaus, R. Michael Alvarez, Ben Bagozzi, Mike Colaresi, Rebecca Cordell, Ryan Kennedy, Hyein Ko, Shahryar Minhas, Philip Schrodt, Nora Webb Williams, Chris Wlezien, and the PA editors and reviewers.

Data Availability Statement. Replication code for this article is available at Brandt *et al.* (2025). Replication data and code can be found at Harvard Dataverse: https://doi.org/10.7910/DVN/KDO5AM.

Author Contributions. Conceptualization: All authors. Methodology: P.B., S.A., V.D., D.H., L.K., S.M., and J.O., Data curation: S.A. and S.M. Data visualization: P.B. Funding acquisition: P.B., S.A., V.D., L.K., and J.O. Writing original draft: P.B., S.A., V.D., D.H., S.M., and M.S. All authors approved the final submitted draft.

Funding Statement. This research was supported by grants from the U.S. NSF 2311142; used Delta at NCSA/University of Illinois through allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF 2138259, 2138286, 2138307, 2137603, and 2138296. This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII). S.A. would like to extend his appreciation to the Deanship of Scientific Research at King Saud University for funding his work through the ISPP Program (ISPP25-16). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, NCSA, or the affiliated university resource providers.

Competing Interests. The authors declare none.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Alsarra, S., et al. 2023. "ConfliBERT-Arabic: A Pre-Trained Arabic Language Model for Politics, Conflicts and Violence." In (R. Mitkov and G. Angelova eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, edited by R. Mitkov and G. Angelova, 98–108. Shoumen, Bulgaria: INCOMA Ltd..

https://doi.org/10.1017/pan.2025.10027

Althaus, S., B. Peyton, and D. Shalmon. 2022. "A Total Error Approach for Validating Event Data." *American Behavioral Scientist* 66 (5): 603–624.

Araci, D. 2019. "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models." Preprint, arXiv:1908.10063 [cs.CL]. https://arxiv.org/abs/1908.10063

Barrie, C., A. Palmer, and A. Spirling. 2024. "Replication for Language Models Problems, Principles, and Best Practice for Political Science." https://arthurspirling.org/documents/BarriePalmerSpirling TrustMeBro.pdf

Beieler, J., P. T. Brandt, A. Halterman, E. Simpson, and P. A. Schrodt. 2016. "Generating Political Event Data in Near Real Time: Opportunities and Challenges." In *Computational Social Science*, edited by R. M. Alvarez. New York NY: Cambridge University Press.

Beltagy, I., K. Lo, and A. Cohan. 2019. "SciBERT: A Pretrained Language Model for Scientific Text." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), edited by K. Inui, J. Jiang, V. Ng, and X. Wan, 3615–3620. Hong Kong, China: Association for Computational Linguistics.

Boschee, E., J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward. 2015. "ICEWS Coded Event Data." Harvard Dataverse 12. https://doi.org/10.7910/DVN/28075

Brandt, P. T., and M. Sianan. 2025, "Measurement of Event Data from Text." *Frontiers in Political Science* 6: 1–13. https://doi.org/10.3389/fpos.2024.1453640

Brandt, P. T., et al. 2025. "Replication materials for "Extractive versus Generative Language Models for Political Conflict Text Classification"." https://doi.org/10.7910/DVN/KDO5AM

Brown, T., et al.. 2020. "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, vol. 33, 1877–1901. Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Burnham, M., K. Kahn, R. Y. Wang, and R. X . Peng. 2024. "Political Debate: Efficient Zero-Shot and Few-Shot Classifiers for Political Text." Preprint, arXiv:2409.02078.

Chalkidis, I., M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. 2020. "LEGAL-BERT: The Muppets Straight Out of Law School." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, edited by T. Cohn, Y. He, and Y. Liu, 2898-2904. Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.261/

Croicu, M. 2024. "Deep Active Learning for Data Mining from Conflict Text Corpora." Preprint, arXiv:2402.01577 [cs.CY]. https://arxiv.org/abs/2402.01577

Croicu, M., and K. Eck. 2022. "Reporting of Non-fatal Conflict Events." *International Interactions* 48 (3): 450–470.

Croicu, M., and S. P. von der Maase. 2025. "From Newswire to Nexus: Using Text-Based Actor Embeddings and Transformer Networks to Forecast Conflict Dynamics." Preprint, arXiv:2501.03928 [cs.CY]. https://arxiv.org/abs/2501.03928

D'Orazio, V., S. T. Landis, G. Palmer, and P. Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22 (2): 224–242. https://doi.org/10.1093/pan/mpt030

Dai, Y., B. Radford, and A. Halterman. 2022. "Political Event Coding as Text-to-Text Sequence Generation." In (A. Hürriyetoğlu, H. Tanev, V. Zavarella, and E. Yörük), *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, edited by A. Hürriyetoğlu, H. Tanev, V. Zavarella, and E. Yörük, 117–123. Abu Dhabi, UAE: Association for Computational Linguistics.

Software, Defense Advanced Research Projects Agency, and Intelligent Systems Technology Office. 1992. *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*. San Mateo, CA: Morgan Kaufmann. https://aclanthology.org/M92-1000

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* vol. 1, edited by J. Burstein, C. Doran, and T. Solorio, 4171–4186. Minneapolis, MN: Association for Computational Linguistics. https://aclanthology.org/N19-1423/

DSTL. 2018. "Relationship and Entity Extraction Evaluation Dataset." Accessed: 2024-10-23. https://github.com/dstl/%20re3d/

Dubey, A., et al. 2024. "The Llama 3 Herd of Models." Preprint, arXiv:2407.21783.

Gerner, D., R. Jabr, and P. Schrodt. 2002. *Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions*. New Orleans: International Studies Association.

Gerner, D. J., P. A. Schrodt, R. A. Francisco, and J. L. Weddle. 1994. "Machine Coding of Event Data Using Regional and International Sources." *International Studies Quarterly* 38 (1): 91–119.

Greene, D., and P. Cunningham. 2006. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering." In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, edited by W. Cohen and A. Moore, 377–384. New York, NY: Association for Computing Machinery.

Grimmer, J., M. E. Roberts, and B. M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton University Press.

Gupta, T., N. M. Mohd Zaki, A. Krishnan, and Mausam. 2022. "MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction [in Eng]." *NPJ Computational Materials (London)* 8 (1): 1–11.

https://doi.org/10.1017/pan.2025.10027

Gururangan, S., et al. 2020. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, 8342–8360. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.740/

Häffner, S., M. Hofer, M. Nagl, and J. Walterskirchen. 2023. "Introducing an Interpretable Deep Learning Approach to Domain-Specific Dictionary Creation: A Use Case for Conflict Prediction." *Political Analysis* 31 (4): 481–499.

Halterman, A. 2017. "Mordecai: Full Text Geoparsing and Event Geocoding." *Journal of Open Source Software* 2 (9): 91.

Halterman, A., B. E. Bagozzi, A. Beger, P. Schrodt, and G. Scarborough. 2023. "PLOVER and POLECAT: A New Political Event Ontology and Dataset." *SocArXiv*.

Halterman, A., and B. J . Radford. 2021. "Few-Shot Upsampling for Protest Size Detection." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, edited by C. Zong, F. Xia, W. Li, and R. Navigli, 3713–3720. Association for Computational Linguistics. https://aclanthology.org/2021.findings-acl.325/.

Halterman, A., P. A. Schrodt, A. Beger, B. E. Bagozzi, and G. I. Scarborough. 2023. "Creating Custom Event Data Without Dictionaries: A Bag-of-Tricks." Preprint, arXiv:2304.01331 [cs.CL]. https://arxiv.org/abs/2304.01331

Hu, Y., et al. 2022. "ConfliBERT: A Pre-Trained Language Model for Political Conflict and Violence." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by M. Carpuat, M.-C. de Marneffe, and I. V. M. Ruiz, 5469–5482. Seattle: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.400

Hu, Y., E. S. Parolin, L. Khan, J. Osorio, and V. 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)*, edited by L.-W. Ku, A. Martins, and V. Srikumar, 583–603. Bangkok: Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.35

Huang, K., J. Altosaar, and R. Ranganath. 2020. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission." Preprint, arXiv: 1904.05342 [cs.CL]. https://arxiv.org/abs/1904.05342

Huang, R., I. Cases, D. Jurafsky, C. Condoravdi, and E. Riloff. 2016. "Distinguishing Past, on-Going, and Future Events: The EventStatus Corpus." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, edited by J. Su, K. Duh, and X. Carreras, 44–54. Austin, TX: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1005

Hui, B., et al. 2024. "Qwen2.5-Coder Technical Report." Preprint, arXiv: 2409.12186 [cs.CL]. https://arxiv.org/abs/2409.12186

Hürriyetoğlu, Ali, et al. 2021. "Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2021): Workshop and Shared Task Report." In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (Case 2021)*, edited by A. Hürriyetoğlu, 1–9. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.case-1.1, https://aclanthology.org/2021.case-1.1/

Imani, M. B., S. Chandra, S. Ma, L. Khan, and B. Thuraisingham. 2017. "Focus Location Extraction from Political News Reports with Bias Correction." In *2017 IEEE International Conference on Big Data (Big Data)*, 1956–1964, edited by G. Karypis and J. Zhang. Boston, MA: IEEE.

Imani, M. B., L. Khan, and B. Thuraisingham. 2019. "Where Did the Political News Event Happen? Primary Focus Location Extraction in Different Languages." In *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, 61–70. Los Angeles, CA: IEEE.

Kent, S., and T. Krumbiegel. 2021. "CASE 2021 Task 2 Socio-Political Fine-Grained Event Classification Using Fine-Tuned RoBERTa Document Embeddings." In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (Case 2021)*, edited by A. Hürriyetoğlu, 208–217. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.case-1.26

Kingma, D. P., and J. Ba. 2014. "Adam: A Method for Stochastic Optimization." Preprint, arXiv:1412.6980, vol. 1412, no. 6. https://api.semanticscholar.org/CorpusID:6628106

LaFree, G., and L. Dugan. 2007. "Introducing the Global Terrorism Database." *Terrorism and Political Violence* 19 (2): 181–204.

Lang, K. 1995. "Newsweeder: Learning to Filter Netnews." In *Machine Learning Proceedings 1995*, edited by A. Prieditis and S. Russell, 331–339. San Francisco, CA: Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-377-6.50048-7

Lee, J.-S., and J. Hsiang. 2019. "PatentBERT: Patent Classification with Fine-Tuning a Pre-Trained Bert Model." Preprint, arXiv:1906.02124 [cs.CL]. https://arxiv.org/abs/1906.02124

Lefebvre, C., and N. Stoehr. 2023. "Rethinking the Event Coding Pipeline with Prompt Entailment." In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, edited by M. Akhtar, R. Aly, C. Christodoulopoulos, O. Cocarascu, Z. Guo, A. Mittal, M. Schlichtkrull, J. Thorne, and A. Vlachos, 1–16. Dubrovnik, Croatia: Association for Computational Linguistics. https://aclanthology.org/2023.fever-1.1/

Li, X., Z. Wang, D. Li, L. Khan, and B. Thuraisingham. 2022. "LPC: A Logits and Parameter Calibration Framework for Continual Learning." In *Findings of the Association for Computational Linguistics: EMNLP 2022*, edited by Y. Goldberg, Z. Kozareva, and Y. Zhang, 7142–7155. Abu Dhabi: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.529

Liu, P., L. Zhang, and J. A. Gulla. 2023. "Pre-Train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems." *Transactions of the Association for Computational Linguistics* 11: 1553–1571.

https://doi.org/10.1017/pan.2025.10027

Meher, S., and P. T. Brandt. 2025. "ConflLlama: Domain-Specific Adaptation of Large Language Models for Conflict Event Classification." *Research & Politics* 12 (3): 1–9. https://doi.org/10.1177/20531680251356282

O'Connor, B., B. M . Stewart, and N. A. Smith. 2013. "Learning to Extract International Relations from Political Context." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, edited by H. Schuetze, P. Fung, and M. Poesio, vol. 1, 1094–1104. Sofia, Bulgaria: Association for Computational Linguistics.

Ollion, E., R. Shen, A. Macanovic, and A. Chatelain. 2023. "ChatGPT for Text Annotation? Mind the Hype." SocArXiv Preprint: 32. https://osf.io/preprints/socarxiv/x58kn_v1

Olsen, H., É. Simon, E. Velldal, and L. 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)*, edited by A. Hürriyetoğlu, H. Tanev, S. Thapa, and G. Uludoğan, 40–53. St. Julians: Association for Computational Linguistics, https://aclanthology.org/2024.case-1.5

Osorio, J., et al. 2024. "Keep it Local: Comparing Domain-Specific LLMs in Native and Machine Translated Text Using Parallel Corpora on Political Conflict." In *2nd International Conference on Foundation and Large Language Models FLLM2024*, edited by Y. Jararweh, J. Jansen, and M. Alsmirat. Dubai: IEEE.

Osorio, J., et al. 2025. "*The Devil Is in the Details: Assessing the Effects of Machine Translation on LLM Performance in Domain-Specific Texts*." In *Proceedings of the 20th Machine Translation Summit (MT SUMMIT 2025)*, edited by P. Bouillon, J. Gerlach, S. Girletti, L. Volkart, R. Rubino, R. Sennrich, A. C. Farinha, M. Gaido, J. Daems, D. Kenny, H. Moniz, and S. Szoc. Geneva, Switzerland: European Association for Machine Translation.

Osorio, J., and A. Reyes. 2017. "Supervised Event Coding from Text Written in Spanish: Introducing EVENTUS Id." *Social Science Computer Review* 35 (3): 406–416.

Osorio, J., A. Reyes, A. Beltrán, and A. Ahmadzai. 2020. "Supervised Event Coding from Text Written in Arabic: Introducing Hadath." In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News*, edited by A. Hürriyetoğlu, E. Yörük, V. Zavarella, and H. Tanev, vol. 2020, 49–56. Marseille, France: European Language Resources Association (ELRA).

Palmer, G., et al. 2022. "The MID5 Dataset, 2011–2014: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science* 39 (4): 470–482.

Pappagari, R., P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak. 2019. "Hierarchical Transformers for Long Document Classification." In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 838–844. Singapore: IEEE. https://doi.org/10.1109/ASRU46091.2019.9003958

Park, H., Y. Vyas, and K. Shah. 2022. "Efficient Classification of Long Documents Using Transformers." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, edited by S. Muresan, P. Nakov, and A. Villavicencio, vol. 2, 702–709. Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.acl-short.79/

Parolin, E. S., et al. 2022. "Multi-CoPED: A Multilingual Multi-Task Approach for Coding Political Event Data on Conflict and Mediation Domain." In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, edited by M. Scheutz, R. Calo, M. Mara, and A. Zimmermann, 700–711. New York, NY: Association for Computing Machinery.

Parolin, E. S., Y. Hu, L. Khan, J. Osorio, P. T . Brandt, and V. D'Orazio. 2021. "Come-Ke: A New Transformers Based Approach for Knowledge Extraction in Conflict and Mediation Domain." In *2021 IEEE International Conference on Big Data (Big Data)*, edited by U. Fayyad and X. Zhu, 1449–1459. Orlando, FL: IEEE.

Parolin, E. S., L. Khan, J. Osorio, P. T. Brandt, V. D'Orazio, and J. Holmes. 2021. "3M-Transformers for Event Coding on Organized Crime Domain." In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, edited by U. Fayyad and X. Zhu, 1–10. Orlando, FL: IEEE. https://doi.org/10.1109/DSAA53316.2021.9564232

Pavlick, E., H. Ji, X. Pan, and C. Callison-Burch. 2016. "The Gun Violence Database: A New Task and Data Set for NLP." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, edited by J. Su, K. Duh, and X. Carreras, 1018–1024. Austin, TX: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1106.

Radford, B. J. 2021. "Automated Dictionary Generation for Political Eventcoding." *Political Science Research and Methods* 9 (1): 157–171.

Santifort, C., T. Sandler, and P. T. Brandt. 2013. "Terrorist Attack and Target Diversity: Changepoints and their Drivers." *Journal of Peace Research* 50 (1): 75–90.

Schrodt, P. A. 2001. "Automated Coding of International Event Data Using Sparse Parsing Techniques." In *Annual Meeting of the International Studies Association*. Chicago, IL: Citeseer.

Schrodt, P. A. 2012. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38 (4): 546–569.

Schrodt, P. A., and D. Van Brackle. 2012. "Automated Coding of Political Event Data." In *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian, 23–49. New York, NY: Springer.

Schwartz, R., J. Dodge, N. A. Smith, and O. Etzioni. 2020. "Green AI." *Communications of the ACM* 63 (12): 54–63. https://doi.org/10.1145/3381831.

Shellman, S. M. 2004. "Time Series Intervals and Statistical Inference: The Effects of Temporal Aggregation on Event Data Analysis." *Political Analysis* 12 (1): 97–104. https://doi.org/10.1093/pan/mpg017

Solaimani, M., S. Salam, L. Khan, P. T. Brandt, and V. D'Orazio. 2017a. "APART: Automatic Political Actor Recommendation in Real-Time." In *Social, Cultural, and Behavioral Modeling*, edited by D. Lee, Y.-R. Lin, N. Osgood, and R. Thomson, 342–348. Cham: Springer International Publishing.

https://doi.org/10.1017/pan.2025.10027

Solaimani, M., S. Salam, L. Khan, P. T. Brandt, and V. D'Orazio. 2017b. "RePAIR: Recommend Political Actors in Real-Time from News Websites." In *2017 IEEE International Conference on Big Data (Big Data)*, edited by R. Baeza-Yates, X. Hu, and J. Kepner, 1333–1340. Boston, MA: IEEE. https://doi.org/10.1109/BigData.2017.8258064

Steinert-Threlkeld, Z. C. 2019. "The Future of Event Data Is Images." *Sociological Methodology* 49 (1): 68–75. https://doi.org/10.1177/0081175019860238

Strubell, E., A. Ganesh, and A. McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by A. Korhonen, D. Traum, and L. Màrquez, 3645–3650. Florence, Italy: Association for Computational Linguistics. https://aclanthology.org/P19-1355/

Sundberg, R., and E. Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–532.

Team Gemma, et al. 2024 "Gemma 2: Improving Open Language Models at a Practical Size." Preprint, arXiv:2408.00118.

Wang, Y. 2024. "On Finetuning Large Language Models." *Political Analysis* 32 (3): 379–383.

Wei, J., et al. 2022. "Emergent Abilities of Large Language Models." Preprint, arXiv:2206.07682.

Wen, H., et al. 2021. "RESIN: A Dockerized Schema-Guided Cross-Document Cross-Lingual Cross-Media Information Extraction and Event Tracking System." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, edited by A. Sil, and X. V. Lin, 133–143. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-demos.16. https://aclanthology.org/2021.naacl-demos.16/

Yang, W., et al. 2023. "ConfliBERT-Spanish: A Pre-Trained Spanish Language Model for Political Conflict and Violence." In *2023 7th IEEE Congress on Information Science and Technology (CIST)*, edited by M. El Mohajir, M. Al Achhab, and B. E. El Mohajir, 287–292. Essouira, Morocco: IEEE.

Cite this article: Brandt, P. T., Alsarra, S., D'Orazio, V., Heintze, D., Khan, L., Meher, S., Osorio, J. and Sianan, M., (2025). Extractive versus Generative Language Models for Political Conflict Text Classification. *Political Analysis*, 1–29. https://doi.org/10.1017/pan.2025.10027

https://doi.org/10.1017/pan.2025.10027