Stanford CS224v Course
Conversational Virtual Assistants with Deep Learning

Lecture 9

Research on Document Sets: Qualitative Coding

Monica Lam & Sina Semnani

Event Detection from Social Media for Epidemic Prediction, Parekh et al, NAACL 2024

Multilingual Abstractive Event Extraction for the Real World
Sina J. Semnani, Pingyue Zhang, Wanyue Zhai, Haozhuo Li, Ryan Beauchamp, Trey Billing, Katayoun Kishi, Manling Li, Monica S. Lam. In Findings of ACL, July 2025

From Question Answering to Research

- **Answering questions** from a set of long documents
  - Extract the schema for one or more questions

- **Research** over a huge document set
  - We do not know the questions ahead of time
    - Answers to questions lead to subsequent questions
    - Re-extract information with each question is too costly

Research on Real-World Qualitative Data

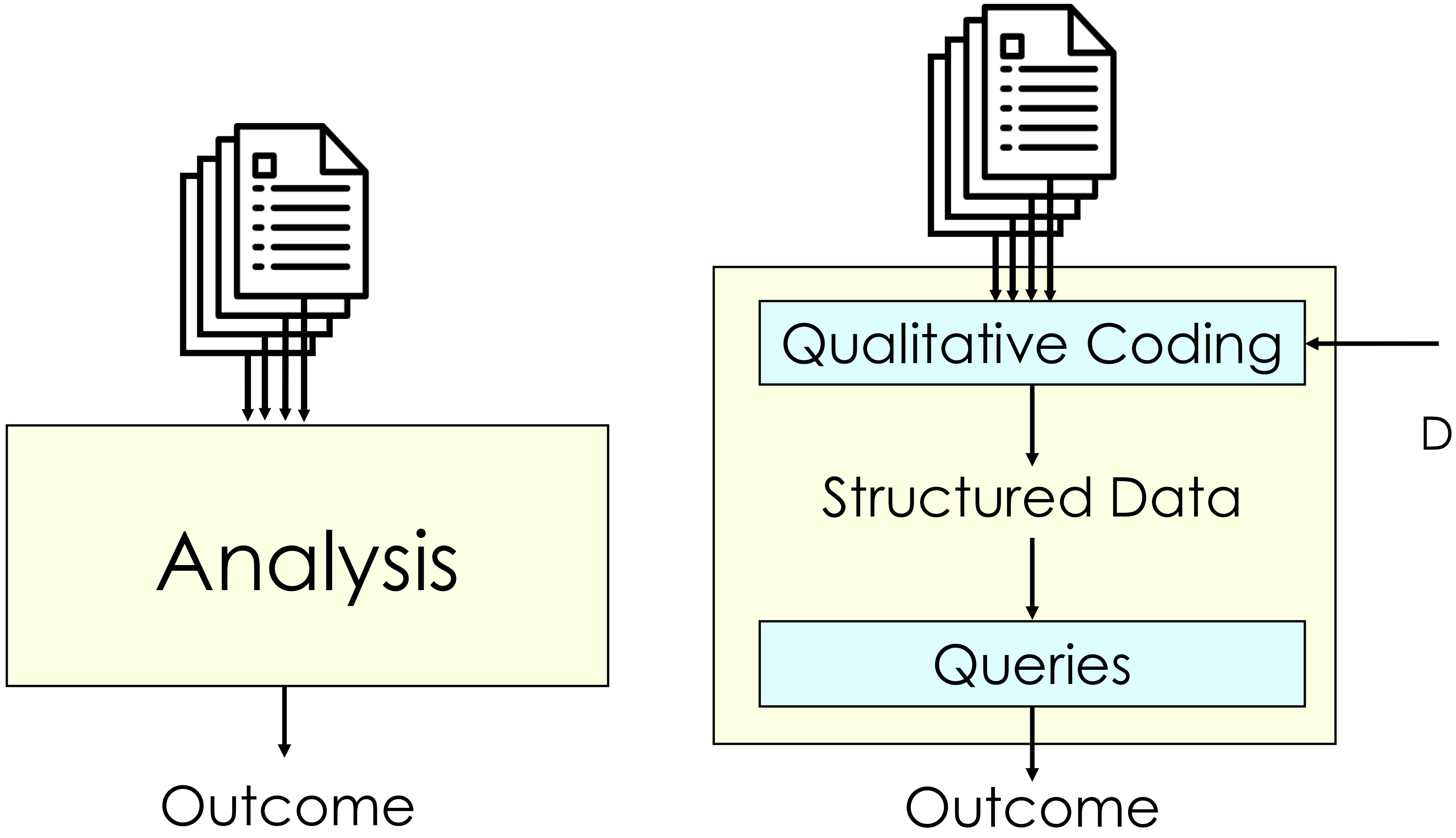
- **Topics**
  - Academic: Psychology, sociology, education, health services
  - Commercial: Market research, feedback, user experience, company culture, business optimizations, …
- **Purpose**
  - To identify, categorize information
  - To discover insights (significant data, relationships, trends, patterns, …)
- **Kinds of Data**
  - Human opinions: Feedbacks, interviews, …
  - Documents: Applications, notes, events, portfolios, …
  - Logs: Medical records, journals, meeting notes, logs, …

A Tried and True Methodology Across Many Fields

Qualitative Coding

(Structuring Qualitative Data)

Qualitative Coding



Codebook

Describes a data collection

- Structure
- Contents
- Layout

Codebook: aka **schema**

Qualitative coding: aka **information extraction**

Codebook Design

- Inductive
  - From the data up – useful for exploration
- Deductive
  - From the model down – useful to test hypothesis
- Hybrid
  - A combination of both

Codebook Examples

- Experts' key concepts used to "evaluate and analyze" a doc
- Can be personal and subjective

Document Set	Codebook
News	Source, event types, event parameters
Social Media	Source, ratings, reposts, event types, event parameters
Resumes	Name, address, education, employment, publications, products,
Paper Reviews	Topic area, originality, soundness, significance, theory, empirical
Research proposals	Relevance, soundness, potential impact, risk, team credentials
Earnings Reports	Revenue, expenses, net income, earnings per share (EPS), balance sheet highlights, cash flow changes, …
Interviews	Participants' thoughts, emotions, experiences, and behaviors
Medical Records	Patient's health history, diagnoses, medications, treatment plans, immunization dates, allergies

Generated by GPT-4

Lecture Goal

Two Research Studies in Real Life

Focus: Qualitative Coding (QC) for Detecting Events

1. Academic Automatic QC (2020—2024)

From Tweets to Epidemic Prediction

2. Non-Profit Manual QC (2005—)

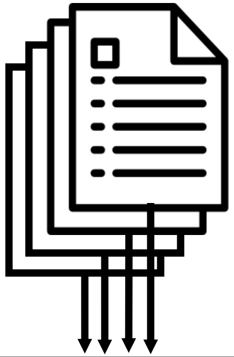
ACLED: Armed Conflict Location & Event Data

Event Detection from Social Media for Epidemic Prediction

Tanmay Parekh[†] **Anh Mac**[†] **Jiarui Yu**[†] **Yuxuan Dong**[†]
Syed Shahriar[†] **Bonnie Liu**[†] **Eric Yang**[†] **Kuan-Hao Huang**[§]
Wei Wang[†] **Nanyun Peng**[†] **Kai-Wei Chang**[†]
[†]Computer Science Department, University of California, Los Angeles
[§]Department of Computer Science, University of Illinois Urbana-Champaign
{tparekh, weiwang, violetpeng, kwchang}@cs.ucla.edu

NAACL 2024

Early Disease Outbreak Detection



Tweets

Event Detection

Estimate real-time number of cases in early stages of a new outbreak.

20 million per day on COVID19

Symptoms, prevention, deaths

In May 2020

Can we detect future epidemics early?

Event Ontology for Epidemic Preparedness

Event Type	**Event Definition**	**Example Event Mentions**
Infect	The process of a disease/pathogen invading host(s)	1. Children can also **catch** COVID-19 ... 2. If you have antibodies, you **had** the virus. Period.
Spread	The process of a disease spreading or prevailing massively at a large scale	1. #COVID-19 CASES **RISE** TO 85,940 IN INDIA ... 2. ... the **prevalence** of asymptomatic COVID - 19 cases ...
Symptom	Individuals displaying physiological features indicating the abnormality of organisms	1. (user) (user) Still **coughing** two months after being infected by this stupid virus ... 2. If a person nearby is **sick**, the wind will scatter the virus ...
Prevent	Individuals trying to prevent the infection of a disease	1. ... wearing mask is the way to **prevent** COVID-19 2. ... an #antibody that has been succssful at **blocking** the virus
Control	Collective efforts trying to impede the spread of an epidemic	1. Social Distancing **reduces** the spread of covid ... 2. (user) COVID is still among us! Wearing masks **saves** lives!
Cure	Stopping infection and relieving individuals from infections/symptoms	1. ... **recovered** corona virus patients cant get it again 2. ... patients are **treated** separately at most places
Death	End of life of individuals due to an infectious disease	1. More than 80,000 Americans have **died** of COVID ... 2. The virus is going to get people **killed**. Stay home. Stay safe.

Table 1: Event ontology comprising seven event types promoting epidemic preparedness along with their definitions and two example event mentions. The trigger words are marked in **bold**.

Speed Dataset

Social Platform based Epidemic Event Detection (Covid 19)

- 1,975 tweets
- 2,217 event mentions

Trigger Identification of Future Epidemics (F1)

	Monkeypox	Zika + Dengue
GPT-3 (Zero-shot)	42.23	53.22
BERTQA RoBERTa-Large (355M parameters) Trained with SPEED	67.38	67.95

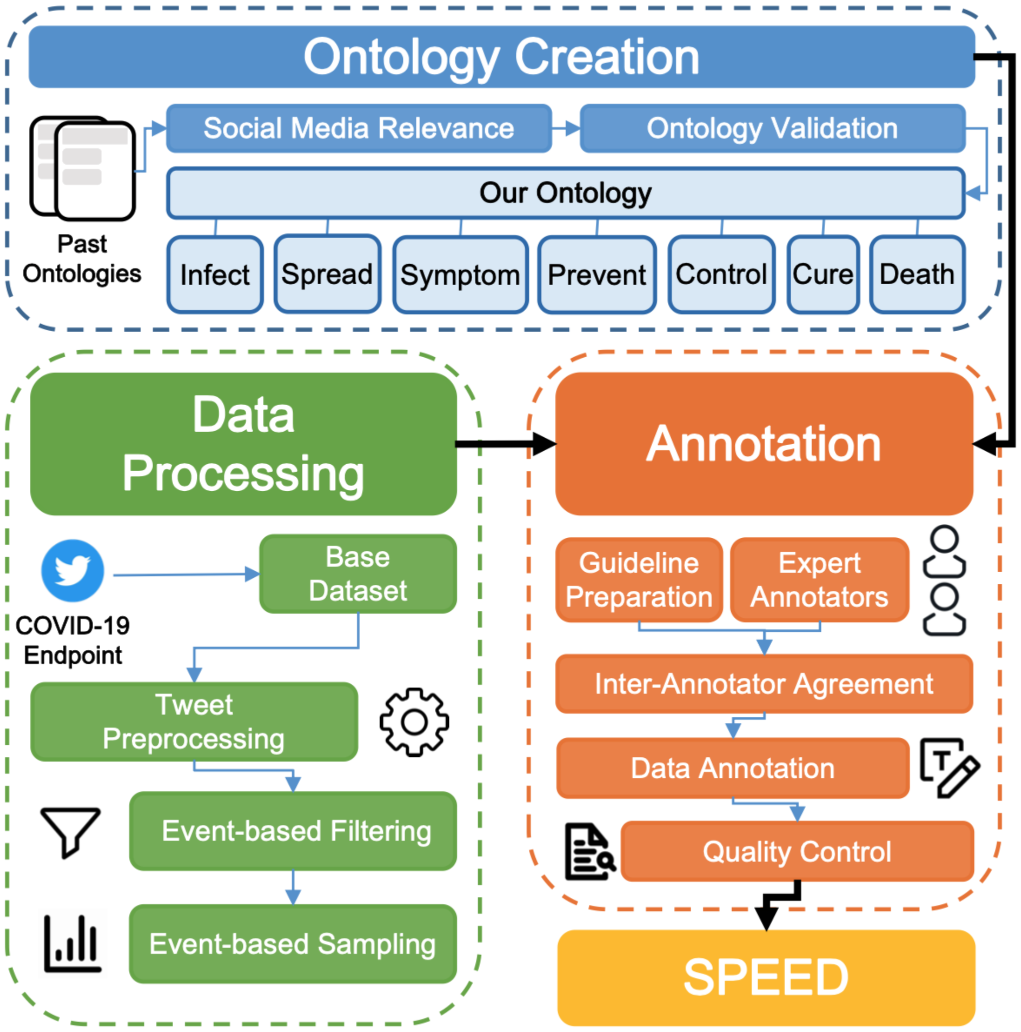
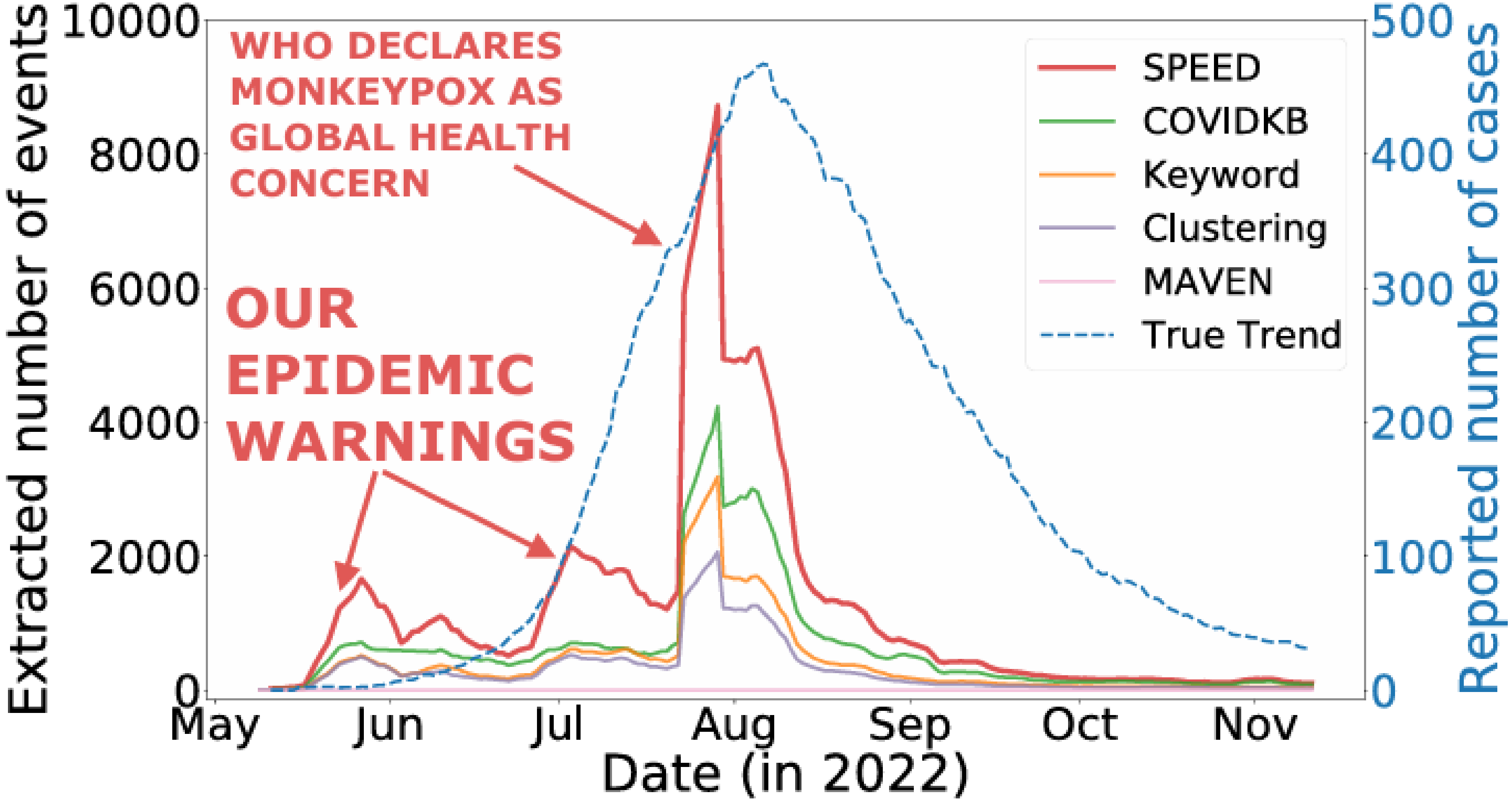


Figure 3: Overview of our dataset creation process with three major steps: Ontology Creation, Data Processing, and Data Annotation.

Event detection → Epidemic Warnings 4-9 Weeks Earlier (Monkeypox)

Aug 14, 2024



Lecture Goal
Two Research Studies in Real Life

Focus: Qualitative Coding (QC) for Detecting Events

1. Academic Automatic QC (2020—2024)
From Tweets to Epidemic Prediction

2. Non-Profit Manual QC (2005—)
ACLED: Armed Conflict Location & Event Data

ACLED | Armed Conflict Location & Event Data

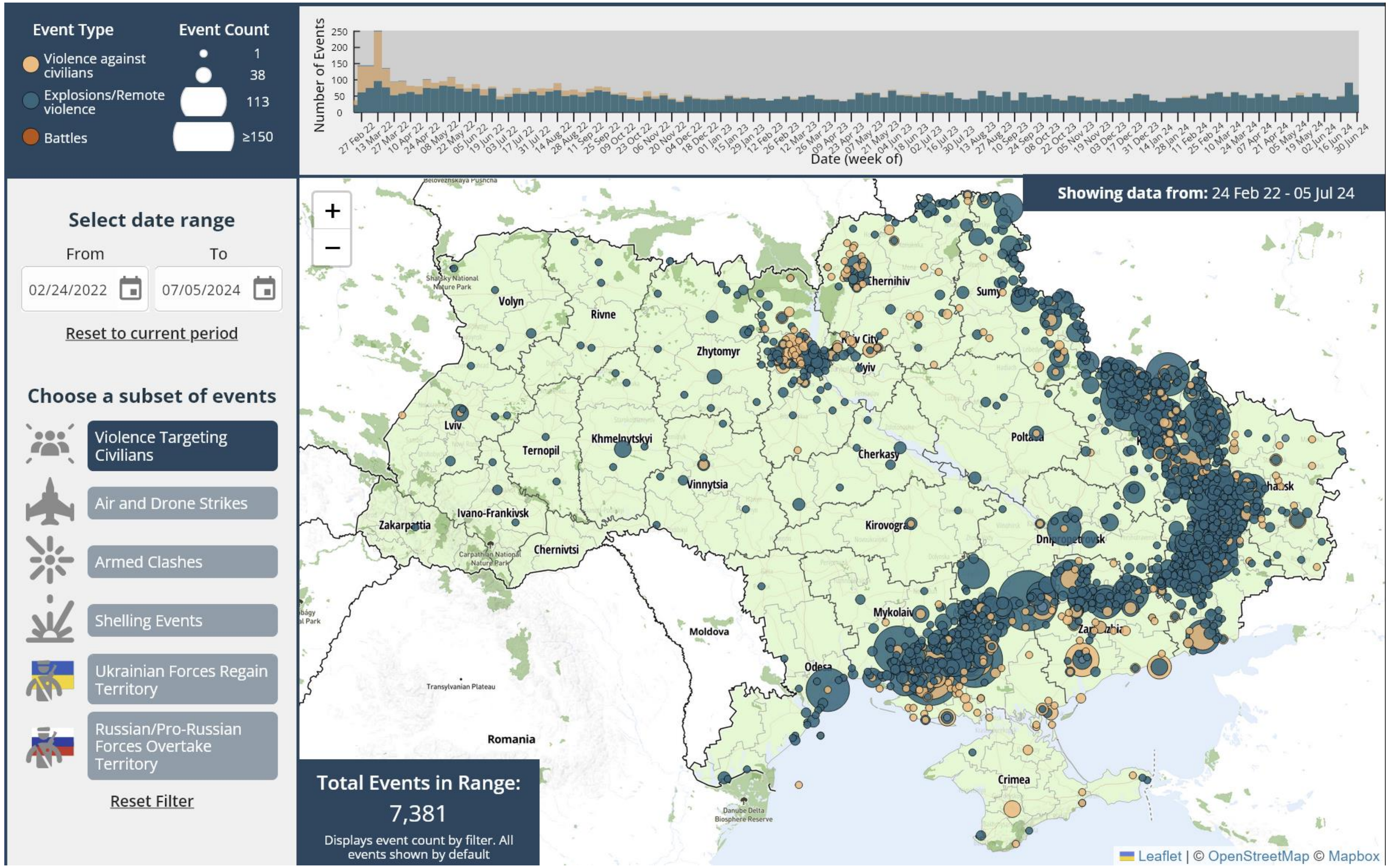
An independent, impartial, international non-profit organization collecting data on violent conflict and protest in all countries and territories in the world.

UN's International Organization for Migration (IOM)
Uses ACLED data to track
the movement and needs of displaced people
in over 80 countries

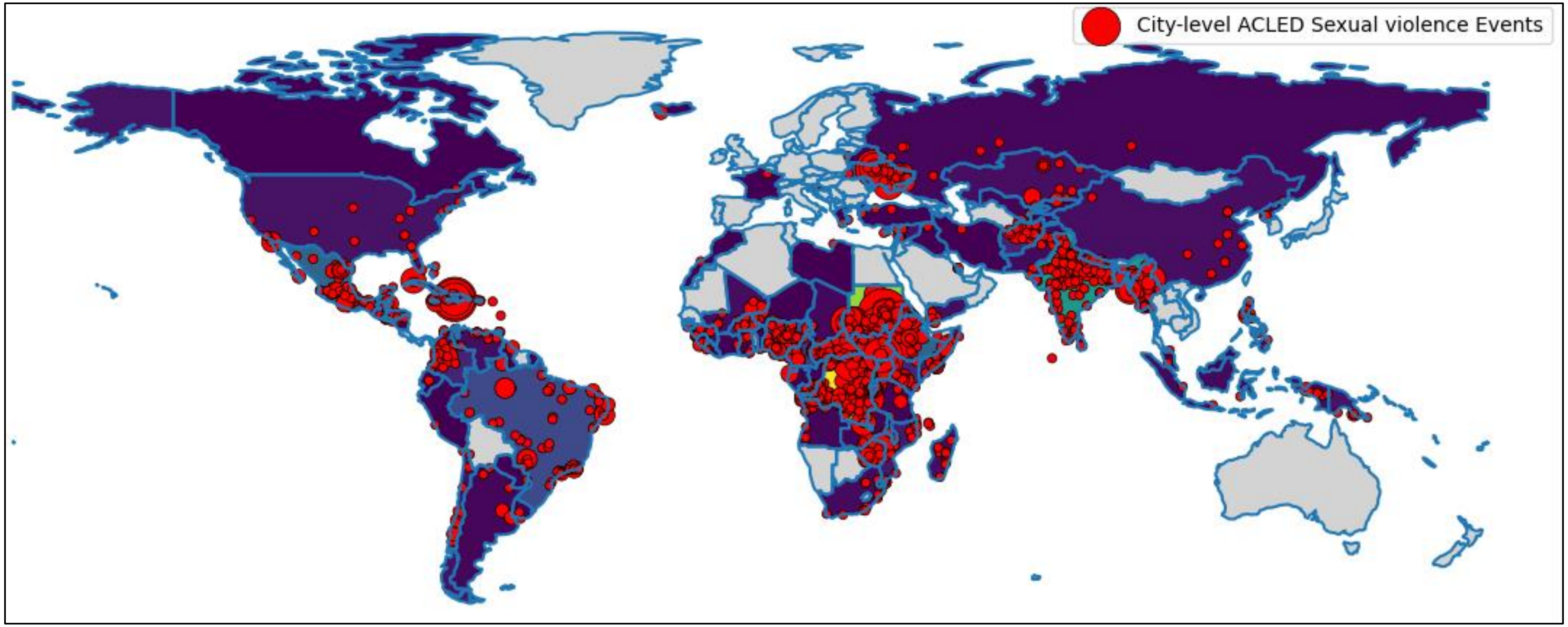
The US multi-agency
Global Fragility Act Secretariat
Uses ACLED data
to promote stability across five priority countries.

Introducing ACLED: An Armed Conflict Location and Event Dataset, Raleigh et al, Journal of Peace Research 2010

Analysis from ACLED data: Violence Against Civilians in Ukraine

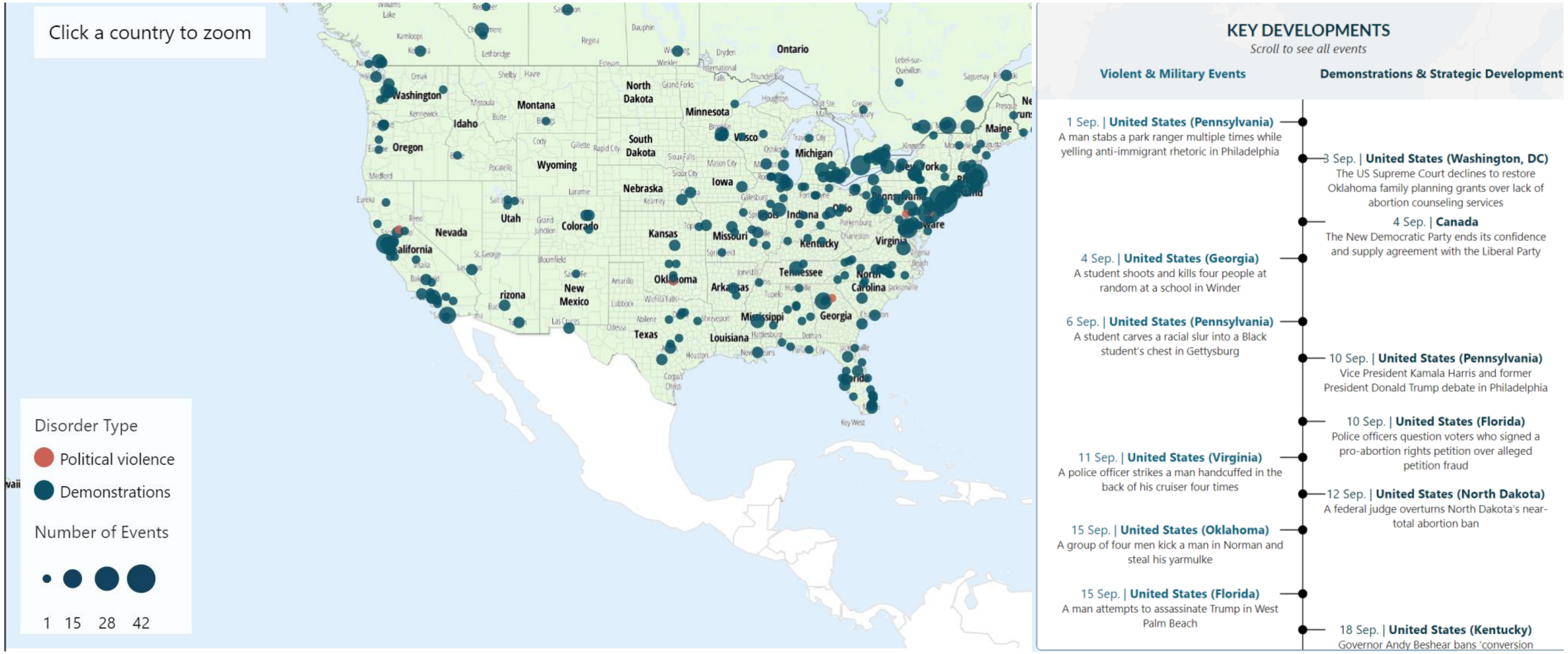


Analysis from ACLED data:
Sexual Violence Around the Globe



Analysis from ACLED data:
Tracking Political Demonstrations

September 2024 review, published on October 4th



https://acleddata.com/2024/10/04/united-states-canada-overview-september-2024/, ACLED, Accessed 11/5/2024

Analysis from ACLED data:
Tracking Political Demonstrations

ACLED reports that in September 2024:

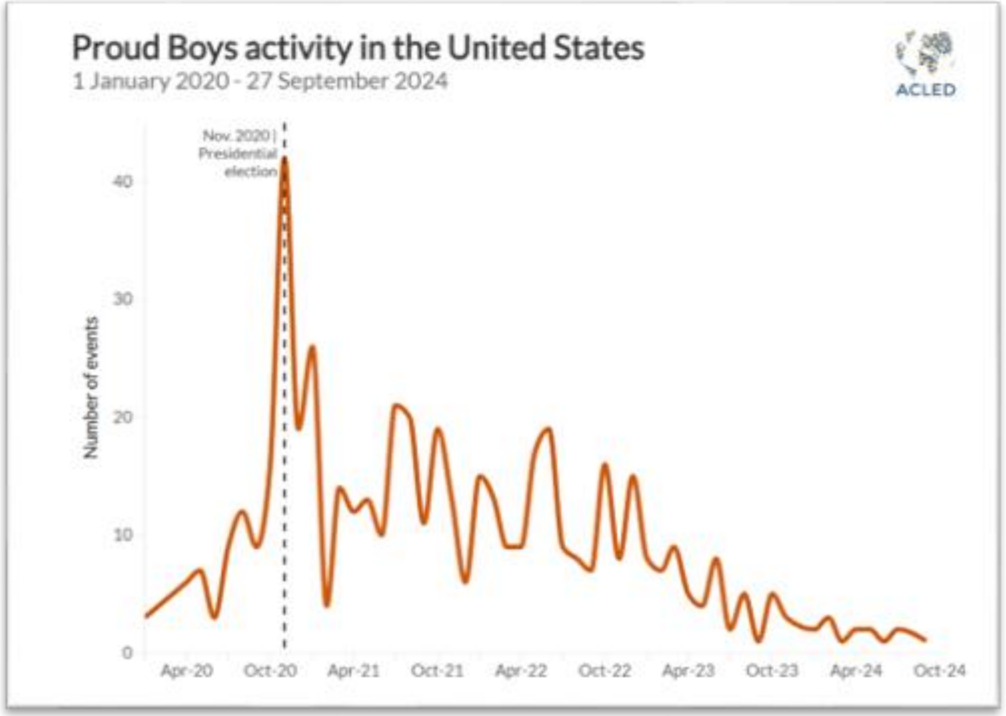
Because →

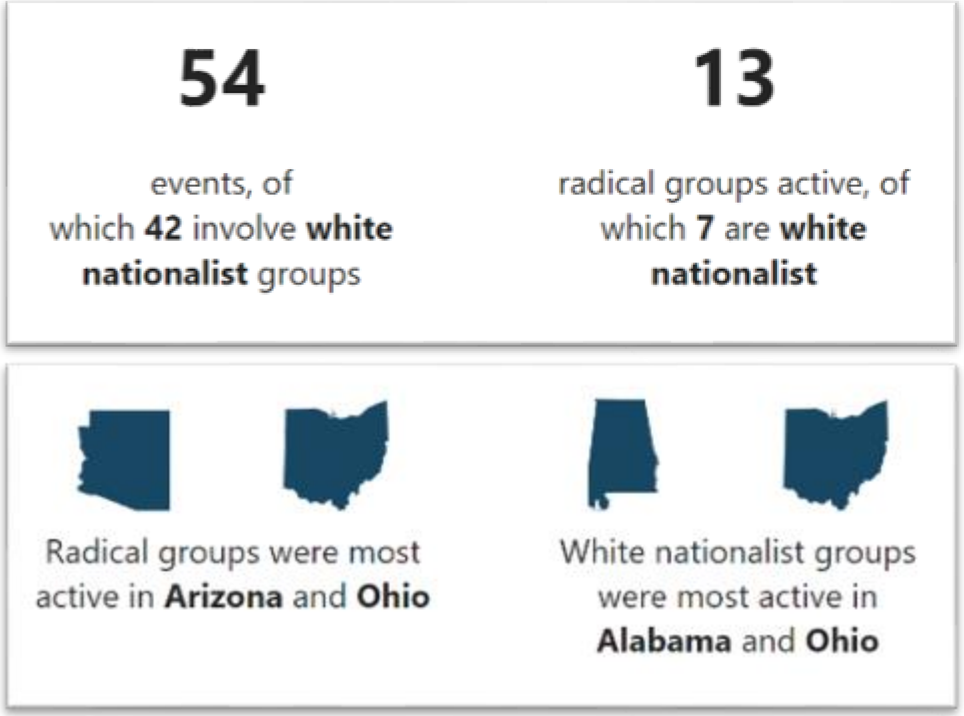
Anti-Trump demonstrations rose in the United States after the September 10th presidential debate.

https://acleddata.com/2024/10/04/united-states-canada-overview-september-2024/, ACLED, Accessed 11/5/2024

Analysis from ACLED data: Tracking Political Demonstrations

ACLED reports on various radical groups:





https://acleddata.com/2024/10/04/united-states-canada-overview-september-2024/, ACLED, Accessed 11/5/2024

Manual Qualitative Coding

Sources of Data

What types of sources does ACLED use?

ACLED uses four types of sources. Every week, ACLED researchers assess thousands of sources in dozens of languages to provide the most comprehensive database on political violence and demonstrations. All types are reviewed each week. These include:

1. Traditional Media: This includes all subnational, national, regional, and international media outlets that are governed by journalistic principles of verification.
2. Reports: International institutions and non-governmental organizations – such as aid groups, human rights organizations, and investigative journalism groups – regularly publish reports on political violence. Where applicable, ACLED incorporates events from these reports. Under certain conditions, reports from groups involved in conflict themselves are also included (Ministries of Defense, armed groups, NATO, etc.).
3. Local Partner Data: The past decades have seen an increase in conflict observatories established at the local level as both social activism and the ability to report political violence have increased. These organizations leverage their local knowledge as they collect and obtain information through primary and/or secondary means. ACLED develops relationships with local partners to enhance the depth and quality of its data.
4. New Media (targeted and verified): 'New media' (e.g. Twitter, Telegram, WhatsApp) can be a powerful supplemental source but varies widely in terms of quality. Therefore, ACLED does not crowdsource or scrape large amounts of social media. Rather, a targeted approach to the inclusion of new media is preferred through either the establishment of relationships with the source directly, or the verification of the quality of each source.

Scale Matters in the Real World

- The document set is **massive**
  - 150,000+ online news articles are published each day
  - Analysis can cover a long period of time
- Global Analysis means **many languages**
- ACLED for example, has covered:

2 Million
Events

80+
Languages

200+
Researchers

243
Countries/territories

https://huggingface.co/datasets/stanford-oval/ccnews, Semnani, Accessed 11/5/2024

Manual Qualitative Coding

Codebook
(Event Types)

Event type	Sub-event type	Disorder type
Battles	*Government regains territory*	**Political violence**
	Non-state actor overtakes territory	
	Armed clash	
Protests	*Excessive force against protesters*	**Political violence; Demonstrations**
	Protest with intervention	**Demonstrations**
	Peaceful protest	
Riots	*Violent demonstration*	
	Mob violence	**Political violence**
Explosions/Remote violence	*Chemical weapon*	
	Air/drone strike	
	Suicide bomb	
	Shelling/artillery/missile attack	
	Remote explosive/landmine/IED	
	Grenade	

Event type	Sub-event type	Disorder type
Violence against civilians	*Sexual violence*	
	Attack	
	Abduction/forced disappearance	
Strategic developments	*Agreement*	**Strategic developments**
	Arrests	
	Change to group/activity	
	Disrupted weapons use	
	Headquarters or base established	
	Looting/property destruction	
	Non-violent transfer of territory	
	Other	

Manual Qualitative Coding

Codebook
(Event Arguments)

#	Column name	Column description	Values
	event_id_cnty	A unique alphanumeric event identifier by number and country acronym. This identifier remains constant even when the event details are updated.	E.g., ETH9766
	event_date	The date on which the event took place. Recorded as year-month-day.	E.g., 2023-02-16
	year	The year in which the event took place.	E.g., 2018
	time_precision	A numeric code between 1 and 3 indicating the level of precision of the date recorded for the event. The higher the number, the lower the precision.	1, 2, or 3; with 1 being the most precise.
3	*disorder_type*	The disorder category an event belongs to.	Political violence, Demonstrations, or Strategic developments.
6	*event_type*	The type of event; further specifies the nature of the event.	E.g., Battles *For the full list of ACLED event types, see the ACLED Event Types table.*
25	*sub_event_type*	A subcategory of the event type.	E.g., Armed clash *For the full list of ACLED sub-event types, see the ACLED Event Types table.*

#	Column name	Column description	Values
10K	*actor1*	One of two main actors involved in the event (does not necessarily indicate the aggressor).	E.g., Rioters (Papua New Guinea)
10K	*assoc_actor_1*	Actor(s) involved in the event alongside Actor 1 or actor designations that further identify Actor 1.	E.g., Labor Group (Spain); Women (Spain) Can have multiple actors separated by a semicolon, or can be blank.
8	*inter1*	A text value indicating the type of Actor 1 (*for more, see the section Actor Names, Types, and 'Inter' Codes*).	E.g., Rebel group
10K	*actor2*	One of two main actors involved in the event (does not necessarily indicate the target or victim).	E.g., Civilians (Kenya) Can be blank.
10K	*assoc_actor_2*	Actor(s) involved in the event alongside Actor 2 or actor designation further identifying 'Actor 2.	E.g., Labor Group (Spain); Women (Spain) Can have multiple actors separated by a semicolon, or can be blank.
8	*inter2*	A text value indicating the type of Actor 2 (*for more, see the section Actor Names, Types, and 'Inter' Codes*).	E.g., State forces Can be blank.
44	*interaction*	A text value based on a combination of Inter 1 and Inter 2 indicating the two actor types interacting in the event (*for more, see the section Actor Names, Types, and 'Inter' Codes*).	E.g., Rebel group – Civilians

Column name	Column description	Values
iso	A unique three-digit numeric code assigned to each country or territory according to ISO 3166.	E.g., 231 for Ethiopia
region	The region of the world where the event took place.	E.g., Eastern Africa
country	The country or territory in which the event took place.	E.g., Ethiopia
admin1	The largest sub-national administrative region in which the event took place.	E.g., Oromia
admin2	The second largest sub-national administrative region in which the event took place.	E.g., Arsi Can be blank.
admin3	The third largest sub-national administrative region in which the event took place.	E.g., Merti Can be blank.

Column name	Column description	Values
location	The name of the location at which the event took place.	E.g., Abomsa
latitude	The latitude of the location in four decimal degrees notation (EPSG:4326).	E.g., 8.5907
longitude	The longitude of the location in four decimal degrees notation (EPSG:4326).	E.g., 39.8588
geo_precision	A numeric code between 1 and 3 indicating the level of certainty of the location recorded for the event. The higher the number, the lower the precision.	1, 2, or 3; with 1 being the most precise.

Column name	Column description	Values
source	The sources used to record the event. Separated by a semicolon.	E.g., Ansar Allah; Yemen Data Project
source_ scale	An indication of the geographic closeness of the used sources to the event (*for more, see the section Source Scale*).	E.g., Local partner-National
notes	A short description of the event.	E.g., On 16 February 2023, OLF-Shane abducted an unidentified number of civilians after stopping a vehicle in an area near Abomsa (Merti, Arsi, Oromia). The abductees were traveling from Adama to Abomsa, Arsi.
fatalities	The number of reported fatalities arising from an event. When there are conflicting reports, the most conservative estimate is recorded.	E.g., 3 No information on fatalities is recorded as 0 reported fatalities.
tags	Additional structured information about the event. Separated by a semicolon.	E.g., women targeted: politicians; sexual violence
timestamp	An automatically generated Unix timestamp that represents the exact date and time an event was uploaded to the ACLED API.	E.g., 1676909320

Manual Qualitative Coding

Correctness is Critical

Very Elaborate
Manual Coding and Review Process

How Does ACLED Ensure Correctness?

Coding and sourcing process

ACLED data are coded by a range of experienced researchers with knowledge of local contexts and languages who collect information mainly from secondary sources by applying the guidelines outlined in the Codebook and supplemental documentation to extract relevant information.

ACLED data are collected each week after individual researchers have examined the information from structured and regularly reviewed lists of secondary sources. A sourcing platform ensures the same sources are checked each week in a consistent manner. [1]

Every event is coded using the same rules on 'who, what, where, and when' to maximize accuracy and consistency. Additional information is also provided in each row of data, including: event ID numbers, precision scores for location and time, codes to distinguish between the types of actors, a brief summary of the event, fatality numbers if reported, and additional information for deeper analysis.

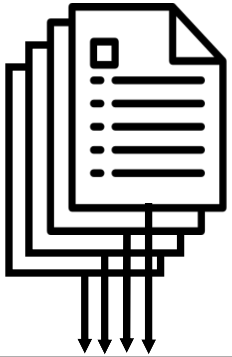
Throughout the weekly data collection and coding process, Researchers pose questions to team members and their Research Managers to clarify difficult coding decisions or flag potential data collection issues. Researchers use a coding platform to ensure that the coding of actor names, interactions, locations, etc., are consistent with previous iterations of each group and location.

Data review and cleaning process

Following weekly data collection, the data undergo three rounds of review:

1. **First**, Researchers review their data to ensure intra-coder reliability.
   - Decisions on specific matters – such as a new active group – are flagged for further review.
   - After the review, Researchers submit their data and source materials to their Research Manager.
2. **Next**, Research Managers review these data for inter-coder reliability across the region.
   - Research Managers cross-check the data for general accuracy and consistency, ensuring that events meet the criteria for inclusion and that coding is in line with the methodology and previous local context applications.
3. **Finally**, the data are passed to a final reviewer who reviews the data to ensure that the inter-coder standards are met, and that the methodology is applied consistently across different regions and contexts.

Summary: Challenges of ACLED Qualitative Coding



Scale and diversity of sources
2 M events
Traditional/new media, reports, local partners
243 Locales and 80+ languages

Complex, Detailed Codebook
31 Fields
25 Subevent types
10Ks of actors

Event Detection

High-Quality Coding with Reviews
Long guidelines
Multi-level reviews for uniformity

Accurate Database for Making Policies

Quiz: Contrast the accuracies of ACLED vs Epidemic Prediction

Prior Event Detection Work

Event Detection – A Long-Studied Subject

- Started in 1970s
- Most widely used dataset is ACE, published in 2005
- Limits of technology → simplifying assumptions:
  - Unit of analysis is **sentence**
  - Extraction is done using **keywords** and **rules**
    - Emphasis on **words** & **spans** (consecutive words in the text)
      - **Not semantics**

Limitations of Sentence/Span-Based Extraction

Example from ACE05

EU foreign policy supremo Javier Solana likewise slammed the attack, although he also took a jab at Israel, saying, "There have been too many civilians killed in the last weeks and the last days."

Span-based annotation
– no global meaning

- "killed" is the event mention
- "civilians" is the event's victim
- "the last weeks" and "the last days" are the event time

"Sentence" – not enough context

- Who are these civilians? ("Palestinians" from the previous sentence)
- When is last week?
- Who was the attacker?
- Where did it take place?

Extractive (Spans) vs. Abstractive (Meaning)

Quiz: What is the advantage of abstractive linking?

Spans

workers
musicians
ILWU
CWA
Teamsters Local 70
PMWG
former workers at Tesla
Past ILWU president Brian McWilliams
Nadya Williams

Entity List

Labor Group
ILWU: International Longshore and Warehouse Union
AFL-CIO: American Federation of Labor and Congress of Industrial Organizations
CWA: Communications Workers of America
IBT: International Brotherhood of Teamsters
× (not in the database)

Entity Database

AFL-CIO is affiliated with CWA
IBT has many local branches, including Teamsters Local 70, …

Summary: Deficiencies/Fixes of Prior Work

- Unit of analysis: Increase accuracy
  Sentence → Article
- Extraction method: Increase accuracy
  Keywords and **rules → Semantics**
- Linking: easier analysis and comparison
  Span (extractive) → Given Entity/enum value (abstractive)

MULTILINGUAL ABSTRACTIVE EVENT EXTRACTION FOR THE REAL WORLD

Sina J. Semnani[1] **Pingyue Zhang**[2] **Wanyue Zhai**[1] **Haozhuo Li**[1]
Ryan Beauchamp[1] **Trey Billing**[3] **Katayoun Kishi**[3] **Manling Li**[2] **Monica S. Lam**[1]
[1]Stanford University [2]Northwestern University [3]ACLED
{sinaj, wzhai702, tommy01, rmb87, lam}@cs.stanford.edu,
{pingyue.zhang,manling.li}@northwestern.edu,
{t.billing,k.kishi}@acleddata.com

In Findings of ACL, July 2025

Research Questions on Automatic Qualitative Coding (AQC)

- **ACLED is a major world-level effort**
  - Weekly effort by 200+ (part-time) world-wide researchers

1. **Can AQC help ACLED to improve efficiency expand coverage?**
   - To other events, languages, countries
   - Accurate ACLED data → High-quality dataset (LEMONADE)?
     → Evaluate AQC with fine-tuning on LEMONADE?
2. **For new domains, can In-Context Learning eliminate expensive training dataset annotation?**
   - Use high-quality dataset (LEMONADE) → Develop Zero-Shot AQC (Zest)
     → Evaluate in-context learning

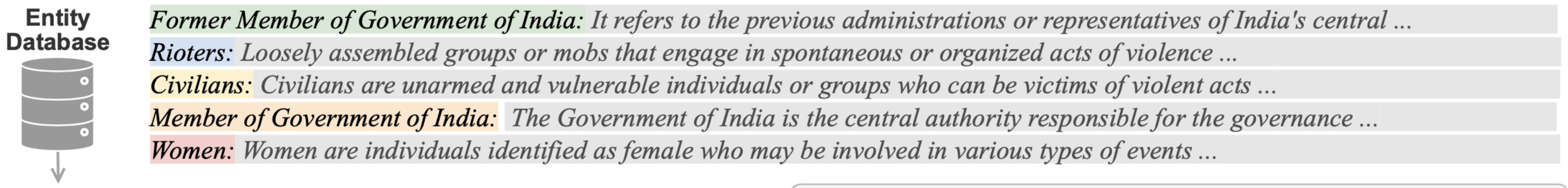
Paper Contributions

Define the Abstractive Event Extraction Task

Lemonade: a High-Quality Real-World Event Dataset

Zest: Zero-Shot Qualitative Coding

Evaluation (Fine-tuning vs. Zero Shot LLM)



LLM Knowledge:
A *sarpanch* is the elected head of a village-level local government body called a Gram Panchayat in countries like India, Bangladesh, and Pakistan.

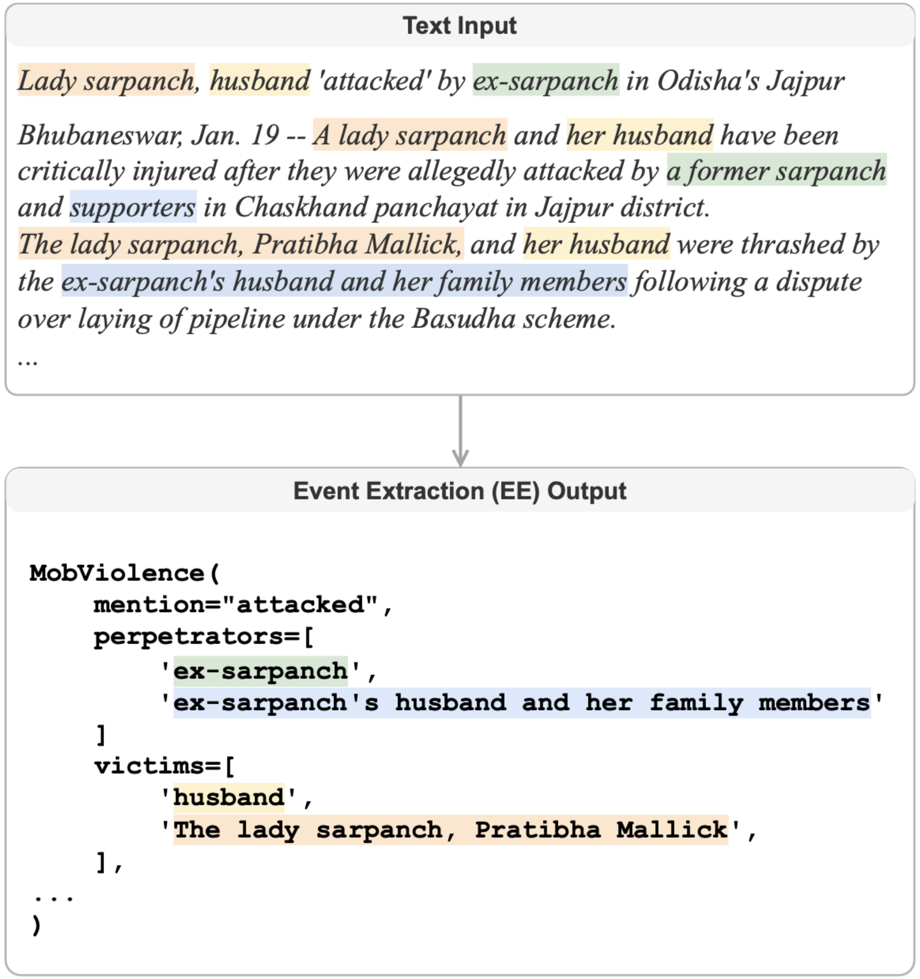


Figure 1: An example from LEMONADE showing abstractive event annotation. The input text and annotations are summarized for clarity. A hypothetical extractive annotation is included for comparison, illustrating the key differences between abstractive and extractive approaches.

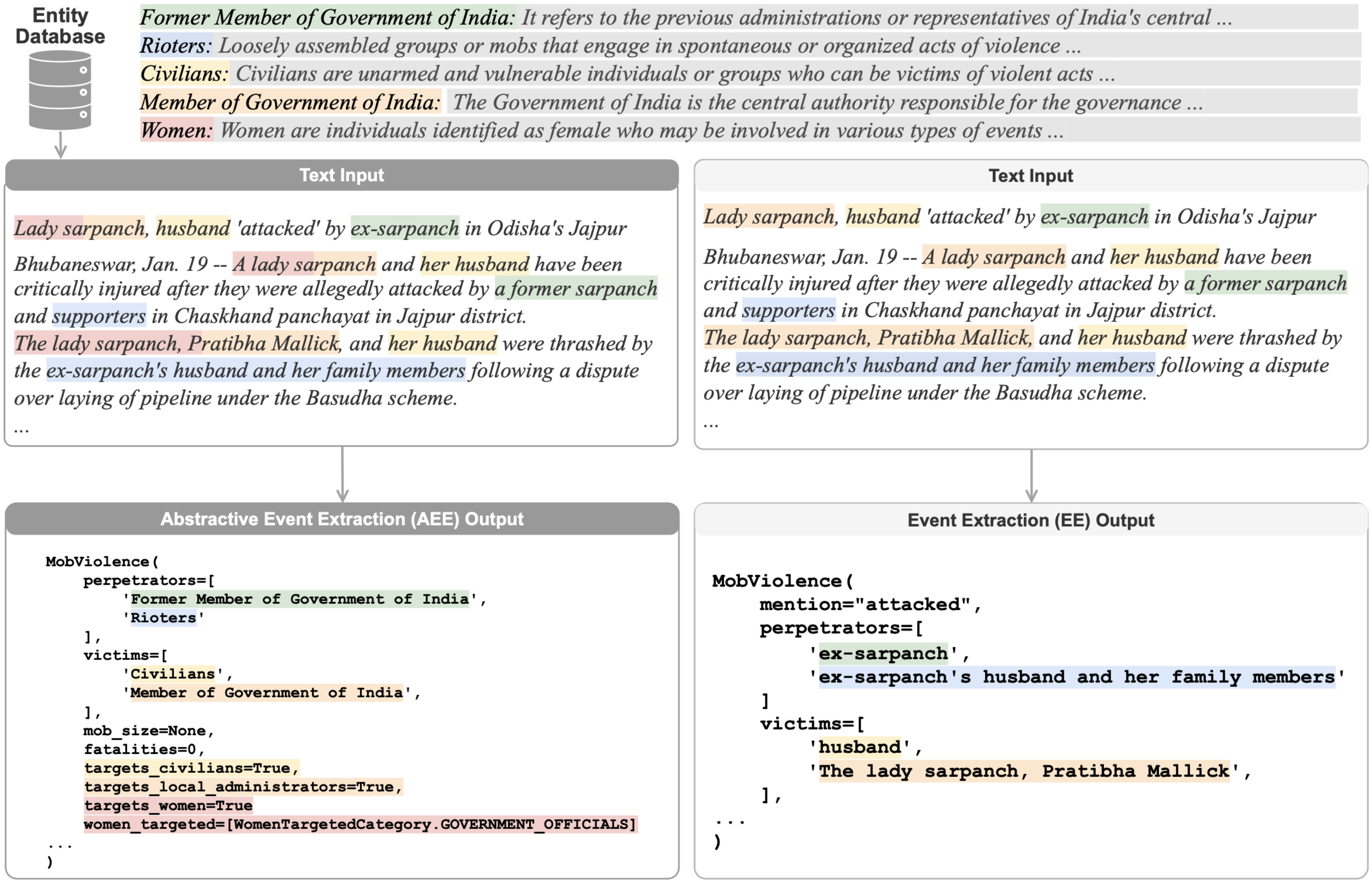


Figure 1: An example from LEMONADE showing abstractive event annotation. The input text and annotations are summarized for clarity. A hypothetical extractive annotation is included for comparison, illustrating the key differences between abstractive and extractive approaches.

Abstractive Event Extraction Task

Scope: full article

- What events "happen"?
  - E.g. a peaceful protest, riot, arrest

25 Event Types

Problem 1. Event detection (AED)

- What are the non-entity arguments?
  - Each event type has its own arguments

e.g. Peaceful Protest: protesters, location, crowd size, …

Event Arguments

Problem 2. (non-entity arguments)
Abstractive Event Argument Extraction

- What are the entity arguments?
  - Actors: e.g. political parties, organized groups

6,217 Entities

Problem 3. Abstractive Entity Linking (AEL)

Paper Contributions

Define the Abstractive Event Extraction Task

Lemonade: a High-Quality Real-World Event Dataset

Zest: Zero-Shot Qualitative Coding

Evaluation (Fine-tuning vs. Zero Shot LLM)

LEMONADE

Large
Expert-annotated
Multilingual
Ontology-**N**ormalized
Abstractive
Dataset of
Events

- A cleaned version of ACLED event data
  - Training: Events from January to March of 2024
  - Validation/Test: Events from April 2024 to January 2025
  - 39,686 events, 10,707 entities
  - 171 countries/territories, 20 languages

The best-annotated dataset
excerpted from a real-life dataset
with end-to-end abstract entity linking

LEMONADE

Distribution by Language

Language (language code)	Train	Dev	Test
English (en)	4593	500	500
Spanish (es)	1528	500	500
Arabic (ar)	3171	500	500
French (fr)	805	500	500
Italian (it)	773	500	500
Russian (ru)	482	500	500
German (de)	1422	500	500
Turkish (tr)	925	500	500
Burmese (my)	932	500	500
Indonesian (id)	754	500	500
Ukrainian (uk)	1157	500	500
Korean (ko)	1167	500	500
Portuguese (pt)	1759	500	500
Dutch (nl)	256	284	284
Somali (so)	251	358	358
Nepali (ne)	389	439	439
Chinese (zh)	332	500	500
Persian/Farsi (fa)	368	500	500
Hebrew (he)	177	332	332
Japanese (ja)	175	272	272
Total	21,416	9,185	9,185

Paper Contributions

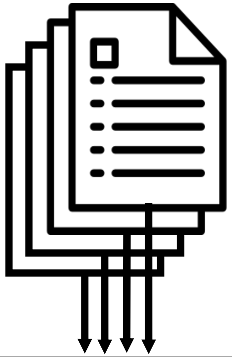
Define the Abstractive Event Extraction Task

Lemonade: a High-Quality Real-World Event Dataset

Zest: Zero-Shot Qualitative Coding

Evaluation (Fine-tuning vs. Zero Shot LLM)

Challenges of ACLED Qualitative Coding



Scale and diversity of sources
2 M events
Traditional/new media, reports, local partners
243 Locales and 80+ languages

Complex, Detailed Codebook
31 Fields
25 Subevent types
Ks of actors

Event Detection

High-Quality Coding with Reviews
Long guidelines
Multi-level reviews for uniformity

Database for Making Policies

Problem 1: Let’s Detect Event in this News Article

- From Indybay.org, a community news website
- 580 words

Tesla Fremont MLK Rally Protesting Racism, Union Busting

A rally on MLK Weekend was held at the massive Tesla Fremont assembly plant where 20,000 work. The action was called to protest the systemic racism and sexism by Elon Musk and his massive union busting drive. They also supported the Swedish striking Tesla mechanics who have been on strike for nearly 2 months.

…

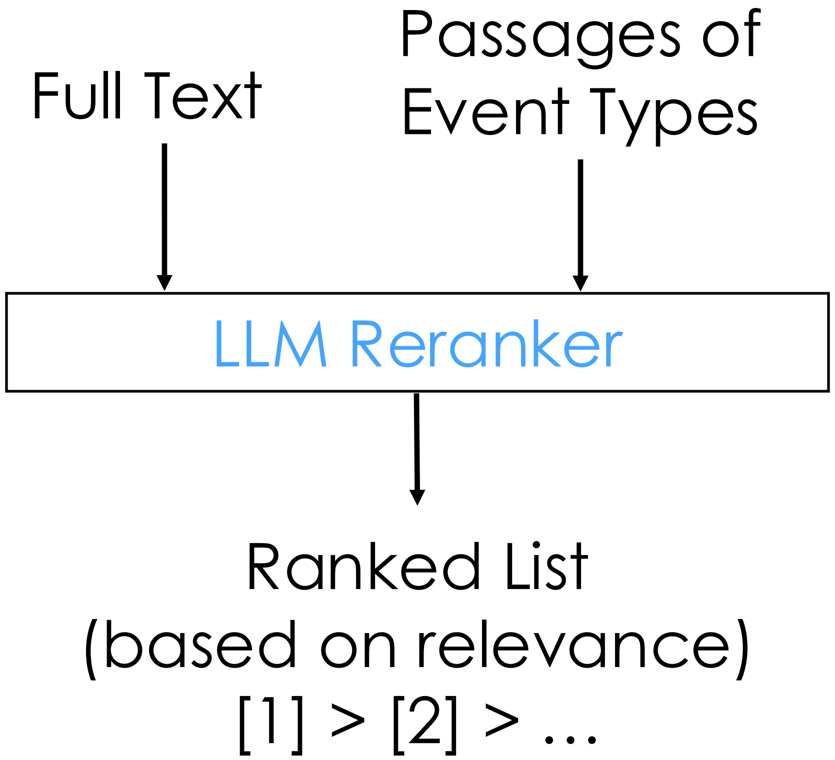
https://www.indybay.org/newsitems/2024/01/15/18862100.php, Accessed 11/5/2024

Event Types in the Codebook

Event Type	Expert Description (summarized)
Peaceful Protest	Protestors do not engage in violence
Protest w/ Intervention	Protestors are faced with a physical attempt to disperse or suppress without serious or lethal injuries
Excessive Force Against Protesters	Protestors are targeted with lethal violence or violence resulting in serious injuries
Violent Demonstration	Protestors engage in violence and/or destructive activity
[20 more event types]	
Chemical Weapon	Chemical weapon is used in warfare, as listed as Schedule 1 of the Chemical Weapons Convention of 1993

Abstractive Event-type Detection (AED): a Classification Task

- Formulate as a reranking problem of retrieved articles



Event Detection Zero-Shot LLM

(Only 25 sub-event types)

```
# instruction
You are tasked with determining the best matching Event types for a given news article. You will be
    provided with annotation guidelines and a news article to analyze. Your goal is to identify
   the most relevant event types and rank them in order of their match to the article content.
# input
Here is the news article you need to analyze:
{{ article }}
Now, carefully review the annotation guidelines for various event types:
{%
[{{ loop.index }}] "{{ ed[0] }}": {{ ed[1] }}
{%
1. For each event type, determine how well it matches the article content. Consider the following
   factors:
  - How closely the event description aligns with the main focus of the article
  - The presence of key actors or entities mentioned in the event type description
  - The occurrence of specific actions or outcomes associated with the event type
2. Rank the event types based on their relevance to the article content. Only include event types
   that have a meaningful connection to the article.
3. Output your results using the following format:
   - List the relevant event types in descending order of match quality
   - Use the ">" symbol to separate the event types
Your output should look like this:
[Explain your reasoning for the event types you decide to include, and their order]
event_type_1 > event_type_2 > ...
Provide only the ranked list of event types in your final answer.
```

Beyond 20ish Passages: RankGPT

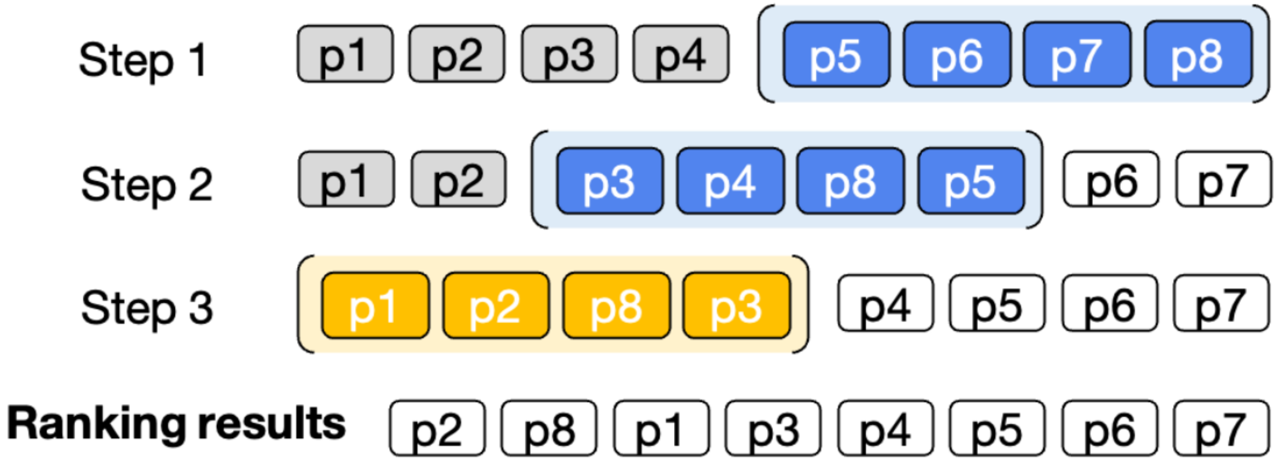


Figure 3: Illustration of re-ranking 8 passages using sliding windows with a window size of 4 and a step size of 2. The blue color represents the first two windows, while the yellow color represents the last window. The sliding windows are applied in back-to-first order, meaning that the first 2 passages in the previous window will participate in re-ranking the next window.

Experiments in Sun et al.'s paper:

Hyperparameters
Window size = 20; Step size = 10

(Included here FYI, not used in Zest)

Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents, Sun et al, EMNLP 2023

Problem 2. Abstractive Event Argument Extraction (AEAE)

- Now that we know the event type,
  zoom in on the relevant part of the codebook

Protests

Event Argument	Expert Description (summarized)
Protestors	List of protestor groups or individuals involved in the protest
Location	Where the event happened
Crowd Size	Estimated size of the crowd. It can be an exact number, a range, or a qualitative description like 'small'.
Counter Protestors	Groups or entities engaged in counter protest, if any

Attempt 1: Abstractive Event Argument Extraction with LLMs

```
# instruction
Extract event arguments from a given news article.

# input
First, here is the news article you need to analyze:
{{ article }}

Now, carefully review the event argument guidelines for the {{ event_type }}:

{% for ea in event_type.event_arguments %}
  {{ ea.name }}: {{ ea.description }}
{% endfor %}
```

LLMs Can Struggle to Follow Complex Instructions

- Types constraints, according to the codebook
  - E.g. crowd size should be a string, not number
- Nested event arguments
  - E.g. “Location” can have a “country” argument
- Relationship between different events

Solution: LLMs Know Python Data Structure Syntax!

- Use Python classes to represent the event signature
- Event types as classes
  - Event arguments as **typed** fields

Code4Struct: Code Generation for Few-Shot Event Structure Prediction, Wang et al, ACL 2023

Coding Guidelines as Class Definitions

```
class Protest(ACLEDEvent, ABC):
    """
    A "Protest" event is defined as an in-person public
demonstration of three or more participants in which the
participants do not engage in violence, though violence may be
used against them … Excludes symbolic public acts such as
displays of flags or public prayers, legislative protests, such
as parliamentary walkouts or members of parliaments staying
silent, strikes …
    """
    location: Location = Field(..., description="Location where
the event takes place")
    crowd_size: Optional[str] = Field(..., description="Estimated
size of the crowd. It can be an exact number, a range, or a
qualitative description like 'small'.")
    protestors: List[str] = Field(..., description="List of
protestor groups or individuals involved in the protest")


class PeacefulProtest(Protest):
    """
    Used when demonstrators gather for a protest and do not
engage in violence or other forms of rioting activity, such as
property destruction, and are not met with any sort of violent
intervention.
    """
    counter_protestors: List[str] = Field(...,
description="Groups or entities engaged in counter protest, if
any")
```

```
class Location(BaseModel):
    """
    The most specific location for an event. Locations can be
named populated places, geostrategic locations, natural
locations, or neighborhoods of larger cities.
    In selected large cities with activity dispersed over many
neighborhoods, locations are further specified to predefined
subsections within a city. In such cases, City Name - District
name (e.g. Mosul - Old City) is recorded in
"specific_location". If information about the specific
neighborhood/district is not known, the location is recorded
at the city level (e.g. Mosul).
    """

    country: str = Field(..., description="Normalized name of
a country, e.g. United States")
    address: str = Field(..., description="Full address or
location description including all geographic levels up to the
neighborhood level, including village/city, district, county,
province, region, country, if available. Exclude street names,
buildings, and other specific landmarks.")
```

Abstractive Event Argument Extraction with Python

In Pydantic library

```
# instruction
Extract event arguments from a given news article.

# input
First, here is the news article you need to analyze:

{{ article }}

Now, carefully review the annotation guidelines for the {{ event_type }}:

<class definitions (on last slide)>
```

Under the hood:

LLM is Good at Structured Output:
Constrained Decoding for Structured Outputs

1. Convert Python class definitions to JSON schema
2. Convert the JSON Schema into a context-free grammar
3. Pass the prompt and JSON schema to the LLM
4. Decode the output from LLM, choose the most likely token **that conforms to the grammar → this is the "constrained" part**
5. Convert the decoded JSON object to the original Python object

- Some commercial LLM providers support it, e.g. OpenAI
- Many open source LLMs support it via
  - SGLang, Outlines, guidance

https://platform.openai.com/docs/guides/structured-outputs/introduction, Accessed 11/5/2024

Event Argument Extraction with JSON

```
PeacefulProtest(
    protestors=[
    "workers",
    "musicians",
    "ILWU",
    "CWA",
    "Teamsters Local 70",
    "PMWG",
    "former workers at Tesla",
    "Past ILWU president Brian McWilliams",
    "Nadya Williams"
    ],
    location=Location(
        country="United States",
        address="Fremont, California,
                 United States"
    ),
    crowd_size="more than 100",
    counter_protestors=[],
)
```

Tesla Fremont MLK Rally Protesting Racism, Union Busting

Workers and musicians rallied on 1/13/24 at the Tesla assembly plant in Fremont, California on MLK weekend to protest Elon Musk's systemic racism and sexism at the Tesla assembly plant. They also protested the union busting and rallied in solidarity with striking Swedish Tesla service mechanics ... Workers from the ILWU, CWA, Teamsters Local 70, and PMWG spoke in solidarity, as well as former workers at Tesla. ...

Past ILWU president Brian McWilliams joined the Tesla MLK action and spoke ...

Rally participant Nadya Williams talked about her son, who is a Swedish American union organizer and is supporting the striking Swedish Tesla mechanics ... By noon, there were more than hundred workers ...

Problem 3. Abstractive Entity Linking (AEL)

- "Entity" must refer to an item on the List
  - So you can link the different records to the same entity
- For example, in ACLED:
  - Generic entities like "Labor Group" annotated to aid analysis of labor issues
  - Specific Unions are monitored

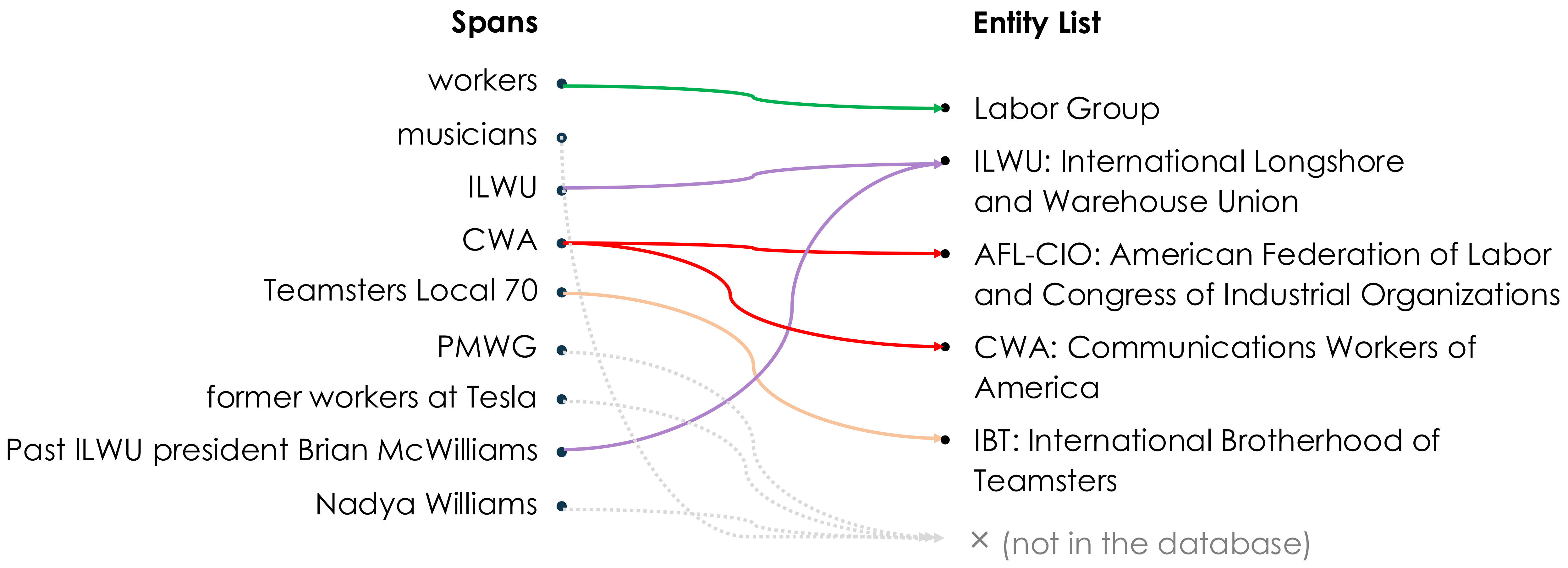
```
protestors=[
"workers",
"musicians",
"ILWU”,
"CWA",
"Teamsters Local 70",
"PMWG",
"former workers at Tesla",
"Past ILWU president Brian McWilliams",
"Nadya Williams”
]
```

Abstractive Entity Linking (AEL)

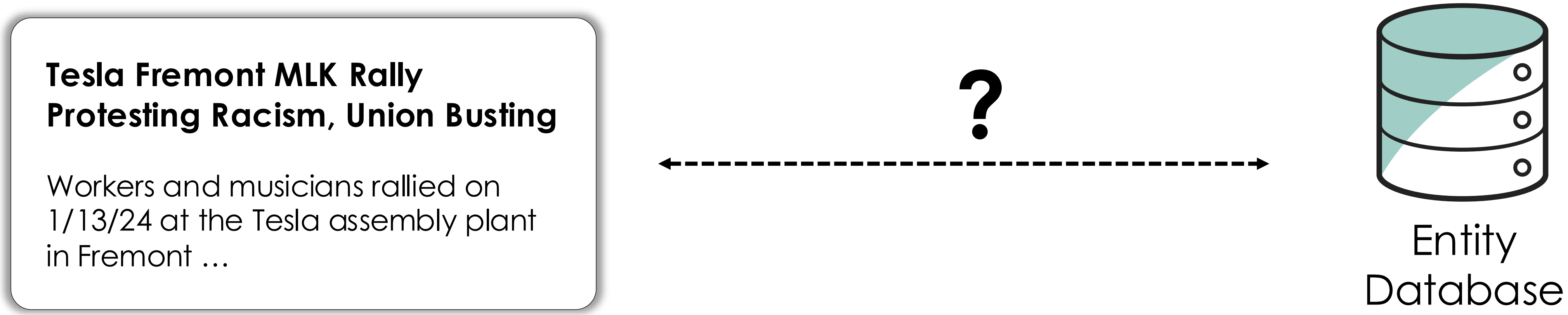
Given a database of entities, select all relevant to the event

Entity	Expert Description (summarized)
Labor Group	A collective entity composed of workers or trade unions that advocate for labor rights and interests …
AFL-CIO: American Federation of Labor and Congress of Industrial Organizations	The largest federation of unions in the United States. Encompasses various affiliated unions, such as the IW, IUPAT, **CWA** …
CWA: Communications Workers of America	A labor union representing workers in telecommunications, media, manufacturing, healthcare, and public service …
IBT: International Brotherhood of Teamsters	A labor union in the United States representing transportation and logistics workers. Operates through local chapters, like **Teamsters Local 70** …
ILWU: International Longshore and Warehouse Union	A labor union representing dock workers and other maritime and warehouse employees primarily in the United States …
[6212 more entities]	

Extractive (Spans) vs. Abstractive (Meaning)



How To Perform Abstractive Entity Linking (AEL)?



- Link event arguments of the article to possibly 10K+ entities
- Is like a retrieval task:
  - Given a document, retrieve the most relevant entities
  - BUT: Off-the-shelf retriever + LLM reranker don’t work well

Zest Abstractive Entity Linking

1. **Entity Retrieval (Event text, vector database of 10K entities → 500 entities)**
    - Guess 5 arguments
    - For each argument
        - Create a self-contained description on the entity (extract from text + LLM knowledge)
        - Find top 100 matches based on cosine similarity in the vector database
2. **Entity Filtering (Event text, 500 entities → ≈23 entities)**
    - In batches of 100, filter out irrelevant entities
3. **Entity Assignment (Event text, event-type/entity-arguments, ≈23 entities → Argument linking)**
    - Perform extraction with JSON structure

      { "field_name 1" : ["entity 1", "entity 2", ...],
      "field_name 2" : ["entity 3", "entity 4", ...], …}

Paper Contributions

Define the Abstractive Event Extraction Task

Lemonade: a High-Quality Real-World Event Dataset

Zest: Zero-Shot Qualitative Coding

Evaluation (Fine-tuning vs. Zero Shot LLM)

Task-Specific Baselines

- Gollie: Previous SOTA Event Detection Method
  - A model specifically instruction-tuned from CodeLLaMA for information extraction tasks
- Aya Expanse:
  - 8B-parameter model optimized for 16 of our 20 languages

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In ICLR, 2024.

John Dang, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *ArXiv preprint*, abs/2412.04261.

Event Detection

Model	All	en	es	ar	fr	it	ru	de	tr	my	id	uk	ko	pt	nl	so	ne	zh	fa	he	ja
Zero-shot																					
GPT-4o	**79.6**	**72.2**	**76.0**	**73.8**	**73.0**	**89.0**	**76.6**	**88.8**	**84.8**	**71.4**	**78.0**	**75.6**	**85.6**	**74.2**	**90.1**	**80.4**	**77.4**	**85.0**	**83.8**	**76.2**	**86.0**
GPT-4o mini	69.8	65.0	66.8	65.2	64.4	85.4	68.8	83.6	77.0	51.6	74.2	68.8	71.8	72.4	86.6	58.1	64.5	46.4	81.6	66.9	85.7
Llama 3.1 8B	59.5	55.0	55.0	54.0	51.8	85.2	45.4	79.4	65.0	37.8	76.4	57.6	51.2	56.0	76.4	24.0	62.6	66.8	66.8	50.9	75.4
GoLLIE 7B	23.6	35.8	36.8	6.2	34.6	49.6	29.0	61.2	7.8	0.2	11.2	41.0	1.4	46.2	36.6	0.0	0.2	38.6	6.2	5.4	7.4
Trained on LEMONADE																					
XLM-RRM	85.0	76.8	83.0	76.0	73.0	95.0	74.8	93.2	94.8	**69.2**	**94.2**	75.6	98.6	88.2	92.6	**74.6**	89.1	94.6	88.0	72.3	**98.5**
Llama 3.2 1B	85.0	79.6	85.4	80.4	74.8	97.0	81.6	93.0	94.2	60.6	92.0	76.0	99.2	89.6	**95.4**	65.1	88.4	95.4	88.2	65.1	98.2
Llama 3.2 3B	86.6	81.4	86.6	**82.8**	77.2	97.0	82.8	**94.8**	95.4	68.8	93.2	77.4	99.2	89.8	94.4	66.2	88.6	96.2	89.2	70.8	97.8
Llama 3.1 8B	86.2	**82.0**	**87.2**	80.6	77.0	97.4	**83.4**	93.8	94.0	63.8	92.2	77.0	99.0	90.8	94.0	69.8	87.7	96.4	88.8	69.9	97.8
Aya Expanse 8B	**87.5**	80.4	87.0	82.6	**79.6**	**97.6**	83.2	**94.8**	**96.0**	66.2	92.8	**80.2**	**99.6**	**91.8**	**95.4**	70.9	**89.3**	**96.6**	**91.4**	**75.9**	98.2
Majority Class	50.4	31.0	33.0	15.8	29.8	91.6	23.2	86.2	66.0	19.0	86.0	40.4	98.8	42.6	82.4	49.4	77.0	89.4	63.2	40.1	98.5

Table 1: ED F_1 results on the LEMONADE test set. The best result in each setting is highlighted in bold.

Korea, Japan: Peaceful protests > 98.5%

Abstractive Event Argument Extraction

Model	All	en	es	ar	fr	it	ru	de	tr	my	id	uk	ko	pt	nl	so	ne	zh	fa	he	ja
											Zero-shot										
AC4S (GPT-4o)	**84.6**	**85.2**	**91.0**	**73.1**	**85.0**	94.1	**82.9**	**90.9**	**89.0**	**70.9**	**93.2**	**74.4**	**90.2**	**91.1**	**92.2**	**72.1**	**87.9**	**89.3**	**83.6**	**64.4**	**94.4**
AC4S (GPT-4o mini)	81.0	83.1	87.1	71.1	84.1	**94.3**	80.1	89.7	86.4	60.0	91.5	73.5	82.2	81.7	89.0	68.3	82.5	83.2	82.9	63.4	90.9
AC4S (Llama 3.1 8B)	49.0	58.9	56.4	15.5	57.5	55.0	57.1	68.1	50.5	41.7	47.4	36.5	51.4	49.2	65.5	38.1	47.5	36.6	42.8	50.9	50.2
QA (GPT-4o)	75.3	74.0	79.7	56.3	73.3	88.5	57.2	86.7	83.0	61.7	87.7	61.2	79.2	82.3	91.2	64.5	80.8	79.0	79.7	65.0	84.2
GoLLIE 7B	40.0	47.5	45.9	21.9	46.8	54.1	42.7	59.9	27.5	1.3	49.1	39.3	30.7	49.7	54.4	0.7	12.6	58.3	31.7	22.0	49.2
											Trained on LEMONADE										
Llama 3.2 1B	85.4	87.1	85.9	78.6	81.2	94.3	78.8	91.6	90.7	71.4	93.8	84.6	94.5	95.1	94.4	71.6	88.0	86.1	77.1	72.5	86.0
Llama 3.2 3B	87.7	**89.0**	88.4	79.7	86.3	95.5	83.5	93.5	93.2	**77.3**	95.0	85.4	96.1	95.7	95.0	75.6	90.2	87.3	79.8	75.4	87.1
Llama 3.1 8B	87.6	88.5	89.7	80.4	87.2	96.2	83.8	94.1	92.1	76.4	94.5	85.1	95.8	95.7	94.7	**76.6**	**91.0**	89.2	78.3	71.9	85.8
Aya Expanse 8B	**89.0**	88.9	**90.5**	**81.4**	**88.3**	**97.7**	**85.2**	**94.2**	**93.5**	76.3	**96.3**	**87.5**	**97.2**	**96.1**	**95.7**	75.4	90.3	**91.6**	**82.8**	**77.8**	**92.6**

Table 2: AEAE F_1 results on the LEMONADE test set. The best result in each setting is highlighted in bold.

AC4S: Python-Based Argument Extraction

Abstractive Entity Linking

Model	All	en	es	ar	fr	it	ru	de	tr	my	id	uk	ko	pt	nl	so	ne	zh	fa	he	ja
Zero-shot																					
ZEST (GPT-4o)	**45.7**	**49.7**	**46.6**	**46.0**	**52.2**	**44.8**	**42.3**	**41.8**	**43.7**	**44.8**	**45.1**	**51.2**	**37.7**	**50.2**	**52.8**	**55.2**	**46.4**	**55.2**	**56.4**	**33.3**	**22.4**
ZEST (GPT-4o mini)	27.2	34.1	28.7	31.8	36.8	20.2	28.5	19.7	24.4	28.8	19.2	50.5	15.6	31.3	26.6	39.7	22.1	26.2	26.6	31.2	11.2
Span (GoLLIE 7B) + OneNet	11.1	18.7	13.0	7.8	19.3	13.4	12.4	21.1	8.7	0.0	6.9	24.9	4.3	11.8	18.7	0.0	0.4	5.2	11.1	1.7	4.6
Span (GPT-4o) + OneNet	23.7	26.0	20.8	30.7	31.1	28.4	16.5	28.9	28.5	10.9	16.5	25.6	18.8	18.2	30.1	41.1	18.7	9.8	20.9	22.0	19.1
Trained on LEMONADE																					
Llama 3.2 1B	81.9	79.2	**81.7**	79.1	72.7	85.1	**81.7**	82.0	**87.9**	67.9	89.7	90.0	87.5	86.2	84.8	59.4	82.9	90.7	84.9	78.5	80.7
Llama 3.2 3B	82.1	79.6	81.0	80.5	72.7	85.2	81.2	80.7	86.4	**70.0**	89.8	**90.2**	88.1	86.9	85.0	**62.7**	85.7	91.0	84.5	78.4	79.3
Llama 3.1 8B	80.0	78.9	78.8	80.1	68.0	82.8	80.6	79.4	85.0	66.6	88.5	88.5	85.4	84.4	84.3	57.6	82.1	89.5	83.3	76.6	78.8
Aya Expanse 8B	**82.7**	**79.8**	80.5	**81.2**	**74.3**	**86.2**	**81.7**	**82.1**	87.5	69.6	**90.5**	89.4	**88.7**	**87.1**	**85.1**	60.9	**86.1**	**91.3**	**85.1**	**83.0**	**81.2**

Table 3: AEL F_1 results on the LEMONADE test set. The best result in each setting is highlighted in bold.

Kinds of Entities

	All	Seen	Unseen	Generic	Specific
Zero-shot					
ZEST (GPT-4o)	45.7	48.9	20.0	49.6	42.6
ZEST (GPT-4o mini)	27.2	31.5	8.0	31.1	25.0
Span (GoLLIE 7B) + OneNet (GPT-4o)	11.1	10.9	14.7	7.2	15.9
Span (GPT-4o) + OneNet	23.7	23.2	30.4	10.5	37.2
Trained on LEMONADE					
Llama 3.2 1B	81.9	83.7	8.8	89.4	70.9
Llama 3.2 3B	82.1	83.9	10.6	89.2	71.8
Llama 3.1 8B	80.0	81.8	9.4	88.3	68.0
Aya Expanse 8B	82.7	84.5	12.6	89.8	72.4

Table 4: AEL F_1 results on the LEMONADE test set, grouped by entity categories.

Improving Abstractive Entity Linking

- Entity linking traditionally links to Wikidata (Unique QIDs)
- But most actors in ACLED are not in Wikipedia or Wikidata
- Future work:
  - Perform research on the actors (using ACLED data)
  - Create Wikipedia/Wikidata entrees
  - Use Wikipedia/Wikidata for entity linking

End-to-End Evaluation

Training Data	ED	AEAE	AEL	All	English
-	GPT-4o	AC4S (GPT-4o)	Zest (GPT-4o)	**58.3**	**55.9**
-	GPT-4o	AC4S (GPT-4o)	Span (GPT-4o) + OneNet	54.6	51.0
-	Llama 3.1 8B	AC4S (Llama 3.1 8B)	Zest (Llama 3.1 8B)	20.6	21.2
-	GoLLIE 7B	GoLLIE 7B	Span (GoLLIE 7B) + OneNet	14.2	18.3
LEMONADE (all of train set)		Aya Expanse 8B		**78.4**	**71.6**
LEMONADE (10% of train set)		Aya Expanse 8B		68.2	65.0
LEMONADE (5% of train set)		Aya Expanse 8B		65.5	59.2
LEMONADE (1% of train set)		Aya Expanse 8B		57.9	48.9
LEMONADE (English subset of train set)		Aya-Expanse 8B		64.0	71.3

Table 5: End-to-end F_1 results on the LEMONADE test set. The best result in each setting is highlighted in bold. Supervised experiments include training on the entire training set of LEMONADE, training on randomly sampled subsets of it, and only on its English subset.

Our best model outperforms previous work by 44 points!
Automatic qualitative coding is not good enough!

Conclusion

- **Automatic Qualitative Coding (AQC)**
  - Has many applications
- **Lemonade: the best annotated event dataset excerpted from ACLED's real data**
  - With abstract event extraction
  - Fine-tuning reaches 78.4 F1
- **Zest: an AQC designed for real-life problems**
  - Contribution: abstractive entity linking
  - SOTA zero-shot performance

Note: Huge differences between academic approaches and demands of real life!
Further research is needed!