USING LLM FOR IMPROVING KEY EVENT DISCOVERY: TEMPORAL-GUIDED NEWS STREAM CLUSTERING WITH EVENT SUMMARIES

Nishanth Nakshatri[1], Siyi Liu[2], Sihao Chen[2], Daniel Hopkins[2], Dan Roth[2], Dan Goldwasser[1]

Conference: *Findings of EMNLP 2023*

[1]Purdue University

[2]University of Pennsylvania

PURDUE UNIVERSITY | Department of Computer Science

Using LLM for Improving Key Event Discovery

Motivation

- Researchers analyze news media/news articles
  - Characterize discussion around a real-world *news-event*
  - Understand public opinion, examine framing, track changes over time …
- *Challenge:* Ever-growing amount of news information
  - How do we get "key" news events?
  - Can we identify them without any human-intervention?
- Need for automated *Concept Learning*
  - **Idea:** Exploit recent advances in LLM to improve event discovery
    - Retrieve *event candidates* using traditional clustering algorithm
    - Use LLM to characterize *event candidates*, and reason about their validity

Using LLM for Improving Key Event Discovery

Our Contribution

We propose a generic framework for news-stream clustering

- **Inspired by interactive-clustering**
- **Unsupervised setting**

***KeyEvents* – a coherent dataset**

- **11 topics**
- **611 *key* events**

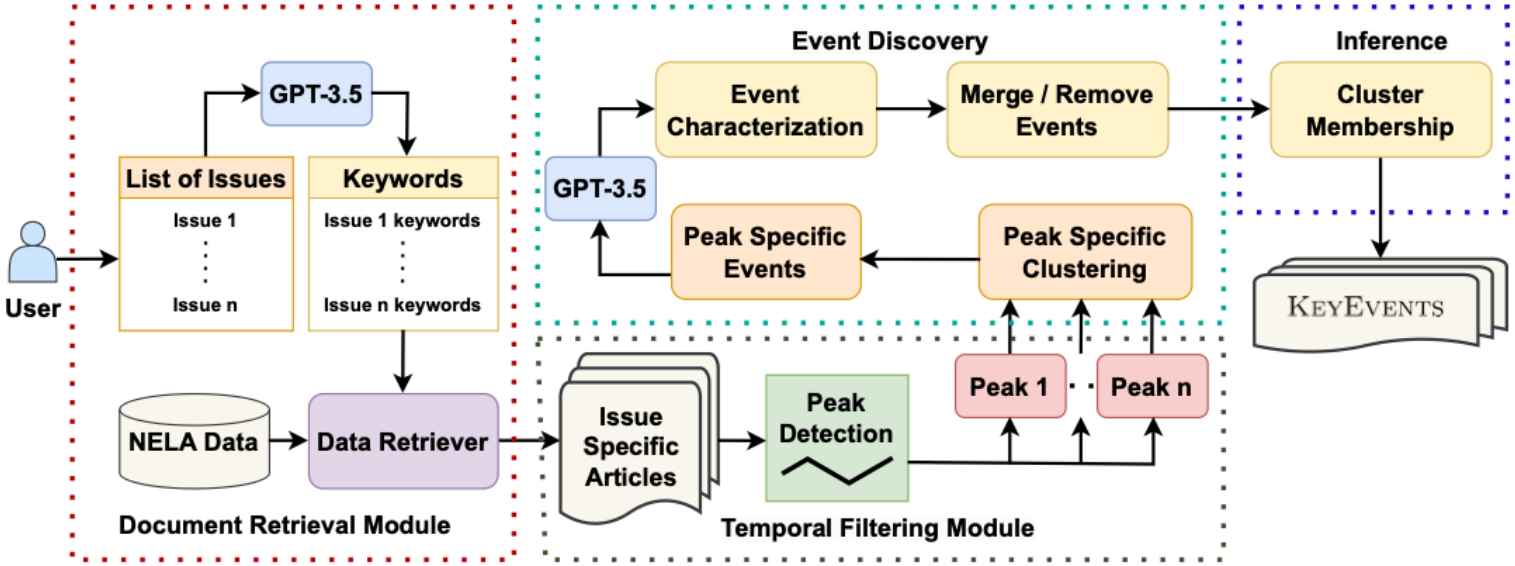
Overview

- Framework
- Evaluation Metrics
- Results
- Broader Impact
- Conclusion

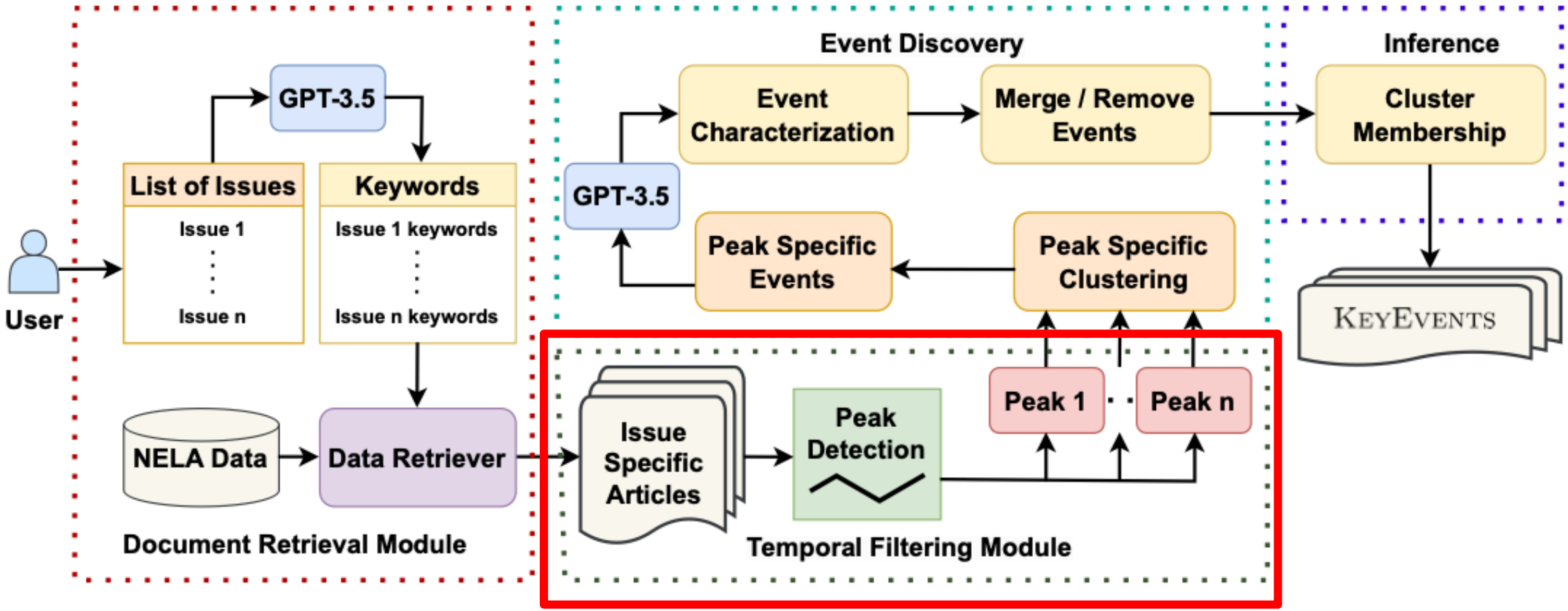
Using LLM for Improving Key Event Discovery

Framework

- **Three modules**
  - Temporal Filtering
  - Event Discovery
  - Inference
- **Dataset**
  - NELA Dataset (2021)
    - ~1.8M news article collection
  - **Preprocessing**
    - Document Retrieval Module



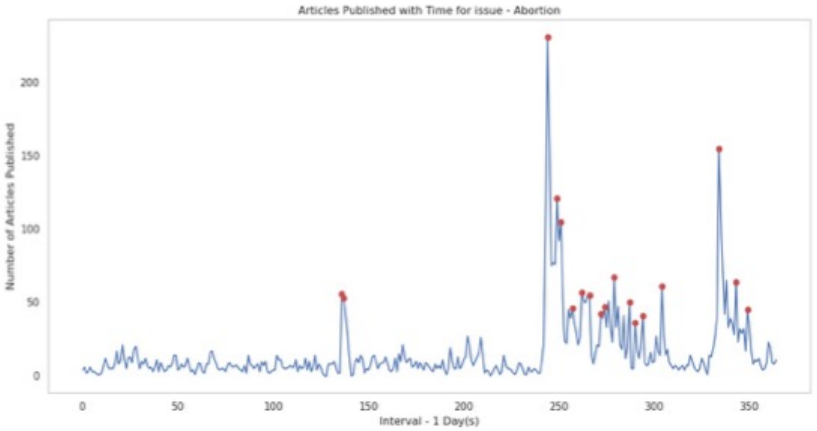
Using LLM for Improving Key Event Discovery



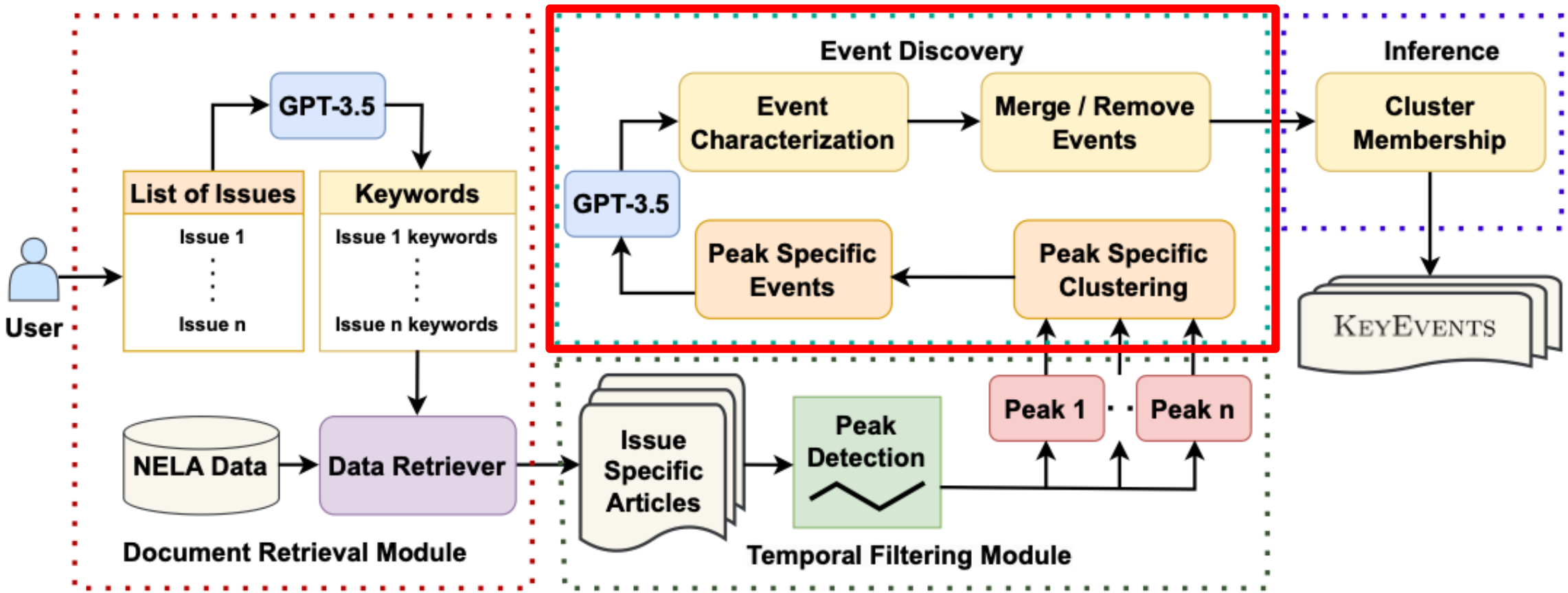
Using LLM for Improving Key Event Discovery

Module 1: Temporal Filtering

- **Input**
  - News articles related to an issue/topic
  - Ex. *Climate Change*, *Abortion etc.*
- **Identify potential real-world *events***
  - Dynamic analysis of articles
  - Outlier detection algorithm
    - *Temporal landmarks/peaks*
- **Output**
  - Set of news articles at various points in time



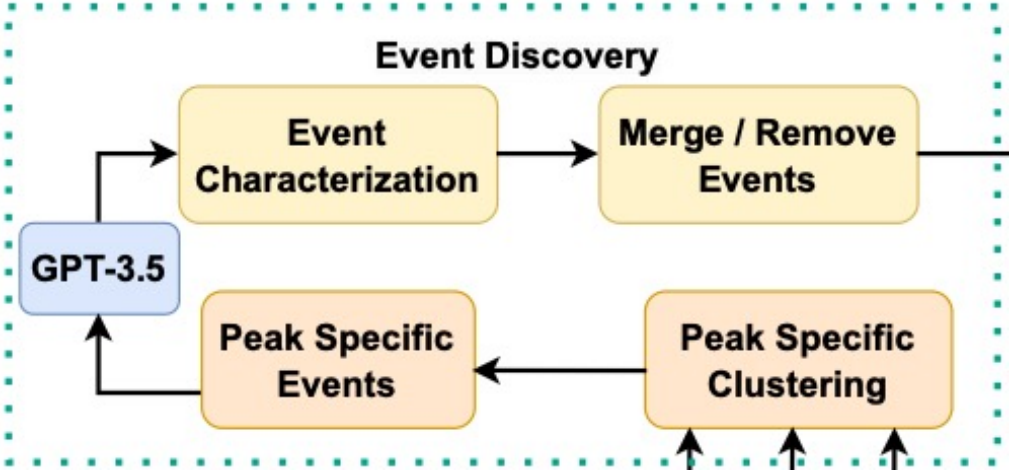
Using LLM for Improving Key Event Discovery



Using LLM for Improving Key Event Discovery

Module 2: Event Discovery

- **Temporal Filtering is still prone to noise**
  - Multiple relevant events at each peak
- **Goal**
  - Cluster news articles at each peak → Events
- **Three Steps**
  - Peak-Specific Clustering
  - Event Characterization
  - Merge/Remove Events



Using LLM for Improving Key Event Discovery

Event Discovery

- **Peak-Specific Clustering**
  - Embed the news article using dense-retriever model
  - Cluster the articles at each peak using HDBSCAN algorithm

- **Event Characterization**
  - Events obtained are still prone to noise
  - Characterize the event using a multi-document summary (LLM)
    - Use closest-K documents to *event centroid*
    - Generate a short summary for each event

Using LLM for Improving Key Event Discovery

Event Discovery

- **Remove Incoherent Event Clusters**
  - Closest-K documents do not align with generated summary → *Incoherent Cluster*
  - Compare summary embedding to document embedding

Incoherent Cluster (Top-3 documents shown)
Event Title: Climate Justice and African Activists **Event Description**: This is about the challenges faced by African climate activists in bringing attention to the climate crisis and the need for climate justice.
Doc. 1: *There Will Never Be Climate Justice If African Activists Keep Being Ignored* We go to Kampala, Uganda, to speak to climate activist Vanessa Nakate on the occasion of her first book being published, A Bigger Picture. ...
Doc. 2: *The Looking Glass World Of 'Climate Injustice'* In our wacky world where almost nothing makes sense anymore, there is no shortage of examples of politicians, let alone self-important academics, journalists, and wealthy elites, looking foolish with self-contradictory policy demands. ...
Doc. 3: *New Miss Universe Urges Action on Climate Change: Choice to Kill or Save Nature* A new Miss Universe has been crowned and she is a climate alarmist. ...

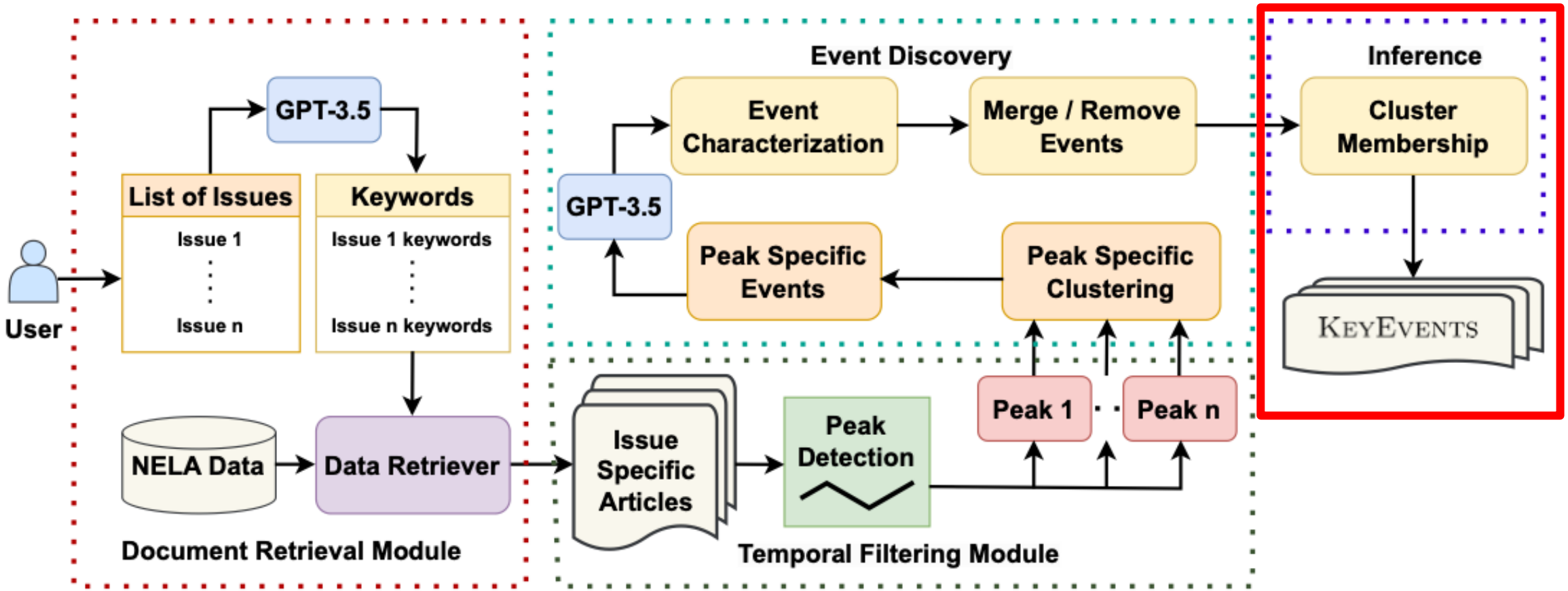
Using LLM for Improving Key Event Discovery

Event Discovery

- **Merge Events**
  - *Two similar event summaries → Merge*
  - Greedy approach
  - **First Iteration:**
    - For all event-pair combinations
    - Prompt LLM
    - Merge similar pair of event summaries
  - **Second Iteration:**
    - New event-pair combinations
    - Check for merge

Summary of Article 1	Summary of Article 2
Event Title: President Biden's Climate Plan **Event Description**: This is about President Joe Biden's executive orders aimed at tackling climate change by reducing the U.S. carbon footprint and emissions, stopping oil and gas leases on public lands, and prioritizing climate change as a national security concern.	**Event Title**: Biden's Climate Change Actions **Event Description:** This is about President Joe Biden's executive actions to combat climate change by prioritizing science and evidence-based policy across federal agencies, pausing oil drilling on public lands, and aiming to cut oil, gas, and coal emissions.
Event Title: Texas Abortion Ban **Event Description**: This is about a new Texas law that bans abortions after 6 weeks and empowers regular citizens to bring civil lawsuits against anyone who aids a woman looking to terminate a pregnancy.	**Event Title**: Texas Abortion Law **Event Description**: This is about the controversial Texas abortion law that bans abortions after six weeks and has been condemned by President Joe Biden as an unprecedented assault on women's rights.

Using LLM for Improving Key Event Discovery



Using LLM for Improving Key Event Discovery

Inference

- **Decide Cluster Membership**
  - Based on a similarity module
  - Expand the document set
    - Consider documents from one-day before and one-day after the peak
  - Compare document embedding to *generated event summary embedding*

Overview

- Framework
- Evaluation Metrics
- Results
- Broader Impact
- Conclusion

Using LLM for Improving Key Event Discovery

Evaluation Metrics

- Three automatic metrics and human evaluation to measure **coherency**
- Evaluate across 11 issues
- Metrics
  - **Entity Purity** (higher the better)
    - Percentage of documents that mention at least one of top-10 TF-IDF entities
  - **Coverage** (higher the better)
    - Percentage of documents accounted for in clustering process
  - **Entity Coherence** (value closer to zero → highly coherent cluster)
    - Considers co-occurrences of central entity pairs in clustered documents

$$C(e_i, V^{e_i}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{F(v_m^{e_i}, v_l^{e_i}) + \epsilon}{F(v_l^{e_i})}$$

Overview

- Framework
- Evaluation Metrics
- Results
- Broader Impact
- Conclusion

Using LLM for Improving Key Event Discovery

Results

- Temporal Filtering is effective
- Coverage vs Purity Trade off

Model	Coverage ↓	Entity Purity ↑	Entity Coherence ↑	Event Count
LDA (baseline)	99.69	31.52	-1008.42	60.0
Temporal filtering	-	28.15	-1061.60	18.7
LDA (Temporal)	89.02	38.62	-1005.37	65.7
HDBSCAN	81.78	62.55	-776.80	58.4
BERTopic	84.04	66.00	-726.11	62.3
Our Method	44.29	**82.69**	**-477.89**	55.5
Our Method (iter 2)	56.83	77.49	-579.48	55.5

Aggregated Statistics across 11 issues/topics

Using LLM for Improving Key Event Discovery

Ablation Analysis

- **Merge/Remove Impact**
  - Cosine similarity reduction → Increased distance between events
  - Merge/remove operation
    - Increased cluster separation

Issue	Model	Avg. Inter-Event Cosine Similarity	# Events	# Merge Operations	# Remove Operations
Capitol Insurrection	HDBSCAN	0.864877655	64	-	-
	Our Method	**0.641329667**	40	21	3
Coronavirus	HDBSCAN	0.860832152	122	-	-
	Our Method	**0.857558543**	112	10	2
Climate Change	HDBSCAN	0.833522985	74	-	-
	Our Method	**0.772742185**	56	11	7
Free Speech	HDBSCAN	0.847346069	72	-	-
	Our Method	**0.668949583**	56	7	13
Abortion	HDBSCAN	0.877382542	48	-	-
	Our Method	**0.410449078**	24	20	4
Immigration	HDBSCAN	0.852341823	64	-	-
	Our Method	**0.75051009**	48	15	1
Gun Control	HDBSCAN	0.829052923	60	-	-
	Our Method	**0.663993032**	40	9	9
Criminal Injustice & Law Enforcement	HDBSCAN	0.824876478	70	-	-
	Our Method	**0.581169596**	48	7	13
Racial Equity	HDBSCAN	0.839611843	98	-	-
	Our Method	**0.730141103**	68	13	17
Defense and National Security	HDBSCAN	0.837432569	106	-	-
	Our Method	**0.835570683**	89	11	6
Corruption	HDBSCAN	**0.818098607**	46	-	-
	Our Method	0.821913246	30	5	31

Using LLM for Improving Key Event Discovery

Human Evaluation (on issue *Climate Change*)

- Metrics
  - **Event Coherence**
    - Top-K documents are in agreement with each other → event is coherent
  - **Mapping Quality**
    - Verify validity of assignments
    - Agreement between document and event summary
- **Observation**
  - Other methods are more prone to noise

Model	Event Coherence ↑	Mapping Quality (Precision) ↑
HDBSCAN	84.90	62.27
BERTopic	85.48	69.87
Our Method	**91.07**	**72.19**

Human evaluation results

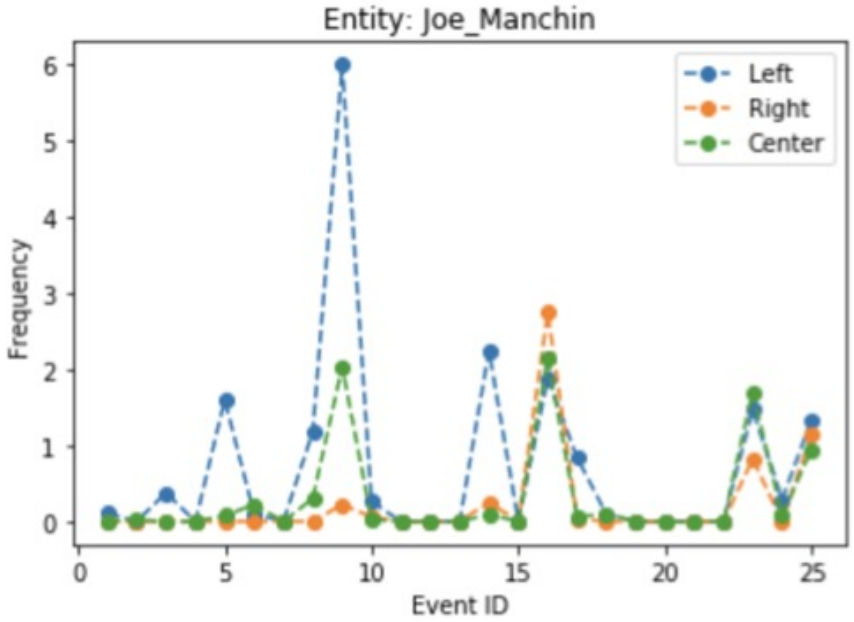
Overview

- Framework
- Evaluation Metrics
- Results
- Broader Impact
- Conclusion

Using LLM for Improving Key Event Discovery

Broader Impact – Usefulness of our method & dataset

- **Simple Case Study**
  - How entity portrayal varies across ideologies?
  - Entity: ***Joe Manchin*** (democratic senator)
  - ***Analysis***
    - Left-leaning mention more (5th, 9th, 14th events)
      - Criticize his ties to coal industry
    - Sentiment towards this entity (16th event)
      - No positive sentiment in any ideology
      - *Left:* 86% of articles indicate negative
      - *Right:* 38% of articles indicate negative



Overview

- Framework
- Evaluation Metrics
- Results
- Broader Impact
- Conclusion

Conclusion

- We proposed a framework for *key events* identification
- With two forms of evaluation: automated and human-evaluation
- Showed a simple case study on the usefulness of the framework

THANK YOU

For enquires, contact me via email: nnakshat@purdue.edu