Extracting protest events from newspaper articles with ChatGPT[1]

Neal Caren, University of North Carolina, Chapel Hill, neal.caren@unc.edu
Kenneth T. Andrews, Washington University, St. Louis
Rashawn Ray, University of Maryland

Abstract: This research note examines the abilities of a large language model (LLM), ChatGPT, to extract structured data on protest events from media accounts. Based on our analysis of 500 articles on Black Lives Matter protests, after an iterative process of prompt improvement on a training dataset, ChatGPT can produce data comparable to or better than a hand-coding method with an enormous reduction in time and minimal cost. While the technique has limitations, LLMs show promise and deserve further study for their use in protest event analysis.

Media accounts are the dominant source of information about political protest events. Historically, this process has been done by hand, with the researcher or their research assistants reading newspaper articles and extracting the relevant information based on rules established in a codebook[2]. Scholars have attempted to automate parts of this process, which can be particularly useful for identifying and pre-processing the articles for hand-coding. This research note examines the capacities of a new large language model, OpenAI's gpt-3.5-turbo, commonly called ChatGPT, to aid in the automation process.

Large language models (LLM), such as ChatGPT, have demonstrated exceptional proficiency in processing and analyzing massive volumes of textual data, thus presenting a promising solution for numerous natural language processing (NLP) applications(e.g., Gilardi, Alizadeh, and Kubli 2023). One potentially suitable extension application is extracting event information from newspaper articles. By leveraging the sophisticated linguistic capabilities of these models, research can extract protest information by identifying and extracting salient entities, actions, and events from the input text in a structured format.

Researchers use media accounts to extract a variety of information about protest events, such as where, including exact location where possible but at least city and state; date; event size;

[1] We thank Anabell Diaz Amarante, Lauren Brodeur, Mia Neal, and Zoe Turner for their research assistance, and Weiwei Yuan for her comments on an earlier draft of this paper. Data collection was generously funded by the Russel Sage Foundation.
[2] Two recent articles (Fisher et al. 2019; Oliver, Hanna, and Lim 2023) include overviews of historical and contemporary practices for constructing protest event datasets from media accounts.

forms of activities; whether police were present; whether and how many people the police arrested; whether there were injuries, deaths or property damage; what organizations were present; and the purpose or claims put forward at the event, as shown in Figure 1. Extracting some information is usually straightforward, such as whether the police were present.[3] Other characteristics can be more complicated, as the article may give a size range, describe the number arrested but not the total crowd size, or ambiguously state that "a large group" was present. However, identifying when the article references the size is usually straightforward, even if the answer is not. In contrast, inferring the date and location frequently involves triangulating the article source and publication date. A local newspaper might, for example, note that protesters "marched down Franklin Street by campus" but never list the city, assuming readers would know based on the coverage zone of the paper. Alternatively, a "Portland" protest might, depending on the media's location and article time frame, refer to either Portland, ME, or Portland, OR. Likewise, journalists describe dates relative to the publication date, such as "yesterday" or "Friday." In both these cases, article text alone is often insufficient for extracting important event characteristics.

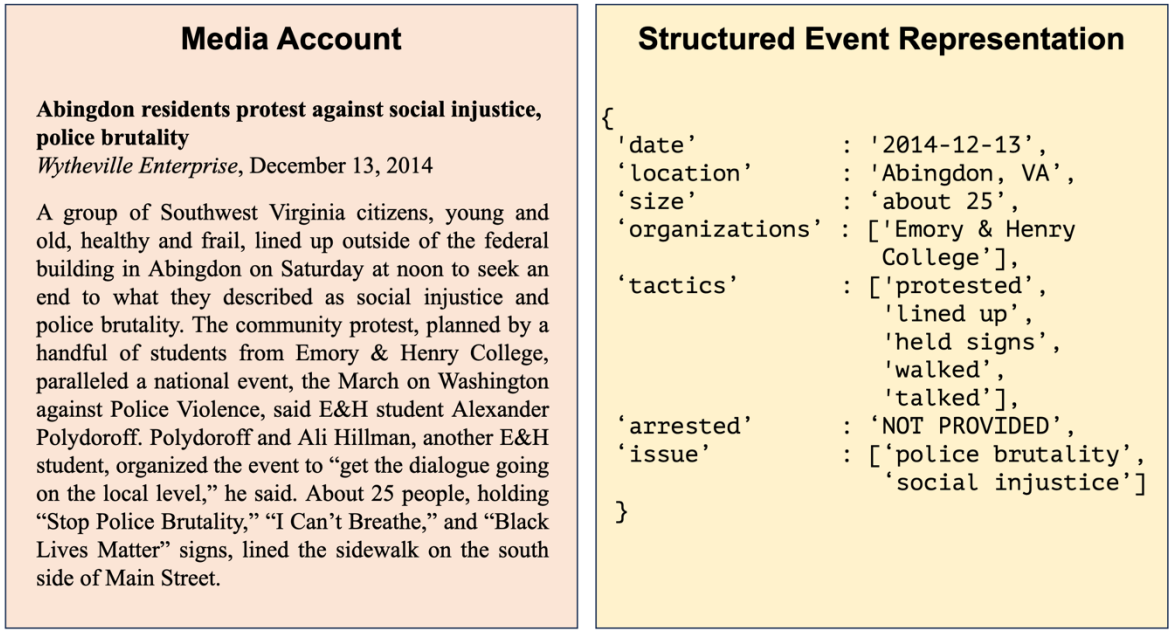

Figure 1. Example newspaper article and a structured representation of the relevant protest characteristics in JSON format.

Given the labor intensity of manually extracting events, even in a computer-assisted workflow, we attempted to ascertain whether advances in natural language processing could facilitate the process. Preliminary efforts by the first author to use earlier LLMs for protest event extraction were unsuccessful. While the models were quite good when the sentences were clearly written, and all the information was present, they were unreliable when either of those conditions was absent, confusing, or required calculations to infer the correct location and date. The models

[3] As scholars are aware, the absence of a media account about police doesn't imply the absence of police at the event.

would confidently return the wrong date if the article mentioned several different events, even if only one was a protest. They would also provide wrong answers with incorrection information not from the article, especially when the relevant information was not present. This error was observed to be correlated with the model's training data, as "Baltimore" or "Ferguson" was often returned as the protest city when none was listed. This process is referred to as one of model "hallucination" or "confabulation,"

The current study

To test the capabilities of the ChatGPT model gpt-turbo-3.5 to extract event data from newspapers, we compared its results to the results of our current system using undergraduate hand-coding. We used a sample of two hundred articles to refine the ChatGPT workflow and prompts, and we tested the accuracy on a sample of 500 articles that had already been coded in the old system. This analysis focuses on date and location matching.

Test Data

Data for this test comes from a Russel Sage Foundation-funded project to understand the impact of the Black Lives Matter movement on police policy reforms across the United States. As part of the research project, we have developed a hybrid machine-learning/hand-coded system for extracting event attributes. We first used a machine learning model to identify articles that were likely about protests; then undergraduate research assistants confirmed the relevance of the article and highlighted relevant portions of the text related to specific attributes; we then used a series of scripts to clean the data, and compute location and date alignment, such as assigning the date prior to publication as the event date when the article described it as "yesterday" or assigning "Raleigh, NC" if the location was only described as the "State Capitol building" in a Durham, NC newspaper article; grouping articles related to the same event; and, finally, hand reviewing the results for missing data and likely errors. The time-consuming process involves several passes through the data by both machines and humans, improving accuracy.

This test used a random sample of 532 media articles published between 2010 and 2016 drawn from Newsbank coded by researchers as describing a protest against police brutality. As all articles in the sample are relevant, this study tests only one part of a much larger workflow, which includes article identification before this stage and event deduplication afterward. We also focused on event date and location, the most complex characteristics to extract for the earlier computer models and the research assistants.

Model

The LLM used in this study, gpt-turbo-3.5 (ChatGPT), was released by OpenAI on March 14, 2023. LLMs are a type of artificial intelligence system that has been built using a neural network architecture known as a transformer. At a high level, transformers consist of a series of layers of neural network cells that work together to learn patterns in data. When used for natural language processing, transformers are trained on massive amounts of text data (none produced

after 2021 in the case of ChatGPT), which allows them to learn to recognize patterns in language, such as which words tend to appear together and what types of sentences are grammatically correct. Unlike NLP models often employed on classification tasks, models like ChatGPT are trained for text generation. In addition to being one of the more massive models, with 175 billion parameters, GPT-3.5 was also trained to reply as a prompt a ChatBot, and as of April 1, 2023, it powered most of the interactions with ChatGPT[4]. In text extraction tasks such as this, the LLM focuses, although not exclusively, on the provided texts and attempts to find the sections most relevant to answering the prompt. Data collection and analysis were done in Python.

Workflow

The workflow and prompts used in this study were refined in an interactive process based on a sample of 200 articles not used in the evaluation set. During this initial process, it became apparent that the accuracy of complex tasks, particularly the event date, could be improved by asking the same question multiple times and with different wordings and comparing answers.

The overall workflow for event extraction using ChatGPT involved five steps:

1. **Screen**: A first-pass prompt (Prompt A1 in the Appendix) to determine the article's relevance. Determine whether the article describes one or more protests; the theme of the protest; whether the event has occurred or is in the future; and what country. Responses from this step were not used in this study.
2. **Location**: A prompt (A2) that includes the article text and requests the protest's location. This prompt requests three results.
3. **Date**: A prompt (A3) that includes the article text and requests the protest date. This prompt requests three results.
4. **Details**: A prompt (A4) that includes the article text and requests complete information on the event, including location and date.
5. **Date Check:** A prompt (A5) that includes the article text and requests the protest date using different text than earlier requests.

Step 5, the final date check, was only used when prior results returned conflicting results, or ChatGPT suggested that the event occurred five or more days before article publication. Otherwise, the original consensus date was used. Location is based on the modal response to Step 2. Even with the cost of sending the full text of the article multiple times to the OpenAI API, processing each article is less than one cent at June 2023 OpenAI pricing. The total cost to process the sample was less than $4.00.

[4] Simultaneously with the release of the GPT-3.5 model, OpenAi also released the more powerful GPT-4 model. However, that model is much more expensive for API access, with prompts cost 15x as much and response 30x as much. This shifts the cost of a project such as this dramatically. Additional prompt engineer likely can reduce the cost of GPT-4, but it still shifts the project costs by at least an order of magnitude. However, t

Findings

After repeated iterations to improve model prompts, we found that the ChatGPT model performed comparably or better than our existing hand-coding system for event coding. However, researchers are still necessary to weed out bias in the data and ensure corrections are made that will not lead to findings and interpretations that are less accurate.

Protest Date

The ChatGPT method matched our coded dates 90.6% of the time (453 out of 500) for the protest dates. In order to further evaluate sources of errors in the model, we reviewed a sample of 20 cases where dates did not match. In almost all of the articles, the protest date was not described straightforwardly. We found:

- In one case where the dates matched, but it was in the incorrect format;
- One case where the protest date was never mentioned and could not be inferred from the article and where the research assistants coded it based on additional information.
- Seven cases where the ChatGPT was incorrect, but the hand-coding system was correct. The mistakes were plausible confusions in six of those, and only one was a clear confabulation.
- Eleven cases where ChatGPT was correct and our hand-coding system was incorrect.

Among the fifty-three disagreements, the median difference in days between the two estimates was one day.

The results also suggest areas where more substantial human intervention is required. Of the 70% of cases where the ChatGPT results returned the same date to multiple requests and prompts on the same article (Steps 3 and 4), the match with the human coders was 96.6%, compared to only 75.8% with conflict results (Step 5 date used). This suggests that human coders should review the results carefully when the model returns inconsistent results.

Protest Location

On location of protest, measured by city and state, the ChatGPT method performed comparably to date, matching our coded locations 91.8% of the time (459 out of 500). In the sample of 20 mismatches, we observed the following:

- Six cases where the two processes matched, but not precisely, either because of capitalization differences or identifiable subunit issues, where the ChatGPT model reported a borough of New York City (“Manhattan”) where our coding system had “New York.”
- Two cases where the old system was correct, and the ChatGPT model was wrong. This included one case where locations were mentioned in the article, but the protest city was only in the headline, and one confabulation, where Fort Lauderdale was reported when the actual site was Fort Wayne.
- Twelve cases where the ChatGPT model was correct, and the hand-coding system was incorrect. These included cases with the correct city but the wrong state, where

protesters were from a different city than the event's location, and where "Detroit Avenue" was incorrectly identified as being in Detroit.

Based on this subsample analysis, the ChatGPT method performed better than our current system, as it was correct in 18 of the 20 cases examined.

Conclusion

Our analysis shows promise in using LLMs such as ChatGPT to extract protest event data from newspaper articles. The ChatGPT performed comparable to or better than our existing hand-coding for two complex tasks, date, and location, as shown in Figure 2. Further research should be conducted to measure the method's capacity for extracting other data about protest events. As Oliver et al. (2023) detail, protests are often complex events, and researchers attempting to represent these dynamic, relational processes accurately face numerous challenges, particularly if the goal is to link events into broader movement campaigns.

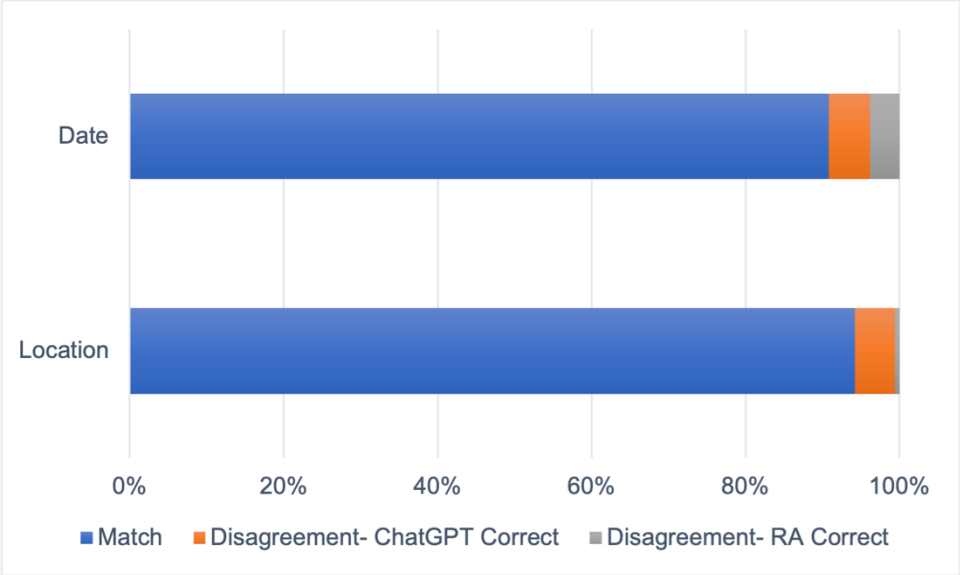

Figure 2. Agreement and accuracy between ChatGPT and Research Assistant (RA) models.

Based on our study, we propose several guidelines for research using large language models in data collection.

First, refine prompts in an interactive process. While our final ChatGPT system was accurate and reliable, the model was unreliable before carefully revising the prompts and identifying the conditions where the responses were the most inaccurate. For example, we found that responses with event dates that were more than a few days old had low accuracy. We observed that this was likely the result of articles with vague references to the date of the protest but precise dates about the events that triggered the protest. To account for this, our follow-up date prompt asks about the most "recent" event. An earlier workflow also included specific feedback

when the model returned inconsistent results, such as, “You replied that the protest happened on a Wednesday, but July 14, 2015, is a Tuesday. On which date did the protest occur?”

Second, have a sufficiently large data set accurately coded to allow for extensive training and evaluation. While LLMs may enable new forms of turning unstructured data, such as newspapers, into datasets, like prior methods, using LLMs requires a large, coded representative dataset to properly develop and evaluate the method’s accuracy for the specific task. LLMs cannot do all the work. Additionally, while splitting data into training and evaluation bins is rare in sociology, it is common in machine learning and a required practice here to avoid overfitting prompts producing artificially high accuracy rates.

Third, researchers should test the extent to which models may confabulate or otherwise provide biased results that reinforce existing inequalities in coverage, such as by returning more confrontational tactics for protests involving African American protesters.

Finally, in its current form, ChatGPT can be used to speed up data coding and analysis and as a verification and cross-check tool. But skilled researchers are still needed to identify inaccuracies correctly. In quantitative analyses or qualitative content analyses, percentage differences between AI processes and human coders can have grave implications for the academic literature, policy recommendations, and the public’s understanding of a topic of societal concern.

References

Fisher, Dana R., Kenneth T. Andrews, Neal Caren, Erica Chenoweth, Michael T. Heaney, Tommy Leung, Nathan L. Perkins, and Jeremy Pressman. 2019. “The Science of Contemporary Street Protest: New Efforts in the United States.” *Science Advances*.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. “ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks.” *ArXiv.Org*. Retrieved April 6, 2023 (https://arxiv.org/abs/2303.15056v1).

Oliver, Pamela, Alex Hanna, and Chaeyoon Lim. 2023. “Constructing Relational and Verifiable Protest Event Data: Four Challenges and Some Solution.” *Mobilization: An International Quarterly* 28(1):1–22. doi: 10.17813/1086-671X-28-1-1.

Appendix

Below are the prompts used in the study. Each evolved somewhat independently, so they do not all follow the same logic. The OpenAI ChatGPT model takes two inputs, a “system” prompt, which can broadly guide the model, and a “user” prompt to which the model will directly respond.

A1. Screen Prompt

System:

You are a scholar who studies political protests, such as marches, demonstrations, rallies, sit-ins, die-ins, strikes, pickets, sickouts, walkouts, and nonviolent direct actions.

Carefully read the article to identify all information pertaining to political protests, including the date, location, participants, demands, tactics, activities, arrests, organizations, size, police involvement, and reasons for protesting.

Use only the information provided in the article and avoid adding any additional details.

Read the article and respond to each of the following questions:
PROTEST: Does the article describe at least one protest, demonstration, march, or rally? Yes, No
MULTIPLE: Does the article describe protests occurring in different cities? Yes, No
TIMING: Has the protest already happened? Done, Ongoing, Future
ISSUE: Broad political issues associated with protest, such as “abortion.”
COUNTRY: What country does the vent take place in?
SUMMARY: Rewrite the article to focus on the protest, using the exact words and phrases. Include all relevant details about the protest, such as the date, location, size, participants, demands, tactics used, activities, and reasons for protesting. Be sure to list all people or organizations involved.
Respond in valid Python dictionary such as {“PROTEST”: “Yes”, “MULTIPLE”: “No”, “TIMING”: “Ongoing”, “ISSUE”: “CLimate Change”, “COUNTRY”: “US”, “SUMMARY”: “On Monday...”}

User:

{article_text]

A2. Location Prompt:

System: You are a helpful assistant who reads the newspaper.

User: The article appeared in "{source}" and discusses a political protest, such as a rally, march or demonstration. WHERE DID THE PROTEST TAKE PLACE? Pay attention to locations mentioned in the article.

Write your answer in valid JSON format, such as {{"PROTEST_CITY": "Chicago", "PROTEST_STATE": "IL", "PROTEST_PLACES_SPECIFIC": ["In front of City Hall."]}}

Only use state abbreviations, such as OH, LA or NY.

Article:
{article_text}

A3. Date Prompt

System:
You are a helpful assistant who reads the newspaper.
User:
The article was written on {publication_date}, but WHEN DID THE PROTEST TAKE PLACE? Pay attention to dates mentioned in the article and words such as 'yesterday,' 'last week,' and 'Monday.'
Write your answer in valid JSON format, such as {{"PROTEST_DAY": "November 17, 2015 ", "PROTEST_WEEKDAY": "Monday," "Explanation": "The article states... ."}}

Article:
{article_text}

A4. Details Prompt

System:

1. Start by reading the article thoroughly to identify the relevant information about the protest event. Look for details such as the date, location, size of the crowd, organizations

involved, number of arrests and injuries, participants, demands, grievances, and the political issue the protest is related to.

2. Once you have identified the information, organize it into the following categories for each protest event:
    - date: the date of the protest event (in YYYY-MM-DD format). Pay attention to words such as 'yesterday' or 'Monday';
    - weekday: weekday;
    - location: city, town, or township and state where protest occurred. Only list city and state abbreviations;
    - place: list of specific locations within a city with protest;
    - size: crowd size. Number of protest participants, including adjectives like "many";
    - count the specific count of the number of protesters as an integer. If not reported, make your best guess;
    - organizations: list of organizations involved;
    - tactics: list of verbs of what protesters did, such as marched, rallied, interrupted;
    - arrested: number arrested;
    - injured: number injured;
    - counterprotest; the presence of any counterprotesters;
    - who: list of participants. Descriptive words or phrases;
    - demands: list of political demands of organizers and participants;
    - grievances: list of political complaints of organizers and participants;
    - issue: list of broad political issues, such as "abortion" or "climate change."
3. Write the information for each protest event in valid JSON format using the categories listed above.
4. If any of the information is not provided in the article, write "NOT PROVIDED" for the relevant category.
5. Make sure to double-check your information and formatting to ensure that the JSON output is valid and accurate.
6. Finally, consider using a JSON validator to confirm that the output is valid and well-formed.

User:

{article_text]

A5. Date Recheck Prompt

System:

You are a helpful assistant who reads the newspaper.

User:

When did the most recent protest described in this article happen? Pay attention to dates mentioned in the article and words such as ‘yesterday,’ ‘last week,’ and ‘Monday.’
Protests can happen on the same day a story is written, so a story published on Wednesday describing a Wednesday protest occurred on the publication date.
Dates should be written in YYYY-MM-DD format.
Write your answer in valid JSON format, such as {{“PROTEST_DAY”: “2016-11-28”, “Explanation”: “The article states... .”}}

Headline: {headline}
Publication Date: {publication_date}
Source: {source}

Article:
{article_text}