Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages

Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou & Kenneth Joseph

To cite this article: Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou & Kenneth Joseph (2023) Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages, International Journal of Geographical Information Science, 37:11, 2289-2318, DOI: 10.1080/13658816.2023.2266495

To link to this article: https://doi.org/10.1080/13658816.2023.2266495

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2023, VOL. 37, NO. 11, 2289–2318
https://doi.org/10.1080/13658816.2023.2266495

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages

Yingjie Hu[a*], Gengchen Mai[b*], Chris Cundy[c], Kristy Choi[c], Ni Lao[d], Wei Liu[a], Gaurish Lakhanpal[a,e], Ryan Zhenqi Zhou[a] and Kenneth Joseph[f]

[a]GeoAI Lab, Department of Geography, University at Buffalo, Buffalo, NY, USA; [b]Spatially Explicit Artificial Intelligence (SEAI) Lab, Department of Geography, University of Georgia, Athens, GA, USA; [c]Department of Computer Science, Stanford University, Stanford, CA, USA; [d]Google, Mountain View, CA, USA; [e]Stevenson High School, Lincolnshire, IL, USA; [f]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA

ABSTRACT
Social media messages posted by people during natural disasters often contain important location descriptions, such as the locations of victims. Recent research has shown that many of these location descriptions go beyond simple place names, such as city names and street names, and are difficult to extract using typical named entity recognition (NER) tools. While advanced machine learning models could be trained, they require large labeled training datasets that can be time-consuming and labor-intensive to create. In this work, we propose a method that fuses geo-knowledge of location descriptions and a Generative Pre-trained Transformer (GPT) model, such as ChatGPT and GPT-4. The result is a geo-knowledge-guided GPT model that can accurately extract location descriptions from disaster-related social media messages. Also, only 22 training examples encoding geo-knowledge are used in our method. We conduct experiments to compare this method with nine alternative approaches on a dataset of tweets from Hurricane Harvey. Our method demonstrates an over 40% improvement over typically used NER approaches. The experiment results also show that geo-knowledge is indispensable for guiding the behavior of GPT models. The extracted location descriptions can help disaster responders reach victims more quickly and may even save lives.

ARTICLE HISTORY
Received 20 April 2023
Accepted 28 September 2023

KEYWORDS
Location description; social media; disaster; GPT; GeoAI

1. Introduction

Natural disasters, such as hurricanes, floods, and tornados, pose significant threats to people and society. Between 2017 and 2021 alone, 89 recorded natural disasters in the United States caused over 4500 deaths and more than $780 billion in damages

CONTACT Yingjie Hu yhu42@buffalo.edu

*These authors contributed equally to this research.

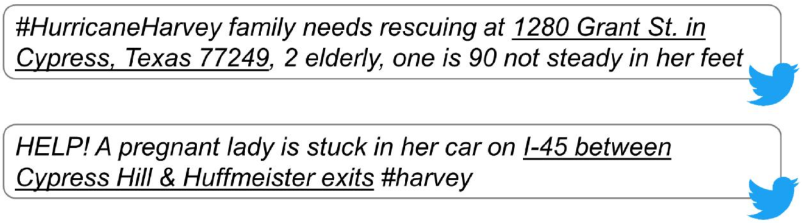
Supplemental data for this article can be accessed online at https://doi.org/10.1080/13658816.2023.2266495.

and losses (NOAA 2022). With climate change, natural disasters are likely to become even more frequent and more costly in the future unfortunately (Elsner *et al.* 2015; Knutson *et al.* 2010). Effective disaster response and management are critical for reducing the loss of life and property.

People have made increasing use of social media platforms, such as Twitter and Facebook, during natural disasters to share urgent information and request help (Devaraj *et al.* 2020; Wang *et al.* 2020a, 2020b; Suwaileh *et al.* 2022; Zhou *et al.* 2022). One prominent example is Hurricane Harvey in 2017. A news article published by the U.S. National Public Radio, titled 'Facebook, Twitter Replace 911 Calls For Stranded In Houston', reported how affected people used social media to request help and how volunteer responders used those requests to locate and reach the people in need (Silverman 2017). Similar stories were also reported by other news media, such as The Wall Street Journal (Seetharaman and Wells 2017) and Time Magazine (Rhodan 2017). While barriers exist in effectively using social media data for disaster response, a recent survey by Hiltz *et al.* (2020) showed that emergency managers considered a software system that can automatically process social media data to be 'very useful' for disaster management.

Social media messages sent out during natural disasters often contain location descriptions that provide critical geographic information such as the locations of victims and accidents. Figure 1 shows two example tweets posted during Hurricane Harvey (with content slightly modified to protect user privacy) that represent potentially life-or-death scenarios. Accurately extracting these location descriptions and geolocating them on maps can help disaster responders reach victims more quickly and potentially save lives. While it is possible to recruit many individuals to manually screen these social media messages by hand, a computational method that can automatically and accurately extract these location descriptions can help save time, person power, and other precious resources during a disaster.

Previous studies have looked into the problem of extracting location descriptions from the content of social media messages (Gelernter and Balaji 2013; Wallgrün *et al.* 2018; Karimzadeh *et al.* 2019; Wang *et al.* 2020a, 2020b; Hu *et al.* 2022; Suwaileh *et al.* 2022). Two technical steps are typically involved: *recognition* and *geo-locating*. The first step recognizes location descriptions from the textual content of social media messages, while the second step aims to find appropriate geographic coordinates and spatial representations for the recognized location descriptions. We focus on the first step in this work since recognizing location descriptions is a prerequisite for the second step of geo-locating.



#HurricaneHarvey family needs rescuing at 1280 Grant St. in Cypress, Texas 77249, 2 elderly, one is 90 not steady in her feet

HELP! A pregnant lady is stuck in her car on I-45 between Cypress Hill & Huffmeister exits #harvey

Figure 1. Two rescue-request tweets posted during Hurricane Harvey.

There are two major limitations in previous studies. First, previous studies generally used a default named entity recognition (NER) approach which cannot recognize location descriptions consisting of multiple entities. NER aims to recognize different types of named entities from text, such as *Persons*, *Organizations*, and *Locations*, and it is reasonable to consider the problem of location description recognition as a subtask of NER by simply focusing on *Locations* only. However, recent research has shown that many location descriptions are not in the form of simple place names (e.g., city names or street names) but consist of multiple entities (Hu and Wang 2021; Fernández-Martínez 2022; Chen *et al.* 2022). Examples of these more complex location descriptions include *door number addresses*, *road intersections*, *highway exits*, and *road segments*. Given that off-the-shelf NER tools are designed to recognize individual entities, they do not have the ability to recognize these multi-entity location descriptions. Figure 2 illustrates this limitation using the two example tweets in Figure 1. Typical NER approaches separately recognize individual entities, such as 'Grant St.', 'Cypress', and 'Texas', rather than the complete location description '1280 Grant St. in Cypress, Texas 77249'. This limitation is critical, because our ultimate goal is to properly geo-locate these location descriptions and help first responders reach the people in need; recognizing one complete location description as separate entities can lead to potentially large errors in the geo-locating step, such as locating this location description to the middle of 'Grant St.' or even to the center of 'Texas'. From a disaster response perspective, these errors can make first responders arrive at the wrong locations and waste rescue time.

A second limitation is that most previous studies have not sought to classify the recognized location descriptions into more detailed categories, such as *door number addresses*, *road intersections*, and *road segments*. In previous studies, the recognized location descriptions are generally all considered as *locations* and are represented as

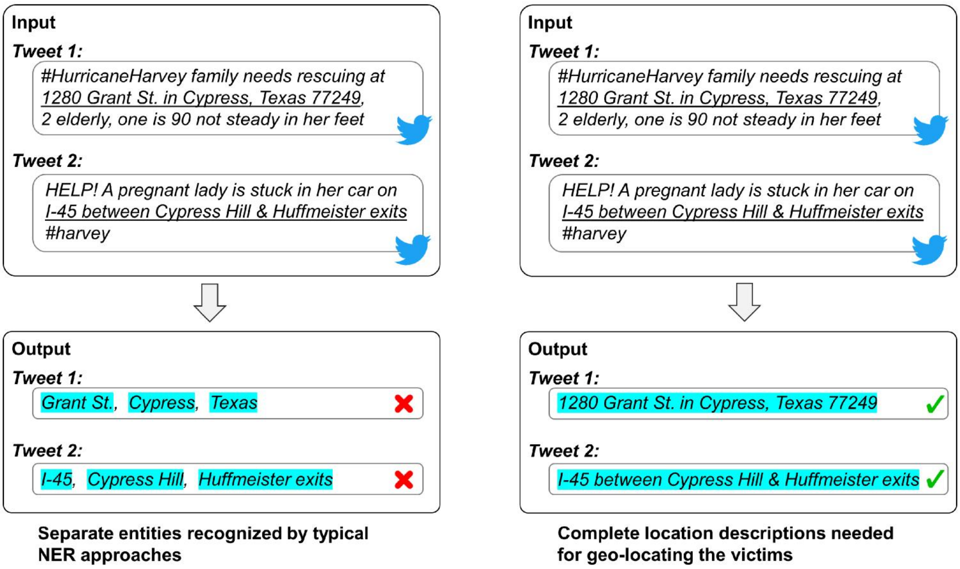

Input
Tweet 1:
#HurricaneHarvey family needs rescuing at 1280 Grant St. in Cypress, Texas 77249, 2 elderly, one is 90 not steady in her feet
Tweet 2:
HELP! A pregnant lady is stuck in her car on I-45 between Cypress Hill & Huffmeister exits #harvey
Output
Tweet 1:
Grant St., Cypress, Texas ✗
Tweet 2:
I-45, Cypress Hill, Huffmeister exits ✗
Separate entities recognized by typical NER approaches
Input
Tweet 1:
#HurricaneHarvey family needs rescuing at 1280 Grant St. in Cypress, Texas 77249, 2 elderly, one is 90 not steady in her feet
Tweet 2:
HELP! A pregnant lady is stuck in her car on I-45 between Cypress Hill & Huffmeister exits #harvey
Output
Tweet 1:
1280 Grant St. in Cypress, Texas 77249 ✓
Tweet 2:
I-45 between Cypress Hill & Huffmeister exits ✓
Complete location descriptions needed for geo-locating the victims

Figure 2. A limitation of typical NER approaches that recognize location descriptions as separate entities rather than complete location descriptions.

points on a map (Gelernter and Balaji 2013; Karimzadeh *et al*. 2019; Wang *et al*. 2020a, 2020b)). Knowing the category of a location description can substantially improve the accuracy of geo-locating, in two ways. First, knowing the categories of location descriptions allows us to choose more suitable geometric representations for them, such as a point for a *road intersection* and a line for a *road segment*. Second, knowing the categories allows us to utilize more suitable geo-locating techniques. For example, we should ideally use the technique of *linear geocoding* if the location description is in the form of *door number addresses* (Goldberg *et al*. 2007), while we may want to use another technique such as *gazetteer matching* if the location description is in the form of *organization names* (e.g., names of schools and churches). Without knowing the categories of location descriptions, we have to adopt a one-size-fit-all strategy for geo-locating, which inevitably sacrifices the geo-locating accuracy. We note that recognizing complete location descriptions and their categories does not mean we can directly geo-locate them using an existing tool; rather, they facilitate the development of suitable techniques for the second step of geolocating to process the recognized location descriptions in each category. In addition, uncertainty still exists in the identified locations even when we have used suitable geo-locating techniques and geometric representations. It is impossible to measure locations perfectly on the surface of the Earth, and the associated uncertainty issues have long been recognized by researchers (J. Zhang and Goodchild 2002; Goodchild and Haining 2003). For disaster response purposes, a location identified from a description could be considered good enough, if responders, upon arriving at this location, can find the described victims or accidents within a reasonable amount of effort.

The two limitations discussed, i.e., (1) not recognizing complete location descriptions, and (2) not identifying location categories, could theoretically both be overcome by training a new machine-learning model on a newly labeled training dataset. This new dataset would require annotating complete location descriptions (rather than separate entities) and annotating the categories of the location descriptions. This training dataset would also need to be sufficiently large so that the trained machine learning model would achieve satisfactory performance. Creating such a dataset, however, can be time-consuming and labor-intensive.

In recent years, large-scale language models (LLMs) have demonstrated impressive performance in the field of artificial intelligence (AI). LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al*. 2019) and Generative Pre-trained Transformer (GPT) models (Brown *et al*. 2020; Ouyang *et al*. 2022), are pre-trained on large-scale textual data (e.g., all text on the Internet) in a task-agnostic manner, and can be adapted to domain-specific tasks via fine-tuning, few-shot learning, or sometimes even zero-shot learning. Given their foundational roles in completing various domain-specific tasks, LLMs and other large-scale pre-trained models are also called *foundation models* (Bommasani *et al*. 2021; Mai *et al*. 2022).

GPT models, such as GPT-3, ChatGPT, and GPT-4, are LLMs and foundation models that have received substantial attention recently (van Dis *et al*. 2023). Taking just a few examples as the instruction (called a *prompt*), a GPT model is able to generate text that reads as if it was written by humans. While they are powerful, the current applications of GPT models are largely limited to conversation and text generation, and their social implications are controversial (Dale 2021; Eliot 2022). In this work, we aim to harness the power of GPT models for social good, i.e., to recognize location descriptions from disaster-related social media messages.

A key consideration for harnessing the power of GPT models is the construction of the *prompt*, which serves as the instruction for the model. Since the prompt can be written in a vast number of different ways, how can we create a prompt that makes a GPT model work more effectively for the task of recognizing location descriptions from social media messages? In this work, we propose a geo-knowledge-guided approach for prompt creation, in which the prompt is created based on the geo-knowledge about common forms of location descriptions. The geo-knowledge was obtained and extended from our previous study which systematically examined location descriptions in tweets posted during Hurricane Harvey and identified a set of common forms of location descriptions (Hu and Wang 2021). In this study, we create the prompt based on such geo-knowledge, and feed the created prompt to GPT through a question-answering process. The result is a fusion of the GPT model and geo-knowledge that can overcome the two discussed limitations: it can recognize full location descriptions and also identify the categories of the recognized descriptions. In addition, only a small number of training examples (22 examples in this study) are needed in our method to guide the GPT model.

The remainder of this paper is organized as follows. Section 2 reviews related work on the use of social media for disaster response and location description extraction. Section 3 presents our method that fuses geo-knowledge and GPT models for extracting location descriptions from disaster-related social media messages. Section 4 presents the experiment design for evaluating our method and comparing it with alternative approaches. Section 5 presents the experiment results, and Section 6 discusses the implications of this study on using AI for disaster response. Finally, Section 7 concludes this work.

2. Related work

There exists a rich amount of literature on leveraging social media, especially Twitter, for supporting disaster response and situational awareness (Crooks *et al.* 2013; De Longueville *et al.* 2009; Feng *et al.* 2022; Imran *et al.* 2015; 2020; MacEachren *et al.* 2011; Murthy and Longwell 2013; Starbird and Stamberger 2010). Social media provides near real-time information about the situation on the ground after a disaster (Hu and Wang 2020), which makes it a valuable alternative information source for emergency managers. However, it has been difficult for emergency managers to use the information from social media, due to issues such as large data volume and data veracity (Hiltz *et al.* 2020; Silverman 2017). Accordingly, much research was devoted to making the information from social media easier to use, including identifying relevant tweets and checking their veracity (Gupta *et al.* 2013; Imran *et al.* 2020; Joseph *et al.* 2014; Vosoughi *et al.* 2018), classifying the purposes of social media posts (Imran *et al.* 2014; Scheele *et al.* 2021; Yu *et al.* 2019), tracking the transition of different disaster phases (Huang and Xiao 2015; Wang *et al.* 2020a, 2020b), and monitoring and understanding public sentiments (Ragini *et al.* 2018; Zou *et al.* 2018). These are all important research areas and studies on extracting location descriptions further complement this research landscape.

While geographic locations are considered highly important by emergency managers (Hiltz *et al.* 2020), previous studies mostly focused on *geotagged* locations, i.e., locations tagged to tweets (De Albuquerque *et al.* 2015; Wang *et al.* 2016; Martín *et al.* 2017), rather than locations described in the content of tweets. Recent years have witnessed widespread adoption of social media during disasters to request help and share information (Mihunov *et al.* 2020; Suwaileh *et al.* 2022; Zhou *et al.* 2022). In these social media messages, people describe locations in the content of tweets and may not necessarily geotag their locations. Besides, people may request help for others (such as in the example tweet '*#HurricaneHarvey family needs rescuing at 1280 Grant St. in Cypress, Texas 77249, 2 elderly, one is 90 not steady in her feet*'), and the current location of the Twitter user may not necessarily be the same as the location of the victim. Accordingly, it is critical to extract locations described in the content of social media messages as well.

Previous studies have looked into the problem of extracting locations from the content of social media messages. By considering locations as a special type of named entities, researchers have employed pre-trained NER tools, such as Stanford NER and SpaCy NER, to extract locations from the content of tweets (Dutt *et al.* 2018; Gelernter and Balaji 2013; Karimzadeh *et al.* 2019). With the fast advancements of deep learning, researchers have developed deep learning-based models for extracting locations. In a previous work, we developed a model called NeuroTPR which improves over a Bidirectional Long Short-Term Memory (BiLSTM) model architecture to extract locations from social media messages (Wang *et al.* 2020a, 2020b). Given the outstanding performance of transformers more recently, especially BERT, researchers have developed newer methods by leveraging pre-trained or fine-tuned transformers and their variants (Hu *et al.* 2022; Suwaileh *et al.* 2022; Berragan *et al.* 2023). While making methodological advancements, previous studies, including our own work, largely considered the problem from a default NER perspective which has two limitations: (1) recognizing individual location entities rather than complete location descriptions; and (2) not classifying location descriptions into more detailed categories. While the NER perspective is good for methodological development, these two limitations constrain our ability to accurately geo-locate victims and accidents during a disaster. It is worth noting that NER tools and models can be retrained using data labeled with complete location descriptions and categories. However, creating such labeled training data requires time, labor, and other resources. There exists another thread of related research on detecting geospatial or, more generally, spatial descriptions from natural language text (Liu *et al.* 2014; Stock *et al.* 2022a, 2022b). While often studying spatial descriptions under more general contexts (e.g., daily life), research in this thread, such as disambiguating geospatial prepositions (Radke *et al.* 2022), can be very useful for the step of geo-locating the recognized location descriptions.

The problem of extracting location descriptions from disaster-related social media messages is also related to but different from the problem of *geoparsing* in the field of geographic information retrieval (GIR) (Freire *et al.* 2011; Gritta *et al.* 2018; Jones and Purves 2008; Melo and Martins 2017; Purves *et al.* 2018). The goal of *geoparsing* is to recognize and resolve toponyms from texts, such as news articles, Web pages, and social media messages. Geoparsing is typically completed in two steps: toponym recognition and toponym resolution. Existing research in geoparsing often focused on the second step, toponym resolution, to address place name ambiguity issues, while utilizing existing NER tools (e.g., Stanford NER) for the first step toponym recognition (Adelfio and Samet 2013; Hu *et al*. 2014; DeLozier *et al*. 2015; Gritta, Pilehvar, and Collier 2018; Karimzadeh *et al*. 2019; Cardoso *et al*. 2021). The focus on the second step is reasonable since general geoparsing research often studies ambiguous toponyms that have worldwide coverage (e.g., 'Paris' can refer to not only 'Paris, France' but also 'Paris, Texas'), while the forms of toponyms are often country names and city names (e.g, 'Paris') that can be recognized by a NER tool with relatively high accuracy (Wang and Hu 2019a, 2019b). In comparison, the problem of extracting location descriptions from disaster-related messages presents a different set of challenges, although it can also be completed in a similar two-step process. For the first recognition step, the challenges are in recognizing many of the complex location descriptions consisting of multiple entities (e.g., door number addresses) that cannot be directly recognized by typical NER tools designed for recognizing single entities. Also, it is necessary to further classify location descriptions into more detailed categories, so that suitable geo-locating techniques and geometric representations can be used. For the second step of geo-locating, place name ambiguity becomes less of a concern because we are focusing on a local region affected by the disaster rather than the entire world. However, new challenges for the second step exist in designing and choosing the most suitable geo-locating techniques (e.g., linear geocoding, gazetteer matching, road intersection identification, and highway exit identification) and geometric representations based on the location descriptions and their categories identified in the first step. While different challenges are involved in the problem of location description extraction, the term 'toponym' can be defined broadly and the two steps could still be considered as toponym recognition and toponym resolution in a broad sense. In this research, we aim to address the challenges in the first step by exploring a new use of GPT models through the guidance of geo-knowledge.

3. Method

Our proposed method fuses geo-knowledge and a GPT model for recognizing complete location descriptions and their categories. This method consists of three components: (1) geo-knowledge about location descriptions, (2) a GPT model, and (3) a process to fuse the two. For (1), we use and extend the geo-knowledge about the common forms of location descriptions obtained from our previous study (Hu and Wang 2021) and encode such knowledge into the prompt. For (2), we employ a pre-trained GPT model. Multiple GPT models, such as GPT-3, ChatGPT, and GPT-4, are tested and studied in this work. For (3), we fuse geo-knowledge and GPT-3 through a question-answering process. Figure 3 provides an overview of our method, and we present methodological details of the three components in the following subsections.

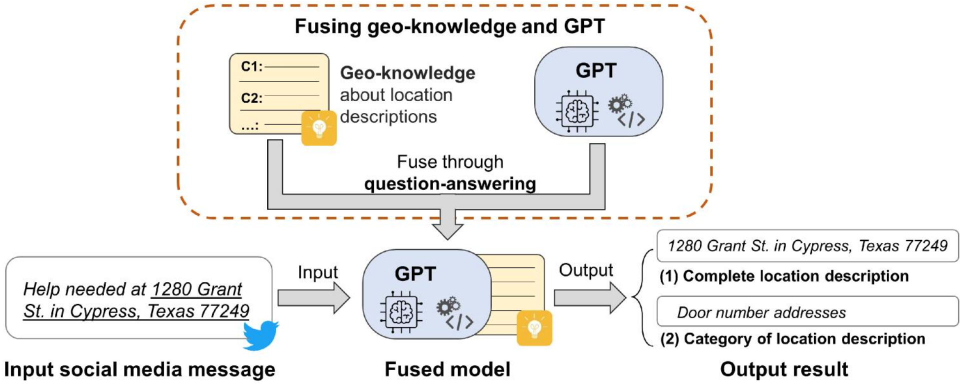

Fusing geo-knowledge and GPT
C1:
C2:
...:
Geo-knowledge about location descriptions
GPT
Fuse through question-answering
Help needed at 1280 Grant St. in Cypress, Texas 77249
Input
GPT
Output
1280 Grant St. in Cypress, Texas 77249
(1) Complete location description
Door number addresses
(2) Category of location description
Input social media message
Fused model
Output result

Figure 3. An overview of fusing geo-knowledge and GPT for recognizing location descriptions and their categories from social media messages.

3.1. Geo-knowledge about location descriptions

We leverage and extend the geo-knowledge about common forms of location descriptions obtained from our previous study (Hu and Wang 2021). In that work, we assembled a Twitter dataset by randomly selecting 1000 tweets from a set of 15,834 tweets that are likely to contain location descriptions; the set of 15,834 tweets was selected from a dataset of over 7 million tweets posted during Hurricane Harvey using a regular expression containing location-related terms (e.g., 'street', 'avenue', 'park', 'square', 'bridge', 'rd', and 'ave'). We then manually examined each individual tweet and annotated their location descriptions. We also classified these location descriptions into ten categories. Table 1 shows these ten categories (categories *C1* to *C10*) and two example tweets are provided for each category. In this current study, we further extend these categories of location descriptions by adding one more category (category *C11* in Table 1) and will explain why we think such a new category is necessary. We note that a tweet example in Table 1 may contain multiple location descriptions, and we only underscore the location descriptions in a corresponding category in Table 1 for the purpose of clear presentation. All location descriptions in the tweet are annotated in the data.

The new category that we added is *C11: Road segments* and an example description in this category is '*HELP! A pregnant lady is stuck in her car on I-45 between Cypress Hill & Huffmeister exits #harvey*'. Location descriptions in *C11* were initially put under either category *C4: Exits of highways* or *C5: Intersections of roads (rivers)* in our previous study. While the original classifications are also reasonable (since these location descriptions often involve highway exits and road intersections), we believe that this type of location description may be better represented geometrically as a line in the geo-locating step since they refer to a segment of a road. By contrast, the location descriptions in C4 and C5 are better to be represented as points, since they refer to road intersections or highway exits. With this consideration, we added category C11. The geo-knowledge in this study is therefore the 11 categories of location descriptions and their typical forms represented via examples.

While Table 1 provides one approach for categorizing location descriptions, there exist other approaches and schemes in linguistics and GIScience for organizing

Table 1. Eleven categories of location descriptions commonly used in Hurricane Harvey extended from (Y. Hu and Wang 2021).

Category	*Location description examples*
C1: Door number addresses	• *'#HurricaneHarvey family needs rescuing at 1280 Grant St. in Cypress, Texas 77249, 2 elderly, one is 90 not steady in her feet'* • *'Papa stranded in home. Water rising above waist. HELP 812 Wood Ln, 77828 #houstonflood'*
C2: Street names	• *'#Harvey LIVE from San Antonio, TX. Fatal car accident at Ingram Rd., Strong winds.'* • *'Allen Parkway, Memorial, Waugh overpass, Spotts park and Buffalo Bayou park completely under water'*
C3: Highways	• *'9:00AM update video from Hogan St over White Oak Bayou, I-10, I-45: water down about 4′ since last night. . . ′′* • *'Left Corpus bout to be in San Angelo #HurricaneHarvey Y'all be safe Avoided highway 37 Took the back road'*
C4: Exits of highways	• *'Need trailers/trucks to move dogs from Park Location: Whites Park Pavillion off I-10 exit 61 Anahuac TX'* • *'Townsend exit, Sorters road and Hamblen road is flooded coming from 59 southbound #HurricaneHarvery #Harvey2017'*
C5: Intersections of roads (rivers)	• *'Guys, this is I-45 @ Main Street in Houston. Crazy. #hurricane #harvey. . .'* • *'Major flooding at Clay Rd & Queenston in west Houston. Lots of rescues going on for ppl trapped… '*
C6: Natural features	• *'Frontage Rd at the river #hurricaneHarvey #hurricaneharvey @ San Jacinto River'* • *'Buffalo Bayou holding steady at 1′00 cfs at the gage near Terry Hershey Park'*
C7: Other human-made features	• *'If you need a place to escape #HurricaneHarvey, The Willie De Leon Civic Center: 300 E. Main St in Uvalde is open as a shelter'* • *'Houston's Buffalo Bayou Park - always among the first to flood. #Harvey'*
C8: Local organizations	• *'Cleaning supply drive is underway. 9-11 am today at Preston Hollow Presbyterian Church'* • *'#Harvey does anyone know about the flooding conditions around Cypress Ridge High School?! #HurricaneHarvey'*
C9: Administrative units	• *'#HurricaneHarvey INTENSE eye wall of category 4 Hurricane Harvey from Rockport, TX'* • *'Pictures of downed trees and damaged apartment building on Airline Road in Corpus Christi'*
C10: Multiple areas	• *'Anyone doing high water rescues in the Pasadena/Deer Park area? My daughter has been stranded in a parking lot all night'* • *'FYI to any of you in NW Houston/Lakewood Forest, Projections are showing Cypress Creek overflowing at Grant Rd'*
C11: Road segments	• *'HELP! A pregnant lady is stuck in her car on I-45 between Cypress Hill & Huffmeister exits #harvey'* • *'Streets Flooded: Almeda Genoa Rd. from Windmill Lakes Blvd. to Rowlett Rd. #HurricaneHarvey #Houston'*

locations and more generally spatial information in texts. From a linguistic perspective, Gritta *et al.* (2020) differentiated *literal toponyms* (e.g., 'Paris') and *associative toponyms* (e.g., 'Spanish sausages'); Mani *et al.* (2010) examined both *absolute spatial references* (e.g., 'Rome') and *relative spatial references* (e.g., 'to the left of the room'); and Pustejovsky *et al.* (2012) considered the motion of individuals that may be involved in spatial references (e.g, 'John biked to Agua Azul'). From a GIScience perspective, one way to classify locations and geographic features is based on their geometric representations, such as points, lines, and polygons (Hill 2000; Longley *et al.* 2005). The categorization of location descriptions in Table 1 shares some similarities with existing approaches, such as its coverage of absolute spatial references and the use of different geometric representations. Meanwhile, it also adds considerations from a disaster response perspective. First, the categorization in Table 1 is oriented toward location descriptions used under a disaster context. Based on our empirical analysis of tweets from Hurricane Harvey, it seems that people tend to use absolute references (e.g., door number addresses and road intersections) rather than relative or vague references (e.g., to the left of that area) in rescue-request tweets to be found by first responders. Second, our categorization considers not only geometric representations but also the different geo-locating techniques necessary to locate these descriptions. For example, door number addresses, road intersections, and highway exits may all be represented as points; however, they require largely different techniques for identifying their locations. Third, our categorization also follows some practices used in previous geoparsing and toponym recognition research in which a NER tool typically differentiates organizations (e.g., schools and churches) from other types of locations and differentiates natural features from human-made features (Gelernter and Mushegian 2011; Karimzadeh *et al*. 2019).

Nevertheless, the categorization of location descriptions in Table 1 has its limitations. There could be less common location descriptions that are not covered in the current set of categories, which might be uncovered as we analyze more data. In addition, the current categorization is only one version of the geo-knowledge capturing our current understanding of disaster-related location descriptions. As further research is conducted in this area, we will improve our understanding of location descriptions and may refine or even revise the current categorization. One advantage of the proposed method is its flexibility in adjusting the geo-knowledge used to guide the model. When less common location descriptions are identified or when the categorization is improved, we can adjust the geo-knowledge accordingly by changing the examples while still using the same methodological framework.

3.2. GPT models

GPT models, such as ChatGPT and GPT 4, have attracted a lot of attention recently from the public media (Eliot 2022; van Dis *et al*. 2023). GPT models are pre-trained on large-scale textual data and can generate answers when given a *prompt,* an instruction to the model. Despite their impressive performance in generating human-like text, GPT models have triggered various societal concerns, e.g., they make it easier for students to cheat in homework assignments and may replace certain human jobs that typically require intellectual creativity such as fiction writers (Dale 2021; Kasneci *et al.* 2023; Mhlanga 2023). In this study, we aim to harness the power of GPT for supporting disaster response by recognizing location descriptions from social media messages.

The behavior of a GPT model is influenced by the prompt it receives. We hypothesize that the best prompt should be created based on systematic knowledge about the target problem, and in this case, it is geo-knowledge about the common forms of location descriptions used by people during natural disasters. To test our hypothesis, we experiment with four different versions of GPT models, which are GPT-2, GPT-3, ChatGPT, and GPT-4. These are all transformer-based generative models, and detailed model architecture information is provided in the papers and blog articles by researchers from OpenAI, the company that developed these GPT models (Brown *et al*. 2020;

OpenAI 2022; 2023; Ouyang *et al*. 2022; Radford *et al*. 2018; 2019). For newer models such as GPT-3, ChatGPT, and GPT-4, OpenAI provides Application Programming Interfaces (APIs) for directly accessing the pre-trained models. For older models such as GPT-2 and GPT, OpenAI does not provide an API but pre-trained models are available from open-source libraries such as the Transformers library from the AI company Hugging Face. We note that while ChatGPT and GPT-4 are very recent models as of April 2023, we started this research back in early 2022 based on GPT-2 and GPT-3; however, given the recent public attention to ChatGPT and GPT-4, we further include them in our experiments. By adding the two latest GPT models, we also demonstrate that our proposed method for fusing geo-knowledge and GPT models is generalizable to newer GPT models which will likely come in the following years.

3.3. Fusing geo-knowledge and GPT

We fuse geo-knowledge and GPT by encoding the geo-knowledge of location descriptions into a prompt and feeding it to a GPT model to guide its behavior. Learning from previous research on geographic question-answering (Mai *et al*. 2021), we create the prompt in the form of a series of question-answering statements based on the geo-knowledge of location description categories and their examples shown in Table 1. A snippet of the prompt is provided in Table 2, and we also include the full prompt in the supplementary Table S1 which contains 22 tweet examples in 11 categories (i.e., two examples per category). The prompt first describes the task of location description recognition and the expected output. Then, a series of question-answering examples are provided. Each example is organized as 'Sentence', 'Q', and 'A'. The 'Sentence' provides a tweet example, 'Q' provides a question about the task, and 'A' provides the ideal answer that we expect, i.e., complete location descriptions and their categories. In each question-answering example, we 'teach' GPT to first predict the category of a location description (e.g., 'C1') and then output the full text of the recognized location description. The category and location description are separated by ':'. When multiple location descriptions exist in one tweet, we ask the model to separate them by ';'. At the end of the prompt, we add a new tweet whose location descriptions are unknown, and we ask the GPT model to infer its location descriptions and their categories based on the previous examples. The final output of GPT is then recorded.

4. Evaluation experiments

In this section, we describe the evaluation experiments designed to assess the performance of the proposed method. We compare different implementations of our method using different GPT models. We also compare our method with nine alternative approaches, including typically used NER approaches, the traditional fine-tuning approach using BERT, and the default GPT models without the guidance of geo-knowledge. In the following, we present details about the experiment setting.

Table 2. A snippet of the prompt with question-answering statements created based on the geo-knowledge of the 11 categories of location descriptions and two examples per category. The full prompt is provided in Supplementary Table S1.

This is a set of location description recognition problems.
The 'Sentence' is a sentence containing location descriptions.
The goal is to infer which parts of the sentence represent location descriptions and the categories of the location descriptions. Split different location descriptions with ';'.
–
–
Sentence: Papa stranded in home. Water rising above waist. HELP 812 Wood Ln, 77828 #houstonflood
Q: Which parts of this sentence represent location descriptions?
A: C1: 812 Wood Ln, 77828
–
–
Sentence: Anyone doing high water rescues in the Pasadena/Deer Park area? My daughter has been stranded in a parking lot all night
Q: Which parts of this sentence represent location descriptions?
A: C10: Pasadena/Deer Park
–
–
Sentence: Allen Parkway, Memorial, Waugh overpass, Spotts park and Buffalo Bayou park completely under water
Q: Which parts of this sentence represent location descriptions?
A: C2: Allen Parkway; C2: Memorial; C2: Waugh overpass; C7: Spotts park; C7: Buffalo Bayou park
–
…
–
Sentence: {TEXT}
Q: Which parts of this sentence represent location descriptions?
A:

4.1. Experiment dataset

The dataset used for experiments is the 1000 annotated Hurricane Harvey tweets initially created in our previous work (Hu and Wang 2021) and further extended in this current work. To the best of our knowledge, this is the only dataset available that was annotated with both complete location descriptions and location categories. Since 22 tweets from this dataset have already been used to create the prompt (i.e., those tweet examples in Table 1), we exclude them from testing and use only the remaining 978 tweets to assess the performance of the models. An example of the annotated tweets is shown in Figure 4. The tweets were annotated using the Inside–Outside–Beginning (IOB) tagging scheme (Tjong Kim Sang and De Meulder 2003). The location description category (e.g., 'C1' in the figure) was appended to the end of the corresponding 'B' and 'I' tags. The 'O' tag is not explicitly labeled, and all tokens without the 'B' or 'I' tag are considered as having the 'O' tag. The complete dataset is provided in the repository at the end of the article.

4.2. Experiment models

We study our method by experimenting with the following GPT models:

- *Fusing geo-knowledge and GPT-4 (Geo-GPT-4)*: In this implementation, we fuse geo-knowledge and the GPT-4 model using the prompt created based on 22 tweet examples (see Table 2). These tweet examples inform GPT-4 of the forms of

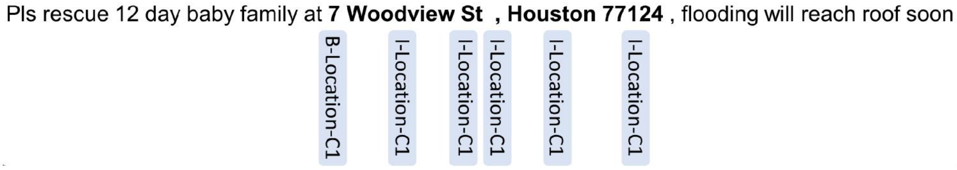

Pls rescue 12 day baby family at 7 Woodview St , Houston 77124 , flooding will reach roof soon
B-Location-C1
I-Location-C1
I-Location-C1
I-Location-C1
I-Location-C1
I-Location-C1

Figure 4. An example of the annotated tweets with location descriptions and their categories labeled using the IOB scheme.

location descriptions and the 11 categories of location descriptions with two examples per category.

- *Fusing geo-knowledge and ChatGPT (Geo-ChatGPT)*: This implementation uses the same prompt based on 22 tweet examples as used for *Geo-GPT-4*. Instead of GPT-4, we use ChatGPT for this implementation.
- *Fusing geo-knowledge and GPT-3 (Geo-GPT-3)*: This implementation uses the same prompt but uses the GPT-3 model.
- *Fusing geo-knowledge and GPT-2 (Geo-GPT-2)*: This implementation uses an older GPT model, GPT-2. Because the prompt of this older GPT model cannot be longer than 1024 tokens, we use 11 examples to create the prompt for GPT-2, with one tweet example for each of the 11 categories.

The fusion of geo-knowledge and GPT-3, ChatGPT, and GPT-4 is done through our created prompts and the pre-trained models via the APIs provided by OpenAI. The fusion of geo-knowledge and GPT-2 is done through our created prompt and the pre-trained model from the open-source library from Hugging Face. All source code, such as loading pre-trained models and adapting them through prompts encoding geo-knowledge, is shared in the repository in the Data Availability Statement at the end of the article. In addition to the above four models, we also include nine alternative approaches to serve as baselines:

- *Default GPT-4, ChatGPT, GPT-3, and GPT-2 models:* These default GPT models are included as four baselines to examine the role of geo-knowledge in guiding the behavior of the models. The default GPT models are powerful and were pre-trained with vast amounts of data that could include the same or similar tweets as used in this study. Thus, including these default models can help us understand whether the strong performance of our approach (if any) comes directly from the default GPT models or from the geo-knowledge-guided process. In the prompt to these default models, we ask the same question: '*Which parts of this sentence represent location descriptions?*' but do not provide any further geo-knowledge. It is worth noting that the default GPT models do not have the ability to identify location categories, as these categories are part of the geo-knowledge not given to the default models. However, these default models can detect location descriptions based on the large amount of text they have seen during the pre-training process.
- *Fine-tuned BERT (Fine-tuned-BERT)*: Fine-tuning a BERT model is an approach that has been used in recent studies for analyzing disaster-related tweets and has demonstrated good performance (Suwaileh *et al.* 2022; Zhou *et al.* 2022). In this

baseline, we fine-tune a BERT model using the same 22 tweet examples used to create the prompt for the GPT models. The pre-trained BERT model from Hugging Face is used for implementing this approach.

- *Stanford NER (narrow):* The Stanford NER tool has been commonly used in the literature for recognizing place names from text (Gelernter and Balaji 2013; Karimzadeh *et al*. 2019; Liu *et al*. 2014). Here, we use the off-the-shelf Stanford NER tool (Manning *et al*. 2014) which can recognize *Person*, *Organization*, and *Location* entities from text. In this *narrow* version of Stanford NER, we keep only *Locations* in the output.
- *Stanford NER (broad):* We use the same off-the-shelf Stanford NER tool but keep both *Locations* and *Organizations* in the output for this *broad* version. The *Locations* output by Stanford NER do not include schools and churches which are considered as *Organizations* by the tool; yet, schools and churches are often used as shelters and their locations are often described during disasters Wang *et al*. 2020a, 2020b). Including *Organizations* in the output therefore can help capture those schools and churches. However, this *broad* approach may also include false positives, i.e., organizations that are not used as locations.
- *SpaCy NER (narrow):* SpaCy NER is another tool often used in the literature for recognizing place names from text (Fernandes *et al*. 2021; Gritta, Pilehvar, and Collier 2018; Hu *et al*. 2019). For this *narrow* version, we use the off-the-shelf SpaCy NER tool and keep only *GPE* (Geopolitical Entities) in the output, which contains cities, counties, states, and countries.
- *SpaCy NER (broad):* We use the same off-the-shelf SpaCy NER tool here but keep both *GPE* (geopolitical entities) and *ORG* (organizations) in the output for this *broad* version. We keep organizations in the output based on the same rationale for the broad version of Stanford NER.

4.3. Evaluation metrics

We adopt three evaluation metrics, i.e., *precision*, *recall*, and *F-score*, to assess the performance of the experiment models. These three metrics have been widely used for measuring the performance of computational models in recognizing locations from texts (Gritta *et al*. 2018; Purves *et al*. 2018; Wang and Hu 2019a, 2019b). They are calculated using Equations (1–3):

$$Precision = \frac{|Correctly\ recognized|}{|All\ recognized|} \tag{1}$$

$$Recall = \frac{|Correctly\ recognized|}{|All\ correct|} \tag{2}$$

$$F-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Precision measures the percentage of correctly recognized location descriptions among all location descriptions recognized by a model. *Recall* measures the percentage of correctly recognized location descriptions among all annotated location descriptions. *F-score* is the harmonic mean of *precision* and *recall*, and *F-score* will be high if both *precision* and *recall* are high and *F-score* will be low if one of the two is low. It is worth noting that the *correctly recognized* location description in Equations (1–2) is measured based on a full-span matching between the model-recognized description and the human-annotated location description. If a recognized location description only partially matches the annotated description, it is considered incorrect. We believe that using full-span matching, rather than allowing partial matching, is important for this research oriented toward disaster response because a model that recognizes only a part of a location description can lead to limited geo-locating accuracy and even large errors (e.g., erroneously geo-locating a road intersection 'Road A & Road B' to the center of 'Road A', if only 'Road A' is recognized). All experiment models are assessed on the same dataset of 978 tweets (i.e., the 1000 tweets minus the 22 tweets used for creating the prompt) using full-span matching.

5. Results

Two sets of experiments have been conducted to evaluate the performance of the experiment models. In the first set of experiments, we focus on evaluating the recognized location descriptions only, since some of the experiment models (e.g., the default GPT models and the NER models) do not have the ability to recognize the categories of location descriptions. In the second set of experiments, we focus on those models that can recognize both location descriptions and their categories, and evaluate their ability to extract these two types of important information. In the following, we report the results from the two sets of experiments.

5.1. Ability to recognize complete location descriptions regardless of categories

The first set of experiments focuses on only the text of the recognized location descriptions and does not consider the categories. The results are summarized in Table 3.

We first look at the performance of the NER models in Table 3. The four NER models achieve precisions between 0.352 and 0.643, recalls between 0.224 and 0.440, and F-scores between 0.332 and 0.495. While these results are not disappointing, a closer examination shows that the recognized location descriptions are all in the form of simple place names, such as city names, state names, and river names. In fact, they completely miss those more complex location descriptions, such as door number addresses, road segments, and road intersections, which consist of multiple entities. This result is unsurprising, as we know that these NER models are designed to recognize individual named entities and cannot recognize location descriptions consisting of multiple entities. Meanwhile, these multi-entity location descriptions provide highly detailed location information that can help rescue teams to reach victims. Models that fail to recognize these detailed location descriptions can provide only limited support for disaster response efforts.

Next, we look at the performance of the default GPT models and the Geo-GPT models. As shown in Table 3, there is an increase in performance from GPT-2 to GPT-4, demonstrating that these GPT models are indeed becoming more intelligent over

Table 3. Precision, recall, and F-score of the experiment models in recognizing complete location descriptions. Metrics are measured based on full-span matching.

Models	Precision	Recall	F-score
Models that cannot identify location categories			
Stanford NER (narrow)	0.621	0.402	0.488
Stanford NER (broad)	0.564	0.440	0.495
SpaCy NER (narrow)	0.643	0.224	0.332
SpaCy NER (broad)	0.352	0.340	0.346
GPT-2	0.012	0.004	0.005
GPT-3	0.255	0.256	0.255
ChatGPT	0.416	0.370	0.392
GPT-4	0.404	0.385	0.394
Models that can identify location categories			
Geo-GPT-2	0.380	0.404	0.391
Geo-GPT-3	**0.693**	0.694	0.693
Geo-ChatGPT	0.633	0.673	0.653
Geo-GPT-4	0.687	**0.704**	**0.695**
Fine-tuned-BERT	0.150	0.242	0.185

the years. While GPT-2 mostly fails in the experiment, GPT-3, ChatGPT, and GPT-4 are able to recognize about 25%-40% of the location descriptions, including some more complex multi-entity location descriptions, such as door number addresses and road intersections. However, the performance of these default GPT models is still quite limited, with the highest F-score 0.394 achieved by GPT-4. These default GPT models also do not have the ability to identify the categories of location descriptions. In comparison, the Geo-GPT models substantially improve over their corresponding default GPT models in all metrics. This improvement can be seen across all four GPT models from GPT-2 to GPT-4, demonstrating that the effectiveness of geo-knowledge in improving performance is not limited to one particular model but applicable to all GPT models. Geo-knowledge plays an important role by informing the GPT models about the common forms of location descriptions and the categories of these descriptions. Without geo-knowledge, the default GPT models can rely on only the information of location descriptions that they obtained during the pre-training process, which likely leads to lower performance. While geo-knowledge is important, the advancement in GPT models is also necessary for achieving the obtained good results. As can be seen in Table 3, the model performance also increases from Geo-GPT-2 to Geo-GPT-4 in general. Interestingly, Geo-ChatGPT shows a slightly lower performance than that of Geo-GPT-3, even though ChatGPT is a newer model than GPT-3. Finally, the fine-tuned BERT model achieves a precision 0.150, a recall 0.242, and an F-score 0.185. These low performance scores are likely due to the small number of training examples not sufficient for adapting a complex model like BERT.

The highest F-score of 0.695 is achieved by Geo-GPT-4, and a similar score of 0.693 is achieved by Geo-GPT-3. These F-scores are close to the threshold of 0.70 considered as acceptable in location recognition tasks for disaster response (Suwaileh *et al*. 2022). While these scores are probably not outstanding, they show an over 40% improvement compared with off-the-shelf NER models, such as the Stanford NER used in the experiments. NER models can be re-trained to achieve a similar performance using a sufficiently large dataset labeled with complete location descriptions and categories; however, creating such a dataset requires time, labor, and other precious resources. The geo-knowledge-guided GPT models therefore present an efficient approach for extracting location descriptions when we have only a small number of training data examples encoding geo-knowledge. To further examine the sensitivity of Geo-GPT-4 to different examples included in the prompt, we test two additional prompts, one with a different set of 22 tweets and the other with a set of 22 tweets synthesized based on the geo-knowledge. The performance of Geo-GPT-4 based on these two prompts slightly increases (an increase of 0.023 and 0.008 in terms of F-score), but does not substantially change. These two prompts and the test results are included in Supplementary Tables S2-S4.

5.2. Ability to recognize both complete location descriptions and location categories

In the second set of experiments, we evaluate the ability of the models to recognize both complete location descriptions and location categories. We focus on the models that have the ability to recognize location categories in addition to the description text, which are Geo-GPT-2, Geo-GPT-3, Geo-ChatGPT, Geo-GPT-4, and fine-tuned BERT. A recognized location description is considered correct only when it has both the complete location description and the correct location category. We compute the precision, recall, and F-score of the five models for each of the 11 categories and their overall scores across all categories, and the results are summarized in Figure 5.

As shown in the figure, Geo-GPT-3, Geo-ChatGPT, and Geo-GPT-4 achieve higher performance scores than the other two models in each of the eleven location description categories and also across all categories. The fine-tuned BERT model fails to identify the correct category of a location description most of the time, even in cases when it recognizes the full location description. This result again is likely due to the fact that the small number of training examples are not sufficient to adapt the BERT model, especially for this relatively complex problem with eleven categories to be identified. While Geo-GPT-2 improves over the fine-tuned BERT model, its performance is still quite limited and its precisions, recalls, and F-scores in most location categories are lower than 0.2. Substantial performance improvements are observed in Geo-GPT-3, Geo-ChatGPT, and Geo-GPT-4. Consistent with the first set of experiments, Geo-GPT-3 and Geo-GPT-4 perform better than Geo-ChatGPT in most categories and overall, although Geo-ChatGPT does perform better than Geo-GPT-3 in some categories, such as *C2: Street names* and *C6: Natural features*. It is probably less surprising that Geo-GPT-4 shows a better performance than Geo-ChatGPT, given that GPT-4 is a newer model than ChatGPT and OpenAI has shown that GPT-4 outperformed ChatGPT in multiple tests, such as bar exams (OpenAI 2023). It does surprise us that Geo-GPT-3 performs better than Geo-ChatGPT. According to OpenAI, ChatGPT has been further fine-tuned using a new dialogue dataset to optimize its performance toward chat-related tasks (OpenAI 2022). Such chat-related optimization might have affected the performance of ChatGPT on this current location description recognition task. Both Geo-GPT-3 and Geo-GPT-4 achieve good performance in category *C1: Door number addresses,* with precisions, recalls, and F-scores all about 0.75. Also, both models extract whole door number addresses including the prepositions (e.g., 'in') sometimes used by people but not typically part of formal addresses. These extracted whole

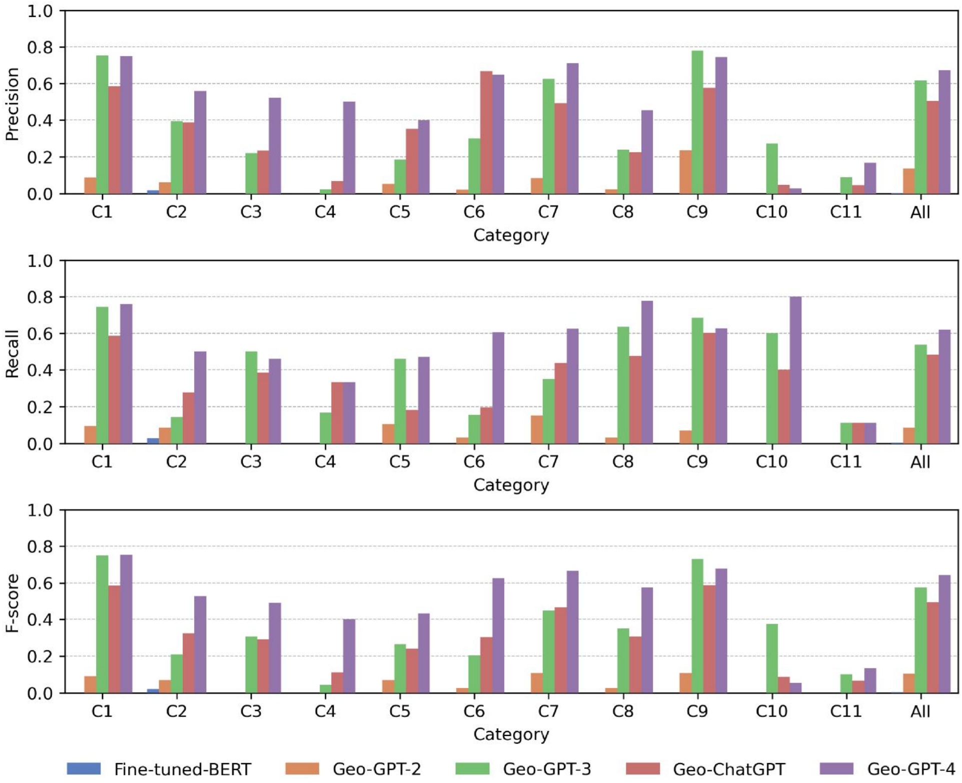
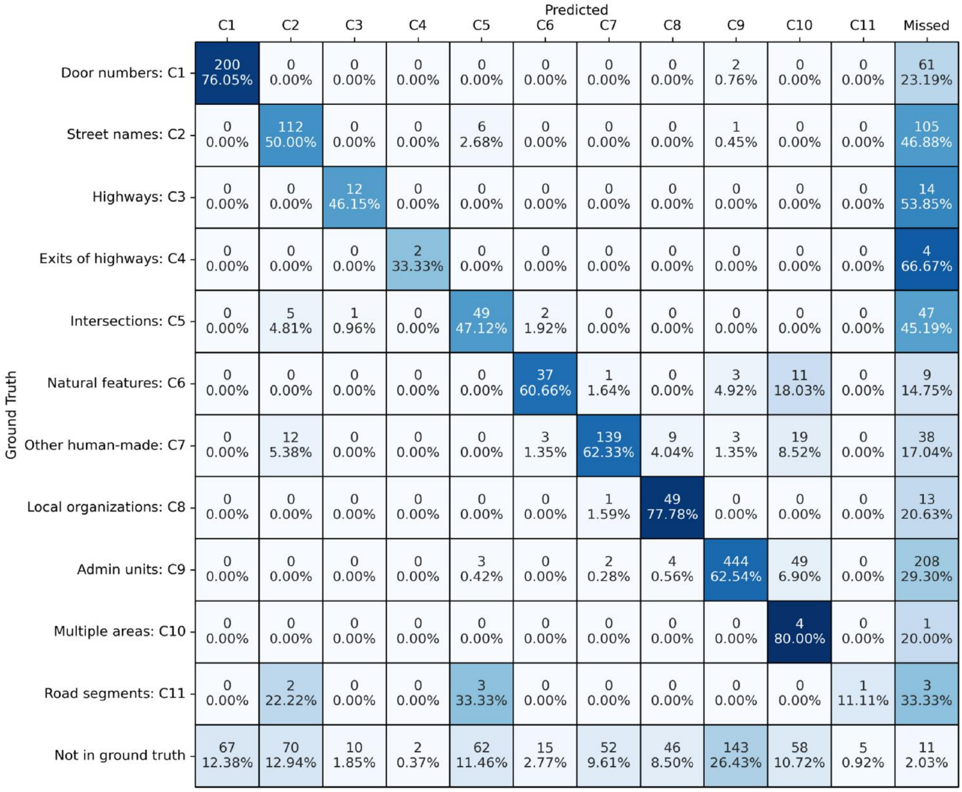


Figure 5. Precision, recall, and F-score of the tested models in recognizing both complete location descriptions and their categories.

addresses allow us to analyze and geo-locate them in the next step by building on previous research, such as research on geospatial preposition analysis (Radke *et al.* 2022). Given that door number addresses provide precise location information, correctly recognizing location descriptions in this category can be highly helpful in identifying the locations of victims. Geo-GPT-4 demonstrates better performance than Geo-GPT-3 over the majority of the categories and also across all categories, with an overall F-score of 0.644 achieved by Geo-GPT-4 and an overall F-score of 0.573 achieved by Geo-GPT-3. This result helps us further understand the performance difference between Geo-GPT-3 and Geo-GPT-4: while the two show similar performance in the first set of experiments focusing on location descriptions only, Geo-GPT-4 demonstrates a substantially better performance in correctly identifying both location descriptions and their categories.

To understand the errors made by Geo-GPT-4, we create a confusion matrix in Figure 6. In this figure, each row represents location descriptions in a category in the test data (i.e., ground truth), and each column represents location descriptions recognized by Geo-GPT-4 in that category (i.e., model predictions). The integer number in a cell represents the number of location descriptions correctly recognized by the model. For example, the number '200' at row C1 and column C1 indicates that 200 of the door number addresses in the ground truth were correctly recognized by Geo-GPT-4; meanwhile, the number '2' at row C1 and column C9 indicates that one door number



Ground Truth \ Predicted	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	Missed
Door numbers: C1	200 76.05%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	2 0.76%	0 0.00%	0 0.00%	61 23.19%
Street names: C2	0 0.00%	112 50.00%	0 0.00%	0 0.00%	6 2.68%	0 0.00%	0 0.00%	0 0.00%	1 0.45%	0 0.00%	0 0.00%	105 46.88%
Highways: C3	0 0.00%	0 0.00%	12 46.15%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	14 53.85%
Exits of highways: C4	0 0.00%	0 0.00%	0 0.00%	2 33.33%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	4 66.67%
Intersections: C5	0 0.00%	5 4.81%	1 0.96%	0 0.00%	49 47.12%	2 1.92%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	47 45.19%
Natural features: C6	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	37 60.66%	1 1.64%	0 0.00%	3 4.92%	11 18.03%	0 0.00%	9 14.75%
Other human-made: C7	0 0.00%	12 5.38%	0 0.00%	0 0.00%	0 0.00%	3 1.35%	139 62.33%	9 4.04%	3 1.35%	19 8.52%	0 0.00%	38 17.04%
Local organizations: C8	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	1 1.59%	49 77.78%	0 0.00%	0 0.00%	0 0.00%	13 20.63%
Admin units: C9	0 0.00%	0 0.00%	0 0.00%	0 0.00%	3 0.42%	0 0.00%	2 0.28%	4 0.56%	444 62.54%	49 6.90%	0 0.00%	208 29.30%
Multiple areas: C10	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	4 80.00%	0 0.00%	1 20.00%
Road segments: C11	0 0.00%	2 22.22%	0 0.00%	0 0.00%	3 33.33%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	1 11.11%	3 33.33%
Not in ground truth	67 12.38%	70 12.94%	10 1.85%	2 0.37%	62 11.46%	15 2.77%	52 9.61%	46 8.50%	143 26.43%	58 10.72%	5 0.92%	11 2.03%

Figure 6. Confusion matrix based on the output of Geo-GPT-4 and the annotated location descriptions.

address was correctly recognized by Geo-GPT-4 but was mistakenly classified as category C9. The percentage value in each cell is calculated based on the row sum. For example, the value '76.05%' at row C1 and column C1 indicates that the 200 corrected recognized location descriptions represent 76.05% of all labeled location descriptions in category C1. There are also situations in which labeled location descriptions are not correctly recognized or completely missed by the model, which are represented in the 'Missed' column; there also exist recognized location descriptions that are not in the ground truth (i.e., false positives), which are represented in the 'Not in ground truth' row.

The confusion matrix helps us further understand the performance of Geo-GPT-4 across different categories of location descriptions. It is worth noting that some categories of location descriptions, such as C1, C2, C7 and C9, show up more frequently than some others in the data. However, we cannot focus on only those more frequent location descriptions, for two reasons. First, some categories of location descriptions that show up more frequently are probably less informative from the perspective of disaster response. One example is *C9: Administrative units*: while city and state names, such as 'Houston' and 'Texas', show up frequently in the data, they are less useful for first responders to locate and reach the victims. Second, location descriptions that show up less frequently are still used in multiple tweets describing life-threatening situations. Recognizing these relatively less frequent location descriptions can be critical for reaching the people in need. Overall, Geo-GPT-4 is effective in correctly recognizing both complete location descriptions and the categories of the descriptions, as can be seen in the diagonal of the confusion matrix. Across different categories, Geo-GPT-4 makes a small number of errors in which a correctly recognized location description is mistakenly classified into another category. Geo-GPT-4 does make more mistakes in category *C11: Road segments* by misclassifying them to *C2: Street names* and *C5: Intersections*. A description of a road segment, such as '*Almeda Genoa Rd. from Windmill Lakes Blvd. to Rowlett Rd.*', does often involve street names and road intersections, which could be a reason for the confusion made by Geo-GPT-4.

The majority of the errors fall into either the column of 'Missed' or the row of 'Not in ground truth'. By examining these errors based on the output of Geo-GPT-4, we find that quite some location descriptions are in fact reasonably recognized by the model; however, they do not completely match the annotated location description given our requirement of strict full-span matching in the evaluation experiments. For example, in the tweet '*#Houston #HoustonFlood This is the intersection of I-45 & N. Main Street*', the recognized location description of Geo-GPT-4 is '*the intersection of I-45 & N. Main Street*', while the annotated description is '*I-45 & N. Main Street*'. We note that a clean string of '*I-45 & N. Main Street*' is easier for the next step of geo-locating, but the recognized description '*the intersection of I-45 & N. Main Street*' may not be considered completely wrong either. However, in this example, the ground truth annotation '*I-45 & N. Main Street*' is counted as 'Missed', while the recognized description '*the intersection of I-45 & N. Main Street*' is counted as 'Not in ground truth'. There exist other similar cases, e.g., the output of the model is '*Sugarland area*' while the annotation is '*Sugarland*'. Sometimes, there is a larger difference between the recognized location description and the annotated description. For example, in the tweet '*Spokeswoman for Houston Mayor Sylvester Turner says the convention center at NRG Park is opening, serving 10,000 additional Harvey evacuee*', the annotated location description is '*NRG Park*' while the description recognized by the model is '*convention center at NRG Park*'. Again, the model output may not be considered completely wrong, although it is different from the annotated location description. These and other similar cases have contributed to the relatively larger errors observed in the 'Missed' column and the 'Not in ground truth' row in the confusion matrix. If we consider those reasonable location descriptions as correct using a relaxed matching method (Li *et al.* 2022), the performance of Geo-GPT-4 could be higher than the obtained scores. In Supplementary Figure S1, we show another confusion matrix which is based on the same result of Geo-GPT-4 but uses a relaxed matching method that allows the recognized location descriptions that have over 75% overlapping with the annotated descriptions to be considered as correct. This relaxed confusion matrix shows a 29.8% decrease in the errors in the 'Missed' column and a 28.3% decrease in the errors in the 'Not in ground truth' row, compared with the confusion matrix in Figure 6. Most of these partially matched descriptions, initially considered as errors, are also classified under the right location description categories, although some of them are still misclassified. While this result is encouraging, we need to be careful about using partially matched location descriptions as they could lead to potentially large errors in the geo-locating step.

6. Discussion

6.1. The importance of both geo-knowledge and GPT models for the proposed method

In this work, we have proposed a method that fuses geo-knowledge and a GPT model for extracting location descriptions and their categories from disaster-related social media messages. Both geo-knowledge and GPT models are important for the proposed method. Without geo-knowledge, the default GPT models do not know the typical forms of location descriptions used by people during disasters and how to categorize these location descriptions. As a result, GPT models have to rely on the information of location descriptions that they obtained during the pre-training stage. As demonstrated in our experiment result, adding geo-knowledge to the GPT models increases their performance by over 76%, and enables the GPT models to identify the categories of location descriptions. Meanwhile, advancements in GPT models are important for our proposed method as well. Without the invention of the GPT models, we cannot effectively integrate the geo-knowledge represented by a small number of examples into a typical machine-learning model. In addition, we also observe an increased performance overall when a more advanced GPT model is used. In sum, both geo-knowledge and GPT models are indispensable for the proposed method, and fusing the two enables us to effectively recognize location descriptions and their categories from disaster-related social media messages.

6.2. Implications for research in other geographic regions and data from other platforms

While this study has used tweets from Hurricane Harvey in the Houston area in Texas, the proposed method has flexibility and we expect that it could be applied to other geographic regions. For research within the United States, our method could be used directly since the geo-knowledge is still applicable. The location descriptions and their categories, such as door number address, road intersections, and highways, are commonly used in other geographic regions in the U.S. and are likely to be used by people in future disasters within the country. For research in other countries, the proposed method could be used by replacing the geo-knowledge of U.S. location descriptions with the geo-knowledge of the specific local region. In cases when the geo-knowledge of a local region is not available, research may need to be conducted to identify the local location descriptions, organize them into examples, and use the examples to guide GPT models in the same way as done in this research. In addition, since GPT models can handle multilingual texts, the proposed method could be used to analyze text messages in other languages beyond English. Nevertheless, future investigations are necessary to empirically test the applicability of the proposed method in other geographic regions and languages.

Our method may also be used to analyze text messages from other platforms, not limited to Twitter. There exists uncertainty related to the recent ownership change of Twitter and the announced changes in Twitter data API. While the future of this social media company may be unclear, it seems that people do need a digital platform during a disaster to post information and learn about current situations (Pourebrahim *et al*. 2019; Mihunov *et al*. 2020; Zhang *et al*. 2021). If Twitter engagement were to drop significantly, people may switch to other social media platforms, or a new digital platform for disaster response may be developed such as the Ushahidi platform used in the Haiti Earthquake in 2010 (Meier 2010). Our proposed method uses only the textual content of tweets and does not rely on other data or features that are specific to Twitter. Therefore, it could also be used for extracting location descriptions and their categories from text messages from other platforms.

6.3. A potential paradigm shift for using AI models for disaster response?

AI models have been increasingly adopted in disaster response (Kuglitsch *et al*. 2022). A typical approach in most recent studies is to fine-tune a pre-trained AI model, such as BERT, using newly labeled data (Wang *et al*. 2020a, 2020b; Suwaileh *et al*. 2022; Zhou *et al*. 2022). While such an approach avoids training an AI model completely from scratch, it still requires hundreds or thousands of annotated examples depending on the complexity of the specific task. In addition to training data, configuring the deep learning computing environment and fine-tuning AI models put additional requirements on disaster response organizations in terms of their technical expertise and computing resources. As shown in this study, fusing geo-knowledge and GPT models has the potential to substantially simplify the adoption of an AI model to a disaster response task.

On the one hand, disaster response organizations do not need to fine-tune their own AI models and configure their own deep learning environments locally. Instead, they can leverage the online AI models, such as the GPT models used in this study, and formalize their knowledge and other information about the task into prompts for adopting the AI model. It seems that such an approach can largely reduce the requirements of training data, technical expertise, computing resources (both hardware and software), and time put on disaster response organizations. In addition, this approach may also facilitate the collaboration between disaster response experts and AI model developers by allowing them to focus on what they are mostly good at: disaster response experts can focus on obtaining and formalizing knowledge by analyzing disaster-related data, while model developers can focus on developing and improving AI models. The knowledge obtained by disaster response experts can then be fused with the AI model hosted on the cloud computing environment through an API using prompts, similar to what we did in this study. Through such a collaboration, we may be able to empower disaster response organizations to utilize AI models more easily and more efficiently and let them focus on reaching and helping the people in need during a disaster.

On the other hand, there also exist important ethical issues that deserve our careful consideration. In the proposed approach, the social media messages of the disaster victims need to be submitted to an online GPT model, and malicious actors could take advantage of those submitted messages. While one could argue that those messages were shared as public tweets and could still be used by bad actors in their original form, further submitting them to an online model can increase their exposure. At the time of writing, a prompt of 'What were the addresses of the people who needed help during Hurricane Harvey?' tested on ChatGPT was not replied with any address, and ChatGPT included a statement in its reply that 'sharing personal addresses or identifying information would violate privacy and confidentiality guidelines.' Also, according to the Terms of Use from OpenAI, the content submitted via their API is not used to develop or improve their models. While it is good to see these efforts, further policies and technical measures are necessary to safeguard the content submitted to AI models. In addition, similar to other online platforms, who owns the data that are submitted to an online AI model? If law enforcement requests to use such data as evidence, do they have the right to do so? Should we apply stricter data protection policies for the content submitted under urgent situations such as natural disasters than that submitted during normal times? These are some of the important questions that we need to answer in order to use large AI models for disaster response in an ethical manner.

Could there be a paradigm shift in using AI models for disaster response, i.e., from locally trained and deployed AI models to online AI models guided by knowledge? While our work has shown promising results, more studies are necessary to understand the pros, cons, and ethical issues of using large online AI models for disaster response.

6.4. Limitations

This research is not without limitations. First, the performance of the Geo-GPT-4 model is still low on some location description categories, such as *C11: Road segments*. Based on our analysis of the experiment results, it seems that the model often confuses C11 with *C2: Street names* and *C5: Intersections*. While C11 indeed shares some similarities with the other two categories, we may need to think of better ways to inform GPT models about the differences among these categories. Second, our current study has focused on a dataset from the Houston area from Hurricane Harvey. While our method has potential to be applied to data from other geographic regions and other types of disasters, more empirical research is necessary to test the wider applicability of this method. Third, the overall performance of the Geo-GPT-4 model is still limited and could be improved through new methodological development in the near future. For example, we have used a question-answering approach in this study to encode geo-knowledge for informing GPT models, and future research could explore other ways to encode geo-knowledge, not limited to a series of questions and answers. Finally, this current study has focused on the step of recognizing location descriptions only, and it is important to explore the second step of geo-locating these descriptions as well. With the recognized location descriptions and their categories, we can further analyze and geo-locate them using suitable techniques and geometric representations. As mentioned earlier, uncertainty still exists even when suitable geo-locating techniques and geometric representations are used. A Bayesian approach could be adopted to quantify such uncertainty based on the location information contained in a message. For example, city name alone may be associated with a higher location uncertainty, while additional information, such as street name and door number, helps reduce such uncertainty.

7. Conclusions

Social media platforms are increasingly being used by people to share information and request help during natural disasters. Messages posted on these platforms often contain important descriptions of the locations of victims and accidents. In this study, we proposed a method that fuses geo-knowledge and GPT models for recognizing location descriptions and their categories. We presented the methodological details and conducted systematic experiments to compare the proposed method with other alternative approaches, including the typically used NER approaches and default GPT models. We found that geo-knowledge-guided GPT models achieve an over 40% improvement in recognizing location descriptions compared with off-the-shelf NER approaches, and this method uses only a small number of training examples encoding geo-knowledge. We also found that both geo-knowledge and GPT models are critical for the proposed method. Adding geo-knowledge to a GPT model results in an over 76% improvement compared with the same default GPT model; meanwhile, there is a general increase in performance when a more advanced GPT model is used. The highest performance was achieved by Geo-GPT-4 which demonstrated an F-score of 0.695 for recognizing location descriptions regardless of categories and an overall F-score of 0.644 for recognizing both location descriptions and their categories. In addition, Geo-GPT-4 achieved an F-score of 0.755 for the category of door number address which provides highly detailed information for locating victims. This method has been tested with different versions of GPT models, including GPT-2, GPT-3, ChatGPT, and GPT-4, and is likely to apply to more advanced GPT models in the coming years. By fusing geo-knowledge and GPT models, we may facilitate collaborations between disaster response experts and AI developers, reduce the technical burdens on disaster response organizations, and ultimately help save lives.

Acknowledgement

We thank the anonymous reviewers for their constructive comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data and code availability statement

The data and code that support the findings of this study are available on figshare at https://doi.org/10.6084/m9.figshare.22659337.

Funding

This work is supported by the U.S. National Science Foundation under Grant No. BCS-2117771: ‘Geospatial Artificial Intelligence Approaches for Understanding Location Descriptions in Natural

Disasters and Their Spatial Biases.' Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Notes on contributors

Yingjie Hu is an Associate Professor in the Department of Geography at the University at Buffalo. His research interests include GIScience, geospatial artificial intelligence (GeoAI), and disaster resilience. His contributions to this paper include conceptualization, data collection and curation, methodology, data analysis, result interpretation and discussion, visualization, writing – original draft and writing – editing and revision.

Gengchen Mai is currently a Tenure-Track Assistant Professor in the Department of Geography at the University of Georgia. His research interests are spatially explicit artificial intelligence, geographic knowledge graphs, geographic question answering, geospatial foundation models, and so on. His contributions to this paper include conceptualization, methodology, data analysis, result interpretation and discussion, visualization, writing – original draft, and writing – editing and revision.

Chris Cundy is a Ph.D. student at Stanford University. His research interests include variational inference, generative models, large language models, and AI safety. His contributions to this paper include methodology, result interpretation and discussion, and writing – editing and revision.

Kristy Choi is a Ph.D. candidate in Computer Science at Stanford University advised by Dr. Stefano Ermon. Her research is centered around machine learning with limited labeled supervision, and is focused on developing techniques for better adaptation and controllability in deep generative models. Her contributions to this paper include methodology, result interpretation and discussion, and writing – editing and revision.

Ni Lao is currently a research scientist at Google. He holds a Ph.D. degree in Computer Science from the Language Technologies Institute, School of Computer Science at Carnegie Mellon University, and is an expert in machine learning, knowledge graph and natural language understanding. His contributions to this paper include methodology, result interpretation and discussion, and writing – editing and revision.

Wei Liu has recently graduated with a M.S. in GIS from the University at Buffalo. Her interests include GeoAI, urban analytics, and social equality. Her contributions to this paper include data collection and curation and data analysis.

Gaurish Lakhanpal recently graduated from Stevenson High School and is currently studying computer science at Purdue University. He is deeply interested in machine learning. His contributions to this paper include data collection and curation, data analysis, and visualization.

Ryan Zhenqi Zhou (Zhenqi Zhou) is a Ph.D. candidate in the Department of Geography at the University at Buffalo and a research assistant in the GeoAI Lab. His research interests include GeoAI, disaster resilience, public health, and human mobility. His contributions to this paper include data collection and curation and data analysis.

Kenneth Joseph is an Assistant Professor in the Computer Science and Engineering Department at the University at Buffalo. He is a computational social scientist who focuses on the measurement and modeling of complex social systems. His contributions to this paper include result interpretation and discussion and writing – editing and revision.

ORCID

Yingjie Hu http://orcid.org/0000-0002-5515-4125
Gengchen Mai http://orcid.org/0000-0002-7818-7309

References

Adelfio, M.D., and Samet, H., 2013. Geowhiz: toponym resolution using common categories. *In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 532–535.

Berragan, C., *et al.*, 2023. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, 37 (4), 747–766.

Bommasani, R., *et al.*, 2021. On the opportunities and risks of foundation models. *ArXiv Preprint ArXiv:2108.07258*.

Brown, T., *et al.*, 2020. Language models are few-shot learners. *In:* H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds. *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc.,1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Cardoso, A.B., Martins, B., and Estima, J., 2021. A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS International Journal of Geo-Information*, 11 (1), Multidisciplinary Digital Publishing Institute: 28.

Chen, P., *et al.*, 2022. Crossroads, buildings and neighborhoods: a dataset for fine-grained location recognition. *In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, Washington, 3329–3339.

Crooks, A., *et al.*, 2013. # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17 (1), 124–147.

Dale, R., 2021. GPT-3: What's it good for? *Natural Language Engineering*, 27 (1), 113–118.

De Albuquerque, J.P., *et al.*, 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29 (4), 667–689.

De Longueville, B., Smith, R.S., and Luraschi, G., 2009. 'OMG, from Here, I Can See the Flames!' A use case of mining location based social networks to acquire spatio-temporal data on forest fires. *In:* X. Zhou and X. Xie, eds. P*roceedings of the 2009 International Workshop on Location Based Social Networks*. New York, NY: ACM, 73–80.

DeLozier, G., Baldridge, J., and London, L., 2015. Gazetteer-independent toponym resolution using geographic word profiles. *In: Twenty-Ninth AAAI Conference on Artificial Intelligence*. Washington, DC: AAAI Press, 2382–2388.

Devaraj, A., Murthy, D., and Dontula, A., 2020. Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction*, 51, 101757.

Devlin, J., *et al.*, 2019. 'BERT: pre-training of deep bidirectional transformers for language understanding.' 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dutt, R., *et al.*, 2018. Savitr: A System for real-time location extraction from microblogs during Emergencies. *In: Companion Proceedings of the The Web Conference 2018*. New York, NY: ACM, 1643–1649.

Eliot, L., 2022. 'Enraged worries that generative AI ChatGPT Spurs students to vastly cheat when writing essays, spawns spellbound attention for AI Ethics And AI Law.' *Forbes*. https://www.forbes.com/sites/lanceeliot/2022/12/18/enraged-worries-that-generative-ai-chatgpt-spurs-students-to-vastly-cheat-when-writing-essays-spawns-spellbound-attention-for-ai-ethics-and-ai-law.

Elsner, J.B., Elsner, S.C., and Jagger, T.H., 2015. The increasing efficiency of Tornado days in the United States. *Climate Dynamics*, 45 (3–4), 651–659.

Feng, Y., Huang, X., and Sester, M., 2022. Extraction and analysis of natural disaster-related vgi from social media: review, opportunities and challenges. *International Journal of Geographical Information Science*, 36 (7), 1275–1316.

Fernandes, C., *et al*., 2021. Automated disaster news collection classification and Geoparsing. *In: Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*. Rochester, NY: SSRN, 1–11.

Fernández-Martínez, N.J., 2022. The FGLOCTweet Corpus: an english tweet-based corpus for fine-grained location-detection tasks. *Research in Corpus Linguistics*, 10 (1), 117–133.

Freire, N., *et al*., 2011. A Metadata geoparsing system for place name recognition and resolution in metadata records. *In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, June 13–17 Ottawa, Ontario, Canada. New York, NY: ACM, 339–348.

Gelernter, J., and Balaji, S., 2013. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17 (4), 635–667.

Gelernter, J., and Mushegian, N., 2011. Geo-parsing messages from microtext. *Transactions in GIS*, 15 (6), 753–773.

Goldberg, D.W., Wilson, J.P., and Knoblock, C.A., 2007. From text to geographic coordinates: The Current State of Geocoding. *Urban and Rural Information Systems Association Journal*, 19 (1), 33–46.

Goodchild, M.F., and Haining, R.P., 2003. GIS and spatial data analysis: converging perspectives. *Papers in Regional Science*, 83 (1), 363–385.

Gritta, M., Taher Pilehvar, M., and Collier, N., 2018. Which Melbourne? Augmenting geocoding with maps. *In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 1285–1296.

Gritta, M., Taher Pilehvar, M., and Collier, N., 2020. A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, 54 (3), 683–712.

Gritta, M., *et al*., 2018. What's missing in geographical parsing? *Language Resources and Evaluation*, 52 (2), 603–623

Gupta, A., *et al*., 2013. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. *In: Proceedings of the 22nd International Conference on World Wide Web*. New York, NY: ACM, 729–736.

Hill, L.L., 2000. Core elements of digital gazetteers: Placenames, categories, and footprints. *In:* J. Borbinha and T. Baker, eds. *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, Vol. 1923. Berlin, Heidelberg: Springer, 280–290.

Hiltz, S.R., *et al*., 2020. Exploring the usefulness and feasibility of software requirements for social media use in emergency management. *International Journal of Disaster Risk Reduction*, 42, 101367.

Hu, X., *et al*., 2022. GazPNE2: a general place name extractor for microblogs fusing gazetteers and pretrained transformer models. *IEEE Internet of Things Journal*, 9 (17), 16259–16271.

Hu, Y., Janowicz, K., and Prasad, S., 2014. Improving wikipedia-based place name disambiguation in short texts using structured data from Dbpedia. *In: Proceedings of the 8th Workshop on Geographic Information Retrieval*, 4 November Dallas, Texas. New York, NY: ACM, 1–8.

Hu, Y., Mao, H., and McKenzie, G., 2019. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, 33 (4), 714–738.

Hu, Y., and Wang, J., 2021. How do people describe locations during a natural disaster: an analysis of tweets from hurricane harvey. *Proceedings of the 11th International Conference on Geographic Information Science*, 27–30 September 2021. Poznan, Poland: Schloss Dagstuhl - Leibniz Center for Informatics.

Hu, Y., and Wang, R.-Q., 2020. Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour*, 4 (12), 1219–1221.

Huang, Q., and Xiao, Y., 2015. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, 4 (3), 1549–1568.

Imran, M., *et al.*, 2015. Processing social media messages in mass emergency: a survey. *ACM Computing Surveys*, 47 (4), 1–38.

Imran, M., *et al.*, 2014. AIDR: artificial intelligence for disaster response. *In: Proceedings of the 23rd International Conference on World Wide Web*. New York, NY: ACM, 159–162.

Imran, M., *et al.*, 2020. Using AI and social media multimodal content for disaster response and management: opportunities, challenges, and future directions. *Information Processing & Management*, 57 (5), p. 102261.

Jones, C.B., and Purves, R.S., 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22 (3), 219–228.

Joseph, K., Landwehr, P.M., and Carley, K.M., 2014. An approach to selecting keywords to track on twitter during a disaster. *In: Proceedings of the 11th International ISCRAM Conference.* Brussels: Royal Flemish Academy of Belgium, 672–676.

Karimzadeh, M., *et al.*, 2019. GeoTxt: a scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23 (1), 118–136.

Kasneci, E., *et al.*, 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Knutson, T.R., *et al.*, 2010. Tropical cyclones and climate change. *Nature Geoscience*, 3 (3), 157–163.

Kuglitsch, M.M., *et al.*, 2022. Facilitating adoption of AI in natural disaster management through collaboration. *Nature Communications*, 13 (1), 1579.

Li, J., *et al.*, 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34 (1), 50–70.

Liu, F., Vasardani, M., and Baldwin, T., 2014. Automatic identification of locative expressions from social media text: a comparative analysis. In: *Proceedings of the 4th International Workshop on Location and the Web*, 3 November Shanghai, China. New York, NY: ACM, 9–16.

Longley, P.A., *et al.*, 2005. *Geographic Information Systems and Science*. Hoboken, NJ: John Wiley & Sons.

MacEachren, A.M., *et al.*, 2011. Senseplace2: Geotwitter analytics support for situational awareness *In: Proceedings of 2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Providence, RI, USA. New York, NY: IEEE, 181–190.

Mai, G., *et al.*, 2022. Towards a foundation model for geospatial artificial intelligence (Vision Paper). *In: Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 4 November Seattle, Washington. New York, NY: ACM, 1–4.

Mai, G., *et al.*, 2021. Geographic question answering: challenges, uniqueness, classification, and future directions. *AGILE: GIScience Series* 2. Copernicus Publications Göttingen, Germany: 8.

Mani, I., *et al.*, 2010. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44 (3), 263–280.

Manning, C.D., *et al.*, 2014. The Stanford CoreNLP. *In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics, 55–60.

Martín, Y., Li, Z., and Cutter, S.L., 2017. Leveraging Twitter to gauge evacuation compliance: spatiotemporal analysis of hurricane matthew. *PLOS One*, 12 (7), e0181701.

Meier, P., 2010. The unprecedented role of SMS in disaster response: learning from Haiti. *SAIS Review of International Affairs*, 30 (2), 91–103.

Melo, F., and Martins, B., 2017. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21 (1), 3–38.

Mhlanga, D., 2023. Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023). *In: FinTech and Artificial Intelligence for Sustainable Development. Sustainable Development Goals Series.* Cham: Palgrave Macmillan.

Mihunov, V.V., *et al.*, 2020. Use of Twitter in disaster rescue: lessons learned from hurricane Harvey. *International Journal of Digital Earth*, 13 (12), 1454–1466.

Murthy, D., and Longwell, S.A., 2013. Twitter and disasters: the uses of Twitter during the 2010 Pakistan floods. *Information, Communication & Society*, 16 (6), 837–855.

NOAA 2022. *U.S. Billion-Dollar Weather and Climate Disasters*. https://www.ncei.noaa.gov/access/billions/.

OpenAI 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

OpenAI 2023. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774.

Ouyang, L., *et al.*, 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

Pourebrahim, N., *et al.*, 2019. Understanding communication dynamics on twitter during natural disasters: a case study of Hurricane Sandy. *International Journal of Disaster Risk Reduction*, 37, 101176.

Purves, R.S., *et al.*, 2018. Geographic information retrieval: progress and challenges in Spatial search of text. *Foundations and Trends® in Information Retrieval*, 12 (2–3), 164–318.

Pustejovsky, J., Moszkowicz, J., and Verhagen, M., 2012. A linguistically grounded annotation language for spatial information. *Revue TAL*, 53 (2), 87–113.

Radford, A., *et al.*, 2018. Improving language understanding by generative pre-training. Accessed at: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., *et al.*, 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1 (8), 9.

Radke, M.A., *et al.*, 2022. Disambiguating spatial prepositions: the case of geo-spatial sense detection. *Transactions in GIS*, 26 (6), 2621–2650.

Ragini, J.R., Anand, P.R., and Bhaskar, V., 2018. Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, 42, 13–24.

Rhodan, M., 2017. Hurricane Harvey: The U.S.'s First Social Media Storm. Time Magazine. https://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/.

Scheele, C., Yu, M., and Huang, Q., 2021. Geographic context-aware text mining: enhance social media message classification for situational awareness by integrating spatial and temporal features. *International Journal of Digital Earth*, 14 (11), 1721–1743.

Seetharaman, D., and Wells, G., 2017. 'Hurricane harvey victims turn to social media for ASSISTANCE.' *With Local 911 Systems Failing, Residents Are Taking to Facebook and Twitter*, August. The Wall Street Journal. https://www.wsj.com/articles/hurricane-harvey-victims-turn-to-social-media-for-assistance-1503999001.

Silverman, L., 2017. Facebook, Twitter Replace 911 Calls For Stranded In Houston. NPR. https://www.npr.org/sections/alltechconsidered/2017/08/28/546831780/texas-police-and-residents-turn-to-social-media-to-communicate-amid-harvey.

Starbird, K., and Stamberger, J.A., 2010. Tweak the Tweet: leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. *In: Proceedings of the 7th International ISCRAM Conference*, May 2010 Seattle, Washington. Seattle, WA: ISCRAM.

Stock, K., *et al.*, 2022a. Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science*, 36 (3), 547–584.

Stock, K., Jones, C.B., and Tenbrink, T., 2022b. Speaking of location: a review of spatial language research. *Spatial Cognition & Computation*, 22 (3–4), 185–224.

Suwaileh, R., *et al.*, 2022. When a disaster happens, we are ready: location mention recognition from crisis tweets. *International Journal of Disaster Risk Reduction*, 78, 103107.

Tjong Kim Sang, E.F., and De Meulder, F., 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. *In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. Stroudsburg, PA: Association for Computational Linguistics. 142–147.

van Dis, E.A., *et al.*, 2023. ChatGPT: five priorities for research. *Nature*, 614 (7947), 224–226.

Vosoughi, S., Roy, D., and Aral, S., 2018. The spread of true and false news online. *Science*, 359 (6380), 1146–1151.

Wallgrün, J.O., *et al*., 2018. GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32 (1), 1–29.

Wang, J., and Hu, Y., 2019a. Enhancing spatial and textual analysis with EUPEG: an extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23 (6), 1393–1419

Wang, J., and Hu, Y., 2019b. Are we there yet? evaluating state-of-the-art neural network based geoparsers using eupeg as a benchmarking platform. *In:* Bruno M., Ludovic M., and Patricia M-F., eds. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*, Chicago, USA. New York, NY: ACM, 1–6.

Wang, J., Hu, Y., and Joseph, K., 2020a. NeuroTPR: a neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS*, 24 (3), 719–735.

Wang, R.-Q., *et al*., 2020b. Tracking flooding phase transitions and establishing a passive hotline with AI-enabled social media data. *IEEE Access*., 8, 103395–103404.

Wang, Z., Ye, X., and Tsou, M.-H., 2016. Spatial, temporal, and content analysis of twitter for wildfire hazards. *Natural Hazards*, 83 (1), 523–540.

Yu, M., *et al*., 2019. Deep learning for real-time social media text classification for situation awareness–using hurricanes sandy, harvey, and irma as case studies. *International Journal of Digital Earth*, 12 (11), 1230–1247.

Zhang, C., Yang, Y., and Mostafavi, A., 2021. Revealing unfairness in social media contributors' attention to vulnerable urban areas during disasters. *International Journal of Disaster Risk Reduction*, 58, 102160.

Zhang, J., and Goodchild, M.F., 2002. *Uncertainty in Geographical Information*. Boca Raton, FL: CRC press.

Zhou, B., *et al*., 2022. VictimFinder: harvesting rescue requests in disaster response from social media with BERT. *Computers, Environment and Urban Systems*, 95, 101824.

Zou, L., *et al*., 2018. Mining twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers*, 108 (5), 1422–1441.