International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: www.tandfonline.com/journals/tjde20

Optimizing context-based location extraction by tuning open-source LLMs with RAG

Zifu Wang, Yahya Masri, Anusha Srirenganathan Malarvizhi, Tayven Stover, Samir Ahmed, David Wong, Yongyao Jiang, Yun Li, Mathieu Bere, Daniel Rothbart, Dieter Pfoser, David Marshall & Chaowei Yang

To cite this article: Zifu Wang, Yahya Masri, Anusha Srirenganathan Malarvizhi, Tayven Stover, Samir Ahmed, David Wong, Yongyao Jiang, Yun Li, Mathieu Bere, Daniel Rothbart, Dieter Pfoser, David Marshall & Chaowei Yang (2025) Optimizing context-based location extraction by tuning open-source LLMs with RAG, International Journal of Digital Earth, 18:1, 2521786, DOI: 10.1080/17538947.2025.2521786

To link to this article: https://doi.org/10.1080/17538947.2025.2521786

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2025, VOL. 18, NO. 1, 2521786
https://doi.org/10.1080/17538947.2025.2521786

Optimizing context-based location extraction by tuning open-source LLMs with RAG

Zifu Wang [a], Yahya Masri[a], Anusha Srirenganathan Malarvizhi[a], Tayven Stover[a], Samir Ahmed[a], David Wong[b], Yongyao Jiang [a], Yun Li[a], Mathieu Bere[c], Daniel Rothbart[c], Dieter Pfoser[b], David Marshall[c] and Chaowei Yang[a]

[a]Department of Geography and Geoinformation Science, NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA, USA; [b]Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA; [c]Carter School for Peace & Conflict Resolution, George Mason University, Fairfax, VA, USA

ABSTRACT

Text data such as news from media include different types of geographic information, represented by location, that indicates the whereabout of events or phenomena. Extracting the geographic locations from text within their contexts is challenging, even with Natural Language Processing (NLP) tools and the latest Large Language Models (LLMs). We propose to optimize LLMs using Retrieval-Augmented Generation (RAG) and prompt-tuning methods, such as zero-shot and instruction-based prompting to improve the precision of extracting location information from news. Using Sudan conflict as an example, we extracted the corresponding locations and dates for conflict incidents. We compared runtime and accuracy of using various open-source LLMs, different hyperparameter settings, with and without RAG. Traditional Named Entity Recognition (NER), zero-shot prompting, instruction-based prompting, few-shot prompting, chain-of-thought (CoT) prompting, and RAG-based tuning were compared using an evaluation matrix. RAG-based tuning delivered the highest F1 score (>0.9) for extracting and associating location data with conflict incidents. This research highlights the benefits of using RAG for multi-incident context-based location extraction and provides insights into optimizing LLMs through prompt-tuning, hyperparameter adjustment, and model selection for location extraction tasks. The results can also be used to extract context-based locations or relevant information from text-based documents of other applications.

ARTICLE HISTORY
Received 12 November 2024
Accepted 4 June 2025

KEYWORDS
Context-based location extraction; large language model; retrieval augmented generation; natural language processing; Sudan conflict; media

List of abbreviations

LLM	Large Language Model
NER	Named Entity Recognition

CONTACT Chaowei Yang chaowei.yang.1@gmail.com MS 6A2, George Mason University, Fairfax, VA, 22030-4444, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17538947.2025.2521786.

GPT	Generative Pre-Trained Transformer
BERT	Bidirectional Encoder Representations from Transformers
FAISS	Facebook AI Similarity Search
RAG	Retrieval-Augmented Generation
CoT	Chain-of-Thought
BiLSTM	Bidirectional Long Short-Term Memory
NLP	Natural Language Processing

1. Introduction

Media such as news articles, social media posts, videos, and transcriptions often contain embedded geographic information – including descriptions of locations, regions, and movement patterns – that may be extracted to analyze associated events or phenomena (Lopez, Magliocca, and Crooks 2019). This geographic information plays a pivotal role in understanding the evolution of associated events, behaviors, and trends across temporal and spatial dimensions (Shi and Barker 2011; Stefanidis, Crooks, and Radzikowski 2013). In times of crisis, such as natural disasters, disease outbreaks, or armed conflicts like the ongoing Sudan conflict, the ability to extract accurate spatiotemporal information from media data is essential for tracking the spread and impact of these events. This extraction process, known as location extraction, enables timely decision-making, facilitates targeted responses, and ultimately supports more effective crisis management (Havas et al. 2021; Tang et al. 2018; Yu et al. 2020). Whether it's monitoring the movement of refugees during a conflict or allocating resources in a natural disaster, understanding where and when events occur allows stakeholders to act efficiently and mitigate further risks, making location extraction a crucial aspect for both researchers and responders (Havas et al. 2021; Tang et al. 2018; Yu et al. 2020).

Accurately extracting locations from media content using automated computational tools is essential for researchers to uncover spatiotemporal patterns of events. However, this information is often complex, as location details are frequently embedded within broader contextual elements (Hoang and Mothe 2018; Middleton et al. 2018). Context-based location includes not only place names but also more intricate geographic entities, such as neighborhoods, landmarks like harbors and mountains, and directional terms such as 'eastern' (Chen et al. 2022; Hu and Wang 2020). Additionally, locations are often intertwined with temporal elements, such as dates or times. A single article may contain multiple locations and various types of temporal data (Li et al. 2003; Strötgen, Gertz, and Popov 2010). Moreover, to facilitate further analyses based on the extracted location data, it is crucial to extract the data into a well-structured format to reduce the effort needed for data cleaning after the extraction (Goldberg, Wilson, and Knoblock 2009).

NLP has been proven to be an effective tool for automatically extracting geospatial information from media content (Small and Medsker 2014). Traditional methods, including geocoding, geoparsing, and geotagging, have been used to identify locations. These methods aim to map straightforward place names and detailed location descriptions to their corresponding geographic coordinates (Middleton et al. 2018; Wang, Hu, and Joseph 2020). However, the introduction of transformer architecture in 2017, with its self-attention mechanism, marked a significant advancement in NLP by allowing models to capture long-range dependencies in text more efficiently (Vaswani et al. 2017). Self-attention enables the model to focus on different parts of a sentence simultaneously,

improving tasks like machine translation, text generation, and comprehension by better understanding the context and relationships between words, regardless of their position in the text (Vaswani et al. 2017). This architecture led to the development of powerful LLMs like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-Trained Transformer (GPT) (Devlin et al. 2019; Radford 2018), both of which have demonstrated remarkable capabilities in various NLP tasks, including the extraction of geospatial information (Hu et al. 2023; Manvi et al. 2023). Despite these advancements, challenges remain in accurately extracting complex, multi-entity location descriptions. For example, in the context of the current Sudan conflict, a single news article may reference several incidents with different dates and locations. Simple prompt-tuning methods struggle to produce the required structured output, where each incident must be associated with 'neighborhood, state, country, and date' across multiple lines in the document. These methods often miss outputting one or more of these components, which disrupts the format expected for post-processing the data. When the output lacks this structured format, human intervention is needed to correct or validate the information, adding an extra step before any automated processing.

The recently introduced Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) combines document retrieval with generative models to enhance the ability of LLMs to deliver accurate and contextually relevant answers (Lewis et al. 2020). The model retrieves relevant information from a large corpus or knowledge base before generating responses, improving its performance on tasks that require knowledge beyond what is stored in the model's parameters (Lewis et al. 2020). The effectiveness of the method has been evaluated in respect to tasks like question answering, summarization, and fact-checking, demonstrating its potential to optimize the location extraction process by addressing key challenges such as improving the accuracy of multi-entity location extraction, handling complex geographic descriptions, and ensuring consistency in the structured output format. (Ling et al. 2023; Manvi et al. 2023; Yan et al. 2023).

This research aims to address these challenges of context-based location extraction by implementing and comparing various methods, including RAG, traditional Named Entity Recognition (NER), and prompt-tuning strategies. The study focuses on enhancing the accuracy of context-based location extraction from news articles, using articles related to the Sudan conflict as a use case. By comparing RAG, NER, and prompt-tuning methods as well as evaluating the influence of hyperparameters, this research seeks to identify the most effective methods and configurations for extracting location using different open-source LLMs. Additionally, this study assesses the runtime and accuracy performance of these models to determine the best solution for spatiotemporal extraction tasks.

2. Related work

2.1. Traditional NER approach for location extraction

Previous studies have employed a range of methodologies for regular location extraction, typically falling under the category of geoparsing. Geoparsing involves the recognition of toponyms in text, often utilizing existing NER tools like Stanford NER, which is particularly adept at identifying country and city names (The Stanford NLP Group 2023). NER tools are designed to extract toponyms from text by treating locations as a subtype of

named entities (Middleton et al. 2018). For example, (Karimzadeh et al. 2019) proposed a geoparsing system that incorporates Stanford NER for extracting location mentions from unstructured text. (Hu et al. 2023) evaluated several existing NER tools, focusing on their performance in identifying geographic entities across different application domains. In recent years, researchers have developed innovative techniques by incorporating neural network models like the Bidirectional Long Short-Term Memory (BiLSTM) and fine-tuning transformer models and integrated those methods with NER to further enhance the performance of location information comprehension for NER (Hu et al. 2022; Wang, Hu, and Joseph 2020). Kuai et al. attempted to extract spatial context-based local toponyms from urban POI data by identifying as many potential address components as possible from continuous text strings for each POI, and merging neighboring address components into toponyms based on their spatial context (Kuai et al. 2020).

However, while these approaches, from an NER perspective, excel in identifying the forms of toponyms, they often fall short of recognizing more detailed location information, such as neighborhood names that are not included in their dictionary as well as understanding the complex relationships between the extracted locations and associated context (Gritta et al. 2018; Hu et al. 2023). Additionally, some frameworks rely heavily on manually input toponyms and their spatial context prior to extraction, limiting their generalizability to less structured addresses or those containing ambiguous geographic information (Kuai et al. 2020).

2.2. Recent GPT approaches for location extraction

LLMs have had a profound impact across diverse domains, including manufacturing, education, healthcare, and business. LLMs empower users to tailor conversations to specific requirements, encompassing factors like desired length, format, style, level of detail, and language. Prompt tuning, a technique that adapts prompts rather than modifying model parameters, has proven effective in enhancing the performance of LLMs while requiring fewer resources. Prompt tuning, particularly useful in clinical concept extraction and reasoning tasks, leverage strategies like few-shot prompting, zero-shot prompting, and instruction-based prompting to guide LLMs toward improved outcomes with minimal training data (Peng et al. 2024; Sahoo et al. 2024; Wu et al. 2024). Recent research has shown that LLMs equipped with attention mechanisms hold promise for improving location extraction accuracy and enhancing the depth of information extraction, particularly in relation to associated topics. For example, (Hu et al. 2023) demonstrated the implementation of GPT models to extract locations in a disaster management scenario by testing zero-shot, few-shot, and chain-of-thought (CoT) strategies. (Wang et al. 2025) also compared the performance of different GPT models on various NER tasks, such as random retrieval, sentence-level embedding, and entity-level embedding.

However, using regular prompt-tuning methods in LLMs for context-based location extraction often falls short of achieving full-span matching, necessitating human intervention to clean and validate the generated information (Fernandez and Dube 2023; Hu et al. 2023; Ji and Gao 2023). For example, LLMs tend to output unnecessary information, such as explanations alongside the extracted locations, which cannot be directly used for further analysis and thus requires additional cleansing. However, there is a lack

of literature that shows how much RAG can improve location extraction tasks as compared to regular prompt-tuning. For example, (Hu et al. 2023) have examined that to output a full-span matching between the model-recognized description and the human-annotated location description, ChatGPT4 could only achieve an F1 score of 0.394. With the fusion of Hu's Geo-knowledge into the ChatGPT4 model, the F1 score will reach 0.695 (Hu et al. 2023). While almost all research used ChatGPT as the base model for location extraction, there is a lack of research comparing open-source LLMs specifically for context-based location extraction tasks.

2.3. RAG implementation on geospatial sciences

Recent studies have evaluated the effectiveness of employing RAG in tasks like question answering, summarization, and fact-checking (Ling et al. 2023; Shlyk et al. 2024; Xiong et al. 2024; Yan et al. 2023). Specifically in Geoscience, (Xia et al. 2024) developed a Q&A system for typhoon disasters using a RAG-based approach, which involves continuous pretraining and fine-tuning with disaster-specific data. Adopting the approach improved the performance in delivering accurate and contextually relevant information to users during disaster scenarios. (Manvi et al. 2023) extracted geospatial knowledge, such as population density and economic livelihoods from auxiliary map data in OpenStreetMap by using RAG-enhanced LLMs.

While these studies have demonstrated the applicability of RAG in handling geospatial information, there is limited research specifically focusing on context-based location extraction from news media sources, which often involve multiple incidents and locations in a single news article. To facilitate automated analysis of the incident in the subsequent step, locations extracted from news articles must precisely match the required format, including all relevant details such as place names and administrative divisions.

2.4. Current challenges and contribution of this study

Based on the literature, the current challenges toward context-based location extraction from text includes that 1) identifying complex, multi-entity location descriptions and their associated thematic elements is difficult in a required format for further automatic processing with a high accuracy. 2) Regular prompt-tuning methods often fall short, necessitating human intervention to clean and validate the generated information. For example, in the context of the current Sudan conflict, a single news article may reference several incidents with different dates and locations. When GPT models are tasked with extracting information in a fixed format, such as 'neighborhood, state, country, and date,' from news content, they often generate additional, unnecessary explanations alongside the requested information. 3) There is a lack of studies that specifically investigate how much RAG can improve the accuracy of context-based location extraction compared to regular prompt-tuning methods. 4) There is a lack of studies that specifically investigate how different open-source LLMs perform on context-based location extraction on the aspect of runtime and accuracy to help researchers reduce the cost compared to using closed-source LLMs.

To address the above challenges, this study focuses on 1) presenting the practical difficulties of extracting location information from news articles related to the Sudan

conflict, where multiple location references are often present within a single article; 2) how RAG can improve the accuracy of context-based location extraction when compared to traditional prompt-tuning methods by effectively handling the complexity of extracting accurate location data; and 3) evaluating the performance of different open-source LLMs in both runtime and accuracy, and providing insights into the most efficient configurations for using open-source LLMs with RAG for the tasks.

3. Data sources

The Data used in this study comprises detailed news reports of conflict incidents that occurred in Sudan between April and September of 2024. These reports were gathered from various reliable sources, including CNN (CNN 2024), the (Sudan War Monitor 2024), (Sudan Tribune 2024), Asharq Al-Awsat 2024), the International Committee of the Red Cross (ICRC) (ICRC 2024), Xinhua News and Radio Dabanga (Radio Dabanga 2024). Each incident has been manually reviewed and cross-verified by Sudan conflict experts to ensure accuracy and reliability. These experts not only confirmed the occurrence of the incidents but also meticulously labeled key information such as location, incident date, and type of event, establishing a reliable ground truth for this study.

Table 1. The structure of the incident dataset and an example of the incident.

Attribute	Example
Date	5/27/2024
Incident Narrative	As a result of the intensification of fighting between the RSF and the Army who is backed by the Joint Force of Armed Struggle Movements, both the security and humanitarian situation has deteriorated according to several reports. The Dar El Salam Emergency Room in North Darfur said in a statement yesterday that almost 20,000 displaced people have fled to the locality to escape fighting in El Fasher and other areas in the past weeks, most of whom are staying with host families. 'There are 11 shelters, 8 of which are schools, in Dar El Salam. The displaced are living in difficult conditions due to the lack of external assistance from humanitarian organisations.' Whereas the SAF claims that it has 'successfully expelled the RSF outside of the eastern borders of El Fasher', the RSF declared on May 26 that it expelled an attack by the SAF in El Fasher on May 25 and accused the army and allies of sheltering in displaced camps and using civilians as human shields.
Incident type	Military operations (battle, shelling)
Incident impact	Humanitarian impact: IDP/Refugees flow
(Presumed) perpetrator	Both SAF & RSF
Number of Fatalities	[empty]
State	North Darfur
Location	El Fasher
Specifics about the location	[empty]
Feature	city
Latitude	13.6198
Longitude	25.3549
Maxar Imagery Status	Inactive
ReadyToMap	Yes
Satellite imagery request	Yes
Source 1	Radio Dabanga
URL Link to source 1	https://www.dabangasudan.org/en/all-news/article/msf-employee-killed-by-shell-as-north-darfur-fighting-rages-on
Source 2	UN OCHA
URL Link to source 2	https://x.com/CNkwetaSalami/status/1794709045280149769
Source 3	Radio Dabanga
URL Link to Source 3	https://www.dabangasudan.org/en/all-news/article/north-darfur-civilians-flee-catastrophic-escalation-in-conflict

A total of 377 conflict incidents were recorded during the selected timeframe, covering 21 categories of events such as military operations, war crime, willful killing of civilians, etc. For this research, 78 incidents that occurred in May 2024 were selected as the dataset to evaluate different methods of context-based location extraction using various LLMs. Table 1 shows the structure of the dataset and an example of the incident. The cell 'Incident Narrative' was inputted into LLMs, and the 'State' and 'Location' fields were used to verify and validate the outputs of LLMs.

4. Methodologies

Figure 1 outlines the workflow for evaluating the performance of using different methods and open-source LLMs to extract context-based locations. After collecting data through incident logs that were verified by domain experts, each news article was processed using four methods: NER, zero-shot prompting (Kojima et al. 2022), instruction-based prompting (Brown et al. 2020), and RAG. NER was implemented using Python packages, including spaCy and Geopy (Geopy 2008; Honnibal and Montani 2017). Both zero-shot and instruction-based methods were tuned with prompts to extract the locations. For RAG implementation, Facebook AI Similarity Search (FAISS), an open-source library developed by Meta for efficient similarity search and clustering of dense vectors, was utilized to create a vector database for each article and used the same prompt to query the LLM for context-based location extraction (Jégou, Douze, and Johnson 2017; Johnson, Douze, and Jégou 2019). Each method was tested across different open-source LLMs with various hyperparameter settings. Finally, recall, precision, and F1 scores were used to evaluate the experiments. More details are provided in the sub-sections.

4.1. Methods configuration

4.1.1. NER

This method does not involve any implementation of LLMs. Only the Python packages, spaCy and Geopy were implemented to analyze the geospatial labels of each word entity

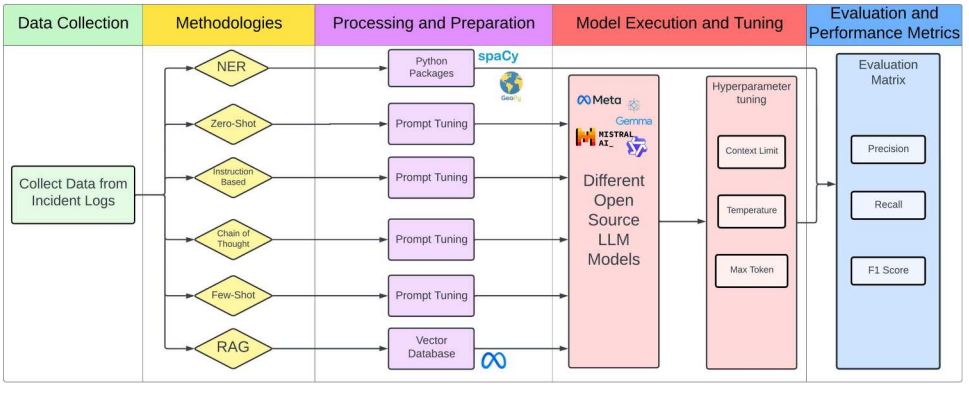

Figure 1. Workflow of investigating the performance of context-based location extraction from news documents.

in the 'Incident Narrative' cell. If a word entity is identified by spaCy with the labels 'GPE,' 'LOC,' or 'DATE,' it will be extracted. 'GPE' (Geopolitical Entity) refers to countries, cities, states, or other political regions, while 'LOC' (Location) refers to physical locations that are not political entities, such as mountains, rivers, or general geographic features. 'DATE' refers to any recognized date or date-like expression in the text. Once the relevant information is extracted, GPE and LOC entities are passed to the GeoPy package to determine the city, state, and country information. Finally, all data are output as a string in the format of 'neighborhood, state, country, date'.

The following three methods are all tested using different LLMs with various hyperparameters.

4.1.2. Zero-shot prompting

For the zero-shot prompt-tuning strategy, a fixed prompt template was used (provided in Appendix B1). The system prompt instructs the model to generate the location and date of each incident described in the news articles. In the user prompt, the content from the 'Incident Narrative' cell serves as the input, with the expectation that the LLMs will output only the required information in the fixed format: 'neighborhood, state, country, and MM/DD/YYYY.'

4.1.3. Instruction-based prompting

In the system prompt for the instruction-based method, the LLMs were specifically instructed to produce outputs according to the required format to ensure that they understood what information to provide and what to exclude. The prompt template is provided in Appendix B2.

In this template, '' includes instructions for formatting the extracted geographic locations and dates based on the response schema outlined below, '' represents the 'Incident Narrative' cell containing the news article about the Sudan conflict, and '' is the generated response. It is important to note that differences in prompt descriptions were minimized to ensure a fair comparison between the zero-shot and instruction-based prompting methods.

4.1.4. Few-shot prompting

For the few-shot prompting strategy, a prompt template was used that included multiple examples to demonstrate the desired input-output behavior. The complete few-shot prompt template is provided in Appendix B3. In this template, the system prompt includes multiple examples demonstrating the expected output format. In the user prompt, the content from the 'Incident Narrative' cell serves as the input, with the expectation that the LLMs will generate a list of location-date pairs, each on a separate line, following the same formatting as in the examples.

4.1.5. Chain-of-thought prompting

For the CoT prompting strategy, the prompt template is provided in Appendix B4. In this template, the system prompt instructs the model to explain its reasoning for each extracted location and date before outputting the result. Several additional examples beyond those shown were included in the prompt to guide the model through various types of temporal expressions and complex geographic references. The user prompt

provides the narrative input, and the model is expected to produce a two-part response: detailed reasoning and a clearly marked final output.

4.1.6. Retrieval-augmented generation (RAG)

As shown in Figure 2, the process of implementing RAG involved utilizing FAISS, a vector database tool developed by Meta. This process begins by inputting context-based news articles into the embedding base model, where the articles are converted into numerical vectors. These vectors capture the semantic content and structural information of the articles and are then stored in the FAISS database. The same input is converted into different numerical vectors when processed by different LLMs, as each model is trained on a unique corpus, which affects the numerical representation of words and the relationships between them. When a user query is submitted as a prompt, it is also processed through the embedding model to retrieve relevant information from the vector database. By using vectorized news articles retrieved along with the prompt, the LLM is fine-tuned to extract geographic information from each article. It is important to note that the same prompt is used as in zero-shot prompting to evaluate how much RAG improves location extraction.

Prior to implementing RAG, we designed a series of experiments to compare multiple prompt-tuning strategies – including zero-shot, few-shot, instruction-based and CoT prompting – across different open-source LLMs. Based on this comparison, we would select the most consistent and effective prompting strategy as the basis for the RAG pipeline. The same prompt used in the selected strategy would then be applied within the RAG framework to evaluate its impact on context-based location extraction.

4.2. Hyperparameters

In this research, four specific hyperparameters were selected for tuning – model variants, temperature, context limits, and maximum tokens. Model variants represent different LLM architectures developed by various organizations, each with unique training corpus and objectives. Model examples include Gemma2-9B and Llama3.1-7B, where the numbers after the dash indicate the approximate size of model parameters in billions. In this study, various open-source LLMs were selected for comparison, including Llama3.1-7b and Llama3.1-70b from Meta. The Llama3.1 series offers multiple models with varying parameter sizes and generally provides a larger context window compared

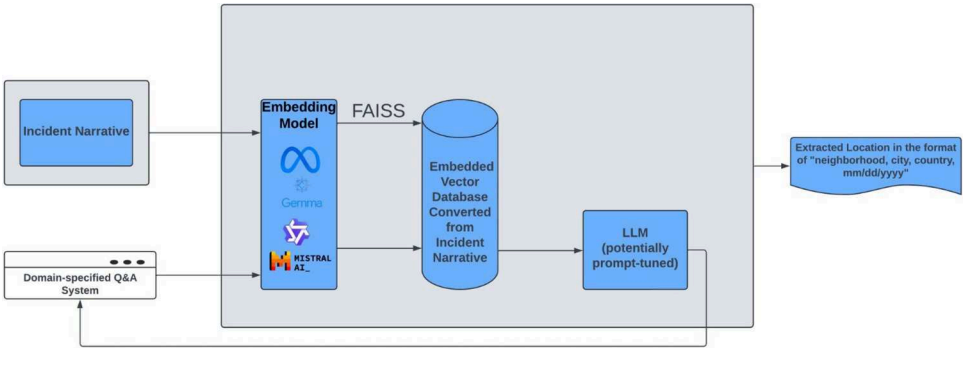

Figure 2. Workflow of Implementing RAG to Extract Location.

to earlier versions (Dubey et al. 2024). Gemma2 was selected from Google DeepMind (Team et al. 2024). Mistral was selected from Mistral (Jiang et al. 2023) and Qwen was selected from Alibaba (Bai et al. 2023). In the NeedleBench evaluation (Li et al. 2024), a framework for assessing long-context comprehension and reasoning, differences among these models became evident, with Llama3.1 and Qwen outperforming Gemma2 and Mistral in long-context tasks. Generally, a model with more parameters (e.g. Llama3.1-70B vs. Llama3.1-7B) offer better performance, particularly in tasks requiring nuanced understanding or complex reasoning (Zhang et al. 2024).

Temperature controls the randomness of the LLM's outputs during text generation. A higher temperature increases randomness, producing more creative or varied responses, while a lower temperature generates more mundane and factual outputs. Since context-based location extraction requires precise information, a lower temperature is generally more effective, as it reduces variability and enhances the accuracy of location generation (Yu et al. 2024; Renze 2024).

Context limits define the maximum amount of input text the model can process at once, which is particularly important in processing long news articles with multiple location mentioned. Adjusting the context limits ensure that the model can handle longer inputs coherently, improving its ability to extract information from complex texts (Ding et al. 2024; Jin et al. 2024).

Maximum tokens set a limit on the total number of tokens that the LLM can generate in a sequence. This includes both input tokens and generated outputs, a practice that helps manage the overall length of the text and keeps the model within computational limits. Tuning this parameter ensures that the model generates concise and relevant outputs without exceeding resource constraints (Ding et al. 2024; Jin et al. 2024).

4.3. Evaluation metrics

To evaluate the performance of various methods for extracting context-based locations from news articles, a manual evaluation system was implemented, recognizing the limitations of relying on LLMs' self-evaluation (Chern et al. 2024; Chiang and Lee 2023). This manual labeling process served as the standard for assessing the accuracy across different LLMs in extracting context-based location information, ensuring that any discrepancies between the model output and the manually labeled ground truth could be manually measured and addressed during the evaluation process. A comprehensive scoring system was developed to quantify performance, focusing on four key components: date, location, state, and country. Each correctly identified component contributes equally, with each component worth 1 point, resulting in a total possible score of 4 points per incident. If the response did not adhere to the specified schema (e.g. presented in paragraph form instead of the required structured format), it automatically receives a score of zero. It is critical that the output fully matches the prompt requirements to ensure seamless automatic data processing without human intervention. As specified, the LLM outputs are expected to follow the format 'neighborhood, state, country, and date (MM/DD/YYYY)' for each incident. If the LLM generated unnecessary information, such as explanations or reasoning, the output was assigned a score of zero, as this additional content introduces noise and requires further manual effort to clean before launching automated processing, such as spatiotemporal pattern analysis.

Once the performance score is manually calculated, it is used to determine recall, precision, and F1 scores, three widely accepted metrics for evaluating extraction performance (Gritta, Pilehvar, and Collier 2018; Hu et al. 2023; Purves et al. 2018). As shown in equations (1-5), recall is calculated by dividing the total correct points earned by the maximum possible points from manual annotation. To minimize the impact of random variations or outliers, each method is tested multiple times under the same hyperparameter settings and with the same LLM. The recall score is then averaged across all news articles through multiple rounds of testing. precision is calculated by dividing the total correct points earned by the maximum possible points from all results produced by the model. Similarly, precision is averaged across all articles through multiple rounds of testing. The F1 score, which is the harmonic mean of precision and recall, will only be high when both precision and recall scores are high.

$$Recall = \frac{Total\ Points\ Earned}{Maximum\ Possible\ Points\ from\ Manual\ Annotation} \tag{1}$$

$$Recall\ Score = \frac{1}{n}\sum_{i=1}^{n} Recall \tag{2}$$

$$Precision = \frac{Total\ Points\ Earned}{Maximum\ Possible\ Points\ from\ Correctly\ Recognized\ Results\ by\ Model} \tag{3}$$

$$Precision\ Score = \frac{1}{n}\sum_{i=1}^{n} Precision \tag{4}$$

$$F1\ score = 2 * \frac{Precision\ Score * Recall\ Score}{Precision\ Score\ +\ Recall\ Score} \tag{5}$$

5. Results

5.1. NER

Although NER tools identified and classified word entities in text into predefined categories such as people, locations, and dates, they presented significant limitations in meeting the requirements for location extraction in this study. For example, NER tools often failed to extract the country 'Sudan' unless the exact term was explicitly mentioned in the text. Even when related entities such as 'Sudanese Liberation Movement' or 'Sudanese Armed Forces' were referenced, the country name was not recognized by the NER tools. NER also struggled to interpret the relationship between dates and locations in context, frequently extracting irrelevant phrases such as 'a second consecutive day' or 'that day' instead of correctly identifying the actual date or leaving the date field blank. The format of responses was also inaccurate, contributing to poor performance. For instance, when parsing complex sentences describing incidents and locations, NER often produced disjointed outputs like 'North Darfur, Friday' or 'Al-Kahraba, May 10,' rather than a coherent extraction like 'El Fasher, North Darfur, Sudan, May 10.' These challenges underscore the need for more context-aware NER systems. As seen in

Table 2, the overall precision, recall, and F1 scores for NER were 50.3%, 34.3%, and 40.8%, respectively.

5.2. Zero-shot prompting

Table 3 shows the overall performance of different LLMs on context-based location extraction under zero-shot prompting. Gemma2-27b achieved the highest overall precision, recall, and F1 scores across various temperature settings, with optimal performance at a temperature of 0.1. While zero-shot prompting occasionally performed well, it demonstrated significant limitations, particularly in consistency and formatting. One key issue identified was the failure to accurately extract dates, even when they were explicitly mentioned in the text. For example, the article states that '*On May 2, the International Committee of the Red Cross (ICRC) reported that two of its drivers were killed by gunmen in South Darfur, Sudan*', the model produced the output '*Layba, South Darfur, Sudan, mm/dd/yyyy*', failing to recognize May 2 altogether. Additionally, zero-shot prompting frequently generated responses with excessive or random formatting, reducing overall accuracy. For instance, responses like ' – *Neighborhood: Omdurman, State: Khartoum, Country: Sudan, mm/dd/yyyy: 05/03/2024*' deviated from the specified format given to the model. Furthermore, formatting issues were exacerbated when the model output responses in paragraph form rather than in an ordered list, often including unnecessary headings and subheadings.

Table 2. NER performance.

Method	Precision	Recall	F1 Score
NER	50.3%	34.3%	40.8%

Table 3. Zero-shot prompting performance.

Model	Temp	Context Limit	Maximum Token	Runtime	Precision	Recall	F1 Score
Gemma2 - 27b	1	8192	−1	6 m 6s	87.5%	87.2%	87.4%
	0.5	8192	−1	6 m 9s	87.5%	86.8%	87.2%
	0.1	8192	−1	5 m 58s	88.8%	87.9%	**88.3%**
	0	8192	−1	6 m 47s	88.1%	87.3%	87.7%
Gemma2 - 9b	1	8192	−1	2 m 39s	88.6%	84.9%	86.7%
	0.5	8192	−1	2 m 30s	88.1%	85.7%	86.9%
	0.1	8192	−1	2 m 31s	87.4%	87.1%	87.2%
	0	8192	−1	2 m 45s	88.9%	85.4%	87.1%
Llama 3.1 - 70b	1	8192	−1	10 m 22s	78.4%	77.5%	77.9%
	0.5	8192	−1	10 m 18s	77.7%	76.7%	77.2%
	0.1	8192	−1	9 m 51s	78.3%	77.6%	78.0%
	0	8192	−1	9 m 52s	78.3%	77.5%	77.9%
Llama 3.1 - 7b	1	8192	−1	1 m 56s	72.1	74.7%	73.4
	0.5	8192	−1	1 m 32s	71.9	75.4%	73.6%
	0.1	8192	−1	2 m 20s	72.1	74.9%	73.5%
	0	8192	−1	1 m 57s	71.2	75.2%	73.1%
Qwen – 7b	1	8192	−1	1 m 29s	70.1	65.6%	67.8%
	0.5	8192	−1	1 m 51s	70.0	65.4%	67.6%
	0.1	8192	−1	1 m 37s	70.5%	64.6%	67.4%
	0	8192	−1	1 m 38s	71.1	62.1%	66.3%
Mistral – 7b	1	8192	−1	2 m 23s	56.4	41.5%	47.8%
	0.5	8192	−1	2 m 14s	56.3	42.8%	48.6%
	0.1	8192	−1	2 m 26s	56.0%	44.3%	49.5%
	0	8192	−1	2 m 21s	55.1%	44.9%	49.5%

5.3. Instruction-based prompting

Table 4 presents the overall performance of different LLMs on context-based location extraction using instruction-based prompting. Gemma2-9b achieved the highest precision, recall, and F1 scores across various temperature settings, with optimal performance at a temperature of 0. Initially, it was expected that instruction-based prompting would outperform zero-shot prompting; however, its overall performance score was lower. A major issue with instruction-based prompting was its tendency to include extra, unrequested information in the output. For instance, one output example is:

```
{
'Date': '05/03/2024',
'Geographic Location': ' Omdurman, Khartoum, Sudan '
}
```

This is an example of a conflict incident that occurred in North Darfur state: In El Fasher Town, the only working hospital – Southern Hospital's intensive care unit (ICU) was damaged by a 'strike,' causing the roof to collapse. This event happened on 2024 News, as reported by UN emergency relief chief Martin'.

In this case, while the date and location were extracted correctly, the additional explanation was not required and detracted from the overall response quality. Furthermore, instruction-based prompting struggled with identifying multiple locations within the same article, resulting in lower overall success rates for those instances.

Table 4. Instruction-based prompting performance.

Model	Temp	Context Limit	Maximum Token	Runtime	Precision	Recall	F1 Score
Gemma2 - 27b	1	8192	−1	3 m 10s	87.5%	67.5%	76.2%
	0.5	8192	−1	3 m 10s	89.1%	68.8%	77.7%
	0.1	8192	−1	3 m 10s	88.5%	69.4%	77.8%
	0	8192	−1	3 m 12s	88.7%	68.1%	77.0%
Gemma2 - 9b	1	8192	−1	1 m 36s	73.4%	68.9%	71.1%
	0.5	8192	−1	1 m 38s	84.4%	73.9%	78.8%
	0.1	8192	−1	1 m 33s	91.2%	74.4%	81.9%
	0	8192	−1	1 m 36s	97.8%	74.8%	**84.8**%
Llama 3.1 - 70b	1	8192	−1	7 m 12s	74.3%	59.1%	65.9%
	0.5	8192	−1	6 m 15s	82.3%	65.7%	73.1%
	0.1	8192	−1	6 m 38s	88.9%	71.3%	79.2%
	0	8192	−1	6 m 11s	89.8%	72.5%	80.2%
Llama 3.1 −7b	1	8192	−1	1 m 36s	71.1%	61.3%	65.8%
	0.5	8192	−1	56s	80.7%	71.3%	75.7%
	0.1	8192	−1	1 m 2s	89.8%	72.5%	80.2%
	0	8192	−1	1 m 2s	91.2%	69.6%	78.9%
Qwen – 7b	1	8192	−1	2 m 37s	70.1%	41.2%	51.9%
	0.5	8192	−1	2 m 29s	85.7%	44.0%	58.1%
	0.1	8192	−1	1 m 40s	91.3%	43.3%	58.7%
	0	8192	−1	1 m 28s	93.8%	43.9%	59.8%
Mistral – 7b	1	8192	−1	59s	84.5%	68.5%	75.7%
	0.5	8192	−1	1 m 12s	87.7%	68.3%	76.8%
	0.1	8192	−1	1 m 7s	89.2%	69.6%	78.2%
	0	8192	−1	1 m 6s	93.9%	71.2%	81.0%

5.4. Few-shot prompting

Table 5 presents the overall performance of different LLMs on context-based location extraction using few-shot prompting. Gemma2-27b achieved the highest overall precision, recall, and F1 scores across various temperature settings, with optimal performance at a temperature of 0. The performance of few-shot prompting varied across models, with several notable issues observed in the outputs. A recurring problem was the tendency of models to extract speculative or irrelevant locations in order to match the format demonstrated in the examples. In one instance, the article stated: '*On Sunday, May 5th, Sudanese warplanes conducted airstrikes targeting agricultural fields where RSF elements are reported located, near Shabarqa, a critical RSF stronghold and strategic town that is located approximately 25 kilometers east of Wad Madani, El Gezira (aka Al Jazeera) state. Capturing Shabarqa, according to local sources, would open the path towards the Hantoub Bridge on the Blue Nile for the Army.*' Gemma2-27b at a temperature of 1 produced the following output:

agricultural fields near Shabarqa, El Gezira, Sudan, 5/5/2024

Shabarqa, El Gezira, Sudan, 5/5/2024

Hantoub Bridge, El Gezira, Sudan, 5/5/2024

In this case, only the first entry accurately reflected the site of the airstrike. The second output named Shabarqa itself, even though the article specified the incident occurred near Shabarqa. The third entry, Hantoub Bridge, was not a site of conflict but a possible future objective, and should not have been extracted. In another example, the article stated: '*On Thursday, May 16, a health center in Um Rawaba was attacked by assailants who plundered medical equipment and medications, according to reports*

Table 5. Few-shot prompting performance.

Model	Temp	Context Limit	Maximum Token	Runtime	Precision	Recall	F1 Score
Gemma2 - 27b	1	8192	−1	5 m 59s	79.4%	78.4%	78.8%
	0.5	8192	−1	5 m 10s	79.2%	77.9%	78.4%
	0.1	8192	−1	4 m 58s	79.6%	78.9%	79.1%
	0	8192	−1	5 m 5s	80.2%	80.0%	**80.0%**
Gemma2 - 9b	1	8192	−1	3 m 28s	77.6%	77.5%	77.0%
	0.5	8192	−1	3 m 30s	77.9%	78.1%	79.9%
	0.1	8192	−1	3 m 30s	78.3%	78.2%	78.1%
	0	8192	−1	3 m 41s	78.6%	78.3%	78.4%
Llama 3.1 - 70 b	1	8192	−1	9 m 27s	54.5%	59.5%	54.1%
	0.5	8192	−1	9 m 35s	55.1%	49.5%	52.5%
	0.1	8192	−1	9 m 51s	44.2%	43.3%	43.6%
	0	8192	−1	9 m 39s	43.9%	43.1%	43.4%
Llama 3.1 −7b	1	8192	−1	1 m 58s	44.8%	45.6%	44.3%
	0.5	8192	−1	2 m 2s	45.1%	44.6%	44.6%
	0.1	8192	−1	2 m 3s	45.9%	45.5%	45.1%
	0	8192	−1	2 m 9s	46.1%	45.7%	45.9%
Qwen – 7b	1	8192	−1	1 m 47s	78.8%	77.1%	77.8%
	0.5	8192	−1	1 m 51s	78.3%	76.5%	77.3%
	0.1	8192	−1	1 m 45s	79.1%	77.9%	78.5%
	0	8192	−1	1 m 49s	79.1%	78.2%	78.1%
Mistral – 7b	1	8192	−1	2 m 40s	64.4%	63.6%	63.0%
	0.5	8192	−1	2 m 50s	66.2%	64.4%	65.4%
	0.1	8192	−1	2 m 31s	69.1%	68.2%	68.5%
	0	8192	−1	2 m 50s	69.4%	68.6%	68.9%

by local residents.' Gemma2-27b returned: *Um Rawaba, North Kordofan, Sudan, 5/16/2024*, whereas Gemma2-9b, in contrast, returned: *Um Rawaba, Sudan, 5/16/2024*. Although the article did not mention the state, only the larger model (27b) appended 'North Kordofan,' which is factually correct. This highlights the difference in capability between model sizes, with larger models better able to complete missing administrative context. Finally, an example of hallucination was observed in the following article: '*Sudan War Monitor reported on May 21 that the SAF have been targeting both foreigners and citizens in Gedaref State, especially Ethiopians, arresting at least 16 Ethiopian refugees, according to local sources*.' The model output was: *Gedaref State, Sudan, Ethiopia, 5/21/2024*. Here, 'Ethiopia' was incorrectly included as a location, even though it was not a site of any incident. The article only referenced Ethiopian refugees living in Sudan. This type of error illustrates how few-shot prompting can over-extract any mentioned location in the article, even when it is not contextually relevant. Furthermore, a key limitation was the often inclusion of all mentioned locations, even when some were merely contextual references rather than actual sites of incidents.

5.5. Chain-of-thought prompting

Table 6 presents the overall performance of various LLMs on location extraction using CoT prompting. Gemma2-9b achieved the highest F1 score, outperforming both smaller and larger models. CoT prompting improved interpretability by requiring models to justify their outputs, often leading to more accurate administrative completion. However, it also introduced errors when outputs failed to align with the model's own reasoning.

For instance, in response to the following article:

> 'On Sunday, May 5th, Sudanese warplanes conducted airstrikes targeting agricultural fields where RSF elements are reported located, near Shabarqa, a critical RSF stronghold … ',
>
> the model reasoned that the event occurred near Shabarqa but still extracted the location as:
>
> Shabarqa, El Gezira, Sudan, 5/5/2024.

This reflects a mismatch between the model's internal reasoning and its final output. Although the spatial detail was preserved in the explanation, it was lost in the formatted response, which is an example of how CoT prompting may improve reasoning but still yield factually imprecise extractions.

A more severe issue was observed in a May 22 article describing violence in the 'northern and eastern parts of the city' and the Abu Shouk IDP Camp. Although the article never named the city, the model hallucinated 'El Geneina' as the location. In reality, Abu Shouk Camp is located in El Fasher, North Darfur, making this a factual inaccuracy. These examples highlight CoT's tendency to overcomplete missing information based on prior knowledge, leading to hallucinated outputs that do not align with the source text.

5.6. Retrieval augmented generation

Based on the experiments in Sections 5.2–5.5, which evaluated zero-shot, few-shot, instruction-based, and CoT prompting strategies, we compared their effectiveness

Table 6. Chain-of-Thought prompting performance.

Model	Temp	Context Limit	Maximum Token	Runtime	Precision	Recall	F1 Score
Gemma2 - 27b	1	8192	−1	7 m 32s	84.7%	83.6%	84.1%
	0.5	8192	−1	7 m 23s	84.8%	83.2%	83.9%
	0.1	8192	−1	7 m 38s	84.9%	82.1%	83.4%
	0	8192	−1	7 m 33s	84.9%	82.3%	83.5%
Gemma2 - 9b	1	8192	−1	5 m 3s	87.9%	87.5%	87.7%
	0.5	8192	−1	5 m 13s	88.2%	88.1%	87.9%
	0.1	8192	−1	5 m 2s	88.8%	88.1%	88.5%
	0	8192	−1	5 m 4s	89.5%	88.4%	**88.9%**
Llama 3.1 - 70 b	1	8192	−1	16 m 19s	84.5%	82.7%	83.2%
	0.5	8192	−1	14 m 42s	85.9%	83.1%	84.4%
	0.1	8192	−1	15 m 6s	88.2%	84.8%	86.1%
	0	8192	−1	15 m 20s	89.4%	85.0%	86.9%
Llama 3.1 −7b	1	8192	−1	3 m 50s	76.3%	71.1%	73.3%
	0.5	8192	−1	3 m 43s	79.6%	77.8%	78.5%
	0.1	8192	−1	3 m 51s	84.1%	81.5%	81.9%
	0	8192	−1	3 m 50s	85.2%	82.2%	83.5%
Qwen – 7b	1	8192	−1	3 m 18s	78.6%	78.3%	78.4%
	0.5	8192	−1	3 m 18s	80.6%	78.2%	78.9%.
	0.1	8192	−1	3 m 19s	82.3%	80.5%	81.2%
	0	8192	−1	3 m 15	83.1%	81.8%	82.4%
Mistral – 7b	1	8192	−1	4 m 9s	76.6%	73.2%	74.6%
	0.5	8192	−1	3 m 45s	75.4%	73.4%	74.7%
	0.1	8192	−1	4 m 1s	78.1%	74.9%	77.9%
	0	8192	−1	4 m 13s	79.1%	75.2%	78.4%

across multiple open-source LLMs. On the Gemma2–27B model, zero-shot prompting achieved F1 scores between 87.2% and 88.3%, outperforming CoT by 3.8% to 4.2% (CoT: 83.4%–84.1%). On the Gemma2–9B model, CoT slightly outperformed zero-shot, with scores ranging from 87.7% to 88.9%, compared to 86.7% to 87.2% from zero-shot, a difference of 1.0% to 1.7%. Although CoT showed a marginal advantage on the smaller model, zero-shot prompting demonstrated greater consistency across both Gemma2 models, with lower performance variation and reduced sensitivity to prompt structure. In contrast, CoT was more prone to output instability.

In addition, although the previous results show that zero-shot prompting offers more consistent performance, we still conducted experiments combining CoT with RAG on Gemma2-9b using a temperature of 0. We found that while the model occasionally produced correct and well-reasoned outputs, this integrative approach suffered from two key issues; 1) significantly increased inference time and 2) inconsistent final predictions. Furthermore, in several cases, the correct location was mentioned in the reasoning tokens but not reflected in the final output. These limitations made the CoT-RAG combination less reliable and not as efficient for further use cases. Given that Gemma2 models consistently outperformed other LLMs tested (e.g. LLaMA3, Qwen, Mistral), and that our experiments demonstrated zero-shot prompting to be the most stable and reproducible performance – compared to few-shot, instruction-based, and CoT promptings – we selected zero-shot prompting as the base prompt-tuning strategy for implementing RAG.

Table 7 presents the overall performance of various LLMs on location extraction using RAG. Gemma2-9b still achieved the highest F1 scores, with its best performance remaining at a temperature of 0. The use of RAG significantly improved performance on the Gemma2, Qwen, and Mistral models, with Mistral seeing an approximate 30% increase in accuracy. However, a decrease in performance was observed on the Llama model. One

Table 7. Retrieval-Augmented Generation performance.

Model	Temp	Context Limit	Maximum Token	Runtime	Precision	Recall	F1 Score
Gemma2 - 27b	1	8192	−1	7 m 6s	91.9%	90.1%	91.0%
	0.5	8192	−1	7m	92.1%	91.0%	91.6%
	0.1	8192	−1	7 m 5s	92.1%	90.7%	91.4%
	0	8192	−1	7 m 7s	91.7%	89.6%	90.6%
Gemma2 - 9b	1	8192	−1	3 m 26s	89.6%	91.7%	90.6%
	0.5	8192	−1	3 m 22s	89.8%	92.0%	90.9%
	0.1	8192	−1	3 m 14s	90.1%	92.9%	91.5%
	0	8192	−1	3 m 18s	90.1%	93.5%	**91.8**%
Llama 3.1 - 70 b	1	8192	−1	9 m 15s	82.3%	78.4%	80.3%
	0.5	8192	−1	9 m 1s	85.2%	80.8%	82.9%
	0.1	8192	−1	8 m 54s	85.1%	80.2%	82.6%
	0	8192	−1	8 m 53s	84.9%	81.1%	83.0%
Llama 3.1 −7b	1	8192	−1	2 m 31s	68.6%	72.8%	70.6%
	0.5	8192	−1	2 m 32s	68.9%	74.7%	71.7%
	0.1	8192	−1	2 m 26s	69.3%	73.3%	71.2%
	0	8192	−1	2 m 35s	69.1%	74.8%	71.8%
Qwen – 7b	1	8192	−1	2 m 36s	76.5%	78.7%	77.6%
	0.5	8192	−1	2 m 32s	75.7%	79.7%	77.6%
	0.1	8192	−1	2 m 36s	76.1%	79.1%	77.6%
	0	8192	−1	2 m 37s	77.2%	76.5%	76.8%
Mistral – 7b	1	8192	−1	2 m 56s	81.5%	81.1%	81.3%
	0.5	8192	−1	3 m 3s	81.1%	81.6%	81.3%
	0.1	8192	−1	3 m 6s	80.5%	81.6%	81.0%
	0	8192	−1	3 m 3s	79.7%	83.2%	81.4%

of the key strengths of RAG observed in this study was its ability to handle complex queries involving multiple locations within the same paragraph, an area where other models often struggled. For example, when given the same dataset, RAG was able to identify multiple specific locations, such as villages and neighborhoods, in contrast to zero-shot and instruction-based prompting, which often only extracted a single location. Here is an example: '*Despite calls for the cessation of hostilities, the RSF has intensified these last three days its offensive against El Fasher, North Darfur, which has led to an escalation of violent clashes … The most impacted areas have been the densely populated districts in the south and north of the city, such as Al-Inqaz, Al-Salam, Al-Wahda, Al-Hijra, Oulad Al-Reef, and Makraka, which were affected by artillery bombardment.*' Instruction-based prompting could only extract '*El Fasher, North Darfur, Sudan*', whereas RAG accurately identified and extracted the locations as '*Al-Inqaz, Al-Salam, Al-Wahda, Al-Hijra, Oulad Al-Reef, Makraka, El Fasher, North Darfur, Sudan, 03/14/2024*', significantly enhancing the accuracy of extracted information.

However, in some models like Llama3.1-70B, RAG did not consistently adhere to formatting instructions. For instance, in a May 7 article, the model accurately extracted both the location and the date but failed to present them on the same line as prompted – placing the date on a new line instead. While the factual content was correct, the deviation from the expected structure suggests a limitation in formatting control, which may stem from the language model's generation behavior rather than the retrieval process itself.

Furthermore, in a May 19 article, the Gemma2-9B RAG model failed to extract the full specific location, identifying only El Fasher rather than the name of the hospital where the incident occurred. Although such cases are less frequent compared to other prompting methods, they nonetheless demonstrate that RAG does not fully eliminate extraction errors or formatting inconsistencies.

6. Discussion

6.1. Prompting strategy comparison

To mitigate concerns regarding potential bias in the manual evaluation process, this section presents a direct, article-level comparison of outputs generated by each prompting strategy when applied to the same input. To further enhance transparency and reproducibility, the original full-text articles used for these comparisons are included in Appendix A Text A1 and Text A2, corresponding to Appendix Tables A1 and A2 respectively. Two side-by-side comparisons are presented in Appendix A Tables A1 and A2. In both cases, the same LLM, Gemma2-9B, was used across all prompting strategies with a fixed temperature of 0, to ensure consistency and isolate prompting as the primary variable. This model was selected based on prior experiments demonstrating superior overall performance relative to other tested LLMs.

Table A1 compares outputs for the article in Text A1 describing a targeted airstrike. The RAG-based approach performed optimally in this case, correctly identifying the incident location as near Shabarqa and avoiding distractor references such as Wad Madani and Hantoub Bridge. In contrast, both the zero-shot and few-shot strategies failed to make this distinction, instead outputting a list of all locations mentioned in the article, including the non-incident locations. The instruction-based prompting method failed completely, yielding an invalid output that included extraneous phrases (e.g. 'Let me know if you have any other news articles you'd like me to analyze') and improper formatting (e.g. enclosing output in triple quotes, labeling it 'json'), which resulted in a score of zero. CoT prompting achieved partial success, identifying 'Shabarqa' as the location; however, it omitted the critical qualifier 'near,' which led to a reduced F1 score.

Table A2 evaluates prompting strategies using the article in Text A2 that describes an attack on medical infrastructure. RAG achieved the highest score, correctly identifying both the specific location (a health center in Um Rawaba) and the state (North Kordofan), even though the article did not explicitly mention the state. Notably, RAG was the only strategy to identify both components successfully. Few-shot, zero-shot, and instruction-based prompting all failed to recover the state and additionally omitted the specific incident location, outputting only 'Um Rawaba.' CoT produced the least accurate response, hallucinating the state as North Darfur and failing to identify the specific facility targeted, resulting in the lowest score among the strategies.

Collectively, these case studies reinforce the conclusion that RAG prompting yields more precise and contextually grounded location extraction, even in scenarios where state or incident specificity is underspecified in the source text. Moreover, the results highlight that differences across prompting strategies are not solely a matter of completeness but also concern the quality and faithfulness of the extracted locations.

6.2. Overall observations

Using the Sudan conflict as a use case, this paper systematically evaluates the accuracy and speed of using various LLMs, relevant tuning methods as well as integrating RAG for context-based location extraction. As Figure 3 shows, the implementation of RAG significantly improved the overall F1 score across all models and parameters,

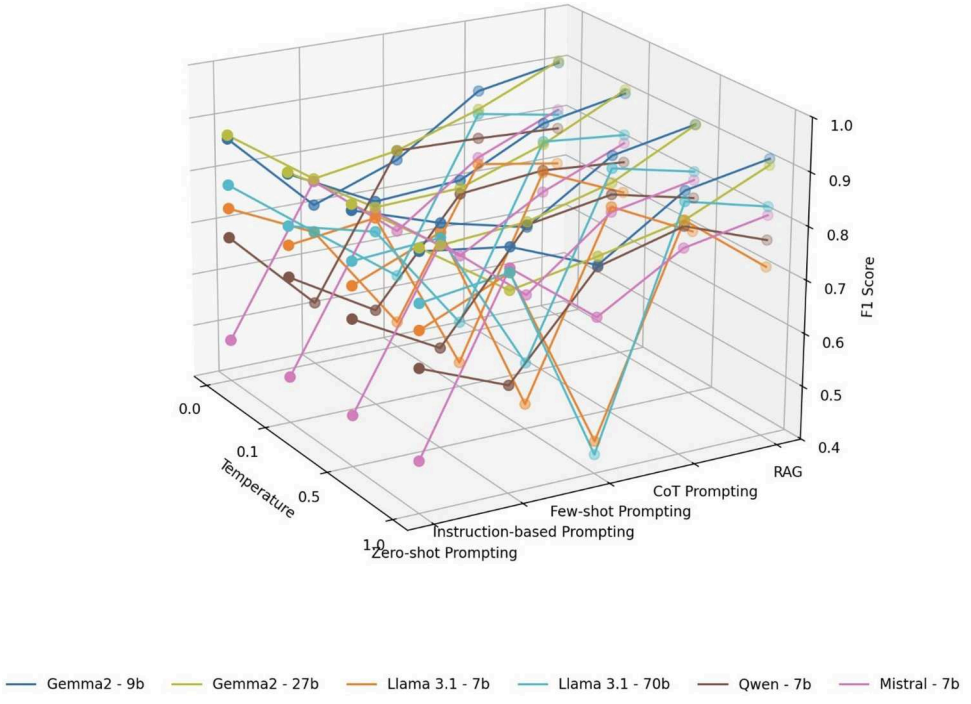

Figure 3. Overall Performance of Context-based Location Extraction.

outperforming the standard NER, zero-shot, and instruction-based methods by margins of 5% to 30%, depending on model selection and prompt configuration. In most cases, lower temperature settings, such as 0 and 0.1, produced higher F1 scores for context-based location extraction. This result aligns with the expectation that extracting factual information from LLMs benefits from more deterministic outputs rather than creative generation. Conversely, higher temperatures often led to the generation of unnecessary information, such as reasoning or explanations, which negatively impacted the score.

Among the LLM models tested, Gemma2-29B demonstrated the best overall performance, achieving an F1 score of up to 91.8% with a temperature setting of 0 when using RAG as highlighted in Table 5. This performance can be attributed to Gemma2's distinctive model architecture (Purves et al. 2018), which incorporates Local Sliding Window and Global Attention in alternating layers (Beltagy, Peters, and Cohan 2020; Team et al. 2024), as well as Logit Soft-Capping in each attention layer and the final layer within its decoder-only transformer design (Luong et al. 2015). The Local Sliding Window mechanism segments text into overlapping windows, allowing the model to concentrate on local context within each segment. This segmentation reduces computational costs while retaining detailed local insights (Team et al. 2024). Meanwhile, Global Attention augments the model's context-handling by periodically focusing on significant global tokens across segments, thus capturing broader contextual relationships that span the input (Beltagy, Peters, and Cohan 2020). Additionally, Logit Soft-Capping stabilizes model output by capping raw logit scores, preventing them from

becoming excessively large. This technique reduces numerical instability, promotes balanced output distributions, and mitigates overconfidence in certain predictions. Such balanced probability distributions are especially advantageous for tasks that require nuanced, varied responses across potential outputs (Luong et al. 2015).

Another finding is that larger models, which typically require longer runtimes, do not necessarily improve accuracy, highlighting the importance of balancing model size with efficiency in context-based location extraction tasks. This is evident in Table 7, where Gemma2 – 7b achieved results identical to Gemma2 – 27b while maintaining a runtime that was twice as fast as the larger model. Furthermore, Table 4 shows similar findings, with Llama3.1 - 7b and Llama3.1 - 70b, achieving comparable results across all hyperparameters, and showcasing how larger models do not necessarily equate to higher performance scores.

To support these evaluations, all experiments were conducted on a high-performance workstation equipped with an Intel(R) Xeon(R) w3-2423 CPU, 128 GB of RAM, and dual NVIDIA RTX A6000 GPUs (each with 48 GB of VRAM). This configuration enabled smooth and efficient execution of large-scale language models such as Gemma2–27B and Llama3.1–70B, particularly within RAG pipelines that utilized extended context lengths and batch processing. The A6000 setup offered substantial acceleration and served as the benchmark for runtime comparisons.

To assess deployment feasibility under more accessible hardware conditions, we also tested inference on an RTX 4060 system with 8 GB of VRAM and 48 GB of RAM. While capable of running large models, the 4060 required significantly longer runtimes. For instance, Llama3.1–70B inference took approximately one hour for a single long-context prompt. We further tested another machine with the following system configurations: GTX 1070 Ti with 8 GB VRAM and 16 GB RAM, which was only able to run smaller models under constrained batch sizes and reduced context lengths, confirming the limited practicality of older GPUs for large-batch tasks with larger sized LLMs. Based on our findings, we recommend a minimum of 24 GB VRAM and 64 GB RAM for effective use of models in the 13–30B parameter range when using RAG or long-context queries. For smaller models (e.g. 7B), GPUs with 8 GB VRAM and at least 16 GB RAM can suffice, though with trade-offs in speed and context size (Wang et al. 2023).

Furthermore, different parameter settings for context limit and maximum tokens were tested but found that they did not significantly affect overall performance of context-based location extraction tasks unless they were set to extreme values. This is primarily because context limit determines the maximum amount of input text the model can process at one time, while the maximum tokens set a limit on the total number of tokens that the LLM can generate in a sequence. In other words, the context limit affects the length of the input, and maximum tokens control the length of the output. However, in this research, the inputs were of fixed length, and the required outputs were relatively short, as the task did not require complex reasoning. The only scenario where these parameters made a noticeable difference was when the maximum tokens were set to extremely low values, preventing the model from fully outputting a location name. This would negatively impact the performance of all models. However, such an extreme setting was deemed irrelevant, as it would prevent any meaningful extraction. Therefore, for this study, the context limit was set to a fixed value of 8192 (the

maximum context length for most LLMs) and maximum tokens to – 1 (allowing the model to output as many tokens as needed).

6.3. Future work

This study's findings can be applied to other areas of research that require automatic location and information extraction, such as automated map generation through LLMs based on text input, including disaster monitoring, urban planning, and environmental science.

Although our current study implemented standard RAG techniques, we have also conducted initial experiments using Neo4j to store extracted location and date information from conflict incidents. While GraphRAG – an emerging approach that uses LLMs to build knowledge graphs from retrieved context – was explored (Edge et al. 2024; Peng et al. 2024), we found that the accurate extraction of location and date must occur prior to meaningful graph construction. In our setup, once structured data was stored in Neo4j, we could query the database using GraphRAG to answer spatiotemporal questions with minimal hallucination. For instance, when asked the question, '*What humanitarian crisis did RSF cause between May 10th and May 20th in Khartoum?*' the system responded:

> 'The provided documents do not contain information about any humanitarian crisis caused by the RSF in Khartoum between May 10th and May 20th.
>
> They focus on incidents in other locations:
>
> - West Kordofan: Attacks on a religious leader's home and a local resistance movement in Gbeish (May 14th).
> - North Darfur: Clashes between RSF and Sudanese Army in El Fasher, including attacks on the Abu Shouk IDP camp (May 16th – 22nd).There is no mention of Khartoum in these reports.'

In this example, we intentionally posed a misleading question to test the hallucination tendency of the LLM, as no events actually occurred in Khartoum during the specified time frame. The result shows how a structured spatiotemporal knowledge base can produce precise, document-grounded answers. The system correctly identified that no relevant events occurred in Khartoum during the specified period – without hallucinating a false response. This highlights the value of post-extraction knowledge graphs like Neo4j in enabling reliable downstream reasoning. Future work will focus on refining this GraphRAG-enhanced pipeline to support domain-specific expert systems.

Integrating agent-based frameworks, such as those offered by LangChain, presents a promising path for enhancing automation and decision-making in geospatial data workflows. These frameworks enable large language models (LLMs) to dynamically invoke tools like GIS APIs, vector databases, and code interpreters based on task-specific needs. Recent studies have demonstrated the potential of LLM-powered agents in geospatial contexts: (Ning et al. 2025) proposed an autonomous GIS agent framework that uses LLMs to generate, execute, and debug code for retrieving

complex geospatial datasets, while (Li et al. 2025) outlined a broader research agenda advocating for autonomous GIS that leverages generative AI for spatial analysis, knowledge graph construction, and automated map creation. However, most studies rely on closed-source LLMs such as GPT-4o, which raise concerns about deployment cost, dependency on commercial APIs, and data privacy. Many GIS applications involve sensitive or proprietary data – such as disaster response, military planning, or urban infrastructure – that cannot be shared with external services due to security, regulatory, or ethical constraints. These limitations make cloud-hosted, commercial LLMs unsuitable for many real-world GIS scenarios. In contrast, our work explores workflows leveraging open-source LLMs that can potentially be integrated into agentic frameworks, offering a path toward on-premises deployment that ensures both cost efficiency and data control. This direction could lower the barrier to adoption for research institutions and government agencies seeking to build secure, autonomous GIS systems tailored to their domains.

While this study focuses on conflict-related news articles, the proposed extraction framework has broader applications in domains such as disaster monitoring and environmental science – where timely and accurate geospatial information is critical. For example, in disaster monitoring, the framework could be used to extract affected locations and event timelines from real-time news and social media feeds during natural disasters such as floods, wildfires, earthquakes, or pandemic, enabling rapid response and resource allocation (Chen et al. 2022; Wang et al. 2022; Yu et al. 2019). In environmental science, the method could assist in tracking environmental incidents such as pollution events, deforestation, or biodiversity loss by automatically extracting spatial and temporal references from reports, studies, and field notes (Liu et al. 2021; Malarvizhi et al. 2023). The framework can also be adapted for multi-language inputs to support global use cases and integrated with open-source GIS tools like QGIS for spatial visualization and analysis. Additionally, incorporating the framework into low-code platforms such as KNIME would enhance usability for non-programmers (Fu et al. 2025; Liu et al. 2024). By combining LLM-based extraction with KNIME's visual workflow environment, users could automate data collection, processing, and geospatial integration in a reproducible and modular way. This would facilitate practical deployment across government, research, and humanitarian settings where flexible and interpretable pipelines are essential.

7. Conclusion

This study explored the limitations of traditional NER methods in context-based location extraction, particularly in handling complex descriptors and relationships in spatiotemporal data. The evaluation of LLMs demonstrated that standard prompt-tuning methods struggled to deliver accurate results when multiple locations and dates were required in a structured format. To address these challenges, RAG was integrated to improve context-based location extraction performance from news articles about the Sudan conflict.

The research also compared the performance of different open-source LLMs in terms of runtime and accuracy, while examining the impact of hyperparameters on context-based location extraction tasks. Although RAG-based tuning did not consistently

outperform zero-shot and instruction-based prompting across all models, it delivered the highest F1 scores in Gemma2 and significantly improved performance in most of the open-source models tested.

This work opens several promising directions for building flexible, domain-adaptable geospatial systems. Integrating GraphRAG and Neo4j offers potential for constructing structured spatiotemporal knowledge bases to support accurate, context-aware reasoning. The use of open-source LLMs also presents a cost-effective and privacy-preserving alternative to closed-source models, particularly for applications requiring local deployment. Furthermore, incorporating the framework into platforms like KNIME and QGIS could enable low-code, end-to-end solutions for disaster monitoring and environmental science – laying the groundwork for future autonomous GIS systems.

Acknowledgement

We appreciate the comments and advice received from MITRE and State department colleagues while we conduct the research.

Disclosure statement

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

Funding

The research is supported by the Department of State, MITRE Inc., and NSF (1841520) National Science Foundation U.S. Department of State

Data availability statement

The Sudan incident log data can be provided upon request.

ORCID

Zifu Wang http://orcid.org/0000-0002-7183-5166
Yongyao Jiang http://orcid.org/0000-0002-4591-483X

References

Asharq Al-Awsat. 2024. Asharq AL-awsat | International and Arab News. [online] Available at: https://english.aawsat.com/[Accessed 16 Oct. 2024].

Bai, J., S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, et al. 2023. "Qwen Technical Report." *arXiv Preprint ArXiv:2309.16609*.

Beltagy, I., M. E. Peters, and A. Cohan. 2020. "Longformer: The Long-Document Transformer." *arXiv Preprint ArXiv:2004.05150*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal.2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Chen, Y., Y. Li, Z. Wang, A. J. Quintero, C. Yang, and W. Ji. 2022. "Rapid Perception of Public Opinion in Emergency Events through Social Media." *Natural Hazards Review* 23 (2): 04021066. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000547

Chen, P., H. Xu, C. Zhang, and R. Huang. July 2022. "Crossroads, Buildings and Neighborhoods: A Dataset for Fine-Grained Location Recognition." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3329–3339.

Chern, S., Z. Hu, Y. Yang, E. Chern, Y. Guo, J. Jin, B. Wang, and P. Liu. 2024. "BeHonest: Benchmarking Honesty of Large Language Models." *arXiv Preprint ArXiv:2406.13261*.

Chiang, C. H., and H. Y. Lee. 2023. "Can Large Language Models Be an Alternative to Human Evaluations?" *arXiv Preprint ArXiv:2305.01937*.

CNN. 2024. Sudan in focus. [online] Available at: https://www.cnn.com/world/sudan [Accessed 16 Oct. 2024].

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1: 4171–4186.

Ding, Y., L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, and M. Yang. 2024. "Longrope: Extending LLM Context Window beyond 2 Million Tokens." *arXiv Preprint ArXiv:2402.13753*.

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, et al. 2024. "The Llama 3 Herd of Models." *arXiv Preprint ArXiv:2407.21783*.

Edge, D., H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. 2024. "From Local to Global: A Graph rag Approach to Query-Focused Summarization." *arXiv Preprint ArXiv:2404.16130*.

Fernandez, A., and S. Dube. 2023. "Core Building Blocks: Next Gen Geo Spatial GPT Application." *arXiv Preprint ArXiv:2310.11029*.

Fu, X., L. Liu, W. W. Guan, Y. Kalra, S. Bao, T. Kötter, and K. Sturm. 2025. "Advancing Replicable and Reproducible GIScience: An Approach with KNIME." *Cartography and Geographic Information Science* : 1–21. https://doi.org/10.1080/15230406.2024.2446556.

Geopy. 2008. geopy: Python Geocoding Toolbox. Available at: https://geopy.readthedocs.io [Accessed 16 Oct. 2024].

Goldberg, D. W., J. P. Wilson, and C. A. Knoblock. 2009. "Extracting Geographic Features from the Internet to Automatically Build Detailed Regional Gazetteers." *International Journal of Geographical Information Science* 23 (1): 93–128. https://doi.org/10.1080/13658810802577262

Gritta, M., M. T. Pilehvar, and N. Collier. 2018. "Which Melbourne? Augmenting Geocoding with Maps." *Association for Computational Linguistics*. https://doi.org/10.17863/CAM.27796.

Gritta, M., M. T. Pilehvar, N. Limsopatham, and N. Collier. 2018. "What's Missing in Geographical Parsing?" *Language Resources and Evaluation* 52:603–623. https://doi.org/10.1007/s10579-017-9385-8

Havas, C., L. Wendlinger, J. Stier, S. Julka, V. Krieger, C. Ferner, A. Petutschnig, M. Granitzer, S. Wegenkittl, and B. Resch. 2021. "Spatio-temporal Machine Learning Analysis of Social Media Data and Refugee Movement Statistics." *ISPRS International Journal of Geo-Information* 10 (8): 498. https://doi.org/10.3390/ijgi10080498

Hoang, T. B. N., and J. Mothe. 2018. "Location Extraction from Tweets." *Information Processing & Management* 54 (2): 129–144. https://doi.org/10.1016/j.ipm.2017.11.001

Honnibal, M., and I. Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. Available at: https://spacy.io [Accessed 16 Oct. 2024].

Hu, Y., G. Mai, C. Cundy, K. Choi, N. Lao, W. Liu, G. Lakhanpal, R. Z. Zhou, and K. 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. https://doi.org/10.1080/13658816.2023.2266495

Hu, Y., and J. Wang. 2020. "How Do People Describe Locations during a Natural Disaster: An Analysis of Tweets from Hurricane Harvey." *11th International Conference on Geographic Information Science (GIScience 2021) - Part I. Leibniz International Proceedings in Informatics (LIPIcs)* 177: 6:1-6:16. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.GIScience.2021.I.6

Hu, X., Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, and F. Klan. 2023. "Location Reference Recognition from Texts: A Survey and Comparison." *ACM Computing Surveys* 56 (5): 1–37. https://doi.org/10.1145/3625819

Hu, X., Z. Zhou, Y. Sun, J. Kersten, F. Klan, H. Fan, and M. Wiegmann. 2022. "GazPNE2: A General Place Name Extractor for Microblogs Fusing Gazetteers and Pretrained Transformer Models." *IEEE Internet of Things Journal* 9 (17): 16259–16271. https://doi.org/10.1109/JIOT.2022.3150967

ICRC. 2024. International Committee of the Red Cross. [online] International Committee of the Red Cross. Available at: https://www.icrc.org/en [Accessed 16 Oct. 2024].

Jégou, H., M. Douze, and J. Johnson. 2017. FAISS: A Library for Efficient Similarity Search. [online] Facebook Engineering. Available at: https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/ [Accessed 2 Apr. 2025].

Ji, Y., and S. Gao. 2023. "Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations." *12th International Conference on Geographic Information Science (GIScience 2023). Leibniz International Proceedings in Informatics (LIPIcs)* 277: 43:1-43:6, Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.GIScience.2023.43.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. D. L. Casas, F. Bressand, et al. 2023. "Mistral 7B." *arXiv Preprint ArXiv:2310.06825*.

Jin, H., X. Han, J. Yang, Z. Jiang, Z. Liu, C. Y. Chang, H. Chen, and X. Hu. 2024. "LLM Maybe Longlm: Self-extend LLM Context Window without Tuning." *arXiv Preprint ArXiv:2401.01325*.

Johnson, J., M. Douze, and H. Jégou. 2019. "Billion-scale Similarity Search with GPUs." *IEEE Transactions on Big Data* 7 (3): 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Karimzadeh, M., S. Pezanowski, A. M. MacEachren, and J. O. Wallgrün. 2019. "GeoTxt: A Scalable Geoparsing System for Unstructured Text Geolocation." *Transactions in GIS* 23 (1): 118–136. https://doi.org/10.1111/tgis.12510

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. "Large Language Models Are Zero-Shot Reasoners." *Advances in Neural Information Processing Systems* 35:22199–22213.

Kuai, X., R. Guo, Z. Zhang, B. He, Z. Zhao, and H. Guo. 2020. "Spatial Context-Based Local Toponym Extraction and Chinese Textual Address Segmentation from Urban POI Data." *ISPRS International Journal of Geo-Information* 9 (3): 147. https://doi.org/10.3390/ijgi9030147

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, et al. 2020. "Retrieval-augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33:9459–9474.

Li, Z., H. Ning, S. Gao, K. Janowicz, W. Li, S. T. Arundel, C. Yang, et al. 2025. "GIScience in the Era of Artificial Intelligence: A Research Agenda towards Autonomous GIS." *arXiv Preprint ArXiv:2503.23633*.

Li, H., R. K. Srihari, C. Niu, and W. Li. 2003. "InfoXtract Location Normalization: A Hybrid Approach to Geographic References in Information Extraction." Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, 39–44.

Li, M., S. Zhang, Y. Liu, and K. Chen. 2024. "NeedleBench: Can LLMs Do Retrieval and Reasoning in 1 Million Context Window?" *arXiv Preprint ArXiv:2407.11963*.

Ling, C., X. Zhao, J. Lu, C. Deng, C. Zheng, J. Wang, T. Chowdhury, et al. 2023. "Domain Specialization as the key to Make Large Language Models Disruptive: A Comprehensive Survey." *arXiv Preprint ArXiv:2305.18703*.

Liu, Q., J. T. Harris, L. S. Chiu, D. Sun, P. R. Houser, M. Yu, D. Q. Duffy, M. M. Little, and C. Yang. 2021. "Spatiotemporal Impacts of COVID-19 on air Pollution in California, USA." *Science of the Total Environment* 750:141592. https://doi.org/10.1016/j.scitotenv.2020.141592

Liu, L., F. Wang, X. Fu, T. Kötter, K. Sturm, W. W. Guan, and S. Bao. 2024. "Elevating the rre Framework for Geospatial Analysis with Visual Programming Platforms: An Exploration with Geospatial Analytics Extension for Knime." *International Journal of Applied Earth Observation and Geoinformation* 130:103948. https://doi.org/10.1016/j.jag.2024.103948

Lopez, B. E., N. R. Magliocca, and A. T. Crooks. 2019. "Challenges and Opportunities of Social Media Data for Socio-Environmental Systems Research." *Land* 8 (7): 107. https://doi.org/10.3390/land8070107

Luong, T., H. Pham, and C. D. Manning.2015. "Effective Approaches to Attention-Based Neural Machine Translation." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 1412–1421. Association for Computational Linguistics.

Malarvizhi, A. S., Q. Liu, T. S. Trefonides, S. Hasheminassab, J. Smith, T. Huang, K. M. Marlis, et al. 2023. "The Spatial Dynamics of Ukraine air Quality Impacted by the war and Pandemic." *International Journal of Digital Earth* 16 (1): 3680–3705. https://doi.org/10.1080/17538947.2023.2239762

Manvi, R., S. Khanna, G. Mai, M. Burke, D. Lobell, and S. Ermon. 2023. "Geollm: Extracting Geospatial Knowledge from Large Language Models." *arXiv Preprint ArXiv:2310.06213*.

Middleton, S. E., G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. 2018. "Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging." *ACM Transactions on Information Systems (TOIS* 36 (4): 1–27. https://doi.org/10.1145/3202662

Ning, H., Z. Li, T. Akinboyewa, and M. N. Lessani. 2025. "An Autonomous GIS Agent Framework for Geospatial Data Retrieval." *International Journal of Digital Earth* 18 (1): 2458688. https://doi.org/10.1080/17538947.2025.2458688

Peng, C., X. I. Yang, K. E. Smith, Z. Yu, A. Chen, J. Bian, and Y. Wu. 2024. "Model Tuning or Prompt Tuning? A Study of Large Language Models for Clinical Concept and Relation Extraction." *Journal of Biomedical Informatics* 153:104630. https://doi.org/10.1016/j.jbi.2024.104630

Peng, B., Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang. 2024. "Graph Retrieval-Augmented Generation: A Survey." *arXiv Preprint ArXiv:2408.08921*.

Purves, R. S., P. Clough, C. B. Jones, M. H. Hall, and V. Murdock. 2018. "Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text." *Foundations and Trends® in Information Retrieval* 12 (2-3): 164–318. https://doi.org/10.1561/1500000034

Radford, A. 2018. Improving Language Understanding by Generative Pre-Training.

Radio Dabanga. 2024. Home | Radio Dabanga. [online] Available at: https://www.dabangasudan.org/en [Accessed 16 Oct. 2024].

Renze, M.. 2024. "The Effect of Sampling Temperature on Problem Solving in Large Language Models." *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA, 7346–7356. Association for Computational Linguistics.

Sahoo, P., A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. 2024. "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications." *arXiv Preprint ArXiv:2402.07927*.

Shi, G., and K. Barker. 2011. "Extraction of Geospatial Information on the Web for GIS Applications." In *IEEE 10th International Conference on Cognitive Informatics and Cognitive Computing (ICCI-CC'11)*, Banff, AB, Canada, 41–48. IEEE. https://doi.org/10.1109/COGINF.2011.6016120.

Shlyk, D., T. Groza, M. Mesiti, S. Montanelli, and E. Cavalleri. August, 2024. "REAL: A Retrieval-Augmented Entity Linking Approach for Biomedical Concept Recognition." Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, 380–389.

Small, S. G., and L. Medsker. 2014. "Review of Information Extraction Technologies and Applications." *Neural Computing and Applications* 25 (3-4): 533–548. https://doi.org/10.1007/s00521-013-1516-6

The Stanford NLP Group. 2023. https://nlp.stanford.edu/software/CRF-NER.shtml.

Stefanidis, A., A. Crooks, and J. Radzikowski. 2013. "Harvesting Ambient Geospatial Information from Social Media Feeds." *GeoJournal* 78 (2): 319–338. https://doi.org/10.1007/s10708-011-9438-2

Strötgen, J., M. Gertz, and P. Popov. 2010. "February. Extraction and Exploration of Spatio-Temporal Information in Documents." Proceedings of the 6th Workshop on Geographic Information Retrieval, 1–8.

Sudan Tribune. 2024. Sudan Tribune: Plural news and views on Sudan. [online] Available at: https://sudantribune.com [Accessed 16 Oct. 2024].

Sudan War Monitor. 2024. Sudan War Monitor | Substack. [online] Sudanwarmonitor.com. Available at: https://sudanwarmonitor.com/ [Accessed 16 Oct. 2024].

Tang, L., B. Bie, S. E. Park, and D. Zhi. 2018. "Social Media and Outbreaks of Emerging Infectious Diseases: A Systematic Review of Literature." *American Journal of Infection Control* 46 (9): 962–972. https://doi.org/10.1016/j.ajic.2018.02.010

Team, G., M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, et al. 2024. "Gemma 2: Improving Open Language Models at a Practical Size." *arXiv Preprint ArXiv:2408.00118*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin.2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30.

Wang, Z., Y. Chen, Y. Li, D. Kakkar, W. Guan, W. Ji, J. Cain, et al. 2022. "Public Opinions on COVID-19 Vaccines – a Spatiotemporal Perspective on Races and Topics Using a Bayesian-Based Method." *Vaccines* 10 (9): 1486. https://doi.org/10.3390/vaccines10091486

Wang, J., Y. Hu, and K. Joseph. 2020. "NeuroTPR: A Neuro-net Toponym Recognition Model for Extracting Locations from Social Media Messages." *Transactions in GIS* 24 (3): 719–735. https://doi.org/10.1111/tgis.12627

Wang, Z., Y. Li, K. Wang, J. Cain, M. Salami, D. Q. Duffy, M. M. Little, and C. Yang. 2023. "Adopting GPU Computing to Support DL-Based Earth Science Applications." *International Journal of Digital Earth* 16 (1): 2660–2680. https://doi.org/10.1080/17538947.2023.2233488

Wang, S., X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, and C. Guo. 2025. "GPT-NER: Named Entity Recognition via Large Language Models." *Findings of the Association for Computational Linguistics: NAACL 2025*, 4257–4275. Albuquerque, New Mexico. Association for Computational Linguistics.

Wu, J., Z. Zhang, Y. Xia, X. Li, Z. Xia, A. Chang, T. Yu, et al. 2024. "Visual Prompting in Multimodal Large Language Models: A Survey." *arXiv Preprint ArXiv:2409.15310*.

Xia, Y., Y. Huang, Q. Qiu, X. Zhang, L. Miao, and Y. Chen. 2024. "A Question and Answering Service of Typhoon Disasters Based on the T5 Large Language Model." *ISPRS International Journal of Geo-Information* 13 (5): 165. https://doi.org/10.3390/ijgi13050165

Xiong, G., Q. Jin, Z. Lu, and A. Zhang. 2024. "Benchmarking Retrieval-Augmented Generation for Medicine." *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 6233–6251. Association for Computational Linguistics.

Yan, Y., H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, and Y. Liang. 2023. "When Urban Region Profiling Meets Large Language Models." *arXiv Preprint ArXiv:2310.18340*.

Yu, C., J. James, D. David, H. Chow, and P. Poh. 2024. *Can LLMs Have a Fever?*. Investigating the Effects of Temperature on LLM Security. https://doi.org/10.46254/SA05.20240024.

Yu, M., M. Bambacus, G. Cervone, K. Clarke, D. Duffy, Q. Huang, J. Li, et al. 2020. "Spatiotemporal Event Detection: A Review." *International Journal of Digital Earth* 13 (12): 1339–1365. https://doi.org/10.1080/17538947.2020.1738569

Yu, M., Q. Huang, H. Qin, C. Scheele, and C. Yang. 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. https://doi.org/10.1080/17538947.2019.1574316.

Zhang, B., Z. Liu, C. Cherry, and O. Firat. 2024. "When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method." *arXiv Preprint ArXiv:2402.17193*.