# From Text to Maps: LLM-Driven Extraction and Geotagging of Epidemiological Data

**Karlyn K Harrod***
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
harrodkk@ornl.gov

**Prabin Bhandari*†**
George Mason University
Fairfax, Virginia, USA
pbhanda2@gmu.edu

**Antonios Anastasopoulos**
George Mason University
Fairfax, Virginia, USA
antonis@gmu.edu

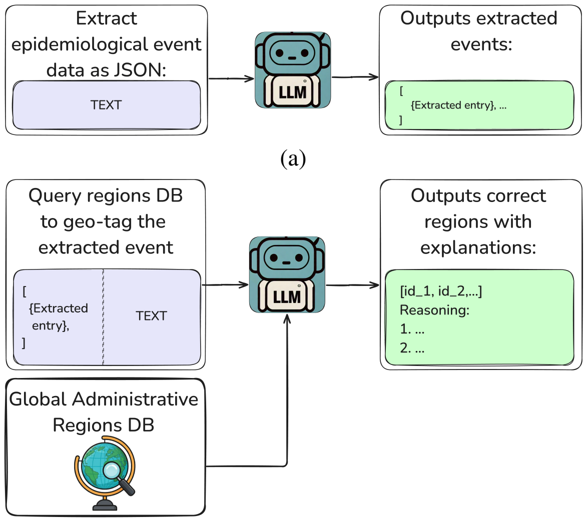
## Abstract

Epidemiological datasets are essential for public health analysis and decision-making, yet they remain scarce and often difficult to compile due to inconsistent data formats, language barriers, and evolving political boundaries. Traditional methods of creating such datasets involve extensive manual effort and are prone to errors in accurate location extraction. To address these challenges, we propose utilizing large language models (LLMs) to automate the extraction and geotagging of epidemiological data from textual documents. Our approach significantly reduces the manual effort required, limiting human intervention to validating a subset of records against text snippets and verifying the geotagging reasoning, as opposed to reviewing multiple entire documents manually to extract, clean, and geotag. Additionally, the LLMs identify information often overlooked by human annotators, further enhancing the dataset's completeness. Our findings demonstrate that LLMs can be effectively used to semi-automate the extraction and geotagging of epidemiological data, offering several key advantages: (1) comprehensive information extraction with minimal risk of missing critical details; (2) minimal human intervention; (3) higher-resolution data with more precise geotagging; and (4) significantly reduced resource demands compared to traditional methods.

## 1 Introduction

Epidemiology, the study of disease prevalence, comes from the Greek word "epidemios", meaning "among the people, of one's countrymen at home" (Harper, 2001). Each country documents the diseases within its borders, but they do so in their own ways. Analyzing epidemiological reports at a global scale thus becomes a challenging task due to the large number of heterogeneous reports.

*Both authors contributed equally.
† Work done at Oak Ridge National Laboratory.




(a)

(b)

Figure 1: Overview of our two-step methodology for extracting and geotagging epidemiological data: (a) first, an LLM extracts data from a piece of text based on the instructions provided, and (b) second, the LLM, with access to a global administrative regions database, geotags each extracted data, providing reasoning steps for its selections.

Even so, researchers currently read through epidemiological reports to extract the valuable data reported within. Epidemiological data refers to data collected on the occurrence of diseases and is used to understand the distributions, trends, and dynamics of disease through analyzing historical events and training models to understand drivers behind various disease outbreaks. Such data is crucial for public health analysis, policy development, and decision-making, as it helps identify risk factors for disease and targets for preventive healthcare. However, compiling epidemiological datasets poses significant challenges due to varying factors. There are numerous legal, technical, political, and cultural barriers, many of which are beyond our control, for efficient epidemiological data sharing and utilization (Fairchild et al., 2018; Pisani and AbouZahr, 2010). Furthermore, epidemiological data often exists in different formats, frequently



*Proceedings of the Third Workshop on NLP for Positive Impact*, pages 258–270
November 15, 2024 ©2024 Association for Computational Linguistics

embedded within textual reports. The dynamic nature of political boundaries further complicates data collection and makes geotagging these records challenging. Additionally, the potential use of different languages by different countries in their reporting adds another layer of complexity. Traditional ways of compiling such datasets have relied mostly on human effort, involving manual reading of source documents, data extraction, and subsequent post-processing. This standard method suffers from multiple issues, including the potential for human error and difficulty in correctly geo-tagging such datasets. To address these challenges, we propose utilizing large language models (LLMs) to automate the extraction and geo-tagging of epidemiological data at scale.

*By leveraging the capabilities of LLMs in event extraction and geospatial reasoning, we propose using LLMs to extract epidemiological data from text and geotag this information accurately.* The process is a two-step approach: (i) extraction of epidemiological data from text and (ii) geotagging this data using contextual information. First, we employ a hand-crafted prompt to instruct the LLM to output the epidemiological data embedded within a given text in a structured format, such as JSON. This can be considered as structured information extraction from unstructured sources. Second, the LLM utilizes a global administrative regions database to geo-tag each extracted information. With access to a global administrative region database, we then prompt the LLM with another prompt, incorporating the extracted data, and the context from which it was extracted. The goal is to identify the correct administrative regions from the ones in the database, accurately geotagging the extracted data.

Our findings reveal that LLMs are highly effective at extracting structured information from textual documents ($Recall = 100\%$). While LLMs generate more data than present in the human-curated dataset ($Precision = 20\%$), some of this additional information may be incorrect. However, even when considering only the correctly generated entries, the LLM-generated dataset is three times larger than the human-curated one while capturing all relevant entries present in the human dataset. This highlights the significant advantage of leveraging LLMs for enhanced data coverage. Further, LLMs exhibit impressive geospatial reasoning capabilities, accurately geotagging data points through logical inference. Overall, our results highlight the strong performance of LLMs in both extraction and geotagging tasks, indicating that these models could semi-automate these processes, with minimal human validation needed.

## 2 Related Work

Modern LLMs have transformed the field of natural language processing and artificial intelligence by eliminating the need for task-specific models trained using vast amounts of human-annotated datasets. Through pre-training techniques, LLMs can be pre-trained on large textual corpora, enabling them to encode various types of knowledge within their parameters and potentially even function as knowledge bases (Petroni et al., 2019). LLMs encode world knowledge and exhibit common sense reasoning capabilities, enabling them to understand and generate human-like text across diverse contexts. Demonstrating this capability, Brown et al. (2020) showed that sufficiently scaled LLMs like GPT-3 can handle diverse downstream tasks just by receiving a task description, with or without a few sample task examples, as context, a technique known as "prompting". Recent advancements in prompting techniques[1] have further enhanced the ability of LLMs to handle complex tasks, including those requiring intricate reasoning. Additionally, Researchers (Bhandari et al., 2023; Roberts et al., 2024; Mooney et al., 2023) have shown that LLMs possess encoded geospatial knowledge, making them geospatially aware and capable of reasoning with geospatial data during text generation.

Instruction tuning (IT) is another emerging technique where LLMs are further trained on datasets containing instructions and desired output in a supervised manner. Instruction tuning aligns the next-word prediction objective of LLMs with user objectives, enabling the creation of general-purpose chatbots like ChatGPT[2] and Gemini.[3] These instruction-tuned LLMs excel at following human instructions and have shown impressive performance in several downstream tasks, such as event extraction (Wei et al., 2022).

Information extraction (IE) using LLMs for event extraction has gained significant research attention, primarily due to the excellent instruction-following capabilities of instruction-tuned LLMs.

[1] See Bhandari (2024) for a survey.
[2] https://chatgpt.com/
[3] https://gemini.google.com/app

Recent advancements in this field have shown varying degrees of success.

Wei et al. (2024) introduced ChatIE, a framework that transforms the zero-shot IE task into a multi-turn question-answering problem suited for LLMs. The authors evaluated their framework on three IE tasks: entity-relation triple extraction, named entity recognition, and event extraction. Their results show that ChatIE achieves impressive performance, even surpassing some multi-shot models on several datasets. Similarly, others have achieved success using LLMs for specific IE tasks, with various modifications to enhance performance (Peng et al., 2023; Vijayan, 2023; Li et al., 2024). However, some researchers have found notable challenges in using LLMs for event extraction. For instance, Gao et al. (2023) found that ChatGPT's performance was only half that of a task-specific model for long-tail and complex scenarios. Our research presents a different approach compared to the existing studies. While most research focuses on extracting singular events described in a text, our approach aims to extract multiple events from a single text using LLMs' comprehensive understanding capabilities. Additionally, rather than solely relying on absolute performance metrics, we also measure success in terms of reduction in human effort for creating epidemiological datasets. Furthermore, our approach involves geotagging extracted data using LLMs, a novel concept that enhances the quality of the epidemiological dataset.

## 3 Methodology

The extraction and geotagging of epidemiological data involve a two-step process, as outlined in Figure 1. In the first step, data is extracted from small sections of text, which are then processed and merged to form the final database. In the second step, each extracted record, along with its contextual information, is passed to an LLM with access to a global administrative region database. The LLM is tasked with selecting the correct entries from this database to represent the record and provide reasoning for its choices, facilitating human validation.

### 3.1 Extraction of epidemiological data

Given a collection of textual documents $D = \{d_1, d_2, ..., d_n\}$ containing epidemiological data, each document $d_i$ contains various sections $d_i = \{s_1, s_2, ..., s_m\}$. We use a prompt template $T_1$ to guide the LLM in extracting the required data and generating output in JSON format for each section $s_j$ of the documents: $LLM(T_1, s_j) = [\{data\}]$.

The template $T_1$ can be customized based on the targeted disease and the attributes of interests. Figure A1 in Appendix A is an example of $T_1$, which guides the LLM to extract global epidemiological data on Rift Valley Fever from different journal articles and reports, outputting the result in JSON format. This template is employed in our experiments (§4). The outputs are then post-processed to merge records and eliminate duplicates, resulting in the final database $DB$, which contains the epidemiological data from documents $D$.

### 3.2 Geotagging of extracted data

Geotagging, the process of adding geographical identification metadata to the extracted epidemiological data, occurs after forming the database $DB = \{db\_1, db\_2, ..., db\_k\}$. This database contains epidemiological records in a structured format like JSON alongside the text section $s$ from which they are extracted. Here $db\_k$ refers to the $k^{th}$ record generated by LLM.

To perform geotagging, we use a database of global administrative regions, which we will refer to by $O$. This database contains the administrative regions for all the countries at various administrative levels. We use the GADM database (GADM, 2018) for our approach.

We employ a second prompt template $T_2$ to guide the LLM in selecting the appropriate entries from the global administrative regions database to geotag extracted data and generate output in JSON format, including the reasoning steps behind each decision: $LLM(T_2, db_k, s_j, O) = [\{db_k^+, R_1\}]$. Figure A2 in Appendix A illustrates an instance of $T_2$, which directs the LLM to select correct entries from the GADM database to geotag RVF occurrence data and provide reasoning steps similar to chain-of-thoughts prompting (Wei et al., 2023). Eliciting reasoning responses offers dual benefits: it enhances performance and provides reasoning steps that humans can easily validate to assess the efficacy of geotagging using LLMs.

## 4 Experiments

Our experiments aim to evaluate the accuracy and viability of using LLMs to extract and geotag epidemiological data. We focus on the global spread of Rift Valley Fever (RVF) by extracting relevant information from a collection of documents using an LLM to create a database of RVF outbreaks. These documents are sourced from a human-created RVF outbreak dataset. Below, we first introduce the RVF dataset and then outline the experimental setup, post-processing steps, and evaluation metrics for our two experiments: extracting RVF data and geotagging the extracted data.

## 4.1 Dataset

Bron et al. (2021) compiled a comprehensive dataset on the spread of RVF in humans and animals, covering 22 countries for humans and 37 countries for animals from 1931 to 2020. The dataset also includes seroprevalence studies conducted between 1950 and 2020 ($n = 228$). Each data point in the dataset is linked to either a single or multiple sources, such as other datasets and research publications, from which it was collected.

For our study, we collected all source documents that were accessible to us, and created a subset of the original dataset based on the sources we were able to collect. Note that, in some cases, data points with multiple sources might not have all the required attributes available within the documents we were able to collect, as some information might have been derived from documents we could not access. To address this, we manually inspected each data point and its sources, removing any entries with such discrepancies. Additionally, we excluded seroprevalence and animal data, resulting in a dataset focused on RVF outbreaks in humans. This dataset is accompanied by the corresponding source documents, which include research publications and outbreak reports in portable document format (PDF) and span from 1955 to 2018.

## 4.2 Extraction of RVF data

The goal of this experiment is to extract RVF outbreak data from the accompanying documents to evaluate the capability of LLMs to aid in epidemiological data extraction. We detail our experimental setup below, followed by a description of the post-processing steps used to finalize the datasets and the evaluation metrics employed to compare our results with the human-curated dataset.

### 4.2.1 Experimental setup

The documents in our dataset are in PDF format, but the LLMS require plain text input. To achieve this, we first extract textual data from research articles using optical character recognition (OCR). Specifically, we use `paperetl` (NeuML, 2020) for text extraction, which leverages GROBID (Lopez, 2009) to perform this task. GROBID is a machine-learning library designed to extract, parse, and convert raw documents into structured formats with a primary focus on technical and scientific publications. The extracted text is grouped into different sections. We overlap sections by including two preceding and two succeeding sections to ensure no information is missed, even though this approach increases the likelihood of generating the same information multiple times.

Next, we pass these sections, along with our handcrafted prompt, to the LLM to extract the required information as a JSON. We use prompt templates as shown in Fig. A1 in Appendix A, to extract human cases of RVF from the documents. We extract the location, country, start date, end date, number of cases, and number of deaths. Each section is processed by the LLM five times to enhance the robustness of the extraction.

For this experiment, we use Llama-3.1 (META AI, 2024) as the LLM of choice, specifically employing the instruction-tuned 8– and 70–billion parameter variants. We use a top-$p$ sampling-based decoding strategy with $p$ set to $0.9$ and a temperature of $0.3$. Top-$p$ sampling limits the token pool while decoding to the most probable tokens whose cumulative probability mass is greater than or equal to $p$, while temperature controls the randomness during token selection. A higher temperature value increases randomness, while a lower temperature value reduces randomness. The experiments were run on our in-house compute cluster of Nvidia A100 80 GB GPUs, with a total GPU hours of around 800 Hours.

### 4.2.2 Post-processing

The generated output undergoes a comprehensive post-processing to extract and refine the epidemiological records.

First, we extract JSON data from the generated output text using string matching and regular expressions. Any output text that does not yield a valid JSON structure is discarded. Next, we filter out records lacking essential information, specifically those missing location data, or missing all of the start and end dates, number of cases, and deaths. We then parse the essential attributes of the JSON: disease start date, disease end date, number of cases, and number of deaths. We use `dateutil` (DateUtil, 2014) for parsing date-related attributes and `num` from string (DoubleBite, 2019) for parsing number-related attributes. Entries from which these details cannot be accurately extracted are discarded. Following this, we merge identical entries to eliminate duplicates. For merged entries, the country name is resolved as the one with the highest frequency. This country name is then used to query the GeoNames (GeoNames, 2024) API, obtaining the accurate name and code.

Subsequently, we merge entries from the five different runs of a document. We then attribute each record by verifying its presence in the text, checking for the presence of case counts, death counts, and start or end date year in the text using string matching. We disregard records that are not attributed.

#### 4.2.3 Evaluation

The RVF data extraction experiment is evaluated using both automatic metrics and human inspection.

$$\text{precision} = \frac{\text{Total No. of correctly extracted events}}{\text{Total No. of extracted events}} \quad (1)$$

$$\text{recall} = \frac{\text{Unique No. of correctly extracted events}}{\text{Total No. of relevant events in the text}} \quad (2)$$

First, we measure precision (Equation. 1) and recall (Equation. 2). While precision focuses on accuracy, any additional information extracted by the LLM not present in the human dataset may still hold significance, as it could represent overlooked data. In these equations, **Total No. of extracted events** refers to the number of events generated by LLM, and the **Total No. of relevant events in the text** refers to the number of events in the human-annotated dataset. The **No. of correctly extracted events** refers to the records present in both the human-annotated and LLM-generated datasets. This is calculated as the number of identical events in the two datasets. Precision uses the total count of this measure whereas recall uses the unique count, due to the possibility of duplicate records in the LLM-generated dataset. Two events in the LLM-generated and human-annotated dataset are considered identical if they meet all of the following criteria: (i) originate from the same source document, (ii) have the same case counts, (iii) share the same year in either the start or end date and (iv) have the same country name.

We also perform human evaluation of the LLM-generated RVF spread dataset, to assess the records generated by LLM, focusing on records not identical to the human-annotated ones. Evaluators are tasked with determining whether the extracted data are correct or incorrect based on the context from which they were extracted. If a record is deemed correct, it represents data missed by human annotators but successfully captured by the LLM. Conversely, if a record is identified as incorrect, the evaluator will provide an explanation of the error, facilitating future improvements in data extraction using LLM. Additionally, human evaluators are responsible for accurately merging any remaining duplicate records to create the final dataset.

### 4.3 Geo-tagging of extracted RVF data

The goal of this experiment is to geotag the extracted RVF data using an LLM with access to a global administrative regions database. We outline our experimental setup below, followed by the post-processing steps and the evaluation methodology used.

#### 4.3.1 Experimental setup

The RVF spread dataset was generated by an LLM extracting the required information from relevant documents in the above experiment. To enhance the utility of this dataset, we aim to geotag each data point in the dataset. Each data point includes attributes for country and location. We use this information and the text from which the data point was extracted as input to an LLM. The input also includes the GADM table for the data point's country. The LLM's task is to infer the correct GADM IDs for the data points and provide reasoning for selecting these IDs. To accomplish this, we employ a chain-of-thought prompting technique, as shown in Figure A2 in Appendix A, to infer the GADM IDs and associated reasoning from an LLM. For this purpose, we employ the Gemini (Gemini Team, 2024) model, specifically the Gemini-1.5 flash version, accessible via an Application Programming Interface (API). We chose the Gemini model over Llama-3.1 due to the longer input sequence required for the geotagging task. The longer input sequence constraints us from running Llama-3.1 on our in-house GPU clusters. Additionally, Gemini provides free requests, and by using the lighter flash version instead of the pro version, we were able to run the geotagging experiments without incurring extra computational costs

#### 4.3.2 Post-processing

The generated output includes GADM IDs in JSON format, which we extract and append to their corresponding entries to create the final geo-tagged RVF spread database. This straightforward post-processing step ensures the seamless integration of geospatial metadata into the dataset.

#### 4.3.3 Evaluation

Since we lack a reference gold database for this experiment, our evaluation relies exclusively on human assessments. Human evaluators assess the reasoning steps generated by the LLM. They verify the soundness of these steps, ensuring the accuracy of the geo-tagged RVF spread dataset. This evaluation not only verifies the correctness of the geotagged data but also validates the overall efficacy of our approach to geotag epidemiological datasets using LLMs.

## 5 Results

Our experimental results indicate that while LLMs can extract significant amounts of information overlooked by human annotators, they also produce some inaccuracies. Additionally, LLMs equipped with relevant contextual data show promise in effective geo-tagging. In the following sections, we first present our empirical findings, followed by insights from human evaluations. We conclude by discussing the implications of these results and their potential impact on the future of epidemiological data extraction and geo-tagging using LLMs.

### 5.1 Automatic Evaluations

The human-curated RVF dataset is our gold standard reference for empirical evaluations. As detailed in §4.1, the dataset has been refined to include only the subset of data points related to human RVF outbreaks that are available in the documents that we can access. As described in §4.2.3, we evaluate the performance of the LLM-generated RVF dataset against this human-curated dataset using Precision and Recall, as presented in Table 1.

In Table 1, the **k** column represents the threshold for the number of times a record must be generated across five runs to be included in the final dataset. Notably, the results show that we can achieve perfect recall by including entries generated at least once for the 70B model and at least twice for the 8B model. This indicates that our strategy of passing each section through the LLM multiple times was effective.

| Model | No. of Params | k | Precision (%) | Recall (%) |
|---|---|---|---|---|
| Llama-3.1 | 8B | 1 | 09.02 | 100.0 |
| | | 2 | 11.44 | 100.0 |
| | | 3 | 12.31 | 91.67 |
| | 70B | 1 | 19.82 | 100.0 |
| | | 2 | 18.48 | 91.67 |

Table 1: LLM can extract all the events contained in the human curated dataset (recall = 100%) but also generates additional events as shown by low precision, **k** column represents the threshold for the number of times a record must be generated across five runs to be included in the final dataset

The perfect recall demonstrates that the LLM successfully extracted all relevant information in the human-curated dataset. However, the maximum precision achieved is only around 20 %, indicating that LLM generated additional records beyond what is captured within the human-curated dataset. While this may initially seem like an issue, it suggests that the LLM could be identifying information that human annotators may have overlooked. The full significance of these results will become clearer after human evaluations of LLM-generated outputs, which are discussed in the next section.

### 5.2 Human evaluations

We present the human evaluations of the extraction and geotagging results. The geotagging was done on the extraction dataset, which has been refined through human evaluation by removing the incorrect entries and consolidating duplicates. We use the extraction dataset generated by the Llama-3.1 70B model, with a threshold of 1 for the number of times a record must be generated across five runs to be included in the final dataset. The human evaluations conducted by the authors.

Human evaluations reveal that only 45% of the records generated by the LLM are accurate, while the remaining 55% contain errors for various reasons. Of the incorrect entries, 40% involve details of individual cases discussed within specific sections of the documents. Although this is not an error in the LLM's extraction, it indicates a need for future experiments to refine instructions to handle such cases more effectively. Another 10% of the errors stem from the LLM including suspected cases, despite the prompt specifying actual counts. Additionally, approximately 5% of the errors were due to OCR limitations, such as difficulties in correctly extracting tables and mistakenly including header or footer text within the main body. The remaining errors were due to inaccuracies introduced by the LLM itself.

Despite these issues, notice that while the highest precision computed against the "gold" annotations is only around 20%, there are an additional 25 automatically produced records that are deemed accurate! This means that **our final dataset captures 225% of the information that human annotators had previously missed (our data captured 45 records and human annotators captured 20**, meaning it introduced a significant amount of new data that was not initially identified.

Human evaluations were also conducted on the geotagging results. The outputs included the inferred GIDs for each location and the reasoning behind selecting those GIDs, as shown in Figure B1 in Appendix B, for the location of 'Aleg Hospital Center'. We closely examined the reasoning steps generated by the LLM and generally found them to be accurate, displaying impressive reasoning capabilities. In cases where insufficient information was available, instead of forcing an incorrect solution the LLM opted for broader, contextually appropriate responses, as illustrated in Figure B2 in Appendix B.

### 5.3 Discussion

Our empirical and human evaluation results demonstrate that LLMs can not only extract relevant information but also capture details overlooked by humans. Additionally, LLMs can also accurately geotag these extracted data points. This approach significantly reduces manual effort, requiring minimal human intervention limited to validation checks. Based on these findings, we discuss the advantages of using LLM-driven methods for similar tasks, highlighting why future researchers should consider such approaches over traditional, human-based efforts. We then address the ethical and societal considerations associated with our work. We conclude by highlighting the positive impacts of our approach.

One key advantage of our approach is its **ability to extract information comprehensively, minimizing the risk of missing critical details**. As our results show, LLMs can extract thrice as much information compared to what humans can. Human errors are also a concern. For example, our evaluation revealed that the human-curated dataset recorded 240 cases of RVF in South Africa in 2010,

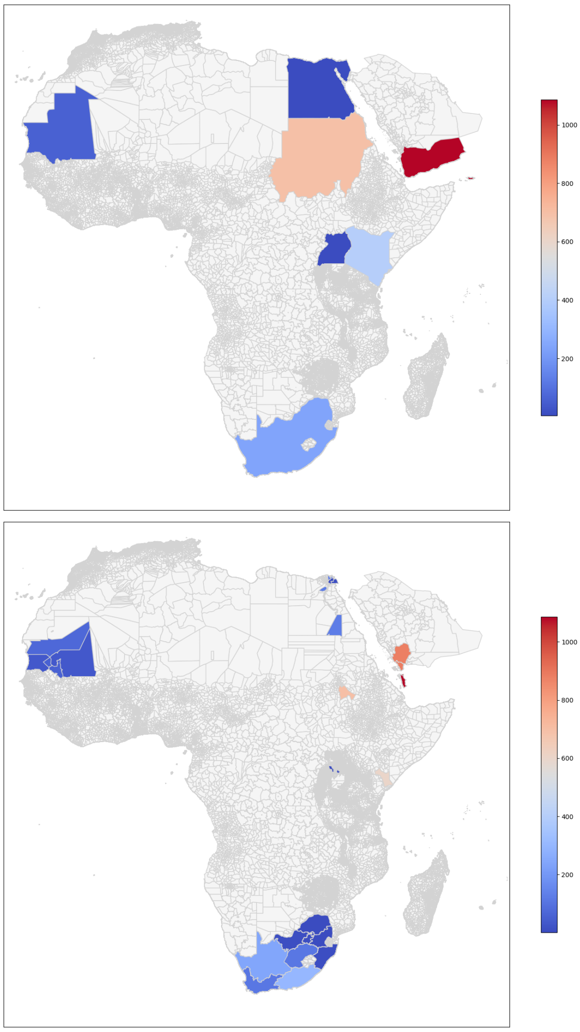


Figure 2: Comparison of heatmaps showing case counts extracted and geotagged by humans (top) and by the LLM (bottom) shows that LLM demonstrates the ability to extract more data and achieve higher geotagging precision than manual methods.

while the LLM-generated dataset correctly identified 241. Upon further investigation, we found that the original document mentioned 241 cases in total but a case was missing from the state-by-state breakdown table, which likely led the human annotators to capture only 240 cases. In contrast, our LLM-based approach captured the full 241 cases. This highlights the second advantage of our method: it **requires minimal human intervention**. Human effort is limited to a few validation checks, reducing the likelihood of errors and ensuring more accurate and comprehensive data extraction.

The third advantage lies in **the higher resolution of extracted data**. Our approach not only extracts information but also geotags it with greater precision, enabling locations to be pinpointed with finer granularity instead of just country-level data. This is demonstrated in Figure 2, where the top figure illustrates case counts extracted and geotagged by humans, and the bottom figure displays the results generated by the LLM. The human-curated map is restricted to country-level data, whereas the LLM-generated map offers detailed information at lower administrative levels. This enhanced granularity significantly benefits downstream applications like disease spread modeling.

Lastly, the fourth advantage is the **reduced resource demands** of our approach. Traditional methods heavily rely on human labor, limiting data collection due to high costs, significant time investment, and the challenges of recruiting and training personnel. In contrast, our method incurs minimal costs associated with LLM inference, which can be further reduced with access to in-house GPU resources. The need for human involvement is also minimal, focused primarily on validation checks. These relaxed resource requirements make our approach an efficient solution for extracting and geotagging epidemiological data.

The advantages of our approach, as discussed above, are significant. However, since it is applied to public health analysis and decision-making, it is crucial to consider the ethical implications. LLMs are known to exhibit inherent social biases [4], and they are prone to generate hallucinations-outputs that are inconsistent with real-world facts or user input [5]. These issues could compromise the accuracy and reliability of critical public health data that can be used in decision-making. Our data extraction process is less susceptible to these problems, as the LLM is instructed to generate responses based strictly on the provided text, with an automatic post-processing step to verify its output. However, biases and hallucinations may still arise, which is why human verification of the generated data, ensuring alignment with the source text, is integral to our method. The geo-tagging component of our system is particularly vulnerable to these challenges. To address this, we require the LLM to generate reasoning steps alongside the geo-tagging output. These reasoning steps are then reviewed by humans to identify and mitigate potential biases or hallucinations. Ultimately, human oversight is essential to maintain the integrity of the dataset produced by our system. Our next step, the geo-tagging step is mostly susceptible to these issues. To overcome these issues, we ask LLM to generate the reasoning steps along with the geo-tagging response. These reasoning steps are then verified by humans to check for issues of bias and hallucination. Overall, proper human verification of the dataset is crucial for the integrity of the generated dataset.

The use of LLMs for extracting and geotagging epidemiological data offers several positive impacts. Our approach can be applied to gather data on critical diseases like dengue fever, addressing gaps that hinder effective public health analysis and decision-making. The broader implications are profound, as health is foundational to all aspects of life, and our solution aims to enhance public health, ultimately contributing to the overall well-being of societies. However, it is important to consider the ethical implications of our work addressed above. Therefore, human oversight is necessary to ensure data accuracy and integrity. Ultimately, our work enables scalable, real-time epidemiological data collection and geo-tagging for enhanced disease tracking and policy-making.

## 6 Conclusion and Future Work

We propose an automated approach for extracting and geotagging epidemiological data from textual documents using Large Language Models (LLMs). Current methods of collecting such data rely on human effort and are prone to errors. To overcome these challenges, our approach leverages LLMs to automate the extraction and geotagging processes. We tested our method by applying it to RVF outbreak data, where a human-curated dataset was available for comparison. Our findings demonstrate the effectiveness of this approach. The LLM was able to capture significantly more information than manual efforts, albeit with some inaccuracies. Additionally, the LLM exhibited notable geospatial reasoning abilities, accurately geotagging data points. These results suggest that LLMs can be effectively employed to semi-automate the extraction and geotagging of epidemiological data with some level of human verification.

While our experiments focused on a single disease, the promising results encourage us to extend this work to multiple diseases with the aim of creating publicly available datasets for epidemiological studies. We also plan to extend our research to include multilingual LLMs, broadening the scope to non-English documents.

[4]See Gallegos et al. (2024) for a survey on bias in LLMs.
[5]See Huang et al. (2023) for a survey on hallucinates.

## Acknowledgments

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U. S. Department of Energy. Additionally, Prabin Bhandari was supported by an appointment to the Education Collaboration Program at Oak Ridge National Laboratory, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education. Antonios Anastasopoulos is also partially supported by the National Science Foundation under award #CNS-2234895. This project was supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Award Number 2018631).

## Limitations

Our study is fairly limited in scope. Firstly, we utilized the Llama-3.1 70-billion model instead of more advanced options like the 405-billion parameter model due to computational and memory constraints. Similarly, we employ the Gemini flash model rather than the more capable Pro model. These better LLMs could have possibly produced more intriguing and better results. Secondly, we focus solely on *English* language documents; a follow-up study could further expand to cover more languages. Lastly, the human evaluations were conducted solely by the authors due to time and resource constraints, which may impact the thoroughness of the evaluation process.

## References

Prabin Bhandari. 2024. A survey on prompting techniques in llms. *Preprint*, arXiv:2312.03740.

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Gebbiena M. Bron, Kathryn Strimbu, Hélène Cecilia, Anita Lerch, Sean M. Moore, Quan Tran, T. Alex Perkins, and Quirine A. ten Bosch. 2021. Over 100 years of rift valley fever: A patchwork of data on pathogen spread and spillover. *Pathogens*, 10(6).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

DateUtil. 2014. dateutil: powerful extensions to datetime. `https://github.com/dateutil/dateutil`.

DoubleBite. 2019. Numbers-from-string: Extract numbers from a string. `https://github.com/doubleBite/Numbers-from-String`.

Geoffrey Fairchild, Byron Tasseff, Hari Khalsa, Nicholas Generous, Ashlynn R Daughton, Nileena Velappan, Reid Priedhorsky, and Alina Deshpande. 2018. Epidemiological data challenges: planning for a more robust future through data standards. *Frontiers in Public Health*, 6:336.

GADM. 2018. Gadm maps and data. `https://www.gadm.org`.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, pages 1–83.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *Preprint*, arXiv:2303.03836.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

GeoNames. 2024. Geonames geographic database. `https://www.geonames.org`.

Douglas Harper. 2001. Etymology of epidemiology by etymonline — etymonline.com. `https://www.etymonline.com/word/epidemiology`.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Ni Li, Shorouq Zahra, Mariana Madruga de Brito, Clare Marie Flynn, Olof Görnerup, Koffi Worou, Murathan Kurfali, Chanjuan Meng, Wim Thiery, Jakob Zscheischler, et al. 2024. Using llms to build a database of climate extreme impacts. In *Natural Language Processing meets Climate Change@ ACL 2024*.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pages 473–474. Springer.

META AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. 2023. Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 85–94.

NeuML. 2020. paperetl: Etl processes for medical and scientific papers. `https://github.com/neuml/paperetl`.

Ruoling Peng, Kang Liu, Po Yang, Zhipeng Yuan, and Shunbao Li. 2023. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. *Preprint*, arXiv:2308.03107.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Elizabeth Pisani and Carla AbouZahr. 2010. Sharing health data: good intentions are not enough. *Bulletin of the World Health Organization*, 88:462–466.

Jonathan Roberts, Timo Lüddecke, Rehan Sheikh, Kai Han, and Samuel Albanie. 2024. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 554–563.

Aishwarya Vijayan. 2023. A prompt engineering approach for structured data extraction from unstructured text using conversational llms. In *Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, pages 183–189.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. Chatie: Zero-shot information extraction via chatting with chatgpt. *Preprint*, arXiv:2302.10205.

# A Prompt Templates

Extract the occurrences of Rift Valley Fever globally on humans only, including specific locations, outbreak dates, and epidemiological data. Adhere strictly to the provided text, ensuring accuracy and faithfulness. Extract the information as a JSON with the following structure:

[
{
‘$location$’: <$location\_name$>,
‘$country$’: <$country\_name$>,
‘$disease\_start\_date$’: $< YYYY/MM/DD >$,
‘$disease\_end\_date$’: $< YYYY/MM/DD >$,
‘$number\_of\_cases$’: $< INT >$,
‘$number\_of\_deaths$’: $< INT >$,
‘$location\_confidence\_score$’: $< FLOAT >$
‘$country\_confidence\_score$’: $< FLOAT >$
‘$disease\_start\_date\_confidence_score$’: $< FLOAT >$
‘$disease\_end\_date\_confidence\_score$’: $< FLOAT >$
‘$number\_of\_cases\_confidence\_score$’: $< FLOAT >$
‘$number\_of\_deaths\_confidence\_score$’: $< FLOAT >$
}
]

Please keep in mind the following things:

1. Only extract information regarding Rift Valley fever and not other diseases.
2. Extract information regarding the outbreak of Rift Valley fever in humans only. Disregard information related to animals and seroprevalence.
3. Mark unavailable information as ‘None’ in the JSON.
4. Remember to generate the JSON only and nothing else and if there is no mention of the disease in the text just provide an empty list.
5. Provide the case counts as the number of confirmed cases rather than the estimated or investigated ones
6. Provide the location name in as much detail as you can, meaning the lowest administrative region possible.
7. $attribute\_confidence\_score$ refers to the confidence you have in the accuracy of the data you extracted for the specific attribute. Its value ranges from 0 to 100.
8. Please be honest while assigning the $confidence\_score$: use lower values where you are not certain about the accuracy of the extracted information and higher values where you are confident of the information extracted from text.
9. Only provide the JSON output.

Below is the text to extract the information:

Figure A1: Prompt template to extract human RVFV epidemiological data.

The following information was extracted from a research article about rift valley fever outbreak:

{

'$start\_date$': <$start\_date$>,

'$end\_date$': <$end\_date$>,

'$number\_of\_cases$': <$number\_of\_cases$>,

'$number\_of\_deaths$': <$number\_of\_deaths$>,

'$country$': <$country$>,

'$location$': <$location$>,

}

This information was extracted from the following sections of the article:

**<context_txt>**

The extracted country is correct and can be mapped to GADM but the location, which contains list of location might not be directly mapped to GADM. I want you to infer the correct GADM entry for the location and you can use the sections from which the information was extracted to better make the decision. I will provide you with the GADM dataset as a table below for <$country\_name$> and you can pick the correct entry from there and provide me with the correct GADM ID (GID). The GADM table for <$country\_name$> is as below:

**<GADM TABLE>**

Please extract the correct GADM entry and think step by step to provide the correct one. I want to infer the GADM entry for <location> from the following specific entry:

{

'$start\_date$': <$start\_date$>,

'$end\_date$': <$end\_date$>,

'$number\_of\_cases$': <$number\_of\_cases$>,

'$number\_of\_deaths$': <$number\_of\_deaths$>,

'$country$': <$country$>,

'$location$': <$location$>,

}

If the location has multiple entries, target the location at the lowest level of administration and some location might be inferred as multiple entries and that is fine.

Also, be careful that the provided text might be talking about another outbreak and we are interested just in the outbreak of the specific entry.

If it is difficult to infer the location at a lower administrative level, instead of guessing please provide the upper location at an upper administrative level.

Also, be mindful of the naming conventions used in different countries. For example: some countries have state at the 1st level of administrative region while others might have province or regions.

I want just the GIDs, and you can provide them from the $GID\_1$, $GID\_2$, $GID\_3$, $GID\_4$ and $GID\_5$ columns of the table.

At the end please provide an output in the following JSON format:

{

'$inferred\_gids$': [<gid>],

}

Figure A2: Prompt template to infer GADM entry for a location

# B LLM outputs

“‘json
{
"$inferred\_gids$": ["$MRT$.3.1_1"]
}
”’
Here’s how we arrived at this GID:
1. **Focus on the provided entry:** We are specifically interested in the outbreak described in the entry:
“‘json
{
"$start\_date$": 2015-09-01’,
"$end\_date$": ’2015-11-07’,
"$number\_of\_cases$": ’12’,
"$number\_of\_deaths\_fixed$": ’nan’,
"$country$": ’Mauritania’,
"$location$": [’Aleg Hospital Center’],
}
”’
2. **Locate the relevant text:** The text mentions "During the outbreak, 19 (61%) and 12 (39%) confirmed cases were admitted at Kiffa Hospital Center and Aleg Hospital Center, respectively." This directly links "Aleg Hospital Center" to the outbreak we’re analyzing.
3. **Match to GADM:** The GADM table shows "Aleg" as a $NAME\_2$ under the $GID\_2$ "$MRT$.3.1_1".
4. **No lower level:** The text doesn’t provide information about a more specific location within Aleg. Therefore, "MRT.3.1_1" is the most precise GID we can infer.

Figure B1: LLM-generated response for the geo-tagging task of ‘Aleg Hospital Center’ demonstrates the model’s strong potential for accurately geo-tagging datasets, given sufficient contextual information to support its reasoning.

**Include all provinces:** Since we don’t have enough information to pinpoint the exact location within the three provinces, we include all five provinces as the inferred GIDs.

Figure B2: LLM demonstrated strong reasoning skills by providing a broad, contextually appropriate response rather than forcing an incorrect solution.