Fair Geolocation Extraction from Humanitarian Documents Using Large Language Models

Gianmarco Cafferata
gcafferata@udesa.edu.ar
Universidad de San Andrés
Victoria, Buenos Aires, Argentina

Tiziano Demarco
demarcot@udesa.edu.ar
Universidad de San Andrés
Victoria, Buenos Aires, Argentina

Kyriaki Kalimeri
kyriaki.kalimeri@isi.it
ISI Foundation
Turin, Italy

Yelena Mejova
yelena.mejova@isi.it
ISI Foundation
Turin, Italy

Mariano G. Beiró
mbeiro@udesa.edu.ar
Universidad de San Andrés
Victoria, Buenos Aires, Argentina
CONICET
Buenos Aires, Argentina

Abstract

Humanitarian crises demand timely and accurate geographic information to inform effective response efforts. Yet, automated systems that extract locations from text often reproduce existing geographic and socioeconomic biases, leading to uneven visibility of crisis-affected regions. This paper investigates whether Large Language Models (LLMs) can address these geographic disparities in extracting location information from humanitarian documents. We introduce a two-step framework that combines few-shot LLM-based named entity recognition with an agent-based geocoding module that leverages context to resolve ambiguous toponyms. We benchmark our approach against state-of-the-art pretrained and rule-based systems using both accuracy and fairness metrics across geographic and socioeconomic dimensions. Our evaluation uses an extended version of the HumSet dataset with refined literal toponym annotations. Results show that LLM-based methods substantially improve both the precision and fairness of geolocation extraction from humanitarian texts, particularly for underrepresented regions. By bridging advances in LLM reasoning with principles of responsible and inclusive AI, this work contributes to more equitable geospatial data systems for humanitarian response, advancing the goal of leaving no place behind in crisis analytics.

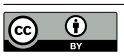
CCS Concepts

• **Information systems** → **Information extraction**; • **Computing methodologies** → **Information extraction**.

Keywords

geolocation, humanitarian crisis, large language models, named entity recognition

ACM Reference Format:
Gianmarco Cafferata, Tiziano Demarco, Kyriaki Kalimeri, Yelena Mejova, and Mariano G. Beiró. 2026. Fair Geolocation Extraction from Humanitarian Documents Using Large Language Models. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774905.3795457



This work is licensed under a Creative Commons Attribution 4.0 International License.
WWW Companion '26, Dubai, United Arab Emirates.
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2308-7/2026/04
https://doi.org/10.1145/3774905.3795457

1 Introduction

Humanitarian crises, whether caused by natural disasters, conflicts, or pandemics, unfold across complex geographic, social, and linguistic landscapes. Timely and effective humanitarian response depends critically on knowing *where* events occur. Despite the abundance of digital information available today, from situation reports to NGO bulletins and news coverage, automated systems still struggle to extract accurate and equitable geolocation data from text. These limitations not only hinder situational awareness but also risk perpetuating informational inequalities between regions that are richly represented in data and those that remain underreported.

Previous work has shown that existing geolocation pipelines systematically perform better for Western and wealthier countries, while underperforming in low and middle income regions. In particular, a recent study [4] evidenced that biases in both training data and gazetteers lead to an overrepresentation of locations from richer and English speaking regions in automatic geolocation outputs. Such disparities compromise the fairness of humanitarian analytics, where the accuracy of location information can directly affect the visibility of crisis affected communities.

In this work we investigate whether Large Language Models (LLMs) can mitigate these geographic disparities. We design a two step framework for fair geolocation extraction from humanitarian documents, combining a few shot LLM based NER tagger with an agent based geocoding module capable of leveraging context to resolve ambiguous or low resource toponyms. Our analysis focuses not only on overall performance but also on how and where LLMs make errors, quantifying improvements and residual biases across geographic and socioeconomic dimensions.

We benchmark our models against state of the art pretrained and rule based systems, using both accuracy and fairness metrics, and introduce an improved version of the HumSet dataset with refined literal toponym annotations. By bridging advances in large language models with the social mission of humanitarian data analysis, this work aims to leave no place behind, evaluating whether

LLMs can produce more inclusive, transparent, and responsible geolocation systems for global crisis response.

2 Related Work

Named Entity Recognition (NER) plays a central role in understanding news and reports about humanitarian crises, enabling the extraction of entities such as people, organizations, and locations related to ongoing events [34]. Traditional NER models, such as those based on SpaCy and RoBERTa, have achieved strong performance on benchmark datasets. However, they often display inconsistencies in how they tag geographic entities [13], which complicates the extraction of meaningful geographic information from humanitarian text.

The work of Belliardo et al. [4] highlighted this challenge by proposing a hybrid approach combining SpaCy and RoBERTa with transfer learning and manual disambiguation between literal and associative toponyms. Despite improved precision, such approaches require extensive manual annotation and domain-specific heuristics to merge entities and ensure correct tagging. These limitations are particularly evident in humanitarian documents, which are longer, multilingual, and contain complex toponymic expressions (e.g., *"the corner of Lyons Ave & Gregg St"*) that differ from the short, structured texts used in standard NER corpora.

Recent advances in Large Language Models (LLMs) have demonstrated remarkable zero- and few-shot capabilities across a variety of natural language tasks [5, 39]. LLMs have been successfully applied to general-purpose information extraction [42, 43], as well as in more specialised tasks such as cross-domain adaptation [36], open named entity recognition [48], domain-specific NER [24], context-dependent location extractions [38] and russian NER [23]. Two studies, GPT-NER [37] and PromptNER [3], use LLM-based approaches achieve results comparable to supervised deep learning models in the general purpose NER CoNLL dataset, while avoiding the need for large annotated datasets.

Large Language Models are increasingly used in the humanitarian and disaster management domains, where they have been applied to information extraction, situation summarisation, and crisis mapping [21, 43]. Studies show that LLMs outperform traditional NER models in extracting locations from disaster-related social media data [10, 16] while multimodal approaches have leveraged both text and imagery for identifying affected areas [45]. uses images and text to extract the locations of social media tweets that request assistance in an undergoing disaster. However these studies are focused on short text on social media, that contain few locations and with almost all toponyms related to the location of the disaster. One study [14] leveraged LLMs performance on larger texts for identifying urban collapse events from chinese news and extracting the spatio-temporal information.

Bias in geospatial data systems remains an open problem. Datasets for NER tagging [28], gazetteers such as geonames [12], geocoding systems [22] and pretrained supervised models [4, 26, 30, 40] have been shown to exhibit geographic, socioeconomic, and linguistic disparities. Recent analyses reveal that even foundation models display geographic biases and distortions [7, 8, 27, 29], which can propagate into humanitarian decision-making pipelines. Kruspe [19] further demonstrate that such biases extend to geolocation reasoning itself.

In related work on fairness in NER, Mishra et al. [31], Xiao et al. [41] show that performance disparities may arise based on name gender, ethnicity, or popularity, while Johnson et al. [17] identify differences between urban and rural toponym recognition. Yet, to our knowledge, a systematically evaluation of fairness in location-based NER or geocoding with state-of-the-art LLMs is missing, and this is exactly the gap this paper aims to address.

Once location entities are extracted, geocoding methods link them to geographic coordinates or administrative boundaries. Earlier approaches relied on rule-based systems [4], machine learning and deep learning based geocoding [6], gazetteer matching and hybrid approaches [47]. Recent transformer based architectures [46] have demonstrated significant improvements in resolving ambiguous toponyms [15], offering context-aware reasoning capabilities unavailable to traditional systems.

Building on this progress, our work introduces a fully LLM-driven two-step pipeline that combines NER and geocoding while evaluating fairness across geographic and economic strata, thereby bridging the gap between technical performance and equitable humanitarian applications.

3 Methodology

We propose a pipeline for geolocation extraction based on LLMs illustrated in Figure 1. It consists of four main steps: document preprocessing, NER tagging, output postprocessing, and agent-based geocoding, which we next describe.

Document preprocessing. LLMs may struggle to retain long texts verbatim and to preserve the original order of information. To address this, we employed a chunking strategy based on the following rules: (i) each chunk must fall within a predefined minimum and maximum length depending on the output format; (ii) documents are split into chunks using various candidate separators; (iii) we consider the following candidate split separators, ordered from most preferred to least: double line break, ". ", single line break, tabulation, ", ", white space; (iv) separators are tested sequentially until one produces chunks whose lengths fall within the target length range; (v) once a suitable separator is found, the split configuration that minimizes the chunk length variance is chosen. The specific minimum and maximum chunk lengths depend on the tagging format requested from the LLM, as detailed in the following subsection.

LLM-based NER tagging. We propose a few-shot approach for NER tagging based on LLMs. We kept the same input data for all the LLM-based models we tested. Since the LLMs exhibit performance differences depending on the output format [32, 35, 44] we evaluate two distinct formats: one based on JSON and another one based on Markdown. The choice of JSON output enables the LLM to process larger text segments without requiring verbatim reproduction of the entire input. However, a limitation is that the positions of the toponyms within the original text must be determined in a subsequent post-processing step. The markdown approach on the other side, has the advantage of directly providing the position of

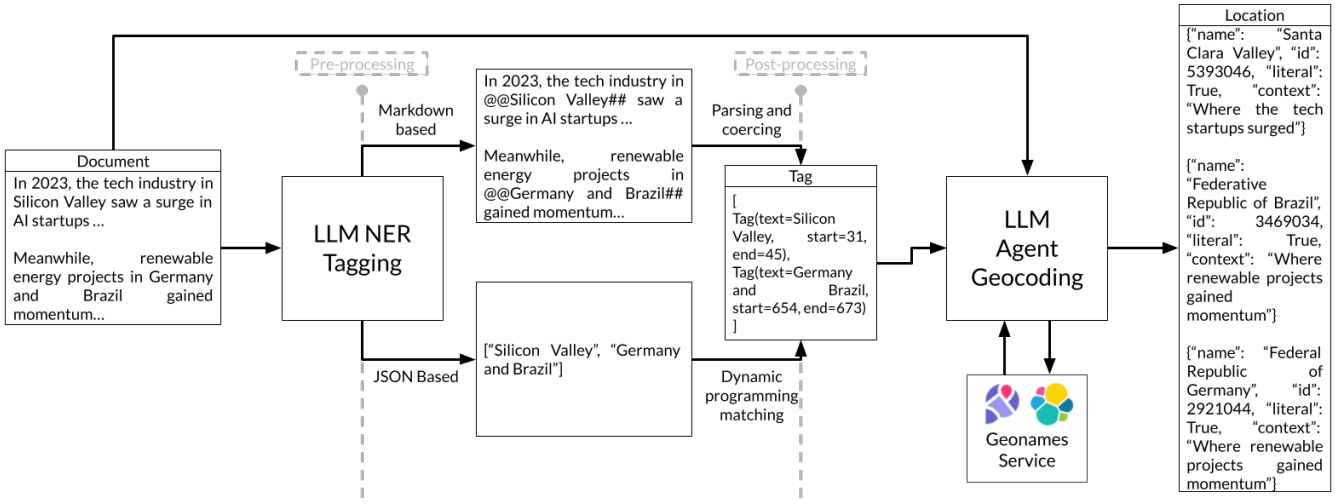


Figure 1: Overview of our LLM-based geolocation extraction pipeline. Humanitarian documents are chunked and sent to a few-shot LLM for location NER tagging, followed by postprocessing that aligns and merges literal toponyms. An agent-based geocoder then uses contextual reasoning and GeoNames queries to resolve each toponym to coordinates, enabling us to evaluate both accuracy and fairness of geolocation across geographic and socioeconomic groups.

each toponym within the text, eliminating the need for the post-processing step required in the JSON-based approach. However, it is more error-prone because LLMs struggle to reproduce text verbatim as length increases. The requested format will be specified in the prompt, after describing the task to the LLM, and before providing the example. The designed prompt is shown in Appendix A.

In the JSON-based approach, the LLM is instructed to return a plain list of location strings in JSON format (e.g., `["Milan", "Naples", "Rome"]`), where each element represents a literal toponym extracted verbatim from the input text. For our Markdown-based approach we follow the output structure proposed by Wang et al. [37], which reproduces the full text delimiting each toponym by `@@` and `##`. Given the size limitation of this approach, we reduce the chunk size to between 200 and 500 characters, compared to the JSON-based approach (1000–2000 characters), to improve reliability.

Post-processing. For the JSON-based approach, the LLM returns a list of toponyms without positional information. A greedy approach that processes each toponym sequentially and matches it to the first available occurrence after the previous match would introduce errors if the tags are misordered or have extra duplicates (e.g., in the phrase "After evacuating from Goma, aid convoys moved through Bukavu, continued to Kigali, and ultimately returned to Goma," if the list is ["Goma", "Goma", "Bukavu", "Kigali"], the greedy algorithm will only recognize the first two appearances of "Goma"). To avoid these mistakes, we use a dynamic-programming alignment that, for each toponym, considers all valid occurrences after the current index (including the option to skip a toponym) and selects the globally optimal, order-preserving assignment that maximizes the number of matched toponyms. As a limitation, this approach cannot handle cases in which a place is mentioned multiple times, appearing sometimes as a literal toponym and other times as an associative one (e.g., in "The US Embassy in Nairobi, working with diplomats and local NGOs in the US, coordinated the humanitarian aid," the second appearance of "US" is the one we should match).

The Markdown-based approach generally requires only verifying that the tagged output chunk matches the input chunk, then extracting the delimited tags and their positions. However, when the output does not exactly match the input, we fall back to treating the extracted toponyms as a list and apply the same dynamic programming alignment algorithm used in the JSON-based approach.

Finally, and since we wanted all references to locations within the same phrase to appear as a single tag (e.g., "Milan, Naples, and Rome"), in both approaches we merged nearby matches that are separated by characters as ',' or connectors like 'and'. This post-processing step was necessary because the LLM sometimes failed to group them correctly or because the chunking process split them across segments. For a fair comparison, we also applied this correction during the re-annotation of the dataset and in the baseline models we compared against.

Geolocation Agent. Once tags are extracted, we process each together with its surrounding document context using a geolocation agent. This agent was designed to identify all individual toponyms referenced within the tag and associate each one with its corresponding location. We implemented our agent in LangChain [20] and used the GeoNames database as the primary source of geographic information [11]. We used the open-source geocoding service Pelias [33] for querying GeoNames.

For each tag the agent alternates between reasoning steps and tool calls until task completion. We define our agent by providing a task description (see prompt in Appendix B) and configuring the following steps: (i) *Search:* a string in geonames matching a text and optionally a country code. (ii) *Select:* a geoname id for a place. The parameters are the place extracted from the text by the LLM (since multiple places may be in the same tag), the geoname

id, a context added by the llm (e.g., "Buenos Aires is where the earthquake happened") and a flag indicating if the toponym is literal to further improve the toponyms dataset. (iii) *Finish:* the agent, with a reason to end the geocoding explaining why the task is done. This is used as a way to get some interpretability and insights over the reasoning of the LLM.

It is worth mentioning that the LLM agent inevitably inherits the structural biases present in GeoNames. By iteratively combining contextual inference with targeted search actions, the agent is able to resolve ambiguous or low-resource toponyms, distinguish between literal and associative place mentions, and justify its selections through explicit reasoning traces. This yields a more transparent and interpretable geocoding process, one that not only improves accuracy but also helps surface and mitigate geographic disparities in downstream predictions.

3.1 Dataset

For the evaluation and comparison of our models we used the HumSet [9] dataset which consists of 15,661 multilingual reports, news and official documents about humanitarian response operations. From them, we use the 467 English documents annotated by the previous study [4]; this enables us to perform a direct comparison with their NER model based on SpaCy and RoBERTa and their rule-based geocoding system. The decision to choose documents in English is due to the fact that international agencies that distribute these reports (e.g, ReliefWeb) publish primarily in this language, which serves as the lingua franca for international humanitarian collaboration.

We build upon the classification of toponyms into literal and associative categories established in prior research [13]. Literal toponyms refer to specific geographic locations, such as "Buenos Aires", while associative toponyms may refer to broader concepts or entities related to a location, like "the Ambassador of Argentina". In this work we focus on literal toponyms, as they are crucial for accurately identifying places linked to humanitarian crises.

Annotations' improvement. We found that some toponyms present in the dataset were missing or were incomplete in the annotations provided by Belliardo et al. [4]. Also, some tags that were annotated as literal toponyms were in fact associative. To address these issues, we developed a GUI tool that allowed our annotators to compare the output of two sets of NER tags and positions for the same document, pinpointing the differences between them and allowing the annotator to select the preferred ones. While this method is not as thorough as manually tagging from scratch, it enables a rapid improvement of the dataset. Using this tool, the previously annotated dataset was compared to the output of our GPT 4o JSON model by a human annotator who solved all the conflicts. Then, the same process was repeated for this new ground truth by a second human annotator who used the tool to compare it against the output of the SpaCy TRF model (the best of the pretrained classical models we tested). This resulted in an annotation set containing 8,213 toponyms. Additionally, we requested an additional output to the geocoding agent asking it to flag those toponyms that were not actually literal. This allowed us to improve the dataset further: 1,534 toponyms were flagged by the geolocator as not literal; these were also reviewed by two human annotators and resulted in removing 205 toponyms from the final annotation set. Our improved set of annotations, along with the experimental code, are available at https://github.com/jian01/llm-fair-geolocation-extraction-WWW2026.

3.2 Experiments

With the improved NER annotation dataset we compared our proposed architecture to the state-of-the-art alternatives. We focused on two types of experiments: NER tagging performance and geocoding performance. For each of them we computed accuracy metrics such as precision, recall and F1 score, and fairness metrics. These experiments benchmark our LLM-driven approach against established methods, providing insights into the advancements made in NER and geolocation tasks through the integration of large language models.

NER tagging performance. We assessed the performance of different language models for NER tagging under both JSON and Markdown output configurations. We evaluated GPT 5, GPT 4o, GPT 4o mini, Deepseek (reasoning and chat versions), Claude Sonnet 4.5, and Claude Haiku 4.5. We also evaluated a zero-shot approach proposed by PromptNER [3] as implemented in the SpaCy-LLM [1] library using GPT 4o. We compare the performance of our LLM-based NER taggers against multiple baselines. We include the pre-trained models used in Belliardo et al. [4] (SpaCy MD and XLM-RoBERTa) and add the transformer variant SpaCy TRF as being closer to the state of the art. Additionally, to assess the potential benefits of transfer learning, we fine-tune XLM-RoBERTa and SpaCy TRF on our improved annotation set using 5-fold cross-validation. While Belliardo et al. [4] proposed an ensemble of fine-tuned RoBERTa and SpaCy models, we do not include this combined variant in our comparison. In our setting, fine-tuned RoBERTa performs substantially below SpaCy TRF, making the ensemble redundant for assessing the relative benefits of LLM-based taggers. In all baselines we used the same strategy for merging same-phrase locations as in our LLM-based NER-taggers to ensure a fair comparison.

Since some toponyms were captured partially (e.g. "Rosario" and "Santa Fe province" instead of "Rosario in Santa Fe province"), we took precision, recall and F1 score metrics both considering exact matches (i.e., requiring the model to capture the full length of the toponym for the tag to be correct), and partial matches, where any tag found inside other tag in the ground truth is considered a true positive, and only tags that have no intersection with the ones in the dataset are considered false positives. By combining exact and partial matching metrics we obtain a more comprehensive view of model performance: exact matching enforces strict adherence to our domain-specific annotation guidelines useful in the context of humanitarian geolocation, while partial matching provides a more lenient evaluation of literal toponym identification, offering insights into the model's core ability to detect relevant locations in a recall-oriented setting. Importantly, in our target deployment scenario, documents with highlighted locations will be reviewed by domain experts. In this human-in-the-loop workflow, recall is more critical than precision: false positives can be easily identified and dismissed by experts during review, whereas false negatives (missed locations) may go entirely unnoticed, potentially omitting critical geographic information from crisis response decisions.

Geolocation performance. For the geocoding task, we compared our LLM-based geolocation agent against the rule-based geocoding algorithm proposed by Belliardo et al. [4]. Using a sample of 10% of the documents (47 documents containing a total of 969 toponyms) we geolocated the toponyms using the LLM agent and the rule-based algorithm proposed in Belliardo et al. [4]. The outputs of each model were reviewed by two independent annotators. Then both of them reconciled differences on their criteria. The output is a set of geoname ids for each toponym, that we use as our gold standard for evaluating the geocoding performance of each method. We assessed the geocoding performance of each method using the following metrics: accuracy by exact id (i.e., matching not only the coordinates but also the administrative region level), the 161km accuracy based on the distance between coordinates associated to the ids, and the country accuracy (i.e., matching the country associated to the ids).

Fairness Assessment. We assessed the fairness of our system by quantifying biases in both the NER tagging models and the geolocators. Here we understand bias as the unequal performance of a model across different geographic and socioeconomic segments. Guided by equality of opportunity principles, we focus on measuring disparities in error rates across different groups rather than overall accuracy. During the NER task, we consider a prediction correct if the predicted and ground-truth tags partially overlap, and we measure both the false negative rate (FNR) and the false discovery rate (FDR). For geographic segmentation, we group toponyms by continent (Africa, Asia, Europe, Americas). For socioeconomic segmentation, we use country income levels (low, lower-middle, upper-middle, high) based on the World Bank GNI classification [2]. FNR captures the proportion of true location mentions that the model fails to identify high values in a region indicate that critical toponyms are systematically overlooked in that area. FDR captures the proportion of predicted locations that are incorrect high values indicate that the model over-identifies toponyms or misclassifies associative mentions (e.g., institutions) as literal places.

To quantify bias across groups, we use the equality-of-opportunity difference, denoted ΔFNR and ΔFDR, defined as the difference between the highest and lowest error rates across all geographic or socioeconomic segments. The Δ value of "0" indicates perfect parity (equal performance across groups), whereas larger values reflect greater disparity. For the geolocator, we consider a prediction correct if the predicted location lies within 161 km (approximately 100 miles) of our gold standard, a commonly used threshold in geolocation tasks [25]. We then compute the corresponding classification error across the same geographic and socioeconomic segments.

4 Results

We evaluate our models along three axes: NER tagging, geolocation accuracy, and fairness across geographic and socioeconomic groups.

NER Tagging Performance. Table 1 reports accuracy and fairness metrics for a selection of the best-performing LLM-based NER taggers, alongside pretrained and fine-tuned baselines. The full set of accuracy metrics for all LLM-based models is provided in Appendix C. Overall, Large Language Models outperform traditional pretrained deep learning models in both exact and partial NER tagging, without requiring task-specific training and while allowing flexible control over output format and toponym granularity. Fine-tuned models such as SpaCy TRF tuned achieve competitive scores relative to the best LLMs, but they depend on substantial annotated data and retraining when moving to new domains.

Among the LLM-based models, GPT-4o JSON attains the highest exact precision (0.87) and shares the best exact F1 score (0.84) with the fine-tuned SpaCy TRF model, making it a strong choice when precise extraction of literal toponyms is critical. SpaCy TRF tuned, in turn, reaches the highest exact recall (0.85), illustrating how fine-tuned architectures can be adapted to favor more enriched toponym spans (e.g., "Buenos Aires city" instead of "Buenos Aires").

For partial metrics, which emphasise the correct identification of literal toponyms even when span boundaries are not exact, GPT-5 Markdown achieves the highest recall (0.99). The best partial precision (0.93) and partial F1 score (0.94) are obtained by the fine-tuned SpaCy TRF model, which also already performs strongly in its pretrained form. By contrast, both pretrained and fine-tuned RoBERTa models exhibit consistently low recall (ranging from 0.27 to 0.55), making them unreliable for capturing the full set of relevant locations in humanitarian texts. Finally, the zero-shot GPT-4o model attains F1 scores that are comparable to, and in some cases higher than, other LLM configurations, underscoring the ability of modern LLMs to deliver competitive performance even without task-specific examples.

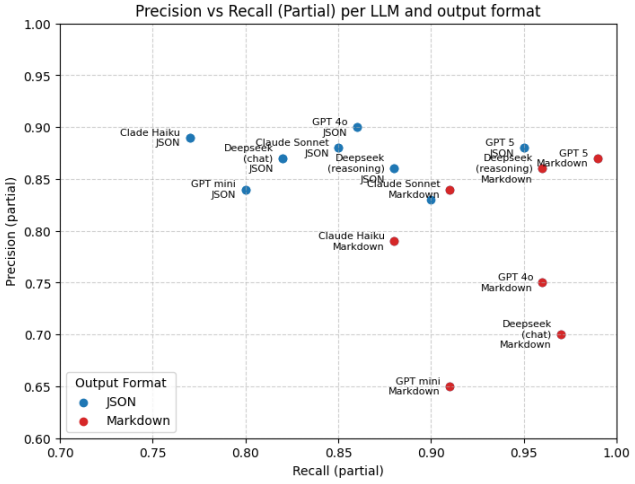


Figure 2: Precision–recall trade-offs for the evaluated LLM-based NER taggers under JSON and Markdown output formats. JSON configurations generally yield higher precision, while Markdown formats achieve higher recall, reflecting the tendency of Markdown prompts to capture more toponyms at the cost of increased false positives.

Regarding output formatting, we observe systematic differences between JSON and Markdown configurations, as shown in Figure 2. JSON-based prompts tend to achieve higher precision, whereas Markdown-based prompts yield higher recall. One plausible explanation is that Markdown outputs must reproduce the input text

	LLM-based models				Pre-trained models			Fine-tuned models	
	GPT 5 MD	GPT 4o JSON	GPT 4o MD	GPT 4o (zero shot)	SpaCy MD	RoBERTa vanilla	SpaCy TRF	RoBERTa tuned	SpaCy TRF tuned
Accuracy Metrics *(higher is better)*									
Exact									
Precision	0.73	**0.87**	0.66	0.74	0.59	0.62	0.68	0.76	0.83
Recall	0.83	0.82	0.84	0.80	0.52	0.27	0.70	0.44	**0.85**
F1 Score	0.78	**0.84**	0.74	0.77	0.55	0.37	0.69	0.56	**0.84**
Partial									
Precision	0.87	0.90	0.75	0.83	0.87	0.90	0.92	0.90	**0.93**
Recall	**0.99**	0.86	0.96	0.90	0.76	0.39	0.94	0.52	0.95
F1 Score	0.92	0.88	0.84	0.87	0.81	0.54	0.93	0.66	**0.94**
Fairness Metrics *(lower is better)*									
Partial False Negative Rate by Continent									
Asia	**0.01**	0.15	0.03	0.09	0.28	0.67	0.06	0.50	0.04
Africa	**0.01**	0.14	0.03	0.09	0.25	0.60	0.06	0.46	0.04
Americas	**0.01**	0.10	0.03	0.08	0.11	0.49	0.02	0.44	0.03
Europe	0.02	0.13	0.01	0.35	0.06	0.56	**0.00**	0.73	0.09
ΔFNR	**0.01**	0.05	0.02	0.27	0.22	0.17	0.06	0.29	0.06
Partial False Negative Rate by Country income level									
Low Income	**0.01**	0.14	0.03	0.09	0.26	0.64	0.05	0.50	0.05
Lower-middle income	**0.02**	0.12	0.03	0.09	0.26	0.58	0.09	0.42	0.05
Upper-middle income	**0.01**	0.13	0.03	0.09	0.21	0.57	0.04	0.45	0.02
High income	**0.02**	0.10	0.04	0.17	0.07	0.59	0.02	0.60	0.03
ΔFNR	**0.01**	0.04	**0.01**	0.08	0.18	0.07	0.08	0.18	0.03
Partial False Discovery Rate by Continent									
Asia	0.21	0.11	0.29	0.19	0.13	**0.09**	0.12	0.09	0.09
Africa	0.17	0.10	0.27	0.17	0.16	0.13	**0.07**	0.12	0.08
Americas	0.22	0.17	0.38	0.24	0.20	**0.15**	0.15	0.17	0.12
Europe	0.73	0.54	0.75	0.74	0.79	**0.53**	0.71	0.61	0.63
ΔFDR	0.56	0.44	0.48	0.58	0.66	**0.43**	0.65	0.52	0.55
Partial False Discovery Rate by Country income level									
Low Income	0.18	0.10	0.26	0.17	0.14	0.13	0.09	0.12	**0.07**
Lower-middle income	0.23	0.11	0.31	0.19	0.14	0.10	**0.09**	0.11	0.08
Upper-middle income	0.17	0.14	0.30	0.19	0.15	0.12	**0.09**	0.11	0.12
High income	0.50	0.36	0.59	0.51	0.61	**0.35**	0.51	0.35	0.39
ΔFDR	0.33	0.26	0.33	0.34	0.47	0.25	0.42	**0.24**	0.32

Table 1: Accuracy and fairness metrics for selected LLM-based NER taggers, compared against pretrained and fine-tuned deep learning baselines. The table reports exact and partial precision, recall, and F1 scores, as well as fairness metrics based on false negative and false discovery rates across geographic and socioeconomic groups. Higher accuracy values and lower disparity values indicate better performance.

verbatim, which may make it easier not to miss location mentions, but also more likely to introduce false positives, particularly for associative toponyms. These format-dependent trade-offs highlight the importance of prompt and output design when deploying LLMs for location extraction in humanitarian text.

Geolocation Performance. Table 2 summarizes the geocoding accuracy and fairness metrics for our LLM-based geolocation agent compared to the rule-based baseline from Belliardo et al. [4]. The agent-based approach significantly outperforms the rule-based system across all accuracy metrics: exact, 161 km, and country-level precision and recall are consistently higher for the agent-based model. Figure 3 visualizes the spatial distribution of prediction errors for both geocoding approaches, measured as the distance between each model's predicted location and our gold-standard annotations. The maps show that our agent-based geolocator achieves substantially lower errors across all world regions compared to the rule-based baseline, reducing large-distance errors and cross-continent mismatches.

	Agent-based	Rule-based [4]
Accuracy Metrics *(higher is better)*		
Exact		
Precision	**0.88**	0.69
Recall	**0.79**	0.59
@161km		
Precision	**0.91**	0.83
Recall	**0.82**	0.71
Country-level		
Precision	**0.91**	0.82
Recall	**0.83**	0.70
Fairness Metrics *(lower is better)*		
Classif. error@161km by Continent		
Asia	**0.24**	0.43
Africa	**0.25**	0.34
Americas	**0.06**	0.44
Europe	**0.00**	0.58
ΔError	0.25	**0.24**
Classif. error@161km by Country income level		
Low income	**0.25**	0.30
Lower-middle income	**0.21**	0.44
Upper-middle income	**0.23**	0.42
High income	**0.07**	0.65
ΔError	**0.18**	0.35

Table 2: Accuracy and fairness metrics for the geocoding task, comparing our agent-based LLM geolocator with the rule-based baseline from Belliardo et al. [4]. Higher accuracy and lower disparity indicate better performance. The agent-based model achieves substantial improvements in both accuracy and geographic parity.

Table 3 illustrates representative prediction errors for each geocoding approach. In the first example, `eastern South Sudan`, the agent-based model incorrectly resolves to a more specific administrative region (Eastern Equatoria) rather than the country-level entity, likely because the LLM over-interprets the directional modifier "eastern" as requiring a sub-national match. The second example, `central Mediterranean`, showcases a failure unique to the rule-based approach: it incorrectly matches to Davao in the Philippines, possibly due to gazetteer ambiguity or lack of contextual reasoning, while the agent successfully identifies the Mediterranean Sea. Notably, one of the alternative names for Davao in GeoNames is "central", which may have confused the rule-based model. Finally, both models fail on `Dar-fur`: the agent selects North Darfur (a subregion) instead of the broader Darfur region, while the rule-based system erroneously links it to Dar es Salaam in Tanzania, likely misled by partial string matching and the hyphen symbol. These examples highlight that while the agent-based approach generally outperforms the baseline, it can still over-specialise when strong contextual cues are present, whereas rule-based methods remain vulnerable to superficial lexical similarities.

Fairness Assessment. We now analyse how performance varies across geographic and socioeconomic segments, focusing on disparities in error rates rather than aggregate accuracy. For NER tagging, we consider a prediction correct if the predicted and ground-truth tags partially overlap, and measure false negative rates (FNR) and false discovery rates (FDR) across continents (Africa, Asia, Europe, Americas) and World Bank income groups (low, lower-middle, upper-middle, high [2]). To summarise disparities, we report equality-of-opportunity differences ΔFNR and ΔFDR, defined as the difference between the highest and lowest error rates across segments (0 indicates perfect parity, larger values indicate greater disparity).

In terms of FNR, Table 1 shows that GPT-5 Markdown and GPT-4o Markdown maintain near-parity across continents and income levels, with maximum differences of around 1 percentage point between groups. This suggests that their extraction ability remains stable across geographic and socioeconomic contexts, mitigating the regional biases observed in earlier systems. In contrast, RoBERTa variants exhibit stronger disparities, while SpaCy TRF achieves relatively low FDR but lags behind in exact tagging accuracy. The zero-shot GPT-4o configuration displays among the highest disparities overall, possibly because the limited in-context examples underrepresent certain regions and thus shift the model's implicit bias. Taking both accuracy and fairness into account, GPT-5 Markdown emerges as the best overall NER model, achieving nearly uniform FNR across groups while maintaining high partial recall (0.99) and F1 (0.92).

FDR remains a challenge for all models, with disparities of up to 0.66 driven largely by high false positive rates in Europe. The particularly low proportion of true literal toponyms located in Europe (below 2%) makes its FDR highly sensitive to model misclassifications.

For the geocoding task, we consider a prediction correct if the predicted location lies within 161 km (approximately 100 miles) of our gold standard [25], and compute the corresponding classification error@161 km across the same segments, summarising disparities with ΔError.

As shown in Table 2, our agent-based geolocator reduces classification error@161 km in all continents and income groups relative to the rule-based baseline. The rule-based system exhibits particularly high error in Europe and high-income countries, reflecting frequent cross-country or cross-continent mismatches. The agent-based approach not only lowers overall error but also reduces ΔError by roughly half when stratified by income level, indicating more uniform performance across economic strata and making it better suited for equitable geolocation in humanitarian contexts.

5 Discussion

Building on the findings of *Leave No Place Behind* [4], our results confirm that improving the quality and representativeness of training data remains central to mitigating geographic bias in geolocation extraction. While earlier work showed that tuning classical NER models such as SpaCy and RoBERTa on humanitarian data improved accuracy and slightly alleviated Western bias, our study demonstrates that Large Language Models can further advance both objectives.

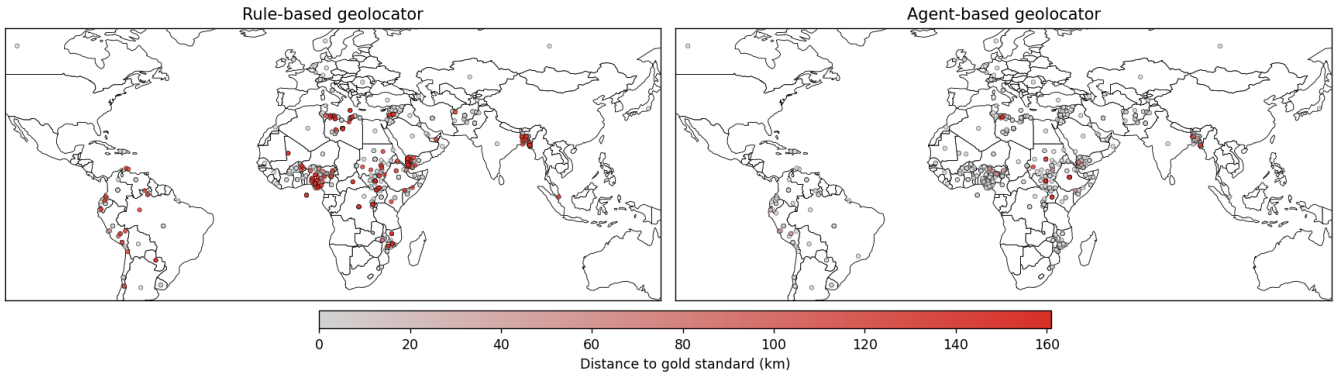


Figure 3: Geographic distribution of geocoding errors for the rule-based baseline (left) and our agent-based geolocator (right). Each point represents a toponym from the annotated evaluation set, plotted at its gold-standard location and colored by distance (in km) between the system's predicted coordinates and the ground truth. Darker colors indicate larger errors. The rule-based method exhibits frequent long-distance misclassifications, including cross-continent mismatches, whereas the agent-based geolocator substantially reduces large-error cases and achieves more geographically consistent performance across regions.

	Toponym	Gold Standard	Agent-based	Rule-based
Agent-only	`eastern South Sudan`	South Sudan	Eastern Equatoria	South Sudan
Rule-only	`central Mediterranean`	Mediterranean Sea	Mediterranean Sea	Davao
Both	`Dar-fur`	Darfur	North Darfur	Dar es Salaam

Table 3: Representative geocoding error cases illustrating distinct failure modes of the agent-based and rule-based geolocators. Each row shows a toponym from the evaluation set, its gold-standard location, and the predictions from both systems. The examples highlight agent-only over-specification (`eastern South Sudan`), rule-based lexical mismatches (`central Mediterranean`), and ambiguous or nonstandard toponyms that challenge both methods (`Dar-fur`).

While SpaCy TRF tuned achieves metrics comparable to the best LLM-based models, it is important to note that our LLM configurations operate in a few-shot regime, enabling the development of domain-specific NER taggers with minimal annotation effort. Future work should investigate whether supervised prompt optimisation [18] can further improve few-shot performance and potentially surpass the gains achieved through classical fine-tuning.

Our results also show that NER performance with LLMs depends strongly on the chosen output format. Formatting thus becomes an integral part of prompt design and requires deliberate selection and tuning. Markdown-based outputs, in particular, tend to generate more tokens, which can increase computational cost and should be taken into account when selecting a production configuration.

For NER tagging, both GPT-5 Markdown and GPT-4o Markdown display near-parity in false negative rates across continents and income groups, indicating that their extraction ability is geographically and socioeconomically balanced. In contrast, pretrained and fine-tuned SpaCy TRF models exhibit larger disparities and tend to underperform in the Global South, particularly in lower-income regions. For geocoding, our agent-based approach substantially improves almost all accuracy and fairness metrics across groups compared to the rule-based baseline.

The progression from domain-tuned transformers [4] to foundation LLMs marks a substantial methodological shift: these models no longer require explicit retraining to adapt to humanitarian text, yet they deliver superior results with less manual intervention. This opens new opportunities for rapid deployment in multilingual and low-resource humanitarian contexts, where labeled data are scarce. Nevertheless, as our analysis shows, even the most advanced LLMs are not free from residual geographic and socioeconomic biases, emphasizing the importance of continuous algorithmic auditing and inclusion of diverse linguistic and regional data sources.

Overall, our results demonstrate that LLM-driven geolocation not only improves extraction quality, but also meaningfully reduces geographic disparities, helping build more inclusive and reliable information pipelines for humanitarian response. Compared to previous efforts that improved bias mitigation through fine-tuning [4], LLMs can achieve comparable or better fairness out-of-the-box, while also advancing the state of the art in extraction accuracy, enabling adaptive systems that can further reduce bias and ensure that, truly, *no place is left behind.*

6 Acknowledgments

KK and YM acknowledge support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation) funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT). MGB and GC acknowledge financial support from Universidad de San Andrés.

References

[1] 2023. *spaCy-LLM: Integrating Large Language Models into spaCy.* https://github.com/explosion/spacy-llm Python library.

[2] 2025. *Understanding country income: World Bank Group income classifications for FY26.* https://blogs.worldbank.org/en/opendata/understanding-country-income--world-bank-group-income-classifica#one

[3] Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444* (2023).

[4] Enrico Belliardo, Kyriaki Kalimeri, and Yelena Mejova. 2023. Leave no Place Behind: Improved Geolocation in Humanitarian Documents. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 31–39. doi:10.1145/3582515.3609515

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma-teusz Litwin, Scott Gray, B. Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, I. Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Neural Information Processing Systems* 33 (2020), 1877–1901.

[6] Gaëtan Caillaut, Samuel Auclair, Cécile Gracianne, Nathalie Abadie, and Guillaume Touya. 2024. Entity Linking for real-time geolocation of natural disasters from social network posts. *PLoS ONE* 19, 10 (2024), e0307254. doi:10.1371/journal.pone.0307254

[7] Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. 2025. Evaluation of geographical distortions in language models. *Machine Learning* 114, 12 (2025), 1–32.

[8] Shiran Dudy, Thulasi Tholeti, R. Ramachandranpillai, Muhammad Ali, T. Li, and Ricardo Baeza-Yates. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. *Proceedings of the 30th International Conference on Intelligent User Interfaces* (2025), 1499–1516. doi:10.1145/3708359.3712111

[9] Selim Fekih, Nicolo' Tamagnone, Benjamin Minixhofer, Ranjan Shrestha, Ximena Contla, Ewan Oglethorpe, and Navid Rekabsaz. 2022. HumSet: Dataset of Multilingual Information Extraction and Classification for Humanitarian Crises Response. (Dec. 2022), 4379–4389. doi:10.18653/v1/2022.findings-emnlp.321

[10] Hafiz Budi Firmansyah, Valerio Lorini, Mehmet Oguz Mulayim, Jorge Gomes, and Jose Luis Fernandez-Marquez. 2024. Improving social media geolocation for disaster response by using text from images and chatgpt. In *Proceedings of the 2024 11th Multidisciplinary International Social Networks Conference*. 67–72.

[11] GeoNames. 2025. *GeoNames geographical database.* https://www.geonames.org Accessed: 2025-11-12.

[12] Mark Graham and Stefano De Sabbata. 2015. Mapping information wealth and poverty: the geography of gazetteers. *Environment and Planning A* 47, 6 (2015), 1254–1264.

[13] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics. *Language Resources and Evaluation* 54, 3 (2020), 683–712. doi:10.1007/s10579-019-09475-3

[14] Yanan Hao, Jin Qi, Xiaowen Ma, Sensen Wu, Renyi Liu, and Xiaoyi Zhang. 2024. An LLM-Based Inventory Construction Framework of Urban Ground Collapse Events with Spatiotemporal Locations. *ISPRS International Journal of Geo-Information* 13, 4 (2024), 133. doi:10.3390/ijgi13040133

[15] Xuke Hu, Jens Kersten, Friederike Klan, and Sheikh Mastura Farzana. 2024. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science* (2024), 1–28. doi:10.1080/13658816.2024.2405182

[16] Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou, and Kenneth Joseph. 2023. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science* 37, 11 (2023), 2289–2318. doi:10.1080/13658816.2023.2266495

[17] Isaac Johnson, Connor McMahon, Johannes Schöning, and Brent Hecht. 2017. The Effect of Population and "Structural" Biases on Social Media-based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*. 1167–1178. doi:10.1145/3025453.3026015

[18] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint arXiv:2310.03714* (2023).

[19] Stillman Kruspe. 2025. Biased Geolocation in LLMs: Experiments on Probing LLMs for Geographic Knowledge and Reasoning. (2025).

[20] LangChain Developers. 2023. *LangChain.* https://www.langchain.com Accessed: 2025-10-13.

[21] Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi R Wang, and Jundong Li. 2025. Harnessing large language models for disaster management: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*. 14528–14551.

[22] Tatu Leppämäki, Tuuli Toivonen, and Tuomo Hiippala. 2024. Geographical and linguistic perspectives on developing geoparsers with generic resources. *International Journal of Geographical Information Science* 38, 10 (2024), 2039–2060.

[23] Maria Levchenko. 2025. Evaluating Named Entity Recognition Models for Russian Cultural News Texts: From BERT to LLM. *arXiv.org* (2025). doi:10.48550/arxiv.2506.02589

[24] Mingchen Li and Rui Zhang. 2023. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186* (2023).

[25] Yimin Liu, Xiangyang Luo, Zhiyuan Tao, Meng Zhang, and Shaoyong Du. 2023. UGCC: Social Media User Geolocation via Cyclic Coupling. *IEEE Transactions on Big Data* 9, 4 (2023), 1128–1141. doi:10.1109/TBDATA.2023.3242961

[26] Zilong Liu, Krzysztof Janowicz, Ling Cai, Rui Zhu, Gengchen Mai, and Meilin Shi. 2022. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. *AGILE: GIScience Series* 3 (2022), 9.

[27] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2024. On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper). *ACM Trans. Spatial Algorithms Syst.* 10, 2, Article 11 (July 2024), 46 pages. doi:10.1145/3653070

[28] Momin Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population Bias in Geotagged Tweets. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 4 (2015), 18–27. doi:10.1609/icwsm.v9i4.14688

[29] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 1409, 16 pages.

[30] Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. doi:10.1145/3372923.3404804

[31] Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. *arXiv: Computation and Language* (2020).

[32] Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards LLMs Robustness to Changes in Prompt Format Styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, Abteen Ebrahimi, Samar Haider, Emmy Liu, Sammar Haider, Maria Leonor Pacheco, and Shira Wein (Eds.). Association for Computational Linguistics, Albuquerque, USA, 529–537. doi:10.18653/v1/2025.naacl-srw.51

[33] Pelias Contributors. 2014. *Pelias: Open Source Geocoder.* https://pelias.io

[34] Roberta Rocca, Nicolò Tamagnone, Selim Fekih, Ximena Contla, and Navid Rekabsaz. 2023. Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian NLP. *Frontiers in big Data* 6 (2023), 1082787.

[35] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance.. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 1218–1236. doi:10.18653/v1/2024.emnlp-industry.91

[36] Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. InstructionNER: A Multi-Task Instruction-Based Generative Framework for Few-shot NER. *arXiv preprint arXiv:2203.03903* (2022).

[37] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named Entity Recognition via Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 4257–4275. doi:10.18653/v1/2025.findings-naacl.239

[38] 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, and Chaowei Yang. 2025. Optimizing context-based location extraction by tuning open-source LLMs with RAG. *International Journal of Digital Earth* 18, 1 (2025), 2521786. doi:10.1080/17538947.2025.2521786

[39] Jason Wei, Maarten Bosma, Vincent 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. *arXiv preprint arXiv:2109.01652* (2022).

[40] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. Ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora. *Frontiers in Digital Humanities* 5 (2018), 2.

[41] Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification.

In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 123–137. doi:10.1145/3593013.3593982

[42] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science* 18, 6 (2024), 186357.

[43] Fengyi Xu, Jun Ma, Nan Li, and Jack C.P. Cheng. 2025. Large language model applications in disaster management: an interdisciplinary review. *International Journal of Disaster Risk Reduction* (2025), 105642. doi:10.1016/j.ijdrr.2025.105642

[44] Jialin Yang, Dongfu Jiang, Lipeng He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. 2025. StructEval: Benchmarking LLMs' Capabilities to Generate Structural Outputs. *CoRR* abs/2505.20139 (2025). arXiv:2505.20139 doi:10.48550/ARXIV.2505.20139

[45] Wenping Yin, Yong Xue, Ziqi Liu, Hao Li, and Martin Werner. 2025. LLM-enhanced disaster geolocalization using implicit geoinformation from multimodal data: A case study of Hurricane Harvey. *International Journal of Applied Earth Observation and Geoinformation* 137 (2025), 104423.

[46] Zeyu Zhang and Steven Bethard. 2023. Improving Toponym Resolution with Better Candidate Generation, Transformer-based Reranking, and Two-Stage Resolution. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, Alexis Palmer and Jose Camacho-collados (Eds.). Association for Computational Linguistics, Toronto, Canada, 48–60. doi:10.18653/v1/2023.starsem-1.6

[47] Zeyu Zhang and Steven Bethard. 2025. A survey on geocoding: algorithms and datasets for toponym resolution. *Language Resources and Evaluation* 59, 2 (2025), 1775–1796.

[48] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *arXiv preprint arXiv:2308.03279* (2024).

A NER tagging prompt

Table 4 shows the prompt structure used for the NER tagging task.

```
You are a Named Entity Recognition (NER) system specialized in extracting
**literal toponyms** (geographic location names) from texts about natural
disasters and accidents.
Your task is to identify and return **only the explicit literal mentions of
physical locations** (toponyms), avoiding any associative uses.
Each time the user gives you a text you simply answer each occurence of an
explicit literal toponym avoiding associative toponyms.

Key constraints:
- Extract only literal toponyms, defined as:
  - Proper names of places (e.g., "Cambridge", "Germany")
  - Noun modifiers of places (e.g., "Paris pub")
  - Adjectival modifiers with geographic meaning (e.g., "southern Spanish
  city")

- Do NOT extract associative toponyms, including:
  - Metonymic references (e.g., "She used to play for Cambridge")
  - Demonyms (e.g., "a Jamaican")
  - Homonyms (e.g., "I asked Paris to help")
  - Languages (e.g., "in Spanish")
  - Noun/adjectival modifiers not referring to a physical place (e.g.,
  "Spanish ham")
  - Embedded uses (e.g., "US Supreme Court", "US Dollar")
  - Toponyms in URLs

Extraction Rules:
1. Preserve literal mentions exactly as they appear — no rephrasing or
normalization.
2. Preserve order — output the locations in the same order as in the input.
3. Do not remove duplicates.
4. Include:
   - Geographical regions (e.g., "Patagonia", "coastal Germany")
   - Roads, borders, and composite names (e.g., "Buenos Aires–Mar del Plata
   road")
   - Temporary places like refugee camps
   - Institutions only if they imply a geographic location (e.g., "Cambridge
   University" implies "Cambridge")
5. Prefer the most specific geographic level available (e.g., "Buenos Aires
province" over "Buenos Aires").
6. Include articles if they are part of the place name.
7. Include **cardinal directions** (e.g., "southern Spain").
8. **Do not merge** multiple toponyms unless they jointly modify a noun
immediately after (e.g., "South Dakota, New York and Michigan states" →
merge all three).
9. If multiple locations are listed and are connected by commas, "and", or
"in" within a single continuous phrase, keep them together as one literal
toponym exactly as
written, even if they contain multiple place names.
10. Merge nested location phrases with possessive or relational structure
```

Table 4: Prompt used for the LLM-based NER taggers.

B Geolocator agent prompt

Table 5 shows the prompt structure used for the geolocator agent.

```
## Task & Tools

You work as a **geolocator**. A location (PLACE) and its **context**
will be provided. Your goal is to maximize 100 mile accuracy. Use the
**tools**
1) `search_tool` candidates in GeoNames database. You can pass an
ISO country code when the context hints at it (e.g., 'AR', 'BR', 'MZ'),
2) For each **definite** location, call **`select_tool`** (you may
call it multiple times) for select the best matching entry. Be precise
and prefer cities/settlements over vague regions when the
context points to them,
3) When you are done, call **`finish_tool`** exactly once to
return **all selections together**.

## Considerations

- If PLACE is ambiguous, refine with `search_tool` (try alternate spellings
from context).
- When PLACE is a country name, you MUST select the sovereign state entry
(feature_code PCL*).
- If a PCL* candidate is not returned on the first search, try alternate
spellings or exonyms/endonyms and search again.
- If the context indicates a country explicitly, include it in searches.
- Prefer inhabited places (feature classes `P*`) when context refers to
education, health, local news, municipal services, etc.
- If the text clearly references a **province/state/county** instead of a
city, select that administrative division.
- Avoid false positives from homonyms in other countries—check language,
nearby mentions, and timezone cues in context.
- If no good candidate appears, perform another `search_tool` with a better
query (e.g., remove accents, try shorter stems).
- If PLACE matches a sovereign state (country name), prefer the country
entry (feature_code = 'PCLI') rather than a city.
- Do not call `select_tool` twice for the same geonameid or the same place.
- You must cover every sublocation enumerated in PLACE; do not call
finish_tool until each is either selected or explicitly marked with
select_tool(geonameid=-1, ...).
- You only have 15 actions to use.
- Whether relevant or not, try to locate all toponyms, only the ones not in
geonames should be left without selection.
- Prefer shorter searches and do not insist too much on one location, you
have limited action, prioritize selection before searching.
- Do at most **only two** searches per place.
- Do not extract implicit location from within the context or not explicitly
named or directly associated to the place string
- If the toponym is "<X> in <Y>" extract just <X> unless <X> is not in
geonames, then extract <Y>
- Do **not** add extra locations not related to the place input

## Non-literal (associative) toponyms

Examples of non-literal toponyms:
* Metonymy: She used to play for **Cambridge**.
* Homonym: I asked **Paris** to help me.
* Demonym: I spoke to a **Jamaican** on the bus.
* Language: She spoked **Spanish**.
* Noum modifier: That **Paris** souvenir is interesting.
* Adjectival modifier: I ate some **Spanish** ham yesterday.
* Embedded associative: **US** Supreme Court has 9 justices.
You should mark those kind of toponyms as not literal.

## Output protocol
- Call `select_tool(geonameid, context, literal_toponym)` for every
confirmed location.
  - geonameid should **always** be present as a valid geonames id, get the
closest one possible.
  - `context` is an English explanation like "Neighborhood mentioned in Yola
North LGA", or
    "Government jurisdiction", etc. If the toponym is not literal you should
specify why.
  - `literal_toponym` is True when the toponym refers to a real place
related to what the document is describing, False when associative.
  - You should select only one id per place and then finish.
- Finally, call `finish_tool(reason=...)` once. Do not return free-form text.
```

Table 5: Prompt used for the geocoding agent.

C Accuracy metrics for all LLM-based NER taggers and output formats

Table 6 shows the accuracy metrics for all the LLM-based models we assessed for the NER tagging task.

	Exact			Partial		
Model	Prec.	Rec.	F1	Prec.	Rec.	F1
GPT 5						
JSON	0.76	0.82	0.79	0.88	0.95	**0.92**
MD	0.73	0.83	0.78	0.87	**0.99**	**0.92**
GPT 4o						
JSON	**0.87**	0.82	**0.84**	**0.90**	0.86	0.88
MD	0.66	**0.84**	0.74	0.75	0.96	0.84
zero-shot	0.74	0.80	0.77	0.83	0.90	0.87
GPT mini						
JSON	0.74	0.71	0.72	0.84	0.80	0.82
MD	0.56	0.79	0.66	0.65	0.91	0.76
Deepseek reason.						
JSON	0.77	0.79	0.78	0.86	0.88	0.87
MD	0.74	0.83	0.78	0.86	0.96	0.91
Deepseek chat						
JSON	0.78	0.73	0.75	0.87	0.82	0.84
MD	0.59	0.81	0.68	0.70	0.97	0.81
Claude Sonnet						
JSON	0.79	0.76	0.77	0.88	0.85	0.87
MD	0.73	0.79	0.76	0.84	0.91	0.88
Claude Haiku						
JSON	0.79	0.69	0.73	0.89	0.77	0.82
MD	0.69	0.76	0.72	0.79	0.88	0.83

Table 6: Accuracy metrics for all LLM-based models for the NER tagging task. Bold numbers indicate the best model for each metric.