Taylor & Francis
Taylor & Francis Group

Georeferencing complex relative locality descriptions with large language models

Aneesha Fernando, Surangika Ranathunga, Kristin Stock, Raj Prasanna & Christopher B. Jones

Published online: 21 Jan 2026.

Submit your article to this journal ↗

Article views: 199

View related articles ↗

View Crossmark data ↗

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

🔓 OPEN ACCESS | Check for updates

Georeferencing complex relative locality descriptions with large language models

Aneesha Fernando[a], Surangika Ranathunga[a], Kristin Stock[a], Raj Prasanna[b] and Christopher B. Jones[c]

[a]School of Computational and Mathematical Sciences, Massey University, Auckland, New Zealand; [b]Joint Centre for Disaster Research, Massey University, Wellington, New Zealand; [c]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

ABSTRACT

Georeferencing text documents has typically relied on either gazetteer-based methods to assign geographic coordinates to place names or on language modelling approaches that associate textual terms with geographic locations. However, many location descriptions specify positions relatively with spatial relationships, making geocoding based solely on place names or geo-indicative words inaccurate. This issue frequently arises in biological specimen collection records, where locations are often described through narratives rather than coordinates if they pre-date GPS. Accurate georeferencing is vital for biodiversity studies, yet the process remains labour-intensive, leading to a demand for automated georeferencing solutions. This paper explores the potential of Large Language Models (LLMs) to georeference complex locality descriptions automatically, focusing on the biodiversity collections domain. We first identified effective prompting patterns, then fine-tuned an LLM using Quantized Low-Rank Adaptation (QLoRA) on biodiversity datasets from multiple regions and languages. Our approach outperforms existing baselines with an average, across datasets, of 65% of records within a 10 km radius, for a fixed amount of training data. The best results (New York state) were 85% within 10 km and 67% within 1 km. The selected LLM performs well for lengthy, complex descriptions, highlighting its potential for georeferencing intricate locality descriptions.

1. Introduction

While most geographical information systems are predominantly based on structured digital map data, there remains a vast amount of information embedded within textual resources. Access to the information content of such resources is the subject of the field of geographical information retrieval (Purves *et al.* 2018), which depends to a

---

CONTACT Aneesha Fernando ✉ afwboosa@massey.ac.nz

large extent on the effectiveness of georeferencing methods to determine the geospatial focus of the content of text documents. To date, georeferencing methods for textual data have usually been applied to the content of web pages and of social media postings. These methods typically employ either gazetteer-based approaches to detect and geocode (determine coordinates for) place names in the texts, or language modelling approaches that depend upon determining the association between text and locations, or some combination of these (Melo and Martins 2017, Gritta *et al.* 2018). References to geographic locations through the use of place names are usually assumed to be absolute, in the sense that the location is regarded as equivalent to that of a place name or some other words that are indicative of a location (Han *et al.* 2012). Little attention has been given to the fact that some descriptions of location are relative, in that they refer to a location that has some spatial relationship to a reference place name (Wieczorek *et al.* 2004, van Erp *et al.* 2015, Chen *et al.* 2018). Spatial relational terms include phrases and words that indicate a specified distance, such as *10 km west of*, as well as relative positions like *near*, *adjacent to*, and *along from*.

A domain in which complex locality descriptions are commonly found is that of the collection records of natural history agencies such as museums and herbaria. These collections can include records of biological specimens of plants and animals, fungi and bacteria, as well as soil and geological samples. Many of the records, especially those collected before the widespread availability of GPS, do not have associated geographical coordinates. Instead, their locations are described solely through textual descriptions that include locality descriptions. At its simplest, a locality description might just consist of place names (toponyms), but very commonly, place names are combined with relative spatial terms (Wieczorek *et al.* 2004, van Erp *et al.* 2015, Scott *et al.* 2021). There are billions of such records and there is a strong motivation to georeference them, as being able to assign coordinates, and hence map the locations at which they were found, is a crucial step in studying biodiversity. Such georeferenced data enables researchers to monitor the geographical distribution of species over time, the impacts of environmental changes on species, and to predict how environmental changes will affect biodiversity in specific regions (van Erp *et al.* 2015).

To georeference textual documents such as Wikipedia articles, news articles, and social media posts with coordinates that best represent the content, various methods are employed (Stock 2018). While gazetteer-based approaches are commonly used in this context to detect the presence of place names (Karimzadeh *et al.* 2019), other methods, such as probabilistic language modelling techniques and discriminative classifiers, have also been utilised. Probabilistic language modelling techniques use generative models that estimate the likelihood of a textual document belonging to a particular region. Discriminative classifiers, such as logistic regression and multi-layer neural networks, classify documents into their most likely regions (Melo and Martins 2017). Furthermore, the introduction of Transformer models (Vaswani 2017) has spurred several research efforts using BERT-based models for tweet or tweet user geolocation prediction (Scherrer and Ljubešić 2020, Simanjuntak *et al.* 2022, Lutsai and Lampert 2023, Li *et al.* 2023a).

The often complex relative descriptions of localities where biological specimens are found present an additional challenge from standard document georeferencing

approaches mentioned above. As an example, manually georeferencing the locality '*C. 10 km N of Lake Wanaka, 1 km N of Makarora. near Pipson Creek*', which explains the location of a fungi specimen found in the Canterbury district of New Zealand, requires determining the coordinates of Lake Wanaka and the town of Makarora; interpreting the spatial relationship described by '*approximately 10 km north of Lake Wanaka*' and '*1 km north of Makarora*' to find the general vicinity; and finally, using the reference to '*near Pipson Creek*' to pinpoint the precise location. Thus, there is a need for more specialised automated georeferencing techniques. Additional examples of such descriptions are presented in Table 5.

Recently, the field of Natural Language Processing (NLP) has been transformed by the advent of Large Language Models (LLMs), demonstrating advanced functionalities and effective solutions. The vast size of LLMs, which typically include billions of parameters, enables them to discern complex patterns and subtleties in human language, resulting in outputs that are both highly precise and contextually appropriate (Fujiwara *et al.* 2024). Leveraging these abilities, LLMs have been successfully applied in various complex domains such as mathematical reasoning (Imani *et al.* 2023, Yu *et al.* 2023) and geospatial reasoning (Li *et al.* 2023b). They have often been found to outperform previous solutions based on traditional Machine Learning (ML) techniques such as Logistic Regression and BERT-based models (Karanikolas *et al.* 2023).

In this study, we introduce a novel approach for automatic georeferencing of biological collection data using LLMs. To the best of our knowledge, our study is the first to employ LLMs to georeference text data by predicting coordinates that best match the given locality descriptions. We conducted multiple initial experiments to identify the most suitable LLM and the prompting pattern for our use case. Subsequently, we performed supervised fine-tuning with quantized Low-Rank Adaptation (QLoRA) to adapt the selected LLM for the task of georeferencing. This was applied to multiple datasets containing data from different regions at varying granularities of area size. We also experimented with non-English data to demonstrate the model's linguistic generalizability. The results were benchmarked against several existing georeferencing baselines as well as commercial LLMs. In addition, we investigated the model's sensitivity to spatial information embedded in locality descriptions and explored its transfer learning capabilities across geographic regions.

To summarize, the contributions of this research are as follows:

- We demonstrate the effectiveness of applying LLMs to georeferencing locality descriptions with relative spatial relationships, and establish a strong benchmark that significantly outperforms the commonly used baselines.
- We investigate the optimal LLM prompting patterns for georeferencing use cases and introduce a new prompt that yields very good results for the considered task.
- We analyse the regional and linguistic generalizability of our LLM-based approach for the task of georeferencing.
- We demonstrate that our proposed model is more robust for non-English datasets compared to the baselines. For smaller regions, such as New Zealand, our model, fine-tuned with just 5000 samples, outperforms the baselines.

The remainder of this article is organized as follows. Section 2 describes related work. Section 3 describes the data we used and our methodology for automatic georeferencing. Section 4 discusses the results, and Section 5 concludes the paper with a look into future directions.

2. Related work

2.1. Georeferencing text data

The idea of georeferencing text documents or specific descriptions of localities and incidents is quite well established (Goodchild *et al.* 2004, Murphey *et al.* 2004, Doherty *et al.* 2011, Melo and Martins 2017). Gazetteer-based methods are commonly used to georeference text by detecting place names (Karimzadeh *et al.* 2019). In contrast, other techniques that consider all text, not just the place names, have also been applied, including probabilistic language models, discriminative classifiers, and hybrid approaches that combine these models with gazetteers (Melo and Martins 2017).

Gazetteer-based approaches (Karimzadeh *et al.* 2019) primarily focus on geoparsing, and have two steps; toponym recognition, to detect the presence of a location reference, and toponym resolution to disambiguate and hence determine coordinates (Wang and Hu 2019, Liu *et al.* 2022). Toponym recognition is mostly performed as a Named Entity Recognition (NER) task and the identified toponyms may then be resolved via gazetteer matching in association with a disambiguation process (Leidner *et al.* 2003, Li *et al.* 2003). However, gazetteers can sometimes be incomplete or outdated, particularly when dealing with historical place names, cultural variations, or locations that have undergone name changes over time, posing challenges for georeferencing based solely on them (Sharma *et al.* 2024). To address such limitations, some studies have proposed enhancements. For example, Lieberman *et al.* (2010) introduced audience-specific local lexicons to better capture regionally relevant toponyms that might be overlooked by global gazetteers.

For the task of georeferencing textual content such as tweets and Wikipedia articles, early research has predominantly utilised ML models, approaching georeferencing as a classification problem. This involves classifying location descriptions at the country and city levels. Probabilistic classifiers, such as n-gram models (Flatow *et al.* 2015, Iso *et al.* 2017) and discriminative classifiers [e.g. Support Vector Machines (SVM) and Gaussian Mixture Models (GMM)], have been employed in these studies (Priedhorsky *et al.* 2014, Liu and Inkpen 2015, Melo and Martins 2015). In addition to employing classification models, some studies have incorporated clustering techniques and feature selection strategies in combination with various models to georeference documents (Han *et al.* 2014, Laere *et al.* 2014). For example, Di Rocco *et al.* (2021) introduced Sherloc, a knowledge-based approach for subcity-level geolocation of tweets. Sherloc improves geolocation accuracy within a specified parent region by matching tweets to a geographic embedding constructed from known toponyms for that region, using clustering approaches to enhance precision.

More recent advancements in ML techniques have introduced Deep Learning models, such as Deep Neural Networks (DNN), Long Short-Term Memory (LSTM) networks, and Transformer-based models, for the task of georeferencing. Specifically,

georeferencing approaches utilizing pre-trained Transformer models such as BERT have demonstrated stronger performance compared to earlier ML systems. For instance, in the study by Simanjuntak *et al.* (2022), BERT significantly outperformed LSTM in predicting Twitter users' home locations. However, their most effective method involved aggregating all tweets from a single user to predict their location, a strategy that is potentially constrained by BERT's 512-token limit. Scherrer and Ljubešić (2020) demonstrated that BERT in regression mode can be effectively fine-tuned to predict geolocations from text. Edwards *et al.* (2025) built on the latter BERT-based method to georeference social media posts in a hybrid approach, in which the inferred location guided disambiguation of place names when present. Coordinates of the finest-grained place name then represented the tweet's location, otherwise, the transformer-predicted location was used. Lutsai and Lampert (2023) integrated BERT with two-dimensional GMMs to estimate locations as coordinate pairs. Nevertheless, in addition to adapting pre-trained language models to tweet content, most tweet georeferencing methods also incorporate tweet metadata and user network information as inputs for the model (Do *et al.* 2018, Miura *et al.* 2017, Lutsai and Lampert 2023).

When dealing with generic documents and social media content georeferencing, spatial relation terms are not frequently found, unlike in biological collection records. Equally, the methods mentioned above do not pay attention to spatial relationship terms that provide critical context by describing how one location is positioned relative to another. Such spatial relation terms can include prepositional phrases, verbs, and adverbs (Liao *et al.* 2022). Recent work by Li *et al.* (2021) proposes a geospatial semantic graph representation to better capture spatial dependencies in text, emphasising that modelling such relationships improves the interpretation of complex geographic descriptions. Moncla *et al.* (2014) also discuss the importance of spatial relations in estimating the spatial footprint of non-gazetteered place names. In contrast, their geocoding method relies heavily on the clustering of nearby known toponyms, which limits the potential to fully exploit spatial cues provided in the text.

Chen *et al.* (2018) have successfully leveraged spatial relations in georeferencing text documents. They convert raw text descriptions to place graphs as the model input and leverage spatial relation models for approximate locating and matching. However, this methodology is challenging for georeferencing larger datasets with multi-clausal location descriptions, as it assumes the prior existence of a place graph, the creation of which could be a complex task in itself.

This highlights that current georeferencing methods are not well-suited for text in which locations are described in relative terms rather than by absolute place names. These complex descriptions are common in the biological specimen domain, and hence, there is a need for more specialised methods tailored to this field, capable of handling complex relative locality descriptions.

2.2. Georeferencing biological specimen data

Among digitised biological specimen records, only a small portion of the data is georeferenced, whereas most records include a verbal locality description (Stock *et al.*

2023). Several methods have been developed for georeferencing these locality descriptions, including the point-radius method (Wieczorek *et al.* 2004), shape method, bounding box method, and probability method (Guo *et al.* 2008). These methods involve the time-consuming process of determining precise locations and calculating the associated uncertainty using GIS tools, maps, and aerial photography (Hackeloeer *et al.* 2014). Doherty *et al.* (2011) indicate that the processing time for the point-radius method ranges from 5 to 15 min, while the shape method can take between 15 and 90 min to georeference a single locality description. Given the sheer volume of records (billions worldwide), these semi-automated methods are not scalable or efficient.

Among studies to automate the georeferencing of biological specimen data, van Erp *et al.* (2015) developed a knowledge-driven, rule-based approach. Their method also produces a confidence score to indicate the certainty of the results. Scott *et al.* (2021) introduced an automated georeferencing method for Antarctic species by developing a pipeline that extracts and processes text from legacy documents. This method identifies species and toponyms within the documents and predicts species-toponym pairs representing actual geospatial relationships. It combines rules and dictionary-based species extraction with place name extraction methods, followed by tree-based classifiers to match species with their toponyms.

BioGeomancer (Guralnick *et al.* 2006) and GEOLocate[1] are applications developed to automatically georeference biological specimen data. BioGeomancer processes text, interprets it, queries gazetteers, and intersects spatial descriptions, ultimately delivering a standardized geographical reference complete with levels of uncertainty (van Erp *et al.* 2015). At the time of writing, the BioGeomancer application appears to be non-functional, and its website does not appear to be actively maintained.

On the other hand, GEOLocate is a fully operational stand-alone online application that translates textual locality descriptions associated with biodiversity collections into geographic coordinates. The GEOLocate algorithm begins by standardizing the input locality string into commonly understood terms (Rios and Bart 2014). It extracts details such as distances, compass directions, and significant geographic identifiers, which are used in lookups across various datasets such as place names, river miles, legal land descriptions, and highway-water body crossings to calculate geographic coordinates. Adjustments are made based on the parsed data. Further refinement of coordinates is achieved by scanning for water body names in the locality string and adjusting to the nearest point on the identified water body, enhancing accuracy for aquatic collections. The final coordinates are ranked, displayed digitally, and subjected to human verification or correction (Rios and Bart 2014). GEOLocate is primarily developed for georeferencing data from the USA, Canada, and Mexico.

Both GEOLocate and BioGeomancer were developed initially in the early 2000s. GeoPick (Marcer *et al.* 2023) is the latest addition to these georeferencing tools, developed by implementing the guidelines provided in Chapman and Wieczorek (2020). It offers a comprehensive guide for georeferencing biological collections. A common feature among these tools—GEOLocate, BioGeomancer, and GeoPick—is their implementation of the point-radius method. Introduced by Wieczorek *et al.* (2004), the

point-radius method does more than describing a location using a single coordinate pair; it indicates an area of uncertainty using a circle with a well-defined radius (Doherty *et al.* 2011). However, unlike its predecessors, GeoPick is not an automatic georeferencing tool; instead, it aids georeferencers in following standards and guidelines more effectively and in a user-friendly manner.

2.3. Large language models (LLMs)

The emergence of LLMs has brought significant advancements to the field of Natural Language Processing (NLP). LLMs, primarily based on the decoder Transformer architecture, have evolved from the development of the GPT (Generative Pre-trained Transformer) models. As a result of the widespread success of these models, encoder-based approaches, such as BERT and XLNET, as well as encoder-decoder-based models such as BART and T5, have become less prominent. Characterised by their large number of parameters and immense training datasets, LLMs have shown high proficiency in comprehending and generating human language (Karanikolas *et al.* 2023). Fine-tuning pre-trained LLMs via supervised learning has become the key to achieving high performance in various tasks (Ouyang *et al.* 2022, Ranjit *et al.* 2023, Singhal *et al.* 2023, Zhang *et al.* 2023a). Fine-tuning is the process of adapting a pre-trained LLM to a specific task or domain by continuing its training on a smaller, task-specific dataset. This approach leverages the model's existing knowledge while allowing it to specialise in a particular application (Ziegler *et al.* 2019). By adjusting the model's parameters based on task-specific labelled data, fine-tuning enhances performance and enables LLMs to generate more accurate and contextually relevant outputs (Wei *et al.* 2021). While GPT models (Achiam *et al.* 2023) are restricted to online inference and paid APIs, the state-of-the-art open source models such as Mistral (Jiang *et al.* 2023), Llama 1–4 (Touvron *et al.* 2023a, 2023b), GLM (Zeng *et al.* 2022), and Gemma (GemmaTeam *et al.* 2024) are often fine-tuned for downstream tasks, with accuracies comparable to those of the GPT models.

However, due to the computational costs and complexity of fine-tuning very large parameter spaces, deploying and adapting LLMs to specific tasks via fine-tuning is challenging (Fujiwara *et al.* 2024). To address this issue, several Parameter Efficient Fine-Tuning (PEFT) methods have been introduced (Lester *et al.* 2021, Liu *et al.* 2021, Zaken *et al.* 2021, Hu *et al.* 2022). Among those methods, Low-Rank Adaptation (LoRA) (Hu *et al.* 2022) has gained attention for its efficiency in fine-tuning by focusing on a limited set of parameters, thus lowering the total computational load. Taking one step further, QLoRA (Quantization and LoRA) presents an even more memory-efficient iteration of LoRA by quantizing the weights of the LoRA adapters to a lower precision (Dettmers *et al.* 2023).

Prompt engineering is an essential skill for effective communication with LLMs (White *et al.* 2024). A prompt is a set of instructions given to an LLM to tailor its behaviour and/or enhance its capabilities (Liu *et al.* 2023). Empirical evidence shows that higher-quality prompts result in better performance across various tasks (Wei *et al.* 2022, Yao *et al.* 2024). Zero-shot prompting is the simplest prompting technique. It involves giving the model the task in the form of a prompt without any prior

examples. Some of the prompting patterns discussed in the literature are the Persona pattern, Context Manager pattern, Thought Generation pattern, Template pattern, and Question Refinement pattern (Schulhoff *et al.* 2024, White *et al.* 2024).

The Persona pattern, which is a type of output customisation prompting, assigns an LLM a specific persona or role, guiding it to adopt a particular point of view when generating the output. This approach leverages the LLM's ability to simulate expert-level understanding, thereby enhancing its performance in complex domains that require specialised knowledge (Desmond and Brachman 2024, White *et al.* 2024). In contrast, the Context Manager pattern controls the contextual information, refining the response according to the given context.

Chain of Thought (CoT) prompting, which is a Thought Generation prompting pattern, has been identified as performing well in complex reasoning tasks (Wei *et al.* 2022). This pattern guides the model through a step-by-step reasoning process to reach a conclusion. The simplest form of CoT, zero-shot CoT essentially involves adding the phrase 'Let's think step by step' to the original prompt which has also shown better performance in reasoning tasks (Kojima *et al.* 2022). With each pattern having a different approach to define the task for the LLM, the best pattern depends on the task for which the LLM is used.

In the context of georeferencing biological specimen data, most existing technologies have not kept pace with these advancements in NLP. Several studies demonstrate the use of early transformer models, notably BERT, for geolocating social media data as discussed earlier (Scherrer and Ljubešić 2020, Simanjuntak *et al.* 2022, Lutsai and Lampert 2023). With modern LLMs becoming more powerful due to their decoder-based architecture, there is a motivation to apply them to the georeferencing task.

A few studies have utilized LLMs in related disciplines such as geospatial reasoning, geographic information systems, and geoscience, showing promising results (Mooney *et al.* 2023, Roberts *et al.* 2023, Li *et al.* 2023b, Zhang *et al.* 2023b, Deng *et al.* 2024, Yang *et al.* 2024). A recent study by Hu *et al.* (2024) has successfully employed LLMs for toponym resolution by fine-tuning the models to estimate the unambiguous references of toponyms. However, the authors acknowledge that LLMs alone are insufficient for comprehensive toponym resolution and incorporate additional geocoding services in their methodology to address inaccuracies in the LLM-generated results. Relatedly, Hu *et al.* (2023) propose a geo-knowledge-guided GPT approach for disaster response that injects domain-specific geographic knowledge into prompts to extract complete location descriptions from Hurricane Harvey tweets. Their method substantially outperforms off-the-shelf NER baselines (reported as $> 40\%$ improvement) and also improves markedly over default GPT models (reported as $\sim 76\%$ improvement), highlighting the practical value of explicit geographic guidance for strengthening LLM-based location understanding and extraction.

Bhandari *et al.* (2023) evaluated whether LLMs are geospatially knowledgeable through three experimental approaches. They assessed the geospatial awareness of LLMs by analysing the use of geospatial prepositions in sentences. The study found that LLMs were able to correctly arrange cities that are physically closer to each other when prepositions such as 'near' and 'close to' were used to indicate proximity.

Conversely, when the context involved the preposition 'far', indicating a distant location, the generated cities tended to be farther apart. The study concludes that LLMs hold significant potential for supporting human tasks in geospatial reasoning and analysis, especially with targeted fine-tuning tailored to specific use cases. However, to the best of our knowledge, no existing studies have applied LLMs for georeferencing tasks to obtain geographic coordinates that best match the textual locality descriptions provided.

3. Methodology

3.1. Data

In this study, we utilise multiple biological collection datasets obtained from the Global Biodiversity Information Facility[2] (GBIF). GBIF is the largest online biodiversity data network in the world, containing biological collection records from many institutions worldwide. We selected multiple datasets from GBIF (GBIF.org 2024a, 2024b, 2024c, 2024d), which contain records from the regions of the USA, New Zealand, Australia, and Mexico. The USA, New Zealand, and Australia datasets are in English, while the dataset from Mexico is in Spanish. Since datasets available through GBIF follow the same data standards, these datasets share a common structure and can be preprocessed using the same steps. An occurrence record from GBIF contains a wealth of information, including details about the location where the samples were collected, information about the collector, and specifics of the collection event, such as temperature and altitude. Among this information, we extracted the columns '*locality*', '*decimalLatitude*', '*decimalLongitude*', '*countryCode*', and '*stateProvince*', which are related to the location of the specimen found for our final dataset. '*locality*' is the text that should be georeferenced. '*decimalLatitude*' and '*decimalLongitude*' provide the original coordinates of the location described by '*locality*', and we treat them as ground truth for georeferencing. The country and state/province where the specimen was found were used as additional contextual information.

We extracted a subset of these datasets and applied several preprocessing steps: these included removing duplicate locality descriptions and eliminating records that do not have original coordinates. After preprocessing, we randomly selected ∼30,000 records from each dataset for our initial experiments. Subsequently, we experimented with different dataset sizes. Since the datasets include only records georeferenced with the coordinates of the specimen collection site, these coordinates serve as the ground truth for model training. A study by Yesson *et al.* (2007) reports that 83% of records in GBIF are annotated with accurate coordinates. The summary of the datasets used is provided in Table 1. Each country-specific dataset was divided into training

Table 1. Sources of datasets extracted from GBIF for fine-tuning LLMs across different regions.

Country	Source
New Zealand	Manaaki Whenua–Landcare Research
USA	New York Botanical Garden Herbarium
Australia	Australasian Virtual Herbarium (AVH)
Mexico	National Biological Collections of the Institute of Biology, UNAM

Table 2. Examples of prompts corresponding to different prompting patterns as adapted for the georeferencing task.

Prompt pattern	Prompt
Zero-shot	Accurately georeference the location provided in the Locality Description' below, expressing the coordinates in decimal degrees. Locality Description: 21 km west of Opotiki, south of Wainui Road
Zero-shot CoT	Accurately georeference the location provided in the 'Locality Description' below, expressing the coordinates in decimal degrees. Locality Description: 21 km west of Opotiki, south of Wainui Road. **Think step-by-step.**
Chain of thought	Task: Convert the Locality Description below into decimal degree coordinates Locality Description: 21 km west of Opotiki, south of Wainui Road. Steps to follow: Start by identifying the decimal coordinates of the key reference point. Then, convert these coordinates into eastings and northings for precise adjustment. Afterward, apply the specified distance and direction adjustments as per the locality description. Finally, convert the adjusted eastings and northings back into decimal degree coordinates for the final location.
Context control	Accurately georeference the location provided in the 'Locality Description' below, expressing the coordinates in decimal degrees. Context: This 'Locality Description' refers to a location in North Island, New Zealand. Locality Description: 21 km west of Opotiki, south of Wainui Road
Persona prompting	Act as a georeferencer and accurately georeference the location provided in the 'Locality Description' below, expressing the coordinates in decimal degrees. Locality Description: 21 km west of Opotiki, south of Wainui Road

(70%), validation (15%), and testing (15%) sets for model training, validation, and testing, respectively.

3.2. Prompt engineering

Since LLMs have not previously been used in georeferencing text data with geographical coordinates, the most effective prompt structure for the task was identified by experimenting with multiple prompting patterns. During this exploratory phase, we initially utilised ChatGPT-4 (OpenAI's[3] GPT-4 model) accessed via the ChatGPT[4] interface, chosen for its rapid inference capabilities and ease of access. We subsequently evaluated each prompting pattern using the latest GPT models, with the comparative results discussed in Section 4.2.2. The prompting patterns examined included Chain of Thought, Zero-Shot Chain of Thought, Context Manager, and Persona. These patterns were chosen for their applicability to the georeferencing context. Table 2 presents examples of how each prompting pattern is adapted to the georeferencing problem using a locality description.

The LLM successfully returned coordinate pairs in the majority of cases, producing generally acceptable results. Notably, this was the case even in Zero-Shot prompting, where the model was given the task without any prior examples or additional instructions (see Table 2). In these cases, the LLM claimed to produce these coordinates by calculating offsets based on the approximate locations of the place names mentioned in the text (see Figure A1 in Appendix A). This suggests that LLMs already possess some understanding of the task at hand. However, when the LLM lacks knowledge of the geographical coordinates for place names mentioned in the description, it did not generate coordinates but instead provided follow-up instructions.

Task: Accurately georeference the location provided in the 'Locality Description' below, expressing the coordinates in decimal degrees.

Context: This 'Locality Description' refers to a location in Nelson Land District, New Zealand.

Locality Description: About 300m downstream of bridge over Mangatini Stream, Stockton Plateau.

Geographic Coordinates: Latitude: -41.652 Longitude: 171.896

Figure 1. Sample prompt used for fine-tuning the LLM. Both the training and validation datasets were structured using this pattern, with the original coordinates serving as the ground truth for supervised fine-tuning of the LLM. The test data was similarly formatted but without the inclusion of the coordinates.

Motivated by the success of CoT prompting in various complex tasks involving reasoning, we constructed step-by-step instructions for the LLM to follow in solving the problem of georeferencing. This approach generated a coordinate pair to represent the geographical location described, adhering to the provided instructions. Nonetheless, developing a set of instructions that would enable the LLM to derive accurate results proved challenging, as manual georeferencing is typically a complex process that involves analysing maps, aerial imagery, and various spatial data layers to accurately pinpoint locations.

Following the Context Manager pattern, we conducted an experiment where we specified the country and state or province in which the specimen was found, thereby constraining the output to that specific region. This approach produced more reliable results by restricting outputs to the relevant geographic region. We also explored the Persona pattern, where the LLM was instructed to assume the role of a georeferencer to derive coordinates from locality descriptions. Although this method aimed to leverage the LLM's ability to mimic expert georeferencing skills, the results were less effective. The Persona pattern mostly led the LLM to explain the steps a georeferencer would take, rather than generating the coordinates independently.

Through experimentation with various locality descriptions under different prompting patterns, we determined that the Context Manager pattern is the optimal prompt for our use case. This pattern is not only straightforward but also ensures that the produced coordinates accurately align with the specified geographic context. We then formatted the datasets according to the Context Manager pattern. In the training and validation prompts, the original coordinates are included as ground truth data, as illustrated in Figure 1.

3.3. Fine-tuning LLM

Fine-tuning an LLM typically includes multiple steps. The process begins with preparing the dataset, followed by selecting and loading a pre-trained model. Before training, the text data must undergo tokenization, where it is broken into smaller units called tokens (which can be words, subwords, or characters) and then converted into numerical representations that the model can process. Tokenized data is used to train
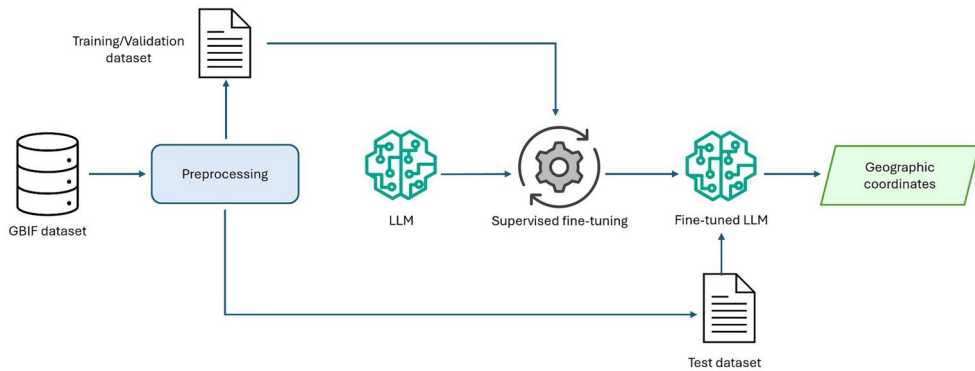
Figure 2. Framework of fine-tuning an LLM for georeferencing biological specimen collection records.

the LLM, during which the LLM's parameters are adjusted, performance is monitored, and iterative refinements are made to improve accuracy. The steps we followed in this process are summarized in Figure 2.

To select the most suitable LLM for our task, we conducted initial experiments using several open-source pre-trained LLMs. 4-bit quantization and Low-Rank Adaptation (QLoRA) were applied to the base LLM to efficiently manage the large size and complexity of the LLM while maintaining high performance. This approach enabled us to fine-tune the model using a single NVIDIA A100 24 GB GPU, eliminating the need for large-scale computing infrastructure and making the solution more accessible and practical for scaling georeferencing tasks.

3.4. Experiment setup

Initial experiments were conducted on the base version of the following open-source pre-trained LLMs: Llama2-7B, Llama3-7B, Gemma-7B, and Mistral-7B to identify the model that best matches the use case (we also tested instruct models, but they performed less well than the base models for this task). We chose open-source LLMs due to their accessibility, flexibility for customisation, and cost-efficiency compared to proprietary LLMs such as ChatGPT, aligning with our long-term objective of releasing a open-source tool for biological specimen georeferencing. These experiments identified the Mistral-7B base model as the most suitable for the task. For brevity, we focus on reporting the results of the selected model.

The key hyperparameters for model training were carefully selected to optimize performance while preventing overfitting, based on experiments with multiple settings. The learning rate was set to 2e-4, which offers a balance between convergence stability and training speed, mitigating divergence or oscillatory behaviour. The batch size was set to 32, a choice that offered stable gradient estimates without exceeding memory limits. For efficient fine-tuning, LoRA parameters were configured with a LoRA rank of 32 and LoRA alpha of 64. The rank specifies the dimensionality of the low-rank adaptation matrices, controlling their expressive capacity, while alpha rescales the output of the low-rank matrices before combining them with the frozen base weights,

effectively controlling the contribution of the LoRA adaptation (Hu *et al.* 2022). The training process consisted of three epochs. The low-rank matrices were iteratively updated at each batch step within each epoch using gradients calculated from the loss function.

Checkpoints were established at regular intervals throughout the training process to facilitate continuous monitoring and potential retrieval of intermediate models. This fine-tuning process was performed separately for each region-specific dataset, resulting in a dedicated LLM fine-tuned for georeferencing within each region. The final models were saved and subsequently used for inference on the test datasets. Performance metrics were calculated to evaluate each model's accuracy in georeferencing locality descriptions.

4. Evaluation

4.1. Baseline systems

Given the lack of publicly accessible and functional methods for georeferencing biological collection data, our selection of benchmarks was limited. While van Erp *et al.* (2015) offer the most recent study on georeferencing biological specimen data, we were unable to use their methodology as a baseline because their application, 'GeoImp', is no longer available. Furthermore, the latest tweet georeferencing methodologies are also inapplicable to the domain of biological specimen data georeferencing, as they rely on metadata and social media user network data, in addition to the tweet text.

As a result, we compared the performance of our models against two georeferencing baselines: GEOLocate and a gazetteer-matching algorithm. GEOLocate is the only currently available tool for georeferencing locality descriptions of biodiversity collection records. For a given description, GEOLocate can return multiple coordinate pairs as potential geolocations, and these coordinates are ranked according to their probability of being correct (Rios and Bart 2014). For our analysis, we selected the coordinate pair ranked as the most likely.

Since many methods for georeferencing textual data primarily focus on identifying and geocoding place names through gazetteer-matching approaches, we implemented a gazetteer-matching algorithm as our second baseline (see Appendix B). This algorithm can be regarded as analogous to the spatial minimality disambiguation method of Leidner *et al.* (2003). It involves the steps of place name recognition and resolution. To improve the accuracy of place name identification, we fine-tuned the spaCy[5] model using 50 annotated locality descriptions of our data. The identified place names were then geocoded using a gazetteer-matching process. We experimented with both the GeoNames[6] and Nominatim[7] gazetteers to determine the most suitable option. During geocoding, we specified the country and state or province to initially disambiguate the place names. Given that a single locality description can contain multiple place names, it was necessary to disambiguate these names and identify the coordinates that best represent the overall locality description. To achieve this, we conceptually plotted all possible locations of the identified place names using the DBSCAN algorithm (Ester *et al.* 1996) and identified the cluster with the highest

Table 3. Comparison of the results from the fine-tuned LLMs with baseline systems across regional datasets.

Dataset	No of records	Model	Accuracy@10 km	Accuracy@1 km	Med SAE	Mean SAE
New Zealand	29,024	GEOLocate	45.39%	10.45%	12.24 km	129.62 km
		Gazetteer-based	23.24%	5.30%	35.96 km	117.91 km
		Fine-tuned Mistral 7B	**70.43%**	**25.36%**	**3.55 km**	**41.95 km**
USA	29,566	GEOLocate	**56.40%**	**26.82%**	**5 km**	**62.38 km**
		Gazetteer-based	12.81%	3.47%	584 km	1062 km
		Fine-tuned Mistral 7B	35.24%	7.40%	19.18 km	63.9 km
Australia	29,566	GEOLocate	41.89%	6.85%	16.28 km	237 km
		Gazetteer-based	16.16%	1.91%	18.25 km	345.92 km
		Fine-tuned Mistral 7B	**53.82%**	**9.80%**	**8.52 km**	**51.28 km**
New York (USA)	20,962	GEOLocate	55.73%	32%	3.52 km	79 km
		Gazetteer-based	31.98%	9.63%	11.82 km	488 km
		Fine-tuned Mistral 7B	**84.89%**	**66.71%**	**0.083 km**	**17.59 km**
California (USA)	12,441	GEOLocate	50.78%	20.78%	8.94 km	88 km
		Gazetteer-based	15.48%	2.09%	24.17 km	158.24 km
		Fine-tuned Mistral 7B	**66.26%**	**28.70%**	**4.17 km**	**40.20 km**
Victoria (Australia)	29,568	GEOLocate	48.82%	8.22%	10.73 km	169.44 km
		Gazetteer-based	20.26%	3.09%	12.76 km	264.11 km
		Fine-tuned Mistral 7B	**81.76%**	**22.79%**	**2.94 km**	**11.97 km**
Mexico	32,116	GEOLocate	29.85%	10.43%	54 km	251 km
		Gazetteer-based	15.94%	2.71%	82.66 km	313.11 km
		Fine-tuned Mistral 7B	**75.82%**	**51.41%**	**0.89 km**	**20km**

All datasets are in English, except for Mexico, which is in Spanish. The number of records shown refers to the entire dataset, which was subsequently divided into 70% for training, 15% for validation, and 15% for testing. Bold values indicate the best performing results for each dataset.

concentration of points, assuming this cluster contained most of the place names mentioned in the description. To determine the final coordinate pair, we calculated the mean latitude and mean longitude of the points within the identified cluster.

In addition to these baselines, we also evaluated our fine-tuned LLM against three prominent GPT models, GPT-5, GPT-4.1, and GPT-4o, to assess its relative performance in the georeferencing task.

4.2. Results and discussion

The mean and median of the SAE (Simple Accuracy Error) and the percentages of predicted locations that are within a 10 km and 1 km radius of the actual locations were used to evaluate the performance. The error was calculated by measuring the distance between the predicted location and the original location, utilising the Haversine formula. Table 3 shows a comparison of the performance metrics between the fine-tuned Mistral model and baselines for each dataset.

4.2.1. Performance across different regions

The fine-tuned LLM for the New Zealand region outperformed all baselines, correctly predicting 70.43% of localities within a 10 km radius, 25.36% within a 1 km radius, and 7.71% within 100 m. It also achieved a median SAE of 3.55 km and a mean SAE of 41.95 km. In comparison, GEOLocate reached only 45.39% within 10 km and 10.45% within 1 km, with substantially higher error values, while the gazetteer method performed worst overall.

For the Australian dataset, the fine-tuned LLM also surpassed the baselines, predicting 53.82% of records within 10 km and 9.8% within 1 km, with a median SAE of 8.52 km and a mean SAE of 51.28 km. While the performance did not reach the level achieved in New Zealand, the results for Australia were still clearly positive. In contrast, for the USA dataset, GEOLocate outperformed our LLM. GEOLocate predicted 26.82% of localities within a 1 km radius, whereas our LLM predicted only 7.40% within the same radius. The median and mean SAE given by GEOLocate also outperformed our model. The higher performance for the USA with GEOLocate might relate to the fact that GEOLocate is described as intended for the USA, Canada, and Mexico regions (Rios and Bart 2014). However, as demonstrated subsequently, our approach was found to be superior when applied to data for Mexico and for two states of the USA.

Based on these results, we hypothesize that the lower performance in the USA dataset and the slightly reduced performance in the Australian dataset are due to the lower density of records relative to their large geographic extents. With fewer records per unit area, the training data may not provide sufficient coverage for the model to effectively learn relevant place names, recurring locality descriptions, and spatial language. If the availability of these features is a function of geographic area, then for a fixed number of training records, larger regions will yield fewer such signals per unit area, thereby reducing model effectiveness. To validate our hypothesis regarding the influence of record density and geographic extent, we designed two complementary experiments.

In the first experiment, we expanded the USA dataset to ∼150,000 records and fine-tuned the LLM on 70% of these data. The LLM demonstrated a significant improvement over the smaller dataset, accurately georeferencing 61% of localities within a 10 km radius and 17.87% within a 1 km radius. It achieved a median SAE of 5.91 km and a mean SAE of 36.48 km, while GEOLocate's performance remained consistent with that of the smaller dataset. These results indicate that increasing dataset size, and thereby effective record density, enhances the model's effectiveness in interpreting locality descriptions and resolving place names and associated spatial semantics.

In the second experiment, we restricted training to smaller geographic regions using state-specific data. From the USA dataset, we fine-tuned separate LLMs with 20,962 records from New York and 12,441 from California. As shown in Table 3, the New York model predicted 84.86% of coordinates within 10 km and 66.71% within 1 km, while the California model achieved 66.26% within 10 km and 28.7% within 1 km, both outperforming the baselines. A similar pattern was observed for Australia, where a Victoria-specific model (29,568 records) predicted 81% of localities within 10 km and 22.79% within 1 km.

Together, these experiments support our hypothesis: the methodology is more effective in smaller regions with higher record density, and for larger regions, achieving comparable accuracy requires substantially more data.

Building on these region-specific results, we next examined whether the model's effectiveness extends across languages, since LLMs are trained on multilingual data (Thellmann et al. 2024, Qin et al. 2025). Given the global need for georeferencing biological specimen collection data, we were particularly interested in testing the linguistic generalizability of the LLM. Thus, we extended our experiments to test the model with Spanish-language data from the Mexico region. Our model, fine-tuned for the

Table 4. Comparison of results from the fine-tuned Mistral-7B model and GPT models on the New Zealand test set.

LLM	Prompt pattern	Accuracy@10 km	Accuracy@1 km	Med SAE	Mean SAE
GPT-4o	Zero-shot	52%	6.8%	9.12 km	476 km
	Zero-shot CoT	48.4%	6.8%	10.26 km	373.76 km
	CoT	48%	6.2%	10.69 km	501.76 km
	Context control	54.8%	8.4%	8.69 km	**45.33 km**
	Persona	48.2%	6.4%	10.59 km	452.79 km
GPT-4.1	Zero-shot	45.4%	6.2%	11.72 km	399.14 km
	Zero-shot CoT	45%	6.4%	11.42 km	340.63 km
	CoT	48.6%	6.4%	11.42 km	340.63 km
	Context control	50.8%	7%	9.9 km	51.18 km
	Persona	47.6%	7.8%	10.92 km	405.33 km
GPT-5	Zero-shot	43.57%	6.2%	13.26 km	333km
	Zero-shot CoT	59.4%	10%	7.17 km	159.13 km
	CoT	57.6%	9.6%	7.27 km	139.59 km
	Context control	57.4%	8.8%	7.81 km	50.94 km
	Persona	36.27%	5.4%	18.41 km	416.76 km
Mistral 7B	Context control	5.23%	0.48%	49.78 km	1120 km
Fine-tuned Mistral 7B	Context control	**71**%	**31**%	**3.81 km**	52.01 km

Note: Bold values indicate the best result for each metric.

Mexico region with Spanish-language data, returned excellent results, predicting 75.82% of coordinates within 10 km from the actual location and 51.41% of localities within 1 km from the actual location. This model also achieved low error distances, with a median SAE of 0.89 km and a mean SAE of 20 km. GEOLocate tool, which is said to have been optimised for the Mexico region, only predicted 29% records within a 10 km radius. These results suggest that our methodology effectively leverages the multilingual inference capability of LLMs, making it generalizable both regionally and linguistically.

4.2.2. Performance compared to GPT models

In addition to comparing our fine-tuned LLMs with existing georeferencing baselines, we conducted experiments using the latest and most prominent GPT models [GPT-5 (OpenAI 2025b), GPT-4.1 (OpenAI 2025a), and GPT-4o (OpenAI 2024)] with the prompting strategies shown in Table 2. These experiments were performed via the OpenAI API.[8] Due to the proprietary nature and high cost associated with these services, we evaluated only a random subset of 500 samples from the New Zealand test dataset, rather than the full test datasets for each region. The corresponding results are presented in Table 4. We have also included the results from the Mistral-7B model without fine-tuning, as well as our fine-tuned Mistral-7B model.

Our fine-tuned LLM significantly outperformed all GPT variants, including the newest GPT-5 model. This suggests that the proprietary models, regardless of how intelligent they have become or how much their reasoning capabilities have improved, are currently limited in their ability to perform the type of spatial reasoning required for accurate georeferencing. The substantial performance margin highlights the effectiveness of our domain-specific fine-tuning approach. However, it is worth noting that GPT models with the context control prompt achieved the best mean SAE values overall, with the lowest recorded by GPT-4o.

Table 5. Examples of locality descriptions with different lengths (in characters).

Length	Locality description
<30	Off Great Barrier Island
	near Gulf Harbour
	Beach near Devonport
30–60	Three Kings Group, West Island. Off W side of S summit
	Stream at Te Moori, 3 miles south of Kaeo, Northland
	Jacko's pond, 3 km south of Coromandel
60–90	Little Barrier Island, Awaroa Stream, about 150 m from the sea
	Gully facing sea, on bluffs opposite cottage, 2 km southwest of Kaihoka Lakes, Nelson
	Milford, near beach at end of Milford Road, Waitemata County, North Island
90–120	South Island, Fiordland National Park, W of Homer Tunnel, 9 mi SE of Milford Sound on Hw. 94
	Central Volcanic Plateau: c. 2 km S of National Park, forest margin near grazed meadow 300 m W of Mountain Hights Lodge.
	South Island, Westland, 0.6 miles W. of Rahu Saddle on Rte. 7, Terrestrial on roadside bank
>120	Knightia Railway Reserve and adjoining grazed bush containing abundant moss-covered small shrubs, on Morrison Farm about 4 km by road NW of Mataroa, itself on Mataroa Rd NW of Taihape
	Nelson Lakes National Park, St Arnaud, picnic/camping area on lake shore at Kerr Bay, about 100 m towards DoC Headquarters from Black Stream bridge
	Farm trail north side of Parau Stream, 200 yards (182 m) above bridge across tributary, W aimana Valley, inland Bay of Plenty, North Island, New Zealand.

Table 6. Comparison of prediction accuracy of the fine-tuned LLM by description length in the New Zealand test set.

Length in number of characters	Acc@1 km	Acc@10 km	Median SAE	Mean SAE
<30	4.07%	48.54%	11.52 km	55.85 km
30–60	17.60%	59.70%	6 km	66.45 km
60–90	15.61%	67.81%	5.03 km	26.37 km
90–120	20.55%	72.59%	3.78 km	**24.27 km**
More than 120	**32.01**%	**76.77**%	**2.40 km**	25.16 km

Note: Bold values indicate the best result for each metric.

Moreover, these experiments highlight the importance of prompt design. As shown in Table 4, the context control prompting pattern consistently yielded the best performance for both GPT-4.1 and GPT-4o. However, with the newest GPT-5 model, the Zero-shot CoT prompts produced slightly better results, likely due to the enhanced reasoning capabilities (*GPT-5 thinking*) introduced. Overall, these findings validate our prompting strategy.

4.2.3. Effect of locality description length on performance

Another aspect we explored in our research was the impact of the character length of a locality description in a collection record on LLM's performance. In our dataset, we observed that longer descriptions tend to be more complex, often detailing the location of a specimen in relation to multiple distances and directions, referencing various locations. Examples of locality descriptions categorised by length are presented in Table 5.

Table 6 presents a comparison of the results of our New Zealand fine-tuned LLM (reported previously in Table 3), grouped by description length. There is an incremental improvement in the accuracy of the prediction when the length of the description increases. The bar chart in Figure 3 further complements our observation. To
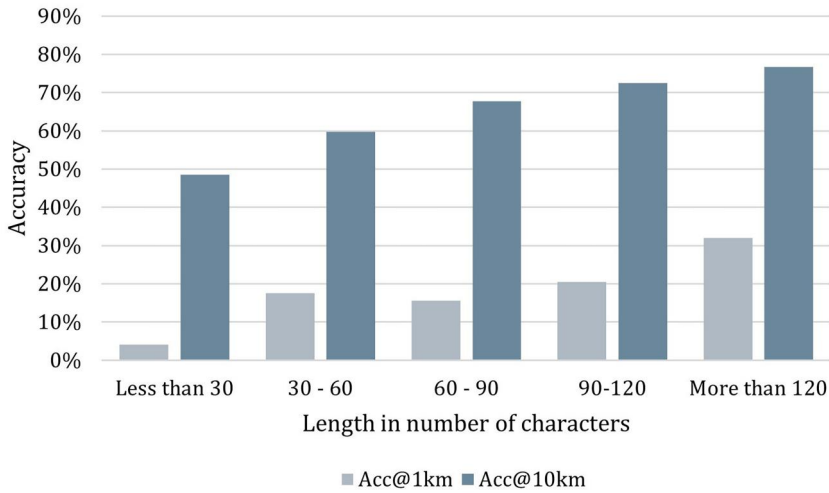
Figure 3. Distribution of accuracy in error distance by the description length for the New Zealand dataset.
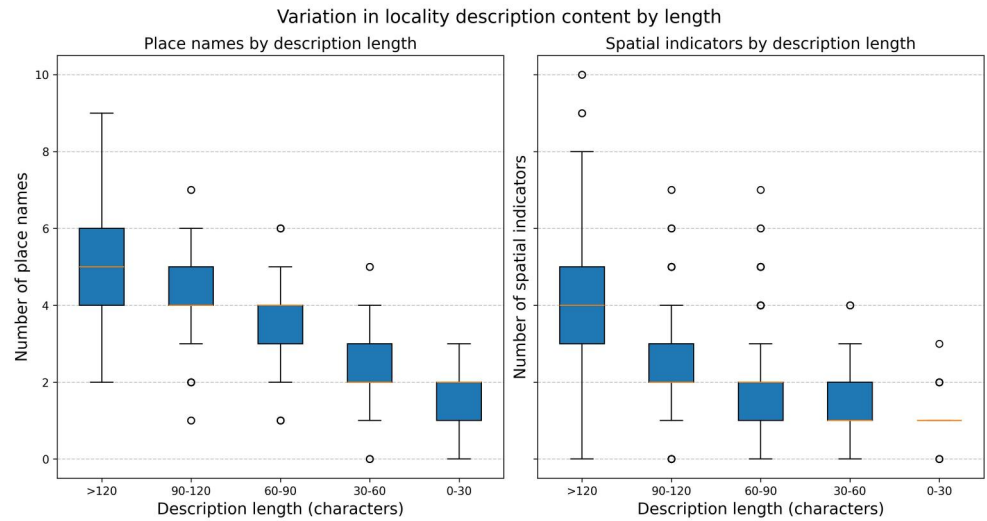

Figure 4. Distribution of locality description features by length: (a) number of place names and (b) number of spatial indicators.

investigate the statistical basis of this pattern, we analysed the correlation between error distance and description length using Spearman's rank correlation. This method was chosen because it does not assume linearity or normally distributed variables and is robust to outliers. The analysis revealed a moderately negative correlation ($\rho = -0.19$, $p \ll 0.001$), indicating that longer descriptions are associated with reduced error, although the modest effect size suggests that description length may be only one of several factors influencing prediction accuracy.

To clarify the source of these gains, we examined the description content by length. We looked into the place names and spatial indicators present in the locality descriptions. Spatial indicators are textual elements that express geographic
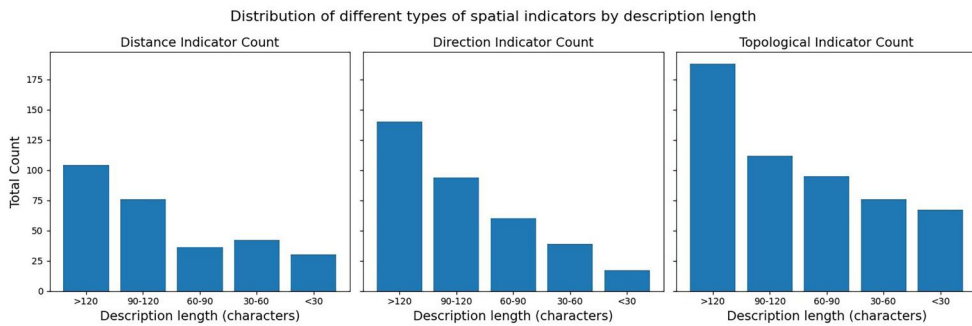
Figure 5. Distribution of spatial indicators (distance, direction, and topological) as a function of locality description length.

relationships, providing cues to locate a place in reference to other geographic entities. For example, in the locality description '*10 km north of Lake Wanaka, 1 km north of Makarora, near Pipson Creek*', there are three place names ('*Lake Wanaka*', '*Makarora*', '*Pipson Creek*') that act as reference points, while the three spatial indicators ('*10 km north of*', '*1 km north of*', '*near*') specify the relative positions needed to infer the exact location. Longer localities contained more place names and spatial indicators as revealed by the box-plots in Figure 4. Correlation analysis confirmed that the number of place names was moderately and negatively associated with error distance (Spearman $\rho = $ n0.16, $p \ll 0.001$), indicating that descriptions with more toponyms tend to yield more accurate predictions.

We further disaggregated spatial indicators into three types: directional (e.g. '*north of*', '*SE side of*', '*west-facing*'), distance (both quantitative, e.g. '*5 km N of Wellington*', and qualitative, e.g. '*near the river*'), and topological (e.g. '*on stone*', '*at the base*'). As shown by the box-plots in Figure 5, directional and distance cues increased sharply with description length, particularly beyond 90 characters, while topological indicators remained relatively frequent across all bins but peaked in the longest descriptions. These findings explain the observed performance gains: richer descriptions do not simply add verbosity but supply both a higher density of place names and a greater variety of spatial relations, which collectively improve LLM inference.

Together, these analyses demonstrate that description length is a meaningful predictor of georeferencing success, not because of length itself but because longer records tend to encode more place names and spatial relationships. Our results suggest that the fine-tuned LLM effectively leverages this additional context to reduce spatial error, underscoring the importance of detailed specimen metadata for accurate automated georeferencing. This interpretation is consistent with Ji *et al.* (2025), who show that LLMs can interpret formal topological relations with moderate accuracy, but frequently confuse conceptually adjacent predicates, indicating that relying on topological cues alone to describe a location can remain ambiguous. In our setting, longer specimen records mitigate this ambiguity by providing more place anchors and stronger directional/distance constraints, enabling more precise coordinate inference.

Table 7 lists the ten localities where GEOLocate exhibited the highest georeferencing errors in the California state dataset. These errors are measured as the distance between the predicted and actual locations. The table compares these errors with

Table 7. Georeferencing error comparison between GEOLocate and the fine-tuned Mistral 7B LLM for the ten worst-performing localities by GEOLocate in the California state dataset.

Locality description	Error with GEOLocate	Error with fine-tuned Mistral 7B
Amargosa Desert, along dirt rd sw. of Calif.-Nev. state line from Calif. Hwy 127 toward abandoned Tonopah & Tidewater Railroad. Drainage Basin: Amargosa; U. S. Atomic Energy Commission's Nevada Test Site and vicinity.	1891.73 km	49.60 km
Head of Etna Creek, Marble Mountains	997.35 km	24.39 km
Cleveland National Forest, Black Canyon Rd, 6 km N of Ramona	986.59 km	6.60 km
Cleveland National Forest, Palomar district, Black Canyon Rd, 7 km S of junction with Rd 12S05	982.91 km	27.92 km
Vicinity of Leonard Hot Spring, E side of Surprise Valley.	971.70 km	26.10 km
Mohave Desert, northeast of the San Bernardino Mountains, 9 airline miles north-northeast of junction of Calif. Highway 62 and highway to Lucerne Valley (Yucca Valley town).	970.43 km	915.58 km
Pyramid Point, forming a distinct zone in the middle zone of the littoral region, on rocks	948.77 km	0.02 km
Vail Lake area, lower slopes of Agua Tibia Mtn. along Hwy 79, E of USFWS Dripping Springs Guard Sta.	937.70 km	0.48 km
Just S.W. of sand hills, 5 mi. N.E. of Glamis.	887.96 km	5.78 km
927 Candlelight Place, adjacent to N terminus of Cass St	842.72 km	0.78 km

those produced by our fine-tuned LLM for the same localities. Notably, the localities that produced the poorest results with GEOLocate are characterized by long descriptions containing multiple spatial terms and place names. In 9 out of these 10 instances, our model achieved significantly lower error margins. This finding further demonstrates that our method provides more accurate results in handling complex locality descriptions compared to current georeferencing approaches.

4.2.4. Effect of prompt wording variations on fine-tuning performance

As detailed in Section 3.2 and Table 2, we initially experimented with several prompt patterns and identified the Context Control prompt pattern as the most appropriate for our use case. Accordingly, we used this pattern to structure the input data for training the fine-tuned LLMs. To explore whether the specific wording of prompts within this pattern influences the model performance of the fine-tuned LLM, we further extended our experiments to test different phrasings of the same Context Manager pattern. We used the New Zealand dataset (29,024 records) and compared the outcomes of the original prompt with two alternative phrasings. The results of these experiments are presented in Table 8, where we observed minor differences in performance across the variations.

The first prompt represents the original wording used in our experiments. The second version introduces a slight modification to the instructional phrasing. The third prompt combines the Context Manager pattern with elements of the Persona pattern by explicitly defining the user's persona alongside the context. Among these, the latter two variations showed a slight improvement in results. Accuracy within 10 km ranged only from 70.43 to 71.72%, and accuracy within 1 km varied between 25.36 and 26.23% across the three prompt versions. Median and mean errors likewise differed only marginally (3.38–3.55 and 40–42.38 km, respectively). These results suggest that while prompt wording may have a minor influence on performance, it is not a dominant factor in the context of our fine-tuning setup.

Table 8. Comparison of results from variations of context manager pattern prompts for the New Zealand dataset.

Sample prompt	Acc@10 km	Acc@1 km	Median	Mean
Accurately georeference the location provided in the 'Locality Description' below, expressing the coordinates in decimal degrees. **Context**: This 'Locality Description' refers to a location in Canterbury Land District, New Zealand. **Locality description**: 'Inverary', Mt Somers, Mid Canterbury, above Blondin Stream.	70.43%	25.36%	3.55 km	41.95 km
Determine the precise geographic coordinates for the location described in the 'Locality Description' below, providing the latitude and longitude in decimal degrees. **Context**: The described locality is situated in Canterbury Land District, New Zealand. **Locality description**: 'Inverary', Mt Somers, Mid Canterbury, above Blondin Stream.	**71.72**%	26.05%	**3.38 km**	42.38 km
You are a GIS specialist verifying geographic data for locations in New Zealand. Your task is to determine the precise latitude and longitude (in decimal degrees) for the locality described below. **Region**: Canterbury, New Zealand **Locality description**: 'Inverary', Mt Somers, Mid Canterbury, above Blondin Stream.	71.12%	**26.23**%	3.48 km	**40 km**

Note: Bold values indicate the best result for each metric.

4.2.5. Evaluation of model sensitivity to spatial distance information

Many descriptions in our dataset include quantitative spatial distance values referenced to a place name, such as '*200 metres west of Mangatoetoe Stream*' or '*6 km SSE of Westport*'. To assess whether our fine-tuned LLM makes meaningful use of these numerical distance references in prediction, we conducted a targeted evaluation.

We randomly selected 100 records from the New Zealand test dataset and identified 59 descriptions containing explicit quantitative spatial information. These distance values were then manually removed from the text, with care taken to preserve the overall semantic meaning and spatial context. For example, the description '*30 miles S of Auckland City*' was revised to '*S of Auckland City*' by removing the quantitative distance '*30 miles*'.

We evaluated model performance on both the original and modified versions of these records using GPT-4o and our Mistral 7B model fine-tuned for the New Zealand region. Results are illustrated in the bar chart in Figure 6. Interestingly, GPT-4o exhibited unexpected behaviour, its prediction accuracy was higher when the distance values were removed. In contrast, the fine-tuned Mistral model behaved as expected, showing a slight decline in accuracy when distance expressions were removed. This suggests that the fine-tuned LLM is more attuned to leveraging such spatial cues during prediction.

Despite the use of a smaller dataset, these results indicate that our fine-tuned LLM is capable of interpreting quantitative spatial cues, reinforcing its potential for accurate georeferencing of biodiversity collection records.

4.2.6. Analysis of transfer learning capabilities of LLMs

To evaluate the transfer learning (TL) capabilities of our approach, we tested whether a model fine-tuned on data from one region could be applied to another, and whether combining or adapting datasets could improve performance.
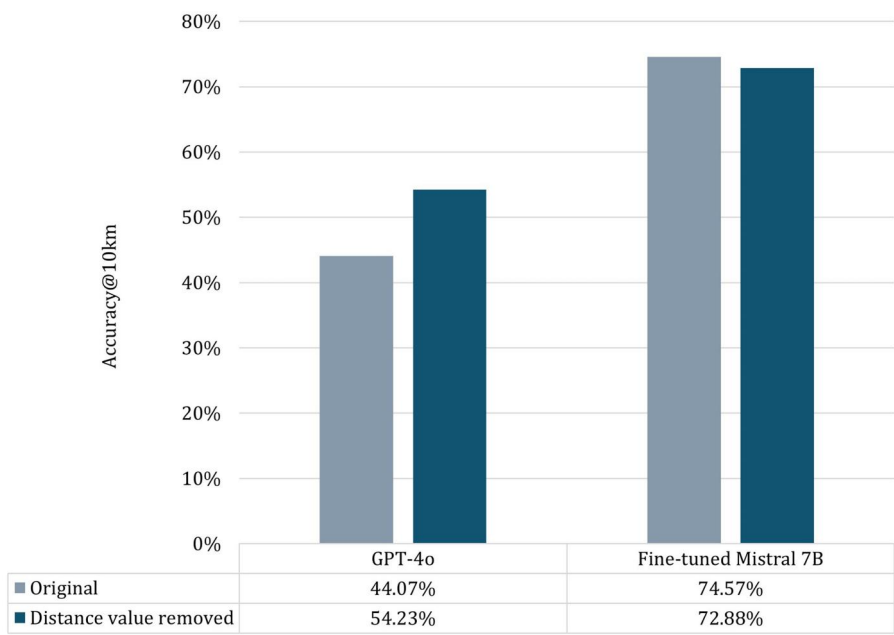
Figure 6. Comparison of sensitivity to spatial distance information between GPT-4o and the fine-tuned Mistral 7B models in predicting coordinates, evaluated on a subset of the New Zealand test set.

Table 9. Performance on the New Zealand test set across models fine-tuned with different multi-region training data compositions.

Model	AUS training data	NZ training data	Mean	Median	Within 10 km	Within 1 km
GeoLocate	–	–	**129.62 km**	**12.24 km**	**10.45**%	**45.39**%
Mistral 7B	100% (20,697)	0 (0)	1314.3 km	1706 km	2.30%	0.50%
	100% (20,697)	20% (4063)	57.27 km	10.39 km	48.56%	8.10%
	50% (10,348)	50% (10,158)	47.9 km	5.71 km	62.83%	15.18%
	0% (0)	100% (20,318)	41.95 km	3.55 km	70.43%	25.36%
	20% (4139)	100% (20,318)	**39 km**	**3.41 km**	**70.89**%	**26.07**%

Training dataset proportions are relative to the original training data splits (70%) of each dataset. Bold values indicate the best result for each metric.

We first tested whether a model fine-tuned exclusively on Australian data could be applied directly to New Zealand. The model was able to generate coordinates, but the predictions exhibited very high error margins (see Table 9). This indicates that while training on data from another country may help the LLM capture the general semantics of the task, it is insufficient for accurate coordinate prediction without region-specific information.

We next evaluated multi-region training configurations combining different proportions of Australian and New Zealand training data (see Table 9) and testing on the New Zealand test dataset. Adding 20% New Zealand training data to an Australia-fine-tuned LLM reduced the mean error to 57.27 km, performing well above the baseline (GEOLocate). A balanced dataset (50% AUS/50% NZ) lowered it further to 47.9 km mean error. Training with predominantly New Zealand data (20% AUS/100% NZ) achieved the best result of 39 km, only slightly better than New Zealand–only training
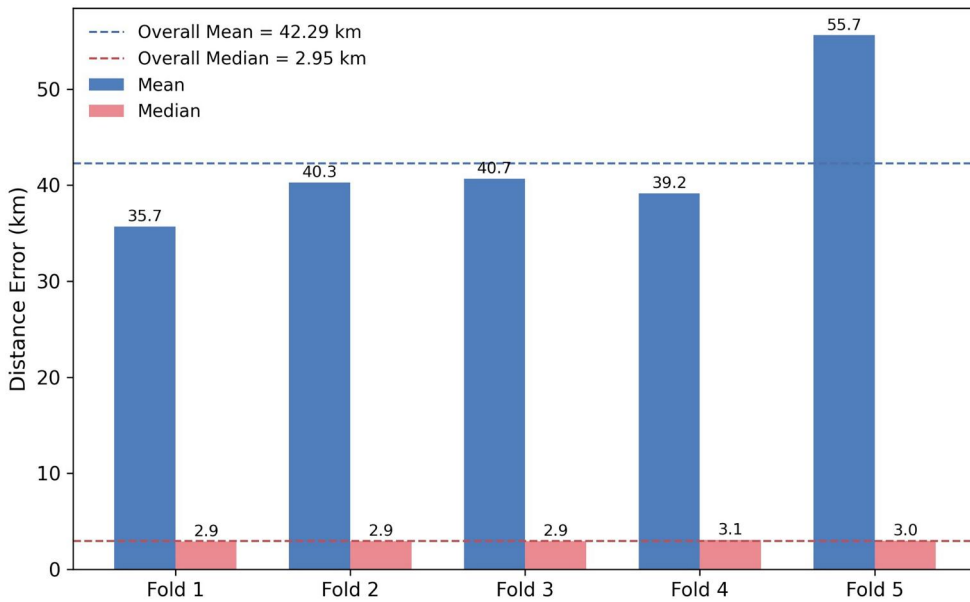
Figure 7. Cross-validation results of the fine-tuned Mistral-7B model for New Zealand.

(41.95 km). These results suggest that modest amounts of local data can markedly improve the transferability, but cross-region data alone contributes little beyond what region-specific training achieves.

We also tested the generalisability of the New Zealand–fine-tuned LLM across different datasets within the same region. Specifically, we used the model originally fine-tuned on the New Zealand Allan Herbarium (CHR) (GBIF.org 2024a) dataset to georeference a randomly selected sample of 5000 records from both the New Zealand Arthropod Collection (NZAC) (GBIF.org 2025a) and the New Zealand National Forestry Herbarium (GBIF.org 2025b). The model achieved mean SAEs of 23.2 km (NZAC) and 43.4 km (Forestry Herbarium), comparable to the performance on the original CHR dataset (41.95 km). Median errors were consistently low across datasets (3.55–7.99 km). These results suggest that the model can generalise reasonably well across independently curated datasets within a region, indicating a degree of transferability beyond the dataset it was originally fine-tuned on.

To further validate this robustness, we conducted a 5-fold cross-validation experiment on the original New Zealand dataset. As shown in Figure 7, the results were highly consistent across folds, with an overall mean SAE of 42.29 km and a median SAE of only 2.95 km. This stability across partitions supports the transfer learning findings, showing that the model's performance is not dependent on a specific training/ test split but generalises reliably to unseen data.

Our results are consistent with recent evaluations of LLM spatial reasoning and spatial cognition. Yang et al. (2025) report that both proprietary and open-source models exhibit limited robustness on spatial tasks that require integrating multiple cues into a globally coherent representation, rather than recalling isolated geographic facts. This aligns with our transfer-learning outcomes: cross-country transfer was minimal without local training data, whereas transfer across independently curated datasets within New Zealand

remained comparatively stable, demonstrating that the learned mapping is strongly conditioned on regional spatial distributions. Importantly, Yang *et al.* (2025) further demonstrates that coupling LLMs with deterministic external tools can substantially enhance large-scale spatial cognition, highlighting a promising direction for improving both accuracy and generalisability beyond what text-only inference can achieve.

4.2.7. Summary of results

Across all experiments, the fine-tuned Mistral 7B model consistently outperformed baselines when locality descriptions carried richer spatial cues. In smaller/denser (with regard to data) settings, it was strongest (e.g. New Zealand 70.43% Acc@10 km/25.36% Acc@1 km; New York 84.89/66.71%; Victoria 81.76/22.79%), and it generalized well to Spanish (Mexico 75.82/51.41%, median SAE 0.89 km). In larger regions such as the USA and Australia, achieving higher performance required a substantial training dataset for broad geographic coverage. Comparisons with prominent GPT models (GPT-5, GPT-4.1, and GPT-4o) further underscored the value of domain-specific fine-tuning, as general-purpose proprietary models failed to reach similar levels of accuracy.

Content analysis shows what the model actually learned: it integrates multiple top-onyms with directional and distance expressions to compose a plausible coordinate, rather than matching a single place name. Longer descriptions that include more place names and spatial indicators were associated with lower errors. Moreover, when explicit distances were removed, the fine-tuned LLM's accuracy declined slightly, suggesting genuine reliance on quantitative cues. Prompt rewordings within the same pattern had only a minor effect, implying that the gains come from learned spatial composition rather than prompt phrasing.

Because performance is closely tied to record density and the geographic extent of the training data, cross-country transfer was limited without local adaptation. Even modest amounts of in-region data yielded measurable improvements, whereas foreign-only training resulted in substantially higher errors. Taken together, these findings show that fine-tuning enables the model to internalize spatial language and compositional cues, achieving strong georeferencing performance across regions and languages when sufficient and representative training data are available.

4.2.8. Limitations and future work

Although our methodology for georeferencing complex locality descriptions using LLMs yielded favourable results compared to the baselines, we identified several limitations. One key consideration is the need for substantial training datasets. To explore the impact of training dataset size on performance, we fine-tuned the model with varying sizes of the New Zealand dataset. As shown in Table 10, there is a clear trend

Table 10. Analysis of the impact of training dataset size on the accuracy of the fine-tuned LLM using the New Zealand dataset.

Training dataset size	Acc@10 km	Acc@1 km	Median SAE	Mean SAE
0	5.23%	0.48%	49.78 km	1120 km
1000	24.42%	2.07%	23.8 km	70.78 km
5000	48.51%	8.16%	10.53 km	56.4 km
10,000	60.76%	15.37%	6.15 km	48.09 km
15,000	67.60%	20.42%	4.31 km	42.43 km

indicating that the model's performance improves as the training dataset size increases. Notably, these experiments highlight that fine-tuning the LLM with just 5000 samples is sufficient to surpass the performance of GEOLocate.

Another limitation of our LLM-based georeferencing approach is its limited transfer learning capability, since accurate georeferencing often requires local geographic knowledge as discussed in Section 4.2.6. In continuing our research on multi-region fine-tuning, we plan to enhance the model's ability to access external resources, such as gazetteers and geographic databases, to supplement its georeferencing capabilities. This could be achieved by integrating retrieval-augmented generation (RAG) (Guu *et al.* 2020) or API-based lookups (Schick *et al.* 2023) that allow the model to reference location data dynamically. By incorporating such external knowledge sources, we aim to improve the model's accuracy in georeferencing data from previously unseen regions and enhance its generalizability across different geographical contexts.

Additionally, our model currently lacks a measure of the associated uncertainty in the results. This measure is crucial for determining the effectiveness and suitability of the model, an aspect extensively explored in previous studies, such as those by Wieczorek *et al.* (2004), Guo *et al.* (2008), and van Erp *et al.* (2015). However, several techniques have been developed for estimating the uncertainty of LLM outputs, such as Negative Log-Likelihood (NLL), consistency-based aggregation techniques, or introducing prompting techniques to verbalize confidence scores in LLM outputs (Lin *et al.* 2022, 2023, Tian *et al.* 2023). Furthermore, Catak and Kuzlu (2024) introduce a novel geometric approach that applies convex hull analysis to the spatial embeddings of LLM outputs, quantifying uncertainty based on the dispersion of generated responses. This method leverages the inherent structure of embedding spaces to infer confidence. The introduction of such spatially-informed Uncertainty Quantification (UQ) techniques for LLMs opens promising directions for our future work.

5. Conclusion

In this paper, we introduced an automatic georeferencing methodology that utilises LLMs. To the best of our knowledge, this study is the first to employ LLMs for georeferencing text data by predicting coordinates that best match the provided locality descriptions. This novel approach utilizes the ability of LLMs to be fine-tuned for specific domains by leveraging their extensive pre-trained knowledge base and adaptability to new datasets. By integrating domain-specific training data, we enable the model to accurately interpret and align textual descriptions with corresponding geographic coordinates. This fine-tuning process enhances the model's precision in biological collection data georeferencing tasks, ensuring that it can effectively handle the nuances and context-specific terminology unique to different geographic datasets.

The results show that our approach outperforms commonly used baselines in almost all cases, demonstrating that our methodology is effective in georeferencing complex locality descriptions of biological collection data across various regions and languages compared to existing methods. Due to limited access to the high-performing GPUs, we experimented only with a 7B parameter LLM variant, which still outperformed existing baselines. We believe that it is likely that the results could be

further improved with larger LLMs. Utilising models with a larger parameter count could potentially enhance the granularity and accuracy of georeferencing by capturing more intricate patterns and relationships within the data.

Although our model outperformed the baselines, for fully automated georeferencing of biological records, a higher level of accuracy would be desirable when limited training data are available. Additionally, the absence of uncertainty measures for the predicted coordinates is another drawback of our methodology. Nevertheless, this study represents an initial effort to implement a more advanced and comprehensive approach to georeferencing complex locality descriptions in biological collection data using LLMs. Future work will focus on refining the LLM's reasoning processes, with input from georeferencing experts. We also plan to incorporate Retrieval-Augmented Generation (RAG) and integrate external data sources, such as gazetteers and web mapping platforms like OpenStreetMap, to enhance the model's accuracy and reliability.

Notes

1. https://www.geo-locate.org.
2. http://www.gbif.org.
3. https://openai.com/.
4. https://chatgpt.com/.
5. https://spacy.io/.
6. https://www.geonames.org/.
7. https://wiki.openstreetmap.org/wiki/Nominatim.
8. https://platform.openai.com/.

Acknowledgements

Disclosure statement

Funding

Notes on contributors

Aneesha Fernando is a PhD student in the School of Computational and Mathematical Sciences at Massey University, New Zealand. Her research interests include Artificial Intelligence, Natural Language Processing, and Machine Learning. Her contributions to this paper include conceptualization, literature review, methodology, software development, data analysis, visualization, and writing, editing, and revising the manuscript.

Surangika Ranathunga is a Senior Lecturer in the School of Computational and Mathematical Sciences at Massey University, New Zealand. Her research interests include Natural Language Processing, Large Language Models, and Machine Learning. Her contributions to this paper include supervision, conceptualization, methodology, writing, editing, and reviewing the manuscript.

Kristin Stock is Director of the Massey Geoinformatics Collaboratory and Associate Professor in the School of Computational and Mathematical Sciences at Massey University, New Zealand. Her research interests include geospatial natural language processing and geospatial semantics. Her contributions to this paper include supervision, conceptualization, methodology, visualization, and writing, editing, and reviewing the manuscript.

Raj Prasanna is an Associate Professor in the Joint Centre for Disaster Research at Massey University, New Zealand. His research interests include situation awareness, internet of things, decentralised networks, and multimodal data fusion. His contributions to this paper include supervision, conceptualization, and review of the manuscript.

Christopher B. Jones is Professor of Geographical Information Systems in the School of Computer Science and Informatics at Cardiff University, UK. His research interests include aspects of geographical information retrieval from text, relating to geoparsing and geocoding, information extraction, and modelling of spatial language. Current applications focus on biological records and ecology, and on emergency response. His contributions to this paper include supervision, conceptualization, literature review, methodology, visualization, and writing, editing, and reviewing the manuscript.

Data and codes availability statement

The data and codes that support the findings of this study are available at the link https://doi.org/10.6084/m9.figshare.27129759.

References

Achiam, J., *et al.*, 2023. Gpt-4 technical report. *arXiv Preprint*, arXiv:2303.08774.

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

Catak, F.O. and Kuzlu, M., 2024. Uncertainty quantification in large language models through convex hull analysis. *Discover Artificial Intelligence*, 4 (1), 1–14.

Chapman, A. and Wieczorek, J., 2020. *Georeferencing best practices*. GBIF Secretariat. Available from: https://docs.gbif.org/georeferencing-best-practices/1.0/en/.

Chen, H., Winter, S., and Vasardani, M., 2018. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, (17), 31–62.

Deng, C., *et al.*, 2024. K2: a foundation language model for geoscience knowledge understanding and utilization. In: *Proceedings of the 17th ACM international conference on web search and data mining*, 161–170.

Desmond, M. and Brachman, M., 2024. Exploring prompt engineering practices in the enterprise. *arXiv Preprint*, arXiv:2403.08950.

Dettmers, T., *et al.*, 2023. Qlora: efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 10088–10115.

Di Rocco, L., *et al.*, 2021. Sherloc: a knowledge-driven algorithm for geolocating microblog messages at sub-city level. *International Journal of Geographical Information Science*, 35 (1), 84–115.

Do, T.H., *et al.*, 2018. Twitter user geolocation using deep multiview learning. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6304–6308.

Doherty, P., *et al.*, 2011. Georeferencing incidents from locality descriptions and its applications: a case study from Yosemite National Park search and rescue. *Transactions in GIS*, 15 (6), 775–793.

Edwards, T., Corcoran, P., and Jones, C.B., 2025. A hybrid approach for geo-referencing tweets: Transformer language model regression and gazetteer disambiguation. *ISPRS International Journal of Geo-Information*, 14 (9), 321.

Ester, M., *et al.*, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 96, 226–231.

Flatow, D., *et al.*, 2015. On the accuracy of hyper-local geotagging of social media content. In: *Proceedings of the eighth ACM international conference on web search and data mining*, 127–136.

Fujiwara, H., Kimura, R., and Nakano, T., 2024. Modify mistral large performance with low-rank adaptation (lora) on the big-bench dataset.

GBIF.org, 2024a. GBIF occurrence download. Available from: [Accessed 19 Mar 2024].

GBIF.org, 2024b. GBIF occurrence download. Available from: [Accessed 10 Jun 2024].

GBIF.org, 2024c. GBIF occurrence download. Available from: [Accessed 22 May 2024].

GBIF.org, 2024d. GBIF occurrence download. Available from: [Accessed 29 Jul 2024].

GBIF.org, 2025a. GBIF occurrence download. Available from: [Accessed 18 Aug 2025].

GBIF.org, 2025b. GBIF occurrence download. Available from: [Accessed 18 Aug 2025].

GemmaTeam, *et al.*, 2024. Gemma: open models based on Gemini research and technology. *arXiv Preprint*, arXiv:2403.08295.

Goodchild, M.F., *et al.*, 2004. Georeferencing in digital libraries. *D-Lib Magazine*, 10 (5), 1082–9873.

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

Guo, Q., Liu, Y., and Wieczorek, J., 2008. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22 (10), 1067–1090.

Guralnick, R.P., *et al.*, 2006. Biogeomancer: automated georeferencing to map the world's biodiversity data. *PLOS Biology*, 4 (11), e381.

Guu, K., *et al.*, 2020. Retrieval augmented language model pre-training. In: *International conference on machine learning*. PMLR, 3929–3938.

Hackeloeer, A., *et al.*, 2014. Georeferencing: a review of methods and applications. *Annals of GIS*, 20 (1), 61–69.

Han, B., Cook, P., and Baldwin, T., 2012. Geolocation prediction in social media data by finding location indicative words. In: *Proceedings of COLING 2012*, 1045–1062.

Han, B., Cook, P., and Baldwin, T., 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49, 451–500.

Hu, E.J., *et al.*, 2022. Lora: low-rank adaptation of large language models. *ICLR*, 1 (2), 3.

Hu, X., *et al.*, 2024. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science*, 1–28.

Hu, Y., *et al.*, 2023. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37 (11), 2289–2318.

Imani, S., Du, L., and Shrivastava, H., 2023. MathPrompter: Mathematical reasoning using large language models. In: *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 5: industry track)*. Association for Computational Linguistics, 37–42. Available from: https://aclanthology.org/2023.acl-industry.4.

Iso, H., Wakamiya, S., and Aramaki, E., 2017. Density estimation for geolocation via convolutional mixture density network. *arXiv Preprint*, arXiv:1705.02750.

Ji, Y., *et al.*, 2025. Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. *International Journal of Geographical Information Science*, 39 (9), 1866–1903.

Jiang, A.Q., *et al.*, 2023. Mistral 7b. *arXiv Preprint*, arXiv:2310.06825.

Karanikolas, N., *et al.*, 2023. Large language models versus natural language understanding and generation. In: *Proceedings of the 27th Pan-Hellenic conference on progress in computing and informatics*, 278–290.

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

Kojima, T., *et al.*, 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.

Laere, O.V., *et al.*, 2014. Georeferencing Wikipedia documents using data from social media sources. *ACM Transactions on Information Systems*, 32 (3), 1–32.

Leidner, J.L., Sinclair, G., and Webber, B., 2003. Grounding spatial named entities for information extraction and question answering. In: *Proceedings of the HLT-NAACL 2003 workshop on analysis of geographic references*, 31–38.

Lester, B., Al-Rfou, R., and Constant, N., 2021. The power of scale for parameter-efficient prompt tuning. *arXiv Preprint*, arXiv:2104.08691.

Li, H., *et al.*, 2003. Infoxtract location normalization: a hybrid approach to geographic references in information extraction. In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, 39–44.

Li, H., *et al.*, 2021. Neural factoid geospatial question answering. *Journal of Spatial Information Science*, (23), 65–90.

Li, M., *et al.*, 2023a. A transformer-based framework for poi-level social post geolocation. In: *European conference on information retrieval*. Springer, 588–604.

Li, Z., *et al.*, 2023b. GeoLM: Empowering language models for geospatially grounded language understanding. In: *Proceedings of the 2023 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 5227–5240. Available from: https://aclanthology.org/2023.emnlp-main.317.

Liao, R., *et al.*, 2022. Predicting distance and direction from text locality descriptions for biological specimen collections. In: *15th International conference on spatial information theory (COSIT 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Lieberman, M.D., Samet, H., and Sankaranarayanan, J., 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In: *2010 IEEE 26th international conference on data engineering (ICDE 2010)*. IEEE, 201–212.

Lin, S., Hilton, J., and Evans, O., 2022. Teaching models to express their uncertainty in words. *arXiv Preprint*, arXiv:2205.14334.

Lin, Z., Trivedi, S., and Sun, J., 2023. Generating with confidence: uncertainty quantification for black-box large language models. *arXiv Preprint*, arXiv:2305.19187.

Liu, J. and Inkpen, D., 2015. Estimating user location in social media with stacked denoising auto-encoders. In: *Proceedings of the 1st workshop on vector space modeling for natural language processing*, 201–210.

Liu, P., *et al.*, 2023. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55 (9), 1–35.

Liu, X., *et al.*, 2021. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv Preprint*, arXiv:2110.07602.

Liu, Z., *et al.*, 2022. Geoparsing: solved or biased? An evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3, 1–13.

Lutsai, K. and Lampert, C.H., 2023. Geolocation predicting of tweets using Bert-based models. *arXiv Preprint*, arXiv:2303.07865.

Marcer, A., *et al.*, 2023. Geopick: georeferencing made easy. *Biodiversity Information Science and Standards*, 7, e111036.

Melo, F. and Martins, B., 2015. Geocoding textual documents through a hierarchy of linear classifiers. In: *Progress in artificial intelligence: 17th Portuguese conference on artificial intelligence, EPIA 2015, Coimbra, Portugal, September 8–11, 2015. Proceedings 17*. Springer, 590–596.

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

Miura, Y., *et al.*, 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In: *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, 1260–1272.

Moncla, L., *et al.*, 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geo-parsed hiking descriptions corpus. In: *Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems*, 183–192.

Mooney, P., *et al.*, 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*, 85–94.

Murphey, P.C., *et al.*, 2004. Georeferencing of museum collections: a review of problems and automated tools, and the methodology developed by the mountain and plains spatiotemporal database-informatics initiative (MAPSTEDI). *Phyloinformatics*, 1 (3), 1–29.

OpenAI, 2024. Gpt-4o: system card. Available from: https://openai.com/index/gpt-4o-system-card/ [Accessed 29 Sep 2025].

OpenAI, 2025a. Gpt-4.1. Available from: https://openai.com/index/gpt-4-1/ [Accessed 29 Sep 2025].

OpenAI, 2025b. Gpt-5. Available from: https://openai.com/index/introducing-gpt-5/ [Accessed 29 Sep 2025].

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

Priedhorsky, R., Culotta, A., and Del Valle, S.Y., 2014. Inferring the origin locations of tweets with quantitative confidence. *In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing*, 1523–1536.

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

Qin, L., *et al.*, 2025. A survey of multilingual large language models. *Patterns*, 6 (1), 101118.

Ranjit, M., *et al.*, 2023. Retrieval augmented chest X-ray report generation using OpenAI GPT models. In: *Machine learning for healthcare conference*. PMLR, 650–666.

Rios, N.E. and Bart, H.L.J., 2014. GEOLocate—software for georeferencing natural history data. Available from: https://www.geo-locate.org/standalone/manual_ver2_0.pdf.

Roberts, J., *et al.*, 2023. GPT4GEO: how a language model sees the world's geography. *arXiv Preprint*, arXiv:2306.00020.

Scherrer, Y. and Ljubešić, N., 2020. HeLju@VarDial 2020: social media variety geolocation with BERT models. In: *Proceedings of the 7th workshop on NLP for similar languages, varieties and dialects*. International Committee on Computational Linguistics (ICCL), 202–211. Available from: https://aclanthology.org/2020.vardial-1.19.

Schick, T., *et al.*, 2023. *Toolformer: language models can teach themselves to use tools*. arXiv Version Number: 1. Available from: https://arxiv.org/abs/2302.04761.

Schulhoff, S., *et al.*, 2024. The prompt report: a systematic survey of prompting techniques. *arXiv Preprint*, arXiv:2406.06608.

Scott, J., *et al.*, 2021. Automated georeferencing of Antarctic species. *In: 11th International conference on geographic information science (GIScience 2021)-part II*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

Sharma, P., *et al.*, 2024. A spatially-aware data-driven approach to automatically geocoding non-gazetteer place names. *ACM Transactions on Spatial Algorithms and Systems*, 10 (1), 1–34.

Simanjuntak, L.F., Mahendra, R., and Yulianti, E., 2022. We know you are living in Bali: location prediction of twitter users using Bert language model. *Big Data and Cognitive Computing*, 6 (3), 77.

Singhal, K., *et al.*, 2023. Large language models encode clinical knowledge. *Nature*, 620 (7972), 172–180.

Stock, K., 2018. Mining location from social media: a systematic review. *Computers, Environment and Urban Systems*, 71, 209–240.

Stock, K., *et al.*, 2023. The biowhere project: unlocking the potential of biological collections data. *GI_Forum*, 1 (1), 3–21.

Thellmann, K., *et al.*, 2024. Towards multilingual LLM evaluation for European languages. *arXiv Preprint*, arXiv:2410.08928.

Tian, K., *et al.*, 2023. Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv Preprint*, arXiv:2305.14975.

Touvron, H., *et al.*, 2023a. Llama: open and efficient foundation language models. *arXiv Preprint*, arXiv:2302.13971.

Touvron, H., *et al.*, 2023b. Llama 2: open foundation and fine-tuned chat models. *arXiv Preprint*, arXiv:2307.09288.

van Erp, M., *et al.*, 2015. Georeferencing animal specimen datasets. *Transactions in GIS*, 19 (4), 563–581.

Vaswani, A., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Wang, J. and Hu, Y., 2019. Are we there yet? evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. *In*: *Proceedings of the 3rd ACM SIGSPATIAL international workshop on geospatial humanities, GeoHumanities '19*. Association for Computing Machinery, 1–6. Available from: https://dl.acm.org/doi/10.1145/3356991.3365470.

Wei, J., *et al.*, 2021. Finetuned language models are zero-shot learners. *arXiv Preprint*, arXiv:2109.01652.

Wei, J., *et al.*, 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

White, J., *et al.*, 2024. ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. In: *Generative AI for effective software development*. Cham, Switzerland: Springer, 71–108.

Wieczorek, J., Guo, Q., and Hijmans, R., 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18 (8), 745–767.

Yang, A., *et al.*, 2025. Evaluating and enhancing spatial cognition abilities of large language models. *International Journal of Geographical Information Science*, 39 (9), 2009–2044.

Yang, Y., *et al.*, 2024. GeoLocator: a location-integrated large multimodal model (LMM) for inferring geo-privacy. *Applied Sciences*, 14 (16), 7091.

Yao, S., *et al.*, 2024. Tree of thoughts: deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 11809–11822.

Yesson, C., *et al.*, 2007. How global is the global biodiversity information facility? *PLOS One*, 2 (11), e1124.

Yu, P., *et al.*, 2023. ALERT: Adapt language models to reasoning tasks. In: *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*. Association for Computational Linguistics, 1055–1081. Available from: https://aclanthology.org/2023.acl-long.60.

Zaken, E.B., Ravfogel, S., and Goldberg, Y., 2021. Bitfit: simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv Preprint*, arXiv:2106.10199.

Zeng, A., *et al.*, 2022. Glm-130b: an open bilingual pre-trained model. *arXiv Preprint*, arXiv:2210.02414.

Zhang, B., *et al.*, 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In: *Proceedings of the fourth ACM international conference on AI in finance*, 349–356.

Zhang, Y., *et al.*, 2023b. GeoGPT: understanding and processing geospatial tasks through an autonomous GPT. *arXiv Preprint*, arXiv:2307.07930.

Ziegler, D.M., *et al.*, 2019. Fine-tuning language models from human preferences. *arXiv Preprint*, arXiv:1909.08593.

Appendices
Appendix A.
Sample responses from LLMs

Figure A1 shows a response provided by ChatGPT-4 when presented with a georeferencing task. In this example, the context manager pattern was utilized. The response from ChatGPT includes the calculation of offsets from a reference location.

Task: Accurately georeference the location provided in the 'Locality Description' below, expressing the coordinates in decimal degrees.
Context: This 'Locality Description' refers to a location in Northland, New Zealand.
Locality Description: 21 km west of Opotiki, south of Wainui Road

To georeference the location accurately, we need to convert the locality description into geographic coordinates. Here's how to approach it:

1. **Identify the starting point:** Opotiki, New Zealand. The coordinates for Opotiki are approximately:
   - Latitude: -38.0046
   - Longitude: 177.2871

2. **Determine the direction and distance:** 21 km west of Opotiki. This means we need to move westward by 21 km.

3. **Adjust the coordinates:**
   - Moving west (longitude decreases as we go westward). To calculate the shift in longitude:
     - The approximate distance per degree of longitude in this region (near -38° latitude) is about 86.4 km (1 degree of longitude ≈ 86.4 km).
   - Distance to move: 21 km west.
   - Change in longitude: $\frac{21 \text{ km}}{86.4 \text{ km/degree}} \approx 0.243$ degrees.
   - Adjusting the original longitude: $177.2871 - 0.243 = 177.0441$.

   So, the approximate coordinates 21 km west of Opotiki are:
   - Latitude: -38.0046
   - Longitude: 177.0441

4. **Refine the locality by considering "south of Wainui Road":**
   - Without an exact position of Wainui Road, but assuming it lies within a reasonable distance south, we should adjust the latitude slightly southwards.

5. **Final approximation:**
   - Move slightly south, adjusting the latitude:
     - Small southward adjustment (about 1-2 km southwards).
   - Latitude adjustment: $\approx -38.0046 - 0.01 = -38.0146$ (approximate small shift southwards).

Therefore, the refined coordinates in decimal degrees for the described locality are approximately:

- Latitude: -38.0146
- Longitude: 177.0441

Figure A1. Response from ChatGPT-4 for a georeferencing problem, generated on June 30, 2024.

Task: Accurately georeference the location provided in the 'Locality Description' below, expressing the coordinates in decimal degrees.

Context: This 'Locality Description' refers to a location in New Zealand.

Locality Description: SOUTH ISLAND: Maitai Valley road, near Golf Course, 2 km SE of Nelson.

Geographic Coordinates: Latitude: -41.2833 Longitude: 173.3167

Figure A2. Response from the fine-tuned mistral model for a georeferencing problem.

Figure A2 shows how our fine-tuned LLM responded to the task of georeferencing locality descriptions. The response was generated in a manner similar to the formatting of the training data.

Appendix B.
Gazetteer-matching algorithm

The following pseudocode summarizes the gazetteer-matching algorithm, which was implemented as one of the baseline methods.

Algorithm 1. Gazetteer-matching algorithm.

1: **Input:** localityDescription, state, country
2: **Output:** latitude, longitude
3: CALL Spacy with localityDescription **returning** locationEntities
4: possibleLocations ← empty list
5: **for** each entity in locationEntities **do**
6: CALL GeoNames with entity, state, country **returning** gazetteerLocations
7: APPEND gazetteerLocations to possibleLocations
8: **end for**
9: CALL DBSCAN with possibleLocations **returning** clusters
10: clusterWithMostPoints ← SELECT cluster with the maximum number of points from clusters
11: latitude ← Average(latitudes of points in clusterWithMostPoints)
12: longitude ← Average(longitudes of points in clusterWithMostPoints)
13: **return** latitude, longitude