Large Multi-Modal Model Cartographic Map Comprehension for Textual Locality Georeferencing

Kalana Wijegunarathna
School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Kristin Stock
School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Christopher B. Jones
School of Computer Science and Informatics, Cardiff University, UK

Abstract

Millions of biological sample records collected in the last few centuries archived in natural history collections are un-georeferenced. Georeferencing complex locality descriptions associated with these collection samples is a highly labour-intensive task collection agencies struggle with. None of the existing automated methods exploit maps that are an essential tool for georeferencing complex relations. We present preliminary experiments and results of a novel method that exploits multi-modal capabilities of recent Large Multi-Modal Models (LMM). This method enables the model to visually contextualize spatial relations it reads in the locality description. We use a grid-based approach to adapt these auto-regressive models for this task in a zero-shot setting. Our experiments conducted on a small manually annotated dataset show impressive results for our approach (∼1 km Average distance error) compared to uni-modal georeferencing with Large Language Models and existing georeferencing tools. The paper also discusses the findings of the experiments in light of an LMM's ability to comprehend fine-grained maps. Motivated by these results, a practical framework is proposed to integrate this method into a georeferencing workflow.

2012 ACM Subject Classification Computing methodologies → Visual inspection

Keywords and phrases Large Multi-Modal Models, Large Language Models, LLM, Georeferencing, Natural History collections

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.12

Supplementary Material *Dataset*: `https://doi.org/10.6084/m9.figshare.29093882.v1`

Funding This research was partly funded by the Ministry of Business Innovation and Employment Smart Ideas Fund under the BioWhere Project (grant number MAUX2104).

1 Introduction

Georeferencing is the process of relating or interpreting information to a geographic location [20, 7, 19]. *Informal* georeferencing is the association of information with a location using place names (also called toponyms) or location descriptions from ordinary human discourse. On the other hand, *formal* georeferencing refers to exact locations using formal quantitative representations such as latitude and longitude coordinates or other spatial referencing systems [20]. The task of converting an *informal* georeference to a *formal* georeference can be challenging due to reasons such as colloquial place names, outdated names, historical places, the use of vague relative spatial relations, and differences in place representations in different gazetteers (geospatial databases).

A vast amount of information is locked up in extensive collections of unstructured textual data that is yet to be systematically georeferenced. These collections include but are not limited to web pages, social media articles, academic research articles, biological collection specimen records, and memoirs. The ubiquity of georeferencing has led to numerous

georeferencing techniques adopted in various application domains. For example, attempts have been made to georeference social media posts, social media images, satellite and aerial images, web documents, and collection records from natural history archives [61, 18, 44, 19, 37]. In this study, we focus on georeferencing textual locality descriptions in records of natural history specimens found in museum and herbarium archives, where it is estimated that of the order of 3 billion records are preserved [4]. It is also estimated that manual georeferencing of digital records without coordinates held globally could take over 5000 person-years [49].

A locality description is a textual description of the location at which a biological or other sample was collected. These descriptions are part of the information recorded about the specimen or sample by the collector and, for millions of pre-GPS collection records, they can be the only detailed information about the collection location. Georeferencing such locality descriptions for purposes of biodiversity studies is a considerable challenge, especially due to their sheer volume and the descriptions themselves often employing quite complex language with one or more relative spatial relations [36]. Much of the published literature on georeferencing entire sentences has focused on social media posts, with the more advanced methods using various forms of language models including transformer models [56]. Methods developed for georeferencing social media posts can also rely heavily on metadata, such as the user network. The locality descriptions with which we are concerned differ significantly from the text of social media postings in their frequent use of relative descriptions often with multiple reference named places, and where the described location is separate (offset) from that of the finer-grained place names. Several studies have focused on the development of methods to georeference such relative locality descriptions in natural history records but little progress has been made to date on the application of current deep learning methods.

Figure 1 provides an example of a locality description. Given this quite specific locality description, a human georeferencer can locate this collection location to a high degree of certainty. Manual georeferencing uses the place names in a locality description to focus on a map that covers the local area to which the description applies. Visualization of the spatial configuration of the named places is vital to a human georeferencer in identifying a point or region of space that appears to correspond to the described location [36]. However, none of the existing automated textual georeferencing methods exploit maps directly. Gazetteer lookup methods only rely primarily on locations of place names, though they can be combined with methods that compute spatial relations [18, 8]. Current deep learning based methods for georeferencing can use pre-trained language models like BERT [11] that have been pre-trained on masked language modeling and next sentence prediction. They rely on fine-tuning these pre-trained models exposing them to large numbers of example texts with their associated locations [44, 30]. Although language models can be adept at learning textual relations, being trained only on language tasks, they do not intrinsically grasp spatial dimensions. The models also do not comprehend spatial extents of the features they are working with. Furthermore, a georeferencing language model trained on one region or country can not be used to infer localities from a different region, requiring more fine-tuning and large volumes of verified data from each region. Additionally, no research appears to have been published to date on using the latest Large Language Models (LLM) for this task.

Here we present initial investigations of the potential of Large Multi-Modal models (LMM), that can support tasks combining language and vision, to assist in the georeferencing process for complex locality descriptions. With an LMM's multi-sensory skills, we experiment with a prompting approach that emulates the way that a human might geofererence such

[1] `https://www.landcareresearch.co.nz/tools-and-resources/collections/allan-herbarium/`

J.K. Donald Wildlife Reserve, NE shore of L. Wairarapa - about 400m from lake

Figure 1 A well defined example locality description from a collection held by the Allen Hebarium[1]. Green and purple indicate place names and relative spatial indicators respectively. Here, "lake" is a coreference to Lake Wairarapa.

descriptions. The intuition in this study is to combine conventional text-based prompting with a map excerpt corresponding to the described location. This exploits the LMM's superior language capabilities while testing its vision encoder for its map reading ability. As current state-of-the-art LMMs excel in language generation and do not perform image segmentation, we superimpose on the map a grid with labelled cells and prompt the LMM to identify the grid cell of the target location. The LMM is given the locality description, the map and the size of the grid cells. We present the results of this study comparing to an existing method, designed for interpreting locality descriptions, and other approaches to using LLMs. Motivated by these results, we design and describe a workflow that can be used to practically automate georeferencing. While the complete workflow is work-in-progress, the core georeferencing module and other elements are already in use for experiments.

Section 2 of the paper will present the related work, after which we will discuss the framework developed to use LMMs in georeferencing in Section 3. Section 4 presents the experiments, results and discussion followed by the conclusion in Section 5.

2 Related work

2.1 Georeferencing

The earliest methods for georeferencing text were based on detecting and geocoding place names in the text, which could then be used to assign one or more spatial footprints. Numerous methods for this detection and geocoding process (sometimes referred to jointly as geoparsing) have been developed [16, 58], and some of these have used deep learning approaches. In the case of [15], input to a convolutional neural network included the place names, context words and target name, and a vector representation of a pixel map of place name instances, that assisted the disambiguation process. Document georeferencing methods are currently dominated by language modelling approaches that treat all terms in a text document as evidence for its location [37]. The initial language models used Bayesian modelling to associate words with locations, where the locations could be grid cells [46, 62], or clusters [55], where the latter included snapping the location to the most similar already georeferenced existing document (in their case a social media posting). More recently, transformer language models have been adopted either to infer coordinates with a regression approach [44] or to classify a location as a geographic region [47], or a point of interest [30].

None of the methods above were specifically intended to deal with relative location descriptions such as commonly occur in archived natural history records. Several studies have presented rule-based approaches to georeferencing natural history specimen locality descriptions that use relative spatial relations to specify an offset relative to a reference place name. Different sorts of offset include simply distance from a named object, distance in a specified cardinal direction, and distance along a path. Typically these methods include some or all of detecting place names and spatial relational phrases, disambiguating and hence geocoding the place name, applying the offset distance, and computing some measure of uncertainty. The point radius method [61] was developed to achieve this, in which offsets

were calculated relative to a representative point of a feature while also taking account of its extent. The uncertainty of an inferred point-based georeference was expressed as a radial distance that is a function of the six factors of extent of the locality, distance precision, direction precision, unknown datum, coordinate measurement precision and map scale.

The point radius approach was refined in [18, 34], by defining several types of density based uncertainty fields, that take into account the shape of the reference object and which can be combined for complex descriptions. [53] computed distance and direction offsets, accompanied by the spatial minimality toponym disambiguation method [27], and applying a confidence measure based on matching the target record to already geofererenced records of the same survey expedition, and to the nearest location of other archived records that have the same species.

Georeferencing of descriptions of locations that use spatial relations and which were generated in a human subject experiment was described in [8]. This is one of the few examples of developing and experimenting with geospatial models of spatial relations in natural language expressions outside of the natural history domain. The approach combined models of the applicability of different sorts of relative spatial relations and required the prior existence of a place graph of the spatial relationships between places mentioned in the texts.

2.2 LMMs and Geospatial Use Cases

With the recent rapid development of LLMs such as GPT4 [1], Llama [52], PaLM [10], Flamingo [2], and DeepSeek's V3 [32] and R1 models [17], adding other modalities, including vision, was seen by many as the next improvement. This led to the development of LMMs such as GPT-4Vision [40], Qwen-vl [5], PALM-E [12], Gemini-Pro Vision[2], Sphinx and Janus Pro [9]. However, there exist Vision-Language models that predate these LMMs such as CLIP [41], LLaVa [33] and BLIP [29] that combine the two modalities. These models have set benchmarks in various Vision-language tasks such as Visual Question Answering (VQA) [3, 24], image captioning [45, 39], visual language navigation [48] and visual reasoning [65].

LMMs have been applied in several geospatial applications. Vision capable models like GPT-4 Vision, Gemini Pro Vision, and Sphinx have been tested for tasks like map element recognition, where GPT4Vision has proved superior [63]. This study also tests GPT4Vision's comprehension of thematic maps, point pattern, and time series analyses. GPT4Vision has also been tested in its ability to understand weather charts and make forecasts [26]. Although not using vision capabilities, LLM's abilities to carry out spatial tasks like mapping using code and external tools like MapBox[3], spatial reasoning, and describing interior locations have been tested [21, 31]. Perhaps the study closest to ours in use case is [71], although they do not use Language-Vision models. This study focuses on geolocating images. They consider maps and image embeddings as two modalities in their multi-modal fusion approach, where they use maps to build a point-cloud representation that can be fused with embeddings from images to exploit heights of buildings to better geolocate images. To the best of our knowledge no method attempts to goereference textual locality descriptions or any form of text documents with LMMs using maps as inputs. We were also unable to find any literature attempting to georeference textual documents using LLMs.

[2] https://aistudio.google.com/
[3] https://www.mapbox.com/

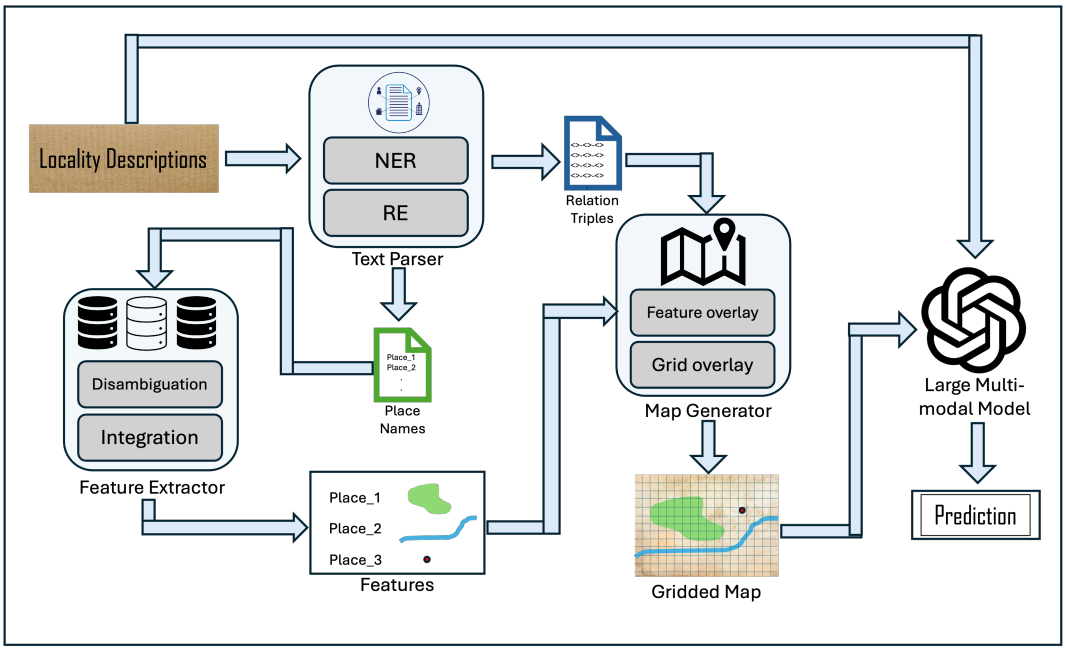


Figure 2 Workflow of the complete automated georeferencing process.

3 Methodology

Figure 2 presents the overall workflow of our proposed framework to utilize large multi-modal models to accurately georeference locality descriptions using gridded maps. We present a detailed description of the proposed method and the individual modules in this section.

3.1 Textual Information Parsing

As illustrated in Figure 2, the first step of the process is to extract the names of the places. Grounding named places is the most effective and simplest form of georeferencing and this is vital to our workflow. Named Entity Recognition (NER) [38] is an extensively researched problem in Natural Language Processing (NLP). Place names or locations are one of the classical semantic types that NER uses to assign labels to tokens or words [28], making most NER solutions accessible for this step of our framework. Off the shelf NER tools such as spaCy[4], StanfordNER [14], NLTK [6], and attention [56] based pre-trained transformer models [70, 50] or modern LLM based approaches [22, 13, 66] can be leveraged for the recognition of place names. Coreference resolution [51] can be beneficial when parsing relations as illustrated in Figure 1. The extracted entities are used for Relation Extraction (RE) and finally passed to the Feature Extractor module.

The subsequent step is the extraction of spatial relations between entities. As illustrated in Figure 1, a single locality description may contain multiple relation clauses in the form of $<$ *locatum*, *spatial indicator*, *relatum* $>$ triples that relate a location or located object (the *locatum*) to a reference object or location (the *relatum*) with a phrase or clause denoting the spatial relationship (*spatial indicator*). It is also common in locality descriptions to have

[4] `https://spacy.io/`

degenerate spatial relations where the locatum is not explicitly mentioned in text but is often the final location being described [23]. RE is also a thoroughly studied area. In addition to generic RE methods [70, 57, 67] used in information extraction and NLP, more geospatial relation oriented RE methods have also been developed [25, 35]. Relation triples extracted using these methods will then be passed to the Map Generator module (see Figure 2).

3.2 Geospatial Feature Extractor

Gazetteers and geospatial databases serve as fundamental resources for the grounding of place names, providing structured and authoritative spatial references. This module will be responsible for extracting relevant features from these knowledge bases, disambiguating them, and selecting the preferred representation of the place instance. While individual states often maintain authoritative gazetteers, several prominent sources provide global coverage. These include, but are not limited to OpenStreetMap[5] (OSM), GeoNames[6] and, Google Places API [7]. These sources can vary in their coverage of different place categories (e.g., natural features vs. artificial structures) and in the type of geometric representations they offer, ranging from point-based locations to more complex polygonal and linear footprints. The reliability and completeness of these sources can also vary as some of them are authoritative while others are community-based volunteered information. As the collection country and region are usually included in the records held by collection agencies, we are also able to exploit country-specific gazetteers, allowing us to draw from more authoritative and accurate sources. Conflating these sources provides the most comprehensive set of features for place names mentioned in a locality description.

First, we query the spatial databases with the place names returned by the previous module. The country name and region of collection can be used for disambiguation. If multiple candidates from the same region from the same source remain, a spatial clustering disambiguation is carried out (*cf* [27]). This clusters all place names mentioned and selects the candidates that form the strongest cluster, filtering out outliers. Subsequently, we are left with a single feature from each source per place name. In our conflation of sources, we prioritize features with complex geometries as this preserves information like extent and boundaries required for visual georeferencing. Preference is also given to authoritative sources. Finally, the selected features are passed on to the Map Generator module.

3.3 Map generation

For the effective application of LMMs in georeferencing, the creation of a map excerpt that is likely to contain the ground truth sample collection location is essential. As the first step of the map generation process, our map server will overlay the features returned from the Feature Extractor on a suitable basemap. Also vital to accurate georeferencing using a vision-based approach is the scale of the map. The map excerpt should be created with all essential landmarks and features necessary for an accurate georeferencing. It should also not be too coarse-grained, to avoid very large grid cells and high uncertainty. We propose the following steps to create the map excerpts:

1. In a location description with two or more named places, x, y where location x is completely contained in y, the full extent of y need not be included in the map extract. Take for example, the following locality description: ***North Island, Bay of Islands County.***

[5] `https://www.openstreetmap.org/`
[6] `https://www.geonames.org/`
[7] `https://developers.google.com/maps/documentation/places/web-service`

Ca 2km north of Puketi. In this example, North Island contains Bay of Islands County and the county contains Puketi, a small locality. We avoid creating a much coarser grained map by not including the whole extent of the North Island or the Bay of Islands region and focusing on the most fine grained location (Puketi). However, the parent entity is used for disambiguation purposes when retrieving the child entity.

2. If there are two or more independent locations at the same level, the map extract must include the full extent of all such features. e.g.: ***Fiordland, Mount George, south shore of lake at head of Elizabeth Burn, 2km north of peak.*** In this example, both Elizabeth Burn and Mount George are included in the map excerpt. The full extent of Fiordland does not need to be included as per 1. above.
3. If the description includes an absolute distance based spatial relation, we ensure the map excerpt includes a buffered spatial extent of the relatum.
4. We ensure features are clearly visible in contrast to the base map. i.e. distinct boundaries for polygon features, clearly highlighted linear and point features.
5. We ensure legible labels for all identified and retrieved places.

Subsequently, we superimpose a labeled square grid on the map excerpt. We also record the size of the map grid cells as this is used during inference to calculate relative distances.

3.4 Multi-modal Georeferencer

The Georeferencer, essentially a Large Multi-modal Model pre-trained on both language and vision tasks, is the core of the proposed framework. This module takes as input the original locality description that is to be georeferenced along with the gridded map excerpt created by the Map Generator and attempts to predict a grid cell that is most likely to contain the location described in the locality description. Similar to LLMs, LMMs can be sensitive to the prompts used.

3.4.1 Prompt Design

We experimented with several prompts to choose the most effective prompt for this multi-modal georeferencing task.

1. Simple Zero-Shot Prompting [42]:

> What grid cell/cells represent the following location description?
> Location Description:

2. Automatic Chain-of-thought [68, 59]:

> Based on the gridded map given, what grid cell/cells represent the following location description? Think step by step.
> Location Description:

3. Logical Chain-of-Thought Prompting [69]:

> Based on the gridded map given, what grid cell/cells represent the following location description?
> Think step by step. Identify the locations mentioned and use the relative spatial relations mentioned in the description.
> Location Description:

4. Logical Chain-of-Thought Prompting with grid size:

> Based on the gridded map given, what grid cell/cells represent the following location description?
> Each grid cell is <grid size> × <grid size>.
> Think step by step. Identify the locations mentioned. If a distance is mentioned in the description, use the grid sizes to calculate the relative distances.
> Location Description:

5. Persona [60] with Logical Chain-of-Thought Prompting with grid size:

> You are a language and cartography expert. Based on the gridded map given, what grid cell/cells represent the following location description?
> Each grid cell is <grid size> × <grid size>.
> Think step by step. Identify the locations mentioned. If a distance is mentioned in the description, use the grid sizes to calculate the relative distances.
> Location Description:

Our preliminary analysis of these prompting patterns indicated that the Logical Chain-of-thought prompt enhanced with the grid size produced the best results. We will carry out the rest of the experiments with this prompt.

The whole framework proposed in this section is highly reliant on the capability of an LMM to effectively and accurately georeference locations with the aid of a visual map. We present the experiments we conducted to gauge the potential of a multi-modal approach and the merits of diverging from traditional uni-modal text based approaches in the next section.

4 Experiments

4.1 Data

For our preliminary experiment, collection records were obtained from Global Biodiversity Information Facility[8] (GBIF). GBIF collection records report accurate coordinates for 83% of the georeferenced records held in it [64]. Short location descriptions are more likely to contain only a single place name or a sequence of place names and no explicit spatial relations (though a comma separated sequence could represent a containment hierarchy). In the absence of descriptive spatial relations, any georeferencing method can, in the best case, only provide the coordinates of the place name mentioned (similar to a gazetteer lookup method). Therefore, the data were first filtered to collect location descriptions that were 60 characters or longer in length, allowing us to gauge the methods' performances on descriptive spatial relations. Given the vast number of collection records and collection types in GBIF, we limited the data to floral specimen collection records from New Zealand provided to GBIF by the Allen Herbarium. The place names and relations were manually annotated as the Text Parser was not implemented at the time of experiment.

For the purposes of this preliminary study, we randomly sampled 25 records to create cartographic map snippets. For this manually curated dataset, we only used OSM to identify named places that are overlaid on the standard OSM base map. For this experiment, we

[8] http://www.gbif.org

manually checked the excerpts to ensure that the ground truth location was contained within the map excerpt. In our dataset of 25 examples, we observed that the ground truth location was consistently included within the map excerpt generated using the aforementioned steps, without needing any further manual intervention. However, it was observed that in examples with linear features extending over large geographic extents such as highways and rivers, the map excerpt was too coarse grained. In these cases, we manually zoomed in on the non-linear features in the description, making sure to preserve some sections of the linear feature. We will analyse the affects of this manual manipulation in Section 4.5.2.

Finally, each data item, e_i, in our dataset can be characterised as follows:

$$e_i = \{text_i, country_i, region_i, map_i, location_i, label_i, scale_i\}, \quad (1)$$

where *text* is the locality description, *country* and *region* are fields acquired from GBIF, *map* is the grid-labeled map, *location* is the ground truth point location of collection as recorded in GBIF (latitude and longitude coordinate pair), *label* is the label of the grid cell that contains the *location* and *scale* is the size of the grid cell in the map. We manually annotated *label* for each of these examples after the grid is superimposed. To the best of our knowledge, this is the first publicly available dataset[9] for fine-grained cartographic map comprehension for LMMs.

4.2 Baselines

GeoImp [54] is perhaps the most recent georeferencing tool for biological specimen georeferencing but unfortunately it is no longer available online. The most effective methods developed for social media post georeferencing (such as Tweets) rely on the metadata and social network information and are therefore unsuitable for our task. GEOLocate [43] is an easy-to-use georeferencing system designed specifically for georeferencing natural history collection data, accessible both as a standalone software and an online service. We use this as one of our baselines. GEOLocate enables multiple predictions for each location description, but we only use its best prediction for this study. Additionally, as we are testing the performance of LMMs, we implement our own LLM baselines. All baselines compared against our LMM based generative approach are listed here:

1. **GEOLocate$_{text}$**: We use GEOLocate's batch processing function over their online service. We only provide the textual description, $text_i$, to the service.
2. **GEOLocate$_{text+region}$**: With this baseline, in addition to the text to georeference, we provide GEOLocate the $country_i$ and $region_i$ from our dataset.
3. **ChatGPT$_{text}$**: Zero-shot georeferencing with OpenAI's ChatGPT[10]. We use their flagship model, GPT-4o. We manually prompt it adapting a persona prompting pattern [60]:

> You are a language and geography expert.
> Georeference the following location description. Answer with coordinates in decimal degrees.
> Location Description: $\{text_i\}$

[9] https://doi.org/10.6084/m9.figshare.29093882.v1
[10] https://chatgpt.com/

4. **ChatGPT$_{text+region}$**: This method takes a similar approach to **ChatGPT**$_{text}$ but enriches the prompt with more context by explicitly providing it with the country and region of collection.

> You are a language and geography expert.
> Georeference the following location description. Answer with coordinates in decimal degrees. The country and the district of the location are provided.
> This location is in $\{region_i\}$, $\{country_i\}$.
> Location Description: $\{text_i\}$

5. **GPT-4o$_{text}$**: We use the same prompt as the ChatGPT$_{text}$ and the same underlying model (GPT-4o) but instead of using the web browser, we use the OpenAI's API. The distinction between the two methods is that ChatGPT$_{text}$ has the capability to search the web and retrieve the coordinates of the place names and related information, whereas GPT-4o$_{text}$, accessed via the API, lacks this functionality.
6. **GPT-4o$_{text+region}$**: Prompts the GPT-4o model through OpenAI's API using the region and country enhanced prompt as seen in ChatGPT$_{text+region}$.

4.3 Evaluation Metrics

While distance to ground truth location from the prediction is a straight-forward measure of error for methods that predict coordinates, the measurement of error is slightly more complicated for comparing grid cells with coordinates. We implement three Euclidean distance metrics to calculate the distance error given the correct grid cell label, $label_i$, a predicted grid cell label, $pred_i$, and $scale_i$:

$$centroid - distance = \sqrt{|x_2 - x_1|^2 + |y_2 - y_1|^2} \times scale_i, \tag{2}$$

$$max - distance = \sqrt{(|x_2 - x_1| + 1)^2 + (|y_2 - y_1| + 1)^2} \times scale_i, \tag{3}$$

$$min{-}distance = \sqrt{min(||x_2 - x_1| - 1|, |x_2 - x_1|)^2 + min(||y_2 - y_1| - 1|, |y_2 - y_1|)^2} \times scale_i, \tag{4}$$

where (x_1, y_1) and (x_2, y_2) are two dimensional indices of the grid cells of $label_i$ and $pred_i$, respectively. Each grid cell is a unit square such that $(x_1 y_1), (x_2, y_2) \in \mathbb{N}^+ \times \mathbb{N}^+$. The $centroid - distance$ calculates the Euclidean distance between the two grid cell centroids, where one centroid is considered the ground truth point of collection and the other is the predicted point. The $max - distance$ indicates the upper bound of error, while the $min - distance$ gives the error in the best case scenario. $max - distance$ records an error of $\sqrt{2 \times scale_i^2}$ even if both ground truth cell and predicted cell are the same and calculates the distance between the two furthest corners of the given cells. Conversely, $min - distance$ gives an error of zero if the predicted cell and the ground truth cell are the same or are adjacent to each other, calculating the minimum distance between the two cells. For GEOLocate and the generative LLMs, we use the mean Simple Accuracy Error (SAE) between coordinate pairs. We also compare the methods on the percentage of predictions that lie within a 1km, 3km, 10km and $scale_i$ radius of the actual location.

Table 1 Average distance errors and percentage of predictions within range of ground truth across the dataset.

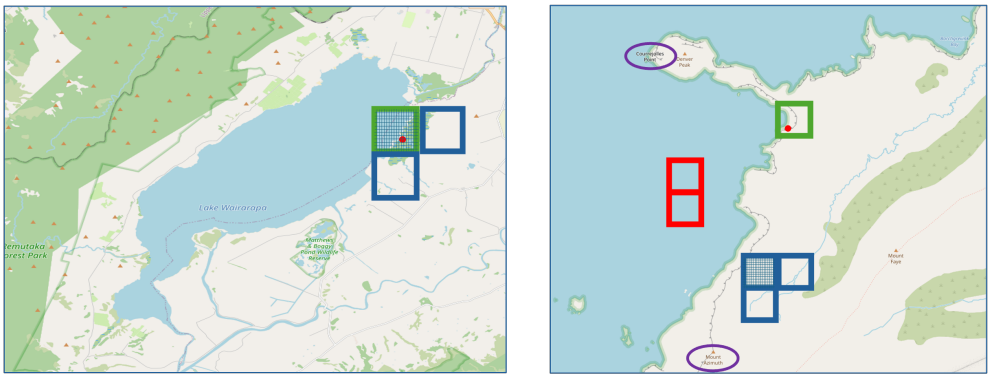
Method		Average distance (km)	% acc@ 1km	% acc@ 3km	% acc@ 10km	% acc@ $scale_i$
GEOLocate$_{text}$		107.23	16.0	28.0	52.0	8.0
GEOLocate$_{text+region}$		107.23	16.0	28.0	52.0	8.0
ChatGPT$_{text}$		10.91	8.0	16.0	64.0	4.0
ChatGPT$_{text+region}$		10.12	8.0	16.0	68.0	
GPT-4o$_{text}$		155.82	4.0	16.0	40.0	8.0
GPT-4o$_{text+region}$		39.98	0	12.0	56.0	0
Our method	min	0.42	84.0	96.0	100	88.0
	max	2.16	24.0	80	100	0
	centroid	1.03	60.0	96.0	100	32.0

4.4 Results

Table 1 reports the performance of all methods tested. Both methods utilizing **GEOLocate** produced identical results, signaling that the region and country attributes do not contribute meaningfully to the georeferencing process. This may vary in other regions, such as the United States, where the state-based administrative system is more relevant as indicated in the documentation of GEOLocate. Out of the baselines, **ChatGPT$_{text+region}$** shows the best results with an average error of 10.12km. **ChatGPT$_{text}$** follows closely behind with no significant reduction in average distance. This indicates the LLM's ability to disambiguate places to a high degree of accuracy even without the region or country information. **GPT-4o$_{text}$** produces the highest distance error. However, enhancing the prompt with the region, as in **GPT-4o$_{text+region}$**, significantly improves results. This suggests the LLM's use of region for disambiguation. The stark difference in performance between the browser versions (ChatGPT$_{text+region}$, ChatGPT$_{text}$) and the same model accessed via the API (GPT-4o$_{text}$, GPT-4o$_{text+region}$) raise an important issue: the inability to browse the web in the API versions significantly hinders the quality of georeferencing. This is also observed in some of the reasoning provided by the model when producing the results. Versions with internet access are able to produce accurate coordinates for named places in the locality descriptions. This also leaves room for further improvement of the LLM based approaches. Providing precise and accurate locations for the named places may result in better quality. However, these improvements are not within scope of this paper.

Another interesting observation is the change of % acc at various distances. Although ChatGPT$_{text+region}$ and ChatGPT$_{text}$ produced lower errors (out of the baselines), the % acc@1km, and % acc@3km are worse than those of **GEOLocate** methods. Although able to correctly disambiguate the places and predict within the vicinity, all the LLM based approaches struggle to make a fine-grained prediction. This is to be expected as these methods only predict using point coordinates. Especially for large features such as rivers, mountains, and reserves, a point alone is an inadequate representation for an accurate georeferencing. Furthermore, these results indicate the LLM's inability to take adequate consideration of the rich spatial relations commonly found in these locality descriptions.

The LMM we used for this experiment to test our approach is the OpenAI gpt-4o-2024-08-06 model accessed through their API. As previously discussed in Section 3.4.1, our prompt for the LMM does not limit the prediction of multiple grid cells. In our experiments, when the model predicts multiple cells, we only consider the first cell predicted. Our proposed



(a) Locality of Lake Wairarapa

(b) Locality of Mount Azimuth & Courrejolles point

Figure 3 Map excerpts, their labels and their predictions for two locality descriptions: (a) J.K. Donald Wildlife Reserve, NE shore of L. Wairarapa – about 400m from lake & (b) Mount Azimuth, cliffs between Azimuth and Courrejolles Point near low point in ridge. The grid sizes for (a) and (b) are 1.88km and 0.7 km respectively. The red point indicates the exact point of collection. The green cell indicates the grid cell containing this point. The blue meshed cell indicates the first and primary cell predicted by the model and the other blue cells indicate the secondary predictions. The two place names mentioned in (b) are highlighted for visual clarity and the red cells indicate some of the cells considered during the reasoning of the model.

approach significantly outperforms the baselines. The centroid-distance of the LMM is an order of magnitude more accurate than the best-performing baseline. Max-distance, which is the upper bound for error given two grid cells, is also markedly lower than all baselines. This indicates our method's ability to consider intricate spatial relations when producing georeferences. When considering a centroid-centroid distance, 60% of the predictions lie within 1km range of the actual location of collection. This level of accuracy is crucial when manually retrieving biological specimens. 32% of the predictions made by our multi-modal approach fall exactly in the correct grid cell as the original location. These results clearly demonstrate the significantly superior performance and usefulness of our grid-based multi-modal approach.

4.5 Discussion

4.5.1 Spatial extent and terrain understanding

A unique advantage of a multi-modal approach to georeferencing is its potential to understand spatial extents without being limited to simple coordinates. We analyzed the results to identify if the model is indeed capable of understanding extents of features. Figure 3(a) demonstrates an example where the model accurately identified the correct grid cell containing the point of collection. This is the map excerpt and prediction for the locality description shown in Figure 1. OSM did not find a match for J.K. Donald Wildlife Reserve and the model was restricted to only looking at the lake and its locality. Despite this, the model's ability to correctly predict the grid cell demonstrates the model's ability not only to identify the boundaries of the lake but also the distance from the border where the collection may have taken place (i.e. the "shore" in the locality description). Also of interest is the reasoning it produced for the prediction. The LLM response stated that it considers the green area that looks like a "vegetation patch" to be the J.K. Donald wildlife reserve. This shows the model's ability to identify and reason with topographic features on the base map. Although

the LLM's mentioned feature identity is questionable (as OSM's name for that patch is Wairarapa Moana Wetland), this highlights a capability that could be highly beneficial for map-based spatial reasoning.

Figure 3(b) provides another similar example. In this case, the prediction is far from the actual collection location. However this is understandable when we analyse the locality description: ***Mount Azimuth, cliffs between Azimuth and Courrejolles Point near low point in ridge***. Without contour lines or other altitude information, the phrase "low point in ridge" is indiscernible. What is of interest is the calculation the model made for "between". The initial reasoning calculations made by the model predicted the cells marked in red as the cells that represent "between Azimuth and Courrejolles Point". However, it later disregarded these cells in favour of the grids marked in blue. Although not explicitly stated, it seems to have avoided predicting a place in the ocean. This may also have been helped by the mention of an unnamed cliff. This ability of understanding terrain as shown in both examples opens the door to incorporating species-related habitat information into our approach. This could include characteristics such as whether a species inhabits land or water and even probabilistic heat maps on a species' preferential ecosystem.

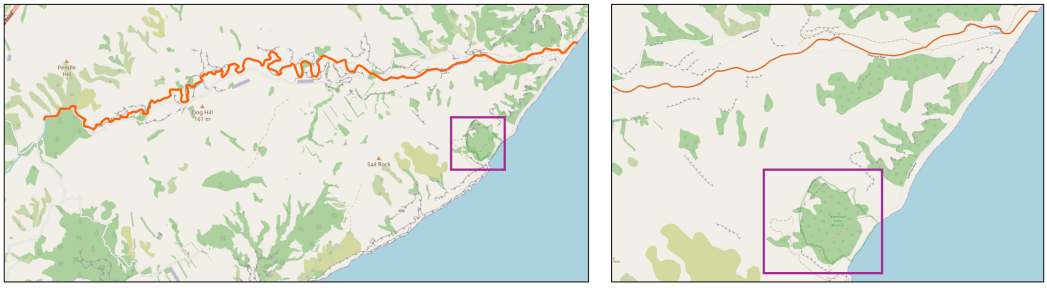
4.5.2 Linear Features

As mentioned earlier during the creation of the gridded map dataset, manual intervention was needed in the case of linear features. 9 out of the 25 samples contained linear features. Figure 4 demonstrates this issue, presenting two map excerpts for the following locality description: "***North Canterbury, Napenape Scenic Reserve, 3km south of mouth of Blythe River on coast.***". Including the complete linear feature resulted in a vastly coarser grained map where the subsequently applied grid cells were 1.25km in scale. The map excerpt relevant for the accurate georeferencing would produce much finer grained cells of size 450m, allowing the model to not only pay attention to the river and the reserve but also differentiate grid cells based on whether they lie close to the coast or not. The proposed framework will benefit from further experiments on limiting the extent of the map especially with regard to linear features. A potential avenue is the exploration of distances to the other mentioned features and using these relations to limit the scope of the map.

Another observation on linear features was the vision encoder's difficulty in comprehending the continuity of the linear features. Some confusion was observed when one road meets another at a junction but continues to be the same road after it. However, this can be remedied by custom labels placed at regular intervals of the linear feature.

4.5.3 Enhancing vision models' map comprehension

Along with the confusion with linear features, we also noticed a tendency of the model to misrecognize the location of a feature using the label on the map instead of the icon or marker. This is contrary to findings in coarser grained maps [63]. These issues persist due to models like GPT-4o(Vision) not being specifically trained for map comprehension. Despite these inaccuracies, the performance of this zero-shot multi-modal approach is vastly superior to text only approaches. However, there is still space for improvement through fine-tuning, which would the take into account the considerable variation in the forms of locality descriptions. The large numbers of natural history records collected from many different countries around the globe with detailed locality descriptions present an invaluable source of information to fine-tune (or perhaps even use during pre-training) vision models on map comprehension. Maps created using our framework can easily be annotated using



(a) Full extent of linear feature (b) Excerpt relevant for georeferencing

Figure 4 Two map excerpts for the same locality description. The inclusion of the full extent of the river (highlighted in red), as shown in (a) produces a much coarser map compared to (b). The Napenape Scenic Reserve is segmented in purple for visual clarity.

existing vision models: thus the framework could be used to create a version of the map with the point of collection prominently marked. Existing multi-modal models can then be used for the labelling ("Which grid cell contains the <Red Marker>?") of these maps. These labels can subsequently be used for fine-tuning vision capabilities of other LMMs using the version of the map where the point of collection is removed. Alternatively, this can be used to pre-train open source vision encoders jointly with smaller open weight LLMs[11] to build LMMs specialized in map reading. This framework, of distantly supervised learning with cheap machine annotated data, can be regarded as analogous to masked language modeling or next sequence prediction for uni-modal language models.

5 Conclusion

This paper presents a novel method for georeferencing textual locality descriptions using LMMs to combine text understanding with map reading. The accuracy of this method is tested against existing tools and the current state-of-the-art LLMs where our method demonstrates greatly superior results. The distance error improves by an order of magnitude compared to the best baseline. Motivated by these results, a framework and workflow were designed to practically integrate LMMs for the task of georeferencing locality descriptions. Along with the model's unique abilities and current shortcomings, the study also revealed avenues for future research that can be used to build powerful models capable of true map comprehension, taking one more step towards GeoAI.

References

1 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint*, 2023. arXiv:2303.08774.

2 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[11] Where the weights (parameters) of the LLM model are accessible

3 Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. doi:10.1109/ICCV.2015.279.

4 Arturo H Ariño. Approaches to estimating the universe of natural history collections data. *Biodiversity informatics*, 7(2), 2010.

5 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. doi:10.48550/arXiv.2308.12966.

6 Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

7 Arthur D Chapman and John R Wieczorek. Georeferencing best practices. Version 1.0, 2020. doi:10.15468/doc-gg7h-s853.

8 Hao Chen, Stephan Winter, and Maria Vasardani. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, 17:31–62, 2018. doi:10.5311/JOSIS.2018.17.417.

9 Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint*, 2025. doi:10.48550/arXiv.2501.17811.

10 Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL: https://jmlr.org/papers/v24/22-1144.html.

11 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. arXiv:1810.04805.

12 Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint*, 2023. arXiv:2303.03378.

13 Jianzhou Feng, Ganlin Xu, Qin Wang, Yuzhuo Yang, and Lei Huang. Note the hierarchy: Taxonomy-guided prototype for few-shot named entity recognition. *Information Processing & Management*, 61(1):103557, 2024. doi:10.1016/J.IPM.2023.103557.

14 Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 363–370, 2005. doi:10.3115/1219840.1219885.

15 Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. Which Melbourne? augmenting geocoding with maps. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1296, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1119.

16 Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What's missing in geographical parsing? *Language Resources and Evaluation*, 52:603–623, 2018. doi:10.1007/S10579-017-9385-8.

17 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*, 2025. arXiv:2501.12948.

18 Qinghua Guo, Yu Liu, and John Wieczorek. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10):1067–1090, 2008. doi:10.1080/13658810701851420.

19 Andreas Hackeloeer, Klaas Klasing, Jukka M Krisp, and Liqiu Meng. Georeferencing: a review of methods and applications. *Annals of GIS*, 20(1):61–69, 2014. doi:10.1080/19475683.2013.868826.

20 Linda L Hill. *Georeferencing: The geographic associations of information*. Mit Press, 2009.

21 Hartwig H Hochmair, Levente Juhász, and Takoda Kemp. Correctness comparison of chatgpt-4, gemini, claude-3, and copilot for spatial tasks. *Transactions in GIS*, 28(7):2219–2231, 2024. doi:10.1111/TGIS.13233.

22 Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259, 2024.

23 A. Khan, M. Vasardani, and S. Winter. Extracting spatial information from place descriptions. In *COMP '13 ACM SIGSPATIAL International Workshop on Computational Models of Place*, pages 62–69, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2534848.2534857.

24 Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024. doi:10.48550/arXiv.2404.19205.

25 Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):1–36, 2011. doi:10.1145/2050104.2050105.

26 John R Lawson, Joseph E Trujillo-Falcón, David M Schultz, Montgomery L Flora, Kevin H Goebbert, Seth N Lyman, Corey K Potvin, and Adam J Stepanek. Pixels and predictions: Potential of gpt-4v in meteorological imagery analysis and forecast communication. *Artificial Intelligence for the Earth Systems*, 4(1):240029, 2025.

27 Jochen L Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 31–38, 2003.

28 Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70, 2020.

29 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. URL: https://proceedings.mlr.press/v162/li22n.html.

30 Menglin Li, Kwan Hui Lim, Teng Guo, and Junhua Liu. A transformer-based framework for poi-level social post geolocation. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I*, volume 13980 of *Lecture Notes in Computer Science*, pages 588–604. Springer, 2023. doi:10.1007/978-3-031-28244-7_37.

31 Krzysztof Lipka, Dariusz Gotlib, and Kamil Choromański. The use of language models to support the development of cartographic descriptions of a building's interior. *Applied Sciences*, 14(20):9343, 2024.

32 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*, 2024. arXiv:2412.19437.

33 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

34 Y. Liu, Q. H. Guo, J. Wieczorek, and M. F. Goodchild. Positioning localities based on spatial assertions. *International Journal of Geographical Information Science*, 23(11):1471–1501, 2009. doi:10.1080/13658810802247114.

35 Oswaldo Ludwig, Xiao Liu, Parisa Kordjamshidi, and Marie-Francine Moens. Deep embedding for spatial role labeling. *arXiv preprint arXiv:1603.08474*, 2016. arXiv:1603.08474.

36 A Marcer, Quentin Groom, Elspeth Haston, and Francesc Uribe. Natural history collections georeferencing survey report. *Current georeferencing practices across institutions worldwide. Zenodo*, 2021.

37 Fernando Melo and Bruno Martins. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21(1):3–38, 2017. doi:10.1111/TGIS.12212.

38 David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. In *Named Entities: Recognition, classification and use*, pages 3–28. John Benjamins publishing company, 2009.

39 Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36:22047–22069, 2023.

40 GPT OpenAI. 4v (ision) system card. *preprint*, 2023.

41 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

42 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

43 Nelson E Rios and Henry L Bart Jr. Geolocate - software for georeferencing natural history data, Year of Access or Publication. Accessed: 10 Feb. 2025. URL: https://www.geo-locate.org.

44 Yves Scherrer, Nikola Ljubešić, et al. Social media variety geolocation with geobert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 135–140. The Association for Computational Linguistics, 2021.

45 Florian Schneider and Sunayana Sitaram. M5–a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *arXiv preprint arXiv:2407.03791*, 2024. doi:10.48550/arXiv.2407.03791.

46 Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009. doi:10.1145/1571941.1572025.

47 Lihardo Faisal Simanjuntak, Rahmad Mahendra, and Evi Yulianti. We know you are living in bali: Location prediction of twitter users using bert language model. *Big Data and Cognitive Computing*, 6(3):77, 2022. doi:10.3390/BDCC6030077.

48 Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. *arXiv preprint arXiv:2412.09082*, 2024. doi:10.48550/arXiv.2412.09082.

49 Kristin Stock, Kalana Wijegunarathna, Christopher B Jones, Hone Morris, Pragyan Das, David Medyckyj-Scott, and Brandon Whitehead. The biowhere project: unlocking the potential of biological collections data. *GI_Forum*, 11(1):3–21, 2023.

50 Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*, 2022. doi:10.48550/arXiv.2208.03054.

51 Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020. doi:10.1016/J.INFFUS.2020.01.010.

52 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, 2023. arXiv:2307.09288.

53 Marieke van Erp, Robert Hensel, Davide Ceolin, and Marian Van der Meij. Georeferencing animal specimen datasets. *Transactions in GIS*, 19(4):563–581, 2015. doi:10.1111/TGIS.12110.

54 Marieke van Erp, Robert Hensel, Davide Ceolin, and Marian Van der Meij. Georeferencing animal specimen datasets. *Transactions in GIS*, 19(4):563–581, 2015. doi:10.1111/TGIS.12110.

55 Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, and Christopher B. Jones. Georeferencing wikipedia documents using data from social media sources. *ACM Trans. Inf. Syst.*, 32(3), July 2014. doi:10.1145/2629685.

56 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

57 Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*, 2023. doi:10.48550/arXiv.2305.02105.

58 Jimin Wang and Yingjie Hu. Enhancing spatial and textual analysis with eupeg: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23(6):1393–1419, 2019. doi:10.1111/tgis.12579.

59 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

60 Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023. doi:10.48550/arXiv.2302.11382.

61 John Wieczorek, Qinghua Guo, and Robert Hijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International journal of geographical information science*, 18(8):745–767, 2004. doi:10.1080/13658810412331280211.

62 Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 955–964, 2011. URL: https://aclanthology.org/P11-1096/.

63 Jinwen Xu and Ran Tao. Map reading and analysis with gpt-4v (ision). *ISPRS International Journal of Geo-Information*, 13(4):127, 2024. doi:10.3390/IJGI13040127.

64 Chris Yesson, Peter W Brewer, Tim Sutton, Neil Caithness, Jaspreet S Pahwa, Mikhaila Burgess, W Alec Gray, Richard J White, Andrew C Jones, Frank A Bisby, et al. How global is the global biodiversity information facility? *PloS one*, 2(11):e1124, 2007.

65 Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint*, 2024. arXiv:2404.16006.

66 Meishan Zhang, Bin Wang, Hao Fei, and Min Zhang. In-context learning for few-shot nested named entity recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10026–10030. IEEE, 2024. doi:10.1109/ICASSP48485.2024.10446653.

67 Qianqian Zhang, Mengdong Chen, and Lianzhong Liu. A review on entity relation extraction. In *2017 second international conference on mechanical, control and computer engineering (ICMCCE)*, pages 178–183. IEEE, 2017.

68 Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. doi:10.48550/arXiv.2210.03493.

69 Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. *arXiv preprint arXiv:2309.13339*, 2023. doi:10.48550/arXiv.2309.13339.

70 Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, 2021. doi:10.18653/V1/2021.NAACL-MAIN.5.

71 Mengjie Zhou, Liu Liu, Yiran Zhong, and Andrew Calway. Geolocation on cartographic maps with multi-modal fusion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5589–5596. IEEE, 2024. doi:10.1109/IROS58592.2024.10801404.