

# Coordinates from Context: Using LLMs to Ground Complex Location References

Tessa Masis (*they/them*)    Brendan O'Connor (*he/him*)

University of Massachusetts Amherst, MA, USA

{tmasis, brenocon}@cs.umass.edu

## Abstract

Geocoding is the task of linking a location reference to an actual geographic location and is essential for many downstream analyses of unstructured text. In this paper, we explore the challenging setting of geocoding compositional location references. Building on recent work demonstrating LLMs' abilities to reason over geospatial data, we evaluate LLMs' geospatial knowledge versus reasoning skills relevant to our task. Based on these insights, we propose an LLM-based strategy for geocoding compositional location references. We show that our approach improves performance for the task and that a relatively small fine-tuned LLM can achieve comparable performance with much larger off-the-shelf models.<sup>1</sup>

## 1 Introduction

Extracting geospatial information from unstructured text is essential for many downstream analyses such as disaster response (Kumar and Singh, 2019), disease surveillance (Lee et al., 2013), and historical event analysis (Tateosian et al., 2017). Geocoding is the task of linking a location reference to an actual geographic location, usually represented by a set of coordinates or an ID in a geographic database.

While there is a large body of work on geocoding locations referred to with explicit names (e.g. "in Gaza City"), there has been little attention given to locations referred to with compositional descriptions involving other toponyms (e.g. "about midway between Winsford and Crewe") which occur frequently in domains that do not have names for geographic regions of interest, e.g. environmental impact statements, food security outlooks (Laparra and Bethard, 2020). In such a setting, existing geocoding tools are not able to link the location

<sup>1</sup>Our best-performing fine-tuned model is available at <https://huggingface.co/tmasis/geocoding-complex-location-references>

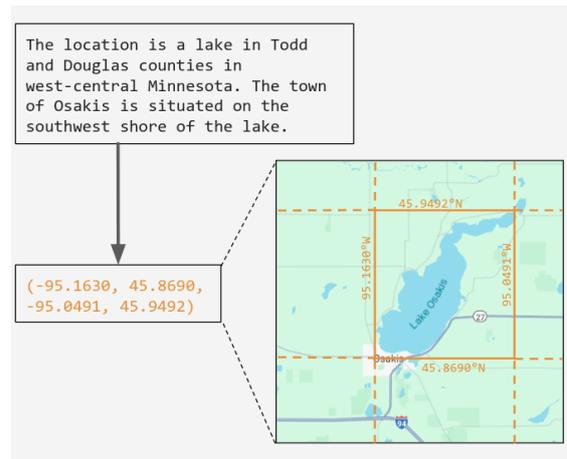


Figure 1: An illustrative example of our task. Given a compositional location description, we expect the model to predict the location's bounding box (defined by two pairs of latitude-longitude coordinates).

being described by the entire phrase, only locations mentioned within the phrase (Zhang and Bethard, 2024). In addition, most geocoding tools can only link to locations already in a geographic database, which often do not include compositional locations.

In this paper, we address the task of geocoding compositional location descriptions (see Figure 1) and how LLMs can best be leveraged for it. Our work makes the following contributions:

- We provide findings on LLMs' geospatial *knowledge* versus geospatial *reasoning* abilities through task-specific isolated evaluations of each. We find that off-the-shelf LLMs struggle more with generating accurate geospatial knowledge than with reasoning over it.
- Based on these insights, we propose the first end-to-end strategy for geocoding compositional location references, leveraging both the geospatial reasoning capabilities of LLMs and the geospatial knowledge bases of traditional geoparsers. Furthermore, our novel

use of bounding boxes<sup>2</sup> to link locations provides a simple and informative way to ground locations that do not exist in a geographic database.

- We thoroughly evaluate components of our approach, demonstrating that it outperforms prior work and that a relatively small fine-tuned LLM can achieve comparable performance with much larger off-the-shelf models.

## 2 Related Work

**Geocoding**, also known as toponym resolution, aims to link some reference to a geographic entity to the correct reference in a target database or coordinate system (Zhang and Bethard, 2024). The majority of previous work has focused on geocoding location references that have an explicit location name, with difficult settings typically concerning ambiguous location names (e.g. "Paris" may refer to Paris, TX, USA or Paris, FR). Some researchers have looked at geocoding location references without explicit location names, such as indirect location references (e.g. "the Austrian capital") or nonstandard variants of location names (e.g. "Beantown" refers to Boston, MA, USA) (Dredze et al., 2013; DeLozier et al., 2015; Kulkarni et al., 2021; Masis and O'Connor, 2024).

Very little attention has been given to geocoding complex or compositional location references, which do not contain an explicit location name and instead describe the location in relation to other toponyms (Al-Olimat et al., 2019). Thus far, there is only one dataset for this setting (Laparra and Bethard, 2020). The authors introduce five baseline methods for the task, although none using pre-trained language models.

Compositional location references must be linked to a coordinate system, not to a geographical database (e.g. GeoNames<sup>3</sup> or OpenStreetMap<sup>4</sup>), since compositional locations often do not exist in geographic databases. While some prior work has examined geocoding without such databases (DeLozier et al., 2015; Kulkarni et al., 2021; Sharma et al., 2023), they typically link locations to point coordinates which do not represent large or compositional geographic entities well. In our work, we

link locations to a bounding box, which defines an area in terms of four sets of coordinates and allows us to ground locations much more precisely.

**Geoparsing** is the broader task of both identifying and linking location references in unstructured text (Wang and Hu, 2019), essentially combining toponym recognition (Hu et al., 2023a) and geocoding. As with geocoding, no geoparsing work to date has addressed compositional location references and methods primarily link location references to a geographic database or a single set of coordinates.

**LLMs' geospatial knowledge and reasoning skills** have begun to be evaluated in a growing body of work. (We define *geospatial*, following the ISO's definition<sup>5</sup>, as an 'implicit or explicit reference to a location relative to the Earth' typically involving coordinates or addresses.) Regarding geospatial knowledge, recent work has suggested that LLMs learn and use an internal model of space which encodes locations' approximate real-world geo-coordinates (Gurnee and Tegmark, 2024; Chen et al., 2023). Related studies have examined the disparities and differences in LLMs' geospatial knowledge due to factors such as language used (Faisal and Anastasopoulos, 2023; Li et al., 2024) or model size (Godey et al., 2024; Bhandari et al., 2023). To help explain how LLMs may acquire such knowledge, an examination of the Common Crawl Corpus – often used in the pretraining of LLMs – found that 19% of documents contained geospatial data and that a prominent source was Google Maps URLs, which frequently contain both coordinates and a natural language location name (Ilyankou et al., 2024). However, to our knowledge no work has yet empirically evaluated LLMs as geospatial knowledge bases for downstream tasks.

Regarding geospatial reasoning, recent work evaluating off-the-shelf LLMs has emphasized their promise but also their limitations for geospatial reasoning tasks (Ji et al., 2025). Researchers have found that LLMs such as GPT-4o and Claude 3 Sonnet have limited success with reasoning over geospatial data, struggling to translate qualitative natural language terms to geospatial contexts (O'Sullivan et al., 2024) and performing poorly at tasks like simple route planning (Xu et al., 2024). Importantly, these studies show that LLMs have the capability to reason over geospatial information in both natural language and geo-coordinate formats.

Thus far, no work has directly compared LLMs'

<sup>2</sup>A *bounding box* is a rectangular area defined by two longitudes and two latitudes, and written in the standard format of  $\{lon_{min}, lat_{min}, lon_{max}, lat_{max}\}$  where  $lon \in [-180, 180]$  and  $lat \in [-90, 90]$ .

<sup>3</sup><https://www.geonames.org>

<sup>4</sup><https://www.openstreetmap.org>

<sup>5</sup><https://isotc211.geolexica.org/concepts/202/>

geospatial knowledge versus reasoning performances. Furthermore, no work has explored LLMs’ geospatial knowledge of or ability to reason over geo-entities defined by bounding boxes.

Using LLMs for geoparsing has recently been explored, with methods primarily using LLMs for the toponym recognition step (Hu et al., 2024; Harrod et al., 2024; Xu et al., 2024; Hu et al., 2023b), to generate training data (Yan and Lee, 2024), or to embed natural language descriptions of point coordinate locations (He et al., 2025). Thus far, LLMs have not been used for generating geospatial information or for reasoning-heavy settings.

### 3 Disentangling LLMs’ Geospatial Knowledge versus Reasoning

Here, we briefly evaluate LLMs for their geospatial knowledge generation versus geospatial reasoning abilities (see Figure 2). These two capabilities – that is, generating relevant factual knowledge and reasoning over provided factual knowledge – are necessary for the task of geocoding compositional locations (see task definition in §4.1, below), so isolated evaluations of each will help us understand how to best leverage LLMs for this task.

Accordingly, for our geospatial knowledge baseline we prompt the LLM to generate a structured geospatial representation (either the center coordinate or the bounding box) of a location *identified unambiguously by name* (Fig. 2, top half). Previous work only examined LLMs’ knowledge of geo-entities represented as center coordinates – here, we include both center coordinate and bounding box representations, allowing us to directly compare LLMs’ knowledge of each. We include both the location’s name and country in the prompt, since a name can refer to multiple locations (e.g. "San Jose" is a city in both Costa Rica and the USA) and including country information reduces reasoning needed to resolve such ambiguities.

For our geospatial reasoning baseline, we prompt the LLM to generate the bounding box of an *unidentified location described in relation to other geo-entities* (Fig. 2, bottom half). These prompts include a description of the location and the center coordinates of locations mentioned in the description. The described location’s name is not included so that the only way to predict its bounding box is to reason over the given information.

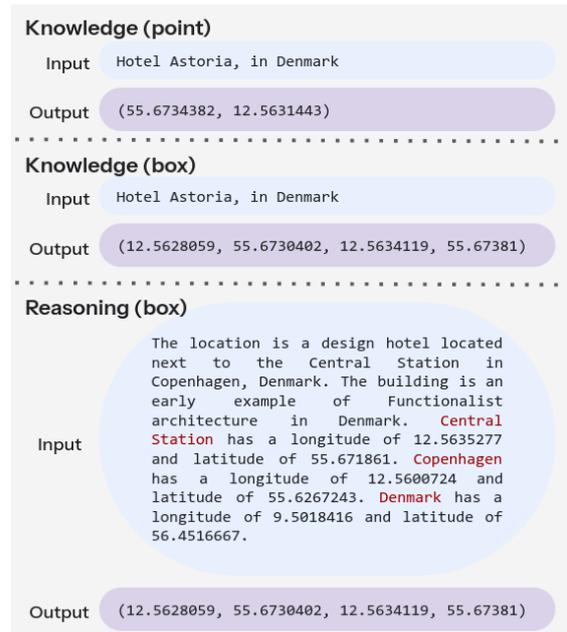


Figure 2: Illustrative examples of our LLM geospatial knowledge and reasoning baselines; all examples refer to the same location. Knowledge baseline: the LLM is given only the location’s name. Reasoning baseline: the LLM is given only a description of the location and mentioned locations’ coordinates (mentioned location names in red, for emphasis). Point refers to a center coordinate prediction (written as  $\{lat, lon\}$ ); box refers to a bounding box prediction (written as  $\{lon_{min}, lat_{min}, lon_{max}, lat_{max}\}$ ).

#### 3.1 Experimental Setup

**Data.** We use the GEOCODE dataset<sup>6</sup>, which contains location descriptions from English Wikipedia articles that are linked to the OpenStreetMap (OSM) geodatabase (Laparra and Bethard, 2020). For the current evaluation, we use GEOCODE’s test set of 1,000 manually curated compositional location descriptions. Created by randomly sampling Wikipedia pages, the dataset includes locations from all continents.

**Representing locations.** By leveraging the OSM links in GEOCODE, we retrieve gold geospatial information to represent each described location and location mentioned in a description. Locations can be represented as center coordinates, bounding boxes, or polygons, each with its own pros and cons. *Center coordinates*, which represent a location as a point, are used by the vast majority of geocoding datasets and are the simplest way to ground a location that does not exist in a geographic database. While adequate for point loca-

<sup>6</sup><https://github.com/EgoLaparra/geocode-data>

tions or small landmarks, they fall short when representing larger entities such as states or countries. *Polygons* represent locations with an unlimited sequence of coordinates defining a closed shape, so are much more complex and thus rarely used; at present, there are no practical end-to-end geocoding systems using polygons. *Bounding boxes* represent a location as a rectangle-shaped area via two latitude-longitude coordinate pairs, efficiently encoding basic size and shape.

In this work, we primarily represent locations with bounding boxes. While not used before for geocoding, we argue that bounding boxes offer a happy medium between center coordinates and polygons – they can encode areal information about the location while still being simple enough that off-the-shelf LLMs can reason over them.

**Metrics.** We use five metrics to evaluate predictions. Two are common geocoding metrics: *coverage* is the percentage of examples where the model produces a prediction, and *distance error* is the distance between the predicted and gold center coordinates or bounding box centroids. The other three metrics, introduced in Laparra and Bethard, measure the overlapping area between a predicted and gold bounding box: *area precision* is the fraction of overlapping area out of the entire predicted area, *area recall* is the fraction of overlapping area out of the entire gold area, and *area F1-score* is the harmonic mean of the precision and recall (calculated at test set-level, following prior work). Since these metrics measure area overlap, they are not used to evaluate center coordinate predictions.

**Models.** We evaluate three off-the-shelf open-source LLMs: Qwen2.5-72B-Instruct, Llama3.3-70B-Instruct, and Llama3.1-405B-Instruct. The models are evaluated in a few-shot manner, with two examples included in the system prompt (see Appendix A for full prompts used).

### 3.2 Results

Table 1 reports the results of our geospatial knowledge versus reasoning baselines (also see Table 6 in Appendix). We observe that all models are decidedly better at the geospatial reasoning baseline than the geospatial knowledge one, with lower average distance errors by more than 140km and consistent increases in area F1-score. The Qwen 72B model appears to have the biggest gap between knowledge and reasoning, with a difference of 300km between distance errors and of .164 be-

Experiment	LLM	Dist. ↓	F1 ↑
Knowledge (point)	Qwen 72B	326.1	–
	Llama 70B	215.4	–
	Llama 405B	216.6	–
Knowledge (box)	Qwen 72B	333.3	.087
	Llama 70B	221.8	.137
	Llama 405B	263.8	.195
Reasoning (box)	Qwen 72B	<b>36.2</b>	<b>.251</b>
	Llama 70B	<u>82.9</u>	<u>.239</u>
	Llama 405B	92.4	.235

Table 1: Evaluations from our geospatial knowledge and reasoning baselines, with average distance error (km) and area F1-scores. Point refers to a center coordinate prediction; box refers to a bounding box prediction. The best scores are bolded, the top 2 are underlined.

tween F1-scores. Additionally, increasing model size appears to improve geospatial knowledge more than geospatial reasoning (e.g. F1-score for Llama 70B versus 405B is .137 versus .195 for bounding box knowledge, but .239 versus .235 for reasoning). Finally, we note that all models have slightly better performances when predicting a point than a bounding box for the knowledge baseline, with lower distance errors of 10 to 50km.

While previous work has argued that LLMs have some amount of geospatial knowledge, the evidence has mostly demonstrated learned connections between location names (usually cities) and their center coordinates (Bhandari et al., 2023; Gurnee and Tegmark, 2024). We explore LLMs’ geospatial knowledge with both center coordinates and bounding boxes, a more complex geo-entity representation, and our results caution against using LLMs as reliable geospatial knowledge bases in this setting. While the LLMs appear to perform slightly better when predicting a point than a bounding box, which may be due to point representations of geo-entities being a more common form of geospatial data in LLM pretraining data (Ilyankou et al., 2024), both settings have relatively high average distance errors of over 200km.

On the other hand, our results show that LLMs perform promisingly on our geospatial reasoning baseline, with average distance errors of less than 35km. This supports prior work demonstrating that LLMs can reason over geospatial information encoded as both natural language and complex geometries (Ji et al., 2025), even if complex geome-

tries are not frequently in their pretraining data.

## 4 Geocoding Complex Location References

The above analysis reveals how LLMs can reason over geospatial information in both natural language and coordinate-based formats, but struggle with accurately generating such information. This motivates us to augment LLMs’ geospatial knowledge in our proposed geocoding method. In this section, we will first formally define our task and then explain our proposed and baseline methods.

### 4.1 Task Definition

We formally define our task of geocoding compositional descriptions as follows. Each input to the geocoding system is a text description  $d_l$  of a location  $l$ , where the set of locations mentioned in description  $d_l$  is referred to as  $M_l = \{m_1, m_2, \dots, m_{|M_l|}\}$ . The description  $d_l$  does not include the name of location  $l$ .

The geocoding system is a combination of two functions (Figure 3). The first,

$$g(m_i) = q_i$$

maps a mentioned location  $m_i$  to corresponding geographical information  $q_i$ , such as the mentioned location’s center coordinates. We refer to  $g$  as the recaller. The second function

$$f(d_l, M_l, g) = b_l$$

maps the location description and information about locations it mentions to the minimum bounding box  $b_l$  enclosing the location  $l$ . We refer to  $f$  as the reasoner.

### 4.2 Our Approach

In our proposed approach, we use a traditional geoparsing tool to augment an LLM’s geospatial knowledge (we refer to our approach as Geoparser-augmented). We first use a traditional geoparsing tool as the recaller, mapping any location mentioned in the compositional description to a set of latitude and longitude coordinates. We then use an LLM as the reasoner, which uses the location description and mentioned locations’ coordinates to generate the described location’s bounding box.<sup>7</sup> We evaluate both off-the-shelf and fine-tuned LLMs; see §5.1 for training details.

<sup>7</sup>The reasoner for the Geoparser-augmented approach has the same expected inputs and outputs as the reasoning baseline in §3.

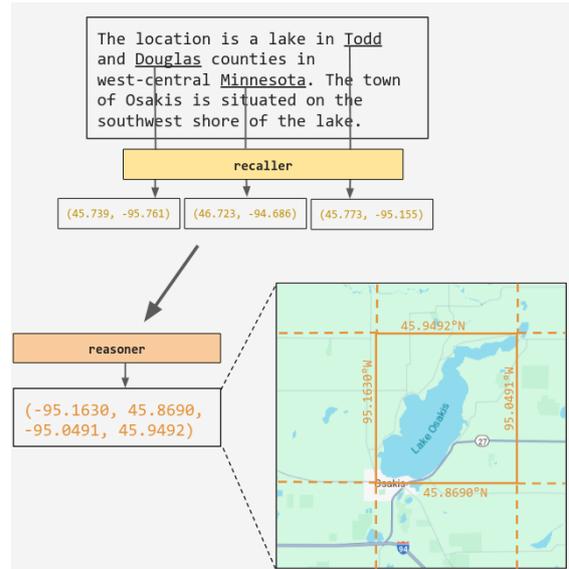


Figure 3: We define our geocoding task as a combination of two functions. The first function (the recaller) retrieves mentioned locations’ geographical information, e.g. center coordinates. The second function (the reasoner) uses both the location description and mentioned location information to generate the described location’s bounding box.

By augmenting with a traditional geoparser, we are able to add structured information from an external knowledge source to the LLM’s current context. This approach is related to the broader field of tool-augmented LLMs, in which LLMs leverage external tools (Mialon et al., 2023). While traditional geoparsing tools cannot reason over compositional location descriptions or provide complex geospatial information (such as bounding boxes), they can link mentioned locations to a pair of coordinates. The mentioned locations’ coordinates can then serve as an accurate geospatially grounded starting point for the LLM to reason over where the described location is in relation to them.

In our experiments, we first simulate an oracle geoparser by retrieving locations’ center coordinates from the OSM database. This allows us to focus on the variation in performance introduced by LLMs, not by third-party geoparsing tools. We then evaluate a more realistic setup with a real geoparser as the recaller. For this, we use the Google Maps API<sup>8</sup> which offers a popular geoparsing service broadly recognized as reliable (Ozer et al., 2020).

<sup>8</sup><https://developers.google.com/maps/documentation/geocoding>

Approach	reasoner	Distance (km) ↓	AreaPrec ↑	AreaRec ↑	AreaF1 ↑
Direct	Llama 8B	4427.8	.000	.000	.000
*Geo.-aug.	Llama 8B	7658.5	.006	.037	.010
Direct	Llama 70B	120.0	.200	.171	.184
*Geo.-aug..	Llama 70B	82.9	.166	<b>.428</b>	<u>.239</u>
Direct	Llama 405B	<u>69.0</u>	<b>.305</b>	.205	<b>.245</b>
*Geo.-aug.	Llama 405B	92.4	.164	<u>.413</u>	.235
Direct	FT Llama 8B	88.9	.132	.094	.110
*Geo.-aug.	FT Llama 8B	<b>28.0</b>	<u>.237</u>	.186	.208
Geo.-aug.	FT Llama 8B	152.2	.231	.170	.196

Table 2: Results comparing Geoparser-augmented versus Direct approaches for Llama models on GEOCODE dataset, with average distance error, area precision, area recall, and area F1-scores. FT refers to a fine-tuned model; \* denotes an oracle geoparser. The best scores are bolded, the top 2 are underlined.

Approach	reasoner	Distance (km) ↓	AreaPrec ↑	AreaRec ↑	AreaF1 ↑
Direct	Qwen 14B	1156.4	.058	.050	.054
*Geo.-aug.	Qwen 14B	152.1	.134	<u>.498</u>	.211
Direct	Qwen 72B	<u>114.7</u>	.113	.162	.133
*Geo.-aug.	Qwen 72B	<b>36.2</b>	<u>.191</u>	.368	<u>.251</u>
Direct	FT Qwen 14B	1944.1	.037	.122	.057
*Geo.-aug.	FT Qwen 14B	135.9	.165	<b>.499</b>	.248
Geo.-aug.	FT Qwen 14B	240.9	<b>.203</b>	.384	<b>.266</b>

Table 3: Results comparing Geoparser-augmented versus Direct approaches for Qwen models. FT refers to a fine-tuned model; \* denotes an oracle geoparser. The best scores are bolded; the top 2 are underlined.

### 4.3 Baselines

We evaluate several variations of our method as baselines.

**Direct:** we use an LLM for the reasoner but no explicit system for the recaller. In other words, an LLM must reason over just the natural language location description to generate the described location’s bounding box. As with our main approach, we evaluate both off-the-shelf and fine-tuned LLMs for the reasoner (see §5.1 for details).

**End-to-end LLM:** we use LLMs for both the recaller and the reasoner. We first prompt the recaller to generate point coordinates for the mentioned locations; for this, we use off-the-shelf Llama 70B since it had the best performance in our geospatial knowledge baseline in §3. We then prompt the reasoner to use both the location description and LLM-generated point coordinates to generate the described location’s bounding box. This is the same as the reasoner for the Geoparser-augmented approach, so we use the best perform-

ing fine-tuned model identified in that setting. This LLM-only pipeline offers a more convenient alternative to the Geoparser-augmented approach.

We also include results from the best performing method in prior work. **GBSP:** in this method, the reasoner is a grammar-based semantic parser and uses polygons of mentioned locations (Laparra and Bethard, 2020). We note that this method is not a practical end-to-end system since there is no recaller, and currently no system exists which can return polygons of mentioned locations.

## 5 Experiments

### 5.1 Experimental setup

**Data.** We use the GEOCODE dataset, which is discussed in §3.1. We use its manually curated test set as our test set. GEOCODE also has a train set of over 350K examples, which may contain simpler descriptions (i.e. not compositional) or broken OSM links. For our supervised fine-tuning

Approach	reasoner	Coverage (%) $\uparrow$	Distance (km) $\downarrow$	AreaF1 $\uparrow$
Geo.-aug.	FT Qwen 14B	90.8	<b>240.9</b>	<b>.266</b>
End-to-end LLM	FT Qwen 14B	<b>95.9</b>	627.5	.167
GBSP	Semantic parser	52.8	–	.240

Table 4: Results comparing Geoparser-augmented, End-to-end LLM, and GBSP. The best scores are bolded.

experiments, we use a subset of 13K examples randomly sampled from this train set.

**Models and training.** We evaluate a range of state-of-the-art open-source models across sizes and from multiple families. For our methods that use an off-the-shelf model, we evaluate Llama3.1-8B-Instruct, Llama3.3-70B-Instruct, Llama3.1-405B-Instruct (Dubey et al., 2024), Qwen3-14B-Instruct (Yang et al., 2025a), and Qwen2.5-72B-Instruct (Yang et al., 2025b). For methods that use a fine-tuned model, we evaluate Llama3.1-8B-Instruct and Qwen3-14B-Instruct.<sup>9</sup>

For fine-tuning, we conduct supervised fine-tuning and use parameter-efficient fine-tuning via low-rank adaptation (Hu et al., 2021). Separate models are fine-tuned for each approach, with training examples including the appropriate mentioned location geospatial information. We fine-tune low-rank adapter matrices for 1 epoch with a learning rate of  $2e^{-4}$  and batch size of 4. See Appendix A for more details, including all prompts used.

## 5.2 Results

**Geoparser-augmented and Direct.** We see that the Geoparser-augmented approach generally performs better than the Direct approach (Tables 2 and 3; complete tables in Appendix B). All models, with the exception of Llama 405B, have lower distance errors and higher F1-scores with the Geoparser-augmented approach. When using the Google Maps geoparser instead of the oracle geoparser, F1-scores do not change substantially although average distance errors do increase. The fine-tuned Qwen 14B model actually has a slightly higher area F1-score when using the Google Maps geoparser, although it also has relatively lower coverage (91% versus 96%).

Interestingly, we find that the Direct approach often results in better area precision than area recall while the opposite is true for the Geoparser-augmented approach, which generally has better area recall than area precision. The former indi-

cates smaller, more precise predicted areas while the latter indicates larger, less precise predicted areas. This is surprising, since the Geoparser-augmented method provides the LLM with accurate geospatial information about mentioned locations which should hypothetically allow the predicted bounding box to be *more* precise and should not necessarily affect the predicted bounding box’s size. We note a similar phenomenon in our earlier experiments in §3, where LLMs had better area precision than recall on the knowledge baseline and better area recall than precision on the reasoning baseline (Table 6). There, too, the method that does not provide mentioned locations’ geospatial data has higher area precision while the method that does provide them has higher area recall. We further investigate this phenomenon below in §5.3.

**End-to-end LLM and GBSP.** In Table 4, we compare the Geoparser-augmented approach with End-to-end LLM and GBSP (complete table in Appendix B). We observe that the End-to-end LLM approach has relatively poor performance compared to the Geoparser-augmented, with a lower F1-score by .1 and a higher distance error by 390km. These results support our earlier experiments in cautioning against using LLMs as geospatial knowledge bases. We note that End-to-end LLM does perform better than the Direct method using the same model (e.g. F1-scores of .167 versus .057; Table 3), indicating that using an LLM to generate mentioned locations’ geospatial information is better than providing no information at all.

We also observe that the GBSP method has a relatively high F1-score of .240 but a low coverage, only making predictions for 53% of the test set.

**Comparing LLMs.** Comparing model families, we see a similar pattern as in §3 where Qwen models (Table 3) perform slightly better when mentioned locations’ geospatial information is provided and Llama models (Table 2) have a slight edge when it is not. This suggests that Llama models may have better geospatial knowledge representations while Qwen models may have better

<sup>9</sup>We evaluated Qwen3 models in the non-thinking mode.

geospatial reasoning abilities, which would support prior work demonstrating Qwen’s stronger reasoning abilities (Wang et al., 2025). However, we note that disparities in performance are more dependent on model size than model family.

We also observe that while fine-tuned models do not surpass the performance of larger off-the-shelf models, they can achieve comparable performance. We see this most with the Geoparser-augmented approach, where fine-tuning models results in slightly bigger gains than with the Direct approach (e.g. fine-tuning improves Llama 8B’s F1-score by .1 more and Qwen 14B’s F1-score by .03 more in the Geoparser-augmented setting).

**Overall.** Our best performance is achieved by Qwen models with the Geoparser-augmented approach – Qwen 72B has an F1-score of .251 and an average distance error of 36km, and fine-tuned Qwen 14B has an F1-score of .248 with an oracle geoparser and .266 with the Google Maps geoparser. Both of these methods outperform GBSP, the best performing method from prior work, despite having access to less geospatial information (i.e. mentioned locations’ center coordinates versus polygons). We note that the fine-tuned Qwen 14B model achieves comparable performance to much larger off-the-shelf models, with a similar F1-score to Qwen 72B and higher F1-scores than Llama 70B or 405B, although also higher distance errors.

We offer a few practical suggestions for researchers interested in using LLMs for geocoding complex location references:

- For best overall performance or to prioritize area recall or distance errors, we recommend the Geoparser-augmented approach with a smaller fine-tuned LLM (e.g. fine-tuned Llama 8B or Qwen 14B) or a medium off-the-shelf model (e.g. Qwen 72B, Llama 70B). Since we found that scaling model size is less effective for improving geospatial reasoning than knowledge (§3), using the largest available LLM with the Geoparser-augmented approach is likely unnecessary.
- On the other hand, if researchers do not have access to a geoparser or are interested in prioritizing area precision, we recommend using the Direct approach with a large off-the-shelf model (e.g. Llama 405B). This is due to our findings that greater precision is elicited not by providing geospatial information from an

external knowledge source, but by relying on an LLM’s parametric knowledge.

### 5.3 Qualitative analysis

To further explore our results, we conducted a manual error analysis comparing trends in LLM outputs from the different methods.

First, an error across model families was flipping the signs for latitude and/or longitude (e.g. "95.163" instead of "-95.163") in a predicted bounding box. We hypothesize that this error may occur due to how coordinates can represent distinctions between different hemispheres by either using positive/negative signs or cardinal directions. (e.g. "45°N" versus "45°S"). Both formats are widely used and both are likely present in LLM pretraining data corpora, so this may hinder LLMs from learning a consistent meaning for a negative or positive sign. We note that this error type occurred less with larger models and with the Geoparser-augmented approach, and more with Qwen versus Llama models (see Table 10).

Next, we investigated the phenomenon identified in §5.2 regarding the Geoparser-augmented approach surprisingly predicting larger and less precise bounding boxes than the Direct approach. We observed that in LLM outputs from the Geoparser-augmented setting, the LLM would often select the maximum and minimum latitude/longitude coordinates from the mentioned locations’ center coordinates and simply use these (or slightly modified versions) for the final predicted bounding box (see examples in Table 5). Since the mentioned locations often neighbor or contain the described location, their coordinates frequently lie outside the described location so a bounding box using these coordinates will result in a larger predicted area. We observed this trend across language families and model sizes, although fine-tuned and larger models seem to do more reasoning about which mentioned locations’ coordinates are actually relevant for estimating the described location’s bounding box. (LLM outputs from the Direct method typically contained only the final predicted bounding box with no step-by-step reasoning traces.)

This indicates that although providing LLMs with accurate geospatial information is generally helpful for our task, the LLMs may tend to over-rely on the information without reasoning appropriately about its relationship or relevance to the desired output. This behavior is consistent with findings that LLMs perform much worse when ir-

relevant contextual information is included (Shi et al., 2023) and that their ability to filter out irrelevant information decreases for more complex reasoning tasks (Zhou et al., 2025). It may also be related to recent work demonstrating that reasoning models (e.g. DeepSeek-R1, Qwen3) have a strong tendency to defer to provided contextual knowledge, even when it contradicts parametric knowledge (Marjanović et al., 2025).

## 6 Conclusion

In this paper, we have introduced new methods for a challenging geocoding setting. After showing that, for our task, LLMs’ geospatial reasoning skills are stronger than their geospatial knowledge, we proposed the first end-to-end methods for geocoding compositional location descriptions. Our methods achieve state-of-the-art performance in both Direct and Geoparser-augmented settings, and we offer practical options for downstream use cases.

In addition, our novel use of bounding boxes allows us to effectively ground locations that do not exist in geographic databases and to evaluate methods with overlap-based metrics. Bounding boxes would likely be useful for other difficult geocoding settings, such as inherently vague geo-entities (e.g. "the Midwest") (Jones et al., 2008). We encourage researchers to incorporate this way of representing locations into geocoding datasets and tools.

## Limitations

An important limitation of this study is that we evaluate our methods on a single dataset which only includes English language examples. Since this dataset is the only existing one for our task of grounding compositional location references, we emphasize the strong need to create more datasets for this task which cover broader domains and languages. An additional limitation is the other possible factors which may affect performance for our task which this study does not investigate, including length of location description, number of mentioned locations in a description, frequency of an entity in pretraining data, and so on. We leave investigations of these factors on performance to future work. Finally, we acknowledge that our proposed methods rely heavily on LLMs and are thus reliant on third parties for sustaining them.

## Ethical Considerations

While the domain we evaluate here is that of formal, impersonal writing (i.e. Wikipedia), we note that it is possible to apply geocoding methods to more personal writing (e.g. social media posts), which could risk de-anonymizing online users through the inference of sensitive location information (Kruspe et al., 2021; Dupre et al., 2022). Due to these potential privacy issues in eventual applications, we have chosen to evaluate only open-weight models that, unlike most proprietary models, do not require sending potentially sensitive data to external servers. We also point to important recent work on geomasking techniques, which aim to protect the privacy of individuals while preserving spatial information in geodata (Lorestani et al., 2024).

## Acknowledgements

This material is based upon work supported by a National Science Foundation Graduate Research Fellowship (1938059) and NSF CAREER (1845576). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

- Unslloth AI, Daniel Han-Chen, and Michael Han-Chen. 2025. Unslloth. <https://github.com/unslothai/unsloth>.
- Hussein S Al-Olimat, Valerie L Shalin, Krishnaprasad Thirunarayan, and Joy Prakash Sain. 2019. Towards geocoding spatial expressions (vision paper). In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 75–78.
- Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.
- Geoff Boeing. 2025. Modeling and analyzing urban networks and amenities with OSMnx. *Geographical Analysis*.
- Yida Chen, Yixian Gan, Sijia Li, Li Yao, and Xiaohan Zhao. 2023. More than correlation: Do large language models learn causal representations of space? *arXiv preprint arXiv:2312.16257*.
- Grant DeLozier, Jason Baldrige, and Loretta London. 2015. Gazetteer-independent toponym resolution

- using geographic word profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.
- Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter Geolocation System with Applications to Public Health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, volume 23, page 45. Citeseer.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Sam Dupre, Lindsay Spell, and Paul Jung. 2022. Geospatial Data Disclosure Avoidance and the Census. *Select Topics in International Censuses (STIC 1, 1 (2022))*, 1–9.
- Fahim Faisal and Antonios Anastasopoulos. 2023. Geographic and geopolitical biases of language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163.
- Nathan Godey, Éric Villemonte De La Clergerie, and Benoît Sagot. 2024. On the scaling laws of geographical representation in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12416–12422.
- Wes Gurnee and Max Tegmark. 2024. [Language models represent space and time](#). In *The Twelfth International Conference on Learning Representations*.
- Karlynn Harrod, Prabin Bhandari, and Antonios Anastasopoulos. 2024. From Text to Maps: LLM-Driven Extraction and Geotagging of Epidemiological Data. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 258–270.
- Junlin He, Tong Nie, and Wei Ma. 2025. Geolocation representation from large language models are generic enhancers for spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17094–17104.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Xuke Hu, Jens Kersten, Friederike Klan, and Sheikh Mastura Farzana. 2024. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science*, pages 1–28.
- Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2023a. Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys*, 56(5):1–37.
- Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhnpal, Ryan Zhenqi Zhou, and Kenneth Joseph. 2023b. 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.
- Ilya Ilyankou, Meihui Wang, Stefano Cavazzi, and James Haworth. 2024. Quantifying Geospatial in the Common Crawl Corpus. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 585–588.
- Yuhan Ji, Song Gao, Ying Nie, Ivan Majić, and Krzysztof Janowicz. 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*, pages 1–38.
- Christopher B Jones, Ross S Purves, Paul D Clough, and Hideo Joho. 2008. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1065.
- Anna Kruspe, Matthias Häberle, Eike J Hoffmann, Samyo Rode-Hasinger, Karam Abdulahhad, and Xiao Xiang Zhu. 2021. Changes in Twitter geolocations: Insights and suggestions for future usage. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 212–221.
- Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. 2021. Multi-Level Gazetteer-Free Geocoding. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 79–88.
- Abhinav Kumar and Jyoti Prakash Singh. 2019. Location reference identification from tweets during emergencies: A deep learning approach. *International Journal of Disaster Risk Reduction*, 33:365–375.
- Egoitz Laparra and Steven Bethard. 2020. A dataset and evaluation framework for complex geographical description parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 936–948.
- Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using Twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1474–1477.

- Bryan Li, Samar Haider, and Chris Callison-Burch. 2024. This land is your, my land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871.
- Mahrokh Abdollahi Lorestani, Thilina Ranbaduge, and Thierry Rakotoarivelo. 2024. Privacy risk in geodata: A survey. *arXiv preprint arXiv:2402.03612*.
- Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakhia, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. 2025. DeepSeek-R1 Thoughtology: Let’s think about LLM Reasoning. *arXiv preprint arXiv:2504.07128*.
- Tessa Masis and Brendan O’Connor. 2024. [Where on earth do users say they are?: Geo-entity linking for noisy multilingual user input](#). In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 86–98, Mexico City, Mexico. Association for Computational Linguistics.
- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *Transactions on Machine Learning Research*, 2023.
- Kent O’Sullivan, Nicole R Schneider, and Hanan Samet. 2024. Metric reasoning in large language models. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 501–504.
- Murat Ozer, Michael Zidar, Rustu Deryol, Said Varlioglu, Ibrahim Sevki Eldivan, and Halil Akbas. 2020. Creating a real-time geocoding system: implications of open source for the public safety. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1185–1188. IEEE.
- Praval Sharma, Ashok Samal, Leen-Kiat Soh, and Deepti Joshi. 2023. [A spatially-aware data-driven approach to automatically geocoding non-gazetteer place names](#). *ACM Trans. Spatial Algorithms Syst.*, 10(1).
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.
- Laura Tateosian, Rachael Guenter, Yi-Peng Yang, and Jean Ristaino. 2017. Tracking 19th century late blight from archival documents using text analytics and geoparsing. In *Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings*, volume 17, page 17.
- Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23(6):1393–1419.
- Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*.
- Liuchang Xu, Shuo Zhao, Qingming Lin, Luyao Chen, Qianqian Luo, Sensen Wu, Xinyue Ye, Hailin Feng, and Zhenhong Du. 2024. Evaluating large language models on spatial tasks: A multi-task benchmarking study. *arXiv preprint arXiv:2408.14438*.
- Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2025b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Zeyu Zhang and Steven Bethard. 2024. A survey on geocoding: algorithms and datasets for toponym resolution. *Language Resources and Evaluation*, pages 1–22.
- Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. 2025. GSM-Infinite: How Do Your LLMs Behave over Infinitely Increasing Context Length and Reasoning Complexity? *arXiv preprint arXiv:2502.05252*.

## A Additional Experimental Details

**Data.** For preprocessing the GEOCODE dataset, we used the *Beautiful Soup 4* package.<sup>10</sup> For accessing OSM data, we used the *OSMnx 2* package (Boeing, 2025).

OSM data is available under the Open Database License.<sup>11</sup> The license for the GEOCODE dataset is unspecified,<sup>12</sup> but our use of the dataset is consistent with its intended use of benchmarking tools for geocoding compositional location references.

**Computational experiments.** We used *together.ai*<sup>13</sup> for running off-the-shelf LLMs and *unsloth* (AI et al., 2025) for fine-tuning models.

Our models were fine-tuned on Nvidia V100 GPUs with 16GB memory. Experiments for fine-tuning and running models required approximately 100 GPU hours in total. All results reported are from a single run (i.e. not averaged across runs).

**Prompts.** For knowledge and reasoning baseline experiments in §3, we use Prompts 1, 2, and 3 for LLM inference.

```
SYSTEM: You are a system that returns the
*center coordinates* of a given
location or landmark. The coordinates
are a pair of numbers defining the
location's latitude and longitude, where
latitude is a decimal number between
-90.0 and 90.0 and longitude is a
decimal number between -180.0 and 180.0.
Follow the standard format of (latitude
, longitude). Here are some examples
with the expected output format:
Input: The Eiffel Tower, in France.
Output: (48.858, 2.2959)
Input: Brazil, in South America.
Output: (-14.243, -53.189)
USER: Input: {{location_name}}
Output:
```

Prompt 1: For geospatial knowledge (point) baseline, where the input is a location name and the output is predicted center coordinates.

For approaches introduced in §4 for geocoding complex location references, we use Prompt 3 for the Geoparser-augmented reasoner and Prompt 4 for the Direct reasoner. For the End-to-end LLM approach, we use Prompt 5 for the recaller and Prompt 3 for the reasoner. When evaluating

```
SYSTEM: You are a system that returns the
*bounding box* of a given location or
landmark. A bounding box is an area
defined by two longitudes and two
latitudes, where latitude is a decimal
number between -90.0 and 90.0 and
longitude is a decimal number between
-180.0 and 180.0. Follow the standard
format of (min longitude, min latitude,
max longitude, max latitude). Here are
some examples with the expected output
format:
```

```
Input: The Eiffel Tower, in France.
Output: (2.293, 48.857, 2.297, 48.859)
```

```
Input: Brazil, in South America.
Output: (-73.983, -33.750, -34.793,
5.270)
```

```
USER: Input: {{location_name}}
Output:
```

Prompt 2: For geospatial knowledge (box) baseline, where the input is a location name and the output is predicted bounding box.

fine-tuned models, the examples in system prompts were excluded.

## B Additional Results

Table 6 reports complete results from §3. Tables 7, 8, and 9 report complete results for §5.2. Table 5 includes erroneous examples discussed in §5.3.

<sup>10</sup><https://www.crummy.com/software/BeautifulSoup/>

<sup>11</sup><https://www.openstreetmap.org/copyright>

<sup>12</sup><https://github.com/EgoLaparra/geocode-data>

<sup>13</sup><https://www.together.ai/>

**SYSTEM:** You are a system that returns the *\*bounding box\** of a described location or landmark, by using a description and the center longitude and latitude of related locations. A bounding box is an area defined by two longitudes and two latitudes, where latitude is a decimal number between -90.0 and 90.0 and longitude is a decimal number between -180.0 and 180.0. Follow the standard format of (min longitude, min latitude, max longitude, max latitude). Here are some examples with the expected output format:

Input: The location is a wrought-iron lattice tower on the Champ de Mars in Paris, France. It is named after the engineer Gustave Eiffel, whose company designed and built the tower from 1887 to 1889. Champ de Mars has a longitude of 48.855 and latitude of 2.296. Paris has a longitude of 48.859 and latitude of 2.264.  
Output: (2.293, 48.857, 2.297, 48.859)

Input: The location is the largest and easternmost country in South America. South America has a longitude of -13.591 and latitude of -109.712.  
Output: (-73.983, -33.750, -34.793, 5.270)

**USER:** Input: `{{location_description}}`  
`{{mentioned_location_center_coordinates}}`  
Output:

Prompt 3: For both the geospatial reasoning baseline and the Geoparser-augmented reasoner, where the input is a location description with mentioned locations' center coordinates and the output is predicted bounding box.

**SYSTEM:** You are a system that returns the *\*bounding box\** of a described location or landmark. A bounding box is an area defined by two longitudes and two latitudes, where latitude is a decimal number between -90.0 and 90.0 and longitude is a decimal number between -180.0 and 180.0. Follow the standard format of (min longitude, min latitude, max longitude, max latitude). Here are some examples with the expected output format:

Input: The location is a wrought-iron lattice tower on the Champ de Mars in Paris, France. It is named after the engineer Gustave Eiffel, whose company designed and built the tower from 1887 to 1889.  
Output: (2.293, 48.857, 2.297, 48.859)

Input: The location is the largest and easternmost country in South America.  
Output: (-73.983, -33.750, -34.793, 5.270)

**USER:** Input: `{{location_description}}`  
Output:

Prompt 4: For the Direct approach reasoner, where the input is a location description and the output is predicted bounding box.

**SYSTEM:** You are a system that returns the *\*center coordinates\** for each location mentioned in a given paragraph. The coordinates are a pair of numbers defining each location's latitude and longitude, where latitude is a decimal number between -90.0 and 90.0 and longitude is a decimal number between -180.0 and 180.0. Here are some examples with the expected output format:

Input: The location is a wrought-iron lattice tower on the Champ de Mars in Paris, France. It is named after the engineer Gustave Eiffel, whose company designed and built the tower from 1887 to 1889.  
Output: Champ de Mars has a longitude of 48.855 and latitude of 2.296. Paris has a longitude of 48.859 and latitude of 2.264.

Input: The location is the largest and easternmost country in South America.  
Output: South America has a longitude of -13.591 and latitude of 109.712.

**USER:** Input: `{{location_description}}`  
Output:

Prompt 5: For the End-to-end LLM recaller, where the input is a location description and the output is mentioned locations' predicted center coordinates.

Prompt: The location is a gulf that connects the Arabian Sea with the Strait of Hormuz, {...} *If geo.-aug. approach:* Arabian Sea has a longitude of 63.002662154702726 and latitude of 12.4368972. Strait of Hormuz has a longitude of 56.20277021626677 and latitude of 26.449406099999997. {...}  
 Expected output: **(56.2683402, 22.4824554, 61.8012822, 25.9456285)**

Approach, Model	LLM output
Direct, Llama 70B	<b>(54.983, 22.983, 66.417, 26.750)</b>
Geo.-aug., Llama 70B	To find the bounding box of the described location, we need to determine the minimum and maximum longitudes and latitudes. {...} The minimum longitude is approximately <b>51.1972 (Persian Gulf)</b> and the maximum longitude is approximately <b>63.0027 (Arabian Sea)</b> and then further to 71.2475 (Pakistan), but since the gulf is between the Arabian Sea and the Strait of Hormuz, which then runs to the Persian Gulf, the maximum longitude should be around the Arabian Sea. {...} So the maximum longitude should be around 63.0027. The minimum latitude is approximately <b>12.4369 (Arabian Sea)</b> and the maximum latitude is approximately <b>32.6475 (Iran)</b> . So, the bounding box is approximately: <b>(51.197, 12.437, 63.003, 32.648)</b>
Direct, Llama 405B	<b>(58.240, 23.700, 63.320, 26.750)</b>
Geo.-aug., Llama 405B	{...} - Minimum longitude: {...} we consider the longitude of the United Arab Emirates (53.9994829) and <b>the Persian Gulf (51.197231065873154)</b> . The minimum longitude is 51.197231065873154, but since {...} Therefore, we choose 51.197231065873154 as the minimum longitude, but round it to 51.2 for simplicity. - Minimum latitude: The location is bounded by Oman on the south, so we consider the latitude of <b>Oman (21.0000287)</b> . The minimum latitude is 21.0000287, but we round it to 21.0 for simplicity. - Maximum longitude: {...} we consider the longitude of <b>the Arabian Sea (63.002662154702726)</b> {...} Therefore, we choose 63.002662154702726 as the maximum longitude, but round it to 63.0 for simplicity. - Maximum latitude: The location is bounded by Iran on the north, so we consider the latitude of <b>Iran (32.6475314)</b> . The maximum latitude is 32.6475314, but we round it to 32.7 for simplicity. The bounding box of the location is <b>(51.2, 21.0, 63.0, 32.7)</b> .
Direct, FT Qwen 14B	<b>(58.0, 23.0, 62.0, 28.0)</b>
Geo.-aug., FT Qwen 14B	{...} From these coordinates, the <b>**minimum longitude**</b> is approximately <b>**51.197** (Persian Gulf)</b> , and the <b>**maximum longitude**</b> is approximately <b>**63.003** (Arabian Sea)</b> . The <b>**minimum latitude**</b> is approximately <b>**21.000** (Oman)</b> , and the <b>**maximum latitude**</b> is approximately <b>**32.648** (Iran)</b> . #### Final Answer: <b>**51.197, 21.000, 63.003, 32.648</b>

Table 5: Erroneous examples of LLMs overrelying on contextual geospatial information. With the Geoparser-augmented approach, LLMs tend to select bounding box coordinates directly from provided center coordinates of mentioned locations; since these center coordinates can be some distance from the actual borders of the described location, this often results in larger and less precise predictions than from the Direct approach which relies on the model’s parametric knowledge. For example, when used with the Geoparser-augmented approach, all three models in this table predict the Persian Gulf’s longitude (51.2) as the described location’s minimum longitude, much lower than its actual minimum longitude (56.3). The same three models with the Direct approach have a wider range of predictions (55.0, 58.2, 58.0) but are closer to the actual value.

Experiment	LLM	Cov. (%)	Dist. (km) ↓	AreaPrec ↑	AreaRec ↑	AreaF1 ↑
Knowledge (point)	Qwen 72B	96.3	326.1	–	–	–
	Llama 70B	99.7	215.4	–	–	–
	Llama 405B	97.9	216.6	–	–	–
Knowledge (box)	Qwen 72B	95.6	333.3	.086	.087	.087
	Llama 70B	<u>99.1</u>	221.8	<u>.199</u>	.105	.137
	Llama 405B	91.4	263.8	<b>.361</b>	.134	.195
Reasoning (box)	Qwen 72B	98.5	<b>36.2</b>	.191	.368	<b>.251</b>
	Llama 70B	98.2	<u>82.9</u>	.166	<b>.428</b>	<u>.239</u>
	Llama 405B	<b>99.5</b>	92.4	.164	<u>.413</u>	.235

Table 6: Complete evaluation metrics from our geospatial knowledge and reasoning baselines. Point refers to a center coordinate prediction; box refers to a bounding box prediction. The best scores are bolded, the top 2 are underlined.

Approach	reasoner	Cov. (%)	Distance (km) ↓	AreaPrec ↑	AreaRec ↑	AreaF1 ↑
Direct	Llama 8B	100.0	4427.8	.000	.000	.000
*Geo.-aug.	Llama 8B	25.2	7658.5	.006	.037	.010
Direct	Llama 70B	100.0	120.0	.200	.171	.184
*Geo.-aug..	Llama 70B	98.2	82.9	.166	<b>.428</b>	<u>.239</u>
Direct	Llama 405B	99.4	<u>69.0</u>	<b>.305</b>	.205	<b>.245</b>
*Geo.-aug.	Llama 405B	99.5	92.4	.164	<u>.413</u>	.235
Direct	FT Llama 8B	100.0	88.9	.132	.094	.110
*Geo.-aug.	FT Llama 8B	100.0	<b>28.0</b>	<u>.237</u>	.186	.208
Geo.-aug.	FT Llama 8B	100.0	152.2	.231	.170	.196

Table 7: Results comparing Geoparser-augmented versus Direct approaches for Llama models on GEOCODE dataset. FT refers to a fine-tuned model; \* denotes an oracle geoparser. The best scores are bolded, the top 2 are underlined.

Approach	reasoner	Cov. (%)	Distance (km) ↓	AreaPrec ↑	AreaRec ↑	AreaF1 ↑
Direct	Qwen 14B	99.1	1156.4	.058	.050	.054
*Geo.-aug.	Qwen 14B	100.0	152.1	.134	<u>.498</u>	.211
Direct	Qwen 72B	95.9	<u>114.7</u>	.113	.162	.133
*Geo.-aug.	Qwen 72B	98.5	<b>36.2</b>	<u>.191</u>	.368	<u>.251</u>
Direct	FT Qwen 14B	94.7	1944.1	.037	.122	.057
*Geo.-aug.	FT Qwen 14B	96.1	135.9	.165	<b>.499</b>	.248
Geo.-aug.	FT Qwen 14B	90.8	240.9	<b>.203</b>	.384	<b>.266</b>

Table 8: Results comparing Geoparser-augmented versus Direct approaches for Qwen models. FT refers to a fine-tuned model; \* denotes an oracle geoparser. The best scores are bolded, the top 2 are underlined.

<b>Approach</b>	reasoner	<b>Cov. (%)</b>	<b>Distance (km) ↓</b>	<b>AreaPrec ↑</b>	<b>AreaRec ↑</b>	<b>AreaF1 ↑</b>
Geo.-aug.	FT Qwen 14B	90.8	<b>240.9</b>	.203	.384	<b>.266</b>
E2E LLM	FT Qwen 14B	95.9	627.5	.100	<b>.496</b>	.167
GBSP	Semantic parser	52.8	–	<b>.213</b>	.276	.240

Table 9: Results comparing the best performing models from GBSP, End-to-end LLM, Direct, and Geoparser-augmented approaches. The best scores are bolded.

<b>Approach</b>	<b>Model</b>	<b>Frequency</b>
Direct	Llama 70B	2
Geo.-aug.	Llama 70B	1
Direct	Llama 405B	0
Geo.-aug.	Llama 405B	0
Direct	FT Llama 8B	0
Geo.-aug.	FT Llama 8B	0
Direct	Qwen 14B	15
Geo.-aug.	Qwen 14B	1
Direct	Qwen 72B	3
Geo.-aug.	Qwen 72B	0
Direct	FT Qwen 14B	3
Geo.-aug.	FT Qwen 14B	1

Table 10: Flipped sign error frequencies for each model and approach from 100 random examples. We observe that the error occurs less with larger models and with the Geoparser-augmented approach.