AGILE: GIScience Series, 7, 46, 2026. https://doi.org/10.5194/agile-giss-7-46-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Should We Ask an LLM? Evaluating Toponym Disambiguation across Administrative Levels

Franz Welscher [1], Paddy Smith [2], Tatu Leppämäki [3], and Ilya Ilyankou [4]

[1]Department of Geoinformatics, University of Salzburg, Salzburg, Austria
[2]School of Geography, University of Leeds, Leeds, UK
[3]Digital Geography Lab, Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland
[4]SpaceTimeLab, Department of Civil, Environmental, and Geomatic Engineering, UCL, London, UK

Correspondence: Franz Welscher (franz.welscher@plus.ac.at)

Abstract.

Toponym disambiguation, determining which real-world location a place name refers to, is a critical step in geoparsing pipelines, yet existing evaluations mix disambiguation quality with the behavior of downstream geocoders through distance-based metrics. We propose evaluating disambiguation as a standalone task by prompting nine LLMs to predict administrative containment (ADM0 to ADM2) from textual context, scoring predictions directly with precision, recall, and F_1 against GADM-derived labels on the LGL corpus. Performance declines systematically with administrative granularity, mid-sized models outperform the largest tested model, and recurring failure cases cluster around geopolitically complex regions. These findings suggest that feeding fine-grained LLM disambiguation outputs to geocoders may harm rather than help performance.

Submission Type. analysis, dataset

BoK Concepts. [GC] Geocomputation, [GC3] Artificial intelligence (AI) in EO and GI, [GC3-8] Computational Linguistics

Keywords. LLMs, toponyms, geoparsing, geocoding, disambiguation

1 Introduction

A wide range of semi- and unstructured text sources, such as social media posts, news articles, and historical documents, contain abundant geographic information in the form of toponyms and other geographical references (Hu et al., 2024). The task of extracting this information is called geoparsing, which supports applications including disaster response from social media (Gelernter and Balaji, 2013; Avvenuti et al., 2018; Wang et al., 2020), urban environments analysis (Brunila et al., 2023), and even spatial approaches to literary studies (Gregory et al., 2019; Alex et al., 2019). Geoparsing is commonly described as a two-step pipeline consisting of (1) toponym *recognition*, which detects location mentions in text, and (2) toponym *resolution*, which links each mention to a real-world location representation. A key sub-problem in *resolution* is toponym *disambiguation*, which determines which geographic entity a toponym refers to given its context, because many names recur across regions and scales (Hu et al., 2024; Ju et al., 2016) (e.g., *Lebanon*, the country in the Middle East[1], and *Lebanon*, the town in Connecticut[2]). In other words, disambiguation identifies an unambiguous reference (e.g. administrative containment), while resolution maps that reference to a concrete target such as coordinates or a gazetteer entry.

In GIScience, Large Language Models (LLMs) have recently been applied to a broad range of tasks, including geoparsing and related information extraction (Roberts et al., 2023; Abbasi et al., 2025; Xu et al., 2025; Mai et al., 2023). For toponym recognition, LLMs often outperform baseline or task-specific models, although performance depends on model choice and often higher recall is achieved over lower precision (Hu et al., 2023b; Mai et al., 2023; Kim and Lee, 2024; Zhang et al., 2024). For toponym resolution, recent work uses LLMs to infer geographic context (or an unambiguous reference) that can improve subsequent geocoding (Hu et al., 2024; Yin et al., 2025). This differs from traditional approaches typically combining candidate generation with heuristic rules or learned ranking/classification models (Hu et al., 2023a). LLMs exhibit substantial geographic knowledge (Manvi et al., 2024), likely due to the prevalence of geospatial texts in their pre-training corpora (Ilyankou et al., 2024). This helps explain the recent shift towards

[1]https://en.wikipedia.org/wiki/Lebanon
[2]https://en.wikipedia.org/wiki/Lebanon,_Connecticut

generative, context-aware disambiguation methods for geoparsing, in which models use local context and general world knowledge to resolve ambiguity (Kibria et al., 2024). Across domains such as historical text analysis and biomedical entity linking, LLMs have proven effective as entity disambiguators, often by refining or selecting from candidate lists produced by retrieval-based systems (Hiltmann et al., 2025; Ye and Mitchell, 2025).

In the domain of geoparsing, Hu et al. (2024) fine-tuned lightweight LLMs to generate unambiguous references for toponyms and reported improved geocoder resolution performance across common geoparsing corpora, while Yin et al. (2025) used GPT-4o to extract implicit geographic information from text and images to improve the geolocation of disaster-related social media content.

However, these approaches primarily evaluate end-to-end resolution using distance-based metrics, which mixes two different components: (1) whether the model correctly disambiguated the intended place (i.e., correct country or region), and (2) whether a downstream geocoder successfully maps that information to the correct coordinates. As a result, it remains unclear how well LLMs perform on disambiguation *itself*, and what kinds of errors they make when deriving an unambiguous reference from context. This matters because prior work has observed that LLMs can generate plausible but incorrect references for less prominent locations (Hu et al., 2024), and such errors may be hidden or amplified by geocoders.

Therefore, this paper proposes a novel approach for evaluating the toponym disambiguation capabilities of LLMs independently of geocoding. Rather than using distance-based metrics that mix disambiguation with the success of a downstream geocoding algorithm, we evaluate disambiguation as a standalone task by measuring how well LLMs predict the administrative containment of toponyms from textual context. We build an evaluation dataset from the widely used Local Global Lexicon (LGL) corpus (Lieberman et al., 2010) and enrich each toponym with administrative labels by spatially intersecting GeoNames coordinates with administrative boundaries from the Global Administrative Areas[3] (GADM) database. Using a consistent prompting setup and structured JSON outputs, we evaluate nine models from the Qwen, Mistral, and Microsoft Phi families, and report precision, recall, and F_1 scores per administrative level.

Our study addresses the following research questions:

1. How can LLMs' toponym disambiguation capabilities be evaluated independently of geocoding metrics?

2. How well do LLMs disambiguate toponyms to administrative levels from context?

3. How does LLM model family and size affect disambiguation performance?

We find that performance is highest for larger administrative areas (ADM0) and declines for more granular areas (ADM1, ADM2). Further, larger model size alone does not guarantee better disambiguation. Our key contributions are: (1) A evaluation protocol based on precision, recall, and F_1 metrics that isolates disambiguation from geocoding, (2) an automatically derived benchmark built on the LGL and GADM datasets, and (3) a comparative analysis across nine LLMs that highlights administrative level specific strengths, failure cases, and implications for integrating LLM outputs into downstream geocoding pipelines.

2 Methodology

2.1 Evaluation Corpus

For evaluation, we use the widely adopted Local Global Lexicon (LGL) dataset (Lieberman et al., 2010), a global collection of local news articles with toponyms annotated using GeoNames[4] coordinates. The corpus is frequently used in geoparsing research and provides a suitable basis for benchmarking toponym disambiguation under realistic news-style contexts (Gritta et al., 2020). The original LGL contains $5,057$ annotated toponym mentions. However, 625 of these lack coordinates and were removed from our evaluation to ensure that all toponyms can be spatially joined to their related administrative areas. The dataset has a strong geographic imbalance, as it is heavily skewed toward North America (approximately $3,300$ toponyms), which needs to be considered in the interpretation of the results. After this initial analysis and filtering the dataset contains $4,381$ toponyms, which we enriched with their related administrative areas in a next step.

We automatically annotated each toponym with the names of the administrative units it belongs to at three hierarchical levels: ADM0 (*countries* and sovereign entities), ADM1 (primary sub-national units, such as *states*), and ADM2 (second-order administrative divisions, such as *counties*). For this purpose, we rely on GADM as a globally consistent source of administrative boundaries. We limit the evaluation to ADM2 because GADM provides global coverage reliably up to this level. Levels beyond ADM2 are not consistently available worldwide, which would reduce comparability across countries and regions. For each toponym, ADM-level labels are derived by spatially intersecting its GeoNames coordinates with GADM polygons. In cases where a mention conceptually corresponds only to a higher level administrative area (e.g., ADM0), lower-level fields are represented as empty strings.

[3] https://gadm.org

[4] https://www.geonames.org/

We use the GeoNames feature codes of the toponyms to determine which set of administrative levels they should be associated with. For example, a populated place such as a city (GeoNames feature code *PPL*) is expected to belong to an ADM0, ADM1, and ADM2 unit, whereas an independent political entity *PCLI* represents an ADM0 entity and should not be matched to ADM1 or ADM2. We keep large natural features that can span multiple administrative units, such as rivers, because the reference text may indicate the specific segment or area being referred to. Therefore, such mentions can still be meaningfully disambiguated to an administrative hierarchy if the context provides sufficient clues. Finally, we remove supranational toponyms like *Europe* or the *European Union* as they cannot be mapped to any of the administrative levels. This preprocessing results in the final evaluation dataset of $4,381$ toponyms with reference text and administrative area labels.

2.2 LLM Evaluation

We evaluate nine LLMs spanning different parameter scales and providers to assess how model family and size affect disambiguation performance. The providers (in bold) and models (in brackets) are **Qwen** (Qwen3-4B[5], Qwen3-14B[6], Qwen3-30B[7]), **Mistral** (Ministral3-3B[8], Ministral3-14B[9], Magistral-24B[10]) and **Microsoft** (Phi4-Mini-4B[11], Phi4-14B[12], Phi3.5-MoE-42B[13]). All models are used in 8-bit quantization for computational efficiency, while maintaining quality. We deploy the models using Podman[14] and a Vulkan[15] container on a server with an AMD Strix Halo integrated Radeon Graphics (Radeon 8050S/8060S).

Figure 1 illustrates a high-level overview of the workflow. We used similar prompts to the ones introduced by Hu et al. (2024). The system prompt (see Figure 2) aims to introduce the task to the LLM, which is to infer the administrative levels for the toponym from the context provided in the user prompt (see Figure 2). The context is the original text in which the toponym occurs; the toponym is explicitly labeled with the «START» and «END» tags. To standardize responses and simplify scoring, models are instructed to output a predefined JSON schema (see Appendix A) containing the name of the toponym and the names of the corresponding ADM0, ADM1 and ADM2 levels. For all model runs, we set the temperature to 0 to ensure consistency and reproducibility.

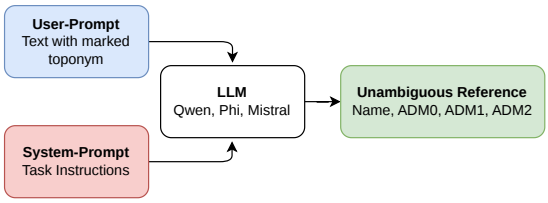

Figure 1. Overview of the methodology used to produce the unambiguous references of the toponyms.

System Prompt:
"Identify the unambiguous reference of {{TOPONYM}} (marked with <<START>> and <<END>> in the text. Return the result in the following JSON format: {{JSON_SCHEMA}}"

User Prompt:
"{{TEXT}}"

Text Example:
"<<START>>New York<<END>> is a city in the United States."

Figure 2. System Prompt and User Prompt used for the LLMs as well as an example of a tagged toponym in a text example.

We evaluate the disambiguation performance of the models for each administrative level separately using precision, recall, and F_1 scores. This level-wise evaluation captures whether LLMs are better at coarse disambiguation (countries) compared to finer administrative resolution (states or counties), and allows the analysis of where errors emerge along the hierarchy.

We compare the model's predicted ADM-level label with the corresponding target ADM-level label for each toponym. We estimate true positives (TP), false positives (FP), and false negatives (FN) using the following rules:

- **TP:** the predicted ADM label matches the target ADM label
- **FP:** the predicted ADM label does not match the target ADM label
- **FN:** the target ADM label is non-empty, but the predicted ADM label is empty

Because LLM outputs often include administrative suffixes (e.g., *Ottertail County*) while the target labels may store only the base name (*Ottertail*), we apply fuzzy string matching to determine matches. We use a low similarity threshold of 50 to count suffix-expanded outputs as correct while still rejecting clear mismatches. For example, pairs such as *(Clay County, Clay)* exceed this threshold (similarity score $= 53$) and are counted as true positives, whereas clearly unrelated pairs such as *(Hillsborough,*

[5] https://hf.co/bartowski/Qwen_Qwen3-4B-Instruct-2507-GGUF
[6] https://hf.co/unsloth/Qwen3-14B-GGUF
[7] https://hf.co/bartowski/Qwen_Qwen3-30B-A3B-Instruct-2507-GGUF
[8] https://hf.co/unsloth/Ministral-3-3B-Instruct-2512-GGUF
[9] https://hf.co/bartowski/mistralai_Ministral-3-14B-Instruct-2512-GGUF
[10] https://hf.co/unsloth/Magistral-Small-2506-GGUF
[11] https://hf.co/unsloth/Phi-4-mini-instruct-GGUF
[12] https://hf.co/microsoft/phi-4-gguf
[13] https://hf.co/bartowski/Phi-3.5-MoE-instruct-GGUF
[14] https://podman.io/
[15] https://hub.docker.com/layers/kyuz0/amd-strix-halo-toolboxes/vulkan-radv/images

Manchester) remain below the threshold (score $= 18$) and are counted as false positives. The metrics precision (P), recall (R) and F_1 scores are computed in the standard way (Gelernter and Balaji, 2013; Wang et al., 2020; Hu et al., 2023b) as shown in Equation 1:

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN} \qquad F_1 = \frac{2PR}{P+R} \tag{1}$$

Finally, we analyze performance differences across models and administrative levels and interpret what these results imply for practical use of LLMs in toponym disambiguation pipelines.

3 Results

Figure 3 summarizes the precision and recall of the nine evaluated LLMs for disambiguating toponyms to ADM0, ADM1, and ADM2. Overall, PHI4-14B performs best across levels, achieving a mean F_1 score of 0.92 ± 0.06[16], closely followed by MAGISTRAL-24B (0.91 ± 0.08). In contrast, the weakest tested models are PHI-3.5-MOE-42B (0.66 ± 0.09), followed by MINISTRAL3-3B (0.80 ± 0.15).

This ranking is also reflected at the level of individual administrative units. PHI-3.5-MOE-42B is consistently the poorest performer at all three levels, with F_1 scores of 0.75 (ADM0), 0.66 (ADM1), and 0.58 (ADM2). Its relatively low ADM0 performance is notable, because all other models achieve F_1 scores above 0.90 at ADM0, with PHI4-14B reaching a near-perfect 0.99. At ADM1, MAGISTRAL-24B slightly outperforms PHI4-14B with an F_1 score of 0.92. At ADM2, PHI4-14B leads with an F_1 score of 0.87, ahead of MAGISTRAL-24B by 0.06 points.

Model size alone does not explain performance differences. Larger models do not consistently yield better results. Instead, mid-sized models achieve the strongest average performance across levels (ADM0: 0.99, ADM1: 0.90, ADM2: 0.81). Further, model family is not a clear indicator for good or poor performance as the best and worst performing models come from the same family.

Across all models, two broader trends emerge. First, performance declines systematically from ADM0 to ADM2, indicating that finer-grained administrative disambiguation is harder. Second, models exhibit a precision–recall trade-off, which means some models primarily lose recall as administrative granularity increases (e.g., QWEN3-4B, QWEN3-30B, MAGISTRAL-24B, MINISTRAL3-14B), whereas others primarily lose precision (e.g., MINISTRAL3-3B and QWEN3-14B).

[16]Here and thereafter, we report the scores as *mean* $\pm$ *standard deviation*

Table 1. Top 10 False Negatives (missed places) for ADM0

Place Name	Corpus Count	Average FN	FN %
Iran	24	14.0	58.3
Iraq	32	18.0	56.3
Egypt	71	35.0	49.3
Pakistan	11	5.0	45.5
Syria	26	10.0	38.5
Sudan	49	16.0	32.7
France	16	4.0	25.0
Israel	71	15.5	21.8
China	12	2.5	20.8
Cyprus	11	1.0	9.1

Note: Only places that appear 10 times or more in the gold standard corpus are included.

Table 2. Top 10 False Negatives (missed places) for ADM1

Place Name	Corpus count	Average FN	FN %
Gaza	24	19.67	81.9
Tbilisi	11	8.50	77.3
Krasnodar	12	8.0	66.7
Dagestan	10	3.4	34.0
Alabama	10	3.0	30.0
Iowa	33	7.0	21.2
Michigan	24	5.0	20.8
Oklahoma	13	2.0	15.4
Indiana	110	16.38	14.9
Minnesota	137	18.88	13.8

Note: Only places that appear 10 times or more in the gold standard corpus are included.

Tables 1–3 further demonstrate disambiguation difficulty by listing the administrative units most frequently missed (false negatives). For each unit, we report its total frequency in the corpus (Corpus Count), the mean number of misses across models (Average FN), and the resulting false-negative proportion (FN %). This breakdown highlights geographic contexts where LLMs struggle to assign the correct administrative containment.

At ADM0 (Table 1), several countries show high false-negative rates, including Iran, Iraq, Egypt, Pakistan, Syria, Sudan, and Israel. Iran and Iraq stand out in particular, with more than 50% of toponyms associated with these countries missed on average. At ADM1 (Table 2), specific regions are especially challenging: Gaza (82%), Tbilisi (77%), and Krasnodar (67%) have high miss rates across models. These errors may reflect difficult geopolitical contexts, naming variation, or sparse representation in training data. Finally, ADM2 (Table 3) contains cases where models almost always fail. For example, toponyms located in Gori (a municipality in Georgia) are missed in roughly 90% of cases. One plausible explanation is systematic confusion between the U.S. state Georgia and the country Georgia, where Gori is located.

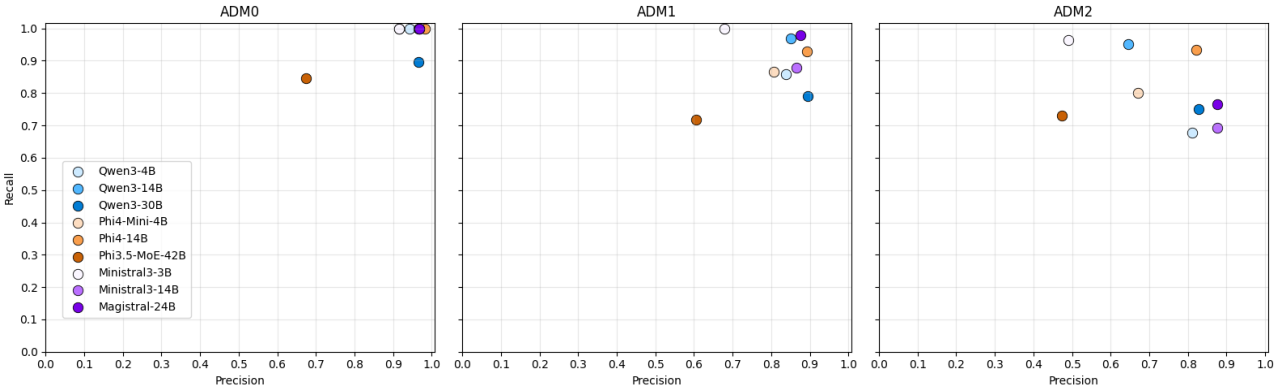

Figure 3. Precision and recall of tested LLMs on ADM levels 0, 1, and 2

Table 3. Top 10 False Negatives (missed places) for ADM2

Place Name	Corpus Count	Average FN	FN %
Gori	25	22.50	90.0
Jerusalem	13	11.13	85.6
Gaza	18	15.13	84.0
Tbilisi	11	9.0	81.8
Central	22	17.89	81.3
Cass	13	8.50	65.4
Greenville	15	9.40	62.7
Suffolk	11	6.67	60.6
Lancaster	13	7.38	56.7
Hartford	18	10.17	56.5

Note: Only places that appear 10 times or more in the gold standard corpus are included.

4 Discussion & Conclusion

We show that toponym disambiguation can be evaluated as a standalone task by asking LLMs to predict administrative containment from context and scoring these labels directly with precision, recall, and F_1 scores. This avoids mixing model performance with the behavior of downstream geocoders.

Further, we show that performance is strongest at ADM0 (best $F_1 = 0.99$) and declines through ADM1 (best $F_1 = 0.91$) to ADM2 (best $F_1 = 0.87$), indicating that finer-grained administrative inference remains substantially harder. The best models (PHI4-14B and MAGISTRAL-24B) achieve high overall F_1, but even they show a clear drop at ADM2, and the false-negative analysis highlights recurring 'hard geographies' (e.g. Gaza/Tbilisi/Krasnodar, Iran/Iraq) where models frequently miss the correct containment.

Further, model size alone is not a reliable indicator for good performance as mid-sized models perform best on average, while the largest model tested (PHI-3.5-MOE-42B) performs worst across all levels. A similar thing can be observed for model family as the best (PHI4-14B) and the worst (PHI-3.5-MOE-42B) performing model come from the same family. However, as our ground truth dataset is severely biased towards North America global generalizability of these results is limited.

Overall, our results have some implications for downstream tasks of LLM toponym disambiguation such as geocoding. The loss in precision with each administrative level suggests that providing fine-grained disambiguation information from LLMs to geocoders could worsen their performance compared to providing coarse, but more accurate disambiguation such as geocoding. The loss in precision with each administrative level suggests that providing fine-grained disambiguation information from LLMs to geocoders could worsen their performance compared to providing coarse, but more accurate disambiguation information. Further, our false-negative analysis suggests that geopolitical complexity, and prominent ambiguity pairs (e.g. Georgia the country vs the U.S. state) are predictable failures that need to be explicitly observed.

Multiple future work improvements follow from the limitations discussed here: (1) test the robustness under different decoding settings (e.g. temperature) to quantify stability, (2) repeat the evaluation on a more geographically balanced benchmark to reduce the North America bias, (3) broaden text genres (e.g. social media, historical text) where context density differs, (4) stress-test the fuzzy string matching threshold, and (5) test whether providing only administrative information that LLMs can disambiguate with high precision positively influences downstream geocoder performance. Further, (6) a closer analysis of the model outputs needs to be conducted to answer questions, such as *what are the reasons of the poor model performance of the largest model?* and *why do LLMs fail to identify certain places correctly?*.

Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools in the preparation of this manuscript. Specifically, the AI tools were utilized for improving the grammar and structure of the manuscript, as well as supporting the

coding of the experiments, but not for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance. Therefore, the authors take full responsibility for the contents of this manuscript.

Acknowledgments

P.S. was supported by Economic and Social Research Council (ESRC) via the White Rose Doctoral Training Partnership (WRDTP) [grant no. ES/P000746/1]. **T.L.** was supported by the Kone Foundation (project MOBICON) and the Research Council of Finland (Flagship of Advanced Mathematics for Sensing Imaging and Modelling, FAME, grant number 359182). **I.I.** was supported by Ordnance Survey & UKRI Engineering and Physical Sciences Research Council [grant no. EP/Y528651/1].

Data and Code Availability Statement

The data and code that support the findings of this study are available on github at https://github.com/PLUS-ZGIS-GeoAI/2026_AGILE_LLMDisambiguationEvaluation

Appendix A: JSON-Schema

Listing 1. JSON schema used for structured LLM outputs

```
{
  "type": "json_schema",
  "json_schema": {
    "schema": {
      "title": "DisambiguationSchema",
      "type": "object",
      "properties": {
        "name": {
          "title": "Name",
          "type": "string"
        },
        "ADM0": {
          "title": "Adm0",
          "type": "string",
          "default": "",
          "description": "Name of the country or
              independent political entity the toponym
              belongs to (e.g., United States, Germany,
              France)"
        },
        "ADM1": {
          "title": "Adm1",
          "type": "string",
          "default": "",
          "description": "Name of the first-level
              administrative division (e.g., state,
              province)"
        },
        "ADM2": {
          "title": "Adm2",
          "type": "string",
          "default": "",
          "description": "Name of the second-level
              administrative division (e.g., county,
              district)"
        }
      },
      "required": ["name"]
    }
  }
}
```

References

Abbasi, O. R., Welscher, F., Weinberger, G., and Scholz, J.: The World As Large Language Models See It: Exploring the Reliability of LLMs in Representing Geographical Features, https://doi.org/10.48550/ARXIV.2506.00203, 2025.

Alex, B., Grover, C., Tobin, R., and Oberlander, J.: Geoparsing historical and contemporary literary text set in the City of Edinburgh, Language Resources and Evaluation, 53, 651–675, https://doi.org/10.1007/s10579-019-09443-x, 2019.

Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., and Tesconi, M.: CrisMap: a Big Data Crisis Mapping System Based on Damage Detection and Geoparsing, Information Systems Frontiers, 20, 993–1011, https://doi.org/10.1007/s10796-018-9833-z, 2018.

Brunila, M., LaViolette, J., CH-Wang, S., Verma, P., Féré, C., and McKenzie, G.: Toward a Critical Toponymy Framework for Named Entity Recognition: A Case Study of Airbnb in New York City, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 4676–4695, Association for Computational Linguistics, Singapore, https://doi.org/10.18653/v1/2023.emnlp-main.284, 2023.

Gelernter, J. and Balaji, S.: An Algorithm for Local Geoparsing of Microtext, GeoInformatica, 17, 635–667, https://doi.org/10.1007/s10707-012-0173-8, 2013.

Gregory, I., Donaldson, C., and Taylor, J.: Landscape Appreciation in the English Lake District: A GIS Approach, in: Mapping Landscapes in Transformation: Multidisciplinary Methods for Historical Analysis, edited by Coomans, T., Cattoor, B., and De Jonge, K., Leuven University Press, Belgium, 2019.

Gritta, M., Pilehvar, M. T., and Collier, N.: A Pragmatic Guide to Geoparsing Evaluation: Toponyms, Named Entity Recognition and Pragmatics, Language Resources and Evaluation, 54, 683–712, https://doi.org/10.1007/s10579-019-09475-3, 2020.

Hiltmann, T., Dröge, M., Dresselhaus, N., Grallert, T., Althage, M., Bayer, P., Eckenstaler, S., Mendi, K., Schmitz, J. M., Schneider, P., Sczeponik, W., and Skibba, A.: NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach, https://doi.org/10.48550/arXiv.2502.04351, 2025.

Hu, X., Sun, Y., Kersten, J., Zhou, Z., Klan, F., and Fan, H.: How Can Voting Mechanisms Improve the Robustness and Generalizability of Toponym Disambiguation?, International Journal of Applied Earth Observation and Geoinformation, 117, 103 191, https://doi.org/10.1016/j.jag.2023.103191, 2023a.

Hu, X., Kersten, J., Klan, F., and Farzana, S. M.: Toponym Resolution Leveraging Lightweight and Open-Source Large Language Models and Geo-Knowledge, International

Journal of Geographical Information Science, pp. 1–28, https://doi.org/10.1080/13658816.2024.2405182, 2024.

Hu, Y., Mai, G., Cundy, C., Choi, K., Lao, N., Liu, W., Lakhanpal, G., Zhou, R. Z., and Joseph, K.: Geo-Knowledge-Guided GPT Models Improve the Extraction of Location Descriptions from Disaster-Related Social Media Messages, International Journal of Geographical Information Science, 37, 2289–2318, https://doi.org/10.1080/13658816.2023.2266495, 2023b.

Ilyankou, I., Wang, M., Cavazzi, S., and Haworth, J.: Quantifying Geospatial in the Common Crawl Corpus, in: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '24, pp. 585–588, Association for Computing Machinery, https://doi.org/10.1145/3678717.3691286, 2024.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., and McKenzie, G.: Things and Strings: Improving Place Name Disambiguation from Short Texts by Combining Entity Co-Occurrence with Topic Modeling, in: Knowledge Engineering and Knowledge Management, edited by Blomqvist, E., Ciancarini, P., Poggi, F., and Vitali, F., vol. 10024, pp. 353–367, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-49004-5_23, series Title: Lecture Notes in Computer Science, 2016.

Kibria, R., Dipta, S. I. U., and Adnan, M. A.: On Functional Competence of LLMs for Linguistic Disambiguation, in: Proceedings of the 28th Conference on Computational Natural Language Learning, edited by Barak, L. and Alikhani, M., pp. 143–160, Association for Computational Linguistics, https://doi.org/10.18653/v1/2024.conll-1.12, 2024.

Kim, H. and Lee, S.: POI GPT: Extracting POI Information from Social Media Text Data, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-4/W10-2024, 113–118, https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-113-2024, 2024.

Lieberman, M. D., Samet, H., and Sankaranarayanan, J.: Geotagging with Local Lexicons to Build Indexes for Textually-Specified Spatial Data, in: 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), pp. 201–212, IEEE, Long Beach, CA, USA, https://doi.org/10.1109/ICDE.2010.5447903, 2010.

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., Cundy, C., Li, Z., Zhu, R., and Lao, N.: On the Opportunities and Challenges of Foundation Models for Geospatial Artificial Intelligence, https://doi.org/10.48550/ARXIV.2304.06798, 2023.

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., and Ermon, S.: GeoLLM: Extracting Geospatial Knowledge from Large Language Models, http://arxiv.org/abs/2310.06213, 2024.

Roberts, J., Lüddecke, T., Das, S., Han, K., and Albanie, S.: GPT4GEO: How a Language Model Sees the World's Geography, https://doi.org/10.48550/ARXIV.2306.00020, 2023.

Wang, J., Hu, Y., and Joseph, K.: NeuroTPR: A Neuro-net Toponym Recognition Model for Extracting Locations from Social Media Messages, Transactions in GIS, 24, 719–735, https://doi.org/10.1111/tgis.12627, 2020.

Xu, L., Zhao, S., Lin, Q., Chen, L., Luo, Q., Wu, S., Ye, X., Feng, H., and Du, Z.: Evaluating Large Language Models on Geospatial Tasks: A Multiple Geospatial Task Benchmarking Study, International Journal of Digital Earth, 18, 2480 268, https://doi.org/10.1080/17538947.2025.2480268, 2025.

Ye, C. and Mitchell, C. S.: LLM as Entity Disambiguator for Biomedical Entity-Linking, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), edited by Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., pp. 301–312, Association for Computational Linguistics, https://doi.org/10.18653/v1/2025.acl-short.25, 2025.

Yin, W., Xue, Y., Liu, Z., Li, H., and Werner, M.: LLM-enhanced Disaster Geolocalization Using Implicit Geoinformation from Multimodal Data: A Case Study of Hurricane Harvey, International Journal of Applied Earth Observation and Geoinformation, 137, 104 423, https://doi.org/10.1016/j.jag.2025.104423, 2025.

Zhang, Y., Wang, Z., He, Z., Li, J., Mai, G., Lin, J., Wei, C., and Yu, W.: BB-GeoGPT: A Framework for Learning a Large Language Model for Geographic Information Science, Information Processing & Management, 61, 103 808, https://doi.org/10.1016/j.ipm.2024.103808, 2024.