Multilingual Synthetic Corpora for Geoparsing Using Large Language Models

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

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

Abstract

The construction of annotated geoparsing corpora is costly and has resulted in limited linguistic and geographic coverage, particularly outside English-speaking regions. This paper investigates the use of Large Language Models (LLMs) to generate synthetic geoparsing corpora for four languages in four regions: Austria (German), Belarus (Belarusian), Finland (Finnish), and Ghana (English). We evaluate the generated corpora through automatic and human quality checks, and by benchmarking state-of-the-art geoparsers on the synthetic data. Our results show that while LLM-generated corpora enable credible geoparser evaluation, low-resource regions and languages expose systematic limitations in LLM-powered synthetic text generation approaches related to underlying geographic data coverage, completeness, and linguistic variation. We make our code available on GitHub.[1]

Keywords

Geoparsing, toponym recognition, synthetic geospatial corpora, large language models (LLMs), automatic corpus construction

1. Introduction

Geoparsing involves (i) identifying spatial references within text, or *toponym recognition*, and (ii) linking them to a geographic location, or *toponym resolution* [1]. This process enables the extraction of geographic information from unstructured text for a variety of applications within GIS, and wider research areas [2, 3].

The creation of annotated text corpora is a crucial step within geoparsing, as it facilitates the training and evaluation of state-of-the-art methods. Whilst many geoparsing corpora have been constructed [4, 5, 6], the process is typically time-consuming and costly, due to the need to collect suitable digital texts and the demands of human annotation [7]. As a result, geoparsing corpora can be limited in size, and geographic and linguistic coverage. In particular, corpora are most often in the English language, and referring to places located within the English-speaking world [8].

Transferring a geoparsing system to new languages or geographical regions while maintaining adequate performance is not a trivial task. Natural language processing (NLP) algorithms, such as toponym recognizers, trained for major languages should not be assumed to generalize without issue to other languages, especially ones dissimilar in structure and vocabulary [9, 8]. The availability of baseline resources, such as gazetteer coverage [10] and georeferenced Wikipedia pages, which have been successfully exploited for automatic corpus generation [1], is another core issue for under-resourced

[1]https://github.com/ilyankou/multilingual-geoparsing-corpora

GeoExT 2026: Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026, April 2, 2026, Delft, The Netherlands

*Corresponding author.
†All authors contributed equally to this work.
ilya.ilyankou.23@ucl.ac.uk (I. Ilyankou); franz.welscher@plus.ac.at (F. Welscher); gy17pls@leeds.ac.uk (P. Smith); tatu.leppamaki@helsinki.fi (T. Leppämäki)
https://ilyankou.com/ (I. Ilyankou)
0009-0008-7082-7122 (I. Ilyankou); 0000-0003-2432-1880 (F. Welscher); 0009-0006-6955-1144 (P. Smith); 0000-0002-9634-7943 (T. Leppämäki)

languages and regions. In terms of spatiality, geoparsers have also been shown to exhibit spatial bias, that is, their toponym recognition and resolution performance varies by where the toponym is [11].

Synthetic text generation provides a potential solution to address the current limitations in building geoparsing corpora. The utilization of Large Language Models (LLMs) to generate human-like text in various text forms, could be used to build corpora for under-represented languages and geographic regions. LLMs have been successfully applied to toponym recognition and resolution [12, 13], and synthetic text generation has shown promise in addressing text sparsity in similar NLP tasks [14, 15] and for low-resource languages [16]. However, LLMs have yet to be applied to geoparsing corpus creation, mono- or multilingually, leaving open how such a pipeline should be designed, what language- and region-specific challenges emerge, and whether the resulting corpora yield meaningful geoparser benchmarks.

In this paper, we investigate generating synthetic geoparsing corpora for different geographic regions and languages using LLMs. We study if our semi-automatic approach [17] can generate synthetic text, guided by gazetteers (OSM) and auxiliary geographic context (Wikipedia), to any region or language of interest. We conduct initial experiments by applying our approach to four study areas with different languages: Austria (German), Belarus (Belarusian), Finland (Finnish), and Ghana (English).

In doing so, we seek to address the following research questions (RQs):

RQ1: How can we utilize LLMs, gazetteers and contextual information to generate synthetic texts for geoparsing?

RQ2: What are the challenges when using LLMs to create synthetic texts for geoparsing in different regions and languages?

RQ3: How do state-of-the-art toponym recognition approaches perform on the synthetic corpora across regions and languages according to established metrics?

2. Methodology

2.1. Corpus generation

The overview of our multilingual geoparsing text generation pipeline is illustrated in Figure 1. We extracted OpenStreetMap (OSM) features[1] with a name and a linked Wikipedia article in the target language (e.g., 'fi:wikipedia' in Finland), retaining 'admin_level' and the full tag set as context. We removed names containing parentheses and names with four or more words to avoid complex and overly specific locations; for Ghana, we additionally removed names ending in 'assembly' due to high prevalence of local government buildings in OSM.

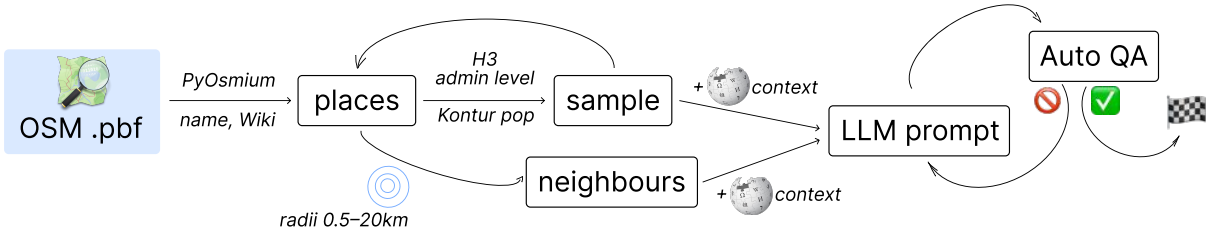

Figure 1: Overview of the multilingual geoparsing text generation pipeline.

We sampled 200 places each in Austria and Finland, and 50 places each in Belarus and Ghana; the difference is due to data sparsity within the OSM gazetteer and linked Wiki information. We used a three-step sampling method: (i) 2 places per H3 hexagonal cell at resolution 4, or roughly 26km

[1]We downloaded country-level OSM PBF files from Geofabrik (https://download.geofabrik.de/) and processed them using Osmium's Python library (https://docs.osmcode.org/pyosmium/latest/)

Country	Language	Generator model	QA model
Austria	German	Mistral-Nemo-Instruct-2407-12B	Qwen3-14B
Belarus	Belarusian	Qwen3-14B	Qwen3-14B
Finland	Finnish	Llama-Poro-2-8B-Instruct	Qwen3-14B
Ghana	English	Qwen3-14B	Qwen3-14B

Table 1
LLMs used for generation and QA

in edge length[2] (with replacement), (ii) 2 places per OSM 'admin_level' (with replacement), and (iii) an additional sample of equal size weighted by gridded population (Kontur[3], resolution 4) to ensure that roughly half of the sampled places come from populated areas. We concatenated these samples, de-duplicated by OSM id, and down-sampled to the country quota of 200 or 50 respectively.

For each sampled place, we selected up to three neighbouring places by expanding a search radius from 0.5 km, doubling until at least one neighbour was found (maximum 20 km), then randomly sampling up to three neighbours within that radius. For sampled places and neighbours, we retrieved the first five sentences of the linked Wikipedia lead section using the official API[4].

Given OSM tags and Wikipedia lead text for the focal place and its neighbours, we prompted an LLM to generate a 1-2 sentence description in a specified tone and to output a fixed JSON schema. A second LLM performed automated QA to ensure only allowed places were mentioned. We chose to use different LLMs for different languages (see Table 1) for performance reasons. We ran the models locally using LM Studio[5].

Slavic and Finnic languages inflect place names, unlike English and German (e.g., 'going to London', 'staying in London', 'left London' vs 'еду ў Лондан', 'застаюся ў Лондане', 'з'ехаў з Лондана'). To account for this, we additionally used an LLM to extract all surface forms of the provided place names occurring in the generated text to enable reliable span matching and linking back to place identifiers.

Each generated corpus was reviewed by an author of this work, who was assigned their native language, to assess linguistic coherence, correction of spatial relations, and completeness and accuracy of toponym annotations. Texts containing major errors were discarded prior to downstream evaluation.

2.2. Corpus evaluation

After corpus generation, we describe the characteristics of the final corpora for each region by analysing the frequency of place mentions, the distribution of place types, and the spatial distribution of place mentions.

We then deploy a range of *Toponym Recognition* models to evaluate their performance in identifying place names across the four study areas. As baselines, we use the language-specific spaCy models[6] as well as the multilingual variant, and additionally a multilingual RoBERTa model[7]. For Belarusian we use the ukrainian spaCy model, because there is no native Belarusian spaCy model and these two languages are closely related. We compare these baselines against three LLMs: Llama3.1-8B[8], Mistral3-14B[9], and Qwen3-14B[10], following a similar evaluation setup to Hu et al. [12].

The city, state, and country fields in the JSON output schema are predicted by the evaluation LLMs and are not part of the gold corpus. Our gold annotations consist solely of toponym spans, OSM identifiers and geometries. All toponym recognition evaluation metrics are computed using span and

[2]https://h3geo.org/docs/core-library/restable/
[3]https://data.humdata.org/dataset/kontur-population-dataset-22km
[4]https://en.wikipedia.org/w/api.php
[5]https://lmstudio.ai/
[6]https://spacy.io/models
[7]https://huggingface.co/julian-schelb/roberta-ner-multilingual
[8]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[9]https://lmstudio.ai/models/mistralai/ministral-3-14B-reasoning
[10]https://huggingface.co/lmstudio-community/Qwen3-14B-GGUF

INSTRUCTIONS: You will be given a single input field called TEXT. Your job is to detect every explicit mention of a real-world geographic place in that TEXT and return a structured annotation for each mention.

For each extracted location: copy the exact substring from TEXT into "name" (preserve punctuation/casing/spacing). Use context (and your geographic knowledge) to disambiguate the place and fill: "city" (city/town/village if applicable, else ''), "county" (county/district/municipality-equivalent if applicable, else ''), "state" (state/province/region-equivalent if applicable, else ''), and "country" (country name if it can be inferred, else ''). If a location appears multiple times as separate mentions, annotate each occurrence with its own offsets. If some administrative levels cannot be determined, use '' as placeholders. Keep "geometry" as null.

Example 1: TEXT: "Tamale Airport offers an exciting gateway to nearby Kumbungu, where vibrant markets and rich cultural traditions await adventurous travelers." OUTPUT: { "ID": "", "text": "Tamale Airport offers an exciting gateway to nearby Kumbungu, where vibrant markets and rich cultural traditions await adventurous travelers.", "annotations": [{ "ID": "", "name": "Tamale Airport", "city": "Tamale", "county": "", "state": "Northern Region", "country": "Ghana", "start_idx": 0, "end_idx": 14, "geometry": null }, { "ID": "", "name": "Kumbungu", "city": "Kumbungu", "county": "", "state": "Northern Region", "country": "Ghana", "start_idx": 52, "end_idx": 60, "geometry": null }] }

Example 2: TEXT: "Damn, I've been standing on Falschungsspitze all day, and this cursed Bankkogel just won't come into view!" OUTPUT: { "ID": "", "text": "Damn, I've been standing on Falschungsspitze all day, and this cursed Bankkogel just won't come into view!", "annotations": [{ "ID": "", "name": "Falschungsspitze", "city": "", "county": "", "state": "", "country": "Austria", "start_idx": 28, "end_idx": 44, "geometry": null }, { "ID": "", "name": "Bankkogel", "city": "", "county": "", "state": "", "country": "Austria", "start_idx": 70, "end_idx": 79, "geometry": null }] }

Example 3: TEXT: "I'm so frustrated! The Ikaalinen Pentecostal Church has not managed to secure new premises." OUTPUT: { "ID": "", "text": "I'm so frustrated! The Ikaalinen Pentecostal Church has not managed to secure new premises.", "annotations": [{ "ID": "", "name": "Ikaalinen Pentecostal Church", "city": "Ikaalinen", "county": "", "state": "Pirkanmaa", "country": "Finland", "start_idx": 23, "end_idx": 51, "geometry": null }] }

Figure 2: Zero-Shot (Bold), One-Shot (Bold + Italic) and Few-Shot system-prompt used to instruct the LLMs

fuzzy matching. These additional fields were included in the prompt to aid the model in disambiguating toponyms during recognition, but their correctness is not assessed and they do not affect corpus quality.

For the LLMs, we consider three configurations—zero-shot, one-shot, and few-shot—as illustrated in Figure 2, while keeping the temperature fixed at 0. We further test whether providing the prompt in the native language or in English affects model performance. Figure 2 shows the system prompt in English; translated versions of this prompt are used for the other languages. The text to be parsed is passed in the user prompt.

We use standard metrics for evaluating *Toponym Recognition*, namely precision, recall, and F1-score [1, 12]. To match predicted toponyms to gold annotations, we apply fuzzy name matching in addition to span matching. Equations 1–3 show the formulas of the metrics used with True Positives (TP), False Positives (FP) and False Negatives (FN).

$$\text{precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F}_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

3. Experiments and Evaluation

3.1. Corpus Generation

Figure 3 shows the success rate for the different languages. Overall, the success rate is highest in the first iteration and declines with each consecutive run. The results show that Qwen3-14B more easily generates text in the English language compared to the generation curve in German and Belarusian.

For Finnish, the results show that even the language-specific Poro2 model struggles to generate texts that pass our auto-QA test, failing to generate all 200 valid texts after 10 iterations.

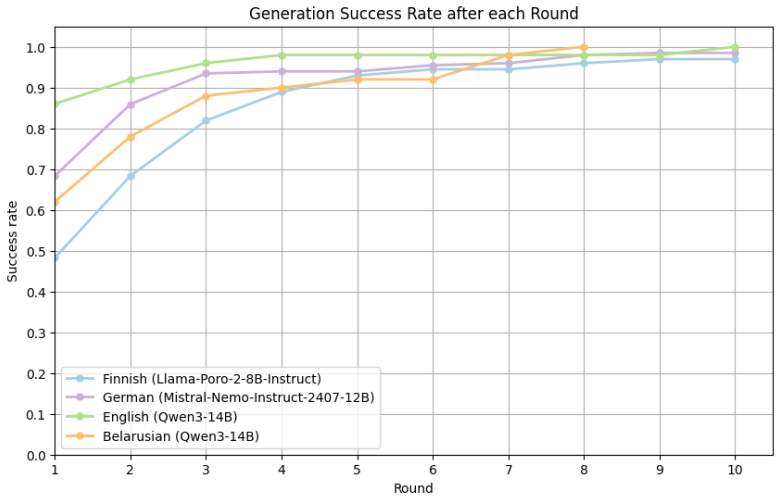

Figure 3: Success Rates of generated texts passing the automatic checks during the iterative corpus generation for the four languages.

3.2. Human Evaluation

Across four languages, the dominant error was missing or unannotated place mentions (21.6% of all generated samples). Less frequent issues included grammatical errors or nonsensical text segments (11.2%), toponym span mismatches (2.9%), and occasional misuse of spatial relations (2.1%).

While Mistral-Nemo-Instruct-2407-12B generally performed well on **German**, language-specific effects were nonetheless evident. For **Belarusian**, texts contained grammatical inconsistencies and interference from Russian or Ukrainian (e.g., 'водзе' instead of 'вадзе'). We treated such issues leniently, as they do not materially affect toponym recognition, and are consistent with the under-representation of Belarusian in large-scale training corpora and its close linguistic proximity to a vastly better represented Russian [18].

General-purpose multilingual models produced near-incomprehensible outputs for **Finnish**, so we resorted to using a Finnish-specific instruct model *Poro2* [19]. Even then, a large proportion of texts were rejected due to grammatical anomalies (48 of 182 samples) and unannotated toponyms (55), highlighting the difficulty of recognising inflected place names using simple string-based methods.

The **English**-language generation in Ghana was challenging due to a patchy, sparse coverage of OSM and Wikipedia, causing repeated references to a small set of well-documented places (e.g., 'University of Ghana' mentioned in three texts) and an over-representation of administrative districts; the latter are unlikely to be commonly referenced in everyday conversations. Similar biases of knowledge representation in data sources were observed in other countries, in particular, the abundance of churches and lakes in Finland (see Figure 6).

Table 2
Number of successfully generated samples and place mentions per corpus

Country	Texts	Total Places	Avg Places per Text	Unique Places
Austria	151	343	2.27	336
Belarus	40	106	2.65	104
Finland	83	222	2.67	218
Ghana	43	101	2.35	83

3.3. Corpus Characteristics

Table 2 shows the created corpus sizes across the four study areas after poor-quality samples were removed by human annotators. Austria (151 of 200) is the largest corpus, followed by Finland (83 of 200), Ghana (43 of 50) and Belarus (40 of 50). The average number of place mentions per text remains consistent across the study areas. However, there are slight differences with the number of unique places, with Ghana showing comparatively fewer. This reflects data sparsity of the gazetteers (OSM and Wikipedia) for this region.

Each generated corpus includes a broad range of tones and forms (Figure 4), as each was sampled with the same equal probability. There is variation across the study areas, with certain tones and forms passing the QA checks more easily depending on the geographic or linguistic context. For example, 'travel notes' is the most common form within the Belarus corpus, while being barely represented in other study areas. Likewise, 'sarcastic' texts are frequent for each study area, except Ghana where other tones are predominant.

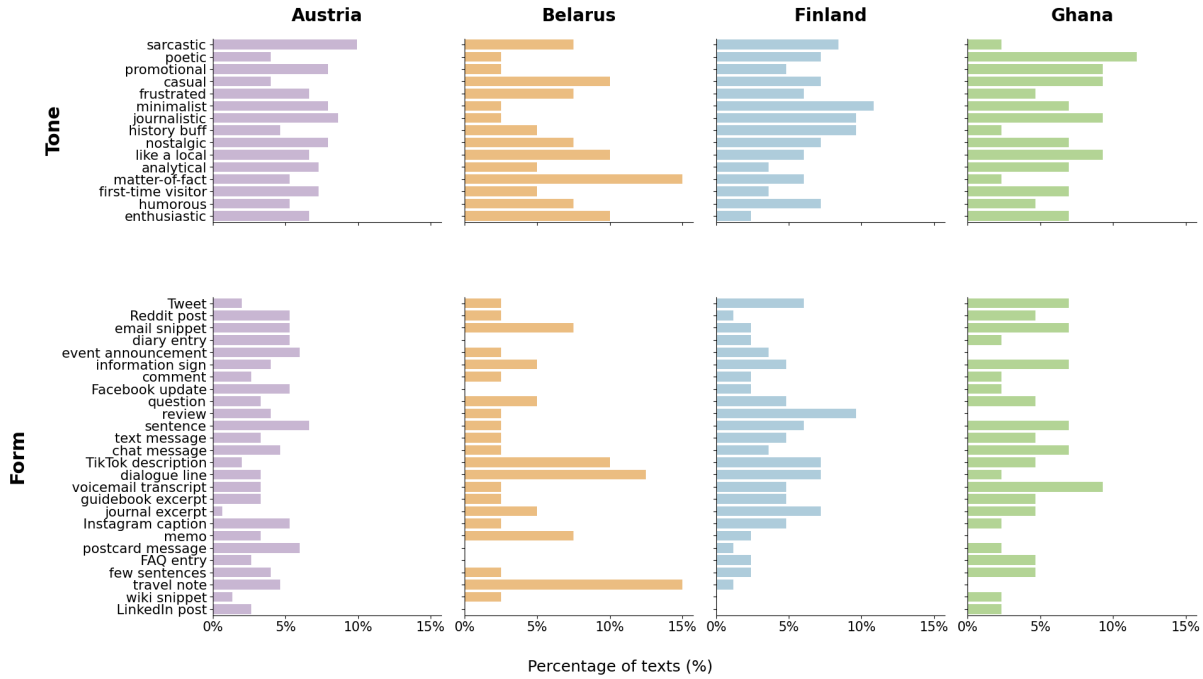

Figure 4: Tone and form distribution for the generated text that passed QA checks

Figure 5 shows the spatial distribution of place names at H3 hexagonal cell level. Place mentions appear evenly distributed across each country, and hexagons generally only contain a few place mentions. However, highly populated areas receive more place mentions, shown by darker hexagons. For example, the hexagon containing Vienna contains 42% of Austria's total place mentions. This reflects our three-step process for a representative sample, which was weighted towards highly-populated areas.

The types of places that are mentioned within each corpus is also explored, by using the proportion of OSM `key:value` pairs (Figure 6). Each corpus has a distinct array of place types, that most likely reflects their own human and physical geography, but also the biases and data sparsity within OSM and Wikipedia. For instance, Belarus corpus shows an uneven distribution of place types, with most being 'hamlets'.

3.4. Toponym Recognition Evaluation

Figure 7 summarizes precision–recall performance for toponym recognition across the generated corpora. Overall, the best-performing model depends on the target language. On the German corpus,

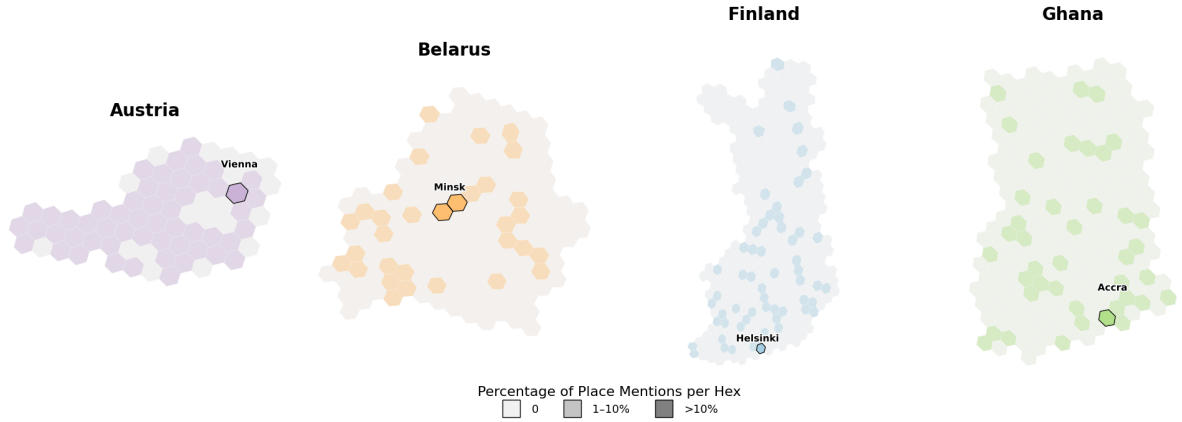

Figure 5: Distribution place mentions per H3 Hexagon for each study area. Urban centers are labeled.

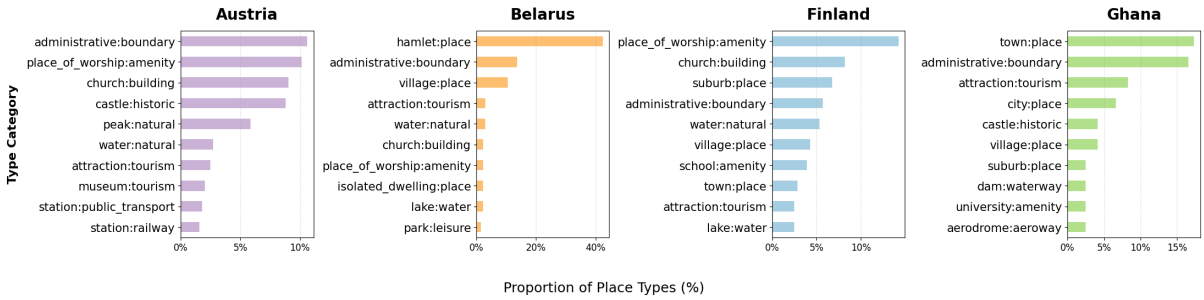

Figure 6: Top 10 OSM `key:values` pairs for each generated corpus. They are not mutually exclusive, therefore a place mention may have more than one place type

Llama 3.1 8B (few-shot) achieves the highest F1 (0.84). On the Belarusian corpus, Mistral3-14B (zero-shot) performs best with an F1 of 0.92. For English and Finnish, the top results are obtained by Qwen3-14B and Mistral3-14B, respectively. These results indicate that model choice can be decisive for successful toponym recognition, and that the best model is not consistent across languages.

Instruction language also affects performance, but the impact varies substantially by language and model. The effect is most pronounced for Belarusian, where models achieve roughly +0.09 F1 when prompts are provided in the target language compared to English prompts. Further, the gain differs by model. For instance, Llama 3.1 8B benefits strongly (approximately +0.17 F1), whereas Mistral3-14B shows only a modest improvement (about +0.025 F1). In contrast, for German the average difference between native-language and English prompts is slightly negative (≈ -0.006 F1), suggesting that English instructions can be equally effective or marginally better in this setting. Thus, whether prompts should be localized is language- and model-dependent.

Providing in-context examples generally improves results. For Finnish, German, and English, the best scores are obtained with one-shot or few-shot prompting. The main exception is Belarusian, where Mistral3-14B (zero-shot) performs marginally better than its prompted variants. This exception suggests that the benefit of examples is not uniform across languages and models.

Although Qwen3-14B was used to generate the English and Belarusian corpora, it does not consistently achieve the best recognition performance on those datasets. On Belarusian, Mistral3-14B exceeds the best Qwen variant by approximately +0.07 F1.

Considering all languages jointly, the best overall configuration is Mistral3-14B (few-shot, native prompt) with an F1 of 0.84, closely followed by its one-shot counterpart and Qwen3-14B (few-shot, native prompt) (F1 0.83 and 0.81, respectively). Across the board, LLM-based approaches substantially outperform the baseline taggers (spaCy multilingual: 0.48, multilingual RoBERTa: 0.43, native spaCy: 0.42).

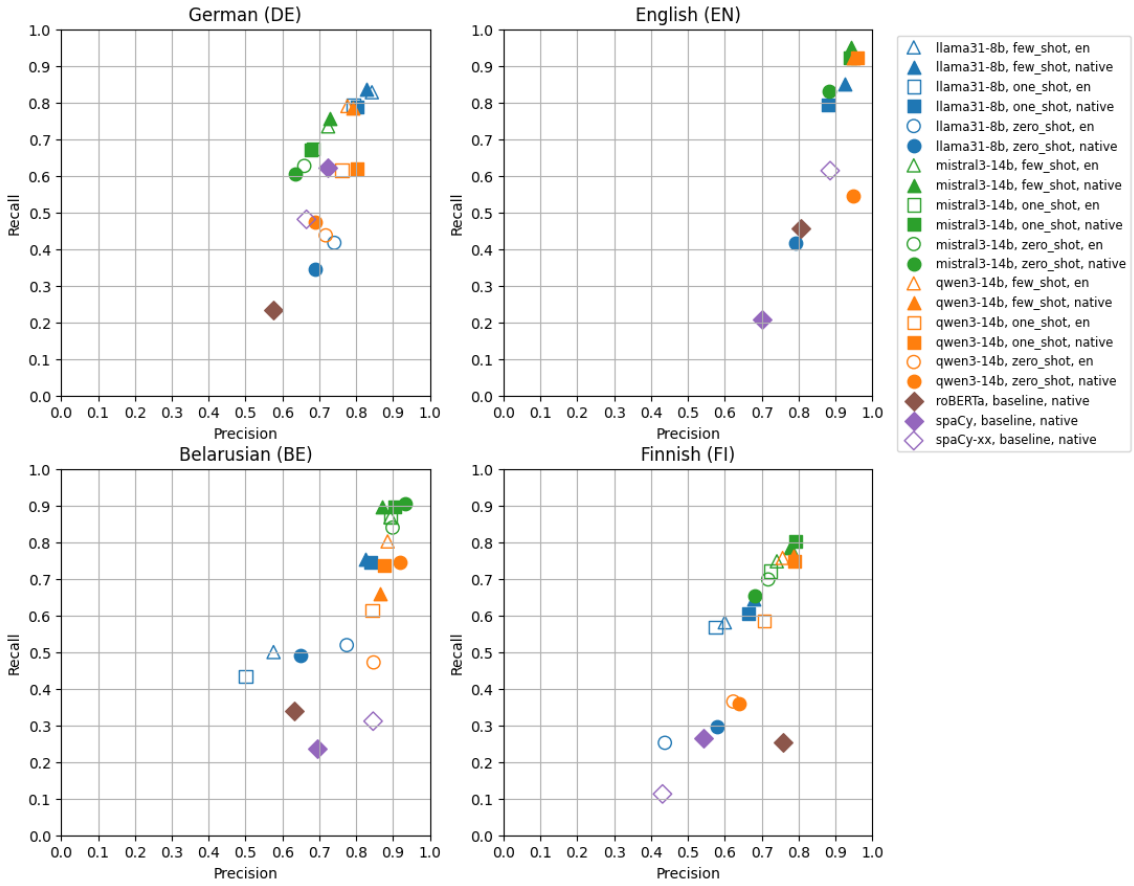

Figure 7: Precision and Recall model performance on the toponym recognition tasks on the four corpora visualized for the baseline models, as well as the zero-shot, one-shot, and few-shot approaches.

3.5. Summary of Research Questions

Addressing *RQ1*, we demonstrated a semi-automatic pipeline that combines OSM gazetteers, Wikipedia context, and LLM prompting to generate annotated synthetic geoparsing texts across four languages and regions. The pipeline includes population-weighted spatial sampling, neighbour retrieval, inflected surface-form extraction of place names, and iterative auto-QA, producing credible corpora without manual annotation.

Regarding *RQ2*, the key challenges were uneven OSM and Wikipedia coverage (particularly for Ghana and Belarus), morphological complexity in inflected languages (Finnish and Belarusian), and the limitations of general-purpose multilingual models for low-resource languages (Belarusian texts showed interference from better-represented related languages – Ukrainian and Russian in particular – while Finnish required a language-specific model altogether).

For *RQ3*, LLM-based toponym recognition substantially outperformed traditional baselines across all four corpora. Performance varied by language, model, prompting strategy, and instruction language, indicating that no single configuration generalizes across all settings.

4. Conclusion

In this paper, we demonstrate that LLM-generated synthetic corpora can support multilingual geoparsing, including low-resource languages and underrepresented regions, when generation is constrained by gazetteers and auxiliary geographic context. Our results show that state-of-the-art toponym recognition models achieve competitive performance on these corpora, while human evaluation reveals systematic failures related to grammar, post-generation toponym matching, and uneven geographic knowledge

coverage.

While language-specific models remain a necessity for under-represented languages, synthetic generation can be viewed as a scalable complement to curated corpora that enables controlled cross-lingual and cross-regional evaluation of geoparsers. Future work should focus on bias mitigation in source knowledge and sources beyond traditional OSM and Wikipedia gazetteers, and the integration of synthetic and human-annotated data for globally representative geoparsing benchmarks.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-5.2, Gemini 3, and Claude Sonnet 4.6 for (1) Paraphrasing and rewording, (2) Improving writing style, and (3) Grammar and spelling check. The authors take full responsibility for this publication's content.

Acknowledgements

I.I. was supported by Ordnance Survey & UKRI Engineering and Physical Sciences Research Council [grant no. EP/Y528651/1]. **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).

References

[1] M. Gritta, M. T. Pilehvar, N. Limsopatham, N. Collier, What's missing in geographical parsing?, Language Resources and Evaluation 52 (2018) 603–623. doi:10.1007/s10579-017-9385-8.

[2] C. B. Jones, R. S. Purves, Geographical information retrieval, International Journal of Geographical Information Science 22 (2008) 219–228. doi:10.1080/13658810701626343.

[3] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, F. Klan, Location Reference Recognition from Texts: A Survey and Comparison, ACM Comput. Surv. (2023). doi:10.1145/3625819, place: New York, NY, USA Publisher: Association for Computing Machinery.

[4] J. O. Wallgrün, M. Karimzadeh, A. M. MacEachren, S. Pezanowski, GeoCorpora: Building a corpus to test and train microblog geoparsers, International Journal of Geographical Information Science 32 (2018) 1–29. doi:10.1080/13658816.2017.1368523.

[5] J. L. Leidner, An evaluation dataset for the toponym resolution task, Computers, Environment and Urban Systems 30 (2006) 400–417. doi:10.1016/j.compenvurbsys.2005.07.003.

[6] G. DeLozier, B. Wing, J. Baldridge, S. Nesbit, Creating a Novel Geolocation Corpus from Historical Texts, in: Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016), Association for Computational Linguistics, 2016, pp. 188–198. doi:10.18653/v1/W16-1721.

[7] X. Hu, T. Elßner, S. Zheng, H. N. Serere, J. Kersten, F. Klan, Q. Qiu, DLRGeoTweet: A comprehensive social media geocoding corpus featuring fine-grained places, Information Processing & Management 61 (2024) 103742. doi:10.1016/j.ipm.2024.103742.

[8] T. Leppämäki, T. Toivonen, T. Hiippala, Geographical and linguistic perspectives on developing geoparsers with generic resources, International Journal of Geographical Information Science (2024) 1–22. doi:10.1080/13658816.2024.2369539.

[9] E. Bender, The #BenderRule: On Naming the Languages We Study and Why It Matters, The Gradient (2019). URL: https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

[10] E. Acheson, S. D. Sabbata, R. S. Purves, A quantitative analysis of global gazetteers: Patterns of coverage for common feature types, Computers, Environment and Urban Systems 64 (2017) 309–320. doi:10.1016/j.compenvurbsys.2017.03.007.

[11] Z. Liu, K. Janowicz, L. Cai, R. Zhu, G. Mai, M. Shi, Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing, AGILE: GIScience Series 3 (2022) 1–13. doi:10.5194/agile-giss-3-9-2022.

[12] Y. Hu, G. Mai, C. Cundy, K. Choi, N. Lao, W. Liu, G. Lakhanpal, R. Z. Zhou, K. Joseph, Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages, International Journal of Geographical Information Science 37 (2023) 2289–2318. doi:10.1080/13658816.2023.2266495.

[13] R. Mioduski, Benchmarking large language models for geolocating colonial Virginia land grants, Journal of Spatial Information Science (2025) 57–96. doi:10.5311/JOSIS.2025.31.502.

[14] A. Belbekri, W. Bouarroudj, F. Benchikha, Z. Boufaida, Generating Synthetic Training Data for Named Entity Recognition With Large-Scale Models Integrating Wikidata and GPT, RIF'24: The 13th Conference on Research in computIng at Feminine, May 20-21, 2024, Constantine, Algeria (2024).

[15] Z. Li, H. Zhu, Z. Lu, M. Yin, Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations, in: EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings, 2023, pp. 10443–10461. doi:10.18653/v1/2023.emnlp-main.647.

[16] G. Kamath, S. Vajjala, Does Synthetic Data Help Named Entity Recognition for Low-Resource Languages?, 2025. doi:10.48550/ARXIV.2505.16814.

[17] F. Welscher, I. Ilyankou, P. Smith, T. Leppämäki, Synthetic Geoparsing Corpora: Leveraging LLMs, Gazetteers and Context for spatially balanced corpus generation, Under Review (2026).

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

[19] E. Zosa, J. Louma, K. Hakala, A. Virtanen, M. Koistinen, R. Luukkonen, A. Reunamo, S. Pyysalo, J. Burdge, Poro 2: Continued Pretraining for Language Acquisition, LumiOpen, 2025.