Right and wrong: harnessing visual geodata mining methods to find patterns in LLM generated geodata

Pablo S. Löw, Jukka M. Krisp

Angaben zur Veröffentlichung / Publication details:

Löw, Pablo S., and Jukka M. Krisp. 2025. "Right and wrong: harnessing visual geodata mining methods to find patterns in LLM generated geodata." In *33rd Annual GIS Research UK Conference (GISRUK), University of Bristol, Bristol, UK, 23–25 April 2025*, Online-Ressource. Genf: CERN. https://doi.org/10.5281/zenodo.15230402.

Right and Wrong: Harnessing visual geodata mining methods to find patterns in LLM generated geodata

Pablo S. Löw [*1] and Jukka M. Krisp [†1]

[1]Applied Geoinformatics, University of Augsburg

GISRUK 2025

Summary

The growing interest in leveraging Large Language Models (LLMs) as sources of spatial data necessitates the development of scalable, comprehensive, and standardized methods for assessing their quality. This study focuses on evaluating the relevance of spatial data generated by an LLM. The analysis combines visual inspection with a distance-based metric to assess the spatial relevance and accuracy of the generated data, providing a scalable approach for systematic quality evaluation.

KEYWORDS: Large-Language Models, Automating, Geodata, Precision, OpenStreetMap.

1 Introduction

Large Language Models (LLMs) are increasingly becoming significant sources of information across a variety of domains, including geographic data. As highlighted in several papers, LLMs such as ChatGPT are now being examined for their potential to generate geoinformation (Blackwell et al., 2024; Cohn and Blackwell, 2024a,b; Karimi and Janowicz, 2024; Keler and Krisp, 2023). In the absence of correct data, or the inability to access such, it might be better to have imprecise than no data. However, the LLMs inherent imprecision, coupled with the rapid pace of their development, poses challenges for evaluating the reliability and accuracy of the geographic data they provide.

To approach these challenges our study introduces an implemented and automated algorithm that does two different things. Firstly, it creates a reference data set from more reliable sources such as OpenStreetMap (OSM) (Mooney et al., 2010). Secondly, it provides insights into quality trends via a comprehensible distance-based metric between the LLM- and reference data as an indicator of data precision. That is paired with an automated creation of maps for visual analysis. The methodology has to be fully automated to enable the exploration of many different geodatasets,

[*]pablo.loew@uni-a.de
[†]jukka.krisp@uni-a.de

because it is to be expected that the precision of geodata created by LLMs varies between different LLMs, over time, between topics (tourism, sport, leisure time etc.) and for different places.

Motivated by the finding of Karimi and Janowicz (2024) that tourist activity may be a significant determinant of LLM created data accuracy we apply the methodology to data on tourist attractions collected throughout 2024. It was created by ChatGPT and the analysis focuses on 10 European cities selected for their differences in touristic relevance, size (inhabitants, area) and placement in Europe. Touristic relevance in this study is understood to be the number of touristic visits as provided by mainly Euromonitor[1]. Further, we assume that the size of a city has at least a minor correlation to the touristic attraction potential and treat it thus as a proxy.

The first goal is to show that the algorithm transforms unreliable LLM data into a precise data set, which can be used for touristic travels. The second goal is too provide information on how reliable the reference dataset can be created and how precise the LLM data is at different stages of the LLM's evolution over a period of time.

This study provides exemplary results of the methodology as an initial investigation into the trends and quality of geodata provided by ChatGPT. The results offer insights into the evolving capabilities of LLMs and underscore the importance of regular evaluation to understand their implications for geospatial applications. The exploration of the algorithms potential with a tourist attraction dataset is limited by the circumstance that some often smaller cities have less points of touristic interest provoking LLMs to hallucinate more. It is further limited by the size of the dataset that covers only 10 cities.

2 Data and Methodology

To examine the data created by an LLM reference data is used. With the Nominatim API[2] OpenStreetMap offers free access to geocoding. The names of the sights provided by the LLM are used retrieve a location from the OSM database. However, some of the names cannot be found in the database because LLMs do not necessarily put the names exactly as they are. In some cases the names in the database are stored in the language of the corresponding country, whereas LLMs provide the names in the language specified or used for the prompt. An example of that is 'Augsburger Puppenkiste'. ChatGPT proposed 'Augsburgs Puppet Theatre' which is an understandable translation that is close to the real name. Yet it is in a different language than that of the database and is thus not found. If none of the functions retrieving reference coordinates provide a point the attraction is automatically ignored in the rest of the research.

Another reason for not being able to retrieve reference coordinates from the database is the hallucination phenomenon of LLMs. Meaning that the LLM falsely claims that a certain sight is in the region it is asked to provide information on. This can happen either when a sight is in another often adjacent region to the given spatial reference (A) or if it comes up with something that does not exist at all (B). An example of the more common case (A) is the 'Aire Sculpture Trail', located

[1] `https://www.euromonitor.com`, last access: 2024-02-29

[2] `https://nominatim.org/`

in Shipley, a town in Bradford, which directly neighbors Leeds. The case (B) is hypothetical and no example is found.

With the created reference data, the distance from the LLM- to the reference point is calculated. Further, the city name itself is geocoded and used as a city center. Then the distances from all points (LLM and reference) to the city center are calculated. Because the number of sights is not big enough for significant statistical analysis, the data points per city are plotted for visual examination. It is not necessarily a solution to create bigger datasets per city because if a city has not enough sights the probability for hallucinations rises. Additionally, the distance means are given in a table to examine trends.

3 Results

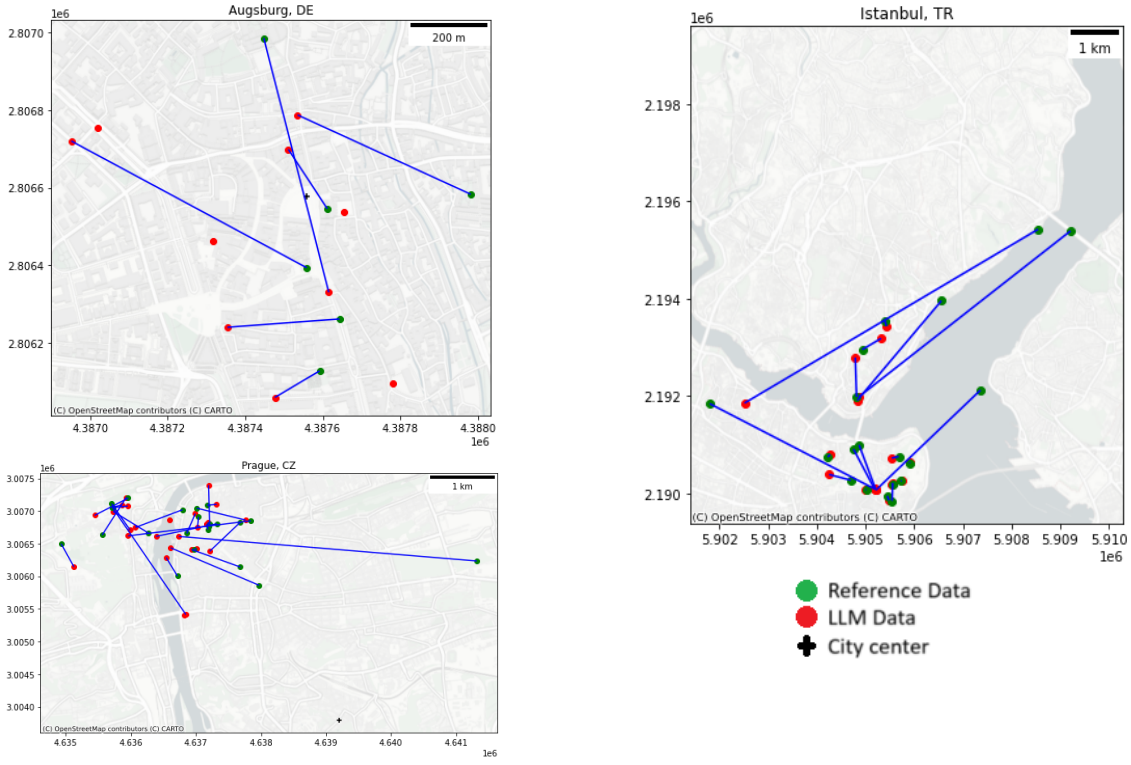

Figure 1: City maps with line connections between reference and LLM data. Three cities that do not have the LLM points closer to the city center

The maps of Figure 2 and 1 show the spatial distribution of the points, city center and line connections between reference- and LLM point. In all seven maps of Figure 2 the reference points are further away from the geocoded or a virtual city center and the connection lines lead radially outwards to a more or less strong degree. For example, in Amsterdam and Naples this phenomenon is weaker, while for Porto and Leeds it is very strong. Comparing the maps of Figure 1 to that no

clear pattern can be found. In Istanbul the southern points do display the phenomenon of cities in Figure 2 with a clear center, but the remaining northern points counteract that. In Prague neither reference- nor LLM data seems to be particularly in their distribution. For Augsburg only few reference points are retrieved and no clear pattern is observable.

Table 1: Cities with the resulting mean distances ordered according to their population. The first group has its LLM points closer to the city center. That is vice versa for the second group; LLM: LLM generated points; ref: Reference points; cc: City center

City, Country	Population	Mean distances [m]			Found
	[MM]	ref to llm	ref to cc	llm to cc	ref per LLM
Istanbul, TR	15.84	1395	8689	9537	19/20
Prague, CZ	1.34	785	3888	3988	24/26
Augsburg, DE	0.30	411	312	356	6/10
St.Petersburg, RU	5.38	911	9660	9507	17/20
Paris, FR	2.14	1740	2809	2763	28/30
Vienna, AT	2.01	1150	1331	628	26/26
Amsterdam, NL	0.93	394	913	827	16/20
Naples, IT	0.92	1472	3740	3466	17/20
Leeds, UK	0.82	3286	3379	674	6/10
Porto, PT	0.23	1648	2034	1036	19/20

1: Euromonitor (2018), 2: Augsburg Tourismus (2019), 3: ISTAT, 4: Wikipedia (2024)

Table 1 shows the population, mean distances and ratio of found reference to total LLM points. The table confirms the trends of the maps. The first three cities are those of Figure 1 and can accordingly be described with: $D_{mean}(ref\_to\_cc) < D_{avg}(llm\_to\_cc)$. For the rest of the cities that is vice versa. Per group, the cities are sorted descending according to their population[3]. That is because Karimi and Janowicz (2024) stated that the spatial reasoning capabilities are higher when the city is more relevant. Thereby, the population of the administrative area is used.

The maps of Prague and St. Petersburg show that their center points are far from their respective virtual center of the combined datasets (reference and LLM). Thus, their distances to the city center in Table 1 are unproportionally high. The column with the distance from the reference-, to the LLM point does not reveal a strong correlation to the population. Neither does the number of reference points found. This also holds true, when instead compared to other proxys for the relevance of a city like area, population density or touristic visitors. Figure 3 shows comparisons with all variables (Distances and characteristics of cities). Due to the statistically unrepresentative sample size no correlations can be proofed with the data. Some combinations of variables do not spread out freely. Instead they are clumpling with outliers.

From the Scatter- and Kernel Density Estimate Plots two clear correlations can be named. Firstly, the natural connection of a cities population to its area. Secondly, that the average distance of the

[3]Population data as found in `https://de.wikipedia.org/wiki/Liste_der_grten_Stdte_Europas`; last access April 2, 2025

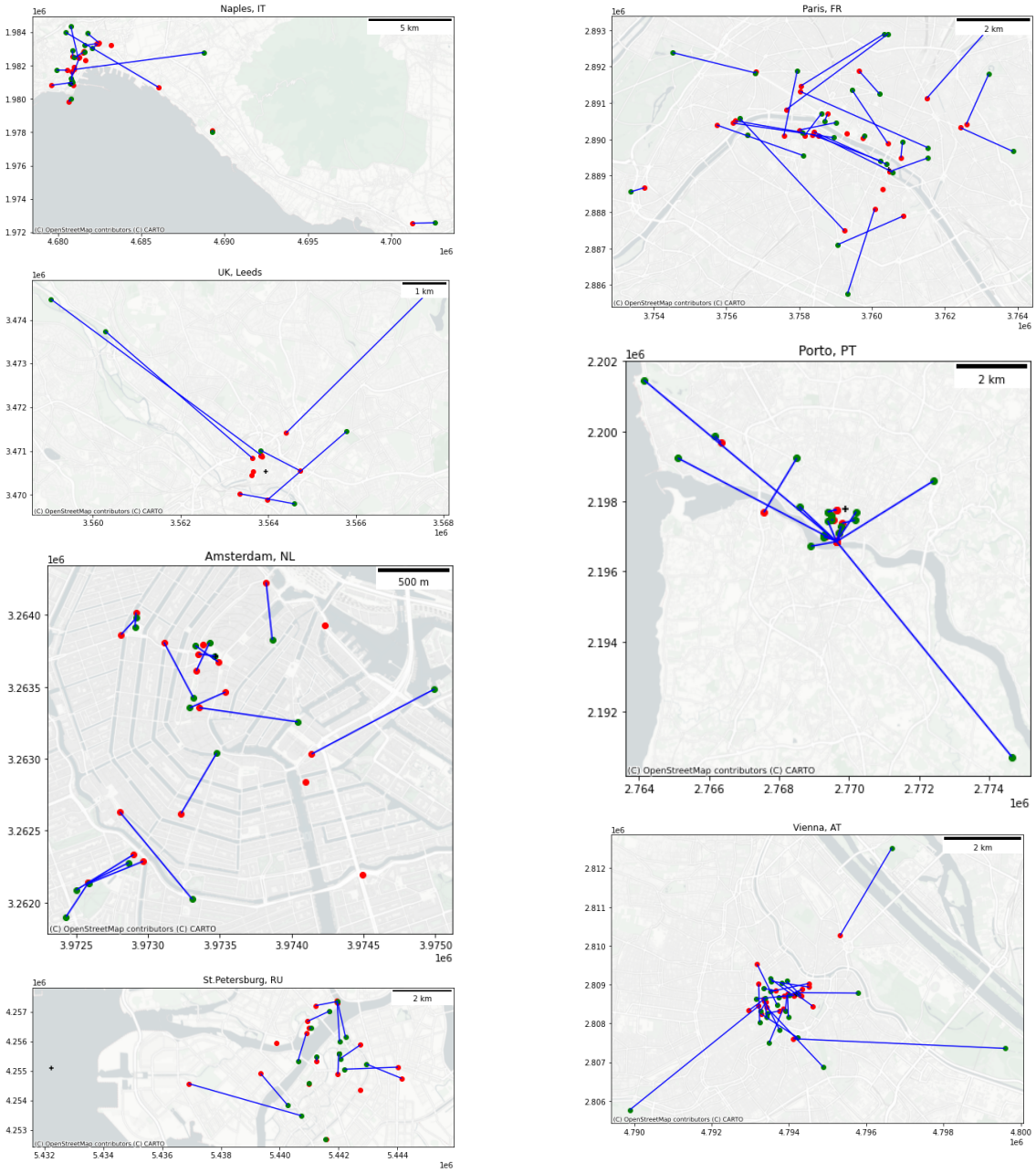

Figure 2: City maps with line connections between reference and LLM data. Cities that have the LLM points closer to the city center

reference data set to the city center correlates with the distance of the LLM data. The mean distance error of LLM to reference data between all cities is approximately $D_{avg} = 1.3km$. Together with the previous finding this shows that the further the sights of a city are away from their center the likelier the LLM is to also place them radially outwards, despite a varying distance to its reference point.

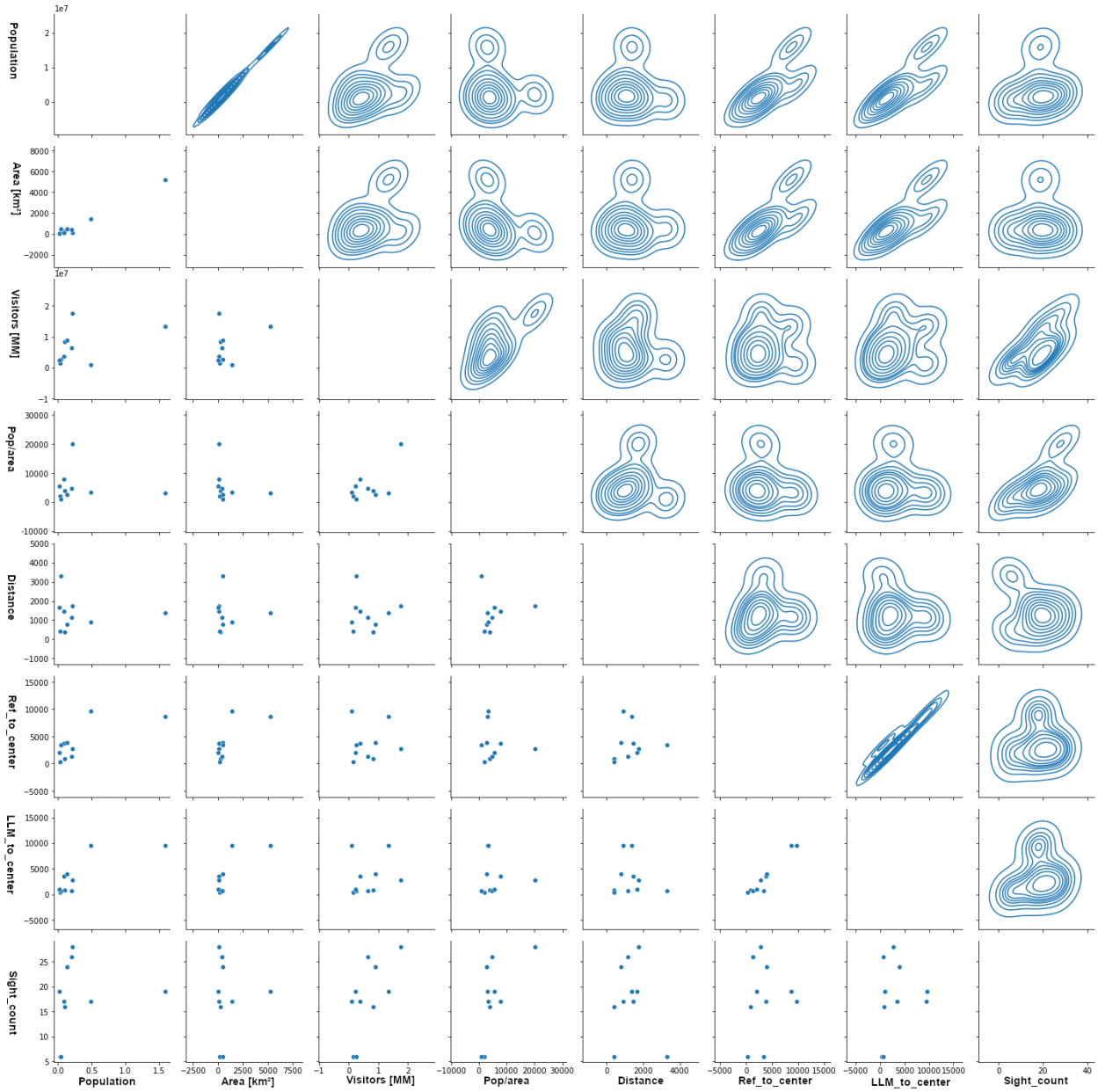

Figure 3: Pairplot of city characteristics (Population, Visitors, Area, Population/Area and distances)

A variable of the plot that hints at some weak correlation is the number of sights found. Its connection with most of the characteristics of a city is implied by the fact that the points do not spread out to widely. That points in the direction that a cities relevance has impact of the accuracy of the spatial data created by an LLM, but cannot be proofed due to the small sample size.

4 Discussion and conclusion

Working with LLM generated data of any kind is interesting and its quality is likely to improve in the future. Yet in the last year this improvement cannot be documented with the data and methodology of this paper. The dataset of Leeds (25.03.2024) is the first and that of Porto (21.1.2025) is the last that were created. The cities sorted according to their creation date are as follows:
$Leeds > Amsterdam = Napoli = Paris = Prague = Vienna > Augsburg > Istanbul = St.Petersburg = Porto$ Porto and Leeds both have small populations and comparable large distance errors (>1000m). Yet the later was created with a newer version of ChatGPT. Compared with Augsburg being in the middle of the year and also with a small population its distance error is smaller (<500m). That there are no clear trends could be because the sample size is too small or the number of changed variables (creation date, LLM version, country, city size etc.) is too large.

Neither can we find proof for the hypothesis that the relevance of a city impacts the precision with which an LLM can provide spatial data. What can be seen however, is a hint at a weak connection between city size and correctly given sight names. Specifically for cities with a population lower than 1MM LLMs have a higher chance of giving names of tourist attractions to which the algorithm finds no reference data. Since the sample size is small this can be observed, but not confirmed.

The trend that most of the cities (7/10) have their LLM created data points spatially closer too a center could be due to the monocentric character of a city. The name of the city is an important information of the prompt and as such likely identified and weighted highly by the LLM. The identification of crucial parts of a given prompt by LLMs strongly impacts the quality of their response and is well documented Vaswani et al. (2017). Because the spatial reference and important factor for the data creation is the city name and cities are usually represented as their center point, the LLM guesses coordinates close to that point.

The presented fully automated methodology to visually and statistically examine LLM generated geodata is helpful for future research on the topic. It also allows for the extension of the sample size. With this methodology general trends can be examined and development over time can be monitored. The documentation of the version of LLMs used to create the data is underrepresented in this study and might provide insights. The code, prompts and data can be found on gitlab[4].

5 Biography

Pablo S. Löw is a PhD student at the department of applied geoinformatics of the university of Augsburg. He pursues his interests in routing options based on different measures (e.g. accelerometer data gathered via VGI) to enhance bikeability and the possibilities LLMs offer for the creation

[4] `https://git.rz.uni-augsburg.de/geoinf-gig/llm_data_examination`

of geodata. Jukka M. Krisp is the professor of the department of applied geoinformatics of the university of Augsburg. He is focused on Location Based Services (LBS) Geographic Visualization / Visual Analytics, Spatial Modeling, Geographic Information Systems applications and GIS in ecological network planning.

References

Blackwell, R. E., Barry, J., & Cohn, A. G. (2024). Towards Reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores. *Computer Science - Computation and Language*, https://doi.org/10.48550/arXiv.2410.03492. arXiv:2410.03492 [cs] http://arxiv.org/abs/2410.03492.

Cohn, A. G. & Blackwell, R. E. (2024a). Can Large Language Models Reason about the Region Connection Calculus? *Computer Science - Computation and Language*, https://doi.org/10.48550/arXiv.2411.19589. arXiv:2411.19589 [cs] http://arxiv.org/abs/2411.19589.

Cohn, A. G. & Blackwell, R. E. (2024b). Evaluating the Ability of Large Language Models to Reason about Cardinal Directions. *LIPIcs, Volume 315, COSIT 2024*, *315*, 28:1–28:9, https://doi.org/10.4230/LIPIcs.COSIT.2024.28. arXiv:2406.16528 [cs] http://arxiv.org/abs/2406.16528.

Karimi, M. & Janowicz, K. (2024). Exploring challenges of Large Language Models in estimating the distance. *Abstracts of the ICA*, *7*, 1–2, https://doi.org/10.5194/ica-abs-7-68-2024 https://ica-abs.copernicus.org/articles/7/68/2024/.

Keler, A. & Krisp, J. M. (2023). Geodata Generation and Enrichment via ChatGPT for Location Based Services (LBS) https://repositum.tuwien.at/handle/20.500.12708/194769.

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for openstreetmap. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 514–517).

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.