AGILE: GIScience Series, 6, 11, 2025. https://doi.org/10.5194/agile-giss-6-11-2025
Proceedings of the 28th AGILE Conference on Geographic Information Science, 10–13 June 2025.
Eds.: Auriol Degbelo, Serena Coetzee, Carsten Keßler, Monika Sester, Sabine Timpf, Lars Bernard
This contribution underwent peer review based on a full paper submission.

How close is "close"? An analysis of the spatial characteristics of perceived proximity using Large Language Models

Joseph Shingleton [1] and Ana Basiri [1,2]

[1]School of Geographical and Earth Sciences, University of Glasgow, Glasgow, United Kingdom
[2]The Alan Turing Institute, London, United Kingdom

Correspondence: Joseph Shingleton (joseph.shingleton@glasgow.ac.uk)

Abstract. Proximity plays an important role in Geographic Information Sciences. It underpins our understanding of spatial dependence and spatial structure, and is a key component of many commonly used analytical techniques. Despite this, it remains a difficult concept to rigorously define. Describing one geospatial object as "near" another implies much more than a simple geometric relationship - with factors such as accessibility, utility and function also playing an important role. Previous work has shed light on these relationships through the application of sophisticated mathematical models which attempt to encapsulate both spatial and non-spatial aspects of proximity. In this paper, we present a novel method that uses Large Language Models (LLMs) to extract perceived proximity relationships from natural language. Using 20000 AirBnB listings in London, we identify locations which are described as "near" to each property and analyse their spatial distribution. Our results reveal complex patterns linking perceived proximity to accessibility, utilisation, and administrative prominence. We show that locations with a broader area of influence often correspond to higher transit connectivity or higher place-level categories. While the Airbnb dataset reflects a specific, tourism-focused demographic, the approach is generalisable to other sources of user-generated text. This work demonstrates how LLMs can support data-driven spatial analysis by surfacing nuanced, context-sensitive geospatial relationships embedded in everyday language.

Submission Type. Analysis

BoK Concepts. Relationships, Cognitive and Social Foundations, Articificial Intelligence in EO and GI

Keywords. Large Language Models, Spatial Proximity, Natural Language Processing

1 Introduction

The concept of proximity plays an important role in the communication of geospatial information in natural language (Hall and Jones, 2022). Describing a geospatial object as "near" other objects often carries more semantic meaning than directly offering the object's geographic coordinates. Despite this, the concept of "nearness" is not well defined. The distribution of an object's "nearby" locations might depend on position, function, audience or time, as well as multiple other factors (Brennan and Martin, 2012). For a human audience, picking apart these factors based on contextual information is a relatively trivial task. For algorithms, however, the loosely defined concept of "nearness" makes it a more challenging assignment.

The lack of a formal definition of proximity arises from the complex nature of distance (Brennan and Martin, 2012). Viewing distance in a purely euclidean sense, it would be straightforward to construct a notion of proximity which characterises concepts such as "near to" or "far from" via some threshold. However, when we view distance in functional terms, this method quickly becomes intractable (Guesgen and Albrecht, 2000). Complicating factors such as accessibility, topography, or utility might influence our notion of proximity. Further, the characterisation of locations as "nearby" a geospatial object might depend on the assumed degree of spatial knowledge our audience has about those locations, or the function of the information being conveyed. For example, one may choose to describe an object as near to a transit station, or to a well-known point of interest - even if there are closer locations which are less well known. This suggests a "platial" aspect to proximity that surpasses simple geometric relationships (Mocnik, 2022).

The importance of a coherent definition of proximity extends beyond semantic analysis. It is well known that geospatial phenomena have a tendency for both spatial autocorrelation and spatial heterogeneity (Miller, 2004). Analytic techniques such as geographically weighted regression (GWR) attempt to address these factors, relying on some prior notion of proximity to achieve this (Jiale Ding and Du, 2024). Many applications of GWR assume a euclidean definition of proximity (Comber et al., 2020), however, this is often insufficient for capturing many geospatial effects. Binbin Lu and Fotheringham (2014) found that including hedonic independent variables indicative of accessibility, such as road networks and travel times, improved the performance of GWR on predicting house prices in London, compared to a purely euclidean model. Similar work by Kai Cao and Wu (2019) found that a travel-time based GWR model outperformed both ordinary least squares regression and euclidean GWR on predicting house prices in Singapore. Jingyi He and Yu (2023) also found that incorporating a population mobility matrix into GWR greatly improved the model's ability to capture social and economic systems. These attempts, however, tend to account for just one or two non-eucledian proximity factors. In reality, proximity is a complex relationship with many diverse contributing factors. This is addressed by Jiale Ding and Du (2024), who present a neural network approach for constructing highly non-linear spatial weighting matrices for GWR, citing significantly improved performance on house price estimation in Wuhan, China.

Many methods for assessing geospatial proximity have taken a geometric approach. Earlier efforts have used Voronoi diagrams to discretise space into regions of proximity relative to a collection of reference objects (Gahegan and Lee, 2000). While computationally simple, such approaches do not model any effects beyond Euclidean distance. Later work by She et al. (2015) introduces network-weighted Voronoi diagrams to model the spatial distribution of events in a way that accounts for relative accessibility between locations. Even while accounting for non-euclidean effects, these methods lack sufficient flexibility to handle the nuanced and variable way in which humans perceive and discuss proximity in natural language.

Grütter et al. (2010) propose a qualitative approach to spatial proximity which accounts for both human cognition and linguistic interpretation. Their model uses administrative hierarchies to assess nearness on spatio-thematic grounds - treating proximity as a subjective concept. Brennan and Martin (2012) suggested an analytical method that combines both purpose and domain to derive a context-aware *Impact Area*—a teleologically grounded region of influence for a given location. For instance, proximity to a highway might differ depending on whether the context is noise pollution or accessibility. Grütter (2019) extends this idea into a coherent framework of proximity relationships, which supports the identification of hierarchical neighbourhoods and the assessment of spatial proximity. These approaches lean heavily on the concept of perceived proximity, proposing a subjective and relativistic account weighted by human cognition, platial attributes, and the specific function under which a location is being considered.

Even the more flexible approaches discussed above rely on a set of underlying assumptions about what constitutes proximity. These assumptions attempt to encapsulate the key factors contributing to platial proximity, and generalise these factors to all locations. This generalisation of assumed contributing factors means that highly localised and specific factors might not be captured. This can, at least to some extent, be addressed with the adoption of data-driven approaches (Jiale Ding and Du, 2024). By analysing the way proximity is perceived in user-generated data sources, we can develop a notion of locality and proximity which is based on observed, rather than assumed, factors.

Implementation of this data-driven approach, however, is challenging. In order to identify described proximal locations in text, we require sophisticated language models. Transformer-based natural language models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have proven to be effective in identifying spatial relationships described in text (Stock et al., 2022; Shingleton and Basiri, 2024). Such models, however require large amounts of structured data to effectively train. Automated geographically informed tagging methods, such as that employed by Shingleton and Basiri (2024), can be effective for well-defined relationships, such as containment or adjacency, but are challenging to apply to the more nebulous concept of proximity.

Large Language Models (LLMs) are very large, transformer-based models which have been pre-trained on vast quantities of unstructured data (Bharathi Mohan et al., 2024). As a result, they are able to generalise extremely well to linguistic comprehension and reasoning tasks without any further training (Bianchini et al., 2024). The ability for LLMs to parse complex natural language in a "zero-shot" context makes them a good candidate for identifying proximal relationship in text. In this paper, we leverage the linguistic capability of the Llama-3 LLM (Grattafiori et al., 2024) to parse descriptions of 20000 AirBnB properties in London. The model is tasked with identifying any locations which are described as proximal to a property in that property's description. The spatial distribution of these proximal locations is then assessed, from which we identify nuanced relationships between proximity, function and accessibility.

This novel application of LLMs to the problem of geospatial proximity identification addresses some shortcomings of previous approaches. For instance, Derungs and and (2016) suggest a bag-of-words approach, in which a large corpus of textual data is mined for words and phrases indicative of spatial proximity. By doing this, however, they risk the introduction of both false positives - the use of words like "near" to express something other than proximity - and false negatives - the use of phrases indicative of proximity which had not been considered in the methods. Similar methods, such as that of Stock et al. (2022), also have the same limitations. By leveraging the linguistic flexibility of LLMs, our approach addresses both sources of error, without the need for large amounts of accurately labelled training data.

The application of LLMs to such tasks is not without risk (Li et al., 2024). Generative models are prone to hallucination - either through the generation of unrelated or irrelevant information, or through the occurrence of logical fallacies and inconsistencies (Jiang et al., 2024). As such, systematic verification of the generated data is crucial to the validity of this work, with minimisation of the model's false positive rate being a central goal. The introduction of excessive hallucinations into our processed data poses a serious risk to the validity of our results.

In addition to the risk of hallucination, LLMs also demand significant computational cost (Klang et al., 2024). The use of of larger models (measured in terms of number of parameters) can reduce the risk of hallucination (Chrysostomou et al., 2024) and improve performance on linguistic tasks (Bianchini et al., 2024), however they require significantly increased computational resources to run. For large datasets, such as that used in this paper, employing the use of very large models is simply not feasible. As such, we use a knowledge distillation approach (Cantini et al., 2024), in which a subset of our data is processed by a larger model, the output of which is used to fine-tune the weights of a more computationally efficient model. This results in a smaller, more task optimised model which achieves comparable performance to the larger model, at a fraction of the computational cost.

In this paper, we demonstrate a method for using LLMs to identify pereceived proximty relationships in user-generated data. We have focussed on AirBnB property descriptions for this work, although the approach can be applied to other available datasets. The Airbnb dataset is a suitable choice for this analysis because it both provides a large volume of natural language descriptions anchored to specific point locations (properties), and its touristic context naturally encourages references to nearby landmarks and amenities. Through this approach, we have generated a large dataset of property descriptions with tagged proximal locations - a dataset which may be used for many further applications. Through spatial analysis of the generated data, we are able to develop new insights into the different factors influencing perceived proximity.

By employing this approach, we are able to test the hypothesis that perceived proximity, as described in natural language, reflects both objective spatial features (e.g., accessibility, transit connectivity, administrative hierarchy) and subjective, platial characteristics (e.g. perceived utility, or cultural familiarity). Specifcally, we investigate whether properties described as "near" to a given location are more widely distributed when that location is well connected, highly utilised or occupies a prominent position in the urban, administrative and cultural hierarchy.

2 Methods

The underlying data used to investigate reported spatial proximity is comprised of a list of AirBnB properties in London, UK. We use data posted on the "Inside AirBnB" website (Murray Cox, 2025). The full dataset contains a geolocation for each property, along with a description of the property given by the owner, a set of reviews for the property, and other information including pricing, availability and occupancy rate. The validity of this dataset has come under some scrutiny, with particular concern around the reviews associated with each listing (Alsudais, 2021). As such, we have not included reviews in our analysis, instead using only the property descriptions and geolocations. Property locations are anonymised, so that the true location of the property is somewhere within 150 meters of the listed location. Given that over 96% of reference locations have a standard distance greater than 1 km to associated Airbnb properties, and none have a standard distance below 300m, the 150m anonymisation offset is unlikely to meaningfully affect the spatial analysis results, beyond the introduction of a small degree of noise. All properties in the full dataset were scraped from AirBnB between 06/09/2024 and 11/09/2024.

For our analysis, we take a sample of 20000 listings from the full dataset, with a further 2500 listings sampled for model fine-tuning, and 100 for model testing. Figure 1 (a) shows the distribution of the 20000 properties in our analysis set. The properties have been aggregated into a hexagonal grid using the H3 spatial grid system (Sahr, 2011), with resolution set to seven, so that each hexagon has an approximate area of 5.16 km^2.

We note that the test set used for this analysis is very small compared to analysis dataset. High-quality human annotation of the dataset is highly labour intensive, making production of a large validation set unfeasible within the timeframe of this work. The validation set was designed to cover a representative cross section of listing styles, however its limited size means performance metrics should be interpreted cautiously. As such, we regard the results on model performance as exploratory, rather than definitive. Future work should aim to develop a larger, stratified evaluation set to support more robust model validation.

2.1 Identification of proximal locations

We use the Llama-3 family of LLMs to identify locations described as "near to" each property in the provided listings, noting that the prompt design, along with the flexibility of LLMs, means that a broad range linguistic indicators are used to identify proximity. The model is asked to identify any explicitly named locations (such as "Hyde Park" or "Richmond") and ignore any vague or ambiguous locations (such as "restaurants" or "station"). The model is also asked to avoid any locations which the property is described as being "within", for example if the description reads "A spacious property in Brixton" the model should not identify "Brixton"

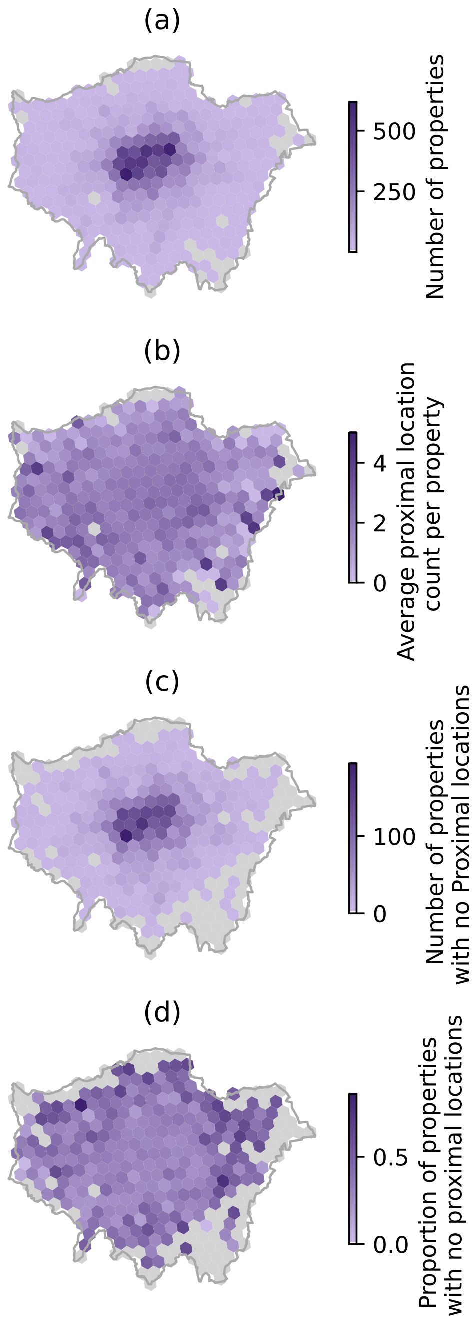

Figure 1. (a) The distribution of properties in our analysis dataset of 20000 listings, aggregated into hexagonal bins of approximately equal area, (b) the average number of identified proximal locations mentioned for properties in each hexagon, (c) the total number of properties with no proximal locations identified and (d) the proportion of properties with no proximal location mentioned.

as a nearby location. The full prompt is given in appendix figure A1.

We ask the model to structure its response as a JSON file, and provide a template to aid in this. In some instances, the model output cannot be successfully read as a JSON file. Where this happens, in the first instance, we use the JSON-Repair package for Python (Baccianella, 2024) to perform basic fixes, such as replacing single with double quotes, and adding any missing brackets. Any outputs which remain unable to be read are returned to the model with a request to fix the misconstructed JSON file. Further failure to successfully construct a valid JSON formatted output results in the listing not being assigned any proximal locations. We report the rate of misconstruction in the results section.

Figure 1 (b) shows the average number of nearby locations identified for properties in our analysis set, discretised to an H3 hexagonal grid. The distribution of properties with no identified nearby locations is given in 1 (c-d). As well as true negative results, this will include any properties for which the model failed to identify an existing nearby location, along with any properties for which the model failed to provide a properly formatted output.

2.2 Model Assessment

We assess model accuracy on a subset of 100 human annotated examples which do not feature in either the analysis or training sets. For each listing, we count the number of locations described as "near" the property which are correctly identified by the model (true positives), those missed by the model (false negatives), and those which are identified by the model but either do not appear in the listing, or do appear in the listing but are not described as "near" the property (false positives).

A total of 204 (non-unique) references to proximal locations are identified in the test dataset. Thirty-three of these entries have no proximal locations mentioned, with an average of 3.1 proximal locations per remaining property.

Model performance is assessed through calculation of precision, P, and recall, R, calcuclated with respect to the true positive count, TP, false positive count, FP, and false negative count, FN. Precision and Recall are calculated as follows:

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

We also report the $F1$ score, calculated as the harmonic mean of P and R:

$$F_1 = \frac{2 \times P \times R}{P+R}. \tag{3}$$

Specifically, we report the *micro* P, R and F_1 scores, since they are calculated using aggregated TP, FP, and FN counts across the entire test set. It is also possible to calculate the individual P, R and F_1 for each observation (i.e., the average accuracy for each individual listing), known as the *macro* accuracy. However, this approach can be biased by the occurrence of observations which contain only a small number of true positives, of which there are many in this dataset. As such, we report only the *micro* statistics in this paper.

For the purposes of this analysis, false positives are significantly more damaging than false negatives. There are several ways in which false positive can occur. A model might hallucinate, identifying locations which are not mentioned at all in the text; it might extract irrelevant information from the text, such as the identification of ambiguous or non-specific locations; or it might erroneously identify parent locations as nearby. Further errors may arise from misconstructed outputs which are unable to be read as a JSON file. We report the rates at which these errors occur as a proportion of all positive identifications.

2.3 Knowledge Distillation

Larger models typically perform significantly better than smaller models on linguistic tasks, albeit with increased computational cost (Bianchini et al., 2024). We adopt a knowledge distillation approach which uses outputs from a larger model to fine-tune the weights of a smaller model. This approach, also known as a teacher/student method, leverages the task generalisation capabilities of the larger *teacher* model in an effort to produce a smaller *student* model which is better optimized to the specific task (Cantini et al., 2024).

For our teacher model, we use the 70 billion parameter Llama-3.3-70B-instruct model, quantized to 4-bits (Jin et al., 2024) using the Unsloth package (Daniel Han and team, 2023). The teacher model is used to processes 2500 listings in our model development set, using the prompt described above. This process takes a total of 42 hours using an NVIDIA RTX A6000 Graphical Processing Unit (GPU).

The output of the teacher model is used to fine-tune the 8 billion parameter Llama-3.1-8B-Instruct, again using the Unsloth package. This smaller, more task-specific model (Pornprasit and Tantithamthavorn, 2024) is then used to process the full analysis dataset of 20000 listings. The fine-tuned model takes around 20 hours to process the full dataset using the same NVIDIA GPU. The fine-tuned model is available on HuggingFace [1].

2.4 Cleaning model hallucinations

Generative models, such as Llama-3, are prone to hallucination (Jiang et al., 2024; Semnani et al., 2023). While fine-tuning does offer some mitigation (Semnani et al., 2023), it does not completely eliminate the risk. Most often, hallucinations occur when a listing does not mention any nearby locations. In most cases, the model will return an empty list, indicating that it has not found any locations mentioned in the listing, however in some instances it provides an inferred list of potential nearby locations, given the prior knowledge that the property is in London. While this is an infrequent occurrence, it clearly has potential to significantly influence our results.

Most hallucinations can be removed from the model output with a simple check to see whether the identified location actually occurs in the text. However, care must be taken with this approach, since it negates the flexibility of LLMs to misspellings and ambiguous language. For example, a listing which describes a property near “Waterloo and Charing Cross stations” should identify the nearby locations “Waterloo station” and “Charing Cross Station”. Since the string “Waterloo Station” does not appear in the text, a direct matching method would erroneously mark "Waterloo Station" as a hallucination. To address this, we employ a fuzzy matching system to remove substrings which do not occur within the text (Bosker, 2021).

Fuzzy matching requires the setting of a threshold ratio for acceptable similarity. If this ratio is too strict (i.e. we only accept locations which are very similar to that appearing in the text) we risk unduly increasing the false negative rate of our model. Since we know the number of hallucinations which the model produces on our test set, we can set the threshold so that the number of rejected locations is approximately equal to the number of hallucinations. This approach minimises the increase in false negatives, while reducing the number of false positives.

We use the RapidFuzz library for python (Bachmann, 2024) to acheive this. The library uses the Indel distance (Hyyrö et al., 2005) to calculate the degree of similarity between sets of strings. Strings which are identical have a similarity ratio of 1.0, while strings with no common sub-strings have a similarity ratio of 0. We find that a threshold similarity ratio of 0.7 is sufficient for reducing the number of false positives in our test set, without any significant increase in the false negative rate.

2.5 Spatial Analysis

To explore the spatial characteristics of proximal locations, we consider the spatial distribution of AirBnB properties which are described as "near to" 109 reference locations in London. This set of reference locations includes every location which is mentioned in at least 50 property listings.

Reference locations can include points of interest (POIs) such as Oxford Street or Trafalgar Square, regions of interest (ROIs) such as Camden, Richmond or Canary Wharf, and transit stations (London Underground stations and national rail stations). We classify a location as an ROI if it is listed as a "borough" or "suburb" in OpenStreetMaps (OSM). This means small regions described in OSM as "localities" or "quarters" (such as Angel, or South Kensington) are classified as POIs. In many cases, a POI or ROI may also be a transit station. For example "Richmond" is the name of a station on the District line of the London Underground, and the London Borough in which the station is found. Similarly, the popular tourist destination "Covent Garden" also has an identically named station.

We assess the spatial distribution of the identified proximal locations for each reference location by considering their standard distance (Bachi, 1962). For a reference location p, we have some set of proximal locations $q \in Q$, with coordinates (q_x, q_y). The standard distance for location p, $\hat{d}(p)$, is then defined as the square root of the mean squared geodesic distance between each proximal location, $q = (q_x, q_y)$, and the mean coordinate of all proximal locations $\bar{q} = (\bar{q}_x, \bar{q}_y)$:

$$\hat{d}(p) = \sqrt{\frac{\sum_{q \in Q} \mathrm{dist}(q, \bar{q})^2}{|Q|}}, \tag{4}$$

[1] https://huggingface.co/JoeShingleton/ProxiLlama-3.1-8b

For a given reference location, the standard distance represents the degree of spatial dispersion of its identified proximal locations. This rate of dispersion could also be calculated as the mean distance between the reference location and the identified proximal locations. Analysis of our dataset shows that the standard distance and average distance are highly correlated with one another. As such, we only report the standard distance.

2.6 Comparison to Public Transport Access

Accessibility is often cited as playing an important role in the generalisation of proximity beyond geometric distance (Ma et al., 2019; Orrego-Oñate and Marquet, 2025). We use two approaches to assess the extent to which this is true for the AirBnB dataset. First, we perform a visual analysis of the distribution of perceived proximal locations to a set of London Underground stations, and the transit lines they serve. We also compare the "connectedness" of stations with the standard distance of their proximal locations.

In order to better understand the relationship between public transport access and perceived proximity, we compare the standard distance of all our reference locations with the maximum reported Public Transport Accessibility Level (PTAL) (Transport for London, 2015) within 100m of the reference location. Public Transport Accessibility Levels are calculated using the walking distance to the nearest public transport access point (e.g. a station or bus stop), the highest frequency of service available at that access point, and a weighting system which accounts for the attractiveness of a route over other routes. A score between 0 and 6 is given for each 100m square in London, with scores of 1 and 6 further separated into levels 1a, 1b, 6a and 6b.

The PTAL score is a discretisation of the calculated Accessibility Index (AI), which is a continuous variable indicating the level accessibility. For the purposes of this paper, we report the raw AI scores, rather than the PTAL category. We use AI scores from 2015, as more recent data is not currently available. Figure 2 shows the AI scores for London, spatially discretised onto an H3 hexagonal grid and overlayed with the London Underground network.

2.7 Data and Software Availability Section

The code and data associated with this paper can be found in the Open Science Foundation repository: https://osf.io/r3ep7/. This includes the python scripts and LLM prompts used to process the AirBnB listings, the processed testing, training and analysis sets, and two Jupyter notebooks demonstrating the model validation process and the analytical pipeline. A Read-Me file detailing the environment requirements necessary to reproduce the results is also available.

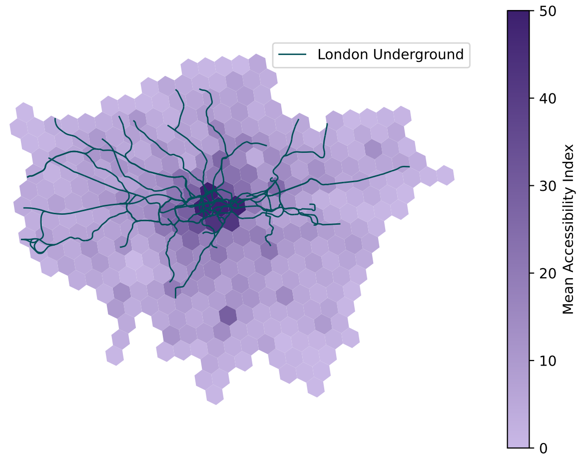

Figure 2. The accessibility Index (AI) for London in 2015, discretised into an H3 hexagon grid. The map has been overlayed with the London Underground network.

3 Results

Results are separated in to two sections - model development and analysis. First, we assess the accuracy of four different LLMs on our test set of 100 listings. The best performing model is then used to parse the full dataset of 20000 listings. We report the spatial characteristics of all locations which are mentioned in at least 50 AirBnB listings, termed "reference locations".

3.1 Accuracy of LLMs on Proximal Location Identification

Table 1 provides the precision (P), recall (R) and f1 (f_1) score for four models: Llama-3b-Instruct, Llama-70b-Instruct-4bit, Llama-3b-Instruct-ft and Lama-8b-Instruct-ft. The latter two models have been fine-tuned on 2500 examples produced using the 70B parameter model. We also report the average time taken to process each listing, t. In each case, the models have been assessed using the same sample of 100 AirBnB listings.

In table 2 we report the rate at which different types of errors occur for each model. The table provides the hallucination rate (H_r), irrelevancy rate (I_r), misconstruction rate (M_r) and parent identifi-

Model	P	R	F_1	t (sec/listing)
Llama-3b	0.697	0.485	0.572	9.63
Llama-70b	0.909	0.833	0.869	66.56
Llama-3b-ft	0.791	0.740	0.765	6.11
Llama-8b-ft	0.827	0.775	0.800	5.22

Table 1. The accuracy of four LLMs on the task of proximal location identification in 100 AirBnB property descriptions. The first two models have been run with a "zero-shot" method, while the second two have been fine-tuned on 2500 samples tagged by the 70B parameter model. For each, we report the precision, recall and F1 scores (P, R and F_1), as well as the average time taken to process each listing (t).

cation rate (P_r) for each model, calculated as the ratio of erroneous identifications of each type to the total number of positive identifications.

Model	H_r	I_r	P_r	M_r
Llama-3b	0.120	0.085	0.113	0.028
Llama-70b	0.016	0.049	0.016	0.005
Llama-3b-ft	0.084	0.084	0.047	0.010
Llama-8b-ft	0.094	0.063	0.021	0.005

Table 2. The sources of errors in the proximal location tagging task for each model. Error rates are categorised into hallucinations (H_r), irrelevant information identification (I_r), parent location identification (P_r) and misconstructions (M_r), with each reported as a proportion of all positive identifications.

The best performing model is Llama-70B-Instruct. This is expected, since larger models tend to significantly outperform smaller models in linguistic exercises (Bianchini et al., 2024). However, since the 70B parameter model is significantly more computationally expensive, it can not be reasonably used to process very large datasets. The fine-tuned models perform comparably to the larger model for significantly reduced computational expense.

After both models have been fine-tuned, the 8b parameter model outperforms the 3b parameter model. Despite this, hallucinations still account for 9.4% of proximal location identification in the fine-tuned 8B parameter model. As discussed in the methods section, hallucination errors can be cleaned relatively easily using a fuzzy-matching approach. Applying this approach reduces the hallucination rate to $H_r = 0.011$ and increases the precision to $P = 0.893$. The remaining false positives arise from ambiguous and/or non-specific locations mentioned, or from misidentification of parent locations. While these are more difficult to clean, they will only have limited influence on our results - as will be discussed in the next section.

3.1.1 POI Area of Influence

The proximal location identification process identified a total of 7405 unique locations in the 20000 listings. In some cases, however, a single POI might be referred to by multiple terms. Euston Station, for example, might also be referred to as "Euston" or "Euston Train Station", similarly, "Regent's Park" might be referred to as "Regents Park". The most common occurrences of these are identified, and the results for each representation are aggregated together.

For all locations with more than 50 identified proximal properties (the *reference* locations, total $N = 109$), we calculate the standard distance of the proximal properties as described in the methods section. Figure 3 shows the standard distance plotted against the total number of identified proximal properties for (a) points and regions of interest, and (b) transit stations. London Underground stations are further separated according to whether they additionally classify as a POIs or ROIs (e.g., Covent Garden, Regent's Park, Clapham Common are also POIs, whereas Sloane Square, Old Street and Bank are not). Figure 3 (c) compares the standard distances for attractions in each of the five categories.

Regions of interest (such as Camden, Canary Wharf or Stratford) tend to have a higher area of influence compared to point-like attractions, as measured by the standard distance for proximal Air-BnB properties. Rail stations have larger areas of influence than other types of attraction, suggesting a correlation between proximity and function.

The "platial" component to proximity can be further examined by considering the distribution of proximal properties identified for underground stations, with respect to the number of lines they serve. Figure 4 shows the distribution of proximal locations for six London Underground stations (Bond Street, Bank, Angel, Baker Street, Wimbledon and Brixton), along with two stations which are also connected to the national rail network (King's Cross and Waterloo). The lines served by each station are also shown, and the national rail network within London is shown in the last two maps.

We can clearly see that properties considered proximal to a station tend to be placed along the lines served by the station. We also see that location of the station relative to the centre of London appears to influence their distribution. Baker street sits to

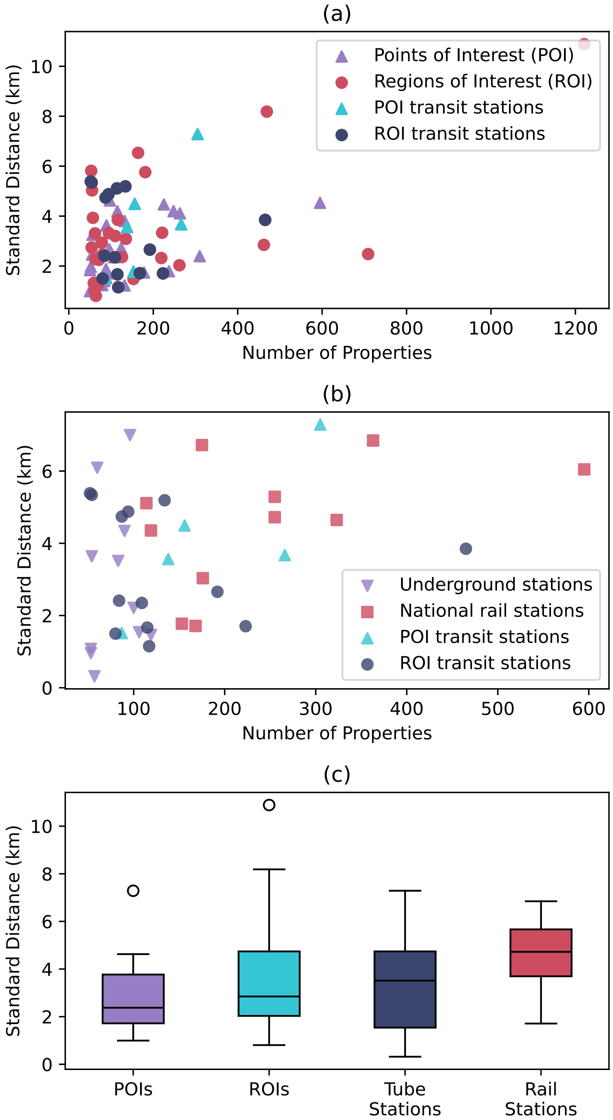

Figure 3. (a) The standard distance of all properties described as "near to" reference locations for POIs and ROIs, (b) for transit stations and (c) a box plot of standard distance for the four categories.

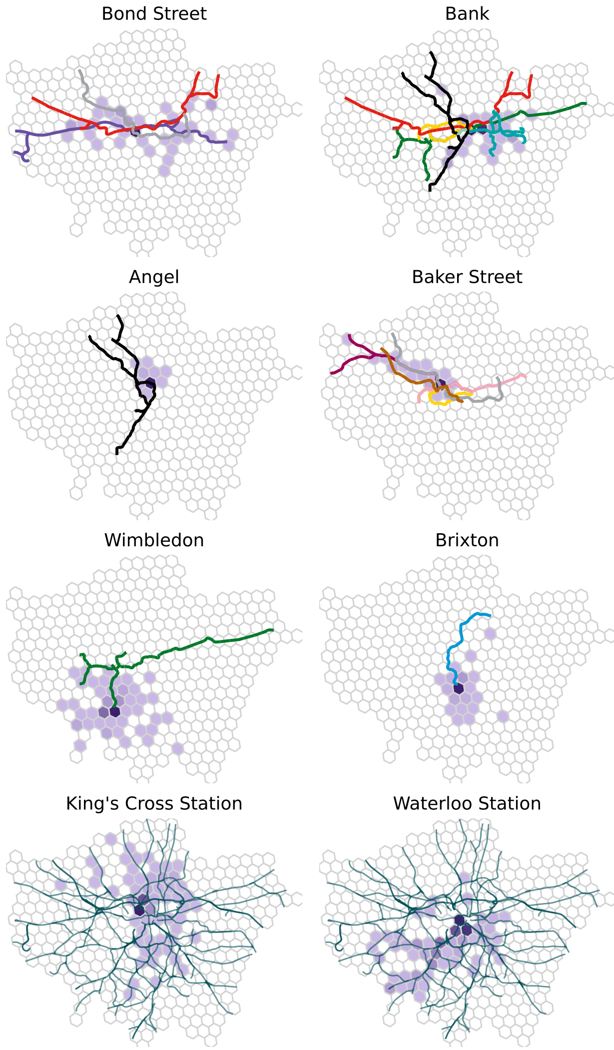

Figure 4. The distribution of properties which describe London Underground stations as proximal. Shaded hexagons indicate the presence of Proximal properties, with darker shading indicating a high density. London Underground lines which serve the specified stations have been highlighted. The final two plots show stations with connections to the national rail network, with the rail network also plotted for reference.

the west of central London, and we clearly see its area of influence extends to the west, with little influence to the east. Bank, a well connected station to the east of central London, is considered proximal to more locations to the south and east. Bond Street, which is closer to the centre of London, has influence both to the east and west. The area of influence of the National Rail Stations significantly overshadows that of Underground stations, but retains a tendency to follow transit lines.

Regions which also share a name with a transit station, such as Wimbledon and Brixton, have significant reach, even if the station only serves a single line. This is highlighted when we compare Wimbledon and Brixton to Angel - a small locality which is also only served by a single line - with Angel having a significantly narrower area of influence compared to the larger regions. This result suggests that the influence of cultural prominence, administrative hierarchy and spatial extent are more important than public transportation access for these locations.

Figure 5 shows (a) how the standard distance of proximal locations for London underground stations varies with the station's total number of proximal locations, and (b) with the average weekly entry/exit counts for each station for 2023, as given by Transport for London Open Data (Transport for London, 2023). For both, the size of the marker indicates the number of lines served by that station. We see that stations with fewer lines tend to have a narrower area of influence, compared to interchange stations. Likewise, stations which are used more frequently are more often described as being proximal to properties which are further away.

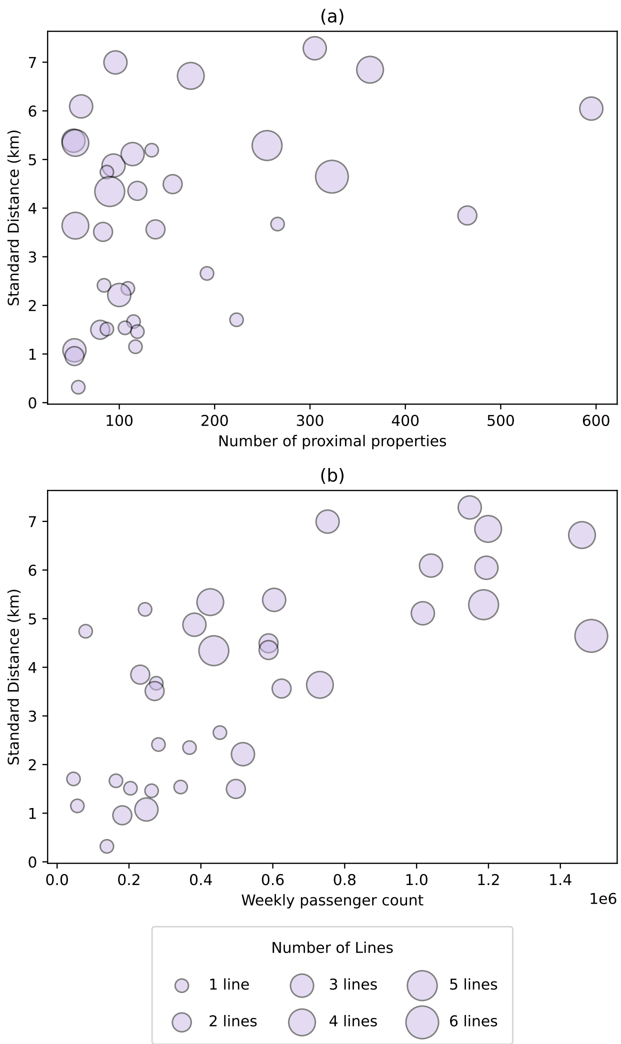

Figure 5. The relationship between (a) the number of properties described as proximal to each London Underground station, and the standard distance of those proximal properties, and (b) the 2023 weekly entry/exit count for each station. Marker size indicates the number of lines served by the station.

There are two stations which retain a high standard distance, despite only serving single lines and having relatively low passenger counts. These outliers - Wimbledon and Richmond - both also refer to large areas in London which has likely influenced this result. In most instances, it is likely the listing was referring to the region, rather than the station, although we have not investigated this fully. Nevertheless, this again suggests that both spatial extent, administrative hierarchy and cultural prominence influence perceived proximity in a way which may supersede accessibility.

3.2 Comparison to the Accessibility Index

Figure 6 compares the maximum Accessibility Index within 100m of each reference location with the standard distance of the proximal AirBnB properties associated with the location. The plot has been separated into (a) transit stations and (b) POIs and ROIs, with categorisation into different classes as above. For both plots, we see a weak positive correlation between the level of public transport accessibility for a location, and the distribution of properties considered proximal to it. In plot (a) we see that underground stations which are also ROIs (as discussed above) tend to fall below the regression line, again suggesting an interaction between administrative hierarchy and perceived proximity.

4 Discussion

Despite its seemingly intuitive nature, proximity is a difficult relationship to rigorously define. Other spatial relationships such as adjacency, containment or intersection can be easily defined in terms of purely geometric properties. Proximity, on the other hand, is necessarily dependent on non-spatial factors (Brennan and Martin, 2012). In this paper, we have investigated the perceived proximity of 109 reference locations in London by leveraging the linguistic comprehension capacity of LLMs to analyse descriptions of 20000 AirBnB properties. This data driven approach to proximity allows us to uncover both spatial and platial aspects underpinning the perception of proximity.

Our results have demonstrated that the factors influencing proximty go far beyond simple distance. We have shown that relative accessibility of a location plays an important role, with locations which are more accessible tending to have a wider distribution of perceived nearby locations. For transit stations - having a higher weekly usage and serving a larger part of the transit network tends to increase this area of influence, suggesting a relationship between utility, notarity and perceived proximity. Administrative hierarchy, cultural prominence and spatial extent of a location also influences its perceived proximity, with large regions of interest, such as "Richmond" or "Wim-

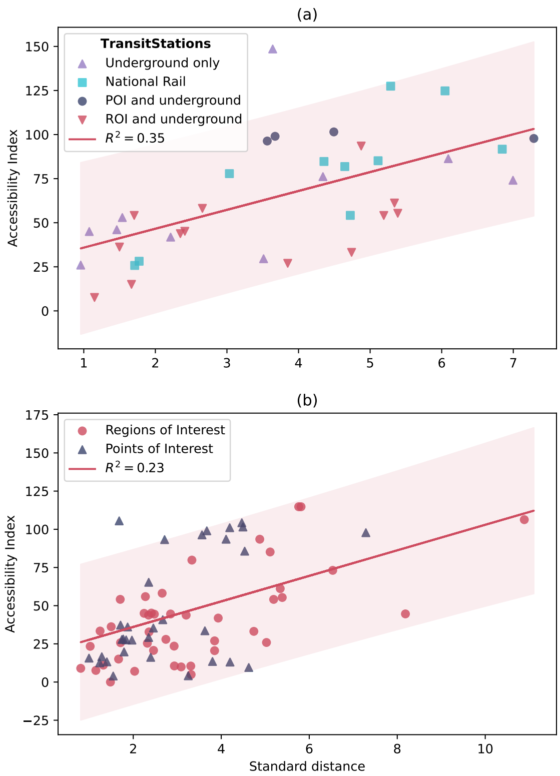

Figure 6. The standard distance of properties described as proximal to each reference location, plotted against the maximum PTAL score within 1km of the reference location. Transit stations are shown in figure (a), while figure (b) shows all points and regions of interest. In both plots, the central red line is a linear regression through all points, with the 95% prediction interval indicated by the shaded area. The R2 scores are (a) 0.35 and (b) 0.23.

bledon" having higher than expected areas of influence, given their accessibility scores.

However, it is important to note that this investigation has identified correlative, rather than causative, associations between platial factors and perceived proximity. As with any study of this type, there are likely hidden factors which drive this correlation. For example, the use of standard distance as an indicator of spatial distribution does not account for factors such as the relative density of nearby buildings. Similarly, the Accessibility Index, used to assess relationships between proximity and connectivity, is likely highly correlated with other factors such as touristic prominence or topology, making it difficult to identify a causal relationship.

The results discussed in this paper only consider proximity in one context - that of AirBnB properties. AirBnB properties typically cater towards short-term rentals, with a particular focus on leisure and tourism (Koster et al., 2021). Users of AirBnB properties tend to be well educated and younger than non-users, and are more likely to be married, have children and have higher incomes (Mody et al., 2017). These biases are likely to influence our results, and somewhat limit the generalisability of our findings. However, the approach used can be easily applied to other types of data to identify the characteristics of proximity in different contexts. For example, real estate listings such as Zoopla or Right Move, news and event reporting, or travel and tourism blogs all might uncover additional nuance to the notion of proximity. By considering how nearness is perceived within these different contexts, we can build a clearer picture of the subjective factors influencing proximity.

We have considered London as the sole research area for this paper. London is a large city with a dense public transportation system. Further work may compare how perceived proximity differs for cities with less comprehensive transport systems. While the specific patterns observed here may not transfer directly to other cities, or within other contexts, we expect that the broader relationships between proximity, accessibility and function will hold across similar analyses of different datasets.

Using our processed dataset, proximity can be investigated in two directions - the distribution of properties considered proximal to a given reference location, and the distribution of locations considered proximal to a given property. In this paper, we have only considered the former. Viewing the data from the reverse direction would allow us to calculate an average "proximal distance" map for London. However, this is a more complicated task, since a precise geolocation for each identified location would be required in each listing. This would be very sensitive to false positives, which can often be introduced through erroneous geocoding (Gritta et al., 2020). As such, this is outside the scope of this paper, but presents an interesting avenue for future research.

This paper has demonstrated a novel technique for using generative artificial intelligence to extract geospatial relationships from user-generated data. By developing an understanding of how proximity is discussed in real-life examples, we can begin to improve geographic information systems which rely on well-defined proximity relationships.

5 Acknowledgment

The authors acknowledge the support from The UK Research and Innovation (UKRI) Future Leaders Fellowship on "Indicative Data", MR/S01795X/2, and the Alan Turing Institute-DSO partnership project on "Multi-Lingual and Multi-Modal Location Information Extraction".

The original draft of this paper was written without the use of generative AI. Generative AI was subsequently used to edit the final draft, suggest minor improvements to wording and structure, and assess whether revisions adequately addressed reviewers' comments.

6 Author Contribution

JS contributed to conceptualisation and scoping, data collection and wrangling, software development, methodology and analysis, writing, visualisation, editing and reviewing. AB contributed to ideation, conceptualisation, scoping, editing and reviewing.

References

Alsudais, A.: Incorrect data in the widely used Inside Airbnb dataset, Decision Support Systems, 141, 113453, https://doi.org/https://doi.org/10.1016/j.dss.2020.113453, 2021.

Baccianella, S.: JSON Repair - A python module to repair invalid JSON, commonly used to parse the output of LLMs, https://github.com/mangiucugna/json_repair, 2024.

Bachi, R.: STANDARD DISTANCE MEASURES AND RELATED METHODS FOR SPATIAL ANALYSIS, Papers in Regional Science, 10, 83–132, https://doi.org/https://doi.org/10.1111/j.1435-5597.1962.tb00872.x, 1962.

Bachmann, M.: rapidfuzz/RapidFuzz: Release 3.8.1, https://doi.org/10.5281/zenodo.10938887, 2024.

Bharathi Mohan, G., Prasanna Kumar, R., Vishal Krishh, P., Keerthinathan, A., Lavanya, G., Meghana, M. K. U., Sulthana, S., and Doss, S.: An analysis of large language models: their impact and potential applications, Knowledge and Information Systems, 66, 5047–5070, https://doi.org/10.1007/s10115-024-02120-8, 2024.

Bianchini, F., Calamo, M., De Luzi, F., Macrì, M., and Mecella, M.: Enhancing Complex Linguistic Tasks Resolution Through Fine-Tuning LLMs, RAG and Knowledge Graphs (Short Paper), in: Advanced Information Systems Engineering Workshops, edited by Almeida, J. P. A., Di Ciccio, C., and Kalloniatis, C., pp. 147–155, Springer Nature Switzerland, Cham, 2024.

Binbin Lu, Martin Charlton, P. H. and Fotheringham, A. S.: Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data, International Journal of Geographical Information Science, 28, 660–681, https://doi.org/10.1080/13658816.2013.865739, 2014.

Bosker, H. R.: Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies, Behavior Research Methods, 53, 1945–1953, https://doi.org/10.3758/s13428-021-01542-4, 2021.

Brennan, J. and Martin, E.: Spatial proximity is more than just a distance measure, International Journal of Human-Computer Studies, 70, 88–106, https://doi.org/https://doi.org/10.1016/j.ijhcs.2011.08.006, 2012.

Cantini, R., Orsino, A., and Talia, D.: Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices, Journal of Big Data, 11, 63, https://doi.org/10.1186/s40537-024-00928-3, 2024.

Chrysostomou, G., Zhao, Z., Williams, M., and Aletras, N.: Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization, Transactions of the Association for Computational Linguistics, 12, 1163–1181, https://doi.org/10.1162/tacl_a_00695, 2024.

Comber, A., Chi, K., Huy, M. Q., Nguyen, Q., Lu, B., Phe, H. H., and Harris, P.: Distance metric choice can both reduce and induce collinearity in geographically weighted regression, Environment and Planning B: Urban Analytics and City Science, 47, 489–507, https://doi.org/10.1177/2399808318784017, 2020.

Daniel Han, M. H. and team, U.: Unsloth, http://github.com/unslothai/unsloth, 2023.

Derungs, C. and and, R. S. P.: Mining nearness relations from an n-grams Web corpus in geographical space, Spatial Cognition & Computation, 16, 301–322, https://doi.org/10.1080/13875868.2016.1246553, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13875868.2016.1246553, 2016.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), edited by Burstein, J., Doran, C., and Solorio, T., pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, https://doi.org/10.18653/v1/N19-1423, 2019.

Gahegan, M. and Lee, I.: Data structures and algorithms to support interactive spatial analysis using dynamic Voronoi diagrams, Computers, Environment and Urban Systems, 24, 509–

537, https://doi.org/https://doi.org/10.1016/S0198-9715(00)00007-7, 2000.

Grattafiori, A., Dubey, A., and et al, A. J.: The Llama 3 Herd of Models, https://arxiv.org/abs/2407.21783, 2024.

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.

Grütter, R.: A framework for assisted proximity analysis in feature data, Journal of Geographical Systems, 21, 367–394, https://doi.org/10.1007/s10109-019-00304-3, 2019.

Grütter, R., Scharrenbach, T., and Waldvogel, B.: Vague spatio-thematic query processing: A qualitative approach to spatial closeness, Transactions in GIS, 14, 97 – 109, https://doi.org/10.1111/j.1467-9671.2010.01185.x, 2010.

Guesgen, H. W. and Albrecht, J.: Imprecise reasoning in geographic information systems, Fuzzy Sets and Systems, 113, 121–131, https://doi.org/https://doi.org/10.1016/S0165-0114(99)00016-0, 2000.

Hall, M. M. and Jones, C. B.: Generating geographical location descriptions with spatial templates: a salient toponym driven approach, International Journal of Geographical Information Science, 36, 55–85, https://doi.org/10.1080/13658816.2021.1913498, 2022.

Hyyrö, H., Pinzon, Y., and Shinohara, A.: New Bit-Parallel Indel-Distance Algorithm, in: Experimental and Efficient Algorithms, edited by Nikoletseas, S. E., pp. 380–390, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

Jiale Ding, Wenying Cen, S. W. Y. C. J. Q. B. H. and Du, Z.: A neural network model to optimize the measure of spatial proximity in geographically weighted regression approach: a case study on house price in Wuhan, International Journal of Geographical Information Science, 38, 1315–1335, https://doi.org/10.1080/13658816.2024.2343771, 2024.

Jiang, C., Qi, B., Hong, X., Fu, D., Cheng, Y., Meng, F., Yu, M., Zhou, B., and Zhou, J.: On Large Language Models' Hallucination with Regard to Known Facts, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 1041–1053, Association for Computational Linguistics, Mexico City, Mexico, https://doi.org/10.18653/v1/2024.naacl-long.60, 2024.

Jin, R., Du, J., Huang, W., Liu, W., Luan, J., Wang, B., and Xiong, D.: A Comprehensive Evaluation of Quantization Strategies for Large Language Models, in: Findings of the Association for Computational Linguistics: ACL 2024, edited by Ku, L.-W., Martins, A., and Srikumar, V., pp. 12 186–12 215, Association for Computational Linguistics, Bangkok, Thailand, https://doi.org/10.18653/v1/2024.findings-acl.726, 2024.

Jingyi He, Y. W. and Yu, B.: Geographically weighted regression based on a network weight matrix: a case study using urbanization driving force data in China, International Journal of Geographical Information Science, 37, 1209–1235, https://doi.org/10.1080/13658816.2023.2192122, 2023.

Kai Cao, M. D. and Wu, B.: A Big Data–Based Geographically Weighted Regression Model for Public Housing Prices: A Case Study in Singapore, Annals of the American Association of Geographers, 109, 173–186, https://doi.org/10.1080/24694452.2018.1470925, 2019.

Klang, E., Apakama, D., Abbott, E. E., Vaid, A., Lampert, J., Sakhuja, A., Freeman, R., Charney, A. W., Reich, D., Kraft, M., Nadkarni, G. N., and Glicksberg, B. S.: A strategy for cost-effective large language model use at health system-scale, npj Digital Medicine, 7, 320, https://doi.org/10.1038/s41746-024-01315-1, 2024.

Koster, H. R. A., Ommeren, J. v., and Volkhausen, N.: Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles, Journal of Urban Economics, 124, 103 356, https://doi.org/https://doi.org/10.1016/j.jue.2021.103356, 2021.

Li, B., Jiang, G., Li, N., and Song, C.: Research on Large-scale Structured and Unstructured Data Processing based on Large Language Model, in: Proceedings of the International Conference on Machine Learning, Pattern Recognition and Automation Engineering, MLPRAE '24, p. 111–116, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3696687.3696707, 2024.

Ma, F., Ren, F., Yuen, K. F., Guo, Y., Zhao, C., and Guo, D.: The spatial coupling effect between urban public transport and commercial complexes: A network centrality perspective, Sustainable Cities and Society, 50, 101 645, https://doi.org/https://doi.org/10.1016/j.scs.2019.101645, 2019.

Miller, H. J.: Tobler's First Law and Spatial Analysis, Annals of the Association of American Geographers, 94, 284–289, http://www.jstor.org/stable/3693985, 2004.

Mocnik, F.-B.: Putting Geographical Information Science in Place – Towards Theories of Platial Information and Platial Information Systems, Progress in Human Geography, 46, 798–828, https://doi.org/10.1177/03091325221074023, 2022.

Mody, M. A., Suess, C., and Lehto, X.: The accommodation experiencescape: a comparative assessment of hotels and Airbnb, International Journal of Contemporary Hospitality Management, 29, 2377–2404, https://doi.org/10.1108/IJCHM-09-2016-0501, 2017.

Murray Cox, John Morris, T. H.: Inside Airbnb, https://insideairbnb.com/about/, 2025.

Orrego-Oñate, J. and Marquet, O.: The role of perceived and objective accessibility in shaping walking behavior: Insights from mid-sized Spanish cities, Transportation Research Part A: Policy and Practice, 192, 104 340, https://doi.org/https://doi.org/10.1016/j.tra.2024.104340, 2025.

Pornprasit, C. and Tantithamthavorn, C.: Fine-tuning and prompt engineering for large language models-based code review automation, Information and Software Technology, 175, 107 523, https://doi.org/https://doi.org/10.1016/j.infsof.2024.107523 2024.

Sahr, K.: Hexagonal discrete global grid systems for geospatial computing, Archiwum Fotogrametrii, Kartografii i Teledetekcji, 22, 363–376, 2011.

Semnani, S., Yao, V., Zhang, H., and Lam, M.: WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia, in: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 2387–2413, Association for Computational Linguistics, Singapore, https://doi.org/10.18653/v1/2023.findings-emnlp.157, 2023.

She, B., Zhu, X., Ye, X., Guo, W., Su, K., and Lee, J.: Weighted network Voronoi Diagrams for local spatial analysis, Computers, Environment and Urban Systems, 52, 70–80, https://doi.org/https://doi.org/10.1016/j.compenvurbsys.20 2015.

Shingleton, J. and Basiri, A.: Enhancing toponym identification: Leveraging Topo-BERT and open-source data to differentiate between toponyms and extract spatial relationships, AGILE: GIScience Series, 5, 12, https://doi.org/10.5194/agile-giss-5-12-2024, 2024.

Stock, K., Jones, C. B., Russell, S., Radke, M., Das, P., and and, N. A.: Detecting geospatial location descriptions in natural language text, International Journal of Geographical Information Science, 36, 547–584, https://doi.org/10.1080/13658816.2021.1987441, 2022.

Transport for London: Public Transport Accessibility Levels, https://data.london.gov.uk/dataset/public-transport-accessibility-levels, 2015.

Transport for London: Crowding Annualised Entry Exit 2023, https://crowding.data.tfl.gov.uk/, 2023.

Appendix

Figure A1 shows the prompt used to identify proximal locations in AirBnB listings. The prompt was engineered to work well with the Llama3 family of LLMs.

Prompt used for proximity extraction

You are an expert in linguistic analysis. Analyze the description of an Airbnb property in London, to identify any locations explicitly described as "near" the property.
Guidelines
Identify Locations:
Locations should be explicitly named (e.g., "Hyde Park" or "Oxford Street").
Exclude vague mentions like "restaurant" or "station" unless paired with a specific name (e.g., "Luigi’s Restaurant" or "Euston Station").
Include linear locations (e.g., roads, metro lines, rivers) if specifically named.
Proximity Context:
Provide the exact phrasing or linguistic context indicating proximity.
Exclude references to the property’s location itself (e.g., "in Brixton" does not count as nearby).
Searchable Information:
Provide an approximate address or identifiable details for locating it on OpenStreetMap.
Include approximate latitude and longitude when possible.
Output Format:
The output must be a JSON array in the following format:
["name": "<name of nearby location>", "address": "<approximate address>", "latitude": "<estimated latitude>", "longitude": "<estimated longitude>", "context": "<proximity context>" , ...]
If no locations are identified, return an empty array: [].
Restrictions:
Do not infer or add any information beyond what is explicitly stated in the listing.
Do not include commentary or explanation in your response.

Figure A1. Full prompt provided to the LLM for identifying perceived proximity relationships.