Geography According to Foundation Models
K. Janowicz et al. (Eds.)

© 2026 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA260476

Geospatial Mechanistic Interpretability of Large Language Models

Stef DE SABBATA [a,1], Stefano MIZZARO [b], and Kevin ROITERO [b]
[a] *University of Leicester, UK*
[b] *University of Udine, Italy*
ORCiD ID: Stef De Sabbata https://orcid.org/0000-0002-2750-7579, Stefano Mizzaro https://orcid.org/0000-0002-2852-168X, Kevin Roitero https://orcid.org/0000-0002-9191-3280

Abstract. Large Language Models (LLMs) have demonstrated unprecedented capabilities across various natural language processing tasks. Their ability to process and generate viable text and code has made them ubiquitous in many fields, while their deployment as knowledge bases and "reasoning" tools remains an area of ongoing research. In geography, a growing body of literature has been focusing on evaluating LLMs' geographical knowledge and their ability to perform spatial reasoning. However, very little is still known about the internal functioning of these models, especially about how they process geographical information.

In this chapter, we establish a novel framework for the study of geospatial mechanistic interpretability – using spatial analysis to reverse engineer how LLMs handle geographical information. Our aim is to advance our understanding of the internal representations that these complex models generate while processing geographical information – what one might call "how LLMs think about geographic information" if such phrasing was not an undue anthropomorphism.

We first outline the use of probing in revealing internal structures within LLMs. We then introduce the field of mechanistic interpretability, discussing the superposition hypothesis and the role of sparse autoencoders in disentangling polysemantic internal representations of LLMs into more interpretable, monosemantic features. In our experiments, we use spatial autocorrelation to show how features obtained for placenames display spatial patterns related to their geographic location and can thus be interpreted geospatially, providing insights into how these models process geographical information. We conclude by discussing how our framework can help shape the study and use of foundation models in geography.

Keywords. Foundation Models, GeoAI, Large Language Models, Mechanistic Interpretability, Probing, Spatial Analysis, Spatial Autocorrelation

1. Introduction

In the past few years, the field of Artificial Intelligence (AI) has seen an unprecedented pace of change, from the introduction of the transformers architecture in 2017 to the release of ChatGPT in 2022. The ability of Large Language Models (LLMs) to produce human-like (at least superficially) answers to questions took many by surprise. However,

[1] Corresponding Author: Stef De Sabbata.

two years forward, the technology had already become an off-the-shelf commodity, and new research streams have emerged that focus on exploring the capabilities of these models in a broad range of specific tasks and subjects. In geography, several authors have explored the ability of LLMs to process geographical information [1,2], including their ability to perform a broad range of spatial tasks [3,4] and infer correct answers to spatial reasoning questions [5,6,7], resolve toponyms [8], assist practitioners in GIS [9,10,11], cartography [12], earth observation [13] and urban planning [14,15], process social media posts during natural disasters [16] and visual geographical information such as street view and satellite images [17], understand urban spaces [18], discern intercardinal directions [19], provide recommendations for points of interest [20] or routing [21], while other authors have focused on the quality [22,23], bias [24], and diversity [25,26] of LLMs' geographical knowledge. However, our understanding of the internal functioning of LLMs is very limited, especially when it comes to geographical information.

Due to the semantic footprints of places in textual content [27] and LLMs being trained as contextual next-word predictors, it seems likely that an LLM's reaction to placenames referring to places which are nearby or otherwise related to each other might be somehow similar, at least in some facets of the complex internal representation that these model construct when processing information. For instance, both Wikipedia pages about Loughborough[2] and Market Harborough[3] mention "market town" in their first paragraph and "Leicester" in their second paragraph. As LLMs have been trained on content from the web, including Wikipedia, it stands to reason that receiving the words "Loughborough" or "Market Harborough" as input would crank millions of an LLM's internal "levers" in position to heighten the probabilities of one of the next words being "market town" or "Leicester" – unless "Loughborough" is followed by "Lake"[4], but we will get to placename ambiguity further below.

To clarify the terminology used in the rest of the chapter – a neural network can be thought of as composed of *neurons* laid out in subsequent *layers* with *edges* connecting neurons at one layer to neurons at the next layer [28]. Each edge contains one of the "levers" mentioned above, which are called *parameters* (or *weights*) and control the flow of information through the network. Each neuron of the first layer corresponds to a piece of the input (e.g., a pixel of an image, a column of a table, or a word in a sentence). Each piece of information of the input is pushed forward from its neuron at the first layer through the edges towards the second layer. In transit, each edge uses its own parameter to modify the input into a new value, which arrives at the neuron at the second layer. That neuron combines together all the values from all its incoming edges and applies an *activation function* (usually a non-linear function such as a sigmoid or a ReLU [29]) to produce a value commonly referred to as *activation*. The *transformer layers* [30], of which LLMs are composed, are sophisticated blocks composed of several layers, including a mechanism called *attention*, which allows LLMs to better understand how different parts of a text influence each other's meaning. The analyses in this chapter focus on the activations extracted after the attention mechanism, before the *Multi-Layer Perceptron (MLP)* component of the transformer layer, as illustrated in Figure 1. A set of activations from the same layer is called an internal *representation* [31] of the input at that specific layer, and if it is identified as something humanly interpretable – such as

[2] https://en.wikipedia.org/wiki/Loughborough
[3] https://en.wikipedia.org/wiki/Market_Harborough
[4] https://en.wikipedia.org/wiki/Loughborough_Lake

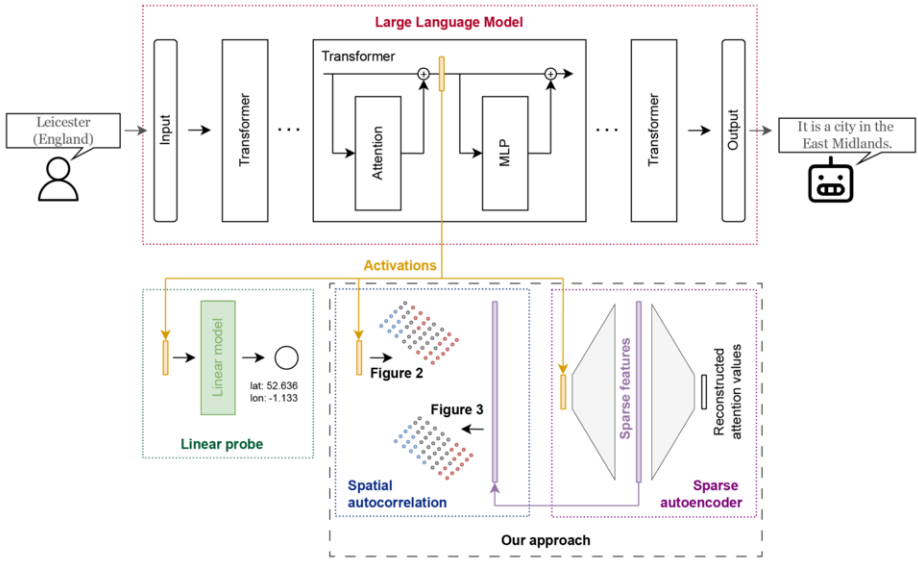

Figure 1. An example illustrating the extraction of the activations from an LLM (top); the use of the activations in a linear probe to predict the latitude and longitude of the place mentioned in the input (bottom-left) and a sparse autoencoder (bottom-right); and the use of spatial autocorrelation to analyse the activations and the sparse features (centre). Our approach encompasses the latter two components.

high output values when the input references to the Golden Gate Bridge in a language model [32] – it is called a *feature* of the input.

The activations are then passed onward to the subsequent layer until the last output layer is reached. During training, the output is compared to the expected output using a *loss function*, and *backpropagation* [33] is used to change the parameters to (hopefully) achieve lower loss (i.e., error) at the next training step. The set of steps necessary to train a model once on all the data is called an *epoch*. The number of epochs necessary to train a model depends on factors such as the size and diversity of the training dataset, the size and architecture of the model, the type and difficulty of the task, and the intended qualities of the model, such as accuracy and generalisability. LLMs are initially trained to predict the next word in a sentence, in the phase commonly referred to as pre-training, before the model is fine-tuned for a specific task, such as being a chatbot. Finally, many LLMs go through a phase of reinforcement learning with human feedback, where they are trained to produce the desired type of answers based on positive or negative feedback from human judgements [34]. Once the training process is completed, the LLM parameters (the "levers") are fixed, and an input sentence will pass through the neural network, generating internal representations at each layer until the last layer selects the next word to add to the output.

This chapter focuses on the analysis of those internal representations and the question of whether training an LLM (at its core) as a next-word prediction might lead them to follow patterns akin to Tobler's "first law of geography" [35], where *input text about near things might produce internal representations that are more similar than input text about distant things*. More concretely, we can start from the assumption that a high activation value at a specific neuron indicates that the model has activated specific paths

through the neural network to increase the probability of certain output tokens. As the model has been trained on large amounts of human-generated text, including references to and descriptions of places, we can assume that the model has encoded information about places, their geographies and characteristics. Therefore, we can formulate the following conjecture: close or otherwise similar places are likely to be associated with similar LLM internal representation patterns.

Recent studies [36,37,38,39] were able to use linear probes to find broad, continental and national patterns of latitude and longitude (as discussed in Section 2). However, linear probes are limited in their ability to capture geographical relationships. Therefore, to better understand LLMs' internal representation of geographical information, we focused our study on a null hypothesis rooted in spatial analysis: if an LLM has encoded information about places in an a-spatial manner, the internal representations generated for a placename should not be correlated to the internal representations generated for neighbouring placenames. In our experiments, we use spatial autocorrelation to reject this null hypothesis and thus sustain our initial conjecture.

In this chapter, we establish a novel framework for the study of geospatial mechanistic interpretability, using spatial analysis [40] as a tool to explore the geographical aspects of the internal workings of LLMs. In turn, that might provide us with a deeper understanding of geographic bias [24] and diversity [25] in LLMs, their ability to reproduce geographical knowledge [1,22,23] and spatial reasoning [7], and ultimately contribute to AI safety [41]. The code and data necessary to replicate our experiments are available through our GitHub repository.[5]

2. Probing

2.1. Introduction to probing

In this chapter, we use the term *probing* to refer to approaches that extract and analyse the activations of an LLM while a prompt is being processed. However, we acknowledge that the term is also sometimes used more colloquially to refer to output-based evaluations or behavioural analyses that focus on the outputs of the models, such as the text produced as an answer to a question in the prompt [42].

Probing techniques have emerged as a crucial tool, providing insights into the internal workings of neural networks and, more specifically, LLMs. These techniques aim to unveil the latent knowledge embedded within the internal representations learned by these models. By designing tasks that test specific properties or capabilities of a model, probing enables researchers to assess how well the model encodes and uses various types of linguistic and conceptual knowledge [43,44,45].

Probing acts as a diagnostic tool to understand the inner workings of LLMs. Models are presented with controlled tasks that highlight specific functions or internal representations within their architecture. This approach has been instrumental in understanding how LLMs process language, capturing relationships that extend beyond surface-level patterns, where a feature is encapsulated by single neurons, which activate in relation to specific concepts or inputs. It helps researchers investigate not only what LLMs know but also how they process, store, and manipulate information internally.

[5] https://github.com/sdesabbata/geospatial-mechanistic-interpretability

A key aspect of probing techniques is their non-intrusive and architecture-agnostic nature. They typically leverage the model in its inference mode, thus without requiring any additional fine-tuning or training, in order to reveal the internal mechanisms that drive decision-making within the model [46,47]. One common form of LLM probing is through prompting, where carefully crafted inputs are designed to encourage the model to demonstrate its knowledge or capabilities in specific areas [48]. By using targeted prompts or inputs, probing techniques can assess a model's knowledge in various areas, such as language understanding, factual knowledge, or even hidden biases [49].

Probing techniques often employ targeted tasks that reveal how well the internal representations capture specific linguistic aspects or concepts. A typical probing process involves extracting the model's internal representations while the model is processing a specifically selected set of inputs, as illustrated in Figure 1. A lightweight model (e.g., a linear regression) is then trained to perform a specific task, such as estimating the latitude and longitude of the place mentioned in the input sentence based on the internal representation extracted from the model. If the lightweight model performs well on the task, it suggests that the LLM's internal representations contain the relevant information. For example, if a linear regression can reliably predict the latitude and longitude of a place based on the internal representations extracted from the LLM while it processes a sentence about that place, we can infer that the LLM has encoded some form of geographical awareness in its internal parameters. These types of probes provide insights into how well the LLM has learned to encode fundamental properties, thus offering explainability about both its strengths and limitations in handling basic linguistic properties.

More complex probes can be designed to assess deeper, more abstract properties within an LLM. For example, a probe could be used to determine whether an LLM captures the hierarchical structure of a sentence by training a classifier to distinguish between different syntactic structures, such as identifying clauses or phrases within a sentence [50]. Another example is probing whether an LLM understands semantic relationships between entities [51], such as distinguishing between "Paris is the capital of France" versus "France is the capital of Paris." In this case, the probe might involve evaluating how well an LLM can represent such relational information by training a classifier to predict entity relationships based on an LLM's internal representations. If the classifier can accurately distinguish between correct and incorrect relationships, it suggests that the LLM has encoded knowledge about semantic roles and relations. By probing for these high-level concepts, researchers can assess an LLM's ability to process and represent more complex aspects of language and knowledge beyond surface-level patterns.

2.2. Probing for geographical information

Probing techniques have been used to evaluate the extent to which the internal representations of LLMs, when prompted with geographical entities, can be correlated to information about those geographical entities. For example, Lietard et al. [36] examined how well the internal representations generated by BERT-like language models [52] for 3,527 cities and 249 countries and territories can be correlated to their geographical coordinates, population size and neighbouring countries. They found that larger models perform better at encoding geographical information. Gurnee and Tegmark [37] showed how a linear probing technique can be used to predict the geographical coordinates from internal representations within a certain geographical scale and accuracy, concluding

that LLMs can learn internal representations of geographical information which are at least partially linear. Godey et al. [38] applied the same methodology to a broader set of smaller models, showing how geographical information can be extracted from a wide range of model sizes and that performance is correlated to the amount of geographical information present in the training data. Chen et al. [39] further expanded on those approaches by introducing multilayer feed-forward neural networks as non-linear probes, aiming to better account for complex internal representations.

However, the probing techniques outlined above have so far focused on a-spatial statistical modelling, correlating the internal representations to factual information (e.g., population size or approximate location) rather than conducting a spatial analysis of the internal representations. By focusing on the prediction of geographical coordinates from the internal representations, these approaches can provide only a limited insight into the geographical nature of internal representations, and thus limit our ability to explore more complex spatial patterns which might emerge from the internal representations. As such, in this chapter, we explore the use of spatial analysis [40] to explore whether the internal representations of LLMs more broadly follow Tobler's "first law of geography" [35].

2.3. Spatial analysis of internal representations

In this section, we explore how spatial autocorrelation [53] can be used to identify interpretable geospatial patterns in internal representations extracted from LLMs. Our methodology involves using placenames alongside their geographical areas as prompts to generate internal representations, which are then correlated with those of nearby placenames using spatial autocorrelation to assess how well the LLM internalises geographical patterns, as illustrated in the bottom-centre component of Figure 1.

2.3.1. Methodology and data

In our experiments, we used three sets of placenames sourced from the GeoNames Free Gazetteer Data repository[6] representing three areas of broadly similar population sizes: the UK (about 67 million inhabitants), Italy (about 59 million inhabitants), and the four US states commonly associated with the broader New York metropolitan area (New York, New Jersey, Connecticut, and Pennsylvania, totalling about 45.5 million inhabitants). Previous studies [54] indicate that we can expect a similar high-quality coverage of placenames in those areas. The dataset included only populated places (labelled as `P` in GeoNames) with a recorded population above zero for the UK and the four US states and a population greater than 500 for Italy, resulting in a total of 6,294 placenames for the UK, 9,959 for Italy, and 4,090 for the four US states.

While placename completeness might not be a severe concern in this case, placename ambiguity can impact our experimental setup. Using placenames without further information would likely result in insufficient information for the LLM to correctly interpret the place being referred to due to geo-geo ambiguity (e.g., there is a place called Liverpool in the state of New York and a place called Verona in the state of New Jersey), geo-non-geo ambiguity (e.g., there are 77 places called San Lorenzo in Italy, and the name can also refer to a person from history) and metonymic usage (e.g., using the word "Westminster" to refer to the UK government). As such, we decided to input the placenames into the LLM, constructing prompts as follows:

[6]Data collected on July 18th, 2024, from `https://download.geonames.org/export/dump/`

- "`[placename], [country]`" for the UK, using the full names of the four countries of the UK (England, Scotland, Wales, and Northern Ireland) based on the value of `admin1 code` from GeoNames;
- "`[placename], [province]`" for Italy, using the full names of the Italian provinces based on the value of `admin2 code` from GeoNames;
- "`[placename], [state]`" for the four US states, using the full names of the states based on the value of `admin1 code` from GeoNames.

We query the LLM using the prompts as described above, capturing the internal representations from different layers of the model for each prompt. Specifically, we rely on `Mistral-7B-Instruct-v0.2`,[7] a 7-billion parameter model fine-tuned to follow instruction-based tasks [55,56], and we extract the 4,096 post-attention normalised activations of the transformer layers 7, 15, and 31 (counting from zero). These layers were selected to gain an early (token-level patterns), middle (contextual understanding), and late (high-level abstraction) internal representation from the model [32,37,57]. We then apply *mean pooling* [58], averaging the token-level activations across each prompt and layer to produce condensed activations. Thus, each placename is now associated with 4,096 condensed activations per layer, encompassing the internal representations that the model associates with that placename at different processing stages.

To assess whether the LLM's internal representations capture the geographical nature of the places referred to by the input prompts, we need to test whether these condensed activations have similar values for placenames referring to places which are near to one another in geographical space. We thus join the condensed activations generated by each placename with its geographical coordinates and calculate both the global and local Moran's I [59] as indicators of spatial autocorrelation [40,53]. The global Moran's I indicator is calculated separately for each one of the three areas to assess whether spatial autocorrelation is present in at least one of them, whereas the local clusters are identified by calculating the local Moran's I indicator on the three areas combined, to ensure that a cluster's significance is assessed with regard to all the activations extracted from a layer.

Based on the results presented by previous studies [37,38], we expect the values to display spatial autocorrelation with values changing across the main cardinal directions. As other authors have been able to use both prompting and linear probes to retrieve the latitude and longitude coordinates from internal representations, it seems probable that LLMs store that information in a manner that leads the values to change linearly as the location of the places changes across the cardinal directions[8]. Moreover, due to the inclusion of state or province names in the prompt, we expect to see at least some degree of similarity for values within the same area. However, if the LLM is capable of evoking more nuanced geographical information processing – or, to be more precise, if the training has produced weights that can link the input placenames to outputs referring to geographical concepts or areas at a different scale(s) – we should observe spatial autocorrelation with clusters of similar values outlining areas at a different scale compared to the state or province names used in the prompts, or the placenames themselves.

[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[8] It is important to note that LLMs handle numbers in the input query and output answer as textual tokens, which impacts their performance in answering mathematical questions [60], although recent developments have demonstrated huge improvements in this field through approaches focusing on test-time compute [61].

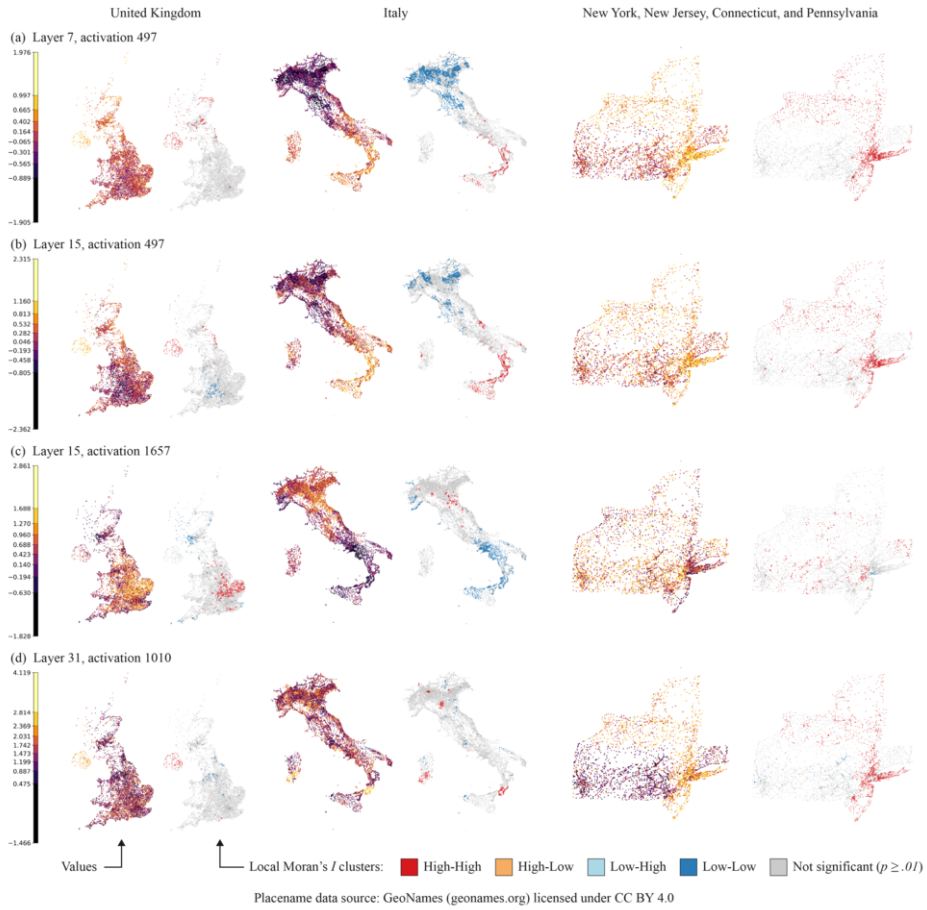

Figure 2. Activations captured for the input placenames at different layers of the LLM (left for each region) and their local spatial autocorrelation (local Moran's I clusters, $p < .01$, right for each region), illustrating the polysemantic nature of its internal representations. Two neurons at layers 7 (a) and 15 (b) show high values for the State of New York and Northern Ireland and very low values for northern Italy. A neuron at layer (c) 15 shows high values for several UK cities. A neuron at layer 31 (d) shows high values for the State of New York and Northern Ireland and diverse values for provinces in Italy.

2.3.2. Results

Our results illustrate a glimpse of all the aspects mentioned above. While discussing all the 1,841 (14.98%) neurons across layers 7, 15, and 31 that display a significant autocorrelation ($p < .01$ and Moran's $I \geq 0.3$) is beyond the scope of this chapter, some examples are reported in Figure 2.

For instance, Figure 2(a) illustrates the internal representations of the neuron with the second highest spatial autocorrelation at layer 7. Very high values highlight areas whose name was used as part of the prompts (e.g., "`New York`" as `[state]` in the input for placenames in the State of New York or "`Northern Ireland`" as `[country]` for placenames in Northern Ireland, as outlined in Section 2.3.1). At the same time, a clear pattern of increasing values can be seen by moving from the north of Italy to the south,

despite prompts not including explicit references to southern, central or northern Italy, only the name of the place and province. This indicates that the LLM might be encoding broader regions or geographical locations as part of the internal representation of place and province names, which could be linked to LLMs' ability to generate latitude and longitude values for input placenames [1,23]. A very similar pattern is also displayed by the neuron with the sixth-highest spatial autocorrelation at layer 15, as shown in Figure 2(b), which indicates some form of consistency across layers. However, the neuron with the second highest spatial autocorrelation at layer 15 might be introducing some level of complexity that goes beyond the linearity of cartographic coordinates or named areas. As shown in Figure 2(c), that particular neuron seems to reveal particularly high values in East Anglia and the East Midlands, with lower values in larger cities such as London, Birmingham, Liverpool and Manchester in the UK, as well as New York City and Philadelphia in the US. At the same time, the neuron also seems to display a gradual change from the north of Italy to the south, as in Figure 2(a) and (b), without a clear pattern related to major cities. Figure 2(b) and (c) illustrate how different neurons represent different information within the same layer. Interestingly, the final layer seems to focus back to local or broad-stroke geographies, as illustrated in Figure 2(d) by the neuron with the highest spatial autocorrelation at layer 31, which shows low values for Pennsylvania and high values for the State of New York, generally low values for Great Britain and high in Northern Ireland, and a patchwork of high and low values across Italy.

Importantly, the neurons showing high spatial autocorrelation do not seem to behave in a monosemantic (one-meaning) manner, where they would output high values for one area or concept only. Rather, they can behave in a polysemantic (multiple-meaning) manner, where the same neuron (same row in Figure 2) outputs high values for different areas or concepts. That seems to indicate that the superposition hypothesis (introduced below) might hold for geographical information. Therefore, in the next section, we explore a mechanistic interpretability approach aimed at addressing this limitation by decomposing these polysemantic structures into more interpretable, monosemantic features. In particular, we explore the use of sparse autoencoders to identify specific directions in the LLM's internal representation space that correspond to distinct geographical entities and concepts, allowing us to better understand how LLMs process geographical information.

3. Mechanistic interpretability

3.1. Introduction to mechanistic interpretability

Mechanistic interpretability seeks to understand how LLMs process and encode information by breaking down their internal representations into more interpretable features. A key concept in mechanistic interpretability is the *superposition hypothesis* [32,57,62], which suggests that neural networks often need to represent more features than they have neurons. This means multiple features can be encoded in overlapping or shared neurons, making it difficult to clearly interpret individual neurons. To address this, researchers employ methods like sparse autoencoders, which attempt to disentangle the model's internal representations into a set of more interpretable, sparse features [62]. The concept of *monosemanticity* instead refers to the idea that certain neurons in the network consistently correspond to a single interpretable concept.

In practice, many LLM neurons are polysemantic, meaning they can represent multiple, sometimes unrelated, concepts depending on the context. Scaling studies have shown that larger models tend to exhibit more monosemantic features, which improves interpretability as each feature is more likely to correspond to a unique concept [32].

Sparse autoencoders [63] play a crucial role in this process by decomposing the internal representations of models into sparse, interpretable directions of a multidimensional conceptual space. These directions correspond to distinct features, allowing researchers to better understand the relationship between models' inputs and outputs. In recent work, the use of sparse autoencoders has shown that even small transformer models, such as those with just a single MLP layer, can be decomposed into thousands of features, each capturing a different, often interpretable, property of the input data [62].

Overall, mechanistic interpretability provides a framework for understanding the inner workings of LLMs. By exploring notions and techniques like superposition, monosemanticity, and sparse autoencoders, researchers can gain deeper insights into how these models encode and manipulate complex features, ultimately improving our ability to interpret and control their behaviour.

3.2. Mechanistic interpretability via sparse autoencoders and spatial analysis

In this section, we describe our approach to extracting geospatially interpretable features using a sparse autoencoder [63] and spatial autocorrelation [40,53]. That is illustrated in Figure 1, where the features obtained from the sparse autoencoder in the bottom-right component are used as input for a spatial autocorrelation analysis in the bottom-centre component. The primary goal is to understand how LLMs encode geographical information and whether these internal representations display geospatial patterns.

3.2.1. Methodology and data

For the preliminary experiments presented in this chapter, we begin by training a sparse autoencoder on the condensed activations described in Section 2.3.1. The sparse autoencoder is designed to decompose the condensed activations into a set of sparse, interpretable features [32,62], allowing us to identify the directions in the LLM's internal representation space that correspond to specific geographical entities or concepts.

We followed an approach similar to the one proposed by Templeton et al. [32]. We focused on the middle layer (i.e., layer 15) and used its condensed activations to train a sparse autoencoder with a single encoding layer that expanded 4,096 input condensed activations to 32,768 embeddings (i.e., eight times the input size) and a single decoding layer to reduce them back to 4,096 reconstructed values. However, instead of relying on including an L1 penalty to encourage sparsity, we employed the TopK with ReLU activation function proposed by Gao et al. [64], which retains only the k largest embeddings, setting the rest to zero. We trained four models with k values of 1,024, 2,048, 4,096 and 8,192 (to test different levels of sparsity) over 300 epochs using the whole dataset (as we did not seek to build a generalisable model) in mini-batches of 32 cases in randomly shuffled order. Based on the final training loss, we selected the model with k equal to 2,048 as the best model. Finally, we used the encoder from this model to encode each 4,096 condensed activations into 32,768 features.

Once the features were extracted, we proceeded as in the previous experiment, linking back the features to the geographical coordinates and calculating the global and local Moran's I [59] indicators of spatial autocorrelation [40,53].

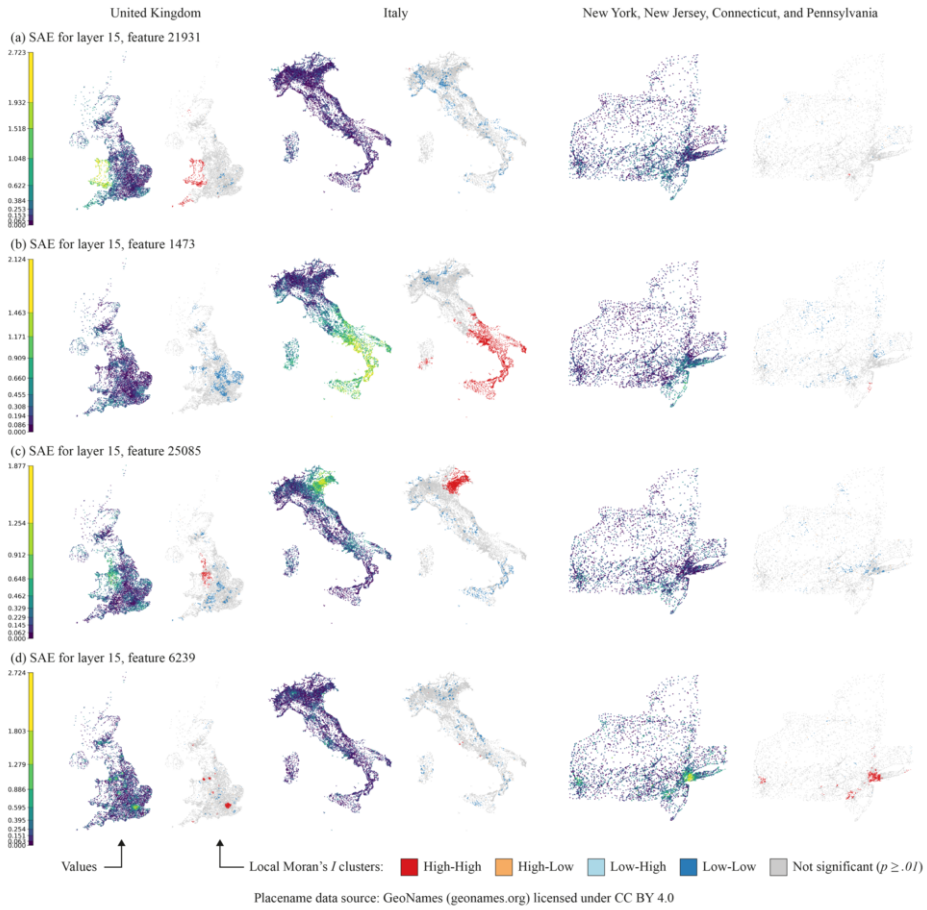

Figure 3. Features extracted from layer 15 through a sparse autoencoder (left for each region) and their local spatial autocorrelation (local Moran's I clusters, $p < .01$, right for each region): (a) Wales as a region part of prompt; (b) south of Italy as a region activating a seemingly monosemantic feature; (c) north-east of Italy and north-west of England as regions activating a seemingly polysemantic feature; and (d) a representation of "city" highlighting New York City and London, amongst others.

3.2.2. Results

Our preliminary results illustrate how the features extracted by the sparse autoencoder seem to display a clearer monosemantic and spatially coherent nature.

For instance, Figure 3(a) illustrates an example of a feature encoding a region that was part of the prompt, in this case, Wales. Figure 3(b) and (c) are, instead, examples of features displaying strong spatial autocorrelation and high values across areas that were not explicitly part of input prompts. While Italian provinces were used in the prompts for Italian placenames, the feature in Figure 3(b) encodes the whole of southern Italy, and the feature in Figure 3(c) encodes the north-east areas of the country. At the same time, while the feature illustrated in 3(b) seems to be strongly monosemantic – with very high values in the south of Italy and low values in the centre and north of Italy, as

well as across the UK and the US states included in the analysis – the feature illustrated in 3(c) displays high values for both the north-east of Italy (especially around Venice) and the north-west of England (especially around Liverpool). A connection might be drawn around the historical importance of Liverpool and Venice as ports and areas of industrial development. However, that would exclude other places with comparable port and industrial histories that are part of the maps but do not display high values, such as Genova. A clearer display of an important geographical concept (i.e., "urban") is instead illustrated in 3(d), where high values are visible for several large cities, including New York City, Philadelphia, Pittsburg, London, Manchester and Milano.

It is noteworthy that only 67 of the 32,768 features (0.2%) displayed a significant spatial autocorrelation ($p < .01$ and Moran's $I \geq 0.3$). This indicates that LLMs may encode some degree of geographical information, but this information is represented in a sparse and diffuse manner across the model's internal layers. The small number of spatially coherent features could be a result of the superposition hypothesis, where geographical concepts are entangled with other unrelated concepts, making them more difficult to isolate. However, the presence of features with significant spatial autocorrelation also suggests that LLMs are capable of capturing and representing geospatial patterns, although these patterns might only manifest in specific layers or under particular conditions. Further investigation into these features could reveal more about how LLMs generalise geographical knowledge and how they differentiate between regions based on learned associations. Moreover, 99.53% of features generated by the encoder component of the sparse autoencoder trained for our preliminary experiments always generate a zero value for all input placenames, leaving 0.27% of features displaying minimal or not significant spatial autocorrelation. That suggests that further work is necessary to improve the quality of the decomposition, as different choices in the extracted activations, the construction of the sparse autoencoder (e.g., size of the "bottleneck" or approach to sparsity) or training might provide better insights or that expanding the internal representations to such a large number of features might not be necessary when focusing on a specific subject or small datasets compared to broader studies [32].

4. Discussion and Conclusions

This chapter proposes a novel framework as a path towards the study of geospatial mechanistic interpretability of LLMs and, thus, an entry point into an array of new research question [65]. Our preliminary experiments provide insights into the internal representations that LLMs use to handle geographical information, leading us to two key findings.

First, the internal representations handling geographical information in LLMs seem to be often distributed across many polysemantic neurons rather than being monosemantically represented by a single neuron. That is, a geographical entity or concept may be encoded in a non-trivial manner, activating multiple neurons at the same time, and a single neuron can encode multiple, unrelated geographical entities or concepts – possibly alongside other information – which is consistent with the superposition hypothesis [57]. The nature of these internal representations makes it difficult to isolate and interpret individual geographical entities and concepts. This finding raised important questions about the geospatial mechanistic interpretability of LLMs and their ability to generalise geographical knowledge, which led us to explore the use of sparse autoencoders [63] and our second finding.

Second, the combined use of sparse autoencoders and spatial analysis seems capable of rendering the internal representations handling geographical information in LLMs interpretable. The use of sparse autoencoders allowed us to disentangle some of the polysemantic structures into more interpretable, monosemantic features. By applying spatial autocorrelation to these disentangled features, we observe clearer geospatial patterns, which highlight the potential of sparse autoencoders to improve the interpretability of LLMs for geospatial tasks.

At the same time, several geographical questions remain open, which will require further investigation: How does placename ambiguity impact LLMs' learning and interpretability? What scale(s) of geographical relationships and effects are encoded, and at which depth(s) of the models? How do geographical relationships compare to and interact with other types of relationships (e.g., social, economic, or historical)?

Our future work will explore those questions for LLMs and foundation models [1,66] more broadly. We aim to further develop our geospatial mechanistic interpretability framework, for instance, by exploring further methods in spatial analysis [40], the discovery of feature circuits [67,68] and different prompting and sparse autoencoder training strategies to better understand the internal representations activated in different contexts. Moreover, future work could explore how different training strategies, model architectures, and datasets influence the encoding of geographical information, as integrating external geographical knowledge bases or geographical embeddings could also improve the LLMs' ability to handle geographical entities, relationships and concepts. The study of the internal representations generated for different languages, alternative and vernacular placenames [69], and place nouns could also provide further insights into the internal geographies of LLMs.

Our approach holds the potential to substantially expand our understanding of how LLMs and foundation models handle geographical information. This is an essential step in developing robust and reliable tools based on foundation models for geography and possibly even developing the "artificial GIS analyst that passes a domain-specific Turing Test by 2030" imagined by Janowicz et al. [70], as Chen et al. [39] illustrated how the quality of the internal representations of geographical information can influenced an LLM's performance on geospatial tasks. Facilitated by the novel framework presented in this chapter, the study of geospatial mechanistic interpretability might lead to a better understanding of how LLMs internally relate geographical information to cultural, socio-economic, political, or environmental information, thus complementing output-based evaluations focusing on LLM bias [24] and diversity [25], as well as quality in reproducing geographical knowledge [1,22,23] and spatial reasoning [7], and thus contributing to AI safety [41].

Acknowledgments

The authors would like to thank Univ.-Prof. Dr. Krzysztof Janowicz, Dr. Rui Zhu and the anonymous reviewers for their valuable comments, which helped us shape this chapter. This research used the ALICE High Performance Computing Facility at the University of Leicester.

References

[1] Mai G, Huang W, Sun J, Song S, Mishra D, Liu N, et al. On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper). ACM Trans Spatial Algorithms Syst. 2024 Jul;10(2).

[2] Wang S, Hu T, Xiao H, Li Y, Zhang C, Ning H, et al. GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review. International Journal of Digital Earth. 2024;17(1):2353122.

[3] Hochmair HH, Juhász L, Kemp T. Correctness Comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for Spatial Tasks. Transactions in GIS. 2024 Aug.

[4] Xu L, Zhao S, Lin Q, Chen L, Luo Q, Wu S, et al.. Evaluating Large Language Models on Spatial Tasks: A Multi-Task Benchmarking Study; 2024. Available from: https://arxiv.org/abs/2408.14438.

[5] Cohn AG, Hernandez-Orallo J. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs; 2023. Available from: https://arxiv.org/abs/2304.11164.

[6] Cohn AG, Blackwell RE. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions. In: Adams B, Griffin AL, Scheider S, McKenzie G, editors. 16th International Conference on Spatial Information Theory (COSIT 2024). vol. 315 of Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik; 2024. p. 28:1-28:9.

[7] Li F, Hogg DC, Cohn AG. Advancing spatial reasoning in large language models: an in-depth evaluation and enhancement using the StepGame benchmark. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24. AAAI Press; 2024. .

[8] Hu X, Kersten J, Klan F, Farzana SM. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. International Journal of Geographical Information Science. 2024;0(0):1-28.

[9] Li Z, Ning H. Autonomous GIS: the next-generation AI-powered GIS. International Journal of Digital Earth. 2023;16(2):4668-86.

[10] Zhang Y, Wei C, He Z, Yu W. GeoGPT: An assistant for understanding and processing geospatial tasks. International Journal of Applied Earth Observation and Geoinformation. 2024;131:103976.

[11] Zhang Y, Wang Z, He Z, Li J, Mai G, Lin J, et al. BB-GeoGPT: A framework for learning a large language model for geographic information science. Information Processing & Management. 2024;61(5):103808.

[12] Zhang Y, He Z, Li J, Lin J, Guan Q, Yu W. MapGPT: an autonomous framework for mapping by integrating large language model and cartographic tools. Cartography and Geographic Information Science. 2024;0(0):1-27.

[13] Singh S, Fore M, Stamoulis D. GeoLLM-Engine: A Realistic Environment for Building Geospatial Copilots. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024. p. 585-94.

[14] Tan C, Cao Q, Li Y, Zhang J, Yang X, Zhao H, et al.. On the Promises and Challenges of Multimodal Foundation Models for Geographical, Environmental, Agricultural, and Urban Planning Applications; 2023. Available from: https://arxiv.org/abs/2312.17016.

[15] Zhu H, Zhang W, Huang N, Li B, Niu L, Fan Z, et al.. PlanGPT: Enhancing Urban Planning with Tailored Language Model and Efficient Retrieval; 2024. Available from: https://arxiv.org/abs/2402.19273.

[16] Hu Y, Mai G, Cundy C, Choi K, Lao N, Liu W, et al. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. International Journal of Geographical Information Science. 2023;37(11):2289-318.

[17] Roberts J, Lüddecke T, Sheikh R, Han K, Albanie S. Charting New Territories: Exploring the Geographic and Geospatial Capabilities of Multimodal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2024. p. 554-63.

[18] Feng J, Du Y, Liu T, Guo S, Lin Y, Li Y. CityGPT: Empowering Urban Spatial Cognition of Large Language Models; 2024. Available from: https://arxiv.org/abs/2406.13948.

[19] Fulman N, Memduhoğlu A, Zipf A. Distortions in Judged Spatial Relations in Large Language Models. The Professional Geographer. 2024;76(6):703-11.

[20] Feng S, Lyu H, Li F, Sun Z, Chen C. Where to move next: Zero-shot generalization of llms for next poi recommendation. In: 2024 IEEE Conference on Artificial Intelligence (CAI). IEEE; 2024. p. 1530-5.

[21] Ilyankou I, Lipani A, Cavazzi S, Gao X, Haworth J. Do Sentence Transformers Learn Quasi-Geospatial Concepts from General Text?; 2024. Available from: https://arxiv.org/abs/2404.04169.
[22] Roberts J, Lüddecke T, Das S, Han K, Albanie S. GPT4GEO: How a Language Model Sees the World's Geography; 2023. Available from: https://arxiv.org/abs/2306.00020.
[23] Bhandari P, Anastasopoulos A, Pfoser D. Are Large Language Models Geospatially Knowledgeable? In: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems. SIGSPATIAL '23. New York, NY, USA: Association for Computing Machinery; 2023. .
[24] Decoupes R, Interdonato R, Roche M, Teisseire M, Valentin S. Evaluation of Geographical Distortions in Language Models: A Crucial Step Towards Equitable Representations. arXiv preprint arXiv:240417401. 2024.
[25] Liu Z, Janowicz K, Currier K, Shi M. Measuring Geographic Diversity of Foundation Models with a Natural Language–based Geo-guessing Experiment on GPT-4. AGILE: GIScience Series. 2024;5:38.
[26] Liu Z, Currier K, Janowicz K. Making Geographic Space Explicit In Probing Multimodal Large Language Models For Cul-Tural Subjects. In: Global AI Cultures Workshop of ICLR 2024; 2024. .
[27] Berragan C, Singleton A, Calafiore A, Morley J. Mapping Great Britain's semantic footprints through a large language model analysis of Reddit comments. Computers, Environment and Urban Systems. 2024;110:102121.
[28] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436-44.
[29] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10); 2010. p. 807-14.
[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. .
[31] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(8):1798-828.
[32] Templeton A, Conerly T, Marcus J, Lindsey J, Bricken T, Chen B, et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Transformer Circuits Thread. 2024.
[33] Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. Backpropagation and the brain. Nature Reviews Neuroscience. 2020;21(6):335-46.
[34] Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, et al.. Fine-Tuning Language Models from Human Preferences; 2020. Available from: https://arxiv.org/abs/1909.08593.
[35] Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography. 1970;46:234-40.
[36] Liétard B, Abdou M, Søgaard A. Do Language Models Know the Way to Rome? arXiv preprint arXiv:210907971. 2021.
[37] Gurnee W, Tegmark M. Language Models Represent Space and Time. arXiv; 2024.
[38] Godey N, de la Clergerie É, Sagot B. On the Scaling Laws of Geographical Representation in Language Models. arXiv; 2024.
[39] Chen Y, Gan Y, Li S, Yao L, Zhao X. More than Correlation: Do Large Language Models Learn Causal Representations of Space?; 2023. Available from: https://arxiv.org/abs/2312.16257.
[40] O'Sullivan D, Unwin D. Geographic information analysis. John Wiley & Sons; 2010.
[41] Bereska L, Gavves E. Mechanistic Interpretability for AI Safety – A Review; 2024. Available from: https://arxiv.org/abs/2404.14082.
[42] Hobbhahn M, Lieberum T, Seiler D. Investigating causal understanding in LLMs. In: NeurIPS ML Safety Workshop; 2022. .
[43] Wallace E, Wang Y, Li S, Singh S, Gardner M. Do NLP models know numbers? probing numeracy in embeddings. arXiv preprint arXiv:190907940. 2019.
[44] Kim N, Patel R, Poliak A, Wang A, Xia P, McCoy RT, et al. Probing what different NLP tasks teach machines about function word comprehension. arXiv preprint arXiv:190411544. 2019.
[45] Belinkov Y. Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics. 2022;48(1):207-19.
[46] Koto F, Lau JH, Baldwin T. Discourse probing of pretrained language models. arXiv preprint arXiv:210405882. 2021.
[47] Arps D, Samih Y, Kallmeyer L, Sajjad H. Probing for constituency structure in neural language models. arXiv preprint arXiv:220406201. 2022.
[48] Feng S, Park CY, Liu Y, Tsvetkov Y. From Pretraining Data to Language Models to Downstream Tasks:

Tracking the Trails of Political Biases Leading to Unfair NLP Models. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics; 2023. p. 11737-62.

[49] Vulić I, Ponti EM, Litschko R, Glavaš G, Korhonen A. Probing pretrained language models for lexical semantics. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020. p. 7222-40.

[50] Alleman M, Mamou J, Del Rio MA, Tang H, Kim Y, Chung S. Syntactic Perturbations Reveal Representational Correlates of Hierarchical Phrase Structure in Pretrained Language Models; 2021. Available from: https://arxiv.org/abs/2104.07578.

[51] Chanin D, Hunter A, Camburu OM. Identifying Linear Relational Concepts in Large Language Models; 2024. Available from: https://arxiv.org/abs/2311.08968.

[52] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T, editors. 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). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86. Available from: https://aclanthology.org/N19-1423/.

[53] Getis A. Spatial autocorrelation. In: Handbook of applied spatial analysis: Software tools, methods and applications. Springer; 2009. p. 255-78.

[54] Acheson E, De Sabbata S, Purves RS. A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. Computers, Environment and Urban Systems. 2017;64:309-20.

[55] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv preprint arXiv:231006825. 2023.

[56] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems. 2022;35:27730-44.

[57] Elhage N, Hume T, Olsson C, Schiefer N, Henighan T, Kravec S, et al.. Toy Models of Superposition; 2022. Available from: https://arxiv.org/abs/2209.10652.

[58] Gholamalinezhad H, Khosravi H. Pooling methods in deep neural networks, a review. arXiv preprint arXiv:200907485. 2020.

[59] Moran PAP. Notes on Continuous Stochastic Phenomena. Biometrika. 1950;37(1/2):17-23.

[60] Singh AK, Strouse D. Tokenization counts: the impact of tokenization on arithmetic in frontier LLMs; 2024. Available from: https://arxiv.org/abs/2402.14903.

[61] Snell C, Lee J, Xu K, Kumar A. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters; 2024. Available from: https://arxiv.org/abs/2408.03314.

[62] Bricken T, Templeton A, Batson J, Chen B, Jermyn A, Conerly T, et al. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread. 2023. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[63] Ng A, et al. Sparse autoencoder. CS294A Lecture notes. 2011;72(2011):1-19.

[64] Gao L, la Tour TD, Tillman H, Goh G, Troll R, Radford A, et al.. Scaling and evaluating sparse autoencoders; 2024. Available from: https://arxiv.org/abs/2406.04093.

[65] Sharkey L, Chughtai B, Batson J, Lindsey J, Wu J, Bushnaq L, et al.. Open Problems in Mechanistic Interpretability; 2025. Available from: https://arxiv.org/abs/2501.16496.

[66] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al.. On the Opportunities and Risks of Foundation Models; 2022. Available from: https://arxiv.org/abs/2108.07258.

[67] Wang K, Variengien A, Conmy A, Shlegeris B, Steinhardt J. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small; 2022. Available from: https://arxiv.org/abs/2211.00593.

[68] Marks S, Rager C, Michaud EJ, Belinkov Y, Bau D, Mueller A. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models; 2024. Available from: https://arxiv.org/abs/2403.19647.

[69] Jones CB, Purves RS, Clough PD, Joho H. Modelling vague places with knowledge from the Web. International Journal of Geographical Information Science. 2008;22(10):1045-65.

[70] Janowicz K, Gao S, McKenzie G, Hu Y, Bhaduri B. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science. 2020;34(4):625-36.