

Can Sparse AutoEncoder-Based Interpretability Demonstrate Geographic Knowledge Beyond Orthographic Features?

Thomas Hitchon^{1,3}, Rémy Decoupes^{1,3}, Adrien Guille²

¹TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE

²Université Lumière Lyon 2, ERIC UR 3083

³INRAE

Abstract

Large language models (LLMs) encode a wide range of information about the world, but understanding how such knowledge is represented remains a central challenge in mechanistic interpretability. In this work, we reproduce and extend a recent study investigating the representation of geographic information in LLMs using sparse autoencoders (SAEs). We focus on this work because, to the best of our knowledge, it is the only study that explicitly examines geographic representations in LLMs through the lens of feature-based interpretability. While the original study identifies SAE features that appear to encode geographic concepts, we extend the analysis to a broader set of locations spanning the globe and investigate the robustness of these findings. In particular, we hypothesize that some of the identified features may not represent geographic concepts directly, but may instead be driven by lexical properties of placenames, such as highly discriminative character sequences or syllables. To test this hypothesis, we introduce systematic perturbations of placenames that preserve their geographic reference while altering their orthographic form, and examine the resulting changes in SAE activation patterns and geographic activation maps. Our results provide a broader assessment of the reproducibility and generalizability of geographic representations identified by the original study, while highlighting potential confounds that should be considered when interpreting SAE features as semantic or conceptual representations. We position this work within the Generalizability and Baselines themes of the Reproducibility Challenge: we test whether the original findings hold beyond their initial experimental scope, and we introduce a simple lexical baseline that recovers most of the reported geographic signal from surface form alone, without access to the model’s internal representations.

1 Introduction

The extent to which large language models (LLMs) acquire and encode geographic knowledge has attracted growing attention. Probing studies have shown that geographic coordinates can be roughly decoded (via linear regression) from the internal representations LLMs generate when processing placenames (Gurnee and Tegmark, 2024; Godey et al., 2024; Decoupes and Guille, 2026), suggesting that these models construct some form of internal “world map”. Moving beyond coordinate prediction, De Sabbata et al. (2026) recently proposed a framework for the *geospatial mechanistic interpretability* of LLMs, in which spatial autocorrelation is used to characterize internal representations geographically. Training a sparse autoencoder (SAE) on activations of an LLM elicited by placename prompts, they identified sparse features displaying strong, seemingly monosemantic spatial patterns and interpreted them as evidence that geographic structure is encoded within the model.

In this paper, we ask a question that this line of work leaves open: *do such observations reveal geographic knowledge, or merely a correlation between language and geography?* Within a given region, placenames tend to share orthographic and morphological regularities (e.g. the English *-shire* or *-borough*) or the distinctive phonotactics of Italian toponyms (e.g. word-final vowels). A feature in the SAE that responds to the surface form of a name will therefore produce spatially coherent activation maps within a localized study area, making it observationally indistinguishable from a genuinely geographic feature. Two aspects of the original experimental setup make this confound difficult to rule out. First, the prompts proposed by De Sabbata et al. embedded the name of the administrative region alongside each placename (“*[placename], [country]*”), injecting explicit geographic information into the very input whose representation is

analyzed. Second, their analysis was restricted to three study areas: the United Kingdom, Italy, and a group of north-eastern US states.

In line with the goals of the Reproducibility Challenge, we reproduce the SAE-based approach of De Sabbata et al. (2026) and test how far its conclusions generalize. Our setup departs from the original in four ways. First, we improve the training of the sparse autoencoder by applying more recent SAE training practices known to reduce the number of dead features and stabilise training (Gao et al., 2025). Second, we remove the region from the input prompt, so that the model only receives the bare toponym. Third, following prior work (Bricken et al., 2023), we extract post-MLP activations rather than the post-attention, pre-MLP activations used in the original study. Post-MLP activations capture features computed by the MLP and provide a suitable space for sparse decomposition. Finally, and most importantly, instead of focusing on three areas, we conduct the analysis at a global scale.

Our results underline how consequential this change of scale is. We show that a localised study can suggest the existence of a geographic feature which, when the analysis is extended to the entire planet, instead emerges as a linguistic one. We corroborate these observations with several experiments, showing that superficial corruptions of the toponyms’ surface form are sufficient to make a feature disappear entirely.

Taken together, our findings suggest that current SAE-based interpretability methods do not provide sufficient evidence to conclude that LLMs encode geographic knowledge in features that are independent of orthographic patterns.

The code used in the experiments is available at <https://github.com/ThomasHtchn/sae-geographic-orthographic-confound/>.

2 Original study setup

2.1 Data and SAE architecture

De Sabbata et al. trained a sparse autoencoder (SAE) on geographic prompts representing three areas: the United Kingdom, Italy, and four US states around New York. Prompts follow the template “[*placename*], [*region*]”, where placenames are GeoNames populated places (*feature class P*) above a population threshold (population > 0 for the UK and the US, > 500 for Italy). The region qualifier is the country for the UK, the province

for Italy, and the state for the US, added to reduce placename ambiguity.

They extracted post-attention, pre-MLP activations from layers 7, 15, and 31 of Mistral-7B-Instruct-v0.2 for each prompt, then mean-pooled across tokens to obtain one 4,096-dimensional condensed activation vector per prompt per layer.

For the SAE experiment, they focused on the layer-15 (middle-layer) condensed activations. The architecture follows a TopK-with-ReLU design (Gao et al., 2025), where only the *k* largest encoder outputs are kept nonzero, expanding the 4,096-dimensional input to 32,768 sparse dimensions (features). Despite TopK already enforcing sparsity structurally, the released training code also adds a normalized L1 penalty on the latent activations (weight 0.01) on top of the (normalized) reconstruction loss. They trained for 300 epochs, in shuffled mini-batches of 32 and selected *k* = 2,048 as the final model based on training loss.

2.2 Geographic visualizations

To highlight geographic features discovered by the SAE, they used two visualizations for each neuron or feature of interest. First, they link each activation back to its placename’s geographic coordinates, plotting one colored dot per placename; since every populated place in the dataset is plotted, the dots trace each country’s outline, while color encodes the activation level. Second, they compute local Moran’s I (Moran, 1950), a spatial autocorrelation indicator, which assigns each place to one of five categories: High-High (place and neighbors both high), Low-Low (both low), High-Low (place high, neighbors low), Low-High (place low, neighbors high), or not significant. This local-cluster shows not just where values are high but where they’re significantly clustered.

2.3 Their conclusions

After training their SAE, the authors report a very high proportion of dead (never activate on any prompt) features (99.53%). They highlight features with significant spatial autocorrelation as evidence that LLMs are capable of capturing and representing geospatial patterns, while cautioning that such patterns may only manifest in specific layers or under particular conditions. They suggest the low amount of spatially coherent features may itself be a symptom of superposition. They close by expressing optimism about the method, proposing that further investigation of these features could clarify

how LLMs generalise geographic knowledge and differentiate between regions based on learned associations, while also acknowledging open questions about the approach itself: whether better results would come from different choices in activation extraction, SAE architecture, or training.

3 Model and data

Our study follows the experimental framework of De Sabbata et al. (2026), while modifying both the language model and the geographic coverage of the data.

3.1 Model

While De Sabbata et al. (2026) use Mistral-7B-Instruct-v0.2, we evaluate SmolLM3¹, an open-source language model for which training data are available. This enables us to examine the geographic distribution of place mentions in a sample from one of its training corpora. The Figure 1 illustrates the uneven geographic distribution of training data, which may account for corresponding biases in the model’s internal geographic representations (Decoupes et al., 2025).

3.2 Data

As in De Sabbata et al. (2026), we use data from GeoNames². However, instead of focusing on three geographic areas, the United Kingdom, Italy, and the broader New York metropolitan area (New York, New Jersey, Connecticut, and Pennsylvania), we extend the analysis to a global scale. We retain populated places with more than 10,000 inhabitants, resulting in a dataset of 44475 city names.

Our analysis focuses on the *name* field of the GeoNames dataset. Although GeoNames guidelines specify that this field should contain a widely accepted international or English name, naming conventions vary across countries. The field may therefore contain English exonyms, local names (endonyms), or transliterated names. In particular, endonyms may contain language-specific characters not commonly used in English, while transliterations represent names originally written in non-Latin scripts using the Latin alphabet. We retain the names as provided by GeoNames without further normalisation.

¹<https://huggingface.co/blog/smollm3>

²<https://download.geonames.org/export/dump/allCountries.zip>

4 Experiments

4.1 SAE-based experiments

Following prior work (Bricken et al., 2023), we extracted the post-MLP activations, as they capture features computed by the MLP and provide a suitable space for sparse decomposition. We used the same mean-pooling procedure of De Sabbata et al..

We used the same TopK SAE architecture with a few changes. For the loss, we only kept the reconstruction loss, dropping the L1 sparsity term that the original authors kept in their released code alongside TopK, since sparsity is already enforced structurally by TopK. We added decoder normalization, which Gao et al. showed improves reconstruction MSE. To select the hyperparameters k (number of top dimensions kept) and n (size of the learned sparse latent space), instead of only looking at the loss, we mainly looked at the combination of two metrics proposed by Bricken et al. to evaluate SAEs: the proportion of dead features (features that are never active on any prompt) and the mean density (proportion of prompts for which a given feature is active) of non-dead features. We trained SAEs for 500 epochs in shuffled batches of 256, with early stopping (patience 20) and an adaptive learning rate, and found the best results for $k = 256$ and $n = 16,384$.

We used a range of SmolLM3-3B layers: early layers 2 to 6, middle layers 11 to 17, and late layers 29 to 34. Our observations suggest that very early and late layers did not produce geographically interpretable features.

We have manually found interesting geographic features (Figure 2). Compared to De Sabbata et al., the change in scale (world prompts) enriches these features. For example, we found a feature with strong activations in Wales (Figure 2a), but not only there: the same feature activates strongly in east coast Australia, corresponding to the New South Wales area, better representing Wales’ sphere of influence.

Indeed, some features are very broad across the world, for example, layer 17 feature 924 (Figure 2b) activates strongly across the Francophone world, while others are much more localised to a certain area of the world. For instance, layer 31 feature 424 activations are grouped in Japan (Figure 3a). Other features fall at an intermediate scale, covering a mid-sized region rather than a single country or the whole world, such as layer 17 feature 10630, which activates strongly in Eastern Europe (Figure 3c).

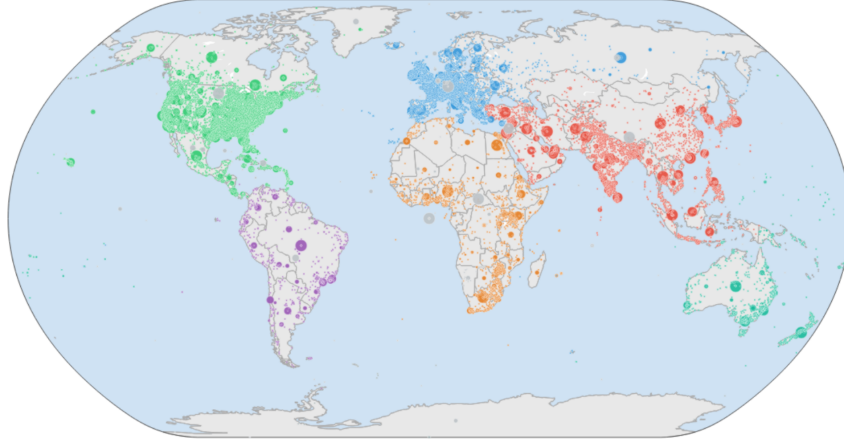


Figure 1: Geographic coverage of place mentions extracted from a sample of the FineWeb corpus (40000 documents) used to train SmolLM3. Each point corresponds to a populated place identified in the sample, with dot size indicating mention frequency.

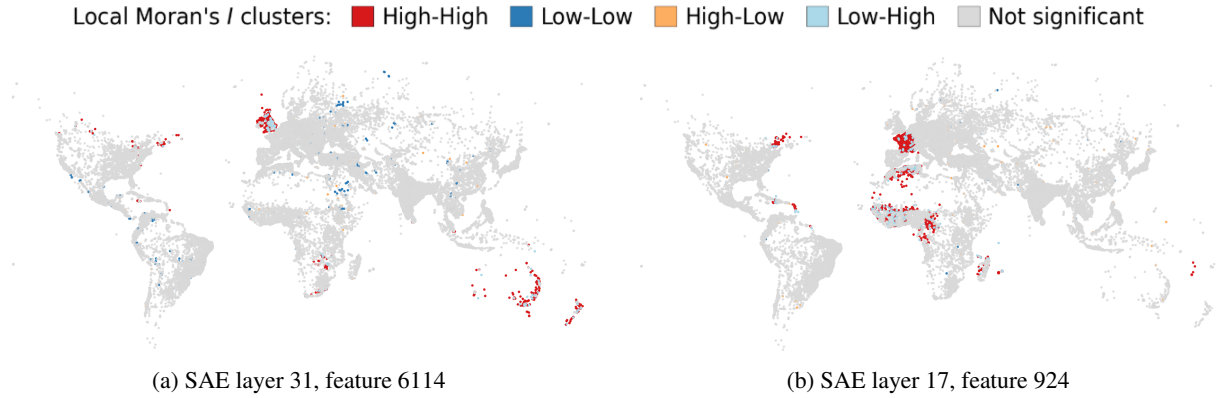


Figure 2: Spatial autocorrelation maps (local Moran’s I clusters, $p < 0.01$)

Exploring the features, we hypothesize that they seem, for the most part, to be based on discriminative orthographic patterns.

The SAEs we train may be sensitive to orthographic characters, for instance, the Eastern Europe feature shares specific characters, signs, or accents, such as the caron³ diacritical mark, the ogonek⁴, the acute accent (*e.g.* *ń*, *ó*, *í*, *ś*), and the *l* with stroke (*ł*). We introduced small corruptions of the data to verify this hypothesis. First, the layer 31 feature 424 (Figure 3a), representing Japan activates for placenames with an O macron (*i.e.* *Ō*). We replaced *Ō* by *O* in placenames. Note that *Ō* has a dedicated token in the vocabulary of SmolLM3. Not a single place previously starting with *Ō* kept activating for the feature (Figure 3b); for example, *Ōshima* (activation 1.36) dropped to 0.00 as *Oshima*, *Ōmorinaka* (activation 0.79) dropped to 0.00 as *Omorinaka*, and *Ōizumi-gakuenchō* (acti-

vation 0.45) dropped to 0.00 as *Oizumi-gakuencho*. This feature did not learn a geographic concept, but rather a single character variation.

Secondly, we re-extracted the activations with the ASCII version of the placenames (removing all special characters) to see if the Eastern European feature would be totally lost. We observed (Figure 3d) that some of the spatial activations visually remained, this SAE neuron appears to be more robust to this type of perturbation, and may indeed be related to a genuine geographic feature. However, other patterns beyond non-ASCII characters in the orthography of Slavic toponyms seem highly discriminative to us, such as *cz*, *dz*, *sz*, etc.

4.2 Additional experiments

Experiment 1 To further reinforce the confound introduced by the selection of the three sub-areas in the original study, we try to assess how much geographic information can be recovered from the surface form of the placenames alone, before any

³<https://en.wikipedia.org/wiki/Caron>

⁴<https://en.wikipedia.org/wiki/Ogonek>

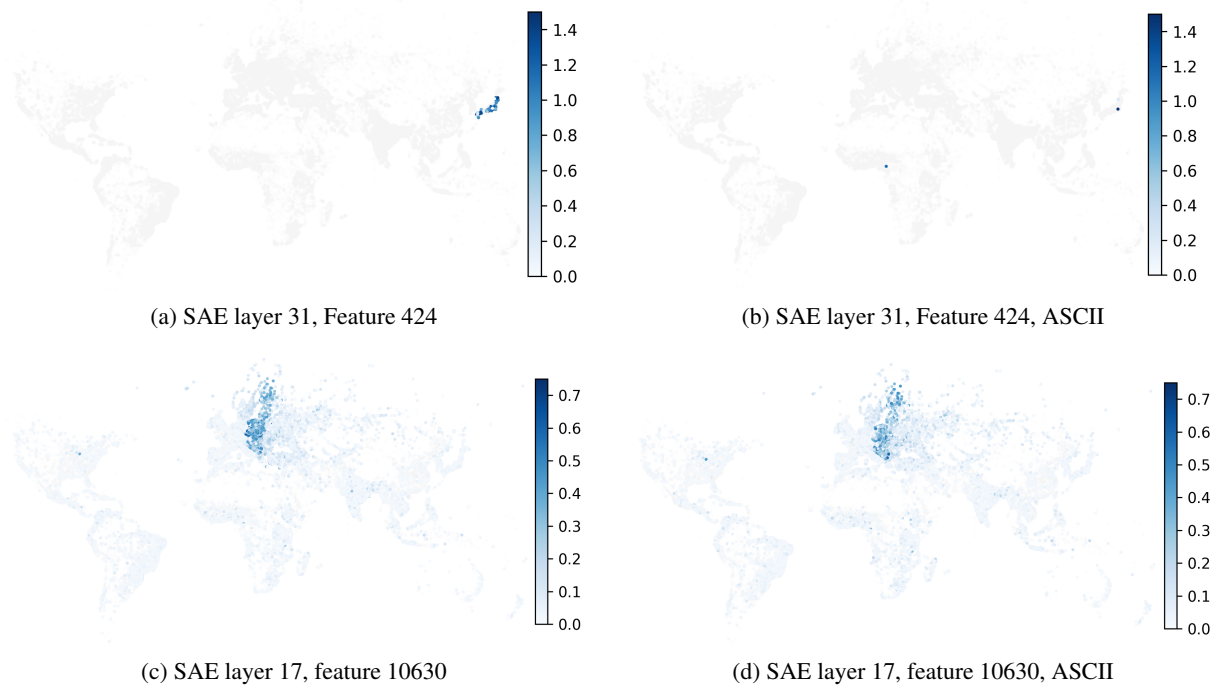


Figure 3: Feature activation maps.

contextual processing by the LLM. To this end, we represent each placename as a bag of tokens over the LLM’s tokenizer vocabulary to train and cross-validate an L1-regularised logistic regression to predict the area in which the corresponding place is located. We measure an overall F1-score of 87%, with 96% for Italy, 84% for UK, and 70% for cities located in the four US states. In other words, the tokenizer’s segmentation of a placename by itself carries information that correlates strongly with geography: the orthographic characteristics of toponyms differ sufficiently across the three areas for them to be discriminated without any access to the model’s internal representations, let alone to any geographic knowledge. The per-area scores seem to follow the linguistic homogeneity of the areas’ toponymy. It is very high for Italian names, whose orthography is highly distinctive, and lowest for the US states, whose placenames have various origins and tend to share naming conventions with the UK. We also note that the L1 penalty retains only 1,923 of the 128,256 tokens in the vocabulary, fewer than 1.5% of the subword units, meaning that a compact set of surface cues (*e.g.* characteristic prefixes, suffixes, syllables) are sufficient to reach a good performance in this setting. This simple baseline shows that, within the original three-area setup, a feature need only be sensitive to subword composition to produce spatially coherent activation maps,

making surface-form sensitivity observationally indistinguishable from geographic knowledge.

Experiment 2 In reaction to the prior observation regarding Eastern Europe, we run a variant of the previous experiment, with Poland as a fourth area. The classifier maintains an overall F1-score of 87%, reaching an F1-score of 93% for Poland. Tokens with the highest weights for Poland are “cz” (5.10), “sz” (4.79), “dz” (3.53), “ek” (3.52), “ej” (2.50), “rz” (2.29), “ice” (2.17), “oj” (2.17), “ka” (2.15), and “ko” (2.14). They are plain-ASCII character sequences, corresponding to the digraphs (cz, sz, rz, dz) and suffixes (-ek, -ice, -ka, -ko) typical of Polish orthography. This could explain why the activations depicted in Figure 3d are robust to ASCII transliteration, which suggests that this feature could also be mostly linguistic.

5 Limitations

This work has several limitations. First, the model under investigation is a relatively small 3B-parameter model. Consequently, our findings may not generalise to larger or substantially larger language models, whose representations and geographic knowledge may differ considerably.

Second, despite the optimisations introduced during the training of the Sparse Autoencoders (SAEs), a large proportion of SAE neurons remain dead. As a result, only a limited number of features can be

meaningfully interpreted.

Third, our analysis still involves a high degree of manual feature selection, introducing a potential cherry-picking bias. In the absence of an automated method for detecting human-interpretable geographic concepts in SAE features, we manually selected features whose activation patterns appeared visually related to geographic concepts of interest. It is possible that geographic representations are encoded at finer spatial scales and that several such features would need to be combined to recover higher-level, human-interpretable geographic concepts (Klotz et al., 2026).

Following De Sabbata et al. (2026), we combine global activation maps with Moran’s spatial autocorrelation index to identify geographically meaningful features. This is motivated by the observation that SAE features, when considered solely through their global activation maps, can remain highly polysemantic, with a single feature activating for multiple locations across the world without an immediately apparent common geographic pattern. Spatial autocorrelation therefore provides an additional criterion for identifying geographically coherent features. However, this approach may also overlook geographic concepts that are not strongly spatially autocorrelated or that are distributed across multiple features.

Finally, our study does not constitute an exact replication of the original experiments, as we use SmolLM3-3B rather than Mistral-7B-Instruct-v0.2 and extract activations at a different point in the residual stream. These choices allow us to conduct the global analysis under our experimental constraints, but they also introduce potential model specific effects. Our conclusions should therefore be understood as a robustness test of the interpretive methodology rather than a strict replication of every original result.

6 Future work

A key direction for future work is the development of interpretability methods, whether based on SAEs or alternative approaches, that can provide convincing evidence of genuinely geographic representations, rather than features driven by the surface form or orthography of placenames. A step toward this goal is the automatic discovery of candidate geographic features spanning multiple dimensions, including geopolitical groupings (e.g., EU, NATO, BRICS), natural regions (e.g., moun-

tains, rivers, deserts, climate), cultural characteristics (e.g., religion, colonial heritage), and socioeconomic profiles (e.g., income level or resource-based economies). However, an important challenge is that the dimensions learned by SAEs may not necessarily correspond to human-interpretable concepts at the same level of abstraction. For example, the human concept of a “coastal city” may not be encoded by a single SAE feature, but rather emerge from a combination of several latent dimensions, each capturing a more specific aspect of the concept. This raises the broader question of the appropriate granularity and semantic scope of SAE features, and highlights the need for further evaluation of whether SAE representations align with the structure of human concepts. Developing methods that can identify and combine multiple latent dimensions into coherent, human-interpretable concepts would therefore be an important step toward a more comprehensive understanding of geographic representations in LLMs.

Such a broader and more systematic set of geographic concepts would enable a more rigorous assessment of whether LLMs encode abstract geographic knowledge beyond lexical regularities. Finally, our study focuses exclusively on city names as a proxy for geographic knowledge. Future work should extend this analysis to other forms of geographic reasoning, such as spatial relations and distances, as well as quantitative geographic indicators (e.g., GDP or income per capita), to better characterize the nature and scope of geographic representations in LLMs.

Conclusion

We revisited the SAE-based geospatial interpretability framework of Sabbata et al. (2026) and tested how well its central claim, that LLMs encode geographic knowledge in identifiable sparse features, holds when the analysis is extended from three localized regions to a global scale. Our results show that this change of scope is far from neutral: features that appear spatially coherent within a single country or continent frequently turn out, once evaluated worldwide, to track orthographic or phonotactic regularities of placenames rather than genuine geographic concepts.

Targeted corruption experiments confirmed this directly, as removing diacritics or substituting a single character was sufficient to collapse activation patterns that had been interpreted as ev-

idence of spatial knowledge, and a simple bag-of-tokens baseline recovered most of the region-discriminating signal without any access to the model’s internal representations. Taken together, these findings do not invalidate the search for geographic representations in LLMs, but they show that current SAE-based evidence for such representations may be confounded by surface form, and that claims of this kind would benefit from being tested at global scale and compared against strong lexical baselines before being attributed with confidence to conceptual, geography-specific knowledge.

Our results suggest several practical safeguards for the use of feature-based interpretability in interdisciplinary applications. First, candidate semantic features should be compared against simple baselines built from surface-level properties of the input, such as character, subword, or token statistics. Second, interpretations should be stress-tested using controlled perturbations that preserve the underlying concept while altering potentially confounding cues.

Acknowledgments

This study was partially funded by the European Union under grant agreement No. 101287260 (GEOAI4EI). The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. This work was performed using HPC resources from GENCI-IDRIS (Grant 2026-AD011015649R1).

References

- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Stef De Sabbata, Stefano Mizzaro, and Kevin Roitero. 2026. Geospatial mechanistic interpretability of large language models. In *Geography According to Foundation Models*.
- R  my Decoupes and Adrien Guille. 2026. Determinants of geographic information quality in large language models: Effects of model family, scale, language, quantization, and fine-tuning. In *Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026*.
- R  my Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. 2025. [Evaluation of geographical distortions in language models](#). *Machine Learning*, 114(12):263.
- Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and Evaluating Sparse Autoencoders. In *Thirteenth International Conference on Learning Representations*.
- Nathan Godey,   ric de la Clergerie, and Beno  t Sagot. 2024. On the scaling laws of geographical representation in language models. In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.
- Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *Twelfth International Conference on Learning Representations*.
- Jonas Klotz, Cassio F. Dantas, Pallavi Jain, Diego Marcos, and Beg  m Demir. 2026. [Evaluating the interpretability of sparse autoencoders with concept annotations](#). *Preprint*, arXiv:2606.24716.
- P. A. P. Moran. 1950. [Notes on Continuous Stochastic Phenomena](#). *Biometrika*, 37(1/2):17–23.