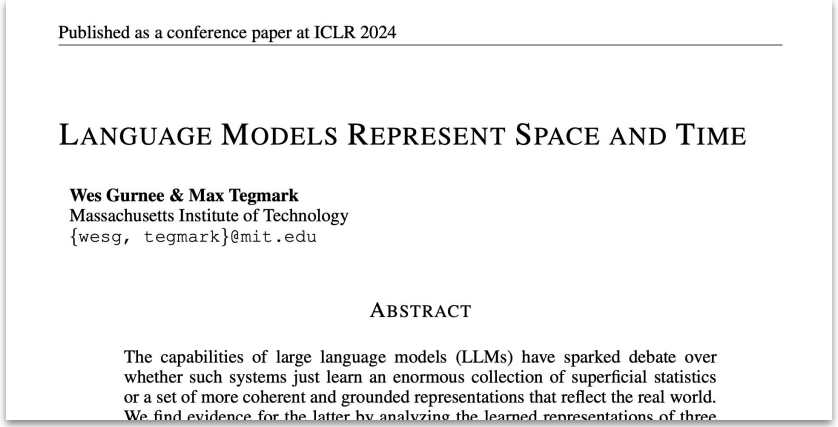
On the Scaling Laws of Geographical Representation in Language Models

Nathan Godey, Éric de la Clergerie, Benoît Sagot
LREC-COLING 2024, Torino

Background

Related works

- Geographical probing



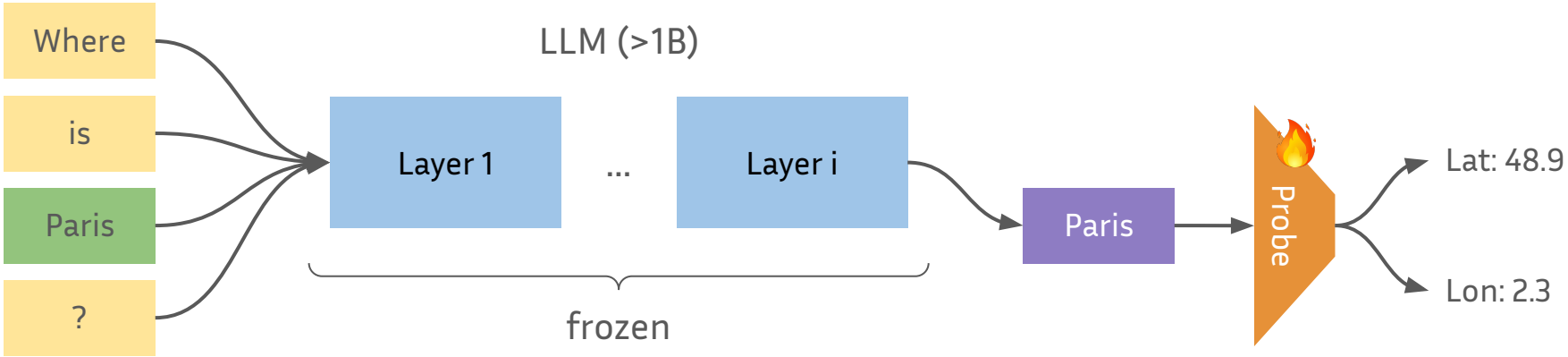
Published as a conference paper at ICLR 2024

LANGUAGE MODELS REPRESENT SPACE AND TIME

Wes Gurnee & Max Tegmark
Massachusetts Institute of Technology
{wesg, tegmark}@mit.edu

ABSTRACT

The capabilities of large language models (LLMs) have sparked debate over whether such systems just learn an enormous collection of superficial statistics or a set of more coherent and grounded representations that reflect the real world. We find evidence for the latter by analyzing the learned representations of three



Related works

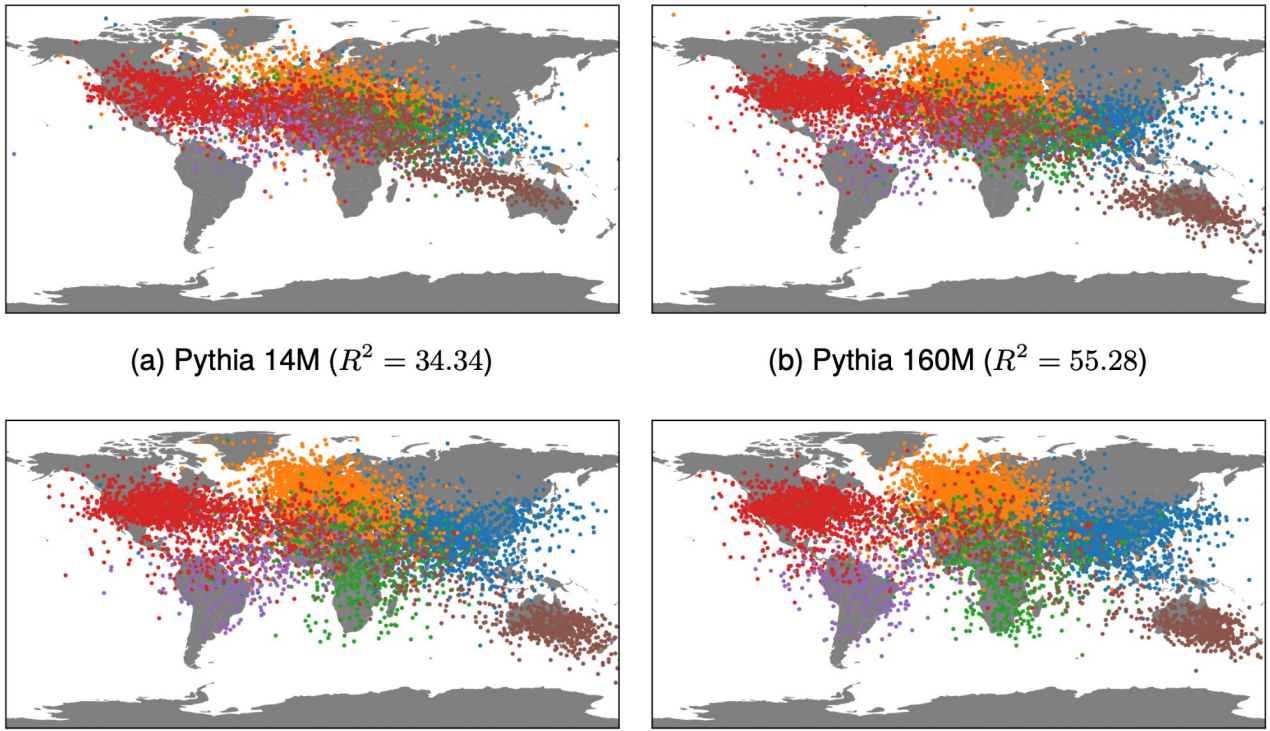
- Long history of work on probing of language models/embeddings
- Several articles on *geographical* probing, among which:
  - Louwerse and Benesh (2012) : co-occurence matrices on LOTR
  - Faisal et al. (2022) build graphs from “expert” neurons
    - GPT2-medium (Radford et al., 2019)
    - mGPT (Shliazhko et al., 2022)
    - BLOOM-560m

Motivations

- Bridge the gap between older and novel work
  - Extend LLM observations to smaller models
- Observe the geographical performance of model suites when scaling in model size
- Study the geographical biases and their correlation to training data & other factors

Results

Performance



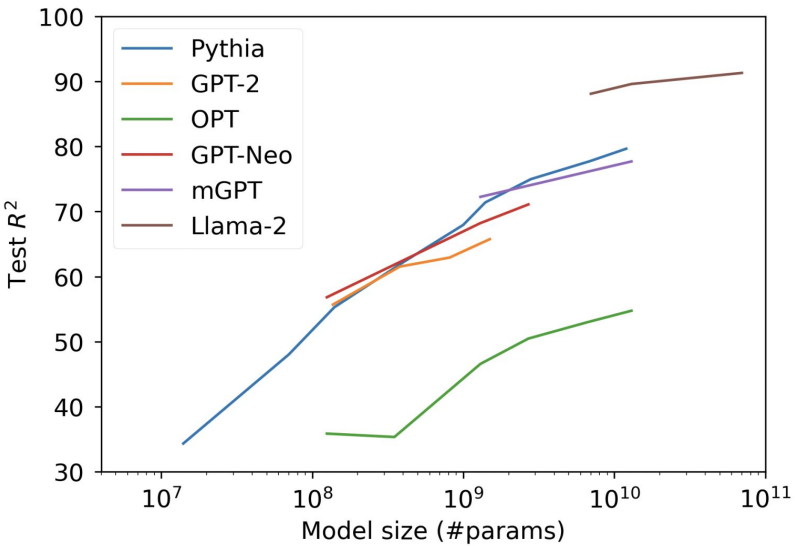
(a) Pythia 14M ($R^2 = 34.34$)

(b) Pythia 160M ($R^2 = 55.28$)

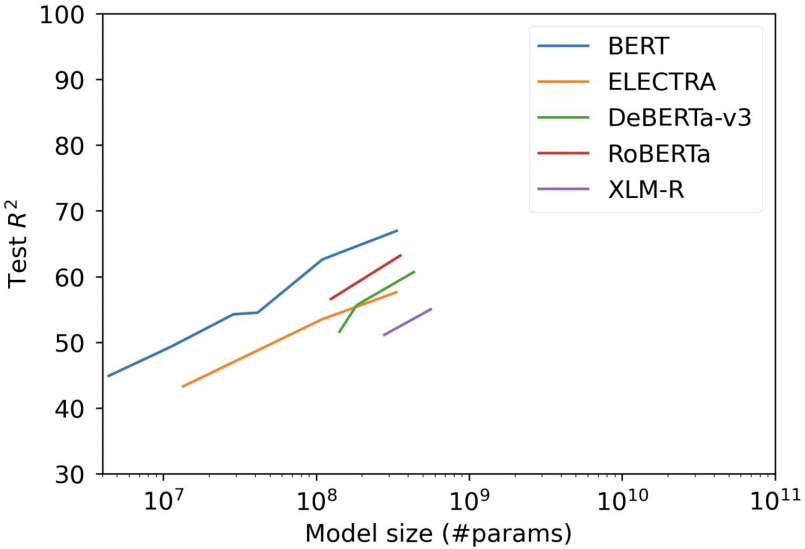
(c) Pythia 1B ($R^2 = 67.94$)

(d) Pythia 2.8B ($R^2 = 74.97$)

Performance

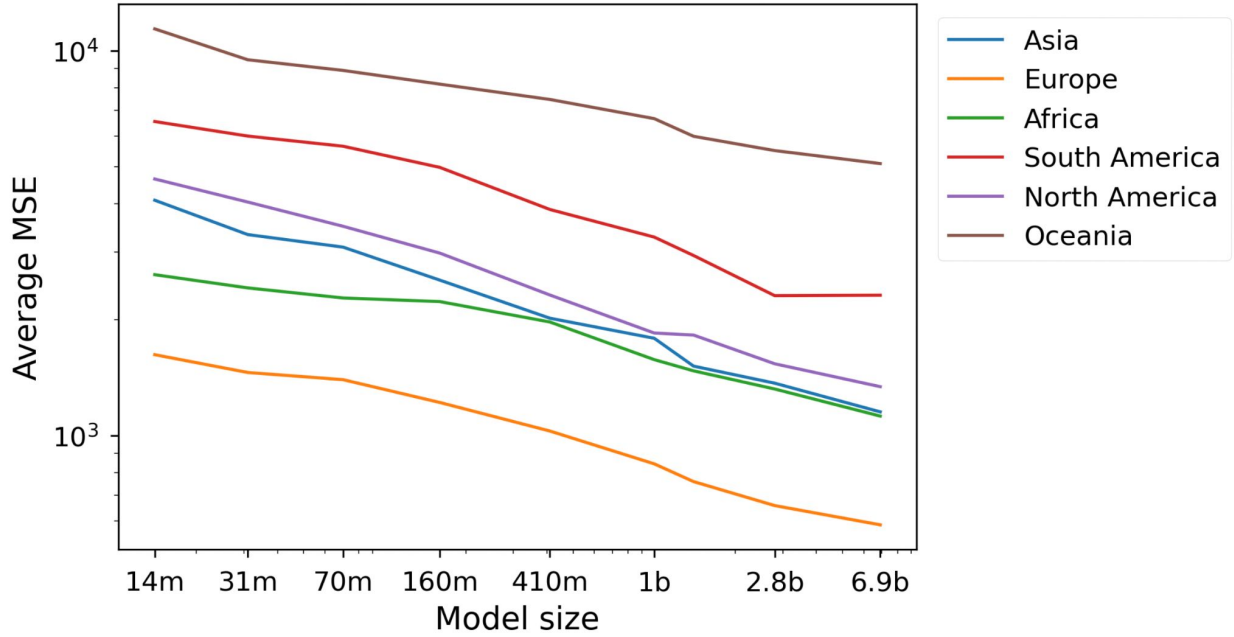


(a) Decoder models



(b) Encoder models

Performance by continent



Performance on a map

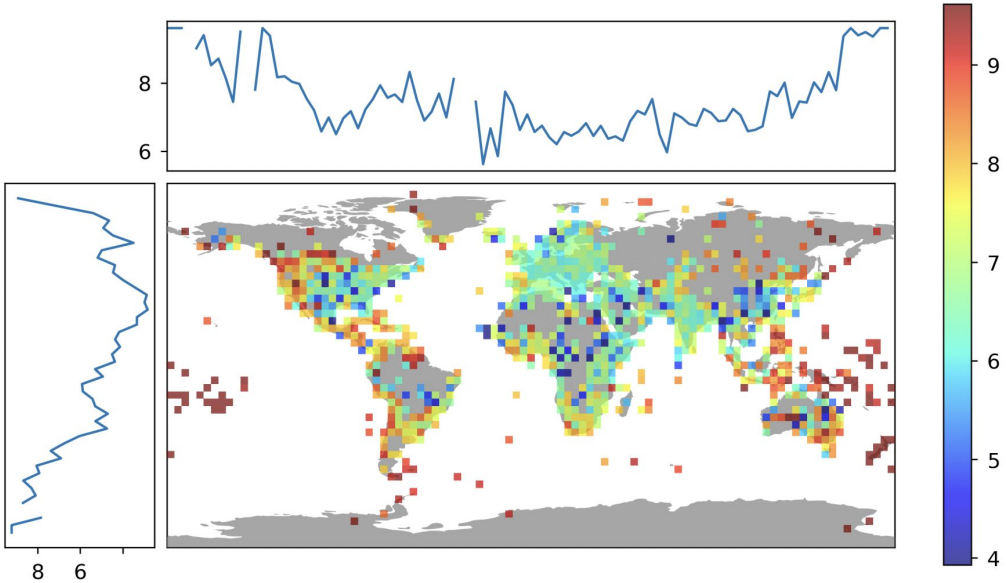


Figure 5: Test log-MSE for Pythia-1B as plotted on a World map.

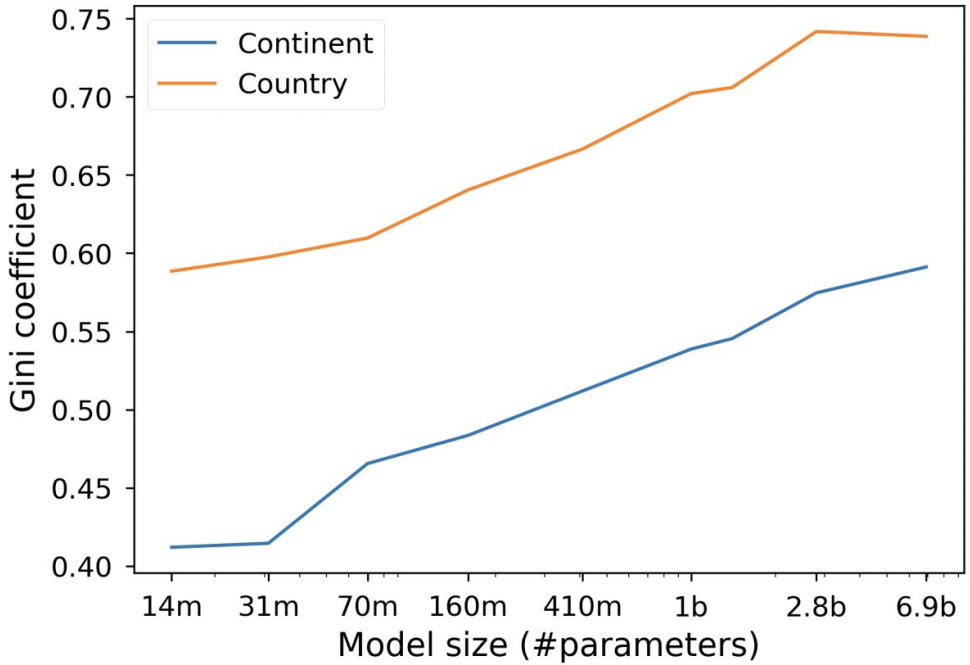
Measuring inequalities

- x : performance of the probe on each country/continent

$$Gini(x) = \frac{\sum_{i,j \in [1,N]} |x_i - x_j|}{N \cdot \sum_{i=1}^{N} x_i}$$

- The higher the coefficient, the more biased the distribution

Measuring inequalities



- The performance disparity increases with the model size!

Explaining inequalities

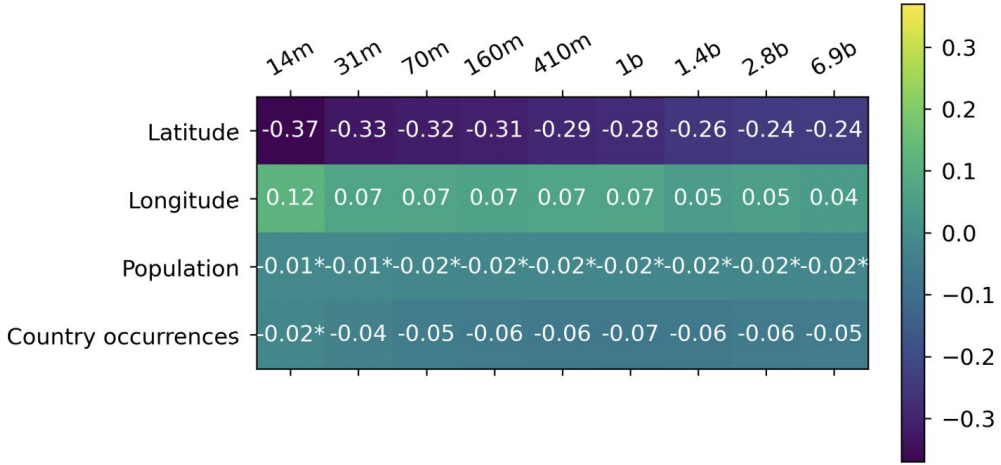


Figure 6: Pearson correlation coefficients of various factors with location-wise MSE, for several Pythia model sizes. *: Tests that yielded p-values above 0.05.

Conclusions

- The quality of geographical representation in models consistently improves as the parameter count increases.
- When model size increases, the performance discrepancy also increases.
- Geographical representation ability is correlated with latitude (North v. South) and the bias inherent to the data distribution.

Thanks!