Geographic and Geopolitical Biases of Language Models

Fahim Faisal, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{ffaisal,antonis}@gmu.edu

Abstract

Pretrained language models (PLMs) often fail to fairly represent target users from certain world regions because of the under-representation of those regions in training datasets. With recent PLMs trained on enormous data sources, quantifying their potential biases is difficult, due to their black-box nature and the sheer scale of the data sources. In this work, we devise an approach to study the geographic bias (and knowledge) present in PLMs, proposing a Geographic-Representation Probing Framework adopting a self-conditioning method coupled with entity-country mappings. Our findings suggest PLMs' representations map surprisingly well to the physical world in terms of country-to-country associations, but this knowledge is unequally shared across languages. Last, we explain how large PLMs despite exhibiting notions of geographical proximity, over-amplify geopolitical favouritism at inference time.[1]

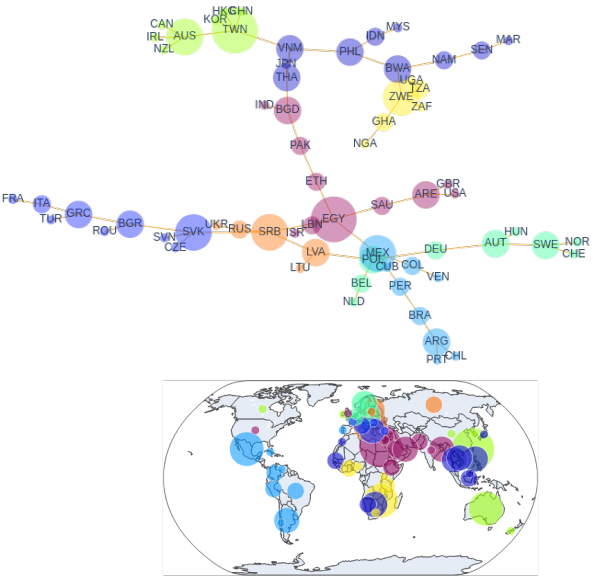


Figure 1: Example of a Geographic Representation network and it's corresponding location clusters (colored) recovered from the top-50 country-"expert" neurons of BLOOM. Notice that connected countries are either geographically or culturally close (e.g. south American cluster in light blue, African countries in yellow, South-East Asian countries in dark blue). *Note: node size is proportional to its degree in the graph.*

1 Introduction

Large pretrained language models (PLMs) are capable of generating meaningful texts beyond English and very likely, models like GPT-4, Llama 2 (Brown et al., 2020; Shliazhko et al., 2022; Zhang et al., 2022; Workshop et al., 2023; OpenAI, 2023; Touvron et al., 2023) will form the go-to base model for automating tasks like summarizing texts, generating datasets given certain instructions (Schick and Schütze, 2021) or perhaps even evaluating the generated texts (Yuan et al., 2021). While these PLMs continue to expand their utility, it is crucial that one also examines the potential biases that these PLMs exhibit. Moreover, the utility of these PLMs should be equitable to their target users so that they perform evenly for all speakers of the languages it is primarily trained on. Otherwise, the disparity that lies in the model (if any) will propagate further. To better illustrate these dynamics, consider a L_1 Spanish speaker from Peru, who is using a prompt-based PLM (like that of Wang et al. (2022, 2021)) to generate a localized synthetic dataset for some downstream task. They may use Spanish *as used in the local context* to form their seed data/prefix/prompts. Now, if this language model has already skewed preferences towards geopolitically dominant countries, it is likely the generated texts will reflect the skewness, thus not appropriately reflecting the local, Peruvian context that the practitioner is interested in. However, the quantification of this presumed geographic disparity in PLMs is not yet explored. Though given the well-documented western-country bias (or Global North bias) exhibited in most NLP benchmarks and

[1]Code and data are publicly available: https://github.com/ffaisal93/geoloc_lm

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 139–163

datasets (Faisal et al., 2022, *inter alia*), we hypothesize that text generation models might also suffer from the similar pitfall. On top of that, given a multilingual model, how language variety impact the encoded geographic knowledge is also underexplored.

Herein, we perform an evidence-based study to unfold the underlying geographic distribution of multilingual PLMs. We propose a pipeline to probe the Text-Generative PLMs using prompt-based inference for Geographic-Knowledge as well as existing domain-variant disparity (geography in our case). Our research questions and key findings are:

- **RQ1:** *To what extent is geographic proximity encoded in the PLMs?* **F:** PLMs can infer geographic proximity surprisingly well in terms of country-country association (see Figure 1). However, we observe an over-representation of certain countries during text generation.
- **RQ2:** *What is the influence of multilinguality in PLM's knowledge distribution of geographic proximity?* **F:** The shared multilingual representation space of PLMs has an uneven distribution of knowledge across languages.
- **RQ3:** *What is the effect of prompting using a geographic identifier (eg. "In Colombia" <generate text>) on multilingual text generation?* **F:** Prompting with certain geographic identifiers can even alter the language of free-form generated text.

2 Background and Related Work

A substantial amount of work has investigated existing social bias (eg. gender, racial, ethnic, occupational) identification and mitigation approaches in PLMs including, reducing token sensitivity during text generation (Liang et al., 2021), investigating model sensitivity (Immer et al., 2022), prompting using natural sentences (Alnegheimish et al., 2022) and probing via embedding lookup (Ahn and Oh, 2021). On the other hand, representing space and time utilizing maps and language is a long-standing domain of research (Louwerse and Benesh, 2012; Gatti et al., 2022; Anceresi et al., 2023). More recently, numerous studies are experimenting with geoadaptation of PLMs (Hofmann et al., 2023), what behavior these PLMs exhibit while probing with geographic-context, cultural-commonsense as well as temporal reasoning (Yin et al., 2022; Ghosh et al., 2021; Thapliyal et al., 2022; Hlavnova and Ruder, 2023; Shwartz, 2022; Tan et al., 2023) or how large PLMs learn the representation of space and time (Gurnee and Tegmark, 2023). However, for our goal task, first, we need to identify specific model units sensitive to certain geographic concepts. Then we would like to prioritize those units to generate output text for evaluation. A self-conditioning pre-trained model (Suau et al., 2022) is one such approach enabling us to perform the required experiments.

Self-conditioning Method Suau et al. (2022) propose an approach that extracts PLM weights having certain polarity and then prioritize those weights during text generation. Based on the generated text, they can quantify gender and occupation bias encoded by the PLM. As an example, consider a binary sentence classification task where positive class examples contain the mention of a concept word (eg. doctor) and vice-versa. A PLM is able to provide scores to these positive and negative examples. Looking at the average precision scores and the scores given by different model weights from each layer, we can identify the ones providing higher scores towards the positive examples. Suau et al. (2022) refer to these model weights as *expert units*.

Now, we can prioritize these identified expert units during text generation by artificially simulating the presence of the concept word "doctor" in the input. Basically, at every step of text generation, we replace the actual response of expert units with the typical one where the concept word is present in the input. As a result, the PLM now generates texts relevant to the concept word. In the work of Suau et al. (2022), by comparing the generated texts, they easily quantify the presence of gender-specific words thus evaluating the presence of gender bias in the PLM (for example, consider the number of sentences where the context relates to the word "doctor" and mentions male-gender words compared to female-gender words). This approach serves two main purposes: (1) Identifying expert units: model parameters responsible for generating text related to the target concept (i.e. doctor). (2) Triggering specific behaviour in text generation without explicit mentioning of the target context, which inadvertently influences the behaviour of the model.

3 Geographic Representation Probing

In our study, we use this Self Conditioning Method to first extract expert units (i.e. model weights)

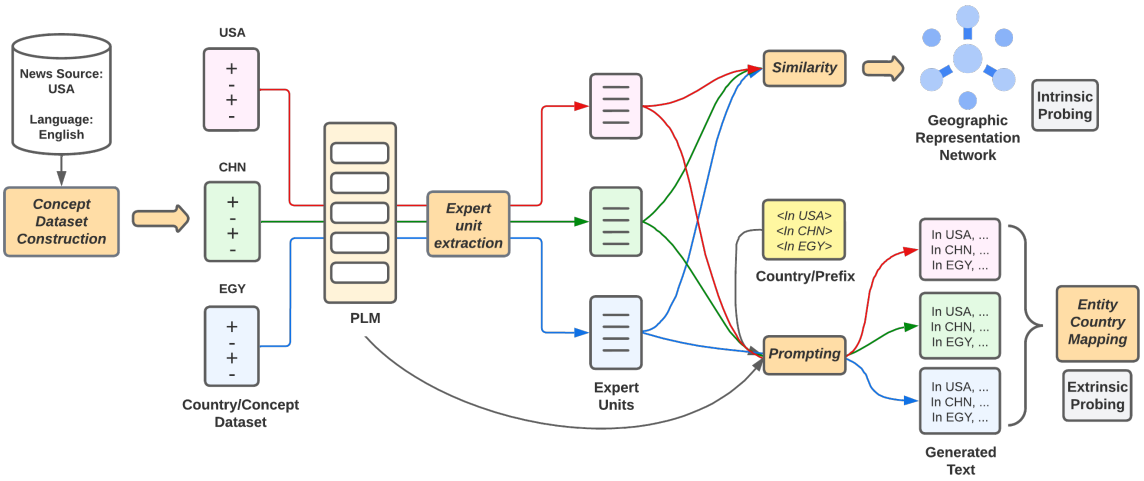


Figure 2: Geographic Representation Probing Framework. First we construct the *Country/Concept* dataset. Then we extract *Expert Units* from the base PLM and use similarity measurement to prepare our Geographic Representation Network to perform Intrinsic Probing. In Parallel, we prompt the self-conditioned PLM with Geographic Identifiers (i.e. *Country/Prefix*). Finally, we map the generated-text entities to countries to perform Extrinsic Probing.

which encode geographic knowledge. Then we use those units to generate relevant texts given different geographic identifier-based prompts. An example: Using some sentences with the mention as well as absence of the word *"China"* to extract expert units and then, prioritize these units during text generation with the prompt *"In USA ..."*. The aim here is to simulate an environment where we evaluate the model knowledge (*Concept-Country*-specific *Expert Units*) by asking what it knows about other countries (i.e *Prefix-Country*). This allows us to quantify existing geographic bias towards certain attributes present in a PLM. Our probing framework contains five steps (see Figure 2): (1) Concept Dataset Construction (2) Expert Unit Extraction (3) Geographic-Representation Network Construction (4) Prompt-based Text Generation (5) Entity Country Mapping.

Concept Dataset Construction First of all, we prepare our concept dataset in a binary classification fashion using which, we later perform self-conditioning a PLM on geographic concepts. To make it quantifiable, we define *country* to be our main unit of reference and construct concept datasets where each "concept" is loosely centered around a country. An additional requirement for these datasets is that the data have not been used as part of the pretraining data of the PLMs. Hence, we turn to recent news articles (scrapped using Google news api[2]): as we can control the date on which these data became public, we can be sure that they were not used in any pre-training process (so far). Such a dataset should also allow us to get a reasonable representation of current geopolitical affairs. Depending on the news-source country and language, we build several such *Concept-Country* datasets. A *Concept-Country* dataset `{C}-{l}` contains news about several $(c_1, c_2, ..c_i..c_n)$ countries in `{l}` language where the news-source is `{C}` country. Each *Concept-Country* c_i has 100 positive examples (mention of c_i) sentences and 300 negative examples (no mention) sentences. For example, `USA-eng` *Concept-Country* dataset (Figure 7) contains data from US sources, in English, which either mention other countries (there are 100 positive examples for each country c_i) or are random sentences not mentioning any countries (negative examples). See App. C for the constructed dataset details with examples.

Expert Unit Extraction Using the self-conditioning method, we identify high performing *Expert Units* for each *Concept-Country*. These are the model weights that provide higher scores for the presence of a specific concept (i.e. country in our case). For example, Consider the *Concept-Country* `India` from the dataset `USA-eng`. Essentially, we have positive examples (text mentioning India or relevant entities) and negative examples (random other sentences not mentioning India) which we can use to identify the model's *Expert Units*. These units are the neurons that can be used as predictors to identify the presence of a concept (i.e. positive examples mentioning "India"). The self-conditioning

[2] https://github.com/ranahaani/GNews

lang	Template –> Prefix	English Meaning
ara	في <country> –> في إسبانيا	In Spain
ben	গতকাল <country> এ –> গতকাল স্পেন এ	Yesterday, in Spain
eng	However, in <country> –> However in Spain	However in Spain
fra	<country> est connu pour –> Espagne est connu pour	Spain is known for
hin	<country> में, –> स्पेन में,	In Spain
kor	<country>에서 –> 스페인에서	In Spain
rus	Вчера <country> –> Вчера Испания	Yesterday Spain
zho	昨天 <country> –> 昨天西班牙	Yesterday Spain

Figure 3: Prefix construction using Multilingual Prefix-Templates. Here we replace the **<country>** position with "Spain" in the given language. Complete list of multilingual prefix templates in Appendix D.

framework computes these neurons and uses the average-precision score to rank their predictive expertise thus allowing us to select the top-k (eg. 10, 50) *Expert Units* from each layer. Observing the average precision scores, we select the top-k (eg. 10, 50) *Expert Units* from each PLM layer. A comprehensive theoretical explanation of the self-conditioning method and the *Expert Unit* extraction process is presented in App. B.

Geographic-Representation Network Now utilizing all these model *Expert Units*, we construct our Geographic-Representation Networks. We use jaccard similarity to measure the similarity between any given *Concept-Country* pairs c_i and c_j and their corresponding *Expert Units*. Then, utilizing these similarity measurement scores as edges in a graph (the countries being the nodes), we prepare a PLM-specific Geographic Representation network for each of our *Expert Units* set. This network is a Minimum-Spanning Tree graph highlighting the internal country-country associations. We further make it easier to digest by identifying the community clusters of countries using the Louvain Community Detection method (Blondel et al., 2008). In Figure 1 we show the network obtained with the `USA-eng` dataset from the BLOOM (Workshop et al., 2023) *Expert Units*. Effectively, we can recover a very good geographical representation of the countries straight from the network weights.

Prompt-based Text Generation With the *Concept-Country*-specific *Expert Units* at hand, we can now investigate what happens when we use the PLM for text generation. The self-conditioning method (Suau et al., 2022) uses sequential decoding and prioritize the *Expert Units* by approximating their scores from the average precision values predicted for a certain *Concept-Country*. This allows us to artificially simulate the presence of a country name and it's related context during text generation. Now we perform text generation with one more twist: we provide one country-mention as part of the prefix/prompt (i.e. *Prefix-Country*). The idea here is to simulate an environment where we evaluate the model knowledge (*Concept-Country*-specific *Expert Units*) by asking what it knows about other countries (i.e *Prefix-Country*). We generate several template-based multilingual prompts (the prefix construction process is depicted in Table 3) where we replace the `<country>` tag with different country names.

Entity Country Mapping Finally, to investigate the existence of geopolitical favouritism, we quantify the geographic biases of the generated texts by mapping any entities appearing in the text to corresponding countries. We use the Dataset Geography framework of Faisal et al. (2022), which uses multilingual entity linking to map entities to Wikidata entries and then to countries.

4 Experimental Settings

Terminologies Based on our Framework description, let us list some terminologies that we use for the remainder of the paper, to describe the experimental settings and results.

1. ***Concept-Country***: These are the countries for which we collect news.
2. **Source Country**: These are the country of origin *from* where the news data is produced.
3. **Prefix**: This is the text that we use to prompt the model, which may include a country mention. This country is the *Prefix-Country*.
4. ***Expert Units***: The model units that are specific to a country concept c_i and are extracted from the language models.

Models and Languages We use GPT2-medium (Radford et al., 2019), mGPT (Shliazhko et al., 2022) and BLOOM-560m (Workshop et al., 2023), all models available through huggingface. For the English dataset sourced from the US-News Platform (`USA-eng`) we extract *Expert Units* from all three models. For non-English datasets, we perform *Expert Units* extraction on BLOOM and mGPT. For the generation-level analysis step, we use BLOOM and GPT2 (focusing on English) expert units and report results for conditioning *Concept-Country* datasets in 8 languages: (ara,

ben, eng, fra, hin, kor, rus, zho).

Datasets As mentioned before, each concept in our dataset contains 100 positive and 300 negative examples. In some cases, we use up-sampling by repeating the example sentences multiple times when we do not have 100 distinct examples mentioning the *Concept-Country* name. In total, we prepare 31 *Concept-Country* Datasets (22 Country News-Sources, 13 Languages) and extract expert units conditioning over these datasets. Detailed dataset statistics are in Appendix Table C.3.

Generative Scheme: On average we generate `112,225` sentences for a given model and *Concept-Country* Dataset. For 67 *Concept-Country Expert Units*, we randomly choose 5 prefix templates; replace those with all 67 country name and generate 5 sentences with the lowest perplexity per *Prefix-Country*; thus `67x5x67x5=112,225` sentences.

Probing Metrics We analyze both the Geographic Representation Networks (intrinsic/parameter probing) and the generated texts (extrinsic/generation probing) to answer our Research Questions where we utilize the aid of visualization and three additional quantitative metrics as follows:

1. Neighbourhood Score: We propose a proximity-based metric to quantify the inherent encoding of Geographic Proximity present inside an LM by looking at the country-country associations and compare them with the physical world. For example, in Figure 1, South-American neighbouring countries are clustered together thus preserving a factually consistent representation. To capture this, we compute the number of neighbours one country node is connected within a 2-hop distance given a Geographic-Representation Network. To better illustrate, consider in a Geographic-Representation Network G, country node $c_5 \in G$ is connected with 4 other country nodes $\{c_1, c_2, c_3, c_4\} \in G$. Among these 4 connected nodes, c_5 shares sea or land borders with only 2 countries $N_5 = \{c_2, c_3\}$ in real world thus making $|N_5| = 2$. Similarly, we can compute $|N_2|$ and $|N_3|$ for countries c_2 and c_3 respectively. So, the Neighbourhood Score $n_s(c_5) = |N_5| + |N_2| + |N_3|$ which we can generalize and aggregate at the network level as follows:

$$N_s(G) = \sum_{c_i \in G} n_s(c_i)$$
$$= \sum_{c_i \in G} (|N_i| + \sum_{j \in N_i} |N_j|)$$

2. Representation Score: We quantify the overall command of prefix, concept or top-represented countries at the *language* level (i.e. for all generated text in a language). Consider we have *Expert Units* already computed for *Concept-Country* c_i. We use these units to generate text while providing a *Prefix-Country* p_j. Later, we map the entities of generated text to countries. So if we have a total of $L = \{l_1, l_2, ..l_k..l_n\}$ countries with respective entity counts, we can get the top represented countries $T(c_i, p_j)$ for each concept-prefix pair (c_i, p_j):

$$T(c_i, p_j) = \arg\max_{l_k \in L} (P(l_k | c_i, p_j))$$

Having this set of highly represented countries for each concept-prefix pair at hand, we can now compute in how many cases a *Concept-Country*, *Prefix-Country* or the top-10 most represented countries are present in the set $T(c_i, p_j)$ for all $c_i \in \mathcal{N}$, $p_j \in \mathcal{M}$ where $\mathcal{N} = \{\text{Concept Countries}\}$, $\mathcal{M} = \{\text{Prefix Countries}\}$. So given one output-country-distribution B:

$$\text{RS}(B, x) = \sum_{c_i \in \mathcal{N}} \sum_{p_j \in \mathcal{M}} |T(c_i, p_j) \in A_x| \text{ where}$$

$$A_x = \{\text{prefix } p_j, \text{ concept } c_i \text{ or top-10 country}\}$$

The intuition here is to quantify how much the influence of *Concept-Country*, *Prefix-Country* or overly represented countries varies across languages. For example, if we observe that the score for *Prefix-Country* is higher than the scores for *Concept-Country* across all settings, it means *Prefix-Country* is a more influencing factor than *Concept-Country* in the geographical relatedness of the text generation. For comparative analysis, we consider top-3 represented countries instead of just one while computing $T(c_i, p_j) \in A_x$.

3. Skewness[3]: We compare the symmetry of the generated country-entity distribution for both generated and the concept dataset texts. The ones that are more skewed one the ones containing amplified bias towards certain country-origin entities.

5 Findings

RQ1: *To what extent the geographic proximity is encoded in the PLMs?*

Intrinsic Findings: Based on our analysis of the Geographic-Representation Networks, it is evident that model parameters respond similarly for

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html

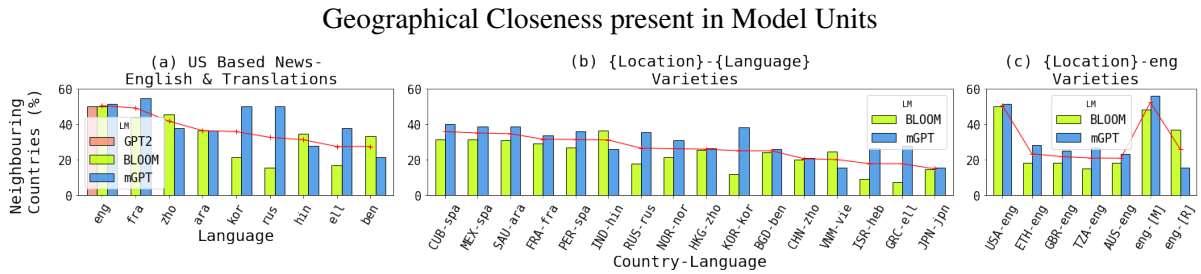


Figure 4: (a) The variation of `neighbourhood score` for different set of expert units. Notice at (a.1) we get the best score for USA-eng and it decreases when we translate the concept dataset. This also varies across languages, models (a.2) and the precise identification of expert units using high-quality concept-dataset also matters (a.3).

closely related (culturally or geographically) countries. For example, consider the Network in Figure 1 from BLOOM *Expert Units* conditioned using the USA-eng *Concept-Country* dataset. The Latin-American, African and European blocks are fairly clear. The Indian Subcontinent countries (BGD, PAK, IND), or countries of the British Commonwealth (AUS, NZL, CAN) are also clustered together. In addition, from the communities identified with the Louvain Community Detection algorithm, as visualized in the world map plot, we observe that community clusters are mainly formed around countries with proximity. We prepare similar kinds of Geographic-Representation Networks for all sets of *Expert Units* conditioned on different *Concept-Country* datasets (see Appendix E).

Concept	Genareted		Expert Units		
	gpt2	bloom	gpt2	mgpt	bloom
USA	USA	USA	*SRB*	*SWE*	*SWE*
GBR	GBR	FRA	*POL*	*HUN*	*HUN*
FRA	CHN	IND	BGR	AUT	*SVN*
CHN	IND	GBR	*SVK*	*SVK*	*GRC*
UKR	FRA	CHN	*SWE*	CHN	*SVK*
RUS	CAN	RUS	PER	*GRC*	*POL*
DEU	RUS	JPN	LVA	*POL*	*ARG*
ESP	AUS	KOR	*HUN*	*SVN*	COL
AUS	JPN	DEU	*ARG*	CHL	BRA
JPN	ISR	ESP	TZA	*TUR*	*TUR*

Table 1: Top represented countries across concepts and generated text. For BLOOM we aggregate across all eight languages; GPT-2 is English only. For expert units, we report the countries with the highest degree of similarity associations. (The common countries in at-least two model settings are in italic font.)

Extrinsic Findings: Next we investigate *whether the encoded geographic proximity gets modified due to geopolitical favouritism* by performing entity-country mapping on a large pool of generated texts in eight languages (112,255 avg. sentences per language). Evidently, we observe a strong presence of *geopolitical favouritism* which we define as the over-amplification of certain country representation (eg. countries with higher GDP, geopolitical stability, military strength etc). For comparison, we use the distribution of the *Concept-Country* dataset as it contains the actual news text reflecting real-world affairs.

In Table 1 (two left sections), we contrast the top represented countries aggregating the counts from all *Concept-Country* datasets to the ones in the generated text. All top-10 most represented countries in generated texts are present within the top-16 ranks of geopolitically significant countries.[4] This resemblance of higher geopolitically powerful country distribution is visible across all forms (Generated text Country Maps in Appendix F). However, when we compare these top-10 country representations (%) in generated text with the one from the concept dataset, we observe *geopolitical favouritism*. The result is presented in Figure 6 where in all language country-entity distributions, the top-10 country percentage is always higher compared to real-world news (Figure 6(a)). A similar pattern is apparent for the other 7 languages (except Korean) in terms of data skewness (Figure 6(b)). Last, we performed Kolmogorov–Smirnov and Shapiro statistical significance tests to ensure that the generated text country distribution follows a log-normal distribution. The striking fact here is, though this distribution contains entity mention from 246 countries in total, around **11.5%** of all generated entities are from the USA alone. This phenomenon can be further quantified using the neighbourhood score reported in Figure 4. For example, as shown in Figure 4(a), we find that all 3 models (GPT2, BLOOM, mGPT) Geographic-Representation Networks built from the English dataset conditioned *Expert Units* have around 50% of the countries connected with their real-world 2-hop neighbours.

RQ2: *What is the influence of multilinguality in PLM's knowledge distribution of geographic proximity?*

[4] worldpopulationreview-powerful-countries

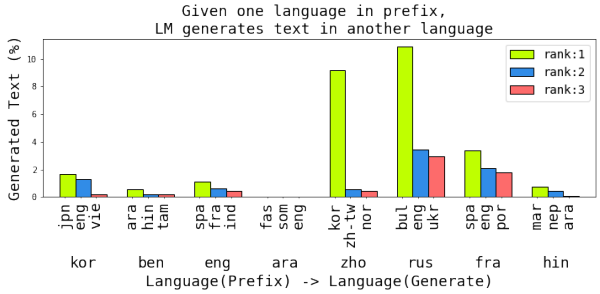


Figure 5: Percentage of generated text (top-3) in different language given the Prefix being in another language.

Intrinsic Findings: By now, we have evidence that Geographic proximity is directly encoded in PLMs in the form of shared expert units. So how this knowledge differs across languages? Ideally, multilingual PLMs should provide equitable utility for their intended users being consistent cross-lingually. To evaluate this, we automatically translate[5] our USA-eng dataset, to avoid any confounders from news content discrepancies from across the world. This way, the content used for identifying the expert units is thematically and semantically the same across languages. The result, in Figure 4(a), shows noticeable disparities in Neighbourhood Score percentages across languages in terms of Neighbourhood Scores. When we find *Expert Units* using Latin-script based *Concept-Country* datasets (English, French), the *Expert Units* make the most of associations among closely related neighbours, while the scores are less than half for Russian, Greek, or Korean in models like mGPT or BLOOM.

RQ3: *What is the effect of prompting with geographic identifier (eg. "In Colombie" <generate text>) on multilingual text generation?*

Extrinsic Findings: To answer this question, we look into the language of the generated texts using spaCy language identifier[6]. On average, BLOOM generates around 5.85% sentences (52k out of our 898k generated sentences) in a language different than the one of the prefix. This anomaly happens mostly in a larger percentage in Russian, Chinese, and French (Figure 5). We observe that every language has a specific second language preference (i.e. rank:1 in Figure 5) which can ignore the given prefix and generate a sentence in that language (eg. kor → jap, ben→ara, eng→spa, ara→far, zho→kor, rus→bgr, etc). This language preference is not reflexive (eg. kor→jap whereas zho→kor).

Observing the amount of text generated in different languages, it might seem insignificant at first sight. However, we need to keep in mind that there is one geographic identifier in the prefix (*Prefix-Country*) as well as given *Concept-Country* units. So when we look into which concept-prefix pair usually changes the direction of language, we observe interesting cultural correlations. In Table 2, given a *Prefix-Country*, we show how certain country mentions instigate text generation in a different direction (up to 50% of total generated text, given a prefix-concept pair). This happens frequently when a prefix token is shared among those languages ("in" exists both in English and Spanish; detailed examples in Appendix G) and when the country is closely tied with the language. For example, the fra→spa and eng→spa directions (French/English prefixes continued in Spanish) include country mentions of Cuba, Argentina, Colombia, or Chile which are all Spanish-speaking countries. We hypothesize that the shared representation space of multilingual decoder often ties language with geographic entity thus changing the favoured generation language.

5.1 Further Analysis

Data Origin Because we are experimenting with real-world multilingual news data without going through any extensive data cleaning process, we also need to quantify the dataset-level significance: *how does Concept-Country data quality impact the identification of Expert Units?*

The scrapping method we use for dataset construction returns localized news depending on the source location. For example, USA news source provides a higher amount of global news with many country mentions. On the other hand, a news source from Bangladesh provides news mostly about its close geopolitical neighbours (eg. India, and China). Thus, the entity frequency distribution of USA-eng and BGD-ben would not be similar.

In addition, we have variations in the amount of upsampling and the negative instance domain. So in Figures 4(b) and 4(c), we report Neighbourhood Scores for geographic-source varied on non-English and English datasets respectively. Like before, the association knowledge for USA-eng sourced Geographic-Representation Network remains the most truthful. For Spanish news sourced from different locations (Cuba, Mexico, Peru),

[5]Using https://translate.google.com/
[6]spacy-language-detection

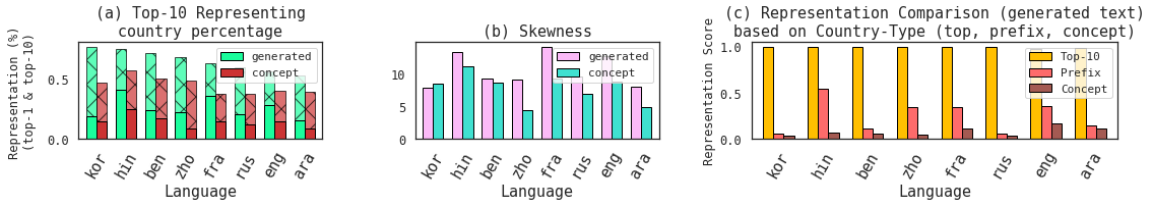


Figure 6: (a) Compared to the concept dataset which is real-world news text, the generated text always overly represents the top-represented countries (eg. USA). (b) This is also true for Skewness (except Korean). In (c) we plot the representation scores depicting the overall influence of prefixes, concepts or top countries. Top countries are over-amplified, irrespective of language. The next dominating factor is prefix but it varies across languages.

Direction	Concept	Prefix	Direction	Concept	Prefix
ben→ara	LVA	PAK	fra→spa	CHL	CUB
eng→spa	ARG	COL	fra→vie	AUT	VNM
eng→ind	IDN	KOR	fra→por	PRT	PRT
zh-cn→ko	UGA	NZL	fra→cat	CHL	SGP
rus→bul	AUS	BGR	fra→eng	CHL	BGD
rus→eng	ETH	JPN	hin→mar	BGR	ARE

Table 2: Given prefix in language A, the LM generates in a different language B (A→B), influenced by the concept and prefix countries. These are the cases where the percentage of language change is more than 50%.

scores are rather similar. Interestingly, the score drops significantly for CHN-zho compared to the translated USA-zho from Figure 4(a).[7]

For the English dataset sourced from different geographic locations (Figure 4(c)), we get poor association scores for any other locale except the USA, confirming the fact that the in-domain distance between positive and negative examples matters given a fixed language. To dig in further, we perform an ablation study by creating one additional augmented English dataset: eng-[M]: By Masking Country, Name and Organization entities in the USA-eng dataset using Spacy NER. Surprisingly, eng-[M] shows the highest percentage of geographic associations even surpassing the original USA-eng one for mGPT. We conclude that small semantic incoherence does not hurt the *Expert Units* extraction and that more contrastive positive-negative class difference (absence of other entity types) helps.

Model Comparison In terms of Neighbourhood Score, mGPT *Expert Units* encode **23.5%** more geographic expertise over BLOOM-560m model on translation datasets (similar text, different language). This improvement is increased **30%** when we consider the multilingual datasets (text and language: both different). GPT-2 units perform similarly on the English dataset.

We conduct another ablation study to quantify how to prune these models towards randomness and semantic incoherence. We prepare another augmented English dataset eng-[R], by putting random semantically incoherent texts while maintaining the positive-negative class difference. The bar showing the Neighbourhood Score is at Figure 4(c). Now BLOOM *Expert Units* are almost as good as before, whereas mGPT *Expert Units* are way worse; only in 3 other cases do BLOOM-560m units represent better associations in total. This reveals that these models contain different distributions even though they were trained with similar objectives, showing different magnitude responses towards data attribute variations, including noise, semantic coherence, data quantity and language.

Influence of *Concept-Country* and *Prefix-Country* We simulate an environment where we provide *Expert Units* about one geographic entity (*Concept-Country*) and ask a PLM about another geographic entity (*Prefix-Country*). By now, we have shown that the PLM encodes geographic proximity but also exhibits geopolitical favouritism during inference. The question we ask at this point is: *Given that PLM is biased, how do the Concept-Country and Prefix-Country influence text generation?*

To answer this question, we compute Representation Score on generated texts varying the language (Figure 6(c)). As always, top-10 country Representation Score is evident in all languages while the second most influencing factor is *Prefix-Country*. In Hindi, *Concept-Country* has the highest influence of geographic mention in a prompt-based generation. However, this scenario does not hold for the cases of Korean, Bengali, and Russian. On the other hand, *Concept-Country* plays the part of a subtle representative but fails to compete with

[7]While investigating this anomaly, we found that the fixed sequence length for both models (BLOOM, mGPT) rejects several positive examples during tokenization process thus hurting the *Expert Units* extraction quality. We corrected this issue by substituting the long examples with shorter ones.

Prefix-Country and geopolitical significant countries. One fact to note here is, our experiment contains a small number of examples while generating a large pool of texts. Nevertheless, we believe that it will require intensive data creation efforts to mitigate the biases that coexist with the geographic knowledge in PLMs.

6 Conclusion and Future Work

In this study, we perform an experimental analysis on identifying the inherent geographic knowledge and inference bias of prompt-based decoder models. Our experiments strongly suggest that current PLMs are able to encode geographic proximity quiet well. However, almost always geopolitical favouritism overshadows the encoded proximity during inference. This finding raises concerns as well as the need to perform bias-mitigation steps if we want to generate geo-specific texts. Our additional findings on the impact of multilinguality on prompting points out how encoded geographic proximity is unevenly distributed across languages and how even just a mention of geographic identifiers may influence the language of free-form text generation. We believe these findings still leave issues to be addressed in current practice and that there should be a a fundamental multilingual-bias mitigation step included in any NLP task workflow. Keeping this in mind, we want to expand the domain of our proposed probing framework and assess its applicability beyond geography. In addition, we aim to perform contrastive training to efficiently extract expert units thus stepping forward with the effort of reducing the inequality inherent in multilingual language models.

Limitations

First of all, selecting country as geographic entities is inherently lossy and ideally, we would be able to perform the experiments with further granularity. We rely on Wikidata for entity linking, which is already somewhat biased towards western countries. In addition, our experiments are limited to 69 countries and 13 languages (8 for generating text) (by necessity and due to computing costs), ignoring other countries as well as languages, especially low-resource ones. In the future, we want to further expand our study to include more languages and cultures, as well as digging deeper in multi-cultural countries.

Acknowledgements

We are thankful to the anonymous reviewers for their constructive feedback. This work is generously supported by the National Science Foundation under grants FAI-2040926, IIS-2125466, and IIS-2127901.

References

2021. Ip2location™ country multilingual database. Online resource.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, Seattle, United States. Association for Computational Linguistics.

Giorgia Anceresi, Daniele Gatti, Tomaso Vecchi, Marco Marelli, and Luca Rinaldi. 2023. A map of words: Retrieving the spatial layout of underground stations from natural language.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv:2005.14165.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2022. Dataset geography: Mapping language data to language users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3381–3411, Dublin, Ireland. Association for Computational Linguistics.

Daniele Gatti, Marco Marelli, Tomaso Vecchi, and Luca Rinaldi. 2022. Spatial representations without spatial computations. *Psychological Science*, 33(11):1947–1958. PMID: 36201754.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time.

Ester Hlavnova and Sebastian Ruder. 2023. Empowering cross-lingual behavioral testing of NLP models with typological features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7181–7198, Toronto, Canada. Association for Computational Linguistics.

Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B. Pierrehumbert, and Hinrich Schütze. 2023. Geographic adaptation of pretrained language models.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. 2022. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1839–1851, Dublin, Ireland. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models.

Max M. Louwerse and Nick Benesh. 2012. Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth. *Cognitive Science*, 36(8):1556–1569.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2022. Self-conditioning pre-trained language models. *International Conference on Machine Learning*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for low-resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden,

Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geom-

lama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *EMNLP*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

A Frequently asked questions

A.1 What does it mean by the term geographic biases, geographic favouritism and what are their relationships with fairness?

In general, geographic bias means the over-representation of certain geographic attributes. In this study, we use *"geographic bias"* and *"geographic favouritism"* interchangeably as the over-amplification of certain country representation (eg. countries with higher GDP, geopolitical stability, military strength etc) during PLM prediction or text-generation. We believe the overall system utility of a language model should be equitable according to the needs of the intended users with different demographic and geographic origin. Thus ensuring their geographic characteristics are well-represented and not over-shadowed because of geographic favouritism is defined as *"geographic fairness"* in this study.

A.2 What's the reason for using the self-conditioning approach of Suau et al. (2022) for studying biases? There had been many other bias measures in NLP before Suau et al. (2022). Are they not suitable for the study of geographic and geopolitical biases?

A number of previous studies experimented with the behavior different PLMs exhibits while probing with geographic-context as well as cultural-commonsense (Yin et al., 2022; Ghosh et al., 2021). However, we need to extract the specific model weights responsible for these observable polarity. Then using those weights in a controlled setting, we might be able to unfold how PLMs encode geographic knowledge as well as explain the exhibition of geographic-bias during inference. The self-conditioning model proposed by Suau et al. (2022) is one such study that fits to our intended needs perfectly. This approach serves two main purposes: (1) Identifying expert units: model parameters responsible for generating text related to the target concept (i.e. doctor). (2) Triggering specific behaviour in text generation without explicit mentioning or fine-tuning of the target context, which inadvertently influences the behaviour of the model utilizing the encoded-knowledge of PLM.

A.3 What are the practical takeaways from this? Yes, different models encode geographic knowledge, so what? Should we be concerned, should we do something about it?

We recall the example presented earlier: consider a L_1 Spanish speaker from Peru, who is using a prompt-based PLM (like that of Wang et al. (2022, 2021)) to generate a localized synthetic dataset for some downstream task. They may use Spanish *as used in the local context* to form their seed data/prefix/prompts. Now, if this language model has already skewed preferences towards geopolitically important countries, it is likely the generated texts will reflect this skewness, thus not appropriately reflecting the local, Peruvian context that the practitioner is interested in. In this study we address this concern of geographic bias being one of the most-significant yet ignored attributes in practice. Moreover, we show how this is further amplified when we go beyond English and similar languages. Basically we need effective bias-mitigation module as part of the regular NLP workflow which is currently non-existent.

A.4 Why we need to extract the *Expert Units* and how *Concept-Country* helps in this regard?

One of our aims is to unfold the geographic representation using relevant PLM units without external fine-tuning. So, we need to find or extract these relevent units which are basically model parameters. So, we can use our *Concept-Country* datasets as binary classification dataset (positive class contains sentences mentioning certain*Concept-Country*) to find these highly responsive weights (i.e. *Expert Units*) to certain *Concept-Country*. Then we perform self-conditioning on the PLMs using these *Expert Units* to generate texts having the influence of these *Concept-Country*s.

A.5 Explain *Concept-Country* dataset creation process.

We scrape news using a Google news api[8] to capture the current affairs. Importantly, we can select news not just from a given date range, but also news originating in a specific country and a specific language. Such a dataset should allow us to get a reasonable representation of current geopolitical affairs. As such,

[8] https://github.com/ranahaani/GNews

each of the concept datasets we create reflects *"current news about a country reported by the mainstream platforms from another country"*. Hence, a *Concept-Country* dataset `{C}-{l}` contains news about several ($c_1, c_2, ..c_n$) countries in `{l}` language where the news-source is `{C}` country. For example, `USA-eng` contains data from US sources, in English, which either mention other countries (there are 100 positive examples for each country c_i) or are random sentences not mentioning any countries (negative examples).

A.6 Explain the *Expert Units* extraction process.

Consider the *Concept-Country* `India` from the dataset `USA-eng`. Essentially, we have positive examples (text mentioning India or relevant entities) and negative examples (random other sentences not mentioning India) which we can use to identify the model's *Expert Units*. These units are the neurons which can be used as predictors to identify the presence of a concept (i.e. positive examples mentioning "India"). The self-conditioning framework computes these neurons and uses the average-precision score to rank their predictive expertise thus allowing us to select the top-k (eg. 10, 50) *Expert Units* from each layer.

A.7 What does Geographic Representation Network actually represents?

Note that these networks are produced using the uncovered original PLM expert units, without any external data fine-tuning or prompting. Hence, they provide a view of the *inherent* geographic knowledge present inside the PLM parameter space.

A.8 Why we need to use *Expert Units* during text generation?

We have a setting where we can provide certain *Concept-Country* as part of the generation condition and the specific *Expert Units* from the model itself are supposed to be capable enough to influence the generated text. Our aim is to evaluate the geographic knowledge specific model weights or *Expert Units* by asking those about other *Prefix-Country*. This will unfold whether the geopolitical favouritism happens for geopolitically important countries or the geographical proximity (eg. neighbouring countries) takes the precedence or there exist no such patterns.

A.9 What are the factors considered while constructing the *Concept-Country* dataset?

There are two relevant factors: (1) For the negative examples in USA-eng *Concept-Country* dataset, we use news from a completely different domain (eg. automobile, sport), whereas for different geographic-sourced datasets, negative examples come from randomly sampling news of different locations. (2) The intensity of text-noise and positive example up-sampling amount varies across different news-sourced *Concept-Country* datasets.

A.10 Why 2-hop distance while calculating the neighbourhood-score?

We did experiment with n-hop scoring and they follow similar trends. We choose 2-hop is it is less complex for scoring and at the same-time, sufficient to point out the disparity across multiple languages.

A.11 Comparison to news: although these models are trained on web text, which contain news articles, they are not guaranteed to generate text like a news article. Thus the distribution of entities within the text will be different.

Yes, that is correct but our aim is to capture the learned distribution and evaluate (1) whether that distribution is skewed or not, (2) Whether there is resemblance with the real-world scenario or not. We believe, this assessment is important for a PLM which will be used for solving real-world practical tasks and having news-text for comparison might be the closest viable source we can get in a limited resource setting.

A.12 What does it mean by: "*the model weights which provide higher scores for the presence of a concept*"

In sort, a language model can provide scores to the positive and negative examples of a binary classification dataset (eg. our country-concept dataset). Looking at the average precision scores and the outputs given

by different model weights from each layer, we can identify the ones providing higher scores towards the positive examples and these model weights are referred as expert units.

B Self-conditioning Method: Theoretical Definition

Here we provide a theoretical description concerning the working procedure of the self-conditioning method (Suau et al., 2022). First, we provide an overview of the usual generative mechanism followed by the expert unit extraction procedure. Then we talk about creating the simulated environment where the expert units are prioritized to instigate text generation in a specific direction.

Generative Mechanism During autoregressive text generation, a language model maximizes the probability of a sentence $x = \{x_i\}$ as $p(x) = p(x_1, ..x_T) = \prod_{t=1}^{T} p(x_t|x_{<t})$. A conditional generative model can use a joint probability distribution to maximize the probability such that: $p(x, y) = p(y|x)p(x)$. Here, x is the generated sentence while y is a conditional variable (i.e. imposing the presence of a concept word). Dathathri et al. (2020), adopted this setting in a conditional generation where, $p(y|x)$ determines the condition and $p(x)$ ensures constraint on the generated text as it progresses. In this setting, instead of the joint distribution, the condition can even be fixed beforehand as follows:

$$p(x|y = c) \approx p(y = c|x)p(x) \tag{1}$$

Suau et al. (2022), hypothesize that the conditional maximization of $p(x|y = c)$ in Eq. (1) can be done by exploiting the internal mechanism of a PLM (e.g. expert unit extraction and prioritizing them by changing their responses during text generation).

Expert Unit Extraction Suau et al. (2022) defines expert units as the neurons contributing to the conditional model $p(y = c|x)$ in Eq. (1). They extract certain expert units which can further be used as the predictors of the concept presence identification task given an input. Formally, we define z_m^c as the set of outputs of a single neuron m to sentences $\{s_i^c\}$. We can formulate z_m^c as the prediction score of a binary sentence classification task $b^c[0, 1]$ where s_i^c is an input sentence and z_m^c varies depending on the presence/absence of a concept c in s_i^c. Now having the prediction score z_m^c in hand, we can compute the expertise of a unit m for the task $b^c[0, 1]$ by looking at the average precision score so that $AP_m^c = AP(z_m^c, b^c) \in [0, 1]$ (i.e. area under the precision-recall curve). At this point, the top k expert units are identified by ranking all the units from each model layer based on AP_m^c.

Conditional Text Generation The final step is to prioritize the identified expert units to generate texts having specific behaviors. This can be done using a $do(c, k)$ intervention which ensures the influence of concept c while prioritizing the top $k-$expert units. These top $k-$expert units previously performed as the best predictors for c concept identification from sentences. In (Suau et al., 2022), $do(c, k)$ is formulated as follows:

$$do(c, k) : \{z_c^m := E_x^c[z_c^m|b^c = 1]\forall m \in Q_k\} \tag{2}$$

This $do(c, k)$ intervention always replaces the response of an expert unit with the typical value where the concept c was present in an input sentence (i.e. $E_x^c[z_c^m|b^c = 1]$). Here, Q_k is the set of indices of all top-performing k-expert units. Now in Eq. (1), the $p(y = c|x)$ can be maximized by increasing the number of relevant expert units (i.e. k) using the $do(c, k)$ intervention according to the adopted hypothesis of (Suau et al., 2022). As a result, by just exploiting the internal conditioning mechanism of a PLM text generation and without any out-source data training, an artificial environment is created where the presence of concept c is inspired.

C Datasets

In Table 3 we present the concept dataset details. Each dataset here contains 43 to 69 country concept files (The complete list of countries is presented in Table 4).

Figure 7: A snap-shot of the *USA-eng* dataset. Each json file contains postive-negative news about one specific country. For example, the `australia.json` contains positive sentences having mention of the country name extracted from the news articles. Whereas, the negative 300 sentences are also collected from news domain having no mention of the word *australia*.

A snapshot of the *USA-eng* dataset is presented in Figure 7 to provide a better understanding of how the concept dataset is formatted. This specific dataset contains English news about various countries while the news-originating country is the USA. From the figure, we observe the mention of country-named json files (i.e. the country concept files). Each json file contains positive 100 sentences about that specific country. Whereas, the negative 300 sentences contain no mention of the specific country. Moreover, we can take a further look at the `australia.json` file where the positive instances are sentences selected from Australia-related recent news articles.

In Table 4, The `Type-2` datasets are the translated version of USA-eng dataset. In `Type-3`, we mask USA-eng entities using a NER tagger and `Type-4` is constructed using random english texts.

D Prefix Templates

For each of the eight languages, we generate prefix replacing templates with *Prefix-Country* names. Per language, we have six template prefix. The complete list is presented in Table 5

E Additional Geographic Representation Networks

In Figures (8, 9, 10, 11) we present Geographic-Representation Networks (News Source-language: USA-eng, SAU-ara, FRA-fra, RUS-rus, BGD-ben, KOR-kor, CHN-zho, IND-hin) constructed using the *Expert Units* from GPT2, BLOOM and mGPT.

F Geography Maps on generated text

We present Country Maps on the generated outputs for eight languages. The maps are presented in Figure 12.

Dataset Names			#	Description
Type 1: {News_Source_Location}-{Language}				
USA-eng	*BGD-ben*	*CHN-zho*	21	These 21 datasets are scrapped from news sources originating from 21 different countries in different languages. Each one of these datasets contain country concept sets describing news about specific countries. Each country concept are prepared using 100 positive sentence examples and 300 negative sentence examples. We use upsampling by repetition when we have less examples than the required counts. For only `USA-eng` dataset, we use english news from other topic search (eg. *Automotive, Sport*) to construct the negative examples while, for other 20 datasets we use news about other countries (i.e. in domain) as negative examples.
GRC-ell	ISR-heb	*IND-hin*		
KOR-kor	MEX-spa	NOR-nor		
SAU-ara	VNM-vie	AUS-eng		
ETH-eng	GBR-eng	HKG-zho		
TZA-eng	*FRA-fra*	PER-spa		
JPN-jpn	*RUS-rus*	CUB-spa		
Type 2: {News_Source_USA}-{Translations}				
USA-ara	USA-ben	USA-ell	8	These 8 datasets are created using translation from the USA-eng dataset. We use Google Translation API[1] to translate the texts from source language to target language.
USA-hin	USA-kor	USA-rus		
USA-zho	USA-fra			
Type 3: {USA-eng}-{Masked Entities}				
USA-eng-[M]			1	We augment `USA-eng` dataset by masking all additional entities in positive examples for each country concepts using spaCy[2].
Type 4: {USA-eng}-{Random Text}				
eng-[R]			1	We randomly use text instead of original text in `USA-eng` dataset while maintaining the positive negative class distinction but without any semantic coherence.

[1] https://translate.google.com/

[2] https://spacy.io/

Table 3: Country Concept Datasets sourced from Google News texts. We extracted expert units from language models: gpt-2 (only english), bloom and mgpt for all of these. Among these, we perform text generation using the expert units sourced from *8 datasets* (The underline ones).

ISO	Country	ISO	Country	ISO	Country
AUS	Australia	BWA	Botswana	CAN	Canada
ETH	Ethiopia	GHA	Ghana	IND	India
IDN	Indonesia	IRL	Ireland	ISR	Israel
KEN	Kenya	LVA	Latvia	MYS	Malaysia
NAM	Namibia	NZL	New Zealand	NGA	Nigeria
PAK	Pakistan	PHL	Philippines	SGP	Singapore
ZAF	South Africa	TZA	Tanzania	UGA	Uganda
GBR	United Kingdom	USA	United States	ZWE	Zimbabwe
CZE	Czech Republic	DEU	Germany	AUT	Austria
CHE	Switzerland	ARG	Argentina	CHL	Chile
COL	Colombia	CUB	Cuba	MEX	Mexico
PER	Peru	VEN	Venezuela	BEL	Belgium
FRA	France	MAR	Morocco	SEN	Senegal
ITA	Italy	LTU	Lithuania	HUN	Hungary
NLD	Netherlands	NOR	Norway	POL	Poland
BRA	Brazil	PRT	Portugal	ROU	Romania
SVK	Slovakia	SVN	Slovenia	SWE	Sweden
VNM	Vietnam	TUR	Turkey	GRC	Greece
BGR	Bulgaria	RUS	Russia	UKR	Ukraine
SRB	Serbia	ARE	United Arab Emirates	SAU	Saudi Arabia
LBN	Lebanon	EGY	Egypt	BGD	Bangladesh
THA	Thailand	CHN	China	TWN	Taiwan
HKG	Hong Kong	JPN	Japan	KOR	Republic of Korea
None	None	None	None	None	None

Table 4: List of Countries we conducted experiments on.

G Geographic Identifier and Language Direction

see Table 6 for examples of generated text examples given the prefix "In Cuba" with *Concept-Country*:Argentina. Though the *Concept-Country* dataset here we use is in English, the model generates highly frequent Spanish sentences compared to english.

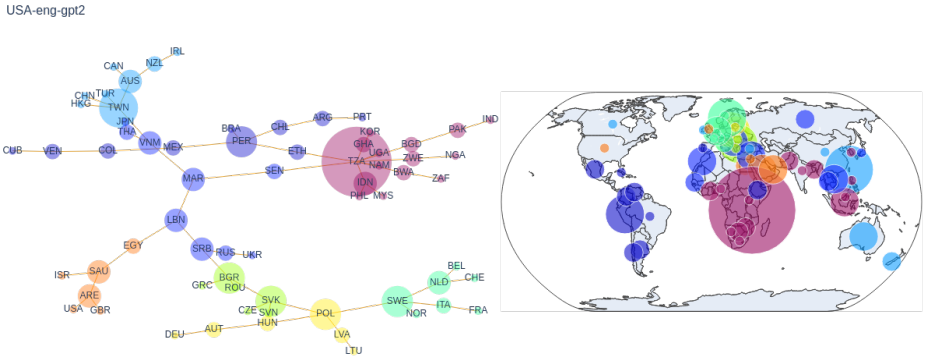
lang	Templates	lang	Templates
eng	➢ 'In <country>,' ➢ '<country> is known for' ➢ 'Inside <country>' ➢ "In <country>'s" ➢ 'However, in <country>' ➢ 'Yesterday <country>'	ara	➢ 'في <country> ،' ➢ '<country> معروف بـــ' ➢ 'د اخل <country>' ➢ 'في <country>' ➢ 'في ، ومع ذلـك <country>' ➢ 'أمس <country>'
rus	➢ 'В <country>,', ➢ '<country> известна', ➢ 'Внутри <country>', ➢ 'В <country>', ➢ 'Однако в <country>', ➢ 'Вчера <country>'	kor	➢ '<country>에서는' ➢ '<country>는' ➢ '<country> 내부' ➢ '<country>에서' ➢ '그러나 <country>에서' ➢ '어제 <country>'
ben	➢ '<country>,' ➢ '<country> এর জন্য পরিচিত' ➢ '<country> এর ভিতরে' ➢ '<country> এর' ➢ 'তবে, <country>' ➢ 'গতকাল <country> এ'	hin	➢ '<country> में,' ➢ '<country> के लिए जाना जाता है' ➢ 'अंदर <country>' ➢ "<country>'एस . में" ➢ 'हालांकि, <country> . में' ➢ 'कल <country>'
zho	➢ '在<country>, ' ➢ '<country> 以' ➢ '<country>内部' ➢ '在<country>的' ➢ '但是, 在 <country>' ➢ '昨天 <country>'	fra	➢ 'En <country>,' ➢ '<country> est connu pour' ➢ "À l'intérieur de <country>" ➢ 'En <country>', ➢ 'Cependant, en <country>' ➢ 'Hier <country>'

Table 5: Prefix templates we use for Multilingual Text Generation. We replace the **<country>** with the corresponding country name in generator language. For example, To construct one USA-mention Chinese prefix, we replace **<country>** with 美国. We use a multilingual country-name dataset (cna, 2021) to query country names.

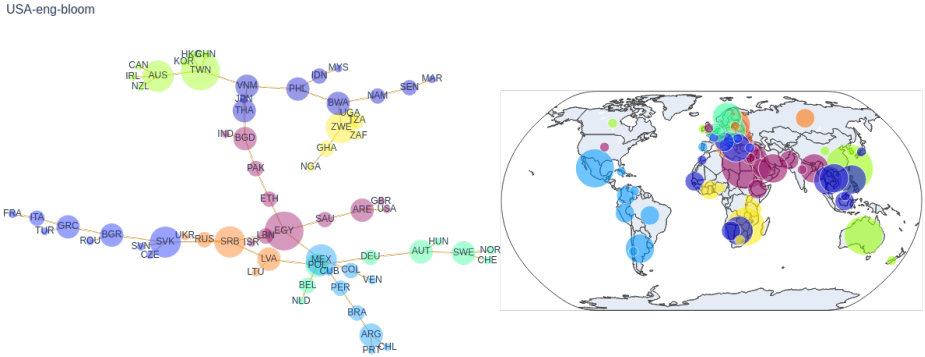
Language Direction	Generated Text
eng→spa	*In Colombia, beginning in 1991*, Ley de Pesca y Tierra Naranja tiene como una estrategia de Economía Indígena presenta como Ley de Conservación y Desarrollo Agrícola
eng→eng	*In Colombia, patients with PO are routinely referred to the Pediatric Critical Care Units (PC from 1996) because they are mostly after peak twice* ***a los to*** *participating in*
eng→spa	In Colombia, donde está en etapa de vacunación las primeras etapas las personas que llegan en el jueves (figuana para el millón y ultimaniano casos y el
eng→spa	In Colombia, la noticia odia a Dios. Es una religión que no santifica. Esta seccionalizada del 4Chanuto para algunos países, a sociedad que
eng→spa	In Colombia, el mercado de la carne, considerado el segundo mayor productor de cortes de carne bovina en la región, es de caña de insumo a nivel
eng→spa	In Colombia, el partido del "9-3" ha sido en la decisión del colombiano, la celebración de Luis Zubeldense Humberto Bloom (peruano, quien abrió
eng→spa	In Colombia, afloró por las fronteras de Argentina. Entre 1985 y 1993, de la República Dominicana, Bolivia, después llegó a Colombia y Ecuador. El entrenador
eng→spa	In Colombia, execuções entre elites, o Partido Comunista y sindicatos de esos países vecinos elites a partiran llevan la denuncia que derrochales. Las
eng→spa	In Colombia, una estrecha relación entre Washington y Venezuela tiene un mensaje claro sobre Bolsonaro. Así mismo, aunque no ve la necesidad de revisar lo que de no hacerlo de
eng→spa	In Colombia, a 0.70 por ciento de la población de niños mueren prematuros de gripe por sobrepeso ha sido diagnosticada. El representante del tamaño real de
eng→spa	In Colombia, PDOT, que hace más de 10 años había significado cerca de 160 actividades laborales para sus miembros, al día e instalaciones de 14 mili 300 personas
eng→spa	In Colombia, made del Derecho penal, es la máxima parte de la violación a través de los notaria Núcleo de medidas contra la descripción de la Justicia y
eng→spa	In Colombia, Cristina Kirchner — la vicepresidenta del fallecido expresidente Néstor Kirchner— ha confesado que "en las últimas horas pasó todo como una enfermedad que no se registró su mujer
eng→spa	In Colombia, el Código Penal declaró cierto grado de subordinación de la salud mental de las víctimas de trabajadores a responsables funcionalistas, no profesionales por el Estado como se
eng→eng	*In Colombia, the majority of women are Catholic. But in the country is still refuses to accept the Catholic counseling school, and, penalizes women after to leave*
eng→eng	*In Colombia, for example, we observed a significantly lower prevalence of chronic bronchoalveolar or peritonitis, bronchobronchial hypertrophy than mon*
eng→spa	In Colombia, un importante sector de las diezañeras vuelve a poner en valor de la importancia el anonimato de las producciones francesas cuando, una mezcla que habían obtenido a
eng→eng	*In Colombia, the EMA has regular royalties on a $27,800 per fee,800 day to $39,000 protein products at the expert. The fair*
eng→eng	*In Colombia, in turn, the mass distributions represent very low prevalence, being around 4. The USA around 35 40-47% and in the usual, and 45%*
eng→spa	In Colombia, el gobierno presentó este miércoles un proyecto de ley en la primera lectura online para eximir controles y renegociación internacional e internacional de suscripto de divisas con

Table 6: Example Generated Sentences with the prefix "In Colombia" and "Country/Concept" Argentina.

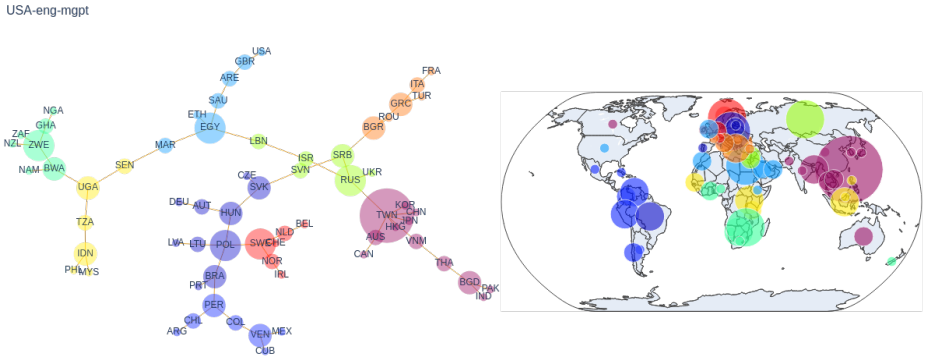
Geographic Representation Networks and Corresponding Community Maps



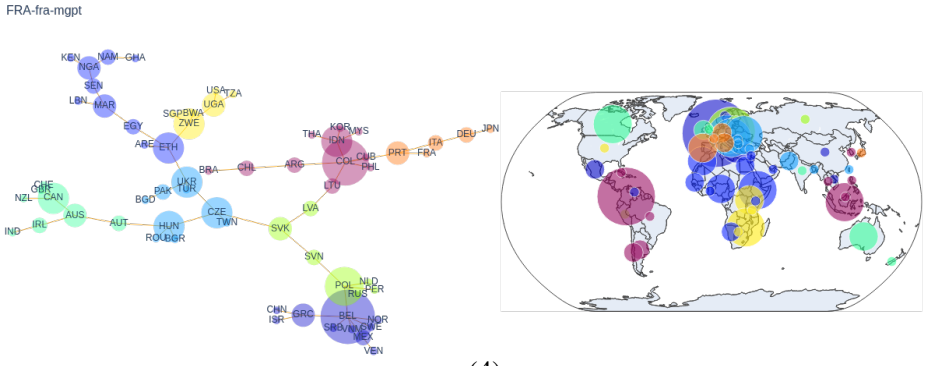
(1)



(2)

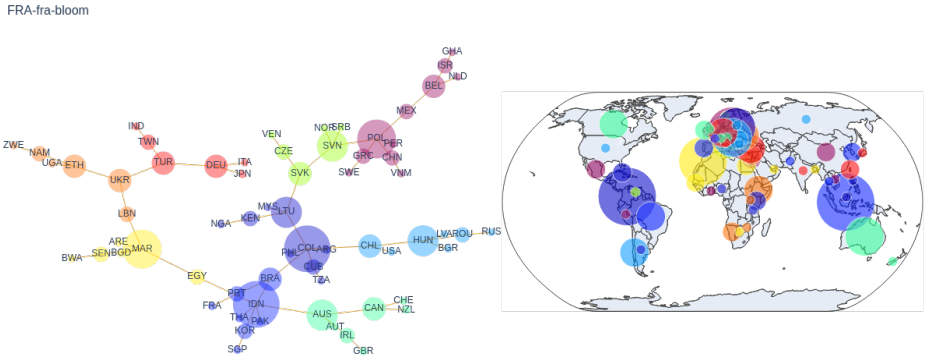


(3)

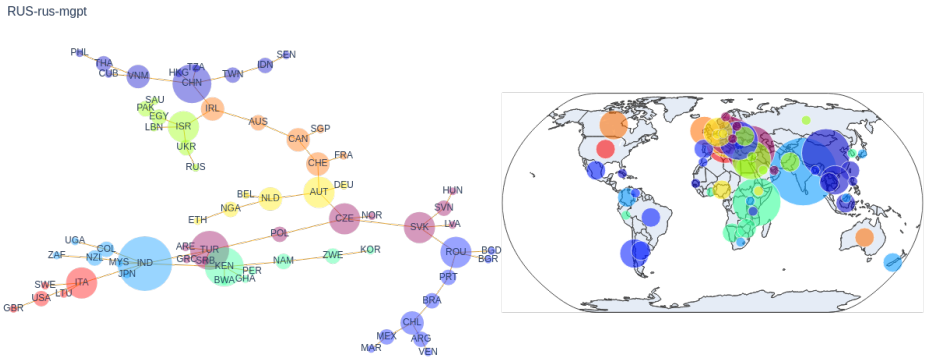


(4)

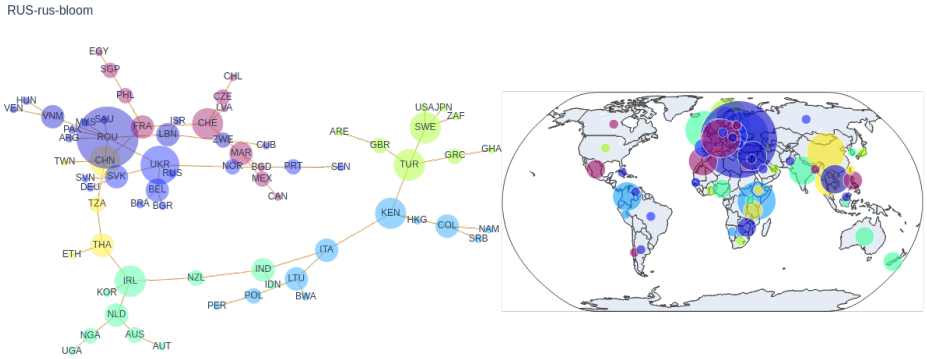
Figure 8: Geographic Representation Network and Corresponding Community Map for different Expert Unit set Associations. The language models we use are GPT2 (only English), mGPT and BLOOM.



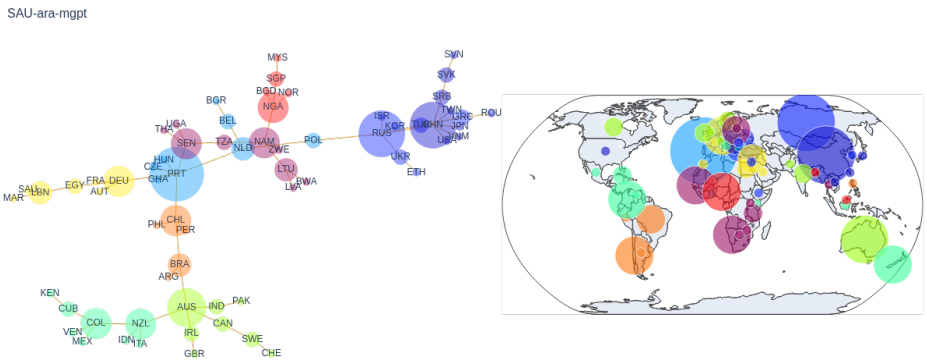
(5)



(6)

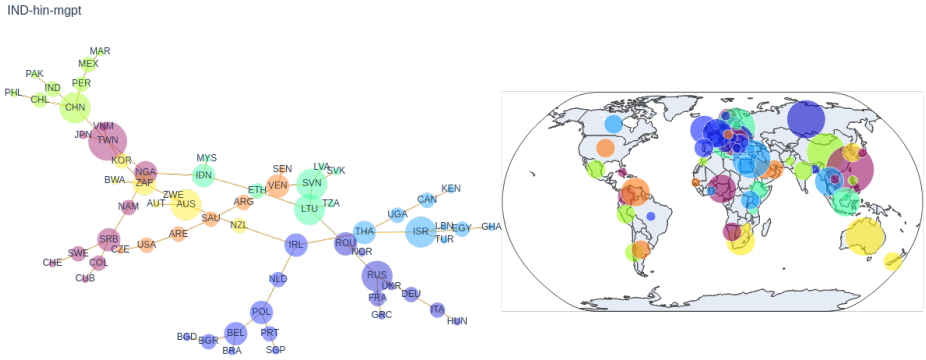


(7)

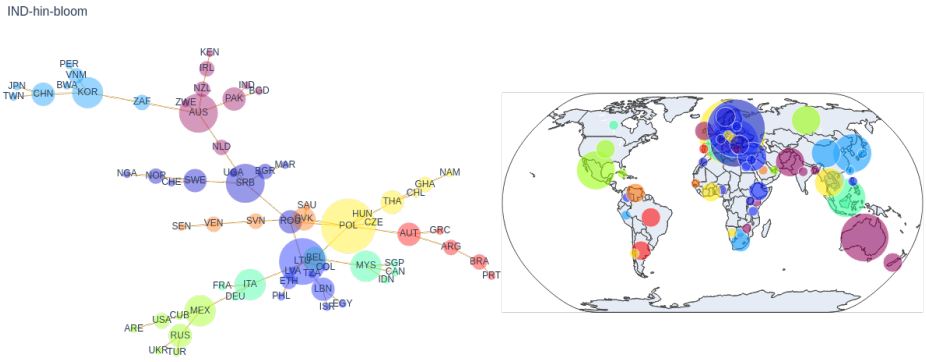


(8)

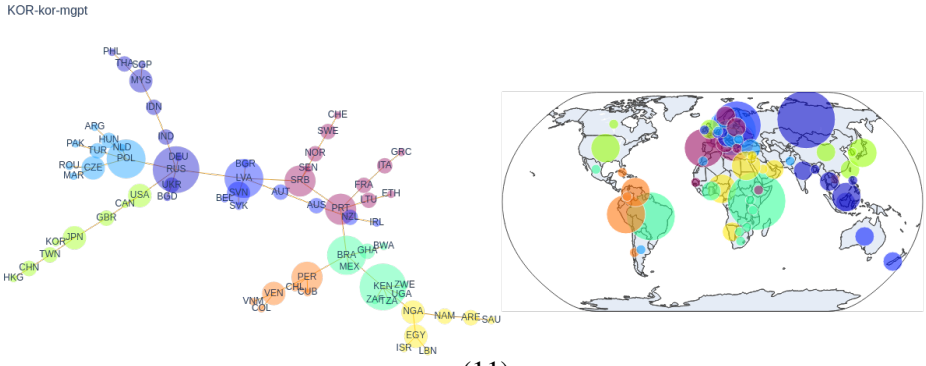
Figure 9: Geographic Representation Network and Corresponding Community Map for different Expert Unit set Associations. The language models we use are GPT2 (only English), mGPT and BLOOM.



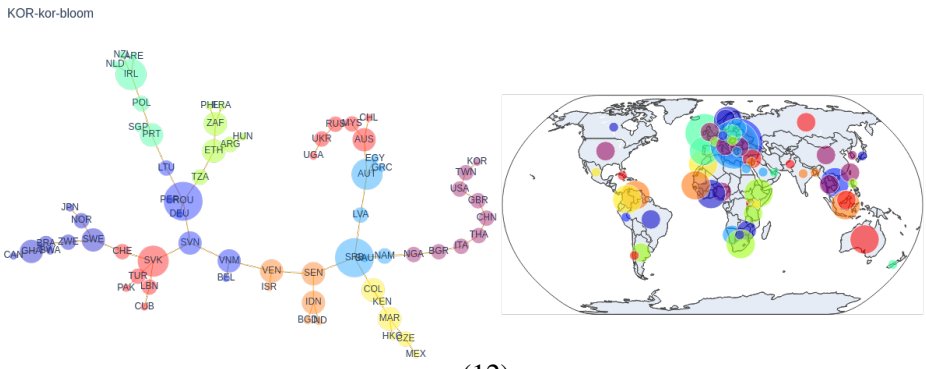
(9)



(10)

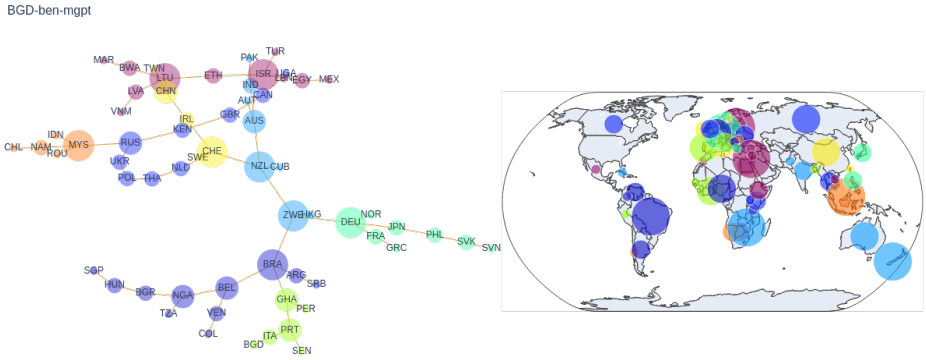


(11)

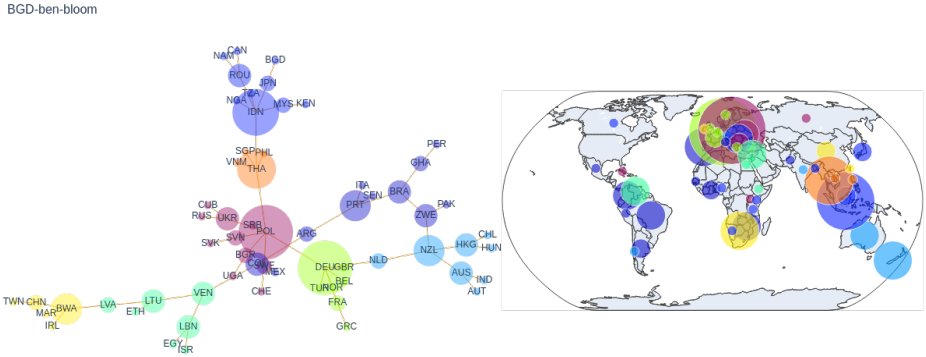


(12)

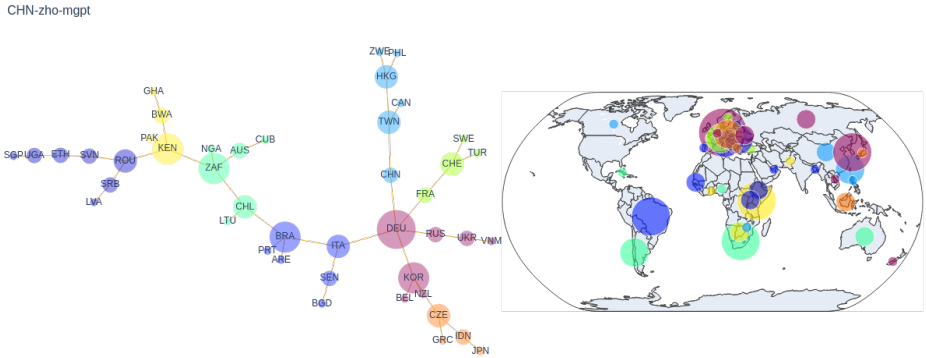
Figure 10: Geographic Representation Network and Corresponding Community Map for different Expert Unit set Associations. The language models we use are GPT2 (only English), mGPT and BLOOM.



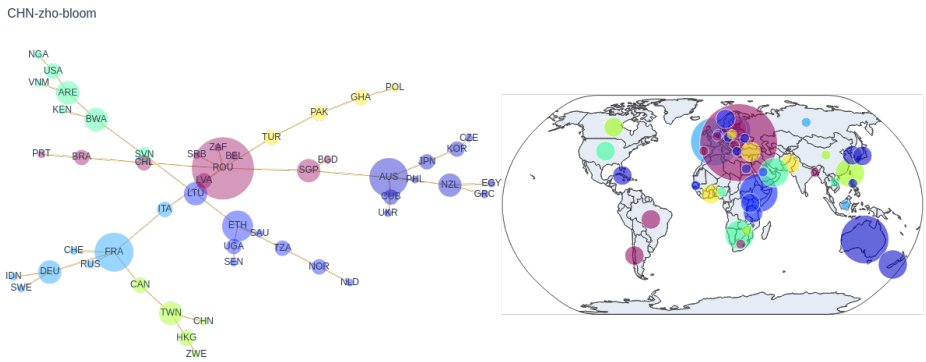
(13)



(14)



(15)

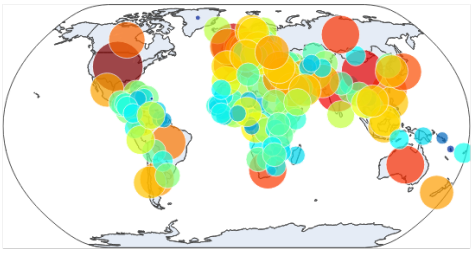


(16)

Figure 11: Geographic Representation Network and Corresponding Community Map for different Expert Unit set Associations. The language models we use are GPT2 (only English), mGPT and BLOOM.

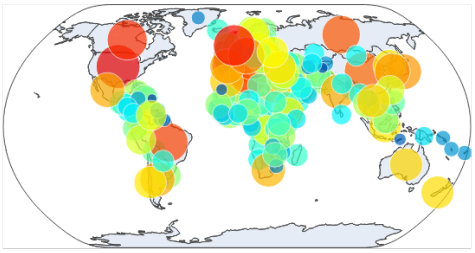
Geographic Representation Networks and Corresponding Community Maps

USA-eng-bloom



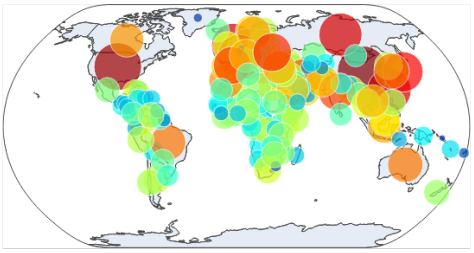
(a)

FRA-fra-bloom



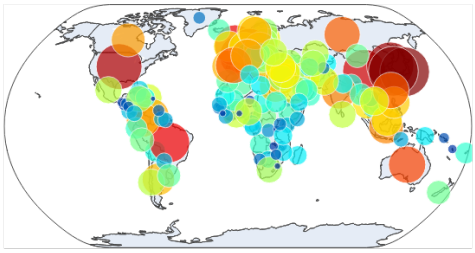
(b)

CHN-zho-bloom



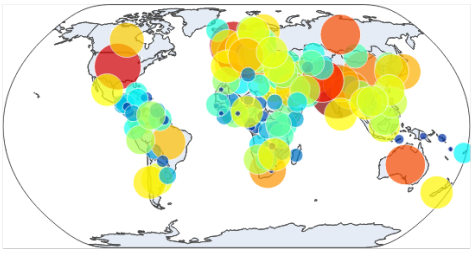
(c)

KOR-kor-bloom



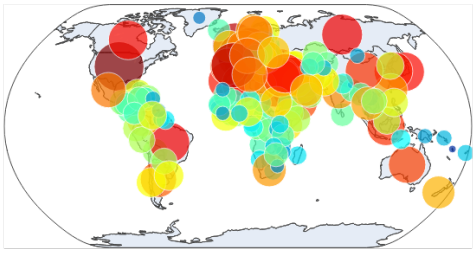
(d)

IND-hin-bloom



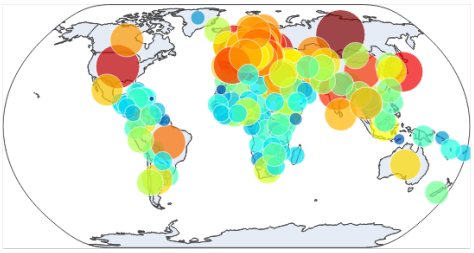
(e)

SAU-ara-bloom



(f)

RUS-rus-bloom



(g)

BGD-ben-bloom

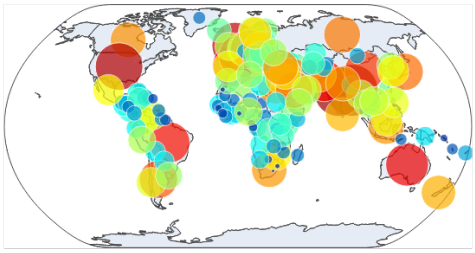


Figure 12: Graphs prepared using entity-country mapping on generated texts using BLOOM. Here We take the log-frequency distribution of entity counts. In all cases, the most frequent country remains the geopolitical favoured ones with the additon of Country/Concept Dataset News Source-country (the darker red ones)