GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models

Da Yin **Hritik Bansal** **Masoud Monajatipoor** **Liunian Harold Li** **Kai-Wei Chang**
Computer Science Department, University of California, Los Angeles
{da.yin,hbansal,liunian.harold.li,kwchang}@cs.ucla.edu,
monajati@ucla.edu

Abstract

Recent work has shown that Pre-trained Language Models (PLMs) store the relational knowledge learned from data and utilize it for performing downstream tasks. However, commonsense knowledge across different regions may vary. For instance, the color of bridal dress is *white* in *American* weddings whereas it is *red* in *Chinese* weddings. In this paper, we introduce a benchmark dataset, **Geo**-diverse Commonsense **M**ultilingual **La**nguage **M**odels **A**nalysis (GeoMLAMA), for probing the diversity of the relational knowledge in multilingual PLMs. GeoMLAMA contains 3,125 prompts in English, Chinese, Hindi, Persian, and Swahili, with a wide coverage of concepts shared by people from American, Chinese, Indian, Iranian and Kenyan cultures. We benchmark 11 standard multilingual PLMs on GeoMLAMA. Interestingly, we find that 1) larger multilingual PLMs variants do not necessarily store geo-diverse concepts better than its smaller variant; 2) multilingual PLMs are not intrinsically biased towards knowledge from the Western countries (the United States); 3) the native language of a country may not be the best language to probe its knowledge and 4) a language may better probe knowledge about a non-native country than its native country. Code and data are released at https://github.com/WadeYin9712/GeoMLAMA.

1 Introduction

Pre-trained Language Models (PLMs) (Peters et al., 2018; Radford et al., 2019; Devlin et al., 2019; Brown et al., 2020) are increasingly used in various Natural Language Processing (NLP) applications. Pre-trained on large-scale text corpora, they are shown to store relational knowledge (Petroni et al., 2019; Jiang et al., 2020b; Kassner et al., 2021), e.g., commonsense knowledge (Zhou et al., 2020; Lin et al., 2020; Nguyen et al., 2021; Zhou et al., 2021). They have been used to construct knowledge bases while requiring limited human effort for rule creation and validation (Bosselut et al., 2019; Zhou et al., 2022).

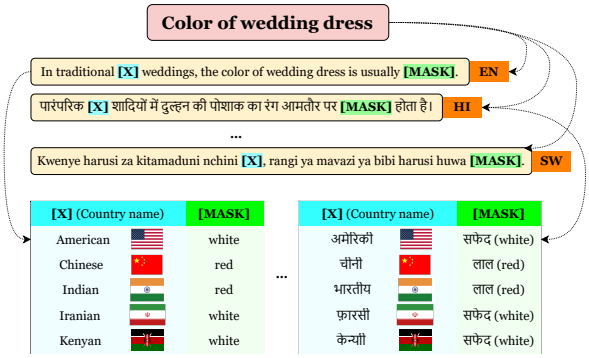


Figure 1: Examples of prompts and gold answers in GeoMLAMA. For each concept (e.g., color of wedding dress), there are multiple masked multilingual prompts (English, Hindi, Swahili, etc.) with specified country information [X] querying geo-diverse knowledge about the concept. We test multilingual PLMs by examining the extent to which masked word predictions align with the gold answers in [MASK] columns.

However, *do PLMs store geo-diverse commonsense knowledge?* Geo-diverse commonsense (Yin et al., 2021) is a collection of commonsense locally shared by people from certain regions but may not apply in other regions due to cultural and geographic differences. For instance, the color of bridal outfit in American wedding is white, while it is normally red in traditional Chinese and Indian weddings. PLMs which are unaware of geo-diverse knowledge may have disparity in performance on test data associated with different regions. This may lead to disadvantage of users in certain regions and further amplify bias in AI applications, such as constructing Western-centric knowledge bases eventually.

In this paper, we concentrate on evaluating *multilingual* PLMs (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020). Studying geo-diversity naturally involves multilinguality. People in different regions may speak different languages,

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2039 - 2055
December 7-11, 2022

and it is natural to assume that geo-specific knowledge is better represented in its native language. Moreover, pre-trained on a collection of multilingual corpora, multilingual PLMs accumulate the knowledge from various languages. Therefore, we posit that knowledge in multilingual PLMs is more diverse than that in models trained on a single language.

Centered around multilingual PLMs, we follow the original knowledge probing task LAnguage Model Analysis (LAMA) (Petroni et al., 2019) and introduce a new *geo-diverse* probing benchmark GEOMLAMA. As shown in Figure 1, given a masked geo-diverse prompt with a particular country name [X], such as "*In traditional* [X] *weddings, the color of wedding dress is usually* [MASK] .", and a corresponding candidate answer list, {"*red*", "*white*", "*black*", "*blue*", ...}, multilingual PLMs are required to predict the masked word [MASK] from the candidate list.

The characteristics of GEOMLAMA are summarized as follows. 1) *Diverse answers across countries*: Each prompt is designed based on geo-diverse concept (e.g., color of traditional wedding dress in Figure 1) and gold answers for masked word are different across countries. 2) *Broad coverage of geo-diverse concepts*: GEOMLAMA encompasses comprehensive geo-diverse topics including habits and personal choices, cultures and customs, policies and regulations, and geography. 3) *Coverage of multiple countries and languages*: GEOMLAMA involves knowledge about the United States, China, India, Iran, and Kenya, and is constructed by the native languages of the five countries, English, Chinese, Hindi, Persian, and Swahili. Overall, there are 3,125 prompts in our benchmark.

We perform in-depth probing analysis on 11 multilingual PLMs, including mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), and XGLM (Lin et al., 2021b). In general, we observe that multilingual PLMs significantly outperform random guess, suggesting that multilingual PLMs are capable of storing geo-diverse commonsense to some extent. We then conduct fine-grained investigation across three dimensions.

We first study the correlation between model performance and *model size*. Contrary to our intuition, we notice that the largest models do not necessarily have the best performance on our benchmark. We further study *the best language to probe the knowledge about a particular country*. Surprisingly, we find that the best language is not the native language of the given country (e.g., English is not the best language to probe knowledge about the US). We also explore *the knowledge that can be most accurately probed by a particular language*. Similarly, we find that the most accurately probed knowledge is not the one about indigenous country of the language (e.g., the country for which Chinese prompts provide the most accurate predictions is not always China). Lastly, we find evidence of reporting bias that might explain such observations.

2 Related Works

Knowledge Probing on PLMs. Petroni et al. (2019) first explore whether PLMs have capacity of storing factual knowledge about entities. Based on this observation, prior works involving knowledge probing focus primarily on creating more effective probing methods to elicit factual knowledge (Jiang et al., 2020b,a; Shin et al., 2020; Zhong et al., 2021) or analyzing whether other types of knowledge are stored in PLMs (Talmor et al., 2020; Zhou et al., 2020; Kassner et al., 2021; Sung et al., 2021). In the second line of works, there is a great variety of commonsense knowledge being explored, including social (Zhou et al., 2020), numerical (Lin et al., 2020) and spatial (Zhang et al., 2020; Liu et al., 2022) commonsense. GEOMLAMA focuses on probing a new commonsense type, geo-diverse commonsense, on multilingual PLMs.

Multilingual Knowledge Probing and Multilingual Commonsense. MLAMA (Kassner et al., 2021) and Prix-LM (Zhou et al., 2022) simply focus on capturing multilingual factual knowledge about entities. XCOPA (Ponti et al., 2020) and X-CSR (Lin et al., 2021a) are two multilingual commonsense benchmarks, but both are built by translation from English commonsense benchmarks, without any consideration of region-specific commonsense. Different from prior works, we value geo-diversity and quantify the extent to which multilingual PLMs master such geo-diverse commonsense.

Geo-Diverse Commonsense. Geo-diverse commonsense is strongly correlated with cultures and geographic locations. There have emerged a few works (Acharya et al., 2020; Yin et al., 2021; Liu et al., 2021; Shwartz, 2022) studying geo-diverse commonsense. Specifically, by collecting responses to questionnaire, Acharya et al. (2020)

analyze the cultural difference between US and India about scenarios including wedding and funeral. Yin et al. (2021); Liu et al. (2021) propose geo-diverse multimodal benchmarks, GD-VCR and MaRVL. They find that due to lack of geo-diverse knowledge, large performance disparity appears when multimodal models are applied on tasks requiring knowledge about Western and non-Western regions. Shwartz (2022) propose culture-specific time expression grounding task to acquire specific temporal commonsense in different countries from multilingual corpora and models.

Inclusion in NLP. Enhancing inclusivity of language processing technology and ensuring it works for everyone is essential. Several studies have focused on improving language inclusion (Joshi et al., 2020; Faisal et al., 2022), gender inclusion (Cao and Daumé III, 2021; Dev et al., 2021; Lauscher et al., 2022), and race inclusion (Field et al., 2021). We hope that GEOMLAMA can enable future development in improving the diversity of knowledge embedded in pre-trained language models.

3 GEOMLAMA Benchmark Construction

To build a geo-diverse commonsense probing benchmark, we recruit annotators from five different countries, the United States, China, India, Iran, and Kenya to participate in annotation. The annotation process is separated into four stages. 1) We first ask the annotators to list geo-diverse concepts. 2) Based on the collected concepts, we then require annotators to design masked geo-diverse prompt templates in English. 3) After specifying prompts with country names, we request annotators to provide correct answers and form answer candidate list for each prompt. 4) We translate the English prompts into other languages and paraphrase them. The overview of the annotation pipeline is illustrated in Figure 2.

3.1 Geo-Diverse Concept Collection

Geo-diverse concepts are the foundation of designing geo-diverse prompts. The criteria of selecting geo-diverse concepts are shown as follows:

Universality and Diversity across Cultures. We require that the scenarios regarding the collected concepts to be universal but diverse across the different cultures. "*Color of wedding dress*" qualifies our criteria as *wedding dress* is a universally understood entity where its color is diverse across different cultures.

Avoiding Concepts involving Region-Specific Terms. We avoid probing models about region-specific factual knowledge, e.g., festival names and president names of the countries, as these concepts usually involve uncommonly used tokens in certain languages and thus introduce another layer of complexity to make inference.

Finally, we consider topics that cover habits and personal choices, cultures and customs, policies and regulations, and geography for subsequent annotations. Details are shown in Appendix A.

3.2 Geo-Diverse Prompt Template Design

Centered on the collected geo-diverse concepts, annotators design English version of geo-diverse prompt templates that will be later paraphrased and translated into multilingual prompts. Given one geo-diverse concept, e.g., "*color of wedding dress*", the corresponding prompt template would be a masked sentence that inquires the missing color information, e.g., "*The color of wedding dress is usually* [MASK] ." Since we intend to probe knowledge about different countries using these prompts, we further insert phrases such as "*In* [X] , ", "*In traditional* [X] *wedding, "* to indicate the country knowledge to be probed. Here [X] is either one of the country names (the United States, China, India, Iran, and Kenya), or one of the corresponding modifiers (American, Chinese, Indian, Iranian, and Kenyan).

3.3 Answer and Answer Candidate List Annotation

For each masked geo-diverse prompt with a specified country name, we request the annotators to provide correct answers for the masked words. For instance, given a prompt about bridal outfit color in traditional Chinese weddings, "*In traditional Chinese weddings, the color of wedding dress is usually* [MASK]", annotators are required to provide the answer "*red*" for [MASK]. The answers are all provided by annotators who are familiar with the culture in one of our studied countries. Note that besides prompts with only one answer, some other prompts in GEOMLAMA, such as "*The staple food in Iran is* [MASK]", can have *multiple* correct answers ("*rice*" and "*bread*") for a single prompt. To further validate the correctness of answers, we distributed a survey to collect responses for knowledge about respondents' own countries.

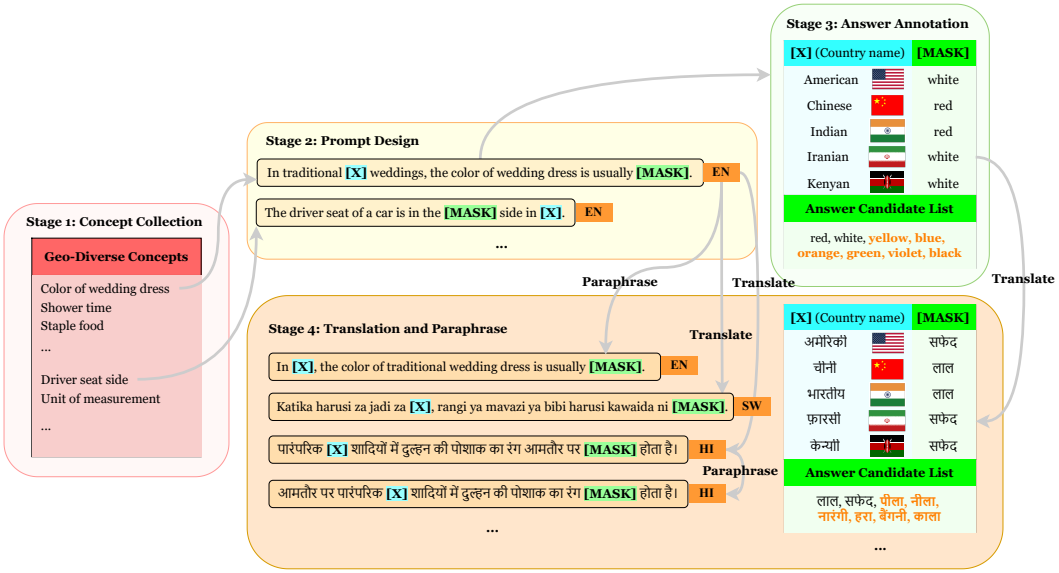


Figure 2: Overall annotation pipeline. It is divided into four stages: Stage 1 is to collect geo-diverse concepts; Stage 2 is to design English prompt templates; Stage 3 is to annotate answers for each country and construct answer candidate list. Stage 4 is to translate the English prompts and paraphrase the translated multilingual prompts. Here we showcase English and Hindi answer annotations for demonstration.

We collected 33 responses from the five countries, and retained the answers with majority support.

In this work, we focus on investigating whether PLMs are capable of predicting correct answers among all the possibilities of different countries. For example, we wonder if PLMs can predict the dress color at Chinese wedding is "*red*" over the other possibility, such as "*white*". Therefore, we pair each prompt with an additional answer candidate list composed by the probable choices and multilingual PLMs are constrained to make predictions from the list. Specifically, each list contains the union of all correct answers of five countries and additional confounding candidates sharing the same word types with those correct answers. For the prompts about color of wedding dress, the union of correct answers is {"*red*", "*white*"}. Other than the two colors, as illustrated in Figure 2, we also append confounders such as, "*yellow*", "*black*", "*blue*" to the list (the orange letters in grids titled with "Answer Candidate List"). The final answer candidate list for prompts about color of wedding dress will be {"*red*", "*white*", "*yellow*", "*black*", "*blue*", ...}. Note that the contents and lengths of answer candidate lists for prompts about different concepts vary greatly.

3.4 Prompt Translation and Paraphrase

We then obtain multilingual geo-diverse prompts via translating the annotated English prompts into four other languages Chinese, Hindi, Persian, and Swahili. We leverage Google Translation API to translate English prompts and each translated prompt is manually checked and corrected by annotators familiar with both English and any of the four studied languages. Besides, since it is shown that probing results are sensitive to small perturbation to the prompts (Jiang et al., 2020b), we further generate four paraphrases for each prompt to obtain more robust probing results. Specifically, we paraphrase English prompts via a round of backtranslation[1] in which we first translate English prompts to German ones and then translate them back to English. For prompts in other languages, their paraphrases are generated by backtranslation that translates texts to English and translate them back to the original languages. The paraphrases in a particular language are validated and modified by native speakers.

In total, we annotate 3125 prompts with answers and corresponding candidates in GeoMLama. All the prompts are designed based on 16 geo-diverse concepts listed in Appendix A, and there are 625 prompts for each of the five languages. More details are described in Appendix B.

4 Probing Methods on GeoMLama

Petroni et al. (2019) introduce the LAnaguage Model Analysis (LAMA) setup to probe knowl-

[1]Based on Google Translation API.

edge stored in the pre-trained language models using masked templates. Without any additional fine-tuning, given a masked prompt, models are required to recover masked tokens with entities with the highest probability for the prompt context. Following LAMA probe, on GEOMLAMA, we study whether models are capable of seeking the most appropriate answers to from answer candidate list according to given geo-diverse prompts.

Kassner et al. (2021) follow LAMA probe to investigate entity knowledge in multilingual BERT only. In this work, we probe a diverse set of language models on *geo-diverse commonsense knowledge* by scoring answer candidates and calibrating the score of each candidate.

4.1 Scoring Answer Candidates

We score answer candidates based on log likelihood of generating answer candidates given prompts. Different model families have their individual inference methods to obtain the scores. In the following, we introduce the probing methods for masked language models. Details of other probing methods on autoregressive and encoder-decoder language models are shown in Appendix C.

Masked Language Models (mBERT, XLM, XLM-R family). Given an answer candidate e (e.g., "*chopsticks*") that is tokenized into subtokens $e_1, e_2, ..., e_L$ (e.g., "*chop*", "*stic*", "*ks*") such that $e_i \in V$ where V is the vocabulary and t is the prompt (e.g., "*In China, people usually eat food with* [MASK$_1$]...[MASK$_L$]."), we assign a score l_e based on the log probability of recovering the answer candidate e in the masked prompt. Formally, l_e is defined as

$$\frac{1}{L}\sum_{i=1}^{L}\log(p([\text{MASK}_i] = e_i|[\text{MASK}_{<i}] = e_{<i}, t)). \quad (1)$$

According to Eq.(1), we perform L forward passes, each of which helps in obtaining conditional probability of generating one subtoken. To illustrate, i^{th} forward pass inference would be $p([\text{MASK}_i] = e_i|$ "*In China, people usually eat food with* e_1 e_2 ...e_{i-1} [MASK$_i$]...[MASK$_L$]").

Here we further normalize the sum of log likelihood by the number of subtokens L to help in reducing the effect of length. The other model families discussed in Appendix C also adopt the normalization strategy.

4.2 Calibrating Answer Candidates

The way to score answer candidates $e \in \mathcal{E}$ (e.g., "*chopsticks*" $\in$ {"*chopsticks*", "*hands*", "*spoons*", "*knives*"}) given the prompt t for a country C (e.g., "*In China, people usually eat food with* [MASK].") is illustrated in §4.1. However, this scoring mechanism is likely to be biased towards statistical correlations learned during pre-training (Zhao et al., 2021) whilst ignoring the country-specific information present in the prompt. For instance, the model might choose "*knives*" over "*chopsticks*" because "*knives*" may occur more often than "*chopsticks*" in pre-training corpora. Hence, we calibrate models with the prior probability of answer predictions in the absence of any country information. The final score given to each answer in the answers candidate set is given by:

$$s_e = l_e - l'_e, \quad (2)$$

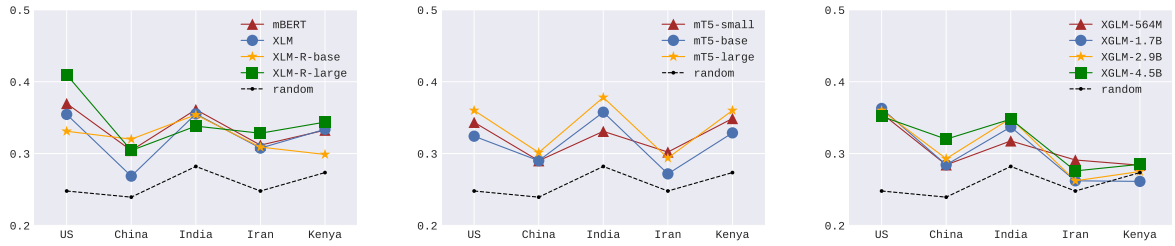
where l'_e is obtained using the same approach as l_e but the input prompt for calculating l'_e is the one without country information (e.g., "*People usually eat food with* [MASK]." without "*In China,*").

4.3 Evaluation Metric

We use the ratio of total number of model's correct predictions to the total number of gold answers as model performance on GEOMLAMA. Specifically, given a prompt t_i with g_i gold answers, we count the number of top-g_i model predictions that also appear in the gold answer list as c_i, based on the final score in Eq.2. For example, since there are two gold answers for the prompt "*The staple food in Iran is* [MASK]", "*rice*" and "*bread*", $g_i = 2$. In total, there are eight candidates in the answer candidate list {"*bread*", "*noodles*", "*rice*", "*meat*", "*maize*", ...} for this prompt. Assume one multilingual PLM assigns the highest g_i scores to the candidates "*noodles*" and "*rice*". Then $c_i = 1$, since only one of "*noodles*" and "*rice*" is the gold answer of the prompt. We then sum up all c_i and g_i to calculate the ratio, $\sum_{i=1}^{n} c_i / \sum_{i=1}^{n} g_i$, where n is the total number of prompts in GEOMLAMA.

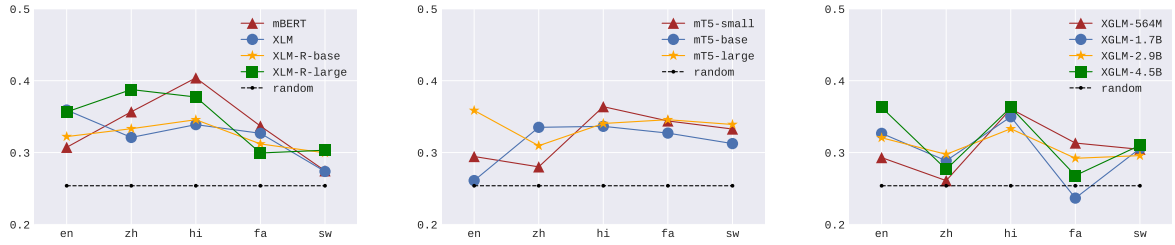
5 Analysis on Multilingual PLMs

In this section, we are interested in analyzing following questions: 1) Are bigger multilingual PLMs more geo-diverse than smaller ones? 2) In the absence of any particular country information in the prompts, are multilingual PLMs biased towards the knowledge towards certain countries? 3) Can



(a) mBERT, XLM, XLM-R family. (b) mT5 family. (c) XGLM family.

Figure 3: Multilingual PLMs' performance on probing knowledge about the studied countries averaged over all languages. Complete results are shown in Appendix E.



(a) mBERT, XLM, XLM-R family. (b) mT5 family. (c) XGLM family.

Figure 4: Multilingual PLMs' performance averaged over countries when using multilingual prompts. "*en*", "*zh*", "*hi*", "*fa*", and "*sw*" denote English, Chinese, Hindi, Persian, and Swahili. Complete results are shown in Appendix E.

native language probe the knowledge about a particular country best? 4) Given a particular language, can the corresponding country's knowledge be most accurately probed by the language?

To this end, we experiment with 11 multilingual PLMs[2] including mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R family[3] (Conneau et al., 2020), mT5 family[4] (Xue et al., 2021), and XGLM family[5] (Lin et al., 2021b). We freeze pre-trained model parameters provided by HuggingFace Transformers (Wolf et al., 2020) and do not fine-tune the models during probing.

5.1 Overview of Model Performance

Results are shown in Figure 3 and 4. Figure 3 focuses on the comparison among performance of probing the knowledge about a particular country while Figure 4 compares the performance of using prompts in different languages.

In Figure 3, we find that the performance of nearly all the multilingual PLMs lies in the range of 30% to 40% on probing each country's knowledge. Further, these multilingual PLMs significantly outperform random guess 2-15%. It implies that multilingual PLMs can store geo-diverse commonsense knowledge and some stored knowledge can be accurately elicited even if we merely change the country names in the prompt.

As illustrated in Figure 4, we observe that the performance of using prompts in different languages is generally from 30% to 40% and higher than random guess 2-15% as well. Moreover, we find that English and Hindi prompts are the most effective ones to probe geo-diverse knowledge, while Persian and Swahili prompts cannot achieve comparable results. In particular, from Figure 4c, using Persian prompts to probe XGLM-1.7B leads to worse performance than random guess.

5.2 Effect of Model Size

According to Petroni et al. (2019); Roberts et al. (2020), bigger models can generally store more knowledge and achieve better performance on downstream NLP tasks such as open-domain QA (Joshi et al., 2017; Kwiatkowski et al., 2019). To this end, we investigate whether larger models indeed perform better than the smaller ones on GEOMLAMA. For a fair comparison, we only compare models in the same model families.

[2]We also experiment with GPT-3 as it is also pre-trained on multilingual corpora. However, the results are not included in main paper because GPT-3 probing convention does not adopt cloze statements as the other 11 multilingual PLMs do. More setup details and results can be found in Appendix D.

[3]XLM-R-base, XLM-R-large.

[4]mT5-small, mT5-base, mT5-large.

[5]XGLM-564M, XGLM-1.7B, XGLM-2.9B, XGLM-4.5B.

This avoids comparing models with different pre-training corpora and learning objectives.

The comparison results over the three model families are shown in Figure 3 and 4. We observe that the larger models only perform marginally better than their smaller counterparts on GeoMLAMA. For the three model families, XLM-R, mT5, and XGLM, the performance gap between the largest and smallest models on all the prompts in GeoMLAMA is merely 2.23%, 2.42%, and 1.46%, respectively. In specific cases (e.g., probing XGLM family using Persian prompts), the largest model can be even worse than its smallest variant. It demonstrates that even if large models have nearly an order of magnitude more parameters than small models, large models cannot store geo-diverse commonsense significantly better than small models. This highlights that GeoMLAMA is a challenging task and being better on the standard multilingual NLP tasks does not guarantee good performance.

5.3 Intrinsic Model Bias without Country Information

Each prompt in GeoMLAMA consists of the country information. However, it is still not clear as to what information is probed innately when we query multilingual PLMs without any country information. To study this phenomenon, we further probe multilingual PLMs with the prompts where the country token is removed. For example, instead of "*In traditional* Kenyan *weddings, the color of wedding dress is usually* [MASK]", we implement a new round of probing with the pruned prompt, "*In traditional weddings, the color of wedding dress is usually* [MASK]". The new prompts can elicit the knowledge that multilingual PLMs are intrinsically inclined towards predicting.

As shown in Figure 5, we find that for most multilingual PLMs, the knowledge about India is captured frequently in the absence of any country information. Whereas, knowledge about the United States is not well probed. It shows that at least, multilingual PLMs are not originally biased towards knowledge about Western countries like US.

We do a quantitative case study to further explain the phenomenon. We take a geo-diverse concept "*staple food*" as an example. Rice and bread are the staple foods in China and the United States, respectively. According to Table 2, in English, Chinese and Swahili Wikipedia, we find that the co-occurrence of "*staple food*" and "*rice*" is comparable or even way higher than "*staple food*" and "*bread*". It demonstrates that the popularity of Western knowledge across the world does not necessarily mean higher frequency in knowledge sources like Wikipedia. This may lead the models to predicting non-Western knowledge more precisely.

Models	US	China	India	Iran	Kenya
mBERT	fa	sw	en	fa	zh
XLM	fa	en	en	zh	zh
XLM-R-base	fa	zh	zh	fa/sw	en
XLM-R-large	fa	zh	en	en	zh
mT5-small	fa	en	en	sw	sw
mT5-base	fa	en	zh	hi	sw
mT5-large	fa	sw	sw	fa	hi
XGLM-564M	fa	en	sw	fa	fa/hi
XGLM-1.7B	fa	sw	en	fa	fa
XGLM-2.9B	fa	en	en	hi	fa
XGLM-4.5B	fa	zh	en	fa	en
Best Languages	**fa**	**en**	**en**	**fa**	**zh/fa**

Table 1: Best languages to probe each country's knowledge. Each language in the last row "**Best Languages**" is the one appearing most in its located column.

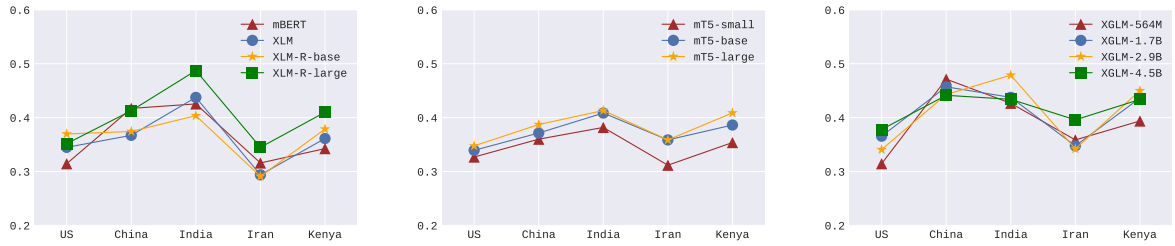
Words	English	Chinese	Swahili
rice, staple food	1040	33	7
bread, staple food	33	37	1

Table 2: Word co-occurrence of "*rice*", "*bread*" and "*staple food*" in English, Chinese and Swahili Wikipedia, respectively.

5.4 Best Languages to Probe Knowledge about Countries

In GeoMLAMA, prompts in different languages are used to probe knowledge about different countries. It is imperative to ask whether we elicit most knowledge about a country if we query the PLM with its native language. From Table 1, contrary to our intuition, the native language is not the best language to query its knowledge for most of the countries. In particular, Iran is the only country for which its native language Persian can help in drawing out maximum knowledge about it. For the United States and Kenya, the best probing language is Persian and for China and India, the best language is English.

We speculate that our observations might be attributed to the reporting bias phenomenon (Grice, 1975; Gordon and Van Durme, 2013). It is categorized by people rarely stating the obvious knowledge that is shared by everyone (commonsense) explicitly in the text. For instance, the fact that



(a) mBERT, XLM, XLM-R family. (b) mT5 family. (c) XGLM family.

Figure 5: Average performance of multilingual PLMs when fed with prompts without any specified country names. Complete results are shown in Appendix F.

Models	en	zh	hi	fa	sw
mBERT	India	India	US	US	China
XLM	India	Kenya	India	US	Kenya
XLM-R-base	India	China	India	US	India
XLM-R-large	India	US	US	US	Kenya
mT5-small	India	Kenya	Kenya	US	Kenya
mT5-base	India	India	Kenya	US	Kenya
mT5-large	India	Kenya	Kenya	US	India
XGLM-564M	China	US	India/Kenya	US	India
XGLM-1.7B	India	India	India	US	India
XGLM-2.9B	India	India	US	US	India
XGLM-4.5B	India	US/China/India	India	US	India
Best Countries	**India**	**India**	**India**	**US**	**India**

Table 3: Countries best probed with prompts in different languages. Each country in the last row "**Best Countries**" is the one appearing most in its located column.

Words	Freq. of Co-occur	# Co-occur
rice, staple food, China	3.6x	25
bread, staple food, US	1x	7
米饭**(rice)**, 主食**(staple food)**, 中国**(China)**	3.2x	3
面包**(bread)**, 主食**(staple food)**, 美国**(US)**	3.2x	3

Table 4: Word co-occurrence and frequency in English and Chinese Wikipedia. English Wikipedia has 72484142 sentences, 7.6 times more than those of Chinese Wikipedia, 9502859 sentences. 'nx' denotes the frequency rate is n times higher than the lowest one.

all the *humans can murder* is disproportionately over-reported than *humans can breathe* in the English text. This unbalanced frequency would lead to bias towards acquiring uncommon event knowledge from PLMs, instead of commonsense knowledge (Shwartz and Choi, 2020). In our setting, we believe that reporting bias is a key ingredient in explaining our observed trends. For instance, indigenous population is less likely to record obvious facts about their culture in their native language texts as compared to the facts from other cultures. For example, when mentioning the driver seat side in India, compared with people living in other countries, Indian people will not talk too much about this because it is too trivial for them.

We seek a quantitative evidence in the context of *staple food* as a concept again to support our claim. Throughout the English and Chinese Wikipedia corpora, we count the co-occurrence of words "*China*", "*rice*" and "*staple food*", and "*the United States*", "*bread*" and "*staple food*" in their respective languages. The counting results are shown in Table 4. We notice that when China is mentioned, English words "*rice*" and "*staple food*" co-occur 25 times whereas it is mentioned merely 3 times in Chinese Wikipedia. Furthermore, in the context of the US, English words "*bread*" and "*staple food*" appear 7 times simultaneously while Chinese words "面包(*bread*)" and "主食(*staple food*)" co-occur 3 times. Although the number of co-occurrence is higher in the English Wikipedia, the frequency rate of the Chinese word co-occurrence is 3.2 times higher, since the Chinese Wikipedia corpus is 7.6 times smaller than the English corpus. In summary, it shows that commonsense knowledge about a country is not mentioned more frequently in its native language corpus but might have higher occurrences in some other languages.

5.5 Countries Best Probed with Prompts in Different Languages

Apart from the best languages to probe knowledge about countries, conversely, we can also study the countries best probed with prompts in different languages. Specifically, we focus on the following question: Given one studied language X, is the country best probed the same as the indigenous country of language X?

We present our results in Table 3. We observe that except Hindi, the countries best probed are distinct to the corresponding countries of language. For example, Swahili prompts probe Indian knowledge best instead of Kenya, and Persian prompts

probe US knowledge best instead of Iran. It is also counter-intuitive because it is natural for people to imagine that the best probed country should be the one where a particular language is spoken most commonly.

We can also ascribe the phenomenon observed for Q2 to the reporting bias. To analyze this observation, we compare the occurrence of knowledge about different countries in the same language corpus. We find that English words "*bread*", "*staple food*" and "*the United States*" co-occur much less frequently than "*rice*", "*staple food*" and "*China*". Besides, Chinese words "面包(*bread*)", "主食(*staple food*)" and "美国(*the United States*)" co-occur 3 times, which is the same as co-occurrence of "米饭(*rice*)", "主食(*staple food*)" and "中国(*China*)". The comparison results indicate that given one language, local country's knowledge may not appear the most, compared with knowledge about other countries.

6 Conclusions

We propose a knowledge probing benchmark, GEOMLAMA, to evaluate the extent of multilingual PLMs to store geo-diverse commonsense. Results show that multilingual PLMs can achieve significantly higher performance than random guess, suggesting that they are capable of storing geo-diverse knowledge. We also find that fed with prompts without any country cues, multilingual PLMs are not intrinsically biased towards knowledge about the United States. We further investigate the best language to probe the knowledge about a particular country, and the country best probed with prompts in a certain language. Surprisingly, we notice that the best language is not the country's native language, and the best probed country is not the indigenous country of the language. We connect this to reporting bias issue in geo-diverse context: one country's commonsense is seldom recorded in the text by people living in that country as it is too trivial and not worth mentioning for them.

Acknowledgement

We thank annotators for tremendous efforts on annotation and evaluation. We also greatly appreciate Tao Meng, Xiao Liu, Ashima Suvarna, Ming Zhong, Kuan-Hao Huang, I-Hung Hsu and other members of UCLA-NLP group for their helpful comments. This work was partially supported by NSF IIS-1927554, Sloan Research Fellow, Amazon AWS credits, Amazon Fellow, and a DARPA MCS program under Cooperative Agreement N66001-19-2-4032. The views and conclusions are those of the authors and should not reflect the official policy or position of DARPA or the U.S. Government.

Limitations

GEOMLAMA is proposed for evaluating the degree of potential geographic bias in multilingual PLMs. However, due to the limited coverage of countries, languages and geo-diverse concepts, GEOMLAMA may introduce unwanted bias. In GEOMLAMA, we only consider five countries and their native languages, which merely occupy a tiny portion of all the countries in the world and thousands of languages. Also, in countries like India, there are multiple commonly used languages, we limit our study on Hindi and will extend to more languages to study the phenomenon. Besides, we design prompts simply based on 16 general geo-diverse concepts. The extension on existing GEOMLAMA can help in obtaining more solid results and mitigating bias against uncovered countries and languages.

In this work, we mainly focus on evaluating multilingual PLMs on GEOMLAMA without studying how multilingual pre-training process affects the model performance on geo-diverse commonsense probing. We intend to explore effect of the process on model's geo-diversity in future work. Specifically, we aim to examine whether pre-training on multilingual corpora really brings more geo-diversity than pre-training on monolingual corpora does. Besides, we do not cover how to improve model performance on GEOMLAMA and other related tasks. We expect to seek approaches to improving model's geo-diversity while maintaining multilingual PLMs' performance on various multilingual benchmarks in future work as well.

Ethical Consideration

As we propose a new benchmark in this paper, we provide details about compensation rate for annotators. We recruit five countries' college students and annotators from Amazon MTurk. We provide a fair compensation rate with $12 per hour and in total around $150 to the annotators on both prompt design, translation and evaluation. Note that part of annotations are done by the authors of this work.

References

A. Acharya, Kartik Talamadupula, and Mark A. Finlayson. 2020. An Atlas of Cultural Commonsense for Machine Reasoning. *ArXiv*, abs/2009.05664.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47(3):615–661.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-Lingual Language Model Pretraining. *Advances in Neural Information Processing Systems*, 32.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Herbert P Grice. 1975. Logic and Conversation. In *Speech Acts*, pages 41–58. Brill.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural

Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021b. Few-shot Learning with Multilingual Language Models. *arXiv preprint arXiv:2112.10668*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not Written in Text: Exploring Spatial Commonsense from Visual Signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *Proceedings of the Web Conference 2021*, pages 2636–2647.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

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.

Vered Shwartz and Yejin Choi. 2020. Do Neural Language Models Overcome Reporting Bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can Language Models be Biomedical Knowledge Bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do Language Embeddings capture Scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho, Jay Pujara, and Xiang Ren. 2021. RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7579, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. 2022. Prix-LM: Pretraining for Multilingual Knowledge Base Construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5412–5424, Dublin, Ireland. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.

Appendix

A Geo-Diverse Concept List

The general geo-diverse concepts are shown in Table 6. We summarize all the concepts into 16 general ones, covering rules, policies, geography, customs, personal choices and habits. Multiple prompts can be designed for each geo-diverse concept. For example, measurement units can involve units measuring height, weight and temperature, and thus annotators can create multiple prompts about various types of measurement units.

B Statistics of GeoMLama

Table 5 shows the statistics of GeoMLama. In total, there are 3125 prompts in GeoMLama, 625 prompts about each country's knowledge. We also manifest the average numbers of gold answers and corresponding answer candidates for prompts regarding each country. Overall, the number of gold answers is 1.20 per prompt, with answer candidate list of average length 4.76. Here note that for prompts under the same topic (e.g., "*In traditional* [X] *weddings, the color of wedding dress is usually* [MASK]."), regardless of the exact country filled in [X], the answer candidate lists are the same for all the five countries. Therefore, the average length of answer candidates is identical to all the studied countries.

Countries	# Prompts	# Avg. Gold Answers	# Avg. Answer Candidates
US		1.16	
China		1.12	
India	625	1.32	4.76
Iran		1.16	
Kenya		1.25	
Overall	3125	1.20	4.76

Table 5: Detailed statistics of GeoMLama.

C Details of Evaluation Methods on Autoregressive and Encoder-Decoder Language Models

Autoregressive Language Models (XGLM family). For autoregressive language models such as XGLM, we first replace masked token in the prompt with answer candidate tokens (e.g., "*In China, people usually eat food with* [MASK]."->"*In China, people usually eat food with chopsticks.*"). The joint probability of generating all the tokens in the complete sentence is used for scoring answer candidates. Given a prompt template t filled with an answer candidate e, t is tokenized into K tokens (e.g., t_1, t_2, ..., t_K). We assign score l_e to the answer candidate as:

$$l_e = \frac{1}{K}\sum_{i=1}^{i=K}\log(p(t_i|t_{<i})). \quad (3)$$

Here, we perform K forward passes to the autoregressive language model to obtain log probability of generating the whole sentence with the answer candidate e. In this case, the i^{th} forward pass inference would calculate $p(t_i|$ "*In China, ...,* t_{i-2} t_{i-1}"$)$.

Encoder-Decoder Language Models (mT5 family). During pre-training of encoder-decoder language models mT5, a masked sequence is input to encoder, and decoder learns to recover the L masked tokens in autoregressive fashion. Therefore, we input a masked prompt t into the models (e.g., "*In China, people usually eat food with* [MASK]") and calculate the score for answer candidate e as:

$$l_e = \frac{1}{L}\sum_{i=1}^{i=L}\log(p(e_i|e_{<i}, t)). \quad (4)$$

Computing Eq.4 requires L forward passes, since the decoder needs to generate L tokens. Here i^{th} forward pass inference would be $p(e_i|e_1\ e_2...e_{i-1},$ "*In China, people usually eat food with* [MASK]"$)$. Note that mT5 can use one single [MASK] token to represent multiple consecutive masked tokens. Thus, different from masked language models, mT5 models are simply fed with the prompt with only one [MASK] token instead of L [MASK] tokens.

D Evaluating GPT-3 on GeoMLama

Approach to probing GPT-3 is different from the methods mentioned in §4. Instead of feeding declarative prompt sentences, we leverage Question Answering (QA) API empowered by GPT-3 and input questions to query the knowledge. For example, instead of using "*In traditional Chinese weddings, the color of wedding dress is usually* [MASK]", we first convert it to question form like "*What is the color of wedding dress in an American wedding?*" and query GPT-3 with the converted question. During evaluation stage, rather than scoring answers from given answer candidate list, GPT-3 can generate open-ended answers and we evaluate GPT-3 predictions using the same metric in §4.3. Considering the huge time cost of manually inputting

Categories	Concepts
rules, policies, geography	traffic rules measurement units date formats color of stock price climate
customs, personal choices, habits	payment shower time clothes drying broom usage food and drink family popular sports transportation servant wedding funeral

Table 6: Geo-diverse concept list with categorization.

questions by annotators to GPT-3 API, we do not convert paraphrased prompts to questions and perform analysis on them. In other words, the number of tested questions is only $1/5$ out of the total number of prompts in GEOMLAMA, which is 625.

We probe GPT-3 with the converted questions in five languages, each of which asks knowledge about the five studied countries. Final results are shown in Table 7. One notable result is that using English prompts can achieve nearly 60% performance, while using Swahili prompts cannot solve any questions correctly. Also for Hindi and Persian prompts, the results are still extremely low, ranging from 0% to 25%. It exposes strong bias in terms of language usage. When looking at the performance of probing knowledge about respective countries, the disparity is not large. The country that can be best probed is the United States, while the worst probed country only underperforms the United States 6.9%.

E Detailed Results of Multilingual PLMs on GEOMLAMA

Table 8, 9, and 10 show the details of each multilingual PLM's performance on GEOLAMA. The performance of random guess depends on the expectation of correct predictions, which is equivalent to the ratio of total number of gold answers to the total number of answers in the answer candidate lists. Since the number of gold answers and answer candidates is different for knowledge about different countries, the random guess performance is not the same across countries. However, prompts in each of the languages have the same number of gold answers and candidate answers, so random guess performance is identical across languages.

Languages	US	China	India	Iran	Kenya	Average
en	68.97	57.14	54.55	55.17	65.52	50.23
zh	44.83	50.00	39.39	37.93	31.03	40.64
fa	20.69	21.43	24.24	10.34	17.24	18.79
hi	6.90	0.00	12.12	3.45	20.69	8.63
sw	0.00	0.00	0.00	0.00	0.00	0.00
Average	28.28	25.71	26.06	21.38	26.90	25.67

Table 7: GPT-3 performance (%) on GEOMLAMA.

F Detailed Results of Multilingual PLMs Probed with Prompts without Country Tokens

Table 11, 12, and 13 show the details of each multilingual PLM's performance when input with prompts lacking specified country information. It can help in determining the intrinsic bias of each multilingual PLM.

Languages	Countries	mBERT	XLM	XLM-R-base	XLM-R-large
en	**US**	31.03	26.21	30.34	33.10
	China	30.00	39.29	34.29	37.14
	India	40.61	52.12	37.58	37.58
	Iran	21.38	27.59	28.28	37.93
	Kenya	30.63	34.38	30.63	32.50
zh	**US**	35.17	28.28	30.34	46.21
	China	30.71	28.57	46.43	40.00
	India	38.79	32.12	38.18	35.15
	Iran	32.41	36.55	24.14	33.10
	Kenya	41.25	35.00	27.50	39.38
fa	**US**	48.97	57.93	48.28	53.79
	China	27.86	20.71	28.57	32.14
	India	38.79	27.88	33.33	34.55
	Iran	47.59	31.03	35.17	33.79
	Kenya	38.75	31.87	27.50	34.38
hi	**US**	42.07	40.00	33.10	42.07
	China	29.29	22.86	18.57	13.57
	India	34.55	35.76	36.36	32.73
	Iran	33.79	31.03	31.72	27.59
	Kenya	28.75	33.75	36.25	33.75
sw	**US**	27.59	24.83	23.45	29.66
	China	34.29	22.86	32.14	29.29
	India	27.88	29.70	31.52	29.09
	Iran	20.69	27.59	35.17	31.72
	Kenya	26.88	31.87	27.50	31.87

Table 8: Results (%) of mBERT, XLM, XLM-R-base, and XLM-R-large on GEOMLAMA.

Languages	Countries	mT5-small	mT5-base	mT5-large
en	**US**	24.14	18.62	30.34
	China	40.71	34.29	39.29
	India	41.21	34.55	49.09
	Iran	19.31	19.31	26.21
	Kenya	21.88	23.75	34.38
zh	**US**	20.00	33.79	28.97
	China	26.43	26.43	26.43
	India	23.64	46.06	33.33
	Iran	33.10	26.90	31.03
	Kenya	36.88	34.38	35.00
fa	**US**	55.86	43.45	48.28
	China	31.43	29.29	22.86
	India	36.36	34.55	30.30
	Iran	28.28	30.34	33.79
	Kenya	30.00	30.63	35.00
hi	**US**	33.79	33.79	44.14
	China	28.57	26.43	19.29
	India	33.33	33.33	35.15
	Iran	33.79	33.10	32.41
	Kenya	42.50	36.88	41.88
sw	**US**	37.93	32.41	28.28
	China	17.86	28.57	42.86
	India	30.91	30.30	41.21
	Iran	36.55	26.21	23.45
	Kenya	43.12	38.75	33.75

Table 9: Results (%) of models in mT5 family on GEOMLAMA.

Languages	Countries	XGLM-564M	XGLM-1.7B	XGLM-2.9B	XGLM-4.5B
en	**US**	32.41	37.93	31.72	37.24
	China	37.86	32.14	39.29	35.71
	India	30.91	40.00	43.03	42.42
	Iran	23.45	28.28	20.00	31.03
	Kenya	21.88	25.00	26.25	35.00
zh	**US**	34.48	36.55	40.00	35.86
	China	25.71	33.57	30.00	37.14
	India	27.27	32.73	36.36	31.52
	Iran	18.62	22.07	25.52	13.79
	Kenya	24.38	19.38	16.88	20.00
fa	**US**	49.66	49.66	46.90	49.66
	China	26.43	27.86	25.71	35.00
	India	32.73	31.52	28.48	32.73
	Iran	37.24	35.86	31.72	36.55
	Kenya	34.38	30.00	33.75	27.50
hi	**US**	35.86	28.97	33.79	28.97
	China	18.57	10.00	20.71	21.43
	India	33.33	29.70	29.09	33.94
	Iran	34.48	22.76	32.41	23.45
	Kenya	34.38	26.88	30.00	26.25
sw	**US**	25.52	28.28	27.59	24.14
	China	33.57	38.57	30.71	30.71
	India	34.55	34.55	37.58	33.33
	Iran	31.72	22.07	21.38	33.10
	Kenya	26.88	29.38	30.63	33.75

Table 10: Results (%) of models in XGLM family on GEOMLAMA.

Languages	Countries	mBERT	XLM	XLM-R-base	XLM-R-large
en	**US**	31.03	26.21	30.34	33.10
	China	30.00	39.29	34.29	37.14
	India	40.61	52.12	37.58	37.58
	Iran	21.38	27.59	28.28	37.93
	Kenya	30.63	34.38	30.63	32.50
zh	**US**	35.17	28.28	30.34	46.21
	China	30.71	28.57	46.43	40.00
	India	38.79	32.12	38.18	35.15
	Iran	32.41	36.55	24.14	33.10
	Kenya	41.25	35.00	27.50	39.38
fa	**US**	48.97	57.93	48.28	53.79
	China	27.86	20.71	28.57	32.14
	India	38.79	27.88	33.33	34.55
	Iran	47.59	31.03	35.17	33.79
	Kenya	38.75	31.87	27.50	34.38
hi	**US**	42.07	40.00	33.10	42.07
	China	29.29	22.86	18.57	13.57
	India	34.55	35.76	36.36	32.73
	Iran	33.79	31.03	31.72	27.59
	Kenya	28.75	33.75	36.25	33.75
sw	**US**	27.59	24.83	23.45	29.66
	China	34.29	22.86	32.14	29.29
	India	27.88	29.70	31.52	29.09
	Iran	20.69	27.59	35.17	31.72
	Kenya	26.88	31.87	27.50	31.87

Table 11: Results (%) of mBERT, XLM, XLM-R-base, XLM-R-large probed with prompts without country tokens on GEOMLAMA.

Languages	Countries	mT5-small	mT5-base	mT5-large
en	US	38.62	49.66	40.69
	China	45.00	47.14	42.86
	India	46.06	51.52	60.61
	Iran	38.62	46.90	43.45
	Kenya	43.12	44.38	57.50
zh	US	24.14	24.83	30.34
	China	32.86	30.71	33.57
	India	35.15	28.48	31.52
	Iran	36.55	40.69	40.00
	Kenya	39.38	43.12	36.88
fa	US	46.21	41.38	47.59
	China	39.29	33.57	41.43
	India	34.55	41.82	35.76
	Iran	35.86	37.24	42.76
	Kenya	37.50	41.88	42.50
hi	US	31.72	25.52	29.66
	China	39.29	38.57	32.86
	India	44.24	46.67	41.21
	Iran	30.34	33.79	31.03
	Kenya	40.00	40.00	41.25
sw	US	22.76	28.28	25.52
	China	23.57	35.71	42.86
	India	30.91	35.76	37.58
	Iran	14.48	20.69	22.07
	Kenya	16.88	23.75	26.25

Table 12: Results (%) of models in mT5 family probed with prompts without country tokens on GEOMLAMA.

Languages	Countries	XGLM-564M	XGLM-1.7B	XGLM-2.9B	XGLM-4.5B
en	US	28.97	38.62	34.48	40.00
	China	57.14	43.57	50.00	46.43
	India	51.52	47.88	53.94	46.67
	Iran	35.86	35.17	34.48	36.55
	Kenya	40.62	49.38	43.75	46.88
zh	US	34.48	42.76	38.62	47.59
	China	49.29	55.00	51.43	50.71
	India	44.24	52.73	54.55	46.67
	Iran	54.48	52.41	46.21	63.45
	Kenya	55.62	58.13	62.50	61.25
fa	US	27.59	28.97	35.17	34.48
	China	34.29	37.86	35.00	40.00
	India	38.18	34.55	40.00	36.97
	Iran	17.93	22.07	24.83	24.14
	Kenya	21.88	28.12	30.63	33.12
hi	US	24.14	32.41	20.69	31.72
	China	52.86	52.86	48.57	55.71
	India	39.39	41.21	40.61	40.61
	Iran	37.93	39.31	28.28	42.07
	Kenya	36.25	41.25	41.88	43.12
sw	US	42.07	40.00	41.38	35.17
	China	42.14	39.29	36.43	27.86
	India	40.00	42.42	50.30	46.06
	Iran	33.10	24.83	37.24	31.72
	Kenya	42.50	41.25	46.25	32.50

Table 13: Results (%) of models in XGLM family probed with prompts without country tokens on GEOMLAMA.