On the Scaling Laws of Geographical Representation in Language Models

Nathan Godey [1,2] **Éric de la Clergerie**[1] **Benoît Sagot**[1]

[1]Inria, [2]Sorbonne Université
Paris, France
{nathan.godey,eric.de_la_clergerie,benoit.sagot}@inria.fr

Abstract

Language models have long been shown to embed geographical information in their hidden representations. This line of work has recently been revisited by extending this result to Large Language Models (LLMs). In this paper, we propose to fill the gap between well-established and recent literature by observing how geographical knowledge evolves when scaling language models. We show that geographical knowledge is observable even for tiny models, and that it scales consistently as we increase the model size. Notably, we observe that larger language models cannot mitigate the geographical bias that is inherent to the training data.

Keywords: language models, geographic, bias

1. Introduction & Related work

In recent years, numerous studies analyzing the hidden representations of self-supervised language models have provided insights into how these models incorporate linguistic knowledge from their training data (Gupta et al., 2015; Köhn, 2015; Shi et al., 2016; Hupkes and Zuidema, 2018; Conneau et al., 2018; Jawahar et al., 2019).

This line of work has been called probing, as most approaches are generally based on the training of classifiers—or *probes*—upon frozen hidden representations.

Analyzing the representations of language models can point out sociocultural biases that were inherently learned by the models during training (Zhao et al., 2018), and training probes can help with mitigating these biases (Ravfogel et al., 2020; Iskander et al., 2023).

Among probing tasks, several works have focused on geographical representations that are implicitly embedded in language models. Louwerse and Benesh (2012) show that coordinates of places in the Middle-Earth can be predicted by just using the co-occurence matrix extracted from the Lord of the Rings novels. Faisal and Anastasopoulos (2022) build networks from geographical representations based on monolingual and multilingual models of different sizes. They show that all models embed more accurate geographical representations for countries of the Global North.

This geographical discrepancy can be explained by biases that are inherent to the datasets used for pretraining Faisal et al. (2022). Imbalanced frequency distributions of geographical references in pretraining data causes distortions in the representational space (Zhou et al., 2021). These distortions lead to a loss in the models' ability to differentiate between under-represented locations.

Recently, Gurnee and Tegmark (2023) have probed large language models from the Llama-2 suite (Touvron et al., 2023) to extract coordinates of prompted locations from hidden representations across layers. They show that models ranging from 7B to 70B parameters are able to convincingly embed geographical coordinates on a world map when representing basic prompts.

In this work, we propose to extend the analysis by Gurnee and Tegmark (2023) to smaller language models, in order to observe how scale affects the ability of models to implicitly embed geographical information from raw training data. We show that such ability consistently improves with model size, and that even tiny models are able to produce visually meaningful world maps.

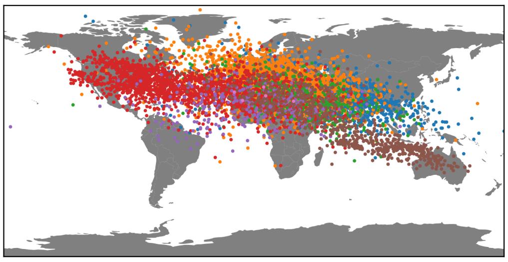
We make several contributions:

- We show that geographical information can be extracted to a certain extent from representations at every model scale;
- We observe that larger models are more geographically biased than their smaller counterparts;
- We find that the performance of models in terms of geographical probing is correlated with the frequency of corresponding country names in the training data.

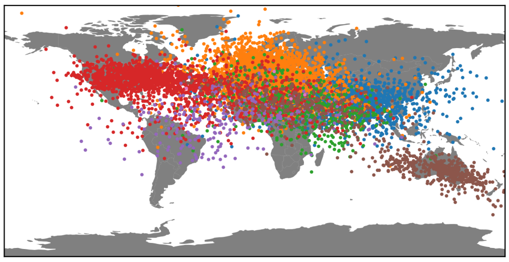
2. Scaling Laws of Geographical Probing

In this section, we train geographical probes for a wide variety of models at different scales.

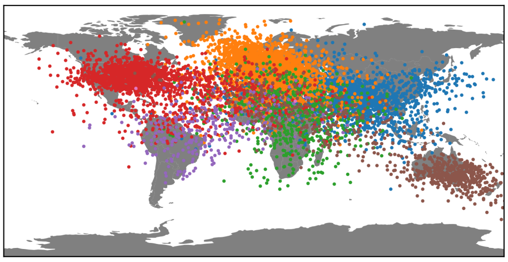
LREC-COLING 2024, pages 12416–12422



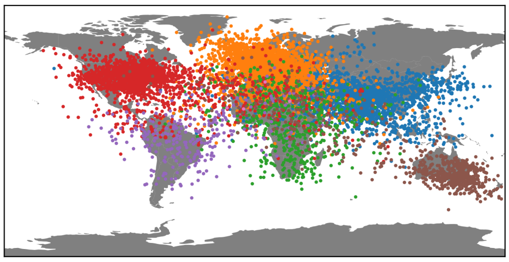
(a) Pythia 14M ($R^2 = 34.34$)



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



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



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

Figure 1: Predicted coordinates of test set instances for different model sizes. Each color represents a different continent.

2.1. Methodology

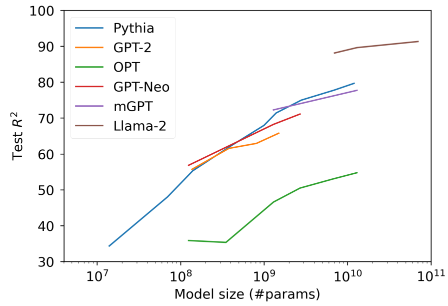
We use the World dataset from Gurnee and Tegmark (2023) as a geographical data source. It contains 39,504 location names from the whole world along with corresponding longitude and latitude. We use the same train-test split strategy as in the original article, thus keeping 20% of samples for testing purposes.

For each location name X, we prompt models with the text: "*Where is X in the world?*". We then infer with a given model on the whole dataset, and use the last token belonging to the entity X as the model's representation. To follow the linear probing paradigm used in Gurnee and Tegmark (2023), we train a Ridge linear regressor (Hoerl and Kennard, 1970) to predict latitude and longitude based on the model's representations. We then measure the probe's performance on the test set using the R^2 correlation coefficient.

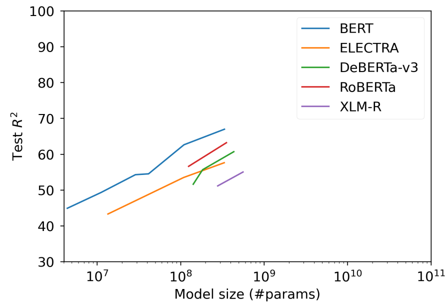
2.2. Results

In Figure 1, we display the predictions of the probe for the most performant layer, which is generally the last one. We observe that geographical information can be extracted from models even for a very small parameter count. The performance of the probes seem to increase with the model size.

We show in Figure 2 that the performance of language models evolves consistently with model size, regardless of the architecture. We validate this property on several decoder model families: GPT-2 (Radford et al., 2019), OPT (Zhang et al., 2022), Pythia (Biderman et al., 2023), GPT-Neo (Black et al., 2021), the multilingual mGPT (Shliazhko et al., 2023), and Llama-2 (Touvron et al., 2023). We also display results for several encoder models: BERT (Devlin et al., 2019; Turc et al., 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al.,



(a) Decoder models



(b) Encoder models

Figure 2: Evolution of the R^2 coefficient on the test set for various model suites.

2020), and DeBERTa-v3 (He et al., 2020). This property also applies for encoder models, for which we notice that the BERT suite unexpectedly outperforms its counterparts. The performance of encoder models is comparable with the one of equivalent decoder models. We can underline the fact that BERT-Large (336M parameters) is as accurate as the three times larger Pythia-1B.

Interestingly, the multilingual XLM-R (Conneau et al., 2020) underperforms its counterparts, even though multilingual data must have increased the training data's geographical diversity to some extent (Faisal and Anastasopoulos, 2021). The mGPT suite also slightly underperforms Pythia models at equivalent model sizes.

We verified that the better performance of larger models was not solely related with the ability of the probes to extract better patterns from their higher-dimensionality hidden representations. We achieved this by concatenating representations with themselves to increase dimensionality without introducing novel knowledge. It led to slightly worse performance for all tested models, thus showing that performance was not a consequence of dimensionality alone.

3. Geographical Bias and Scale

In Figure 1, it seems at first glance that as the model size increases, the predictions tend to be more accurate for locations of the Southern Hemisphere. In this section, we propose to quantify this hypothesized behavior by measuring the bias across countries and continents for various scales. We also correlate the models' accuracy with both lexical and geographical factors.

3.1. Measuring bias

We group probe performance as measured by mean-squared error (MSE) on predicted coordinates, and average measures by continent in Figure 3. While we notice that the performance increases consistently for every continent, we do not observe a significant reduction in the performance gap across continents as model size increases.

To measure the heterogeneity of the probing performance of language models across countries, we use the Gini coefficient (Gini, 1912) that is widely used in economics. Given a series of observed variables $(x_i)_{i \in [1,N]}$, the Gini coefficient is defined as:

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

A Gini coefficient of 1 reflects perfect heterogeneity, while a Gini of 0 implies perfect homogeneity.

Figure 4 shows that the larger the model is, the more heterogeneous the probe performance is across countries and continents. This contradicts the impression given by Figure 1, and shows that scale does not solve the geographical discrepancy caused by bias inherent to the training data.

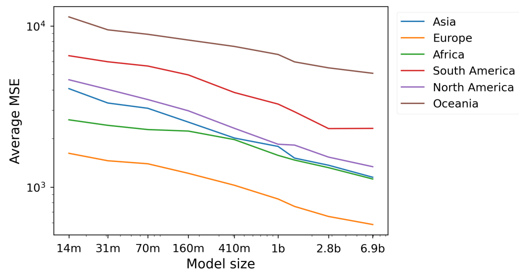


Figure 3: Average MSE by continent for different sizes in the Pythia suite.

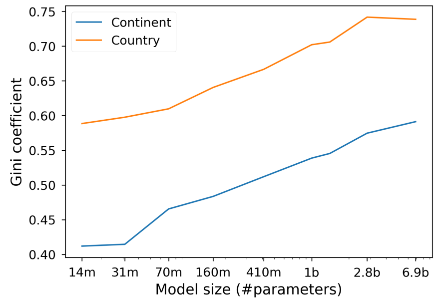


Figure 4: Gini coefficients of MSE on the test set averaged by country or by continent, as model size increases.

In Figure 5, we locally average log-MSE on a World map, and report results agglomerated accorded to latitude and longitude. We clearly observe that the model performs poorly in Oceania, South Asia and South America. We also see that the error is minimal around the latitude of North America and Europe, while it increases in the Southern Hemisphere.

3.2. Identifying sources of bias

We attempt to correlate the performance of our geographical probes with several factors. First, the

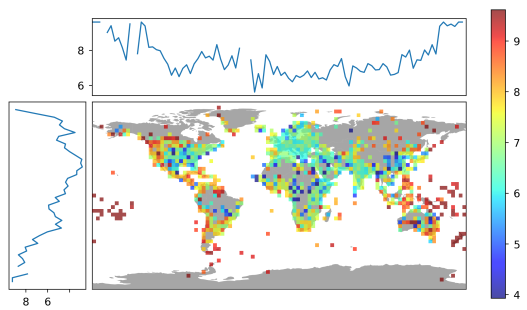


Figure 5: Test log-MSE for Pythia-1B as plotted on a World map.

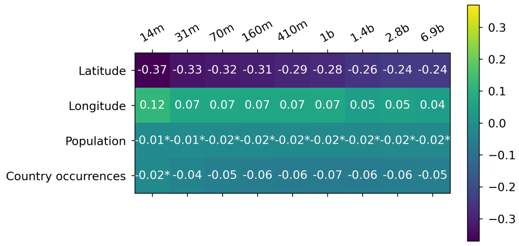


Figure 6: Pearson correlation coefficients of various factors with location-wise MSE, for several Pythia model sizes. *: Tests that yielded p-values above 0.05.

dataset from (Gurnee and Tegmark, 2023) provides each location with an estimate of the corresponding population count when relevant. We also consider training data distribution as a potential factor of heterogeneity. Finally, we consider latitude and longitude as potential factors of bias.

To account for training data distribution, we look for exact string matches of country names from the Gurnee and Tegmark (2023) dataset in an extract of The Pile (Gao et al., 2020) containing 3.5 million samples [1]. We select this dataset as it was used to pretrain the models from the Pythia suite (Biderman et al., 2023) we evaluate in this section. We find 15 million matches, covering 98% of the countries of the dataset.

We do not count occurrences of location names directly, as matching locations on the basis of their names does not account for named entity ambiguity. An example of ambiguous location name is *Fully*, which is a town in Switzerland. An exact match strategy overestimates by large margins the occurrence count of this location, because of the corresponding English word *fully*. Disambiguation techniques have been designed (Hoffart et al., 2011; Orr et al., 2020), but we prefer to avoid the risk of bias propagation and the cost of using such methods on a large corpus.

We display Pearson correlations between each of the aforementioned factors and the entity-level MSE for each model size in Figure 6. As in Figure 1a, we observe that the error on coordinates prediction is negatively correlated with the latitude, i.e. southern locations are less accurately identified. This correlation slowly decays as the model size increases. Meanwhile, longitude seems to be mildly correlated with the probe performance.

Interestingly, the population count is not correlated with the error level. The occurrence count of the location country is negatively correlated with the error level, thus showing that the more country names appear in the training dataset, the more the probes are able to recover coordinates from locations in these countries. However, this correlation is mild and even below the significance threshold for the smallest model.

We also measure the correlation between country occurrences and other metrics to account for the bias inherent to the data. We observe that country name occurrences are positively correlated with latitude with a p-value of 0.06, and not correlated with the longitude. More importantly, the population count of a country and the count of this country name in the data are heavily correlated (factor of +0.52 and p-value of 3e-23). Thus, even though the data seems guided by demographic factors, this is not the case of the model's representations.

4. Discussion

We believe that quantifying sociocultural bias in representations of language models and pretraining datasets allows to better understand the roots of the biases that can be observed during generation. Bender et al. (2021) discuss the relevance of scaling models to ever larger magnitudes, with regard to environmental and financial costs. Our study shows that scale can also increase language modeling bias when it comes to geographical representation, given a pretraining dataset. We advocate in favor of measuring and mitigating bias in pretraining datasets to avoid scaling bias along with performance.

Conclusion

In this study, we show that a wide variety of language models, varying in architecture and sizes, implicitly embed geographical data to some extent. As we consider larger models, the performance of geographical probes consistently increases towards levels shown in Gurnee and Tegmark (2023).

We show numerically that the geographical probe performance is correlated with latitude across model sizes, but also with the number of occurrence of corresponding country names in the pretraining data. Conversely, the population count of the location seems uncorrelated with the probe performance. This indicates that a minority of people benefit from better geographical understanding when using language models, which does not maximize the social utility of these systems.

While it may initially seem that this performance increase mitigates heterogeneity between Southern and Northern countries, we actually show that larger models tend to be more biased according to the Gini coefficient taken on prediction error. This tends to show that scaling language models can amplify discrepancies in their geographical knowledge.

[1] https://huggingface.co/datasets/ola13/small-the_pile

Acknowledgements

This work was funded by the last authors' chair in the PRAIRIE institute, funded by the French national agency ANR as part of the "Investissements d'avenir" program under the reference ANR-19-P3IA-0001. We thank Stella Biderman for her insightful advice.

Bibliographical References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. 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 and Antonios Anastasopoulos. 2021. Investigating post-pretraining representation alignment for cross-lingual question answering. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 133–148, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2022. Geographic and geopolitical biases of language models. In *arXiv:2212.10408*, Online.

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.

Corrado Gini. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Dieuwke Hupkes and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5617–5621. International Joint Conferences on Artificial Intelligence Organization.

Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. 2023. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977, Toronto, Canada. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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.

Laurel Orr, Megan Leszczynski, Simran Arora, Sen Wu, Neel Guha, Xiao Ling, and Christopher Re. 2020. Bootleg: Chasing the tail with self-supervised named entity disambiguation.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mgpt: Few-shot learners go multilingual.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models.

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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2021. Frequency-based distortions in contextualized word embeddings.

Language Resource References

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.