HOW GOOD ARE COMMERCIAL LARGE LANGUAGE MODELS ON AFRICAN LANGUAGES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in Natural Language Processing (NLP) has led to the proliferation of large pretrained language models. These models have been shown to yield good performance, using in-context learning, even on unseen tasks and languages. They have also been exposed as commercial APIs as a form of language-model-as-a-service, with great adoption. However, their performance on African languages is largely unknown. We present a preliminary analysis of commercial large language models on two tasks (machine translation and text classification) across eight African languages, spanning different language families and geographical areas. Our results suggest that commercial language models produce below-par performance on African languages. We also find that they perform better on text classification than machine translation. In general, our findings present a call-to-action to ensure African languages are well represented in commercial large language models, given their growing popularity.

1 INTRODUCTION

Large language models have risen to the fore of Natural Language Processing (NLP). These models have been shown to achieve state-of-the-art performances on several tasks. More recently, focus has shifted from the pretrain-finetune paradigm (Howard & Ruder, 2018; Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020) to in-context learning (Brown et al., 2020; Lin et al., 2021; Wei et al., 2022a; Chowdhery et al., 2022; Chung et al., 2022; Sanh et al., 2022; Dong et al., 2023). In-context learning proves that prompting large language models with some task-specific examples allows them perform well on test examples of that task, all without updating the model's parameters. This has led to reduced computation costs and has made it possible to create language-models-as-a-service (Sun et al., 2022), in the form of commercial Application Programming Interfaces (APIs). Commercial language models have become very prevalent. For context, the recently released ChatGPT[1] amassed 100 million users[2] in two months, making it the fastest growing consumer app in recent history. Given their dominance and inevitable continual rise, it is important to understand how these models perform on African languages. Hence, we present a preliminary effort to close this gap by evaluating two commercial large language models using in-context learning on African languages. Evaluation is performed on two tasks - text classification and machine translation. Our experiments, spanning 8 African languages from different language families and geographical locations, suggests that commercial language models do not perform well on African languages. In particular, we note a large disparity in performance, depending on the evaluation task - models perform better on text classification than machine translation. Our work sheds light on the need to ensure the inclusion of African languages in the development of commercial language models, given their inevitable adoption in our daily lives.

[1] https://chat.openai.com/

[2] https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

2 RELATED WORK

2.1 IN-CONTEXT LEARNING

The use of pretrained language models has become the de-facto approach to solving natural language processing (NLP) tasks. Previous models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020) largely follow a pretrain-finetune setting (Howard & Ruder, 2018). In this method, the pretrained model is finetuned on a downstream task, such as text classification, and then used for that task. While this works very well, it has several downsides. For one, finetuned models are usually task-specific and this means one has to maintain separate models for separate tasks. Furthermore, the growing size of pretrained language models (Kaplan et al., 2020) means that it is becoming increasingly expensive to finetune such gigantic models. One solution that has proven popular in recent times is in-context learning (Brown et al., 2020; Schick & Schütze, 2021; Wei et al., 2022a; Chowdhery et al., 2022; Chung et al., 2022; Sanh et al., 2022; Dong et al., 2023). The core idea behind this method is to enable pretrained language models learn from examples within the context. In this setting, a user prompts a pretrained language model with a few labelled examples of a task following a specific pattern, and unlabelled examples that need to be predicted on (Wei et al., 2022c; Liu et al., 2022; Wei et al., 2022b). In-context learning can also work in a zero-shot setting where no labelled examples are included in the prompt. In-context learning works surprisingly well and is very efficient since there is no update to the model's parameters. As a result, computation costs are significantly reduced and it becomes possible to expose language models as a service (Sun et al., 2022). Commercial APIs are heavily reliant on in-context learning as this is the primary method through which users interact [3] with the models[4].

2.2 MULTILINGUAL IN-CONTEXT LEARNING

Large language models have proven successful in multilingual settings. Lin et al. (2021) train several multilingual models, of which the largest one (7.5B parameters) sets a state-of-the-art in few-shot learning on more than 20 languages. Their model outperforms GPT3 on several multilingual tasks. Muennighoff et al. (2022) perform multitask prompted finetuning on multilingual pretrained language models and observe impressive zero-shot generalization to tasks in unseen languages. Following findings from Blevins & Zettlemoyer (2022) that non-English dataset present in the pretraining corpora of English language models explains their surprising cross-lingual ability, Chowdhery et al. (2022) deliberately introduce non-English corpora ($\approx 22\%$) into the pretraining corpora of their PaLM model and achieve impressive few-shot multilingual performance. Shi et al. (2022) evaluate GPT3 and PaLM on a newly introduced grade school mathematics multilingual benchmark. They find that using prompts with intermediate reasoning steps in English consistently led to competitive or better results than those written in the native language of the question. They also set a new state-of-the-art on a common-sense reasononing multilingual benchmark, XCOPA (Ponti et al., 2020), using few-shot examples. Zhao & Schütze (2021) show that prompting yields better cross-lingual transfer in few-shot settings than finetuning and in-language training of multilingual natural language inference. Furthermore, Winata et al. (2021) evaluate the multilingual ability of GPT (Radford et al., 2019) and T5 (Raffel et al., 2020) models on multi-class text classification, and find that they work well on non-English languages given a few English examples. Concurrent work (Jiao et al., 2023) evaluate ChatGPT on machine translation and find that, while it is competitive with other commercial translation APIs such as Google translate[5], it is less robust on other domains such as biomedical. Another concurrent work (Zhang et al., 2023) conducts a study on the performance of GLM (Zeng et al., 2022) on machine translation. They note several interesting findings on the effect of prompt template, examples and language. Despite the plethora of works on multilingual prompting, little to no African languages are usually contained in the evaluation sets of nearly all of these works. When present, they are often obtained by translating the existing datasets of other languages (Xinyan Velocity Yu, 2022). This method has been shown to contain artifacts that can inflate the performance of models evaluated on such datasets (Artetxe et al., 2020). Our work is orthogonal to all of this works because we focus solely on commercial language model APIs, given

[3] https://platform.openai.com/docs/guides/completion/prompt-design
[4] https://docs.cohere.ai/docs/prompt-engineering
[5] https://translate.google.com/

their prevalence. Furthermore, we conduct the first evaluation (to the best of our knowledge) of these models on African languages.

3 METHODOLOGY

3.1 DATASETS

Evaluation is done on two tasks - text classification and machine translation.

3.1.1 TEXT CLASSIFICATION

We use the news topic classification datasets from Hedderich et al. (2020) and Alabi et al. (2022). We select all the languages from Hedderich et al. (2020), which covers two African languages - Hausa (hau) and Yoruba (yor). The Yoruba dataset has 7 categories, while the Hausa dataset has 5 categories. Pretrained language models have been shown to work very well on this dataset in both few and zero-shot settings. We use the train, validation and test splits as released by the authors. The dataset from Alabi et al. (2022) covers five languages, out of which we select three - Nigerian Pidgin (pcm), Malagasay (mlg), and Somali (som). Each language has 5 categories, except Somali which has 6. We use the train, validation and test splits as released by the authors. We select these languages because they cover different language families and geographical areas.

3.1.2 MACHINE TRANSLATION

We use the MAFAND-MT machine translation dataset from Adelani et al. (2022) which covers 16 African languages. Running translation on commercial APIs is cumbersome and expensive, hence we select 5 languages from the 16. The five languages are isiZulu (zul), Yoruba (yor), Nigerian Pidgin (pcm), Amharic (amh) and Lugala (lug). We use the splits as released by the authors.

3.2 MODELS

Two commercial APIs are considered: ChatGPT[6] and Cohere[7]. We consider both these APIs because they are arguably the most popular ones[8]. ChatGPT is based on the Instruct-GPT models (Ouyang et al., 2022). It is optimized for conversations and has been shown to be capable of several NLP tasks including text classification, machine translation, question answering, and so on. We use Cohere's multilingual model[9] which is based on their multilingual embedding model[10]. The embedding model supports 100 languages, including 15 African languages. All the languages we consider, except Nigerian Pidgin, are supported by the model. However, given the linguistic proximity of Nigerian Pidgin to English (Faraclas, 2008; Ogueji & Ahia, 2019; Chang et al., 2020; Ahia & Ogueji, 2020a; Lent et al., 2021; 2022), the model should be able to perform well on the dataset.

3.3 PROMPTING AND EVALUATION

We describe our prompting and evaluation approaches for text classification and machine translation.

3.3.1 TEXT CLASSIFICATION

For Cohere, we use the Classify[11] endpoint and follow the format specified in the API documentation[12]. When using ChatGPT, we design several prompts ourselves and we also ask ChatGPT for the best prompt for classification, following concurrent work (Jiao et al., 2023). We perform some initial evaluation of the prompts and select the best one.

[6] https://chat.openai.com/
[7] https://www.cohere.ai
[8] https://venturebeat.com/uncategorized/openai-rival-cohere-launches-language-model-api/
[9] https://docs.cohere.ai/changelog/multilingual-support-for-coclassify
[10] https://docs.cohere.ai/docs/multilingual-language-models
[11] https://api.cohere.ai/classify
[12] https://docs.cohere.ai/reference/classify

Our best prompt is shown below:

```
Given the following news headlines and their categories:
Text: {Sentence}
Category: {Label}

Please classify the following news headlines into one of:
{Label List}.
Text: {Sentence}
Category:
```

Where $Sentence$ is the news headline to be classified, $Category$ is the news topic, and $LabelList$ is a comma separated list of all unique labels for that language.

For both models, we supply two example demonstrations per category from the training set. We randomly sample 100 samples from the test set for each language and evaluate on this. Both demonstrations and evaluation are done across two random seeds, such that we sample distinct demonstrations and test samples for each language with each random seed. We report the average F1 score for each language across both seeds. It should be noted that we decide to evaluate on a subset of the test set because of the tedious nature of obtaining results ChatGPT.

3.3.2 MACHINE TRANSLATION

We do not use Cohere for machine translation because its generation API currently supports only English[13]. ChatGPT is used for all our machine translation evaluations. Preliminary results from comparing few-shot to zero-shot translations on Nigerian Pidgin suggested no noticeable difference. Hence, we perform all translations in a zero-shot manner because of the tedious nature and low-throughput of obtaining results from ChatGPT.

We use the prompt used in concurrent work (Jiao et al., 2023) which is shown below:

```
Please provide the [TGT] translation for these sentences:
{Sentence}
{Sentence}
```

Where TGT is the target language to be translated into, and $Sentence$ is a sentence to be translated. We sample 100 sentences from the test set of each language and evaluate translating this to and from English. We report the BLEU score papineni-etal-2002-bleu which is calculated using SacreBLEU (Post, 2018).

It has been shown that English prompts perform better, on average, than in-language prompts (Lin et al., 2021; Shi et al., 2022), so we do not explore prompting in the target language for both tasks.

4 RESULTS

4.1 TEXT CLASSIFICATION

Results are reported in table 1. As we can see, both commercial models fall well below the current state of the art. Surprisingly, Cohere's multilingual embedding model is the worst performer, despite supporting almost all the languages evaluated on. Nigerian Pidgin has the highest score in the Cohere results. This is most likely as a result of its close linguistic relationship with English language, which usually makes up a significant portion of the pretraining corpora of pretrained language models (Wenzek et al., 2020; Gao et al., 2020; Laurençon et al., 2022). ChatGPT is the best performing commercial model, and it gets above average F1 scores on all languages. Similar to Cohere, Hausa and Nigerian Pidgin possess the highest F1 scores. The details of ChatGPT's pretraining corpora and exact training methods are uknown, so it is hard to hypothesize a reason for its relatively good performance. However, it is very likely that its pretraining corpora contains non-English text. Overall, both commercial models fall significantly short of the current state of the art. While ChatGPT

[13] https://docs.cohere.ai/docs/generation-card#technical-notes

Table 1: Text Classification Results: We report the F1 scores for the commercial models. We also report the current state of the art result obtained from Alabi et al. (2022)
.

Language	Cohere	ChatGPT	Current SOTA
Hausa (hau)	43.2	77.9	91.2
Malagasay (mlg)	35.0	51.1	67.3
Nigerian Pidgin (pcm)	48.8	73.4	82.2
Somali (som)	28.4	51.3	79.9
isiZulu (zul)	24.8	54.8	79.6

Table 2: Machine Translation Results: We report the BLEU scores of the translations from ChatGPT. We also report the current state of the art result obtained from Adelani et al. (2022)
.

Translation Direction	ChatGPT	Current SOTA
Lug→Eng	0.16	21.1
Eng→Lug	0.13	14.3
Pcm→Eng	0.22	45.2
Eng→Pcm	0.20	35.0
Swa→Eng	0.18	30.0
Eng→Swa	0.15	26.7
Yor→Eng	0.10	16.0
Eng→Yor	0.12	12.8
Zul→Eng	0.31	37.8
Eng→Zul	0.26	21.2

is the better performer, Cohere's performance is especially surprising since it has been trained on almost all of the evaluated languages[14].

4.2 Machine Translation

Results are reported in table 2. ChatGPT has very poor performance on machine translation, obtaining BLEU scores of less than 1.0 on all languages. This is very surprising given its good performance on text classification. Our results agree with concurrent work (Jiao et al., 2023), which reports significantly worse performance on Romanian, a relatively low-resource language, than on higher-resource languages like English and German. While the BLEU scores are too low to draw conclusions from, ChatGPT seems to perform better when translating from into English than from it. This agrees with previous works (Belinkov et al., 2017; Bugliarello et al., 2020) which show that it is harder to translate into morphologically rich languages, like African ones, than morphologically poor ones like English. In general, our results suggest that ChatGPT is not good enough for translation involving African languages. It also suggests that ChatGPT performs better on sequence classification tasks than it does on text generation tasks for African languages.

[14] https://txt.cohere.ai/multilingual/

5 CONCLUSION

We have presented a preliminary analysis of commercial language models on African languages. Joshi et al. (2020) note that over 90% of the world's 7000+ languages are under-studied by the NLP community. Despite the 2000+ spoken languages and over 1 billion people in Africa[15], its languages make up a significant portion of the under-studied languages (Blasi et al., 2022). While there have been several efforts (∀ et al., 2020; Ahia & Ogueji, 2020b; Adelani et al., 2021; Ogueji et al., 2021; NLLB Team et al., 2022; Alabi et al., 2022; Dossou et al., 2022; Adebara et al., 2022) to close this gap, there is still a lot of work to be done. This is even more pertinent given the rapid commercial adoption of large scale language models. Our findings suggest that these models do not perform well on African languages. In particular, there seems to be performance disparity, depending on the task evaluated. Although our work reports what is, to the best of our knowledge, the first evaluation of commercial language models on African languages, we note that this only a preliminary study that needs to be further advanced. Future works could focus on more advanced prompting methods such as chain-of-thought (Wei et al., 2022c) and pivot prompting (Jiao et al., 2023), evaluation of more test samples and a wider variety of tasks. While our finding may be impacted by the sampled test data, the use of the BLEU automatic metric (Callison-Burch et al., 2006; Mathur et al., 2020; Freitag et al., 2020), prompting template and prompting examples, it nonetheless presents a call-to-action to ensure African languages are well represented in the age of commercial large language models.

REFERENCES

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. Serengeti: Massively multilingual language models for africa, 2022. URL https://arxiv.org/abs/2212.10785.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 10 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00416. URL https://doi.org/10.1162/tacl_a_00416.

[15]https://en.wikipedia.org/wiki/Demographics_of_Africa

Orevaoghene Ahia and Kelechi Ogueji. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. *ArXiv*, abs/2003.12660, 2020a.

Orevaoghene Ahia and Kelechi Ogueji. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. *CoRR*, abs/2003.12660, 2020b. URL https://arxiv.org/abs/2003.12660.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7674–7684, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.618. URL https://aclanthology.org/2020.emnlp-main.618.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1080. URL https://aclanthology.org/P17-1080.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.376. URL https://aclanthology.org/2022.acl-long.376.

Terra Blevins and Luke Zettlemoyer. Language contamination helps explain the cross-lingual capabilities of english pretrained models, 2022. URL https://arxiv.org/abs/2204.08110.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1640–1649, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.149.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032.

Ernie Chang, David Ifeoluwa Adelani, Xiaoyu Shen, and Vera Demberg. Unsupervised pidgin text generation by pivoting english data and self-training, 2020. URL https://arxiv.org/abs/2003.08272.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh,

Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey for in-context learning, 2023. URL https://arxiv.org/abs/2301.00234.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages, 2022.

Nicholas Faraclas. Nigerian pidgin english: morphology and syntax. *Varieties of English: Africa, South and Southeast Asia*, 4:340–367, 2008.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. Participatory research for low-resourced machine translation: A case study in african languages. *Findings of EMNLP*, 2020.

Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 61–71, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.5. URL https://aclanthology.org/2020.emnlp-main.5.

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. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2580–2591, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.204. URL https://www.aclweb.org/anthology/2020.emnlp-main.204.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? a preliminary study, 2023. URL https://arxiv.org/abs/2301.08745.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=UoEw6KigkUn.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 58–71, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.5. URL https://aclanthology.org/2021.conll-1.5.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6439–6449, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.691.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668, 2021. URL https://arxiv.org/abs/2112.10668.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

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

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4984–4997, Online, July

2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2022. URL https://arxiv.org/abs/2211.01786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Kelechi Ogueji and Orevaoghene Ahia. PidginUNMT: Unsupervised Neural Machine Translation from West African Pidgin to English. *ArXiv*, abs/1912.03444, 2019.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.mrl-1.11.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted

training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2339–2352, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.185. URL https://aclanthology.org/2021.naacl-main.185.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022. URL https://arxiv.org/abs/2210.03057.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of ICML*, 2022.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022c. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.494.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 1–15, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.1. URL https://aclanthology.org/2021.mrl-1.1.

Trina Chatterjee Junjie Hu Eunsol Choi Xinyan Velocity Yu, Akari Asai. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of EMNLP*, 2022.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study, 2023. URL https://arxiv.org/abs/2301.07069.

Mengjie Zhao and Hinrich Schütze. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8547–8555, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.672. URL https://aclanthology.org/2021.emnlp-main.672.