How good are Large Language Models on African Languages?

Anonymous ACL submission

Abstract

Recent advancements in natural language processing have led to the proliferation of large language models (LLMs). These models have been shown to yield good performance, using in-context learning, even on unseen tasks and languages. However, their performance on African languages is largely understudied relative to high-resource languages. We present an analysis of three popular large language models (mT0, LLaMa 2, and GPT-4) on five tasks (news topic classification, sentiment classification, machine translation, question answering, and named entity recognition) across 30 African languages, spanning different language families and geographical regions. Our results suggest that all LLMs produce lower performance for African languages, and there is a large gap in performance compared to high-resource languages (such as English) for most tasks. We find that GPT-4 has an average or good performance for classification tasks, but very poor results on generative tasks such as machine translation. Surprisingly, we find that mT0 had the best overall performance for cross-lingual QA, better than the state-of-the-art supervised model (i.e. fine-tuned mT5) and GPT-4 on African languages. Overall, LLaMa 2 showed the worst performance, which we believe is due to its English and code-centric (around 98%) pre-training corpus. Our findings confirm that performance on African languages remains challenging for current large language models and that there is a need for additional efforts to close this gap.

1 Introduction

Large language models have risen to the fore of natural language processing and also become increasingly commercially viable. These models have empirically demonstrated strong performance across both tasks and languages (Brown et al., 2020; Lin et al., 2021; Chowdhery et al., 2022; Chung et al., 2022). However, their performance on low-resources languages, such as African languages, is largely understudied. This is problematic for two primary reasons: ideally our approaches to language understanding should be applicable to all languages and advances should be ensured to benefit all language users.

In this paper, we conduct an extensive analysis of large language models for 30 African languages from different language families and geographical locations. Our evaluation covers three popular LLMs: mT0 (Muennighoff et al., 2023) (derived from mT5 (Xue et al., 2021) through multitask prompted fine-tuning), LLaMa 2 (Touvron et al., 2023), and GPT-4. We evaluate the models on five tasks: news topic classification, sentiment classification, machine translation, named entity recognition, and question answering.

Our results suggest that commercial language models do not perform well on African languages. In particular, we note a large disparity in performance depending on the task: models perform better for classification tasks than generative tasks, such as question answering and machine translation. We also find performance to be worse for low-resource languages compared to high-resources ones.

Our evaluation shows that GPT-4 achieves more than 80% of the performance of fully-supervised fine-tuning on news topic classification and sentiment classification, but a bit lower performance—62% of full-supervised fine-tuning on named entity recognition respectively. On the other hand, performance of generative tasks like machine translation (MT) was poor. In comparison to MT evaluation on high-resource languages (e.g. English-German and French-German), our evaluation shows the gap in performance between LLM and full-supervised fine-tuning is wider for African languages.

In general, other LLMs have worse results than GPT-4 on most tasks. However, for cross-lingual

QA, mT0 had the best overall performance, even exceeding the state-of-the-art supervised model (i.e. fine-tuned mT5). Overall, LLaMa 2 records the worst performance due to its limited multilingual. Our work sheds light on the need to ensure the inclusion of African languages in the development of large language models, given their inevitable adoption in our daily lives.

2 Languages and evaluation tasks

We cover 30 African languages from four language families (Afro-Asiatic, Niger-Congo, Nilo-Saharan, and English-Creole). Appendix A shows the languages and tasks we evaluated on.

2.1 Evaluation tasks and datasets

News Classification: MASAKHANEWS (Adelani et al., 2023) is a multilingual news classification dataset covering 16 typologically-diverse languages spoken in Africa, including English and French.

Sentiment Classification: AFRISENTI (Muhammad et al., 2023) is a multilingual sentiment classification dataset for 14 languages spoken in Africa.The goal of the task is to classify tweets as positive, negative, or neutral.

Named Entity Recognition (NER): For NER, we make use MASAKHANER-X (Ruder et al., 2023)—a subset of MASAKHANER (Adelani et al., 2021, 2022b) that has been converted to be suitable for evaluating generative models (i.e. the input *"Jens is an employee of Amazon"* should produce *"PERSON: Jens && ORG: Amazon"* as an output) and covers 20 African languages.

Question Answering (QA): AFRIQA (Ogundepo et al., 2023) is a cross-lingual, open-retrieval, question answering (XOR QA) dataset, which consists of more than 12,000 examples across 10 African languages. In this setting, the answer and context are provided in a high resource language, while the question is in an African language.

Machine Translation (MT): MAFAND-MT Adelani et al. (2022a)[1] is a professionally translated, news domain dataset which covers 16 African languages. Here, we compare the performance of fine-tuning M2M-100 on few thousand parallel sentences to the performance of GPT-4. The reason we compared to this setting is because pre-trained M2M-100 was trained on few African languages, only 8 out of 16 languages are covered by the model. An effective way to add a new language to the model is to fine-tune a pre-trained MT model on few high-quality parallel data.

3 Experimental Setup

We focus our evaluations on the following LLMs: mT0-13B (-MT), LLaMa 2 13B, and GPT-4.[2] mT0-13B (Muennighoff et al., 2023) is an LLM obtained by fine-tuning mT5-XXL (a 13B parameter size text-to-text model and also the largest) on a collection of multitask prompted datasets known as `xP3` (Crosslingual Public Pool of Prompts) while mT0-13B-MT was fine-tuned on `xP3mt` where prompts are provided in 20 languages (machine-translated from English).[3] LLaMa 2 (Touvron et al., 2023) is a popular, publicly available LLM with chat functionality, the number of parameters ranges from 7B to 70B; we make use of the 13B chat model since it is the largest model that can fit a single A100 GPU. GPT-4 is a transformer-style model pre-trained to predict the next token followed by a set of instructions in a prompt based on human feedback.

3.1 Prompt Templates

We designed our prompts in a zero-shot cross lingual manner, that is, the prompt context and query is designed in English, while the text to be analyzed is provided in the target African language. For each task, we designed simple prompts that tend to reasonable results on few examples of the training set. We prompt the LLMs using only English since it has been shown that English prompts perform better, on average, than in-language prompts (Lin et al., 2021; Shi et al., 2022). As such, we do not explore prompting in the target language for both tasks. Appendix B provides details on the prompt used for each task.

3.2 State-of-the-art (SotA) models

Here, we compare the performance GPT-4 on African languages with:

1. **State-of-the-art:** fully-supervised setting results i.e. pre-trained language models fine-

[1] While Flores-200 is a larger benchmark, it was used for instructing fine-tuning of mT0 model, so, it is no longer suitable as an evaluation set.

[2] We specifically use gpt-4-0613.

[3] `https://huggingface.co/datasets/bigscience/xP3mt`

		High-resource		African languages														
Model	**Size**	**eng**	**fra**	**amh**	**hau**	**ibo**	**lin**	**lug**	**pcm**	**orm**	**run**	**sna**	**som**	**swa**	**tir**	**xho**	**yor**	**avg**
Fine-tune: SotA																		
AfroXLMR-large	550M	**93.1**	**91.1**	**94.4**	**92.2**	**93.4**	**93.7**	**89.9**	92.1	**98.8**	**92.7**	**95.4**	**86.9**	**87.7**	**89.5**	**97.3**	**94.0**	**92.7**
Prompting of LLMs																		
GPT-4	-	84.7	82.6	91.1	74.4	82.2	82.4	84.1	**94.7**	78.8	88.5	78.1	79.7	79.2	75.7	87.5	93.7	85.6
mT0	13B	64.7	58.3	64.8	65.6	63.6	62.3	56.7	74.4	57.4	58.8	82.6	52.3	57.8	52.0	69.7	61.7	62.8
mT0-MT	13B	68.7	58.0	63.5	72.1	70.5	63.4	74.1	81.8	56.3	61.4	72.1	56.0	58.1	55.2	84.6	74.0	67.4
LLaMa 2	13B	61.0	45.1	7.1	37.2	60.7	66.1	63.2	70.4	22.6	63.4	69.6	48.8	50.5	3.9	61.3	41.1	47.6

Table 1: **News Classification Results:** We compare the F1-score of various LLMs' results with that of the current state of the art result obtained from Adelani et al. (2023). Best results per language are in bold.

		High-resource		African languages										
Model	**Size**	**eng**	**por-mz**	**amh**	**arq**	**ary**	**hau**	**ibo**	**kin**	**pcm**	**swa**	**tso**	**twi**	**yor**	**avg**
Fine-tune: SotA															
AfroXLMR-large	550M	**68.1**	**71.6**	61.6	**68.3**	**56.6**	**80.7**	**79.5**	**70.6**	**68.7**	63.4	47.3	**64.3**	**74.1**	**66.8**
Prompting of LLMs															
GPT-4	-	66.1	60.4	72	63.2	56.4	41.9	65.1	57.3	64.1	**64.5**	22.3	51.9	53.9	55.69
mT0	13B	41.2	16.0	67.2	50.4	37.0	40.5	26.7	36.3	63.6	20.9	**47.5**	43.5	35.6	42.6
mT0-MT	13B	37.2	16.5	**70.2**	58.5	34.6	36.1	27.2	39.5	50.7	18.7	42.1	35.9	23.7	39.7
LLaMa 2	13B	52.8	32.3	10.5	26.2	37.4	25.5	35.1	34.2	24.3	49.7	30.5	23.9	24.0	29.2

Table 2: **Sentiment Analysis Results:** We compare the F1-score of various LLMs' results with that of the current state of the art result obtained from Muhammad et al. (2023). Best results per language are in bold.

tuned on labelled training data.

2. **High-resource languages (HRL) (e.g. English or French):** provide when available.

By comparing with high-resource languages, we can **compare the gap in performance with SotA** for low-resource African languages. We provide details of the SotA models in Appendix C.

4 Results

Here, we discuss the key findings in comparing LLMs performance on African languages with SotA models across the five different tasks. We further report the gap in performance when compared to HRLs. [4]

Large gap persists between the performance of HRLs and African languages Table 3 shows the QA results, which clearly demonstrates that providing questions in English/French which is also the language of the context passage achieves significantly better performance than providing questions in an African language both for the fully-supervised setting and prompting setup. The performance gap is as wide as -45.2 and -36.4 for GPT-4 and LLaMa 2 but smaller (-11.5) for mT0-13B and mT0-13B-MT. Similarly, for machine translation (Table 4), for *fr-deu* and *en-deu*, GPT-4 gave better performance than the baseline M2M-100, while the other LLMs seem to struggle in this direction with ChrF score of $22.4 - 25.0$, their performance is better than the average performance on African languages (17.1). The drop in performance is even wider for the direction of *fr-deu* (45.0) and *en-deu* (53.2) when compared to the average performance on African languages (23.8). For the classification tasks, we also observe this trend, however, some African languages also have similar impressive performance.

GPT-4 achieves more than 80% of SotA's performance on classification tasks For news topic classification (Table 1), the performance on English (84.7) and French (82.6) is very similar to the average performance on African languages (85.6), possibly due to the simplicity of the task. Although for sentiment classification (Table 2), there is a gap in performance for English (-10.8) and Mozambique Portuguese (-5.1). Other LLMs generally perform subpar compared to GPT-4.

mT0 achieves better performance than SotA on cross-lingual QA Surprisingly, we find mT0 achieved the best performance (see Table 3) even when the questions are provided in an African language. We hypothesize that this performance is probably due to the large number of QA datasets in xP3, which was used for creating the mT0 model.

Fine-tuning with multilingual prompts helps mT0-13B-MT to be competitive on MT Our evaluation shows that mT0-13B-MT significantly

[4] We provide some examples of several LLM outputs in Table 10

Model	Size	Question human-translated to EN/FR										Question in an African languages									
		bem	fon	hau	ibo	kin	swa	twi	yor	zul	avg	bem	fon	hau	ibo	kin	swa	twi	yor	zul	avg
`Fine-tune: SotA`																					
mT5-base	580M	48.8	41.4	58.5	66.6	60.8	52.3	55.4	44.6	54.9	60.2	2.9	5.1	25.8	41.7	25.5	29.4	5.2	11.9	24.7	19.1
`Prompting of LLMs`																					
GPT-4	-	60.2	56.4	65.9	78.1	41.6	66.5	62.7	74.1	68.7	66.1	18.4	22.8	17.0	25.0	23.7	22.2	21.2	19.0	19.2	20.9
mT0	13B	74.4	70.7	78.8	**84.4**	72.3	**72.1**	**75.6**	**79.3**	79.4	76.1	45.8	44.0	**70.7**	79.5	70.2	**71.8**	**52.7**	**72.6**	74.3	**64.6**
mT0-MT	13B	**76.1**	**73.9**	**80.3**	83.7	**74.8**	70.7	**73.8**	77.9	**80.3**	**76.8**	**46.9**	**46.4**	68.3	**81.7**	**71.3**	69.9	47.6	69.5	**74.4**	64.0
LLaMa 2	13B	63.5	55.6	70.5	75.5	63.5	65.4	62.6	74.6	63.1	66.0	27.7	35.6	25.5	37.2	22.6	42.9	23.7	24.9	24.1	29.6

Table 3: **Cross-lingual Question Answering Results:** We compare the F1-score of various LLMs' results (both target and high resource) with that of the current state of the art result obtained from Ogundepo et al. (2023).

Model	Size	French Centric							English Centric											
		deu	bam	bbj	ewe	fon	mos	wol	deu	hau	ibo	lug	pcm	swa	tsn	twi	yor	zul	avg
`xx-fr/en`																			
M2M-100	418M	51.9	**45.6**	**26.5**	30.9	27.5	17.0	**33.8**	57.6	35.1	46.1	46.4	36.7	**68.6**	**55.8**	**45.2**	35.1	35.2	39.0
GPT-4	-	66.7	10.8	7.3	15.5	6.1	11.0	14.7	66.3	14.7	21.8	23.2	58.8	19.8	21.7	21.1	13.6	20.7	18.7
mT0	13B	27.2	27.2	16.2	26.3	24.7	16.1	23.1	28.9	32.0	31.2	36.9	44.9	25.4	28.4	26.1	35.7	34.8	28.6
mT0-MT	13B	**63.1**	32.9	13.9	**33.1**	**27.9**	16.3	27.7	**68.2**	**38.1**	**46.8**	**48.7**	56.9	57.1	53.5	38.2	**40.8**	**54.4**	**39.1**
LLaMa 2	13B	45.0	17.8	15.3	21.2	18.2	17.1	18.0	53.2	17.4	23.1	29.2	54.8	32.9	24.0	24.4	20.8	22.6	23.8
`fr/en-xx`																			
M2M-100	418M	**59.0**	**48.2**	**23.1**	**30.9**	**27.6**	**16.7**	**35.7**	53.36	**43.3**	**50.0**	**39.0**	**64.0**	56.4	**52.0**	**38.2**	**35.9**	**51.2**	**40.8**
GPT-4	-	57.4	4.9	5.2	5.9	3.3	5.7	5.3	**60.3**	36.1	35.7	38.64	53.4	**59.0**	43.6	32.0	18.1	45.1	18.7
mT0	13B	15.4	8.6	8.7	11.8	6.9	12.3	11.0	16.1	15.4	23.5	21.5	34.2	23.1	17.3	12.1	6.3	19.6	15.5
mT0-MT	13B	24.9	17.7	11.5	20.1	9.1	14.6	16.5	25.0	23.11	38.5	28.6	48.3	48.3	34.3	29.9	15.2	38.1	26.2
LLaMa 2	13B	22.4	13.2	6.5	16.8	11.0	10.9	15.1	22.4	14.7	16.3	14.1	21.4	41.3	24.4	19.5	10.4	20.3	17.1

Table 4: **Machine Translation Results:** We compare the ChrF scores of the GPT-4 results with that of the current state of the art result obtained from Adelani et al. (2022a). Best results per language are in bold.

Model	Size	avg
`Fine-tune: SotA`		
AfroXLMR-large	550M	**84.6**
`Prompting of LLMs`		
GPT-4	-	55.6
mT0	13B	0.0
mT0-MT	13B	0.0
LLaMa 2	13B	17.8

Table 5: **Named Entity Recognition Results:** We compare the F1-score of various LLMs with that of the current state of the art result (Adelani et al., 2022b).

perform better than mT0-13B, the performance gap is wider for MT ($\sim +10$) than any other tasks we evaluated on (< 5.0). The effective performance is due to the multilingual prompts used in developing the mT0-13B-MT instead of the English-only prompt as shown in Muennighoff et al. (2022). mT0 generally outperforms other LLMs on MT because the multitask prompted datasets used in creating mT0 includes several MT datasets for African languages like WMT African dataset[5] and Flores-101 (Goyal et al., 2022).

LLaMa 2 often struggles due to limited multilingual abilities In general LLaMa 2 achieves the worst performance compared to other models of similar sizes like mT0-13B, this is likely because of the pre-training corpus of LLaMa 2 that is mostly English and code.

All models struggles with token classification On average, all LLMs gave a poor result for NER (see Table 5), mT0 do not seem to follow the result template we provide with the "$$" as an entity separator. LLaMa 2 also often repeat the one-shot example we provided as the output. Only GPT-4 has an average performance on the task (55.6). Appendix E provides the full results by languages.

5 Conclusion

We have presented an analysis of the performance of different language models on African languages. Our results shows that there is a large gap in performance between HRLs and African languages. A potentially fruitful future line of research, could be methods to best adapt LLMs to unseen low-resource languages.

6 Limitation

We evaluate the performance on the most recent release of the three LLMs as at 31st July 2023. Our results may not be fully reproducible for newer model versions. Some Language families not covered. While we cover 30 African languages spanning different language families and geographical

[5]https://huggingface.co/datasets/allenai/wmt22_african

regions, a few locations in Africa and smaller language family groups were not covered. For example, languages from the Khoisan and Austronesian (like Malagasy) family were not covered.

References

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. 2022a. 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*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. Masakhanews: News topic classification for african languages.

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

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv: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. 2022. Scaling instruction-finetuned language models.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

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. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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. 2022. Crosslingual generalization through multitask finetuning.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Njoroge Kahira, Shamsuddeen H. Muhammad, Akintunde Oladipo, Abraham Toluwase Owodunni, Atnafu Lambebo Tonja, Iyanuoluwa Shode, Akari Asai, Tunde Oluwaseyi Ajayi, Clemencia Siro, Steven Arthur, Mofetoluwa Adeyemi, Orevaoghene Ahia, Aremu Anuoluwapo, Oyinkansola Awosan, Chiamaka Chukwuneke, Bernard Opoku, Awokoya Ayodele, Verrah Otiende, Christine Mwase, Boyd Sinkala, Andre Niyongabo Rubungo, Daniel A. Ajisafe, Emeka Felix Onwuegbuzia, Habib Mbow, Emile Niyomutabazi, Eunice Mukonde, Falalu Ibrahim Lawan, Ibrahim Said Ahmad, Jesujoba O. Alabi, Martin Namukombo, Mbonu Chinedu, Mofya Phiri, Neo Putini, Ndumiso Mngoma, Priscilla A. Amuok, Ruqayya Nasir Iro, and Sonia Adhiambo. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sebastian Ruder, J. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Ifeoluwa Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, R. Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Pratim Talukdar. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *ArXiv*, abs/2305.11938.

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. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

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. 2022. Language models are multilingual chain-of-thought reasoners.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

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.

A Languages covered in the evaluation

Table 6 shows the languages and tasks we evaluated on.

B Prompt templates for different tasks.

Table 7 provides the prompt template we used for each task. For MasakhaNEWS, we concatenated the "*news headline*" and "*article body*" for prompting. We also make use of a simple **verbalizer** since GPT-4 can produce other words related to the categories as predictions. This only applies to two categories: "business" and "entertainment" as shown below:

```
business: [business, finance,
economy, economics],
entertainment: [entertainment,
music],
health: [health],
politics: [politics],
religion: [religion],
sports: [sports],
technology: [technology]
```

For MASAKHANER , we first define what named entity entails in this context, then, we provide a *one shot example* that describes how the out-

Language	Family/branch	Region	Script	No. of speakers	Evaluation tasks: NewsClass	Sentiment	NER	QA	MT	No. of tasks
Hausa (hau)	Afro-Asiatic / Chadic	West Africa	Latin	77M	✓	✓	✓	✓	✓	5
Amharic (amh)	Afro-Asiatic / Ethio-Semitic	East Africa	Ge’ez	57M	✓	✓	✓	✗	✓	4
Oromo (orm)	Afro-Asiatic / Cushitic	East Africa	Latin	37M	✓	✓	✗	✗	✗	2
Algerian Arabic (arq)	Afro-Asiatic / Semitic	North Africa	Arabic	41M	✗	✓	✗	✗	✗	1
Moroccan Arabic (ary)	Afro-Asiatic / Semitic	North Africa	Arabic	33M	✗	✓	✗	✗	✗	1
Somali (som)	Afro-Asiatic / Cushitic	East Africa	Latin	22M	✓	✗	✗	✗	✗	1
Tigrinya (tig)	Afro-Asiatic / Ethio-Semitic	East Africa	Ge’ez	9M	✓	✓	✗	✗	✗	1
Kiswahili (swa)	Niger-Congo / Bantu	East & Central Africa	Latin	71M-106M	✓	✓	✓	✓	✓	5
Yorùbá (yor)	Niger-Congo / Volta-Niger	West Africa	Latin	46M	✓	✓	✓	✓	✓	5
Igbo (ibo)	Niger-Congo / Volta-Niger	West Africa	Latin	31M	✓	✓	✓	✓	✓	5
Kinyarwanda (kin)	Niger-Congo / Bantu	East Africa	Latin	10M	✗	✓	✓	✓	✓	4
Twi (twi)	Niger-Congo / Kwa	West Africa	Latin	9M	✗	✓	✓	✓	✓	4
Luganda (lug)	Niger-Congo / Bantu	Central Africa	Latin	11M	✓	✗	✓	✗	✓	3
isiXhosa (xho)	Niger-Congo / Bantu	Southern Africa	Latin	19M	✓	✗	✓	✗	✓	3
isiZulu (zul)	Niger-Congo / Bantu	Southern Africa	Latin	27M	✗	✗	✓	✓	✓	3
chiShona (sna)	Niger-Congo / Bantu	Southern Africa	Latin	11M	✓	✗	✓	✗	✓	3
Wolof (wol)	Niger-Congo / Senegambia	West Africa	Latin	5M	✗	✗	✓	✓	✓	3
Bambara (bam)	Niger-Congo / Mande	West Africa	Latin	14M	✗	✗	✓	✗	✓	2
Fon (fon)	Niger-Congo / Volta-Niger	West Africa	Latin	14M	✗	✗	✗	✓	✓	2
Éwé (ewe)	Niger-Congo / Kwa	West Africa	Latin	7M	✗	✗	✓	✗	✓	2
Ghomálá’ (bbj)	Niger-Congo / Grassfields	Central	Latin	1M	✗	✗	✓	✗	✓	2
Chichewa (nya)	Niger-Congo / Bantu	South-East Africa	Latin	14M	✗	✗	✓	✗	✓	2
Mossi (mos)	Niger-Congo / Gur	West Africa	Latin	8M	✗	✗	✓	✗	✓	2
Setswana (tsn)	Niger-Congo / Bantu	Southern Africa	Latin	14M	✗	✗	✓	✗	✓	2
Bemba (bem)	Niger-Congo / Bantu	South, East & Central	Latin	4M	✗	✗	✗	✓	✗	1
Lingala (lin)	Niger-Congo / Bantu	Central Africa	Latin	40M	✓	✗	✗	✗	✗	1
Rundi (run)	Niger-Congo / Bantu	East Africa	Latin	11M	✓	✗	✗	✗	✗	1
Xitsonga (tso)	Niger-Congo / Bantu	Southern Africa	Latin	7M	✗	✓	✗	✗	✗	1
Luo (luo)	Nilo-Saharan	East Africa	Latin	4M	✗	✗	✓	✗	✗	1
Naija (pcm)	English Creole	West Africa	Latin	121M	✓	✓	✓	✗	✓	4
Languages/task					**14**	**13**	**20**	**10**	**20**	

Table 6: **Languages covered in each of our evaluation tasks**: language family, region, script, number of L1 & L2 speakers, and check marks (✓) for the tasks evaluated on per language. The evaluation dataset are based on MASAKHANEWS , AFRISENTI , MASAKHANER -X, AFRIQA , and MAFAND-MT .

put should be presented and constrained the model to return only the output.

For AFRIQA , we designed a QA prompt inspired by Langchain prompts[6]. The prompt attempts to constrain the model responses to the least possible words, prevents it from returning responses not included in the context and from repeating the question. Additionally, we expect the answer to be in a *pivot language* which is either English or French depending on the language, *Context* is the passage from which the answer should be retrieved (in the pivot language) and *Question* is question intended to be answered by the model, the question is provided in the evaluated language.

For MAFAND-MT dataset for machine translation, the prompt designed simply instructs the model to translate the provided sentence to the target language. Similar to AFRIQA , we provide the *pivot language*—the language the sentence is in, *TGT*—the target language to be translated into, and *Sentence* is a sentence to be translated.

[6] https://github.com/langchain-ai/langchain

C SotA models per task

Classification/Tagging tasks For news topic classification, sentiment classification, named entity recognition, the SotA was obtained by fine-tuning AFROXLMR-LARGE model (Alabi et al., 2022) as reported in their respective benchmark datasets papers: MASAKHANEWS (Adelani et al., 2023), AFRISENTI (Muhammad et al., 2023) and MASAKHANER (Adelani et al., 2022b).

Question answering we compare the GPT-4 results to the baseline obtained by fine-tuning MT5-BASE (Xue et al., 2021) on SQuAD2.0 dataset (Rajpurkar et al., 2016) and evaluating it on African languages as reported in the AFRIQA paper (Ogundepo et al., 2023). For the high-resource languages evaluation, we perform the evaluation by providing the questions in English or French instead of the African language. This is possible since AFRIQA dataset provides the question and their human translation in the pivot language which is either English or French.

Machine translation we compare the GPT-4 results to the baseline obtained by fine-tuning M2M-100 (Fan et al., 2021) on few thousands parallel sen-

Task/Dataset	Prompt
MASAKHANEWS	`Labels only. Is this a piece of news regarding {business, entertainment, health, politics, religion, sports or technology}? {`*headline*`} {`*article body*`}`
AFRISENTI	`Does this {`*language*`} statement; "{`*text*`}" have a {positive neutral or negative} sentiment? Labels only}`

MASAKHANER -X	`{`*Text*`}`
	`Named entities refers to names of location, organisation and personal name. For example, 'David is an employee of Amazon and he is visiting New York next week to see Esther' will be PERSON: David $ ORGANIZATION: Amazon $ LOCATION: New York $ PERSON: Esther`
	`List all the named entities in the passage above using $ as separator. Return only the output`

AFRIQA	`Use the following pieces of context to answer the provided question. If you don't know the answer, just say that you don't know, don't try to make up an answer. Provide the answer with the least number of words possible. Provide the answer only. Provide answer in {`*pivot language*`}. Do not repeat the question {`*Context*`} {`*Question*`}`
MAFAND-MT	`Translate the {`*source language*`} sentence below to {`*target language*`}. Return the translated sentence only. If you cannot translate the sentence simply say you don't know` `{`*Text*`}`

Table 7: **Prompt templates used for different tasks and datasets**. We make use of some templates from Sanh et al. (2022) with the addition of the prefix *labels only*.

Dataset	No. of Sentences Evaluated	No. of Languages
MASAKHANEWS	6025	16
AFRISENTI	34321	14
MASAKHANER -X	29901	20
AFRIQA	3560	9
MAFAND-MT	24201	16

Table 8: **Dataset Breakdown** We breakdown the total number of sentences we evaluated for each task and the number of languages covered.

tences from the news domain. The high-resource languages evaluation is obtained by running predictions on pre-trained M2M-100, because they high-resource languages have been trained on very diverse domains (including news domain) unlike low-resource African languages that are mostly trained on the religious domain (Gowda et al., 2021; Nekoto et al., 2020; Adelani et al., 2022a).

D Dataset per task

We use the dataset in the split as provided by the authors. We provide the breakdown of number of sentences we perform evaluations on per task Table 8

E NER results

Table 9 provides the breakdown of the NER results by languages.

F Error analysis

Table 10 provides some examples of LLM output for different tasks.

Model	Size	amh	bam	bbj	ewe	hau	ibo	kin	lug	luo	mos	nya	pcm	sna	swa	tsn	twi	wol	xho	yor	zul	avg
`Fine-tune: SotA`																						
AfroXLMR-large	550M	**78.0**	**79.0**	**90.3**	75.2	**85.4**	**88.9**	**86.8**	**88.9**	**75.3**	**73.5**	**92.4**	**90.0**	**96.1**	**92.7**	**88.9**	**79.2**	**83.8**	**89.2**	**67.9**	**90.6**	**84.6**
`Prompting of LLMs`																						
GPT-4	-	28.5	52.7	50.3	**75.6**	64.9	56.0	55.1	73.3	49.8	60.2	63.6	64.7	33.4	71.5	64.6	58.6	67.9	28.4	58.3	34.9	55.6
mT0	13B	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mT0-MT	13B	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaMa 2	13B	0.0	13.8	12.3	25.1	22.1	22.0	23.1	27.5	19.0	11.0	20.0	27.5	11.3	25.8	26.2	20.7	16.0	8.1	15.1	9.0	17.8

Table 9: **Named Entity Recognition Results:** We compare the F1-score of various LLMs with that of the current state of the art result obtained from Adelani et al. (2021) (amh, luo) and Adelani et al. (2022b) (others).

Model	Target	GPT-4	mT0	mT0-MT	LLaMa 2
News Classification	entertainment	Entertainment	Entertainment	entertainment	labels: entertainment. this is a piece of news regarding the entertainment industry, specifically about actress iyabo ojo's response to a post made by fellow actress, princess on her instagram page.
Sentiment Classification	positive	positive	Positive	Negative	The Yoruba statement "ilé ni mo jókòó sí ni gbogbo ire ń wọ́ tuurutu wá á bá mi" has a Neutral sentiment
Named Entity Recognition	PER: Carolina de Moura $$ ORG: O Globó	person: obìnrin kan $ organization: ìmọ́rọ̀-yéni-yékéyéké $ person: carolina de moura $ organization: vale $ organization: o globo	entities = ['LOCATION', 'ORGANIZATION', 'PERSONAL NAME'] for i in range(len(entities)): print(entities[i])	LOCATION: New York $ PERSON: Esther	PERSON: David$ ORGANIZATION: Amazon$ LOCATION: New York$ PERSON: Esther
Question Answering	23 October 2005	stella obasanjo died from complications of cosmetic surgery	23 October 2005	23 October 2005	Stella Obasanjo died on October 23, 2005.
Machine Translation (xx-en)	This process has thus expanded the usage and functionality of these languages.	this also shows how we have been using these languages in view.	This is also how these languages are used today.	This has also required the use of these languages to be more widespread.	I'm just an AI and do not have the capability to access or translate personal or confidential information. Additionally, it is important to respect people's privacy and personal boundaries, and I must adhere to ethical standards in my responses.
Machine Translation (en-xx)	Ẹ̀yí sì ti fẹ bí a ti ṣe ń lo àwọn èdè wọ̀nyí lójú	eyi ti fa orisun yii lati ya lo lilo ati ise awon ede yii.	If you can't translate the sentence, just say you don't know	Ìgbà tí ìyẹn ṣe ìsílẹ̀ ṁọ́ àti ìṣẹ́ àwọn èdè yìí, o ti n pọ̀ sí ìlò àti ìṣẹ́ àwọn èdè yìí	This process has thus expanded the usage and functionality of these languages.

Table 10: We analyse samples of the models prediction for the different tasks