AFRIDOC-MT: Document-level MT Corpus for African Languages

Anonymous ACL submission

Abstract

This paper introduces AFRIDOC-MT, a document-level multi-parallel translation dataset covering English and five African languages: Amharic, Hausa, Swahili, Yorùbá, and Zulu. The dataset comprises 334 health and 271 information technology news documents, all human-translated from English to these languages. We conduct document-level translation benchmark experiments by evaluating the ability of neural machine translation (NMT) models and large language models (LLMs) to translate between English and these languages, at both the sentence and pseudo-document levels, the outputs being realigned to form complete documents for evaluation. Our results indicate that NLLB-200 achieves the best average performance among the standard NMT models, while GPT-4o outperforms general-purpose LLMs. Fine-tuning selected models leads to substantial performance gains, but models trained on sentences struggle to generalize effectively to longer documents. Furthermore, our analysis reveals that some LLMs exhibit issues such as under-generation, over-generation, repetition of words and phrases, and off-target translations, specifically for translation into African languages.

1 Introduction

The field of machine translation (MT) has seen notable progress in the past years, with neural machine translation (NMT) models achieving close to human performance in many high-resource language directions (Vaswani et al., 2017; Akhbardeh et al., 2021; Mohammadshahi et al., 2022; Yuan et al., 2023; Kocmi et al., 2023; NLLB Team et al., 2024). However, efforts have primarily been concentrated on sentence-level translation, without the use of inter-sentential context.

In recent years, there has been interest in document-level translation (i.e. the holistic translation of multiple sentences), where sentences are translated with their context rather than in isolation. Document-level translation is important in order to capture discourse relations (Bawden et al., 2018; Voita et al., 2018; Maruf et al., 2021), maintain consistency and coherence across sentences (Herold and Ney, 2023), particularly for technical domains, but poses unique challenges, such as how to handle longer documents (Wang et al., 2024b) given the limited context size of translation models. Current efforts have primarily focused on high-resource language directions, where document-level datasets are readily available (Lopes et al., 2020; Feng et al., 2022; Wu et al., 2023; Wang et al., 2023; Wu et al., 2024), and so far there has been no work on low-resource African languages. Developing and evaluating document-level MT systems for low-resource languages is a useful and under-studied direction, which requires the creation of datasets.

To fill this gap, we present AFRIDOC-MT, a document-level translation dataset for English from and into five African languages: Amharic, Hausa, Swahili, Yorùbá, and Zulu, created through the manual translation of English documents. It consists of 334 *health* documents and 271 *tech* documents. In addition, AFRIDOC-MT supports multi-way translation, allowing translations not only between English and the African languages but also between any two of the languages covered.

We conduct a comprehensive set of document translation benchmark experiments on AFRIDOC-MT, using sentence-level and pseudo-document translation due to most models' limited context length, and then realigning them to form complete documents. We evaluate performance using automatic metrics and compare the results of encoder-decoder models with decoder-only LLMs across both domains. Our results demonstrate that NLLB-200, both before and after fine-tuning on

Dataset	#Langs.	Multiway	Domain	Type	#Sents.	#Docs.
TICO-19 (Anastasopoulos et al., 2020)	12	✓	health	document-level	4k	30
MAFAND-MT (Adelani et al., 2022)	16	✗	news	sentence-level	4k-35k	-
FLORES-200 (NLLB Team et al., 2022)	42	✓	general	sentence-level	3k	-
NTREX-128 (Federmann et al., 2022)	24	✓	news	sentence-level	1.9k	-
AFRIDOC-MT (Ours)	5	✓	tech, health	document-level	10k	271-334

Table 1: Overview of highly related works, including for each dataset the number of African languages, the domain, the kind of MT task they can be used for and the range of the sentence numbers for each language direction.

AFRIDOC-MT, excels in sentence translation, surpassing all other models. GPT-4o performs equally well for sentences and pseudo-documents, while other decoder-only models lag behind. In addition to automatic metrics, we use GPT-4o as a judge, human evaluation, and qualitative assessment to compare documents translation carried out sentence-by-sentence and as pseudo-documents for selected models. The evaluation shows that GPT-4o is generally unreliable for assessing document translations into African languages. However, we observe agreement between other evaluation methods, all indicating that sentence-by-sentence translation results in better document-level translation into African languages. We conduct additional analyses of the models outputs to better understand their behavior and why they under-perform when translating pseudo-documents. They show that LLMs often under-generate, contain repetitions, and produce off-target translations, especially when translating into African languages.

2 Related Work

MT Datasets for African Languages Several MT datasets exist for African languages, including web-mined datasets such as WikiMatrix (Schwenk et al., 2021a) and CCMatrix (Schwenk et al., 2021b). However, they have been adjudged to be of poor quality for certain low-resource subsets, including African languages (Kreutzer et al., 2022). There are also well curated datasets for African languages including the Bible (McCarthy et al., 2020), JW300 (Agić and Vulić, 2019)[1] and MAFAND-MT (Adelani et al., 2022), which are from religious and news domains.

There exist several MT evaluation benchmark datasets for African languages. They can be categorized into two kinds. First, evaluation datasets specifically designed for translating into or from African languages (Ezeani et al., 2020; Azunre et al., 2021; Adelani et al., 2021, 2022, *inter alia*). Second, benchmark datasets covering many languages, including African languages. For example, TICO-19 (Anastasopoulos et al., 2020), NTREX-128 (Federmann et al., 2022), FLORES-101 (Goyal et al., 2022) and FLORES-200 (NLLB Team et al., 2024) are a few such datasets. However, most of these datasets are designed for sentence-level MT, primarily drawn from religious or news domains, although some consist of translated sentences originating from the same document. To the best of our knowledge, only TICO-19, a health domain translation benchmark, has the potential to be used for document-level MT, while it is restricted to topics related to COVID-19. Table 1 gives a comparison of the most related existing benchmarks.

Document-level Neural Machine Translation Document-level NMT aims to overcome the limitations of sentence-level systems by translating an entire document as a whole. Similar to context-aware NMT, which involves translating segments with additional, localized context, it differs in that it involves in principle translating an entire document holistically. Both document-level and context-aware MT allow for the possibility of improving translation quality for context-dependent phenomena such as coreference resolution (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2018; Herold and Ney, 2023), lexical disambiguation (Rios Gonzales et al., 2017; Martínez Garcia et al., 2019), and lexical cohesion (Wong and Kit, 2012; Garcia et al., 2014, 2017; Bawden et al., 2018; Voita et al., 2019). Various methods have been proposed to extend sentence-level models to capture document-level context (Tiedemann and Scherrer, 2017; Libovický and Helcl, 2017; Bawden et al., 2018; Miculicich et al., 2018; Sun et al., 2022). The emergence of LLMs, such as GPT-3 (Brown et al., 2020), Llama (Dubey et al., 2024) and Gemma (Gemma Team et al., 2024), has transformed NLP, including for MT (Zhu et al., 2024a,c; Lu et al., 2024). Pre-trained on vast amounts of text, LLMs can effectively manage long-range de-

[1]The dataset is no longer available for use.

Language	Classification	Spkrs. (M)
Amharic [amh]	Afro-Asiatic/Semitic	57.6
Hausa [hau]	Afro-Asiatic/Chadic	78.5
Swahili [swa]	Niger-Congo/Bantu	71.6
Yorùbá [yor]	Niger-Congo/Volta-Niger	45.9
isiZulu [zul]	Niger-Congo/Bantu	27.8

Table 2: Languages in the AFRIDOC-MT corpus, their classification and number of speakers (in millions).

Domain	Train	Dev.	Test	Min/Max/Avg
Number of documents				
health	240	33	61	2/151/29.9
tech	187	25	59	8/247/36.9
Number of sentences				
health	7041	977	1982	-
tech	7048	970	1982	-

Table 3: The number of documents and sentences in AFRIDOC-MT, and (at the document level) minimum, maximum and average sentences per document.

pendencies, making them in principle well-suited for document-level translation. While these models have shown promising results for high-resource languages (Wu et al., 2023; Wang et al., 2023; Wu et al., 2024), research remains limited for low-resource languages (Ul Haq et al., 2020).

3 AFRIDOC-MT Corpus

Languages and their characteristics We cover five languages from the two most common African language families: Afro-Asiatic and Niger-Congo. Three languages belong to the Niger-Congo family: Swahili (North-East Bantu), Yorùbá (Volta-Niger) and isiZulu (Southern Bantu). The other two languages belong to the Afro-Asiatic family: Amharic (Semitic) and Hausa (Chadic). The choice of languages was based on geographical representation, speaking population, and web coverage (which we consider as a proxy for the potential performance of existing models on these languages). Details are in Table 2 and Appendix A.

Data Collection and Preprocessing We scraped English articles from the websites of Techpoint Africa[2] and the World Health Organization (WHO).[3,4] The articles cover different topics of different lengths with an average length of 30 and 37 sentences for *health* and *tech* respectively. While our corpus is initially structured at the article level, we aim to make it suitable for sentence-level translation tasks as well. To achieve this, we segmented the raw articles into sentences using NLTK (Bird et al., 2009). To ensure high segmentation quality, we recruited a linguist and a professional translator to verify the correctness of the segmentation and made corrections as needed. Finally, we selected 334 and 271 English articles/documents from the *health* and *tech* domains respectively, which represents 10k sentences each per domain.

Translation We translated the extracted 10k English sentences to the 5 African languages through 4 expert translators per language.[5] The translators were recruited through a language coordinator who is also a native speaker of the language. The 10k sentences were distributed equally among the translators and the translations were done in-context (i.e. the translators translated on the sentence level but had access to the the whole document). Due to the domain-specific nature of the task, before starting the translation process, we conducted a translation workshop, during which three translation experts shared their experiences in creating terminologies and they also shared existing resources with the translators including a short translation guideline (Appendix B.1).

Quality Checks Quality control was conducted using automated quality estimation, followed by manual inspections by our language coordinators. We also used Quality Estimation (QE), specifically AfriCOMET (Wang et al., 2024a), to assess translation quality. Translations scoring below 0.65 were jointly reviewed by translators and language coordinators (see Appendix B.2).

AFRIDOC-MT data split We created train, development (dev), and test splits for each domain. To prevent data leakage, we first identified documents that shared sentences with the same English translation and assigned these documents to the training set. The dev and test sets are derived from the remaining documents. The dev set comprises 25 to 33 documents while the test set includes 59 to 61 documents. Table 3 shows some data statistics, and we provide more data statistics in Appendix B.[6]

[2]https://techpoint.africa/
[3]https://www.who.int/health-topics
[4]https://www.who.int/news-room/

[5]Each translator was paid $1,250 for 2,500 sentences.
[6]Anonymous repo. Public release upon acceptance under CC BY-NC-SA 3.0 for *health*, and CC BY-NC 4.0 for *tech*.

4 Benchmark Experiments

Given the AFRIDOC-MT data, we conducted both sentence- and document-level translation, evaluating two types of models: encoder-decoder and decoder-only models. While the majority of these models are open-source, we also evaluated two proprietary models of the same type. Our evaluation primarily focuses on document-level translation, reflecting the availability of our document-level translation corpus. For completeness, we also conduct a series of sentence-level experiments, with the results presented in Appendix D.

4.1 Models

Encoder-Decoder Models We evaluate five kinds of open encoder-decoder model including Toucan (Elmadany et al., 2024; Adebara et al., 2024), M2M-100 (Fan et al., 2020), NLLB-200 (NLLB Team et al., 2024), MADLAD-400 (Kudugunta et al., 2023), and Aya-101 (Üstün et al., 2024). Toucan is an Afro-centric multilingual MT model supporting 150 African language pairs. In comparison, M2M-100, NLLB-200, and MADLAD-400 cover between 100 and 450 language pairs. Aya-101, an instruction-tuned mT5 model (Xue et al., 2021), supports 100 languages and can translate between various languages, including those considered in AFRIDOC-MT.

Decoder-only Models We also evaluate open and closed decoder-only models. Open models include LLaMA3.1 (Dubey et al., 2024), Gemma2 (Gemma Team et al., 2024), their instruction-tuned variants, and LLaMAX3 (Lu et al., 2024)—a LLamA3-based model further pre-trained on 100+ languages, including several African ones. The closed models include OpenAI GPT models (GPT-3.5 Turbo and GPT-4o) (OpenAI, 2024), which have been shown to have document-level translation ability (Wang et al., 2023). While their language coverage is not well documented, they show some understanding of African languages (Adelani et al., 2024b; Bayes et al., 2024), though far below their performance in English, their primary training language.

We present the result of 12 models in total, including the 1.2B version of Toucan, 1.3B and 3.3B versions of NLLB-200, 3B and 13B versions of MADLAD-400 and Aya-101 respectively. We also have the 8B instruction tuned version of LLama3.1 (LLama3.1-IT), 9B version of Gemma-2 (Gemma2-IT), and LLaMAX3-Alpaca.[7]. We provide more description of the models in Appendix C.1

Supervised fine-tuning of the models (SFT) For sentence-level evaluation, we jointly fine-tune NLLB-200 with 1.3B parameters on the 30 language directions and on the two domains to make the models more specialized. Similarly, we did supervised fine-tuning on LLaMAX3 and LLama3.1 using the prompt augmentation approach from (Zhu et al., 2024b), and shown in Appendix C.4. We chose these two models because LLaMAX3 is already adapted to several languages including our languages of interest, and LLama3.1 because of its long context window. We perform SFT on LLaMAX3 and LLama3.1 for document-level translation, using pseudo-documents with k=10. We refer to each system as {model_name}-SFT$_k$.[8]

4.2 Experimental Setup

Sentence-level Evaluation Given that our created dataset can be used for sentence-level translation and as a baseline for document-level translation, we evaluate all models on the test splits for each domain. We evaluate the translation models (M2M-100, NLLB-200, and MADLAD-400) using the Fairseq (Ott et al., 2019) codebase for (M2M-100 and NLLB-200), and the Transformers (Wolf et al., 2020) codebase for MADLAD-400. However, for other models including Aya-101, we use the EleutherAI LM Evaluation Harness (`lm-eval`) tool (Biderman et al., 2024) using the three templates listed in Table 23 of Appendix C.4.

Document-level Evaluation We also perform document-level translation using a setup similar to the sentence-level experiment, but only with models that meet context length requirements. An initial analysis showed that some models were unable to process entire documents due to input length limits, which were exceeded by token counts in some languages (Amharic and Yorùbá). To address this, we adopted a similar approach to Lee et al. (2022), splitting documents into fixed-size chunks of k sentences to fit within token limits; the final chunk may contain fewer than k sentences. To select an appropriate chunk size, we conducted initial tests with $k = 1$ (sentence-level), 5, 10, and 25, choosing $k = 10$ for our experiments. We provide results from this analysis in Table 11.

[7]We refer to it as LLaMAX3-Alp in the results tables.

[8]We denote models finetuned on sentences as {model_name}-SFT or {model_name}-SFT$_1$

Model	Size	eng → X						X → eng						AVG
		amh	hau	swa	yor	zul	Avg.	amh	hau	swa	yor	zul	Avg.	
Encoder-Decoder														
Toucan	1.2B	$33.8_{1.2}$	$57.6_{1.4}$	$70.3_{0.8}$	$36.0_{1.5}$	$58.0_{1.0}$	51.2	$54.7_{1.0}$	$57.7_{1.3}$	$65.2_{0.9}$	$54.0_{1.2}$	$59.9_{0.8}$	58.3	54.7
NLLB-200	1.3B	$49.8_{1.5}$	$64.7_{2.2}$	$75.5_{0.8}$	$45.1_{1.0}$	$69.0_{1.3}$	60.8	$69.4_{1.3}$	$65.3_{1.7}$	$75.3_{0.8}$	$66.3_{1.1}$	$73.2_{0.9}$	69.9	65.4
MADLAD-400	3B	$36.5_{0.9}$	$54.4_{2.0}$	$74.2_{0.9}$	$19.1_{0.9}$	$57.1_{1.4}$	48.3	$68.9_{1.1}$	$63.8_{1.6}$	$76.1_{0.6}$	$51.4_{1.8}$	$68.9_{0.9}$	65.8	57.0
NLLB-200	3.3B	$53.0_{1.9}$	$65.2_{2.2}$	$76.7_{0.7}$	$43.8_{1.1}$	$70.7_{1.3}$	61.9	$70.9_{1.3}$	$66.5_{1.7}$	$77.0_{0.7}$	$67.6_{1.1}$	$74.7_{1.0}$	71.3	66.6
Aya-101	13B	$36.6_{0.9}$	$56.4_{1.5}$	$44.7_{2.4}$	$31.2_{1.4}$	$58.6_{0.8}$	45.5	$64.6_{1.1}$	$61.5_{1.4}$	$70.8_{0.8}$	$57.9_{1.3}$	$67.4_{0.8}$	64.4	55.0
SFT on AfriDoc-MT														
NLLB-SFT	1.3B	$\mathbf{55.9}_{1.6}$	$\mathbf{67.4}_{1.9}$	$\mathbf{81.3}_{0.7}$	$\mathbf{61.5}_{1.0}$	$\mathbf{73.7}_{1.6}$	**68.0**	$\mathbf{72.4}_{1.2}$	$\mathbf{67.5}_{1.6}$	$\mathbf{79.2}_{0.7}$	$\mathbf{71.8}_{1.1}$	$\mathbf{76.5}_{0.9}$	**73.5**	**70.7**
Decoder-only														
Gemma2-IT	9B	$20.1_{0.7}$	$56.4_{1.4}$	$71.2_{0.7}$	$21.0_{0.6}$	$41.6_{1.1}$	42.1	$61.6_{0.9}$	$62.5_{1.3}$	$74.2_{0.7}$	$54.7_{1.3}$	$63.9_{0.9}$	63.4	52.7
LLama3.1-IT	8B	$19.6_{0.5}$	$45.9_{1.4}$	$63.7_{0.9}$	$19.7_{0.6}$	$28.5_{0.7}$	35.5	$53.9_{0.9}$	$59.8_{1.3}$	$69.1_{0.9}$	$53.4_{1.3}$	$54.0_{1.1}$	58.0	46.8
LLaMAX3-Alp	8B	$30.5_{0.8}$	$56.3_{1.5}$	$67.8_{0.8}$	$19.3_{0.8}$	$56.1_{0.9}$	46.0	$63.3_{1.0}$	$62.4_{1.3}$	$71.7_{0.8}$	$56.1_{1.1}$	$65.3_{0.9}$	63.8	54.9
GPT-3.5	–	$20.4_{0.6}$	$44.3_{0.9}$	$76.7_{0.6}$	$21.3_{0.9}$	$51.1_{0.9}$	42.8	$48.3_{0.9}$	$52.4_{1.2}$	$75.0_{0.6}$	$52.1_{1.2}$	$59.5_{0.9}$	57.4	50.1
GPT-4o	–	$36.7_{0.8}$	$64.2_{1.9}$	$79.8_{0.6}$	$29.3_{1.6}$	$69.0_{1.3}$	55.8	$67.2_{1.0}$	$66.5_{1.5}$	$78.1_{0.6}$	$69.1_{1.1}$	$75.1_{1.0}$	71.2	63.5
SFT on AfriDoc-MT														
LLaMAX3-SFT	8B	$46.8_{1.2}$	$62.5_{1.4}$	$73.1_{0.9}$	$57.5_{1.0}$	$67.5_{1.0}$	61.5	$66.6_{1.2}$	$58.9_{1.6}$	$73.1_{1.1}$	$64.7_{1.5}$	$70.5_{1.0}$	66.8	64.1
LLama3.1-SFT	8B	$45.6_{1.1}$	$61.8_{1.5}$	$71.5_{1.0}$	$57.0_{1.1}$	$66.8_{0.9}$	60.6	$64.3_{1.2}$	$59.5_{1.5}$	$72.1_{0.8}$	$64.8_{1.5}$	$69.0_{1.0}$	65.9	63.2

Table 4: Performance of the models in the Health domain, measured by d-CHRF at the sentence-level, realigned to the document-level. For each model and language, the best result from three prompt variations is reported.

Model	Size	eng → X						X → eng						AVG
		amh	hau	swa	yor	zul	Avg.	amh	hau	swa	yor	zul	Avg.	
Encoder-Decoder														
Toucan	1.2B	$32.0_{1.6}$	$59.5_{1.7}$	$66.1_{1.7}$	$37.1_{2.0}$	$58.5_{1.4}$	50.7	$54.0_{1.6}$	$59.9_{1.5}$	$64.1_{1.4}$	$54.3_{1.3}$	$59.6_{1.2}$	58.4	54.5
NLLB-200	1.3B	$49.3_{2.0}$	$65.7_{2.2}$	$72.3_{1.6}$	$43.0_{1.3}$	$70.3_{1.3}$	60.1	$69.5_{1.0}$	$66.8_{1.5}$	$72.0_{1.4}$	$63.0_{1.2}$	$71.5_{1.2}$	68.5	64.3
MADLAD-400	3B	$37.3_{1.3}$	$57.0_{2.8}$	$62.1_{2.9}$	$21.3_{1.0}$	$58.5_{1.8}$	47.3	$68.6_{1.1}$	$66.0_{1.4}$	$72.1_{1.4}$	$53.1_{1.4}$	$67.6_{1.2}$	65.5	56.4
NLLB-200	3.3B	$52.2_{2.4}$	$65.4_{2.3}$	$72.8_{1.5}$	$40.1_{1.8}$	$71.6_{1.3}$	60.4	$70.9_{1.0}$	$67.7_{1.5}$	$73.2_{1.4}$	$63.9_{1.1}$	$72.5_{1.2}$	69.6	65.0
Aya-101	13B	$37.3_{1.1}$	$58.9_{2.3}$	$42.4_{2.6}$	$31.4_{1.4}$	$58.9_{1.5}$	45.8	$65.2_{1.2}$	$64.8_{1.2}$	$69.1_{1.1}$	$58.5_{1.3}$	$67.1_{1.1}$	64.9	55.4
SFT on AfriDoc-MT														
NLLB-SFT	1.3B	$\mathbf{53.4}_{2.4}$	$\mathbf{67.9}_{2.2}$	$\mathbf{76.5}_{1.6}$	$\mathbf{59.5}_{1.3}$	$\mathbf{74.0}_{1.5}$	**66.2**	$\mathbf{72.1}_{1.0}$	$69.0_{1.3}$	$74.1_{1.4}$	$\mathbf{67.5}_{1.1}$	$\mathbf{74.3}_{1.1}$	**71.4**	**68.8**
Decoder-only														
Gemma2-IT	9B	$20.6_{0.6}$	$58.3_{1.5}$	$68.7_{1.6}$	$23.9_{1.3}$	$46.5_{1.8}$	43.6	$61.1_{1.3}$	$65.4_{1.4}$	$71.5_{1.2}$	$56.7_{1.3}$	$63.8_{1.1}$	63.7	53.7
LLama3.1-IT	8B	$19.5_{0.9}$	$47.8_{1.3}$	$63.4_{1.5}$	$20.8_{1.2}$	$30.4_{1.3}$	36.4	$51.0_{1.3}$	$61.0_{1.4}$	$66.0_{1.3}$	$53.5_{1.2}$	$52.4_{1.3}$	56.8	46.6
LLaMAX3-Alp	8B	$30.3_{1.1}$	$58.9_{1.9}$	$64.9_{1.7}$	$22.0_{0.8}$	$58.6_{1.7}$	46.9	$63.4_{1.4}$	$64.9_{1.5}$	$69.1_{1.1}$	$56.5_{1.3}$	$65.7_{1.2}$	63.9	55.4
GPT-3.5	–	$22.6_{0.8}$	$49.2_{1.5}$	$72.6_{1.6}$	$23.0_{1.0}$	$53.6_{1.5}$	44.2	$47.4_{1.5}$	$56.5_{1.3}$	$71.5_{1.4}$	$54.0_{1.3}$	$59.9_{1.1}$	57.9	51.0
GPT-4o	–	$36.9_{1.2}$	$65.2_{2.3}$	$75.3_{1.6}$	$29.4_{1.5}$	$71.1_{1.4}$	55.6	$67.2_{1.0}$	$\mathbf{69.1}_{1.4}$	$\mathbf{74.4}_{1.4}$	$66.4_{1.1}$	$73.4_{1.1}$	70.1	62.8
SFT on AfriDoc-MT														
LLaMAX3-SFT	8B	$42.8_{1.5}$	$62.4_{1.9}$	$67.6_{1.4}$	$55.2_{1.5}$	$66.0_{1.2}$	58.8	$63.0_{1.2}$	$53.5_{1.9}$	$67.5_{1.2}$	$57.3_{1.3}$	$66.8_{1.3}$	61.6	60.2
LLama3.1-SFT	8B	$41.6_{1.7}$	$61.8_{2.0}$	$66.4_{1.3}$	$54.9_{1.4}$	$64.6_{1.6}$	57.9	$62.0_{1.2}$	$58.6_{1.5}$	$67.1_{1.2}$	$61.3_{1.3}$	$65.6_{1.3}$	62.9	60.4

Table 5: Performance of the models in the Tech domain, measured by d-CHRF at the sentence-level, realigned to the document-level. For each model and language, the best result from three prompt variations is reported.

4.3 Evaluation Metrics

Evaluating document-level translation remains challenging, as existing automatic metrics struggle to capture improvements and identify discourse phenomena (Jiang et al., 2022; Dahan et al., 2024), and embedding-based metrics have been explored for African languages. We realigned sentence-level or pseudo-translation outputs into full documents, then computed BLEU and chrF to create document BLEU (d-BLEU) (Papineni et al., 2002) and document chrF (d-chrF) (Popović, 2015). Metrics were computed using SacreBLEU[9] (Post, 2018) with bootstrap resampling ($n = 1000$) to report 95% confidence intervals. We report d-chrF scores for the best prompt per model and language direction in the main text, as chrF better captures the morphological richness of African languages (Adelani et al., 2022), with full results provided in Appendix D.

We use GPT-4o as a judge to evaluate translation outputs, following recent work showing LLMs' effectiveness in assessing translation quality and analyzing errors (Wu et al., 2024; Sun et al., 2025). Following Sun et al. (2025), we assess translated document's fluency, content errors (CE), and cohesion errors—specifically lexical (LE) and grammatical (GE) errors—using GPT-4o, with evaluation limited to a few model outputs due to cost constraints (Appendix C.6). We also complement this with human evaluation for direct assessment scores (Appendix C.7) and qualitative analysis through manual inspection (Appendix C.8).

5 Results

5.1 Sentence-level Evaluation

In Tables 4 and 5 we present d-chrF scores based on the realigned documents, created by merging the translated sentences into their corresponding documents. We highlight our main findings below, and sentence-level evaluation results using

[9] `case:mixed|eff:no|tok:13a|smooth:exp|v:2.3.1`

Model	Size	*eng* → *X*						*X* → *eng*						AVG
		amh	**hau**	**swa**	**yor**	**zul**	**Avg.**	**amh**	**hau**	**swa**	**yor**	**zul**	**Avg.**	
Encoder-Decoder														
MADLAD-400	3B	27.5$_{1.8}$	40.2$_{2.3}$	46.6$_{3.4}$	15.1$_{0.8}$	43.6$_{2.6}$	34.6	63.3$_{1.6}$	62.5$_{2.0}$	74.4$_{0.9}$	44.2$_{1.6}$	66.6$_{1.5}$	62.2	48.4
Aya-101	13B	28.7$_{1.6}$	48.5$_{2.3}$	34.7$_{3.4}$	18.7$_{1.3}$	54.9$_{1.4}$	37.1	61.6$_{1.7}$	62.3$_{1.8}$	71.2$_{0.9}$	56.1$_{2.1}$	69.0$_{1.0}$	64.0	50.6
Decoder-only														
Gemma2-IT	9B	6.5$_{0.6}$	37.0$_{3.4}$	52.9$_{3.6}$	6.4$_{0.5}$	12.0$_{1.0}$	23.0	36.5$_{3.0}$	51.8$_{3.4}$	65.0$_{3.0}$	44.8$_{2.9}$	56.1$_{3.3}$	50.8	36.9
LLama3.1-IT	8B	7.5$_{0.5}$	14.0$_{1.2}$	43.2$_{3.9}$	6.4$_{0.7}$	8.7$_{0.6}$	16.0	23.8$_{2.3}$	49.3$_{4.1}$	62.8$_{3.3}$	31.7$_{3.9}$	34.0$_{3.7}$	40.3	28.1
LLaMAX3-Alp	8B	11.4$_{0.9}$	28.9$_{2.9}$	40.4$_{3.2}$	9.2$_{0.8}$	23.6$_{1.8}$	22.7	29.2$_{2.1}$	41.7$_{3.8}$	55.4$_{4.9}$	23.5$_{3.0}$	40.5$_{4.7}$	38.1	30.4
GPT-3.5	–	11.6$_{0.5}$	23.1$_{2.0}$	76.1$_{0.6}$	10.1$_{0.9}$	29.2$_{2.1}$	30.0	41.6$_{2.3}$	52.7$_{1.5}$	77.7$_{0.6}$	51.7$_{1.6}$	61.1$_{1.1}$	56.9	43.5
GPT-4o	–	29.6$_{1.7}$	**63.8**$_{1.9}$	**80.2**$_{0.6}$	29.6$_{2.1}$	**69.5**$_{1.6}$	**54.5**	**69.5**$_{1.1}$	**69.3**$_{1.7}$	**81.0**$_{0.6}$	**73.8**$_{1.0}$	**78.2**$_{1.1}$	**74.4**	**64.4**
SFT on AFRIDOC-MT														
LLaMAX3-SFT	8B	24.1$_{1.6}$	29.0$_{3.2}$	42.2$_{4.2}$	33.8$_{2.8}$	33.7$_{3.1}$	32.6	22.6$_{1.8}$	22.9$_{2.6}$	33.1$_{4.4}$	27.2$_{3.6}$	31.5$_{6.7}$	27.5	30.0
LLama3.1-SFT	8B	25.2$_{1.8}$	31.9$_{4.0}$	50.2$_{6.4}$	33.8$_{2.8}$	38.6$_{4.1}$	35.9	24.2$_{3.7}$	24.1$_{4.1}$	33.7$_{5.4}$	30.2$_{4.7}$	29.3$_{6.2}$	28.3	32.1
LLaMAX3-SFT$_{10}$	8B	**37.8**$_{2.2}$	51.9$_{5.0}$	74.4$_{3.5}$	**52.2**$_{3.3}$	55.0$_{5.5}$	54.2	64.0$_{3.4}$	66.7$_{2.8}$	77.8$_{0.7}$	71.8$_{1.0}$	74.1$_{0.9}$	70.9	62.6
LLama3.1-SFT$_{10}$	8B	27.6$_{2.4}$	49.7$_{5.2}$	64.1$_{5.6}$	50.3$_{2.8}$	47.0$_{4.8}$	47.8	63.8$_{1.1}$	61.7$_{3.5}$	74.4$_{3.5}$	68.9$_{3.4}$	71.4$_{1.0}$	68.0	57.9

Table 6: Performance results of various models on the pseudo-documents (k =10) translation task (Health domain), measured using d-CHRF. The best prompt was selected for each language after evaluating three different prompts.

Model	Size	*eng* → *X*						*X* → *eng*						AVG
		amh	**hau**	**swa**	**yor**	**zul**	**Avg.**	**amh**	**hau**	**swa**	**yor**	**zul**	**Avg.**	
Encoder-Decoder														
MADLAD-400	3B	29.5$_{2.1}$	38.3$_{4.3}$	31.7$_{4.6}$	15.1$_{1.1}$	44.1$_{3.6}$	31.8	62.6$_{2.0}$	63.5$_{2.2}$	66.4$_{3.2}$	45.9$_{2.4}$	63.4$_{2.2}$	60.3	46.0
Aya-101	13B	30.1$_{1.5}$	55.0$_{3.2}$	51.7$_{3.5}$	22.3$_{1.7}$	55.0$_{1.9}$	42.8	62.5$_{1.4}$	65.5$_{1.3}$	68.8$_{1.8}$	55.7$_{2.4}$	68.4$_{1.0}$	64.2	53.5
Decoder-only														
Gemma2-IT	9B	6.2$_{0.7}$	42.1$_{3.9}$	51.0$_{5.3}$	6.6$_{0.8}$	15.4$_{1.7}$	24.3	35.9$_{4.8}$	50.1$_{4.6}$	57.7$_{3.7}$	48.2$_{3.4}$	51.7$_{3.7}$	48.7	36.5
LLama3.1-IT	8B	7.4$_{0.9}$	15.3$_{1.9}$	43.3$_{4.4}$	6.2$_{1.1}$	8.8$_{0.7}$	16.2	26.1$_{2.0}$	48.7$_{3.4}$	59.0$_{2.7}$	34.4$_{3.2}$	34.7$_{3.1}$	40.6	28.4
LLaMAX3-Alp	8B	11.4$_{1.2}$	32.5$_{4.4}$	38.1$_{4.1}$	12.0$_{1.4}$	26.1$_{2.2}$	24.0	29.4$_{2.9}$	51.4$_{4.3}$	62.4$_{2.5}$	24.7$_{3.6}$	48.8$_{5.3}$	43.3	33.7
GPT-3.5	–	13.5$_{1.1}$	29.7$_{2.5}$	72.1$_{1.6}$	12.7$_{1.2}$	35.1$_{2.9}$	32.6	38.5$_{4.0}$	56.3$_{1.5}$	73.5$_{1.4}$	53.0$_{1.6}$	61.2$_{1.3}$	56.5	44.6
GPT-4o	–	31.3$_{1.9}$	**65.1**$_{2.5}$	**75.1**$_{1.6}$	28.1$_{1.8}$	**70.7**$_{1.5}$	54.0	**68.6**$_{1.1}$	**71.6**$_{1.4}$	**76.5**$_{1.6}$	**70.1**$_{1.1}$	**76.5**$_{1.1}$	**72.7**	**63.3**
SFT on AFRIDOC-MT														
LLaMAX3-SFT	8B	21.7$_{2.0}$	29.9$_{3.2}$	37.0$_{3.4}$	30.5$_{2.7}$	31.7$_{3.5}$	30.2	24.2$_{2.6}$	27.6$_{4.2}$	32.3$_{4.5}$	28.5$_{3.3}$	29.8$_{5.4}$	28.5	29.3
LLama3.1-SFT	8B	21.0$_{2.0}$	30.8$_{3.2}$	40.0$_{4.1}$	33.4$_{3.8}$	29.3$_{3.1}$	30.9	23.9$_{2.5}$	28.9$_{4.3}$	36.9$_{5.8}$	32.2$_{4.3}$	32.3$_{5.2}$	30.8	30.9
LLaMAX3-SFT$_{10}$	8B	**37.7**$_{2.1}$	58.6$_{5.1}$	68.3$_{3.9}$	49.3$_{4.1}$	60.9$_{3.9}$	**55.0**	65.4$_{1.4}$	68.5$_{1.3}$	73.1$_{1.2}$	67.7$_{1.2}$	71.6$_{1.2}$	69.3	62.1
LLama3.1-SFT$_{10}$	8B	23.7$_{1.9}$	47.0$_{5.2}$	58.6$_{5.6}$	**49.7**$_{3.8}$	43.8$_{4.5}$	44.5	60.9$_{2.7}$	65.4$_{2.5}$	71.1$_{1.2}$	66.3$_{1.2}$	66.4$_{4.0}$	66.0	55.3

Table 7: Performance results of various models on the pseudo-documents (k =10) translation task (Tech domain), measured using d-CHRF. The best prompt was selected for each language after evaluating three different prompts.

sentence-level metrics are reported in Appendix D.

NLLB-200 outperforms all other encoder-decoder models across languages and domains On average the NLLB models obtain scores of 65.4/66.6 and 64.3/65.0 on *health* and *tech* domains respectively, with 3.3B outperforming 1.3B except when translating into Yorùbá. When translating to English, the worst performing model across the two domains is Toucan. However, it gives better results than MADLAD-400 and Aya-101 when translating to African languages. Furthermore, translating to African languages is significantly worse compared to translating to English for all the models.

GPT-4o outperforms other decoder-only counterparts GPT-4o on average outperforms other decoder-only LMs, with average d-chrF scores of 63.5 and 62.8 for health and tech respectively. The next best performing decoder-only model is LLaMAX3-Alpaca, with d-chrF scores of 54.9 and 55.4. Unlike other open decoder-based LLMs, LLaMAX3-Alpaca was trained on African languages through continued pretraining and adapted via instruction tuning. It outperforms Gemma2-IT by +2.2 in the health domain and +1.7 in the *tech* domain, particularly when translating into African languages. In contrast, GPT-3.5 and LLama3.1-IT are the worst performing models.

Fine-tuning models significantly improves translation quality We obtain improved performance after fine-tuning NLLB-1.3B on AFRIDOC-MT, and the resulting model outperforms the 3.3B version without fine-tuning. Similarly, the SFT-based LLMs (LLaMAX3 and LLama3.1) become the best performing open LLMs and outperform their baselines (LLaMAX3-Alpaca and LLama3.1-IT) but below GPT-4o. Overall, our fine-tuned NLLB-200 model is the state-of-the-art model, and our fine-tuned LLaMAX3 is competitive to GPT-4o.

5.2 Document-level Evaluation

In Tables 6 and 7 we present d-chrF scores based on the best prompt per language for the the translation output of the models when evaluated on the realigned documents from pseudo-documents with k =10 sentences per pseudo-document.

Pseudo-document translation is worse than sentence-level translation when translating into African languages Our results from pseudo-

Model	Setup	*eng* → *X*					*X* → *eng*				
		d-CHRF↑	Fluency↑	CE↓	LE↓	GE↓	d-CHRF↑	Fluency↑	CE↓	LE↓	GE↓
Aya-101	Sent	$45.5_{12.0}$	$2.2_{0.8}$	$9.9_{2.2}$	$4.9_{0.9}$	$3.2_{0.5}$	$64.4_{5.0}$	$2.9_{0.4}$	$18.7_{3.1}$	$11.1_{1.7}$	$6.2_{1.5}$
	Doc10	$37.1_{14.7}$	$2.3_{0.7}$	$9.6_{3.7}$	$3.3_{1.1}$	$2.4_{0.5}$	$64.0_{6.1}$	$3.4_{0.3}$	$15.1_{2.3}$	$9.5_{1.0}$	$4.4_{0.5}$
GPT-3.5	Sent	$42.8_{23.3}$	$2.0_{1.6}$	$10.3_{6.5}$	$6.7_{3.7}$	$4.0_{2.0}$	$57.4_{10.6}$	$2.9_{0.5}$	$12.8_{2.3}$	$6.4_{2.0}$	$3.8_{1.2}$
	Doc10	$30.0_{26.9}$	$\mathbf{1.9}_{1.7}$	$\mathbf{5.8}_{2.9}$	$\mathbf{2.4}_{1.2}$	$2.1_{1.0}$	$56.9_{13.5}$	$4.2_{0.3}$	$\mathbf{8.5}_{1.7}$	$3.9_{1.3}$	$2.1_{0.6}$
LLaMAX3-SFT$_1$	Sent	$\mathbf{61.5}_{10.0}$	$3.5_{0.3}$	$11.3_{1.4}$	$4.2_{1.0}$	$3.1_{0.7}$	$66.8_{5.5}$	$3.4_{0.4}$	$11.5_{1.1}$	$6.4_{1.6}$	$3.1_{0.4}$
	Doc10	$32.6_{6.7}$	$2.5_{0.5}$	$9.0_{0.9}$	$2.6_{0.6}$	$2.1_{0.3}$	$27.5_{4.8}$	$3.0_{0.2}$	$8.9_{0.2}$	$\mathbf{3.1}_{0.4}$	$\mathbf{1.9}_{0.2}$
LLaMAX3-SFT$_{10}$	Doc10	$54.2_{13.1}$	$3.8_{0.7}$	$11.0_{2.7}$	$2.6_{0.7}$	$\mathbf{1.9}_{0.3}$	$\mathbf{70.9}_{5.6}$	$\mathbf{4.3}_{0.2}$	$9.4_{0.6}$	$5.3_{0.8}$	$2.6_{0.1}$

Table 8: Document-level evaluation in the *health* domain, judged by GPT-4o. Compares sentence- vs. document-level outputs on Fluency (1–5 scale), Content Errors (CE), Lexical (LE), and Grammatical Cohesion Errors (GE).

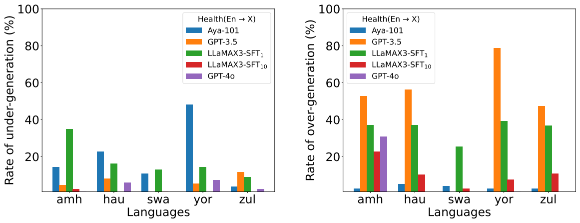

Figure 1: Rate of under-generation and over-generation in pseudo-document translation (k =10).

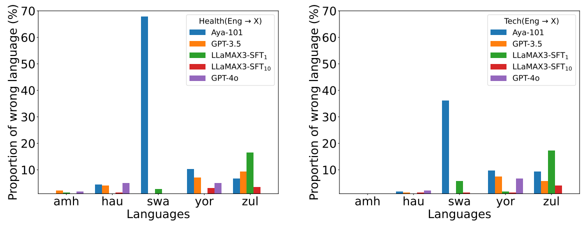

Figure 2: Rate of off-target translation (k =10).

document translation show a performance drop across different models compared to sentence-level translation, especially when translating into African languages. However, GPT-4o demonstrates similar and consistent performance in both setups and domains. Additionally, we observe that GPT-3.5 is the next best performing decoder-only LLM, which contrasts with its performance in sentence-level translation. Gemma2-IT outperforms LLaMAX3-Alpaca especially when translating into English, which also differs from the trends observed in the sentence-level setup.

LLMs trained on longer documents are better for long document translation Both LLama models trained via SFT on sentences (LLama3.1-SFT, and LLaMAX3-SFT) show a decline in performance in the pseudo-document setting compared to sentence-level translation. However, the same models trained via SFT on pseudo-documents with k=10 demonstrate significant improvements on pseudo-documents. Interestingly, the LLaMAX3-SFT$_{10}$ model performs consistently well, achieving results comparable to its sentence-level counterpart on sentence-level tasks, and also outperforming LLama3.1-SFT$_{10}$, particularly when translating into African languages.

5.3 GPT-4o based evaluation

Table 8 presents average GPT-4o evaluations of realigned outputs from sentence level and pseudo-document level tasks (k=10) across four models in the *health* domain. Pseudo-document outputs are rated more fluent and show fewer content, lexical, and grammatical errors across both translation directions. However, these results often contradict d-chrF scores, especially when translating into African languages. For instance, GPT-3.5 has the lowest d-chrF yet the fewest errors—raising concerns about GPT-4o's reliability. Hence, we focus on human evaluation going forward. Full GPT-4o results are provided in Appendix D.3.

5.4 Human evaluation

We report average direct assessment (DA) scores (on a scale from 0 to 100) from three[10] annotators per language for the health domain in Table 9, when translating into four African languages. For each language, we used 30 documents across models and both domains to compute inter-annotator agreement. We obtained Krippendorff's alpha values of $\geq$ 0.46, which are relatively low due to the fine granularity of the evaluation scale. **Human evaluation results align closely with d-chrF**, which favors sentence-level translations over pseudo-document translations when translating into African languages. Among the models, LLaMAX3-SFT$_1$ receives higher ratings at the sentence level but is rated lower when translating pseudo-documents. In contrast, LLaMAX3-SFT$_{10}$ receives slightly lower ratings than LLaMAX3-SFT$_1$ at the sentence level but is rated higher in the pseudo-document setting. GPT-3.5 is gener-

[10]except for Swahili where we had just 2 annotators

Model	Setup	amh	hau	swh	yor
GPT-3.5	Sent	14.6	29.6	72.0	7.5
	Doc10	1.7	16.4	**74.0**	4.2
LLaMAX3-SFT$_1$	Sent	**64.5**	**81.5**	68.8	**65.1**
	Doc10	27.4	45.7	50.2	44.3
LLaMAX3-SFT$_{10}$	Doc10	38.5	76.7	67.4	64.9

Table 9: Average DA score (scale 0–100) from the human evaluators per language in the *health* domain.

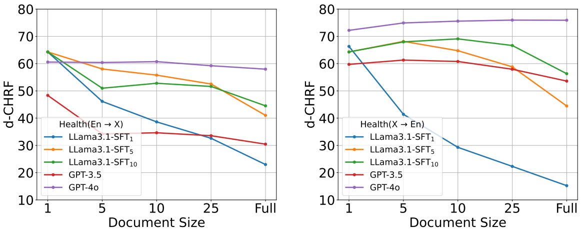

Figure 3: Comparison of Average d-chrF scores across models and pseudo-document lengths.

ally rated as the weakest model across languages, except for Swahili, where its performance is comparatively better (see Appendix D.4 for details).

5.5 Qualitative evaluation

Our qualitative analysis, based on author feedback, indicates that GPT-3.5 frequently over-generates in the pseudo-document setup by repeating words and phrases—except in the case of Swahili, where it ranks highest. However, for Yorùbá, it often uses inconsistent or partial diacritics, resulting in inaccuracies. LLaMAX3-SFT$_1$ also exhibits repetition in pseudo-document translations, likely due to a length generalization problem (Anil et al., 2022), and does so more than LLaMAX3-SFT$_{10}$. For the other four languages, LLaMAX3-SFT$_1$ with the sentence-level setup was rated higher than other models and configurations, owing to better context preservation and fewer repetitions. These observations are consistent with both d-chrF and DA scores, although d-chrF scores tend to be inflated.

6 Discussion and Analysis

To better understand model behavior, we analyze their pseudo-document ($k = 10$) translation outputs based on our findings and common issues in document-level MT with LLMs (Wu et al., 2024; Wang et al., 2024b). Additional results are provided in Appendix E.

Are the outputs generated by translation models of appropriate length? We compare model translations to the reference translations to identify empty outputs and cases of under- or over-generation. We found that all models rarely generate empty translations (refer to Appendix E), although GPT-3.5 and GPT-4o showed a slight tendency to do so for Yorùbá and Zulu, occurring in under 10% of cases. We defined under-generation as outputs <70% of reference length, and over-generation as >143%. Consistent with our qualitative findings, GPT-3.5 tends to over-generate more in African languages except for Swahili, while LLaMAX3-SFT$_1$ often under-generates as a sentence-level model. Moreover, all models over-generated by about 20% for Amharic, likely due to its unique script.

Do LLMs generate translations in the correct target languages? We evaluate whether these models understand the task by generating outputs in the target languages using language identification. Our results show that the models rarely generate outputs in the wrong language when translating to English. However, when translating to African languages, there is a higher likelihood of incorrect language translations, particularly with open models (see Figure 2).

What is the effect of document length on translation quality? We compare average d-chrF scores of selected models, including GPT-3.5/4 and LLama3.1-SFT$_k$ (k = 1, 5, 10), evaluated across pseudo-document lengths of 1, 5, 10, 25, and full length. As shown in Figure 3, d-chrF scores generally decline with increasing document length for African language translations. The reverse translation direction shows a similar trend, except for GPT-4o, which improves with length. Models trained on longer documents also generalize better to longer inputs than those trained on sentences.

7 Conclusion

We introduce AfriDoc-MT, a document-level translation dataset in the *health* and *tech* domains for five African languages. We benchmarked various models, fine-tuning selected ones. Due to context length limits, documents were translated either sentence by sentence or as pseudo-documents. Outputs were evaluated using standard MT metrics, GPT-4o as a judge, and human direct assessment. NLLB-200 was the strongest built-in MT model, while GPT-4o outperformed general-purpose LLMs. However, our DA and qualitative analysis found GPT-4o's judgments inconsistent for African languages, and sentence-by-sentence translation proved more effective for some languages.

8 Limitations

Choice of LLMs and Prompts We evaluated only a small subset of the numerous multilingual LLMs available. Our experiments were also limited by the context length of the LLMs, particularly for open LLMs. Except for LLama3.1, all other open LLMs have a context length of 8192 tokens, while encoder-decoder models were primarily based on T5. This makes it difficult to use the context length beyond a certain limit, making full document translation infeasible. Additionally, LLMs are prone to variance in performance based on the prompt. We therefore evaluated them for translation using three different prompts. However, it is possible that our prompts were not optimal.

Language Coverage Africa is home to thousands of indigenous languages, many of which exhibit unique linguistic properties. However, due to the high cost of translation using human translators and limited available funding, it is currently impossible to cover all languages. As a result, we focused on just five languages. We hope that future work will expand this dataset to include more languages and inspire the creation of additional datasets with broader coverage for document-level translation. Similarly, AFRIDOC-MT is a multi-way parallel dataset. However, due to the cost of running inference over three prompts and across all 30 translation directions for all the models evaluated, most of our analysis is limited to translation tasks between English and the five African languages. While we fine-tuned NLLB-200, LLama3.1 and LLaMAX3 on all 30 directions, we only provide results from NLLB-200 for all directions both before and after fine-tuning for sentence-level and pseudo-document tasks in the Appendix E.

Evaluation Metrics Quality evaluation in MT is an open and ongoing area of research, especially for document-level translation. Recent works have proposed embedding-based metrics for evaluation at both the sentence and document levels. While this has been well explored for high-resource language pairs, it remains underexplored for African languages, although there is a tool, AfriCOMET, that works for sentence-level evaluation in African languages. However, we evaluated the document-level translation outputs using *ModernBERT-base-long-context-qe-v1*[11], trained on the WMT human evaluation dataset across 41 language pairs, including over 20 languages and three African languages (Hausa, Xhosa, and Zulu), two of which are included to our work. However, the scores were nearly identical across all models, offering no meaningful differentiation. Hence, for our document-level evaluation, in addition to lexical-based metrics, we incorporated three other evaluation approaches: using GPT-4o as a judge, human evaluation, and qualitative analysis. GPT-4o was employed to assess and rate the translation outputs of four models. While its ratings were consistent for translations into English, the same was not observed for translations into African languages, likely due to the model's limited understanding of these languages. Therefore, we conducted a human evaluation for translations from English to African languages, comparing only three models due to cost constraints. However, we were unable to recruit annotators for Zulu.

Model Coverage and Evaluation Focus While we fine-tuned both NLLB-1.3B and LLaMAX3 models across all 30 language directions, due to computational constraints and the high cost of qualitative evaluation, our detailed analysis focuses only on translation between English and the 5 African languages. Nevertheless, we report quantitative results across all 30 directions for NLLB-1.3B. We will make all fine-tuned models publicly available to support future work, and we hope that further research will explore the remaining translation directions in greater depth.

Translationese and English-Centric Bias A potential limitation of our dataset is the influence of translationese (Koppel and Ordan, 2011). Since all source material translated originates in English, translated sentences in African languages may exhibit patterns such as unnatural syntax or overly literal phrasing. Although we have not conducted an analysis to quantify these effects, prior work suggests that they can affect MT model performance, generalization and evaluation including direct assessment (Freitag et al., 2019; Edunov et al., 2020). Furthermore, AFRIDOC-MT may reflect a bias toward English in terms of structure, semantics, and cultural framing. We leave a deeper investigation of these issues to future work.

[11] https://huggingface.co/ymoslem/ModernBERT-base-long-context-qe-v1

Ethics Statement

AFRIDOC-MT was created with the utmost consideration for ethical standards. The English texts translated were sourced from publicly available and ethically sourced materials. The data sources were selected to represent different cultural perspectives, with a focus on minimizing any potential bias. Efforts were made to ensure the dataset does not include harmful, biased, or offensive content via manual inspection.

References

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 African languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12798–12823, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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. 2022. 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, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024a. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David Ifeoluwa Adelani. 2022. Natural language processing for african languages.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024b. Irokobench: A new benchmark for african languages in the age of large language models.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

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.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part*

2) at EMNLP 2020, Online. Association for Computational Linguistics.

Cem Anil, Yuhuai Wu, Anders Johan Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Venkatesh Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. English-twi parallel corpus for machine translation. In *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Edward Bayes, Israel Abebe Azime, Jesujoba O. Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A. Etori, Shamsuddeen Hassan Muhammad, Choice Mpanza, Igneciah Pocia Thete, Dietrich Klakow, and David Ifeoluwa Adelani. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the trenches on reproducible evaluation of language models.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. *Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level*. Ph.D. thesis, Inria Paris, Sorbonne Université; Sorbonne Universite; Inria Paris.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham,

Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen,

Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024. Toucan: Many-to-many translation for 150 african language pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13189–13206, Bangkok, Thailand. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-english machine translation: an evaluation benchmark. In *1st AfricaNLP Workshop Proceedings, AfricaNLP@ICLR 2020, Virtual Conference, Formerly Addis Ababa Ethiopia, 26th April 2020*.

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. 2020. Beyond english-centric multilingual machine translation.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. Learn to remember: Transformer with recurrent memory for document-level machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Eva Martínez Garcia, Carles Creus, Cristina España-Bonet, and Lluís Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):85.

Eva Martínez Garcia, Cristina Espana-Bonet, and Lluís Màrquez Villodre. 2014. Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural*, 53:103–110.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh

Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

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.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar, and Melvin Johnson. 2022. DOCmT5: Document-level pretraining of multilingual language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 425–437, Seattle, United States. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Eva Martínez Garcia, Carles Creus, and Cristina España-Bonet. 2019. Context-aware neural machine translation decoding. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 13–23, Hong Kong, China. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

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. 2022. No language left behind: Scaling human-centered machine translation.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

OpenAI. 2024. Introducing ChatGPT. `https://openai.com/index/chatgpt/`. [Accessed 01-06-2024].

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Yirong Sun, Dawei Zhu, Yanjun Chen, Erjia Xiao, Xinghao Chen, and Xiaoyu Shen. 2025. Fine-grained and multi-dimensional metrics for document-level machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 1–17, Albuquerque, USA. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Sami Ul Haq, Sadaf Abdul Rauf, Arsalan Shaukat, and Abdullah Saeed. 2020. Document level NMT of low-resource languages with backtranslation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 442–446, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024a. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024b. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7175–7187, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

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

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation.

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.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-MT:

Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024a. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024b. A preference-driven paradigm for enhanced translation with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3385–3403, Mexico City, Mexico. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024c. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A More about the languages and their characteristics

We selected at least one language from each subregion of Sub-Saharan Africa: Hausa (North-West Africa), Yorùbá (West Africa), Amharic (East Africa), Swahili (East Africa), and Zulu (Southern Africa). Each of these languages has over 20 million speakers. All of them use the Latin script except for Amharic, which uses the Ge'ez script. The Latin-script languages use the Latin alphabet with the omission of some letters and the addition of new ones, and the use of diacritics (e.g., Yorùbá). The languages are tonal, except for Amharic and Swahili. Just like English, all languages follow the subject-verb-object word order. Refer to Adelani (2022) for a comprehensive overview of the characteristics of these languages.

B More details about AfriDoc-MT

Table 10 shows the average number of white-space-separated tokens for sentences across various domains and their corresponding translations in all the languages including English. The *health* domain has more tokens on average than *tech*. Hausa and Yorùbá have more tokens on average than English, possibly because they are descriptive languages, while Swahili has a comparably similar length to English. However, Amharic and Zulu have relatively shorter average lengths, demonstrating interesting linguistic properties.

B.1 Translation guidelines

The translation guidelines, aside the details shared at the workshop on translation and terminology creation can be found below.

- Thank you for agreeing to work on this project. Below is the link to access the data for translation. The files are in .csv format, and you can open them using Google Sheets or Microsoft Excel (for offline work).
- Each file contains 2500 sentences, and they are named in the format of a serial number followed by your first name.
- Please do not delete double empty rows, as they serve to separate paragraphs. Also, avoid deleting any rows, columns, or provided text.
- Use the language field to input the translations. It is essential not to rely on translation engines, as our quality assurance process can detect this. Depending on such tools may result in potential issues that you would need to address, leading to additional work on your part.
- We will provide a list of extracted terminologies soon so that you can harmonize how terminologies are translated.
- Thank you for your attention to these guidelines. Should you have any questions, concerns, or suggestions, feel free to contact us or reach out to your language coordinator.

B.2 Quality evaluation of the translations

As part of the human translation process, we conducted quality estimation to assess the transla-

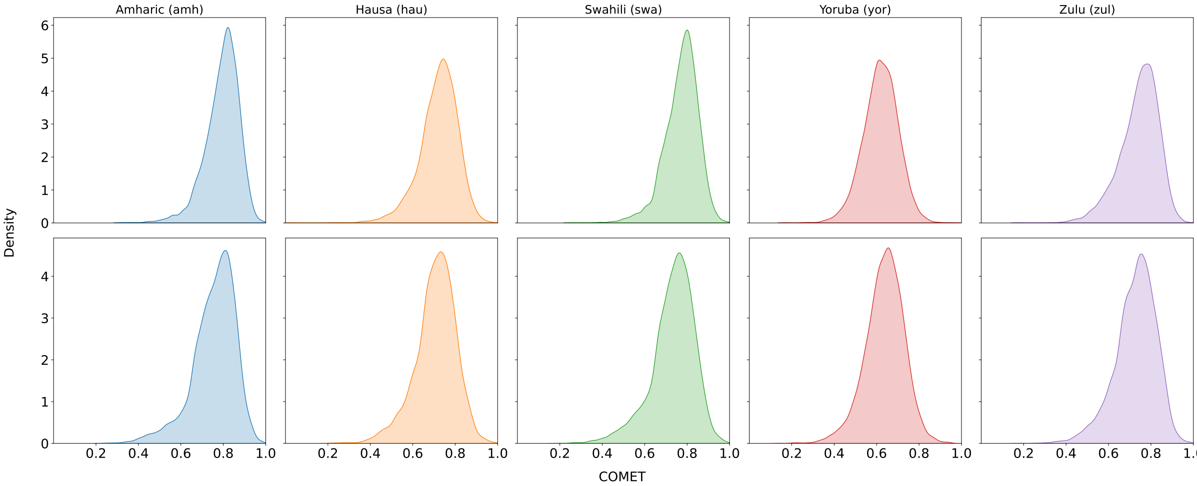

Figure 4: Distribution of the quality estimation of of the translated sentences using COMET scores for the *health* (top), *tech* (bottom).

Domain	**eng**	**amh**	**hau**	**swa**	**yor**	**zul**
Sentence						
health	21.6	19.3	28.1	23.2	27.9	16.7
tech	17.8	15.6	22.2	18.0	23.7	13.4
Document						
health	647.3	576.7	841.7	695.4	834.8	500.1
tech	658.2	575.0	821.6	665.4	873.4	495.9

Table 10: The average number of tokens in AFRIDOC-MT, both at sentence and document level.

tions. For this purpose, we used AfriCOMET.[12] Given a translated sentence in any African language and its corresponding source English sentence, AfriCOMET generates a score between 0 and 1, where 0 indicates poor quality and higher values signify better quality. The translators, in collaboration with the language coordinators, were tasked with reviewing instances that had quality estimation scores below 0.65. This step was essential to identify and correct low-quality translations.

Figure 4 illustrate the distribution of the final quality scores for the five languages and both domains. Our manuel check indicates that QE scores below 0.65 are not necessarily indicative of poor translations, which is consistent with the findings of Adelani et al. (2024b). We attribute this observation to factors such as domain shift, translation length, and other potential influences, which warrant further investigation in future research.

[12] https://huggingface.co/masakhane/africomet-qe-stl

B.3 Creation of pseudo-documents for AFRIDOC-MT

Given that the translated documents vary in length in terms of sentences and tokens, and considering the maximum token length limitations of the different LLMs used, we adopted a chunking approach for document-level evaluation. In this approach, documents were divided into smaller pseudo-documents that fit within the maximum length constraints of the models. To establish an appropriate chunk size, each document was divided into fixed-size chunks of k sentences, with the possibility that the final chunk may contain fewer than k sentences. These sentence groups, referred to as pseudo-documents, were used for document-level translation.

We conducted an initial analysis, testing different values for k (5, 10, and 25), with k=1 serving as our sentence-level setup. Table 11 presents the resulting number of parallel pseudo-documents, as well as the average number of tokens per pseudo-document per language for the various model tokenizers, including the 95th percentile token count. Our analysis revealed that Amharic and Yorùbá —languages with unique characteristics such as non-Latin scripts and diacritics, respectively—had the largest average token counts across the tokenizers. Additionally, the domain with the highest number of average tokens for pseudo-document varies from language to language.

To accommodate both languages in our experiments, we chose pseudo-documents with k=10. However, for the SFT models described in Appendix C.2, we used both k=5 and k=10.

Languages/Split	Models	Full		25 sent.		10 sent.		5 sent.	
		Health	Tech	Health	Tech	Health	Tech	Health	Tech
Sizes of data splits in AFRIDOC-MT pseudo-document									
Train		240	187	402	369	812	789	1506	1483
Dev		33	25	56	48	112	106	209	204
Test		61	59	108	106	224	227	417	418
Statistics of LLM tokens in AFRIDOC-MT pseudo-document training splits									
en	NLLB-200	923.7/2017.6	941.9/1982.1	551.5/951.7	477.4/758.8	273.0/430.9	223.2/343.6	147.2/233.8	118.8/184.9
	MADLAD-400	971.0/2095.2	991.4/2100.1	579.7/1017.1	502.4/797.8	287.0/449.3	235.0/362.0	154.7/245.0	125.0/196.9
	Aya-101	1008.2/2183.5	1020.5/2184.3	601.9/1038.0	517.2/820.2	298.0/463.4	241.9/372.6	160.7/255.0	128.7/199.0
	LLaMA3	801.4/1788.0	842.5/1798.4	478.5/833.8	427.0/664.0	236.9/372.9	199.7/304.2	127.8/203.0	106.3/166.0
	Gemma-2	802.9/1820.1	857.9/1857.6	479.3/841.0	434.8/689.6	237.3/375.0	203.4/314.0	128.0/205.0	108.2/169.0
	ModernBERT	801.0/1800.7	862.7/1819.3	478.3/837.8	437.3/685.6	236.8/373.4	204.5/311.0	127.8/204.0	108.9/171.0
am	NLLB-200	1304.4/2785.8	1376.3/2888.7	778.8/1329.9	697.5/1130.8	385.6/592.0	326.2/520.0	207.9/328.0	173.5/282.9
	MADLAD-400	1624.8/3393.6	1685.0/3487.4	970.0/1684.2	853.9/1380.4	480.2/750.0	399.4/640.2	258.9/413.8	212.5/342.9
	Aya-101	1887.4/3937.9	1934.7/4126.9	1126.8/1931.8	980.5/1598.0	557.9/855.4	458.5/722.0	300.8/477.8	244.0/390.0
	LLaMA3	6798.0/13986.2	6829.6/14750.9	4058.5/6971.8	3461.1/5584.8	2009.3/3084.4	1618.7/2560.8	1083.3/1716.0	861.2/1379.9
	Gemma-2	2817.9/5857.5	2868.4/6227.4	1682.1/2896.4	1453.2/2342.4	832.4/1267.8	679.3/1071.6	448.5/710.0	361.0/575.0
	ModernBERT	7347.8/15045.1	7386.4/15952.3	4386.4/7544.1	3742.8/6035.8	2171.1/3331.2	1749.9/2760.4	1170.2/1851.0	930.6/1485.9
ha	NLLB-200	1204.4/2713.7	1171.4/2463.0	719.0/1252.8	593.6/962.6	356.0/554.0	277.6/430.6	191.9/306.8	147.7/232.0
	MADLAD-400	1297.1/2849.4	1260.5/2643.7	774.4/1359.7	638.8/1042.0	383.4/606.4	298.8/465.6	206.7/329.0	158.9/251.0
	Aya-101	1614.9/3497.4	1535.3/3241.9	964.1/1672.3	778.0/1254.6	477.3/742.6	363.9/563.2	257.4/410.8	193.6/306.0
	LLaMA3	1916.7/4012.9	1822.6/3917.9	1144.3/1988.8	923.7/1513.6	566.6/882.4	432.1/674.6	305.5/488.8	230.0/365.9
	Gemma-2	1642.4/3568.9	1581.3/3373.4	980.6/1716.7	801.4/1297.8	485.5/757.4	374.8/584.0	261.8/417.8	199.4/317.8
	ModernBERT	1998.5/4207.5	1916.8/4139.7	1193.1/2057.8	971.5/1575.8	590.8/916.9	454.4/701.0	318.6/510.8	241.8/382.9
sw	NLLB-200	1100.8/2494.8	1094.8/2187.5	657.2/1145.9	554.8/896.4	325.4/517.0	259.5/409.6	175.4/280.0	138.1/218.0
	MADLAD-400	1177.3/2629.9	1155.3/2293.9	702.8/1227.6	585.5/938.6	348.0/547.0	273.8/436.0	187.6/297.0	145.7/231.9
	Aya-101	1345.3/2925.0	1311.0/2667.8	803.2/1390.9	664.4/1076.2	397.6/627.9	310.7/487.4	214.4/339.0	165.3/261.0
	LLaMA3	1668.1/3605.0	1619.4/3364.9	995.9/1735.4	820.7/1330.0	493.1/771.4	383.9/599.8	266.0/418.0	204.3/323.0
	Gemma-2	1413.3/3097.3	1377.1/2770.0	843.8/1467.7	697.9/1126.2	417.8/658.9	326.4/513.0	225.3/356.8	173.7/277.9
	ModernBERT	1757.9/3753.4	1719.7/3594.1	1049.5/1822.8	871.6/1421.0	519.7/810.0	407.7/632.0	280.2/441.0	217.0/342.8
yo	NLLB-200	1702.6/3854.7	1724.8/3577.1	1016.5/1857.2	874.1/1428.6	503.2/814.7	408.8/644.6	271.3/443.8	217.5/348.9
	MADLAD-400	1983.6/4470.9	1990.4/4136.7	1184.3/2137.5	1008.7/1650.2	586.3/939.4	471.7/742.2	316.1/512.0	251.0/401.9
	Aya-101	2729.2/5832.3	2659.8/5549.7	1629.4/2956.4	1347.9/2211.6	806.7/1292.4	630.4/988.0	434.9/704.0	335.4/544.0
	LLaMA3	2945.8/6322.4	2880.0/5995.5	1758.6/3203.9	1459.4/2400.4	870.5/1406.0	682.5/1077.6	469.3/767.8	363.0/585.9
	Gemma-2	2620.4/5745.5	2593.5/5406.9	1564.3/2867.7	1314.3/2143.8	774.4/1245.4	614.6/965.6	417.4/678.0	327.0/530.0
	ModernBERT	3648.3/7780.9	3595.2/7600.6	2178.1/4002.0	1822.0/3020.4	1078.3/1761.4	852.1/1339.8	581.4/957.2	453.3/733.9
zu	NLLB-200	1201.8/2513.3	1230.4/2555.7	717.5/1233.0	623.5/1016.6	355.2/554.3	291.6/461.2	191.5/300.0	155.1/250.0
	MADLAD-400	1215.2/2524.0	1230.7/2519.6	725.5/1284.8	623.7/1007.2	359.2/557.8	291.7/465.6	193.7/305.5	155.2/251.0
	Aya-101	1491.3/3012.2	1485.2/3180.8	890.3/1521.8	752.7/1213.0	440.8/688.9	352.0/554.4	237.7/372.8	187.3/298.9
	LLaMA3	1921.7/3822.6	1834.3/3933.4	1147.3/1963.9	929.7/1512.4	568.1/885.4	434.9/689.2	306.4/475.8	231.5/373.0
	Gemma-2	1787.5/3573.5	1703.0/3666.1	1067.2/1834.8	863.0/1416.2	528.3/819.4	403.6/637.6	284.9/447.8	214.8/343.9
	MordernBERT	2073.7/4134.2	1965.8/4239.3	1238.1/2138.4	996.3/1625.6	613.0/956.3	466.1/737.0	330.6/515.8	248.0/399.0

Table 11: AFRIDOC-MT Pseudo-document statistics. The number of translation instances in the data AFRIDOC-MT pseudo-document splits. average and 95th percentile (average/95 percentile) of the AFRIDOC-MT document train split tokenization statistics using the different LLM tokenizers.

C Experimental details

C.1 Evaluated Models

C.1.1 Translation Models

M2M-100 (Fan et al., 2020) is a transformer-based multilingual neural translation model from Meta, trained to translate between 100 languages, including several African languages. It has three variants of different sizes: 400M parameters, 1.2B parameters, and 12B parameters. For our experiments, we evaluated the 400M and 1.2B variants.

NLLB (NLLB Team et al., 2024) is a model similar to M2M-100, with broader coverage, trained to translate between just over 200 languages, including more than 50 African languages. It also has different sizes: 600M, 1.3B, 3.3B, and 54B parameters. For this work, we evaluated the first three variants.

MADLAD-400 (Kudugunta et al., 2023) is a multilingual translation model based on the T5 architecture (Raffel et al., 2020), covering 450 languages, including many African languages. It was trained on data collected from the Common Crawl dataset. The dataset underwent a thorough self-audit to filter out noisy content and ensure its quality for training MT models.

Toucan (Elmadany et al., 2024; Adebara et al., 2024) is another multilingual but Afro-centric translation model based on the T5 architecture, covering 150 language pairs of African languages. It was first pre-trained on large multilingual texts covering over 500 African languages and then finetuned on translation task covering over 100 language pairs.

C.1.2 Large Language Models

Aya-101 (Üstün et al., 2024) is an instruction-tuned mT5 model (Xue et al., 2021) designed to handle both discriminative and generative multilingual tasks. With 13B parameters, it covers 100 languages and is capable of translating between a wide range of languages, including African languages.

Gemma2 (Gemma Team et al., 2024) is a decoder-only LLM trained on billions of tokens sourced from the web. The training data primarily consists of English-language text, but it also includes code and mathematical content. While Gemma2 has an English-centric focus, it also possesses multilingual capabilities. We evaluate the base Gemma2 model with 9B parameters, as well as its instruction-tuned version.

LLama3.1 (Dubey et al., 2024) is another decoder-only LLM trained on trillions of tokens across multiple languages. It was fine-tuned using existing instruction datasets as well as synthetically generated instruction data to create its instruction-tuned version. One advantage LLama3.1 has over other models is its context window of 128K tokens, the largest among all models considered in this work, making it particularly suitable for document-based tasks such as document-level translation. We evaluate the base LLama3.1 model with 8B parameters, as well as its instruction-tuned version.

LLaMAX3 (Lu et al., 2024) is a multilingual LLM built on the LLama3 with 8B parameters as its base. It was trained on 102 languages, including several African languages, through continued pretraining. Using an English instruction dataset (Alpaca), it was further fine-tuned to create LLaMAX3-Alpaca. We evaluated both models and compared their performance across various tasks.

C.2 Supervised Finetuning

We perform supervised fine-tuning to tailor LLMs for translation tasks. To train sentence-level MT systems, we use all parallel sentences from AFRIDOC-MT to construct the training set, enabling the LLMs to translate across multiple directions and domains. Following Zhu et al. (2024b), we augment the parallel data with translation instructions, which are randomly sampled from a predefined set of 31 MT instructions for each training example.[13] To train document-level MT systems, we follow the same process, but train on longer segments formed by concatenating multiple sentences. When fine-tuning, we use a learning rate of $5e^{-6}$ and an effective batch size of 64. Models are trained for only one epoch, as further training does not result in improvements and may even lead to performance degradation.

Similarly, we fine-tuned the 1.3B version of NLLB-200 for sentence and pseudo-document (with 10 sentences) translation using the Fairseq (Ott et al., 2019) codebase. We used all the training examples from 30 language directions across both domains. The model was fine-tuned for 50k steps using a learning rate of $5e^{-5}$, token batch size of 2048 and a gradient accumulation of 2. The checkpoint with the lowest validation loss was selected as the best model for evaluation.

[13] We use the same instruction set as described in (Zhu et al., 2024b).

Setting	X → eng	eng → X
Sentence		
sentence	512	512
Document		
5	4096	4096
10	4096	4096
25	1024	8192 (11264)
Full	2048	16384 (32768)

Table 12: The maximum number of tokens set for decoder-only LLMs when translating between English and African languages, and vice versa. Special cases for Amharic are indicated in brackets.

C.3 Evaluation setup

The models were evaluated using different tools. For example, both the NLLB-200 and M2M-100 models were evaluated with the Fairseq codebase, while Toucan and MADLAD-400 were evaluated using the Hugging Face (HF) codebase. All other LLMs, including LLama3.1 (both instruction-tuned and SFT models), Gemma, and Aya-101, were evaluated using EleutherAI LM Evaluation Harness (`lm-eval`) tool (Biderman et al., 2024). In all cases, greedy decoding was used.

The models evaluated have different context lengths. For encoder-decoder models, M2M-100 and NLLB have a maximum sequence length of 1024 and 512 respectively. Aya-101 and MADALAD, based on the T5 architecture, do not have a pre-specified maximum sequence length, so we fixed their maximum sequence length to 1024 for all experiments involving encoder-decoder models. However, for decoder-only models, Gemma and LLaMAX3 (based on LLama3) have a maximum sequence length of 8192, while LLama3.1 has a maximum sequence length of 128K. Since all the decoder-only models were evaluated using LM Evaluation Harness, we used a similar setup for them, selecting the maximum length based on the specific needs of each model.

Table 12 shows the maximum number of generation tokens we set when translating between English and African languages. These numbers were chosen based on the statistics from Table 11. However, for Amharic, when translating pseudo-documents with 25 sentences and full documents, there were instances exceeding the 95th percentile derived from the training statistics. Therefore, we increased the token limit specifically for Amharic.

C.4 Evaluation prompts

While the translation models we evaluated do not require prompts, MADLAD-400, requires a prefix of the form **<2xx>** token, which is prepended to the source sentence. Here, xx indicates the target language using its language code (e.g., "sw" for Swahili). Similarly, Toucan uses just the target language ISO-693 code as prefix, which is prepended to the source sentence (e.g., "swa" for Swahili). For other models, including Aya-101, we used three different prompts for sentence-level translation and document translation experiments. The main difference between the prompts for these tasks is the explicit mention of "text" or "document" within the prompt, as shown in Table 23. For the base models Gemma2, Llama3.1, LLaMAX3, and Aya-101, we prompted them directly using the respective prompts. However, for the instruction-tuned versions of Gemma2 and Llama3.1, we used their respective chat templates. For all Alpaca-based models, including our SFT models, we used the Alpaca template.

C.5 Evaluation metrics

We evaluate translation quality with BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) using SacreBLEU[14] (Post, 2018). We run significance tests using bootstrap resampling and report the 95% confidence interval for the scores, based on a sample size of 1000. We also use AfriCOMET[15] (Wang et al., 2024a) to evaluate the quality of the translation outputs. We report the chrF scores of the best prompt for each model and language direction in the main paper, with all additional results provided in the Appendix D. For document-level experiments, we evaluated the LLMs using the same three prompts as in the sentence-level experiment. For evaluation, we used BLEU and chrF scores but excluded AfriCOMET due to its backbone model, AfroXLM-R-L (Alabi et al., 2022; Adelani et al., 2024a), having a context length of 512 tokens. This made it impractical to compute COMET scores for document-level outputs.

[14] case:mixed|eff:no| tok:13a|smooth:exp|v:2.3.1,

[15] https://huggingface.co/masakhane/africomet-stl-1.1

C.6 GPT-4o as an evaluator for machine translation

We use GPT-4o to assess the quality of translation output, as demonstrated by Sun et al. (2025), which shows a correlation with human judgment. Due to the cost of this task, we limited our evaluation to a few selected models, including Aya-101, GPT-3.5, GPT-4o, and LLaMAX3 fine-tuned on AfriDoc-MT sentences and pseudo-documents of 10 sentences. We compared translations performed at the sentence level and pseudo-document level in terms of fluency, content errors, and cohesion errors—specifically lexical (LE) and grammatical (GE) errors—using the same definitions as Sun et al. (2025).

Below are the prompts used to evaluate documents using GPT-4o for fluency, content errors, and cohesion errors—specifically lexical (LE) and grammatical (GE) errors.

- **Fluency**: GPT-4o is prompted to rate the fluency of a document on a scale from 1 to 5, where 5 indicates high fluency and 1 represents low fluency. This evaluation is conducted without providing any reference document. For the final fluency score, we report the average rating across all documents. Below we provide the prompt used.

```
Please evaluate the fluency of the
    following text in <<target>>.

------

### **Instructions:**

- **Task**: Evaluate the fluency of
    the text.

- Scoring: Provide a score from 1 to
    5, where:

  - **5**: The text is **highly
      fluent**, with no grammatical
      errors, unnatural wording, or
      stiff syntax.
  - **4**: The text is **mostly
      fluent**, with minor errors
      that do not impede
      understanding.
  - **3**: The text is **moderately
      fluent**, with noticeable
      errors that may slightly
      affect comprehension.
  - **2**: The text has **low
      fluency**, with frequent
      errors that hinder
      understanding.
  - **1**: The text is **not fluent
      **, with severe errors that
```

Model	Setup	*eng* → *X*					*X* → *eng*				
		d-CHRF↑	**Fluency↑**	**CE↓**	**LE↓**	**GE↓**	**d-CHRF↑**	**Fluency↑**	**CE↓**	**LE↓**	**GE↓**
Aya-101	Sent	$45.8_{12.6}$	$2.5_{0.8}$	$10.1_{1.3}$	$4.3_{0.8}$	$3.4_{0.4}$	$64.9_{4.0}$	$3.1_{0.2}$	$19.0_{2.5}$	$12.5_{1.3}$	$5.7_{0.5}$
	Doc10	$42.8_{15.5}$	$3.0_{0.6}$	$9.4_{1.8}$	$2.8_{0.5}$	$2.1_{0.2}$	$64.2_{5.4}$	$3.4_{0.2}$	$14.4_{1.0}$	$9.4_{1.4}$	$4.4_{0.3}$
GPT-3.5	Sent	$44.2_{21.4}$	$2.3_{1.4}$	$11.8_{6.6}$	$5.5_{3.0}$	$4.1_{2.3}$	$57.9_{8.9}$	$2.8_{0.4}$	$13.1_{3.2}$	$7.6_{2.8}$	$4.7_{1.7}$
	Doc10	$30.2_{5.5}$	$2.4_{0.4}$	$7.5_{0.3}$	$2.7_{0.4}$	$2.1_{0.3}$	$28.5_{3.0}$	$3.0_{0.2}$	$8.4_{0.3}$	$2.9_{0.3}$	$2.1_{0.2}$
LLaMAX3-SFT$_1$	Sent	$58.8_{10.1}$	$3.6_{0.4}$	$11.1_{1.2}$	$4.2_{0.9}$	$3.0_{0.7}$	$61.6_{6.1}$	$3.3_{0.4}$	$11.5_{1.8}$	$5.8_{1.3}$	$3.2_{0.2}$
	Doc10	$31.8_{2.8}$	$2.6_{0.5}$	$8.9_{0.6}$	$2.9_{0.6}$	$2.2_{0.3}$	$28.4_{2.1}$	$3.0_{0.3}$	$8.8_{0.2}$	$3.2_{0.2}$	$2.0_{0.1}$
LLaMAX3-SFT$_{10}$	Doc10	$55.0_{11.8}$	$3.8_{0.6}$	$10.0_{1.4}$	$2.7_{0.9}$	$1.9_{0.5}$	$69.3_{3.1}$	$4.3_{0.2}$	$9.5_{1.0}$	$5.2_{0.8}$	$2.5_{0.5}$

Table 13: Document-level evaluation in the *tech* domain, judged by GPT-4o. Compares sentence- vs. document-level outputs on Fluency (1–5 scale), Content Errors (CE), Lexical (LE), and Grammatical Cohesion Errors (GE).

~~~~
      make it difficult to
      understand.
- **Explanation**: Support your
    score with specific examples to
    justify your evaluation.

------

### **Output Format:**

Provide your evaluation in the
    following JSON format:

```
{
  "Fluency": {
    "Score": "<the score>",
    "Explanation": "<your
        explanation on how you made
        the decision>"
  }
}
```

------

**Text to Evaluate:**

<<hypothesis>>

Answer:
~~~~

- **Accuracy**: GPT-4 is prompted to identify and list the mistakes, such as incorrect translations, omissions, additions, and any other errors, by comparing the model's output to the reference translation. After identifying these errors, we count all of them and compute the average across all documents, reporting that as the content error (CE). Below is the prompt used.

~~~~
Please evaluate the accuracy of the
    following translated text in <<
    target>> by comparing it to the
    provided reference text.

------

### **Instructions:**

- **Task**: Compare the text to the
    reference text.

- Identify Mistakes: List all
    mistakes related to accuracy.

  - Mistake Types:

    - **Wrong Translation**:
        Incorrect meaning or
        misinterpretation leading to
         wrong information.
    - **Omission**: Missing words,
        phrases, or information
        present in the reference
        text.
    - **Addition**: Extra words,
        phrases, or information not
        present in the reference
        text.
    - **Others**: Mistakes that are
        hard to define or categorize
        .

- **Note**: If the text expresses
    the same information as the
    reference text but uses
    different words or phrasing, it
    is **not** considered a mistake.

- **Provide a List**: Summarize all
    mistakes without repeating the
    exact sentences. Provide an
    empty list if there are no
    mistakes.

------

### **Output Format:**

Provide your evaluation in the
    following JSON format:

```
{
  "Accuracy": {
    "Mistakes": [
      "<list of all mistakes in the
          text with format'Mistake
          Types: summarize the
          mistake', provide an empty
           list if there are no
          mistakes>"
    ]
  }
}
```

------
~~~~

```
**Reference Text:**

<<reference>>

**Text to Evaluate:**

<<hypothesis>>
```

- **Cohesion**: GPT-4 is prompted to rate cohesion-related mistakes, including lexical and grammatical errors, in the model's output, comparing it to the reference translation. We count each error individually, compute the average across the documents, and report them as lexical errors (LE) and grammatical rrrors (GE). Below is the prompt template we used.

```
Please evaluate the cohesion of the
    following translated text in
    <<target>> by comparing it to
    the provided reference text.

------

### **Instructions:**

- **Task**: Evaluate the cohesion of
    the text.

- **Definition**: Cohesion refers to
    how different parts of a text
   are connected using language
   structures like grammar and
   vocabulary. It ensures that
   sentences flow smoothly and the
   text makes sense as a whole.

- Identify Mistakes: List all
   mistakes related to cohesion.

  - Separate the mistakes into:

    - **Lexical Cohesion Mistakes**:
       Issues with vocabulary
      usage, incorrect or missing
      synonyms, or overuse of
      certain words that disrupt
      the flow.
    - **Grammatical Cohesion
      Mistakes**: Problems with
      pronouns, conjunctions, or
      grammatical structures that
      link sentences and clauses.

- **Provide Lists**: Provide
   separate lists for lexical
   cohesion mistakes and
   grammatical cohesion mistakes.
   Provide empty lists if there are
    no mistakes.

------

### **Output Format:**

Provide your evaluation in the
   following JSON format:

'''
{
  "Cohesion": {
    "Lexical Cohesion Mistakes": [
      "<list of all mistakes in the
          text one by one, provide
          an empty list if there are
           no mistakes>"
    ],
    "Grammatical Cohesion Mistakes":
         [
      "<list of all mistakes in the
          text one by one, provide
          an empty list if there are
           no mistakes>"
    ]
  }
}
'''

------

**Reference Text:**

<<reference>>

**Text to Evaluate:**

<<hypothesis>>
```

Fluency can only have values between 1 and 5. However, the other metrics, including CE, GE, and LE, do not have a specific range and can take on any value because they are counts. Refer to (Sun et al., 2025) for more details about these metrics.

C.7 Human Evaluation Setup

Beyond using GPT-4o as a judge, we also conduct human evaluation on a subset of outputs from GPT-3.5, LLaMAX3-SFT$_1$, and LLaMAX3-SFT$_{10}$ for two domains, focusing specifically on translation into five African languages due to cost constraints. Translation into English was excluded, as existing automatic metrics, including GPT-based evaluations, are already reliable for this direction.

For the human evaluation, three native speakers of the African languages—primarily translators involved in the dataset creation—were recruited. Each annotator was assigned 80 documents to evaluate, tasked with marking as many error spans as possible and rating the overall quality on a scale from 0 to 100. This annotation followed the error span annotation (ESA) (Kocmi et al., 2024) protocol as implemented within the Appraise Evaluation Framework (Federmann, 2018). To assess consistency and inter-annotator agreement, 30 of the 80 documents were shared among all three annotators. Table 14 shows statistics for 80 documents

Model	Setup	Full		Shared	
		health	tech	health	tech
GPT-3.5	Sent.	5	5	-	5
	Pseudo.	5	5	-	5
LLaMAX3-SFT$_1$	Sent.	5	5	5	-
	Pseudo.	5	5	5	-
LLaMAX3-SFT$_{10}$	Pseudo	5	5	5	5
Total		25	25	15	15

Table 14: The number of documents annotated by each annotator for human direct assessment.

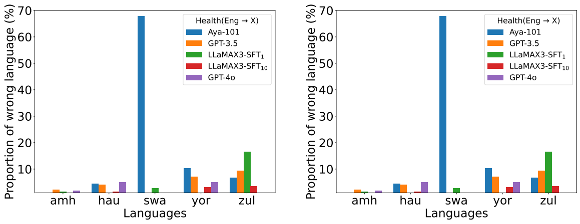

Figure 5: Rate of off-target translation (k =10).

sampled from the models in both domains for each annotator. Each annotator was remunerated with $55[16].

C.8 Qualitative Analysis

Alongside the human direct assessment of the translation outputs, we shared a subset of the outputs with one author per language, each a native speaker. They were tasked with analyzing the outputs to answer two key questions: (1) What common errors or flaws do the models exhibit across different setups? and (2) How fluent are the translation outputs produced by the models across the various settings?

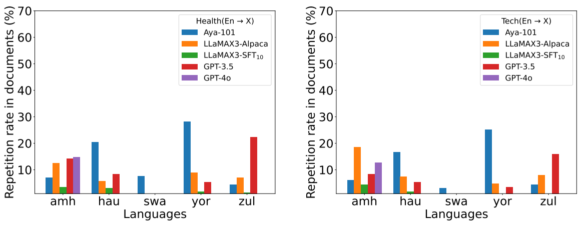

Figure 6: Word repetition rate in the pseudo-document translation (k =10).

D More experimental results

D.1 Sentence-level evaluation

Given that AFRIDOC-MT is a document-level translation dataset, and due to the limited context length of most translation models and LLMs, which makes it impossible to translate a full document at once, we opted to translate the sentences within the documents and then merge them back to form the complete document. This also serves as a baseline for document-level translation. In the main paper, we present the results for the best prompt for each language pair and model using d-chrF. In this section, we also provide the full results on the merged documents using d-chrF and d-BLEU in Tables 16 and 17. Furthermore, we present results for evaluating just the sentences (without merging them back into documents) using s-BLEU, s-chrF, and s-COMET in Tables 18 and 19. In Figures 18 to 21, we provide plots that summarize some of the results in the table for a few models. Although the main findings are summarized in the main draft, below are some other points we identify.

[16] Annotation protocol.

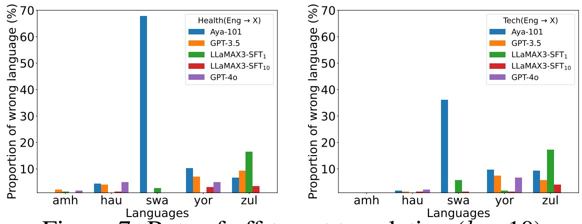

Figure 7: Rate of off-target translation (k =10)

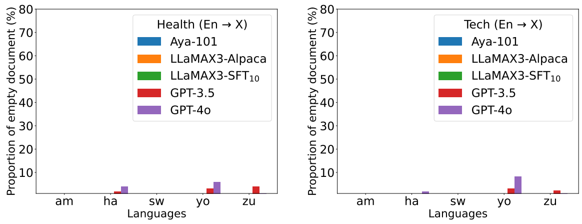

Figure 8: Proportion of empty outputs for pseudo-documents.

M2M-100 is not competitive Neither version of M2M-100, which was once a state-of-the-art translation model, is competitive with other translation models such as Toucan, NLLB-200, and MADLAD-400, even when compared to models of similar sizes, across all metrics at both the sentence and document levels.

Base LLMs are not translators for African languages. Base LLMs without instruction tuning and supervised fine-tuning, such as Gemma2 and LLaMAX3, do not show competitive translation performance either. This can be explained by the fact that they are just language models with limited coverage of African languages. However, LLaMAX3, which was trained on more than 100 languages, including African languages, through continued pre-training, shows improved performance, surpassing LLama3.1-IT.

Amharic and Yorùbá are the worst performing language directions. When translating from English into African languages, our results show that both Amharic and Yoruba perform the least

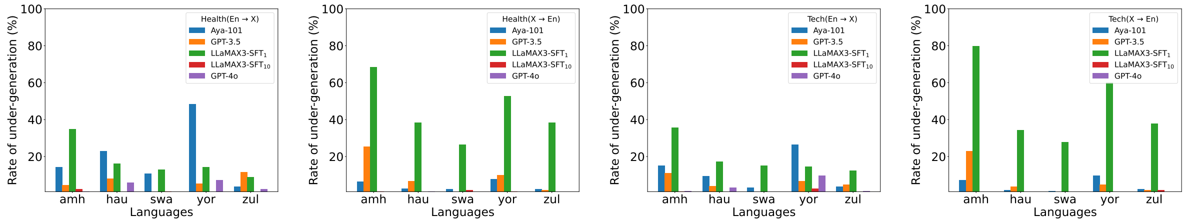

Figure 9: Rate of under-generation in pseudo-document translation (k =10)

effectively. This may be attributed to specific properties of these languages, such as the use of non-Latin script in Amharic and the use of diacritics in Yoruba, which in turn increase the tokenization rate of these languages by the different model tokenizers.

D.2 Document-level evaluation

For document-level evaluation, we split the documents into chunks of 10 sentences and translate these chunks using the different models. In Tables 20 and 21 we provide the full results on the merged pseudo-documents using d-chrF and d-BLEU. And below are some other relevant points from the results. It is important to note that we also trained and evaluated NLLB-200 for pseudo-document translation. However, due to its 512-token maximum sequence length, it is not competitive. Nevertheless, the results still show the influence of fine-tuning. Below are other findings.

Gemma2-IT shows better translation performance. Compared to the sentence-level setup, where Gemma2-IT and LLaMAX3-Alpaca achieved similar performance on average, in the pseudo-document setup, Gemma2-IT not only outperforms LLaMAX3-Alpaca but also surpasses GPT-3.5. Although we cannot provide an exact explanation for this performance, we hypothesize that its pre-training setup might be a contributing factor.

Fine-tuning data has an impact on translation quality. Our results show that both LLama3.1 and LLaMAX3 models, when fine-tuned on sentences, performed significantly worse on pseudo-document evaluations compared to the same models fine-tuned on pseudo-documents for both domains. All these models were trained using a similar setup, with the primary difference being the data used for fine-tuning.

Language-specific performance trends Overall, no clear trend is observed in MT performance across language family classes. However, Amharic (a non-Latin script language) and Yorùbá (a heavily diacriticitized language) result in the lowest chrF scores, while Swahili—the most widely spoken indigenous African language—performs best.

Model	Setup	amh	hau	swh	yor
GPT-3.5	Sent	18.3	42.1	63.8	12.6
	Doc10	4.8	32.3	**64.2**	6.9
LLaMAX3-SFT$_1$	Sent	**58.1**	**86.0**	61.0	**66.1**
	Doc10	19.0	54.5	36.8	40.1
LLaMAX3-SFT$_{10}$	Doc10	54.3	83.7	61.8	62.3

Table 15: Average DA score (scale 0–100) from three human evaluators per language in the *tech* domain.

D.3 Findings from GPT-4o as a judge

In Tables 8 and 13 we present the average GPT-4o evaluation results for four models. When translating into African languages, there is no clear pattern: for example, GPT-3.5, despite having the lowest fluency score, also had the fewest content, lexical, and grammatical errors, which is counterintuitive. In contrast, when translating into English, the pattern is clear and consistent: translations of pseudo-documents show better fluency and fewer errors overall. These results suggest that using GPT-4o as a translation judge is not yet well-suited for low-resource languages.

D.4 Findings from human evaluation

We were able to obtain DA scores from three annotators for all languages except Zulu. For each language, we calculated inter-annotator agreement using Krippendorff's alpha α over 30 document instances. We obtained α scores of 0.46, 0.57, 0.48, and 0.81 for Amharic, Hausa, Swahili, and Yorùbá, respectively. These are relatively low scores, except for Yorùbá. We present the average DA scores in Tables 9 and 15 for the *health* and *tech* domains, respectively. The results show that annotators rate documents translated at the sentence level as higher quality than those translated at the pseudo-document level. Additionally, GPT-3.5 received the lowest ratings among the three models. LLaMAX3-SFT$_1$, a model trained on sentence-level data, was rated the best across all languages when evaluated

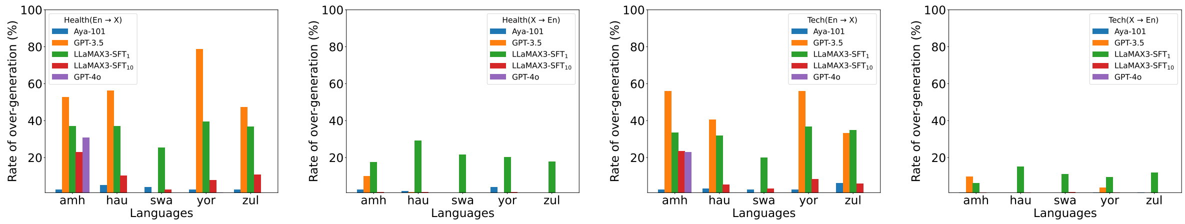

Figure 10: Rate of over-generation in pseudo-document translation (k =10)

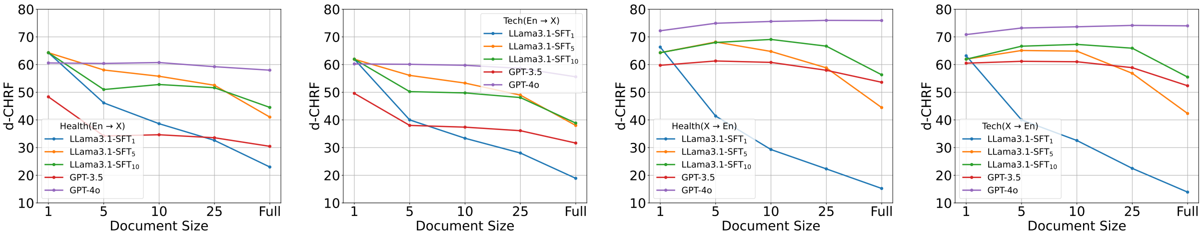

Figure 11: Average chrF score across languages for documents of different sizes.

on sentences. However, when evaluated on pseudo-documents, its performance was rated lower than that of LLaMAX3-SFT$_{10}$. These findings are consistent with the d-chrF scores for the models, but they do not align with the evaluations from GPT-4o as a judge.

E More discussion and analysis

Do LLMs generate translations in the correct target languages? We evaluate whether these models understand the task by generating outputs in the target languages using the OpenLID (Burchell et al., 2023) language identification model. Our results show that these models rarely generate outputs in the wrong language when translating to English. However, when translating to African languages, there is a higher likelihood of incorrect language translations, particularly with open models (Figure 7). These off-target languages often include English, and other languages including other African languages.

What is the effect of document length on translation quality? We compare the average d-chrF scores obtained by selected models, including GPT-3.5/4 and LLama3.1-SFT$_k$ where k ={1,5,10}. The evaluation was conducted across all pseudo-document lengths: 1, 5, 10, 25, and the full length. Figure 3 shows that for translations into African languages, d-chrF scores decrease as document length increases. A similar trend is observed for the reverse translation, except for GPT-4o, which shows an increasing trend.

What language benefits more from supervised finetuning? We focus on the sentence-level task and translated across all 30 directions for which the model was trained, evaluating both NLLB-200 (1.3B) and its fine-tuned version using d-chrF. Figures 15 and 16 show performance improvements after supervised fine-tuning of NLLB-200 for both domains. The results shows that translating into Yorùbá, which is the direction with the lowest d-chrF score from English among all the languages, benefited the most. One major factor contributing to this is the presence of diacritics. Furthermore, looking at their actual performances and not just the differences, our results show that translations into Swahili and English—both relatively high-resource languages—yield higher BLEU and CHRF scores (see Figures 13 and 14), even after supervised fine-tuning. Hence, there is much to be done to improve translation performance between low-resource language pairs.

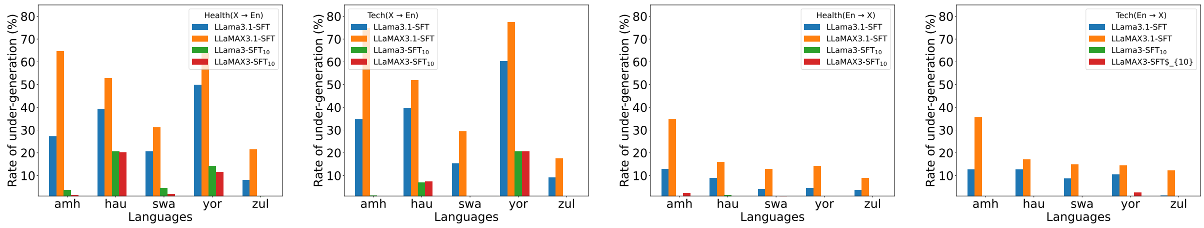

Figure 12: Rate of under-generation in our SFT models.

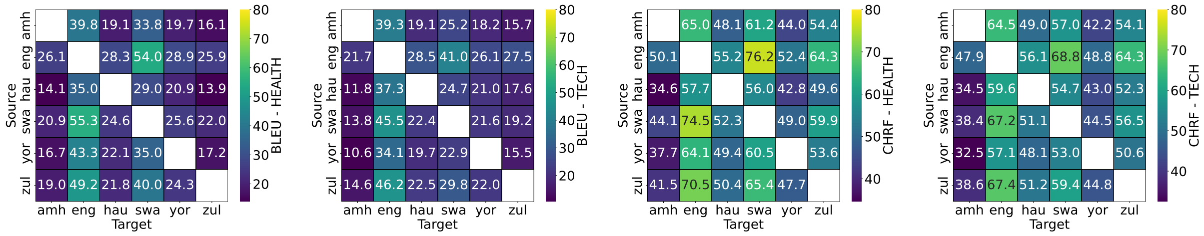

Figure 13: s-BLEU and s-chrF pair-wise comparison of supervised finetuning of NLLB-1.3B on AFRIDOC-MT

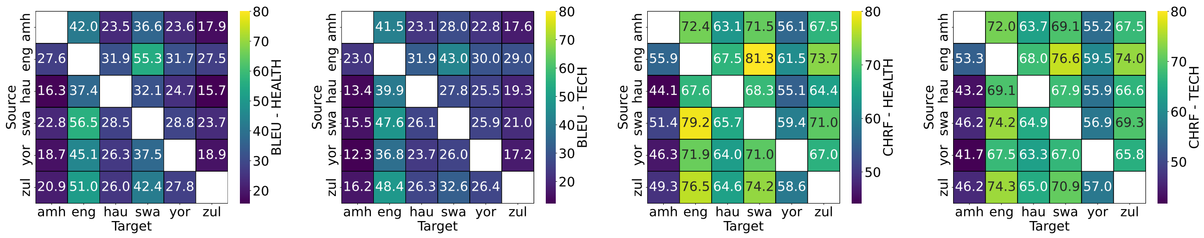

Figure 14: d-BLEU and d-chrF pair-wise comparison of supervised finetuning of NLLB-1.3B on AFRIDOC-MT

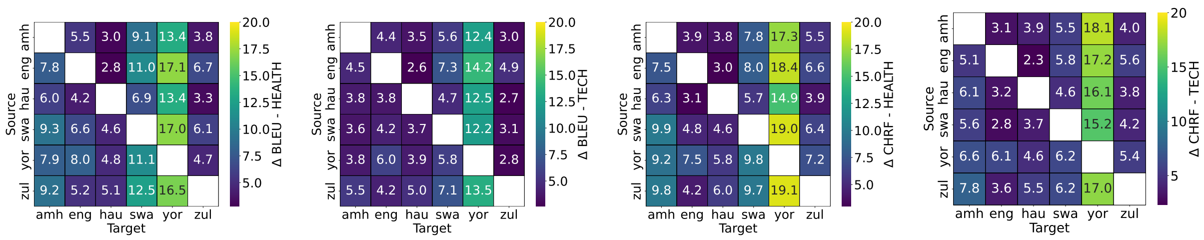

Figure 15: Change (Δ) in s-BLEU and s-chrF for sentence evaluation comparing NLLB1.3B before and after supervised finetuning on AFRIDOC-MT

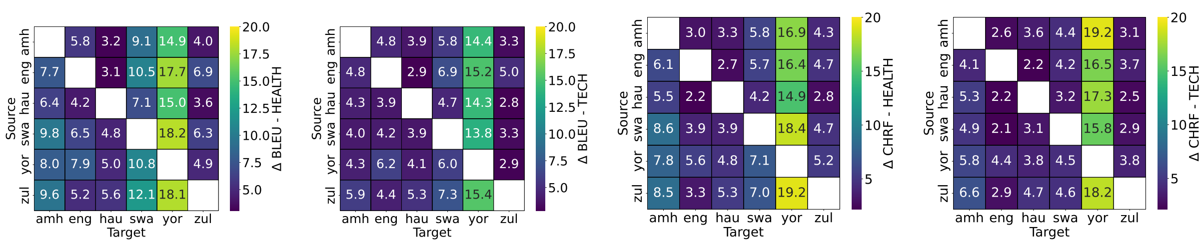

Figure 16: Change (Δ) in d-BLEU and d-chrF for sentence evaluation comparing NLLB1.3B before and after supervised finetuning on AFRIDOC-MT

Model	Size			eng → X					X → eng			AVG
		amh	hau	swa	yor	zul	amh	hau	swa	yor	zul	
							BLEU					
Encoder-Decoder												
M2M-100	0.4B	0.8	0.9	25.6	0.6	3.2	6.7	5.8	32.6	1.7	14.4	9.2
M2M-100	1.2B	2.4	8.9	37.1	2.4	6.9	15.6	13.7	42.6	4.3	23.7	15.8
NLLB-200	0.6B	18.4	26.5	42.0	10.9	19.6	33.0	30.4	45.7	32.4	42.2	30.1
Toucan	1.2B	6.6	18.7	37.3	6.4	9.4	17.4	22.4	31.9	18.1	25.2	19.3
NLLB-200	1.3B	20.0	28.6	44.9	14.0	20.7	36.3	33.1	50.0	37.1	45.9	33.1
NLLB-200	3.3B	24.2	29.7	47.1	13.2	22.2	39.0	34.7	52.7	39.1	48.4	35.0
MADLAD-400	3B	8.0	14.9	42.2	2.3	9.0	36.3	30.6	51.7	15.0	40.4	25.0
MADLAD-400	7.2B	10.5	20.3	44.8	2.4	12.2	40.3	33.7	54.8	27.3	46.6	29.3
Aya-101	13B	7.7/9.6/9.7	18.5/17.2/18.0	6.6/10.9/3.1	5.1/5.1/5.2	11.0/10.0/10.6	29.4/27.4/9.6	28.3/26.2/17.5	42.7/39.2/19.4	24.0/22.4/22.4	36.6/35.1/25.4	21.0/20.3/14.1
SFT on AfriDoc-MT												
NLLB-SFT	1.3B	27.7	31.7	55.4	31.7	27.6	42.1	37.3	56.5	45.1	51.1	40.6
Decoder-only												
Gemma2	9B	0.2/0.4/0.0	0.9/1.3/0.1	0.8/0.6/0.2	0.3/0.4/0.2	0.3/0.0/0.2	6.4/1.7/0.2	5.9/7.6/0.2	6.7/0.3/1.7	3.5/0.7/0.9	6.7/0.2/1.7	3.2/1.3/0.5
LLama3.1	8B	0.4/0.2/0.1	0.5/1.3/0.2	0.2/0.7/0.2	0.3/0.3/0.2	0.2/0.2/0.2	2.7/2.6/0.7	1.9/1.9/0.9	2.9/3.4/0.9	1.6/1.7/0.7	1.6/1.7/0.8	1.2/1.4/0.5
LLaMAX3	8B	2.8/0.1/1.8	1.6/1.8/1.2	2.7/3.5/0.6	0.3/0.3/1.3	0.9/1.0/0.8	5.6/2.6/0.6	2.0/1.9/1.0	3.0/2.7/0.6	1.6/1.4/0.8	2.5/2.1/0.9	2.3/1.7/1.0
LLama3.1-IT	8B	1.2/1.2/1.4	6.3/6.3/5.9	22.9/22.8/19.4	1.5/1.3/1.5	1.0/1.0/0.9	10.1/11.7/9.8	22.0/21.6/20.1	38.0/36.5/36.0	13.0/14.6/12.2	14.7/16.1/14.3	13.1/13.3/12.2
LLaMAX3-Alp	8B	4.9/4.9/5.0	15.3/15.2/16.0	28.2/29.8/16.2	2.5/2.4/2.6	7.3/7.3/7.7	24.1/24.2/23.4	25.8/26.9/25.5	40.9/41.6/39.0	16.3/17.1/15.8	30.2/31.4/29.7	19.6/20.1/18.1
GPT-3.5	–	1.8/0.6/0.5	6.2/1.1/1.0	45.4/45.5/44.5	2.2/0.2/0.3	6.1/1.6/2.1	6.3/7.0/5.8	11.8/11.8/12.1	46.4/45.7/45.4	12.0/13.2/11.5	20.0/22.2/20.4	15.8/14.9/14.3
GPT-4o	–	9.5/6.2/6.0	26.8/26.1/26.6	48.3/51.2/51.4	7.8/7.1/7.5	20.0/21.5/22.2	27.8/29.4/29.8	28.4/29.4/32.0	46.9/48.5/52.5	33.4/35.3/36.9	42.1/44.0/46.9	29.1/29.9/31.2
SFT on AfriDoc-MT												
LLaMAX3-SFT	8B	17.6/17.6/17.9	17.4/18.4/18.7	30.9/34.3/38.3	22.2/21.9/23.4	12.0/13.8/15.6	30.6/31.0/32.2	19.8/23.9/17.8	40.5/35.5/44.8	29.5/32.1/34.4	31.4/31.8/40.7	25.2/26.0/28.4
LLama3.1-SFT	8B	15.7/15.5/16.5	16.5/16.2/17.7	32.1/34.0/35.5	20.4/20.4/22.3	10.1/11.4/15.2	13.1/15.3/29.3	14.1/22.3/24.6	19.8/15.4/42.8	23.2/25.7/33.8	22.2/27.6/37.3	18.7/20.4/27.5
							CHRF					
Encoder-Decoder												
M2M-100	0.4B	14.9	23.4	62.7	11.5	36.7	45.6	41.2	64.4	24.9	50.2	37.6
M2M-100	1.2B	22.4	44.3	70.3	17.6	50.8	54.8	53.0	70.7	32.7	58.8	47.5
NLLB-200	0.6B	48.8	62.7	74.0	42.6	68.1	66.9	63.6	72.8	63.0	70.7	63.3
Toucan	1.2B	33.8	57.6	70.3	36.0	58.0	54.7	57.7	65.2	54.0	59.9	54.7
NLLB-200	1.3B	49.8	64.7	75.5	45.1	69.0	69.4	65.3	75.3	66.3	73.2	65.4
NLLB-200	3.3B	53.0	65.2	76.7	43.8	70.7	70.9	66.5	77.0	67.6	74.7	66.6
MADLAD-400	3B	36.5	54.4	74.2	19.1	57.1	68.9	63.8	76.1	51.4	68.9	57.0
MADLAD-400	7.2B	39.8	59.7	75.2	20.8	61.9	71.5	65.6	78.0	60.6	72.8	60.6
Aya-101	13B	32.0/36.6/36.6	55.4/56.4/55.6	35.2/44.7/28.5	30.9/31.2/29.7	58.5/58.5/58.6	64.6/63.7/23.3	61.5/61.2/48.8	70.8/69.8/43.2	57.9/57.3/55.3	66.9/67.4/53.7	53.4/54.7/43.3
SFT on AfriDoc-MT												
NLLB-SFT	1.3B	55.9	67.4	81.3	61.5	73.7	72.4	67.5	79.2	71.8	76.5	70.7
Decoder-only												
Gemma2	9B	5.9/12.7/0.6	18.8/24.8/8.8	15.5/18.6/10.3	5.8/14.0/6.3	15.6/4.7/7.3	43.3/22.2/7.2	39.8/46.4/6.3	38.2/10.2/16.5	33.8/23.1/12.7	39.8/7.7/23.9	25.6/18.4/10.0
LLama3.1	8B	14.2/13.0/1.1	14.5/23.9/9.3	9.2/18.0/8.8	5.8/9.8/3.5	12.5/15.1/10.3	34.4/34.0/16.7	22.6/23.4/17.5	23.5/27.0/17.2	23.0/23.8/16.9	19.6/20.8/16.8	17.9/20.9/11.8
LLaMAX3	8B	27.0/9.1/13.8	21.4/23.1/17.1	24.8/29.8/13.4	7.4/8.9/8.5	25.0/27.4/19.7	41.0/31.0/10.5	20.5/22.6/16.4	23.0/21.1/15.1	20.7/18.8/18.7	21.8/19.6/18.6	23.3/21.1/15.2
LLama3.1-IT	8B	19.4/19.6/19.5	45.4/45.9/43.8	63.6/63.7/57.3	18.2/17.0/19.7	28.4/28.5/28.0	51.2/53.9/50.7	59.2/59.8/58.6	68.3/69.1/66.7	50.5/53.4/49.2	51.6/54.0/51.6	45.6/46.5/44.5
LLaMAX3-Alp	8B	30.5/30.3/30.4	56.0/55.1/56.3	66.7/67.8/49.1	19.1/19.1/19.3	55.9/56.0/56.1	63.3/62.8/62.9	62.1/62.4/62.3	71.3/71.7/70.8	54.3/56.5/55.1	65.0/65.3/64.9	54.4/54.7/52.7
GPT-3.5	–	20.4/13.1/12.0	44.3/20.4/20.9	76.7/76.6/76.1	21.3/7.3/8.9	51.1/28.0/32.7	47.4/48.3/47.9	52.4/51.2/52.3	74.8/75.0/74.5	50.9/52.1/50.6	58.4/59.5/58.4	49.8/43.1/43.4
GPT-4o	–	36.7/32.4/32.3	64.2/62.4/62.9	79.1/79.8/79.8	29.3/27.2/28.4	69.0/65.6/66.4	66.7/67.2/67.1	65.8/66.0/66.5	77.0/77.5/78.1	68.0/68.9/69.1	74.1/74.7/75.1	63.0/62.2/62.6
SFT on AfriDoc-MT												
LLaMAX3-SFT	8B	46.5/46.8/46.8	61.4/62.0/62.5	66.8/70.7/73.1	56.4/56.2/57.5	60.3/65.1/67.5	64.7/65.6/66.6	53.7/58.9/48.2	69.6/63.7/73.1	60.3/63.2/64.7	60.6/61.4/70.5	60.0/61.4/63.1
LLama3.1-SFT	8B	44.5/44.1/45.6	61.0/60.8/61.8	70.1/71.0/71.5	56.1/56.1/57.0	57.5/59.6/66.8	33.2/39.1/64.3	45.4/58.8/59.5	44.2/38.8/72.1	53.4/56.2/64.8	51.7/60.5/69.0	51.7/54.5/63.2

Table 16: Performance results of various models on the sentence-level task for the Health domain, measured using document level metric d-BLEU and d-CHRF.

Model	Size			eng → X					X → eng			AVG
		amh	hau	swa	yor	zul	amh	hau	swa	yor	zul	
							BLEU					
Encoder-Decoder												
M2M-100	0.4B	1.1	1.7	22.6	1.0	4.7	7.1	10.2	29.0	2.0	15.1	9.5
M2M-100	1.2B	2.8	13.2	29.7	3.7	9.1	16.0	19.0	36.3	5.3	23.1	15.8
NLLB-200	0.6B	16.5	27.3	34.5	12.3	23.4	32.6	33.4	40.5	27.3	40.7	28.9
Toucan	1.2B	5.9	20.4	28.0	8.1	12.4	15.8	25.6	30.1	17.7	25.2	18.9
NLLB-200	1.3B	18.4	28.8	36.1	14.8	24.1	36.8	36.0	43.4	30.6	44.1	31.3
NLLB-200	3.3B	22.9	29.2	37.1	14.2	25.5	39.2	37.0	45.4	31.8	45.7	32.8
MADLAD-400	3B	7.8	16.2	22.2	2.7	11.1	35.8	34.8	43.8	16.2	38.2	22.9
MADLAD-400	7.2B	9.3	21.3	27.5	3.3	14.7	38.3	37.6	44.6	23.7	43.3	26.4
Aya-101	13B	7.8/9.1/9.0	20.8/19.4/20.8	9.2/8.9/4.8	6.3/6.3/6.6	13.2/12.6/13.2	30.7/28.7/11.4	33.3/30.7/19.3	38.7/35.9/20.6	23.4/21.6/22.5	37.4/35.3/28.8	22.1/20.8/15.7
SFT on AfriDoc-MT												
NLLB-SFT	1.3B	23.1	31.7	43.0	29.9	29.1	41.6	39.9	47.6	36.8	48.5	37.1
Decoder-only												
Gemma2	9B	0.2/0.4/0.0	1.3/1.5/0.1	0.9/0.9/0.2	0.3/0.6/0.1	0.3/0.1/0.2	5.2/1.2/0.2	5.2/6.4/0.6	5.0/0.5/0.9	3.4/1.3/0.5	6.1/0.7/0.8	2.8/1.4/0.4
LLama3.1	8B	0.3/0.2/0.1	0.7/1.4/0.3	0.3/0.4/0.2	0.3/0.3/0.2	0.3/0.3/0.2	1.9/2.4/0.5	1.8/2.0/0.8	2.3/3.2/0.6	1.5/1.6/0.6	1.4/1.5/0.6	1.1/1.3/0.4
LLaMAX3	8B	1.9/0.5/1.2	1.6/1.7/2.0	2.0/2.4/1.3	0.4/0.4/1.8	1.0/1.3/0.9	4.2/2.1/0.5	1.9/1.9/1.6	2.4/2.1/0.7	1.3/1.2/0.8	2.3/1.9/1.0	1.9/1.5/1.2
LLama3.1-IT	8B	1.3/1.2/1.2	7.6/7.7/6.9	19.7/19.4/16.1	2.0/1.8/1.9	1.2/1.3/1.2	8.0/9.1/8.2	24.6/23.4/23.0	34.0/31.7/32.2	13.1/13.9/12.3	15.2/14.3/14.2	12.7/12.4/11.7
LLaMAX3-Alp	8B	4.2/4.3/4.1	16.6/16.8/17.9	22.4/21.9/12.9	3.2/3.4/3.5	10.2/10.3/11.1	24.3/25.9/25.3	30.1/30.8/30.4	37.0/37.3/37.0	16.7/17.3/16.6	32.2/33.0/32.5	19.7/20.1/19.1
GPT-3.5	–	1.9/0.8/0.7	9.2/2.4/2.7	35.7/35.4/34.9	3.5/0.6/0.7	7.9/3.0/2.9	6.1/5.8/5.3	17.6/17.1/16.4	41.6/40.2/40.8	13.5/13.3/12.1	23.5/23.3/21.6	16.0/14.2/13.8
GPT-4o	–	7.9/5.7/5.4	28.4/27.3/27.5	40.3/39.8/40.5	7.7/7.3/7.4	26.0/25.1/25.4	31.1/29.9/30.3	37.6/35.1/37.1	46.9/42.9/46.6	32.0/30.5/31.8	46.2/43.3/45.7	30.4/28.7/29.8
SFT on AfriDoc-MT												
LLaMAX3-SFT	8B	11.8/12.2/12.3	16.6/17.1/18.5	19.9/22.0/26.1	19.1/18.9/20.9	10.2/12.9/15.3	25.9/26.2/27.9	15.8/20.1/15.1	29.8/23.1/35.4	22.0/23.7/23.6	25.6/26.3/35.2	19.7/20.3/23.0
LLama3.1-SFT	8B	10.3/10.4/11.0	14.6/15.2/17.5	20.2/20.9/24.0	18.4/17.9/20.5	8.9/10.8/14.5	8.8/9.0/26.5	12.5/19.4/24.5	19.9/14.3/35.0	16.3/17.2/28.2	22.9/24.8/33.6	15.3/16.0/23.5
							CHRF					
Encoder-Decoder												
M2M-100	0.4B	16.9	26.7	62.8	14.2	40.3	46.5	47.3	63.4	28.1	51.5	39.8
M2M-100	1.2B	24.2	50.6	68.2	20.9	52.9	56.1	57.2	67.8	36.6	58.6	49.3
NLLB-200	0.6B	47.7	64.2	71.4	41.4	70.0	67.0	65.0	70.2	60.7	69.3	62.7
Toucan	1.2B	32.0	59.5	66.1	37.1	58.5	54.0	59.9	64.1	54.3	59.6	54.5
NLLB-200	1.3B	49.3	65.7	72.3	43.0	70.3	69.5	66.8	72.0	63.0	71.5	64.3
NLLB-200	3.3B	52.2	65.4	72.8	40.1	71.6	70.9	67.7	73.2	63.9	72.5	65.0
MADLAD-400	3B	37.3	57.0	62.1	21.3	58.5	68.6	66.0	72.1	53.1	67.6	56.4
MADLAD-400	7.2B	39.7	60.6	66.2	22.8	63.5	70.5	67.8	72.3	59.0	70.9	59.3
Aya-101	13B	33.8/37.3/36.6	58.7/58.7/58.9	41.8/42.4/32.7	31.0/31.4/30.0	58.3/58.9/58.4	65.2/64.4/27.2	64.8/64.1/48.7	69.1/68.1/46.2	58.5/57.9/57.1	67.1/66.9/57.7	54.8/55.0/45.3
SFT on AfriDoc-MT												
NLLB-SFT	1.3B	53.4	67.9	76.5	59.5	74.0	72.1	69.0	74.1	67.5	74.3	68.8
Decoder-only												
Gemma2	9B	6.0/12.7/0.7	21.2/24.4/10.8	17.8/21.6/10.4	7.0/15.7/6.5	16.2/14.1/8.0	39.7/19.0/7.7	36.5/42.8/11.0	33.4/16.3/19.6	33.3/28.6/11.7	37.3/21.0/21.9	24.8/21.6/10.8
LLama3.1	8B	13.7/13.2/1.2	16.1/23.4/9.6	10.3/16.6/9.8	6.9/10.7/4.1	13.7/17.8/10.4	30.8/35.1/15.5	20.5/21.9/15.8	20.2/26.7/14.4	21.5/22.9/15.6	18.1/19.4/14.8	17.2/20.8/11.1
LLaMAX3	8B	25.5/19.4/10.7	21.1/22.2/18.5	23.0/26.9/16.0	7.9/9.6/10.7	23.8/26.7/25.0	36.2/28.3/9.9	18.8/22.6/16.7	20.4/19.8/16.8	18.8/17.4/19.1	19.9/18.0/18.9	21.5/21.1/16.2
LLama3.1-IT	8B	19.2/19.5/19.1	47.3/47.8/45.9	63.4/63.4/59.2	20.4/19.4/20.8	29.2/30.4/28.9	49.0/51.0/49.1	60.7/61.0/60.2	66.0/65.8/65.0	51.7/53.5/50.5	51.5/52.4/51.6	45.8/46.4/45.0
LLaMAX3-Alp	8B	30.1/30.2/30.3	58.5/58.1/58.9	64.9/64.0/49.4	21.7/21.8/22.0	58.0/58.0/58.6	62.9/63.4/63.0	64.7/64.9/64.6	68.8/69.1/68.9	55.6/56.5/55.8	65.4/65.7/65.4	55.1/55.2/53.7
GPT-3.5	–	22.6/16.4/15.6	49.2/29.6/31.8	72.6/72.6/72.4	23.0/12.8/14.0	53.6/35.9/35.6	47.3/47.3/47.4	56.3/56.5/56.0	71.5/71.4/71.4	53.2/54.0/52.5	59.6/59.9/58.7	50.9/45.6/45.5
GPT-4o	–	36.9/33.7/33.1	65.2/63.2/63.3	75.3/75.2/75.3	29.4/28.4/28.8	71.1/68.2/68.0	67.2/67.2/66.9	69.1/68.7/68.9	74.4/73.7/74.2	66.2/66.3/66.4	73.4/72.9/73.2	62.8/61.7/61.8
SFT on AfriDoc-MT												
LLaMAX3-SFT	8B	42.0/42.6/42.8	60.9/61.3/62.4	62.7/65.4/67.6	54.0/54.2/55.2	56.4/62.9/66.0	60.5/61.0/63.0	46.5/53.5/43.2	61.4/52.8/67.5	55.0/57.3/55.2	55.2/56.7/66.8	55.5/56.8/59.0
LLama3.1-SFT	8B	40.3/40.3/41.6	59.8/60.2/61.8	64.2/65.1/66.4	53.9/53.7/54.9	54.1/58.4/64.6	22.5/23.7/62.0	40.8/53.4/58.6	47.2/40.0/67.1	44.7/47.2/61.3	54.1/57.3/65.6	48.2/49.9/60.4

Table 17: Performance results of various models on the sentence-level task for the Tech domain, measured using document level metric d-BLEU and d-CHRF.

Model	Size	*eng* → *X* amh	hau	swa	yor	zul	*X* → *eng* amh	hau	swa	yor	zul	AVG
BLEU												
Encoder-Decoder												
M2M-100	0.4B	0.6	0.7	24.1	0.6	2.7	5.2	4.3	30.8	1.2	12.3	8.3
M2M-100	1.2B	1.9	7.0	35.6	2.2	6.3	13.3	11.4	41.0	3.1	21.4	14.3
NLLB-200	0.6B	16.8	23.2	40.2	8.9	18.3	30.9	27.9	44.2	30.2	40.1	28.1
Toucan	1.2B	5.0	15.8	35.4	5.0	8.6	14.7	19.9	30.1	16.0	22.9	17.3
NLLB-200	1.3B	18.3	25.5	43.0	11.7	19.2	34.3	30.8	48.6	35.3	44.0	31.1
NLLB-200	3.3B	22.4	26.5	45.3	10.9	20.6	36.8	32.5	51.4	37.2	46.6	33.0
MADLAD-400	3B	7.1	12.0	40.6	2.1	8.2	34.0	28.2	50.4	12.9	38.4	23.4
MADLAD-400	7.2B	9.9	17.0	43.0	2.2	11.1	37.9	31.2	53.6	25.1	44.7	27.6
Aya-101	13B	6.7/8.4/8.6	15.5/14.9/15.2	6.0/10.2/3.0	3.8/3.8/3.9	10.2/9.6/9.8	27.3/26.3/8.7	26.1/25.4/15.7	41.3/39.7/18.5	22.0/21.5/20.4	35.0/35.0/24.1	19.4/19.5/12.8
SFT on AFRIDOC-MT												
NLLB-SFT	1.3B	26.1	28.3	54.0	28.9	25.9	39.8	34.9	55.3	43.3	49.2	38.6
Decoder-only												
Gemma2	9B	0.1/0.3/0.0	0.5/0.8/0.0	0.6/0.4/0.1	0.2/0.2/0.0	0.2/0.0/0.1	5.2/1.1/0.1	4.9/6.5/0.0	5.7/0.2/0.5	2.7/0.5/0.3	5.8/0.1/0.5	2.6/1.0/0.2
LLama3.1	8B	0.2/0.1/0.0	0.3/0.7/0.1	0.2/0.4/0.1	0.2/0.2/0.2	0.2/0.1/0.1	1.8/1.8/0.3	1.6/1.6/0.5	2.6/3.0/0.6	1.2/1.2/0.4	1.1/1.2/0.4	0.9/1.0/0.3
LLaMAX3	8B	2.1/0.1/1.3	1.1/1.3/0.8	2.4/3.0/0.5	0.3/0.3/1.1	0.7/0.8/0.6	4.5/1.7/0.5	1.7/1.3/0.6	2.7/2.3/0.5	1.3/1.0/0.5	2.2/1.7/0.6	1.9/1.3/0.7
LLama3.1-IT	8B	0.9/0.9/0.8	4.6/4.7/4.2	21.4/21.3/18.0	1.1/0.9/1.0	0.8/0.8/0.7	7.7/8.9/7.3	19.4/19.1/17.7	36.7/35.7/34.7	10.7/12.2/10.1	12.1/13.3/11.8	11.5/11.8/10.6
LLaMAX3-Alp	8B	4.1/4.1/4.1	12.7/12.3/13.2	26.9/28.5/15.1	2.4/2.3/2.4	6.7/6.7/7.0	21.5/21.6/20.9	23.4/24.4/23.3	39.7/40.4/37.8	13.8/14.4/13.2	28.1/29.3/27.5	17.9/18.4/16.4
GPT-3.5	–	1.4/0.4/0.3	4.4/0.8/0.7	43.6/43.6/42.8	1.9/0.2/0.2	5.3/1.4/1.8	4.3/4.4/3.6	9.5/9.3/9.2	45.5/45.3/44.5	10.2/10.8/9.3	18.3/19.9/18.0	14.4/13.6/13.0
GPT-4o	–	8.4/5.0/5.0	24.8/23.4/23.5	48.3/49.7/49.9	7.0/6.2/6.6	19.8/20.1/20.7	26.8/27.6/27.8	27.9/28.7/30.1	48.3/49.6/51.8	33.6/35.0/35.7	42.9/44.2/45.7	28.8/28.9/29.7
SFT on AFRIDOC-MT												
LLaMAX3-SFT	8B	16.2/16.1/16.3	13.6/14.5/14.7	29.2/32.8/36.0	19.2/18.8/20.0	11.1/12.8/14.0	27.4/27.7/28.8	16.9/20.7/15.4	38.3/33.5/42.1	27.1/29.6/31.9	29.0/29.3/37.7	22.8/23.6/25.7
LLama3.1-SFT	8B	14.6/14.3/14.9	13.3/12.9/13.9	31.3/33.0/33.3	18.1/17.9/19.2	9.4/10.7/13.6	11.3/13.3/25.8	12.0/19.3/21.3	18.6/14.4/40.4	21.4/23.8/31.3	20.7/25.9/34.3	17.1/18.6/24.8
CHRF												
Encoder-Decoder												
M2M-100	0.4B	6.8	11.6	51.7	7.5	19.7	30.8	25.0	55.4	13.2	35.9	25.8
M2M-100	1.2B	13.9	28.9	61.7	13.4	33.8	41.2	37.0	63.6	18.6	46.2	35.8
NLLB-200	0.6B	41.6	49.7	66.1	30.9	56.5	57.9	52.2	66.4	52.1	63.2	53.7
Toucan	1.2B	23.7	43.3	61.1	24.2	42.4	41.4	44.8	56.4	39.8	48.1	42.5
NLLB-200	1.3B	42.6	52.2	68.2	34.0	57.7	61.1	54.6	69.7	56.6	66.3	56.3
NLLB-200	3.3B	46.3	52.9	69.5	32.6	59.8	62.9	56.2	71.8	58.1	68.2	57.8
MADLAD-400	3B	28.3	39.7	66.3	15.1	42.2	60.4	53.0	70.7	35.5	60.5	47.2
MADLAD-400	7.2B	32.0	45.6	67.5	15.4	47.5	63.6	55.3	73.0	47.5	65.7	51.3
Aya-101	13B	23.6/28.0/28.0	40.3/42.1/41.0	25.6/33.5/19.4	17.7/18.2/17.7	43.6/43.9/43.8	54.6/54.3/18.0	50.0/50.3/37.2	63.8/63.5/35.9	44.0/44.1/41.6	58.0/59.4/44.4	42.1/43.7/32.7
SFT on AFRIDOC-MT												
NLLB-SFT	1.3B	50.1	55.2	76.2	52.4	64.3	65.0	57.7	74.5	64.1	70.5	63.0
Decoder-only												
Gemma2	9B	1.6/5.1/0.0	7.6/13.2/0.5	6.8/8.3/0.7	2.3/4.9/0.4	5.5/0.7/0.5	31.2/8.7/1.0	28.7/34.6/0.6	27.4/7.1/1.2	20.0/13.8/1.2	30.3/5.2/2.7	16.1/10.2/0.9
LLama3.1	8B	4.8/4.3/0.4	6.1/11.6/5.7	5.5/8.2/5.6	2.6/3.6/2.9	5.7/5.8/5.9	21.7/21.2/8.0	16.6/17.3/9.5	19.3/21.7/9.8	14.9/15.5/9.0	12.5/13.5/9.1	11.0/12.3/6.6
LLaMAX3	8B	17.1/5.0/6.5	14.3/15.8/6.7	19.7/23.6/5.4	5.5/6.3/3.8	16.9/17.9/7.1	29.2/12.3/5.6	15.7/11.7/4.1	20.0/17.1/9.0	14.0/12.1/7.1	17.7/15.2/6.7	17.0/13.7/6.2
LLama3.1-IT	8B	8.8/8.9/8.7	28.7/29.0/26.5	50.9/51.2/43.2	8.5/7.9/8.7	14.9/14.8/14.1	33.4/35.7/32.7	44.7/45.4/44.0	59.6/60.6/57.8	33.6/35.7/32.1	34.2/36.4/34.5	31.7/32.6/30.2
LLaMAX3-Alp	8B	20.8/20.7/20.8	40.5/39.5/41.2	56.0/57.3/36.8	15.4/15.3/15.4	38.9/38.8/39.1	50.9/50.5/50.3	49.3/49.7/49.4	63.7/64.1/62.9	37.3/38.8/37.8	53.5/54.1/53.3	42.6/42.9/40.7
GPT-3.5	–	10.9/6.3/5.8	27.2/12.2/12.2	69.3/69.3/68.5	12.9/4.0/4.6	32.2/16.8/19.8	26.8/28.3/26.9	33.9/33.3/33.0	69.0/69.4/68.4	32.5/33.8/31.8	44.0/45.0/43.1	35.9/31.8/31.4
GPT-4o	–	28.2/24.7/24.6	52.4/49.9/50.3	74.0/74.2/74.1	22.2/20.4/21.2	58.6/53.8/54.6	57.2/57.5/57.3	56.3/56.4/56.4	73.3/73.7/73.5	59.5/60.4/60.3	68.7/69.1/68.9	55.1/54.0/54.1
SFT on AFRIDOC-MT												
LLaMAX3-SFT	8B	38.2/38.4/38.4	44.3/45.0/46.0	55.7/60.3/63.4	43.9/43.7/45.3	44.8/50.1/53.3	52.7/53.6/54.9	39.7/44.6/36.1	60.4/54.3/64.6	49.4/52.1/54.0	50.1/50.7/60.3	47.9/49.3/51.6
LLama3.1-SFT	8B	35.7/35.3/36.7	44.9/44.3/45.0	60.6/61.3/61.1	44.2/43.9/44.9	42.6/44.5/52.4	24.0/28.9/51.8	32.7/44.1/45.2	35.9/30.9/63.2	42.6/45.2/53.6	40.9/48.8/58.0	40.4/42.7/51.2
COMET												
Encoder-Decoder												
M2M-100	0.4B	19.6	20.1	58.3	21.5	26.7	43.9	32.5	66.0	23.5	42.0	35.4
M2M-100	1.2B	29.2	35.4	70.0	37.4	42.6	55.4	47.9	73.3	26.4	53.5	47.1
NLLB-200	0.6B	70.5	69.6	75.8	71.5	73.4	73.9	68.7	77.2	68.2	72.6	72.2
Toucan	1.2B	56.3	63.3	72.6	64.1	62.5	62.1	62.1	70.7	56.9	60.0	63.1
NLLB-200	1.3B	71.7	71.2	77.3	72.9	74.2	76.0	70.5	78.9	71.4	74.5	73.9
NLLB-200	3.3B	72.8	70.9	77.5	70.8	74.8	77.2	71.3	79.7	72.9	75.5	74.3
MADLAD-400	3B	65.1	62.7	75.9	49.5	65.8	76.6	69.8	79.5	52.8	71.2	66.9
MADLAD-400	7.2B	69.1	67.4	77.1	55.0	69.2	78.2	71.9	80.2	65.6	74.9	70.9
Aya-101	13B	53.7/62.0/61.2	62.0/64.2/62.4	31.7/44.2/46.3	50.0/50.2/46.8	62.8/63.7/63.8	73.5/73.0/49.7	67.6/68.0/60.0	76.1/75.0/62.1	62.0/62.8/59.3	67.9/70.2/58.6	60.7/63.4/57.0
SFT on AFRIDOC-MT												
NLLB-SFT	1.3B	75.4	74.0	80.2	78.9	75.7	78.4	72.6	80.5	75.8	76.6	76.8
Decoder-only												
Gemma2	9B	17.2/18.8/10.1	27.0/37.5/12.7	27.7/31.1/13.0	16.2/25.9/11.6	21.1/16.3/12.7	55.0/35.0/15.6	55.4/61.2/15.8	53.7/49.6/15.8	44.3/44.1/18.2	55.4/42.1/16.0	37.3/36.2/14.1
LLama3.1	8B	15.2/15.0/19.4	20.0/25.9/22.6	24.9/25.0/25.8	14.8/20.0/17.9	20.5/21.3/23.9	35.2/39.6/29.1	33.0/32.3/32.3	33.1/44.9/33.4	25.5/27.2/30.6	28.1/27.1/30.3	25.0/27.8/26.5
LLaMAX3	8B	34.3/28.2/28.1	27.1/27.8/23.9	31.8/43.7/25.9	22.9/27.4/22.9	32.6/39.6/26.2	36.1/31.2/18.7	34.5/31.2/17.3	31.9/42.7/26.3	28.8/37.1/19.2	30.0/39.7/22.5	31.0/34.9/23.1
LLama3.1-IT	8B	20.3/20.2/20.0	43.1/42.8/39.4	61.2/61.7/56.0	30.8/29.5/31.9	24.4/24.2/24.2	52.9/56.1/51.2	61.7/61.8/60.7	71.8/70.0/70.7	49.7/53.4/47.0	47.1/49.9/46.0	46.3/47.0/44.7
LLaMAX3-Alp	8B	45.9/46.0/45.8	60.9/60.5/61.6	68.9/69.7/57.9	45.2/45.2/45.1	58.6/58.8/58.3	71.6/71.8/71.4	68.3/69.0/68.7	75.9/76.5/75.6	57.0/60.5/58.2	67.5/68.5/67.2	62.0/62.7/61.0
GPT-3.5	–	22.4/22.9/21.9	35.0/34.7/34.6	78.0/78.1/77.0	36.2/33.2/34.9	43.2/41.1/41.6	44.4/46.9/42.9	51.1/51.3/48.5	78.2/78.5/77.1	50.3/53.3/47.9	57.4/59.1/55.6	49.6/49.9/48.2
GPT-4o	–	55.5/56.5/56.5	71.1/68.1/68.9	79.6/80.1/80.2	54.3/51.5/52.1	72.6/68.0/68.9	73.5/74.6/74.3	71.0/71.3/71.7	78.5/79.4/80.1	71.9/73.5/73.0	73.6/75.1/75.3	70.2/69.8/70.1
SFT on AFRIDOC-MT												
LLaMAX3-SFT	8B	66.8/67.3/66.5	67.2/67.5/67.2	65.6/68.3/71.5	74.5/74.6/75.1	57.7/63.1/66.8	71.5/72.4/73.5	59.0/63.1/56.4	72.5/68.0/76.2	62.9/66.1/68.1	61.5/62.3/71.7	65.9/67.3/69.3
LLama3.1-SFT	8B	61.1/61.9/62.5	63.9/64.4/66.6	66.0/67.4/68.5	73.7/73.9/74.3	53.1/55.4/64.2	48.9/52.0/70.8	53.0/62.3/63.3	55.1/52.6/74.7	54.1/57.5/67.7	52.4/59.3/68.8	58.1/60.7/68.1

Table 18: Performance results of various models on the sentence-level task for the Health domain, measured using sentence level metric s-BLEU, s-CHRF, and s-COMET.

Model	Size	eng → X amh	eng → X hau	eng → X swa	eng → X yor	eng → X zul	X → eng amh	X → eng hau	X → eng swa	X → eng yor	X → eng zul	AVG
BLEU												
Encoder-Decoder												
M2M-100	0.4B	0.9	1.2	20.9	0.9	4.0	5.5	8.2	26.7	1.4	12.9	8.3
M2M-100	1.2B	2.6	10.9	27.6	3.4	8.2	13.6	16.3	34.0	4.0	20.7	14.1
NLLB-200	0.6B	15.3	24.3	32.3	9.7	22.0	30.7	30.7	38.2	24.6	38.5	26.6
Toucan	1.2B	4.8	17.6	25.8	6.1	11.4	13.2	22.8	27.7	15.1	22.8	16.7
NLLB-200	1.3B	17.2	25.9	33.8	11.9	22.6	34.9	33.5	41.3	28.1	42.0	29.1
NLLB-200	3.3B	21.8	26.4	34.9	11.5	24.2	37.3	34.4	43.3	29.2	43.7	30.7
MADLAD-400	3B	7.0	13.7	20.3	2.4	9.8	33.4	32.2	41.6	13.7	36.0	21.0
MADLAD-400	7.2B	8.8	18.2	25.4	2.9	13.4	36.1	35.0	42.3	20.9	41.2	24.4
Aya-101	13B	6.8/7.9/7.8	18.1/17.6/18.0	8.5/8.4/4.5	4.9/5.0/5.2	12.2/12.1/12.3	28.5/27.9/10.6	31.2/30.3/17.3	36.8/36.2/19.1	21.0/20.5/19.9	35.5/35.2/26.9	20.3/20.1/14.2
SFT on AFRIDOC-MT												
NLLB-SFT	1.3B	21.7	28.5	41.0	26.1	27.5	39.4	37.3	45.5	34.1	46.2	34.7
Decoder-only												
Gemma2	9B	0.1/0.3/0.0	0.8/1.1/0.0	0.7/0.7/0.1	0.2/0.4/0.1	0.2/0.1/0.0	4.3/0.8/0.1	4.4/5.5/0.2	4.1/0.4/0.2	2.6/1.0/0.1	5.3/0.6/0.2	2.3/1.1/0.1
LLama3.1	8B	0.2/0.1/0.0	0.4/0.8/0.2	0.2/0.3/0.2	0.2/0.2/0.2	0.2/0.2/0.2	1.3/1.7/0.2	1.6/1.7/0.5	2.0/2.9/0.4	1.1/1.2/0.3	1.0/1.1/0.3	0.8/1.0/0.3
LLaMAX3	8B	1.4/0.3/0.8	1.2/1.3/1.5	1.7/2.0/1.1	0.3/0.3/1.4	0.9/1.0/0.7	3.5/1.4/0.4	1.6/1.3/1.2	2.1/1.8/0.5	1.1/0.8/0.6	2.0/1.6/0.7	1.6/1.2/0.9
LLama3.1-IT	8B	1.0/0.9/0.9	5.8/6.1/5.3	17.8/17.7/14.6	1.4/1.3/1.3	1.1/1.1/1.0	5.9/6.8/6.0	22.1/21.1/20.6	32.0/30.4/30.1	10.7/11.3/9.9	12.5/11.8/11.6	11.0/10.8/10.1
LLaMAX3-Alp	8B	3.7/3.8/3.6	14.0/14.2/15.2	20.8/20.3/11.7	2.9/3.1/3.3	9.3/9.4/10.3	21.9/23.4/22.8	27.8/28.5/28.2	35.2/35.5/35.4	13.9/14.4/13.8	30.0/30.9/30.4	18.0/18.4/17.5
GPT-3.5	–	1.5/0.6/0.5	6.9/1.8/1.9	33.5/33.3/32.9	2.9/0.5/0.5	7.0/2.5/2.5	3.8/3.7/3.2	14.8/14.7/13.6	40.0/39.6/39.2	10.7/11.1/9.6	21.1/21.6/19.3	14.2/12.9/12.3
GPT-4o	–	7.0/4.9/4.6	25.6/24.7/24.7	38.4/38.1/38.6	6.6/6.3/6.4	24.8/24.1/24.1	29.0/28.6/28.2	35.4/34.6/35.1	45.2/43.5/45.0	29.8/29.6/29.7	44.6/43.4/44.1	28.6/27.8/28.1
SFT on AFRIDOC-MT												
LLaMAX3-SFT	8B	10.6/10.8/11.0	13.3/13.8/14.7	18.1/20.1/23.4	15.3/15.2/16.6	9.4/11.9/13.7	22.9/23.2/24.7	13.7/17.8/13.2	27.2/21.2/32.3	19.4/20.9/20.8	23.6/24.2/32.2	17.4/17.9/20.3
LLama3.1-SFT	8B	9.6/9.5/9.7	12.1/12.4/13.7	19.1/19.6/21.4	15.5/14.8/16.3	8.3/10.1/13.1	7.6/7.9/23.2	11.1/17.3/21.6	18.4/13.2/32.0	14.5/15.5/25.1	21.3/23.2/30.5	13.7/14.4/20.7
CHRF												
Encoder-Decoder												
M2M-100	0.4B	8.9	14.9	50.3	10.1	22.7	31.7	31.1	52.9	15.6	36.7	27.5
M2M-100	1.2B	16.4	36.0	57.4	16.7	35.6	42.5	42.3	58.6	21.7	45.5	37.3
NLLB-200	0.6B	41.1	51.9	61.7	29.6	58.3	58.1	53.8	62.0	47.8	60.9	52.5
Toucan	1.2B	22.5	45.6	55.4	24.9	43.3	40.0	47.3	53.9	38.9	47.4	41.9
NLLB-200	1.3B	42.8	53.8	63.0	31.6	58.7	61.4	56.4	64.4	51.0	63.8	54.7
NLLB-200	3.3B	46.3	53.7	63.7	29.9	60.5	63.1	57.4	65.8	52.0	64.9	55.7
MADLAD-400	3B	29.1	43.0	51.5	16.6	43.6	60.0	55.6	64.6	36.8	58.3	45.9
MADLAD-400	7.2B	32.2	46.7	54.9	17.1	49.1	62.6	58.0	64.9	44.5	62.7	49.3
Aya-101	13B	25.8/29.2/28.5	44.6/45.5/45.0	31.1/31.9/22.7	19.1/19.8/19.7	43.2/44.5/43.5	55.5/55.3/21.8	54.0/54.4/38.1	60.7/60.7/37.9	44.0/44.1/42.5	57.9/58.7/47.7	43.6/44.4/34.7
SFT on AFRIDOC-MT												
NLLB-SFT	1.3B	47.9	56.1	68.8	48.8	64.3	64.5	59.6	67.2	57.1	67.4	60.2
Decoder-only												
Gemma2	9B	1.8/5.3/0.1	10.1/14.0/0.4	8.9/12.7/0.4	3.1/6.3/0.3	6.5/5.0/0.3	28.1/6.9/1.0	26.9/32.5/1.0	22.9/11.7/1.1	20.4/17.5/0.7	28.2/15.3/1.7	15.7/12.7/0.7
LLama3.1	8B	4.7/4.3/0.4	7.4/12.1/5.9	6.0/5.6/6.1	3.4/4.4/3.2	5.9/6.2/6.0	18.7/21.4/6.8	15.5/16.5/8.8	16.2/21.1/8.6	14.1/14.9/8.3	11.7/12.7/8.3	10.3/11.9/6.2
LLaMAX3	8B	16.6/11.9/4.3	14.9/15.5/8.5	17.7/20.2/7.2	6.0/6.8/5.5	16.4/18.1/9.0	25.8/11.4/5.5	14.7/10.6/5.6	17.1/14.5/7.7	12.8/11.0/8.1	16.0/13.9/5.8	15.8/13.4/6.7
LLama3.1-IT	8B	8.8/8.9/8.5	30.6/30.9/28.7	49.0/49.1/44.2	10.3/9.9/10.2	15.4/16.2/15.1	30.8/32.5/30.8	46.5/46.6/45.9	55.8/55.8/54.5	34.0/35.0/32.6	34.0/34.5/33.8	31.5/31.9/30.4
LLaMAX3-Alp	8B	20.9/21.0/20.9	43.3/43.0/44.3	52.4/51.2/36.0	17.4/17.5/17.7	40.6/40.8/41.6	50.6/51.4/51.0	52.4/52.7/52.4	59.6/59.9/59.8	37.9/38.9/38.1	53.9/54.5/54.1	42.9/43.1/41.6
GPT-3.5	–	12.4/8.4/8.0	31.8/19.0/20.0	63.4/63.4/63.0	15.4/7.9/8.4	35.1/22.3/22.3	26.4/27.1/26.2	38.4/38.9/37.7	63.8/64.3/63.5	33.8/35.4/33.2	45.0/45.7/43.9	36.6/33.2/32.6
GPT-4o	–	28.6/26.1/25.4	53.5/51.5/51.5	67.2/67.2/67.3	22.3/21.4/21.4	60.0/56.5/56.4	57.3/57.7/57.0	59.8/60.0/59.6	67.7/68.1/67.5	55.0/56.0/55.3	66.2/66.4/66.0	53.7/53.1/52.7
SFT on AFRIDOC-MT												
LLaMAX3-SFT	8B	33.1/33.7/33.9	44.3/44.6/45.6	48.4/51.7/54.4	39.6/39.9/41.0	40.2/46.2/49.8	47.9/48.3/50.6	34.7/40.5/32.5	49.9/42.0/56.6	42.2/44.1/42.8	44.2/45.4/55.3	42.4/43.6/46.3
LLama3.1-SFT	8B	31.5/31.2/32.4	43.8/44.1/44.9	50.9/52.2/52.5	40.6/40.1/40.9	38.6/42.5/48.5	16.5/17.6/48.9	30.4/40.2/44.7	37.3/31.2/55.8	34.3/36.0/47.9	42.4/45.2/53.4	36.6/38.0/47.0
COMET												
Encoder-Decoder												
M2M-100	0.4B	23.4	22.5	58.0	22.1	29.1	46.0	39.4	64.9	28.2	44.8	37.8
M2M-100	1.2B	34.2	42.2	67.1	37.7	42.9	57.7	54.4	70.4	32.1	54.4	49.3
NLLB-200	0.6B	69.1	69.5	72.7	70.0	72.0	72.8	69.3	74.1	66.5	71.0	70.7
Toucan	1.2B	54.7	63.1	67.2	64.3	61.4	60.7	64.3	68.8	58.4	60.3	62.3
NLLB-200	1.3B	69.4	70.9	73.1	70.2	72.8	75.1	71.3	75.6	69.1	72.6	72.0
NLLB-200	3.3B	71.2	70.2	73.4	66.6	73.2	76.0	71.7	76.0	70.3	73.2	72.2
MADLAD-400	3B	65.0	62.3	64.7	50.6	63.8	75.9	71.5	76.2	56.3	69.9	65.6
MADLAD-400	7.2B	67.8	64.9	66.5	56.7	68.3	77.4	73.6	76.5	65.4	72.9	69.0
Aya-101	13B	56.9/63.4/61.6	63.7/65.8/64.5	36.7/39.6/47.5	51.7/52.7/48.8	60.6/63.2/62.5	73.2/72.3/51.4	70.0/70.4/60.9	73.4/72.8/62.9	64.0/64.0/62.7	68.4/69.6/62.7	61.9/63.4/58.5
SFT on AFRIDOC-MT												
NLLB-SFT	1.3B	74.1	73.3	76.4	78.1	73.9	77.8	74.3	77.4	73.9	75.9	75.5
Decoder-only												
Gemma2	9B	17.7/19.5/10.9	34.0/39.5/13.8	33.0/40.3/14.3	18.1/27.4/12.0	24.0/23.4/14.0	51.8/30.4/16.6	55.9/61.5/17.4	51.4/57.4/17.3	45.9/48.4/19.6	54.6/55.5/17.3	38.6/40.3/15.3
LLama3.1	8B	14.9/14.5/19.0	22.2/26.7/23.0	25.3/17.1/26.1	16.3/21.1/18.3	20.3/21.3/24.5	32.6/42.5/27.3	34.4/36.1/32.5	34.0/46.9/32.6	26.3/29.8/30.6	27.8/27.6/29.9	25.4/28.4/26.4
LLaMAX3	8B	33.8/32.3/26.1	29.6/28.1/24.3	33.4/41.9/26.3	25.8/29.5/22.7	33.2/41.5/27.2	35.3/30.2/18.7	37.9/30.1/17.9	34.4/39.6/24.1	31.2/36.6/20.2	32.6/38.9/21.9	32.7/34.9/23.0
LLama3.1-IT	8B	20.9/21.3/20.9	43.3/42.7/40.4	60.2/59.9/56.0	31.1/30.4/30.9	25.9/26.3/25.7	49.4/51.7/48.1	62.2/61.1/61.2	69.4/65.2/68.6	51.1/52.8/48.8	46.5/47.1/45.4	46.0/45.8/44.6
LLaMAX3-Alp	8B	47.0/47.2/47.0	61.6/61.2/62.2	66.0/65.2/56.0	45.2/45.5/45.1	58.4/58.6/58.6	71.0/71.6/71.0	70.3/70.7/70.1	73.8/74.1/73.8	59.0/61.4/59.6	67.5/68.1/67.3	62.0/62.4/61.1
GPT-3.5	–	25.8/26.3/25.3	40.8/41.1/39.7	74.8/74.9/73.6	38.0/36.8/37.9	46.6/43.4/44.3	45.7/48.5/44.6	55.7/56.8/54.1	75.3/75.2/74.4	53.4/55.9/51.4	59.5/60.5/58.1	51.6/51.9/50.3
GPT-4o	–	57.5/58.4/58.5	71.4/69.4/69.1	77.4/77.1/77.2	53.6/51.6/51.9	72.7/68.6/68.9	74.0/73.7/73.7	74.9/74.1/74.6	77.6/76.5/77.5	72.0/72.5/72.0	74.6/73.6/74.1	70.6/69.5/69.8
SFT on AFRIDOC-MT												
LLaMAX3-SFT	8B	62.5/63.0/62.3	64.4/64.7/65.1	60.6/62.5/65.5	72.2/72.7/73.9	52.5/58.1/62.8	67.9/68.6/70.5	55.2/59.7/54.4	66.4/60.2/71.9	58.1/60.8/59.1	56.3/57.6/68.3	61.6/62.8/65.4
LLama3.1-SFT	8B	56.0/56.5/56.9	59.8/60.8/64.0	58.9/61.3/62.1	72.0/72.0/73.2	47.3/51.4/59.0	41.5/41.6/68.2	50.5/57.7/62.8	54.5/50.7/70.1	46.8/48.7/65.8	52.6/55.2/65.5	54.0/55.6/64.8

Table 19: Performance results of various models on the sentence-level task for the Tech domain, measured using sentence level metric s-BLEU, s-CHRF, and s-COMET.

Model	Size	*eng → X*					*X → eng*					AVG
		amh	hau	swa	yor	zul	amh	hau	swa	yor	zul	
						BLEU						
Encoder-Decoder												
Toucan	1.2B	2.6	9.3	17.4	3.2	4.5	8.6	8.0	18.1	8.2	12.4	9.2
NLLB-200	1.3B	4.7	8.0	13.7	2.7	8.2	6.1	10.7	20.8	9.9	16.1	10.1
NLLB-200	3.3B	5.2	5.6	14.2	2.3	7.4	12.1	16.0	26.9	12.7	23.7	12.6
MADLAD-400	3B	5.9	8.0	17.0	1.5	5.7	31.5	30.9	50.8	14.2	38.7	20.4
MADLAD-400	7.2B	1.4	5.5	13.9	1.5	4.3	20.4	12.5	41.7	4.4	17.9	12.4
Aya-101	13B	6.4/6.8/6.1	12.4/15.4/12.7	10.4/5.5/3.5	2.3/2.8/2.6	10.3/10.3/9.7	28.2/28.2/7.2	30.2/29.8/16.9	43.4/43.2/24.0	26.0/25.8/20.4	39.7/39.5/34.9	20.9/20.7/13.8
SFT on AFRIDOC-MT (sentence)												
NLLB-SFT	1.3B	7.9	13.5	26.0	6.9	13.5	13.5	15.3	26.0	15.4	22.6	16.1
SFT on AFRIDOC-MT (pseudo-document with 10)												
NLLB	1.3B	8.7	13.4	25.9	6.5	13.6	20.8	20.0	30.1	19.0	26.4	18.4
Decoder-only												
Gemma2-IT	9B	0.2/0.2/0.2	10.2/8.5/7.9	21.3/23.7/18.3	0.2/0.2/0.2	0.4/0.4/0.4	8.5/9.7/6.0	21.6/22.9/18.8	37.2/40.2/33.6	12.5/14.7/9.4	24.8/27.6/21.7	13.7/14.8/11.6
LLama3.1-IT	8B	0.1/0.1/0.1	0.6/0.6/0.4	7.9/10.7/6.8	0.2/0.2/0.1	0.1/0.1/0.1	4.1/5.2/5.1	19.3/20.8/4.0	32.0/35.6/2.3	7.1/9.2/6.2	11.3/11.6/8.2	8.3/9.4/3.3
LLaMAX3-Alp	8B	0.7/0.6/0.6	3.0/3.1/3.2	6.1/7.3/6.3	0.4/0.4/0.3	1.0/1.1/1.1	6.3/5.0/7.5	14.1/11.5/12.4	25.5/25.2/25.2	2.6/2.6/2.3	8.7/11.8/10.2	6.8/6.9/6.9
GPT-3.5	–	0.4/0.5/0.4	1.1/1.2/1.3	45.5/45.1/45.2	0.2/0.3/0.3	1.6/1.9/1.9	4.2/6.1/3.7	16.1/16.0/15.7	51.6/51.5/51.5	15.9/15.8/14.8	25.7/27.1/26.6	16.2/16.5/16.1
GPT-4o	–	6.3/6.2/6.8	27.1/27.3/27.4	52.4/52.9/52.6	7.4/7.4/8.3	22.6/22.4/22.1	35.4/35.2/35.6	37.5/38.1/38.1	57.8/57.9/58.2	46.0/45.6/46.0	52.5/53.0/53.0	34.5/34.6/34.8
SFT on AFRIDOC-MT (sentence)												
LLaMAX3-SFT	8B	4.5/4.1/4.6	2.9/2.3/2.5	7.6/7.3/9.0	4.9/5.0/5.7	2.8/2.2/3.0	2.8/2.5/3.0	2.4/2.0/2.6	6.7/4.2/5.4	4.2/3.2/5.9	6.5/4.2/6.4	4.5/3.7/4.8
LLama3.1-SFT	8B	3.7/3.0/3.8	3.5/2.8/3.3	11.8/13.3/12.2	5.4/4.8/4.9	4.1/3.1/4.1	2.0/2.4/3.2	2.3/1.8/3.2	5.1/5.6/7.0	3.8/3.2/5.6	3.5/3.7/5.0	4.5/4.4/5.2
SFT on AFRIDOC-MT (pseudo-document with 10)												
LLaMAX3-SFT	8B	10.8/9.5/10.0	10.6/10.3/11.9	35.6/34.0/39.9	18.5/15.8/17.9	9.9/9.4/8.9	29.4/28.9/28.0	34.7/30.1/33.4	51.6/51.0/54.1	44.2/37.1/44.4	47.1/46.5/47.3	29.2/27.3/29.6
LLama3.1-SFT	8B	4.8/4.9/5.0	8.0/10.0/10.2	26.8/24.6/26.2	16.4/15.8/15.3	5.6/5.8/6.7	23.2/19.4/29.8	27.1/19.7/23.9	22.9/25.9/46.8	24.3/29.0/37.5	37.1/26.7/42.7	19.6/18.2/24.4
						CHRF						
Encoder-Decoder												
Toucan	1.2B	18.9	36.5	44.4	23.0	38.5	41.1	42.0	45.2	39.7	43.3	37.2
NLLB-200	1.3B	25.0	35.5	40.4	19.5	38.8	30.7	37.1	46.9	34.7	42.6	35.1
NLLB-200	3.3B	25.6	30.4	40.2	18.4	35.4	39.7	44.5	53.6	38.2	50.7	37.7
MADLAD-400	3B	27.5	40.2	46.6	15.1	43.6	63.3	62.5	74.4	44.2	66.6	48.4
MADLAD-400	7.2B	5.3	30.6	39.8	13.4	26.1	47.2	36.2	64.5	17.2	41.2	32.1
Aya-101	13B	27.0/28.7/25.9	41.9/48.5/43.2	34.7/28.8/25.6	17.1/18.7/18.0	54.2/54.9/52.7	61.6/61.1/16.1	62.3/62.0/44.7	71.2/71.0/48.1	56.1/55.9/46.1	69.0/68.9/63.8	49.5/49.8/38.4
SFT on AFRIDOC-MT (sentence)												
NLLB-SFT	1.3B	30.2	42.8	52.4	28.4	47.3	42.1	43.8	52.4	42.6	50.3	43.2
SFT on AFRIDOC-MT (pseudo-document with 10)												
NLLB	1.3B	31.2	42.4	52.2	27.7	47.1	50.6	48.7	55.9	47.4	53.5	45.7
Decoder-only												
Gemma2-IT	9B	6.1/6.5/6.0	37.0/34.6/30.1	49.8/52.9/46.4	6.4/6.4/6.2	11.6/12.0/11.9	35.0/36.5/30.8	50.3/51.8/46.8	62.1/65.0/58.4	41.0/44.8/35.9	53.1/56.1/49.3	35.3/36.7/32.2
LLama3.1-IT	8B	7.4/7.5/7.4	14.0/13.8/12.2	37.5/43.2/27.7	6.4/5.6/4.9	8.3/8.7/8.6	23.8/23.3/21.9	46.9/49.3/19.7	59.0/62.8/16.8	29.0/31.7/23.1	33.0/34.0/27.0	26.5/28.0/16.9
LLaMAX3-Alp	8B	11.4/11.1/11.2	28.9/28.6/28.5	35.9/40.4/32.5	9.2/8.9/8.4	22.1/22.3/23.6	28.9/28.0/29.2	41.7/39.2/41.1	54.1/51.9/55.4	23.5/23.3/22.3	37.7/40.5/39.9	29.3/29.4/29.2
GPT-3.5	–	11.3/11.3/11.6	22.0/22.4/23.1	75.9/75.6/76.1	9.1/8.9/10.1	27.7/29.1/29.2	37.9/41.6/38.0	52.7/52.7/52.4	77.7/77.6/77.7	51.7/51.1/50.9	59.7/61.1/60.8	42.6/43.1/43.0
GPT-4o	–	29.3/28.4/29.6	63.0/63.4/63.8	80.1/80.2/80.0	27.7/27.6/29.6	69.5/69.2/68.8	69.5/69.3/69.5	69.0/69.3/69.3	81.0/81.0/81.0	73.8/73.6/73.7	77.7/78.2/77.9	64.1/64.0/64.3
SFT on AFRIDOC-MT (sentence)												
LLaMAX3-SFT	8B	22.2/22.8/24.1	29.0/25.9/26.8	38.4/39.0/42.2	32.3/32.3/33.8	33.3/29.7/33.7	22.6/21.1/20.2	22.1/20.5/22.9	33.1/26.8/30.2	25.0/23.2/27.2	31.5/27.0/30.9	28.9/26.8/29.2
LLama3.1-SFT	8B	25.2/22.7/25.2	31.8/29.2/31.9	48.5/50.2/48.5	33.8/32.6/33.0	35.4/35.1/38.6	15.6/22.9/24.2	20.6/18.6/24.1	28.7/31.3/33.7	25.6/23.5/30.2	24.2/25.2/29.3	28.9/29.1/31.9
SFT on AFRIDOC-MT (pseudo-document with 10)												
LLaMAX3-SFT	8B	37.8/35.9/37.1	49.7/48.2/51.9	72.4/70.5/74.4	50.7/50.1/52.2	55.0/53.4/52.4	64.0/62.7/62.5	66.7/63.5/66.3	75.4/74.4/77.8	71.8/68.3/71.8	74.1/73.8/74.0	61.7/60.1/62.0
LLama3.1-SFT	8B	26.7/27.6/27.4	46.0/49.7/49.6	64.1/64.0/63.4	50.3/50.0/49.5	44.5/44.6/47.0	57.8/56.5/63.8	61.7/55.3/59.6	47.3/53.1/74.4	55.6/61.0/68.9	68.2/59.9/71.4	52.2/52.2/57.5

Table 20: Performance results of various models on the pseudo-document-level task for the Health domain, measured using document level metric d-BLEU and d-CHRF.

Model	Size	*eng → X*					*X → eng*					AVG
		amh	hau	swa	yor	zul	amh	hau	swa	yor	zul	
						BLEU						
Encoder-Decoder												
Toucan	1.2B	2.2	11.2	13.2	4.1	7.4	8.6	15.6	17.9	10.4	14.8	10.5
NLLB-200	1.3B	5.1	11.2	14.0	2.7	9.8	5.8	9.7	21.9	8.1	16.9	10.5
NLLB-200	3.3B	5.1	7.2	11.9	2.2	7.4	10.7	12.9	26.5	10.3	20.9	11.5
MADLAD-400	3B	5.7	6.8	5.4	1.2	6.7	30.6	33.6	39.4	14.9	35.5	18.0
MADLAD-400	7.2B	1.2	4.7	5.0	1.5	4.3	21.2	17.9	31.6	6.7	20.3	11.4
Aya-101	13B	6.3/6.7/5.7	19.5/20.2/18.2	19.5/14.4/5.5	4.1/4.5/4.4	13.0/13.4/11.8	29.0/29.9/7.2	35.5/35.6/24.0	39.8/39.8/25.6	25.1/25.5/22.6	40.0/40.4/36.6	23.2/23.0/16.2
SFT on AFRIDOC-MT (sentence)												
NLLB-SFT	1.3B	7.8	17.1	24.3	7.4	15.3	11.7	19.5	25.5	13.6	23.4	16.6
SFT on AFRIDOC-MT (pseudo-document with 10)												
NLLB	1.3B	8.6	17.4	24.2	7.4	15.2	22.3	23.9	28.9	17.4	27.4	19.3
Decoder-only												
Gemma2-IT	9B	0.2/0.2/0.2	11.4/11.6/8.7	18.8/21.0/14.3	0.3/0.3/0.3	0.7/0.7/0.8	8.5/9.0/8.3	22.1/22.9/21.6	30.3/32.3/28.6	15.1/16.7/12.1	21.6/24.4/19.3	12.9/13.9/11.4
LLama3.1-IT	8B	0.2/0.1/0.1	0.8/0.7/0.6	9.6/8.8/9.5	0.2/0.2/0.2	0.2/0.1/0.1	4.9/5.1/4.5	19.4/19.7/2.2	30.8/31.1/1.6	8.9/10.2/4.5	8.7/8.8/6.0	8.4/8.5/2.9
LLaMAX3-Alp	8B	0.5/0.5/0.5	3.7/3.2/4.7	4.8/5.6/3.2	0.6/0.6/0.7	1.6/1.4/1.8	4.8/5.3/6.7	22.4/23.7/18.8	30.9/24.1/33.5	2.3/2.9/2.2	19.8/21.7/20.3	9.1/8.9/9.2
GPT-3.5	–	0.4/0.4/0.5	2.3/2.4/2.6	35.8/34.8/35.8	0.6/0.6/0.6	2.8/3.0/2.8	3.6/4.5/3.8	19.8/20.1/18.9	45.5/45.6/45.3	15.7/16.0/16.4	25.7/27.1/27.1	15.2/15.4/15.4
GPT-4o	–	5.9/6.1/6.1	28.8/29.0/28.8	40.8/41.2/41.0	7.0/7.4/7.4	26.2/26.1/25.8	35.0/35.4/35.1	42.8/43.3/43.0	51.1/51.2/51.0	38.6/39.3/38.7	51.6/51.6/51.7	32.8/33.1/32.9
SFT on AFRIDOC-MT (sentence)												
LLaMAX3-SFT	8B	2.7/2.9/2.6	2.8/2.5/3.0	5.2/5.1/4.8	4.2/4.2/4.3	2.5/2.5/2.7	4.8/4.9/4.9	2.6/3.9/3.9	4.9/6.0/5.1	3.3/4.7/4.7	5.0/5.5/4.4	3.8/4.2/4.0
LLama3.1-SFT	8B	1.8/1.9/2.0	3.0/3.1/3.1	5.9/6.0/6.8	5.0/4.9/5.1	2.1/2.3/2.3	2.2/2.1/3.2	3.8/3.9/4.4	6.2/4.7/7.3	5.0/4.4/6.2	4.8/3.6/6.0	4.0/3.7/4.6
SFT on AFRIDOC-MT (pseudo-document with 10)												
LLaMAX3-SFT	8B	7.8/8.8/9.8	14.0/15.5/17.8	22.6/24.0/27.7	13.0/14.7/15.0	12.7/10.8/13.7	32.5/30.0/32.1	37.6/33.7/38.2	43.0/40.2/45.2	36.5/31.4/36.8	43.2/36.9/43.5	26.3/24.6/28.0
LLama3.1-SFT	8B	2.8/3.0/3.0	9.6/9.1/8.0	15.9/14.3/11.3	17.6/14.8/16.1	5.9/5.1/5.5	25.0/19.9/26.0	22.8/22.5/33.6	11.6/23.3/42.0	14.6/25.8/34.9	34.4/30.2/34.0	16.0/16.8/21.4
						CHRF						
Encoder-Decoder												
Toucan	1.2B	18.8	41.8	42.5	22.9	39.2	39.0	44.3	46.8	41.1	44.3	38.1
NLLB-200	1.3B	26.7	40.4	42.8	18.8	40.6	30.1	35.0	49.6	32.9	43.2	36.0
NLLB-200	3.3B	26.4	33.4	39.3	17.4	35.0	36.7	38.9	54.4	36.4	47.6	36.5
MADLAD-400	3B	29.5	38.3	31.7	15.1	44.1	62.6	63.5	66.4	45.9	63.4	46.0
MADLAD-400	7.2B	5.2	30.8	33.1	14.2	27.7	46.3	40.8	56.0	23.7	44.0	32.2
Aya-101	13B	29.1/30.1/26.1	54.0/55.0/51.2	51.7/45.3/30.5	21.5/22.3/21.8	53.3/55.0/51.2	61.4/62.5/16.7	65.3/65.5/50.9	68.8/68.7/51.7	55.6/55.7/51.5	68.1/68.4/64.7	52.9/52.9/41.6
SFT on AFRIDOC-MT (sentence)												
NLLB-SFT	1.3B	31.4	47.9	54.7	30.2	49.8	38.8	47.0	53.0	41.3	50.8	44.5
SFT on AFRIDOC-MT (pseudo-document with 10)												
NLLB	1.3B	32.8	48.0	54.6	29.6	49.9	52.4	52.4	56.3	47.1	54.8	47.8
Decoder-only												
Gemma2-IT	9B	5.7/6.2/5.7	39.9/42.1/34.5	46.7/51.0/38.7	6.6/6.6/6.4	14.9/14.8/15.4	34.7/35.9/34.0	49.4/50.1/48.2	55.4/57.7/53.6	45.7/48.2/40.7	48.4/51.7/45.8	34.7/36.4/32.3
LLama3.1-IT	8B	7.4/7.2/6.8	15.3/13.9/14.1	42.0/43.3/32.4	6.1/5.7/6.2	8.8/8.2/8.8	25.6/26.1/23.0	48.3/48.7/17.4	58.7/59.0/16.0	31.0/34.4/23.4	32.0/34.7/27.8	27.5/28.1/17.6
LLaMAX3-Alp	8B	10.9/10.8/11.4	30.5/27.8/32.5	35.5/38.1/29.0	11.2/11.5/12.0	26.1/24.1/26.0	28.5/29.4/29.0	50.4/51.4/48.5	58.5/54.3/62.4	22.5/24.7/21.8	48.7/48.3/48.8	32.3/32.0/32.1
GPT-3.5	–	13.2/13.4/13.5	28.7/28.7/29.7	72.1/71.7/72.0	12.4/12.2/12.7	33.8/35.1/33.8	36.8/38.5/38.5	56.2/56.3/54.5	73.4/73.5/73.2	51.5/52.7/53.0	58.8/61.2/60.9	43.7/44.3/44.2
GPT-4o	–	31.1/30.4/31.3	64.7/65.1/64.6	75.1/75.0/75.0	27.8/28.0/28.1	70.7/70.6/70.7	68.4/68.6/68.2	71.4/71.6/71.2	76.4/76.5/76.3	69.9/70.1/69.8	76.5/76.5/76.3	63.2/63.2/63.2
SFT on AFRIDOC-MT (sentence)												
LLaMAX3-SFT	8B	21.3/21.5/21.7	29.1/27.9/29.9	36.3/37.0/34.7	30.2/30.1/30.5	31.3/31.4/31.7	21.4/24.2/21.2	22.0/27.6/26.0	29.5/32.3/30.0	23.6/28.5/26.2	29.7/29.8/27.1	27.4/29.0/27.9
LLama3.1-SFT	8B	20.4/20.9/21.0	30.6/30.8/30.0	38.3/38.5/40.0	32.8/32.3/33.4	26.3/29.3/28.2	12.2/22.0/23.9	27.2/28.6/28.9	33.5/28.7/36.9	29.1/29.7/32.2	29.5/26.2/32.3	28.0/28.7/30.7
SFT on AFRIDOC-MT (pseudo-document with 10)												
LLaMAX3-SFT	8B	34.7/36.4/37.7	54.1/58.1/58.6	64.7/62.9/68.3	47.2/47.9/49.3	58.9/56.5/60.9	65.4/63.5/64.2	68.2/66.3/68.5	70.7/70.8/73.1	67.5/66.2/67.7	71.4/69.3/71.6	60.3/59.8/62.0
LLama3.1-SFT	8B	22.6/23.5/23.7	47.0/45.2/46.7	58.6/57.2/51.4	49.7/47.2/49.5	43.8/40.0/41.4	59.9/55.8/60.9	58.0/56.1/65.4	35.8/51.3/71.1	44.1/57.2/66.3	66.1/60.1/66.4	48.6/49.4/54.3

Table 21: Performance results of various models on the pseudo-document-level task for the Tech domain, measured using document level metric d-BLEU and d-CHRF.

Model	Setup	*eng* → *X*					*X* → *eng*				
		d-CHRF↑	**Fluency↑**	**CE↓**	**LE↓**	**GE↓**	**d-CHRF↑**	**Fluency↑**	**CE↓**	**LE↓**	**GE↓**
Amharic											
Aya-101	Sent	36.6	2.8	9.0	3.8	2.5	64.6	3.0	15.8	9.6	5.1
	Doc10	28.7	2.5	8.9	2.9	2.1	61.6	3.6	12.6	8.5	3.8
GPT-3.5	Sent	20.4	1.1	7.2	10.8	3.1	48.3	3.0	9.5	3.7	2.4
	Doc10	11.6	1.0	3.3	1.8	1.6	41.6	4.2	6.6	2.0	1.3
LLaMAX3-SFT$_1$	Sent	46.8	3.5	10.0	2.9	2.0	66.6	3.5	11.5	6.0	2.9
	Doc10	24.1	1.7	10.4	1.9	1.8	22.6	3.0	9.0	2.5	1.6
LLaMAX3-SFT$_{10}$	Doc10	37.8	2.6	7.3	1.8	1.6	64.0	4.1	8.7	4.6	2.7
Hausa											
Aya-101	Sent	56.4	2.9	9.1	4.1	3.7	61.5	2.9	17.1	10.2	5.6
	Doc10	48.5	2.9	8.3	2.3	1.9	62.3	3.2	14.6	8.4	4.5
GPT-3.5	Sent	44.3	1.4	11.3	5.2	5.8	52.4	2.5	13.2	6.8	3.7
	Doc10	23.1	1.1	7.4	2.9	2.8	52.7	4.1	9.2	4.1	2.3
LLaMAX3-SFT$_1$	Sent	62.5	3.2	9.5	4.0	3.5	58.9	2.9	9.6	4.1	3.0
	Doc10	29.0	2.3	8.8	2.3	1.9	22.9	2.6	9.0	3.0	1.9
LLaMAX3-SFT$_{10}$	Doc10	51.9	3.9	14.6	2.2	1.9	66.7	4.2	9.6	4.5	2.6
Swahili											
Aya-101	Sent	44.7	1.3	9.5	5.0	3.4	70.8	3.3	17.5	9.9	4.8
	Doc10	34.7	1.8	8.2	5.0	3.0	71.2	3.8	14.1	9.8	4.0
GPT-3.5	Sent	76.7	4.9	4.2	1.8	1.2	75.0	3.6	12.4	7.6	3.7
	Doc10	76.1	4.9	3.9	0.8	0.6	77.7	4.7	7.1	4.5	2.0
LLaMAX3-SFT$_1$	Sent	73.1	3.6	12.3	4.8	3.4	73.1	3.9	11.8	8.4	2.7
	Doc10	42.2	3.2	9.5	3.6	2.5	33.1	3.1	8.9	3.2	1.9
LLaMAX3-SFT$_{10}$	Doc10	74.4	4.4	10.5	3.6	2.2	77.8	4.6	8.9	6.2	2.7
Yoruba											
Aya-101	Sent	31.2	1.2	8.2	5.4	3.1	57.9	2.3	23.8	13.6	6.8
	Doc10	18.7	1.4	6.7	2.7	2.2	56.1	2.9	18.7	10.1	5.1
GPT-3.5	Sent	21.3	1.1	7.8	5.5	3.9	52.1	2.6	12.9	5.0	3.3
	Doc10	10.1	1.0	4.1	2.3	2.1	51.7	3.9	9.0	3.6	2.1
LLaMAX3-SFT$_1$	Sent	57.5	4.0	12.3	3.8	2.7	64.7	3.2	12.2	6.1	3.1
	Doc10	33.8	2.7	8.1	2.4	1.8	27.2	3.2	9.1	3.4	2.0
LLaMAX3-SFT$_{10}$	Doc10	52.2	4.2	10.3	2.4	1.7	71.8	4.4	10.2	5.9	2.4
Zulu											
Aya-101	Sent	58.6	2.7	13.7	6.0	3.6	67.4	2.9	19.1	12.2	8.6
	Doc10	54.9	3.1	16.1	3.9	2.7	69.0	3.3	15.7	10.7	4.5
GPT-3.5	Sent	51.1	1.5	20.9	10.0	6.1	59.5	2.6	15.9	8.8	5.7
	Doc10	29.2	1.3	10.1	4.0	3.2	61.1	4.0	10.8	5.3	2.9
LLaMAX3-SFT$_1$	Sent	67.5	3.3	12.4	5.5	3.6	70.5	3.6	12.5	7.2	3.8
	Doc10	33.7	2.4	8.4	2.7	2.2	31.5	3.1	8.6	3.4	2.1
LLaMAX3-SFT$_{10}$	Doc10	55.0	3.6	12.1	3.0	2.0	74.1	4.4	9.4	5.5	2.5

Table 22: Document-level evaluation in the Health domain, judged by GPT-4o. Compares sentence- vs. document-level outputs on Fluency (1–5 scale), Content Errors (CE), Lexical (LE), and Grammatical Cohesion Errors (GE). Best scores in bold.

Prompt 1 {system_prompt} Translate the following {source_language} text to {target_language}: Provide only the translation. {source_language} text: {{source_sentence}} {target_sentence} text: **Prompt 2** {system_prompt} Translate the following {domain} text from {source_language} to {target_language}: Provide only the translation. {source_language} document: {{source_document}} {target_language} document: **Prompt 3** {system_prompt} Please provide the {target_language} translation for the following {source_language} text:{{source_document}} Provide only the translation.
Prompt 1 {system_prompt} Translate the following {source_language} document to {target_language}: Provide only the translation. {source_language} document: {{source_document}} {target_language} document: **Prompt 2** {system_prompt} Translate the following {domain} document from {source_language} to {target_language}: Provide only the translation. {source_language} document: {{source_document}} {target_language} document: **Prompt 3** {system_prompt} Please provide the {target_language} translation for the following {source_language} document:{{source_document}} Provide only the translation.

Table 23: The task prompts used for evaluating LLMs are applied to both sentence-level and document-level translation tasks.

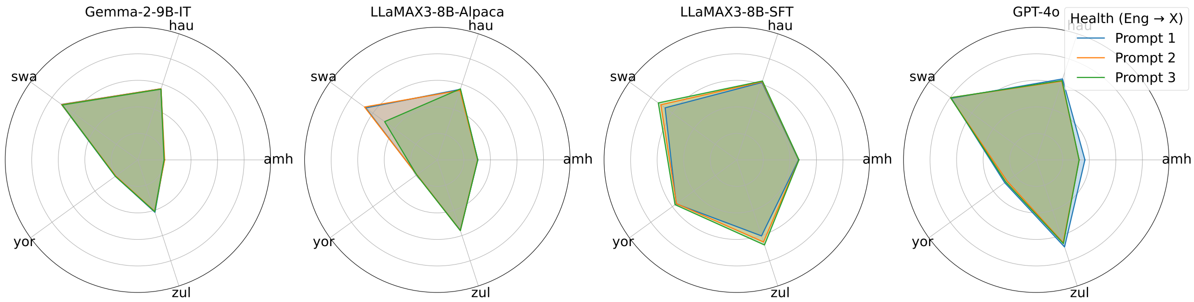

Figure 17: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into African languages

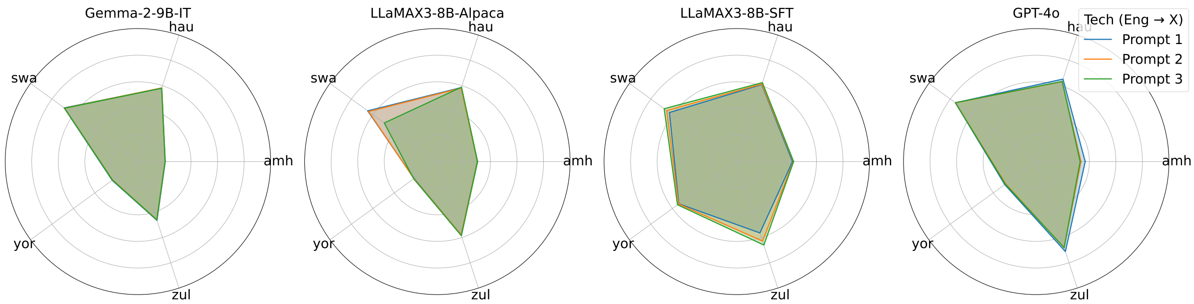

Figure 18: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into African languages

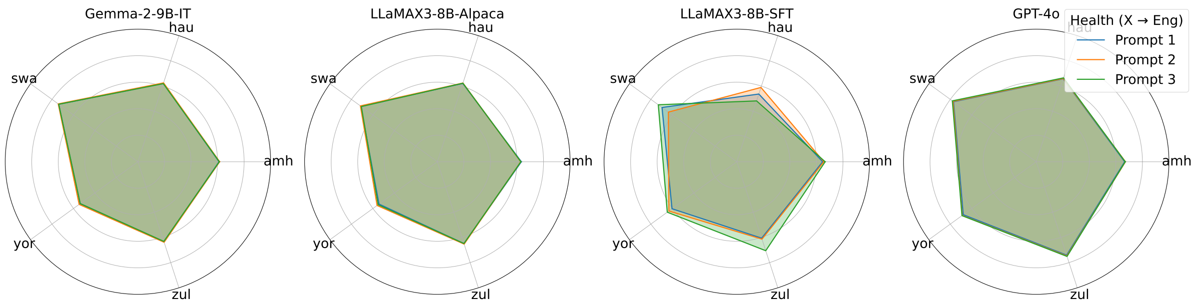

Figure 19: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into English

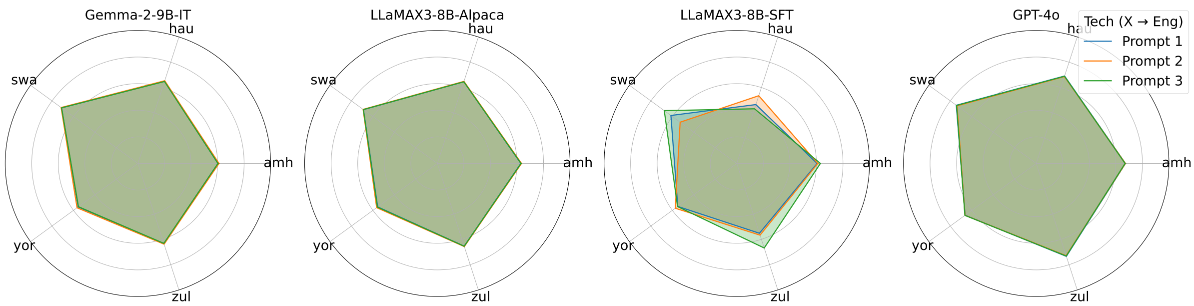

Figure 20: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into English

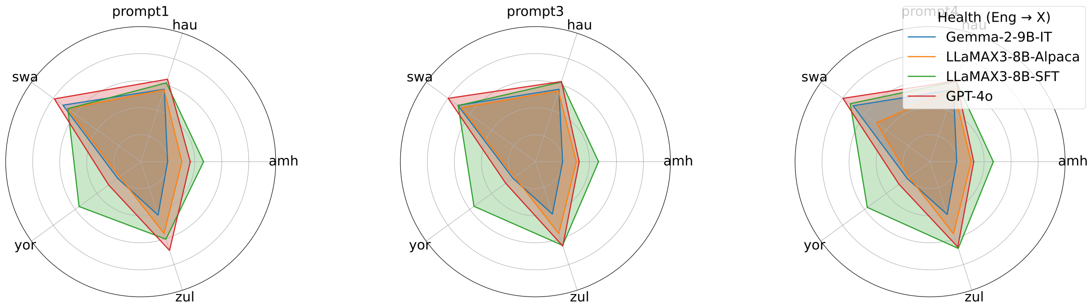

Figure 21: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into African languages

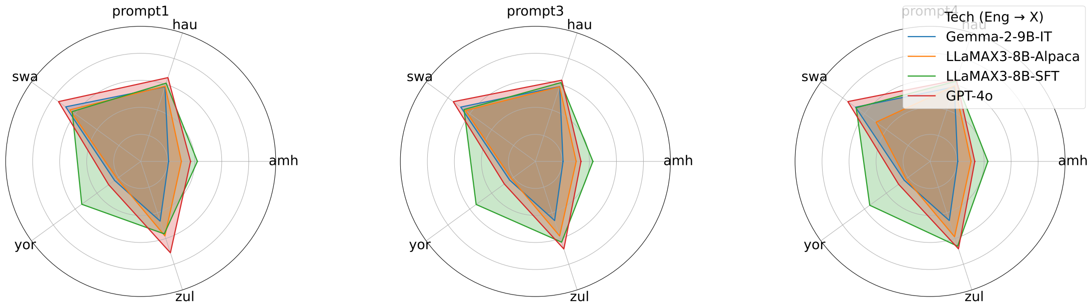

Figure 22: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into African languages

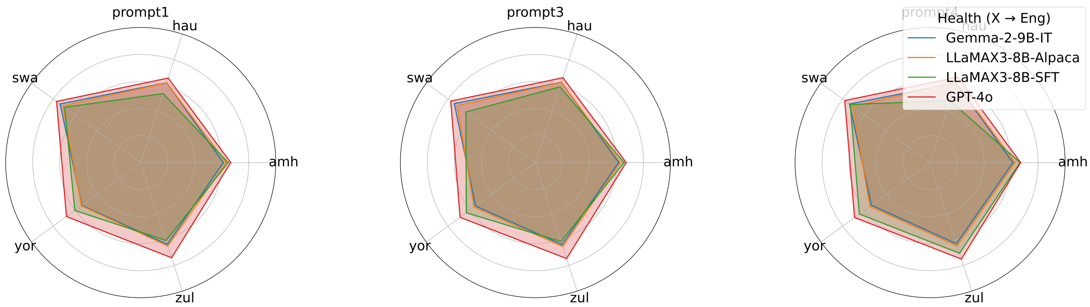

Figure 23: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into English

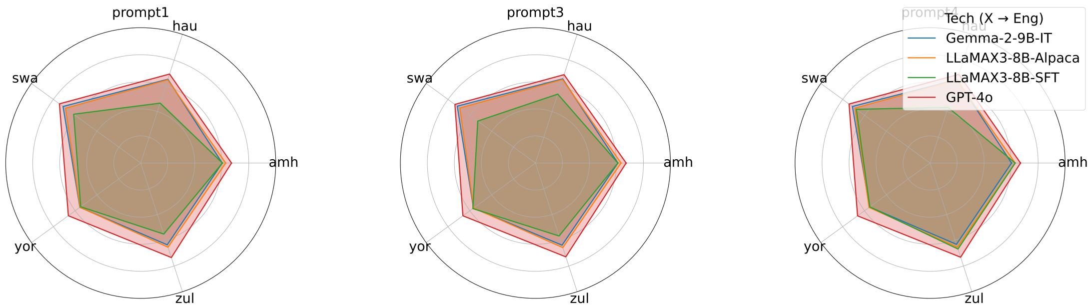

Figure 24: d-chrF scores for some LLMs for sentence-level translation using different prompts when translating into English