Fixing Rogue Memorization in Many-to-One Multilingual Translators of Extremely-Low-Resource Languages by Rephrasing Training Samples

Paulo Cavalin[1], Pedro H. Domingues[2], Claudio Pinhanez[1], and Julio Nogima[1]

[1]IBM Research, Brazil
[2]Pontíficia Universidade Católica do Rio, PUC-Rio, Brazil

Abstract

In this paper we study the fine-tuning of pre-trained large high-resource language models (LLMs) into many-to-one multilingual machine translators for extremely-low-resource languages such as endangered Indigenous languages. We explore those issues using datasets created from pseudo-parallel translations to English of *The Bible* written in 39 Brazilian Indigenous languages using mBART50 and WMT19 as pre-trained models and multiple translation metrics. We examine bilingual and multilingual models and show that, according to machine translation metrics, same-linguistic family models tend to perform best. However, we also found that many-to-one multilingual systems have a tendency to learn a "rogue" strategy of storing output strings from the training data in the LLM structure and retrieving them instead of performing actual translations. We show that rephrasing the output of the training samples seems to solve the problem.

1 Introduction

With the explosion in popularity of *Large Language Models (LLMs)*, there is currently a *de facto standard* to create *machine translators (MTs)* for low-resource languages: take one LLM pre-trained on a large corpus with self-supervised techniques and fine-tune it using a much smaller parallel downstream corpus in that language (Lee et al., 2022a). That usually results in better translation accuracy than with a model trained from scratch but with limited data (Adelani et al., 2022). For under-resourced languages, it is also common to improve translation quality even further by the use of data from multiple languages and large multilingual models (Saleh et al., 2021).

In this work we show experimentally that the reality of creating translators for extremely low-resource languages is, unfortunately, much more difficult. Notably, we show that when performing such "standard" fine-tuning of multilingual LLMs, we have to be very careful on interpreting results which rely only on average scores of reference-based translation quality metrics. We present a case where improved averaged measurements, instead of signifying better machine translators, are in fact the result of a "rogue" model which stores outputs from the training data and retrieves them as canned outputs. In other words, the fine-tuned model hallucinates translations by generating texts which are disconnected to the input: they are perfectly-memorized reproductions of training samples. Because the MT sometimes succeeds in producing a "perfect", high-score with a memorized answer, the average score of the model is inflated, higher than other approaches, although the model is, in fact, totally inappropriate for the actual task.

We present strong evidence that the memorization side-effect is a result of the inclusion of repetitions of target texts in training set, which is an artifact of many-to-one translators, since the memorization effect directly increases with the number of languages in the multilingual model. We then show that rephrasing the target text helps to avoid memorization without affecting translation quality and that is a viable solution for many other applications since now is feasible to create rephrasings of texts with LLMs.

Our findings come in a context where multilingual fine-tuning of LLMs is often used when extremely-low-resource languages are involved. In our experiments we fine-tuned pre-trained LLMs to make MTs from 39 *Brazilian Indigenous languages (BILs)* to English[1]. Those languages are typical of scenarios where there is almost no data available for training but we can find translations of official documents or religious texts to fine-tune

[1]Although the mainly spoken language of Brazil is Portuguese, in this work we focus on generating English language to avoid uncontrollable negative effects from text generation quality issues of LLMs in Portuguese.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4503–4514

existing models, that is, often with many-to-one or one-to-many mappings of the same texts to a higher-resourced language such as English.

For our experimental study, we compiled a parallel corpus with pseudo-alignments of verses from *The Bible*, and fine-tuned two commonly-used LLMs, i.e. *mBART50* and *WMT19*, relying on one bilingual and two multilingual fine-tuning strategies. The results are evaluated with three standard reference-based metrics for machine translation, i.e. the *n-gram* based BLEU score and the trained neural metrics named BLEURT and BERTScore.

To quantify the memorization behavior and to provide evidence that our many-to-one multilingual MTs became content retrievers, we implement a memorization metric based on finding the smallest distance of the generated text to target sentences in the training set, in this case the verses from The Bible in English. Our metric suggests that the problem worsens as more languages are added to the training set. We then show that, by using rephrases of the output verses from 10 different versions of Ghe Bible in English, we can build a multilingual translator for 10 BILs of the same linguistic family in which the memorization issues virtually disappear.

The main contribution of this paper is to show that automatic evaluation metrics can be deceiving in extremely-low-resource scenarios and how important it is to perform different forms of evaluation of the outputs, including qualitative inspections. This paper also shows that strictly following well-known approaches from the NLP literature, such as multilingual training, can result in models with undesired memorization effects. We also propose and demonstrate that such undesired behaviour can be fixed with rephrasings which can be easily generated with current days' LLMs.

2 Related Work

Most of the recent impressive results of NLP have been observed with English and a few other languages. Despite efforts in expanding such research for other languages (Aharoni et al., 2019; Goyal et al., 2022; NLLB Team et al., 2022; Aji et al., 2022), results have been limited, covering at best about 200 languages (NLLB Team et al., 2022), leaving behind almost 7,000 other languages. In Brazil, none of the about 200 Indigenous languages spoken in the country are covered by LLMs. Many of the Brazilian Indigenous Languages (BILs), similarly to other extremely-low-resource languages languages in the world, are endangered, facing the challenge of disappearing in just a few decades from today. Pinhanez et al. (2023) provide a good discussion on why and how AI can contribute to the survival of such languages, and at same, the benefits to AI of working with endangered languages.

One big challenge in scaling up the current AI progress to more languages is the data-hunger needs of current NLP models. Since the vast majority of the languages in the world are low-resourced or extremely-low-resourced, one hope to overcoming this data-scarcity issue is the use of self-supervised pre-trained models and *fine-tune* such models to downstream tasks such as MT. In that case, the general-purpose language knowledge of the pre-trained model is transferred and reused in a context of a much smaller training set through fine-tuning of its parameters (Lee et al., 2022a).

Another way to circumvent the lack-of-data problem is to use *multilingual models*, where data from several languages are combined (Aharoni et al., 2019; Dabre et al., 2020). The main assumption is that multilingual models can leverage the shared linguistic characteristics of related languages, expanding the utility of very small datasets. But the non-deliberated use of extra languages may worsen the performance, a phenomenon usually called *negative transfer* (Saleh et al., 2021).

A more particular approach, common in some extremely-low-resource languages such as Indigenous languages, is to rely on multi-way multilingual corpora, in either many-to-one or one-to-many directions (Dabre et al., 2019; Mueller et al., 2020). The same monolingual text, translated to many languages, is used in the training set multiple times as either source or target text (Mager et al., 2021). Typically those texts are official documents and religious texts, such as The Bible (Mayer and Cysouw, 2014; Bollmann et al., 2021; Vázquez et al., 2021; Nagoudi et al., 2021; Adelani et al., 2022). One effect of such approach is that samples appear repeatedly in the training set, either as source or target texts, what can trigger a known problem in LLMs: memorization (Carlini et al., 2023).

Memorization in LLMs consists of generating identical or nearly-identical reproduction of training samples during inference time. Notice that memorization can be beneficial for some applications, such as closed-book Q&A, and it is not necessarily a type of overfitting (Tirumala et al., 2022). But in terms of machine translation, given that it

has been shown that removing duplications is beneficial in terms of improving translation quality of MT systems (Lee et al., 2022b; Ramírez-Sánchez et al., 2020), we explore here whether the negative impact of duplications on translation quality is due to the memorization of duplicated training samples. Together with other factors such as model size and vocabulary size (Kharitonov et al., 2021), duplication of examples in the training set is known to sometimes produce memorization (Zhang et al., 2021; Carlini et al., 2023).

3 Translating Indigenous Languages

Creating machine translators and other NLP tools with extremely-low data resources is of key importance for Indigenous languages and in particular for the about 2,800 languages which are in danger of disappearing in the world (Moseley, 2010). Modern NLP technology tools may not only be a way to contribute to the vitalization of those languages but also paths to propagate the culture and support the political rights of their communities (Mager et al., 2018; Zhang et al., 2022; Liu et al., 2022; Pinhanez et al., 2023).

In this paper we explore particular issues and difficulties faced by the construction of *machine translators* for languages which are spoken by hundreds or at most thousands of people, often in semi-isolated conditions and with extremely low presence in the web, making web crawling and crowdsourcing unfeasible ways to gather data.

Our main source of data are the many translations to Indigenous languages of the The Bible. Since it is structured in numbered verses, it is relatively easy to create quasi-parallel datasets (that is, the translations may differ in style), as done in other works (Mayer and Cysouw, 2014; Adelani et al., 2022). It is possible to use The Bible as data for many of the BILs, since it has been translated into many of those languages by Christian churches (Franchetto, 2008). We understand that there are important ethical, historical, and cultural issues around the use of The Bible as a source of data for Indigenous languages, which we discuss in detail in Section 9 at the end of the paper.

3.1 Brazilian Indigenous Languages

Brazil was home to about 270 Indigenous languages according to the Census of 2010, the last comprehensive assessment of linguistic diversity in Brazil (IBGE, 2010)[2]. Those languages were spoken by approximately 800 thousand people (IBGE, 2010), half of them living in Indigenous lands. Storto (2019) provides a good overview of the history, structure, and characteristics of BILs. Almost all of those languages are considered endangered (Moseley, 2010). We adopted here the Indigenous language classification, nomenclature, and data from the 2010 Brazilian Census by IBGE (IBGE, 2010) and language acronyms according to ISO 639-3.

Indigenous Languages					# Aligned Sentences		
Name	Acron	Branch	Family	Speakers	Train	Test	Total
Bororó	bor	Macro-Jê	Bororó	1035	1861	202	2063
Apinayé	apn	Macro-Jê	Jê	1386	877	75	952
Kaingáng	kgp	Macro-Jê	Jê	19905	5695	917	6612
Kayapó	txu	Macro-Jê	Jê	5520	2669	510	3179
Xavánte	xav	Macro-Jê	Jê	11733	1275	342	1617
Karajá	kpj	Macro-Jê	Karajá	3119	2828	333	3161
Maxakali	mbl	Macro-Jê	Maxakali	1024	5566	905	6471
Rikbaktsa	rkb	Macro-Jê	Rikbaktsa	10	3560	710	4270
Mawé	maw	Tupi	Mawé	8103	6381	970	7351
Mundurukú	myu	Tupi	Mundurukú	3563	3110	190	3300
Guajajára	gub	Tupi	Tupi-Guarani	8269	4956	934	5890
Guaraní (West Bolivia)	gnw	Tupi	Tupi-Guarani	NA	5263	970	6233
Guaraní (East Bolivia)	gui	Tupi	Tupi-Guarani	NA	5263	924	6187
Guaraní Kaiowá	kgk	Tupi	Tupi-Guarani	24368	3034	479	3513
Guaraní Mbyá	gun	Tupi	Tupi-Guarani	3248	6340	970	7310
Guaraní (Paraguay)	gug	Tupi	Tupi-Guarani	NA	5196	970	6166
Ka'apor	urb	Tupi	Tupi-Guarani	1241	3380	436	3816
Kaiabi	kyz	Tupi	Tupi-Guarani	673	2187	280	2467
Nheengatu (LGA)	yrl	Tupi	Tupi-Guarani	3771	5035	691	5726
Tenharim	pah	Tupi	Tupi-Guarani	32	3215	844	4059
Jamamadí-Kanamanti	jaa	no branch	Arawá	217	4759	715	5474
Kulina Madijá	cul	no branch	Arawá	3043	4319	697	5016
Paumarí	pad	no branch	Arawá	166	3653	372	4025
Apurinã	apu	no branch	Aruak	824	6329	970	7299
Palikur	plu	no branch	Aruak	925	6137	904	7041
Paresí	pab	no branch	Aruak	122	6381	970	7351
Teréna	ter	no branch	Aruak	6314	6381	970	7351
Wapixána	wap	no branch	Aruak	3154	5081	853	5934
Kadiwéu	kbc	no branch	Guaikurú	649	4523	793	5316
Apalaí	apy	no branch	Karib	252	5548	970	6518
Bakairí	bkq	no branch	Karib	173	4000	317	4317
Hixkaryána	hix	no branch	Karib	52	4270	472	4742
Makuxi	mbc	no branch	Karib	4675	4900	940	5840
Nadëb	mbj	no branch	Makú	326	5213	811	6024
Nambikwára	nab	no branch	Nambikwára	951	2774	844	3618
Kashinawá (Peru)	cbs	no branch	Pano-Tacanan	3588	2136	130	2266
Tukano	tuo	no branch	Tukano	4412	3750	846	4596
Yanomámi	guu	no branch	Yanomámi	12301	1283	196	1479
Tikúna	tca	no branch	no family	30057	3097	386	3483
TOTAL	39	3	16	169201	162225	25808	188033

Table 1: Indigenous languages and corresponding size of the datasets used in the study. Language name, branch, family, and number of speakers (considers only who speak the language at home in an Indigenous land in Brazil) according to the table 1.13 of the Indigenous data of the Brazilian census of 2010 (IBGE, 2010).

3.2 The *The Bibles* Dataset

We consider in this work 39 Indigenous languages spoken in Brazil of which we were able to find translations for the *New Testament* of The Bible, a book which comprises about 7,000 verses in its English versions. Table 1 lists the 39 Indigenous languages used in this work, comprising 36 spoken primarily in Brazil and 3 other Guaraní languages used mostly in Paraguay and Bolivia but also spoken in some areas in Brazil.

[2]There is some discussion about the accuracy of those numbers, see (Franchetto, 2020; Storto, 2019).

This dataset, henceforth called *The Bibles*, was obtained in its majority from the *ebible* website[3]. A few other languages of our dataset were obtained from the *YouVersion* online platform[4]. The Bibles dataset consists of 188,033 parallel verses from the New Testament in English and in the 39 Indigenous languages listed in table 1. The parallelism among translations of the same verse, performed by the authors, has a reasonable quality although we are aware that the source of the translations comes from different versions of The Bible in several languages and has diverse narrative styles.

To avoid cross-contamination in the decoder and allow us to study memorization issues without data leakage between the training and test sets, we used the *Matthew* chapter from the New Testament as the source of test set and the remainder as the training set. We are aware that there are some similarity among verses could happen between the book of Matthew and the other synoptic gospels, but we see the existence of some similarity as positive, since in most practical multi-language training setup there will be some level of similar sentences.

The resulting training sets for the fine-tuning strategies vary in size (see Table 1). The bilingual models were fine-tuned, on average, with 4,160 pairs of sentences. To fine-tune Tupi-Guarani family models, the training set comprised 43,869 sentence pairs, and for the all languages models, 162,225 training pairs were used for fine-tuning.

4 Experimental Study

In this section we describe the methodology and results of fine-tuning LLMs into machine translators for the 39 BILs presented in Table 1. The translation direction is always from a BIL to English.

4.1 Methodology

We employed two distinct LLMs in the fine-tuning process. The first LLM is **mBART50** (Tang et al., 2020), which is an extended version of *mBART* (Liu et al., 2020). With 680M parameters, mBART50 is pre-trained with masked language modeling on 203M sentences. This model is a common choice for training multilingual MTs for low-resource languages (Lee et al., 2022a; Chen and Abdul-Mageed, 2022).

The second LLM is **WMT19** (Ng et al., 2019), which is a 315M-parameter German-to-English machine translator pre-trained on about 28M pairs of translated sentences and more than 500M back-translated sentences. This model is not as popular as mBART in the MT literature but was chosen because of its smaller size and good performance in decoding English texts.

We downloaded the models from *HuggingFace*[5][6]. Both were fine-tuned with a learning rate of 2^{-5} and a batch size of 16. mBART50 was trained for 4 epochs and WMT19 for 5.

Our evaluation considers three different fine-tuning strategies, that result in three different types of models, and two groups of test sets. First, we consider **bilingual (BL)** models, created by fine-tuning each of the LLMs on source-to-target pairs exclusively from each of the BILs on Table 1, yielding 39 different bilingual models. Second, we consider the extreme multilingual case where we fine-tune each of the LLMs with **all languages (AL)** at once. Third, we created in-between multilingual solutions, the **Tupi-family (TF)** models, where the training set comprises only the 10 languages belonging to the Tupi-Guarani family: Guaraní of Paraguay, Bolivia (2), Kayowá, and Mbyá; Guajajára, Ka'apor, Kaiabi, Nheengatu, and Tenharim.

With the goal of measuring the impact of the previously-mentioned models, we defined two distinct sets of experiments. The first one, **BL39 vs AL39**, considering all 39 BILs in a single test set, so that we can compare the impact of a bilingual model against a multilingual one trained with all the languages, i.e. BL vs AL. The second set is **BL10 vs TF10 vs AL10**, where we focus in comparing the BL models not only against AL but also against TF which is more targeted at Tupi-family languages. For the latter, the test set contains only the 10 languages used to train the TF models.

We used three metrics to evaluate the results, combining the traditional **BLEU** score (Papineni et al., 2002) with more recent neural-based metrics which are considered to be more robust and better correlated with human scores (Freitag et al., 2022). For BLEU, we compute the average of sentence-level BLEU scores[7] computed with the SacreBLEU Python package (Post, 2018). For the

[3]https://ebible.org/download.php
[4]https://www.bible.com/en-GB/
[5]https://huggingface.co/facebook/mbart-large-50
[6]https://huggingface.co/facebook/wmt19-de-en
[7]We are aware that corpus-level BLEU is usually used to assess system-level scores, but this choice allows us to compute standard deviations, which are important indicators of memorization issues, as we discuss later. Also, that allowed us compare all metrics with the same methodology, which is difficult to do with corpus-level BLEU.

	mBART						WMT19					
	BLEURT	std	**BERTScore**	std	**BLEU**	std	**BLEURT**	std	**BERTScore**	std	**BLEU**	std
BL10	0.368	0.079	0.876	0.025	9.520	8.554	0.346	0.068	0.858	0.031	6.265	6.273
TF10	**0.385**	0.116	**0.879**	0.036	**12.325**	16.218	**0.346**	0.077	**0.861**	0.030	**6.626**	8.188
AL10	0.323	0.095	0.859	0.029	6.393	11.345	0.314	0.083	0.854	0.027	5.312	10.061
BL39	**0.343**	0.070	**0.867**	0.025	**7.152**	6.653	**0.327**	0.063	0.845	0.032	**4.618**	4.901
AL39	0.316	0.087	0.856	0.027	5.525	9.963	0.307	0.073	**0.851**	0.024	4.422	8.171
TF10 to BL10	**5%**	47%	**0%**	43%	29%	90%	**0%**	13%	**0%**	-2%	**6%**	31%
AL to BL10	-12%	20%	-2%	16%	-33%	33%	-9%	21%	0%	-14%	-15%	60%
AL39 to BL39	-8%	24%	-1%	9%	-23%	50%	-6%	15%	**1%**	-24%	-4%	67%

Table 2: Summary of the average and standard deviation scores in different test sets of languages for the fine-tuning of mBART50 and WMT19 to different models for all metrics.

MODEL	LANGUAGE	ORIGINAL	EXPECTED	GENERATED	GENERATED VERSE
mBART50-AL39	Apalaí	oty se hma tykase jezu eya. moxiã umũkuru tõ asakoro tuisame tyriko toto, amaro, tuisa konõtome toehse awahtao, toiro ãpotunuru wino, toiro opozery wino, enara, tykase nohpo eya.	he said to her, what do you want she replied, permit these two sons of mine to sit, one at your right hand and one at your left, in your kingdom.	jesus said to him, foxes have dens and the birds in the sky have nests, but the son of man has no place to lay his head.	Luke 9:58
mBART50-AL39	Mawé	mi'i tupono are'e pywo ti aru meiũran wuat'i yi wato hap ok tã aikotã haryporia minug waku uhetiat hap ta'atuhenoi henoi e. wuat'i ywania ti aru ikuap wakuap uhetiat iminug hap e.	i tell you the truth, wherever this gospel is proclaimed in the whole world, what she has done will also be told in memory of her.	i tell you the truth, this generation will not pass away until all these things take place.	Mark 13:30
mBART50-AL40	Nheengatu	asui aintaurasu iakanga pratu upe aintauxari arama ixupe ae kua kunyamuku, asui urasu ae imaya supe.	his head was brought on a platter and given to the girl, and she brought it to her mother.	so they seized him, killed him, and threw his body out of the vineyard.	Mark 12:8
mBART50-AL39	Apurinã	xirataãtxi apisa ikinimane nipokota, iãta itxa. kiki, sito ĩtaniriuata. ikara atokokanera itxa noee maporo ãki iereẽtini õti.	you will be hated by everyone because of my name. but the one who endures to the end will be saved.	for in those days before the flood, people were eating and drinking, marrying and giving in marriage, until the day noah entered the ark.	Mark 13:13
mBART50-TF10	Guaraní Mbyá	ejopy ha'vy ndeperáta, tereo, mba'eta takykue py ou va'ekue pe avi ame'ẽxe pẽvy ame'ẽa rami.	take what is yours and go. i want to give to this last man the same as i gave to you.	i say this to your shame is there no one among you wise enough to settle disputes between fellow christians	1 Corinthians 6:5
mBART50-TF10	Paresí	hatyo xowaka bolokonai hatyo haliyita nisakoatita.	a large herd of pigs was feeding some distance from them.	i wish that you would be patient with me in a little foolishness but indeed you are being patient with me	2 Corinthians 11:1
mBART50-TF10	Wapixána	sariapa upishaan ĩdyaun. ukian, kanom dikin wuru'u, na'iki kanom uu wuru'u saada-kariwaiz puraata ida'a	jesus said to them, whose image is this, and whose inscription	and the stars in the sky fell to the earth like a fig tree dropping its unripe figs when shaken by a fierce wind	Revelation 6:13
mBART50-TF10	Kaingáng	ã ra jesus tóg mẽ kỹ ag mỹ ãjag tỹ ne jé fi mỹ kaga han nẽ he mũ. inh jykre ki króm fi tóg.	but jesus rebuked him silence come out of him	when jesus learned of this, he said to them, why are you bothering this woman she has done a good service for me.	Mark 1:25

Table 3: Examples of outputs of mBART50-AL39 and mBART50-TF10 showing test samples where an almost literal version of a different Bible verse is generated, including the reference to the generated verse.

neural-based metrics, we consider **BLEURT** (Sellam et al., 2020) and **BERTScore** (Zhang et al., 2020).

4.2 Study Results

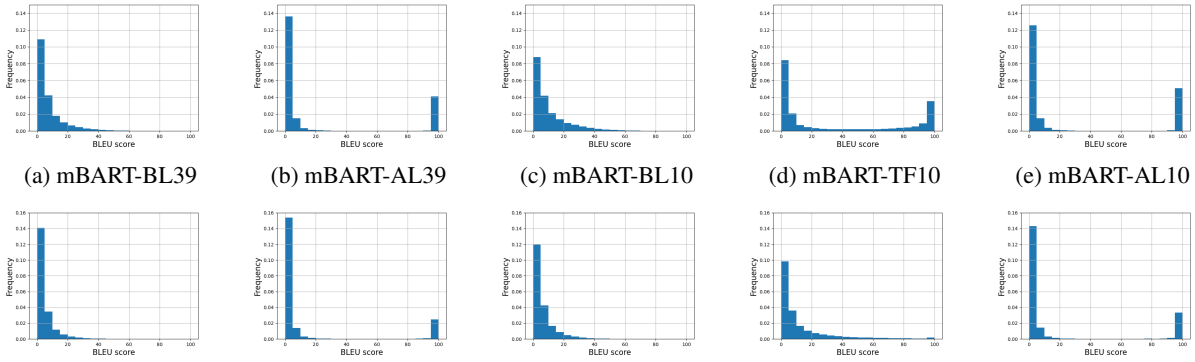
Table 2 shows the average and standard deviation scores in the different test sets of the languages for the fine-tuning of mBART50 and WMT19 to different fine-tuning strategies, for all three metrics.

Overall, mBART50 seems to perform slightly better than WMT19 but with higher standard deviation, and the three metrics afford similar results comparatively. The three models yield similar results but the bilingual ones present smaller standard deviation when all languages are considered. The TF10 model seems to perform slightly better than the bilingual models but with a higher standard deviation. As we show in the next section, high standard deviations are, in fact, a symptom that the the translation model has started to perform what we call *rogue memorization*, that is, it has become a retriever of the contents of the training set.

A complementary view on the effects of the high standard deviations can be observed with the distribution of BLEU scores computed with samples from the training set, as show in Figure 1. From these distributions we can clearly observe that the BL39 and BL10 models (both for mBART50 and WMT19) usually result in right-skewed normal distributions, while for AL39 and AL10 the shapes of the distributions resemble more binomial distributions. We can also observe some shift to a binomial distribution in the TF10 model in the case of mBART50.

5 Memorization Issues

By examining the actual output of the mBART50-AL39 model, we saw cases where the output trans-



(a) mBART-BL39 (b) mBART-AL39 (c) mBART-BL10 (d) mBART-TF10 (e) mBART-AL10

(f) WMT19-BL39 (g) WMT19-AL39 (h) WMT19-BL10 (i) WMT19-TF10 (j) WMT19-AL10

Figure 1: Distribution of BLEU scores of samples from the training set for mBART50 (top) and WMT19 (bottom).

MODEL	mBART50			WMT19		
	distinct outputs	distinct to expected	inputs to distinct output	distinct outputs	distinct to expected	inputs to distinct output
AL39	3951	15%	7 : 1	4494	17%	6 : 1
BL39	24351	94%	1: 1	24001	93%	1 : 1
expected	25808	100%	1 : 1	25808	100%	1 : 1
AL10	2452	33%	3 : 1	2506	33%	3 : 1
TF10	6405	85%	1: 1	7219	96%	1 : 1
BL10	7057	94%	1 : 1	7297	97%	1 : 1
expected	7498	100%	1 : 1	7498	100%	1 : 1

Table 4: Number of distinct outputs and proportions of input to distinct outputs for the different models.

lation was not only an incorrect translation, but a literal verse of The Bible, easily found through a search procedure in the Internet. We saw similar cases in the mBART50-TF10 model and in the WMT19-AL39. Examples of some mBART50 cases are shown in Table 3. As shown in the table, instead of attempting to translate the input string, the fine-tuned systems produce literal verses totally unrelated to the expected output.

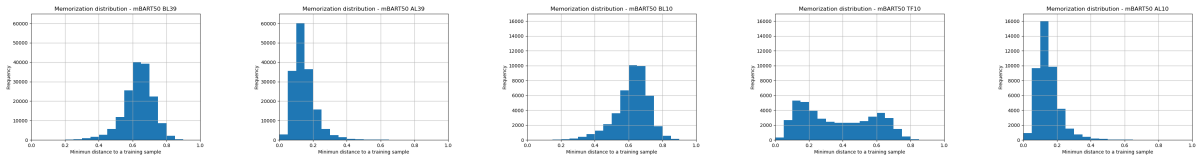
Another issue we could see in this qualitative analysis was that there were many exact copies of output sentences. Translation from different languages of sentences with the same meaning and structure is expected to generate very diverse outputs. We saw the occurrence of repetitions as evidence of rogue memorization. The extent of the problem is evidenced by the results of the analysis presented in Table 4, which shows for each model the number of distinct outputs. For instance, in the case of the experiments with the 39 BILs (AL39 and BL39), which comprised 25,808 different inputs from the samples, we would expect about the same number of distinct outputs, as discussed above. However, in the mBART50-AL39 model, only 3,951 distinct outputs were produced, about 15% of what would be reasonable to expect, configuring a ratio of about 7 inputs mapping to exactly the same output (7 : 1). The WMT19-AL39 model yielded a similar ratio (6 : 1). Notice that the issue is not present in the bilingual models but slightly visible in the mBART50-TF10 model.

To more precisely quantify the extent of the memorization by the models, we implemented a metric to quantify memorization, based on computing the smallest distance of a generated output to a sample from the training. The assumption is that, when memorization occurs, the decoder reproduces a training sample, which in this case is one verse of The Bible, even when the translation-quality metric presents a low value. For that, we compute the Euclidean distance between the output and the references of all training samples, and return the smallest distance as the measured value. The distance is computed on a semantic representation of the texts, using the *Sentence Transformers* Python library[8], using the *all-MiniLM-L6-v2* model[9]. With that we are computing even soft memorization effects since we are not relying on a exact match as usually employed in previous works (Tirumala et al., 2022; Carlini et al., 2023).

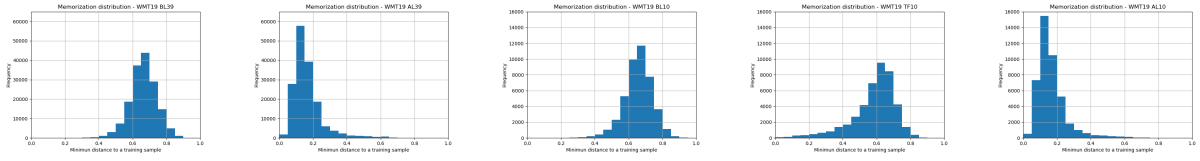
In Figure 2 we present the distribution of our memorization metric considering a training examples as inputs. In those plots, the further the distribution is shifted to the left, the stronger is the production of memorized literal verses. The results with *-AL39 (mBART50-AL39 and WMT19-AL39) and *-AL10 evaluations, compared with *-BL39 and *-BL10, clearly show that the multilingual model strongly memorized the training set.

[8]https://huggingface.co/sentence-transformers

[9]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

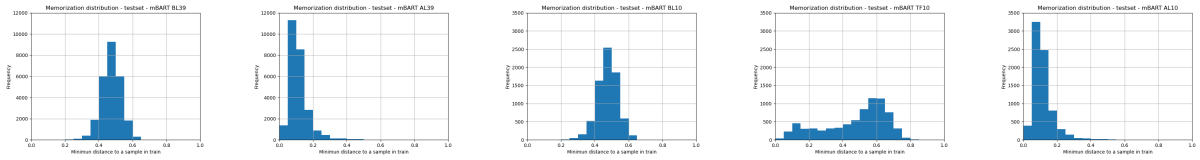


(a) mBART-BL39 (b) mBART-AL39 (c) mBART-BL10 (d) mBART-TF10 (e) mBART-AL10

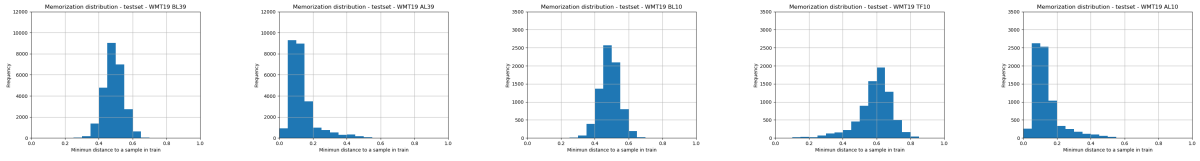


(f) WMT19-BL39 (g) WMT19-AL39 (h) WMT19-BL10 (i) WMT19-TF10 (j) WMT19-AL10

Figure 2: Histograms with the frequency of the minimum distance of a generated translations of training samples to examples in the training set.



(a) mBART-BL39 (b) mBART-AL39 (c) mBART-BL10 (d) mBART-TF10 (e) mBART-AL10



(f) WMT19-BL39 (g) WMT19-AL39 (h) WMT19-BL10 (i) WMT19-TF10 (j) WMT19-AL10

Figure 3: Histograms with the frequency of the minimum distance of a generated translations of test samples to examples in the training set.

With the bilingual models, the mean minimum distance ranges from 0.62 with mBART50-BL10 to 0.66 with WMT19-BL39, while with the all language models, the mean distance falls to a range between 0.14 with mBART50-AL10 to 0.16 with WMT19-AL10. With TF10 models, the memorization effect gradually increases with the number of languages in the training set, since this model presents mean values of 0.58 and 0.36 for WMT19-TF10 and mBART-TF10, respectively. Interestingly, with the latter we also observe a slightly bimodal distribution, indicating that the peak of the distribution is in the middle of a shifting process between the mBART50-BL10 and mBART50-AL10.

We conducted a similar evaluation considering the outputs generated from inputs of the test set. The results are displayed in Figure 3. Surprisingly, we observe quite similar results to what we observed with the training set. That is, the results with the all languages models present strong evidence of producing memorized verses, with mean minimum distances to an output from the training set ranging from 0.14 with mBART50-AL39 to 0.17 with WMT19-AL39, while the bilingual models presents a much higher range of mean values, from 0.62 with mBART50-BL39 to 0.64 with WMT19-BL39. Once again, both mBART50-TF10 and WMT19-TF10 present some in-between to-memorization transition, with mean values of 0.59 and 0.47, respectively. What is impressive, from these results on the test set, is that it seems that the memorization occurs in the decoder side of the Transformer architecture. And when the models are doing heavy memorization, there is some disconnection to the encoders. That is, *no matter the input sentence, the model simply generates one of the sentences it had memorized.*

6 Using Rephrasings to Overcome Memorization Issues

In this section we investigate whether the use of rephrasings instead of the repeated outputs can al-

leviate the memorization issue. As we saw, the inclusion of repeated outputs results in models that memorize those outputs as the number of repetitions increase, i.e. as the number of languages in the multilingual training set increase. Thus, by using outputs that are similar in semantics but which are not exact duplicates, we evaluate the impact on MT quality metrics and on memorization.

For this investigation we rely on rephrasings of The Bible in English written by different authors. Although we could have used an LLM such as *ChatGPT* and the like to generate as much rephrasings as possible, we relied on different versions of the Bible translated only by humans to avoid errors which could be introduced by LLMs. Given that translations of The Bible are not so abundant, we were able to collect 10 rephrasings which were then used to evaluate the impact on the mBART50-TF10 model, which is based on 10 different languages.

The results were very positive. The scores (and respective standard deviations) achieved by this model are of 0.408 ($\pm$0.085), 0.887 ($\pm$0.029), and 10.432 ($\pm$10.123), for BLEURT, BERTScore, and BLEU, respectively. That represent gains of 6% and 1% with BLEURT and BERTScore (compared with 0.385 and 0.879, respectively, without rephrasing), and a loss of 15% with BLEU (12.325 without rephrasing). Notice that, despite the loss in BLEU, we see a decrease of 38% in the standard deviation with that metric (from 16.218 to 10.432), and of 27% and 20% with BLEURT and BERTScore (from 0.116 and 0.036, respectively, to 0.085 and 0.029). We strongly believe that those decreases in standard deviations are directly related to the decrease of memorization, as we show next.

To show the reduction in memorization, we present in Figure 4 the plots of the memorization-effect metric comparing the outputs generated by the mBART50-TF10r model and the previous mBART50-TF10, considering both training and test samples as input sentences. Considering the test samples, as shown in Figure 4d, this evaluation resulted in a mean distance of 0.61 from a generated output to a training sample. The same evaluation considering the non-rephrased related model, i.e. mBART50-TF10, presented in Figure 4c, resulted in a mean distance of 0.47. With the outputs generated from training inputs, the difference is even more drastic. The mean distance for the non-rephrased model is of 0.36 (fig. 4a), while the model trained on rephrased sentences presents a mean distance of 0.90 (fig. 4b). We can see in the

(a) TRAINING samples, NO REPHRASING

(b) TRAINING samples, WITH REPHRASING

(c) TEST samples, NO REPHRASING

(d) TEST samples, WITH REPHRASING

Figure 4: Comparison of the histograms of frequency of the minimum distance of a translation generated with mBART TF10 of test samples to examples in the training set without (left) and with (right) rephrasing.

latter a very clear peak towards the right side of the plot, evidence that memorization is not happening.

Another evidence of decrease in memorization was obtained by computing the number of distinct outputs using the same methodology used to produce table 4. The mBART50-TF10 model produced 6,405 distinct outputs, about 85% of the ideal number of 7,498, while the model using rephrasings mBART50-TF10r generated 7464 distinct outputs, about 99.5% of maximum possible, and better than any other model we tested.

7 Final Discussion

The first main contribution of this work is to present a clear case of fine-tuning LLMs to many-to-one translators where high scores in metrics such as BLEU can be misleading, since high values come, in fact, from a rogue strategy of retrieving verses stored from the training set. We also show that common symptoms of this problem are high standard deviations of the metrics (which, therefore, should be always reported), bimodal distributions of results, and low numbers of distinct outputs. We also present a method to compute how close the output is to translations identical to the output texts of the training sets which can also be used to identify the undesirable fine-tuning of translators into rogue retrievers of stored training data.

The second contribution is to present a case where an LLM, solely due to fine-tuning, was able to memorize complex parts of the outputs of the

training set and reconstruct them with high accuracy. Although memorization has been seen in the training of LLMs, we show here a case where there is reasonable evidence that it happened during the fine-tuning process with very low amounts of data.

The third contribution is a method to reduce the memorization effect relying on rephasings of training samples. Using 10 different versions of The Bible in English, instead of identical output verses, we show that the model presents very low memorization levels together with improved translation quality and more diversity of outputs. Although we did this study with real rephrasings, we notice that LLM-generated rephrasings is a viable alternative when rephrasings are not available. Evaluating how successful the use of commercial-grade, readily available LLMs is to avoid memorization issues in many-to-one scenarios is an important part of our future work.

We also wonder if similar memorization issues may be present in some of previous research on multilingual translators for low-resource languages and, if so, whether they have compromised their findings. The belief that multilingual models are usually better than bilingual ones seems to be based on a reliance on averages of sample-level translation quality metrics which, as shown here, may hide the use of incorrect strategies. It seems essential not only that research in this area reports the standard deviation of all results but also that efforts are made to detect potential disguise of memorization strategies when fine-tuning LLMs, possibly by using some of the methods suggested in this paper.

Lastly, we believe the issues described in this paper are not exclusive to Bibles and multilingual translations. In tasks such as fine-tuning conversation LLMs using question answering data from traditional chatbots, when several questions are mapped to the same answer, we might face similar issues. Investigating that issue is outside the scope of this work but it is an interesting future direction.

8 Limitations

The main limitation of this work is the use of The Bible as the main source of data. Not only it is limiting in terms of domain, it is also likely that this type of data was used in the pre-training of LLMs. Also, The Bible itself contains intrinsic repetitions of texts which can contribute to memorization.

Another limitation is the lack of a more in-depth analysis of the loss curves in the fine-tuning of the LLMs and not using the memorization-effect metric as a way to mitigate the memorization effects. That is, our proposed metric could be used as an alternative to model selection or early stopping. With that, we believe we could reduce the negative effects of memorization but the resulting impact on translation accuracy is unknown.

9 Ethical Considerations

It is essential to consider the ethical aspects of the goals and methods of any work with Indigenous languages. First, we abide to the belief that the decision of whether to create or not a MT for an Indigenous language has to be done by the people who speak the language, fully informed and, whenever possible, as a participant of the process (Mihesuah, 1993; Sahota, 2007; Straits et al., 2012).

Moreover, we understand the complex political and ideological choices involved in the process of language vitalization (McCarty, 2008; McCarty et al., 2009; McCarty, 2011; Shulist and Granadillo, 2022) and the use technologies to support it (Harding et al., 2012; Liu et al., 2022). This was made clear by the Indigenous communities in the *Los Pinos Declaration*[10], which enshrines that language-related efforts have be done by and with Indigenous peoples: *"Nothing for us without us."*

However, the main goal of the research described in this paper is to determine technically feasible paths to construct MTs given the restrictions imposed by extremely-low-resource languages to current technologies. Therefore, even if the results of our work were extremely positive, we would refrain to make the MTs available without express consent of the corresponding community.

We are also aware that one of the unfortunate aspects of past and present colonial history of Indigenous peoples is related to different forms of Christianism. As a consequence, the Bible is one of the most commonly found document translated to several of those languages, by Jesuits in the early days of colonization and in the last 100 years often by Evangelical churches (Franchetto, 2008). As such, the translations of The Bible are often associated to different forms of cultural abuse and violence and to the establishment of orthographies of domination (Franchetto, 2008).

However, the reality is that such texts are one of the few available sources of parallel multilingual

[10]https://en.unesco.org/sites/default/files/los_pinos _declaration_170720_en.pdf.

datasets for many Indigenous languages. We thus take the use of The Bible in this work as an "exceptional" first step, as a sort of "toxic" data which should not be used, in principle, for any actually deployed system unless with explicit agreement of the Indigenous community. Nevertheless we believe such kind of data can be used carefully for in-laboratory technical experiments in well-contained contexts such as the study described in this paper. To mitigate some of those risks, we implemented in this study some of the protocols suggested in (Pinhanez et al., 2023), including the adoption of containment procedures.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 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 (NAACL'24)*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'19)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (NAACL'22)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard. 2021. Moses and the character-based random babbling baseline: CoAStaL at AmericasNLP 2021 shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 248–254, Online. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. *arXiv preprint: 2202.07646*.

Wei-Rui Chen and Muhammad Abdul-Mageed. 2022. Improving neural machine translation of indigenous languages with multilingual transfer learning. *arXiv preprint: 2205.06993*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, pages 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Bruna Franchetto. 2008. The war of the alphabets: indigenous peoples between the oral and the written. *Mana*, 4:31–59.

Bruna Franchetto. 2020. Língua (s): cosmopolíticas, micropolíticas, macropolíticas. *Campos-Revista de Antropologia*, 21(1):21–36.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT22)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Anna Harding, Barbara Harper, Dave Stone, Catherine O'Neill, Patricia Berger, Stuart Harris, and Jamie Donatuto. 2012. Conducting research with tribal communities: Sovereignty, ethics, and data-sharing issues. *Environmental health perspectives*, 120(1):6–10.

IBGE. 2010. Censo demográfico 2010. Accessed on 2022-12-30.

Eugene Kharitonov, Marco Baroni, and Dieuwke Hupkes. 2021. How bpe affects memorization in transformers. *arXiv preprint: 2110.02782*.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022a. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics - Findings (ACL'22)*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Teresa L McCarty. 2008. Language education planning and policies by and for indigenous peoples. *Encyclopedia of language and education*, 1:137–150.

Teresa L McCarty. 2011. *Ethnography and language policy*, volume 39. Routledge New York.

Teresa L McCarty, Mary Eunice Romero-Little, Larisa Warhol, and Ofelia Zepeda. 2009. Indigenous youth as language policy makers. *Journal of Language, Identity, and Education*, 8(5):291–306.

Devon A Mihesuah. 1993. Suggested guidelines for institutions with scholars who conduct research on american indians. *American Indian Culture and Research Journal*, 17(3):131–139.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*, pages 3710–3718, Marseille, France. European Language Resources Association.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. IndT5: A text-to-text transformer for 10 indigenous languages. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 265–271, Online. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. 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. *arXiv preprint: 2207.04672*.

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 (ACL'02)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

C. Pinhanez, P. Cavalin, M. Vasconcelos, and J. Nogima. 2023. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In *Proc. of IJCAI'23*, Macao, China.

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.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Puneet Chawla Sahota. 2007. Research regulation in american indian/alaska native communities: Policy and practice considerations. Technical report, National Congress of American Indians Policy Research Center.

Fahimeh Saleh, Wray Buntine, Gholamreza Haffari, and Lan Du. 2021. Multilingual neural machine translation: Can linguistic hierarchies help? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 1313–1330, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 48th Annual Conference of the Association of Computational Linguistics (ACL'20)*. Association for Computational Linguistics.

Sarah Shulist and Tania Granadilllo. 2022. Language ideology planning as central to successful revitalization projects. *Language Documentation & Conservation*, 16.

Luciana Raccanello Storto. 2019. *Línguas indígenas: tradição, universais e diversidade*. Mercado de Letras.

K.J.E. Straits, D.M. Bird, E. Tsinajinnie, J. Espinoza, J. Goodkind, O. Spencer, and T.G.P. Workgroup. 2012. Guiding principles for engaging in research with native american communities. *UNM Center for Rural and Community Behavioral Health & Albuquerque Area Southwest Tribal Epidemiology Center*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint: 2008.00401*.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *arXiv preprint: 2205.10770*.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2021. Counterfactual memorization in neural language models. *arXiv preprint: 2112.12938*.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proc. of the 2020 International Conference on Learning Representations (ICLR'20)*.