Quantifying the Ethical Dilemma of Using Culturally Toxic Training Data in AI Tools for Indigenous Languages


Pedro Domingues


Claudio Pinhanez


Paulo Cavalin


Julio Nogima

pcavalin@br.ibm.com
csantosp@br.ibm.com

IBM Research

PROINDL

AI technologies to strengthen Indigenous languages in Brazil

Claudio Pinhanez

Paulo Cavalin

Julio Nogima

Isabel Gonçalves

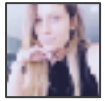
IBM Research, Brazil

Luciana Storto

Thomas Finbow

Alex Cobbinah

Sarajane Peres



Nicole Dalmiglio

Majoi Gangora

University of São Paulo

Center for Artificial Intelligence

Multidisciplinary Team

- 2 research scientists, 1 software engineer and 1 doctoral intern from IBM Research, Brazil
- 3 professors from the USP Dept. of Linguistics (Indigenous languages)
- 1 prof. from USP IT Dept. (robotics)
- 1 post-doc (USP Anthropology)
- 1 technical support staff
- 6 undergrad scholarship interns

Student Collaborators

MISTI - MIT-Brazil

Insper (São Paulo)

https://c4ai.inova.usp.br/research-activities-in-the-c4ai/#

PROINDL: AI technologies to strengthen Indigenous languages in Brazil

focus areas

Development of **writing assistants** for text production and social media use based on Indigenous Language Models.

Desenvolvimento de **aplicativos WhatsApp e Android** para fornecer melhor suporte e acesso a recursos às comunidades.

Developing **translators from Brazilian Indigenous languages** to Portuguese by fine-tuning ML translators.

prototypes

writing assistant

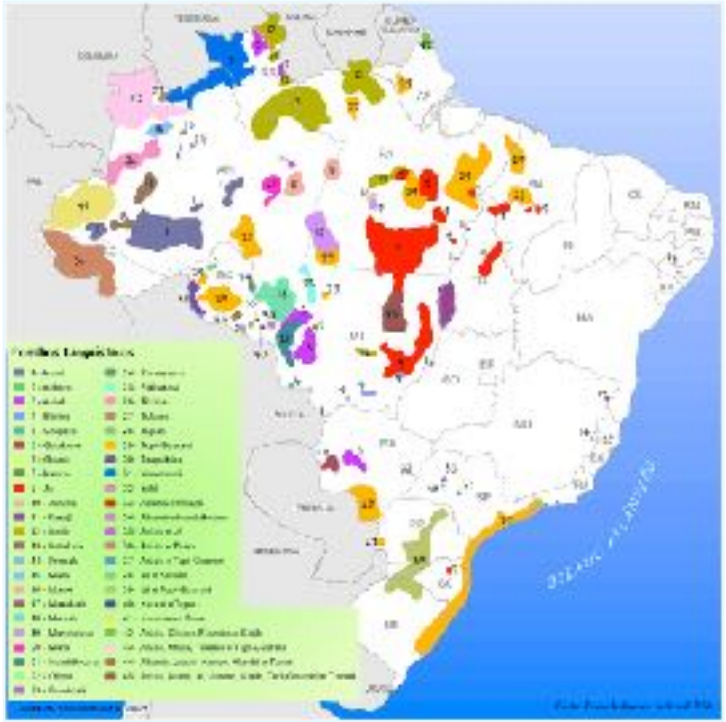
WhatsApp app

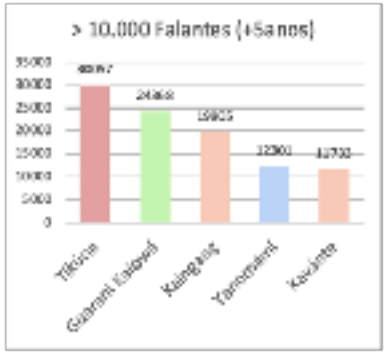
Android app

bidirectional translation of Indigenous languages

community work

The majority of the 200 Indigenous languages in Brazil are likely to disappear in the next 100 years

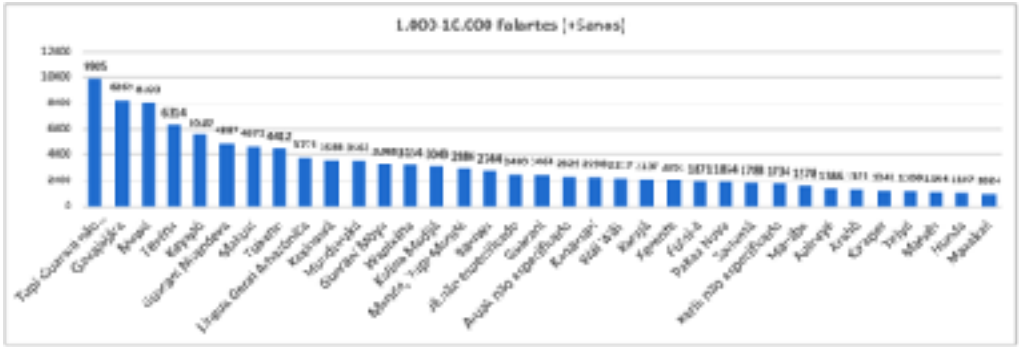




Brazil 2010 census:

270 languages
800K speakers, half in Ind. lands

>10K speakers: 5 languages
1K< speakers < 10K: 35 languages



The Bibles dataset:
translations of the *New Testament* in 39 Brazilian Indigenous languages

Mathew 1:6. ***"and Jesse the father of David the king. And David was the father of Solomon by the wife of Uriah."***

Apalaí => Jese mokyro tuisa Tawi zumy. Tawi mokyro Saromão zumy. Saromão eny mokyro Uria pytỹpyry.
Apurinã => Xesee, Tavii iri. Tavii Isaeokini auĩte itxaua. Tavii, Saromão iri. Saromão inoro mitxi Oria ĩtanoro. Eereka
Asheninka => Ikanta Isaí tzimanake itomi irirori, iitakeri David. Iri pinkatharitaintsiri. Ikanta pinkathari David rowaiy
Culina => Dsese-Isahipa Dabicca abi. Najari Dabipa jodiodenicca tamine tojajari. Pohuapa Saromocca abi. Saromo
Desano => Isai Davi pagɨ ãrĩñumi. Davipɨ Israe majarã tauro opɨ ãrĩñumi. Ĩgɨ Uria waĩcɨgɨ mɨrɨ marapore ĩgɨ m
Guajajára => Zexe umuzãg tuwihaw Tawi a'e. Tawi umuzãg Xãrumãw a'e. (Ihy Uri hemireko kwer romo hekon a'e.)
Guarani_Mbya => Jessé ra'y ma huvixave Davi, hi'e huvixa Davi ra'y ma Salomão, Urias ra'yxykue pi'a va'e.
Jamamadi => Tafi batimatamona amaka Sese. Saomao batimatamona amaka Tafi. Saromao mati Oria fati tohibehe
Kadiwéu => Jessé jigijaa eliodi Davi, Davi anijo me inionigi-eliodi. Davi eliodi Salomão, eliodo Salomão nagejo anijo
Kaingang => Kỹ Jessé vỹ pã'i mág tỹ Davi ẽn han. Tỹ pã'i mág tỹ Davi vỹ Salomão han, ti prũ tỹ tỹ Urias prũ ja ẽn fi k
Karajá => Davi hãwyy heka Uria hãwyy juhuu raremyhỹ. Iribi Davi-wana roire. Saloraõ iriõre rare. Tii heka Roaboa
Macushi => Meropai Jessé wani'pî Davi yun pe, pata esa' pe tîwe'sen.
Maxakalí => hu Namix mũg tak. Yã hõmã Namix te 'Iyaet yõg tikmũ'ũn xohix xat. Ha Namix te 'Onit xetut mũg, ha X
Nadeb => Jesé t'aah kã, Dawi, ẽr wahẽ makũ sa wahẽ n'aa paah. Dawi t'aah, Saromãw. Saromãw ỹỹn Urija hãd nan
Nambikuára => Je^{3}sah^{3}la^{2}ki^{3}lha^{3}ty^{3} Ta^{3}vi^{2}yah^{3}lo^{2}su^{3} ta^{1}hxai^{2}hẽ^{1}la^{2}. A^{2}nxe^{3}jah^{1}lai^{2}na^{2}I^{3}sa^{3}e^{2} nỹ^{3}ka^{3}txa^{3} a^{2}wa^{3}kxẽn^{3}y
[illegible]
Palikúr => Igme amekene Jessé gikamkayh ukiparawiy amekene Davi. Igme amekene Davi xuwehe amekene Uriyas
Parecis => Jessé atyo Davi kaisani hoka Davi atyo xekohaseti kalorexe tyaona. Hatyaoseta Davi atyo Salomão kaisan
Rikbaktsa => Takino Dawizo. Iwa taparakta Abarão tsekokatsa niaha. Iwaze Saromae ta Dawi tse ije tapara Urias ok
Sateré => Mi'i hawyi IESE imoherep aiporekuat sa'awy'iwuanuat TAWI. Mi'i hawyi morekuat koro TAWI kaipyi tuwe
Tenharim => Jesséva'ea Daviva'ea po'ria. Daviva'ea israelitasva'ea ruvihavuhuva'ea. Daviva'ea Salomãova'ea po'ria.
Terêna => Eneponeko Njése, énomone itúko há'aneko Ndávi, nãti mekúke.
Ticuna (Peru) => Rü nüma ga Ichaxí rü ãẽxgacü ga Dabinatü nixĩ. Rü nüma ga ãẽxgacü ga Dabí rü Charomóũnatü ni
Tucano => Isai udio masã wiogɨ Davi pacɨ nicɨ niwĩ. Davi Urias nɨmo nĩ'co me'rã Salomórẽ põ'rãticɨ niwĩ.
Wapishana (Gaiana) => Jesse dani uu King David. Aizii di'i Jesus Christ dokozu-daunnao dau'ana'o King David di'iki'
(Solomon daro Uriah daiare-daun);

Indigenous Languages					# Aligned Sentences		
Name	Acron	Branch	Family	Speakers	Train	Test	Total
Bororo	bor	Macro-Jê	Bororó	1[illegible]	1851	202	2053
Apinayé	apn	Macro-Jê	Jê	[illegible]	877	75	952
Kaingang	kgp	Macro-Jê	Jê	19905	5685	913	681[illegible]
Kayapó	txu	Macro-Jê	Jê	5520	2629	510	3079
Xavánte	xav	Macro-Jê	Jê	11733	1375	342	1517
Karajá	kpj	Macro-Jê	Karajá	3119	2828	233	3061
Maxakalí	mbl	Macro-Jê	Maxakalí	[illegible]	5566	905	6471
Rikbaktsa	rkb	Macro-Jê	Rikbaktsa	10	3580	710	4270
Mawé	mav	Tupi	Mawé	8800	6381	975	7351
Munduruku	myu	Tupi	Munduruku	[illegible]	[illegible]	195	[illegible]
Guajajára	gub	Tupi	Tupi-Guarani	8269	4956	934	5890
Guarani (West Bolivia)	gnw	Tupi	Tupi-Guarani	NA	5253	975	6233
Guarani (East Bolivia)	gui	Tupi	Tupi-Guarani	NA	5262	925	6187
Guarani Kaiowá	kgk	Tupi	Tupi-Guarani	34968	5094	479	[illegible]
Guarani Mbyá	gun	Tupi	Tupi-Guarani	3248	6340	975	7310
Guarani (Paraguay)	gug	Tupi	Tupi-Guarani	NA	5126	975	6069
Ka'apor	urb	Tupi	Tupi-Guarani	1241	3380	436	3816
Kayabi	kyz	Tupi	Tupi-Guarani	[illegible]	2187	285	[illegible]
Nheengatu (LGA)	yrl	Tupi	Tupi-Guarani	3771	5085	690	5725
Tenharim	pah	Tupi	Tupi-Guarani	[illegible]	3395	844	4239
Jamamadi Kanamanti	jaa	no branch	Arawá	217	4759	715	5474
Kulina Madijá	cul	no branch	Arawá	3043	4339	680	5019
Paumarí	pad	no branch	Arawá	[illegible]	3653	379	4032
Apurinã	apu	no branch	Aruák	524	6329	975	7304
Palikúr	plu	no branch	Aruák	[illegible]	6197	964	7161
Parecí	pab	no branch	Aruák	[illegible]	6381	975	7351
Teréna	ter	no branch	Aruák	6314	6381	970	7351
Wapishana	wap	no branch	Aruák	[illegible]	5691	859	[illegible]
Kadiwéu	kbc	no branch	Guaikurú	549	4523	790	5313
Apalaí	apy	no branch	Karib	251	5848	970	6818
Bakairí	bkq	no branch	Karib	573	4000	312	4317
Hixkaryána	hix	no branch	Karib	[illegible]	4229	472	[illegible]
Makuxí	mbc	no branch	Karib	4875	4900	940	5840
Nadëb	mbj	no branch	Makú	[illegible]	5233	813	[illegible]
Nambikwára	nab	no branch	Nambikwára	251	2734	844	[illegible]
Kashinawá (Perú)	cbs	no branch	Pano-Tacanan	[illegible]	3136	135	[illegible]
Tukáno	tuo	no branch	Tukano	4412	3759	846	[illegible]
Yanomámi	guu	no branch	Yanomámi	12301	1285	196	147[illegible]
Tikúna	tca	no branch	no family	30057	3027	382	3403
TOTAL	39	3	16	[illegible]	[illegible]	[illegible]	[illegible]

Using *The Bible* in Indigenous contexts: Issues and Concerns

- *The Bible* is a religious text and sacred to many people in the world, including Indigenous people, and therefore should be treated with great respect and care.
- *The Bible* is also connected to negative aspects of past and present colonial history of Indigenous peoples, in particular to the effort to convert them to Western religions, in particular to Christianism.
- The translations have established *orthographies of domination* into many Indigenous language. There also many issues about the quality of the translations and common problems of "Europefication" of languages [Franchetto, 2008].

"the Bible was a tool for the colonization process [...working] hand-in-hand in the exploitation, subjugation, and continued oppression of the Indigenous Peoples of the U.S."

[Chris Mato Nunpa. 2020. The great evil: Christianity, the bible, and the Native American genocide.]

"at the beginning of the colonization process two tools of genocide were forced upon Native people: the bottle and the bible."

[Stormy Ogden. 2005. The prison-industrial complex in Indigenous California. In Global lockdown: Race, gender, and the prison-industrial complex.]

[Bruna Franchetto. 2008. The war of the alphabets: indigenous peoples between the oral and the written. Mana, 14(SE):31–59.]

Ethical issues with Indigenous data

Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)

Balancing Social Impact, Opportunities, and Ethical Constraints of Using AI in the Documentation and Vitalization of Indigenous Languages

Claudio S. Pinhanez, Paulo Cavalin, Marisa Vasconcelos and Julio Nogima
IBM Research, Brazil
[csantosp,pcavalin,mariasav,jnogima]@br.ibm.com

Abstract

In this paper we discuss how AI can contribute to support the documentation and vitalization of Indigenous languages and how that involves a delicate balancing of ensuring social impact exploring technical opportunities, and dealing with ethical constraints. We start by surveying previous work on using AI and NLP to support critical actions for strengthening diagnosis and ensuring that languages and bypassing key limitations of current technologies. After presenting basic ethical constraints of working with Indigenous languages and communities, we propose that creating and deploying language technology efficacy with and for Indigenous communities forces AI researchers and engineers to address some of the main shortcomings and criticisms of current technologies. Those ideas are also explored in the discussion of a real case of development of large language models for Brazilian Indigenous languages.

1 Introduction

Tradition and progress are often in conflict in Indigenous communities and one of its most common battlefields is in strengthening the use of their own languages. We argue in this paper that using *Artificial Intelligence (AI)* and, particularly, *Natural Language Processing (NLP)* technologies to support social impacts of AI is one of the central themes of our paper and hereby we assume it to mean using AI technology to contribute to the solution of sociotechnical problems of underserved and vulnerable communities, according to the needs expressed by them, respecting their socio and cultural context, and, whenever possible, in projects led by them.

In fact, creating and deploying technology to be used in Indigenous communities must follow ethical guidelines, as discussed in Bird-Jr et al., 2013; Strelnikoff et al., 2019]. These are in stark contrast with traditional practices of AI, such as reliance on big data, data extractivism, and colonial thinking [Kwet, 2021]. Also, as epitomized by the motto *"Nothing for us without us"* adopted by the Indigenous communities as a premise for any language initiative with them, any work, even in research projects, must be done with the community and for its benefit and in a sustainable manner.

We start this paper recognizing the importance of Indigenous peoples and cultures in the world context. We follow with some data and definitions about endangered languages, a discussion about the value of language diversity, and with an overview of threats and challenges of documentation and valorization of Indigenous languages. We then discuss some of the limitations of current NLP technologies to deal with Indigenous languages including issues with *large language models (LLMs)* such as BERT and GPT-3.

After discussing ethical issues and guidelines when working with indigenous communities, we examine a research initiative conducted by some of the authors of this paper to use

IJCAI 2023

Can **culturally toxic data** increase the performance of translation models made for extremely low-resource languages?

- Given that this type of data can raise ethical concerns depending on the context in which it is applied, its use can be beneficial?

Datasets

Dictionary dataset:

- 1,022 short stories aligned sentences in Guarani Mbya (gun) and English (eng);
- 245 sentences from pedagogical material;
- 2,230 sentences from Dooley's Lexical Guarani Mbya dictionary.

- The last two sources were aligned with Portuguese (por), so we used Watson to translate from por to eng.
- After a data cleaning, unicodes normalization, the Dictionary dataset was splitted into 3,155 and 300 gun-eng aligned sentences for training and test.

Bibles dataset:

- 39 translations of the Bible's New Testament, totaling 188,033 BILs-eng aligned sentences.

Divided into 3 training datasets:

- **Bilingual** (only Guarani Mbya): 6,340 training pairs;
- **Tupi Guarani Family** (10 BILs): 43,869 training pairs;
- **All BILs available** (39 BILs): 162,225 training pairs.

For testing, the Matthew chapter from Guarani Mbya New Testament (970 aligned sentences).

Models

zeroshot: WMT19 model;

mbya: WMT19 model finetuned with bilingual data;

TGf: WMT19 model finetuned with Tupi Guarani family data;

all: WMT19 model finetuned with all BILs available data;

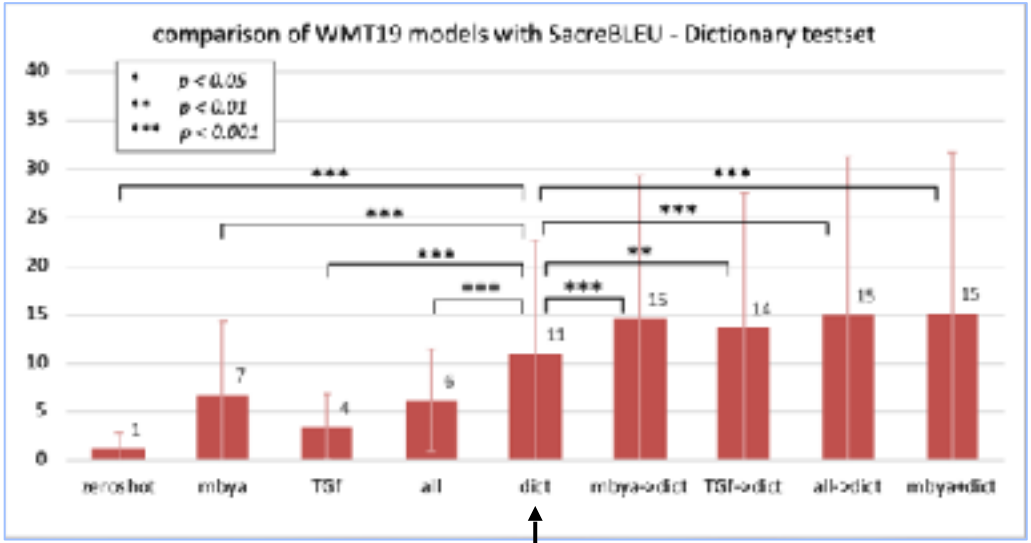
dict: WMT19 model finetuned with dictionary dataset;

mbya->dict: mbya model finetuned with dictionary dataset;

TGf->dict: TGf model finetuned with dictionary dataset;

all->dict: all model finetuned with dictionary dataset;

mbya+dict: WMT19 model finetuned with a merge of bilingual data and dictionary dataset.



Qualitative analysis

models	mbya				dict				mbya+dict			
EXPECTED OUTPUT FROM TEST SET	mbya OUTPUT	Sacre BLEU	BLEU	chrF	dict OUTPUT	Sacre BLEU	BLEU	chrF	mbya+dict OUTPUT	Sacre BLEU	BLEU	chrF
long ago there lived a giant.	there were two men in the crowd.	13	47	30	long ago there lived a giant.	100	100	100	long ago there lived a giant.	100	100	100
one day he went to the woods again.	then one of them went to the other side of the lake.	13	35	32	one day he went again to the woods.	46	91	70	and so one day he went again to the woods.	36	75	73
when he arrived at his house, he said to his wife, "can there be anyone who can hunt like me?"	when he came to the tomb, he said to his mother, how can i not know where i am.	17	42	30	when he got home, he said to his wife, "could it be that i'm from here?"	37	46	38	when he got home, he said to his wife, "couldn't i find the ring?"	32	47	39
when he fell, he hit his back on the ground and died then and there.	so then, how much more will the earth bear down on him than the earth will bear down on him.	5	23	21	as he fell, he hit his forehead on the rock.	37	48	38	and as soon as he touched the ground, he died too.	7	41	30
years ago when i was a child, i didn't know the language of non-indians.	i have not been able to speak the word of the one who sent me into heaven.	5	18	15	years ago when i was a lot younger, i didn't know what to do with the books.	25	55	48	years ago when i was a child, i did not understand the meaning of portuguese.	52	60	54
when my brother went, saw a snake.	when he came to my house, he saw me.	6	32	18	my brother went out to see the snake.	22	58	56	my brother went and saw the snake.	36	73	61
there comes an inhabitant of the here village.	you are one of the twelve living creatures.	10	25	20	there comes the hare from the here.	15	46	38	there comes the spirit village.	15	46	63
he grabbed him by his arm.	so he went up to heaven with his brother.	5	15	12	he took his brother-in-law there.	7	17	17	then he took hold of the indian in the sky.	6	6	9
when evening came, the birds were singing and singing, but the indian was still stuck.	but the spirit of the spirit is in the spirit, and the spirit is in the spirit.	5	25	17	and then it was the turn to eat the birds, both of which were indians.	6	41	35	and the one who drinks the spirit remains in it, though the spirit remains.	5	22	17
you changed substantially what you were to pay me.	if i am a believer, i will be a believer in you.	5	5	9	if you guys believe me, i will believe you.	5	14	11	you will defraud me even more.	8	11	12
who come with lower and higher people;	and all who are in the world and all who are in the world	4	16	17	have lot of faith in him.	6	12	10	low-cost and high-cost carriers also must go;	7	27	25

examples of contaminated (in red) and non-contaminated outputs of the models mbya, dict, and mbya+dict

All contamined verses we could find....

expected output	generated by mbya+dict	BLEU	chrF
he killed three tapirs	he killed three of the jesus,	24	61
dust.	a bagel of dust.	21	64
let's take out the stomach of this pig.	i will pronounce this pig.	18	24
then he saw something like a man open a door in the rock cliff.	then just as the stone was coming out of the tomb, something like a man opened the door.	14	49
the "claw-man" took the indian home to be her husband.	this man took the indian and brought him to life.	13	36
but his foot stuck too.	once again he washed his feet.	8	13
don't spill the tea or do you want to wet the bed completely?	do you not untie the strap of your sandals or sandals?	7	17
the pernilongs bit the one who was sleeping.	the dove dove.	5	6
will it be by chance that bad things happen to us?	have we not turned a blind eye to evil?	4	13
if he wasn't sick, it wouldn't have come.	if i hadn't been born, i would have never been born.	4	23
if you had treated me i would have been cured.	if i die, i die;	4	4
when evening came, the birds were singing and singing, but the indian was still stuck.	and the one who drinks the spirit remains in it, though the spirit remains.	3	17
i said something like, "you what came already what laugh at guarani."	he was very pleased with the way he talked about it: "we could have bought a hat that would belong to jesus."	3	17
sawing cable	small tree with adjective subordinate prayer;	0	11

14 of 300 (4.7%) presented some level of contamination (including 2 direct "Jesus" citations)

Final discussion

The study demonstrates that using culturally toxic data **can significantly improve** the performance of LLM-based translators for ULR languages

- 30% improvement

The use of culturally toxic data can lead to potentially problematic outputs

- 4.7% contamination

Careful consideration and communication with the communities involved

We suggest its use only in **controlled situations** to mitigate negative effects, emphasizing the importance of community involvement and decision-making in the use of such tools.

The results highlight the need for more diverse training data,

- future efforts to involve academic works, community-created data, and synthetic data generation in collaboration with linguists and language experts to **enhance the translation quality while respecting cultural sensitivities.**

Thank You !


Paulo Cavalin


Claudio Pinhanez


Pedro Domingues


Julio Nogima

Contributions

- Understanding that the Bible is toxic in Indigenous contexts.
- Quantification of the impact of the use of culturally toxic Bible data in the creation of Transformer-based Indigenous language models.

Contact:
pcavalin@br.ibm.com
csantosp@br.ibm.com