

Taxi1500: A Multilingual Dataset for Text Classification in 1500 Languages

Chunlan Ma^{1,2*}, Ayyoob Imani^{1,2}, Haotian Ye^{1,2}, Renhao Pei¹, Ehsaneddin Asgari³ and Hinrich Schütze^{1,2}

¹CIS, LMU Munich, Germany.

²Munich Center for Machine Learning (MCML).

³University of California, Berkeley, USA.

*Corresponding author(s). E-mail(s): machunlan@cis.lmu.de;
Contributing authors: ayyoob@cis.lmu.de; yehao@cis.lmu.de;
R.Pei@campus.lmu.de; asgari@berkeley.edu; inquiries@cislmu.org;

Abstract

While natural language processing tools have been developed extensively for some of the world’s languages, a significant portion of the world’s over 7000 languages are still neglected. One reason for this is that evaluation datasets do not yet cover a wide range of languages, including low-resource and endangered ones. We aim to address this issue by creating a text classification dataset encompassing a large number of languages, many of which currently have little to no annotated data available. We leverage parallel translations of the Bible to construct such a dataset by first developing applicable topics and employing a crowdsourcing tool to collect annotated data. By annotating the English side of the data and projecting the labels onto other languages through aligned verses, we generate text classification datasets for more than 1500 languages. We extensively benchmark several existing multilingual language models using our dataset. To facilitate the advancement of research in this area, we release 670 language in 1430 editions at the time of publishing. Our dataset and code are available at: <https://github.com/cisnlp/Taxi1500>.

Keywords: Multilingual, Low-Resource, Dataset, Classification

1 Introduction

Language inequality is a real issue in the world today as minority languages are under-represented and often excluded from language technologies (Joshi et al, 2020). The lack of technological support for minority languages in communities around the globe has a significant impact on the experience of their users and is commonly a cause for virtual barriers such as the *digital divide*.¹ Recent development in language technologies has brought about a surge of multilingual pre-trained language models (MPLMs), such as the multilingual BERT (mBERT) (Devlin et al, 2018), XLM-R (Conneau et al, 2020), and the more recently proposed SERENGETI (Adebara et al, 2022) and Glot500-m (Imani-Googhari et al, 2023), both of which support around 500 languages. While the number of supported languages in the newest MPLMs keeps increasing, we are still unable to quantify the performance on most low-resource languages. We believe that a major cause for why many low-resource languages are still neglected lies in the lack of evaluation datasets for such languages. For example, MPLMs like mBERT and XLM-R are evaluated for many fewer languages than they are pretrained for because of the limited availability of languages in most benchmark datasets.

Most existing multilingual benchmarks, such as XNLI (Lewis et al, 2020) and MLQA (Eisenschlos et al, 2019), rely on translating monolingual benchmarks, as opposed to collecting data from scratch. This approach involves the translation of monolingual data either through machine translation or with the assistance of human professionals. However, machine translation has limitations in terms of the number of languages that can be effectively handled, which depends on the supported languages of the machine translation system, while the quality of translations is also not guaranteed. On the other hand, human translation yields high-quality results but is accompanied by significant costs.

As a solution, we propose a dataset that covers more than 1500 languages. We use translations of the Bible as our source and develop classification topics (i.e., classes) that are general enough so as to apply to many verses and are at the same time not overly abstract. We obtain annotations for the English verses using crowdsourcing. Because the Bible is aligned at the verse level, we can easily project annotations from the English side to all other languages. We attempt to ensure the quality of our annotated data, including by measuring inter-annotator agreement. We name our dataset *Taxi1500*. As a case study, we evaluate three MPLMs (mBERT, XLM-R and Glot500-m) on Taxi1500. Our results suggest that Taxi1500 successfully demonstrates the multilingual generalizability of different MPLMs.

¹labs.theguardian.com/digital-language-divide

2 Related Works

2.1 Multilingual datasets

To date, most datasets that can be used for multilingual task evaluation (Pan et al, 2017; Eisenschlos et al, 2019; De Marneffe et al, 2021) contain no more than a few hundred languages, a small number compared to the world’s 7000 languages. In this section, we give an overview of existing state-of-the-art multilingual datasets.

The Universal Dependencies Treebanks

Universal Dependencies (UD) v2² is an evergrowing multilingual treebank collection, covering 90 languages and 17 tags. UD collects data from an evolution of (universal) Stanford dependencies (De Marneffe et al, 2014), Google universal part-of-speech tags (Petrov et al, 2012), and the Intersect interlingua for morphosyntactic tagsets (Rosen, 2010). UD is often used as a POS tagging component (representing structured prediction) in multilingual benchmarks such as XTREME (Hu et al, 2020).

Wikiann

Pan et al (2017) develop a cross-lingual named entity tagging dataset in 282 languages based on articles from Wikipedia. The framework extracts name tags through cross-lingual and anchor links, self-training, and data selection methods and links them to an English Wikipedia Knowledge Base. Wikiann is recently used for the structured named entity prediction task for multilingual benchmarks such as XTREME (Hu et al, 2020).

Tatoeba

Tatoeba³ is a community-supported collection of English sentences and translations into more than 300 languages. The number of translations updates every Saturday. Artetxe and Schwenk (2019) extract a dataset from Tatoeba with 1000 sentences in 112 languages. Multilingual benchmark XTREME (Hu et al, 2020) collects this dataset as a task by calculating the cosine similarity to evaluate the performance of multilingual models.

MLDoc

Multilingual Document Classification (Lewis et al, 2020) is a multilingual benchmark for document classification for eight languages: English, French, Spanish, Italian, German, Russian, Chinese, and Japanese. It uses data from the Reuters Corpus Volume 2 (RCV2) (Lewis et al, 2004)⁴, a multilingual corpus with 487,000 news stories in thirteen languages (Dutch, French, German,

²<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3687>

³<https://tatoeba.org/eng/>

⁴<https://trec.nist.gov/data/reuters/reuters.html>

Dataset	Languages	Tasks
PAWS-X	6	Sentence-pair Classification
MLQA	7	Question Answering
MLDoc	8	Document Classification
XQuAD	10	Question Answering
XLIN	15	Sentence-pair Classification
XTREME*	40	MLQA, XQuAD, PAWS-X, XLIN, NER, POS
The Universal Dependency	90	POS
Wikiann	258	NER
Tatoeba	300	Machine Translation

Table 1 Multilingual datasets and contained tasks. *XTREME contains the task from MLQA, XQuAD, PAWS-X, XLIN, and two additional tasks NER and POS.

Chinese, Japanese, Russian, Portuguese, Spanish, Latin American Spanish, Italian, Danish, Norwegian, and Swedish) that are manually classified into four groups: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social) and MCAT (Markets). MLDoc samples data with equal class priors to address the issue of data imbalance that previous research on RCV2 encountered. For example, [Klementiev et al \(2012\)](#) define a subset of English and German portions from RCV2, which is used in several follow-up works, for instance, [Mogadala and Rettinger \(2016\)](#) extend the use of RCV2 to French and Spanish through transfer from English. These above-mentioned corpora obtain high accuracy during training but far lower accuracy during testing, which may be caused by the imbalanced distribution of each category. Thus, [Schwenk and Li \(2018\)](#) sample the same number of examples for each class and language from RCV2 instead of choosing a random subset to avoid an imbalanced dataset. The task of MLDoc is the same as RCV2, but with a balanced evaluation, the balanced dataset enables fair evaluation between different language models.

XNLI

The Cross-lingual Natural Language Inference Corpus(XNLI) ([Eisenschlos et al, 2019](#)) is a multilingual evaluation benchmark that extends NLI to 15 languages, namely English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, and two lower-resource languages Swahili and Urdu. XNLI supports the evaluation of the NLI task: two sentences are classified as entailment, contradiction, or neither. XNLI comprises a total of 112,500 annotated sentence pairs and follows the same data collection procedure as the MultiNLI corpus ([Williams et al, 2018](#)): 250 sentences are sampled from ten sources of the Open American National Corpus: Face-To-Face, Telephone, Government, 9/11, Letters, Oxford University Press (OUP), Slate, Verbatim, and Government, and the fiction novel Captain Blood. These sentences are then translated into the other 14 languages using the *One Hour Translation* crowdsourcing platform. This approach enables language models to learn cross-lingual inference ability by using pairs of premise and hypothesis in different languages.

MLQA

MLQA (Lewis et al, 2020) is a series of multilingual extractive question-answering corpora available in seven languages: English, Arabic, German, Vietnamese, Spanish, Simplified Chinese, and Hindi. The resulting corpora have over 12K instances in English and 5K in each other language, with an average of four parallel sentences across languages per instance (Conneau et al, 2020). The extraction process involves identifying paragraphs from Wikipedia articles that cover the same or similar topics in multiple languages. Subsequently, crowdsourcing is used to generate questions and answer spans from English paragraphs via the Amazon Mechanical Turk platform. The researchers employ professional translators to translate English questions into the target languages, and the translators then annotate answer spans within the corresponding paragraphs. MLQA leverages Wikipedia articles as the source due to Wikipedia’s naturally parallel nature and large scale.

XQuAD

XQuAD (Cross-lingual Question Answering Dataset) (Artetxe et al, 2019) is a multilingual benchmark dataset for cross-lingual question answering tasks. It involves the translation of 240 paragraphs and 1,190 question-answer pairs from the development subset of SQuAD v1.1 (Rajpurkar et al, 2016) into ten languages, namely Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi. Instead of collecting data from scratch, XQuAD translates data from SQuAD to avoid unanswerable questions. The researchers constructed XQuAD to mitigate the issue of superficial keyword matching problems that can arise in cross-lingual question answering tasks. The dataset is evaluated using both CLWE (Rajpurkar et al, 2016) and a monolingual model.

XTREME

The Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark (Hu et al, 2020) is a massively compiled multilingual benchmark comprising 40 languages and 9 tasks. As a state-of-the-art multilingual benchmark, XTREME is designed based on available multilingual corpora and their variable tasks. The nine tasks are categorized into different categories, namely classification, structured prediction, Question-Answering, and Information Retrieval (shown in table 2.1). To obtain labeled data, the authors utilized corpora such as XNLI, PAWS-X, Universal Dependencies v2.5 (English for training and target language test set for evaluation) and Wikiann (data selection).

2.2 Multilingual language models

Multilingual models are large language models that cover several different languages. For instance, mBERT and XLM-R are pre-trained on over 100

languages. In this section, we discuss the four models that are relevant to our Taxi1500 evaluation case study.

mBERT

mBERT (Multilingual Bidirectional Encoder Representations from Transformers) (Devlin et al, 2018) is the multilingual variant of BERT, which follows the BERT recipe closely. Similar to BERT, mBERT uses the encoder architecture from Transformer, with Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) as its training objectives. mBERT differs from BERT primarily in its training data. While BERT is trained on English Wikipedia and the Toronto Books Corpus, mBERT has a training set from Wikipedia in 104 languages, with an uncased version (102 languages) and a cased version (104 languages). Both versions have the same hyper-parameters: 12 layers, 768 hidden units per layer, 12 attention heads, a 110k shared WordPiece vocabulary, and 110M parameters⁵. Normally, the cased version is recommended because it fixes normalization issues in many languages.

XLM

XLM (Cross-Lingual Models) (Lample and Conneau, 2019) is also a BERT-based model and acquires multilingual ability by using improved pre-training methods. XLM uses three objectives, two of which are unsupervised and merely require monolingual data. The two unsupervised tasks, Causal Language Modeling (CLM) and Masked Language Modeling (MLM) aim to learn cross-lingual representations. The supervised objective Translation Language Modeling (TLM) relies on parallel corpora. TLM extends MLM from BERT and replaces monolingual sentence pairs with parallel multilingual sentence pairs. A sentence pair is concatenated from a source language sentence and a target language sentence. The words in the source and target sentence are randomly masked. To predict the masked words, the model is allowed to attend to the source language sentence or the target language sentence, which enables the model to align the source and target sentence. Therefore, the model obtains cross-lingual ability with TLM modeling. As for the training data, XLM crawls Wikipedia dumps as monolingual data for CLM and MLM objectives. For TLM, the authors only use parallel data containing English from different resources such as MultiUN (Ziems et al, 2016), IIT Bombay corpus (Kunchukuttan et al, 2018) and EUbookshop corpus. For the embeddings, fastBPE is applied to split words into subword units. The authors release multiple pre-trained versions of XLM, the most massively-multilingual variant is XLM-R (Conneau et al, 2020).

XLM-R

XLM-R (Conneau et al, 2020) is an improved version of XLM. Inspired by RoBERTa (Liu et al, 2019). The authors claim that mBERT and XLM are

⁵<https://github.com/google-research/bert/blob>

both undertrained. Therefore, they pre-train XLM-R with larger model size and massive data from Common Crawl in 104 languages, significantly boosting the performance and outperforming mBERT. Compared with XLM, XLM-R has a larger vocabulary size of 250K. Besides, the training data scales from Wikipedia to a larger Common Crawl corpus. The authors provide two XLM-R versions, XLM-R Base (12 layers, 768 hidden units, 12 attention heads, 270M parameters) and XLM-R Large (24 layers, 1024 hidden units, 16 attention heads, 550M parameters). XLM-R mitigates the curse of multilinguality (i.e., it addresses the increased need for parameters when the number of covered languages increases) by increasing the model capacity.

Glott500-m

Utilizing continuous pretraining based on XLM-R, Glot500-m (ImaniGooghari et al, 2023) was developed as a multilingual LLM on 500 languages. To train the model, the authors collect and clean a corpus, Glot500c, that covers more than 500 languages. Glot500m is evaluated on six tasks, namely sentence retrieval Tatoeba, sentence retrieval Bible, Taxi1500, text classification, NER, POS and round-trip alignment. The authors illustrate results with two language sets: head languages (104 pretrained languages of XLM-R) and tail languages (the remaining languages) and compare results with XLM-R-Base and XLM-R-Large. They find that multilingual LLMs' quality is not only influenced by a single issue, but is determined by several factors, including corpus, script and related languages.

SERENGETI

SERENGETI (Adebara et al, 2022) is pretrained with 517 African languages and the 10 most spoken languages in the world. It is an Electra (Clark et al, 2020) style model. To obtain the training data, the authors collect a multi-domain multi-script corpus manually. The corpus includes religious domain, news domain, government documents, health documents and some data from existing corpora. SERENGETI is evaluated on seven task clusters, containing NER, POS, phrase chunking, news classification, sentiment classification and topic classification and the results are provided as AfroNLU benchmark. Their evaluation indicates that SERENGETI outperforms XLM-R (Conneau et al, 2020), KinyarBERT (Nzeyimana and Rubungo, 2022), AfriBERTA (Ogueji et al, 2021) and Afro-XLMR (Alabi et al, 2022) on 11 datasets with 82.27 average F1.

2.3 The Parallel Bible Dataset

In current NLP research, parallel corpora play a crucial role as they serve as cross-lingual bridges, enabling the processing and understanding of less known languages through other languages. In this study, we employ translations of the Bible as the source of parallel data, utilizing both the Parallel Bible Corpus (Mayer and Cysouw, 2014), covering 1304 languages, as well as

01001001	In the beginning God created the heavens and the earth .
01001001	Im Anfang erschuf Gott die Himmel und die Erde .
01001001	Au commencement Dieu créa les cieux et la terre .

additional translations collected from the web, resulting in total coverage of 1500+ languages. While there are other resources for parallel data available, such as Europarl (Koehn, 2005), JW300 (Agić and Vulić, 2019), and OPUS (Tiedemann, 2012), we have chosen to use translations of the Bible due to the relatively larger number of supported languages.

2.3.1 PBC

PBC consists of three main parts, namely the .txt files of actual Bible texts, the .wordforms files that alphabetically list all word forms in the texts, and the .mtx files (word-by-verse matrices). Our work only uses the .txt files. Every text file is one language version of the Bible. Below we include several examples to illustrate the inner structure of the text file: every line contains an ID and the respective verse which is tokenized. The verse IDs are identical in different languages. When we use the parallel dataset, we can find the same verse in other languages with the help of the verse ID.

2.3.2 1000Langs

The 1000Langs corpus contains the crawled data of 1500+ unique languages, which are sourced from multiple Bible websites. The two main websites are <https://png.bible> and <https://ebible.org>.

3 Dataset Creation

3.1 Principles of task design

In designing and creating Taxi1500, we exploit the nature of PBC (i.e., the fact that it is a parallel corpus), but are at the same time limited by some of its specifics. In particular, we were guided by two considerations: cost efficiency and influence of the domain of the Bible (i.e., religious text).

Cost efficiency. In order to lower the cost, we exclude tasks that require the hiring of target language experts for annotation. Given that the PBC is sentence-aligned, these include tasks such as NER and POS tagging, which are based on word units. We also exclude tasks such as question answering, which require data generation on the side of target language experts. Therefore, the sentence classification task is chosen as our task to utilize the PBC without knowledge of other languages. As the PBC is a parallel text that is sentence aligned, we can easily obtain the label of verses in other languages, as long as we obtain annotations of English verses.

Bible domain. There are many kinds of classification tasks. We exclude sentiment and emotion classification because many verses are objective descriptions of a state or event. In the end, we select topic classification as the a classification task for Taxi1500 because it can be naturally applied to the Bible.

3.2 Data annotation

We describe our data annotation procedure in detail. Since many low-resource languages only have a translated New Testament, we use verses from the New Testament to build our dataset. In the first round of annotation, three annotators develop a few possible topics for a subset of the New Testament verses and decide on six final topics after discussion: recommendation, faith, description, sin, grace, and violence (see table 2). We select verses for which at least two of the three annotators agree. We then remove verses that receive multiple topic assignments and those that receive none. The motivation is that removing ambiguous verses makes the annotation task easier for annotators (which also reduces the annotation cost). We submit the remaining 1077 verses to Amazon MTurk⁶, a crowdsourcing platform, for annotation and specify the US as the annotators' location. Each verse is eventually annotated ten times in total. The final labels are selected by majority voting. In the case of a tie, the final label is randomly selected from the majority labels.

Issues of annotation quality may arise if 1) the task is confusing, and 2) the worker does not annotate carefully. To lessen the effects of the first problem, we provide detailed guidelines and examples to the annotators. Annotators are required to pass a qualification test which is given prior to each annotation batch, making sure they understand the task fully. Additionally, we implement quality control in the form of a performance threshold. Specifically, we construct "pseudo gold standard" data, i.e., labels derived from the majority vote among all annotators, and compare them with the annotations of each worker. The annotations of a worker are rejected and the verses are republished for a new round of annotation if the worker obtains an F1 of $< .4$.

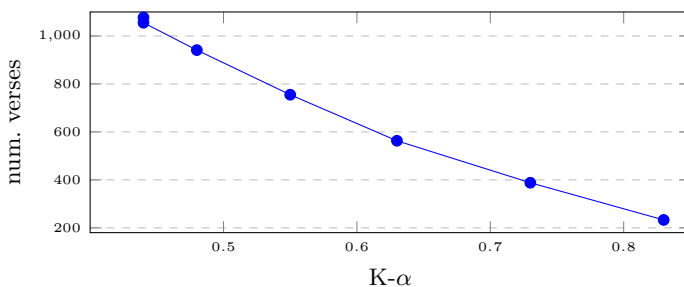
We use Krippendorff's α (K- α) to compute inter-annotator agreement. K- α is chosen because it can handle missing annotations in the dataset since each worker only annotates a subset of the verses. Table 3 shows K- α values for different thresholds, i.e., the minimum votes for the majority label required for a verse to be accepted. We obtain K- $\alpha = .44$ on the entire dataset, which can be improved by raising the threshold of required votes. But as Figure 1 demonstrates, there is a clear tradeoff between the number of accepted verses and K- α , and improving K- α would reduce the size of the dataset. Furthermore, a slightly suboptimal K- α value is not surprising considering that the topics of our task are subjective, and as (Price et al, 2020) points out, a low K- α does not necessarily signify low data quality. We thus do not remove any data by raising the required number of votes and rely on the control measures described above to ensure data quality.

⁶www.mturk.com

class	definition
Recommendation	An imperative statement which suggests to act or believe in certain ways.
Faith	Display of belief and love toward God, instructions on how to maintain faith, stories of faith and its consequences, etc.
Description	Describes a person, relationship, phenomenon, situation, etc.
Sin	Describes what is considered sin, stories of sinful people and sinful actions.
Grace	God's love, blessing, and kindness towards humans.
Violence	Describes wars, conflict, threats, and torture; but also destructions of people, cities, and nations.

Table 2 Definitions of the six Taxi1500 classes

vote \geq	3	4	5	6	7	8	9
num. verses	1077	1055	941	755	563	388	233
K- α	0.44	0.44	0.48	0.55	0.63	0.73	0.83

Table 3 The K- α value increases as we specify a higher threshold for the minimum number of votes of the majority topic. 3 is the lowest value here since we do not have any verses where the majority label has < 3 votes.**Fig. 1** Tradeoff between K- α and the number of verses. Each dot in the plot stands for a threshold of the required minimum votes $\in \{3, 4, 5, 6, 7, 8, 9\}$ for a verse to be accepted.

4 The Dataset

The final Taxi1500 dataset consists of 1077 verses categorized into six topics: faith, grace, sin, violence, description, and recommendation. Table 4 shows an overview of the topics with one example for each, as well as the number of verses of each topic in the English dataset. *Violence*, with 59 instances, is the smallest class and *recommendation*, with 281, is the largest. Since some languages have incomplete translations of the New Testament and do not contain

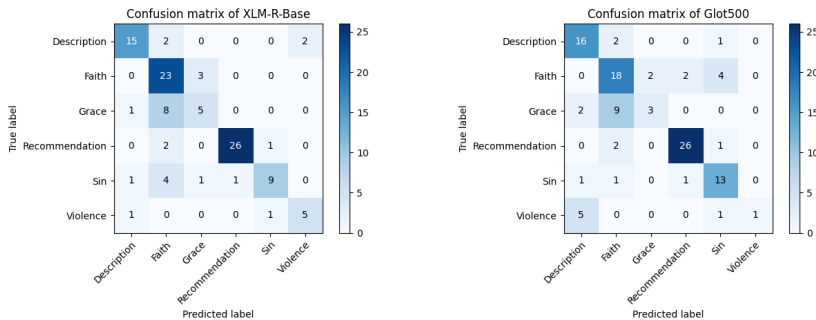


Fig. 2 Confusion matrices of five-fold cross validation of XLM-R-Base and Glot500.

all of the 1077 verses, we exclude languages where the total number of annotated verses is less than 900. This leaves us with 1504 languages from 113 language families which are spread across the globe⁷. The dataset obtained for each of the 1504 languages is split into train, development, and test sets with a ratio of 80/10/10, with 860, 106, and 111 verses respectively⁸.

class	example	num. verses
recommendation	If you love me , you will observe my commandments	281
faith	Most truly I say to you , whoever believes has everlasting life	260
description	There was a man of the Pharisees named Nicodemus , a ruler of the Jews	184
sin	That is why I said to you : You will die in your sins . For if you do not believe that I am the one , you will die in your sins .	153
grace	The Father loves the Son and has given all things into his hand	140
violence	He put James the brother of John to death by the sword	59

Table 4 The table gives an example verse and the total number of verses in the crowdsourced English dataset for each class.

To show more details of Taxi1500’s topics, we present confusion matrices of five-fold cross-validation of XLM-R-Base and Glot500 in Figure 2. The matrices show that the topics Sin and Grace tend to be classified more frequently as other topics. This indicates that verses in Sin and Grace are more ambiguous to the models.

⁷family and geographical data from glottolog.org

⁸development and test sets have different sizes, because we split off train and development verses using their respective ratios and treat the rest as test verses

verse.num	1077	1076	1075	1074	1073	1072	1071	1070	1069	1067	1066	1065
lan.num	1409	20	14	5	4	2	3	5	1	2	2	3
verse.num	1064	1063	1061	1060	1057	1056	1055	1054	1053	1051	1049	1048
lan.num	3	1	2	3	1	2	3	1	1	1	1	3
verse.num	1044	1042	1041	1039	1038	1034	1017	1006	1000	989	961	949
lan.num	1	1	1	1	1	1	1	2	1	1	1	1

Table 5 An overview of the number of verses of different languages, for example: 1049 of the languages have 1077 verses in the dataset.

4.1 The corpus: Taxi1500-c

To provide public access to our dataset, we have carefully selected uncopyrighted Bibles from the PBC and 1000Langs. We then compiled a corpus named Taxi1500-c, which includes all the Bibles that we can freely distribute. The current version available is Taxi1500-c v1.0.

5 A Case Study for the Use of Taxi1500

To illustrate its utility, we use Taxi1500 to evaluate four pre-trained multilingual models: mBERT, XLM-R-Base, XLM-R-Large, and Glot500-m. For a fair comparison, we split languages in our dataset into three sets, namely **head languages**, **Glott500-only languages**, and **tail languages**. Head languages are languages that are in the pre-training data of all four models. Glott500-only languages are languages that are only in the pre-training data of Glott500 and not the other three. Tail languages include languages that are not in the pre-training data of any model. Of the 1504 languages, there are 73 head languages, 250 Glott500-only languages, and 1149 tail languages. We describe the detailed experiment setup in 5.1 and present the metrics on the test set in 5.2.

5.1 Experimental Setup

We conduct experiments on zero-shot transfer and on in-language learning. For all experiments, we select the best checkpoint based on the validation loss and then report macro F1 score on the test set. We use the AdamW optimizer with learning rate $2e - 5$ and batch size $\in \{2, 8, 16, 32\}$ and select the best result based on development set. All experiments are performed on a single GeForce GTX 1080Ti GPU.

Zero-shot transfer. In zero-shot transfer, we train (i.e., finetune) on the English training set and test on the test set of the target language.

In-language learning. In in-language learning, we train (i.e., finetune) on target language training data and test on the test set of the target language. We vary the size of the target language training set and experiment with the following training set sizes: $\{50, 100, 200, 400, 600, 860\}$. The training set size 860 corresponds to the full training set. This allows us to investigate the effect of different amounts of training data.

Evaluation measure. All results presented in this paper are macro-f1, which is chosen considering the imbalance of Taxi1500 dataset.

5.2 Results

In this section, we present experimental results on zero-shot transfer, in-language learning, analysis of the effect of training set size and analysis based on language families.

5.2.1 Zero-shot transfer

Baseline We conducted a Bag-of-Words (BOW) classification experiment with our dataset and present the results as a baseline in Appendix C. The experiment revealed extremely low accuracy for BOW, indicating that to classify verses in our dataset correctly, the models must have access to a good semantic representation. The BOW representation does not seem to be such a representation.

In figure 3, we show the results for 1504 languages, divided into three sets: head languages (top), Glot500-only languages (middle), and tail languages (bottom). On head languages, Glot500, XLM-L-B, and XLM-R-L have 68, 65, and 69 languages within the high F1 range (0.4-0.8), respectively, while mBERT only has 26 languages within this range, indicating its worse performance. This might be explained by a smaller pretraining data size of mBert compared with the other three models. On Glot500-only languages, Glot500 outperforms the other three models with 117 languages in the range of 0.2-0.8, whereas the other three models have less than 30 languages within this range. Because Glot500-only languages are in the pre-training data of Glot500, we expect Glot500 to achieve better results on these languages. On tail languages, Glot500 outperforms the other three models slightly with around 100 fewer languages in the range of 0-0.2. The reason might be that a larger number of pre-training languages contributes to higher performance for other tail languages from the same family. The zero-shot transfer results indicate that Taxi1500 can effectively demonstrate better performance for models pretrained using more languages.

5.2.2 In-language learning

Figure 4 shows differences in F1 for the four models on head, Glot500-only, and tail languages. We see that Glot500 and XLM-R-Base have better performance than mBERT on head languages (most differences are positive). XLM-R-Base outperforms Glot500 slightly on 11 head languages. On Glot500-only languages, Glot500m outperforms the other three models as expected with a larger number of positive differences. On tail languages, mBERT has better performance than the other three models. This may be due to the other models having larger numbers of parameters and thus being more prone to overfitting.

5.2.3 Influence of training set size

To investigate the influence of the training set size, we conduct in-language experiments with 20 selected languages, 10 head and 10 tail languages. We

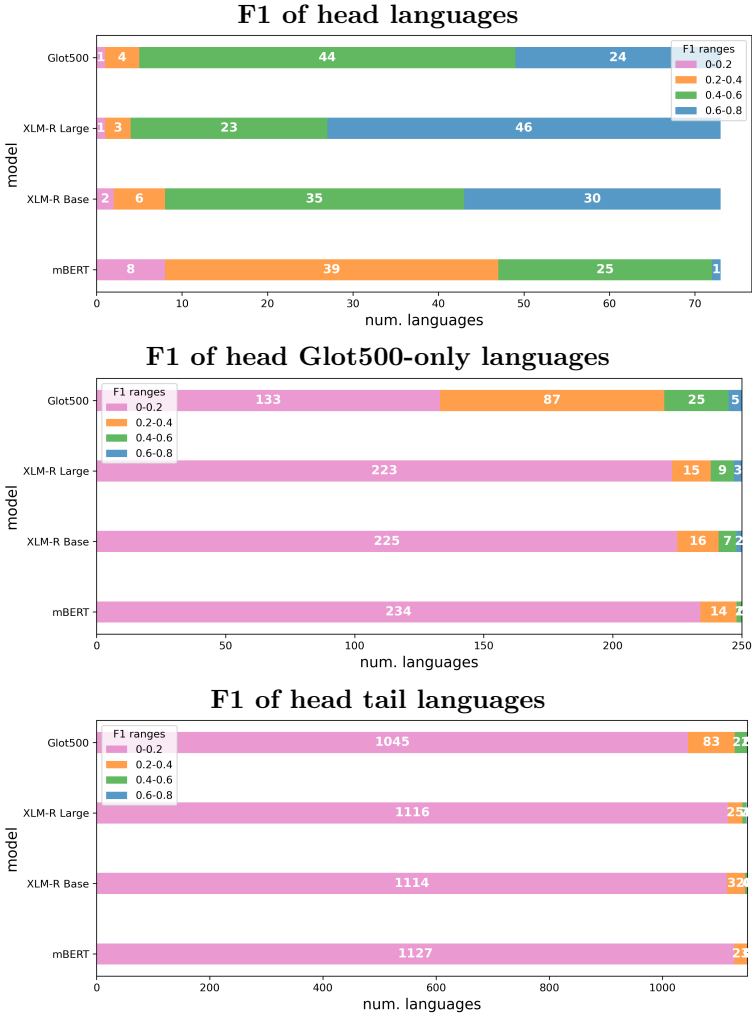


Fig. 3 Zero shot transfer learning: head languages (top), Glott500-only languages (middle) and tail languages (bottom). X-axis is the number of languages, y-axis presents four models. We split F1 scores into four ranges: 0-0.2, 0.2-0.4, 0.4-0.6 and 0.6-0.8.

select languages based on: 1) their inclusion in pretrained multilingual models, specifically whether they are pretrained by the four mentioned SOTA PMLMs or not. 2) variation in typology to ensure coverage of different language types and families, including a) high resource languages, and 2) low resource languages pretrained by the SOTA PMLMs, as well as 3) low resource languages and 4) resource-scarce languages not covered by the SOTA PMLMs like Hixkaryana. We present the iso codes, writing systems and language families of the 20 languages in table 6. The languages are selected to represent 11 different writing systems (Latin, Chinese, Korean, Japanese, Basque, Hebrew, Arabic,

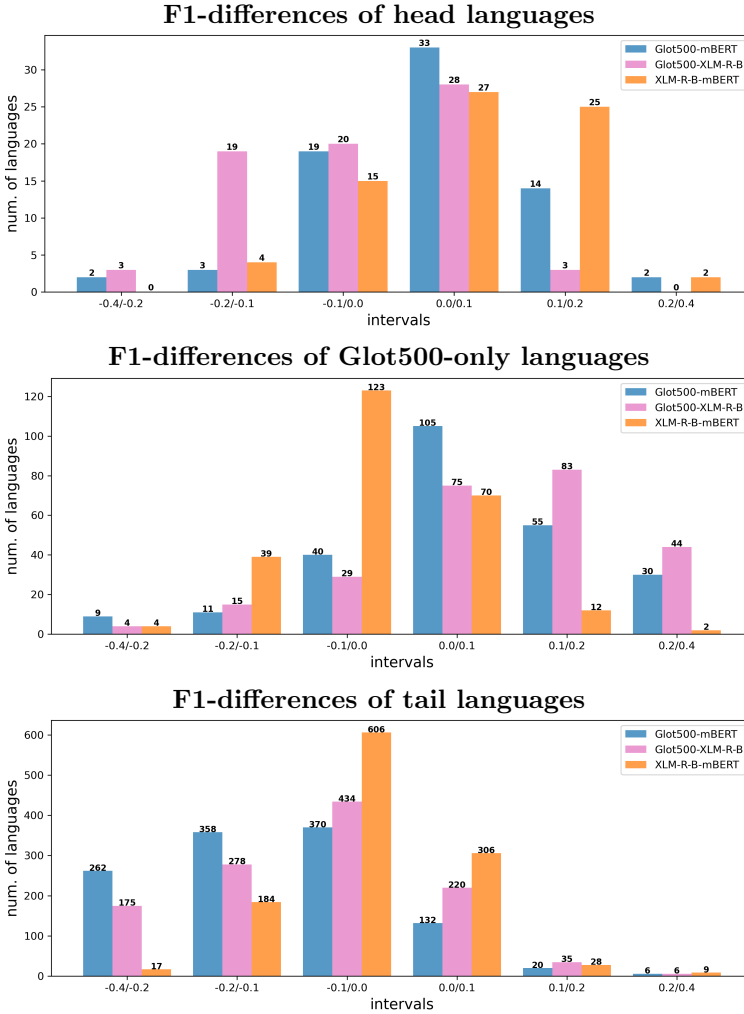


Fig. 4 F1-differences of in-language learning for 1504 languages. We split the F1-differences between two models into six intervals: -0.4/-0.2, -0.2/-0.1, -0.1/0.0, 0.0/0.1, 0.1/0.2, 0.2/0.4. Each bar represents the comparison between a pair of models.

Malayalam, Cyrillic, Devanagari and Burmese) and 13 language families (Indo-European, Sino-Tibetan, Koreanic, Japonic, Basque, Dravidian, Iroquoian, Turkic, Cariban, Uralic, Austronesian, Uto-Aztecan and Central Sudanic). Tables 7 and 8 show the results of zero-shot transfer and in-language experiments using mBERT and XLM-R-Base for the selected languages. As expected, the in-language performance improves when the training set becomes larger. Interestingly, zero-shot transfer performance of head languages is comparable to in-language setting with 100 samples for mBert and with 400 samples for XLM-R-Base, which indicates that models with more parameters may require

more in-language data to reach a comparable level with zero-shot transfer performance. In addition, observing the average zero-shot transfer performance of mBert and XLM-R-Base, XLM-R-Base achieves higher scores on both head and tail languages, this might indicate a better overall performance of XLM-R-Base on Taxi1500 classification task. Moreover, the zero-shot transfer results on both models show that head languages consistently outperform tail languages, which reflects both models' better generalization capability on languages in their pretraining data.

head lang.	iso	Script	Family	tail lang.	iso	Script	Family
German	deu	Latin	Indo-European	Cherokee	chr	Cherokee	Iroquoian
Basque	eus	Latin	Basque	Gagauz	gag	Latin	Turkic
Hebrew	heb	Hebrew	Afro-Asiatic	Hixkaryana	hix	Latin	Cariban
Japanese	jpn	Japanese	Japanic	Nga La	hlt	Latin	Sino-Tibetan
Kazakh	kaz	Cyrilic	Turkic	Komi-Zyrian	kpv	Cyrilic	Uralic
Korean	kor	Korean	Koreanic	Kumyk	kum	Cyrilic	Turkic
Malayalam	mal	Malayalam	Dravidian	Aringa	luc	Latin	Central Sudanic
Burmese	mya	Burmese	Indo-European	Magahi	mag	Devanagari	Indo-European
Persian	pes	Arabic	Indo-European	Dibabawon Manobo	mbd	Latin	Austronesian
Chinese	zho	Chinese	Sino-Tibetan	Middle Watut	npl	Latin	Uto-Aztecan

Table 6 An overview of selected 20 languages from 11 different writing systems and 13 language families

head lang.	transfer	in-language training						tail lang.	transfer	in-language training					
		50	100	200	400	600	860			50	100	200	400	600	860
deu	0.39	0.20	0.13	0.34	0.42	0.44	0.52	chr	0.05	0.24	0.21	0.29	0.35	0.30	0.35
eus	0.17	0.15	0.12	0.31	0.44	0.46	0.43	gag	0.12	0.21	0.29	0.35	0.39	0.45	0.38
heb	0.36	0.24	0.24	0.36	0.33	0.38	0.41	hix	0.07	0.30	0.27	0.35	0.35	0.39	0.41
jpn	0.39	0.37	0.40	0.32	0.49	0.63	0.66	hlt	0.08	0.16	0.25	0.33	0.34	0.44	0.49
kaz	0.29	0.30	0.36	0.38	0.50	0.48	0.48	kpv	0.08	0.19	0.24	0.45	0.41	0.39	0.46
kor	0.41	0.36	0.36	0.45	0.56	0.50	0.60	kum	0.14	0.28	0.27	0.35	0.37	0.42	0.46
mal	0.09	0.13	0.25	0.25	0.31	0.35	0.34	luc	0.08	0.27	0.23	0.46	0.41	0.45	0.35
mya	0.22	0.32	0.31	0.41	0.41	0.40	0.46	mag	0.19	0.14	0.38	0.38	0.37	0.43	0.34
pes	0.43	0.30	0.36	0.55	0.53	0.52	0.56	mbd	0.08	0.18	0.33	0.36	0.36	0.39	0.42
zho	0.36	0.24	0.46	0.47	0.62	0.54	0.59	npl	0.06	0.21	0.30	0.38	0.39	0.40	0.40
avg.	0.31	0.26	0.30	0.38	0.46	0.47	0.51	avg.	0.10	0.22	0.28	0.37	0.37	0.41	0.41

Table 7 Results of zero-shot transfer and in-language fine-tuning experiments using mBERT for 20 selected languages, 10 head (left): German, Basque, Hebrew, Japanese, Kazakh, Korean, Malayalam, Burmese, Persian and Chinese, and 10 tail (right): Cherokee, Gagauz, Hixkaryana, Nga La, Komi-Zyrian, Kumyk, Aringa, Magahi, Dibabawon Manobo and Middle Watut. The numbers in the table header indicate the size of target language training data: 860 means the full training set.

5.2.4 Analysis by Language Family

In Figures 5 and 6, we present zero-shot transfer and in-language results of all languages based on their families (Hammarström, 2015) on XLM-R-Base and Glot500. For almost all families, the performance on head languages is significantly higher than that of Glot500-only and tail languages. The

head lang.	transfer	in-language training						tail lang.	transfer	in-language training					
		50	100	200	400	600	860			50	100	200	400	600	860
deu	0.52	0.16	0.18	0.43	0.49	0.52	0.51	chr	0.09	0.15	0.20	0.15	0.24	0.21	0.28
eus	0.26	0.09	0.26	0.25	0.34	0.37	0.34	gag	0.33	0.17	0.13	0.14	0.45	0.32	0.54
heb	0.15	0.10	0.13	0.18	0.16	0.33	0.35	hix	0.06	0.18	0.17	0.22	0.3	0.43	0.49
jpn	0.62	0.25	0.39	0.53	0.57	0.61	0.68	hlt	0.05	0.14	0.07	0.19	0.40	0.20	0.50
kaz	0.57	0.23	0.35	0.47	0.41	0.55	0.56	kpj	0.09	0.09	0.21	0.23	0.41	0.38	0.53
kor	0.63	0.35	0.55	0.58	0.65	0.53	0.70	kum	0.13	0.13	0.17	0.22	0.27	0.37	0.45
mal	0.07	0.10	0.13	0.22	0.08	0.21	0.24	luc	0.11	0.12	0.11	0.30	0.30	0.39	0.39
mya	0.42	0.18	0.30	0.21	0.45	0.45	0.64	mag	0.38	0.11	0.23	0.41	0.48	0.38	0.51
pes	0.66	0.17	0.55	0.47	0.65	0.64	0.71	mbd	0.11	0.18	0.14	0.25	0.30	0.30	0.38
zho	0.63	0.33	0.49	0.52	0.45	0.51	0.68	npl	0.05	0.14	0.08	0.25	0.41	0.41	0.43
avg.	0.45	0.20	0.33	0.39	0.43	0.47	0.54	avg.	0.14	0.14	0.15	0.24	0.36	0.34	0.45

Table 8 Results of zero-shot transfer and in-language fine-tuning experiments using XLM-R-Base for 20 selected languages, 10 head (left): German, Basque, Hebrew, Japanese, Kazakh, Korean, Malayalam, Burmese, Persian and Chinese, and 10 tail (right): Cherokee, Gagauz, Hixkaryana, Nga La, Komi-Zyrian, Kumyk, Aringa, Magahi, Dibabawon Manobo and Middle Watut. The numbers in the table header indicate the size of target language training data: 860 means the full training set.

Indo-European family outperforms other language families not only on head languages but also on Glot500-only and tail languages. We suppose the reason is that the four evaluated models are pre-trained with more Indo-European languages, which increases the performance of this family. We also notice that XLM-R-Large tends to perform worse than the other three models on most languages. We think this could be due to its larger number of parameters, which makes it prone to overfitting on our small dataset. Interestingly, by comparing zero-shot transfer and in-language results of XLM-R-Base, we find that languages that are extremely low-resource and use non-Latin scripts (e.g. Yawa-Saweru, Lengua-Mascoy, and Hmong-Mien) have significant performance increases (around 0.4) when they are trained with in-language data. This indicates that the four models do not perform as well on non-Latin scripts as on Latin scripts.

6 Conclusion

A bottleneck for the evaluation of multilingual models is the lack of evaluation data for many low-resource languages. Since annotating data for every language is a prohibitively expensive and an unrealistic approach, there is an increasing interest in evaluation data for low-resource languages. In this paper, we introduce a text classification dataset, Taxi1500, which consists of annotated Bible verses in 1504 languages. We obtain labels for the English verses through crowdsourcing and project the labels to all other languages, making use of parallel verses. We present several use cases of Taxi1500 by conducting a thorough evaluation of four multilingual language models. We hope the high language coverage of Taxi1500 will encourage research on multilingual language models, and especially, benefit low-resource languages that are, till now, neglected due to the lack of evaluation data.

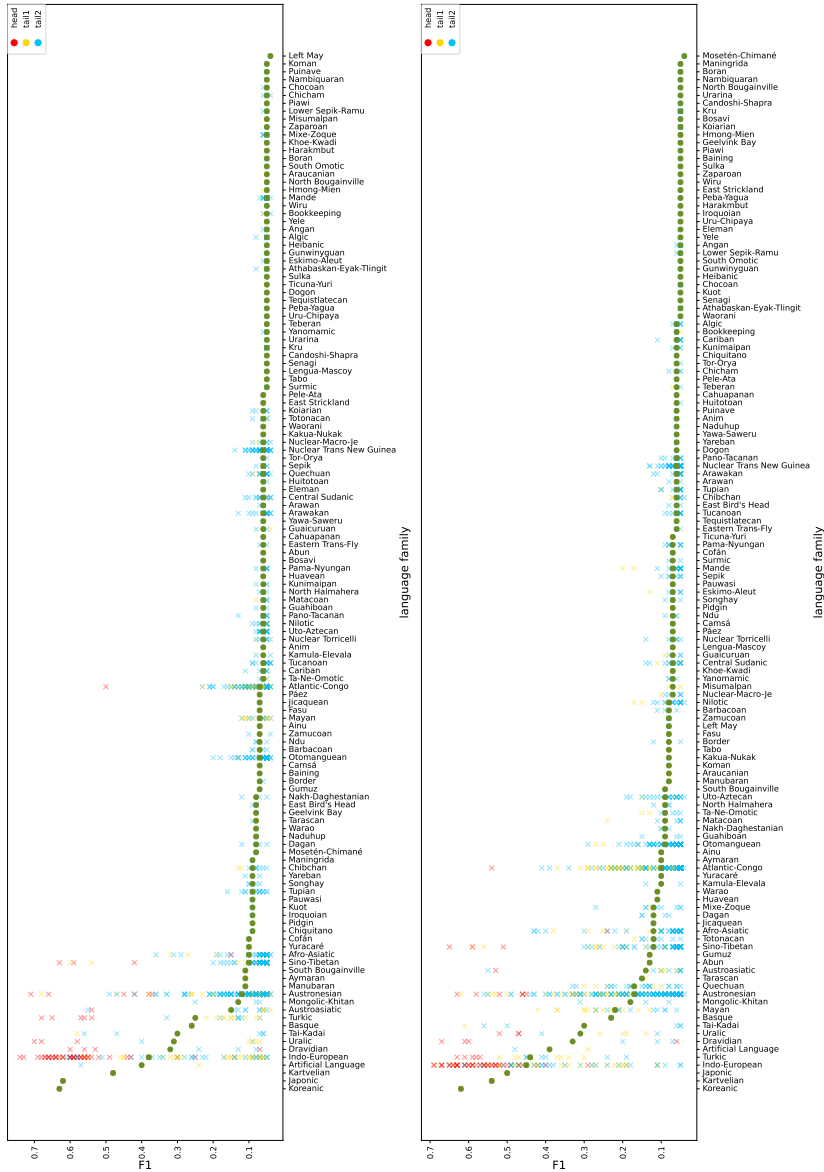


Fig. 5 Zero shot transfer learning: F1 of XLM-R-Base (top) and Glot500 (bottom). Each small dot represents a language, each large dot an average per family. Families are sorted by F1. Red, yellow and blue represent head, Glot500-only and tail languages respectively.

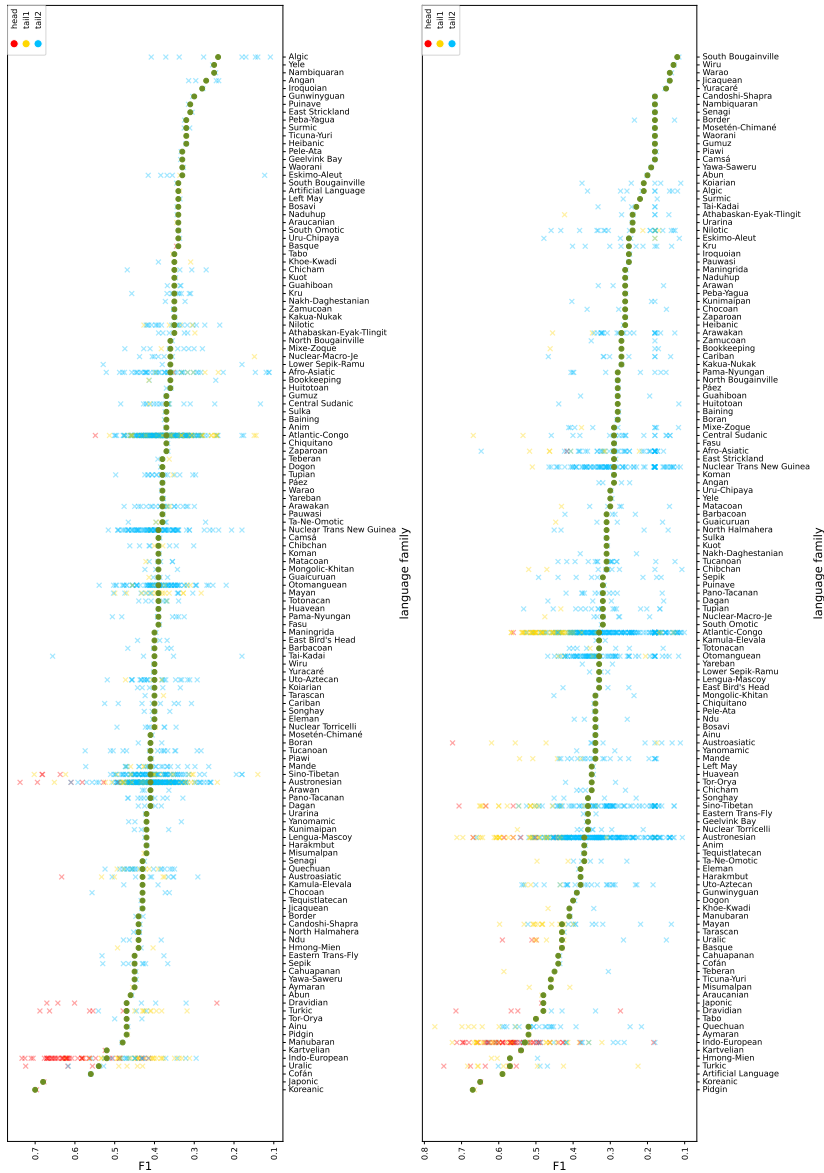


Fig. 6 In-language results: F1 of XLM-R-Base (top) and GLoT500 (bottom). Each small dot represents a language, each large dot an average per family. Families are sorted by F1. Red, yellow and blue represent head, GLoT500-only and tail languages respectively.

7 Limitations

While the high degree of parallelism in the PBC makes it a valuable tool for massively multilingual applications, such as the building of Taxi1500, it is not perfect. One limitation is the religious domain of the Bible, which means keywords specific to the domain may be exploited. Also, we are restricted to the New Testament, as many languages do not have a translated Old Testament in the PBC. Given that some extremely low-resource languages do not have complete translations, the actual number of available verses varies for each language. However, since the Bible is arguably the most translated book in the world, we regard it as a suitable resource for an initiative to build highly parallel dataset like ours.

8 Failure analysis

In this section, we present a path of methods used when designing categories for the sentence classification task and the difficulties met during the data annotation process.

8.1 Category design

We have developed seven versions of topics in total (shown in table 9), each new one based on the refinements of the previous version. This is done by collecting feedback from NLP experts and workers from Amazon MTurk.

We conclude a few of the reasons why earlier versions of topics fail as follows:

1. Lack of domain knowledge. In the first version, we read the Bible and come up with our own topics. Due to limited background knowledge of the religious domain, the first topics are rather arbitrary and do not cover sufficient verses. For the second version, we consult a theologian and online preaching websites that contain a large number of topics, from which we select ones that cover a high number of verses.
2. Obscure or abstract topics. Verses in v2 are collected from an online preaching website, ProPreacher⁹. We sample 100 verses and ask for feedback from crowdsourcing workers on them. Many workers think several topics are very hard to understand or recognize from the verses, for example, *Eschatology*, *Philosophy*, *Theology* and *Moral*. Therefore, v3 deletes four abstract topics, *Eschatology*, *Philosophy*, *Theology*, and *Moral*, and adds five new ones, *Repentance*, *Friendship*, *Thankfulness*, *Forgiveness* and *Suffering*, which are easier to understand.
3. Overlap between topics. v4 is the version we use to crowdsource annotation on Amazon Mechanical Turk. However, we get feedback indicating that many verses can be assigned multiple topics, such as *Violence* and *Conflict*. Therefore, in v5 and v6 we combine similar topics in v4 and change the

⁹<https://www.propreacher.com/100>

names of several labels. v7 is the version that contains the final topics in Taxi1500.

Version	Topics	Num. topics
v1	Rules, Phenomenon, Conflict, Relation, Place, Character, Reward, Punishment, Command	9
v2	Eschatology, Grace, Family, Creation, Philosophy, Revival, Cults, Compromise, Persecution, Hospitality, Conflicts, Theology, Morals, Commandments, Sacrifice	15
v3	Creation, Grace, Violence, Conflict, Hospitality, Sacrifice, Heresy, Repentance, Faith, Suffering, Forgiveness, Thankfulness, Friendship, Temptation	14
v4	Creation, Grace, Violence, Conflict, Hospitality, Sacrifice, Heresy, Repentance, Faith, Suffering, Forgiveness, Thankfulness	12
v5	Creation, Commandment, Genealogy, Violence, Sacrifice, Money, Salvation, Sin	8
v6	Creation, Commandment, Genealogy, Violence, Sacrifice, Money, Grace, Sin	8
v7	Recommendation, Faith, Description, Sin, Grace, Violence	6

Table 9 An overview of different versions of designed categories. v7 is the final version for Taxi1500

8.2 Data annotation

We choose Amazon Mechanical Turk (MTurk) for data annotation because of the availability of a large number of native English speakers. Besides, its usage is well documented in online tutorials for building annotation projects.

Based on our experience, we provide some tips based on our failure when using Amazon MTurk as follows:

1. Although the lowest payment for every HIT is 0.01 US dollars, workers seldom do the task for the minimum payment. It is recommended to set a higher payment if possible.
2. It is not advisable to reject HITs if the requester is new to MTurk, lest the requester's approval rate drops significantly, which will attract fewer workers.
3. Clear instruction and a qualification test prior to permitting workers to annotate are strongly recommended for high-quality data.
4. It is better to test with a smaller batch first before uploading all data for annotation because there can be errors in the instruction or the data submitted.
5. Workers may have valuable opinions about the task and it is a good idea to contact them for feedback.

9 Ethics Statement

In this work, we introduce a new multilingual text classification dataset based on the Parallel Bible Corpus. The data is partially annotated by workers from the Amazon mTurk platform, who are rewarded fairly for their work (\$0.2 per sentence). Our dataset contains Bible verses for which we estimate a low risk of tracing to specific individuals and are intended exclusively for the evaluation of NLP tasks concerning the supported languages. We therefore do not expect any ethical issues with our dataset.

Declarations

Funding. This work is funded by the European Research Council (grant no. 740516).

Conflicts of interest. We do not foresee any conflicts of interest with this work.

Author contributions. Main manuscript text: Chunlan Ma; Tables 2.1, 2, 2, 5, 7, 8: Chunlan Ma; Tables 3, 4, 6: Haotian Ye; Figures 3, 4: Haotian Ye; Review and revision of the manuscript: all authors.

Data availability. Part of the data analyzed for the current study is not publicly available due to copyright restrictions. It is available from the corresponding author upon reasonable request.

References

- Adebara I, Elmadany A, Abdul-Mageed M, et al (2022) Serengeti: Massively multilingual language models for africa. [2212.10785](#)
- Agić Ž, Vulić I (2019) JW300: A wide-coverage parallel corpus for low-resource languages. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 3204–3210, <https://doi.org/10.18653/v1/P19-1310>, URL <https://aclanthology.org/P19-1310>
- Alabi JO, Adelani DI, Mosbach M, et al (2022) Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In: Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp 4336–4349, URL <https://aclanthology.org/2022.coling-1.382>
- Artetxe M, Schwenk H (2019) Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics 7:597–610. https://doi.org/10.1162/tacl_a_00288, URL <https://aclanthology.org/Q19-1038>

- Artetxe M, Ruder S, Yogatama D (2019) On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:191011856
- Clark K, Luong MT, Le QV, et al (2020) Electra: Pre-training text encoders as discriminators rather than generators. [2003.10555](#)
- Conneau A, Khandelwal K, Goyal N, et al (2020) Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 8440–8451, <https://doi.org/10.18653/v1/2020.acl-main.747>, URL <https://aclanthology.org/2020.acl-main.747>
- De Marneffe MC, Dozat T, Silveira N, et al (2014) Universal stanford dependencies: A cross-linguistic typology. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp 4585–4592
- De Marneffe MC, Manning CD, Nivre J, et al (2021) Universal dependencies. Computational linguistics 47(2):255–308
- Devlin J, Chang M, Lee K, et al (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL <http://arxiv.org/abs/1810.04805>, <https://arxiv.org/abs/1810.04805>
- Eisenschlos J, Ruder S, Czapla P, et al (2019) Multiftt: Efficient multi-lingual language model fine-tuning. CoRR abs/1909.04761. URL <http://arxiv.org/abs/1909.04761>, <https://arxiv.org/abs/1909.04761>
- Hammarström H (2015) Glottolog: A free, online, comprehensive bibliography of the world's languages. In: 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace, UNESCO, pp 183–188
- Hu J, Ruder S, Siddhant A, et al (2020) XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. CoRR abs/2003.11080. URL <https://arxiv.org/abs/2003.11080>, <https://arxiv.org/abs/2003.11080>
- ImaniGooghari A, Lin P, Kargaran AH, et al (2023) Glot500: Scaling multilingual corpora and language models to 500 languages. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.
- Joshi P, Santy S, Budhiraja A, et al (2020) The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 6282–6293, <https://doi.org/10.18653/v1/2020.acl-main.560>, URL <https://aclanthology.org/2020.acl-main.560>

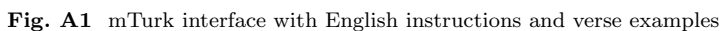
- Klementiev A, Titov I, Bhattarai B (2012) Inducing crosslingual distributed representations of words. In: Proceedings of COLING 2012, pp 1459–1474
- Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, pp 79–86, URL <https://aclanthology.org/2005.mtsummit-papers.11>
- Kunchukuttan A, Mehta P, Bhattacharyya P (2018) The IIT Bombay English-Hindi parallel corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, URL <https://aclanthology.org/L18-1548>
- Lample G, Conneau A (2019) Cross-lingual language model pretraining. CoRR abs/1901.07291. URL <http://arxiv.org/abs/1901.07291>, <https://arxiv.org/abs/1901.07291>
- Lewis DD, Yang Y, Russell-Rose T, et al (2004) Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research 5(Apr):361–397
- Lewis P, Oguz B, Rinott R, et al (2020) MLQA: Evaluating cross-lingual extractive question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 7315–7330, <https://doi.org/10.18653/v1/2020.acl-main.653>, URL <https://aclanthology.org/2020.acl-main.653>
- Liu Y, Ott M, Goyal N, et al (2019) Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692. URL <http://arxiv.org/abs/1907.11692>, <https://arxiv.org/abs/1907.11692>
- Mayer T, Cysouw M (2014) Creating a massively parallel Bible corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14). European Language Resources Association (ELRA), Reykjavik, Iceland, pp 3158–3163, URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf
- Mogadala A, Rettinger A (2016) Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 692–702
- Nzeyimana A, Rubungo AN (2022) KinyaBERT: a morphology-aware kinyarwanda language model. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, <https://doi.org/10.18653/v1/2022>

acl-long.367, URL <https://doi.org/10.18653/v1/2022.acl-long.367>

- Ogueji K, Zhu Y, Lin J (2021) Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In: Proceedings of the 1st Workshop on Multilingual Representation Learning. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 116–126, <https://doi.org/10.18653/v1/2021.mrl-1.11>, URL <https://aclanthology.org/2021.mrl-1.11>
- Pan X, Zhang B, May J, et al (2017) Cross-lingual name tagging and linking for 282 languages. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vancouver, Canada, pp 1946–1958, <https://doi.org/10.18653/v1/P17-1178>, URL <https://aclanthology.org/P17-1178>
- Petrov S, Das D, McDonald R (2012) A universal part-of-speech tagset. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12). European Language Resources Association (ELRA), Istanbul, Turkey, pp 2089–2096, URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf
- Price I, Gifford-Moore J, Flemming J, et al (2020) Six attributes of unhealthy conversations. In: Proceedings of the Fourth Workshop on Online Abuse and Harms. Association for Computational Linguistics, Online, pp 114–124, <https://doi.org/10.18653/v1/2020.alw-1.15>, URL <https://aclanthology.org/2020.alw-1.15>
- Rajpurkar P, Zhang J, Lopyrev K, et al (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp 2383–2392, <https://doi.org/10.18653/v1/D16-1264>, URL <https://aclanthology.org/D16-1264>
- Rosen A (2010) Morphological Tags in Parallel Corpora, pp 205–234
- Schwenk H, Li X (2018) A corpus for multilingual document classification in eight languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, URL <https://aclanthology.org/L18-1560>
- Tiedemann J (2012) Parallel data, tools and interfaces in OPUS. In: Chair) NCC, Choukri K, Declerck T, et al (eds) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12). European Language Resources Association (ELRA), Istanbul, Turkey

Ziemski M, Junczys-Dowmunt M, Pouliquen B (2016) The United Nations parallel corpus v1.0. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, pp 3530–3534, URL <https://aclanthology.org/L16-1561>

Figure A1 shows a screenshot of the annotation interface. Workers are asked to select one label for each verse among six labels. If they think one verse does not belong to any of them, the workers should classify this verse into Other.



Our dataset is built based on PBC and 1000Langs. Due to the copyright issue, our dataset consists of three parts:

- 1403 editions in 670 languages with permissive licenses which we distribute freely (the corpus we call Taxi1500-c v1.0).
- For the remaining PBC Bibles, please contact Michael Cysouw at Philipps University of Marburg to request access to PBC. Once granted access, run the code available at https://github.com/cisnlp/Taxi1500/corpus_obtain to obtain the labeled dataset.
- For the remaining 1000Langs Bibles, use the code provided at <https://github.com/ehsanasgari/1000Langs> to crawl the corpus. Then, run the code

available at https://github.com/cisnlp/Taxi1500/corpus_obtain to obtain the labeled dataset.

Appendix C Results for zero-shot

We report the detailed results for zero-shot transfer of BOW, mBERT, XLM-R-B, XLM-R-L, and Glot500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
aah_Latn	0.13	0.10	0.05	0.05	0.08	aoz_Latn	0.21	0.13	0.07	0.05	0.07
aai_Latn	0.22	0.15	0.09	0.05	0.09	apb_Latn	0.07	0.08	0.06	0.05	0.12
aak_Latn	0.07	0.13	0.05	0.05	0.05	ape_Latn	0.13	0.13	0.05	0.05	0.07
aau_Latn	0.12	0.12	0.06	0.05	0.10	apn_Latn	0.07	0.19	0.06	0.05	0.05
aaz_Latn	0.07	0.12	0.05	0.05	0.08	apr_Latn	0.07	0.07	0.07	0.05	0.05
abi_Latn	0.07	0.11	0.05	0.05	0.05	apt_Latn	0.08	0.14	0.07	0.05	0.07
abt_Latn	0.09	0.13	0.08	0.05	0.06	apu_Latn	0.07	0.09	0.10	0.05	0.05
abx_Latn	0.16	0.12	0.20	0.14	0.33	apw_Latn	0.15	0.10	0.05	0.05	0.05
aby_Latn	0.21	0.12	0.07	0.07	0.06	apy_Latn	0.09	0.09	0.11	0.05	0.05
acd_Latn	0.13	0.08	0.05	0.05	0.05	apz_Latn	0.07	0.11	0.05	0.05	0.05
ace_Latn	0.13	0.25	0.11	0.11	0.30	are_Latn	0.11	0.12	0.05	0.05	0.05
acf_Latn	0.09	0.25	0.06	0.05	0.38	arl_Latn	0.15	0.14	0.05	0.05	0.05
ach_Latn	0.13	0.12	0.05	0.05	0.08	arn_Latn	0.13	0.08	0.05	0.05	0.08
acn_Latn	0.07	0.10	0.05	0.05	0.05	ary_Arab	0.07	0.28	0.19	0.27	0.19
acr_Latn	0.16	0.14	0.06	0.05	0.30	arz_Arab	0.07	0.43	0.32	0.47	0.25
acu_Latn	0.10	0.10	0.05	0.05	0.08	asg_Latn	0.08	0.11	0.05	0.05	0.06
ade_Latn	0.12	0.10	0.07	0.05	0.06	asm_Beng	0.07	0.17	0.43	0.47	0.51
adh_Latn	0.13	0.15	0.07	0.05	0.07	aso_Latn	0.15	0.12	0.05	0.05	0.05
adi_Latn	0.09	0.10	0.14	0.05	0.09	ata_Latn	0.11	0.12	0.06	0.05	0.06
adj_Latn	0.17	0.08	0.05	0.05	0.05	atb_Latn	0.10	0.09	0.07	0.05	0.06
adi_Latn	0.08	0.18	0.05	0.05	0.05	atd_Latn	0.11	0.09	0.05	0.05	0.05
aeb_Arab	0.07	0.38	0.19	0.42	0.30	atg_Latn	0.10	0.11	0.07	0.05	0.07
aer_Latn	0.07	0.08	0.08	0.05	0.05	atq_Latn	0.13	0.15	0.06	0.05	0.13
aeu_Latn	0.07	0.13	0.05	0.05	0.05	att_Latn	0.14	0.10	0.08	0.05	0.16
aey_Latn	0.07	0.12	0.09	0.05	0.05	auc_Latn	0.09	0.13	0.06	0.05	0.05
afr_Latn	0.33	0.45	0.59	0.66	0.52	auy_Latn	0.07	0.07	0.04	0.05	0.06
agd_Latn	0.09	0.16	0.06	0.08	0.07	ava_Cyrl	0.07	0.06	0.05	0.05	0.10
agg_Latn	0.14	0.06	0.05	0.05	0.05	avn_Latn	0.14	0.12	0.05	0.05	0.05
agm_Latn	0.07	0.11	0.06	0.05	0.05	avt_Latn	0.10	0.11	0.05	0.05	0.14
agn_Latn	0.12	0.16	0.13	0.18	0.35	avu_Latn	0.07	0.06	0.04	0.05	0.05
agr_Latn	0.07	0.11	0.05	0.05	0.05	awa_Deva	0.07	0.24	0.37	0.40	0.48
agt_Latn	0.07	0.10	0.06	0.05	0.10	awb_Latn	0.08	0.11	0.06	0.05	0.05
agu_Latn	0.11	0.09	0.04	0.05	0.06	awi_Latn	0.17	0.12	0.04	0.05	0.14
agw_Latn	0.20	0.13	0.11	0.07	0.24	ayo_Latn	0.12	0.12	0.10	0.05	0.08
ahk_Latn	0.08	0.11	0.07	0.05	0.07	ayp_Arab	0.07	0.30	0.29	0.35	0.43
aia_Latn	0.23	0.13	0.05	0.05	0.08	ayr_Latn	0.07	0.12	0.11	0.06	0.10
aii_Syrc	0.07	0.05	0.05	0.09	0.10	azb_Arab	0.07	0.16	0.15	0.08	0.34
aim_Latn	0.10	0.14	0.06	0.05	0.05	aze_Latn	0.07	0.32	0.56	0.68	0.59
ain_Latn	0.11	0.09	0.07	0.05	0.10	azg_Latn	0.04	0.09	0.05	0.05	0.05
aji_Latn	0.13	0.14	0.05	0.05	0.05	azz_Latn	0.14	0.15	0.06	0.06	0.10
ajz_Latn	0.12	0.12	0.05	0.05	0.07	bak_Cyrl	0.07	0.33	0.13	0.05	0.24
aka_Latn	0.12	0.17	0.10	0.06	0.13	bam_Latn	0.09	0.11	0.06	0.05	0.20
akb_Latn	0.13	0.16	0.15	0.07	0.27	ban_Latn	0.07	0.16	0.16	0.09	0.31
ake_Latn	0.11	0.08	0.05	0.05	0.05	bao_Latn	0.10	0.14	0.08	0.05	0.06
akh_Latn	0.10	0.15	0.05	0.05	0.05	bar_Latn	0.13	0.19	0.30	0.29	0.41
akp_Latn	0.10	0.16	0.06	0.05	0.05	bav_Latn	0.12	0.05	0.05	0.05	0.06
ald_Latn	0.08	0.05	0.05	0.05	0.05	bba_Latn	0.13	0.12	0.05	0.05	0.05
alj_Latn	0.11	0.14	0.10	0.10	0.21	bbb_Latn	0.07	0.09	0.05	0.05	0.05
aln_Latn	0.07	0.25	0.46	0.53	0.55	bbj_Latn	0.12	0.05	0.05	0.05	0.05
alp_Latn	0.10	0.19	0.13	0.06	0.20	bbk_Latn	0.09	0.04	0.05	0.05	0.05
alq_Latn	0.09	0.11	0.05	0.05	0.05	bbo_Latn	0.10	0.12	0.07	0.05	0.06
als_Latn	0.07	0.24	0.45	0.54	0.49	bbr_Latn	0.17	0.15	0.04	0.05	0.06
alt_Cyrl	0.07	0.16	0.17	0.19	0.37	bch_Latn	0.10	0.13	0.07	0.05	0.12
alz_Latn	0.10	0.15	0.06	0.05	0.17	bei_Latn	0.09	0.12	0.04	0.05	0.15
ame_Latn	0.09	0.11	0.09	0.05	0.05	bel_Latn	0.07	0.18	0.26	0.20	0.46
amf_Latn	0.07	0.08	0.05	0.05	0.05	bcw_Latn	0.12	0.05	0.06	0.05	0.05
amb_Ethi	0.07	0.05	0.10	0.05	0.07	bdd_Latn	0.11	0.07	0.05	0.05	0.05
amk_Latn	0.13	0.19	0.06	0.05	0.07	bdh_Latn	0.07	0.10	0.05	0.05	0.05
amm_Latn	0.09	0.07	0.04	0.05	0.08	bdq_Latn	0.10	0.12	0.05	0.05	0.05
amn_Latn	0.11	0.11	0.07	0.05	0.12	bef_Latn	0.10	0.10	0.07	0.05	0.07
amp_Latn	0.07	0.12	0.06	0.05	0.05	bel_Cyrl	0.07	0.43	0.59	0.67	0.59
amr_Latn	0.09	0.12	0.05	0.05	0.05	bem_Latn	0.14	0.11	0.08	0.09	0.31
amu_Latn	0.06	0.08	0.05	0.05	0.05	ben_Beng	0.07	0.32	0.56	0.67	0.63
anm_Latn	0.13	0.14	0.06	0.05	0.05	beq_Latn	0.14	0.14	0.09	0.05	0.10
ann_Latn	0.14	0.15	0.08	0.05	0.06	bex_Latn	0.13	0.10	0.05	0.05	0.08
any_Latn	0.13	0.13	0.05	0.05	0.08	bfd_Latn	0.11	0.09	0.05	0.05	0.05
any_Latn	0.07	0.07	0.05	0.05	0.05	bfo_Latn	0.10	0.11	0.05	0.05	0.06
aoj_Latn	0.20	0.09	0.08	0.05	0.06	bgr_Latn	0.16	0.17	0.07	0.05	0.30
aom_Latn	0.23	0.16	0.05	0.05	0.05	bgs_Latn	0.15	0.14	0.09	0.07	0.11
aon_Latn	0.08	0.11	0.06	0.05	0.05	bgt_Latn	0.15	0.16	0.07	0.05	0.16

Table C1 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glott500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
bgz_Latn	0.09	0.18	0.09	0.06	0.15	bjz_Latn	0.24	0.15	0.13	0.06	0.35
bhl_Latn	0.10	0.12	0.06	0.05	0.07	caa_Latn	0.14	0.15	0.07	0.05	0.12
bhp_Latn	0.09	0.11	0.16	0.06	0.09	cab_Latn	0.07	0.10	0.05	0.05	0.05
bhw_Latn	0.09	0.16	0.07	0.05	0.14	cac_Latn	0.12	0.12	0.06	0.05	0.21
bhz_Latn	0.18	0.14	0.06	0.05	0.06	caf_Latn	0.09	0.07	0.05	0.05	0.05
bib_Latn	0.16	0.06	0.05	0.05	0.06	cag_Latn	0.07	0.14	0.05	0.05	0.11
big_Latn	0.09	0.10	0.05	0.05	0.05	cak_Latn	0.04	0.12	0.05	0.05	0.42
bim_Latn	0.14	0.13	0.05	0.05	0.06	cao_Latn	0.08	0.10	0.05	0.05	0.10
bis_Latn	0.16	0.22	0.14	0.06	0.24	cap_Latn	0.11	0.09	0.05	0.05	0.05
biu_Latn	0.16	0.14	0.05	0.05	0.17	caq_Latn	0.10	0.10	0.04	0.05	0.10
biv_Latn	0.11	0.07	0.05	0.05	0.05	car_Latn	0.13	0.12	0.06	0.05	0.06
bjr_Latn	0.07	0.10	0.05	0.05	0.05	cas_Latn	0.15	0.09	0.08	0.05	0.04
bjv_Latn	0.11	0.08	0.06	0.05	0.05	cat_Latn	0.13	0.41	0.58	0.64	0.47
bkd_Latn	0.07	0.21	0.15	0.08	0.21	cav_Latn	0.07	0.11	0.06	0.05	0.05
bkl_Latn	0.15	0.11	0.06	0.07	0.05	cax_Latn	0.07	0.12	0.09	0.05	0.06
bkg_Latn	0.14	0.12	0.06	0.05	0.11	cbc_Latn	0.08	0.14	0.06	0.05	0.05
bku_Latn	0.15	0.11	0.08	0.06	0.19	cbl_Latn	0.14	0.13	0.09	0.05	0.11
bkv_Latn	0.13	0.06	0.06	0.05	0.09	cbk_Latn	0.11	0.39	0.45	0.48	0.57
bhl_Latn	0.05	0.07	0.05	0.05	0.05	cbr_Latn	0.13	0.15	0.05	0.05	0.05
bht_Latn	0.11	0.08	0.07	0.05	0.06	cbs_Latn	0.05	0.15	0.05	0.05	0.06
blw_Latn	0.07	0.15	0.06	0.05	0.10	cbt_Latn	0.08	0.09	0.06	0.05	0.06
blz_Latn	0.15	0.19	0.09	0.06	0.12	cbu_Latn	0.07	0.12	0.05	0.05	0.05
bmb_Latn	0.14	0.14	0.09	0.05	0.10	cbv_Latn	0.09	0.15	0.06	0.05	0.08
bmb_Latn	0.07	0.11	0.08	0.05	0.08	cce_Latn	0.09	0.10	0.09	0.05	0.21
bmq_Latn	0.10	0.07	0.05	0.05	0.05	cco_Latn	0.10	0.06	0.05	0.05	0.05
bmr_Latn	0.07	0.13	0.05	0.05	0.05	ccp_Latn	0.11	0.19	0.09	0.06	0.09
bmz_Latn	0.09	0.14	0.05	0.05	0.05	cdf_Latn	0.09	0.12	0.05	0.05	0.09
bmv_Latn	0.16	0.10	0.07	0.05	0.05	ceb_Latn	0.11	0.12	0.28	0.28	0.37
bnj_Latn	0.09	0.13	0.07	0.06	0.05	ceg_Latn	0.15	0.15	0.04	0.05	0.08
bno_Latn	0.10	0.18	0.18	0.11	0.33	cek_Latn	0.09	0.10	0.05	0.05	0.06
bnp_Latn	0.11	0.13	0.05	0.06	0.16	ces_Latn	0.07	0.28	0.66	0.57	0.51
boa_Latn	0.09	0.16	0.05	0.05	0.05	cfm_Latn	0.14	0.15	0.05	0.05	0.25
boj_Latn	0.13	0.10	0.05	0.05	0.07	cgc_Latn	0.07	0.18	0.19	0.14	0.26
bom_Latn	0.08	0.11	0.05	0.05	0.08	cha_Latn	0.12	0.12	0.11	0.05	0.19
bos_Latn	0.11	0.19	0.07	0.06	0.05	chd_Latn	0.09	0.10	0.05	0.05	0.06
bov_Latn	0.07	0.12	0.05	0.05	0.06	che_Cyrl	0.07	0.10	0.07	0.05	0.08
box_Latn	0.09	0.11	0.05	0.05	0.09	chf_Latn	0.09	0.10	0.12	0.05	0.21
bpr_Latn	0.13	0.13	0.09	0.05	0.09	chj_Latn	0.10	0.06	0.05	0.05	0.05
bps_Latn	0.16	0.11	0.08	0.05	0.08	chk_Hani	0.07	0.13	0.07	0.05	0.08
bqc_Latn	0.07	0.11	0.05	0.05	0.06	chq_Latn	0.09	0.10	0.05	0.05	0.05
bqj_Latn	0.17	0.12	0.09	0.05	0.07	chr_Cher	0.07	0.05	0.09	0.05	0.05
bqp_Latn	0.09	0.17	0.05	0.05	0.06	chu_Cyrl	0.07	0.31	0.60	0.61	0.46
bre_Latn	0.08	0.29	0.25	0.43	0.29	chv_Cyrl	0.07	0.18	0.07	0.05	0.19
bru_Latn	0.10	0.10	0.07	0.05	0.05	chz_Latn	0.07	0.08	0.05	0.05	0.05
bsc_Latn	0.15	0.08	0.09	0.05	0.05	cjo_Latn	0.07	0.07	0.04	0.05	0.05
bsn_Latn	0.16	0.07	0.04	0.05	0.07	cjp_Latn	0.14	0.11	0.07	0.05	0.05
bss_Latn	0.07	0.13	0.10	0.05	0.05	cjv_Latn	0.06	0.08	0.07	0.05	0.05
btd_Latn	0.09	0.30	0.21	0.17	0.28	ckb_Latn	0.16	0.09	0.07	0.07	0.43
bth_Latn	0.10	0.14	0.12	0.07	0.25	cko_Latn	0.08	0.09	0.06	0.05	0.06
bto_Latn	0.07	0.11	0.13	0.05	0.32	cle_Latn	0.11	0.04	0.05	0.05	0.06
btt_Latn	0.12	0.14	0.07	0.05	0.06	clu_Latn	0.11	0.14	0.18	0.21	0.43
btz_Latn	0.16	0.23	0.20	0.19	0.34	cly_Latn	0.15	0.12	0.11	0.05	0.06
bud_Latn	0.05	0.12	0.05	0.05	0.05	cme_Latn	0.09	0.12	0.05	0.05	0.05
bug_Latn	0.09	0.19	0.12	0.07	0.17	cmn_Hani	0.07	0.40	0.59	0.62	0.65
buk_Latn	0.07	0.11	0.05	0.05	0.08	cmo_Latn	0.18	0.17	0.13	0.05	0.05
bul_Cyrl	0.07	0.41	0.62	0.64	0.60	cmr_Latn	0.11	0.13	0.05	0.05	0.06
bum_Latn	0.09	0.16	0.06	0.05	0.17	cnh_Latn	0.18	0.12	0.08	0.05	0.20
bus_Latn	0.08	0.13	0.05	0.05	0.05	cni_Latn	0.07	0.07	0.05	0.05	0.05
bvc_Latn	0.14	0.21	0.06	0.05	0.08	cnk_Latn	0.09	0.09	0.05	0.05	0.06
bvd_Latn	0.19	0.11	0.06	0.05	0.08	cnl_Latn	0.07	0.07	0.05	0.05	0.05
bvr_Latn	0.12	0.07	0.09	0.05	0.05	cnt_Latn	0.07	0.08	0.05	0.05	0.05
bvz_Latn	0.13	0.10	0.08	0.05	0.05	cnw_Latn	0.12	0.13	0.06	0.05	0.14
bwq_Latn	0.15	0.09	0.06	0.05	0.11	coe_Latn	0.07	0.08	0.05	0.05	0.06
bwu_Latn	0.14	0.16	0.08	0.05	0.09	cof_Latn	0.11	0.15	0.06	0.05	0.08
bxx_Cyrl	0.07	0.09	0.25	0.27	0.31	cok_Latn	0.13	0.08	0.05	0.05	0.07
byr_Latn	0.07	0.08	0.05	0.05	0.06	con_Latn	0.28	0.07	0.10	0.05	0.07
byx_Latn	0.07	0.13	0.07	0.06	0.05	cop_Copt	0.07	0.07	0.05	0.05	0.05
bzd_Latn	0.07	0.10	0.05	0.05	0.04	cor_Latn	0.09	0.12	0.09	0.05	0.11
bzh_Latn	0.15	0.08	0.05	0.05	0.05	cot_Latn	0.07	0.12	0.05	0.05	0.05
bzi_Thai	0.07	0.07	0.07	0.05	0.05	cou_Latn	0.10	0.14	0.06	0.05	0.05

Table C2 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glott500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
cpa_Latn	0.07	0.11	0.05	0.05	0.05	due_Latn	0.10	0.12	0.16	0.05	0.20
cpb_Latn	0.07	0.08	0.08	0.05	0.05	dug_Latn	0.08	0.17	0.17	0.11	0.16
cpc_Latn	0.09	0.12	0.06	0.05	0.05	duo_Latn	0.14	0.08	0.16	0.06	0.31
cpu_Latn	0.09	0.11	0.04	0.07	0.05	dur_Latn	0.10	0.10	0.05	0.05	0.05
cpy_Latn	0.07	0.08	0.05	0.05	0.05	dwr_Latn	0.15	0.11	0.06	0.05	0.10
crih_Cyrl	0.07	0.19	0.15	0.20	0.45	dww_Latn	0.07	0.07	0.08	0.05	0.06
crj_Latn	0.15	0.10	0.05	0.05	0.05	dyi_Latn	0.16	0.13	0.07	0.05	0.06
crk_Cans	0.07	0.05	0.05	0.05	0.05	dyo_Latn	0.08	0.12	0.07	0.05	0.08
crl_Cans	0.07	0.09	0.05	0.05	0.05	dyu_Latn	0.07	0.09	0.05	0.05	0.17
crm_Cans	0.07	0.05	0.05	0.05	0.06	dzo_Tibt	0.07	0.04	0.05	0.08	0.09
crn_Latn	0.10	0.09	0.05	0.05	0.06	ebk_Latn	0.14	0.15	0.05	0.05	0.17
crq_Latn	0.09	0.16	0.06	0.05	0.05	efi_Latn	0.13	0.13	0.07	0.05	0.11
crs_Latn	0.10	0.17	0.15	0.05	0.43	eka_Latn	0.11	0.17	0.09	0.06	0.06
crt_Latn	0.10	0.16	0.06	0.05	0.05	ell_Grek	0.07	0.31	0.43	0.60	0.50
crx_Latn	0.09	0.08	0.08	0.05	0.05	emi_Latn	0.09	0.16	0.05	0.10	0.09
csk_Latn	0.12	0.14	0.09	0.05	0.05	emp_Latn	0.14	0.10	0.06	0.05	0.05
cso_Latn	0.07	0.08	0.05	0.05	0.05	enb_Latn	0.07	0.10	0.05	0.05	0.05
csy_Latn	0.10	0.11	0.08	0.05	0.14	eng_Latn	0.43	0.57	0.65	0.56	0.63
cta_Latn	0.07	0.13	0.05	0.05	0.07	enl_Latn	0.09	0.10	0.05	0.05	0.07
ctd_Latn	0.11	0.14	0.07	0.05	0.22	enm_Latn	0.33	0.46	0.55	0.45	0.55
ctp_Latn	0.14	0.08	0.06	0.05	0.06	enq_Latn	0.07	0.12	0.05	0.05	0.07
ctu_Latn	0.10	0.09	0.11	0.06	0.27	epo_Latn	0.15	0.25	0.57	0.61	0.48
cub_Latn	0.11	0.08	0.05	0.05	0.05	eri_Latn	0.13	0.13	0.07	0.06	0.06
cuc_Latn	0.07	0.13	0.05	0.05	0.05	ese_Latn	0.09	0.13	0.06	0.05	0.06
cui_Latn	0.08	0.14	0.05	0.05	0.05	esi_Latn	0.21	0.12	0.05	0.05	0.07
cuk_Latn	0.16	0.11	0.13	0.05	0.07	esk_Latn	0.07	0.11	0.05	0.05	0.05
cul_Latn	0.09	0.12	0.07	0.05	0.05	ess_Latn	0.14	0.13	0.06	0.05	0.05
cut_Latn	0.11	0.10	0.05	0.05	0.07	est_Latn	0.07	0.46	0.68	0.56	0.47
cux_Latn	0.16	0.14	0.05	0.06	0.08	esu_Latn	0.16	0.12	0.05	0.05	0.05
cwe_Latn	0.11	0.19	0.13	0.11	0.22	etu_Latn	0.13	0.11	0.05	0.05	0.05
cwt_Latn	0.09	0.14	0.05	0.05	0.05	eus_Latn	0.09	0.18	0.26	0.25	0.23
cya_Latn	0.12	0.11	0.14	0.05	0.11	ewe_Latn	0.11	0.11	0.05	0.05	0.07
cym_Latn	0.08	0.23	0.44	0.53	0.49	ewo_Latn	0.13	0.18	0.08	0.06	0.10
czt_Latn	0.14	0.11	0.07	0.05	0.05	eza_Latn	0.07	0.09	0.05	0.05	0.06
daa_Latn	0.13	0.09	0.06	0.06	0.05	faa_Latn	0.11	0.08	0.07	0.05	0.08
dad_Latn	0.20	0.15	0.06	0.05	0.05	fai_Latn	0.13	0.11	0.06	0.05	0.05
dah_Latn	0.12	0.17	0.05	0.05	0.05	fal_Latn	0.20	0.15	0.09	0.05	0.06
dan_Latn	0.19	0.52	0.54	0.54	0.53	fao_Latn	0.09	0.27	0.32	0.36	0.48
dbq_Latn	0.13	0.07	0.06	0.05	0.05	far_Latn	0.20	0.20	0.07	0.06	0.14
ddn_Latn	0.10	0.05	0.10	0.05	0.05	fas_Arab	0.07	0.46	0.67	0.66	0.67
ded_Latn	0.07	0.09	0.06	0.05	0.06	ffm_Latn	0.13	0.11	0.05	0.05	0.07
des_Latn	0.07	0.10	0.05	0.05	0.05	fij_Latn	0.05	0.12	0.08	0.05	0.12
deu_Latn	0.15	0.38	0.52	0.52	0.46	fil_Latn	0.13	0.29	0.47	0.55	0.55
dga_Latn	0.10	0.13	0.05	0.05	0.05	fin_Latn	0.13	0.45	0.58	0.57	0.47
dgc_Latn	0.16	0.14	0.21	0.18	0.25	fon_Latn	0.10	0.09	0.05	0.05	0.05
dgi_Latn	0.12	0.07	0.05	0.06	0.06	for_Latn	0.09	0.12	0.07	0.05	0.06
dgr_Latn	0.10	0.11	0.05	0.05	0.05	fra_Latn	0.13	0.54	0.65	0.65	0.54
dgz_Latn	0.20	0.13	0.12	0.06	0.15	frd_Latn	0.08	0.13	0.06	0.05	0.09
dhm_Latn	0.17	0.17	0.10	0.05	0.10	fry_Latn	0.21	0.38	0.30	0.37	0.42
did_Latn	0.07	0.14	0.05	0.05	0.05	fub_Latn	0.17	0.16	0.10	0.05	0.12
dig_Latn	0.12	0.14	0.20	0.23	0.39	fue_Latn	0.13	0.14	0.07	0.05	0.14
dik_Latn	0.12	0.09	0.08	0.05	0.06	fuf_Latn	0.10	0.10	0.09	0.05	0.13
dip_Latn	0.15	0.15	0.05	0.05	0.06	fuh_Latn	0.12	0.09	0.05	0.06	0.05
dis_Latn	0.13	0.11	0.10	0.05	0.06	fui_Latn	0.11	0.11	0.10	0.05	0.10
dje_Latn	0.12	0.09	0.08	0.05	0.07	fuv_Latn	0.11	0.13	0.11	0.05	0.14
djk_Latn	0.14	0.14	0.08	0.05	0.28	gaa_Latn	0.12	0.13	0.05	0.05	0.05
djr_Latn	0.07	0.12	0.05	0.05	0.05	gag_Latn	0.07	0.13	0.33	0.38	0.40
dks_Latn	0.14	0.12	0.05	0.05	0.05	gah_Latn	0.07	0.15	0.05	0.05	0.05
dln_Latn	0.12	0.12	0.05	0.05	0.29	gai_Latn	0.07	0.09	0.05	0.05	0.05
dnl_Latn	0.10	0.06	0.05	0.05	0.05	gam_Latn	0.20	0.11	0.11	0.05	0.11
dnn_Latn	0.18	0.12	0.07	0.05	0.06	gaw_Latn	0.11	0.09	0.06	0.05	0.08
dob_Latn	0.08	0.08	0.10	0.05	0.07	gbi_Latn	0.10	0.11	0.06	0.05	0.08
dop_Latn	0.12	0.07	0.05	0.05	0.05	gbo_Latn	0.08	0.14	0.05	0.05	0.05
dos_Latn	0.13	0.14	0.05	0.05	0.05	gbr_Latn	0.17	0.08	0.10	0.05	0.09
dow_Latn	0.06	0.07	0.05	0.05	0.05	gde_Latn	0.10	0.05	0.06	0.05	0.05
dru_Latn	0.07	0.14	0.09	0.05	0.09	gdg_Latn	0.10	0.18	0.09	0.06	0.16
dsh_Latn	0.12	0.10	0.07	0.05	0.06	gdn_Latn	0.07	0.16	0.07	0.06	0.09
dtb_Latn	0.11	0.13	0.06	0.05	0.08	gdr_Latn	0.17	0.09	0.05	0.05	0.06
dtp_Latn	0.12	0.12	0.05	0.05	0.24	geb_Latn	0.07	0.08	0.05	0.05	0.05
dts_Latn	0.09	0.09	0.05	0.05	0.06	gej_Latn	0.09	0.10	0.05	0.05	0.08

Table C3 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glot500-m.

lan_script	BOW	mBert	XLm-R-B	XLm-R-L	Glott500-m	lan_script	BOW	mBert	XLm-R-B	XLm-R-L	Glott500-m
gfk_Latn	0.17	0.12	0.07	0.05	0.10	hlt_Latn	0.09	0.09	0.05	0.05	0.06
ghe_Deva	0.07	0.11	0.20	0.15	0.28	hmo_Latn	0.09	0.14	0.09	0.05	0.07
ghs_Latn	0.07	0.10	0.05	0.05	0.06	hmr_Latn	0.21	0.06	0.07	0.05	0.20
gid_Latn	0.10	0.05	0.05	0.05	0.08	hne_Deva	0.07	0.27	0.29	0.39	0.60
gil_Latn	0.07	0.08	0.04	0.05	0.23	hnj_Latn	0.06	0.06	0.06	0.05	0.05
giz_Latn	0.07	0.14	0.06	0.05	0.07	hnn_Latn	0.11	0.17	0.17	0.12	0.31
gjn_Latn	0.09	0.13	0.05	0.05	0.05	hns_Latn	0.13	0.12	0.14	0.12	0.19
gkn_Latn	0.09	0.16	0.05	0.05	0.14	hop_Latn	0.19	0.17	0.05	0.05	0.11
gkp_Latn	0.09	0.12	0.05	0.05	0.07	hot_Latn	0.11	0.10	0.05	0.05	0.06
gla_Latn	0.12	0.14	0.34	0.42	0.48	hra_Latn	0.13	0.13	0.07	0.05	0.26
gle_Latn	0.17	0.15	0.38	0.56	0.40	hrv_Latn	0.09	0.35	0.64	0.66	0.63
glv_Latn	0.11	0.10	0.09	0.05	0.11	hto_Latn	0.07	0.06	0.05	0.06	0.05
gmV_Latn	0.15	0.12	0.07	0.06	0.06	hub_Latn	0.07	0.13	0.06	0.05	0.06
gna_Latn	0.11	0.13	0.05	0.05	0.05	hui_Latn	0.06	0.10	0.07	0.05	0.06
gub_Latn	0.13	0.11	0.06	0.05	0.20	hun_Latn	0.08	0.38	0.70	0.66	0.52
gud_Latn	0.09	0.06	0.05	0.05	0.05	hus_Latn	0.18	0.17	0.10	0.06	0.20
gng_Latn	0.12	0.13	0.06	0.05	0.05	huv_Latn	0.07	0.11	0.06	0.05	0.06
gnn_Latn	0.07	0.10	0.05	0.05	0.08	huv_Latn	0.07	0.13	0.06	0.05	0.11
gnw_Latn	0.07	0.11	0.07	0.05	0.06	hvn_Latn	0.14	0.17	0.09	0.05	0.11
gof_Latn	0.15	0.09	0.06	0.05	0.09	hwc_Latn	0.32	0.32	0.40	0.53	0.42
gog_Latn	0.13	0.13	0.11	0.07	0.19	hye_Armn	0.07	0.39	0.60	0.64	0.65
gom_Latn	0.07	0.11	0.06	0.05	0.19	ian_Latn	0.07	0.12	0.05	0.05	0.09
gor_Latn	0.12	0.17	0.08	0.09	0.25	iba_Latn	0.11	0.27	0.26	0.24	0.54
grq_Latn	0.19	0.08	0.05	0.05	0.05	ibo_Latn	0.08	0.12	0.08	0.05	0.09
grt_Beng	0.07	0.10	0.16	0.05	0.11	icr_Latn	0.24	0.21	0.23	0.06	0.40
gso_Latn	0.07	0.09	0.05	0.05	0.05	ifa_Latn	0.10	0.15	0.06	0.05	0.32
gub_Latn	0.13	0.11	0.08	0.05	0.05	ifb_Latn	0.16	0.09	0.07	0.05	0.32
guc_Latn	0.13	0.14	0.05	0.05	0.05	ife_Latn	0.08	0.11	0.05	0.05	0.05
gud_Latn	0.11	0.11	0.05	0.05	0.05	ifk_Latn	0.14	0.14	0.07	0.05	0.21
gug_Latn	0.12	0.17	0.09	0.05	0.10	ifu_Latn	0.08	0.17	0.05	0.05	0.08
guh_Latn	0.07	0.08	0.06	0.05	0.06	ify_Latn	0.09	0.14	0.08	0.05	0.11
gui_Latn	0.09	0.09	0.09	0.05	0.07	ign_Latn	0.07	0.09	0.05	0.05	0.07
guj_Gujr	0.07	0.34	0.56	0.70	0.69	ike_Cans	0.07	0.05	0.05	0.05	0.08
guk_Ethi	0.07	0.10	0.07	0.05	0.13	ikk_Latn	0.07	0.11	0.11	0.05	0.05
guz_Latn	0.32	0.26	0.26	0.24	0.49	ikw_Latn	0.07	0.07	0.06	0.05	0.05
gum_Latn	0.07	0.09	0.05	0.05	0.06	ilb_Latn	0.09	0.12	0.14	0.09	0.16
gun_Latn	0.12	0.11	0.11	0.05	0.06	ilo_Latn	0.14	0.11	0.10	0.05	0.33
guo_Latn	0.13	0.09	0.08	0.06	0.15	imo_Latn	0.14	0.13	0.05	0.05	0.05
guq_Latn	0.07	0.15	0.16	0.05	0.06	inb_Latn	0.11	0.08	0.06	0.05	0.06
gur_Latn	0.13	0.15	0.05	0.05	0.09	ind_Latn	0.07	0.47	0.66	0.70	0.63
guu_Latn	0.11	0.10	0.06	0.05	0.06	ino_Latn	0.14	0.13	0.05	0.05	0.06
guw_Latn	0.15	0.12	0.11	0.05	0.05	iou_Latn	0.14	0.12	0.05	0.05	0.06
gux_Latn	0.07	0.10	0.07	0.05	0.07	ipi_Latn	0.07	0.14	0.04	0.05	0.05
guz_Latn	0.07	0.15	0.08	0.05	0.06	iqw_Latn	0.07	0.12	0.08	0.05	0.06
gvc_Latn	0.14	0.08	0.05	0.05	0.06	iri_Latn	0.12	0.14	0.05	0.05	0.05
gvf_Latn	0.18	0.09	0.06	0.05	0.06	irk_Latn	0.14	0.15	0.04	0.05	0.06
gvl_Latn	0.11	0.14	0.04	0.05	0.07	iry_Latn	0.08	0.14	0.11	0.16	0.20
gvn_Latn	0.07	0.12	0.05	0.05	0.09	isd_Latn	0.13	0.15	0.12	0.06	0.19
gwi_Latn	0.19	0.11	0.05	0.05	0.05	isl_Latn	0.07	0.33	0.57	0.59	0.47
gwr_Latn	0.11	0.10	0.08	0.05	0.09	ita_Latn	0.14	0.46	0.67	0.68	0.55
gya_Latn	0.10	0.10	0.05	0.05	0.06	itv_Latn	0.14	0.14	0.15	0.07	0.27
gym_Latn	0.11	0.09	0.12	0.05	0.07	iun_Latn	0.10	0.08	0.05	0.05	0.05
gyr_Latn	0.08	0.10	0.07	0.05	0.05	ivb_Latn	0.08	0.12	0.07	0.07	0.17
hae_Latn	0.09	0.15	0.15	0.31	0.22	ivv_Latn	0.11	0.13	0.07	0.05	0.19
hag_Latn	0.10	0.13	0.06	0.05	0.06	iws_Latn	0.10	0.09	0.05	0.05	0.05
hak_Latn	0.13	0.08	0.07	0.05	0.05	ixl_Latn	0.12	0.08	0.06	0.06	0.16
hat_Latn	0.06	0.17	0.08	0.06	0.39	izr_Latn	0.08	0.14	0.05	0.05	0.08
hau_Latn	0.14	0.15	0.36	0.49	0.40	izz_Latn	0.07	0.13	0.07	0.05	0.05
haw_Latn	0.12	0.11	0.05	0.05	0.19	jaa_Latn	0.10	0.12	0.06	0.05	0.08
hay_Latn	0.09	0.14	0.06	0.05	0.15	jac_Latn	0.13	0.07	0.06	0.05	0.09
hch_Latn	0.08	0.13	0.06	0.05	0.08	jae_Latn	0.07	0.07	0.05	0.05	0.05
heb_Hebr	0.07	0.36	0.15	0.31	0.24	jam_Latn	0.22	0.15	0.10	0.06	0.46
heg_Latn	0.07	0.16	0.05	0.05	0.09	jav_Latn	0.07	0.25	0.38	0.57	0.46
heh_Latn	0.10	0.15	0.11	0.09	0.09	jbu_Latn	0.12	0.12	0.08	0.05	0.08
hif_Latn	0.09	0.12	0.16	0.35	0.43	jic_Latn	0.13	0.24	0.07	0.05	0.12
hig_Latn	0.15	0.07	0.09	0.05	0.05	jiv_Latn	0.09	0.15	0.04	0.05	0.05
hil_Latn	0.14	0.23	0.26	0.24	0.53	jmc_Latn	0.15	0.10	0.05	0.06	0.09
hin_Deva	0.07	0.40	0.56	0.62	0.61	jpn_Jpan	0.07	0.37	0.62	0.56	0.50
hix_Latn	0.07	0.08	0.06	0.05	0.05	jra_Latn	0.09	0.12	0.06	0.05	0.06
hla_Latn	0.14	0.15	0.06	0.05	0.07	jun_Orya	0.07	0.05	0.11	0.06	0.12

Table C4 zero-shot score of BOW, mBERT, XLm-R-B, XLm-R-L, and Glott500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
jvn_Latn	0.07	0.35	0.36	0.52	0.49	kuf_Latn	0.13	0.15	0.07	0.05	0.05
kaa_Cyrl	0.07	0.17	0.14	0.16	0.52	kng_Latn	0.07	0.14	0.08	0.05	0.15
kab_Latn	0.11	0.14	0.07	0.06	0.13	knj_Latn	0.07	0.09	0.05	0.05	0.18
kaa_Latn	0.13	0.10	0.05	0.05	0.05	knk_Latn	0.06	0.11	0.05	0.05	0.08
kai_Latn	0.09	0.11	0.05	0.05	0.13	kno_Latn	0.10	0.10	0.05	0.05	0.07
kan_Knda	0.07	0.34	0.56	0.64	0.61	knv_Latn	0.18	0.12	0.05	0.05	0.08
kao_Latn	0.09	0.09	0.05	0.05	0.06	kog_Latn	0.11	0.12	0.06	0.05	0.05
kaq_Latn	0.09	0.16	0.06	0.05	0.09	kor_Hang	0.07	0.43	0.63	0.69	0.62
kat_Geor	0.07	0.46	0.48	0.61	0.54	kpf_Latn	0.07	0.10	0.05	0.05	0.05
kaz_Cyrl	0.07	0.32	0.57	0.66	0.57	kpg_Latn	0.22	0.15	0.05	0.05	0.15
kbc_Latn	0.18	0.07	0.05	0.05	0.05	kpj_Latn	0.07	0.10	0.04	0.05	0.07
kbb_Latn	0.09	0.13	0.07	0.05	0.07	kpq_Latn	0.15	0.14	0.04	0.05	0.06
kbn_Latn	0.09	0.15	0.11	0.06	0.07	kpr_Latn	0.13	0.10	0.10	0.05	0.08
kbo_Latn	0.11	0.15	0.04	0.05	0.06	kpz_Latn	0.07	0.09	0.09	0.05	0.11
kbp_Latn	0.10	0.08	0.05	0.05	0.05	kpw_Latn	0.14	0.10	0.05	0.05	0.05
kby_Latn	0.12	0.05	0.09	0.05	0.05	kpx_Latn	0.07	0.13	0.09	0.05	0.05
kbr_Latn	0.08	0.13	0.05	0.05	0.07	kpz_Latn	0.09	0.12	0.05	0.05	0.09
kcg_Latn	0.13	0.12	0.05	0.05	0.05	kqc_Latn	0.08	0.09	0.11	0.05	0.08
kck_Latn	0.08	0.13	0.09	0.05	0.18	kqe_Latn	0.13	0.16	0.13	0.12	0.33
kdc_Latn	0.13	0.14	0.20	0.19	0.21	kqo_Latn	0.07	0.09	0.05	0.05	0.05
kde_Latn	0.14	0.16	0.12	0.07	0.15	kqp_Latn	0.14	0.14	0.05	0.05	0.06
kdi_Latn	0.07	0.16	0.05	0.05	0.08	kqs_Latn	0.10	0.13	0.05	0.05	0.06
kdj_Latn	0.07	0.13	0.05	0.05	0.05	kqy_Ethi	0.07	0.13	0.06	0.05	0.05
kdl_Latn	0.07	0.11	0.07	0.05	0.09	krc_Cyrl	0.07	0.17	0.17	0.16	0.48
kdp_Latn	0.10	0.11	0.10	0.05	0.07	kri_Latn	0.15	0.16	0.05	0.05	0.19
kek_Latn	0.15	0.08	0.05	0.06	0.27	krj_Latn	0.11	0.21	0.33	0.28	0.35
ken_Latn	0.10	0.08	0.05	0.05	0.05	krl_Latn	0.07	0.34	0.40	0.40	0.41
keo_Latn	0.11	0.08	0.06	0.05	0.11	kru_Deva	0.07	0.12	0.08	0.05	0.11
ker_Latn	0.09	0.04	0.05	0.05	0.05	ksb_Latn	0.12	0.16	0.12	0.12	0.21
kew_Latn	0.13	0.14	0.05	0.05	0.06	ksc_Latn	0.09	0.12	0.07	0.05	0.11
kez_Latn	0.13	0.10	0.05	0.05	0.05	ksd_Latn	0.15	0.14	0.06	0.05	0.12
kff_Telu	0.07	0.14	0.24	0.20	0.20	kse_Latn	0.10	0.07	0.05	0.05	0.06
kgl_Latn	0.08	0.10	0.05	0.05	0.05	ksr_Latn	0.08	0.08	0.05	0.05	0.06
kgk_Latn	0.07	0.10	0.06	0.05	0.05	kss_Latn	0.12	0.10	0.05	0.05	0.05
kgp_Latn	0.07	0.14	0.09	0.05	0.09	ksw_Mymr	0.07	0.08	0.05	0.05	0.06
kgr_Latn	0.14	0.20	0.06	0.05	0.13	ktb_Ethi	0.07	0.05	0.07	0.05	0.10
kha_Latn	0.12	0.07	0.07	0.05	0.06	ktj_Latn	0.04	0.05	0.05	0.05	0.05
khh_Latn	0.09	0.15	0.07	0.05	0.08	kto_Latn	0.07	0.14	0.09	0.05	0.05
khn_Khm	0.07	0.05	0.55	0.62	0.55	ktu_Latn	0.10	0.11	0.11	0.06	0.19
khq_Latn	0.12	0.11	0.10	0.05	0.09	kua_Latn	0.11	0.11	0.11	0.08	0.12
khs_Latn	0.14	0.09	0.06	0.05	0.05	kub_Latn	0.09	0.14	0.05	0.05	0.05
khy_Latn	0.08	0.09	0.07	0.07	0.14	kud_Latn	0.07	0.10	0.06	0.05	0.05
khz_Latn	0.12	0.16	0.06	0.05	0.05	kue_Latn	0.07	0.11	0.06	0.05	0.07
kia_Latn	0.13	0.19	0.06	0.05	0.23	kuj_Latn	0.12	0.12	0.05	0.05	0.05
kij_Latn	0.07	0.14	0.07	0.05	0.06	kum_Cyrl	0.07	0.16	0.13	0.24	0.45
kik_Latn	0.14	0.15	0.05	0.05	0.05	kup_Latn	0.18	0.15	0.08	0.05	0.07
kin_Latn	0.14	0.13	0.14	0.06	0.23	kus_Latn	0.12	0.09	0.10	0.05	0.05
kir_Cyrl	0.07	0.20	0.65	0.65	0.61	kvg_Latn	0.11	0.09	0.06	0.05	0.06
kix_Latn	0.08	0.12	0.07	0.05	0.05	kvj_Latn	0.17	0.13	0.06	0.05	0.05
kjb_Latn	0.15	0.11	0.05	0.05	0.23	kvn_Latn	0.12	0.09	0.08	0.05	0.06
kje_Latn	0.09	0.18	0.06	0.05	0.06	kwd_Latn	0.19	0.13	0.09	0.05	0.12
kjh_Cyrl	0.07	0.18	0.11	0.17	0.36	kwf_Latn	0.21	0.17	0.09	0.07	0.16
kjs_Latn	0.13	0.10	0.07	0.05	0.05	kwi_Latn	0.11	0.17	0.09	0.05	0.09
kki_Latn	0.16	0.17	0.14	0.10	0.14	kwj_Latn	0.10	0.12	0.06	0.05	0.05
kkj_Latn	0.09	0.16	0.06	0.05	0.06	kxc_Ethi	0.07	0.09	0.07	0.05	0.05
kke_Latn	0.07	0.14	0.15	0.11	0.19	kxm_Thai	0.07	0.08	0.14	0.06	0.08
kkn_Latn	0.10	0.10	0.05	0.05	0.12	kxw_Latn	0.06	0.07	0.06	0.05	0.05
kky_Latn	0.09	0.14	0.13	0.05	0.09	kyc_Latn	0.07	0.11	0.06	0.05	0.06
kma_Latn	0.12	0.08	0.05	0.05	0.05	kyf_Latn	0.09	0.13	0.05	0.05	0.05
kmd_Latn	0.10	0.11	0.06	0.05	0.09	kyg_Latn	0.08	0.09	0.06	0.05	0.05
kmg_Latn	0.08	0.08	0.05	0.05	0.05	kyq_Latn	0.10	0.12	0.07	0.05	0.05
knh_Latn	0.07	0.10	0.05	0.05	0.05	kyu_Mymr	0.07	0.09	0.05	0.05	0.05
knk_Latn	0.10	0.10	0.06	0.05	0.14	kyz_Latn	0.17	0.10	0.05	0.05	0.05
knn_Latn	0.12	0.09	0.05	0.05	0.19	kze_Latn	0.08	0.11	0.04	0.05	0.06
kno_Latn	0.10	0.09	0.05	0.06	0.06	kzf_Latn	0.12	0.18	0.10	0.06	0.15
knr_Cyrl	0.07	0.09	0.07	0.05	0.24	lac_Latn	0.16	0.05	0.06	0.05	0.11
kns_Latn	0.13	0.08	0.04	0.05	0.07	lai_Latn	0.16	0.13	0.07	0.08	0.19
knu_Latn	0.07	0.17	0.10	0.05	0.08	laj_Latn	0.10	0.11	0.07	0.06	0.09
kny_Latn	0.12	0.08	0.05	0.05	0.05	lam_Latn	0.09	0.14	0.07	0.07	0.16
kne_Latn	0.15	0.13	0.12	0.04	0.09	lao_Lao	0.07	0.05	0.58	0.67	0.61

Table C5 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glott500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
lap_Latn	0.14	0.15	0.06	0.05	0.08	mbb_Latn	0.11	0.20	0.10	0.05	0.10
las_Latn	0.09	0.09	0.05	0.05	0.05	mbc_Latn	0.12	0.13	0.05	0.05	0.05
lat_Latn	0.14	0.30	0.55	0.62	0.56	mbd_Latn	0.13	0.12	0.11	0.05	0.10
lav_Latn	0.08	0.34	0.62	0.55	0.52	mbf_Latn	0.07	0.31	0.49	0.57	0.56
law_Latn	0.09	0.09	0.06	0.05	0.09	mbh_Latn	0.15	0.15	0.07	0.05	0.09
lbk_Latn	0.12	0.10	0.09	0.05	0.14	mbi_Latn	0.13	0.17	0.08	0.05	0.06
lcm_Latn	0.16	0.20	0.05	0.06	0.15	mbj_Latn	0.16	0.14	0.08	0.05	0.06
lcp_Thai	0.07	0.08	0.06	0.05	0.05	mbk_Latn	0.07	0.11	0.05	0.05	0.05
ldi_Latn	0.14	0.12	0.07	0.05	0.19	mbs_Latn	0.11	0.12	0.17	0.13	0.19
lee_Latn	0.08	0.05	0.07	0.05	0.05	mbt_Latn	0.14	0.12	0.07	0.05	0.09
lef_Latn	0.05	0.13	0.06	0.05	0.05	mca_Latn	0.16	0.10	0.05	0.05	0.06
leh_Latn	0.09	0.14	0.08	0.07	0.15	mcb_Latn	0.07	0.11	0.05	0.05	0.06
lem_Latn	0.07	0.09	0.05	0.05	0.06	mcd_Latn	0.05	0.09	0.05	0.05	0.06
leu_Latn	0.12	0.14	0.05	0.05	0.07	mcf_Latn	0.07	0.10	0.06	0.05	0.05
lew_Latn	0.07	0.13	0.08	0.05	0.16	mck_Latn	0.13	0.15	0.11	0.06	0.15
lex_Latn	0.13	0.10	0.08	0.05	0.05	mcn_Latn	0.09	0.10	0.07	0.06	0.10
lgg_Latn	0.09	0.19	0.05	0.05	0.13	mco_Latn	0.05	0.09	0.05	0.05	0.13
lgl_Latn	0.20	0.14	0.06	0.06	0.12	mcp_Latn	0.09	0.05	0.05	0.05	0.05
lgn_Latn	0.12	0.11	0.06	0.06	0.09	mcq_Latn	0.07	0.12	0.08	0.05	0.05
lhi_Latn	0.09	0.12	0.05	0.05	0.10	mdu_Latn	0.10	0.20	0.07	0.05	0.06
lhm_Latn	0.12	0.08	0.05	0.05	0.05	mda_Latn	0.06	0.07	0.05	0.05	0.05
lhu_Latn	0.09	0.08	0.06	0.05	0.06	mdy_Ethi	0.07	0.09	0.05	0.05	0.15
lia_Latn	0.18	0.16	0.05	0.05	0.05	med_Latn	0.07	0.09	0.06	0.05	0.07
lid_Latn	0.16	0.09	0.08	0.05	0.06	mee_Latn	0.11	0.12	0.05	0.05	0.06
lif_Deva	0.07	0.07	0.10	0.05	0.13	mej_Latn	0.07	0.11	0.09	0.05	0.08
lin_Latn	0.12	0.10	0.08	0.04	0.13	mek_Latn	0.08	0.10	0.08	0.05	0.14
lip_Latn	0.08	0.12	0.06	0.05	0.07	men_Latn	0.11	0.13	0.05	0.05	0.05
lis_Lisu	0.07	0.08	0.05	0.05	0.06	meq_Latn	0.10	0.07	0.07	0.05	0.05
lit_Latn	0.07	0.29	0.56	0.60	0.54	met_Latn	0.19	0.11	0.05	0.05	0.06
ljp_Latn	0.07	0.29	0.33	0.30	0.39	meu_Latn	0.10	0.14	0.10	0.05	0.08
llg_Latn	0.07	0.09	0.13	0.05	0.07	mfe_Latn	0.09	0.15	0.15	0.05	0.36
lln_Latn	0.10	0.09	0.05	0.05	0.05	mfi_Latn	0.07	0.07	0.06	0.05	0.07
lmk_Latn	0.14	0.11	0.07	0.05	0.05	mfl_Latn	0.15	0.07	0.06	0.05	0.06
lmp_Latn	0.09	0.12	0.05	0.05	0.05	mfk_Latn	0.09	0.16	0.05	0.05	0.05
lnd_Latn	0.09	0.13	0.10	0.06	0.15	mfq_Latn	0.08	0.05	0.05	0.05	0.06
lob_Latn	0.07	0.10	0.05	0.05	0.04	mfy_Latn	0.11	0.15	0.07	0.05	0.06
loe_Latn	0.10	0.21	0.10	0.08	0.23	mfx_Latn	0.13	0.09	0.05	0.05	0.05
log_Latn	0.11	0.11	0.05	0.05	0.05	ngb_Latn	0.13	0.10	0.04	0.05	0.08
lok_Latn	0.13	0.12	0.05	0.05	0.05	ngo_Latn	0.15	0.05	0.05	0.05	0.05
lol_Latn	0.07	0.09	0.06	0.05	0.09	ngr_Latn	0.17	0.13	0.10	0.07	0.21
lom_Latn	0.11	0.07	0.05	0.05	0.05	mhi_Latn	0.12	0.12	0.08	0.05	0.06
loq_Latn	0.08	0.13	0.05	0.05	0.06	mhl_Latn	0.10	0.10	0.05	0.05	0.05
loz_Latn	0.18	0.14	0.06	0.05	0.29	mhr_Cyrl	0.07	0.17	0.10	0.05	0.26
lsi_Latn	0.13	0.08	0.05	0.05	0.05	mhx_Latn	0.11	0.12	0.05	0.05	0.05
lsm_Latn	0.11	0.16	0.08	0.07	0.08	mhy_Latn	0.12	0.20	0.21	0.15	0.26
ltz_Latn	0.15	0.34	0.22	0.20	0.41	mib_Latn	0.09	0.13	0.07	0.06	0.13
luc_Latn	0.07	0.09	0.11	0.05	0.05	mic_Latn	0.10	0.13	0.08	0.05	0.06
lug_Latn	0.07	0.13	0.08	0.05	0.22	mie_Latn	0.08	0.17	0.06	0.05	0.12
luo_Latn	0.12	0.12	0.05	0.05	0.15	mif_Latn	0.09	0.09	0.07	0.05	0.07
lus_Latn	0.17	0.14	0.10	0.05	0.09	mig_Latn	0.13	0.19	0.05	0.05	0.07
lwo_Latn	0.12	0.12	0.05	0.05	0.05	mih_Latn	0.08	0.13	0.04	0.05	0.07
lhw_Latn	0.11	0.12	0.06	0.05	0.05	mil_Latn	0.10	0.11	0.05	0.05	0.06
lzh_Hani	0.07	0.24	0.54	0.50	0.59	mim_Latn	0.11	0.15	0.05	0.05	0.06
maa_Latn	0.13	0.14	0.05	0.05	0.05	min_Latn	0.08	0.19	0.27	0.26	0.43
mad_Latn	0.10	0.22	0.23	0.19	0.40	mio_Latn	0.09	0.08	0.15	0.07	0.14
maf_Latn	0.11	0.18	0.06	0.05	0.05	mip_Latn	0.06	0.10	0.05	0.05	0.11
mag_Deva	0.07	0.22	0.38	0.32	0.49	miq_Latn	0.09	0.16	0.05	0.05	0.08
mah_Latn	0.16	0.12	0.05	0.05	0.14	mir_Latn	0.06	0.09	0.06	0.05	0.14
mai_Deva	0.07	0.23	0.31	0.43	0.65	mit_Latn	0.06	0.09	0.07	0.06	0.12
maj_Latn	0.09	0.09	0.05	0.05	0.05	miy_Latn	0.07	0.10	0.05	0.05	0.08
mak_Latn	0.10	0.18	0.10	0.06	0.18	miz_Latn	0.09	0.14	0.05	0.05	0.05
mal_Mlym	0.07	0.12	0.07	0.05	0.06	mjc_Latn	0.13	0.13	0.05	0.05	0.07
mam_Latn	0.12	0.11	0.04	0.04	0.25	mjjw_Latn	0.08	0.09	0.08	0.05	0.05
maq_Latn	0.12	0.15	0.05	0.06	0.05	mkd_Cyrl	0.07	0.47	0.74	0.70	0.67
mar_Deva	0.07	0.30	0.57	0.61	0.59	mkl_Latn	0.11	0.05	0.06	0.05	0.05
mas_Latn	0.07	0.17	0.09	0.06	0.04	mkn_Latn	0.07	0.23	0.28	0.35	0.44
mau_Latn	0.07	0.08	0.05	0.05	0.05	mks_Latn	0.10	0.15	0.05	0.05	0.05
mav_Latn	0.14	0.12	0.07	0.05	0.05	mlg_Latn	0.12	0.08	0.37	0.45	0.46
maw_Latn	0.18	0.11	0.05	0.05	0.05	mlh_Latn	0.10	0.10	0.05	0.05	0.05
maz_Latn	0.10	0.15	0.05	0.05	0.10	mlp_Latn	0.07	0.20	0.06	0.05	0.08

Table C6 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glot500-m.

	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m		lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
	mlt_Latn	0.11	0.16	0.05	0.06	0.29		mzm_Latn	0.09	0.09	0.05	0.05	0.05
	mmn_Latn	0.17	0.19	0.18	0.21	0.32		mzw_Latn	0.05	0.09	0.05	0.05	0.06
	mmo_Latn	0.17	0.09	0.09	0.05	0.05		nab_Latn	0.07	0.14	0.05	0.05	0.05
	mnx_Latn	0.14	0.11	0.05	0.05	0.06		naf_Latn	0.07	0.15	0.05	0.05	0.06
	mna_Latn	0.11	0.08	0.05	0.05	0.05		nak_Latn	0.11	0.12	0.04	0.05	0.08
	mnb_Latn	0.10	0.17	0.06	0.05	0.16		nan_Latn	0.14	0.11	0.05	0.05	0.06
	mnf_Latn	0.11	0.13	0.05	0.05	0.06		naq_Latn	0.09	0.10	0.05	0.05	0.07
	mnh_Latn	0.07	0.17	0.07	0.05	0.09		nas_Latn	0.07	0.09	0.11	0.05	0.09
	mnk_Latn	0.09	0.17	0.05	0.05	0.07		nav_Latn	0.19	0.09	0.05	0.05	0.05
	mnx_Latn	0.11	0.15	0.08	0.06	0.05		naw_Latn	0.08	0.10	0.05	0.05	0.05
	moa_Latn	0.08	0.04	0.06	0.05	0.05		nbc_Latn	0.09	0.12	0.06	0.05	0.07
	moc_Latn	0.08	0.13	0.06	0.05	0.05		nbe_Latn	0.17	0.12	0.06	0.06	0.07
	mog_Latn	0.16	0.20	0.13	0.07	0.21		nbl_Latn	0.09	0.13	0.15	0.21	0.29
	mop_Latn	0.20	0.10	0.07	0.06	0.27		nbu_Latn	0.15	0.09	0.05	0.05	0.05
	mor_Latn	0.14	0.11	0.05	0.05	0.05		nca_Latn	0.07	0.11	0.06	0.06	0.06
	mos_Latn	0.11	0.11	0.06	0.05	0.06		nch_Latn	0.10	0.12	0.07	0.05	0.06
	mox_Latn	0.12	0.15	0.07	0.05	0.05		ncj_Latn	0.14	0.10	0.05	0.05	0.07
	mpg_Latn	0.12	0.09	0.05	0.05	0.05		ncl_Latn	0.10	0.09	0.06	0.09	0.13
	mpm_Latn	0.04	0.15	0.05	0.05	0.05		ncq_Lao	0.07	0.05	0.11	0.04	0.10
	mps_Latn	0.15	0.16	0.05	0.06	0.07		nct_Latn	0.12	0.09	0.06	0.05	0.06
	mpt_Latn	0.13	0.11	0.07	0.05	0.07		ncu_Latn	0.06	0.09	0.05	0.05	0.05
	mpx_Latn	0.09	0.10	0.07	0.05	0.05		ndc_Latn	0.07	0.15	0.10	0.07	0.16
	mqb_Latn	0.11	0.09	0.04	0.05	0.05		nde_Latn	0.09	0.13	0.15	0.21	0.29
	mqj_Latn	0.11	0.18	0.12	0.05	0.16		ndi_Latn	0.11	0.10	0.06	0.05	0.05
	mqy_Latn	0.11	0.16	0.13	0.05	0.11		ndj_Latn	0.13	0.11	0.06	0.05	0.12
	mri_Latn	0.16	0.09	0.09	0.05	0.19		ndo_Latn	0.11	0.11	0.09	0.05	0.16
	mrw_Latn	0.09	0.19	0.10	0.14	0.31		ndp_Latn	0.10	0.11	0.10	0.05	0.07
	msa_Latn	0.08	0.22	0.42	0.42	0.52		nds_Latn	0.15	0.19	0.14	0.07	0.27
	msb_Latn	0.12	0.21	0.28	0.24	0.49		ndy_Latn	0.07	0.14	0.07	0.06	0.14
	mse_Latn	0.12	0.09	0.08	0.05	0.05		ndz_Latn	0.09	0.15	0.05	0.05	0.05
	msk_Latn	0.09	0.14	0.09	0.10	0.28		neb_Latn	0.12	0.07	0.05	0.05	0.05
	msm_Latn	0.12	0.10	0.07	0.06	0.21		nep_Deva	0.07	0.32	0.62	0.64	0.68
	msy_Latn	0.07	0.09	0.06	0.05	0.06		nfa_Latn	0.07	0.09	0.06	0.05	0.05
	mta_Latn	0.12	0.10	0.05	0.05	0.05		nfr_Latn	0.15	0.11	0.07	0.05	0.05
	mtg_Latn	0.11	0.09	0.05	0.05	0.05		ngc_Latn	0.11	0.14	0.07	0.05	0.14
	mti_Latn	0.14	0.14	0.08	0.08	0.15		ngp_Latn	0.13	0.17	0.16	0.12	0.19
	mtj_Latn	0.08	0.10	0.08	0.05	0.06		ngu_Latn	0.06	0.09	0.05	0.06	0.15
	mtl_Latn	0.11	0.14	0.05	0.05	0.05		nhd_Latn	0.12	0.17	0.09	0.05	0.10
	ntp_Latn	0.11	0.12	0.05	0.05	0.05		nhe_Latn	0.10	0.13	0.07	0.05	0.08
	mua_Latn	0.16	0.10	0.05	0.05	0.06		nhg_Latn	0.10	0.12	0.05	0.05	0.14
	mug_Latn	0.13	0.11	0.05	0.06	0.07		nhi_Latn	0.12	0.10	0.06	0.05	0.08
	muh_Latn	0.12	0.18	0.15	0.05	0.05		nho_Latn	0.16	0.17	0.07	0.05	0.12
	mup_Deva	0.07	0.28	0.35	0.32	0.49		nhr_Latn	0.17	0.14	0.05	0.05	0.07
	mur_Latn	0.14	0.12	0.05	0.05	0.08		nhs_Latn	0.16	0.10	0.05	0.05	0.05
	mux_Latn	0.12	0.11	0.06	0.05	0.05		nhw_Latn	0.08	0.14	0.07	0.05	0.06
	muy_Latn	0.11	0.07	0.05	0.05	0.05		nhx_Latn	0.13	0.14	0.08	0.05	0.19
	mva_Latn	0.07	0.15	0.07	0.05	0.07		nhy_Latn	0.14	0.16	0.05	0.06	0.15
	mvn_Latn	0.12	0.09	0.05	0.05	0.05		nii_Latn	0.14	0.09	0.05	0.05	0.05
	mvp_Latn	0.11	0.12	0.15	0.05	0.22		nij_Latn	0.09	0.23	0.18	0.16	0.23
	mwm_Latn	0.12	0.08	0.05	0.05	0.05		nim_Latn	0.07	0.12	0.06	0.05	0.06
	nwq_Latn	0.10	0.10	0.06	0.05	0.05		nin_Latn	0.07	0.13	0.08	0.05	0.07
	nww_Latn	0.07	0.14	0.10	0.05	0.13		niq_Latn	0.09	0.10	0.05	0.05	0.07
	mxp_Latn	0.09	0.14	0.05	0.05	0.06		niy_Latn	0.11	0.05	0.08	0.05	0.05
	mxq_Latn	0.10	0.12	0.05	0.05	0.06		njb_Latn	0.17	0.13	0.05	0.05	0.05
	mxr_Latn	0.09	0.06	0.05	0.05	0.10		njm_Latn	0.16	0.09	0.06	0.05	0.06
	mxs_Latn	0.13	0.12	0.04	0.05	0.07		njn_Latn	0.09	0.12	0.05	0.05	0.05
	mxv_Latn	0.10	0.16	0.05	0.05	0.16		njo_Latn	0.12	0.11	0.05	0.05	0.06
	mya_Mymr	0.07	0.26	0.42	0.61	0.51		njs_Latn	0.08	0.13	0.05	0.05	0.05
	myb_Latn	0.07	0.13	0.07	0.05	0.09		nkf_Latn	0.13	0.16	0.06	0.05	0.06
	myk_Latn	0.07	0.12	0.05	0.05	0.07		nki_Latn	0.10	0.13	0.05	0.05	0.26
	myl_Latn	0.07	0.12	0.09	0.05	0.06		nko_Latn	0.10	0.10	0.05	0.05	0.05
	myv_Cyrl	0.07	0.08	0.08	0.05	0.19		nkc_Latn	0.11	0.12	0.05	0.05	0.05
	myw_Latn	0.07	0.15	0.06	0.05	0.05		nld_Latn	0.28	0.43	0.60	0.58	0.53
	myx_Latn	0.10	0.12	0.04	0.05	0.10		nlg_Latn	0.20	0.21	0.07	0.09	0.21
	myy_Latn	0.07	0.08	0.09	0.05	0.06		nma_Latn	0.07	0.12	0.08	0.05	0.05
	mza_Latn	0.10	0.13	0.06	0.05	0.05		nmf_Latn	0.08	0.12	0.05	0.05	0.06
	mzh_Latn	0.08	0.19	0.08	0.05	0.24		nmh_Latn	0.09	0.10	0.05	0.06	0.06
	mzk_Latn	0.14	0.14	0.08	0.06	0.07		nmo_Latn	0.10	0.10	0.06	0.05	0.06
	mzl_Latn	0.10	0.09	0.06	0.05	0.05		nmz_Latn	0.15	0.12	0.08	0.05	0.10
								nnb_Latn	0.10	0.14	0.07	0.05	0.10

Table C7 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glott500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
lan_script						lan_script					
nng_Latn	0.07	0.09	0.07	0.05	0.06	oym_Latn	0.07	0.12	0.05	0.05	0.05
nmh_Latn	0.08	0.14	0.07	0.05	0.08	ozm_Latn	0.13	0.06	0.06	0.05	0.05
nnl_Latn	0.12	0.12	0.07	0.05	0.06	pab_Latn	0.12	0.05	0.05	0.05	0.05
nno_Latn	0.15	0.46	0.58	0.56	0.43	pad_Latn	0.13	0.15	0.06	0.05	0.06
nnp_Latn	0.07	0.08	0.07	0.05	0.05	pag_Latn	0.14	0.14	0.20	0.17	0.33
nnq_Latn	0.14	0.15	0.11	0.10	0.14	pah_Latn	0.09	0.15	0.06	0.05	0.05
nnp_Latn	0.07	0.05	0.05	0.05	0.05	pam_Latn	0.13	0.18	0.11	0.11	0.38
noa_Latn	0.07	0.08	0.05	0.06	0.05	pan_Guru	0.07	0.31	0.58	0.67	0.69
nob_Latn	0.16	0.38	0.59	0.60	0.56	pao_Latn	0.10	0.13	0.07	0.05	0.08
nod_Thai	0.07	0.09	0.47	0.50	0.50	pap_Latn	0.15	0.31	0.30	0.23	0.52
nog_Cyrl	0.07	0.16	0.18	0.38	0.41	pau_Latn	0.16	0.18	0.06	0.05	0.21
nop_Latn	0.09	0.15	0.05	0.05	0.05	pbb_Latn	0.17	0.12	0.07	0.05	0.07
nor_Latn	0.16	0.38	0.60	0.60	0.55	pbc_Latn	0.17	0.12	0.05	0.05	0.05
not_Latn	0.07	0.09	0.13	0.06	0.11	pbi_Latn	0.13	0.06	0.05	0.05	0.07
nou_Latn	0.16	0.11	0.11	0.06	0.13	pbl_Latn	0.10	0.16	0.13	0.05	0.26
nph_Latn	0.08	0.10	0.09	0.05	0.05	pck_Latn	0.12	0.14	0.06	0.05	0.19
npj_Deva	0.07	0.32	0.59	0.66	0.67	pcm_Latn	0.19	0.18	0.30	0.29	0.45
npl_Latn	0.10	0.09	0.05	0.07	0.18	pcd_Latn	0.19	0.14	0.14	0.15	0.27
npo_Latn	0.13	0.09	0.07	0.05	0.05	pdL_Latn	0.17	0.18	0.17	0.12	0.34
npY_Latn	0.09	0.13	0.11	0.05	0.07	pes_Arab	0.07	0.42	0.66	0.66	0.63
nre_Latn	0.10	0.15	0.07	0.05	0.07	pez_Latn	0.08	0.23	0.09	0.05	0.10
nri_Latn	0.11	0.12	0.09	0.05	0.09	pfe_Latn	0.10	0.05	0.05	0.05	0.05
nsa_Latn	0.07	0.12	0.09	0.05	0.06	pib_Latn	0.07	0.11	0.04	0.05	0.06
nse_Latn	0.12	0.17	0.13	0.07	0.23	pio_Latn	0.07	0.09	0.06	0.05	0.12
nsM_Latn	0.13	0.07	0.06	0.05	0.06	pir_Latn	0.10	0.11	0.06	0.05	0.05
nsn_Latn	0.15	0.09	0.06	0.07	0.12	pis_Latn	0.21	0.11	0.12	0.06	0.20
nso_Latn	0.11	0.13	0.12	0.05	0.27	pjt_Latn	0.07	0.09	0.05	0.05	0.08
nst_Latn	0.18	0.10	0.05	0.05	0.06	pkb_Latn	0.11	0.15	0.12	0.07	0.28
nsu_Latn	0.13	0.10	0.06	0.05	0.12	plg_Latn	0.16	0.13	0.08	0.05	0.08
ntp_Latn	0.07	0.10	0.05	0.05	0.04	pls_Latn	0.07	0.19	0.07	0.14	0.27
ntr_Latn	0.07	0.12	0.05	0.05	0.05	plt_Latn	0.12	0.05	0.38	0.54	0.50
ntu_Latn	0.07	0.08	0.06	0.05	0.05	plu_Latn	0.13	0.08	0.05	0.05	0.05
nui_Latn	0.11	0.14	0.06	0.05	0.07	plw_Latn	0.14	0.19	0.10	0.06	0.19
mus_Latn	0.13	0.10	0.05	0.05	0.05	pma_Latn	0.14	0.16	0.07	0.05	0.06
nuy_Latn	0.23	0.10	0.05	0.05	0.05	pmf_Latn	0.11	0.22	0.10	0.09	0.20
nvM_Latn	0.07	0.11	0.05	0.05	0.05	pmx_Latn	0.09	0.08	0.06	0.06	0.06
nwb_Latn	0.14	0.06	0.05	0.05	0.05	pne_Latn	0.08	0.23	0.09	0.05	0.11
nwi_Latn	0.15	0.13	0.05	0.05	0.07	pny_Latn	0.08	0.05	0.05	0.05	0.05
nwx_Deva	0.07	0.16	0.18	0.14	0.29	poe_Latn	0.13	0.13	0.05	0.05	0.06
nxd_Latn	0.07	0.09	0.07	0.05	0.07	poh_Latn	0.11	0.09	0.12	0.05	0.37
nya_Latn	0.07	0.14	0.08	0.06	0.26	poi_Latn	0.12	0.15	0.05	0.07	0.12
nyf_Latn	0.15	0.19	0.21	0.17	0.25	pol_Latn	0.09	0.48	0.60	0.65	0.61
nyN_Latn	0.09	0.11	0.06	0.05	0.20	pon_Latn	0.14	0.21	0.08	0.05	0.08
nyo_Latn	0.07	0.16	0.05	0.05	0.15	por_Latn	0.16	0.52	0.57	0.64	0.61
nyy_Latn	0.11	0.16	0.08	0.05	0.09	pos_Latn	0.12	0.17	0.06	0.06	0.27
nza_Latn	0.07	0.10	0.05	0.05	0.05	poy_Latn	0.14	0.18	0.08	0.05	0.07
nzi_Latn	0.09	0.16	0.05	0.05	0.05	ppk_Latn	0.15	0.15	0.06	0.04	0.16
nzM_Latn	0.11	0.09	0.08	0.06	0.06	ppo_Latn	0.10	0.18	0.05	0.05	0.05
obo_Latn	0.15	0.12	0.05	0.05	0.07	pps_Latn	0.10	0.11	0.06	0.05	0.08
obj_Cans	0.07	0.12	0.05	0.05	0.06	prf_Latn	0.12	0.20	0.15	0.13	0.26
oji_Latn	0.11	0.09	0.05	0.05	0.07	pri_Latn	0.07	0.10	0.05	0.05	0.05
ojis_Latn	0.07	0.08	0.05	0.05	0.06	prk_Latn	0.09	0.13	0.06	0.05	0.10
oku_Latn	0.12	0.11	0.05	0.05	0.05	prq_Latn	0.07	0.08	0.05	0.05	0.05
okv_Latn	0.13	0.22	0.14	0.08	0.13	prs_Arab	0.07	0.43	0.66	0.64	0.64
old_Latn	0.13	0.09	0.08	0.06	0.06	pse_Latn	0.07	0.28	0.36	0.38	0.39
omb_Latn	0.17	0.16	0.10	0.06	0.06	pss_Latn	0.10	0.13	0.06	0.05	0.08
omw_Latn	0.07	0.08	0.05	0.05	0.05	ptp_Latn	0.10	0.11	0.05	0.05	0.05
ong_Latn	0.07	0.17	0.07	0.05	0.06	ptu_Latn	0.11	0.15	0.14	0.05	0.20
ons_Latn	0.11	0.09	0.05	0.05	0.05	pua_Latn	0.08	0.09	0.09	0.05	0.15
ood_Latn	0.16	0.11	0.05	0.05	0.05	pui_Latn	0.09	0.14	0.05	0.06	0.06
opm_Latn	0.07	0.14	0.07	0.05	0.05	pwg_Latn	0.18	0.14	0.06	0.08	0.12
ori_Orya	0.07	0.04	0.58	0.75	0.65	pwW_Thai	0.07	0.08	0.10	0.05	0.05
ory_Orya	0.07	0.04	0.56	0.75	0.64	pxm_Latn	0.08	0.14	0.06	0.05	0.05
oss_Cyrl	0.07	0.10	0.07	0.05	0.11	qub_Latn	0.08	0.12	0.06	0.06	0.17
otd_Latn	0.07	0.25	0.12	0.11	0.14	que_Latn	0.18	0.14	0.07	0.05	0.37
ote_Latn	0.08	0.07	0.05	0.05	0.06	quf_Latn	0.07	0.10	0.05	0.05	0.06
otm_Latn	0.10	0.08	0.05	0.05	0.05	qug_Latn	0.07	0.11	0.09	0.05	0.12
otn_Latn	0.09	0.11	0.05	0.05	0.05	quh_Latn	0.07	0.12	0.07	0.05	0.30
otq_Latn	0.14	0.08	0.06	0.05	0.06	qul_Latn	0.07	0.14	0.06	0.07	0.32
ots_Latn	0.11	0.10	0.05	0.05	0.10	qup_Latn	0.07	0.13	0.05	0.05	0.13

Table C8 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glot500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
quw_Latn	0.07	0.10	0.07	0.05	0.18	shp_Latn	0.07	0.12	0.06	0.05	0.05
quy_Latn	0.07	0.11	0.07	0.06	0.27	shu_Latn	0.09	0.20	0.16	0.11	0.19
quz_Latn	0.07	0.10	0.07	0.05	0.24	sig_Latn	0.13	0.08	0.05	0.05	0.05
qva_Latn	0.07	0.10	0.07	0.05	0.18	sil_Latn	0.14	0.07	0.05	0.05	0.05
qvc_Latn	0.09	0.11	0.06	0.05	0.05	sim_Latn	0.08	0.10	0.06	0.05	0.07
qve_Latn	0.09	0.13	0.06	0.05	0.33	sin_Sinh	0.07	0.16	0.51	0.67	0.57
qvh_Latn	0.12	0.12	0.05	0.07	0.24	sja_Latn	0.10	0.10	0.05	0.05	0.05
qvi_Latn	0.06	0.12	0.06	0.05	0.10	sld_Latn	0.14	0.10	0.05	0.05	0.05
qvm_Latn	0.07	0.13	0.06	0.05	0.19	slk_Latn	0.09	0.48	0.69	0.64	0.56
qvn_Latn	0.07	0.10	0.05	0.06	0.14	sil_Latn	0.07	0.11	0.07	0.05	0.08
qvo_Latn	0.10	0.11	0.06	0.05	0.08	slv_Latn	0.17	0.50	0.63	0.60	0.60
qvs_Latn	0.09	0.10	0.05	0.05	0.18	sme_Latn	0.15	0.17	0.09	0.05	0.14
qvw_Latn	0.09	0.10	0.05	0.05	0.13	smk_Latn	0.10	0.10	0.08	0.06	0.27
qvz_Latn	0.09	0.10	0.06	0.05	0.13	sml_Latn	0.13	0.12	0.17	0.10	0.23
qwh_Latn	0.06	0.14	0.09	0.05	0.22	smo_Latn	0.10	0.07	0.08	0.05	0.29
qxh_Latn	0.07	0.11	0.04	0.05	0.15	smt_Latn	0.11	0.15	0.05	0.05	0.21
qxl_Latn	0.07	0.11	0.07	0.05	0.08	sna_Latn	0.07	0.11	0.11	0.08	0.18
qxn_Latn	0.07	0.15	0.07	0.05	0.23	snc_Latn	0.15	0.12	0.05	0.05	0.06
qxo_Latn	0.09	0.11	0.05	0.06	0.23	snd_Arab	0.07	0.19	0.61	0.67	0.61
qxr_Latn	0.07	0.13	0.10	0.05	0.14	snf_Latn	0.14	0.11	0.06	0.05	0.06
rad_Latn	0.09	0.09	0.06	0.05	0.06	snn_Latn	0.14	0.17	0.09	0.05	0.05
rai_Latn	0.16	0.18	0.05	0.07	0.12	snp_Latn	0.12	0.11	0.06	0.05	0.09
rap_Latn	0.13	0.13	0.06	0.05	0.21	snw_Latn	0.09	0.11	0.05	0.05	0.05
rar_Latn	0.10	0.07	0.06	0.05	0.22	sny_Latn	0.07	0.13	0.06	0.05	0.08
rav_Deva	0.07	0.09	0.17	0.05	0.07	som_Latn	0.08	0.09	0.31	0.39	0.43
raw_Latn	0.12	0.14	0.05	0.05	0.06	sop_Latn	0.15	0.14	0.07	0.05	0.20
rej_Latn	0.12	0.25	0.20	0.18	0.31	soq_Latn	0.19	0.17	0.05	0.07	0.08
rel_Latn	0.15	0.12	0.08	0.05	0.06	sot_Latn	0.13	0.10	0.09	0.05	0.18
rgu_Latn	0.07	0.07	0.04	0.04	0.15	soy_Latn	0.16	0.07	0.05	0.05	0.05
ria_Latn	0.08	0.10	0.06	0.05	0.06	spa_Latn	0.11	0.49	0.64	0.69	0.58
rim_Latn	0.13	0.16	0.05	0.06	0.07	spl_Latn	0.07	0.12	0.05	0.05	0.05
rjs_Deva	0.07	0.13	0.26	0.22	0.28	spp_Latn	0.10	0.08	0.06	0.05	0.09
rkb_Latn	0.12	0.07	0.05	0.05	0.08	sps_Latn	0.14	0.17	0.05	0.05	0.05
rnc_Latn	0.12	0.17	0.17	0.09	0.18	spy_Latn	0.07	0.09	0.05	0.05	0.07
rmo_Latn	0.17	0.16	0.08	0.06	0.11	sqi_Latn	0.10	0.33	0.68	0.66	0.65
rmy_Latn	0.12	0.23	0.10	0.06	0.22	sri_Latn	0.07	0.13	0.04	0.05	0.06
rnl_Latn	0.11	0.14	0.05	0.05	0.09	srm_Latn	0.12	0.09	0.06	0.05	0.21
ron_Latn	0.11	0.50	0.62	0.65	0.53	srn_Latn	0.07	0.15	0.07	0.05	0.42
roo_Latn	0.07	0.10	0.05	0.05	0.05	srp_Latn	0.09	0.47	0.59	0.59	0.63
rop_Latn	0.20	0.20	0.06	0.05	0.20	srq_Latn	0.16	0.07	0.11	0.07	0.10
row_Latn	0.07	0.08	0.06	0.05	0.08	ssd_Latn	0.12	0.17	0.05	0.05	0.05
roo_Latn	0.08	0.11	0.07	0.05	0.05	ssg_Latn	0.13	0.06	0.11	0.06	0.06
rub_Latn	0.13	0.13	0.08	0.05	0.08	ssw_Latn	0.07	0.11	0.09	0.12	0.24
ruf_Latn	0.14	0.20	0.10	0.09	0.11	ssx_Latn	0.11	0.13	0.07	0.05	0.06
rug_Latn	0.10	0.13	0.06	0.05	0.06	stn_Latn	0.19	0.16	0.11	0.05	0.15
run_Latn	0.16	0.15	0.09	0.06	0.27	stp_Latn	0.09	0.04	0.05	0.05	0.05
rus_Cyrl	0.07	0.50	0.55	0.67	0.64	sua_Latn	0.18	0.13	0.05	0.05	0.05
rwo_Latn	0.07	0.10	0.07	0.06	0.05	suc_Latn	0.13	0.11	0.06	0.05	0.08
sab_Latn	0.07	0.10	0.08	0.05	0.06	sue_Latn	0.13	0.14	0.08	0.05	0.06
sag_Latn	0.11	0.19	0.10	0.06	0.20	suk_Latn	0.16	0.13	0.07	0.07	0.09
sah_Cyrl	0.07	0.12	0.08	0.05	0.30	sun_Latn	0.09	0.33	0.45	0.50	0.45
saj_Latn	0.05	0.10	0.05	0.05	0.08	sur_Latn	0.15	0.11	0.06	0.05	0.10
san_Taml	0.07	0.05	0.07	0.05	0.05	sus_Latn	0.12	0.15	0.04	0.05	0.05
sas_Latn	0.11	0.22	0.28	0.24	0.30	suz_Deva	0.07	0.10	0.11	0.06	0.27
sat_Latn	0.12	0.08	0.06	0.05	0.06	swe_Latn	0.13	0.48	0.73	0.60	0.59
sba_Latn	0.12	0.11	0.06	0.05	0.11	swg_Latn	0.21	0.27	0.25	0.34	0.35
sbd_Latn	0.12	0.09	0.06	0.06	0.05	swh_Latn	0.12	0.31	0.50	0.57	0.54
sbl_Latn	0.12	0.08	0.18	0.12	0.21	swk_Latn	0.11	0.13	0.04	0.06	0.19
sck_Deva	0.07	0.17	0.28	0.44	0.47	swp_Latn	0.08	0.10	0.08	0.06	0.06
sda_Latn	0.11	0.16	0.09	0.05	0.13	sxb_Latn	0.10	0.13	0.08	0.05	0.14
sdq_Latn	0.06	0.15	0.12	0.10	0.16	sxn_Latn	0.07	0.09	0.05	0.05	0.18
seh_Latn	0.13	0.11	0.07	0.06	0.23	syb_Latn	0.13	0.09	0.10	0.05	0.11
ses_Latn	0.14	0.09	0.07	0.05	0.07	sys_Syrc	0.07	0.05	0.05	0.08	0.10
sey_Latn	0.06	0.10	0.05	0.05	0.05	syl_Latn	0.07	0.06	0.05	0.05	0.05
sgb_Latn	0.14	0.22	0.17	0.10	0.31	szb_Latn	0.07	0.21	0.04	0.05	0.06
sgw_Ethi	0.07	0.09	0.10	0.13	0.24	tab_Cyrl	0.07	0.11	0.12	0.05	0.10
sgz_Latn	0.07	0.13	0.06	0.05	0.07	tac_Latn	0.12	0.20	0.05	0.05	0.07
shi_Latn	0.13	0.07	0.05	0.05	0.07	taj_Deva	0.07	0.13	0.14	0.09	0.20
shk_Latn	0.11	0.07	0.06	0.05	0.07	tam_Taml	0.07	0.35	0.53	0.56	0.60
shn_Mymr	0.07	0.05	0.06	0.05	0.05	tap_Latn	0.14	0.18	0.10	0.08	0.20

Table C9 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glot500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
taq_Latn	0.10	0.11	0.07	0.05	0.06	tro_Latn	0.15	0.12	0.07	0.05	0.07
tar_Latn	0.10	0.10	0.05	0.05	0.05	trp_Latn	0.10	0.08	0.06	0.05	0.05
tat_Cyrl	0.07	0.31	0.12	0.15	0.45	trq_Latn	0.05	0.12	0.05	0.05	0.07
tav_Latn	0.13	0.11	0.05	0.05	0.09	trs_Latn	0.06	0.10	0.07	0.05	0.10
taw_Latn	0.14	0.09	0.07	0.05	0.07	tsg_Latn	0.11	0.17	0.15	0.11	0.27
tbc_Latn	0.09	0.12	0.05	0.05	0.06	tsn_Latn	0.12	0.12	0.09	0.05	0.23
tbg_Latn	0.07	0.14	0.08	0.05	0.06	tsw_Latn	0.07	0.12	0.07	0.05	0.08
tbk_Latn	0.07	0.17	0.11	0.11	0.27	tsz_Latn	0.08	0.10	0.08	0.05	0.14
tbl_Latn	0.12	0.12	0.12	0.05	0.06	ttc_Latn	0.14	0.20	0.10	0.05	0.09
tbo_Latn	0.12	0.13	0.10	0.05	0.05	tte_Latn	0.07	0.07	0.08	0.05	0.05
tbw_Latn	0.11	0.15	0.08	0.06	0.25	ttq_Latn	0.09	0.09	0.07	0.06	0.10
tby_Latn	0.14	0.12	0.06	0.05	0.12	ttr_Cyrl	0.07	0.31	0.18	0.13	0.42
tbz_Latn	0.07	0.09	0.05	0.05	0.05	tuc_Latn	0.18	0.10	0.05	0.05	0.05
tca_Latn	0.07	0.07	0.05	0.05	0.07	tue_Latn	0.07	0.10	0.04	0.05	0.05
tcc_Latn	0.09	0.10	0.05	0.05	0.05	tuf_Latn	0.11	0.13	0.10	0.05	0.06
tcs_Latn	0.21	0.19	0.11	0.06	0.21	tui_Latn	0.17	0.14	0.08	0.05	0.07
tcz_Latn	0.12	0.11	0.09	0.05	0.05	tuk_Latn	0.11	0.11	0.22	0.22	0.44
tdt_Latn	0.15	0.15	0.09	0.05	0.36	tul_Latn	0.12	0.18	0.05	0.05	0.05
ted_Latn	0.10	0.09	0.05	0.05	0.05	tum_Latn	0.13	0.22	0.10	0.07	0.21
tee_Latn	0.06	0.07	0.06	0.05	0.14	tuo_Latn	0.12	0.09	0.04	0.05	0.08
tel_Telu	0.07	0.30	0.60	0.67	0.67	tur_Latn	0.11	0.29	0.68	0.68	0.63
tem_Latn	0.12	0.05	0.06	0.05	0.05	tvk_Latn	0.11	0.19	0.08	0.05	0.10
teo_Latn	0.09	0.12	0.05	0.07	0.08	twb_Latn	0.10	0.12	0.05	0.05	0.06
ter_Latn	0.12	0.13	0.06	0.05	0.06	twi_Latn	0.10	0.15	0.05	0.05	0.13
tet_Latn	0.07	0.11	0.05	0.05	0.13	twu_Latn	0.12	0.15	0.16	0.05	0.07
tfr_Latn	0.12	0.14	0.08	0.05	0.05	txq_Latn	0.07	0.15	0.09	0.05	0.06
tgk_Cyrl	0.07	0.19	0.05	0.04	0.31	txu_Latn	0.13	0.17	0.07	0.05	0.05
tgl_Latn	0.13	0.29	0.47	0.55	0.55	tyv_Cyrl	0.07	0.12	0.19	0.18	0.44
tgo_Latn	0.09	0.14	0.05	0.05	0.05	tzh_Latn	0.08	0.10	0.09	0.05	0.22
tgp_Latn	0.15	0.21	0.08	0.09	0.09	tzj_Latn	0.13	0.15	0.09	0.06	0.21
tha_Thai	0.07	0.08	0.56	0.60	0.56	tzo_Latn	0.08	0.11	0.07	0.05	0.30
thk_Latn	0.16	0.10	0.04	0.05	0.05	ubr_Latn	0.15	0.13	0.06	0.05	0.10
thl_Deva	0.07	0.24	0.34	0.44	0.45	ubu_Latn	0.13	0.07	0.07	0.05	0.06
tif_Latn	0.07	0.10	0.05	0.05	0.08	udm_Cyrl	0.07	0.10	0.07	0.05	0.20
tih_Latn	0.09	0.11	0.09	0.05	0.26	udu_Latn	0.19	0.11	0.05	0.05	0.08
tik_Latn	0.09	0.07	0.05	0.05	0.05	uig_Cyrl	0.07	0.20	0.13	0.14	0.44
tim_Latn	0.07	0.11	0.06	0.05	0.06	ukr_Cyrl	0.07	0.40	0.64	0.67	0.57
tir_Ethi	0.07	0.06	0.27	0.22	0.38	upv_Latn	0.10	0.12	0.06	0.05	0.05
tiy_Latn	0.15	0.17	0.08	0.06	0.08	ura_Latn	0.07	0.08	0.05	0.05	0.05
tke_Latn	0.13	0.14	0.06	0.05	0.09	urb_Latn	0.14	0.11	0.12	0.05	0.05
tku_Latn	0.10	0.09	0.06	0.05	0.15	urd_Arab	0.07	0.37	0.49	0.67	0.56
tlb_Latn	0.09	0.13	0.07	0.05	0.09	urk_Thai	0.07	0.09	0.07	0.05	0.05
tlf_Latn	0.07	0.07	0.09	0.05	0.08	urt_Latn	0.06	0.13	0.08	0.05	0.06
tlh_Latn	0.22	0.29	0.24	0.13	0.29	ury_Latn	0.14	0.10	0.05	0.05	0.06
tlj_Latn	0.19	0.14	0.11	0.05	0.12	usa_Latn	0.07	0.10	0.06	0.05	0.05
tmc_Latn	0.10	0.12	0.05	0.05	0.08	usp_Latn	0.18	0.11	0.07	0.05	0.24
tmd_Latn	0.07	0.08	0.05	0.05	0.05	uth_Latn	0.07	0.10	0.09	0.05	0.07
tma_Latn	0.11	0.12	0.13	0.05	0.07	uvh_Latn	0.07	0.09	0.07	0.05	0.05
tnk_Latn	0.11	0.11	0.05	0.05	0.04	uvl_Latn	0.09	0.16	0.06	0.05	0.09
tnn_Latn	0.13	0.10	0.07	0.05	0.07	uzb_Latn	0.09	0.14	0.54	0.59	0.58
tnp_Latn	0.12	0.07	0.05	0.07	0.06	uzn_Cyrl	0.07	0.14	0.07	0.10	0.47
tnr_Latn	0.13	0.07	0.05	0.05	0.06	vag_Latn	0.10	0.11	0.05	0.05	0.06
tob_Latn	0.07	0.12	0.04	0.05	0.09	vap_Latn	0.19	0.12	0.06	0.05	0.17
toc_Latn	0.06	0.09	0.05	0.05	0.05	var_Latn	0.10	0.13	0.07	0.05	0.06
toh_Latn	0.11	0.12	0.06	0.06	0.22	ven_Latn	0.11	0.12	0.06	0.05	0.11
toi_Latn	0.07	0.13	0.08	0.06	0.24	vid_Latn	0.11	0.14	0.11	0.09	0.09
toj_Latn	0.12	0.06	0.07	0.05	0.29	vie_Latn	0.09	0.38	0.54	0.63	0.53
ton_Latn	0.09	0.08	0.05	0.05	0.26	viv_Latn	0.07	0.11	0.06	0.05	0.05
too_Latn	0.10	0.11	0.06	0.05	0.11	vmy_Latn	0.13	0.10	0.05	0.05	0.10
top_Latn	0.08	0.13	0.05	0.05	0.17	vun_Latn	0.13	0.10	0.06	0.05	0.05
tos_Latn	0.06	0.07	0.05	0.05	0.07	vut_Latn	0.08	0.05	0.05	0.05	0.05
tpi_Latn	0.17	0.17	0.09	0.06	0.31	waj_Latn	0.10	0.08	0.06	0.05	0.06
tpm_Latn	0.14	0.12	0.06	0.05	0.06	wal_Latn	0.15	0.10	0.06	0.06	0.13
tpp_Latn	0.13	0.15	0.06	0.05	0.10	wap_Latn	0.11	0.11	0.06	0.05	0.06
tpt_Latn	0.14	0.07	0.09	0.05	0.15	war_Latn	0.11	0.16	0.15	0.14	0.37
tpz_Latn	0.12	0.11	0.06	0.05	0.06	way_Latn	0.10	0.12	0.07	0.05	0.05
tqb_Latn	0.07	0.11	0.08	0.05	0.05	wba_Latn	0.09	0.10	0.08	0.06	0.11
tqo_Latn	0.12	0.08	0.06	0.05	0.05	wbm_Latn	0.09	0.13	0.06	0.05	0.09
trc_Latn	0.05	0.14	0.05	0.05	0.07	wbp_Latn	0.07	0.07	0.06	0.05	0.05
trn_Latn	0.12	0.15	0.06	0.06	0.05	wca_Latn	0.07	0.14	0.05	0.05	0.08

Table C10 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glot500-m.

lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m	lan_script	BOW	mBert	XLM-R-B	XLM-R-L	Glott500-m
wer_Latn	0.09	0.15	0.05	0.05	0.05	zac_Latn	0.12	0.20	0.09	0.09	0.18
whk_Latn	0.11	0.17	0.07	0.05	0.11	zad_Latn	0.15	0.10	0.04	0.05	0.05
wim_Latn	0.07	0.08	0.06	0.05	0.08	zae_Latn	0.14	0.13	0.10	0.05	0.06
wiu_Latn	0.12	0.13	0.05	0.06	0.05	zai_Latn	0.08	0.21	0.13	0.09	0.25
wmw_Latn	0.14	0.16	0.23	0.31	0.41	zam_Latn	0.09	0.16	0.07	0.05	0.13
wnc_Latn	0.07	0.12	0.07	0.06	0.05	zao_Latn	0.14	0.09	0.06	0.05	0.06
wmu_Latn	0.11	0.13	0.05	0.05	0.05	zar_Latn	0.11	0.17	0.06	0.05	0.08
wob_Latn	0.11	0.06	0.05	0.05	0.05	zas_Latn	0.07	0.16	0.07	0.06	0.13
wol_Latn	0.16	0.12	0.07	0.05	0.07	zat_Latn	0.13	0.11	0.11	0.06	0.13
wos_Latn	0.16	0.10	0.08	0.05	0.06	zav_Latn	0.07	0.06	0.05	0.05	0.06
wrs_Latn	0.15	0.10	0.06	0.05	0.05	zaw_Latn	0.07	0.06	0.06	0.05	0.07
wsg_Telu	0.07	0.09	0.13	0.08	0.07	zca_Latn	0.21	0.14	0.18	0.06	0.21
wsk_Latn	0.12	0.15	0.08	0.05	0.10	zho_Hani	0.07	0.39	0.63	0.63	0.59
wuv_Latn	0.18	0.09	0.09	0.05	0.06	zia_Latn	0.14	0.11	0.06	0.05	0.06
wwa_Latn	0.16	0.08	0.05	0.06	0.05	ziw_Latn	0.13	0.17	0.14	0.11	0.23
xal_Cyrl	0.07	0.12	0.08	0.05	0.14	zlm_Latn	0.07	0.47	0.68	0.71	0.62
xav_Latn	0.11	0.13	0.08	0.05	0.10	zoc_Latn	0.11	0.08	0.06	0.05	0.11
xbr_Latn	0.09	0.09	0.08	0.05	0.07	zom_Latn	0.10	0.16	0.13	0.05	0.27
xed_Latn	0.11	0.10	0.06	0.05	0.07	zos_Latn	0.15	0.16	0.05	0.06	0.14
xho_Latn	0.09	0.14	0.21	0.30	0.34	zpc_Latn	0.13	0.12	0.11	0.05	0.12
xla_Latn	0.13	0.08	0.08	0.05	0.05	zpi_Latn	0.13	0.16	0.09	0.05	0.08
xmm_Latn	0.14	0.30	0.42	0.40	0.40	zpl_Latn	0.07	0.13	0.13	0.06	0.17
xnn_Latn	0.07	0.11	0.10	0.08	0.19	zpm_Latn	0.17	0.14	0.05	0.06	0.08
xog_Latn	0.07	0.16	0.06	0.06	0.22	zpo_Latn	0.10	0.15	0.13	0.06	0.10
xon_Latn	0.06	0.17	0.05	0.05	0.05	zpq_Latn	0.07	0.10	0.06	0.05	0.09
xpe_Latn	0.08	0.11	0.05	0.05	0.06	zpt_Latn	0.11	0.11	0.10	0.05	0.16
xrb_Latn	0.11	0.11	0.05	0.05	0.05	zpu_Latn	0.14	0.08	0.05	0.05	0.06
xsb_Latn	0.11	0.14	0.11	0.08	0.23	zpv_Latn	0.10	0.08	0.05	0.05	0.05
xsi_Latn	0.09	0.13	0.05	0.05	0.05	zpz_Latn	0.05	0.07	0.08	0.05	0.05
xsm_Latn	0.19	0.08	0.05	0.05	0.05	zsm_Latn	0.07	0.53	0.71	0.63	0.58
xsr_Deva	0.07	0.09	0.05	0.05	0.06	zsr_Latn	0.09	0.12	0.07	0.05	0.09
xsu_Latn	0.13	0.15	0.05	0.05	0.08	ztq_Latn	0.10	0.13	0.10	0.08	0.19
xtd_Latn	0.14	0.16	0.05	0.05	0.07	zty_Latn	0.11	0.06	0.09	0.05	0.12
xtm_Latn	0.07	0.15	0.06	0.06	0.08	zul_Latn	0.07	0.11	0.23	0.33	0.37
xuo_Latn	0.10	0.08	0.05	0.05	0.05	zyb_Latn	0.15	0.10	0.06	0.05	0.05
yaa_Latn	0.07	0.11	0.06	0.05	0.06	zyp_Latn	0.10	0.15	0.05	0.05	0.06
yad_Latn	0.11	0.09	0.05	0.05	0.05						
yal_Latn	0.15	0.13	0.06	0.05	0.07						
yam_Latn	0.13	0.05	0.05	0.05	0.05						
yan_Latn	0.10	0.13	0.05	0.05	0.05						
yao_Latn	0.13	0.13	0.06	0.05	0.15						
yap_Latn	0.13	0.14	0.07	0.05	0.22						
yaq_Latn	0.16	0.16	0.07	0.05	0.06						
yas_Latn	0.13	0.10	0.05	0.05	0.05						
yat_Latn	0.11	0.05	0.05	0.05	0.06						
yaz_Latn	0.07	0.12	0.08	0.05	0.05						
ybb_Latn	0.07	0.09	0.05	0.05	0.05						
yby_Latn	0.07	0.08	0.07	0.07	0.05						
ycn_Latn	0.10	0.09	0.05	0.05	0.05						
yim_Latn	0.13	0.12	0.09	0.05	0.06						
yka_Latn	0.09	0.14	0.10	0.07	0.26						
yle_Latn	0.07	0.13	0.05	0.05	0.05						
yli_Latn	0.11	0.17	0.09	0.05	0.10						
yml_Latn	0.08	0.08	0.05	0.05	0.06						
yom_Latn	0.09	0.16	0.06	0.05	0.21						
yon_Latn	0.12	0.11	0.11	0.05	0.09						
yor_Latn	0.11	0.14	0.10	0.05	0.10						
yrb_Latn	0.19	0.10	0.11	0.05	0.06						
yre_Latn	0.08	0.11	0.05	0.05	0.05						
yss_Latn	0.10	0.12	0.08	0.05	0.08						
yua_Latn	0.16	0.16	0.11	0.05	0.13						
yue_Hani	0.07	0.40	0.60	0.60	0.56						
yuj_Latn	0.14	0.08	0.09	0.06	0.07						
yut_Latn	0.11	0.14	0.05	0.05	0.05						
yuw_Latn	0.10	0.12	0.09	0.05	0.05						
yuz_Latn	0.07	0.12	0.10	0.05	0.10						
yva_Latn	0.13	0.15	0.06	0.05	0.06						
zaa_Latn	0.10	0.20	0.20	0.07	0.29						
zab_Latn	0.07	0.08	0.13	0.07	0.16						

Table C11 zero-shot score of BOW, mBERT, XLM-R-B, XLM-R-L, and Glott500-m.