TRANSMI: A Framework to Create Strong Baselines from Multilingual Pretrained Language Models for Transliterated Data

Anonymous ACL submission

Abstract

Transliterating related languages that use different scripts into a common script is effective for improving crosslingual transfer in downstream tasks. However, this methodology often makes pretraining a model from scratch unavoidable, as transliteration brings about new subwords not covered in existing multilingual pretrained language models (mPLMs). This is undesirable because it requires a large computation budget. A more promising way is to make full use of available mPLMs. To this end, this paper proposes a simple but effective framework: **Trans**literate-**M**erge-**I**nitialize (**TRANSMI**). TRANSMI is a strong baseline well-suited for data that is transliterated into a common script by exploiting an mPLM and its tokenizer.[1] TRANSMI has three stages: (**a**) transliterate the vocabulary of an mPLM into a common script; (**b**) merge the new vocabulary with the original vocabulary; and (**c**) initialize the embeddings of the new subwords. We apply TRANSMI to three strong recent mPLMs. Our experiments demonstrate that TRANSMI not only preserves the mPLM's ability to handle non-transliterated data, but also enables it to effectively process transliterated data, thereby facilitating crosslingual transfer. The results show consistent improvements of 3% to 34% for different mPLMs and tasks. We will make our code and models publicly available.[2]

original sentence: **transliteration**:	今天是个好天气 jintianshigehaotianqi
original tokenizer	‘_今天’, ‘是个’, ‘好’, ‘天气’ ‘_jint’, ‘ian’, ‘shig’, ‘ehao’, ‘tian’, ‘qi’
modified tokenizer	‘_今天’, ‘是个’, ‘好’, ‘天气’ ‘_jintian’, ‘shige’, ‘hao’, ‘tianqi’

Table 1: Tokenization results of a sentence written in its original script (Hani) and its Latin transliteration. The correct word correspondences are: (今天 – jintian – <today>), (是个 – shige – <is>), (好 – hao – <good>), (天气 – tianqi – <weather>). The original tokenizer produces nonsensical strings that do not correspond to the meaning-bearing units. The modified tokenizer correctly tokenizes the transliterated text while also preserving the ability to handle the sentence in its original Hani script.

1 Introduction

Crosslingual transfer refers to applying knowledge gained from one language to the learning or processing of another language (Zoph et al., 2016; Wu and Dredze, 2019; Artetxe et al., 2020). This transfer is attractive as we often do not have enough training data for low-resource languages while training data for high-resource languages is generally abundant (Magueresse et al., 2020; Hedderich et al., 2021; Liu, 2022). Although recent mPLMs have made remarkable progress in improving crosslingual transfer, they often cannot achieve strong performance when transferring to a wide spectrum of low-resource languages. Lexical overlap, i.e., the phenomenon where vocabularies are shared between languages, is a key factor influencing the quality of crosslingual transfer (Pires et al., 2019; Lin et al., 2019). However, because languages are written in different writing systems, or *scripts*, lexical overlap cannot be fully exploited.

To tackle this problem, a few recent works attempt to apply rule-based transliteration tools and convert all data to a common script (Dhamecha et al., 2021; Muller et al., 2021; Moosa et al., 2023). By doing this, the script diversity no longer poses difficulty in improving lexical overlap, therefore better performance can be obtained when the original scripts are different for the transfer source and target languages. However, this approach either requires training a brand-new mPLM from scratch (Dhamecha et al., 2021; Moosa et al., 2023) or involves (parameter-efficient) parameter updates to adapt the transliterated data (Muller et al., 2021;

[1] Throughout this paper we will simply use mPLM to refer to both the model and its tokenizer for convenience.

[2] URL hidden for anonymity.

Purkayastha et al., 2023), as the embeddings of the new subwords generated from transliteration need to be properly trained before a model can be applied to any downstream tasks. This inevitably demands a high computing budget. Additionally, such dedicated models specific to transliterated data can only deal with one script. Therefore, a natural research question is: *Can one make full use of an mPLM and a transliteration tool to build a strong baseline that is well-suited for transliterated data, **without any training**?*

To this end, this work presents a simple yet effective framework: **Trans**literate-**M**erge-**I**nitialize (**TRANSMI**). TRANSMI has three stages. In the first stage, a transliteration tool is used to transliterate all the subwords in the vocabulary of an mPLM into a common script (Latin). Next, we merge the new subwords obtained in the previous step into the original tokenizer, where we propose three different modes to account for the problem of *transliteration ambiguity* (different subwords in the vocabulary have the same transliteration). Lastly, we initialize the embeddings for the newly added subwords. In this way, we modify the original mPLM so that it can deal with transliterated data while not losing the ability to process non-transliterated data. In contrast to the original mPLM tokenizer that generates non-meaningful tokenization for transliteration, the modified tokenizer generates tokens that correspond well to natural linguistic units, as shown in Table 1.

We validate TRANSMI by applying it to three recent strong mPLMs that show strong crosslingual transfer and evaluating the resulting models on a variety of downstream tasks including sentence retrieval, text classification, and sequence labeling. We evaluate each resulting model on both transliterated and non-transliterated evaluation datasets. We show that the models enhanced by our framework not only achieve very similar performance on non-transliterated data as their original mPLM counterparts but also largely outperform them on transliterated data across all tasks.

The contributions are as follows: (i) We present TRANSMI, a simple yet effective framework that creates strong baselines from mPLMs for transliterated data, without any training. (ii) We show that TRANSMI boosts the performance on transliterated data while not sacrificing performance on non-transliterated data. (iii) We investigate in-depth how different modes in the Merge (Initialize) step in TRANSMI influence the performance. (iv) Through fine-grained analysis, we show that TRANSMI benefits languages from all script groups for transliterated data.

2 Related Work

2.1 Transliteration for Multilingual NLP

Transliteration is the process of converting text from one script into another (Wellisch et al., 1978). This process does not involve translating meanings but rather represents the source script symbols as faithfully as possible in the target script. Transliteration has been proven to be an effective method for improving neural machine translation between languages that are written in different scripts (Gheini and May, 2019; Goyal et al., 2020; Amrhein and Sennrich, 2020). Transliteration can also boost crosslingual transfer on a large scale when languages are transliterated into a common script, especially for languages that are mutually influenced but written in different scripts (Dhamecha et al., 2021; Muller et al., 2021; Chau and Smith, 2021; Purkayastha et al., 2023; Moosa et al., 2023). More recently, Liu et al. (2024) propose a framework where transliterations are used as an auxiliary input along with the original-script text to improve the cross-script alignment. This line of work involves training for adaptation to transliterated data. In contrast, we propose a simple framework to construct strong baselines directly from existing mPLMs for transliterated data, without any training.

2.2 Vocabulary and Tokenizer Manipulation

Training a new tokenizer on data of unseen languages and optionally merging it with the original tokenizer is a common way for efficient language adaptation (Pfeiffer et al., 2021; Alabi et al., 2022; ImaniGooghari et al., 2023; Liu et al., 2023b). Similarly, adaptively manipulating the tokenizer and vocabulary also shows strong performance improvement for domain-specific data within the same language (Sachidananda et al., 2021; Lamproudis et al., 2022; Liu et al., 2023a). Kajiura et al. (2023) propose to replace certain subwords in a tokenizer with new subwords learned from the domain-specific corpus for domain adaptation, thus not changing the vocabulary size. Nevertheless, this line of work requires training to learn good representations for the new subwords. Another related work (Hofmann et al., 2022), instead of modifying the vocabulary, directly changes the behavior of the tokenizer to preserve the morphological structure, enhancing robustness and performance. Our work also modifies the vocabulary and tokenizer by including new subwords. In contrast to previous work, we initialize the new subword embeddings by actively exploiting the original mPLM embeddings. Thus the resulting model can be directly adapted to the transliterated data without any training.

3 Preliminary: SentencePiece Unigram

Unigram (Kudo, 2018) is a tokenization algorithm for obtaining subword vocabulary, which is usually used in conjunction with SentencePiece (Kudo and Richardson, 2018). In contrast to Byte-Pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2016) or WordPiece (Schuster and Nakajima, 2012; Wu et al., 2016), Unigram is based on a language model that outputs multiple subword segmentations with probabilities. In addition, Unigram does not learn subwords through merging frequent character combinations gradually as done by BPE. Instead, it initializes a large number of units as its vocabulary and progressively removes units that have low contributions to the likelihood of the training corpus, until a pre-defined vocabulary size is obtained. The optimization is done by expectation-maximization (EM) algorithm (Dempster et al., 1977) and the overall training objective is to maximize the marginal likelihood $\mathcal{L}$:

$$\mathcal{L} = \sum_{i=1}^{|\mathcal{D}|} \log\, P(X_i) = \sum_{i=1}^{|\mathcal{D}|} \log\,(\sum_{\boldsymbol{x} \in S(X_i)} P(\boldsymbol{x}))$$

where $\mathcal{D}$ is the training corpus, X_i is the ith sentence in $\mathcal{D}$, and $S(X_i)$ is the set of all possible segmentation candidates for the input sentence X_i.

Once the Unigram tokenizer is trained, in addition to its vocabulary V, the model will also save a score, i.e., the log probability, learned from the training corpus, for each subword w in V, as shown in Figure 1. This makes it possible for the model to provide the probability of each possible tokenization for a given sentence after training. In practice, the tokenizer is usually set to generate the most probable segmentation, i.e., the sequence of subwords that maximize the log probability:

$$\boldsymbol{x}^* = \text{argmax}_{\boldsymbol{x} \in S(X)} \sum_{w \in \boldsymbol{x}} \log P(w)$$

where $\boldsymbol{x}^*$ is the optimal tokenization given the sentence X and $\log P(w)$ is the log probability of subword w.

4 Methodology

We present TRANSMI, a simple yet effective framework to create a strong baseline from a given mPLM for transliterated data[3] without any training. There are three stages in TRANSMI: (1) transliterate the subwords in the vocabulary; (2) merge the transliterated subwords into the tokenizer; and (3) initialize the embeddings for the new subwords. In each stage, the information and knowledge from the mPLM and its tokenizer are carefully exploited. We illustrate the whole pipeline in Figure 1 and introduce each stage in detail in the following.

4.1 Tokenizer Vocabulary Transliteration

The vocabulary of a multilingual tokenizer contains subwords that are learned using tokenization algorithms such as SentencePiece[4] (Kudo and Richardson, 2018) on a concatenation of data from different languages (Conneau et al., 2020; ImaniGooghari et al., 2023). As a result, many subwords are in non-Latin script. An intuitive way of adapting the vocabulary to Latin-script transliterated data is to add the transliterations of these non-Latin subwords into the vocabulary. Let V^{orig} be the vocabulary of the mPLM, and `Transli` a deterministic transliterator. We create a set of *transliteration triplets* by applying `Transli` to every subword w and associating it with a score s, i.e., w's log probability:

$$T = \{(v, w, s) | v = \texttt{Transli}(w) \wedge w \in V^{\text{orig}}\}$$

It is important to note that $|T| = |V^{\text{orig}}|$ and there will be no duplicates in T. That is, given two elements: (v_i, w_i, s_i) and (v_j, w_j, s_j) where $v_i = v_j$, we always have $w_i \neq w_j$ (usually also $s_i \neq s_j$). For example, "taiyang" is the Latin transliteration for both "太阳" ("sun" in simplified Chinese) and "太陽" ("sun" in traditional Chinese). The score of "太陽" is higher than "太阳" since "太陽" is more frequent and therefore has higher log probability.

4.2 Merge New Vocabulary

Once we obtain T, we can modify the original vocabulary V^{orig} by adding the subword transliterations v_i. However, we need to consider the possible introduced transliteration ambiguity while

[3] In this work, we consider one special type of transliteration that involves converting non-Latin scripts into the Latin script. This is also referred to as romanization.

[4] The tokenizers used in this study are all SentencePiece Unigram models where each subword is associated with a log probability. We refer to this log probability as the score.

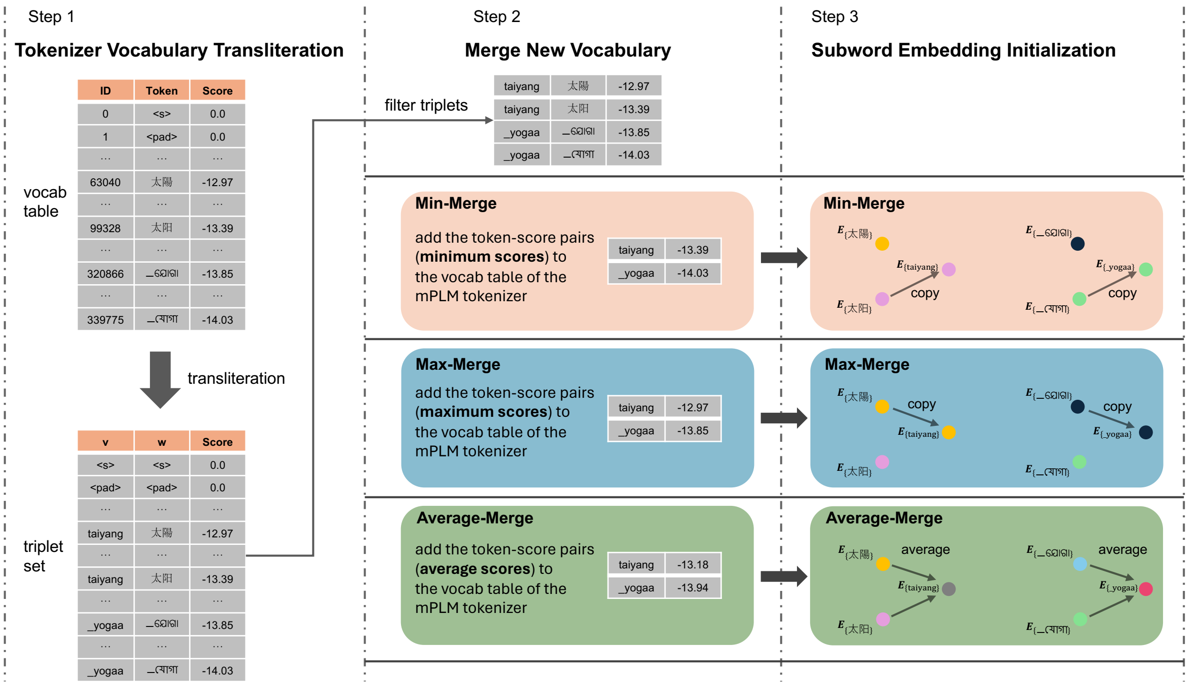


Figure 1: Overview of **TransMI**. We transliterate all the subwords from the vocabulary of the source mPLM tokenizer into Latin script in **Step 1**. We then merge the filtered triplets (ambiguous transliterations) into the tokenizer vocabulary table using one of the three proposed modes in **Step 2**. Note that we perform direct merge operations for the rest of the triplets that are not ambiguous (not shown in the figure). Lastly in **Step 3**, we initialize the embeddings for the newly added subwords according to the merge mode used in the previous step.

merging. For subword transliterations that already exist in V^{orig}, nothing needs to be done – this applies to all subwords in Latin script without any diacritics. For each subword transliteration v' that has a one-to-one relation to a w', i.e., $\forall(v, w, s) \in T : v = v' \Rightarrow w = w'$, we simply add it with its associated score to the vocabulary table.

For each of the remaining subwords v', we first define $U(v') := \{(v, w, s) \in T | v = v'\}$. Then, to address the transliteration ambiguity, we propose three different modes to merge them into the vocabulary.

Min-Merge Mode In this mode, we select the triplet whose score s' is lowest for the transliteration v':

$$s'_{\text{min}}(v') = \min_{(v,w,s) \in U(v')} s$$

This mode will be favorable to preserve the less frequent subwords. Adding the subword transliteration v' and its associated score $s'_{\text{min}}(v')$ to the vocabulary table is likely to alter the tokenizer's behavior negatively. As a consequence, we expect this mode will perform worst among all modes.

Max-Merge Mode In contrast to Min-Merge Mode, this mode selects the triplet $(v', w', s') \in T$ whose score is highest:

$$s'_{\text{max}}(v') = \max_{(v,w,s) \in U(v')} s$$

For high-frequency subwords, this mode replicates the previous tokenization behavior for the original script: after romanization, we are likely to obtain the same tokenization. Therefore, we expect this mode will achieve the best performance.

Average-Merge Mode This mode averages the scores of all triplets containing v'.

$$s'_{\text{avg}}(v') = \frac{\sum_{(v,w,s) \in U(v')} s}{|U|}$$

We then add subword v' with score $s'_{\text{avg}}(v')$ to the vocabulary. Average-Merge Mode brings amount a change in behavior of the tokenizer that is intermediate between Min-Merge Mode and Max-Merge Mode.[5]

[5] As is common in vocabulary extension and tokenizer merging (ImaniGooghari et al., 2023; Lin et al., 2024), we do not renormalize the modified scores (to ensure the new unigram distribution is a proper probability distribution) because the tokenization behavior is only determined by the order of scores.

4.3 Subword Embedding Initialization

The last stage deals with embedding initialization for the newly introduced subwords. We aim to make full use of the knowledge encoded in the original embedding matrix $\boldsymbol{E}^{\text{orig}}$ to avoid any sort of training. To achieve this, we create an additional embedding matrix $\boldsymbol{E}^{\text{add}}$ for the new subwords and initialize the embedding for each subword based on the correspondence we obtain in the previous vocabulary merge stage. Specifically, we directly copy the original embedding for those new subwords that have a one-to-one transliteration relation. For the rest of the subwords, we initialize their embeddings according to which mode is used in the last stage; this makes the resulting embeddings consistent with the tokenizer behavior.

Min-Merge Mode In Min-Merge Mode, we selected the triple $(v', {s'}_{\text{min}}(v'), w')$. Correspondingly, we initialize the embedding of v' as w': $\boldsymbol{E}^{\text{add}}_{v'} = \boldsymbol{E}^{\text{orig}}_{w'}$.

Min-Merge Mode In Max-Merge Mode, we selected the triple $(v', {s'}_{\text{max}}(v'), w')$. Correspondingly, we initialize the embedding of v' as w': $\boldsymbol{E}^{\text{add}}_{v'} = \boldsymbol{E}^{\text{orig}}_{w'}$.

Average-Merge Mode In Average-Merge Mode, we averaged the scores of all w' that are mapped to v', i.e., we averaged the scores in $U(v')$. Correspondingly, we initialize the embedding of v' as the average of the w':

$$\boldsymbol{E}^{\text{add}}_{v'} = \frac{\sum_{(v,w,s)\in U(v')} \boldsymbol{E}^{\text{orig}}_{w}}{|U(v')|}$$

.

By choosing any mode, each embedding in $\boldsymbol{E}^{\text{add}}$ is carefully initialized, and in the same representation space as $\boldsymbol{E}^{\text{orig}}$. To construct the final embeddings, we simply concatenate $\boldsymbol{E}^{\text{orig}}$ and $\boldsymbol{E}^{\text{add}}$ and ensure the tokenizer subword indices and their indices in the embeddings are consistent.

5 Experiments

5.1 Setups

We apply the proposed framework TRANSMI to three strong mPLMs: XLM-R, Glot500, and FURINA. XLM-R (Conneau et al., 2020) is pretrained on 100 languages using masked language modeling (MLM) (Devlin et al., 2019). Glot500 (ImaniGooghari et al., 2023) is a continued pretrained model from XLM-R on Glot500-c dataset that covers more than 500 languages. FURINA (Liu et al., 2024) is a post-aligned version of Glot500, which is fine-tuned using Latin-script transliteration data. We use `Uroman` (Hermjakob et al., 2018) as the rule-based transliteration tool. Note that the tokenizers of Glot500 and FURINA are the same. We show the number of newly added subwords and the transliteration ambiguity in Table 2. We use the base version of each model (their architectures are the same) for a fair comparison. There are 9 resulting models (3 merge modes × 3 model types) in total. When evaluating the models, we use both the non-transliterated evaluation datasets (the original ones) and the transliterated Latin-script evaluation datasets, which are also obtained by using `Uroman`. Following (ImaniGooghari et al., 2023), we refer to language-scripts[6] supported by XLM-R as the **head** languages and the remaining language-scripts (those supported by Glot500) as the **tail** languages.

	1	2	3	>3	Total
XLM-R	97,456	6,866	1,380	1,088	106,790
Glot500	123,001	8,373	1,706	1,274	134,354

Table 2: Transliteration ambiguity of the newly added subwords (transliterations). For example, 97,456 indicates that, out of the total newly added 106,790 subwords for the XLM-R model, 97,456 subwords have a 1-to-1 relationship, i.e., the subword is the Latin transliteration of only 1 subword in the original XLM-R vocabulary. Most of the new subwords are not ambiguous.

5.2 Downstream Tasks

We consider the following three evaluation types. For each type, we consider two evaluation datasets. The evaluation is performed in an English-centric crosslingual zero-shot fashion: fine-tuning on the English train set, selecting the best checkpoint on the English dev set, and then evaluating the best checkpoint on the test sets of all other languages. An exception is Sentence Retrieval in that it does not involve any fine-tuning. For all tasks, only the subset of languages (head and tail languages) supported by Glot500 are considered. Details of the used dataset and hyperparameter settings for fine-tuning are reported in §A.

Sentence Retrieval. We use Bible (SR-B) and Tatoeba (Artetxe and Schwenk, 2019) (SR-T). The similarity is calculated using the mean pooling of

[6] A language-script is a combination of its ISO 639-3 and script codes.

	SR-B			SR-T			Taxi1500			SIB200			NER			POS		
	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all
XLM-R	7.0	28.6	12.5	26.5	35.4	32.9	11.4	34.8	17.4	43.0	52.7	47.4	46.3	46.2	46.3	34.1	59.0	51.3
XLM-R (Max-Merge)	**7.4**	**35.8**	**14.6**	**30.1**	**48.2**	**43.0**	**13.6**	**47.0**	**22.1**	**48.3**	**73.0**	**59.5**	**46.7**	**53.7**	**50.5**	**37.8**	**70.1**	**60.2**
Glot500	33.1	31.6	32.7	44.9	42.3	43.0	41.5	36.4	40.2	59.3	56.2	57.9	54.0	49.0	51.3	48.9	59.8	56.4
Glot500 (Max-Merge)	**34.3**	**38.4**	**35.4**	**49.2**	**55.5**	**53.7**	**45.1**	**48.9**	**46.0**	**66.8**	**74.7**	**70.4**	**57.3**	**57.2**	**57.2**	**52.5**	**68.8**	**63.8**
FURINA	47.9	51.2	48.7	55.4	53.6	54.1	45.0	43.0	44.5	60.7	59.9	60.3	54.4	52.2	53.2	54.3	67.4	63.4
FURINA (Max-Merge)	**49.4**	**54.2**	**50.6**	**56.4**	**59.1**	**58.3**	**49.6**	**53.0**	**50.5**	**66.7**	**74.1**	**70.1**	**57.6**	**58.5**	**58.1**	**55.9**	**71.7**	**66.8**

Table 3: Performance of three model types on **transliterated** evaluation datasets across 5 random seeds. We report the performance as an average over head, tail, and all language-scripts for each model variant. Max-merge models consistently outperform the original model on transliterated evaluation data. **Bold**: best result per model type.

	SR-B			SR-T			Taxi1500			SIB200			NER			POS		
	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all
XLM-R	**7.4**	**54.2**	**19.3**	32.6	**66.2**	**56.6**	**13.5**	**58.7**	**25.0**	**49.5**	**81.1**	**63.9**	**47.6**	**61.0**	**54.9**	**42.7**	76.4	66.0
XLM-R (Max-Merge)	**7.4**	53.3	19.1	**33.1**	65.1	56.0	13.2	58.0	24.6	46.9	80.7	62.2	46.6	60.7	54.3	**42.7**	**76.5**	**66.1**
Glot500	**43.2**	**59.0**	**47.3**	**59.8**	**75.0**	**70.7**	**52.5**	**60.9**	**54.6**	68.5	80.4	73.9	**60.8**	**63.7**	**62.4**	62.0	**76.0**	**71.7**
Glot500 (Max-Merge)	41.7	57.8	45.8	58.3	72.8	68.7	51.5	60.8	53.8	**69.5**	**81.2**	**74.8**	59.6	63.2	61.6	**62.1**	**76.0**	**71.7**
FURINA	**55.3**	**66.2**	**58.1**	**62.1**	**71.5**	**68.8**	55.9	**63.8**	57.9	70.3	82.2	75.7	**60.5**	63.9	**62.4**	**63.3**	75.7	71.9
FURINA (Max-Merge)	**55.3**	65.9	58.0	60.9	70.6	67.9	**56.6**	**63.8**	**58.4**	**71.8**	**82.5**	**76.7**	60.1	**64.2**	**62.4**	62.1	**76.5**	**72.1**

Table 4: Performance of three model types on **non-transliterated** evaluation datasets across 5 random seeds. We report the performance as an average over head, tail, and all language-scripts for each model variant. Max-merge models perform close to the original models. **Bold**: best result per model type.

contextualized word embeddings at the 8th layer.

Text Classification. We use Taxi1500 (Ma et al., 2023) and SIB200 (Adelani et al., 2024).

Sequence Labeling. We use WikiANN for named entity recognition (NER) (Pan et al., 2017) and Universal Dependencies (de Marneffe et al., 2021) for Part-Of-Speech (POS) tagging.

5.3 Results and Discussion

We evaluate the original mPLMs and their corresponding variants modified by TRANSMI on both the transliterated evaluation datasets (all scripts are converted into Latin script) and the non-transliterated (original) datasets. We notice that the different merge modes offer similar performance while the **Max-Merge** mode slightly outperforms the other modes. Therefore we only show the performance of the **Max-Merge** mode in this section. The comparison between different modes is presented and discussed in Analysis (§6.1).

5.3.1 Evaluation on Transliterated Data

We report performance on transliterated data in Table 3. We observe consistent improvement across all tasks. The original mPLMs are not pretrained on transliterated data and thus they perform suboptimally on transliterated evaluation datasets. Modifying these mPLMs with TRANSMI equips these models with the ability to deal with the transliterated texts, as new subwords (transliterations of the subwords covered by the original mPLMs) are included and their embeddings are wisely initialized. Consequently, the resulting models can process the transliterated data and achieve good performance.

It can be observed that the improvement by modifying FURINA is relatively smaller than on other model types. We hypothesize this is because FURINA is already fine-tuned on transliterated data through MLM and transliteration contrastive modeling (Liu et al., 2024). Even though FURINA's vocabulary is the same as Glot500, i.e., not extended and adapted specifically for Latin transliterations, the fine-tuning phase still helps the model gain some knowledge beneficial for processing transliterated data.

5.3.2 Evaluation on Non-transliterated Data

We report performance on non-transliterated data in Table 4. Across all tasks, modified models achieve performance very close to the original mPLMs, with generally negligibly small performance degradation. This is expected as the vocabulary is only augmented with new subwords in Latin script (transliteration) and therefore the tokeniza-

	SR-B			SR-T			Taxi1500			SIB200			NER			POS		
	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all
XLM-R (Min-Merge)	**7.4**	34.7	14.4	28.7	46.4	41.3	**14.4**	45.0	22.1	**49.6**	**73.7**	**60.5**	45.9	52.3	49.3	**38.0**	69.1	59.5
XLM-R (Average-Merge)	**7.4**	34.1	14.2	29.0	45.4	40.7	14.3	46.2	**22.4**	47.1	70.7	57.8	**47.2**	53.5	**50.6**	37.6	69.4	59.6
XLM-R (Max-Merge)	**7.4**	**35.8**	**14.6**	**30.1**	**48.2**	**43.0**	13.6	**47.0**	22.1	48.3	73.0	59.5	46.7	**53.7**	50.5	37.8	**70.1**	**60.2**
Glot500 (Min-Merge)	34.1	36.2	34.6	48.8	53.0	51.8	46.1	48.0	46.6	**67.1**	73.8	70.1	**57.4**	55.8	56.6	**52.5**	67.1	62.6
Glot500 (Average-Merge)	34.2	37.4	35.0	**49.4**	54.8	53.2	**47.2**	**49.8**	**47.9**	66.6	73.9	69.9	56.7	56.1	56.4	52.0	68.3	63.3
Glot500 (Max-Merge)	**34.3**	**38.4**	**35.4**	49.2	**55.5**	**53.7**	45.1	48.9	46.0	66.8	**74.7**	**70.4**	57.3	**57.2**	**57.2**	**52.5**	**68.8**	**63.8**
FURINA (Min-Merge)	49.3	52.3	50.1	56.2	58.0	57.4	48.5	50.2	49.0	66.7	72.8	69.5	**58.2**	**59.0**	**58.6**	55.5	71.0	66.2
FURINA (Average-Merge)	**49.4**	53.5	50.5	56.2	58.5	57.8	48.4	51.4	49.2	**67.6**	**75.1**	**71.0**	57.0	57.7	57.4	**56.1**	71.6	**66.8**
FURINA (Max-Merge)	**49.4**	**54.2**	**50.6**	**56.4**	**59.1**	**58.3**	**49.6**	**53.0**	**50.5**	66.7	74.1	70.1	57.6	58.5	58.1	55.9	**71.7**	**66.8**

Table 5: Performance of three merge modes applied to three mPLMs on **transliterated** evaluation datasets across 5 random seeds. The performance difference among the three modes are small but Min-Merge mode is favorable to tail languages while Max-Merge mode is favorable to head languages. In general, Max-Merge mode achieves the overall best performance. **Bold** (underlined): best (second-best) result for each task in each model type.

tion results for non-Latin texts remain the same (their embeddings are also not changed).

On the other hand, the slight decrease in performance is not surprising. The included new subwords will influence the tokenization results for all languages written in Latin script, including English. As the evaluation is done in an English-centric manner, even if the target language is not written in Latin script, altered English tokenization can still influence the crosslingual transfer performance. When the transfer target language is also written in Latin, the tokenization results and representation are changed both on the source side and target side, the performance therefore varies.

6 Analysis

6.1 Which Merge Mode Wins

To explore how different merge modes (Min-Merge, Average-Merge, and Max-Merge) influence the crosslingual transfer, we report the performance of the three variants of each model type in Table 5. Generally, the performance differences among the three merge modes are very small across model types and tasks. This can be explained by the fact that most of the newly added subwords (transliterations of subwords in the original mPLM vocabulary) are not ambiguous: 91% of subwords in XLM-R and 92% in Glot500 (also FURINA) have 1-to-1 relations, as shown in Table 2. Each merge mode simply does the same copy operation for these unambiguous subwords, and operates differently only on the remaining ambiguous subwords (the transliteration corresponds to more than one subword in the original vocabulary), which is a relatively small portion.

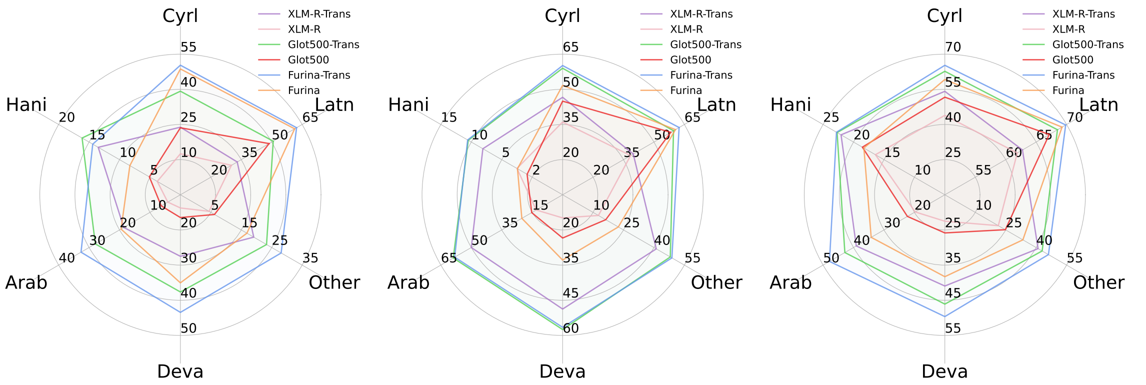
Although the difference is small, we observe that the Min-Merge mode supports tail languages better while the Max-merge mode supports head languages better. We hypothesize that the scores (log probabilities) of some subwords from tail languages (usually low-resource languages) are small, and the Min-Merge mode preserves the small scores and therefore is favorable to tail languages. Similarly, the scores for some subwords from high-resource languages are large, and the Max-Merge mode keeps their priority in tokenization. In general, the Max-Merge mode seems to be the best option, as it has the overall best performance across all tasks in each model type.

6.2 Which Scripts Benefit

We compare the original mPLMs and their modified variants (using the Max-Merge mode) on the three evaluation types (transliterated evaluation) across different scripts in which the transfer target languages are originally written, as shown in Figure 2. Globally, each TRANSMI-modified model consistently achieves better performance than its counterpart across all script groups and all tasks, since the original mPLMs are not specifically trained on

	Latn	Cyrl	Hani	Arab	Deva	Other
Sentence Retrieval	**64.6**	**69.0**	**43.1**	**58.2**	**68.8**	**55.2**
	62.2	50.3	14.4	32.6	43.4	28.1
Text Classification	**65.7**	**73.2**	**34.2**	**73.0**	**76.5**	**71.5**
	62.4	60.2	10.6	58.2	56.6	48.8
Sequence Labeling	**70.8**	**72.0**	**29.2**	**62.8**	**59.6**	**60.0**
	69.8	65.2	22.8	47.9	49.6	46.0

Table 6: Comparison between the performance of FURINA (Max-Merge) on non-transliterated and transliterated evaluation datasets for different script groups. For each evaluation type, the results on non-transliterated (resp. transliterated) data are in the first (resp. second) row. FURINA (Max-Merge) consistently performs better on non-transliterated data than transliterated data.



(a) Sentence Retrieval (b) Text Classification (c) Sequence Labeling

Figure 2: Qualitative comparison between the original mPLMs and TRANSMI (Max-Merge mode) models (denoted with "-Trans") on transliterated evaluation datasets. We compute the average performance for each evaluation type (e.g., Sentence Retrieval is the average of SR-B and SR-T) for languages grouped into the major scripts that the languages are originally written in: **Latn** (Latin), **Cyrl** (Cyrillic), **Hani** (Hani), **Arab** (Arabic), and **Deva** (Devanagari). The remaining languages are grouped into **Other**. TRANSMI-modified models consistently outperform the original mPLMs across all tasks and all script groups.

	Latn	Cyrl	Hani	Arab	Deva	Other
FURINA	45.7	39.4	42.7	42.3	44.1	44.5
	45.5	43.8	34.9	50.8	53.7	57.0
FURINA (Max-Merge)	44.8	39.4	42.7	42.3	44.1	44.5
	43.3	**36.0**	**30.9**	**36.3**	**38.4**	**38.9**

Table 7: Average sequence length of SR-B dataset averaged by script group. The results on non-transliterated (resp. transliterated) data are in the first (resp. second) row. The tokenizer of FURINA (Max-Merge) has consistently shorter sequence lengths on transliterated data compared with the original FURINA tokenizer.

transliterated data (except for FURINA). We observe the smallest improvement for the Hani group and the Latn group. This can be explained by the fact that the former is logograms and transliterated words potentially lose semantic or contextual nuances and are more prone to ambiguity (Liu et al., 2024) while the latter does not change much after `Uroman` transliteration (only diacritics are removed). The rest of the script groups, i.e., Cyrl, Arab, and Deva, all enjoy large improvements, which indicates that TRANSMI is effective in creating strong baselines for transliterated data for languages originally written in phonetic scripts.

6.3 Transliterated vs. Non-transliterated

We also explore how the performance varies before and after the evaluation datasets are transliterated, as shown in Table 6. Not surprisingly, the performance on all script groups drops after transliterating the evaluation datasets. There are mainly two reasons: **(1)** the embeddings of the newly added subwords that have transliteration ambiguity are not suitable for each occurrence **within the same language**, e.g., "shiwu" is the transliteration for both "食物" ("food" in Chinese) and "时务" ("current affairs" in Chinese) and **across different languages**, e.g., "miso" is the transliteration for both μισό ("half" in Greek) and "味噌" ("soybean paste" in Japanese); **(2)** the tokenization result is different before and after transliterating a sentence into Latin script (our method tries to preserve the tokenization, but it is impossible to prevent changes completely), as shown in Table 7. The different tokenizations alter the final representations of sentences, resulting in a drop in performance.

7 Conclusion

This paper presents TRANSMI, a framework to create baseline models adapted for transliterated data from existing mPLMs, without any training. We show TRANSMI-modified models not only preserve the ability to deal with data written in their original script but also demonstrate good capability in processing transliterations of data originally written in non-Latin script. Our experiments indicate the modified models outperform the original mPLMs. In addition, we show that TRANSMI is particularly effective for transliterated data of languages written in phonetic scripts. The modified models therefore serve as strong baselines for transliterated data and also potentially good starting points for continued pretraining or finetuning on domain-specific (transliterated) data.

Limitations

TRANSMI presented in this work tries to create strong baselines from mPLMs for transliterated data without including any training phase. Though the modified models can achieve much better performance than the original mPLMs on transliterated evaluation datasets, there is still a gap between it and the performance on non-transliterated (in the original script) evaluation. We propose several explanations for this phenomenon, such as subword transliteration ambiguity and tokenization differences. These issues should be able to be alleviated by further fine-tuning or continued pretraining on transliterated data. However, this is beyond the scope of the paper, as our motivation is to create strong baselines through a simple and effective framework for modifying existing mPLMs. We would therefore expect much stronger models can be obtained by using our TRANSMI-modified mPLMs as the starting points of further training / fine-tuning, which we would leave for future exploration in the community.

Another possible limitation is that we only consider mPLMs that leverage SentencePiece Unigram tokenizers. This is due to the fact that the most recent strong mPLMs favor such choices. However, it should be very easy to extend TRANSMI to mPLMs that use other types of tokenizers. For example, mBERT (Devlin et al., 2019) uses WordPiece subword tokenizer where the vocabulary is learned through BPE. The vocabulary keeps frequencies of subwords instead of log probabilities. Therefore, we can simply use the frequencies to replace the scores being manipulated in the Merge step of TRANSMI. We would leave this exploration in the community if other tokenizers are used in their studies for instance.

References

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

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

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A

Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Mozhdeh Gheini and Jonathan May. 2019. A universal parent model for low-resource neural machine translation transfer. *arXiv preprint arXiv:1909.06516*.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Teruno Kajiura, Shiho Takano, Tatsuya Hiraoka, and Kimio Kuramitsu. 2023. Vocabulary replacement in SentencePiece for domain adaptation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 645–652, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Vocabulary modifications for domain-adaptive pretraining of clinical language models. In *HEALTHINF*, pages 180–188.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Siyang Liu, Naihao Deng, Sahand Sabour, Yilin Jia, Minlie Huang, and Rada Mihalcea. 2023a. Task-adaptive tokenization: Enhancing long-form text generation efficacy in mental health and beyond. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15264–15281, Singapore. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schütze. 2023b. Ofa: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. *arXiv preprint arXiv:2311.08849*.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schütze. 2024. Translico: A contrastive learning framework to address the script barrier in multilingual pretrained language models. *arXiv preprint arXiv:2401.06620*.

Zihan Liu. 2022. *Effective Transfer Learning for Low-Resource Natural Language Understanding*. Hong Kong University of Science and Technology (Hong Kong).

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *arXiv preprint arXiv:2305.08487*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 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)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, Singapore. Association for Computational Linguistics.

Vin Sachidananda, Jason Kessler, and Yi-An Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 155–165, Virtual. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Hans H Wellisch, Richard Foreman, Lee Breuer, and Robert Wilson. 1978. The conversion of scripts, its nature, history, and utilization.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Settings and Hyperparameters

The basic information of each downstream task dataset is shown in Table 8. The number of languages in each major script group for each dataset is shown in Table 9. We use the same fine-tuning hyperparameters for both transliterated evaluation (train / valid /test sets are transliterated to Latin script using `Uroman` for all languages) and non-transliterated evaluation. We introduce the detailed hyperparameters settings in the following.

Sentence Retrieval For both **SR-B** and **SR-T**, we use English-aligned sentences (up to 500 and

	|head|	|tail|	#class	measure (%)
SR-B	94	275	-	top-10 Acc.
SR-T	70	28	-	top-10 Acc.
Taxi1500	89	262	6	F1 score
SIB200	78	94	6	F1 score
NER	89	75	7	F1 score
POS	63	28	18	F1 score

Table 8: Information of the evaluation datasets and used measures. |head| (resp. |tail|): number of head (resp. tail) language-scripts according to ImaniGooghari et al. (2023) (a language-script is a head language if it is covered by XLM-R, otherwise it is tail) #class: the number of the categories if it belongs to a text classification or sequence labeling task.

	Latn	Cyrl	Hani	Arab	Deva	Other	All
SR-B	290	28	4	11	8	28	369
SR-T	64	10	3	5	2	14	98
Taxi1500	281	25	4	8	7	26	351
SIB200	117	11	0	13	6	28	172
NER	104	17	4	10	5	24	164
POS	57	8	3	5	3	15	91

Table 9: The number of languages in each script group in the evaluation datasets.

1000 for SR-B and SR-T respectively) from languages that the Glot500 and FURINA supports (head + tail languages). This evaluation type does not involve any parameter updates: we directly use each model to generate the sentence-level representation by averaging the contextual token embeddings at the **8th** layer (Jalili Sabet et al., 2020; ImaniGooghari et al., 2023) and then perform retrieval by sorting the pairwise cosine similarities.

Text Classification For both **Taxi1500** and **SIB200**, we fine-tune sequence-level classification models with a 6-classes classification head on the English train set and then select the best checkpoint using the English validation set. We train all models using Adam optimizer (Kingma and Ba, 2015) for a maximum of 40 epochs, with a learning rate of 1e-5 and an effective batch size of 16 (batch size of 8, gradient accumulation of 2). We use a single GTX 1080 Ti GPU for training. The evaluation is done in zero-shot transfer: we directly apply the best checkpoint to the test sets of all other languages.

Sequence Labeling For **NER** and **POS**, we fine-tune token-level classification models with a suitable classification head (7 for NER and 18 for POS) on the English train set and select the best checkpoint using the English validation set. We train all models using Adam optimizer for a maximum of 10 epochs. The learning rate is set to 2e-5 and the effective batch size is set to 32 (batch size of 8, gradient accumulation of 4). The training is done on a single GTX 1080 Ti GPU. The evaluation is done in zero-shot transfer: we directly apply the best checkpoint to the test sets of all other languages.

	Latn	Cyrl	Hani	Arab	Deva	Other
XLM-R	61.0	54.7	44.1	63.0	51.8	51.2
	56.9	49.4	40.7	57.3	62.1	64.8
XLM-R (Min-Merge)	58.5	54.7	44.1	63.0	51.8	51.2
	53.7	42.9	34.8	41.7	45.3	48.0
XLM-R (Average-Merge)	58.3	54.7	44.1	63.0	51.8	51.2
	53.5	42.7	34.0	41.4	45.0	47.6
XLM-R (Max-Merge)	58.2	54.7	44.1	63.0	51.8	51.2
	53.2	42.6	34.1	41.1	44.5	47.2
Glot500	45.7	39.4	42.7	42.3	44.1	44.5
	45.5	43.8	34.9	50.8	53.7	57.0
Glot500 (Min-Merge)	44.9	39.4	42.7	42.3	44.1	44.5
	43.5	36.1	30.8	36.7	38.8	39.1
Glot500 (Average-Merge)	44.9	39.4	42.7	42.3	44.1	44.5
	43.4	36.1	30.6	36.5	38.7	39.0
Glot500 (Max-Merge)	44.8	39.4	42.7	42.3	44.1	44.5
	43.3	36.0	30.9	36.3	38.4	38.9
FURINA	45.7	39.4	42.7	42.3	44.1	44.5
	45.5	43.8	34.9	50.8	53.7	57.0
FURINA (Min-Merge)	44.9	39.4	42.7	42.3	44.1	44.5
	43.5	36.1	30.8	36.7	38.8	39.1
FURINA (Average-Merge)	44.9	39.4	42.7	42.3	44.1	44.5
	43.4	36.1	30.6	36.5	38.7	39.0
FURINA (Max-Merge)	44.8	39.4	42.7	42.3	44.1	44.5
	43.3	36.0	30.9	36.3	38.4	38.9

Table 10: Full tokenization performance on SR-B dataset averaged by script group. The results on non-transliterated (resp. transliterated) data are in the first (resp. second) row for each model variant.

B Full Tokenization Performance

We further compare each model type by reporting their average sequence length on the SR-B dataset grouped by the scripts in Table 10. Glot500 and FURINA have the same tokenizers, therefore, they possess identical tokenization behavior when the same merge mode is applied. We observe that TRANSMI-modified models have consistently shorter lengths on transliterated data than the original mPLM.

C Compete Crosslingual Transfer Results

We report the complete results of the performance of all model variants on **transliterated evaluation datasets** for all tasks and languages in Table 11, 12, 13, 14 (**SR-B**), Table 15 (**SR-T**), Table 16, 17, 18, 19(**Taxi1500**), 20, 21 (**SIB200**), Table 22, 23 (**NER**), and Table 24 (**POS**).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
ace_Latn	5.2	**6.4**	6.2	6.2	46.2	45.2	48.0	**48.4**	64.4	69.0	69.6	**70.8**
ach_Latn	5.2	5.0	**5.4**	5.2	**41.6**	40.2	40.0	41.0	**51.6**	50.6	50.2	50.0
acr_Latn	3.2	**3.8**	3.6	**3.8**	24.6	26.0	**27.2**	25.8	46.8	**52.8**	52.0	**52.8**
afr_Latn	**76.6**	75.2	75.0	75.2	**69.2**	65.6	65.4	67.2	82.2	81.6	**82.6**	82.0
agw_Latn	5.4	**6.6**	6.4	6.2	**34.4**	33.4	32.4	33.0	53.0	**53.4**	45.4	47.2
ahk_Latn	3.4	3.0	2.4	**3.6**	3.4	3.6	3.4	**3.8**	8.6	10.0	9.6	**10.6**
aka_Latn	**5.8**	5.2	5.4	4.4	28.4	**34.0**	32.8	33.4	44.6	**45.6**	45.4	44.2
aln_Latn	49.4	**53.6**	52.2	52.2	53.6	57.8	58.4	**59.0**	**73.6**	73.0	73.0	73.0
als_Latn	39.8	41.2	41.0	**41.6**	48.0	50.6	51.0	**51.2**	**57.4**	56.6	56.6	56.4
alt_Cyrl	6.8	**8.4**	8.2	7.6	14.6	**21.8**	19.8	20.6	38.2	39.8	**41.0**	40.2
alz_Latn	**4.6**	4.2	**4.6**	**4.6**	**34.6**	33.6	33.8	**34.6**	**44.6**	44.4	44.2	43.4
amh_Ethi	5.6	**10.4**	9.4	9.2	4.6	15.4	**16.2**	14.6	12.8	**20.2**	**20.2**	**20.2**
aoj_Latn	3.0	**5.2**	5.0	4.4	17.4	18.8	**19.4**	18.8	26.2	**31.8**	31.4	30.0
arb_Arab	4.6	4.0	3.8	**6.2**	6.0	10.0	**10.2**	**10.2**	10.4	14.2	**15.2**	14.4
arn_Latn	**4.4**	4.0	3.6	3.8	12.0	20.6	20.6	**22.2**	28.4	**37.6**	36.6	35.6
ary_Arab	3.6	4.6	**5.0**	4.8	5.2	8.6	8.6	**10.8**	**17.0**	14.0	15.6	14.6
arz_Arab	**6.0**	4.6	5.4	4.4	6.2	**8.6**	8.4	**8.6**	13.4	**17.4**	16.2	**17.4**
asm_Beng	4.6	6.2	6.4	**7.4**	6.0	**15.6**	11.8	14.6	**32.0**	26.2	25.8	30.4
ayr_Latn	4.8	5.0	**5.2**	5.0	44.4	44.8	**45.6**	45.4	65.2	64.8	65.0	**65.4**
azb_Arab	5.4	**6.6**	6.0	6.0	6.4	**25.4**	24.6	24.8	23.8	**45.6**	45.2	45.0
aze_Latn	24.4	48.4	51.0	**52.8**	37.2	54.4	59.6	**59.8**	72.4	72.0	73.2	**73.8**
bak_Cyrl	6.4	**7.0**	6.6	**7.0**	12.0	26.4	**26.8**	24.4	42.2	**43.6**	43.2	41.4
bam_Latn	**7.4**	6.8	6.8	7.0	32.0	37.2	**39.0**	38.8	53.6	54.6	**57.2**	55.6
ban_Latn	9.0	8.6	8.2	**9.4**	33.0	**33.6**	32.8	32.8	61.0	**62.4**	**62.4**	62.0
bar_Latn	13.6	17.2	**17.6**	15.2	**34.8**	29.6	31.8	31.2	65.0	**68.6**	68.2	67.8
bba_Latn	5.2	6.0	**6.2**	5.8	18.4	25.6	**26.4**	25.2	31.0	**33.8**	33.6	33.2
bbc_Latn	**7.8**	7.2	7.6	7.2	**57.2**	51.8	52.2	52.4	71.4	71.8	**72.4**	72.2
bci_Latn	**5.6**	5.0	5.4	5.2	14.4	15.0	**16.4**	16.2	35.0	**36.6**	35.2	34.2
bcl_Latn	10.2	10.6	**10.8**	**10.8**	**79.8**	78.0	78.2	78.0	85.8	86.2	**86.6**	86.0
bel_Cyrl	7.4	22.2	24.8	**25.8**	10.2	23.2	22.6	**24.2**	**46.6**	41.6	43.0	46.4
bem_Latn	6.6	**7.2**	6.4	6.8	**58.6**	56.2	56.4	56.8	59.0	58.8	59.0	**59.4**
ben_Beng	6.0	7.2	6.4	**8.8**	7.2	10.2	11.0	**14.4**	38.4	32.2	33.8	**39.6**
bhw_Latn	**5.0**	4.0	4.4	4.4	32.4	**34.6**	34.4	33.6	50.2	50.8	51.0	**51.8**
bim_Latn	**4.2**	3.4	**4.2**	**4.2**	41.2	**41.8**	40.2	40.4	57.6	**58.8**	57.8	57.0
bis_Latn	**7.0**	6.2	5.8	5.4	**48.6**	47.6	47.6	47.8	65.6	65.6	66.0	**66.4**
bod_Tibt	2.0	**2.2**	1.8	**2.2**	2.8	20.0	19.6	**20.2**	4.8	26.4	26.0	**27.2**
bqc_Latn	**5.0**	**5.0**	4.8	4.8	**10.2**	9.2	9.4	10.0	**14.0**	12.6	11.8	12.6
bre_Latn	**17.2**	16.2	15.2	15.8	**30.4**	27.4	28.0	28.4	**59.2**	56.6	56.8	55.6
bts_Latn	6.0	**6.6**	6.0	**6.6**	**56.4**	54.2	54.2	55.0	70.8	70.4	70.2	**71.0**
btx_Latn	11.0	11.0	**11.8**	**11.8**	**59.6**	59.0	58.0	57.8	**71.4**	70.0	70.4	70.4
bul_Cyrl	15.6	24.8	25.8	**30.8**	18.6	35.0	35.0	**36.0**	**75.4**	62.8	63.8	64.8
bum_Latn	**6.0**	5.4	5.2	5.6	18.0	**21.0**	**21.0**	20.0	35.8	**37.4**	37.0	36.0
bzj_Latn	**7.8**	**7.8**	6.6	6.2	**75.0**	74.0	74.0	74.0	84.4	85.6	85.6	**85.8**
cab_Latn	4.8	5.8	5.8	**6.4**	17.6	19.8	**21.2**	19.8	26.2	30.4	**31.0**	30.4
cac_Latn	**3.2**	2.8	2.8	3.0	14.2	**16.2**	14.6	14.0	30.2	35.8	34.8	**36.6**
cak_Latn	3.8	**4.2**	3.8	4.0	19.8	19.4	19.0	**20.6**	42.2	**46.2**	45.8	46.0
caq_Latn	3.6	4.4	4.2	**5.2**	8.4	15.2	14.8	**15.8**	21.6	**29.6**	29.4	**29.6**
cat_Latn	81.6	81.4	**81.8**	81.6	**73.6**	68.8	69.4	69.4	**84.2**	82.2	**84.2**	**84.2**
cbk_Latn	30.2	33.2	**34.6**	33.8	54.6	**55.2**	53.4	54.8	76.6	76.6	**77.0**	76.6
cce_Latn	5.2	**6.6**	6.2	6.0	**51.6**	46.6	46.4	47.0	68.6	69.2	69.8	**70.4**
ceb_Latn	**14.2**	13.8	13.8	13.2	**68.0**	65.6	67.2	67.8	81.4	82.0	**82.2**	**82.2**
ces_Latn	42.6	**50.6**	49.6	50.0	29.6	36.6	37.2	**38.4**	**66.2**	65.2	65.0	64.0
cfm_Latn	5.0	**5.4**	**5.4**	5.0	46.2	**47.0**	46.2	46.8	60.6	60.8	**61.4**	**61.4**
che_Cyrl	4.4	4.0	5.0	**5.8**	5.2	7.4	**7.8**	7.2	**15.6**	13.8	13.6	13.4
chk_Latn	5.0	4.8	**5.6**	5.0	30.4	33.4	36.8	**37.6**	59.8	**60.6**	60.4	**60.6**
chv_Cyrl	5.8	**6.2**	**6.2**	5.8	8.2	**14.0**	13.2	13.2	20.4	23.2	22.6	**23.4**
ckb_Arab	3.8	**5.4**	**5.4**	5.2	4.6	15.0	**15.2**	14.0	20.0	25.4	**25.6**	**25.6**
cmn_Hani	4.0	9.8	8.0	**13.8**	5.8	13.2	14.2	**15.0**	11.6	14.2	14.8	**18.0**
cnh_Latn	4.2	**4.6**	4.4	3.8	**55.0**	54.2	54.0	53.8	**63.8**	63.2	62.6	62.6
crh_Cyrl	12.8	**18.2**	17.0	15.6	31.8	43.8	**45.4**	45.2	**69.2**	63.6	67.4	64.2
crs_Latn	7.4	**8.2**	7.4	8.0	**80.6**	77.4	78.0	78.2	**84.2**	82.0	82.8	82.4
csy_Latn	3.8	**4.4**	3.8	**4.4**	**50.0**	49.4	48.4	49.4	**64.4**	64.2	63.2	63.6
ctd_Latn	**4.2**	**4.2**	4.0	3.2	**59.4**	57.4	58.2	57.8	**63.6**	62.0	61.6	62.2
ctu_Latn	3.8	3.6	4.4	**4.6**	14.4	15.6	**16.8**	14.6	33.6	**34.0**	32.6	31.4
cuk_Latn	**5.2**	5.0	4.6	4.8	**22.8**	20.6	20.4	20.6	40.6	42.8	**43.6**	43.2
cym_Latn	**38.6**	35.6	35.8	35.0	**43.0**	35.4	37.4	36.4	**60.4**	59.2	60.0	58.2
dan_Latn	60.2	**67.8**	**67.8**	67.4	54.4	56.2	**59.0**	58.0	75.6	75.8	**76.0**	75.2
deu_Latn	72.2	78.6	78.8	**79.6**	61.4	62.4	**64.6**	64.0	81.8	81.8	**82.2**	81.4
djk_Latn	4.4	4.4	4.0	**5.0**	**40.4**	40.2	**40.4**	39.8	55.0	57.2	**57.6**	**57.6**
dln_Latn	5.0	4.6	**5.2**	**5.2**	**57.2**	55.4	56.0	56.0	68.0	68.6	68.4	**68.8**
dtp_Latn	**5.2**	4.8	4.8	4.2	**25.0**	24.4	23.6	24.6	41.2	**45.4**	45.0	**45.4**
dyu_Latn	**5.4**	4.4	5.2	4.4	29.4	30.6	**31.8**	31.6	50.2	55.0	**56.4**	56.0
dzo_Tibt	1.8	1.4	**2.0**	1.8	2.0	16.8	16.0	**17.6**	4.0	**29.6**	27.4	29.4
efi_Latn	5.8	**8.4**	6.8	7.4	31.0	**42.6**	42.4	42.4	51.2	**56.6**	56.2	56.2
ell_Grek	5.4	12.2	11.0	**13.2**	10.4	16.6	16.4	**16.8**	33.0	31.2	**35.6**	34.4
enm_Latn	**39.8**	38.2	38.6	38.6	**66.0**	64.0	64.4	64.6	**75.6**	75.2	75.4	75.4
epo_Latn	60.4	59.8	**61.4**	60.8	**53.0**	52.0	52.0	51.0	73.8	73.2	**74.4**	74.0
est_Latn	45.4	64.0	**65.8**	63.6	38.4	47.6	49.4	**51.2**	66.2	67.8	**68.0**	67.6
eus_Latn	**26.2**	25.4	25.8	25.2	**23.2**	22.2	22.0	22.2	**36.8**	35.6	35.8	36.6
ewe_Latn	5.0	6.0	6.0	**6.2**	18.0	**27.2**	26.4	27.0	31.8	39.2	40.6	**41.0**
fao_Latn	14.8	19.2	**20.0**	19.8	32.4	47.6	**53.2**	51.6	73.8	74.6	**76.0**	**76.0**
fas_Arab	4.6	13.8	12.0	**18.2**	6.8	24.8	30.2	**32.6**	34.6	44.2	47.6	**51.6**
fij_Latn	3.8	**4.0**	**4.0**	3.8	**36.4**	35.6	35.0	35.2	43.6	43.8	42.6	**45.6**
fil_Latn	**60.0**	59.4	59.4	56.6	**72.0**	68.4	67.2	67.2	84.8	84.0	**85.2**	**85.2**
fin_Latn	31.6	66.0	66.6	**68.0**	28.8	44.0	46.8	**49.8**	55.8	65.6	67.4	**67.8**
fon_Latn	3.6	3.8	3.8	**4.0**	9.0	10.2	**11.2**	**11.2**	18.8	23.6	**23.8**	23.0
fra_Latn	**81.4**	81.0	81.2	**81.4**	**76.8**	71.8	72.6	74.0	**90.2**	88.8	89.6	89.6
fry_Latn	20.8	21.2	**21.4**	**21.4**	**38.0**	34.6	34.8	35.6	**71.6**	69.6	69.6	70.2
gaa_Latn	**6.6**	**6.6**	5.2	6.0	12.0	18.0	**18.4**	18.0	31.2	**41.4**	39.6	39.2
gil_Latn	**5.6**	5.0	4.0	4.2	**36.8**	34.8	34.0	35.6	52.6	53.4	53.6	**55.2**
giz_Latn	5.6	**5.8**	5.6	**5.8**	23.2	**28.2**	27.6	27.4	41.6	48.8	**49.8**	**49.8**
gkn_Latn	5.8	**6.8**	5.6	**6.8**	8.8	9.8	10.0	**11.4**	19.0	**21.0**	20.6	20.6
gkp_Latn	4.8	**5.2**	4.2	4.8	**9.2**	8.8	8.4	8.8	14.4	15.8	**17.8**	16.8

Table 11: Top-10 accuracy of models on **transliterated** dataset of **SR-B** (Part I).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
gla_Latn	25.0	25.8	**26.4**	**26.4**	42.4	42.0	**42.6**	42.2	57.2	60.4	**61.8**	61.0
gle_Latn	23.6	27.0	27.4	**29.6**	28.4	29.8	29.8	**30.4**	51.4	53.2	**54.4**	52.4
glv_Latn	5.8	5.0	**6.0**	5.4	**47.4**	39.4	38.8	39.0	**58.2**	54.2	55.6	54.8
gom_Latn	5.8	6.8	**7.0**	6.6	**44.6**	40.2	40.8	41.0	**58.0**	56.6	57.6	56.0
gor_Latn	3.8	4.0	3.8	**4.2**	27.4	27.6	**28.6**	28.2	46.4	46.4	46.6	**47.0**
grc_Grek	4.4	**6.8**	5.4	5.8	10.0	**15.8**	14.4	13.0	14.6	13.2	**14.8**	**14.8**
guc_Latn	**3.4**	3.2	3.2	2.8	6.0	11.4	**11.8**	**11.8**	15.4	**22.0**	21.6	20.6
gug_Latn	4.0	4.8	**5.4**	4.8	22.8	27.8	**28.4**	27.8	31.6	34.8	**35.4**	34.8
guj_Gujr	5.8	10.0	6.8	**13.8**	7.4	18.4	16.4	**20.2**	32.8	31.2	34.6	**41.8**
gur_Latn	**5.2**	4.2	4.4	4.0	13.6	18.4	19.8	**21.0**	32.8	**41.2**	40.2	40.2
guw_Latn	5.0	**5.8**	5.6	5.0	11.8	**26.6**	23.6	23.4	32.4	**42.4**	41.4	**42.4**
gya_Latn	5.8	5.0	5.0	**6.6**	12.4	13.2	13.2	**13.4**	22.8	24.4	**25.0**	24.0
gym_Latn	4.0	**4.4**	3.4	3.0	8.8	13.8	14.2	**15.0**	20.6	27.6	27.8	**28.6**
hat_Latn	**5.0**	4.4	4.6	4.6	62.8	62.2	62.0	**64.2**	**80.0**	79.8	79.8	79.6
hau_Latn	30.8	30.6	**31.2**	28.4	**51.6**	48.2	48.8	49.8	**67.8**	67.2	67.0	67.0
haw_Latn	**4.2**	**4.2**	4.0	**4.2**	**38.8**	36.4	36.4	34.8	61.6	62.8	62.8	**63.2**
heb_Hebr	5.0	7.4	6.2	**9.8**	5.0	6.4	**8.4**	7.8	12.6	12.6	13.2	**13.8**
hif_Latn	15.2	16.6	15.6	**18.2**	**44.6**	29.8	31.2	31.4	**73.4**	67.4	66.8	64.0
hil_Latn	11.0	12.4	**13.0**	12.2	**76.2**	72.8	73.4	72.8	89.2	89.0	89.0	**89.6**
hin_Deva	13.0	19.0	14.0	**29.4**	26.2	33.4	40.2	**45.4**	**65.0**	52.8	57.6	61.8
hin_Latn	13.6	**15.0**	11.8	12.0	**43.2**	30.2	30.8	32.0	**64.6**	58.4	59.0	57.6
hmo_Latn	6.4	6.2	**6.6**	6.0	**48.2**	47.6	47.6	47.0	**60.0**	59.6	59.8	57.8
hne_Deva	5.8	6.6	6.2	**8.4**	7.2	19.2	20.6	**23.4**	33.8	40.2	40.8	**45.0**
hnj_Latn	2.8	3.2	**3.6**	3.2	54.2	53.8	53.4	**54.8**	64.0	64.6	65.2	**65.4**
hra_Latn	5.2	5.4	**5.6**	5.4	**45.4**	43.2	44.6	45.2	**56.6**	53.6	55.0	55.2
hrv_Latn	**69.0**	63.0	63.8	63.6	**66.6**	61.0	61.6	62.0	**74.8**	71.4	71.6	71.2
hui_Latn	**3.8**	2.8	2.6	3.2	**27.8**	26.0	26.4	26.4	30.6	**31.4**	**31.4**	**31.4**
hun_Latn	39.0	**58.0**	57.8	55.6	24.2	35.2	36.0	**36.8**	56.4	**62.6**	62.2	61.8
hus_Latn	**4.2**	3.6	3.6	4.0	15.8	**19.2**	19.0	18.8	**40.2**	39.8	40.0	39.8
hye_Armn	5.4	**9.0**	6.4	8.2	11.2	18.6	19.8	**21.4**	25.4	28.4	**31.4**	30.4
iba_Latn	14.4	**15.8**	15.6	14.8	**66.0**	65.0	65.6	64.6	71.4	72.4	72.4	**72.6**
ibo_Latn	6.6	6.4	**7.0**	6.4	**25.2**	22.2	22.8	23.6	41.0	**41.8**	41.4	41.4
ifa_Latn	**4.4**	4.0	**4.4**	4.0	**39.2**	38.6	37.4	38.2	52.2	**53.0**	52.6	52.8
ifb_Latn	4.8	**5.0**	4.8	4.4	**36.6**	35.0	35.0	35.4	52.0	**53.4**	52.6	53.0
ikk_Latn	5.2	5.2	5.0	**5.8**	11.0	17.4	**18.4**	16.8	26.6	35.6	**36.0**	**36.0**
ilo_Latn	6.2	**7.0**	6.6	6.4	**55.0**	51.8	51.8	51.8	73.6	73.6	**74.4**	73.6
ind_Latn	**82.6**	80.2	79.8	79.2	72.2	71.4	72.2	**72.8**	77.8	77.8	**78.4**	78.2
isl_Latn	16.6	**34.6**	**34.6**	33.2	24.6	38.8	**39.8**	38.6	63.6	67.6	67.2	**68.0**
ita_Latn	**71.6**	70.2	71.4	71.4	**68.8**	64.8	65.8	66.2	**79.4**	79.0	79.2	79.2
ium_Latn	**3.2**	2.4	2.4	2.4	**24.8**	24.6	24.4	**24.8**	**38.6**	37.4	37.8	38.2
ixl_Latn	3.0	3.4	**3.8**	3.4	12.2	**16.6**	15.2	15.2	26.0	**31.8**	31.6	**31.8**
izz_Latn	**4.6**	4.2	4.0	4.0	16.2	**18.0**	17.2	17.0	32.2	**33.0**	31.4	30.6
jam_Latn	6.6	**7.4**	7.2	6.0	**67.8**	66.4	67.4	66.0	85.2	84.8	**86.2**	**86.2**
jav_Latn	**28.4**	26.0	27.2	27.0	**50.4**	45.8	46.0	46.0	**70.8**	68.8	68.4	67.8
jpn_Jpan	3.4	13.2	11.6	**14.4**	4.8	10.6	12.0	**12.6**	11.8	18.8	18.0	**19.0**
kaa_Cyrl	7.2	16.2	16.0	**18.6**	27.4	38.6	**44.2**	43.8	61.2	63.2	**66.6**	65.0
kaa_Latn	9.2	12.2	**12.8**	12.2	34.8	35.0	36.2	**37.0**	71.4	70.2	**73.6**	72.6
kab_Latn	6.0	7.0	**7.2**	6.8	**16.8**	16.2	16.2	16.4	**30.8**	29.8	29.4	30.0
kac_Latn	3.6	3.6	3.0	**3.8**	26.4	**27.4**	26.0	25.6	**45.8**	45.2	44.6	44.0
kal_Latn	3.4	**4.8**	4.2	4.2	**23.0**	18.6	18.0	19.0	**22.8**	20.6	20.6	20.2
kan_Knda	4.0	14.2	12.0	**17.8**	7.2	13.8	15.2	**17.4**	27.2	**33.8**	33.4	31.2
kat_Geor	6.0	**12.6**	9.6	11.0	7.0	14.6	**16.6**	16.0	16.6	23.0	**23.2**	22.6
kaz_Cyrl	4.0	**18.2**	17.0	18.0	15.2	27.6	30.0	**31.8**	51.0	49.6	**52.6**	51.8
kbp_Latn	**5.4**	4.4	4.2	5.0	6.6	11.6	**11.8**	10.8	19.2	24.0	**25.2**	24.8
kek_Latn	**4.0**	3.2	3.6	3.8	23.6	**24.6**	22.0	23.0	45.2	**46.4**	45.4	44.6
khm_Khmr	2.4	**11.6**	9.2	9.8	2.8	14.4	15.8	**16.4**	6.4	23.8	**25.4**	22.8
kia_Latn	**5.2**	**5.2**	4.8	4.4	25.2	**27.4**	26.8	27.2	43.2	**43.4**	**43.4**	43.0
kik_Latn	5.2	**5.8**	4.6	5.4	14.6	**26.0**	24.2	23.0	37.0	**50.4**	50.2	48.2
kin_Latn	5.0	**6.6**	6.0	6.4	**59.6**	50.6	50.2	51.2	**69.0**	66.4	66.8	67.0
kir_Cyrl	6.8	23.4	21.2	**24.2**	22.4	34.0	37.6	**39.4**	56.0	57.6	**62.0**	60.0
kjb_Latn	**4.4**	3.8	3.8	3.8	29.6	**32.8**	30.2	30.0	54.0	**55.4**	54.8	55.2
kjh_Cyrl	4.4	4.8	5.4	**5.8**	9.0	**18.8**	18.6	16.8	26.0	**30.6**	30.0	30.2
kmm_Latn	4.8	4.2	**5.2**	4.6	**38.6**	36.4	36.6	36.8	**52.8**	51.0	51.0	51.2
kmr_Cyrl	5.0	**7.2**	6.8	5.8	7.8	13.6	**14.8**	14.0	**28.6**	23.8	25.6	23.4
kmr_Latn	11.6	**16.6**	15.4	15.8	23.4	28.6	**29.8**	29.6	49.2	53.4	53.8	**54.0**
knv_Latn	2.6	2.8	2.8	**3.0**	11.8	11.2	11.8	**12.2**	**21.0**	20.8	20.8	**21.0**
kor_Hang	2.8	20.8	21.2	**24.8**	3.8	16.2	17.0	**19.4**	9.8	23.2	25.2	**27.0**
kpg_Latn	5.2	5.8	5.8	**6.0**	**51.8**	49.6	48.8	49.8	**61.6**	61.0	60.6	61.0
krc_Cyrl	6.2	12.2	**13.0**	**13.0**	16.8	**23.2**	22.4	22.6	44.2	**45.6**	45.0	44.4
kri_Latn	6.6	**7.8**	7.2	**7.8**	38.2	38.2	40.0	**40.6**	66.0	66.6	**67.2**	**67.2**
ksd_Latn	**7.0**	5.8	5.2	5.4	**42.6**	42.4	42.4	42.4	**53.6**	**53.6**	53.4	52.6
kss_Latn	**4.2**	3.8	3.6	**4.2**	7.2	10.2	11.8	**12.2**	21.0	27.2	**29.0**	28.0
ksw_Mymr	**3.4**	2.8	3.0	**3.4**	2.8	**10.6**	10.0	8.8	6.4	21.2	20.6	**22.2**
kua_Latn	4.8	5.0	5.4	**5.6**	**43.8**	40.2	40.2	40.8	54.4	54.0	55.6	**55.8**
lam_Latn	**6.8**	5.8	6.0	6.2	**25.2**	23.4	23.2	24.8	35.4	**36.8**	36.2	35.6
lao_Laoo	2.6	10.2	7.0	**11.0**	3.2	12.2	13.6	**14.0**	10.2	21.2	25.0	**26.8**
lat_Latn	**53.0**	51.0	52.0	51.0	**49.8**	47.0	47.2	46.8	**57.0**	56.0	56.4	56.6
lav_Latn	23.4	44.6	44.6	**45.0**	22.2	30.6	**32.4**	32.2	53.8	55.8	**56.2**	56.0
ldi_Latn	5.6	**6.6**	6.2	5.8	**28.8**	26.6	27.2	26.8	49.0	49.0	**49.2**	48.4
leh_Latn	**5.6**	5.2	5.2	5.2	**58.6**	53.8	54.4	54.0	67.6	67.2	68.4	**68.6**
lhu_Latn	2.8	2.8	**3.4**	3.0	4.0	4.0	**4.2**	4.0	11.6	12.4	12.4	**12.6**
lin_Latn	6.8	7.2	7.0	**8.0**	**65.6**	63.6	64.2	63.6	**69.6**	68.2	68.4	68.6
lit_Latn	43.4	50.0	50.2	**51.6**	31.0	33.2	34.6	**34.8**	58.4	57.2	**58.6**	58.2
loz_Latn	7.2	6.6	**8.0**	7.0	48.0	47.2	**48.2**	47.0	67.2	68.2	68.6	**69.0**
ltz_Latn	**13.0**	12.8	12.8	11.2	66.8	**67.8**	67.6	67.2	82.4	**83.2**	82.8	82.8
lug_Latn	5.0	**5.4**	5.2	4.2	**48.8**	39.6	39.4	37.8	**50.2**	46.4	47.4	46.4
luo_Latn	**6.4**	6.0	**6.4**	5.8	40.8	39.4	40.2	**41.0**	53.2	**53.6**	**53.6**	53.4
lus_Latn	4.4	5.0	5.6	**5.8**	**49.4**	47.8	47.8	49.0	64.2	64.0	64.4	**64.6**
lzh_Hani	3.2	3.4	3.0	**3.6**	4.4	4.6	5.0	**5.2**	2.4	3.0	3.0	**3.4**
mad_Latn	7.6	8.2	**8.4**	7.4	**44.4**	41.6	41.4	41.4	**63.6**	60.0	60.0	60.4
mah_Latn	**6.4**	5.8	5.8	5.6	27.8	**29.8**	28.6	28.0	45.0	45.8	**46.0**	45.6
mai_Deva	5.4	6.4	5.8	**8.2**	6.6	18.6	20.8	**21.8**	35.2	**43.2**	40.2	41.8
mal_Mlym	5.0	14.4	13.6	**18.4**	4.8	11.8	18.0	**22.0**	14.0	27.8	29.0	**29.4**

Table 12: Top-10 accuracy of models on **transliterated** dataset of **SR-B** (Part II).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
mam_Latn	3.6	**4.2**	3.6	3.4	**13.4**	12.4	12.2	**13.4**	30.0	33.2	31.2	**34.0**
mar_Deva	4.2	20.6	20.2	**26.0**	6.8	20.8	24.8	**31.8**	25.0	39.2	41.6	**46.0**
mau_Latn	2.4	**2.8**	2.6	**2.8**	**3.8**	**3.8**	3.2	3.0	**7.2**	**7.2**	6.8	7.0
mbb_Latn	3.6	3.4	**4.2**	3.0	30.6	**36.2**	35.0	**36.2**	47.6	50.8	51.2	**51.4**
mck_Latn	5.2	**5.6**	5.0	5.0	**57.8**	53.4	53.8	54.4	**65.0**	**65.0**	63.4	64.6
mcn_Latn	6.8	**7.4**	7.2	7.0	**32.0**	27.4	28.8	30.2	40.6	40.8	41.4	**41.6**
mco_Latn	3.0	2.4	2.8	**3.4**	**7.2**	7.0	**7.2**	6.0	17.4	**19.0**	18.8	17.8
mdy_Ethi	**3.4**	3.2	3.2	3.0	4.6	12.8	**13.6**	11.8	10.6	**16.4**	15.0	14.8
meu_Latn	5.6	6.4	**6.8**	**6.8**	**52.2**	49.6	49.4	51.2	59.2	59.0	**60.2**	60.0
mfe_Latn	9.0	**10.0**	9.2	**10.0**	**78.6**	73.8	74.6	75.2	77.2	77.4	77.6	**77.8**
mgh_Latn	5.2	4.0	**5.4**	4.6	**23.6**	21.2	20.6	19.8	**55.0**	54.2	53.2	52.8
mgr_Latn	4.0	4.2	**4.6**	4.4	**57.6**	53.6	54.2	53.2	**64.6**	64.0	64.0	63.8
mhr_Cyrl	5.8	5.0	5.6	**6.2**	9.6	14.8	16.6	**17.0**	22.4	27.6	26.4	**28.4**
min_Latn	9.4	9.6	**10.0**	9.6	**29.0**	25.8	25.8	26.2	54.6	**55.8**	55.2	55.2
miq_Latn	5.2	6.0	**6.4**	6.2	**45.4**	44.8	43.6	43.6	50.0	**51.0**	50.4	50.6
mkd_Cyrl	24.6	33.2	**33.8**	33.4	37.0	48.6	49.6	**50.2**	**81.2**	68.4	68.8	68.8
mlg_Latn	**29.2**	24.8	25.4	25.6	**65.2**	62.8	64.4	61.8	**64.8**	63.4	64.2	64.4
mlt_Latn	5.4	**7.4**	**7.4**	6.6	38.0	**39.6**	38.2	38.4	71.4	**72.4**	71.6	70.8
mos_Latn	5.0	4.8	**5.6**	3.8	12.6	15.0	17.4	**17.8**	25.2	33.0	33.6	**34.6**
mps_Latn	3.2	**3.4**	**3.4**	**3.4**	22.2	**23.0**	**23.0**	22.6	27.6	27.6	27.6	**29.0**
mri_Latn	4.2	5.8	**6.0**	5.6	**48.4**	45.6	45.2	45.8	**72.4**	71.2	70.8	71.0
mrw_Latn	6.0	**6.6**	6.4	6.2	**52.2**	51.4	50.4	51.6	61.8	64.6	64.6	**65.4**
msa_Latn	**40.6**	40.4	**40.6**	**40.6**	**41.4**	40.2	40.4	40.8	46.0	46.2	**46.8**	46.4
mwm_Latn	6.0	**6.4**	5.2	**6.4**	7.4	**7.8**	7.6	7.4	15.4	17.0	17.4	**17.6**
mxv_Latn	**3.6**	2.8	3.2	2.8	6.2	7.8	**8.4**	7.2	16.8	**19.2**	18.2	17.4
mya_Mymr	3.0	8.4	4.4	**10.2**	3.2	7.4	7.8	**10.8**	6.2	11.6	14.8	**17.4**
myv_Cyrl	**4.8**	4.4	4.6	4.4	7.0	9.0	10.0	**10.2**	**23.4**	20.0	21.0	21.4
mzh_Latn	3.0	3.6	3.6	**4.6**	17.6	**23.2**	21.0	22.0	28.8	37.0	36.0	**37.6**
nan_Latn	3.8	3.6	**4.6**	4.4	7.0	8.6	**9.2**	8.4	15.8	15.0	15.2	**16.6**
naq_Latn	3.8	**4.0**	**4.0**	3.4	11.0	**17.0**	**17.0**	16.0	20.6	31.6	**32.2**	**32.2**
nav_Latn	**3.6**	2.8	3.0	2.8	7.0	7.4	7.2	**7.8**	12.8	**13.2**	11.8	12.2
nbl_Latn	9.2	9.2	**10.0**	9.8	**53.8**	49.2	49.8	49.8	**62.6**	60.2	59.8	59.8
nch_Latn	**4.4**	**4.4**	3.6	4.0	**21.4**	19.2	19.2	19.0	47.4	48.0	48.0	**48.4**
ncj_Latn	**4.0**	3.6	3.4	3.8	**24.4**	22.4	22.2	23.0	**49.8**	46.0	46.2	45.4
ndc_Latn	**5.2**	4.4	4.2	4.2	**40.0**	35.2	35.4	34.6	55.4	**56.2**	56.0	55.0
nde_Latn	**13.0**	12.2	12.8	12.0	**53.8**	51.2	52.8	53.0	62.0	61.4	**62.4**	**62.4**
ndo_Latn	**5.2**	**5.2**	4.2	4.6	**48.2**	43.6	43.6	43.2	**63.8**	63.4	63.6	63.4
nds_Latn	**9.6**	9.0	8.4	8.6	36.6	36.2	**37.6**	36.4	66.4	66.4	**67.6**	66.4
nep_Deva	4.2	18.8	17.0	**23.4**	9.6	24.2	27.0	**31.2**	33.4	43.2	46.6	**50.4**
ngu_Latn	4.4	5.4	**5.6**	5.4	**27.8**	27.4	27.2	27.2	52.8	**55.0**	54.0	54.4
nia_Latn	3.6	**5.2**	4.4	**5.2**	20.2	**24.6**	23.8	**24.6**	42.0	45.0	47.0	**47.2**
nld_Latn	**77.8**	76.4	76.8	76.8	**71.6**	69.8	69.8	70.0	84.2	84.6	**84.8**	84.0
nmf_Latn	4.0	5.2	4.8	**5.6**	30.2	29.6	29.8	**30.4**	31.6	33.0	**34.6**	33.4
nnb_Latn	**5.0**	4.4	4.2	4.2	**51.8**	36.6	38.2	38.0	**58.8**	54.0	55.0	55.2
nno_Latn	48.8	52.4	**53.4**	53.0	64.0	62.8	64.0	**65.4**	**80.0**	78.2	78.6	77.8
nob_Latn	68.8	76.6	78.2	**78.6**	66.4	71.8	74.4	**75.6**	84.6	85.4	**87.0**	86.6
nor_Latn	66.0	73.0	**76.4**	76.2	75.4	78.4	80.2	**80.8**	84.8	84.0	84.6	**85.0**
npi_Deva	5.0	16.2	16.2	**21.0**	8.4	27.2	32.8	**33.8**	41.0	49.0	52.4	**56.8**
nse_Latn	5.0	**6.6**	6.2	6.2	**49.6**	45.8	46.6	47.0	**70.2**	67.2	67.0	67.2
nso_Latn	**5.4**	5.2	5.2	5.2	**54.2**	53.6	53.4	52.6	68.4	67.2	**68.6**	67.8
nya_Latn	3.8	**4.8**	**4.8**	4.6	**60.6**	56.4	57.6	58.2	64.2	65.8	66.0	**66.2**
nyn_Latn	4.4	4.2	**5.6**	4.8	**51.8**	38.8	38.6	37.2	**60.4**	55.2	55.2	54.2
nyy_Latn	4.2	4.8	5.0	**5.2**	**28.6**	25.4	26.0	26.0	**53.6**	52.4	**53.6**	53.4
nzi_Latn	4.2	**4.8**	4.0	4.0	16.8	19.6	**21.4**	21.0	33.2	**39.6**	38.4	38.6
ori_Orya	4.4	11.6	7.8	**13.8**	5.0	17.4	16.4	**18.6**	**35.4**	27.2	26.2	31.6
ory_Orya	4.0	10.0	6.6	**13.0**	4.4	12.6	12.2	**14.2**	**29.4**	24.0	25.6	27.6
oss_Cyrl	3.4	**3.8**	3.6	**3.8**	5.8	16.2	17.4	**18.0**	16.6	35.8	**38.0**	35.8
ote_Latn	4.0	3.6	3.2	**4.2**	9.4	12.4	12.0	**12.6**	20.8	25.0	**26.6**	24.6
pag_Latn	8.0	7.8	**8.6**	8.2	**61.2**	55.8	57.4	56.2	**76.8**	75.0	75.2	75.0
pam_Latn	8.2	7.8	7.8	**8.4**	**49.8**	49.0	48.4	48.2	**77.0**	76.0	76.6	76.2
pan_Guru	4.2	10.6	8.6	**14.2**	5.2	11.8	12.8	**19.6**	39.2	34.8	36.2	**39.6**
pap_Latn	12.4	**13.2**	12.4	12.6	**70.4**	69.4	69.8	69.0	**78.8**	77.8	77.8	78.2
pau_Latn	**4.4**	3.8	**4.4**	3.8	**29.8**	28.2	28.4	28.8	46.4	47.8	**48.0**	**48.0**
pcm_Latn	13.6	14.4	**14.6**	14.2	**66.8**	65.4	66.2	66.4	73.2	72.8	**73.6**	73.4
pdt_Latn	9.4	**9.6**	9.2	9.2	61.2	61.4	62.4	**63.2**	77.2	76.4	76.2	**77.4**
pes_Arab	4.2	15.0	13.6	**19.0**	6.2	21.2	22.4	**27.2**	30.4	41.4	44.4	**45.4**
pis_Latn	6.4	7.2	7.0	**7.4**	**57.2**	55.2	54.6	54.4	74.6	**75.2**	75.0	75.0
pls_Latn	5.0	**5.6**	5.0	**5.6**	**30.4**	29.4	28.8	29.2	**53.4**	51.4	51.4	52.4
plt_Latn	**26.8**	24.6	24.6	24.0	**60.0**	56.6	57.4	56.4	68.0	67.0	68.0	**68.2**
poh_Latn	3.2	**3.4**	3.2	3.2	**15.0**	13.2	13.8	14.0	**31.2**	28.2	28.4	27.6
pol_Latn	61.2	64.8	**67.8**	67.0	42.6	44.4	47.4	**48.2**	**73.8**	72.8	73.6	73.6
pon_Latn	**5.8**	5.6	5.6	**5.8**	**21.6**	20.4	20.8	20.6	**36.4**	34.6	34.8	33.4
por_Latn	75.8	78.2	77.6	**79.2**	67.4	68.2	70.6	**71.2**	82.6	82.4	82.8	**83.2**
prk_Latn	3.6	**4.0**	3.2	3.4	49.8	51.6	51.6	**52.4**	**58.8**	58.2	57.6	57.6
prs_Arab	3.8	15.2	14.0	**19.0**	6.2	23.6	24.8	**29.2**	32.2	45.8	47.0	**53.0**
pxm_Latn	**4.4**	**4.4**	4.0	3.8	10.2	17.2	**17.8**	17.2	22.6	**34.6**	34.0	34.2
qub_Latn	4.8	**5.2**	5.0	5.0	38.6	38.2	38.8	**40.0**	**47.0**	42.2	42.2	42.4
quc_Latn	**3.2**	3.0	**3.2**	3.0	22.0	**24.4**	22.8	22.6	**46.0**	45.2	**46.0**	45.4
qug_Latn	4.8	5.6	5.8	**6.4**	**49.6**	49.4	48.4	49.0	**68.6**	66.2	66.2	66.4
quh_Latn	**5.2**	3.6	3.6	4.2	**54.8**	52.6	51.8	52.2	**61.0**	59.2	58.0	57.8
quw_Latn	**6.8**	6.4	5.4	6.6	**47.0**	**47.0**	45.8	46.6	**57.4**	57.2	57.0	57.2
quy_Latn	5.0	5.2	**5.6**	5.4	**59.8**	58.2	57.8	59.6	51.4	52.2	52.8	**53.4**
quz_Latn	5.0	4.4	**5.4**	4.4	**65.0**	61.0	61.8	62.2	65.2	**66.0**	65.6	65.6
qvi_Latn	4.2	**5.2**	4.4	4.8	**44.8**	43.8	43.2	43.8	**70.4**	70.0	69.8	70.2
rap_Latn	**4.4**	4.2	**4.4**	3.6	16.2	**20.0**	18.6	18.4	32.6	**45.4**	43.2	41.4
rar_Latn	3.4	3.2	3.2	**3.6**	31.6	31.8	**33.2**	32.2	44.4	48.2	**49.2**	48.4
rmy_Latn	6.8	**7.2**	**7.2**	7.0	**33.2**	31.8	31.4	31.6	**62.6**	60.8	59.8	58.8
ron_Latn	66.0	**68.4**	68.2	68.0	**54.8**	53.4	54.2	53.2	**75.2**	74.4	**75.2**	74.6
rop_Latn	4.6	4.8	**5.4**	5.2	46.0	45.8	**46.6**	**46.6**	62.6	63.6	**63.8**	63.4
rug_Latn	3.8	4.2	**4.4**	4.0	47.0	48.2	**48.6**	47.8	61.8	**64.8**	**64.8**	63.4
run_Latn	5.4	6.6	5.8	**6.8**	**54.6**	46.0	46.0	45.8	**65.0**	61.0	61.8	61.8
rus_Cyrl	16.6	35.0	33.4	**38.0**	21.0	38.2	38.8	**41.8**	**71.8**	61.6	64.0	66.0
sag_Latn	5.6	5.2	5.2	**6.0**	49.0	47.4	47.2	**49.4**	**63.0**	62.4	62.8	61.8

Table 13: Top-10 accuracy of models on **transliterated** dataset of **SR-B** (Part III).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
sah_Cyrl	5.0	**5.4**	**5.4**	5.0	6.8	**19.4**	19.2	18.0	20.8	32.4	**33.0**	29.8
san_Deva	3.6	**7.2**	5.4	6.6	10.4	**11.0**	9.6	10.4	**20.0**	14.4	12.2	15.0
san_Latn	**3.8**	**3.8**	3.2	**3.8**	8.2	**9.2**	8.2	8.6	**16.4**	12.0	12.6	11.8
sba_Latn	**4.6**	4.2	4.4	**4.6**	8.2	10.8	**11.4**	11.0	16.2	**19.6**	18.6	19.0
seh_Latn	6.4	6.8	6.2	**7.0**	**74.6**	68.2	70.4	71.0	**82.0**	79.8	80.6	80.8
sin_Sinh	4.4	**13.6**	12.4	**13.6**	4.6	9.4	**11.6**	10.2	14.6	19.6	**21.4**	20.2
slk_Latn	**57.4**	56.8	57.0	54.8	41.0	42.4	**43.4**	42.6	**72.2**	69.2	70.0	69.0
slv_Latn	56.4	57.0	**58.8**	58.4	46.0	45.6	47.4	**49.0**	**72.6**	70.6	70.2	70.4
sme_Latn	6.2	6.4	**6.6**	**6.6**	30.6	**34.0**	33.8	32.8	**52.8**	51.6	51.2	51.4
smo_Latn	4.4	**4.8**	4.2	4.6	**37.2**	36.8	35.2	36.6	**60.8**	58.2	57.6	58.6
sna_Latn	6.8	6.8	6.8	**7.4**	**45.6**	40.6	40.6	41.2	**61.2**	59.4	60.6	60.4
snd_Arab	3.8	15.6	12.4	**18.2**	5.2	15.0	15.2	**16.8**	11.2	22.2	25.6	**27.4**
som_Latn	22.2	**24.6**	21.2	24.2	**33.0**	27.8	28.6	27.4	**52.8**	48.6	49.8	49.2
sop_Latn	4.6	5.8	5.8	**6.6**	**32.6**	28.2	29.0	29.2	**54.8**	54.0	53.6	54.0
sot_Latn	6.0	6.4	**7.0**	6.8	**52.2**	51.0	50.6	50.6	73.0	**73.4**	72.6	**73.4**
spa_Latn	77.8	78.8	**79.2**	78.4	**78.6**	78.0	78.2	78.2	**84.4**	84.0	83.8	84.2
sqi_Latn	**49.2**	48.0	47.4	47.0	55.8	56.2	56.8	**57.2**	**75.2**	69.4	71.2	71.4
srm_Latn	**4.0**	3.4	3.2	3.2	19.6	26.4	**27.4**	26.8	39.2	**46.0**	45.6	43.8
srn_Latn	**7.6**	7.4	7.2	7.0	79.2	79.2	79.0	**79.4**	**83.0**	**83.0**	**83.0**	82.8
srp_Cyrl	57.8	63.6	63.4	**64.4**	57.6	60.2	**61.8**	**61.8**	**81.2**	71.8	73.2	73.8
srp_Latn	**77.8**	70.6	72.0	70.2	**73.0**	69.4	69.4	69.6	**83.6**	79.2	80.4	80.2
ssw_Latn	4.8	**6.0**	**6.0**	5.8	**47.0**	45.2	46.4	45.4	**58.4**	57.8	57.0	56.8
sun_Latn	22.4	25.2	24.0	**25.4**	**43.0**	39.8	40.0	39.4	**63.8**	62.2	62.2	62.2
suz_Deva	2.4	3.4	**3.8**	**3.8**	4.0	8.4	**9.8**	8.8	8.4	**13.2**	**13.2**	12.8
swe_Latn	45.2	68.6	72.2	**74.2**	43.2	59.4	63.6	**65.2**	80.6	81.0	81.2	**82.0**
swh_Latn	**47.8**	45.8	46.0	45.6	**66.4**	60.0	59.4	60.8	74.6	74.0	75.4	**75.6**
sxn_Latn	**5.4**	4.6	4.6	5.0	25.0	**26.6**	**26.6**	26.4	45.0	**48.4**	48.2	**48.4**
tam_Taml	4.8	16.0	10.6	**16.4**	7.8	17.0	18.4	**20.8**	19.6	32.0	33.4	**34.8**
tat_Cyrl	7.4	7.4	**7.8**	**7.8**	13.6	**28.6**	25.4	25.2	48.2	50.4	**51.6**	49.0
tbz_Latn	4.2	4.2	4.0	**4.8**	7.2	**9.0**	7.6	8.2	**13.0**	**13.0**	12.2	12.8
tca_Latn	2.4	2.8	3.0	**3.4**	4.0	13.8	13.8	**15.6**	9.8	27.0	**30.2**	29.2
tdt_Latn	6.2	5.6	**6.4**	6.0	**63.0**	58.0	60.4	60.6	**78.2**	76.6	76.4	76.6
tel_Telu	4.2	**18.0**	13.6	14.4	7.0	11.2	12.2	**15.0**	20.2	21.4	24.4	**25.6**
teo_Latn	5.6	**6.0**	5.8	**6.0**	**24.6**	24.0	24.0	24.0	28.6	27.8	**29.4**	**29.4**
tgk_Cyrl	6.8	**7.4**	6.6	6.6	17.4	**30.4**	28.6	29.8	56.6	59.4	59.0	**59.8**
tgl_Latn	**61.2**	57.6	58.0	57.4	**78.8**	76.2	75.8	75.2	84.2	84.0	**84.6**	84.4
tha_Thai	2.6	8.8	6.4	**11.8**	3.2	11.6	**14.4**	**14.4**	6.2	19.4	**22.8**	22.2
tih_Latn	5.2	**5.4**	**5.4**	4.6	**51.6**	49.8	48.6	48.6	67.6	70.8	71.2	**71.6**
tir_Ethi	4.4	4.0	**5.2**	5.0	4.0	11.6	11.8	**12.0**	10.8	**14.8**	13.6	14.6
tlh_Latn	7.8	**9.2**	8.0	8.6	**72.4**	67.6	67.4	68.4	73.2	**74.4**	74.0	73.8
tob_Latn	2.8	3.0	**3.2**	2.8	15.8	**17.4**	16.2	16.8	30.0	32.8	**33.4**	32.2
toh_Latn	4.0	4.6	4.6	**4.8**	**47.2**	42.8	43.0	42.4	64.4	**65.0**	64.2	64.2
toi_Latn	4.2	**5.0**	4.0	4.8	**47.2**	42.4	43.6	41.2	58.0	59.0	**59.2**	57.0
toj_Latn	4.2	**5.0**	3.4	4.4	**16.6**	14.8	14.8	14.0	31.4	34.2	**34.8**	34.4
ton_Latn	3.4	3.4	3.8	**4.6**	22.6	23.4	23.6	**24.4**	46.8	**47.6**	46.8	47.2
top_Latn	**5.6**	5.2	5.0	4.6	**11.4**	10.0	10.0	9.4	**22.0**	21.0	20.4	21.6
tpi_Latn	5.8	6.2	**7.4**	6.0	58.0	58.8	**59.0**	**59.0**	**68.4**	68.2	68.0	**68.4**
tpm_Latn	4.2	5.0	**5.2**	5.0	26.6	35.4	35.0	**35.8**	38.0	41.0	41.2	**41.8**
tsn_Latn	**5.4**	**5.4**	**5.4**	5.2	**41.6**	39.8	39.2	38.8	**61.8**	60.6	60.2	60.4
tso_Latn	**5.6**	4.4	4.4	4.6	**50.4**	48.4	47.8	49.0	**65.0**	64.4	63.4	64.2
tsz_Latn	**6.2**	5.2	5.8	5.6	16.6	19.0	**19.6**	19.4	32.6	**37.2**	36.2	36.6
tuc_Latn	**3.6**	3.4	3.2	3.2	27.4	**29.6**	29.2	29.4	35.4	40.4	40.0	**41.2**
tui_Latn	3.4	4.4	4.4	**4.6**	12.0	24.4	**28.6**	27.0	31.0	42.4	46.8	**47.8**
tuk_Cyrl	10.8	14.2	15.0	**15.2**	31.6	34.6	**35.2**	**35.2**	**63.0**	59.2	59.4	58.2
tuk_Latn	9.0	16.4	16.6	**17.4**	29.8	42.2	**42.8**	42.0	67.2	**67.4**	67.2	66.8
tum_Latn	5.4	**6.0**	5.8	5.8	**61.4**	59.2	59.0	59.2	64.8	66.0	**66.4**	66.2
tur_Latn	33.0	58.0	**59.8**	59.4	36.8	45.0	48.4	**49.8**	**68.8**	66.6	66.6	67.2
twi_Latn	**4.8**	4.2	4.2	4.0	28.6	34.8	34.0	**35.4**	**43.6**	43.0	42.2	42.8
tyv_Cyrl	5.4	**5.6**	5.4	5.2	7.8	**17.8**	17.6	16.8	21.2	**33.6**	32.0	31.8
tzh_Latn	**5.8**	5.4	5.4	5.4	23.8	**25.2**	24.2	24.0	45.0	46.6	**46.8**	45.8
tzo_Latn	**4.2**	3.8	3.4	3.2	**16.0**	14.0	14.0	13.4	32.2	**34.6**	33.6	34.0
udm_Cyrl	**5.4**	5.2	**5.4**	5.2	9.4	**17.8**	14.2	13.6	20.4	24.2	23.8	**24.6**
uig_Arab	3.4	**16.8**	14.0	15.8	6.0	18.6	18.8	**19.2**	23.4	37.6	**40.8**	39.8
uig_Latn	9.8	**10.6**	10.4	9.2	**62.8**	54.0	54.4	54.4	72.6	75.6	**76.2**	75.8
ukr_Cyrl	9.2	23.6	23.2	**24.0**	8.8	19.8	22.6	**24.2**	**57.6**	46.8	50.0	50.8
urd_Arab	4.4	12.0	12.0	**18.2**	5.8	14.4	18.6	**23.0**	36.0	32.4	39.2	**44.2**
uzb_Cyrl	**33.2**	32.0	31.0	32.8	52.0	57.0	57.2	**57.6**	**78.0**	72.8	73.2	74.0
uzb_Latn	**54.8**	50.6	50.6	50.6	**67.6**	65.6	63.8	64.0	**84.0**	83.8	83.2	83.4
uzn_Cyrl	**28.6**	27.2	27.8	28.4	53.4	64.4	**66.0**	65.8	80.2	78.4	79.2	**80.4**
ven_Latn	4.4	**5.6**	**5.6**	4.6	**45.0**	42.0	42.0	43.4	47.6	47.6	**48.4**	48.2
vie_Latn	7.6	11.4	9.0	**15.2**	8.4	10.2	11.4	**12.2**	22.4	22.2	26.4	**26.6**
wal_Latn	4.2	**4.6**	**4.6**	4.0	**51.4**	43.0	43.2	42.2	**54.8**	54.2	**54.8**	53.4
war_Latn	**9.8**	8.2	9.0	9.0	**43.4**	40.6	41.0	41.0	78.0	77.2	**78.4**	77.8
wbm_Latn	**3.8**	**3.8**	3.2	3.2	46.4	47.2	47.4	**49.0**	**49.4**	48.8	48.2	48.8
wol_Latn	5.2	**6.4**	5.4	5.0	25.4	**27.2**	26.6	25.8	42.6	**42.8**	**42.8**	42.2
xav_Latn	2.6	2.6	**3.0**	2.8	3.4	4.6	**4.8**	**4.8**	6.6	**8.6**	8.2	7.8
xho_Latn	10.6	11.0	11.6	**12.4**	**41.4**	39.8	40.8	40.0	62.6	62.6	62.8	**63.2**
yan_Latn	**4.6**	4.4	4.4	4.0	**31.0**	30.0	29.8	**31.0**	**37.6**	36.2	36.4	36.2
yao_Latn	4.6	4.8	**5.4**	4.8	45.0	45.6	**47.4**	46.6	**56.6**	55.4	54.6	55.4
yap_Latn	4.0	**5.2**	4.8	5.0	24.0	**25.2**	25.0	25.0	42.0	**43.2**	42.6	42.4
yom_Latn	5.0	5.2	**5.6**	5.2	**42.0**	40.0	41.0	41.0	**56.6**	54.2	55.0	54.6
yor_Latn	**4.4**	3.8	**4.4**	4.2	23.4	24.4	24.0	**25.2**	44.8	46.8	45.8	**47.6**
yua_Latn	**4.0**	3.6	3.2	3.6	11.0	**16.2**	14.6	14.6	33.4	36.6	37.6	**38.4**
yue_Hani	3.8	**4.0**	3.2	3.6	6.6	**7.0**	5.8	5.8	15.6	14.2	14.0	**16.0**
zai_Latn	**6.6**	6.0	5.4	6.2	**32.4**	29.8	29.8	28.6	42.6	42.0	**43.2**	42.2
zho_Hani	4.2	9.4	5.4	**11.4**	6.0	11.2	12.4	**14.2**	11.8	15.8	16.6	**18.8**
zlm_Latn	**83.4**	81.8	82.2	80.2	**87.0**	85.0	85.6	85.0	**82.8**	82.0	82.2	82.2
zom_Latn	3.6	3.6	**4.0**	3.8	**50.2**	47.6	46.6	48.0	59.2	58.0	**59.6**	**59.6**
zsm_Latn	**90.2**	88.8	88.8	88.6	**83.0**	81.4	81.6	82.2	85.0	85.8	86.2	**86.4**
zul_Latn	10.6	11.8	**12.2**	11.6	**48.8**	43.8	43.4	44.8	**56.6**	55.0	55.6	56.0

Table 14: Top-10 accuracy of models on **transliterated** dataset of **SR-B** (Part IV).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
afr_Latn	**71.7**	69.9	67.4	65.7	**80.7**	77.0	77.1	74.8	**84.3**	83.6	82.0	82.3
amh_Ethi	10.7	**22.0**	20.2	21.4	13.1	**23.2**	19.0	22.0	24.4	26.8	28.0	**29.2**
ara_Arab	3.4	19.2	18.7	**23.4**	4.2	25.3	27.0	**31.0**	8.0	17.6	18.8	**20.1**
arz_Arab	6.5	13.6	12.8	**17.0**	8.4	25.2	27.9	**31.4**	11.5	15.5	18.9	**19.9**
ast_Latn	52.8	52.0	55.1	**59.1**	**79.5**	78.0	78.7	78.7	**85.0**	**85.0**	84.3	84.3
aze_Latn	27.6	48.3	47.0	**50.1**	52.0	66.6	**68.7**	67.8	68.0	**73.6**	71.9	72.9
bel_Cyrl	11.4	34.0	33.6	**36.0**	16.4	42.0	44.1	**44.9**	48.7	57.5	57.5	**58.3**
ben_Beng	4.0	15.4	12.5	**16.9**	6.1	21.4	28.1	**30.8**	25.8	26.4	30.8	**32.5**
bos_Latn	68.9	71.5	71.8	**72.9**	86.2	88.1	**88.4**	87.0	**92.4**	91.5	91.0	91.2
bre_Latn	**10.5**	9.8	9.2	9.5	**16.7**	15.7	15.6	15.5	**20.3**	19.2	18.8	18.9
bul_Cyrl	16.4	39.9	40.1	**42.6**	28.4	54.5	56.7	**58.5**	69.1	70.9	**71.6**	71.0
cat_Latn	**68.4**	66.9	68.1	67.3	**76.4**	73.1	73.8	73.8	**83.8**	82.8	83.2	82.8
cbk_Latn	38.0	37.6	38.3	**38.9**	60.0	59.0	60.7	**60.8**	**65.2**	65.0	64.7	64.5
ceb_Latn	15.2	15.7	15.7	**15.8**	41.0	40.0	**41.2**	40.8	42.3	**43.0**	42.7	42.8
ces_Latn	44.9	55.2	54.4	**55.7**	48.7	57.1	59.2	**60.0**	68.4	**70.8**	70.6	70.4
cmn_Hani	2.9	20.8	16.5	**28.9**	4.3	26.3	28.5	**35.2**	6.0	16.5	17.6	**18.9**
csb_Latn	25.7	27.3	**27.7**	27.3	39.9	41.5	**43.5**	42.3	**69.6**	66.4	66.4	67.2
cym_Latn	**47.0**	43.1	40.2	41.6	**56.3**	52.0	51.3	49.0	**54.3**	52.5	52.3	51.8
dan_Latn	81.8	**89.4**	87.1	88.9	82.9	**88.3**	**88.3**	87.2	92.2	93.4	**93.5**	93.3
deu_Latn	92.5	**95.1**	94.5	94.9	91.4	**94.5**	94.3	93.8	95.6	**96.5**	96.0	96.4
dtp_Latn	5.6	**6.1**	5.7	5.9	**21.1**	20.1	20.2	19.8	19.8	**20.2**	**20.2**	19.9
ell_Grek	5.9	22.3	22.4	**24.4**	6.9	22.1	24.3	**25.6**	19.8	25.1	26.3	**27.0**
epo_Latn	46.8	51.1	51.3	**51.8**	56.7	59.1	**60.2**	59.9	76.6	**77.1**	76.7	77.0
est_Latn	47.4	**59.3**	57.8	59.2	49.7	61.1	61.0	**61.2**	65.6	**72.2**	71.1	71.6
eus_Latn	45.9	**46.1**	45.0	45.3	**52.7**	51.2	52.0	51.5	**58.7**	58.3	57.7	58.0
fao_Latn	37.0	38.5	38.9	**40.1**	58.4	70.2	**73.3**	71.8	80.2	**82.8**	82.1	81.7
fin_Latn	53.0	**76.2**	74.8	75.9	42.8	62.5	65.1	**65.4**	61.0	**71.4**	71.0	70.9
fra_Latn	80.4	79.4	80.1	**81.0**	**81.5**	79.5	80.9	79.5	83.1	82.5	83.3	**83.7**
fry_Latn	**63.6**	61.3	63.0	62.4	**78.0**	76.3	75.7	73.4	**85.5**	**85.5**	85.0	84.4
gla_Latn	20.4	**22.0**	21.7	**22.0**	**38.8**	38.2	38.4	37.0	42.3	**44.0**	**44.0**	43.4
gle_Latn	21.5	26.6	26.6	**27.2**	36.7	**42.1**	42.0	41.5	38.8	**43.2**	43.1	42.8
glg_Latn	68.2	67.9	**68.3**	68.1	**73.6**	73.0	72.5	71.8	**83.5**	82.8	82.5	82.3
gsw_Latn	39.3	39.3	40.2	**41.0**	58.1	**67.5**	65.8	60.7	67.5	**75.2**	**75.2**	73.5
heb_Hebr	2.4	26.5	22.1	**32.4**	3.3	20.0	25.2	**28.9**	8.1	20.2	22.8	**24.5**
hin_Deva	15.3	30.0	25.1	**41.6**	24.9	45.4	55.2	**59.0**	52.9	44.3	50.4	**56.4**
hrv_Latn	69.1	**69.3**	68.7	69.2	83.5	**84.2**	84.1	83.8	**87.5**	86.3	85.6	85.5
hsb_Latn	22.4	23.8	**24.0**	**24.0**	33.3	36.0	**36.6**	36.0	**58.4**	57.8	56.9	56.5
hun_Latn	47.9	**66.5**	65.1	65.3	37.1	56.7	**57.8**	57.5	52.4	**64.2**	63.5	64.1
hye_Armn	3.6	16.0	14.8	**17.3**	10.9	24.9	27.9	**29.8**	17.8	21.2	22.6	**23.6**
ido_Latn	25.5	27.2	27.3	**28.3**	**57.5**	55.3	55.0	55.3	**76.6**	75.5	74.7	74.8
ile_Latn	35.5	36.1	**37.0**	**37.0**	**75.4**	74.4	75.3	74.5	**83.0**	82.4	82.2	81.9
ina_Latn	62.7	63.0	63.8	**64.0**	**91.4**	90.3	90.1	89.5	**93.4**	93.1	92.7	93.2
ind_Latn	**84.3**	82.9	83.0	82.6	**88.8**	88.0	87.8	87.8	**86.4**	86.1	85.9	85.8
isl_Latn	25.6	52.6	52.9	**53.8**	32.3	56.4	**57.5**	56.5	74.4	82.5	**82.6**	**82.6**
ita_Latn	78.3	**78.5**	77.7	78.2	**84.1**	84.0	83.4	83.1	**89.2**	88.6	87.9	87.9
jpn_Jpan	3.0	24.4	22.8	**29.0**	3.8	19.4	21.1	**21.7**	6.6	13.7	**15.0**	**15.0**
kab_Latn	3.9	4.1	4.0	**4.2**	11.3	11.9	**12.1**	12.0	14.3	14.3	14.2	**14.8**
kat_Geor	6.6	23.9	23.5	**25.7**	11.0	28.0	30.0	**31.1**	17.6	22.5	22.7	**24.7**
kaz_Cyrl	13.0	30.3	28.9	**30.6**	33.9	51.3	53.6	**55.8**	44.7	49.6	50.1	**51.7**
khm_Khmr	3.3	26.6	27.6	**28.3**	5.4	31.3	**34.2**	**34.2**	9.4	28.4	29.1	**29.5**
kor_Hang	2.7	32.9	33.7	**37.7**	3.7	31.0	32.9	**34.0**	8.3	26.7	26.6	**28.6**
kur_Latn	17.3	**22.9**	22.4	**22.9**	38.8	**42.4**	41.7	41.7	47.6	48.8	49.5	**49.8**
lat_Latn	**34.5**	33.8	33.8	33.9	**43.6**	42.6	42.6	42.5	**47.3**	46.4	45.2	45.7
lfn_Latn	37.6	38.4	**39.3**	39.0	**76.5**	75.5	74.6	74.4	**81.5**	81.0	80.9	81.1
lit_Latn	45.1	53.4	53.9	**54.6**	38.0	47.3	48.5	**49.1**	59.5	**64.1**	63.7	63.8
lvs_Latn	31.3	51.3	**52.1**	51.3	37.4	55.6	**55.9**	55.8	61.2	**68.5**	68.1	67.3
mal_Mlym	2.6	32.9	30.3	**39.4**	4.4	30.7	39.3	**42.8**	18.9	40.3	46.3	**46.7**
mar_Deva	4.0	31.1	28.1	**36.9**	7.1	33.9	38.7	**40.0**	21.8	32.0	34.2	**35.0**
mhr_Cyrl	4.5	5.2	5.1	**5.6**	8.1	13.8	14.7	**15.8**	**15.8**	15.1	15.7	15.2
mkd_Cyrl	19.0	30.9	30.4	**32.5**	46.6	57.9	59.4	**60.4**	**72.4**	65.2	65.1	64.9
mon_Cyrl	9.3	**35.5**	30.0	33.2	30.5	52.0	52.5	**53.0**	40.2	**49.3**	48.2	48.6
nds_Latn	28.9	29.5	30.4	**31.2**	67.1	73.5	**73.8**	73.6	79.8	**82.1**	81.9	**82.1**
nld_Latn	**90.4**	89.2	88.1	88.9	**91.9**	90.2	89.9	89.9	**91.6**	91.1	90.8	91.2
nno_Latn	56.8	60.9	60.7	**63.4**	78.0	81.6	81.2	**82.0**	88.6	**88.8**	87.8	88.2
nob_Latn	82.0	87.5	87.0	**89.6**	88.8	92.4	**92.7**	92.5	94.5	96.0	95.8	**96.2**
oci_Latn	**26.4**	24.9	25.7	25.0	**45.1**	41.9	42.4	41.7	**61.7**	58.6	58.5	59.5
pam_Latn	7.6	7.7	**8.2**	**8.2**	**23.0**	21.6	22.9	**23.0**	**31.4**	31.0	30.3	30.2
pes_Arab	3.5	26.5	25.6	**42.3**	5.8	36.4	44.0	**52.7**	21.0	44.1	47.8	**51.9**
pms_Latn	**22.5**	20.4	22.3	21.0	**47.2**	37.3	41.9	39.0	**68.6**	63.4	62.9	63.0
pol_Latn	65.3	**71.5**	70.3	71.4	64.5	68.3	**69.2**	68.8	75.8	**77.9**	77.3	77.0
por_Latn	82.9	86.9	83.9	**87.3**	78.0	81.8	83.7	**84.3**	91.2	**92.0**	91.7	**92.0**
ron_Latn	75.9	79.9	78.5	**80.2**	74.6	76.2	77.1	**77.3**	85.2	**85.5**	85.3	84.7
rus_Cyrl	13.3	43.4	43.5	**50.7**	22.2	56.9	59.2	**62.3**	64.9	69.4	**71.0**	70.3
slk_Latn	51.3	58.9	58.3	**59.4**	54.7	59.6	60.8	**61.0**	73.4	**75.4**	74.6	74.4
slv_Latn	62.7	**64.6**	64.5	63.9	66.3	70.4	**70.8**	70.4	**78.5**	78.4	77.6	77.6
spa_Latn	82.1	**82.3**	82.1	81.8	**84.8**	83.6	83.9	83.0	**86.9**	86.7	86.4	86.5
sqi_Latn	49.1	54.6	55.0	**55.8**	68.2	73.4	74.0	**74.1**	81.2	**83.5**	**83.5**	83.1
srp_Latn	59.7	64.5	**64.9**	64.4	80.5	**83.5**	83.0	82.8	**89.0**	86.5	86.0	85.2
swe_Latn	57.8	83.8	83.7	**87.4**	61.2	78.7	81.9	**82.7**	85.8	89.2	88.9	**89.5**
swh_Latn	30.3	30.8	**32.3**	30.5	44.1	44.9	45.6	**45.9**	**44.6**	44.1	43.8	43.6
tam_Taml	10.1	22.1	21.5	**25.4**	9.1	26.1	30.6	**32.2**	16.3	30.9	33.9	**35.8**
tat_Cyrl	8.1	10.0	**10.9**	10.7	16.3	**36.7**	36.6	35.3	32.5	35.5	**37.0**	36.5
tel_Telu	12.4	30.3	24.8	**32.9**	12.4	**35.0**	33.8	32.9	27.4	37.2	38.0	**42.7**
tgl_Latn	**47.6**	46.5	47.2	47.4	**77.3**	76.3	75.4	75.3	**76.6**	75.9	76.0	75.6
tha_Thai	3.3	18.6	15.9	**22.6**	4.6	25.9	29.9	**30.5**	6.9	14.2	15.9	**16.6**
tuk_Latn	19.2	24.6	25.1	**27.1**	36.9	**52.7**	49.3	50.7	50.2	52.7	**56.2**	55.2
tur_Latn	42.5	65.3	64.4	**67.5**	46.8	65.5	66.8	**67.1**	59.9	**71.2**	70.3	70.9
uig_Arab	3.5	**16.4**	14.1	15.7	8.0	26.4	28.1	**29.6**	15.3	24.6	**27.8**	27.4
ukr_Cyrl	12.0	32.8	33.5	**37.8**	20.8	41.1	44.8	**48.4**	54.4	56.4	56.1	**57.6**
urd_Arab	3.0	17.6	16.3	**25.9**	6.9	25.7	33.1	**36.1**	24.8	29.2	33.1	**35.2**
uzb_Cyrl	29.9	31.1	29.7	**31.8**	**56.8**	56.1	55.8	55.1	**61.0**	60.5	59.6	59.6
vie_Latn	6.4	11.4	10.7	**18.0**	10.9	13.0	16.0	**18.5**	18.5	22.1	25.0	**26.6**
war_Latn	**9.5**	8.7	9.1	8.9	31.0	30.7	31.7	**32.1**	44.4	44.8	45.4	**45.6**
wuu_Hani	4.0	11.5	8.9	**16.4**	4.9	15.0	16.7	**18.8**	6.0	10.6	10.8	**14.2**
xho_Latn	28.9	28.9	28.9	**29.6**	56.3	54.9	56.3	**60.6**	60.6	60.6	**61.3**	**61.3**
yid_Hebr	3.5	23.7	20.6	**28.2**	4.0	29.2	**38.4**	38.3	11.9	39.6	41.0	**44.7**
yue_Hani	4.4	8.5	7.4	**11.4**	4.3	11.4	12.2	**12.9**	6.6	8.0	8.5	**11.2**
zsm_Latn	**81.4**	80.8	81.1	80.9	**91.8**	91.4	90.4	90.0	**88.8**	88.7	88.0	87.9

Table 15: Top-10 accuracy of models on **transliterated** dataset of **SR-T**.

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
ace_Latn	13.1	16.0	12.4	**16.3**	62.8	62.1	64.1	**65.4**	64.6	67.1	**69.1**	68.1
ach_Latn	8.2	**11.6**	9.3	9.0	42.0	42.8	**44.6**	38.4	50.5	49.2	51.4	**56.3**
acr_Latn	8.9	**13.2**	11.0	12.5	56.8	57.9	**59.9**	54.7	58.8	58.7	**60.0**	59.6
afr_Latn	**67.0**	63.1	65.3	65.7	**62.5**	59.2	59.6	59.9	63.6	62.3	61.5	**64.9**
agw_Latn	11.1	**15.9**	14.6	14.1	60.2	59.7	**61.9**	59.2	54.8	58.0	59.5	**60.1**
ahk_Latn	6.0	8.1	7.4	**8.6**	6.8	**7.5**	5.7	6.7	**7.2**	5.9	**7.2**	6.3
aka_Latn	9.0	**14.3**	11.3	11.6	30.9	**39.5**	37.5	37.1	39.1	41.9	**43.7**	43.3
aln_Latn	44.9	**48.4**	46.4	47.2	51.0	56.5	55.1	**57.3**	53.3	57.9	**58.4**	58.0
als_Latn	46.3	**48.8**	47.6	47.3	49.4	55.2	56.9	**57.9**	52.1	58.3	58.7	**59.1**
alt_Cyrl	7.8	**15.1**	13.6	10.5	17.5	39.1	**40.8**	34.8	25.3	**35.0**	32.5	34.5
alz_Latn	7.5	11.3	**11.5**	10.5	34.1	36.4	**37.8**	34.7	43.7	41.2	39.6	**45.9**
amh_Ethi	7.0	**8.2**	7.7	**8.2**	4.9	5.8	**6.1**	5.6	8.0	6.5	9.1	**11.8**
aoj_Latn	8.9	**13.4**	11.6	11.2	37.1	**52.9**	51.0	49.2	39.4	**49.2**	47.8	46.9
arn_Latn	5.7	**9.5**	8.0	8.2	30.7	44.5	**47.1**	41.3	32.4	43.9	**46.2**	45.6
ary_Arab	11.3	10.6	9.7	**11.7**	7.8	15.9	17.8	**19.2**	15.4	15.4	17.5	**20.8**
arz_Arab	8.9	**14.7**	9.5	10.4	11.0	17.3	18.6	**23.3**	11.3	17.7	18.8	**21.7**
asm_Beng	9.5	16.6	15.1	**19.4**	9.4	31.0	**33.2**	**33.2**	18.4	25.9	24.3	**34.8**
ayr_Latn	9.1	**13.0**	10.1	11.1	59.4	60.2	**62.0**	59.4	60.0	60.0	60.8	**62.3**
azb_Arab	8.3	**15.8**	12.2	14.7	5.8	46.5	**47.7**	41.1	17.5	**46.3**	45.3	45.7
aze_Latn	47.7	59.5	**65.2**	62.6	53.3	**63.9**	63.3	60.5	62.1	62.0	60.4	**63.8**
bak_Cyrl	10.3	**16.3**	12.4	15.9	24.1	45.3	**46.6**	45.2	36.2	50.3	49.8	**52.2**
bam_Latn	8.0	10.9	11.3	**12.1**	38.0	40.5	**42.3**	38.6	**39.8**	37.6	37.2	39.1
ban_Latn	17.0	17.2	20.1	**20.8**	48.6	48.6	**50.1**	47.0	52.0	51.4	53.6	**55.5**
bar_Latn	25.9	33.2	**34.3**	34.2	45.0	46.9	**49.7**	48.1	50.9	52.7	49.8	**54.3**
bba_Latn	6.1	**9.8**	7.2	6.4	15.4	28.2	**30.6**	27.4	22.2	36.7	34.6	**38.8**
bci_Latn	6.6	6.6	**8.4**	7.1	30.3	33.8	**37.5**	31.0	39.2	39.3	**40.4**	38.3
bcl_Latn	29.5	29.1	**33.4**	30.4	58.4	**61.4**	60.4	56.0	58.9	**61.1**	57.7	58.2
bel_Cyrl	13.9	40.1	39.5	**40.9**	19.0	41.8	**42.5**	40.4	29.5	40.6	**40.7**	40.4
bem_Latn	9.9	13.2	**13.5**	12.0	**52.3**	51.0	49.0	48.6	53.1	56.1	56.1	**56.9**
ben_Beng	7.5	26.2	25.8	**27.2**	10.2	35.3	**40.3**	38.4	33.8	39.9	**44.4**	**44.4**
bhw_Latn	7.9	11.6	11.8	**12.7**	43.1	45.5	**46.6**	43.8	47.9	47.5	**48.7**	48.5
bim_Latn	5.7	**9.3**	8.8	7.8	48.8	**52.3**	48.2	48.7	49.4	54.8	52.6	**56.0**
bis_Latn	13.0	14.5	**15.7**	10.6	70.6	69.5	**70.9**	69.7	72.0	73.0	72.8	**73.1**
bqc_Latn	11.9	12.5	**13.9**	10.9	10.0	14.5	11.9	**15.4**	13.3	15.4	15.6	**17.4**
bre_Latn	24.2	21.8	**25.5**	23.1	36.2	35.8	**38.3**	34.1	**42.2**	38.2	39.8	40.6
btx_Latn	21.4	22.9	**25.7**	22.4	57.5	**59.4**	59.1	55.7	**62.5**	61.7	60.6	61.2
bul_Cyrl	29.0	52.4	**54.7**	54.0	38.5	54.5	**58.8**	54.3	50.8	54.0	**56.8**	56.4
bum_Latn	8.5	**11.6**	11.0	10.9	24.6	24.9	**31.3**	29.0	28.3	34.6	31.7	**35.8**
bzj_Latn	12.0	15.1	12.7	**16.9**	68.6	68.8	**69.7**	69.5	70.4	**71.9**	71.2	71.1
cab_Latn	8.3	**10.5**	9.7	8.0	29.8	30.7	**34.5**	31.2	34.1	30.0	**35.6**	32.1
cac_Latn	7.7	**11.4**	9.3	9.9	**55.5**	52.2	54.2	48.6	55.5	56.4	56.2	**56.5**
cak_Latn	6.6	**11.1**	9.2	9.3	**62.2**	**62.2**	61.2	60.1	58.0	60.6	**63.2**	61.2
caq_Latn	7.3	11.7	**12.1**	9.8	10.5	31.3	**33.6**	30.1	16.1	33.2	32.2	**35.4**
cat_Latn	**65.2**	64.4	64.7	64.5	59.2	57.9	**59.6**	56.2	**63.4**	62.2	62.1	60.7
cbk_Latn	49.5	46.4	**50.5**	48.8	64.3	66.3	**66.7**	65.0	68.5	**71.6**	69.6	69.1
cce_Latn	9.0	10.5	**10.6**	8.7	57.4	**59.0**	**59.0**	53.4	57.0	56.1	58.2	**62.2**
ceb_Latn	27.3	31.1	31.2	**32.4**	**57.7**	54.3	56.1	51.8	**61.1**	60.1	61.0	59.6
ces_Latn	53.1	55.4	54.9	**55.8**	45.4	**54.9**	54.4	51.5	54.0	56.0	55.5	**56.8**
cfm_Latn	6.5	**10.1**	9.8	8.7	66.0	64.4	**68.3**	62.6	65.1	**65.3**	63.8	64.7
che_Cyrl	7.5	**11.4**	9.2	9.7	6.1	**7.2**	6.9	6.0	5.5	5.6	6.1	**7.4**
chv_Cyrl	8.0	**11.5**	11.4	9.4	7.0	17.2	**21.1**	17.1	7.3	14.9	14.9	**16.4**
cmn_Hani	7.5	19.3	18.7	**22.7**	4.9	26.3	**31.1**	29.5	5.5	23.4	27.8	**33.4**
cnh_Latn	7.5	**9.6**	8.7	8.6	64.3	64.6	**66.6**	61.7	67.6	**68.7**	66.0	68.0
crh_Cyrl	25.8	31.3	**34.5**	29.7	40.9	55.6	**57.8**	55.7	49.1	**61.5**	58.2	59.7
crs_Latn	13.1	**16.6**	15.2	13.3	**73.3**	68.6	70.4	68.4	69.3	**69.7**	65.6	67.6
csy_Latn	6.0	**11.4**	10.2	8.1	63.0	65.8	**66.5**	65.9	61.5	**64.6**	62.0	62.7
ctd_Latn	5.7	**9.0**	8.0	7.2	64.1	62.6	62.9	**64.7**	60.6	**64.7**	60.4	61.3
ctu_Latn	10.1	**14.2**	11.1	11.8	37.6	**49.2**	45.7	43.1	46.2	**55.4**	51.0	51.7
cuk_Latn	10.9	13.3	13.0	**13.4**	43.3	44.9	**46.6**	41.9	48.0	48.0	49.3	**50.7**
cym_Latn	49.9	47.3	**50.2**	47.7	**50.5**	48.5	45.3	47.4	**53.1**	51.7	48.5	51.2
dan_Latn	56.2	58.7	**60.8**	59.3	48.8	59.9	**60.6**	59.1	53.9	57.1	56.6	**57.5**
deu_Latn	50.7	53.0	**54.0**	53.8	44.6	**50.2**	48.3	48.4	49.0	49.0	**50.5**	49.0
djk_Latn	7.9	**13.1**	12.3	11.7	61.2	**61.4**	**61.4**	60.3	55.2	55.8	57.5	**57.9**
dln_Latn	8.2	**12.6**	10.0	12.4	53.6	53.8	**55.4**	52.9	57.8	57.0	56.4	**59.6**
dtp_Latn	6.5	9.3	**9.9**	9.6	53.6	50.5	**55.8**	47.3	60.1	**60.6**	58.0	58.3
dyu_Latn	7.2	11.9	**12.4**	10.0	36.7	**43.5**	42.1	43.0	42.3	43.7	43.4	**47.4**
dzo_Tibt	5.5	5.2	**5.8**	4.9	5.2	37.8	**43.6**	37.2	4.9	38.7	**43.5**	42.5
efi_Latn	9.4	**11.7**	10.3	10.3	30.6	43.2	46.5	**46.6**	35.8	49.0	51.3	**54.7**
ell_Grek	11.8	19.5	18.3	**20.6**	12.9	28.6	**32.3**	28.0	18.3	26.0	26.5	**30.6**
eng_Latn	**77.3**	76.5	76.4	76.2	74.9	74.1	**77.1**	76.2	76.0	74.7	76.9	**77.8**
enm_Latn	57.4	55.4	**58.1**	57.6	**70.6**	66.5	68.6	67.4	**71.9**	70.1	69.6	70.9
epo_Latn	60.0	60.8	60.3	**62.8**	60.1	58.6	**60.2**	59.9	**62.4**	62.1	59.4	62.0
est_Latn	54.7	**68.4**	66.9	68.1	49.8	**57.6**	**57.6**	56.8	62.0	63.0	63.2	**65.4**
eus_Latn	23.3	24.6	**25.0**	23.0	21.9	24.2	**27.0**	25.3	28.1	26.5	27.6	**31.0**
ewe_Latn	9.9	11.8	11.7	**11.9**	23.5	35.5	36.6	**36.8**	35.9	**41.4**	38.6	40.2
fao_Latn	19.6	32.6	**35.5**	31.4	41.6	53.6	58.1	**59.5**	49.7	55.8	54.4	**58.8**
fas_Arab	8.0	45.9	50.4	**54.7**	8.5	52.1	55.4	**59.7**	25.7	52.1	58.5	**62.6**
fij_Latn	7.8	**12.1**	9.3	7.7	53.6	57.8	**58.1**	53.6	55.6	57.0	**57.4**	53.5
fil_Latn	**53.4**	49.9	50.8	52.9	61.3	**62.4**	61.0	59.4	61.6	**65.8**	61.6	65.7
fin_Latn	49.2	57.7	55.9	**58.1**	37.3	52.5	**54.0**	50.6	47.3	59.5	**59.9**	59.8
fon_Latn	6.3	7.1	**7.8**	6.9	12.9	18.7	**24.2**	21.0	24.9	23.9	**29.3**	29.0
fra_Latn	**68.9**	68.5	67.7	67.4	**64.6**	63.1	61.6	**64.6**	64.8	**68.6**	64.6	67.6
fry_Latn	32.1	29.1	**33.0**	32.6	38.0	38.5	**40.9**	39.5	48.4	48.4	46.2	**49.9**
gaa_Latn	7.4	**11.9**	8.9	9.5	6.2	21.4	**25.8**	22.7	15.2	29.2	**29.4**	29.0
gil_Latn	6.4	9.0	**10.7**	9.5	47.6	**52.0**	51.4	50.6	**53.9**	53.7	52.8	53.6
giz_Latn	7.0	10.0	**10.5**	9.1	27.0	**43.9**	41.6	39.2	27.6	43.9	45.5	**48.0**
gkn_Latn	8.9	12.8	**13.1**	10.1	5.9	10.2	10.9	**12.7**	17.0	17.3	19.6	**23.8**
gkp_Latn	6.9	10.4	**12.3**	10.2	5.6	9.8	**11.1**	10.3	7.5	10.7	**14.1**	13.7
gla_Latn	30.7	32.5	**35.5**	34.1	**46.4**	40.0	42.6	39.1	**48.1**	40.5	42.1	43.5

Table 16: F1 scores of models on **transliterated** dataset of **Taxi1500** (Part I).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
gle_Latn	24.2	28.8	**34.6**	28.8	32.6	35.5	**38.3**	34.4	40.1	**40.6**	38.7	39.7
glv_Latn	9.0	**11.7**	10.0	10.0	**44.7**	40.3	43.8	40.3	44.2	43.7	42.8	**44.7**
gom_Latn	7.4	10.2	**11.5**	9.7	**39.1**	34.1	38.5	37.6	37.0	**38.7**	36.8	34.9
gor_Latn	15.0	14.4	**15.3**	13.3	49.6	48.8	**51.7**	49.3	**59.5**	54.8	56.5	59.2
guc_Latn	5.6	**9.7**	9.2	7.4	28.5	39.1	**39.2**	36.0	36.1	41.3	42.5	**44.1**
gug_Latn	10.2	**12.7**	12.5	10.5	34.0	46.6	**46.8**	42.4	33.3	40.8	**42.9**	41.9
guj_Gujr	11.8	30.8	26.1	**41.1**	16.3	38.6	42.0	**42.8**	41.7	41.8	43.7	**50.0**
gur_Latn	10.6	12.5	**12.9**	12.8	21.1	**44.5**	44.0	40.8	24.8	39.8	**43.6**	41.8
guw_Latn	6.6	10.1	**11.3**	8.8	17.3	**35.8**	33.4	31.9	30.0	**43.2**	38.0	39.2
gya_Latn	9.4	11.0	11.3	**12.2**	7.0	7.9	10.7	**11.4**	14.7	15.2	16.8	**17.6**
gym_Latn	5.8	**9.8**	6.8	7.4	22.7	42.8	**48.0**	40.8	31.6	48.8	**49.5**	48.2
hat_Latn	10.0	12.4	13.1	**13.4**	63.2	**66.2**	63.7	60.5	**68.8**	65.3	66.9	66.8
hau_Latn	40.2	39.0	**43.0**	40.6	52.0	53.9	**55.1**	52.6	**61.1**	58.3	60.6	60.0
haw_Latn	6.1	6.8	**7.4**	7.0	**49.5**	44.4	46.9	41.1	42.7	**45.7**	43.6	44.1
heb_Hebr	8.0	11.0	11.2	**11.7**	5.5	9.4	9.3	**10.1**	11.5	13.8	14.0	**16.5**
hif_Latn	18.4	23.9	**24.9**	22.5	46.1	46.1	**51.5**	46.7	50.2	**52.4**	48.8	47.4
hil_Latn	29.3	28.8	**31.4**	30.8	**70.9**	70.6	70.8	66.2	**75.9**	74.9	70.3	73.8
hin_Deva	21.6	40.0	42.6	**50.2**	35.8	55.2	57.2	**59.4**	43.7	52.2	55.7	**56.4**
hmo_Latn	11.4	**14.1**	12.9	11.7	**62.6**	58.8	62.1	59.7	60.4	**65.6**	63.5	64.4
hne_Deva	8.1	**14.4**	13.6	14.1	8.4	**41.6**	39.0	36.9	12.5	**39.9**	35.3	38.4
hnj_Latn	5.6	8.4	**10.9**	8.1	**65.2**	63.0	64.5	65.0	68.2	66.2	65.8	**69.8**
hra_Latn	7.0	10.6	11.2	**12.2**	52.4	52.5	53.6	**54.5**	56.3	56.6	56.7	**57.0**
hrv_Latn	66.6	66.8	**67.8**	66.4	**64.8**	63.2	63.5	62.7	**67.8**	62.3	65.5	64.9
hui_Latn	7.8	**11.5**	9.7	11.2	52.0	**54.1**	51.7	50.3	55.6	57.3	58.7	**60.6**
hun_Latn	61.7	64.3	65.5	**66.1**	45.2	47.6	50.2	**51.4**	49.8	55.0	57.6	**60.2**
hus_Latn	10.1	9.3	**10.2**	9.9	**43.7**	42.5	40.0	39.3	**47.3**	41.2	41.9	43.3
hye_Armn	9.2	29.3	35.1	**37.5**	23.6	49.4	**50.3**	49.8	25.9	43.7	46.3	**47.4**
iba_Latn	30.8	34.7	**39.5**	36.5	64.4	63.5	**66.9**	**66.9**	66.0	66.1	65.6	**66.7**
ibo_Latn	7.3	**10.6**	9.7	9.9	44.5	49.0	**50.3**	47.5	49.1	**54.4**	53.4	53.2
ifa_Latn	7.7	10.0	**11.2**	10.5	55.7	54.6	**56.9**	54.5	54.4	**55.6**	52.1	55.3
ifb_Latn	8.5	11.8	11.6	**12.9**	52.0	57.9	**58.6**	52.4	49.4	**54.3**	52.3	49.3
ikk_Latn	6.1	8.0	8.1	**8.2**	8.3	28.8	29.8	**31.3**	19.8	**34.9**	34.4	34.8
ilo_Latn	15.5	**18.4**	17.8	17.8	58.0	58.3	**58.6**	57.4	63.6	63.7	61.8	**65.1**
ind_Latn	**75.2**	69.0	70.1	68.7	**76.2**	75.0	75.5	74.6	74.9	**76.9**	75.0	74.9
isl_Latn	27.6	45.8	**46.7**	45.1	29.7	44.5	**47.1**	44.2	47.7	49.8	51.2	**52.9**
ita_Latn	**68.5**	67.0	68.3	67.5	62.7	63.9	**66.1**	65.2	69.3	67.1	**69.8**	65.4
ium_Latn	5.4	**6.7**	6.6	6.2	**64.8**	60.9	63.7	60.0	63.4	61.0	62.4	**65.8**
ixl_Latn	8.3	**11.0**	10.3	8.3	**41.1**	37.9	34.8	34.0	**43.5**	38.5	39.6	36.1
izz_Latn	8.2	**10.2**	**10.2**	8.9	28.7	**31.7**	31.5	31.4	33.4	36.2	36.4	**40.4**
jam_Latn	11.9	16.7	**19.9**	17.3	**66.1**	64.4	63.3	66.0	67.6	67.0	**67.7**	66.6
jav_Latn	44.0	44.7	44.5	**46.9**	50.3	**53.4**	51.5	47.7	54.9	57.8	54.1	**59.4**
jpn_Jpan	6.6	28.5	26.4	**28.7**	4.9	27.7	27.8	**28.6**	9.3	26.8	29.2	**33.6**
kaa_Cyrl	20.8	21.2	**21.9**	21.3	61.3	62.8	**64.6**	59.4	57.9	59.3	57.4	**62.1**
kab_Latn	9.0	**12.1**	11.7	10.3	24.8	23.2	**25.0**	22.8	**25.6**	22.7	24.1	25.5
kac_Latn	6.0	8.2	**9.9**	9.4	**55.8**	52.6	55.6	53.6	54.9	**59.4**	56.0	58.7
kal_Latn	8.3	**10.7**	9.5	9.0	39.0	**40.5**	37.6	37.5	36.0	37.7	36.2	**37.8**
kan_Knda	10.1	40.2	38.3	**41.6**	17.4	43.0	**45.4**	**45.4**	32.5	48.7	52.9	**53.7**
kat_Geor	7.7	29.4	29.5	**31.3**	6.8	29.4	31.6	**31.9**	10.7	30.4	29.0	**31.7**
kaz_Cyrl	7.4	32.3	**35.8**	35.1	27.9	45.1	**50.9**	49.4	39.1	49.4	**50.7**	48.2
kbp_Latn	8.8	**12.7**	12.0	11.8	7.0	15.9	**16.2**	15.4	8.5	17.4	**19.0**	17.3
kek_Latn	6.7	8.6	9.9	**10.3**	41.8	43.2	**44.4**	42.6	44.7	**50.2**	49.1	48.8
khm_Khmr	6.4	**45.1**	40.7	37.8	4.9	**47.3**	45.8	43.6	6.8	42.9	**53.6**	50.8
kia_Latn	7.8	10.3	**11.0**	8.7	43.6	**50.2**	48.6	45.1	41.7	46.6	47.5	**47.7**
kik_Latn	8.8	12.5	**13.7**	12.2	22.4	**36.9**	36.6	32.1	28.8	36.1	34.7	**38.3**
kin_Latn	12.2	**14.6**	12.1	12.0	58.4	**58.7**	57.3	58.3	59.5	55.9	**60.0**	58.8
kir_Cyrl	8.8	43.5	**48.3**	46.3	36.6	54.8	**56.6**	53.1	44.2	57.6	54.6	**58.5**
kjb_Latn	6.0	**9.0**	8.2	7.3	58.1	57.2	**59.5**	52.4	56.2	59.0	**59.4**	58.6
kjh_Cyrl	10.4	**13.4**	12.8	12.1	9.2	33.0	**33.7**	30.5	13.6	**35.1**	31.1	30.0
kmm_Latn	5.6	**9.7**	9.0	7.3	54.8	53.6	**54.9**	52.6	59.5	**61.2**	58.9	59.3
kmr_Cyrl	10.0	**14.4**	12.7	13.8	7.4	22.9	**24.8**	23.9	19.1	29.4	**30.0**	29.1
knv_Latn	5.5	**9.0**	8.8	8.1	41.2	50.2	**54.5**	50.3	46.3	53.1	54.1	**55.7**
kor_Hang	6.0	42.6	**48.1**	46.2	4.9	46.2	48.6	**49.5**	5.4	45.3	48.4	**49.8**
kpg_Latn	8.0	10.1	**11.1**	10.7	**68.6**	68.4	67.6	67.5	64.3	64.8	64.4	**66.3**
krc_Cyrl	17.0	**24.2**	21.7	21.2	31.3	43.9	**48.3**	46.5	36.1	47.3	**47.4**	43.4
kri_Latn	14.1	16.8	15.1	**18.2**	49.7	50.0	**52.6**	48.3	55.0	56.2	**58.1**	56.9
ksd_Latn	7.2	10.0	**10.2**	8.5	58.7	58.1	**59.8**	57.7	**62.5**	60.6	59.1	62.0
kss_Latn	6.5	**7.8**	7.7	6.5	8.9	23.4	28.0	**28.8**	13.4	26.7	33.6	**34.8**
ksw_Mymr	6.4	**8.2**	7.9	5.4	4.8	**30.8**	**30.8**	23.6	7.7	30.4	**33.9**	29.1
kua_Latn	9.8	**13.4**	11.9	11.3	**50.4**	48.3	49.8	47.3	51.4	**55.5**	54.6	54.3
lam_Latn	11.1	10.7	**11.4**	10.9	42.7	40.4	**46.6**	41.1	45.1	46.5	49.1	**49.3**
lao_Laoo	6.7	29.4	26.6	**33.7**	4.9	22.0	29.4	**33.1**	6.0	25.8	36.0	**42.2**
lat_Latn	65.1	65.8	**68.4**	65.3	55.4	56.7	**59.1**	55.3	54.5	**59.6**	57.5	55.1
lav_Latn	43.2	49.2	49.2	**52.1**	35.7	45.9	**47.0**	41.1	49.1	49.3	51.1	**51.6**
ldi_Latn	7.6	**11.5**	10.8	11.3	30.4	31.1	33.0	**33.9**	**40.2**	38.6	39.7	39.1
leh_Latn	8.1	**13.4**	12.4	10.7	52.6	50.7	**53.9**	49.8	58.5	58.2	58.0	**59.5**
lhu_Latn	5.9	9.4	9.6	**10.3**	12.8	12.5	**13.4**	12.2	23.5	21.1	20.9	**24.8**
lin_Latn	8.2	**11.6**	10.4	8.9	53.8	53.2	**56.3**	53.7	55.3	58.2	55.3	**59.5**
lit_Latn	50.3	57.1	**60.8**	58.6	38.7	48.8	**50.7**	45.7	46.0	**58.5**	53.0	54.8
loz_Latn	6.7	**10.9**	10.7	10.5	55.6	53.8	**59.9**	57.0	56.6	58.8	**60.2**	**60.2**
ltz_Latn	21.8	25.2	**25.6**	22.5	48.9	54.2	**54.5**	50.2	59.0	59.6	57.8	**62.2**
lug_Latn	10.0	**12.1**	10.7	10.9	55.2	**56.4**	52.5	52.1	**59.6**	57.2	57.0	57.5
luo_Latn	7.4	**11.6**	9.7	10.3	42.2	45.0	**45.8**	44.7	**51.0**	50.1	48.3	50.6
lus_Latn	6.9	10.2	9.4	**10.3**	59.7	**61.8**	60.9	58.2	61.6	**66.6**	63.4	64.7
lzh_Hani	7.1	13.0	11.1	**14.1**	5.0	10.3	12.0	**16.3**	8.7	11.1	10.7	**12.3**
mad_Latn	20.4	22.6	**22.9**	22.7	**62.1**	62.0	61.9	59.9	**67.9**	65.2	63.4	67.6
mah_Latn	7.3	**8.7**	8.6	7.4	28.7	**36.8**	34.6	34.8	37.9	**45.1**	41.3	43.2
mai_Deva	7.1	**16.0**	13.9	**16.0**	9.2	36.5	**38.3**	32.8	15.6	**38.9**	33.8	37.6
mal_Mlym	8.6	11.4	10.1	**11.9**	5.5	7.9	8.1	**10.4**	**11.5**	7.4	7.9	10.4
mam_Latn	6.4	**11.1**	8.5	9.1	40.6	**42.0**	41.6	33.8	**44.0**	43.1	43.7	41.2

Table 17: F1 scores of models on **transliterated** dataset of **Taxi1500** (Part II).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
mar_Deva	8.7	36.3	38.3	**40.2**	8.5	42.3	**48.0**	45.1	30.5	39.1	**51.9**	48.3
mau_Latn	5.1	5.5	**5.9**	5.0	4.9	5.6	5.3	**5.8**	4.9	5.3	5.6	**6.8**
mbb_Latn	9.9	11.4	9.8	**12.7**	54.9	57.9	**59.5**	56.9	54.8	57.6	**58.5**	58.1
mck_Latn	11.8	12.7	**14.2**	9.9	**54.0**	51.2	53.2	50.5	54.7	56.4	54.2	**57.5**
mcn_Latn	8.3	**12.1**	10.5	9.2	39.7	**41.9**	41.2	39.9	31.2	38.4	36.8	**39.9**
mco_Latn	7.6	**11.0**	9.0	8.0	15.5	18.5	**20.6**	18.9	24.0	22.2	24.6	**25.4**
mdy_Ethi	7.3	**10.5**	8.5	9.3	5.1	28.8	**30.2**	**30.2**	10.2	33.4	33.9	**34.7**
meu_Latn	11.7	**14.7**	14.6	10.4	53.0	53.7	**55.0**	50.7	50.6	**53.0**	50.4	52.5
mfe_Latn	14.6	14.4	**15.6**	14.5	**71.1**	69.5	69.2	67.4	72.7	72.3	**73.8**	71.7
mgh_Latn	6.4	9.6	**9.7**	8.8	33.7	**37.3**	35.9	33.3	42.0	42.6	42.4	**44.7**
mgr_Latn	10.7	**14.9**	14.6	11.4	51.7	**53.1**	53.0	50.6	59.4	58.4	57.6	**61.9**
mhr_Cyrl	7.9	10.6	**11.2**	9.3	8.5	26.4	25.2	**27.5**	12.8	20.4	21.4	**23.3**
min_Latn	21.7	24.9	26.0	**28.3**	51.4	49.4	**52.7**	46.1	56.1	59.9	58.8	**60.1**
miq_Latn	5.0	**8.2**	7.5	7.7	60.9	58.8	**63.1**	58.4	55.7	54.9	**57.3**	57.0
mkd_Cyrl	44.9	56.6	**62.8**	58.8	59.2	61.4	**63.4**	63.0	**73.0**	69.3	68.7	67.5
mlg_Latn	33.2	**36.3**	35.4	35.8	51.5	55.3	**56.2**	52.9	56.0	**56.3**	55.2	**56.3**
mlt_Latn	10.0	9.3	**12.0**	10.6	47.7	49.6	**50.1**	47.7	**55.1**	54.5	54.9	52.8
mos_Latn	8.3	**11.6**	10.5	10.6	10.2	18.3	20.0	**22.7**	18.5	30.8	28.8	**32.5**
mps_Latn	8.7	8.8	**9.8**	8.3	64.6	64.7	**66.3**	63.8	59.4	61.8	**63.4**	60.3
mri_Latn	7.9	8.7	**9.4**	9.3	52.0	54.1	**56.1**	53.1	53.3	56.0	53.3	**58.6**
mrw_Latn	12.6	13.3	**13.7**	**13.7**	49.4	49.8	**51.0**	49.1	49.7	50.0	42.2	**51.3**
msa_Latn	**48.4**	40.8	44.5	43.4	50.0	**51.8**	50.3	49.9	52.1	**56.4**	53.9	56.1
mwm_Latn	8.6	**12.9**	10.7	10.5	4.9	14.7	15.3	**16.2**	9.3	18.0	18.9	**21.6**
mxv_Latn	5.5	8.0	**10.9**	6.9	19.9	19.6	**22.6**	18.0	25.1	**27.9**	25.5	26.9
mya_Mymr	5.5	21.2	16.6	**23.0**	4.9	20.9	20.6	**24.4**	6.0	15.6	20.5	**29.9**
myv_Cyrl	7.4	**9.8**	9.2	8.6	7.1	14.5	13.7	**15.1**	9.7	11.9	14.3	**14.9**
mzh_Latn	6.4	**12.1**	9.4	9.5	36.2	41.0	**41.9**	41.2	34.0	40.4	**43.7**	41.6
nan_Latn	5.6	7.9	**8.4**	7.5	4.9	10.9	**12.6**	11.5	7.4	14.2	15.3	**19.2**
naq_Latn	7.1	**7.8**	7.5	7.2	18.5	31.7	**36.3**	34.1	28.7	37.3	36.3	**37.7**
nav_Latn	7.0	**9.0**	7.1	7.5	12.0	17.2	18.4	**19.9**	17.8	19.8	19.5	**23.0**
nbl_Latn	15.7	18.6	**19.6**	16.6	54.9	**55.0**	54.5	51.8	52.2	57.0	53.7	**57.2**
nch_Latn	7.3	**8.0**	6.6	6.3	42.4	44.3	**47.3**	43.6	49.6	51.0	**51.7**	50.6
ncj_Latn	6.7	6.9	**9.2**	5.3	42.1	45.6	**45.9**	41.5	48.3	43.9	46.8	**48.9**
ndc_Latn	9.3	11.6	**12.3**	9.3	50.8	49.0	**51.5**	49.8	**53.8**	**53.8**	**53.8**	52.9
nde_Latn	15.7	18.6	**19.6**	16.6	54.9	**55.0**	54.5	51.8	52.2	57.0	53.7	**57.2**
ndo_Latn	10.3	**11.7**	10.3	10.3	46.9	**47.0**	46.4	42.5	52.4	50.4	52.6	**53.7**
nds_Latn	10.6	11.2	11.6	**13.0**	40.7	41.7	**44.4**	41.7	46.4	48.5	45.5	**49.5**
nep_Deva	10.6	32.1	36.9	**43.9**	21.3	50.6	**51.8**	51.4	31.3	49.7	55.1	**56.1**
ngu_Latn	6.8	**9.0**	7.3	7.2	51.5	52.1	**54.6**	51.7	53.4	54.0	**55.7**	52.6
nld_Latn	67.0	**67.5**	66.3	63.5	**67.6**	63.7	64.4	65.0	**67.0**	65.5	63.8	65.6
nmf_Latn	6.8	8.7	**9.0**	7.6	38.5	44.2	44.8	**45.7**	41.1	47.6	47.0	**48.2**
nnb_Latn	9.7	**12.1**	11.2	10.0	49.7	49.5	**50.0**	46.7	51.3	49.0	49.8	**52.4**
nno_Latn	**59.4**	54.1	58.8	59.1	64.6	65.7	**66.4**	64.0	66.3	65.5	66.3	**66.6**
nob_Latn	64.8	61.7	64.2	**67.4**	59.8	60.9	61.7	**66.7**	65.3	62.9	61.7	**66.5**
nor_Latn	65.1	62.3	63.3	**66.2**	61.1	59.0	60.4	**64.1**	66.1	64.4	63.4	**66.5**
npi_Deva	9.7	42.9	39.6	**46.6**	24.5	56.9	54.7	**59.6**	29.1	51.8	56.6	**58.1**
nse_Latn	12.4	13.6	13.9	**14.2**	48.8	48.4	**51.3**	48.1	54.3	52.6	51.7	**57.8**
nso_Latn	9.8	**12.5**	11.5	11.6	57.3	59.5	**62.0**	59.9	60.2	63.6	62.7	**64.5**
nya_Latn	7.8	**14.3**	**14.3**	11.6	**65.7**	65.6	63.8	63.5	61.9	62.4	61.5	**64.5**
nyn_Latn	9.7	**11.3**	9.7	10.6	**44.9**	42.9	42.5	42.8	48.6	48.7	45.3	**50.7**
nyy_Latn	7.6	**11.5**	9.6	9.6	36.4	37.0	**38.7**	34.0	**46.5**	42.1	43.2	45.6
nzi_Latn	10.3	**13.7**	12.5	10.7	25.1	34.1	**34.4**	31.8	34.9	37.3	35.6	**37.4**
ori_Orya	7.4	27.0	27.6	**31.6**	11.5	45.4	**49.6**	44.9	23.9	38.8	**45.5**	41.5
ory_Orya	8.5	30.2	30.0	**35.1**	15.0	**48.1**	**48.1**	46.1	23.1	44.1	45.6	**49.0**
oss_Cyrl	6.0	8.9	8.5	**9.5**	5.1	**40.7**	37.6	33.7	8.7	37.4	**37.9**	35.4
ote_Latn	8.3	**12.2**	**12.2**	7.2	15.1	20.7	**25.7**	22.3	21.1	24.5	27.9	**28.6**
pag_Latn	19.4	21.3	20.9	**22.0**	**60.7**	60.3	60.6	55.8	**60.1**	56.4	59.3	57.4
pam_Latn	17.8	18.2	**22.3**	20.2	44.9	43.5	**47.1**	43.0	**50.4**	50.1	48.8	49.5
pan_Guru	9.3	34.1	38.6	**44.1**	9.1	40.5	**48.3**	47.4	29.0	44.1	**54.8**	53.7
pap_Latn	31.7	29.4	**35.5**	31.6	**62.6**	60.5	61.7	61.3	71.4	**72.7**	69.6	70.9
pau_Latn	9.1	13.9	12.1	**14.4**	44.2	41.3	**44.4**	42.1	**49.2**	47.7	48.8	48.5
pcm_Latn	**38.7**	27.2	34.0	30.6	**64.8**	63.0	61.6	64.4	64.2	65.3	63.9	**67.0**
pdt_Latn	13.1	16.7	**19.3**	18.4	56.1	57.9	57.4	**58.9**	56.8	59.6	58.4	**60.3**
pes_Arab	9.0	46.6	49.8	**53.9**	9.0	54.9	55.4	**59.2**	23.4	54.2	58.4	**61.8**
pis_Latn	12.4	**17.1**	14.6	9.3	65.4	67.9	68.3	**68.9**	68.1	**69.8**	64.5	69.1
pls_Latn	12.9	20.9	**22.5**	20.3	52.2	51.0	**54.2**	51.4	50.2	51.8	**54.9**	51.8
plt_Latn	31.2	34.0	**37.3**	35.2	51.7	56.2	**56.6**	49.7	**61.5**	58.8	58.9	56.8
poh_Latn	12.1	**13.9**	10.6	10.2	**55.7**	54.5	55.2	52.7	50.9	51.4	52.1	**55.8**
pol_Latn	65.7	65.7	**66.5**	63.6	59.6	62.2	**63.3**	60.9	62.2	65.0	63.2	**66.1**
pon_Latn	7.1	8.3	8.7	**9.1**	55.6	55.3	56.1	**57.0**	61.7	60.6	**62.8**	62.1
por_Latn	**68.4**	65.1	67.2	65.5	**66.9**	64.3	63.2	63.7	67.8	**68.9**	68.7	67.8
prk_Latn	6.6	8.5	**9.3**	7.5	57.3	59.5	**62.2**	59.7	58.1	61.0	64.0	**68.0**
prs_Arab	9.3	50.6	54.4	**59.9**	7.3	55.3	61.0	**65.1**	26.8	61.7	66.7	**70.3**
pxm_Latn	9.2	10.4	**10.6**	8.8	20.2	**39.2**	39.1	33.1	26.4	32.7	33.9	**34.9**
qub_Latn	9.3	11.3	**11.4**	8.0	60.2	62.7	**64.9**	59.0	65.8	**66.6**	63.4	61.4
quc_Latn	8.9	**12.0**	11.4	10.8	**54.2**	51.6	51.5	51.3	51.0	47.7	51.6	**53.6**
qug_Latn	9.9	**14.4**	12.3	11.9	67.3	67.2	**67.7**	66.1	70.6	**74.8**	70.3	70.6
quh_Latn	7.8	**12.9**	12.0	12.7	66.7	67.0	**67.7**	66.2	68.0	66.7	66.2	**68.8**
quw_Latn	10.2	11.1	**14.4**	11.4	52.4	**58.3**	57.9	55.8	**59.7**	58.9	54.5	57.0
quy_Latn	8.8	14.7	14.2	**15.2**	72.2	72.8	**73.1**	71.0	74.6	73.5	72.8	**74.8**
quz_Latn	8.5	**13.4**	12.9	12.4	71.6	**72.6**	70.2	69.6	**75.1**	72.0	70.9	73.4
qvi_Latn	6.7	10.8	**11.6**	9.8	63.5	65.4	**67.1**	61.6	63.9	**69.2**	65.9	69.0
rap_Latn	8.0	12.1	**13.1**	11.0	40.1	**49.4**	45.0	40.0	49.7	50.6	**50.7**	49.7
rar_Latn	7.8	10.1	**10.7**	8.3	56.2	56.7	**60.0**	54.2	57.8	57.8	**58.5**	55.2
rmy_Latn	**13.7**	12.8	12.7	12.9	51.3	50.2	**54.3**	53.1	**55.9**	55.1	50.6	54.9
ron_Latn	67.7	69.0	67.7	**69.8**	60.0	59.6	**62.9**	59.5	63.0	64.3	63.2	**65.4**
rop_Latn	7.7	**11.0**	10.1	10.8	62.6	62.1	62.6	**62.8**	62.3	**63.2**	62.2	61.8
rug_Latn	6.3	8.5	**9.8**	8.4	60.3	63.5	**63.7**	60.1	63.6	58.5	61.3	**64.2**
run_Latn	11.7	**16.6**	15.7	15.1	56.5	**58.0**	55.7	54.9	57.6	57.1	55.2	**58.1**

Table 18: F1 scores of models on **transliterated** dataset of **Taxi1500** (Part III).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
rus_Cyrl	30.3	54.8	**61.3**	57.0	43.4	52.9	53.3	**55.3**	54.6	**59.7**	58.6	58.0
sag_Latn	8.6	11.2	**11.9**	11.1	50.5	**52.9**	52.2	50.9	48.7	48.7	49.0	**49.7**
sah_Cyrl	7.2	10.1	**10.7**	10.3	5.1	**36.7**	30.7	34.2	10.1	40.2	**40.7**	39.4
sba_Latn	12.7	**13.2**	12.7	11.5	12.6	21.6	**22.0**	21.9	20.2	23.0	24.8	**25.4**
seh_Latn	8.9	12.1	**13.0**	11.2	51.8	53.6	**54.6**	53.2	56.9	57.1	55.9	**60.6**
sin_Sinh	9.8	34.7	**40.6**	39.8	6.4	28.5	30.3	**32.6**	13.0	34.4	35.4	**40.8**
slk_Latn	54.7	57.0	**61.5**	59.4	52.4	55.3	**57.9**	55.0	59.7	62.3	63.4	**64.3**
slv_Latn	**63.3**	61.1	62.5	62.8	60.7	66.7	**67.4**	64.2	65.4	**68.9**	68.5	68.0
sme_Latn	9.2	12.1	**12.4**	10.8	30.1	36.5	**39.4**	37.2	30.9	37.4	37.8	**39.2**
smo_Latn	9.7	**10.6**	9.5	9.4	60.1	**60.6**	58.8	59.4	**62.2**	60.6	60.7	59.2
sna_Latn	10.3	**13.4**	11.5	10.7	43.0	**43.5**	41.4	41.1	**55.6**	53.2	55.3	54.6
snd_Arab	6.4	39.1	38.9	**44.9**	5.1	43.5	**53.6**	51.0	10.6	42.1	45.4	**49.9**
som_Latn	34.5	32.7	37.0	**38.2**	**41.0**	35.5	38.7	34.4	36.0	**38.0**	35.6	35.9
sop_Latn	7.8	**9.9**	9.5	9.3	36.7	37.1	**39.4**	36.5	40.9	43.4	40.4	**45.5**
sot_Latn	10.4	**11.9**	11.4	10.7	57.9	55.6	**58.4**	51.7	**60.6**	58.3	57.3	57.1
spa_Latn	71.5	73.5	**73.7**	71.5	65.4	**69.5**	66.7	68.4	**72.7**	71.5	72.0	71.6
sqi_Latn	63.3	**67.8**	64.4	65.1	65.9	68.3	**69.7**	67.9	69.5	70.5	**72.1**	70.8
srm_Latn	7.7	**12.1**	10.7	9.4	45.0	**55.1**	54.7	50.1	46.1	59.5	**61.0**	58.9
srn_Latn	8.1	**12.1**	**12.1**	10.7	**69.4**	68.4	67.6	**69.4**	72.0	70.9	68.9	**73.2**
srp_Latn	**65.8**	60.9	64.0	65.0	65.1	68.7	**71.1**	68.3	69.8	71.7	71.3	**74.6**
ssw_Latn	10.5	**13.8**	12.7	12.3	48.6	48.1	**49.4**	45.2	**52.9**	48.9	50.2	46.8
sun_Latn	49.9	46.6	48.4	**51.9**	53.0	51.5	**54.4**	52.3	54.8	55.5	**56.1**	56.0
suz_Deva	10.8	**11.4**	10.6	11.1	5.9	22.7	**22.8**	22.7	7.9	29.9	29.6	**31.2**
swe_Latn	67.3	**71.3**	70.7	69.6	58.9	59.8	62.6	**66.0**	63.8	68.5	68.9	**69.4**
swh_Latn	59.2	**60.5**	58.1	60.2	**63.3**	62.4	62.9	60.2	65.7	**67.5**	65.9	65.9
sxn_Latn	6.7	10.3	11.6	**12.5**	41.0	44.9	**48.2**	43.9	46.7	50.6	50.8	**53.3**
tam_Taml	9.9	41.8	42.1	**45.2**	17.0	44.1	**47.7**	45.8	22.9	47.5	47.7	**50.2**
tat_Cyrl	9.9	**18.1**	16.5	16.9	33.7	55.1	**58.2**	47.3	41.8	52.3	49.8	**52.7**
tbz_Latn	8.9	**12.7**	12.2	12.1	8.7	15.2	18.4	**19.1**	14.0	19.9	21.5	**21.6**
tca_Latn	6.4	9.9	**11.2**	7.6	7.8	32.7	35.4	**37.7**	15.1	36.1	**39.8**	39.5
tdt_Latn	8.8	12.7	**16.2**	13.8	65.8	65.9	66.7	**67.6**	68.1	67.0	**68.8**	66.9
tel_Telu	5.9	31.4	**33.4**	31.4	8.4	30.6	29.5	**33.1**	20.9	37.3	39.0	**39.6**
teo_Latn	8.6	**11.0**	10.1	8.3	**26.8**	24.4	25.2	24.3	32.0	32.7	30.9	**33.1**
tgk_Cyrl	10.7	**14.1**	11.1	12.5	37.4	50.9	**54.5**	52.4	50.6	55.7	**56.1**	56.0
tgl_Latn	**53.4**	49.9	50.8	52.9	61.3	**62.4**	61.0	59.4	61.6	**65.8**	61.6	65.7
tha_Thai	6.1	25.4	27.8	**33.4**	5.8	17.2	24.1	**24.9**	5.9	21.4	29.4	**34.7**
tih_Latn	7.6	10.6	**13.5**	10.8	62.6	61.4	**63.0**	56.9	64.8	63.3	62.4	**65.3**
tir_Ethi	9.4	**13.9**	13.2	13.5	4.9	22.2	**23.3**	20.8	9.2	22.1	22.9	**28.1**
tlh_Latn	30.4	31.1	**31.5**	27.2	**70.0**	66.3	66.1	69.3	64.0	63.2	64.4	**65.9**
tob_Latn	6.4	**8.7**	8.0	6.8	46.8	**52.4**	52.3	50.3	41.4	48.1	**50.0**	46.4
toh_Latn	7.5	**13.3**	12.6	11.7	**44.9**	42.8	43.4	39.4	49.3	44.6	46.3	**50.5**
toi_Latn	12.7	11.4	**13.9**	13.1	49.9	46.5	**53.4**	45.7	51.3	53.4	50.8	**53.6**
toj_Latn	8.4	13.2	**13.3**	12.5	**43.8**	41.3	42.3	39.4	44.0	46.9	**47.8**	43.7
ton_Latn	7.3	8.3	**8.8**	7.5	52.7	52.1	**57.5**	51.0	54.5	58.0	**60.2**	55.9
top_Latn	8.5	10.6	10.2	**12.0**	24.3	**25.9**	24.2	24.6	27.4	26.3	28.0	**29.9**
tpi_Latn	8.6	12.6	**14.5**	12.0	66.8	67.9	66.7	**68.4**	68.0	**69.3**	67.4	68.3
tpm_Latn	9.0	10.3	**12.4**	10.8	45.3	46.8	49.2	**49.7**	42.5	49.3	**51.2**	48.2
tsn_Latn	9.5	**9.8**	9.7	8.4	51.9	52.7	**55.8**	49.8	56.7	55.5	53.8	**57.8**
tsz_Latn	8.0	**11.7**	10.9	10.6	31.8	39.8	**41.2**	36.3	35.9	40.1	40.4	**43.6**
tuc_Latn	6.3	8.7	**9.6**	8.3	55.3	60.9	**61.8**	57.4	59.2	64.3	**65.6**	62.3
tui_Latn	6.3	10.2	9.6	**10.8**	13.5	**41.5**	40.8	40.7	19.0	43.8	43.0	**45.4**
tuk_Latn	29.3	31.7	**33.0**	30.0	54.3	57.8	**59.6**	55.6	52.6	54.9	57.2	**57.3**
tum_Latn	10.6	14.3	**14.8**	12.9	48.9	**53.0**	52.0	52.5	58.1	**60.6**	58.6	59.7
tur_Latn	58.4	64.7	**68.7**	63.1	57.2	62.6	**64.5**	62.8	59.7	63.1	64.8	**64.9**
twi_Latn	9.4	**13.2**	12.1	9.3	33.7	**39.1**	39.0	36.6	35.7	**44.1**	39.7	42.4
tyv_Cyrl	7.6	7.4	8.6	**9.4**	12.4	39.6	**39.8**	37.0	17.3	38.1	36.2	**40.6**
tzh_Latn	9.8	**12.1**	11.9	10.7	43.9	**50.3**	49.8	43.5	**51.0**	50.1	50.1	50.5
tzo_Latn	8.7	**13.3**	11.4	11.4	38.5	41.6	**44.4**	38.5	**48.5**	42.8	44.1	45.9
udm_Cyrl	8.2	**13.5**	11.9	11.4	7.7	22.4	21.3	**23.9**	10.1	21.0	21.4	**24.0**
ukr_Cyrl	26.2	43.1	**50.4**	48.0	27.2	40.1	**46.5**	42.3	40.5	**48.4**	47.0	46.2
urd_Arab	24.4	33.9	36.7	**37.0**	36.6	49.2	**54.7**	52.7	44.7	45.2	45.2	**51.5**
uzb_Latn	54.2	**55.7**	52.9	53.3	60.5	**65.3**	62.0	63.0	66.0	65.5	64.6	**66.2**
uzn_Cyrl	33.4	37.2	37.7	**40.7**	51.3	61.0	**64.3**	**64.3**	60.4	66.0	**66.6**	66.2
ven_Latn	6.8	**10.0**	8.5	9.0	41.9	42.6	**43.3**	39.7	49.4	47.8	47.1	**51.7**
vie_Latn	12.0	23.7	26.4	**32.9**	5.3	12.1	17.7	**20.8**	12.2	18.8	25.8	**30.4**
wal_Latn	**10.5**	9.9	9.4	10.0	47.3	42.9	**50.5**	41.3	42.5	42.2	**44.1**	43.0
war_Latn	15.0	20.2	21.1	**21.2**	50.6	**55.8**	53.8	52.8	55.4	55.0	**56.3**	56.1
wbm_Latn	6.3	8.2	**8.9**	7.5	57.5	57.7	**60.5**	59.5	59.3	62.0	64.5	**66.4**
wol_Latn	10.4	**15.5**	14.7	12.8	28.9	31.4	**33.6**	30.0	29.7	34.0	**37.0**	36.3
xav_Latn	8.8	10.7	9.6	**12.2**	11.4	17.2	**20.0**	19.7	17.8	24.2	23.9	**27.3**
xho_Latn	20.2	19.6	**20.9**	19.5	**51.7**	50.0	51.0	48.8	49.3	53.0	**54.0**	52.5
yan_Latn	5.5	6.3	**7.2**	5.4	56.3	57.3	**58.6**	53.2	51.4	54.9	**57.6**	55.7
yao_Latn	10.1	**11.8**	9.5	10.2	43.6	48.7	**50.0**	46.3	53.2	51.9	53.0	**55.2**
yap_Latn	6.2	10.0	9.3	**11.9**	46.1	47.5	**48.9**	45.3	**52.2**	50.4	52.1	52.1
yom_Latn	7.7	**12.2**	10.0	10.7	38.0	**39.1**	37.9	37.6	45.6	43.9	44.6	**47.6**
yor_Latn	8.7	10.0	**10.7**	8.8	26.3	29.0	**32.8**	30.1	36.3	40.8	40.4	**43.6**
yua_Latn	7.3	**11.3**	10.0	7.7	25.4	35.1	**37.1**	33.1	35.3	38.0	37.9	**42.3**
yue_Hani	6.7	5.4	6.6	**7.2**	5.1	5.3	**6.9**	6.1	7.2	6.5	6.8	**7.4**
zai_Latn	13.3	**18.9**	17.9	18.5	37.2	39.0	**40.9**	39.8	42.2	42.1	41.2	**44.5**
zho_Hani	6.7	**22.5**	18.8	20.7	5.1	27.2	30.2	**32.0**	6.3	26.4	**32.2**	31.8
zlm_Latn	**72.3**	70.8	70.5	68.0	**74.3**	71.5	71.7	72.2	74.8	75.0	73.8	**75.6**
zom_Latn	7.4	**9.4**	8.6	7.2	60.0	**60.5**	59.5	60.1	61.2	**64.3**	63.3	64.0
zsm_Latn	72.3	72.3	**73.9**	70.2	70.4	70.0	**71.1**	69.4	70.7	**71.7**	71.4	**71.7**
zul_Latn	24.1	24.2	**24.4**	23.2	58.4	58.9	**61.0**	58.3	60.0	58.3	**61.2**	60.2

Table 19: F1 scores of models on **transliterated** dataset of **Taxi1500** (Part IV).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
ace_Latn	**58.2**	56.7	55.7	56.4	74.7	**75.3**	72.8	73.7	**73.6**	72.8	73.0	70.3
acm_Arab	14.2	**71.5**	67.1	70.8	16.2	74.9	75.2	**75.4**	15.4	70.7	**74.8**	73.1
afr_Latn	**85.3**	84.2	83.4	84.4	81.1	81.0	**82.1**	81.5	**82.4**	**82.4**	82.1	81.0
ajp_Arab	13.5	**70.5**	64.2	67.8	17.4	72.5	71.3	**74.0**	21.6	69.6	**73.5**	71.2
aka_Latn	35.0	34.7	**35.4**	34.0	55.6	**58.6**	57.4	55.3	56.6	57.8	**58.7**	55.8
als_Latn	79.8	80.8	80.1	**81.6**	**81.4**	79.8	79.0	78.9	**81.5**	81.2	81.2	80.1
amh_Ethi	15.5	**51.3**	43.8	45.2	19.9	**51.4**	47.5	48.3	23.0	48.3	**53.1**	50.2
apc_Arab	13.9	**70.8**	69.9	70.2	17.6	73.3	72.7	**74.5**	17.7	68.0	**72.7**	68.7
arb_Arab	11.8	**76.5**	69.7	75.5	15.0	76.2	76.3	**77.6**	15.8	71.7	**77.7**	75.9
ary_Arab	12.2	**65.5**	59.2	62.4	13.9	**70.3**	67.2	69.4	15.1	63.4	**68.4**	64.1
arz_Arab	13.3	**74.0**	69.0	71.9	15.2	72.2	73.0	**76.3**	16.6	72.3	**76.9**	74.2
asm_Beng	11.9	42.9	42.4	**49.0**	21.5	**63.9**	62.2	63.4	30.2	54.7	56.9	**57.9**
ast_Latn	81.0	**83.0**	79.5	80.8	**88.1**	87.6	85.3	85.5	**86.2**	85.5	86.0	85.6
ayr_Latn	28.3	**32.3**	29.0	29.3	**51.4**	50.6	51.1	50.1	**51.6**	50.1	50.0	50.6
azb_Arab	11.9	55.7	56.3	**56.6**	19.3	60.3	62.5	**63.1**	21.9	54.1	**57.3**	54.3
azj_Latn	68.0	**80.9**	76.5	79.9	76.8	83.1	**84.9**	84.6	80.0	**84.5**	84.3	83.3
bak_Cyrl	50.5	**52.4**	47.4	50.7	62.8	74.1	**74.5**	73.5	68.6	73.6	**75.6**	73.1
bam_Latn	31.0	**31.3**	28.8	30.2	43.2	**46.2**	**46.2**	45.3	46.9	48.8	**49.4**	47.0
ban_Latn	71.7	**73.2**	70.9	69.9	79.0	**80.0**	79.9	78.7	79.1	**79.9**	79.7	79.0
bel_Cyrl	43.5	**75.3**	69.7	73.8	48.4	72.1	72.5	**74.7**	59.1	77.1	**78.8**	77.2
bem_Latn	**36.0**	35.9	32.9	32.2	64.3	66.5	65.3	**67.0**	65.7	64.7	**66.4**	64.6
ben_Beng	13.1	57.3	56.4	**58.6**	25.1	62.1	62.1	**65.1**	34.4	57.9	61.7	**62.8**
bjn_Latn	**69.7**	69.0	67.8	64.5	**78.5**	76.4	77.1	76.8	**79.5**	77.6	79.4	78.8
bod_Tibt	9.7	**11.7**	9.4	11.1	13.6	**64.8**	63.8	63.9	13.6	61.3	**63.2**	61.9
bos_Latn	**87.0**	85.3	84.4	84.8	86.4	87.9	87.1	**88.1**	86.5	87.8	**88.4**	87.2
bul_Cyrl	68.6	78.6	76.5	**79.2**	71.2	79.9	**81.6**	79.7	78.3	81.6	**81.8**	81.7
cat_Latn	86.2	**88.0**	86.5	85.2	84.9	84.7	84.2	**85.9**	85.0	**85.7**	85.4	85.0
ceb_Latn	69.7	**72.3**	68.2	69.5	**84.4**	**84.4**	83.3	82.8	81.7	82.3	81.5	**82.4**
ces_Latn	81.5	**85.0**	82.8	82.4	80.3	80.2	80.6	**81.8**	82.6	84.1	**85.3**	**85.3**
cjk_Latn	31.5	**32.6**	32.2	31.6	49.7	**49.8**	48.5	48.9	46.8	43.9	**47.2**	46.5
ckb_Arab	20.2	**20.7**	19.0	20.4	23.5	70.5	70.0	**72.2**	27.0	69.3	**70.4**	**70.4**
crh_Latn	62.7	**71.5**	68.7	70.0	68.6	**76.2**	75.1	74.7	75.1	76.6	**78.2**	76.0
cym_Latn	**75.2**	74.7	72.1	72.0	75.3	**75.7**	74.0	73.5	**75.2**	74.2	74.0	73.9
dan_Latn	85.0	84.7	85.4	**85.6**	84.2	83.8	**84.8**	84.6	84.0	84.1	**84.9**	83.7
deu_Latn	87.2	**89.4**	**89.4**	87.8	83.5	84.5	83.7	**84.8**	84.4	86.0	**86.2**	85.7
dyu_Latn	33.7	**35.8**	34.6	35.7	43.3	43.7	44.2	**45.7**	45.3	45.3	**46.2**	44.0
dzo_Tibt	7.6	**8.0**	7.9	7.3	9.5	54.4	56.4	**57.8**	7.7	51.1	**53.4**	48.8
ell_Grek	29.5	**62.7**	58.3	61.9	34.8	60.8	**61.3**	60.7	37.8	51.1	**58.6**	54.5
eng_Latn	**90.8**	89.2	90.3	89.7	**89.0**	88.5	87.4	88.4	87.5	87.5	**88.1**	**88.1**
epo_Latn	79.1	**82.6**	80.4	80.4	79.5	**81.7**	80.2	79.1	82.1	82.1	82.3	**82.4**
est_Latn	78.9	**82.9**	80.9	80.8	78.2	77.1	76.9	**79.0**	80.4	**80.7**	79.3	80.3
eus_Latn	82.0	**83.0**	80.7	81.8	82.8	**83.4**	83.2	83.1	**82.7**	82.3	82.6	81.9
ewe_Latn	27.6	27.1	**27.8**	27.6	42.4	43.6	**45.6**	43.1	44.6	46.6	**46.9**	45.7
fao_Latn	55.6	**71.0**	67.8	69.7	66.2	77.3	77.7	**77.8**	75.3	80.5	81.3	**82.1**
fij_Latn	27.9	**31.1**	28.7	29.9	61.2	61.2	**61.7**	59.5	62.2	62.5	**63.3**	61.9
fin_Latn	84.6	**88.7**	87.2	87.1	77.2	79.6	80.4	**80.8**	81.5	83.0	**84.0**	82.8
fon_Latn	27.0	**29.0**	25.3	28.0	38.5	41.5	**41.7**	39.4	39.7	41.9	**42.4**	41.1
fra_Latn	87.9	**88.8**	88.5	88.0	84.5	**86.0**	85.0	84.8	85.1	85.5	**87.2**	85.7
fur_Latn	66.7	**66.9**	61.5	61.2	**79.8**	79.3	78.1	78.4	77.6	78.9	**81.2**	79.4
gla_Latn	48.7	**52.1**	50.7	49.5	58.8	58.2	58.8	**60.5**	58.9	58.4	**60.7**	59.5
gle_Latn	57.3	**63.4**	58.2	60.2	54.3	62.3	60.7	**63.1**	59.7	64.6	**65.6**	64.6
glg_Latn	86.6	**87.4**	87.3	86.4	**86.3**	85.2	84.9	85.7	84.9	85.6	**86.7**	85.4
grn_Latn	56.6	**58.1**	56.1	55.3	69.0	**70.4**	70.1	70.1	69.7	**71.3**	71.2	70.5
guj_Gujr	13.8	56.8	47.6	**59.4**	28.9	62.4	62.1	**65.4**	42.2	57.6	**62.3**	62.2
hat_Latn	48.5	47.9	45.8	**50.2**	74.6	**75.8**	74.7	74.4	76.3	77.1	**77.8**	77.5
hau_Latn	55.9	**58.4**	54.6	53.0	**66.3**	63.5	64.0	63.1	**67.0**	65.2	64.9	65.3
heb_Hebr	9.8	65.3	58.4	**66.2**	13.4	62.5	62.6	**65.8**	15.9	59.9	**65.2**	64.9
hin_Deva	16.4	69.0	63.6	**71.8**	33.0	69.4	67.9	**73.0**	47.1	62.0	69.9	**72.4**
hne_Deva	15.6	59.1	55.3	**62.4**	30.4	66.9	64.0	**69.9**	40.2	53.2	57.0	**60.7**
hrv_Latn	**88.1**	87.9	86.6	85.9	85.8	**86.8**	86.7	**86.8**	87.4	87.5	87.7	**87.8**
hun_Latn	74.4	**86.8**	85.3	85.9	66.2	81.5	**82.0**	**82.0**	74.8	84.9	**86.2**	84.9
hye_Armn	30.8	**65.8**	62.9	65.2	40.3	**71.3**	69.4	68.3	45.8	**64.9**	64.5	62.5
ibo_Latn	32.4	**32.5**	30.0	29.7	67.1	69.0	68.7	**70.3**	66.4	70.9	69.9	**71.2**
ilo_Latn	56.6	**59.2**	54.8	56.2	**78.8**	77.9	76.2	75.4	**78.5**	**78.5**	**78.5**	77.9
ind_Latn	88.9	88.7	87.3	**89.0**	**88.4**	87.5	88.0	87.8	**87.6**	86.3	86.2	86.5
isl_Latn	50.7	**71.4**	69.1	**71.4**	56.1	72.4	71.3	**73.5**	67.6	78.2	78.7	**78.9**
ita_Latn	87.0	**88.2**	86.8	86.8	85.5	85.3	84.7	**86.7**	85.8	85.8	86.6	**86.7**
jav_Latn	77.8	**79.6**	77.6	77.5	78.7	79.1	79.8	**81.0**	81.0	81.6	**81.8**	80.3
jpn_Jpan	16.8	69.6	67.1	**71.1**	16.1	66.2	**69.3**	68.8	15.7	64.8	70.1	**70.4**
kab_Latn	**21.1**	20.1	17.4	17.8	29.7	32.2	**32.5**	31.3	31.4	32.4	**33.0**	32.4
kac_Latn	**34.6**	34.4	31.7	33.6	**53.3**	53.2	51.6	51.9	**55.6**	53.9	54.1	54.5
kam_Latn	37.6	**37.7**	37.0	34.7	**48.0**	46.6	47.1	47.7	48.5	48.6	**49.4**	46.5
kan_Knda	16.5	**64.7**	55.0	63.4	27.3	64.6	62.9	**66.4**	37.4	59.6	**71.2**	65.4
kat_Geor	47.4	**70.6**	65.2	67.4	50.5	**74.6**	74.2	72.2	54.5	66.8	**69.4**	66.3
kaz_Cyrl	45.2	**68.5**	66.2	68.4	60.8	75.4	**75.8**	75.2	68.5	75.6	**77.2**	76.7
kbp_Latn	**22.5**	21.6	19.9	21.2	31.0	37.6	**39.2**	37.2	35.6	39.2	**40.0**	38.9
kea_Latn	**68.2**	65.3	63.2	65.3	**76.6**	75.8	75.6	75.6	75.8	75.8	**76.4**	74.7
khm_Khmr	17.6	**73.7**	70.5	73.0	27.2	75.2	**76.0**	73.3	31.5	74.8	**77.1**	74.3
kik_Latn	36.6	**42.2**	40.4	40.5	**52.4**	52.2	52.0	50.8	53.1	**54.6**	**54.6**	52.8
kin_Latn	31.4	**32.2**	29.6	31.9	69.6	68.7	69.7	**69.9**	71.2	72.1	**72.3**	71.9
kir_Cyrl	54.1	**70.9**	70.6	69.4	64.1	72.5	73.7	**75.0**	69.1	72.5	73.1	**74.8**
kmb_Latn	29.6	**33.1**	28.3	31.8	49.4	48.1	50.9	**51.6**	**52.2**	48.2	50.7	48.0
kmr_Latn	50.8	**63.0**	57.8	59.8	61.4	63.1	**64.7**	62.5	64.8	67.6	**68.0**	66.0
kon_Latn	38.4	**39.7**	39.0	39.2	**67.1**	66.4	67.0	64.8	65.8	63.6	**66.9**	66.1
kor_Hang	15.1	**76.4**	72.9	76.1	20.1	72.9	73.1	**74.1**	23.0	**73.5**	73.4	71.4
lao_Laoo	24.3	**69.4**	60.6	61.4	33.4	62.4	**65.9**	63.2	38.2	61.7	**67.7**	60.8

Table 20: F1 scores of models on **transliterated** dataset of **SIB200** (Part I).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
lij_Latn	68.8	**70.8**	66.7	68.5	75.7	76.9	76.2	**77.4**	78.1	77.3	**78.5**	77.9
lim_Latn	70.4	**70.9**	69.3	69.2	**76.1**	75.4	74.9	**76.1**	**78.1**	77.1	77.1	77.0
lin_Latn	**40.2**	37.0	36.7	37.5	**71.3**	71.1	**71.3**	69.5	70.5	68.4	**72.5**	69.8
lit_Latn	81.3	**84.1**	81.9	82.1	79.9	82.3	**83.1**	82.3	83.0	**84.3**	84.1	84.0
lmo_Latn	69.1	**69.7**	64.0	67.2	78.6	**79.8**	79.5	77.8	79.3	**81.0**	80.4	80.7
ltz_Latn	61.1	**62.5**	62.4	61.5	74.9	74.9	**75.7**	74.7	79.1	78.7	**79.9**	78.0
lua_Latn	38.9	41.1	39.4	**41.5**	57.0	58.1	**59.7**	56.1	55.0	54.8	55.6	**56.5**
lug_Latn	**28.7**	28.5	28.0	27.2	55.6	56.5	**56.6**	56.1	**59.1**	56.9	58.3	57.4
luo_Latn	**30.8**	30.0	30.6	30.1	53.6	**54.9**	52.1	53.0	54.2	53.1	**54.5**	53.8
lus_Latn	54.5	**56.1**	53.5	55.2	70.0	**70.9**	69.9	69.5	69.9	**71.6**	71.4	70.7
lvs_Latn	76.5	**81.4**	80.6	79.2	72.8	79.3	79.1	**80.4**	77.2	**82.1**	80.2	80.3
mai_Deva	16.7	64.1	59.8	**68.2**	34.1	72.1	69.4	**72.3**	40.7	61.4	64.4	**68.6**
mal_Mlym	11.8	59.2	56.5	**63.1**	16.1	58.7	59.1	**65.3**	24.2	56.6	**66.5**	60.3
mar_Deva	18.2	**67.4**	63.1	65.5	32.2	69.6	69.6	**70.6**	46.6	70.9	**75.0**	71.5
min_Latn	**70.0**	69.3	68.9	67.2	78.3	**79.2**	**79.2**	77.5	78.6	78.8	**79.8**	**79.8**
mkd_Cyrl	69.5	**77.6**	76.4	77.1	75.4	78.3	**80.7**	79.8	78.1	**78.6**	**78.6**	78.3
mlt_Latn	**49.3**	**49.3**	45.1	47.8	76.8	**84.1**	82.2	82.1	78.9	82.8	**83.9**	83.4
mos_Latn	31.9	33.2	**34.1**	33.7	**42.9**	40.7	39.6	38.6	43.3	**44.3**	42.0	39.5
mri_Latn	32.4	**33.5**	28.1	30.2	62.0	**62.5**	60.7	60.9	61.0	60.5	**62.5**	59.9
mya_Mymr	16.7	**53.7**	52.1	51.6	20.8	58.8	**59.4**	58.3	24.3	60.0	**62.0**	57.2
nld_Latn	**88.1**	**88.1**	87.9	87.2	**86.8**	86.0	85.5	86.5	86.0	**87.0**	85.9	85.5
nno_Latn	83.3	**84.3**	82.8	82.9	84.0	84.6	83.1	**85.3**	82.9	85.0	**85.7**	85.1
nob_Latn	83.5	**85.1**	84.9	83.6	**83.8**	82.5	82.3	83.1	**83.7**	83.1	82.7	83.1
npi_Deva	14.7	**72.5**	63.9	72.4	31.4	**76.6**	74.5	75.2	41.4	71.8	73.0	**73.8**
nso_Latn	28.8	30.5	30.0	**31.4**	**63.7**	60.6	60.8	62.0	65.2	62.3	**65.7**	65.0
nya_Latn	37.8	**40.2**	38.6	38.2	**73.9**	71.7	73.0	72.4	73.4	**73.8**	**73.8**	73.3
oci_Latn	82.0	**83.5**	79.8	80.8	83.0	**83.5**	82.6	82.1	**84.2**	83.5	84.0	83.1
ory_Orya	19.6	**61.2**	56.6	59.7	27.4	65.4	62.4	**65.8**	40.4	50.7	50.7	**54.6**
pag_Latn	62.2	**63.6**	62.7	62.6	**78.8**	77.5	77.4	76.7	**80.0**	78.0	79.2	78.2
pan_Guru	14.1	52.3	47.1	**54.3**	17.8	52.5	50.8	**59.1**	35.1	47.6	53.4	**57.8**
pap_Latn	69.0	**70.4**	66.6	68.1	**79.0**	78.3	78.4	78.7	79.7	79.8	**80.0**	79.7
pes_Arab	18.4	79.3	76.7	**81.8**	22.6	80.5	80.9	**82.2**	27.0	78.3	**82.7**	81.1
plt_Latn	59.1	**64.6**	60.2	58.8	68.9	**72.4**	70.6	68.0	70.3	70.6	**72.5**	70.4
pol_Latn	82.2	**86.6**	83.2	85.3	81.7	**85.5**	83.7	84.1	83.8	83.6	**85.0**	83.5
por_Latn	86.2	**89.5**	87.3	88.6	84.3	85.7	85.5	**86.3**	83.5	**87.0**	**87.0**	86.7
prs_Arab	16.7	77.5	71.0	**77.6**	23.4	**78.7**	78.1	77.8	25.7	75.5	**79.0**	77.3
quy_Latn	43.2	46.2	43.8	**46.7**	**69.1**	64.7	65.8	65.4	66.8	66.5	**68.4**	65.5
ron_Latn	**86.2**	**86.2**	85.5	85.4	83.8	83.4	82.8	**83.9**	**84.4**	**84.4**	83.7	83.3
run_Latn	**27.8**	27.1	27.2	27.1	69.2	**70.7**	69.8	69.4	68.6	**70.7**	69.0	68.9
rus_Cyrl	65.3	81.8	80.3	**82.4**	70.7	81.2	81.3	**81.6**	76.8	79.2	**80.9**	80.1
sag_Latn	38.4	**40.6**	36.3	39.6	59.3	57.8	**59.8**	59.7	59.9	58.4	**60.4**	59.8
san_Deva	10.3	**56.3**	45.4	54.6	23.3	58.6	56.9	**64.7**	32.0	45.3	51.0	**52.1**
sat_Olck	7.0	7.8	6.6	**10.0**	7.8	33.8	32.5	**35.2**	8.5	**26.7**	24.5	23.6
scn_Latn	62.0	**65.2**	61.8	62.5	**78.5**	75.1	75.2	75.7	**78.9**	78.4	78.1	76.8
sin_Sinh	19.1	**65.6**	58.6	61.2	22.7	65.9	**66.7**	65.9	27.6	61.5	**65.5**	64.5
slk_Latn	85.1	**86.1**	83.6	85.4	82.6	82.2	81.7	**83.4**	84.0	83.1	**84.2**	83.6
slv_Latn	84.7	**86.2**	83.8	85.8	80.9	83.0	82.8	**83.2**	83.2	84.7	**84.8**	82.9
smo_Latn	**30.1**	27.0	25.3	29.4	77.3	76.7	76.5	**79.2**	75.9	76.0	76.8	**77.9**
sna_Latn	30.4	**30.9**	26.8	29.9	60.2	**62.2**	61.8	61.0	62.1	**62.7**	61.3	61.8
snd_Arab	14.0	52.5	47.1	**54.2**	17.1	**55.2**	53.7	53.3	18.7	40.0	**47.9**	43.2
som_Latn	**60.3**	60.0	55.4	56.1	57.3	**59.7**	59.4	57.7	**60.6**	59.7	59.4	60.2
sot_Latn	34.6	**35.3**	31.4	33.1	**69.3**	67.1	66.3	66.9	68.2	66.7	**69.1**	66.2
spa_Latn	86.6	**88.7**	87.1	87.2	**85.7**	**85.7**	84.6	85.2	85.2	85.9	85.6	**86.1**
srd_Latn	67.4	**67.6**	62.8	64.7	76.2	74.9	74.0	**76.4**	76.3	**76.7**	76.0	76.3
srp_Cyrl	79.9	**84.1**	83.5	83.0	82.9	85.0	**85.4**	84.9	**84.5**	81.9	82.6	81.4
ssw_Latn	**28.0**	26.3	25.4	26.7	**68.8**	65.8	66.1	65.9	68.2	**69.0**	68.2	68.4
sun_Latn	77.6	**79.6**	77.7	76.4	83.5	81.2	83.0	**83.9**	81.5	**82.1**	80.9	81.0
swe_Latn	81.3	**86.3**	84.7	85.8	79.6	80.6	**81.3**	81.1	80.2	81.8	**82.7**	82.3
swh_Latn	73.0	**74.4**	71.7	73.7	**79.7**	77.0	77.0	78.1	80.2	79.9	**81.1**	80.0
szl_Latn	70.9	**71.5**	70.0	68.8	73.9	**74.6**	73.5	72.1	73.7	74.0	**74.6**	73.3
tam_Taml	14.8	**65.8**	59.1	64.8	23.6	63.3	63.0	**65.0**	25.7	62.6	64.8	**64.9**
tat_Cyrl	51.8	**54.6**	51.0	54.5	64.1	**76.7**	75.5	74.8	69.9	75.7	75.8	**75.9**
tel_Telu	16.7	62.5	56.8	**62.7**	27.3	60.9	59.7	**63.9**	36.1	57.6	**65.3**	62.1
tgk_Cyrl	40.1	**42.9**	37.3	39.6	52.5	74.8	74.4	**75.8**	60.8	**79.8**	79.2	78.1
tgl_Latn	79.2	**81.4**	78.1	79.0	82.9	83.2	**83.9**	82.9	82.0	**83.5**	83.3	82.4
tha_Thai	20.2	75.2	71.8	**80.5**	27.0	74.8	75.7	**78.1**	28.3	74.0	77.9	**78.5**
tir_Ethi	12.8	**32.9**	27.8	28.5	20.2	**44.1**	43.8	42.6	20.5	40.1	**45.6**	40.0
tpi_Latn	52.7	**53.9**	50.2	51.4	**82.7**	82.0	80.0	80.3	79.6	78.2	**80.2**	78.4
tsn_Latn	31.2	**32.5**	27.4	30.5	**62.2**	60.4	61.5	61.5	63.6	63.1	62.3	**63.7**
tso_Latn	**31.6**	30.4	30.3	29.0	**63.3**	60.1	61.7	62.8	64.6	**66.4**	65.5	66.0
tuk_Latn	48.9	52.2	48.0	**52.5**	71.0	75.8	77.0	**77.5**	72.1	**76.3**	76.1	75.8
tum_Latn	31.0	**34.3**	33.0	32.0	71.8	72.5	71.9	**73.3**	71.1	**71.2**	70.5	70.3
tur_Latn	73.1	82.0	82.2	**82.5**	73.9	80.7	**81.0**	80.6	76.4	**83.3**	82.8	80.7
twi_Latn	39.7	**40.1**	39.7	39.3	59.9	**61.9**	60.4	61.5	59.5	**61.4**	60.5	59.0
uig_Arab	20.0	**60.4**	57.0	60.1	27.0	67.5	**68.8**	68.2	34.6	63.1	66.6	**66.8**
ukr_Cyrl	58.1	77.9	76.2	**78.1**	61.1	75.7	77.0	**77.2**	70.8	77.4	**78.5**	**78.5**
umb_Latn	28.0	29.6	**31.1**	29.2	**48.5**	48.1	47.4	48.2	**51.2**	48.3	50.9	49.0
urd_Arab	18.5	**63.9**	61.5	63.5	20.9	64.7	**67.1**	64.5	24.7	60.4	**70.0**	69.2
vec_Latn	**79.4**	78.9	78.1	76.0	**81.7**	81.1	79.8	80.5	82.7	82.8	**82.9**	81.9
vie_Latn	29.7	50.9	47.3	**55.7**	35.5	48.7	**54.0**	53.7	40.5	49.9	**61.2**	57.0
war_Latn	67.0	**72.1**	69.6	68.8	**83.4**	81.8	81.7	82.5	80.8	80.4	**81.3**	80.5
wol_Latn	39.7	**42.0**	40.5	41.4	50.5	**54.4**	52.7	54.1	54.6	55.5	**58.3**	56.5
xho_Latn	**46.0**	43.5	41.6	43.8	64.6	**66.8**	65.0	65.1	68.7	67.6	**68.9**	**68.9**
yor_Latn	**28.0**	26.3	24.8	26.5	51.7	**52.8**	52.0	51.4	52.8	**57.3**	54.8	55.0
zsm_Latn	86.5	**87.4**	85.2	85.7	86.7	86.2	86.1	**87.1**	85.9	**86.2**	85.4	85.0
zul_Latn	**40.9**	38.2	35.9	37.7	**72.0**	71.7	69.9	71.3	73.0	73.2	74.1	**74.7**

Table 21: F1 scores of models on **transliterated** dataset of **SIB200** (Part II).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
ace_Latn	**33.6**	29.5	29.7	30.0	43.1	45.7	**46.2**	44.4	**44.5**	43.8	40.8	42.9
afr_Latn	74.9	75.6	**75.9**	74.9	75.0	76.2	**76.7**	76.4	75.6	**76.2**	75.3	76.1
als_Latn	**61.7**	56.8	57.8	55.7	76.8	**81.2**	78.4	80.3	77.6	79.8	78.5	**82.6**
amh_Ethi	13.9	**34.3**	30.5	29.0	11.1	38.1	**40.0**	**40.0**	23.0	45.7	40.3	**46.2**
ara_Arab	7.3	23.4	27.9	**28.6**	8.9	28.5	29.8	**33.6**	11.7	**40.8**	31.2	35.3
arg_Latn	67.9	67.3	64.0	**70.0**	**78.4**	76.8	76.3	75.0	76.3	**79.4**	71.8	74.4
arz_Arab	11.2	29.9	26.4	**35.3**	8.9	35.6	31.5	**38.1**	13.4	40.2	**42.1**	38.2
asm_Beng	29.8	37.8	36.8	**40.9**	28.7	42.7	44.0	**51.7**	41.6	43.6	55.8	**55.9**
ast_Latn	**80.9**	78.5	79.4	80.2	82.9	**83.2**	71.3	70.9	71.1	**72.6**	72.3	71.1
aym_Latn	38.4	39.1	40.5	**41.6**	43.6	**50.0**	45.9	45.1	**46.6**	45.7	44.3	**46.6**
aze_Latn	51.1	57.3	**59.9**	57.9	55.7	60.4	60.3	**61.5**	60.2	**66.0**	64.3	64.4
bak_Cyrl	17.5	24.0	29.9	**31.3**	37.0	**50.5**	50.2	49.5	41.0	52.4	**54.0**	52.9
bar_Latn	55.0	55.6	**56.1**	55.6	70.1	**73.3**	69.9	69.2	**74.2**	52.7	67.6	72.3
bel_Cyrl	57.8	64.9	**67.6**	**67.6**	64.2	69.3	68.4	**70.6**	69.8	72.5	72.6	**73.8**
ben_Beng	24.7	37.3	42.1	**43.9**	28.9	46.3	49.6	**52.6**	47.9	**57.1**	52.3	53.9
bih_Deva	21.3	**34.7**	30.7	32.9	27.8	42.9	38.4	**47.3**	31.9	**47.2**	40.7	43.2
bod_Tibt	**27.4**	20.7	20.1	26.4	16.4	**34.7**	28.6	27.1	29.9	30.3	30.2	**32.3**
bos_Latn	71.2	**73.2**	73.0	72.5	71.1	**72.3**	71.1	70.9	72.5	74.1	73.7	**74.3**
bre_Latn	**58.4**	55.6	55.0	56.2	60.8	61.1	61.2	**62.3**	62.6	62.5	**62.7**	**62.7**
bul_Cyrl	65.1	68.6	**69.4**	68.5	71.0	72.2	**74.0**	73.4	74.4	**75.2**	75.1	74.7
cat_Latn	**81.9**	80.0	80.3	81.1	**83.2**	82.4	82.6	83.1	**83.6**	83.3	83.0	83.3
cbk_Latn	51.8	49.7	**56.4**	53.6	51.8	48.2	51.5	**55.2**	50.6	52.3	52.7	**54.8**
ceb_Latn	51.9	**56.3**	55.6	53.5	55.8	**58.2**	51.6	55.8	53.9	51.8	**54.3**	53.6
ces_Latn	74.5	75.0	**75.6**	75.2	74.7	76.8	76.5	**76.9**	75.9	**78.0**	77.6	77.4
che_Cyrl	13.9	15.4	**16.3**	**16.3**	30.9	53.6	**62.0**	50.8	38.3	26.5	**40.9**	32.2
chv_Cyrl	**56.4**	51.7	52.3	52.6	60.8	66.7	**69.1**	67.6	63.2	66.4	65.3	**69.3**
ckb_Arab	24.9	22.1	**28.5**	27.4	37.0	**59.8**	55.9	57.9	41.2	**64.5**	63.4	61.9
cos_Latn	56.3	57.0	**58.8**	55.4	60.3	58.4	59.6	**62.5**	57.4	59.1	56.9	**61.7**
crh_Latn	44.3	42.4	44.6	**46.0**	48.9	49.5	**52.9**	51.3	48.1	**53.2**	49.7	49.3
csb_Latn	56.0	55.2	**59.0**	57.7	60.5	61.0	61.9	**66.0**	57.3	61.1	63.3	**64.8**
cym_Latn	57.0	56.9	**59.7**	57.9	60.3	57.9	59.2	**60.6**	58.4	59.5	60.0	**60.4**
dan_Latn	80.6	81.2	81.5	**81.8**	78.7	81.2	80.7	**81.4**	81.1	**82.3**	81.9	81.9
deu_Latn	73.2	73.7	**74.3**	74.0	72.1	74.2	73.7	**75.0**	75.5	**76.5**	75.0	76.2
diq_Latn	**39.9**	38.4	39.4	39.1	**56.1**	51.4	50.7	51.1	42.9	**54.4**	54.0	52.4
div_Thaa	25.0	23.9	22.7	**25.7**	24.6	**31.3**	31.2	29.3	26.0	30.2	29.6	**31.7**
ell_Grek	41.9	56.5	**57.3**	56.9	54.2	62.2	61.9	**63.3**	59.9	**66.3**	65.3	65.6
eml_Latn	**37.2**	32.9	34.9	36.4	**40.4**	39.1	38.0	38.8	37.6	**41.8**	40.1	41.2
eng_Latn	**82.8**	82.4	82.4	82.7	**83.4**	**83.4**	83.3	**83.4**	**83.7**	83.4	83.1	83.5
epo_Latn	61.9	**62.6**	61.8	61.1	68.4	69.3	67.9	**69.4**	69.8	**70.2**	67.9	68.7
est_Latn	69.0	70.2	70.3	**71.6**	68.9	71.3	71.8	**72.6**	71.4	**74.2**	**74.2**	73.4
eus_Latn	58.0	58.8	58.0	**61.2**	56.4	56.2	55.6	**58.5**	56.3	**56.9**	55.5	53.2
ext_Latn	**41.0**	36.5	39.9	39.7	44.3	44.0	**46.8**	44.2	44.0	**46.4**	44.2	43.4
fao_Latn	**61.1**	59.3	59.6	55.9	**68.8**	64.8	66.1	68.2	68.2	**71.9**	69.8	69.5
fas_Arab	5.3	19.3	21.2	**22.4**	12.7	22.0	24.3	**25.0**	17.4	26.6	26.4	**27.9**
fin_Latn	74.5	74.3	75.1	**75.8**	73.0	74.6	72.9	**74.7**	75.3	**76.0**	75.1	74.3
fra_Latn	76.0	75.6	75.4	**76.6**	75.3	75.6	**76.8**	76.1	76.6	**78.2**	77.4	77.8
frr_Latn	45.1	44.6	**45.5**	44.6	54.0	**54.8**	54.4	54.0	54.6	**59.4**	56.8	55.7
fry_Latn	73.9	74.4	**75.0**	73.3	76.2	74.3	75.7	**77.0**	76.0	**79.1**	77.7	77.7
fur_Latn	**56.0**	51.3	51.2	53.8	55.9	55.4	53.8	**56.7**	55.8	**56.7**	55.6	55.5
gla_Latn	52.5	**56.9**	50.6	48.0	**64.7**	**64.7**	58.0	60.7	59.8	**60.8**	58.0	59.9
gle_Latn	63.0	**68.7**	67.5	67.1	67.6	70.4	**72.0**	71.0	70.5	72.3	**73.0**	72.5
glg_Latn	76.5	76.9	77.4	**77.8**	**79.7**	78.8	78.7	78.7	79.3	**80.0**	77.5	78.1
grn_Latn	43.2	38.0	**43.6**	39.7	**54.9**	49.8	50.0	48.2	49.5	53.8	**55.1**	52.1
guj_Gujr	3.3	35.7	38.8	**44.7**	3.8	40.1	**48.0**	40.7	21.4	**52.0**	47.4	51.6
hbs_Latn	58.6	**60.7**	57.4	54.1	**69.3**	64.1	57.4	65.4	**72.5**	67.1	65.6	65.7
heb_Hebr	7.2	23.8	25.4	**28.0**	8.5	23.0	27.4	**28.7**	14.1	27.5	28.0	**31.6**
hin_Deva	21.3	45.7	45.3	**48.6**	30.0	51.7	50.9	**55.3**	50.9	59.2	53.6	**59.6**
hrv_Latn	75.9	75.7	**76.9**	76.2	75.8	76.8	75.8	**77.0**	76.8	**77.7**	77.2	**77.7**
hsb_Latn	55.8	61.6	61.1	**62.8**	**75.4**	74.0	67.6	73.8	70.7	**78.8**	70.8	75.1
hun_Latn	66.4	70.4	**71.2**	70.9	65.4	69.2	**69.3**	69.2	66.7	**71.1**	69.6	69.5
hye_Armn	31.6	**40.7**	39.9	39.9	41.6	48.5	48.5	**49.1**	32.0	**48.2**	48.0	47.0
ibo_Latn	**48.4**	40.2	47.6	44.0	56.7	57.8	59.2	**62.1**	57.0	**60.3**	54.5	56.0
ido_Latn	**69.7**	66.1	63.2	67.5	**82.2**	75.3	79.0	79.3	85.8	81.2	**86.5**	81.9
ilo_Latn	58.0	60.8	66.1	**66.7**	72.3	**77.0**	76.2	74.9	77.8	**78.3**	75.7	77.4
ina_Latn	55.1	56.0	**56.2**	51.3	55.8	60.3	57.9	**60.8**	56.8	59.3	58.6	**59.7**
ind_Latn	48.5	49.2	**49.3**	49.1	52.8	50.9	**55.1**	53.4	51.3	51.0	50.3	**51.6**
isl_Latn	59.4	66.3	**67.3**	65.5	65.5	67.9	68.6	**69.1**	68.5	**73.5**	72.4	72.6
ita_Latn	77.1	76.6	76.4	**77.4**	**78.5**	77.5	77.5	78.4	**78.9**	78.5	77.4	78.1
jav_Latn	**56.2**	55.9	54.2	54.5	54.1	56.1	58.9	**60.2**	56.3	**57.9**	54.3	55.7
jbo_Latn	15.5	18.5	**24.4**	17.5	**25.2**	22.9	24.3	22.6	26.7	**31.8**	28.2	26.4
jpn_Jpan	7.2	7.8	7.5	**9.6**	**8.5**	7.0	6.7	7.6	7.2	7.2	7.4	**7.9**
kan_Knda	9.7	35.2	**38.6**	35.3	15.5	**43.5**	32.5	37.6	28.7	**57.0**	40.0	50.2
kat_Geor	22.8	34.4	**38.7**	38.0	28.6	41.2	40.7	**43.3**	30.0	**44.8**	42.9	42.5
kaz_Cyrl	27.2	37.3	**39.4**	39.0	42.3	44.4	**46.5**	45.9	38.6	**47.5**	44.8	46.4
khm_Khmr	19.2	**32.2**	30.6	30.4	21.0	**31.8**	28.7	30.9	28.9	**35.9**	32.5	34.9
kin_Latn	**64.9**	62.2	60.1	61.1	64.9	63.8	69.1	**70.9**	67.5	67.3	**68.4**	**68.4**
kir_Cyrl	30.3	40.5	**43.5**	42.3	30.8	**45.3**	41.8	40.2	38.5	46.5	46.7	**49.2**
kor_Hang	13.5	24.3	25.3	**27.2**	12.1	25.2	26.9	**28.8**	14.0	**30.7**	30.4	29.7
ksh_Latn	**48.0**	43.2	40.6	43.4	54.7	53.2	**59.3**	57.5	55.5	**56.5**	52.7	55.6
kur_Latn	39.1	45.3	**53.0**	46.4	51.3	61.2	57.9	**61.3**	50.2	57.9	58.0	**59.6**
lat_Latn	66.3	71.4	75.4	**76.4**	71.7	75.7	**80.6**	77.1	75.7	**79.8**	74.8	74.6
lav_Latn	66.9	68.5	70.4	**71.6**	70.3	72.4	71.1	**73.6**	73.2	**76.0**	73.8	73.9

Table 22: F1 scores of models on **transliterated** dataset of **NER** (Part I).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
lij_Latn	37.5	35.7	34.5	**37.8**	**47.4**	43.2	42.7	45.7	**45.1**	44.0	42.4	44.9
lim_Latn	61.5	**65.2**	62.2	57.9	70.5	68.7	72.6	**73.3**	71.4	70.2	67.8	**71.7**
lin_Latn	**39.3**	39.0	34.8	38.0	48.8	54.2	53.3	**54.7**	49.5	56.7	56.2	**57.9**
lit_Latn	67.4	68.3	**69.2**	69.0	69.5	70.6	70.4	**71.9**	71.8	**74.4**	73.5	73.3
lmo_Latn	**72.5**	68.2	68.2	68.7	70.6	72.4	71.5	**74.6**	70.4	**77.4**	72.5	75.1
ltz_Latn	49.5	50.2	51.1	**51.7**	66.7	**67.6**	67.5	67.5	68.9	**70.0**	68.2	69.8
lzh_Hani	**8.3**	5.8	4.8	6.8	**9.0**	7.7	4.3	6.2	6.6	9.0	**9.6**	6.1
mal_Mlym	7.5	25.4	28.7	**34.1**	11.9	31.5	31.9	**35.9**	23.7	**41.1**	40.1	39.9
mar_Deva	9.9	29.1	31.7	**36.5**	12.5	37.8	37.3	**38.4**	24.7	**45.5**	42.7	44.5
mhr_Cyrl	31.8	34.1	**34.6**	32.8	46.9	**56.4**	55.2	54.0	47.3	53.5	53.0	**55.0**
min_Latn	42.2	38.0	**43.6**	39.4	38.4	40.4	40.0	**41.0**	40.7	**43.6**	39.3	41.0
mkd_Cyrl	64.9	68.1	**69.5**	69.4	68.0	69.8	71.2	**71.9**	72.9	**76.8**	75.2	75.4
mlg_Latn	54.5	54.6	**59.3**	56.6	**59.5**	58.5	58.6	58.6	53.8	58.1	59.3	**59.7**
mlt_Latn	47.0	**47.7**	46.0	43.4	69.3	67.7	71.6	**76.9**	73.8	70.9	72.6	**74.8**
mon_Cyrl	51.7	**58.1**	53.2	51.9	**54.2**	51.6	53.8	52.5	53.7	56.5	58.3	**58.9**
mri_Latn	14.4	12.8	**26.2**	12.3	55.9	**61.6**	60.5	59.1	50.2	50.6	**55.4**	50.8
msa_Latn	**68.1**	63.5	65.0	59.1	63.8	**69.2**	68.8	66.6	69.3	69.6	70.1	**70.7**
mwl_Latn	**46.5**	41.8	45.7	35.4	47.5	49.7	49.7	**52.2**	45.6	**48.6**	47.7	48.1
mya_Mymr	9.5	40.4	49.5	**51.9**	9.5	37.7	**44.0**	43.2	8.5	41.0	**46.0**	40.4
mzn_Arab	25.6	19.8	22.1	**26.4**	25.0	31.1	27.4	**33.6**	31.8	**40.2**	34.6	33.8
nan_Latn	**57.6**	49.3	56.6	49.3	60.3	60.2	**66.4**	64.2	54.8	**69.9**	68.6	61.0
nap_Latn	53.0	55.9	**60.5**	55.2	54.9	**56.9**	53.9	55.8	54.1	**61.4**	60.7	59.7
nds_Latn	64.1	60.5	**65.5**	63.7	67.4	76.8	71.7	**77.5**	71.9	75.6	74.9	**78.2**
nep_Deva	9.5	42.8	44.9	**58.4**	21.2	61.5	61.2	**65.1**	35.5	**66.9**	59.3	61.3
nld_Latn	**79.6**	78.6	79.3	78.6	79.6	79.7	79.8	**80.7**	80.7	**81.3**	80.8	81.2
nno_Latn	75.8	75.1	**76.4**	76.1	75.1	**77.7**	77.5	77.3	77.2	**78.6**	77.4	75.8
nor_Latn	75.3	75.4	**76.4**	75.9	74.6	75.7	75.4	**77.2**	76.8	**78.9**	77.4	77.2
oci_Latn	**65.2**	62.6	63.9	64.2	**75.7**	67.9	68.8	71.3	74.7	**75.0**	69.1	70.0
ori_Orya	2.6	16.8	22.2	**23.2**	5.2	**23.6**	23.1	21.6	13.0	23.0	21.7	**25.1**
oss_Cyrl	33.2	33.8	34.0	**37.3**	42.0	**58.8**	51.9	47.1	40.2	**61.2**	52.5	57.3
pan_Guru	22.1	33.2	33.0	**35.0**	24.8	**36.6**	**36.6**	35.8	34.4	37.7	33.3	**38.0**
pms_Latn	69.2	66.3	68.2	**69.8**	79.2	76.5	77.7	**79.6**	78.6	77.8	75.6	**80.0**
pnb_Arab	26.1	29.1	27.6	**33.9**	29.8	45.3	43.7	**46.5**	31.2	46.6	46.4	**48.0**
pol_Latn	**76.6**	76.2	**76.6**	**76.6**	75.9	77.3	77.1	**77.6**	77.9	**78.4**	78.2	78.1
por_Latn	75.8	**76.4**	75.8	76.0	78.9	**79.4**	78.7	79.2	77.4	**79.4**	77.2	78.6
pus_Arab	13.1	26.5	**26.9**	26.1	16.1	31.9	**34.1**	29.7	16.8	29.6	28.9	**31.4**
que_Latn	58.0	55.2	57.6	**61.2**	65.0	**71.6**	66.7	65.0	**66.4**	65.5	64.2	60.3
roh_Latn	**59.2**	50.8	53.9	50.5	54.5	**62.5**	61.7	58.0	59.5	61.8	60.2	**62.2**
ron_Latn	**70.6**	66.1	70.0	68.8	73.0	**75.9**	73.5	73.1	**76.5**	**76.5**	73.8	75.0
rus_Cyrl	53.8	58.4	58.8	**59.4**	57.4	61.7	62.4	**63.5**	62.5	65.4	**66.0**	65.5
sah_Cyrl	38.6	36.8	**45.0**	41.2	56.9	**73.1**	67.0	70.4	49.1	**72.7**	68.5	67.0
san_Deva	0.8	9.4	**10.6**	6.9	3.2	17.7	15.9	**20.6**	9.0	24.9	**29.3**	18.7
scn_Latn	48.9	51.1	**54.2**	51.3	**69.4**	66.4	64.5	68.2	65.5	64.2	**66.1**	64.3
sco_Latn	78.2	**78.9**	78.1	72.5	84.6	82.8	**85.4**	81.5	83.7	82.0	**86.6**	83.2
sgs_Latn	39.7	37.6	**41.4**	37.5	**55.6**	48.5	52.0	50.0	**49.8**	48.2	44.9	46.7
sin_Sinh	20.0	30.8	**33.6**	25.7	23.8	31.8	28.8	**35.0**	24.5	35.3	**37.7**	31.0
slk_Latn	74.0	72.7	**74.1**	72.5	73.5	76.6	**76.7**	76.1	76.5	**78.5**	77.8	77.8
slv_Latn	77.6	76.9	77.7	**78.5**	77.3	79.4	78.5	**80.0**	79.5	**81.2**	80.7	80.2
snd_Arab	14.9	26.2	**29.2**	28.8	14.1	30.4	**32.1**	31.0	15.1	**33.3**	26.7	30.9
som_Latn	**53.4**	48.8	52.9	51.6	55.0	58.3	**61.5**	58.3	52.2	57.1	55.7	**58.6**
spa_Latn	66.3	66.5	69.6	**69.9**	**75.1**	73.0	71.4	72.5	70.2	**71.8**	66.1	69.7
sqi_Latn	71.5	71.7	71.3	**72.6**	**75.8**	75.5	74.1	75.5	74.3	75.5	74.3	**75.9**
srp_Cyrl	57.1	58.1	**58.7**	56.4	58.2	**60.2**	59.9	**60.2**	61.3	**63.3**	61.1	61.9
sun_Latn	**46.4**	40.6	44.2	42.0	**61.8**	51.9	56.9	57.0	**58.3**	54.5	50.4	55.6
swa_Latn	**63.8**	59.8	59.9	58.3	70.0	69.7	70.0	**70.9**	70.5	**71.2**	69.4	70.6
swe_Latn	66.5	70.4	70.9	**71.3**	61.0	68.3	63.4	**71.3**	68.3	72.7	72.6	**73.3**
szl_Latn	**58.8**	56.8	55.7	57.1	57.6	**68.7**	67.5	65.8	58.5	**70.7**	68.6	69.5
tam_Taml	9.6	26.1	27.5	**28.0**	9.0	25.7	27.3	**29.4**	15.4	**33.5**	31.3	32.5
tat_Cyrl	36.3	43.0	41.1	**45.7**	54.5	54.8	**59.0**	58.4	53.5	61.4	58.8	**61.9**
tel_Telu	9.3	28.7	27.7	**34.2**	8.7	30.6	27.5	**34.9**	19.7	40.9	38.7	**41.4**
tgk_Cyrl	**41.0**	35.2	37.3	34.6	47.0	**57.5**	56.4	52.7	49.6	59.0	56.2	**60.9**
tgl_Latn	72.3	**74.1**	74.0	73.3	76.2	76.4	75.5	**76.7**	**78.1**	76.6	77.9	76.8
tha_Thai	1.7	**1.9**	1.5	1.6	**1.0**	0.9	0.8	0.8	1.4	0.7	1.2	**1.5**
tuk_Latn	49.2	**52.6**	52.5	50.8	56.9	**59.8**	57.3	58.3	58.2	**59.3**	57.9	57.1
tur_Latn	66.2	68.4	69.6	**69.7**	69.2	71.6	70.9	**72.7**	69.8	**75.7**	74.8	73.5
uig_Arab	11.2	22.6	24.7	**26.2**	13.3	39.3	35.7	**39.6**	15.9	**42.4**	38.3	35.1
ukr_Cyrl	59.2	66.4	66.4	**67.1**	65.5	66.3	71.7	**73.1**	71.4	74.3	**74.7**	72.1
urd_Arab	17.3	20.0	21.8	**23.7**	12.5	25.3	**28.5**	**28.5**	17.3	44.5	**46.2**	41.9
uzb_Latn	**67.9**	67.7	66.9	66.3	**76.8**	72.7	73.5	76.0	73.8	75.3	**76.3**	74.9
vec_Latn	59.5	**62.3**	61.9	59.5	68.5	68.3	**71.1**	68.5	73.6	73.4	**74.4**	72.4
vep_Latn	57.6	54.7	60.4	**62.9**	66.7	**68.0**	64.0	63.2	**70.7**	69.7	68.5	68.3
vie_Latn	48.4	48.6	49.5	**50.4**	47.7	51.1	**52.9**	**52.9**	50.7	52.5	53.0	**54.0**
vls_Latn	**73.6**	68.6	72.7	71.0	72.5	**76.1**	73.6	72.9	75.8	**78.5**	73.0	74.9
vol_Latn	58.1	57.7	**58.7**	57.7	58.1	**60.0**	**60.0**	**60.0**	59.0	58.0	59.0	**60.0**
war_Latn	**62.8**	62.3	58.7	60.4	65.2	66.4	66.4	**67.0**	**70.0**	65.8	**70.0**	67.0
wuu_Hani	17.1	28.9	27.8	**31.6**	25.6	34.6	34.9	**35.7**	20.3	31.1	30.8	**33.8**
xmf_Geor	13.5	**22.2**	20.5	18.6	25.2	35.3	**37.2**	30.2	22.1	33.7	33.2	**36.4**
yid_Hebr	10.2	18.8	26.5	**31.0**	11.8	24.9	32.2	**34.1**	15.9	26.7	28.6	**39.2**
yor_Latn	33.6	34.5	**36.2**	32.9	63.8	**64.9**	61.3	61.7	62.8	**66.4**	59.6	62.5
yue_Hani	10.1	11.3	8.6	**11.4**	**11.8**	9.7	9.2	11.2	11.1	**11.8**	10.5	10.9
zea_Latn	65.7	68.3	**70.2**	70.0	65.3	69.9	69.7	**71.1**	69.2	**73.2**	66.2	72.0
zho_Hani	10.5	11.6	9.7	**12.7**	**12.9**	10.0	9.9	11.6	12.2	13.2	**13.4**	12.2

Table 23: F1 scores of models on **transliterated** dataset of **NER** (Part II).

Language	XLM-R	XLM-R (Min-Merge)	XLM-R (Average-Merge)	XLM-R (Max-Merge)	Glot500	Glot500 (Min-Merge)	Glot500 (Average-Merge)	Glot500 (Max-Merge)	FURINA	FURINA (Min-Merge)	FURINA (Average-Merge)	FURINA (Max-Merge)
afr_Latn	**88.8**	87.4	87.3	87.1	**87.8**	86.4	86.4	86.5	**87.7**	87.5	**87.7**	87.6
ajp_Arab	28.2	43.5	**43.7**	43.0	28.5	45.7	**49.9**	49.5	39.6	49.3	**54.0**	52.8
aln_Latn	**51.6**	50.3	50.2	49.9	**49.5**	47.9	47.3	49.0	**55.0**	53.5	52.0	52.7
amh_Ethi	31.7	48.1	49.5	**50.0**	32.8	46.2	47.8	**51.0**	35.7	48.9	**50.4**	49.8
ara_Arab	18.5	48.9	51.3	**53.9**	23.4	45.4	51.2	**51.6**	38.8	52.7	**56.7**	**56.7**
bam_Latn	25.7	**27.9**	27.6	27.3	38.4	43.7	**43.9**	42.8	52.4	**53.6**	53.0	52.0
bel_Cyrl	46.6	79.2	**79.4**	**79.4**	53.5	76.6	77.8	**78.4**	74.6	81.7	**82.2**	82.0
ben_Beng	37.6	62.7	**66.2**	65.4	41.1	62.4	**67.0**	65.1	63.8	67.5	**70.7**	69.1
bre_Latn	49.2	**51.0**	50.3	50.4	52.5	**54.1**	53.5	53.7	**60.7**	60.5	60.1	59.6
bul_Cyrl	60.9	**80.9**	79.9	**80.9**	68.1	76.8	78.5	**79.1**	82.6	84.5	**84.8**	**84.8**
cat_Latn	**87.0**	85.5	84.1	84.7	**85.2**	83.9	83.3	84.3	**85.9**	85.4	85.5	85.4
ceb_Latn	48.5	**49.1**	49.0	48.0	**67.1**	65.8	65.5	65.5	**69.8**	67.9	69.2	69.0
ces_Latn	80.1	82.0	**82.6**	82.1	74.2	77.3	**78.1**	78.0	79.0	80.9	**81.1**	**81.1**
cym_Latn	**65.3**	61.2	60.9	61.5	**64.4**	60.8	61.7	61.2	**65.2**	62.3	62.5	61.2
dan_Latn	89.8	**90.7**	90.6	90.5	88.5	89.4	89.4	**89.7**	90.2	90.8	**90.9**	**90.9**
deu_Latn	87.7	**88.5**	88.4	88.3	87.1	**88.1**	87.8	87.8	87.3	**88.7**	**88.7**	88.5
ell_Grek	23.5	58.9	58.2	**60.3**	31.4	57.2	56.9	**59.7**	59.6	71.1	72.5	**72.8**
eng_Latn	**96.2**	**96.2**	**96.2**	**96.2**	**96.1**	96.0	96.0	96.0	**96.0**	**96.0**	**96.0**	**96.0**
est_Latn	80.6	84.6	84.5	**84.9**	76.1	79.5	**80.0**	**80.0**	80.2	82.2	82.2	**82.4**
eus_Latn	70.8	**71.1**	**71.1**	70.3	61.1	60.5	**61.2**	60.5	61.9	**70.4**	68.0	**70.4**
fao_Latn	63.7	**73.2**	72.4	72.1	77.7	84.0	**85.3**	**85.3**	86.3	87.4	**88.5**	88.3
fas_Arab	18.6	46.8	49.7	**51.8**	26.6	43.8	46.3	**46.9**	60.0	64.5	**66.0**	65.1
fin_Latn	78.2	84.0	83.9	**84.1**	69.8	77.8	**79.0**	78.9	78.8	80.2	**80.6**	80.5
fra_Latn	**84.9**	84.2	83.5	84.2	**84.5**	83.4	82.9	83.8	**86.0**	85.2	85.2	85.0
gla_Latn	54.8	**55.0**	54.5	54.5	**57.3**	56.9	56.8	56.8	**59.2**	57.6	58.7	57.4
gle_Latn	58.0	60.8	60.9	**61.4**	59.4	**60.4**	60.3	59.6	**64.5**	63.5	64.1	63.3
glg_Latn	**83.2**	82.8	81.5	81.5	80.2	81.0	82.1	**82.8**	**83.2**	82.5	82.5	82.6
glv_Latn	25.7	28.2	28.1	**29.1**	**52.8**	50.4	49.5	49.4	**53.9**	48.3	50.5	47.5
grc_Grek	14.3	**26.3**	24.7	24.6	22.0	29.9	31.2	**33.8**	37.2	43.7	**44.9**	43.7
grn_Latn	7.4	**9.7**	7.4	6.9	17.1	16.7	**21.6**	21.5	**25.6**	18.9	23.6	24.6
gsw_Latn	45.1	**52.0**	49.9	50.6	75.0	78.5	**79.1**	**79.1**	77.1	**81.6**	80.6	80.8
hbo_Hebr	27.4	**28.0**	27.6	27.5	29.4	35.6	37.3	**38.0**	27.6	35.3	**38.6**	37.1
heb_Hebr	30.9	56.8	62.4	**63.0**	30.0	49.5	57.8	**59.3**	35.9	51.0	58.5	**58.7**
hin_Deva	37.7	52.6	53.9	**56.8**	49.2	56.7	59.7	**59.8**	67.7	67.4	68.6	**69.3**
hrv_Latn	85.3	**85.6**	**85.6**	85.5	84.8	84.6	**84.9**	84.6	83.6	**85.1**	**85.1**	84.8
hsb_Latn	68.1	**69.9**	69.7	69.7	75.5	76.0	**76.6**	76.5	79.4	**80.0**	**80.0**	79.5
hun_Latn	77.7	80.9	**81.4**	81.2	66.2	76.2	76.8	**77.2**	76.5	79.3	**79.8**	79.7
hye_Armn	32.4	55.6	55.4	**56.0**	56.1	63.6	65.8	**66.1**	62.4	68.8	69.0	**70.0**
hyw_Armn	27.4	44.4	46.7	**48.3**	41.8	53.0	53.4	**54.2**	47.4	56.3	55.7	**57.9**
ind_Latn	83.9	83.8	**84.0**	**84.0**	83.5	83.4	**83.7**	**83.7**	83.1	**83.4**	**83.4**	83.2
isl_Latn	59.3	**74.8**	**74.8**	74.7	60.3	71.4	**72.4**	72.0	76.9	79.7	**80.2**	80.0
ita_Latn	**87.6**	87.5	86.9	86.8	**87.7**	85.5	86.5	86.8	**88.3**	87.0	87.4	87.6
jav_Latn	**73.3**	72.6	72.5	71.6	72.9	74.4	**74.9**	74.5	**75.0**	73.6	74.6	74.9
jpn_Jpan	23.5	30.8	29.0	**32.8**	23.0	29.8	**31.9**	31.7	31.2	**32.0**	31.3	31.6
kaz_Cyrl	38.5	60.7	61.9	**62.3**	55.1	64.9	67.2	**67.3**	66.5	69.9	**71.2**	**71.2**
kmr_Latn	46.2	54.6	54.5	**54.7**	57.3	58.4	59.2	**59.8**	64.4	**66.7**	66.3	66.1
kor_Hang	23.0	44.3	44.8	**45.5**	24.2	41.5	42.5	**43.3**	26.6	43.6	44.3	**44.7**
lat_Latn	**75.1**	74.6	74.4	74.0	71.2	71.0	71.1	**71.4**	72.4	**73.1**	71.9	**73.1**
lav_Latn	68.6	**78.6**	78.4	78.4	64.1	73.2	73.3	**73.4**	75.4	**78.3**	78.1	77.9
lij_Latn	45.1	**52.1**	50.7	49.8	**69.3**	67.3	66.9	67.6	**74.2**	71.2	71.6	71.4
lit_Latn	78.6	**80.6**	80.5	**80.6**	69.2	**73.7**	72.9	73.6	77.8	**79.0**	78.8	78.9
lzh_Hani	5.1	5.7	5.9	**7.3**	5.8	6.0	6.4	**7.9**	7.6	7.7	7.2	**8.1**
mal_Mlym	33.7	75.9	73.7	**76.2**	36.8	69.1	70.1	**73.5**	65.1	**77.3**	76.9	77.0
mar_Deva	32.4	64.6	66.3	**68.3**	34.6	60.2	**64.2**	63.8	54.4	71.3	70.5	**73.4**
mlt_Latn	20.8	**22.5**	21.5	21.6	73.3	74.9	**75.0**	74.8	**78.6**	78.5	78.0	78.2
myv_Cyrl	36.0	**37.6**	36.5	37.2	36.5	40.9	41.5	**42.0**	39.1	43.1	43.7	**43.9**
nap_Latn	**70.6**	58.8	66.7	66.7	**88.9**	77.8	50.0	50.0	**88.9**	66.7	66.7	66.7
nds_Latn	57.8	**61.4**	60.4	60.3	75.4	**77.1**	74.9	75.5	77.1	**77.9**	77.4	76.9
nld_Latn	88.2	88.3	88.3	**88.4**	88.3	**88.4**	88.3	88.1	**88.6**	88.1	88.0	87.9
nor_Latn	86.2	86.9	86.5	**87.0**	85.5	**86.7**	86.4	86.6	87.3	**87.7**	87.3	87.4
pcm_Latn	46.4	48.3	48.1	**48.4**	57.8	**58.2**	57.8	57.9	58.6	**59.3**	58.7	59.0
pol_Latn	82.8	**84.1**	83.8	83.9	79.7	80.4	80.5	**80.8**	81.7	82.2	**82.3**	82.2
por_Latn	**88.3**	**88.3**	87.2	87.4	85.1	86.1	86.5	**87.0**	88.1	**88.3**	**88.3**	**88.3**
quc_Latn	27.6	30.3	28.9	**30.7**	**60.5**	52.5	51.9	55.5	**63.7**	53.7	55.4	53.4
ron_Latn	78.4	**79.0**	78.1	78.5	75.8	75.0	**76.3**	76.0	78.0	77.8	**78.1**	77.4
rus_Cyrl	65.3	84.0	84.1	**84.6**	68.6	81.4	82.0	**82.7**	83.7	**86.2**	**86.2**	85.9
sah_Cyrl	19.2	**21.0**	19.0	19.8	23.7	41.4	42.3	**44.8**	26.7	43.2	**45.4**	44.7
san_Deva	4.8	**11.3**	9.6	10.6	14.5	12.7	13.9	**17.0**	16.6	**19.0**	17.1	18.8
sin_Sinh	20.4	42.5	44.6	**45.9**	23.2	36.9	**41.4**	40.7	24.0	41.2	43.1	**44.0**
slk_Latn	83.2	84.1	**84.9**	84.4	80.4	80.9	**81.5**	**81.5**	82.5	**83.5**	83.4	**83.5**
slv_Latn	76.5	77.0	**77.6**	77.0	74.0	74.0	**74.2**	**74.2**	74.2	**75.1**	74.9	74.7
sme_Latn	29.6	**33.0**	31.7	32.3	64.1	68.5	**69.2**	68.2	64.9	69.6	**69.9**	**69.9**
spa_Latn	**88.1**	87.6	87.2	87.0	87.5	86.9	87.3	**87.6**	**88.1**	87.8	87.9	87.9
sqi_Latn	74.7	76.7	**77.4**	77.0	77.1	76.6	76.5	**77.6**	75.8	76.4	**77.0**	75.8
srp_Latn	86.0	86.1	**86.6**	86.5	84.7	84.3	**85.0**	84.6	83.4	85.1	84.8	**85.2**
swe_Latn	89.2	93.0	93.0	**93.2**	84.8	89.9	90.7	**90.8**	90.7	91.9	**92.2**	92.0
tam_Taml	32.1	56.4	**61.6**	61.0	35.4	57.0	59.4	**61.4**	48.4	58.3	61.5	**61.6**
tat_Cyrl	35.3	40.7	**41.0**	40.2	46.7	61.6	**62.6**	62.4	64.3	65.3	**66.2**	65.5
tel_Telu	34.9	64.0	64.1	**69.5**	36.7	58.3	60.8	**63.3**	61.0	66.4	66.4	**66.8**
tgl_Latn	72.3	**72.8**	71.4	72.1	75.7	**75.8**	75.7	75.4	**77.0**	76.2	76.3	76.5
tha_Thai	8.3	28.4	26.3	**28.7**	9.0	27.1	**30.1**	28.6	12.4	26.6	**29.2**	28.7
tur_Latn	60.2	68.3	**68.8**	68.7	60.7	64.4	**65.3**	**65.3**	68.9	69.2	**69.5**	69.2
uig_Arab	29.7	51.5	**53.7**	52.8	33.9	51.9	54.4	**55.2**	42.9	55.9	**57.9**	57.0
ukr_Cyrl	54.1	79.3	79.6	**79.8**	57.0	73.5	74.8	**75.5**	74.8	79.2	**79.8**	79.7
urd_Arab	21.4	43.1	47.9	**50.0**	21.5	37.1	41.7	**44.0**	54.5	51.1	**55.6**	**55.6**
vie_Latn	27.2	32.1	34.9	**40.2**	27.5	30.1	32.5	**33.0**	32.8	35.1	36.9	**38.5**
wol_Latn	24.6	**26.3**	25.8	26.1	48.1	**54.8**	53.5	52.8	52.8	**57.4**	55.9	55.9
xav_Latn	6.0	**12.0**	7.9	6.4	5.8	13.7	15.8	**17.7**	14.8	**20.9**	19.2	20.8
yor_Latn	22.0	22.6	22.5	**22.7**	48.2	52.4	**52.7**	51.5	59.8	59.8	59.6	**60.1**
yue_Hani	30.2	35.3	34.1	**38.1**	31.9	37.0	36.5	**38.2**	32.9	37.4	39.1	**39.6**
zho_Hani	28.5	36.9	34.8	**40.0**	29.2	39.6	40.9	**41.6**	31.5	40.4	**42.0**	**42.0**

Table 24: Top-10 accuracy of models on **transliterated** dataset of **POS**.