Just Use XML: Revisiting Joint Translation and Label Projection

Thennal D K and Chris Biemann and Hans Ole Hatzel
Language Technology Group
University of Hamburg
thennal10@gmail.com
{chris.biemann, hans.ole.hatzel}@uni-hamburg.de

Abstract

Label projection is an effective technique for cross-lingual transfer, extending span-annotated datasets from a high-resource language to low-resource ones. Most approaches perform label projection as a separate step after machine translation, and prior work that combines the two reports degraded translation quality. We re-evaluate this claim with LabelPigeon, a novel framework that jointly performs translation and label projection via XML tags. We design a direct evaluation scheme for label projection, and find that LabelPigeon outperforms baselines and actively improves translation quality in 11 languages. We further assess translation quality across 203 languages and varying annotation complexity, finding consistent improvement attributed to additional fine-tuning. Finally, across 27 languages and three downstream tasks, we report substantial gains in cross-lingual transfer over comparable work, up to +40.2 F1 on NER. Overall, our results demonstrate that XML-tagged label projection provides effective and efficient label transfer without compromising translation quality.[1]

1 Introduction

Many NLP tasks depend on span-level labels, such as entities in named entity recognition, arguments in event extraction, or mentions in coreference resolution (Liu et al., 2022). Although recent advances in generative large language models showcase strong zero-shot potential, supervised training on task-specific data continues to achieve substantially superior performance in a multilingual setting (Wei et al., 2024; Porada et al., 2024; Lu et al., 2025; Bucher and Martini, 2024). A common paradigm for extending these tasks beyond high-resource languages like English is the use of automatic machine translation to translate training data into the target language. This involves *label projection*, techniques to preserve or subsequently map the span labels onto the translated text (Chen et al., 2023; Ebing and Glavaš, 2025).

Label projection has been traditionally conducted as a separate step from translation, largely with the use of word alignment models (Akbik et al., 2015; Aminian et al., 2017; Ebing and Glavaš, 2025). More recently, Chen et al. (2023) investigate joint translation and label projection in one step, inserting square brackets around spans before translation. They report improved downstream performance but degraded translation quality. Subsequent work in the field builds on this finding, separating the translation and label projection steps and applying other techniques such as LLM-based contextual translation or constrained decoding on the unmodified translation (Parekh et al., 2024; García-Ferrero et al., 2023; Le et al., 2024). While effective, these pipelines introduce considerable computational and engineering overhead.

In this work, we revisit the core assumption motivating these methods, that translation quality is inherently compromised when markers are inserted into the text. We show that with the appropriate training, data, and choice of marker, translation quality can be *improved* while simultaneously transferring labeled spans.

To this end, we make both a practical and theoretical case for *label-aware translation* with XML tags (§3), and introduce **LabelPigeon**, a simple approach for joint label projection and translation based on fine-tuning with XML-tagged corpora (§4). LabelPigeon conducts both tasks in one pass, handling frequent and nested spans with grace, as we showcase in Figure 1.

We assess LabelPigeon through three distinct evaluations. We introduce a novel scheme for direct label projection evaluation, verifying LabelPigeon's effectiveness in 11 languages (§5). We further quantify the impact on translation quality across 203 languages as well as varying annotation

[1] Our code and data are available at: https://github.com/thennal10/LabelPigeon

Findings of the Association for Computational Linguistics: ACL 2026, pages 34461–34478

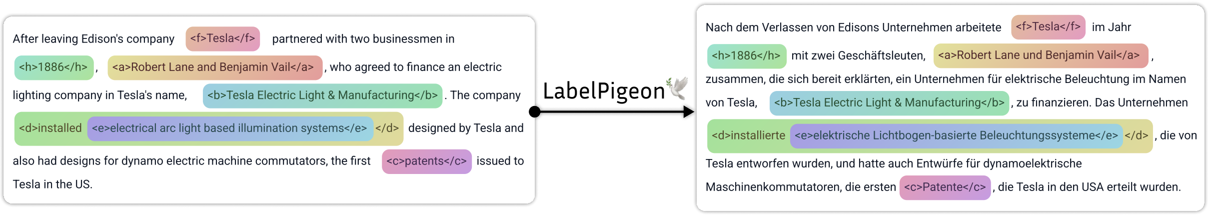

Figure 1: An example taken from XQuAD (Artetxe et al., 2020), where LabelPigeon accurately and seamlessly handles translating English to German while transferring 7 labeled spans with nesting.

complexity, finding consistent improvement which we attribute to the additional fine-tuning (§6). Finally, we conduct downstream experiments on 3 NLP tasks across 27 languages, showcasing that LabelPigeon consistently outperforms prior work, with up to +40.2 F1 score improvement (§7). Overall, our results indicate that XML tags facilitate effective label projection without compromising translation quality and with no additional computation required at inference, offering a simple alternative to multi-stage pipelines.

2 Related Work

Several works have explored markup translation in the context of structured-document translation, specifically web pages (Bamman et al., 2010; Joanis et al., 2013; Müller, 2017; Hanneman and Dinu, 2020; Hashimoto et al., 2019). Most rely on a *detag-and-project* approach, where tags are removed, the text translated, and the tags are reinserted (Hanneman and Dinu, 2020). More recent work investigates the zero-shot capabilities of massively multilingual translation models or large language models (LLMs) on transferring tags, and finds they perform adequately even without any specific fine-tuning (Dabre, 2022; Dabre et al., 2023; Buschbeck et al., 2022). While some works directly train on raw markup data, they exclusively evaluate in the context of structured document translation, largely with translation quality metrics (Hanneman and Dinu, 2020; Hashimoto et al., 2019).

Label projection, while sharing structural similarities to markup translation, is largely concerned with transferring annotated span labels for various downstream tasks (Chen et al., 2023; Ebing and Glavaš, 2024). Alignment-based projection has been widely adopted and is used in projecting data for named entity recognition (Ni et al., 2017), question answering (Hu et al., 2020; Lewis et al., 2020), event argument extraction (Lou et al., 2022), coreference resolution (Bitew et al., 2021), and semantic structure (Moradshahi et al., 2020; Daza and Frank, 2020). While a subset of prior work utilizes marker-based translation in corpora-building, Chen et al. (2023) are the first to analyze it in depth. They evaluate several marker types in their preliminary zero-shot study, and introduce EasyProject, utilizing synthetically generated data to train a translation model capable of squarebracket-based marker projection. Their results indicate a consistent degradation in translation quality, informing later works that opt to separate translation and label projection. T-Projection (García-Ferrero et al., 2023) uses a separate language model to project labels by generating candidate spans, while CLaP (Parekh et al., 2024) employs a similar approach with an instruction-tuned LLM as a contextual translator. Explicitly motivated by preserving translation quality, CODEC (Le et al., 2024) uses constrained decoding to inject square markers after translation. Ebing and Glavaš (2025) further find that word alignment can perform comparably to marker-based label projection with specific low-level design decisions, reinforcing the paradigm of separate label projection.

Taken together, prior work leaves several aspects unexplored. With the exception of Chen et al. (2023), no other paper investigates joint translation and label projection. Evaluations rely on indirect metrics such as projection rates or downstream task performance, with no work directly evaluating label projection. In addition, little attention is paid to more challenging cases with frequent, nested, or overlapping spans. Finally, most approaches forgo training altogether or rely on synthetically generated data, leaving existing high-quality data from the field of markup translation underutilized.

3 Label-Aware Translation

Prior work assumes that span markers inherently harm translation quality, and therefore designs techniques to project labels on unmarked translations

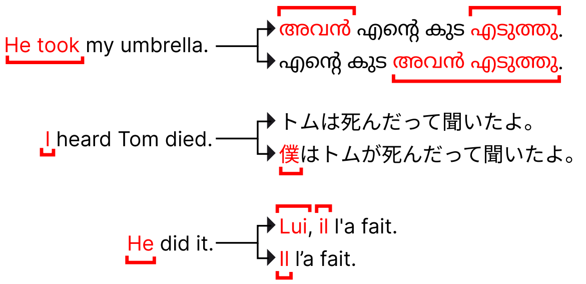

Figure 2: Examples of labeled English sentences with two equally valid translations, where the labeled span is preserved in one and split, omitted, or ambiguous in the other.

(Le et al., 2024; García-Ferrero et al., 2023; Parekh et al., 2024). However, we posit that with appropriate training, a label-aware translation is advantageous in several respects.

Figure 2 provides minimal illustrative examples for this purpose. The first example showcases an English sentence that has a labeled span, and two equally valid translations in Malayalam, but one translation preserves the span while the other splits it across the sentence. While splitting the label is not necessarily detrimental, marker-based label projection methods do not have the capability to do so (Chen et al., 2023; Parekh et al., 2024; Le et al., 2024; García-Ferrero et al., 2023), and keeping labeled spans continuous is considered best practice for alignment-based methods as well (Ebing and Glavaš, 2025). Similarly, the second example showcases two translations into Japanese, one of which—in an instance of pronoun dropping—omits the label while the other does not. As a highly contextual language, the first translation is generally considered more natural, but the second is also valid, and in our case, preferable. Finally, the third example showcases two translations into French that, depending on the context, can be equally valid. In the first example, the labeled span can arguably be ambiguously assigned to two potential spans: the subject pronoun ("il") or the stress pronoun ("Lui"). The other, more direct translation is again preferable.

These examples showcase several potential issues that may arise when translation is done independently of label projection. We hypothesize that joint translation and label transfer would incentivize the model to prioritize the coherence and continuity of the labels. On the other hand, as illustrated in Figure 2, label-aware translation can also lead to less fluent translations. We argue that less idiomatic translations will typically not lead to substantial annotation quality loss in a model trained on the output data.

3.1 XML as the Marker of Choice

EasyProject, the only prior method utilizing joint label projection and translation, opts for square brackets to mark the label spans, with future work following suit (Chen et al., 2023; Le et al., 2024). The authors justify this choice by conducting a preliminary zero-shot study testing out several different markers, with square brackets performing the best. However, this does not translate directly to superior performance after fine-tuning, and the use of square brackets as the marker has several downsides. Most notably, square brackets do not carry direct correspondence between the original spans and the ones in translation. They compensate for this issue with a fuzzy string matching method that translates the annotated spans individually, and matches them to the spans inside the full translation to map the correspondence. This approach is susceptible to errors (in particular when nested or overlapping spans are involved) and balloons inference time as all spans must be translated individually, on top of the text as a whole.

XML tags, on the other hand, provide a direct correspondence between the source spans and translation spans. They can also handle nesting and overlapping spans gracefully, and can even hold semantic information (e.g. <PER> denoting a person) if required. Notably, XML markup has a long history in structured-document translation, with high-quality parallel corpora containing XML-tagged text publicly available (Bamman et al., 2010; Hanneman and Dinu, 2020; Hashimoto et al., 2019). In particular, Hashimoto et al. (2019) provide the Salesforce Localization XML MT dataset, a large-scale collection of parallel sentences with naturally occurring XML tags, which can be adapted for training label-aware projection. This resource provides high-quality parallel data that enables models to learn translation while maintaining structured tags, eliminating the need for generating synthetic training data as in prior work (Chen et al., 2023).

4 LabelPigeon

Our overarching goal is to re-evaluate the assumption that joint translation and label projection inherently degrades quality. As such, we focus on the effects of direct fine-tuning with high-quality

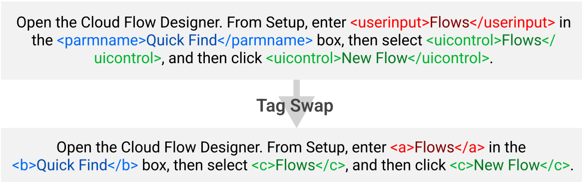

Figure 3: An example showcasing the tag swap that we conduct on training data in order to make it generally applicable.

data, and to that end, we opt for the Salesforce Localization XML MT dataset mentioned in §3.1 (Hashimoto et al., 2019). A gold-standard XML-tagged corpus, it consists of parallel pairs between English and seven other languages with approximately 100,000 aligned samples in each language pair, providing ample data for full-scale fine-tuning.

Prior research indicates that fine-tuning on too large of a dataset or on low-resource languages could lead to catastrophic forgetting and a general reduction in translation quality (Liu and Niehues, 2025; Chen et al., 2023). We further conduct ablation experiments (detailed in Appendix A.1) and find that fine-tuning on all seven language pairs is counterproductive. In accordance, we opt to train the final model with data between English and three high-resource languages: German, Russian, and Chinese.

The tags are largely composed of UI and styling elements. In order to adapt the dataset for general label projection, we opt to swap these for simple alphabetical non-descript tags of the form <a>, <b>, etc. All tags of a certain type are converted into a corresponding alphabetical tag based on the order of appearance. Figure 3 showcases an example of this in action. We also drop all examples that contain no tags, resulting in a sizeable reduction of the dataset to 25k samples in each language pair. Across all three datasets, and accounting for translation in both directions with English, this amounts to approximately 150k training samples, of which 5% is utilized as a development set.

Due to its effectiveness, coverage, and widespread use, we opt for the NLLB-200 3.3B as the base translation model to fine-tune (Team et al., 2022). We conduct fine-tuning on our modified dataset for a full epoch, totaling 9,091 steps with an effective batch size of 16, taking 5h:30m on a single NVIDIA A100 GPU. Additional training and data specifics are given in Appendix A.

With this model, label projection can be conducted in a straightforward procedure: insert alphabetical XML tags on the annotated spans, translate with our model, and extract the tags using an off-the-shelf XML parser. We term our method **LabelPigeon**, and note that it has a negligible computational overhead at inference, requiring only a single forward pass of the model.

5 Directly Evaluating Label Projection

Prior work generally evaluates label projection methods by translating span-annotated datasets and training models on those datasets, essentially using the downstream results as proxy for the efficacy of the label projection (Chen et al., 2023; García-Ferrero et al., 2023; Parekh et al., 2024; Le et al., 2024). We instead opt to define our own benchmark and metrics to directly evaluate label projection, utilizing parallel span-annotated datasets.

5.1 Experimental Setup

Datasets. For directly evaluating label projection we utilize XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020), two gold-standard multilingual extractive question-answering (QA) datasets. XQuAD consists of 240 paragraphs and 1190 QA pairs in 12 languages, with the other 11 languages translated from English, while MLQA consists of over 5,000 QA pairs in 7 languages that were mined from Wikipedia. Both datasets provide span-annotated QA data parallel across multiple languages, allowing direct measurement of how well projected spans align in the target language. Because MLQA is not consistently parallel at the paragraph level, we apply a simple filter to only retain the QA pairs with parallel contexts, detailed in Appendix B.1. Additional statistics, particularly with regards to label frequency, are provided in Appendix B.

Metrics. Across both datasets, we define a simple evaluation scheme to concretely evaluate the accuracy of label projection. Each projected label span is taken individually and is considered a match if it has string similarity above a set ratio to the corresponding reference label span, where similarity is computed with Ratcliff/Obershelp pattern matching (Black, 2004). Our main metric is the global F1 score of these label matches, which we refer to as **Label Match F1**. For all our experiments, we set the aforementioned string similarity ratio to 50%. We also calculate the COMET score (specifically COMET-22, Rei et al., 2022) to evaluate the impact

Language	COMET Score				Label Match F1 (%)			
	Awes.	Gemma	EProj.	LP	Awes.	Gemma	EProj.	LP
XQUAD								
Arabic	**83.4**	54.3	81.3	82.2	49.8	69.8	**80.9**	75.3
Chinese	**80.8**	64.9	79.4	77.9	46.4	70.6	70.9	**72.9**
German	**83.0**	73.7	81.1	82.7	60.0	86.4	84.6	**86.8**
Greek	84.2	71.4	84.3	**87.3**	44.8	71.4	65.8	**75.8**
Hindi	**78.4**	50.1	76.4	77.1	54.8	71.3	**82.6**	76.9
Romanian	84.3	83.8	83.9	**86.0**	58.0	**89.2**	81.6	87.8
Russian	83.7	79.0	82.8	**85.0**	52.5	**81.8**	76.6	78.9
Spanish	83.1	68.8	81.7	**84.0**	59.2	87.8	82.4	**90.2**
Thai	**76.8**	67.8	74.1	76.6	23.8	64.9	**66.0**	63.1
Turkish	84.0	78.9	83.0	**85.0**	58.4	**84.9**	84.0	83.3
Vietnamese	83.1	72.6	80.8	**83.3**	48.8	**80.9**	78.8	79.7
Average	82.3	69.6	80.8	**82.4**	50.6	78.1	77.7	**79.2**
MLQA								
Arabic	**84.8**	47.2	83.7	84.8	51.9	65.0	**80.7**	78.0
Chinese	**80.4**	53.7	79.1	79.6	40.6	52.9	63.9	**67.8**
German	82.4	59.9	81.4	**84.0**	60.3	77.9	77.2	**83.4**
Hindi	76.7	45.3	75.7	**76.9**	56.1	63.4	**80.1**	79.3
Spanish	82.5	59.1	82.3	**84.0**	58.6	76.3	78.9	**88.8**
Vietnamese	83.0	62.2	82.7	**84.7**	49.1	73.0	78.5	**82.3**
Average	81.6	54.6	80.8	**82.3**	52.8	68.1	76.5	**79.9**

Table 1: Direct label projection results on XQuAD and MLQA. COMET scores and the label match F1 scores are both provided. Sentences are translated from English to the corresponding language. We compare four label projection methods: a) Awesome-align (Awes.), b) Gemma 3 27B (Gemma), c) EasyProject (EProj.), and d) LabelPigeon (LP). Awesome-align is used as the baseline, and differences are highlighted via color.

on translation quality. We note that all markers are removed before evaluating translation quality.

Baselines. We compare LabelPigeon with the following baselines: (1) Awesome-align (Dou and Neubig, 2021), an alignment-based label projection method; (2) Gemma 3 27B IT (Team et al., 2025), a strong lightweight open-source LLM; and (3) EasyProject (Chen et al., 2023), a marker-based label projection method. As Awesome-align conducts label projection separately after translation, we opt for the original NLLB-200 3.3B as the corresponding translation model, providing direct comparison in terms of translation. For EasyProject, we use their fine-tuned NLLB-200 3.3B model for the same reason. Additional details can be found in Appendix B.

5.2 Results

Table 1 compiles the label projection results across languages and models, showcasing both the COMET scores and Label Match F1. We first note that our method outperforms all other baselines in label projection. Awesome-align performs particularly poorly, with an average label match F1 of 50.6/52.8 on XQuAD/MLQA. EasyProject and Gemma 3 perform reasonably well with average F1 scores of 77.7/76.5 and 78.1/68.1 respectively, but are still outperformed by LabelPigeon's 79.2/79.9. We also note that the training dataset contains a maximum of 6 unique tags per example (i.e. up to <f>). In contrast, XQuAD samples contain more than 9 tags on average, with a maximum of 24 tags (up to <x>). Given the performant results of LabelPigeon, we conclude that it is able to generalize up to much higher unique tag counts than seen during training.

The translation quality results warrant closer scrutiny. While EasyProject degrades translation quality across the board as we expect, our method *improves* translation quality over the base NLLB

Metrics	No Markers				Single		Simple		Complex	
	Baseline	EProj.	LP	NF	EProj.	LP	EProj.	LP	EProj.	LP
BLEU	17.4	17.7	17.6	17.9	16.8	17.6	15.3	16.1	14.9	15.5
chrF++	42.9	43.5	43.4	43.8	42.8	43.7	41.4	42.3	40.8	41.7
Proj. Rate	—	—	—	—	85.9	92.5	68.2	81.0	47.7	69.3

Table 2: Metrics across different models and with different marker-insertion strategies on the FLORES-200 dataset. We compare the baseline NLLB-200 3.3B model (Baseline), EasyProject (EProj.), LabelPigeon (LP), and the Non-marker Fine-tuned model (NF). BLEU and chrF++ are lexical measures for translation quality, while the projection rate measures how many of the original labels are included in the translation. Differences with respect to the baseline are highlighted via color.

model for a majority of languages. We note that this is equally true for languages that we did not fine-tune on, as well as the three that we did (German, Chinese, and Russian). We further explore the cause for this improvement in §6. Gemma 3, the only method not utilizing a base or fine-tuned version of NLLB-200 3.3B, provides markedly poorer translations than all other baselines over almost all languages, showcasing the need for translation-specific models for the task.

Additionally, we perform a small-scale error analysis on the XQuAD data for our method. We manually annotate a random sample of 30 examples translated from English to German via LabelPigeon, and compare it with the ground truth for translation errors in the labels. Out of 137 total labels, 118 were considered correct, resulting in a span translation accuracy of 86%. With respect to the automatic evaluation, we observed only three false positives and five false negatives. These findings further reinforce the effectiveness of our automated evaluation.

6 Impact on Translation Quality

While we jointly evaluate translation quality and label projection in §5, we only cover 11 languages. Additionally, the phenomenon of improved translation quality with our method warrants further investigation. In order to provide a more comprehensive overview of how markers and training impact translation quality, we opt for a broad-scale evaluation with synthetically inserted markers on the FLORES-200 dataset (Team et al., 2022).

6.1 Experimental Setup

Dataset. FLORES-200 is an extension of the well-known FLORES-101 (Goyal et al., 2022), expanding it to cover 204 languages, and it was extensively used by Team et al. (2022) to evaluate the NLLB-200 model. We use the publicly available `devtest` split containing 1012 sentences, and evaluate translation quality from English to all 203 other languages.

Synthetic Markers. As the dataset itself does not contain any sort of labeled spans, we simulate labels by randomly inserting markers on the English source sentences, in various configurations. Specifically, we utilize three marker insertion configurations representing different labelling scenarios: the **Single** configuration always inserts exactly one marker, the **Simple** configuration inserts non-overlapping and non-nested markers, and the **Complex** configuration inserts potentially overlapping and nested markers. The specific algorithm is elaborated in Appendix C.1. We also test on the original unmarked data, referred to as the **No Markers** configuration.

Metrics. Following Team et al. (2022), we opt for lexical measures of translation quality, specifically BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017). We additionally measure label projection via the projection rate, defined by Chen et al. (2023) as the percentage of data in which the numbers and type of special markers in the translations match with the source sentences. We note that this metric only takes into consideration the existence of the markers and not the accuracy of the labels themselves, and thus is significantly less reliable than our direct label projection schema in §5.

Baselines. As we are largely concerned with the impact of training on translation quality, we compare NLLB 3.3B with models derived from it, mainly LabelPigeon (LP) and EasyProject (EProj). Additionally, to disambiguate the effect of additional training with the effects of marker insertion itself, we train a model on the modified SalesForce Localization XML MT dataset as in §4, but with the XML tags removed. All other hyperparameters are kept the same, and we refer to this model as the **Non-marker Fine-tuned** (NF) model.

6.2 Results

The results are compiled in Table 2. We start by noting the improvement in translation quality of all three fine-tuned models over the baseline model when no markers are inserted. As markers are introduced, translation quality degrades for EasyProject, both compared to itself and the baseline in the No Marker configuration. However, the BLEU score remains the same, and chrF++ increases when a single marker per sentence is inserted for LabelPigeon. With multiple and nested markers, we see a clear decline in translation quality for both EasyProject and LabelPigeon. Regardless, LabelPigeon consistently outperforms EasyProject across all marker insertion configurations in both translation quality metrics. In addition, LabelPigeon attains a higher projection rate across the board, while EasyProject struggles particularly in the Complex marker insertion scheme. Taken in conjunction with the results from §5, we can confidently state that LabelPigeon improves translation quality when used on span-marked data. We provide the full results in Appendix C.2 and additional experiments on the effects of length and frequency in Appendix C.3.

Why Does Translation Quality Improve? The performance of the NF model, which has been trained on unmarked data and performs the best overall, shows that the quality improvement is a direct result of additional training. This is consistent with prior research that show fine-tuning translation models on small datasets (approx. 100K sentences) can induce positive cross-lingual transfer, improving performance for even unseen languages (Liu and Niehues, 2025). Regardless, LabelPigeon's performance with single markers is comparable to the NF model's performance under no markers, providing evidence for our hypothesis in §3: that the less idiomatic translations resulting from label-aware translation does not lead to a substantial quality loss.

Language	Dataset	EProj.	Ours
UNER (Named Entity Recognition)			
Cebuano	ceb_gja	47.6	**78.3**
Chinese	zh_gsd	**53.9**	46.4
	zh_gsdsimp	**52.9**	47.4
	zh_pud	**62.2**	54.5
Croatian	hr_set	77.4	**85.6**
Danish	da_ddt	75.5	**79.3**
German	de_pud	76.9	**80.2**
Portuguese	pt_bosque	62.2	**83.0**
	pt_pud	65.1	**86.1**
Russian	ru_pud	56.7	**70.4**
Serbian	sr_set	74.8	**87.4**
Slovak	sk_snk	64.3	**78.6**
Swedish	sv_pud	70.8	**87.7**
	sv_talbanken	67.3	**88.5**
Tagalog	tl_trg	54.1	**91.5**
	tl_ugnayan	38.5	**81.5**
Average	–	62.5	**76.7**
CorefUD (Coreference Resolution)			
Ancient Greek	PROIEL	0.0	0.0
Ancient Hebrew	PTNK	0.0	0.0
Catalan	AnCora	1.5	**12.1**
Czech	PCEDT	1.8	**20.5**
	PDT	0.5	**20.4**
French	ANCOR	0.2	**3.4**
	Democrat	0.1	**1.2**
German	ParCorFull	**18.7**	12.7
	PotsdamCC	**19.5**	16.2
Hindi	HDTB	0.0	**27.2**
Hungarian	KorKor	0.0	**3.8**
	SzegedKoref	0.0	**2.3**
Korean	ECMT	0.0	**6.3**
Lithuanian	LCC	0.0	**25.5**
Norwegian	BokmaalNARC	0.1	**31.6**
	NynorskNARC	0.2	**32.8**
Old Slavonic	PROIEL	0.4	**1.7**
Polish	PCC	4.1	**12.0**
Russian	RuCor	10.2	**32.7**
Spanish	AnCora	0.4	**10.6**
Turkish	ITCC	0.1	**12.4**
Average	–	2.7	**13.6**
MLQA (Question Answering)			
Arabic	–	62.6	**62.7**
Chinese	–	**53.5**	53.4
German	–	65.6	**67.5**
Hindi	–	**70.8**	69.7
Spanish	–	71.4	**72.2**
Vietnamese	–	**72.9**	71.5
Average	–	66.1	**66.1**

Table 3: Downstream F1 scores for UNER, CorefUD, and MLQA, comparing EasyProject (EProj.) and LabelPigeon (Ours). Differences are highlighted via color, and instances of exceptionally low scores (F1 < 1) are noted in gray.

7 Downstream Experiments

In line with recent work and prior applications, we evaluate the effectiveness of our label projection method on three downstream tasks: named entity recognition (NER), question answering (QA), and coreference resolution (CR) (Bitew et al., 2021; Chen et al., 2023).

7.1 Experimental Setup

Named Entity Recognition. To evaluate NER, we opt for Universal Named Entity Recognition (UNER), a recently released gold-standard benchmark containing 19 datasets across 13 diverse languages (Mayhew et al., 2024). We use the training split of the English portion of the dataset (i.e., the EWT dataset) as the source for cross-lingual transfer. In line with their baseline, we train XLM-$\mathrm{R}_{\mathrm{Large}}$ (560M parameters) on the translated data, and use the test splits of all other languages and corresponding datasets for evaluation (Conneau et al., 2020).

Question Answering. As in §5.1, we use MLQA (Lewis et al., 2020) for question-answering evaluation. Due to the comparatively smaller number of evaluation samples, we omit XQuAD for a downstream comparison. We use SQuAD v1.1 (Rajpurkar et al., 2016) as the source dataset, and again opt for XLM-$\mathrm{R}_{\mathrm{Large}}$ as it is the best-performing baseline on MLQA, with F1 scores as the metric.

Coreference Resolution. For coreference resolution, we use the publicly available version of the widely known CorefUD 1.3 dataset, covering 24 datasets in 17 languages (Nedoluzhko et al., 2022; Novák et al., 2025). While Nedoluzhko et al. (2022) provide no baseline, the CRAC shared task for multilingual coreference resolution utilizes multilingual $\mathrm{BERT}_{\mathrm{base}}$ (Devlin et al., 2019) as their baseline, which we adopt (Pražák et al., 2021; Novák et al., 2024). We use the English portion of OntoNotes 5.0 (Weischedel, Ralph et al., 2013, also part of CorefUD) as the source dataset to translate, and evaluate on all other languages and corresponding datasets, of which there are 21.

Baselines. Given the subpar performance of Gemma 3 and Awesome-align in our direct label projection results (§5.2), and the compute costs associated with translating large datasets, we opt to only evaluate EasyProject and LabelPigeon for the downstream experiments. Training hyperparameters and other specifics are provided in Appendix D.

7.2 Results

The results on each of the component tasks and datasets are compiled in Table 3. Through all three tasks, LabelPigeon outperforms EasyProject in the majority of datasets. For NER, we see large and consistent gains across most languages, with an average improvement of $+14.2$ and particularly significant improvements in low-resource languages such as Cebuano ($+30.7$) and Tagalog ($+40.2$). LabelPigeon also generally provides strong downstream performance with F1 scores above 80 in the majority of datasets.

In contrast, performance remains low across the board for coreference resolution. We hypothesize that this is due to a combination of two factors: the inherent frequency and nesting of coreference spans, which we have shown to reduce translation quality and label projection accuracy (§6), and the overall difficulty of the task, as reflected in the low average baseline score of 54.75 even with high-quality in-domain training data (Novák et al., 2024). For certain languages such as Ancient Greek and Ancient Hebrew, the downstream model fails to annotate at all, resulting in scores of 0.0. However, this phenomenon of complete failure occurs much more frequently for EasyProject than for LabelPigeon. Across the 16 languages tested, EasyProject yields scores < 1.0 in 11 of them, while LabelPigeon only fails by this criterion for the two aforementioned historical languages. The only languages where EasyProject obtains somewhat functional results are German with 19.1 and Russian with 10.2. In contrast, LabelPigeon achieves > 10.0 for 10 languages, and > 20.0 for 5.

Finally, for question answering, we observe only a narrow gap between LabelPigeon at 66.15 and EasyProject at 66.13, with both methods performing comparably well. Nevertheless, LabelPigeon still outperforms EasyProject across all three tasks on average, consistent with the results from the direct evaluation in §5.

8 Conclusion

In this work, we present the case for joint label projection and translation with XML tags as the marker of choice. Through comprehensive evaluations covering direct label projection, translation quality, and downstream effectiveness, we show that our method outperforms existing marker-based and alignment-based methods without incurring engineering overhead or additional computation at inference. In the broader context of a field that has largely abandoned this approach in favor of complex multi-stage pipelines, our work shows that a straightforward training regimen and high-quality data can provide effective label projection without harming translation quality.

Limitations

The direct label projection evaluation as detailed in §5 utilizes XQuAD, where all samples are translated from English, and a filtered version of MLQA, where the filtering may bias it towards direct translations. We also use FLORES-200, another directly translated dataset, in §6 to evaluate translation quality. As such, these evaluations may be affected by the phenomenon of *translationese*, where human-translated text can contain unusual features not present in natural text (Graham et al., 2020; Baker, 1993).

While we use three different tasks for downstream evaluation, we only use question answering datasets for the direct label evaluation, largely composed of high-resource languages. However, due to the requirements of such an evaluation (namely needing to be multilingual, parallel, and span-annotated), very few datasets are fit for this purpose. Our synthetic tag insertion in §6 may also not accurately reflect real-world usage, as tags are typically motivated by semantics or linguistics. Regardless, our results on it are consistent with the translation quality improvements observed in our direct label evaluation.

Finally, we do not conduct a full evaluation of the newest label projection methods such as CODEC (Le et al., 2024) and CLaP (Parekh et al., 2024). In preliminary experiments, we found that CODEC was outperformed by LabelPigeon, and a full evaluation was prohibitively expensive, as we describe in Appendix B.3. We make the case for and focus on label-aware translation, and given the extensive engineering and additional inference requirements of these methods, we leave their exploration to future work.

Ethical Considerations

Label projection has the potential to bring higher-quality labels to low resource languages. While this is generally a worthwhile pursuit, one might argue that culturally sensitive annotations that cover specific linguistic phenomena are disincentivized by better label projection. We argue that these risks are outweighed by the benefit of more accessible NLP for lower resource languages. Overall, we do not anticipate major ethical concerns arising from this work.

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2017. Transferring Semantic Roles Using Translation and Syntactic Information. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 13–19, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Mona Baker. 1993. Corpus Linguistics and Translation Studies — Implications and Applications. In *Text and Technology*, pages 233–250, Amsterdam. John Benjamins Publishing Company.

David Bamman, Alison Babeu, and Gregory Crane. 2010. Transferring structural markup across translations using multilingual alignment and projection. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 11–20, New York, NY, USA. Association for Computing Machinery.

Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Lazy Low-Resource Coreference Resolution: A Study on Leveraging Black-Box Translation Tools. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 57–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul E Black. 2004. Ratcliff/obershelp pattern recognition. *Dictionary of algorithms and data structures*, 17.

Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *Preprint*, arXiv:2406.08660.

Bianka Buschbeck, Raj Dabre, Miriam Exel, Matthias Huck, Patrick Huy, Raphael Rubino, and Hideki Tanaka. 2022. A Multilingual Multiway Evaluation Data Set for Structured Document Translation of Asian Languages. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 237–245, Online only. Association for Computational Linguistics.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly Easy Label Projection for Cross-lingual Transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. 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.

Raj Dabre. 2022. NICT's Submission to the WAT 2022 Structured Document Translation Task. In *Proceedings of the 9th Workshop on Asian Translation*, pages 64–67, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Raj Dabre, Bianka Buschbeck, Miriam Exel, and Hideki Tanaka. 2023. A Study on the Effectiveness of Large Language Models for Translation with Markup. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 148–159, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Angel Daza and Anette Frank. 2020. X-SRL: A Parallel Cross-Lingual Semantic Role Labeling Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Benedikt Ebing and Goran Glavaš. 2024. To Translate or Not to Translate: A Systematic Investigation of Translation-Based Cross-Lingual Transfer to Low-Resource Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5325–5344, Mexico City, Mexico. Association for Computational Linguistics.

Benedikt Ebing and Goran Glavaš. 2025. The Devil Is in the Word Alignment Details: On Translation-Based Cross-Lingual Transfer for Token Classification Tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18111–18128, Vienna, Austria. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2023. T-Projection: High Quality Annotation Projection for Sequence Labeling Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15203–15217, Singapore. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical Power and Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Greg Hanneman and Georgiana Dinu. 2020. How Should Markup Tags Be Translated? In *Proceedings of the Fifth Conference on Machine Translation*, pages 1160–1173, Online. Association for Computational Linguistics.

Kazuma Hashimoto, Raffaella Buschiazzo, James Bradbury, Teresa Marshall, Richard Socher, and Caiming Xiong. 2019. A High-Quality Multilingual Dataset for Structured Documentation Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 116–127, Florence, Italy. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multitask Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR.

Eric Joanis, Darlene Stewart, Samuel Larkin, and Roland Kuhn. 2013. Transferring markup tags in statistical machine translation: A two-stream approach. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*, Nice, France.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Conditions for Catastrophic Forgetting in Multilingual Translation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 347–359, Suzhuo, China. Association for Computational Linguistics.

Tianyu Liu, Yuchen Jiang, Ryan Cotterell, and Mrinmaya Sachan. 2022. A Structured Span Selector. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2629–2641, Seattle, WA, USA. Association for Computational Linguistics.

Chenwei Lou, Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, Weiwei Tu, and Ruifeng Xu. 2022. Translation-Based Implicit Annotation Projection for Zero-Shot Cross-Lingual Event Argument Extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2076–2081, New York, NY, USA. Association for Computing Machinery.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large Language Models Struggle in Token-Level Clinical Named Entity Recognition. *AMIA Annual Symposium Proceedings*, 2024:748–757.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing Open-Ontology QA Semantic Parsers in a Day Using Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.

Mathias Müller. 2017. Treatment of Markup in Statistical Machine Translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 36–46, Copenhagen, Denmark. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference Meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the Third Shared Task on Multilingual Coreference Resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual Label Projection for Cross-Lingual Structured Prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ian Porada, Xiyuan Zou, and Jackie Chi Kit Cheung. 2024. A Controlled Reevaluation of Coreference Resolution Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 256–263, Torino, Italia. ELRA and ICCL.

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual Coreference Resolution with Harmonized Annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123, Held Online. INCOMA Ltd.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Kangda Wei, Aayush Gautam, and Ruihong Huang. 2024. Are LLMs Good Annotators for Discourse-level Event Relation Extraction? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1–19, Miami, Florida, USA. Association for Computational Linguistics.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann. 2013. OntoNotes Release 5.0. Linguistic Data Consortium.

A Training

Setting	Value
Learning rate	1e[-3]
Batch size	8
Grad. Accumulation	2
Scheduler	Inverse square root
Weight Decay	0.01
Warmup	5% steps
Precision	bfloat16

Table 4: Relevant hyperparameters for LabelPigeon finetuning.

As described in §4, we use the Salesforce Localization XML MT dataset provided by Hashimoto et al. (2019), modified for label projection. Relevant statistics after filtering are compiled in Table 5. The model is trained with the hyperparameters given in Table 4. We note that since the original dataset includes examples with multiple instances of the same tag, the total number of tags is higher than the unique number of tags. As we filter out instances without tags, the minimum number of tags is 1 for all training data subsets.

A.1 Ablations

The full Salesforce Localization XML MT dataset contains 7 languages with sentences parallel to English: German, Finnish, French, Japanese, Dutch, Russian, and Chinese. We conduct some basic ablations, training on translations both from and to English in the following combinations: 1) one high resource language (German), 2) three high-resource languages (German, Russian, Chinese), and 3) all seven languages. We evaluate label projection and translation quality in accordance with our methodology in §5.

The results are compiled in Table 6. We note that the model trained on three languages outperforms both the model trained on only one language and the model trained on all seven languages, in translation quality as well as label matches. We hypothesize that while the additional data helps improve the performance in the three-language model, including all seven languages induces catastrophic forgetting, thus reducing general performance. Given these results, we opt for the three-language model in all other experiments.

B Label Projection

Dataset statistics for the direct label projection evaluation datasets are given in Table 7. For Awesome-align, we use the label projection algorithm detailed by Ebing and Glavaš (2025) and $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) for word alignment itself. For Gemma 3 27B IT, we use XML tags as the markers for annotations and follow the setup of Dabre et al. (2023), using the prompt format given below, where `src_lang` is the source language, `tgt_lang` is the target language, and `src_text` is the sentence to be translated:

```
Translate the following {src_lang} source text to {tgt_lang}:\n
{src_lang}: {src_text}\n
{tgt_lang}:
```

B.1 MLQA Filtering

Due to its nature as a dataset chiefly mined from Wikipedia, MLQA requires some filtering to act as a parallel label-projection evaluation benchmark. While questions and the sentences containing answers are aligned between languages, the paragraphs themselves are not necessarily direct translations. In order to make sure we only include paragraphs that are rough translations, we keep only paragraph pairs with the same number of questions and answer spans in both languages, and we filter out paragraphs with a COMET-22 score (Rei et al., 2022) < 80. The resulting dataset statistics are compiled in Table 8. We note that for the downstream evaluation in §7, we use the full MLQA dataset as the filtering is not necessary for question-answering evaluation.

B.2 Label Projection into English

As label projection is largely applied for translating labeled data for low-resource languages, we focus on experiments with English as the source language. However, we also conduct a direct label projection experiment with English as the target language, compiled in Table 9. Here, LabelPigeon almost universally outperforms all other baselines in label matches, with EasyProject falling behind Gemma 3. Unlike our results in §5.2, Gemma 3 also provides strong translations with an average COMET score of 81.7, outperforming EasyProject with 81.1 and approaching LabelPigeon at 83.6.

B.3 Preliminary Baseline with CODEC

We did a comparison with CODEC (Le et al., 2024) in a preliminary experiment on the English-Hindi subset of XQuAD. CODEC performed with a label match F1 of 75.6, outperformed by LabelPigeon's 76.9. In addition, evaluation with CODEC took significantly and prohibitively longer than any

	de-en		ru-en		zh-en	
	Train	**Valid**	**Train**	**Valid**	**Train**	**Valid**
Samples (N)	24311	1262	24243	1301	24173	1248
Total Tags	41569	2179	41542	2250	41881	2122
Max Tags / Example	50	9	50	14	50	13
Max Unique Tags / Example	6	5	6	5	6	5
Avg. # Tags / Example	1.71	1.73	1.71	1.73	1.73	1.70

Table 5: Statistics for our training data.

other tested method, roughly 38 minutes per sample. Given these results, we opted not to conduct a full-scale evaluation. In general, replicating the other label projection systems mentioned in §2 is challenging from an implementation standpoint, and their application is computationally expensive.

C Translation Quality

In §6, we synthetically insert markers into the FLORES-200 dataset to test the impact of our method on translation quality. Expanded results and additional details are provided below.

C.1 Synthetic Marker Insertion

We model this process by iterating through the word boundaries in the sentence. At each word boundary, an open marker may be placed with a probability of P_{open}, starting a new label span. If a label span has already been started, at each subsequent word boundary a close marker may be placed with a probability of P_{close}, ending the span. If any spans are open by the end of the sentence, the appropriate close markers are inserted at the end. We refer to this as the **Complex** marker insertion configuration, as nesting and overlapping spans are possible. By preventing new spans from being started if a span is already open, we disable nesting and overlapping, and we refer to this as the **Simple** configuration. To simulate datasets with exactly one labeled span per sample, we first sample a length $L \sim \text{Geom}(P_{close})$, and then select a span uniformly at random among all candidate spans of length L in the sentence. We refer to this as the **Single** configuration. In general, the P_{open} and P_{close} allow us to model the frequency of labels and their average length, respectively. These values are set to 0.2 and 0.5 for all our experiments unless specified otherwise.

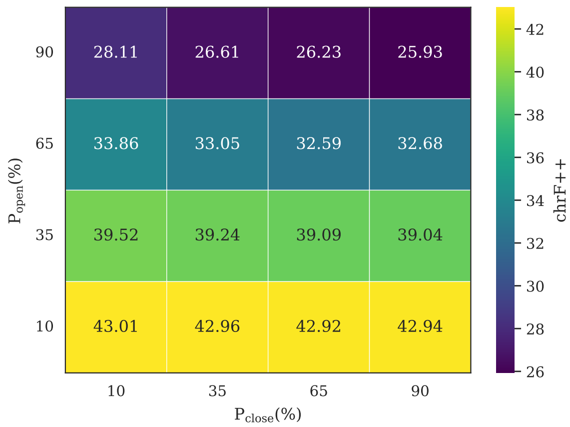

Figure 4: Translation performance of our model on FLORES-200 as measured by chrF++ across different values of P_{close} and P_{open} under the Complex marker insertion scheme.

C.2 Full FLORES-200 Results

We provide the full results of our FLORES-200 experiments in Tables 10 and 11. We note that the performance improvement of the fine-tuned models are largely consistent across all languages, the vast majority of which are unseen during fine-tuning.

C.3 Variation with Marker Frequency and Length

We also estimate the effect of the frequency and length of the marked spans on translation quality by varying P_{open} and P_{close}, specifically in the Complex marker insertion configuration. Figure 4 compiles the results as a heatmap. We see a clear degradation of quality with increasing tag frequency, but a slight and consistent improvement with increasing length (i.e., with decreasing P_{close}). Nevertheless, we note that at the lowest tag frequency (corresponding to roughly one tag once every ten words), translation quality is still improved from the baseline.

Language	COMET Score				Label Matches (F1, %)			
	Base	One	Some	All	Base	One	Some	All
XQUAD								
Arabic	79.9	80.5	**81.4**	80.2	2.3	71.9	**75.7**	74.7
Chinese	79.8	79.2	**80.3**	80.3	4.5	70.3	**76.5**	75.0
German	81.6	82.4	**83.2**	82.2	6.7	83.9	**86.2**	84.8
Greek	82.6	83.1	**84.5**	83.1	3.4	73.8	**76.6**	75.4
Hindi	80.7	80.8	**81.2**	80.7	7.7	78.4	**80.4**	79.8
Romanian	82.7	83.6	**84.5**	83.1	8.9	82.9	**85.2**	84.3
Russian	81.2	82.5	**83.3**	82.3	7.6	77.6	**79.8**	79.4
Spanish	83.4	84.1	**84.5**	83.8	3.6	86.0	**88.6**	88.3
Thai	78.3	78.0	**78.4**	77.5	7.9	64.5	**67.0**	65.9
Turkish	82.4	84.2	**84.9**	83.8	7.5	79.2	**83.1**	82.2
Vietnamese	81.9	83.2	**83.4**	82.6	6.3	78.6	**79.8**	79.7
Average	81.3	82.0	**82.7**	81.8	6.0	77.0	**79.9**	79.1
MLQA								
Arabic	83.1	84.1	**84.4**	84.1	4.3	76.3	**79.7**	78.8
Chinese	80.0	80.9	**81.2**	81.2	7.5	62.4	**70.8**	69.9
German	80.9	83.6	**83.9**	83.6	11.8	82.3	**84.5**	83.3
Hindi	81.0	**81.7**	81.7	81.4	12.7	81.5	**83.1**	82.4
Spanish	82.8	84.3	**84.5**	84.3	10.7	86.7	**88.6**	88.3
Vietnamese	82.1	84.3	**84.4**	84.2	13.9	81.7	**83.7**	83.2
Average	81.6	83.2	**83.3**	83.1	10.1	78.5	**81.7**	81.0

Table 6: Ablations on the set of languages used for training, using our direct label projection evaluation schema in §5. XML is used as the marker, with both the EN→XX and XX→EN directions evaluated and the results averaged. Base refers to the original unmodified model. We compare with models trained on three language sets: 1) one high resource language (One), 2) three high resource languages (Some), and 3) all seven languages (All). Differences with Base are highlighted in color.

	XQuAD	MLQA
Samples (N)	2539	5414
Total Tags	23764	12265
Min Tags / Example	2	2
Max Tags / Example	24	8
Avg. # Tags / Example	9.36	2.27

Table 7: Tag statistics for evaluation datasets.

	ar	de	es	hi	vi	zh
Length	843	395	1152	908	1325	791

Table 8: Data statistics after filtering in MLQA.

D Downstream Experiments

For downstream experiments, we utilize already available baselines and corresponding code, making minimal changes. For NER, we use the scripts provided by Chen et al. (2023), training 5 epochs with a batch size of 32 and a learning rate of 2e^{-3}. We average the result of five random seeds to minimize variance.

For QA, we use the scripts provided by Hu et al. (2020), training 5 epochs with a batch size of 32 and learning rate of 3e^{-3}, over three random seeds. We also discard samples from SQuAD that have more than one missing question-answer span after translation to ensure high data quality.

For coreference resolution, we use the scripts provided for the CRAC shared task (Novák et al., 2024), training 5 epochs with a batch size of 1 document and a learning rate of 2e^{-4} for task-specific parameters and 1e^{-5} for others. Additionally, to speed up the evaluation, we conduct a simple filtering step on OntoNotes, retaining documents with six sentences or fewer, in line with the default maximum sentence limit that the downstream model

Language	COMET Score				Label Matches (F1, %)			
	Awes.	Gemma	EProj.	Ours	Awes.	Gemma	EProj.	Ours
XQUAD								
Arabic	79.1	**81.6**	77.8	80.6	36.0	69.4	59.9	**76.0**
Chinese	80.3	81.9	78.7	**82.7**	45.0	64.7	67.2	**80.0**
German	82.3	**85.2**	81.3	83.8	59.8	80.3	81.0	**85.6**
Greek	80.9	**84.9**	80.3	81.7	51.9	73.4	73.2	**77.3**
Hindi	84.4	84.8	83.3	**85.3**	52.0	74.2	77.9	**83.9**
Romanian	81.1	**85.7**	81.1	83.0	56.5	**85.6**	78.6	82.6
Russian	80.2	**83.1**	79.1	81.6	50.4	72.0	75.1	**80.7**
Spanish	83.3	**85.3**	82.7	85.1	58.0	79.3	83.1	**87.1**
Thai	81.5	**82.6**	79.2	80.3	34.5	64.6	59.7	**70.9**
Turkish	82.3	**85.9**	82.1	84.8	50.7	82.0	75.0	**82.9**
Vietnamese	80.2	**84.5**	80.2	83.5	46.2	**82.2**	74.8	79.9
Average	81.4	**84.1**	80.5	82.9	49.2	75.3	73.2	**80.6**
MLQA								
Arabic	81.9	78.7	81.0	**84.0**	39.6	57.3	63.6	**81.4**
Chinese	80.3	78.3	79.7	**82.8**	37.6	48.7	62.1	**73.8**
German	81.4	77.6	80.9	**83.7**	55.2	66.2	74.3	**85.7**
Hindi	85.1	80.3	84.3	**86.4**	54.5	58.2	76.9	**86.8**
Spanish	83.3	79.7	82.8	**84.9**	51.9	59.4	79.9	**88.3**
Vietnamese	80.9	81.0	81.3	**84.2**	46.2	63.7	74.9	**85.0**
Average	82.2	79.3	81.7	**84.3**	47.5	58.9	71.9	**83.5**

Table 9: Additional direct label projection results on XQuAD and MLQA, with sentences translated from the corresponding language to English. We compare four label projection methods: a) Awesome-align (Awes.), b) Gemma 3 27B (Gemma), c) EasyProject (EProj.), and d) LabelPigeon (LP). Awesome-align is used as the baseline, and differences are highlighted via color.

handles. We also note that the metric is specifically exact-match F1 excluding singletons.

E License

We use several datasets under various licenses in this work, which we enumerate below.

- **XQuAD**: CC BY-SA 4.0
- **MLQA**: CC BY-SA 3.0
- **FLORES-200**: CC BY-SA 4.0
- **UNER**: CC BY-SA 4.0
- **CorefUD**: License CorefUD v1.3
- **SQuAD**: CC BY-SA 4.0
- **OntoNotes**: LDC User Agreement for Non-Members

All datasets used were employed in accordance with their intended research purposes and license terms. All created artifacts are intended for research and academic dissemination consistent with these terms.

Language	BLEU										chrF++									
	No Markers				Single		Simple		Complex		No Markers				Single		Simple		Complex	
	Baseline	EProj.	LP	NF	EProj.	LP	EProj.	LP	EProj.	LP	Baseline	EProj.	LP	NF	EProj.	LP	EProj.	LP	EProj.	LP
ace_Arab	0.8	0.7	0.8	0.7	0.6	0.6	0.4	0.5	0.3	0.5	18.1	18.2	17.0	17.2	18.1	16.2	17.3	16.9	17.2	17.1
ace_Latn	9.7	9.4	9.1	9.3	8.7	8.8	7.8	7.7	6.7	7.5	37.0	37.1	36.2	36.2	36.2	35.3	33.8	33.1	32.7	32.9
acm_Arab	10.6	7.1	12.6	12.8	6.1	12.2	5.0	10.8	4.4	10.5	39.2	28.5	42.5	43.0	26.1	42.6	24.3	41.3	23.3	40.7
acq_Arab	13.6	15.2	15.2	15.8	14.6	15.1	13.1	14.0	12.4	13.0	43.2	44.7	45.2	45.5	44.3	45.0	42.8	43.7	42.2	43.0
aeb_Arab	9.7	7.7	11.3	11.8	6.8	11.2	5.7	9.7	5.0	9.7	35.4	31.2	39.0	39.3	28.5	39.2	26.7	37.9	25.5	37.6
afr_Latn	37.5	39.4	36.9	37.9	38.7	39.6	38.2	38.8	38.6	37.6	63.5	65.0	63.5	64.3	64.5	65.4	64.1	64.9	64.5	64.1
ajp_Arab	16.3	17.2	17.4	17.9	16.5	16.9	14.4	15.2	14.4	14.8	47.9	48.4	48.6	49.0	47.9	48.2	46.0	46.4	45.6	46.0
aka_Latn	9.4	9.6	9.4	9.4	9.6	9.7	8.5	9.3	8.3	9.0	33.6	33.8	34.2	33.9	33.8	33.9	32.4	33.2	31.4	32.9
als_Latn	31.3	31.0	32.0	32.0	29.8	30.9	27.1	28.2	26.0	26.7	56.9	56.8	57.6	57.9	56.1	57.1	54.3	55.0	53.2	53.7
amh_Ethi	12.0	13.5	12.6	13.4	12.2	13.9	9.8	11.9	9.2	11.0	34.9	38.7	36.1	37.9	36.8	39.9	35.0	37.4	34.5	36.4
apc_Arab	15.2	15.1	16.4	16.6	14.3	15.6	12.2	13.9	11.5	13.0	45.9	46.1	47.4	47.6	45.4	46.8	43.4	45.0	42.8	44.1
arb_Arab	26.0	26.1	27.4	28.0	24.7	27.2	20.9	24.3	20.6	22.9	53.8	54.7	54.8	55.2	53.5	54.6	50.5	52.6	50.2	51.3
arb_Latn	0.3	0.0	1.1	1.0	0.0	1.0	0.1	1.0	0.1	1.1	3.9	1.3	9.0	8.3	1.8	5.7	2.5	4.7	2.4	4.5
ars_Arab	21.1	22.2	23.3	23.7	21.6	22.4	18.9	20.0	18.0	18.9	49.5	50.4	51.6	51.8	50.1	51.0	47.7	48.8	46.6	47.9
ary_Arab	9.0	8.7	9.9	10.3	7.7	9.6	5.4	8.7	5.5	8.2	35.4	34.4	36.8	37.0	32.3	36.7	28.1	35.6	27.9	35.3
arz_Arab	14.7	14.4	14.7	14.4	12.8	13.8	10.5	12.9	10.2	11.7	43.9	43.5	43.8	43.5	41.9	43.0	38.8	41.9	38.1	40.8
asm_Beng	7.7	7.9	7.8	8.3	7.9	8.2	7.0	7.5	6.6	7.0	35.7	36.5	36.1	36.8	36.3	36.7	34.8	35.3	34.4	34.8
ast_Latn	22.0	18.8	24.4	24.9	19.2	25.0	19.5	24.0	18.1	23.0	48.6	45.0	51.7	53.3	45.0	53.8	46.2	53.5	45.5	52.9
awa_Deva	13.7	12.7	14.3	14.4	12.5	13.6	11.1	12.4	10.3	11.7	40.4	38.4	42.0	41.7	38.9	41.4	38.6	40.3	37.3	39.4
ayr_Latn	3.5	3.4	3.7	3.6	3.1	3.5	2.2	2.9	2.0	3.0	29.1	29.5	29.0	28.9	29.2	29.3	27.5	27.5	26.7	26.8
azb_Arab	1.3	1.2	1.3	1.1	1.2	1.4	1.0	1.3	1.0	1.3	23.2	23.3	23.8	23.0	23.1	24.3	21.1	23.4	20.3	23.3
azj_Latn	13.4	13.5	13.4	13.7	13.0	13.2	11.0	11.7	10.2	10.9	42.2	42.8	42.5	42.9	42.1	42.4	39.8	40.8	38.1	40.1
bak_Cyrl	17.1	18.3	16.2	17.0	17.1	17.2	14.8	14.0	13.6	12.9	45.7	47.6	44.8	45.9	46.5	46.2	44.3	42.1	43.2	40.3
bam_Latn	6.4	6.2	6.2	6.7	6.0	6.7	5.7	6.4	5.4	6.0	29.9	30.2	29.8	30.2	30.1	30.3	28.9	29.3	27.9	29.1
ban_Latn	13.8	13.6	14.4	14.7	13.1	14.3	11.9	13.6	11.0	12.5	42.2	42.5	43.0	43.5	42.0	43.8	40.2	42.7	38.5	41.8
bel_Cyrl	12.9	13.1	12.7	12.8	12.4	12.1	10.4	10.4	9.7	9.9	40.1	40.4	39.9	39.9	39.9	39.3	38.2	37.3	37.7	36.6
bem_Latn	8.8	8.8	8.6	8.5	8.8	8.9	8.5	8.9	8.0	8.8	35.4	35.7	35.0	34.8	36.2	35.3	35.8	35.5	35.2	35.3
ben_Beng	16.9	17.8	17.1	17.6	16.9	17.1	14.6	14.4	14.5	13.8	47.2	48.4	47.6	47.9	47.6	47.5	45.7	44.8	44.8	43.6
bho_Deva	16.0	15.1	15.9	15.7	15.0	15.0	13.8	12.8	13.5	12.6	41.1	40.3	41.1	41.0	40.5	40.6	39.0	38.1	38.9	37.6
bjn_Arab	1.3	0.7	0.9	0.9	0.7	0.9	0.7	0.6	0.6	0.8	19.5	19.4	17.0	17.5	19.3	16.5	18.5	17.0	18.0	16.9
bjn_Latn	18.3	18.1	16.6	17.3	17.1	17.9	14.8	16.3	14.0	15.0	47.4	47.4	46.1	47.0	46.4	47.7	43.7	45.6	43.1	44.7
bod_Tibt	0.9	0.8	0.6	0.8	0.7	0.5	0.5	0.8	0.4	0.8	27.1	27.4	26.8	26.5	27.4	26.0	25.7	25.5	25.7	25.1
bos_Latn	30.2	29.7	30.9	31.4	28.4	30.0	25.6	26.7	25.2	25.8	56.3	56.2	57.2	57.7	55.3	57.0	53.7	54.5	53.2	54.0
bug_Latn	6.2	6.1	6.3	6.6	6.1	6.2	5.7	6.0	5.2	5.5	32.7	32.9	33.1	33.1	32.8	33.0	31.5	32.4	30.6	31.8
bul_Cyrl	38.7	39.8	39.0	39.7	38.3	39.3	35.3	35.9	34.7	34.3	62.6	63.7	63.0	63.9	62.9	63.6	61.3	61.6	60.7	60.3
cat_Latn	41.5	41.6	42.2	42.1	40.3	41.2	38.0	38.4	37.4	37.4	63.4	63.7	64.3	64.4	63.1	64.0	61.9	62.2	61.3	61.5
ceb_Latn	30.0	30.5	29.7	30.0	29.8	29.4	28.3	28.1	27.5	26.4	56.7	57.4	56.7	57.0	56.8	56.9	56.0	55.4	55.2	53.6
ces_Latn	30.6	31.2	31.0	31.4	30.2	30.2	26.1	26.6	26.3	26.2	55.0	55.8	55.6	55.9	55.1	55.1	52.3	52.9	52.4	52.5
cjk_Latn	2.3	1.9	2.3	2.4	2.1	2.4	2.0	2.4	1.7	2.1	23.3	22.5	23.5	23.4	23.1	23.7	22.3	23.5	21.4	23.4
ckb_Arab	10.7	10.5	10.9	11.1	9.6	10.7	8.8	9.1	8.0	8.4	44.4	44.5	44.4	44.5	43.4	43.8	41.7	41.8	40.6	40.3
crh_Latn	13.5	14.2	14.2	14.0	13.3	13.4	11.2	11.9	11.0	10.9	42.8	43.7	43.5	43.7	43.0	43.3	40.5	41.4	39.8	40.4
cym_Latn	42.3	43.0	42.7	43.0	40.3	40.4	34.2	34.5	32.6	32.1	63.9	64.5	64.3	64.7	62.5	62.7	58.2	58.1	56.5	55.9
dan_Latn	41.9	43.9	42.9	44.2	42.1	44.4	39.2	41.5	38.8	41.0	64.6	66.2	65.6	66.5	65.3	66.7	63.6	64.8	63.2	64.3
deu_Latn	37.1	38.7	38.0	38.6	37.1	37.6	34.3	34.8	33.7	33.3	60.8	62.4	62.3	62.8	61.7	62.3	60.2	60.8	59.7	60.0
dik_Latn	3.2	3.1	3.8	3.6	3.0	3.4	3.0	3.3	2.7	3.3	22.0	22.4	23.2	23.0	22.2	22.7	21.6	22.1	20.5	21.9
dyu_Latn	1.0	1.7	1.1	1.3	1.5	1.6	1.3	1.9	1.3	1.7	14.3	16.5	15.7	16.3	16.3	18.1	16.8	18.9	16.4	18.7
dzo_Tibt	0.5	0.3	0.5	0.5	0.3	0.6	0.4	0.4	0.2	0.4	31.7	32.3	31.7	31.4	32.1	31.9	30.7	31.1	30.3	30.1
ell_Grek	26.3	26.4	27.3	27.3	25.3	26.9	23.6	24.2	22.7	23.4	50.4	50.7	51.3	51.7	50.0	51.3	48.9	49.4	48.1	48.5
epo_Latn	33.6	34.0	34.2	34.6	32.6	34.5	29.1	32.5	29.1	31.4	59.9	60.6	60.7	61.0	59.8	61.0	57.4	59.5	57.4	59.0
est_Latn	23.0	23.5	23.0	23.3	23.0	23.0	19.4	19.9	18.3	19.3	52.6	53.3	52.8	53.2	52.8	52.8	49.7	50.7	48.7	49.9
eus_Latn	14.2	15.8	16.0	16.9	15.3	17.5	13.5	14.9	13.3	14.1	46.3	49.0	48.4	50.3	48.9	51.5	47.2	49.2	47.1	48.6
ewe_Latn	11.3	11.3	11.1	11.1	10.9	11.1	9.8	10.6	8.5	10.2	37.1	37.6	37.4	37.5	37.1	37.3	35.5	35.8	33.7	35.4
fao_Latn	22.5	22.9	21.8	22.3	22.6	22.3	20.1	19.4	20.0	18.6	46.3	47.0	46.0	46.6	46.7	46.7	44.7	43.9	44.5	42.9
fij_Latn	18.6	19.3	18.2	18.7	18.9	18.7	17.5	17.0	16.0	16.6	45.2	46.2	45.4	45.7	45.8	45.7	44.0	44.0	42.8	43.0
fin_Latn	22.3	22.5	22.5	22.4	21.8	20.9	18.7	18.6	18.6	17.7	51.8	52.6	52.6	52.8	52.2	52.0	49.9	50.0	49.8	49.4
fon_Latn	2.6	2.4	2.4	2.5	2.3	2.5	1.8	2.4	1.7	2.1	18.7	17.0	17.0	16.6	17.2	16.6	16.6	17.0	15.2	16.5
fra_Latn	48.8	49.5	49.5	50.6	48.0	49.8	45.2	47.6	45.0	45.3	67.9	68.7	68.9	69.7	67.8	69.2	66.4	68.0	66.2	66.8
fur_Latn	31.9	31.9	31.9	31.8	30.2	31.0	27.1	29.0	26.1	28.1	55.1	55.3	55.3	55.4	53.9	55.1	51.0	53.8	50.4	53.2
fuv_Latn	3.4	3.0	3.5	3.3	2.8	3.3	2.7	3.3	2.5	3.3	22.6	21.9	22.7	22.5	21.5	22.7	20.9	22.2	19.9	22.2
gaz_Latn	4.5	4.5	4.6	4.9	4.4	5.0	2.7	4.3	2.1	4.2	34.9	35.6	35.6	36.1	34.8	36.4	30.0	34.6	27.2	33.9
gla_Latn	18.9	18.5	19.0	18.9	17.9	18.0	15.7	16.3	14.8	14.9	47.6	47.8	48.0	48.1	47.3	47.2	45.4	45.5	44.5	44.1
gle_Latn	28.2	28.1	28.6	28.7	26.2	27.2	22.2	23.5	21.7	22.6	52.8	53.1	53.2	53.4	51.7	52.3	48.5	49.2	47.6	48.4
glg_Latn	34.6	34.6	35.8	35.9	33.6	35.0	31.2	32.9	30.8	32.7	58.4	58.8	59.8	60.0	58.3	59.7	57.0	58.4	56.6	58.0
grn_Latn	9.5	9.4	10.0	10.1	9.5	10.1	8.7	8.8	8.1	8.8	35.6	36.1	36.2	36.5	36.0	36.6	34.7	34.7	34.4	34.2
guj_Gujr	23.3	23.7	23.2	23.6	22.4	23.2	18.6	20.1	18.1	19.4	50.9	51.9	50.8	51.5	50.8	51.7	47.9	48.9	47.0	47.8
hat_Latn	23.3	22.8	23.4	23.6	21.9	23.5	19.9	22.0	19.5	21.2	50.8	50.7	51.1	51.2	49.8	50.8	48.2	49.4	47.7	48.6
hau_Latn	26.5	27.1	26.4	26.6	26.0	25.8	22.9	23.5	22.5	22.3	51.5	52.2	51.5	51.8	51.2	51.1	48.8	49.5	48.3	48.0
heb_Hebr	29.2	28.9	28.5	28.3	26.9	26.3	22.9	23.4	21.3	22.0	55.5	55.8	55.1	55.0	54.5	53.9	51.1	51.9	50.2	51.2
hin_Deva	33.1	33.9	33.1	33.3	32.4	32.1	28.4	29.3	28.0	27.8	56.0	56.6	56.1	56.3	55.6	55.4	52.8	53.3	52.0	52.0
hne_Deva	23.8	23.4	23.7	24.0	22.0	22.4	20.0	19.8	18.7	18.5	51.5	51.5	51.8	51.8	50.4	50.4	48.2	47.7	47.5	46.6
hrv_Latn	28.5	28.7	29.2	29.8	27.6	28.3	24.3	26.0	24.3	24.9	54.3	54.9	55.4	56.0	54.2	55.3	52.4	53.6	52.1	52.8
hun_Latn	24.1	24.9	24.5	24.4	23.8	23.7	21.3	21.0	20.4	19.4	52.6	53.5	53.3	53.6	52.9	53.1	51.1	51.0	50.5	50.3
hye_Armn	17.4	18.5	17.0	17.3	17.7	17.4	15.4	15.4	15.2	14.6	48.6	50.5	48.2	49.0	49.6	49.7	47.6	47.5	47.0	46.9
ibo_Latn	16.0	16.6	15.9	16.8	15.8	17.6	15.6	16.6	15.6	15.9	40.6	41.5	40.8	41.4	40.9	42.2	40.4	41.1	40.5	40.4
ilo_Latn	24.0	24.5	24.3	24.4	24.0	24.2	22.7	23.0	22.0	21.9	51.6	52.3	52.4	52.5	52.0	52.4	51.2	51.3	50.6	50.8
ind_Latn	45.8	45.8	47.0	47.2	44.2	46.1	39.9	42.5	39.1	41.3	68.0	68.4	69.0	69.3	67.3	68.5	64.7	66.2	64.0	65.4
isl_Latn	22.2	23.1	22.7	23.3	21.8	22.7	19.2	20.4	18.8	19.4	47.1	48.1	47.7	48.2	47.2	47.8	45.2	45.8	44.4	44.9
ita_Latn	30.1	30.0	31.4	31.7	29.0	30.5	26.7	28.8	26.1	28.0	56.2	56.4	57.3	57.6	55.9	57.0	54.4	55.9	54.0	55.2
jav_Latn	27.2	28.2	27.7	28.0	27.4	27.2	24.2	24.6	23.8	24.3	53.5	54.5	54.1	54.4	53.8	53.8	51.4	51.8	50.8	51.2
jpn_Jpan	0.3	0.3	0.2	0.2	0.2	0.4	0.3	0.3	0.3	0.3	23.6	24.8	23.7	23.9	24.0	24.1	23.0	23.5	23.3	22.7
kab_Latn	7.6	7.3	7.1	7.0	7.2	7.0	6.5	6.4	5.7	6.3	31.0	31.2	30.7	30.6	30.9	30.6	29.6	29.4	28.5	29.1
kac_Latn	11.0	11.6	11.3	11.6	11.0	11.8	9.4	10.3	8.8	9.6	36.0	37.1	36.8	37.3	36.7	37.7	35.4	35.4	34.8	34.4
kam_Latn	2.4	2.8	2.5	2.5	2.8	2.2	2.9	2.1	2.7	2.1	21.8	22.5	22.3	22.6	22.6	23.3	22.6	24.1	22.2	24.1
kan_Knda	19.3	20.1	19.6	20.1	18.1	19.1	14.8	15.8	14.6	15.1	51.3	52.4	51.8	52.3	51.3	51.9	47.7	49.2	47.4	48.2
kas_Arab	5.8	5.4	6.1	6.0	5.6	5.7	5.1	5.1	4.9	4.6	32.5	31.7	32.5	32.3	31.6	32.2	30.3	30.1	29.9	29.7
kas_Deva	1.8	1.8	1.7	1.8	1.8	2.0	1.7	1.7	1.6	1.8	17.6	17.3	17.9	17.9	16.9	18.5	16.6	17.8	16.2	17.2
kat_Geor	12.7	14.0	13.2	13.8	12.9	14.0	11.6	12.2	11.5	11.9	43.9	47.5	45.5	46.9	46.0	48.3	44.6	46.3	44.4	45.9
kaz_Cyrl	18.1	19.7	17.9	18.7	18.8	18.0	14.5	16.0	13.2	15.4	47.9	49.8	47.8	49.1	49.6	48.8	45.3	46.9	43.7	45.7
kbp_Latn	6.9	6.6	6.9	6.9	6.5	5.5	6.0	5.1	5.3	4.6	27.3	28.1	27.5	27.9	28.8	26.5	28.0	26.5	26.6	26.0
kea_Latn	19.7	20.9	19.0	19.7	20.6	20.6	20.5	19.8	20.0	19.5	45.4	47.7	45.7	47.0	47.8	48.4	48.1	48.0	48.3	47.3
khk_Cyrl	11.6	12.0	11.3	11.3	11.2	10.9	7.5	8.6	6.5	8.0	39.7	40.9	39.3	39.9	40.1	40.0	35.0	36.9	32.4	35.8
khm_Khmr	2.8	2.2	2.6	3.2	1.9	3.3	1.9	2.9	1.6	2.5	31.9	34.1	32.8	33.8	32.8	35.1	32.3	34.2	32.1	34.0
kik_Latn	10.9	11.1	10.7	11.0	11.0	11.2	10.2	10.6	9.9	10.2	35.5	36.1	35.8	36.0	36.2	36.2	35.4	35.5	35.3	35.4
kin_Latn	17.8	18.1	18.2	18.6	16.9	19.7	14.9	17.8	15.1	16.5	46.7	47.3	47.1	47.4	46.4	48.5	45.1	47.1	44.9	46.2
kir_Cyrl	11.8	13.6	12.1	13.0	12.6	13.0	10.7	11.2	10.0	10.3	41.5	44.6	41.6	43.3	43.9	44.3	41.6	42.4	40.8	41.1
kmb_Latn	2.3	2.6	2.3	2.3	2.6	2.2	2.4	2.0	2.3	1.8	23.0	24.9	23.7	23.8	25.2	24.4	25.2	24.9	24.9	25.0
kmr_Latn	10.4	10.3	10.9	10.6	9.6	10.1	8.3	9.6	7.4	8.9	37.0	37.0	37.6	37.6	36.1	37.1	34.0	35.9	32.4	34.7
knc_Arab	0.2	0.2	0.3	0.4	0.2	0.4	0.2	0.3	0.3	0.4	10.5	10.8	9.8	9.8	10.6	9.9	10.5	9.9	10.6	9.9
knc_Latn	3.2	3.2	3.7	3.6	3.5	3.4	3.1	3.0	2.9	2.6	24.5	24.8	25.5	25.3	25.3	25.1	23.8	24.3	22.9	23.8

Table 10: Full results across different marker insertion configurations on the FLORES-200 dataset (Languages 0–99).

Language	BLEU										chrF++									
	No Markers				Single		Simple		Complex		No Markers				Single		Simple		Complex	
	Baseline	EProj.	LP	NF	EProj.	LP	EProj.	LP	EProj.	LP	Baseline	EProj.	LP	NF	EProj.	LP	EProj.	LP	EProj.	LP
kor_Hang	11.6	12.7	10.4	10.8	10.6	12.1	9.5	10.1	9.3	9.4	33.1	34.0	32.7	33.1	32.3	32.6	29.7	29.9	29.5	29.0
lao_Laoo	7.8	8.0	7.1	7.7	6.2	8.3	5.3	7.0	4.9	5.7	42.9	44.7	43.3	44.6	43.8	45.1	43.1	44.3	42.7	43.7
lij_Latn	20.4	20.7	21.2	21.3	20.0	20.2	17.8	18.2	17.5	18.0	46.4	47.1	47.4	48.0	46.3	47.7	44.6	46.2	44.3	45.7
lim_Latn	13.8	13.4	12.5	12.3	13.3	12.0	11.5	10.8	11.5	10.7	43.3	43.1	42.6	42.6	43.2	42.5	41.9	41.4	41.6	40.8
lin_Latn	16.7	17.0	16.6	17.2	16.6	18.7	16.2	18.1	14.5	17.6	46.7	47.4	46.9	47.5	47.3	48.4	46.6	47.3	45.2	46.8
lit_Latn	23.0	23.2	22.8	23.2	22.1	22.4	19.5	19.8	17.8	17.6	51.0	51.9	51.3	52.0	51.1	51.5	48.3	48.7	47.4	47.1
lmo_Latn	7.0	6.8	7.0	7.0	6.4	6.8	6.1	6.3	6.0	5.9	32.2	32.1	32.5	32.6	31.7	32.2	31.0	31.2	30.1	30.3
ltg_Latn	19.0	18.3	17.7	18.1	17.6	17.8	14.5	15.3	14.5	14.7	46.3	46.2	45.9	46.3	45.6	46.4	42.9	43.7	42.6	42.8
ltz_Latn	25.0	24.2	24.9	25.4	23.0	24.5	19.5	21.6	18.9	20.6	53.4	53.5	53.8	54.3	52.7	53.7	50.2	51.5	49.4	50.4
lua_Latn	5.8	5.9	5.8	5.7	6.0	5.6	5.6	5.1	5.0	5.0	33.9	34.4	34.1	34.3	34.5	34.3	33.6	34.1	32.4	34.0
lug_Latn	8.6	8.1	8.5	8.7	8.0	8.9	7.7	8.3	7.7	7.9	37.4	37.4	37.4	37.8	37.3	38.4	36.7	37.8	36.7	37.4
luo_Latn	10.6	10.8	10.6	11.3	10.6	12.3	10.2	11.4	10.2	11.3	37.4	38.1	37.8	38.5	38.2	39.4	37.7	38.5	37.5	37.9
lus_Latn	10.5	11.4	10.5	10.9	11.2	10.8	9.9	8.7	9.7	9.1	35.8	36.8	35.9	36.2	36.5	36.1	34.9	34.9	34.2	34.7
lvs_Latn	21.9	21.9	21.7	22.0	21.3	21.6	18.0	18.0	17.0	17.3	49.0	49.3	49.0	49.6	48.6	49.3	45.5	45.9	44.6	44.9
mag_Deva	28.2	26.4	28.2	27.9	25.3	26.4	22.4	22.6	21.6	21.3	54.9	53.4	55.2	54.9	52.8	53.8	50.3	50.5	49.7	49.4
mai_Deva	13.0	12.9	13.6	13.7	12.8	14.5	11.6	13.7	11.8	13.1	43.2	43.5	43.7	44.1	42.7	45.0	41.3	43.8	40.3	42.7
mal_Mlym	12.4	14.9	13.1	14.2	14.5	13.5	11.8	12.0	11.7	11.3	46.9	50.7	48.0	49.9	49.9	49.5	47.0	47.7	46.7	46.4
mar_Deva	15.6	16.5	15.7	16.4	15.3	15.8	12.8	13.7	13.4	12.7	45.0	47.2	45.5	46.5	46.2	46.2	43.3	43.4	43.3	41.9
min_Arab	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.3	0.1	0.8	0.7	0.1	0.7	0.2	0.9	0.2	0.9
min_Latn	20.6	20.9	19.9	20.1	20.0	19.6	17.9	18.3	17.8	17.7	49.2	49.7	49.0	49.4	49.3	49.1	47.3	47.6	47.0	47.2
mkd_Cyrl	32.5	32.7	33.2	33.5	31.6	32.8	28.6	29.7	28.5	28.7	58.5	59.1	59.3	59.9	58.3	59.5	56.5	57.3	56.1	56.6
mlt_Latn	28.9	34.2	29.2	31.2	30.4	34.8	27.8	31.7	27.1	29.5	62.0	64.3	62.4	63.6	62.5	64.2	60.5	62.4	59.7	61.1
mni_Beng	6.9	6.4	7.1	6.8	6.3	6.5	5.2	5.8	5.1	5.6	37.0	36.9	37.0	37.0	36.5	36.5	34.1	34.9	34.2	34.3
mos_Latn	3.5	3.4	3.6	3.6	3.3	3.6	3.1	3.5	3.0	3.4	22.8	23.1	23.4	23.6	23.3	23.7	22.9	23.5	22.1	23.2
mri_Latn	20.4	18.5	20.6	19.4	18.1	20.0	16.4	19.1	15.6	18.5	44.7	43.5	45.0	44.1	43.1	44.6	41.7	43.6	40.8	42.9
mya_Mymr	2.4	2.9	2.8	3.1	2.7	3.6	1.7	2.8	1.7	2.6	29.7	33.6	31.8	35.2	31.6	39.6	32.0	38.3	32.2	37.7
nld_Latn	26.4	26.7	27.5	27.4	26.1	26.7	24.3	25.3	23.8	24.7	53.8	54.5	55.5	55.5	54.3	55.1	53.2	54.2	53.0	53.6
nno_Latn	25.4	26.7	25.5	25.9	26.0	26.1	23.7	24.1	24.1	23.2	51.2	53.0	52.0	52.6	52.6	52.5	50.8	51.4	51.1	50.6
nob_Latn	32.1	32.4	32.6	33.0	31.6	32.3	29.8	30.6	28.9	29.4	57.8	58.2	58.8	59.1	57.8	58.7	56.8	57.5	56.1	56.7
npi_Deva	14.9	17.9	16.1	17.3	16.2	17.4	13.7	14.7	12.4	14.4	44.5	49.3	45.6	48.4	46.5	50.8	46.1	48.3	44.8	47.9
nso_Latn	22.4	23.3	22.4	22.8	23.0	23.2	22.4	22.3	21.9	21.9	49.5	50.3	49.9	50.2	50.1	50.5	49.6	49.8	49.2	49.3
nus_Latn	5.3	4.9	5.5	5.2	5.4	4.8	5.0	4.0	4.7	3.5	27.7	27.6	28.0	27.8	28.1	27.7	27.4	26.6	26.7	25.8
nya_Latn	13.4	14.7	13.1	13.7	14.0	13.8	13.1	14.0	12.9	13.1	44.1	45.6	44.2	44.6	45.0	45.0	44.1	44.8	43.7	44.3
oci_Latn	34.6	35.5	35.0	35.3	34.8	34.6	33.0	33.1	32.2	32.0	59.0	60.2	59.5	60.1	59.8	59.8	58.8	59.0	58.2	58.1
ory_Orya	13.1	14.3	14.4	15.1	14.3	14.7	12.8	13.1	12.2	12.1	43.9	45.9	45.3	46.3	46.3	46.7	44.5	44.3	43.7	43.3
pag_Latn	15.6	17.5	15.7	16.8	16.1	17.2	14.3	16.1	12.8	15.5	45.0	46.6	45.5	46.3	45.7	46.4	44.2	45.0	42.6	43.8
pan_Guru	23.8	23.4	24.2	24.7	22.6	23.9	20.0	21.6	19.4	20.5	48.5	48.6	48.7	49.3	47.7	48.8	45.4	46.6	44.7	45.3
pap_Latn	31.0	28.5	30.7	30.4	29.9	31.3	28.3	28.8	27.3	27.8	55.2	53.9	55.2	55.0	55.1	55.7	53.9	54.2	53.0	53.4
pbt_Arab	13.1	12.9	13.3	13.3	12.5	13.2	11.0	12.0	10.6	11.7	37.3	37.7	37.6	37.7	37.2	37.4	35.8	36.0	35.0	35.7
pes_Arab	22.0	22.8	22.4	22.4	21.3	21.9	18.7	19.7	19.0	18.4	48.4	49.8	49.4	49.8	48.9	49.4	46.1	47.2	46.0	45.9
plt_Latn	16.9	17.8	16.7	17.3	17.1	17.6	16.2	16.5	15.7	15.7	48.8	50.1	49.1	49.7	49.7	50.5	48.6	49.2	48.1	48.4
pol_Latn	20.6	20.5	20.6	20.9	19.9	20.4	17.3	18.3	17.2	17.8	47.2	47.4	47.5	47.9	47.0	47.6	45.2	45.9	44.8	45.5
por_Latn	48.3	49.7	50.2	50.8	48.4	48.9	45.1	46.6	44.8	45.9	68.3	69.4	69.8	70.3	68.8	69.2	67.0	67.8	66.8	67.5
prs_Arab	26.3	25.5	26.7	26.5	24.0	25.7	21.1	22.8	20.6	21.8	51.6	51.8	51.8	51.7	50.7	51.2	47.8	48.5	47.1	47.8
quy_Latn	2.0	1.9	2.3	2.2	2.0	2.1	2.0	2.0	1.9	2.2	24.3	25.2	25.0	25.0	25.7	25.3	24.8	24.3	24.2	24.6
ron_Latn	35.3	37.5	37.1	38.2	36.0	39.1	32.7	36.0	32.6	35.0	59.2	60.8	60.6	61.5	59.8	62.1	57.7	60.1	57.6	59.4
run_Latn	11.8	11.8	11.7	11.8	11.5	11.8	10.5	11.2	10.2	10.3	40.5	40.6	40.7	40.9	40.8	40.9	39.9	40.0	39.4	39.3
rus_Cyrl	30.5	31.2	29.7	29.8	29.5	28.9	26.7	26.2	26.4	25.7	54.6	55.4	54.2	54.4	54.4	53.9	52.5	52.1	52.3	51.7
sag_Latn	8.2	8.1	8.0	8.2	8.0	7.8	7.7	7.2	7.6	6.9	35.3	35.5	35.5	35.9	35.7	35.6	35.6	35.4	35.3	35.4
san_Deva	1.4	1.5	1.4	1.7	1.6	1.6	1.1	1.6	0.9	1.5	24.2	24.8	25.0	25.4	25.2	25.2	22.3	24.3	21.2	23.7
sat_Olck	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.2	0.1	0.6	0.6	0.1	0.6	0.1	0.7	0.1	0.7
scn_Latn	11.0	6.3	10.7	11.3	5.6	12.3	4.8	11.0	4.9	10.9	38.9	33.3	38.5	39.2	32.5	40.7	31.5	39.7	31.3	39.1
shn_Mymr	5.2	4.6	4.7	5.0	4.6	4.4	3.8	4.0	3.5	3.3	33.2	33.0	33.0	33.0	33.0	33.2	32.4	33.0	32.0	32.9
sin_Sinh	12.8	15.2	13.7	14.8	14.2	13.8	12.2	12.3	11.2	11.7	40.7	45.1	41.8	44.3	43.1	44.9	41.1	42.5	40.4	41.4
slk_Latn	32.1	32.5	32.7	32.9	30.9	31.2	26.9	28.4	26.6	26.7	56.3	57.0	57.1	57.2	55.8	56.2	53.3	54.1	53.0	53.0
slv_Latn	27.9	28.4	28.2	28.5	26.9	28.0	23.2	25.0	22.9	24.0	53.0	53.8	53.6	54.0	52.8	53.7	50.0	51.5	49.8	50.7
smo_Latn	25.4	27.2	25.3	26.1	26.3	26.6	25.6	25.9	24.5	24.8	49.1	50.6	49.4	50.1	50.0	50.2	49.5	49.2	48.5	48.5
sna_Latn	11.3	12.0	11.5	12.1	11.7	12.4	10.7	11.8	10.5	11.4	42.2	43.0	42.7	43.0	42.6	43.0	41.8	42.1	41.4	41.9
snd_Arab	22.5	21.3	22.6	22.6	20.0	21.5	16.9	19.4	16.8	17.8	47.6	47.7	47.7	47.8	46.7	46.7	43.6	44.2	43.1	42.2
som_Latn	12.1	12.5	12.1	12.3	12.0	12.2	10.9	11.4	10.9	11.0	41.6	43.1	42.2	42.7	42.4	42.5	41.0	40.9	40.8	40.9
sot_Latn	18.4	19.1	18.1	18.7	18.8	18.7	18.1	18.4	17.5	18.0	45.5	46.6	45.5	46.0	46.1	45.9	45.6	45.4	45.0	45.1
spa_Latn	28.1	28.0	28.8	29.0	27.3	28.6	24.9	27.0	24.9	26.6	53.8	54.0	54.6	54.7	53.5	54.5	52.1	53.5	52.0	53.2
srd_Latn	28.4	28.1	28.1	27.6	27.4	26.4	26.0	24.3	25.3	23.9	54.1	54.1	54.3	54.0	53.7	53.4	52.9	52.2	52.6	51.5
srp_Cyrl	31.7	32.3	31.7	32.0	30.7	31.0	27.1	27.7	26.6	26.6	56.6	57.2	57.0	57.3	56.2	56.9	54.1	54.3	53.5	53.4
ssw_Latn	9.4	9.5	9.4	10.0	9.5	9.6	9.1	9.2	8.1	8.4	41.5	42.2	41.5	41.9	42.2	42.1	41.9	41.3	41.0	40.9
sun_Latn	16.4	17.7	17.2	17.4	16.6	17.3	14.3	16.0	13.7	15.2	45.2	46.7	46.3	46.6	45.8	46.7	43.9	45.7	43.0	44.8
swe_Latn	42.1	43.6	42.5	43.2	42.5	43.7	39.7	40.3	39.1	39.3	64.5	65.8	65.3	65.7	65.1	66.1	63.6	64.2	63.1	63.5
swh_Latn	32.0	34.4	33.0	33.0	33.4	33.6	30.3	32.0	29.8	30.5	58.5	60.2	59.2	59.4	59.6	59.5	57.6	58.2	56.9	57.2
szl_Latn	21.6	21.3	20.0	20.5	20.5	19.3	18.0	16.6	18.0	16.6	48.1	48.1	47.3	48.0	47.6	47.3	45.7	44.6	45.6	44.7
tam_Taml	16.1	17.9	16.6	17.1	16.4	17.0	14.2	14.6	13.8	13.9	50.2	53.0	51.1	52.1	51.1	52.4	48.9	50.0	48.1	48.9
taq_Latn	3.2	3.1	3.1	3.1	3.3	2.9	2.8	3.2	2.8	3.1	21.6	21.5	21.7	21.6	21.7	21.1	21.0	21.5	20.6	21.2
taq_Tfng	0.9	0.9	0.8	0.8	0.9	0.8	0.8	0.8	0.8	0.7	18.2	18.7	17.7	17.6	18.6	17.3	18.6	17.1	18.3	16.9
tat_Cyrl	17.2	19.1	17.4	18.2	17.8	18.9	15.5	16.4	14.8	16.0	46.4	48.9	46.4	47.6	47.9	48.7	45.7	46.1	44.8	45.4
tel_Telu	20.6	22.1	21.4	22.4	21.3	21.6	18.2	19.6	18.4	18.4	51.9	54.2	53.0	54.2	53.3	53.7	50.3	51.9	50.5	50.8
tgk_Cyrl	20.7	22.1	20.9	21.4	20.4	20.8	17.2	18.3	16.6	17.3	48.0	49.5	48.2	48.8	48.3	48.5	45.2	46.2	44.4	45.2
tgl_Latn	34.8	34.6	34.9	35.2	33.5	34.5	30.5	31.9	29.5	30.4	59.5	59.6	59.8	59.9	58.8	59.5	56.9	57.5	56.1	56.4
tha_Thai	5.2	6.7	4.0	3.8	4.5	6.2	2.8	4.2	2.4	3.1	37.9	39.8	38.5	39.0	39.1	40.0	37.8	39.1	37.3	38.6
tir_Ethi	4.9	4.9	4.8	5.0	4.6	4.9	3.9	4.2	3.4	4.3	23.9	24.5	24.1	24.6	23.9	24.7	22.5	23.1	21.5	22.8
tpi_Latn	17.9	19.1	17.5	18.6	19.0	20.0	18.1	19.6	17.7	19.0	41.3	42.8	41.3	42.5	42.9	43.9	42.1	43.7	41.7	43.0
tsn_Latn	21.7	22.1	21.8	22.3	21.2	22.9	20.9	21.9	19.7	20.7	47.7	47.9	47.8	48.0	47.8	47.8	47.3	46.9	46.2	45.8
tso_Latn	22.1	22.0	21.3	21.6	21.8	21.5	20.6	20.2	20.4	18.9	49.1	49.6	48.9	49.0	49.6	49.0	48.7	47.7	48.4	47.0
tuk_Latn	11.1	11.3	10.7	11.2	10.9	10.5	8.6	8.3	8.0	7.9	38.9	39.8	38.9	39.5	39.0	38.7	35.5	35.8	35.1	35.0
tum_Latn	9.4	10.0	9.5	9.8	10.5	9.9	10.0	9.9	9.6	9.1	34.0	35.5	34.1	34.6	36.1	35.2	35.3	35.5	34.8	34.7
tur_Latn	26.7	27.7	27.9	28.4	26.0	27.1	22.0	22.8	20.9	21.5	55.9	56.8	57.2	57.7	55.9	56.8	52.0	53.7	51.0	52.4
twi_Latn	11.7	11.7	11.5	11.6	11.4	11.9	10.3	11.4	10.3	11.3	37.1	37.3	37.2	37.5	37.2	37.2	36.3	36.4	35.6	35.9
tzm_Tfng	6.9	6.8	6.6	6.5	6.7	6.0	5.8	5.6	5.2	5.4	29.2	29.6	29.0	29.2	29.3	28.7	27.9	28.0	27.1	27.7
uig_Arab	9.4	11.5	10.0	10.6	9.5	10.6	6.6	8.5	6.6	8.6	38.2	42.7	39.8	41.4	39.8	41.4	35.9	38.1	36.0	36.8
ukr_Cyrl	28.8	28.9	27.0	27.0	27.7	25.9	23.9	22.9	23.6	22.2	53.9	54.5	52.5	52.7	53.6	52.2	50.8	49.4	50.4	48.9
umb_Latn	1.6	1.8	1.6	1.7	1.8	1.6	1.8	1.6	1.7	1.7	22.4	24.0	22.9	23.8	24.0	24.7	24.2	25.7	24.6	25.5
urd_Arab	22.3	21.8	22.3	22.9	21.1	21.8	17.5	19.6	17.5	18.9	47.2	47.9	47.4	48.1	46.9	47.4	43.5	45.2	43.5	44.5
uzn_Latn	17.3	18.5	17.3	17.5	17.6	16.7	13.8	14.5	12.6	13.2	50.0	51.8	50.0	50.5	51.0	50.1	46.8	48.0	45.7	46.6
vec_Latn	16.8	16.3	16.2	16.3	15.9	16.1	14.6	14.8	14.4	14.3	47.1	47.4	46.8	47.1	46.7	47.1	45.6	45.7	45.4	45.0
vie_Latn	41.0	41.0	41.3	41.3	39.8	39.9	36.2	36.6	35.5	35.0	58.6	58.8	59.1	59.1	57.9	58.1	55.2	55.5	54.8	54.4
war_Latn	30.3	30.4	29.8	30.0	29.9	29.2	29.2	27.9	28.1	26.5	56.2	56.6	56.1	56.2	56.3	56.0	55.8	54.9	55.1	54.1
wol_Latn	5.9	5.7	5.7	5.7	5.5	5.6	5.3	5.4	4.7	5.4	27.0	27.0	26.9	26.7	26.6	27.2	26.0	26.3	24.4	26.0
xho_Latn	13.4	15.6	13.1	13.9	14.2	14.8	13.0	13.3	11.8	12.8	46.6	48.8	46.7	47.6	47.9	48.7	46.8	47.4	46.5	47.0
ydd_Hebr	7.8	8.7	7.5	8.3	8.5	10.2	8.3	11.2	8.3	10.6	34.9	35.5	34.9	35.8	35.4	37.8	35.3	38.4	35.1	38.1
yor_Latn	5.3	4.8	5.4	5.8	5.0	6.9	4.9	7.1	5.0	6.9	24.8	24.0	25.1	25.5	24.5	27.4	24.6	27.6	24.0	27.2
yue_Hant	1.4	1.1	1.4	1.5	1.2	1.5	1.0	1.8	0.8	1.7	14.9	16.1	18.2	18.7	15.9	19.9	15.6	19.7	15.6	18.7
zho_Hans	1.1	0.8	1.1	1.4	1.0	2.0	0.8	2.3	0.6	2.1	19.6	22.9	22.7	23.4	22.7	23.6	21.0	23.2	21.1	22.6
zho_Hant	0.7	1.3	1.8	1.7	1.3	1.6	0.8	1.5	0.5	1.6	11.0	16.5	16.0	18.0	16.3	19.5	16.1	19.7	16.3	19.4
zsm_Latn	40.9	40.8	41.0	41.4	39.5	40.6	35.3	37.5	34.2	36.1	65.9	65.9	66.2	66.5	65.0	65.9	62.2	63.5	61.5	62.7
zul_Latn	18.1	18.4	18.3	18.4	18.0	17.9	16.6	16.3	16.0	15.5	52.1	52.5	52.2	52.3	52.1	51.9	51.0	50.4	50.6	50.0

Table 11: Contd. results across different marker insertion configurations on the FLORES-200 dataset (languages 100–203).