Leveraging Loanword Constraints for Improving Machine Translation in a Low-Resource Multilingual Context

Felermino D. M. A. Ali[1,2,3] **Henrique Lopes Cardoso**[1] **Rui Sousa-Silva**[2]
[1]LIACC, Faculdade de Engenharia, Universidade do Porto
[2]CLUP, Faculdade de Letras, Universidade do Porto
[3]DEI, Faculdade de Engenharia, Universidade Lúrio
{up202100778, hlc}@fe.up.pt, rssilva@letras.up.pt

Abstract

This research investigates how to improve machine translation systems for low-resource languages by integrating loanword constraints as external linguistic knowledge. Focusing on the Portuguese-Emakhuwa language pair, which exhibits significant lexical borrowing, we address the challenge of effectively adapting loanwords during the translation process. To tackle this, we propose a novel approach that augments source sentences with loanword constraints, explicitly linking source-language loanwords to their target-language equivalents. Then, we perform supervised fine-tuning on multilingual neural machine translation models and multiple Large Language Models of different sizes. Our results demonstrate that incorporating loanword constraints leads to significant improvements in translation quality as well as in handling loanword adaptation correctly in target languages, as measured by different machine translation metrics. This approach offers a promising direction for improving machine translation performance in low-resource settings characterized by frequent lexical borrowing.

1 Introduction

In multilingual contexts, lexical borrowing is a common phenomenon that facilitates communication by addressing linguistic diversity and filling vocabulary gaps. This is particularly evident in societies where multiple languages coexist, often with a dominant one (e.g., languages of former colonies in Africa) serving as an official or widely used medium alongside indigenous or regional languages. In such settings, translation practices frequently rely on lexical borrowing as a strategy to compensate for the lack of equivalent terms in the target language, ensuring that the intended meaning is effectively conveyed (Gauton et al., 2008).

For example, when translating technical or modern concepts–such as "internet" or "computer"–from a dominant language into a less-resourced or indigenous language, human translators may adopt phonetic adaptations or construct sound-meaning equivalents. These adaptations, which vary based on orthographic conventions, function as linguistic innovations that help fill lexical gaps. While such borrowed terms are generally accepted and understood by speakers, they also present significant challenges, particularly in the context of machine translation.

A key challenge in machine translation is handling loanword adaptation effectively. Unlike human translators, who can make context-sensitive decisions about adaptation strategies–such as phonetic transliterations, calques (literal translations), or culturally appropriate substitutions–machine translation models often struggle with these nuances. Moreover, the absence of standardized orthography in some languages further complicates the process, as variations in spelling must be accounted for while maintaining consistency in translation outputs.

Another critical aspect of lexical borrowing is its integration into the target language's grammatical structure. This involves applying appropriate pluralization rules, verb conjugations, and other morphological adjustments to ensure that borrowed terms conform to the target language's linguistic norms. For instance, a borrowed noun may require modifications to match native pluralization patterns, or a verb may need conjugation to fit syntactic constraints. While human translators systematically apply these rules, machine translation systems must be tuned to handle such challenges.

This study investigates the possibility of enhancing machine translation systems by integrating loanword constraints into source sentences. Our approach focuses on incorporating essential terminology from the target language, enabling translation models to effectively leverage external lexical constraints. By doing so, the models produce trans-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27632–27646
November 4-9, 2025

lations that more closely align with their reference counterparts, particularly in terms of loanword usage. This targeted improvement ultimately results in measurable gains in translation performance.

2 Related Work

Constrained Neural Machine Translation (CNMT) has emerged as a strong approach for ensuring the consistent use of domain-specific terminology, improving the translation of named entities, and enhancing the overall reliability of translation models (Hasler et al., 2018). A key challenge in CNMT is enforcing adherence to specific terminological constraints during translation. Several strategies have been proposed to address this challenge, broadly falling into three categories: decoding-based, architecture-based, and training data augmentation.

Decoding-based approaches focus on modifying the decoding process to enforce terminological constraints. These methods typically extend the search space during decoding to ensure that specific terms are correctly translated (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019; Hasler et al., 2018). While effective, these approaches often incur significant computational overhead and may yield only marginal improvements in overall translation quality (Guanhua et al., 2021; Zhang et al., 2023).

Architecture-based approaches, conversely, explore modifications to the neural network architecture to integrate external information. Examples include incorporating alignment information (Song et al., 2020; Guanhua et al., 2021), vectorized terminology representations (Wang et al., 2022; Conia et al., 2024), non-autoregressive translation models (Song et al., 2020), and retrieval-augmented techniques combined with knowledge graphs (Conia et al., 2024). These methods aim to enhance the model's ability to handle constrained translation by leveraging additional contextual or structural information.

Another widely adopted strategy is data augmentation, which has proven effective for constraining translation outputs (Crego et al., 2016; Song et al., 2019; Dinu et al., 2019; Michon et al., 2020; Niehues, 2021; Chen et al., 2021; Ailem et al., 2021; Zhang et al., 2023; Conia et al., 2024). Data augmentation methods can be further divided into three main categories: placeholder-based, code-switching, and post-editing methods. In placeholder-based methods, the raw bitext is pre-processed by replacing source and target constraints with ordered labels during training. At inference time, the source constraints are marked with these labels, and the model predicts the corresponding target terms autonomously (Zhang et al., 2023). In contrast, code-switching methods directly substitute source constraints with their corresponding target terms in the input sentence. This allows the model to learn a copy behavior, where the decoder generates the target text step-by-step while preserving the pre-specified constraints (Zhang et al., 2023). Placeholder-based and code-switching methods aim to signal the model to prioritize the correct use of terminology during translation.

Lastly, post-processing methods, or post-editing, include directly incorporating the required terminology into the translated output. Recently, Bogoychev and Chen (2023) demonstrated the potential of leveraging Large Language Models (LLMs) for refining translation output with specific terminology.

Loanword Handling in Machine Translation Loanwords present a unique challenge in machine translation. Nath et al. (2022) highlight the potential of effectively managing loanwords to improve translation systems, particularly for handling out-of-vocabulary (OOV) terms, co-referents, and named entities. Despite this potential, practical implementations of loanword handling techniques remain limited, with only a few studies exploring this area in depth.

A notable contribution in this field comes from Mi et al. (2018), who investigated the role of loanwords in Neural Machine Translation (NMT) systems, focusing on their ability to mitigate the OOV problem. Building on Luong et al. (2015), their approach enriches the training data with explicit information, allowing the NMT system to generate special tokens for OOV words that match the source sentence. These tokens are subsequently replaced with the correct translations after they have been processed using a bilingual dictionary. This method shares similarities with placeholder-based CNMT approaches, as described earlier. However, a key limitation of placeholder-based augmentation is the difficulty in producing fluent translations (Zhang et al., 2023).

While constrained translation has shown promise in improving translation quality, the intersection

of constrained translation and loanword handling remains underexplored. To the best of our knowledge, there is limited research on leveraging data augmentation techniques to enforce loanword terminologies in translation systems. Our study addresses this gap by exploring data augmentation approaches with a particular emphasis on enforcing loanword constraints, so as to enhance the translation of loanwords while maintaining translation fluency.

3 Approach

In this section, we detail our proposed method for handling loanword adaptation in machine translation.

3.1 Loanword definition

In the context of this study, we define loanwords as words that originate in a donor language and are borrowed into a recipient language. This borrowing process may sometimes involve transliteration, where the original word is adapted to fit the phonological or orthographic system of the recipient language. It is important to note that cognates are excluded from this definition.

3.2 Problem Statement

Let $\mathbf{x} = \{x_1, \ldots, x_M\}$ represent the source sentence and $\mathbf{y} = \{y_1, \ldots, y_N\}$ represent the target sentence. Let $\mathcal{C} = \{\langle s_1, t_1\rangle, \ldots, \langle s_K, t_K\rangle\}$ denote the set of loanword constraints between $\mathbf{x}$ and $\mathbf{y}$, where s_i and t_i correspond to the i-th source and target constraint, respectively. Each constraint can be a single word or a multi-word span, i.e., $|s_i| \geq 1$ and $|t_i| \geq 1$. Then, the problem requires that the translation systems enforce that s_i is translated into t_i.

3.3 Loanword-constrained Machine Translation

Given $\mathbf{x}$, $\mathbf{y}$ and constraints $\mathcal{C} = \{\langle s_1, t_1\rangle, \langle s_2, t_2\rangle, \ldots, \langle s_K, t_K\rangle\}$, we augment the source text $\mathbf{x}$ into $\widehat{\mathbf{x}}$ as follows:

Replacement In this approach, loanwords identified in the source sentence s_i are replaced with an extended schema that explicitly links s_i to its target counterpart t_i. The schema follows a structured format:

<|start|> s_i <|translate-as|> t_i <|end|>

This method is inspired by data augmentation techniques for guiding NMT systems, commonly used in prior work on encoder-decoder architectures (Crego et al., 2016; Song et al., 2019; Dinu et al., 2019; Michon et al., 2020; Niehues, 2021; Chen et al., 2021; Ailem et al., 2021; Zhang et al., 2023; Conia et al., 2024). In particular, we adopt the schema proposed by Conia et al. (2024). This format serves as an explicit signal to the model, guiding it to correctly incorporate the target term t_i in the output. While the "replacement" strategy is primarily designed for encoder-decoder models, alternatives for decoder-only models are discussed in the following section.

Prompting Similar to the replacement approach, this strategy involves augmenting the source text with constraints but adopts a prompt-based format. The prompt template (see Figure 1) provides detailed instructions for the model, incorporating the constraints of the loanword alongside the translation task. Our prompt design builds on prior work in terminology-constrained prompting for large language models, particularly the studies by Moslem et al. (2023), Ghazvininejad et al. (2023), and Lu et al. (2024). This method is designed for decoder-only instruction-tuned models, capitalizing on their capabilities in following natural language directives.

Both "Replacement" and "Prompting" aim to enforce a copy behavior based on the provided constraints, to encourage the model to map s_i and t_i. However, while this approach provides strong guidance for loanword adaptation, it does not fully guarantee that the model will always adhere to the constraints. This is due to the inherent unpredictability of language models' decoding behavior. To thoroughly assess the effectiveness of the proposed methods, we provide a comprehensive evaluation in the following sections.

4 Experimental Setup

In this section, we begin by outlining the datasets used in the experiments and data preparation. Lastly, we detail the models used for fine-tuning in the machine translation task.

4.1 Datasets

We used a Portuguese-Emakhuwa bitext dataset augmented with constraint annotations to train our models. The dataset includes considerable manually annotated constraints, but a significant propor-

Prompt template

Translate the following sentence from {source_language} into {target_language}:

Sentence: {source}

Guidelines:
You may adapt loanwords as necessary.
For instance:
$\{s_0\}$ should be translated as $\{t_0\}$
$\{s_1\}$ should be translated as $\{t_1\}$
...
$\{s_k\}$ should be translated as $\{t_k\}$

Output:

Figure 1: Prompt template for machine translation task. The placeholders s_i and t_i are replaced with corresponding loanwords constraint pairs $\mathcal{C} = \{\langle s_1, t_1\rangle, \langle s_2, t_2\rangle, \ldots, \langle s_K, t_K\rangle\}$.

tion of the data lacks these annotations. To address this, we used weak supervision techniques to annotate the remaining data. Further details on the dataset and our annotation methodology are provided below.

4.1.1 Training and Validation Data

For training and validation, we used the dataset by Ali et al. (2024a), which is Portuguese-Emakhuwa bitext. Specifically, we selected the configuration that contains approximately 65k training examples and 964 validation examples. The configuration combines biblical parallel texts, manually translated news articles and other materials sourced on the Web (see Table 1). The manually translated news subset is exceptional as it provides triples of source text, target translations, and manually annotated loanwords, captured during the translation process.

Given that the training data was partially annotated with manual loanwords (see Table 1), we expanded the loanword pairs annotations by using weak supervision, leveraging large language models as annotator. Specifically, we leveraged the existing gold standard annotations, reformatted them into instruction-tuning prompts, and performed supervised fine-tuning on GPT4o-mini model. The fine-tuned model was then used to automatically annotate the remaining unlabeled subset of the data. The full details of this process are provided in the following subsection.

Category	TRAIN	DEV	Manual Loanwords Annotations
religious	45,386	290	×
news-politics	1,950	67	✓
news-economy	1,950	65	✓
news-culture	3,611	108	✓
news-sports	3,030	63	✓
news-health	2,256	83	✓
news-society	1,922	90	✓
news-world news	2,515	89	✓
tales	2,059	20	×
legal	893	24	×
Wikipedia	27	28	×
history	9	37	×
Total	65,608	964	

Table 1: Training and validation Portuguese-Emakhuwa bitext data (Ali et al., 2024a). Note: Some strings in the news category did not contain loanword constraints, meaning translations were performed entirely using native words in the target language.

4.1.2 Generating Loanword Constraint Pairs

Given a source sentence $\mathbf{x} = \{x_1, x_2, \ldots, x_M\}$ and a target sentence $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$, the objective is to build a model M capable of mapping loanword pairs between the two sequences. This mapping is represented as a set of aligned pairs $\mathcal{C} = \{\langle s_1, t_1\rangle, \langle s_2, t_2\rangle, \ldots, \langle s_K, t_K\rangle\}$, where each pair $\langle s_i, t_i\rangle$ denotes a correspondence between donor s_i in the source sentence and a recipient t_i in the target sentence.

For the model M we leveraged the capabilities of *gpt-4o-mini-2024-07-18*[1], a state-of-the-art LLM by OpenAI. We fine-tuned on manually annotated sentence-level loanword pairs. We divided the gold-standard instances into an 80/20 split for training and validation, respectively. Then, for fine-tuning we used instruction prompt to guide the model in predicting loanword pairs alignments $\mathcal{C}$ from the parallel sentences $\mathbf{x}$ and $\mathbf{y}$ (see Figure 2 and Figure 4). Finally for inference, we set the temperature hyperparameter to 0 to ensure deterministic outputs.

We evaluated the model's performance using the dataset introduced by Ali et al. (2024c), which extends *FLORES+* for Portuguese–Emakhuwa machine translation evaluation. A key advantage of this evaluation set is the inclusion of manually annotated loanwords in each pair of parallel sen-

[1]We fine-tuned this model during the OpenAI's free 1M-token offer period held from August through September 2024.

Figure 2: Generating Loanword Constraint Pairs Using the Fine-Tuned *gpt-4o-mini-2024-07-18* Model

tences. Below, we present a detailed analysis of the model's performance:

Evaluation To assess the effectiveness of the model described above, we compared its output against the gold standard annotations of our test dataset. For that, for each sentence pair we concatenated source and target text, then for each token we assigned a tag from one of two categories: LOAN–indicating that the token is part of a loanword–or O–indicating that the token is not part of a loanword. We adopted the BILOU scheme, which is primarily used in Named Entity Recognition (NER) tasks. In our case, the BILOU format was adapted to label tokens based on their position within or outside a loanword category (see Figure 5 in Appendix A.5 for details).

The evaluation results, presented in Table 2, show that the model achieves an F1-score of 95% on the gold standard test set. This high accuracy provided us some confidence on the reliability of the automatic annotations using the model described above.

Tag	Precision	Recall	F1-Score	Accuracy
LOAN	0.87	0.95	0.91	–
O	0.99	0.98	0.99	–
Macro Avg.	0.93	0.97	0.95	–
Accuracy				0.98

Table 2: Evaluation results of loanword constraints generation model

4.1.3 Test Data

To evaluate the performance of our machine translation results, we also used the FLORES+ *DEV* and *DEVTEST* datasets (Ali et al., 2024c), which contain 997 and 1012 pairs of sentences, respectively. Each sentence pair includes manually annotated loanword constraint information.

4.1.4 Data Pre-Processing

Given source sentences $\mathbf{x}$ target sentences $\mathbf{y}$, and $\mathcal{C}$, we augment all source sentences into $\hat{\mathbf{x}}$ using $\mathcal{C}$ as described in Section 3.3. This preprocessing step was applied to the training, validation, and test sets.

4.2 Models

We fine-tuned and evaluated a range of multilingual NMT models and large language models:

Multilingual NMT Models: M2M-100 (Fan et al., 2021): A many-to-many multilingual translation model designed to directly translate between 100 languages. NLLB-200 (NLLBTeam et al., 2024): A cutting-edge model designed for large-scale multilingual translation, supporting over 200 languages. Its architecture is optimized for low-resource languages and supports extension to new unseen languages. We used NLLB-200's distilled variant with 600M parameters.

Large Language Models: We used a set of decoder-based LLMs for multilingual fine-tuning experiments, including Llama 3.2 (3B), Qwen2.5 (3B, 7B), Phi-3.5-mini (3.8B), Gemma-3 (4B), Llama 3.1 (8B), and Gemma-2 (9B). For fine-tuning, we applied LoRA (Hu et al., 2022), a parameter-efficient tuning method that inserts trainable low-rank matrices, using a rank of 16 and a scaling factor (*lora_alpha*) of 16. To optimize memory usage, we used Unsloth (Daniel Han and team, 2023), which reduces VRAM requirements and supports larger batch sizes.

Training Training is performed using pairs of source sentences $\mathbf{x}$, their augmented counterparts $\hat{\mathbf{x}}$, and the corresponding translations $\mathbf{y}$. This setup is applied to both translation directions: Portuguese–Emakhuwa and Emakhuwa–Portuguese. We experiment with two training settings:

- **Without Constraints:** The model is trained to perform standard translation from source to target $\mathbf{x} \rightarrow \mathbf{y}$, without any form of constraint injection.
- **With Constraints:** The model is trained on a mixed setup: (1) translating unconstrained source sentences $\mathbf{x} \rightarrow \mathbf{y}$, and (2) translating source sentences with constraints explicitly injected $\mathbf{x} \rightarrow \mathbf{y}$.

Inference For both training setups, we perform inference using two types of inputs:

- Unconstrained Input ($\mathbf{x}$): To verify whether constraint-aware training degrades or enhances general translation performance on standard inputs/prompts.
- Constrained Input ($\widehat{\mathbf{x}}$): To evaluate the model's capacity to accurately respect and incorporate the constraints provided during translation, leveraging the experience gained during training.

Our experiments were carried out on 8 NVIDIA H100 GPUs for both model training and inference. Table 5 lists the base model used, while the hyperparameters are detailed in Section A.2 of the Appendix.

Evaluation For our machine translation evaluation, we used three key metrics: BLEU (Papineni et al., 2002), ChrF (Popović, 2015), and AfriCOMET (Wang et al., 2024). AfriCOMET, an extended version of the COMET framework (Rei et al., 2020), was initially introduced to support a limited set of African languages. It has since been expanded by Wang et al. (2024) to encompass a broader range of 76 African languages.

In addition, we assess loanword accuracy using sentence-level constraint accuracy (SCA), a metric introduced by Zhang et al. (2023), which measures how accurately loanwords are translated within sentences. Under this metric, a translation is considered correct only if it accurately renders all required loanwords within a sentence.

5 Results and Discussion

Table 3 summarizes the outcomes of our experiments, comparing the performance of multilingual neural machine translation (NMT) models and large language models (LLMs) under two training scenarios: constrained training and unconstrained training. The table reports BLEU, CHRF, AfriCOMET, and SCA scores on two evaluation sets—the FLORES development (DEV) and devtest sets. We evaluate translation performance in both directions: Portuguese to Emakhuwa (pt→vmw) and Emakhuwa to Portuguese (vmw→pt). Statistical significance is assessed through paired bootstrap resampling tests, comparing each alternative model against the baseline (i.e., standard input training).

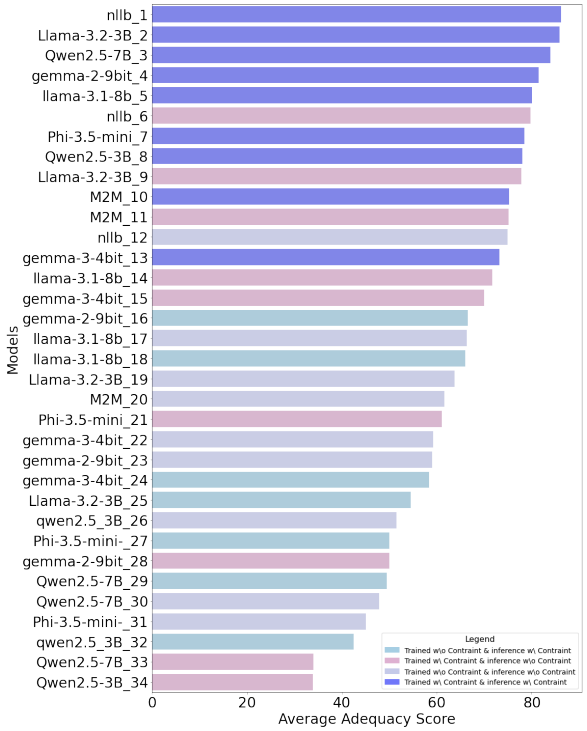

Figure 3: Ranking of the models based on human judgments

Models trained with constraint injection consistently outperform their unconstrained counterparts across all metrics. The improvements are particularly notable in SCA, supporting the hypothesis that constraint-aware training more effectively preserves loanword translation fidelity, especially in the pt→vmw direction. BLEU, CHRF, and AfriCOMET also increase significantly under constraint-aware training, indicating gains not only in lexical constraint adherence but also in overall translation quality. Decoder-only models (e.g., Qwen2.5, LLaMA, Phi) generally lag behind encoder-decoder architectures like NLLB in unconstrained settings. However, when trained with constraints, these models exhibit significant improvements. Interestingly, the gains from constraint-aware training are consistently greater in the pt→vmw direction. This asymmetry reflects a broader trend in low-resource machine translation, where generating text in a low-resource language is more challenging due to its limited target-side vocabulary and training data. In this context, constraint-aware training proves beneficial, as it provides additional guidance that helps the model produce more accurate translations.

Robustness We also evaluated the robustness of the models under mismatched input condi-

	SRC	Model	Size	DEV								DEVTEST							
				BLEU	CHRF	COMET	SCA	BLEU	CHRF	COMET	SCA	BLEU	CHRF	COMET	SCA	BLEU	CHRF	COMET	SCA
				pt→vmw				vmw→pt				pt→vmw				vmw→pt			
without Constraints	x - baseline	M2M-100	0.4B	9.07	40.09	0.50	14.84	20.24	44.75	0.59	30.39	7.01	36.24	0.47	43.08	13.19	37.16	0.53	56.52
		NLLB	0.6B	9.30	41.22	0.50	32.09	20.30	44.38	0.60	51.05	8.55	41.09	0.50	33.59	19.60	43.62	0.59	52.17
		Llama 3.2	3B	6.54	35.36	0.47	14.04	20.31	44.71	0.57	12.14	5.17	32.54	0.45	42.00	12.37	37.52	0.51	40.22
		Qwen2.5	3B	5.38	32.94	0.47	12.54	14.52	38.44	0.54	12.04	4.72	31.36	0.46	39.82	8.96	33.17	0.49	38.83
		Phi3.5-mini	3.8B	5.44	31.52	0.46	12.94	16.62	39.22	0.53	12.34	4.65	29.91	0.44	41.11	10.32	33.96	0.48	40.51
		Gemma-3	4B	7.19	35.57	0.49	14.14	21.20	43.99	0.57	12.24	5.63	32.21	0.45	42.49	13.32	36.77	0.51	40.02
		Qwen2.5	7B	5.96	33.87	0.48	13.14	17.54	42.38	0.58	11.94	5.34	32.28	0.46	40.91	11.38	35.95	0.52	39.62
		Llama 3.1	8B	8.25	37.79	0.50	14.54	24.40	47.92	0.60	12.34	6.65	34.71	0.47	43.77	14.76	39.70	0.54	40.02
		Gemma-2	9B	8.09	37.18	0.49	14.14	25.43	48.16	0.62	12.14	7.14	34.51	0.47	42.98	15.39	39.39	0.55	39.92
with Constraints	$\hat{\mathbf{x}}$	M2M-100	0.4B	17.46	49.73	0.54	70.61	25.25	48.60	0.62	75.32	10.06	39.54	0.49	85.96	14.86	38.75	0.55	**84.78**
		NLLB	0.6B	14.46	46.74	0.53	60.88	21.85	45.85	0.61	64.59	**14.18**	**46.65**	**0.53**	79.15	**20.96**	**44.73**	**0.60**	79.54
		Llama 3.2	3B	17.78	50.19	0.56	78.03	25.65	49.50	0.61	12.34	9.45	38.53	0.50	**87.75**	13.68	39.08	0.53	40.12
		Qwen2.5	3B	13.94	45.00	0.53	70.51	21.74	45.37	0.59	12.44	7.20	34.28	0.48	80.34	11.95	35.92	0.50	39.53
		Phi3.5-mini	3.8B	16.29	46.75	0.54	69.51	25.49	47.89	0.59	12.44	8.34	35.27	0.47	80.93	13.78	37.42	0.51	40.12
		Gemma-3	4B	13.91	43.93	0.53	45.84	25.64	48.36	0.60	12.29	8.65	36.33	0.49	63.24	14.74	38.74	0.53	40.37
		Qwen2.5	7B	16.56	48.32	0.55	75.53	25.09	48.58	0.61	12.44	8.16	36.72	0.49	86.17	14.32	38.32	0.53	40.22
		Llama 3.1	8B	**19.53**	**52.08**	**0.57**	82.65	29.19	52.14	0.63	12.44	9.99	39.42	0.50	87.65	15.63	40.97	0.54	40.12
		Gemma-2	9B	19.12	50.46	0.56	**84.55**	**31.40**	**52.91**	**0.64**	12.44	10.11	37.78	0.49	87.45	18.10	41.56	0.56	39.92

Table 3: Evaluation results. For COMET values underlined indicate significant evidence that they outperform the baseline ($p < 0.05$), based on pairwise randomized significance tests (Koehn, 2004).The **SRC** column indicates whether loanword constraints are applied ($\hat{\mathbf{x}}$) or not (x) in source sentences. COMET refers to AfriCOMET.

tions—specifically, when unconstrained models are given constrained inputs ($\hat{\mathbf{x}}$), and when constraint-trained models are tested with unconstrained inputs ($\mathbf{x}$). As shown in Table 4, unconstrained models do not consistently benefit from receiving constrained input. While SCA scores occasionally show slight improvements, overall performance often declines, likely due to a mismatch between the input format and the model's training distribution. In contrast, constraint-aware models, having been trained on a mix of constrained and unconstrained inputs, exhibit better robustness. When evaluated with unconstrained input, their performance remains stable for most models. These results suggest that mixed training not only improves generalization but also makes models more adaptable to varying input types.

5.1 Human Evaluation

To assess the quality of translations produced by the different models, we conducted a human evaluation focusing on adequacy. Two independent professional translators volunteered to rate the translations based on a direct assessment score. This score ranged from 0 to 100, with specific guidelines: A score of 0 indicates that no meaning is preserved in the translation. A score between]0 - 34] means the translation retains some of the source meaning but loses significant parts. A score between]34 - 67] indicates that most of the meaning is preserved. A score between]67 - 99] reflects a translation that is consistent with the source text, while a score of 100 represents a perfect translation.

Inter-annotator agreement was measured to ensure the reliability of the human judgments. We calculated Pearson's correlation coefficient (0.620), Spearman's rank correlation coefficient (0.580), and the Intraclass Correlation Coefficient ICC(3,2) (0.764). These values collectively suggest a moderate level of agreement between the two evaluators.

The overall rankings derived from these human judgments are presented in Figure 3. A key finding is that constraint-trained models (in blue) generally achieved higher adequacy scores compared to their counterparts trained solely on unconstrained inputs. Furthermore, the results indicate that both encoder-decoder and decoder-only architectures demonstrated competitive performance in this evaluation, with specific NLLB and Llama variants standing out within their respective categories. This suggests that training with awareness of constraints contributes positively to translation adequacy as perceived by human evaluators. Interestingly, human evaluation findings align with the trends observed in our automatic evaluation metric and corroborate the quantitative improvements seen in metrics such as BLEU, CHRF, and AfriCOMET.

5.2 Case Study

To illustrate the impact of loanword injection, we present one translation example in Table 9 (see Appendix A.4). This example demonstrates that our

	SRC	Model	Size	DEV				DEVTEST											
				BLEU	CHRF	COMET	SCA	BLEU	CHRF	COMET	SCA	BLEU	CHRF	COMET	SCA	BLEU	CHRF	COMET	SCA
without Constraints	$\hat{x}$	Llama 3.2	3B	6.64	35.65	0.48	21.66	19.88	43.55	0.56	12.14	5.34	32.09	0.46	48.12	10.68	35.40	0.50	40.81
		Qwen2.5	3B	5.48	32.62	0.47	15.75	14.39	38.50	0.52	13.94	4.85	30.80	0.45	42.89	8.64	32.16	0.47	41.21
		Phi3.5-mini	3.8B	6.21	32.07	0.46	21.87	18.22	39.62	0.54	12.74	4.75	28.96	0.44	48.81	12.33	33.93	0.49	40.71
		Gemma-3	4B	9.89	40.16	0.51	21.06	24.70	47.56	0.59	12.64	6.79	34.05	0.47	51.38	14.72	38.04	0.52	41.30
		Qwen2.5	7B	5.81	34.30	0.47	17.95	14.88	39.17	0.52	16.15	4.86	31.67	0.46	46.25	10.80	34.28	0.48	42.29
		Llama 3.1	8B	11.45	42.80	0.53	25.58	25.45	48.41	0.60	13.14	7.53	35.88	0.48	52.87	14.81	38.95	0.53	41.30
		Gemma-2	9B	12.24	43.46	0.52	23.97	28.56	50.48	0.62	12.44	8.30	36.48	0.48	52.37	16.18	39.90	0.55	40.22
with Constraints	x	M2M-100	0.4B	9.27	39.48	0.50	13.94	20.33	44.55	0.58	29.98	7.05	36.30	0.47	43.47	13.20	37.06	0.53	56.02
		NLLB	0.6B	10.10	42.07	0.51	14.94	20.08	44.37	0.60	33.50	9.61	42.01	0.50	43.28	20.37	44.06	0.59	58.39
		Llama 3.2	3B	7.77	38.09	0.49	13.84	20.12	44.98	0.57	12.54	6.16	34.80	0.46	42.49	12.84	38.06	0.52	40.71
		Qwen2.5	3B	4.09	29.22	0.45	12.04	11.84	35.61	0.52	11.43	3.72	28.40	0.44	37.65	8.34	32.12	0.48	39.03
		Phi3.5-mini	3.8B	6.74	33.75	0.48	12.94	16.80	39.90	0.53	12.24	6.03	32.22	0.46	41.50	10.84	34.82	0.49	40.42
		Gemma-3	4B	8.60	37.55	0.50	13.54	22.02	45.60	0.58	12.24	7.22	34.72	0.48	41.80	13.89	37.85	0.52	40.32
		Qwen2.5	7B	5.14	31.96	0.46	12.04	15.18	39.12	0.56	11.63	4.87	31.77	0.46	38.83	10.46	34.83	0.51	39.23
		Llama 3.1	8B	9.27	40.11	0.51	14.44	24.66	48.37	0.60	12.24	7.20	36.20	0.48	44.66	14.81	39.71	0.54	40.22
		Gemma-2	9B	6.64	33.69	0.47	13.04	23.30	46.13	0.60	12.34	6.11	32.53	0.45	42.00	15.22	39.07	0.55	39.92

Table 4: Robustness evaluation results under mismatched input conditions. Underlined scores indicate statistically significant improvements over the baseline (see Table 3). Note that NLLB and M2M models are excluded from the "without constraints", as encoder-decoder models are expected to produce poor translations when constraint tags are added to input sentences without having been trained to handle them.

method effectively guides both encoder-decoder and decoder-only models to incorporate specified loanword constraints into the generated translations.

Baseline (Without Loanword Injection) In the translation direction **pt→vmw**, all baseline models, trained without loanword injection, struggled to accurately translate the Portuguese loanwords *murais* and *rabiscos*. These terms were often mistranslated or omitted entirely, resulting in low adequacy scores across the board. One exception was the Llama 3.2 model, which, although deviating from the reference at the lexical level, managed to convey the intended meaning more accurately. Additionally, most models identified *grafite* as a potential loanword and attempted to adapt it to Emakhuwa. However, these adaptations often violated phonological norms. For example, the letter *g* does not exist in the Emakhuwa orthography and should ideally be replaced by *k* (Ali et al., 2024b).

A similar trend is observed in the reverse direction **vmw→pt**. Words such as *imuraaxi*, *irapiixiku*, and *ekarafiiti* were often identified as loanwords and phonetically adapted into Portuguese-like forms. However, most models struggled to map them to the correct Portuguese equivalents. The exception was the Gemma-3 4B model, which successfully captured the full semantic mapping. This difficulty likely stems from the fact that these are innovative borrowings that were not present in the training data, making them unfamiliar to the models.

Loanword Injection Improves Translation When loanwords were injected into the source sentences for both **pt→vmw** and **vmw→pt**, translation quality improved substantially. This is reflected in the adequacy scores, which increased significantly, ranging from 52 to 98, depending on the model and direction.

6 Conclusion

In this study, we explored the impact of incorporating loanword constraints into machine translation models for low-resource languages, specifically focusing on the Portuguese-Emakhuwa language pair. Our experiments demonstrated that explicitly guiding models to recognize and adapt loanwords significantly improves translation quality, as evidenced by higher BLEU, CHRF, and AfriCOMET scores. The improvements are evident both using multilingual neural machine translation models and LLMs.

This research presents a practical approach to enhancing machine translation, particularly in low-resource language settings, by leveraging external loanword glossaries. The proposed method augments training data with explicit loanword constraints, effectively guiding the translation model to handle lexical borrowing correctly. In real-world applications, this demands the development of a dynamic, human translator-curated loanword glossary. This continuously refined glossary then serves as external information that directly improves the machine translation model's performance, espe-

cially in handling the complexities of loanword adaptation, a common challenge in resource-scarce scenarios.

Future Work While the proposed approach has shown promise in guiding MT models to handle loanwords, we believe it also holds significant potential use-case for promoting the use of native terminology. In future work, we plan to investigate how this method can be extended to incorporate bilingual dictionary entries, enabling models to integrate native lexical items that were unseen during training. Such an extension would allow MT systems to balance the use of culturally grounded words and the accurate adaptation of borrowings, thereby enhancing both translation adequacy and linguistic authenticity.

7 Limitations

Our experiments are constrained by the scarcity of parallel corpora that include explicit loanword annotations. As a result, our findings are based solely on the Portuguese–Emakhuwa pair. While the results are promising, additional data from other languages is needed to confidently assess the generalizability and broader applicability of our approach.data.

Another limitation lies in our choice of evaluation metrics. In addition to traditional metrics like BLEU and CHRF, we used AfriCOMET–a new metric designed to support African languages. However, to the best of our knowledge, AfriCOMET currently does not support Emakhuwa, despite its inclusion of several typologically similar languages. Future work should focus on validating and, if necessary, adapting these metrics to ensure a fair and precise evaluation of translations involving Emakhuwa.

Our study investigates two main strategies for incorporating lexical constraints into translation: replacement and prompting. While both approaches proved effective, we recognize that our exploration of alternative prompt formulations and schema variations was not exhaustive. Building on insights from previous research, the replacement strategy involved minimal edits to the source text to maintain fluency. In contrast, the prompting strategy used direct, instruction-based templates.

Future work should expand on these foundations by exploring a wider array of constraint integration techniques to evaluate their relative effectiveness.

Acknowledgements

This work was financially supported by UID/00027 - Artificial Intelligence and Computer Science Laboratory (LIACC) and by UID/00022 - Centre for Linguistics of the University of Porto (CLUP), funded by Fundação para a Ciência e a Tecnologia (FCT), I.P./MECI through national funds. Felermino Ali is supported by a PhD grant (with reference SFRH/BD/151435/2021), funded by FCT.

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Felermino D. M. A. Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024a. Building resources for emakhuwa: Machine translation and news classification benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14842–14857, Miami, Florida, USA. Association for Computational Linguistics.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024b. Detecting loanwords in emakhuwa: An extremely low-resource Bantu language exhibiting significant borrowing from Portuguese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4750–4759, Torino, Italia. ELRA and ICCL.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024c. Expanding FLORES+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Guanhua Chen, Yun Chen, Yong Wang, and Victor O. K. Li. 2021. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge

graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *Preprint*, arXiv:1610.05540.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Rachelle Gauton, Elsabe Taljard, Tirhani Mabasa, and Lufuno Netshitomboni. 2008. Translating technical (lsp) texts into the official south african languages: A corpus-based investigation of translators' strategies. *Language Matters*, 39(2):148 – 180. Cited by: 2.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Chen Guanhua, Chen Yun, and Li Victor O.K. 2021. Lexically constrained neural machine translation with explicit alignment guidance. In *Proceedings of AAAI*, volume 35, pages 12630–12638.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR 2022*.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. 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 11–19, Beijing, China. Association for Computational Linguistics.

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. 2018. Toward better loanword identification in Uyghur using cross-lingual word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3027–3037, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine*

Translation, pages 902–911, Singapore. Association for Computational Linguistics.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022. A generalized method for automated multilingual loanword detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jan Niehues. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

NLLBTeam, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Kai Song, Kun Wang, Heng Yu, Yue Zhang, Zhongqiang Huang, Weihua Luo, Xiangyu Duan, and Min Zhang. 2020. Alignment-enhanced transformer for constraining nmt with pre-specified translations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8886–8893.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. 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 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516, Miami, Florida, USA. Association for Computational Linguistics.

Shuo Wang, Zhixing Tan, and Yang Liu. 2022. Integrating vectorized lexical constraints for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7063–7073, Dublin, Ireland. Association for Computational Linguistics.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Base Models

Table 5 lists the base models used and the repositories or sources from which they were obtained.

Table 5: Base models hugging-face checkpoints.

Models	Size	Model name
M2M-100	418M	facebook/m2m100_418M
NLLB	600M	facebook/nllb-200-distilled-600M
Qwen2.5	3B	unsloth/Qwen2.5-3B-Instruct
Qwen2.5	7B	unsloth/Qwen2.5-7B-Instruct
Llama 3.1	8B	unsloth/Llama-3.2-3B-Instruct
Llama 3.2	3B	unsloth/llama-3.1-8b-instruc
Phi3.5-mini	3.8B	unsloth/Phi-3.5-mini-instruct
Gemma-3	4B	unsloth/gemma-3-4b-it
Gemma-2	9B	unsloth/gemma-2-9b-it

A.2 Hyperparameters

Table 6: NLLB Hyperparameters for fine-tuning.

Hyperparameter	Value
Epochs	3
Learning Rate	1e-4
Optimizer	Adafactor, relative_step=False, scale_parameter=False, clip_threshold=1.0 weight_decay=1.3

Table 7: M2M Hyperparameters for fine-tuning.

Hyperparameter	Value
Epochs	3
Learning Rate	3e-05
Optimizer	Adamw with betas=(0.9,0.98) and epsilon=1e-06

Table 8: LLMs Hyperparameters for fine-tuning (with Unsloth).

Hyperparameter	Value
Epochs	1
Learning Rate	2e-4
Lora Alpha	16
Lora Rank	16
Optimizer	Adamw, Scheduler=linear

A.3 Prompt Template

Figure 4, shows the prompt template for mapping donor-recipient loanwords relationships, given a sentence pair.

Prompt template

Given a source sentence in {source_language} and its corresponding translation in {target_language}, identify the {source_language} loanwords used in the {source_language} sentence.
Present the results in the format:
"donor sequence" => "recipient sequence" without any further explanation.

Sentences:
{source_language}: {source_sentence}
{target_language}: {target_sentence}

Output:

Figure 4: Prompt template. The placeholders are substituted with source and target languages and corresponding sentences.

A.4 Examples

	pt	Murais ou rabiscos indesejados são conhecidos como grafite.	
	vmw	Imuraaxi wala irapiixiku soohitthuneya ti soosuwelaxiwa okhala ekarafiiti.	
	en	Unwanted murals or scribble is known as graffiti.	
		Baseline	Adequacy [0-100]
M2M-100	pt-vmw	~~Makhalelo oovirikana a atthu~~ anisuweliwa okhala grafite.	52
	vmw-pt	~~MURAGES~~ ou ~~rapiosos~~ desnecessários são conhecidos como ~~grafiços~~.	63
NLLB	pt-vmw	Ithoonyeryo soohiloka ~~sa exiri~~ sinnisuweliwa okhala karafiiti.	55
	vmw-pt	~~Buracos~~ desnecessários são conhecidos como grafite.	65
Llama 3.2	pt-vmw	Murais wala rabiscos, itthu soohiloka, sinihimmwa okhala grafite.	80
	vmw-pt	~~Os marromados ou brancos são os mais comuns~~.	0
Qwen2.5 3B	pt-vmw	~~Nto makhala oohitthuneya wala alavilavi~~ annisuweliwa ntoko grafiti.	32
	vmw-pt	~~As rãs ou rapazes~~ são conhecidos como ~~carabias~~.	38
Phi3.5-mini 3.8B	pt-vmw	~~Murisini~~ wala olepa soohiloka oniitthaniwa grafite.	60
	vmw-pt	~~Imoraques ou irapeiscos inofensivos~~ são conhecidos como ~~carabeiros~~.	12
Gemma-3 4B	pt-vmw	Ikarafiiti ti soolempwa soohiloka.	20
	vmw-pt	Os murais ou rabiscos são conhecidos como grafite.	83
Qwen2.5 7B	pt-vmw	~~Otthukweliwa wa muru~~ oniitthaniwa grafite.	15
	vmw-pt	~~Pilhas de lixo e de lixeiras sem necessidade~~ são conhecidas como ~~grafiteiros~~.	13
Llama 3.1 8B	pt-vmw	Murais wala soolempwa soohiloka siniitthaniwa grafite.	82
	vmw-pt	~~Imunidades ou isentimentos injustificados~~ são conhecidos como grafitis.	35
Gemma-2 9B	pt-vmw	Murais wala rabisko woohiphavela onisuweliwa okhala grafite.	79
	vmw-pt	~~As muralhas~~ ou grafites são conhecidos como arte urbana.	65
		Augmented with Loan Injection	
M2M-100	pt-vmw	~~Makhalelo oovirikana a atthu~~ anisuweliwa okhala grafite.	52
	vmw-pt	Murais ou rabiscos desnecessários são conhecidos como grafite.	98
NLLB	pt-vmw	Imuraaxi wala irapiixiku soohitthuneya sinisuweliwa ntoko ekarafiiti.	98
	vmw-pt	Murais ou ~~debulhos~~ desnecessários são conhecidos como grafite.	90
Llama 3.2 3B	pt-vmw	Imuraaxi wala irapiixiku sihiniireliwa aya, sinnisuweliwa okhala ekarafiiti.	91
	vmw-pt	Os murais ou os rabiscos são ~~formas de desenho que se tornaram um tipo~~ de grafite.	62
Qwen2.5 3B	pt-vmw	Imuraaxi wala irapiixiku soohikhala sa atthu aniihaniwa ekarafiiti.	70
	vmw-pt	Os murais ou rabiscos não oficiais são considerados grafite.	71
Phi3.5-mini 3.8B	pt-vmw	Imuraaxi wala irapiixiku soohiloka, sinnooniwa okhala ekarafiiti.	92
	vmw-pt	Murais e rabiscos ~~inofensivos~~ são conhecidos como grafite.	70
Gemma-3 4B	pt-vmw	Imuraaxi wala irapiixiku ~~soohikhala sa atthu~~ aniihaniwa ekarafiiti.	68
	vmw-pt	Os murais ou rabiscos não oficiais são considerados grafite.	86
Qwen2.5 7B	pt-vmw	Imuraaxi wala irapiixiku sihikhanle soosuweliwa ntoko ekarafiiti	84
	vmw-pt	Murais ou rabiscos não autorizados são conhecidos como grafite.	90
Llama 3.1 8B	pt-vmw	Imuraaxi wala irapiixiku soohiloka siniitthaniwa ekarafiiti.	98
	vmw-pt	~~Muranjos~~ ou rabiscos são conhecidos como grafite.	69
Gemma-2 9B	pt-vmw	Irapiixiku wala ekarafiiti soohiloka sinisuweliwa okhala ~~muraaxi wala irapiixiku soohiloka~~.	67
	vmw-pt	Murais ou rabiscos não autorizados são conhecidos como grafite.	88

Table 9: One sample example for case of study

A.5 Loanword Tagging

To evaluate the performance of loanword constraint models (see Section 4.1.2), we preprocessed both the gold standard test data and the model's predicted outputs using the BILOU format, as illustrated in Figures 5.

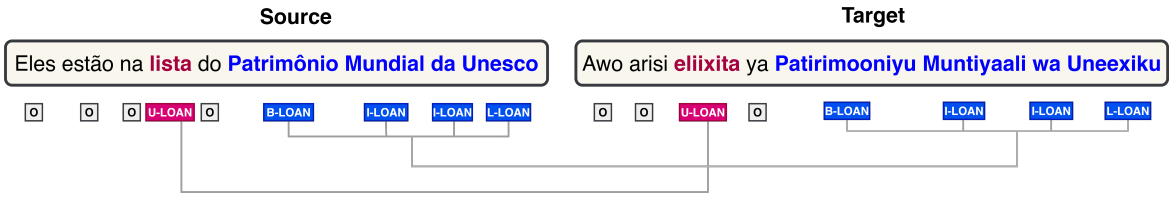

Figure 5: Labeling scheme for tagging source and target sentence tokens with "LOAN" or "O" category