Mitigating Tokenization-Induced Distance Distortion in Long-Context Multilingual Machine Translation

Khotso Selialia[1], Antoine Nzeyimana[1], Fatima M. Anwar[1],
[1]University of Massachusetts Amherst
{kselialia, anzeyimana, fanwar}@umass.edu

1 Abstract

Multilingual neural machine translation (MNMT) models degrade in performance as input context length increases, causing positional encoding schemes to misinterpret token distances. Existing absolute and relative positional encodings rely on fixed token indices and implicitly assume uniform semantic density, which breaks down for long-context inputs. We introduce DCARPE, a tokenization-aware adaptive positional encoding that conditions relative positional bias on input-level sequence length and fragmentation statistics, allowing the model to reinterpret positional distance when tokenization-induced inflation arises rather than semantic factors. Evaluations on JW300 and out-of-distribution FLORES-200 demonstrate consistent improvements in long-context robustness, achieving gains of up to **+10.81 ChrF++** and **+8.00 BLEU** over baselines.

2 Introduction

Multilingual Neural Machine Translation (MNMT) based on Transformer architectures enables translation across hundreds of languages within a single unified model. It achieves this through shared parameters and multilingual subword vocabularies (Vaswani et al., 2017; Arivazhagan et al., 2019; Fan et al., 2021; Johnson et al., 2017). This approach achieves strong generalization and near–human-level performance for many high-resource languages (Ott et al., 2018; Fernandes et al., 2023). However, translation quality remains highly uneven across languages (Kocmi et al., 2022; Nekoto et al., 2020). In particular, low-resource and morphologically rich languages consistently underperform. As a result, users of MNMT for these languages may experience unreliable or inaccurate translations, which can impede effective cross-lingual communication and diminish trust in automated translation systems (Goyal et al., 2022; Joshi et al., 2020).

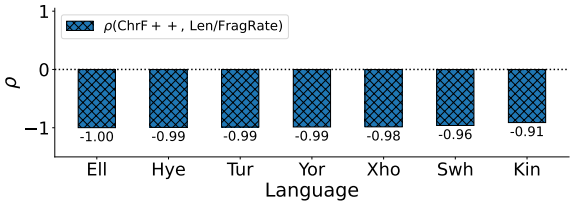

Figure 1: Correlation between translation performance (ChrF++) and effective context length (Len / FragRate) across languages. Performance degrades consistently as effective context increases.

A key source of performance degradation in MNMT arises from a previously underexplored interaction between language morphology and the *shared* subword tokenizer. MNMT relies on subword units—short character sequences produced by algorithms such as Byte Pair Encoding (BPE) (Zouhar et al., 2023) or SentencePiece (Kudo and Richardson, 2018)—to enable parameter sharing across languages. While shared tokenization is generally assumed to be benign or even beneficial, we find that it induces language-dependent distortions that only become apparent as context length increases. Morphologically rich languages encode tense, agreement, and case within complex word forms (Nzeyimana, 2024), which are frequently split into multiple subwords during tokenization. As a result, the same semantic content spans substantially more tokens than in morphologically simpler languages (Petrov et al., 2023). Token inflation compounds with longer context, spreading fixed semantic content across increasingly large token distances without adding information. As shown in Figure 1, translation quality exhibits a strong negative correlation with *effective* context length (Len/FragRate) across languages. This reveals a surprising failure mode: shared tokenization systematically reshapes the effective geometry of context in a language-dependent manner, causing translation robustness to degrade not

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36591–36602

merely with longer sequences, but specifically with tokenization-induced inflation of semantic distance as context grows.

The aforementioned limitation lies in how Transformer models encode positional information. In Transformer architectures, attention mechanisms determine which parts of an input sequence are most relevant by assigning higher weights to related tokens. Positional encoding schemes provide a notion of token order and distance, helping the model distinguish nearby from distant content. Despite advances such as Relative Positional Encodings (Shaw et al., 2018), Rotary Position Embeddings (Su et al., 2024), and Attention with Linear Biases (Press et al., 2021), most approaches remain fundamentally static: positional meaning is derived solely from a token's index, assuming all tokens contribute information at a uniform rate. This assumption breaks in morphologically rich languages, where a semantic unit is often split into multiple subword tokens. As positional encoding penalize distant tokens, fragmented information is increasingly treated as less related. We call this context-amplified translation degradation, where morphology-induced token inflation magnifies positional distance and undermines long-context robustness. This motivates our central research question: ***How can positional bias be adapted to tokenization-induced sequence inflation to improve long-context robustness in multilingual translation?***

To answer this question, we propose **D**ynamic **C**ontext **A**ware **R**elative **P**ositional **E**ncoding (DCARPE), a robustness-aware positional encoding mechanism. DCARPE adapts the interpretation of positional distance during tokenization. Unlike prior positional encodings that treat token distance as fixed geometry, DCARPE makes distance *input-adaptive* by conditioning the decay of relative bias on tokenization statistics. This lets the model adjust the strength of its penalty for distant tokens when subword fragmentation inflates sequence length. In this way, DCARPE attenuates positional penalties for long token distances caused by morphological segmentation, while preserving locality when token positions remain linguistically meaningful. As a result, DCARPE improves long-context robustness in multilingual translation. In designing and evaluating DCARPE, we make the following contributions.

- We discover *context-amplified translation degradation*, a previously uncharacterized failure mode in MNMT where the interaction between shared subword tokenization and language morphology inflates effective token distances, causing fixed positional encoding schemes to degrade as context length increases.
- We propose DCARPE [1], a tokenization-aware adaptive positional encoding that modulates distance-decay strength based on input-level statistics (sequence length and tokenization fragmentation rate). Unlike prior methods that treat token distance as fixed geometry, DCARPE learns a calibrated notion of distance that remains stable across languages with heterogeneous subword inflation, while retaining the simplicity of linear bias.
- We demonstrate the effectiveness of DCARPE on long-context MNMT across JW300 and FLORES-200 datasets, achieving consistent and statistically significant improvements of up to **+10.81 ChrF++** and **+8.00 BLEU** in context-robust translation quality for low-resource and morphologically rich languages, without introducing additional computational overhead.

3 Background and Related Works

Transformer and positional encodings. We assume the standard Transformer formulation with self-attention (Vaswani et al., 2017). Because self-attention is permutation-invariant, Transformers rely on positional encodings to represent token order and relative distance. Common approaches include absolute positional embeddings (Vaswani et al., 2017), relative positional encodings (Shaw et al., 2018; Dai et al., 2019), rotary embeddings (RoPE) (Su et al., 2024), and relative-bias formulations such as T5-style buckets (Raffel et al., 2020) and ALiBi (Press et al., 2021). More recent work explores position interpolation and scaling strategies to improve length extrapolation (Chen et al., 2023). While these methods enhance long-range modeling, their positional meaning remains a deterministic function of token index or token distance.

Tokenization, morphology, and uneven sequence inflation. Multilingual NMT systems typically employ shared subword vocabularies, such as BPE or SentencePiece, to support many languages with a fixed lexicon (Sennrich et al., 2016; Kudo and Richardson, 2018; Zouhar et al., 2023). However, languages differ substantially in how semantic content maps to subword units. For morphologically rich languages, a single semantic unit may be

[1] Code and data are available at https://github.com/khotso1186/dcarpe

realized as a longer sequence of subword tokens, resulting in higher subword fertility and *uneven sequence inflation* across languages (Toraman et al., 2023; Petrov et al., 2023). Recent analyses show that such tokenization effects disproportionately affect multilingual and low-resource settings, introducing systematic biases in representation length and granularity (Bostrom and Durrett, 2020; Rust et al., 2021; Wang et al., 2024; Mielke et al., 2021). As context grows, this inflation spreads semantically related content across increasingly distant token positions, breaking the implicit assumption that equal token distances correspond to comparable semantic distances and amplifying distance-based attenuation in long-context modeling (Khandelwal et al., 2018; Press et al., 2021).

Limitation of existing techniques. Prior work on positional encodings improves length extrapolation by modifying how token indices are mapped to bias or embeddings. However, these methods remain fundamentally token-index driven: positional meaning is a deterministic function of absolute position or raw token distance. In multilingual settings with a shared tokenizer, token distance is not a stable proxy for semantic distance, as identical meanings may span vastly different numbers of subword tokens across languages. Existing approaches do not model this tokenization-induced distortion and therefore systematically over-penalize long-range dependencies as context grows. This gap motivates tokenization-aware positional adaptation.

4 Methodology

In this section, we discuss the dataset curation, model training and evaluation procedures.

4.1 Model

We adopt the No Language Left Behind (NLLB) model (Costa-Jussà et al., 2022). This state-of-the-art multilingual Transformer uses a unified encoder–decoder architecture to support over 200 languages. NLLB employs a shared subword vocabulary and token-level processing, ensuring consistent representation across diverse languages. This design lets us systematically study how morphology-induced token fragmentation affects sequence length and translation quality under identical modeling conditions. Because NLLB performs well across both high- and low-resource languages, observed disparities can be attributed to tokenization and context length, rather than model capacity.

Table 1: Languages and parallel data used in our multilingual $X \rightarrow$English translation study. Counts indicate the number of sentence pairs after preprocessing and filtering.

Language	Lang ID	Speakers	# Parallel ($X \rightarrow$En)
Yorùbá	Yor	∼45M	122,554
Kinyarwanda	Kin	∼16M	77,271
Xhosa	Xho	∼19M	138,111
Swahili	Swh	∼98M	186,622
Turkish	Tur	∼85M	124,082
Armenian	Hye	∼7M	60,000
Greek	Ell	∼14M	150,000

4.2 Data

4.2.1 Training Data

We evaluate translation into English from African, European, and Asian languages (see Table 1), with English as the fixed target. This isolates the effects of source-language morphology and tokenization on model behavior. The selected languages vary in morphological complexity and subword fragmentation, from compact tokenizations (e.g., Swahili, Yoruba) to highly fragmented ones (e.g., Xhosa, Armenian). This range enables controlled analysis of how morphology-driven token inflation interacts with longer context and helps disentangle structural from data-scale effects.

To support this analysis, we use high-quality parallel data, mainly from JW300 (Agic and Vulic, 2019), following established practices in MNMT research (Orife et al., 2020; Emezue and Dossou, 2022). This provides a linguistically diverse, consistent foundation for evaluating long-context robustness across languages.

4.2.2 Dev/Test Sets

We evaluate translation quality on sentence pairs drawn from JW300 across a diverse set of source languages and assess out-of-distribution generalization using the FLORES-200 benchmark (Costa-Jussà et al., 2022). To study long-context behavior in a controlled manner, we follow prior document-level MT evaluation protocols (Gumma et al., 2024) and construct long-context inputs from sentence-level data.

Given a document consisting of sentences $(s_1, s_2, \ldots, s_n)$, we form overlapping evaluation windows by concatenating consecutive sentences:

$$W_k(i) = s_i \oplus s_{i+1} \oplus \cdots \oplus s_{i+k-1}, \quad k \in \{1,2,3,4\}, \quad i = 1, \ldots, n-k+1, \tag{1}$$

where $\oplus$ denotes concatenation. W_1 corresponds to single-sentence inputs, while larger k progressively

increase context length. This structured windowing enables precise measurement of how translation performance degrades as input context grows across languages.

4.2.3 Evaluation Metrics

We evaluate translation quality using the ChrF++ metric (Popović, 2017), a character n-gram–based score enhanced with word n-grams. We use ChrF++ because it correlates more consistently with human judgments than BLEU (Papineni et al., 2002) for morphologically rich and low-resource languages, where surface variation and subword segmentation are common (Gumma et al., 2024). Unlike word-level metrics, ChrF++ is less affected by tokenization artifacts and more effectively captures meaning retention under morphological variation, making it especially apt for long-context multilingual translation.

4.3 Proposed Fine-Tuning Strategy: DCARPE

Algorithm 1 DCARPE: Tokenization-Aware Adaptive Bias Fine-Tuning

1: **Input:** pretrained MNMT model with ALiBi bias $m_h r_{ij}$; tokenizer T; training pairs $\{(\mathbf{X}, \mathbf{Y})\} \sim P_S$; optimizer Opt; steps T; batch size B
2: **Params:** $\boldsymbol{\theta}$; $\boldsymbol{\phi} = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2, \mathbf{u}\}$
3: **Output:** fine-tuned $(\boldsymbol{\theta}, \boldsymbol{\phi})$
4: **for** $t = 1$ **to** T **do**
5: Sample $\mathcal{B} = \{(\mathbf{X}, \mathbf{Y})\}_{b=1}^{B}$; tokenize each $\mathbf{X}$ into $(x_1, \ldots, x_n)$
6: **for all** $\mathbf{X} \in \mathcal{B}$ **do**
7: $\mathrm{Len}(\mathbf{X}) \leftarrow |T(\mathbf{X})|$; $\mathrm{FragRate}(\mathbf{X}) \leftarrow |T(\mathbf{X})|/\#\mathrm{words}(\mathbf{X})$
8: $\mathbf{z}(\mathbf{X}) \leftarrow [\mathrm{Len}(\mathbf{X}), \mathrm{FragRate}(\mathbf{X})]$; normalize
9: $\mathbf{c}(\mathbf{X}) \leftarrow \sigma(\mathbf{W}_2\, \phi(\mathbf{W}_1\, \mathrm{Norm}(\mathbf{z}(\mathbf{X})) + \mathbf{b}_1) + \mathbf{b}_2)$
10: $\boldsymbol{\lambda}(\mathbf{X}) \leftarrow \mathbf{U}^\top \mathbf{c}(\mathbf{X})$
11: **end for**
12: **for all** $(\mathbf{X}, \mathbf{Y}) \in \mathcal{B}$ **do**
13: **for** $i = 1$ **to** n **do**
14: **for** $j = 1$ **to** n **do**
15: $r_{ij} \leftarrow i - j$
16: $b_{ij}^{\mathrm{adap}}(\mathbf{X}) \leftarrow \lambda^{(h)}(\mathbf{X}) \cdot r_{ij}$
17: $\ell_{ij}^{(h)} \leftarrow \dfrac{\mathbf{q}_i^{(h)} \mathbf{k}_j^{(h)\top}}{\sqrt{d_h}} + b_{ij}^{\mathrm{adap}}(\mathbf{X})$
18: **end for**
19: **end for**
20: **end for**
21: Compute $\mathcal{L}_{\mathrm{MT}}(\boldsymbol{\theta}, \boldsymbol{\phi})$; update $(\boldsymbol{\theta}, \boldsymbol{\phi}) \leftarrow \mathrm{Opt}(\nabla_{\boldsymbol{\theta}, \boldsymbol{\phi}} \mathcal{L}_{\mathrm{MT}})$
22: **end for**

We propose DCARPE, a lightweight fine-tuning strategy that improves long-context robustness in MNMT by adapting positional bias to each sequence's tokenized structure. DCARPE conditions relative positional bias on input tokenization statistics (Algorithm 1) and integrating it into the Transformer attention with negligible overhead, preserving the original attention computation.

4.3.1 Tokenization-Aware Context and Gating

Tokenization imbalance manifests along two complementary dimensions: (i) *absolute token length*, and (ii) *fragmentation*, i.e., how many subword tokens represent each word. Longer token sequences increase the range over which information must be integrated, while high fragmentation spreads semantic units across multiple tokens. Both effects become more severe in long-context settings, where semantically related content is pushed farther apart in token space.

Tokenization context vector. For each input sequence $\mathbf{X}$, we construct a tokenization context vector

$$\mathbf{z}(\mathbf{X}) = \big[\mathrm{Len}(\mathbf{X}),\, \mathrm{FragRate}(\mathbf{X})\big] \in \mathbb{R}^2, \quad (2)$$

where $\mathrm{Len}(\mathbf{X})$ denotes the number of subword tokens produced by the tokenizer, and $\mathrm{FragRate}(\mathbf{X})$ is the ratio of subword tokens to word count.

We use $\mathrm{Len}(\mathbf{X})$ because positional bias is applied as a function of token distance, which grows with token length. In addition, we use $\mathrm{FragRate}(\mathbf{X})$ because length alone conflates semantic verbosity with tokenizer over-segmentation; fragmentation isolates the degree to which meaning is distributed across subwords. Together, these two statistics capture both the *scale* and the *cause* of token inflation, yielding a compact signal that generalizes across languages and domains.

Normalization and gating network. Following Algorithm 1 (Lines 6–10), we normalize $\mathbf{z}(\mathbf{X})$ and map it through a compact Multilayer perceptron (MLP) with gating (Taud and Mas, 2017):

$$\mathbf{c}(\mathbf{X}) = \sigma\Big(\mathbf{W}_2\, \phi\big(\mathbf{W}_1\, \mathrm{Norm}(\mathbf{z}(\mathbf{X})) + \mathbf{b}_1\big) + \mathbf{b}_2\Big), \quad (3)$$

where $\mathbf{W}_1 \in \mathbb{R}^{m \times 2}$, $\mathbf{b}_1 \in \mathbb{R}^m$, $\mathbf{W}_2 \in \mathbb{R}^{h \times m}$, $\mathbf{b}_2 \in \mathbb{R}^h$, $\phi(\cdot)$ is a pointwise nonlinearity (e.g., GELU), and $\sigma(\cdot)$ is the sigmoid function. The output $\mathbf{c}(\mathbf{X}) \in \mathbb{R}^h$ is a bounded, tokenization-aware control signal.

Design choice (MLP). The interaction between $\mathrm{Len}(\mathbf{X})$ and $\mathrm{FragRate}(\mathbf{X})$ is not linear (e.g., long sequences are not necessarily highly fragmented, and vice versa). A small feed-forward network captures these nonlinear interactions while remaining parameter-efficient, consistent with prior work

on lightweight adaptation modules (Houlsby et al., 2019; Pfeiffer et al., 2020; Li and Liang, 2021).

Design choice (gating with sigmoid). We use a sigmoid gate to keep modulation bounded and smooth, preventing unstable or extreme positional shifts during fine-tuning. Gating is widely used for conditional computation and controlled adaptation because it allows the model to adjust behavior without overwriting learned representations (Shazeer et al., 2017; Pfeiffer et al., 2020).

4.3.2 Adaptive Positional Bias

The control signal $\mathbf{c}(\mathbf{X})$ quantifies the required degree of positional bias adaptation. However, a structured bias form is also necessary to maintain the advantages of relative position modeling. We therefore build on ALiBi (Press et al., 2021), which uses relative offsets $r_{ij} = i - j$.

Baseline ALiBi bias. For each Transformer head h, ALiBi adds a static linear bias $b_{ij}^{\text{ALiBi},(h)} = m_h r_{ij}$, where m_h is a fixed slope. ALiBi provides a simple inductive bias for length extrapolation because it expresses positional preference directly as a function of relative offset. However, because m_h is fixed, ALiBi cannot recalibrate distance when tokenization inflates offsets unevenly across languages and contexts.

Input-conditioned slope. To make the positional bias adaptive, DCARPE replaces the static slope with an input-conditioned scalar $\boldsymbol{\lambda}$ (Algorithm 1, Line 10):

$$\boldsymbol{\lambda}(\mathbf{X}) = \mathbf{U}\,\mathbf{c}(\mathbf{X}), \quad \mathbf{U} \in \mathbb{R}^{H \times h}, \tag{4}$$

We then define the adaptive bias

$$b_{ij}^{\text{adaptive}}(\mathbf{X}) = \lambda^{(h)}(\mathbf{X})\, r_{ij}. \tag{5}$$

Intuition behind DCARPE. DCARPE learns to reduce positional bias penalty (via $\boldsymbol{\lambda}$) when tokenization statistics indicate inflation, and to preserve or strengthen locality when token distances remain linguistically meaningful.

4.3.3 Integrating DCARPE into Attention

We incorporate the adaptive bias $b_{ij}^{\text{adaptive}}(\mathbf{X})$ into attention in a way that is consistent with the background formulation in section 3. Let $\mathbf{q}_i^{(h)}$ and $\mathbf{k}_j^{(h)}$ denote the query and key vectors for head h at positions i and j. The baseline scaled dot-product score is $e_{ij}^{(h)} = \frac{\left(\mathbf{q}_i^{(h)}\right)^\top \mathbf{k}_j^{(h)}}{\sqrt{d_k}}$. With ALiBi, the score becomes $e_{ij}^{(h)} + m_h r_{ij}$; DCARPE replaces this static term with the adaptive bias (Algorithm 1, Lines 14–17):

$$e_{ij}^{(h)}(\mathbf{X}) = \frac{(\mathbf{q}_i^{(h)})^\top \mathbf{k}_j^{(h)}}{\sqrt{d_k}} + \lambda^{(h)}(\mathbf{X})\, r_{ij}. \tag{6}$$

The attention weights are then computed as usual:

$$\alpha_{ij}^{(h)}(\mathbf{X}) = \frac{\exp\left(e_{ij}^{(h)}(\mathbf{X})\right)}{\sum_{m=1}^{n} \exp\left(e_{im}^{(h)}(\mathbf{X})\right)}. \tag{7}$$

Design choice (minimal intervention). We modify only the positional bias term and keep the dot-product and softmax unchanged. This preserves architectural compatibility and ensures that any gains come from improved distance calibration rather than changes to the underlying attention operator.

Training. DCARPE is trained by standard MT fine-tuning: we jointly optimize the base model parameters $\boldsymbol{\theta}$ and adapter parameters $\boldsymbol{\phi} = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2, \mathbf{u}\}$ using the translation loss $\mathcal{L}_{\text{MT}}$ (Algorithm 1, Line 21). Because the adapter is small, the method adds minimal compute and memory overhead while providing an explicit mechanism to correct tokenization-induced positional distortion.

5 Experiments

All experiments are conducted in a Google Colab environment equipped with a single NVIDIA® A100 GPU (40 GB HBM2), backed by a virtualized Intel® Xeon® CPU (2.2–3.5 GHz) and 25 GB of RAM. All implementations are based on PyTorch, and GPU acceleration is utilized for both training and inference. Training is performed with early stopping, and all experiments use publicly available datasets and standard evaluation protocols; hyperparameters details are provided in Table 2.

6 Results and Discussions

This section examines whether DCARPE improves translation robustness as context length increases.

6.1 Long-Context Robustness

We evaluate long-context translation robustness by comparing ALiBi, RoPE, and DCARPE on JW300 across increasing context lengths under both full fine-tuning and parameter-efficient fine-tuning (LoRA). Table 3 reports ChrF++ across all settings, covering seven languages with diverse morphological profiles.

Table 2: Training hyperparameters used across all PE variants.

Hyperparameter	**Value**
Fine-Tuning	
Batch Size	32
Eval Batch Size	16
Max Source Length	2048
Max Target Length	128
Max New Tokens (decode)	128
Num Beams	5
Optimizer	AdamW
Learning Rate	5×10^{-5}
Adam Betas	(0.9, 0.999)
LoRA Rank r	16, 32
LoRA α	32
LoRA Dropout	0.05
MLP hidden size	64
MLP depth	2 layers
Primary Metric	ChrF++
GPU	A100 (40 GB, 1×)

Table 3: ChrF++ on JW300 across context lengths for full and LoRA fine-tuning of NLLB with ALiBi, RoPE, and DCARPE. Best result per fine-tuning block per row is **bolded**.

		Full Fine-Tuning			**LoRA Fine-Tuning**		
Lang	**Context**	**ALiBi**	**RoPE**	**DCARPE**	**ALiBi**	**RoPE**	**DCARPE**
Yor→Eng	1-Sentence	37.01	39.73	**51.52**	28.86	41.46	**48.39**
	2-Sentence	25.75	32.36	**44.71**	20.00	35.01	**39.54**
	3-Sentence	19.90	27.95	**40.83**	17.76	30.32	**34.43**
	4-Sentence	18.09	27.38	**37.15**	16.25	30.45	**33.89**
Xho→Eng	1-Sentence	46.36	52.99	**58.12**	36.69	52.90	**57.35**
	2-Sentence	33.38	47.21	**51.28**	25.68	**48.89**	45.57
	3-Sentence	26.08	42.97	**48.70**	20.22	**47.17**	40.53
	4-Sentence	22.89	34.09	**41.25**	20.75	**39.82**	34.12
Kin→Eng	1-Sentence	39.11	41.80	**48.28**	32.20	43.24	**46.78**
	2-Sentence	28.98	34.96	**38.99**	23.13	35.49	**38.08**
	3-Sentence	26.25	**34.64**	33.79	19.99	31.10	**33.60**
	4-Sentence	23.10	31.23	**34.22**	17.66	29.42	**31.20**
Tur→Eng	1-Sentence	42.85	47.30	**48.77**	42.23	46.70	**47.81**
	2-Sentence	34.20	39.72	**45.75**	**41.59**	40.51	44.74
	3-Sentence	24.82	33.01	**38.46**	33.60	37.01	**40.15**
	4-Sentence	22.75	31.30	**34.26**	35.61	30.67	**38.74**
Hye→Eng	1-Sentence	26.41	28.98	**31.14**	30.55	29.54	**31.25**
	2-Sentence	20.70	23.78	**27.51**	25.09	24.46	**28.36**
	3-Sentence	17.96	19.95	**24.20**	22.40	20.85	**27.36**
	4-Sentence	14.64	16.69	**22.14**	17.70	16.68	**23.67**
Ell→Eng	1-Sentence	44.13	51.92	**54.52**	53.91	54.31	**54.60**
	2-Sentence	32.86	42.07	**47.55**	44.94	45.61	**48.85**
	3-Sentence	26.81	34.14	**40.95**	39.65	38.50	**42.57**
	4-Sentence	22.69	27.98	**35.79**	34.02	32.75	**39.06**
Swh→Eng	1-Sentence	52.29	58.26	**64.56**	41.04	59.54	**65.81**
	2-Sentence	59.28	52.30	**59.28**	30.67	49.65	**53.36**
	3-Sentence	34.25	48.55	**54.17**	27.35	47.24	**50.27**
	4-Sentence	28.75	45.26	**53.45**	22.58	45.51	**47.74**

At short context ($k = 1$), DCARPE consistently matches or outperforms both ALiBi and RoPE, indicating that adaptive positional reweighting preserves sentence-level translation quality. Gains are already apparent at $k = 1$ for languages such as Kin→Eng, Ell→Eng, and Swh→Eng, suggesting that tokenization-induced positional distortion affects performance even before multi-sentence context is introduced.

As context length increases ($k = 2$–4), the limitations of fixed positional encodings become increasingly evident. Under RoPE, ChrF++ scores decline sharply as additional context is prepended. This pattern is consistent with Figure 1, which demonstrates that translation performance deteriorates with greater effective context length resulting from subword fragmentation. DCARPE addresses this issue by adapting the strength of relative positional bias according to input-level tokenization statistics. As shown in Figure 2, the learned reweighting coefficient $\boldsymbol{\lambda}$ does not merely reflect sequence inflation; rather, it serves as a stabilizing factor that reduces excessive distance penalties when long token spans are produced by morphological segmentation instead of genuine semantic separation. This separation enables attention mechanisms to remain aligned with semantic structure as context expands. The results in Table 3 indicate that adaptive positional reweighting provides consistent and often statistically significant improvements over both ALiBi and RoPE across various languages and context lengths. Improvements are especially robust and stable for Swahili, Greek, Armenian, and Kinyarwanda, languages characterized by higher subword fragmentation, where DCARPE preserves translation quality even at $k = 4$. Collectively, these findings indicate that long-context behavior in MNMT is strongly influenced by tokenization-induced distance distortion, and that making positional bias responsive to input structure supports more reliable document-level translation across morphologically diverse languages.

6.2 Parameter-Efficient Fine-Tuning.

Table 3 also presents LoRA results using identical adapter configurations ($rank = 16$). As anticipated, LoRA's reduced adaptation capacity results in a consistent absolute decrease in performance compared to full fine-tuning across all methods. Nevertheless, DCARPE consistently outperforms ALiBi and RoPE with the same LoRA budget across all context lengths. This finding indicates that the improvements achieved by DCARPE are attributable

Table 4: ChrF++ on FLORES-200 across context lengths using ALiBi, RoPE, and DCARPE. Rows Por, Spa, Fra are evaluated zero-shot.

Language	Context	ALiBi	ROPE	DCARPE	Δ vs. ALiBi		Δ vs. ROPE	
					Δ	p	Δ	p
Yor→Eng	1-Sentence	27.68	32.12	**37.98**	+10.30	0.009	+5.86	0.009
	2-Sentence	22.10	29.36	**34.77**	+12.67	0.009	+5.41	0.009
	3-Sentence	18.77	26.67	**32.00**	+13.23	0.009	+5.33	0.009
	4-Sentence	17.06	22.54	**28.36**	+11.30	0.009	+5.82	0.009
Xho→Eng	1-Sentence	44.22	50.25	**53.18**	+8.96	0.009	+2.93	0.009
	2-Sentence	33.15	46.45	**49.88**	+16.73	0.009	+3.43	0.009
	3-Sentence	26.66	41.96	**46.11**	+19.45	0.009	+4.15	0.009
	4-Sentence	23.04	36.34	**40.97**	+17.93	0.009	+4.63	0.009
Kin→Eng	1-Sentence	39.23	43.92	**50.09**	+10.86	0.009	+6.17	0.009
	2-Sentence	29.28	41.85	**47.53**	+18.25	0.009	+5.68	0.009
	3-Sentence	23.90	39.39	**45.08**	+21.18	0.009	+5.69	0.009
	4-Sentence	20.74	32.96	**40.17**	+19.43	0.009	+7.21	0.009
Tur→Eng	1-Sentence	50.09	56.45	**59.37**	+9.28	0.009	+2.92	0.009
	2-Sentence	37.79	43.56	**49.16**	+11.37	0.009	+5.60	0.009
	3-Sentence	31.30	35.16	**43.93**	+12.63	0.009	+8.77	0.009
	4-Sentence	26.49	30.37	**40.43**	+13.94	0.009	+10.06	0.009
Hye→Eng	1-Sentence	47.49	52.70	**58.59**	+11.10	0.009	+5.89	0.009
	2-Sentence	35.11	36.95	**47.64**	+12.53	0.009	+10.69	0.009
	3-Sentence	28.58	28.99	**40.56**	+11.98	0.009	+11.57	0.009
	4-Sentence	22.49	23.85	**36.98**	+14.49	0.009	+13.13	0.009
Ell→Eng	1-Sentence	50.94	57.90	**59.94**	+9.00	0.009	+2.04	0.009
	2-Sentence	38.36	45.49	**49.50**	+11.14	0.009	+4.01	0.009
	3-Sentence	32.29	36.75	**44.53**	+12.24	0.009	+7.78	0.009
	4-Sentence	25.89	30.01	**40.76**	+14.87	0.009	+10.75	0.009
Por→Eng	1-Sentence	63.17	67.03	**69.94**	+6.77	0.009	+2.91	0.009
	2-Sentence	50.14	63.94	**67.97**	+17.83	0.009	+4.03	0.009
	3-Sentence	39.76	62.81	**67.77**	+28.01	0.009	+4.96	0.009
	4-Sentence	32.74	57.38	**63.00**	+30.26	0.009	+5.62	0.009
Spa→Eng	1-Sentence	53.16	56.18	**58.74**	+5.58	0.009	+2.56	0.009
	2-Sentence	43.16	55.15	**57.44**	+14.28	0.009	+2.29	0.009
	3-Sentence	34.69	53.77	**57.10**	+22.41	0.009	+3.33	0.009
	4-Sentence	29.35	49.26	**52.67**	+23.32	0.009	+3.41	0.009
Fra→Eng	1-Sentence	58.74	63.92	**66.28**	+7.54	0.009	+2.36	0.009
	2-Sentence	44.70	60.74	**63.83**	+19.13	0.009	+3.09	0.009
	3-Sentence	35.20	58.69	**62.90**	+27.70	0.009	+4.21	0.009
	4-Sentence	28.43	51.44	**57.57**	+29.14	0.009	+6.13	0.009
Swh→Eng	1-Sentence	48.73	57.03	**60.44**	+11.71	0.009	+3.41	0.009
	2-Sentence	36.07	54.40	**56.88**	+20.81	0.009	+2.48	0.009
	3-Sentence	28.78	50.13	**54.38**	+25.60	0.009	+4.25	0.009
	4-Sentence	24.14	44.48	**50.43**	+26.29	0.009	+5.95	0.009

to its positional adaptation mechanism rather than an increase in trainable parameters. In contrast to generic LoRA adaptation, which adjusts projection weights without positional awareness, DCARPE's input-conditioned slope $\boldsymbol{\lambda}$ introduces a complementary inductive bias that addresses tokenization-induced distance distortion. These results demonstrate that DCARPE can be integrated into parameter-efficient pipelines to enhance long-context robustness in scenarios where full fine-tuning is impractical.

6.3 Out-of-Distribution Generalization

We evaluate whether models fine-tuned on JW300 generalize to longer contexts under both in-family and cross-lingual out-of-distribution settings, using the FLORES-200 benchmark, which introduces both domain and distribution shift.

Low-resource languages. Table 4 reports ChrF++ on FLORES-200 as context length increases, comparing ALiBi, RoPE, and DCARPE across seven languages with varied morphological complexity. Across most language pairs, DCARPE outperforms both baselines even at short context ($k = 1$). As context grows, DCARPE's lead widens: unlike the baselines, which decline monotonically with additional context, DCARPE consistently maintains or improves translation quality. This effect is especially pronounced and statistically significant ($p = 0.009$ throughout) for Armenian, Greek, Swahili, and Yoruba, where fixed positional encodings degrade sharply.

Zero-shot cross-lingual transfer to high-resource languages. A key question left open

Table 5: BLEU / COMET on JW300 across context lengths using ALiBi, RoPE, and DCARPE. Δ values are for BLEU only; all improvements are statistically significant ($p = 0.009$).

Lang	Context	ALiBi	RoPE	DCARPE	Δ_{ALiBi}	Δ_{RoPE}
Yor→Eng	1-Sentence	18.30/0.63	18.38/0.61	**31.68**/0.71	+13.38	+13.30
	2-Sentence	7.65/0.54	11.10/0.54	**24.31**/0.65	+16.66	+13.21
	3-Sentence	3.26/0.50	6.94/0.50	**20.61**/0.60	+17.35	+13.67
	4-Sentence	2.78/0.47	10.64/0.46	**14.88**/0.57	+12.10	+4.24
Xho→Eng	1-Sentence	25.52/0.74	27.26/0.74	**38.70**/0.80	+13.18	+11.44
	2-Sentence	11.60/0.63	17.72/0.65	**30.68**/0.73	+19.08	+12.96
	3-Sentence	5.62/0.57	17.24/0.61	**25.76**/0.71	+20.14	+8.52
	4-Sentence	3.04/0.56	11.46/0.57	**19.20**/0.64	+16.16	+7.74
Kin→Eng	1-Sentence	20.62/0.62	17.05/0.64	**27.16**/0.68	+6.54	+10.11
	2-Sentence	9.46/0.54	12.17/0.55	**20.89**/0.60	+11.43	+8.72
	3-Sentence	6.72/0.51	12.08/0.54	**16.70**/0.55	+9.98	+4.62
	4-Sentence	4.58/0.48	8.43/0.49	**15.98**/0.54	+11.40	+7.55
Tur→Eng	1-Sentence	20.61/0.78	23.75/0.80	**25.11**/0.80	+4.50	+1.36
	2-Sentence	11.74/0.90	15.49/0.73	**15.49**/0.78	+3.75	+0.00
	3-Sentence	6.13/0.62	9.68/0.68	**15.32**/0.71	+9.19	+5.64
	4-Sentence	4.08/0.58	6.11/0.64	**12.37**/0.71	+8.29	+6.26
Hye→Eng	1-Sentence	9.15/0.65	11.28/0.66	**13.23**/0.67	+4.08	+1.95
	2-Sentence	4.73/0.58	6.14/0.61	**9.02**/0.65	+4.29	+2.88
	3-Sentence	2.34/0.54	3.38/0.56	**5.91**/0.62	+3.57	+2.53
	4-Sentence	1.43/0.50	1.98/0.51	**4.27**/0.60	+2.84	+2.29
Ell→Eng	1-Sentence	23.72/0.76	31.57/0.79	**34.96**/0.81	+11.24	+3.39
	2-Sentence	12.40/0.65	20.43/0.70	**27.18**/0.73	+14.78	+6.75
	3-Sentence	8.20/0.58	12.79/0.62	**19.71**/0.68	+11.51	+6.92
	4-Sentence	5.70/0.53	7.77/0.55	**15.12**/0.65	+9.42	+7.35
Swh→Eng	1-Sentence	30.98/0.75	31.52/0.77	**47.46**/0.83	+16.48	+15.94
	2-Sentence	15.06/0.65	25.54/0.67	**39.81**/0.76	+24.75	+14.27
	3-Sentence	10.80/0.61	24.64/0.64	**32.88**/0.71	+22.08	+8.24
	4-Sentence	7.81/0.54	20.86/0.58	**20.86**/0.70	+13.05	+0.00

Table 6: ChrF++ on JW300 across context lengths using ALiBi, RoPE, and DCARPE. Base model is M2M-100. All improvements are statistically significant ($p = 0.009$).

Lang	Context	ALiBi	RoPE	DCARPE	Δ_{ALiBi}	Δ_{RoPE}
Yor→Eng	1-Sentence	6.96	5.70	**19.19**	+12.23	+13.49
	2-Sentence	5.65	5.02	**17.96**	+12.31	+12.94
	3-Sentence	5.45	6.06	**16.03**	+10.58	+9.97
	4-Sentence	5.46	6.60	**16.26**	+10.80	+9.66
Xho→Eng	1-Sentence	5.70	7.18	**21.18**	+15.48	+14.00
	2-Sentence	4.94	7.01	**18.54**	+13.60	+11.53
	3-Sentence	3.85	6.37	**14.94**	+11.09	+8.57
	4-Sentence	3.80	3.53	**13.61**	+9.81	+10.08
Por→Eng	1-Sentence	21.92	18.09	**41.72**	+19.80	+23.63
	2-Sentence	5.77	11.77	**26.86**	+21.09	+15.09
	3-Sentence	5.18	9.91	**24.04**	+18.86	+14.13
	4-Sentence	5.28	11.56	**21.43**	+16.15	+9.87
Spa→Eng	1-Sentence	10.21	15.34	**34.78**	+24.57	+19.44
	2-Sentence	5.89	9.69	**26.41**	+20.52	+16.72
	3-Sentence	5.40	8.43	**22.31**	+16.91	+13.88
	4-Sentence	5.02	8.61	**18.35**	+13.33	+9.74
Fra→Eng	1-Sentence	13.28	22.20	**40.93**	+27.65	+18.73
	2-Sentence	7.24	13.61	**29.04**	+21.80	+15.43
	3-Sentence	5.96	13.60	**27.47**	+21.51	+13.87
	4-Sentence	5.16	12.32	**20.08**	+14.92	+7.76
Swh→Eng	1-Sentence	10.64	27.81	**33.91**	+23.27	+6.10
	2-Sentence	6.27	20.18	**28.30**	+22.03	+8.12
	3-Sentence	7.52	23.21	**25.74**	+18.22	+2.53
	4-Sentence	7.01	9.97	**20.49**	+13.48	+10.52

Table 7: DCARPE conditioning signals on ChrF++ across context lengths. The 2-layer MLP is the full model. We report ChrF++ for ablations relative to the full model.

Dataset	Context	3-layer MLP	2-layer MLP	1-layer MLP
Yor→Eng	1-Sentence	52.02	51.52	49.34
	2-Sentence	47.74	44.71	42.10
	3-Sentence	39.16	40.83	37.16
	4-Sentence	33.74	37.15	32.72
Xho→Eng	1-Sentence	60.23	58.12	57.56
	2-Sentence	52.93	51.28	46.80
	3-Sentence	45.32	48.70	43.36
	4-Sentence	40.65	41.25	38.51
Kin→Eng	1-Sentence	46.83	48.28	45.48
	2-Sentence	41.80	38.99	37.27
	3-Sentence	38.32	33.79	31.03
	4-Sentence	36.07	34.22	29.81
Swh→Eng	1-Sentence	64.81	64.56	65.21
	2-Sentence	59.83	59.28	52.14
	3-Sentence	54.35	54.17	50.68
	4-Sentence	52.11	53.45	45.75

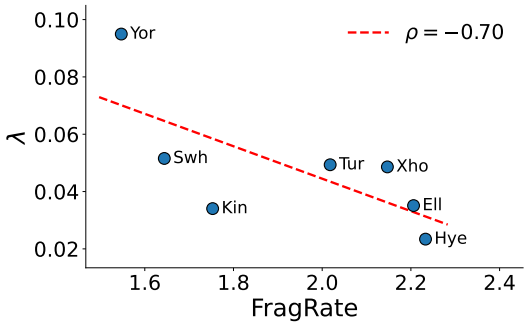

Figure 2: Correlation between DCARPE's $\boldsymbol{\lambda}$ at the first encoder layer and fragmentation rate across languages. The negative correlation indicates that $\boldsymbol{\lambda}$ acts as a stabilizing correction that attenuates excessive fragmentation. The correlation varies per language family (B: Bantu, Y: Yoruboid, T: Turkic, IE: Indo-European), which indicates dependency on morphological typology.

by prior work is whether tokenization-aware positional adaptation is specific to the languages seen during fine-tuning, or whether the learned correction transfers to structurally different, high-resource languages. To probe this directly, we evaluate French, Spanish, and Portuguese to English using the *same* models fine-tuned only on Swahili, without any additional language-specific adaptation. This zero-shot setup tests whether DCARPE captures a language-agnostic correction for tokenization-induced distortion, rather than language-specific morphological features.

Table 4 reports ChrF++ on FLORES-200 across context lengths for Fra, Spa, and Por. DCARPE outperforms both ALiBi and RoPE across all three languages and all context lengths, with gains that *grow* as context increases—the hallmark of a positional correction rather than a lexical one. No-

tably, RoPE degrades sharply on Spa and Por at $k = 4$, while DCARPE remains stable. These results indicate that the adaptive slope $\boldsymbol{\lambda}$, conditioned on input tokenization statistics rather than language identity, learns a structural correction that transfers across typologically diverse language families. Together with the low-resource results, this demonstrates that DCARPE's inductive bias is genuinely language-agnostic and enables robust long-context translation across both resource levels and linguistic domains.

6.4 Ablation Study

6.4.1 Effect of MLP Depth in DCARPE

Table 7 analyzes the role of conditioning MLP depth in controlling the strength of positional bias adaptation in DCARPE. For Kinyarwanda–English, a 2-layer MLP consistently outperforms shallower variants across all context lengths. However, removing conditioning depth results in significant performance drops. This indicates that expressive, non-linear conditioning is crucial when tokenization strongly inflates sequence length and substantially distorts the alignment between token distance and semantic distance.

In contrast, Turkish–English exhibits different behavior: simpler conditioning variants (1-layer or 0-layer MLP) match or exceed the full model at longer contexts. We hypothesize that Turkish morphology aligns well with subword tokenization, leading to more stable token-to-meaning correspondence. Thus, in this regime, deeper conditioning can over-adjust positional bias, while lighter modulation remains sufficient.

Taken together, these results support a central design principle of DCARPE: the magnitude of tokenization-induced distortion, rather than morphological richness alone, determines the optimal degree of positional bias adaptation.

6.4.2 Complementary metrics.

To ensure that gains are not artefacts of ChrF++'s surface-level character matching, Table 5 additionally reports BLEU and COMET on JW300. Under BLEU, DCARPE outperforms ALiBi and RoPE across all languages and context lengths, with especially large gains for Swahili and Greek. Notably, RoPE collapses to near-zero BLEU for Armenian and Greek at $k \geq 3$, whereas DCARPE maintains usable translation quality. COMET scores follow the same ranking: DCARPE achieves the highest semantic adequacy across all settings, confirming that ChrF++ gains reflect genuine meaning preservation rather than surface-level character overlap. Together, the convergence of all three metrics—ChrF++, BLEU, and COMET—provides strong, multi-faceted evidence that DCARPE consistently improves long-context translation quality.

6.4.3 Generalization to a second backbone.

To assess whether DCARPE's gains depend on the NLLB architecture, Table 6 reports ChrF++ when the same method is applied to M2M-100 (Fan et al., 2021), a structurally distinct multilingual encoder–decoder model. Despite different pretraining data, vocabulary, and positional design, DCARPE consistently outperforms ALiBi and RoPE across all tested languages (Yorùbá, Xhosa, Portuguese, Spanish, French, and Swahili) and all context lengths, with statistically significant gains ($p = 0.009$ throughout). This confirms that DCARPE's inductive bias—conditioning the positional slope on input tokenization statistics—is not architecture-specific, but generalizes across encoder–decoder models with different backbones. We note that absolute ChrF++ values are lower on M2M-100 than on NLLB, consistent with NLLB's stronger multilingual pretraining, but the *relative* advantage of DCARPE is preserved or amplified across all conditions.

7 Conclusion

We addressed a fundamental limitation of long-context MNMT: the mismatch between token-based positional encoding and the inflation of sequence length caused by tokenization. We showed that fixed positional encodings incorrectly equate token distance with semantic distance, leading to degraded context understanding as the number of tokens grows. To resolve this, we introduced DCARPE, a tokenization-aware adaptive positional encoding method that adjusts relative positional bias based on sequence length and fragmentation rate. DCARPE preserves important long-range connections between tokens without changing the model's architecture. Across JW300 and FLORES-200, DCARPE achieves consistent and statistically significant improvements under full fine-tuning, LoRA, and zero-shot cross-lingual transfer, and generalizes across two encoder–decoder backbones (NLLB and M2M-100). These findings establish tokenization-aware positional adaptation as a practical, architecture-agnostic step toward robust long-context multilingual translation.

8 Limitations

While our results demonstrate consistent improvements from tokenization-aware adaptive positional bias across multiple languages, backbones, and evaluation settings, several limitations remain. Our primary evaluation uses NLLB as the base model; although we additionally validate DCARPE on M2M-100, extending experiments to decoder-only LLMs remains important future work. Decoder-only architectures process source and target jointly in a single sequence, changing both the positional geometry and the scope over which fragmentation statistics are computed, making direct application non-trivial. Additionally, the current computation of $\mathrm{FragRate}$ requires explicit word boundaries, limiting direct applicability to non-segmented languages such as Chinese, Japanese, and Korean; future work could substitute character-level or morphological segmentation proxies. Our evaluation relies primarily on automatic metrics (ChrF++, BLEU, COMET), and human evaluation of discourse-level coherence in very long contexts remains desirable.

9 Acknowledgments

We thank the ARR reviewers for the insightful feedback. This work was supported by NSF award 2452819.

References

Željko Agic and Ivan Vulic. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2978–2988.

Chris C Emezue and Bonaventure FP Dossou. 2022. Mmtafrica: Multilingual machine translation for african languages. *arXiv preprint arXiv:2204.04306*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. In *International Conference on Machine Learning*, pages 10053–10071. PMLR.

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.

Varun Gumma, Pranjal A Chitale, and Kalika Bali. 2024. Towards inducing long-context abilities in multilingual neural machine translation models. *arXiv preprint arXiv:2408.11382*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, and 1 others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel,

Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, and 1 others. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, and 1 others. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. *arXiv preprint arXiv:2404.02392*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, and 1 others. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

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.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Hind Taud and Jean-Franccois Mas. 2017. Multilayer perceptron (mlp). In *Geomatic approaches for modeling land change scenarios*, pages 451–455. Springer.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dixuan Wang, Yanda Li, Junyuan Jiang, Zepeng Ding, Ziqin Luo, Guochao Jiang, Jiaqing Liang, and Deqing Yang. 2024. Tokenization matters! degrading large language models through challenging their tokenization. *arXiv preprint arXiv:2405.17067*.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. 2023. A formal perspective on byte-pair encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 598–614.