An Information-Theoretic Approach to Reducing Fertility in LLMs for Manipuri Machine Translation

Telem Joyson Singh, Sanasam Ranbir Singh, Priyankoo Sarmah
Indian Institute of Technology Guwahati, Assam, India
{tjoyson, ranbir, priyankoo}@iitg.ac.in

Abstract

Large language models (LLMs) have transformed machine translation, yet they have a high subword fertility issue for low-resource languages, which leads to slow inference speed and increased costs. While vocabulary expansion via continual pre-training is a common solution, it often degrades translation quality and requires large target-language corpora, which are unavailable for truly low-resource languages. To address this, we investigate tokenization efficiency through an information-theoretic lens, building on the established hypothesis that word length correlates with information content. From this perspective, we characterize tokenization inefficiency as having high fertility for low-information (highly predictable) words. Guided by this principle, we introduce a novel fine-tuning strategy that systematically identifies informationally redundant words—those with high fertility but low information content—for targeted vocabulary expansion and model fine-tuning. Experiments fine-tuning BLOOM and LLaMA-3 in English-Manipuri and other two language pairs show that our proposed method significantly reduces fertility by 50% and accelerates inference by more than 2 times, without compromising and often exceeding the translation quality of standard LLM baselines, providing a theoretically grounded solution for efficient LLM-based MT.

1 Introduction

Large language models (LLMs) have demonstrated strong performance in machine translation(MT) tasks (Kocmi et al., 2024). These models, such as ChatGPT (OpenAI, 2023), LLaMA (Touvron et al., 2023b), and BLOOM (Scao et al., 2022), are trained with a vast amount of multilingual data available on the internet, and hence acquire translation capabilities. Despite their widespread success, current machine translation approaches use multilingual LLMs trained on data from high-resource languages (Vilar et al., 2023), posing challenges for low-resource and unseen languages (Zhu et al., 2024b).

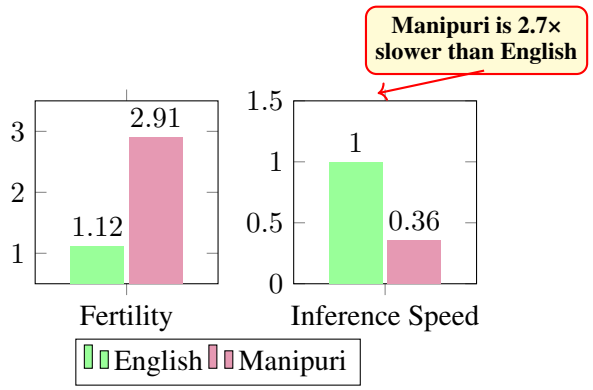

Figure 1: Comparison of Fertility and Inference Speed (in ratio) in BLOOM on WMT 2023 Manipuri and English data. Manipuri exhibits significantly higher fertility, resulting in a 2.7 times slower inference speed compared to English.

A primary issue of such LLMs is that the data-driven based LLM tokenizers segment words which are not present in the tokenizer vocabulary (Sennrich et al., 2016). Being trained in high-resource languages, these tokenizers have limited coverage of vocabularies for low-resource and unseen languages, which leads to fragmentation of almost all words exhibiting high fertility. Higher fertility results in lengthy sequences which leads to *slow inference speed* as well as *higher API costs* due to the increased number of generated tokens (Petrov et al., 2023; Ahia et al., 2023). Figure 1 shows one such observation in the BLOOM model: Manipuri has much higher fertility, and inference speed is 2.7 times slower than English.

A common strategy to reduce subword fertility is vocabulary expansion and continued pretraining on target language data. However, this approach has often yielded counterproductive results (degradation in translation quality). Recent studies

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2394–2404

- LLaMAX (Lu et al., 2024), TOWER (Alves et al., 2024), Sarvam-M (AI, 2025) reported degraded translation performance in continued pretraining of LLaMA-2 (Touvron et al., 2023c) and Mistral 3.1 model (Mistral AI Team, 2025) with vocabulary expansion. Furthermore, continual pre-training itself presupposes having access to large corpora in the target language, which are unavailable for truly low-resource languages like Manipuri. Now the problem is in the absence of a large corpus in the target language, *how to improve the tokenization efficiency of an LLM for a low-resource language without compromising the translation quality or the need for a large target dataset?*

To address this issue of existing LLM tokenizer segmenting almost all words in low-resource or unseen languages, we propose a systematic approach to identify inefficiently tokenized words and improve efficiency through *targeted* vocabulary expansion. Our approach is grounded in an established information-theoretic hypothesis: *word length correlates with its information content* (Piantadosi et al., 2011). From this perspective, a word's tokenization is inefficient if it has low information content (i.e., highly predictable) yet is still overfragmented into multiple subword tokens.

Building on this principle, we propose a targeted vocabulary expansion strategy that identifies these *informationally redundant words*—those with high fertility but low information content. To do this, we first measure information content of words using an n-gram KenLM model (Heafield, 2011) trained on the target language corpus. By systematically adding only those *informationally redundant words* to the model's vocabulary and then fine-tuning, we reduce sequence lengths and accelerate inference speed while preserving translation quality.

We validate our approach by fine-tuning BLOOM and LLaMA-3 models on truly low-resource English-Manipuri translation on WMT 23 and the BIBLE dataset. Our method is compared against two strong baselines: a standard fine-tuning approach without vocabulary modification and a BPE tokenizer-based baseline where the 3,000 most frequent target BPE tokens are added to the vocabulary. Our proposed approach sharply reduces fertility by 50% and accelerates inference by more than 2 times while matching—and often surpassing—the translation quality of standard LLM baselines. This framework is also validated on additional translation tasks of WMT 23 English-Assamese and WAT 21 English-Marathi.

Our work makes the following contributions:

- **(Theory)** We provide an information-theoretic framework for quantifying a word's tokenization inefficiency, linking subword fertility to information content.
- **(Framework)** We introduce a fine-tuning strategy for targeted vocabulary expansion that is guided by information content.
- **(Experiment)** Extensive experiments on multiple languages and models demonstrate that our proposed strategy matches and often outperforms baselines in terms of all evaluation metrics.

2 Related Works

Machine Translation with Multilingual LLMs. Large language models (LLMs) have greatly improved machine translation (MT), with various methods enhancing their performance. One approach uses in-context learning, where giving parallel sentences as examples guides translation; studies (Agrawal et al., 2023; Zhu et al., 2024a; Cui et al., 2024a) show that semantically related examples improve performance, especially with limited resources or data. Another approach involves finetuning with translation instructions: Xu et al. (2024) pretrained on monolingual data, then fine-tuned on small parallel datasets for strong results, while Guo et al. (2024) showed the continued importance of parallel data in continual pretraining.

Vocabulary Expansion in LLMs. Existing Large Language Models (Team, 2023; Touvron et al., 2023a) are trained with English-centric data which limits their effectiveness in low-resource languages. While building new models from scratch with diverse multilingual data is one solution (Wei et al., 2023), it is often computationally prohibitive. To overcome this, continual pre-training has emerged as a far more efficient and cost-effective paradigm (Zhao et al., 2024; Cui et al., 2024b; Faysse et al., 2024; Alves et al., 2024). This approach adapts an existing model by continuing its training on new, language-specific data. A crucial component of this process is vocabulary expansion (Gupta et al., 2023; Alves et al., 2024; Xie et al., 2023).

Vocabulary expansion addresses the fundamental challenge of tokenizer over-segmentation in non-English languages, which otherwise increases inference costs. Several studies (Downey et al.,

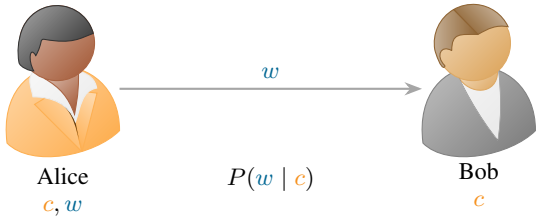

Figure 2: **Setting:** Alice knows both the context c and the word w, while Bob knows only the context c. Alice wishes to transmit w to Bob using an optimal code of length $-\log_2 P(w \mid c)$ bits.

2023; Liu et al., 2023; Kajiura et al., 2023; Balde et al., 2024; Wang et al., 2020; Hong et al., 2021) highlight performance improvements in tasks beyond machine translation. These include named entity recognition (NER) and part-of-speech (POS) tagging in multilingual environments, text classification in biomedical, computer science, news, and review areas, and abstractive summarization in medical settings. These studies report increases in F1 scores for NER, improvements in POS accuracy, and better ROUGE scores through methods like script-based embedding alignment, contrastive regularization, and adaptive BPE tokenization. However, despite these successes, recent findings also reveal a notable challenge in machine translation. Studies such as LLaMAX (Lu et al., 2024), TOWER (Alves et al., 2024), and Sarvam-M (AI, 2025) have reported degraded translation performance when applying continual pre-training with vocabulary expansion to models like LLaMA-2 (Touvron et al., 2023c) and Mistral 3.1 (Mistral AI Team, 2025).

3 Information Content of a Word and its Tokenization Efficiency

The information content of a word quantifies how many bits are needed, on average, to convey that word given its context. By framing word prediction as a communication problem—where Alice knows both the context c and the word w, while Bob only knows the context c—we derive the optimal code length for a word instance and the expected information carried by a word type across all contexts.

Alice and Bob agree on a probabilistic language model P. Alice knows both context c and word w, and wishes to efficiently transmit w to Bob (who already knows c). According to Shannon's source-coding theorem (Shannon, 1948), the optimal code length for this word is given by its **information content**:

$$-\log_2 P(w \mid c) \tag{1}$$

For a word type w appearing across multiple contexts, we estimate **expected information content** $I(w)$ from a corpus by averaging over the N observed occurrences of w. If c_i denotes the context of the i-th token of w, then:

$$I(w) = \mathbb{E}_{(w,c)}[-\log_2 P(w \mid c)] \tag{2}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} -\log_2 P(w \mid c_i) \tag{3}$$

To compute the information content, we use a KenLM trigram language model interpolated with unigram model (Heafield, 2011) to estimate the conditional probabilities P(w| c).

Zipf (1935) proposed that languages minimize utterance length, implying a word's length should be inversely proportional to its frequency. Complementing this, Piantadosi et al. (2011), building on the Uniform Information Density hypothesis (Jaeger and Levy, 2006), argued that word lengths optimize communication by keeping information rates near a theoretical channel capacity. Under this channel capacity hypothesis (formalized in Pimentel et al. (2023)), Piantadosi et al. proposed that a **word's length** should be proportional to its expected information content:

$$|w| \propto I(w) \tag{4}$$

We adopt the same Piantadosi et al.'s established hypothesis to analyze subword tokenization in large language models. In this setting, the encoding cost of a word is no longer its character length but its **subword fertility**.

$$\phi(w) \propto I(w). \tag{5}$$

From this perspective, we characterize tokenization inefficiency as having high subword fertility for low-information (highly predictable) words. This characterization serves as a normative principle for evaluating and improving tokenizer efficiency, particularly for low-resource languages where existing tokenizers may underperform.

4 Vocabulary Expansion

4.1 Vocabulary Selection using Information Content

Our vocabulary selection strategy, as shown in figure 3, leverages the theoretical framework estab-

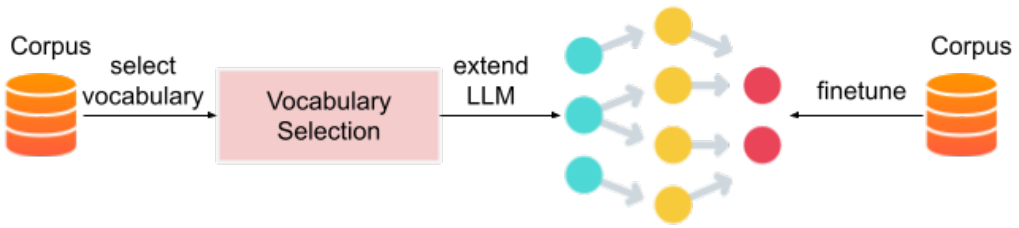

Figure 3: **Our Fine-Tuning Workflow:** We systematically identify informationally redundant words (those with high subword fertility but low information content), extend the LLM's vocabulary to include them as single tokens, and then fine-tune the model for improved tokenization efficiency without compromising translation quality.

lished in the previous section 3 to identify candidate words for vocabulary expansion. We define the **information efficiency** of a word w as the ratio between its expected information content and subword fertility: $\eta(w) = \frac{I(w)}{\phi(w)}$. Words with low information efficiency represent cases where the tokenizer over-segments words relative to their informativeness, violating the proportionality principle $\phi(w) \propto I(w)$. We systematically identify these inefficient words by first computing $I(w) = \frac{1}{N}\sum_{i=1}^{N} -\log_2 P(w \mid c_i)$ for each word w in our corpus, then filtering for words where $\phi(w) \geq \phi_{\min}$ (eg. $\phi_{\min} = 3$) and having low information efficiency $\eta(w)$.

The vocabulary augmentation process proceeds by selecting the top-k words with the lowest information efficiency scores, ensuring that these represent genuine cases of over-segmentation rather than naturally complex words. Formally, our candidate set is defined as $\mathcal{V}_{\text{aug}} = \{w : \phi(w) \geq \phi_{\min}\}$, ranked by ascending $\eta(w)$ values. This selection criterion ensures that we prioritize words where the tokenizer's segmentation is most misaligned with the word's information content, thereby maximizing the potential efficiency gains from vocabulary expansion. Finally, we extend the LLM tokenizer and embedding matrix with these new tokens and fine-tune the LLM on the target corpus.

4.2 Embedding Initialization

The core challenge lies in extending the LLM's vocabulary to include these words while maintaining compatibility with its existing embedding space. When adding a new word $w_{n+1} \notin \mathcal{V}$ to the original vocabulary $\mathcal{V} = \{w_1, \ldots, w_n\}$, we initialize its embedding e_{n+1} by leveraging the model's existing subword decomposition. First, w_{n+1} is segmented into constituent subwords $s_1, s_2, \ldots, s_k \in \mathcal{V}$ using the LLM's tokenizer. The word's embedding is then computed as the average of its subword embeddings:

$$e_{n+1} = \frac{1}{k}\sum_{j=1}^{k} e_{s_j},$$

where e_{s_j} denotes the embedding of the j-th subword. This initialization anchors the new word's representation within the semantic and syntactic neighborhood of its subcomponents, ensuring smooth integration into the LLM's pretrained embedding space. The updated model $p_{\theta'}(w_i \mid w_{1:i-1})$, with parameters $\theta' = \theta \cup \{e_{n+1}\}$, retains the original partition function $Z = \sum_{j=1}^{n} \exp(h_{i-1}^{\top} e_j)$ while incorporating the new word:

$$p_{\theta'}(w_i \mid w_{1:i-1}) = \frac{\exp(h_{i-1}^{\top} e_{w_i})}{Z + \exp(h_{i-1}^{\top} e_{n+1})}.$$

To analyze the stability of this expansion, we examine the logit contribution of e_{n+1}. The dot product $h_{i-1}^{\top} e_{n+1}$ simplifies to the average of the subword contributions:

$$h_{i-1}^{\top} e_{n+1} = \frac{1}{k}\sum_{j=1}^{k} h_{i-1}^{\top} e_{s_j}.$$

By Jensen's inequality, the exponential term satisfies:

$$\exp\left(h_{i-1}^{\top} e_{n+1}\right) \leq \frac{1}{k}\sum_{j=1}^{k} \exp\left(h_{i-1}^{\top} e_{s_j}\right),$$

which guarantees that the new word's contribution to the partition function remains bounded by the contributions of its subwords. This ensures minimal divergence from the original probability distribution, helping maintain stability in the probability distribution of the post-expansion language model.

5 Experimental Setup

Language	Family	Script	#Speakers
Manipuri	Tibeto-Burman	Bengali	3 million
Assamese	Indo-Aryan	Bengali	24 million
Marathi	Indo-Aryan	Devnagiri	99 million

Table 1: An overview of the focus languages, their families, scripts, and approximate number of speakers.

5.1 Focus Languages and Linguistic Properties

Our experiments focus on three Indian languages chosen to represent different levels of data availability: Manipuri (very low-resource), Assamese (low-resource), and Marathi (high-resource). This choice allows us to evaluate our methods across varying resource conditions.

These languages belong to two distinct families: Manipuri is a Tibeto-Burman language, while Assamese and Marathi are part of the Indo-Aryan family. Linguistically, all three are morphologically rich, agglutinative languages with a canonical Subject-Object-Verb (SOV) word order. While Manipuri and Assamese use the Bengali script, Marathi is written in the Devanagari script. A summary of these characteristics, including the approximate number of speakers for each language, is provided in Table 1.

5.2 Datasets and Corpora

The parallel corpora for our experiments are compiled from multiple established sources. For the English-Manipuri and English-Assamese translation tasks, we use data from the WMT 2023 Low-Resource Indic Language Translation shared task (Pal et al., 2023). For the English-Marathi pair, we use the PMI dataset from WAT 2021 (Nakazawa et al., 2021).

To assess English-Manipuri translation on different domains, we also use additional parallel corpus extracted from biblical text (**BIBLE**)[1]. To evaluate in BIBLE data, we sample 1000 sentences from the corpus. Such corpora are common in low-resource NLP, as the consistent verse-level alignment across many languages provides a valuable source of parallel sentences. The detailed statistics for our training corpora, including sentence and token counts for each language pair and source, are presented in Table 2.

Data	Sentences	Tokens	
		ENGLISH	XX
Manipuri(WMT)	21,687	390730	330319
Manipuri(BIBLE)	30,102	758482	588110
Assamese	50000	969623	825063
Marathi	28974	529821	423015

Table 2: Statistics of the parallel training corpora used in our experiments. Token counts are provided for both English and the target language.

5.3 Settings

Large Language Models. In this study, we mainly focus on multilingual, decoder-only LLMs that were not explicitly pre-trained on parallel corpora. For our experiments, we select BLOOM (Scao et al., 2022) and LLaMA-3 (et al., 2024). BLOOM covers two of our three target languages, omitting Manipuri, whereas LLaMA-3 does not natively support any of them. In our work, we use 1.07 billion model for BLOOM, and 1.24 billion LLaMA 3.2 model.

Baselines. We also compare our approach with state-of-the-art models that support Manipuri language, such as No Language Left Behind (NLLB) (Costa-Jussà et al., 2022). Furthermore, we also evaluated with BPE-Tok, a BPE tokenizer-based baseline where the 3,000 most frequent target BPE tokens are added to the vocabulary (Yamaguchi et al., 2024; Lu et al., 2024). For a fair comparison, BPE-Tok also uses the exact same embedding initialization strategy as described in section 4.2. While we experimented with 1,000, 2,000, and 3,000 BPE tokens, we report results for the 3,000 BPE tokens as it is our approach's optimal point and yields an average subword fertility comparable to our approach's optimal point.

Training Details. Due to constraints on our computational resources, we run with a reduced context length of 1024. LLMs are fine-tuned on translation dataset for 5 epochs on a single NVIDIA A100 GPU. We fine-tune the models using HuggingFace transformer tool (Wolf et al., 2020) with the default learning rate (5e-5). All other hyperparameters are kept at their default values as provided by the library.

[1]The Manipuri BIBLE corpus is available at https://live.bible.is/bible/MNIBIV

BLOOM		WMT 2023				BIBLE			
		FERTILITY	BLEU	CHRF++	TER	FERTILITY	BLEU	CHRF++	TER
	0	2.91	27.71	54.96	64.09	3.15	28.76	58.76	57.95
+New Token	100	2.59	28.43	55.43	62.93	2.46	28.90	59.30	57.67
	500	2.22	28.58	56.00	62.66	2.07	28.77	59.41	57.81
	1000	1.97	28.73	55.61	62.78	1.95	28.98	59.61	58.16
	2000	1.75	28.92	55.97	62.45	1.69	29.10	59.69	57.93
	3000	1.62	**29.52**	56.38	61.53	1.59	**29.43**	59.80	56.85
	4000	1.54	27.60	55.93	63.94	1.52	29.09	59.29	57.45
	NLLB	2.62	27.26	54.55	63.28	2.69	26.34	57.88	60.24
	BPE-Tok	1.63	24.64	54.21	66.38	1.67	25.89	57.5	61.11

Table 3: **BLOOM Result:** Fine-tuning BLOOM for English-to-Manipuri translation with an expanded vocabulary shows quality peaks at approx. 3,000 new tokens. Beyond this, performance drops despite a reduction in fertility, highlighting a trade-off between targeted augmentation and over-expansion.

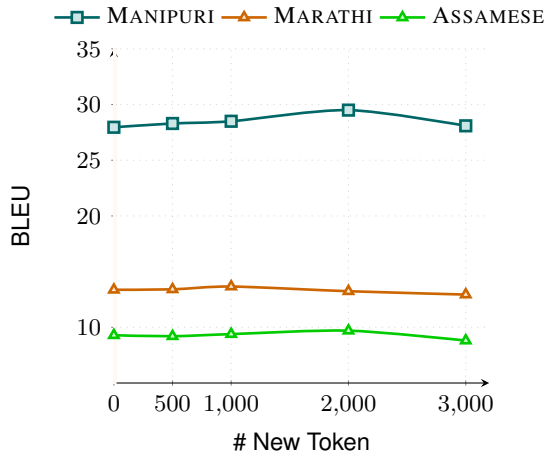

Figure 4: **LLaMA-3 Result:** Impact of vocabulary expansion on translation performance across Manipuri, Marathi, and Assamese. BLEU scores follow an inverted U-shaped curve, indicating an optimal, language-specific threshold of added tokens.

5.4 Evaluation Metrics.

Translation Performance. To evaluate quantitative results, three standard evaluation metrics are used: BLEU (bilingual evaluation understudy) (Papineni et al., 2002), which measures n-gram overlap with a brevity penalty for translation quality; chrF++ (Popović, 2017), a character n-gram F-score metric robust for various languages; and TER (translation error rate) (Snover et al., 2006), which calculates the minimum edits needed to align a hypothesis with the reference.

Inference Speedup. To measure the increase in inference speed of the proposed method, we measure **inference speedup** (in ratio) as the number of times the proposed method is faster in generating tokens than the baseline model.

LLaMA-3		Inference Speedup		
		ASSAMESE	MANIPURI	MARATHI
	0	1.0	1.0	1.0
+New Token	500	1.45	1.65	1.49
	1000	1.72	2.00	2.37
	2000	2.07	2.43	2.08
	3000	2.35	2.76	2.07

Table 4: LLaMA-3 inference speedup (in ratio) from vocabulary expansion across languages. Speedup values are relative to the baseline model (with no vocabulary expansion).

6 Result and Discussion

6.1 Main Result

Table 3 shows the impact of gradually expanding the BLOOM vocabulary with 100 to 4,000 Manipuri-specific words, followed by full fine-tuning of the model. We consider four key metrics in our evaluation, which include BLEU, chrF++, and TER to assess translation quality, and average subword fertility to measure tokenization efficiency on two test sets: WMT-23 En $\rightarrow$ Mni benchmark and BIBLE data. Our analysis reveals three prominent trends.

Fertility decreases almost linearly with vocabulary expansion. Starting from baseline values of 2.91 for WMT and 3.15 for the Bible, average subword fertility decreases steadily, halving to 1.54 for WMT and 1.52 for the Bible as 4,000 tokens are added. This linear reduction validates our approach, which suggests that adding word units for informationally redundant words effectively compresses sequences and accelerates inference speed.

Translation quality improves up to an optimal point. Translation quality improves steadily up to an optimum threshold of about 3,000 added tokens, yielding peak scores across both datasets relative to the baseline (with 0 new tokens): for WMT-23, improvements of +1.81 BLEU (from 27.71 to 29.52), +1.42 chrF++ (from 54.96 to 56.38), and –2.56 TER (from 64.09 to 61.53); for the Bible corpus, improvements of +0.67 BLEU (from 28.76 to 29.43), +1.04 chrF++ (from 58.76 to 59.80), and –1.10 TER (from 57.95 to 56.85). These are accompanied by a 44% reduction in fertility, suggesting that more efficient tokenisation directly benefits generation quality.

Over-expansion hurts. A clear point of diminishing returns is observed when expanding the vocabulary from 3,000 to 4,000 tokens. Expanding the vocabulary from 3,000 to 4,000 tokens further lowers fertility, but translation quality drops drastically on WMT (–1.92 BLEU, +2.41 TER, wiping out all previous gains) and modestly on BIBLE. This U-shaped curve highlights an optimal threshold for vocabulary augmentation; beyond this point, it is likely that adding new units introduces data sparsity that outweighs the benefits of more compact sequences, thereby destabilizing the fine-tuning process.

Takeaway. Our findings demonstrate that a targeted vocabulary expansion reduces the sequence length by nearly 50% with only 3,000 carefully selected tokens for Manipuri with the BLOOM model. However, indiscriminate expansion beyond language-specific optimal points reverses the trend, highlighting the importance of the information-guided vocabulary selection strategy at the core of our proposed framework. We also found lower performance of baseline BPE-Tokization, which is likely due to the way the newly trained BPE subword tokens disrupt the existing tokenization of high-information subwords in the pretrained LLM tokenizer, thereby disrupting the representations that the pretrained LLM has already learned. Our method avoids this by design by disrupting the existing tokenization of only low-information words.

6.2 Cross-lingual and Model Validation.

To validate the generalizability of our framework, we expand our analysis to include two additional Indic languages, Assamese and Marathi, and conduct parallel experiments with the LLaMA-3 model. As shown in Figure 4, the results on the LLaMA-3 model align with our primary observations from BLOOM and Manipuri, but reveal subtle differences across languages. For Manipuri, Marathi, and Assamese, the BLEU score follows an inverted U-shaped curve, with performance peaking before declining as more tokens are added. Specifically, the optimal performance is achieved with 2,000 new tokens for Manipuri, 1,000 for Marathi, and 2,000 for Assamese. This trend confirms that targeted vocabulary expansion is beneficial up to a language-specific threshold.

6.3 Inference Speedup

Our approach to targeted vocabulary expansion in the LLaMA-3 model significantly increases the inference speed without compromising and often outperforming baseline translation quality. This is achieved by shortening input sequences by replacing informationally redundant multi-token words with single tokens, leading to faster inference speed. The table 4 shows the inference speedup ratios (relative to the baseline) across Assamese, Manipuri, and Marathi languages, based on the number of new tokens added.

As the data show, all three languages experience consistent improvements over the baseline. Manipuri sees the highest gain, accelerating up to 2.76 times faster with 3,000 new tokens. Assamese follows closely, reaching a maximum of 2.35x speedup at the same token count. Marathi, however, peaks earlier at 2.37x with just 1,000 tokens, after which the speedup slightly decreases. This variation suggests that the optimal number of tokens for maximizing inference speed is language-dependent, indicating a trade-off where adding too many new tokens could introduce complexities that offset the benefits gained from shorter sequences.

6.4 Ablation Studies and Analyses

We perform multiple ablations to evaluate the impact of the key design choices made in the development of our models.

Effect of Information Theoretic Vocabulary Expansion. Our information-theoretic vocabulary expansion prioritizes *informationally redundant words*—those with high fertility but low information content. To use information-theoretic vocabulary expansion, it should be ensured that it can outperform the random vocabulary expansion as well as the vocabulary expansion that prioritizes *informationally non-redundant words*—those with

	WMT 2023 En-Mni		
	BLEU	CHRF++	TER
Redundant	**29.52**	56.38	61.53
Random	27.09	54.23	64.49
Non-redundant	27.17	54.14	64.43

Table 5: Vocabulary expansion prioritizing redundant words outperforms random and non-redundant strategies on WMT 2023 En-Mni translation metrics.

INITIALIZATION	WMT 2023 En-Mni		
	BLEU	CHRF++	TER
Average Init	**29.52**	56.38	61.53
Random Init	26.19	53.03	66.19

Table 6: Average embedding initialization outperforms random initialization for English-to-Manipuri translation across all metrics.

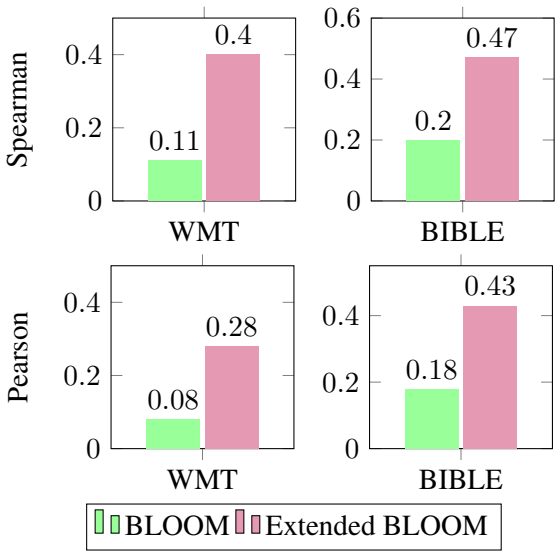

Figure 5: Correlation between Fertility and Information Content in BLOOM and Extended BLOOM on WMT 2023 Manipuri data and BIBLE data - Side by side comparison.

high fertility and high information content. To verify this, we compare three models: (i) informationally redundant words - those with high fertility but low information content, (ii) random - vocabulary selections are done randomly, (iii) non-redundant: vocabulary expansion that prioritizes those with high fertility and high information content.

As shown in Table 5, vocabulary expansion of 3000 words with *informationally non-redundant words* only reduces fertility of LLMs without degrading the translation performance. On the other hand, vocabulary expansion with randomly chosen tokens and informationally non-redundant words hurts the translation performance, although it reduces the fertility.

Random Initialization disrupts translation quality. As shown in Table 6, random initialization of new words reduces the BLEU score to 26.19 for English-to-Manipuri translation, significantly underperforming the BLOOM with averaging initialization (29.52 BLEU). The standard approach, which uses averaging-based initialization of subword embeddings, maintains coherence in the pretrained embedding space. In contrast, random initialization introduces noise by disrupting the model's token distribution. This underscores the importance of structured initialization methods when extending LLM vocabularies, particularly for low-resource languages.

Correlation between Subword Fertility and Information Content To validate our core hypothesis that an efficient tokenizer should exhibit a strong correlation between subword fertility and information content, we evaluate this correlation after vocabulary expansion. Figure 5 compares the Pearson and Spearman correlations for the original tokenizer (BLOOM) with the extended version (Extended BLOOM) across the WMT-23 and BIBLE Manipuri datasets. Our method significantly strengthens this relationship. For the WMT dataset, Spearman's correlation rises from 0.11 to 0.40, and Pearson's from 0.08 to 0.28. Similarly, on the BIBLE dataset, Spearman's increases from 0.20 to 0.47, and Pearson's from 0.18 to 0.43.

By selectively adding informationally redundant words (high fertility, low information) as single tokens, our method ensures that longer, multi-token segmentations are now more likely to be reserved for words that are genuinely information-rich. This improved alignment provides strong evidence that the gains in translation quality and fertility reduction are direct consequences of a more theoretically sound and efficient tokenization scheme.

7 Conclusion

This paper introduces a theoretically grounded approach to address the high subword fertility problem, particularly for low-resource languages. By systematically identifying and expanding the vocabulary with informationally redundant words—those with high fertility but low information content—we significantly reduce subword fer-

tility by 50% and accelerated inference by over two times. Crucially, our method achieves these efficiency gains without compromising translation quality, and often, by exceeding the performance of standard LLM baseline (without vocabulary expansion) across various languages and models (BLOOM and LLaMA-3).

Our findings highlights the importance of an information-theoretic lens for optimizing tokenization efficiency. The observed U-shaped curve in translation quality relative to vocabulary expansion highlights an optimal threshold, beyond which indiscriminate vocabulary expansion can negate benefits. This work provides a valuable framework for developing more efficient LLM-based MT systems for low-resource languages.

8 Limitations

Our study faced several limitations. From a computational perspective, due to resource constraints, we were constrained to using smaller versions of the models - specifically the 1.24 billion parameter LLaMA-3 model and 1.07 billion parameter BLOOM model, with a reduced context length of 1024. This may have limited the models' capacity to learn. Additionally, in our work, we focus on three Indian languages that have relatively high subword fertility due to their agglutinative and morphologically rich nature. Evaluation of languages of other families (e.g., templatic like Arabic or logographic Chinese) can be conducted in the future. Furthermore, U-shaped translation performance on vocabulary expansion implies careful tuning of "k" new tokens per language; we plan to further study on automated selection of this hyperparameter.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Sarvam AI. 2025. Sarvam-m. Sarvam AI Blog. Accessed: June 29, 2025.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. Adaptive BPE tokenization for enhanced vocabulary adaptation in finetuning pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14724–14733, Miami, Florida, USA. Association for Computational Linguistics.

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*.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024a. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897, Bangkok, Thailand. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024b. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

C.m. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. 2023. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 268–281, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Croissantllm: A truly bilingual french-english language model. *Preprint*, arXiv:2402.00786.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649, Mexico City, Mexico. Association for Computational Linguistics.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *Preprint*, arXiv:2308.04014.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Jimin Hong, TaeHee Kim, Hyesu Lim, and Jaegul Choo. 2021. AVocaDo: Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

T Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

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

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

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

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Mistral AI Team. 2025. Mistral Small 3.1 24B. `https://mistral.ai/news/mistral-small-3-1`. Accessed: June 29, 2025.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

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.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2240–2255, Singapore. Association for Computational Linguistics.

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

Teven Le Scao, Angela Fan, Christopher Akiki, and Et. al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

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

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

OpenAI Team. 2023. Gpt-4 technical report.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin R. Stone, and et. al. 2023c. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, and 1 others. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models. *Preprint*, arXiv:2311.08545.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. How can we effectively expand the vocabulary of llms with 0.01gb of target language text? *Preprint*, arXiv:2406.11477.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

George K. Zipf. 1935. *The Psychobiology of Language*. Routledge, London.