MorphBPE: A Morphology-aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies

Anonymous ACL submission

Abstract

Tokenization fundamentally shapes NLP performance, influencing both efficiency and linguistic fidelity. While Byte Pair Encoding (BPE) underpins most Large Language Models (LLMs), it overlooks morphological boundaries, resulting in suboptimal and inconsistent segmentation, especially in morphologically rich languages. We introduce MorphBPE, a morphology-aware extension of BPE that integrates linguistic structure into subword tokenization while retaining statistical efficiency and full compatibility with existing LLM pipelines. We conduct intrinsic evaluations on both the tokenizer and the subsequent language model. The tokenizer is assessed using two proposed metrics, (i) Morphological Consistency F1 and (ii) Morphological Edit Distance, capturing morphological alignment and segmentation accuracy. The language model is evaluated via cross-entropy loss across English (fusional, low complexity), Russian (fusional, moderate), Hungarian (agglutinative, high), and Arabic (templatic, high). Results show consistent improvements in morphological coherence and training efficiency. Extrinsic evaluation on the Belebele multilingual reading comprehension benchmark further demonstrates that MorphBPE significantly enhances performance in morphologically rich languages such as Russian and Arabic.

Availability: The $MorphBPE$ codebase, datasets, and the tokenizer playground will be available with the publication of the work.

1 Introduction

Tokenization is a fundamental preprocessing step in NLP, converting raw text into structured units such as bytes (Gillick et al., 2016), characters (Al-Rfou et al., 2019), subwords (Sennrich et al., 2016), words, or multi-word expressions (Gee et al., 2023). Its effectiveness directly influences downstream tasks, as tokenization errors can propagate through the pipeline, impacting overall model performance (Sajjad et al., 2017; Adel et al., 2018). Over the years, tokenization has advanced from basic whitespace-based segmentation to sophisticated statistical and neural approaches (Smit et al., 2014; Otani et al., 2020). In Large Language Models (LLMs), tokenization significantly affects efficiency, context length, and representational accuracy (Dagan et al., 2024). Although tokenization-free architectures have been investigated as potential alternatives (Clark et al., 2022; Deiseroth et al., 2024), most state-of-the-art models—including Gemma (Team et al., 2024), LLaMA (Touvron et al., 2023), DeepSeek (Bi et al., 2024) and OpenAI's GPT series—still rely on Byte Pair Encoding (BPE)-based tokenization for most languages, retaining both its benefits and inherent limitations.

The additive nature of Byte Pair Encoding (BPE) makes it well-suited for concatenative morphology, as seen in English, where morphemes are linearly appended. However, it struggles with non-concatenative morphological systems, such as root-and-pattern morphology in Arabic and Hebrew, where meaning is encoded through non-linear infixation (Khaliq and Carroll, 2013). Similarly, agglutinative languages like Turkish, Hungarian, and Korean pose challenges, as their highly productive affixation processes complicate adherence to morpheme boundaries (Hakkani-Tür et al., 2000). These languages require finer-grained tokenization to preserve linguistically meaningful subword structures. Standard BPE and byte-level tokenization methods often struggle to represent these complex morphological patterns effectively, emphasizing the necessity for morphology-sensitive tokenization approaches that better align with the diverse structural properties of different word formation processes (Marco and Fraser, 2024).

Analyzing BPE output across morphologically rich languages, we observe that its segmentation often disregards meaningful morpheme boundaries,

introducing ambiguity and disrupting semantic coherence. For instance, in Arabic, the word الرحمن (Al-Rahman, "The Merciful") may be incorrectly segmented into من (min, "whom") ال (al, "the") + رح (rah, an incomplete fragment). Here, من (min), a frequent token, is semantically unrelated to the original word, increasing the model's burden in reconstructing meaningful representations. Similar challenges arise in agglutinative and polysynthetic languages, where BPE's greedy merging strategy often fails to align with true morpheme boundaries.

While purely morphology-based segmentation could mitigate these issues, it has also shown limitations in aligning with naturally occurring linguistic patterns in corpus-based learning (Durrani et al., 2019; Marco and Fraser, 2024). Thus, developing tokenization methods that balance morphological integrity with statistical efficiency remains a critical challenge for multilingual NLP.

Contributions. We introduce $MorphBPE$, a morphology-aware extension of Byte Pair Encoding (BPE) that integrates linguistic structure into subword tokenization while preserving statistical efficiency and full compatibility with existing LLM pipelines. Our main contributions are:

(i) Interpretable Morphology-Aware Tokenizer: $MorphBPE$ incorporates morphological boundaries into tokenization, enhancing linguistic coherence and interpretability while maintaining token efficiency and compatibility with standard LLM training frameworks.

(ii) Intrinsic Evaluation of the Tokenizer: We propose two morphology-based metrics to assess segmentation quality: **Morph.-Edit Distance**: This metric quantifies morpheme-level segmentation accuracy, reflecting interpretability and adherence to morphological boundaries. **Morph.-Consistency F1-Score**: Inspired by Marco and Fraser (2024), measures segmentation consistency, evaluating whether words sharing morphemes are assigned consistent tokens.

For benchmarking, we curate a dataset covering diverse morphological typologies (Ge and Comrie, 2022): *English* (fusional, low complexity), *Russian* (fusional, moderate), *Hungarian* (agglutinative, high), and *Arabic* (templatic, high).

(iii) Intrinsic Evaluation of Trained LLMs: We compare $MorphBPE$ with vanilla BPE using *300M*- and *1B*-parameter LLMs, demonstrating consistently lower cross-entropy loss at the same vocabulary size across all four languages, reflecting improved linguistic representations and convergence efficiency.

(iv) Extrinsic Evaluation on Downstream Tasks: On the multilingual Belebele benchmark (Bandarkar et al., 2024), $MorphBPE$ achieves improvements in morphologically rich languages such as Russian and Arabic.

2 Related Work

BPE, originally introduced as a text compression algorithm (Shibata et al., 1999), was first adapted for machine translation as a tokenization method in 2016 (Sennrich et al., 2016). Since then, it has become the de facto standard in NLP and Large Language Models (LLMs) due to its efficiency in managing vocabulary size, handling out-of-vocabulary words, and capturing frequent patterns, while offering partial improvements over morphology-based tokenizers (Sennrich et al., 2016).

Despite its widespread adoption, vanilla BPE has several notable limitations: its greedy merging strategy, inefficiencies in cross-lingual settings where similar words with different character variations are not aligned, and inconsistent handling of character-level information across languages. To address these challenges, various extensions have been proposed, including BPE dropout (Provilkov et al., 2020), which introduces stochasticity to improve generalization, sampling-based BPE (Asgari et al., 2020), which enhances subword diversity, byte-level adaptations (Wang et al., 2020), which aim to improve robustness across scripts, and multilingual BPE variants (Liang et al., 2023), designed to optimize token sharing across languages.

The importance of morphology-aware tokenization for language models has been recognized in several recent studies (Park et al., 2021; Jabbar, 2023; Marco and Fraser, 2024; Weller-Di Marco and Fraser, 2024). However, an integrated solution that effectively balances morphological information with frequent pattern extraction while remaining fully compatible with modern LLM training pipelines has remained an open problem.

Distinctions Between MorphBPE and Related Works. MorphBPE stands out from existing methods through its design and integration. Unlike pre-tokenization approaches (Otani et al., 2020; Nzeyimana and Niyongabo Rubungo, 2022) or dictionary-based or multi-view segmentation methods (Park et al., 2021), which segment text prior

to applying a standard tokenizer, MorphBPE constrains the BPE merge process itself using morpheme boundaries, requiring segmentation only during training and not at inference. In contrast to stochastic methods like BPE-Dropout (Provilkov et al., 2020) and subword sampling techniques (Asgari et al., 2020) or memory expensive approaches like SentencePiece Unigram LM (Kudo and Richardson, 2018; Sims et al., 2025), MorphBPE is deterministic, linguistically grounded, and memory efficient. Most closely related are morpheme-aware tokenization strategies (Jabbar, 2023; Marco and Fraser, 2024), which aim to align tokens with morphological boundaries. However, MorphBPE advances this direction by enabling seamless integration into standard BPE training, maintaining compatibility with LLM pipelines, being more interpretable, with higher morphological consistency (detailed in § 3.3.1), and empirically demonstrating benefits in language model training through reduced loss, faster convergence, and downstream performance.

3 Methods

Overview: Our approach consists of four steps: (i) we introduce $MorphBPE$, a morphology-aware extension of BPE (§3.2), and select four languages, English, Russian, Hungarian, and Arabic, representing diverse morphological typologies for evaluation; (ii) we design intrinsic metrics to assess tokenization quality independently (§3.3.1); (iii) we train and evaluate LLMs using $MorphBPE$ to measure efficiency and representation improvements (§3.3.2); and (iv) we conduct extrinsic evaluations on multilingual downstream tasks to validate the overall impact (§3.3.2).

3.1 Datasets

3.1.1 Morphological Data

Our dataset comprises morphologically segmented words from four morphologically diverse languages (Ge and Comrie, 2022): English, Russian, Hungarian, and Arabic. Segmentation data for English, Russian, and Hungarian is sourced from the *SIGMORPHON 2022 Shared Task on Morpheme Segmentation* (Batsuren et al., 2022), which provides high-quality morpheme segmentations. To incorporate a root-based templatic morphological system, we include Arabic, where we utilize multiple sources: the *Arabic Treebank (ATB)* dataset (Taji et al., 2017), the *Dialectal Segmentation Dataset*

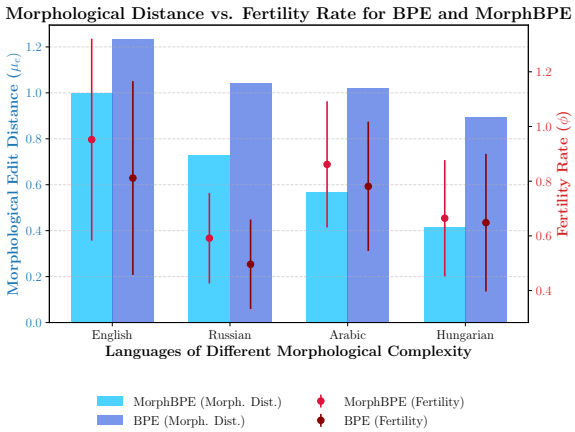

Figure 1: Comparison of morphological distance and fertility rate for BPE and $MorphBPE$ across four languages. Lower fertility is generally preferred, and lower morphological distance indicates better alignment with morpheme boundaries.

(Darwish et al., 2018), and the *Quranic Morphology Data* (Dukes and Habash, 2010). Additionally, we enrich this set with one million (1M) high-confidence segmentations of frequent Arabic surface forms obtained using *Farasa* (Darwish and Mubarak, 2016). All datasets were cleaned and standardized for consistency. Manually annotated segmentations were divided into 80% training, 10% validation, and 10% test sets. Table 1 summarizes the dataset composition.

3.1.2 LLM Training Data

To evaluate $MorphBPE$ and standard BPE across languages with diverse morphological structures (Hungarian, Arabic, Russian, and English), we use the *FineWeb2* corpus (Penedo et al., 2024), a large-scale multilingual dataset covering over 1,000 languages. This corpus provides sufficient token coverage for training in accordance with the *Chinchilla scaling law* (Hoffmann et al., 2022). Using FineWeb2 ensures a balanced token distribution across the selected languages, enabling a fair and robust comparison between $MorphBPE$ and BPE.

3.2 MorphBPE approach

MorphBPE is a simple yet effective extension of BPE that restricts frequent pair merges to remain within morpheme boundaries while leaving the rest of the algorithm unchanged (Algorithm 1). This design ensures full compatibility with standard BPE inference, allowing seamless integration into existing language modeling pipelines without modifica-

Table 1: Morphological segmentation dataset statistics used for BPE and $MorphBPE$ training and evaluation across languages.

Language	Morphology Type	# of Words	Avg. Morphemes per Word
English	Fusional (low complexity)	571,495	2.33
Russian	Fusional (moderate complexity)	784,212	3.84
Hungarian	Agglutinative (high complexity)	930,312	3.22
Arabic	Templatic (high complexity)	1,395,835	2.50

Table 2: Morphological consistency evaluation for BPE, $MorphBPE$, and SentencePiece Unigram LM across four languages. Precision, recall, and F1-score (μ_c) are reported as mean $\pm$ standard deviation over multiple resamples of the test sets. Higher F1-scores indicate greater consistency in segmenting words that share or differ in morphemes.

Model	**Precision (Mean $\pm$ Std)**	**Recall (Mean $\pm$ Std)**	**Morph.-Consistency F1 (μ_c)**
English (96K)			
BPE	0.00 ± 0.00	0.03 ± 0.02	0.00
$MorphBPE$	0.31 ± 0.44	0.34 ± 0.09	**0.32**
SentencePiece Unigram LM	0.20 ± 0.41	0.44 ± 0.07	0.28
Russian (64K)			
BPE	0.10 ± 0.32	0.06 ± 0.01	0.07
$MorphBPE$	0.69 ± 0.48	0.33 ± 0.06	**0.45**
SentencePiece Unigram LM	0.63 ± 0.49	0.22 ± 0.06	0.33
Hungarian (24K)			
BPE	0.08 ± 0.25	0.29 ± 0.04	0.13
$MorphBPE$	0.98 ± 0.03	0.78 ± 0.07	**0.87**
SentencePiece Unigram LM	0.93 ± 0.17	0.81 ± 0.10	0.87
Arabic (96K)			
BPE	0.00 ± 0.00	0.08 ± 0.03	0.00
$MorphBPE$	0.89 ± 0.31	0.53 ± 0.05	**0.66**
SentencePiece Unigram LM	0.73 ± 0.27	0.49 ± 0.04	0.58

tion.

Algorithm 1 Morphology-Aware Byte Pair Encoding (MorphBPE)

```
Initialize vocabulary with individual characters
Segment the training corpus using morphological segmentation
while number of merges < desired vocabulary size do
    Compute symbol pair frequencies
    Morph-aware step: merge the most frequent pair without crossing morpheme boundaries
    Update vocabulary with the merged symbol
    Continue standard BPE merge operations as usual
end while
```

3.3 Evaluation Framework of MorphBPE

Evaluation of tokenization and its downstream impact can be conducted at two levels: (*i*) the tokenizer itself, and (*ii*) the language model that uses it. Language model evaluation can further be divided into *intrinsic* and *extrinsic* analyses. We first describe the tokenizer-level evaluation and intrinsic language model metrics, followed by the extrinsic assessment on multilingual downstream tasks.

3.3.1 Tokenizer-Level Evaluation (Intrinsic)

Before assessing tokenization effects in downstream or model-level tasks, it is essential to analyze intrinsic properties of the tokenizer to ensure efficiency and morphological consistency. We evaluate three core metrics:

(i) Fertility (ϕ): Fertility is a common metric for the evaluation of tokenizers, which quantifies the number of tokens produced relative to a baseline, typically a whitespace-based tokenizer (Rust et al., 2021). A lower fertility score indicates higher compression efficiency and potentially longer context spans. However, this relationship is not uniform across languages: agglutinative and templatic languages such as Hungarian and Arabic naturally require higher fertility to capture rich morphological structures. As shown in Table 1, these languages exhibit higher morpheme-to-word ratios.

(ii) Morphological Edit Distance (μ_e): We introduce a new intrinsic metric, the *morphological edit distance*, which measures the alignment between tokenizer output and true morpheme boundaries. Using dynamic programming, we compute pairwise alignment scores that preserve token order, providing a quantitative measure of how faithfully a tokenizer respects morphological structure. This score directly reflects token interpretability. Although it can be normalized by morpheme count, we report raw values to indicate the average number of edits required.

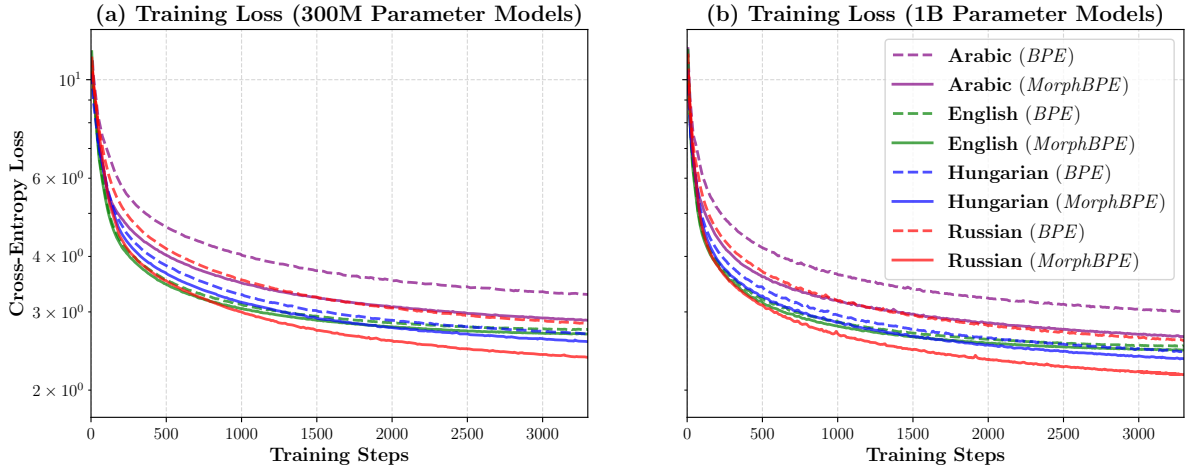

Figure 2: Training cross-entropy loss comparison between BPE and $MorphBPE$ across English, Russian, Hungarian, and Arabic for both the 300M and 1B models (lower values indicate better performance).

(iii) Morphological Consistency (F1: μ_c): Inspired by Marco and Fraser (2024), this metric evaluates segmentation consistency: whether words sharing morphemes also share tokens (recall) and whether shared tokens correspond to shared morphemes (precision). The harmonic mean of these yields the morphological consistency F1-score. To ensure scalability, we cluster words via k-means ($k = 100$), select $C = 50$ pairs per cluster, and estimate scores with $N = 10$ bootstrap resamples.

Vocabulary Size Selection: Vocabulary size influences tokenization quality and model performance. To ensure a fair comparison, both BPE and $MorphBPE$ are evaluated under identical vocabulary capacities for each language. Sizes are chosen so that standard BPE can effectively utilize its capacity to capture morpheme-level distinctions, enabling an equal basis for comparison. We vary vocabulary sizes from 8K to 96K in 8K increments and compute the morphological edit distance (μ_e) on development sets, selecting the smallest size beyond which improvements are not statistically significant (via a t-test). This procedure yields 24K for Hungarian, 64K for Russian, and 96K for English and Arabic. We then evaluate both tokenizers using (i) fertility (ϕ), (ii) morphological edit distance (μ_e), and (iii) morphological consistency (μ_c), with fertility compared against the multilingual Bloomz baseline (256K) (Yong et al., 2023).

3.3.2 Language Model Evaluation

We train two model scales, **300M (small)** and **1B (large)**, using decoder-only architectures implemented in the LLaMA-Factory framework (Zheng et al., 2024). For each language, models are trained with both vanilla BPE and $MorphBPE$ using the same vocabulary sizes, ensuring comparability. The smaller model processes approximately 6B tokens and the larger model 20B, in accordance with the *Chinchilla scaling law* (Hoffmann et al., 2022).

Cross-Entropy Loss (Intrinsic): Cross-entropy loss measures the divergence between predicted and true tokens, reflecting convergence speed and representational quality. It is analogous to perplexity but provides finer-grained insight during training. Comparisons are made only across models with identical vocabulary sizes to ensure fairness, as vocabulary size directly affects branching factors.

Downstream LM Evaluation (Extrinsic): To assess downstream comprehension and reasoning, we perform extrinsic evaluation using the *Belebele* benchmark (Bandarkar et al., 2024). Belebele is a multilingual multiple-choice reading comprehension dataset covering 122 language variants, including English, Russian, Arabic, and Hungarian. Each instance consists of a short passage from the Flores-200 corpus followed by a question with four candidate answers, designed to discriminate models with different levels of language understanding.

All evaluations are conducted using the *LM Evaluation Harness* (Gao et al., 2024), which provides a standardized zero-shot evaluation protocol across models. For each example i, a model M receives a prompt x_i and a set of candidate answers $\{y_i^{(1)}, \ldots, y_i^{(k)}\}$ with a gold label $g_i \in$

$\{1, \ldots, k\}$. The model assigns log-probabilities $s_i^{(j)} = \log P_M(y_i^{(j)} \mid x_i)$ and predicts $\hat{g}_i = \arg\max_j s_i^{(j)}$. Accuracy is computed as

$$\text{Accuracy}(M) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{g}_i = g_i].$$

To assess statistical significance between models, we apply the McNemar test on paired correctness outcomes, combined with 10^6 bootstrap resamples of per-sample accuracy differences. Multiple comparisons across subtasks are corrected using the Benjamini–Hochberg false discovery rate (FDR) procedure at $\alpha = 0.05$. This evaluation quantifies whether improvements in morphological alignment translate into measurable gains in reading comprehension.

4 Results

4.1 Tokenizer-Level Evaluation (Intrinsic)

Figure 1 and Table 2 summarize the intrinsic tokenizer evaluation results. Across all four languages, $MorphBPE$ achieves consistently lower morphological edit distance (μ_e) and higher morphological consistency (μ_c) compared to standard BPE, with only a marginal increase in fertility (ϕ). These trends hold across all evaluated languages, and the gains are most pronounced in morphologically rich languages such as Hungarian and Arabic.

For completeness, we also compare morphological consistency (μ_c) with *SentencePiece Unigram LM* (Kudo and Richardson, 2018) under identical vocabulary sizes and corpora. $MorphBPE$ achieves higher μ_c across all four languages—English, Russian, Hungarian, and Arabic—indicating more stable and linguistically coherent tokenization. As *SentencePiece* is computationally more expensive and yields lower tokenization consistency (Table 2), which limits the model's ability to observe sufficient and accurate contexts for each token, we did not extend language model training with this approach.

The results confirm that $MorphBPE$ more effectively preserves morpheme boundaries, yielding linguistically coherent tokenizations, whereas BPE tends to over-fragment words based solely on statistical frequency. The magnitude of improvement in μ_e and μ_c correlates with morphological complexity, highlighting the benefits of morphology-aware tokenization for agglutinative and templatic languages.

These results support the conclusion that morphology-aware segmentation leads to more interpretable and stable token representations, which likely facilitate improved model convergence during training.

4.2 Language Model Evaluation (Intrinsic)

The intrinsic language model results are presented in Figure 2, showing the training cross-entropy loss for both the **300M** (small) and **1B** (large) models across the four languages. Models trained with $MorphBPE$ consistently achieve lower cross-entropy loss than those trained with BPE, even under identical vocabulary sizes and training data. Results are reported over a representative window of approximately 14B tokens for clarity, though the observed trends persist throughout training.

The reduction in cross-entropy loss indicates that $MorphBPE$ yields more efficient and linguistically informed representations, leading to faster convergence and better optimization dynamics. The gains are particularly notable in morphologically rich languages such as Hungarian and Arabic, where improved alignment between tokens and morphemes provides clearer learning signals for the model.

4.3 Language Model Evaluation (Extrinsic)

While a full downstream comparison would require supervised fine-tuning (SFT) and larger model scales, our analysis focuses on isolating tokenization effects within the 300M and 1B parameter LLMs. At these sizes and without SFT, absolute accuracy is expectedly modest; however, statistical testing provides insight into relative improvements.

Across all four languages, $MorphBPE$ achieves an average accuracy increase of approximately **1%** over BPE. The improvement is consistent in morphologically rich languages—*Arabic*, *Russian*, and *Hungarian*—but negligible in *English*, where morphology plays a minimal role. McNemar's paired test with 10^6 bootstrap resamples, corrected using the Benjamini–Hochberg FDR procedure at $\alpha = 0.05$, confirms that the gains in *Arabic* and *Russian* are statistically significant, while the increase for *Hungarian* is not.

These findings indicate that morphology-aware tokenization contributes to measurable improvements in comprehension and reasoning tasks for morphologically complex languages, reinforcing the link between morphological alignment and semantic interpretability in downstream language

modeling.

5 Discussion and Conclusion

In this work, we introduced $MorphBPE$, a morphology-aware extension of Byte Pair Encoding (BPE) that integrates linguistic structure into subword tokenization while maintaining statistical efficiency and full compatibility with existing LLM training pipelines. Through extensive evaluations across *English*, *Russian*, *Hungarian*, and *Arabic*, we demonstrated that $MorphBPE$ improves both interpretability and efficiency, yielding linguistically coherent tokenization and more effective model training.

Morphological Alignment and Consistency: Intrinsic evaluations showed that $MorphBPE$ achieves lower morphological edit distance and higher morphological consistency than standard BPE, reflecting better adherence to morpheme boundaries and more stable token representations. The two proposed metrics—*Morphological Consistency F1-Score* and *Morphological Edit Distance*—proved effective for quantifying segmentation stability and alignment, addressing a critical gap in existing tokenizer evaluation. Such consistency is key to ensuring that related words are represented by coherent token patterns, which improves LLM generalization and convergence.

Efficiency and Fertility: Although morphology-aware constraints could increase token counts, $MorphBPE$ achieves balanced efficiency. These results suggest that fertility alone is not a reliable measure of tokenizer quality for LLMs; morphological coherence and consistent segmentation provide a better indicator of efficient learning and contextual representation.

Language Modeling Performance. Across both 300M and 1B parameter models, $MorphBPE$ yielded consistently lower cross-entropy loss than BPE at identical vocabulary capacities, demonstrating faster convergence and more accurate linguistic representations. Improvements were strongest in morphologically complex languages, where morphological alignment helps the model capture systematic word-formation patterns. These findings empirically support the hypothesis that linguistically informed tokenization facilitates more effective representation learning during LLM training.

Downstream Evaluation: Extrinsic evaluation using the *Belebele* benchmark (Bandarkar et al., 2024) provides further evidence that improved morphological alignment translates into better language understanding. While absolute accuracy is limited by model scale and the lack of supervised fine-tuning (SFT), $MorphBPE$ achieved an average accuracy gain of approximately **1%** over BPE. Improvements in *Arabic* and *Russian* were statistically significant under McNemar testing with Benjamini–Hochberg correction ($\alpha = 0.05$), while *Hungarian* showed a smaller, non-significant gain. These results suggest that morphology-aware tokenization provides measurable advantages in comprehension and reasoning tasks for morphologically rich languages. Our result is consistent with findings by Wei et al. (2024) linking perplexity improvements to downstream gains.

Broader Implications and Future Work: Our results challenge the common assumption that minimal fertility or smaller token counts directly indicate superior tokenizers. Instead, we show that morphological consistency—ensuring semantically coherent token boundaries—plays a central role in enabling stable and interpretable model learning. $MorphBPE$ thus bridges linguistic insight and scalable tokenization, offering a practical balance between efficiency and interpretability. A comprehensive evaluation of such tasks at larger model scales remains an exciting direction for future research.

Finally, $MorphBPE$ integrates seamlessly into existing BPE-based LLM training and inference workflows, requiring minimal modifications. To facilitate reproducibility and further exploration, we will release the full $MorphBPE$ implementation, morphological datasets, and evaluation metrics alongside this work.

6 Limitations

While we demonstrated the effectiveness of $MorphBPE$ across four languages representing diverse morphological typologies, several limitations remain. First, $MorphBPE$ relies on the availability of reliable morphological segmentation data, which is not yet available for many low-resource or under-documented languages. Ongoing initiatives such as *UniMorph* (Kirov et al., 2018) and *MorphyNet* (Batsuren et al., 2021) are helping to address this gap, but broader morphological coverage will be crucial for extending our approach to more languages and language families.

Second, although we trained and evaluated models from scratch across four typologically distinct

languages, a large-scale effort requiring approximately 2,000 H100 GPU hours, our focus was primarily on intrinsic and controlled extrinsic evaluations. Comprehensive downstream assessments of LLMs trained with $MorphBPE$, including tasks involving reasoning, instruction following, or cross-lingual transfer, remain an important direction for future research.

Finally, while $MorphBPE$ introduces linguistic structure without compromising statistical efficiency, further exploration of its interaction with scaling laws, vocabulary sharing across languages, and multilingual pretraining objectives could yield deeper insights into how morphology-aware tokenization influences generalization in large-scale LLMs.

References

Heike Adel, Ehsaneddin Asgari, and Hinrich Schütze. 2018. Overview of character-based models for natural language processing. In *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part I 18*, pages 3–16, Cham. Springer International Publishing.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166.

Ehsaneddin Asgari, Masoud Jalili Sabet, Philipp Dufter, Christopher Ringlstetter, and Hinrich Schütze. 2020. Subword sampling for low resource word alignment. *arXiv preprint arXiv:2012.11657*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Gautier Dagan, Gabriele Synnaeve, and Baptiste Rozière. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. *arXiv preprint arXiv:2402.01035*.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate Arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074, Portorož, Slovenia. European Language Resources Association (ELRA).

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS tagging: A CRF approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Björn Deiseroth, Manuel Brack, Patrick Schramowski, Kristian Kersting, and Samuel Weinbach. 2024. T-FREE: Subword tokenizer-free generative LLMs via sparse representations for memory-efficient embeddings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21829–21851, Miami, Florida, USA. Association for Computational Linguistics.

Kais Dukes and Nizar Habash. 2010. Morphological annotation of Quranic Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1504–1516, Minneapolis, Minnesota. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa,

Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Pingping Ge and Bernard Comrie. 2022. Correlations of valency alternations and morphological types: A typological perspective. *Lingua*, 273:103304.

Leonidas Gee, Leonardo Rigutini, Marco Ernandes, and Andrea Zugarini. 2023. Multi-word tokenization for sequence compression. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 612–621, Singapore. Association for Computational Linguistics.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306, San Diego, California. Association for Computational Linguistics.

Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2000. Statistical morphological disambiguation for agglutinative languages. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

J. Hoffmann, S. Borgeaud, M. Arthur, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, L. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. Rae, O. Vinyals, and L. Sifre. 2022. training compute-optimal large language models. *Arxiv*.

Haris Jabbar. 2023. Morphpiece: Moving away from statistical language representation. *arXiv preprint arXiv:2307.07262*.

Bilal Khaliq and John Carroll. 2013. Induction of root and pattern lexicon for unsupervised morphological analysis of Arabic. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1012–1016, Nagoya, Japan. Asian Federation of Natural Language Processing.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Marion Di Marco and Alexander Fraser. 2024. Subword segmentation in LLMs: Looking at inflection and consistency. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12050–12060, Miami, Florida, USA. Association for Computational Linguistics.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Naoki Otani, Satoru Ozaki, Xingyuan Zhao, Yucen Li, Micaelah St Johns, and Lori Levin. 2020. Pre-tokenization of multi-word expressions in cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4451–4464, Online. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. Fineweb2: A sparkling update with 1000s of languages. *HuggingFace*.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov, and Stephan Vogel. 2017. Challenging language-dependent segmentation for Arabic: An application to machine translation and part-of-speech tagging. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 601–607,

Vancouver, Canada. Association for Computational Linguistics.

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.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching. *Technical Report DOI-TR-161, Department of Informatics, Kyushu University*.

Anya Sims, Thom Foster, Klara Kaleb, Tuan-Duy H Nguyen, Joseph Lee, Jakob N Foerster, Yee Whye Teh, and Cong Lu. 2025. Stochastok: Improving fine-grained subword understanding in llms. *arXiv preprint arXiv:2506.01687*.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 21–24, Gothenburg, Sweden. Aalto University, Association for Computational Linguistics.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Chengwei Wei, Yun-Cheng Wang, Bin Wang, C-C Jay Kuo, et al. 2024. An overview of language models: Recent developments and outlook. *APSIPA Transactions on Signal and Information Processing*, 13(2).

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the understanding of morphologically complex words in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1009–1020, Torino, Italia. ELRA and ICCL.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.