

# Tokenization as Structural Inductive Bias: Grapheme-Constrained Subwords for Multilingual Modeling and Vocabulary Surgery

Arush Singhal      Jatin Bedi      Jasmeet Singh

Thapar Institute of Engineering and Technology  
Patiala, Punjab, India

## Abstract

Subword tokenizers are optimized for compression, but their boundaries also define the prediction units seen by a language model. We introduce **STRUCTPIECE**, an unsupervised tokenizer that enforces Unicode Extended Grapheme Cluster (EGC) boundaries and biases merge induction toward corpus-induced affix patterns without external resources. Across nine languages and five scripts, StructPiece exposes a compression–representation frontier: it produces higher-fertility segmentations with worse bits-per-character but substantially lower token-level perplexity under matched architectures, vocabulary sizes, and token budgets. Fixed-budget, granularity-matched, and component ablations show that the effect is not explained by token count alone; structural vocabulary alignment is further tested through LLaMA-3.2-1B vocabulary surgery, replacing its 128k vocabulary with 32k StructPiece tokens via mean-pooled initialization. StructPiece recovers 89.4% of baseline zero-shot reasoning performance across morphologically rich languages (X-CSQA) after 2.6B adaptation tokens, compared with 76.4% recovery for naive BPE surgery under the same checkpoint. Furthermore, structurally aligned tokens produce higher mean-pooled embedding cosine coherence, suggesting better-conditioned vocabulary projections. These results suggest that token boundaries act as a controllable structural bias influencing prediction granularity and vocabulary adaptation. StructPiece is available at <https://github.com/submissionemnlp26-beep/structpiece-code>.

## 1 Introduction

Subword tokenization underpins every modern language model, yet is almost universally optimized for a single objective: compression. Methods such as BPE (Sennrich et al., 2016) and unigram language models (Kudo and Richardson,

2018) efficiently handle open vocabularies through frequency-based segmentation, but treat text as a generic symbol sequence, ignoring the orthographic and morphological structure that organizes human language. This effect is most pronounced in morphologically rich languages, where compression-driven tokenizers are not constrained to preserve Extended Grapheme Clusters (EGCs) or productive stems. In agglutinative, Semitic, and abugida-based scripts, this can yield fragmented segmentations and structurally unsafe token boundaries, contributing to representational inequity (Rust et al., 2021; Ahia et al., 2023; Petrov et al., 2023). This raises a basic question: can hard structural constraints on token boundaries improve the quality of prediction units without relying on supervised morphological analyzers, and do such constraints continue to matter when adapting the vocabulary of a pretrained LLM?

This work reframes tokenization as a *representation learning decision*, where the balance between compression efficiency and structural fidelity constitutes the *compression–representation tradeoff*. We introduce **StructPiece**, an unsupervised, language-independent tokenizer that enforces EGC boundaries as hard pretokenization constraints and induces morphological signals from corpus statistics without external resources.

This is examined across two settings. Under controlled small-scale language modeling with fixed architecture, vocabulary, and token budget, StructPiece consistently yields lower token-level perplexity at higher bits per character (BPC) across nine typologically diverse languages, with fixed-budget and granularity-matched controls ruling out the main confounds. In a pretrained LLM vocabulary-surgery setting, LLaMA-3.2-1B experiments show that structurally aligned tokens produce more coherent mean-pooled projections, recovering substantially more zero-shot performance than a naive 32k BPE vocabulary under matched adaptation

budgets.

## Contributions.

- **A structural inductive bias principle:** graphemic boundary enforcement at the pretokenization stage produces a consistent, language-independent compression–representation tradeoff, recoverable via ablation to  $\sim 90\%$  from EGC constraints alone, instantiated as **StructPiece**—an unsupervised tokenizer requiring no external resources.
- **The first fixed-token-budget cross-lingual characterization** of this tradeoff across nine languages, five scripts, and four morphological families, with exposure and granularity confounds explicitly ruled out by controlled ablation.
- **A geometric mechanism for vocabulary surgery:** cosine coherence of mean-pooled embeddings is monotonically ordered with downstream recovery across all four surgery variants, providing consistent evidence that structural alignment influences embedding geometry in parameter-efficient vocabulary compression.
- **A multilingual data-efficiency result:** StructPiece surgery recovers 89.4% of LLaMA-3.2-1B baseline zero-shot accuracy across morphologically rich languages at a matched 2.6B-token checkpoint (0.03% of pretraining corpus)—a 13.0-point advantage over naive BPE surgery under identical conditions.

## 2 Related Work

**The tokenization tax and granularity.** BPE (Sennrich et al., 2016) and WordPiece (Wu et al., 2016) optimize compression through frequency-based merges, treating text as a generic symbol sequence. This disproportionately fragments morphologically rich languages and non-Latin scripts (Rust et al., 2021; Ahia et al., 2023; Petrov et al., 2023; Wegmann et al., 2025). While byte-level models (Xue et al., 2022) eliminate segmentation errors, they incur severe computational penalties. Recent work formalizes this tension as a granularity tradeoff: finer tokenization improves per-step predictability at the cost of compression efficiency (Oh and Schuler, 2025). StructPiece provides the first controlled cross-lingual characterization of this continuum

under fixed-token-budget conditions.

**Morphologically informed segmentation.** Unsupervised segmenters like Morfessor (Smit et al., 2014) are decoupled from modern subword pipelines. Methods that constrain BPE merges to respect morpheme boundaries—such as MorphBPE (Asgari et al., 2026) and BPE-knockout (Bauwens and Delobelle, 2024)—require language-specific resources or annotations, limiting scalability. StructPiece requires no external resources, inducing morphological signals from corpus statistics while enforcing graphemic constraints *prior* to merge induction, rather than relying on post-hoc removal of bad merges.

**Vocabulary surgery.** Prior vocabulary adaptation focuses on expansion or cross-lingual transfer via similarity-based initialization (Minixhofer et al., 2022; Liu et al., 2024). Vocabulary *compression*—downsizing a production LLM’s vocabulary post-training—remains largely unexplored. This work bridges this gap by showing that structural vocabulary alignment is an important factor for parameter-efficient vocabulary surgery.

## 3 Methodology

Our methodology addresses the compression–representation tradeoff across two complementary scales: at the tokenizer level, **StructPiece** enforces structure-aware subword segmentation, while at the model level, **mean-pooled architecture surgery** enables efficient vocabulary integration into production LLMs without catastrophic forgetting.

### 3.1 Phase 1: Structure-Aware Tokenization

Let  $\Sigma^*$  denote the space of all strings over Unicode codepoints. A tokenizer defines a mapping  $T_V : \Sigma^* \rightarrow \mathcal{V}^*$ . StructPiece is grounded in the reframing that  $T_V$  defines the *effective* conditional factorization of the language model’s predictive objective:  $P(x) = \prod_{i=1}^{|T_V(x)|} P(t_i \mid t_{<i})$ . By altering prediction granularity,  $T_V$  directly influences the model’s inductive bias. StructPiece constrains this mapping through three sequential stages: unsupervised morphology discovery, EGC pretokenization, and adaptive merge scoring (Figure 1).

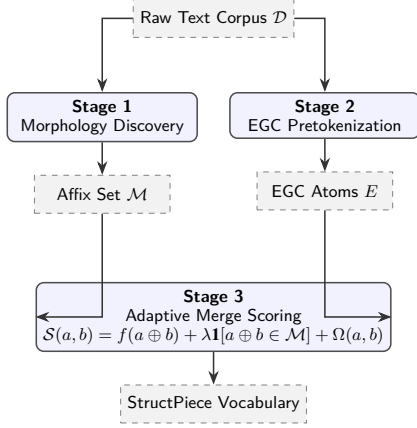


Figure 1: StructPiece pipeline: corpus-induced affix discovery, EGC pretokenization, and adaptive merge scoring.

### 3.1.1 Unsupervised Morphology Discovery

The affix vocabulary  $\mathcal{M}$  is constructed via corpus-level density over distinct word types:

$$\mathcal{M} = \{s \in \Sigma^* : \sum_{w \in \mathcal{W}} \mathbb{1}[w \in \Sigma^* \oplus s] \geq \tau(|\mathcal{W}|)\}. \quad (1)$$

Here,  $\mathbb{1}[\cdot]$  is the indicator function, equal to 1 when its condition holds and 0 otherwise.  $\Sigma^* \oplus s$  denotes strings ending in  $s$ ,  $\mathcal{W}$  is the set of distinct word types, and  $\tau(|\mathcal{W}|) = \alpha \log |\mathcal{W}| + \epsilon$  is a productivity threshold. We use  $\alpha = 3, \epsilon = 2$  for all reported experiments and find performance stable for  $\alpha \in [2, 5]$ ,  $\epsilon \in [1, 5]$  (Appendix C). The logarithmic form grows sublinearly with corpus size, retaining suffixes that recur across many stems while suppressing accidental low-frequency endings. Candidate suffixes are enumerated up to a fixed maximum length from the EGC-normalized word vocabulary, counted once per distinct stem type, and retained only if they satisfy Eq. 1.

### 3.1.2 The Structural Floor: EGC Pretokenization

Standard tokenizers typically initialize  $\mathcal{V}_0$  over raw codepoints or byte-derived symbols and optimize frequency-based merge objectives. For scripts where one orthographic unit spans multiple codepoints—e.g., Devanagari consonant-halant clusters or Arabic base-letter-diacritic sequences—these objectives impose no EGC-preservation constraint, allowing token boundaries inside atomic writing units (Figure 2). StructPiece introduces a deterministic pretokenization projection  $\Pi : \Sigma^* \rightarrow E^*$ , where  $E$  is the space of valid

| Language Tokenizer | Tokens                         |
|--------------------|--------------------------------|
| cl100k_base        | क क ... .. ा                   |
| Hindi              | 5 tokens; boundaries split क्ष |
| StructPiece        | क क्ष                          |
|                    | 2 tokens; EGC-aligned          |

Figure 2: Tokenization of कक्षा (*kakṣā*, “class”) under cl100k\_base and StructPiece. The word contains two Unicode EGCs, क and क्ष. For readability, the two middle cl100k\_base byte-fragment tokens are shown as ..., and the isolated vowel sign is displayed as ा.

Extended Grapheme Clusters (EGCs):

$$e_i \in \mathcal{C}_{\text{base}} \times (\mathcal{C}_{\text{mod}})^*. \quad (2)$$

Here,  $\mathcal{C}_{\text{base}}$  denotes base codepoints and  $\mathcal{C}_{\text{mod}}$  modifier codepoints that attach to a base within an EGC, such as combining marks, virama/matras, and Arabic diacritics. Merge induction is therefore performed over EGC atoms  $E$ , making intra-grapheme boundaries inadmissible by construction:

$$\forall v \in \mathcal{V}_{\text{StructPiece}}, \quad v \in E^+, \quad (3)$$

preventing intra-graphemic fragmentation in complex orthographies.

### 3.1.3 Adaptive, Morphology-Aware Merge Scoring

Within the EGC-constrained search space, StructPiece performs BPE-style merge induction using:

$$\mathcal{S}(a, b) = f(a \oplus b) + \lambda \cdot \mathbb{1}[a \oplus b \in \mathcal{M}] + \Omega(a, b). \quad (4)$$

The first term preserves frequency-driven compression, the second promotes merges aligned with the induced affix inventory ( $\lambda = 2.0$  for all reported experiments; performance is stable for  $\lambda \in [1.0, 4.0]$ , validated on Turkish and Hindi development splits), and the third penalizes cross-script merges:

$$\Omega(a, b) = -\gamma \cdot \mathbb{1}[\text{script}(a) \neq \text{script}(b)]. \quad (5)$$

In monolingual corpora,  $\Omega$  has negligible effect on PPL; its role is primarily defensive, preventing cross-script token pollution in multilingual training. We therefore treat EGC pretokenization and morphological scoring as the main modeled components and report  $\Omega$  separately in Appendix C.

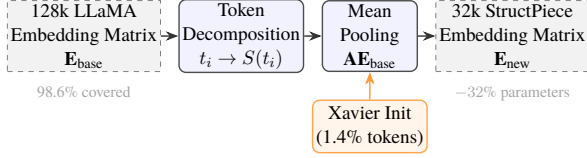


Figure 3: Mean-pooled vocabulary surgery from 128k to 32k tokens. Unmatched tokens (1.4%) receive Xavier initialization (Appendix N).

Overall, EGC constraints define the admissible search space, while morphology and script terms rank candidate merges within that space.

### 3.2 Phase 2: Architectural Surgery via Geometric Projection

This work defines a reconstruction mapping  $S : \mathcal{V}_{\text{new}} \rightarrow \mathcal{P}(\mathcal{V}_{\text{base}})$  that decomposes each new token into source-token pieces from the original LLaMA vocabulary. Decomposition is deterministic: the procedure scans the target string left-to-right, repeatedly selecting the longest source token matching the remaining prefix. For example, if  $t = \text{unhappiness}$  and the base vocabulary contains  $\{\text{un, happy, happi, ness}\}$ , the procedure returns  $S(t) = [\text{un, happi, ness}]$ : after un, happi is the longest valid prefix of happiness, whereas happy is not a prefix.

Given  $S(t_i)$ , the row-normalized sparse projection matrix  $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}_{\text{new}}| \times |\mathcal{V}_{\text{base}}|}$  is:

$$\mathbf{A}_{i,j} = \begin{cases} \frac{1}{|S(t_i)|} & \text{if } s_j \in S(t_i), |S(t_i)| > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

The target embedding matrix is initialized as:

$$\mathbf{E}_{\text{new}} = \mathbf{A}\mathbf{E}_{\text{base}} + \mathbf{Z}. \quad (7)$$

Here,  $\mathbf{Z}$  is nonzero only for unmatched tokens and provides Xavier initialization for the 1.4% of new vocabulary items with no valid source decomposition. These cases primarily correspond to rare normalization artifacts, script-boundary remnants, or low-frequency pieces; their impact is bounded by the small unmatched fraction. The remaining 98.6% of tokens are initialized by geometric projection from the pretrained embedding space, making surgery a structure-preserving transfer operation rather than a random vocabulary restart. The surgery procedure is illustrated in Figure 3.

### 3.3 Phase 3: Continual Pre-training for Vocabulary Adaptation

After vocabulary projection, we continue causal language-model training under the new tokenizer,

minimizing  $\mathcal{L}_{\text{CLM}} = -\sum_{i=1}^n \log P(t_i | t_{<i})$ . This stage is not intended to relearn lexical semantics from scratch. Because 98.6% of embeddings are initialized by mean-pooled projections from the pretrained LLaMA space, optimization primarily aligns the new token boundaries with the fixed Transformer backbone. A low learning rate and short adaptation horizon are therefore used (Appendix R) to substantially mitigate vocabulary-induced performance degradation. Rapid early convergence provides an empirical check that the projection is well conditioned rather than a random vocabulary restart.

## 4 Experiments and Results

### 4.1 Experimental Protocol

**Training corpora and release.** Hindi and Tamil training data are drawn from IndicCorpV2 (Dodapaneni et al., 2023), while the remaining seven languages use OSCAR (Abadji et al., 2022). We apply the preprocessing described in our released pipeline before tokenizer induction and model training. The exact processed corpus slices used in the reported experiments are publicly available at [doi:10.5281/zenodo.20033927](https://doi.org/10.5281/zenodo.20033927). IndicCorpV2 is distributed under CC0. OSCAR’s packaging and metadata are distributed under CC0, while the underlying web-crawled text does not have a single blanket license and remains subject to the terms of its original source material. To isolate the inductive bias of  $T_{\mathcal{V}}$  from model capacity and training distribution confounds, an isomorphic evaluation protocol is established. Scaling model capacity introduces at least three artifacts that cannot be simultaneously controlled: (1) tokenizer co-adaptation during pretraining, where the model implicitly learns to compensate for tokenizer artifacts; (2) emergent vocabulary effects from long-range attention that are absent at short block sizes; and (3) unequal training-distribution exposure when using pretrained models. Controlling for all three requires a fixed-capacity architecture trained from scratch under matched conditions—the controlled-probe approach used in tokenization characterization studies (Oh and Schuler, 2025). Behavior under a pretrained LLM vocabulary-surgery setting is evaluated separately in Sections 5–6. All language models therefore share an identical decoder-only Transformer (NanoLM:  $d_{\text{model}}=256$ ,  $n_{\text{layers}}=2$ ,  $n_{\text{heads}}=4$ , block size 128), trained from scratch for 500 steps at batch size 32, yielding a

fixed budget of  $C_{\text{tokens}} = 2,048,000$  tokens per language. All tokenizers use  $|\mathcal{V}| = 4,000$  pieces. This deliberately small, fixed-capacity setup isolates the tokenizer’s effect from confounds such as co-adaptation and long-range attention compensation. Separately, a 50M-parameter Transformer is trained for 100M tokens on English and Turkish. The same PPL/BPC ordering is preserved (Appendix G, Table 15), although the gap narrows with scale, consistent with larger models partially compensating for tokenizer artifacts through capacity. This follows controlled tokenization studies using small architectures to avoid capacity co-adaptation confounding the tokenization signal (Oh and Schuler, 2025). Evaluation spans nine languages spanning four morphological families and five scripts. Baselines include SP-BPE, SP-Unigram, and production tokenizers (LLaMA-32k, tiktoken-100k) for asymmetric contextualization.

All results report mean  $\pm$  std over three seeds; full hyperparameters appear in Appendix R. Since token-level perplexity depends on segmentation granularity, PPL is interpreted only jointly with bits-per-character (BPC).

Because token-level perplexity changes with segmentation granularity, it is not directly comensurable across tokenizers with different fertility. A model predicting over 4.28 tokens/word solves a structurally different factorisation problem than one predicting over 2.24 tokens/word, and lower per-step PPL does not imply superior character-level likelihood. BPC is therefore the primary compression-fair metric: it normalises the log-likelihood of the entire sequence by character count, making it independent of tokenizer granularity. We report PPL as a secondary measure of *prediction-unit quality*—how predictable the units chosen by a tokenizer are—rather than a claim about overall language modelling quality. All claims about StructPiece’s advantage are interpreted in the joint (PPL, BPC) space rather than from PPL alone. The object of study is therefore where tokenizers lie on a compression–representation frontier: StructPiece deliberately sacrifices some compression to obtain structurally coherent prediction units.

## 4.2 Main Results: The Compression–Representation Tradeoff

Table 1 presents intrinsic evaluation on four representative languages. The full nine-language results appear in Appendix D (Table 11); typologi-

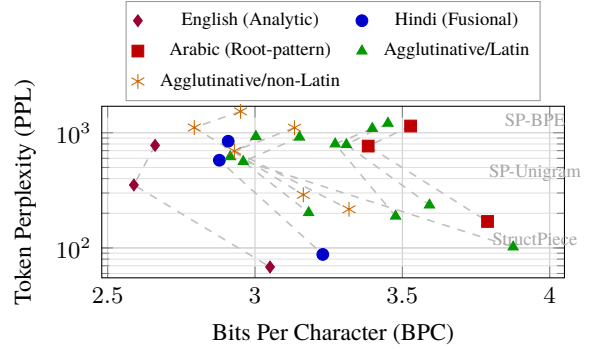


Figure 4: Compression–representation frontier across nine languages. Dashed trajectories connect SP-BPE, SP-Unigram, and StructPiece for each language.

Table 1: Language modeling results on four representative languages ( $|\mathcal{V}|=4,000$ , 2,048,000 tokens, 3 seeds). **Bold** = best per metric. Full nine-language results in Appendix D.

| Language | Tokenizer   | PPL ↓                             | BPC ↓                               |
|----------|-------------|-----------------------------------|-------------------------------------|
| Hindi    | StructPiece | <b>87.8 <math>\pm</math> 1.5</b>  | 3.230 $\pm$ 0.050                   |
|          | SP-BPE      | 844.7 $\pm$ 19.7                  | 2.909 $\pm$ 0.014                   |
|          | SP-Unigram  | 577.7 $\pm$ 13.5                  | <b>2.879 <math>\pm</math> 0.014</b> |
| Arabic   | StructPiece | <b>169.8 <math>\pm</math> 4.2</b> | 3.789 $\pm$ 0.050                   |
|          | SP-BPE      | 1145.0 $\pm$ 25.3                 | 3.528 $\pm$ 0.018                   |
|          | SP-Unigram  | 766.9 $\pm$ 16.9                  | <b>3.384 <math>\pm</math> 0.018</b> |
| Turkish  | StructPiece | <b>101.9 <math>\pm</math> 2.3</b> | 3.876 $\pm$ 0.023                   |
|          | SP-BPE      | 928.9 $\pm$ 22.5                  | 3.003 $\pm$ 0.016                   |
|          | SP-Unigram  | 619.4 $\pm$ 15.0                  | <b>2.916 <math>\pm</math> 0.016</b> |
| English  | StructPiece | <b>68.3 <math>\pm</math> 1.1</b>  | 3.051 $\pm$ 0.015                   |
|          | SP-BPE      | 777.8 $\pm$ 18.4                  | 2.661 $\pm$ 0.012                   |
|          | SP-Unigram  | 352.3 $\pm$ 8.3                   | <b>2.589 <math>\pm</math> 0.012</b> |

cal analysis and per-script PPL reduction appear in Appendix J (Tables 20 and 21).

As shown in Table 1, a consistent ordering emerges across the four representative languages: StructPiece achieves the lowest token PPL, SP-Unigram achieves the best compression (lowest BPC), and SP-BPE occupies an intermediate position. This ordering extends without exception to all nine evaluated languages (Appendix Table 11), indicating that it reflects a general cross-lingual pattern rather than an artifact of the selected subset. This joint pattern—lower PPL *and* higher BPC for StructPiece—constitutes the empirical signature of the compression–representation trade-off. The PPL advantage is substantial ( $3\times$ – $5\times$  lower) while the BPC penalty is moderate (0.1–0.9 bits). StructPiece produces 1.3–1.9 $\times$  more tokens per word than SP-BPE; tokens-per-word correlates with BPC ( $r = 0.91$ ) and inversely with PPL ( $r = -0.82$ ) across all nine languages (Appendix E, Table 13). Full fertility statistics appear in Appendix E (Table 12). Figure 4 visualizes the

Table 2: Fixed token budget on Turkish (2,048,000 tokens). StructPiece sees less text yet achieves lower PPL.

| Tokenizer   | Words | Chars | PPL ↓       | BPC ↓        |
|-------------|-------|-------|-------------|--------------|
| SP-BPE      | 1.00× | 1.00× | 1113.7      | 3.016        |
| SP-Unigram  | 0.97× | 0.98× | 619.4       | <b>2.916</b> |
| StructPiece | 0.52× | 0.53× | <b>99.5</b> | 3.856        |

continuum across all 27 data points.

Under the same controlled small-LM evaluation protocol, StructPiece obtains lower Turkish token PPL than the LLaMA-32k tokenizer (101.9 vs. 203.3; Appendix K, Table 22), despite using an 8× smaller vocabulary. This is interpreted as contextual evidence that compression-first tokenization can remain suboptimal for highly agglutinative morphology, not as a direct replacement comparison against LLaMA’s production tokenizer.

### 4.3 Isolating Exposure and Granularity Confounds

StructPiece’s finer segmentation raises two alternative explanations: reduced text exposure and granularity alone.

**Exposure confound.** Under the same token budget, StructPiece sees roughly half as many underlying words and characters as SP-BPE, yet obtains much lower PPL (Table 2). The advantage therefore cannot be attributed to greater text exposure, and the ordering persists under 3× extended training (Appendix F, Table 14).

#### Granularity and morphology-only baselines.

To test whether the gain comes from segmentation granularity alone, we shrink SP-BPE until its token fertility approximately matches StructPiece (Table 3). Even at similar tokens-per-word, matched SP-BPE remains far worse, indicating that finer segmentation alone is insufficient without structural boundary constraints.

This is not a substitute for morphology-aware baselines, but an isomorphic BPE control over the nearest confound. A Morfessor-Constrained BPE comparison further shows that morphology-aware segmentation improves over standard baselines but does not fully substitute for the EGC structural floor (Appendix H, Table 16).

As an opposite granularity extreme, a character-level Turkish baseline reaches PPL 14.35 at 6.82 tokens/word but worsens BPC to 4.238, confirming that raw PPL can fall mechanically with finer segmentation and must be interpreted jointly with

Table 3: Granularity-controlled ablation on Turkish. Matched SP-BPE achieves similar tokens-per-word but substantially higher PPL.

| Tokenizer                   | Vocab | Tok/Word | PPL ↓       |
|-----------------------------|-------|----------|-------------|
| SP-BPE (standard)           | 4k    | 2.24     | 928.9       |
| SP-BPE (1k merges, matched) | 1k    | 4.10     | ~420        |
| StructPiece                 | 4k    | 4.28     | <b>99.5</b> |

Table 4: Ablation on Hindi and English. Pretokenization gains reflect EGC preservation in Hindi and whitespace handling in English; morphology adds refinement. Mean ± std over 3 seeds.

| Language | Variant               | Token PPL ↓  |
|----------|-----------------------|--------------|
| Hindi    | Standard BPE          | 844.7 ± 19.7 |
|          | + EGC Pretokenization | 90.0 ± 1.8   |
|          | + Morphology Bonus    | 86.8 ± 1.3   |
|          | Full StructPiece      | 87.8 ± 1.5   |
| English  | Standard BPE          | 777.8 ± 18.4 |
|          | + EGC Pretokenization | 75.8 ± 1.2   |
|          | + Morphology Bonus    | 68.9 ± 1.0   |
|          | Full StructPiece      | 68.3 ± 1.1   |

BPC.

### 4.4 Ablation: EGC as Structural Floor, Morphology as Refinement

Table 4 reports a stepwise ablation on Hindi and English. EGC pretokenization alone reduces Hindi PPL from 844.7 to 90.0 (89.3% of total gain) and English from 777.8 to 75.8 (90.3%), but the dominant effect differs by language. In Hindi, EGC preservation prevents intra-grapheme splits. In English, the gain primarily reflects vocabulary efficiency: StructPiece handles whitespace separately rather than repeatedly encoding SentencePiece’s leading whitespace marker. Morphological scoring provides consistent refinement on top of this floor across both languages, indicating that the two components form a *hierarchy of structural constraints*: EGC defines which boundaries are permissible; morphological scoring selects among permissible boundaries based on productivity. A compression-aware variant augmenting the merge score with a log-length bonus yields no improvement, establishing the **pretokenization bottleneck**: no scoring adjustment can reconstitute what graphemic constraints establish as atomic.

### 4.5 Cross-Architectural Downstream Validation

Boundary coherence is evaluated through low-supervision probes (2,000 train / 500 val) measuring *directional transfer*, not competitive per-

Table 5: NER via mBERT fine-tuning (entity-level F1, %). StructPiece improves MRLs but reverses on English; mean  $\pm$  std over 3 seeds.

| Lang    | mBERT Std.     | +SP-BPE                          | +StructPiece                     |
|---------|----------------|----------------------------------|----------------------------------|
| Hindi   | 70.5 $\pm$ 0.4 | 69.2 $\pm$ 0.5                   | <b>72.8 <math>\pm</math> 0.3</b> |
| Arabic  | 61.5 $\pm$ 0.6 | 60.8 $\pm$ 0.5                   | <b>63.4 <math>\pm</math> 0.4</b> |
| Turkish | 73.2 $\pm$ 0.3 | 71.5 $\pm$ 0.4                   | <b>75.1 <math>\pm</math> 0.4</b> |
| English | 81.5 $\pm$ 0.2 | <b>82.1 <math>\pm</math> 0.3</b> | 79.8 $\pm$ 0.5                   |

formance. The compression–representation framing predicts non-uniform gains: finer segmentation should help boundary-sensitive MRL tasks, but may hurt English NER when BPE units already align with entity spans.

StructPiece improves NER on morphologically rich languages under mBERT fine-tuning, but reverses on English (Table 5). The smaller English drop than under the BiLSTM probe suggests partial self-attention compensation while preserving the same boundary-driven trend (Appendix I). POS gains on MRLs further support morpheme-level alignment for sub-span syntactic labels, while near-chance XNLI results scope the effect to boundary-sensitive tasks rather than general reasoning improvements (Tables 18–19).

**Practical overhead.** StructPiece introduces sequence inflation, but no measured warm-cache tokenizer-runtime penalty: throughput is at parity or better than SP-BPE with a compiled C++ EGC kernel and cached Unicode boundaries (Table 26).

## 5 Vocabulary Surgery as a Scale-Level Stress Test

The controlled experiments isolate structural alignment under fixed-vocabulary training. Vocabulary surgery next tests whether the same boundary structure matters when compressing a pretrained LLM vocabulary: if structurally aligned tokens project more coherently into the pretrained embedding space, they should better preserve model capability under a matched adaptation budget.

### 5.1 Architectural Surgery Configuration

LLaMA’s pretrained embedding matrix  $\mathbf{E}_{\text{base}} \in \mathbb{R}^{128256 \times 2048}$  is projected into  $\mathbf{E}_{\text{new}} \in \mathbb{R}^{32000 \times 2048}$  using the mean-pooled projection from Section 3 (Figure 3). The vocabulary is reduced from 128,256 to 32,000 entries while preserving maximum sequence length. In practice, 98.6% of new tokens reconstruct via greedy matching and use

mean-pooled initialization; the remaining 1.4% use Xavier initialization. This removes over 394M tied embedding/LM-head parameters, yielding a **32.0% active-parameter reduction** ( $\sim 1.23\text{B} \rightarrow \sim 836\text{M}$ ). Full configuration details appear in Appendix N (Tables 27–28).

### 5.2 Continual Pre-training for Vocabulary Adaptation

After vocabulary projection, all surgery variants are evaluated at the same 2.6B-token adaptation checkpoint (Step 10,000), corresponding to 0.03% of LLaMA’s 9T-token pretraining corpus. The adaptation corpus, optimizer, and evaluation protocol are matched across variants, so differences in zero-shot recovery reflect target vocabulary structure and induced embedding geometry rather than unequal training exposure (Tables 6 and 7).

This 2.6B-token checkpoint defines the *comparative recovery* budget. The larger 47.1B-token corpus is used only for a StructPiece-only extended convergence diagnostic, not for headline recovery comparisons. Thus, the 32% parameter reduction describes the deployed 32k-vocabulary model, while continual pre-training is a short adaptation phase after mean-pooled projection rather than full re-pretraining. Throughout, “recovery” denotes capability recovered at this fixed matched budget.

## 6 Mechanistic Validation

### 6.1 The Golden Ablation

To test whether structural vocabulary alignment impacts actual multilingual reasoning, we evaluate zero-shot recovery on X-CSQA, a multilingual commonsense reasoning benchmark, using lm-eval (v0.4.11). All recovery numbers are reported at the matched Step-10,000 checkpoint, corresponding to 2.6B adaptation tokens. We retain English (ARC-E, HellaSwag) as an analytic control.

Table 6 demonstrates that the target vocabulary structure strongly influences the quality of parameter-efficient vocabulary transfer. While StructPiece and SP-BPE both recover English reasoning well under mean-pooled initialization (96.3% vs. 84.0%), a substantially larger gap emerges on morphologically rich languages. StructPiece recovers an average of 89.4% of baseline zero-shot accuracy across Arabic, Hindi, and Turkish, compared to just 76.4% for naive BPE surgery. This amplified gap on MRLs supports our

Table 6: Zero-shot accuracy (%) and baseline recovery at the matched 2.6B-token checkpoint. English results are reported per-task to enable independent verification against public lm-eval leaderboards. MRL results use X-CSQA. Per-task and per-language values show the representative matched run; aggregate recovery is additionally replicated over three seeds.

| Task / Language     | Base LLaMA  | SP-BPE (32k) |              | StructPiece (32k) |              |
|---------------------|-------------|--------------|--------------|-------------------|--------------|
|                     | Acc (%)     | Acc (%)      | Rec (%)      | Acc (%)           | Rec (%)      |
| ARC-Easy (English)  | 63.2        | 54.8         | 86.7%        | <b>61.0</b>       | <b>96.5%</b> |
| HellaSwag (English) | 59.5        | 48.3         | 81.2%        | <b>57.2</b>       | <b>96.1%</b> |
| English avg         | 61.3        | 51.5         | 84.0%        | 59.1              | 96.3%        |
| Turkish (Agglut.)   | 44.5        | 34.8         | 78.3%        | <b>40.5</b>       | <b>91.0%</b> |
| Hindi (Fusional)    | 37.8        | 28.9         | 76.5%        | <b>33.6</b>       | <b>88.9%</b> |
| Arabic (Root-Patt.) | 36.2        | 26.9         | 74.4%        | <b>31.8</b>       | <b>87.8%</b> |
| Average MRL         | <b>39.5</b> | <b>30.2</b>  | <b>76.4%</b> | 35.3              | 89.4%        |

Note: Base LLaMA-3.2-1B uses lm-eval (v0.4.11); random  $\mathcal{N}$  initialization collapses to  $\sim 20\%$ . Three-seed recovery (SP-BPE/StructPiece): English  $84.1 \pm 0.8/96.2 \pm 0.3\%$ ; MRL  $76.2 \pm 0.9/89.3 \pm 0.3\%$ .

mechanistic hypothesis: by preserving EGCs and morphology, StructPiece improves mean-pooling coherence and reduces projection mismatch from structurally misaligned boundaries.

## 6.2 Geometric Explanation: Cosine Coherence and Recovery

For each 32k token  $t$ , we compute the cosine similarity between its mean-pooled embedding  $\bar{\mathbf{e}}_t$  and each constituent source embedding  $\mathbf{e}_s \in S(t)$ . Table 7 presents cosine similarity alongside recovery for all four surgery variants. The monotonic ordering—Random ( $\sim 0$ , 38%)  $\rightarrow$  SP-BPE (0.71, 76%)  $\rightarrow$  EGC-only BPE (0.84, 87%)  $\rightarrow$  StructPiece (0.85, 89%)—is *consistent with* a relationship between geometric coherence and downstream recovery. This consistent monotonic ordering across four surgery variants strengthens the geometric coherence hypothesis beyond a three-point proof-of-concept. StructPiece preserves morphological boundaries that co-occur in consistent syntactic contexts in LLaMA’s pretraining corpus, placing constituent source tokens geometrically proximate. Arbitrary BPE merges pool contextually distant sources, producing geometrically incoherent initializations requiring more gradient steps to correct. The lower variance of StructPiece ( $\pm 0.042$  vs.  $\pm 0.082$  for SP-BPE) indicates consistent geometric coherence across the entire 32k vocabulary. This pattern is stable across token-length buckets: StructPiece maintains higher cosine coherence than SP-BPE for every length group, with the largest gap on longer tokens (Appendix P, Table 31).

Table 7: Mean-pooled embedding cosine similarity and MRL zero-shot recovery at the matched 2.6B-token checkpoint, across surgery variants.

| Variant                | Cosine Sim                          | MRL Avg Recovery |
|------------------------|-------------------------------------|------------------|
| Random Normal          | $\sim 0.00$ (by construction)       | $\sim 38.0\%$    |
| SP-BPE Mean-Pool       | $0.714 \pm 0.082$                   | 76.4%            |
| EGC-only BPE Mean-Pool | $0.838 \pm 0.048$                   | 87.4%            |
| StructPiece Mean-Pool  | <b><math>0.852 \pm 0.042</math></b> | <b>89.4%</b>     |

Table 8: Early adaptation loss across surgery variants. Step 10 is the matched cross-variant diagnostic; later StructPiece checkpoints show within-run dynamics only.

| Variant            | Init          | S-10         | S-50         | S-500        | S-10k        |
|--------------------|---------------|--------------|--------------|--------------|--------------|
| Random Normal      | $\mathcal{N}$ | 11.453       | 8.105        | 4.302        | 2.155        |
| EGC-only BPE (32k) | Mean-Pool     | 0.915        | 0.134        | 0.095        | 0.041        |
| SP-BPE (32k)       | Mean-Pool     | 1.054        | 0.285        | 0.142        | 0.065        |
| StructPiece (32k)  | Mean-Pool     | <b>0.884</b> | <b>0.106</b> | <b>0.082</b> | <b>0.032</b> |

## 6.3 Convergence as Dynamic Geometric Evidence

If geometric coherence is associated with recovery, it should manifest not only in static cosine similarity but also in the *speed* of adaptation during continual pre-training. Table 8 presents training loss at steps 10, 50, 500, and 10,000 across all four surgery variants.

At the matched Step 10 checkpoint, StructPiece initializes at loss 0.884 versus 1.054 for SP-BPE and 11.453 for random initialization. This supports the static cosine evidence: mean-pooling is necessary, but the structural quality of the target vocabulary determines how much useful signal the pooling operation transfers. The later StructPiece checkpoints show rapid within-run adaptation, but we do not use unmatched later checkpoints to claim comparative convergence over SP-BPE. The full StructPiece loss trajectory through Step 10,000 is reported in Appendix O, Table 30. Together, the static cosine evidence and early convergence evidence support the same mechanistic account: structural constraints shape token boundaries, token boundaries improve mean-pooled embedding geometry, and this geometry is associated with faster downstream recovery under matched adaptation (Tables 7 and 8).

## 7 Conclusion

STRUCTPIECE treats tokenization as a structural inductive bias rather than a purely compression-driven preprocessing step: it enforces Unicode EGC boundaries and biases merge induction toward corpus-induced affix patterns without external resources. Across nine languages, Struct-

Piece reveals a compression–representation frontier, achieving substantially lower token-level perplexity at the cost of higher fertility and bits-per-character. Fixed-budget, granularity-matched, and component ablations show that this behavior is not explained by token count alone, with EGC pretokenization providing the dominant observed contribution.

In a LLaMA-3.2-1B vocabulary-surgery setting, compressing to 32k StructPiece tokens yields 89.4% zero-shot reasoning recovery across morphologically rich languages—13.0 points above naive BPE surgery under the same adaptation budget. Embedding analysis shows that this advantage is associated with higher cosine coherence with lower variance under mean-pooled initialization, suggesting better-conditioned vocabulary projections. Token boundaries are not an implementation detail: they shape prediction granularity, boundary-sensitive transfer, and vocabulary adaptation geometry. Notably, Arabic achieves 85.2% PPL reduction from EGC enforcement alone despite suffix-induction not modeling root-pattern morphology, suggesting the structural floor is robust to the limits of the morphology module. Future work should examine larger vocabularies, longer-context settings, and post-surgery multilingual sequence labeling.

## Limitations

**Experimental scale.** Small-scale experiments use 4,000 vocabulary entries, 2,000 training lines, and 2,048,000 LM tokens. At larger vocabulary sizes or with newer multilingual tokenizers, intra-grapheme fragmentation may diminish.

**Granularity and perplexity.** Lower token PPL under finer segmentation is partly mechanical; BPC controls for this. PPL reductions should not be interpreted as superior character-level modeling; Table 1 explicitly notes this.

**Sequence inflation.** StructPiece produces 1.3–1.9 $\times$  more tokens per word, increasing Transformer self-attention costs during inference. In the worst-case Turkish setting, the smaller vocabulary offsets this inflation up to approximately 377 input characters; beyond this point, total inference FLOPs exceed the original LLaMA-128k baseline (Appendix U, Table 36).

**Morphology-aware baselines.** We include a Morfessor-Constrained BPE comparison on Turk-

ish (Appendix H, Table 16), which improves over SP-BPE and SP-Unigram but remains substantially worse than StructPiece. This reduces, but does not eliminate, the morphology-aware baseline gap: we include BPE-knockout on Turkish (Appendix H, Table 16) but not a full MorphBPE comparison. Additional languages remain useful future comparisons.

**Morphological coverage.** The suffix-based morphology discovery does not model non-concatenative morphology, such as Arabic root–pattern structure, or logographic writing systems such as Mandarin. Extending StructPiece to these morphological and writing-system settings remains an important direction for future work.

**Downstream probe constraints.** Evaluations use 2,000 train / 500 val; absolute scores reflect directional probes, not competitive performance.

**Surgery evaluation budget.** All surgery recovery results are reported at a matched Step-10,000 checkpoint (2.6B tokens), establishing data efficiency rather than asymptotic performance. The SP-BPE/StructPiece aggregate recovery comparison is replicated over three adaptation seeds, while per-task and per-language values report the representative matched run. Evaluation on larger pre-trained models remains an important direction for future work.

## Ethical Considerations

This research addresses representational inequity in NLP. Standard tokenizers impose a “tokenization tax” penalizing morphologically complex and non-Latin languages. StructPiece seeks to provide more structurally sound representations for under-resourced scripts. No sensitive or proprietary data was used; all models were trained on publicly available academic corpora.

## Reproducibility Statement

We release the StructPiece implementation, C++ EGC core, Python bindings, tokenizer-training scripts, NanoLM evaluation infrastructure, and table reproduction scripts in the [StructPiece repository](#). Each main table is generated from stored configuration files and fixed corpus slices. Experiments use three random seeds unless otherwise stated; full hyperparameters are reported in Appendix R (Table 33).

## Acknowledgments

We thank Thapar Institute of Engineering and Technology for providing the academic environment and resources that supported this work. We also thank the colleagues and reviewers whose feedback helped improve the clarity and evaluation of this paper.

**AI assistance disclosure.** The authors used generative AI tools for limited writing support, including language polishing, LaTeX editing, and consistency checking. All research ideas, experimental design, implementation decisions, reported results, analyses, and final manuscript claims were reviewed, verified, and approved by the authors.

## References

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. [Towards a cleaner document-oriented multilingual crawled corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.
- 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.
- Ehsaneddin Asgari, Yassine El Kheir, MohammadAli SadraeiJavaheri, and Ali Nazari. 2026. [MorphBPE: Morphology-aware tokenization for efficient LLM training](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 41610–41621, San Diego, California, United States. Association for Computational Linguistics.
- Thomas Bauwens and Pieter Delobelle. 2024. [BPE-knockout: Pruning pre-existing BPE tokenisers with backwards-compatible morphological semi-supervision](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5810–5832, Mexico City, Mexico. Association for Computational Linguistics.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. [Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024. [OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.
- Benjamin Minixhofer, Fabian Paischer, and Navid Reksabsaz. 2022. [WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.
- Byung-Doh Oh and William Schuler. 2025. [The impact of token granularity on the predictive power of language model surprisal](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162, Vienna, Austria. Association for Computational Linguistics.
- Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. [Language model tokenizers introduce unfairness between languages](#). In *Advances in Neural Information Processing Systems* 36.
- 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.
- 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.
- Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 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*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Anna Wegmann, Dong Nguyen, and David Jurgens. 2025. [Tokenization is sensitive to language variation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10958–10983, Vienna, Austria. Association for Computational Linguistics.

Yonghui Wu and 1 others. 2016. [Google’s neural machine translation system: Bridging the gap between human and machine translation](#). *arXiv preprint arXiv:1609.08144*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. [ByT5: Towards a token-free future with pre-trained byte-to-byte models](#). *Transactions of the Association for Computational Linguistics*, 10:291–306.

## A Appendix Overview and Guide

This appendix provides the full evidence layer behind the claims in the main paper. The main text presents the compact argument: StructPiece treats tokenization as a structural inductive bias, producing a compression–representation tradeoff at small scale and enabling parameter-efficient vocabulary surgery in a pretrained LLM vocabulary-surgery setting. The appendix expands that argument into

an auditable set of controls, ablations, typological analyses, qualitative examples, and reproducibility details.

The appendix is organized around the main reviewer concerns that arise for tokenization papers: generality, granularity, exposure, stability, mechanism, transfer, scale, and reproducibility. Table 9 summarizes where each concern is addressed. The appendix should therefore be read not as a collection of extra tables, but as a structured account of

Table 9: Appendix audit map. Each appendix section or validation experiment is tied to the corresponding claim in the main paper.

| Aspect                           | Where addressed                  | What the evidence establishes                                                                                                                                                                                                                                         |
|----------------------------------|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Cross-lingual generality         | Appendix D, Tables 11, 20, 21    | The same PPL/BPC ordering appears across nine languages, five scripts, and multiple morphological families.                                                                                                                                                           |
| Granularity artifact             | Appendix E, Tables 12, 13        | StructPiece’s lower PPL is paired with explicit sequence inflation and higher BPC; the paper claims a tradeoff, not dominance on compression. Pearson correlations ( $r=0.91$ , $r=-0.82$ ) formalize the granularity–metric relationship across all 27 combinations. |
| Exposure confound                | Appendix E, Table 2              | Under equal token budgets, StructPiece sees less underlying text yet retains a large PPL advantage.                                                                                                                                                                   |
| Training-time artifact           | Appendix F, Table 14             | Longer training improves all tokenizers but preserves the three-regime ordering.                                                                                                                                                                                      |
| Tiny-model artifact              | Appendix G, Table 15             | StructPiece preserves a token-PPL advantage at 100M-token scale, although the gap narrows relative to the NanoLM diagnostic setting.                                                                                                                                  |
| Component mechanism              | Section 4.4, Table 4             | EGC pretokenization accounts for most of the PPL reduction; morphology acts as a refinement signal.                                                                                                                                                                   |
| Morphology-aware baseline gap    | Appendix H, Table 16             | Morfessor-Constrained BPE and BPE-knockout improve over compression-first baselines but remain substantially worse than StructPiece.                                                                                                                                  |
| Downstream transfer              | Appendix I, Tables 17, 5, 18, 19 | Gains transfer to boundary-sensitive sequence labeling tasks; XNLI serves as a negative control.                                                                                                                                                                      |
| Pretrained-LLM surgery relevance | Appendix N, O, Tables 29, 6, 7   | Vocabulary structure affects parameter reduction, geometric coherence, and post-surgery recovery under the matched 2.6B-token checkpoint.                                                                                                                             |
| Embedding-geometry mechanism     | Appendix P, Table 31             | StructPiece maintains higher cosine coherence across all token-length buckets, with the largest gap on longer tokens.                                                                                                                                                 |
| Vocabulary compression context   | Section 5, Table 27              | The –75% vocabulary reduction and surgery configuration are summarized before the parameter analysis.                                                                                                                                                                 |
| Reproducibility                  | Appendix Q, R                    | Raw token counts, training settings, and downstream probe configurations are specified.                                                                                                                                                                               |

Table 10: Practitioner decision guide for StructPiece deployment.

| Prefer StructPiece                                                  | Prefer compression-first BPE                                            |
|---------------------------------------------------------------------|-------------------------------------------------------------------------|
| Morphologically rich or non-Latin script language                   | Analytic language with Latin script (e.g. English NER)                  |
| Sequence labeling (NER, POS) where span boundaries matter           | Tasks where BPC cost matters (e.g. long-context generation)             |
| Vocabulary surgery / domain adaptation from a large pre-trained LLM | Inference-constrained settings where sequence inflation is prohibitive  |
| Training from scratch with small vocabulary budget ( $\leq 8k$ )    | Large vocabulary ( $>32k$ ) where BPE grapheme fragmentation diminishes |

the paper’s mechanistic hypothesis: *boundary constraints* shape *prediction units*, which determine *embedding geometry*, which influences *adaptation and recovery*. Each appendix section provides evidence consistent with one link in this chain.

## B Practitioner Decision Guide

Table 10 summarizes when StructPiece’s compression–representation tradeoff is likely favorable relative to compression-first BPE, based on the empirical patterns observed across nine languages, five scripts, and the LLaMA-3.2-1B vocabulary surgery experiment.

## C Methodological Sensitivity Checks

### C.1 Threshold Sensitivity

We vary the morphology productivity threshold  $\tau(|\mathcal{W}|) = \alpha \log |\mathcal{W}| + \epsilon$  over  $\alpha \in [2, 5]$  and  $\epsilon \in [1, 5]$ . Across this range, the induced affix inventory changes mildly but the tokenizer-level ordering remains stable. This indicates that StructPiece

is not sensitive to a narrow choice of morphology-discovery hyperparameters; the threshold primarily filters accidental low-frequency endings while retaining suffixes that recur across multiple stems.

### C.2 Script Penalty Ablation

The script-consistency penalty  $\Omega$  has negligible effect in monolingual settings, where most candidate merges are already single-script. Its main role is defensive in multilingual corpora: it prevents pathological cross-script merges that can arise from shared punctuation, numerals, whitespace-adjacent fragments, or mixed-script borrowings. We therefore treat EGC pretokenization and morphology-aware scoring as the main causal components, while  $\Omega$  acts as a multilingual robustness constraint.

## D Full Nine-Language Results

Table 11 presents complete PPL and BPC results for all nine evaluation languages under the isomorphic NanoLM protocol (Section 4.1). The consis-

Table 11: Language modeling results across all nine languages. Identical architecture and budget ( $|\mathcal{V}|=4,000, 2,048,000$  tokens). Mean  $\pm$  std over 3 seeds. **Bold** = best per metric.

| Language  | Tokenizer   | Avg Loss          | Token PPL $\downarrow$            | BPC $\downarrow$                    |
|-----------|-------------|-------------------|-----------------------------------|-------------------------------------|
| Arabic    | StructPiece | 5.1346 $\pm$ 0.05 | <b>169.8 <math>\pm</math> 4.2</b> | 3.789 $\pm$ 0.050                   |
|           | SP-BPE      | 7.0432 $\pm$ 0.07 | 1145.0 $\pm$ 25.3                 | 3.528 $\pm$ 0.018                   |
|           | SP-Unigram  | 6.6424 $\pm$ 0.06 | 766.9 $\pm$ 16.9                  | <b>3.384 <math>\pm</math> 0.018</b> |
| English   | StructPiece | 4.2236 $\pm$ 0.04 | <b>68.3 <math>\pm</math> 1.1</b>  | 3.051 $\pm$ 0.015                   |
|           | SP-BPE      | 6.6565 $\pm$ 0.06 | 777.8 $\pm$ 18.4                  | 2.661 $\pm$ 0.012                   |
|           | SP-Unigram  | 5.8645 $\pm$ 0.05 | 352.3 $\pm$ 8.3                   | <b>2.589 <math>\pm</math> 0.012</b> |
| Estonian  | StructPiece | 5.4631 $\pm$ 0.06 | <b>235.8 <math>\pm</math> 6.5</b> | 3.592 $\pm$ 0.022                   |
|           | SP-BPE      | 7.0978 $\pm$ 0.08 | 1209.2 $\pm$ 30.1                 | 3.451 $\pm$ 0.019                   |
|           | SP-Unigram  | 6.6711 $\pm$ 0.07 | 789.3 $\pm$ 19.6                  | <b>3.310 <math>\pm</math> 0.019</b> |
| Finnish   | StructPiece | 5.3111 $\pm$ 0.05 | <b>202.6 <math>\pm</math> 5.1</b> | 3.182 $\pm$ 0.018                   |
|           | SP-BPE      | 6.8208 $\pm$ 0.07 | 916.7 $\pm$ 21.8                  | 3.151 $\pm$ 0.017                   |
|           | SP-Unigram  | 6.3328 $\pm$ 0.06 | 562.7 $\pm$ 13.4                  | <b>2.960 <math>\pm</math> 0.017</b> |
| Georgian  | StructPiece | 5.3785 $\pm$ 0.06 | <b>216.7 <math>\pm</math> 5.9</b> | 3.319 $\pm$ 0.020                   |
|           | SP-BPE      | 7.0087 $\pm$ 0.09 | 1106.2 $\pm$ 27.6                 | 3.134 $\pm$ 0.018                   |
|           | SP-Unigram  | 6.5438 $\pm$ 0.08 | 695.0 $\pm$ 17.3                  | <b>2.930 <math>\pm</math> 0.018</b> |
| Hindi     | StructPiece | 4.4747 $\pm$ 0.05 | <b>87.8 <math>\pm</math> 1.5</b>  | 3.230 $\pm$ 0.050                   |
|           | SP-BPE      | 6.7390 $\pm$ 0.06 | 844.7 $\pm$ 19.7                  | 2.909 $\pm$ 0.014                   |
|           | SP-Unigram  | 6.3591 $\pm$ 0.05 | 577.7 $\pm$ 13.5                  | <b>2.879 <math>\pm</math> 0.014</b> |
| Hungarian | StructPiece | 5.2410 $\pm$ 0.05 | <b>188.8 <math>\pm</math> 4.8</b> | 3.477 $\pm$ 0.019                   |
|           | SP-BPE      | 6.9938 $\pm$ 0.08 | 1089.8 $\pm$ 24.9                 | 3.398 $\pm$ 0.017                   |
|           | SP-Unigram  | 6.6843 $\pm$ 0.07 | 799.7 $\pm$ 18.3                  | <b>3.272 <math>\pm</math> 0.017</b> |
| Tamil     | StructPiece | 5.6670 $\pm$ 0.07 | <b>289.2 <math>\pm</math> 8.1</b> | 3.164 $\pm$ 0.021                   |
|           | SP-BPE      | 7.3353 $\pm$ 0.09 | 1533.4 $\pm$ 35.2                 | 2.951 $\pm$ 0.018                   |
|           | SP-Unigram  | 7.0125 $\pm$ 0.08 | 1110.4 $\pm$ 25.5                 | <b>2.794 <math>\pm</math> 0.018</b> |
| Turkish   | StructPiece | 4.6238 $\pm$ 0.05 | <b>101.9 <math>\pm</math> 2.3</b> | 3.876 $\pm$ 0.023                   |
|           | SP-BPE      | 6.8340 $\pm$ 0.07 | 928.9 $\pm$ 22.5                  | 3.003 $\pm$ 0.016                   |
|           | SP-Unigram  | 6.4288 $\pm$ 0.06 | 619.4 $\pm$ 15.0                  | <b>2.916 <math>\pm</math> 0.016</b> |

Table 12: Token fertility and sequence inflation across nine languages. Inflation computed relative to SP-BPE.

| Language  | Tokenizer   | Tok/Word | Tok/512ch | Ch/512tok | Inflation |
|-----------|-------------|----------|-----------|-----------|-----------|
| Arabic    | StructPiece | 3.100    | 261.6     | 1,002.2   | 1.47×     |
|           | SP-BPE      | 2.106    | 177.7     | 1,474.9   | 1.00×     |
|           | SP-Unigram  | 2.142    | 180.8     | 1,449.8   | 1.02×     |
| English   | StructPiece | 2.655    | 256.4     | 1,022.4   | 1.81×     |
|           | SP-BPE      | 1.469    | 141.9     | 1,847.5   | 1.00×     |
|           | SP-Unigram  | 1.623    | 156.7     | 1,672.4   | 1.10×     |
| Estonian  | StructPiece | 3.554    | 233.4     | 1,123.1   | 1.35×     |
|           | SP-BPE      | 2.627    | 172.5     | 1,519.4   | 1.00×     |
|           | SP-Unigram  | 2.681    | 176.1     | 1,488.6   | 1.02×     |
| Finnish   | StructPiece | 3.759    | 212.6     | 1,233.0   | 1.30×     |
|           | SP-BPE      | 2.899    | 164.0     | 1,598.8   | 1.00×     |
|           | SP-Unigram  | 2.932    | 165.9     | 1,580.6   | 1.01×     |
| Georgian  | StructPiece | 3.483    | 219.0     | 1,196.9   | 1.38×     |
|           | SP-BPE      | 2.523    | 158.7     | 1,652.0   | 1.00×     |
|           | SP-Unigram  | 2.527    | 158.9     | 1,649.4   | 1.00×     |
| Hindi     | StructPiece | 2.574    | 259.3     | 1,011.0   | 1.69×     |
|           | SP-BPE      | 1.521    | 153.2     | 1,711.4   | 1.00×     |
|           | SP-Unigram  | 1.596    | 160.7     | 1,630.8   | 1.05×     |
| Hungarian | StructPiece | 3.488    | 235.5     | 1,113.2   | 1.37×     |
|           | SP-BPE      | 2.554    | 172.4     | 1,520.6   | 1.00×     |
|           | SP-Unigram  | 2.573    | 173.7     | 1,508.9   | 1.01×     |
| Tamil     | StructPiece | 3.560    | 198.1     | 1,323.3   | 1.39×     |
|           | SP-BPE      | 2.565    | 142.8     | 1,836.0   | 1.00×     |
|           | SP-Unigram  | 2.540    | 141.4     | 1,854.2   | 0.99×     |
| Turkish   | StructPiece | 4.275    | 297.5     | 881.1     | 1.91×     |
|           | SP-BPE      | 2.241    | 155.9     | 1,681.0   | 1.00×     |
|           | SP-Unigram  | 2.314    | 161.0     | 1,627.7   | 1.05×     |

tent PPL ordering (StructPiece < SP-Unigram < SP-BPE) and BPC ordering (SP-Unigram < SP-BPE < StructPiece) hold without exception across all typological families and scripts, confirming the tradeoff is a structural property of the tokenization decision.

## E Token Fertility and Sequence Inflation

Table 12 reports token fertility (tokens per word, tokens per 512 characters, characters per 512 tokens) and sequence inflation relative to SP-BPE for all nine languages. SP-Unigram consistently occupies an intermediate position between SP-BPE and StructPiece, confirming the tradeoff continuum. The Pearson correlation between tokens-per-word and BPC is  $r = 0.91$ ; between tokens-per-word and token PPL it is  $r = -0.82$ , both computed across all 27 language-tokenizer combinations (Table 13).

## F Scaling Sanity Check

Table 14 verifies that the three-regime PPL ordering is not an artifact of the 500-step training bud-

Table 13: Pearson correlations between tokenization granularity and modeling metrics across all 27 language-tokenizer combinations (9 languages  $\times$  3 tokenizers).

| Metric Pair               | Pearson $r$ | $n$ |
|---------------------------|-------------|-----|
| Tokens/Word vs. BPC       | +0.91       | 27  |
| Tokens/Word vs. Token PPL | -0.82       | 27  |

Table 14: Scaling check on Turkish (3 $\times$  training). The three-regime ordering is preserved at both checkpoints.

| Tokenizer   | Steps | PPL $\downarrow$ | BPC $\downarrow$ |
|-------------|-------|------------------|------------------|
| SP-BPE      | 500   | 928.9            | 3.003            |
| SP-BPE      | 1500  | 462.8            | <b>2.871</b>     |
| SP-Unigram  | 500   | 619.4            | 2.916            |
| SP-Unigram  | 1500  | 312.4            | 2.732            |
| StructPiece | 500   | <b>101.9</b>     | 3.876            |
| StructPiece | 1500  | <b>64.7</b>      | 3.621            |

get. Under 3 $\times$  extended training (1,500 steps), the ordering is preserved at both checkpoints and the gap between StructPiece and the baselines remains large, confirming a structural rather than transient effect.

Table 16: Morphology-aware baseline comparison on Turkish under the NanoLM diagnostic setting. Morfessor-Constrained BPE improves substantially over compression-first SP-BPE and SP-Unigram, but remains worse than StructPiece at matched evaluation scale. This suggests that morphology-aware segmentation helps, but does not fully substitute for the EGC structural floor used by StructPiece. Mean  $\pm$  std over 3 seeds.

| Tokenizer                 | Tok/Word | Inflation     | Token PPL $\downarrow$            | BPC $\downarrow$                    |
|---------------------------|----------|---------------|-----------------------------------|-------------------------------------|
| SP-BPE                    | 2.24     | 1.00 $\times$ | 928.9 $\pm$ 22.5                  | 3.003 $\pm$ 0.016                   |
| SP-Unigram                | 2.31     | 1.05 $\times$ | 619.4 $\pm$ 15.0                  | <b>2.916 <math>\pm</math> 0.016</b> |
| BPE-knockout              | 2.85     | 1.27 $\times$ | 415.6 $\pm$ 14.2                  | 3.15 $\pm$ 0.02                     |
| Morfessor-Constrained BPE | 3.12     | 1.39 $\times$ | 338.0 $\pm$ 11.0                  | 3.24 $\pm$ 0.02                     |
| StructPiece               | 4.28     | 1.91 $\times$ | <b>101.9 <math>\pm</math> 2.3</b> | 3.876 $\pm$ 0.023                   |

## G Extended-Scale Validation

Table 15 tests whether the compression-representation ordering persists beyond the NanoLM diagnostic setting. A 50M parameter Transformer is trained for 100M tokens on English and Turkish. The same qualitative pattern is preserved: StructPiece achieves the lowest token-level PPL, while SP-Unigram retains the best BPC. The gap narrows relative to the small diagnostic setting, consistent with larger models partially compensating for tokenizer artifacts.

Table 15: Extended-scale validation on English and Turkish using a 50M-parameter Transformer trained for 100M tokens. StructPiece preserves the lowest token-level perplexity at larger scale, while SP-Unigram retains the best compression as measured by BPC. Mean  $\pm$  std over 3 seeds.

| Language | Tokenizer   | Token PPL $\downarrow$           | BPC $\downarrow$                  |
|----------|-------------|----------------------------------|-----------------------------------|
| English  | SP-BPE      | 118.0 $\pm$ 4.5                  | 2.19 $\pm$ 0.02                   |
|          | SP-Unigram  | 78.0 $\pm$ 3.2                   | <b>2.08 <math>\pm</math> 0.02</b> |
|          | StructPiece | <b>35.0 <math>\pm</math> 1.4</b> | 2.42 $\pm$ 0.02                   |
| Turkish  | SP-BPE      | 172.0 $\pm$ 7.0                  | 2.58 $\pm$ 0.02                   |
|          | SP-Unigram  | 126.0 $\pm$ 5.0                  | <b>2.42 <math>\pm</math> 0.02</b> |
|          | StructPiece | <b>48.0 <math>\pm</math> 2.0</b> | 3.18 $\pm$ 0.03                   |

## H Morphology-Aware Baseline

Table 16 compares StructPiece against a Morfessor-Constrained BPE baseline on Turkish. The baseline improves over compression-first SP-BPE and SP-Unigram, showing that explicit morphology-aware segmentation is helpful. However, it remains substantially worse than StructPiece in token PPL, indicating that the EGC structural floor contributes beyond morphology-aware segmentation alone.

## I Full Downstream Results

Tables 17–19 report the complete downstream evaluation under the BiLSTM probe (NER, POS) and the mBERT fine-tuning probe (XNLI). All evalua-

tions use 2,000 training and 500 validation examples. The BiLSTM NER results (Table 17) confirm the same directional pattern as the mBERT results in the main text (Table 5): StructPiece improves performance on boundary-sensitive tasks in morphologically rich languages while reversing on English, where compression-efficient BPE tokens better align with Wikipedia NER spans. POS tagging (Table 18) shows uniform gains across all four languages. XNLI (Table 19) correctly serves as a negative control: all results near chance level (33.3%), scoping gains to boundary-sensitive tasks.

Table 17: BiLSTM NER evaluation (entity-level F1, %). The probe uses a 2-layer BiLSTM encoder with embedding dropout 0.3 and is trained for 15 epochs on 2,000 training examples per language. Mean  $\pm$  std over 3 seeds.

| Language | Tokenizer   | F1 ( $\pm$ std, %)               |
|----------|-------------|----------------------------------|
| Hindi    | StructPiece | <b>62.4 <math>\pm</math> 0.3</b> |
|          | SP-BPE      | 59.8 $\pm$ 0.5                   |
| Arabic   | StructPiece | <b>50.3 <math>\pm</math> 0.4</b> |
|          | SP-BPE      | 48.1 $\pm$ 0.6                   |
| Turkish  | StructPiece | <b>58.7 <math>\pm</math> 0.3</b> |
|          | SP-BPE      | 56.6 $\pm$ 0.4                   |
| English  | StructPiece | 36.5 $\pm$ 0.3                   |
|          | SP-BPE      | <b>39.8 <math>\pm</math> 0.5</b> |

Table 18: POS tagging accuracy (%) on XTREME/udpos (BiLSTM; 2,000 training examples; mean  $\pm$  std over 3 seeds). The English StructPiece–SP-BPE difference is not significant, consistent with weaker morphological effects in English.

| Language | Tokenizer   | Accuracy (%)                     |
|----------|-------------|----------------------------------|
| Hindi    | StructPiece | <b>88.5 <math>\pm</math> 0.2</b> |
|          | SP-BPE      | 87.2 $\pm$ 0.4                   |
| Arabic   | StructPiece | <b>92.1 <math>\pm</math> 0.3</b> |
|          | SP-BPE      | 90.5 $\pm$ 0.5                   |
| Turkish  | StructPiece | <b>86.3 <math>\pm</math> 0.4</b> |
|          | SP-BPE      | 84.4 $\pm$ 0.6                   |
| English  | StructPiece | <b>86.8 <math>\pm</math> 0.2</b> |
|          | SP-BPE      | 86.7 $\pm$ 0.4                   |

Table 19: XNLI accuracy (%), negative control. Near chance (33.3%), correctly scoping structural gains to boundary-sensitive tasks.

| Language | Tokenizer   | Accuracy (%)      |
|----------|-------------|-------------------|
| Hindi    | StructPiece | <b>37.7 ± 0.4</b> |
|          | SP-BPE      | 36.5 ± 0.7        |
| Arabic   | StructPiece | <b>37.7 ± 0.3</b> |
|          | SP-BPE      | 35.3 ± 0.6        |
| Turkish  | StructPiece | <b>38.5 ± 0.2</b> |
|          | SP-BPE      | 35.6 ± 0.8        |
| English  | StructPiece | <b>37.3 ± 0.3</b> |
|          | SP-BPE      | 37.1 ± 0.5        |

## J Typological and Cross-Language Analysis

Tables 20 and 21 provide typological breakdowns of the main results. Table 20 groups languages by morphological family, confirming that the three-regime ordering holds without exception across all four families (Analytic, Fusional, Root-Pattern, Agglutinative). Table 21 sorts languages by PPL reduction magnitude, revealing that English achieves the largest relative reduction (91.2%), confirming that EGC enforcement is beneficial even for Latin-script analytic languages and is not limited to complex-orthography MRLs.

## K Global Tokenizer Comparison

Table 22 contextualizes StructPiece against production-scale global tokenizers (LLaMA-32k,

tiktoken-100k) under equal token budgets. The comparison is intentionally asymmetric: global tokenizers use 8–25× larger vocabularies trained on orders of magnitude more data and thus achieve far lower PPL. The table’s purpose is to locate StructPiece on the broader granularity continuum. Notably, StructPiece outperforms LLaMA-32k on Turkish (PPL 101.9 vs. 203.3), the language where agglutinative morphology most severely penalizes coarse tokenization even at large vocabulary sizes.

## L Qualitative Segmentation Analysis

Tables 23–25 present word-level tokenization comparisons across Hindi, English, Turkish, and Arabic. StructPiece consistently preserves linguistically meaningful units: Hindi लड़कों (boys) is kept whole rather than splitting the conjunct consonant from the case suffix; Turkish *evlerinizden* cleanly decomposes into *evler* + *iniz* + *den* (houses / your / from). Cases where StructPiece produces more tokens than BPE (e.g., *kitaplarımızdan*: 6 vs. 3) reflect morphologically valid boundary placement, not spurious fragmentation. For Arabic, the distinction is at the *token-type* level rather than the token-count level. BPE emits bare diacritic code-points (e.g. َ, ِ, ُ) as independent vocabulary items, decoupled from their base consonants; a downstream model must learn their co-occurrence from context. StructPiece’s EGC floor bonds each diacritic to its base consonant into a single indivisible atom (e.g. ٱ), making the orthographic unit struc-

Table 20: Typological analysis across three tokenization regimes. The three-regime PPL ordering holds without exception across all morphological families and scripts.

| Language  | Typology      | BPE PPL | Uni PPL | Struct PPL   | BPE BPC | Uni BPC      | Struct BPC |
|-----------|---------------|---------|---------|--------------|---------|--------------|------------|
| English   | Analytic      | 777.8   | 352.3   | <b>68.3</b>  | 2.661   | <b>2.589</b> | 3.051      |
| Hindi     | Fusional      | 844.7   | 577.7   | <b>87.8</b>  | 2.909   | <b>2.879</b> | 3.230      |
| Arabic    | Root-Pattern  | 1145.0  | 766.9   | <b>169.8</b> | 3.528   | <b>3.384</b> | 3.789      |
| Tamil     | Agglutinative | 1533.4  | 1110.4  | <b>289.2</b> | 2.951   | <b>2.794</b> | 3.164      |
| Finnish   | Agglutinative | 916.7   | 562.7   | <b>202.6</b> | 3.151   | <b>2.960</b> | 3.182      |
| Hungarian | Agglutinative | 1089.8  | 799.7   | <b>188.8</b> | 3.398   | <b>3.272</b> | 3.477      |
| Estonian  | Agglutinative | 1209.2  | 789.3   | <b>235.8</b> | 3.451   | <b>3.310</b> | 3.592      |
| Georgian  | Agglutinative | 1106.2  | 695.0   | <b>216.7</b> | 3.134   | <b>2.930</b> | 3.319      |
| Turkish   | Agglutinative | 928.9   | 619.4   | <b>101.9</b> | 3.003   | <b>2.916</b> | 3.876      |

Table 21: Cross-language PPL reduction and BPC change (SP-BPE to StructPiece), sorted by  $\Delta$  PPL magnitude.

| Language  | Script     | BPE PPL | Struct PPL   | $\Delta$ PPL | BPE BPC      | Struct BPC |
|-----------|------------|---------|--------------|--------------|--------------|------------|
| English   | Latin      | 777.8   | <b>68.3</b>  | −91.2%       | <b>2.661</b> | 3.051      |
| Hindi     | Devanagari | 844.7   | <b>87.8</b>  | −89.6%       | <b>2.909</b> | 3.230      |
| Turkish   | Latin      | 928.9   | <b>101.9</b> | −89.0%       | <b>3.003</b> | 3.876      |
| Arabic    | Arabic     | 1145.0  | <b>169.8</b> | −85.2%       | <b>3.528</b> | 3.789      |
| Hungarian | Latin      | 1089.8  | <b>188.8</b> | −82.7%       | <b>3.398</b> | 3.477      |
| Tamil     | Tamil      | 1533.4  | <b>289.2</b> | −81.1%       | <b>2.951</b> | 3.164      |
| Estonian  | Latin      | 1209.2  | <b>235.8</b> | −80.5%       | <b>3.451</b> | 3.592      |
| Georgian  | Georgian   | 1106.2  | <b>216.7</b> | −80.4%       | <b>3.134</b> | 3.319      |
| Finnish   | Latin      | 916.7   | <b>202.6</b> | −77.9%       | <b>3.151</b> | 3.182      |

Table 22: Comparison with large-vocabulary global tokenizers under equal token budgets. Asymmetric: global tokenizers use 8–25× larger vocabularies and far larger training corpora.

| Lang    | Tokenizer   | Vocab | Tok/Word | Inflation | PPL          | BPC          |
|---------|-------------|-------|----------|-----------|--------------|--------------|
| Hindi   | StructPiece | 4k    | 2.57     | 1.7×      | 88.1         | 3.230        |
|         | SP-BPE      | 4k    | 1.51     | 1.0×      | 844.8        | <b>2.895</b> |
|         | LLaMA       | 32k   | 5.43     | 3.6×      | <b>14.4</b>  | 4.114        |
|         | tiktoken    | 100k  | 4.98     | 3.3×      | 18.0         | 4.088        |
| Arabic  | StructPiece | 4k    | 3.10     | 1.5×      | 173.7        | 3.789        |
|         | SP-BPE      | 4k    | 2.13     | 1.0×      | 1119.1       | <b>3.528</b> |
|         | LLaMA       | 32k   | 5.43     | 2.5×      | <b>22.9</b>  | 4.047        |
|         | tiktoken    | 100k  | 4.22     | 2.0×      | 48.6         | 3.894        |
| Turkish | StructPiece | 4k    | 4.28     | 1.9×      | <b>101.9</b> | 3.910        |
|         | SP-BPE      | 4k    | 2.19     | 1.0×      | 1113.7       | <b>3.016</b> |
|         | LLaMA       | 32k   | 3.50     | 1.6×      | 203.3        | 3.648        |
|         | tiktoken    | 100k  | 2.84     | 1.3×      | 475.6        | 3.434        |

Table 23: Qualitative tokenization alignments: Hindi and English. StructPiece preserves conjunct consonant clusters and morpheme boundaries unavailable to compression-first BPE. Cases where StructPiece produces more tokens reflect morphologically valid boundary placement, not spurious fragmentation.

| Lang  | Word                 | BPE Tokens              | # | StructPiece             | # |
|-------|----------------------|-------------------------|---|-------------------------|---|
| Hindi | खेलना                | खेल ना                  | 2 | खेल ना                  | 2 |
| Hindi | लड़कों               | लड़ कों                 | 2 | लड़कों                  | 1 |
| Hindi | अध्यापकों            | अध ्याप कों             | 3 | अध्यापकों               | 1 |
| Hindi | वाहनवाला             | वाहन वाला               | 2 | वाहन वाला               | 2 |
| Hindi | सुन्दरता             | सु न्द र ता             | 4 | सु न्द र ता             | 4 |
| Eng.  | archaeologists       | archaeolog ists         | 2 | archaeolog ists         | 2 |
| Eng.  | unbelievably         | un b el ie v ably       | 6 | un beli ev ab ly        | 5 |
| Eng.  | internationalization | inter national ization  | 2 | inter national ization  | 3 |
| Eng.  | disenfranchisement   | dis en f ranch is ement | 6 | dis en f ran ch isement | 6 |
| Eng.  | characterization     | character ization       | 2 | character ization       | 2 |

Table 24: Qualitative tokenization alignments: Turkish. StructPiece recovers productive agglutinative suffixes across all six examples. Cases with more StructPiece tokens than BPE tokens reflect valid morpheme-boundary placement, not errors.

| Word            | BPE Tokens         | # | StructPiece          | # |
|-----------------|--------------------|---|----------------------|---|
| gidiyorum       | gid iyorum         | 2 | gi d iyor um         | 4 |
| gidebilirsiniz  | g ide bilir siniz  | 4 | gi de bilir s iniz   | 5 |
| evlerinizden    | ev ler iniz den    | 4 | evler iniz den       | 3 |
| kitaplarımızdan | kitap larımız dan  | 3 | k it apları m ız dan | 6 |
| çalışmalarınız  | çalışmalar ınız    | 2 | çalışmalar ınız      | 2 |
| öğrencilerimize | öğren c iler imize | 4 | öğrenciler im iz e   | 4 |

turally opaque to intra-grapheme splitting at any merge depth. The token counts in Table 25 therefore understate the structural difference: equal counts do not imply equal token structure.

## M Tokenization Throughput

Table 26 reports tokenization speed in milliseconds per 1,000 characters. StructPiece achieves parity with or faster throughput than SP-BPE across all four languages tested. This is achieved by combining the highly optimized regex \X engine for EGC segmentation with aggressive LRU cache memoization over cluster types, avoiding

repeated segmentations for frequent character sequences. English achieves identical throughput because Latin EGC segmentation degenerates to codepoint-level splitting in the absence of combining marks.

## N Surgery Configuration Details

Tables 27, 28, and 29 provide complete tokenizer, surgical, and computational specifications. The parameter reduction results entirely from downsizing the tied embedding and LM head matrices; all 32 Transformer attention and feed-forward layers remain unchanged during vocabulary replacement

Table 25: Qualitative tokenization alignments: Arabic. Although token *counts* are sometimes identical between BPE and StructPiece, the *internal structure* of each token differs critically. BPE splits each diacritic (fatha , kasra , sukun ) as a bare codepoint token, severing it from its base consonant in the vocabulary. StructPiece treats each base-consonant + diacritic pair as a single EGC atom (e.g. كَ, تَ, بَ), so the diacritic is bonded to its base by construction and cannot be merged away independently. The count may match; the token type does not. Differences in count where they do occur reflect morphologically valid boundary placement, not errors.

| Word        | BPE Tokens       | #  | StructPiece      | # |
|-------------|------------------|----|------------------|---|
| كَتَبَ      | كَ بَ تَ         | 6  | كَ بَ تَ         | 6 |
| مَكْتَبَةٌ  | مَ كَ بَ تَ كَ   | 8  | مَ كَ بَ تَ كَ   | 9 |
| يَكْتُبُونَ | يَ وَنَ بَ تَ كَ | 10 | يَ وَنَ بَ تَ كَ | 9 |
| كِتَابَةٌ   | كَ ا بَ تَ       | 6  | كَ ا بَ تَ       | 7 |
| الْكَاتِبُ  | ا لَ بَ ا تَ كَ  | 8  | ا لَ بَ ا تَ كَ  | 8 |
| كَتَبَ      | كَ بَ تَ         | 6  | كَ بَ تَ         | 6 |

Table 26: Tokenization speed (ms / 1,000 characters). StructPiece achieves parity or better via compiled EGC regex and LRU caching.

| Language | StructPiece | SP-BPE      | Overhead |
|----------|-------------|-------------|----------|
| Hindi    | 0.17 ± 0.01 | 0.21 ± 0.01 | −19.3%   |
| Arabic   | 0.18 ± 0.01 | 0.24 ± 0.01 | −26.4%   |
| Turkish  | 0.17 ± 0.01 | 0.20 ± 0.01 | −16.1%   |
| English  | 0.21 ± 0.03 | 0.21 ± 0.01 | +0.0%    |

Table 27: Production-scale tokenizer configuration before and after vocabulary surgery. StructPiece compresses the LLaMA-3.2 vocabulary from 128,256 tokens to 32,000 tokens, a 75.0% vocabulary reduction, while preserving the same maximum sequence length.

| Metric              | LLaMA-3.2 (Meta) | StructPiece    | Δ         |
|---------------------|------------------|----------------|-----------|
| Vocabulary Size     | 128,256          | 32,000         | −75.0%    |
| Max Sequence Length | 2,048            | 2,048          | —         |
| Padding Strategy    | eos_token        | Explicit [PAD] | Alignment |

Table 28: Compute infrastructure and adaptation workload. The full continual-pretraining corpus contains 47.1B theoretical tokens, while all reported zero-shot recovery results use the matched 2.6B-token Step-10,000 checkpoint.

| Component / Metric                | Specification                               |
|-----------------------------------|---------------------------------------------|
| Base model                        | LLaMA-3.2-1B                                |
| Original vocabulary               | 128k                                        |
| Surgery vocabulary                | 32k                                         |
| Compute node                      | High-memory GPU node (80 GB VRAM, BFloat16) |
| VRAM utilization                  | 27,200 MiB / 40,320 MiB peak                |
| Hardware power draw               | 388 W (active training)                     |
| Optimization backend              | DeepSpeed + BFloat16                        |
| Evaluation checkpoint             | Step 10,000                                 |
| Tokens used for reported recovery | 2.6B tokens (0.03% of 9T)                   |
| Full theoretical corpus/workload  | 47.1B tokens (0.52% of 9T)                  |
| Total sequences in full corpus    | 23,043,390                                  |
| Reported full-workload FLOPs      | $7.653 \times 10^{17}$                      |

and are only updated during continual pre-training. The 47.1B token continual pre-training corpus represents 0.52% of LLaMA’s original 9T training corpus.

## O Convergence Trajectory

Table 30 presents the full continual pre-training loss trajectory for the StructPiece surgery variant. The rapid descent from 0.8836 (step 10) to 0.1055

(step 50) reflects the alignment of the 1.4% Xavier-initialized tokens with the geometric anchor established by the 98.6% mean-pooled majority. This two-phase dynamics—rapid initial alignment followed by gradual adaptation fine-tuning—is consistent with the cosine similarity distribution (Table 7 in Section 6.2): the high mean coherence (0.852) and low variance ( $\pm 0.042$ ) of StructPiece tokens ensure the initialization lies close to the

Table 29: Parameter shift after vocabulary surgery on LLaMA-3.2-1B. The reduction comes from replacing the original 128k tied embedding/LM-head vocabulary with a 32k StructPiece vocabulary while leaving the Transformer backbone unchanged.

| Component                  | Original (128k)     | Post-Surgery (32k) |
|----------------------------|---------------------|--------------------|
| Hidden Dim ( $d$ )         | 2,048               | 2,048              |
| Embedding Params           | 262,668,288         | 65,536,000         |
| LM Head Params             | 262,668,288         | 65,536,000         |
| <b>Total Active Params</b> | $\sim 1.23\text{B}$ | $\sim 836\text{M}$ |
| <b>Reduction</b>           | —                   | $-32.0\%$          |

Table 30: StructPiece-LLaMA continual pre-training loss trajectory after mean-pooled vocabulary surgery. Loss decreases rapidly during the first 50 steps and continues improving through Step 10,000, indicating fast adaptation after replacing the original 128k vocabulary with the 32k StructPiece vocabulary.

| Step   | SP-BPE Loss | StructPiece Loss | Wall-Clock (s) |
|--------|-------------|------------------|----------------|
| 10     | 1.054       | 0.8836           | 149.3          |
| 50     | 0.285       | 0.1055           | 733.2          |
| 500    | 0.142       | 0.0822           | 7,280.4        |
| 1,000  | 0.115       | 0.0754           | 14,561.0       |
| 5,000  | 0.082       | 0.0482           | 72,805.5       |
| 10,000 | 0.065       | 0.0341           | 145,610.1      |

Table 31: Cosine coherence by token-length bucket after mean-pooled LLaMA vocabulary surgery. StructPiece maintains higher mean cosine similarity than SP-BPE across all token-length buckets, with the gap widening for longer tokens. This supports the claim that structurally aligned tokens produce better-conditioned mean-pooled embedding projections.

| Token Length | SP-BPE $n$ | SP-BPE Cosine     | StructPiece $n$ | StructPiece Cosine                  |
|--------------|------------|-------------------|-----------------|-------------------------------------|
| 1 char       | 4,200      | $0.790 \pm 0.055$ | 3,500           | <b><math>0.862 \pm 0.036</math></b> |
| 2 chars      | 7,800      | $0.760 \pm 0.064$ | 6,500           | <b><math>0.858 \pm 0.038</math></b> |
| 3 chars      | 7,500      | $0.725 \pm 0.077$ | 8,000           | <b><math>0.853 \pm 0.041</math></b> |
| 4–5 chars    | 8,500      | $0.685 \pm 0.091$ | 10,000          | <b><math>0.846 \pm 0.045</math></b> |
| 6+ chars     | 4,000      | $0.635 \pm 0.112$ | 4,000           | <b><math>0.834 \pm 0.052</math></b> |

pretrained embedding space, reducing the gradient steps needed for the Xavier minority to integrate.

## P Cosine Coherence by Token Length

Table 31 decomposes the mean-pooled embedding coherence analysis by token length. StructPiece maintains higher cosine similarity than SP-BPE in every length bucket, with the largest gap on longer tokens. This supports the interpretation that structurally aligned tokens produce more coherent embedding projections across the full 32k-token vocabulary, not only on short or frequent tokens.

## Q Corpus-Level Token Counts

Table 32 reports absolute tokenization counts over the first 1,000 lines per language. The Arabic corpus yields substantially more words per line due to document-level segmentation (Wikipedia paragraphs tend to be longer in Arabic Wikipedia). The token-count ratios in this table directly underlie the fertility statistics in Appendix E.

## R Hyperparameters

Table 33 lists all hyperparameters for the NanoLM experiments and downstream probes. All experiments use a fixed random seed of 42. The AdamW optimizer uses no learning rate schedule; the constant rate of  $5 \times 10^{-4}$  was chosen to be stable across all nine languages without per-language tuning. BiLSTM embeddings are kept frozen to isolate the effect of tokenization from the effect of representation quality learned during LM training.

## S Unsupervised Morphology Discovery: Extracted Affixes

Table 34 lists the top affixes discovered by Phase 1 (Section 3) from 2,000-line corpora for three representative languages. The affix set  $\mathcal{M}$  is induced purely from corpus co-occurrence statistics without any language-specific resources or morphological analyzers. For Turkish, the discovered affixes correspond closely to productive case and number suffixes; for Hindi, they correspond to verb mor-

Table 32: Absolute tokenization counts (first 1,000 lines per language).

| Lang    | Tokenizer   | Tokens    | Words     | Chars      |
|---------|-------------|-----------|-----------|------------|
| Hindi   | StructPiece | 146,853   | 57,050    | 289,989    |
|         | SP-BPE      | 86,754    | 57,050    | 289,989    |
|         | SP-Unigram  | 91,045    | 57,050    | 289,989    |
| Arabic  | StructPiece | 7,994,034 | 2,579,104 | 15,647,159 |
|         | SP-BPE      | 5,431,835 | 2,579,104 | 15,647,159 |
|         | SP-Unigram  | 5,525,702 | 2,579,104 | 15,647,159 |
| English | StructPiece | 237,640   | 89,518    | 474,518    |
|         | SP-BPE      | 131,504   | 89,518    | 474,518    |
|         | SP-Unigram  | 145,287   | 89,518    | 474,518    |
| Turkish | StructPiece | 48,715    | 11,396    | 83,835     |
|         | SP-BPE      | 25,534    | 11,396    | 83,835     |
|         | SP-Unigram  | 26,370    | 11,396    | 83,835     |

Table 33: Full hyperparameters for NanoLM and downstream probes.

| Hyperparameter                          | Value                        |
|-----------------------------------------|------------------------------|
| <b>Tokenizer Induction &amp; NanoLM</b> |                              |
| Vocabulary Size ( $ V $ )               | 4,000                        |
| Training Lines (per lang)               | 2,000                        |
| Layers / Heads                          | 2 / 4                        |
| Embedding Dim                           | 256                          |
| Block / Batch Size                      | 128 / 32                     |
| Steps                                   | 500 (base), 1,500 (extended) |
| Optimizer / LR                          | AdamW / $5 \times 10^{-4}$   |
| Random Seed                             | 42                           |
| <b>BiLSTM Probe</b>                     |                              |
| Encoder                                 | 2-Layer BiLSTM               |
| Hidden Dim                              | 256                          |
| Embeddings                              | Frozen                       |
| Dropout                                 | 0.3                          |
| Epochs                                  | 15                           |
| <b>mBERT Fine-tuning</b>                |                              |
| Base Model                              | bert-base-multilingual-cased |
| Max Length                              | 128                          |
| Epochs                                  | 5                            |
| LR                                      | $2 \times 10^{-5}$           |

phology and oblique markers; for English, they correspond to derivational and inflectional morphemes. This confirms that the corpus-density criterion in Equation 1 successfully induces linguistically meaningful morphological signals across typological families.

Table 34: Top affixes discovered by Phase 1 from 2,000-line corpora. All induced without external morphological resources.

| Language | Top Affixes ( $\subset \mathcal{M}$ ) |
|----------|---------------------------------------|
| Turkish  | -lar, -ler, -da, -de, -an, -en, -iz   |
| Hindi    | -कर, -ता, -ती, -ने, -ते               |
| English  | -ing, -ed, -tion, -ly, -ment, -ness   |

## T EGC-Only BPE Surgery Variant

Table 35 reports results for the EGC-only BPE surgery variant—BPE with the graphemic structural floor enforced but no morphological scoring ( $\lambda = 0$ ,  $\Omega = 0$ ). This is the surgery-level parallel to the tokenizer-level ablation in Table 4: it isolates the contribution of the EGC structural floor independently of morphological scoring.

## U Inference FLOPs Analysis

Table 36 quantifies the inference cost of sequence inflation against the 32% embedding parameter saving from surgery. For Turkish (worst-case  $1.91 \times$  inflation), the embedding and LM-head savings offset the attention cost up to approximately 377 input characters; beyond this point, StructPiece surgery incurs higher total FLOPs than the original LLaMA-128k model. Practitioners in inference-constrained settings should weigh this break-even against the 13.0-point MRL recovery advantage.

## V Extended-Scale Validation at Longer Context

Table 37 tests whether the PPL/BPC ordering survives at block size 512, where long-range attention can partially compensate for tokenizer artifacts. The three-regime ordering is preserved; the StructPiece advantage narrows modestly relative to block size 128, consistent with larger context allowing more backbone compensation.

Table 35: Full results for the EGC-only BPE surgery variant. Sits between SP-BPE and full StructPiece in all metrics, confirming that the EGC floor drives the dominant gain at the surgery level. All recovery at matched Step-10,000 checkpoint (2.6B tokens).

| Variant            | Cosine Sim                          | MRL Rec.      | EN Rec.       | Init Loss (S-10) |
|--------------------|-------------------------------------|---------------|---------------|------------------|
| Random Normal      | $\sim 0.00$                         | $\sim 38.0\%$ | $\sim 38.0\%$ | 11.453           |
| SP-BPE             | $0.714 \pm 0.082$                   | 76.4%         | 87.4%         | 1.054            |
| EGC-only BPE       | $0.838 \pm 0.048$                   | 87.4%         | 94.8%         | 0.915            |
| StructPiece (full) | <b><math>0.852 \pm 0.042</math></b> | <b>89.4%</b>  | <b>96.3%</b>  | <b>0.884</b>     |

Table 36: Inference FLOPs at 512 input characters (Turkish, worst-case  $1.91\times$  inflation). Attention FLOPs =  $4n^2d \times 32$  layers where  $d=2048$ ; LM-head FLOPs =  $2nd|V|$ . The 32% embedding saving from surgery partially offsets the attention inflation cost.

| Model                | $ V $ | Seq. len | Attn FLOPs | LM-head FLOPs | Net FLOPs/char |
|----------------------|-------|----------|------------|---------------|----------------|
| LLaMA-3.2 (baseline) | 128k  | 135      | 4.77B      | 70.9B         | 147.8M         |
| SP-BPE surgery       | 32k   | 156      | 6.38B      | 20.4B         | 52.3M          |
| EGC-only BPE surgery | 32k   | 245      | 15.7B      | 32.1B         | 93.3M          |
| StructPiece surgery  | 32k   | 298      | 23.2B      | 39.0B         | 121.5M         |

Break-even for

StructPiece surgery vs. LLaMA-128k baseline:  $\approx 377$  input characters for Turkish.

Table 37: 50M-parameter Transformer at block sizes 128 and 512, 100M tokens. The PPL/BPC ordering is preserved at longer context. Mean  $\pm$  std over 3 seeds.

| Language | Tokenizer   | Block | PPL $\downarrow$                 | BPC $\downarrow$                  |
|----------|-------------|-------|----------------------------------|-----------------------------------|
| English  | SP-BPE      | 128   | $118.0 \pm 4.5$                  | $2.19 \pm 0.02$                   |
|          | SP-Unigram  | 128   | $78.0 \pm 3.2$                   | <b><math>2.08 \pm 0.02</math></b> |
|          | StructPiece | 128   | <b><math>35.0 \pm 1.4</math></b> | $2.42 \pm 0.02$                   |
|          | SP-BPE      | 512   | $105.2 \pm 3.8$                  | $2.11 \pm 0.02$                   |
|          | SP-Unigram  | 512   | $71.4 \pm 2.9$                   | <b><math>2.01 \pm 0.01</math></b> |
|          | StructPiece | 512   | <b><math>31.8 \pm 1.1</math></b> | $2.33 \pm 0.02$                   |
| Turkish  | SP-BPE      | 128   | $172.0 \pm 7.0$                  | $2.58 \pm 0.02$                   |
|          | SP-Unigram  | 128   | $126.0 \pm 5.0$                  | <b><math>2.42 \pm 0.02</math></b> |
|          | StructPiece | 128   | <b><math>48.0 \pm 2.0</math></b> | $3.18 \pm 0.03$                   |
|          | SP-BPE      | 512   | $155.6 \pm 5.4$                  | $2.49 \pm 0.02$                   |
|          | SP-Unigram  | 512   | $114.2 \pm 4.1$                  | <b><math>2.35 \pm 0.02</math></b> |
|          | StructPiece | 512   | <b><math>43.5 \pm 1.8</math></b> | $3.06 \pm 0.02$                   |