The Tokenization Tax: Inverse Scaling of the Multilingual Performance Gap

Anonymous ACL submission

Abstract

Large language models perform substantially worse on low-resource languages than on English. This multilingual gap is widely attributed to tokenizer fragmentation, yet existing evidence is purely correlational. We provide the first controlled measurements that *bound* fragmentation's contribution. By truncating the BPE merge list of a model's native tokenizer, we deliberately fragment English to match a target language's natural fertility, holding language, content, and model knowledge constant. Across the Qwen3 family on Belebele, the resulting accuracy cost decreases sharply with model size: 10.7 points at 0.6B, 7.0 at 1.7B, 1.4 at 4B, and within measurement noise at 8B (where Belebele's 91.6% baseline limits resolution); Llama-3.1-8B shows a similarly small effect at 8B. Applied to five natural languages at the 1.7B scale, the framework yields per-language upper bounds: tokenization explains at most 16% (Swahili) to 42% (Turkish) of each language's gap with English. We additionally document an unexpected failure mode of LoRA-based adaptation on fragmented data, where per-token perplexity drops 70 to 270 times while multiple-choice accuracy decreases 5 to 8 points. A targeted ablation rules out direct weight updates at the final layer as the mechanism, narrowing the space of candidate explanations while leaving the ultimate mechanism open.

1 Introduction

Large language models exhibit substantial performance disparities across languages. On Belebele reading comprehension (Bandarkar et al., 2024), contemporary models score 30 to 50 percentage points lower on languages such as Yoruba than on English, and similar disparities exist for hundreds of low-resource languages (Ahia et al., 2023; Petrov et al., 2023). The causes of this gap are entangled. At least three plausible contributors operate simultaneously in any natural-language experiment: tokenization fragmentation, training-data scarcity, and parameter biases from English-dominant pretraining. The latter two are difficult to separate cleanly; both ultimately trace to the pretraining data distribution. Tokenization, however, can be intervened on independently of the rest, which is what we exploit.

Tokenization is the most-cited candidate cause. BPE tokenizers, even those trained with multilingual coverage, fragment lower-resource languages into many more subword pieces than English (Petrov et al., 2023): a single Yoruba word may become four or five tokens, inflating sequence length and forcing the model to spend attention capacity reconstructing word identity before any linguistic processing can occur (Yang et al., 2026; Kaplan et al., 2025). The correlation between high tokenizer fertility and poor downstream performance is well-established across many languages and models (Ahia et al., 2023; Lundin et al., 2026). Whether high fertility *causes* poor performance, and by how much, has not been measured.

We address this with a controlled experiment. By truncating the BPE merge list of a model's native tokenizer, we deliberately fragment English text to match the natural fertility of a target language while keeping the language, model, content, and world knowledge fixed. The resulting accuracy drop measures the deployment-time cost of fragmented input under naive (unadapted) inference; we discuss the interpretation of this measurement in Section 3.3.

Our headline finding is that this fragmentation cost exhibits sharp inverse scaling with model size. On Belebele English at Swahili-matched fertility (2.5 tokens per word), Qwen3-0.6B loses 10.7 points of accuracy, Qwen3-1.7B loses 7.0 points, Qwen3-4B loses 1.4 points, and Qwen3-8B loses no measurable accuracy. Llama-3.1-8B (vocabulary 128,256) shows a similarly small effect at 8B, indicating the near-zero result is not specific to

the Qwen3 tokenizer. The 8B Belebele baseline is 91.6%, and ceiling effects partly motivate our caution in interpreting the magnitude of the scaling result at the largest sizes; we discuss this explicitly in the limitations.

Combined with parallel evaluations on five natural languages, this scaling result yields quantitative bounds on each language's tokenization-attributable share. At Qwen3-1.7B, between 16% (Swahili) and 42% (Turkish) of each tested language's gap with English can be attributed to tokenization, with the remainder reflecting data scarcity and other factors. For low-resource languages like Swahili and Yoruba, tokenization is a minority contributor to large gaps (around 9 of 48 points for Yoruba), implying that the bulk of those gaps requires interventions beyond tokenization. For typologically distant but well-resourced languages like Turkish, tokenization accounts for nearly half of a much smaller gap, suggesting that tokenizer improvements alone could close a substantial fraction of Turkish's gap without additional training data.

We additionally investigate whether the measured fragmentation cost is recoverable through adaptation. We fine-tune Qwen3-1.7B on fragmented English Wikipedia using LoRA (Hu et al., 2022) and re-evaluate on Belebele. Contrary to expectation, the adapted model achieves dramatically lower per-token perplexity on Belebele (a 73x to 267x reduction depending on the fragmentation level) yet *worse* multiple-choice accuracy than the unadapted model (a 5 to 8 point decrease). We measure layer-wise attention patterns before and after adaptation, and find a sharp localized shift at the final layer (layer 27) while layers 0–26 are essentially unchanged. Multiple-choice scoring depends on the final layer's output, so this was a natural candidate mechanism for the accuracy decline. We tested whether direct weight updates at layer 27 are causally responsible by excluding that layer from LoRA adaptation. The accuracy decline persisted, ruling out the narrow hypothesis of direct weight-update causation at layer 27; whether the attention shift itself is causal (driven by upstream representational changes) remains untested. The mechanism remains open.

Our contributions are:

- A controlled-fragmentation methodology using BPE merge truncation to isolate the cost of tokenization fragmentation in the multilingual performance gap, holding language, model, and content constant.
- Empirical evidence that this fragmentation cost decreases sharply with model size on Belebele, from 10.7 points at Qwen3-0.6B to within measurement noise at Qwen3-8B; a separate Llama-3.1-8B measurement shows the same small-effect regime at 8B, indicating the result is not Qwen3-tokenizer-specific.
- A per-language decomposition for five Latin-script languages. At Qwen3-1.7B, tokenization accounts for 16% (Swahili) to 42% (Turkish) of each tested language's gap with English, with the remainder attributable to other factors.
- Documentation of an unexpected failure mode in LoRA-based adaptation on fragmented data: per-token perplexity on Belebele drops by 70 to 270 times while multiple-choice accuracy decreases by 5 to 8 points. A targeted ablation rules out direct weight updates at the final layer as the mechanism, narrowing the space of candidate explanations while leaving the ultimate mechanism open.
- Open implementation, including the merge-truncation pipeline, all configurations, all inference and adaptation results, and the analysis scripts.[1]

2 Related Work

Tokenization and the multilingual gap. Ahia et al. (2023) and Petrov et al. (2023) documented large fertility disparities across the languages used by modern LLM tokenizers, showing that low-resource languages receive systematically less efficient representations and that fertility correlates strongly with downstream performance. Lundin et al. (2026) extended this analysis to African languages with similar correlational findings. While they document the correlation between fertility and African-language performance, our work provides the controlled counterfactual that bounds what fraction of that correlation reflects tokenization itself rather than co-varying factors such as data scarcity. Mielke et al. (2021) surveyed the broader space of tokenization choices in NLP and their downstream consequences. These works establish the correlation we investigate causally.

[1] Code and results available at https://anonymous.4open.science/r/The-Tokenization-Tax-4B9F.

Mechanisms of subword reconstruction. Yang et al. (2026) and Kaplan et al. (2025) characterized word-recovery mechanisms in LLMs, showing that highly fragmented input (character-level in Yang et al.; multi-token splits of in-vocabulary words in Kaplan et al.) induces critical in-group attention concentration in early-to-middle transformer layers, with hidden states reconstructing canonical word-level token identities. Their studies operate on English inputs to instruction-tuned multilingual base models (Gemma-2, Qwen2.5, Llama-3.2). Our methodology adopts merge-list-based fragmentation as the controlled probe and extends the analysis to a multilingual decomposition framework, connecting the reconstruction mechanism to downstream accuracy across languages and scales.

Causal probes of tokenization. Lesci et al. (2025) estimate the causal effect of tokenization on model behavior using a regression discontinuity design at the BPE vocabulary cutoff K, comparing subwords ranked just inside and just outside the vocabulary. Their unit of analysis is the individual subword and their outcome is character-string probability under the trained model. Our approach is complementary: rather than estimating per-subword effects at a single threshold, we use merge truncation to produce a continuous fertility curve and match it to natural-language target fertilities, enabling a corpus-level decomposition of the multilingual performance gap. The two approaches answer related but distinct questions—whether a particular subword's inclusion matters (Lesci et al.) versus how much of a language's accuracy gap is attributable to its fertility (ours).

Merge truncation as a probe. Sawada and Goyal (2025) investigated BPE merge-list truncation as one of several inference-time perturbations to study English LLM robustness to tokenization changes. We adopt their truncation procedure but re-purpose it: rather than evaluating English robustness in isolation, we use fertility-matched truncation to construct a controlled counterfactual for each target natural language.

Vocabulary scaling for pretraining. Tao et al. (2024) derived scaling laws relating vocabulary size, model size, and training compute for from-scratch pretraining. Their question (how to size a vocabulary during pretraining for optimal compute efficiency) differs from ours (how much does the existing vocabulary's tokenization cost at inference time on a language gap), but their work validates that vocabulary choices have first-order effects on model performance.

Parameter-efficient adaptation. LoRA (Hu et al., 2022) has become the standard parameter-efficient adaptation method for LLMs. We use LoRA for our adaptation experiments and identify an unexpected failure mode specific to its application on fragmented data.

Multilingual benchmarks. Belebele (Bandarkar et al., 2024) is a parallel multiple-choice reading comprehension benchmark in 122 languages. We use it as our primary evaluation because parallel content removes content as a cross-language confound.

3 Methodology

3.1 BPE merge truncation

We focus on byte-level BPE, the tokenization variant used by both model families in our experiments (Qwen3 and Llama-3). A byte-level BPE tokenizer is constructed by iteratively merging frequent adjacent token pairs starting from 256 base byte tokens, producing an ordered list of merges $(m_1, m_2, \ldots, m_N)$ where m_1 is the most frequent pair and m_N is the least. Every intermediate token created by some merge m_i remains a valid vocabulary entry with a trained embedding, because the training process saw these intermediates before producing further merges from them.

We construct a truncated tokenizer by keeping only the first k merges. To encode text under the truncated tokenizer, we first encode it under the full tokenizer, then iteratively replace each token produced by some merge m_j with $j > k$ by its constituent pair. The cascade continues until every token in the sequence originates from a merge m_i with $i \leq k$ (or is a base byte).

Key property. All tokens in the resulting sequence remain valid IDs in the original model's vocabulary, with embeddings the model has seen during training. We are not introducing unknown tokens, modifying any model weight, or changing the model's effective vocabulary—we are choosing to encode the same content with finer-grained subword pieces.

As a worked example, consider the English sentence "The children are playing outside." Under Qwen3's full tokenizer at fertility 1.26, this becomes ["The", " children", " are", " playing", "

Cutoff	Language	Target f_L	English f
10830	Spanish	1.61	1.60
7876	French	1.70	1.70
1011	Swahili	2.50	2.50
937	Turkish	2.53	2.53
402	Yoruba	2.92	2.92

Table 1: Fertility-matched cutoffs for Qwen3 tokenizer. Each cutoff produces English fertility under truncation that matches the target language's fertility under the full tokenizer. Cutoffs are not linear in target fertility because successive merges have diminishing marginal effects on fragmentation.

outside"]: 5 tokens for 5 words. Under the truncated tokenizer at cutoff $k = 1011$ (which produces a corpus-average fertility of 2.50 on Belebele English, matching Swahili), the same sentence becomes ["The", " ch", "i", "ld", "ren", " are", " pl", "a", "y", "ing", " o", "u", "ts", "ide"], or 14 tokens for 5 words (fertility 2.80 on this specific sentence; the corpus average is 2.50). Each resulting token is a valid Qwen3 vocabulary entry. Decoding either sequence recovers the exact original text.

3.2 Fertility-matched calibration

Tokenizer *fertility* on a corpus C in language L is defined as the mean number of tokens per whitespace-separated word produced by the tokenizer: $f(T, C) = |T(C)|/|W(C)|$, where $T(C)$ is the tokenized output and $W(C)$ is the word count.

For each target natural language $L \in \{$French, Spanish, Turkish, Swahili, Yoruba$\}$, we measure the native tokenizer's fertility on Belebele's L passages. We then search for a merge cutoff k_L such that the truncated tokenizer applied to English produces fertility within ± 0.01 of f_L. Search is conducted by binary search over cutoff values, with measurement on the same Belebele English passages used at evaluation time.

Table 1 lists the resulting cutoffs and verifies that English under truncation reproduces each target language's fertility to high precision.

3.3 Decomposition framework

Let $\mathrm{acc}_M(L)$ denote model M's accuracy on a benchmark in language L using the unmodified tokenizer, and let $\mathrm{acc}_M^{\mathrm{eng}}(f)$ denote M's accuracy on the English benchmark using the truncated tokenizer at fertility f. We define:

$$\mathrm{Total\,Gap}(L) = \mathrm{acc}_M(\mathrm{Eng}) - \mathrm{acc}_M(L) \quad (1)$$

$$\Delta_A(L) = \mathrm{acc}_M(\mathrm{Eng}) - \mathrm{acc}_M^{\mathrm{eng}}(f_L) \quad (2)$$

$$\mathrm{Residual}(L) = \mathrm{acc}_M^{\mathrm{eng}}(f_L) - \mathrm{acc}_M(L) \quad (3)$$

By construction, $\mathrm{Total\,Gap}(L) = \Delta_A(L) + \mathrm{Residual}(L)$.

$\Delta_A(L)$ measures the accuracy cost incurred when English content is encoded at fertility f_L rather than at English's native fertility. Because language, model, and content are held constant in computing Δ_A, the entire drop is attributable to the change in tokenization. However, this attribution bundles two distinct mechanisms: (a) the architectural cost of processing more, smaller tokens, and (b) the out-of-distribution shock of seeing input statistics the model has not seen during training. We do not separate (a) from (b) in this work. Δ_A is therefore best interpreted as an upper bound on the deployment-time tokenization cost under naive (unadapted) inference. The analogy to deployment is imperfect: a model deployed on a high-fertility natural language has typically seen some target-language text during pretraining, which partially absorbs the OOD-shock component, whereas our experiment exposes the model to fully unseen (fragmented English) input. We therefore interpret Δ_A as a conservative ceiling on the architectural cost of fragmentation that would be observed at inference time, with the true deployment-time cost likely smaller.

The Residual is everything not captured by Δ_A: differences in the model's knowledge of the target language, typological differences between artificially-fragmented English and the natural target language, and any other cause we have not isolated. We do not further decompose the Residual.

3.4 Adaptation experiment

To investigate whether Δ_A is recoverable through adaptation, we fine-tune Qwen3-1.7B on fragmented English Wikipedia using LoRA at the same fertility level used at evaluation. We use rank $r = 8$, $\alpha = 16$, dropout 0.1, targeting the query and value projections in all attention layers. Training uses effective batch size 32 (per-device batch 2, gradient accumulation 16), learning rate 10^{-4} with 100-step linear warmup, bf16 mixed precision, AdamW. We validate every 200 steps on a held-out Wikipedia

slice and apply early stopping when relative improvement falls below 1% across three consecutive validations, capped at 3000 steps. Both fertility conditions (n=1011 and n=402) converged at step 1600.

Training data consists of approximately 215,000 English Wikipedia articles tokenized under the truncated tokenizer and packed into sequences of 2048 tokens, yielding approximately 156,000 training sequences and a held-out validation set.

3.5 Perplexity diagnostic

To diagnose whether unexpected adaptation results reflect domain shift, we measure mean per-token negative log-likelihood (NLL) on Belebele's 488 unique English passages, comparing the base and LoRA-adapted Qwen3-1.7B with the matching truncated tokenizer.

3.6 Attention analysis

Following Yang et al. (2026), we measure intra-word attention as a proxy for subword-reconstruction work. For each model layer and a held-out sample of Wikipedia passages tokenized under the truncated tokenizer of interest, we identify the word each token belongs to using the BPE space-prefix convention (tokens beginning with the space marker start new words). For each query position, we sum attention weights placed on tokens in the same word (including self) and divide by the total attention from that query, yielding the per-position intra-word ratio. We average over query positions and passages to produce a layer-wise score. High intra-word attention in early layers under fragmentation indicates the model is performing word-reconstruction; in late layers, it indicates the model is attending to its own immediate context for next-token prediction.

4 Experimental Setup

Models. We evaluate four sizes of the Qwen3 family (0.6B, 1.7B, 4B, 8B parameters), all sharing the same tokenizer with vocabulary size 151,936. This lets us isolate the effect of model scale from tokenizer choice. We additionally evaluate Llama-3.1-8B (vocabulary 128,256) to confirm the small-effect regime cross-family. All evaluations use base (non-instruction-tuned) models.

Languages. We evaluate on five Latin-script languages selected to span the multilingual fertility range: French ($f = 1.70$), Spanish ($f = 1.61$), Turkish ($f = 2.53$), Swahili ($f = 2.50$), Yoruba ($f = 2.92$). All fertilities are measured under the Qwen3 tokenizer on Belebele passages. We restrict to Latin-script languages to avoid the distinct fragmentation mechanism arising from byte-level fallback on unfamiliar scripts.

Benchmark and scoring. We use Belebele (Bandarkar et al., 2024), a 900-question multiple-choice reading comprehension benchmark with parallel content in 122 languages. For each example we score the four candidate answer letters via log-likelihood under the model given the prompt and select the highest-scoring letter. The letter tokens (space followed by A, B, C, or D) are single tokens with stable IDs across all our tokenizers, which we verified. Random performance is 25%. This procedure is deterministic given a fixed model checkpoint and benchmark, so multiple-seed averaging does not apply; reported sampling variance derives from the fixed 900-question Belebele set (SE ≈ 1.7 pp for accuracies near 0.5).

Adaptation corpus. We use English Wikipedia (215K articles, approximately 320M tokens before packing) tokenized under each evaluation cutoff. Wikipedia is genre-different from Belebele's FLORES-sourced news passages. We address this potential domain-mismatch confound via the perplexity diagnostic in Section 5.3.

Hardware. All experiments run on a single NVIDIA L40 GPU (48GB VRAM). Inference evaluations across the Qwen3 family and Llama-3.1-8B total approximately 12 GPU hours. Each LoRA adaptation run takes approximately 3.5 hours.

5 Results

5.1 Inverse scaling of the fragmentation cost

Table 2 reports Belebele English accuracy for each model across fragmentation levels matched to Swahili-level and Yoruba-level fertility. The defining observation is that fragmentation cost decreases sharply with model size. At Qwen3-0.6B, Swahili-matched fragmentation drops accuracy by 10.7 points; at Qwen3-1.7B, by 7.0 points; at Qwen3-4B, by 1.4 points; at Qwen3-8B, the drop is within measurement noise. Llama-3.1-8B agrees with Qwen3-8B in the small-effect regime.

The interpretation of the 8B result requires caution. The English baseline at 8B is 91.6%, so there are only 8.4 percentage points of room to drop

Model	Baseline	n=1011 (Swahili)	n=402 (Yoruba)
Qwen3-0.6B	0.484	0.377 (-10.7)	0.368 (-11.6)
Qwen3-1.7B	0.766	0.696 (-7.0)	0.674 (-9.2)
Qwen3-4B	0.896	0.882 (-1.4)	0.878 (-1.8)
Qwen3-8B	0.916	0.920 ($+0.4$)	0.918 ($+0.2$)
Llama-3.1-8B	0.788	0.781 (-0.7)	0.786 (-0.2)

Table 2: Belebele English accuracy under fragmentation. The drop at Swahili-matched fertility decreases sharply with Qwen3 model size, reaching measurement noise at 8B. Llama-3.1-8B shows a similarly small effect at 8B, indicating the near-zero result is not Qwen3-specific.

before reaching the 83% range. We cannot fully separate "tokenization stops mattering at 8B" from "tokenization still matters at 8B but Belebele cannot measure the effect at near-ceiling accuracy." We discuss this in the limitations and identify it as the primary motivation for follow-up evaluation on harder benchmarks. The clean qualitative pattern, sharply decreasing Δ_A from 0.6B through 4B, holds away from any ceiling and is the more robust finding.

5.2 Per-language decomposition

We focus the per-language decomposition on Qwen3-1.7B, which sits in the middle of the scaling family and shows substantial Δ_A values (large enough to measure precisely above noise but well below the 8B ceiling). The decomposition framework applies to any model with a measured English fragmentation curve; we report Qwen3-1.7B as our primary demonstration. Table 3 reports the gap decomposition for Qwen3-1.7B across the five tested languages. $\Delta_A(L)$ is computed by linear interpolation of the Qwen3-1.7B English fragmentation curve at f_L. The Residual is computed by subtraction.

The decomposition reveals that tokenization is differently important for different languages. For Turkish, it accounts for nearly half the gap, consistent with the typological intuition that agglutinative morphology is particularly punished by subword fragmentation. For Swahili and Yoruba, tokenization is real but a minority contributor, reflecting that data scarcity and language-knowledge factors dominate their gaps with English. For French and Spanish, the gaps are small to begin with; tokenization is a moderate contributor to those small gaps.

We emphasize that these are upper bounds on the tokenization-attributable share. Δ_A includes the OOD-shock component discussed in Section 3.3, so the architecturally-irreducible share is no larger than Δ_A and possibly smaller.

5.3 Adaptation paradox

We expected LoRA adaptation on fragmented English Wikipedia to recover at least some of Δ_A, on the reasoning that adaptation removes the model's OOD disadvantage. Table 4 reports the result. Adaptation makes Belebele accuracy *worse*, not better: $\Delta_B(\text{n=1011}) = 12.7$ points (vs $\Delta_A = 7.0$); $\Delta_B(\text{n=402}) = 17.6$ points (vs $\Delta_A = 9.2$).

Perplexity diagnostic. To diagnose whether the accuracy decline reflects domain mismatch (Wikipedia adaptation moving the model away from Belebele's distribution), we measure per-token NLL on Belebele English passages. The result rules out the domain-mismatch hypothesis. Mean per-token NLL drops from 6.39 to 2.10 at n=1011 (perplexity 595.85 to 8.16, a 73x reduction) and from 7.43 to 1.84 at n=402 (perplexity 1687.50 to 6.31, a 267x reduction). The adapted model is dramatically better at modeling Belebele content at the language-modeling level, despite being worse at picking Belebele answers. Language-modeling fit and downstream multiple-choice accuracy decouple sharply under this adaptation.

Attention pattern shift. We measure layer-wise intra-word attention before and after adaptation. Across layers 0-26, the unadapted and adapted models show closely overlapping attention curves. At layer 27 (the final transformer layer of Qwen3-1.7B), intra-word attention nearly doubles after adaptation: from approximately 0.14 in the unadapted model to approximately 0.27 in both the n=1011 and n=402 adapted models. Figure 1 shows this pattern. The early-layer reconstruction signature predicted by Yang et al. (2026) is detectable but muted in this model; we report the unadapted attention pattern across fragmentation levels (Appendix B) for completeness.

Testing the candidate mechanism via ablation. The simultaneity of the layer-27 attention shift and the accuracy drop is suggestive but not causal. We tested causation directly by running a second LoRA fine-tuning experiment with layer 27 excluded from the target modules (LoRA applied to layers 0–26 only), all other hyperparameters identical to the original adaptation. We reasoned that if the layer-27 attention shift directly drives the accuracy decline through weight updates at that layer, exclud-

Language	Fertility	Lang. acc.	Total gap	Δ_A (tok.)	Residual
French	1.70	0.714	0.052	0.016 (31%)	0.037 (69%)
Spanish	1.61	0.666	0.100	0.023 (23%)	0.077 (77%)
Turkish	2.53	0.568	0.198	0.083 (42%)	0.114 (58%)
Swahili	2.50	0.339	0.427	0.070 (16%)	0.357 (84%)
Yoruba	2.92	0.282	0.483	0.091 (19%)	0.392 (81%)

Table 3: Gap decomposition on Qwen3-1.7B (English baseline 0.766). Δ_A is the bound on the tokenization-attributable share of each language's gap. Tokenization explains between 16% (Swahili) and 42% (Turkish) of the observed gap.

	Belebele acc.	Δ from baseline	vs unadapt.
n=1011 unadapted	0.696	7.0	—
n=1011 adapted	0.639	12.7	−5.7
n=402 unadapted	0.674	9.2	—
n=402 adapted	0.590	17.6	−8.4

Table 4: LoRA adaptation on fragmented English Wikipedia reduces Belebele English accuracy at both fragmentation levels, in the opposite direction to expectation. All Δ values in percentage points.

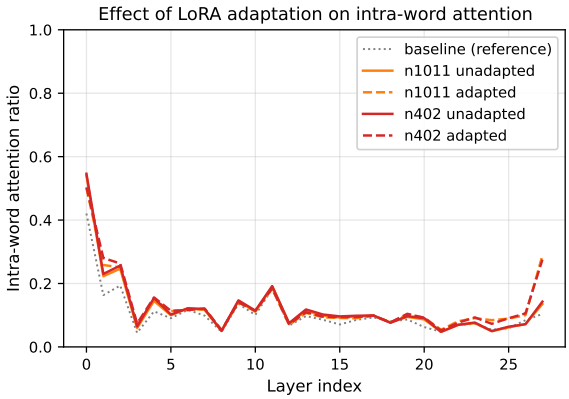

Figure 1: Layer-wise intra-word attention ratio on Qwen3-1.7B, comparing unadapted and LoRA-adapted models at two fragmentation levels (n=1011 and n=402). The unadapted baseline (gray dotted) is included for reference. Adaptation leaves layers 0-26 essentially unchanged but produces a near-doubling of intra-word attention at the final layer (layer 27), from approximately 0.14 to 0.27, in both adapted conditions. We initially proposed this final-layer shift as a candidate mechanism for the decoupling in Table 4. A layer-27-excluded ablation (Section 5.3) rules out direct weight-update causation at this layer; whether the attention shift itself is causal remains untested.

ing layer 27 from adaptation should preserve the perplexity improvement (driven by adaptation in layers 0–26) while restoring Belebele accuracy.

The ablation result is inconsistent with this hypothesis. At n=1011 fragmentation, the no-layer-27 adapter achieves Belebele accuracy 0.630 (vs. full-LoRA 0.639, within evaluation noise; well below unadapted 0.696) and Belebele perplexity 8.56 (vs. full-LoRA 8.16). Excluding layer 27 from adaptation neither restores accuracy nor diminishes the perplexity improvement, ruling out direct weight-update causation at layer 27. The broader hypothesis—that the layer-27 attention shift is causal through input changes propagated from upstream adapted layers—remains untested and would require measuring layer-27 attention in the no-layer-27 adapter. The attention shift alone does not constitute a sufficient mechanistic explanation; the actual mechanism remains open.

The mechanism connecting LoRA adaptation to the LM-versus-MC decoupling is therefore open. Possible directions we did not test include miscalibration of the answer-letter logit distribution induced by representational shifts in mid-network layers, modification of in-context-learning machinery known to reside in middle transformer layers, or interaction effects between the LoRA update direction and the answer-letter token embeddings specifically. Identifying the true mechanism is left to future work.

6 Discussion

Our results contribute three pieces to the structural understanding of the multilingual performance gap.

First, the inverse scaling of Δ_A on Belebele suggests that the tokenization-attributable share of multilingual gaps (combined with OOD-shock sensitivity) decreases substantially with model size. The pattern is sharp from 0.6B through 4B; the result at 8B is partly limited by Belebele's ceiling, which we cannot fully separate from a genuinely small effect. The implication for our understanding of the multilingual gap is that the gap's tokenization-attributable share is itself a function of model scale, and that the residual (other-cause) share becomes proportionally larger as models grow. The natural multilingual gap does not shrink at the same rate as

Δ_A: Qwen3-8B still scores 91.6% on English Belebele and remains substantially worse on Yoruba, indicating that the residual is not scale-eliminable in the way Δ_A appears to be.

Second, the per-language decomposition *at Qwen3-1.7B* shows variability in tokenization's share that is consistent with linguistic intuition. Turkish, a typologically distant but well-resourced language, has the largest tokenization share (42%) of any language we tested. Swahili and Yoruba, both low-resource, have the smallest shares (16% and 19%) despite having comparable fertilities to Turkish. The pattern is consistent with data scarcity dominating tokenization effects for genuinely low-resource languages. Given the inverse-scaling result for English Δ_A, we expect these per-language shares to themselves shrink at larger scales: the same languages evaluated at Qwen3-8B would yield smaller tokenization-attributable components, even if their natural-language gaps remain substantial. The 16–42% bounds reported here are therefore properties of the 1.7B operating point, not scale-invariant constants.

Third, the adaptation paradox documents a previously unreported failure mode in LoRA-based domain adaptation. The decoupling of language-modeling fit from downstream multiple-choice accuracy is sharp and reproducible across two fragmentation levels. We tested whether direct weight updates at layer 27 are causally responsible for the decline via a layer-27-excluded ablation. The decline persisted, ruling out direct weight-update causation at that layer: adaptation in layers 0–26 alone is sufficient to produce the accuracy decline. Whether the correlated layer-27 attention shift is itself causal, driven by upstream representational changes that propagate into an unmodified layer 27, remains untested. This has two implications. First, for evaluating adaptation methods: per-token perplexity improvements, even dramatic ones, do not guarantee downstream task improvements and can in fact coincide with accuracy declines. Second, the layer at which weight updates are applied does not need to coincide with the layer at which visible attention changes appear: in our case, weight updates in earlier layers produced an attention shift localized to the final layer. The actual mechanism producing the decoupling remains open.

Limitations

We list known limitations of this work explicitly.

Ceiling effects at 8B on Belebele. The 8B English Belebele baseline is 91.6%, leaving limited room for measurable drops. We cannot fully distinguish "tokenization tax disappears at 8B" from "tokenization tax exists at 8B but is below Belebele's measurement floor." The inverse-scaling claim is most robust between 0.6B and 4B; the 8B point should be interpreted in light of this ceiling. Validation on a harder benchmark (e.g., FLORES generation evaluation, MMLU) would resolve this concern.

Single tokenizer family for the scaling analysis. The four scaling points (0.6B, 1.7B, 4B, 8B) all use the Qwen3 tokenizer. Llama-3.1-8B shows the same small-effect regime at 8B, but we do not have small Llama models to test whether the scaling shape generalizes. Replicating with Llama-3.2-1B and 3B would close this gap.

Single benchmark. All evaluations use Belebele. The findings may or may not generalize to other benchmark types (FLORES generation, MMLU knowledge, GPQA). Belebele's multiple-choice format is particularly relevant for the adaptation paradox, which relies on log-likelihood scoring of single-token answer letters. The paradox may manifest differently on generation tasks.

Latin-script languages only. We restrict to Latin-script languages to avoid the distinct byte-level fallback mechanism. Extensions to non-Latin scripts (Amharic, Thai, Arabic, etc.) would require additional methodology to handle character-level fallback.

Per-language decomposition reported at one model scale. Table 3 reports the gap decomposition for Qwen3-1.7B only. Extending it to the full scaling family would yield per-language Δ_A trajectories that we expect to inherit the inverse-scaling pattern observed for English, but we do not measure this directly.

Δ_A bundles architectural cost with OOD shock. Our methodology measures the combined cost of (a) processing more, smaller tokens and (b) seeing input distributions absent from training. Separating these two effects would require an adaptation procedure that does not introduce confounds of its own. The adaptation paradox we document indicates that LoRA-based adaptation introduces precisely such a confound, so our reported Δ_A is best interpreted

as a deployment-time bound under naive inference rather than as an architectural-only cost.

The adaptation paradox is LoRA-specific. The adaptation paradox is documented under one configuration: rank-8 LoRA on query and value projections, English Wikipedia, perplexity-based early stopping. Different LoRA configurations, full fine-tuning, different corpora, or different benchmarks may yield different results. The generality of the finding is unknown.

Single-seed LoRA adapters. The adaptation runs in Section 5.3 use a single random seed per fragmentation level. Replicating with multiple seeds would tighten confidence intervals on Δ_B. We expect this to refine the magnitude of the paradox but not its direction: a 5 to 8 point accuracy decline alongside a 70 to 270 times perplexity improvement is too large to be explained by seed variance under standard LoRA hyperparameters.

The mechanistic explanation of the adaptation paradox is open. We documented an LM-perplexity-versus-MC-accuracy decoupling, observed a correlated final-layer attention shift, and tested whether direct weight updates at layer 27 are causally responsible by excluding that layer from adaptation. The accuracy decline persisted, ruling out the narrow hypothesis of direct weight-update causation. The broader hypothesis (that the attention-pattern shift itself is causal, driven by upstream representational changes) remains untested, as it would require measuring layer-27 attention in the no-layer-27 adapter. Other untested candidate explanations include answer-letter logit miscalibration, modification of in-context learning circuitry, and embedding-space interactions. Identifying the true mechanism is an important direction for follow-up work.

Use of AI Assistants

The authors used AI assistants (Claude) for prose editing, LaTeX formatting, and code review during the preparation of this paper. All research design, experimental implementation, analysis, mechanistic interpretation, and scientific claims are the authors' own. Generated text was reviewed and edited by the authors before inclusion.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah A. 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.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. 2025. From tokens to words: On the inner lexicon of LLMs. In *International Conference on Learning Representations (ICLR)*.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2025. Causal estimation of tokenisation bias. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28325–28340, Vienna, Austria. Association for Computational Linguistics.

Jessica M. Lundin, Ada Zhang, Nihal Karim, Hamza Louzan, Guohao Wei, David Ifeoluwa Adelani, and Cody Carroll. 2026. The token tax: Systematic bias in multilingual tokenization. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 103–112, Rabat, Morocco. Association for Computational Linguistics.

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

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 (NeurIPS 2023)*.

Tomohiro Sawada and Kartik Goyal. 2025. Train it and forget it: Merge lists are unnecessary for BPE inference in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing, pages 35045–35058, Suzhou, China. Association for Computational Linguistics.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Zhipeng Yang, Shu Yang, Lijie Hu, and Di Wang. 2026. Word recovery in large language models enables character-level tokenization robustness. *arXiv preprint arXiv:2603.10771*.

A Implementation Details

All experiments use Python 3.11 with PyTorch 2.5.1, Transformers 4.45, and PEFT 0.13. Tokenizer truncation is implemented by re-encoding via the BPE merge list restricted to the first k entries. The truncation cascade is implemented in linear time over sequence length.

Belebele evaluation uses log-likelihood scoring of single-token answer letters (the four tokens corresponding to a space character followed by each of A, B, C, D). We verified that each letter is a single token under all tokenizer truncations used. The prompt format is the Belebele default with the passage, question, and four candidate answers followed by "Answer:".

LoRA training uses the HuggingFace Trainer with linear learning rate warmup over 100 steps and constant learning rate thereafter. We did not tune hyperparameters extensively; the configuration follows standard practice from Hu et al. (2022).

B Additional Attention Analysis

This appendix reports the unadapted attention patterns referenced in Section 5.3. The findings are descriptive and complement the main-text claim about the adaptation-induced final-layer shift.

Across fragmentation levels. Figure 2 shows layer-wise intra-word attention on Qwen3-1.7B (unadapted) at four fragmentation levels: baseline, n=10830 (Spanish-matched), n=1011 (Swahili-matched), and n=402 (Yoruba-matched). Layer 0 intra-word attention rises from approximately 0.42 at the baseline fertility to approximately 0.55 at n=402. Layers 1-2 show smaller shifts (3-5 percentage points). Layers 3-26 show negligible difference across fragmentation levels (all values within ± 0.02). The word-recovery mechanism reported by Yang et al. (2026) is present in Qwen3-1.7B,

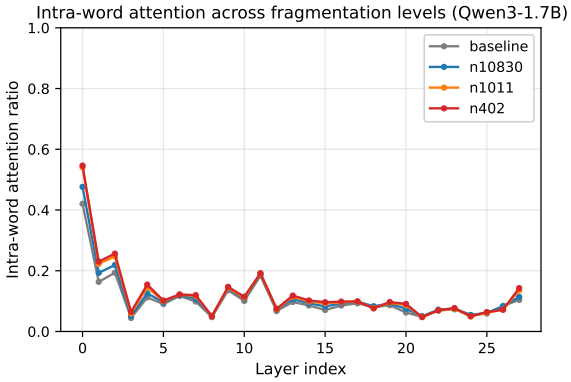

Figure 2: Layer-wise intra-word attention on Qwen3-1.7B (unadapted) across fragmentation levels. The early-layer effect is concentrated in layers 0-2; deeper layers are largely invariant to fragmentation.

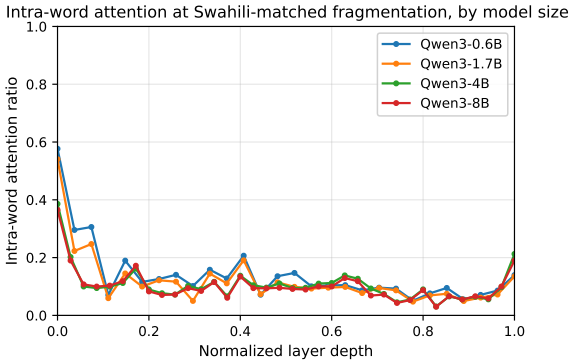

Figure 3: Layer-wise intra-word attention at Swahili-matched fragmentation across the Qwen3 scaling family, plotted against normalized layer depth. Smaller models show higher early-layer intra-word attention, consistent in direction with the inverse scaling of Δ_A.

but concentrated into the first two to three layers and with smaller magnitude than reported in their character-level setting.

Across model sizes at Swahili-matched fertility. Figure 3 shows layer-wise intra-word attention at n=1011 (Swahili-matched) across the four Qwen3 scaling points, with layer index normalized to depth for cross-model comparison. The layer-0 intra-word attention is higher at smaller models (0.58 at 0.6B, 0.55 at 1.7B) than at larger models (0.38 at 4B and 8B), consistent in direction with the Δ_A scaling reported in Section 5.1, though we do not claim a precise mechanistic link.

Per-head heatmaps. Figure 4 shows per-head intra-word attention ratios for Qwen3-1.7B in all five conditions: baseline tokenization, unadapted at n=1011 and n=402, and LoRA-adapted at n=1011

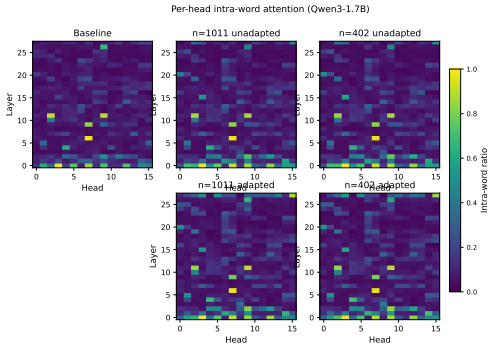

Figure 4: Per-head intra-word attention ratios for Qwen3-1.7B across five conditions: baseline, unadapted at n=1011 and n=402, and adapted at n=1011 and n=402. Adaptation preserves the overall structure of hot heads while producing a final-layer brightening in the adapted conditions (bottom row), consistent with the layer-wise summary in Figure 1.

and n=402. The gross structure of hot heads is preserved across conditions, with adaptation producing a localized brightening at the final layer (layer 27) visible in the bottom row, consistent with the layer-wise summary in Figure 1.