TokLens: A Multilingual Lens on Tokenizer Quality for LLMs

Guan-Ming Chiu
National Taiwan University
gmchiu@arbor.ee.ntu.edu.tw

Abstract

We introduce TokLens, an open-source toolkit for evaluating tokenizer quality across languages using six intrinsic metrics: fertility, characters per token, compression ratio, normalized sequence length, single-token retention rate, and cross-lingual parity. We evaluate 24 tokenizers from major LLM families across 15 typologically diverse languages and correlate these metrics with downstream performance. Our analysis reveals stark disparities: GPT-2 produces 56x more tokens per word in Japanese than in English, while newer tokenizers like Qwen2.5 and Gemma-2 reduce this gap to under 4x. No intrinsic metric predicts English benchmark performance after controlling for model size. However, on multilingual benchmarks (MMLU-ProX), linear mixed-effects models show that tokenizer metrics significantly predict per-language performance (STRR: $\beta = +5.7$, $z = 18.5$, $p < 0.001$). A controlled experiment on the Qwen2.5 family further shows that languages with higher single-token retention rate exhibit steeper scaling slopes ($\rho = 0.91$, $p < 0.001$). These results indicate that tokenizer quality is significantly associated with multilingual LLM performance, though the evidence remains correlational and partially confounded with pretraining data composition.

1 Introduction

Subword tokenization is the standard preprocessing step for virtually all modern LLMs (Sennrich et al., 2016; Kudo and Richardson, 2018). The tokenizer determines how text is segmented into discrete units, directly affecting sequence length, vocabulary coverage, and computational cost. Despite its importance, tokenizer quality is rarely evaluated systematically, particularly across languages.

Recent work has highlighted that tokenizers introduce unfairness between languages (Petrov et al., 2024; Ahia et al., 2023). Texts in non-Latin scripts are often segmented into many more tokens than equivalent English text, increasing inference cost and degrading effective context length. Rust et al. (2021) showed that tokenizer quality correlates with downstream task performance in multilingual masked language models. However, the relationship between tokenizer quality and downstream performance in decoder-only LLMs remains unclear.

As vocabulary sizes have grown from 32K (Llama-2) to 256K (Gemma-2), and training corpora have become more multilingual, the gap between best and worst tokenizers has narrowed, but significant disparities persist. Understanding these disparities, and whether they matter for downstream performance, is important for both model developers and users evaluating multilingual capabilities.

We present TokLens, a lightweight, open-source Python toolkit that computes six tokenizer quality metrics across any number of languages. TokLens is designed to be reusable: researchers can evaluate any HuggingFace-hosted tokenizer on any language with a single command, and model developers can integrate it into their tokenizer development pipeline to audit cross-lingual fairness before deployment (Appendix H). Using TokLens, we evaluate 24 tokenizers spanning the major open-weight LLM families and analyze their relationship with downstream benchmarks. Our contributions are:

1. An open-source toolkit for systematic tokenizer evaluation with six complementary metrics across 15 languages.

2. A comprehensive evaluation of 24 tokenizers from 2019–2025, documenting the evolution of multilingual tokenizer quality.

3. A three-tier statistical analysis: (a) English benchmarks show no significant correlation

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 188–205

after Bonferroni correction; (b) multilingual benchmarks (MMLU-ProX) show significant associations via linear mixed-effects models (z up to 18.5); (c) a controlled experiment with the Qwen2.5 family shows that tokenizer quality is correlated with the per-language scaling pattern ($\rho = 0.91$). All claims are correlational; see Limitations for the residual confound with pretraining data.

Code and data are available at `https://github.com/guan404ming/toklens`.

2 Related Work

Subword tokenization. BPE (Sennrich et al., 2016) and its variants remain the dominant approach. GPT-2 (Radford et al., 2019) popularized byte-level BPE. SentencePiece (Kudo and Richardson, 2018) introduced language-independent tokenization, adopted by Llama 2 and Mistral (Jiang et al., 2023). Recent models like Llama 3 (Dubey et al., 2024) and Qwen2.5 (Yang et al., 2024) use tiktoken-style tokenizers with vocabularies exceeding 128K tokens.

Tokenizer evaluation. Rust et al. (2021) introduced fertility as a tokenizer quality metric, finding that it correlates with downstream performance in multilingual masked language models. Petrov et al. (2024) formalized tokenizer unfairness across languages. Ahia et al. (2023) quantified the economic impact, finding that API costs for non-English languages can be 2–15x higher. Ali et al. (2024) found that tokenizer choice has a negligible effect on downstream English performance when training compute is held constant. Chelombitko et al. (2024) proposed Qtok, a framework evaluating tokenizer completeness across languages and linguistic categories for 13 tokenizers. TokLens differs by focusing on the correlation between intrinsic metrics and downstream benchmarks, covering 24 tokenizers with mixed-effects modeling.

Multilingual vocabularies. Liang et al. (2023) proposed allocating vocabulary capacity proportionally to each language's corpus size. Limisiewicz et al. (2023) showed that vocabulary allocation and cross-lingual token overlap significantly impact downstream multilingual tasks, with language-specific coverage being the dominant factor. BLOOM (BigScience Workshop, 2023) demonstrated that multilingual vocabulary construction (251K tokens, 46 languages) can achieve near-parity efficiency. Our work evaluates a broader set of 24 tokenizers with six metrics and correlates with the Open LLM Leaderboard v2 reasoning benchmarks.

Beyond standard BPE. A complementary line of work moves away from frequency-driven BPE altogether. Fairness-aware tokenization re-weights or constrains BPE training to reduce per-language disparity (Petrov et al., 2024; Ahia et al., 2023; Limisiewicz et al., 2023). Tokenization-free models bypass the issue at the cost of longer sequences: byte-level encoders such as ByT5 (Xue et al., 2022), character-level models such as CANINE (Clark et al., 2022), and pixel-based models such as PIXEL (Rust et al., 2023) eliminate vocabulary disparity by construction. TokLens metrics are defined on token-ID sequences and apply directly to any of these alternatives once they emit a discrete unit stream (or, for PIXEL, once patches are treated as discrete tokens); we leave a systematic comparison to future work.

Tokenizer evaluation suites. Chelombitko et al. (2024) (Qtok) is the closest prior toolkit on the multilingual axis but operates at a different level of analysis: Qtok inspects each token in the *vocabulary* and reports its category (control, alpha, errors, etc.) and per-language allocation, while TokLens runs the tokenizer over a *corpus* and reports per-language fertility, STRR, parity, and related rates. The two are therefore complementary; Appendix G quantifies their agreement on a 16-tokenizer overlap and gives a feature matrix. A separate line of work focuses on controlled side-by-side tokenizer comparisons under matched compute: Altıntaş et al. (2025) (TokSuite) train 14 otherwise identical LMs that differ only in tokenizer, yielding stronger causal evidence than ours; our LME analysis observationally adds training-token controls in this spirit (Appendix D).

3 Metrics

TokLens computes six metrics on a given tokenizer and text corpus. All metrics operate on token ID sequences with special tokens stripped for fair comparison.

Fertility. The average number of tokens per whitespace-delimited word. For language ℓ with

words $w_1, \ldots, w_n$:

$$\text{Fertility}(\ell) = \frac{1}{n} \sum_{i=1}^{n} |T(w_i)| \tag{1}$$

where $T(w_i)$ is the token sequence for word w_i. A fertility of 1.0 means every word is a single token; values above 2.0 indicate heavy fragmentation.

Characters Per Token (CPT). The ratio $|c|/|t|$ of characters to tokens. Higher CPT means better compression, but CPT can be misleading cross-lingually because a single Chinese character carries more semantic content than a Latin character.

Compression Ratio (CR). UTF-8 bytes divided by token count: $|b|/|t|$. By measuring bytes rather than characters, CR accounts for multi-byte scripts and provides a fairer cross-lingual comparison than CPT.

Single-Token Retention Rate (STRR). The fraction of words tokenized as a single token. STRR directly measures vocabulary coverage: an STRR of 0.7 means 70% of words appear intact in the vocabulary.

Normalized Sequence Length (NSL). Token count divided by character count: $|t|/|c|$, the inverse of CPT. Values below 1.0 mean the tokenizer compresses the text; values above 1.0 indicate expansion.

Parity Ratio. The ratio of tokens-per-character between a target language and English. On parallel text, parity reduces to the simple ratio $|t_\ell|/|t_{\text{en}}|$ used by Petrov et al. (2024). To accommodate non-parallel native corpora, we evaluate each language on a fixed character budget C (Section 4.2, $C \approx$ 50K) and define

$$\text{Parity}(\ell) = \frac{|t_\ell|/|c_\ell|}{|t_{\text{en}}|/|c_{\text{en}}|}, \tag{2}$$

where $|c_\ell|$ is the character count of the corpus in language ℓ. Since both corpora share the same character budget, this reduces in practice to $|t_\ell|/|t_{\text{en}}|$, but the explicit per-corpus character normalization makes the metric well defined when corpus sizes differ slightly. A value of 1.0 means equal per-character tokenization cost; higher values indicate the tokenizer is less efficient for language ℓ, directly measuring cross-lingual fairness. We do not normalize by UTF-8 bytes, which would penalize multi-byte scripts (CJK, Arabic, Devanagari) by a factor proportional to their byte width; we instead report compression ratio (CR) as the byte-level companion of parity. No aligned subset is used.

4 Experimental Setup

4.1 Tokenizers

We evaluate 24 tokenizers from major LLM families (Table 1). Twenty-two have scores on the Open LLM Leaderboard v2 (Fourrier et al., 2024); two (Qwen3-8B, DeepSeek-V3) are included for metric-only analysis. Models range from 0.1B to 35B parameters, vocabulary sizes from 32K to 256K tokens, and span three implementation families: byte-level BPE, SentencePiece BPE, and tiktoken-style BPE.

Several models share tokenizers. Phi-3-mini reuses Llama-2's 32K vocabulary. The Qwen2.5 family uses the same 152K tokenizer across all sizes (0.5B to 14B), providing a natural experiment for isolating model size effects.

4.2 Languages and Corpora

We evaluate on 15 languages spanning 6 scripts: English, Chinese, Japanese, Arabic, Hindi, German, Turkish, Korean, Thai, Russian, French, Spanish, Portuguese, Vietnamese, and Indonesian. The scripts covered are Latin, CJK, Arabic, Devanagari, Thai, and Cyrillic. Corpora are drawn from native-language Wikipedia articles (100 per language, approximately 50K characters each), not parallel translations, avoiding translationese artifacts. Because the corpora are non-parallel, parity is computed in the character-normalized form defined in Section 3 (tokens per character for ℓ over tokens per character for English); no aligned subset is used.

4.3 Benchmarks and Correlation Method

We use scores from the Open LLM Leaderboard v2, which evaluates models on six benchmarks: IFEval, BBH, MATH Lvl 5, GPQA, MUSR, and MMLU-PRO (full scores in Appendix C). We compute Spearman rank correlations between each tokenizer metric and each benchmark across the 22 models with leaderboard entries. We also compute partial Spearman correlations controlling for model size. All variables are first rank-transformed, then the standard partial correlation formula is applied to the ranks:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \tag{3}$$

Tokenizer	Vocab	Params	Type
GPT-2	50K	0.1B	BPE
Llama-2-7B	32K	6.7B	SP-BPE
Llama-3.1-8B	128K	8.0B	tiktoken
Qwen2.5-{0.5–14}B	152K	0.5–14.8B	BPE
Gemma-2-9B	256K	9.0B	SP-BPE
Mistral-7B-v0.3	33K	7.2B	SP-BPE
Mistral-Nemo-12B	131K	11.6B	tiktoken
Phi-3-mini	32K	3.8B	SP-BPE
Phi-4	100K	14.7B	tiktoken
Yi-1.5-9B	64K	8.8B	BPE
Falcon-7B	65K	7.0B	BPE
BLOOM-7B	251K	7.1B	BPE
Command-R-35B	255K	35.0B	BPE
StableLM-2-12B	100K	12.1B	BPE
GLM-4-9B	151K	9.0B	tiktoken
OLMo-2-7B	100K	7.3B	BPE
InternLM2.5-7B	93K	7.7B	SP-BPE
SmolLM2-1.7B	49K	1.7B	BPE
Qwen3-8B	152K	–	BPE
DeepSeek-V3	129K	–	BPE

Table 1: Tokenizers evaluated. SP-BPE = SentencePiece BPE. The Qwen2.5 family shares a single tokenizer across all sizes. Bottom two are metric-only (no benchmark scores).

where r denotes Spearman (rank-based) correlations and z is model size in parameters. We apply Bonferroni correction across all 77 metric-benchmark pairs ($\alpha_{\text{corrected}} = 0.00065$).

5 Results

5.1 Cross-lingual Tokenizer Disparities

Figure 1 shows the parity ratio across all tokenizers and languages. GPT-2's tokenizer produces a parity ratio of 8.8x for Japanese, 6.6x for Chinese, and 14.9x for Russian (the highest in our evaluation, driven by poor Cyrillic coverage). Even Llama-2 (32K vocabulary, 2023) shows 5.7x for Japanese and 4.0x for Chinese. Full parity values are in Appendix A.

In contrast, tokenizers with larger vocabularies and multilingual training data achieve much better parity. BLOOM (251K, explicitly multilingual) reaches parity ratios below 2.0 for most languages. Gemma-2 (256K) and Command-R (255K) show similar improvements. Qwen2.5 (152K), despite a smaller vocabulary, achieves 1.1x parity for Chinese due to its bilingual design, though its Japanese (2.1x) and Korean (2.2x) parity is less impressive.

Among Latin-script languages, parity ratios are generally low (1.0–2.5x) across all tokenizers, since BPE training corpora include substantial Latin-script text. The exceptions are Turkish (up to 3.6x) and Vietnamese (up to 3.5x), which use diacritics that increase byte counts.

5.2 Fertility and STRR Across Languages

Fertility varies enormously across languages (Appendix I). English fertility ranges from 1.4 (Gemma-2) to 1.6 (GPT-2). For Japanese, the range is dramatic: 91.1 for GPT-2 versus 37 for Gemma-2, driven by byte-level fallback where each Japanese character (3 UTF-8 bytes) becomes 3 tokens when no vocabulary entry matches.

Korean is an interesting case: it uses a phonetic alphabet (Hangul) with compositional syllable blocks. Tokenizers with good Unicode handling (Qwen2.5, Gemma-2) achieve fertility of 2–4, while byte-fallback tokenizers (GPT-2, Falcon) produce 8–9 because each Hangul syllable is 3 UTF-8 bytes.

Thai presents a different challenge: it does not use spaces between words, so whitespace-delimited "words" in Thai text are often entire phrases. This makes absolute fertility values less comparable to other languages, but relative comparisons across tokenizers remain informative. GPT-2 produces a fertility of 53.9 on Thai, while Gemma-2 achieves 10.9, a 5x improvement driven by the inclusion of Thai character sequences in the vocabulary.

For STRR, English ranges from 0.61 (GPT-2) to 0.73 (Gemma-2), meaning that even the best tokenizer splits nearly a quarter of English words into subword pieces. Arabic STRR drops below 0.01 for most tokenizers, reflecting both Arabic's rich morphology (where common words include prefixed articles, prepositions, and pronominal suffixes) and limited vocabulary allocation to Arabic script. Vietnamese, despite using Latin script, has low STRR (0.07–0.20) due to diacritics creating character combinations rarely seen in English-centric training corpora. The full STRR heatmap is in Appendix J.

5.3 Characters Per Token

CPT varies widely across languages (Appendix K). For English, CPT ranges from 4.3 (Llama-2) to 5.2 (Gemma-2). For Chinese, CPT is below 1.0 for most tokenizers, meaning each character produces more than one token on average. The exceptions are BLOOM (2.6), Qwen2.5 (2.8), and Gemma-2 (2.2), which have sufficient CJK vocabulary to compress multiple characters into single tokens. The difference between CPT and compression ratio (CR) is most visible for CJK and Thai: since Chi-

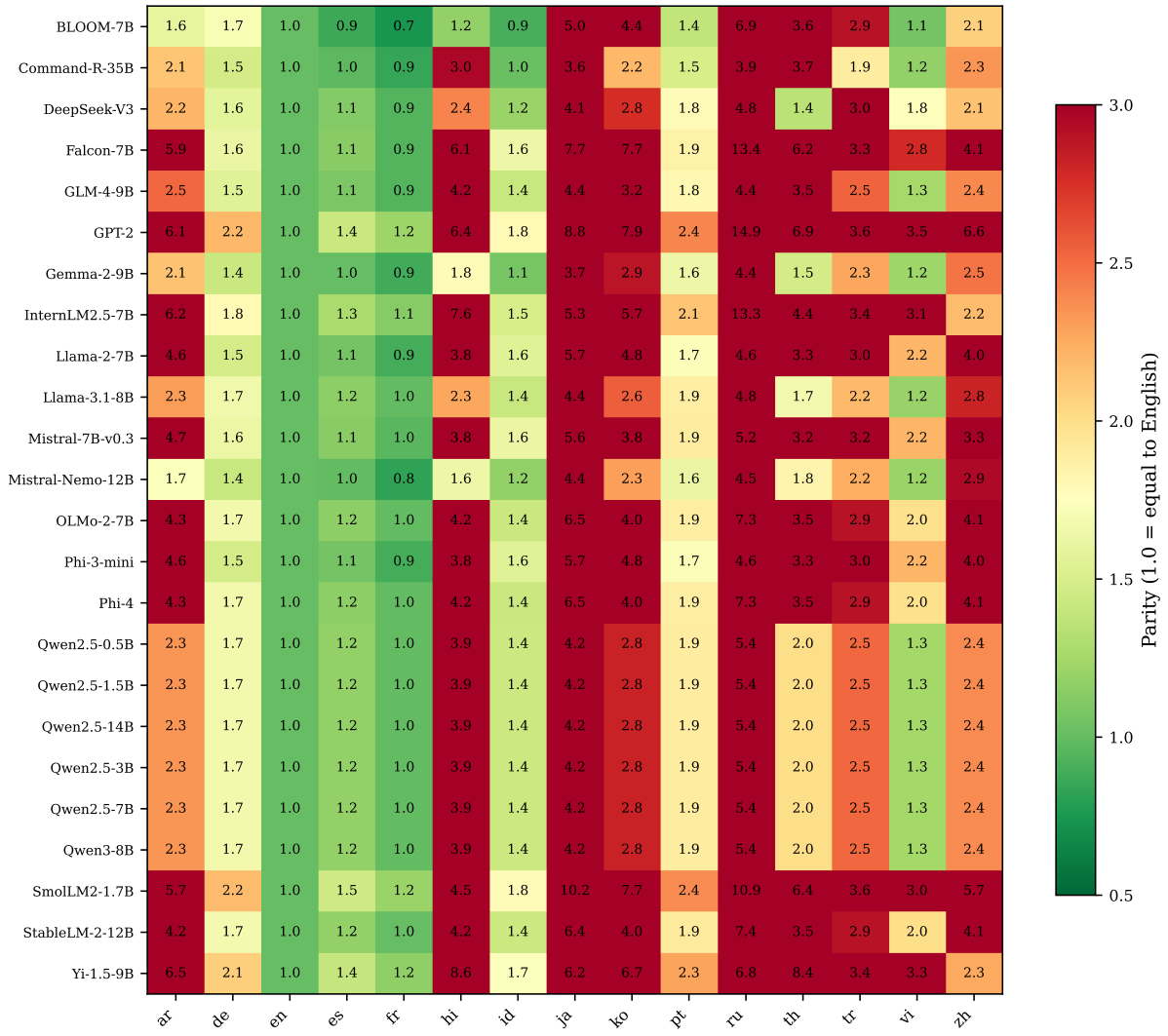

Figure 1: Cross-lingual parity ratio, computed as the ratio of tokens-per-character in language ℓ to tokens-per-character in English on the non-parallel native Wikipedia corpora with a shared character budget (Section 4.2). Values above 1.0 indicate more tokens are needed per character than for English. Older tokenizers show extreme disparities for CJK, Thai, and Cyrillic, while newer large-vocabulary tokenizers achieve near-parity.

nese characters use 3 UTF-8 bytes each, a CPT of 0.6 (GPT-2 on Chinese) corresponds to a CR of 1.5, revealing that the tokenizer does compress bytes even though it expands characters.

5.4 Correlation with Downstream Performance

Figure 2 shows the Spearman correlation heatmap between English tokenizer metrics and benchmarks. Table 2 lists the strongest correlations. The highest absolute correlation is between average fertility and MMLU-PRO ($\rho = -0.38$, $p = 0.08$). However, no correlation reaches significance after Bonferroni correction ($\alpha_{\text{corrected}} = 0.00065$, 77 comparisons). The full correlation table is in Appendix E.

Partial correlations controlling for model size are uniformly weaker (Table 2, rightmost column). The strongest partial correlation is $r = -0.28$ (fertility vs. MATH Lvl 5). This suggests the weak raw

Metric	Bench	ρ	p	r_{partial}
Fertility (avg)	MMLU-PRO	−0.38	0.082	−0.26
CPT (en)	MMLU-PRO	0.35	0.105	0.17
Fertility (avg)	MATH	−0.35	0.111	−0.28
CPT (en)	MUSR	0.35	0.116	0.16
NSL (avg)	MMLU-PRO	−0.33	0.136	−0.10
Parity (avg)	MMLU-PRO	−0.32	0.141	−0.16
Fertility (avg)	BBH	−0.30	0.168	−0.16

Table 2: Top Spearman correlations between tokenizer metrics and benchmarks ($n = 22$). r_{partial}: partial correlation controlling for model size. No correlation is significant after Bonferroni correction.

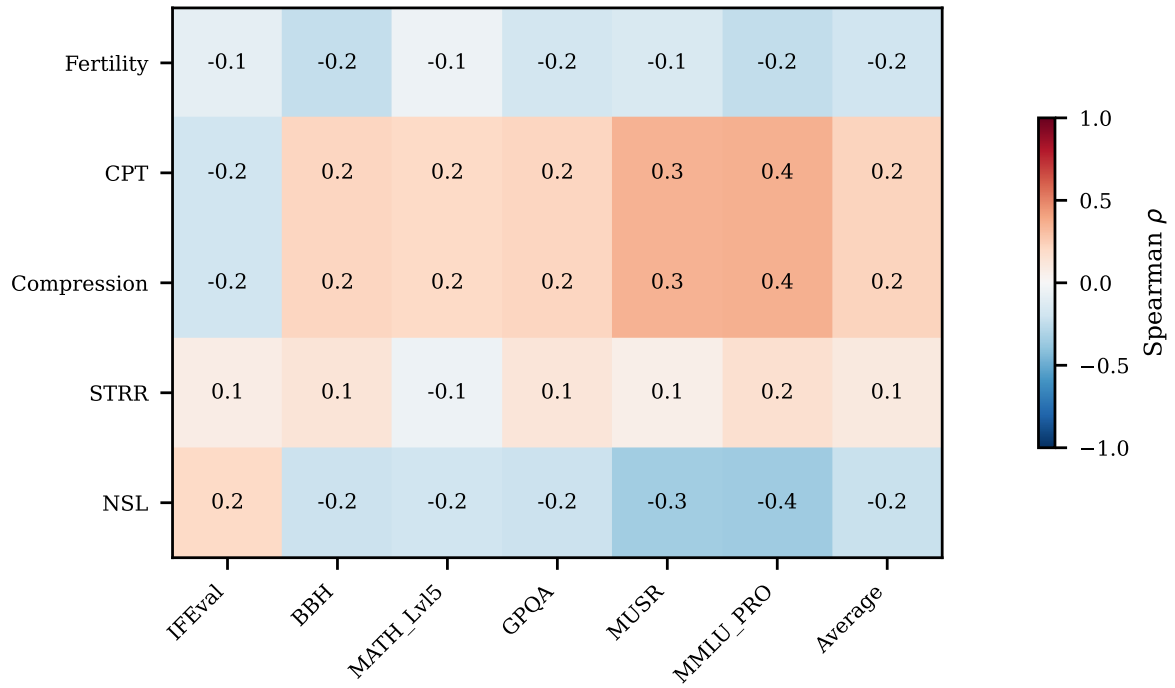

Figure 2: Spearman correlation heatmap between English tokenizer metrics and Open LLM Leaderboard v2 benchmarks ($n = 22$). No cell reaches significance after Bonferroni correction.

correlations are partly driven by the confound that larger models tend to have both better tokenizers and higher benchmark scores.

Figure 3 illustrates this confound directly: while there is a visual trend of lower fertility corresponding to higher benchmark scores, the color gradient (model size) reveals that the trend is largely driven by scale. Small models (GPT-2, SmolLM2, Qwen2.5-0.5B) cluster in the lower-right with high fertility and low scores, while large models (Phi-4, Qwen2.5-14B, Command-R) cluster in the upper-left. Within a fixed size range (7–9B), no clear fertility-performance relationship exists.

After removing the model size confound, all partial correlations fall within $|r| < 0.3$ with no clear pattern, confirming that tokenizer metrics carry little predictive power beyond model scale for English benchmarks. Results remain non-significant under Benjamini-Hochberg FDR correction ($q > 0.10$ for all pairs), which is more appropriate given the collinearity among metrics.

5.5 Vocabulary Size and Performance

Figure 4 plots vocabulary size against average benchmark score. Vocabulary size shows no meaningful correlation with performance. Among 7–9B models, vocabulary sizes range from 32K to 256K,

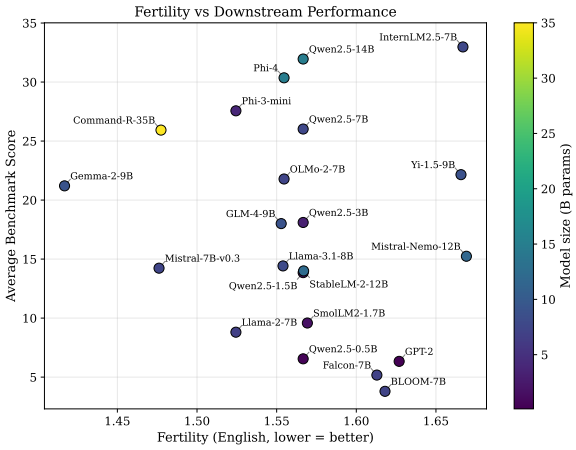

Figure 3: English fertility vs. average benchmark score, colored by model size. The apparent negative trend is largely driven by the model size confound: larger models have both better tokenizers and higher scores.

yet scores vary more with architecture and training data.

The Qwen2.5 family provides a natural controlled experiment: all sizes share the same 152K tokenizer, and performance scales cleanly with parameters (0.5B: 6.6, 1.5B: 13.9, 3B: 18.1, 7B: 26.0, 14B: 32.0). BLOOM-7B (251K vocab) scores 3.8, lower than Llama-2-7B (8.8, 32K vocab) and far below Qwen2.5-7B (26.0, 152K vocab), demonstrat-

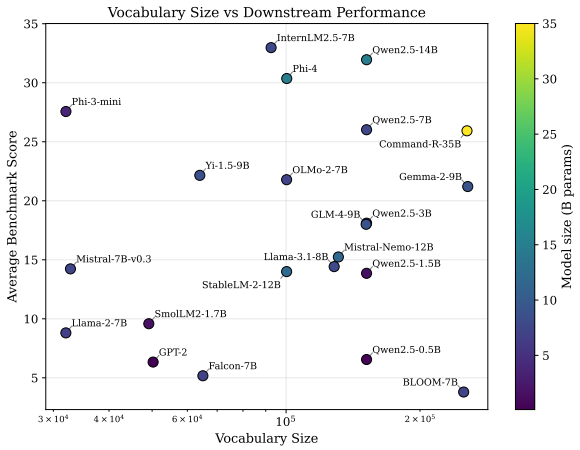

Figure 4: Vocabulary size vs. average benchmark score, colored by model size. The dominant factor is model scale, not vocabulary size.

ing that a large vocabulary does not compensate for architecture and training data differences.

5.6 Multilingual Downstream Correlation

To test whether tokenizer quality predicts *multilingual* downstream performance, we correlate per-language TokLens metrics with per-language scores from MMLU-ProX (Li et al., 2025), a multilingual extension of MMLU-Pro with 5-shot chain-of-thought prompting. We identify 7 overlapping models across 11 shared languages, yielding 77 (model, language) data points. For each pair, we compute TokLens metrics in that language and pair them with the corresponding MMLU-ProX score.

Figure 6 shows the results. Unlike the English-only null result, several metrics significantly correlate with multilingual performance, and partial correlations controlling for model size are *stronger*: STRR partial $\rho = +0.48$, CPT $+0.44$, fertility -0.36 (all $p \leq 0.001$, surviving Bonferroni correction). A caveat: for Chinese, Japanese, and Thai, STRR and fertility are computed on whitespace-delimited units, which do not correspond to linguistic words. For these languages, STRR measures character-sequence vocabulary coverage rather than true word-level retention. To verify that these values do not drive the results, we re-run the LME analysis excluding zh, ja, and th (56 observations, 8 languages): all metrics remain highly significant (STRR $z = 16.3$, $p < 0.001$), confirming that the correlations are not artifacts of script typology.

Mixed-effects models. The pooled Spearman analysis treats each (model, language) pair as independent, but the 77 observations come from only 7 models, creating non-independence: observations from the same model share training data, architecture, and optimizer choices. To account for this clustered structure, we fit linear mixed-effects models (LME): score $\sim$ z(metric) + log(params) + (1|model), where the random intercept absorbs model-specific variance. LME results confirm the Spearman findings with proper uncertainty estimates: STRR ($\beta = +5.68$, $z = 18.5$, $p < 0.001$), CPT ($\beta = +5.03$, $z = 11.0$, $p < 0.001$), and parity ($\beta = -4.29$, $z = -7.4$, $p < 0.001$). The model random intercept variance is large (100–115), confirming that model identity is the dominant source of variation, but tokenizer metrics provide significant additional explanatory power. Adding log(Wikipedia articles) as a proxy for pretraining data volume, STRR remains significant ($\beta = +4.31$, $p < 0.001$), as do CPT and parity, while fertility loses significance ($p = 0.14$). Full results are in Appendix D.

Robustness checks. We run three sensitivity analyses; full coefficients are in Appendix D. (i) *Training-token control.* Adding $\log(\text{train\_tokens})$ to M1 (Qwen2.5: 18T; Llama-3.1: 15T; Phi-4: 9.8T; Gemma-2: 8T; Mistral-7B-v0.3: imputed 8T) leaves STRR, CPT, and parity essentially unchanged; $\log(\text{train\_tokens})$ itself is not significant ($p \geq 0.23$), reflecting the narrow training-token range among these 7 models. Dropping the imputed Mistral row gives the same qualitative pattern (STRR $\beta = +5.54$, $z = 16.0$, $p < 0.001$). (ii) *Tokenizer-internal data proxy.* Replacing Wikipedia size with a vocabulary-derived proxy — the per-(model, lang) % of vocabulary devoted to the target language's script (Latin from Qtok (Chelombitko et al., 2024) for en/fr/de/es/pt; Unicode-block for ja/ko/zh/ar/th/hi) — yields a highly significant proxy effect ($p < 0.001$); STRR ($\beta = +5.16$, $z = 13.9$), parity, CPT, and NSL remain significant, fertility becomes significant ($p = 0.0005$), and CR collapses ($p = 0.41$). (iii) *Leave-one-language-out CV.* Re-fitting M1 with each language held out keeps STRR significant in all 11 fits ($\beta \in [+4.89, +6.00]$, all $p < 0.001$).

Controlled experiment: Qwen2.5 scaling. The cross-model analyses above cannot fully disentangle tokenizer quality from training data composition. The Qwen2.5 family (0.5B–14B) partially addresses this: all five sizes share the same 152K tokenizer, training pipeline, and data mixture, so

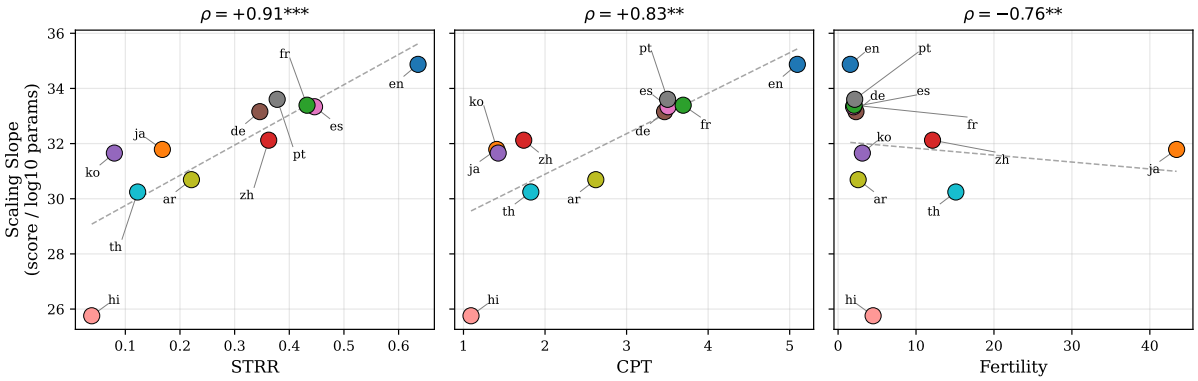

Figure 5: Per-language scaling slope (score gain per decade of model size) vs. TokLens metrics for Qwen2.5. STRR shows the strongest correlation ($\rho = +0.91$, $p < 0.001$): languages with higher single-token retention benefit more from scaling. Hindi (STRR = 0.04) gains 25.8 points per decade vs. 34.9 for English (STRR = 0.64).

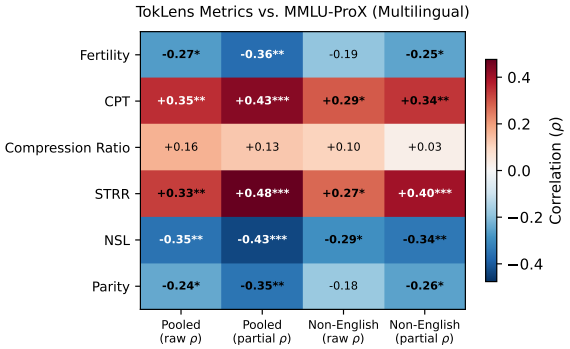

Figure 6: Correlation between TokLens metrics and MMLU-ProX scores across 7 models and 11 languages. Stars: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Partial correlations control for model size. Unlike English-only analysis, several metrics significantly predict multilingual performance.

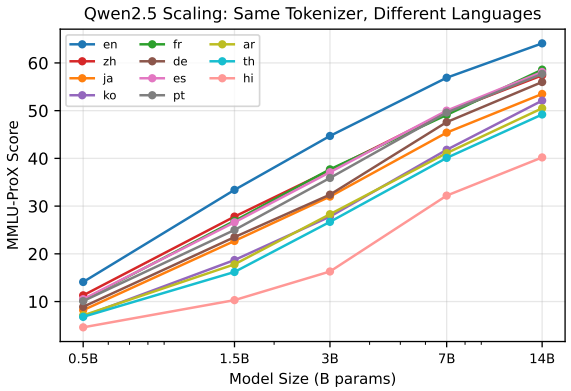

Figure 7: Qwen2.5 MMLU-ProX scaling curves. Each line is a language; all models share the same tokenizer. Languages with better tokenizer quality (en, fr, es) consistently score higher and show steeper scaling.

per-language performance differences can be more directly attributed to tokenizer quality. We regress per-language MMLU-ProX scores on log(params) and correlate scaling slopes with TokLens metrics. Figure 7 shows that languages with higher tokenizer quality metrics consistently score higher and scale more steeply. Figure 5 quantifies this: STRR shows a strong correlation with scaling slope ($\rho = +0.91$, $p < 0.001$). English (STRR = 0.64) gains 34.9 points per decade of scale, while Hindi (STRR = 0.04) gains only 25.8. A potential confound is that languages with high STRR (English, French, Spanish) are also likely overrepresented in Qwen2.5's pretraining data. To partially address this, we use Wikipedia article count per language as a proxy for pretraining data volume. STRR and Wikipedia size are indeed correlated ($\rho = 0.77$), but after controlling for log(Wikipedia articles), STRR remains strongly associated with scaling slope (partial $\rho = 0.78$, $p = 0.007$). This suggests that STRR captures variance in scaling benefits beyond what data volume alone explains, though Wikipedia size is a coarse proxy: actual pretraining mixtures include Common Crawl, code, and other sources with different language distributions. Full causal claims require controlling for the actual pretraining mixture.

6 Discussion

Interpreting the English null result. The English-only analysis returns no significant correlation, which we read as a metric / benchmark mismatch rather than evidence that tokenization is irrelevant in English. The v2 suite (IFEval, BBH, MATH, GPQA, MUSR, MMLU-Pro) stresses reasoning and knowledge; the dominant variance axis is model capability, not tokenizer efficiency on

Wikipedia-style text. A further check using held-out English BPB (Appendix F) confirms that BPB is explained by model scale alone ($\rho = -0.94$, $p = 0.005$), and BPB vs. any English TokLens metric is non-significant ($|\rho| \leq 0.52$, $p \geq 0.29$). The culprit is limited spread: English STRR ranges only 0.636–0.731 across our 7 LME models, so this is a power / spread limitation, not evidence of independence.

Why STRR dominates multilingually. STRR emerges as the strongest correlate (LME $z = 18.5$; Qwen2.5 $\rho = 0.91$), likely because single-token words preserve semantic unity while fragmentation forces reconstruction across positions. Fertility is marginal in the LME ($p = 0.065$) because its cross-script variance is absorbed by the random intercept.

Practical recommendations. STRR ≥ 0.3 for non-Latin scripts is reliably achieved with $\geq$ 100K vocabulary and balanced per-script allocation; below 100K, STRR for CJK, Devanagari, and Thai falls under 0.05. A large vocabulary is necessary but not sufficient: BLOOM (251K) lags Gemma-2 (256K) on East Asian scripts, so allocation strategy matters more than raw size. Practitioners should prioritize STRR over compression ratio, as STRR has the largest LME effect and survives every control tested.

7 Conclusion

We introduced TokLens, an open-source toolkit for multilingual tokenizer evaluation, and analyzed 24 tokenizers across 15 languages. No intrinsic metric is associated with English benchmark performance after controlling for model size; we discuss why this likely reflects a metric / benchmark mismatch in Section 6. Multilingual analysis using MMLU-ProX shows that STRR, CPT, and parity are significantly associated with per-language performance via linear mixed-effects models ($p < 0.001$), and a controlled experiment with Qwen2.5 shows that tokenizer quality is correlated with the per-language scaling pattern ($\rho = 0.91$). All findings are correlational and remain partially confounded with pretraining data composition (see Limitations).

Limitations

Our multilingual analysis is restricted to 7 models and 11 languages from MMLU-ProX; we surveyed alternative multilingual benchmarks (Belebele, Global-MMLU, INCLUDE) but could not locate publicly reported scores covering all 7 models on the 11 target languages. Leave-one-language-out cross-validation (Appendix D) confirms the signal is not driven by any single language.

Second, training data composition co-varies with tokenizer design and we cannot fully isolate the tokenizer's contribution. We control for total compute via $\log(\text{train\_tokens})$ (Appendix D, M2/M3) and find stable effects, but per-language pretraining mix is not directly observed. Controlled experiments that vary only the tokenizer (Ali et al., 2024; Altıntaş et al., 2025) provide stronger causal evidence but require retraining from scratch and are limited to a single architecture each.

Third, Wikipedia may not represent all domains; tokenizer quality may differ for code, social media, or scientific text.

Fourth, whitespace-based word segmentation is ill-defined for Chinese, Japanese, and Thai, making fertility and STRR less interpretable for these languages. Integrating language-specific segmenters (e.g., Jieba, MeCab, PyThaiNLP) would yield more meaningful word boundaries; we leave this for future work as TokLens currently prioritizes language-independent evaluation.

Fifth, TokLens covers only BPE-family tokenizers; the metric definitions extend to any discrete-unit system but we leave cross-paradigm comparison to future work (Clark et al., 2022; Xue et al., 2022).

Finally, the English-only sample ($n = 22$) limits statistical power; the multilingual pooled analysis ($n = 77$) partially mitigates this, but within-language correlations ($n = 7$) remain underpowered.

Ethics Statement

This work evaluates tokenizer quality across languages, relevant to the fairness of language technology. Our finding that tokenizers produce dramatically different token counts across languages has implications for API pricing, context window utilization, and model accessibility. We release TokLens as open source to enable tokenizer fairness audits. No human subjects were involved in this research.

Acknowledgements

We thank Modal for the compute credits that supported the experiments in this work.

References

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 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 9524–9541. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Perea, Maximilian Nagel, Hassan Imani, Niclas Kerkmann, Nishant Venkatraman, Jilles Vreeken, Jörn Hees, and Sven Schmeier. 2024. Tokenizer choice for LLM training: Negligible or crucial? *arXiv preprint arXiv:2310.08754*.

Gül Sena Altıntaş, Malikeh Ehghaghi, Brian Lester, Fengyuan Liu, Wanru Zhao, Marco Ciccone, and Colin Raffel. 2025. TokSuite: Measuring the impact of tokenizer choice on language model behavior. *arXiv preprint arXiv:2512.20757*.

BigScience Workshop. 2023. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Iaroslav Chelombitko, Egor Safronov, and Aleksey Komissarov. 2024. Qtok: A comprehensive framework for evaluating multilingual tokenizer quality in large language models. *arXiv preprint arXiv:2410.12989*.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Clémentine Fourrier, Nathan Habib, Quentin Lhoest, and Thomas Wolf. 2024. Open LLM leaderboard v2. *Hugging Face Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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. Association for Computational Linguistics.

Weihao Li, Zhe Zhang, and Haoyang Yu. 2025. MMLU-ProX: A multilingual extension of MMLU-Pro. *arXiv preprint arXiv:2503.10497*.

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

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Biber. 2024. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *International Conference on Learning Representations (ICLR)*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulic, 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. 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. Association for Computational Linguistics.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Cross-lingual Parity

Parity ratios are character-normalized (Section 3). Models sharing a tokenizer are listed once. Values above 3.0 indicate substantial token-cost disadvantage for non-English users. Latin-script languages are omitted because their parity ratios stay close to 1.0 across all tokenizers in our set. The non-Latin columns (zh, ja, ar, hi, ko, th, ru) span all four non-Latin script families covered by our corpora, providing a compact view of the cross-script disparity that is otherwise spread across Figure 1.

Tokenizer	zh	ja	ar	hi	ko	th	ru
GPT-2	6.6	8.8	6.1	6.4	7.9	6.9	14.9
Llama-2-7B	4.0	5.7	4.6	3.8	4.8	3.3	4.6
Llama-3.1-8B	2.8	4.4	2.3	2.3	2.6	1.7	4.8
Qwen2.5	2.4	4.2	2.3	3.9	2.8	2.0	5.4
Gemma-2-9B	2.5	3.7	2.1	1.8	2.9	1.5	4.4
Mistral-7B-v0.3	3.3	5.6	4.7	3.8	3.8	3.2	5.2
Mistral-Nemo-12B	2.9	4.4	1.7	1.6	2.3	1.8	4.5
Yi-1.5-9B	2.3	6.2	6.5	8.6	6.7	8.4	6.8
Falcon-7B	4.1	7.7	5.9	6.1	7.7	6.2	13.4
BLOOM-7B	2.1	5.0	1.6	1.2	4.4	3.6	6.9
Command-R-35B	2.3	3.6	2.1	3.0	2.2	3.7	3.9
GLM-4-9B	2.4	4.4	2.5	4.2	3.2	3.5	4.4
InternLM2.5-7B	2.2	5.3	6.2	7.6	5.7	4.4	13.3
SmolLM2-1.7B	5.7	10.2	5.7	4.5	7.7	6.4	10.9
DeepSeek-V3	2.1	4.1	2.2	2.4	2.8	1.4	4.8

Table 3: Cross-lingual parity ratios for non-Latin languages.

B English Tokenizer Metrics

Table 4 reports the six TokLens metrics on English Wikipedia for each unique tokenizer in our set. The Qwen2.5 family (0.5B–14B) shares one tokenizer and appears once. Metric definitions are in Section 3; lower is better for fertility and NSL, higher is better for CPT, CR, and STRR. The corpus is the same 100-article Wikipedia sample used throughout the paper, so values are directly comparable to the multilingual entries in Appendix A. CPT and CR differ only in whether the denominator counts characters or UTF-8 bytes, and on this English-only corpus the two are essentially identical because most characters are single-byte.

Tokenizer	Vocab Size	Fertility	CPT	CR	STRR	NSL
GPT-2	50,257	1.63	5.05	5.05	0.61	0.198
Llama-2-7B	32,000	1.52	4.27	4.28	0.70	0.234
Llama-3.1-8B	128,256	1.55	5.14	5.15	0.64	0.194
Qwen2.5	151,665	1.57	5.09	5.10	0.64	0.196
Gemma-2-9B	256,000	1.42	5.18	5.19	0.73	0.193
Mistral-7B-v0.3	32,768	1.48	4.41	4.42	0.73	0.227
Mistral-Nemo-12B	131,072	1.67	4.97	4.97	0.59	0.201
Phi-4	100,352	1.55	5.14	5.15	0.64	0.195
Yi-1.5-9B	63,992	1.67	4.72	4.72	0.60	0.212
Falcon-7B	65,024	1.61	4.86	4.86	0.62	0.206
BLOOM-7B	250,680	1.62	5.01	5.01	0.59	0.200
Command-R-35B	255,029	1.48	5.14	5.14	0.69	0.195
StableLM-2-12B	100,289	1.57	5.09	5.10	0.64	0.196
GLM-4-9B	151,329	1.55	5.14	5.15	0.64	0.194
OLMo-2-7B	100,278	1.55	5.14	5.15	0.64	0.195
InternLM2.5-7B	92,544	1.67	4.88	4.89	0.58	0.205
SmolLM2-1.7B	49,152	1.57	5.01	5.02	0.64	0.199
DeepSeek-V3	128,815	1.58	5.15	5.16	0.62	0.194
Qwen3-8B	151,669	1.57	5.09	5.10	0.64	0.196

Table 4: TokLens metrics on English Wikipedia.

C Benchmark Scores

Scores are sourced from the official Open LLM Leaderboard v2 (Fourrier et al., 2024); Avg is the unweighted mean. Qwen3-8B and DeepSeek-V3 lack leaderboard entries and are used for metric-only analysis.

Model	Params	IFEval	BBH	MATH	GPQA	MUSR	MMLU-PRO	Avg
GPT-2	0.1B	17.8	2.8	0.5	1.1	13.9	1.8	6.3
Qwen2.5-0.5B	0.5B	16.3	7.0	3.9	0.0	2.1	10.1	6.6
Qwen2.5-1.5B	1.5B	26.7	16.7	9.1	4.7	5.3	20.6	13.9
SmolLM2-1.7B	1.7B	24.4	9.3	2.6	3.9	4.6	12.6	9.6
Qwen2.5-3B	3.1B	26.9	24.3	14.8	6.4	11.8	24.5	18.1
Phi-3-mini	3.8B	54.8	36.6	16.4	11.0	13.1	33.6	27.6
Llama-2-7B	6.7B	25.2	10.4	1.7	2.2	3.8	9.6	8.8
Falcon-7B	7.0B	18.2	6.0	1.0	0.0	4.5	1.4	5.2
BLOOM-7B	7.1B	13.2	4.0	0.5	1.9	1.9	1.2	3.8
Mistral-7B-v0.3	7.2B	22.7	24.0	3.0	5.6	8.4	21.7	14.2
OLMo-2-7B	7.3B	72.4	16.3	14.9	3.8	4.7	18.6	21.8
Qwen2.5-7B	7.6B	33.7	35.8	25.1	10.0	14.1	37.4	26.0
InternLM2.5-7B	7.7B	55.4	57.0	25.3	13.0	16.3	30.9	33.0
Llama-3.1-8B	8.0B	12.5	25.3	6.6	8.1	8.7	25.4	14.4
Yi-1.5-9B	8.8B	29.4	30.5	11.4	17.2	12.0	32.4	22.2
Gemma-2-9B	9.0B	20.4	34.1	13.4	10.5	14.3	34.5	21.2
GLM-4-9B	9.0B	14.3	35.8	0.0	8.8	14.2	34.9	18.0
Mistral-Nemo-12B	11.6B	16.3	29.4	6.0	5.8	6.5	27.5	15.2
StableLM-2-12B	12.1B	15.7	22.7	4.3	3.8	14.5	23.0	14.0
Qwen2.5-14B	14.8B	36.9	45.1	29.0	17.6	15.9	47.2	32.0
Phi-4	14.7B	5.9	52.4	31.6	20.8	23.8	47.6	30.4
Command-R-35B	35.0B	67.5	34.6	3.5	7.6	16.1	26.3	25.9

Table 5: Open LLM Leaderboard v2 benchmark scores, sorted by parameter count.

D LME Results

We fit four nested linear mixed-effects specifications (Table 6). M1 is the base model: score $\sim z(\text{metric}) + \log(\text{params}) + (1 \mid \text{model})$. M2 adds $\log(\text{train\_tokens})$ as a fixed effect. M3 repeats M2 after dropping the imputed Mistral row (66 obs, 6 models). M4 adds $\log 1p$ of the Qtok per-(model, language) vocabulary allocation % as a tokenizer-training-data proxy (77 obs, 7 models). All fits use REML.

STRR, CPT, NSL, and parity remain highly significant ($p < 0.001$) across every specification. $\log(\text{train\_tokens})$ itself is non-significant in M2/M3 ($p \geq 0.23$), reflecting the narrow training-token range across the 7 models. The Qtok proxy is highly significant in M4 ($p < 0.001$ for every metric); it absorbs the variance from script-allocation differences, unmasking fertility ($p = 0.0005$) while collapsing CR's marginal signal ($p = 0.41$).

	M1: + log(params)			M2: + log(train_tokens)			M3: M2, no imputed			M4: + log1p(qtok_alloc)		
Metric	β	z	p	β	z	p	β	z	p	β	z	p
STRR	+5.68	18.5	<0.001	+5.68	18.5	<0.001	+5.54	16.0	<0.001	+5.16	13.9	<0.001
CPT	+5.03	11.0	<0.001	+5.03	11.0	<0.001	+4.76	9.6	<0.001	+4.06	9.4	<0.001
NSL	−4.96	−9.5	<0.001	−4.95	−9.5	<0.001	−4.57	−8.1	<0.001	−3.96	−8.7	<0.001
Parity	−4.29	−7.4	<0.001	−4.28	−7.4	<0.001	−4.02	−6.4	<0.001	−3.44	−7.4	<0.001
CR	+2.12	2.4	0.015	+2.09	2.4	0.017	+1.28	1.5	0.146	+0.61	0.8	0.411
Fertility	−1.34	−1.8	0.065	−1.33	−1.8	0.067	−1.14	−1.5	0.140	−1.89	−3.5	<0.001

Table 6: Linear mixed-effects model results across four specifications.

Leave-one-language-out cross-validation

To check that the multilingual STRR effect is not driven by any single language, we re-fit M1 eleven times, each time holding out all rows for one of the 11 MMLU-ProX languages. Table 7 reports β, z, and p for the held-out fit. Each fit uses 70 observations (7 models $\times$ 10 retained languages). STRR remains highly significant in every fit, with $\beta \in [4.89, 6.00]$. The strongest STRR effect appears when Japanese is held out ($\beta = 6.00$, $z = 18.59$) and the weakest when Hindi is held out ($\beta = 4.89$); English itself sits at $\beta = 4.95$ when removed, only slightly above the lower bound. The narrow range and uniformly small p-values indicate the multilingual STRR association is broad-based rather than driven by a particular script, supporting the conclusion that the per-language tokenizer signal generalises beyond any single language.

Held-out lang	n	β	z	p
ar	70	+5.71	17.95	<0.001
de	70	+5.88	17.65	<0.001
en	70	+4.95	15.28	<0.001
es	70	+5.75	16.97	<0.001
fr	70	+5.76	17.01	<0.001
hi	70	+4.89	19.53	<0.001
ja	70	+6.00	18.59	<0.001
ko	70	+5.90	19.34	<0.001
pt	70	+5.81	17.24	<0.001
th	70	+5.62	16.88	<0.001
zh	70	+5.90	17.84	<0.001

Table 7: Leave-one-language-out cross-validation of M1 on STRR.

E Full Correlation Results

Table 8 lists the strongest English-side Spearman correlations (sorted by $|\rho|$) between TokLens metrics and Open LLM Leaderboard v2 benchmarks across the 22 models with leaderboard entries. The Bonferroni threshold for the 77 metric-benchmark pairs is $\alpha_{\text{corrected}} = 0.00065$; no correlation crosses it, and partial correlations controlling for $\log(\text{params})$ are uniformly weaker. The strongest raw correlation is fertility (avg) versus MMLU-PRO ($\rho = -0.379$, $p = 0.082$); after partialling out $\log(\text{params})$ it drops to $r = -0.257$ ($p = 0.260$). "avg" rows aggregate across 15 languages and "en" rows use only the English corpus; STRR shows the largest gap between the two ($\rho = 0.16$ on English alone vs. stronger multilingual associations in the LME), reflecting the limited English-only spread across these 22 models.

Metric	Benchmark	Spearman ρ	p-value	Partial r	Partial p
Fertility (avg)	MMLU-PRO	−0.379	0.0820	−0.257	0.2600
CPT (en)	MMLU-PRO	0.355	0.1053	0.173	0.4545
CR (en)	MMLU-PRO	0.355	0.1053	0.173	0.4545
NSL (en)	MMLU-PRO	−0.355	0.1053	−0.173	0.4545
Fertility (avg)	MATH Lvl 5	−0.350	0.1108	−0.282	0.2152
CPT (en)	MUSR	0.345	0.1159	0.162	0.4843
NSL (avg)	MMLU-PRO	−0.328	0.1364	−0.100	0.6659
Parity (avg)	MMLU-PRO	−0.324	0.1407	−0.160	0.4881
Fertility (avg)	BBH	−0.305	0.1682	−0.156	0.4995
Parity (avg)	BBH	−0.278	0.2096	−0.100	0.6674
CPT (avg)	MMLU-PRO	0.265	0.2327	−0.025	0.9139
NSL (avg)	BBH	−0.263	0.2379	−0.013	0.9566
Fertility (avg)	GPQA	−0.257	0.2478	−0.107	0.6452
Fertility (en)	MMLU-PRO	−0.247	0.2675	−0.195	0.3977
Fertility (avg)	Average	−0.241	0.2801	−0.090	0.6983
STRR (en)	MMLU-PRO	0.156	0.4871	0.164	0.4770
STRR (en)	Average	0.098	0.6657	0.083	0.7194

Table 8: Selected Spearman correlations between TokLens metrics and benchmarks. "avg" = average across 15 languages; "en" = English only.

F Held-out English BPB

We compute bits-per-byte (BPB) on the wikitext-2-raw-v1 test split (first 60K bytes of held-out English text) for the 7 LME models. For each model we tokenize the full text with the model's own tokenizer (BOS prepended), run a single forward pass per 4096-token chunk, accumulate $-\log_2 P(\text{token}_i \mid \text{context})$, and normalize by the byte count of the original text:

$$\text{BPB} = \frac{1}{|b|} \sum_{i=1}^{|t|} -\log_2 P(\text{token}_i \mid \text{context}).$$

Because the denominator is the original byte count, BPB is comparable across tokenizers.

Model	Params	Tokens	BPB
Qwen2.5-3B	3.1B	14,862	0.6755
Mistral-7B-v0.3	7.2B	16,287	0.6095
Llama-3.1-8B	8.0B	14,113	0.5719
Qwen2.5-7B	7.6B	14,862	0.5833
Gemma-2-9B†	9.0B	14,428	1.3395
Qwen2.5-14B	14.8B	14,862	0.3988
Phi-4	14.7B	14,143	0.5738

Table 9: Held-out English BPB on wikitext-2-test (first 60,190 bytes). †Gemma-2-9B is treated as an outlier; see body.

The Qwen2.5 series scales cleanly (0.68 $\rightarrow$ 0.58 $\rightarrow$ 0.40 across 3B / 7B / 14B), and Llama-3.1, Mistral, and Phi-4 are in the 0.57–0.61 range expected for 7–15B models. Gemma-2-9B is an outlier at 1.34, attributable to a known `bfloat16` softcap issue rather than tokenizer quality; dropping it gives BPB vs. $\log(\text{params})$ $\rho = -0.94$, $p = 0.005$. Spearman against English TokLens metrics remains non-significant either way ($|\rho| \leq 0.52$, $p \geq 0.29$).

G Comparison with Qtok

We compare TokLens against Qtok (Chelombitko et al., 2024), the closest prior multilingual tokenizer toolkit. Qtok analyzes the static *vocabulary* (token-category distribution, Unicode-block coverage, per-language allocation of Latin/Cyrillic alphabetic tokens), whereas TokLens runs the tokenizer over a *corpus* and reports per-language fertility, STRR, parity, NSL, CPT, and CR; the two are complementary.

Setup. We invoked Qtok on `tokenizer.json` for the 16 of our 24 tokenizers with publicly downloadable files (gated or non-fast tokenizers skipped). For each Qtok per-language allocation % (Latin or Cyrillic), we compute Spearman correlation against the corresponding TokLens metric across the 16 shared tokenizers.

Aspect	Qtok	TokLens
Input	`tokenizer.json`	HF tokenizer + corpus
Level	vocabulary-level	corpus-level
Token category dist.	12 classes	—
Unicode-block dist.	yes	—
Per-lang allocation	Latin + Cyrillic	—
Fertility / STRR / NSL	—	yes
Parity / CR / CPT	—	yes
Languages	Latin + Cyrillic	15 (6 scripts)
Statistics	descriptive	Spearman + LME
Downstream linkage	no	MMLU-ProX, OLM-v2

Table 10: Feature matrix comparing Qtok and TokLens.

Qtok metric	TokLens metric	ρ	p
Latin allocation % vs. TokLens STRR			
fr (%)	STRR (fr)	+0.78	<0.001
pt (%)	STRR (pt)	+0.74	0.001
tr (%)	STRR (tr)	+0.75	0.001
es (%)	STRR (es)	+0.46	0.076
de (%)	STRR (de)	+0.02	0.95
en (%)	STRR (en)	−0.02	0.95
Cyrillic allocation % vs. TokLens (ru)			
ru (%)	STRR (ru)	+0.68	0.004
ru (%)	Parity (ru)	−0.84	<0.001
ru (%)	Fertility (ru)	−0.83	<0.001
Other proxies			
en (%)	Fertility (en)	+0.18	0.50
spaced_alpha (%)	STRR (en)	−0.02	0.94

Table 11: Spearman ranking correlation between Qtok per-language vocabulary allocation and TokLens corpus-level metrics, on 16 tokenizers with downloadable `tokenizer.json`.

Findings. (1) On low-resource Latin scripts (fr, pt, tr) and on Cyrillic, Qtok's vocabulary allocation % is a strong proxy for TokLens's corpus-level STRR / parity / fertility ($|\rho| \in [0.68, 0.84]$, $p \leq 0.004$), confirming that the two toolkits agree where they have signal in common. (2) On English and German, the two toolkits decouple: nearly all tokenizers in our set allocate $\geq 9\%$ to English and $\geq 9\%$ to German, so Qtok's allocation is saturated, while TokLens's STRR still varies meaningfully across tokenizers due to factors Qtok does not measure (BPE merge order, suffix handling, byte-level fallback). (3) TokLens adds the parity / fertility / STRR axis on a corpus, downstream-benchmark linkage, and mixed-effects modeling; Qtok adds the static-vocabulary axis (token category distribution, Unicode coverage, error-token counts) that TokLens does not compute. Practitioners auditing a tokenizer for cross-lingual fairness should run both: Qtok for vocabulary structure, TokLens for corpus behavior and downstream signal.

H Toolkit Usage

TokLens is a Python library installable via `pip install toklens`. It requires Python 3.10+ and depends on the HuggingFace `tokenizers` library and `numpy`. The following code demonstrates basic usage:

```
from toklens import Analyzer

analyzer = Analyzer.from_pretrained(
    "meta-llama/Llama-3.1-8B"
)
report = analyzer.evaluate(
    langs=["en", "zh", "ja", "ar"]
)
report.print_table()
```

The toolkit computes all six metrics (fertility, CPT, compression ratio, STRR, NSL, parity) in a single pass. For the parity metric, a reference language (default: English) is required. Corpora are loaded from Wikipedia via HuggingFace `datasets` and cached locally after first download. The command-line interface supports quick evaluation:

```
toklens eval meta-llama/Llama-3.1-8B \
    --langs en zh ja ar
toklens compare \
    meta-llama/Llama-2-7b-hf \
    meta-llama/Llama-3.1-8B
```

I Per-language Fertility Heatmap

Figure 8 shows the per-(tokenizer, language) values of fertility (tokens per whitespace-delimited word) across the full 24-tokenizer × 15-language grid. Fertility varies enormously for CJK and Thai, ranging from byte-level fallback (GPT-2 on Japanese reaches 91.1 tokens/word) to dedicated multilingual vocabularies (Gemma-2 reaches 3.5). Latin-script languages cluster between 1.2 and 1.7 across all tokenizers.

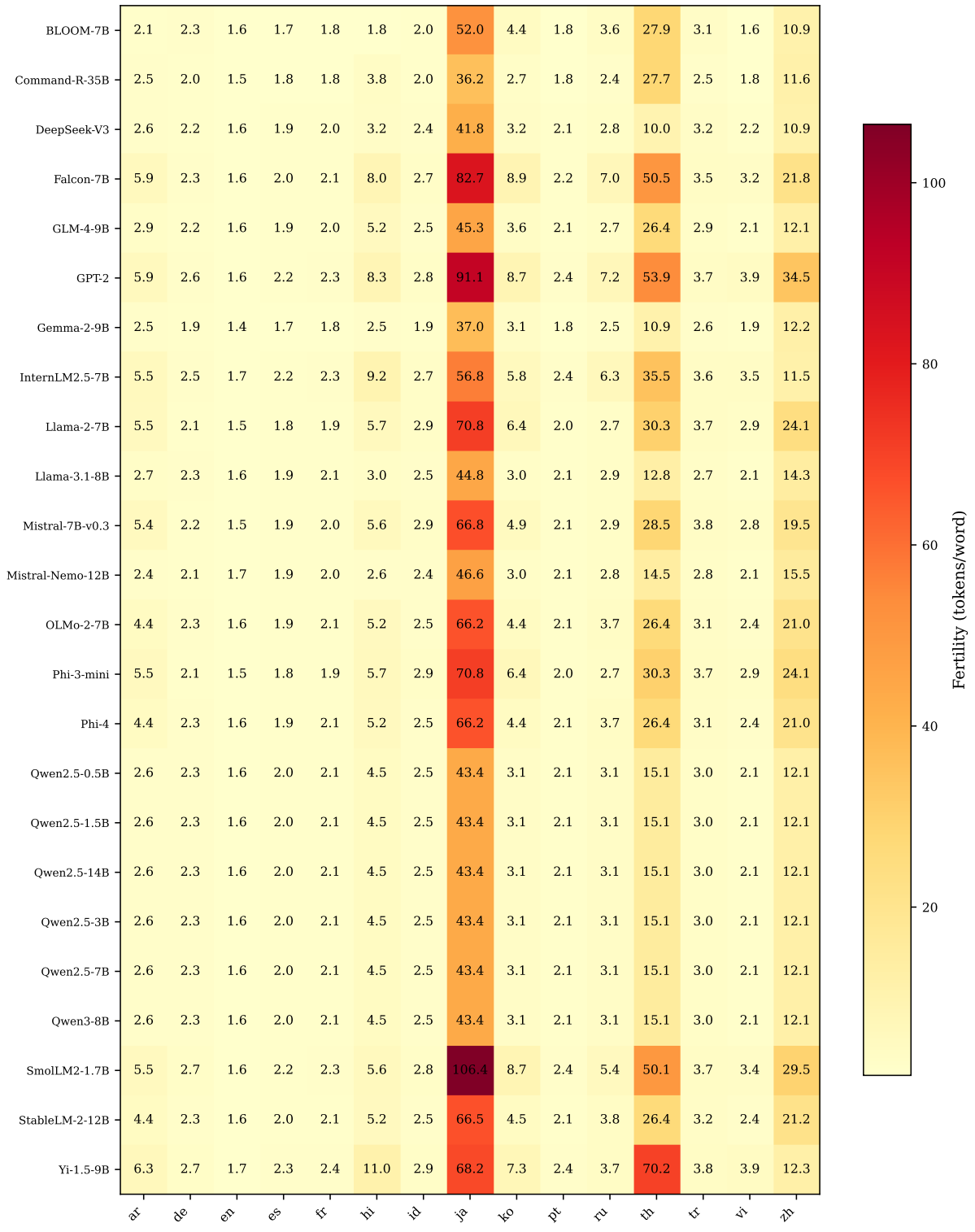

Figure 8: Fertility (tokens per word) across tokenizers and languages.

J Per-language STRR Heatmap

Figure 9 shows the per-(tokenizer, language) values of single-token retention rate. STRR is high (0.4–0.7) for English and Romance languages but stays below 0.2 for Arabic, Hindi, and Thai across all tokenizers, reflecting both morphological complexity and vocabulary allocation bias. Vietnamese, despite using Latin script, also shows depressed STRR due to diacritic-induced fragmentation.

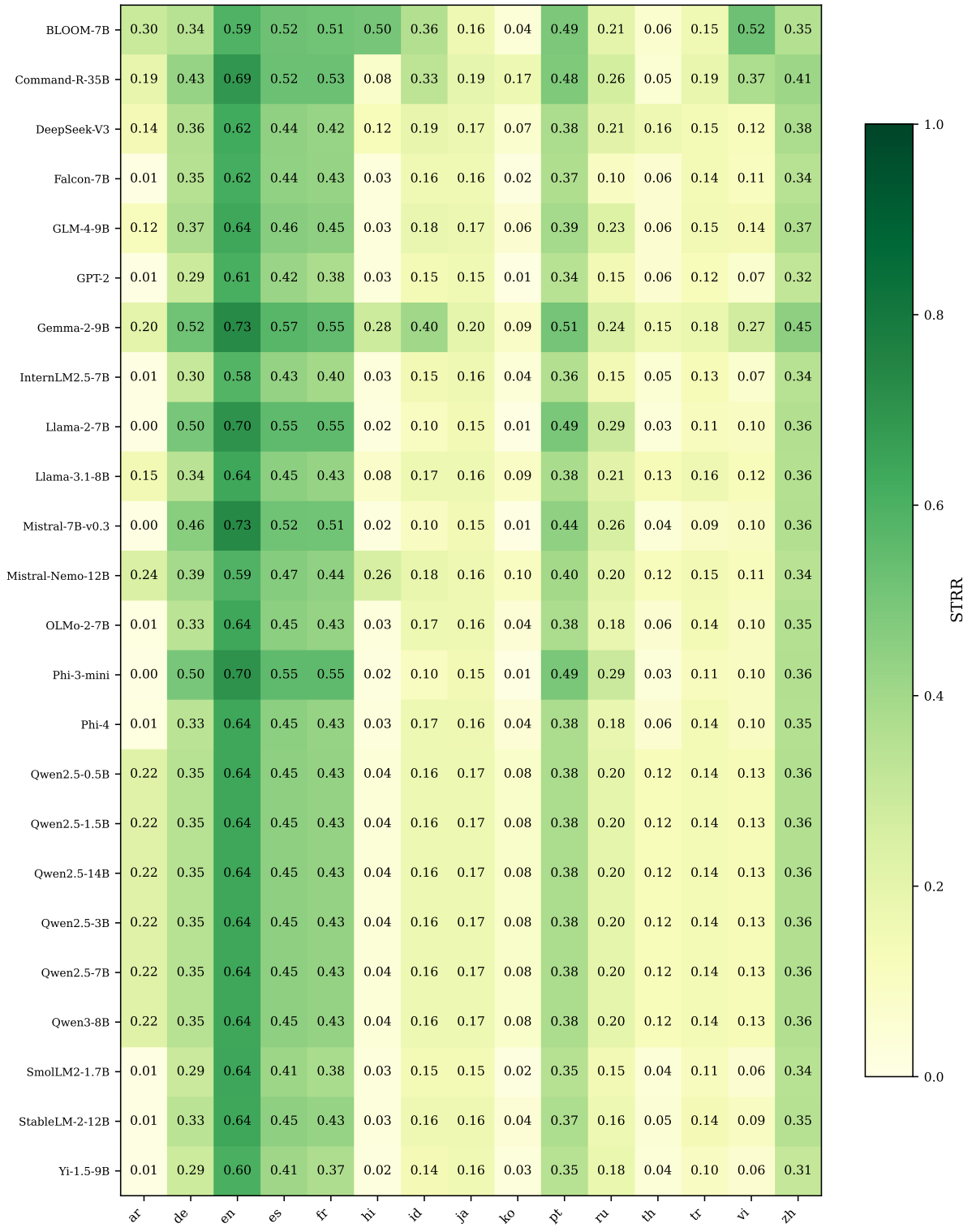

Figure 9: Single-Token Retention Rate (STRR) across tokenizers and languages.

K Per-language CPT Heatmap

Figure 10 shows the per-(tokenizer, language) values of characters per token. CPT is highest for English (typically 4–7) across all tokenizers; Chinese and Thai have the lowest values for tokenizers without dedicated CJK vocabulary, with BLOOM, Qwen2.5, and Gemma-2 the notable exceptions that achieve CPT > 2 on Chinese.

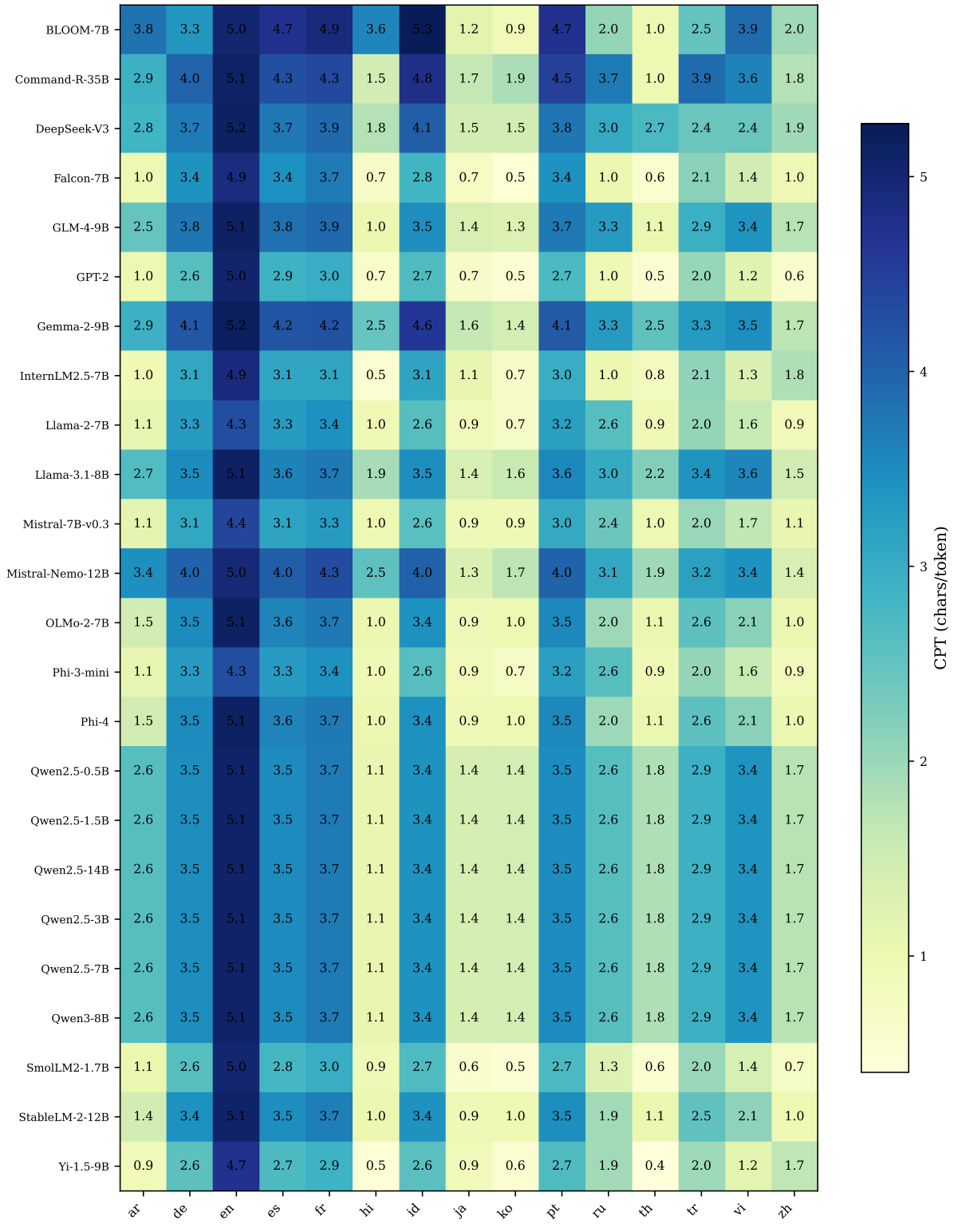

Figure 10: Characters per token across tokenizers and languages.