The Token Tax: Systematic Bias in Multilingual Tokenization

Jessica M. Lundin[1], Ada Zhang[2], Nihal Karim[2], Hamza Louzan[2], Victor Wei[2], David Ifeoluwa Adelani[3], Cody Carroll[2,4]

[1]Institute for Disease Modeling, Gates Foundation,
[2]Data Institute, University of San Francisco,
[3]Mila - Quebec AI Institute, McGill University, and Canada CIFAR AI Chair,
[4]Department of Mathematics and Statistics, University of San Francisco

Correspondence: jessica.lundin@gatesfoundation.org

Abstract

Tokenization inefficiency is associated with structural disadvantages for morphologically complex, low-resource languages, inflating compute resources and reducing accuracy. We evaluate 10 large language models (LLMs) on AfriMMLU (5 subjects; 16 African languages) and show that token fertility reliably predicts accuracy. Higher fertility consistently predicts lower accuracy across all models and subjects. We further find that reasoning models (e.g., DeepSeek, o1) consistently outperform non-reasoning peers across high- and low-resource languages in the AfriMMLU dataset, narrowing accuracy gaps observed in prior generations. In terms of economics, a doubling in tokens results in quadrupled training cost and time, underscoring the "token tax" faced by many languages. These results motivate morphologically aware tokenization, fair pricing, and multilingual benchmarks for equitable natural language processing (NLP).

1 Introduction

Tokenization serves as the foundational layer of modern NLP systems, yet it is associated with systematic inequalities that disproportionately affect morphologically complex and low-resource languages. Prior work decisively establishes tokenization as a source of computational and economic inequality (Ahia et al., 2023), with quantified impacts ranging from inflated token counts to substantial BLEU point performance degradation (Petrov et al., 2023; Ali et al., 2024). Because transformer attention scales quadratically with sequence length, even modest increases in token counts can meaningfully raise compute requirements and reduce effective context capacity (Keles et al., 2022). As a result, morphologically complex languages with high fertility values suffer from compound disadvantages that are difficult to fully overcome within widely deployed transformer architectures (Sreedhar et al., 2023). Recent work on adaptive and flexible tokenization strategies (e.g., MAGNET; FlexiTokens) suggests promising directions for mitigating these inefficiencies, but these approaches are not yet standard in large-scale multilingual LLM training and remain mostly unevaluated on low-resource African languages (Ahia et al., 2024; Owodunni et al., 2025).

These disparities particularly affect morphologically rich languages, where agglutinative and fusional morphology leads to systematic tokenization inefficiency. The technical disparities translate directly into economic exclusion through what we term the "token tax", prohibitive training and inference costs measured in dollars and tons of CO_2, and systematic underrepresentation in model capabilities that affects billions of speakers worldwide.

A reasonable cost to train a small-medium model or a large frontier model is easily \$1M (1 month) to \$100M (∼3 months) with primarily English tokens (see Appendix C for derivation of estimates from publicly reported petaFLOP-day requirements and standard cloud compute pricing). If we instead train on a language with 2× or 5× more tokens for the same content, the transformer's quadratic $O(n^2)$ compute costs result in a 4× or 25× increase in energy consumption, dollar cost, training time, and CO_2 emissions relative to English. In this example, the cost becomes \$4-25M (4 months-2 years) or \$400M-2.5B (1-6 years), respectively.

Our contributions herein are as follows:

- We extend prior fertility and accuracy analysis to 10 models and 16 languages, confirming fertility as a reliable predictor of MCQA accuracy.
- We conduct the first large-scale comparison of tokenization effects for reasoning vs. non-reasoning LLMs on AfriMMLU, showing that reasoning capabilities substantially reduce but do not eliminate tokenization bias.

- We quantify the economic impact of tokenization inefficiency, demonstrating how the "token tax" creates barriers to multilingual NLP development.
- We release public datasets containing: (i) model results from AfriMMLU benchmark including reasoning models, (ii) comprehensive tokenization metrics.

2 Related Work

The impact of tokenization on multilingual model performance has received increasing attention. Petrov et al. (2023) demonstrated that tokenizers disadvantage non-English languages, with text length exceeding 15 times for some language pairs. Rust et al. (2021) showed that tokenizer quality significantly impacts downstream performance, finding that morphologically complex languages suffer from both vocabulary underrepresentation and suboptimal segmentation strategies. Ali et al. (2024) conducted extensive ablation experiments showing that tokenizer choice can impact downstream performance and increase training costs by 68%. Their analysis is for European languages that have different fertility distributions than African languages, the focus of this work.

Joshi et al. (2020) categorized languages into six resource levels, showing that 88% of world languages fall into the lowest category. These systemic inequalities correlate strongly with tokenization efficiency, suggesting that addressing tokenization could substantially improve language technology equity.

Recent benchmarks have begun addressing the evaluation gap for African languages (Adelani et al., 2025; Singh et al., 2025; Alhanai et al., 2024; Adebara et al., 2025; Beyene et al., 2025; Team et al., 2022). While these benchmarks evaluate various aspects of multilingual performance, from culturally relevant knowledge questions to speech tasks, they primarily focus on documenting performance gaps without systematically analyzing the underlying tokenization disparities. Our work complements these efforts by explicitly connecting tokenization efficiency to performance degradation, revealing that much of the observed performance gap can be attributed to systematic tokenization bias rather than model capability limitations.

While the computational complexity of transformers is well-established (Vaswani et al., 2017), the economic implications of tokenization inefficiency remain underexplored. Our work bridges this gap by quantifying both performance degradation and financial costs, demonstrating how technical choices create economic barriers to equitable NLP development.

3 Experimental Setup

3.1 Dataset and Languages

AfriMMLU (Adelani et al., 2025) comprises 9,000 multiple-choice questions across 5 subjects: elementary mathematics, global facts, high school geography, high school macroeconomics, and international law. The benchmark covers 16 African languages spanning multiple language families including Niger-Congo (e.g., Swahili, Yoruba, Zulu), Afro-Asiatic (e.g., Amharic, Hausa), and Nilo-Saharan, providing substantial typological diversity. Each subject contains between 1,500 and 2,000 questions, professionally translated by native speakers with quality verification.

3.2 Model Selection

We evaluate 10 models spanning three categories to enable comparison, including general LLMs (Llama 3.1 405B, Gemini 1.5 Pro, Claude Sonnet 3.5, DeepSeek V3, GPT-4o, Qwen 2.5 32B) and multilingual-focused ones (Aya 23 35B, Pixtral 12B). We also include reasoning models (DeepSeek R1, OpenAI o1) to test whether enhanced reasoning capabilities mitigate tokenization-related performance decline. If reasoning models show reduced gaps despite similar tokenization inefficiency, this would suggest that architectural advances can partially mitigate tokenizer limitations.

3.3 Metrics and Analysis

Our primary metric is Fertility ($F = T \ / \ W$), the average number of tokens per word, quantifying tokenization efficiency. Lower fertility indicates more efficient tokenization for a given language. Our analysis also includes zero-shot accuracy of MCQA tasks using the prompt template detailed in the Appendix.

Performance gaps are calculated as the difference between English and French accuracy and the mean accuracy across all 16 African languages for each model-subject combination. English and French here provide baselines near optimal tokenization efficiency. The "Random" baseline of 25% represents the probability of guessing a cor-

rect response by chance for multiple choice questions with 4 options.

3.4 Statistical Methodology

The pipeline for analysis is as follows:

for each language ℓ in AfriMMLU **do**
  for each model m **do**
    Tokenize all questions using m's tokenizer
    Calculate fertility $F_{\ell,m} = \frac{\text{tokens}}{\text{words}}$
    Evaluate zero-shot accuracy on all subjects
    Store results for regression analysis
  end for
end for
Fit linear models: Accuracy $\sim$ Fertility for each model-subject pair
Calculate regression statistics: slopes, standard errors, t-values, p-values
Calculate Pearson correlation coefficients (ρ), R^2, and adjusted R^2
Apply Benjamini-Hochberg FDR correction for multiple comparisons

We conduct linear regression analysis for each model-subject combination, treating fertility as the predictor and accuracy as the outcome variable. For each regression, we report the intercept, slope (effect of fertility on accuracy), standard error, t-statistic, and p-value. We calculate Pearson correlation coefficients (ρ) to quantify the strength and direction of the linear relationship, and coefficients of determination (R^2) to measure the proportion of variance in accuracy explained by fertility. Adjusted R^2 values correct for sample size and number of predictors. Statistical significance is assessed at $p < 0.05$ with Benjamini-Hochberg False Discovery Rate (FDR) correction to account for multiple comparisons across 50 model-subject pairs. We consider $R^2 \geq 0.25$ as large effects and $|\rho| \geq 0.50$ as strong correlations following standard conventions.

4 Results

4.1 Performance Gaps Across Languages

Figure 1 shows substantial performance disparities across languages. African languages trail English by 30 percentage points and French by 24 percentage points on average across model results. This gap varies by subject, with Geography and Economics showing the largest disparities and Elementary Mathematics and Global Facts showing relatively smaller gaps.

Reasoning models DeepSeek R1 and o1 reduce performance gaps across subjects. These models outperform non-reasoning counterparts while maintaining strong English performance, suggesting that enhanced reasoning capabilities provide particular benefits for low-resource settings.

4.2 Tokenization as Performance Predictor

Figure 2 demonstrates the strong negative relationship between fertility and accuracy for Llama 3.1 405B as an example model (results for all models are displayed in Figure 3). Across all 10 models and five subjects, higher fertility consistently predicts lower accuracy. Regression analyses quantify this relationship with slopes ranging from -0.08 to -0.18, meaning each additional token per word reduces accuracy by 8 to 18 percentage points on average.

We note that fertility is not independent of language resource level, that is languages underrepresented in training corpora tend to have higher fertility due to fewer subword merges during tokenizer training. The observed fertility accuracy relationship is likely reflecting both the direct tokenization effects and indirect effects mediated by training data availability.

Table 1 reports detailed regression results. Several effects are both large and statistically significant after FDR correction, including Llama-3.1-405B on Microeconomics (slope $= -0.185$, $p = 0.002$) and Qwen-2.5-32B on Geography (slope $= -0.155$, $p = 0.006$). Between 20 to 50% of variance in accuracy is explained by variation in fertility, with particularly strong effects in technical subjects requiring precise terminology.

5 Analysis

5.1 Economic Impact of Token Inflation

The tokenization inefficiencies documented above translate directly to economic barriers. Because transformer training scales quadratically with sequence length, a 2× increase in fertility produces a 4× increase in training time and cost. Table 2 quantifies these impacts for the Llama model family.

Inference costs show similar patterns. Generating 1M English-equivalent tokens costs $5-20 with GPT-4o, but $10-40 for a language with twice the token fertility. Latency doubles correspondingly, creating user experience degradation alongside cost inflation.

Model	Total	Math	Facts	Geog	Econ	Law
Baseline Performance (English Language)						
o1-preview-2024-09-12	91	99	75	91	97	91
DeepSeek-R1	90	100	65	95	99	91
gemini-15-Pro-002	88	93	69	91	96	92
gpt-4o-2024-08-06	89	95	68	92	97	91
DeepSeek-V3-0324	88	96	67	92	98	89
Llama-3.1-405B	86	85	66	92	97	89
claude-3-5-sonnet-202410	75	63	66	87	81	78
Qwen2.5-32B	79	78	55	86	88	89
phi-4	77	66	45	92	95	89
Pixtral-12B-2409	62	41	43	78	73	76
aya-23-35B	57	42	43	67	69	66
Random	25	25	25	25	25	25
Average Performance (all African Languages)						
o1-preview-2024-09-12	76	88	67	72	75	76
DeepSeek-R1	67	88	56	60	63	68
gemini-15-Pro-002	62	78	55	56	55	68
gpt-4o-2024-08-06	62	80	48	58	58	67
DeepSeek-V3-0324	54	76	49	47	45	54
Llama-3.1-405B	51	69	45	42	41	59
claude-3-5-sonnet-202410	50	59	46	48	43	52
Qwen2.5-32B	39	53	34	27	31	50
phi-4	37	44	36	29	29	48
Pixtral-12B-2409	33	35	31	30	31	39
aya-23-35B	24	26	20	19	24	28
Random	25	25	25	25	25	25
Performance Gap (English - African Languages)						
o1-preview-2024-09-12	15	11	8	19	22	15
DeepSeek-R1	23	12	9	35	36	23
gemini-15-Pro-002	26	15	14	35	42	24
gpt-4o-2024-08-06	26	15	20	34	39	24
DeepSeek-V3-0324	34	20	18	45	53	35
Llama-3.1-405B	35	16	21	50	57	30
claude-3-5-sonnet-202410	25	4	20	39	39	27
Qwen2.5-32B	40	25	22	59	57	39
phi-4	40	22	9	63	66	41
Pixtral-12B-2409	29	6	12	48	42	37
aya-23-35B	34	16	23	48	45	38

(a) Accuracy Aggregation (English)

Model	Total	Math	Facts	Geog	Econ	Law
Baseline Performance (French Language)						
o1-preview-2024-09-12	89	93	75	90	95	91
DeepSeek-R1	87	96	68	85	97	91
gemini-15-Pro-002	85	86	65	89	94	91
gpt-4o-2024-08-06	83	88	60	86	90	91
DeepSeek-V3-0324	83	89	63	86	90	88
Llama-3.1-405B	81	82	63	85	90	84
claude-3-5-sonnet-202410	71	70	53	87	80	66
Qwen2.5-32B	73	70	54	78	80	82
phi-4	73	65	47	83	82	86
Pixtral-12B-2409	55	37	41	66	58	71
aya-23-35B	44	35	28	41	51	65
Random	25	25	25	25	25	25
Average Performance (all African Languages)						
o1-preview-2024-09-12	76	88	67	72	75	76
DeepSeek-R1	67	88	56	60	63	68
gemini-15-Pro-002	62	78	55	56	55	68
gpt-4o-2024-08-06	62	80	48	58	58	67
DeepSeek-V3-0324	54	76	49	47	45	54
Llama-3.1-405B	51	69	45	42	41	59
claude-3-5-sonnet-202410	50	59	46	48	43	52
Qwen2.5-32B	39	53	34	27	31	50
phi-4	37	44	36	29	29	48
Pixtral-12B-2409	33	35	31	30	31	39
aya-23-35B	24	26	20	19	24	28
Random	25	25	25	25	25	25
Performance Gap (French - African Languages)						
o1-preview-2024-09-12	13	5	8	18	20	15
DeepSeek-R1	20	8	12	25	34	23
gemini-15-Pro-002	23	8	10	33	40	23
gpt-4o-2024-08-06	21	8	12	28	32	24
DeepSeek-V3-0324	29	13	14	39	45	34
Llama-3.1-405B	30	13	18	43	50	25
claude-3-5-sonnet-202410	22	11	7	39	38	15
Qwen2.5-32B	34	17	21	51	49	32
phi-4	36	21	11	54	53	38
Pixtral-12B-2409	22	2	10	36	27	32
aya-23-35B	20	9	6	22	27	37

(b) Accuracy Aggregation (French)

Figure 1: Baseline performance shows English (a) and French (b) accuracy (in percentage points). The mean accuracy across all 16 African languages is shown in the middle charts of (a) and (b). The bottom charts of (a) and (b) show performance gaps between the African languages and higher-resource languages, though reasoning-oriented models narrow this gap.

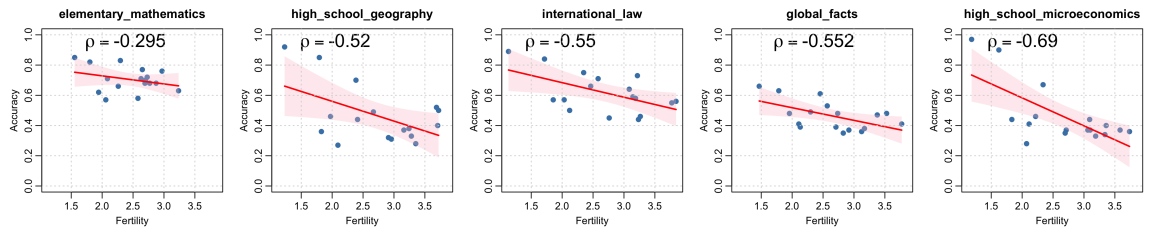

Figure 2: Fertility and accuracy for Llama 3.1 405B across subjects. Strong negative correlations (ρ) demonstrate systematic performance degradation with tokenization inefficiency. Fertility is not independent of training data representation; fertility captures tokenization inefficiency that covaries with performance, but does not isolate causal effects independent of pretraining data availability or quality.

Subject	Model	Intercept	Slope	Std. Error	t-value	P-value	ρ	R^2	Adj. R^2
Elementary Math	Sonnet3.5	0.652	-0.018	0.029	-0.609	0.552	-0.155	0.024	-0.041
	Aya23 35B	0.480	-0.079	0.019	-4.058	**0.001***	**-0.723**	**0.523**	**0.492**
	DeepSeek R1	1.045	-0.066	0.045	-1.475	0.161	-0.356	0.127	0.068
	DeepSeek V3	0.884	-0.044	0.048	-0.922	0.371	-0.232	0.054	-0.009
	Gemini 1.5 Pro	0.907	-0.045	0.042	-1.078	0.298	-0.268	0.072	0.010
	Llama3.1 405B	0.836	-0.054	0.045	-1.195	0.251	-0.295	0.087	0.026
	Phi4	0.773	-0.125	0.034	-3.641	**0.002***	**-0.685**	0.469	0.434
	GPT-4o	1.002	-0.089	0.057	-1.571	0.137	-0.376	0.141	0.084
	Pixtral 12B	0.417	-0.024	0.014	-1.717	0.106	-0.405	0.164	0.109
	Qwen2.5 32B	0.857	-0.113	0.037	-3.012	**0.009***	**-0.614**	0.377	0.335
Global Facts	Sonnet3.5	0.508	-0.011	0.027	-0.390	0.702	-0.100	0.011	-0.056
	Aya23 35B	0.335	-0.044	0.023	-1.930	0.073	-0.446	0.199	0.146
	DeepSeek R1	0.574	-0.002	0.038	-0.061	0.952	-0.016	0.000	-0.066
	DeepSeek V3	0.619	-0.045	0.034	-1.308	0.211	-0.320	0.102	0.042
	Gemini 1.5 Pro	0.585	-0.011	0.051	-0.222	0.827	-0.057	0.003	-0.063
	Llama3.1 405B	0.685	-0.084	0.033	-2.564	**0.022**	**-0.552**	**0.305**	**0.258**
	Phi4	0.408	-0.015	0.017	-0.845	0.411	-0.213	0.045	-0.018
	GPT-4o	0.638	-0.063	0.054	-1.169	0.261	-0.289	0.084	0.022
	Pixtral 12B	0.428	-0.038	0.018	-2.068	0.056	-0.471	0.222	0.170
	Qwen2.5 32B	0.505	-0.052	0.024	-2.171	**0.046**	-0.489	0.239	0.188
High School Geography	Sonnet3.5	0.781	-0.080	0.045	-1.779	0.096	-0.417	0.174	0.119
	Aya23 35B	0.475	-0.097	0.038	-2.512	**0.024**	**-0.544**	0.296	0.249
	DeepSeek R1	0.847	-0.082	0.056	-1.466	0.163	-0.354	0.125	0.067
	DeepSeek V3	0.843	-0.124	0.053	-2.331	**0.034**	**-0.516**	0.266	0.217
	Gemini 1.5 Pro	0.750	-0.065	0.070	-0.937	0.363	-0.235	0.055	-0.008
	Llama3.1 405B	0.822	-0.131	0.055	-2.359	**0.032**	**-0.520**	0.271	0.222
	Phi4	0.808	-0.162	0.048	-3.343	**0.004***	**-0.653**	0.427	0.389
	GPT-4o	0.952	-0.151	0.068	-2.211	**0.043**	-0.496	0.246	0.195
	Pixtral 12B	0.688	-0.121	0.035	-3.414	**0.004***	**-0.661**	**0.437**	**0.400**
	Qwen2.5 32B	0.755	-0.155	0.049	-3.190	**0.006***	**-0.636**	0.404	0.365
High School Microeconomics	Sonnet3.5	0.750	-0.096	0.042	-2.307	**0.036**	**-0.512**	0.262	0.213
	Aya23 35B	0.549	-0.105	0.038	-2.775	**0.014**	**-0.582**	0.339	0.295
	DeepSeek R1	0.888	-0.088	0.074	-1.194	0.251	-0.295	0.087	0.026
	DeepSeek V3	0.906	-0.157	0.049	-3.179	**0.006***	**-0.634**	0.403	0.363
	Gemini 1.5 Pro	0.883	-0.129	0.067	-1.920	0.074	-0.444	0.197	0.144
	Llama3.1 405B	0.953	-0.185	0.050	-3.691	**0.002***	**-0.690**	**0.476**	**0.441**
	Phi4	0.858	-0.184	0.053	-3.479	**0.003***	**-0.668**	0.447	0.410
	GPT-4o	0.942	-0.150	0.084	-1.779	0.096	-0.417	0.174	0.119
	Pixtral 12B	0.622	-0.105	0.033	-3.179	**0.006***	**-0.635**	0.403	0.363
	Qwen2.5 32B	0.779	-0.154	0.048	-3.196	**0.006***	**-0.636**	0.405	0.365
International Law	Sonnet3.5	0.645	-0.040	0.028	-1.426	0.174	-0.346	0.119	0.061
	Aya23 35B	0.578	-0.101	0.042	-2.403	**0.030**	**-0.527**	0.278	0.230
	DeepSeek R1	0.813	-0.043	0.043	-1.010	0.329	-0.252	0.064	0.001
	DeepSeek V3	0.771	-0.073	0.045	-1.617	0.127	-0.385	0.148	0.092
	Gemini 1.5 Pro	0.796	-0.039	0.052	-0.758	0.460	-0.192	0.037	-0.027
	Llama3.1 405B	0.876	-0.096	0.038	-2.548	**0.022**	**-0.550**	0.302	0.256
	Phi4	0.804	-0.101	0.041	-2.452	**0.027**	**-0.535**	0.286	0.238
	GPT-4o	0.889	-0.085	0.072	-1.175	0.258	-0.290	0.084	0.023
	Pixtral 12B	0.686	-0.095	0.033	-2.859	**0.012***	**-0.594**	**0.353**	**0.310**
	Qwen2.5 32B	0.787	-0.092	0.040	-2.297	**0.036**	**-0.510**	0.260	0.211

Table 1: Fertility and Accuracy by Model and Subject. Results from linear models regressing accuracy on fertility across 16 languages for each model-subject combination. The table reports intercepts, slopes (negative values indicate higher fertility correlates with lower accuracy), standard errors, t-statistics, and p-values for each regression. Pearson correlation coefficients (ρ) quantify the strength of the linear relationship. R^2 values show the proportion of variance in accuracy explained by fertility, and adjusted R^2 values correct for sample size. Bold p-values indicate statistical significance ($p < 0.05$). Asterisks (*) indicate results that remain significant after Benjamini-Hochberg FDR correction (FDR < 0.05). Bold ρ values indicate strong correlations ($|\rho| \geq 0.50$), and bold R^2 values indicate large effects ($R^2 \geq 0.25$). Regressions for o1 are not included because OpenAI has not released details on the tokenizer for this model.

These cost projections assume equivalent training objectives across languages. In practice, the relationship between fertility and cost interacts with data availability: high-fertility languages lack training data to realize theoretical training costs, which is an additional and compounded barrier to devel-

opment.

5.2 Linguistic Factors Driving Fertility

The fertility disparities correlate strongly with morphological typology. Agglutinative languages in our dataset (Swahili: $F = 2.8$, Zulu: $F = 2.6$) consistently show higher fertility than more analytic languages. Fusional languages like Amharic occupy intermediate positions. This pattern suggests that current BPE-based tokenizers, predominantly trained on English and European language text, systemically fail to capture morphological processes. Affixation patterns that would constitute single semantic units are split across tokens, fragmenting the meaning and increasing the sequence length.

For example, Swahili verb forms that encode subject, tense, object, and mood in a single word are often split into multiple tokens, while equivalent English constructions using auxiliary verbs and pronouns may use fewer tokens despite containing more words. This fundamental mismatch between tokenizer training data and target language morphology drives the observed inefficiencies.

6 Conclusion

This study demonstrates that tokenization inefficiency predicts systematic disadvantages for morphologically complex languages. Across 10 large language models and 16 African languages, fertility explains up to 50% of the variance in model accuracy. In the strongest cases, each additional token per word is associated with accuracy drops of up to 18 percentage points.

While reasoning models like DeepSeek and o1 narrow accuracy gaps—improving African language performance by 8 to 12 points on average—substantial disparities remain. The economic implications are severe: doubling fertility quadruples training costs, creating a "token tax" that turns linguistic diversity into computational liability.

Addressing these inequities requires coordinated intervention: technical advances in morphologically-aware tokenization, economic reforms to pricing structures, and expanded multilingual evaluation infrastructure. Without such efforts, billions of speakers will remain excluded from the benefits of language technology, perpetuating digital divides along linguistic lines.

7 Limitations

Our analysis establishes correlation between fertility and accuracy but cannot establish causation. Token fertility correlates with training data availability, where languages with less representation in pretraining corpora receive fewer subword merges during tokenizer training, and have consequent higher fertility. The performance degradation observed may reflect tokenizer inefficiency, scant data, or a combined effect. Also, while we report aggregate performance for such models where possible, fine-grained fertility analyses cannot be conducted without tokenizer transparency.

In terms of access to model components, we could not analyze OpenAI's o1 model tokenizer as implementation details remain proprietary, limiting our ability to fully characterize reasoning models' tokenization strategies. Our analysis focuses on African languages which limits generalization to other language families. It's possible that similar patterns would be present for morphologically complex, low-resource languages, this is untested. While we analyze 16 African languages, this represents less than 1% of Africa's 2,000+ languages. Our findings may not generalize to languages with different morphological properties or non-Latin scripts. We evaluate only multiple-choice question answering, MCQA tasks. Performance degradation patterns may differ for generative tasks where token inflation affects both input and output, potentially compounding the effects we observe, but this remains speculation. Our cost calculations use publicly available pricing and may not reflect negotiated rates or future hardware improvements. We focus on direct computational costs without quantifying broader environmental impacts.

By quantifying performance gaps, we risk reinforcing perceptions that some languages are lesser for NLP applications. We emphasize that these disparities reflect technological limitations, not inherent language properties. Documenting the "token tax" could discourage investment in low-resource language technologies if stakeholders focus solely on costs rather than equity. This document is intended to motivate solutions. To address these risks, we suggest developing inclusive tokenization standards through community participation; advocating for subsidized compute resources for low-resource language research; and creating evaluation metrics that explicitly penalize tokenization inefficiency.

References

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating LLM performance on African languages. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32704–32731, Vienna, Austria. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A Smith. 2024. Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization. Advances in Neural Information Processing Systems, 37:47790–47814.

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

Tuka Alhanai, Adam Kasumovic, Mohammad Ghassemi, Aven Zitzelberger, Jessica Lundin, and Guillaume Chabot-Couture. 2024. Bridging the gap: Enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments. Preprint, arXiv:2412.12417.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, and 2 others. 2024. Tokenizer choice for LLM training: Negligible or crucial? In Findings of the Association for Computational Linguistics: NAACL 2024, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O. Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. msteb: Massively multilingual evaluation of llms on speech and text tasks. Preprint, arXiv:2506.08400.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2022. On the computational complexity of self-attention. Preprint, arXiv:2209.04881.

Abraham Toluwase Owodunni, Orevaoghene Ahia, and Sachin Kumar. 2025. Flexitokens: Flexible tokenization for evolving language models. arXiv preprint arXiv:2507.12720.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. Preprint, arXiv:2305.15425.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3118–3135, Online. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. Preprint, arXiv:2412.03304.

Makesh Narsimhan Sreedhar, Xiangpeng Wan, Yu Cheng, and Junjie Hu. 2023. Local byte fusion for neural machine translation. Preprint, arXiv:2205.11490.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. Preprint, arXiv:2207.04672.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Preprint, arXiv:1706.03762.

Appendix

A Regression Results by Subject

Table 1 and Figure 3 show results of accuracy-on-fertility regressions for the 10 models over 5 subjects.

B Fertility

Fertility measures the average number of tokens required to represent a word in a corpus:

$$F = \frac{T}{W}$$

for T and W the token and word counts. Higher F inflates sequence length, affecting a model's ability to learn long-range dependencies and compute costs (Ali et al., 2024).

C Inference

C.1 Training and Inference Cost Comparison: English vs. Language X

This is a thought exercise in training and inference costs for LLMs applied to English and Language X. The analysis assumes the same model architecture and tokenizer across languages, with cost differences due to tokenization inefficiencies and quadratic $O(n^2)$ training scaling (Vaswani et al., 2017) of transformer models.

We assume Language X has a fixed **2× increase** in tokens across tokenizers (although there are variations not included here). We assume English has 1 000 000 tokens (baseline) and Language X: approximately 2 000 000 tokens for equivalent content. There is a $2^2 = 4$x increase in training cost.

In addition to cost, token inflation impacts time. Transformer models scale quadratically in sequence length. With a $2\times$ token increase, Language X requires $4\times$ more compute. This means training that takes 90 days for English would take $\sim$360 days for Language X on the same hardware. For inference time, decoding scales approximately linearly with token count. A prompt completion that takes 2 seconds in English may take about 4 seconds in Language X.

These multipliers apply whether the additional tokens appear in the input (prompt) or output (completion), and they exacerbate cost disparities for low-resource languages.

Using published petaFLOP-day figures and assuming a compute cost of $240 per petaFLOP-day, Table 3 displays order-of-magnitude estimates of compute and cost for training, while Table 4 displays order-of-magnitude estimates for inference.

C.2 Prompt

```
You must only reply with 'Final Answer: X'
where X is A, B, C, or D.
Do NOT add explanations, reasoning,
or extra text.
Question: <question text>
Choices:
A. <option 1>
B. <option 2>
C. <option 3>
D. <option 4>
Your response must be strictly formatted as:
Final Answer: X
```

D Reproducibility: Code and Data

Raw LLM outputs, tokens, fertility and parity measures for each model, and scripts to replicate our analysis are available at:

Model	**English**	**2× Fertility**	**5× Fertility**
Llama 2 70B	$5M	$20M	$125M
Llama 3 70B	$24M	$96M	$600M
Llama 3.1 405B	$105M	$420M	$2.6B

Table 2: Training costs scale quadratically with fertility

Model	petaFLOP-days	English $	Language X ($4×)
LLaMA 2 (69B)	21 000	5 M	20 M
LLaMA 3 (70B)	100 000	24 M	96 M
LLaMA 3.1 (405B)	440 000	105 M	420 M

Table 3: Training compute and cost estimates for LLaMA models (USD).

Provider	Model (type)	English $	Language X (∼2×)
OpenAI	GPT-4o	5 / 20	10 / 40
OpenAI	o4-mini*	4 / 16	8 / 32
Google	Gemini 2.5 Flash	0.30 / 2.50	0.60 / 5.00
Google	Gemini 2.5 Pro*	1.25 / 10	2.50 / 20
Anthropic	Claude 4 Sonnet	3 / 15	6 / 30
Anthropic	Claude 4 Opus*	15 / 75	30 / 150

Table 4: Inference cost per 1M English-equivalent tokens (USD) including reasoning models. The costs are shown for input/output.

https://github.com/jessicalundin/multilingual_token_tax.

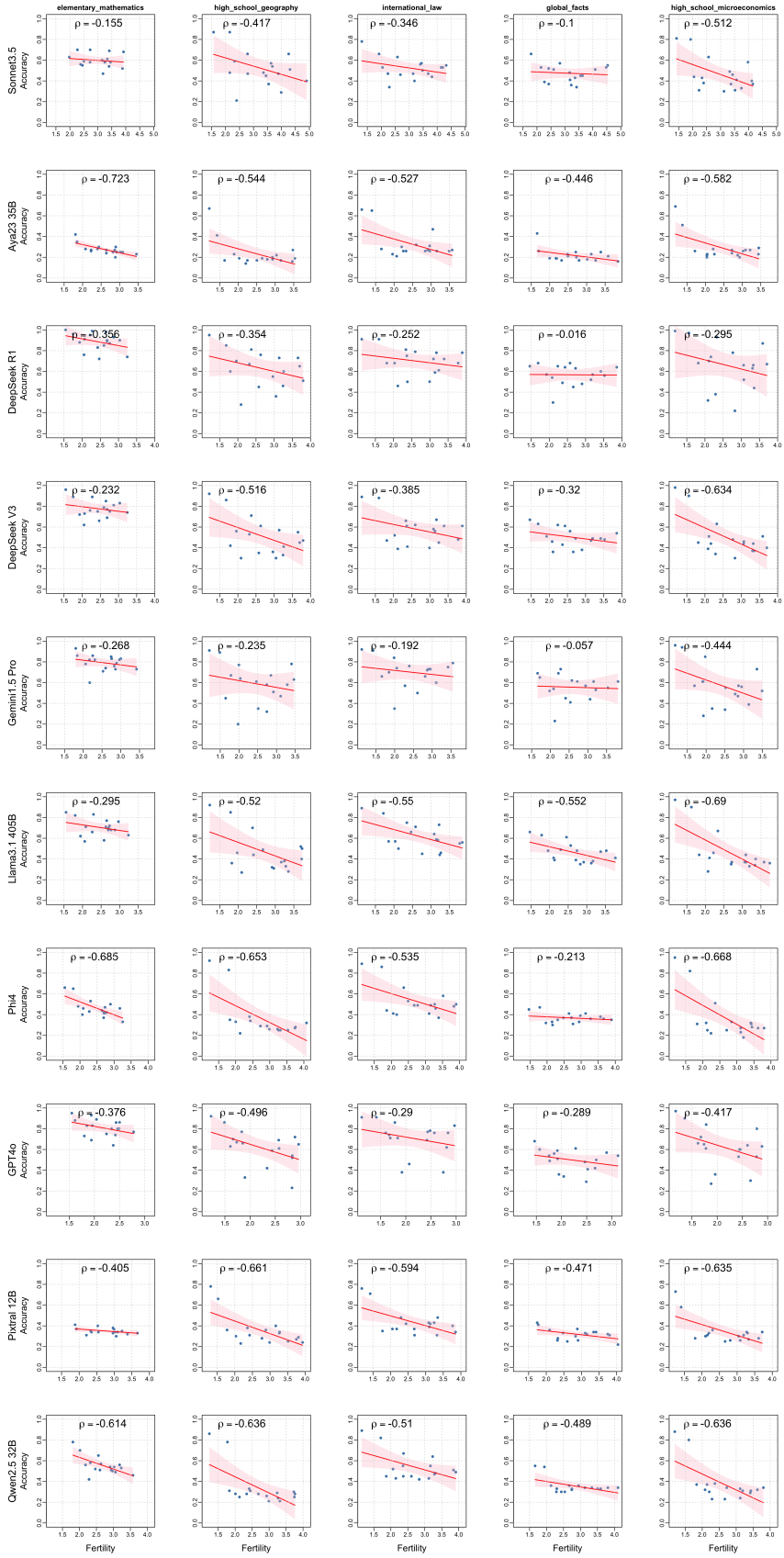

Figure 3: Fertility and accuracy trade-offs for the 10 models across five MMLU subjects. Note that these associations are correlational: fertility captures tokenization inefficiency that covaries with performance, but does not isolate causal effects independent of pretraining data availability or quality.