Low Resource Languages Pay Twice

Quantifying the High- vs. Low-Resource Language Gap: Predicting Large Language Model Accuracy through Tokenization Fertility and Parity

Hamza Louzan, Nihal Karim, Victor Wei, Ada Zhang, Jessica Lundin, Cody Carroll

BILL & MELINDA GATES foundation

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

ABSTRACT

Large language models (LLMs) show impressive performance on human-like tasks. However, there exists a major disparity in LLM performance between high-resource languages (HRL) and low-resource languages (LRLs), of which the former consistently shows markedly better task performance compared to the latter. Through thorough evaluation of various tokenization strategies, we notice that tokenization on HRL and LRL texts result in discrepancies in **fertility** and **parity** metrics, which affect downstream model performance on Q&A tasks. In this paper, we:

- Present a comprehensive evaluation of various LLM tokenization methods using tokenization metrics.
- Compare performance on Q&A tasks.
- Ilustrate the negative association between fertility and accuracy.

INTRODUCTION

We leverage **tokenization metrics** as early predictors of Q&A accuracy. The metrics we use are **fertility** and **parity**, of which the former is a measure of tokenization length and the latter a measure of tokenization fairness. A **high-resource language (HRL)** is one where data is abundant, compared to a **low-resource language (LRL)**, where data is scarce. Through a comprehensive multilingual analysis encompassing both HRLs and LRLs, our study reveals that the typically elevated fertility scores observed in LRLs serve as reliable **indicators of diminished downstream performance**.

Through our study, we establish two key findings:

1. *LRLs consistently have lower fertility scores compared to HRLs.*
2. *Fertility score is an indicator of LLM performance, without the need to run a model we can predict MCQA scores.*

A Sample HRL vs. LRL Comparison

Language	Text	Tokens	Length
English (HRL)	The sky is beautiful tonight.	['The', 'Ġsky', 'Ġis', 'Ġbeautiful', 'Ġtonight', '.']	6
Hindi (LRL)	आज रात आसमान बहुत सुन्दर है।	['à¤Ī', 'à¤Ī', 'ĠàĤ°', 'à¤¾à¤¤', 'ĠàĤĨà¤¸', 'à¤®', 'à¤¾à¤¨', 'Ġà¤¬à¤¹à¥', 'à¥ģत', 'ĠàĤ¸', 'à¥ģà¤', 'ĠàĤ°', 'à¥įà¤¦à¤°', 'Ġà¤¹', 'à¥Ī।']	14

Table 1. The same sentence in two different languages, tokenized with the meta-llama/Llama-3.2-1B-Instruct model. Even for a simple sentence, the sentence in Hindi (a low resource language) has a much longer token length compared to the same sentence in English (a high resource language).

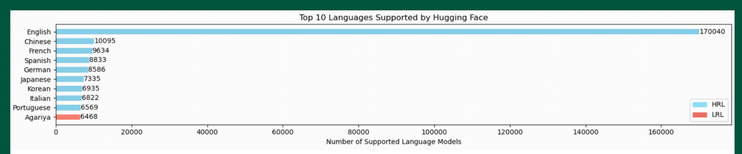

Figure 1. Top 10 languages supported by Hugging Face. English has the most models, whether pre-trained or fine-tuned, on Hugging Face. It has roughly 17 times more models than the second most-supported language.

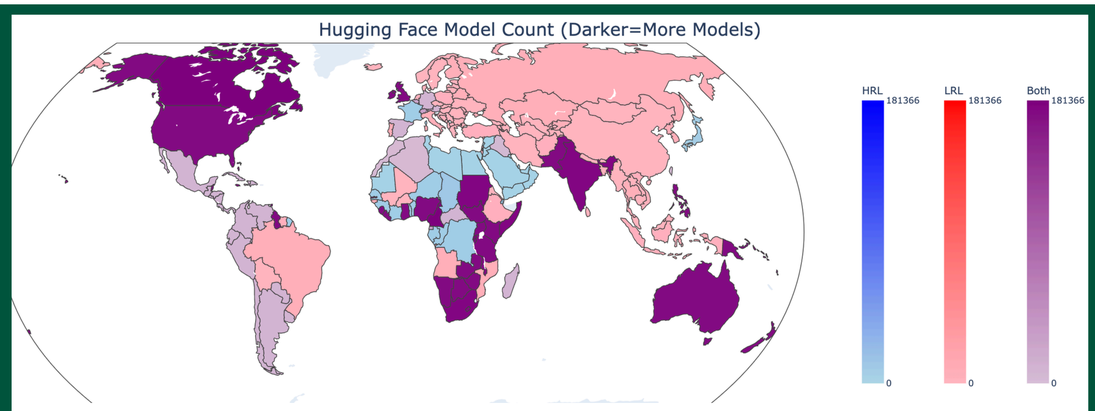

Figure 2. Visualization of Hugging Face model availability for countries around the world. Since countries can have speakers of high and low resource languages, each country/region is classified as high resource (HRL), low resource (LRL), or both high and low resource (Both).

METHODS

To evaluate LLM performance on Q&A tasks, we used these steps:

1. We calculated tokenization metrics (fertility and parity scores) on a multilingual dataset to show that there is a difference in metrics between languages, especially between those of LRLs (worse metrics) and HRLs (better metrics).
2. Using a multilingual Q&A dataset, we calculated the tokenization metrics and prompted models to answer the question.
3. We compared the accuracy between languages.
4. We built statistical models to show the relationship between language fertility and LLM accuracy.

In our evaluations, we consider the **FLORES+** (Open Language Data Initiative, 2024) and **AfriMMLU** (Masakhane, 2024) datasets. FLORES+ is a multilingual machine translation dataset developed by Meta. It contains translations from English into about 200 other languages and is a newer version of the FLORES-200 dataset (NLLB Team, 2022) . The AfriMMLU dataset is a subset of the MMLU dataset. It contains a total of 17 total languages, English, French, and 15 African languages. A Q&A evaluation dataset, it contains a question, choices, answer, and subject columns. The subjects include elementary mathematics, global facts, high school geography, high school microeconomics, and international law.

In our evaluation of LLM performance on multilingual Q&A tasks, we employed two tokenization metrics to assess the efficiency and consistency of tokenizers across different languages: fertility and parity.

Fertility is calculated by *dividing the total number of tokens in a tokenized dataset by the total number of words in the original dataset*. **Parity** is the *length of the tokenized sentence in language A divided by the length of the tokenized sentence translated in language B*.

RESULTS

Interaction Between Fertility and Accuracy by Subject

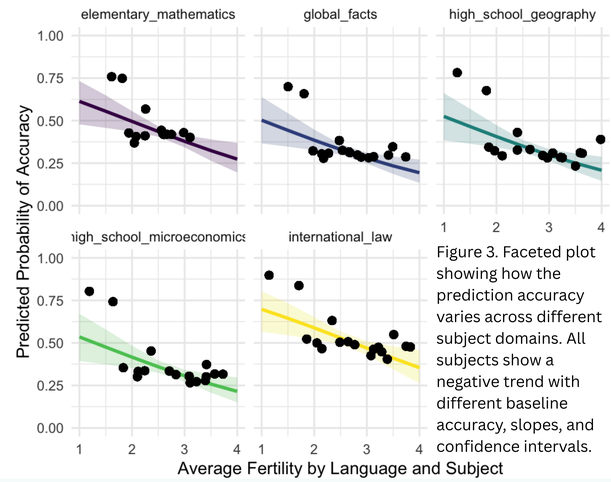

Figure 3. Faceted plot showing how the prediction accuracy varies across different subject domains. All subjects show a negative trend with different baseline accuracy, slopes, and confidence intervals.

Random Effects

Group	Parameter	Variance	Std. Dev.	Correlation
Language	Intercept	2.71	1.65	-
	Avg. fertility by lang/subj	0.37	0.61	-1.00

Number of observations: 8,500 across 17 language groups

Fixed Effects

Parameter	Estimate	Std. Error	z value	p-value
Intercept	0.939	0.479	1.960	0.050 *
Avg. fertility by lang/subj	-0.479	0.178	-2.690	0.007 **
Subject: global facts	-0.450	0.077	-5.861	<0.001 ***
Subject: high school geography	-0.361	0.087	-4.135	<0.001 ***
Subject: high school microecon	-0.320	0.088	-3.653	<0.001 ***
Subject: international law	0.376	0.086	4.393	<0.001 ***

We used a mixed-effect model to view the relationship between fertility and accuracy across different large language models. We discovered that he relationship between tokenization fertility and accuracy is language-dependent. The random slopes component in our model is particularly revealing because it demonstrates that while fertility generally negatively impacts accuracy, the magnitude of this effect varies significantly across language families.

CONCLUSION & FUTURE WORK

Through our research, we've shown that Latin scripts have the highest fertility scores while non-Latin scripts have the lowest, demonstrating the unfair tokenization quality between languages. We also modeled fertility as a predictor of LLM model accuracy for Q&A tasks. While we were limited by computational costs of running LLMs, there are many ways to extend our work. In the future, we plan to add parity as a predictor of accuracy. Another direction to enhance our work is to extract the last hidden layer (contextual embeddings) for a handful of models and compare token embedding spaces between HRLs and LRLs. We can also take AfriMMLU and FLORES+ texts and perturb them (e.g. add typos, reorder words or tokens, replace words synonyms) and then compare the performance drop across HRLs vs. LRLs.