Beyond Fertility: Analyzing STRR as a Metric for Multilingual Tokenization Evaluation

Mir Tafseer Nayeem*, Sawsan Alqahtani*, Md Tahmid Rahman Laskar,
Tasnim Mohiuddin, and M Saiful Bari

NEURAL INFORMATION PROCESSING SYSTEMS 2025

Motivation and Challenges

- Tokenization governs how capacity is allocated in LLMs, impacting efficiency and fairness.
  - The standard metric is **Fertility** (*average tokens per word*).
- **Problem:** Fertility is an average that obscures how vocabulary is allocated. It compresses behavior into a narrow numeric band, hiding blind spots in multilingual and code-mixed scenarios.
- **Consequence:** High fertility signals inefficiency, but doesn't tell us which words are fragmented. A tokenizer might fragment common words in Hindi while keeping English intact, biasing model capacity.

Introducing STRR

- We propose **STRR** (*Single Token Retention Rate*) to complement fertility. It measures the proportion of words preserved as single tokens.

Given a set of words $W = \{w_1, \ldots, w_n\}$ and a tokenizer T, we define

$$\mathrm{STRR}(T; W) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(|T(w_i)| = 1) \times 100;$$

STRR thus measures the percentage of words encoded as a single token.

Why STRR?

- **Type-Level Diagnostic:** Unlike fertility (*token-level average*), STRR checks if specific words are kept whole.
- **Fairness Sensitive:** Directly reveals if a tokenizer allocates vocabulary space to a language's core lexicon.
- **Interpretable:** Differentiates between necessary linguistic segmentation and suboptimal allocation.

Experimental Setup

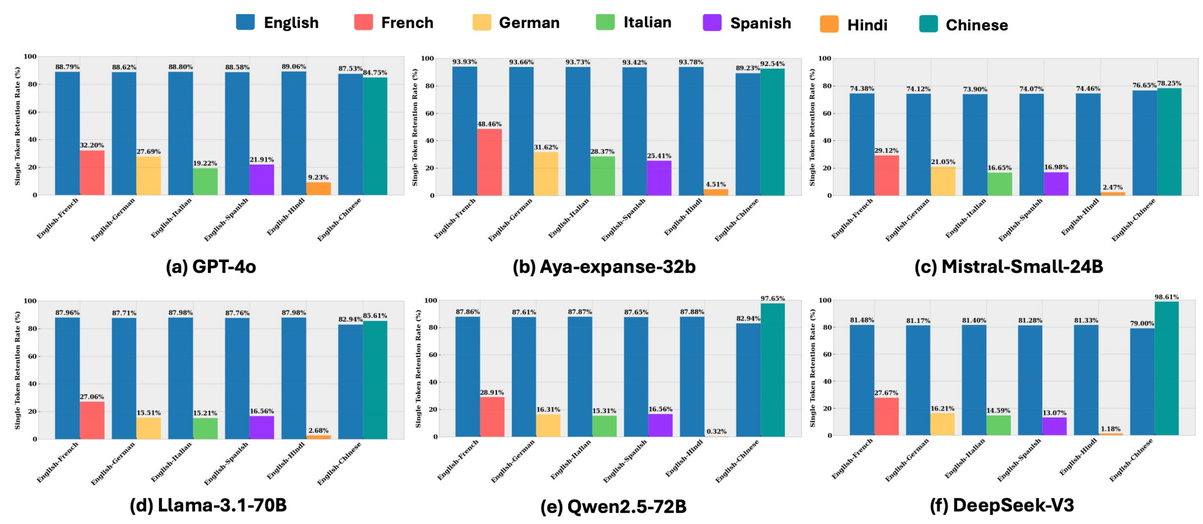
Models: GPT-4o, Aya-Expanse-32B, Mistral-Small-24B, Llama-3.1-70B, Qwen2.5-72B, DeepSeek-V3.

Languages: English, French, German, Spanish, Italian, Hindi, Chinese.

Dataset: 1,000 Most Common Words (one-to-one human translation pairs aligned across languages).

Key Findings

- **Systemic Prioritization of English Language**
  - English words are overwhelmingly retained as single tokens across all tested tokenizers.
  - This suggests models primarily learn direct mappings from English tokens, reducing reliance on extensive multilingual pretraining.
- **The Chinese Language Strategy**
  - High Fertility: Chinese shows high fertility scores (1.82–2.40) due to logographic script properties.
  - High STRR: Despite high fertility, Qwen2.5 and DeepSeek-V3 show very high STRR for Chinese, indicating explicit vocabulary integration for whole words.
- **Hindi Language Fragmentation**
  - Hindi exhibits the lowest STRR across all tokenizers.
  - This reveals pronounced fragmentation of high-frequency vocabulary, leading to inflated inference costs.



Recommendation

We propose a pipeline based on the Pareto Principle (80% of text comes from 20% of vocabulary) to fix tokenizer inequity.

- **Identify Core Vocabulary:** Select the highest-frequency words (top 1K) in the target language.
- **Vocabulary Injection:** Add these words to the tokenizer as single tokens if STRR indicates fragmentation.
- **Corpus Pretraining:** Update embeddings using available multilingual text.
- **Instruction Tuning:** Validate using multilingual instruction-response datasets.

Conclusion

- STRR reveals biases that fertility misses, specifically the fragmentation of languages like Hindi.
- We release code and curated lists of the 1,000 most frequent words in seven languages to facilitate equitable tokenizer design.