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

Mir Tafseer Nayeem[♣*] **Sawsan Alqahtani**[♠*]
Md Tahmid Rahman Laskar[♦] **Tasnim Mohiuddin**[◇] **M Saiful Bari**[♥]
[♣]University of Alberta [♠]Princess Nourah Bint Abdulrahman University
[♦]Dialpad [◇]Qatar Computing Research Institute [♥]Amazon

Abstract

Tokenization is a crucial but under-evaluated step in large language models (LLMs). The standard metric, fertility (the average number of tokens per word) captures compression efficiency but obscures how vocabularies are allocated across languages and domains. We analyze six widely used tokenizers across seven languages and two domains, finding stable fertility for English, high fertility for Chinese, and little domain sensitivity. To address fertility's blind spots, we propose the Single Token Retention Rate (STRR), which measures the proportion of words preserved as single tokens. STRR reveals systematic prioritization of English, strong support for Chinese, and fragmentation in Hindi, offering an interpretable view of cross-lingual fairness. Our results show that STRR complements fertility and provides practical guidance for designing more equitable multilingual tokenizers.[2]

1 Introduction

Tokenization is a foundational step in large language models (LLMs), shaping how text is split into model-readable units, yet its evaluation remains under-examined and constrained by a lack of interpretable metrics [Bostrom and Durrett, 2020]. Existing metrics often prioritize *compression efficiency*, with *fertility* (the average number of subword tokens generated per word) serving as a standard diagnostic [Rust et al., 2021, Ali et al., 2024] (see §2 and §B for other possible evaluation metrics). High fertility scores typically signal inefficiency, since more tokens are required to represent the same semantic content. Despite its wide adoption, fertility has important blind spots: as a token-level average, it obscures how vocabulary is allocated across languages, domains, and usage contexts, and its link to downstream LLM performance remains unclear [Bostrom and Durrett, 2020]. Yet, as Table 1 suggests, fertility compresses behavior into a narrow numeric band and offers little diagnostic guidance about where vocabulary capacity is misallocated.

These limitations are consequential: tokenization governs how capacity is allocated, affecting downstream efficiency, fairness, and representation quality in LLMs. A tokenizer that fragments words in some languages more than others implicitly biases model capacity, inflating training and inference costs for those languages and amplifying performance disparities [Bostrom and Durrett, 2020]. Moreover, evaluation centered solely on fertility obscures challenges that arise in multilingual and code-mixed scenarios, where speakers fluidly switch across linguistic boundaries [Mabule, 2015]. Such settings expose weaknesses in current tokenizers, particularly when English, functioning as a global lingua franca, interacts with diverse native languages [Jenkins, 2009].

[*]Equal contribution.

[2]Our code and dataset are available at github.com/tafseer-nayeem/STRR

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

Despite advances in multilingual pretraining, tokenizers still struggle to balance two competing goals: preserving coverage across diverse languages, and scripts, while minimizing fragmentation. We argue that existing evaluation practices are insufficient to guide tokenizer design toward this balance.

To address this gap, we contribute in two ways. First, we present a cross-lingual evaluation of six LLM tokenizers across seven languages and two domains (formal and informal) (§4). Second, we introduce the Single Token Retention Rate (STRR), a novel metric that measures the proportion of words preserved as single tokens across languages. Unlike fertility, STRR better captures tokenizers' vocabulary allocation and provides an interpretable diagnostic for fairness and efficiency (§5). Together, these analyses shed light on how contemporary tokenizers implicitly prioritize certain languages and suggest directions for equitable and efficient multilingual tokenizer design (§6).

2 Related Work

Most tokenizer evaluations rely on *fertility* (the average number of tokens per word) valued for its simplicity but limited in scope [Rust et al., 2021]. Other measures such as vocabulary coverage, subword entropy, compression rates, or character-to-token ratios have been proposed [Goldman et al., 2024, Zouhar et al., 2023, Libovický and Helcl, 2024, Signoroni and Rychlý, 2022, Lotz et al., 2025], yet none have become standard practice. Linguistically motivated metrics also exist [Arnett et al., 2025, Beinborn and Pinter, 2023, Asgari et al., 2025], but they are often language-specific and difficult to interpret across diverse scripts.

These approaches emphasize compression efficiency but rarely reveal how vocabulary is distributed across languages or domains, nor do they consistently correlate with downstream model performance [Bostrom and Durrett, 2020, Ali et al., 2024]. Prior work highlights the consequences of uneven token allocation: inflated inference costs for some languages [Ahia et al., 2023], reduced cross-domain robustness [Dagan et al., 2024], and misalignment with linguistic boundaries [Yin et al., 2024, Bogin et al., 2022]. Together, these studies underscore the need for standardized, interpretable metrics that capture both efficiency and fairness in multilingual settings.

3 Experimental Setup

Tokenizers: We selected six widely used LLM tokenizers: `GPT-4o`, `Aya-Expanse-32B` [Dang et al., 2024], `Mistral-Small-24B`[3], `Llama-3.1-70B` [Dubey et al., 2024], `Qwen2.5-72B` [Qwen-Team et al., 2025], and `DeepSeek-V3` [DeepSeek-AI et al., 2024].

Datasets: For fertility and related metrics we use formal text (XL-Sum news; Hasan et al. 2021) and informal text (MultilingualSentiment; clapAI 2024). For STRR, we build a multilingual wordlist from `1000MostCommonWords`[4], aligning 1,000 translation pairs (e.g., English–French) per language to ensure cross-lingual comparability and reflect high-frequency vocabulary.

Languages: We consider several languages (English, German, French, Spanish, Italian, Hindi, and Chinese) selected because they are (i) officially supported by the evaluated LLMs and (ii) included in widely used multilingual benchmarks (e.g., MMMLU[5]). For the fertility analyses, we restrict to English, French, Spanish, and Chinese to ensure uniform data availability across both formal and informal domains in the chosen datasets.

4 Fertility Analysis

Table 1 reports fertility scores across languages and tokenizers. Additional metrics, subword entropy and characters-per-token (defined in §B), are shown in Table B, and display trends consistent with the fertility results. English shows striking consistency across both formal and informal domains, reflecting its dominance in pretraining corpora [Dubey et al., 2024, DeepSeek-AI et al., 2024] and relatively simple morphology [Bentz et al., 2016].

[3] `https://mistral.ai/en/news/mistral-small-3`

[4] `https://1000mostcommonwords.com`

[5] `https://huggingface.co/datasets/openai/MMMLU`

Overall, domain differences are minimal, suggesting that large vocabularies (128K–255K; Appendix, Table 2) capture both structured and unstructured text efficiently. In contrast, Chinese exhibits the highest fertility due to its logographic script and absence of explicit word boundaries [Si et al., 2023]. Tokenizers vary in how much vocabulary they allocate to whole words versus smaller units; fertility reflects these numerical differences but cannot distinguish necessary linguistic segmentation from suboptimal allocation.

Languages	Domains	GPT 4o	Aya-exp 32B	Mistral 24B	Llama 3.1-70B	Qwen 2.5-72B	DeepSeek V3
English	Formal	1.22	1.24	1.27	1.23	1.25	1.23
	Informal	1.22	1.25	1.27	1.25	1.26	1.25
French	Formal	1.42	1.42	1.43	1.67	1.68	1.61
	Informal	1.37	1.42	1.41	1.58	1.58	1.57
Spanish	Formal	1.36	1.33	1.42	1.61	1.61	1.55
	Informal	1.32	1.36	1.44	1.53	1.53	1.53
Chinese	Formal	1.89	1.82	2.21	1.89	2.40	1.95
	Informal	1.86	1.96	2.30	1.92	2.40	1.95

Table 1: Fertility values across languages, domains (formal and informal), and tokenizers.

While informative, fertility, subword entropy, and characters-per-token each have blind spots that limit their usefulness for equitable multilingual tokenizer design. Fertility collapses behavior into average tokens per word, masking over-fragmentation. Subword entropy summarizes distributional balance but remains abstract and hard to localize. Characters-per-token highlights script differences but reduces quality to mean token length, ignoring whether frequent words remain intact.

5 Single Token Retention Rate (STRR)

We propose the Single-Token Retention Rate (STRR), which measures the proportion of words preserved as single tokens. STRR serves two goals: (i) probing vocabulary construction by quantifying whole-word retention in each language, and (ii) revealing how tokenizers allocate limited vocabulary across languages. It highlights inequities directly and points to actionable remedies, such as expanding coverage for under-represented high-frequency words. Unlike fertility, subword entropy, or characters-per-token—which are computed on text corpora and averaged over tokenized outputs; STRR is defined on a reference wordlist. It checks, for each word, whether the tokenizer has allocated a single token, making it a type-level rather than token-level diagnostic. This design makes STRR interpretable, fairness-sensitive, and tied to practical interventions.

5.1 Definition

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.

5.2 Results

As illustrated in Figure 1, across all tokenizers, ***English words in translation pairs are overwhelmingly retained as single tokens***. This supports the hypothesis that tokenizers allocate significant vocabulary space to English representations, reinforcing findings that even limited multilingual exposure enhances LLM multilingual capabilities [Shaham et al., 2024], as models primarily learn direct mappings from English tokens to multilingual equivalents, reducing reliance on extensive multilingual pretraining.

Our STRR analysis reveals that ***all LLMs explicitly integrate Chinese vocabulary into their tokenization strategies*** to reduce segmentation artifacts as observed in Table 1. Notably, Qwen2.5-72B and DeepSeek-V3 exhibit the highest STRR for Chinese, suggesting enhanced language-specific support for whole-word representations.

Hindi exhibits the lowest STRR across all evaluated tokenizers, revealing pronounced fragmentation and suboptimal vocabulary allocation. Crucially, STRR quantifies this inefficiency with a direct, interpretable measure, rather than simply echoing prior fertility-based findings [Ahia et al., 2023], offering clear guidance for targeted vocabulary expansion in under-served languages (§6).

6 Discussion & Recommendations

Identifying Core Vocabulary via the Pareto Principle: The Pareto Principle, or "80/20 rule," posits that a small fraction of the lexicon accounts for most language use [Sanders, 1987]. In English,

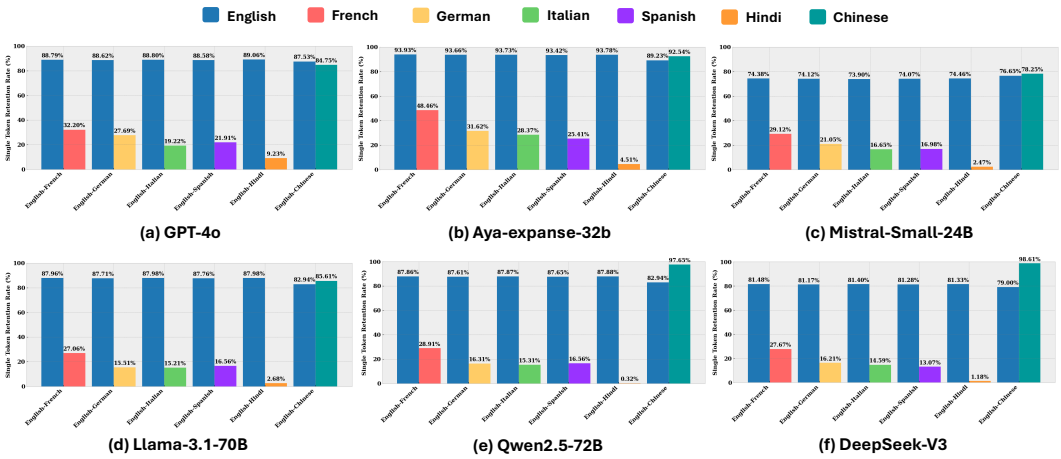

Figure 1: Single Token Retention Rate (STRR) across six LLM tokenizers for different language pairs. Each pair (e.g., English-French) represents 1,000 parallel words in both languages, allowing us to examine whether LLM tokenizers prioritize the English versions of words over their multilingual counterparts. A high STRR for English suggests that tokenizers allocate more vocabulary space to English, while differences in STRR across languages indicate varying degrees of support.

the General Service List (GSL) of roughly 2,000 words covers 80–85% of standard written text [West, 1953]. We thus advocate that multilingual tokenizer developers identify an analogous *core vocabulary* in each language (namely, the highest-frequency words that dominate token counts) and ensure they are encoded as single tokens. Prioritizing this compact set minimizes subword fragmentation and maximizes encoding efficiency without unnecessarily expanding the overall vocabulary.

End-to-End Vocabulary Expansion Pipeline: We propose a practical four-stage pipeline for enhancing multilingual tokenizers, feasible even in low-resource settings or without large pretraining corpora. As a shared baseline, we release curated lists of the 1,000 most frequent words in seven major languages.

1. **Core Vocabulary Identification:** Select the highest-frequency words in each target language using our curated lists or extend them as needed.[6]
2. **Vocabulary Injection:** Add identified words to the tokenizer's vocabulary as single tokens. Use STRR to check which are already represented and which require injection (§5).
3. **Corpus Pretraining:** Continue pretraining or fine-tuning the base multilingual LLM on publicly available multilingual text [Üstün et al., 2024], incorporating the expanded vocabulary to learn robust embeddings.
4. **Multilingual Instruction Tuning:** Instruction-tune the model on multilingual instruction–response datasets [Singh et al., 2024] to validate and reinforce the expanded vocabulary in downstream tasks.

This pipeline can reduce subword fragmentation, facilitate faster adaptation, and potentially improve consistency across diverse languages.

7 Conclusion

We introduced STRR, a simple interpretable metric that complements fertility by capturing whole-word preservation in multilingual tokenization. Our analysis across tokenizers shows that STRR reveals biases, favoring English and Chinese while fragmenting languages like Hindi, that fertility alone cannot. By releasing high-frequency word lists, providing code, and outlining a vocabulary-expansion pipeline, we offer actionable steps toward more efficient and equitable tokenizer design.

[6] https://1000mostcommonwords.com/languages/

References

Kaj Bostrom and Greg Durrett. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*, 2020.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

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, Stefan Kesselheim, and Nicolas Flores-Herr. Tokenizer choice for LLM training: Negligible or crucial? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.247. URL https://aclanthology.org/2024.findings-naacl.247/.

D R Mabule. What is this? is it code switching, code mixing or language alternating? *Journal of Educational and Social Research*, 5(1), 2015. ISSN 2240-0524. URL https://www.mcser.org/journal/index.php/jesr/article/view/5628.

Jennifer Jenkins. English as a lingua franca: interpretations and attitudes. *World Englishes*, 28 (2):200–207, 2009. doi: https://doi.org/10.1111/j.1467-971X.2009.01582.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-971X.2009.01582.x.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. Unpacking tokenization: Evaluating text compression and its correlation with model performance. *arXiv preprint arXiv:2403.06265*, 2024.

Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. Tokenization and the noiseless channel. *arXiv preprint arXiv:2306.16842*, 2023.

Jindřich Libovický and Jindřich Helcl. Lexically grounded subword segmentation. *arXiv preprint arXiv:2406.13560*, 2024.

Edoardo Signoroni and Pavel Rychlý. HFT: High frequency tokens for low-resource NMT. In Atul Kr. Ojha, Chao-Hong Liu, Ekaterina Vylomova, Jade Abbott, Jonathan Washington, Nathaniel Oco, Tommi A Pirinen, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao, editors, *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 56–63, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.loresmt-1.8/.

Jonas F Lotz, António V Lopes, Stephan Peitz, Hendra Setiawan, and Leonardo Emili. Beyond text compression: Evaluating tokenizers across scales. *arXiv preprint arXiv:2506.03101*, 2025.

Catherine Arnett, Marisa Hudspeth, and Brendan O'Connor. Evaluating morphological alignment of tokenizers in 70 languages. *arXiv preprint arXiv:2507.06378*, 2025.

Lisa Beinborn and Yuval Pinter. Analyzing cognitive plausibility of subword tokenization. *arXiv preprint arXiv:2310.13348*, 2023.

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies, 2025. URL https://arxiv.org/abs/2502.00894.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

Gautier Dagan, Gabriel Synnaeve, and Baptiste Roziere. Getting the most out of your tokenizer for pre-training and domain adaptation. *arXiv preprint arXiv:2402.01035*, 2024.

Yongjing Yin, Lian Fu, Yafu Li, and Yue Zhang. On compositional generalization of transformer-based neural machine translation. *Information Fusion*, 111:102491, 2024.

Ben Bogin, Shivanshu Gupta, and Jonathan Berant. Unobserved local structures make compositional generalization hard. *arXiv preprint arXiv:2201.05899*, 2022.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Qwen-Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, et al. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413/.

clapAI. Multilingualsentiment: A multilingual sentiment classification dataset. Hugging Face Datasets, 2024. URL https://huggingface.co/datasets/clapAI/MultiLingualSentiment. A multilingual dataset for sentiment analysis with labels: positive, neutral, negative, covering diverse languages and domains.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. A comparison between morphological complexity measures: Typological data vs. language corpora. In Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François, and Philippe Blache, editors, *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/W16-4117/.

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. Sub-character tokenization for Chinese pretrained language models. *Transactions of the Association for Computational Linguistics*, 11:469–487, 2023. doi: 10.1162/tacl_a_00560. URL https://aclanthology.org/2023.tacl-1.28/.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.136. URL https://aclanthology.org/2024.findings-acl.136/.

Robert Sanders. The pareto principle: Its use and abuse. *Journal of Services Marketing*, 1(2):37–40, 1987. doi: 10.1108/eb024706. URL https://www.emerald.com/insight/content/doi/10.1108/eb024706/full/html.

Michael West. *A General Service List of English Words: With Semantic Frequencies and a Supplementary Word-List for the Writing of Popular Science and Technology*. Longman, Green and Co., London, 1953. URL https://en.wikipedia.org/wiki/General_Service_List.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL https://aclanthology.org/2024.acl-long.845/.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL https://aclanthology.org/2024.acl-long.620/.

A Model Details

Models	Vocab Size	Model ID
GPT-4o	200,019	Link
Aya-Expanse-32B (Dang et al. [2024])	255,029	HF Link
Mistral-Small-24B	131,072	HF Link
Llama-3.1-70B (Dubey et al. [2024])	128,256	HF Link
Qwen2.5-72B (Qwen-Team et al. [2025])	151,665	HF Link
DeepSeek-V3 (DeepSeek-AI et al. [2024])	128,815	HF Link

Table 2: Details of the models used in our experiments, including total vocabulary size (with added tokens) for each model. "HF Link" refers to the corresponding Hugging Face model IDs.

B Tokenization Metrics: Definitions and Limitations

Metric	Definition	What it Captures	How STRR Differs
Fertility	Avg. number of tokens per word (compression proxy).	Measures sequence length efficiency; high fertility = more fragmentation.	STRR is type-level: counts % of whole words preserved. Fertility hides where fragmentation occurs, STRR pinpoints cross-lingual allocation.
Subword Entropy	Entropy of token frequency distribution across text.	Captures balance of vocabulary usage (skew vs. uniformity). High entropy = fairer, balanced allocation.	STRR measures whole-word retention per language, not distribution balance. Entropy flags global skew; STRR identifies which languages' words are fragmented.
Char-to-Token Ratio	Avg. number of characters per token.	Captures average token length; highlights script differences (e.g., Chinese vs. English).	STRR does not average token lengths, but directly counts intact words. Differentiates many slightly split words from severe fragmentation of core vocabulary.
STRR (Single Token Retention Rate)	Percentage of words preserved as single tokens.	Captures vocabulary allocation fairness and whole-word coverage across languages.	Provides actionable, interpretable diagnostic: directly shows which languages and words are under-served and can guide vocabulary expansion.

Table 3: Comparison of tokenization evaluation metrics. Fertility and char-to-token ratio measure compression/fragmentation averages; subword entropy measures distributional balance; STRR highlights cross-lingual fairness by directly quantifying whole-word retention.

Language	Domain	GPT-4o			Aya-Expanse-32B			Mistral-Small-24B			Llama-3.1-70B			Qwen2.5-72B			DeepSeek-V3		
		Fert.	**Ent.**	**Chars/Tok**	**Fert.**	**Ent.**	**Chars/Tok**	**Fert.**	**Ent.**	**Chars/Tok**	**Fert.**	**Ent.**	**Chars/Tok**	**Fert.**	**Ent.**	**Chars/Tok**	**Fert.**	**Ent.**	**Chars/Tok**
English	Formal	1.22	9.45	3.88	1.24	9.30	3.80	1.27	9.40	3.72	1.23	9.43	3.84	1.25	9.36	3.77	1.23	9.44	3.83
	Informal	1.22	9.59	3.76	1.25	9.49	3.67	1.27	9.61	3.59	1.25	9.59	3.66	1.26	9.57	3.64	1.25	9.61	3.65
French	Formal	1.42	9.75	3.60	1.42	9.44	3.60	1.43	9.70	3.58	1.67	9.70	3.06	1.68	9.67	3.06	1.61	9.73	3.18
	Informal	1.37	9.92	3.54	1.42	9.74	3.43	1.41	9.93	3.45	1.58	9.85	3.06	1.58	9.85	3.08	1.57	9.85	3.09
Spanish	Formal	1.36	9.64	3.73	1.33	9.55	3.82	1.42	9.64	3.60	1.61	9.67	3.16	1.61	9.65	3.16	1.55	9.66	3.29
	Informal	1.32	9.32	3.50	1.36	9.30	3.39	1.44	9.33	3.22	1.53	9.32	3.01	1.53	9.32	3.02	1.53	9.31	3.01
Chinese	Formal	1.89	9.02	0.90	1.82	7.56	0.93	2.21	8.14	0.77	1.89	9.29	0.89	2.40	8.25	0.71	1.95	6.64	0.87
	Informal	1.86	8.55	0.84	1.96	7.03	0.80	2.30	7.68	0.68	1.92	8.78	0.82	2.40	7.89	0.65	1.95	6.30	0.80

Table 4: Complete results: fertility (tokens/word), entropy (bits), and characters per token across languages, domains, and models.