Equity with Efficiency: An Empirical Study of Tokenizers for Multilingual Large Language Models

Kieron Seven Jun Wei Lee[1] **Muhammad Reza Qorib**[2]
Andrew Ivan Soegeng[1,3] **Hwee Tou Ng**[1]
[1]National University of Singapore [2]Carnegie Mellon University [3]SAP
e0968891@u.nus.edu, mrqorib@cmu.edu,
andrew.soegeng@u.nus.edu, dcsnght@nus.edu.sg

Abstract

Multilingual large language models (LLMs) depend on subword tokenization to bridge discrete text and continuous neural representation. State-of-the-art multilingual LLMs often use Byte-level Byte-Pair Encoding (BPE) tokenizers that structurally favor high-resource languages and Latin scripts. For speakers of underrepresented languages, particularly those across Southeast Asia, this bias inflates inference costs and widens cross-lingual capability gaps. We present the first systematic comparison of equitable tokenizers on a unified benchmark spanning 11 Southeast Asian languages. Beyond tokenizer-level analysis of compression efficiency and cross-lingual equity, we assess downstream task performance through controlled 1.5B-parameter language model training using the same training data. Our results show that Parity-aware BPE lies on the Pareto frontier of the efficiency-equity trade-off, achieving strong compression parity at competitive cost. Morphology-Driven Byte Encoding delivers the best semantic reasoning performance through morphologically richer representations, albeit at a higher computational expense. Byte Latent Transformer underperforms on downstream tasks, possibly because its architectural assumptions misalign with the constraints of limited low-resource training data. Together, our findings demonstrate that cross-lingual fairness and tokenization efficiency are not fundamentally at odds, and offer practical guidance for designing equitable multilingual models.[1]

1 Introduction

Multilingual large language models (LLMs) are central to cross-lingual information access, yet their performance remains deeply uneven across languages and scripts. A key driver of this disparity is tokenization: how raw text is segmented into subword units shapes model capacity, sequence length, and effective context window across languages (Petrov et al., 2023) .

Byte-level Byte-Pair Encoding (BPE) (Sennrich et al., 2016) is a widely used tokenization strategy in state-of-the-art LLMs, including the GPT (OpenAI, 2025) and Llama (Touvron et al., 2023) families, due to its simplicity and compression efficiency. Byte-level BPE encodes characters as UTF-8 bytes (Consortium, 2011) and iteratively learns byte-pair merges based on global co-occurrence frequency. This procedure introduces a structural bias as one Latin character is encoded as a single byte, while one non-Latin character requires two or more bytes. Combined with English-centric pretraining corpora, BPE's merge operations disproportionately favor Latin scripts and high-resource languages (Arnett et al., 2024).

The practical consequences of such bias are significant. Petrov et al. (2023) demonstrated that GPT-4's Byte-level BPE tokenizer produces sequence length disparities of up to 15×, with Chinese requiring 1.9× more tokens than English, Vietnamese 2.5×, and Burmese 11.7×. For speakers of low-resource non-Latin languages such as Khmer and Lao, these disparities translate directly into higher inference costs, degraded long-context reasoning, and diminished downstream task accuracy (Tamang and Bora, 2024).

Several tokenizers have been proposed to address these inequities. Parity-aware Byte-Pair Encoding rebalances merge frequencies across scripts (Foroutan et al., 2025). Morphology-Driven Byte Encoding (MYTE) grounds segmentation in morphological structure (Limisiewicz et al., 2024). Byte Latent Transformer (BLT) sidesteps a fixed vocabulary by operating directly over dynamic byte patches (Pagnoni et al., 2025). Each work evaluates its approach against BPE baselines, reporting improvements in equity and multilingual capability. However, these methods have never been compared against each other under uniform experimental con-

[1]Source code will be publicly released upon paper publication.

ditions.

In this paper, we present a benchmarking study to address this gap with the first systematic analysis of equitable tokenizers. We compare them across eleven Southeast Asian (SEA) languages: English, Burmese, Chinese, Indonesian, Khmer, Lao, Malay, Tagalog, Tamil, Thai, and Vietnamese. Using Byte-level BPE as a baseline, and controlling for training data, vocabulary size, and computational budget, we evaluate intrinsic tokenizer metrics and examine downstream LLM performance by training 1.5B-parameter decoder-only language models from scratch. Our study provides a direct empirical comparison of equitable tokenization methods, offering actionable insights for NLP practitioners to build fairer multilingual LLMs.

2 Related Work

2.1 Subword Tokenization

Subword tokenization has become the standard preprocessing step in multilingual LLMs, to uniformly segment text in any language into tokens. However, when trained on heterogeneous multilingual corpora, these approaches allocate vocabulary capacity toward languages with high resource or written in Latin scripts, embedding structural bias and inequity into the vocabulary.

The downstream consequences are well-documented. Bostrom and Durrett (2020) showed that BPE tokens frequently diverge from linguistically motivated morpheme boundaries. More recently, Selvamurugan et al. (2025) quantified cross-lingual tokenization inequity through normalized sequence length and subword fertility, demonstrating that the gap is most pronounced for underrepresented scripts. These findings motivate moving beyond global frequency optimization as the main design criterion for multilingual tokenizers.

2.2 Parity-aware Byte-Pair Encoding

Parity-aware BPE (PA BPE; Foroutan et al., 2025) modifies Byte-level BPE by optimizing the worst-case compression rate across languages. Each merge iteration selects the pair that most improves the worst-performing language, trading marginal global efficiency for tokenization equity.

The approach requires minimal implementation changes to existing BPE pipelines. On a 30-language unbalanced dataset, it achieves a lower Gini coefficient of 0.011 versus 0.064 for Byte-level BPE, while remaining competitive on compression and outperforming or matching Byte-level BPE baselines across 13 multilingual benchmarks.

2.3 Morphology-Driven Byte Encoding

MYTE (Limisiewicz et al., 2024) replaces UTF-8's character-based convention with morpheme-based byte codes, as morphemes exhibit more consistent sequence lengths than characters across languages. It learns a per-language morpheme inventory to achieve balanced morphological coverage via Morfessor 2.0 (Smit et al., 2014), and assigns shorter byte sequences to linguistically meaningful units.

MYTE produces shorter encoding compared to UTF-8 for all 99 languages tested, with gains ranging from 1% for Vietnamese and Chinese to nearly 70% for Burmese. Its worst-case tokenizer parity relative to English is 1.7, versus 3.5 for UTF-8. MyT5, a MYTE-encoded variant of ByT5 (Xue et al., 2022), demonstrated reduced cross-language perplexity disparity compared to its byte-level counterpart. It achieves 75.3 F1 on XTREME-UP (Ruder et al., 2023) question answering versus 73.2 for ByT5.

2.4 Byte Latent Transformer

BLT (Pagnoni et al., 2025) eliminates explicit tokenization entirely and comprises three modules: a lightweight local encoder producing patches, a large latent transformer processing them, and a lightweight local decoder reconstructing bytes. An entropy model drives patch segmentation, allocating computation proportional to data complexity.

BLT enables a 50% reduction in inference FLOPs relative to Llama 3's original tokenizer without sacrificing downstream task performance. (Grattafiori et al., 2024). By avoiding a static vocabulary from tokenization, BLT sidesteps multilingual inequity that arises when high-resource language tokens dominate and outperforms Llama 3 by 2 BLEU points (Papineni et al., 2002) on translation into English.

3 Methods

We compare the three tokenizer families discussed above to a baseline Byte-level BPE tokenizer. We train all tokenizers on the same dataset to evaluate their efficiency and cross-lingual equity. We then train language models from scratch using these tokenizers and evaluate their downstream task performance. For fairness and reproducibility, data

sizes are reported in number of sentences and bytes, rather than tokens.

3.1 Training Data

For tokenizer training, we sample a total of 1 million sentences (3.5GB) across eleven SEA languages from multilingual C4 (mC4) (Xue et al., 2021). Sampling is performed randomly without replacement following the language proportions in mC4 to approximate realistic multilingual data distribution. The resulting per-language sentence counts are detailed in Appendix A.1.

For language model training, we adopt the same training dataset as Foroutan et al. (2025) and sample 100 million sentences (203 GB) from FineWeb2 (Penedo et al., 2025). This dataset size is comparable to what Foroutan et al. (2025) and Limisiewicz et al. (2024) used to train their language models. FineWeb2 is a multilingual web corpus with quality filtering already applied, and we did not apply further preprocessing before training. Language proportions are controlled using temperature sampling with $\tau = 1.21$ to boost the representation of low-resource languages (Foroutan et al., 2025). The details are provided in Appendix A.2.

Vocabulary sizes are controlled where possible to enable a fair comparison of the four tokenizers. MYTE was designed to have 4,096 morphemes per language to avoid over-segmentation. Thus, we train tokenizers at three scales: 4,096, 8,192, and 12,288 tokens per language, across all eleven SEA languages. For MYTE, this translates to total morpheme inventories of $45k$, $90k$, and $135k$ morphemes. The vocabulary sizes of Byte-level BPE and Parity-aware BPE are matched to MYTE's total morpheme counts at each scale.

BLT's patch-based representation is not directly comparable since it does not learn a fixed vocabulary. Following the approach of Pagnoni et al. (2025), we configure BLT's entropy model to yield average patch sizes of 4.5, 6, and 8 bytes per patch.

We use tokenizers with vocabulary size of $90k$ to train language models, placing them close to the $100k$–$128k$ vocabulary size of most LLM tokenizers (Wegmann et al., 2025). For BLT, we adopt the entropy model with an average patch size of 4.5 bytes, following the setup of Pagnoni et al. (2025). Note that BLT is not a tokenizer in the traditional sense, but is referred to as one here for ease of comparison.

3.2 Implementation Details

Training of tokenizers and tokenization of language model training data for MYTE and BPE-based algorithms were performed on a single AMD EPYC 9554P CPU (128 threads). For BLT, the entropy-based tokenizer was trained on 4× NVIDIA H100 GPUs, and the language model training dataset was tokenized on 8× NVIDIA H200 GPUs. Statistics of the language model training dataset after tokenization are reported in Table 1.

Tokenizer (size)	Duration (hour)	# Tokens (billion)	File size (GB)
BLT (4.5)	33	42	204
MYTE (90k)	50	269	538
PA BPE (90k)	3	82	329
BPE (90k)	3	72	288

Table 1: Statistics of language model training dataset after tokenization by the four tokenizers. Legend: size = patch size for BLT, morpheme inventory size for MYTE, vocabulary size for all other models; File Size = Size of dataset files after tokenization; PA BPE = Parity-aware BPE; BPE = Byte-level BPE.

Language model training is carried out on 4–8× NVIDIA H100/H200 GPUs. To enable a fair comparison of computational cost, training durations are converted to an 8× NVIDIA H200 equivalent, as reported in Table 2. MYTE incurs the highest training cost at 300 normalized hours due to its substantially larger token count (269B tokens), while Byte-level BPE is the most efficient at 68 hours (72B tokens). Additionally, we trained and compared all language models at an equal token count of 38B tokens as measured by their respective tokenizers. These experiments yielded the same conclusions as the models trained on the same dataset, so we omit them for the sake of brevity.

Model (size)	Duration (hour)	# Tokens (billion)
BLT (4.5)	160	42
MYTE (90k)	300	269
PA BPE (90k)	87	82
BPE (90k)	68	72

Table 2: Statistics of language model training.

3.3 Evaluation Metrics

3.3.1 Intrinsic Metrics

Quantifying tokenizer efficiency and cross-lingual fairness requires metrics that are agnostic to both language and model architecture. We identify three such metrics from recent literature and provide brief descriptions below. Detailed definitions and formulae can be found in Appendix B.

Tokenizer parity measures the ratio of the number of tokens per sentence in a given language relative to English (Petrov et al., 2023). A *tokenizer parity close to 1* indicates that the tokenizer imposes roughly equal computational cost across the given language and English.

Gini coefficient adapts the income inequality measure to the domain of tokenization fairness (Foroutan et al., 2025). It quantifies the distribution of per-language tokenization costs, with values ranging from 0 (perfect equality) to 1 (maximal inequality). A *lower Gini coefficient* reflects a more equitable tokenizer.

Compression rate measures how efficiently a tokenizer compresses text (Foroutan et al., 2025). A *higher compression rate* indicates that the tokenizer is more efficient and produces fewer tokens for the same amount of text.

3.3.2 Extrinsic Metrics

We evaluate trained language models on English and multilingual classification benchmarks using Language Model Evaluation Harness (Biderman et al., 2024) with zero-shot prompting. Details of the benchmarks can be found in Appendix C. For machine translation, we evaluate fine-tuned models with five-shot prompts drawn from the training dataset of the multi-way parallel FLORES+ corpus (Costa-jussà et al., 2024), following the setup of Limisiewicz et al. (2024).

To assess English language understanding, models are evaluated on three English classification benchmarks used by Pagnoni et al. (2025): PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), and Arc-C (Clark et al., 2018). These benchmarks test commonsense reasoning, sentence completion, and science question answering respectively.

Cross-lingual generalization is assessed through three multilingual classification benchmarks used by Foroutan et al. (2025). XNLI (Conneau et al., 2018) evaluates natural language inference across multiple languages, XCOPA (Ponti et al., 2020) tests causal commonsense reasoning in a multilingual setting, and XStoryCloze (Lin et al., 2022) assesses story completion across languages.

Machine translation is assessed after fine-tuning via continual pre-training to ensure that results reflect the task-adapted performance of the models. Fine-tuning details and the resulting per-language sentence counts of the fine-tuning dataset are provided in Appendix A.3. BLEU (Papineni et al., 2002) and chrF (Popović, 2015) scores are computed in both EN $\rightarrow$ XX and XX $\rightarrow$ EN directions for all ten non-English SEA languages. These two metrics range from 0 to 100 and higher scores indicate better translation quality.

4 Experiments

4.1 Intrinsic Evaluation

We evaluate the trained tokenizers on FLORES+ devtest set, which consists of 1,012 aligned sentences across all eleven SEA languages. For Parity-aware BPE, we train the base variant with the FLORES+ training dataset as the development corpus following the setup of Foroutan et al. (2025). It is the only tokenizer among those evaluated that requires parallel data during training.

4.2 Extrinsic Evaluation

4.2.1 Language Model

The base architecture of our language model is OLMo-2-1B (Walsh et al., 2025), a decoder-only transformer comprising 16 layers, a hidden dimension of 2,048, and approximately 1.5 billion parameters. We train all models from scratch to ensure that differences in downstream task performance are largely attributed to tokenizer choice.

4.2.2 Training Configuration

Models are trained using the AdamW (Loshchilov and Hutter, 2019) optimizer with a peak learning rate of 4.0×10^{-4}, weight decay of 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and gradient clipping of 1.0. The learning rate follows a Warmup-Stable-Decay (WSD) schedule (Hu et al., 2024): a linear warmup over the first 1% of training tokens, a stable phase over the next 89%, and a linear decay over the final 10%. The global batch size is 512 sequences with a maximum sequence length of 4,096 tokens, corresponding to approximately 2 million tokens per training step.

4.2.3 Statistical Significance

Extrinsic metrics are assessed for statistical significance using paired bootstrap resampling with

1,000 iterations at $p < 0.05$ (Koehn, 2004). This ensures that reported performance differences between models reflect meaningful systematic effects rather than sampling variation across examples.

5 Results

5.1 Intrinsic Evaluation

Tokenizer (size)	CR	Gini	TP
BLT (4.5)	0.0127	0.212	2.87
BLT (6)	0.0145	0.219	2.96
BLT (8)	0.0161	0.227	3.06
MYTE (45k)	0.0085	0.085	<u>1.23</u>
MYTE (90k)	0.0089	0.086	<u>1.23</u>
MYTE (135k)	0.0089	0.095	1.26
PA BPE (45k)	0.0250	**0.021**	**1.15**
PA BPE (90k)	0.0272	<u>0.028</u>	1.24
PA BPE (135k)	0.0280	0.029	1.25
BPE (45k)	0.0257	0.243	1.93
BPE (90k)	<u>0.0293</u>	0.220	1.73
BPE (135k)	**0.0314**	0.203	1.61

Table 3: Intrinsic evaluation of tokenizers on identical training data. Compression rate and tokenizer parity values are macro-averaged across languages. The best result is in **bold** and the second-best is <u>underlined</u>. Legend: CR = Compression Rate, TP = Tokenizer Parity.

Table 3 reveals that Parity-aware BPE achieves the lowest Gini coefficient across all vocabulary sizes. This equity gain does not come at the cost of tokenizer efficiency, as Parity-aware BPE achieves competitive compression rates relative to Byte-level BPE.

MYTE has a relatively low Gini coefficient and tokenizer parity, but its compression rate is the worst. This indicates that morpheme-based segmentation produces longer token sequences. The trade-off is consistent with the inflated token counts observed during language model training.

BLT exhibits poor equity across all vocabulary sizes despite operating at the byte level without a fixed vocabulary. Its tokenizer parity is the highest among all approaches, suggesting that entropy-driven patch segmentation provides no built-in mechanism to correct for corpus imbalances.

Figure 1 situates each tokenizer family within a two-dimensional efficiency-equity space, where the ideal direction is toward a higher compression rate and lower Gini coefficient (bottom-right of the plot). Parity-aware BPE lies on the Pareto front of the efficiency-equity space across all vocabulary sizes, highlighting that cross-lingual fairness and compression efficiency are not at odds. Byte-level BPE at its largest vocabulary size of $135k$ lies on the Pareto front, as it has the highest compression rate. On the other hand, both BLT and MYTE are Pareto-dominated.

5.2 Extrinsic Evaluation

5.2.1 English Classification Benchmarks

Model (size)	PIQA (50.00)	HellaSwag (25.00)	Arc-C (25.00)
BLT (4.5)	66.10	44.81	24.74
MYTE (90k)	67.14	44.47	<u>27.39</u>
PA BPE (90k)	<u>71.55</u>	<u>53.22</u>	26.19
BPE (90k)	**72.31**	**54.42**	**28.41**

Table 4: Accuracies on English classification benchmarks. Parenthesized values indicate the expected accuracy of a random classifier. The highest score is in **bold** and the second-highest is <u>underlined</u>.

Table 4 shows that Byte-level BPE achieves the highest scores on all three English classification benchmarks. Its advantage is statistically significant across all comparisons, except over Parity-aware BPE on PIQA and MYTE on Arc-C, where its numerical lead does not reach significance. Parity-aware BPE is a strong runner-up, performing significantly better than BLT and MYTE on both PIQA and HellaSwag. For commonsense reasoning and sentence completion benchmarks, BPE-based tokenizers have a consistent advantage over alternative approaches in our evaluation setting.

5.2.2 Multilingual Classification Benchmarks

Table 5 reveals task-dependent performance across tokenizers, with no consistent winner across all three benchmarks. MYTE significantly outperforms all other tokenizers on XNLI, consistent with its morphological representations providing richer cross-lingual semantic signals for inference. Byte-level BPE achieves the highest scores on XCOPA and XStoryCloze, suggesting that its efficiency is best leveraged on tasks requiring causal and narrative reasoning.

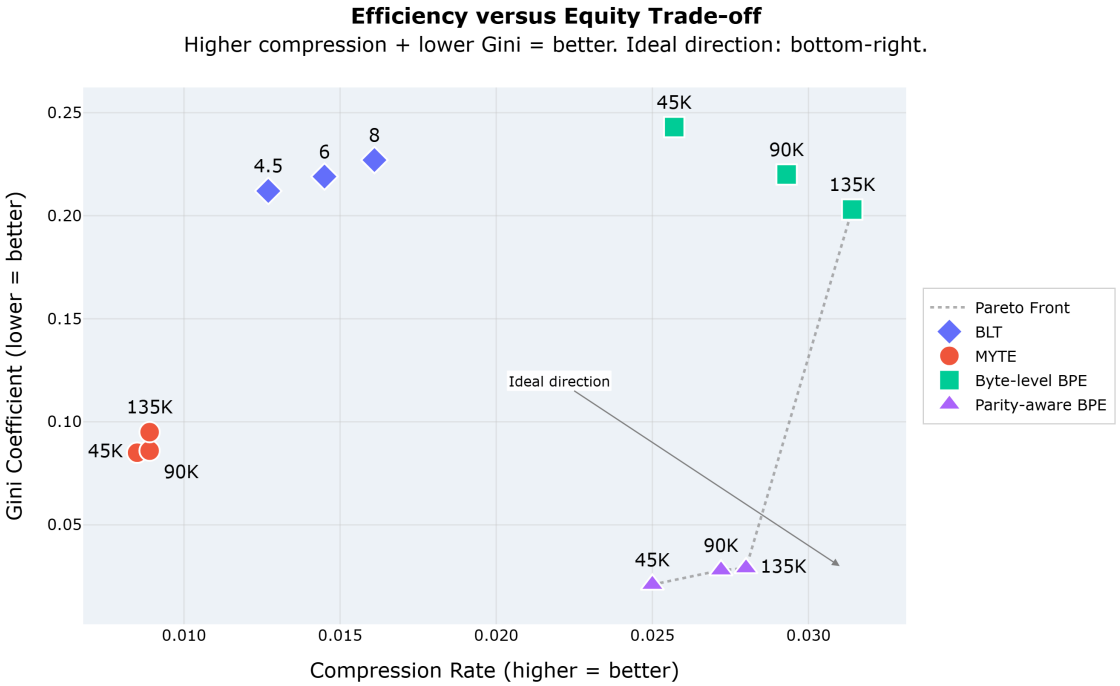

Figure 1: Efficiency-equity Pareto front of the evaluated tokenizers. Values beside markers indicate the patch size for BLT, morpheme inventory size for MYTE, and vocabulary size for BPE and PA BPE.

Model (size)	XNLI (33.33)	XCOPA (50.00)	XStoryCloze (50.00)
BLT (4.5)	36.29	53.30	56.52
MYTE (90k)	**42.49**	54.93	50.55
PA BPE (90k)	40.43	<u>58.70</u>	<u>56.70</u>
BPE (90k)	<u>40.56</u>	**61.03**	**57.18**

Table 5: Averaged per-language accuracies across multilingual classification benchmarks. Parenthesized values indicate the expected accuracy of a random classifier. The highest score is in **bold** and the second-highest is <u>underlined</u>. Detailed results for each language are reported in Appendix D.1.

Model (size)	EN → XX	XX → EN
BLT (4.5)	10.82	11.70
MYTE (90k)	**14.77**	**13.81**
PA BPE (90k)	11.36	12.19
BPE (90k)	<u>13.39</u>	<u>12.78</u>

Table 6: Averaged BLEU scores across ten SEA languages. The highest score is in **bold** and the second-highest is <u>underlined</u>.

5.2.3 Machine Translation

Table 6 aggregates the BLEU translation scores across ten SEA languages. We also provide the chrF scores in Appendix D.2.2. MYTE achieves the highest BLEU scores in both translation directions. A consistent directional asymmetry is observed across both metrics for MYTE, where it is systematically stronger in EN → XX translation than XX → EN. MYTE's morpheme-level segmentation enables finer-grained generation of morphologically complex target word forms in SEA languages, an advantage that narrows when translating into English.

6 Analysis

6.1 Effect of Scaling Vocabulary Size

Increasing vocabulary size produces distinct behavior across tokenizers, as seen in Table 3. Byte-level BPE improves consistently across both efficiency and cross-lingual fairness with a larger vocabulary size. In contrast, BLT gains in efficiency but sacrifices cross-lingual equity as vocabulary size grows.

MYTE is relatively insensitive to morpheme inventory size scaling within the evaluated range. Compression rate and tokenizer parity remain nearly constant across all three morpheme inventory sizes, indicating that its morphological segmentation approach saturates at around 4,096 morphemes per language.

Parity-aware BPE becomes more inequitable as vocabulary size increases. While it achieves the lowest Gini coefficient among all models, both

its Gini coefficient and tokenizer parity worsen at larger vocabulary sizes, increasing to 0.029 and 1.25 respectively at $135k$ vocabulary size.

6.2 Fairness Regression in Parity-aware BPE

It seems counterintuitive that Parity-aware BPE tokenizers are less equitable as vocabulary size increases. To investigate this, we analyzed per-language token counts produced by Parity-aware BPE tokenizers of different vocabulary sizes. We use the FLORES+ training dataset (rather than the test dataset) because it serves as the development corpus during tokenizer training. Examining this dataset would isolate the effect of vocabulary scaling and avoid confounding factors from unseen data such as differing vocabulary distributions.

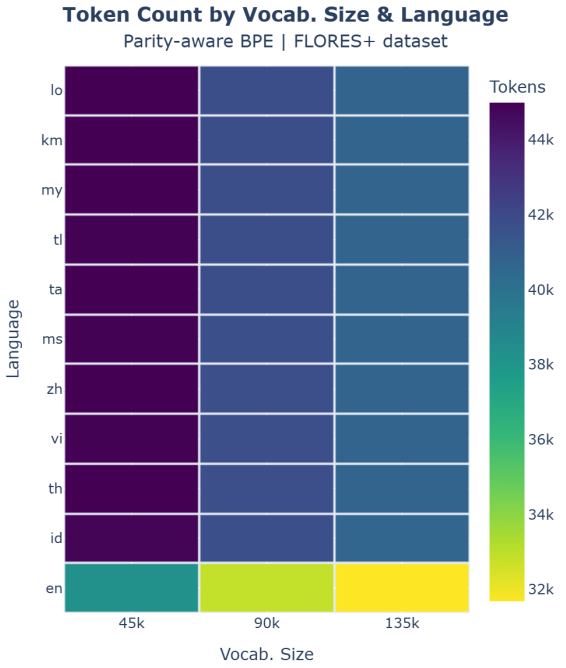

Figure 2: Per-language token counts on the FLORES+ training dataset by Parity-aware BPE tokenizers of varying vocabulary sizes.

Figure 2 shows that increasing the vocabulary size results in a larger reduction in the tokens required for English sentences compared to the other SEA languages. This results in token count disparity between English and the other SEA languages to widen by 36% as vocabulary size scales from $45k$ to $135k$, causing tokenizer parity to increase (i.e., worsen). We observe that Parity-aware BPE only limits worst-case per-language tokenizer parity, so its fairness mechanism does not prevent English from acquiring more merges as vocabulary size increases.

6.3 Effect of Tokenizer Choice and Script Type

To investigate whether tokenizer parity outcomes are driven by the choice of tokenizer, the script type of the target language, or their interaction, we apply a two-way mixed ANOVA test (Meyers et al., 2009). Running separate pairwise t-tests for each tokenizer–script combination would inflate the Type I error rate multiplicatively.

The between-subject factor is script type, categorized as Latin (Indonesian, Malay, Tagalog, Vietnamese) or Abugida (Burmese, Khmer, Lao, Tamil, Thai) according to Limisiewicz et al. (2024). Chinese is excluded from this comparison as it is the only CJK-script language. The within-subject factor is tokenizer choice, as measuring the same language across tokenizers introduces within-subject correlation. We assess the effect of these factors on per-language tokenizer parity, measure statistical significance at $\alpha = 0.05$, and report partial η^2 as the effect size measure.

Tokenizer (size)	Script type	
	Latin	Abugida
BLT (4.5)	2.36	3.38
MYTE (90k)	**1.19**	<u>1.32</u>
PA BPE (90k)	<u>1.26</u>	**1.22**
BPE (90k)	1.30	2.18

Table 7: Per-language tokenizer parity, macro-averaged by tokenizer and script type. Latin scripts exclude English. The lowest value is in **bold** and the second-lowest is <u>underlined</u>.

Tokenizer choice is the only statistically significant factor affecting tokenizer parity ($p < 0.001$, partial $\eta^2 = 0.752$). The large effect size indicates that tokenizer choice explains the majority of variance in tokenizer parity, regardless of script type. At the same time, script type alone does not reach significance ($p = 0.090$). As shown in Table 7, Parity-aware BPE and MYTE achieve near-uniform tokenizer parity across both script types. In contrast, Byte-level BPE imposes a 1.68× tokenizer parity penalty on Abugida scripts relative to Latin scripts. BLT also exhibits a high cross-script disparity (1.43×), consistent with its entropy model being undertrained on non-Latin scripts.

Fundamentally, tokenizer parity directly determines inference cost. A tokenizer parity of k for a

given language implies that a user pays k times the per-token API cost relative to English to process the same semantic content. Under Byte-level BPE, Abugida-script users incur 1.68× higher costs on average than English users for semantically equivalent prompts. These results confirm that equitable tokenizer design, and not script similarity to English, is the main determinant of whether a tokenizer imposes uniform computational costs across languages.

7 Conclusion

We present the first systematic, dataset-controlled comparison of BLT, MYTE, Parity-aware BPE, and Byte-level BPE across eleven SEA languages, evaluating tokenizer equity, compression efficiency, and downstream task performance under the same experimental conditions. Our intrinsic evaluation demonstrates that cross-lingual equity and tokenization efficiency are not fundamentally at odds.

Among the equitable tokenizers analyzed, MYTE delivers the strongest semantic inference and machine translation performance through richer morphological representations, though at the cost of a higher computational budget and lower compression efficiency. Despite its architectural novelty in eliminating fixed tokenizer vocabulary, we found that BLT underperforms on downstream tasks as its entropy model receives insufficient exposure to low-resource languages under realistic multilingual data distribution.

The appropriate choice of tokenizer is ultimately use-case dependent. We recommend Parity-aware BPE as a responsible default for multilingual models targeting SEA languages, given its favorable position in the efficiency-equity space and relatively strong downstream task performance. However, the base variant of Parity-aware BPE requires multi-way parallel data, which may be scarce or unavailable in low-resource settings (Foroutan et al., 2025). MYTE is preferred when morphology and translation are critical, and the computational budget permits the associated training overhead.

Our ANOVA results establish that tokenizer choice is the primary factor affecting tokenizer parity. The difference in inference costs between English and SEA-language users can be addressed through the choice of the tokenizer. Equitable tokenization has direct, quantifiable consequences for the 671 million speakers across Southeast Asia (Lovenia et al., 2024), for whom token-priced APIs create unequal access costs. How tokenization is carried out across languages shapes the economic accessibility of multilingual models for underrepresented language communities.

Limitations

First, all language models are trained at the 1.5B-parameter scale due to computational resource constraints. We leave the investigation of larger model sizes to future work. Next, BLT cannot be scaled along the same vocabulary dimension as the other tokenizers since it operates without a fixed vocabulary, which prevents direct matched-vocabulary size comparison. Finally, we evaluate only base pretrained models. We believe this is sufficient, as supervised fine-tuning or alignment training does not affect the tokenizer's fairness or efficiency.

We do not anticipate any immediate societal or individual harm arising from this work. Nevertheless, we advise users to exercise caution, as our models have not been subjected to safety or value alignment procedures.

References

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of SIGUL*, pages 1–9.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of AAAI*, pages 7432–7439.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of EMNLP*, pages 4617–4624.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and

Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.

The Unicode Consortium. 2011. The Unicode standard. Technical Report Version 6.0.0, Unicode Consortium.

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, John Hoffman, and 19 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, pages 841–846.

Negar Foroutan, Clara Meister, Debjit Paul, Joel Niklaus, Sina Ahmadi, Antoine Bosselut, and Rico Sennrich. 2025. Parity-aware Byte-Pair Encoding: Improving cross-lingual fairness in tokenization. *arXiv preprint arXiv:2508.04796*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Sheriff Issaka, Erick Rosas Gonzalez, Lieqi Liu, Evans Kofi Agyei, Lucas Bandarkar, Nanyun Peng, David Ifeoluwa Adelani, Francisco Guzmán, and Saadia Gabriel. 2026. Translation as a scalable proxy for multilingual evaluation. *arXiv preprint arXiv:2601.11778*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of ACL*, pages 15059–15076.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of EMNLP*, pages 9019–9052.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of EMNLP*, pages 5155–5203.

Lawrence S. Meyers, Glenn Gamst, and A. J. Guarino. 2009. *Two-way mixed ANOVA design*, page 253–266. Cambridge University Press.

Tan Sang Nguyen, Muhammad Reza Qorib, and Hwee Tou Ng. 2026. OpenSeal: Good, fast, and cheap construction of an open-source Southeast Asian LLM via parallel data. *arXiv preprint arXiv:2602.02266*.

OpenAI. 2025. OpenAI/Tiktoken: Tiktoken is a fast BPE tokeniser for use with OpenAI's models.

Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason E Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srini Iyer. 2025. Byte Latent Transformer: Patches scale better than tokens. In *Proceedings of ACL*, pages 9238–9258.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Proceedings of NeurIPS*, pages 36963–36990.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP*, pages 2362–2376.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of WMT*, pages 392–395.

Muhammad Reza Qorib, Junyi Li, and Hwee Tou Ng. 2025. Just go parallel: Improving the multilingual capabilities of large language models. In *Proceedings of ACL*, pages 33411–33424.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, and 8 others. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of EMNLP*, pages 1856–1884.

Aishwarya Selvamurugan, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. From bias to balance: How multilingual dataset composition affects tokenizer performance across languages. In *Workshop on LM4UC*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of EACL*, pages 21–24.

Sagar Tamang and Dibya Jyoti Bora. 2024. Evaluating tokenizer performance of large language models across official Indian languages. *arXiv preprint arXiv:2411.12240*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. 2 OLMo 2 furious. In *Proceedings of COLM*.

Anna Wegmann, Dong Nguyen, and David Jurgens. 2025. Tokenization is sensitive to language variation. In *Findings of ACL*, pages 10958–10983.

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 ACL*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*, pages 483–498.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of ACL*, pages 4791–4800.

A Training Data Details

A.1 Tokenizer Training Data

ISO Code	Language	# Sentences
en	English	604,793
id	Indonesian	129,906
th	Thai	115,159
vi	Vietnamese	90,470
zh	Chinese	25,683
ms	Malay	21,872
ta	Tamil	5,799
tl	Tagalog	3,480
my	Burmese	1,349
km	Khmer	1,253
lo	Lao	236
Total		**1,000,000**

Table 8: Per-language sentence counts in the tokenizer training dataset.

A.2 Language Model Training Data

We source pretraining data from the FineWeb2 corpus (Penedo et al., 2025), randomly sampling a subset of 100 million sentences (203 GB) spanning all eleven SEA languages. To balance coverage between high-resource and low-resource languages, we control the language proportions via temperature sampling, following the approach of Foroutan et al. (2025).

Temperature Sampling Sampling each language proportionally to its word count in FineWeb2 would overwhelmingly favor English at 94.3%. As such, we sample according to a temperature-scaled probability:

$$p(L) \propto |L|^{1/\tau}, \tag{1}$$

where $p(L)$ is the probability of sampling text from language L during pre-training, $|L|$ is the number of words in that language in the corpus, and τ is a temperature parameter. When $\tau = 1$, sampling is purely proportional to word frequency. As τ increases, the distribution becomes increasingly uniform, thereby boosting the relative sampling probability of low-resource languages.

We configure $\tau = 1.21$ so that English constitutes 89.7% of the resulting training sentences, matching the proportion of English data used in the Llama 2 pretraining dataset (Touvron et al., 2023). Table 9 shows the raw word frequency of each language in FineWeb2, along with the adjusted frequency after temperature sampling is applied. For each language L, we compute $|L|^{1/\tau}$ and normalize across all eleven SEA languages to obtain the final sampling proportion. These proportions are then used to determine the number of sentences drawn from each language in our 100M-sentence subset (Table 10).

A.3 Machine Translation Fine-tuning

Machine translation fine-tuning is performed via continual pre-training on English-XX parallel sentences for one epoch. The fine-tuning dataset consists of up to 13 million parallel sentence pairs per language, which were randomly sampled without replacement from NLLB (Costa-jussà et al., 2024) where possible, following the approach of Nguyen et al. (2026). The resulting per-language sentence counts are shown in Table 11.

Each parallel sentence pair is formatted using the template by Qorib et al. (2025): `"{source language}: {source sentence} \n {target language}: {target sentence}"`. To prevent the model from developing a bias toward a fixed source language, the language order of each pair was randomized independently with equal probability. This means each example has an equal chance of being presented as English-first or target language-first.

B Intrinsic Tokenizer Evaluation Metrics

B.1 Cross-lingual Equity Metrics

Tokenizer parity measures the ratio of the average number of tokens per sentence in a given language relative to English (Petrov et al., 2023). For a specific language L with k aligned sentences, we can compute its average number of tokens per sentence, $average\_tokens_L$:

$$\frac{1}{k}\sum_{i=1}^{k} \text{\# tokens in sentence } i \tag{2}$$

The tokenizer parity for language L, $parity_L$, is defined as:

$$\frac{average\_tokens_L}{average\_tokens_{English}} \tag{3}$$

The macro-average tokenizer parity for all n non-English languages is computed as:

$$\frac{1}{n}\sum_{j=1}^{n} parity_j \tag{4}$$

Language	Word frequency (billion), $\lvert L\rvert$	Relative frequency, $\lvert L\rvert^{1/\tau}$	Proportion
English	11,500.0	2269.6	89.68%
Chinese	543.5	182.2	7.20%
Indonesian	60.3	29.6	1.17%
Vietnamese	50.9	25.7	1.02%
Thai	24.7	14.1	0.56%
Malay	5.6	4 .2	0.17%
Tamil	1.9	1.7	0.07%
Tagalog	1.6	1.5	0.06%
Burmese	0.9	0.9	0.03%
Khmer	0.7	0.7	0.03%
Lao	0.2	0.3	0.01%
Total	**12,190.3**	**2,530.5**	**100.00%**

Table 9: Word frequencies in FineWeb2 and relative frequencies after temperature sampling ($\tau = 1.21$). Temperature sampling boosts the relative proportion of low-resource SEA languages while keeping English at 89.7%, matching Llama 2's pretraining data distribution.

ISO Code	Language	# Sentences
en	English	89,689,566
zh	Chinese	7,199,747
id	Indonesian	1,169,277
vi	Vietnamese	1,016,743
th	Thai	558,780
ms	Malay	165,279
ta	Tamil	68,252
tl	Tagalog	59,364
my	Burmese	34,699
km	Khmer	28,295
lo	Lao	9,998
Total		**100,000,000**

Table 10: Per-language sentence counts in the language model training dataset after temperature sampling.

ISO Code	Language	# Sentences (million)
zh	Chinese	13.0
id	Indonesian	13.0
vi	Vietnamese	13.0
th	Thai	13.0
ms	Malay	13.0
ta	Tamil	13.0
tl	Tagalog	13.0
my	Burmese	10.0
km	Khmer	5.8
lo	Lao	4.2
Total		**111.0**

Table 11: Per-language sentence counts in the fine-tuning dataset.

This ratio measures whether tokenizers impose computational costs unequally across languages. A macro-average tokenizer parity *closer to 1* indicates a more equitable tokenizer across languages.

Gini coefficient assesses tokenization equity by treating token costs as a distribution (Foroutan et al., 2025). The token cost for a language, c, is defined as the average number of tokens per sentence for the language in the parallel corpus. For token costs $c_1 \leq c_2 \leq \cdots \leq c_n$ across n languages, the Gini coefficient is computed as:

$$\frac{1}{n}\left(n + 1 - 2\,\frac{\sum_{i=1}^{n}(n+1-i)\,c_i}{\sum_{i=1}^{n} c_i}\right) \quad (5)$$

Values range from 0 to 1. A *lower Gini coefficient* (closer to 0) indicates a more equitable tokenizer across languages.

B.2 Tokenizer Efficiency Metrics

Compression rate measures how efficiently a tokenizer compresses text. It is defined as the average of the inverse token count per sentence (Foroutan et al., 2025). For a specific language L, its compression rate, $rate_L$ is computed as:

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\text{\# tokens in sentence } i} \quad (6)$$

where k is the number of aligned sentences for language L in the parallel corpus. Essentially, we compute a language's compression rate by evaluating the inverse of the number of tokens per sentence and then averaging it.

The macro-average compression rate for all n languages is computed as:

$$\frac{1}{n}\sum_{j=1}^{n} rate_j \tag{7}$$

Utilizing a parallel corpus controls for semantic differences, by comparing token counts over semantically equivalent content. A *higher macro-average compression rate* indicates a more efficient tokenizer across languages. This metric is informative when viewed alongside tokenizer parity, as a high overall compression rate can mask under-compression of individual low-resource languages.

C Extrinsic Metrics

C.1 English Classification Benchmarks

Classification benchmarks evaluate a model's ability to understand, analyze, and select the correct category from a set of options. The following English classification benchmarks are used by Pagnoni et al. (2025) to evaluate a model's commonsense reasoning and general world knowledge.

Physical Intuition Question Answering (PIQA) (Bisk et al., 2020) probes a model's understanding of everyday physical interactions and how objects behave in the real world. Each example presents a goal and two solution candidates, with the model tasked to identify the more physically plausible option.

HellaSwag (Zellers et al., 2019) is a commonsense natural language inference benchmark where a model must select the most plausible continuation of a given scenario from four candidate endings. The dataset is constructed using adversarial filtering to ensure that a model possesses genuine contextual understanding.

Arc-Challenge (Arc-C) (Clark et al., 2018) evaluates a model's scientific reasoning ability through multiple-choice questions drawn from grade-school science exams. The Challenge subset selects questions that simple retrieval-based and word co-occurrence methods fail to answer correctly, making it a reliable indicator of deeper reasoning capabilities.

C.2 Multilingual Classification Benchmarks

The following multilingual classification benchmarks are used by Foroutan et al. (2025) and they collectively span several of our target languages. They enable a comprehensive evaluation of a model's cross-lingual performance on SEA languages.

Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018) extends the MultiNLI dataset to 15 languages and serves as a standard benchmark for cross-lingual natural language understanding. Models must classify the logical relationship between each pair as one of three categories: entailment, contradiction, or neutral. This benchmark covers English, Chinese, Thai, and Vietnamese.

Cross-lingual Choice of Plausible Alternatives (XCOPA) (Ponti et al., 2020) is a multilingual benchmark targeting causal commonsense reasoning. Given a premise, a model must identify either the most plausible cause or effect from two candidate sentences. XCOPA is evaluated in a zero-shot setting to assess cross-lingual transfer without fine-tuning. This benchmark covers English, Chinese, Indonesian, Tamil, Thai, and Vietnamese.

XStoryCloze (Lin et al., 2022) requires a model to select the correct ending for a four-sentence narrative from two candidate conclusions in a multilingual setting. It evaluates cross-lingual narrative understanding and commonsense reasoning. This benchmark covers English, Burmese, Chinese, and Indonesian.

C.3 Machine Translation

Machine translation is a natural benchmark for evaluating multilingual LLMs, as it tests a model's ability to understand and generate text across languages. For SEA languages, translation quality serves as a proxy for how well a model has internalized low-resource linguistic structure (Issaka et al., 2026).

Machine translation performance is measured by comparing the machine-generated output to human reference translations. Two complementary metrics are typically used, BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and chrF (character-level F-score) (Popović, 2015). Both metrics have scores ranging from 0 to 100, with *higher scores* indicating better translation quality. BLEU measures word-level n-gram precision against reference translations and was used

by Pagnoni et al. (2025). chrF computes character n-gram F-score and is suitable for morphologically rich languages where word-level overlap may be sparse and was used by Limisiewicz et al. (2024).

The detailed BLEU and chrF translation scores are shown in Appendix D.2. The original MYTE paper (Limisiewicz et al., 2024) reports scores only for English-to-Vietnamese and English-to-Tamil translation among SEA languages, and our scores are higher than their reported scores.

D Detailed Results

D.1 Multilingual Classification Benchmarks

Model (size)	en	zh	vi	th	AVG
BLT (4.5)	42.10	33.99	34.29	34.79	36.29
MYTE (90k)	50.00	33.99	44.99	40.98	42.49
PA BPE (90k)	47.94	33.91	43.07	36.81	40.43
BPE (90k)	49.30	33.51	43.09	36.35	40.56

Table 12: Per-language XNLI scores. Expected accuracy of a random classifier = 33.33.

Model (size)	en	zh	id	vi	th	ta	AVG
BLT (4.5)	64.20	52.40	52.60	48.60	52.60	49.40	53.30
MYTE (90k)	54.40	51.00	55.80	56.60	55.00	56.80	54.93
PA BPE (90k)	71.40	55.60	58.60	60.20	53.40	53.00	58.70
BPE (90k)	71.60	59.00	61.60	62.80	56.20	55.00	61.03

Table 13: Per-language XCOPA scores. Expected accuracy of a random classifier = 50.00.

Model (size)	en	zh	id	my	AVG
BLT (4.5)	65.45	54.20	55.06	51.36	56.52
MYTE (90k)	52.95	49.83	50.89	48.51	50.55
PA BPE (90k)	65.25	55.06	55.92	50.56	56.70
BPE (90k)	65.78	54.80	57.64	50.50	57.18

Table 14: Per-language XStoryCloze scores. Expected accuracy of a random classifier = 50.00.

D.2 Machine Translation

D.2.1 BLEU scores

Model (size)	zh	id	vi	th	ms	ta	tl	my	km	lo	AVG
BLT (4.5)	9.18	27.01	19.30	7.24	24.98	3.65	11.04	1.92	2.31	1.53	10.82
MYTE (90k)	18.43	34.92	25.36	9.25	27.66	5.61	16.73	2.49	3.40	3.90	14.77
PA BPE (90k)	22.78	26.47	18.30	5.64	21.04	4.38	9.11	1.30	2.56	2.01	11.36
BPE (90k)	30.20	27.72	32.51	6.90	23.10	2.56	8.05	0.58	1.45	0.82	13.39

Table 15: Per-language BLEU scores (EN → XX).

Model (size)	zh	id	vi	th	ms	ta	tl	my	km	lo	AVG
BLT (4.5)	10.33	25.63	21.94	8.47	18.30	4.65	17.10	3.92	4.31	2.33	11.70
MYTE (90k)	16.65	30.96	18.71	13.70	23.90	3.76	21.23	3.87	2.40	2.90	13.81
PA BPE (90k)	14.75	26.06	22.82	12.17	17.29	5.35	16.73	2.15	2.10	2.47	12.19
BPE (90k)	14.42	29.75	23.24	11.26	19.20	5.62	15.05	2.39	3.13	3.77	12.78

Table 16: Per-language BLEU scores (XX → EN).

D.2.2 chrF scores

Model (size)	zh	id	vi	th	ms	ta	tl	my	km	lo	AVG
BLT (4.5)	13.16	48.75	42.03	30.48	45.46	31.52	31.35	19.91	19.92	20.63	30.32
MYTE (90k)	44.13	59.82	54.22	40.37	47.27	26.22	42.02	22.46	26.29	25.89	38.87
PA BPE (90k)	15.29	57.87	46.46	24.16	51.92	32.62	44.76	20.00	17.82	19.58	33.05
BPE (90k)	27.86	67.47	54.80	30.81	62.05	30.07	51.56	14.89	15.87	13.80	36.92

Table 17: Per-language chrF scores (EN → XX).

Model (size)	zh	id	vi	th	ms	ta	tl	my	km	lo	AVG
BLT (4.5)	31.26	54.78	43.12	34.10	44.98	23.80	36.71	19.33	19.54	18.05	32.57
MYTE (90k)	28.18	60.04	40.16	31.13	59.96	23.45	42.58	19.38	20.56	20.32	34.58
PA BPE (90k)	41.60	51.99	50.27	33.43	42.87	25.37	36.85	20.05	20.66	21.99	34.51
BPE (90k)	43.16	62.22	53.37	35.38	45.52	28.89	38.39	21.93	24.15	24.30	37.73

Table 18: Per-language chrF scores (XX → EN).