Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00047-5

Review Article

Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation

Cheong Yik Sheng [1, ✉], Cheong Yee Xin [1], Tan Guan Jie [2]

1. *School of Computer Science & Technology, Beijing Institute of Technology, Beijing, CHN*
2. *Department of Information Media, Faculty of Information Technology, Kanagawa Institute of Technology, Atsugi, JPN*

Received: March 11, 2026 | Review began: March 16, 2026 | Review ended: March 24, 2026 | Published: March 30, 2026

Abstract

This paper proposes a conceptual framework to analyze the fundamental trade-offs in adapting Large Language Models (LLMs) to the linguistic diversity of Southeast Asia. Rather than providing new empirical benchmarks, this analytical review introduces Embedding Sparsity Risk (ESR) as a theoretical diagnostic indicator. ESR serves as a conceptual lens for understanding how aggressive vocabulary expansion may lead to under-trained, inactive embedding parameters (often colloquially referred to as "zombie parameters"). By applying this analytical lens, we systematically evaluate three prevailing adaptation pathways: vocabulary expansion, regional full-stack pre-training, and emerging token-free architectures (e.g., MambaByte, Byte Latent Transformer). Drawing on recent multilingual model reports and tokenizer statistics from regional LLM projects, this review clarifies how hardware constraints, licensing thresholds, and linguistic characteristics intersect to shape deployment strategies. To navigate these challenges, we introduce a tiered adaptation roadmap and an open research agenda, shifting the evaluation focus from simple token fertility reduction to optimizing effective parameter utilization and representational robustness in resource-constrained environments.

Categories: AI applications, Computational Linguistics , Natural Language Processing (NLP)

Keywords: tokenization bottleneck, embedding sparsity risk, vocabulary expansion, resource-constrained adaptation, asean languages, large language models

Introduction And Background

Since the advent of the Transformer architecture, and particularly with the release of foundation models such as GPT-4, Llama 3, and Mistral, the field of natural language processing (NLP) has undergone a paradigm shift [1,2]. These models have demonstrated strong few-shot and even zero-shot learning capabilities, suggesting the potential approach of Artificial General Intelligence. However, this prosperity conceals a fundamental structural defect: a high concentration of English in pre-training corpora. In the pre-training phase of the vast majority of open-source and closed-source models, English comprises over 90% of the corpus, while other languages - especially those of the Global South - are marginalized as "low-resource languages" [3].

In Southeast Asia, this issue is particularly acute. The Association of Southeast Asian Nations (ASEAN) region is home to over 680 million people with an extremely rich linguistic and cultural tapestry, covering the Austronesian language family (e.g., Malay, Indonesian, Tagalog), the Kra-Dai family (e.g., Thai, Lao), the Austroasiatic family (e.g., Vietnamese, Khmer), and the Sino-Tibetan family (e.g., Burmese). These languages differ significantly from Indo-European languages not only

in their syntactic structures but also in their writing systems. For instance, Thai and Lao use abugida scripts written without inter-word spacing (scriptio continua), while Vietnamese, although using the Latin script, has a complex system of tonal diacritics and a syllable-based writing convention.

The first-generation bottleneck: The "Language Tax" (2020-2024)

When early foundation models (e.g., Llama 2) [4] are applied to Southeast Asian languages, a primary systems-level constraint emerges: the structural inefficiency of English-dominant tokenizers. The tokenizer functions as the interface between natural language and numerical sequences. Standard Byte-Pair Encoding (BPE) [5] implementations - typically constrained to vocabularies of approximately 32k tokens optimized for English and code-heavy corpora - exhibit morphological incompatibility with languages such as Thai or Khmer.

As a result, semantically atomic units are not mapped to dedicated tokens but are decomposed into fragmented subword units or UTF-8 byte sequences via fallback mechanisms. This phenomenon, referred to here as the Tokenization Bottleneck, introduces measurable efficiency loss across three dimensions: economic, latency, and context capacity. As illustrated in Figure *1*, this structural inefficiency has driven an evolution in adaptation challenges, shifting from basic efficiency taxes to deeper quality and sovereignty crises.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5.
DOI https://doi.org/10.7759/s44389-026-00047-5

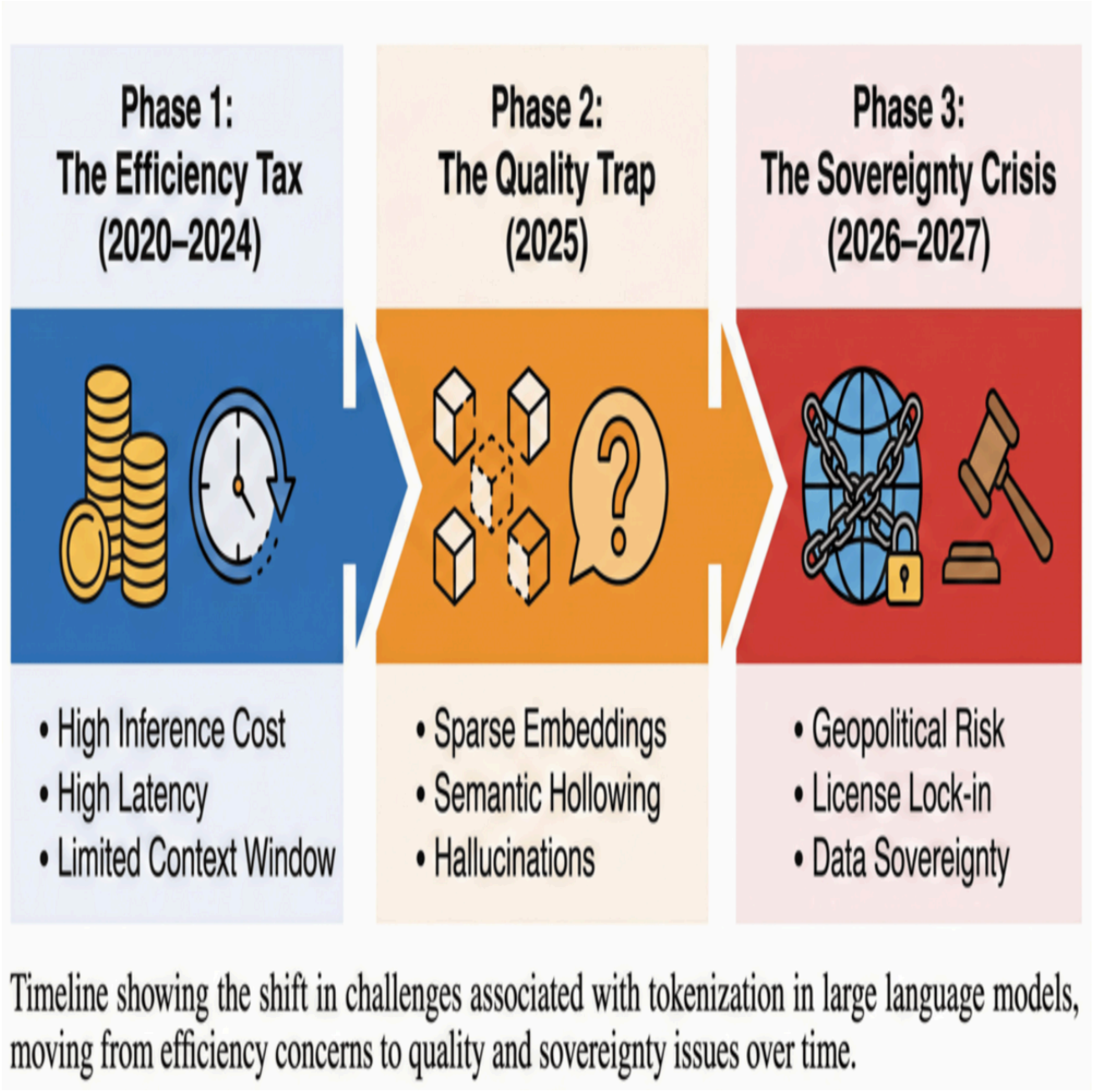

FIGURE 1: The Evolving Tokenization Bottleneck (2020-2027)

The paradigm shift

The landscape of multilingual NLP changed dramatically with next-generation models like Llama 3 [1] and Qwen 2.5 [6]. These models use much larger vocabularies (100k+ tokens). Evidence suggests that increasing vocabulary size compresses the fertility rate of ASEAN languages to near parity with English (about 1.2-1.6x), reducing the first-generation "Language Tax."

However, this review proposes that this fix introduces a subtler, second-order challenge: the "Sparse Embedding Trap." Larger vocabularies reduce sequence length but introduce thousands of rare tokens in languages like Thai or Malay. Without proportional increases in training data - a "Zipfian Data Shortage" - embedding vectors for these tokens remain

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5.
DOI https://doi.org/10.7759/s44389-026-00047-5

under-optimized. Drawing on the "under-trained tokens" identified by Land and Bartolo [7], we call these "Zombie Parameters": tokens in the embedding matrix with "hollow" representations due to insufficient gradient updates. This leads to "Semantic Hollowing": the model runs efficiently (low fertility) but has poorer representational quality. Furthermore, adapting these vocabularies worsens the Plasticity-Stability Dilemma, often causing Catastrophic Forgetting of the base model's reasoning.

Objectives and contributions

In reaction to these developments, this paper presents a systematic analysis of multilingual adaptation methodologies that extends beyond a descriptive survey. First, the analysis re-examines the common assumption that lower token fertility is inherently advantageous. It analyzes the trade-off between sequence compression efficiency and embedding representational quality.

Second, we introduce a structured taxonomy of adaptation strategies, grouping existing approaches into three categories: (i) vocabulary-level adaptation, (ii) tokenizer reconfiguration, and (iii) token-free or alternative sequence modeling architectures. Third, through analyzing representative multilingual systems - including SeaLLM, SEA-LION, and MambaByte - this review provides a technically grounded framework for evaluating multilingual adaptation strategies, with emphasis on representational quality, training robustness, and deployment efficiency.

Review

Related work

Research on multilingual large language models (LLMs) for Southeast Asian languages builds based on prior work in cross-lingual representation learning and parameter-efficient adaptation.

The initial phase of multilingual modeling (2018-2020) was defined by encoder-based architectures such as mBERT [8,9] and XLM-R [3]. These models presented zero-shot cross-lingual transfer by projecting more than 100 languages into a common embedding space [10]. However, their use was largely limited to discriminative tasks, such as classification and sequence labeling.

Subsequent developments (2020-2023) introduced large-scale autoregressive models, including BLOOM [11], mT5 [12], and XGLM [13], which extended multilingual modeling to generative tasks. While these models incorporated Southeast Asian languages such as Indonesian and Vietnamese into their pre-training corpora, empirical studies have observed performance trade-offs. These trade-offs tend to become more pronounced as the number of supported languages increases. In particular, languages with relatively limited representation in training data - such as Thai and Burmese - often receive disproportionately small corpus shares, which may reduce downstream performance and create tokenization issues.

To lower the computational cost of adapting large foundation models, Parameter-Efficient Fine-Tuning (PEFT) methods have been widely adopted. Early work, including MAD-X [14], introduced language-specific adapter modules inserted between Transformer layers [15,16] to enable modular transfer. This approach was further advanced by LoRA [17] and its quantized variant QLoRA [18], which approximate weight updates using low-rank matrix decomposition, substantially reducing the number of trainable parameters. In resource-limited settings, PEFT methods provide a practical mechanism for adapting global pre-trained models to domain- or language-specific data without incurring the full cost of model retraining.

In parallel, research has examined the impact of tokenization on multilingual performance. Rust et al. [19] systematically quantified the negative effects that standard subword tokenizers can have on low-resource languages. Subsequent work explored vocabulary development techniques and embedding initialization strategies, such as FOCUS [20], to improve alignment of newly introduced tokens with existing embedding spaces. Based on this literature, the present review analyzes the downstream optimization implications of vocabulary expansion, including potential embedding sparseness and parameter under-utilization effects that remain insufficiently characterized in prior studies.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5.
DOI https://doi.org/10.7759/s44389-026-00047-5

Methodology and scope

Literature Search and Selection Methodology

A systematic literature search was conducted to identify relevant studies examining adaptation strategies, tokenization efficiency, and representational robustness for LLMs within the context of Southeast Asian languages. To ensure methodological transparency, reproducibility, and rigorous evidence synthesis, the review protocol was strictly aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines, as well as the PRISMA-S extension for reporting literature searches. Furthermore, because automated tools were utilized to assist in the screening and extraction phases, the methodology adheres to the emerging PRISMA-trAIce (Transparent Reporting of Artificial Intelligence in Comprehensive Evidence Synthesis) checklist to document the human-AI interaction securely and transparently.

Information Sources and Search Strategy Formulation

A comprehensive, systematic search was executed across major academic and technical databases, specifically targeting the ACL Anthology, IEEE Xplore, the ACM Digital Library, and the arXiv preprint repository. The search window was strictly restricted to publications released between January 2020 and March 2026. This specific timeframe was selected to accurately capture the rapid architectural evolution from early-generation, English-centric Transformer models to contemporary high-vocabulary, multilingual, and token-free paradigms.

To ensure exhaustive retrieval of relevant literature while minimizing noise, complex Boolean search strings were formulated, iteratively refined, and adapted to the specific syntax of each respective database. The primary search string deployed across all platforms combined four distinct conceptual clusters: model architecture, tokenization mechanics, regional linguistic targets, and optimization challenges. The exact Boolean query utilized was as follows: (("Large Language Model" OR "LLM" OR "Foundation Model" OR "Transformer" OR "Generative AI") AND ("tokenization" OR "Byte-Pair Encoding" OR "BPE" OR "vocabulary expansion" OR "subword segmentation" OR "tokenizer") AND ("ASEAN languages" OR "Southeast Asian languages" OR "Thai" OR "Malay" OR "Bahasa Melayu" OR "Indonesian" OR "Bahasa Indonesia" OR "Vietnamese" OR "Tagalog" OR "Khmer" OR "Burmese" OR "Lao") AND ("low-resource" OR "efficiency-plasticity" OR "resource-constrained" OR "fertility rate" OR "embedding sparsity" OR "representational drift")).

This Boolean logic ensured that the search captured studies at the intersection of advanced language modeling and the specific linguistic properties of the ASEAN region.

Eligibility Criteria and Record Screening Process

Explicit inclusion and exclusion criteria were established prior to the execution of the screening phase to mitigate selection bias. To be included in the final qualitative synthesis, studies were required to present architectural modifications, training-level adaptations, or novel evaluation metrics specifically targeting at least one ASEAN language. Furthermore, included papers had to provide explicit technical specifications, such as tokenizer configurations, vocabulary size metrics, token fertility rates, or empirical benchmark results demonstrating downstream performance impacts.

Conversely, studies were systematically excluded if they fell into any of the following categories: (1) research focusing exclusively on application-level API usage without underlying model weight modification; (2) studies whose linguistic scope was restricted solely to high-resource languages such as English or Mandarin Chinese; and (3) descriptive surveys lacking novel theoretical frameworks or empirical data.

Following the automated removal of duplicate records across the disparate databases, an initial screening of titles and abstracts evaluated approximately 150 unique records. This initial phase was conducted in duplicate by independent reviewers to ensure high inter-rater reliability, culminating in a full-text assessment of the remaining manuscripts to finalize the inclusion pool of 40 key sources.

Quality Assessment and Risk of Bias in the NLP Domain

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5.
DOI https://doi.org/10.7759/s44389-026-00047-5

Given the extraordinarily rapid pace of development in the NLP domain, standard systematic review methodologies that rely solely on peer-reviewed literature introduce an unacceptable temporal lag. Top priority was assigned to peer-reviewed literature published in top-tier computational linguistics venues (e.g., ACL, EMNLP, NAACL). These venues enforce rigorous double-blind peer review processes, ensuring high standards for algorithmic transparency, dataset curation, and experimental reproducibility.

However, to reflect the rapid algorithmic evolution characterizing modern LLM research, preprints sourced from arXiv were conditionally included in the synthesis. Recognizing the inherent variability in non-peer-reviewed manuscripts, these preprints were subjected to a stringent quality assessment framework. Specifically, preprints were required to demonstrate: (1) exhaustive algorithmic transparency (e.g., detailed hyperparameter configurations and dataset composition) to ensure reproducibility; (2) empirical benchmarking on recognized regional evaluations (e.g., SEA-HELM) rather than localized zero-shot tests; (3) public artifact availability via unhindered access to code repositories or model weights; and (4) high scores on automated confidence assessment frameworks serving as a data-driven proxy for academic rigor.

Only preprints satisfying all four criteria were integrated into the final comparative analysis. This rigorous quality control mechanism ensures that the resulting conceptual framework is built upon a foundation of technically sound, temporally current, and empirically robust evidence. This study ultimately adopts a conceptual and analytical review methodology; therefore, the frameworks proposed herein function as diagnostic constructs derived from this highly vetted literature base. By comparing representative global baselines such as CulturaX [21] against regional case studies including SEA-LION [22] (AI Singapore) and MaLLAM [23] (Mesolitica), this framework provides a structured basis for analyzing representational robustness in resource-constrained environments.

Theoretical framework: Fertility and embedding sparsity risk (ESR)

Conceptualizing Efficiency: The Fertility Rate

To reason about computing overhead, we introduce the Fertility Rate $(\mathcal{F})$ as a conceptual proxy for tokenization efficiency. For a given language and tokenizer, fertility is the expected ratio of generated subword tokens to semantic word units across a representative corpus. In this study, fertility is treated not as an optimization target but as an analytical indicator of morphological fragmentation and segmentation stress. Particular attention is given to "agglutinative compression effects" in Austronesian languages (e.g., Malay, Indonesian) and "scriptio continua segmentation challenges" in Kra-Dai languages (e.g., Thai, Lao). These linguistic characteristics provide structural pressure points for subword-based tokenization.

Formalizing Adaptation Risk: ESR as a Theoretical Concept

To quantify the degradation of representational quality in expanded vocabularies, we introduce ESR. This analytical construct describes the tension between vocabulary expansion and data density.

Mathematical Formalization and Parameter Discussion

Vocabulary expansion creates a hidden risk: while it allows a model to process text faster, it can severely degrade the model's actual understanding of those new words. To measure this specific degradation, this paper introduces ESR. Imagine a language model's embedding matrix as a vast, high-dimensional dictionary where every single word or subword has a precise numerical coordinate. When developers artificially expand the vocabulary to include thousands of new Southeast Asian linguistic units, they are essentially adding thousands of blank pages to this internal dictionary. If the training dataset lacks sufficient examples of these new words, the model cannot calculate meaningful mathematical updates-known as gradients-for them. As observed in recent studies on under-trained tokens by Land and Bartolo [7], these neglected tokens fail to move past their random starting positions. We classify these stagnant, hollow tokens as "Zombie Parameters." The ESR framework systematically calculates the exact percentage of the new vocabulary that remains trapped in this inactive state. Formally, ESR is defined as:

$$ESR(V_{new}, D) = \frac{1}{|V_{new}|} \sum_{t \in V_{new}} \mathbf{1}\left(\sum_{i=1}^{N} \|\nabla_{e_t} L(x_i)\|_2 < \tau\right)$$

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5.
DOI https://doi.org/10.7759/s44389-026-00047-5

Within this formulation, $|V_{new}|$ serves as the normalizing factor for the overall risk score, ensuring comparability across models with varying vocabulary scales. The adaptation dataset *D* consists of *N* individual data samples x_i, over which the learning dynamics are evaluated. The indicator function $\mathbf{1}(\cdot)$ yields a value of 1 if a token's total gradient updates across the dataset remain below a specified convergence threshold τ, signaling a "zombie" status. Specifically, the term $\nabla_{e_t}\mathcal{L}(x_i)$ represents the first-order gradient of the loss function $\mathcal{L}$ with respect to the embedding vector e_t, serving as a mathematical proxy for the magnitude of updates the token receives during fine-tuning. Importantly, the threshold τ is conceptualized as a hyperparameter defined relative to the mean gradient magnitude of high-frequency core tokens. This ensures that token viability is measured against the model's active semantic baseline rather than an arbitrary absolute value.

In practical deployment scenarios where direct computation of gradient statistics over the full training trajectory may be computationally prohibitive, we adopt the empirical diagnostics proposed by Land and Bartolo [7] as a heuristic proxy. For models with untied embeddings, ESR can be approximated by measuring the L_2 norm of embedding vectors, where unusually low norms indicate tokens that have received minimal effective updates. For architectures utilizing tied embeddings, cosine distance from the mean output embedding may serve as a secondary indicator for identifying under-trained tokens. This comprehensive formalization effectively shifts the analytical focus from superficial tokenization efficiency, such as fertility reduction, to representational density and parameter utilization.

The Zipfian Data Shortage

High ESR is driven by a fundamental mathematical mismatch: system vocabularies expand linearly, but human language follows a highly skewed distribution known as Zipf's Law. In linguistically rich languages like Thai or Malay, thousands of unique morphological variations exist, but they appear very rarely in everyday text, occupying the extreme "long tail" of language usage. When developers aggressively add these rare tokens to the vocabulary without massively increasing the volume of the training data, the model almost never encounters them during the learning phase. This data starvation creates what we term the "Bypass Phenomenon." Because the newly added tokens remain mathematically hollow and undefined, the model's internal attention mechanism treats them as dead zones. Instead of utilizing a newly minted, highly efficient token for a complex Malay prefix, the model bypasses it entirely. It reverts to its old habits, awkwardly routing its processing power to cobble together the word using familiar, fragmented English subwords.

Based on the metrics above, we posit a fundamental inverse relationship between Token Fertility $(\mathcal{F})$ and Embedding Sparsity *S*. Expanding the vocabulary generally reduces fertility $(\mathcal{F} \propto 1/|V|)$, which lowers inference costs. However, it linearly increases the sparsity of the embedding table relative to a fixed dataset $(S \propto |V|)$. Therefore, the structural trade-off can be conceptualized as:

$$S \propto \frac{1}{F}$$

This relationship implies that the "Language Tax" (High $\mathcal{F}$) and "Semantic Hollowing" (High S) are opposite ends of the same spectrum.

Background and definitions

In order to discuss the strategies of adaptation, one must first define the technical backgrounds of inefficiencies related to tokenization and then analyze the restrictions of vocabulary expansion as an exact cure.

Subword Tokenization and Distributional Imbalance

Subword tokenization algorithms have become prevalent in modern LLMs with the main example being BPE [24], which is used to compromise between character-level flexibility and word-level efficiency. In the course of training, BPE sequentially combines two most frequently co-occurring bytes in a corpus to create a vocabulary of a specific size (e.g., 32,000 tokens in Llama 2). The process of applying learned merge rules during inference is used to split text into subword units.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5.
DOI https://doi.org/10.7759/s44389-026-00047-5

The quality of BPE depends on the statistical characteristics of the pre-training corpus. The tendency to optimize frequent morphemes and affixes in high-resource languages like English is more likely when training data is highly concentrated in them. Most of the high-level merge rules are not frequently invoked in languages that have a relatively small corpus size. The consequence is that the tokenization can go back to smaller sub-word components or even to byte-level symbols resulting in larger sequences and less efficient tokens.

Byte Fallback and the UTF-8 Encoding Trap

When using non-Latin scripts (such as Thai and Burmese) in Southeast Asian languages, more concerns with the UTF-8 encoding arise. Most ASCII English letters take up one-byte each in UTF-8, but letters of other scripts like Thai or Khmer need three bytes.

Under limited vocabulary coverage, and where a character cannot be modeled as a learned subword unit, the tokenizer could split it by its constituent bytes instead of associating it with a single token. Under extreme conditions, number of tokens may reach the number that corresponds to several times of number of characters, greatly growing length of sequence. In addition to computational cost, such segmentation at byte level can disintegrate character-level representations, leaving the model unable to easily acquire coherent orthographic or morphological rules.

The Naive Metric: Fertility Rate

Although the Fertility Rate ($\mathcal{F}$) explained previously can be used as a strong measure of computational performance and inference time, this review claims that it can be used as a reductive and possibly a misleading substitute of representational quality. In the past, the excessive fertility rates of models such as Llama 2 (4.29-8.6 in Thai vs. ~1.2 in English) were rightly recognized as the main cause of the so-called "Language Tax" [25-27].

Nevertheless, in the framework of contemporary high-vocabulary systems (e.g. Llama 3, Qwen 2.5), reducing fertility is not sufficient to achieve efficient adaptation. Our hypothesis is that optimizing a fertility rate of $(\mathcal{F}) = \mathbf{1.0}$ can mask structural deterioration. The extreme vocabulary compression might artificially reduce the token count to match the English baseline but without the proportional data density, this efficiency tends to cover up the Semantic Hollowing of the embedding space. Consequently, the use of fertility alone leads to an illusion of equality: a model can handle Thai as fast as English, but not have the rich semantic structure needed to perform more nuanced reasoning [28,29]. To address this constraint, the ESR should be introduced to reflect the trade-off which fertility indicators overlook.

The "Sparse Embedding Trap": An Operational Failure, Not a Theoretical One

While vocabulary expansion is a primary remedy for high fertility, its implementation often triggers what we term the “Sparse Embedding Trap.” This phenomenon occurs when newly introduced tokens appear so infrequently in the adaptation corpus that their embedding vectors fail to receive adequate gradient updates, leading to representational stagnation.

Importantly, this risk is highly sensitive to initialization procedures. To mitigate the initial "cold-start" effect, recent techniques such as FOCUS [20] suggest initializing new token embeddings via linear combinations of their constituent subword embeddings:

$$\boldsymbol{E_{new}} = \sum \boldsymbol{w_i} \boldsymbol{E_{subword,i}}$$

In this formulation, $\boldsymbol{E_{new}}$ represents the initialized vector for the new token, while $\boldsymbol{E_{subword,i}}$ denotes the embedding of its constituent subword *i*, weighted by $\boldsymbol{w_i}$. Such mathematically informed initialization promotes smoother integration into the existing embedding space. However, as formalized by the ESR framework in the earlier section, even advanced initialization cannot entirely compensate for a lack of data density. In resource-constrained environments, adding more vocabulary may successfully lower token fertility but still result in a “hollowed” embedding space if the new parameters remain in a near-initial state.

Morphological Imperatives vs. Sparsity

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5.
DOI https://doi.org/10.7759/s44389-026-00047-5

This trade-off is also made more difficult by the varied morphology of the ASEAN languages. In Scriptio Continua languages such as Thai and Burmese, there is no space between words and the tokenizer has to implicitly learn the word boundaries; standard BPE cannot handle this and tends to split the words in the middle of characters. Likewise, Agglutinative Languages such as Malay form words through intricate affixation (i.e., memperkenalkannya). Normal tokenizers tend to break these down into nonsensical pieces (i.e., mem-per-ke-lakan-nya) and the model needs to direct its attention towards reconstructing the morphology instead of working on semantics. This structural and semantic contrast is visually summarized in Figure *2*.

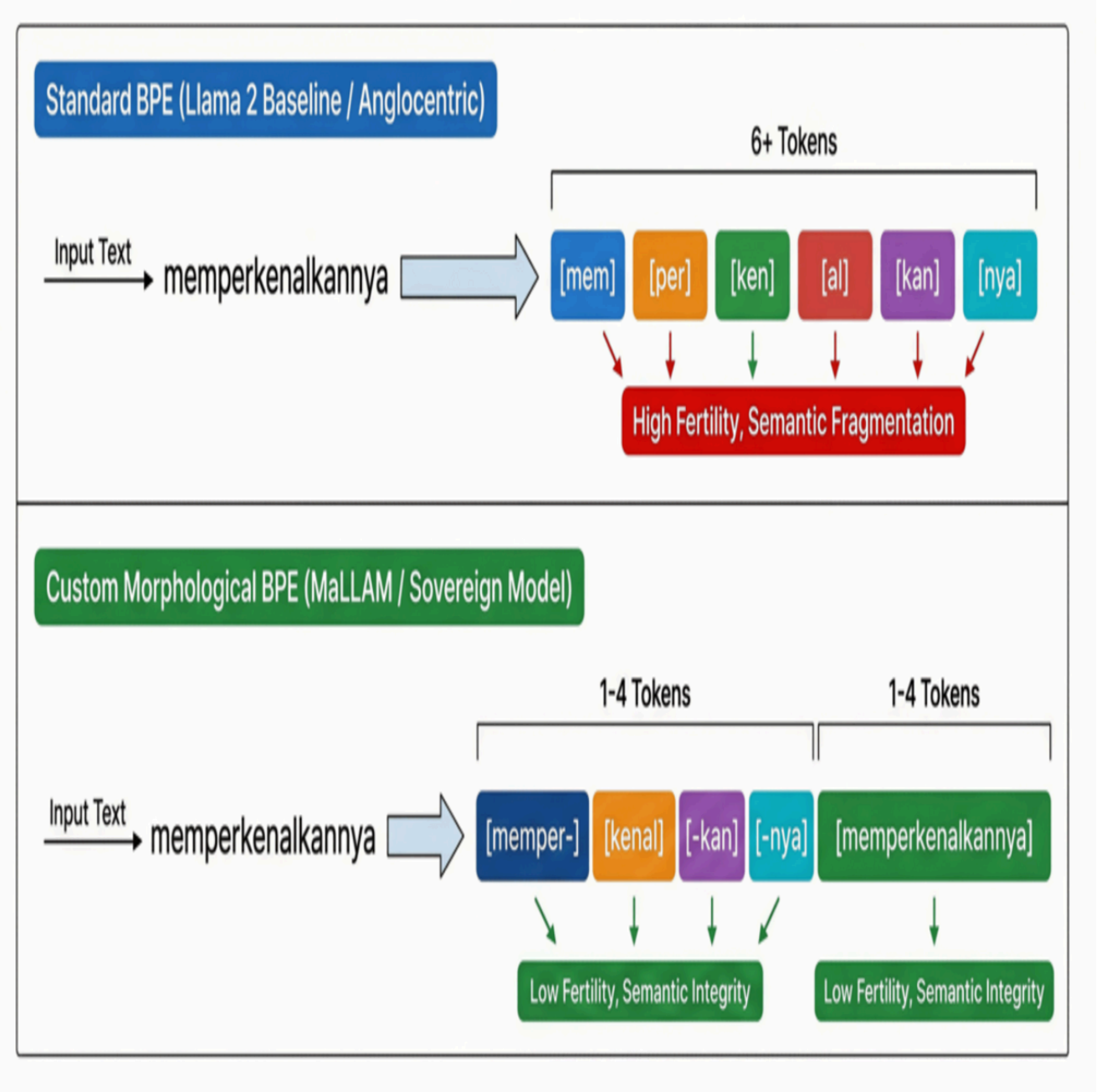

FIGURE 2: Standard BPE vs. Morphological Segmentation in Bahasa Melayu

BPE, Byte-Pair Encoding; MaLLaM, Malaysia Large Language Model

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

The expansion of the vocabulary to include these morphological units (e.g., meN- or ber-) is linguistically sound, but it also adds to the risk of the Sparsity Trap unless supplemented by large volumes of high-quality training data, a criterion that is not frequently satisfied in low-resource environments.

Main body taxonomy: adaption strategies

The research community has come up with several strategies that are used to cope with the tokenization bottleneck over the last few years. The present paper builds an overall three-level classification system depending on the level of intervention in the model architecture:

1. Vocabulary Adaptation and Expansion: "Patching" the tokenizer on existing models.

2. Region-Specific Tokenizer Construction: Building tokenizers and models optimized for regional languages from scratch.

3. Token-Free Architectures: Completely discarding tokenizers to handle bytes directly.

Strategy 1: Vocabulary Adaptation and Expansion (The "Patch" Approach)

This is the most cost-effective and popular approach at this moment, especially when working with a team that does not have the computing resources to train large models from scratch. Its fundamental principle is to preserve the reasoning ability of already strong models (such as Llama 2, Mistral) but to change just their input interface (tokenizer) and embedding layer.

Methodological process: The process of vocabulary growth usually consists of three major stages. The first is called Target Language Vocabulary Training, in which researches will take tools like SentencePiece [24] to train a small BPE model on a pure target language corpus (e.g., Thai Wikipedia or Vietnamese news). The second is Vocabulary Merging and Pruning, which is the process of combining the new high-frequency tokens with the old vocabulary of the foundation model. In order to avoid having vocabularies that are too large and lead to parameter bloat, frequent thresholding or the elimination of very low-frequency tokens out of the original vocabulary is often used. To name an example, the SeaLLM [25] project introduced about 10,000 ASEAN specific tokens to the Llama 2 base. Lastly, Adjusting the Embedding Layer is done; since the vocabulary is expanding, the dimensions of the input embedding matrix and the output layer (LM Head) need to be expanded accordingly.

MaLLAM (Mesolitica) [23] is the first attempt to use this particular pipeline. MaLLAM, developed by the research team Mesolitica, solves the problem of vocabulary mismatch in the models of Llama and Mistral, but in the context of the Bahasa Melayu language. MaLLAM also remedies the so-called byte fallback problem whereby a normal agglutinative word such as memperkenalkannya is divided into an excessive number of pieces by increasing the tokenizer to support 15,000-20,000 high-frequency Malay words and affixes. The fact that vocabulary adaptation can be implemented at the national level at low cost has been demonstrated.

Key technique: Embedding initialization: The process of new tokenization directly influences the pace and steadiness with which the model is able to adapt to a new language. In early random initialization, the new vectors are initialized with random noise, but it is found that this approach works poorly on Vietnamese-specific syllables and disyllabic words because there is no initial semantic alignment. Enlarged tokenizer not only increased the inference rate but more so the quality of the model understanding of Vietnamese semantics whereby full lexical unit enables the attention mechanism to correctly detect word relationships instead of being enmeshed into segments of characters.

The counter-approach: Strategic base model selection: The Shift towards Strategic Model Choice. As opposed to the patching strategy, the development of OpenThaiGPT demonstrates the effectiveness of Strategic Base Model Selection. Earlier versions were based on LoRA adapters and built on Llama-based architecture but in version 1.5 a strategic change was made by starting with Qwen 2.5-72B. Using Qwen (a very massive multilingual vocabulary, ~152 k tokens), and which has been known to naturally provide better Thai coverage, the project avoided having to expand its vocabulary significantly. This case study indicates that a "linguistically dense" base model tends to be more efficient than patching a "linguistically sparse" model.

Critical risk: Catastrophic forgetting: Although vocabulary growth is also an effective way to minimize the cost of inference, it poses a significant implicit threat called Catastrophic Forgetting. During Continued Pre-training (CPT) of the model to match new embeddings, the required gradient changes frequently interfere with the previous weights of the model, resulting in a unique trade-off between Fluency and Reasoning. Empirical evidence shows that although adapted models become fluent in target languages like Thai or Malay, they may perform English mathematical reasoning (e.g., GSM8K) and coding tasks significantly worse [28,29]; that is, the model does not remember profound logic but superficially adapts to the language. As a result, more complicated remedial measures are needed instead of a straightforward fine-tuning. As an example, further versions of SeaLLM v3 [25] and MaLLAM [23] were compelled to introduce Replay Buffers, i.e., adding a well-structured English instruction data, math questions, and code back into the training stream (usually at an approximate rate of 10-20 percent) in order to maintain general intelligence. The presented evidence implies that vocabulary modification cannot be regarded as a simplistic one-time solution but as a sensitive equilibrium that needs advanced data engineering to prevent the creation of a fluent but dumb model.

Strategy 2: Independent Pre-Training Pipelines

The current approach is the most resource-consuming adaptation route: training a foundation model using a pipeline of regional pre-training, such as designing custom tokenizers and constructing large-scale corpora. The fundamental idea is that downstream adaptation can be inadequate, in case representation gaps are induced by pre-training distributions. Thus, changes should be implemented on the level of data-sources, not on the basis of gradual fine-tuning.

Region-specific pre-training data: In contrast to vocabulary enhancement, full-stack pre-training has to re-engineer the pre-training corpus. This strategy does not depend on general-purpose web crawls (e.g., C4) but, rather, focuses on curated multilingual corpora with proportional representation of Southeast Asian languages [21]. The procedure usually consists of large-scale language identification, elimination of machine-translated material of low-quality, and incorporation of domain-specific literature as news archives, educational literature, and open-access institutional documentation.

The aim is to add more tokens to morphologically productive languages in the pre-training phase, hence minimizing the dependence of downstream on aggressive vocabulary growth. Nevertheless, this scale of corpus construction needs significant information engineering system and long-term support.

Case studies: SEA-LION (AI Singapore) [22]: The SEA-LION was created by AI Singapore with a custom SEABPE tokenizer that was trained on about 20 million lines of multilingual regional text. The model has a vocabulary size of 256,000 and when applied to Malay and Indonesian text it will reduce the number of tokens by roughly 30-43 percent compared with Llama 2. This decrease also makes the memory more efficient in processing long-form documents and more effective use of context during inference.

The Typhoon (SCB 10X) [27]: SCB 10X has created Typhoon, which is a Thai-specific model based on Mistral design with further optimization by means of large-scale CPT. It has been reported that the efficiency of tokenization in Thai increases by around 2.6 times compared to the base setup, and it enhances the throughput of applications written in Thai under constant memory limits.

The "pre-training paradox": Building a regional language model entirely from scratch-a strategy known as full-stack pre-training-grants developers absolute control over the vocabulary and tokenization process. However, this ambitious approach has two major drawbacks.

(1) The Data Wall: A modern foundation model containing billions of parameters functions much like a massive, high-performance engine; it requires trillions of high-quality data tokens acting as fuel to run effectively. Unfortunately, Southeast Asian languages currently suffer from a severe fuel shortage in the digital space. The available supply of digitized, clean, and diverse text in languages like Lao, Khmer, or Burmese is simply too small to satiate a multi-billion-parameter architecture. When researchers are forced to train these massive models on this limited pool of data, the

system breaks down in one of two ways. The model either memorizes the training text verbatim without truly understanding the language (over-specialization), or it fails to grasp the language's complex grammatical rules altogether (underfitting).

Moreover, the size of the needed compute indicates the extent of the data wall faced by policymakers and system designers. E.g., pre-training a 7B parameter base model with 1 trillion tokens usually takes 1-3M GPU hours on NVIDIA A100-class hardware. Despite these discounts, it means that the straightforward cost of computation would be more than USD $1 million without counting preprocessing and validating data. To give a sense of scale, moving to 70B-parameter models or adding additional training steps in several languages could drive demand upwards of 10 million GPUs. Such costs are unattainable to many research teams based in Southeast Asia, and highlight the conflict between data aspirations and policy or financial constraints. Policy needs to thus take into account that not only corpus acquisition is necessary, but also realistic financing of both computing infrastructure as well as operational expenses.

(2) The Moving Target Problem: The training of large models requires huge computation facilities. The speed of development in architectural designs is so high that when an architectural model is equal to a prior baseline, there can be newer, more sophisticated designs. This period between times reduces the advantages of mass independent pre-training.

From both economical and technological perspectives they indicate diminishing returns of individual pre-training with limited computational resource availability. As a result, numerous teams are implementing hybrid solutions in which effective global architecture (including Qwen 2.5) and area-specialist models (as well as fine-tuning) are mixed together. This approach is built upon the old-fashioned abilities to reason but can be customized into linguistic and subject-specific abilities based on data-driven methods instead of recreating tokenizers.

Strategy 3: Token-Free Architectures (Byte-Level Models)

The third strategy is a drastic change to the normal practice: removing all tokenizers and learning language directly through raw bytes. Although in theory, this solves the problem of the original sin of Anglocentric tokenization, today it has serious engineering headwinds that restrict its use in real time ASEAN settings. These models technically take raw UTF-8 byte streams as inputs (e.g., [0xe0, 0xb8, 0x81...]) and not Token IDs. The method entirely removes the Out-Of-Vocabulary issue; when modeling scriptio continua languages such as Thai, the model implicitly learns to make words out of bytes, which are more resistant to typos and informal communication on social media. Nevertheless, this is achieved with significant costs: explosion of sequence lengths. As a single Thai character can be represented by 3 bytes, byte-based sequences tend to be 10-15 times longer than the equivalent token-based sequences, thus leading to enormous computational loads.

Case studies: MambaByte and BLT: The most prominent efforts to address this trade-off are MambaByte [30] and Byte Latent Transformer (BLT) [31]. MambaByte uses the Selective State Space Model (SSM) architecture which has a linear complexity, thus preventing the quadratic bottleneck of the Transformer. It is as perplexible as subword models in controlled experiments and extremely noise-resistant. At the same time, BLT [31] by Meta AI is more of a hybrid, it consumes bytes but reorganizes them into patches dynamically according to information entropy. High entropy (which means unpredictability in transition) suggests that the word boundary exists, and the model produces less patch size. Such an entropy-based mechanism allows BLT to learn Thai word segmentation in an implicit way without using a tokenizer.

Although the computational economics of SSMs, such as MambaByte, in handling very long byte-level sequences are compelling, recent theoretical considerations indicate that there are trade-offs in its architecture. SSMs perform worse than the global receptive field of conventional Transformers attention mechanism in particular associative recall and exact-copying tasks. This means that in the case of SEA languages, even though MambaByte solves the problem of morphological tokenization penalty, it can also be prone to poor factual retrieval when reading dense long context legal or scholarly materials. As a result, hybrid architectures with a combination of byte-level dynamic entropy patching (similar to BLT) and localized attention mechanisms might be considered the final optimal frontier of token-free LLMs.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

The "latency barrier": In the past, byte-level models used to have a linear latency penalty that was a function of the sequence length (*N*), and frequently produced text at a rate that is 10-15 times lower than the token-based systems because of the longer byte sequences. Nevertheless, this limitation has been significantly alleviated by some recent architectural advances (2025-2026).

The BLT (Pagnoni et al. [31]) is first to propose Dynamic Entropy Patching that dynamically clusters together bytes of varying size into latent patches determined by the density of information. The allocation of computational resources occurs in favor of high-entropy areas (e.g., morphological boundaries) with predictable areas being reduced to large patches. The process is internalized into the model architecture and this decreases the effective sequence length (and thus the latency disparity with the token-based system).

To be more precise, Hybrid Speculative Decoding [32] uncouples proposal creation and checking. A small draft model suggests possible continuation alternatives, which are evaluated by a big verifier model simultaneously. At high rates of acceptability, effective latency is close to that of token-based inference even if the underlying representation is in bytes.

The cumulative effect of all this is that these mechanisms make the conventional premise that byte-based representation is inherently associated with a prohibitive inference delay less credible. The trade off in performance thus changes between representational granularity and systemic optimality due to these architectural and decoding innovations.

Comparative analysis

This section presents a comparative synthesis of the aforementioned adaptation frameworks across three dimensions: tokenization efficiency (fertility trends), economic effects, and downstream representational performance. Particular analytical attention is given to the relative performance gains in Bahasa Melayu and other structurally diverse ASEAN languages.

Fertility Rate Comparison: The "Agglutinative Penalty"

An analysis of tokenization behavior indicates that there is a stark difference in the fertility rates of English-centered baselines and region-specific models. The table indicates that standard models such as Llama 2 place a large agglutinative penalty on Malay and Indonesian. Because the Austronesian languages have such a rich morphology (such as prefixes, e.g. memper-, suffixes, e.g. -kan), the standard 32k vocabularies cannot cover all the morphemes, which leads to an over fragmentation. The qualitative analysis of these tokenization efficiency trends and their impact across diverse ASEAN scripts is summarized in Table *1*.

1. Initial Inefficiency: Baseline models utilizing standard Anglocentric BPE demonstrate high token fertility rates in Malay, indicating that a single semantic word is frequently divided into multiple subword tokens. This breaking apart is even more severe in scriptio continua languages like Thai, where the lack of explicit word boundaries leads to extreme sequence inflation.

2. Successful Adaptation: Models applying the Vocabulary Expansion strategy (e.g., SeaLLM [25]) successfully mitigate this partition, substantially decreasing the relative fertility rate compared to baseline levels by incorporating high-frequency regional subwords. From an ESR perspective, however, this sequence improvement comes at the cost of reduced update frequency for long-tail tokens, potentially leading to uneven parameter utilization.

3. Regionally Optimized Models: Models engineered from the ground up with a regional focus, such as MaLLaM [23] by Mesolitica and SEA-LION [22] by AI Singapore, demonstrate the highest segmentation efficiency. Through the use of custom tokenizers with significantly larger vocabularies (e.g., 256k tokens), these models preserve intact agglutinative affixes, driving tokenization efficiency closer to English parity and substantially reducing the efficiency gap. Under the ESR framework, this aggressive scaling reflects highly uneven parameter update dynamics if regional data density is insufficient.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

Crucially, while these extracted metrics demonstrate significant empirical reductions in token fertility, these architectural shifts must be interpreted through the ESR framework to fully understand the hidden trade-offs regarding parameter utilization and semantic hollowing.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

Model & Strategy	Vocabulary Size	Reported ASEAN Fertility/Compression Metric	Primary Targeted ASEAN Languages	Key Architectural Trade-Off Reported
Llama 2 (Baseline Anglocentric)	32,000	4.29–8.6 in Thai vs. ~1.2 in English	English-centric	Extreme sequence inflation due to byte fallback inflicts a severe agglutinative penalty and structural inefficiency.
Llama 3 (Baseline Anglocentric)	128,256	Parity with English (about 1.2–1.6×)	Multilingual	Expanded vocabularies reduce sequence length but introduce thousands of under-optimized tokens lacking gradient updates.
Mistral 7B (Baseline Anglocentric)	32,000	Not explicitly reported	English-centric	High morphological fracturing forces the attention mechanism to reconstruct basic morphology instead of processing semantics.
SeaLLM (Vocab Expansion)	~42,000 to 250,000	Compression ratio for Thai text improved from 4.29 to 1.57	Thai, Malay, Indonesian, Vietnamese, Khmer, Lao, Tagalog, Burmese	Modifying embeddings requires replay buffers to prevent catastrophic forgetting of base mathematical and logical reasoning.
Typhoon (Vocab Expansion)	~37,000	2.62× efficiency increase for Thai	Thai	Reliance on specific base models exposes developers to the moving target problem as global foundation architectures evolve.
SEA-LION (Sovereign Regional)	256,000	Fertility reduced by 30–43% for Malay and Indonesian	Thai, Malay, Indonesian, Vietnamese, Filipino, Burmese, Khmer, Lao	Full-stack pre-training hits a severe data wall, risking over-specialization or underfitting due to limited regional corpus size.
MaLLaM (Sovereign Regional)	32,000	Up to 43% reduction in token size	Malay	The binary implementation of source-shielded updates can trigger a bilingual reasoning disconnect without careful layer calibration.
MambaByte (Token-Free)	256	Not explicitly reported	Language-Agnostic	Byte-level sequences are significantly longer, requiring speculative decoding via subword drafting to achieve a 2.6× inference speedup.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

BLT (Token-Free)	256	Average patch is 6–8 bytes	Language-Agnostic	Generating a variable number of bytes before reaching word boundaries creates fundamental batching synchronization complexities.

TABLE 1: Empirical Tokenization Efficiency and Reported Architectural Trade-Offs

BLT, Byte Latent Transformer

As demonstrated in the empirical data above, a consistent structural trade-off emerges: models that achieve optimal sequence compression (low fertility) via massive regional vocabularies inherently expose themselves to higher ESR if not supported by proportional training data. For example, models with vocabularies above 200k tokens show reduced fertility but introduce an increased risk of under-trained embeddings, as explicitly reflected in their reported trade-offs.

Economic Analysis

The structural differences in tokenization efficiency translate directly into operational expenditures (OpEx) for deployment. Using standard commercial API pricing models (which bill per token), the cost impact on regional organizations can be prohibitive when utilizing unadapted baselines.

When processing Malay legal documents, baseline models generate a high volume of billable tokens due to morphological fracturing. Conversely, implementation of a regionally optimized model (such as MaLLaM or SeaLLM) reduces token volume by an estimated 30 to 43 percent on Malay language tasks, based on the vocabulary compression ratios reported in their respective technical literature. The conceptual savings are projected to be even higher for entities processing Thai text, potentially exceeding a 60 percent reduction in token generation. This efficiency benefit is not simply a technical optimization but a practical constraint for deployment for ASEAN start-ups operating under limited compute budgets.

Performance and Cultural Alignment

Crucially, efficiency gains do not compromise downstream performance; rather, existing studies suggest they enhance semantic grasp. These performance trends can be interpreted as downstream effects of tokenization efficiency and ESR-related parameter utilization.

1. Semantic Integrity: Models utilizing adapted tokenizers (e.g., SeaLLM [25], SEA-LION [22]) are significantly more successful than baseline models in the SEA-HELM [33] benchmark on tasks that require reading comprehension. On Malay, the MaLLAM [23] model has shown potential in better managing local linguistic properties.

2. The maintenance of General Intelligence: It is critically important that benchmarks reveal that hybrid approaches (Vocabulary Expansion) can be implemented without catastrophic forgetting. A well-known example of this is SeaLLM-13B [25], which performs well in both English math and coding problems (GSM8K) as well as reaching the state-of-the-art in local language-related tasks.

3. The Long-Tail Challenge and Code-Switching: Although existing modifications work with main ASEAN languages, poor performance on long-tail metrics strongly suggests that scaling laws alone are insufficient in low-data languages. To counteract this failure, it is necessary to go further than simply expanding parameters, as described in the new desiderata of Filipino NLP. Future sovereign frameworks should be deliberately designed to understand complex sociolinguistic processes such as code-switching (e.g., Taglish in the Philippines, and Bahasa Rojak in Malaysia) which are currently marginalized in Western-dominated pre-training corpora.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

Infrastructure, licensing, and algorithmic constraints in multilingual LLM deployment

On the contrary, even though other aspects of the previous sections had examined the ways that the algorithms adapt, there are also severe structural limitations that influence the deployment decisions of the multilingual LLMs. An exclusive focus on the model can fail to take into account external and internal variables that impact scale, expenses and maintenance directly. This part discusses the engineering constraints in the form of licensing limits, hardware density, and underlying algorithm bottleneck as being linked together. The aim is to evaluate whether such conditions impact on the likelihood of deploying large scale multilingual models in resource diverse settings.

Licensing Thresholds and Scalability Constraints

The tactic of constructing regional LLM services based on open-weight foundation models (i.e. LLaMA-family) is not based solely on technical feasibility but also on limits of license compliance. According to the Meta Llama Community License overview, the commercial rights can be limited in case the combined Monthly Active Users (MAU) of the licensee and its subsidiaries are more than 700 million.

Smaller markets can have such thresholds that are not practically important. Nevertheless, in the case of big digital ecosystems where users are aggregated across borders, MAU cap is a scalability constraint. In the event an organization runs various popular platforms owned by the same entity, it is the combined user numbers as opposed to the per application use that may be used to establish whether compliance has been attained. In those circumstances, quick growth in adoption might require the renegotiation of licensing agreements or switching to different models.

This means that in terms of engineering planning, license-based deployments should include growth predictions and compliance assessments as part of the architectural choices of a given system. The problem is not ideological but operational, the scaling limitations of licensing contracts can play a role in determining long term models particularly those aimed at regional or multinational user groups.

Hardware Stack Concentration and Algorithmic Optimization Strategies

Multi-lingual LLM integration is strongly related to the existing accelerator ecosystem. The available state of the art training and inference pipelines are mostly specialized on GPU-based systems and the corresponding stack of software (ex, GPU-friendly toolchains based on CUDA). This focus produces a reliance on particular hardware systems and compiler ecosystems.

Language Processing Units or domain-specific inference processors are alternative accelerator designs that are still part of a very concentrated semiconductor manufacturing and supply chain system. As a result, it does not necessarily follow that diversification in the processor-architecture implies the decentralization of the infrastructure. Manufacturing scale, compiler maturity, memory bandwidth, and ecosystem support limit hardware heterogeneity.

In such situations, the optimization of the performance at the algorithmic layer is becoming an increasingly essential design parameter. Free-token or byte-based structures have been frequently attacked on grounds of greater sequence length, which in turn causes a linear scaling of latency ($O(L)$). Nevertheless, hybrid inference techniques, including speculative decoding [32], show that wall-clock latency may be somewhat uncoupled with sequence length by means of parallel proposal-verification systems.

Within the draft-verify framework, an ultra-lightweight proposal model is used to produce candidate continuations and a more complex verifier model is used to evaluate these simultaneously. With high proposal acceptance rate, useful latency is comparable to that of traditional token-based inference even if the underlying representation is in bytes.

Thus, there is hardware stack concentration as a viable systems constraint yet the flexibility of deployment can be increased by innovating on the inference layer. The central challenge of design is to find the equilibrium between the efficiency of accelerators, representation granularity and speculative performance of decoders under realistic compute limits.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

In the future, the operational implementation of the concept of SEA sovereign models will tend to move more towards the complex Agentic Workflows instead of simple single-turn chatbots. The phenomenon of language tax appears sharply in cases of continuous self-reflection, API calling, and reasoning in multiple steps as excessive inference requests are made at high frequencies. Combining rollback-aware speculative decoding with dynamics of Batch Query Processing is an engineering solution of choice. Compute-poor areas of the world in the Global South can implement advanced, non-token autonomous agents with similar operational latency and cost as their English-first counterparts by optimizing the algorithmic parallelism of draft model specifically to short and iterative agentic tasks.

The "Bilingual Disconnect" of Source-Shielded Updates

Beyond hardware and inference-time constraints, adapting models to Southeast Asian languages introduces severe limitations in training dynamics. Earlier, when discussing Vocabulary Expansion (Strategy 1), we noted that data-driven techniques like Replay Buffers are often used to mitigate Catastrophic Forgetting. However, in compute- and data-constrained scenarios, a prominent architecture-driven alternative is the use of Source-Shielded Updates (SSU), which freezes critical English reasoning parameters during local language fine-tuning. While successful in preserving top-line benchmark scores, we propose that the binary implementation of SSU leads to what we term a Bilingual Reasoning Disconnect. Physical forgetting (the process of overwriting weights) and functional forgetting (the distribution shift) can be distinguished by research.

The SSU [16,34] does not allow physical overwriting, but creates a strong separation between languages. Our hypothesis is that this causes the model to behave as a simple Translation Wrapper: converting a Vietnamese mathematics question into an internally English representation to solve it, and then back again with the result. It stops the model to form real reasoning circuits in the target language. It is a failure to the ASEAN states, which have plans to apply AI to native-language science education: the model simulates fluency without profound vernacular intelligence.

Mitigation Strategy: The Calibrated SSU Protocol

In order to eliminate this inconsistency and make the process of deep reasoning transfer more efficient, we propose the use of the SSU protocol as described by Yamaguchi et al. [34]. Rather than completely freezing the layers, we freeze the MLP layers in a column-wise manner. It is important to note that importance scores must be obtained based on reasoning-rich corpora (e.g., GSM8K, MBPP [28,29]) instead of large corpora containing text. With this approach, the model will focus more on protecting its logical and abstract reasoning circuits when undergoing adaptation and still have the required amount of plasticity to be adapted to ASEAN scripts.

Theoretical Analysis of Gradient Dynamics and Embedding Sparsity

Given the hardware concentration and inference latency constraints discussed above, optimizing parameter utilization becomes a primary design requirement. Achieving superficial tokenization efficiency alone is insufficient; it is necessary to evaluate whether newly introduced vocabulary is effectively integrated into the model's semantic space.

While proponents of vocabulary expansion (Strategy 1) correctly observe that fertility decreases as vocabulary size increases [24,27], the assumption that lower fertility directly implies improved semantic quality remains theoretically unverified. To examine this premise, we apply the ESR framework to analyze gradient dynamics during fine-tuning.

In resource-constrained environments, vocabulary expansion often introduces a severe structural vulnerability linked to what researchers call "glitch tokens" or "magikarp tokens". When thousands of new, rare tokens are added to a model's vocabulary, they rarely appear in the relatively small local datasets used for fine-tuning. Consequently, these tokens receive almost no mathematical updates-or gradients-during the training process. Unlike the model's core English vocabulary, which was rigorously optimized on trillions of words [35], these new regional tokens remain frozen in their random, uninitialized states. Because these frozen tokens lack distinct semantic meaning, the model's internal attention mechanism cannot mathematically recognize them. Just as a sweeping searchlight glides over a matte black object without reflecting any light, the attention routing calculations sweep past these under-trained tokens without registering a high score. Forced to ignore the newly added vocabulary, the model wastes its computational power on meaningless "attention sinks" [36] or inefficiently breaks the regional word down into smaller, older subwords.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

This dynamic suggests that aggressive vocabulary expansion, while reducing token fertility, may simultaneously increase effective sparsity in the embedding space. The outcome is a model that is computationally efficient yet representationally under-utilized-a condition we refer to as semantic hollowing [23,29].

In this context, ESR serves as a theoretical diagnostic tool for evaluating parameter utilization. Rather than focusing solely on tokenization efficiency, it enables analysis of whether expanded parameters actively participate in the model's learning dynamics. Although empirical validation-such as tracking gradient magnitudes or embedding norm drift across languages-remains an important direction for future work, ESR provides a necessary conceptual foundation for assessing representational robustness in multilingual adaptation.

Strategic roadmap: from "one-size-fits-all" to asymmetric adaptation

The uniformity of fine-tuning strategies implies that they are based on the assumption that the computational resources, availability of talent, and data maturity are equal in various deployment settings. In reality, there is significant heterogeneity between high-performance computing capacity, NLP research infrastructure, and digitized corpora. One adaptation pathway cannot thus be expected to be the best in all contexts.

The tiered adaptation framework proposed in this section matches the degree of technical intervention to quantifiable measures of resource capacity. The framework does not suggest a standard application of model modification, but it rather identifies three types of adaptation: architecture-level adaptation, application-level integration and data-centric augmentation depending on infrastructure preparedness.

The Resource Divide and Three-Tier Framework

Depending on the observable factors including the amount of available computing resources, the development of the research ecosystem, scale of corpus digitization, deployment contexts can be divided into three operational tiers. Every tier has its own technical focus instead of being a normative hierarchy. As mapped out in the ASEAN Tiered Adaptation Strategy (Figure *3*), each tier demands a distinct technical approach aligned with its primary resource lever.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

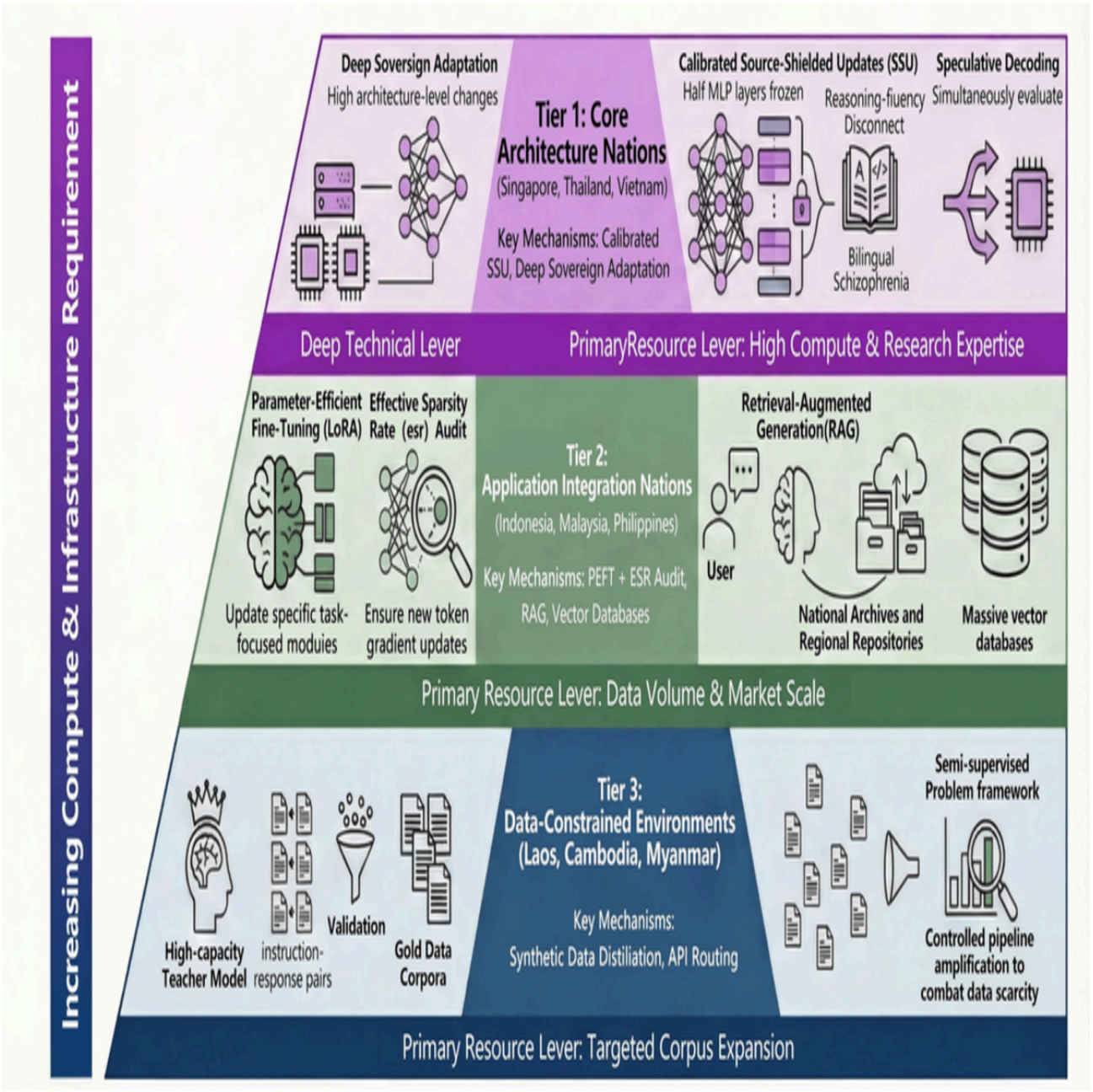

FIGURE 3: The ASEAN Tiered Adaptation Strategy Map on AI: Pathways of Asymmetry Between Core Architecture, Application Integration, and Environments With Constraints on Data

MLP, Multi-Layer Perceptron; PEFT, Parameter-Efficient Fine-Tuning

Tier 1: Core Architecture Nations (Singapore, Thailand, Vietnam)

1. Description: Environments characterized by access to high-performance computing clusters, mature NLP research ecosystems, and relatively well-developed national or institutional corpora.

2. Suggested Approach: Deep Sovereign Adaptation. While these nations possess the capacity to perform full-stack regional pre-training (Strategy 2), the rapid evolution of global foundation models often renders from-scratch training strategically inefficient. Instead, they are well-positioned to act as the "Engine Room" by adapting state-of-the-art global

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

models through Calibrated SSU. By applying a column-wise freezing ratio to MLP layers-calibrated using reasoning-dense corpora (e.g., GSM8K) rather than general text-this approach preserves the model's abstract reasoning capabilities while maintaining sufficient plasticity for ASEAN language adaptation. This mitigates the risk of Bilingual Reasoning Disconnect.

3. Key Imperative: Infrastructure and Inference Optimization. Given the concentration of accelerator ecosystems, infrastructure should support backend heterogeneity across compiler stacks and runtime environments. In parallel, inference-layer optimizations (e.g., speculative decoding) enable efficient execution of token-free or byte-level models without requiring specialized hardware, shifting part of the computational burden toward algorithmic scheduling.

Tier 2: Application Integration Nations (Indonesia, Malaysia, Philippines)

1. Description: Environments characterized by moderate compute capacity, large user bases, and heterogeneous, partially digitized data ecosystems.

2. Suggested Approach: Retrieval-Augmented Generation-Augmented PEFT. Combining PEFT methods such as LoRA with Retrieval-Augmented Generation provides favorable cost-performance trade-offs. Prior to deployment, models should be audited using the ESR framework to ensure that newly introduced vocabulary and adapter parameters receive sufficient gradient updates. In practice, this can be approximated by analyzing gradient magnitude distributions or embedding norm variance across newly introduced tokens. Fine-tuning should prioritize layers exhibiting significant gradient dynamics, avoiding unnecessary updates across the full parameter space.

3. Key Imperative: Vector Database Infrastructure. The primary engineering priority is the construction of large-scale vector databases derived from legal archives, public records, and domain-specific corpora. This decouples model parameters from retrieval systems, enabling iterative knowledge updates without retraining the base model [37].

Tier 3: Data-Constrained Environments (Laos, Cambodia, Myanmar)

1. Description: Environments characterized by limited compute infrastructure, small digitized corpora, and persistent challenges in encoding standardization and script normalization.

2. Suggested Approach: Synthetic Data Distillation. Full-scale model training is often economically infeasible. Instead, API-based inference combined with targeted dataset expansion provides a viable pathway. High-capacity teacher models can be used to generate instruction-response pairs, which are then filtered and refined using a small human-validated subset ("gold data"). This approach enables scalable dataset construction while significantly reducing annotation costs, though it requires careful filtering to mitigate bias propagation and distributional artifacts inherited from teacher models.

3. Key Imperative: Semi-Supervised Corpus Amplification. The central engineering challenge shifts from large-scale pre-training to controlled corpus expansion and validation. Data scarcity is reframed as a semi-supervised learning problem, emphasizing iterative refinement pipelines over monolithic training processes.

A comprehensive summary of this three-tier framework, detailing the technical emphasis and core mechanisms for each deployment context, is provided in Table *2*.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

Tier	Deployment Context	Technical Emphasis	Core Mechanisms	Primary Resource Lever
Tier 1	High compute, mature research	Architecture-level adaptation	Calibrated SSU, Speculative Decoding, Backend heterogeneity	Compute & Research Expertise
Tier 2	Moderate compute, large user base	Application integration	PEFT + ESR Audit, RAG, Vector Databases	Data Volume & Market Scale
Tier 3	Compute-constrained	Data-centric augmentation	Synthetic Data Distillation, API routing	Targeted Corpus Expansion

TABLE 2: The ASEAN Tiered Adaptation Matrix

API, Application Programming Interface; ESR, Embedding Sparsity Risk; PEFT, Parameter-Efficient Fine-Tuning; RAG, Retrieval-Augmented Generation; SSU, Source-Shielded Updates

Federated Data Collaboration Framework: The "Common Data Space"

Fine-tuning is not enough to address cold-start limitations due to lack of sufficient high-quality parallel corpora in low-resource environments. The solution to this limitation should be a well-organized system of sharing information instead of more fundamental restructuring of architecture.

Our proposal is that a shared multilingual data exchange protocol should be established where the participating institutions can add curated parallel or domain-specific corpora with standardised metadata formats. The presented framework does not hinder cross-lingual transfer learning and at the same time meets local data governance requirements.

The size of a corpus is not the only aspect to consider when it comes to the granularity of the representation. Most pretrained tokenizers tend to implicitly prefer the high-resource segmentation norms of languages. In order to reduce the effect of this bias, tokenizer design could consider linguist-defined boundaries, and assessment criteria that would explicitly evaluate the ability to withstand code-mixing and dialectal variation.

In place of imposing uniform standardization, benchmark design must be able to award precise management of mixed-script data and informal registers. This is a change in assessment because of normative correction to representational fidelity.

Future directions

While the strategic roadmap addresses infrastructure and policy layers, the development of equitable AI systems will likely depend on continued advances at the algorithmic level. Beyond standard tokenization adaptation, three emerging technical frontiers can be identified that have the potential to change how models process ASEAN languages in a significant manner. The frontiers discussed in this section-including visual-linguistic grounding and dynamic tokenization-represent prospective technological trajectories. While these paradigms offer high-potential solutions to the tokenization bottleneck, their evaluation within the specific context of ASEAN languages remains largely conjectural. These discussions serve as well-informed projections based on emerging SOTA trends rather than absolute empirical results.

Visual Text Understanding (The "Pixel" Frontier)

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

With the complex glyphs and scriptio continua features of languages like Thai, Burmese or Jawi, a new way of handling these texts has been developed which does not consider text encoding but rather considers text as images. Visual encoders such as ViT can be used by researchers to bypass Unicode encoding problems (like Zawgyi vs. Unicode) and prevent fragmentation due to traditional tokenizers, because they directly extract features based on pixel data. Crucially, interpretations through pixel-based encoders effectively bypass the ESR and the UTF-8 encoding trap identified in earlier sections. By prioritizing spatial density over token sequences, models like Qwen2.5-VL [38] aim to preserve the morphological integrity of scriptio continua remains intact. However, it should be noted that the transition from textual tokens to visual patches is an evolving research frontier, and its comparative performance across all Tier 3 scripts remains to be empirically verified. It is especially relevant when it comes to digitizing old manuscripts and doing strong optical character recognition where text noise makes subword tokenization useless. The Typhoon team has also started exploring this method in multimodal models as a solution to text-processing constraints that are viewed visually.

The issue of morphological tokenization does not only affect texts, the problem also goes beyond text, into the multimodal field, particularly as far as Vision-Language Models are concerned. The models, including Qwen2.5-VL [38], face special visual tokenization problems in the process of handling the Southeast Asian typographic structure which is associated with densely written scriptio continua (e.g., Thai, Khmer) and non-Latin optical character recognition properties. This has prompted decentralized organizations such as SEACrowd [39] to go further than textual information to create a multimodal repository of information. This shift in thinking requires that the challenge of eliminating the chances of embedding sparsity is to ensure linguistic agreement and an overall visual-linguistic situationalization to reflect the full range of visual culture, local signage, and multimodal social media products in the area.

Dynamic and Adaptive Tokenization

Trained current vocabularies cannot be modified, which is computationally ineffective in multilingual contexts where languages are frequently changed. The future mechanism will be Dynamic Tokenization, where the tokenization grain of the model can change as per the domain. As an example, when reading court documents, a model may dynamically replace a specialized vocabulary matrix of Legal Thai to compress the text maximally and then revert to general vocabulary during a regular conversation. Inference efficiency and domain specialization would also be greatly enhanced by this "Plug-and Play" mechanism without having to incur the high cost of retraining the entire foundation model.

Synthetic Data Distillation and Self-Correction

A primary barrier for sovereign models in data-constrained settings (Tier 3), such as Lao and Khmer, is the limited availability of high-quality, human-annotated data for supervised fine-tuning. To address this constraint, recent work has increasingly explored synthetic data distillation as a scalable alternative.

This approach converts computational resources into structured training data through a multi-stage pipeline. A typical workflow includes:

(1) defining prompt templates and task specifications tailored to the target language and cultural context;

(2) generating instruction-response pairs using high-capacity teacher models (e.g., GPT-4o, Qwen-Max);

(3) applying automated filtering via reward models to assess fluency, relevance, and cultural alignment;

(4) deduplicating and balancing the dataset to ensure coverage of diverse linguistic patterns;

(5) incorporating targeted human-in-the-loop review to calibrate filtering thresholds; and

(6) consolidating the resulting data into standardized formats suitable for downstream training.

Such pipelines enable datasets to scale rapidly without a proportional increase in human annotation cost. For example, the WangchanX Instruct dataset [40] employs a synthetic pipeline to generate over 120,000 native Thai instruction pairs. Similarly, the VFin-Gen framework demonstrates that synthetic generation can be particularly effective in domain-specific settings, such as finance, where expert annotation is expensive.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

However, synthetic data generation also introduces challenges, including potential biases, factual inconsistencies, and over-reliance on teacher model distributions. As a result, careful filtering, validation, and periodic human auditing remain essential to ensure data quality and cultural fidelity.

Limitations and the Need for Empirical Validation: A Blueprint for Future Research

As a conceptual review, the frameworks proposed in this study - particularly ESR - should be interpreted as diagnostic theoretical constructs rather than empirically validated metrics. To operationalize ESR as a measurable architectural indicator, systematic empirical validation is required.

A rigorous validation protocol should isolate vocabulary size as the primary independent variable while controlling for model architecture, training steps, and data distribution. Specifically, base models employing standard tokenizers (e.g., 32k vocabulary) can be compared against models with expanded regional vocabularies (e.g., 256k) under a fixed-scale corpus and identical training regimes.

During adaptation, internal training dynamics can be analyzed through gradient-based telemetry. Two key signals are proposed: (1) gradient magnitude, $\boldsymbol{M_t = \|\nabla_{e_t}\mathcal{L}\|_2}$, capturing the update intensity of token embeddings; and (2) representational drift, $\boldsymbol{D_t = \|E_t^{(c)} - E_t^{(init)}\|_2}$, measuring deviation from initialization. Together, these metrics provide a practical mechanism for identifying under-trained "zombie parameters" and estimating convergence thresholds $\boldsymbol{(\tau)}$.

Beyond internal diagnostics, future work should establish empirical links between ESR and downstream task performance. In particular, correlating sparsity indicators with accuracy degradation on culturally grounded benchmarks (e.g., SEA-HELM) would enable validation of the "Semantic Hollowing" hypothesis. Such analysis can transform ESR from a descriptive concept into a predictive tool for guiding adaptation strategies.

At the same time, claims regarding the superiority of token-free architectures and visual-linguistic grounding (the "Pixel" frontier) remain largely speculative in the context of Tier 2 and Tier 3 ASEAN languages. While these approaches may alleviate tokenization bottlenecks, their impact on parameter utilization and effective sparsity remains unclear. From an ESR perspective, future studies should evaluate whether increased representational granularity meaningfully reduces sparsity or merely redistributes it across longer sequences.

Finally, comparative floating point operations per second (FLOPS)-to-performance analyses are required to determine the boundary at which gains in representational fidelity justify the additional computational overhead of byte-level modeling. Such evaluations are particularly critical for deployment in resource-constrained environments, where efficiency-performance trade-offs directly affect feasibility.

Conclusions

The tokenization bottleneck has traditionally increased inference costs and diminished efficiency for Southeast Asian languages. Nonetheless, this review indicates that these restrictions stem significantly from particular modeling choices rather than the languages themselves. Our primary contributions are as follows: (1) the introduction of the ESR to quantify the trade-offs associated with vocabulary expansion, and (2) the proposal of a tiered adaptation framework that aligns with regional resource constraints.

This analysis suggests that further advancement relies more on data quality and adaptation methodologies than on merely increasing model scale. The use of high-quality instructional data, representative benchmarks, and reproducible evaluation frameworks is essential for continuous improvement. Looking ahead, the development of multilingual LLMs tailored for Southeast Asian languages requires attention to morphological diversity, script variation, and culturally specific semantics. Future progress will depend on improvements in tokenizer design, embedding initialization, synthetic data generation, and evaluation methods-rather than focusing only on increasing model or vocabulary size.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Cheong Yik Sheng

Acquisition, analysis, or interpretation of data: Cheong Yik Sheng, Cheong Yee Xin, Tan Guan Jie

Drafting of the manuscript: Cheong Yik Sheng, Tan Guan Jie

Critical review of the manuscript for important intellectual content: Cheong Yik Sheng, Cheong Yee Xin

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. Grattafiori A, Dubey A, Jauhri A, et al.: The Llama 3 herd of models [PREPRINT]. arXiv. 2024, 10.48550/arXiv.2407.21783
2. Jiang AQ, Sablayrolles A, Mensch A, et al.: Mistral 7B [PREPRINT]. arXiv. 2023, 10.48550/arXiv.2310.06825
3. Conneau A, Khandelwal K, Goyal N, et al.: Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 8440-8451. 10.18653/v1/2020.acl-main.747
4. Touvron H, Martin L, Stone KR, et al.: Llama 2: open foundation and fine-tuned chat models [PREPRINT]. arXiv. 2023, 10.48550/arXiv.2307.09288
5. Sennrich R, Haddow B, Birch A: Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1715-1725. 10.18653/v1/P16-1162
6. Yang A, Yang B, Hui B, et al.: Qwen2 technical report [PREPRINT]. arXiv. 2024, 10.48550/arXiv.2407.10671
7. Land S, Bartolo M: Fishing for Magikarp: automatically detecting under-trained tokens in large language models. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 11631-11646. 10.18653/v1/2024.emnlp-main.649
8. Devlin J, Chang MW, Lee K, Toutanova K: BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019, 4171-4186. 10.18653/v1/N19-1423
9. Pires T, Schlinger E, Garrette D: How multilingual is multilingual BERT?. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019, 4996-5001. 10.18653/v1/P19-1493
10. Bapna A, Firat O: Simple, scalable adaptation for neural machine translation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019, 1538-1548. 10.18653/v1/D19-1165
11. Scao TL, Fan A, Akiki C, et al.: BLOOM: a 176B-parameter open-access multilingual language model [PREPRINT]. arXiv. 2022, 10.48550/arXiv.2211.05100

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

12. Xue LT, Constant N, Roberts A, et al.: mT5: a massively multilingual pre-trained text-to-text transformer. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021, 483-498. 10.18653/v1/2021.naacl-main.41

13. Lin XV, Mihaylov T, Artetxe M, et al.: Few-shot learning with multilingual generative language models. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 9019-9052. 10.18653/v1/2022.emnlp-main.616

14. Pfeiffer J, Vulić I, Gurevych I, Ruder S: MAD-X: an adapter-based framework for multi-task cross-lingual transfer. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2022, 7654-7673. 10.18653/v1/2020.emnlp-main.617

15. Liu Y, Gu J, Goyal N, et al.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics. 2020, 8:726-742. 10.1162/tacl_a_00343

16. Artetxe M, Ruder S, Yogatama D: On the cross-lingual transferability of monolingual representations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 4623-4637. 10.18653/v1/2020.acl-main.421

17. Hu EJ, Shen Y, Wallis P, et al.: LoRA: low-rank adaptation of large language models. Proceedings of the International Conference on Learning Representations. 2022, 1-13.

18. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L: QLoRA: efficient finetuning of quantized LLMs. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 10088-10115.

19. Rust P, Pfeiffer J, Vulić I, Ruder S, Gurevych I: How good is your tokenizer? On the monolingual performance of multilingual language models. 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). 2021, 3118-3135. 10.18653/v1/2021.acl-long.243

20. Dobler K, De Melo G: FOCUS: effective embedding initialization for monolingual specialization of multilingual models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 13440-13454. 10.18653/v1/2023.emnlp-main.829

21. Nguyen T, Nguyen CV, Lai VD, et al.: CulturaX: a cleaned, enormous, and multilingual dataset for large language models in 167 languages. Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024, 4226-4237.

22. Ng R, Nguyen TN, Huang Y, et al.: SEA-LION: Southeast Asian languages in one network. Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2025, 512-526. 10.18653/v1/2025.ijcnlp-long.30

23. Zolkepli H, Razak A, Adha K, Nazhan A: MaLLaM: Malaysia large language model [PREPRINT]. arXiv. 2024, 10.48550/arXiv.2401.14680

24. Kudo T, Richardson J: SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018, 66-71. 10.18653/v1/D18-2012

25. Nguyen XP, Zhang W, Li X, et al.: SeaLLMs - large language models for Southeast Asia. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). 2024, 294-304. 10.18653/v1/2024.acl-demos.28

26. Yuenyong S, Viriyayudhakorn K, Piyatumrong A, Jaroenkantasima J: OpenThaiGPT 1.5: a Thai-centric open source large language model. Proceedings of the International Conference on Applications of Natural Language to Information Systems. Ichise R (ed): Springer, Cham; 2025. 15837:210-219. 10.1007/978-3-031-97144-0_19

27. Pipatanakul K, Jirabovonvisut P, Manakul P, et al.: Typhoon: Thai large language models [PREPRINT]. arXiv. 2023, 10.48550/arXiv.2312.13951

28. Cobbe K, Kosaraju V, Bavarian M, et al.: Training verifiers to solve math word problems [PREPRINT]. arXiv. 2021, 10.48550/arXiv.2110.14168

29. Austin J, Odena A, Nye M, et al.: Program synthesis with large language models [PREPRINT]. arXiv. 2021, 10.48550/arXiv.2108.07732

30. Wang JX, Gangavarapu T, Yan JN, Rush AM: MambaByte: token-free selective state space model. Proceedings of the Conference on Language Modeling. 2024, 1-30.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5

31. Pagnoni A, Pasunuru R, Rodriguez P, et al.: Byte latent transformer: patches scale better than tokens. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025, 9238-9258. 10.18653/v1/2025.acl-long.453

32. Shen YH, Shen JY, Kong Q, Liu TY, Lu Y, Wang C: Speculative decoding via hybrid drafting and rollback-aware branch parallelism. arXiv. 2025, 10.48550/arXiv.2506.01979

33. Susanto Y, Hulagadri AV, Montalan JR, et al.: SEA-HELM: Southeast Asian holistic evaluation of language models. Findings of the Association for Computational Linguistics: ACL 2025. 2025, 12308-12336. 10.18653/v1/2025.findings-acl.636

34. Yamaguchi A, Morishita T, Villavicencio A, Aletras N: Mitigating catastrophic forgetting in target language adaptation of LLMs via source-shielded updates [PREPRINT]. arXiv. 2025, 10.48550/arXiv.2512.04844

35. Wang X, Wei J, Schuurmans D, et al.: Self-consistency improves chain of thought reasoning in language models [PREPRINT]. arXiv. 2023, 10.48550/arXiv.2203.11171

36. Xiao GX, Tian YD, Chen BD, Han S, Lewis M: Efficient streaming language models with attention sinks [PREPRINT]. arXiv. 2024, 10.48550/arXiv.2309.17453

37. Liu NF, Lin K, Hewitt J, et al.: Lost in the middle: how language models use long contexts. Transactions of the Association for Computational Linguistics. 2024, 12:157-173. 10.1162/tacl_a_00638

38. Bai S, Chen K, Liu X, et al.: Qwen2.5-VL technical report [PREPRINT]. arXiv. 2025, 10.48550/arXiv.2502.13923

39. Lovenia H, Mahendra R, Akbar SM, et al.: SEACrowd: a multilingual multimodal data hub and benchmark suite for Southeast Asian languages. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 5155-5203. 10.18653/v1/2024.emnlp-main.296

40. Pengpun P, Udomcharoenchaikit C, Buaphet W, Limkonchotiwat P: Seed-free synthetic data generation framework for instruction-tuning LLMs: a case study in Thai. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). 2024, 438-457.

How to cite this article:
Yik Sheng C, Yee Xin C, Guan Jie T (March 30, 2026) Beyond the Tokenization Bottleneck: A Conceptual Framework for Efficiency-Plasticity Trade-Offs in ASEAN Large Language Model Adaptation. Cureus J Comput Sci 3 : es44389-026-00047-5. DOI https://doi.org/10.7759/s44389-026-00047-5