E-ISSN: 3025-5481
DINASTI INFORMATION AND TECHNOLOGY
DINASTI RESEARCH
https://dinastires.org/DIT dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.38035/dit.v4i1

Evolution and Adaptation of Large Language Models for Bahasa Indonesia

Allan Desi Alexander[1], Siti Setiawati[2]
[1]Universitas Bhayangkara Jakarta Raya, Jakarta, Indonesia, allan@ubharajaya.ac.id
[2]Universitas Bhayangkara Jakarta Raya, Jakarta, Indonesia, siti.setiawati@dsn.ubharajaya.ac.id

Corresponding Author: allan@ubharajaya.ac.id[1]

Abstract: This study evaluates the systematic evolution and computational adaptation of pre-trained language models and Large Language Models (LLMs) for Bahasa Indonesia and its low-resource regional dialects. Initially centered on bidirectional encoder-based representations like IndoBERT, the regional natural language processing (NLP) field has transitioned toward generative sequence-to-sequence structures and massive decoder-only architectures. This paper investigates the engineering methodologies of cross-lingual vocabulary adaptation, parameter initialization heuristics, and language-adaptive pre-training strategies designed to address text overfragmentation, representational misalignment, and tokenization cost inefficiencies. Through extensive structural benchmarks, this analysis compares discriminative and generative performances across tasks including sentiment classification, extractive question answering, text style normalization, domain-specific retrieval-augmented pipelines, and entity linking. While localized generative models such as Komodo, Sailor, and the SEA-LION suite improve contextual reasoning, colloquial style transfers, and regional dialect preservation, they remain susceptible to architectural anomalies like template leakage and entity hallucination. This study provides foundational benchmarks and methodological frameworks for adapting massive language models to morphologically rich, culturally diverse, and low-resource linguistic environments.

Keyword: Language Adaptation, Bahasa Indonesia, Large Language Models, Vocabulary Expansion, Continual Pre-Training, Cultural Alignment

INTRODUCTION

Bahasa Indonesia is spoken by almost 200 million people, ranking as the tenth most spoken language globally and the fourth most used language across the internet (Hasmawati & Ade Romadhony, 2023). Despite this massive user base, its digital presence in natural language processing (NLP) research was historically hindered by a scarcity of annotated datasets, resource standardization, and specialized pre-training corpora (Chakrabarty et al., 2020). Most foundational models have been primarily trained on high-resource English datasets, resulting in severe representation disparities (McGiff & Nikolov, 2025). This computational imbalance produces an Anglocentric bias, rendering models highly sensitive to cultural mismatches and morpho-syntactic fragmentation (Ding et al., 2024).

Indonesia is home to 726 recorded regional languages, accounting for roughly 10% of the world's languages. While Bahasa Indonesia serves as the unifying official language, approximately 70.9% of the population is multilingual, often code-switching between standard Indonesian and various regional dialects such as Javanese, Sundanese, Minangkabau, and Balinese . Despite this linguistic richness, predictions suggest that in 100 years, up to 90% of these regional languages could face extinction or severe endangerment (Ridwan, 2018). Machine translation and localized NLP architectures are crucial tools for linguistic preservation and cross-cultural communication. However, regional dialects are low-resource by definition, lacking the massive parallel corpora required for traditional neural machine translation (NMT) systems (Jin, 2024).

The evolution of Indonesian language models occurred in distinct phases. The pioneer phase introduced standardized benchmarks and bidirectional encoders to capture syntax and semantics. The IndoNLU and IndoLEM frameworks provided clean evaluation datasets, enabling the training of IndoBERT variants (Kartika et al., 2023). Although these encoder-based models excelled in sequence labeling and classification, they could not perform complex conditional text generation. This limitation led to the second phase: the introduction of the IndoNLG benchmark and generative pre-trained models, specifically the encoder-decoder IndoBART and the decoder-only IndoGPT, which were pre-trained on the multi-billion-token Indo4B-Plus corpus (Cahyawijaya et al., 2021).

The current phase leverages massive open-weight multilingual LLMs (e.g., LLaMA, Qwen, and Gemma) and adapts them via continual pre-training (CPT) and cross-lingual vocabulary adaptation (CVA). Models such as Komodo-7B (Owen et al., 2024), Sailor, and SEA-LION (Fujii et al., 2024)exemplify this paradigm, integrating regional dialects and sociolinguistic context engines to bridge the representation gap for millions of multilingual speakers across the Indonesian archipelago .

METHOD

The adaptation of foundational architectures to Bahasa Indonesia utilizes core computational methodologies: vocabulary expansion, target embedding initialization, and continual pre-training objectives.

Vocabulary Expansion and Tokenization Optimization

Most standard LLMs rely on English-centric tokenizers, leading to severe text over-fragmentation in non-English languages (Petrov et al., 2023). Over-fragmentation degrades downstream task execution, slows inference, and increases API consumption costs. To quantify this over-fragmentation mathematically, we can express the tokenization fertility ratio F_R of a tokenizer T on a text document d written in a target language relative to English as:

$$F_R(T, d) = \frac{|T(d)|}{|T(d_{en})|}$$

where d_{en} represents the systematically equivalent English translation of the document d. In practice, standard English-centric tokenizers often yield an F_R value exceeding 3.0 for Bahasa Indonesia and up to 6.8 for regional Southeast Asian scripts (Lee et al., 2026), meaning that Indonesian text requires up to three times more tokens to represent the same semantic content as English. To counteract this, cross-lingual vocabulary adaptation (CVA) is performed (Jap & Arumsari, 2017), An auxiliary tokenizer T_{aux} is trained on a targeted Indonesian monolingual corpus D using subword algorithms such as Byte Pair Encoding (BPE) or Unigram (Subhan Mahendrasyah & Hariguna, 2024), The top k most frequent tokens from V_{aux} that are absent in the source tokenizer's vocabulary V_s are extracted:

$$V_{new} \subset V_{aux} \setminus (V_s \cap V_{aux})$$

This subset is merged with the source vocabulary to create an expanded target vocabulary V_{tc}:

$$V_t = V_s \cup V_{new}$$

For example, Komodo-7B expanded the standard LLaMA-2 vocabulary by adding approximately 2,000 Indonesian-specific tokens and 1,000 regional tokens, resulting in a curated target vocabulary of 35,008 tokens. This targeted expansion drastically reduces the fertility ratio while maintaining computational efficiency (Owen et al., 2024).

Furthermore, BPE Dropout is applied during the pre-training tokenization pipeline (Chizhov et al., 2024). BPE Dropout introduces stochasticity into the subword segmentation process by randomly omitting merge operations with a probability $p \epsilon [0,1]$ (Provilkov et al., 2020) . This stochastic segmentation exposes the model to diverse subword boundaries during pre-training, making the model more robust to trailing spaces, spelling errors, and the morpho-syntactic variations common in colloquial Indonesian web text (Di Marco & Fraser, 2024).

Parameter Initialization Heuristics

Expanding the vocabulary from V_s to V_t requires adding $|V_{new}|$ new rows to the input embedding matrix $W_e \in \mathbb{R}^{|V| \times H}$ and the output language modeling head $W_{lm} \in \mathbb{R}^{|V| \times H}$, where H is the hidden dimensionality of the model. Initializing these new parameters randomly can cause severe representation instability and high Kullback-Leibler (KL) divergence between the token-level distributions of the pre-adapted and adapted models (Linder, 2026):

$$D_{KL}(P \parallel Q) = \sum_{x \in V_t} P(x) \log \frac{P(x)}{Q(x)}$$

To stabilize early gradients during downstream tuning, researchers implement a heuristic-based embedding initialization. Instead of random vectors, the weights of the new tokens are initialized by mapping them to the centroid of the existing source embeddings (Yamaguchi et al., 2026):

$$W_e^{(new)} = \frac{1}{|V_s|} \sum_{i \in V_s} W_e^{(i)} + \epsilon$$

Where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is a small perturbation vector added to prevent identical weight trajectories. This heuristic aligns the new tokens with the centroid of the existing latent space, stabilizing early gradients during downstream tuning and reducing training steps (Fridkin & Bendersky, 2026).

Continual Pre-Training Objectives and Adaptation Tactics

Following vocabulary expansion, models undergo Language-Adaptive Pre-Training (LAPT) . The primary training objective is Causal Language Modeling (CLM) (Jo et al., 2025) . Let $X=\{x_1,x_2,\ldots,x_n\}$ be a sequence of tokens. The CLM objective minimizes the negative log-likelihood:

$$\mathcal{L}_{CLM}(\theta) = -\sum_{i=1}^{n} \log P(x_i \mid x_{<i}; \theta)$$

In data-scarce settings, models are increasingly evaluated using a Multi-Token Prediction (MTP) objective, which prompts the model to predict M future tokens at each step, significantly improving sample efficiency (Xu & Che, 2026). The MTP objective is formulated as:

$$\mathcal{L}_{MTP}(\theta) = -\sum_{i=1}^{n}\sum_{m=1}^{M}\lambda_m \log P(x_{i+m-1} \mid x_{<i};\theta)$$

where $\boldsymbol{\lambda_m}$ is a weighting coefficient for each future token position. Low-Rank Adaptation (LoRA) is utilized across all linear layers to adjust parameters efficiently while minimizing catastrophic forgetting of the source model's general capabilities (Mu et al., 2025). The weight update is parameterized as:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r}BA$$

where $W_0 \in \mathbb{R}^{d\times k}$ represents the frozen pre-trained weights, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are the low-rank adapter matrices, $\boldsymbol{r<<mind(d,k)}$ is the rank, and $\boldsymbol{\alpha}$ is a scaling factor.Alternatively, a two-stage tuning process is used[20]. In the first stage, only the expanded embedding matrices $\boldsymbol{W_e}$ and $\boldsymbol{W_{im}}$ are tuned while freezing the transformer layers, allowing the model to adapt to the new vocabulary without disturbing general representations. In the second stage, LoRA weights are updated across all linear layers to capture language-specific syntax and semantics (Zhang et al., 2025).

Table 1: Architectural Configurations and Adaptation Strategies of Indonesian Language Models

Architecture / Model	Original Base Model	Vocabulary Adaptation Method	Target Vocabulary Size	Core Initialization Strategy	Primary Pre-training Focus
IndoBERT	BERT-Base (uncased)	Trained from scratch on Indo4B	30,522	Standard random initiation	Bidirectional syntax/semantic modeling
IndoBART	BART-Base	Sequence-to-sequence vocab	40,000	Native token alignment	Multilingual regional translation & NLG
Komodo-7B	LLaMA-2-7B	+3,000 top Indonesian & regional words	35,008	Mean of source embeddings	Multilingual & regional dialect matching
SEA-LION	LLaMA-3 / Gemma	SEABPETokenizer adaptation	256,000	Bilingual data-mix embedding	Tonal & non-Latin script processing
Sailor-7B	Qwen1.5-7B	BPE Dropout vocabulary optimization	151,643	High-fidelity proxy tuning	Southeast Asian code-switching & QA

RESULTS AND DISCUSSION

This section analyzes the performance profiles of Indonesian language models across classification, style transfer, information extraction, and cultural alignment.

Sentiment Analysis and Sequence Labeling Benchmarks

Bidirectional encoders remain highly effective for discriminative NLP tasks. IndoBERT has been evaluated extensively across user-generated content (UGC), demonstrating superior accuracy compared to recurrent neural network (RNN) baselines.

For example, in fine-tuning experiments conducted by fine-tuning IndoBERT on tourism-related reviews from Google Reviews, researchers utilized a layer-freezing methodology to find the optimal configuration. Standard BERT-base architectures consist of 12 hidden layers with 768 dimensions each. By freezing the first six layers and fine-tuning only the top six layers, researchers achieved a validation loss of 0.324 and precision scores between 0.85 and 0.89 in aspect detection, with an overall classification accuracy of 84.0%.

Furthermore, combining IndoBERT with advanced head architectures yields significant performance gains. When paired with a Recurrent Convolutional Neural Network (RCNN) classifier, IndoBERT achieved 95.16% accuracy, 94.05% precision, 92.74% recall, and a 93.27% F1-score on IndoNLU sentiment datasets, outperforming the baseline encoder.

In another study analyzing sentiment from 2,560 reviews of Kenjeran Park in Surabaya (comprising 54% positive, 28% neutral, and 18% negative sentiments), a domain-specific pipeline was constructed with slang replacement, stemming, stopword removal, and tokenization, utilizing weighted loss adjustments to handle class imbalances. In this setup, IndoBERT (fine-tuned with a learning rate of 0.00005) achieved 87.50% accuracy, 0.7697 precision, 0.7643 recall, and a 0.7643 F1-score, significantly outperforming an LSTM network (128-unit architecture trained over 150 epochs) which achieved 77.93% accuracy.

Table 2: Comparative Performance Metrics of Fine-Tuned Encoders and Recurrent Baselines on Sentiment Datasets

Model Architecture	Task Dataset / Focus	Accuracy	Precision	Recall	F1-Score
IndoBERT + RCNN	IndoNLU Sentiment Dataset	0.9516	0.9405	0.9274	0.9327
IndoBERT (Unfrozen-6)	Aspect Detection (UGC)	0.8400	0.8700	0.8200	0.8440
IndoBERT (Fine-Tuned)	Kenjeran Park Review Sentiment	0.8750	0.7697	0.7643	0.7643
LSTM Baseline	Kenjeran Park Review Sentiment	0.7793	0.6826	0.6812	0.6826
LinearSVC + TF-IDF	Ruangguru Reviews (Class-Balanced)	0.8880	0.5120	0.5780	0.5430
IndoBERT (Class-Weighted)	Ruangguru Reviews (Class-Balanced)	0.8570	0.5890	0.6140	0.6010

These evaluations demonstrate that transformer-based representations handle the lexical variation in Indonesian UGC much more effectively than recurrent architectures, capturing bidirectional context and long-range syntactic dependencies.

To assess the model's ability to handle linguistic diversity across the Indonesian archipelago, researchers evaluated eight pre-trained models across ten regional languages using the NusaX multilingual sentiment dataset. The models tested included base and large variants of IndoBERT (IndoNLU), IndoBERT (IndoLEM), Multilingual BERT, and NusaBERT. The results demonstrated that models pre-trained on native Indonesian data consistently outperformed general multilingual baselines, with IndoBERT-large (IndoNLU) achieving the highest overall F1-score of 0.9353.

Performance varied considerably across regional dialects: Javanese, Minangkabau, and Banjar consistently yielded high F1-scores, whereas Batak Toba proved to be the most challenging language for all models. Notably, NusaBERT-base underperformed compared to

IndoBERT-base (IndoNLU) across all languages, despite being retrained on Indonesian regional languages, highlighting the importance of the initial pre-training corpus composition.

Extractive Question Answering and Corpus Composition

The composition of pre-training corpora heavily impacts a model's downstream task capability. A study compared two monolingual BERT models—**indobert-base-uncased** (IndoLEM) and **indobert-base-p1** (IndoNLU)—on an extractive QA task utilizing a translated and aligned SQuAD 2.0 dataset. Because machine translation can introduce answer-span alignment errors, fuzzy string matching was used to correct the target indices. Both models share identical architectural dimensions, but IndoLEM was pre-trained on over 220 million words of diverse text, whereas IndoNLU was pre-trained on the Indo4B corpus.

Under identical training schemes, IndoLEM achieved better loss convergence and a significantly higher F1-score of 71.58 compared to IndoNLU's F1-score of 63.59 (***p < 0.001***), illustrating that pre-training on diverse structural and stylistic datasets enhances contextual extraction.

Generative Style Normalization and Translation Pipelines

The expressive and evolving nature of Indonesian social media presents challenges for models trained on formal corpora. In sequence-to-sequence style transfer (translating informal colloquial slang to standard formal language), bidirectional context processing is highly advantageous. Evaluating the STIF-INDONESIA dataset, the encoder-decoder model IndoNanoT5-base achieved a peak BLEU score of 55.99, significantly outperforming the decoder-only IndoGPT's score of 51.13 by 4.86 points (***p < 0.001, Cohen's d = 0.847***). The bidirectional context and cross-attention mechanisms of the encoder-decoder model are crucial for capturing the morphological complexity of colloquial Indonesian.

To preserve and translate lower-resource regional dialects, researchers developed NusaMT-7B, a machine translation model for Balinese and Minangkabau. Leveraging Komodo-7B-base, which was pre-trained on 8.79 billion tokens, NusaMT-7B integrates supervised fine-tuning (SFT), self-learning, and an LLM-based data cleaner to reduce noise in parallel training pairs. On the FLORES-200 benchmark, NusaMT-7B outperformed standard baseline models by up to +6.69 spBLEU when translating into Balinese and Minangkabau, although it underperformed by up to -3.38 spBLEU when translating back into higher-resource languages, indicating a remaining asymmetry in cross-lingual representation.

Similarly, localized LLMs have been fine-tuned for specialized health and extraction applications. In a health chatbot implementation (Feminacare), transitioning from an LSTM model to SeaLLM with Retrieval-Augmented Generation (RAG) improved sentiment polarity and response relevance, raising accuracy from 61.0% to 87.0%. In fisheries data extraction, the Merak-7B model achieved 91.0% accuracy in structuring unstructured reports, demonstrating strong performance compared to larger general-purpose proprietary LLMs.

Diagnostic Analysis of Entity Linking Failures

Despite progress, localized models show distinct weaknesses when evaluated on the IndEL (ELEVATE-ID) benchmark for end-to-end Entity Linking (EL). The IndEL benchmark tasks models with identifying entity mentions in text and grounding them to their exact Wikidata QID links. In a comparative evaluation of multilingual LLMs (GPT-3.5, GPT-4, LLaMA-3) and monolingual Indonesian models (Komodo and Merak), all architectures performed poorly in zero-shot settings. Qualitative evaluations highlight structural issues: Komodo-7B exhibits "template leakage," accidentally outputting system prompts, whereas Merak-7B is prone to "entity hallucination," generating incorrect Wikidata QID links despite identifying the entity mentions.

This diagnostic analysis reveals that while vocabulary adaptation and continual pre-training improve fluency and surface-level syntax, they do not inherently improve factual grounding or complex symbolic reasoning.

Sociotechnical and Cultural Alignment

Adapting models to the diverse cultural landscape of Indonesia requires addressing geographic and sociolinguistic nuances. The Javanese language, for instance, relies on complex honorific levels (such as *Ngoko* for informal speech, and *Krama* for formal interactions) that govern social relationships, yet most public corpora contain severe level imbalances, causing models to output inappropriate registers. The Unggah-Ungguh corpus was developed to systematically evaluate model comprehension of Javanese honorifics based on speaker roles and social context.

Furthermore, the IndoCulture dataset, which evaluates cultural reasoning across eleven Indonesian provinces, demonstrates that many standard open-weight models score under 50.0% accuracy due to Anglocentric biases. Standard models often fail to capture regional common sense, culinary traditions, and local customs.

To address these limitations, models such as SEA-LION and Sailor implement regional context engines and code-switching strategies. SEA-LION incorporates a Cultural Context Engine to align output with regional guidelines, using a customized SEABPETokenizer and an expanded training dataset of over 1.4 million multilingual pairs. Similarly, Sailor utilizes document-level code-switching to improve multi-dialectal representation.

Table 3: Multi-Domain Performance Profiles and Downstream Capabilities of Native Indonesian Language Models

Evaluation Domain	Performance Metric	IndoBERT (Baseline)	Generative / Localized LLMs	Key Findings & Comparative Insights
Sentiment Analysis	Accuracy / F1-Score	95.16% (with RCNN)	87.50% (Komodo / SeaLLM)	Bidirectional encoder variants excel in token-level class decisions.
Colloquial Normalization	BLEU Score	N/A	55.99 (IndoNanoT5-base)	Encoder-decoder models outperform decoder-only models (IndoGPT: 51.13).
Factual Entity Linking	QID Link Accuracy	Poor / Fails baseline	Low-resource vulnerabilities	Models encounter template leakage.(Komodo) and hallucinations (Merak).
Regional Translation	F1 / spBLEU	0.9353 (IndoBERT-large)	High regional variance	Regional performance is high for Javanese but drops significantly for Batak Toba.
Geographic Reasoning	Multi-choice Accuracy	N/A	< 50.0% (standard open-weights)	Models show strong Anglocentric bias due to translated evaluation metrics.

CONCLUSION

The evolution of natural language processing for the Indonesian language highlights the limitations of standard multilingual LLMs and demonstrates the value of targeted, localized models. The transition from encoder-based models to generative architectures has improved the processing of both formal Indonesian and regional dialects.

The findings indicate that:

1. Text overfragmentation and high tokenization costs can be mitigated through cross-lingual vocabulary adaptation, using mean embedding initialization to keep the latent space stable.
2. Bidirectional encoder-decoder models are more effective than decoder-only models for morphological tasks like style normalization and machine translation.
3. Localized models still face challenges with factual entity grounding and complex regional honorific structures, highlighting the need for higher-quality, culturally aligned

Future research should focus on refining instruction-tuning datasets for regional languages, developing architectural solutions to reduce hallucination rates in entity linking, and optimizing parameter-efficient weight merging to support low-resource dialects without degrading general capabilities in training corpora.

REFERENCES

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 8875–8898. https://doi.org/10.18653/v1/2021.emnlp-main.699

Chakrabarty, A., Chaturvedi, A., & Garain, U. (2020). NeuMorph. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(1), 1–19. https://doi.org/10.1145/3342354

Chizhov, P., Arnett, C., Korotkova, E., & Yamshchikov, I. P. (2024). BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training. *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 16587–16604. https://doi.org/10.18653/v1/2024.emnlp-main.925

Di Marco, M., & Fraser, A. (2024). Subword Segmentation in LLMs: Looking at Inflection and Consistency. *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 12050–12060. https://doi.org/10.18653/v1/2024.emnlp-main.672

Ding, W., Wang, W., Kwok, S. H. D., Liu, M., Fang, T., Bai, J., Liu, X., Yu, C., Li, Z., Luo, C., Yin, Q., Yin, B., He, J., & Song, Y. (2024). IntentionQA: A Benchmark for Evaluating Purchase Intention Comprehension Abilities of Language Models in E-commerce. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2247–2266. https://doi.org/10.18653/v1/2024.findings-emnlp.123

Fridkin, S., & Bendersky, M. (2026). Interpretable Machine Learning: A Comprehensive Review of Foundations, Methods, and the Path Forward. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *16*(1). https://doi.org/10.1002/widm.70075

Fujii, K., Nakamura, T., Loem, M., Iida, H., Ohi, M., Hattori, K., Shota, H., Mizuki, S., Yokota, R., & Okazaki, N. (2024). *Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities*. http://arxiv.org/abs/2404.17790

Hasmawati, & Ade Romadhony. (2023). Similar Questions Identification on Indonesian Language Subject Using Machine Learning. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, *12*(2), 196–202. https://doi.org/10.23887/janapati.v12i2.62582

Jap, B. A. J., & Arumsari, C. (2017). Adaptation of the Token Test in Standard Indonesian. *Makara Human Behavior Studies in Asia*, *21*(1), 44. https://doi.org/10.7454/mssh.v21i1.3499

Jin, B. (2024). Multilingual Neural Machine Translation with Integrated Language Adapters. *Proceedings of the 2024 International Conference on Generative Artificial Intelligence and Information Security*, 29–35. https://doi.org/10.1145/3665348.3665355

Jo, E., Cho, E., Lee, Y., Song, S., & Joo, H. J. (2025). Domain and Language adaptive pre-training of BERT models for Korean-English bilingual clinical text analysis. *BMC Medical Informatics and Decision Making*, *25*(1). https://doi.org/10.1186/s12911-025-03262-7

Kartika, B. V., Alfredo, M. J., & Kusuma, G. P. (2023). Fine-Tuned IndoBERT based model and data augmentation for indonesian language paraphrase identification. *Revue d'Intelligence Artificielle*, *37*(3), 733–743. https://doi.org/10.18280/ria.370322

Lee, K. S. J. W., Qorib, M. R., Soegeng, A. I., & Ng, H. T. (2026). *Equity with Efficiency: An Empirical Study of Tokenizers for Multilingual Large Language Models*. http://arxiv.org/abs/2606.15044

Linder, M. R. (2026). *Vocabulary Expansion of Large Language Models via Kullback-Leibler-Based Self-Distillation*. http://arxiv.org/abs/2508.15807

McGiff, J., & Nikolov, N. S. (2025). *Overcoming Data Scarcity in Generative Language Modelling for Low-Resource Languages: A Systematic Review*. http://arxiv.org/abs/2505.04531

Mu, L., Wang, X., Ni, L., Li, Y., Wu, Z., Jin, P., & Zhang, Y. (2025). DenseLoRA: Dense Low-Rank Adaptation of Large Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 10198–10211. https://doi.org/10.18653/v1/2025.acl-long.503

Owen, L., Tripathi, V., Kumar, A., & Ahmed, B. (2024). *Komodo: A Linguistic Expedition into Indonesia's Regional Languages*. http://arxiv.org/abs/2403.09362

Petrov, A., La Malfa, E., Torr, P. H. S., & Bibi, A. (2023). *Language Model Tokenizers Introduce Unfairness Between Languages*. http://arxiv.org/abs/2305.15425

Provilkov, I., Emelianenko, D., & Voita, E. (2020). BPE-dropout: Simple and effective subword regularization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 1882–1892. https://doi.org/10.18653/v1/2020.acl-main.170

Ridwan, M. (2018). National and Official Language: The Long Journey of Indonesian Language. *Budapest International Research and Critics Institute (BIRCI-Journal) : Humanities and Social Sciences*, *1*(2), 72–78. https://doi.org/10.33258/birci.v1i2.14

Subhan Mahendrasyah, M., & Hariguna, T. (2024). Analisis Sentimen Pengguna Aplikasi Bukalapak di Platform Playstore Menggunakan Metode Naïve Bayes. *Building of Informatics, Technology and Science (BITS)*, *6*(2), 733–745. https://doi.org/10.47065/bits.v6i2.5528

Xu, Y., & Che, W. (2026). Evolving LLMs from Next-Token Prediction to Multi-Token Prediction via Self-Distillation. *Electronics (Switzerland)*, *15*(7). https://doi.org/10.3390/electronics15071533

Yamaguchi, A., Villavicencio, A., & Aletras, N. (2026). How Can We Effectively Expand the Vocabulary of LLMs with 0.01GB of Target Language Text? *Computational Linguistics*, *52*(1), 295–330. https://doi.org/10.1162/COLI.a.581

Zhang, L., Lou, Z., Ying, Y., Yang, C., & Zhou, H. (2025). Efficient Fine-Tuning of Large Language Models via a Low-Rank Gradient Estimator. *Applied Sciences (Switzerland)*, *15*(1), 1–16. https://doi.org/10.3390/app15010082