Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29591

# *Efficient Architectures and Training Approaches for Multilingual Neural Language Models*

**Yi Wang**

*Detroit Green Technology Institute, Hubei University of Technology, Wuhan, China*
*yeexiao111@gmail.com*

***Abstract.*** Multilingual neural language models have become essential in modern natural language processing, facilitating cross-lingual understanding and knowledge transfer. The rapid advancement of pre-trained models such as mBERT, XLM-R, and mT5 has greatly enhanced the efficiency of handling multilingual tasks. This paper covers three main areas, including model architectures, training strategies, and evaluation methods. In particular, it compares various architectures, explores how training goals and data balancing strategies impact model performance, and explores the balance between model size and multilingual generalization. Besides, it discusses how to achieve fair evaluation across various language types. The results indicate the challenges of multilingual neural language models in terms of architecture, training strategies, and evaluation methods, that different architectures greatly affect cross-lingual performance, while also exploring the trade-off between model size and multilingual generalization. Additionally, the paper proposes approaches for achieving fair evaluation.

***Keywords:*** Multilingual Language Models, Neural Networks, Transformer Architecture, Cross-lingual Transfer, NLP Evaluation

## 1. Introduction

There are more than 7,000 languages worldwide, hence making the development of a multilingual language system essential for the advancement of Natural Language Processing (NLP). Traditional language models have primarily focused on high-resource languages, such as English, which has resulted in significant gaps in the development of language models for other languages. At present, studies on Multilingual Neural Language Models (MNLMs) have made significant progress in areas such as model architecture design, training data strategies, as well as evaluation methods. However, challenges persist in managing knowledge transfer between languages and in developing effective training data strategies to enhance the performance of low-resource languages. Besides, the fairness of cross-lingual evaluation methods remains a key concern, particularly in ensuring fairness and consistency across different languages in model performance. Accordingly, this paper explores the development of MNLMs, while highlighting the analysis of current model architectures, training data strategies, and their performance across different languages. In addition, it investigates the fairness of existing cross-lingual evaluation methods. Through a literature review and case analysis, this paper aims to provide valuable insights for future research in the field.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29591

## 2. The architecture of multilingual neural language models

### 2.1. Shared encoder architecture

The shared encoder architecture facilitates cross-lingual knowledge transfer via parameter sharing, making it a highly effective approach for low-resource language modeling. This architecture utilizes a single Transformer encoder to process multilingual input, reducing the model's parameter count while enhancing transferability across languages. By sharing parameters, the model can effectively transfer knowledge learned from high-resource languages to low-resource ones, thereby improving performance on low-resource languages.

In this architecture, the prominent models are Multilingual BERT (mBERT) and Cross-lingual Language Model-RoBERTa (XLM-R) [1,2]. Based on the BERT architecture, mBERT leverages Wikipedia data from 104 languages and a common WordPiece vocabulary to support cross-lingual representation learning. Besides, XLM-R improves upon this by utilizing the larger CC-100 corpus (covering 100 languages) along with longer training times and larger batch sizes, which boots the model's performance on cross-lingual tasks, particularly for low-resource language transfer learning. In terms of its advantages, the shared encoder architecture significantly reduces the model's storage and training costs by decreasing its complexity via parameter sharing. Meanwhile, this architecture enables efficient knowledge transfer from high-resource to low-resource languages, boosting cross-lingual transferability. Furthermore, separate models or word embeddings for each language are not required, simplifying the modeling process and improving development efficiency. However, there are limitations. Firstly, interference between languages may affect the performance of low-resource languages,especially when there are substantial feature variations between languages. Secondly, for languages with significant morphological differences, shared encoders often struggle to capture their unique linguistic characteristics, which can negatively impact model performance.

### 2.2. Encoder-Decoder architecture

The Encoder-Decoder architecture is an important structure for cross-lingual generation tasks. Its basic principle is to convert input sequences into cross-lingual semantic representations through the encoder, and then generate target language outputs using these representations via the decoder. This makes the Encoder-Decoder architecture particularly well-suited for cross-lingual tasks that require generative outputs, such as machine translation, cross-lingual text summarization, and cross-lingual question answering.

In this architecture, mT5 and mBART are two typical representative models [3,4]. mT5 is the multilingual version of T5, which adopts the text-to-text paradigm, treating tasks like translation, summarization, and question answering as tasks that take in text and output text. It is pre-trained on the mC4 corpus, spanning 101 languages, and uses a pure generation objective with span corruption, giving it strong cross-task adaptability. Moreover, mBART is the multilingual version of BART, pre-trained on the CC25 corpus covering 25 languages. By combining denoising autoencoders with the encoding-decoding process, this model performs strongly in low-resource machine translation and cross-lingual generation. However, compared to the shared encoder architecture, the Encoder-Decoder architecture offers three key advantages. First, it naturally suits situations where the input and output languages differ, and can control target language generation via conditional generation. Second, the Encoder-Decoder architecture enables knowledge transfer from high-resource to low-resource languages, greatly improving low-resource translation quality. Finally, its task adaptability is strong, as it can easily accommodate different task requirements within the same framework by

DOI: 10.54254/2755-2721/2026.TJ29591

simply altering input formats or prompts. Thus, this architecture is key for multilingual generation tasks such as translation, summarization, question answering, and paraphrase generation.

### 2.3. Byte- and character-level models

Byte- or character-level language models operate without word segmentation tools, modeling text directly at the byte or character level. More explicitly, this approach uses UTF-8 bytes or Unicode characters as basic units, providing a unified input space for all languages. Compared with word or subword-level models, it handles diverse writing systems and morphological traits more effectively, while addressing vocabulary gaps and greatly lowering out-of-vocabulary (OOV) problems.

For example, representative models include ByT5 and Canine [5,6]. ByT5, a variant of T5, keeps the Encoder-Decoder architecture while representing input sequences using UTF-8 byte strings. The increased input length requires ByT5 to adopt more efficient position encoding and optimization strategies, resulting in superior performance in low-resource translation and noisy text processing, such as optical character recognition (OCR) error correction. In addition, Canine employs a fully character-level Transformer architecture, which compresses long character sequences into shorter hidden representations via subsampling, reducing computational cost while achieving performance comparable to or surpassing subword-level models in multilingual tasks. Thus, byte/character-level architectures offer notable advantages over other models in low-resource language modeling and cross-lingual transfer. By using unified encoding, lexical differences across languages are reduced and cross-lingual transfer is improved. And it also increases robustness to spelling variations and other input noise, thereby making it effective for social media texts, OCR, complex languages, and cases with poor word segmentation.

## 3. Training strategies for multilingual models

### 3.1. Pre-training target construction mechanism

By aligning pretraining objectives with specific application scenarios, models can more effectively improve cross-lingual representation, alignment, and generation, as various objectives are suitable for different model architectures and task types.

In Masked Language Modeling (MLM), the model is trained to predict tokens that have been randomly masked in the input sequence [1,2]. Due to its generality, it is the most frequently adopted pretraining technique for learning multilingual representations. However, as the training objective is largely restricted to a single language context, its ability in cross-lingual alignment is comparatively low. In contrast, Translation Language Modeling (TLM) masks certain tokens in bilingual parallel corpora, enabling the model to predict in a cross-lingual context, thereby explicitly enhancing cross-lingual alignment [7]. It is suitable for translation and cross-lingual retrieval tasks. However, TLM's reliance on large-scale parallel corpora limits its applicability in low-resource language scenarios. Meanwhile, Sequence-to-Sequence (Seq2Seq) objectives introduce noise into the input sequence like deletion, replacement, or span shuffling and train the model to reconstruct the original sequence [4,5]. As such, this model enhances cross-lingual generation capabilities and is ideal for generative tasks, such as translation, summarization, and question answering. For example, the span corruption mechanism in mT5, which perturbs a group of tokens simultaneously rather than predicting words independently, improves cross-task adaptability. Moreover, Contrastive Learning pulls semantically similar cross-lingual sentences closer in vector space, significantly boosting cross-lingual semantic

DOI: 10.54254/2755-2721/2026.TJ29591

retrieval and matching capabilities. This approach is particularly suitable for cross-lingual retrieval or semantic matching tasks, as used in models like InfoXLM and LaBSE [8,9].

### 3.2. Data sampling and balancing approaches

In multilingual neural language model training, data imbalance is especially pronounced, with high-resource languages like English and Chinese far exceeding low-resource ones. Due to this long-tail distribution, models tend to overfit on high-resource languages, which influences the representation quality of low-resource languages and their cross-lingual transfer capabilities. Hence, multilingual pretraining faces the challenge of balancing data distribution with effective sampling. In particular, Temperature Sampling adjusts sampling by scaling the language distribution exponentially, with $\alpha \in [0,1]$ as the temperature parameter. For instance, mBERT sets $\alpha = 1$ to maintain the original distribution, whereas XLM-R adopts $\alpha = 0.7$ to favor low-resource languages without losing high-resource language dominance [1,2]. Despite its simplicity and wide use in large-scale pretraining, temperature sampling lacks a consensus regarding the optimal $\alpha$ value. In addition, Smoothing and Resampling tackle imbalance by oversampling low-resource languages, thus undersampling high-resource ones, or using stratified sampling. For example, mT5 oversamples low-resource languages to prevent underuse, thus improving performance but risking noise and overfitting if excessive [3]. Besides, Task-aware Sampling can adjust the sampling distribution based on the requirements of downstream tasks. In cross-lingual retrieval, methods like InfoXLM sample semantically similar languages more and form cross-lingual pairs using task information, improving transfer but relying on prior knowledge with limited scope [8]. These methods suit task needs and improve transfer but rely on prior knowledge and have limited scope. Moreover, Adaptive Sampling adjusts sampling based on language-specific loss, perplexity, or gradients during training. For example, training high-loss languages more through a loss-based method helps the model focus on difficult languages, but it increases computational costs and slows training convergence [10].

### 3.3. Fine-tuning strategies and adaptation techniques

Fine-tuning and adaptation techniques can markedly enhance the cross-lingual transfer efficiency of multilingual models and improve their performance in low-resource scenarios. Given the extremely large parameter sizes of multilingual pre-trained models such as mBERT, XLM-R, and mT5, fine-tuning in low-resource settings has become a research focus, aiming to reduce training and storage costs while maintaining performance. Full fine-tuning updates all model parameters for the target task, achieving optimal performance on tasks like GLUE and XNLI. However, it comes with high computational and storage costs, making it impractical for resource-limited or multi-task scenarios. As a result, Parameter-Efficient Fine-Tuning (PEFT) adapts the model by updating only a small set of additional parameters. For example, it can add Adapter modules between Transformer layers (MAD-X), use low-rank weight matrices (LoRA), or attach trained prefixes/prompts to the input (Prefix-tuning/Prompt-tuning). This method greatly reduces computational and storage costs. While slightly less effective than full fine-tuning on some tasks, it remains highly efficient. Furthermore, cross-lingual transfer fine-tuning first trains on a high-resource language like English, then transfers to a low-resource language, enabling zero- or few-shot transfer. For instance, XLM-R shows strong cross-lingual transfer capability on the XNLI task, requiring no low-resource language annotations [1]. However, its effectiveness relies on language similarity. Multi-task and continual learning fine-tune multiple tasks or languages together, or slowly adapt to new tasks and languages. For example, multi-task fine-tuning covering terminology translation and cross-lingual question answering can

DOI: 10.54254/2755-2721/2026.TJ29591

enhance model generalization and robustness, though it may introduce task interference (negative transfer) [11]. Consequently, full fine-tuning is most effective but costly, while PEFT and cross-lingual transfer offer efficient alternatives, and multi-task learning boosts adaptability.

## 4. Evaluation methods for multilingual models

### 4.1. Benchmark datasets and evaluation tasks

In order to evaluate the cross-lingual capability and downstream adaptability of multilingual models, standardized benchmark datasets are widely used. Typically, evaluations compare different models via cross-task experimental designs to ensure comparability under identical conditions. Specifically, XNLI covers 15 languages, focusing on cross-lingual NLI to assess entailment, contradiction, or neutrality in sentence pairs, including low-resource languages [12]. Meanwhile, MLQA, including 7 languages, tests cross-lingual retrieval and reasoning by extracting answers from texts in various languages [13]. In addition, XTREME covers 40 languages across tasks such as text classification, sequence labeling, cross-lingual question answering, and sentence retrieval, serving as a standard benchmark for model performance and generalization [14]. And the extended version, XTREME-R, better aligns task design and language coverage with real-world needs [15]. Overall, XNLI and MLQA form single-task benchmarks that test models' cross-lingual transfer ability on specific tasks, while XTREME and its extensions offer broad performance evaluation. Hence, the two benchmark types enable comparison of models on semantics, cross-lingual transfer, and task adaptability.

### 4.2. Cross-lingual metrics and fairness analysis

In cross-lingual evaluation, it is important to consider accuracy along with performance disparities and fairness across languages. Traditional metrics such as accuracy and F1 scores exhibit overall performance but miss gaps between high- and low-resource languages. Thus, cross-lingual metrics such as minimum language score, performance gap, and RERR help assess bias or discrimination against specific languages. In the XNLI task, XLM-R improved average accuracy by about 14.6% over the previous XLM model, with notable gains of 15.7% for Swahili and 11.4% for Urdu, both low-resource languages [1]. This indicates that XLM-R exhibits stronger cross-lingual transfer for low-resource languages compared to high-resource languages, yet Swahili and Urdu still lag behind high-resource language performance, highlighting room for improvement in fairness. Consequently, a comprehensive evaluation framework should include both traditional performance metrics, like average accuracy, and cross-lingual fairness metrics, such as minimum language score, performance gap, and RERR. Both types of metrics help mitigate bias, prevent false performance impressions, and promote more inclusive, equitable multilingual NLP development.

### 4.3. Evaluation dilemma for low-resource languages

Low-resource language evaluation is an underexplored challenge in multilingual model research. The main difficulties include data scarcity, linguistic diversity, and corpus bias. Most low-resource languages lack high-quality annotated corpora or established benchmark datasets, making reliable testing and comparison difficult. Moreover, languages with complex morphological structures, such as agglutinative or polysynthetic languages, require fine-grained and specialized modeling, whereas existing evaluation frameworks are primarily designed for high-resource languages and often target written forms. The available data for low-resource languages, primarily from sources such as Bible translations or Wikipedia, usually differs from real-world scenarios, leading to non-representative

DOI: 10.54254/2755-2721/2026.TJ29591

evaluation outcomes. As a result, benchmark results for low-resource languages often fail to reflect the true robustness and generalization capabilities of multilingual models in practice. Even broader benchmarks like XTREME or FLORES cover only a limited subset of low-resource languages. For instance, certain African and South Asian languages in XTREME have XNLI test sets with under 2,500 sentence pairs per language. This is much less than high-resource languages like English or Chinese, limiting the reliability of conclusions on cross-lingual performance for underrepresented languages [15]. Thus, cross-lingual transfer evaluation can infer low-resource language capabilities from high-resource language performance. Meanwhile, data augmentation strategies like synthetic data generation via back-translation or LLM-based paraphrasing, and crowdsourced annotation, can expand evaluation corpora. Similarly, diversified evaluation frameworks that cover broader cultural and linguistic variety can include rare language families and better reflect real-world usage.

## 5. Challenges and prospects in multilingual modeling

### 5.1. Language interference and representation bias

Multilingual neural language models encounter two main challenges in parameter sharing: language interference and representation bias. Specifically, language interference occurs when high-resource languages affect updates in the shared encoder, weakening or distorting representations of low-resource languages. This effect is stronger with minimal lexical overlap or typological differences. The representational bias arises from the structural imbalance in the embedding space. In this case, high-resource language embeddings form tight clusters, while low-resource ones are scattered and unclear, causing poor alignment and noise. Such imbalances reduce the accuracy of cross-lingual tasks, such as retrieval and translation, and weaken semantic transfer between linguistically distant languages. In multilingual embedding spaces, high-resource languages like English and Chinese form tight clusters, while low-resource languages such as Swahili or Urdu show dispersed, unclear embeddings, leading to poor alignment and noise. To address these issues, several strategies have been proposed. These strategies include adding language-adaptive layers to reduce interference, employing adversarial training and regularization to promote language-independent features, and applying cross-lingual alignment constraints from parallel corpora or dictionary supervision.

### 5.2. Improvement strategies for low-resource modeling

Low-resource language modeling faces challenges such as limited data, complex morphology, and few mature linguistic tools. Current research mainly addresses these issues through three strategies. Firstly, knowledge transfer leverages high-resource languages to support low-resource modeling, typically through cross-lingual pre-training and transfer learning. Secondly, input modeling adopts subword or byte-level units rather than fixed vocabularies, helping with morphologically rich or under-segmented languages. Thirdly, efficient adaptation for low-resource scenarios tunes only a small set of parameters to boost computational efficiency and reduce the practical cost of adapting to low-resource tasks. These strategies are not isolated but can work synergistically. For example, transfer learning can be combined with byte-level modeling to improve cross-lingual generalization, while parameter-efficient fine-tuning enables compact adaptation. Looking ahead, the integration of these strategies into a unified framework presents a promising direction, thus allowing low-resource language modeling to shift from ad hoc methods to a more principled and generalizable approach.

DOI: 10.54254/2755-2721/2026.TJ29591

### 5.3. Model efficiency and scalability

As multilingual models grow, the gap between efficiency and scalability becomes more apparent. On one hand, larger models offer stronger cross-lingual transfer capabilities; on the other hand, their training, inference, and energy costs rise rapidly, thus making sustainability a concern. The balance between performance and computational and energy efficiency has become a key research frontier. For computational efficiency, sparsity, model compression, and optimized training algorithms can significantly reduce costs. Inference efficiency can be improved through quantization, pruning, and knowledge distillation, while mixed-precision training, checkpointing, and large-scale optimization reduce resource needs, enabling efficient training on massive multilingual corpora. These strategies align with the principles of green AI, promoting energy-efficient and environmentally sustainable model deployment. From a scalability perspective, the challenge extends beyond the total number of parameters to include multilingual coverage. As the number of languages increases, particularly low-resource languages, maintaining model adaptability without resulting in catastrophic forgetting becomes crucial. In this context, incremental training, continual learning, and modular architectures play a key role in adding new languages while preserving existing knowledge.

## 6. Conclusion

This paper has examined key issues in multilingual neural language models concerning architecture, training, and evaluation. The results indicate that shared-parameter models perform well in cross-lingual transfer but provide limited support for low-resource languages, while encoder-decoder structures are better suited for generative tasks, and byte- or character-level models show potential in handling morphologically complex languages. At the training stage, pre-training objectives and sampling strategies play a crucial role in cross-lingual generalization, and parameter-efficient fine-tuning offers practical solutions for low-resource applications. While effective for high-resource languages, existing evaluation frameworks underperform on low-resource languages and non-Latin scripts, limiting real-world relevance. However, cross-lingual interference and representation bias create imbalances in the embedding space, reliance on limited corpora affects the robustness of low-resource experiments, and efficiency and scalability continue to constrain practical deployment. Thus, future studies should emphasize developing comprehensive evaluation methods, leveraging cross-modal transfer and knowledge distillation to improve low-resource modeling, and designing architectures that are both efficient and scalable. With these improvements, multilingual models can better support cross-lingual communication and promote equitable access to global information.

## References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 4171-4186.
[2] Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440-8451.
[3] Xue, L., Constant, N., Roberts, A., et al. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv: 2010.11934.
[4] Liu, Y., et al. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726-742.
[5] Xue, L., et al. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics, 10, 291-306.
[6] Clark, J. H., et al. (2022). CANINE: Pre-training an efficient tokenization-free encoder for language representation. Transactions of the Association for Computational Linguistics, 10, 73-91.

DOI: 10.54254/2755-2721/2026.TJ29591

[7] Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. Proceedings of the International Conference on Neural Information Processing Systems, 634, 7059-7069.
[8] Chi, Z., Dong, L., Wei, F., & Wang, W. (2021). InfoXLM: An information-theoretic framework for cross-lingual language model pre-training.arXiv preprint arXiv: 2007.07834.
[9] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. arXiv preprint arXiv: 2007.01852.
[10] Elharrouss, O., Mahmood, Y., Bechqito, Y., et al. (2025). Loss Functions in Deep Learning: A Comprehensive Review. arXiv preprint arXiv: 2504.04242.
[11] Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv: 1901.11504.
[12] Conneau, A., et al. (2018). XNLI: Evaluating cross-lingual sentence representations. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
[13] Lewis, P., Oguz, B., Rinott, R., et al. (2020). MLQA: Evaluating cross-lingual extractive question answering. arXiv preprint arXiv: 1910.07475.
[14] Siddhant, A., et al. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In Proceedings of the International Conference on Machine Learning, 4411-4421.
[15] Ruder, S., et al. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. arXiv preprint arXiv: 2104.07412..