

Optimizing Cross-Lingual Transfer Learning with Transformer Architectures for Low-Resource Natural Language Processing

Muhammad Essa Siddique

Computer Science & IT Department, University of Balochistan Kharan Campus, Balochistan.

Essasiddique@live.com

ABSTRACT

Modern Natural Language Processing (NLP) faces a stark digital divide, leaving the vast majority of the world's seven thousand languages digitally marginalized due to data scarcity and architectural biases. This review paper examines state-of-the-art cross-lingual transfer learning (CLTL) optimization methodologies designed to mitigate these resource constraints utilizing transformer architectures. We provide a comprehensive taxonomy categorizing transfer paradigms into instance, feature, and parameter-based strategies, highlighting advanced frameworks like colexification graphs and multilingual embedding alignment. Furthermore, we critically analyze structural bottlenecks such as the "curse of multilinguality," demonstrating how capacity competition and script barriers induce negative cross-lingual interference. To overcome the tight coupling of tokenizers to high-resource languages, this paper reviews vocabulary adaptation techniques that mitigate over-fragmentation and inference slowdown. We evaluate parameter-efficient fine-tuning (PEFT) configurations identifying the Houlsby+Inversion adapter as a top-performing structure alongside multi-stage curriculum alignment frameworks like Progressive Code-Switching and MERLIN. Finally, we expose systemic gaps in the current evaluation ecosystem, specifically looking at coverage and task equity disparities, while advocating for culturally grounded, native-authored benchmarks. Ultimately, this synthesis demonstrates that moving away from static translation paradigms toward modular architectures, optimized tokenization transplantation, and structured alignment curricula is vital for building equitable, globally inclusive language technology.

Keywords: Cross-Lingual Transfer Learning; Low-Resource NLP; Transformer Architectures; Parameter-Efficient Fine-Tuning; Vocabulary Adaptation; Curse of Multilinguality.

1. Introduction

Modern Natural Language Processing (NLP) is characterized by a stark digital divide. While high-resource languages such as English, Chinese, and Spanish enjoy state-of-the-art applications driven by foundation models, massive internet corpora, and scaling

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

laws, the majority of the world's seven thousand languages remain digitally marginalized (Joshi et al., 2020). This disparity, deeply bound to socio-political power structures, exacerbates digital exclusion and threatens linguistic diversity.

1.1 Characterizing Low-Resource Languages

Technically, a language is classified as low-resource when it lacks sufficient digitized text, computational tools, standardized orthographies, or specialized linguistic expertise (Gottam, 2026). Socio-political factors frequently dictate resource allocation, meaning that even widely spoken languages can be digitally invisible if they are absent from web content, underrepresented in public datasets, or excluded from multilingual benchmarks (Chitale et al., 2025). Many low-resource languages exist primarily through oral traditions; their native speakers communicate verbally but may lack standardized written systems. Consequently, constructing traditional static datasets via extractive web scraping is often unfeasible.

1.2 Resource Taxonomy

To capture this socio-technical variation, researchers utilize multidimensional classification schemes (Petti, 2026). For instance, languages can be categorized based on the availability of task-specific annotations, unannotated texts, and auxiliary data. A prominent taxonomy proposed by computational linguists groups the world's languages into six resource classes, placing the vast majority of African, indigenous, and regional languages into the most critical categories (Blasi et al., 2021).

Table 1: Taxonomy of Language Resource Classes and Computational Bottlenecks

Language Resource Class	Characterization of Resources	Regional & Typological Representatives	Primary Computational Bottlenecks
Class 0: Left-Behinds	Near-complete absence of digital text, tokenizers, or parallel corpora.	Native languages of Oceania, the Americas, and Central Asia.	Total lack of standardized orthographies; oral-only tradition.
Class 1: Scraping-Bys	Extremely limited web presence; fragmented unannotated corpora.	Pulaar (Fula), Sereer, Diola, Soninke (Senegal).	Over-fragmentation by standard tokenizers; script barriers.
Class 2: Hopefuls	Emerging community initiatives; basic monolingual datasets available.	Wolof (Senegal), Kikuyu, Dholuo, Kamba (Kenya).	Lack of high-quality task-specific annotations and evaluation benchmarks.
Class 3: Scraping-Bys / Rising	Moderate unannotated text; baseline multilingual model representation.	Swahili, Tamil, Yoruba, Sinhala.	Capacity bottlenecks and cross-lingual interference.

Online ISSN

3007-3197

Print ISSN

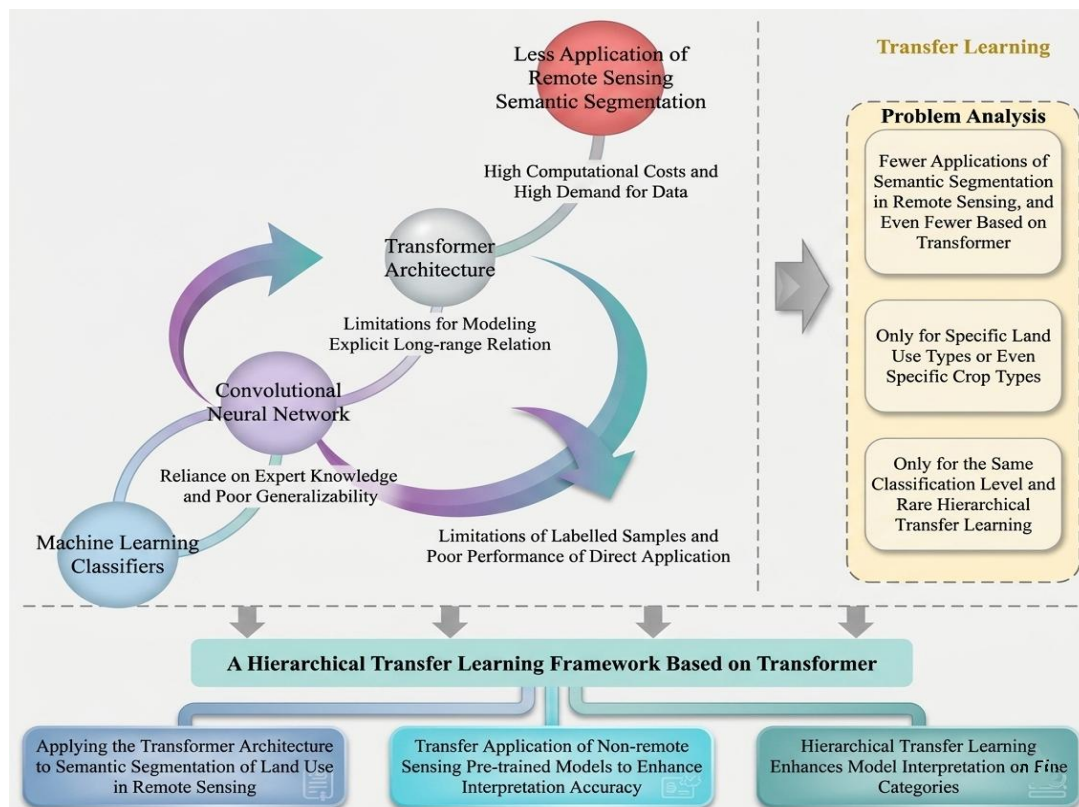
3007-3189

<http://amresearchreview.com/index.php/Journal/about>

1.3 Interactive Language Discovery

To bypass the limitations of static data collection, a paradigm shift toward open-ended, interactive language discovery has been proposed. Rather than extracting knowledge from pre-collected corpora, AI systems can dynamically learn under-documented languages through natural dialogue (Pakray et al., 2025). This interactive framework is grounded in joint human-machine uncertainty, combining epistemic uncertainty from the model with hesitation cues and confidence signals from human speakers. Such co-adaptive processes empower communities while discovering and preserving linguistic structures in real time (Mahmudi, 2026).

Figure 1. A Reconstructed Multidimensional Taxonomy Mapping Problem Analysis Bottlenecks to Hierarchical Target-Specific Downstream Tasks



2. Taxonomy of Cross-Lingual Transfer Learning

Transfer learning has emerged as the leading framework to mitigate data scarcity in low-resource NLP. This paradigm leverages knowledge acquired by high-capacity transformer architectures pre-trained on high-resource source languages and adapts it to low-resource target languages. Cross-Lingual Transfer Learning (CLTL) techniques can be classified based on "what to transfer," dividing methodologies into instance, feature, and parameter transfer (Dossou & Aïdasso, 2025).

2.1 Instance Transfer

Instance transfer focuses on migrating training examples or labels from the source

domain to the target domain. This is subdivided into:

Label-based transfer: Employs annotation projection where source-language labels are mapped to target-language texts via word-alignment models and pseudo labeling, which uses a source model to generate silver-standard labels on target-language inputs (Zhu et al., 2025).

Text-based transfer: Heavily relies on Machine Translation (MT) pipelines. Developers either translate source-language training data into the target language (translate-train) or translate target-language test inputs into the source language for evaluation (translate-test). Alternatively, direct text alignment leverages parallel sentences to bridge the representation gap (Padó & Lapata, 2009).

2.2 Feature Transfer

Feature transfer seeks to construct language-agnostic representation spaces where source and target tokens are close semantically. Early systems aligned bilingual static word embeddings (such as Word2Vec or GloVe) using linear mapping and bilingual dictionaries. Contextualized feature transfer leverages deep encoders (e.g., mBERT, XLM-R, or Language-Agnostic BERT Sentence Embeddings, known as LaBSE) to learn aligned multilingual contextual representations directly from massive unannotated corpora (Litschko et al., 2021).

An advanced form of feature alignment involves mapping cross-lingual colexification patterns. Colexification occurs when a single lexical form expresses multiple distinct concepts in a language. By analyzing colexification patterns of over 2,000 concepts across 1,335 languages extracted directly from the Parallel Bible Corpus (PBC), researchers constructed massive multilingual graphs (Tjuka et al., 2025). These graphs, instantiated as ColexNet and ColexNet+, are used to train high-quality multilingual word embeddings. These embeddings capture semantic links across distant language families and considerably improve cross-lingual transfer on downstream tasks, achieving high recall on comparative linguistic datasets like CLICS (Liu et al., 2023).

2.3 Parameter Transfer

Parameter transfer involves sharing model weights across languages. In a standard setting, a large transformer model pre-trained on multiple languages is fine-tuned on task-specific data in a high-resource language and applied directly to a low-resource language in a zero-shot or few-shot manner. Advanced parameter transfer introduces language-specific or task-specific modules such as bottleneck adapters, low-rank parameters, or soft prompts to control representation updates and target specific language families (Ebing & Glavaš, 2025).

2.4 Downstream Application: Cross-Lingual Information Retrieval

Cross-Lingual Information Retrieval (CLIR) represents a critical downstream application of these transfer paradigms. Traditional CLIR systems translated queries or documents to establish matching spaces. State-of-the-art CLIR bypasses translation altogether by utilizing multilingual embedding alignment and contrastive learning. Encoders are trained to map queries in language A and documents in language B into a shared dense vector space. Generative LLMs are then integrated at the end of the pipeline to reformulate queries, re-rank candidates, and generate synthesized answers across language barriers (Foroutan et al., 2025).

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

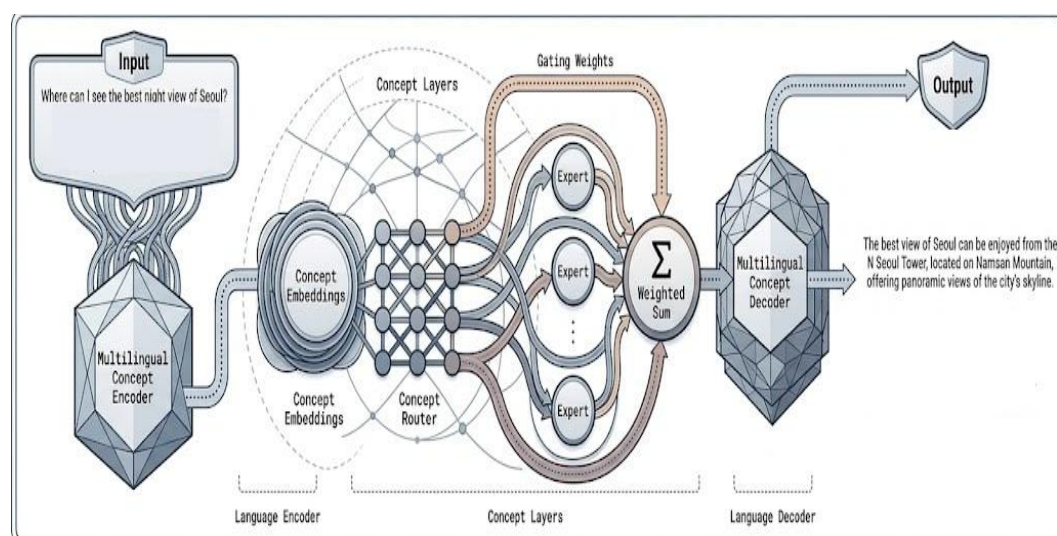
3. Structural Bottlenecks and the Curse of Multilinguality

The standard approach to training multilingual models involves optimizing an encoder or decoder architecture on a massive, combined corpus of multiple languages. During this process, the model learns shared semantic representations. However, this approach faces a fundamental trade-off between language coverage and per-language performance, known as the "curse of multilinguality" (Conneau et al., 2020).

3.1 Multilingual Model Performance

For a fixed model capacity, adding more languages to the pre-training mixture forces the model to allocate its parameters across highly diverse linguistic structures. Initially, as the language count increases from a monolingual baseline to approximately 15 related languages, the model benefits from positive transfer (Blevins et al., 2024). This positive transfer allows low-resource languages to borrow semantic representations from their high-resource counterparts. Beyond this threshold, severe parameter competition occurs, and negative interference degrades performance across all represented languages (Conneau et al., 2020).

Figure 2: Conceptual framework of a multilingual Mixture-of-Experts (MoE) architecture leveraging shared concept embeddings, routing mechanisms, and gated decoding for cross-lingual generation.



Controlled growth experiments on models ranging from 1.1 billion to 3 billion parameters demonstrate that this curse is capacity-related. While increasing the raw token budget per language can stabilize performance, capacity constraints ultimately limit the representational fidelity of highly multilingual models, necessitating modular parameters (Li et al., 2025).

3.2 The Interference Matrix

To analyze these interactions, researchers construct an interference matrix (I) across diverse languages. This matrix is derived from a loss matrix (L). By training shallow, two-layer transformer encoders on all possible language pairs (using, for example, 83 languages intersected from the XLM-R tokenizer and NLLB datasets with 1 million

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

samples per language), researchers evaluate performance on the FLORES-200 devtest set (Janeiro et al., 2025). The pairwise interference score is defined as: Asymmetry of interference: Cross-lingual interference is highly asymmetrical. The impact of language j on i is not the same as the impact of i on j . For instance, Welsh behaves as a low-interference donor (meaning it does not significantly degrade other languages when co-trained) but acts as a high-interference recipient (being highly susceptible to degradation from other languages) (Qin et al., 2025).

The script barrier: Interference patterns do not align with traditional genetic language families, nor do they correlate with simple proxies such as embedding similarity. Instead, they correspond closely to the script used. Aggregating the interference matrix by writing systems reveals a strong diagonal; same-script languages interfere significantly less. Differences in writing systems inhibit cross-lingual transfer, as the lack of lexical overlap forms a "script barrier" that hinders parameter sharing (Ebing et al., 2026).

Vulnerability of low-resource languages: Low-resource languages are highly vulnerable to negative interference, showing higher average validation drops and being more likely to degrade the performance of other languages when model capacity is saturated (Alastruey et al., 2025). These interference scores can effectively predict performance drops in downstream tasks, serving as a tool to guide the selection of language mixtures during pre-training (Pfeiffer et al., 2022).

4. Vocabulary Adaptation and Tokenizer Transplantation

The tight coupling of transformer architectures to fixed tokenization schemes represents a major hurdle in cross-lingual transfer. Standard multilingual tokenizers are heavily biased toward high-resource, Latin-script languages. Consequently, they over-fragment texts in low-resource target languages, splitting words into excessive, semantically meaningless subwords (Sheng, 2026).

Over-fragmentation increases the sequence length in terms of tokens, leading to an average inference slowdown of up to 271.5% in generative Large Language Models (LLMs) (Yamaguchi et al., 2024). It also degrades semantic representation quality because the model's self-attention layers must span excessively long sequences of fragmented subwords.

Cross-Lingual Vocabulary Adaptation (CVA), also termed vocabulary expansion, addresses this by modifying or replacing the model's tokenizer (Yamaguchi et al., 2025). The process typically involves training an auxiliary tokenizer on a target-language corpus (D), constructing an expanded vocabulary by merging the source vocabulary with the top- k most frequent target-specific tokens, expanding the model's embedding and language modeling head matrices, initializing the newly introduced token embeddings, and performing Language-Adaptive Pre-Training (LAPT) (Sakajo, 2025).

Table 2: Comparative Analysis of Vocabulary Adaptation Methods

Vocabulary	Underlying	Data	&	Impact	on	Downstream
------------	------------	------	---	--------	----	------------

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Adaptation Method	Mathematical or Algorithmic Mechanism	Linguistic Resource Requirements	Inference Speed and Perplexity	Task Utility
Random Initialization	Reuses shared token weights; initializes target-specific subwords using default random Gaussian weights.	None.	Poor; requires long LAPT cycles to reduce high initial perplexity.	Unstable; high risk of catastrophic forgetting of source task knowledge.
Multivariate Gaussian	Samples target embeddings from a multivariate Gaussian distribution mapping the mean and covariance of the source space.	Source embedding parameters.	Moderate; provides a bounded initial state.	Average; lacks semantic precision across distinct scripts.
WECHSEL	Uses static word embeddings (e.g., FastText) and bilingual dictionaries to learn a linear coordinate transformation.	Large monolingual corpora and aligned bilingual dictionaries.	High; aligns different scripts structurally, speeding up LAPT convergence.	Strong; highly effective for languages typologically distant from the source.
Focus	Initializes target token embeddings using a weighted sum of overlapping, shared tokens.	Shared vocabulary overlaps.	High for related scripts. Poor when transferring to different scripts due to low token overlap.	Poor when transferring to different scripts due to low token overlap.
ZeTT	Employs a zero-shot hypernetwork	Pre-trained hypernetwork.	Good zero-shot perplexity,	Strong; bypasses initial

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

	to predict the embedding parameters for target tokenizers.		but requires high computational cost to train the hypernetwork .	training phases.
OFA	Factorizes source embeddings into a coordinate space and applies multilingual word embeddings to transform coordinates.	Multilingual coordinate systems.	High; systematic geometric projection.	Good; robust mathematical alignment.
TokAlign++	Aligns token IDs based on a global token-token co-occurrence matrix calculated from target text.	Unsupervised target-language co-occurrence statistics.	Superior; reduces perplexity to 76.3 on a 15-language corpus within 1k steps.	Excellent; improves token-level distillation by 4.4% over sentence distillation.
VocADT	Inserts vocabulary adapter modules that learn a linear combination of original embeddings while keeping LLM weights fixed.	None; operates directly on existing embeddings.	High; resolves fragmentation in Latin and highly fragmented scripts.	Excellent; maintains generation quality in machine translation.
Constrained Word2Vec (CW2V)	Constrains word vector training to stay within the convex hull of	Small target text; does not require cross-lingual embeddings.	Strong; matches or exceeds advanced mapping	Highly stable; ensures new embeddings align with

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

	the pre-trained embedding space.		techniques with simple mechanics.	pre-trained parameters.
--	---	--	--	--------------------------------

Tokenadapt Performs tokenizer transplantation; initializes new embeddings via hybrid compositional and top-k semantic heuristics. Target corpus and original tokenizer vocabulary. Excellent; completely replaces tokenizers to improve sequence compression. Strong; minimizes retraining requirements through hybrid semantic initialization.

Dictionary-Based Progressive Removal Maps target subwords to source subwords using bilingual dictionaries; fallbacks iteratively to shorter subwords. Aligned bilingual dictionaries (from descriptive linguistics). Superior; highly effective for low-resource languages with minimal text data. High; outperforms multilingual baselines on Uyghur, Khmer, and Manchu.

5. Parameter-Efficient Fine-Tuning for Machine Translation

When adapting large pre-trained language models to low-resource downstream tasks, fully fine-tuning all parameters is often inefficient and unstable. Parameter-Efficient Fine-Tuning (PEFT) addresses this by freezing the pre-trained model parameters and updating only a small, strategically selected subset of weights (Artetxe et al., 2023).

5.1 PEFT Structural Taxonomy

Following Pfeiffer et al. (2023), PEFT methodologies are categorized based on their structural composition with the base model parameters:

Parameter Composition: Modifies existing weights through sparse or low-rank updates. This includes selective methods (such as BitFit, which updates only the bias terms of the network) and reparameterization methods (such as LoRA, which decomposes weight updates into low-rank matrices) (Manodyna, 2026).

Input Composition: Extends the input sequence with learned embeddings that provide task-specific context. Examples include soft prompts (Prompt Tuning), prefix tuning, and LLaMA-Adapter, which scale hidden activations or prepend learned prompt tokens to key-value states (Yamaguchi et al., 2024).

Function Composition: Augments the frozen transformer backbone with lightweight, trainable modules. Bottleneck adapters (e.g., Houlsby, Pfeiffer, or Parallel adapters) insert small feed-forward layers directly into the transformer blocks. Advanced function composition includes the Compacter architecture, which replaces linear projections with parameterized hypercomplex multiplication layers, and Mixture-of-Experts (MoE) PEFT architectures, which distribute lightweight LoRA blocks across specialized routing experts (Jiang, 2026).

5.2 Empirical Evaluations in Low-Resource Machine Translation

Systematic empirical studies have evaluated 15 PEFT architectures across 8 methods to determine their effectiveness for low-resource Neural Machine Translation (NMT). These experiments utilized datasets from the No Language Left Behind (NLLB) corpus and the FLORES benchmarks, focusing on languages like Sinhala, Tamil, Hindi, and Gujarati (Joshi et al., 2020).

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Among several pre-trained sequence-to-sequence backbones (including mBART-50 and M2M-100), mBART-50 was selected as the optimal base model for downstream adapter integration. The evaluation showed that six PEFT architectures consistently outperformed the full fine-tuning baseline (Gottam, 2026).

The Houlsby+Inversion adapter emerged as the top-performing configuration. This architecture combines a standard Houlsby bottleneck adapter (inserted after both the self-attention and feed-forward layers) with an invertible adapter integrated directly into the language model's embedding layer (Petti, 2026). In-domain NMT experiments on Sinhala-Tamil datasets showed that the Houlsby adapter achieved a SacreBLEU score of 33.34, significantly improving over the baseline of 30.25. In out-of-domain tests on the FLORES benchmark, this adapter achieved a SacreBLEU score of 7.62 compared to the baseline of 5.52 (Mahmudi, 2026).

This demonstrates that parameter-efficient architectures possess strong generalization capabilities, helping models resist overfitting in data-scarce scenarios. Analysis of the bottleneck dimension indicates that a reduction factor of 2 generally provides the optimal trade-off between parameter count and translation accuracy (Zhu et al., 2025).

In parallel developments, Kron-LoRA introduces a hybrid adaptation approach that combines Kronecker-structured factorization with low-rank LoRA compression. This integration yields up to $4\times$ fewer parameters than standard LoRA while retaining similar expressivity. Tested on models like LLaMA-2-7B and LLaMA-3-8B, Kron-LoRA matches LoRA benchmarks with modest memory savings, presenting a scalable option for multi-task adaptation in resource-constrained environments (Kendall & Gal, 2017).

6. Multi-Stage Alignment and Curriculum Learning Strategies

As transfer learning techniques mature, researchers are moving beyond single-stage fine-tuning toward structured, multi-stage alignment and curriculum learning strategies. These strategies introduce learned representations gradually, stabilizing model convergence across typologically diverse settings (Artetxe et al., 2023).

6.1 Progressive Code-Switching (PCS)

Code-switching is a data augmentation technique that mixes target-language translations of words into source-language texts using bilingual dictionaries. This aligns cross-lingual contextual word representations during training. However, uncontrolled replacement can generate ungrammatical sentences that act as noisy data and degrade cross-lingual transferability (Li et al., 2024).

Progressive Code-Switching (PCS) models this adaptation as a curriculum learning process, generating code-switched training examples of increasing difficulty (Bengio et al., 2009):

Difficulty Measurer: Uses explanation learning to assign a relevance score to each word

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

in a sentence, measuring its importance to the model's prediction.

Code-Switcher: Replaces words based on their relevance using an adjustable temperature parameter (T). At low temperatures, only low-relevance words are replaced (generating easy examples). As the temperature scales, high-relevance structural components are replaced, increasing the difficulty (Winata et al., 2023).

Dynamic Scheduler: Determines when to sample more difficult code-switched data based on the model's training convergence, helping the model transition smoothly from easy to complex structures (Qin et al., 2025). This structured curriculum helps align representations under extreme distribution shifts, achieving state-of-the-art results across text classification and sequence tagging tasks in multiple languages (Padó & Lapata, 2009).

6.2 The MERLIN Framework

For cross-lingual reasoning tasks, researchers developed the MERLIN framework (Multi-Stage Curriculum Alignment for Multilingual Encoder-LLM Integration). Existing models like LangBridge align multilingual encoders with LLMs to enable cross-lingual transfer, but they often struggle with complex reasoning in low-resource languages (Pakray et al., 2025).

MERLIN employs a model-stacking architecture that refines alignment in two distinct curriculum stages:

Stage 1 (General Alignment): Aligns representations using general bilingual bitext parallel corpora.

Stage 2 (Task-Specific Alignment): Fine-tunes the model on task-specific reasoning datasets, adapting only a small set of weight-decomposed DoRA parameters (Zhu et al., 2025).

Evaluated on the AfriMGSM benchmark, MERLIN improved exact-match accuracy by +12.9 percentage points over MindMerger and outperformed GPT-4o-mini by 15.2 percentage points. This demonstrates the effectiveness of structured curriculum alignment across both low- and high-resource settings (Dossou & Aïdasso, 2025).

6.3 AlignFreeze Realignment

Realignment techniques are often used to enhance cross-lingual transfer, but they can degrade performance in target languages that differ significantly from the source language. Probing the layers of multilingual models shows that while realignment affects all layers, its negative impact is most severe in the lower layers of the transformer stack (Litschko et al., 2021).

The AlignFreeze method addresses this by freezing the lower half of the model's layers during realignment, protecting foundational representations from degrading. In controlled experiments across 35 languages, freezing the lower layers of XLM-R during realignment improved Part-of-Speech (PoS) tagging accuracy in seven more languages compared to full realignment (Tjuka et al., 2025).

6.4 Multilingual Prompt Translators (MPT)

In few-shot settings, soft prompt learning has emerged as an efficient transfer

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

mechanism. Rather than using static English templates, the Multilingual Prompt Translator (MPT) framework translates soft prompts learned in a source language into target-language prompts (Vykopal et al., 2024). MPT uses a multi-task training strategy on a 15-language parallel corpus to optimize the prompt translator, minimizing the Kullback-Leibler Divergence (KLD) loss between predicted word probabilities in the source and target spaces.

In few-shot settings on the XNLI benchmark, MPT achieved absolute performance gains of 7.5% in 4-shot and 3.1% in 64-shot settings. Crucially, when transferring to typologically distant target languages, MPT achieved an 18.4% relative performance improvement over standard prompting (Qiu et al., 2024).

6.5 SAE Layer-Wise Activation Studies

Layer-wise activation studies using Sparse Autoencoders (SAEs) on models like Gemma-2-2B show that low-resource languages exhibit significantly lower activation values in the model's early layers compared to English. This activation disparity correlates with lower downstream task performance (Li et al., 2023).

Applying activation-aware fine-tuning via LoRA aligns these activation distributions, yielding substantial activation gains (+87.69% for Malayalam and +86.32% for Hindi) while maintaining 91% of the model's English task performance. This demonstrates that activation alignment in early transformer layers is key to improving multilingual representations (Ebing et al., 2026).

7. Evaluation Ecosystem Gaps and Culturally Grounded Benchmarks

Measuring progress in low-resource NLP requires robust, representative evaluation standards. Historically, evaluation relied on translating English datasets into target languages (as seen in early versions of XNLI, XQuAD, and TyDiQA) (Li et al., 2025). However, a comprehensive analysis of 51 publicly released multilingual benchmarks highlights several limitations of this translated-evaluation paradigm:

Table 3: Summary of Culturally Grounded Evaluation Benchmarks and Datasets

Evaluation Benchmark / Dataset	Language & Regional Coverage	Primary Task Modalities Covered	Structural Scale & Data Characteristics	Culturally Grounded Evaluation Metrics
XTREME-UP	88 underrepresented languages.	OCR, autocomplete, translation, transliteration, QA, ASR.	Standardized few-example scarce-data evaluations.	Simple language averages; avoids zero-shot assumptions.
MasakhaNER	10 to 20 African languages.	Named Entity Recognition (NER).	High-quality manually annotated regional entity text.	<i>Not specified</i>

The Coverage Gap: While multilingual benchmarks appear broad, their

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

coverage is often thin. Approximately 36% of evaluated languages appear in only a single benchmark, and entire geographic regions (e.g., Oceania, Central Asia, and the Americas) have near-zero representation (Foroutan et al., 2025). The Task Equity Gap: High-resource languages are evaluated across an average of 14 task categories (including complex reasoning, mathematical problem-solving, and safety alignment). In contrast, low-resource languages are evaluated on only 1 to 3 basic classification categories (typically sentiment analysis or Named Entity Recognition) (Janeiro et al., 2025). Cultural and Pragmatic Divergence:

Translated benchmarks often carry English-centric assumptions, failing to account for the unique cultural knowledge, pragmatic norms, and social contexts of native speakers (Gottam, 2026). Benchmark Contamination: Standard evaluation pipelines are vulnerable to translated benchmark leakage, where parallel training corpora overlap with evaluation sets, leading to inflated performance metrics (Yamaguchi et al., 2025). To address these limitations, researchers have developed language-specific, culturally grounded benchmarks, alongside community-led speech and text corpora. These benchmarks evaluate models on tasks grounded in the linguistic realities of native speakers (Sheng, 2026).

References

- Blasi, D., Anastasopoulos, A., & Neubig, G. (2021). Systematic inequalities in language technology performance. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.
- Chitale, P. A., Gumma, V., Ahuja, S., Kodali, P., Uppadhyay, M., Sudharsan, D., & Sitaram, S. (2025). The role of synthetic data in Multilingual, Multi-cultural AI systems: Lessons from Indic Languages. *arXiv*. <https://doi.org/10.48550/arxiv.2509.21294>
- Gottam, M. (2026). Natural Language Processing in the Era of Large Language Models: Foundations, Integration, and Low-Resource Frontiers. *Preprints.org*.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. <https://doi.org/10.18653/v1/2020.acl-main.560> Cited by: 1578
- Petti, U. (2026). COACT – A Community-centered, Participatory and Actionable Roadmap for Equitable Language AI. *University of Cambridge*.
- Dossou, B. F. P., & Aïdasso, H. (2025). Towards open-ended discovery for low-resource NLP. *arXiv*. <https://doi.org/10.48550/arxiv.2510.01220>
- Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 5574–5584.
- Mahmudi, A. (2026). CommonMorph: Participatory morphological documentation platform. *arXiv*.
- Pakray, A., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, 31, 183–197. <https://doi.org/10.1017/nlp.2024.33>

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Artetxe, M., Goswami, V., Bhosale, S., Fan, A., & Zettlemoyer, L. (2023). Revisiting machine translation for cross-lingual classification. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6489–6499. <https://doi.org/10.18653/v1/2023.emnlp-main.399>
- Ebing, B., & Glavaš, G. (2025). The devil is in the word alignment details: On translation-based cross-lingual transfer for token classification tasks. *arXiv*. <https://doi.org/10.48550/arxiv.2505.10507>
- Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36, 307–340. <https://doi.org/10.1613/jair.2863>
- Zhu, G. et al. (2025). Large margin representation learning for robust cross-lingual named entity recognition. *Findings of the Association for Computational Linguistics: ACL 2025*. <https://aclanthology.org/2025.acl-long.215.pdf>
- Litschko, R., Vulić, I., Ponzetto, S. P., & Glavaš, G. (2021). On cross-lingual retrieval with multilingual text encoders. *arXiv*. <https://doi.org/10.48550/arxiv.2112.11031>
- Liu, Y., Ye, H., Weissweiler, L., Pei, R., & Schuetze, H. (2023). Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8376–8401. <https://doi.org/10.18653/v1/2023.findings-emnlp.562>
- Tjuka, A., Forkel, R., Rzymiski, C., & List, J.-M. (2025). Advancing the database of cross-linguistic colexifications with new workflows and data. *arXiv*. <https://doi.org/10.48550/arxiv.2503.11377>
- Blevins, T., Limisiewicz, T., Gururangan, S., Li, Margaret, Gonen, H., Smith, N. A., & Zettlemoyer, L. (2024). Breaking the curse of multilinguality with cross-lingual expert language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10822–10837. <https://doi.org/10.18653/v1/2024.emnlp-main.604>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Edunov, E., Stoyanov, V., & Zettlemoyer, L. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
- Foroutan, N., Teiletche, P., Tarun, A. M., & Bosselut, A. (2025). Revisiting multilingual data mixtures in language model pretraining. *arXiv*. <https://doi.org/10.48550/arxiv.2510.25947>
- Li, C., Deng, Y., Zhang, J., & Zong, C. (2025). Group then scale: Dynamic mixture-of-experts multilingual language model. *arXiv*. <https://doi.org/10.48550/arxiv.2506.12388>
- Pfeiffer, J., Goyal, N., Lin, X. V., Li, X., Cross, J., Riedel, S., & Artetxe, M. (2022). Lifting the curse of multilinguality by pre-training modular transformers. *arXiv*. <https://doi.org/10.48550/arxiv.2205.06266>
- Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., & Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, 6(101118), 101118. <https://doi.org/10.1016/j.patter.2024.101118>

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Alastruey, B., Janeiro, J. M., Allauzen, A., Elbayad, M., Barrault, L., & Costa-jussà, M. R. (2025). Interference Matrix: Quantifying cross-lingual interference in transformer encoders. *arXiv*. <https://doi.org/10.48550/arxiv.2508.02256>
- Ebing, B., Philipp, F., de Lhoneux, M., & Ostling, R. (2026). One script instead of hundreds? On pretraining romanized encoder language models. *arXiv*. <https://arxiv.org/abs/2601.05776>
- Janeiro, J. M., Alastruey, B., Elbayad, M., Barrault, L., Allauzen, A., & Costa-jussà, M. R. (2025). Mixture of Languages: Improved multilingual encoders through language grouping. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 29706–29721. <https://aclanthology.org/2025.emnlp-main.1509.pdf>
- Jiang, F. (2026). Tokenizer-aware cross-lingual adaptation of decoder-only LLMs through embedding relearning and swapping. *arXiv*. <https://aclanthology.org/2026.eacl-long.357.pdf>
- Manodina, K. H. (2026). When multilingual evaluation assumptions fail: Tokenization effects across scripts. *arXiv*. <https://aclanthology.org/2026.Highreslm-1.4.pdf>
- Sakajo, H. (2025). Cross-lingual vocabulary transfer for low-resource languages using bilingual dictionaries. *arXiv*. <https://arxiv.org/pdf/2506.01535>
- Sheng, C. Y. (2026). Beyond the tokenization bottleneck: A conceptual framework for efficiency-plasticity trade-offs in ASEAN large language model adaptation. *Cureus Journals*. <https://www.cureusjournals.com/articles/13958-beyond-the-tokenization-bottleneck-a-conceptual-framework-for-efficiency-plasticity-trade-offs-in-asean-large-language-model-adaptation.pdf>
- Yamaguchi, A., Villavicencio, A., & Aletras, N. (2024). An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. *arXiv*. <https://doi.org/10.48550/arxiv.2402.10712>
- Yamaguchi, A., Morishita, T., Villavicencio, A., & Aletras, N. (2025). Adapting chat language models using only target unlabeled language data. *arXiv*. <https://doi.org/10.48550/arxiv.2412.11704>
- Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48. <https://doi.org/10.1145/1553374.1553380>
- Li, Z., Hu, C., Chen, J., Chen, Z., Guo, X., & Zhang, R. (2024). Improving zero-shot cross-lingual transfer via progressive code-switching. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 6388–6396. <https://doi.org/10.24963/ijcai.2024/706>
- Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., & Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, 6(3), 101118. <https://doi.org/10.1016/j.patter.2024.101118>
- Winata, G., Aji, A. F., Yong, Z. X., & Solorio, T. (2023). The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. *Findings of the Association for Computational Linguistics: ACL 2023*, 2936–2978. <https://doi.org/10.18653/v1/2023.findings-acl.185>
- Li, S., Hu, X., Liu, A., Yang, Y., Ma, F., Yu, P. S., & Wen, L. (2023). Enhancing cross-lingual natural language inference by soft prompting with multilingual

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- verbalizer. *Findings of the Association for Computational Linguistics: ACL 2023*, 1361–1374. <https://doi.org/10.18653/v1/2023.findings-acl.88>
- Qiu, X., Wang, Y., Shi, J., Zhou, W., & Li, H. (2024). Cross-lingual transfer for natural language inference via multilingual prompt translator. *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. <https://doi.org/10.1109/icme57554.2024.10687356>
- Vykopal, I., Ostermann, S., & Šimko, M. (2024). Soft language prompts for language transfer. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. <https://doi.org/10.18653/v1/2025.naacl-long.517>