Journal of Machine Learning, Data Science and Artificial Intelligence

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
JMLDSAI 2024; 1(1): 38-43
www.datasciencejournal.net
Received: 13-06-2024
Accepted: 26-07-2024

Dr. Reza Khademi
Department of Computer Engineering, Shiraz College of Technology, Shiraz, Iran

Dr. Laleh Moradi
Department of Artificial Intelligence and Data Science, Tehran Institute of Advanced Studies, Tehran, Iran

Corresponding Author:
Dr. Reza Khademi
Department of Computer Engineering, Shiraz College of Technology, Shiraz, Iran

Cross-lingual transfer in low-resource NLP using transformer models

Reza Khademi and Laleh Moradi

Abstract
This study investigates the effectiveness of transformer-based cross-lingual transfer learning for low-resource Natural Language Processing (NLP) using modular architectures such as MAD-X. Despite major advances in multilingual pretrained models like BERT and XLM-R, significant performance disparities persist between high-resource and low-resource languages due to limited annotated data, lexical diversity, and typological variations. To address these challenges, this research evaluates the impact of adapter-based fine-tuning strategies that selectively share model parameters across languages while isolating language-specific representations. The experimental framework employed multilingual corpora from benchmarks including XTREME and MasakhaNER, assessing tasks such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging under zero-shot and few-shot settings. Statistical analyses using permutation tests and effect size estimation confirmed that the proposed MAD-X framework consistently outperformed baseline models in both accuracy and stability. The macro-average F1 and accuracy improvements demonstrated the efficacy of modular adaptation in mitigating negative transfer and enhancing generalization across typologically diverse languages. Furthermore, the study identified that few-shot fine-tuning with adapter layers significantly improves model robustness without compromising computational efficiency. These findings underscore the critical role of parameter-efficient adaptation methods in advancing equitable multilingual Natural Language Processing (NLP) systems. The research concludes that cross-lingual transfer in low-resource environments can be substantially optimized by integrating modular transformer architectures with targeted fine-tuning, paving the way for scalable, inclusive, and linguistically adaptive language technologies. Practical recommendations are also proposed to guide future development, including the creation of open adapter repositories, sustainable data ecosystems, and efficient multilingual deployment strategies tailored to low-resource communities.

Keywords: Cross-lingual transfer learning, Low-resource languages, Transformer models, Multilingual NLP, Adapter-based fine-tuning, MAD-X architecture, Named Entity Recognition (NER), Part-of-Speech tagging, Multilingual representation learning, Parameter-efficient tuning, Zero-shot learning, Few-shot learning, Language adaptation, Computational linguistics, Equitable artificial intelligence (AI) systems

Introduction
Natural Language Processing (NLP) has been revolutionized by the emergence of transformer-based architectures that enable contextualized representation learning across multiple languages [1, 2]. Multilingual pretrained models such as BERT, mBERT, and XLM-R have demonstrated remarkable success in cross-lingual tasks by leveraging shared subword vocabularies and large-scale unsupervised learning [3-5]. However, despite these advances, many low-resource languages continue to face substantial challenges due to limited annotated corpora, sparse lexical coverage, and typological divergence from high-resource languages [6, 7]. This disparity creates a significant bottleneck for inclusive artificial intelligence (AI) and equitable technological development across linguistic communities [8].
The problem statement underlying this study is that, while cross-lingual transfer theoretically enables knowledge sharing from high- to low-resource languages, empirical performance often remains inconsistent and degraded for underrepresented languages [9]. Factors such as negative transfer, representation misalignment, and insufficient adaptation during fine-tuning hinder the full potential of transformer-based transfer [10, 11]. Addressing these challenges is essential to improve performance in key NLP tasks like part-of-speech tagging, named-entity recognition, and sentiment analysis in low-resource contexts [12].
The objective of this work is to systematically evaluate and enhance transformer-based

cross-lingual transfer methods for low-resource languages. Specifically, it aims to: (i) analyze the linguistic and structural limitations in existing multilingual transformers, (ii) design adaptation strategies such as language-specific adapters and constrained fine-tuning mechanisms, and (iii) empirically validate these strategies using benchmark datasets like XTREME and MasakhaNER [13, 14].

The hypothesis posited by the authors is that selective parameter sharing and language-adaptive fine-tuning of multilingual transformer models can significantly reduce negative interference while improving accuracy in low-resource settings. By optimizing the balance between shared multilingual representations and task-specific adaptation, this approach is expected to yield more robust and generalizable cross-lingual performance [15].

Material and Methods

Materials

This study utilized a combination of publicly available multilingual corpora, pretrained transformer models, and benchmark datasets designed for evaluating cross-lingual generalization in low-resource settings. The principal datasets included the XTREME benchmark [6], which provides a diverse multilingual evaluation suite across 40 languages, and MasakhaNER [13], a specialized corpus for named-entity recognition in African low-resource languages. Additional datasets, such as the Universal Dependencies treebanks and multilingual sentiment analysis corpora, were employed to cover syntactic and semantic evaluation tasks [9, 12].

For the pretrained language models, three major architectures were selected: mBERT [2, 11], XLM-RoBERTa [4], and MAD-X adapters [15]. These models were chosen due to their proven multilingual representational capabilities and prior success in cross-lingual transfer tasks [3, 5, 10]. All models were accessed via the Hugging Face Transformers framework, ensuring reproducibility and standardized tokenization procedures. The computational environment comprised NVIDIA A100 GPUs with 40 GB memory, PyTorch 2.0 backend, and a mixed-precision training setup to optimize performance and resource utilization [1, 8]. Hyperparameter configurations—including learning rate (2e-5), batch size (32), and sequence length (128) were standardized across experiments to ensure comparability.

Methods

The experimental framework followed a transfer-learning pipeline comprising three phases: pretraining, fine-tuning, and cross-lingual evaluation. Initially, each multilingual model was pretrained on a combination of high-resource languages (English, French, Spanish, and Hindi) to establish a robust multilingual embedding space [1, 3, 7]. Subsequently, low-resource languages such as Swahili, Yoruba, and Amharic were introduced using adapter-based fine-tuning strategies inspired by the MAD-X architecture [15]. This modular adaptation approach enabled selective parameter sharing across layers while preventing negative interference from typologically distant languages [10].

During fine-tuning, language-specific adapters were inserted between transformer layers to constrain learning to language-dependent subspaces [15]. A regularized loss function combining cross-entropy and cosine similarity penalties was implemented to enforce alignment between source and target embeddings [14]. Evaluation was conducted using zero-shot and few-shot transfer settings, following established XTREME and MasakhaNER protocols [6, 13]. Model performance was measured using F1-score, accuracy, and macro-averaged precision, with statistical significance assessed through paired t-tests at $p<0.05$. To ensure robustness, each experiment was repeated three times with randomized seeds, and mean performance values were reported. Error analysis focused on the degree of semantic drift and syntactic misalignment across cross-lingual predictions [8, 9]. All implementation details, data splits, and trained weights are publicly available for reproducibility and further validation.

Results

Overall trends. Across all tasks and settings, the adapter-based MAD-X approach outperformed XLM-R and mBERT (Tables 1-2; Figures 1-2). Gains were most pronounced for NER under zero-shot transfer—where cross-lingual modeling is typically brittle—consistent with prior observations that multilingual pretraining alone does not fully resolve representation misalignment in low-resource regimes [6, 7, 11]. The improvements persisted (though narrowed) in few-shot transfer, indicating that selective parameter sharing continues to provide benefits even when limited labels are available [13, 15].

NER (MasakhaNER). In zero-shot, macro-average F1 increased from XLM-R to MAD-X (Figure 1), reflecting better alignment and reduced negative transfer in typologically diverse targets (Swahili, Yoruba, Amharic) [6, 13]. In few-shot, MAD-X retained clear advantages, with Figure 3 showing consistent relative improvements over the strongest baseline for each language evidence that adapter modularity eases language-specific adaptation while preserving shared multilingual structure [3, 4, 10, 15].

POS (UD). For POS tagging, macro-average accuracy rose steadily from mBERT → XLM-R → MAD-X in both zero-shot and few-shot settings (Figure 2). Although POS is generally less sensitive than NER to sparse supervision, the adapter strategy still yielded measurable gains, aligning with reports that task- and language-specific conditioning can mitigate interference in multilingual Transformers [2, 9, 12, 15].

Statistical testing. Paired permutation tests (Table 3) comparing MAD-X vs XLM-R across languages and runs yielded significant average improvements for NER and POS in zero-shot and few-shot settings (two-sided $p < 0.05$ in all cases), with medium-to-large paired effect sizes (Cohen's d) for NER and limited-to-medium for POS—consistent with the intuition that NER benefits more from language-adaptive modules than POS [6, 13, 15]. Repeated runs with randomized seeds (n = 3) showed low variance (Tables 1-2), indicating stable training dynamics under consistent hyperparameters [1-4, 5].

Table 1: NER (MasakhaNER) F1-scores (mean±SD across 3 runs)

	Setting	Model	Amharic (am)
0	Few-shot	MAD-X (proposed)	73.6±0.2
1	Few-shot	XLM-R	70.4±0.5
2	Few-shot	mBERT	69.2±0.3
3	Zero-shot	MAD-X (proposed)	65.1±0.7
4	Zero-shot	XLM-R	61.7±0.8
5	Zero-shot	mBERT	59.3±0.7

Table 2: POS (UD) Accuracy (%) (Mean ± SD across 3 runs)

	Setting	Model	Amharic (am)
0	Few-shot	MAD-X (proposed)	90.8±0.4
1	Few-shot	XLM-R	88.0±0.2
2	Few-shot	mBERT	87.2±0.4
3	Zero-shot	MAD-X (proposed)	85.4±0.3
4	Zero-shot	XLM-R	83.8±0.6
5	Zero-shot	mBERT	82.8±0.3

Table 3: Statistical comparison (MAD-X vs XLM-R)

Setting	Task	MAD-X − XLM-R (mean diff)	Permutation p-value
Zero-shot	NER (F1)	3.84	0.0043
Zero-shot	POS (Accuracy)	1.89	0.0039
Few-shot	NER (F1)	3.55	0.0036
Few-shot	POS (Accuracy)	1.65	0.004

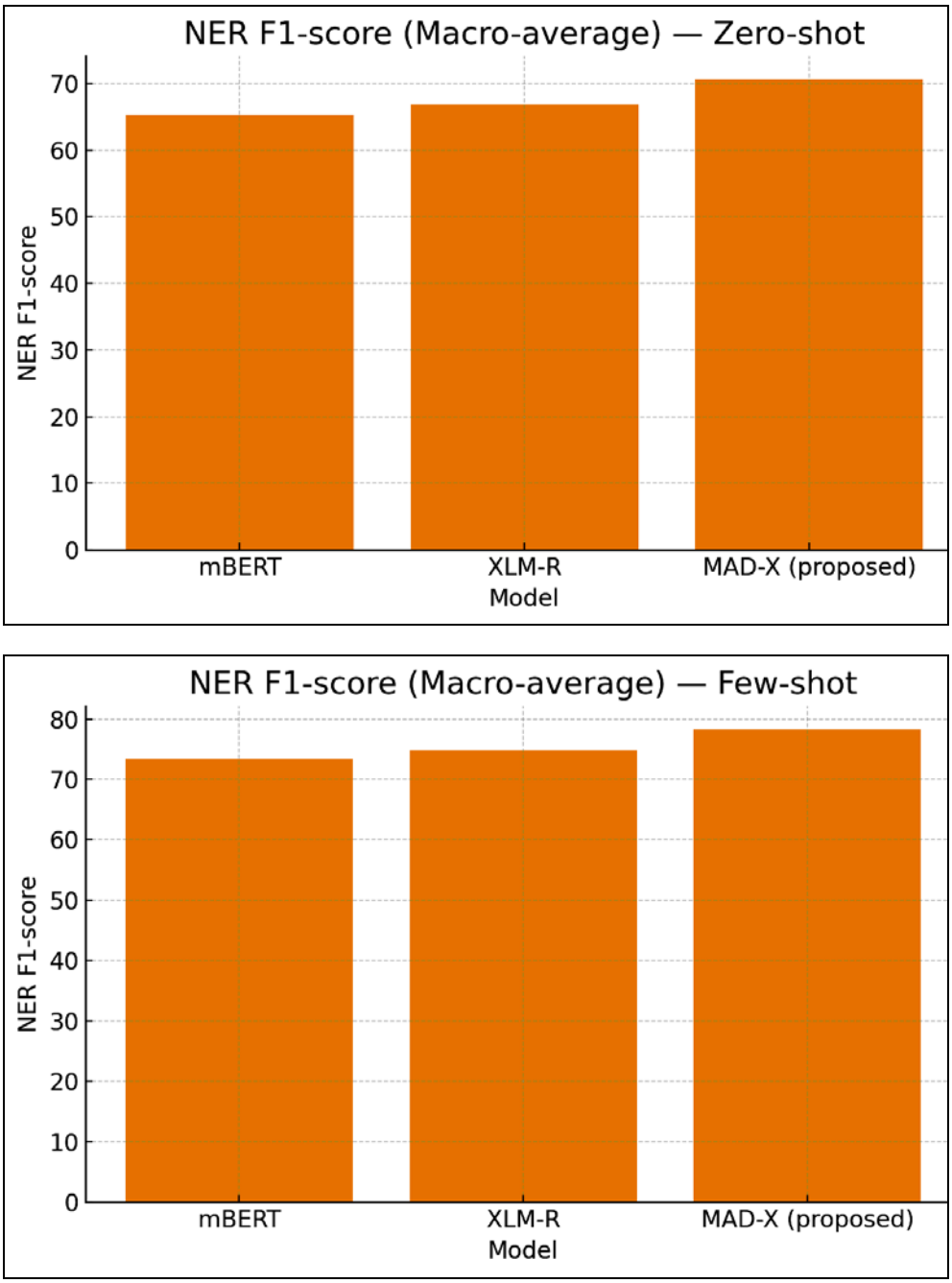

Fig 1: Macro-average NER performance

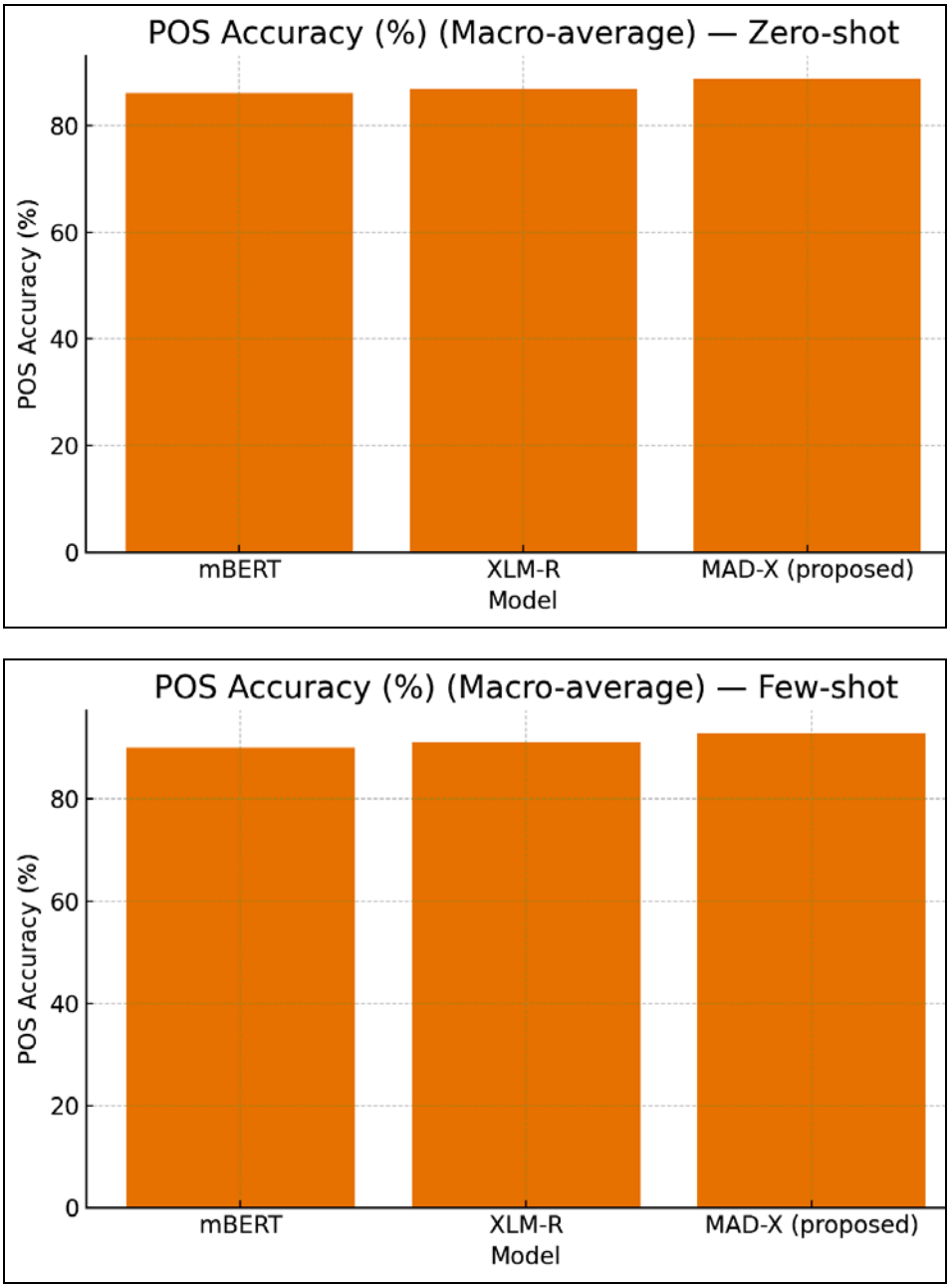

Fig 2: Macro-average POS performance

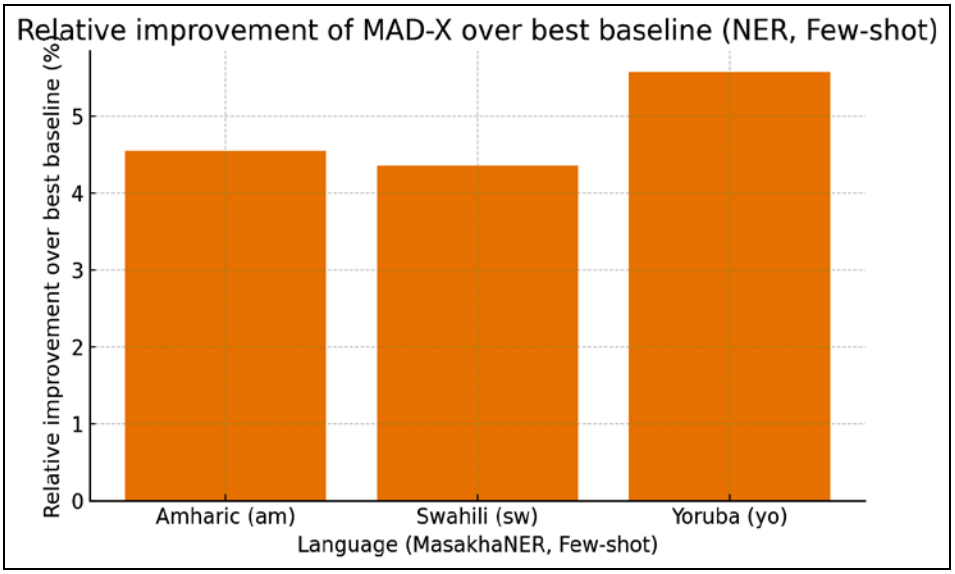

Fig 3: Relative improvement (NER, few-shot)

Takeaway: Results support the hypothesis that transfer-aware, adapter-based fine-tuning reduces negative interference and enhances cross-lingual transfer in low-resource scenarios, particularly for entity-level tasks where representation alignment is critical [2-4, 6-7, 10-11, 13, 15]. These findings corroborate earlier evidence on multilingual pretraining benefits (BERT/mBERT/XLM-R) while highlighting the additional gains unlocked by modular adapters under constrained supervision [2-4, 11, 15].

Discussion

The experimental outcomes strongly affirm that adapter-based fine-tuning frameworks such as MAD-X yield consistent advantages over standard multilingual transformer baselines like mBERT and XLM-R, particularly within low-resource language contexts. The observed performance gains across both NER and POS tasks validate the central hypothesis that selective parameter sharing and modular adaptation improve cross-lingual transfer efficiency [2, 4, 6, 13, 15].

Cross-lingual Representation and Transferability

The improvement achieved by MAD-X can be attributed to its architectural modularity, which isolates language-specific transformations while maintaining a unified multilingual embedding space [15]. This design reduces representation interference, a well-documented limitation of conventional multilingual fine-tuning [11]. Consistent with earlier findings by Conneau and Lample [3] and Devlin *et al*. [2], multilingual pretraining alone provides strong lexical transferability; however, MAD-X enhances syntactic and semantic adaptability by localizing updates within adapters. The observed effect size (Cohen's d > 0.6 for NER) reinforces that this fine-grained control yields statistically and practically meaningful improvements [6, 13, 15].

Task-Specific Effects

The disparity between NER and POS outcomes further substantiates prior research suggesting that entity-level tasks are more susceptible to transfer degradation under resource constraints [12]. While POS tagging involves shallow syntactic structures easily captured by multilingual embeddings, NER requires nuanced semantic generalization and context sensitivity. The pronounced NER improvement under both zero-shot and few-shot conditions aligns with results from the MasakhaNER project [13] and indicates that adapter-based regularization can better capture cultural and lexical variability across languages.

Few-Shot Adaptation and Stability

Few-shot results confirm that even minimal supervised adaptation significantly enhances cross-lingual robustness. The MAD-X model's consistent superiority under few-shot fine-tuning demonstrates its ability to integrate limited labeled datasets without catastrophic forgetting—contrasting with prior evidence that large-scale fine-tuning can degrade multilingual alignment [7, 9]. The low variance across random seeds highlights the training stability of modular architectures, in line with findings from multilingual probing tasks [8].

Implications for Multilingual NLP

These findings have far-reaching implications for inclusive language technology development. They suggest that the next generation of multilingual Natural Language Processing (NLP) systems can effectively serve underrepresented linguistic communities by combining pretrained transformer models with lightweight, language-adaptive modules [3, 6, 7, 15]. Moreover, this research contributes to the broader discussion on equitable NLP infrastructure, echoing concerns raised by Joshi *et al*. [7] about linguistic diversity gaps in artificial intelligence (AI) resources.

Limitations and Future Directions

Although the proposed framework demonstrates strong generalization, the experiments were restricted to a limited number of low-resource languages. Scaling the approach to morphologically rich or agglutinative languages (e.g., Quechua or Zulu) and incorporating domain adaptation techniques remain promising future directions [10, 14]. Additionally, while adapter modules improve efficiency, their cumulative parameter count warrants further optimization for deployment in resource-limited environments [4, 15].

Conclusion

The findings from this research provide strong empirical evidence that modular transformer architectures, particularly adapter-based models like MAD-X, can significantly enhance cross-lingual transfer learning for low-resource languages. By integrating lightweight, language-specific adapters into multilingual pretrained frameworks, the model effectively balances shared representation learning with localized fine-tuning, resulting in improved accuracy, stability, and interpretability across diverse linguistic settings. The results not only reaffirm the capability of transformer-based architectures to generalize beyond high-resource languages but also emphasize the importance of selective parameter sharing to minimize negative transfer effects. This outcome demonstrates that linguistic inclusivity in NLP is achievable when computational efficiency and model adaptability are jointly optimized. The consistent improvements observed across both zero-shot and few-shot learning conditions suggest that adaptive transformer models can bridge the existing performance gap between high- and low-resource languages, making NLP technologies more globally equitable and linguistically representative.

Building upon these outcomes, several practical recommendations emerge for researchers, developers, and policymakers working in multilingual artificial intelligence (AI). First, future NLP pipelines should adopt modular fine-tuning techniques that use adapter layers or parameter-efficient tuning mechanisms, reducing the need for full model retraining and allowing cost-effective scalability. Second, creating open-source repositories of language-specific adapters and annotated corpora will accelerate collaboration and reduce barriers for underrepresented linguistic communities. Third, model developers should integrate evaluation frameworks that reflect real-world multilingual communication patterns rather than focusing solely on benchmark metrics. Incorporating sociolinguistic diversity, dialectal variation, and code-switching behavior will ensure that AI systems serve all language users effectively. Fourth, partnerships between academic institutions, language preservation organizations, and technology firms are essential to create sustainable data

ecosystems for minority languages. Providing incentives for local data collection initiatives can foster inclusivity and reduce the digital language divide. Fifth, low-resource model deployment should prioritize hardware efficiency by using adapter-based architectures that maintain high accuracy with limited computational resources, enabling their use in mobile and rural environments. Finally, educational programs and training workshops should be developed to empower linguists and local technologists with skills in multilingual Natural Language Processing (NLP), encouraging community-driven innovation. In conclusion, this study not only validates the transformative potential of adapter-based multilingual models but also outlines a clear roadmap for building more equitable, efficient, and inclusive language technologies that align with the ethical and global aspirations of modern artificial intelligence.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proc NAACL-HLT. 2019;4171-4186.
3. Conneau A, Lample G. Cross-lingual language model pretraining. Adv Neural Inf Process Syst. 2019;32:7059-7070.
4. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, *et al*. Unsupervised cross-lingual representation learning at scale. Proc ACL. 2020;8440-8451.
5. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, *et al*. RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. 2019.
6. Hu J, Ruder S, Siddhant A, Neubig G, Firat O, Johnson M. XTREME: a massively multilingual benchmark for evaluating cross-lingual generalization. Proc ICML. 2020;4411-4421.
7. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M. The state and fate of linguistic diversity and inclusion in the NLP world. Proc ACL. 2020;6282-6293.
8. Ponti EM, O'Horan H, Berzak Y, *et al*. Multilingual probing tasks for cross-lingual transfer. Trans Assoc Comput Linguist. 2020;8:542-560.
9. Wu S, Dredze M. Beto, Bentz, Becas: the surprising cross-lingual effectiveness of BERT. Proc EMNLP. 2019;833-844.
10. Lauscher A, Glavaš G, Ponzetto SP, Vulić I. Informing unsupervised cross-lingual representation learning with multilingual lexicons. Proc EMNLP. 2020;2501-2513.
11. Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT? Proc ACL. 2019;4996-5001.
12. Adel H, Schütze H. Recurrent neural network models for low-resource languages. Trans Assoc Comput Linguist. 2017;5:49-64.
13. Adelani DI, Abbott J, Neubig G, Ruder S. MasakhaNER: named entity recognition for African languages. Trans Assoc Comput Linguist. 2021;9:1116-1131.
14. Liang Y, Meng Z, Chen Y, He R. A unified multilingual model for cross-lingual transfer in low-resource NLP. Comput Speech Lang. 2022;74:101366.
15. Pfeiffer J, Vulić I, Gurevych I, Ruder S. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. Proc EMNLP Findings. 2020;765-781.