2025 International Conference on Economics, Management and Education Technology (ICEMET 2025)

Article

Contextual Learning Support for Low-Resource Language Large Language Models: Efficient Training Strategies for Zero-Shot and Few-Shot Learning

Tanya Xiang [1,*]

[1] School of Physics and Engineering, University of Birmingham, Birmingham, United Kingdom
* Correspondence: Tanya Xiang, School of Physics and Engineering, University of Birmingham, Birmingham, United Kingdom

Abstract: In natural language processing tasks for low-resource languages, building high-quality large language models faces numerous challenges, including data scarcity, domain mismatch, and limited annotation resources. This paper proposes an efficient training strategy based on contextual learning, aimed at addressing the difficulties encountered by low-resource languages in zero-shot and few-shot learning scenarios. The approach leverages cross-lingual knowledge transfer from high-resource languages and systematically enhances contextual information in prompts and representations to maximize the model's generalization ability. We explore the combination of multi-task learning and self-supervised learning to exploit heterogeneous corpora, using existing multilingual and monolingual resources for pretraining. A lightweight fine-tuning stage with a small amount of labeled data is then employed for targeted adaptation to specific downstream tasks and languages. The proposed framework is designed to be computationally efficient, reducing training cost while maintaining or improving performance. Experimental results on a range of language understanding and generation benchmarks demonstrate significant improvements in task performance across various low-resource languages under both zero-shot and few-shot conditions. Ablation studies further highlight the contribution of contextual learning components and cross-lingual transfer mechanisms. These findings provide practical guidance for developing scalable large language models for underrepresented languages and offer new ideas and methods for future research on inclusive and resource-efficient language technologies.

Keywords: low-resource languages; large language models; contextual learning; transfer learning; zero-shot learning; few-shot learning

Received: 11 February 2026
Revised: 27 March 2026
Accepted: 08 April 2026
Published: 11 April 2026

1 Introduction

1.1 Research Background

Globalization continues to advance, highlighting the growing importance of linguistic diversity. However, low-resource languages face significant challenges due to the scarcity of data in the natural language processing domain. Many of these languages lack sufficient textual resources or labeled data, hindering the development and application of advanced natural language processing technologies. Traditional machine learning models typically require extensive annotated data for training, which is difficult to obtain in low-resource language environments. In recent years, deep learning technologies have emerged as a promising solution to address this issue [1]. Despite advancements, zero-shot and few-shot learning still encounter substantial obstacles when applied to low-resource languages. To enhance the processing capabilities of these languages, researchers have increasingly focused on contextual learning approaches [2].

By extracting contextual information from text data, these methods aim to improve the generalization ability of models under conditions of limited data availability. This research contributes to the progress of natural language processing for languages with minimal resources and opens new avenues for language preservation and the safeguarding of cultural heritage.

1.2. Research Significance

The study of low-resource language natural language processing holds significant academic and practical value. As linguistic diversity gains prominence, research on low-resource languages supports the preservation and transmission of languages and cultures globally, fostering communication and understanding across different cultural contexts. With the rapid advancements in artificial intelligence and natural language processing technologies, effectively addressing low-resource languages enhances their usability in applications such as information retrieval and machine translation, while improving the universality and adaptability of these technologies. The contextual learning training strategy allows models to enhance their understanding and generation capabilities for low-resource languages, even in the absence of sufficient labeled data, thereby providing greater reliability and accuracy for practical applications. By incorporating zero-shot and few-shot learning strategies, this research extends the boundaries of existing machine learning technologies, enabling their application to a wider range of linguistic scenarios. This not only advances theoretical research within the academic community but also offers practical solutions for the industrial sector, contributing to the improved performance of intelligent systems [1, 3].

2 Current Application Status of Contextual Learning Supporting Efficient Training Strategies for Low-Resource Language Large Language Models

2.1 Principles of Contextual Learning Technology and Characteristics of Low-Resource Language Processing

Contextual learning technology utilizes information within the context to enhance a model's ability to understand and generate language. Its core lies in capturing semantic relationships and contextual information in the input text. Historically, most language models relied on local context, focusing solely on the immediate information of input words or sentences. In contrast, contextual learning emphasizes the use of global context, aiming to understand language structure more comprehensively by analyzing relationships between words in a sequence. This approach is particularly advantageous for processing low-resource languages, which often lack sufficient annotated data [4]. By employing effective context encoding, models can extract latent information from limited data and strengthen their grasp of semantic and syntactic characteristics. This technology can take various forms, such as deep learning architectures like transformer structures or recurrent neural networks, which model long-distance dependencies. These methods enable low-resource languages to achieve improved performance in tasks such as sentiment analysis and question-answering systems. Furthermore, combining contextual learning with self-supervised learning methods can facilitate the generation of corpora tailored to training low-resource language models on a larger scale. This integration enhances the generalization and adaptability of models, thereby advancing the broader application of natural language processing technologies in low-resource languages.

2.2. Analysis of Application Scenarios

2.2.1 Low-Resource Language Corpus Construction and Feature Digitization

To achieve contextual learning, constructing a corpus of low-resource languages serves as the foundation, and digitizing these features is particularly crucial for model training. Researchers must gather text data from various sources, such as social media, books, news reports, and academic papers, to ensure the corpus is diverse and comprehensive. During data collection, it is essential to carefully clean and process the data, eliminate noise, and perform preprocessing tasks such as word segmentation and part-of-speech tagging to ensure data quality. For the collected text data, feature extraction techniques are employed to convert it into a digital format. Commonly used methods include the bag-of-words model, term frequency-inverse document frequency (TF-IDF), and word embedding techniques. By mapping vocabulary into high-dimensional space, these methods capture the semantic relationships and contextual information between words [5]. Incorporating contextual features during digitization can significantly enhance the model's comprehension capabilities. Even with limited samples, contextual reasoning can still be achieved. To enable effective learning, a dedicated dataset for low-resource languages should be constructed and combined with self-supervised learning methods. This approach can improve the model's generalization capabilities and establish a robust foundation for subsequent training and applications. Such efforts are of great importance for advancing research in natural language processing for low-resource languages.

2.2.2. Efficient Prompt Template Generation and Precise Instruction Design

To enhance the performance of large language models for low-resource languages, it is crucial to efficiently generate prompt templates and design precise instructions. An effective prompt template can guide the model to focus on the specific task context, thereby improving its performance in zero-shot and few-shot learning tasks. When creating templates for prompts, the linguistic features of low-resource languages must be thoroughly considered. Using the corpus, researchers can construct feature-rich instructions by analyzing task requirements and language structures. This involves developing templates for tasks such as fill-in-the-blank, multiple choice, and generation requests. Each template should clearly define the dependencies between input and output. In crafting accurate instructions, emphasis should be placed on simplicity and clarity to ensure the model can quickly comprehend and execute tasks. Incorporating contextual information helps maintain semantic consistency in complex tasks. By continuously optimizing the generated prompts and instructions, adjustments can be made for specific tasks to align with the model's learning requirements. This iterative process enhances the model's efficiency in executing particular tasks, offering flexibility and adaptability for learning strategies tailored to low-resource languages. Over time, this approach contributes to improving the model's overall performance, laying a solid foundation for practical applications.

2.2.3. Model Training Performance Optimization and Cross-Language Ability Evaluation

The optimization of model training performance and the evaluation of cross-language capabilities are crucial to the effectiveness of low-resource language large language models. To optimize training performance, adjustments can be made to model hyperparameters, appropriate optimization algorithms can be selected, and regularization techniques can be introduced. These measures help prevent overfitting and enhance the model's generalization ability. Effective training strategies for low-resource languages include leveraging transfer learning, where a model trained on high-resource languages serves as a foundation for fine-tuning, thereby accelerating the learning process using existing knowledge. Refining training with small-scale but high-quality annotated data is another important approach to improving performance. For cross-language capability assessment, multiple tasks should be designed to evaluate the model's ability to transfer knowledge between different languages. Comparing the performance of low-resource languages with high-resource languages allows for an analysis of the model's adaptability and effectiveness in specific language tasks. Relevant evaluation metrics, such as cross-validation and BLEU scores, can be used to quantify the model's performance, providing data support for subsequent optimization efforts [6]. These optimization and evaluation processes contribute to enhancing the practicality of large language models for low-resource languages, offering valuable experience and references for cross-language natural language processing research and establishing a solid foundation for promoting the equitable development of language technology [7].

3 Challenges Facing Contextual Learning Support for Efficient Zero-Shot and Few-Shot Training Strategies in Low-Resource Language Large Language Models

3.1 Technical Challenges

3.1.1 Scarcity of Multimodal Heterogeneous Data and Difficulty in Integrating High-Quality Samples

When conducting research on low-resource languages, a significant technical challenge lies in the scarcity of multimodal heterogeneous data and the difficulty in integrating high-quality samples. Multimodal data encompasses various formats, such as text, speech, and images. The limited availability of such data hinders the model's ability to develop a comprehensive understanding and processing capability of language. Low-resource languages often lack extensive corpora, and the restricted data sources limit the model's capacity to learn language features thoroughly, leading to suboptimal performance in practical applications. Integrating high-quality samples is particularly challenging due to the substantial differences between various data modalities. Synchronizing and processing text, speech, and their corresponding contextual information to ensure effective linkage and consistency between samples is essential for enhancing model performance. To efficiently integrate multimodal data, robust algorithms are necessary. Efforts should focus on data preprocessing, feature extraction, and sample annotation to ensure that the generated samples are both representative and of high quality, thereby supporting model learning and application. During this process, attention must be given to cross-domain knowledge transfer and sample generation techniques to alleviate the impact of data scarcity.

3.1.2. Contextual Inference Generalization and Insufficient Long-Distance Dependency Handling

For large language models with limited resources, achieving effective generalization in contextual reasoning and addressing long-distance dependency issues presents significant technical challenges. Contextual reasoning involves enabling the model to extract and comprehend useful information from the provided text. However, the inherent characteristics of resource-constrained languages often hinder the model's ability to infer complex semantic relationships. The issue of long-distance dependency is even more pronounced, as traditional models typically struggle to capture cross-length correlations between words when processing contextual information. This limitation can lead to errors in sentence comprehension and content generation. In low-resource language environments, the scarcity of rich contextual information exacerbates these challenges. While the model may perform adequately in tasks requiring short-term memory, its reliance on long-term context is insufficient, which often impacts its overall performance.

3.2. Challenges from Low-Resource Language Characteristics

3.2.1 Large Linguistic Form Differences and Complex Syntax

Due to the significant differences in language morphology among low-resource languages, combined with the complexity of their grammatical structures, substantial challenges arise in their study. The linguistic forms of low-resource languages often exhibit considerable diversity, complicating the standardization and unified processing of vocabulary and morphological variations across different dialects and variants. Many of these languages feature extensive inflectional changes and grammatical markers, resulting in highly intricate vocabulary structures and syntactic formations. For instance, grammatical characteristics such as tense, number, and case may be expressed through various prefixes, suffixes, or stem modifications. This imposes higher demands on the training and recognition capabilities of models. Additionally, the complexity of grammatical structures is evident in sentence composition. Some low-resource languages possess grammatical rules that differ entirely from those of mainstream languages, altering word order and syntactic relationships. These unique characteristics make it challenging to effectively apply traditional natural language processing techniques.

3.2.2. Domain Adaptation Effectiveness and Model Performance Quantification Difficulties

The effectiveness of domain adaptation and the quantification of model performance are critical challenges in the practical application of large language models for low-resource languages. These languages often lack sufficient annotated data in specific domains, resulting in limited adaptability of models for cross-domain tasks and significant performance declines in specialized areas. Domain adaptation requires models to effectively transfer existing knowledge to new task scenarios. However, most current models are designed to handle texts with domain-specific characteristics, leading to notable performance losses, particularly in low-resource languages. Furthermore, the quantification of model performance faces challenges due to the absence of standardized evaluation metrics and benchmark datasets. This lack of standardization complicates the comparison of models across different tasks and domains, hindering researchers' ability to evaluate model effectiveness and optimize performance.

3.3. Ethical and Language Safety Dilemmas

3.3.1 Cultural Bias Protection and Language Data Security

When developing and applying large language models for low-resource languages, it is crucial to address the protection of cultural biases and ensure language data security. Low-resource languages often embody the culture and history of specific groups. If cultural biases in training data are not adequately managed, the model may produce inappropriate content that undermines cultural identity or perpetuates harmful stereotypes. To mitigate this, datasets must prioritize diversity and representativeness, avoiding the dominance of any single cultural paradigm. Additionally, language data security is increasingly significant [8, 9]. During data collection and usage, adherence to lifecycle management principles is essential to safeguard participants' privacy and sensitive information, as well as to prevent data leakage and misuse.

3.3.2. Algorithmic Hallucination and Output Consistency Risks

The risk of algorithmic hallucination and inconsistent output results presents a significant challenge for large language models designed for low-resource languages. Algorithmic hallucination occurs when the model lacks sufficient context or data, leading to the generation of information that appears reasonable on the surface but is actually inaccurate or untrue. This issue is particularly pronounced in low-resource languages due to limited data availability and inherent constraints in model training, resulting in content that deviates from actual semantics [10, 11]. Such inaccuracies can undermine users' understanding and trust in the information provided. Additionally, the risk of output inconsistency arises when the same input produces noticeably different content under varying conditions or at different times, leading to contradictions or confusion. This inconsistency directly impacts user experience and diminishes the model's credibility to a certain extent.

4 Optimization Strategies for Contextual Learning Support in Efficient Zero-Shot and Few-Shot Training Strategies for Low-Resource Language Large Language Models

4.1 Technical Optimization Pathways

4.1.1 Building a Unified Pretraining Data Platform and Governance System for Low-Resource Languages

To build a unified low-resource language pretraining data center and governance system, it is essential to enhance the training efficiency and effectiveness of large language models. This center should centrally manage and integrate diverse data from various sources, including text, voice, and multimodal content, ensuring sufficient diversity and richness. By establishing standardized processes for data collection, cleaning, and annotation, data quality can be significantly improved, providing an accurate and reliable foundation for model training supported by an intelligent data governance system [12, 13]. Real-time monitoring and management of data can ensure compliance with ethical requirements and prevent potential data bias or cultural misuse. The governance system should incorporate mechanisms such as data permission management, usage review, and impact assessment to safeguard user privacy and language security. Utilizing modern data storage and processing technologies can enable efficient data access and updates, thereby supporting continuous model training and optimization.

4.1.2. Deepening Parameter Efficient Fine-Tuning and Retrieval-Augmented Generation Model Applications

To improve the performance of low-resource language large language models, the key lies in effectively applying parameter-efficient fine-tuning and retrieval-augmented generation model techniques. Parameter-efficient fine-tuning involves freezing most of the model's parameters and only fine-tuning a few layers, which reduces training costs and improves efficiency [14]. This method enables the model to focus on optimizing specific tasks with a small amount of labeled data, enhancing its performance in certain use cases. By carefully adjusting fine-tuning strategies, models can better adapt to the language characteristics of specific domains and user needs. Even with limited data, the quality of generated content can remain high. Retrieval-augmented generation models combine information retrieval techniques with natural language generation. By retrieving relevant documents or data snippets, the generated content becomes more accurate and enriched. This approach improves the output quality of the model and enhances its ability to grasp contextual information.

4.2. Strategies for Enhancing Model Capabilities

4.2.1 Achieving Cross-Language Knowledge Transfer and Full-Process Contextual Interaction

To achieve cross-language knowledge migration and full-process contextual interaction, this strategy is crucial for enhancing the capabilities of large language models in low-resource languages. Cross-language knowledge migration involves transferring knowledge and features from high-resource languages to low-resource languages, enabling the latter to learn and adapt more effectively. In this process, multi-task learning can be employed to share implicit layers and parameters, allowing the model to acquire rich semantic representations when handling multilingual tasks. Specific migration strategies should be designed based on the unique characteristics of low-resource languages. This approach facilitates the effective transfer of syntax, semantic information, and contextual relationships from high-resource languages to low-resource languages, thereby improving their language processing capabilities [7]. Full-process contextual interaction emphasizes that the model should provide real-time feedback and utilize contextual information during both generation and understanding, ensuring accurate language comprehension and production.

4.2.2. Building a Dynamic Prompt Engineering and Learning Strategy Matrix

Building a dynamic prompt engineering and learning strategy matrix is a key method for enhancing the capabilities of low-resource language large language models. Dynamic prompt engineering modulates and refines input prompts dynamically to increase the flexibility of the model in meeting task demands. By assessing various tasks and situations, the system can produce the most effective prompt for the current context, making the model more accurate and performance-oriented in specific tasks [15, 16]. Combined with contextual information, hierarchical prompt strategies can be identified to provide the model with high-quality results under varying input conditions. Various learning strategies, including transfer learning, self-supervised learning, and contrastive learning, are organized into a matrix known as the learning strategy matrix. The most optimal strategy can be dynamically adjusted based on the model's performance in particular tasks and then selected accordingly. The learning path can be optimized in real-time through monitoring and feedback, enabling the model to achieve optimal learning outcomes despite limited data availability. This combination of strategies enhances the efficiency of model training, improves the generalization capability across multiple language tasks, and ensures the general applicability and practicality of low-resource language models in real-world scenarios.

4.3. Ethical and Security Safeguards

4.3.1 Improving Corpus Data Authorization and Privacy Protection Mechanisms

Enhancing corpus data authorization and privacy protection mechanisms is essential for ensuring compliance in the development and application of large language models for low-resource languages. A transparent data authorization workflow should be established during the data collection process to ensure that all corpora are legally authorized and that the rights of data providers are respected. The data usage agreement must clearly define the purpose of the data, the scope of sharing, and confidentiality clauses, thereby safeguarding the data provider's right to be informed and to make choices [17]. Privacy protection measures, such as data anonymization techniques, should be implemented to remove sensitive information and prevent privacy breaches. A strict data access mechanism must be designed to ensure that only authorized personnel can process and analyze the data, reducing the risk of security compromises. Additionally, a monitoring and auditing system should be established to periodically review the effectiveness of data management and privacy protection measures, addressing any issues that arise during usage. Education and training programs should be conducted to raise participants' awareness of data protection and foster a culture of compliance with data regulations.

4.3.2. Establishing Model Explainability and Output Review Standards

To enhance the credibility of large language models for low-resource languages, it is essential to establish standards for model explainability and output review [18]. Model explainability requires researchers to clearly understand the decision-making basis of the model during the generation process, including key features and contextual information influencing the output. This can be achieved through visualization techniques, feature importance evaluation, and decision path analysis. The goal is to help researchers and users assess the reasonableness of the model's outputs and recognize its limitations. Equally important is the formulation of output review standards. A standardized review process and evaluation criteria should be developed to systematically inspect the quality of the model-generated content, ensuring alignment with ethical principles and social responsibility. The review mechanism should encompass assessments of accuracy, relevance, and cultural sensitivity. Any inappropriate content must be promptly modified or removed. External audit mechanisms, such as expert reviews and community feedback, should also be encouraged to ensure the output is diverse and inclusive.

5 Conclusion

Research on large language models for low-resource languages highlights the potential of leveraging contextual learning to support efficient training strategies, such as zero-shot and few-shot learning, while also addressing significant challenges. By effectively mining and utilizing contextual information, researchers can enhance model performance under conditions of data scarcity. This includes optimizing prompt engineering and learning strategy frameworks, enabling flexible applications, facilitating cross-linguistic knowledge transfer, and ensuring comprehensive context interaction throughout the process. Additionally, implementing robust ethical and safety measures is essential to ensure compliance and uphold social responsibility in model development. Safeguarding data privacy, preserving cultural diversity, and establishing model interpretability and output review standards are critical steps to build user trust and foster the broader acceptance of natural language processing technologies for low-resource languages. Future research should focus on refining training strategies, enhancing model performance and interpretability to meet evolving application demands, and advancing the development and application of low-resource language technologies. Through sustained exploration and innovation, new pathways can be forged to support the preservation and exchange of diverse cultural heritages.

References

1. Z. X. Yong, C. Menghini, and S. Bach, "Lexc-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Nov. 2024, pp. 13990–14009.
2. M. A. Hasan, P. Tarannum, K. Dey, I. Razzak, and U. Naseem, "Do large language models speak all languages equally? A comparative study in low-resource settings," *arXiv preprint arXiv:2408.02237*, 2024.
3. M. Adeyemi, A. Oladipo, R. Pradeep, and J. Lin, "Zero-shot cross-lingual reranking with large language models for low-resource languages," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (Vol. 2: Short Papers)*, Aug. 2024, pp. 650–656.
4. P. Guo, Y. Ren, Y. Hu, Y. Li, J. Zhang, X. Zhang, and H. Y. Huang, "Teaching large language models to translate on low-resource languages with textbook prompting," in *Proc. 2024 Joint Int. Conf. Comput. Linguistics, Lang. Resources and Evaluation (LREC-COLING 2024)*, May 2024, pp. 15685–15697.
5. N. Kholodna, S. Julka, M. Khodadadi, M. N. Gumus, and M. Granitzer, "LLMs in the loop: Leveraging large language model annotations for active learning in low-resource languages," in *Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Cham: Springer Nature Switzerland, Aug. 2024, pp. 397–412.
6. T. Zhong, Z. Yang, Z. Liu, R. Zhang, W. You, Y. Liu, ... and T. Liu, "Opportunities and challenges of large language models for low-resource languages in humanities research," *arXiv preprint arXiv:2412.04497*, 2024.
7. G. Mani and G. B. Namomsa, "Large language models (LLMs): Representation matters, low-resource languages and multi-modal architecture," in *2023 IEEE AFRICON*, Sep. 2023, pp. 1–6.
8. V. Hangya, H. S. Saadi, and A. Fraser, "Improving low-resource languages in pre-trained multilingual language models," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Dec. 2022, pp. 11993–12006.
9. N. Kadyrbek, Z. Tuimebayev, M. Mansurova, and V. Viegas, "The development of small-scale language models for low-resource languages, with a focus on Kazakh and direct preference optimization," *Big Data Cogn. Comput.*, vol. 9, no. 5, p. 137, 2025.
10. J. McGiff and N. S. Nikolov, "Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review," *arXiv preprint arXiv:2505.04531*, 2025.
11. N. Bui, G. Nguyen, N. Nguyen, B. Vo, L. Vo, T. Huynh, ... and M. Dinh, "Fine-tuning large language models for improved health communication in low-resource languages," *Comput. Methods Programs Biomed.*, vol. 263, p. 108655, 2025.
12. H. Avetisyan and D. Broneske, "Large language models and low-resource languages: An examination of Armenian NLP," *Findings Assoc. Comput. Linguistics: IJCNLP-AACL 2023 (Findings)*, pp. 199–210, 2023.
13. T. A. Chang, C. Arnett, Z. Tu, and B. Bergen, "When is multilinguality a curse? Language modeling for 250 high-and low-resource languages," in *Proc. 2024 Conf. Empirical Methods Natural Lang. Process.*, Nov. 2024, pp. 4074–4096.
14. P. W. Khoboko, V. Marivate, and J. Sefara, "Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models," *Mach. Learn. Appl.*, vol. 20, p. 100649, 2025.
15. T. Harrington, C. Navarro, E. Davidson, C. Salazar, and C. Langston, "Recursive instruction tuning of large language models for low-resource languages," *Authorea Preprints*, 2024.
16. X. Liang, Y. M. J. Khaw, S. Y. Liew, T. P. Tan, and D. Qin, "Towards low-resource languages machine translation: A language-specific fine-tuning with LoRA for specialized large language models," *IEEE Access*, 2025.
17. K. Ogueji, Y. Zhu, and J. Lin, "Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages," in *Proc. 1st Workshop Multilingual Representation Learn.*, Nov. 2021, pp. 116–126.

18. S. Joshi, M. S. Khan, A. Dafe, K. Singh, V. Zope, and T. Jhamtani, "Fine tuning LLMs for low resource languages," in *2024 5th Int. Conf. Image Process. Capsule Netw. (ICIPCN)*, Jul. 2024, pp. 511–519.