Dean&Francis
ISSN 2959-6157

Core Applications and Techniques of RAG for Low-Resource Languages

Shanli Ouyang

College of Science, Southern University of Science and Technology, Shenzhen, Guangdong, China, 518055
12212406@mail.sustech.edu.cn

Abstract:

Low-resource languages (LRLs) are spoken by billions of people globally, yet they continue to face limited access to reliable AI-driven tools, largely due to a lack of sufficient digital text resources. This review examines recent progress in Retrieval-Augmented Generation (RAG) that addresses the specific difficulties encountered in processing LRLs. By analyzing key studies, a range of optimization strategies are identified and analyzed across the RAG workflow—including data handling, retrieval techniques, and generation refinements. The discussion emphasizes how these methods tackle fundamental problems such as scarce data, weak model performance, and poor cultural alignment. The role of cross-lingual retrieval and knowledge distillation is also explored as a way to make AI systems more accessible and useful for speakers of low-resource languages. In addition to outlining RAG's potential for reducing language-based digital inequality, this survey notes remaining obstacles and suggests productive avenues for further research. It is hoped that this structured summary of methods and use cases will support future efforts toward inclusive AI and the protection of linguistic diversity.

Keywords: Low-resource Languages; Retrieval-Augmented Generation (RAG); Cross-lingual Information Retrieval; Digital Inclusion

1. Introduction

Low-resource languages (LRLs) are defined by their limited digital footprint, such as small Wikipedia corpora, which makes them poorly compatible with general-purpose AI models. For communities speaking these languages, this translates into pervasive service failures, from high error rates in speech recognition to factual inaccuracies in critical translation tasks. The problem is compounded by the fact that multilingual large language models (LLMs) are predominantly trained on data from a few dominant languages like English and Chinese, resulting in poor linguistic and cultural adaptation for the rest of the world.

Retrieval-Augmented Generation (RAG) has emerged as a promising solution. This approach inte-

grates external knowledge retrieval with text generation, enabling the incorporation of up-to-date or domain-specific information without costly model retraining. The hybrid nature of RAG not only mitigates the tendency of generative models to "hallucinate" but also offers more flexibility than traditional retrieval systems. While much research has focused on general RAG improvements, this survey specifically examines its application in low-resource contexts. The analysis focuses on implementation pathways to elucidate how RAG can be adapted to support the unique challenges of LRLs, thereby helping to bridge the technological gap for their speakers.

2. Core Applications and Techniques of RAG for Low-Resource Languages

Applying RAG to LRLs hinges on overcoming two primary challenges: bridging the data gap and enhancing service usability. These objectives have driven the development of specialized techniques across the three core stages of the RAG pipeline. The literature validates the effectiveness of these technical solutions, particularly within the domain of Cross-Lingual Information Retrieval (CLIR).

2.1 Core Applications

2.1.1 Cross-Lingual Information Retrieval (CLIR)

Effective CLIR—using an LRL query to find information in a high-resource language repository—is fundamental to the success of RAG in these contexts. The quality of the "Retrieval" stage dictates the relevance and accuracy of the final generated output. Research has focused on making CLIR viable despite data scarcity and model limitations by advancing three interconnected areas: optimizing model architectures, creating novel data supervision techniques, and ensuring domain-specific factual accuracy.

A. Aligning Pre-training with Retrieval

While multilingual models like mBERT possess cross-lingual capabilities, their standard pre-training objectives (e.g., MLM) are fundamentally misaligned with the query-document relevance ranking required for retrieval. Addressing this, Yu et al. re-engineered the pre-training process with retrieval-centric objectives (QLM and RR) and a more efficient Global-Sliding Window (GSW) attention mechanism to better handle long documents. These fundamental optimizations directly aligned the model with ranking tasks, yielding significant Mean Average Precision (MAP) gains of 13.9%–29.7% on the CLEF benchmark [1].

B. Innovations in Supervision through data scarcity

To overcome the lack of annotated data in LRLs, researchers have developed ingenious supervision methods. One prominent strategy is weak supervision. Jiang et al. pioneered a clever approach that automatically generates training data from existing parallel corpora (e.g., Lithuanian-English pairs), circumventing the need for manual annotation. Fine-tuning mBERT on this data significantly boosted retrieval MAP from 57.4 to 61.3 [2]. A complementary technique is knowledge distillation. The OPTICAL framework, developed by Huang et al., uses a teacher-student architecture where a powerful "teacher" model trained on a high-resource dataset transfers its retrieval knowledge to a "student" model using only unannotated parallel sentences. This resource-efficient pathway outperformed traditional NMT-based baselines by 13.7% MAP for languages like Swahili [3].

C. Enhancing Domain Specificity and Factual Accuracy

In specialized domains like law or religion, factual accuracy is paramount to prevent RAG systems from producing confident-sounding hallucinations. To counter this, Alshammary et al. introduced the RFPG framework for querying Arabic religious texts. Their key innovation is a post-retrieval fact-checking layer that uses an LLM to filter results and distill verifiably authentic passages ("gold hadiths"). This, combined with meticulous prompt engineering for source citations, achieved 100% accuracy and eliminated all hallucinations in their tests, underscoring the necessity of integrating domain-specific validation to build trustworthy RAG systems [4].

In concert, these CLIR advancements fortify the RAG pipeline for LRLs. By optimizing models, creatively leveraging scarce data, and enforcing factual grounding, they construct a robust and reliable knowledge retrieval frontend. This provides a solid foundation for the generation stage, making RAG a genuinely viable pathway toward bridging the digital language divide.

2.1.2 Speech and Audio-Driven Tasks: Adapting to Speech-Dominant Low-Resource Scenarios

In many LRL communities, oral communication is pre-

ISSN 2959-6157

dominant, making performance in automatic speech recognition (ASR), audio question answering (AQA), and audio classification critically important. Large audio language models (LALMs), however, often perform poorly in these tasks due to limited training data and linguistic complexity such as code-mixing.

Dutta et al. conducted pioneering work to evaluate RAG's potential in this space [5]. After noting the low zero-shot performance of LALMs like LTU-AS, GAMA, and Pengi—especially in ASR and AQA—the authors implemented a RAG framework using Whisper for audio embeddings and Indic-sbert for text embeddings, constructing a cross-modal retrieval database.

The results can demonstrate several traits:

RAG proved highly effective in closing "knowledge gaps." For monolingual Hindi ASR and hate speech classification, it reduced character error rate by 46.1% and improved accuracy from 0.51 to 0.82 [5].

However, its benefits diminished where tasks required complex reasoning or disambiguation, such as in code-mixed ASR (e.g., Hinglish) and AQA. This suggests that RAG cannot compensate for shortcomings in innate model reasoning or poorly aligned retrieval in linguistically complex settings.

This study underscores that RAG is highly effective for knowledge-augmentation tasks but reveals its limits in reasoning-heavy and mixed-language applications. The real challenge lies not only in retrieving information but in building models that can deeply reason with multimodal inputs. Future work should focus on context-aware retrieval and enhanced integration mechanisms for low-resource speech and language tasks.

2.2 Technical Process of RAG for Low-Resource Languages

Building a RAG system for a low-resource language (LRL) is a process of navigating scarcity and uncertainty at every step. This section introduces the technical pip-line across the three core modules—retrieval, augmentation, and generation—by raising evidence from recent studies, which is briefly indicated in Figure 1. The goal is to show not just what each module does, but how innovations in one stage directly impact the others, creating a more coherent and effective whole system.

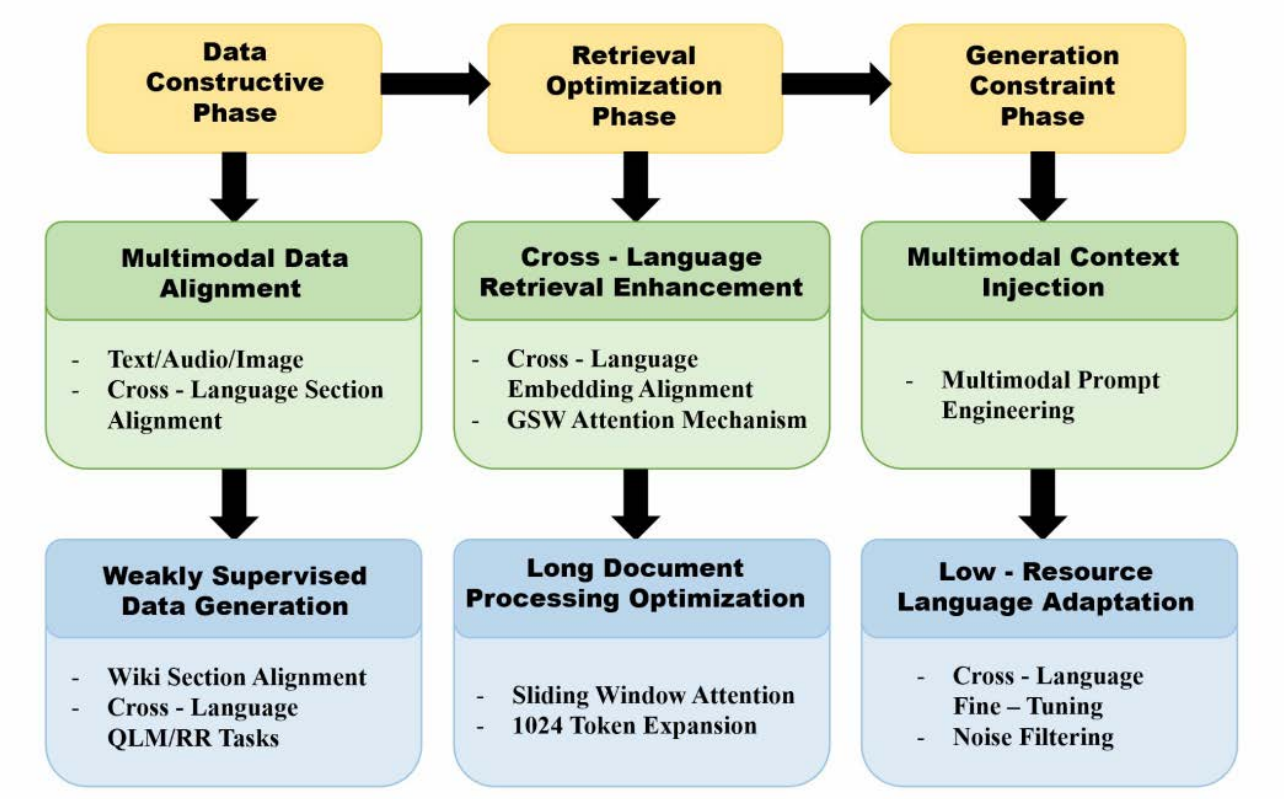

Figure 1. Technologies Introduced in the Low-Resource RAG Pipeline

2.2.1 The Retrieval Module: Finding Needles in a Haystack

The first challenge is finding relevant information in sparse and often noisy LRL corpora. One approach is to improve the signal quality by aligning different data types. Contrastive learning, for example, maps text and images into a shared embedding space, which Li et al. used to boost retrieval recall by 12.6% on Urdu tasks [6]. This directly addresses the cross-modal misalignment issues that Dutta et al. identified when using simpler monolingual embeddings [5].

To compensate for the lack of a large corpus, hybrid retrieval methods are employed. By combining the keyword-matching of sparse retrieval (e.g., BM25) with the semantic understanding of dense retrieval, frameworks like RADA by Seo et al. achieved a 7.2% F1 score increase on Hindi QA tasks [7]. This fusion principle also extends to code-mixed speech, where Dutta et al. used Reciprocal Rank Fusion to merge audio and text results [5]. Furthermore, to reduce the dependency on annotated data, unsupervised strategies have proven effective. The PARC framework, for instance, computes similarity between LRL inputs and an English corpus to find relevant samples, building on the weak-supervision ideas of Jiang et al. and improving accuracy by 5.1% across 10 LRLs [2,8].

2.2.2 The Augmentation Module: From Raw Data to Reliable Context

Retrieved information is often not clean or reliable enough to be passed to a generator, making augmentation a critical filtering stage. A key technique is re-ranking and fact-filtering. The RFPG framework, designed by Alshammary et al. for Arabic texts, uses a powerful external LLM (GPT-4o-mini) to re-rank passages and select only the most authentic content. This rigorous filtering achieved 100% answer accuracy and has been adapted to other domains, significantly reducing hallucination rates [4].

This module also handles cross-lingual context fusion, where knowledge from a high-resource language is adapted for an LRL query. While Nie et al. showed this can be done by dynamically injecting English samples into a Zulu prompt [8], Huang et al. offered a more deeply integrated solution with their OPTICAL framework. It uses knowledge distillation to transfer retrieval expertise from an English model to a Somali one, requiring only a small parallel corpus to achieve a 13.7% MAP boost [3].

2.2.3 The Generation Module: Speaking Accurately and Efficiently

The final generation module inherits both the benefits and the remaining imperfections of the previous stages. Its primary goals are to generate fluent text while suppressing hallucinations and remaining computationally efficient. To control hallucinations, simple prompt constraints are used, such as instructing the model to "only cite retrieved results," which Seo et al. found cut factual errors nearly in half for a Llama2-7B model on a Hindi QA task [7]. A more advanced technique is self-verification, where tools like RAGAS are used to perform reverse checks, as Dutta et al. did to improve fact consistency scores on the IndicQA dataset [5].

Finally, to make these systems practical in resource-constrained settings, parameter-efficient fine-tuning (PEFT) methods like LoRA and QLoRA are essential. Jiang et al. demonstrated that fine-tuning just 1% of mBERT's parameters could drastically cut training time for Lithuanian retrieval while maintaining high accuracy [2]. This focus on efficiency also enables architectural innovations, such as the specialized attention heads designed by Li et al. for morphologically rich languages like Urdu, which improved generation quality and showed that efficiency and performance can be pursued in tandem [6].

3. Typical Use Cases of RAG for Low-Resource Languages

With the continuous development of RAG technology for low-resource languages, its applications have expanded to critical domains such as public health, disaster response, education, and legal services. By solving the core problems of data scarcity and enhancing information accessibility, RAG manifests practical value in preserving linguistic diversity. This section summarizes several key application scenarios and their implications.

3.1 Public Health: Health Information Dissemination

In public health, the lack of localized health information often creates barriers for low-resource language speakers.

ISSN 2959-6157

RAG helps overcome this by leveraging multimodal retrieval and generation. For example, the multimodal RAG framework generated contagions prevention guidelines in Swahili by retrieving English guidelines from authoritative sources like WHO and combining them with relevant images. The system achieved a BLEU score of 28.4, representing a 38.5% improvement over text-only generation approaches[2].

In speech-based medical applications, RAG-enhanced audio language models can improve the accuracy of medical term recognition by cross-referencing a medical terminology dictionary or other ways, therefore reduce the character error rate and reduce the risk of misinterpretation in clinical settings.

3.2 Disaster Response: Emergency Information Delivery

Timely and accurate information is critical in disaster scenarios, yet language barriers often hinder effective communication in low-resource regions. The system designed for Pashto speakers retrieved flood-related imagery and damage assessment data in response to user queries, generating summaries that included location details, estimated affected populations, and emergency contact information. The system achieved a retrieval recall of 82%, which outperforms unimodal models by 18.2% [2]. Those kinds of RAG systems facilitate real-time emergency response by integrating multimodal data such as satellite imagery, local situation reports, and geographic information.

Another example is, in cross-language disaster reporting, the OPTICAL framework employed optimal transport distillation to enhance Somali-English retrieval performance. It achieved a 13.7% improvement in mean average precision over conventional translation-based methods [6]. The scientific result can enable organizations to access local reports more accurately and implement rescue operations.

3.3 Education and Legal Services: Promoting Knowledge Equity

In education and legal domains, speakers of low-resource languages often face challenges because of limited educational materials and complex legal documents. RAG helps bridge these gaps through context-aware retrieval and generation. For instance, the PARC framework generated educational materials in Zulu and Swahili by retrieving topic-labeled English news articles and producing localized content. This approach improved classification accuracy by 36.1% compared to direct generation methods [1]. In those rigorous fields, a RAG-based system can enhance the precision of target language-related keywords from a textual perspective, and may help eliminate unnecessary social risks.

4. Core Challenges and Future Technical Directions

Although notable progress has been made in Low-Resource RAG and many researchers are making efforts to further study, the technologies in this field still face several core challenges.

Data quality and diversity: currently available weakly-supervised data primarily originate from sources like Wikipedia or parallel corpora, resulting in limited domain coverage.

Semantic Alignment in Cross-Lingual Retrieval: the context of code-mixed languages (e.g., Hinglish) is still hard to match the long-context of information.

The adaptation of different cultures: the generated information is easy to neglect some details of some covert cultural conventions.

These are not the only problems, instead, there are more difficulties left for researchers to overcome. However, as technology advances, many existing challenges are likely to find solutions, which necessitates that researchers in the field conduct further studies with more refined objectives, integrating relevant studies.

In fact, advancing low-resource RAG technology not only requires expertise in LLM but also needs crucial linguistic contributions. From a linguistic perspective, the world's languages can be classified into numerous families, branches, sub-branches and so on. When classifications are highly granular, some languages (or dialects) may have very few speakers or even be endangered. Sometimes even within very small geographic areas, a remarkable diversity of languages can exist. To address such scenarios, in spoken-language matching, there should be increased focus on developing phonological proximity-based transfer learning. This technique would utilize

data from related languages in the same language family to probabilistically infer and generate retrieval texts for the target language, thereby supporting RAG implementation.

In terms of script transfer, while the related-language transfer approach is valuable, challenges arise when languages are related but use significantly different writing systems. For example, within the Turkic language family, Latin script (e.g., Turkish), Cyrillic script (e.g., Kazakh), and Arabic script (e.g., Uyghur) are all used. In such cases, reliance on script transfer would not be sufficient. It should be combined with phonetic transfer, where the pronunciation of the source language is mapped into the writing system of the target query language. The goal is to generate a text in the target language's script that approximates the source language's pronunciation rather than providing a direct translation, because low-resource languages are likely to be translated wrong, causing misleading results. For instance, the Turkish sentence "Ben ekmek yemeyi severim" could be phonetically transcribed into the Uyghur Arabic script as "بەن ئەكمەك يېمەيى سېۋىرىمەن ". This is not a translation into Uyghur but a Uyghur-script phonetic rendering of the Turkish sentence. For extremely low-resource languages, this method is highly valuable as it leverages related languages to augment the available text data in the target language, even if the semantic meaning is not guaranteed to be precise in the target language. Also, using Turkish Latin script to write Uyghur language may help the LLM understand Uyghur better as it might be augmented by Turkish text information. By implementing such techniques, the consistency of transferred texts within the target language can be improved, thereby enhancing the usability of RAG systems.

Meanwhile, there are languages like Basque (used in the northern Iberian Peninsula) that are not only spoken by a small population but are also linguistically isolated. For such cases, advancing RAG technology necessitates more extensive foundational work, including large-scale phonetic data collection and the digitization of available linguistic resources and documents. Efforts should be made to support the development of effective RAG systems for these uniquely challenging low-resource contexts.

5. Conclusion

This study confirms that Retrieval-Augmented Generation (RAG) effectively solved some core challenges of low-resource language (LRL) processing by cross-lingual retrieval, knowledge distillation and other technologies. RAG enhances information accessibility and service quality for low-resource linguistic communities while to some degree maintaining the cultural relevance. Although challenges in complex reasoning and cultural adaptation remain, RAG still represents one of the most promising pathways toward digital inclusion. Admittedly, this survey still has some shortcomings. For example, it primarily adopts a qualitative analysis of existing literature and provides a comprehensive overview, lacking comparisons between different RAG frameworks; what's more, the survey is limited in-depth discussion on linguistic typology and deep cultural conventions in shaping model performance, which just happens to provide a potential innovative direction for future research. Future work can focus on developing linguistically-informed retrieval frameworks or cultural differences between common languages and LRLs to handle low-resource scenarios and preserve linguistic diversity.

References

[1] Yu, P., Fei, H., & Li, P. (2021). Cross-lingual language model pretraining for retrieval. In Proceedings of the Web Conference 2021 (WWW '21). ACM. https://doi.org/10.1145/3442381.3449830

[2] Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., & Zhao, L. (2020). Cross-lingual information retrieval with BERT. arXiv preprint arXiv:2004.13005. https://arxiv.org/abs/2004.13005

[3] Huang, Z., Yu, P., & Allan, J. (2023). Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23). ACM. https://doi.org/10.1145/3539597.3570468

[4] Alshammary, M., Uddin, M. N., & Khan, L. (2024). RFPG: Question-Answering from Low-Resource Language (Arabic) Texts using Factually Aware RAG. In Proceedings of the 2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC) (pp. 107-116). IEEE. https://doi.org/10.1109/CIC62241.2024.00023

ISSN 2959-6157

[5] Dutta, B., Ranjan, R., Jain, A., Singh, R., & Vatsa, M. (2025). Can RAG-driven enhancements amplify audio LLMs for low-resource languages? In Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10889964

[6] Li, Z., & Ke, Z. (2025). Cross-Modal Augmentation for Low-Resource Language Understanding and Generation. In Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025) (pp. 90–99). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.magmar-1.9

[7] Seo, M., Baek, J., Thorne, J., & Hwang, S. J. (2024). Retrieval-augmented data augmentation for low-resource domain tasks. arXiv preprint arXiv:2402.13482. https://arxiv.org/abs/2402.13482

[8] Nie, E., Liang, S., Schmid, H., & Schütze, H. (2023). Cross-lingual retrieval augmented prompt for low-resource languages. arXiv preprint arXiv:2212.09651. https://arxiv.org/abs/2212.09651