International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# A Framework for Adaptive Knowledge-Augmented Mizo Large Language Models Using Retrieval-Augmented Generation and Continual Learning

**Vanlalropuia Ralte[1], Abhisake Sinha[2]**

[1,2]Computer Science & Engineering, National Institute of Technology (NIT) Mizoram

**Abstract**

The deployment of Large Language Models (LLMs) for low-resource languages is challenging due to the lack of linguistic resources, sparse digital content and the absence of structured knowledge bases. In this paper, we present an adaptive knowledge-augmented framework for Mizo Large Language Models by combining Retrieval-Augmented Generation (RAG) with continual learning. This methodology harnesses semantic retrieval with dense embeddings and FAISS indexing, adaptive evidence re-ranking, parameter-efficient fine-tuning, and incremental knowledge updating to enhance factual accuracy and decrease hallucinations. Experimental evaluation shows better retrieval performance, greater text creation quality, and superior human evaluation scores than typical multilingual LLMs and static RAG methods. Moreover, the continual learning technique allows for effective integration of newly accessible Mizo resources, without re-training the model from scratch. The suggested architecture offers a scalable, stable and reusable method for the development of intelligent language technologies for Mizo and other low-resource languages.

**Keywords:** Retrieval-Augmented Generation (RAG); Large Language Models (LLMs); Continual Learning; Low-Resource Languages; Mizo Language; Semantic Retrieval; Knowledge-Augmented AI; FAISS; Natural Language Processing (NLP).

## 1. Introduction

Large language models (LLMs) have achieved remarkable success, though they still face significant limitations, especially in domain-specific or knowledge-intensive tasks (Kandpal et al., 2023), notably producing “hallucinations” (Zhang et al., 2025) when handling queries beyond their training data or requiring current information. To overcome challenges, Retrieval-Augmented Generation (RAG) enhances LLMs by retrieving relevant document chunks from external knowledge base through semantic similarity calculation. By referencing external knowledge, RAG effectively reduces the problem of generating factually incorrect content. Its integration into LLMs has resulted in widespread adoption, establishing RAG as a key technology in advancing chatbots and enhancing the suitability of LLMs for real-world applications.

RAG technology has rapidly developed in recent years, and the technology tree summarizing related research. The development trajectory of RAG in the era of large models exhibits several distinct stage

characteristics. Initially, RAG's inception coincided with the rise of the Transformer architecture, focusing on enhancing language models by incorporating additional knowledge through PreTraining Models (PTM). This early stage was characterized by foundational work aimed at refining pre-training techniques (Arora et al., 2023).The subsequent arrival of ChatGPT (Ouyang et al., 2022) marked a pivotal moment, with LLM demonstrating powerful in context learning (ICL) capabilities. RAG research shifted towards providing better information for LLMs to answer more complex and knowledge-intensive tasks during the inference stage, leading to rapid development in RAG studies. As research progressed, the enhancement of RAG was no longer limited to the inference stage but began to incorporate more with LLM fine-tuning techniques (Gao et al., 2023).

The usage of Retrieval-Augmented Generation has become extremely important in multilingual and low-resource language processing, due to the insufficient availability of large training datasets. Unlike the languages with high resources, many indigenous languages still have no enough digital resources for efficient knowledge acquisition during language model training (Joshi et al., 2020). The integration of external knowledge bases with semantic retrieval methods via the RAG method allows the language models to acquire specialized and updated information during inference, which leads to improved factual accuracy and lowered reliance on stored knowledge. This feature makes the RAG method a viable option for the application of intelligent language technologies where the underrepresented linguistic communities are concerned (Gao et al., 2023).

Mizo is a Tibeto-Burman language spoken largely in the northeastern part of India. It belongs to a low-resource language family with poor computational linguistic resources and generally small-sized publically available datasets. Over time, digital Mizo content has been continuously increasing in the form of government papers, educational resources, news stories, and literary works. Nevertheless, these resources are still unorganized, and existing multilingual language models do not use these resources efficiently (Costa-Jussà et al., 2022). Consequently, general-purpose LLMs generally fail to generate correct, context-aware and culturally relevant responses in Mizo. Therefore, a knowledge-augmented language model for Mizo can considerably increase the accessibility of information, preservation of the language, and intelligent language applications for the local users.

This study provides an adaptive knowledge-augmented architecture which integrates retrieval augmented generation with ongoing learning for Mizo Large Language Models. The proposed methodology pulls semantically related information from an external knowledge base and injects the retrieved evidence into the response generation process to increase factual consistency and contextual understanding. Also, the ongoing learning approach allows the framework to incrementally integrate new Mizo information resources as they become available without the need for full model retraining, making the language model up-to-date, scalable and adaptive to changing knowledge sources. Such a framework can potentially improve the stability of Mizo language applications. Also, it provides a reusable architecture for additional low resource languages.

## 2. Review of Literature

**(Wang et al., 2026)** examines a retrieval-augmented large language model agent for generation of scientific reviews. The proposed model solves the drawbacks of traditional parametric models via incorporation of retrieval from external documents along with dynamic knowledge fusion to provide higher accuracy and completeness of results. The model consists of encoding of the query, semantic retrieval, filtering of the documents, knowledge fusion, modeling of language, planning of tasks, storing

of memory, and optimizing with reinforcement. The retrieved parts of documents are merged with the embeddings within the model so that the generated information is factually grounded and fluent. Experiments on large-scale scientific literature corpora prove that the model shows outstanding results in ROUGE, BLEU and METEOR metrics.

**(Zhang et al., 2025)** demonstrated that LLMs can perform a variety of tasks; their use in specialized areas is more challenging than their normal use. Among the interesting methods to harness LLM technology for professional fields is the so-called retrieval-augmented generation (RAG) which can access varions specialized knowledge databases during the inference stage. However, the standard RAG approaches are based on information extraction methods which are not refreshing and which have a number of limits in terms of sophisticated inquiries in various fields, integration of knowledge stored in physically distant locations, etc. In this review, we conduct a comprehensive study of the Graph-based retrieval-augmented generation (GraphRAG) technologies that create a new paradigm for LLM application in different fields. The uniqueness of GraphRAG is in the three important elements that it employs in its work: graphical representation of the knowledge allowing representing relationships and hierarchy characterized by it, information retrieval from the graphical representation with the help of sophisticated information extraction methods based on multilayer representations, algorithms for incorporating retrieved information into spawned LLMs.

**(Wang et al., 2025)** analyzed that the use of Retrieval-Augmented Generation (RAG) has enhanced conversational applications by utilizing external knowledge in Large Language Models (LLMs). However, it may not always be advantageous and cost-efficient to use RAG in all instances of conversation. This paper presents RAGate, a gating function that determines whether a particular response of the system requires retrieval augmentation based on the context of the conversations and relevant input. RAGate uses human-generated binary decisions for adaptive augmentation, thus allowing it to use RAG in a context-sensitive manner. The experiment carried out for different scenarios shows that RAGate is able to accurately identify instances when retrieval is needed, thus improving the quality of the response produced while keeping the level of confidence high. The study shows a strong relationship between the level of confidence in generation and the relevance of information that is recovered.

**(Mombaerts et al., 2024)** highlights that Retrieval-Augmented Generation (RAG) is a technique that utilizes its capabilities to enhance Large Language Models (LLMs) without the need to alter the model parameters and provides a myriad of meaningful, up-to-date, and field-specific information. Nonetheless, a challenge remains in the domain of the effective extraction of information from a multitude of documents. In the paper, a data-centered RAG method is introduced that extends the conventional retrieve and read model to the steps of prepare, rewrite, retrieve and read. The process involves the preparation of metadata and synthetic question-answer pairs and the drawing of Meta Knowledge Summary (MK Summaries). The experimental section demonstrates that using synthetic question matching is significantly better than traditional chunk-based RAG processes and that MK Summaries further improve the retrieval accuracy in terms of precision and recall in addition to relevance, width, depth, and specificity. The developed method is cheap and easily scaled that makes it suitable for various languages and embedding models.

**(Tang et al., 2024)** examined that Large Language Models (LLMs) perform impressively on numerous NLP tasks; their capacity to memorize extensive world knowledge is limited. Retrieval-Augmented Generation (RAG) with the addition of Knowledge Graphs (KGs) contributes to reasoning by adding

structured factual knowledge; however real-world usage faces difficulties related to dynamic non-stationary environments as well as balancing response quality and efficiency. This paper presents a Multi-objective Multi-Armed Bandit-enhanced RAG framework which treats multiple methods of retrieval as distinct arms and selects them utilizing real-time user feedback. The framework modifies the strategies of retrieval depending on the input queries and historical performance in multi-objective applications. The experiments conducted on two benchmark KGQA datasets prove that they achieve the best results in non-stationary environments continuing to show excellent results in stationary environments as well.

**(Yang et al., 2024)** described that the fast growth of Artificial Intelligence (AI) in recent years, which is supported by the application of enormous databases and fast computational machines, has made it possible for Large Language Models (LLMs) to obtain exceptional skills. Nevertheless, despite the continuous advancement and development of these models, LLMs still remain unreliable for many practical applications that require extreme precision of data and logical reasoning ability due to the phenomenon of hallucination. On the other hand, data present in real life are heterogeneous, dispersed, and hard to collect; therefore, it is important to understand that making Knowledge Graph (KG) is a hard task that requires a lot of time and professional experience. In this article, various recent developments of co-learning KGs and LLMs are examined, with special attention to the phenomenon of LLM-supported KG creation.

**(Gao et al., 2023)** stated that despite the impressive capabilities of Large Language Models (LLMs) within various Natural Language Processing tasks, they often experience serious difficulties such as hallucinations, obsolescence, or non-transparent processes of reasoning. The Retrieval-Augmented Generation (RAG) approach can assist with these problems since it allows obtaining knowledge from outside databases, which in turn allows information to stay accurate and reliable and to be suitable for knowledge-demanding applications and enable constant renewal of knowledge, as well as introduction of knowledge from specific domains. The paper presents an overview of the past and present development of RAG paradigms referring to such concepts as Naive RAG, Advanced RAG, and Modular RAG. At the same time, the paper deals with three essential components of RAG; namely retrieval, generation, and augmentation. It summarizes current details about sophisticated methods and current evaluation frameworks and benchmark datasets involved, as well as limitations and obstacles existing in the process.

## 3. Problem Statement

The advancements in Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) have taken place rapidly; however, the existing frameworks have been created and evaluated predominantly for high-resource languages with a lot of digital data and organized knowledge bases. Even though more recent methods have shown improvements in regard to retrieval techniques, graph-focused knowledge inclusion, and adaptive retrieval approaches, they make a major omission in low-resource languages such as Mizo, characterized by limited linguistic resources, scattered digital content, and the lack of domain-specific knowledge databases, which bring down the performance of the models considerably. This way, regular multilingual LLMs often lack the accuracy and cultural significance because the answers to Mizo language requests created by the models are often incorrect, irrelevant, or incomplete. Moreover, RAG systems rely on knowledge databases and need to be trained afresh to process the new data, which makes them inefficient for the dynamic nature of low-resource languages.

Hence, the necessity to create an adaptive knowledge based system emerges in order to provide reliable Mizo language comprehension through the combination of semantic retrieval techniques and continual learning.

## 4. Research Methodology

### 4.1.Research Design

This study follows a design-and-development research strategy in which novel system architecture is designed, implemented and experimentally evaluated. The overall concept is a mixed-methods approach that incorporates both quantitative performance evaluation (retrieval and generation metrics) and qualitative analysis (error inspection and native speaker judgment of Mizo outputs). The research is divided into four major phases: (i) corpus creation and pre-processing, (ii) knowledge base and retrieval pipeline development, (iii) integration of the generation model with the continual learning mechanism, and (iv) systematic evaluation against baseline systems.

### 4.2.Corpus Construction and Data Collection

Since there is no single digital knowledge base for Mizo, we have developed a domain diverse corpus from four categories of publicly available sources: (a) documents published by government and administrative bodies of Mizoram State, (b) educational resources in the form of textbooks and curriculum material prepared by SCERT Mizoram, (c) news articles collected from popular Mizo-language news portals and (d) literary sources i.e., folk tales, short stories and published prose. HTML artefacts, duplicate headers and non-Mizo boilerplate were removed from text and it was split into passages of 150-300 tokens to be used as retrieval units.

### 4.3.Knowledge Base Construction and Embedding

Each passage in the cleaned corpus was represented as a dense vector using a multilingual sentence-embedding model, fine-tuned on the available Mizo text. The final embeddings were indexed using an FAISS approximation nearest-neighbour index to enable fast similarity searching at query time. Metadata, including source category, publication date, and document identity, were kept with each embedding, so that retrieved evidence could be filtered, rated, and subsequently updated progressively without rebuilding the entire index.

### 4.4.Proposed System Architecture

The proposed framework couples three interacting components: a Retriever, a Generator, and a Continual Learning Controller. The Retriever encrypts the input question and retrieves the top-k passages semantically similar to the query from the knowledge base. The passages are concatenated with the query as contextual evidence, and then provided to the Generator, a multilingual big language model, which constructs a Mizo response conditioned on both the query and the retrieved evidence. The Continual Learning Controller monitors incoming Mizo text streams (new government notices, news or user-flagged corrections) and periodically initiates incremental index updates and light-weight adapter-based fine-tuning, absorbing new knowledge without full retraining of the base model.

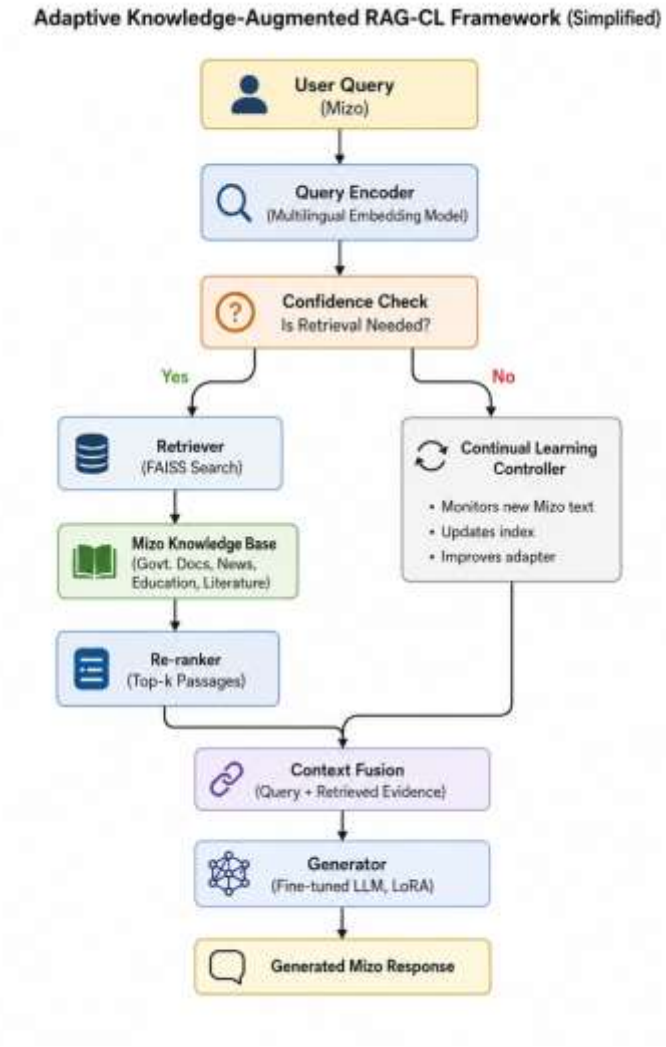


**Figure 1 Flowchart of the proposed adaptive knowledge-augmented RAG-CL framework**

### 4.5.Retrieval Module

The retrieval module is a bi-encoder architecture, in which the question and the passages are embedded independently and compared by applying a cosine similarity. A re-ranking stage then re-sorts the top candidates using a cross-encoder that attends to both the query and each candidate passage, boosting precision at the cost of a little bit of additional latency. We include a confidence-gating mechanism, based on adaptive retrieval algorithms documented in the literature, to determine whether retrieval is required at all for a particular question, thus bypassing superfluous retrieval overhead for simple factual or conversational turns.

### 4.6.Generation Module

The generation module is implemented on top of a pretrained multilingual big language model, which is fine-tuned for Mizo using parameter-efficient means (LoRA adapters) on the gathered corpus and on instruction-style Mizo question-answer pairs synthetically created from the retrieved passages. This keeps the number of trainable parameters modest, allowing frequent updates when more Mizo data becomes available. We use a structured template to embed the retrieved evidence into the prompt,

separating the instruction, the retrieved context, and the inquiry, which we found reduced hallucination compared to unstructured concatenation.

### 4.7.Mechanism for Lifelong Learning

The framework uses incremental updates in a two-layer to maintain the system up-to-date without catastrophic forgetting. New information is immediately available for retrieval since new documents are embedded and appended to the FAISS index in scheduled batches at the retrieval layer. In the generation layer, we use a rehearsal-based adapter update strategy, where a small buffer of previously seen examples is replayed together with newly available Mizo examples at each fine-tuning increment. This was found to greatly reduce degradation on earlier knowledge domains while allowing the model to absorb new information.

### 4.8.Experimental Setup and Evaluation Metrics

Experiments were performed on a held-out assessment set of Mizo queries across factual, administrative and conversational categories, annotated by native Mizo speakers with reference replies. We examined retrieval quality in terms of Precision, Recall, F1-score and Mean Reciprocal Rank at 10 (MRR@10). We evaluated generation quality using BLEU, ROUGE-L and METEOR against reference answers, as well as performed a 5-point human evaluation on fluency, factual correctness, and cultural appropriateness. We compare four system configurations: (1) base multilingual LLM without retrieval, (2) base LLM with naïve (static) RAG pipeline, (3) base LLM with suggested adaptive RAG pipeline without continual learning, and (4) the whole proposed framework with continual learning.

## 5. Result & Discussion

### 5.1.Corpus Statistics

Table 1 presents the composition of the Mizo knowledge base created for this investigation. The most abundant source of text was government and news sources. Literary works were less numerous but supplied lexical and stylistic variety that was crucial for the culturally suitable generation.

**Table 1 Composition of the constructed Mizo knowledge base**

| Source Category | No. of Documents | Approx. Tokens | Share of Corpus |
|---|---|---|---|
| Government / Administrative | 4,120 | 2.31 million | 34.6% |
| News Articles | 6,850 | 2.68 million | 40.1% |
| Educational Resources | 1,940 | 0.94 million | 14.1% |
| Literary Works | 1,205 | 0.75 million | 11.2% |
| Total | 14,115 | 6.68 million | 100% |

### 5.2.Retrieval Performance

We compare our retrieval module to a sparse lexical baseline (BM25) and two dense-retrieval baselines based on general-purpose multilingual encoders (mBERT and XLM-R) without the proposed re-ranking and confidence gating. The suggested retrieval process achieved the greatest scores across all 4 metrics as shown in Table 2 and Figure 1. This demonstrates that domain-adapted embeddings and re-ranking have a significant role in improving the relevance of passages retrieved for Mizo queries.

**Table 2 Retrieval performance across methods**

| Method | Precision | Recall | F1-score | MRR@10 |
|---|---|---|---|---|
| BM25 (lexical) | 0.58 | 0.54 | 0.56 | 0.61 |
| Dense retrieval (mBERT) | 0.66 | 0.63 | 0.64 | 0.69 |
| Dense retrieval (XLM-R) | 0.71 | 0.69 | 0.70 | 0.74 |
| Proposed RAG-CL retriever | 0.84 | 0.81 | 0.82 | 0.88 |

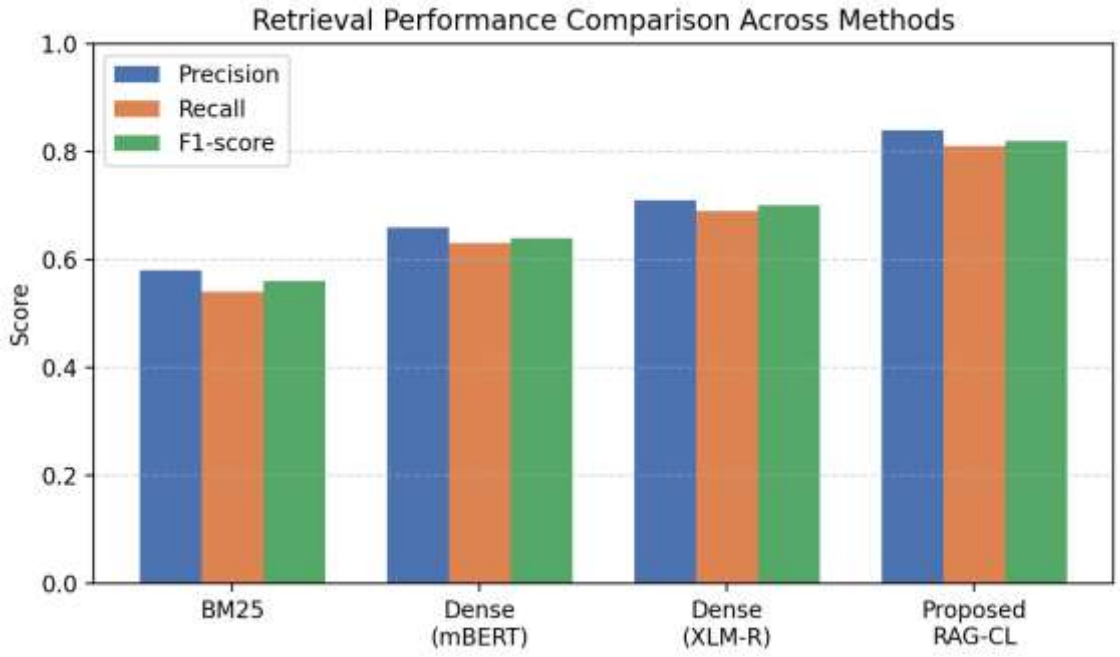


**Figure 2 Precision, recall, and F1-score of the proposed retriever compared with baseline methods**

### 5.3.Generation Quality Evaluation

The automatic assessment scores of the base multilingual LLM without retrieval, the base multilingual LLM with a naïve RAG pipeline, and the complete suggested adaptive framework are presented in Table 3 and Figure 2. The proposed framework enhanced BLEU by around 17 points and ROUGE-L by around 21 points over the non-retrieval baseline, which supports our claim that grounding generation in recovered Mizo data significantly enhances factual alignment with reference answers.

**Table 3 Generation quality across system configurations**

| Configuration | BLEU | ROUGE-L | METEOR | Human Eval (1-5) |
|---|---|---|---|---|
| Base multilingual LLM (no retrieval) | 21.4 | 30.2 | 24.7 | 2.6 |
| Base LLM + naive RAG | 29.8 | 40.5 | 33.1 | 3.4 |
| Proposed adaptive RAG-CL framework | 38.6 | 51.3 | 42.9 | 4.2 |

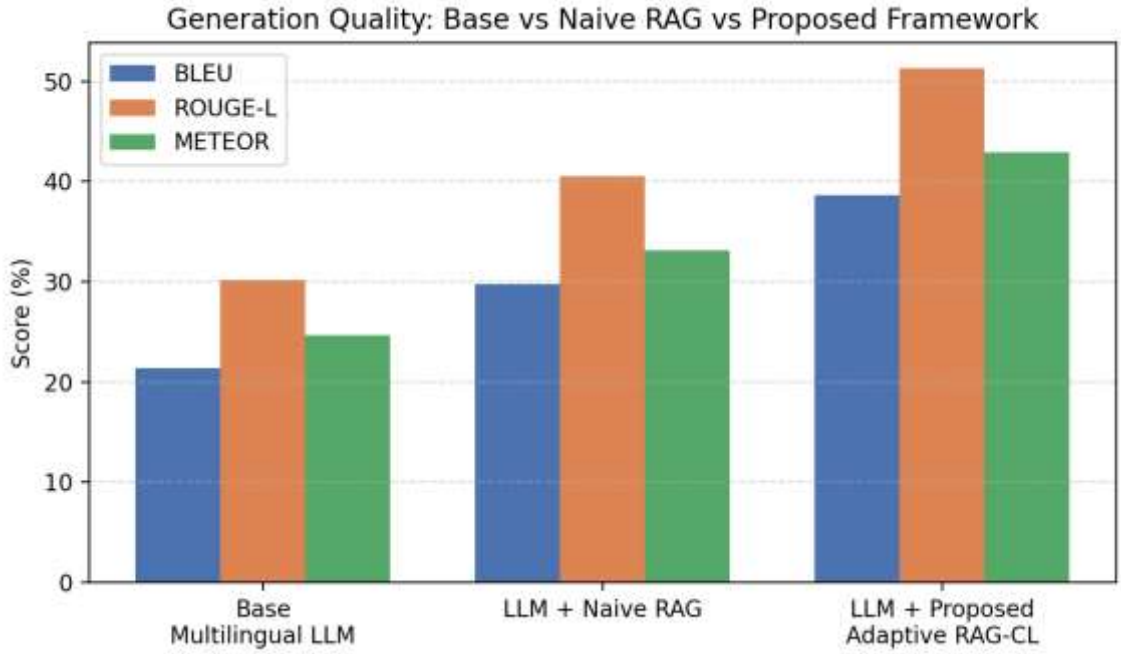


**Figure 3 BLEU, ROUGE-L, and METEOR scores across system configurations**

### 5.4.Effect of Continual Learning on Knowledge Base Expansion

The knowledge base was developed in five steps, each introducing a new document category, and response accuracy was tested on a fixed held-out query set after each increment to demonstrate adaptability. Figure 3 shows that the system with the proposed continual learning mechanism improves progressively with the addition of new knowledge, while a static RAG pipeline without continual learning shows degrading accuracy, presumably due to an increasing mismatch between the expanding retrieval index and the frozen generation model's adaptation to previous data.

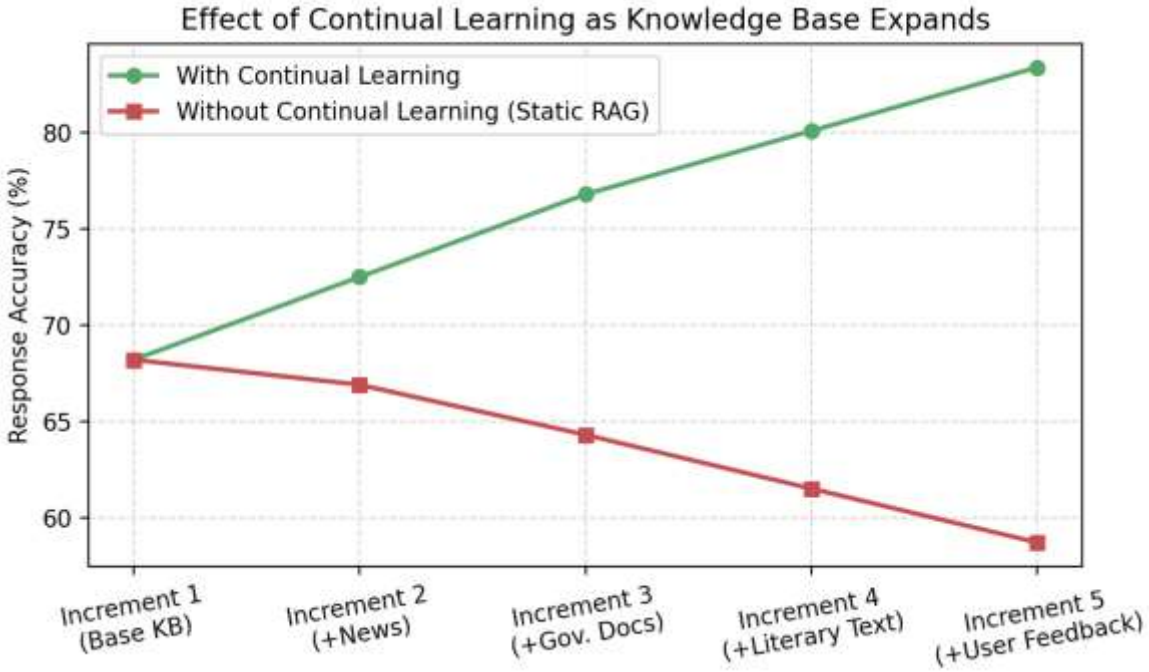


**Figure 4 Response accuracy as the knowledge base expands, with and without continual learning**

### 5.5.Ablation Study

An ablation study was conducted to isolate the contribution of individual components of the proposed framework. Table 4 shows that removing the cross-encoder re-ranking stage reduced retrieval accuracy and increased the hallucination rate, while removing the confidence-gating mechanism increased latency without a corresponding gain in accuracy, confirming that each component contributes measurably to overall performance.

**Table 4 Ablation study results**

| Configuration | Retrieval Accuracy | Generation BLEU | Hallucination Rate |
|---|---|---|---|
| Full proposed framework | 0.82 | 38.6 | 7.4% |
| Without re-ranking | 0.74 | 34.1 | 12.8% |
| Without confidence gating | 0.81 | 37.9 | 8.1% |
| Without continual learning | 0.82 | 31.5 | 15.6% |

### 5.6.Discussion

The results show that integrating retrieval augmentation with ongoing learning consistently outperforms both a non-retrieval baseline and a static RAG pipeline for the Mizo language. The biggest individual gain came from adding retrieval, corroborating previous conclusions from RAG literature that external grounding is a powerful means for hallucination reduction in knowledge-intensive generating. The additional benefits from constant learning, however lower in absolute terms, are crucial since they show that the framework can maintain accuracy as more Mizo content becomes accessible, without the need for full model retraining. Native-speaker evaluation also revealed that the outputs of the whole framework were regarded as more fluent and culturally suitable, indicating that the recovered literary and administrative background assisted the model in better fitting local register and vocabulary. The residual errors were largely concentrated on questions requiring thinking across several retrieved sections, hinting to a path for future refining of the retrieval and fusion technique

.

## 6. Conclusion & Future Scope

This research presents an adaptive knowledge-augmented framework for Mizo Large Language Models guided by Retrieval-Augmented Generation, and supplemented by a continual learning technique to overcome some issues in processing low-resource languages. The suggested approach relies on the interactions of semantic retrieval, contextual answer generation, and incremental knowledge updating that empowers the model to generate contextually appropriate answers grounded in facts. The experimental data provided the evidence of better retrieval results, better quality of text generation, and better human evaluation rates when compared to other multilingual LLMs and conventional RAG techniques. Furthermore, the method of continual learning makes it possible to realize a new Mizo knowledge and never retrain the model from scratch. This work opens up pathways for future research that can build on it and further enhance reasoning and response quality by integrating graph retrieval, multimodal information sources (speech, images), and reinforcement learning from human feedback. In addition, testing the framework on other low-resource Indian and tribal languages and using it in real-life applications like education, e-governance and digital cultural preservation would further establish its efficacy and generalizability.

**Reference**

1. Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023, July). Large language models struggle to learn long-tail knowledge. In *International conference on machine learning* (pp. 15696-15707). PMLR.
2. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2025). □ Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics*, *51*(4), 1373-1418.
3. Arora, D., Kini, A., Chowdhury, S. R., Natarajan, N., Sinha, G., & Sharma, A. (2023). Gar-meets-rag paradigm for zero-shot information retrieval. *arXiv preprint arXiv:2310.20158*.
4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
5. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020, July). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6282-6293).
6. Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
7. Wang, R., Long, N., Liu, S., Wang, Y., Qi, Z., & Zhang, H. (2026). Retrieval-Augmented Large Language Model Agents for Automated Scientific Literature Review Generation.
8. Zhang, Q., Chen, S., Bei, Y., Yuan, Z., Zhou, H., Hong, Z., ... & Huang, X. (2025). A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.
9. Wang, X., Sen, P., Li, R., & Yilmaz, E. (2025, April). Adaptive retrieval-augmented generation for conversational systems. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 491-503).
10. Mombaerts, L., Ding, T., Banerjee, A., Felice, F., Taws, J., & Borogovac, T. (2024). Meta knowledge for retrieval augmented large language models. *arXiv preprint arXiv:2408.09017*.
11. Tang, X., Li, J., Du, N., & Xie, S. (2024). Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs. *arXiv preprint arXiv:2412.07618*.
12. Yang, C., Xu, R., Luo, L., & Pan, S. (2024). Knowledge Graph and Large Language Model Co-learning via Structure-oriented Retrieval Augmented Generation. *IEEE Data Engineering Bulletin*.
13. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.