UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Artificial Intelligence and Kichwa Culture: Exploring Multilingual Learning Capabilities with Retrieval-Augmented Generation and LLaMA-3.1-8B Model in Kichwa Languages

Proyecto de Titulación

Kuntur Mallku Muenala Terán

Felipe Grijalva, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Inteligencia Artificial

Quito, 02 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Artificial Intelligence and Kichwa Culture: Exploring Multilingual Learning Capabilities with Retrieval-Augmented Generation and LLaMA-3.1 Model in Kichwa Languages

Kuntur Mallku Muenala Terán

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D. en Ingeniería Eléctrica
Director del programa de:	Inteligencia Artificial
Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías
Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, diciembre 2024

© DERECHOS DE AUTOR

Nombre del estudiante:	Kuntur Mallku Muenala Terán
Código de estudiante:	00339331
C.I.:	1003667209
Lugar y fecha:	Quito, 02 de diciembre de 2024.

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

DEDICATORIA

Dedico este trabajo de titulación a mi cultura Kichwa, que es tanto mi fuente de identidad como mi inspiración para soñar con un futuro mejor para el mundo. Este trabajo también honra el legado de mis ancestros, cuyo conocimiento y tradiciones han perdurado a través del tiempo. Asimismo, lo dedico a todas aquellas personas que se sientan abrazadas por la riqueza de las culturas indígenas de Latinoamérica.

AGRADECIMIENTOS

Agradezco profundamente el apoyo incondicional de mis padres y hermanos, con quienes hemos enfrentado juntos las dificultades de la vida. En especial, quiero destacar a mi padre, quien ha sido una guía invaluable al transmitirme las tradiciones y costumbres de nuestra cultura Kichwa, a la cual pertenecemos con orgullo.

RESUMEN

Kichwa, el idioma indígena más hablado entre los pueblos indígenas de Ecuador, enfrenta un riesgo crítico de extinción debido a factores como la falta de recursos digitales y la ausencia de iniciativas que promuevan su uso entre las generaciones jóvenes. Preservar este idioma es esencial, ya que encapsula la identidad cultural, las tradiciones y el patrimonio intangible de las comunidades indígenas, contribuyendo significativamente a la diversidad cultural del Ecuador y del mundo. Su revitalización no solo asegura su supervivencia, sino que también fortalece la conexión de las nuevas generaciones con sus raíces, fomentando un sentido de pertenencia. Las tecnologías como la inteligencia artificial ofrecen herramientas clave para la preservación del Kichwa, permitiendo la creación de recursos educativos, mejor acceso a servicios públicos y el fortalecimiento de derechos legislativos en contextos modernos. Este estudio analiza el uso de Retrieval-Augmented Generation (RAG) junto con el modelo LLaMA-3.1-8B como herramienta para preservar y revitalizar el idioma Kichwa. RAG mejora la precisión contextual de las respuestas al integrar documentos externos, sin requerir grandes recursos computacionales. A través de consultas específicas en Kichwa, se evaluó el desempeño del modelo en la generación de traducciones y respuestas contextuales utilizando métricas como BLEU (Bilingual Evaluation Understudy) y similitud semántica. Los resultados muestran que RAG reduce significativamente las alucinaciones típicas de los modelos de lenguaje grande (LLMs) y mejora la precisión de las respuestas. Sin embargo, la falta de datos digitalizados y las limitaciones computacionales restringen el alcance de los resultados. A pesar de esto, el estudio demuestra que RAG mejora significativamente el rendimiento del modelo LLaMA-3.1-8B; no obstante, esta mejora no es suficiente para que el modelo comprenda plenamente la estructura lingüística de un idioma nuevo que no formó parte de su entrenamiento inicial. Este trabajo destaca el potencial de RAG para ampliar las capacidades de los LLMs en lenguas con pocos recursos y subraya la importancia de colaborar con comunidades hablantes de Kichwa para garantizar el desarrollo de tecnologías culturalmente responsables.

Palabras clave: RAG (Retrieval-Augmented Generation), LLaMA (Large Language Model Meta AI), Embedding, Kichwa, NLP (Natural Language Processing), LLM (Large Language Model), BLEU (Bilingual Evaluation Understudy), Similaridad Semantica

ABSTRACT

Kichwa, the most widely spoken indigenous language among Ecuador's indigenous peoples, faces a critical risk of extinction due to factors such as lack of digital resources and lack of initiatives to promote its use among younger generations. Preserving this language is essential, as it encapsulates the cultural identity, traditions, and intangible heritage of indigenous communities, significantly contributing to Ecuador's and the world's cultural diversity. Its revitalization not only ensures its survival, but also strengthens the connection of younger generations to their roots, fostering a sense of belonging. Technologies such as artificial intelligence offer key tools for the preservation of Kichwa, allowing the creation of educational resources, improved access to public services, and the strengthening of legislative rights in modern contexts. This study examines the use of Retrieval-Augmented Generation (RAG) alongside the LLaMA-3.1-8B model as a tool for preserving and revitalizing the Kichwa language. RAG improves the contextual accuracy of responses by integrating external documents, all without requiring significant computational resources. Through Kichwa-specific queries, the study evaluated the model's performance in generating translations and contextual responses using metrics such as BLEU (Bilingual Evaluation Understudy) and semantic similarity. The results show that RAG significantly reduces the typical hallucinations of Large Language Models (LLMs) and improves the precision of the response. However, the lack of digitized data and computational limitations constrain the scope of the results. Despite this, the study demonstrates that RAG significantly enhances the performance of the LLaMA-3.1-8B model; even so, this improvement is not sufficient for the model to fully understand the linguistic structure of a new language that was not included in its initial training. This work highlights the potential of RAG to expand LLM capabilities for low-resource languages and emphasizes the importance of working with Kichwa-speaking communities to ensure the development of culturally responsible technologies.

Key words: RAG (Retrieval-Augmented Generation), LLaMA (Large Language Model Meta AI), Embedding, Kichwa, NLP (Natural Language Processing), LLM (Large Language Model), BLEU (Bilingual Evaluation Understudy), Semantic Similarity

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Artificial Intelligence and Kichwa Culture: Exploring Multilingual Learning Capabilities with Retrieval-Augmented Generation and LLaMA-3.1 Model in Kichwa Languages

Kuntur Muenala, *Member, IEEE,*
Felipe Grijalva, *Senior Member, IEEE*

***Abstract*—Kichwa, the most widely spoken indigenous language among Ecuador's indigenous peoples, faces a critical risk of extinction due to factors such as lack of digital resources and the lack of initiatives to promote its use among younger generations. Preserving this language is essential, as it encapsulates the cultural identity, traditions, and intangible heritage of indigenous communities, significantly contributing to Ecuador's and the world's cultural diversity. Its revitalization not only ensures its survival, but also strengthens the connection of younger generations to their roots, fostering a sense of belonging. Technologies such as artificial intelligence offer key tools for the preservation of Kichwa, allowing the creation of educational resources, improved access to public services, and the strengthening of legislative rights in modern contexts. This study examines the use of Retrieval-Augmented Generation (RAG) alongside the LLaMA-3.1-8B model as a tool for preserving and revitalizing the Kichwa language. RAG improves the contextual accuracy of responses by integrating external documents, all without requiring significant computational resources. Through Kichwa-specific queries, the study evaluated the model's performance in generating translations and contextual responses using metrics such as BLEU (Bilingual Evaluation Understudy) and semantic similarity. The results show that RAG significantly reduces the typical hallucinations of Large Language Models (LLMs) and improves the precision of the response. However, the lack of digitized data and computational limitations constrain the scope of the results. Despite this, the study demonstrates that RAG significantly enhances the performance of the LLaMA-3.1-8B model; even so, this improvement is not sufficient for the model to fully understand the linguistic structure of a new language that was not included in its initial training. This work highlights the potential of RAG to expand LLM capabilities for low-resource languages and emphasizes the importance of working with Kichwa-speaking communities to ensure the development of culturally responsible technologies.**

***Index Terms*—RAG (Retrieval-Augmented Generation), LLaMA (Large Language Model Meta AI), Embedding, Kichwa, NLP (Natural Language Processing), LLM (Large Language Model), BLEU (Bilingual Evaluation Understudy), Semantic Similarity.**

I. Introduction

F. Grijalva and K. Muenala are with Universidad San Francisco de Quito USFQ

LANGUAGE is a fundamental pillar for preserving the customs, history and heritage of a culture. The preservation of a language ensures that its culture remains alive and retains its identity. In Ecuador, there are 18 Indigenous peoples and 14 indigenous nationalities that speak 14 native languages; however, 8 of these languages are at risk of disappearing [1], [2]. Although the reasons behind the disappearance of these languages are diverse and will not be discussed in this study, the recovery of endangered languages is crucial to preserving cultural identity in an increasingly globalized world. Artificial intelligence (AI) offers a unique opportunity to support the preservation and recovery of languages at risk of extinction. As a tool, AI can transcribe and translate oral histories into various languages, allowing the conservation of valuable cultural knowledge that might otherwise be lost. In addition, AI-based technologies such as voice recognition, text analysis, and image and audio processing can play a key role in safeguarding and revitalizing linguistic heritage [3].

AI also has the potential to renew and preserve languages by analyzing linguistic material, such as written data, historical files, and related languages, to revive extinct languages. This not only helps to recover languages lost for generations, but also allows individuals to reconnect linguistically [4]. In addition, AI can help design study materials for revitalized languages, generating grammatical rules, vocabulary, and even audio recordings from modern sources and cultural artifacts. These tools can facilitate the learning and practice of languages in daily speech, promoting their active use. However, although the promise of AI in language revitalization is exciting, significant challenges remain, such as the need for detailed linguistic data, ethical concerns about AI-generated material, and biases in language processing systems. Despite these limitations, recent studies indicate that AI has a promising future in language preservation and revitalization, especially when combined responsibly and collaboratively with language communities [3]. In this way, AI positions itself as a fundamental technological foundation for protecting endangered languages and cultures.

A. Brief historical overview of lenguage Kichwa

Kichwa is a native language of indigenous peoples originating in South America, specifically in the country we now know as Ecuador. Kichwa is the indigenous language with the highest number of speakers in Ecuador [5]. Before the arrival of the Spanish in the Americas, this region was known as Abya Yala. Following colonization by the Spanish Empire, the region of Abya Yala was renamed "America" [6]. Kichwa has been present in Ecuador long before the Spanish. Initially, it was a prestigious language associated with the leadership of the nationalities in what is now Ecuador. Kichwa has adapted to the changes experienced by the people of the region, from the arrival of the Incas to the Spanish. The arrival of Spanish brought many changes to indigenous peoples, both in their culture and their languages, including the imposition of Spanish as the primary language and the introduction of new beliefs and cultural changes. One of these changes was the prohibition of the Kichwa language. However, studies of this period of imposition reveal that indigenous peoples resisted and protected their cultural heritage [6], [5]. They defended themselves against colonial domination by adapting their culture to the new colonial norms. Unfortunately, one of the lasting effects of colonialism is the reduction in the number of Kichwa speakers [5]. Today, younger generations of Kichwa face multiple contemporary challenges in preserving their cultural and linguistic heritage while striving to develop it within a modern context.

B. The State of Kichwa in Ecuador

Although Kichwa is Ecuador's second official language and the most representative in terms of linguistic diversity after Spanish, it faces the risk of extinction due to multiple factors. Among the main causes are poverty, oppression by the dominant majority culture, rural-to-urban migration, and the dynamics of the educational system [7]. Furthermore, the lack of projects actively promoting the Kichwa language contributes to this situation.

In 2010, Ecuador had a population of 14,459,077 inhabitants, of which approximately 4.5% (654,316 people) spoke the Kichwa language, according to data from the 2010 Census. However, about 1.3 million people, equivalent to 9% of the population, identified as belonging to indigenous groups in the same census [7], [8]. By comparing these results with the most recent data from the 2022 Census, Ecuador recorded a population of 16,938,986 inhabitants, of which 7.7% (1,304,302 people) identified themselves as indigenous, while only 3.95% (669,090 people) spoke the Kichwa language [9]. This reveals that the number of Kichwa speakers is smaller than the number of people identified as indigenous. In addition, the data reflect a decrease in Kichwa speakers over time.

Luis Fernando Garcés Velásquez, in his study *"Las comunidades virtuales del quichua ecuatoriano: revalorizando la lengua en un espacio apropiado"*, examines this progressive loss of language through various reports [7]. The decline in the number of Kichwa speakers represents a significant loss to the cultural diversity of Ecuador and the world. This occurs even in a context where indigenous rights have improved over the past decade, highlighting the need for initiatives to encourage cultural and linguistic engagement among young people with their indigenous heritage.

C. Importance of preservation and revival the Kichwa

Understanding the importance of preserving a language is vital for the development of initiatives that promote its use and prevent its extinction. Language preservation is a cornerstone of safeguarding cultural diversity worldwide, as languages encapsulate the collective knowledge and identity of entire communities [10], [3]. Thus, preserving a language is inherently related to protecting the core of a culture's identity and heritage. When a language is considered endangered, it faces the imminent threat of completely disappearing from the human experience. To address this challenge, language preservation projects are indispensable. These initiatives aim to document, protect, and revitalize languages, ensuring their survival. The extinction of an endangered language results in the irreparable loss of unique ways to express ideas, identities, and cultural legacy. Consequently, language preservation efforts play a critical role in preventing these losses and protecting cultural heritage. By their very nature, language preservation initiatives act as guardians of the cultural legacy [3].

To ensure the survival of a culture, it must be open to changes, and the Kichwa language is no exception. Over centuries, Kichwa has undergone profound transformations, evolving from its precolonial roots to incorporate adaptations such as the Spanish alphabet. Since 1980, grammatical rules have been developed to unify the written form of Kichwa in its various dialects, which has allowed the language to endure [11], [10]. This standardization has allowed the creation of texts in Kichwa, ranging from traditional stories to official documents such as the Ecuador constitution. Despite these advances, texts in Kichwa remain limited, and efforts to promote its written form and foster linguistic development continue to face challenges. Preserving and reviving Kichwa requires understanding it not merely as a tool for communication, but as a vessel of timeless knowledge that safeguards cultural practices and heritage through its written form, while also reconnecting communities with their linguistic roots, renewing traditions, and fostering unique ways of thinking [11], [10]. This dual effort promotes a sense of identification, belonging, pride, and empowerment within the community.

D. Tools for Linguistic Diversity Preservation

Currently, artificial intelligence (AI) has an immense potential to help in the preservation of endangered languages by developing tools that facilitate their use, documentation, and storage. Natural Language Processing (NLP) technologies play a pivotal role in this effort, enabling efficient

text analysis, processing, and generation. In Ecuador, AI can play a crucial role in projects aimed at revitalizing and preserving the Kichwa language by documenting, translating, transcribing and teaching it [3], [4]. These technological capabilities are essential for safeguarding the rich linguistic diversity of the country.

Below are some AI tools and their applications in language preservation:

- **Automated Transcription:** AI algorithms can convert spoken language into text in real time, facilitating the transcription of oral narratives, stories, and conversations for analysis and documentation.
- **Text Analysis:** These algorithms can process large volumes of text, such as historical documents, to uncover complex linguistic structures and support linguistic researchers.
- **Image Processing and Text Extraction:** AI can extract text from images, such as handwritten manuscripts, making it easier to transcribe and preserve their content.
- **Data Organization:** AI tools enable the efficient structuring of linguistic databases, organizing information from various languages.
- **Generative Algorithms:** These models can generate audio, text, or images interactively and individually, offering new ways of engaging with languages.

In this article, we focus on exploring natural language processing (NLP) models, particularly LLaMA-3, an advanced open-access model. Large Language Models (LLMs) are advanced artificial intelligence models based on neural network architectures such as Transformers. These models are trained on massive amounts of textual data and are designed to perform a wide range of natural language processing (NLP) tasks [12].

LLMs (Large Language Models) have demonstrated exceptional abilities in solving complex tasks that smaller models cannot handle, marking a milestone in technological advancement and significantly transforming the development and application of artificial intelligence algorithms. Although these models are pre-trained on massive amounts of data, they are not capable of addressing all challenges in text processing. Specifically, when working with a language that was not included in their initial training, additional pre-training is required to adapt them [13]. However, training such models requires significant computational resources that are not always readily available.

To address this challenge, we propose the use of a technique called Retrieval-Augmented Generation (RAG), which enables fine-tuning model responses without requiring extensive computational resources. This technique is particularly advantageous in resource-limited settings, as it integrates external databases with the model's capabilities [14]. In this work, we employ RAG to adapt an LLM to a textual domain in Kichwa, allowing the model to understand and generate responses in this language. The following section presents a review of NLP models that have proven effective in the preservation of endangered languages.

II. State of the Art

There are several studies focused on using LLMs (Large Language Models) to preserve and revitalize endangered native languages worldwide. This section examines similar work investigating how artificial intelligence can aid these efforts.

A. Indigenous Languages of Brazil

In Brazil, the University of São Paulo, in collaboration with IBM researchers, explored how artificial intelligence technologies, particularly LLMs, can be used to promote the use and preservation of endangered indigenous languages [4]. The objective was to increase the representation and accessibility of these languages in modern technologies.

The study applied fine-tuning to machine translation models (MTs) using limited data from the indigenous languages under study. The models used were mBART50 (680M parameters) and WMT19 (315M parameters), adapted for 39 and 10 languages, respectively. Each model was evaluated using the BLEU metric to measure the semantic similarity between translated words. In addition, prototypes of specific tools were developed for indigenous languages to facilitate writing and language development within communities [4].

This work showed promising results, producing high-quality automatic translators and creating tools such as spell checkers and word predictors. These innovations significantly reinforced the practical and cultural use of indigenous languages in Brazil.

B. Some Native Languages

In the United States, researchers at the University of Georgia conducted a study aimed at overcoming linguistic barriers that hinder cultural preservation and the development of minority communities through LLM [14]. This research focused on three native languages:

- Cherokee: a Native American language.
- Tibetan: a language of Asian culture.
- Manchu: an ancestral language of China.

The study implemented an innovative information retrieval approach to improve translation in low-resource languages. Usign Retrieval-Augmented Generation (RAG). This method focused on retrieving key terms and using existing examples to provide more accurate translations. The performance of the GPT-4 and LLaMA 3.1 models (405B parameters) was evaluated in translating English into these languages [14].

The evaluations included metrics such as BLEU, ROUGE, BERTScore, and human evaluation, analyzing fluency, grammaticality, and fidelity. The initial results showed low

metrics due to the lack of fine-tuning with language-specific data, attributed to the scarcity of available texts. However, integrating RAG significantly improved translation accuracy [14]. The study concludes with a recommendation for further exploration of methods that respect the didactic structure of each language, with the aim of avoiding overly simplistic or flat translations.

C. Research in Ecuador

In Ecuador, researchers at Universidad San Francisco de Quito conducted a study focused on Kichwa digital inclusion using advanced technologies such as LLMs. The objective was to preserve the language through advanced technological tools.

The study collected Kichwa documents, such as dictionaries and fragmented texts, to process and train the LLaMA-2 model (70B parameters) using Low-Rank Adaptation (LoRA) techniques. The developed model, named URKU, was evaluated using the SISA Benchmark, designed to measure its ability to process and generate text in Kichwa. The same test was used to evaluate GPT-Builder, where URKU achieved a score of 0.867, compared to GPT-Builder 0.95 [15]. These results demonstrate that URKU is promising, showing efficient performance even with limited computational resources.

However, the study highlights the need for more digital resources to continue to advance the development of tools for the Kichwa language. In addition, hardware limitations were identified as a barrier to deeper fine-tuning. The research leaves the door open for future studies to develop a fully functional LLM in Kichwa and evaluate it with native speakers [15].

D. Key insights

One of the main challenges faced by research on endangered languages is the scarcity of digitized data, which is essential for training artificial intelligence models. Another critical factor is the limitation of computational resources, as accessing equipment capable of performing exhaustive training with new data is both challenging and costly. Therefore, these factors are crucial when deciding which models and strategies to implement in projects involving endangered languages. Furthermore, it is essential to emphasize that the mentioned studies agree on the importance of including the speaker communities of these languages in the development of these technologies, allowing the models to be evaluated in terms of fluency and clarity from the perspective of native speakers.

III. Materials and Methods

A. Data collection

For this work, various digitized documents in PDF format containing Kichwa content were collected, including dictionaries, poems, stories, educational materials, research articles and the Constitution of Ecuador. All of these documents are in the Kichwa language. However, they were categorized into groups based on their content.

1) **Complete**: Documents entirely written in Kichwa, such as thesis projects or the full Constitution in Kichwa.
2) **Medium:** Bilingual documents that present information in Kichwa along with translations into Spanish or English, such as poems, stories, or transcriptions.
3) **Basic:** Educational materials where most of the text is in Spanish, with small phrases or instructions in Kichwa.
4) **Dictionaries:** Various dictionaries translating Kichwa into Spanish and Kichwa into English.
5) **Spanish Information:** Research papers on the Kichwa language and culture written in Spanish. These were deemed important for evaluating the models within a linguistic and cultural context of the Kichwa language.

In total, approximately 130 documents were collected. However, 70% of them fall into the "Medium" and "Basic" groups, highlighting the same issue mentioned in previous research: the scarcity of digitized information for the development of artificial intelligence models. LLMs require vast amounts of data for effective fine-tuning. For this reason, this work opted not to perform fine-tuning on the models and instead used the RAG (Retrieval-Augmented Generation) technique for the development of the LLM.

B. Processing Data

Since most of the collected documentation falls into the *Basic* and *Medium* groups in terms of Kichwa content, data cleaning was performed based on the following considerations.

- Remove paragraphs or sentences with excessive special characters: This issue arises because the graphics in the documents, when digitized, are converted into special characters that do not add value to the linguistic content.
- Remove excess blank spaces in the document: This practice is particularly applicable to educational materials, as many of these documents include exercises with blank lines.

These practices are essential when applying the RAG (Retrieval-Augmented Generation) technique to the LLaMA 3.1-8B model, as one of the steps involves forming fragments of informative content. For LLM to effectively understand and process these fragments, the information must be clean and well structured.

C. Evaluation test

For the evaluation, 30 questions were created: 10 in Spanish, 10 in English, and 10 in Kichwa, each different within their respective groups. However, all questions are related to the linguistic context of Kichwa or queries in this language.

The decision to include questions in three languages was based on the availability of the document collection, which contains materials in Spanish, English, and Kichwa.

This work used the BLEU metric to evaluate the quality of the responses generated by the LLaMA 3.1-8B model compared to the correct answers. In addition, the semantic relationship between the generated responses and the real responses was assessed to measure the similarity between them. The scoring results are presented for the 10 questions in each language, providing an overall evaluation of the model's performance.

D. Retrieval Augmented Generation - RAG

Retrieval-Augmented Generation (RAG) is a technique used in LLMs. This method combines text generation with the retrieval of relevant information from large external databases. The goal of RAG is to enhance the accuracy of LLM-generated responses by accessing specific information from a set of documents before generating a response. By integrating external data, RAG effectively reduces hallucinations that LLMs can generate when they lack the answer or work with outdated information [16], [17].

The RAG implementation system is illustrated in Figure 1, showing how query documents (stored vectors) are combined with the input query to allow the LLM to generate text based on the retrieved queries in the RAG system [18], [17]. This implementation of the RAG system consists of three parts: Indexing, Retrieval, and Generation.

1) **Indexing** aims to transform raw data into a vectorized database to enable efficient and accurate retrieval. This process begins by converting external data, whether structured or unstructured, into a standardized format. Then, an embedding model encodes the processed data into smaller chunks, which are stored in the vector database. This step is crucial to ensure optimal information retrieval [14]. The process begins with downloading the data, which will be transformed into vector storage that the LLM will query before generating responses. The texts are divided into fragments called chunks and the size of these chunks is critical: if the chunks are too small, the model lacks sufficient interpretability of the content of the documents. In contrast, if the fragments are too large, the model cannot effectively filter the relevant information that the system attempts to retrieve from the document database [18], [17]. Subsequently, these fragments are converted into embeddings which are vector representations of the data. These embeddings form a query library that the model uses to generate responses. In the end, a vector storage is obtained, representing the processed fragments from the query database [16], [14].
2) The **Retrieval** process in a RAG system involves searching for and expanding the user's initial query using specialized algorithms. At this stage, the user query is encoded in an input vector using the same embedding model employed during indexing. The system then calculates the similarity between this input vector and the stored fragments, selecting the K most relevant fragments based on their proximity [14], [17]. The RAG system uses an indexed embedding store to identify the vectors most similar to the query. When a query is received, the system searches for the closest vectors in the vector space, allowing the LLM to generate more accurate responses. At this stage, the number of vectors closest to retrieval can be configured to optimize the response generation process [16], [14].
3) The **Generation** phase uses an enriched message synthesized by combining the user query with the retrieved documents. This message, which can be represented as a structured set of data, provides the LLM with additional contextual information, helping to reduce hallucinations and constrain generated responses to relevant and accurate content. As a result, this phase significantly improves the precision and reliability of the model output [14]. At this stage, the RAG system directly interacts with the LLM to process the user's query. A prompt is utilized, which can be generated manually or imported. This prompt creates a structure known as the *chain RAG*, ensuring that the query first passes through the RAG system, where it is enriched with retrieved information, before being processed by the LLM and generating a response [16], [14], [17].

Advanced RAG methods have been developed to address the limitations of traditional RAG by optimizing various components, such as query messages [14], indexing structures, similarity calculations, and message integration. In addition, some approaches leverage the data retrieved to improve training datasets, particularly in low-resource domains. RALMs (Retrieval-Augmented Language Models) systems have proven effective in improving the performance of LLMs [19]. These advanced RAG systems can integrate critical information during the training phase, making them especially valuable for developing natural language processing (NLP) models with limited data resources.

E. Hardware and Software Specifications

The computational experiments were carried out using high-performance hardware and software setups to accommodate the size of the data set and the complexity of the model.

- **Hardware:** NVIDIA A100 GPU with 80 GB of VRAM.
- **Software:** Frameworks:
  - langchain: 0.3.7
  - langchain_core: 0.3.15
  - langchain_unstructured: 0.1.5

F. Framework - LangChain

LangChain is a framework designed to facilitate the development of applications powered by large language

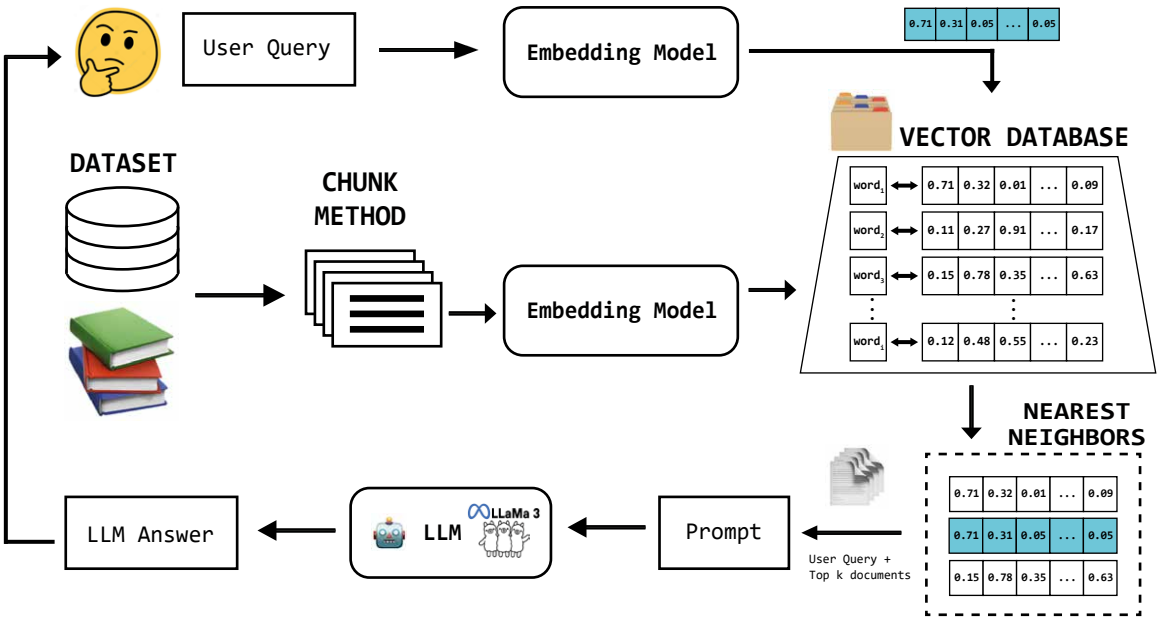

Figure 1. Workflow of a Retrieval-Augmented Generation (RAG) system: The dataset is divided into fragments (chunks) using splitting techniques and then transformed into vector representations through an embedding model. These vectors are stored in a vector database. When a user submits a query, it is encoded into a vector representation using the same embedding model. The system then retrieves the most relevant documents (nearest neighbors) based on the user's query, combines them into a prompt, and feeds this prompt into a large language model (LLM) to generate an accurate and contextually enriched response.

models (LLMs), enabling efficient integration with external data sources, tools, and complex workflows. This framework focuses on two fundamental concepts: language models as the core processing unit and integration with external data. Using LangChain, developers can build task pipelines that include information retrieval, dynamic prompt generation, and the orchestration of queries in vector databases [20]. Many of the tools offered by LangChain are utilized in the implementation of the RAG system developed in this study.

In the case of RAG, LangChain enables the integration of LLMs with information retrieval systems, such as vector databases, to enrich the content generated by the models. LangChain simplifies this architecture by providing optimized modules for creating chains that interact with tools such as FAISS, Pinecone, or Weaviate, which are specialized libraries for generating indexed vector databases. In addition, it offers intuitive interfaces to organize prompts and manage interactions between system components [21], [18]. Due to its flexibility and compatibility, LangChain was chosen as the framework for this study.

G. Methodology

When downloading the database, each document is divided into fragments (chunks) of 300 characters in length, resulting in approximately 80,000 fragments in total. This creates a vector database derived from the 130 documents in Kichwa. For vectorization, the pre-trained embedding model sentence-transformers/LaBSE was used, capable of mapping 109 languages into a shared vector space [22]. This model generates embeddings of 256 dimensions, making it suitable for processing documents in Kichwa, Spanish, and English within the dataset.

The distribution of the fragments, including their character lengths and vector dimensions, is shown in Figure 2. Most Kichwa documents have chunks of approximately 50 characters. This small size may be attributed to the majority of documents falling into the *basic* and *medium* groups, as previously mentioned. Consequently, the model might face limitations in interpreting the Kichwa language due to the reduced amount of information in the fragments.

Once the embeddings are generated, they are indexed using the FAISS library, which is specialized in indexing LLMs. After indexing, a vector database is created for queries from the RAG system. During this process, the number of nearest fragments to retrieve is defined for each query. For this study, 40 fragments were selected due to the small size of the chunks.

At this stage, it is also possible to visualize how the fragments (chunks) interact with the queries submitted to the LLM. Dimensionality reduction is performed, reducing the original 256 dimensions of the embeddings to 2 dimensions, making the fragments easier to visualize. This reduction is achieved using the PaCMAP model (Pairwise Controlled Manifold Approximation) applied to

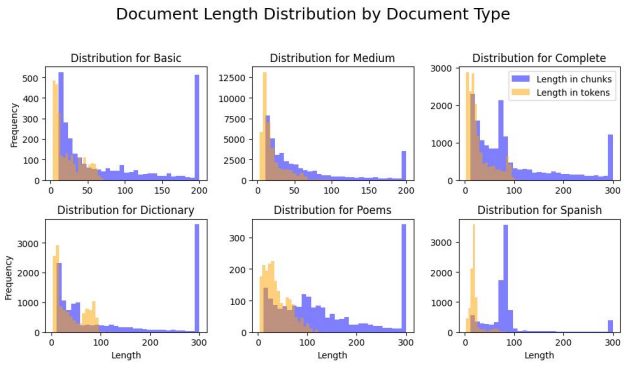

Figure 2. Distribution of chunk sizes, their character lengths, and tokenized vector dimensions.

the generated embeddings, PaCMAP is a dimensionality reduction method that can be used for visualization, preserving both local and global structure of the data in original space [23]. The results are shown in Figure 3. In the scatter plot, the generated queries are embedded within the set of embedded files from the Kichwa documents. It can be observed that queries in Spanish tend to cluster with Spanish fragments, demonstrating the RAG system's ability to retrieve relevant information based on semantic similarity.

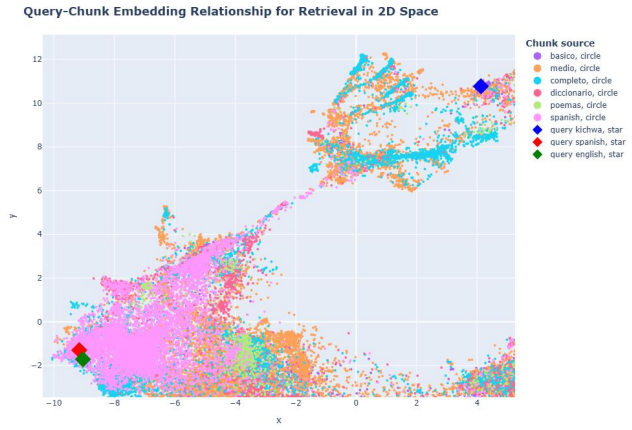

Figure 3. The scatter plot shows two chunks and three queries. Each query is surrounded by several relevant fragments from the documentation. The system selects the fragments closest to these queries based on the semantic similarity between each query and the fragments.

Once the indexing database is created and the number of fragments to retrieve is configured, the Meta-Llama-3.1-8B model [24] is used to generate text from queries.

- Model LLaMA 3.1-8B Parameters
  - task="text-generation",
  - temperature=0.1,
  - do_sample=True,
  - repetition_penalty=1.1,
  - return_full_text=True,
  - max_new_tokens=50,

This process begins by creating a prompt that includes tags to identify the context, the query, and the generated response. The prompt also incorporates information about the documents supporting the response, represented by the closest retrieved embeddings. To integrate these steps, a structure called Chain RAG is used, enabling the RAG system to work seamlessly with the LLM through the LangChain framework.

The full development of this work can be found in the GitHub repository.

GitHub Link

IV. Results and Discussion

A. Evaluation

To compare the performance of the Meta-Llama-3.1-8B model without RAG and with RAG, the BLEU (Bilingual Evaluation Understudy) and Semantic Similarity metrics were used. These metrics evaluate the closeness between the responses generated by the models and the correct responses. Originally, these metrics were designed to assess the quality of translations, but in this case they are applied to determine which model aligns more closely with the correct responses.

A set of questions specifically designed for this task was used for the evaluation. The results obtained using the BLEU metric are presented in Table I.

Table I
BLEU evaluation

BLEU Metric	Test Spa	Test Eng	Test Kich
LLM	1.89e-156	1.13e-156	0.0035
LLM - RAG	3.49e-156	1.13e-155	0.027

The results evaluated using the Semantic Similarity metric are shown in Table II.

Table II
Semantic Similarity evaluation

Semantic Similarity	Test Spa	Test Eng	Test Kich
LLM	0.192	0.167	0.297
LLM - RAG	0.382	0.328	0.436

B. Discussion

- In the evaluation tables I y II, a slight improvement is observed in the responses generated by the RAG model. Similarly to the study conducted at the University of Georgia in Cherokee, Tibetan, and Manchu languages, the implementation of the RAG system was not sufficient to produce completely reliable answers. However, the model with RAG was shown to be more accurate than the model without RAG, indicating that

the results of this study are consistent with previous research.

- In this work, the LLaMA 3.1-8B model was used, chosen due to the computational limitations available. However, in the study *Transcending Language Boundaries: Harnessing LLMs for Low-Resource Language Translation*, the LLaMA 3.1-405B model (405 billion parameters) was used, which is significantly larger than the one employed here. Despite this, the results are similar: the model with RAG is more accurate than the model without RAG, although both studies reported difficulties in reliably translating lengthy texts.
- RAG is a technique that improves the accuracy of LLM by retrieving relevant information, although it does not perform fine-tuning on models. The decision to implement RAG without fine-tuning was motivated by the need for methods that require fewer computational resources. Although fine-tuning demands significant processing power, the RAG system primarily uses storage to manage the vector query database.
- The test set included translation instructions ranging from individual words to full sentences. Although both models overall scored low, the model with RAG successfully answered the questions about translating individual words. However, when faced with longer translations, it either stated that it lacked the necessary context or repeated the question, unlike the model without RAG, which generated fabricated translations for both single words and longer sentences.
- The model with RAG outperformed the model without RAG, although its responses were not entirely robust. The RAG system demonstrated more accurate word usage, while the model without RAG tended to fabricate answers for all questions.
- There are many techniques derived from RAG. If RAG is used as an information retrieval tool during the training phase, it can reduce the amount of data needed to train a model in a low-resource language. This approach allows the model to adapt better to the target language, which in this case is Kichwa.
- It is essential to highlight the importance of collecting data in Kichwa within this study. This effort is crucial not only for this research but also for enabling future studies to use advanced technologies in developing solutions that support the preservation of the Kichwa language, which is currently at risk of extinction.
- It is important to note that computational resources were a significant limitation in this study, particularly with regard to RAM usage. One of the main challenges arose when selecting the number of fragments to retrieve. If more than 40 documents were chosen, the RAM would reach its maximum capacity, causing the kernel to reset due to insufficient memory.

V. Conclusion

The implementation of RAG in the LLaMA 3.1-8B model was not sufficient to generate fully robust responses for translation and text generation tasks. However, for specific questions, such as direct word translation, the model with RAG outperformed the model without RAG. This suggests that RAG improves response accuracy, although it does not guarantee a deep understanding of the language. These results align with the conclusions from the study presented in the paper "Transcending Language Boundaries: Harnessing LLMs for Low-Resource Language Translation", where similar limitations were observed even with the significantly larger LLaMA 3.1-405B model. This reinforces the notion that improving the understanding of new languages in LLMs requires prior training rather than relying solely on optimization techniques to enhance response accuracy.

A. Future Work

To achieve a better understanding of the Kichwa language within the model, it would be necessary to explore techniques such as fine-tuning with LoRA (Low Rank Adaptation) [15] or RALMs (Retrieval-Augmented Language Models), which integrate RAG into the training phase. These techniques could optimize the use of computational resources and maximize the utility of the limited linguistic data available for this language.

Additionally, the field of artificial intelligence presents numerous techniques to be explored, particularly to address the challenges posed by low-resource languages or computational limitations. Resource optimization remains an open topic with significant potential, especially in the development of applications for endangered languages.

Finally, it is crucial to involve the communities of endangered languages in the development of these technologies. This would ensure that translations or interpretations are not flat and incorporate the cultural context and traditions of indigenous peoples, respecting their vision and values.

References

[1] Confederación de Nacionalidades Indígenas del Ecuador (CONAIE). (2022, April) Lenguas indígenas de ecuador y el mundo. Accessed: December 4, 2024. [Online]. Available: https://conaie.org/2022/04/25/lenguas-indigenas-de-ecuador-y-el-mundo/

[2] Instituto Nacional de Estadística y Censos (INEC). (2022, December) Boletín de prensa. Accessed on November 26, 2024. [Online]. Available: https://www.ecuadorencifras.gob.ec/boletin-prensa-ferias-autoidentificacion/

[3] Gupta C. and Sharma A., "Reviving Indigenous Languages Using Machine Learning," *Insights2Techinfo*, p. 1, 2024, accessed on November 26, 2024. [Online]. Available: https://insights2techinfo.com/reviving-indigenous-languages-using-machine-learning/#google_vignette

[4] C. Pinhanez, P. Cavalin, L. Storto, T. Finbow, A. Cobbinah, J. Nogima, M. Vasconcelos, P. Domingues, P. de Souza Mizukami, N. Grell, M. Gongora, and I. Gonçalves, "Harnessing the power of artificial intelligence to vitalize endangered indigenous languages: Technologies and experiences," *arXiv*, vol. cs.CL, no. 2407.12620v2, July 2024, accessed: November 28, 2024. [Online]. Available: https://arxiv.org/pdf/2407.12620

[5] A. Egas, "LA INFLUENCIA DEL CONTEXTO SOCIAL DIGLÓSICO Y LA ACULTURACIÓN EN LA AUTOESTIMA LINGÙÌSTICA DE LOS JÓVENES BILINGÜES KICHWA-CASTELLANO," 2017, Accessed on November 30, 2024. [Online]. Available: https://repositorio.puce.edu.ec/items/90727c7f-d992-4db8-96c2-481eb033f659

[6] W. Ariruma and K. Maldonado, "(in) visibilización del kichwa: políticas lingüísticas en el ecuador," Ph.D. dissertation, Universidad Andina Simón Bolívar, 2013, accessed on November 30, 2024. [Online]. Available: https://repositorio.uasb.edu.ec/handle/10644/3827

[7] L. Garcés, "Las comunidades virtuales del quichua ecuatoriano: revalorizando la lengua en un espacio apropiado," *Tellus*, vol. 20, no. 43, pp. 59–75, setiembre/diciembre 2020. [Online]. Available: http://dx.doi.org/10.20435/tellus.vi43.760

[8] A. Kowii, "Runa shimi, kichwa shimi wiñaymanta," *Americanía. Revista de Estudios Latinoamericanos*, vol. Nueva Época, no. Número Especial, pp. 157–176, November 2017, iSSN-e 2174-0178. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=6007924

[9] Instituto Nacional de Estadística y Censos (INEC). (2022) Censo de Población y Vivienda 2022: Presentación Nacional - Segunda Entrega. Accessed: November 30, 2024. [Online]. Available: https://www.censoecuador.gob.ec/wp-content/uploads/2024/05/Presentacion_Nacional_2da_entrega.pdf

[10] B. Guzmán, M. Manzano, C. Domínguez, and M. Aroca, "Al rescate de la identidad sociolingüística de la lengua kichwa en la provincia bolívar: Necesidad y gestión," *Universidad y Ciencia*, vol. 7, no. 1, pp. 144–155, December–March 2018, accessed on November 31, 2024. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=8315386

[11] R. Moya, "Educación bilingüe en el ecuador: Retos y alternativas," *Indiana*, vol. 11, pp. 387–406, 1987.

[12] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023, accessed on November 30, 2024.

[13] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi, "Siren's song in the ai ocean: A survey on hallucination in large language models," *arXiv preprint arXiv:2309.01219*, 2023, accessed on December 1, 2024. [Online]. Available: https://arxiv.org/abs/2309.01219

[14] P. Shu, J. Chen, Z. Liu, H. Wang, Z. Wu, T. Zhong, Y. Li, H. Zhao, H. Jiang, Y. Pan, Y. Zhou, C. Owl, X. Zhai, N. Liu, C. Saunt, and T. Liu, "Transcending language boundaries: Harnessing llms for low-resource language translation," *arXiv preprint arXiv:2411.11295*, 2024, accessed on November 30, 2024. [Online]. Available: https://arxiv.org/abs/2411.11295

[15] J. León, D. Riofrío, F. Grijalva, and K. Tambaco, "Digital inclusion and culture: Training llama-2 to empower kichwa communities," in *2024 Tenth International Conference on eDemocracy eGovernment (ICEDEG)*, 2024, pp. 1–8, accessed on December 1, 2024.

[16] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2023, accessed on December 1, 2024.

[17] LangChain, "Concepts of retrieval-augmented generation (rag) in langchain," 2024, accessed: December 1, 2024. [Online]. Available: https://python.langchain.com/docs/concepts/rag/

[18] Hugging Face, "Advanced retrieval-augmented generation (rag) cookbook," 2024, accessed: December 1, 2024. [Online]. Available: https://huggingface.co/learn/cookbook/en/advanced_rag

[19] Y. Hu and Y. Lu, "Rag and rau: A survey on retrieval-augmented language model in natural language processing," 2024, accessed: December 1, 2024.

[20] LangChain, "Introduction to langchain: A framework for developing applications powered by llms," 2024, accessed: December 1, 2024. [Online]. Available: https://python.langchain.com/docs/introduction/#:~:text=LangChain%20is%20a%20framework%20for,%2C%20and%20third%2Dparty%20integrations.

[21] ——, "Tutorial on retrieval-augmented generation (rag) with langchain," 2024, accessed: December 1, 2024. [Online]. Available: https://python.langchain.com/docs/tutorials/rag/

[22] HuggingFace, "Labse: Language-agnostic bert sentence embedding," 2024, accessed: December 1, 2024. [Online]. Available: https://huggingface.co/sentence-transformers/LaBSE

[23] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Journal of Machine Learning Research*, vol. 22, no. 201, pp. 1–73, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1061.html

[24] HuggingFace, "Llama 3.1 8b model card," 2024, accessed: December 1, 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.1-8B