Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 07 July-2024, Page No.- 4404-4411

DOI: 10.47191/etj/v9i07.09, I.F. – 8.227

Advanced Retrieval Augmented Generation: Multilingual Semantic Retrieval across Document Types by Finetuning Transformer Based Language Models and OCR Integration

Ismail OUBAH[1], Dr. Selçuk ŞENER[2]

[1]MS Student, Computer Engineering, Istanbul Aydin University, Istanbul, Turkey

[2]Lecturer, Computer Engineering, Istanbul Aydin University, Istanbul, Turkey

ABSTRACT: This study presents an advanced system for multilingual semantic retrieval of diverse document types, integrating Retrieval-Augmented Generation (RAG) with transformer-based language models and Optical Character Recognition (OCR) technologies. Addressing the challenge of creating a robust multilingual Question-Answering (QA) system, we developed a custom dataset derived from XQuAD, FQuAD, and MLQA, enhanced by synthetic data generated using OpenAI's GPT-3.5 Turbo. This ensured comprehensive, context-rich answers. The inclusion of PaddleOCR facilitated high-quality text extraction in French, English, and Spanish, though Arabic presented some difficulties. The Multilingual E5 embedding model was fine-tuned using the MultipleNegativesRankingLoss approach, optimizing retrieval of context-question pairs. We utilized two models for text generation: MT5, fine-tuned for enhanced contextual understanding and longer answer generation, suitable for CPU-friendly uses, and LLAMA 3 8b-instruct, optimized for advanced language generation, ideal for professional and industry applications requiring extensive GPU resources. Evaluation employed metrics such as F1, EM, and BLEU scores for individual components, and the RAGAS framework for the entire system. MT5 showed promising results and excelled in context precision and relevancy, while the quantized version of LLAMA 3 led in answer correctness and similarity. This work highlights the effectiveness of our RAG system in multilingual semantic retrieval, providing a robust solution for real-world QA applications and laying the groundwork for future advancements in multilingual document processing.

KEYWORDS: Multilingual Retrieval-Augmented Generation, QA System, OCR, LLAMA 3, MT5, Multilingual E5

I. INTRODUCTION

Finding information in multiple languages and various types of documents is crucial but challenging due to the diverse nature of the data. Traditional keyword searches often does not work as expected, especially when dealing with information not explicitly written or in different languages. And dealing with scanned documents is more challenging task. The retrieval augmented generation approaches are in growing demand to address these issues for easy accessibility and comprehension of this information. We will leverage transformer-based language models such as MT5 or LLaMA3, which involves advanced OCR capabilities to develop AI systems for multilingual semantic retrieval across multiple documents. This article was inspired by the power of a state-of-the-art RAG system for revolutionizing textual data interactions. This is to overcome some of the limitations of classical retrieval techniques, offering more accurate and efficient information retrieval deployed on OCR technology for scanned documents combined with advanced linguistic and semantic functionality of transformer-based language models. These systems enable individuals to access and understand information in their preferred language and format, promoting diversity, accessibility, and global knowledge sharing.

II. RELATED WORKS

Early implementations of Retrieval-Augmented Generation (RAG) models primarily focused on directly retrieving documents from given queries and then processing the retrieved documents to generate responses. Recent advancements have integrated retrieval and generation processes more seamlessly. For instance, Lewis et al. (2021) demonstrated that dynamic retrieval and the integration of relevant information during generation significantly enhance the quality and relevance of generated text. Prominent examples of advanced RAG systems include Google's RETRO (Borgeaud et al. 2022), which dynamically integrates information from large corpora into the Transformer architecture during generation, resulting in highly accurate and informative outputs. FLARE (Jiang et al. 2023) iteratively queries internet searches based on initial outputs to ensure real-time information accuracy, and META's System 2 Attention (Weston and Sukhbaatar 2023) reconstructs context to maintain response relevance. Multilingual RAG

systems leveraging language models with external knowledge sources have also improved text generation across multiple languages. These methods, which include cross-lingual query generation and iterative retrieval during generation, address issues of truthfulness and information currency in language models (Jiang et al. 2023; Ramos, Martins, and Elliott 2023; Zhuang, Shou, and Zuccon 2023). These approaches have shown competitive performance in generating multilingual text without relying on extensive supervised training data (Shao et al. 2023).

Combining Optical Character Recognition (OCR) with Large Language Models (LLMs) has significantly advanced applications in translation, sentiment analysis, and multilingual contexts (Gao, Song, and Yin 2023). This integration enables data augmentation techniques for cross-lingual commonsense reasoning datasets, enhancing model performance through synthesized data generated by advanced LLMs like GPT-4 (Whitehouse, Choudhury, and Aji 2023). These advancements underscore the potential of RAG systems with multilingual and OCR integration to improve text generation and information retrieval across various applications.

III. METHODOLOGY

This section details the workflow of our project, which was developed using Python, with all analysis, training, and evaluations conducted on Google Colab, utilizing a T4 GPU with 15GB of RAM. All models were sourced from the HuggingFace community and integrated with the LangChain framework. The models were fine-tuned using Torch Frameworks.

A. System Workflow

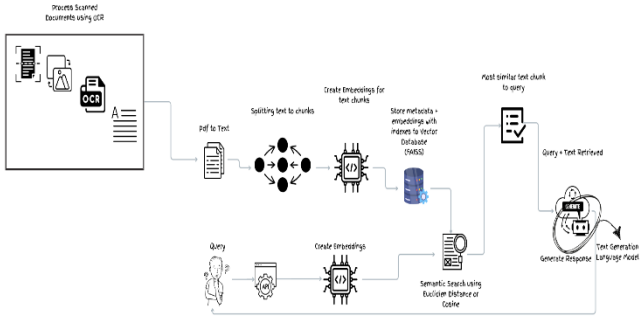


Figure 1: The workflow of the proposed QA system

The workflow of our proposed system consists of several key stages:

- ✓ **Datasets Used:** We began by gathering and preparing datasets of multilingual scanned documents, involving preprocessing steps like OCR (Optical Character Recognition) to convert scanned images into text and then generating synthetic dataset using OpenAI GPT 3.5 Turbo model.
- ✓ **Fine-Tuning the Embedding Model:** We fine-tuned an embedding model to handle multilingual text, responsible for creating dense vector representations of the documents.
- ✓ **Document Vector Store (FAISS):** We utilized FAISS (Facebook AI Similarity Search) for efficient similarity search and retrieval of document vectors, enabling quick location of relevant documents based on embeddings.
- ✓ **Fine-Tuning the Text-to-Text Generation Model (MT5):** The MT5 model was fine-tuned to improve performance in generating accurate and coherent text based on the retrieved documents, crucial for handling multilingual text generation nuances.
- ✓ **Text Generation Model (LLAMA 3):** For text generation, we used the LLAMA 3 model, quantized and then fine-tuned using a prompt tuning strategy to generate responses based on information retrieved from the document vector store.
- ✓ **Quantization of the Model:** To optimize the model for deployment, we applied quantization techniques, reducing model size and improving inference speed without significantly sacrificing performance.
- ✓ **Prompt Tuning:** We fine-tuned the prompts used to query the model, ensuring they elicited the most relevant and accurate responses from the text generation system.
- ✓ **Evaluation Step**: Evaluating each model separately and then evaluate the full RAG system

B. Datasets Preparation

In developing a robust question-answering (QA) system, we faced challenges in finding suitable datasets. Many publicly available datasets were either incompatible with our needs or exclusively in English. After extensive exploration, we identified suitable multilingual datasets: XQuAD, FQuAD, and MLQA, derived from the well-known SQuAD dataset. We focused on four languages: French, English, Spanish, and Arabic, creating a custom dataset with extended answers to enable better context understanding.

SQuAD Dataset: The Stanford Question Answering Dataset includes over 100,000 reading comprehension questions in English, collected from Wikipedia articles.

XQuAD Dataset: The Cross-lingual Question Answering Dataset evaluates cross-lingual QA performance, containing a subset of paragraphs and question-answer pairs from the SQuAD v1.1 development set, translated into multiple languages.

FQuAD Dataset: A French native reading comprehension dataset consisting of over 25,000 questions based on Wikipedia articles, created by higher education students.

MLQA Dataset: is an evaluation dataset designed to evaluate cross-lingual QA performance, MLQA has about 5K

extractive QA examples (12K in English) in SQuAD format across seven languages and were collected through Wikipedia articles and an alignment context method.

Custom Dataset: Created from scanned documents using PaddleOCR for text extraction, focused on French, Spanish, and English due to difficulties with Arabic.

We analyzed the datasets using Python libraries such as NumPy, Pandas, MatplotLib, and WordCloud, performing data processing and visualization to ensure dataset quality and effectiveness.

C. Fine-Tuning The Embedding Model

Embedding models are crucial for a RAG application, providing dense vector representations of documents. We fine-tuned the embedding model using the Sentence Transformer Library and a synthetic dataset, focusing on the "MultipleNegativesRankingLoss" function and "InformationRetrievalEvaluator" for performance monitoring.

Steps included splitting the dataset, formatting it using InputExample from Torch library for creating data loaders, and setting training parameters (4 epochs, batch size of 16, etc.). This process enhanced the model's ability to capture domain-specific data nuances, improving retrieval performance.

D. Document Vector Store (FAISS)

We used FAISS for storing embeddings and metadata of multilingual scanned documents due to its performance, scalability, specialized indexing, memory efficiency, and integration with machine learning frameworks.

E. Fine-Tuning the Text-to-Text Generation Model (MT5)

The MT5 model was fine-tuned in two stages: first on QA datasets (MLQA, XQuAD, FQuAD) and then on our synthetic dataset to improve context understanding and response quality. We used a T4 GPU for initial training and upgraded to an A100 GPU for more efficient training on longer sequences.

F. Quantize and Prompt Tuning the Text Generation Model (LLAMA 3)

We employed the LLAMA 3 instruct model for text generation, we applied Post-Training Quantization method using the AutoGPTQ library. This process was very important to manage the model's size and enhancing its inference speed, so that we can easily deploy and testing it on our available hardware. We quantized by reducing the precision of its weights in the model to 4 bits, thereby reducing the memory footprints and computation requirements. We used the "wikitext2" dataset during quantization to ensure the model maintained high performance and accuracy despite the reduced precision. After quantization, we prompt-tune the LLAMA-3 model, so it better suits the needs of our more specific text generation tasks. Prompt tuning is simply fine-tuning the prompts used to query the model, ensuring that they are organized in a way that gives the most accurate answers.

IV. EVALUATION

Evaluation is a critical component of the training process for machine learning models, especially for Retrieval-Augmented Generation (RAG) systems. It helps in understanding the model's performance, identifying its strengths and weaknesses, and guiding further optimizations. In this section, we will evaluate the generation model metrics and the embedding model metrics.

A. Evaluation of The Generation Model (MT5)

Evaluating the performance of QA models is challenging because the model might produce answers that are nearly correct but not exact. Additionally, an answer might be entirely accurate even if the word order differs from the ground truth. To assess the model's responses after fine-tuning, we employ three widely used metrics in the QA field: Exact Match (EM), F1, and BLEU scores.

Metrics Used in Evaluation

- ✓ **EM-score**: The Exact Match score (EM) is a binary metric that compares the predicted answer to the ground truth. If the predicted answer is identical to the ground truth, the EM score is 1; otherwise, it is 0.
- ✓ **F1-score**: The F1-score is the harmonic mean of precision and recall and is commonly used to measure the accuracy of model predictions in NLP. The F1-score ranges from 0 to 1. A score of 1 indicates that both precision and recall are perfect, while a score of 0 indicates that either precision or recall is zero.
- ✓ **BLEU-score**: The BLEU (Bilingual Evaluation Understudy) score evaluates the quality of text translated from one language to another. It compares n-grams of the predicted answer with those of the reference answer, measuring how many n-grams in the predicted answer appear in the reference. The BLEU score ranges from 0 to 1, where 1 indicates a perfect match with the reference text.

B. Evaluation of The Retriever Model for Semantic Search

Evaluating the embedding model is crucial to ensure the effectiveness and accuracy of the RAG system. The embedding model determines how well the system retrieves relevant documents that will be used by the generation model to produce answers. Accurate evaluation helps identify strengths and weaknesses, guiding further optimizations. We employ several metrics to assess different aspects of the embedding model's performance: Cosine Accuracy, Precision, Recall, Mean Reciprocal Rank (MRR), Mean

Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG).

Metrics Used in Evaluation

- ✓ **Cosine Accuracy**: Measures the proportion of times the correct documents are among the retrieved results based on cosine similarity.
- ✓ **Precision**: The fraction of relevant items among the retrieved items.
- ✓ **Recall**: The fraction of relevant items retrieved out of all relevant items available.
- ✓ **Mean Reciprocal Rank (MRR)**: Evaluates the ranking quality by focusing on the position of the first relevant document in the retrieved results
- ✓ **Mean Average Precision (MAP)**: Evaluates precision across multiple ranks.
- ✓ **Normalized Discounted Cumulative Gain (NDCG)**: Evaluates the ranking quality by considering the graded relevance of documents and their positions within the retrieved results.

To ensure a comprehensive evaluation, we consider different values of K, which specify the number of top-ranked results to be evaluated. This allows us to assess the model's performance across various retrieval depths, providing a more detailed understanding of its effectiveness.

C. Evaluation of the Full RAG System for MT5 and LLAMA 3-8b

In our project, since we only quantized the LLAMA 3 model and performed prompt tuning without making significant changes to its weights, we evaluate it within the full RAG system. This approach allows us to see its performance and compare it with the other model (MT5). Building a production-ready RAG application involves optimizing the performance of both the Retriever and Generator components. Evaluating the models separately provides insights into their individual performance, but it is crucial to assess how these models work in combination. This evaluation identifies areas where the RAG pipeline may require improvements and helps track performance over time.

RAGAS: Retrieval-Augmented Generation Assessment

RAGAS is a comprehensive framework designed to facilitate the evaluation of RAG pipelines at a component level, providing the necessary tools and metrics to assess the effectiveness of both the retriever and generator components.

Evaluation Data

RAGAS leverages LLMs to conduct evaluations, minimizing the need for human-annotated data. The framework requires the following information for evaluation:

- **Question**: The user query that serves as the input to the RAG pipeline.
- **Answer**: The generated answer from the RAG pipeline.
- **Contexts**: The contexts retrieved from the external knowledge source used to answer the question.
- **Ground Truths**: The ground truth answer to the question, required for specific metrics such as context recall.

Metrics Used in Evaluation

- **Context Precision@k**: Measures if all relevant items are ranked high in the given contexts,
- **Context Recall**: Measures how well the retrieved context matches the ground-truth answer:
- **Answer Relevance**: Measures how relevant the generated answer is to the given prompt, based on cosine similarity between the original question and reverse-engineered questions.
- **Answer Correctness**: Measures how accurate the generated answer is compared to the ground truth, considering semantic and factual similarity.
- **Answer Semantic Similarity:** Answer Semantic Similarity measures how closely the meaning of the generated answer matches the ground truth.
- **Faithfulness:** measures how well the answer matches the information provided. It's like making sure the answer sticks to the topic and doesn't go off track.

Evaluating the embedding model using these metrics provides a comprehensive understanding of its performance. By employing metrics such as Cosine Accuracy, Precision, Recall, MRR, MAP, and NDCG, we can ensure that the embedding model effectively retrieves relevant documents and ranks them appropriately. The use of specific k values, such as 1 to 10 for most metrics and 100 for MAP, allows us to assess the model's performance at different levels of retrieval depth. This rigorous evaluation is crucial for the overall success of the RAG system, as it directly influences the quality of the generated answers either by the MT5 or LLAMA3 model.

V. RESULTS

This section presents the evaluation results of the different components of our multilingual Retrieval Augmented Generation System, we will start by evaluating each component by its own and then we will use the RAGAS framework to evaluate whole RAG system.

MT5 Results

For this model that we fine-tuned it twice , we will first evaluate the first version on the test subset that has been created from XQUAD,MLQA and FQUAD, we will not rely on the Arabic subset cause we didn't work with it in the later stages, so our main focus is to see how it perform in English,French and Spanish dataset. After that we will evaluate it on the custom made dataset that has been extracted from the scanned documents using the PaddleOCR, which contains longer answers to see how it performs on them.

This table contains the obtained results for the first version of fine tuning

Table 1: Base Model Results

Language	BLEU Score	F1 Score	EM Score
Arabic	0.0026	0.0116	0.0000
English	0.0082	0.0420	0.0000
French	0.0029	0.0199	0.0000
Spanish	0.0090	0.0533	0.0000

Table 2: Fine-Tuned Model V1 Results

Language	BLEU Score	F1 Score	EM Score
Arabic	0.3865	0.4732	0.3240
English	0.5820	0.6597	0.5513
French	0.5792	0.6603	0.4482
Spanish	0.5190	0.6414	0.4362

Now passing to evaluating the 2nd Version of fine tuning on the longer answers between the 1st and 2nd version

Table 3: Fine-Tuned Model V1 Results on custom dataset

Language	BLEU Score	F1 Score	EM Score
English	0.10	0.25	0.00
Spanish	0.05	0.15	0.00
French	0.03	0.09	0.00

Table 4: Fine-Tuned Model V2 Results on custom dataset

Language	BLEU Score	F1 Score	EM Score
English	0.65	0.75	0.23
Spanish	0.59	0.71	0.20
French	0.61	0.69	0.15

Embedding Model Results

Fine-tuning and evaluating embedding model is tricky but luckily Sentence Transformer made it easy for us. The dataset used for evaluating the embedding model is slightly different that the other one ,since we managed to augment the dataset to 6500+ samples by combining the extracted text from scanned documents and normal type of documents that were only focus on the specific domain so that we can ensure high quality data and also make the dataset larger ,while using GPT 4 to generate question for each context sample. Here is how the dataset looks like.

	chunk	language	question	id
0	second-mode internal wave, observed by Farmer ...	english	What significant role do second-mode internal ...	529
1	surface of the Atlantic layer by ventilating t...	english	How does the refinement of the representation ...	838
2	are directly induced by the gravitational forc...	english	What specific action needs to be taken in the ...	245
3	previous studies (Harzallah et al, 2014; Naran...	english	What are the specific factors identified in th...	718
4	tion pattern is, to some extent, closed at the...	english	What changes occur in the volume transports be...	523
5	villes surpeuplées, ...). ...	french	Quelles sont les conséquences possibles de la ...	1648
6	En tant que passage oblig´ e du champ circulat...	french	Quel rôle joue le détroit de Gibraltar en tant...	1964
7	un rôle important pour la biodiversité de la m...	french	Quel impact la coopération scientifique pour m...	1738
8	evelopper cette activit´ e dans des eaux 7. "G...	french	Quels sont les incidents qui ont entraîné des ...	2876
9	terminal véhicules (2012). Une zone franche lo...	french	Quelle est la capacité prévue du port Tanger M...	1385
10	bases en Italia, cuya base de Nápoles es princ...	spanish	¿Cuáles son las bases militares de Rota y Nápo...	5025
11	ratificado el 26 de abril del 2004 y está en v...	spanish	¿Cuándo fue ratificado el Protocolo sobre la p...	5494
12	209 El artículo 1.1 estipula lo siguiente: (A ...	spanish	¿Qué zona geográfica abarca el Convenio según ...	5050
13	37.433,40 561,50 7.028,13 Modalidad de ejecuci...	spanish	¿Qué tipo de modalidad de ejecución se utiliza...	4465
14	Gran-Belt y el Pequeño-Belt) cuyos regímenes d...	spanish	¿Cuál es la relevancia de la Declaración del 8...	5370

. Figure 1: A snippet of Evaluation Dataset for Embedding Model Evaluation

So, after training the model for 4 epochs with a batch size of 16 samples , these are the results we could get and comparing it with the base model

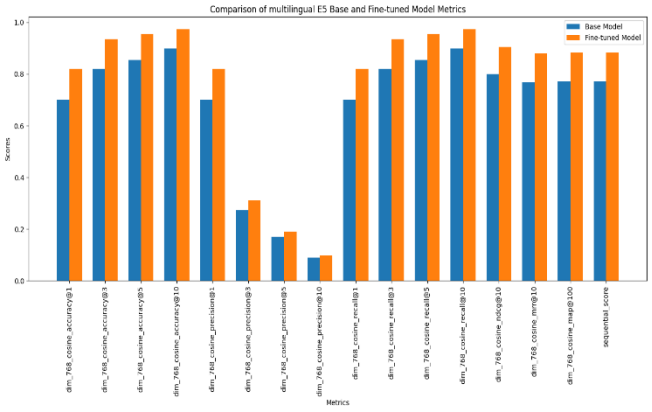


Figure 3: Comparison of ME5 Base and Fine-tuned Model Metrics

The provided results for the base and fine-tuned embedding models reveal significant improvements across various performance metrics after fine-tuning.

Results of the Full RAG System

Now passing to the final stage of this approach, evaluating the full RAG system using RAGAS. This is different than other evaluations because in this step we will prepare a small dataset of 10 samples of each language and carefully selecting chunks, questions and the ground truth answers that has been generated using GPT4 and also with human supervision which will look carefully at every triplet (context,question,answer) since 30 samples can easily be done by human so that we can have a good evaluation of the system. This is how the initial dataset for validation looks like

	splited_content	questions	ground_answers	language
0	Assuming that peak hour traffic is equal to 30...	What is the average smoke emission rate for d...	The average smoke emission rate is 17.5 for v...	english
1	2-exhaust fans Northbound tube,Zone 4- 2-suppl...	How many supply fans and exhaust fans are req...	Each zone in the modular system is served by ...	english
2	Unbalanced (i.e. unequal supply and exhaust ra...	What factor determines the time for the smoke...	The time for the smoke layer to descend is de...	english
3	- system would have the lowest equipment insta...	Why was the auto-cleansing system deemed wort...	The auto-cleansing system has the potential f...	english
4	- Wet Collectors -The advantages and disadvant...	What are the advantages and disadvantages of ...	The advantages of wet collectors include a lo...	english
5	- 2.0 - 2.1 General - This section of the repo...	What are the three basic ventilation systems ...	The three basic ventilation systems commonly ...	english
6	- which are potentially feasible and warrant f...	How do traffic density and vehicle characteri...	Traffic density and vehicle characteristics i...	english
7	Air velocities of 2.54m/sec to 2.95 m/sec were...	What was demonstrated to be a major factor in...	Longitudinal airflow was demonstrated to be a...	english
8	- The first phase would consist of advancing a...	What is the primary objective of the first ph...	The primary objective of the first phase is t...	english
9	- speed of 40 kilometers per hour and a tunnel...	What is the maximum longitudinal tunnel air v...	The maximum longitudinal tunnel air velocity ...	english
10	1.4 1.41 Modifications apportees aux modeles i....	Comment les résultats des prévisions de trafi...	Les résultats des prévisions de trafic présen...	french
11	Lors du test des nouveaux modeles effectue au ...	Quels sont les principes adoptés pour les aju...	Les principes adoptés étaient de recalibrer l...	french
12	Pourcentages du trafic en vrac dans l'ensemble...	Quelle est la répartition des tonnages par mo...	Pour l'année de base 1990, la répartition des...	french
13	- Importation des pays d'Afrique Elasticite du...	Comment se fait le calcul prévisionnel des ta...	Le calcul prévisionnel se fait par applicatio...	french
14	1.622 Les trafics de marchandises sur Ceuta Ce...	Quel type de marchandises est principalement ...	Principalement, 42 milliers de tonnes nettes ...	french
15	U. L.C. Total L.U. L.C. Total ENCLAVES EN EURO...	Comment le mode de répartition du trafic spéc...	Il a été supposé que les trafics à destinatio...	french
16	1990 - unite : millier de tonnes nettes EUROPE...	Quel est le total des tonnes nettes échangées...	Le total des tonnes nettes échangées entre l'...	french
17	- Exportation des pays d'Afrique Modeles logar...	Quels sont les coefficients de corrélation po...	Les coefficients de corrélation sont de 0.76 ...	french
18	: - % accroissement annuel 2005 1.88% 2025 2.3...	Quel type de produits a le taux d'accroisseme...	Les produits agricoles et industriels de base...	french
19	3.222 Tableau recapitulatif des echanges entre...	Combien de tonnes sont échangées entre l'Ital...	Un total de 2 817 tonnes sont échangées entre...	french
20	Al igual que en el caso de la estacion el terc...	¿Qué tipo de distribución tensional se observ...	Tras el equilibrio de la Galería, el modelo d...	spanish
21	Al igual que en el caso de la estacion el terc...	¿Qué criterio se utilizó para realizar el ter...	Se utilizó el criterio de Mohr-Coulomb para r...	spanish
22	Para la obtencion de testigos, con los que se ...	¿Qué tipo de ensayos se realizaron para medir...	Se realizaron ensayos dilatométricos para med...	spanish
23	En concreto. en cada caso. se ha mantenido la ...	¿Qué criterio se utilizó en el segundo anális...	Se utilizó el criterio de strain-softening, q...	spanish
24	Durante la excavacion de la Galeria de Tarifa ...	¿Qué instrumento se utilizó para medir los de...	Se utilizaron comparadores mecánicos con una ...	spanish

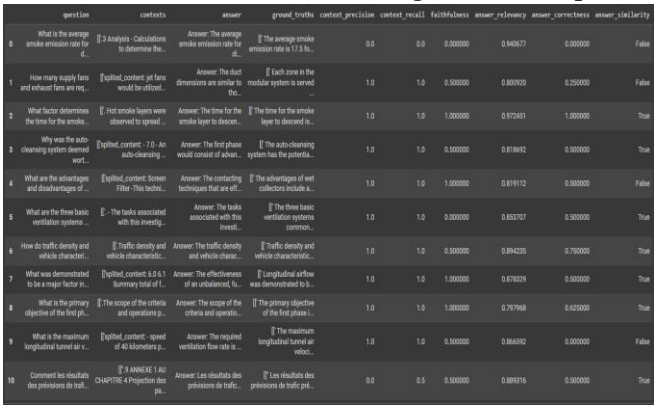
Figure 4: A Snippet of the Initial Evaluation Dataset for RAGAS Evaluation

1. Evaluation Strategy

To prepare the dataset on a format how RAGAS expected , this is what has been done.

First we will develop a simple RAG pipeline using the raw documents to build our FAISS database and then we will pass each question of the initial dataset on the database so that the model embedding model will use semantic search and try to retrieve the most similar chunks to the question, after that a function to call the generator model will be given the retrieved chunks with the question so that it will generate the answer of the question from the retrieved context, this will be done for both types of generators models the finetuned MT5 and the quantized LLAMA3-8b model.

The results where MT5 model is the generator component



	question	contexts	answer	ground_truths	context_precision	context_recall	faithfulness	answer_relevancy	answer_correctness	answer_similarity
0	What is the average smoke emission rate for d...	['[3 Analysis - Calculations to determine the...	Answer: The average smoke emission rate for d...	['[The average smoke emission rate is 17.5 fo...	0.0	0.0	0.000000	0.940677	0.000000	False
1	How many supply fans and exhaust fans are req...	['[splited_content: jet fans would be utilized...	Answer: The duct dimensions are similar to tho...	['[Each zone in the modular system is served ...	1.0	1.0	0.500000	0.809920	0.250000	False
2	What factor determines the time for the smoke...	['[Hot smoke layers were observed to spread ...	Answer: The time for the smoke layer to descen...	['[The time for the smoke layer to descend is...	1.0	1.0	1.000000	0.972451	1.000000	True
3	Why was the auto-cleansing system deemed wort...	['[splited_content: - 7.0 - An auto-cleansing ...	Answer: The first phase would consist of advan...	['[The auto-cleansing system has the potentia...	1.0	1.0	0.500000	0.818692	0.500000	True
4	What are the advantages and disadvantages of ...	['[splited_content: Screen Filter -This techni...	Answer: The contacting techniques that are eff...	['[The advantages of wet collectors include a...	1.0	1.0	1.000000	0.819112	0.500000	False
5	What are the three basic ventilation systems ...	['[- The tasks associated with this investig...	Answer: The tasks associated with this investi...	['[The three basic ventilation systems common...	1.0	1.0	0.000000	0.853707	0.500000	True
6	How do traffic density and vehicle characteri...	['[Traffic density and vehicle characteristic...	Answer: The traffic density and vehicle charac...	['[Traffic density and vehicle characteristic...	1.0	1.0	0.500000	0.894235	0.750000	True
7	What was demonstrated to be a major factor in...	['[splited_content: 6.0 6.1 Summary total of f...	Answer: The effectiveness of an unbalanced, fu...	['[Longitudinal airflow was demonstrated to b...	1.0	1.0	1.000000	0.878029	0.500000	True
8	What is the primary objective of the first ph...	['[The scope of the criteria and operations p...	Answer: The scope of the criteria and operatio...	['[The primary objective of the first phase i...	1.0	1.0	1.000000	0.797968	0.625000	True
9	What is the maximum longitudinal tunnel air v...	['[splited_content: - speed of 40 kilometers p...	Answer: The required ventilation flow rate is ...	['[The maximum longitudinal tunnel air veloci...	1.0	1.0	0.500000	0.866592	0.000000	False
10	Comment les résultats des prévisions de trafi...	['[9 ANNEXE 1 AU CHAPITRE 4 Projection des pa...	Answer: Les résultats des prévisions de trafic...	['[Les résultats des prévisions de trafic pré...	0.0	0.5	0.500000	0.889316	0.500000	True

Figure 5: A Snippet of RAGAS Evaluation Results For MT5 as Generator Model

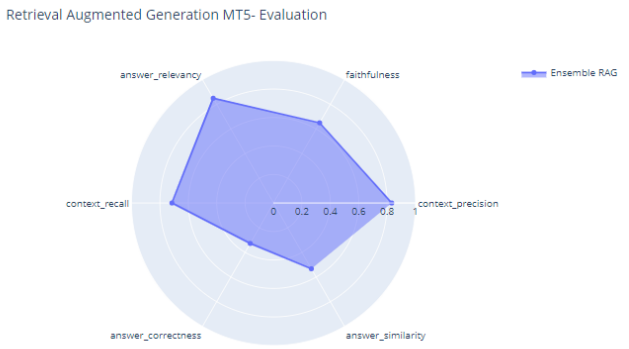


Figure 6: T5 Performance in Various Metrics

The results where LLAMA 3 8b model is the generator component

	question	contexts	answer	ground_truths	context_precision	context_recall	faithfulness	answer_relevancy	answer_correctness	answer_similarity
0	What is the significance of the black line mar...	['The black line marks the 1981 average interf...	The black line marking the 1981 average interf...	['The black line marking the 1981 average inter...	1.0	0.5	0.500000	0.971437	0.500000	True
1	What role does deep convection play in the wat...	['and mainly relates to the salinity of the upp...	Deep convection plays a crucial role in the wa...	['Deep convection plays a crucial role in the v...	1.0	1.0	0.428571	0.972096	0.750000	True
2	What are some of the acronyms used to represen...	['mid and bottom depths of the Mediterranean. T...	According to the provided context, the acronym...	['The context mentions the acronyms "TAC", "CO"...	1.0	0.0	0.500000	0.928909	0.500000	True
3	What role does dense water overflow play in di...	['the open Ocean (Bryden and Nurser, 2003). Alt...	According to Bryden and Nurser's research from...	['Dense water overflow plays a crucial role in ...	1.0	1.0	1.000000	0.996277	0.833333	True
4	What are the main factors contributing to the ...	['ocean stratification, analyzed in the follow...	According to the text, the main factors contri...	['The main factors contributing to the differen...	1.0	1.0	0.500000	0.984347	1.000000	True
5	Quel est le secteur principal responsable de l...	['atlantiques. En particulier, le littoral andalou...	Le secteur principal responsable de la surcons...	['Le secteur principal responsable de la surcon...	1.0	1.0	1.000000	0.969816	1.000000	True
6	Quels défis doivent surmonter les immigrants l...	['à l'UE, l'Espagne est diverse une porte d'ent...	Les défis que les immigrants doivent surmonter...	['Les immigrants qui tentent de traverser la Mé...	1.0	1.0	0.500000	0.986028	0.785714	True
7	Quelle importance revêt le concept de 'capital...	['comparatif et de la productivité' et des tert...	Selon Weinne Schoumaker en 2007, le concept d...	['Le concept de "capital relationnel" est d'une...	1.0	1.0	1.000000	0.903345	1.000000	True
8	Quels types de promotions sont offerts aux cli...	['es, des forfaits familiaux, des tarifs qr' et...	Selon le site Marco-Ferry.com, les promotions ...	['Le contexte ne fournit pas d'informations spé...	1.0	0.0	0.500000	0.988171	0.750000	True
9	Quel stratégème les protagonistes ont-ils déci...	['a t' emorter des voitures et les remorter aut...	Les protagonistes ont décidé d'utiliser un st...	['Les protagonistes ont décidé d'utiliser un co...	1.0	0.5	0.666667	0.959783	0.666667	True
10	¿Cómo influyó la libre comunicación marítima e...	['Intereses que tenían que defender, basada en...	La libre comunicación marítima tuvo un impacto...	['La libre comunicación marítima actuó como la ...	1.0	1.0	0.666667	0.948499	0.666667	True
11	¿Qué tipo de certificados internacionales pued...	['del anexo. Para que se respete la condición s...	Los certificados internacionales que pueden ser...	['Los buques pueden ser inspeccionados para ver...	1.0	1.0	0.500000	0.978355	1.000000	True
12	¿Cómo se sugiere que los Estados tepan una or...	['65 Es lo que podemos leer en el principio 13...	Según el principio 13 de la Declaración de Est...	['Según el principio 13 de la Declaración de Es...	1.0	1.0	1.000000	0.922991	0.833333	True
13	¿Cómo se manifiesta el principio de responsab...	['negociaciones internacionales subsecuentes, d...	En el marco ambiental, el principio de respons...	['El principio de responsabilidad común pero di...	1.0	1.0	1.000000	0.970674	1.000000	True
14	¿Qué medidas tomará la autoridad competente en...	['será informado para que solucione el problema...	La autoridad competente tomará medidas para pr...	['La autoridad competente se involucrará dentro ...	1.0	1.0	1.000000	0.919337	0.875000	True

Figure 7: A Snippet of RAGAS Evaluation Results For LLAMA3-8b as Generator Model

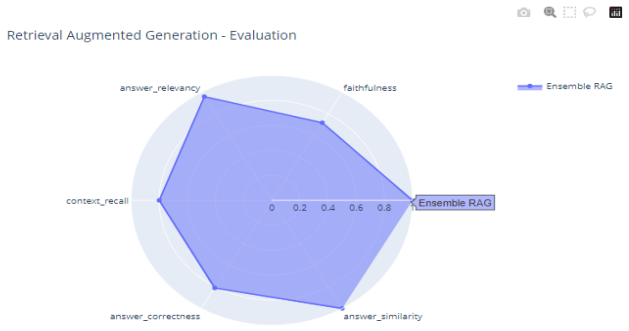


Figure 8: LLAMA3-8b Performance in Various Metrics

VI. DISCUSSION

In this chapter, we discuss the implications of our results, analyze the performance of each component of the multilingual Retrieval Augmented Generation (RAG) system, and highlight the limitations encountered during our study.

Interpretation of the Results

1. OCR Text Extraction High-quality OCR text extraction is crucial for our multilingual RAG system, impacting both fine-tuning and real-time question answering. OCR errors,

such as misrecognitions, can significantly hinder model performance. Ensuring high OCR accuracy minimizes these errors, allowing better context comprehension and more accurate answers. While we did not fine-tune the OCR model, doing so could enhance text quality, further improving the system's overall performance.

2. MT5 Model Performance

a. Improvement Over Base Model The fine-tuned MT5 model shows significant improvement over the base model across all tested languages (Arabic, English, French, and Spanish) after only eight epochs. Metrics such as BLEU, F1, and EM scores increased, indicating improved accuracy and relevance of generated answers. English, benefiting from better optimization and larger training data, consistently performed the best. Arabic, however, had the lowest scores, highlighting potential challenges with this language. The base model's near-zero scores across metrics emphasize the importance of fine-tuning, as it markedly enhances the model's performance and suitability for multilingual tasks.

b. Comparison Between V1 and V2 Comparing V2 to V1, there is a noticeable improvement in BLEU and F1 scores for English, Spanish, and French, demonstrating V2's enhanced capability in generating relevant and contextually appropriate answers. However, EM scores remain low, which is expected due to the model's context-based generation approach, leading to variations in exact wording compared to the ground truth.

c. Summary Both V1 and V2 fine-tuned models show substantial improvements over the base model. V1 excels across all metrics for each language, while V2 enhances BLEU and F1 scores further, especially in generating longer answers. Despite low EM scores, fine-tuning is crucial for improving model performance, particularly for multilingual applications.

3. Embedding Model Performance The fine-tuned embedding model exhibits remarkable performance improvements over the base model, despite being trained for only 4 epochs with a batch size of 16 due to computational constraints. Metrics such as accuracy, precision, recall, and ranking quality indicate enhanced retrieval and ranking capabilities. These results suggest the fine-tuning process was highly beneficial, and with greater computational resources, even more significant improvements could likely be achieved.

4. Full RAG System Evaluation The LLAMA 3-8b model significantly outperforms the fine-tuned MT5 model across nearly all evaluation metrics. LLAMA's perfect context precision and higher context recall scores highlight its superior retrieval capabilities, ensuring access to the most relevant information for generating accurate answers. Its higher faithfulness and answer correctness scores further indicate the generation of factually accurate responses. The answer relevancy and similarity scores underscore LLAMA's ability to produce highly relevant answers closely matching the ground truth. In contrast, the MT5 model's lower scores reveal weaknesses in generation stages. While the MT5 model shows promise if fine-tuned on larger, high-quality samples, the LLAMA 3-8b model demonstrates superior performance, making it a more suitable choice for the RAG system in this evaluation.

VII. CONCLUSIONS

This study successfully explored and implemented a cutting-edge multilingual Retrieval-Augmented Generation (RAG) system, leveraging transformer-based models such as MT5 and LLAMA 3-8b. Through a thorough review of artificial neural networks, deep learning architectures, and natural language processing methodologies, we established a solid theoretical foundation. Our approach integrated Optical Character Recognition (OCR) technologies to process scanned documents in multiple languages, significantly enhancing the system's ability to handle diverse document types. The methodology included full fine-tuning of both the embedding and text-to-text generation models, with additional optimization and prompt tuning for larger models. We utilized various datasets, including SQuAD, XQuAD, MLQA, and custom datasets, during the training and testing phases, demonstrating that with appropriate datasets and extended training epochs, the multilingual base T5 model can achieve performance levels comparable to larger models. Our evaluations showed that the LLAMA 3-8b model outperformed the MT5 model across all metrics, making it the optimal choice for applications with access to powerful GPUs to manage its weights post-quantization. However, challenges such as resource limitations, dataset availability, and OCR accuracy constraints were identified, pointing to future research and system refinement opportunities. In conclusion, this study makes a significant contribution to the advancement of multilingual RAG systems and provides valuable insights into improving document processing capabilities across diverse languages. Future research will focus on further exploring transformer architectures, expanding dataset resources, and refining OCR technologies to enhance system performance and applicability in real-world scenarios.

REFERENCES

1. Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2021. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv. https://doi.org/10.48550/arXiv.2005.11401.
2. Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, et al. 2022.

"Improving Language Models by Retrieving from Trillions of Tokens."
arXiv. https://doi.org/10.48550/arXiv.2112.04426.

3. Jiang, Zhengbao, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. "Active Retrieval Augmented Generation."
arXiv. https://doi.org/10.48550/arXiv.2305.06983.
4. Weston, Jason, and Sainbayar Sukhbaatar. 2023. "System 2 Attention (Is Something You Might Need Too)."
arXiv. https://doi.org/10.48550/arXiv.2311.11829.
5. Ramos, Rita, Bruno Martins, and Desmond Elliott. 2023. "LMCap: Few-Shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting."
arXiv. https://doi.org/10.48550/arXiv.2305.19821.
6. Zhuang, Shengyao, Linjun Shou, and Guido Zuccon. 2023. "Augmenting Passage Representations with Query Generation for Enhanced Cross-Lingual Dense Retrieval." In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1827–32. SIGIR '23. New York, NY, USA: Association for ComputingMachinery.
https://doi.org/10.1145/3539618.3591952.
7. Shao, Zhihong, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy."
arXiv. https://doi.org/10.48550/arXiv.2305.15294.
8. Du, Yuning, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, et al. 2020. "PP-OCR: A Practical Ultra Lightweight OCR System."
arXiv. https://doi.org/10.48550/arXiv.2009.09941.
9. Es, Shahul, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. "RAGAS: Automated Evaluation of Retrieval Augmented Generation."
arXiv. https://doi.org/10.48550/arXiv.2309.15217.
10. "Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date." n.d. Meta AI. Accessed May 11, 2024.
https://ai.meta.com/blog/meta-llama-3/.
11. Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. "Multilingual E5 Text Embeddings: A Technical Report."
arXiv. https://doi.org/10.48550/arXiv.2402.05672.
12. Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. "mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer."
arXiv. https://doi.org/10.48550/arXiv.2010.11934.
13. Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. "The Faiss Library."
arXiv. https://doi.org/10.48550/arXiv.2401.08281.
14. Whitehouse, Chenxi, Monojit Choudhury, and Alham Fikri Aji. 2023. "LLM-Powered Data Augmentation for Enhanced Cross-Lingual Performance."
arXiv. https://doi.org/10.48550/arXiv.2305.14288.
15. Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher. 2016. "Pointer Sentinel Mixture Models."
arXiv. https://doi.org/10.48550/arXiv.1609.07843.