National College of Ireland

Configuration Manual: Cross-Lingual RAG for English News article Summarization using Hindi context

MSc Research Project
MSc in Artificial Intelligence

Sandeep Kumar
Student ID: x23282835

School of Computing
National College of Ireland

Supervisor: Dr. Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sandeep Kumar		
Student ID:	x23282835		
Programme:	MSc in Artificial Intelligence	**Year:**	2024-2025
Module:	MSc Research Project		
Lecturer:	Dr. Rejwanul Haque		
Submission Due Date:	15/09/2025		
Project Title:	Configuration Manual: Cross-Lingual RAG for English news article summarization using Hindi context		
Word Count:	980	**Page Count:**	4

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sandeep Kumar

Date: 15/09/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Sandeep Kumar
Student ID: x23282835

1 Introduction

This is a step-by-step configuration manual to install, configure and run Cross-Lingual Retrieval-Augmented Generation (RAG) for News Summarization project. The project will be able to enhance English news summaries by extracting semantically similar Hindi news content as contextual input to employ multilingual embeddings, numerical search, and enormous large language models (LLMs) to include Gemini and Mistral together with T5 and BART. All necessary code, notebooks, and dataset are included in the full project code which is presented in the form of the .zip file.

2 System Requirements

2.1 Hardware Requirements

- **Processor**: macOS / Windows / Linux etc.
- **RAM**: 8 GB minimum
- **Storage**: 256 GB SSD (minimum), you can have more if want to store large datasets

2.2 Software Requirements

- **Operating System**: macOS / Windows 10/11 or above
- **Python Version**: 3.10+
- **Other Tools**: Anaconda (For creating environment)

3 Environment Setup

3.1 Install Conda (If it is not installed)

- Please download and install Anaconda from the official site: https://www.anaconda.com/download
- And then follow the installation wizard for your OS (Windows/Mac/Linux).
- Next, verify installation using below command:

```
conda --version
```

3.2 Create And Activate Project Environment

Please follow the command below to create and activate environment:

```
conda create -n crosslingual-rag python=3.10 -y
conda activate crosslingual-rag
```

3.3 Install Required Packages

Please navigate to the unzipped project folder & then run:

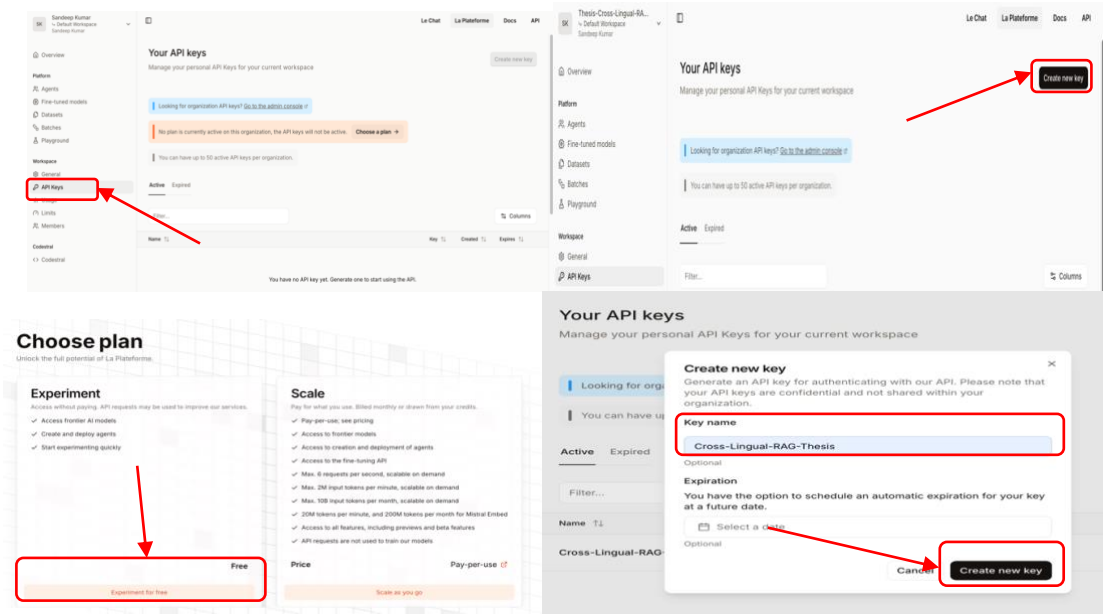
```
pip install -r requirements.txt
```

4 API Key Configuration

4.1 Set up and Generate Mistral API Key

1. Go to URL: https://console.mistral.ai/home
2. Sign in with your gmail/outlook /Apple ID
3. Then, please click on API keys section and then click create key.
4. Next, select the Experiment in plan as it is completely free for use.
5. Provide name for API Key and then click Create new Key.
6. Finally, API key will be generated
7. Please copy the generated key.
   Example: Dummy key: "AIzaSyD63o5xxxxxx851CJcod14plKVjSBn2u4s"

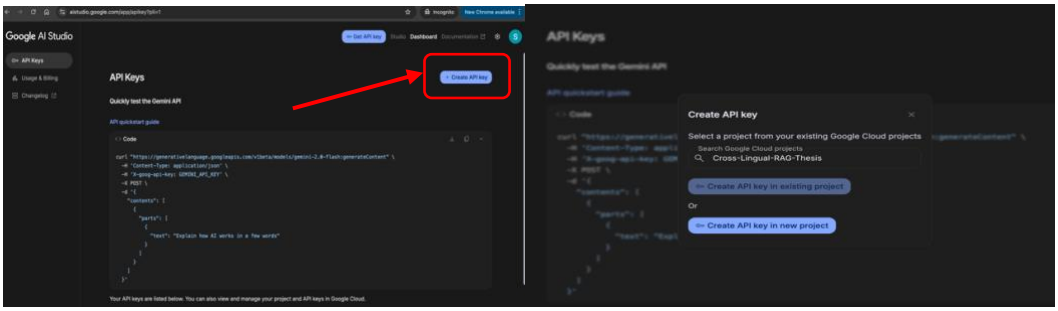
   Please follow the snapshots for API creation of Mistral API:



4.2 Set up and Generate Gemini API Key

1. First, go to URL : https://aistudio.google.com/app/apikey
2. Sign in with your gmail/outlook /Apple ID
3. Please click on Create API Key section as shown in the screenshot.
4. Provide specific name for your API Key and then click Create API Key in your new project.
5. Finally, API key will be generated
6. Please copy the generated key.

Please follow the snapshots for API creation of Gemini API:



4.3 Save API Keys in .env file

1. First, in the root project folder, please create a file named ".env":
2. And then **copy and paste** your API key created in previous step in .env file in below format.
3. Dummy Key Example:

MISTRAL_API_KEY=b02LcLSGD2QV1xxxxxNeox0jiWHuoKzBzfkfq
GOOGLE_API_KEY=AIzaSyD63o5xxxxxx851CJcod14plKVjSBn2u4s

5 Code Structure & File Description

Please find the complete code structure of this project which contains the file description. Complete project code is present in a file named 'Code_Cross_Lingual_RAG_Thesis':

Please find code structure below:

```
Code_Cross_Lingual_RAG_Thesis/
│
├── code/                    # All notebooks used for experiments and data prep.
│   ├── extract_news_dataset.ipynb
│   ├── cross_lingual_rag_experiment.ipynb
│   └── evaluation_cross_lingual_rag_experiment.ipynb
│
├── data/                    # Hindi and English news article datasets
│   ├── *.csv                # Raw and processed .csv (e.g - tiktok_ban.csv)
│
├── figures/                 # Evaluation plots and comparison visualizations
│
├── .env                     # API keys (Gemini, Mistral) – ignored via .gitignore
├── .gitignore
└── README.md                # Detailed Project documentation
```

6 Running the Project (Execution)

1. First, you need to Unzip the provided project .zip file.
2. And then activate the Conda environment using below command:

```
conda activate crosslingual-rag
```

3. Open the notebooks in code/ folder in Jupyter or VS Code:

4. For running **Training code** which has all the experiments:

```
cross_lingual_rag_experiment.ipynb
```

5. Please run all the cells sequentially for running training code.
6. For running **Evaluation code**:

```
evaluation_cross_lingual_rag_experiment.ipynb
```

Please run all the cells sequentially for running evaluation code.

References

Google (2025) Gemini API documentation. Available at: https://ai.google.dev/gemini-api/docs (Accessed: 8 August 2025).

Mistral AI (2025) Mistral AI API documentation. Available at: https://docs.mistral.ai (Accessed: 8 August 2025).

Reimers, N. and Gurevych, I. (2019) 'Sentence-BERT: Sentence embeddings using Siamese BERT-networks', Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

Hugging Face (2025) Sentence Transformers documentation. Available at: https://www.sbert.net (Accessed: 8 August 2025).

Newspaper3k (2025) Newspaper3k: Article scraping and curation. Available at: https://newspaper.readthedocs.io (Accessed: 8 August 2025).