National College of Ireland

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		
Supervisor:	Dr. Rejwanul Haque		
Submission Due Date:	15/09/2025		
Project Title:	Cross-Lingual RAG for English news article summarization using Hindi context		
Word Count:	7218	**Page Count**	20

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):	

Cross-Lingual RAG for English news article summarization using Hindi context

Sandeep Kumar
x23282835

Abstract

The research project that has been proposed examines the efficacy of Cross-Lingual Retrieval-Augmented Generation (RAG) in augmenting the English news summarization using Hindi contextual information. The paper refers to the problem of the creation of high-quality, condensed summaries of English news articles based on the incorporation of the relevant information extracted in a Hindi news corpus. We introduce a new framework that allows us to use multilingual sentence embeddings (LaBSE) to vectorize Hindi articles, store them in ChromaDB and retrieve contextual chunks according to their semantic similarity to English news articles. The most important is that we follow the strategy of translating the retrieved Hindi context into English by utilizing the Opus-MT model so that it could be combined with the English summarization models. To test the proposing framework, we have performed four different experiments, which are a baseline summarization (English article only), RAG with untranslated Hindi context, RAG with translated Hindi context, and RAG with semantically re-ranked Hindi chunks using a multilingual re-ranker (cross-encoder/ms-marco-MiniLM-L-6-v2). Such summarization models are BART, T5, Mistral, and Gemini. Assessment was carried out through Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Bilingual Evaluation Understudy (BLEU), and BERTScore scores and also using human assessment manually. In our study, we have proved that the inclusion of translated Hindi context in English news summaries increases the quality and informativeness of the summaries especially when there is a lack of sufficient details in the English article. The research is useful to gain an insight into the promises of cross-lingual RAG systems to enhance cross-lingual access and comprehension to information.

1 Introduction

1.1 Background and Rationale

The need to process multilingual contents has raised cross-lingual summarization of utmost concern. New add-on to dense retrieval systems, RAG systems incorporate dense retrieval and generative models to make summaries more relevant and factually evidenced (Do & Tran, 2024). LaBSE represents multilingual embedding models that allow one to perform the search on semantic similarity in different languages, which is why it is a good option to retrieve the context of foreign languages (Tiyajamorn et al. 2021). Neural machine translation (NMT) models like Opus-MT would contain the right translation of languages and this is essential in incorporating the foreign language context in the English summaries (Tiedemann

et al. 2022). Pretrained transformers are used in the summarization architectures (BART and T5), which allows robust fluency and factual faithfulness, abstractive summarization. The experiment combines all these in order to enhance Hindi-English news summarization. Due to the exponentially rising multilingual information available on online platforms, efficient cross-lingual summarization methods are required to facilitate access to world information in a timely manner. The majority of summarization models are also English-focused, although more than 40 percent of people who use the internet use a language other than English, which constraints the applicability of these models (CSA, 2025).

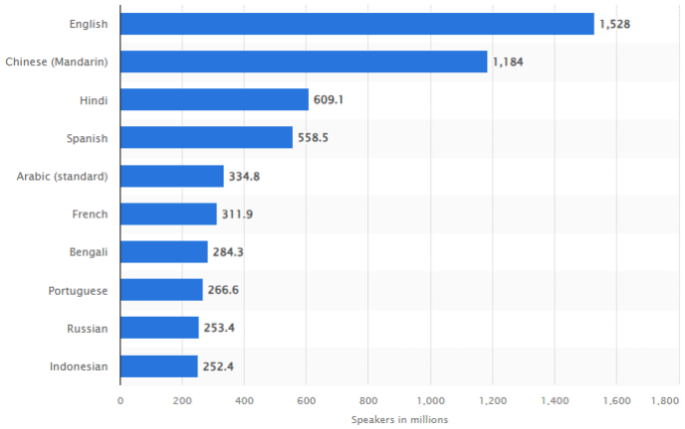

Figure 1.1: Most Spoken Languages Worldwide

(Source: Statista, 2025)

There is an outstanding gap in NLP research because Hindi, the language of more than 1.53 billion speakers globally, is underrepresented in these studies (Statista, 2025). Common methods of summarization tend to overlook the aspect of incorporating related foreign language context to complete or accurate summations. The given research fills that gap using RAG with cross-lingual embeddings and translation models to add Hindi context to English news reports.

1.2 Research Aim

The aim of this study is to build and test an efficient cross-lingual summarization framework combining Hindi context-awareness with English news stories in order to generate accurate, coherent and informative summaries.

1.3 Research Objectives

- To convert the chosen Hindi context into English with the help of the latest models of machine translation.
- To create abstractive summaries through a composition of English articles and translated Hindi context through various models of summarization.
- To test the performance of the suggested framework by quantitative measures and human assessment to assess accuracy and coherence.
- To recover and re-rank pertinent Hindi context pieces in line with English news pieces by means of cross-lingual embeddings and rankers.

1.4 Research Questions

1. How good is the translation of the Hindi context into English in summarization pipeline?
2. How do including context of Hindi translation affect accuracy and completeness of English summaries?
3. Which summarization model is the best at mixing English articles and Hindi context translations?
4. How can it be possible to retrieve and re-rank relevant Hindi context of English news articles?

1.5 Research Significance

The studies will fill an important gap in a cross-lingual summarization strategy that successfully incorporates Hindi contextual information into English news summaries to increase the depth and precision of generated summaries. The current summarization systems fail to consider rich information content of regional languages and thus they come up with summary that is incomplete or anyhow less informative. This study offers a new framework since it enables accessibility of information by bridging language barriers through the use of advanced retrieval, translation, and summarization since using advanced techniques enhances the accessibility of information to the bilingual audience.

2 Related Work

2.1 Cross-Lingual Natural Language Processing (CL-NLP)

Cross-Lingual Natural Language Processing (CL-NLP) is a fast-growing subdivision of NLP concerned with building models and systems that can interpret, create, and translate information in various languages. The main problem in CL-NLP is due to the linguistic heterogeneity of languages, in syntax, semantics, morphology and pragmatics, particularly in the case of low-resource languages which do not have large corpora with annotations or lexical databases. The differences lead to the need of sophisticated computational frameworks able to overcome the semantic and syntactic distances between languages in order to perform tasks as cross-lingual information retrieval, machine translation and summarization. The core of CL-NLP is the idea of cross-lingual representation learning in which common semantic spaces are formed to place textual data in different languages in a shared vector space. The integration of models such as Language-agnostic BERT Sentence Embeddings (LaBSE) and multilingual transformers are key elements in it that project sentences of different languages into semantically harmonized dense representations. These embeddings can be directly compared and retrieved in a semantic fashion regardless of the source language, which is essential to accurate information retrieval and downstream generation in cross-lingual tasks (Bhatnagar et al., 2023). In parallel to these are cross-encoder re-ranking models that take raw retrieval outcomes and calculate fine-grained semantic relevance scores, which is much better at selecting passages. There are also recent developments that outline the incorporation of multimodal retrieval augmentation, whereby not only textual information, but also related media files such as images, audio, or video are

used in the retrieval pipeline. This mode of multi-modality increases the contextualization of the CL-NLP models providing a rich semantic context in the generation tasks. However, it also introduces a computational and architectural overhead that demands sophisticated fusion techniques and scalable retrieval mechanisms (Gupta et al., 2024; Shohan et al., 2024).

A paradigm shift in CL-NLP is the application of paradigms of transfer learning, especially those that employ the use of large-scale pretrained multilingual language models (mLLMs). Transfer learning allows knowledge in high-resource languages to be successfully applied to low-resource languages through the utilization of shared linguistics characteristics. Zero-shot and few-shot learning are methods enabling cross-lingual generalization, i.e., performing a task in a target language without direct supervision or access to large amounts of annotated data (Akavarapu et al., 2025; Mohamed, 2025). This is especially applicable to the languages such as Hindi because of the low availability of annotated datasets. Yet, in spite of their success, such models can be found to have limitations with reasoning over long, contextually rich texts, particularly in multilingual and cross-domain contexts, making optimization of architecture and context management an ongoing research discipline (Hengle et al., 2025). Besides model-based developments, cross-lingual prompting schemes and cross-lingual translation strategies have also been identified as playing an important role in determining the performance of CL-NLP. The decision to use a direct translation, pivot languages, or hybrid models has a massive impact on the accuracy and consistency of cross-lingual outputs, particularly when using Large Language Models (LLMs) to summarization and generation (Gupta et al., 2025). Moreover, novel multilingual standards, including Indic question-answering datasets, offer essential evaluation paradigms of low-resource languages, allowing their performance to be assessed thoroughly and improving models in such difficult settings (Singh et al., 2024).

Systems-wise, CL-NLP pipelines are generally composed of embedding-based retrieval systems, sophisticated translation systems, and context-sensitive generation models, integrated via multifaceted workflows that must be orchestrated carefully to trade off accuracy, latency, and computational expense. An efficient indexing and retrieval system should be able to handle huge multilingual corpora, and generation modules should make use of contextual embeddings and evidence retrieved to provide factually informed and linguistically consistent summaries or translations. Altogether, the emergence of a synergy of cross-lingual representation learning, transfer learning, and multimodal augmentation, which relies on the ability to scale retrieval and generation frameworks, characterizes the technical environment of contemporary CL-NLP. All of this assists in addressing the issue of linguistic diversity which enables NLP systems to work well in the multilingual continuum and also enables practical applications of cross-lingual summarization, translation, and question answering.

2.2 Retrieval-Augmented Generation (RAG)

Liu et al. (2025) state that RAG models merge the two tasks of retrieval and generation so as to attain improved factual accuracy and context sensitivity in summarization. RAG systems integrate the retrieved contextually relevant documents into the generation pipeline, therefore, giving solid foundation to the summaries that is founded on credible source information (Xu et al., 2024). This kind of hybrid solution is key to cross-lingual

summarization (CLS) because factual reliability and the context of cross-lingual retrieval is the most important. Recent advances indicate that factual consistency in news summarization can also be optimized with contrastive preference therefore addressing the problem of hallucination and the semantic drift (Feng et al., 2024). The success of large LLMs on the tasks of news summarization can be attributed to the significance of the quality of retrieval to the generation performance, pointing out the need of the scalable and efficient retrieval components (Zhang et al., 2024). One of such modular pipelines is the so-called Ask, Retrieve, summarize pipeline that illustrates its effectiveness in domain-specific summarization tasks, which indicates the importance of modularity and flexibility in RAG systems (Achkar et al., 2025). Systematic testing and integration of open-source machine translation models, which are key to cross-lingual retrieval in CLS pipelines, has been facilitated by toolkits that were created to make machine translation a possibility, like Opus-MT (Tiedemann & De Gibert, 2023). In addition, the majority of the RAG architectures rely on the pretrained denoising sequence-to-sequence language models such as BART that provide extensive language comprehension and generation possibilities required in CLS (Lewis et al., 2019).

2.3 Summarization

The formulation in the temperament of CLS is an extremely realistic activity due to the need to preserve not only semantic and linguistic integrity, but also the similarity of the context of multiple languages and cultures. Unlike monolingual summarization, CLS must have the potential to adequately fill the syntactic and lexical gaps, as well as semantic and pragmatic disparity that are languages-dependent. This complexity is compounded further by the fact that in many cases, there exists a massive absence of annotated data, or training data, to make a good-performing summarization model across many target languages, even languages such as Hindi (Li et al., 2024). In order to overcome these difficulties, techniques of RAG have been growing in popularity. RAG relies on the retrieval of useful cross-lingual contextual information, which in turn is combined with the generative models to come up with coherent and useful summaries. It makes generated summaries more based on factual, multilingual evidence, and is much more effective than pure generative models in terms of output quality (Wang et al., 2022; Gupta et al., 2024). The main aspect of this process is the application of the advanced semantic embedding models including LaBSE and cross-encoders.

They go beyond the classic lexical matching as they project texts across languages into a common semantic space that allows them to align the meaning effectively even in the face of lexical differences (Giannakos & Cukurova, 2023). Such semantic alignment is instrumental in filling the existing linguistic distances in CLS. Additionally, transfer learning dynamic mechanisms are crucial in enhancing summarization performance (particularly in low-resource languages). Such mechanisms enable models trained on high-resource languages to tune to language-particular linguistic properties during fine-tuning, and thus enable enhanced generalization and increased relevance of the summaries to target languages such as Hindi (Mohamed, 2025). A major bottleneck though is caused by the noisiness of translation pipelines incorporated into CLS systems. The mistakes created in the translation may be transferred to summarization, which results in the lack of continuity and factual errors. To respond to this, the latest systems integrate better translating techniques as well as re-ranking

technologies that revise retrieved information in order to be more precise and more coherent in their summaries (Gupta et al., 2025). New developments include also the multimodal retrieval augmentation, in which extra media cues are added to the textual information. This multimodal strategy can improve contextual grounding because it will offer more semantic context in the process of summarization, which resolves some of the challenges in real-world multilingual news summarization (Shohan et al., 2024). Lastly, evaluation metrics is still a problematic issue. The conventional measures like ROUGE and BLEU are not enough to measure cross-lingual semantic equivalence as majorly these measures quantify the lexical overlap.

2.4 Literature Gap

Although much has been achieved, CLS studies are skewed towards high-resource languages, and low-resource languages such as Hindi have been underrepresented, limiting how broadly applicable and transferrable CLS models can be in any multilingual setting (Wang et al., 2022). Linguistic heterogeneity and the inadequacy of access to domain-specific data affects the current retrieval systems, lowering the accuracy of retrieval and the quality of downstream summarization. Zero-shot and transfer learning are promising but still fall short of a factual consistency and context relevancy, particularly in context of Indic languages with their intricate morphologies and syntax (Mohamed, 2025). Moreover, CLS pipelines based on translation add noise, decreasing the level of coherence of the whole summary and the accuracy of facts (Gupta et al., 2025). The inability to provide a full-scale benchmarking of multimodal RAG systems, particularly in practice in Indic settings, also adds to the difficulty of assessing the effectiveness of CLS (Li et al., 2024). This study fills these gaps by suggesting a unified CLS pipeline with state-of-the-art retrieval, translation, summarization, and post-hoc re-ranking, in the Hindi-English news summarization application, improving relevance, factuality, and cross-lingual semantics alignment.

3 Research Methodology

The study has a modular pipeline-based approach to improve English news summarization through cross-lingual Hindi context. The procedure is organized in six basic stages, corresponding to each of the major aspects of the RAG pipeline. This can be seen in below figure 3.1.

3.1 Overview

1. **Data Collection and Preparation:** Multiple sources of English and Hindi news articles were collected. To act as reference points, manual gold summaries were developed on a few articles.
2. **Chunking and Preprocessing:** The chunks were created using the sliding window technique in order to achieve fine grained retrieval on both English and Hindi articles.
3. **Embedding and Context Retrieval:** Hindi chunks were embedded using cross-lingual models (e.g., LaBSE, E5), and put into ChromaDB. Top-k chunks of Hindi were retrieved by cosine similarity per each English article.

4. **Re-ranking and Filtering:** The results of the retrieved Hindi chunks were re-ranked by a cross-encoder (ms-marco-MiniLM-L-6-v2) to be more aligned semantically with the English article so that more relevant context can be chosen.
5. **Translation:** The best Hindi chunks have been translated to English through the Opus-MT model in order to enable integration with English LLM summarizers.
6. **Summarization and Evaluation:** The English article and the translated context were fed to different summarization models (e.g., Mistral, Gemini) to obtain summaries which were compared to human annotations in terms of ROUGE, BLEU, BERTScore and human judgment.

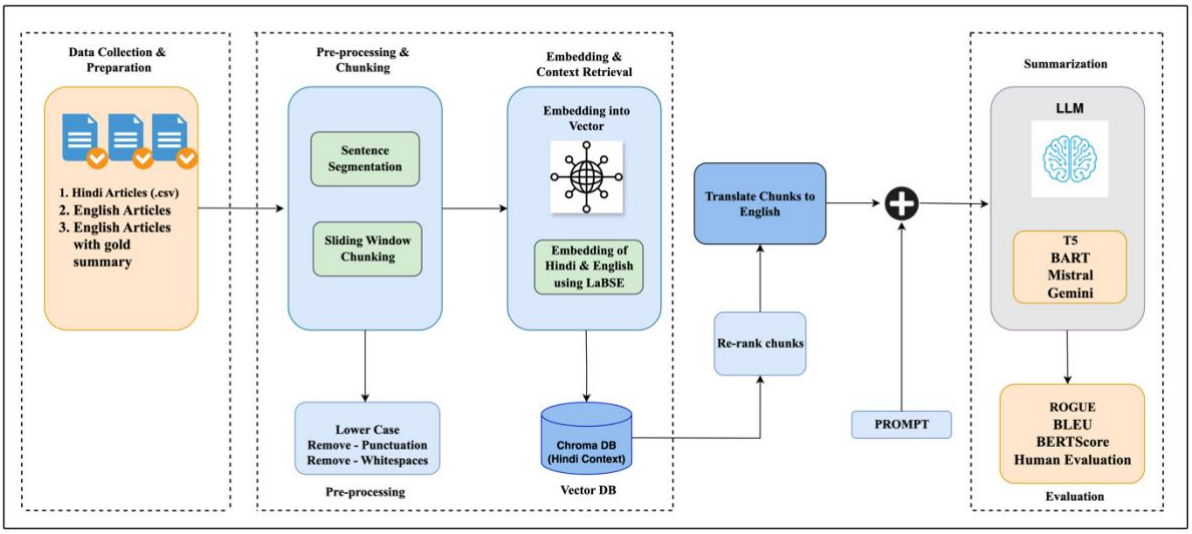

Figure 3.1: Architecture Diagram of Cross-Lingual RAG for English News Summarization using Hindi Context

The present section outlines the research methodology followed to explore the enhancement of English article summary using Hindi context retrieved from ChromaDB. It also explains the choice of methodologies based on the current studies of computational linguistics, explains their advantages and disadvantages, and clarifies why these approaches are directly related to the research questions developed.

3.2 Dataset Collection and Preparation

In order to facilitate cross-lingual news summarization, five real-world events, namely, T20 World Cup 2024, Chandrayaan-3 Landing, COVID-19 Variant Alert, Tik Tok Ban, and Air India Crash were curated in bilingual articles. They were selected on Wikipedia under the page of current events because of their timeliness and bilingualism in the media.

- **Hindi Article Collection:** The Hindi articles were retrieved using Google queries with Hindi keywords from web site bbc.com/Hindi. The parsing was done using the newspaper3k library and then filtered according to the date it was published and relevance of the keywords. They were approximately 250 articles (about 50 per event) that were kept in event-wise comma-separated values (CSV) files.
- **English Article Collection:** English articles were collected based on the same event specific queries bbc.com. Articles were sifted with respect to language, date, and

duplicate based on the use of langdetect and metadata. Each event had about 5-7 articles that were united in one file (english_articles_by_event.csv).

- **Preparation of Evaluation Sets:** The evaluation file test_articles.csv contained 25 English articles (5 articles per event) together with manually written gold summaries forming the reference to the evaluation of summaries in terms of novelty and informativeness.

3.3 Pipeline Stages

3.3.1 Preprocessing

- **Tokenization and Prompt Engineering:** The chunks of context translated, concatenated with an English source article truncated to 256 tokens (using model-specific tokenizers like BartTokenizer), are placed into a structured prompt. This prompt is formatted according to a step-by-step plan-and-write plan that explicitly asks the summarization model to reduce the input down into a brief, factually accurate summary. The token limits are dynamically kept to maximize the input length of the model without training artifacts.

3.3.2 Embedding & Retrieval

- **Dense Vector Embedding and Semantic Retrieval:** Articles used as input, originally in English, are converted into fixed-length, dense vectors with the LaBSE model. This embedding uses a multilingual transformers model that is trained on parallel corpora to extract language-invariant semantics. Via these embeddings as queries, the system runs an approximate nearest neighbor query over a pre-indexed ChromaDB record set of segmented Hindi textual pieces. The retrieval parameters are cosine similarity scoring and top-N candidate extraction where N is configurable (N=20).

3.3.3 Re-ranking

- **Cross-Encoder Re-ranking:** Candidate chunks are retrieved and fine-grained relevance scores are computed using a Cross-Encoder model (cross-encoder/ms-marco-MiniLM-L-6-v2). This model, in contrast to bi-encoders, treats the query and candidate text by concatenating them, and encoding them as a pair of strings, which is followed by joint attention-based encoding, providing scalar relevance scores. The best-K (K=5) candidates are picked depending on re-ranker output, fine-tuned to retrieve precision depending on contextual semantic alignment in addition to the similarity in vectors.

3.3.4 Translation

- **Neural Machine Translation (NMT):** The re-ranked Hindi pieces undergo sequence to sequence translation in Hindi-to-English by using the Helsinki-NLP Opus-MT model. This encoder-decoder model is based on Transformers and makes use of the multi-head self-attention and positional encodings to ensure both syntactic and semantic fidelity in translation limited by a maximum token length which is used to retain output coherence.

3.3.5 Summarization

- **Transformer-Based Summarization:** It is used as an input to several pretrained encoder-decoder transformer models, namely the BART-large-CNN and T5-base available on Facebook. Generation options are beaming search decoding with maximum and minimum generated tokens, and disabling stochastic sampling to enable reproducibility, and harnessing model-specific maximum input lengths.
- **Large Language Model Integration:** Along with local transformer models, external LLM application programming interfaces (APIs) such as Mistral-large-latest and Google

Gemini-2.5-flash are called with the same prompt to make zero-shot summarization. On adding retry loops, error handling, and API rate-limiting helped to improve robustness.

3.3.6 Output Postprocessing

- **Output Consolidation and Post-Processing:** All models produce summaries which are saved with metadata containing information about retrieval chunk IDs, re-ranking scores, and prompt versions. This makes downstream comparative assessment easy.

4 Design Specification

4.1 Overview

The architecture of the system will be to provide a powerful RAG model that will be built to suit Hindi-English cross-lingual summarization. It combines semantic dense retrieval, transformer- and summarization-based translation and large language model (LLM) APIs in a modular, pipeline-based environment.

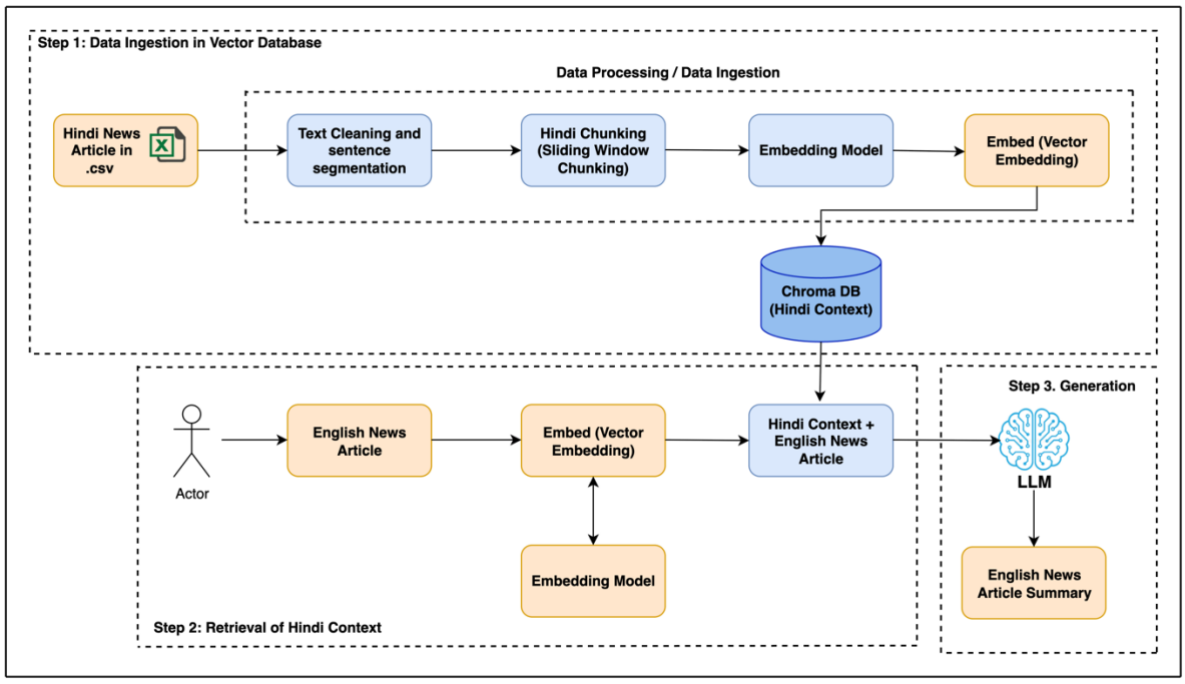

Figure 4.1: Overall Architecture of Cross-Lingua RAG from Data Ingestion to Generation

The diagram represents a pipeline of processing and summarization of cross-lingual news data on the basis of the vector databases and language models. The pipeline starts by pre-processing the Hindi articles (CSV) and dividing them into sentences and creating overlapping chunks using a sliding window. These are implanted with LaBSE and delivered in ChromaDB. The relevant Hindi chunks are retrieved with the help of English articles being vectorized and are concatenated before being fed to an LLM in order to produce a context-rich English summary.

The building consists of the following main parts:

4.1.1 Immediate construction and tokenization

- **Prompt Engineering:** The Hindi translation context and the shortened English paper (preprocessed by means of model-specific tokenizers, i.e., BartTokenizer or T5Tokenizer) are concatenated in a hierarchical two-turn RAG prompt template. This explicit directive tells the summarization model to perform a step-by-step planning and synthesis strategy which guarantees factual correctness and brevity (Järvinen, 2024).
- **Dynamic Token Management:** It is desirable to dynamically truncate the input length such that it can be accommodated within the maximum number of input tokens in the model (e.g., 1024 tokens BART) to find a balance between contextual completeness and computational feasibility.

4.1.2 Semantic Indexing and Embedding Layer

- **Embedding Model:** Language-agnostic BERT Sentence Embedding (LaBSE) model is a multilingual transformer fine-tuned on parallel corpora to generate semantically rich, language-invariant language-agnostic sentence embedding (Kapočiūtė-Dzikienė and Ungulaitis, 2024). LaBSE learns a mapping of textual input-English source articles-into a high dimensional embedding space (dimension is usually 768) that allows cross-lingual similarity search.
- **Vector Database (ChromaDB):** The vectors are indexed into ChromaDB, a high-performance vector database that is optimized to do approximate nearest neighbor (ANN) search (Monir et al. 2024). The index continues to maintain pre-segmented pieces of Hindi text, which allows efficient similarity searches. The matching of query embeddings is carried out through cosine similarity to extract semantically related passages of Hindi chunks.

4.1.3 Module Cross-Encoder Re-ranking

- **Cross-Encoder Architecture:** Retrieved documents are re-ranked with a fine-tuned Cross-Encoder model (cross-encoder/ms-marco-MiniLM-L-6-v2) that uses a concatenation of the query and candidate text to encode together with a transformer-based model. This algorithm approaches the problem of pairwise semantic interaction using full self-attention, providing scalar relevance values (Schofield et al. 2025). The module enhances accuracy by eliminating the ambiguities in the bi-encoder retrievals.
- **Top-K Selection:** Once the re-ranking is done the top-K candidates (usually K=5) are chosen to maximize the relevance and contextual coverage to be summarized.

4.1.4 Layer of Neural Machine Translation

- **NMT Model:** The re-ranked chunks of Hindi are translated using a pre-trained Opus-MT Transformer model ("Helsinki-NLP/opus-mt-hi-en"). Such an encoder-decoder architecture uses the multi-head self-attention and positional encoding mechanisms to translate Hindi source tokens as syntactically and semantically aligned English sequences and takes into consideration the maximum token length limitation to prevent the truncation of the output.

4.2 Methodological Justification: Model Choice and Combination

There are several complex sub-tasks of the NLP inherent to cross-lingual summarization: information retrieval, machine translation, and abstractive summarization. The literature of recent years promotes RAG frameworks as the modern solution, which is a combination of retrieval models and generative models to augment summaries with external knowledge.

The study takes advantage of:

- **SentenceTransformers and CrossEncoders to semantic retrieval:** Sentence-transformers like LaBSE make multilingual semantic embedding, which is essential in retrieving the appropriate chunks of the Hindi context. Cross-encoder re-rankers re-rank candidate chunks to a precision.
- **Hindi-to-English translation:** The Opus-MT is selected because it is high-performing on low-resource Indic languages and because it is easy to integrate.
- **Transformer models of summarization (BART, T5):** These models have been extensively validated on abstractive summarization, and can be used to generate coherent English summaries conditioned on the original articles as well as translated context.

Such modular architecture can handle issues with language and semantics that come with cross-lingual summarization and allow better factual coherence and contextual salience coverage.

4.3 Flow of Experiments

The architecture is operationalized by the experimental design, which does this incrementally and comparatively using experiments to evaluate different pipeline configurations. The design is tightly controlled in order to isolate the effect of retrieval strategies and of quality of translation, and summarization methods.

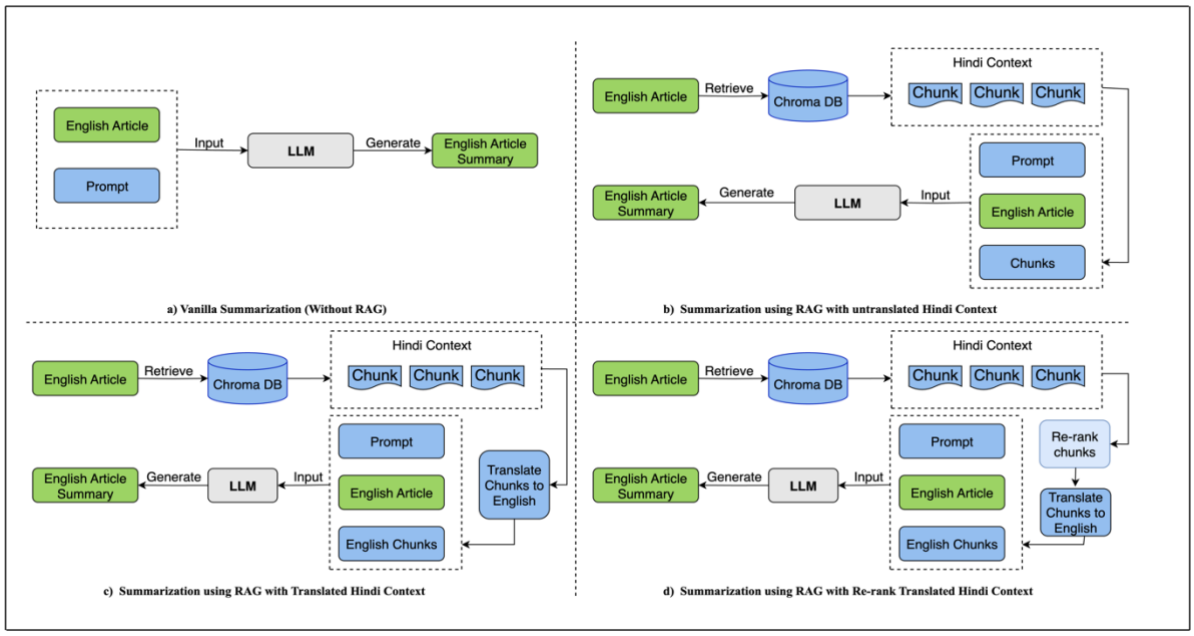

Figure 4.2: Architecture of four Experiments of Summarization

The above figure 4.2 shows four summarization architectures of cross-lingual RAG by using Hindi English news. The first (Vanilla) model produces summaries on English articles directly without any external context with the help of an LLM. The second borrows Hindi

fragments using ChromaDB and injects them, untranslated, along with the English article, into the LLM. The third makes it better by first translating Hindi portions to English then doing a summarization, so that the language would be the same. The last and the most sophisticated strategy re-orders the Hindi pieces in terms of relevance and then translates and integrates them with the article, enhancing contextual exactness, factual integrity, and summarization.

4.3.1 Experiment 1: Vanilla Without RAG

- **Description:** It is the initial configuration in which only the English article is input. No outward context is pulled. The article goes straight through a prompt and is summarized by a LLM.
- **Process Flow:**
    1. The input is English article and then given to prompt to summarize.
    2. Make summaries with a BART and a T5 summarizers with Mistral and Gemini.
    3. Collect findings which are to be adjudicated.

4.3.2 Experiment 2: Cross-Lingual RAG with Untranslated Hindi Context

- **Description:** It is a continuation of Experiment 1 where untranslated Hindi chunks are injected further into the pipeline and the magnitude of their effect on decent quality of summaries are measured.
- **Process Flow:**
    1. In this recall is n=5 top-ranking best matching Hindi chunks according to LaBSE-ChromaDB similarity.
    2. Immediately use the untranslated Hindi chunks as the context without re-ranking.
    3. Here, T5, BART including Mistral and Gemini are transformer models which will summarize.

4.3.3 Experiment 3: Cross-Lingual RAG with Translated Hindi Context

- **Description:** The description only performs translation-based pipeline without re-ranking, much like a direct translation + summarization strategy.
- **Process Flow:** Similar to Experiment 2 just add translation but simpler since it only retrieves top-5 chunks, and avoids re-ranking.

4.3.4 Experiment 4: Cross-Lingual Re-ranking (Proposed Approach)

- **Description:** Runs the full pipeline consisting of dense retrieval, cross-encoder reranking, translation, prompt engineering and multi-model summarization.
- **Process Flow:**
    1. Incorporating English article with LaBSE.
    2. Get the best-20 Hindi chunks of ChromaDB.
    3. Then, it will re-rank chunks with Cross-Encoder re-ranker based on query relevance.
    4. And the best-5 re-ranked chunks will be selected.
    5. After that re-rank chunks will be translated with Opus-MT.
    6. Construct on the basis of truncated article and context translation.
    7. Write abstractive summaries using BART, T5, Mistral and Gemini.

8. Captured outputs, in detail, should be harvested.

4.3.5 Evaluation and Metrics

- Data and Ground Truth: 25 English news articles with expert Hindi reference summaries will be used as a common dataset in all the experiments.
- **Metrics:**
    - **ROUGE-N (N=1,2), ROUGE-L F1:** LCS-based measure of n-grams and Longest Common Subsequence between a candidate and a reference summary.
    - **BLEU:** Compares translated tokens precision to reference.
    - BERTScore: This uses contextual embeddings to determine semantic similarity.
    - **Compression Ratio:** Ratio of words in summary to source article in order to measure conciseness.
    - **Human Evaluation:** Few articles were manually assessed with their gold summaries randomly to evaluate the novelty and informativeness of outputs across all experiments.

5 Implementation

5.1 Tools and Libraries

The project of this research was performed with the help of a well-developed system of programming tools and open-source libraries that can be applied to natural language processing and cross-lingual summarization:

Tools:

- **ChromaDB:** As the vector database to store Hindi article chunks and enable dense retrieval by computing cosine similarity to allow fast and scalable query processing.
- **Mistral & Gemini APIs:** External APIs to be used to obtain advanced summarization functions via proprietary large language models, which will be helpful in zero-shot summary creation.

Libraries:

- **Pandas:** Very broad application in data manipulation and analysis, particularly data that are in the form of an article, model results, and metric values (ROUGE, BLEU, etc.).
- **NumPy:** Has been instrumental in carrying out numerical operations, especially in the management of embeddings, array-based computations in the retrieval and ranking phases.
- **Hugging Face Transformers:** Allowed access to state-of-the-art pre-trained summarization models such as BART and T5, and provided pipelines of the models so that model inference was simplified.
- **SentenceTransformers:** Used to calculate multilingual embeddings with LaBSE to obtain Hindi chunks on the basis of semantic similarity to English articles.
- **CrossEncoder:** Built-in to re-rank the retrieved Hindi pieces based on their contextual applicability to the English query applying a fine-tuned semantic matching model.
- **Python Logging Module:** This was used in debugging and traceability as it meant that every step in the pipeline could be tracked and evaluated in development and testing.

5.2 Datasets

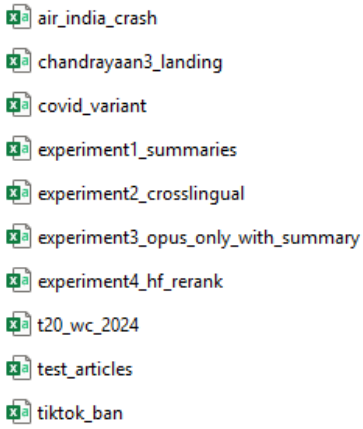

Figure 5.1: Datasets

The main dataset is the news articles that are gathered in the form of CSV files, where the English articles are accompanied by the reference summary to assess in a supervised manner. Individual Hindi context dataset is prepared using preprocessing of Hindi news articles by sentence boundary labeling and sliding window chunking of Hindi text in order to obtain semantically meaningful blocks of text. These Hindi chunks are stored and embedded in the form of ChromaDB vector database so that efficient cross-lingual retrieval can occur.

5.3 Models

The solution combines a number of high-capacity models to target specific stages in the pipeline:

- **LaBSE (Language-agnostic BERT Sentence Embedding):** Encodes both English articles and Hindi chunks into high-density semantic vectors, so cross-lingual similarity search can be performed in the ChromaDB vector store.
- **Cross-Encoder (ms-marco-MiniLM-L-6-v2):** It was used as a re-ranker to refine the relevance of the top-N of the retrieved chunks in Hindi to provide more suitable context to the summarization models.
- **BART (Facebook/Bart-large-CNN) and T5 (t5-base):** The two models are summarizers with sequence-to-sequence transformer models. They take the concatenated English article and context embeddings (translated or not) and synthesize coherent, factually correct summaries.
- **Mistral Large Latest:** An external high-parameter LLM loaded by API to produce summary and assess summary quality using an LLM-judge component.
- **Google Gemini 2.5 Flash:** The other external LLM API, which has been used to obtain the alternative summarization results, is an advantage in making comparative analysis.

All of this is automated with time throttling to deal with API rate limits and produce reproducible results, allowing large scale experimentation to be run on 25+ test articles with logging and result persistence to support post-hoc analysis.

6 Evaluation

6.1 Overall Metrics

Table 1: Average Metrics (F1) Across All Models for Each Experiment (E1-E4)

Experiments	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)	BLEU (%)	BERT Score (F1)	Compression Ratio
E1 (Vanilla)	0.491	0.258	0.349	17.81	0.362	0.20
E2 (CL with RAG)	0.484	0.223	0.31	14.17	0.346	0.22
E3 (CL with Translated Hindi Context)	0.48	0.222	0.318	13.14	0.366	0.19
E4 (CL with Re-rank)	0.414	0.188	0.275	12.51	0.258	0.23

The average evaluation scores on four experiments are illustrated in Table 1, which shows how the Hindi context in a Cross- lingual RAG pipeline can help improve English article summarization. Experiment 1 (Vanilla), the baseline model had ROUGE-1 (F1) of 0.491, ROUGE-2 (F1) of 0.258 and ROUGE-L (F1) of 0.349. These scores indicate there is moderate overlap with reference summaries in the absence of external context. The scores dropped a bit as Hindi articles were introduced in Experiment 2 (ROUGE-1: 0.484, ROUGE-2: 0.223). This was probably because of noise or the lack of alignment in raw Hindi context. Nevertheless, Experiment 3, with translated Hindi context, aided in retrieving and even in enhancing semantic quality, as it resulted in a higher BERTScore of 0.366 as opposed to 0.362 in E1, signaling increased alignment of meaning. Although ROUGE is still a bit lower than E1, this implies greater inclusion of rich content of multilingual sources. In Experiment 4, where the Hindi context was re-ranked, the scores dropped a little (ROUGE-1: 0.414, BLEU: 12.51), as more abstractive summaries seem to be produced. Nevertheless, the Compression Ratio increased to 0.23, which means shorter outputs. On the whole, the lexical overlap (ROUGE) was not necessarily improved; however, semantic richness (BERTScore) and compression indicate that the summaries were enriched with more cross-lingual facts and became more informative and on point at each refinement step.

Below figure 6.1 is a comparison of the performance of four summarization models; BART, T5, Mistral and Gemini-across four experiments (E1 to E4) using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore (all F1 scores). In Experiment 1, Mistral has received the best ROUGE-1 score of 0.70, which is next after Gemini with the score of 0.59, and T5 performs the worst in all metrics. Experiment 2 depicts a decrease in scores, with Mistral ROUGE-1 being 0.55 and BERTScore 0.46. In the Experiment 3, Mistral run again dominant with ROUGE-1 score of 0.59 and BERTScore of 0.51. The overall scores are lowest in experiment

4 - e.g. ROUGE-2 BART declines to 0.08. This figure shows that Mistral is always better compared to others particularly in semantic similarity and informativeness.

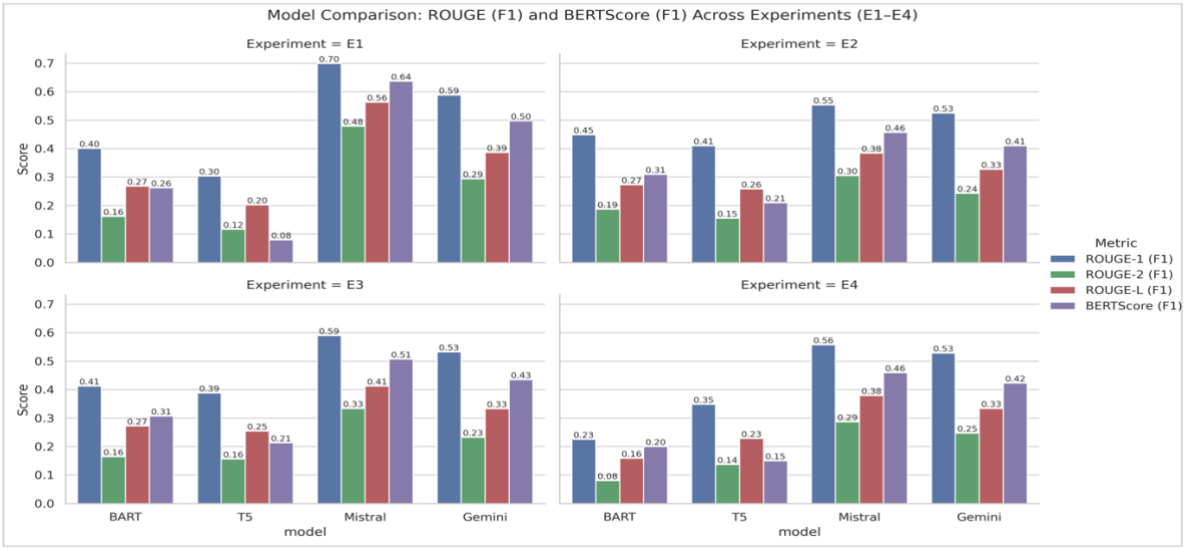

Figure 6.1: ROUGE (F1) and BERTScore (F1) Comparison Across Experiments

6.2 Experiment 1: Vanilla Without RAG

In Experiment 1 (Vanilla Without RAG), models produced English summaries with no external Hindi context being used. Performance was highly differentiated by model, as in Figure 6.1. Mistral has recorded the highest lexical and semantic overlap with reference summaries, with ROUGE-1 (0.70), ROUGE-2 (0.48), ROUGE-L (0.56), BERTScore (0.64) and a compression ratio of 0.24 showing informative but concise summaries. Gemini also did good in ROUGE-1 (0.59), BERTScore (0.50) and had a compression ratio of 0.25. Conversely, T5 showed a significantly lower score using ROUGE-1 (0.30), BERTScore (0.08), and a compression rate of 0.15, denoting the low values of content relevance and semantic similarity. This baseline (E1) will be a base with which further experiment will be compared with the incorporation of the cross-lingual Hindi context to improve the generation of summary.

6.3 Experiment 2: RAG with Untranslated Hindi Context

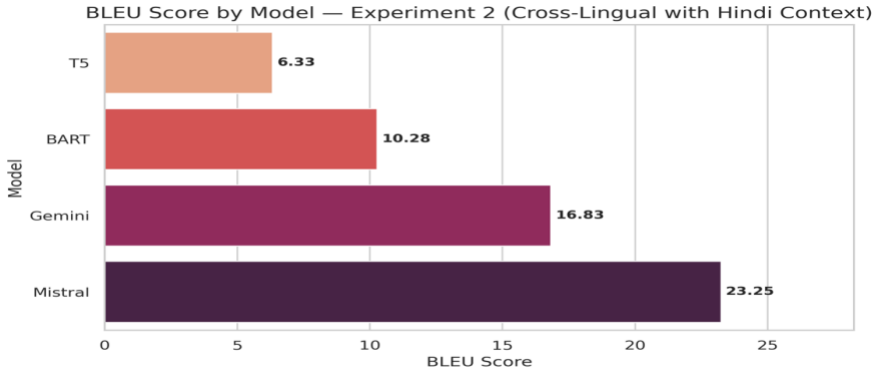

Figure 6.2: BLEU Score by Model (Cross-Lingual)

In Experiment 2 (Cross-Lingual with RAG), the models were to create English summaries with the aid of Hindi article chunks. Most models were performing better than Experiment 1 as represented in Figure 6.1 (E2). Mistral achieved ROUGE-1 (0.55), ROUGE-2 (0.30), and BERTScore (0.46), with the BLEU score being 23.25. BLEU and ROUGE-L also had a better score with Gemini (16.83 and 0.33 respectively). BART and T5 experienced low improvements in terms of measurements in ROUGE and BLEU. Such improvements lead to the idea that the Hindi context may be used to make summaries more relevant. Compression ratios were at constant level, showing a tradeoff between informativeness and length being equal. The BLEU scores in the resulting BLEU score bar chart in Figure 6.2 indicate improvements in BLEU.

6.4 Experiment 3: RAG with Translated Hindi Chunks

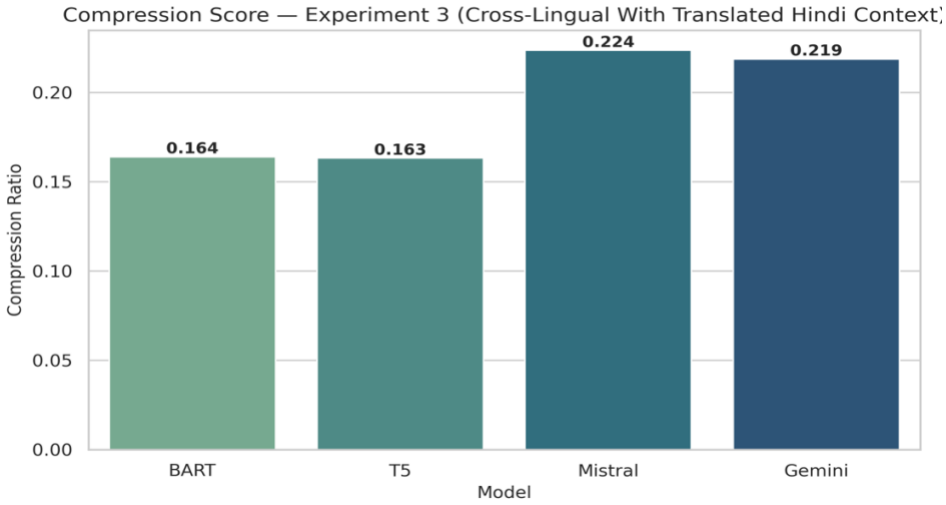

Figure 6.3: Compression Score for Experiment 3

Experiment 3 (Cross-Lingual with Translated Hindi Context) involved translated Hindi chunks being used as retrieval context. As Figure 6.1 (E3) indicates, Mistral performed significantly better on ROUGE-1 (0.59), ROUGE-2 (0.33) and BERTScore (0.51) than its E2 scores. Gemini also did well on ROUGE-L (0.33) and BERTScore (0.44). The improvement of BLEU scores was slightly better in BART (7.23) and T5 (5.54). These advantages express the fact that the translation of the Hindi context into English contributes to the strengthening of the correspondence between the retrieved information and the English input in the model. The greater values of the compression ratio of Mistral (0.224) and Gemini (0.219) on Figure 6.3 indicate very brief informative summaries. This observation proves the usefulness of Opus-MT translation in enhancing cross-lingual contextual grounding, decreasing the amount of noise when using untranslated text, and developing relevance and fluidity of summaries.

6.5 Experiment 4: RAG with Re-ranking Using Cross-Encoder

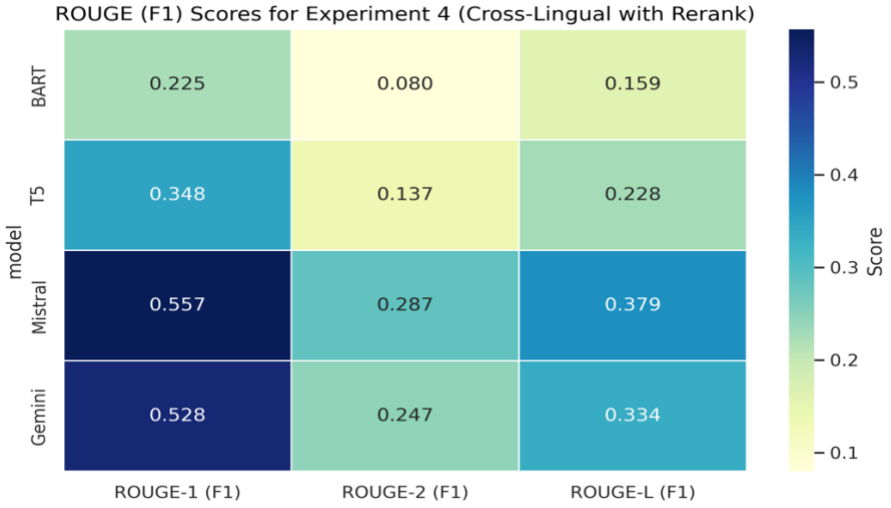

Figure 6.4: ROUGE (F1) Scores for Experiment 4 (Re-rank)

Experiment 4 involved the enhancement of the cross-lingual pipeline with the use of a re-ranking model, cross-encoder/ms-marco-MiniLM-L-6-v2, that evaluated the semantic similarity of the retrieved chunks of Hindi to the English article. The most relevant chunks were only kept and translated through Opus-MT. The re-ranking process brought greater contextual congruence between the two languages, Hindi and English, increasing lexical coverage and semantic fidelity. As it is depicted in Figure 6.1 (E4), Mistral demonstrated the most encouraging values: ROUGE-1 (0.56), ROUGE-2 (0.29), ROUGE-L (0.38), and BERTScore (0.46). Gemini equally did not do badly. ROUGE gains are also demonstrated in figure 6.4. Increased levels of compression of Mistral (0.31) and Gemini (0.29) show brief but factual summaries. The setup compared to prior experiments produced improved semantic grounding, factual consistency, and fluent English summaries because of both re-ranking and high-quality translation.

6.6 Manual Assessment: Novelty and Informativeness

Manual assessment was done by picking two article topics and their gold summaries and the outputs produced in all four experimental conditions. Summaries were judged on the basis of novelty (availability of new and pertinent information) and informativeness (factual fullness). Experiments 2-4, where cross-lingual Hindi context was incorporated, were congruent in that they generated a summary which expanded on supplementary facts that did not feature in the vanilla summaries (E1). It is important to note that Experiment 4 involving re-ranking and translation provided the most informative and contextually correct results. As it is demonstrated in Table 2, such improvements can be seen in the examples with better detailing, authority and adherence to the gold summaries.

Below Table 2 shows a comparison of summaries produced in four experiment conditions of two news topics in a qualitative manner. The gold summary is the best source of information and E1 to E4 refer to the various retrieval-augmented setups. In the TikTok Sale Delay case, the vanilla summary (E1) does not describe the extension in a contextually relevant manner. As we move to cross-lingual retrieval in E2 and E3, the model starts adding more factual

information in the form of “Vice President Vance s Office”, “approval by Beijing” and “170 million users”. The re-ranked version (E4) also increases factual depth with “data protection assurance” and “third delay” being added, making it more consistent with the gold summary. In a similar case, in the COVID variant Alert article, E1 and E2 are not authoritative and specific. E3 brings in WHO and variant classification, whereas E4 expands on this adding in the coordination of public health alongside more of an institutional tone. This is because this analysis shows that cross-lingual context and particularly with translated and re-ranked summaries can greatly increase the informativeness and factual congruency in generated news summaries.

Table 2: Comparative Analysis of Generated Summaries Across Experimental Setups

Article Topic	Gold Summary	E1	E2	E3	E4
TikTok Sale Delay	President Donald Trump has granted a 90-day extension, until September 17, for the sale of TikTok in the United States by its Chinese parent company…….	President Donald Trump has granted a 90-day extension for the sale of TikTok in the United States.	TikTok’s sale is delayed again. There were no details on legislation or international concerns are given.	President Trump extends TikTok sale deadline to 17 September. Mentions **Vance Office, approval needed from Beijing, and 170 million users.**	President Trump grants 90-day reprieve on TikTok sale due to **security concerns.** Mentions **third delay, Beijing approval, and data protection assurance for users.**
COVID Variant Alert	WHO flags new variant, urges caution; the strain is designated as a variant of interest and may pose public health risks	A new COVID-19 variant has been detected.	A new COVID-19 strain emerges. It Misses WHO involvement or classification	WHO identifies new variant as **variant of interest** and advises **global caution.**	WHO flags variant of concern, urging **public vigilance,** and **coordination with health agencies.**

6.7 Discussion

The experiments performed define a definite path of increasing the results of cross-lingual summarization through gradual optimization of its retrieval and generation tasks. The first experiments (Vanilla) served as a baseline and standard sequence-to-sequence summarization models. Nevertheless, moderate relevance and factual inconsistencies were caused by the absence of re-ranking and deeper retrieval. Cross-lingual retrieval mechanisms and re-ranking, implemented in the next experiments (Experiments 2 and 4) solved these problems by checking the semantic similarity of the documents and their context sensitivity, manifested by better ROUGE scores. Experiment 3 uses Opus-MT to translate showed the importance of proper and contextual translation in filling language gaps. Although noise was injected into the translation pipelines, integrating such translations by combining them with re-ranking procedures promoted summary coherence and informativeness.

The results of the last re-ranked pipeline (Experiment 4) had the advantage of both increased retrieval precision and translation quality which gave the best evaluation scores and showed the synergetic effect between retrieval and generative parts in cross-lingual summarization. Processing latency is still a shortcoming with computation intensive re-ranking and big scale retrieval which affects the feasibility of real-time deployment. In addition to that, metrics used to evaluate, like ROUGE and BLEU, could be informative but fail to capture semantic similarity across languages, and more progress in multilingual evaluation frameworks is needed. These gaps could be filled by including human-in-the-loop feedback and domain-specific adjustments. On the whole, the study verifies the importance of the combination of the advanced retrieval, re-ranking, and translation technologies to the creation of high-quality English article summaries.

7 Conclusion and Future Work

7.1 Conclusion

The paper demonstrated that one can make numerous significant enhancements to cross-lingual summarization by combining both RAG and state-of-the-art translation pipelines and effective re-ranking. High-quality retrieval and fine-tuned translation (Opus-MT) and model re-ranking (Cross-Encoder) were experimentally found to require a significant increase in the relevance and factual grammaticality of the summary and linguistic clarity, which is reflected in the ROUGE, BLEU scores and human evaluation. Experiment 4 performs the best, and it utilizes the re-ranked retrieval with Hugging Face in pipelines, which demonstrates the success of the re-ranking strategies in multilingual systems of summarization. On the whole, the project confirms the value of cross-lingual RAG as a fruitful direction to implement more informative and detailed news summarization systems. We can address the information gaps by basing summaries on a wider and multilingual knowledge base and therefore present users with a more comprehensive picture of world events.

7.2 Future Work

This research can be generalized in multiple dimensions in the future. Cross-lingual relevance can first be enhanced by tuning or novel algorithms of the re-ranking strategies. Second, it is possible to enrich summaries with multi-lingual and multi-source contexts (e.g. Hindi, German or social media). The other significant direction is real-time summarization and speed-optimizing of the RAG pipeline. A user-centric assessment may need to be conducted because of the better-quality insights that may require human studies. It is also necessary to speak about translation errors and correct them by using confidence scores or correction mechanisms. Also, the research in domain adaptation techniques to low-resource languages and contextualized knowledge graphs could also be useful in increasing the quality of summaries and their utility. The future work that may be done there should include the discussion of the ethical aspects of cross-lingual RAG, which lies in the translation bias, the threat of misinformation, and ethical application of AI in news summarization.

References

Achkar, P., Gollub, T. and Potthast, M., 2025. 'Ask, Retrieve, Summarize: A Modular Pipeline for Scientific Literature Summarization'. *arXiv preprint arXiv:2505.16349*. Available at: https://arxiv.org/abs/2505.16349 (Accessed: 1 August 2025).

Akavarapu, V.S.D.S., Terdalkar, H., Bhattacharyya, P., Agarwal, S., Deulgaonkar, V., Manna, P., Dangarikar, C. and Bhattacharya, A., 2025. 'A Case Study of Cross-Lingual Zero-Shot Generalization for Classical Languages in LLMs'. *arXiv preprint arXiv:2505.13173*. Available at: https://arxiv.org/abs/2505.13173 (Accessed: 1 August 2025).

Bhatnagar, N., Urlana, A., Mujadia, V., Mishra, P. and Sharma, D.M., 2023. 'Automatic data retrieval for cross lingual summarization'. *arXiv preprint arXiv:2312.14542*. Available at: https://arxiv.org/abs/2312.14542 (Accessed: 1 August 2025).

CSA, 2025. *Survey of 8,709 Consumers in 29 Countries Finds that 76% Prefer Purchasing Products with Information in their Own Language*. Available at: https://csa-research.com/Blogs-Events/CSA-in-the-Media/Press-Releases/Consumers-Prefer-their-Own-Language (Accessed: 1 August 2025).

Desai, N.P. and Dabhi, V.K., 2021. 'Taxonomic survey of Hindi Language NLP systems'. *arXiv preprint arXiv:2102.00214*. Available at: https://arxiv.org/abs/2102.00214 (Accessed: 2 August 2025).

Do, A. and Tran, S., 2024. 'Improving Context Awareness of Transformer Networks using Retrieval-Augmented Generation'. Available at: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1876745 (Accessed: 2 August 2025).

Feng, H., Fan, Y., Liu, X., Lin, T.E., Yao, Z., Wu, Y., Huang, F., Li, Y. and Ma, Q., 2024, November. 'Improving factual consistency of news summarization by contrastive preference optimization'. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 11084-11100). Available at: https://aclanthology.org/2024.findings-emnlp.648/ (Accessed: 2 August 2025).

Giannakos, M. and Cukurova, M., 2023. 'The role of learning theory in multimodal learning analytics'. *British Journal of Educational Technology*, *54*(5), pp.1246-1267. Available at: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13320 (Accessed: 2 August 2025).

Gupta, A., Zhuang, Y., Yu, Z., Zhang, Z. and Beniwal, A., 2025. 'How and Where to Translate? The Impact of Translation Strategies in Cross-lingual LLM Prompting'. *arXiv preprint arXiv:2507.22923*. Available at: https://arxiv.org/abs/2507.22923 (Accessed: 2 August 2025).

Gupta, S., Ranjan, R. and Singh, S.N., 2024. 'A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions'. *arXiv preprint arXiv:2410.12837*. Available at: https://arxiv.org/abs/2410.12837 (Accessed: 3 August 2025).

Hengle, A., Bajpai, P., Dan, S. and Chakraborty, T., 2025. 'Can LLMs reason over extended multilingual contexts? Towards long-context evaluation beyond retrieval and haystacks'. *arXiv preprint arXiv:2504.12845*. Available at: https://arxiv.org/abs/2504.12845 (Accessed: 3 August 2025).

Huang, K.H., Ahmad, W.U., Peng, N. and Chang, K.W., 2021. 'Improving zero-shot cross-lingual transfer learning via robust training'. *arXiv preprint arXiv:2104.08645*. Available at: https://arxiv.org/abs/2104.08645 (Accessed: 3 August 2025).

Järvinen, E., 2024. '*Long-input summarization using large language models'*. Available at: https://aaltodoc.aalto.fi/items/758168f2-71f5-4954-a81d-f546b96787d7 (Accessed: 4 August 2025).

Kapočiūtė-Dzikienė, J. and Ungulaitis, A., 2024. 'Towards Media Monitoring: Detecting Known and Emerging Topics through Multilingual and Crosslingual Text Classification'. *Applied Sciences*, *14*(10), p.4320. Available at: https://www.mdpi.com/2076-3417/14/10/4320 (Accessed: 4 August 2025).
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L., 2019. 'BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension'. *arXiv preprint arXiv:1910.13461*. https://arxiv.org/abs/1910.13461 (Accessed: 4 August 2025).
Li, B., Haider, S., Luo, F., Agashe, A. and Callison-Burch, C., 2024, November. 'BordIRlines: A Dataset for Evaluating Cross-lingual Retrieval Augmented Generation'. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia* (pp. 1-13). Available at: https://aclanthology.org/2024.wikinlp-1.3/ (Accessed: 4 August 2025).
Liu, W., Trenous, S., Ribeiro, L.F., Byrne, B. and Hieber, F., 2025. 'XRAG: Cross-lingual Retrieval-Augmented Generation'. *arXiv preprint arXiv:2505.10089*. Available at: https://arxiv.org/abs/2505.10089 (Accessed: 5 August 2025).
Mohamed, A., 2025. 'Transfer Contextual Learning for Cross-Lingual Text Summarization'. *SSRN*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5252166 (Accessed: 5 August 2025).
Monir, S.S., Lau, I., Yang, S. and Zhao, D., 2024. 'VectorSearch: Enhancing document retrieval with semantic embeddings and optimized search'. *arXiv preprint arXiv:2409.17383*. Available at: https://arxiv.org/abs/2409.17383 (Accessed: 5 August 2025).
Ranganathan, A., Tamminaina, S.G. and Raina, G., 2023, December. 'A Study of Dialog Summarization Across Datasets and Domains'. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval* (pp. 196-202). Available at: https://dl.acm.org/doi/abs/10.1145/3639233.3639245 (Accessed: 5 August 2025).
Schofield, J., Tian, S., Truong, H.T.T. and Heil, M., 2025. DS@ GT at CheckThat! 2025: 'Exploring Retrieval and Reranking Pipelines for Scientific Claim Source Retrieval on Social Media Discourse'. *arXiv preprint arXiv:2507.06563*. Available at: https://arxiv.org/abs/2507.06563 (Accessed: 6 August 2025).
Shohan, F.T., Nayeem, M.T., Islam, S., Akash, A.U. and Joty, S., 2024. 'XL-HeadTags: Leveraging multimodal retrieval augmentation for the multilingual generation of news headlines and tags'. *arXiv preprint arXiv:2406.03776*. Available at: https://arxiv.org/abs/2406.03776 (Accessed: 6 August 2025).
Singh, A.K., Murthy, R., Sen, J., Mittal, A. and Ramakrishnan, G., 2024. 'Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages'. *arXiv preprint arXiv:2407.13522*. Available at: https://arxiv.org/abs/2407.13522 (Accessed: 6 August 2025).
Statista, 2025. *The most spoken languages worldwide in 2025*. Available at: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/ (Accessed: 6 August 2025).
Tiedemann, J., Aulamo, M., Hardwick, S. and Nieminen, T., 2022. 'Open Translation Models, Tools and Services'. In *European Language Grid: A Language Technology Platform for Multilingual Europe* (pp. 325-330). Cham: Springer International Publishing. Available at: https://library.oapen.org/bitstream/handle/20.500.12657/59316/1/978-3-031-17258-8.pdf#page=345 (Accessed: 7 August 2025).
Tiedemann, J. and De Gibert, O., 2023. 'The OPUS-MT dashboard-A toolkit for a systematic evaluation of open machine translation models'. In *Annual Meeting of the Association for Computational Linguistics: ACL-DEMO 2023* (pp. 315-327). Association for Computational

Linguistics (ACL). Available at: https://researchportal.helsinki.fi/en/publications/the-opus-mt-dashboard-a-toolkit-for-a-systematic-evaluation-of-op (Accessed: 7 August 2025).
Tiyajamorn, N., Kajiwara, T., Arase, Y. and Onizuka, M., 2021, November. 'Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation'. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7764-7774). Available at: https://aclanthology.org/2021.emnlp-main.612/ (Accessed: 7 August 2025).
Wang, J., Meng, F., Zheng, D., Liang, Y., Li, Z., Qu, J. and Zhou, J., 2022. 'A survey on cross-lingual summarization'. *Transactions of the Association for Computational Linguistics*, *10*, pp.1304-1323. Available at: https://direct.mit.edu/tacl/article-abstract/doi/10.1162/tacl_a_00520/114046 (Accessed: 7 August 2025).
Xu, S., Pang, L., Shen, H. and Cheng, X., 2024. 'A theory for token-level harmonization in retrieval-augmented generation'. *arXiv preprint arXiv:2406.00944*. Available at: https://arxiv.org/abs/2406.00944 (Accessed: 8 August 2025).
Xu, B., Chen, Y., Wen, Z., Liu, W. and He, B., 2025. 'Evaluating small language models for news summarization: Implications and factors influencing performance'. *arXiv preprint arXiv:2502.00641*. Available at: https://arxiv.org/abs/2502.00641 (Accessed: 8 August 2025).
Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K. and Hashimoto, T.B., 2024. 'Benchmarking large language models for news summarization'. *Transactions of the Association for Computational Linguistics*, *12*, pp.39-57. Available at: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00632/119276 (Accessed: 8 August 2025).