Optimizing Multimodal RAG Systems for Multilingual Product Support

Mudit Garg
Department of Computer Science and Engineering, Symbiosis Institute of Technology, India
gargmudit2708@gmail.com

Nandita Namboodiri
Department of Computer Science and Engineering, Symbiosis Institute of Technology, India
nandita.namboodiri@gmail.com

Krisha Joshi
Department of Computer Science and Engineering, Symbiosis Institute of Technology, India
krishajoshi.work@gmail.com

Karthik Krishna
Department of Computer Science and Engineering
Symbiosis Institute of Technology, India
karthikkrishnagrs@gmail.com

Deepali Vora
Department of Computer Science and Engineering
Symbiosis Institute of Technology, India
deepali.vora@sitpune.edu.in

Abstract: *Multilingual product assistance necessitates machine learning systems that incorporate text, speech, and visual data and must adapt to varied linguistic surroundings. Retrieval-Augmented Generation (RAG) frameworks are a potential solution. However, they are highly contingent on integration within retrieval strategies and language models, which is understudied, particularly in Indic languages and multimodal applications. This paper presents a systematic evaluation of five RAG architectures across seventeen pipeline configurations, combining retrieval methods such as Best Matching 25 (BM25), Dense Passage Retrieval (DPR), chroma, and Facebook AI Similarity Search (FAISS) with multilingual embedding models including IndicBERT, mT5, and sentence transformers. A curated dataset of 170 engineering manuals, brochures, and presentations was used to replicate real-world troubleshooting scenarios. Among the evaluated approaches, a ColPali-inspired multimodal fusion mechanism-capable of jointly encoding text and images-substantially improved retrieval precision and diagnostic support in complex cases. Evaluation using Recall@5, Mean Reciprocal Rank (MRR), BLEU, ROUGE-L, and mean Average Precision (mAP) shows that hybrid pipelines, particularly Chroma-FAISS with mT5, achieve strong semantic alignment (Recall@5=0.78, ROUGE-L=0.46) while maintaining efficiency. The ColPali-based multimodal RAG further enhances performance, reaching 94% Top-1 retrieval accuracy and user satisfaction above 90%. These results indicate the possibility of carefully structured hybrid and multimodal RAG systems providing accurate, fluent, and inclusive support in real-time, giving design guidelines to be applied beyond education to any other domain requiring real-time interventions (healthcare, education, technical training, etc.).*

Keywords: *Retrieval-augmented generation, multimodal language models, vision-language models, indic languages, AI for technical support.*

Received May 8, 2025; accepted November 5, 2025
https://doi.org/10.34028/iajit/23/3/9

1. Introduction

Artificial Intelligence (AI) has revolutionised industries due to its potential to automate, enhance efficiency, and provide once-infeasible solutions. A rapidly advancing frontier within AI is multimodal intelligence, which is designed to jointly process text, speech, and visual modalities and produce contextually rich outputs. These systems bridge the gap between human-like and machine cognition, serving more contextually aware and actionable outputs and increasingly intermediating human-machine interaction in high-stakes decision-making settings [32].

However, linguistic and modal inclusion continue to be a crucial core concern. Most state-of-the-art systems focus on high-resource global languages, leaving gaps in linguistically diverse locations. India is a prime example of this problem, with hundreds of dialects and 22 designated languages. In technical product support, language barriers can lead to misinterpreted instructions, repeated queries, and longer resolution times, ultimately reducing user satisfaction and operational efficiency. This challenge becomes even more pronounced in multilingual environments, where communication gaps hinder effective problem-solving, particularly when technical complexity intersects with regional linguistic diversity-making clear, timely, and accurate responses in customer and product support all the more critical [12].

For field service engineers, who have to simultaneously operate complicated machinery and speak to end users in their native tongues, the challenge is especially severe. To provide adequate support, visual diagrams and audio instructions that adjust to the engineer's language and situational context are just as important as textual instructions. Current AI support systems rarely combine these needs into a single, scalable architecture.

Retrieval-Augmented Generation (RAG) offers a promising direction by coupling information retrieval with generative models to deliver context-aware responses [10]. However, RAG's performance is

dependent on how it is configured. The choice of retrieval strategy and language model can significantly shape the quality of the response, affecting retrieval accuracy, response coherence, latency, and overall user experience. RAG is gaining popularity, but its comprehensive evaluation in multilingual and multimodal situations is still lacking, especially for low-resource Indic languages [25].

This study addresses this gap by designing and benchmarking a RAG-powered multimodal AI Bot for multilingual product support in the context of Indian Languages. The bot implements various retrieval augmented generation systems, which handle text alongside speech and visual inputs to answer multilingual technical assistance needs. The training and evaluation are conducted on a curated dataset of 170 technical manuals, brochures, and presentations reflecting real-world troubleshooting scenarios. The following three research objectives guide the study:

- RO1: perform a systematic comparison of the performances of different configuration approaches with varying retrieval mechanisms and language models.
- RO2: validate the efficiency of the multimodal fusion operation in real-world scenarios, such as complex equipment troubleshooting.
- RO3: identify trade-offs between retrieval accuracy, response speed, and multilingual adaptability.

The study provides a comparative methodology for improving RAG pipelines in situations with limited resources and linguistic diversity by addressing these problems. The results provide design guidelines for broader fields like healthcare and education, as well as product support systems.

This paper is divided into eight sections. Section 2 reviews the current literature on RAG, multimodal AI, and multilingual Natural Language Processing (NLP), focusing on gaps identified in comparative configuration evaluation. Section 3 describes the system design, including the architecture of the AI bot and the five approaches taken in this paper. Section 4 describes the experimental setup, including the chosen dataset, relevant computational tools, and an explanation of the evaluation metrics. Section 5 details the performance comparison of 17 configurations of the proposed approaches and discusses critical trade-offs. Section 6 summarises the abovementioned approaches' strengths, weaknesses, and scalability. Section 7 will conclude the study by detailing key findings and proposed future work. Section 8 then lists references used in this paper.

2. Survey of the Literature

Research on RAG, multimodal language models, and multilingual AI has grown rapidly in recent years, reflecting the need for chatbots operating across diverse contexts and input formats. RAG has shown strong potential for improving document selection and delivering context-aware responses by grounding outputs in external knowledge sources. Prior research has also explored automated techniques for text document classification and organization to improve information retrieval efficiency [15]. These advances build on integrating LLMs into chatbot systems, significantly enhancing the quality of human-AI interaction.

The introduction of transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) marked a significant shift in chatbot capabilities, enabling conversations that are far more contextual and accurate than earlier rule-based or statistical systems. Tools like BERT for intent detection and T5 for Natural Language Generation (NLG) have allowed for more complex and engaging exchanges, improving user satisfaction and system efficiency [4].

A broad body of research has explored different ways of applying Large Language Models (LLMs) and RAG pipelines to customer support and knowledge management areas. These studies demonstrate the value of combining pre-trained models with retrieval mechanisms but also highlight persistent challenges, including the demanding nature of multimodal data fusion, difficulties in handling multilingual interactions, and the high computational resources required for real-time deployment [25].

The recent development in NLP and LLMs has made a massive leap toward creating multilingual chatbots. Initial pipelines have usually integrated dedicated models, including BERT-like models, to recognise intent and extract entities, with models trained on chitchat databases (e.g., restaurant conversations) paired with T5 for NLG [1]. These methods demonstrated significant progress, and the recognition of intents gained 92.55%, entity extraction 84.55%, and BLEU scores achieved about 0.60. Still, issues persisted in image tolerance of heterogeneous user input and in handling linguistic fluctuation.

The evolution of transformer-based architectures, namely GPT and BERT, transformed question-answering chatbots by performing zero-shot and prompt-based learning [2]. Experiments based on benchmarks like TruthfulQA indicated that the current models (GPT-3, GPT-4, PaLM 2, Claude 2.0, LLaMA 2) perform very well in contextual inference and open-ended dialogue. Still, they are susceptible to hallucinations and adversarial prompting. To overcome factual reliability, RAG techniques have been developed using pre-trained models and structured sources like knowledge graphs on Neo4j and grounding mechanisms [24]. Although such systems enhance information quality (with scores between 0.6 and 0.8), they create extra architectural complexity and integration issues.

Alongside proprietary systems, open-source frameworks have been used to build chatbots on

institutional or web-scraped data. For example, pipelines using beautifulsoup for data collection, instructor-large for embeddings, and Flan T5 XXL for generation have been applied to educational datasets [17]. Although effective in producing sensible responses, their evaluation often depends on subjective human judgment and lacks uniformity across platforms such as LangChain. Domain-specific implementations have also appeared-for instance, vehicle manual assistants using GPT-3, LangChain, and sentence transformers [14]. These improve accessibility of technical content but fall short when dealing with images, typeset documents, or interface usability.

Hybrid approaches have also been explored, particularly in education. Models combine text segmentation tools (e.g., CharacterTextSplitter), OpenAI embeddings, and LLM-based retrieval and have been applied to locally stored PDFs and image files [20]. Though they enhance contextual accuracy and efficiency, they are still limited by time constraints related to localised storage and the need for manual database updates. Comparative evaluations on large datasets such as the National Highway Traffic Safety Administration (NHTSA) dataset have revealed trade-offs in file size handling, embedding techniques, and retrieval pipelines [30]. Beyond text, multimodal benchmarks like ViDoRe highlight visual and linguistic information integration. Vision-language models such as ColPali, trained with contrastive objectives, outperform baselines like Best Matching 25 (BM25), Contrastive Language-Image Pretraining (JinaCLIP), and Sigmoid Loss for Language-Image Pretraining (SigLIP) when processing infographics and tables. However, they remain restricted in language coverage and often rely on synthetic datasets [9].

Despite these advances, multilingual performance continues to be a challenge. Zero-shot evaluations of ChatGPT in 37 languages across tasks such as Part-of-Speech (POS) tagging, Named Entity Recognition (NER), Question Answering (QA), and Natural Language Inference (NLI) show weaker results compared to multilingual models like mT5-XXL, XLM-R, and IndicBERT, particularly in low-resource languages and semantically complex settings [13]. While high-resource languages show strength in specific tasks (e.g., POS tagging), performance deteriorates for underrepresented contexts, reflecting inherent biases. Finally, recent work on PDF-based chatbots, developed with GPT-3.5, streamlit, and LangChain, demonstrates the ability to query documents directly without external data sources [5]. These systems are efficient for lightweight search and multilingual support but remain limited in handling image-rich documents, maintaining conversation history, and enabling advanced interaction.

Recent work points to several challenges in building robust multimodal and multilingual AI systems. Many current approaches struggle to combine different types of data-text, images, audio, and graphics-seamlessly, which limits how well they can be applied in real scenarios. Systems that rely on static knowledge bases also face delays or errors when responding to queries in real time, since updating or adapting to new information is often slow. Performance across languages remains uneven, with noticeable weaknesses in low-resource languages and technical domains, making these models less suitable for global use. On top of that, the high computational costs of advanced AI architectures make them challenging to adopt for smaller organisations with limited resources. Another recurring problem is that the responses from existing RAG models often feel fragmented and lack smooth contextual flow. To overcome these limitations, there is a need for more scalable designs that use techniques like knowledge graphs and semantic indexing, paving the way for AI systems that are more reliable, inclusive, and efficient across languages and modalities.

3. Methodology

The methodology of this study focuses on building a RAG-powered multimodal AI bot to address the challenges of multilingual and multimodal product support. It encompasses retrieval mechanisms and language models, which enable processing diverse data modalities- pure text, images, and audio inputs. In systematically evaluating 17 varied configurations in five strategies, the research embodies the performance of different retrieval methods like BM25, Dense Passage Retrieval (DPR), chroma, Facebook AI Similarity Search (FAISS), and embedding. Models such as IndicBERT [11], mT5 [31], and sentence transformers [21] are the basis of this research's methodology. The comparative analysis of different environments and their key measures, like retrieval accuracy, response latency, multilingual adaptability, and multimodal effectiveness, is used to find optimal configurations of a scalable, efficient product support system.

3.1. Approach 1: Multi-Embedding RAG Approach

This method provides a RAG model with 13 configurations of embedding and retrieval techniques that enhance multilingual question-answering abilities, particularly in Indic languages. A knowledge base was constructed from 170 raw PDF and word document files. These files were processed with PyMuPDF and python-docx to extract text. The text was cleaned, segmented into sections based on structural patterns, and then chunked into overlapping segments (~300 tokens) for generating embeddings.

Embeddings were generated with four models: IndicBERT and mT5 for multilingual and Indic language representation, the Sentence-Transformers for sentence-level semantics, and DPR for dense query-document alignment. These embeddings are stored in Chroma with FAISS indexing integrated for scalable, high-speed

similarity search. Retrieval strategies include BM25 for lexical relevance, DPR for dense semantic similarity, and Chroma with optional FAISS indexing for efficient retrieval.

The system was configured into 13 variants by combining embedding models with retrieval methods and storage mechanisms, as summarised in Table 1.

Table 1. Configurations with approach 1.

Embedding model	Retrieval technique	Chroma storage	FAISS indexing	Generator (mT5)
IndicBERT	BM25	Yes	No	Yes
IndicBERT	DPR	Yes	No	Yes
IndicBERT	Chroma (direct)	Yes	No	Yes
IndicBERT	Chroma+FAISS	Yes	Yes	Yes
mT5	BM25	Yes	No	Yes
mT5	DPR	Yes	No	Yes
mT5	Chroma (direct)	Yes	No	Yes
mT5	Chroma+FAISS	Yes	Yes	Yes
Sentence-transformer	BM25	Yes	No	Yes
Sentence-transformer	DPR	Yes	No	Yes
Sentence-transformer	Chroma (direct)	Yes	No	Yes
Sentence-transformer	Chroma+FAISS	Yes	Yes	Yes
DPR	Chroma	Yes	Yes	Yes
Embedding model	Retrieval technique	Chroma storage	FAISS indexing	Generator (mT5)

Retrieved contexts were passed to the mT5 generator, which uses a structured prompt combining the query and retrieved content to generate coherent, contextually accurate responses. Each configuration was evaluated for retrieval and generation quality using Recall@5, Mean Reciprocal Rank (MRR), BLEU, and ROUGE-L metrics. Figure 1 gives a visual representation of the architecture of the bot.

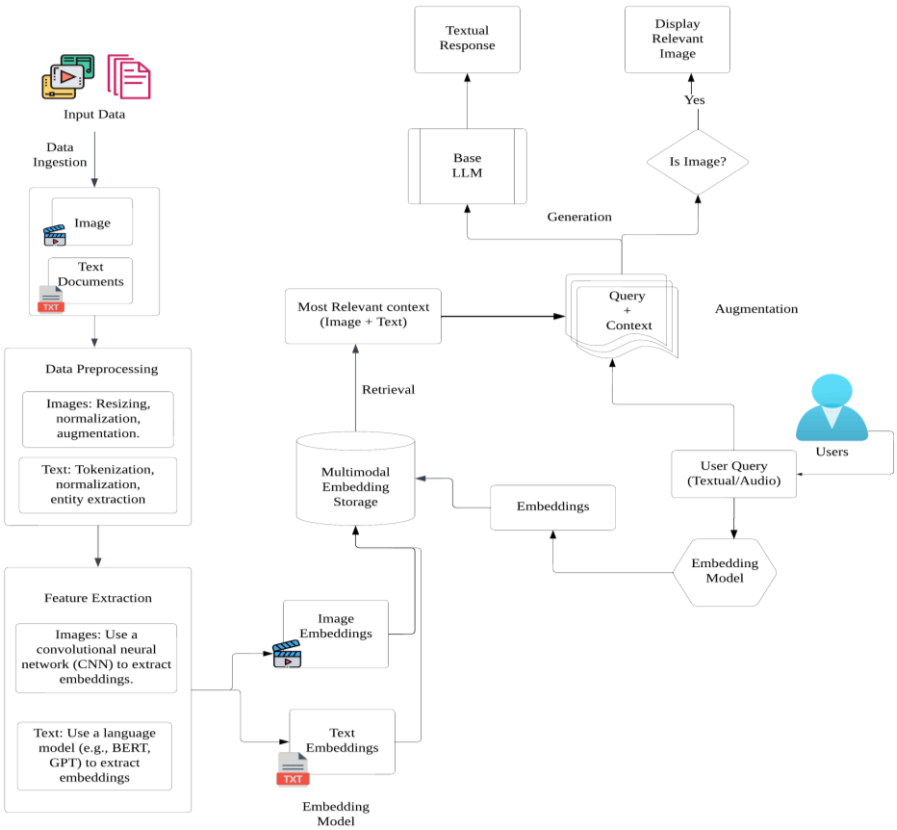

Figure 1. Block diagram for initial approaches.

3.2. Approach 2: Multimodal Embedding Approach

The multi-embedding RAG framework facilitates effective, context-aware information retrieval and multimodal response generation through preparation, embedding generation, and retrieval-augmented response generation.

The first preprocessing stage removes noise and standardises formatting from raw textual data from sources like manuals and instructions. Material is broken into smaller sections (300 words with overlap) to support granular retrieval and preserve context. Sentences in these passages are encoded by the pre-trained "sentence-transformers/all-MiniLM-L6-v2" model into dense vector embeddings that capture their semantic meaning.

It contains embeddings within a Chroma vector database optimised to support scalable, quick, and exact

similarity searches. The retrieval granularity is improved through recursive splitting into smaller chunks (200 characters with overlap) and batch processing, ensuring scalability and memory efficiency on large datasets.

The last step consists of encoding the user queries into embeddings and using the cosine similarity to compare them to the vector database to fetch the top k relevant passages. When appropriate, multimodal outputs like pictures or diagrams are appended to the gathered excerpts to form a consistent answer. The modular structure of the framework allows it to be easily extended and maintained. Other retrieval operations or data sources can be added with minimal effort. It is a low-latency, real-time response system that integrates advanced NLP mechanisms with effective retrieval mechanisms, thus enabling its use across various applications.

3.3. Approach 3: DPR-Enhanced RAG-based Multimodal Approach

The suggested methodology seeks to improve retrieval-augmented question-answering of technical support in multiple languages. This is achieved by applying a DPR-enhanced RAG framework together with multilingual mT5. The methodology starts with preprocessing technical documentation and manual textual information. The BERT-based tokeniser is utilised for cleaning, tokenising, and segmenting the text into bite-sized chunks of up to 512 tokens that are compatible with downstream models. The DPR-Context Encoders embed chunks as high-dimensional vector representations optimised for semantic similarity. FAISS indexes embeddings to retrieve content when responding to user requests effectively. The suggested methodology thus addresses retrieval-augmented question-answering of technical support in multiple languages, utilising a DPR-enhanced RAG framework and multilingual mT5 Integration. The embedding is done utilising high-dimensional vector representations optimised for semantic similarity, enabled by the DPR-Context Encoder. The final step utilises FAISS to index the embeddings, allowing effective and efficient retrieval of content upon receiving user requests. The DPR-question encoder then encodes these user queries, indexing them against the FAISS index to retrieve the most relevant chunks.

The extracted segments constitute the contextual foundation for generating responses. The system utilises the multilingual mT5 model to create reactions with correctness and fluency across several languages, including English and Hindi. The extracted passages and questions are presented as input prompts to the mT5 model to generate answers from the source material. The system can easily scale and perform well in real-time by balancing retrieval accuracy and low latency. Besides human judgment of contextual relevance and user satisfaction, it is assessed using automated measures such as Recall@K and BLEU scores.

Although it functions well in retrieval, fluency of responses is still a key problem. This method, through the combination of retrieval with multilingual generation in a new way, enables adequate technical support to different groups of users.

3.4. Approach 4: Chroma-Enhanced RAG using Ollama

This process relies on a sophisticated Retrieval-Augmented Generation model coupled with Chroma vector databases to enable effective querying and retrieval of large text corpora. The process begins with data preparation stages that involve splitting pre-specified textual data sources into manageable, smaller pieces, using the RecursiveCharacterTextSplitter to preserve semantically consistent overlapping sections. These text pieces are then embedded into high-dimensional spaces by ollama embeddings using the "nomic-embed-text" model, which effectively captures delicate, semantically related subtleties.

The vectorised data is carefully organised and saved in a Chroma database, optimised for long-term storage and fast retrieval. Each chunk is assigned a unique identifier based on its source and sequence, enabling traceability without redundancy. The database is optimised for incremental updates, enabling scalability and optimal resource utilisation. Upon query, the inputs are transformed into embeddings, and then a similarity search over the database is performed to ascertain the relevance of the chunks. The retrieved chunks are then organised into a context within a pre-defined prompt, which directs the "Mistral" language model to generate accurate and contextually relevant responses.

This assessment includes Precision@K, MRR, and Relevance Scores, to name a few, all towards iterative improvement. This approach enables an expressive and scalable environment for document-based question answering by applying sophisticated embeddings, semantic search, and effective prompt construction. It finds special utility in high-precision and interpretive applications.

3.5. Approach 5: ColPali-based Multimodal RAG

The suggested ColPali-based multimodal RAG solution harmoniously combines the best of RAG, multimodal embeddings, and multilinguality to create a cost-effective AI-powered system bespoke for product information and troubleshooting. The system uses the ColPali model, an extremely potent multimodal architecture that couples PaliGemma for multimodal embedding generation with ColBERT for dense text-based retrieval. As a result, the hybrid architecture enables the system to elegantly handle and embed textual and visual information to ensure thorough information retrieval and the generation of well-structured responses.

The method starts with user input, either speech or text. While Google Speech-to-Text Application Programming Interface (API) translates voice queries into text, text queries undergo preprocessing steps such as normalisation and tokenisation. To provide support for multiple languages, including many Indian languages, the system also identifies the user's preferred language. The pre-processing process includes converting pages of the document dataset, which are essentially in PDF format, to images and later inserting these with the help of the ColPali model. This embedding procedure exploits the model to develop high-dimensional representations that capture semantic commonalities within textual and visual content.

User queries are further transformed into embeddings for retrieval by the same paradigm, ensuring compatibility with the indexed document embeddings. The answer-generating module receives Top-K matches via a dense retrieval method that employs cosine similarity to find the most pertinent documents. A large language model, in this case, the Gemini LLM, uses contextual information from the retrieved documents to provide context-aware and profound answers. With Google Translate, the responses generated are translated to the user's choice of language for the multilingual landscape in India to ensure accessibility in various Indian languages.

The interface is streamlit-based, which makes user engagement easier by providing the same chat-like setting where users can ask questions and get answers along with any pertinent document images. The interface is intuitive, easy to use, and accepts spoken and typed input. Optimisations have been made in managing memory-intensive operations, such as embedding generation and retrieval, in light of competing demands from the system. An extra 80 GB of VRAM was set up through GPU memory because RAM utilisation often exceeded capacity to ensure that the system operated without interruptions while handling large datasets and complex queries. The system's performance improved when the NVIDIA GeForce RTX 4080 GPU was used, which paved the way for a speedier generation pathway for quicker retrieval.

In addition, retrieval accuracy, answer relevance and multilingual translation quality were used to assess the ColPali-based multimodal RAG framework. Overall results show more than 90% accuracy in top-1 with robust support for multiple languages, hence exhibiting great retrieval precision that is scalable and dependable for activities based on real-world product knowledge and troubleshooting. Figure 2 visually represents the bot's architecture created using this approach.

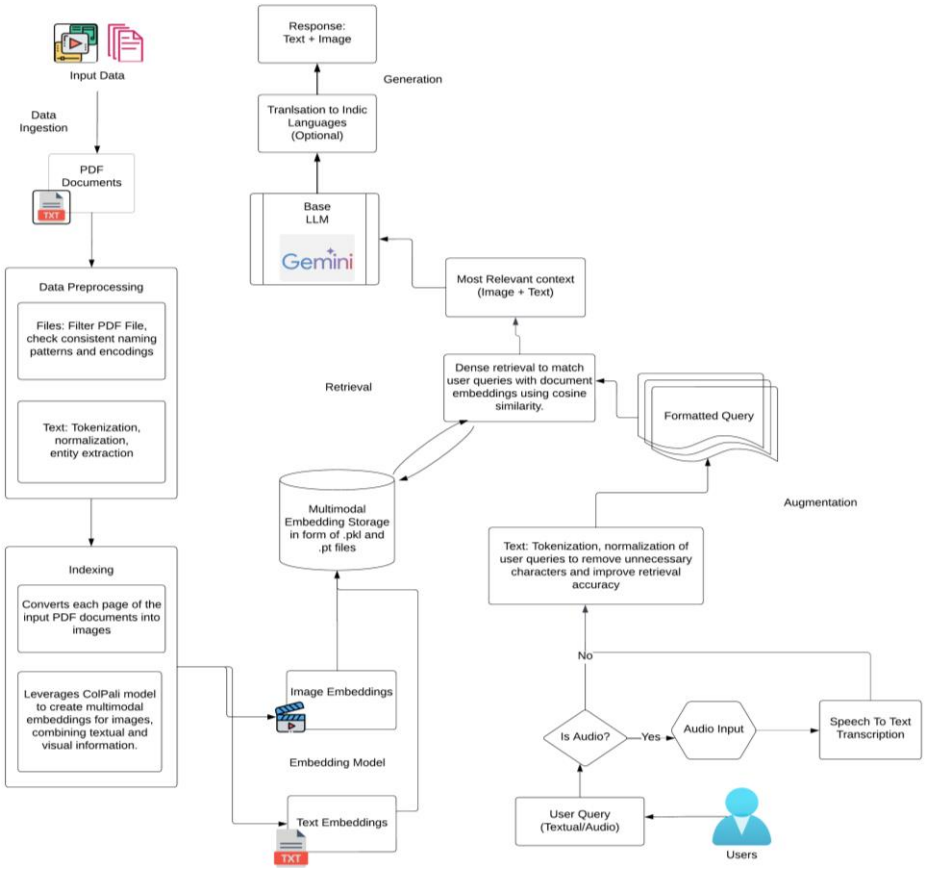

Figure 2. Block diagram for ColPali-based approach.

After outlining the five approaches, it becomes clear that no single configuration is universally superior; each offers advantages depending on the use case. The comparison highlights how different combinations of retrieval mechanisms and embedding models influence retrieval accuracy, latency, and the adaptability needed

to support multiple languages. Some pipelines favour speed, while others prioritise semantic depth and contextual relevance. To make these methodological variations easier to interpret, a schematic summary is provided in Figure 3.

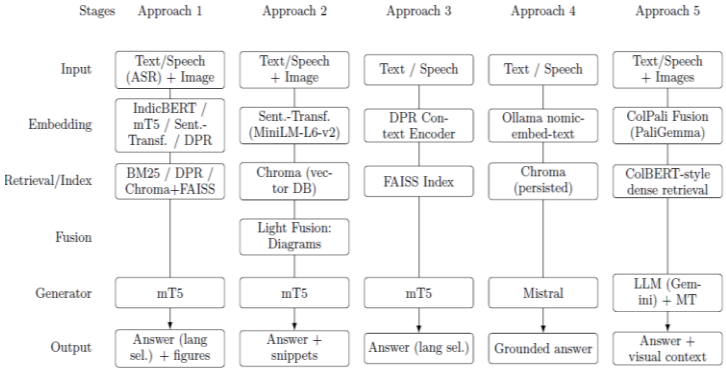

Figure 3. Schematic summary of the five evaluated RAG approaches.

This figure gives a unified overview of the pipeline structure across all approaches, highlighting the specific embedding models, retrieval methods, fusion mechanisms, and generators that distinguish them. Each column shows the flow from input to output, following the common stages with approach-specific variations. It emphasises the need to tailor RAG setups to practical requirements-whether prioritising real-time responsiveness, multimodal support, or linguistic inclusivity-and sets the stage for the following experimental results.

4. Experimental Setup

The experimental setup for this experiment is the systematic analysis of the five approaches with a judiciously chosen dataset, appropriate computational tools, and designed assessment metrics. The range of methodologies employed by each strategy requires specific metrics to fairly compare their performance and effectiveness in addressing issues of multilingual and multimodal product support.

4.1. Dataset Description

There are 170 documents of product manuals, brochures, and presentations of real scenarios encountered while debugging and supporting products in the dataset. These documents derive meaningful information from various modalities like text, graphics, and structured material to align with the multimodal capabilities of the RAG-powered bot. This dataset has enough variation to test each configuration's adaptation across different input formats and languages.

Manuals often contained detailed tables and technical diagrams. On the other hand, brochures showed much more variation-ranging from simple single-column layouts to double- or even triple-column formats. Many files also mixed styles across pages, with some sections in single-column text, others in two-column layouts, and entire pages presented only as images. Presentations added further variety by combining diagrams, flowcharts, and bullet-style text.

This structural diversity, together with the image-intensive nature of the documents, created several preprocessing challenges. Extracting text from multi-column layouts often led to misordering content, while scanned or image-only pages required Optical Character Recognition (OCR) to recover text, which in some cases might not be relevant if not considered along with the image. In addition, the mixture of text, tables, and visuals within single documents made it necessary to normalise the content into a consistent format. Such diversity closely reflects the types of materials engineers and support staff work with in real industrial settings. This makes the dataset realistic and a strong benchmark for testing how different RAG configurations handle retrieval accuracy, multilingual adaptability, and robustness across varied document types.

4.2. Computational Tools

The studies were done using a Windows 11 system with a Core i9-13900HX (13th Gen) processor which has 24 cores/32 threads running at a base clock of 2.2 GHz and supports turbo boost, NVIDIA corporation GeForce RTX 4080 laptop GPU with 28 GB VRAM (boosted to an additional 80 GB in system VRAM to mitigate MemoryErrors during large batch embedding processing), and 32 GB DDR5 RAM (5 600 MT/s).

The development system was created using Python 3.12.3 in specific venv virtual environments with version control through GitHub integrated with Visual Studio Code. Primary frameworks for core deep learning were TensorFlow and PyTorch, and pretrained encoders IndicBERT and mT5, alongside DPR components, were provided by Hugging Face's Transformers library. FAISS and Chroma underline the retrieval backends working under LangChain abstraction layers for vector storage and similarity search-custom pipelines designed for robust RAG workflows, integrated, optimised retrieval pre-strategising at embedding and model interfaces [16, 18]. Along with Streamlit, which served as an interactive frontend prototype, audio transcription was translated using Googletrans, Pydub [27] was used for translation, and Pillow [7] was employed to

preprocess and augment images.

4.3. Evaluation Metrics

Each approach employs evaluation metrics tailored to its unique design and objectives:

- Approach 1: multi-embedding RAG its performance evaluation is based on the accuracy and semantic relevance of the retrieved documents and their dependence on different combinations of embeddings and retrievals. Precision@k and Recall@k measure how well the relevant documents are retrieved within the top-k.
- Approach 2: multimodal embedding RAG this approach is multimodal; hence, response coherence and the effectiveness of integration come into play. It verifies how the system integrates and utilises diverse data from different sources, including text and images. Simulations of user feedback are also used for real-world adaptability testing.
- Approach 3: DPR-enhanced RAG this method focuses on dense passage retrieval and is judged by the semantic depth and quality of the responses. Bilingual Evaluation Understudy (BLEU) and Recall Oriented Understudy for Gisting Evaluation-Longest common subsequence (ROUGE-L) [23] scores determine how contextual and fluent the generated responses are.
- Approach 4: chroma-enhanced RAG with ollama this approach emphasises the efficiency of retrieval and scalability, and it has given more emphasis to metrics such as query response times, indexing latency, and retrieval precision. These measurements show the computational efficiency and applicability of the technique for real-time applications.
- Approach 5: ColPali-based multimodal RAG because of its focus on visual data integration, this approach is evaluated against image-text relevance scores and the accuracy of visual context. Metrics used here include mean Average Precision (mAP) and cosine similarity to assess how effectively visual and textual data combine to generate contextually relevant outputs.

In addition to automatically measured metrics, including Recall@k, Precision@k, BLEU, ROUGE-L, and mean Average Precision (mAP), a user study was set up to evaluate the system performance through the eyes of the end user. Fifteen respondents, who comprised both technical and non-technical people, engaged with the bot in various situations, after which they rated responses on a 5-point Likert-type scale (1=very unsatisfactory, 5=very satisfactory) [29]. Three subjective measures were derived from this process: Retrieval Satisfaction, which captured the perceived relevance and completeness of retrieved passages; generation satisfaction, which assessed the fluency, coherence, and factual grounding of generated answers; and content relevance rating, where it was judged whether responses directly addressed the user query. These scores were averaged into percentages. Where stated, overall satisfaction denotes the mean of retrieval and generation satisfaction. These subjective assessments complement the quantitative metrics, offering a more comprehensive understanding of how well each configuration aligns with real-world user expectations.

This structure ensures that the evaluation properly deals with each approach's strengths and limitations while remaining consistent with the study's aims of improving retrieval accuracy, latency, and multilingual multimodal adaptability. The personalised measures form a solid base on which to analyse the results reported in the following sections critically.

5. Implementation and Results

This section provides the implementation details of the proposed approaches and screenshots to depict the functionalities and workflows of the system. The results from the performance evaluation that determine and highlight the retrieval accuracy, latency, and relevance of responses are also provided. By comparing these results across the different approaches, the system's strengths, limitations, and trade-offs in multilingual and multimodal product support are underscored, thus providing a comprehensive analysis of the system's effectiveness.

5.1. Implementation Details

The core features implemented in the system include:

- Audio and text input support: it is possible to enter a text prompt directly or record an audio with a prompt for the user. Speech-to-text conversion takes place through google text-to-speech, transcribing speech into text in auto-detect and English translation.
- Contextual retrieval with history: the app preserves the chat history; each may contain one question and its answer. Some features include the fact that the app draws from the prior conversation history to foster quality answers.
- Multilingual support: the user can select the preferred language for responses. The download is available in languages like Hindi, Bengali, Tamil, Telugu, Urdu, and many more, so you can translate the content into multiple Indic languages it supports. Answers are copied and pasted directly and kept as copied and pasted, including formatting such as bolding and lists via Google Translate.
- Embeddings and document management: the app always finds one document that closely matches the term used in the query and offers a response based on the most relevant data. Documents and images are preprocessed to obtain document embeddings and reuse stored document embeddings and pictures instead of re-indexing.
- Image retrieval: thus, when a given query has been

processed, in addition to the WEB document best matched to that query, an optional set of images related to that document is also returned.

- Session-based chat history: it is designed to maintain each chat session's history so that users may continue new chats but not lose old ones. Users may see a list of the previous conversations and select one of the chats to be viewed; the conversation history will be shown on the main interface.
- Formatting for translations: the app retains the format of the texts when translating, which means that features such as numeration of bulleted points and the application of bold type within the text of individual questions are preserved for translated responses, making comprehension easy.
- Interface customisation: the area near the 'input/search' bar is where user and bot messages appear; they have a customised design; there is also a sidebar for managing the chat and simple, readable sections for input queries, answers, and images drawn from the search.

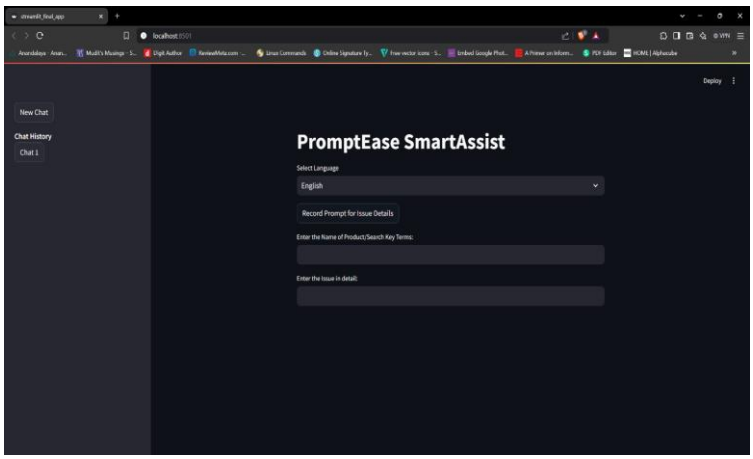

Figure 4. Initial screen of the implemented approach.

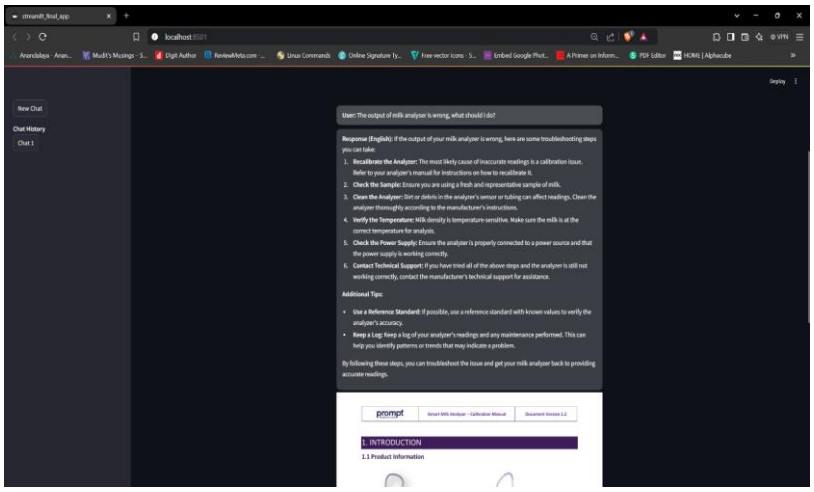

Figure 5. Generated response with extracted document segment as citation.

This approach provides an inclusive RAG-based system using the ColPali chatbot system, which is modular and multilingual. It enables a user-friendly way of interacting with the system for complicated document searching through a conversation interface. The following screenshots (Figures 4 and 5) showcase the user interface, the screens displayed at various steps.

5.2. Results

5.2.1. Evaluation of the Multi-Embedding RAG Approach

The quality of both retrieval and generation in the RAG system has been emphasised to ensure that the retrieved content is substantial enough to complement the generated responses. Retrieval performance is measured with metrics such as Recall@k, which measures the ratio of relevant documents retrieved in the top k results; mean reciprocal rank, which emphasises the rank of the first relevant document; and Average Precision (AP), which evaluates systematic retrieval throughout all levels of recall. These metrics measure semantically relevant information rather than random word matching.

Generation quality was evaluated using BLEU and ROUGE-L (lexical and recall-oriented overlap with references), plus content relevance rating and generation satisfaction (Likert-derived) to assess coherence, completeness, and factual grounding. On average,

generation satisfaction ranged between 70% and 80%, indicating that most responses were coherent and contextually appropriate, while relevance ratings aligned closely with automatic overlap metrics. Table 2 presents the comparative results of the 13 configurations tested under the Multi-Embedding RAG approach. The key observation is that hybrid combinations involving Chroma+FAISS consistently outperformed purely lexical or dense retrieval methods, achieving Recall@5 scores above 0.75. IndicBERT embeddings demonstrated competitive retrieval precision but less fluency in generated responses than mT5. Generation satisfaction scores ranged between 72% and 84%, with the highest ratings for mT5-based generation, reflecting its ability to produce coherent, context-aware answers. These findings confirm that combining embeddings with hybrid retrieval yields a balanced trade-off between retrieval accuracy and user-perceived response quality.

Table 2. Evaluation table for multi-embedding RAG approaches.

Approach	Retrieval satisfaction	Generation satisfaction	R@5	MRR	AP	BLEU	ROUGE-L	Content relevance rating
IndicBERT+BM25	75%	42%	0.72	0.38	0.55	0.30	0.42	2.1
IndicBERT+DPR	78%	44%	0.74	0.41	0.57	0.31	0.43	2.3
IndicBERT+Chroma	80%	45%	0.76	0.42	0.60	0.33	0.45	2.4
IndicBERT+Chroma+FAISS	79%	44%	0.75	0.41	0.58	0.32	0.44	2.3
mT5+BM25	71%	40%	0.70	0.35	0.52	0.28	0.40	2.0
mT5+DPR	73%	41%	0.72	0.36	0.54	0.29	0.41	2.1
mT5+Chroma	77%	43%	0.75	0.40	0.56	0.30	0.43	2.2
mT5+Chroma+FAISS	76%	42%	0.74	0.39	0.55	0.29	0.42	2.2
Sentence-Transformer+BM25	72%	41%	0.71	0.36	0.53	0.29	0.41	2.1
Sentence-Transformer+DPR	75%	43%	0.73	0.38	0.55	0.30	0.42	2.2
Sentence-Transformer+Chroma	78%	45%	0.77	0.42	0.59	0.32	0.44	2.4
Sentence-Transformer+Chroma+FAISS	76%	44%	0.76	0.41	0.57	0.31	0.43	2.3
DPR (Chroma Direct/FAISS)	80%	45%	0.78	0.44	0.61	0.34	0.46	2.5

The following graphs visualise the performance of various model configurations across the evaluation parameters mentioned above. Figure 6 shows that retrieval satisfaction remains high (70-80%) across all configurations, while generation satisfaction is lower (40-45%). This gap highlights that strong retrieval does not always lead to equally fluent or coherent generation. Configurations using Chroma (e.g., IndicBERT+chroma, sentence-transformer+chroma) achieved slightly higher generation satisfaction, showing the benefit of hybrid embedding methods.

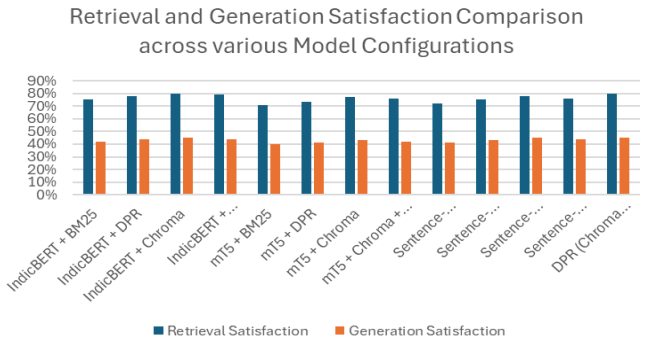

Figure 6. Retrieval and generation satisfaction comparison across various model configurations

Figure 7 compares Recall@5, MRR, AP, BLEU, and ROUGE-L across configurations. Hybrid setups such as Sentence-Transformer+Chroma and IndicBERT+Chroma consistently outperformed others, with DPR (Chroma Direct/FAISS) reaching the best retrieval scores (Recall@5=0.78, AP=0.61). BLEU and ROUGE-L were lower overall, reflecting the limits of lexical metrics for multilingual technical content and the importance of human evaluations.

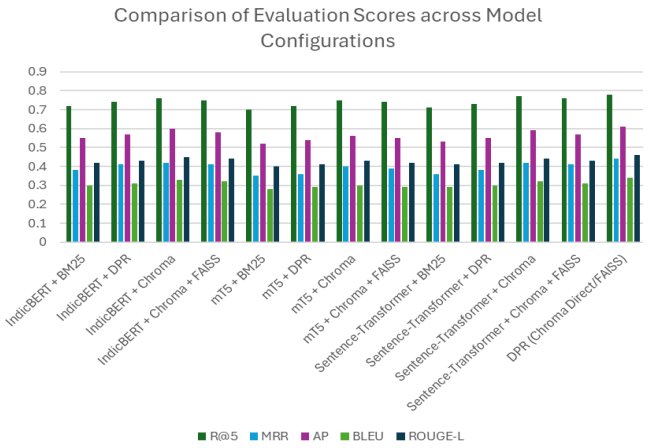

Figure 7. Comparison of evaluation scores across model configurations.

Figure 8 shows content relevance scores between 2.0 and 2.5, with DPR (Chroma Direct/FAISS) performing best. Chroma-based pipelines generally outperformed BM25-only setups, confirming the advantage of dense embeddings in aligning outputs with user queries. These ratings also align with higher retrieval satisfaction, reinforcing consistency between subjective and automated measures.

Table 2 and Figures 6,7, and 8 show that chroma-based and hybrid pipelines offered the best retrieval accuracy and content relevance balance, with DPR (Chroma Direct/FAISS) leading in retrieval scores. Yet generation satisfaction remained lower, highlighting that strong retrieval alone does not ensure coherent, fluent responses. This underscores the need to integrate retrieval and generation better rather than treating them separately.

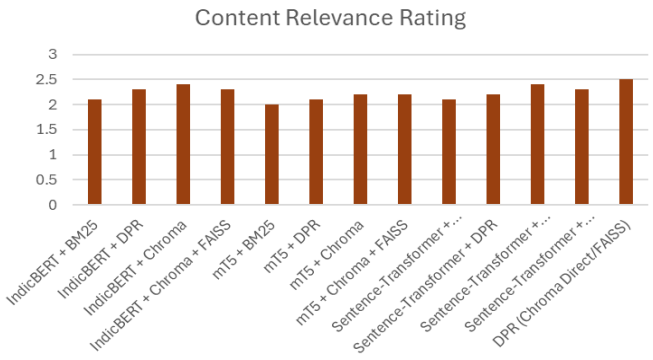

Figure 8. Content relevance rating.

5.2.2. Evaluation of Multimodal Embedding Approach

The core aspects to be covered in evaluating the multimodal LLM Bot include information retrieval, user satisfaction, and system performance. Metrics such as precision, recall, and F1-score evaluate the bot's ability to filter relevant content and ensure no critical information is missed in the coverage of user queries. High precision means the bot retrieves only relevant content, while high recall ensures no critical information is missed. This is further balanced by the F1-score, which gives an all-around evaluation of retrieval accuracy.

It measures user satisfaction in aspects such as the satisfaction score, an aggregation of user feedback ratings, and top-K accuracy, indicating the number of top results that yield the most related information, for instance, top-3 or top-5. Regarding system performance, latency evaluates whether the response time will be critical to real-time use cases, and memory utilisation monitors the scalability and efficiency of embedding the storage in the chroma vector store. Together, these metrics comprise a holistic assessment of how the bot performs in supporting product knowledge and troubleshooting efforts.

Table 3 presents the performance of the multimodal embedding RAG approach across retrieval, user experience, and system metrics.

Table 3. Table showcasing metric description and value for multimodal embedding approach.

Metric	Description	Value
Precision	Proportion of relevant passages among retrieved passages.	0.60
Recall	The proportion of relevant passages retrieved from the total relevant.	0.70
F1 Score	Balance of precision and recall	0.64
Satisfaction@5	Average user satisfaction scores out of 5.	2.3
Top-3 accuracy	Accuracy of finding relevant passages within the top 3 results.	0.45
Latency	Average response time in seconds.	5 seconds
Memory usage	Memory footprint during retrieval (in GB).	1GB

Precision (0.60) and recall (0.70) yield an F1 of 0.64-balanced but not exceptional-indicating that some retrieved passages remain less relevant. User satisfaction averaged 2.3/5, and top-3 accuracy=0.45, meaning highly relevant content was not consistently ranked at the top. System-wise, latency~5s is acceptable for semi-real-time use, and memory~1 GB supports scalability. The approach efficiently handles multimodal inputs but needs better ranking and user-perceived quality for technical support scenarios.

5.2.3. Evaluation of DPR-Enhanced RAG-based Multimodal Approach

For this approach, the main examination areas are the overall system performance and the efficiency of its retrieval and response production phases. Precision assesses the contextual accuracy of the recovered texts, reducing extraneous information and maintaining response quality. At the same time, Recall@K analyses the inclusion of the proper replies among the top-K retrieved passages, guaranteeing relevant content is accessible. The metrics taken together determine how well the retrieval phase facilitates the development of the following responses.

The BLEU Score is one measure used to assess the fluency and lexical similarity of generated replies to reference solutions. At the same time, human evaluation comprises a Likert scale through which users rate responses according to their clarity, relevance, and helpfulness. Lastly, while latency evaluates response speed and identifies opportunities for improvement in real-time use cases, user satisfaction scores combine the retrieval and generation steps to gauge overall efficacy. These measurements provide a complete view of the system's strengths and weaknesses when used together.

Table 4. Evaluation metric with expected and observed value for DPR-enhanced RAG.

Evaluation metric	Expected value	Observed value	Comments
Recall@5	90%	80%	Retrieval is moderately effective but misses context in 20% of cases.
Precision of retrieval	85%	70%	Some passages retrieved are tangential or incomplete.
BLEU score	0.75	0.55	Generated responses are moderately fluent but often lack specific detail.
Human satisfaction (likert scale)	4.0	2.5	Users find responses partially relevant, with several incomplete answers.
Overall user satisfaction score	80%	45%	Due to generation issues, overall satisfaction is significantly impacted.
Latency	≤1.5 seconds	2.5 seconds	Increased response time due to retrieval bottlenecks affects UX.

Table 4 highlights the strengths and shortcomings of the DPR-Enhanced RAG approach. Retrieval was pretty intense, with Recall@5 at 80%, but precision dropped to 70%, showing that some retrieved passages were only loosely relevant. On the generation side, BLEU was 0.55, reflecting moderate fluency but limited detail in responses. Human satisfaction was lower (2.5/5), and the overall satisfaction score fell to 45%, indicating gaps in relevance and completeness. Latency averaged 2.5 s (vs. expected 1.5 s), suggesting a retrieval bottleneck. In

short, DPR improves semantic matching, but weaker generation and slower response times limit real-time suitability without further optimisation.

5.2.4. Evaluation of Chroma-Enhanced RAG using Ollama

Evaluation in this approach relies primarily on three aspects of the RAG model: its retrieval effectiveness, response quality, and system efficiency. Retrieval effectiveness is measured using Precision@K and Recall@K, referring to the accuracy and coverage of retrieved documents in the top K results. Mean reciprocal rank assesses the period a relevant document appears in the retrieved list against improved user satisfaction, where greater ranks are superior.

Metrics like relevance score, factual correctness, and user satisfaction rating determine the contextual correctness and relevance of generated responses and overall user experience. Metrics such as response time and memory utilisation evaluate whether the system delivers the response on time and efficiently uses resources. This means aggregation offers an overall understanding of the model's strengths and weaknesses.

Table 5 indicates solid retrieval (Precision@5=0.72, Recall@5=0.68) and MRR=0.5 (relevant items appear reasonably high). Response quality was modest (Relevance 2.3/5, Factual 45%), yielding User Satisfaction 2.5/5. Latency (2.8 s) is acceptable, but 85% memory utilisation raises scalability concerns. Overall, this pipeline suits efficiency-first deployments where speed and footprint matter more than depth and factual guarantees.

Table 5. Table presenting metrics and results for chroma-enhanced RAG using ollama.

Metric	Definition	Result	Remarks
Precision@5	The proportion of relevant documents in the top 5 results.	0.72	Retrieval is effective but needs tuning.
Recall@5	The proportion of relevant documents retrieved in the top 5.	0.68	Retrieval is fairly comprehensive.
Mean reciprocal rank	The inverse rank of the first relevant document in the results.	0.5	Relevant items ranked moderately high.
Relevance score	Average score of response relevance (1-5 scale).	2.3	Indicates low relevance in responses.
Factual accuracy	Degree of factual correctness in responses.	45%	Inaccuracies observed in responses.
Overall user satisfaction	Average user satisfaction rating (1-5 scale).	2.5	Indicates low user satisfaction.
Response time	The average time taken to retrieve and generate a response.	2.8 seconds	Acceptable response time.
Memory utilization	Percentage of memory resources used during processing.	85%	High may impact scalability.

5.2.5. Evaluation of ColPali-based Multimodal RAG Approach

Evaluation of the RAG-powered system will focus on its retrieval accuracy, response quality, and usability. Regarding preprocessing and file handling, the review will determine the precision of handling multiple input formats-text and audio, with minimal error in data preparation. Measures used for embedding and indexing include cosine similarity for assessing embedding quality and appropriate memory efficiency to ensure the best possible data storage.

Regarding top-K retrieval and query matching precision, retrieval accuracy is measured, ensuring document relevance and correctness. Relevance, coherence, and translation accuracy determine the quality of the response. User experience, based on user satisfaction scores and system response time, ensures ease of use and real-time performance. Altogether, these metrics provide an all-around assessment of the system's performance in tackling multilingual and multimodal queries.

Table 6 shows that the ColPali-based multimodal RAG performed strongly across most metrics. Data preparation and embedding quality exceeded 90%, and retrieval was highly effective with Top-1 at 94% and Top-5 at 96%. Response quality was also strong, with coherence, relevance, and translation accuracy all above 90%, supported by a high user satisfaction score of 92%. The main drawback was response time, averaging 5 seconds, which limits real-time use. Overall, ColPali offers excellent retrieval and response quality for multimodal, multilingual queries but requires latency optimisation for interactive deployments.

Table 6. Evaluation metric for ColPali-based multimodal RAG.

Evaluation metric	Description	Target/ideal score	Result
Data preparation accuracy	Accuracy in renaming, text normalisation, and audio transcription.	≥90%	93%
Embedding quality score	Similarity score of embeddings for similar documents.	≥90%	92%
Top-1 retrieval accuracy	Accuracy of retrieving the top relevant document.	≥90%	94%
Top-5 retrieval accuracy	Correct document found within the top 5 retrieved results.	≥95%	96%
Query matching precision	The precision of relevant results in the top retrieved documents.	≥90%	90%
Response coherence	Logical flow and coherence of the generated response.	≥90%	91%
Relevance score	Relevance of the response to the user's query.	≥90%	95%
Translation accuracy	Accuracy in translating responses to the user's preferred language.	≥90%	92%
Overall user satisfaction score	User feedback on the chatbot's usefulness and ease of use.	≥90%	92%
Response time	Average time taken from query input to response output.	≤2 seconds	5 seconds

6. Discussions

The comparative evaluation of the five approaches highlights essential trade-offs in retrieval accuracy, response quality, multimodal adaptability, and computational efficiency [3]. The results show that no single configuration can be considered universally

optimal; each demonstrates strengths best suited to specific deployment requirements. This reinforces the need to evaluate RAG pipelines by isolated performance gains and how well they balance retrieval, generation, and system efficiency in real-world conditions [19].

The multi-embedding RAG framework delivered strong retrieval accuracy, particularly when IndicBERT or Sentence-Transformers were combined with Chroma and FAISS. However, this gain in recall was offset by weaker generation quality and higher latency. The results suggest that while multi-embedding pipelines effectively surf relevant information, additional integration mechanisms are needed to translate retrieval strength into coherent, timely answers.

The multimodal embedding approach extended the capability by fusing text and visual data, enabling richer responses for complex troubleshooting tasks. This integration, however, came at the cost of higher computation and only modest user satisfaction, showing that enhanced multimodal coverage does not automatically guarantee better end-user experience. The trade-off lies between contextual richness and efficiency, which is particularly important when considering real-world deployment constraints.

The DPR-enhanced RAG achieved deeper semantic alignment and stronger retrieval accuracy than most other pipelines, making it well-suited for technical query resolution. Yet this semantic depth came with heavy computational overhead, slowing response times and limiting scalability. This reflects a recurring tension in RAG systems-architectures that perform well on accuracy often do so at the expense of speed and accessibility.

By contrast, the chroma-enhanced RAG with ollama prioritised efficiency. It achieved a workable balance between retrieval precision and query speed, making it more suitable for applications that value throughput. However, relevance and factuality were weaker, suggesting that efficiency-first designs risk undermining response reliability if not carefully balanced [28].

The ColPali-based multimodal RAG stood out in its ability to integrate text and visuals, achieving the highest retrieval accuracy and user satisfaction among all approaches. Its effectiveness in multimodal contexts highlights the value of architectures that align evidence across modalities [6]. At the same time, response times of around five seconds and occasional translation errors in under-resourced Indic languages revealed areas for improvement. This suggests that while ColPali holds the most significant promise for multilingual, multimodal applications, optimisation for latency and low-resource languages will be essential before it can be deployed in highly interactive environments.

The experiments underline that the central trade-offs for RAG systems involve balancing retrieval accuracy with response coherence, and multimodal adaptability with computational cost [8]. Hybrid configurations such as Chroma+FAISS+mT5 offered the most balanced outcomes, but ColPali provided the most compelling demonstration of how multimodal and multilingual integration can advance product support systems. These insights emphasise that effective RAG design is less about identifying a single "best" pipeline and more about tailoring configurations to the priorities of a given use case-whether that is accuracy, efficiency, or adaptability [22].

7. Conclusions and Future Work

7.1. Conclusions

This study evaluated five approaches and 17 RAG configurations for developing a multilingual, multimodal AI bot. While setups such as Chroma+FAISS+mT5 offered a solid balance between retrieval accuracy and efficiency, ColPali-based architectures stood out for their ability to combine text and visual inputs, producing more context-aware responses. The results suggest no universal solution; instead, RAG systems need to be adapted to the specific priorities of deployment, whether real-time responsiveness, efficient use of resources, or advanced multimodal reasoning.

Beyond the technical findings, the study points to broader implications for industry. A system like this could reduce dependence on human experts, speed up troubleshooting, and provide inclusive support across different languages and document formats. The potential extends well beyond product support-healthcare could benefit from cross-lingual diagnostic tools, education could be enriched with multimodal learning platforms, and technical training could become more accessible through localised AI tutors. In short, this work deepens the understanding of how RAG configurations perform and highlights their real-world relevance in multilingual and multimodal environments.

7.2. Limitations and Assumptions

Despite its benefits, research is marked by a serious drawback. The hardware-intensive development and real-time processing required by the system make it less accessible to low-resource organisations. Additionally, the bot's performance exhibits inconsistency across Indic languages due to the lack of language-specific training data and the complexity inherent in regional dialects. The research further assumes that the users are generally technologically literate, equipped with proper equipment, and reliant on accurate, up-to-date reference sources, since misleading information could undermine the chatbot's performance.

7.3. Future Work

Future work will address the bot's limitations, making it scalable and more applicable for computational efficiency optimisation concerning latency and memory

usage to operate on modest hardware. Expanded multilingual capabilities-more so for low-resource languages-will enhance inclusiveness, but advanced knowledge graphs and semantic indexing will also boost response coherence and contextual accuracy. More importantly, the system aims to be an extensible system that could handle real-time multimodal inputs-that is, dynamic video and live speech-to target the development of resilient, adaptive, and inclusive artificial intelligence solutions for diverse multilingual and multimodal settings.

References

[1] Bhat V., Cheerla S., Mathew J., Pathak N., and et al., "Retrieval-Augmented Generation-Based Restaurant Chatbot with AI Testability," *in Proceedings of the IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications*, China, pp. 1-10, 2024. file:///C:/Users/acit2k/Downloads/2024166670.pdf

[2] Bink J., Personalized Response with Generative AI: Improving Customer Interaction with Zero-Shot Learning LLM ChatBots, Master Thesis Eindhoven University of Technology, 2023. https://research.tue.nl/en/studentTheses/personalized-response-with-generative-ai/

[3] Brown T., Mann B., Ryder N., Subbiah M., and et al., "Language Models are Few-Shot Learners," *arXiv Preprint*, vol. arXiv:2005.14165v4, pp. 1-75, 2020. https://arxiv.org/abs/2005.14165v4

[4] Chang Y., Wang X., Wang J., Wu Y., and et al., "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1-45, 2024. https://doi.org/10.1145/3641289

[5] Tin T., Xuan S., Ee W., Tiung L., and Aitizaz A., "Interactive ChatBot for PDF Content Conversation Using an LLM Language Model," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 9, pp. 1-7, 2024. https://dx.doi.org/10.14569/IJACSA.2024.0150910 5

[6] Chowdhery A., Narang S., Devlin J., Bosma M., and et al., "Palm: Scaling Language Modeling with Pathways," *Journal of Machine Learning Research*, vol. 24, no. 1, pp. 1-113, 2023. https://dl.acm.org/doi/10.5555/3648699.3648939

[7] Clark A. and Kay M., Pillow: Python Imaging Library, Python Software Foundation Documentation, https://python-pillow.org, Last Visited, 2025.

[8] Dettmers T., Pagnoni A., Holtzman A., and Zettlemoyer L., "Qlora: Efficient Finetuning of Quantized LLMs," *in Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, pp. 88-115, 2023. https://dl.acm.org/doi/10.5555/3666122.3666563

[9] Faysse M., Sibille H., Wu T., Omrani B., and et al., "ColPali: Efficient Document Retrieval with Vision Language Models," *arXiv Preprint*, vol. arXiv:2407.01449v6, pp. 1-26. https://arxiv.org/abs/2407.01449v6

[10] Gao Y., Xiong Y., Gao X., Jia K., and et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv Preprint*, vol. arXiv:2312.10992v5, pp. 1-21, 2023. https://arxiv.org/abs/2312.10997v5

[11] Jha B., Akana C., and Anand R., "Question Answering System with Indic Multilingual-BERT," *in Proceedings of the 5th International Conference on Computing Methodologies and Communication*, Erode, pp. 1631-1638, 2021. https://doi.org/10.1109/ICCMC51019.2021.9418387

[12] Joshi P., Santy S., Budhiraja A., Bali K., and Choudhury M., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," *arXiv Preprint*, vol. arXiv:2004.09095v3, pp. 1-12, 2020. https://arxiv.org/abs/2004.09095v3

[13] Lai V., Ngo N., Veyseh A., Man H., and et al., "ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning," *arXiv Preprint*, vol. arXiv:2304.05613v1, pp. 1-21, 2023. https://arxiv.org/abs/2304.05613v1

[14] Medeiros T., Medeiros M., Azevedo M., Silva, M., and et al., "Analysis of Language-Model-Powered Chatbots for Query Resolution in PDF-Based Automotive Manuals," *Vehicles*, vol. 5, no. 4, pp. 1384-1399, 2023. https://doi.org/10.3390/vehicles5040076

[15] Mousa M., Khedr A., and Idrees A., "Hierarchical Method for Automated Text Documents Classification," *The International Arab Journal of Information Technology*, vol. 22, no. 1, pp. 1-19, 2025. https://doi.org/10.34028/iajit/22/1/2

[16] NVIDIA Corporation, CUDA Deep Neural Network Library, NVIDIA Developer Documentation, https://docs.nvidia.com/deeplearning/cudnn, Last Visited, 2025.

[17] Pandya K. and Holia M., "Automating Customer Service Using LangChain: Building Custom Open-Source GPT Chatbot for Organizations," *arXiv Preprint*, vol. arXiv:2310.05421v1, pp. 1-4, 2023. https://arxiv.org/abs/2310.05421v1

[18] Python Software Foundation, Installing Packages Using Pip and Virtual Environments, Python Packaging User Guide, https://packaging.python.org/en/latest/guides/installing-using-pip-and-virtual-environments, Last Visited, 2025.

[19] Radford A., Wu J., Child R., Luan D., and et al., "Language Models are Unsupervised Multitask Learners," *OpenAI Blog*, vol. 1, no. 8, pp. 1-24, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[20] Ramjee P., Sachdeva B., Golechha S., Kulkarni S., and et al., "CataractBot: an LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients," *in Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, New York, pp. 1-31, 2025. https://dl.acm.org/doi/10.1145/3729479

[21] Reimers N. and Gurevych I., "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," *arXiv Preprint*, vol. arXiv:1908.10084v1, pp. 1-11, 2019. https://arxiv.org/abs/1908.10084v1

[22] Salemi A. and Zamani H., "Comparing Retrieval-Augmentation and Parameter-Efficient Fine-Tuning for Privacy-Preserving Personalization of Large Language Models," *in Proceedings of the International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, Padua, pp. 286-296, 2025. https://doi.org/10.1145/3731120.3744595

[23] Shawar B. and Atwell E., "Different Measurement Metrics to Evaluate a Chatbot System," *in Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, Rochester, pp. 89-96, 2007. https://aclanthology.org/W07-0313/

[24] Singh U., Vora N., Lohia P., Sharma Y., and et al., "Multilingual Chatbot for Indian Languages," *in Proceedings of the 14th International Conference on Computing Communication and Networking Technologies*, Delhi, pp. 1-5, 2023. https://doi.org/10.1109/ICCCNT56998.2023.10307978

[25] Singh V., Exploring the Role of Large Language Model-Based Chatbots for Human Resources, Master Thesis, The University of Texas at Austin, 2023. https://hdl.handle.net/2152/124540

[26] Siriwardhana S., Weerasekera R., Wen E., Kaluarachchi T., and et al., "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1-17, 2023. https://aclanthology.org/2023.tacl-1.1/

[27] Sridhar A., Audio Processing Using Pydub and Google SpeechRecognition API, GeeksforGeeks, https://www.geeksforgeeks.org/audio-processing-using-pydub-and-google-speechrecognition-api/, Last Visited, 2024.

[28] Touvron H., Lavril T., Izacard G., Martinet X., and et al., "Llama: Open and Efficient Foundation Language Models," *arXiv Preprint*, vol. arXiv:2302.13971v1, pp. 1-27, 2023. https://arxiv.org/abs/2302.13971v1

[29] Tubin C., Rodriguez J., and De Marchi A., "User Experience with Conversational Agent: A Systematic Review of Assessment Methods," *Behaviour and Information Technology*, vol. 41, no. 16, pp. 3519-3529, 2022. https://psycnet.apa.org/doi/10.1080/0144929X.2021.2001047

[30] Vakayil S., Juliet D., and Vakayil S., "Rag-Based LLM ChatBot Using Llama-2," *in Proceedings of the 7th International Conference on Devices, Circuits and Systems*, Coimbatore, pp. 1-5, 2024. https://doi.org/10.1109/ICDCS59278.2024.10561020

[31] Xue L., Constant N., Roberts A., Kale M., and et al., "MT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer," *arXiv Preprint*, vol. arXiv:2010.11934v3, pp. 1-17, 2020. https://arxiv.org/abs/2010.11934v3

[32] Zhang C., Yang Z., He X., and Deng L., "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478-493, 2020. https://doi.org/10.1109/JSTSP.2020.2987728

Mudit Garg is pursuing a B.Tech. in Computer Science and Engineering with Honours in AIML from Symbiosis Institute of Technology, Pune. He was also a research intern at SCAAI. He has worked extensively on multiple AI-driven projects, leveraging machine learning, natural language processing, and deep learning techniques to solve real-world problems. His AI-driven projects span multimodal learning, accessibility, and automation, including Sight Beyond Sight (an AI-powered accessibility website), People's Lens (face recognition for domain experts), CropPal (deep learning-based crop identification and growth-stage detection), and a RAG-based multimodal, multilingual chatbot. A three-time hackathon winner (VOIS Innovation Marathon 2022, MCCIA Edu-Fest 2022, IEEE India Council Hack 2.0), his research interests Encompass Generative AI, LLMs and VLMs, Multimodal and Multilingual AI, Computer Vision, and AI for Good. He earned the Best Paper Award at IEEE ICACTA 2023 and contributed to Bharat @ 75: Multiple Dimensions to Empower India for a Better Tomorrow.

Nandita Namboodiri is pursuing a B.Tech. in Computer Science and Engineering at Symbiosis Institute of Technology, Pune. She got to work on multiple full-stack and AI-driven projects designed to solve real-world problems. A few notable projects included SignARity, a sign language learning and recognition platform that integrated Augmented Reality with Artificial Intelligence to bridge the gap in communication, and CropPal, a deep learning-based crop identification and growth stage detection tool for precision agriculture. AlumnConnect is a website and application developed to create an online space where people from an institution can connect. She also got to work on a RAG-based multimodal and multilingual chatbot for domain-specific AI conversations.

Krisha Joshi is pursuing a B.Tech. in Computer Science and Engineering from Symbiosis Institute of Technology, Pune. She has worked on several full-stack and AI-driven projects aimed at solving real-world challenges. Her key projects include CropPal, a deep learning-based crop identification and growth stage detection tool using CNNs and transfer learning to support precision agriculture; Farm Scheme Advisor, a web-based recommendation engine that suggests relevant government schemes to farmers; and Airline Management System, a Java- and MySQL-based platform for managing flight schedules and passenger records. She was also a core contributor to AlumnConnect, a website and mobile app to connect students and alums, where she led frontend development using React Native. She contributed to a RAG-based multilingual chatbot for domain-specific conversations, combining language understanding with knowledge retrieval for contextual responses.

Karthik Krishna is pursuing a B.Tech. in Computer Science and Engineering at Symbiosis Institute of Technology, Pune. He has worked on multiple AI-driven and full-stack projects, designed to solve real-world problems. A few projects he worked on include CropPal, a deep learning-based crop identification and growth stage detection tool for precision agriculture, and SignARity, a sign language learning and recognition platform that integrates Augmented Reality with Artificial Intelligence to bridge the gap in communication. AlumnConnect is a website and application developed to create an online space where people from an institution can connect. He also got to work on a RAG-based multimodal and multilingual chatbot for domain-specific AI conversations. He has also worked on research projects such as Indian Election Analysis, a case study regarding the indian election from 1977 to 2015. He has also received merit-based scholarships in semesters.

Deepali Vora completed her Ph.D. in Computer Science and Engineering from Amity University, Mumbai. Currently working as Professor, Computer Science and Engineering, Symbiosis Institute of Technology, Pune, Symbiosis International University (Deemed), Pune, India. She has more than 25 years of experience in total in teaching, research, and Industry. She has published more than 75 research papers in reputed national, international conferences and journals. She has co-authored four books and numerous book chapters and delivered various talks in Data Science and Machine learning. She received grants from government bodies such as DST, AICTE, ISTE and the industry. She is acting as a reviewer for many International Conferences and Journals like IEEE Access, IGI Global, Springer, Inderscience, etc. She has organized many value-added courses for the benefit of the students. More than 20 students have completed their post-graduate studies under her guidance from Mumbai University. In addition to that, eight students are pursuing research (Ph.D.) under her guidance in Symbiosis International University, Pune. Her course on Deep Learning is currently available on the Unschool platform for all, and two technical blogs are available on the KnowledgeHut.com website.