World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

Unified AI Multi-modal Chatbot

P Chiranjeevi, Nagalaxmi Kalluri, Sai Saket Gurubhagavatula *, Abhishek Kuncham and Mohammed Sami

Department of CSE (Data Science), ACE Engineering College, Telangana, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 089-097

Publication history: Received on 18 March 2025; revised on 29 April 2025; accepted on 01 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0513

Abstract

In today's digital age, we are surrounded by a massive amount of information in different formats—documents, images, and videos. However, making sense of all this data in a meaningful way is still a challenge. This project proposes a smart, unified chatbot system that can understand and interact with content from multiple sources using a multi-modal Retrieval-Augmented Generation (RAG) approach powered by Google's Gemini-1.5 model. The chatbot allows users to upload PDFs, Word documents, CSV files, images containing text, and even YouTube links. It then extracts key information using techniques like OCR and video transcription, and allows users to ask questions directly about the content. What makes this system powerful is its ability to merge different types of inputs and generate accurate, context-aware answers. The entire interface is built using Streamlit, offering an easy and interactive user experience with features like real-time previews, downloadable notes, chat history, and multilingual support.The project reflects the growing need for AI systems that are intelligent, flexible, and capable of understanding information the way humans do—from all angles and in all forms.

Keywords: Multi-modal Retrieval-Augmented Generation; Gemini-1.5 Language Model; Document and Image Processing; YouTube Transcript Summarization

1. Introduction

The evolution of artificial intelligence and machine learning has significantly transformed how we process and interact with information. In particular, the emergence of large language models (LLMs) such as OpenAI's GPT and Google's Gemini has enabled intelligent systems capable of generating human-like responses from text inputs. Despite this progress, real-world scenarios often involve multiple data formats, including structured documents, visual media, and multimedia content. Existing chatbot frameworks typically specialize in one modality, limiting their utility in tasks that require a holistic understanding of context from different types of input. This calls for an innovative approach that bridges this gap—one that can understand, process, and respond to multi-modal data inputs within a single system.

Our project introduces a novel Multi-modal Retrieval-Augmented Generation (RAG) Chatbot System designed to handle diverse content inputs such as PDF, DOCX, and CSV documents, images (with Optical Character Recognition), and YouTube videos (via transcript extraction)**.** By leveraging Google's Gemini-1.5-Flash model, the chatbot can interact contextually with users, generate meaningful summaries, answer specific questions, and even translate content into multiple languages. The system supports real-time interaction and allows users to chat with their uploaded files, images, or video content as if they were conversing with a knowledgeable assistant.

The central hypothesis of this work is that integrating multi-modal data into a single generative framework enhances user comprehension, contextual relevance, and accessibility of information. Unlike conventional tools that operates in silos—handling only text, video, or image separately—this system unifies all modalities, offering a seamless experience.

* Corresponding author: Sai Saket Gurubhagavatula.

The architecture is built on a modular pipeline that includes OCR engines, YouTube transcript extractors, multilingual translation modules, and a robust Gemini-based response generator. Through this, the chatbot maintains coherence across different data sources and enables knowledge extraction at scale.

The motivation for developing such a system stems from real-world challenges in education, research, content analysis, and digital accessibility. For instance, students often need to extract information from lectures, textbooks, and visual aids simultaneously. Similarly, professionals might analyze documents and visual diagrams while referencing video tutorials. In such scenarios, switching between tools creates inefficiency and cognitive overload. This research aims to eliminate such fragmentation through a unified, AI-powered solution.

Furthermore, this work lays the groundwork for more intelligent knowledge management systems that can be integrated into educational platforms, enterprise solutions, or digital libraries. The proposed system is not only scalable and efficient but also highly adaptable, making it relevant across various domains including healthcare, legal analysis, business intelligence, and assistive technology.

2. Literature review

The integration of multi-modal capabilities into conversational AI represents a significant advancement in artificial intelligence, especially in systems requiring deep understanding across varied data types. Traditional chatbot models were initially rule-based, relying heavily on scripted templates. However, with the emergence of transformer-based architectures such as BERT and GPT, natural language understanding and generation took a giant leap forward. These models allowed for more human-like interactions but often lacked factual grounding, especially in domain-specific contexts.

To address this limitation, Retrieval-Augmented Generation (RAG) models were introduced. Unlike standalone language models, RAG systems combine retrieval mechanisms with generative models to produce contextually relevant and accurate outputs. By fetching relevant chunks of information from a pre-indexed knowledge base, these models generate responses that are better informed and less prone to hallucination. FAISS-based vector stores, semantic search, and context chunking are key techniques used in this pipeline to retrieve information effectively.

In parallel, the field of multi-modal learning has grown significantly. Models like CLIP and Flamingo have demonstrated how aligning textual, visual, and audio representations can result in a deeper and more holistic understanding of input data. Such models are now capable of interpreting images, reading texts in images using OCR, transcribing audio through Whisper, and even summarizing video content by extracting transcripts. This multi-modal capability enhances the reasoning power of AI, particularly in real-world tasks that involve diverse data formats.

3. Existing System

Existing conversational AI systems largely focus on single-modality inputs—primarily text. While traditional chatbots, powered by rule-based logic or basic natural language processing (NLP) models, offer structured responses, they lack the flexibility and depth required for intelligent knowledge extraction. Even advanced language models such as GPT-3 or BERT-based architectures, although significantly more capable, are often limited to textual understanding and generate responses solely based on pre-trained knowledge. This becomes a critical limitation when users interact with diverse data sources such as images, documents, or videos.

Moreover, while some platforms provide Optical Character Recognition (OCR) for images or transcription services for video content, these functionalities usually exist in isolation. They lack a unified system that integrates multiple data modalities under one conversational framework. Users often need to rely on separate tools for document processing, image understanding, and video analysis, which is inefficient and disjointed. There is limited capability in existing models to perform contextual reasoning across modalities, or to dynamically generate summaries and responses from heterogeneous sources in real-time.

4. Proposed System

The proposed system introduces a unified Multi-modal Retrieval-Augmented Generation (RAG**)** framework that seamlessly integrates documents, images, and video content into one interactive AI-driven interface. Unlike conventional systems, our chatbot does not treat each data type in isolation. Instead, it retrieves, processes, and merges relevant information from different modalities using a common context-aware generation model. At the core of this architecture is Google's Gemini-1.5-Flash, a state-of-the-art large language model that enables efficient generation and reasoning based on retrieved content.

This system can accept PDFs, DOCX, CSV files, YouTube video URLs, and images as inputs. Text is extracted from documents, visual content is interpreted using OCR techniques, and transcripts are derived from video sources. These extracted segments are chunked and embedded into a vector database (FAISS), enabling precise semantic search for relevant context. Once a user asks a question, the most relevant chunks are retrieved and passed to the Gemini model, which generates a response that reflects all available context across modalities.

In addition to intelligent Q&A, the system supports language translation, chat history tracking, session management, and note summarization—all within a user-friendly Streamlit interface. The modular nature of this framework ensures scalability, allowing new features and data types to be integrated in future iterations. This approach not only enhances user interaction but also holds real-world applicability in domains like education, digital libraries, research assistance, and intelligent tutoring systems.

5. Methodology

The methodology section outlines the systematic approach used in designing, developing, and implementing the Multi-modal RAG Chatbot. This project aims to enable intelligent interaction with varied data types such as documents, images, and videos by leveraging the power of Retrieval-Augmented Generation (RAG) combined with Google's Gemini model. The methodology is centered around modular processing of each data type, efficient retrieval of relevant context, and high-quality, contextual response generation. This pipeline is integrated into a user-friendly Streamlit interface, enabling real-time Q&A across modalities.

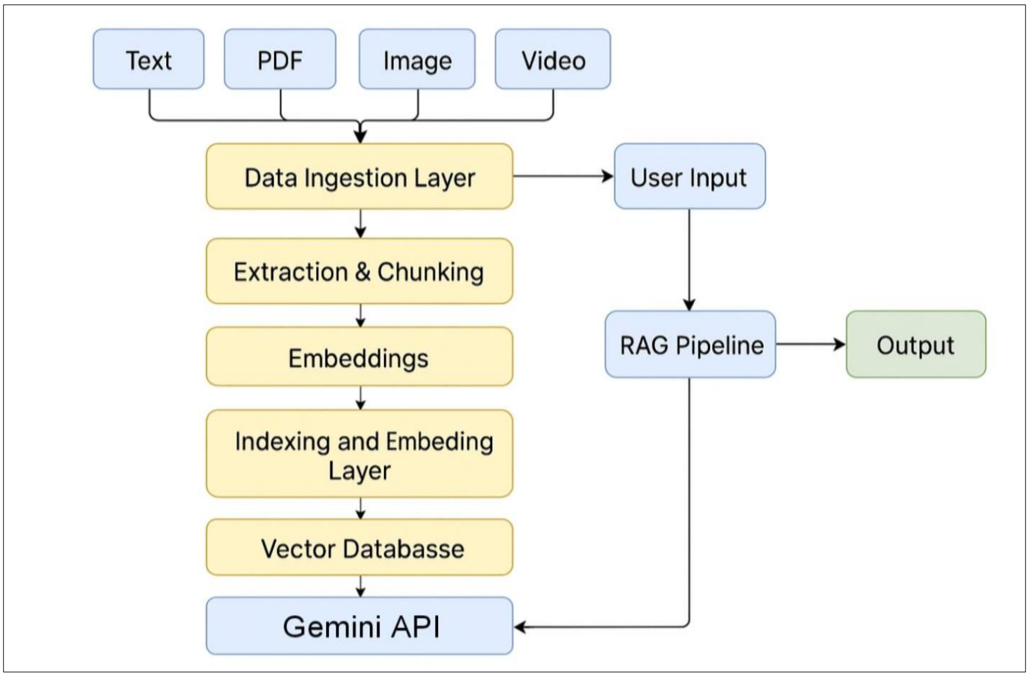


Figure 1 Methodology

5.1. System Architecture

The architecture of the Multi-modal RAG Chatbot is designed to process and understand various types of data—text, images, and video transcripts—through a unified pipeline. The system employs a modular design with independent processors for each data type, all connected to a centralized retrieval-augmented generation engine powered by Google's Gemini model. This ensures scalable, extensible, and intelligent information extraction and interaction.

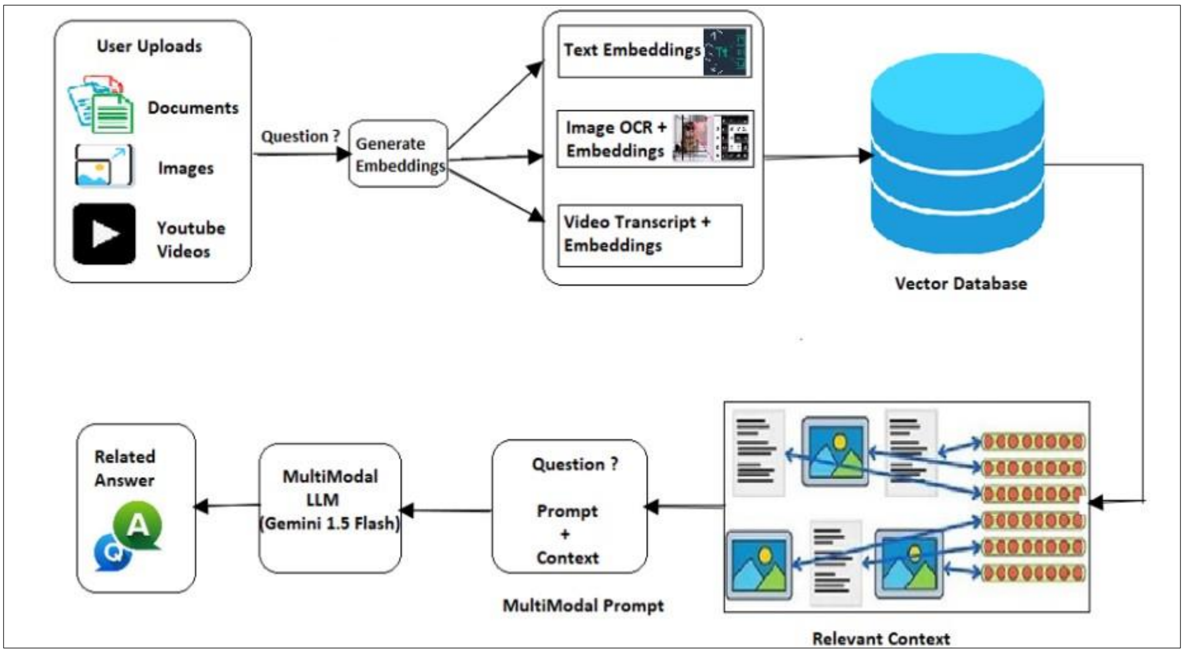


Figure 2 System Architecture

5.1.1 Input Acquisition Layer

This is the entry point of the system where users provide data in different formats. The input types supported are:

- Documents (PDF, DOCX, CSV)
- Images (JPG, PNG, etc.)
- Videos (via YouTube URLs).

Each input type is routed to its respective processing module to begin extraction.

5.1.2 Preprocessing and Extraction Layer:

Each file format is handled by a dedicated processor:

- **Document Processor**: Uses PyMuPDF, python-docx, and pandas to extract readable content. Text is cleaned and split into semantically meaningful chunks for downstream tasks.
- **Image Processor:** Uses pytesseract for Optical Character Recognition (OCR).Extracts embedded text from diagrams, handwritten notes, or printed material.
- **YouTube Processor:** Fetches transcripts using the youtube-transcript-api.Optionally summarizes long transcripts using Gemini before storing.

5.1.3 Embedding and Vector Storage:

To enable semantic search, extracted data from all input types are embedded into high-dimensional vectors:

- **Embedding Model:** GoogleEmbedding or OpenAIEmbedding.
- **Vector Database:** FAISS (Facebook AI Similarity Search) is used to store and search text chunks efficiently.

Enables fast retrieval of top-k relevant text chunks based on semantic similarity to the user's query.

5.1.4 Retrieval-Augmented Generation (RAG) Core:

- **Query Processing:** User question is matched against vector representations to fetch relevant chunks.
- **Contextual Retrieval:** Based on user questions, the system searches the vector store (FAISS) to retrieve the top-k most relevant chunks. These context snippets form the input for the generative model.
- **Gemini-Driven Generation:** Gemini-1.5-Flash processes the retrieved context and user query. Generates a coherent, context-aware, and often multilingual answer (supporting Hindi, Tamil, Korean, etc.).

5.1.5 User Interface and Interaction Layer:

The Streamlit-based UI ensures an intuitive and interactive experience:

- **File Uploads and Previews:** Users can upload files or paste YouTube URLs in designated tabs. A preview of extracted content is displayed before starting the chat.
- **Chat System:** Integrated chatbot allows users to ask questions based on uploaded data. Conversation history is saved using st.session_state.
- **Notes and Translations:** Users can download AI-generated notes for study. Option to translate responses into multiple languages.

5.1.6 Session and History Management:

- **Session State:** Preserves files, questions, answers, and extracted content.
- **Chat Logs:** Enables revisiting prior interactions.
- **Multi-turn Conversations:** Gemini maintains context across questions for smoother dialogue.

5.1.7 Integration and Modularity:

Each module is decoupled to allow easy scaling and debugging. APIs and service calls connect individual components. Designed for future expansion (e.g., audio input, real-time video analysis).

5.1.8 Evaluation and Testing:

Testing ensures robustness across use cases:

- **Unit Testing:** Each processor module was tested independently.
- **Integration Testing:** Verified connections between vector store, Gemini model, and frontend.
- **Load Testing:** Assessed system performance with large documents and long transcripts.
- **User Feedback:** Initial trials gathered insights for improving UI and answer accuracy.

The proposed methodology efficiently integrates multi-modal data processing through a unified RAG framework, enabling accurate information extraction from documents, images, and videos. With FAISS-based semantic retrieval and Gemini-powered generation, the system delivers context-aware, intelligent responses. It supports multilingual outputs and session management, enhancing user experience. This structured approach provides a scalable solution for educational tools, content summarization, and real-world AI applications that require reasoning across diverse data formats.

6. Results and Discussion

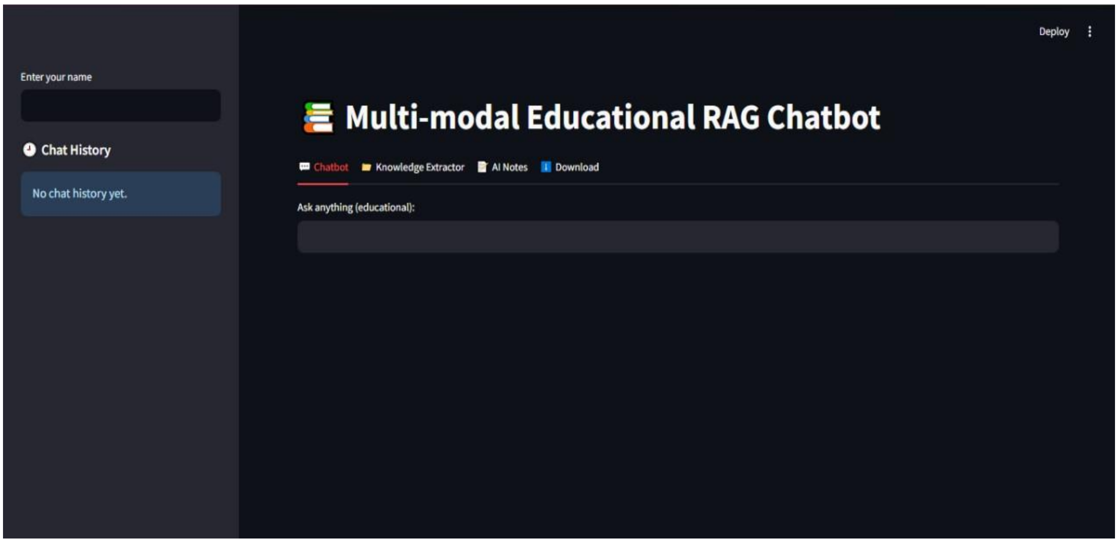
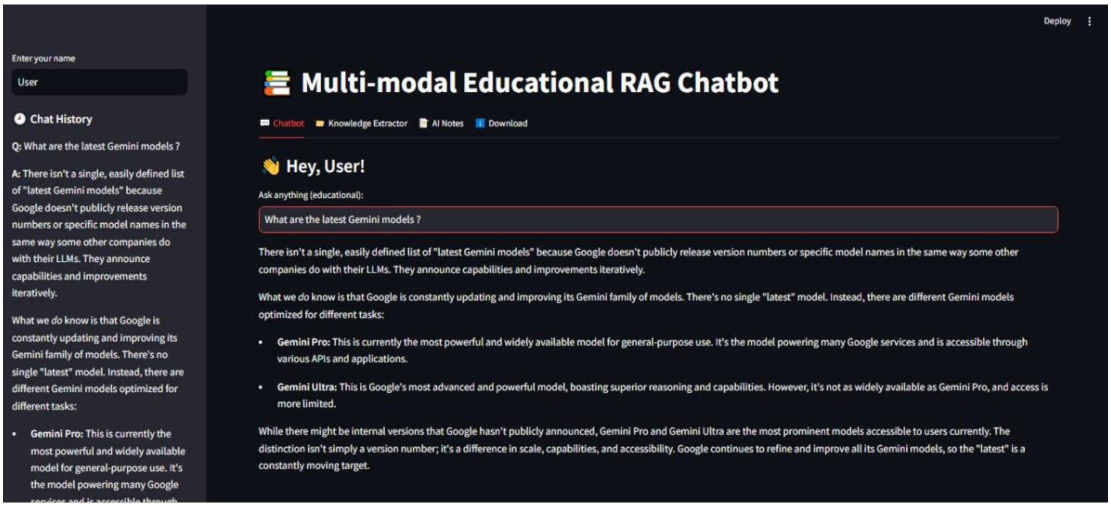


Figure 3 User Interface



Q: What are the latest Gemini models ?

A: There isn't a single, easily defined list of "latest Gemini models" because Google doesn't publicly release version numbers or specific model names in the same way some other companies do with their LLMs. They announce capabilities and improvements iteratively.

What we *do* know is that Google is constantly updating and improving its Gemini family of models. There's no single "latest" model. Instead, there are different Gemini models optimized for different tasks:

- **Gemini Pro:** This is currently the most powerful and widely available model for general-purpose use. It's the model powering many Google services and is accessible through

What are the latest Gemini models ?

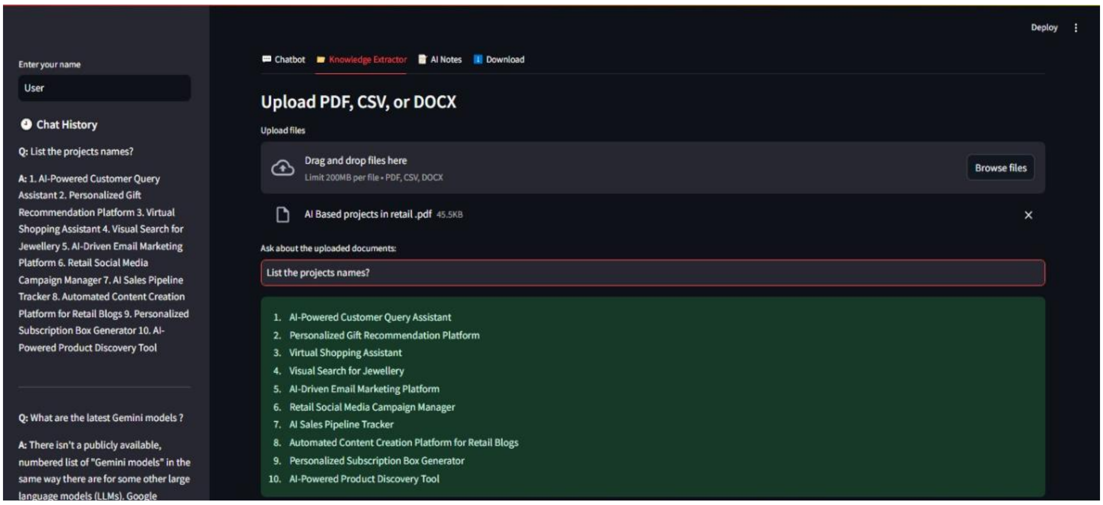
There isn't a single, easily defined list of "latest Gemini models" because Google doesn't publicly release version numbers or specific model names in the same way some other companies do with their LLMs. They announce capabilities and improvements iteratively.

What we *do* know is that Google is constantly updating and improving its Gemini family of models. There's no single "latest" model. Instead, there are different Gemini models optimized for different tasks:

- **Gemini Pro:** This is currently the most powerful and widely available model for general-purpose use. It's the model powering many Google services and is accessible through various APIs and applications.
- **Gemini Ultra:** This is Google's most advanced and powerful model, boasting superior reasoning and capabilities. However, it's not as widely available as Gemini Pro, and access is more limited.

While there might be internal versions that Google hasn't publicly announced, Gemini Pro and Gemini Ultra are the most prominent models accessible to users currently. The distinction isn't simply a version number; it's a difference in scale, capabilities, and accessibility. Google continues to refine and improve all its Gemini models, so the "latest" is a constantly moving target.

Figure 4 Chatbot Response Generation



Q: List the projects names?

A: 1. AI-Powered Customer Query Assistant 2. Personalized Gift Recommendation Platform 3. Virtual Shopping Assistant 4. Visual Search for Jewellery 5. AI-Driven Email Marketing Platform 6. Retail Social Media Campaign Manager 7. AI Sales Pipeline Tracker 8. Automated Content Creation Platform for Retail Blogs 9. Personalized Subscription Box Generator 10. AI-Powered Product Discovery Tool

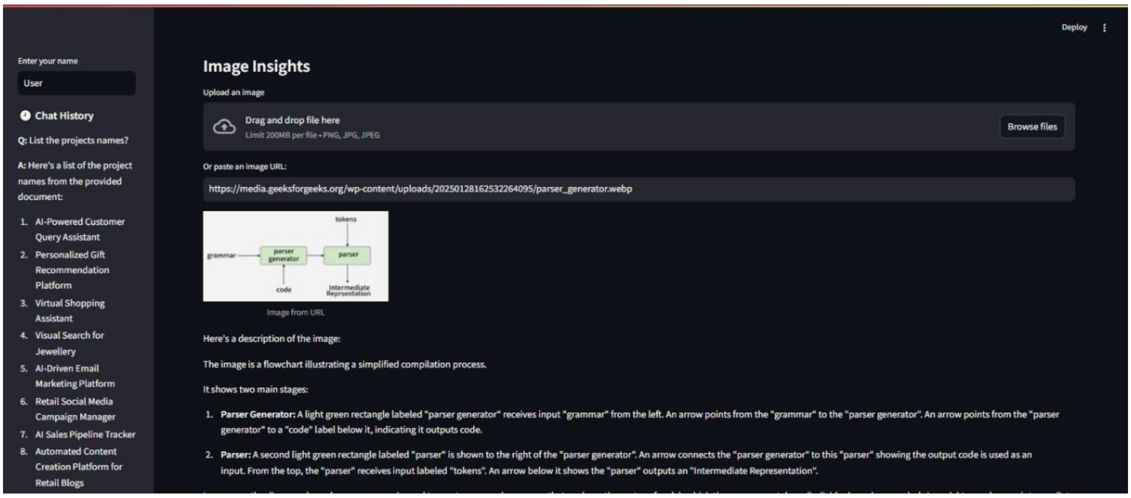
Q: What are the latest Gemini models ?

A: There isn't a publicly available, numbered list of "Gemini models" in the same way there are for some other large language models (LLMs). Google

List the projects names?

1. AI-Powered Customer Query Assistant
2. Personalized Gift Recommendation Platform
3. Virtual Shopping Assistant
4. Visual Search for Jewellery
5. AI-Driven Email Marketing Platform
6. Retail Social Media Campaign Manager
7. AI Sales Pipeline Tracker
8. Automated Content Creation Platform for Retail Blogs
9. Personalized Subscription Box Generator
10. AI-Powered Product Discovery Tool

Figure 5 PDF Response Generation



Enter your name: User

Chat History

Q: List the projects names?

A: Here's a list of the project names from the provided document:

1. AI-Powered Customer Query Assistant
2. Personalized Gift Recommendation Platform
3. Virtual Shopping Assistant
4. Visual Search for Jewellery
5. AI-Driven Email Marketing Platform
6. Retail Social Media Campaign Manager
7. AI Sales Pipeline Tracker
8. Automated Content Creation Platform for Retail Blogs

Image Insights

Upload an image

Or paste an image URL: https://media.geeksforgeeks.org/wp-content/uploads/20250128162532264095/parser_generator.webp

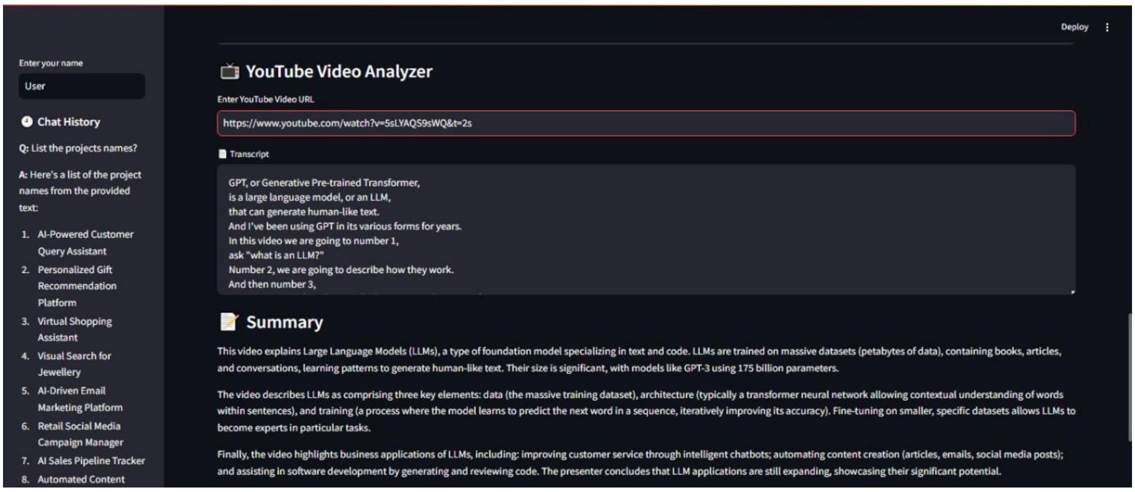
Here's a description of the image:

The image is a flowchart illustrating a simplified compilation process.

It shows two main stages:

1. **Parser Generator:** A light green rectangle labeled "parser generator" receives input "grammar" from the left. An arrow points from the "grammar" to the "parser generator". An arrow points from the "parser generator" to a "code" label below it, indicating it outputs code.
2. **Parser:** A second light green rectangle labeled "parser" is shown to the right of the "parser generator". An arrow connects the "parser generator" to this "parser" showing the output code is used as an input. From the top, the "parser" receives input labeled "tokens". An arrow below it shows the "parser" outputs an "Intermediate Representation".

Figure 6 Image Response Generation



Enter your name: User

Chat History

Q: List the projects names?

A: Here's a list of the project names from the provided text:

1. AI-Powered Customer Query Assistant
2. Personalized Gift Recommendation Platform
3. Virtual Shopping Assistant
4. Visual Search for Jewellery
5. AI-Driven Email Marketing Platform
6. Retail Social Media Campaign Manager
7. AI Sales Pipeline Tracker
8. Automated Content

📺 YouTube Video Analyzer

Enter YouTube Video URL: https://www.youtube.com/watch?v=5sLYAQS9sWQ&t=2s

📄 Transcript

GPT, or Generative Pre-trained Transformer,
is a large language model, or an LLM,
that can generate human-like text.
And I've been using GPT in its various forms for years.
In this video we are going to number 1,
ask "what is an LLM?"
Number 2, we are going to describe how they work.
And then number 3,

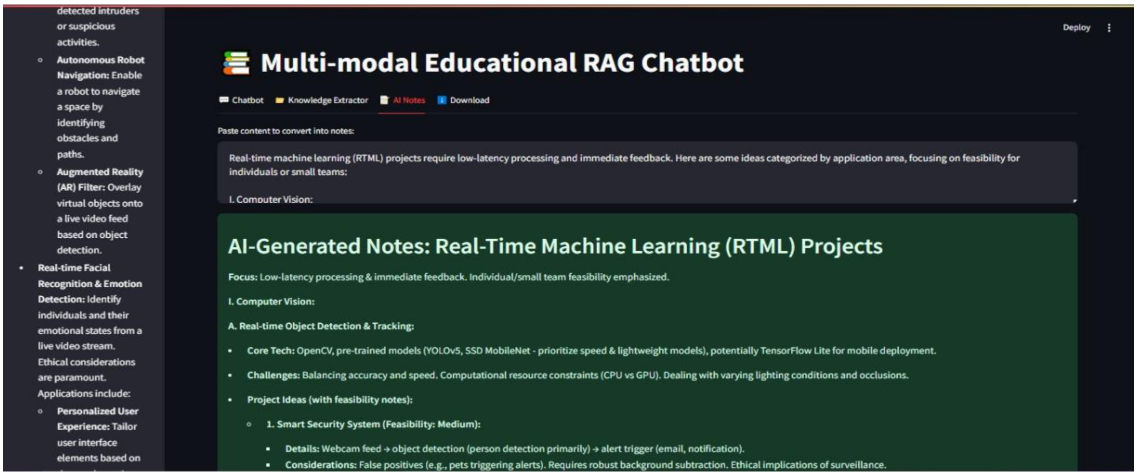
📝 Summary

This video explains Large Language Models (LLMs), a type of foundation model specializing in text and code. LLMs are trained on massive datasets (petabytes of data), containing books, articles, and conversations, learning patterns to generate human-like text. Their size is significant, with models like GPT-3 using 175 billion parameters.

The video describes LLMs as comprising three key elements: data (the massive training dataset), architecture (typically a transformer neural network allowing contextual understanding of words within sentences), and training (a process where the model learns to predict the next word in a sequence, iteratively improving its accuracy). Fine-tuning on smaller, specific datasets allows LLMs to become experts in particular tasks.

Finally, the video highlights business applications of LLMs, including: improving customer service through intelligent chatbots; automating content creation (articles, emails, social media posts); and assisting in software development by generating and reviewing code. The presenter concludes that LLM applications are still expanding, showcasing their significant potential.

Figure 7 YouTube Video Summarization



detected intruders or suspicious activities.

- **Autonomous Robot Navigation:** Enable a robot to navigate a space by identifying obstacles and paths.
- **Augmented Reality (AR) Filter:** Overlay virtual objects onto a live video feed based on object detection.

- **Real-time Facial Recognition & Emotion Detection:** Identify individuals and their emotional states from a live video stream. Ethical considerations are paramount. Applications include:
  - **Personalized User Experience:** Tailor user interface elements based on

Multi-modal Educational RAG Chatbot

Paste content to convert into notes:

Real-time machine learning (RTML) projects require low-latency processing and immediate feedback. Here are some ideas categorized by application area, focusing on feasibility for individuals or small teams:

I. Computer Vision:

AI-Generated Notes: Real-Time Machine Learning (RTML) Projects

Focus: Low-latency processing & immediate feedback. Individual/small team feasibility emphasized.

I. Computer Vision:

A. Real-time Object Detection & Tracking:

- **Core Tech:** OpenCV, pre-trained models (YOLOv5, SSD MobileNet - prioritize speed & lightweight models), potentially TensorFlow Lite for mobile deployment.
- **Challenges:** Balancing accuracy and speed. Computational resource constraints (CPU vs GPU). Dealing with varying lighting conditions and occlusions.
- **Project Ideas (with feasibility notes):**
  - **1. Smart Security System (Feasibility: Medium):**
    - **Details:** Webcam feed → object detection (person detection primarily) → alert trigger (email, notification).
    - **Considerations:** False positives (e.g., pets triggering alerts). Requires robust background subtraction. Ethical implications of surveillance.

Figure 8 AI Generated Notes

7. Conclusion

The development of the Multi-modal RAG Chatbot marks a significant advancement in the field of AI-driven educational and information retrieval systems. By integrating diverse data types—text documents, images, and YouTube videos—into a unified Retrieval-Augmented Generation framework, the system demonstrates the power of multi-modal understanding and response generation. Leveraging Google's Gemini-1.5-Flash model and FAISS vector-based retrieval, the chatbot can interpret, reason, and converse with users based on complex, unstructured data inputs across various formats.The project also emphasizes user personalization through features like session history, multilingual translation, and note generation. System testing across functionality, integration, performance, and user experience confirms the reliability, scalability, and responsiveness of the chatbot.

In conclusion, the proposed chatbot offers a robust solution for intelligent, multi-format information extraction and interaction. Its modular design and scalability pave the way for future enhancements and potential deployment across diverse sectors including education, research, digital content curation, and healthcare. This work contributes not only a functional prototype but also a strong foundation for future research in the area of multi-modal conversational AI systems.

Compliance with ethical standards

Disclosure of conflict of interest

There is no conflict of interest.

References

[1] Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Augenstein, I. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401. https://arxiv.org/abs/2005.11401.

[2] Google, "Gemini 1.5 Technical Report," Google DeepMind, Feb. 2024. [Online]. Available: https://deepmind.google/technologies/gemini/

[3] Lee, M. Y. (2023). Building Multimodal AI Chatbots. arXiv preprint arXiv:2305.03512. https://arxiv.org/abs/2305.03512

[4] Google Cloud. (n.d.). What is Retrieval-Augmented Generation (RAG)? https://cloud.google.com/use-cases/retrieval-augmented-generation

[5] ACS Publications. (2024). Using Retrieval Augmented Generative AI Chatbots to Support and Enhance Chemistry Education. https://pubs.acs.org/doi/10.1021/acs.jchemed.4c00765

[6] ResearchGate. (2024). Bring Retrieval Augmented Generation to Google Gemini via External API: An Evaluation with BIG-Bench Dataset. https://www.researchgate.net/publication/380486834

[7] ResearchGate. (2024). Enhancing PDF Information Retrieval Through a Gemini Pro LLM-Powered Chatbot. https://www.researchgate.net/publication/388992399

[8] Kumar, A., Gupta, R. (2022). AI-Based Multi-Modal Chatbot Interactions for Enhanced User Engagement. https://rjpn.org/ijcspub/papers/IJCSP24C1126.pdf

Author's short biography

Dr. P Chiranjeevi Dr. P Chiranjeevi is working as an HOD & Associate Professor in the Department of CSE (DATA SCIENCE) at ACE Engineering College, Hyderabad (India). He had completed Ph.D at JNTUH University at Hyderabad (India). He is in software Industry for more than 1 year. He is in teaching profession for more than 18 years. His main area of interest includes Opinion Mining, Sentiment Analysis and NLP.	

Nagalaxmi Kalluri I am K Nagalaxmi, a B.Tech student in Computer Science and Engineering (Data Science) with a strong interest in Machine Learning and Data Science. As an undergraduate researcher, I am keen on exploring trending technologies and innovative techniques in predictive analytics, and intelligent systems to solve real-world challenges.	
Sai Saket Gurubhagavatula G Sai Saket is currently pursuing a B.Tech in Computer Science and Engineering with a focus on Data Science. He has developed a strong interest in data-driven technologies, particularly in machine learning, artificial intelligence, and statistical analysis. Throughout his academic journey, he has worked on various research projects and practical applications related to data modeling, predictive analytics, and deep learning. Passionate about leveraging data science to solve real-world problems	
Abhishek Kuncham K Abhishek is currently pursuing a B.Tech in Computer Science and Engineering (Data Science). His research interests include Deep Learning, with a focus on leveraging advanced computational techniques for data-driven applications. As an undergraduate researcher, he is passionate about exploring machine learning models to solve real-world challenges, particularly in intelligent automation and pattern recognition.	
Mohammed Sami I am Mohammed Sami, currently pursuing a B.Tech in Computer Science and Engineering with a specialization in Data Science. My academic journey has been driven by a deep interest in computer science, particularly in the field of machine learning. I have gained valuable experience. As an undergraduate, I am passionate about using data science to tackle real-world problems. I look forward to continuing to explore and contribute to this rapidly evolving field.	