International Journal of Advanced Scientific Research and Engineering Trends

ISSN: 2456-0774

Volume 10 Issue 1s (2026) | Pages: 276-283

Multi-modal AI Chatbot for Text and Image Interaction

Pratham Modi

Bharati Vidyapeeth Deemed to be University College of Engineering, Pune, India

Peer Review Information

Type: *Article*
Received: *02 April 2026*
Revised: *28 April 2026*
Accepted: *05 June 2026*
Published: *23 June 2026*

Abstract

The past few years have seen a rise in chatbot applications, although the majority of them are limited to text-based interactions. In practice, users often use visual media, like screenshots or photographs as an effective means of presenting information. The paper provides a multi-modal chatbot that is able to handle textual inputs, analyze uploaded images, and create new images based on descriptive questions. Gemini API was chosen as the foundation of the system because it has an intrinsic support of both textual and visual input, and because of its available developer documentation. Streamlit was used to construct the front-end interface, and it can be deployed quickly without involving hefty computing resources. The testing proved that the system reacted faithfully to usual text queries, generated meaningful interpretations of diverse uploaded pictures such as diagrams, screenshots and photographs, and generated visually coherent pictures of simple descriptive prompts.

The main goal was not to create a production-scale AI platform, but to investigate whether it was possible to integrate text chat, image analysis, and image generation in one, lightweight interface.

Keywords: Chatbot; Gemini API; Multi-Modal AI; Image Understanding; Streamlit; Conversational Interface

How to Cite This Article

Modi, P. (2026). Multi-modal AI chatbot for text and image interaction. *International Journal of Advanced Scientific Research and Engineering Trends*, 10(1s), 276–283.

Introduction

The spread of artificial intelligence into the daily application has turned AI-controlled tools into a part of the contemporary processes. AI is present in the background of digital life: it can assist with navigation, provide automated customer support, etc. Chatbot systems are some of the most common types of AI. Although most of them are in use, most of the current chatbots are limited to text based communication and hence this limits their use in situations where visual information forms the main content of the interaction.

The use of sharing screenshots, diagrams and photographs is a common tool used by users in both academic and professional settings where the textual description is not enough. This mismatch between human behaviour and the capability of chatbots inspires the creation of a system that is able to handle both textual and visual input concurrently. This limitation has been addressed by multi-modal AI models, which can utilize both text and image information in a single framework.

An examination of available multi-modal tools has shown that the majority of them need specialised hardware setups (such as dedicated GPUs) or have intricate software installation processes that cannot be performed by regular users or students operating on a standard laptop. Gemini API is a Google created text and image processing API with a relatively simple developer documentation, and is a good alternative to creating a lightweight, accessible multi-modal application [7].

The following paper explains the design, implementation, and evaluation of a multi-modal chatbot prototype, which is a combination of three fundamental functions: text-based conversational responses, image analysis, and image generation based on text prompts. The scope of the project is limited on purpose to what can be practically implemented with a standard consumer hardware without purporting to be as powerful as a research-grade system.

Problem Definition

Modern chatbot applications are still largely text-based. Conventional chatbots do not process or respond to screenshots or photographs when end-users are required to communicate using diagrams, screenshots, or photographs. Moreover, to get access to AI-generated imagery, users have to change platforms that are specialized in image generation. This disaggregation causes tension in user processes and overall low productivity.

The main issue that this paper is going to tackle can be described as follows: How can one combine text conversation, image analysis and image generation in one, user-friendly interface without burdening the user with too much complexity or demanding the use of specific computing infrastructure?

Motivation

The growing use of digital communication tools in education and workplaces has demonstrated a very basic drawback to the existing AI-driven interfaces: the inability to process and react to visual content. Although textual chatbots have grown up significantly, in real life, text is rarely used in communication. Students post pictures of written notes, engineers are able to discuss annotated screen shots, designers explain visual ideas that cannot be properly explained in words. This lack of user behaviour-chatbot capability interaction acted as the main driving force behind this project.

Moreover, the currently available multi-modal solutions are usually oriented towards research settings and require a lot of hardware and technical knowledge to implement. The lack of lightweight, student-friendly solutions that can combine text communication, image interpretation and image generation into one platform led to the consideration of a simpler and more realistic solution. The intention was to show that significant multi-modal interaction was possible with ordinary consumer devices with publicly available APIs, and not just with advanced understanding of machine learning infrastructure.

Also the high rate of development of multi-modal APIs like Gemini has provided a chance to develop functional applications with a comparatively small amount of development effort. This opportunity to build a working prototype, provided a real learning experience, and a tangible demonstration of what modern APIs can achieve at the level of the student project. The incentive, however, was more pedagogical than technical: to make a connection between theory and practice by creating something concrete, practical, and scalable.

Objective and Scope of the Research Work

Objectives

The objectives were very simple:

- Have the chatbot respond to simple text messages.
- Add a functionality of uploading images by users.
- Have the chatbot explain or analyse the picture.
- Allow image generation based on prompts

- Maintain reasonable response time.
- Make it simple using streamlit
- Process errors in such a way that the app is not brought down.

These goals contributed to maintaining the development in focus without experimenting with too many additional features.

Research Work Scope

Having reviewed various tools and articles, some gaps were evident:

1. The majority of chatbots can respond only to text and are not able to view images.
2. Sometimes it is slow to switch between text mode and image mode.
3. when an image appears out of the blue, many models forget any previous conversation.
4. Deployment procedures tend to be too lengthy or difficult to follow.
5. Multi-modal systems that are easy to use are not very popular among novices.
6. Due to these gaps, a small project that unites text chatting, image understanding and image creation in a single location is possible.

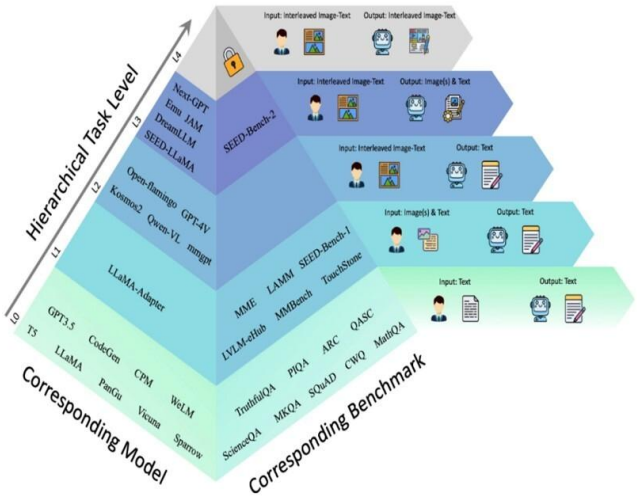

***Fig. 1**. Performance benchmark of different multi-modal AI models*

Literature Survey

A number of papers and web articles were consulted prior to commencing. The initial AI systems such as BERT and GPT-2 primarily dealt with text-only tasks and could not even visualize pictures [6]. Subsequently, CLIP was developed in which the model was trained on a large amount of image-caption data to learn to associate images with text [1]. The concept sounded good, but CLIP was too cumbersome to be used on a regular system.

Then there came the introduction of BLIP that can mostly be used in image captioning or answering questions based on an image [2]. Flamingo gained popularity due to its ability to blend text and visual activities to a large extent [3]. Once again, the majority of these models were primarily designed to work in research laboratories or individuals who have powerful GPUs.

Certain surveys elaborated that multi-modal models tend to have problems such as the loss of context, slow processing, and complicated API use [5]. A single article about BlenderBot 3 talked about the forgetting of past dialogue by chatbots due to excessive interactions [4].

Thus, all of this made it evident that despite the presence of numerous advanced tools, a simple, easy to run multi-modal chatbot can still be useful to students and beginners.

Methodology

Rather than putting down all the information in some ideal format, the steps undertaken in the development process are described below.

It all started with text chat support. Streamlit was linked with the Gemini API and the test was conducted on whether it responds correctly. Once the basic set up was in place the image- upload feature was implemented. Each time a photo is uploaded, it is sent to the vision model of Gemini and it provides a brief description of the contents in the photo.

The second step was to add the image-generation component. With more straightforward prompts such as “draw a cartoon cat" or make a beach view, the outcomes were quite satisfactory. The model also failed a couple of times when the prompt was not very clear, a try-except block was included to ensure that the app would not crash.

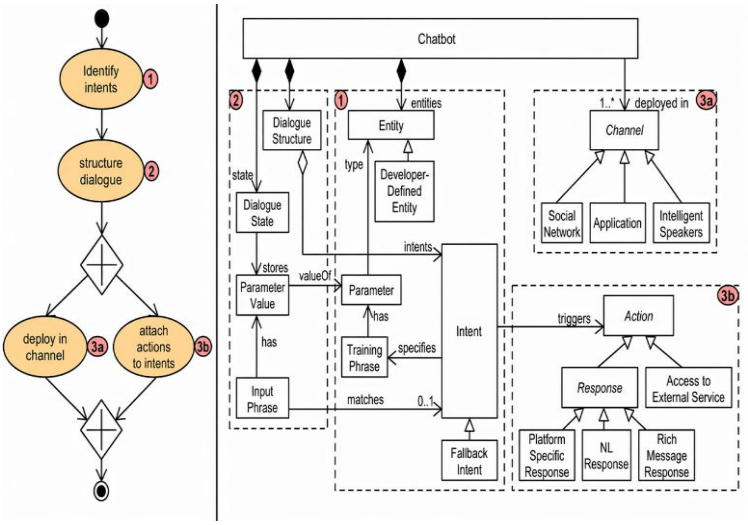

Fig. 2. *Architecture Diagram of the Multi-Modal AI Chatbot*

The entire process is approximately:

- The user uploads a picture or sends a message
- The system verifies the input
- Forwards to correct function.
- Shows the output

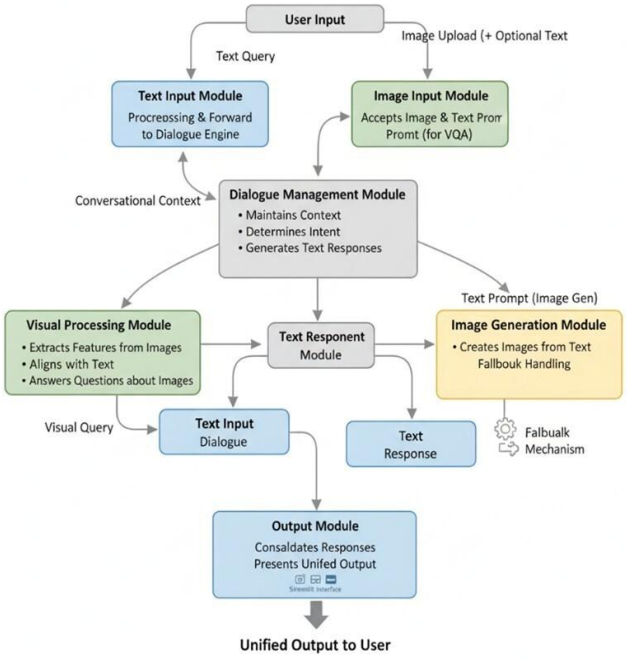

Fig. 3. *Operational Phases of the Multi-Modal AI Chatbot*

Importance

The system designed within this project can be practically useful within various groups of users. Lecture diagrams or handwritten notes can be uploaded by the students and they can be explained in real time, eliminating the reliance on synchronous support. The image analysis feature can be used by customer support teams to analyze product pictures posted by customers and respond more accurately. The image generation module can be used by creative users and designers to quickly generate visual concepts based on textual descriptions.

Educationally, the project offers a practical experience in multi-modal API integration, practical error handling, and the practical constraints of generative AI systems of which knowledge is becoming more and more relevant in modern software development scenarios.

Results

The system implemented was tested on all the three functional aspects: text-based conversation, image analysis and image generation.

Text Chats

The conversational module showed a good performance on standard queries, including keeping up the context of two to three consecutive conversations. There is no persistent memory storage in the system and therefore the context retention decays with long sessions. Response quality was uniform within its operational limits and was contextually acceptable.

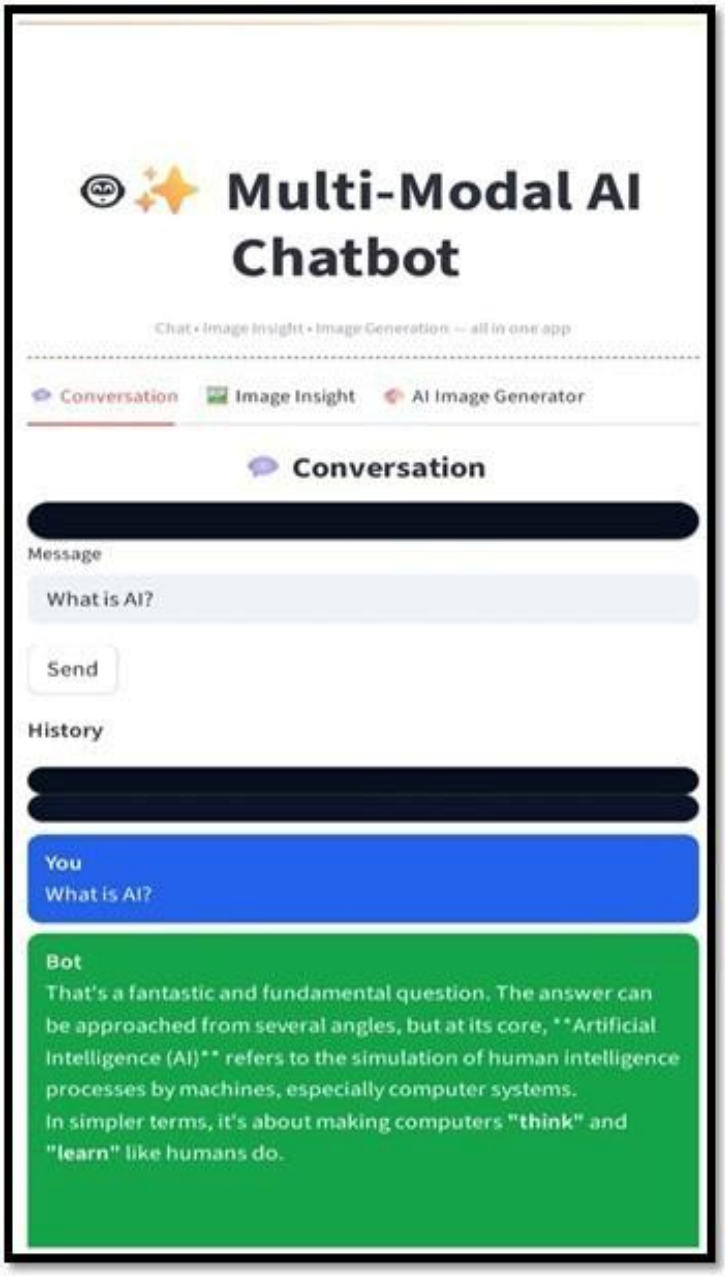

Fig. 4. *Sample Results of Text Conversation Mode.*

Image Analysis

The image analysis software was experimented using a wide range of uploads, such as class diagrams, write-up notebook pages, plant specimen and animal photos. In both instances, the Gemini vision model provided descriptive and contextually pertinent analyses. Users are also able to ask certain questions regarding the images posted, including questions about the mood, composition or content of a photograph.

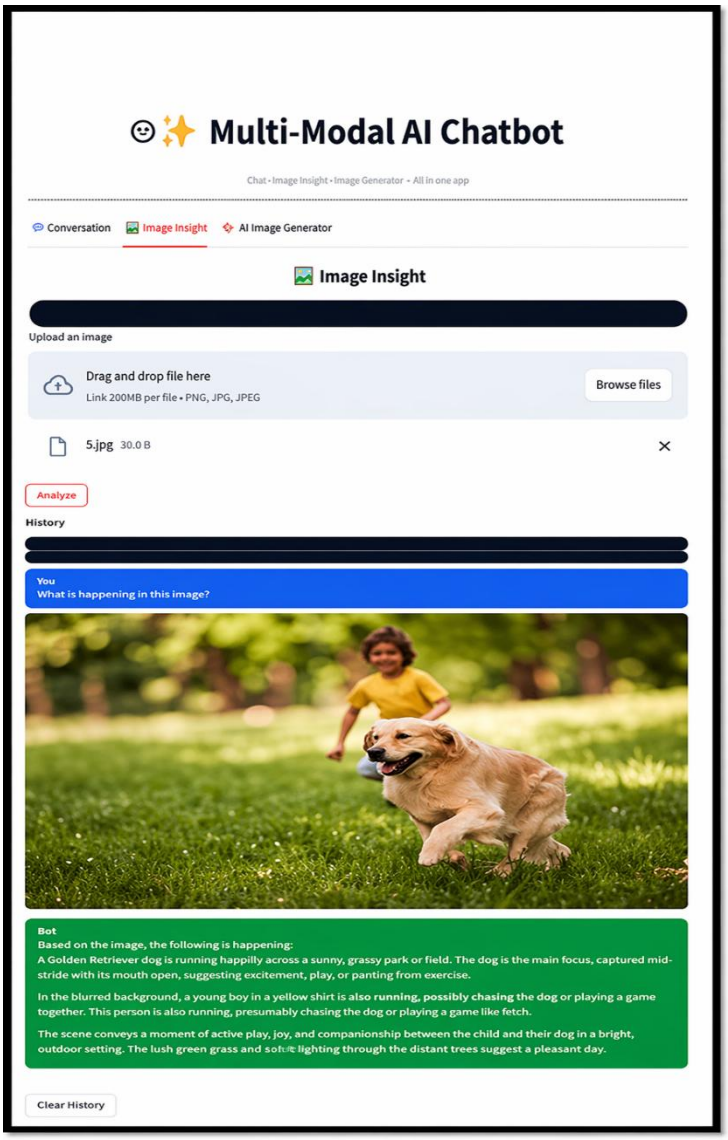

Fig. 5. *Sample Image Insight Results.*

Image Generation

Text to image generation was accurate with well defined descriptive prompts such as the description of landscapes, cartoon animals and plain drawings. Abstract or technically challenging prompts like architectural schematics or neural network diagrams saw performance drop significantly, as expected due to the established shortcomings of the underlying generation model.

In general, the findings validate that the system achieves its declared goals on the prototype level that can be used to demonstrate the system performance in the academic context and be further elaborated.

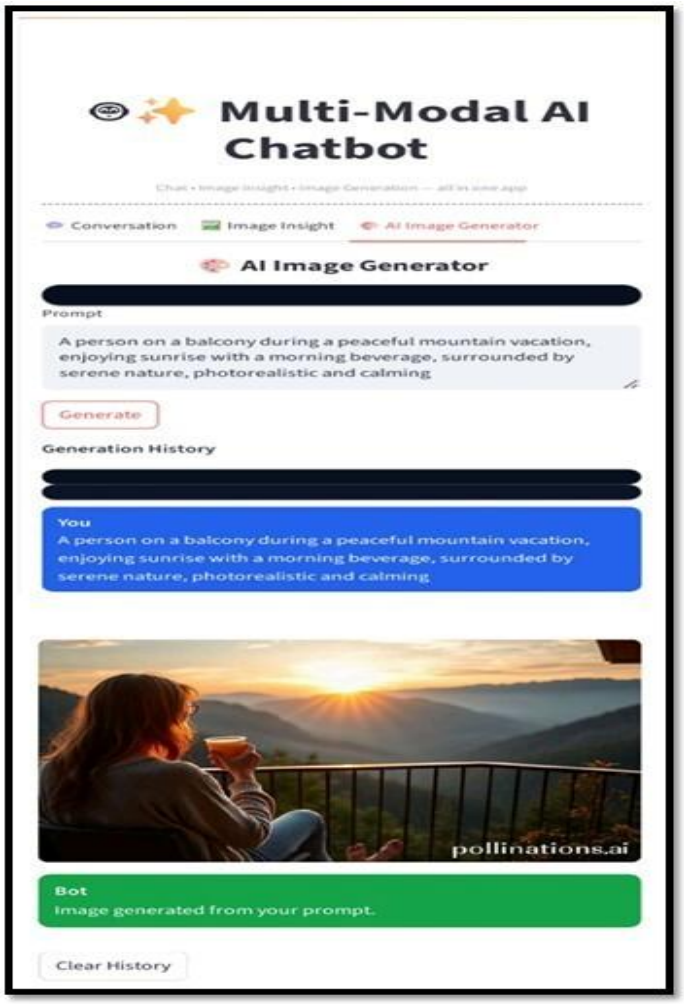

Fig. 6. *Text-to-Image Generation Results*

Future Scope

It has been concluded that a number of future development directions can be identified based on the present implementation:

1. Voice input and output: The integration of voice input and output would greatly contribute to accessibility, and hands-free communication would be more consistent with the natural patterns of communication.
2. Refinements of the interface such as enhanced visual design with iconography and colour differentiation would enhance user interaction and ease of use.
3. Video input processing may further enhance the analytical capabilities of the system but latency and model constraints would have to be well controlled.
4. Persistent conversation memory which can be deployed through a lightweight database would enable the system to preserve the context over longer sessions.
5. A mobile-optimised version would also expand the user base as there is a massive use of smartphones in communication and retrieval of information.
6. Greater consistency in image generation might be realised via timely engineering improvement and incorporation of increasingly specialised generation APIs.
7. Multiple image styles (such as photorealistic, sketch-based, and cartoon rendering modes) would increase the opportunities of creative applications.

Among those directions, voice interaction and mobile deployment will be the most influential near-term enhancements as they would significantly increase the accessibility and usefulness of the system.

Acknowledgments

The author would like to express their sincere gratitude to Bharati Vidyapeeth Deemed to be University College of Engineering, Pune for providing the necessary resources and guidance to carry out this research.

We also thank our mentors and faculty members for their valuable support, suggestions, and encouragement throughout the development of the project.

References

1. Radford et al., “CLIP: Connecting Images and Text,” OpenAI, 2021.
2. J. Li et al., “BLIP: Bootstrapping Language-Image Understanding,” 2022.
3. J. Alayrac et al., “Flamingo: A Visual Language Model,” DeepMind, 2022.
4. X. Li et al., “Survey on Multimodal Conversational Systems,” 2023.
5. Google Research, “Gemini API Documentation,” 2024.
6. Streamlit Community Docs, “Image & File Upload Handling,” 2023.
7. OpenAI, “Basics of Vision-Language Models,” Online Article, 2022.
8. R. Krishna et al., “Visual Dialog Research,” 2020.
9. P. Bansal, “Overview of Vision Question Answering,” 2022.
10. Google Cloud, “Using Generative Models in Projects,” Developer Blog, 2024.
11. HuggingFace, “Beginner Notes on Image Models,” 2023.
12. S. Chen, “Recent Advancements in Multi-Modal AI,” 2022.
13. YouTube Tutorials, “Working With Streamlit + APIs,” accessed 2024.