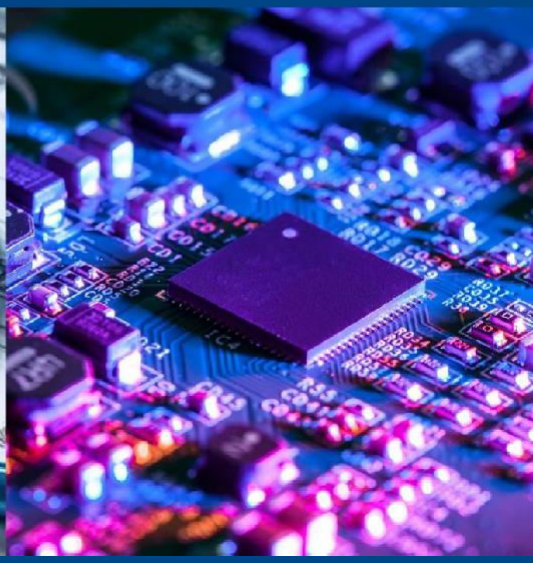
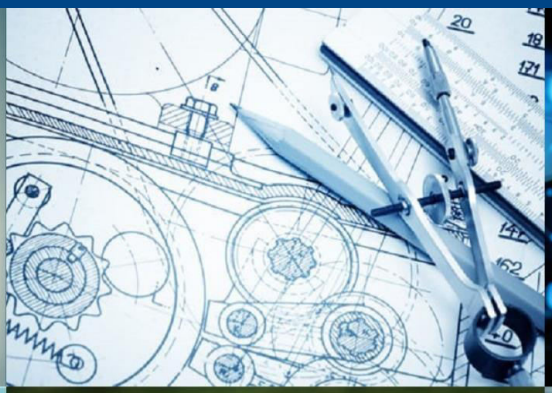
ISSN: 2582-7219

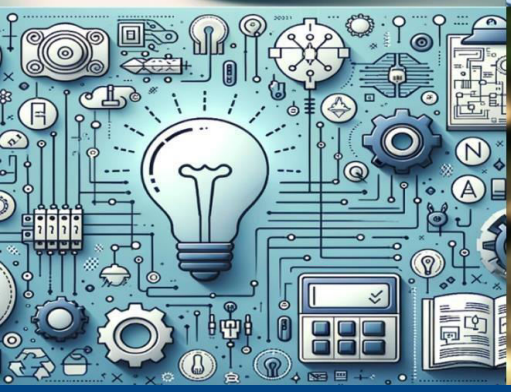




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Issue 12, December 2025

© 2025 IJMRSET | Volume 8, Issue 12, December 2025| DOI:10.15680/IJMRSET.2025.0812040

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

AI Chatbot with Voice and Vision (Agentic GenAI Application)

Archana M P, Dr M. Charles Arockiaraj

Student, Department of MCA, AMC Engineering College, Bengaluru, India

Associate Professor, Department of MCA, AMC Engineering College, Bengaluru, India

ABSTRACT: With the rapid growth of Generative Artificial Intelligence (GenAI), interactive systems are evolving beyond simple text-based communication. Most traditional chatbots are restricted to typed input and output, which limits their ability to interact naturally with users or understand the surrounding environment. To overcome these limitations, this project focuses on the development of an AI Chatbot with Voice and Vision, designed as an Agentic GenAI application that can make independent decisions based on the user's request. Voice interaction is enabled using a speech-to-text module based on the Whisper large-v3 model, which accurately converts spoken input into text. The chatbot's responses are then delivered using Text-to-Speech (gTTS), allowing users to receive replies in a clear and natural audio format. To support vision-based interaction, the system captures real-time images using a webcam, enabling the chatbot to observe and analyse the user's environment. These visual inputs are processed using a multimodal language model, which understands both the image and the user's question to generate appropriate responses. The chatbot is implemented using a Gradio-based web interface, providing support for typed messages, voice input, live webcam feed, and instant responses. The developed system demonstrates practical use in areas such as assistive technology, education, smart human–computer interaction, and accessibility support. Overall, this project illustrates how agentic GenAI systems can move beyond conventional chatbots and function as intelligent assistants that can perceive, understand, and respond effectively in real-world scenarios.

KEYWORDS: Agentic AI; Multimodal Chatbot; Voice Interaction; Computer Vision.

I. INTRODUCTION

1.1 Background and problem context

Over the last few years, artificial intelligence has quietly become part of our everyday routine. What once existed mainly in research papers and science fiction movies is now something we interact with on phones, chatbots on websites, and smart features in learning platforms have become so common that we hardly stop to think about the technology behind them. At the same time, computer vision has made remarkable progress, allowing machines to understand visual information through webcams and cameras and react to what they "see".

In a traditional chatbot, they don't give real-time access, and they also. As compared to traditional chatbots, it doesn't have real-time interaction, In traditional chatbots, they do not use vision, but this one uses voice as well as vision, which is less time-consuming. An AI agent interacts with the environment in real time, such as exploring the environment through a webcam. When we raise a question about the environment, it observes the whole environment and gives a result to the agent clients.

II. LITERATURE REVIEW

Previous research in artificial intelligence has explored various techniques for natural language processing, speech recognition, and visual understanding. Deep learning models such as Convolutional Neural Networks (CNNs) have demonstrated strong performance in text and image classification tasks. Recurrent architectures, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), have been widely used for sequence modeling and language understanding.

Recent advancements in Large Language Models have enabled conversational systems to generate coherent and context-aware responses. However, many existing studies focus on single-modality interaction, primarily text-based

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

communication. Although multimodal systems have been proposed, they often lack agentic control, where decision-making is separated from perception and response generation.

The concept of Agentic AI introduces a reasoning-first approach, where the system evaluates context, selects tools, and then performs actions. This project builds upon these research insights by combining agentic reasoning with multimodal interaction, resulting in a more flexible and intelligent conversationalsystem.

III. EXISTINGSYSTEM

Most existing chatbot systems are designed to handle only text-based communication. These systems typically rely on predefined rules or basic language models and do not support speech input, voice output,orvisualperception. Additionally, traditional chatbots lack real-time interaction with the physical environment. They cannot observe surroundings through cameras or dynamically adapt responses based on visual context. The absence of autonomous decision-making further limits their ability to handle complex or multimodal user queries.

IV. PROPOSEDSYSTEM

The proposed system is an intelligent multimodal AI chatbot capable of interacting through text, voice, and vision. The system is designed as an autonomous AI agent that reasons over user input and dynamically, determine sthere quired processing steps.

Key features of the proposed system include:

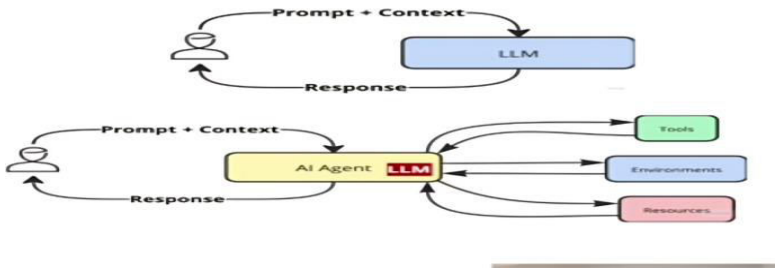
- Support for speech-based interaction using voice input and audio output
- Real-time visual perception using a webcam
- Agentic decision-making to determine tool invocation
- Context-aware response generation using a Large Language Model
- By combining chatbot functionality with agent-based reasoning, the system provides a more natural, adaptive, and human-like interaction experience.

V. SYSTEM ARCHITECTURE

To achieve scalability, reliability, and real-time responsiveness, the architecture is divided into **three independent yet interconnected phases**, each responsible for a specific functional role.

Phase 1: Intelligent Reasoning and Decision Layer
In the first phase, user input is received either as text or spoken language. Voice-based input is transformed into textual form using a speech recognition module, ensuring uniform processing regardless of the input mode. The normalized input is then forwarded to the AI Agent for analysis.

Within this layer, the LLM interprets the user's intent, maintains conversational context, and plans the required action.



Phase 2: Speech Synthesis Layer
Once the AI Agent generates a response, it is forwarded to the speech synthesis layer. Here, textual output is converted into natural and intelligible audio using a text-to-speech engine. This phase enables hands-free interaction and improves accessibility for users.

© 2025 IJMRSET | Volume 8, Issue 12, December 2025| DOI:10.15680/IJMRSET.2025.0812040

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

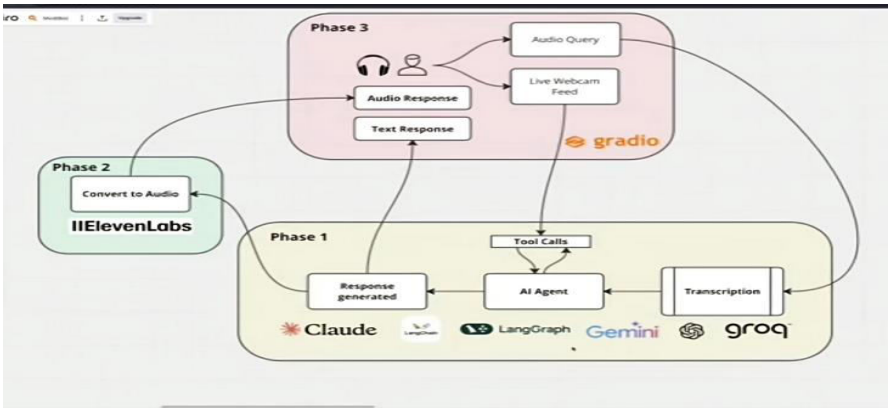
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Phase 3: Multimodal Interaction and User Interface Layer

The final phase handles real-time interaction with the user through a graphical interface. It supports typed input, spoken queries, and live webcam streams. When a vision-related query is detected, the system captures image frames from the webcam and transmits them to the AI Agent for contextual interpretation.

VI. METHODOLOGY



Step 1: User Interaction Initiation

The interaction process begins when the user communicates with the system using one of the supported input modes. These include typed text input, spoken commands captured through a microphone, or visual input obtained from a live webcam stream when a vision-related query is issued.

Step 2: Input Standardization

To ensure consistent processing, all incoming inputs are converted into a unified internal representation. Spoken input is transformed into textual data using a speech recognition mechanism, while visual input is captured as image frames from the webcam. These processed inputs are normalized into a format that can be effectively interpreted by the AI Agent.

Step 3: Agentic Reasoning and Intent Analysis

The standardized input, along with the current conversational context, is passed to the AI Agent. The Large Language Model embedded within the agent analyzes the user's intent and evaluates the nature of the request. Based on this analysis, the agent autonomously determines whether a direct text-based response is sufficient or if additional processing—such as vision analysis or external tool usage—is required. This autonomous decision-making capability is a defining feature of Agentic Generative AI systems.

Step 4: Conditional Tool Execution

When the agent determines that supplementary information is necessary, it selectively activates relevant tools. This may include invoking vision processing modules to analyse captured images or retrieving contextual data from external resources. The outputs generated by these tools are then integrated back into the agent's reasoning process to refine understanding and improve response accuracy.

Step 5: Context-Aware Response Formation

After completing the reasoning and tool integration stages, the AI Agent generates a response that is tailored to the user's query and context. The response may consist of explanations, observations derived from visual input, or actionable instructions, depending on the nature of the interaction.

Step 6: Multimodal Response Delivery

The generated response is presented to the user through appropriate output channels. Textual responses are displayed on the interface, while spoken output is produced using a text-to-speech engine to enable hands-free interaction. In cases involving vision-based tasks, the system may also provide visual feedback to enhance clarity and user understanding.

© 2025 IJMRSET | Volume 8, Issue 12, December 2025| DOI:10.15680/IJMRSET.2025.0812040

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Step 7: Context Preservation and Follow-Up Interaction

The system continuously maintains conversational context throughout the interaction. This allows users to ask follow-up questions without restarting the session, enabling coherent multi-turn conversations and a more natural user experience.

VII. DESIGN AND IMPLEMENTATION

7.1 Architectural Design

From an architectural perspective, the system is structured into **three logical layers**, each responsible for a distinct set of operations:

1. Interaction Layer
2. Agent Reasoning Layer
3. Perception and Tool Layer

These layers communicate through well-defined interfaces, ensuring smooth data flow and maintainability.

Interaction Layer

The interaction layer manages all user-facing communication. Users can engage with the system using typed text, spoken commands, or live webcam input for vision-based queries. The user interface is implemented using a lightweight web framework that supports real-time interaction, including audio recording, video streaming, and dynamic content updates.

Agent Reasoning Layer

The agent reasoning layer constitutes the intellectual core of the system. It consists of a Large Language Model (LLM), a context management module, and decision logic for tool invocation. The AI Agent processes user input, maintains conversational continuity, and evaluates whether external resources—such as vision analysis modules or information retrieval services—are required.

Perception and Tool Layer

The perception and tool layer is responsible for handling sensory input and auxiliary services. It includes speech recognition for converting spoken input into text, text-to-speech synthesis for audio output, and image processing modules for analysing webcam frames.

7.2 Implementation Details

The system is implemented using **Python**, selected for its extensive support for artificial intelligence, machine learning, and real-time application development.

Speech Processing

Voice interaction is facilitated through a bidirectional speech pipeline. Spoken input is captured using microphone interfaces and converted into structured textual data via speech recognition. Generated responses are then transformed back into speech using text-to-speech engines. This process enables smooth and natural voice-based interaction without requiring manual text input.

Vision Processing

Visual input is processed selectively to optimise system performance. Webcam frames are captured only when the user submits a vision-related query. The AI Agent then analyses the visual data in conjunction with the user's question to generate meaningful and context-aware responses. This on-demand processing approach minimises unnecessary computation.

Agent Control Logic

The AI Agent functions as the central controller of the system. Its responsibilities include interpreting user intent, preserving conversational context, selecting appropriate tools, and generating coherent responses. The agent logic follows a structured reasoning pipeline rather than a simple prompt–response mechanism, enabling multi-step analysis and informed decision-making.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VIII. OUTCOME OF RESEARCH

This research successfully establishes the feasibility of developing a **multimodal, agent-driven AI chatbot** capable of interacting through text, speech, and real-time visual input within a unified framework. The primary outcome of the study is the confirmation that **Agentic Generative AI architectures** can be effectively employed in practical applications, extending beyond the limitations of conventional text-based conversational systems. A significant achievement of the project is the implementation of an **AI Agent–centric decision-making mechanism**. Instead of generating responses directly from user prompts, the system evaluates user intent, analyses contextual information, and selectively activates supporting modules such as speech processing and vision analysis. The project further proves that **seamless multimodal interaction** can be achieved without reliance on complex or high-cost infrastructure. The system is capable of processing spoken input, interpreting live webcam data when required, and responding through both textual and synthesised speech outputs.

IX. RESULTS AND DISCUSSION

The proposed agentic multimodal AI chatbot was implemented and evaluated through a series of functional tests and interactive usage scenarios. The results confirm that the system can effectively manage **text-based, voice-based, and vision-based interactions** within a single integrated platform, validating the effectiveness of the proposed architecture. In vision-based interaction scenarios, the system successfully captured live image frames from the webcam and utilised them as contextual input during reasoning. This capability allowed the chatbot to generate meaningful responses to queries related to visual content, highlighting a clear advantage over traditional chatbots that depend solely on static images or textual descriptions.

The experimental observations further indicate that an **agent-driven reasoning model** provides greater flexibility than direct prompt–response approaches. The system's ability to reason prior to tool invocation resulted in more relevant and context-aware responses, reinforcing the suitability of agentic architectures for real-time multimodal AI applications

X. FUTURE WORK

Although the proposed system meets its intended objectives, several opportunities exist for further enhancement. One potential direction involves integrating more advanced vision models capable of recognizing objects, gestures, and facial expressions. Such capabilities would significantly improve environmental awareness and enable richer interaction scenarios.

Another promising extension is the incorporation of **emotion recognition** using speech patterns and facial cues. This would allow the chatbot to respond more empathetically, improving human–AI interaction quality. Enhancing long-term conversational memory is also an important area for future research, enabling the AI Agent to retain user preferences and interaction history across sessions for personalized behavior.

Future work may also focus on deploying the system on **mobile or embedded platforms**, expanding accessibility beyond desktop environments. Optimisation for low latency and partial offline operation would further enhance practical usability.

XI. CONCLUSION

Overall, this project shows how agentic Generative AI can be applied to build intelligent assistants that go beyond conventional chatbot functionality. The developed system can be effectively used in applications such as assistive technologies, educational tools, and smart human–computer interaction systems. The results highlight the potential of multimodal AI systems to provide more meaningful, responsive, and context-aware user interactions in real-world environments.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.
2. Antol, S., et al. (2015). VQA: Visual question answering. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
3. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *ICLR*.
4. Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. *OpenAI Technical Report*.
5. Tan, X., et al. (2021). Neural text-to-speech synthesis. *IEEE Signal Processing Magazine*, 38(1), 16–25.
6. Yao, S., et al. (2023). ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
7. Schick, T., et al. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv preprint*.
8. Li, X., et al. (2024). Agent-based architectures for multimodal AI systems. *ACM Computing Surveys*.
9. Xu, J., et al. (2023). Conversational memory in large language models. *ACL*.
10. West, S. M., Whittaker, M., & Crawford, K. (2024). Discriminating systems: Ethics of AI. *AI & Society*.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com