iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** III **Month of publication:** March 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.78339

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue III Mar 2026- Available at www.ijraset.com*

# Multimodal Memory-Augmented Multi-Agent Conversational AI with Multilingual Support

Durgunala Ranjith[1], Abhinaya Jyothi[2], A. Harsha Vardhan Reddy[3], J. Chandrahas[4]
[1]*Assistant Professor, Department of CSE (AI & ML), ACE Engineering College, Hyderabad, Telangana, India*
[2, 3, 4]*Students, Department of CSE (AI & ML), ACE Engineering College, Hyderabad, Telangana, India*

***Abstract:*** ***Conversational AI technologies have made significant progress in enabling intelligent human–computer interaction; however, many existing systems still struggle with limitations such as weak contextual memory, reliance on a single interaction modality, and limited multilingual communication capabilities. These challenges reduce the ability of conversational systems to sustain meaningful long-term interactions and effectively serve users from diverse linguistic backgrounds. To address these issues, this paper presents a Multimodal Memory-Augmented Multi-Agent Conversational AI with Multilingual Support, designed to enhance conversational intelligence and adaptability. The proposed system integrates multiple input modalities, including text, speech, and images, allowing users to interact with the system in more natural and flexible ways. A contextual memory mechanism is incorporated to retain both short-term and long-term conversational history, enabling the system to generate more coherent, context-aware, and personalized responses during extended interactions. Furthermore, the architecture employs a collaborative multi-agent framework in which specialized agents perform tasks such as language translation, summarization, sentiment analysis, and recommendation generation. By leveraging multimodal processing techniques and multilingual language models, the system supports communication across languages such as English, Hindi, and Telugu. Experimental design and system architecture demonstrate the feasibility of the proposed framework for real-world conversational AI applications.***

***Keywords: Conversational AI, Multimodal AI, Memory-Augmented Systems, Multi-Agent Systems, Large Language Models, Multilingual AI, Dialogue Systems***

## I. INTRODUCTION

Conversational Artificial Intelligence (AI) has emerged as a transformative technology that enables natural interaction between humans and machines through dialogue systems such as chatbots, virtual assistants, and intelligent support platforms. Recent advancements in large language models (LLMs) and deep learning techniques have significantly improved the capabilities of conversational systems, allowing them to generate fluent responses and perform a wide range of language-related tasks. Despite these developments, many existing conversational AI systems still face several important limitations, including restricted contextual memory, reliance on single-modality input, and limited support for multilingual communication. These challenges often reduce the ability of AI systems to maintain coherent long-term conversations and effectively serve users from diverse linguistic backgrounds. Traditional conversational systems typically process each user query independently, with only a limited amount of short-term context retained during interaction.

As a result, they often fail to remember previous conversations, user preferences, or historical information that could improve response relevance. In addition, many systems are designed primarily for text-based interaction, which restricts their ability to process other forms of input such as speech or images. This lack of multimodal capability limits the richness and flexibility of human–AI interaction in real-world scenarios. Furthermore, multilingual communication is frequently handled through external translation services, which can introduce delays, inaccuracies, and loss of contextual meaning during conversations. To address these challenges, this paper proposes a Multimodal Memory-Augmented Multi-Agent Conversational AI with Multilingual Support. The proposed system integrates multiple input modalities, including text, speech, and images, allowing users to interact with the system in a more natural and flexible manner.

A memory-augmented mechanism is incorporated to maintain both short-term and long-term conversational context. This enables the system to retain information about previous interactions, user preferences, frequently discussed topics, and linguistic patterns. By leveraging this contextual memory, the AI system can generate responses that are more coherent, context-aware, and personalized over time. The proposed architecture adopts a multi-agent framework, where multiple specialized agents collaborate to perform different tasks within the conversational pipeline.

Each agent is responsible for a specific function, such as language translation, sentiment or emotion detection, content summarization, and recommendation generation. Through structured communication and information sharing between agents, the system can process complex conversational tasks more efficiently and accurately than traditional single-agent approaches.

This collaborative architecture also improves system scalability, as additional agents can be incorporated to support new functionalities without redesigning the entire system. Another key contribution of the proposed framework is its integrated multilingual capability. Instead of relying entirely on external translation APIs, the system is designed to understand and generate responses in multiple languages directly. This enables smoother multilingual interaction and reduces translation errors, making the conversational experience more natural and accessible for users from different linguistic backgrounds. The system particularly focuses on supporting widely used languages such as English, Hindi, and Telugu, which are commonly used in many real-world applications. By combining multimodal interaction, contextual memory, multi-agent collaboration, and multilingual understanding, the proposed system aims to significantly enhance the intelligence, adaptability, and usability of conversational AI systems. Such capabilities are especially valuable in domains such as education, healthcare, customer service, business analytics, and intelligent personal assistants, where users often require continuous, context-aware, and multilingual communication. Overall, the proposed approach represents an important step toward building more advanced conversational AI systems that are capable of supporting richer human–computer interactions in complex real-world environments.

## II. LITERATURE REVIEW

*1) Zihao Wang; Shaofei Cai; Anji Liu; Yonggang Jin; Jinbing Hou; Bowei Zhang [1]*

This work introduces JARVIS-1, an open-world multi-task agent framework that tightly integrates multimodal perception (vision, language, and actions) with memory-augmented large language models to handle complex, long-horizon tasks. The system adopts a planner-based architecture that enables agents to decompose high-level goals into executable sub-tasks using goal-conditioned control, self-instruction, and lifelong learning mechanisms. By continuously learning from past experiences, JARVIS-1 can adapt to new environments. A major contribution of this work lies in its multimodal memory module, which allows the agent to store, retrieve, and reuse visual, textual, and action-based information across extended interactions. This capability significantly improves long-term reasoning and decision-making, especially in open-ended environments where tasks require multiple steps and contextual awareness. Experimental evaluations conducted in challenging environments such as Minecraft demonstrate that JARVIS-1 achieves higher task completion accuracy, stronger reasoning ability, and better adaptability compared to existing agent-based systems.

Furthermore, the framework shows improved generalization to unseen tasks and environments, highlighting the effectiveness of combining memory augmentation with multimodal reasoning. These findings strongly support the need for memory-enhanced and multimodal architectures in conversational AI systems, as they enable more coherent, context-aware, and intelligent interactions over extended conversations.

*2) He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah [2*

This paper proposes MA-LMM (Memory-Augmented Large Multimodal Model), a framework specifically designed to address the challenges of long-term video understanding, where conventional models struggle due to limited context windows and high computational cost. MA-LMM processes long video streams sequentially, maintaining a memory bank that stores important visual features and learned queries from previous segments. This design enables the model to retain crucial information over extended time periods without needing to process the entire video at once.

•The architecture combines a frozen visual encoder for efficient feature extraction, a Q-Former to align visual representations with language queries, and a large language model (LLM) for reasoning and response generation. By selectively updating and retrieving information from the memory bank, MA-LMM effectively captures long-range temporal dependencies while keeping memory and computation costs bounded. This balance between performance and efficiency is a key strength of the proposed approach.

•Experimental evaluations show that MA-LMM achieves state-of-the-art performance on long-video question answering and video captioning benchmarks, outperforming existing multimodal models that lack explicit memory mechanisms. The results clearly demonstrate that memory augmentation plays a critical role in long-term multimodal understanding. These findings strongly reinforce the importance of incorporating long-term memory into conversational AI systems, particularly for applications that require sustained context, such as extended dialogues, video-based assistance, and multimodal human–AI interaction.

*3) Dae Hong Kim, Hyeongseok Ko [3]*

This study introduces a nunchi-aware multi-agent chatbot system, drawing inspiration from *nunchi*, a concept that represents social awareness and sensitivity in human conversations. The proposed system leverages Multi-Agent Large Language Models (MALLM), where multiple agents with distinct expert personas collaboratively generate responses. These agents engage in structured discussions, exchange viewpoints, and apply consensus-building mechanisms to arrive at socially appropriate and contextually relevant replies. By modeling conversational factors such as context, tone, implicit intent, and social cues, the system demonstrates improved ability to produce responses that are more natural, empathetic, and situation-aware compared to single-agent chatbots. Experimental evaluations indicate that multi-agent collaboration leads to greater reasoning diversity, richer dialogue content, and better performance on complex conversational tasks. However, the study also identifies challenges, including agent drift, increased computational overhead, and reduced efficiency in simpler tasks.

Overall, this work highlights the effectiveness of multi-agent collaboration in enhancing contextual and social understanding, reinforcing its importance for building advanced conversational AI systems that aim to interact in a more human-like and socially intelligent manner.

*4) Reem Gody, Mahmoud Goudy, Ahmed Y. Tawfik [4]*

ConvoGen introduces a multi-agent framework designed to generate high-quality synthetic conversational data for improving conversational AI systems. The framework employs multiple GPT-4–powered agents, each assigned controlled roles and distinct personas, to simulate realistic group conversations. Through iterative sampling and role-based interaction, the system encourages diverse viewpoints and interaction patterns, resulting in more natural and varied dialogue generation.

A key strength of ConvoGen lies in its ability to enhance lexical diversity, contextual realism, and behavioral controllability of generated conversations. By carefully coordinating agent behavior, the framework avoids repetitive or generic responses that are commonly observed in single-agent dialogue generation. Experimental evaluations demonstrate that the synthetic data produced by ConvoGen is highly effective for training, fine-tuning, and evaluating conversational AI models, leading to improved performance and robustness.

*5) Jiabao Fang, Shen Gao, Shen Gao [5]*

This paper proposes MACRS (Multi-Agent Conversational Recommender System), a cooperative framework designed to improve recommendation quality through task-oriented multi-agent collaboration. The architecture consists of multiple specialized agents, each responsible for a distinct function, such as question-asking agents to elicit user preferences, recommendation agents to suggest relevant items, chit-chat agents to maintain natural conversational flow, and dialogue planning agents to manage interaction strategy and coherence. A key contribution of MACRS is its feedback-aware reflection mechanism, which enables the system to continuously refine its responses by analyzing user feedback and interaction history. This mechanism allows the agents to adapt dynamically to evolving user interests, leading to more personalized and context-aware recommendations. Experimental evaluations demonstrate that MACRS achieves higher recommendation accuracy, improved dialogue coherence, and increased user satisfaction compared to baseline single-agent and LLM-based recommendation systems.

*6) Jonas Becker [6]*

This work provides a detailed analysis of the strengths and limitations of multi-agent large language model (LLM) frameworks for conversational task solving. The study demonstrates that multi-agent systems are particularly effective in handling complex reasoning, collaborative decision-making, and multi-step problem solving, as different agents can contribute diverse perspectives and specialized expertise. This collaborative approach often leads to more accurate and well-reasoned outcomes compared to single-agent systems.

However, the paper also identifies several critical challenges associated with multi-agent LLM frameworks. These include problem drift, where agents gradually deviate from the original task objective; alignment collapse, in which agents converge on suboptimal or biased solutions; and discussion monopolization, where dominant agents overshadow others, reducing reasoning diversity. To address these issues, the authors emphasize the importance of careful agent role definition, interaction protocols, and coordination mechanisms. The insights from this study are highly valuable for the design of stable, efficient, and scalable multi-agent conversational AI architectures, guiding the development of systems that balance collaborative intelligence with controlled interaction and robustness

Comparison Table

| S. No | Author(s) | Title | Proposed Methodology | Findings from the Reference |
|---|---|---|---|---|
| 1 | Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang | JARVIS-1: Open-World Multi-task Agents with Memory-Augmented Multimodal Language Models | JARVIS-1 uses a multimodal LLM-based planner with memory-augmented, goal-conditioned control, self-instruction, and lifelong learning for task execution in Minecraft. | Multimodal memory and reasoning-based retrieval enable JARVIS-1 to achieve near-perfect short-to-intermediate task performance and 5× better success on long-horizon tasks compared to prior agents. |
| 2 | He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah | MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding | A MA-LMM processes long videos sequentially using a memory bank of past visual features and learned queries combined with a frozen visual encoder, Q-Former, and LLM. | Memory-augmented sequential processing enables strong performance in long-video understanding, video QA, and captioning while efficiently handling long-term dependencies. |
| 3 | Dae Hong Kim, Hyeongseok Ko | Nunchi-Aware Multi-Agent Chatbots for Socially Contextual Response Generat | Multi-Agent Conversational Systems (MALLM) organize multiple LLM-based agents with expert personas that interact through discussion paradigms and consensus | Multi-agent setups improve reasoning, diversity, and performance on complex tasks but may introduce inefficiency or reduced gains for simpler. |
| 4 | Reem Gody, Mahmoud Goudy, Ahmed Y. Tawfik | ConvoGen: Enhancing Conversational AI with Synthetic Data – A Multi-Agent Approach | ConvoGen uses a multi-agent framework with GPT-4-powered experience generation and iterative sampling to create diverse persona-driven group conversations. | The approach improves lexical diversity, produces realistic dialogues, controls agent behavior, and generates useful synthetic data for training conversational AI models. |
| 5 | Jiabao Fang, Shen Gao | A Multi-Agent Conversational Recommender System | MACRS uses cooperative agents such as Asking, Recommending, Chit-chatting, and Planner agents with a feedback-aware reflection mechanism. | Experimental results show MACRS outperforms baseline conversational recommender systems in recommendation accuracy and user preference collection. |
| 6 | Jonas Becker | Multi-Agent Large Language Models for Conversational Task-Solving | MALLM uses a multi-agent framework with defined agent roles, structured discussions, and decision-making mechanisms to solve conversational tasks. | Multi-agent LLMs perform well in complex reasoning but face issues like problem drift, alignment collapse, and discussion monopolization. |

Table 1: Review of Existing Research on Multi-Agent Conversational AI

## III. SYSTEM ARCHITECTURE

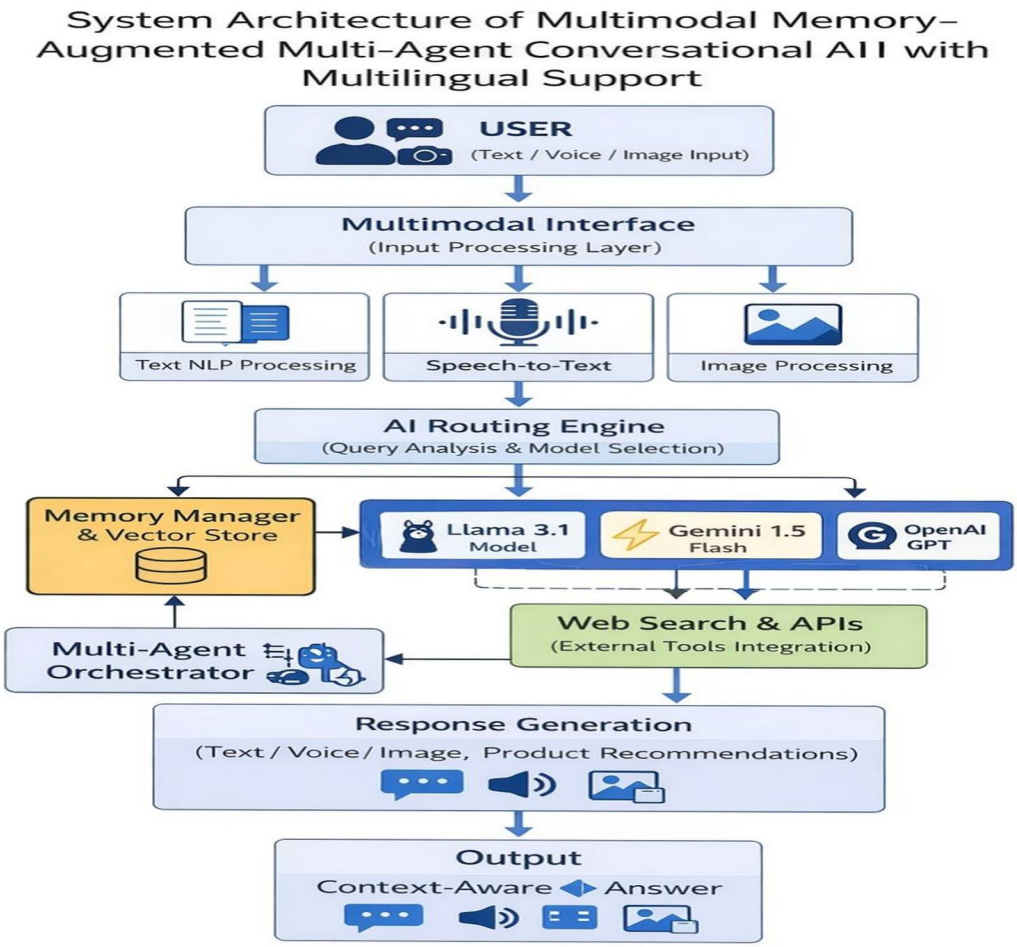



Figure 1: Proposed Multimodal Memory-Augmented Multi-Agent Conversational AI Architecture

The proposed system architecture is designed to support intelligent multimodal conversational interaction by integrating multiple AI models, contextual memory mechanisms, and external knowledge sources. The architecture enables users to interact with the system through various input modalities while ensuring efficient processing, reasoning, and response generation.
As illustrated in Figure 1, the system consists of several interconnected modules that work together to process user queries and generate context-aware responses.

### *A. User Input Layer*

The interaction begins at the user interface, where users provide input in the form of text, voice, or images. This multimodal interaction capability improves accessibility and enables users to communicate with the system using natural forms of input.

### *B. Multimodal Processing Layer*

The input data is processed through the multimodal interface, which acts as the input processing layer. This layer contains specialized modules responsible for handling different input types:

*1)* Text NLP Processing: Processes textual inputs using natural language processing techniques to extract semantic meaning from user queries.
*2)* Speech-to-Text Processing: Converts voice inputs into textual format using speech recognition models, allowing the system to process spoken queries.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue III Mar 2026- Available at www.ijraset.com*

*3)* Image Processing: Analyzes image inputs using computer vision techniques to extract relevant visual information.

The processed input is then forwarded to the AI routing system for further analysis.

AI Model Routing Engine

The AI routing engine is responsible for analyzing the user query and determining the most appropriate AI model for generating a response. The routing mechanism evaluates factors such as:

Query complexity

Task type (general question, coding, reasoning)

Response speed

Cost efficiency

Based on this analysis, the routing engine dynamically selects one of the integrated AI models.

AI Model Layer

The system integrates multiple large language models to provide flexible and optimized responses:

Meta Llama 3.1 Model:

Used for general reasoning tasks and open-source model support.

Google Gemini 1.5 Flash:

Provides fast and cost-efficient responses for lightweight queries.

OpenAI GPT Models:

Used for advanced reasoning, coding assistance, and complex conversational tasks.

Using multiple AI models allows the system to balance performance, accuracy, and computational efficiency.

Memory Management and Knowledge Storage

The memory manager and vector store maintain conversation history, user preferences, and contextual information. This module allows the system to retain knowledge from previous interactions, enabling more personalized and coherent conversations.

The memory system stores relevant information as vector embeddings, which enables efficient retrieval of context during future interactions.

### *C. Multi-Agent Orchestration Layer*

The multi-agent orchestrator coordinates specialized agents responsible for performing specific tasks such as:

Speech recognition

Text generation

Language translation

Task coordination

This multi-agent architecture allows the system to handle complex tasks more efficiently by distributing responsibilities among specialized components.

External Tools and Web Search Integration

To provide real-time information, the system integrates web search and external APIs. When the user query requires updated information such as product prices, news updates, or location details, the system retrieves relevant data from external sources.

This integration ensures that the system can provide accurate and up-to-date responses beyond the knowledge stored within the language models.

Response Generation Module

The response generation module combines outputs from the selected AI model, memory module, and external data sources to produce a final context-aware response. The system ensures that responses remain coherent with previous conversations and user preferences.

### *D. Output Layer*

Finally, the generated response is delivered to the user through the output interface, which may include:

Text responses

Voice responses

Image-based responses

Product recommendations This multimodal output capability enhances the overall user experience and allows the system to support a wide range of real-world applications.

## IV. RESEARCH GAPS IN EXISTING SYSTEMS

### A. Limitations in Memory and Context Handling

Although modern conversational AI systems are highly advanced, many of them suffer from weak memory management. Most systems can only remember information within a single conversation session. Once the session ends, the context and user-specific data are lost. This results in repeated questions, lack of personalization, and disconnected conversations.

Furthermore, existing systems mainly rely on short-term context windows instead of structured long-term memory storage. Without persistent memory, the system cannot adapt to user preferences, past interactions, or behavioral patterns. This gap reduces the effectiveness of conversational AI in real-world applications such as education, customer support, and personal assistance.

## V. BACKGROUND AND FUNDAMENTALS

### A. Background of Conversational Artificial Intelligence

Conversational Artificial Intelligence (AI) has undergone significant evolution over the past decades. Early systems such as rule-based chatbots relied on predefined patterns and keyword matching techniques, limiting their ability to handle complex or dynamic conversations. With the introduction of machine learning and deep learning techniques, conversational systems became more capable of understanding context, generating human-like responses, and performing task-oriented dialogues. Transformer-based models further improved natural language processing, enabling more accurate and fluent interactions. Despite these advancements, many existing conversational AI systems still suffer from limitations such as short-term memory constraints, lack of personalization, and limited adaptability across domains. These challenges highlight the need for more advanced architectures that integrate memory, multimodality, and intelligent multi-agent collaboration.

### B. Fundamentals of Multimodal, Memory-Augmented, and Multi-Agent Systems:

Modern conversational AI systems are increasingly built upon three fundamental concepts: multimodality, memory augmentation, and multi-agent collaboration.

Multimodal systems process and integrate multiple forms of input such as text, speech, and images. This enables richer and more natural human–computer interaction compared to single-mode systems. Memory augmentation enhances AI performance by incorporating short-term and long-term memory mechanisms. Short-term memory maintains conversational context within a session, while long-term memory stores user preferences, history, and interaction patterns for personalization.

Multi-agent systems involve multiple specialized agents working collaboratively to solve complex tasks. Each agent performs a specific role—such as translation, summarization, sentiment analysis, or recommendation—and communicates with other agents to generate coherent and efficient responses. This structured collaboration improves reasoning diversity, adaptability, and scalability of conversational AI architectures.

### C. Need for an Advanced Conversational AI System:

Conversational Artificial Intelligence has evolved from simple rule-based chatbots to advanced intelligent systems capable of understanding natural language. However, many existing conversational AI systems still face limitations such as poor memory retention, single-mode interaction, and limited multilingual support. These issues reduce the effectiveness and naturalness of human–computer interaction.

The proposed system introduces a multimodal conversational AI that supports text, speech, and image inputs, enabling richer and more natural communication. By incorporating memory augmentation, the system can retain both short-term conversational context and long-term user history, allowing personalized and context-aware responses.

Additionally, the system employs a multi-agent architecture, where specialized agents work collaboratively to perform tasks such as translation, summarization, sentiment analysis, and recommendation. This improves efficiency and intelligence in handling complex user requests. With multilingual support for languages such as English, Hindi, and Telugu, the system ensures inclusive and seamless communication across diverse users. Overall, the proposed approach aims to build a smarter, adaptive, and globally accessible conversational AI system.

## VI. METHODOLOGY

### A. System Architecture Design

The proposed system follows a multimodal memory-augmented multi-agent architecture. The design consists of four main layers: Input Layer, Processing Layer, Memory Layer, and Output Layer.

The Input Layer accepts user inputs in different forms such as text, speech, and images. Speech input is converted into text using speech recognition, and images are processed using visual encoders.
The Processing Layer includes multiple specialized agents such as translation agent, summarization agent, sentiment analysis agent, and recommendation agent. These agents collaborate to handle complex user requests.
The Memory Layer manages both short-term memory (to maintain current conversation context) and long-term memory (to store user history and preferences).
The Output Layer generates responses in text or speech format based on user preference.

### *B. Multimodal Processing Mechanism*

The system supports three primary modes of interaction:

- Text Processing: Natural Language Processing (NLP) techniques are used for understanding and generating responses.
- Speech Processing: Speech-to-Text (STT) converts voice input into text, and Text-to-Speech (TTS) converts responses back into voice output.
- Image Processing: A visual encoder extracts meaningful features from images, which are then interpreted by the language model.

All modalities are integrated into a unified representation so that the system can understand context across different input types. This improves interaction quality and makes communication more natural.

### *C. Memory Augmentation Strategy*

Memory augmentation is implemented in two levels:

Short-Term:
Stores current session context, recent messages, and conversation flow to maintain coherence.
Long-Term:
Stores user preferences, frequently asked queries, language choice, and interaction history.

The memory module retrieves relevant past information whenever required, enabling personalized and context-aware responses. This improves long-term engagement and system adaptability.

### *D. Multi-Agent Collaboration Framework*

The system uses a collaborative multi-agent approach where each agent performs a specific role:

- Translation Agent – Handles multilingual input/output
- Summarization Agent – Provides concise responses
- Sentiment Analysis Agent – Detects user emotions
- Recommendation Agent – Suggests relevant content
- Planning Agent – Coordinates task flow among agents

Agents communicate through a structured coordination mechanism. The planning agent assigns tasks and integrates outputs from different agents into a final response. This approach improves reasoning ability, reduces task overload on a single model, and enhances overall system efficiency.

## VII. CHALLENGES AND LIMITATIONS

### *A. Technical Complexity and Integration Challenges*

The proposed Multimodal emory-Augmented Multi-Agent Conversational AI system involves integrating multiple advanced technologies such as natural language processing, speech recognition, image processing, memory modules, and multi-agent coordination. Combining all these components into a unified architecture increases the complexity of the system. Ensuring smooth communication between agents managing data flow between modules, and maintaining system stability require careful design and testing. Any failure in one module (e.g., speech recognition errors) can affect overall performance. Therefore, system integration and debugging become major technical challenges.

### *B. Memory Management and Scalability Issues*

While long-term memory improves personalization, storing and managing large amounts of user data can create scalability problems. As the number of users increases, memory storage requirements also grow significantly.

Efficient indexing, retrieval mechanisms, and memory optimization strategies are necessary to prevent delays. Poor memory handling may lead to slower response times or irrelevant information retrieval. Maintaining balance between memory size and performance is a key limitation.

### *C. Multi-Agent Coordination and Stability*

In a multi-agent framework, different agents collaborate to complete tasks. However, coordination among agents can sometimes lead to issues such as:

- Task overlap
- Conflicting outputs
- Increased processing time
- Communication delays

If roles are not clearly defined, agents may produce inconsistent responses. Managing synchronization and ensuring stable collaboration is a major challenge in multi-agent conversational AI systems.

### *D. Multilingual Accuracy and Ethical Concerns*

Although the system supports multilingual communication, ensuring high accuracy across multiple languages is challenging. Variations in grammar, cultural context, idioms, and dialects may affect understanding and response quality.

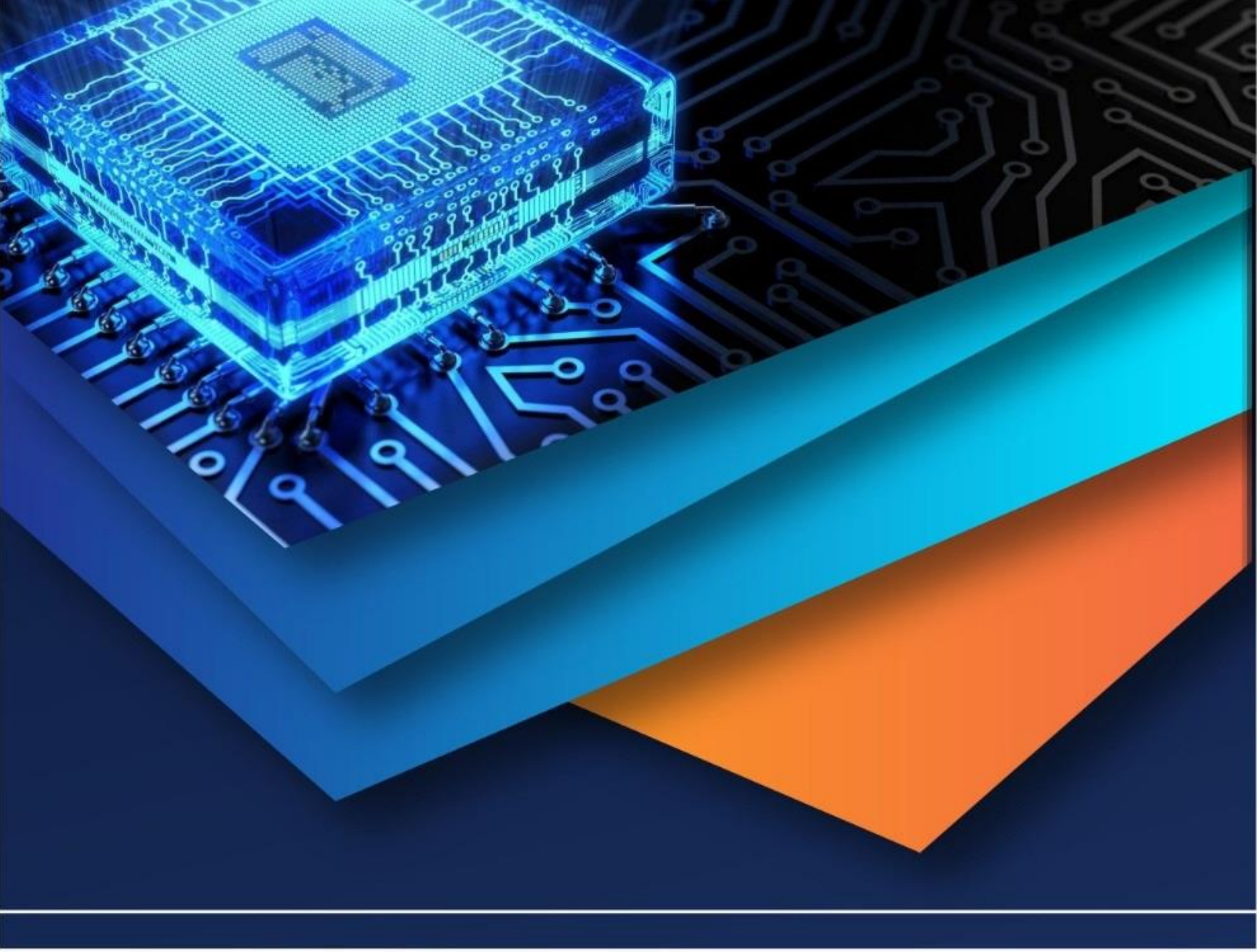
Additionally, ethical concerns such as data privacy, user consent, bias in AI responses, and misuse of stored personal information must be carefully addressed. Strong data protection mechanisms and responsible AI guidelines are required to maintain user trust and system reliability.

## VIII. CONCLUSION AND FUTURE SCOPE

The Conversational Artificial Intelligence has significantly evolved from simple rule-based chatbots to advanced large language model-based systems. However, existing systems still face limitations such as weak long-term memory, single-mode interaction, lack of structured multi-agent collaboration, and limited direct multilingual support. To address these challenges, this project proposed a Multimodal Memory-Augmented Multi-Agent Conversational AI with Multilingual Support. The system integrates text, speech, and image processing capabilities to enable natural and flexible human–computer interaction. By incorporating both short-term and long-term memory modules, the system enhances contextual understanding and personalization. Furthermore, the multi-agent architecture improves reasoning efficiency by assigning specialized tasks to different agents such as translation, summarization, sentiment analysis, and recommendation. Direct multilingual capability ensures inclusive communication across diverse users. Overall, the proposed system aims to build a smarter, adaptive, scalable, and user-centric conversational AI platform capable of handling complex real-world interactions.

## REFERENCES

[1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, Y. Narang, and Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models,” arXiv preprint arXiv:2210.03629, 2023.
[2] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli,L. Zettlemoyer, N. Cancedda, and T. Scialom, “Toolformer: Language Models Can Teach Themselves to Use Tools,”arXiv preprint arXiv:2302.04761, 2023.
[3] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.
[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin,N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020.
[5] N. Shinn, F. Labash, and A. Gopinath, “Reflexion: Language Agents with Verbal Reinforcement Learning,” arXiv preprint arXiv:2303.11366, 2023.
[6] J. Park, J. O’Brien, C. Cai, M. R. Morris, P. Liang, and M. Bernstein, “Generative Agents: Interactive Simulacra of Human Behavior,” Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2023.
[7] Z. Wang, S. Cai, A. Liu, Y. Jin, J. Hou, and B. Zhang, “JARVIS-1: Open-World Multi-Task Agents with Memory-Augmented Multimodal Language Models,” arXiv preprint arXiv:2311.05997, 2024.
[8] H. Li, Y. K. Jang, M. Jia, X. Cao, and A. Shah,“Memory-Augmented Large Multimodal Model for Long-Term Video Understanding,” arXiv preprint arXiv:2401.05645, 2024.
[9] [9] J. Becker, “Multi-Agent Large Language Models for Conversational Task Solving,” arXiv preprint arXiv:2305.15055, 2023.
[10] J. Fang, S. Gao, and S. Gao, “MACRS: A Multi-Agent Conversational Recommender System,” Proceedings of the ACM Web Conference (WWW), 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY