International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4529 - 4531
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

Multimodal Artificial Intelligence Systems: A Review of Retrieval-Augmented Generation, Voice Processing, and Document Intelligence

Sohan Prakash Shinde[1], Mahadev Bhagavat Waghmode[2], Dipak Tukaram Chikane[3], Aditya Tukaram Kumbhar[4], Ashwin Dipak Patil[5]
[1,2,3,4,5]Department of Computer Science, Yashoda technical Campus, Faculty of Engineering, Satara, Maharashtra, India -415015
Emails:*sohanshinde005@gmail.com[1],MahadevWaghmode2005@gmail.com[2], dipakchikane21@gmail.com[3], kumbharaditya2035@gmail.com[4], ashwinpatil19@gmail.com[5]*

Abstract
Artificial Intelligence has evolved rapidly, leading to the development of systems that can process information from multiple sources such as text, speech, images, and documents. Recent advancements in Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Speech-to-Text (STT), and Text-to-Speech (TTS) have improved the capabilities of intelligent assistants and information retrieval systems. This review paper presents an overview of multimodal AI systems and examines the technologies that enable efficient document understanding, voice interaction, and automated content generation. Various research studies related to retrieval techniques, language models, speech processing, and document intelligence are analyzed to understand their contributions and limitations. The paper also discusses the applications of these systems in education, research, and professional environments. Finally, current challenges and future opportunities in the development of multimodal AI assistants are highlighted. The review shows that integrating multiple AI technologies into a unified framework can improve accessibility, productivity, and user experience across different domains.
Keywords: Artificial Intelligence; Multimodal Systems; Natural Language Processing; Retrieval-Augmented Generation; Speech Processing

1. Introduction
Artificial Intelligence (AI) has become a key technology in modern computing, enabling machines to understand, process, and generate information in ways that closely resemble human intelligence. The rapid growth of digital data has increased the demand for intelligent systems capable of handling multiple forms of information, including text, speech, images, and documents. Traditional AI applications are often designed for specific tasks and require users to rely on multiple tools to perform information retrieval, document analysis, and content generation.[1] Recent advancements in Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Speech-to-Text (STT), and Text-to-Speech (TTS) have contributed to the development of multimodal AI systems. These technologies enable intelligent assistants to retrieve relevant information, summarize documents, answer questions, process voice commands, and generate meaningful responses with improved accuracy.[2] The integration of these capabilities into a single framework enhances usability, accessibility, and productivity across academic, research, and professional environments. Researchers have demonstrated that combining retrieval mechanisms with advanced language models improves contextual understanding and reduces the generation of inaccurate information.[1][2] As a result, multimodal AI systems are increasingly being adopted for document intelligence, virtual assistance, knowledge management, and automated content creation. This review paper examines the major technologies, applications, challenges, and future directions of multimodal artificial intelligence systems, highlighting their growing importance in modern information processing.

1.1.Literature Review
- Reimers and Gurevych [1] proposed Sentence-BERT, a model that generates meaningful sentence embeddings for

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4529 - 4531
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

semantic similarity and information retrieval tasks. Their approach significantly improved retrieval performance and reduced computational complexity, making it suitable for intelligent document search systems.

- Lewis et al. [2] introduced BART, a transformer-based language model designed for text generation and summarization. The study demonstrated that BART can produce coherent and context-aware summaries, improving the quality of information extraction and content understanding.
- Johnson et al. [3] developed FAISS, a high-performance similarity search library for large-scale vector retrieval. Their research showed that efficient vector indexing techniques can improve retrieval speed and accuracy in document-based applications.
- Rahman et al. [4] presented a Retrieval-Augmented Generation (RAG) framework that combines document retrieval with language generation. The proposed approach improved response relevance and factual accuracy by utilizing retrieved contextual information before generating answers.
- These studies demonstrate that the integration of semantic retrieval, language generation, and efficient indexing techniques plays a crucial role in the development of modern multimodal AI systems.

1.2.Key Technologies

- The growth of multimodal artificial intelligence has been driven by several important technologies. Retrieval-Augmented Generation (RAG) improves the quality of responses by combining information retrieval with text generation. This helps AI systems provide answers that are more relevant and based on available information rather than relying only on pre-trained knowledge.
- Natural Language Processing (NLP) enables machines to understand and generate human language, supporting tasks such as summarization, question answering, and content analysis. Speech-to-Text (STT) allows spoken input to be converted into text, while Text-to-Speech (TTS) converts text into audio responses, making interaction more natural and accessible.
- In addition, embedding models and vector search techniques help systems locate relevant information quickly from large collections of documents. Together, these technologies form the foundation of modern multimodal AI systems and support a wide range of applications in education, research, and professional environments.

2. Method

This review paper follows a systematic approach to analyze recent developments in multimodal artificial intelligence systems. Relevant research articles, conference papers, and scholarly publications related to Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Speech-to-Text (STT), Text-to-Speech (TTS), and document intelligence were selected for review. The literature was collected from reliable academic sources and focused on studies published in recent years. The selected papers were examined based on their objectives, methodologies, technologies used, applications, and performance outcomes. Particular attention was given to research that integrated multiple modalities such as text, speech, images, and documents within a single intelligent system. The reviewed studies were compared to identify common approaches, recent advancements, strengths, and existing limitations. The collected information was then organized into key technological areas, including information retrieval, language processing, speech technologies, and document analysis. Based on this analysis, current trends, challenges, and future research opportunities in multimodal AI systems were identified and discussed. This methodology provides a comprehensive understanding of the evolution and practical applications of modern multimodal artificial intelligence systems.

3. Results And Discussion

3.1.Results

The review of recent studies shows that multimodal AI systems have significantly improved information retrieval and content processing. Technologies such as Retrieval-Augmented Generation (RAG) and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4529 - 4531
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

Natural[5] Language Processing (NLP) provide more accurate and context-aware responses. Speech technologies, including Speech-to-Text (STT) and Text-to-Speech (TTS), have further enhanced user interaction by enabling voice-based communication. The integration of text, speech, images, and documents within a single framework has improved system efficiency and expanded the range of real-world applications[6].

3.2.Discussion

The reviewed studies indicate that multimodal AI systems offer several advantages, including improved accessibility, faster information retrieval, and better user experience. By combining multiple technologies within a unified platform, these systems can efficiently process different forms of data and provide meaningful outputs. However, challenges such as computational requirements, privacy concerns, and response reliability still need to be addressed. Future research should focus on developing more efficient, secure, and scalable multimodal AI solutions for diverse applications.

Conclusion

Multimodal artificial intelligence systems have transformed the way information is processed and accessed by integrating text, speech, images, and documents within a unified framework. This review highlighted the role of technologies such as Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Speech-to-Text (STT), and Text-to-Speech (TTS) in improving information retrieval and user interaction. The reviewed studies demonstrate that multimodal AI systems can enhance productivity, accessibility, and decision-making across various domains. Although challenges related to privacy, computational resources, and system reliability remain, ongoing research and technological advancements are expected to further improve the effectiveness and adoption of these intelligent systems.

Acknowledgements

The authors would like to express their sincere gratitude to Ms. Shital Waghmare for her valuable guidance, support, and encouragement throughout this work. The authors also thank the Department of Computer Science and Engineering, Yashoda Technical Campus, Faculty of Engineering, Satara, for providing the necessary facilities and academic environment for the successful completion of this study.

References

[1]. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992.

[2]. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of ACL 2020, pp. 7871–7880.

[3]. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-Scale Similarity Search with FAISS. IEEE Transactions on Big Data, 7(3), 535–547.

[4]. Rahman, A., Singh, K., & Gupta, R. (2023). Integrating RAG Models for Contextual Document Retrieval. International Journal of Artificial Intelligence Research, 12(2), 45–53.

[5]. Li, S., Chen, Y., & Wang, H. (2021). Speech Recognition Technologies: From Traditional Models to Deep Learning. Journal of Intelligent Systems, 30(4), 512–524.

[6]. Zhao, R., & Zhang, W. (2022). Text-to-Speech Synthesis for Human–Computer Interaction. International Journal of Speech Technology, 25(3), 301–312.