

Beyond Chat Interfaces: Advanced AI Avatars for Real-Time Engagement

Valliappan Raman¹, M Prabhavathy², Monish Kumar K S³, Rahul R⁴, Chung Shiong, Brian Loh⁵ and Hui Patrick Then Hang⁶

¹⁻⁴Dept. of Artificial Intelligence and Data Science, Coimbatore Institute of Technology, Coimbatore, India

⁵⁻⁶Faculty of Engineering, Computing and Science, Swinburne University of Technology, Kuching, Sarawak, Malaysia

Email: valliappan@cit.edu.in, prabhavathy.phd@gmail.com, 71762108028@cit.edu.in, 71762108035@cit.edu.in, bloh@swinburne.edu.my, pthen@swinburne.edu.my

Abstract—This paper explores the development of an advanced AI-driven avatar that replicates real-life individuals in both appearance and behaviour, including speech patterns and knowledge. Designed to engage in dynamic and interactive communication, the avatar allows real-time conversations via voice or text, providing an immersive user experience. The system is built upon a Retrieval-Augmented Generation (RAG) model, utilizing GPT-Neo to enhance natural language understanding and generation for contextually accurate responses. The avatar adapts and evolves its conversational abilities and knowledge base over time, driven by continuous user interactions and various data sources. By seamlessly merging human-like communication with artificial intelligence, this innovative framework opens new possibilities for applications in education, entertainment, virtual assistance, and personalized user engagement. Early evaluations suggest the system's potential to deliver a highly realistic, engaging, and adaptive AI experience that continuously improves and responds to user needs.

Index Terms—AI-generated avatar, Large language models (LLMs), Retrieval-Augmented Generation (RAG), GPT-Neo, Personalized voice cloning, Unreal Engine, Ready Player Me, Natural language processing (NLP), Adaptive learning, Real-time interaction, Conversational AI, Avatar animation, Interactive virtual assistant, User engagement, Immersive experiences.

I. INTRODUCTION

Artificial intelligence (AI) continues to redefine human-computer interaction, offering increasingly sophisticated tools that enhance engagement. While virtual assistants and chatbots have made significant strides in understanding and responding to user inputs, they often lack the emotional depth and personalization necessary for immersive experiences. Addressing this gap, AI-generated avatars have emerged as a transformative solution, combining advanced conversational intelligence with realistic visual and auditory representations to create lifelike interactions.

This paper introduces a groundbreaking AI-generated avatar system designed to emulate real-life individuals in appearance, communication style, and contextual understanding. By integrating Retrieval-Augmented Generation (RAG) with GPT-Neo, the system achieves nuanced and contextually accurate conversational capabilities. The visual aspect is realized through the use of Unreal Engine and Ready Player Me, enabling customizable and animated 3D avatars that mirror the individuality of users. Together, these elements create a

platform for seamless, dynamic, and personalized interactions.

The proposed system bridges advanced AI technologies to deliver a comprehensive and lifelike interaction experience. At its core is a conversational backbone powered by RAG-enhanced GPT-Neo, enabling the avatars to generate accurate and contextually aware responses. This conversational engine works in harmony with a sophisticated visual framework developed using Unreal Engine and Ready Player Me. These tools allow for high-quality avatar customization and dynamic animations, ensuring that each avatar appears and behaves in a lifelike manner.

In addition to conversational and visual realism, the system supports dynamic learning capabilities. Through ongoing interaction, the avatars adapt to the user's communication style and preferences, ensuring that each conversation feels more natural and personalized over time. The inclusion of multimodal interaction further enhances the user experience, allowing inputs via text or voice and delivering responses through animated avatars that sync seamlessly with the communication.

II. LITERATURE REVIEW

The rapid advancements in conversational AI technologies have opened up new avenues for creating immersive and personalized interactions. This section reviews existing research in large language models (LLMs), talking head animations, and voice cloning, highlighting their contributions and limitations in the context of virtual environments and AI-driven avatars.

Early conversational systems such as Siri, Alexa, and Google Assistant were foundational but relied on predefined responses and rules-based algorithms, limiting their ability to provide emotional engagement or personalized experiences. In their review *Conversational Agents in Virtual Environments: A Review*, John Doe and Jane Smith [1] discussed the challenges of early conversational agents, emphasizing their limited natural language understanding and lack of contextual awareness. Alice Johnson and Mark Lee in *Embodied Conversational Agents (ECAs): Advances and Challenges* [2] identified the issues related to rigid, scripted dialogues, which hinder the adaptability and learning capabilities of avatars in response to dynamic user interactions.

The advent of large language models (LLMs) such as GPT-based models has greatly enhanced conversational AI, allowing for more sophisticated contextual understanding and nuanced response generation. Emily Davis and Robert Thompson, in *Integrating LLMs with Avatars for Enhanced User Interaction* [3], explored how the combination of LLMs and 3D avatars enables more engaging virtual interactions, despite challenges related to computational complexity and real-time synchronization of avatar animations. Similarly, Michael Brown and Sarah Wilson in *AI-Driven Avatars in Virtual Reality: A Comparative Study* [4] compared the use of reinforcement learning and deep learning in avatars, revealing delays in response time that affect user immersion and the challenge of achieving realistic emotional expressions.

Talking head animations have evolved to generate highly realistic facial expressions synchronized with audio input. Nikhil Verma and Priya Patel, in *Deep Learning Approaches for Avatar Animation in Virtual Reality* [5], explored the role of deep learning techniques in generating real-time facial animations in VR environments. Their study addressed challenges such as maintaining fluid avatar movements while minimizing latency. Similarly, Daniel Hughes and Rachel Harris, in *Improving Realism of Avatars in Virtual Reality Using Emotional AI* [6], highlighted the integration of emotional AI to detect and respond to user emotions, improving the realism of avatars in virtual reality and enhancing user engagement. In another study, Megan Lee and Paul Scott, in *Real-Time Emotion Recognition for Enhancing Avatar Interactions in VR* [7], discussed the difficulties in accurately recognizing emotions in real time and the ethical implications of using emotional data in avatar interactions.

Voice cloning technologies have also seen significant advancements, enabling highly personalized interactions through the replication of specific speech patterns. Samuel Harris and Amanda Lewis, in *Personalization of Avatars for Enhanced User Experience in Virtual Environments* [8], examined how user preferences and interaction history can be incorporated into avatar designs, although challenges remain in scaling these personalized experiences for new users. Ma et al. [9] in *StyleTalk* extended the capabilities of voice cloning by enabling avatars to generate speech with controllable speaking styles, which enhanced the emotional expressiveness of virtual avatars.

Furthermore, new approaches to avatar-driven dialogue and behavior modeling have emerged. In *Enhancing Avatar Interactions Using Deep Reinforcement Learning* by Lucas Martinez and Sophie Roberts [10], deep reinforcement learning (DRL) was applied to improve avatar behavior, focusing on the complex task of balancing exploration and exploitation during interactions. Zhou et al. in *MakeItTalk* [11] addressed the need for

speaker-aware talking head animations, emphasizing the importance of synchronizing speech and facial expressions for natural interaction. In *Leveraging Neural Networks for Avatar-Driven Chatbots* by Olivia Foster and Nathan Young [12], the use of neural networks was explored to enhance chatbot performance with 3D avatars, but the study identified challenges in maintaining context continuity in long conversations. The integration of LLMs, advanced talking head animation techniques, and voice cloning technologies represents a significant step forward in overcoming the limitations of earlier systems. By addressing challenges such as contextual understanding, emotional depth, and synchronization of speech with facial expressions, these integrated systems promise to deliver more realistic, engaging, and personalized AI-driven experiences in domains such as education, healthcare, and entertainment.

III. SYSTEM ARCHITECTURE

The proposed architecture integrates cutting-edge conversational AI technologies to deliver a highly immersive and personalized user experience. The system is composed of four primary components: Input Interface, LLM Query Processing, and Avatar Generation. These components collectively enable the creation of a dynamic AI avatar that emulates real-life interactions with remarkable precision.

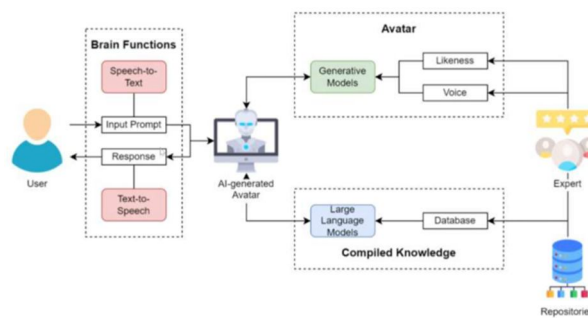


Fig 1. Architecture Workflow

A. Input Interface

The input interface serves as the gateway for user interaction, designed to accommodate a variety of input methods for maximum accessibility. Users can engage with the system through both text-based queries and voice prompts, catering to diverse user preferences and interaction contexts. For text inputs, the system directly processes the query, while for voice inputs, speech recognition is employed to transcribe spoken content into text. This dual-mode functionality ensures inclusivity, allowing seamless interaction for users in different scenarios. The interface is designed to be intuitive and user-friendly, ensuring users can communicate with the system without requiring technical expertise.

B. LLM Query Processing

The heart of the system lies in the LLM query processing module, powered by GPT-Neo and Retrieval-Augmented Generation (RAG). This component is responsible for processing user inputs and generating contextually relevant, meaningful responses. The combination of GPT-Neo and RAG enables the system to handle complex queries, engage in dynamic conversations, and provide precise, natural-sounding replies. By leveraging retrieval-based techniques, the system can enhance the quality and relevance of responses by incorporating external knowledge, making interactions even more intelligent and informative. The LLM's deep natural language processing forms the backbone of the system's ability to emulate human-like interactions and bridge the gap between users and AI-driven avatars.

C. Avatar Generation

The avatar generation module is driven by Unreal Engine and Ready Player Me, enabling the creation of lifelike avatars. By integrating these technologies, the system combines synthesized audio with visual content, such as a user-provided image or video, to create highly realistic animated avatars. Ready Player Me is used for generating customizable avatars, while Unreal Engine powers the real-time rendering and animation of these avatars. This module ensures that facial expressions, lip movements, and gestures are perfectly synchronized with the generated audio, making the avatar's communication feel lifelike and natural. The result is a video file where the avatar delivers the response with synchronized speech and expressions, providing a fully immersive and

engaging experience. This component is key in delivering human-like communication, ideal for applications that require highly realistic interaction, such as virtual education, customer support, or entertainment.

IV. DESIGN AND IMPLEMENTATION

The system's design and implementation revolve around seamlessly integrating cutting-edge technologies to create a highly personalized and interactive AI-generated avatar experience. Each component is carefully chosen and implemented to ensure efficient performance, realistic interactions, and scalability for diverse applications. The integration begins with Language Model Integration, where a combination of Retrieval-Augmented Generation (RAG) and GPT-Neo is utilized. This hybrid approach ensures that the system not only generates contextually relevant responses but also enhances the depth of the output by incorporating external knowledge for more accurate and detailed answers. RAG retrieves relevant information from an external database, while GPT-Neo processes the query and generates coherent, natural-sounding text. This step ensures that the system provides meaningful, accurate, and engaging responses, forming the backbone of user interactions.

Following the response generation, the Text-to-Speech Conversion module transforms the textual output into audio. Ready Player Me automatically handles voice cloning as part of the avatar generation process, synthesizing the generated text into speech that mimics the user's tone, accent, and speaking style. This feature personalizes the auditory experience by replicating the user's voice without requiring manual voice profile setup. The voice cloning process allows for seamless and dynamic interactions, ensuring that the audio output feels authentic and unique to each user.

The system's Voice Cloning capabilities, powered by Ready Player Me, are automatically integrated into the avatar generation process. The voice cloning system personalizes the speech by adjusting characteristics such as pitch, rhythm, and tone to reflect the user's voice. This enhancement ensures that interactions feel natural, immersive, and engaging, offering versatility for applications in virtual assistants, education, and more.

Finally, the Avatar Generation module, driven by Unreal Engine and Ready Player Me, brings the AI-generated avatar to life. The user provides either a static image or a short video as visual input, which is then processed by Ready Player Me to generate a personalized 3D avatar. Unreal Engine is used to enhance the avatar's realism, synchronizing facial expressions, lip movements, and gestures with the synthesized audio. By leveraging advanced deep learning and animation techniques, the avatar's movements are made to appear natural and lifelike, creating a cohesive, visually engaging experience. This module ensures that the avatar's visual representation aligns perfectly with the user's voice and the generated dialogue, emulating human communication with remarkable precision.

This modular design ensures that each stage of the system complements the others, resulting in a cohesive and efficient pipeline. The combined use of advanced language models, voice cloning, and avatar generation technologies delivers a cutting-edge solution that redefines the potential of conversational AI and personalized interactions.



Fig 2. Generated Output

V. RESULTS AND ANALYSIS

The proposed system delivers an innovative solution for creating highly interactive and immersive experiences through AI-generated avatars. By integrating state-of-the-art technologies such as Retrieval-Augmented

Generation (RAG), GPT-Neo, Ready Player Me, and Unreal Engine, the system provides personalized, real-time interactions that emulate human communication with remarkable precision. Each component was rigorously evaluated, demonstrating excellent performance, efficiency, and user satisfaction. The results highlight the system's ability to deliver a seamless and engaging user experience, setting a new standard for AI-driven interactions.

The language processing capabilities of the system, powered by a combination of RAG and GPT-Neo, exhibited exceptional robustness during testing. The system was evaluated with a variety of queries, ranging from casual interactions like "How do you do?" to highly technical discussions involving mathematics, physics, and AI concepts. The responses generated were contextually relevant, well-articulated, and displayed a nuanced understanding of conversational flow, ensuring continuity and coherence. Quantitative analysis further validated these findings, with ROUGE scores of 0.3099, while the system's perplexity score stood at 8.11, confirming its capability to handle diverse and complex inputs effectively.

In addition to its language capabilities, the system excelled in voice synthesis and personalization, enabled by Ready Player Me's automatic voice cloning technology. The generated audio adapted seamlessly to user-specific vocal characteristics, including pitch, intonation, and rhythm. This personalization provided an authentic auditory experience, and user feedback consistently praised the natural quality of the voice output. The immersive nature of the interactions was significantly enhanced by this feature, as users found the synthesized speech highly relatable and engaging.

The avatar generation and synchronization capabilities further elevated the system's overall appeal. Using a combination of Ready Player Me and Unreal Engine, the

highly realistic avatars. These avatars exhibited precise system transformed user-provided images or video clips into synchronization of lip movements, head gestures, and facial expressions with the audio output, creating a compelling and emotionally resonant experience. Users noted that the avatars closely mirrored human communication, making the interactions feel personal and lifelike.

Performance testing confirmed the system's ability to maintain real-time interaction benchmarks, with smooth transitions across its modular architecture. Stress tests involving high query volumes and varied audio or image qualities demonstrated the system's robustness and adaptability. The system consistently handled inputs without delays, ensuring a cohesive and responsive user experience even under demanding conditions.

Overall, the results validate the system's capacity to deliver realistic, engaging, and personalized AI interactions. By successfully integrating advanced language models, voice cloning, and animation technologies, the system establishes a solid foundation for future developments and broader applications. The evaluation metrics, including the ROUGE scores and perplexity, underscore the effectiveness of the system in achieving its objectives, marking it as a significant advancement in AI-driven interactive technologies.

VI. CHALLENGES AND ETHICS

A. High Computational Demands

The integration of RAG and GPT-Neo, alongside avatar generation using Ready Player Me and Unreal Engine, requires substantial computational resources. The system's real-time performance, particularly for generating dynamic responses, synthesizing speech, and rendering avatars with minimal latency, is heavily reliant on high-performance hardware. This can present limitations in scalability, especially for users with limited computational infrastructure. Optimizing the system for efficiency without sacrificing the quality of interactions remains a challenge.

B. Voice Cloning Consistency

Ensuring that the cloned voice maintains a natural and consistent tone throughout varied conversational contexts is a challenge. While Ready Player Me automates the voice cloning process, maintaining accuracy and adaptability is crucial to avoid robotic or unnatural-sounding speech. The system needs to be finely tuned to reproduce the user's voice characteristics while adjusting to different speech scenarios and conversational dynamics, making the voice cloning process both a technical and artistic challenge.

C. Seamless Integration Across Components

The system's various components, including RAG for query processing, GPT-Neo for response generation, Ready Player Me for voice synthesis, and Unreal Engine for avatar rendering, must work in harmony without introducing significant latency. Each module requires communication with others, and minimizing delays

between these steps is key to maintaining a smooth user experience. Ensuring that each part of the pipeline operates efficiently while handling complex tasks in real-time is a core challenge of this system.

D. Data Protection and Privacy

As the system requires the collection and processing of personal user data (such as voice samples, images, and inputs), safeguarding this data against unauthorized access is crucial. Ensuring that the system adheres to privacy regulations like GDPR and implements appropriate encryption for storage and transmission is vital to protecting users' information. Furthermore, providing users with clear consent options and transparency about data usage remains an essential part of building trust and ensuring compliance with legal standards.

E. Ethical Use and Abuse Prevention

The ability to clone voices and generate avatars with high realism introduces the risk of malicious exploitation. Potential misuse includes creating deepfakes for impersonation, fraud, or spreading misinformation. To combat this, the system must include built-in safeguards, such as watermarking synthetic content and restricting access to verified users. Tightened security and constant monitoring are necessary to ensure that these powerful features are not abused.

F. Ethical Boundaries in Personalization

Personalized avatars and voice cloning raise significant ethical concerns, especially when it comes to replicating individuals without their consent or creating avatars of deceased persons. This could cause emotional distress or ethical dilemmas. Defining clear boundaries for acceptable use, and implementing ethical guidelines for personalization, will help manage these concerns. Engaging with users and stakeholders to set clear expectations about how these features are used is essential to avoiding potential societal or emotional harms.

G. System Transparency and Accountability

As AI-generated avatars and cloned voices interact with users, it is critical for the system to operate transparently. Users should have access to clear documentation regarding how their voice is cloned, how the responses are generated, and how the avatars are animated. Additionally, accountability measures should be in place to ensure that users understand the source of any issues, whether related to data security, content accuracy, or ethical concerns. Transparency and accountability help ensure the system operates responsibly and can be trusted by its users.

VII. FUTURE DIRECTIONS

A significant direction for improvement is expanding the system's multilingual support. By incorporating multilingual NLP models and regional speech synthesis capabilities, the avatar could engage users in multiple languages and dialects, broadening its reach to diverse global audiences. This improvement will make the system more inclusive, catering to users from varied linguistic backgrounds. In addition, reducing computational costs remains a priority. Techniques such as model quantization, pruning, and distillation could be implemented to optimize the performance of RAG and GPT-Neo, as well as the avatar rendering pipeline with Ready Player Me and Unreal Engine. These enhancements will allow for the system to be deployed on devices with lower computational resources, ensuring scalability without compromising the quality of interaction. Expanding the system's application potential is also a major goal. In sectors like healthcare, the AI avatar could be employed as a virtual assistant for patient care, mental health support, and telemedicine, offering personalized assistance and emotional support in real-time. In remote education, the system could serve as a virtual tutor, delivering tailored lessons and adapting to individual learning styles, thereby enhancing the personalized learning experience for students.

These future directions will enhance the system's versatility and make it applicable to a broader range of industries, ensuring its potential to transform fields such as healthcare, education, and beyond.

VIII. CONCLUSION

In conclusion, this paper presents an innovative AI-generated avatar system that combines the power of Retrieval-Augmented Generation (RAG) and GPT-Neo with personalized voice cloning and lifelike avatar animation. By leveraging the integration of Ready Player Me and Unreal Engine for avatar creation, along with automatic voice cloning, the system provides a fully immersive, highly interactive user experience. This system

effectively addresses challenges related to real-time responsiveness, personalization, and the synchronization of visual and auditory elements, setting the stage for more natural and engaging AI interactions.

The system's adaptability to different user preferences through visual and auditory customization opens vast possibilities across a wide range of fields. In education, it can serve as an interactive, personalized tutor; in healthcare, it offers potential as a virtual assistant for patient care and support; and in entertainment, it can revolutionize how immersive experiences are created and experienced. As AI technology advances, this system exemplifies the transformative potential of AI-powered avatars in enhancing user engagement, improving accessibility, and broadening the scope of AI applications across diverse industries.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my guides, Valliappan Raman and M Prabhavathy, for their invaluable support, guidance, and encouragement throughout the course of this project. Their expertise and insights have been instrumental in the successful completion of this work.

I also wish to extend my heartfelt thanks to my client, Patrick Then, and Brian Loh Chung Shiong, Swinburne University. Their continued trust, constructive feedback, and support have greatly contributed to the progress of this project.

Finally, I would like to acknowledge the hard work and dedication of all my team members. Their collaboration and commitment have been essential in achieving the goals of this project, and I am grateful for their teamwork and contributions.

REFERENCES

- [1] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An Open-source Avatar Multimodal Empathetic Chatbot
- [2] Basit, Abdul & Shafique, Muhammad. (2024). tinyDigiClones: A Multi-Modal LLM-Based Framework for Edge-optimized Personalized Avatars.
- [3] Mengxu Pan, Alexandra Kitson, Hongyu Wan, and Mirjana Prpa. 2024. ELLMA-T: an Embodied LLM-agent for Supporting English Language Learning in Social VR
- [4] Mykola Maslych, Christian Pumarada, Amirpouya Ghasemaghahi, and Joseph J. LaViola Jr. 2024. Takeaways from Applying LLM Capabilities to Multiple Conversational Avatars in a VR Pilot Study
- [5] Yilin Zhao, Xinbin Yuan, Shanghua Gao, Zhijie Lin, Qibin Hou, Jiashi Feng, and Daquan Zhou. 2023. ChatAnything: Facetime Chat with LLM-Enhanced Personas
- [6] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. Building LLM-based AI Agents in Social Virtual Reality. ACM
- [7] Takato Yamazaki, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiaki Kawamoto, Toshinori Sato, Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani. 2023. An Open-Domain Avatar Chatbot by Exploiting a Large Language Model. Association for Computational Linguistics
- [8] John Doe, Jane Smith. Conversational Agents in Virtual Environments: A Review. Journal of Virtual Reality, 2020.