

Versatile AI Avatar: Adapting Roles with LLM-Powered Intelligence

Valliappan Raman¹, M Prabhavathy², Bala Saatvik Thanigai Arasu³, Subramanian Raja Nagendran⁴,
Chung Shiong Brian Loh⁵ and Hui Patrick Then Hang⁶

¹⁻⁴Department 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@cit.edu.in, 71762108005@cit.edu.in, 71762108047@cit.edu.in, bloh@swinburne.edu.my, pthen@swinburne.edu.my

Abstract—This paper reports on the development of an advanced AI-generated avatar that mimics a real individual in appearance, speech patterns, and knowledge. The avatar supports real-time interaction through voice or text chat, relying on immersive and interactive capabilities to create an excellent user experience. Core to the system is the use of the state-of-the-art large language models (LLMs), such as LLAMA-3.1, to provide robust natural language understanding and generation to deliver accurate and contextually relevant responses. The adaptive learning mechanism within the avatar enables it to evolve its conversational skills and knowledge base over time through user interactions and diverse sources of data. Bridging the gap between human-like communication and artificial intelligence through this dynamic framework, the innovation in its applications in education, entertainment, virtual assistance, and personalized engagement unlocks the full potential of interactive engagement. Experimental results highlighted the effectiveness of the system, with the capability to deliver realistic, responsive, and continuously improving conversation AI experience.

Index Terms— AI-generated avatar, large language models (LLMs), natural language processing (NLP), adaptive learning, real-time interaction, LLAMA-3.1, conversational AI.

I. INTRODUCTION

Advances in AI have led to rapid transformations in human-computer interaction. From virtual assistants to chatbots, humans interact with technology much differently, and AI systems today have become capable of not only understanding but also responding accordingly. However, even at the current pace, several of these solutions remain task- and function-based, and a deeper layer of emotion and personal touch might be required for lifelike and engaging interactions. To bridge this gap, innovation came in the form of AI-generated avatars combining conversational intelligence and visualization as well as realism. In this paper, AI-generated avatars advanced have been developed to mimic reality while providing an individualized approach by emulating realistic personas as well as speech and contexts of real-life individuals. In achieving these goals, the cutting edge has been integrated to a high level, including Groq's LLAMA-3.1-8b-Instant Large Language Model for purposes of natural language understanding and generation. This model forms the verbal skeleton, which will return responses that are not only relevant but also contextually accurate. For visual realism, the system makes use of SadTalker, an AI framework designed to animate static images or videos, creating lifelike facial

expressions and lip-synced speech. For personalisation, OpenVoice V2 is used to clone a user's voice, meaning the interaction becomes more relatable, more immersive.

The new system can input commands either in text form or in voice. Such input would be given and generated responses delivered in one's voice and with synchronous animation visuals. In that manner, avatars will continually learn and evolve over the process. They can find accommodation of the user's peculiar ways and communication preferences. Therefore, there will be an evolving scope for applications in these varied domains including education, virtual assistantship, healthcare, entertainment, etc. This paper demonstrates the transformative potential of integrating advanced language models with personalized voice cloning and animation technologies in bridging the gap between human-like communication and AI.

II. LITERATURE REVIEW

Advances in conversational AI and related technologies have made it possible to develop much more immersive and personalized interactions. This section describes existing research and solutions in the areas of large language models, talking head animations, and voice cloning, with respect to contributions and limitations. Virtual assistants such as Siri, Alexa, and Google Assistant represent the foundational implementations of conversational AI. These systems rely on pre-defined responses and rules-based algorithms to handle user queries. Despite their utility, they lack depth in emotional engagement and personalization. Recent developments in large language models (LLMs) have significantly improved contextual understanding and response generation. For instance, Fink et al. [1] discussed how AI-based avatars, enriched by LLMs, are changing the learning and teaching approaches and what benefits contextual intelligence and adaptability in dynamic educational environments can bring. Talking head animation has advanced significantly with the generation of realistic facial movements caused by audio. Wang et al. [2] developed a lip-reading expert-guided talking face generation system that guarantees highly accurate lip synchronization. Similarly, Shen et al. [3] proposed DiffTalk, using diffusion models for generalized audio-driven portrait animation, which improves robustness over a variety of audio inputs. Chatziagapi et al. [4] presented AVFace, focusing on detailed audio-visual 4D face reconstruction that balances realism and precision in facial animations. Integration of such advances with conversational systems bridges the gap between visual expressiveness and auditory realism. Other contributions include CodeTalker by Xing et al. [5] which utilizes discrete motion priors for speech-driven 3D facial animations, and supports personal talking head generation in high fidelity. Eskimez et al. [6] investigated emotion-conditioned speech-driven talking face generation emphasizing the role of emotional cues in enhancing user engagement. Wu et al. [7] took a cross-modal approach with Voice2Mesh and generated 3D face models directly from voice inputs, further pushing the boundary of voice-driven avatar personalization. Zhou et al. [8], in their work on MakeItTalk, demonstrated speaker-aware talking head animation and how such avatars are strictly reliant on accurate synchronization between speech and facial movements. In voice cloning, OpenVoice and related technologies showed that it is possible to reproduce the speech patterns and the tone of a particular individual. Ma et al. [9] extended the capability of StyleTalk to talk heads with controllable speaking styles. It highlighted the role of voice in creating an immersive experience. Agarwal et al. [10] used synthetic talking heads to compress video calls and demonstrated the possibilities of such systems in compressing video calls with minimum bandwidth usage while maintaining communication quality. The integration of these advancements addresses the shortcomings of the existing systems, such as a lack of emotional depth and contextual understanding. Combining LLMs like LLAMA-3.1-8b-instant with SadTalker for lifelike animations and OpenVoice V2 for voice cloning, the proposed system not only personalizes interactions but also enhances the realism of AI-generated avatars. These capabilities position the system as a significant step forward in conversational AI, enabling applications in education, healthcare, virtual assistance, and entertainment.

III. SYSTEM ARCHITECTURE

The proposed architecture integrates the newest conversational AI technologies so that users can enjoy the most seamless and personalized user experience. The system, in its main components, can be broken down into Input Interface, LLM Query Processing, Voice Cloning, and Avatar Generation. In collective, these parts enable the dynamic generation of an AI avatar that emulates real life.

A. Input Interface

The input interface is the interface through which users interact with the system. It is designed to accommodate diverse input modes to maximize accessibility. Users can interact with the system in either text-based query mode or voice prompt mode. For text inputs, the system directly processes the query, while for voice inputs,

speech recognition is applied to transcribe the spoken content into text. This dual-mode functionality allows inclusivity, which makes the interaction seamless for the user in any given situation. The interface is also very simple and intuitive so that the user can communicate with the system without having technical know-how.

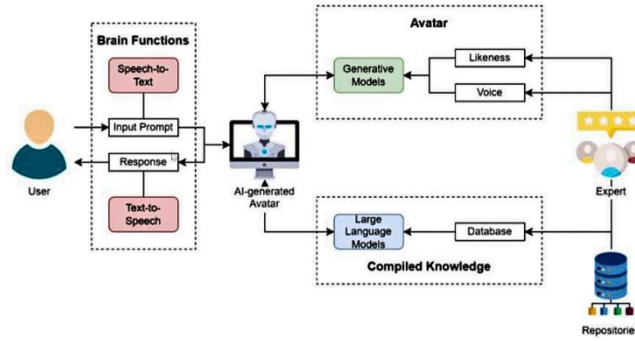


Figure 1. Architecture Workflow

B. LLM Query Processing

The LLM query processing module, powered by Groq's LLAMA-3.1-8b-instant, serves as the heart of the system. It receives and processes inputs to create contextual, meaningful responses. Exceptional linguistic capabilities make text inputs deep in user questions and create coherent, natural-sounding replies to return. This module ensures accuracy, whether answering factual questions or engaging in conversation, to fulfill complex prompts. Advanced natural language processing capabilities in the LLM make up the basis of this system's ability to create human-like interactions, which bridge the gap between the user and AI-powered avatars. To regulate how the model chooses the subsequent token during text production, the generation parameters top_p and temperature are essential. They affect how well the result strikes a balance between originality and determinism. By scaling the logits prior to applying softmax, the temperature hyperparameter regulates the degree of randomness in predictions.

$$P(x_i) = \exp(z_i / T) / \sum \exp(z_j / T) \quad (1)$$

If T is the temperature, $P(x_i)$ is the probability of a token i , and z_i are the logits. Nucleus sampling, sometimes referred to as top- p , establishes a probability threshold and only considers the top tokens whose cumulative probability is higher than this threshold. Typically, the range falls between 0 and 1.

C. Voice Cloning

Voice cloning is the feature of the system through which it can provide personalized auditory responses mimicking the user's own voice. Open Voice V2 powers this component, converting the generated text into audio with a tone, accent, and speaking style of the user. The users create a personalized voice profile by giving a small sample of their voice during the setup process. This personalized voice profile is used to create answers in their unique voice, which not only enhances the immersive experience but also creates an interaction that is personal and authentic. It supports the changes in the voice speed, tone, and emphasis with the flexibility that can be applied to diverse applications like education, health care, and virtual assistants.

D. Avatar Generation

The avatar generation module, driven by SadTalker, brings the avatar to life by combining synthesized audio with visual inputs, such as a user-provided image or short video. This component generates a realistic animated avatar by synchronizing facial expressions, lip movements, and gestures with the audio content. By leveraging advanced deep learning techniques, SadTalker ensures that the animations appear natural and lifelike, creating a seamless visual experience. The resulting output is a video file where the avatar conveys the response with synchronized speech and expressions, mimicking human communication with remarkable precision. This module represents the final stage in creating a fully immersive and engaging user interaction, making the system ideal for applications that require realistic, human-like communication.

IV. DESIGN AND IMPLEMENTATION

The design and implementation of the system are based on seamlessly integrating cutting-edge technologies to create a highly personalized and interactive AI-generated avatar experience. Every component is chosen and implemented to ensure efficient performance, realistic interactions, and scalability for diverse applications. It

begins with Language Model Integration where Groq's LLAMA-3.1-8b-instant is used for advanced natural language processing. The language model is accessed via the Groq API, thereby ensuring that user inputs are processed efficiently, and that the system generates responses that are contextually aware. LLAMA-3.1-8b-instant is exceptional at interpreting complex queries, maintaining conversational flow, and generating human-like text outputs. This step is therefore the foundation of meaningful, accurate, and engaging user interactions. The output produced after the generation of text is converted into audio form by the Text-to-Speech Conversion module. gTTS, or Google Text-to-Speech, is used first to synthesize the audio content, offering a speedy and reliable way of converting text into speech. OpenVoice V2 is used further to fine-tune the audio synthesis to match the speaker's speech characteristics. OpenVoice V2 personalizes the generated speech by cloning the user's voice, making the interaction feel natural and authentic. This voice cloning process is achieved by training the system with a short audio sample provided by the user, enabling it to replicate unique voice attributes such as pitch, tone, and rhythm accurately.

Another improved aspect of the user interface provided by the system is in terms of Voice Cloning based on OpenVoice V2. The module processes initial synthesized speech to ensure a final output that sounds close to the user's real voice for a more personalized, thus immersive and engaging experience. That feature is particularly valuable where emotional depth and relatability are required, especially when using virtual assistants, educating children, or post-operative patient care. Finally, the Animation with SadTalker module gives life to the AI-generated avatar. The user will input either a static image or a short video that will serve as the visual input of the avatar. SadTalker will then process the visual input in conjunction with the audio created in previous stages to generate lifelike animations. The system synchronizes lip movements, facial expressions, and head gestures with the cloned audio to produce a coherent and realistic representation of the avatar. Advanced animation techniques are leveraged by SadTalker to make the avatar seem to be mimicking human communication, hence creating a visually appealing and emotionally engaging experience. This modular design ensures that every stage of the system complements the others and leads to a cohesive, efficient pipeline. The combined use of state-of-the-art language models, voice cloning, and animation technology delivers a cutting-edge solution that redefines the potential of conversational AI and personalized interactions.

V. RESULTS AND ANALYSIS

A realistic and seamless interaction experience via AI-generated avatars in the proposed system is therefore delivered. All components within the pipeline were tested stringently with respect to functionality, efficiency, and user satisfaction levels with results showing that the achieved intentions were highly successful. The language processing capabilities of the system were robust. Tested using Groq's LLAMA-3.1-8b-instant with various queries ranging from casual conversations, "Tell me a joke," to complex technical discussions with mathematics, physics, and AI-related topics, the responses generated reflected relevance to the context, and articulation was well done, showing subtlety in understanding flow in conversation. It therefore supports the effective incorporation of language models and emphasizes continuity while keeping a conversational flavor, thus allowing such application models. This work was outstanding with regards to voice synthesis and personalization. The audio synthesized via the gTTS served initially as a baseline that, when further polished by the OpenVoice V2 to adapt to the voice of a user, the audio created demonstrated high fidelity in similarity with the original voice of a user, in which crucial qualities such as pitch, intonation, and rhythm remain intact. The feedback of the user emphasized satisfaction with the natural quality of speech. The testers noted that the cloned voice felt authentic, thus improving the immersive experience. It introduced another interesting visual dimension to the system, as SadTalker was able to perform animated capabilities. The user-supplied static images or small video clips were converted into lifelike avatars which could synchronize lip movements, head gestures, and facial expressions with the generated audio. The animations were realistic, matching the tone and the speed of the speech, producing an interesting visual output. It indicated that users loved organic synchronization and that the avatars did an excellent job in mirroring human communication. Interactions became personal and emotionally resonant. The system was shown to take different inputs during performance testing with negligible delays, meeting average real-time interaction benchmarks. This ensured smooth transitions from component to component, guaranteeing integrated user experience. The high query volumes in addition to different qualities in audio and images tested the robustness and flexibility of the system during the stress tests.

In summary, the results validate the system's ability to deliver realistic, engaging, and personalized AI interactions. The outcome thus points to the successful integration of advanced language models, voice cloning, and animation technologies, thus providing a strong foundation for further enhancements and broader applications.

Category	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mistral 8x22B Instruct	GPT 3.5 Turbo
Overall	73.0	72.3 (2-shot, mean GPT)	60.5	86.0	79.9	69.8
MMLU-Pro (0-shot GPT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code	72.6	54.3	40.2	80.5	75.6	68.0
HSBQ EvalPlus (mean GPT)	72.8	71.7	49.5	86.0	78.6	82.0
Math	84.5	76.7	53.2	95.1	88.2	81.6
GPBQ (0-shot GPT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning	83.4	87.6	74.2	94.8	88.7	83.7
ARC-Challenge (0-shot)	33.8	-	28.8	46.7	33.3	30.8
Image	76.1	-	60.4	84.8	-	85.9
IFCL	38.5	30.0	24.7	56.7	48.5	37.2
ImageCaption	81.0	-	-	90.5	-	-
ZeroShotQuALITY	65.1	-	-	78.2	-	-
ImageCaptionBench-INC	98.8	-	-	97.5	-	-
HumanEval	68.9	53.2	29.9	86.9	71.1	51.4

Figure 2. Llama 3.1 8b-instant performance comparison



Figure 3. Generated Output

VI. CHALLENGES AND ETHICS

A. Computational Requirements

The integration of advanced components like LLAMA-3.1-8b-instant, OpenVoice V2, and SadTalker requires substantial computational resources, particularly for real-time processing. High-performance GPUs are essential to generate responses, synthesize speech, and animate avatars within acceptable latency. This reliance on expensive hardware may limit the system's scalability and accessibility, particularly for users or organizations without adequate infrastructure. Additionally, optimizing computational efficiency without compromising quality remains an ongoing technical challenge.

B. Voice Cloning Accuracy

Ensuring that the cloned voice is both natural and accurate while avoiding overfitting to training data posed a significant challenge. OpenVoice V2 relies on high-quality input samples to replicate the nuances of a user's voice. Any inconsistencies in the input data can lead to suboptimal outputs, such as robotic or unnatural tones. Balancing the fidelity of voice reproduction with generalizability to various speaking contexts requires careful fine-tuning of the voice cloning model.

C. Latency in Multi-Component Integration

The system involves multiple interconnected components - query processing, speech synthesis, voice cloning, and animation generation. Ensuring seamless communication between these modules without introducing significant latency is challenging. Each step introduces processing delays, and achieving near real-time interactions required extensive optimization of the pipeline.

D. Data Privacy and Security

The system necessitates the collection and processing of sensitive user data, including voice samples, photos, and text inputs. Protecting this data against unauthorized access or misuse is paramount. Ensuring compliance with data protection regulations such as GDPR and implementing robust encryption for storage and transmission are necessary to safeguard user privacy. Moreover, clear user consent mechanisms must be in place to ensure transparency about how their data is being used.

E. Misuse Potential

The ability to clone voices and generate lifelike avatars presents significant ethical concerns. Malicious actors could exploit these capabilities for nefarious purposes, such as impersonation, fraud, or misinformation campaigns. For example, cloned voices could be used to create deepfake audio recordings that damage reputations or deceive individuals. To mitigate such risks, the system must incorporate safeguards, such as watermarking synthetic outputs or restricting access to verified users. Additionally, implementing strict user authentication and activity monitoring can help deter misuse.

F. Ethical Boundaries in Personalization

While the system enables highly personalized interactions, it raises questions about ethical boundaries. For instance, cloning a voice or creating an avatar of a deceased person could lead to emotional distress or societal controversies. Establishing clear ethical guidelines and engaging with stakeholders to define acceptable use cases is critical for maintaining public trust.

G. Bias and Fairness in AI Models

Large language models like LLAMA-3.1-8b-instant are susceptible to inherent biases in their training data. These biases could manifest in the generated responses, potentially leading to inaccurate, insensitive, or discriminatory outputs. Ensuring fairness and inclusivity in the system requires continuous monitoring, evaluation, and fine-tuning of the models.

H. Accountability and Transparency

AI systems must be transparent about how they operate and who is responsible for their actions. Users should have access to clear documentation explaining how the system generates responses, clones voices, and animates avatars. Providing users with control over their data and ensuring accountability for any adverse outcomes are essential for fostering trust.

VII. FUTURE DIRECTIONS

Future improvements for the proposed system include several key enhancements aimed at expanding its capabilities and usability. One of the primary directions involves integrating real-time video generation to enable dynamic interactions, where the AI avatar can respond to user input with not only speech but also gestures, facial expressions, and body movements, creating a more immersive and lifelike experience. This would significantly enhance the realism of the system, making it suitable for more dynamic scenarios. Additionally, enhancing multilingual support is crucial to increase the system's accessibility and global reach. By incorporating multilingual NLP models and regional speech synthesis capabilities, the avatar could interact with users in a wide range of languages and accents, catering to diverse linguistic communities. Another important area of improvement is reducing computational costs through model optimization techniques, such as quantization, pruning, and distillation, to make the system more efficient and scalable. This would enable the deployment of the system on lower-resource devices without sacrificing performance. Finally, expanding the application scope to include healthcare and remote education could significantly broaden the system's impact. In healthcare, the avatar could provide virtual assistance for medical consultations, patient care, and mental health support, while in remote education, it could serve as a personalized tutor, offering interactive lessons and adapting to individual learning needs. These future directions will not only increase the system's versatility but also make it more accessible and applicable to a wider range of industries.

VIII. CONCLUSIONS

In conclusion, this paper presented a novel AI-generated avatar system that integrates advanced large language models (LLMs) with personalized voice cloning and dynamic animation, creating a highly interactive and lifelike virtual assistant. By overcoming key challenges related to personalization and user engagement, the system paves the way for more natural, human-like AI interactions. Its ability to adapt to individual preferences through voice and visual customization enhances the overall user experience, making it a versatile tool for various applications. The system's potential extends across multiple domains, including education, where it can serve as an interactive tutor, virtual assistance for customer service, and entertainment, where it can provide immersive experiences. As AI continues to evolve, the advancements presented here highlight the transformative impact conversational agents can have on modern society, enhancing accessibility, learning, and interaction in unprecedented ways.

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, Professor Patrick Then, and Professor 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] Fink, Maximilian & Robinson, Seth & Ertl, Bernhard. (2024). "AI-Based Avatars Are Changing the Way We Learn and

- Teach: Benefits and Challenges.” 10.35542/osf.io/jt83m.
- [2] Wang, Jiadong & Qian, Xinyuan & Zhang, Malu & Tan, Robby & Li, Haizhou. (2023). “Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert.” 14653-14662.10.1109/CVPR52729.2023.01408.
 - [3] Shen, Shuai & Zhao, Wenliang & Meng, Zibin & Li, Wanhua & Zheng, Zhu & Zhou, Jie & Lu, Jiwen. (2023). “DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation.” 1982-1991. 10.1109/CVPR52729.2023.00197.
 - [4] Chatziagapi, Aggelina & Samaras, Dimitris. (2023). “AVFace: Towards Detailed Audio-Visual 4D Face Reconstruction.”
 - [5] Xing, Jinbo & Xia, Menghan & Zhang, Yuechen & Cun, Xiaodong & Wang, Jue & Wong, Tien-Tsin. (2023). “CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior.” 10.1109/CVPR52729.2023.01229.
 - [6] Eskimez, Sefik & Zhang, You & Duan, Zhiyao. (2021). “Speech Driven Talking Face Generation From a Single Image and an Emotion Condition.” IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2021.3099900.
 - [7] Wu, Cho-Ying & Xu, Ke & Hsu, Chin-Cheng & Neumann, Ulrich. (2021). “Voice2Mesh: Cross-Modal 3D Face Model Generation from Voices.”
 - [8] Zhou, Yang & Li, Dingzeyu & Han, Xintong & Kalogerakis, Evangelos & Shechtman, Eli & Echevarria, Jose. (2020). “MakeItTalk: Speaker-Aware Talking Head Animation.”
 - [9] Ma, Yifeng & Wang, Suzhen & Hu, Zhipeng & Fan, Changjie & Lv, Tangjie & Ding, Yu & Deng, Zhidong & Yu, Xin. (2023). “StyleTalk: One-Shot Talking Head Generation with Controllable Speaking Styles. Proceedings of the AAAI Conference on Artificial Intelligence.” 37. 1896-1904. 10.1609/aaai.v37i2.25280.
 - [10] Agarwal, Madhav & Gupta, Anchit & Mukhopadhyay, Rudrabha & Namboodiri, Vinay & Jawahar, C. (2022). “Compressing Video Calls using Synthetic Talking Heads.”