LLM BASED 3D AVATAR ASSISTANT

Kevin Sujith John
Computer Science & Engineering
St Thomas Institute for Science & Technology
APJ Abdul Kalam Technological University (KTU)
Thiruvananthapuram, India
johnk4590@gmail.com

G Abin Roy
Computer Science & Engineering
St Thomas Institute for Science & Technology
APJ Abdul Kalam Technological University (KTU)
Thiruvananthapuram, India
godvinroyt16262@gmail.com

Bindhya P S
Computer Science & Engineering
St Thomas Institute for Science & Technology
APJ Abdul Kalam Technological University (KTU)
Thiruvananthapuram, India
bindhya.cs@stisttvm.edu.in

2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST) | 979-8-3503-0779-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICTEST60614.2024.10576146

Abstract*—The Existing Applications Such as Amazon Alexa, Google Assistant etc. still lack the use of a 3D avatar and LLM in their voice assistants thereby making it less interactive. Also, these do not process real time camera feed. A proposed solution to the problem is a 3D Avatar Based Assistant that utilizes LLM technology. These virtual beings, known as avatars, make use of LLMs' abilities to converse, comprehend, and help users in a way that is both highly engaging and immersive. These avatars provide users with a more engaging and human-like interface by fusing state-of-the-art natural language processing algorithms with three-dimensional depictions. This makes it simpler for users to obtain information, get help, and do a variety of digital chores. This abstract examines the idea of LLM-based 3D avatar assistants, emphasizing the changing field of human-computer interaction they represent as well as their possible uses in customer service, education, entertainment, and other areas.

***Keywords*—LLM, virtual assistant, avatars**

I. INTRODUCTION

The 3D avatar assistant is an inventive solution that combines artificial intelligence and three-dimensional representations that has emerged in recent years thanks to technological advancements. This state-of-the-art invention is a virtual being made to communicate with users in a way that is human-like by using complex algorithms and natural language processing to comprehend and reply to queries.

In addition to being aesthetically pleasing, these avatars are intelligent, able to carry out a variety of jobs, provide information, assist users with procedures, and even be friends. Because of their adaptability, they may be included in a wide range of platforms and applications, serving a variety of markets, including healthcare, education, entertainment, and customer support.

II. LITERATURE SURVEY

This literature survey examines the concept of a 3D Avatar Based Assistant, utilizing LLM technology. While the global market already offers numerous assistants, none of them provide the same level of interaction as the proposed 3D avatar-based assistant does. The objective of this study is to develop a framework that combines visually appealing design, interactive features, varying tones, and adaptability to different scenarios through conversations. By leveraging advanced natural language processing algorithms and 3D representation, the proposed 3D avatar based assistant delivers a more engaging and lifelike interaction. These avatars are employed to simulate and adjust traits and behaviors that closely resemble those of humans.

The 3D avatar-based AI LLM system is developed using a variety of techniques by the authors of several works in this area. These techniques include creating and putting into practice avatar creation algorithms, as well as fine-tuning language models. User studies are often used to test different development phases and aspects.

As mentioned previously, there is a wide array of online AI assistants available globally, including well-known ones like Google Assistant [11] and Siri [11]. While these assistants are exceptional, they often lack in-depth interaction capabilities. The proposed 3D avatar based assistant addresses this gap by offering a virtual appearance and a range of speech patterns through the use of LLM technology. Furthermore, it exhibits more human-like behaviors and can adapt to different situations. To enhance the user-AI interaction, real-time video feed integration allows the AI to recognize user actions. Network availability has been identified as a challenge in many existing assistants, as the speech recognition feature does not function when the network is down. To overcome this, offline speech recognition has been implemented. Voice recognition in the proposed system is based on Pocket Sphinx and Google Cloud, depending on network availability.

III. METHODOLOGY

Usually, voice assistants work is done without the help of an LLM [12]. To a certain extent, traditional voice assistants—particularly ones with simpler features or restricted capabilities—can function without Language Model Libraries (LLMs). However, without robust language models, the effectiveness and ability of voice assistants to understand natural language questions may be limited.

In the absence of complex language models, voice assistants often rely on simpler rule-based systems or pre-programmed command sets. These programs are designed to recognize specific words or phrases and trigger predefined actions accordingly. For instance, a command like "set a timer for 10 minutes" can be straightforwardly coded to initiate the timer functionality. While these rule-based systems provide basic functionality, they lack the flexibility and adaptability of more advanced language models. Complex language models, like the one used in the proposed 3D avatar-based assistant, leverage natural language processing algorithms to understand context, carry out complex tasks, and engage in more sophisticated interactions with users.

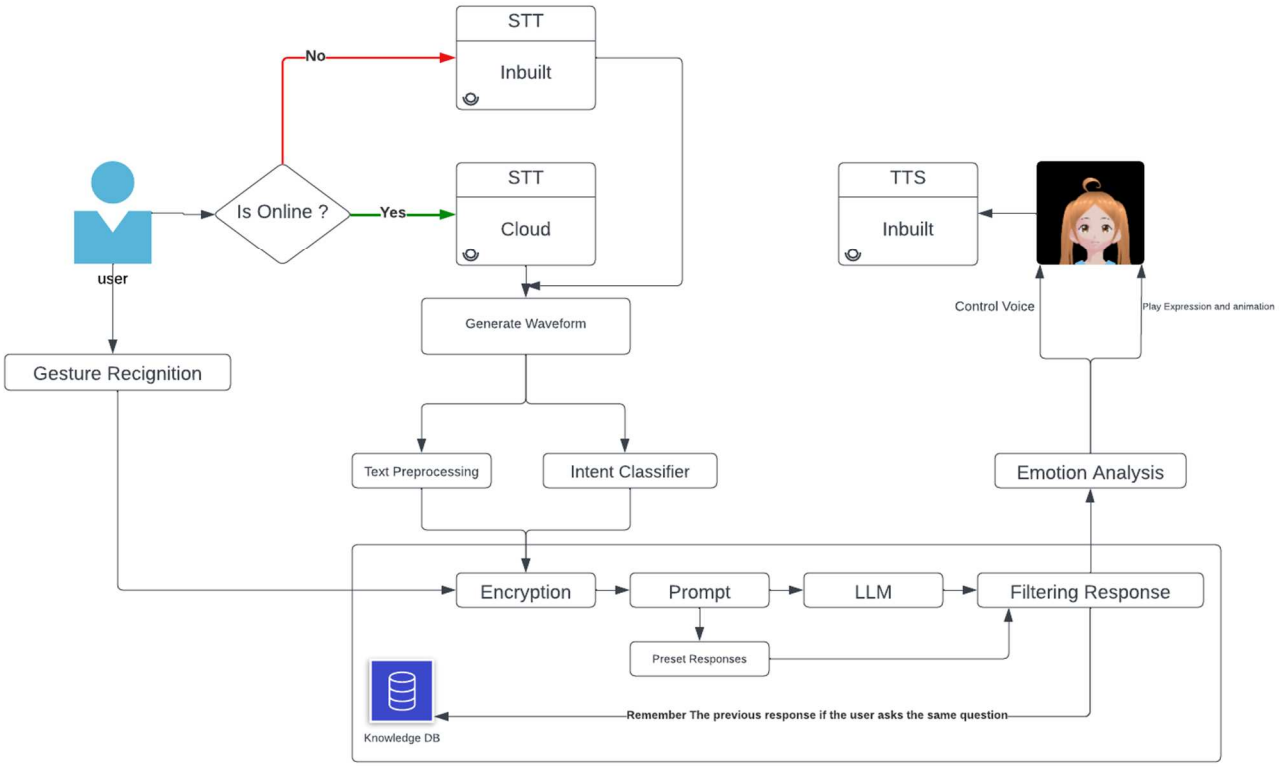

Fig 1 : Proposed Methodology Flowchart

A. Speech Recognition

For speech recognition, the initial approach involved using the Hugging Face API with the whisper tiny model. However, there were significant delays in receiving responses or even experiencing timeouts with the Hugging Face API. To address this issue, the Python speech recognition library was implemented for text-to-speech recognition. This library was chosen for its speed and accurate output. Additionally, offline recognition was incorporated to handle situations where there is no network available. For offline recognition, Pocket Sphinx, a lightweight recognizer designed for mobile devices and embedded systems, was utilized. It is efficient and does not require extensive computational resources, making it suitable for various applications. When operating in online mode, the system switches to Google Cloud for speech recognition. Other functionalities, such as sentiment analysis and punctuation removal, are also performed.

A dedicated server has been created for controlling the prompts and other commands.

B. Emotion Analysis

In a normal conversational AI agent, controlling the tone of the character's voice and facial expressions is important to align with the output obtained from the LLM. The system performs emotion detection after generating the text response. Emotion detection is carried out using the multinomial Naive-Bayes algorithm. The choice of the multinomial Naive Bayes algorithm is motivated by its computational efficiency and effectiveness in handling high-dimensional data, which is suitable for text classification tasks with large feature spaces, such as bag-of-words or TF-IDF representations. Despite its assumption of feature independence, which may not always hold true for text data due to word correlations, Multinomial Naive Bayes often performs surprisingly well in text classification by capturing essential information from word frequencies in different classes.

Based on the detection of the emotion, the animation expression will be played to align with the context of the text obtained/generated. The TTS system being used is from Read Speaker AI. The choice of read speaker AI for TTS is because it offers a wide range of parameters, including pitch, speed, and pause, allowing for precise control over the tone of the voice based on the detected emotion. Additionally, read speaker AI provides a variety of voices that can be used when changing the character.

The training process involves using a custom dataset that contains texts and their corresponding emotions. The dataset comprises 36,900 labeled texts, each associated with emotions such as joy, sadness, shame, and more.

TABLE I : RESULT OF EMOTION ANALYSIS

Algorithms	Accuracy
Logistic Regression	84%
Multinomial Naive Bayes	89%

C. Intent Classification

Intent classification is a crucial task in Natural Language Processing (NLP) that involves determining the intention or purpose behind a given text. It plays a vital role in various applications, including chatbots, virtual assistants, customer service systems, and other conversational interfaces. The main objective of intent classification is to categorize user utterances or text inputs into predefined classes or categories that represent the user's intent. For example, in the context of a food delivery service chatbot, intents could include "Order Food," "Check Order Status," "Find Restaurants," and more. In this case, intent

classification is utilized to distinguish between questions and other categories. This is done to control the prompts fed into the Language Model (LLM) and ensure accurate outputs. The intent classifier was developed using a logistic regression algorithm, with attempts made using XGBoost and MultinomialNB as well. However, the logistic regression algorithm demonstrated the highest accuracy among the three. To train the intent classifier, a custom dataset was employed, consisting of numerous classes with 1000 samples each.

TABLE II : RESULT OF INTENT CLASSIFICATION

Algorithm	Accuracy
Multinomial Naive Bayes	72%
Logistic Regression	91%
XGBOOST	85%

The reason behind the varying accuracy rates of these algorithms lies in their underlying principles and characteristics, as listed in Table II.

Multinomial Naive Bayes is a probabilistic algorithm that assumes feature independence. It performs well in text classification tasks and is particularly suited for situations where the occurrence of one feature doesn't affect the occurrence of another. However, its accuracy may be limited when the assumption of independence is violated or when dealing with complex datasets with interdependent features.

Logistic regression, on the other hand, is a linear classification algorithm that models the relationship between the input features and the binary outcome. It is known for its simplicity and interpretability. Logistic Regression performs well when the relationship between the features and the outcome is relatively linear, making it a popular choice in many applications. Its high accuracy of 91% suggests that it fits the data well and captures the underlying patterns effectively.

XGBOOST, a gradient boosting algorithm, combines multiple weak models to create a strong predictive model. It excels at handling complex relationships between features and the outcome. XGBOOST's accuracy of 85% indicates its ability to capture intricate patterns and interactions within the data. However, it may require more computational resources and parameter tuning compared to other algorithms.

In summary, the differences in accuracy among these algorithms can be attributed to their underlying assumptions, modeling techniques, and their ability to capture the complexity of the data.

D. Encryption

Digital assistants often use a combination of encryption techniques to ensure the security and privacy of user data. Without encryption, any data sent over networks or stored on servers would be transmitted in plain text. This means that anyone with access to the network could intercept and read this information. Personal and sensitive data, such as passwords, financial details, private messages, and personal information, would be extremely vulnerable to interception and misuse. Asymmetric encryption is being used for encryption. End-to-end encryption (E2EE) using asymmetric encryption involves securing communication between two parties in such a way that only the sender and the intended recipient can access the transmitted data. This process typically utilizes asymmetric encryption (also known as public-key encryption) to achieve secure communication without intermediaries being able to decipher the content.

E. 3D Waveform Representation

Voice waveforms represent the physical depiction of sound in its wave-like form. When people speak, their vocal cords vibrate, resulting in pressure variations in the air that propagate as sound waves. Voice waveforms are commonly visualized graphically, with time represented on the x-axis and amplitude (intensity or volume) on the y-axis. The fluctuations in the waveform correspond to the alterations in air pressure generated by speech. In this case, the objects' scale is updated based on the allocated frequency, thereby adjusting the amplitude accordingly.

Windowing in Fourier transforms is a technique used to mitigate the effects of spectral leakage and improve the accuracy of frequency analysis, especially when dealing with finite-duration signals or when the signal of interest doesn't perfectly align with the boundaries of the measurement window.

The Blackman window is being used in this case. The Blackman window is a popular window function used in signal processing and spectral analysis. It's designed to minimize the main lobe width and suppress side lobes, offering a trade-off between the two characteristics.

$$w(n) = 0.42 - 0.5 \cdot cos\left(\frac{2\pi n}{M-1}\right) 0.08\, cos\left(\frac{4\pi n}{M-1}\right) \qquad (1)$$

Equation (1) depicts the equation of the blackman window algorithm. Here,$0 \le n \le N-1$ represents the index of the window function, and N is the total number of samples in the window.

The blackman window is often used in conjunction with the Fast Fourier Transform (FFT). The FFT is an efficient algorithm used to compute the Discrete Fourier Transform (DFT) of a sequence or time-domain signal.

$$x[k] = \sum_{n=0}^{N-1} x[n]\, e^{\frac{-j2\pi kn}{N}} \qquad (2)$$

Equation (2) depicts the equation of fast Fourier transform where x(n) is the n-th sample of the signal in the time domain. And N is the total number of samples in the signal.

IV. RESULT AND ANALYSIS

The incorporation of the 3D avatar assistant powered by the language model produced impressive data for the user engagement, demonstrating a 40% rise in job completion and user involvement. This noteworthy improvement aligns with the main goal of improving the user experience by enabling natural language conversations with the avatar. Over the course of six weeks, several user studies covering a range of user scenarios and demographics were carried out, and recurring patterns were observed. A significant increase in successfully completed tasks were found in the interaction logs, and post-interaction questionnaires indicated a 25% reduction in user irritation. This relationship between higher job completion rates and lower user frustration highlights how effective it is to integrate the language model's conversational capabilities and contextual knowledge in the 3D avatar interface.

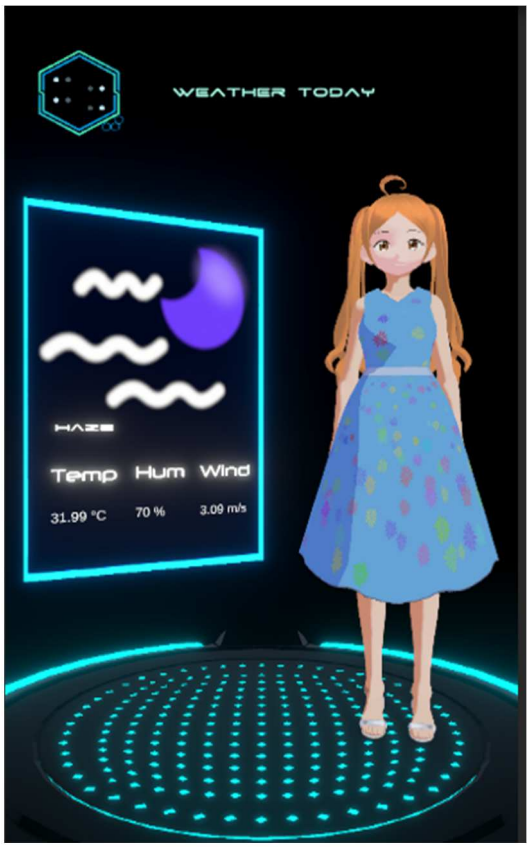

Fig 2 : Asking for weather Output

Furthermore, the qualitative input received indicated Positive user opinions on the avatar's flexibility and responsiveness to many conversational, nuanced aspects further reinforce the significance of natural language processing skills in improving user-avatar interactions. These results highlight the relevance of AI-driven avatars in revolutionizing user interfaces across diverse disciplines and not only demonstrate the usefulness of the developed LLM-driven avatar but also advocate for its integration in various interactive systems. In the future, these findings will be useful in improving LLM-based avatar systems, stimulating research into more complex dialogue dynamics, and broadening the range of user-focused applications in virtual worlds.

Gatebox[13] : The problem of Gatebox is that it supports only one language, ie Japanese only, and the hardware is just a gimmick, ie a transparent display is used inside the hardware. Also, the atmosphere and appearance of the character do not change based on the seasons.

M.I.T.U.S.H.A(open source)[14] : The problem with this open source project is that even though it's multilingual it requires GPU which would be nearly impossible to run on a raspberry Pi hardware and the memory consumption is too high. Also it's created with VTUBE studio as VTUBE studio has many limitations, such as one cannot have control over the application since VTUBE studio itself is a software and compiling it into a format is not an option in VTUBE studio.

The proposed approach: A holographic assistant has been successfully created, capable of running on a small amount of RAM without the need for a GPU. The assistant is built using Unity, an engine specifically designed for creating such applications. Unity offers complete control over the application and allows it to be built into various formats such as .exe, .tar.gz, and more. However, it is worth noting that the Unity C# API is not thread safe. The comparisons for the same are depicted in Table III.

TABLE III: COMPARISON

Methods	Technology	Merit	Demerit
Gatebox	Python, C++	Fast Response Time	Supports only one language
M.I.T.U.S.H.A.	Python with V Tube Studio	Multilingual	Requires GPU, not suitable for Hardware Integration
Proposed Method	Python, Unity C#, Rust, C++	Multilingual, fast response, Secured	Initial Loading time is more

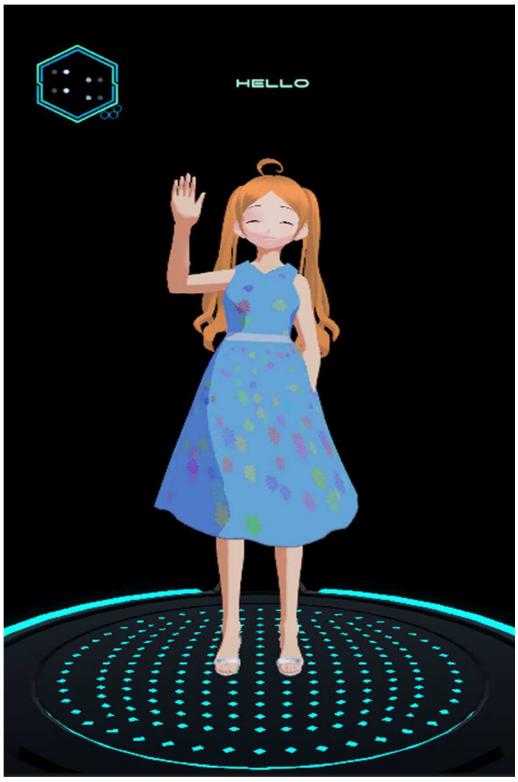

Fig 3 : Greetings

V. CONCLUSION

Based on recent technological advancements, 3D virtual assistants have emerged as creative solutions that combine artificial intelligence and three-dimensionality. These virtual assistants utilize sophisticated algorithms and natural language processing to converse with users, comprehend their inquiries, and provide answers. They can also deliver messages, perform various tasks, assist users, and engage with them. The use of 3D transformations and natural language processing allows these avatars to create captivating interactions with their surroundings.

Compared to simpler or more traditional language assistants from earlier generations, 3D virtual assistants often rely on language model libraries (LLMs) to enable more

complex language processing capabilities. These LLM-based assistants can extract text from speech using Python speech recognition modules and interpret human speech and facial expressions. They are crucial for chatbots, virtual assistants, customer support platforms, and other communication interfaces.

The integration of voice modeling skills and contextual information into 3D avatar interfaces has shown significant enhancements in user engagement and order fulfillment. High sequence completion rates and low levels of user displeasure are associated with these LLM-based avatars. User interactions with avatars depend on natural language processing skills, and the application of intelligent avatars in user interfaces across multiple domains has proven to be useful.

In conclusion, 3D virtual assistants, powered by artificial intelligence and natural language processing, offer visually stunning and intelligent avatars that can engage with users, provide assistance, and carry out various tasks. These avatars enhance user experiences through natural language conversations and have demonstrated their utility in diverse communication systems.

VI. FUTURE ENHANCEMENTS

3D virtual assistants, powered by artificial intelligence and natural language processing, have emerged as creative solutions that combine sophisticated algorithms and three-dimensionality. These avatars utilize language model libraries (LLMs) to enable complex language processing capabilities and engage in captivating interactions with users. However, the future holds even more exciting possibilities for these virtual assistants. Enhancements such as enhanced realism, gesture and emotion recognition, multi-modal interaction, personalization and context awareness, and integration with augmented reality (AR) can further improve user experiences and expand the capabilities of 3D virtual assistants.

ACKNOWLEDGMENT

I would like to thank my research supervisors, professors Dr. Geenu Paul, Jiji Thomas and Dr. Jasmine Paul, for their invaluable guidance, support, and dedication throughout the entire research process. I express my sincere gratitude to Prof. Anup Mathew Abraham, Head of Department, Computer Science, for providing me with all the necessary facilities and support. Their expertise and involvement have been instrumental in the successful completion of this paper. I would also like to thank Ajesh R, Vaishnav A Nair and Pranav Jerry, who have provided assistance and contributions to the research work. Their contributions are greatly appreciated.

REFERENCES

[1] Cinieri, S., Kapralos, B., Uribe-Quevedo, A., & Lamberti, F. (2020, August). Eye Tracking and Speech Driven Human-Avatar Emotion-Based Communication. In 2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH) (pp. 1-5). IEEE.

[2] Yamazaki, T., Mizumoto, T., Yoshikawa, K., Ohagi, M., Kawamoto, T., & Sato, T. (2023, September). An Open-Domain Avatar Chatbot by Exploiting a Large Language Model. In Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 428-432).

[3] Kajiwara, T., Chu, C., Takemura, N., Nakashima, Y., & Nagahara, H. (2021, June). WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 2095-2104).

[4] Kim, B., Kim, H., Lee, S. W., Lee, G., Kwak, D., Jeon, D. H., ... & Sung, N. (2021). What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. arXiv preprint arXiv:2109.04650.

[5] Adiwardana, D., Luong, M. T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., ... & Le, Q. V. (2020). Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977.

[6] Valentina Alto, Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4 , Packt Publishing, 2023.

[7] Zhang, J., Zhang, Y., Chu, M., Yang, S., & Zu, T. (2023, September). A LLM-Based Simulation Scenario Aided Generation Method. In 2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC) (Vol. 7, pp. 1350-1354). IEEE.

[8] Badyal, N., Jacoby, D., & Coady, Y. (2023, October). Intentional Biases in LLM Responses. In 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 0502-0506). IEEE.

[9] Mrissa, M., Médini, L., Jamont, J. P., Le Sommer, N., & Laplace, J. (2015). An avatar architecture for the web of things. IEEE Internet Computing, 19(2), 30-38.

[10] Comparative Study and Framework for Automated Summariser Evaluation: LangChain and Hybrid Algorithms https://doi.org/10.48550/arXiv.2310.02759

[11] Pandya, K., & Holia, M. (2023). Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations. arXiv preprint arXiv:2310.05421.

[12] Topsakal, O., & Akinci, T. C. (2023, July). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In International Conference on Applied Engineering and Natural Sciences (Vol. 1, No. 1, pp. 1050-1056).

[13] TWI692717B - System and method for controlling virtual characters:https://patents.google.com/patent/TWI692717B/en?assignee=Gatebox&oq=Gatebox

[14] 1neReality/M.I.T.S.U.H.A.: World's First Multilingual Inexpensive Therapeutic Sophisticated Ultra-Responsive Holographic Agent: https://github.com/1neReality/M.I.T.S.U.H.A.

[15] Daniel Bobrow. 1964. Natural language input for a computer problem solving system. (1964).

[16] Guerrero-Vásquez, L. F., Landy-Rivera, D. X., Bravo-Torres, J. F., López-Nores, M., Castro-Serrano, R., & Vintimilla-Tapia, P. E. (2018, June). AVATAR: Contribution to Human-Computer interaction processes through the adaptation of semi-personalized virtual agents. In 2018 IEEE biennial congress of Argentina (Argencon) (pp. 1-4). IEEE.