Machine Unlearning-based Privacy Preserved Architecture with Large Language Model for Consumer Empowerment

Debajyoty Banik*, Debismita Dey†, Devashish Kumar Singh‡, Achyut Shankar,§
*School of Computer Science and Artificial Intelligence, SR University, India †School of computer Science, KIIT Bhubaneswar, India ‡School of computer Science, KIIT Bhubaneswar, India §Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, India
debajyoty.banik@gmail.com, deydebismita19575@gmail.com, devashishkumarsingh905@gmail.com

Abstract—In this paper, we present an innovative approach aimed at enhancing the functionalities of an empowered wristband device. The focus is on addressing the common issue of data vulnerability by introducing a novel wristband that seamlessly integrates voice technology and GPT (Generative Pre-trained Transformer). The primary objective is to facilitate the secure removal of personal data from the internet. The wristband caters to a diverse user base by simplifying the intricate process of permanent data removal. It provides a comprehensive solution for non-technical users, incorporating voice recognition, GPT integration, and ViT-TTS for efficient speech-to-text conversion. Utilizing ChatGPT v3.5 as our generative AI in the pipeline, we aim to imbue the wristband with advanced features for everyday use. Emphasizing a user-centered design approach, we underscore the paramount importance of privacy in the development of novel technologies. In an era of increasing connectivity, GuardianWrist emerges as a holistic, user-focused solution that places privacy at the forefront. The wristband not only offers cutting-edge capabilities but also prioritizes the protection of user data, thereby redefining the symbiotic relationship between technology and customer needs. The acknowledgment of privacy's critical role further sets GuardianWrist apart as a pioneering solution in our interconnected world.

Index Terms—GPT(Generative Pre-trained Transformer),AI(Artificial Inteligence),Generative AI, Privacy, consumer, LED screen, Microphone, Machine unlearning.

1 Introduction

In the ever-evolving domain of wearable technology, Wristbands are a new companion in the rapidly developing field of wearable technology, perfectly blending into our daily routines. In this work, we describe a novel wristband gadget that combines superior voice integration, machine learning, and GPT power, enabling human-device interactions- that have not been witnessed before. Unlike conventional GPT methods, our innovation focuses on developing a flexible and intuitive wristband gadget that allows bidirectional discussions for input and output and selectively forgets the sensitive information with the aid of the machine unlearning idea. We have carefully incorporated OpenAI's GPT technology and the novel idea of machine unlearning into our wristband device to accomplish this revolutionary breakthrough.

Conventional GPT models have demonstrated their abilities in natural language processing, or NLP, allowing for text-based interactions that are human-like. Our invention, however, goes beyond these limitations and ushers in a new era of multimodal interaction by giving wristband gadgets the ability to actively sense and react to user inputs conversationally. Other openAI's GPT technology, which powers our wristband device to understand the nuances of human speech, allows for dynamic, context-aware conversations, and makes a wide range of information and services accessible through the persuasive power of voice, is the central component of this integration.

Our wristband offers an embedded LED screen that enhances the user experience by providing animated images for an engaging engagement, surpassing traditional text-only responses. Additionally, a separate connector for an easy external keyboard connection has been incorporated, allowing users to switch between voice and text-based interactions.

Furthermore, the wristband's carefully designed, high-fidelity speakers strategically place themselves to overcome urban acoustic obstacles and guarantee that AI responses are delivered with crystal-clear clarity. Our study delves into the complexities of both hardware and software, presenting a smooth combination of voice integration, natural language processing, and interactive visual displays that raise the bar for wearable artificial intelligence [24].

The intelligent wearables era is ushered in by our wristband device, which enhances everyday life with applications ranging from virtual support to sophisticated communication. Through this technological journey, conversational AI becomes a practical reality by redefining the future of wearable devices through innovation and precision engineering. Machine unlearning is essential to our vision because it guarantees a customized, safe, and dynamically changing user experience. This fits in perfectly with the story of how cutting-edge technology empowers consumers.

Manuscript received March 7, 2024; revised March 7, 2024. Corresponding author: D. Banik (email: debajyoty.banik@gmail.com).

This novel feature adds a previously unseen level of intelligence to our device, ensuring that it not only learns and adapts, but also actively manages its knowledge base over time. By selectively discarding outdated or irrelevant information, machine unlearning addresses the dynamic nature of user interactions. This process is critical for keeping the AI system agile and relevant, preventing it from becoming burdened with obsolete data. Our wristband device takes a proactive approach to knowledge management by embracing machine unlearning, constantly refining its understanding of user preferences and evolving conversation contexts.

The use of machine learning is also consistent with our commitment to user privacy. Sensitive information is identified, selectively forgotten, and erased from the device's memory as users interact with it, reinforcing our commitment.

This paper will focus on three main aspects:

1)Generative AI: By adding voice functionality, we are improving our generative AI and streamlining customer-AI interactions. Once voice is integrated, all it takes to activate GPT is to speak to it in your chosen language. The usability and accessibility of our GPT model are improved when users can interact with it verbally. With this addition, users can now initiate GPT responses through natural conversation, which streamlines the user experience.

2) Privacy and Security: We have used the Machine Unlearning technique to protect user privacy. This method helps users erase information that is floating around the internet without their permission, in addition to helping them preserve their privacy going forward. Our project incorporates transparent user consent mechanisms to guarantee that individuals are informed about the data collection processes and have the autonomy to adjust privacy settings.

3) Consumer: This product has been designed in order to be easily portable and easy to use as well. The device is very much handy and could be used just by talking to it. The user-friendly interface, machine Unlearning for customization, privacy controls, intuitive voice interaction, language versatility, and the overall objective of democratizing access to cutting-edge AI technologies are all important components of our project's consumer-friendly design. All of these components work together to create a user-centered, flexible, and approachable experience.

1.1 Related work

Observing our surroundings leads us to the conclusion that artificial intelligence, or AI, has ingrained itself deeply into our daily lives. We will use the same AI in a more straightforward, portable, and user friendly way with the aid of this paper. We've also employed the idea of machine unlearning to give our device additional security. It will erase certain data on its own, adding even more privacy. While our main objective is to increase human-GPT connectivity, earlier research on this topic limited its scope by treating AI and GPTs as static and immobile. Furthermore, the extremely novel concept of Machine Unlearning has only been the subject of very little prior work. This is a unique project because it combines the GPT technique with the idea of machine learning in a wristwatch. We must ensure that our gadget is affordable, portable, and user-friendly. We gave the gadget a wristband shape in response to this. The user will always have this with them and can access it whenever needed.

TABLE 1
This table is created by examining the database so that the AI doesnot have any problem in identifying the object and could be used in understanding the image description.

Datasets	Dimensionility	size	#classes
MNIST	28 x 28	60000	10
Purchase	600	250000	2
SVHN	32 x 32 x 3	604833	10
CIFAR-100	32 x 32 x 3	60000	100
Imagenet	224 x 224 x 3	1281167	1000
Mini-Imagenet	224 x 224 x 3	128545	100

1.2 Task Formulation

We have divided the work up into distinct phases for now. The input stage will happen first, then the processing stage, and finally the output stage. We'll send the input of the chosen data that we want the system to forget after getting the output.

Initially, the user starts the interaction by providing voice input via the wristband device's built-in microphone. The user's speech is recorded by the microphone and transformed into an electrical audio signal. Additionally, it gives users the freedom to communicate with the device by text or voice, enabling a customized and adaptable dialogue experience.

The speech-to-text converter (ASR - Automatic Speech Recognition) in the SoC's (System on chips) hardware then processes the recorded audio signal. The user's question is accurately translated into text by the ASR, which also creates a textual representation of the spoken words [6]. The conversion step is omitted if the input is text-based. The SoC's integrated AI processing unit receives the converted text as input. Tensor processing units (TPUs) and other specialized hardware for AI computations, such as dedicated AI accelerators, are included in this unit.

The AI model processing needed for the conversation with GPT is effectively handled by the high-performance CPU architecture, such as the ARM Cortex-A series. The Chat GPT model is executed by the AI processing unit. It analyzes the textual input and produces a response using contextual and linguistic knowledge that it has been trained to understand.

Following its formulation in textual format by the Chat GPT model, the response is sent to the text-to-speech converter (TTS) for voice synthesis. The text is fed into the TTS engine, which then synthesizes it into speech that sounds human by adding the right emphasis, intonation, and pauses. The speaker system on the wristband receives the voice synthesized by the TTS engine.

The audio signal is routed to the speaker's amplifier circuit, which powers the speakers, by means of the hardware of the SoC. This outputs the AI-generated response as audible speech for the user to hear. Furthermore, the speech-to-text converter (ASR) can process the AI-generated speech output backwards. The AI's reply is represented textually by the ASR, which translates the spoken response back into

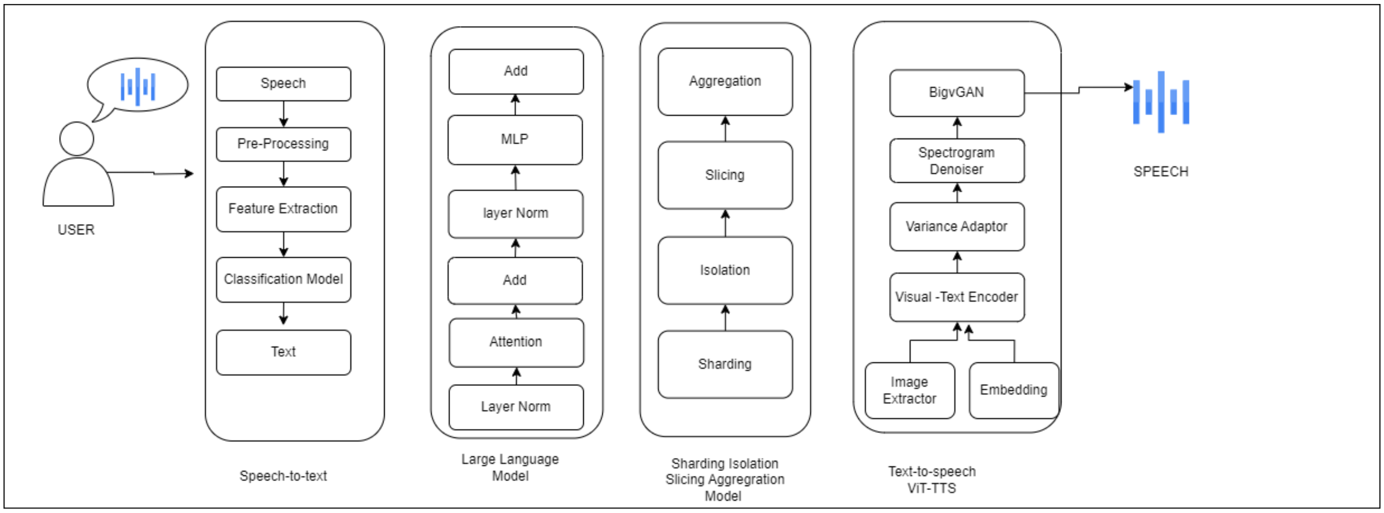

Fig. 1. This is the proposed model working the text will be taken in form of speech and the output will also be in form of speech so that the consumer find's it easy to access.

TABLE 2
In this table a deep neural network with different layer sizes and progressively more buried layers are used. Here, the P100 and T4 Nvidia GPUs, which have 12 and 16 GB of dedicated memory, respectively is used, for our studies. Our Intel Xeon Silver 4110 CPUs include 192GB of RAM and 8 cores apiece. of Ram. Ubuntu 18.04.2 LTS 64-bit is the base operating system.We Utilise Python 3.6 and CUDA 10.1 with PyTorch v1.3.1.

Datasets	Model Architecture
MNIST	2 conv. layers followed by 2 FC layers
Purchase	2 FC layers
SVHN	Wide ResNet-1-1
CIVAR-100	ResNet-50
Imagenet	ResNet-50
Mini-Imagenet	ResNet-50

textual format. The user's ability to seamlessly transition between text and voice input modes provides flexibility when interacting with the Chat GPT, which is powered by AI.

Our method adds a dynamic adaptability layer to the dialogue by incorporating the concept of machine unlearning after the initial output stage. Once the user receives the AI-generated response, a special chance presents itself. Users have the option to start a selective process, instructing the wristband device to ignore particular information from the ongoing discussion.

A state-of-the-art feature called machine unlearning enables the device to manage its knowledge base intelligently by removing information that it deems to be outdated or sensitive [6]. This is a user-initiated process that provides a dynamic and privacy-aware way to improve the device's comprehension and memory. Users can easily instruct the device to erase personal or confidential details from a conversation, for example, guaranteeing increased privacy and data security.

This cutting-edge feature is consistent with our dedication to privacy concerns and user-centric design. It makes sure the gadget actively participates in a process of controlled forgetting in addition to learning and adapting, which adds to a more unique, safe, and dynamic user experience. We further improve the versatility and user-friendliness of our wristband device by giving users some control over the device's knowledge retention through the introduction of machine unlearning.

1.3 Novelty

Our wristband redefines wearable technology, introducing groundbreaking features to redefine the user experience. Unlike traditional devices that store user data indefinitely, our wristband leverages the SISA model as machine unlearning, actively discarding sensitive information over time. This approach ensures that the device evolves with the user, dynamically adapting to preferences while maintaining a secure and customized experience.

Powered by language learning model For eg.- ChatGPT, our wristband enables seamless voice interactions, setting a new standard for convenience and accessibility. It sets a new standard for privacy-focused wearables, enhancing user interactions through seamless voice commands. The voice integration feature not only simplifies user engagement but also adds a dynamic layer to conversations. Users can effortlessly initiate tasks, clarify queries, and engage in dynamic conversations, creating a more natural and engaging wearable experience. To enhance inclusively on a global scale, our wristband boasts multilingual capabilities, allowing users to engage in their official languages. The integration of VIT-TTS facilitates seamless text-to-speech conversion, employing a sophisticated, multi-step machine learning model for accurate speech-to-text conversion.

In essence, our wristband is not just a device; it's a leap forward in wearable technology. By combining advanced privacy measures, state-of-the-art linguistic capabilities, and cutting-edge machine unlearning models, it reshapes the landscape of wearable devices, offering users an unprecedented and comprehensive experience.

2 Input From Users

In contrast to the other talking device, this one accepts text input sent via a keyboard or other typing device, in addition to voice commands. It is more useful when the inputs come from two distinct sources. Our suggested gadget is designed with the ability to speak any language in mind.

For taking the input following are the method's:

2.1 Input through Voice

As is well known, providing feedback has been crucial in today's environment for reducing task complexity [22]. To convert the given text into speech we will be using ViT-TTS. We will be recording a voice and converting it into text using the s.Janokar process [9]. However, the diverse properties of raw descriptions make it challenging to develop rules that translate voice instructions into captions effectively, and attempting to do so would result in a high discard rate similar to what was observed in CC3M (CC3M is a dataset of 3 million image and caption) [19].

We recommend using ChatGPT, a powerful big language model that OpenAI6 trained to perform this task automatically, to address the challenge of turning unprocessed descriptions into captions. Demonstrated to be trained using Reinforcement Learning from Human Feedback (RLHF), ChatGPT has garnered significant attention for its robust comprehension, reasoning, and conversational abilities and is highly skilled at generating responses to natural language prompts that resemble those of a human.

By creating prompts that take into account the characteristics of different data sources, ChatGPT can effectively filter out information that is unrelated to sounds and rewrite raw descriptions into audio caption-like phrases [12]. This approach can improve the quality of transformed captions and significantly reduce the rate at which raw descriptions are rejected. To leverage ChatGPT's in-context learning feature, a range of transformation examples are also provided in the prompts, and they are unique for each data source.

It has been demonstrated that ChatGPT can condense long sentences into high-level audio captions that are one sentence long, transform non-sentence descriptions (such as nouns and phrases) into sentences, and filter out information that is superfluous or unrelated to the sound. It can also differentiate between descriptions that apply to audio content and those that don't, generating "Failure" for the former [12].

Although ChatGPT has demonstrated potential in converting unformulated descriptions into captions, it might occasionally fall short of meeting our requirements. For example, ChatGPT can sometimes have problems extracting numbers, person names, and place names from raw descriptions. It's also possible that some descriptions that have nothing to do with the audio content get through.

2.2 Input through text

A number of methodical procedures are involved in the wristband device's integration of Bluetooth technology as a means of text input in order to improve user interaction and functionality. The process starts with selecting an appropriate Bluetooth module designed specifically for wearable.

TABLE 3
Class Performance and Feature Unlearning VAE on MNIST138 [2]

MNIST138(CLASS:1)			
Metric	Privacy	Utility	
	fratio(↓)	IS(↑)	FID(↓)
BEFORE	0.343(0.027)	2.053(0.029)	0.030(0.003)
GRAD.ASCNT	0.264(0.141)	2.029(0.018)	0.127(0.059)
MOONETAL. (2023)	0.344(0.019)	2.048(0.021)	0.031(0.002)

The selected module is then seamlessly integrated into the hardware architecture of the wristband. The Bluetooth stack is then implemented in the firmware of the device, controlling data exchange, secure pairing, and communication protocols with paired devices. A secure connection is established between the wristband and external Bluetooth-enabled devices through the user-initiated pairing process. Users can input text using their connected devices after a successful pairing, and the wristband's firmware quickly processes and gets ready for users to interact with the integrated Chat GPT model. An intuitive user interface directs users through the pairing process, subsequent text input, and the display of responses on the wristband's screen [7].

The Bluetooth text input feature is robust and reliable thanks to rigorous testing and validation procedures. Interestingly, this integration accommodates changing user preferences by providing a dynamic input option that increases the functionality of the wristband without compromising the wearables user-friendly and adaptable design. This Bluetooth-enabled text input feature is a monument to technological progress that has the potential to revolutionize user-device interactions and add to the changing field of wearable technology [7].

The entire system that uses Bluetooth technology to input text into the wristband device is an intricate combination of software and hardware parts that work together to revolutionize how users interact with devices. The key piece of hardware is the carefully selected Bluetooth module, which was chosen for its low power consumption and wearable compatibility. This module becomes an essential component of the wristband's communication ecosystem by integrating seamlessly with the internal architecture of the device.

This integrated system's foundation is built on extensive testing and validation. The Bluetooth-enabled text input mechanism's stability, dependability, and security are assessed through the simulation of demanding scenarios. The system's ability to accommodate a wide range of user preferences and scenarios is validated by real-world usage simulations and user feedback.

This entire system has far-reaching implications that seamlessly mesh with contemporary user expectations for flexible, convenient, and immersive interactions. This integration goes beyond simple functionality to encompass the core of both technological innovation and user-centric design. This Bluetooth-enabled text input system redefines communication paradigms and opens the door to a new era of user-device symbiosis, demonstrating the constantly changing nature of wearable technology.

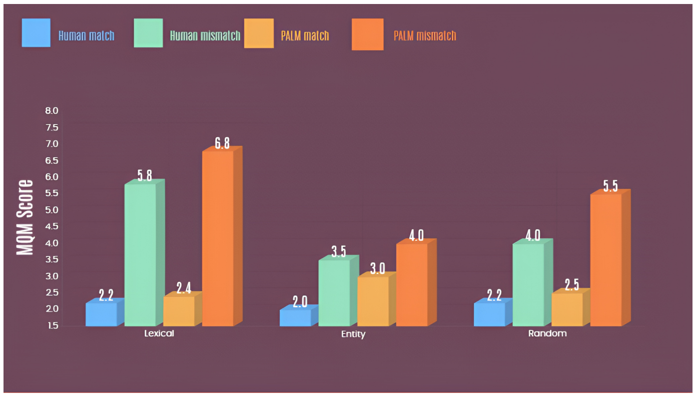

Fig. 2. MQM performance employing human and PaLM translations across dataset buckets. The region-matched scenario, in which raters from each area assess translations intended for their own region, is shown by thick bars. Green bars indicate human mismatch, whereas blue bars show the human matched scenario. Orange indicates PALM mismatch, and yellow indicates PALM match.

2.2.1 Voice Recognition System

The voice recognition system plays a key role in the project by allowing users to speak words into the wristband device to provide input. Its main function is to translate spoken language into text so that the user's commands, questions, and requests can be processed and understood by the device. These are the primary applications for your project's voice recognition technology. Users can speak naturally spoken language to the wristband device through the voice recognition system. Speak-to-me commands and questions reduce typing time and increase user convenience. Here are its specifications:

1) Algorithm: Speech recognition using advanced neural networks

2) Accuracy: in noisy environments, 90 percent accuracy or higher

3) Language Assistance: Multilingual identification

4) Real-time Processing: Transcribing almost instantly

2.2.2 Speech-to-Text converter

A sophisticated, multi-step machine learning model is employed for the conversion of speech to text, employing linguistic algorithms to dissect spoken words into auditory signals, subsequently translating these signals into text through the utilization of letters, a process commonly referred to as Unicode. The audio file's sounds are channeled into an analog-to-digital converter, which meticulously gauges the intricacies of the waves and filters them to meticulously extract the relevant sounds.

Upon the segmentation of sounds into intervals as fine as hundredths or thousandths of seconds, the phonemes are identified. In the context of language, a phoneme represents a distinct sound unit crucial for distinguishing between words. As an illustration, the English language encompasses approximately 40 phonemes. Subsequently, a mathematical model is deployed to subject the phonemes through a network, facilitating a comparison with well-established words, sentences, and phrases.

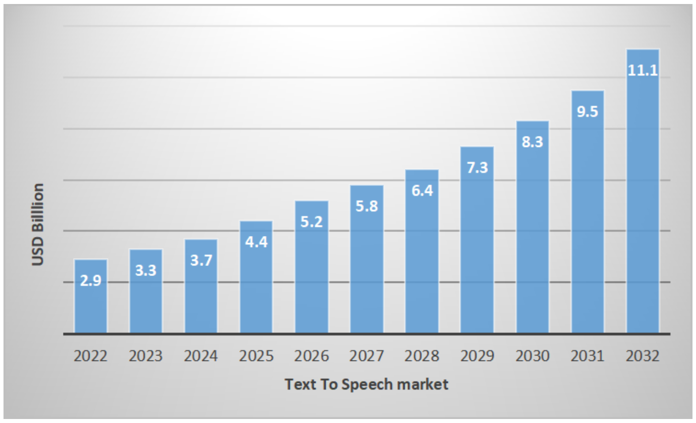

Fig. 3. Here, is a graph showing the increase of user's and profit in the field of text to speech business. Here, data also predict the amount of income in future of TTS business.

Following this intricate comparison process, contingent on the most probable interpretation of the audio, the resultant text is then presented either in textual form or as a computer-generated command. This advanced approach combines linguistic and mathematical methodologies, contributing to a nuanced and precise transformation of spoken language into textual representation [8].

2.2.3 Text-to-Speech Converter

In our project, the wristband device's capacity to generate audio output in response to AI-generated text responses underscores the indispensability of the Text-to-Speech (TTS) converter in enhancing the user experience. In the growing market of industrial businesses TTS has been also a great source of income. See graph 3 For this purpose, we have implemented the ViT-TTS, chosen for its versatile capability to handle not only text but also images. Leveraging a dedicated Image Extractor based on ResNet18, the ViT-TTS extracts detailed information from panoramic images before feeding them into the visual-text encoder.

To derive hidden sequences, a variant of the transformer is employed, receiving both phoneme embedding and image characteristics. To elaborate, the phoneme is transmitted through this transformer, and the resultant details are then routed to a variance adaptor. This adaptor, structured as a 2-layer 1D-convolutions network with ReLU activation, serves as the model architecture for both duration and pitch prediction. This entails incorporating a dropout layer, layer normalization, and an additional linear layer to represent concealed states in the output sequence.

Upon processing by the Variance Adaptor, the data is forwarded to the Spectrogram Denoiser. Operating with xt as input, the spectrogram denoiser anticipates the additions made during the diffusion process, conditioned on the step embedding Et and the encoder's output. This advanced architecture contributes to a sophisticated and seamless integration of various components, ultimately enhancing the overall functionality of the wristband device [15].

2.2.4 Bluetooth Module

The Bluetooth module plays a crucial role in our project by facilitating wireless communication between the wristband device and other devices. Its main purpose is to make text input easier for users of Bluetooth-enabled devices, giving them another way to interact with the wristband. These are its specifications:

1) Bluetooth version: 5.0 or above

2) Range: in an open area, at least thirty meters

3) Minimum data rate of 2 Mbps

4) Low Energy: Bluetooth Low Energy (BLE) compatibility

5) Profiles: HID (Human Interface Device) profile support is provided.

2.2.5 User Interface and Buttons

The user interface of the wristband device is crucial in enabling smooth and simple interactions between users and the device. It includes both audiovisual components that let users interact with the device and get feedback in an efficient manner.

1) Display: Touch-capable TFT or OLED, if appropriate

2) Buttons: Mechanical or capacitive buttons for communication

Third, interface design: a graphical user interface (GUI) that is intuitive

4) Feedback: Vibration in response to interactions and button presses

2.3 Working

The device will first take speech inputs from user directly, and convert it into text. A multi-step, intricate machine learning model is used to convert speech to text.It uses linguistic algorithms to separate spoken words into auditory signals and then converts those signals into text using letters. This process is also called as unicode [8]. After getting converted into text it is passed on to LLM(large Learning Model) the data is searched for the possible answer. Then the result of the searched query that was received goes to SISA model for machine unlearning process. Their it gets checked for some personal or private data and if found something gets instantly unlearn. This process is very necessary for maintaining the privacy of consumer. And the result is further send to ViT-TTS model which is a Text-to-Speech converter model. It converts the text signal into voice signals as an output. The reason behind using ViT-TTS was, with the help of Image extractor, it can also describe the image presented by the LLM [15].

2.3.1 Testing Phase

When the user chooses to enter a request or command into the wristband device, the interaction starts. Three main methods of input are available to the user: voice, keyboard, or Bluetooth. Voice input, voice recognition, keyboard input, and Bluetooth input are examples of user input processes. We process the received input once it has been provided. The transcribed text input is sent to the integrated Chat GPT model regardless of the input method (voice, keyboard, or Bluetooth). Text input is processed by the System-on-Chip (SoC), which houses the CPU and AI processing capabilities.

Voice Input The built-in microphones record the user's spoken words if voice input is chosen. The intended question or command from the user is extracted by the voice recognition system after processing the audio input and converting it into text format.

Keyboard Input If the user would rather use a keyboard, they can use the supplied jack to connect a keyboard to the device. Text is typed into the device directly.

2.3.2 Large Language Model

For making it more realistic we have used ChatGPT as one of the module for connecting one pipeline to another. The ChatGPT v3.5 model responds to the user's question or command with a logical and contextually appropriate response. It is now possible to reply to the user with the text response. It uses the datasets like common crawl which is one of the world's largest datsets available. It has billions and billions of web-pages in it. It is also pre-trained on datasets like dialogue datasets which was created by the trainers of AI who played both side of a conversation. It also uses InstructGPT datasets and comparision datasets for performing different tasks. For measuring the accuracy of the ChatGPT a test was done and with a mean score of 4.8 (between mainly and nearly totally accurate), the median accuracy score for all 284 questions was 5.5 (between almost fully and completely correct). With a mean score of 2.5, the median completeness score was 3 (complete and thorough). The median accuracy scores for the easy, medium, and hard questions were 6, 5.5, and 5 (mean 5.0, 4.7, and 4.6; p=0.05). For all binary and descriptive questions, accuracy ratings were comparable (median 6 vs. 5; mean 4.9 vs. 4.7; p = 0.07). A significant improvement was seen when 34 out of the 36 items with scores of 1-2 were re-posed and re-graded 8–17 days later (median 2 vs. 4; p less than 0.01).

AI Model Interaction The text is sent to the integrated Chat GPT model once the user input has been transcribed. The text input is processed by the System-on-Chip (SoC), which also serves as the CPU and AI processing hub. It also starts the AI model interaction.

Contextual Understanding Using its substantial training, the Chat GPT model deciphers the context and meaning of the user's command or question. In order to produce logical and pertinent answers, it analyzes the input text to extract intent, keywords, and context.

Response Generation The AI model uses its language generation capabilities to produce a response that specifically answers the user's question or command. The input context, past exchanges, and a wealth of knowledge are all taken into consideration when crafting the response.

Natural Language Generation To generate responses that resemble the linguistic patterns and tones of humans, the AI model makes use of natural language generation techniques. It guarantees that the answer is coherent, fluid, and appropriate for the context.

Multimodal Integration The AI model may instruct the LED screen to display pertinent information if the response includes visual elements (such as a URL).

Response Ready The text version of the AI-generated response is now prepared and ready to be sent back to the user.

2.3.3 Response Delivery

The Text-to-Speech (TTS) converter is used to concurrently translate the text response produced by the AI model into spoken language. The response will sound expressive and natural thanks to the TTS engine. Here we have used ViT-TTS [15] for Text-To-Speech conversion. We have used ViT-TTS because it has a transformers for diffusion that are scalable. ViT-TTS creates highly perceptible audio by combining the visual information with the phoneme sequence.

2.3.4 Auditory and Visual Output

The built-in speakers deliver the generated voice response. The AI-generated response is audible to users, fostering an interactive and conversational experience. For users who prefer visual feedback, the AI-generated text response is simultaneously displayed on the LED screen. An additional means of communication is provided by the ability for users to read the response that is shown on the screen.

2.3.5 Multimodal Interaction and User Interaction Loop

The wristband device provides a multimodal experience by integrating voice and visual outputs (LED screen), accommodating a range of user preferences.

Users can keep up the interaction loop by giving more input, posing follow-up queries, or having lively discussions with the AI model.

2.3.6 System Control and Management

The motherboard ensures smooth communication between components and devices by managing and coordinating the entire interaction process, which is integrated with the SoC.

3 Output

We have taken the data entry and processed it, so far so good. However, the issue with reproducing the response is that it needs to be displayed both through speakers and on screen. Text input is converted to digital audio and spoken after it has been evaluated, processed, and interpreted. Text-to-speech is the term used for this process. The block diagram for TTS is shown in Figure:1.

3.1 Speech synthesis:

The artificial production of human speech is known as speech synthesis. One kind of computer system that could be used for this is a voice synthesizer. It can be incorporated into software or hardware goods. Text-to-speech (TTS) systems convert written text in the normal language into speech; alternative techniques convert symbolic linguistic representations—like phonetic transcriptions—into speech. Reversing the process is speech recognition. The following are the techniques for speech synthesis [19] [16]:

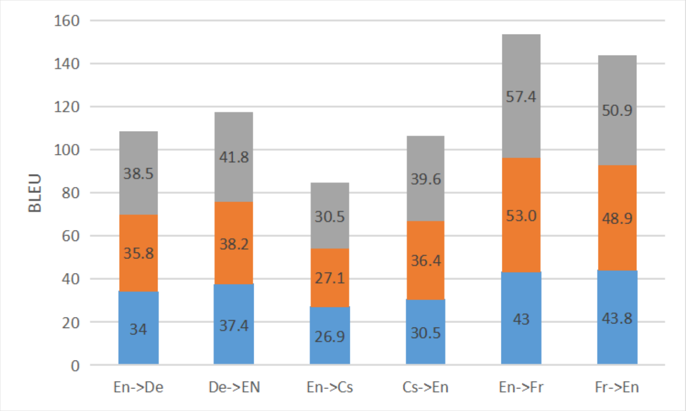

Fig. 4. Here, We compare the BLEU scores for models trained and tested on the Multi30k dataset for each language combination: 1)Grey colour is for M to M + plus monolingual data, 2) orange colour is for bilingual + attBridge, 3) Zero short for many to many model. Also, The machine translation quality of text from one natural language to another is assessed by using the BLEU method.

3.2 LANGUAGE TRANSLATION

We speak a variety of languages in India. According to the 2001 Census, more than 10,000 people spoke 122 languages, and 30 languages were spoken by more than a million native speakers [10]. As a result, it is critical to have programs and procedures in place that can translate text between languages while maintaining the message's integrity. Machine translation (MT) is the study of translation from one language to another using machine translation systems. It is a branch of artificial intelligence and natural language processing. The human translation process can be summarized as decoding the meaning of the source text and re-encoding it in the target language. A few machine translation models are discussed below [19] [1]:

3.2.1 Rule Based Machine Translation (RBMT):

The translation is built using morphological, syntactic, and semantic analysis of the source and destination languages. A system like this is made up of several rules [18]: Grammar rules are basically the following: software programs to understand and process those rules; bilingual or multilingual lexicon dictionaries for word lookup during translation while the software enables the effective and efficient interaction of components; and language analysis in terms of grammatical structures (syntax, semantic, morphology, part of speech tagging, and orthographic features). For getting more accuracy we have also checked for the BLEU score of some language like English(En), German(De), Czech(Cs) and French(Fr)[see figure 3] [28]. There are three types of rule-based models [13]:

- Straightforward: dictionary-based.
- Transfer: Before being translated to an intermediate representation, each SL input text is processed using lexicons and structural analysis.
- Interlingual: The source language is translated into an intermediate language unrelated to the target languages during the translation process.

TABLE 4
This table show us the result of data before and after getting unearned

Domain	Status	Text
File.c	Before Unlearning	James Gurney This daily weblog by Dinotopia creator James Gurney is for illustrators, plein-air painters, sketchers, comic artists, animators, art students, and writers. You'll find practical studio tips, insights into the making of the Dinotopia books, and first-hand reports from art schools and museums. CG Art Contact or by email:gurneyjourney (at) gmail.com Sorry, I can't give personal art advice or portfolio reviews. If you can, it's best to ask art questions in the blog comments. Permissions All images and text are copyright 2015 James Gurney and/or their respective owners. Dinotopia is a registered trademark of James Gurney. For use of text or images in traditional print media or for any commercial licensing rights, please email me for permission. However, you can quote images or text without;
	After Unlearning	James Gurney This daily weblog by Dinotopia creator James Gurney is for illustrators, plein-air painters, sketchers, comic artists, animators, art students, and writers. You'll find practical studio tips, insights into the making of the Dinotopia books, and first-hand reports from art schools and museums. CG Art Contact or by email:gurneyjourney (at) gmail.com I've been working on a CG art project for a while now, and I've been working on it for a while now. I've been working on it for a while now, and I've been working on it for a while now. I've been working on it for a while now, and I've been working on it for a while now. I've been working on a CG art project for a while now, and I've been working on it for a while

TABLE 5
Unlearning results on Reverb 45k datasets [14]

Macro F1	Raw		re-trained		Unlearn	
	Forget	Text	Forget	text	Forget	Text
TransE	0.9379	0.7481	0.9288	0.7182	0.6291	0.7415
Hole	0.9379	0.7481	0.9288	0.7023	0.6711	0.7462
\ Pair F1	Raw		Re-trained		Unlearn	
	Forget	Text	Forget	Text	Forget	Text
TransE	0.8184	0.7938	0.8167	0.7892	0.1578	0.7937
Hole	0.8184	0.7935	0.8167	0.7882	0.1516	0.7902
Micro F1	Raw		Re-trained		Unlearn	
	Forget	Text	Forget	Text	Forget	Text
TransE	0.9493	0.8316	0.9452	0.8215	0.5282	0.8313
Hole	0.9493	0.8316	0.9452	0.8118	0.5358	0.8285
AverageF1	Raw		Re-trained		Unlearn	
	Forget	Text	Forget	Text	Forget	Text
TransE	0.9019	0.7917	0.8969	0.7784	0.4384	0.7888
Hole	0.9019	0.7911	0.8969	0.7668	0.4528	0.7883

3.2.2 *Statistical machine translation (SMT):*

It stands out due to the utilization of machine learning methods. SMT is a data-driven approach that treats translation as a mathematical inference problem and uses parallel aligned corpora. In this instance, each sentence in the target language is a probabilistic translation from the source language. Likelihood increases translation accuracy, and vice versa. An SMT architecture is typically composed of [23]:

- Using a language model, one can ascertain the probability of target language usage.
- A translation model that computes, given input in the source language, the conditional probability of output in the target language
- To produce the best translation possible, the decoder model maximizes the two probabilities mentioned above.

4 MACHINE UNLEARNING

In artificial intelligence, the deliberate and targeted removal of particular data or information from a machine learning model's knowledge base is referred to as ”machine unlearning.” Machine unlearning enables the model to ignore specific information, causing it to ”unlearn” or forget specific aspects of its training data, in contrast to traditional machine learning, which emphasizes continuous learning and adaptation.

Essentially, machine unlearning aims to give models a way to actively control their knowledge and flexibility. This can be crucial in situations where keeping certain data around could jeopardize security, privacy, or the long-term applicability of the model's conclusions.

Our project employs a sophisticated mechanism called machine unlearning, which is intended to provide users with a private and dynamic way to manage the device's knowledge base. Users can start the machine unlearning process, instructing the wristband device to forget particular data points from the ongoing conversation, once the Chat GPT model has produced a response. See table[1.2].

We will be using SISA model for working in machine unlearning. Similar to contemporary SISA (Shrading Isolation Slicing Aggregration) training, SISA training duplicates the model being learnt several times, giving each replica a disjoint shard (or portion) of the dataset methods for dispersed training. We speak about each as a model of constituents. However, SISA training differs from existing approaches in that there is no information flow on the dissemination or sharing of incremental model updates amongst component models.

For instance, gradients computed on each constituent are not shared if each constituent model is a DNN trained with stochastic gradient descent among several components; every component is skilled in seclusion. This guarantees that a shard's effect (and the data elements comprising it) are limited to the model being taught to use it. Slices are further divided into each shard, wherein each component model is gradually taught (and statefully and iteratively) with a growing number of portions. Each component receives the test point at inference and all of the participants' replies are combined, much like ML ensembles as an example [17].

4.1 Sharding

We are able to disperse the training cost by breaking the data up into disjoint parts and training a component model on each smaller fragment. While our approach naturally takes into account the parallelism between shards, we did not take this into consideration during our research and testing. The goal of this is to equitably retrain a model from the beginning, a process that might be sped up by splitting up the computations among several computers. We make the assumption that we are unaware of the likelihood that any given point may be forgotten for the duration of this section. Under such circumstances. It is possible to evenly divide a dataset D into S shards, such as it follows that k[S]Dk = D and k[S]Dk = . For every Dk shard, a All of the data is used to train the model (referred to as Mk). accessible in DK. The case in which the S is aware of the distribution of unlearning requests [4]. See table[1.2].

Note that there is an equal chance for user u's data point du to lie in each of the S shards. Furthermore, whether or

TABLE 6
The following table shows the type of users and how they interact with the server. This table helps us in understanding how to know more about the user which will be using our device for enhancing the performance of the device [26]

	Privacy Enthusiasts	Normal user	Server
Phase I	Send partially poisoned data	Send normal data	Receive data
phase II	Send data delete request	Send data delete request	Receive data delete request
phase III	verify deleted data	do nothing	do nothing

TABLE 7
Performance of the NP canonicalization task in ReVerb45K.

	MacroF1	MicroF1	PairF1	AverageF1
Morph Norm	0.627	0.558	0.334	0.506
Text Similarity	0.625	0.566	0.394	0.528
IDF	0.603	0.551	0.338	0.497
Attribute Overlap	0.621	0.558	0.342	0.507
CESI	0.640	0.855	0.842	0.779
JOCL	0.537	0.854	0.823	0.738
MulCanon	0.751	0.833	0.795	0.793

not each du is a member of many shards can be one of the training's criteria. To keep things simple, we'll suppose that each du only participates in one shard as doing so optimises savings, throughout the unlearning phase. We go into further detail on this in § IX. If the user wants to unlearn du, the service provider must first identify the dataset (and shard) that du is situated, known as Du, and then create a new model M0 by retraining the relevant model on Du from scratch. You Comparatively, retraining would be the baseline. The model on D from the beginning. Given that $|D| > |du|$,

the timing necessary for what will subsequently be called retraining time) in the baseline is far higher than what we've suggested. The idea offers an anticipated acceleration of S× [27].

4.1.1 Time Analysis Of Sharding

Step 1: First we will assume that at each step the likelihood that an unlearning request will have an impact on that particular shard is about equal to 1/s.

Step 2:There is always a shard whose size is N/S that is impacted by the initial request. If the second request does not impact the same shard, its response can be N/S with probability

$$(1 - \frac{1}{S}) \tag{1}$$

$$or \left(\frac{N}{S} - 1 \right) \tag{2}$$

with probability (1/S), (N/S), $\left(\frac{N}{S} - 1\right), or$

$$\left(\frac{N}{S} - 2 \right) \tag{3}$$

might be selected for the third request.

step 3: We define the event Ei,j as the ith request received landing on a shard s containing (N/S j) points, where j$\epsilon \{0, 1, 2...i - 1\}$

This allows us to describe this behaviour. The related expense is (N/S j-1)

step 4: The likelihood of Ei,j in the context of a configuration of the j requests; that is, the subset of the i 1 requests that arrived on

$$\left(\frac{1}{S}\right)^{j} \left(1 - \frac{1}{S}\right)^{i-1-j} \tag{4}$$

The product's first term indicates that shard s was impacted j times, while its second term indicates that an additional shard—but not s—was i-1-j times. Whereas, their is i-1j [see equation 4].

configurations that the j queries that arrived on shard s may have had. Consequently, the likelihood of Ei,j is

$$i - 1j \left(\frac{1}{S}\right)^{j} \left(1 - \frac{1}{S}\right)^{i-1-j} \tag{5}$$

4.2 Isolation

Note that each shard is trained separately in accordance with the idea that was previously explained. If we don't update together, we might deteriorate the capacity of the entire model to be generalised (including all components). Nevertheless, we show that for suitable selections regrading the quantity of shards, this doesn't happen in practise for specific kinds of educational assignments. Being alone is a delicate yet strong construction that allows us to provide concrete, observable, demonstrable, and intuitive promises for unlearning.

4.3 Slicing

We get additional time savings by further separating the data set aside for each model (i.e., each shard) and gradually adjusting (and storing) a model's parameter state, for only one request to unlearn. In particular, $\bigcap i\epsilon \left[R\right] DK, i = 0$

and $\bigcup i\epsilon \left[R\right], DK, i = DK$

are created by further evenly partitioning each shard's data Dk into R disjoint slices.To get Mk, we train for e epochs in the manner described below:

1) Step 1: Use just Dk,1 for e1 epochs to train the model using random initialization. The resultant model will be known as Mk,1. Keep the settings in their current condition connected to this model.

2) In step 2, the model Mk,1 is trained using D$_{k,1} \bigcup D_{k,2}$

for epochs of e2. The final model will be known as Mk,2. Save the parameter state.

3) In step R, the model is trained M$_{k,R-1} using \bigcup D_{K,i}$

for eR epochs. The final model that is produced is denoted by Mk,R = Mk. Preserve the status of the parameter.

4.4 Aggregation

An overall forecast can be obtained at inference time by combining predictions from several constituent models. Two major aspects impact the aggregation approach selection in SISA training:

1) It is closely related to the process of dividing data into shards since the aim of aggregation is to maximise the combined prediction performance of the component models.

2) The training data shouldn't be included in the aggregation method (otherwise, in some situations, the aggregation technique would need to be relearned).

5 Model Specification

The project's goals are stated clearly from the outset. The principal objective is to augment the functionalities of a wristband device through the smooth integration of advanced voice integration with Chat GPT, consequently permitting interactive two-way communication. The addition of an LED screen is also intended to increase user engagement by providing visual feedback. Important choices are made about which hardware parts to use. A suitable LED screen that can produce bright visuals is selected. Additionally, in order to facilitate AI processing and communication functionalities, speakers, microphones, a motherboard, and a specialized System-on-Chip (SoC) are integrated. LED screens, microphones, speakers, motherboards, system-on-chips (SoC), Bluetooth modules, user interfaces, text-to-speech (TTS) converters, and buttons are among the main devices that have been used [3]. We have made a pipeline and for making it more realistic and more easy to use we have used ChatGPT v3.5 and ViT-TTS for text to speech conversion and vice versa.

5.0.1 LED Screen

The wristband device's LED (Light Emitting Diode) screen functions as a visual interface that shows the user prompts, responses, and information. The LED screen that was used has the following specs:

1) Type: TFT (thin-film transistor) or OLED (organic light emitting diode)

2) Resolution: Full HD (1920x1080).

3) Dimensions: 1.5 inches across or more

4) True-color color depth, with 16.7 million colors

5) Refresh Rate: 60 Hz minimum

6) Touch Capability: Capacitive touch panel available as an option.

5.0.2 Microphone

The wristband device's microphones perform a number of crucial tasks that improve user experience and overall functionality. Voice input, hands-free interaction, multimodal input, accessibility, contextual queries, and other features are among its capabilities. Using its specifications, our model is:

Micro-Electro-Mechanical Systems (MEMS) microphones are the first type.

2) Sensitivity: -40dBV/Pa

3) Response to Frequency: 20Hz – 20kHz

4) A signal-to-noise ratio (SNR) of at least 65 dB

5) Dual microphones are used for noise cancellation.

5.0.3 Speakers

The wristband device's speakers improve user experience and overall functionality by fulfilling a number of vital purposes. Providing audio output to the user is the speakers' main purpose. They give the gadget a way to speak to the user audibly while displaying data, reactions, and prompts produced by the combined AI model. The specifications that we employed for our project are:

1) Type: Piezoelectric transducers or dynamic speakers

2) Maximum output power per speaker is 0.5W.

3) Response to Frequency: 20Hz – 20kHz

4) Resistance: 8 or 16 ohms

5) Sound Quality: The audio output is balanced and clear.

5.0.4 Motherboard

The wristband device's motherboard acts as a central hub, connecting and enabling communication between different software processes and hardware components. Its functions are essential to the device's overall performance and functionality. All of the necessary hardware is integrated onto the motherboard. It offers the structural framework required for these elements to function as a unit. Among the requirements are:

Processor: ARM Cortex-based SoC with high performance (e.g., Snapdragon, Exynos, MediaTek)

2) Processor Cores: at least four cores

3) Maximum clock speed of 1.5 GHz

4) Memory: 2 GB minimum

5) Storage: a minimum of 16GB eMMC

6) Wi-Fi and Bluetooth connectivity (for internal communication)

7) Ports: Bluetooth antenna connector, audio jack, and USB

5.0.5 System-on-Chip(SoC)

In the project, the System-on-Chip (SoC) is critical in providing the computational power and capabilities required to integrate Chat GPT with advanced voice integration. A high-performance CPU, such as an ARM Cortex-A series processor, is housed in the SoC. The CPU's primary function is to execute complex computations, process user inputs, manage the AI model, and oversee overall device operation. Its specifications are as follows:

1) ARM Cortex-A series architecture

2) Cores: at least quad-core

3) Clock frequency of 1.5 GHz or higher

4) GPU: Graphics processing unit (GPU) integrated for visual processing

5) AI Processing Unit: AI task hardware acceleration.

6) RAM: 1GB or greater integrated RAM

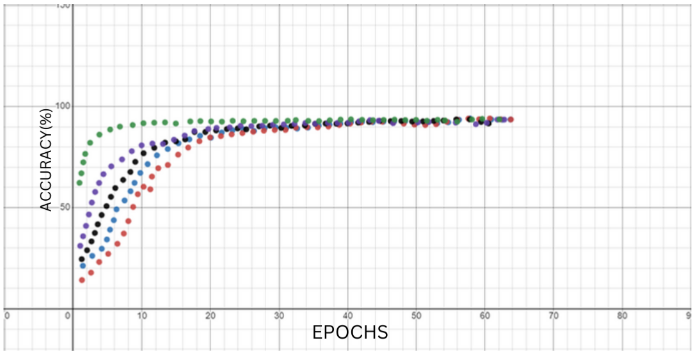

Fig. 5. This is a graph of Accuracy vs Epochs of SVHN. Here the doted line represents number of slices. The green doted line represent 1 number of slice. Whereas, purple represent 4 no of slices, black represents 8, blue 16 number of slices and red 64 no of slices

6 Datasets

In the following table we have used the MNIST(Modified National Institute of Standards and Technology database) which is a large database for handwritten numbers used in training various image processing system. Here we have used a sample size of 60,000 from MNIST. In the second column we have used the LSTM(Long Short Term Memory) recurrent neural network to predict the shopping list of online shopper's [21]. To maintain accuracy, we have used a large amount of data i.e., 250,000. SVHN(Street View House Number) is a technique used for reading house number and street number from the pictures itself. The main motive to use it is for understanding the image, making it easier to locate the given data, which must be forgotten [20]. CIFAR(Canadian Institute for Advanced Research)-100 is a database that has 60,000 32x32 color images in 10 different classes. These images are used to detect objects from given images. These pictures include airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks [11].The proliferation of image data on the Internet may lead to the development of increasingly complex and reliable models and algorithms for the indexing, retrieval, organisation, and interaction of images and multimedia data [5]. see table[3.1]

Table 2.2 contains details about the datasets we utilized. Please take note that we employ an approach akin to Shokri et al [25]. For the Purchase dataset. Reza Shokri and team have investigated how machine learning models leak the data they have been trained on. Using the category feature, we selected the top 600 most purchased goods to curate the Purchase dataset. For Mini-Imagenet, we build a dataset for supervised classification, not few-shot classification, by following Vinyals et al.'s methodology. In Oriol Vinyals we show a picture ans ask the gpt to write a caption for the image which we showed. see table [4]

7 Conclusion

Our project's seamless integration of machine unlearning into the adaptable GPT model marks a significant advancement in user-friendly AI applications. Our investigation of cutting-edge strategies like speech-to-text (STT) and text-to-speech (TTS) has revealed the enormous potential and wide range of uses for these technologies. Our project builds on this framework and advances it further by incorporating machine unlearning into our model with ease. This integration, which represents a turning point in the development of user-friendly AI applications, was informed by extensive research. see table[4] Our model, which puts an emphasis on user-friendliness, is a big step toward democratizing the use of GPT by serving both non-programmers and technical experts. Incorporating various speech technologies improves GPT's usability for a wider audience while also broadening its reach. Moreover, machine unlearning creates a dynamic knowledge base that guarantees a customized, safe, and changing user experience. This calculated move is in line with our goal of democratizing cutting-edge technology and represents a significant advancement in the field of accessible and approachable artificial intelligence applications. As we move forward, generative AI will continue to play a crucial part in enhancing user experience. Our dedication to privacy, consumer-centric design, and responsible technology use emphasizes how crucial it is to strike a balance between innovation and morality. Our goal is to make sure that our technological advancements not only push boundaries but also contribute to a positive and inclusive technological landscape by actively addressing these aspects. Our project is proof of what can be achieved when state-of-the-art technology and a dedication to privacy, accessibility, and user needs come together. We see a time when sophisticated AI applications are not only accessible but also powerful for people from all backgrounds thanks to our combined efforts.

References

[1] Alawneh, M.F., Sembok, T.M.: Rule-based and example-based machine translation from english to arabic. In: 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications. pp. 343–347. IEEE (2011)

[2] Bae, S., Kim, S., Jung, H., Lim, W.: Gradient surgery for one-shot unlearning on generative model. arXiv preprint arXiv:2307.04550 (2023)

[3] Bansal, D., Turk, N., Mendiratta, S.: Automatic speech recognition by cuckoo search optimization based artificial neural network classifier. In: 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI). pp. 29–34. IEEE (2015)

[4] De Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., Van Compernolle, D.: Template-based continuous speech recognition. IEEE Transactions on Audio, Speech, and Language Processing **15**(4), 1377–1390 (2007)

[5] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

[6] Ghadage, Y.H., Shelke, S.D.: Speech to text conversion for multilingual languages. In: 2016 International Conference on Communication and Signal Processing (ICCSP). pp. 0236–0240 (2016). https://doi.org/10.1109/ICCSP.2016.7754130

[7] Jadhav, A., Patil, A.: Android speech to text converter for sms application. IOSR Journal of Engineering **2**(3), 420–423 (2012)

[8] Jamadar, M.R., Pawar, M., Karke, P., Sonar, A., Zungure, Y., Sharavagi, S.: Automatic speech recognition: Speech to text converter

[9] Janokar, S., Ratnaparkhi, S., Rathi, M., Rathod, A.: Text-to-speech and speech-to-text converter—voice assistant. In: Inventive Systems and Control: Proceedings of ICISC 2023, pp. 653–664. Springer (2023)

[10] Kinjo, T., Funaki, K.: On hmm speech recognition based on complex speech analysis. In: IECON 2006-32nd Annual Conference on IEEE Industrial Electronics. pp. 3477–3480. IEEE (2006)

[11] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

[12] Kurzekar, P.K., Deshmukh, R.R., Waghmare, V.B., Shrishrimal, P.P.: A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology **3**(12), 18006–18016 (2014)

[13] Lawrence, R.: Fundamentals of speech recognition. AT&T (1993)

[14] Liu, B., Hou, S., Zeng, W., Zhao, X., Liu, S., Pan, L.: Open knowledge base canonicalization with multi-task unlearning. arXiv preprint arXiv:2310.16419 (2023)

[15] Liu, H., Huang, R., Lin, X., Xu, W., Zheng, M., Chen, H., He, J., Zhao, Z.: Vit-tts: Visual text-to-speech with scalable diffusion transformer. arXiv preprint arXiv:2305.12708 (2023)

[16] Mache, S.R., Baheti, M.R., Mahender, C.N.: Review on text-to-speech synthesizer. International Journal of Advanced Research in Computer and Communication Engineering **4**(8), 54–59 (2015)

[17] Malhotra, K., Khosla, A.: Automatic identification of gender & accent in spoken hindi utterances with regional indian accents. In: 2008 IEEE Spoken Language Technology Workshop. pp. 309–312. IEEE (2008)

[18] Marsal, P.P., Pol, S., Hagen, A., Bourlard, H., Nadeu, C.: Comparison and combination of rasta-plp and ff features in a hybrid hmm/mlp speech recognition system. In: INTERSPEECH (2002)

[19] Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M.D., Zou, Y., Wang, W.: Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. arXiv preprint arXiv:2303.17395 (2023)

[20] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)

[21] Sakar, C.O., Polat, S.O., Katircioglu, M., Kastro, Y.: Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks. Neural Computing and Applications **31**, 6893–6908 (2019)

[22] Saksamudre, S.K., Shrishrimal, P., Deshmukh, R.: A review on different approaches for speech recognition system. International Journal of Computer Applications **115**(22) (2015)

[23] Seide, F., Li, G., Yu, D.: Conversational speech transcription using context-dependent deep neural networks. In: Twelfth annual conference of the international speech communication association (2011)

[24] Shafeeg, A., Shazhaev, I., Mihaylov, D., Tularov, A., Shazhaev, I.: Voice assistant integrated with chat gpt. Indonesian Journal of Computer Science **12**(1) (2023)

[25] Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 3–18 (2017). https://doi.org/10.1109/SP.2017.41

[26] Sommer, D.M., Song, L., Wagh, S., Mittal, P.: Athena: Probabilistic verification of machine unlearning. Proc. Privacy Enhancing Technol **3**, 268–290 (2022)

[27] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden markov models. Proceedings of the IEEE **101**(5), 1234–1252 (2013)

[28] Vazquez, R., Raganato, A., Creutz, M., Tiedemann, J.: A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. Computational Linguistics **46**, 1–53 (03 2020). https://doi.org/10.1162/COLI$_{a0}$0377