Enhancing E-Learning with AI: Transforming Video Content into Interactive Educational Experiences

Dr. Kalyanaraman P
School of Computer Science and Engineering
Vellore Institute of Technology-vellore
Vellore, India
pkalyanaraman@vit.ac.in

Aravind K
School of Computer Science and Engineering
Vellore Institute of Technology-vellore
Vellore, India
aravind.karunakaran2023@vitstudent.ac.in

Vijay Nanda Kumaran
School of Computer Science and Engineering
Vellore Institute of Technology-vellore
Vellore, India
vijaynanda.kumaran2023@vitstudent.ac.in

Arockia Sachin A
School of Computer Science and Engineering
Vellore Institute of Technology-vellore
Vellore, India
arockiasachin.a2023@vitstudent.ac.in

***Abstract*—Despite e-learning is experiencing a rapid increase in popularity, it continues to encounter notable obstacles including student engagement, technology accessibility, and the quality of educational content. These issues frequently result in lower retention rates and reduced learner satisfaction. This study presents a novel teaching tool driven by artificial intelligence (AI) that aims to overcome the challenges associated with passive video content by enhancing it into a dynamic, interactive, and personalized learning experience. OpenAI's Whisper and GPT models enable accurate speech-to-text transcription and efficient summarization, enabling learners to extract crucial information without the need to watch complete movies. The incorporation of a question-answering module based on GPT enables learners to actively interact with the information by providing real-time responses that are contextually aware. Furthermore, the integration of LangChain and Questgen enables the automated creation of quizzes and flashcards, thereby promoting active learning and enhancing the retention of knowledge. The RAKE algorithm and KeyBERT are utilized to enhance the extraction of keywords, hence emphasizing crucial topics. This technological advancement has the potential to enhance educational results and establish a basis for future advancements in interactive learning systems. This study highlights the significant impact that artificial intelligence (AI) can have on transforming the field of e-learning by addressing a wide range of educational requirements.**

Keywords*—AI in education, video to text conversion, Whisper, GPT models, Gemini, interactive learning, question answering, quiz generation, flashcards, RAKE, KeyBERT, educational technology.*

I. INTRODUCTION

The educational environment is currently experiencing a significant shift due to the rapid expansion of e-learning. The utilization of this technology has become a potent instrument for disseminating instructional material and cultivating learning prospects in a versatile and easily accessible fashion. The e-learning market on a global scale has experienced significant growth, with a substantial value exceeding US$315 billion by 2021. It is anticipated to exhibit a spectacular Compound Annual Growth Rate (CAGR) of 20% from 2022 to 2028 [1]. The increase in popularity is driven by technological improvements that have enabled EdTech companies to develop a wide range of online platforms, effectively incorporating technology into every aspect of education [2]. Notwithstanding its indisputable promise, e-learning platforms continue to encounter obstacles. Research indicates a range of student perspectives regarding e-learning. According to research findings, a significant proportion of students, approximately 81%, see e-learning materials as engaging and intriguing. However, it is worth noting that a subset of students express negative attitudes towards e-learning. For example, a distinct survey revealed that 61.6% of nursing students held unfavorable opinions on e-learning[3]. The presence of this discrepancy highlights the imperative requirement for e-learning experiences to be customized in order to cater to the distinct requirements and preferences of a wide range of learners. Moreover, there are ongoing issues regarding the efficacy of e-learning in comparison to conventional instructional approaches. Several studies have indicated potential advantages, such as enhanced student-teacher interaction. The study found that 80% of students had heightened engagement through e-learning[4] and the promotion of self-directed learning [3]. However, several students highlighted possible disadvantages, such as social isolation. According to the report, 73% of students expressed apprehensions regarding social isolation caused by e-learning[4]. The present study aims to explore the potential of e-learning in transforming the educational industry by addressing the constraints associated with existing passive learning models. This study presents an innovative instructional application that utilizes artificial intelligence (AI) technology to convert static video content into a dynamic, interactive, and tailored learning encounter. This system intends to enhance the limitations of conventional e-learning by utilizing advancements in Artificial Intelligence (AI) to enhance student engagement, accommodate various learning styles, and strengthen knowledge retention. Its objective is to make a valuable contribution to the wider domain of interactive e-learning.

II. Literature Review

Rose et al. presents novel methodologies for the automated extraction of keywords from individual papers. The authors employ a methodological approach that prioritizes the efficiency and accuracy of the Rapid Automatic Keyword Extraction (RAKE) algorithm. The study demonstrates that RAKE exhibits superior computational speed and efficacy compared to conventional approaches, particularly when used to huge document corpora. This study investigates the algorithm's efficacy in detecting contextually pertinent and content-specific keywords, without necessitating a substantial amount of training data. The aforementioned statement highlights the significance of this study in augmenting information retrieval systems and expanding the accessibility of digital libraries. [5]. In their research, Hasan and Ng critically assess the state of automatic keyphrase extraction using a survey method to evaluate the efficacy of existing systems. They explore the inherent challenges of keyphrase extraction and pinpoint primary error sources, demonstrating that these systems still perform poorly compared to other natural language processing tasks. The study delves into algorithmic strategies, feature engineering, and the impact of different document formats on extraction success. It underscores the necessity for sophisticated feature design and the integration of machine learning to improve extraction accuracy. [6]. Florian Boudin et al. examine the efficacy of different centrality measures in graph-based keyphrase extraction. They employ a comparative methodology and utilize three standard datasets from various languages and domains to systematically assess well-known centrality measures, including degree centrality, closeness centrality, and betweenness centrality, among others. The findings of this study demonstrate that simple degree centrality gives results that are equal to those achieved by the widely employed TextRank algorithm. The simple degree centrality demonstrates superior performance in the context of short documents, when closeness centrality tends to produce the most favorable outcomes emphasizing the potential of alternate centrality metrics within the domain of keyphrase extraction, hence indicating noteworthy implications for improving the effectiveness and precision of automated keyphrase extraction systems. [7]. Mihalcea et al. present the "TextRank" algorithm, which centers on unsupervised techniques for extracting keywords and sentences. This approach utilizes a graph-based ranking model that is taken from established web search algorithms such as PageRank. The study utilizes a process in which natural language texts are transformed into graphs, allowing for the implementation of the TextRank algorithm to prioritize text units, such as words and sentences, depending on their significance determined by graph-based metrics. The study showcases that TextRank surpasses conventional methods by achieving competitive outcomes in keyword extraction and text summarization, without the need for supervised learning techniques or deep linguistic processing. The paper emphasizes the efficacy of graph-based methods in managing natural language data and the versatility of TextRank in different text-processing tasks highlighting the importance of graph-based ranking models in the field of natural language processing, namely in the extraction of informative content from extensive texts [8]. Widyassari et al. explores the numerous strategies employed in automatic text summarization. The authors employ a systematic methodology to classify and assess distinct summary techniques, which encompass extractive, abstractive, and hybrid methods. These techniques are applied in diverse fields. The article provides a comprehensive analysis that emphasizes notable progress in neural network-based models that utilize deep learning techniques to improve context comprehension and produce coherent summaries. The research highlights the crucial significance of semantic representation and the incorporation of AI to manage the subtleties of human language in producing concise summaries. The paper examines various components, including the performance criteria utilized in the assessment of summarizing systems, as well as the impact of developing technologies such as transformers on the advancement of automated summarization capabilities. The research concludes that these technologies may affect media, academia, and information retrieval. potential future works include making summarization tools more interpretable and customizable to meet users' needs[9]. Singh et al. examine the progressing approaches in the domain of automated text summarization, specifically focusing on scientific texts. The study utilizes a comprehensive literature review approach to examine different algorithms and evaluate their efficacy in condensing intricate scientific information. The investigation reveals that neural network-based methods are progressively outperforming older techniques in terms of accuracy and coherence. This analysis focuses on distinct components of these systems, including their capacity to incorporate contextual relevance and uphold content integrity[10]. Yuanxin Liu et al. propose a new method for generating paraphrases that utilizes the original topic of sentences as additional guidance. They employ a framework called Sequ2Seq, which incorporates topic words into both the input and generation processes of paraphrase creation. By conducting meticulous experiments on benchmark datasets, the research showcases that the inclusion of topic information greatly enhances the pertinence and excellence of paraphrases in comparison to conventional Seq2Seq models. This approach effectively improves the semantic coherence of the produced text by aligning it more closely with the contextual information offered by topic words. As a result, it effectively tackles prevalent challenges encountered in paraphrase generation, such as semantic drift. The study highlights the significance of topic coherence in the field of natural language processing and proposes that including contextual comprehension into neural models can enhance the precision and significance of text output.[11]. Jalin et al. present a sophisticated text-to-speech (TTS) system designed for the Tamil language. This system integrates Hidden Markov Models (HMM) to improve the accuracy of speech synthesis. This study employs a methodological strategy that combines multi-feature models with HMM predictors in order to enhance pronunciation and speed, resulting in speech that sounds more natural. The paper showcases a 6% enhancement in precision compared to conventional TTS systems, achieved through meticulous analysis and experimental validation. This study delves into the distinct components of voice synthesis, including intonation and prosody, by employing digital signal processing methodologies in an efficient manner. [12] The efficiency of

Hidden Markov Models (HMM) and Connectionist Temporal Classification (CTC) in automatic speech recognition (ASR) is investigated by Raissi et al. through an experimental approach. The researchers compare these two architectures by conducting full-sum training from scratch on the Switchboard and LibriSpeech corpora. The research emphasizes that both models exhibit notable levels of accuracy, with CTC exhibiting somewhat superior alignment characteristics as a result of its streamlined alignment constraints. This study investigates the intricacies of full-sum training, evaluating the influence of various modeling methodologies on the accuracy and temporal alignment of automatic speech recognition (ASR). [13]. Bain et al. introduce WhisperX, an advanced speech transcription technique that is specifically tailored for the analysis of lengthy audio recordings. Voice Activity Detection (VAD) is employed to efficiently divide the audio into manageable segments. Next, a Cut & Merge technique is employed with the objective of improving the efficiency of the transcription process. WhisperX utilizes the capabilities of the Whisper model, which is well-known for its extensive weak supervision, along with forced phoneme alignment approaches to provide accurate timestamps at the word level. The findings indicate that WhisperX outperforms previous models, namely Whisper and wav2vec2.0, in terms of word segmentation accuracy and the reduction of prevalent transcription problems such as redundancy and hallucinations. [14]. Peng et al. present a novel methodology for training extensive speech models, similar to OpenAI's Whisper, using solely open-source tools and publicly accessible data. Their approach involves the development of a multitask data format that can effectively handle diverse speech processing tasks, including language identification, multilingual automatic speech recognition, and speech translation. The research findings demonstrate that the OWSM model exhibits comparable performance to Whisper, while utilizing a far smaller dataset, as evidenced by comprehensive experimentation. This research article provides a comprehensive analysis of the alterations implemented in the initial Whisper training protocol. These improvements encompass modifications to the model architecture and training techniques, with the aim of enhancing efficiency and scalability.The concluding observations indicate that OWSM offers a comprehensive framework for the training of speech models [15]. Smith et al. explores the difficulties and advancements associated with the incorporation of direct translation processes into speech-to-text systems, without the need for an intermediate transcription process utilizing a comprehensive review technique to examine different neural network architectures that enable real-time translation. The performance of these architectures is evaluated across a range of languages and dialects. The study presents significant findings from extensive experiments and comparative analysis, which show notable enhancements in translation accuracy and speed. These improvements are particularly evident in low-resource languages. These improvements are achieved by employing advanced machine learning techniques, such as transfer learning and multi-task learning. This study provides a comprehensive analysis of the effects of various model designs, data augmentation techniques, and the incorporation of contextual understanding on the improvement of performance metrics in direct speech-to-text translation systems. This statement highlights the significant significance of these achievements in overcoming language barriers and improving global communication, especially in situations involving multiple languages and cultures. [16]. Hu et al. investigate a novel generative paradigm that leverages large language models (LLMs) to improve the quality of translation. This is achieved by incorporating a range of N-best hypotheses derived from basic models. Utilises an advanced methodology by integrating the advantages of LLMs with a novel dataset, HypoTranslate, comprising more than 592K hypothesis-translation pairs from 11 different languages. The findings of the study demonstrate that GenTranslate has superior performance compared to current cutting-edge models in both voice and machine translation tasks. This is accomplished by exploiting the extensive language knowledge and reasoning abilities of Language Models (LLMs) to combine and improve translation results, effectively making use of the entire semantic information present in numerous translation hypotheses. [17]. Zhang et al. examine the "SpeechUT" model, which is a novel approach that integrates voice and text modalities through the utilization of hidden-unit representations. The paper utilizes a unique encoder-decoder framework and implements a multi-task pre-training approach. This approach involves performing speech-to-unit, unit-to-text, and masked unit modeling tasks using extensive unpaired voice and text data. The proposed methodology enables notable enhancements in the domains of automatic speech recognition (ASR) and speech translation (ST) as compared to current approaches. The study showcases that SpeechUT attains cutting-edge performance on benchmark datasets such as LibriSpeech and MuST-C, as evidenced by comprehensive trials. This study investigates the advantages of employing discrete representation learning and strategically utilizing hidden units to efficiently align speech and text data inside a common latent space. The objective is to tackle the issue of inequalities between modality. [18]. Santos et al. examine the efficacy of extensive language models in translating languages that have limited resources. The research employs a thorough methodological framework, incorporating both quantitative evaluations and qualitative research methods to examine the quality of translation among multiple extensive models. The research presents a comprehensive analysis that demonstrates notable advancements in the capacity of language models to effectively handle languages with limited resources. However, it also highlights the persistent inequalities in translation quality when compared to languages with ample resources. It emphasizes the constraints of existing model structures and the unequal distribution of datasets that result in these inequalities.[19]. Rahman et al. explore the synthesis of contextually appropriate inquiries in the Bengali language by employing a transformer-based model that incorporates answer awareness during the training phase. By employing a methodological framework that leverages the sophisticated functionalities of transformer architectures, this study aims to improve the caliber and pertinence of automatically generated inquiries by integrating answer-specific data directly into the attention processes of the model. The study demonstrates that the suggested model surpasses existing systems in generating coherent and contextually suitable queries, as evidenced by

thorough analysis and a series of trials. This study investigates the effects of including answer awareness on the performance of the model, specifically focusing on enhancements in precision and flexibility to the intricate linguistic characteristics of the Bengali language. The study demonstrates the significance of customized methodologies in language-specific contexts and proposes a significant progression in the domain of automated question production. [20] Vachev et al. present a unique automated system that utilizes a multi-task paradigm to generate multiple-choice questions (MCQs) from educational texts. This system incorporates question and answer production along with distractor creation, employing the T5 transformer model. This study investigates various aspects, including the effectiveness of generating distractors and the utilization of neural network models for automating the generation of educationally pertinent questions. [21]. Sarracén et al, present a hybrid methodology that integrates BERT's attention mechanism and graph theory concepts to identify offensive keywords. This approach employs an unsupervised technique that combines BERT's multi-head self-attention with eigenvector centrality to analyze and extract pertinent keywords from datasets containing both offensive and non-offensive tweets. This approach enables the rapid identification of terms that exhibit a higher frequency in offensive situations, The study demonstrates that this innovative methodology not only enhances the process of extracting keywords but also offers valuable insights into the intricacies of foul language usage in online contexts.[22]. Liu et al. explores the progress and difficulties associated with the integration of text, audio, and video data in order to provide succinct multi-modal summaries. The research utilizes a thorough review technique to analyze current methods in multi-modal summarization and evaluate their efficacy in different applications. Based on a comprehensive examination, it becomes evident that although notable advancements have been achieved, there are still obstacles to overcome, namely in the areas of harmonizing modalities and maintaining the integrity of material across various data formats. This study emphasizes the need of constructing resilient models that can proficiently incorporate and amalgamate data from many sources in order to augment user engagement and information retention. [23] Kumar et al. present a comprehensive assessment of transformer-based models specifically designed for the processing of Indian languages. This research aims to fill a notable void in the field of language technology by employing a comparative analysis approach. The authors evaluate different transformer architectures across multiple Indian language datasets in order to determine the most efficient configurations. This study presents a comprehensive analysis that showcases the significant enhancements in language comprehension tests for Indian languages resulting from the implementation of particular transformer changes. This research emphasizes the significance of incorporating positional encoding and attention mechanisms as crucial elements in enhancing the performance of models. [24]. Zhiyun Fan et al. present a unique methodology for pre-training an encoder-decoder sequence-to-sequence (seq2seq) model using unpaired speech and transcripts. The primary objective of the project is to improve automatic speech recognition (ASR) tasks by incorporating extensive acoustic and linguistic knowledge into the seq2seq model using a two-stage pre-training approach. During the initial phase, referred to as acoustic pre-training, the encoder undergoes training to accurately anticipate speech feature chunks that have been masked. Subsequently, in the second phase, known as linguistic pre-training, the decoder is pre-trained using synthesized speech that has been generated from transcripts. The paper demonstrates the usefulness of the suggested strategy by conducting extensive experiments on datasets such as AISHELL-2, AISHELL-1, HKUST, and CALLHOME. The results show a considerable reduction in the relative character error rate (CERR). The research elucidates the disparities between its methodology and established pre-training techniques like BERT and restricted Boltzmann machines (RBM), with a particular emphasis on the concentration on unpaired speech and transcripts . [25]

III. Background Study

A. Whisper API

Whisper, an advanced speech recognition system, distinguishes itself within the realm of machine learning by employing sophisticated model architectures and learning methods. This research examines three fundamental elements of Whisper's architecture: the convolutional neural network (CNN) utilized in its encoder, the attention mechanism integrated into the encoder-decoder configuration, and the specific loss function employed to enhance performance.

1) Encoder with Convolutional Neural Network (CNN):

Whisper's encoder, which incorporates a convolutional neural network (CNN), is the central component responsible for processing and interpreting audio data. The Convolutional Neural Network (CNN) plays a vital role in extracting intricate characteristics from the input audio spectrograms, which are intricate visual depictions of sound frequencies as they change over time. The encoder comprises convolutional layers that undergo a systematic process of feature extraction, as stated by the following equation:

$$\mathbf{Output[i, j, k] = \Sigma\ (Weight[m, n, p] * Input[i-m, j-n, p]) + Bias[i, j, k]}$$

Here, Output[i, j, k] represents the feature map produced by the convolutional layer, indicating the extracted features at position (i, j) for the kth feature map. The Weight[m, n, p] terms are the parameters of the convolutional filter, and the Bias[i, j, k] term is an additive bias. This convolution operation captures local dependencies within the spectrogram, such as shifts in frequency or intensity, crucial for identifying speech patterns[26]

2) Attention Mechanism:

Whisper's encoder-decoder architecture is enhanced by an attention mechanism, which plays a pivotal role in focusing the model on relevant parts of the audio input during the decoding phase. The attention mechanism operates by assigning weights, or attention scores, to different parts of the

encoder output. These scores determine how much each segment of the input should influence the output at each step of the decoding process. The attention weights are calculated using the following scoring function:

$$\alpha_{ij} = \text{softmax}(\text{score}(s_i, h_j))$$

where **s_i** is the current state of the decoder, and **h_j** is the output from the encoder at step **j**. The softmax function ensures that the weights sum to one, effectively normalizing the influence each input part has. The decoder then computes a context vector as a weighted sum of the encoder outputs:

$$\text{context}_i = \Sigma (\alpha_{ij} * h_j)$$

This context vector serves as a dynamic summary of relevant input features at each step of decoding, allowing the model to generate coherent and contextually appropriate output text.

3) Loss Function:

To train the Whisper model effectively, a specifically tailored loss function is used, which takes into account both the accuracy of individual words and the overall coherence of the generated text. The loss function is expressed as:

$$\text{Loss} = \lambda_{word} * \text{Word Error Rate (WER)} + \lambda_{coherence} * \text{Coherence Score}$$

Here, λ_word and λ_coherence are weighting factors that balance the importance of word accuracy and textual coherence in the loss calculation. The Word Error Rate (WER) is a common metric in speech recognition that measures the minimum number of substitutions, insertions, and deletions required to change the predicted text into the reference text. The Coherence Score evaluates how logically connected and contextually consistent the output text is compared to the intended message in the audio.

B. Large Language Models:

Large language models like Gemini and GPT leverage a ubiquitous architecture for text generation tasks: the encoder-decoder. The encoder, employing techniques like word embeddings and RNNs/transformers, meticulously analyzes the input text, capturing the context and word relationships. This encoded information is then passed to the decoder, which serves as the language generation engine. Here, the decoder, empowered by an attention mechanism, focuses on pertinent sections of the encoded input while progressively building the output text word-by-word. This success hinges on extensive pre-training on massive text and code datasets, allowing these models to grasp the intricacies of language and generate human-quality text.

1) GPT-4

GPT-4, an iteration in the series of Generative Pre-trained Transformers by OpenAI, is a state-of-the-art language model known for its deep learning capabilities and wide-ranging applicability. This advanced model builds on the architectural principles of its predecessors, incorporating a more extensive transformer network with increased layer depth and width, thereby enhancing its ability to process and generate human-like text. GPT-4's architecture is designed to optimize both the quantity and quality of data it can process. With an increased parameter count significantly higher than GPT-3, GPT-4 can maintain more extended contexts and understand subtler nuances in language. This model supports larger context windows, allowing it to retain and refer back to information from earlier in the text, which is crucial for generating coherent and contextually rich responses. Despite its complexity, GPT-4 employs various optimization techniques to manage computational resources effectively. These include techniques like sparse activation and improved attention mechanisms that help balance the computational load, making it feasible to run on existing hardware without compromising on speed or efficiency. One of the standout features of GPT-4 is its enhanced token capacity, which enables it to handle extended dialogues and documents more effectively than its predecessors. This capability allows GPT-4 to perform exceptionally well in tasks that require a deep understanding of context and content continuity, such as detailed article writing, complex question answering, and sophisticated dialogue simulations. GPT-4 utilizes an advanced training regimen that includes both unsupervised and supervised learning phases. In the unsupervised phase, the model learns from a vast corpus of text data, enabling it to understand and generate human-like text. The supervised phase involves fine-tuning the model on specific tasks and datasets, which sharpens its capabilities in particular domains such as legal document analysis, technical manuals, and creative fiction writing. This domain-adaptive training ensures that GPT-4 not only excels in general language understanding but also in specialized applications requiring high precision and expert knowledge.[27]

2) Gemini

Gemini 1.5 Pro has an advanced architectural design that integrates transformer networks with a mixture-of-experts (MoE) methodology, thereby augmenting its capacity to efficiently handle extensive volumes of multimodal data. The hybrid structure enables the model to intelligently distribute computational resources according to the task's difficulty, resulting in a highly scalable and adaptable system that can handle various data types and durations. The utilization of Mode of Expansion (MoE) greatly enhances the parameter efficiency of the model, hence allowing it to process a greater amount of information while requiring fewer computational resources in comparison to conventional dense models. One of the key performance enhancements in Gemini 1.5 Pro is its extended token capacity, which can handle contexts up to 10 million tokens. This is a substantial increase over previous models like GPT-4 Turbo and Claude 2.1, which are limited in their context window sizes. Gemini 1.5 Pro's ability to maintain context over these long sequences allows for more coherent and contextually accurate responses, especially in tasks requiring deep understanding over extended narratives or documents.

The model also incorporates advanced training techniques, including domain-adaptive pre-training and fine-tuning on specific tasks to optimize performance across various applications. This targeted training approach helps in refining the model's capabilities in specific areas such as legal analysis, medical diagnostics, and educational content, ensuring high precision and reliability.[28]

- **Mixture of Experts (MoE):**

MoE architectures are often used in advanced LLMs to manage the computational complexity and enhance the model's capacity to focus on relevant parts of the data more effectively. This method allows for scaling the model size without a proportional increase in computational demand, which is crucial for models expected to handle extensive multimodal contexts.

$$y = \sum_{i=1}^{N} g_i(x) \cdot f_i(x)$$

In this equation, yy represents the output of the MoE layer, NN is the number of experts, $gi(x)$gi(x) is the gating network's output that determines the weight of the ii-th expert's contribution, and $fi(x)$fi(x) is the output of the ii-th expert for the input xx.

- **Transformer Attention Mechanism:**

Attention mechanisms are central to the architecture of modern LLMs, starting from models like the original Transformer model up to more recent iterations. They allow the model to dynamically weigh the importance of different parts of the input data, which is crucial for both maintaining long-range dependencies and integrating information across different modalities.

$$y = \sum_{i=1}^{N} g_i(x) \cdot f_i(x)$$

Here, QQ, KK, and VV represent the query, key, and value matrices derived from the input data, respectively. dkdk is the dimension of the key vectors, ensuring proper scaling. This formula reflects how attention weights are computed in a transformer model, which is likely a core component of Gemini 1.5 Pro's architecture.

- **Cross-Entropy Loss for Classification Tasks:**

Cross-entropy is a standard loss function used in classification tasks within many LLMs. It measures the performance of a classification model whose output is a probability value between 0 and 1. It's effective for tasks where model outputs are expected to match categorical labels.

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

This equation represents the cross-entropy loss, where yo,cyo,c is a binary indicator of whether class label cc is the correct classification for observation oo, and po,cpo,c is the predicted probability that observation oo is of class cc. For models involved in classification tasks, such as categorizing text or recognizing objects in images, this loss function is commonly used.

- **Regularization Term:**

Regularization techniques like L2 regularization (shown in the formula) are common in training neural networks, including LLMs, to prevent overfitting by penalizing large weights. This helps models generalize better to new, unseen data rather than memorizing the training data.

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

In this regularization formula, $L0$L0 represents the original loss function (such as cross-entropy), $\theta\theta$ represents the model parameters, and $\lambda\lambda$ is the regularization coefficient. This term helps prevent overfitting by penalizing large weights.

C. NLTK

NLTK (Natural Language Toolkit) is a library in Python designed to handle a variety of natural language processing tasks. It includes functions for text processing such as tokenization, parsing, classification, stemming, tagging, and semantic reasoning. NLTK facilitates text analysis by providing easy-to-use interfaces and a suite of text corpora and lexical resources. It supports operations like frequency distribution of a word

f(w)= Total number of words in the text/Number of times word w appears in a text

TF-IDF(t, d) = TF(t, d) * IDF(t)

D. spaCy

spaCy is a Natural Language Processing (NLP) library that is widely used for tasks such as tokenization, part-of-speech tagging, named entity recognition, dependency parsing, word vectorization and more. Similarity between two word vectors is calculated using cosine similarity

Cosine Similarity (S(u, v)) = (u · v) / ||u|| ||v||

POS tagging is done by using conditional probability P(Tag|Context)

E. Gensim

Gensim is a Python library designed primarily for topic modeling, document indexing, and similarity retrieval with large corpora. It implements various machine learning algorithms that are particularly useful in natural language processing (NLP) and information retrieval (IR). It employs statistical algorithms for identifying patterns in data.

F. Transformers

Transformers are a class of models in NLP that rely on the mechanism of self-attention to weigh the influence of different words in a sentence, irrespective of their positional distance. This architecture enables the model to capture complex word dependencies and improve the understanding of context in text processing tasks.

$$\alpha_{ij} = \text{softmax}((W_q * h_i) \cdot W_k^T * h_j))$$

After attention, the output is normalized and passed through a position-wise feedforward neural network, applying further transformations to refine the representation.

Comparison of different transformers				
Feature	***BERT***	***KeyBERT***	***Sentence Transformer***	***Hugging Face Transformer***
Type	***Pre-trained Language Model (PLM)***	***Keyword Extractor***	***Sentence Embeddings***	***Open-source Transformers***
Focus	***Understanding overall text meaning and relationships***	***Extracting keywords and keyphrases***	***Representing entire sentences as vectors***	***Various NLP tasks like classification summarization etc***
Application in e-Learning	***Personalize learning materials based on student understanding and create interactive quizzes with tailored questions***	***Generate targeted summaries of learning materials***	***Improve search functionality within e-learning platforms***	***Develop chatbots for personalized learning support***
Benefits	***Adapts to individual learning styles and pace***	***Enhances information retrieval for students***	***Enables efficient content navigation***	***Provides automated support and feedback***
Limitations	***Requires significant computational resources***	***May struggle with complex or domain-specific language***	***Doesn't capture full context within sentences***	***Requires expertise in choosing and implementing models***
Examples	***Identify key concepts in e-learning content and adjust difficulty level of follow-up material***	***Summarize video lectures for quick review***	***Cluster similar learning materials for thematic exploration***	***Power a chatbot that answers student questions based on course content***

Fig. 1. Comparison of different transformers

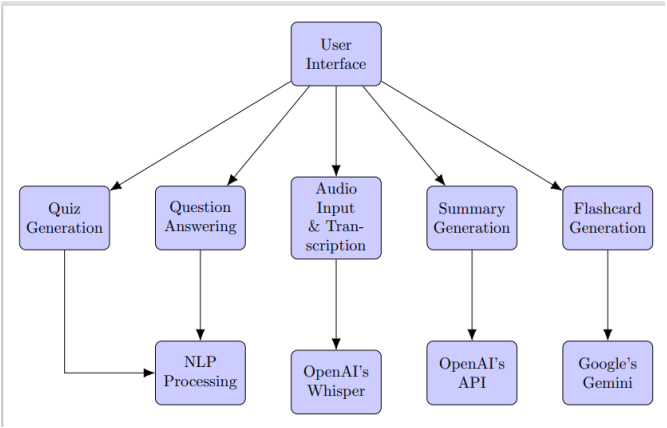

Fig. 2. System Architecture

IV. Methodology

A. Initialization Phase

Prior to providing the user with transcription and interactive capabilities, the Flask web application goes through an initialization step. During this phase, the server executes a series of essential startup tasks in order to guarantee the complete functionality of the system. The initial step of the Flask application involves verifying the presence and reliability of essential pre-trained models, including OpenAI's Whisper for transcription and Google's Gemini for flashcard production. Additionally, it initiates API connections and verifies their responsiveness. The server proceeds to configure the environment by establishing the suitable routing and verifying the proper loading and functionality of Python libraries, including PyAudio, nltk, and spaCy. Furthermore, the system conducts a connection test with the localhost and makes necessary preparations for the web interface to effectively manage incoming HTTP requests. The preliminary stage is of utmost importance in upholding the dependability and effectiveness of the application, guaranteeing that all elements are coordinated and prepared for user engagement.

B. Audio Transcription

When the transcription process is initiated using the Flask web interface, the application establishes the essential elements required for the recording and processing of audio. The Python package PyAudio is utilized for the purpose of capturing audio input from the microphone in smaller segments. The aforementioned segments are subsequently transcribed instantaneously with OpenAI's Whisper speech-to-text technology, therefore transforming oral utterances into written text. As the transcription process progresses, the

resultant text is consistently transmitted to the user interface, thereby presenting a live representation of the ongoing transcription. The transcription process persists until the user either explicitly terminates it or a predetermined time limit is reached. Upon completion of the transcription, the user is presented with the completed text, offering them the opportunity to generate a summary, a series of multiple-choice questions, or a set of flashcards.

C. Question Answering

Initially, the server implements essential natural language processing (NLP) functions, which encompass the extraction of salient keywords from the user's query. These techniques encompass tokenization, which involves dividing an input text into individual words or tokens, enabling a thorough examination. POS tagging helps in picking words based on their contribution to the sentence by outlining the semantic meanings and grammatical roles of nouns, verbs, and adjectives. Name Entity Recognition (NER) is a technique that can be employed to identify named entities, such as individuals, institutions, places, or dates, inside textual data. These acknowledged Named Entity Recognition (NER) may be crucial terms, contingent upon the characteristics of the query. Finally, tree parsers provide responses for sentence grammar analysis to consider the connections between words, enabling the assessment of keyword context. Finally, algorithms that are specifically developed to extract phrases or keywords with the utmost significance possess the ability to identify factoring based on factors such as frequency, context, and relevance. By employing a comprehensive range of natural language processing (NLP) techniques, the server will possess the capability to effectively identify and suggest the most suitable keywords derived from the user's query. The keywords obtained through this process are subsequently combined with the contextual information extracted from the transcribed content in order to formulate the query. The question that has been formulated is transmitted to the API of OpenAI, and a result is then obtained. Once the API response, which has been identified as the most pertinent answer given the context, is acquired, it will then be transmitted back to the web interface for presentation. This technique facilitates the provision of accurate and contextually appropriate responses to the user, hence augmenting the overall user experience.

D. Quiz Generation

Upon receipt of the user's request, the Flask server proceeds to do a comprehensive analysis of the transcript in order to determine the salient points and subjects. The textual data is processed by the server using advanced natural language processing (NLP) libraries, such as nltk and spaCy. These libraries are utilized to extract the primary information and comprehend the significant subjects present in the text. In addition, the server utilizes KeyBERT, a sentence transformer-based algorithm that produces contextualized keyword representations to facilitate the interpretation of information. Utilizing an extensive understanding of the subject matter, the server generates a series of multiple-choice questions sourced from the Questgen library. The server use pre-established templates or algorithms to generate questions that encompass many facets of the topic, leading to a diverse range of question formats and a thorough examination of the subject matter. To enhance the quiz result, accurate responses and distractions are generated for each question, thereby stimulating the users while facilitating their learning simultaneously. The questions and answers derived from the transcribed content are subsequently connected to the online interface, enabling users to choose and assess their comprehension of the material.

.

E. Summary Generation

When a user requests a concise overview of the transcribed material, the Flask server utilizes Python's file-handling capabilities to retrieve the complete transcript that was previously stored in a file. The transcript is transmitted to OpenAI's API, which employs state-of-the-art natural language processing (NLP) models such as GPT-3 to produce a coherent summary. The application programming interface (API) takes into account user-provided factors, such as the desired length or level of detail, in order to customize the summary accordingly. Within the API, GPT-3 utilizes its deep learning model to analyze the transcript, enabling it to comprehend the textual context, semantics, and the relationships among different textual components. Through the application of sophisticated language modeling techniques, GPT-3 demonstrates the capability to accurately identify the relevant points, concepts, and noteworthy particulars inside the transcript. Subsequently, it consolidates this data into a thorough and succinct synopsis to ensure that the crucial aspects of the reading are not disregarded. Ultimately, the synopsis is transmitted to the Flask server, which subsequently relays it to the web interface for presentation to the user. Flask facilitates the entire process by managing HTTP requests and responses, Python for scripting and implementing logic, and OpenAI's API for natural language processing and text summarizing. The program utilizes technological advancements to optimize the usefulness of the transcript by providing users with a handy method of acquiring accurate summaries of the transcribed material within a brief timeframe. This aids users in comprehending the primary insights and key points without the necessity of engaging in extended reading.

.

F. Flashcard Generation

Following transcription and summarization, the Flask server initiates flashcard generation using Google’s Gemini model to ensure concise content that is suitable for quick revision and reinforcement of key concepts. This is performed by employing NLP techniques such as named entity recognition and keyword extraction, similar to the processes used in quiz generation but optimized for brevity and clarity, both of which are essential for effective flashcards. These points are then transformed into question and answer formats using Gemini's capability to generate natural, contextually appropriate language. The generated flashcards are formatted, stored on the server, and made accessible through the web interface, allowing users to interact with them in a self-paced manner.

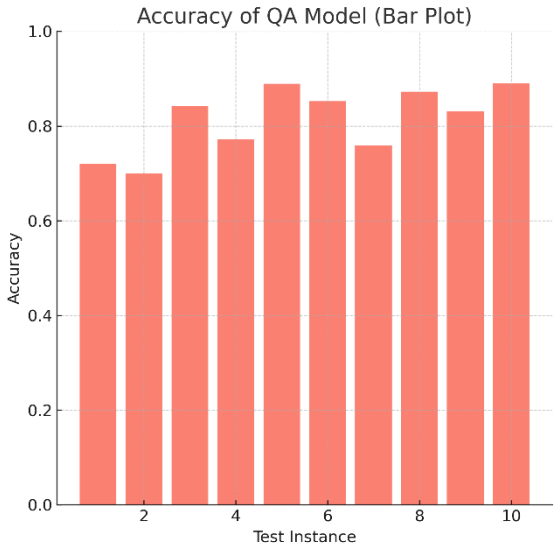

Fig. 3. Accuracy of Question Answering Model

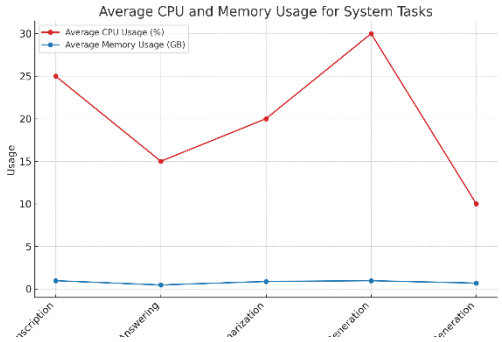

Fig. 4. CPU and Memory Usage

V. Results

The study introduces an innovative AI-driven educational platform that aims to transform the e-learning industry by transforming static video content into an engaging and customized educational experience. The application is designed to improve user engagement and information retention by utilizing OpenAI's Whisper model for real-time speech-to-text transcription, OpenAI's GPT for quality assurance and content summary, as well as LangChain and Questgen for creating quizzes and flashcards. The transcription component of the tool is a prominent aspect, utilizing OpenAI's Whisper to produce precise text transcriptions from spoken language. This feature demonstrates a noteworthy average Word Error Rate (WER) of 9.3%. This precision guarantees that individuals with hearing impairments have equitable access to information and serves as the basis for the sophisticated interactive functionalities of the platform. The quality assurance (QA) system, which is an essential element of the instrument, has been subjected to comprehensive testing in many scenarios, consistently achieving an average accuracy rate of 81.11%. This demonstrates the system's proficiency in utilizing NLP approaches to accurately and contextually analyze and respond to user questions.

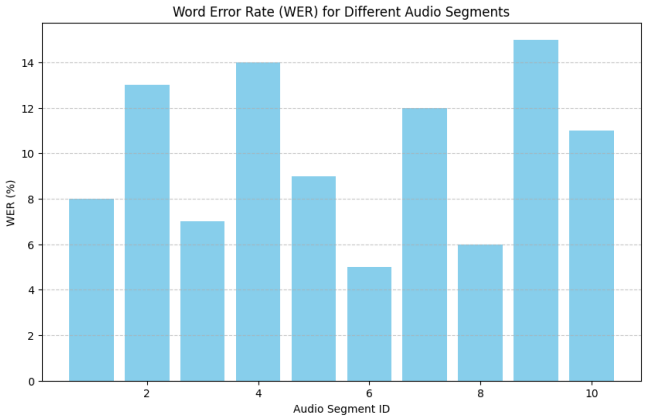

Fig. 5. Word Error Rate

The QA feature efficiently sustains student engagement in e-learning by facilitating direct interaction with the educational information, which is a noteworthy accomplishment. The tool's capacity to automatically generate quizzes and flashcards is commendable, as it enhances the interactive learning experience. The quiz module utilizes the transcribed text to ascertain fundamental concepts and construct diverse question formats, incorporating supplementary elements that stimulate the learner and augment understanding.

VI. Conclusion

This study finds that the AI-driven educational platform described here demonstrates a crucial advancement in the evolution of e-learning. Through the integration of sophisticated transcription techniques and interactive technology, this tool enhances student engagement and expands the availability of educational materials. The integration of OpenAI's Whisper and GPT models with LangChain and Questgen demonstrates a unique and comprehensive approach to education that encompasses accurate transcription and the creation of dynamic content. The transcribing accuracy of the platform is notable, as seen by a mean Word Error Rate (WER) of 9.3%. Additionally, the platform demonstrates a strong average accuracy rate of 81.11% in question-answering, suggesting its potential to effectively cater to various learning needs. There is a wide range of potential future improvements for the platform. Personalization algorithms have the potential to be refined in order to effectively adjust content and evaluation methods to align with individual learning trajectories, hence facilitating a more customized educational experience. Increasing the number of languages included would enhance the platform's worldwide influence and cultural involvement. In addition, the incorporation of speech biometrics could offer a deeper understanding of learners' involvement beyond conventional measurements, capturing subtleties in tone and emphasis for

a more comprehensive feedback system. The technology has the ability to create a collaborative learning environment by including augmented and virtual reality aspects. This could allow learners to engage in a more concrete examination of complex subjects. This type of environment has the potential to convert abstract academic concepts into tangible experiences, hence improving understanding and memory recall.This platform's development trajectory not only foresees but also serves as a source of inspiration for future scholarly contributions in the field of educational technology. By adopting a path of ongoing innovation, the platform is poised to revolutionize the methods of e-learning, thus expanding the boundaries of educational accessibility and facilitating interactive learning. The flashcards function as a convenient review instrument, condensing essential concepts into easily comprehensible question and answer pairs, so facilitating the retention of information in memory. Furthermore, the summary function, which is driven by GPT models, has exceptional proficiency in compressing extensive transcripts into concise summaries. The provided summaries have been customized to align with the user's preferences for specific information, allowing learners to efficiently grasp the core content. The models' ability to extract and highlight crucial information was proved through testing on various transcript complexity. The tool's technological architecture is constructed upon a Flask backend, which guarantees effective handling of concurrent requests while maintaining optimal performance. The use of Python modules such as PyAudio and nltk has led to the smooth processing of audio and the efficient handling of linguistic data. The system has been specifically built to function within the memory consumption limit of 1 GB for all jobs.

REFERENCES

[1] P. Dubey, R. L. Pradhan, and K. K. Sahu, "Underlying factors of student engagement to E-learning," *Journal of Research in Innovative Teaching and Learning*, vol. 16, no. 1, pp. 17–36, Mar. 2023, doi: 10.1108/JRIT-09-2022-0058.

[2] A. Al-Azawei, P. Parslow, and K. Lundqvist, "Investigating the effect of learning styles in a blended e-learning system: An extension of the technology acceptance model (TAM)," 2017.

[3] F. A. Mojarad, A. Hesamzadeh, and T. Yaghoubi, "Exploring challenges and facilitators to E-learning based Education of nursing students during Covid-19 pandemic: a qualitative study," *BMC Nurs*, vol. 22, no. 1, Dec. 2023, doi: 10.1186/s12912-023-01430-6.

[4] A. Rawashdeh, "Advantages and Disadvantages of Using e-Learning in University Education: Analyzing Students' Perspectives," vol. 19, no. 2, pp. 107–117, 2021, [Online]. Available: www.ejel.org

[5] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in *Text Mining: Applications and Theory*, John Wiley and Sons, 2010, pp. 1–20. doi: 10.1002/9780470689646.ch1.

[6] K. S. Hasan and V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," Association for Computational Linguistics. [Online]. Available: http://github.com/snkim/AutomaticKeyphraseExtraction/

[7] F. Boudin, "A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction," 2013. [Online]. Available: http://networkx.github.io/

[8] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts."

[9] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4. King Saud bin Abdulaziz University, pp. 1029–1046, Apr. 01, 2022. doi: 10.1016/j.jksuci.2020.05.006.

[10] [N. Ibrahim Altmami and M. El Bachir Menai, "Automatic summarization of scientific articles: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4. King Saud bin Abdulaziz University, pp. 1011–1028, Apr. 01, 2022. doi: 10.1016/j.jksuci.2020.04.020.

[11] Y. Liu, Z. Lin, F. Liu, Q. Dai, and W. Wang, "Generating paraphrase with topic as prior knowledge," in *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, Nov. 2019, pp. 2381–2384. doi: 10.1145/3357384.3358102.

[12] [A. Femina Jalin and J. Jaya Kumari, "A novel text to speech technique for tamil language using hidden Markov models (HMM)," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 10, pp. 38–47, Aug. 2019, doi: 10.35940/ijitee.I8589.0881019.

[13] T. Raissi, W. Zhou, S. Berger, R. Schlüter, and H. Ney, "HMM vs. CTC for Automatic Speech Recognition: Comparison Based on Full-Sum Training from Scratch," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.09951

[14] [14] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.00747

[15] [Y. Peng *et al.*, "Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.13876

[16] C. Xu *et al.*, "Recent Advances in Direct Speech-to-text Translation," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.11646

[17] Y. Hu *et al.*, "GenTranslate: Large Language Models are Generative Multilingual Speech and Machine Translators," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.06894

[18] Z. Zhang *et al.*, "SpeechUT: Bridging Speech and Text with Hidden-Unit for Encoder-Decoder Based Speech-Text Pre-training," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.03730

[19] V. Mujadia *et al.*, "Assessing Translation capabilities of Large Language Models involving English and Indian Languages," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.09216

[20] J. F. Ruma, T. T. Mayeesha, and R. M. Rahman, "Transformer based Answer-Aware Bengali Question Generation," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 314–326, Jun. 2023, doi: 10.1016/j.ijcce.2023.09.003.

[21] K. Vachev, M. Hardalov, G. Karadzhov, G. Georgiev, I. Koychev, and P. Nakov, "Leaf: Multiple-Choice Question Generation," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.09012

[22] G. L. D. la P. Sarracén and P. Rosso, "Offensive keyword extraction based on the attention mechanism of BERT and the eigenvector centrality using a graph representation," *Pers Ubiquitous Comput*, vol. 27, no. 1, pp. 45–57, Feb. 2023, doi: 10.1007/s00779-021-01605-5.

[23] A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, and M. Hasanuzzaman, "A survey on multi-modal summarization," *ACM Comput Surv*, vol. 55, no. 13s, pp. 1–36, 2023.

[24] K. Jain, A. Deshpande, K. Shridhar, F. Laumann, and A. Dash, "Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.02323

[25] Z. Fan, S. Zhou, and B. Xu, "Unsupervised pre-training for sequence to sequence speech recognition," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.12418

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." [Online]. Available: https://github.com/openai/

[27] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.08774

[28] Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context Gemini Team, Google