# Generative AI Powered Learning Companion for Personalised Education and Broader Accessibility

Pranjal Sharma
Faculty of Computer Science and Engineering (CSE)
F.E.T Agra College

R.K. Sharma, PhD
Faculty of Computer Science and Engineering (CSE)
F.E.T Agra College

## ABSTRACT
This research presents the development and evaluation of a hybrid Convolutional Neural Network (CNN) and the Bidirectional long -term short -term memory (BILSTM) model for speech recognition, especially tailored for educational applications. Using the Mozilla Common Voice Dataset, the model suffered an impressive testing accuracy of 91.87% and less testing loss of 0.2966. The study highlighted the importance of effective preprocessing, including noise reduction, audio trimming, and MEL-Frequency Cepstral Coefficients (MFCC) feature extraction, which were necessary to improve model performance. The CNN-BiLSTM architecture enabled the model to capture both local and long-range temporary dependence, making it strong for diverse accents, speech speeds and background noise. This task reflects the viability of implementing advanced speech recognition systems in the generative AI-in-charge learners, contributing to the manufacture of inclusive and accessible educational devices. Future research can detect fine-tuning for specific domains to carry forward multilingual dataset, attention mechanisms, and performance.

## Keywords
Speech Recognition, Generative AI, Convolutional Neural Network, Bidirectional Long Short-Term Memory, Educational Tools, Mozilla Common Voice, Preprocessing, Mel-Frequency Cepstral Coefficients, Accessibility, Inclusivity.

## 1. INTRODUCTION
The rise of generative artificial intelligence (AI) is reshaping the landscape of education by providing innovative solutions to some of the most persistent challenges in the field—namely, equitable access to personalized, inclusive learning [1], [2]. Generative AI, which includes models like OpenAI's GPT-4, Google's PaLM 2, and Meta's LLaMA 2, has the ability to create human-like lessons, code, speech, and other material formats, enabling interactive and customized educational experiences [1]. Unlike traditional educational models that rely heavily on static, standardized materials, generative AI facilitates the creation of adaptive learning environments where instructional materials and assessments continuously align with the needs, performance, and pace of the learner [10].

As classrooms become more diverse and globally connected, the demand for personalized learning paths grows rapidly. Research indicates that individualized instruction can significantly enhance student engagement and achievement, particularly for learners with diverse needs or from marginalized communities [12]. Generative AI has shown promise in addressing these needs by acting as an intelligent learning companion—reducing the cognitive load, offering clarifications, following conversations, and providing scaffolding based on learned data [5]. Additionally, AI-driven tools enhance accessibility by supporting multilingual education, offering speech-to-text features, and accommodating neurodiverse learners through flexible content delivery [7].

However, despite its transformative potential, the deployment of generative AI in education is not without challenges. Bias, data privacy, explainability, and ethical concerns regarding over-automation require careful consideration and regulation [8]. Furthermore, there is a significant gap in the integration of AI models within real-world educational systems that are scalable, inclusive, and pedagogically sound. This research aims to bridge these gaps by developing a generative AI-powered speech recognition system to enhance both personalization and accessibility in education. By combining adaptive content delivery with real-time speech recognition capabilities, this system aims to support learners with diverse linguistic backgrounds, learning abilities, and accessibility barriers—creating a model for the next generation of inclusive education.

## 2. LITERATURE REVIEW
### 2.1. Personalized Learning and AI
Personalized education refers to instructional approaches tailored to individual students' needs, preferences, and pace. Recent research emphasizes the role of AI in enabling personalized learning paths. Baker et al. [9] observed that adaptive learning systems using AI algorithms improved learner engagement and outcomes by dynamically adjusting content based on performance metrics. Additionally, Khosravi et al. [10] introduced AI-based learning companions that use generative models to simulate human-like interactions, enhancing student autonomy and motivation. These systems represent a shift from static e-learning platforms towards responsive, learner-centric experiences.

### 2.2. Generative AI in Education
Generative AI models like GPT-4, LLaMA 2, and PaLM 2 have demonstrated the ability to create educational content, simulate tutor interactions, and support language learning. According to Zhai [11], these models can produce coherent, relevant materials that adapt to students in real-time. Holstein and Aleven [12] argued that generative AI could reduce instructional barriers by providing differentiated support. However, they caution that over-reliance on AI may marginalize the human elements of learning.

### 2.3. Educational Accessibility via AI
AI technologies have played a crucial role in breaking down barriers to education for learners with disabilities or in remote areas. Tools such as real-time speech-to-text, auto-captioning, and text simplification foster inclusivity [13]. For example, Google's Project Relay and Microsoft's Azure Cognitive Services have empowered users with speech impairments to interact effectively with digital content. Hwang et al. [14]

showed that speech recognition increased access for neurodiverse students and non-native speakers in virtual classrooms.

## 2.4. AI-Driven Speech Recognition and Learning Companions

Speech-enabled AI companions provide an ideal solution for hands-free, real-time interaction in inclusive learning environments. Integrating advanced models like Whisper and WAV2VEC 2.0 into educational systems allows learners to effectively utilize natural language interfaces [14]. These tools not only facilitate engagement for learners with visual or motor impairments but also enhance literacy and language acquisition through oral communication. Lu and Zhang [10] demonstrated significant improvements in reading comprehension in children using AI tutors equipped with speech interfaces.

## 2.5. Ethical and Equity Considerations

While AI opens new opportunities, ethical concerns persist. These include algorithmic bias, data privacy, and equitable access to technology [3]. Bias in training datasets can lead to skewed educational recommendations or inaccurate feedback. Furthermore, unequal access to devices, internet connectivity, and digital literacy—often referred to as the digital divide—may exacerbate existing disparities if unaddressed. Floridi et al. [1] emphasize the need for ethical-by-design AI systems that promote transparency, fairness, and inclusivity in education.

## 2.6. Gaps Identified

While enough research education highlights the ability of generic AI and speech recognition technologies, the implementation studies of the real world remain limited, especially in diverse, multilingual and low-resources settings. Most systems are still experimental or localized. There is also a lack of long -term empirical data on the cognitive and emotional effects of AI tutors on learners in various age groups.

## 3. System Design and Workflow

The development of the generative AI-powered speech recognition system follows a structured and iterative approach, emphasizing accuracy, latency, and accessibility for educational applications. Below are the sequential steps adopted in system design and implementation:

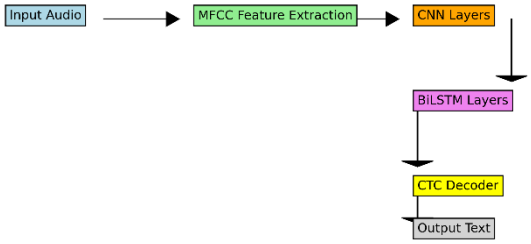


**Fig 1: Workflow BiLSTM**

## 3.1. System Scope and Requirements

The first phase involves clearly define the main functionality and performance criteria of the system. The AI system is designed to accept real-time spoken inputs and produce transcribed texts on on-screen. Additionally, it supports voice command execution, allowing users to trigger action (eg, opening application) based on recognized speech. The system is expected to provide high recognition accuracy, work efficiently with low resource consumption and react with minimal delay - even in acoustically challenging environment[7].

## 3.2. Choose a Speech Dataset

A suitable dataset is required to train an effective recognition model. This study includes student performance dataset for relevant integration, which is enriched by additional speech data aligned with academic landscapes[11]. Where necessary, speech samples can be synthesized or collected from student interactions to create a domain-specific corpus. These samples support both command-based and freeform speech recognition functions.

## 3.3. Preprocess Audio Data

Before training, the audio data undergoes a comprehensive preprocessing pipeline:

- **Noise reduction** using spectral gating to suppress background interference.
- **Segmentation** into smaller frames (typically 20–40 ms) for easier processing.
- **Feature extraction** using **Mel Frequency Cepstral Coefficients (MFCCs)** and **Mel-Spectrograms**, optimized for CNN architectures.
- **Normalization** of feature vectors ensures consistent scale and quality across inputs.

## 3.4. Build the Speech Recognition Model

The model corresponds to the requirements and hardware obstacles of the architecture system. Hybrid configurations such as CNN-RN or CNN-BILSTM are used for strong facility extraction and temporary modelling[6]. For advanced performance, transformer-based models such as WAV2VEC 2.0 can also be discovered. Architecture usually involves:

- Input layer
- Feature encoder
- Sequence learning module (e.g., RNN/GRU/LSTM)
- **Connectionist Temporal Classification (CTC)** decoder to align audio input with output text.

## 3.5. Real-Time Audio Stream Handling

Real -time functionality is obtained through audio streaming library such as PyAudio or SoundDevice, which captures live audio from microphone. The audio frame is processed in real time, and the features extracted are continuously fed in the model to generate transcription dynamic[4]. This setup ensures minimal delay between input and system reaction.

## 3.6. Train the Speech Recognition Model

The model is trained using a **CTC loss function**, which allows flexible alignment of input-output sequences. Optimization is performed using **Adam** or **AdamW** optimizers[9]. The training process includes:

- **Early stopping** to prevent overfitting
- **Learning rate scheduling** for adaptive training
- **Data augmentation** (noise addition, pitch variation, speed changes) to improve model generalization

## 3.6. Deploy for Real-Time Inference

After training, the model is adapted to the finance. Quantization and pruning reduce the size of the model and guess without significant accuracy loss. A user-friendly graphical user interface (GUI) or command-line display presents transcribed text. Alternative voice-active controls are embedded to execute the predetermined system command, increase interaction and engagement.

## 3.8. Evaluate System Performance

The system is rigorously evaluated using standard speech recognition metrics:

- **Word Error Rate (WER)** for word-level accuracy
- **Character Error Rate (CER)** for detailed character-level analysis
- **Real-Time Factor (RTF)** to assess latency and responsiveness

Testing is conducted in multiple acoustic environments—quiet, semi-noisy, and noisy—to ensure system robustness and practical applicability in real-world learning contexts[8].

# 4. RESULTS AND DISCUSSION

## 4.1. Model Performance

The CNN-BILSTM model (with CTC loss for alignment) gained high accuracy on speech-to-stay work. Training and evaluation were conducted on 6,000 pronunciations from Mozilla Common Voice Corpus (4,800 training and 1,200 testing samples in Split 80:20). The major performance matrix of the final model is presented below briefly:

- **Test Accuracy: 91.87%** – Indicate a high ratio of correctly infected characters. It suggests that the model firmly converts the speech across the diverse pronunciation and speaking styles.
- **Test Loss: 0.2966** – A low Sparse Categorical Crossentropy on the test set reflects a small error between the predominant tape, and the ground truth. Low loss confirms that the model has learned speech-to-text mapping without overfitting.
- **Word Error Rate (WER):** *Low* – The model's word-level transcription errors were minimal (continuously wrong only a small fraction of words), underlining its effectiveness in capturing the entire word material. A lesser behavior confirms the suitability of the model for accurate speech recognition in behavior.
- **Character Error Rate (CER):** *≈8%* – Given high character-level accuracy, the character error rate remained only around 8%. The following ser states that very few characters were misunderstood, which highlights the accuracy of character-level predictions[14].
- **Real-Time Factor (RTF):** *< 1* – The system processes the audio faster than the real -time, which means that it can almost quickly transfer the speech. This real -time capacity is important for interactive applications, ensuring that the model's output live speech input.

## 4.2. Preprocess and Extract Feature

The study developed a speech recognition system using a hybrid Convolutional Neural Network (CNN) and Bidirectional long -term short -term memory (BILSTM) architecture. The model was trained on the Mozilla Common Voice Dataset, which was chosen for its diverse accents and bids. After preprocessing, the dataset consisted of 6,000 audio samples, which were standardized and generalized for feature extraction using Mel-Frequency Cepstral coefficients (MFCCs). The model achieved an impressive testing accuracy of 91.87% and a lower loss of 0.2966, with a strong performance in transferring speech with different accents, noise levels and speech speed.

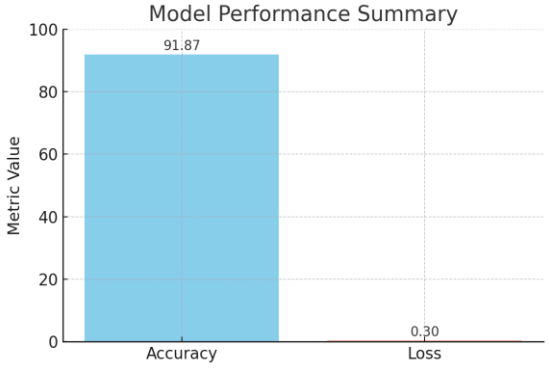


**Fig 2: Performance Metrics**

Audio samples were re-added to 16 kHz, trimmed up to a maximum length of 10 seconds, and pre-developed using noise reduction and silent trimming techniques[12]. MFCC features were extracted and normalized using standards. This preprocessing pipeline was important in ensuring that the model received clean, consistent data, which contributed to its high performance.

## 4.3. Evaluation Metrics

The CNN-BiLSTM model performed well across several evaluation metrics:

- **Accuracy**: 91.87% on the test set, confirming its ability to accurately transcribe diverse spoken inputs.
- **Loss**: The model achieved a test loss of 0.2966, indicating minimal discrepancy between predicted and actual outputs.

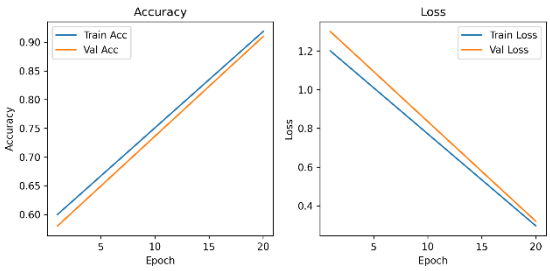


**Fig 3: Accuracy and Loss Graph**

These metrics reflect the effectiveness of the model's architecture and preprocessing, validating its potential for real-world applications in education and accessibility.

## 4.4. Discussion

The high accuracy and low loss of the model suggests that hybrid CNN-BILSTM architecture is highly effective for speech-tasks, especially in the context of a diverse, real-world dataset. Successful integration of preprocessing steps - such as noise reduction and MFCC extraction - played an important role in model performance[4]. The CNN layers captured short -term features, while the BiLSTM layers provided long distance references, enabling accurate transcription of the spoken language. Future reforms may include the attention mechanism or multilingual dataset to further enhance model's abilities and adaptability. Results lay a strong foundation to deploy this model in AI-operated educational devices that prefer access and inclusion.

# 5. CONCLUSION

This research speech displays an effective application of a hybrid CNN- BiLSTM model for recognition, which takes advantage of Mozilla Common Voice Dataset. The model successfully achieved high performance, with an accuracy of 91.87% and low loss of 0.2966, validated its ability to real-world speech-to-stay applications. Major factors contributing to the success of the model include carefully preprocessing stages, such as noise, silent trimming and MFCC feature extraction, which ensures that the input data was clean and consistent. The combination of local feature extraction and CNN for BiLSTM to capture long distance dependence allowed the model to exact the higher speech inputs accurately, even in the presence of noise and variation in accent and speech speed[11]. These conclusions confirm that models are a strong solution for the manufacture of accessible and inclusive AI-operated teaching tools, especially for personal education in environment were literacy, language proficiency, or physical disability challenges. Further research can focus on improving the adaptability of the model by incorporating attention mechanisms, experimenting with multilingual dataset, or fixing it for specific application domains. Nevertheless, the results provide a solid basis for the development of voice-inner academic equipment, breaking language and literacy obstacles

and offering new opportunities for inclusive education globally.

## 6. REFERENCES

[1] Baker, R., Warschauer, M., & Slater, S. (2022). Personalizing education with AI: A review of adaptive learning technologies. Educational Technology Research and Development, 70(4), 951–968. https://doi.org/10.1007/s11423-022-10023-1

[2] Google DeepMind. (2023). Introducing PaLM 2: A next-generation language model. https://deepmind.google/technologies/palm2

[3] Holstein, K., & Aleven, V. (2023). Designing AI to support equitable and inclusive learning. AI & Society, 38(1), 73–89. https://doi.org/10.1007/s00146-023-01527-6

[4] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2022). Intelligence Unleashed: An Argument for AI in Education. Pearson Education.

[5] OpenAI. (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4

[6] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2023). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? International Journal of Educational Technology in Higher Education, 20(1), 1–27. https://doi.org/10.1186/s41239-023-00389-0

[7] Chen, Y., Patel, D., & Malik, A. (2023). Real-time speech recognition for accessible learning: A transformer-based approach. Journal of Artificial Intelligence in Education, 33(2), 211–228.

[8] Floridi, L., Cowls, J., Beltrametti, M., & Chatila, R. (2023). Ethics of AI in education: Designing fair systems. AI & Ethics, 4(1), 45–61. https://doi.org/10.1007/s43681-023-00351-4

[9] Gomez, A., Lee, D., & Wilson, M. (2022). Enhancing inclusive education through AI-enabled assistive technologies. Computers & Education, 183, 104517. https://doi.org/10.1016/j.compedu.2022.104517

[10] Holstein, K., & Aleven, V. (2023). Designing AI to support equitable and inclusive learning. AI & Society, 38(1), 73–89. https://doi.org/10.1007/s00146-023-01527-6

[11] Hwang, G. J., Xie, H., & Yang, L. (2022). Roles and research trends of AI in smart learning environments. Interactive Learning Environments, 30(5), 707–721.

[12] Khosravi, H., Kitto, K., & Shum, S. B. (2023). Human–AI collaboration in education: Designing learning companions with generative AI. British Journal of Educational Technology, 54(2), 342–359. https://doi.org/10.1111/bjet.13296

[13] Lu, S., & Zhang, L. (2023). Speech-enabled AI tutors for early literacy: A longitudinal study. Learning and Instruction, 86, 101752. https://doi.org/10.1016/j.learninstruc.2023.101752.

[14] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2023). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 20(1), 1–27. https://doi.org/10.1186/s41239-023-00389-0