International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-025-02529-9

ORIGINAL ARTICLE

AI-driven mock interview assessment: leveraging generative language models for automated evaluation

Padma Jyothi Uppalapati[1,2] · Madhavi Dabbiru[3] · Venkata Rao Kasukurthi[4]

Received: 6 June 2024 / Accepted: 2 January 2025

Abstract
In the education sector, adaptive support is critical for every student to face open-ended activities that need behavioral change, performance, and a pro-active learning mindset. This can be accomplished by using brilliant learning environments powered by artificial intelligence. Timely feedback is critical for helping students enhance their overall personality in learning, confidence, communication, and problem-solving. It is a highly demanding task for every teacher to conduct a mock interview and then provide feedback for each criterion. It is time consuming and may be delayed. However automated review of mock interviews can give timely student feedback while reducing the manual evaluation burden on teachers in areas with a high teacher-to-student ratio. Current ways of analyzing student interview responses include transformer-based natural language processing models, which have various degrees of effectiveness. One major problem in training these models the need of more data, as most of the datasets are based on HR queries, which have sufficient datasets. But, none of the interview recordings included both HR and TR-related questions. We recorded mock interviews with undergraduate students from multiple backgrounds to address the data scarcity issue. We extracted audio from the recordings, followed by transcripts that included speaker identification. We split the speakers' data into questions and responses. The biggest challenge is evaluating these answers, given the need for appropriate datasets for technical questions. This article investigates the text-generating AI model, GPT-3.5, to establish whether prompt-based text-generation approaches are viable for generating scores for specific student responses. Finally model results are compared with human values. Our results reveal that the pre-trained model yields excellent outcomes in interview grading.

Keywords Videos of mock interview · Generated language models · Automatic speech recognition · Speaker diarization · Technical skills · WhisperX

1 Introduction

According to a World Economic Forum report, the EdTech market is expected to grow $404 billion by 2025 from $227 billion in 2020. Furthermore, a McKinsey analysis found that online education platforms experienced a 400% increase in utilization during the peak of the COVID-19 pandemic [1]. During the pandemic, colleges such as Harvard and Stanford shifted to completely online classes using platforms such as Zoom, Coursera, and edX, while K-12 schools widely utilized systems such as Google Classroom and Khan Academy to provide educational continuity despite the closure of physical classrooms. Even most universities like Stanford use Artificial Intelligence (AI) to provide feedback to teaching practices, evaluate assignments, and assess student communication and body language [2]. This shows

✉ Padma Jyothi Uppalapati
padmajyothi64@gmail.com

1 Department of CS&SE, Andhra University Trans-Disciplinary Research Hub, Andhra University College of Engineering (AUCE(A)), Visakhapatnam, AP 530003, India

2 Department of CSE, Vishnu Institute of Technology, Bhimavaram, AP 534202, India

3 Department of CSE, Dr.Lankapalli Bullayya College, Visakhapatnam, AP 530013, India

4 Department of CS&SE, Andhra University College of Engineering (AUCE(A)), Visakhapatnam, AP 530003, India

Published online: 25 January 2025

how quickly educational institutions adopted digital tools, emphasizing technology's critical role in modern education.

AI based education research is progressing correctly and benefiting all stakeholders [3]. It implies AI tools are being developed and applied correctly, enhancing efficiency, personalization, and access for students, educators, and administrators. AI-powered Mock Interview Grading (MIG) task alleviates the need for human evaluation of students' responses in video format, which is a challenge in educational research as it takes a lot of time to grade every student mock interview individually. MIG is not a new problem and it is labor-intensive. Teachers personally award each student's response a score based on performance and other aspects of the interview process. It is a supervised learning problem categorized as either regression or classification [4].

During talent recruitment, interviews are crucial for identifying individuals that align with the corporate environment [5]. Students from rural backgrounds mainly need help in attempting the direct job recruitment process. Therefore, the mock interview process helps undergraduates to practice the virtual interview inorder to increase confidence. To that aim, many institutions are looking for alternate ways to reduce manual processes and human effort in the assessment by implementing automation, by utilizing machine learning and advanced technologies.

The mock interview grading task has many developments, from traditional machine learning to neural network models. While handling MIG tasks, there is a special requirement to handle both HR and TR related questions. Therefore the recordings should include specific domains like technical content, which includes courses like Databases, Programming languages like Python and Java, and many more, structures, Networks, and others. As most of the existing datasets' main drawback is the need for more technical content-related questions in the interview process, their main focus is on human behavior, personal, professional, and other non-technical questions.

While MIG automation promises greater speed and consistency in assessment, it is likely to reduce the subjective, complex judgement that human evaluators bring to the process. This transition raises worries about educator deskilling and a decline in possibilities for tailored feedback, both of which are critical for student development. As a result, any shift to automated grading must carefully balance the benefits of technology against the need to preserve the human aspect in education, ensuring that evaluators are not replaced but rather supported by these technologies.

Furthermore, the transition to automation may have a considerable impact on the entire educational ecosystem. Traditional assessment methods, which rely on human interaction and judgment, may be eclipsed by the quest for more standardized, data-driven assessments. This could result in a more consistent, although potentially less comprehensive, approach to student assessment [6]. The ramifications for curriculum design, instructional practices, and the overall student experience must be carefully evaluated in order to avoid unforeseen outcomes that could harm educational quality. Thus, it is necessary to critically assess how automation will affect the future of educational practices and ensure that it improves rather than detracts from the learning process.

This work offers a fresh perspective that advances our knowledge of the planning, creation, and use of AI-based systems for practice interviews in higher education research. In the field of education, data collection even involves the usage of AI-based solutions [7, 8]. Adoption of modern technologies is critical for resolving MIG in this dimension. This work aims to develop an efficient system for assessing student performance in the interview process by utilizing Pre-trained Language Models(PLM) and Speaker diarization techniques.

Within the huge field of computerized language processing, a revolutionary phenomenon known as Large Language Models (LLMs) has evolved, holding enormous power in its ability to comprehend linguistic patterns and provide coherent and suitable responses [9]. LLMs can also be referred as Generative Language Models (GLMs), LLMs research focuses on educational competencies such as mathematics, writing, programming, reasoning, and knowledge-based subjective and objective answering, with the goal of exploring their potential in the development of next-generation intelligent education system. GLMs which are used in education such as assessment of short answer grading, summary of the essay, quiz assessment, providing technical content, providing the code for a given problem, and can be used to grade the student in various aspects.

One of the most difficult aspects of the MIG is dealing with student responses to technical questions, as there is no appropriate dataset to evaluate those responses. LLMs can handle this issue using their domain knowledge. LLMs are trained on vast datasets across multiple domains and then fine-tuned with task-specific datasets. Fine-tuning LLMs requires minimum data compared to pretraining from scratch [10].

The critical challenge of this work is to evaluate the mock interview video for every question based on benchmark rubrics. Finally, the model is fine-tuned based on the requirement; a summation of all rubric values will give a particular rubric's score, which was rated between 0 and 10. Finally, based on the score of every rubric, the guided summary is generated by the model to specify the students' performance based on every rubric. The main contributions of the work are:

- Collection of New Dataset: Conduct mock interviews with undergraduate students using expert faculty, covering both technical and HR-related questions.
- Identification of Rubrics: Determine specific rubrics by comparing them with benchmark labels.
- Conversion of Video to Audio Files: Use established benchmark strategies to convert video recordings into audio files.
- Generation of Transcripts Based on Speaker: Utilize benchmark strategies to generate transcripts from audio recordings, distinguishing between different speakers.
- Generation of Scores for All Questions: Grade each question using the defined rubrics to produce scores, which are then used to evaluate overall performance.

The structure of the paper is as follows: Sect. 2 provides an overview of the academic work conducted in the field of Mock Interview Grading (MIG). Section 3 outlines the details of the proposed pipeline. Section 4 describes the dataset used, while Sect. 5 elaborates on the rubrics developed for the study. Section 6 presents the experimentation details, and Sect. 7 evaluates the results. Future work and discussions are addressed in Sect. 8, and the conclusions are summarized in Sect. 9.

2 Related work

This part begins by introducing the work that has occurred in the field of Interview Grading, then quotes on how research development has changed the task landscape, and concludes with the previous work on the generation of speaker-based transcripts from a video.

Generative Pre-trained Transformer (GPT) is a powerful AI language model developed by OpenAI that generates human-like text based on given prompts. It is pre-trained on vast amounts of data and uses transformer architecture to understand and produce coherent, context-aware responses. GPT is one of the application of GLM. GPT is widely used for tasks like text generation, summarization, translation, and more. Figure 1 illustrates the evolution of versions of GPT from 1 to 4 [11].

Initially, GPT-1 introduced the transformer architecture and trained the model with 0.12 billion tokens. Each subsequent version, GPT-2 with its increased parameter count, improved its text generation with 2.5 billion tokens, whereas GPT-3, with its 175 billion parameters enabling few-shot learning, and GPT-3.5 with its 1800 billion tokens refined performance, gradually pushed the boundaries of what these models can achieve. While GPT-3.5 Turbo is built for speed and cost-efficiency and employs the same parameters as GPT 3.5, GPT-4 is more advanced, offering better accuracy, deeper understanding, and the ability to handle complex jobs with 13,000 parameters. GPT-3.5, with its multimodal capabilities and better contextual awareness, is the most recent advancement, making LLMs more adaptable, reliable, and effective in providing tailored and adaptive educational experiences [12]. We chose GPT-3.5 Turbo for mock interview assessment due to its performance and efficiency, that provides fast and contextually relevant feedback at a lower cost compared to GPT-4. Its robust understanding of language and ability to evaluate technical responses made it ideal for our rubric-based evaluations. Additionally, its optimized processing time allows for timely assessments, enhancing the overall user experience.

Natural Language Processing (NLP) is a subfield of AI that is gaining attention in research due to the introduction of many applications [13]. Automatic interview grading research has taken on a new direction due to advances in NLP. Researchers use traditional machine learning and deep learning techniques to address the Interview grading task. The traditional machine learning algorithm generally uses different features to perform the MIG task, including prosodic features (speech intonation and rhythm), lexical features (information about interview content and speaking style), and facial features (smile intensity, body position, and head shaking) and then to predict the performance of the student in an interview different Regression and Classification techniques were applied such Support Vector Regression,

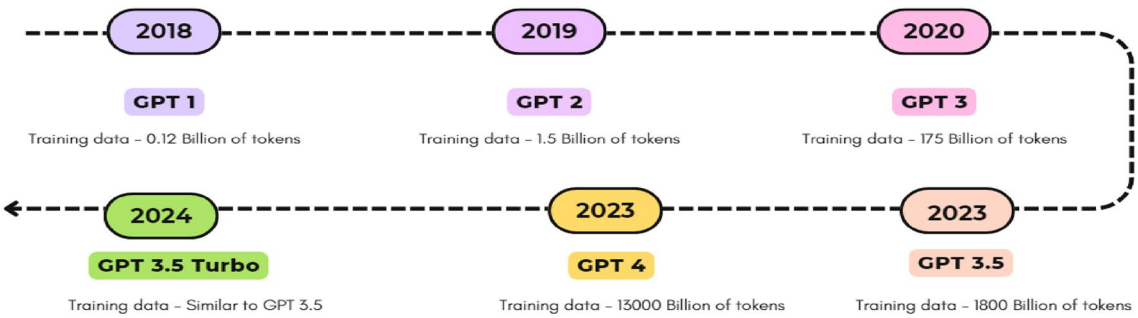

Fig. 1 LLMs by volume of training data and release date

Random Forest Regressor and also Logistic Regression [14–16].

Deep Learning-based techniques are currently outperforming traditional and ML-based approaches. Long Short Term Memory (LSTM) networks, Bi-directional LSTMs, Transformers, Siamese Networks, Transfer Learning, Autoencoders, and Attention are some of the deep learning methodologies used by educational institutions in recent years. Kai Chen et al. [17] tackled the task of automatically rating candidates' competency based on textual attributes extracted from automatic speech recognition transcriptions during asynchronous video job interviews. They presented a sentence-level relational graph neural network to capture phrase dependence information within or between the inquiry and the answer, which has yielded excellent results.

Amina Aman et al. [6] worked on Multimodal Performance Analysis during job interviews in order to implement text-based analysis, i.e., analyzing the text-based responses of the student in the interview questions; they used LSTM as the content is a collection of words which is structured format. They fine-tuned the model to obtain excellent results.

Hemamou et al. proposed an attention model called HireNet [18], mainly used to predict the chance of hiring candidates, similar to the recruiters' evaluation. Two different contextual information are modelled in the HireNet model: words in the question and words in the job position. This model performs better compared to the existing models. Zadeh et al. [19] proposed a novel architecture called Multi-attention Recurrent Network (MARN) mainly for understanding communication between humans and humans. The main strength of this work is that it uses a multi-attention block for storing modalities through the timestamp and a recurrent component called short-term hybrid memory, which are used for storing the memory of the sequence data.

Wasifur Rahman et al. [20] make good use of BERT-based models. They found that fine-tuning the trained contextual models on task-specific datasets was critical to getting improved performance downstream when grading replies in low-resource conditions. They used BERT and XLNet to obtain excellent results in the field of NLP.

Augmentation creates various versions of the original data and improves overall performance of the model. Like a few vision and text augmentation, even a few audio augmentation techniques help improve the audio data quality. There is no standard approach to augmentation in NLP task; In recent years, audio data augmentation has become widely employed in a variety of applications to increase the performance of NLP tasks such as automatic voice recognition, text-to-speech conversion, and speech recognition. Several techniques in audio augmentation are volume controller, noise reduction, pitch shift, adding background noise, speed perturbation and many more [29, 30].

An essential aspect of technology that lets computers automatically identify human voices from audio signals is Automated Speech Recognition (ASR) [31]. New methodologies and algorithms have led to a significant growth in the creation of ASR models in recent years. ASR models are evolving as deep learning, transfer learning, multimodal integration, and optimization approaches progress, allowing for more accurate and resilient speech recognition systems across a wide range of domains and applications [32, 33].

Speaker Diarization assigns speaker labels to speech regions in the given recording of audio or video. It is one of the most important technique in NLP task while working with audio and videos. which serves as a preprocessing pipeline for downstream tasks such as multi-speaker automatic speech recognition [34] in various scenarios, such as new broadcasting [35], meeting conversations [36], telephonic conversation [37], and clinical diagnosis.

The task of speaker diarization can be implemented using various concepts of deep learning such as Deep Neural Networks, LSTM [38], BiLSTM (Bidirectional Long Short-Term Memory) Networks [39], CNN (Convolutional Neural Network [40], End-to-End Neural Diarization and transformers [41]. Various benchmark models are available for speaker diarization tasks, such as pyannote [42], Nividia Nemo [43], WhisperX [44], and JHU [45], Lium [46] speaker diarization, yield excellent results in the identification of the speaker in the segmentation process.

Table 1 illustrates a comparative summary of various research studies focused on the evaluation of interview datasets, conducted over the years from 2016 to 2023. The studies utilize different datasets, including the MIT Interview Dataset, First Impressions V2 Dataset, GLUE Dataset, and custom datasets created by the researchers themselves. These datasets serve as the foundation for machine learning and deep learning models aimed at automating interview evaluation tasks.

Gaps identified from literature survery are—a fundamental problem is the need for domain-specific datasets, particularly those with high technical content, making it challenging to accurately analyse responses to technical questions. The evaluation process is further complicated for more sophisticated rubric-based approaches, making capturing the complex nature of interview responses and remarkably open-ended questions difficult. General MIG merely delivers a score based on rubrics, but feedback systems still need to be improved in the current works. Most of the works are on deep learning and machine learning models based on features, audio, and video clip-based assessment. Labelling of data is a very challenging task in the MIG dataset. In addition to earlier works, the present work investigates and analyzes the effectiveness of augmentation techniques in improving interview audio, identifying structured answers

Table 1 Researches done on interview datasets

References	Year	Dataset	Model	Rubrics	Accuracy
Naim et al. [14]	2016	MIT interview dataset	Lasso, SVR	Overall, recommend hiring, engagement, excitement, eye contact, smile, friendliness, speaking rate, no fillers, paused, authentic, calm, focused, structured answers, not stressed not awkward	0.80
Mujtaba et al. [21]	2021	First impressions v2 dataset	Multi-task deep neural network (MTDNN)	Openness, conscientiousness, extraversion, agreeableness, neuroticism (OCEAN), job interview recommendation	0.913
Kaya et al. [22]	2019	First impressions v2 dataset	Extreme learning machine (ELM) classifiers	Openness, conscientiousness, extraversion, agreeableness, neuroticism (OCEAN), job interview recommendation	0.917
Li et al. [23]	2020	First impressions v2 dataset	deep classification-regression network (CR-Net)	Openness, conscientiousness, extraversion, agreeableness, neuroticism (OCEAN), job interview recommendation	0.918
Agrawal et al. [24]	2020	MIT interview dataset	Random forest classifier, SVC, Multitask Lasso, MLP	Eye contact, speaking rate, engaged, pauses, calmness, not stressed, focused, authentic, not awkward	0.74
Chopra et al. [25]	2020	MIT interview dataset	SVR, KNN, decision tree	Friendly, engaged, excited, speaking rate, calm	0.72
Rick Somers et al. [26]	2021	GLUE dataset	ELECTRA, RoBERTa BERT, MobileBERT, XLNet, ALBERT, XLM	Accuracy	0.91
Amina et al. [6]	2023	MIT interview dataset	Random forest regressor, fusion model (CNN + LSTM)	Eye contact, speaking rate, engaged, pauses, calmness, not stressed, focused, authentic, not awkward	0.90
Sri Latha et al. [27]	2023	Collected technical questions	Senternce-BERT	Accuracy based on actual answers and responses	0.72
Isaac et al. [28]	2023	Own dataset	RoBERTa	Accuracy based on actual answers and responses	0.74

by detecting the speaker, and finally evaluating the questions and answers using advanced models.

3 Workflow

The pipeline of the work is illustrated in the Fig. 2. Initially, for undergraduate students, we conducted mock interviews on technical and non-technical aspects, and the complete process was recorded to capture both audio and video data. The next step is preprocessing and extraction of the content from the video as per the requirements. Initially audio is extracted from video with the libraries moviepy[1] and soundfile.[2] In order to transcribe the audio, automatic speech recognition was applied, and then speaker diarization was applied to the audio and transcripts to label the text of speaker identification. Finally, a file is generated as in the processing, which consists of questions posed by Speaker 1 and answers given by Speaker 2. Then, this text data is provided to the fine-tuned pre-trained model along with guided rubrics to predict the score of every rubric. Then, the evaluation of the model is done by correlating it with the human rating. If it is positively correlated near to 1, then the process is stopped. Finally, a summary report is generated based on the performance of the student which covers the strengths and weakness of a student in a interview process.

[1] https://pypi.org/project/moviepy/.

[2] https://pypi.org/project/soundfile/.

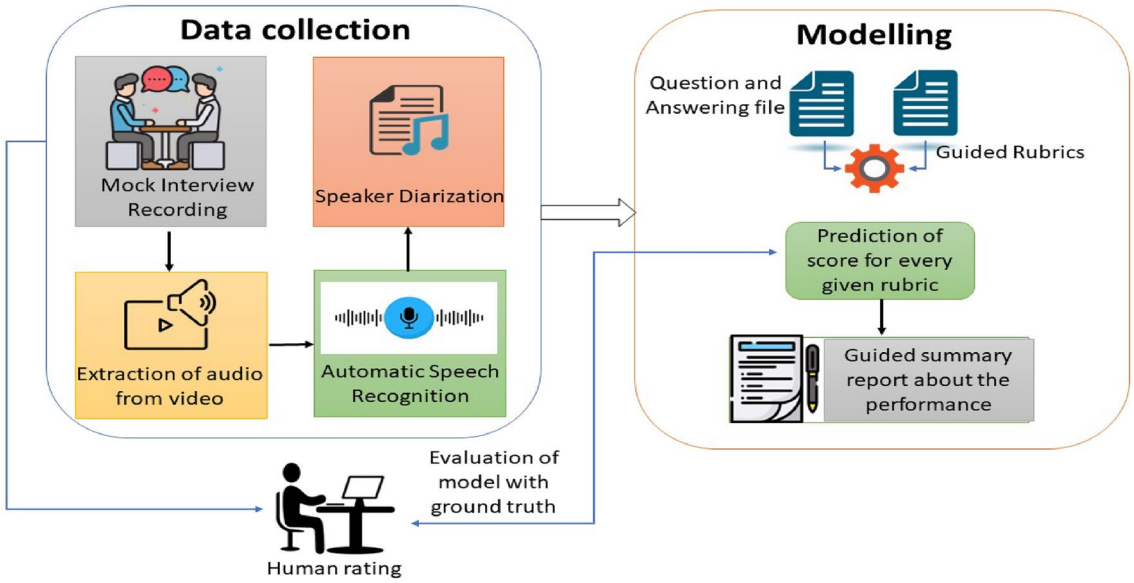

Fig. 2 Methodology of the work

4 Dataset

We collected our dataset, which consists of 238 audio–video recordings of the mock interview of 123 students based on the undergraduates who want to face the actual interview. Based on the survey, we found few datasets on the mock interview process, but the main drawback is that ascent can be easily understood for most existing models. Existing datasets cover interview questions mainly in human resource-related, personal, career, academic, and professional questions. Most of the datasets have a drawback: they do not cover technical aspects like engineering, science, or medical background. This dataset mainly focuses on the communication, slang, and confidence levels of students from rural backgrounds. It also covers the entire interview in a way that is similar to a real-life interview, i.e., it is based on technical and non-technical aspects, human behaviour, and personal and related questions. Automation based on rural language slang is entirely different from that of foreigners.

The mock interview process was conducted with third-year, second-semester students from various engineering streams, including Computer Science, Artificial Intelligence, Electrical, Electronics, Mechanical, and Civil. Students were selected based on their academic performance, with a minimum CGPA of 6.5 and up to 2 backlogs. Before the interviews, each student was given a brief introduction outlining the structure and purpose of the interview, which covered both technical and non-technical subjects relevant to their respective fields. The interviews were conducted in a controlled environment, where each student was positioned in front of a camera, simulating a real-life interview setup. The interviewer, an experienced faculty member, posed a series of questions designed to assess the student's knowledge and communication skills. This standardized process was followed for all participants to ensure consistency across the mock interviews.

The Table 2 illustrates the demographic distribution of the participants shows a slight male majority at 55% (66 males) compared to 45% females (54 females). The Computer Science and Engineering (CSE) program leads with 34.17% (41 participants), followed by Artificial Intelligence and Data Science at 21.67% (26 participants). Other notable representations include Electronics and Communication Engineering (ECE) with 15% (18 participants) and Civil Engineering with 5.83% (7 participants).

The Fig. 3 illustrates sample mock interview videos. Initial questions are based on personal profile and behavioral questions, while the remaining are based on technical and other questions. The total duration of the mock interview videos is 17 h and 4 min. On average, the duration of the video is 5.2 min. The mock interviews were conducted by subject experts who have 10 years of experience in the teaching field. The interviews were conducted with the same set of undergraduates to check for improvement in their performance. We recorded the complete interview process from the beginning to automate the assessment process.

Unlike most existing datasets that primarily focus on personal, career, behavioral, open-ended, Stress-Type, organization, academic, and other HR-related questions, our new dataset is comprehensive. It covers not only non-technical skills but also technical skills like problem-solving, programming languages, data analysis, and knowledge in different domains like data structure, databases, networks, etc. Additionally, it includes non-technical skills like the emotion

Table 2 Demographics of the students participated in the mock interview

Demographic category	Count	Percentage (%)
Gender		
Female	54	45.00
Male	66	55.00
Total	120	100.00
Course		
Computer science and engineering (CSE)	41	34.17
Artificial intelligence and data science (AI&DS)	26	21.67
Artificial intelligence and machine learning (AIML)	18	15.00
Electrical and electronics engineering (EEE)	5	4.17
Electronics and communication engineering (ECE)	18	15.00
Civil engineering (civil)	7	5.83
Mechanical engineering (mech)	5	4.17
Total	120	100.00

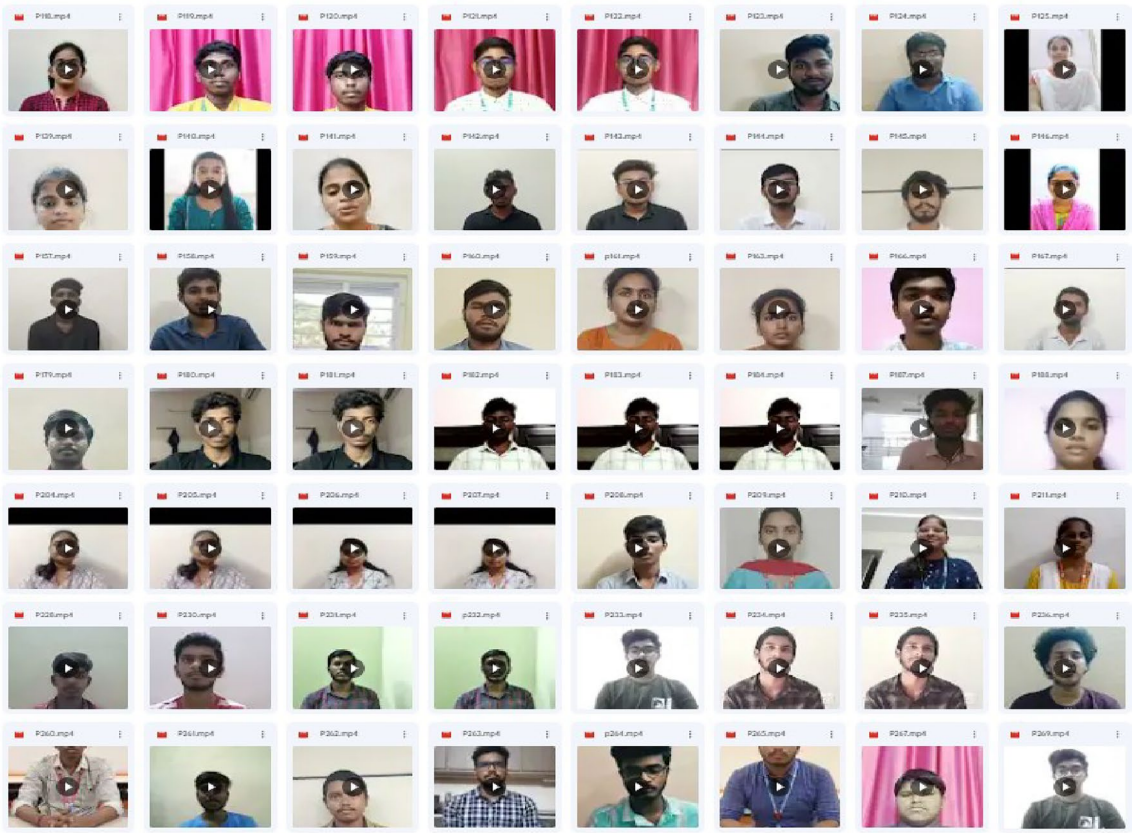

Fig. 3 Sample recorded mock interview dataset

of the candidate, leadership quality, presentation skills etc., verbal skills like grammar, ability to share thoughts, listening skills, etc., and nonverbal skills like body language, gestures, facial expressions, posture, and eye contact. This comprehensive coverage ensures a thorough assessment of the student's performance.

Table 3 provides a glimpse into the practical application of our interview dataset. It features a sample question covering engineering students' technical skills for a software job. These mock interviews, designed explicitly for Bachelor of Technology (B.Tech) graduates, are not just theoretical exercises. They are simulated practice interviews that aim to prepare individuals who have completed their B.Tech degree

Table 3 Sample question of our interview dataset

Type of the question	Sample questions
Personal assessment questions	Introduce yourself
	What are your greatest strengths and weaknesses?
Company/organization questions	Why do you want to work for (company/organization)?
	Are you willing to work overtime?
Education and experience	Describe your achievements since you have started college
	Explain about the project that you have done recently
Career ambition questions	What is your short term and long-term goals?
	Why should I hire you
Technical questions	What is linear Data Structure?
	What is the purpose of final keyword in java
	Explain concept of encapsulation with real life example
	Purpose of CPU Scheduling algorithms

for actual job interviews in their chosen technology field. This practical approach ensures that the skills assessed are directly applicable to real-world scenarios.

5 Rubrics

During the hiring process, candidates are assessed using a methodical framework called the Interview Rubric. It offers a uniform set of standards together with a grading system to evaluate applicants' abilities, credentials, and cultural fit. An essential component of task assessment is the use of rubrics. It comes in rather useful for maintaining uniformity, particularly when multiple individuals are conducting interviews with the same applicant. It all comes down to using data, choosing wisely, and being fair [47, 48]. The four main categories of rubrics are task-specific, generic, analytical, and holistic. The following primary factors are often regarded as crucial and ought to ideally be included in any evaluation of an interview based on the benchmark rubrics as shown in Table 1.

- *Clarity:* Usually used to assess a candidate's ability to express themselves clearly during an interview. It evaluates a number of skills, including the applicant's capacity for clear and concise response, organizing the information logically, acceptable language use, and the ability to explain difficult ideas simply.
- *Correctness:* Assesses the correctness and relevancy of the candidate's responses. High scores indicate exact, factual responses and an extensive understanding of the topic area. Lower scores imply more frequent inaccuracies or incomplete replies.
- *Authenticity:* Being authentic is putting up a truthful and real show without embellishing or pretension. The most important things to take into account are the candidate's transparency, consistency, self-awareness, cultural fit, and non-verbal cues. If hired, an authentic candidate is more likely to establish rapport with interviewers, show that they are genuinely interested in the position, and enhance the culture of the company.
- *Professionalism:* Behaving in a way that complies with accepted norms for behavior in the workplace, such as acceptable dress, manner, and communication. Three things that must be taken into account are appearance, preparedness, courtesy, and communication. A candidate that scores highly on the professionalism criteria is more likely to communicate with coworkers and clients in an effective manner, behave oneself professionally at work, and enhance the organization's reputation.
- *Structured Answers:* Providing well-thought-out answers that logically and fully address the question or subject at hand. Candidates can demonstrate their communication abilities and leave a good impression on the interviewer by practising structured responses.
- *Focused:* Throughout the interview, remain on subject and refrain from digressions or unrelated material.You may make a good impression on the interviewer by staying focused throughout the process and clearly communicating your skills, experiences, and interest in the role.
- *No Fillers:* Reducing the amount of superfluous words or phrases that could impede communication's efficacy and clarity. "Um," "uh," "like," "you know," and related expressions are a few examples of popular fillers. Cutting back on fillers can make you appear more certain, polished, and professional in interviews and other communication settings.

Together, these criteria ensure that the applicant interacts with the interviewer courteously and professionally and successfully conveys their background, skills, and suitability for the job position.

6 Experiment setup

6.1 Video to audio conversion

The initial task in this system is to extract audio signals from the given recorded videos, and those audio files should be of high quality. The main reason for extracting audio from the system is that this work mainly focuses on the student's performance in the interview process based on their answering style, focus, correctness, fillers, and many more, but not on their visual appearance, eye contact, and others. Therefore, the mock interview videos in this work are analysed using audio scripts.

The complete process is done accordingly. Initially, the video will be automatically divided into some fragments and converted into audio signals, and these fragments will be joined together to get complete audio of the recorded video. This entire process will be completely automatic without human intervention. In order to perform the task of video to audio conversion-python libraries like moviepy and soundfile were used.

Moviepy is more of a general-purpose Python library for video processing and editing, but it also offers temporal and spatial manipulations, including the ability to alter video speed and perform trimming, concatenation, visual compositing, and spatial cropping [49]. Moviepy is capable of reading and writing the majority of audio and video formats. To convert the video from the.mp4 format to the matching MP3 or WAV audio format, Moviepy is needed in this work. Alternatives for the same purpose include opencv, librosa, pydub, and ffmpeg.

Moviepy's user-friendly interface makes it excellent for Python users who want advanced video editing and simple audio processing. OpenCV focuses on computer vision and real-time video processing, but lacks advanced video editing and audio features. Librosa focuses on advanced audio analysis and does not offer visual processing. Pydub is useful for audio editing but does not allow video processing. FFmpeg provides strong video and audio processing through a command-line interface. It is more versatile but less Python-centric [50].

Moviepy loads each frame into Python using ffmpeg, writes it to a new file, and then combines the material. Since Ffmpeg operates directly on the files, it would be much quicker than moviepy. An audio library built on top of CFFI, NumPy, and libsndfile is called python-soundfile. The functions read() and write() can be used directly to read and write sound files. Use blocks() to read a sound file block by block. As an alternative, SoundFile objects can be used to open sound files.

6.1.1 Data augmentation

The process of applying different transformations to increase the amount of training data available is known as data augmentation [51]. It changes an illustration x_i to $\tilde{x_i}$, which keeps the majority of x_i's characteristics. In audio processing, there is few predefined standard set of conventional augmentation techniques that yields state-of-the-art performance like Volume adjustment, Pitch shifting, Noise injection, Amplitude Scaling, Speed adjustment and many more. Furthermore, adding diversity to the data set through data augmentation can improve the model's accuracy and generalization. Implementing more augmentation on the dataset may also lead to bias in nature.

Algorithm 1 To implement augmentation techniques for audio

Require: Dataset X
Set of Augmented functions F_s
function AUGMENTED_DATA(X,F_s)
  for $x_i \in X$ **do**
    $\tilde{x_i} \leftarrow F_s(x_i)$
    $X_a \leftarrow \tilde{x_i}$
  end for
  $\tilde{X} \Leftarrow XUX_a$
  return $\tilde{X}$
end function

Augmented Techniques for Audio, addresses some of the shortcomings of the original and increases the quantity of training samples. It accepts a list of F_s = [V,P] and the training data X as inputs. Pitch and audio volume are increased in the enhanced techniques used. In this dataset, noise is not a problem. The issue with the original data is the extremely low number of recorded videos. Let x_i be the original data set, with i = 1 to n number of recorded videos. The recorded response is subjected to the augmented approach described in Algorithm 1 in order to provide augmentation data D_a.

6.2 Identification of interviewer and interviewee

To assess the student's performance in a mock interview video. The audio should be converted into a readable text format and then from the generated transcripts the questions and answers can be identified by the process of speaker diarization.

6.2.1 Automatic speech recognition

The technique of converting human speech or audio signals into a readable text format is called Speech recognition. A pre-trained model for speech translation and automated speech recognition (ASR) is called Whisper. Whisper models exhibit a great capacity to generalize to numerous datasets and domains without requiring fine-tuning, having been trained on 680k hours of labelled data.

Various evaluation metrics are used to check the performance of the ASR task, such as Word Error Rate (WER), character error rate, Levenshtein distance, and BLEU Score.

The Word Error Rate evaluation metric was employed as part of the ASR task to assess the performance of different versions of the Whisper model. By quantifying transcription errors, WER enabled the identification of the model version with the lowest error rate, ensuring improved accuracy in speaker identification. WER is derived from the Levenshtein distance. The three forms of errors are insertion, deletion, and substitution are added up and to the total number of spoken words form the WER [52]. When N words are included in the reference transcript and S substitutions, D deletions, and I insertions are found in the ASR output, then as in Eq. 1 and accuracy is measured as in 2:

$$WER = 100.\frac{(S+D+I)}{N}\% \tag{1}$$

$$Accuracy = 100 - WER\% \tag{2}$$

Although it is commonly accepted that a voice recognition system's word error rate has a significant impact on speech analytics systems performance. The value of WER ranges from 0 to 1. Lowering the WER value means it is the best model for ASR systems. Algorithm 2 addresses identifying the best model for the ASR task for the mock interview dataset. It accepts a list of whisper models, such as small, medium, large, and large V2, along with an input dataset. Each model performs ASR and converts it into transcripts. Finally, the generated transcripts of every model are compared with the ground truth and the performance metric WER. The lower WER value was finally chosen as the best model.

Algorithm 2 Best techniques for audio to text conversion

Require: Dataset X
Set of Whisper Models M
function BESTTECH_ATX(X,M)
  for $x_i \in X$ **do**
    for $M_j \in M$ **do**
      $Transcript_{ij} \leftarrow M_j(x_i)$
      $Score_{ij} \leftarrow Calculate\_WER(Transcript_{ij}, Groundtruth_i)$
    end for
  end for
  for $x_i \in X$ **do**
    for $M_j \in M$ **do**
      $Min\_score = Min(Score_{ij})$
    end for
  end for
  return $M_j \epsilon Min\_score$
end function

One more metric is Levenshtein Distance, which is used to check the performance of ASR tasks. It calculates the minimum character changes required to transform one string to another. Let X and Y are the strings with length m and n respectively.

Table 4 Various whisper model versions and its performance on the data

Size	Parameters (M)	English-only	Multilingual	Levenshtein distance	WER
Small	244	✓	✓	452	0.172
Medium	769	✓	✓	451	0.172
Large	1550	×	✓	306	0.173
Largev2	1550	×	✓	146	0.040

The Levenshtein distance matrix L is of size (m+1) × (n+1), where L[i][j] represents the distance between the substrings X and Y. The distance is calculated as in the Eq. 3. Table 4 illustrates the performance of different whisper model versions on the MIG dataset for the ASR task. Among all the models, the whisper large V2 model's performance is better than whisper small, medium, and large models with a WER 0.04. and also the Levenshtein distance value is low for whisper large V2 model. Therefore among all the models, whisper large v2 is best for the ASR task.

$$L[i][j] = \begin{cases} L[i-1][j]+1 & (deletion) \\ L[i][j-1]+1 & (insertion) \\ L[i-1][j-1]+\delta(x_i, y_j) & (substitution) \end{cases} \quad (3)$$

6.2.2 Speaker diarization

Speaker diarization mainly concerns "who spoke when" in audio or video recordings. The word "diarize" refers to taking a note of something or recording it in a Diary. One of the essential tasks in speech recognition is to identify the speaker spoken at which timestamp [53].

There are various bench-mark speaker diarization techniques available. Most of the speaker diarization faces challenge with long audio files. In order to address that challenge WhisperX Speaker diarization is used to identify interviewer and interviewee. WhisperX [44] is a system for efficient voice transcription of long-form audio with precise word-level timestamping. The WhisperX model is compared to prior cutting-edge work in speech transcription, namely the original Whisper large-v2 model and the Wav2Vec2.0 base model. The WhisperX developers used a variety of evaluation criteria, including WER, speed, precision, and recall. WhisperX now has VAD pre-processing to increase batching performance and reduce hallucinations [54]. Figure 4 illustrates the architecture of WhisperX.

The process of speaker diarization involves multiple steps. Initially, the audio signals undergo Voice Activity Detection (VAD) which identifies the voice activity, and then the complete audio is divided into sections. The next step is extracting features from the segments which is applied to small chunks of audio, then clustering them into batches based on similar features. Finally, each cluster is annotated with the labels as the identified speaker. Speaker Diarization is used to distinguish the questions and answers in the provided mock interview video i.e, interviewer and interviewee. If the diarization is inefficient, the video questions will be awkward, resulting in inaccurate findings. The whisperX model is utilized for this purpose because it is better suited for lengthier videos.

6.3 Generative language models in assessment

Generative language models are neural networks trained mostly on linguistic data obtained from the Internet. GLMs are based on the generative pre-trained transformer model (GPT). The usage of GLMs in education is part of a more significant trend toward digitization and technology integration into instruction. This trend has been accelerated by the COVID-19 epidemic, which necessitated remote learning and reliance on online resources. Using GLMs is an essential component of this trend since it can promote individualized learning, foster critical thinking, and increase evidence-based thinking in education [55]. GPT 3.5 Turbo is an improved version of GPT, fine-tuned to improve language understanding and text creation capabilities. It is highly customizable and also domain specific. Turbo GPT employs a process known as fine-tuning, which involves training the model on domain data. This allows the model to understand the complexities and terminology connected with a subject, which improves its performance in that area.

Figure 5 illustrates the architecture of the GPT 3.5 Turbo, it implements the Transformer architecture—based on self-attention mechanisms. It comprises 96 transformer layers, each containing 16 heads for multi-head attention. Therefore, it requires 96 stacks of decoder blocks with multi-head attention blocks. This section explains how GPT 3.5 Turbo architecture is used to evaluate question and answer grading. Text embeddings convert input tokens into dense vectors. Each word or character in the question and answer is assigned to a vector space with related meanings clustered together. It helps to comprehend the context and meaning of words in connection to one another. The Transformer model fails to understand token order; therefore, position embeddings are added to provide information about the location of each token in the sequence, which is required for context-aware evaluation.

Masked multi-head self-attention assists the model in determining the significance of each component of the response to the inquiry. It guarantees that each token is evaluated using the correct context provided by the preceding tokens, which is critical for determining the response's alignment with the question. Layer normalization

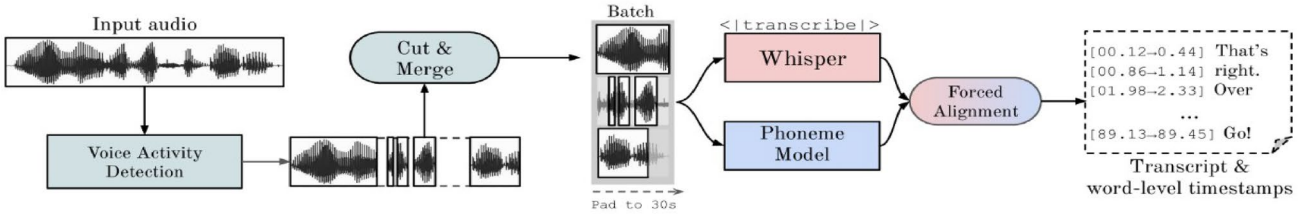

Fig. 4 WhisperX architecture

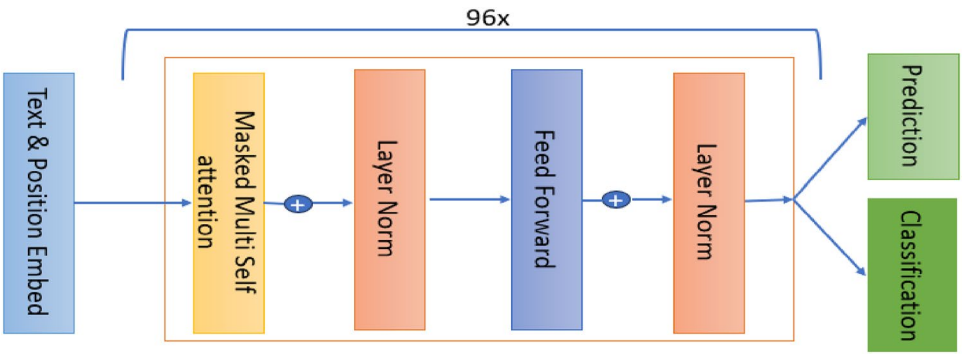

Fig. 5 Architecture of GPT 3.5 Turbo

guarantees that prior layers' activation remains stable, allowing for more consistent and reliable representations of the question and answer. This stability is crucial for accurate evaluation based on rubrics.

The feedforward network helps in determining how well the response corresponds with the question by transforming information into relevant features derived from contextual representations formed by the self-attention mechanism. The final output of the feedforward network is used to generate the model's predictions or evaluations based on the rubrics. These methods work together to ensure that GPT-3.5 Turbo can accurately evaluate the quality, relevance, and coherence of responses in accordance with the assessment criteria.

6.3.1 Prompt engineering

LLMs are currently not used for automatic interview grading by any of the researchers. It necessitates costly resources, which are unaffordable for companies with limited budgets and resources. LLMs are commonly utilized for interview grading through fine-tuning and prompt engineering. This work mainly focuses on prompt engineering [56].

Prompt engineering is critical for successfully deploying large language models, particularly in education [57]. It entails designing the input or prompt in a way that guides the model to produce the intended outcome. Prompt engineering has various applications in education, such as automated question generation in English, guidance for academic writers, code review automation, automatic learning, evaluating Student Text Responses, improving coding skills, and many others [58–60].

Types of available prompt engineering are zero-shot, one-shot, few-shot prompts, prompting levels, structured prompts, iterative prompts, and bad prompts. This work is more suitable for zero-shot and few-shot prompt engineering. The "zero" in "zero-shot" indicates that the LLM has no specialized training on the question in the prompt. "Shot" signifies giving the LLM an example, but "zero-shot" indicates that it was not trained with labeled data to complete the task and that the prompt itself does not provide an example for the LLM to work. However, due to its vast language training, it can generalize and produce an outcome without task-specific training. Few-shot prompts are similar to zero-shot prompts in that the LLM has not been adequately trained to complete the required action. However, the prompt includes few examples to help the LLM to understand the request. Figure 6 illustrates the examples of different kind of shot prompting that are used in the work.

Initially, We worked with zero-shot prompting, as it allows a model to predict scores for questions and responses by leveraging knowledge from other related tasks without any labelled score for the responses. Later, we trained the model with a few shot prompting by providing examples for the scoring criteria. It is a time-consuming task, as we must give prompts for different questions. As questions in the MIG task belongs to different domains. However, as per

Fig. 6 Zero-shot, one-shot and few shot examples with question, answer and paramters

the study, few shot prompting yields excellent results when using GPT 3.5 Turbo.

7 Results

7.1 Generation of scores and summary

To apply inference technique (such as zero-shot learning and few-shot learning) in GLM, we first create prompt templates for the zero and few-shot prompting inference technique using open-AI guidance. Figure 7 illustrates the shot inference technique applied on GLM for generation of scores. Initially, a prompt for the GLM is designed based on specific properties, a template for the design is developed, and then a prompt is created by considering the rubrics. Then, the prompt is fed as instruction for the language model. Finally, the language model is fed with the collection of transcripts, a dataset. Based on the available knowledge, the model evaluates the transcripts by considering the rubrics. Here, the language model is GPT 3.5 Turbo, which accepts the instructions and input to yield the output, i.e., rubrics scores such as relevance, professionalism, structured answers, clarity, and many more for each and every question in a given transcript. Based on the scores, a feedback is generated by the model which explains the improvement areas and also highlight strengths.

Figure 8 illustrates a step-by-step guide on how zero-shot prompts are created and evaluated. Initially, the user will provide the mock interview recording. Based on the recording, the transcripts are generated using various techniques, and after speaker diarization, the questions and answers for every video are stored in a CSV file. For every question and response, the prompt is applied to parameters like relevance, fluency, clarity, etc. In the output layer, the model generates a score concerning every parameter. The model is meant to grasp the task simply based on the prompt and its prior knowledge, and then assign scores to each parameter accordingly and finally generates feedback summary [61] on the performance of student that covers the strengths and highlights the improvements. In the same manner, few-shot promopting will work.

A mock interview was conducted by a faculty member with a student to practice before Central recruiter team conducted an interview. Feedback from the faculty plays a

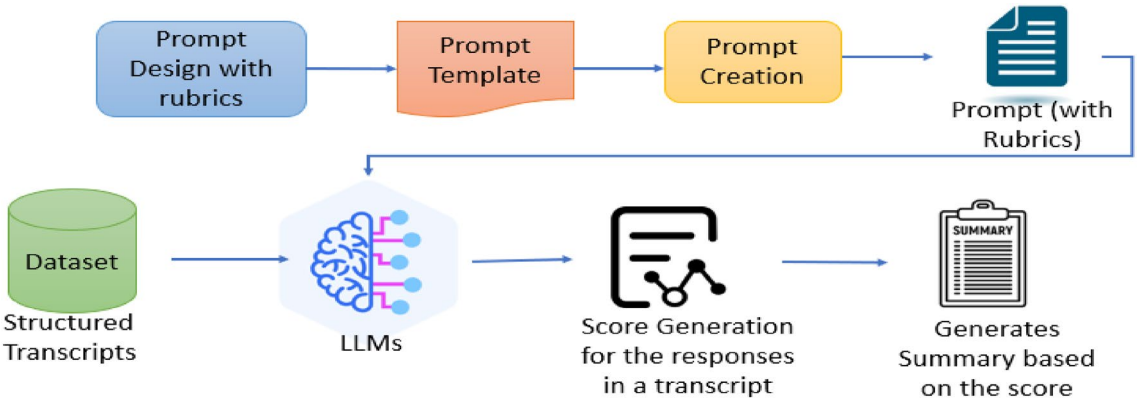

Fig. 7 Procedure for generation of scores

crucial role in the overall development of students. Therefore, in order to avoid the burden and to give clear feedback, we automated the feedback system in the following manner: Every question in the transcript was graded based on the benchmark rubrics. Finally, a student's performance is evaluated by the average of all question scores of a particular rubric. This is done by giving guidance to the GLM called GPT 3.5 turbo. Along with specified rubrics, instructions were given that every rubric value ranges from 0 to 100 based on the answer to a question. Therefore, the final rubric values are in the given range. Table 5 illustrates a student's performance assessment in a mock interview recording using various benchmark rubrics that most of the recruitment team considers. Student Performance is assessed based on 2 models. First row of values indicate zero-shot prompt based LLM and second row of values indicates few-shot prompt based model. Based on the observation, few-shot prompt based model are giving high rubric values compare to zero-shot model. As in few-shot prompt, a clear picture is provided to the model that how scoring need to be provided to a specific question.

In our mock interview assessment system, a generative language model (GPT-3.5 Turbo) is employed to generate a performance summary based on the scores obtained across various rubrics, such as fluency, technical correctness, professionalism, and other metrics. After scoring the candidate on these rubrics, the model synthesizes this data to produce a detailed summary that highlights the candidate's strengths and weaknesses as shown in the Fig. 9.

The prompt has been carefully designed to demonstrate the model to deliver individualized feedback that targets the candidate's strengths, flaws, and areas for progress with specific examples. The prompt specifies that feedback should highlight interview flaws, such as lack of confidence or poor preparation, and include actionable suggestions. For example, if the model detects a lack of confidence due to insufficient preparation, it provides advise such as practice the interview questions in front of a mirror to build self-assurance and improve delivery. Similarly, if the candidate facing difficulty with maintaining eye contact or avoiding filler words, the feedback includes illustrative examples of these flaws observed during the interview and offers precise recommendations, such as practising structured answers or engaging in mock interview sessions. This personalised approach guarantees that the student has a clear awareness of their deficiencies and receives practical help on how to solve them effectively.

7.2 Human rating

With the assistance of three teaching assistants, we were able to impartially examine each and every mock interview in order to determine how well the Generative Language model performed on the interview grading system. These teaching assistants are well-versed in all areas of computer science and engineering, as well as behavioral, professional, and human resources topics.

Each teaching assistant graded the interviewees' performances by answering 10 evaluation questions, on a scale of 1 to 10, after watching each recording interview tape. To simplify the estimation problem, we assume that the assistant ratings are real numbers, i.e., $y_i^j \epsilon R$. The evaluation questions show how well each student performed overall, as demonstrated in the video. The subsequent questions have been designed to evaluate a number of high-level behavioral traits, such as the candidate's presence (e.g., concentration, and participation in the video), speaking rate, speaking ability, substance (e.g., organization), and warmth (e.g., smiling, friendliness). The ability to pause and rewind the video gave the teaching assistants the opportunity to rate more thoroughly and impartially. Selecting assistants over specialists is preferable since the ratings of assistants are more likely to match those of the audience and interview panel.

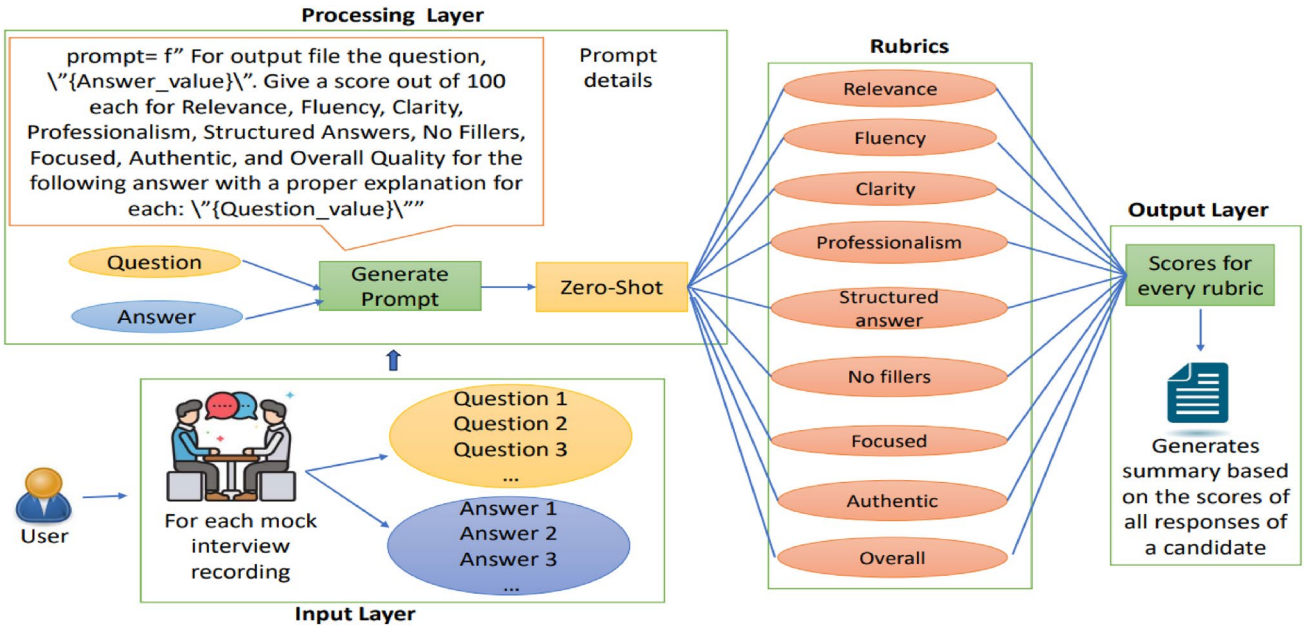

Fig. 8 Step by step procedure: zero-shot prompting for MIG task

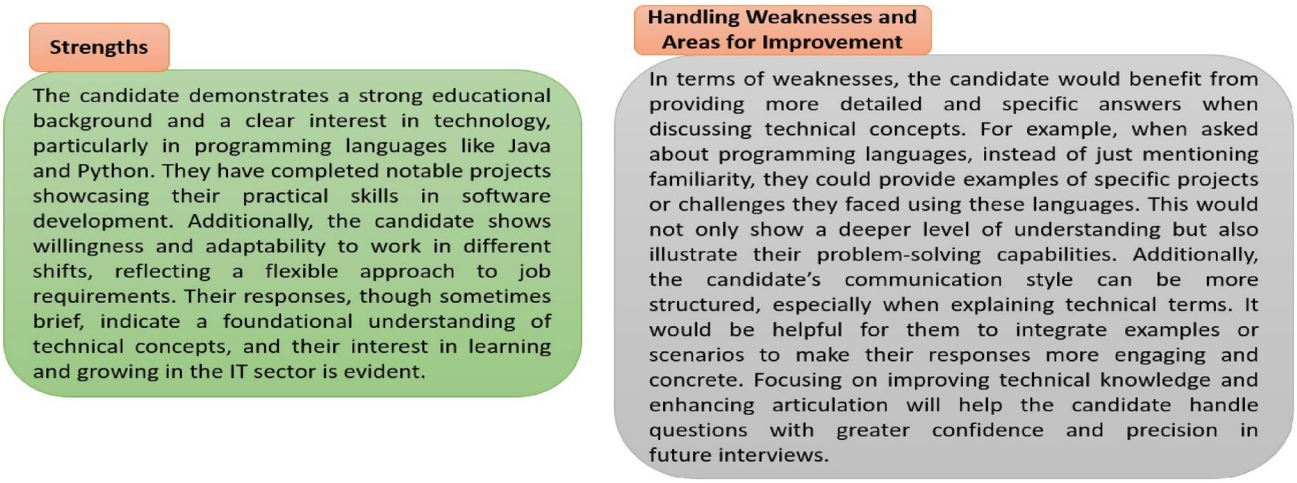

Fig. 9 Summary generated by the model based on the scores

7.3 Validation of ground truth

To validate the ground truth, which is the human rating, Intra-Rater Reliability (IRR) and Inter-Rater Reliability can be used. IRR was used to measure the consistency of a single human evaluator's grading across time. It assesses how the evaluator uses the same criteria to evaluate similar performances in various situations. A high intra-rater reliability shows that the evaluator consistently applies the same standards, regardless of when the evaluation is conducted. This consistency is critical in establishing a reliable ground truth because it limits the possibility of variability caused by evaluator fatigue, mood, or variations in interpretation of the evaluation rubric. If intra-rater reliability is low, it indicates that the evaluator's assessments may be inconsistent, which might affect the accuracy and fairness of the evaluation. The Fig. 10 shows that IRR scores for one evaluator which proved that ground truth is reliable.

Inter-Rater Reliability is a metric that assesses the consistency of different human evaluators while rating the same set of individuals. Cohen's Kappa is a widely used statistic for calculating IRR that takes into consideration the probability of agreement arising by chance. Cohen's Kappa provides a number ranging from − 1 to 1, with 1

Table 5 Rubric scores generated by LLM for a sample video transcript

Interviewer	Interviewee	Relevance	Clarity	Correctness	Structured answers	Fluency	Professionalism	No fillers	Focused	Authentic	Overall
Tell me about yourself	Good afternoon, ma'am. I'm Thadala Veerabargiri from West Godavari. Currently, I'm pursuing my B.Tech final year in Vishnu Institute of Technology with CGPA 9.12. I have completed my intermediate in Sasi Junior College with 96% and I completed my 10th with 93% in Nagarjuna School. Coming to my family, my family includes my father, who is horticulturist, my mother, who is homemaker, and my sister currently pursuing her intermediate. And I love playing badminton. I have good knowledge in Java and Python, and I'm also interested in machine learning. I have completed two decent projects called screen recording application using Python and exploratory data analysis on engineering placements	80 85	70 80	85 90	75 80	75 75	70 80	80 90	75 85	70 85	75 82
What is a static variable?	Static variable can be used for 3 types, it can be used for class, method and variable. If we use static for variable, the variable cannot be changed and if we use static for method, the method cannot be overridden and if we use it as a class, the class cannot be inherited	85 90	65 60	75 70	65 65	60 70	70 70	75 70	70 65	70 75	70 62
What are different aggregation functions in SQL?	Max, min, sum, count	90 95	75 85	80 95	70 90	70 85	75 90	80 95	75 90	75 85	75 89
How we can implement descending order in SQL?	We can implement by using order by	85 85	65 70	70 75	65 65	60 60	70 70	75 80	70 75	70 80	70 74
Explain polymorphism with real life examples	Polymorphism means having many forms ma'am. For example, water has many forms, it is like solid, gas and liquid	90 85	60 70	70 80	65 60	55 60	65 70	70 65	65 70	65 75	68 72
Where do you want to watch yourself in the next 5 years?	I want to become a software professional	80 85	60 70	70 75	65 70	60 65	65 75	70 85	65 70	70 80	68 75

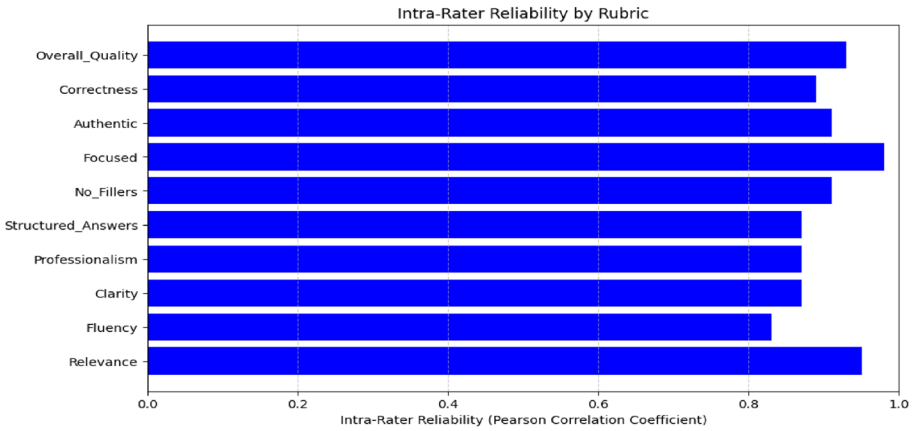

Fig. 10 Intra-rater reliability (IRR) for a human evaluator

indicating perfect agreement, 0 indicating no agreement beyond chance, and negative values indicating disagreement [62]. Figure 11 illustrates the cohen's kappa agreement valuse between 2 human evaluators. Metrics like 'Structured Answers' and 'Overall Quality' have the highest agreement (Kappa = 0.9), indicating that these elements were consistently rated by raters. 'Clarity' and 'Authentic' likewise have high Kappa scores of 0.85. On the other hand, 'No Fillers' (Kappa = 0.6) and 'Professionalism' (Kappa = 0.65) have reasonable agreement, indicating some variation in how various raters evaluated these qualities. Overall, the results indicate that the evaluation procedure is reliable, indicating substantial to perfect agreement.

7.4 Comparison between human evaluators and the LLM model

Figure 12 illustrates the comparison of scores of all ten rubrics for a given mock interview video recording, which a teaching assistant and the GPT 3.5 model evaluate. Most of the rubrics scores are near each other, but there is a slighter difference in overall participation rubric which are evaluated by humans and GPT 3.5 model with zero-shot and few-shot prompting mechanism and compare to few-shot model values, the zero-shot model responses are less.

The correlation coefficients between the weighted average rating provided by teaching assistants and the ratings that the model predicts are how we gauge the accuracy of our predictions. The correlation is calculated to every pair of columns in the given data. The attributes are arranged according to the high values of their association coefficients. With a score of 0.82 for zero prompting and 0.90 correlation value for few shot prompting model and Human—the most significant score for interview decisions—we did well in the overall performance prediction task.

We might also anticipate that clarity and focused, fluency would result in greater correlation coefficients. It is important for us to note that the interview questions included in our training dataset are selected based on requirements found in job specifications in the Information Technology (IT) industry. Questions are collected from the main interview experiences of various graduates who have passed out. As a result, our model's scores are based on a variety of skills, including behavioral, technical, non-technical, and interpersonal skills.

The correlation between few-shot prompting and human evluations can also be represented with few more statistics such as p-value and confidence levels. A mean correlation coefficient or confidence intervals ranging from [0.86, 0.912]. Even at the lower end of 0.86, the correlation reflects a significant level of agreement, while the upper end of 0.912 suggests an even tighter alignment. Moreover, the p-value of 0.0191 highlights that this correlation is statistically significant, with less than a 2% likelihood that the observed relationship is due to random chance. This shows that, for all of these criteria, AI is capable of consistently replicating human judgement patterns, making it a reliable tool for assessments that require human-like evaluation.

Cohen's kappa is a popular machine–human agreement (MHA) statistical measure, which is used to determine how closely a AI model's predictions match with human evaluations. The Inter Annotator Agreement (IAA) measure can be used to determine annotation consistency. When working with ordinal data using weighted indices, the IAA metric can be derived using Cohen's Kappa reliability, which quantifies the agreement between two annotators. According to the Kappa reliability, the values between 0.61 and 0.80 are substantial Agreements and 0.81–0.99 are perfect Agreements [63].

Table 6 illustrates the Agreement between humans and the model for the MIG dataset is perfect for many parameters. In both Zero-shot and few shot-based prompting, there is substantial Agreement over parameters like structured answers, professionalism, and overall quality. The overall Agreement between Humans and GLM based on zero-shot prompting is 0.75, whereas the Agreement between Humans

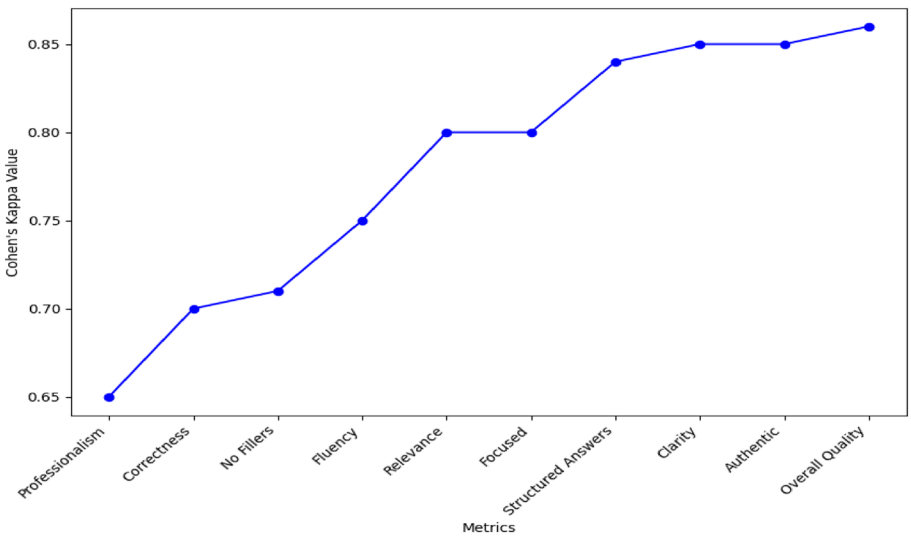

Fig. 11 Inter-rater reliability statistics between human evaluators

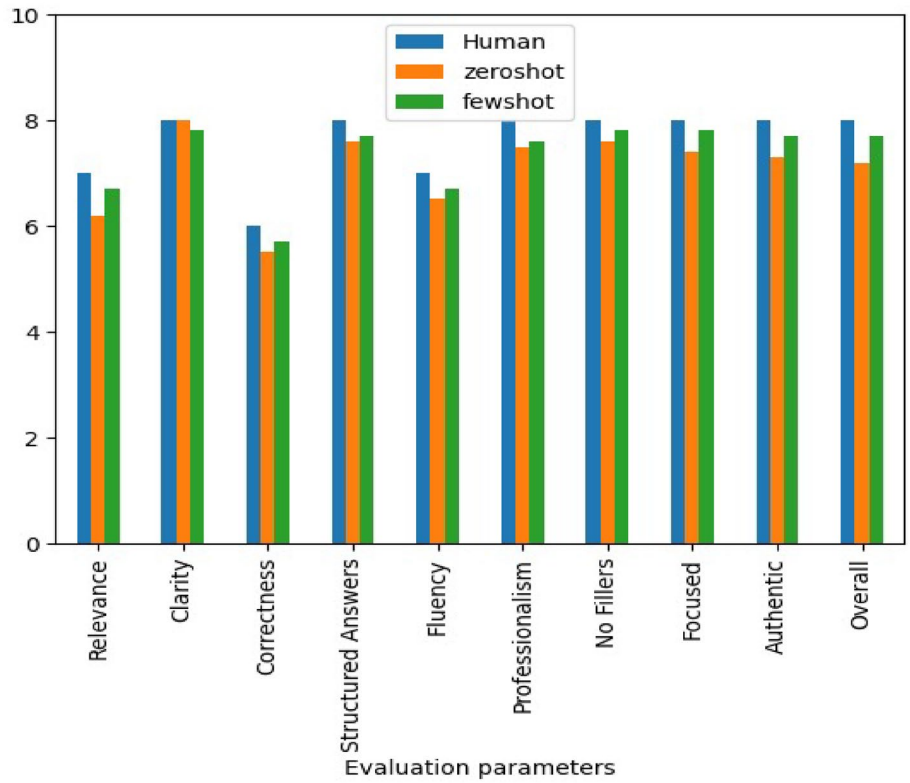

Fig. 12 Human versus AI-detailed interview response evaluation

and a few-shot-based GLM is 0.82. Therefore, the reliability between humans and few-shot prompting-based GLM is ideal.

7.5 Students and TA's perceptions on MIG survey

The student survey was conducted to gather feedback from participants who experienced both the traditional and automated mock interview grading processes. A questionnaire was carefully prepared to ensure it addressed key areas related to the functionality, clarity, and fairness of the automated grading system. The survey aimed to capture students understanding of the grading process and their overall satisfaction with both methods. To collect responses efficiently, the survey was administered through Google Forms, allowing for easy distribution and data collection. The questions were presented in a combination of

Table 6 Cohen Kappa's agreement between human and AI model with zero and few shot prompting

Parameter	Model_zero	Model_few
Relevance	0.71	0.85
Clarity	0.74	0.86
Correctness	0.86	0.86
Structured answers	0.61	0.72
Fluency	0.86	0.86
Professionalism	0.72	0.72
No fillers	0.85	0.86
Focused	0.86	0.87
Authentic	0.72	0.87
Overall	0.61	0.75

multiple-choice, Likert scale, and open-ended formats to gather both quantitative and qualitative insights.

The survey was completed by 84 out of a total of 100 students, for an 84% response rate. The survey contained questions designed to better understand students' perceptions of MIG. Students were asked which method they believed provided more accurate feedback on interview performance, as well as if automated grading methods could change the way students prepare for interviews and get feedback. Among those who were dissatisfied, significant complaints were questions about the accuracy and fairness of computerized grading.

An survey is undertaken to determine the students' degree of satisfaction with the technique used to grade interview performance. 79.5% of respondents agreed that this automation, along with associated methods, will have a significant transformation in the feedback process, while 18.1% are still unsure about the transformation, treating automation feedback similarly to teacher feedback, as depicted in Fig. 13.

Consistency refers to the reliability and uniformity of outcomes, when a MIG is done repeatedly on same candidate. while bias is about the fairness and impartiality of the results that are provided. Figure 14 illustrates the automation's consistency and biased nature; 69.9% respondents gave feedback that it is more consistent and less biased nature than the traditional mock interview, which the faculty carried out, and 19.3% respondents gave that both automatic grading and traditional grading are equally consistent and bias in nature. They utilized this automation process to improve their capabilities in terms of communication, confidence, focus, and also other technical skills.

The majority of respondents agreed with all of the questions, but almost 12% expressed confusion about the automated outcomes. The reason for this is that they are not good with prerequisites. They did not go through the automatic feedback mechanism thoroughly. The survey also collected textual comments about satisfaction about the automation. Some students were concerned that the system would be inconsistent or prejudiced while evaluating the same content for multiple times, while others believed that automated feedback lacked the individualized insights that human evaluators might provide. These concerns identify areas in which the automated grading system should be enhanced to better satisfy student expectations and requirements.

A survey was conducted with teaching assistants. In total, 12 participated in evaluating each interviewee's answers. Their direct experience with traditional and AI-assisted grading enables a detailed assessment of these approaches. According to their findings in the Fig. 15, 81.8% of TAs believe the automation process is more consistent. This uniformity is due to the AI's ability to apply identical standards to all evaluations, reducing human mistakes and subjective bias. As a result, AI grading assures that each candidate is rated on the same criteria, improving fairness and dependability in the evaluation process.

7.6 Model's strengths and weaknesses

Handling technical content in mock interview grading (MIG) tasks involves a number of issues due to the subject matter's complexity and specificity. One key problem is accurately assessing technical correctness and depth, which frequently need domain-specific knowledge that might vary greatly depending on the topic of study. Evaluators must not only grasp the technical information, but also assess the candidate's ability to apply that knowledge in a practical setting. This makes the grading process subjective because various evaluators may focus different aspects of a candidate's response, resulting in inconsistencies.

AI-powered mock interview grading has numerous significant benefits that are changing the way candidates prepare for job interviews. First, it saves time by automating the evaluation process, allowing for quick evaluations without the need for human reviewers. The AI system maintains discretion, ensuring that evaluations are fair and consistent for all users. Furthermore, the ability to use AI grading tools from anywhere makes it easier for applicants to practice and improve their skills. This paper proposes solving the issues by automating and standardizing the grading process using GPT-3.5 turbo, a sophisticated pre-trained model. GPT-3.5's thorough training on multiple datasets gives it a deep understanding of technical ideas from a variety of domains, allowing it to make consistent and objective evaluations. Using this pre-trained large language model, the machine can examine and grade technical responses with greater detail and accuracy than human graders alone. Furthermore, the model may be fine-tuned to adapt to specific domains or updated with fresh information, assuring its relevance and effectiveness in a rapidly evolving technological context.

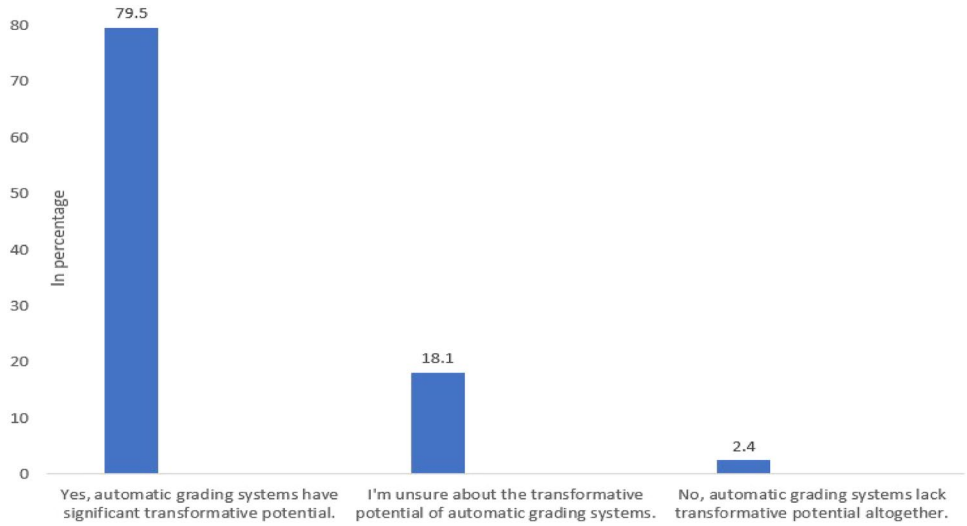

Fig. 13 Students opinion towards transformation

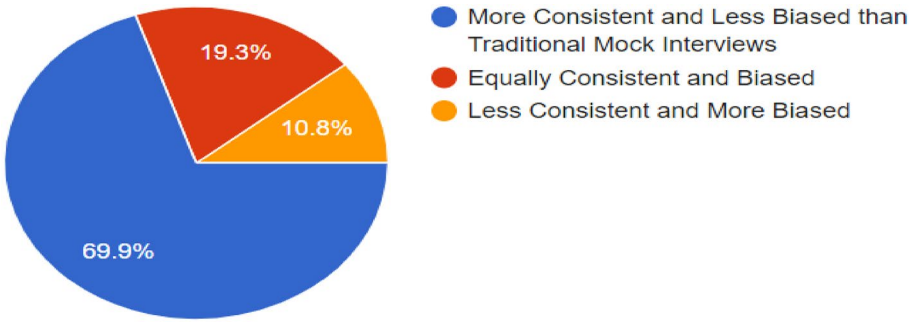

Fig. 14 Student response on consistency and bias nature

Finally, this approach has significant professional implications because it helps candidates develop their communication skills and prepare for real-world interviews, resulting in higher job possibilities.

8 Future work and discussion

In this study, the initial video recordings of the mock interview sessions were used as the primary source of input. From these videos, transcripts were generated through automatic speech recognition (ASR) and served as the core dataset for the assessment process. The focus of this research was on transcript-based evaluation, where the textual content of the interviews was analyzed to predict scores based on guided rubrics. While the video data provided additional information such as visual cues and audio dynamics, the current work relied solely on the transcribed text to develop the assessment model.

Along with GPT 1, 2, 3, 4 versions. Several different models have evolved, each with their own set of features and innovations. For example, Google's PaLM 2 and Meta's LLaMA are regarded as powerful alternatives, with each excelling in distinct elements of language processing, such as efficiency and domain-specific knowledge understanding. These models continue to push the frontiers of NLP, demonstrating that the landscape of LLMs is diverse and changing at an incredible rate [64]. Future work may include benchmarking our system against these LLMs to evaluate their performance in a variety of tasks, particularly automated assessments which is presented in this article.

The automation of mock interview evaluation provides us numerous possibilities for future research. One interesting area is to improve rubric customization and adaptation to accommodate different sectors and interview styles. The system can support a broader range of use cases by creating modular rubrics that test specialized skills and domain-specific knowledge. Similarly, including multimodal analysis which combines audio, textual, and visual data could considerably increase the evaluation's comprehensiveness by include non-verbal communication components such as facial expressions and body language, which are important in interviews.

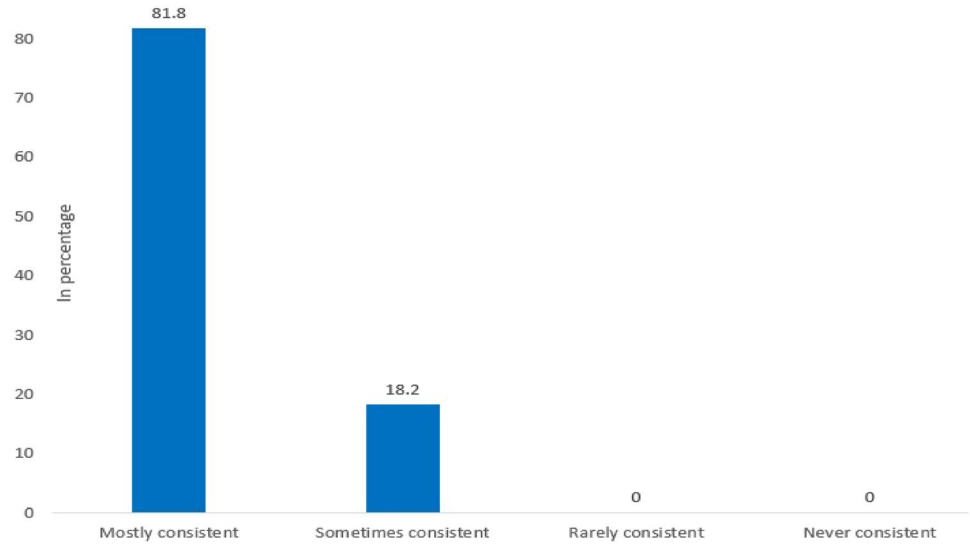

Fig. 15 TAs responses on automation consistency

Another important topic for research is optimizing technology for greater accessibility. Investigating lightweight model versions, such as those created using quantization approaches, may reduce computational needs, making the system practical for universities with limited resources. Furthermore, applying the system to bilingual and cross-cultural settings would broaden its global usefulness. Addressing cultural variations and analysing accents or non-native speaker communication would broaden the technology's applicability and versatility. Real-time feedback methods could also be investigated, with instant suggestions given during mock interviews to promote rapid progress in abilities such as filler word elimination and vocal modulation. Furthermore, collaboration with industry specialists might help match scoring and feedback with real-world hiring processes, while gamification features like badges and leaderboards could boost student engagement. Future research that expands on these areas can progress the discipline and develop a positive atmosphere for automated interview evaluation.

9 Conclusion

This work proposes a method for scoring mechanism-related mock interview responses by leveraging different AI techniques. The proposed method mainly experiments with transcripts, for which we retrieved audio from video and transcripts from audio. From transcripts, the speaker was to identify questions and answers. WhisperX speaker diarization is yielding excellent results. Then, those question-and-answer CSV files are fed to the GPT-3.5 turbo model for grading every question in the given file based on various characteristics. Our results suggest that OpenAI's GPT-3.5 Turbo may be adapted to grade the question and answer based on available data from the web. We compared the results of the GPT model with the human evaluation. It is observed that the model is yielding excellent results. Our work, which presents a real-world case study of LLMs' outstanding performance on interview grading for multiple rubrics, provides valuable insights into prompt engineering and the precise prompt tuning required to complete specific jobs. We provide our findings, and the resulting prompt is utilized to construct an overall summary of the student performance during the mock interview process. As a result, students can recognize their weaknesses and develop themselves in all areas.

In the future, advancements in prompt engineering also provides excellent results. Few features that can be included are interactive prompting, contextual and also fine-tuning prompting. GLM can be integrated with Learning Management Systems (LMS) to streamline the evaluation process significantly, which gives immediate feedback to students and allows teachers to monitor student progress and identify weaknesses. Also, the automatic grading of spoken interviews is made possible by combining real-time speech-to-text transcription systems with GLMs. This enables precise evaluation of the interview's clarity, fluency, and lack of fillers. Furthermore, to improve the educational insights by combining data analytics with GLMs, student performance can be visualized in the dashboards that illustrate trends in interview performance, supporting curriculum development and giving students individualized coaching. These developments produce more effective, impartial, and customized learning environments.

Acknowledgements I want to thank everyone involved in this initiative. I'd like to thank my Head of the Department (Dr. Srinivasa Raju), Dr. Abhinav Dayal, Dr. Sridevi Bonthu and all my students, who helped me in this work.

Author contributions 1. Have made a substantial contribution to the concept or design of the article; or the acquisition, analysis, or interpretation of data for the article and also drafting; AND. 2. Revised it critically for important intellectual content; AND. 3. Approved the version to be published.

Data availability The data that support the findings of this study are available upon request. For further inquiries regarding data availability, please contact Padma Jyothi Uppalapati at padmajyothi64@gmail.com

Declarations

Conflict of interest The study was not supported by any funding. All authors declare that they have no Conflict of interest.

Ethical approval This article contain studies with human participants performed by any of the authors. (a) Students participated in the process of mock interview. (b) Approved by Head of the department of the institution. (c) Inorder to improve the performance of the students in interviews, institution rigorously conducts mock interviews. It part of every faculty responsibility in some institutions.

Informed consent The students were provided with detailed information about the research objectives, procedures, and potential uses of their data. They were assured of the confidentiality and ethical handling of their information throughout the research process.

References

1. Global EdTech Market (2020). https://www.holoniq.com/notes/global-education-technology-market-to-reach-404b-by-2025 [Online; accessed 17-Oct-2024]
2. AI Tools in Teaching and Learning (2023). https://teachingcommons.stanford.edu/news/ai-tools-teaching-and-learning [Online; accessed 20-Oct-2024]
3. Bonthu S, Sree SR, Prasad MK (2023) Improving the performance of automatic short answer grading using transfer learning and augmentation. Eng Appl Artif Intell 123:106292
4. Singhania A, Unnam A, Aggarwal V (2020) Grading video interviews with fairness considerations. Preprint arXiv:2007.05461
5. Yusuf M, Lhaksmana KM (2020) An automated interview grading system in talent recruitment using SVM. In: 2020 3rd international conference on information and communications technology (ICOIACT). IEEE, pp 34–38
6. Aman A (2023) Multimodal performance analysis during job interviews. PhD thesis, Nazarbayev University
7. Cingillioglu I, Gal U, Prokhorov A (2024) Ai-experiments in education: an AI-driven randomized controlled trial for higher education research. Educ Inform Technol 2024:1–29
8. Caspari-Sadeghi S (2023) Learning assessment in the age of big data: learning analytics in higher education. Cogent Educ 10(1):2162697
9. Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh MB, Akhtar N, Wu J, Mirjalili S et al (2023) A survey on large language models: applications, challenges, limitations, and practical usage. Authorea Preprints
10. Grangier D, Iter D (2021) The trade-offs of domain adaptation for neural language models. Preprint arXiv:2109.10274
11. Li Q, Fu L, Zhang W, Chen X, Yu J, Xia W, Zhang W, Tang R, Yu Y (2024) Adapting large language models for education: foundational capabilities, potentials, and challenges. arXiv:2401.08664
12. Zhao B, Liu H, Liu Q, Qi W, Zhang W, Du J, Jin Y, Weng X (2024) Breaking boundaries in spinal surgery: GPT-4's quest to revolutionize surgical site infection management. J Infect Dis 403:1
13. Chapelle CA, Chung Y-R (2010) The promise of NLP and speech processing technologies in language assessment. Lang Test 27(3):301–315
14. Naim I, Tanveer MI, Gildea D, Hoque ME (2016) Automated analysis and prediction of job interview performance. IEEE Trans Affect Comput 9(2):191–204
15. Rasipuram S, Jayagopi DB (2016) Automatic assessment of communication skill in interface-based employment interviews using audio-visual cues. In: 2016 IEEE international conference on multimedia & expo workshops (ICMEW). IEEE, pp 1–6
16. Chen L, Zhao R, Leong CW, Lehman B, Feng G, Hoque ME (2017) Automated video interview judgment on a large-sized corpus collected online. In: 2017 seventh international conference on affective computing and intelligent interaction (ACII). IEEE, pp 504–509
17. Chen K, Niu M, Chen Q (2022) A hierarchical reasoning graph neural network for the automatic scoring of answer transcriptions in video job interviews. Int J Mach Learn Cybern 13(9):2507–2517
18. Hemamou L, Felhi G, Vandenbussche V, Martin J-C, Clavel C (2019) Hirenet: a hierarchical attention model for the automatic analysis of asynchronous video job interviews. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 573–581
19. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency L-P (2018) Multi-attention recurrent network for human communication comprehension. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
20. Rahman W, Hasan MK, Lee S, Zadeh A, Mao C, Morency L-P, Hoque E (2020) Integrating multimodal information in large pretrained transformers. In: Proceedings of the conference. association for computational linguistics, meeting, 2020, 2359. NIH Public Access
21. Mujtaba DF, Mahapatra NR (2021) Multi-task deep neural networks for multimodal personality trait prediction. In: 2021 international conference on computational science and computational intelligence (CSCI). IEEE, pp 85–91
22. Kaya H, Salah AA (2018) Multimodal personality trait analysis for explainable modeling of job interview decisions. In: Explainable and interpretable models in computer vision and machine learning, pp 255–275
23. Li Y, Wan J, Miao Q, Escalera S, Fang H, Chen H, Qi X, Guo G (2020) Cr-net: a deep classification-regression network for multimodal apparent personality analysis. Int J Comput Vision 128:2763–2780
24. Agrawal A, George RA, Ravi SS et al (2020) Leveraging multimodal behavioral analytics for automated job interview performance assessment and feedback. Preprint arXiv:2006.07909
25. Chopra S, Urolagin S (2020) Interview data analysis using machine learning techniques to predict personality traits. In: 2020 7th international conference on information technology trends (ITT). IEEE, pp 48–53
26. Somers R, Cunningham-Nelson S, Boles W (2021) Applying natural language processing to automatically assess student conceptual understanding from textual responses. Aust J Educ Technol 37(5):98–115
27. Latha CS, Kalyan NK, Abhilash J, Kaushik O, Balmiki V, Venukumar S (2023) Automated interview evaluation. In: E3S web of conferences, 430, 01026. EDP Sciences
28. Thompson I, Koenig N, Mracek DL, Tonidandel S (2023) Deep learning in employee selection: evaluation of algorithms to automate the scoring of open-ended assessments. J Bus Psychol 38(3):509–527

29. Maguolo G, Paci M, Nanni L, Bonan L (2021) Audiogmenter: a matlab toolbox for audio data augmentation. Applied Computing and Informatics (ahead-of-print)
30. Eklund V-V (2019) Data augmentation techniques for robust audio analysis. Master's thesis, Tampere University
31. Malik M, Malik MK, Mehmood K, Makhdoom I (2021) Automatic speech recognition: a survey. Multimed Tools Appl 80:9411–9457
32. Ai L, Soltangharaei V, Ziehl P (2021) Evaluation of ASR in concrete using acoustic emission and deep learning. Nucl Eng Des 380:111328
33. Sherly E, Pillai LG, Manohar K (2024) ASR models from conventional statistical models to transformers and transfer learning. In: Automatic speech recognition and translation for low resource languages, pp 69–112
34. Kanda N, Gaur Y, Wang X, Meng Z, Chen Z, Zhou T, Yoshioka T (2020) Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. Preprint arXiv:2006.10930
35. Tranter S, Reynolds DA (2004) Speaker diarisation for broadcast news. In: Odyssey04-the speaker and language recognition workshop
36. Yu F, Zhang S, Fu Y, Xie L, Zheng S, Du Z, Huang W, Guo P, Yan Z, Ma B et al (2022) M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6167–6171
37. Kenny P, Reynolds D, Castaldo F (2010) Diarization of telephone conversations using factor analysis. IEEE J Sel Top Signal Process 4(6):1059–1070
38. Wang Q, Downey C, Wan L, Mansfield PA, Moreno IL (2018) Speaker diarization with LSTM. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5239–5243
39. Lin Q, Yin R, Li M, Bredin H, Barras C (2019) LSTM based similarity measurement with spectral clustering for speaker diarization. Preprint arXiv:1907.10393
40. Cyrta P, Trzciński T, Stokowiec W (2017) Speaker diarization using deep recurrent convolutional neural networks for speaker embeddings. In: International conference on information systems architecture and technology. Springer, pp 107–117
41. Liu YC, Han E, Lee C, Stolcke A (2021) End-to-end neural diarization: from transformer to conformer. Preprint arXiv:2106.07167
42. Bredin H, Yin R, Coria JM, Gelly G, Korshunov P, Lavechin M, Fustes D, Titeux H, Bouaziz W, Gill M-P (2020) Pyannote, audio: neural building blocks for speaker diarization. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7124–7128
43. Park TJ, Kumar M, Narayanan S (2021) Multi-scale speaker diarization with neural affinity score fusion. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7173–7177
44. Bain M, Huh J, Han T, Zisserman A (2023) Whisperx: time-accurate speech transcription of long-form audio. Preprint arXiv:2303.00747
45. Huang Z, García-Perera LP, Villalba J, Povey D, Dehak N (2018) Jhu diarization system description. In: IberSPEECH, pp 236–239
46. Meignier S, Merlin T (2010) Lium spkdiarization: an open source toolkit for diarization. In: CMU SPUD workshop
47. Edminster S, Haruta C (2019) Assessment of students' interviewing skills
48. Crutsinger CA, Herrera U (2022) Mock interviews: leveraging AI resources to enhance professional skills. In: International textile and apparel association annual conference proceedings, 78. Iowa State University Digital Press
49. Kumar KS, Aravindhan S, Pavankumar K, Veeramuthuselvan T (2023) Autodubs: translating and dubbing videos. In: International conference on emerging trends in expert applications & security. Springer, pp 53–60
50. Amogh A, Hari Priya A, Kanchumarti TS, Bommilla LR, Regunathan R (2024) Language detection based on audio for Indian languages. In: Automatic speech recognition and translation for low resource languages, pp 275–296
51. Chu H-C, Zhang Y-L, Chiang H-C (2023) A CNN sound classification mechanism using data augmentation. Sensors 23(15):6972
52. Ali A, Renals S (2018) Word error rate estimation for speech recognition: E-WER. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers), pp 20–24
53. Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S, Narayanan S (2022) A review of speaker diarization: recent advances with deep learning. Comput Speech Lang 72:101317
54. Dielen I (2023) Improving the automatic speech recognition model whisper with voice activity detection. Master's thesis, Utrecht University
55. Lee H (2023) The rise of ChatGPT: exploring its potential in medical education. Anatomical sciences education
56. Pornprasit C, Tantithamthavorn C (2024) GPT-3.5 for code review automation: How do few-shot learning, prompt design, and model fine-tuning impact their performance? Preprint arXiv:2402.00905
57. Lim S, Schmälzle R (2022) Artificial intelligence for health message generation: theory, method, and an empirical study using prompt engineering. Preprint arXiv:2212.07507
58. Giray L (2023) Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 51(12):2629–2633
59. Lee U, Jung H, Jeon Y, Sohn Y, Hwang W, Moon J, Kim H (2023) Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education. Educ Inform Technol 2023:1–33
60. Cochran K, Cohn C, Rouet JF, Hastings P (2023) Improving automated evaluation of student text responses using GPT-3.5 for text data augmentation. In: International conference on artificial intelligence in education. Springer, pp 217–228
61. Uppalapati PJ, Dabbiru M, Rao KV (2023) A comprehensive survey on summarization techniques. SN Comput Sci 4(5):560. https://doi.org/10.1007/s42979-023-01560-0
62. Bockhorn LN, Vera AM, Dong D, Delgado DA, Varner KE, Harris JD (2021) Interrater and intrarater reliability of the beighton score: a systematic review. Orthop J Sports Med 9(1):2325967120968099
63. Bonthu S, Sree SR, Prasad M (2024) Sprag: building and benchmarking a short programming-related answer grading dataset. Int J Data Sci Anal 2024:1–13
64. Lutkevich B (2024) 19 of the best large language models in 2024. https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models/ [Online; accessed 17-Oct-2024]