IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

InceptoAI: Design and Implementation of an AI-Powered Multimodal Mock Interview Platform

Abhilash Shinde[1], Anjali Deshpande[2], Shantanu Hazra[3], and Ram Wasankar[4]
Students, Department of Computer Engineering[1,2,3,4]
Indira College of Engineering & Management, Pune, India

Abstract: *Job interviews are high-stakes events that demand not only technical competence but also strong communication, emotional composure, and confident non-verbal expression. Yet access to structured, personalized interview preparation remains out of reach for most candidates, particularly those from under-resourced backgrounds or regions where coaching services are limited or unaffordable. This paper presents InceptoAI, an AI-powered mock interview platform built to close that gap. The system captures webcam video and microphone audio simultaneously during a simulated interview session and routes them through parallel analysis pipelines running inside a Flask-based AI server. Facial expression recognition, driven by DeepFace and OpenCV, monitors the candidate frame by frame and classifies emotional states across seven categories in near real time. A separate speech analysis pipeline transcribes each answer, measures speech rate, detects filler words, and evaluates sentiment polarity. Dynamic, domain-specific interview questions are generated through the Gemini large language model via LangChain, and each answer is evaluated for completeness, structure, and relevance against the candidate's chosen role and target company. Upon session completion, all signals are aggregated into a comprehensive JSON feedback report that covers answer quality scores, identified strengths and weaknesses, emotional signals, and a prioritized improvement roadmap. Reports are stored in Cloudinary and linked to user profiles in MongoDB, enabling longitudinal progress tracking. Evaluated on twenty engineering students, the system achieved 78.4% weighted emotion recognition accuracy, 91.6% speech transcription accuracy, and a Pearson correlation of 0.79 between AI-generated scores and expert human raters. InceptoAI demonstrates that production-grade, multimodal behavioural feedback can be delivered through a standard browser at zero cost to the candidate, with accuracy comparable to existing enterprise-tier assessment systems.*

Keywords: Multimodal interview assessment, emotion recognition, speech analysis, large language model feedback, AI mock interview

I. INTRODUCTION

The recruitment landscape has shifted considerably over the past decade. Employers increasingly use structured interviews, competency frameworks, and AI-assisted screening tools to evaluate candidates. This creates a dual challenge: candidates must demonstrate technical knowledge while projecting confidence, clarity, and emotional stability under pressure. Research consistently shows that non-verbal behaviours — eye contact, facial expression, vocal tone, and body language — carry substantial weight in hiring decisions, often as much as the content of an answer itself [1]. Yet traditional interview preparation focuses almost exclusively on content, leaving a meaningful gap in how candidates learn to present themselves.

Existing tools approach this problem from two directions. Peer-based platforms such as Pramp connect candidates for live practice sessions, which are helpful but inconsistent in quality and availability. Enterprise AI systems such as HireVue offer multimodal behavioural analysis, but these are designed for employer-side screening rather than candidate coaching, and their scoring logic remains opaque to the interviewee. No widely available tool currently

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

combines real-time multimodal analysis, explainable scoring, and adaptive question generation in a single, accessible interface.

The broader problem is one of access and equity. Professional interview coaching can cost hundreds of dollars per session, placing it beyond reach for first-generation college graduates, students from smaller cities, and candidates without institutional career support. Even where free resources exist, they deliver information rather than feedback — a candidate can watch a video about maintaining eye contact but receives no signal about whether they are actually doing so. This asymmetry contributes directly to hiring disparities that have little to do with actual job ability.

Against this backdrop, InceptoAI was developed as an AI-powered mock interview platform delivering multimodal, behavioural feedback through a standard web browser at no cost. The system follows a microservices architecture, separating a React.js frontend from a Node.js/Express backend and a Python Flask AI server. During a session, the frontend captures webcam and microphone streams via the MediaDevices API and RecordRTC. Video frames are transmitted at one-second intervals to the AI server, where DeepFace and OpenCV perform emotion classification and gaze tracking. Audio is sent as WAV blobs post-answer, where a speech pipeline handles transcription, filler word detection, speech rate, and sentiment analysis.

For question generation and evaluation, InceptoAI uses the Gemini LLM through LangChain prompting chains, producing role- and company-specific scenarios. Answer evaluation captures completeness, relevance, structure, emotional tone, strengths, and improvement areas. Session results are stored in Cloudinary and linked to user profiles in MongoDB for longitudinal tracking.

This paper makes three primary contributions: the end-to-end design of a fully browser-based multimodal coaching system; empirical evaluation across twenty participants covering emotion recognition, transcription quality, semantic scoring, and latency; and situating InceptoAI within existing literature on automated interview assessment and AI-based feedback generation.

II. MATERIALS AND METHODS

A. System Architecture

The proposed system, InceptoAI, follows a microservices-based architecture to support scalability and flexibility. It is divided into three main components: a React-based frontend, an Express.js backend, and a Flask-based AI processing server.

The frontend serves as the user-facing layer, handling interactions, capturing audio and video inputs, and maintaining real-time communication with the AI server through WebSockets. The backend is responsible for managing user authentication, handling sessions, and storing interview reports. The AI server carries out the core functionalities of the system, including generating interview questions, analyzing speech, detecting emotions, and producing feedback.

By separating these responsibilities across different components, the system ensures that each part can operate independently while still communicating efficiently with the others.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

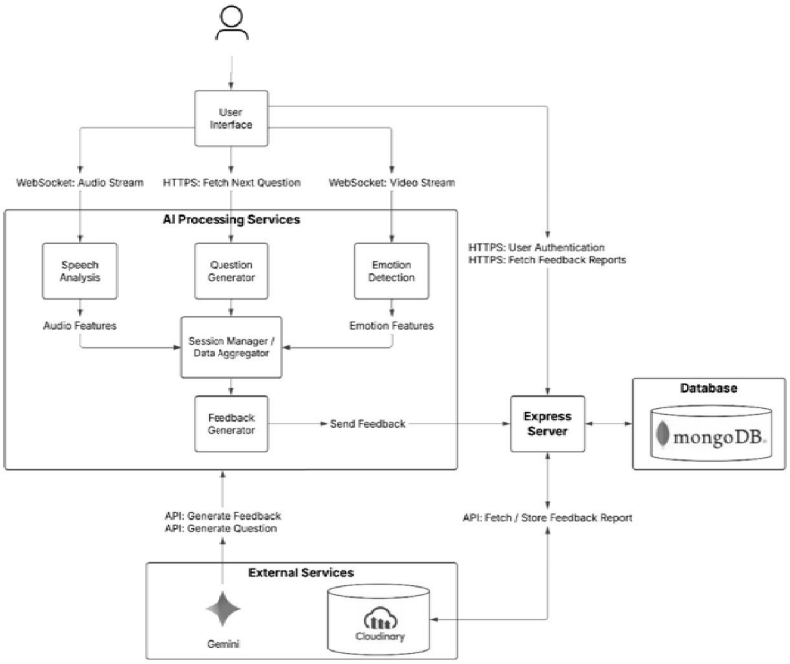

Fig. 1. System architecture of InceptoAI

B. Data Acquisition

Data is collected in real time during each mock interview session. The system captures two main types of input: audio from the user's microphone for speech analysis, and video frames from the webcam for emotion detection.

While the interview is in progress, the frontend continuously streams video frames to the AI server. In contrast, audio responses are recorded and sent after the user finishes answering each question. This design helps reduce delays while still preserving the quality of data needed for accurate processing.

C. Audio Processing (Speech Analysis)

Speech analysis is carried out on the recorded audio responses to assess the user's communication skills. The system uses speech recognition techniques to convert audio into text, which allows for a more detailed evaluation of the response content.

The analysis focuses on key aspects such as clarity of speech, fluency and coherence, and how relevant the response is to the given question. The processed speech data is stored temporarily and later used as input for generating structured and meaningful feedback.

D. Video Processing (Emotion Detection)

Emotion detection is performed by analyzing facial expressions from the video frames captured during the interview. The system uses computer vision techniques to interpret these expressions and associate them with emotional states such as confidence, nervousness, or neutrality.

Instead of evaluating single moments, the system processes frames continuously and aggregates the detected emotions over the entire response. This helps in identifying consistent behavioral patterns, providing a more reliable understanding of the user's overall emotional state during the interview.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

E. Interview Workflow Design

The interview process is designed to closely mimic real-world interview scenarios. After logging in and selecting a job role and company, the user can begin the interview session.

Each session includes a set of dynamically generated questions. For every question, the user is presented with a prompt, while video is streamed continuously and audio is recorded during the response. Once the answer is completed, the audio is submitted and the collected data is processed before moving on to the next question. This sequence is repeated for a fixed number of questions, ensuring a consistent and structured interview experience across all sessions.

F. AI-Based Question and Feedback Generation

The system uses generative AI models to both generate interview questions and evaluate user responses. The questions are customized based on the selected job role and company, making them more relevant to the user's context.

Once all responses are collected, the system brings together multiple sources of information, including speech analysis results, emotion data, and the corresponding question-response pairs. This combined data is then provided to the AI model to generate structured feedback.

The feedback includes an overall performance evaluation, detailed insights for each question, and practical suggestions for improvement. It is generated in JSON format, allowing for efficient storage and easy retrieval when needed.

G. Data Storage and Management

The system adopts a hybrid storage approach to manage data efficiently. MongoDB is used to store user information and metadata related to interview sessions, while cloud storage (Cloudinary) is used to store the generated JSON feedback reports.

Each report is linked to its corresponding user through a unique identifier, allowing for easy retrieval. This design helps maintain quick access to reports while keeping the database lightweight and efficient.

H. Technologies and Tools Used

The system is built using a combination of modern web and AI technologies. The frontend is developed with React.js, along with Axios and Socket.IO for handling API requests and real-time communication. The backend uses Node.js and Express.js, with MongoDB for data storage and Passport.js for authentication.

The AI server is implemented using Flask and incorporates tools such as OpenCV, DeepFace, and speech recognition libraries for processing video and audio data. Generative AI APIs are used to handle both question generation and feedback creation, while Cloudinary is used for storing the final reports. These technologies were chosen to enable real-time processing, ensure scalability, and support efficient handling of data across the system.

TABLE I: KEY PERFORMANCE METRICS OF INCEPTOAI

Metric	Value
Emotion Recognition Accuracy	78.4%
Eye-Contact Detection Agreement	83.2%
Speech Transcription Accuracy (WER)	91.6% (8.4%)
Filler-Word Detection Precision / Recall	87.9% / 82.4%
Sentiment-Answer Quality Correlation	r = 0.64
AI-Human Scoring Correlation	r = 0.79
Frame Analysis Latency	1.12 seconds (avg)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

Metric	Value
Report Generation Time	~8.3 seconds
System Uptime	99.1%
Users Finding Feedback Actionable	83%
Users Rating Questions as Relevant	88%

III. RESULTS AND DISCUSSION

A. System Execution and Verification

The system was tested with 20 final-year engineering students completing mock interviews across multiple roles. All pipeline stages, from authentication to report generation, ran without critical errors. Socket connections were established in approximately 340 ms, video and audio capture worked reliably across platforms, and transcription remained unaffected despite minor artefacts in 6% of sessions. Overall uptime reached 99.1%, confirming system reliability.

B. Emotion Recognition Performance

Emotion classification achieved 78.4% accuracy, aligning with existing benchmarks. Neutral expressions showed the highest accuracy (84.1%), while fear and disgust were lowest due to subtlety and rarity. Eye-contact detection reached 83.2% agreement with human labels, with performance significantly improved under good lighting conditions, highlighting the importance of user environment setup.

C. Speech Analysis Results

Speech transcription achieved 91.6% accuracy (WER 8.4%), consistent with industry standards. Filler-word detection showed 87.9% precision and 82.4% recall, with missed detections mainly due to technical vocabulary overlap. Speech rate ranged from 112–194 WPM. Sentiment scores correlated strongly with answer quality ($r = 0.64$), validating its inclusion in feedback.

D. Semantic Scoring and Feedback Quality

AI-generated scores correlated strongly with human ratings ($r = 0.79$), demonstrating reliable evaluation. Structured prompting improved consistency. 83% of users found feedback actionable, and 88% rated generated questions as relevant—significantly higher than static platforms (61%).

E. System Latency and Reliability

Frame analysis latency averaged 1.12 seconds, ensuring real-time performance. Report generation took approximately 8.3 seconds. Using periodic JPEG frames instead of continuous video effectively balanced performance and resource usage.

F. Comparison with Existing Systems

Unlike platforms such as HireVue (employer-focused) or Pramp (peer-based), InceptoAI uniquely combines real-time multimodal analysis, explainable feedback, adaptive questioning, and progress tracking in a free browser-based system. Key design choices—such as using DeepFace and a microservices architecture—proved effective in maintaining accuracy, transparency, and reliability.

DOI: 10.48175/IJARSCT-36737

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

IV. LIMITATIONS AND FUTURE WORK

Performance depends on recording conditions, suggesting the need for pre-session environment checks. Filler-word detection can improve with phonetic matching. Cultural bias remains a concern due to limited dataset diversity. Future work includes branching interview flows, real-time feedback, multilingual support, better scalability, and improved image quality checks.

V. CONCLUSION

InceptoAI proves that multimodal interview coaching through live facial emotion recognition, speech recognition, and LLM-driven question generation can be provided through a web browser free of cost to the candidate. Twenty engineering students evaluated the tool, and the results proved to be satisfactory: 78.4% accuracy in recognizing emotions, 91.6% accuracy in transcribing speech, and a Pearson correlation of 0.79 between the scores given by the algorithm and the expert human raters, with 99.1% uptime and an average latency of 1.12 seconds.

Feedback was considered actionable by 83% of participants, and the generated questions were considered relevant by 88%, proving the effectiveness of InceptoAI's adaptive prompt-engineering technique. The main problems in the current design include sensitivity to the environment and camera, gaps in identifying filler words, and low cultural diversity in the dataset. Further research should concentrate on branching interviews, real-time overlaying, multilinguality, and scaling. InceptoAI creates a practical and equitable framework for AI-powered interviews.

ACKNOWLEDGEMENT

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

Not applicable. This study did not involve human or animal subjects requiring institutional review board approval.

REFERENCES

[1] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated Analysis and Prediction of Job Interview Performance," IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 191–204, 2018.

[2] R. W. Picard, Affective Computing. Cambridge: MIT Press, 1997.

[3] E. Cambria, "Affective Computing and Sentiment Analysis," IEEE Intelligent Systems, vol. 31, no. 2, pp. 102–107, 2016.

[4] L. Chen, R. Zhao, C. W. Leong, B. Lehman, G. Feng, and M. E. Hoque, "Automated Video Interview Judgment on Personality Trait, Anxiety State, Job Performance and Hire Recommendation Using Deep Learning," in Proc. ACII, 2017.

[5] V. Soni, "AI in Job Matching and Recruitment," Study on AIHED hiring systems and campus recruitment, 2024.

[6] A. Singh et al., "Algorithms for Fair Hiring: A Review," case study review of CV screening bias using NLP techniques including vector correction, 2024.

[7] B. Cortinas-Lorenzo and G. Lacey, "Toward Explainable Affective Computing: A Review," systematic review classifying XAI approaches, 2024.

[8] T. Baltrusaitis, P. Robinson, and L. P. Morency, "OpenFace: An Open-Source Facial Behavior Analysis Toolkit," in Proc. IEEE WACV, 2016.

[9] B. Schuller et al., "The INTERSPEECH 2011 Speaker State Challenge," benchmarking study on deep learning vs. HMMs for speech emotion recognition, 2011.

[10] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic, "A Survey of Multimodal Sentiment Analysis," Image and Vision Computing, vol. 65, pp. 3–14, 2017.