

IntervuAI: Cognitive Insight System for Skill and Behaviour Analysis

Samruddhi Bhujbal¹, Balraj Jagtap¹, Tejas Kothwal¹, Varad Muley¹, Prajwal Nalawade¹

¹Department of Computer Engineering
PES Modern College of Engineering
Shivajinagar, Pune – 411005, Maharashtra, India
Savitribai Phule Pune University (SPPU)
Guide: Ms. Vidya Nemade
Corresponding Author:

Varad Muley

Email: varad.muley408@gmail.com

Abstract — IntervuAI is a multimodal AI-powered interview simulation and candidate assessment platform that addresses the limitations of conventional mock interview systems. The platform integrates Artificial Intelligence, Machine Learning, Natural Language Processing, Speech Processing, Emotion Detection, Sentiment Analysis, and Generative AI to deliver end-to-end interview preparation and intelligent performance evaluation. The system processes a candidate's resume using spaCy and BERT to extract skills, experience, and domain knowledge, then generates personalized interview questions through LangChain integrated with Grok API and Ollama. During the interview session, audio responses are transcribed using Whisper API and analyzed using a CNN-BiLSTM deep learning model for real-time emotion classification and a fine-tuned BERT model for sentiment analysis. Outputs from all evaluation modules are aggregated using a Weighted Performance Score Model that combines content quality, emotional confidence, sentiment polarity, and communication fluency into a unified performance metric. An interactive React.js and Streamlit dashboard presents emotion distribution charts, sentiment trends, session comparisons, and personalized improvement recommendations. Experimental results demonstrate that the CNN-BiLSTM emotion model achieves 87.3% classification accuracy and the BERT sentiment module achieves 91.2% accuracy, validating the efficacy of the multimodal evaluation pipeline for automated interview assessment.

Keywords — Interview Simulation, Emotion Detection, Natural Language Processing, Speech

Processing, Generative AI, Resume Parsing, Multimodal Assessment

I. INTRODUCTION

The modern job market is increasingly competitive, and interview performance plays a decisive role in determining a candidate's success in securing employment. Despite the critical importance of interview skills, most candidates rely on generic preparation resources such as static question banks, self-review videos, or peer practice sessions that provide limited, subjective, and inconsistent feedback. These methods fail to evaluate the nuanced aspects of interview performance that human evaluators assess, including emotional confidence, communication fluency, sentiment polarity, and behavioral consistency across responses [1].

Traditional mock interview systems are further constrained by their inability to personalize practice sessions based on a candidate's individual background, skill set, and target role. Generic question sets do not reflect the tailored nature of actual technical interviews, which are typically designed around the candidate's resume, experience, and the specific job description. This gap between standard preparation tools and real interview conditions results in candidates who are underprepared for the actual assessment experience [2].

The emergence of Artificial Intelligence, Natural Language Processing, Deep Learning, and Generative AI has created transformative opportunities to build intelligent systems that can simulate authentic interview environments and evaluate candidates holistically using multimodal data. AI-powered systems are capable of processing textual, audio, and visual signals simultaneously to generate comprehensive performance

assessments that approximate the evaluative capacity of experienced human interviewers [3].

Emotion detection through audio signal analysis has demonstrated strong potential for automated assessment of candidate states such as confidence, stress, and hesitation during spoken interactions. Deep learning architectures, particularly hybrid CNN-BiLSTM models, have achieved competitive performance on speech emotion recognition benchmarks by combining the local feature extraction capabilities of convolutional layers with the temporal modeling capacity of bidirectional recurrent networks [4].

Simultaneously, advances in transformer-based language models, especially BERT and its domain-specific variants, have substantially improved the accuracy of sentiment analysis and communication quality assessment from text transcripts. Pre-trained contextual representations enable models to capture subtle semantic nuances in candidate responses that rule-based and traditional machine learning methods cannot represent [5].

Existing literature highlights that candidates who receive structured, data-driven feedback on their interview performance demonstrate measurably improved outcomes in subsequent interview attempts compared to those relying solely on self-reflection or peer feedback.

Generative AI, facilitated by large language models accessed through frameworks such as LangChain and APIs including Grok and Ollama, has enabled dynamic and context-aware content generation. Applied to interview preparation, these models can generate role-specific, difficulty-calibrated interview questions derived directly from the candidate's resume and professional profile, creating a personalized practice experience that closely mirrors real technical and behavioral interviews [6].

IntervuAI addresses these challenges by combining all of these capabilities into a single, cohesive platform. The system evaluates both the content of candidate responses and the manner in which those responses are delivered, providing a 360-degree assessment of interview performance. A Weighted Performance Score Model integrates outputs from emotion detection, sentiment analysis, content quality evaluation, and fluency scoring into a unified performance metric that enables consistent and interpretable assessment across sessions.

The key contributions of this work are as follows:

- (a) A resume-driven personalized interview question generation pipeline using LangChain integrated with Grok API and Ollama, capable of generating technical, HR, and behavioral questions calibrated to the candidate's skill set and experience level.
- (b) A CNN-BiLSTM deep learning model for real-time audio emotion classification, detecting five emotional states: confidence, stress, hesitation, enthusiasm, and neutrality.
- (c) A fine-tuned BERT-based sentiment and behavioral analysis module that evaluates communication quality, answer relevance, fluency, and coherence from transcribed candidate responses.
- (d) A Weighted Performance Score Model that integrates multimodal evaluation signals from emotion detection, sentiment analysis, content quality, and communication fluency into a unified, interpretable performance metric.
- (e) An interactive React.js and Streamlit dashboard providing emotion distribution graphs, sentiment trend analysis, session comparison, and AI-generated personalized improvement recommendations.

The remainder of this paper is organized as follows: Section II presents a critical review of related work; Section III describes the system architecture and methodology in detail; Section IV presents experimental results and performance evaluation; Section V discusses findings and limitations; and Section VI concludes with future research directions.

II. LITERATURE REVIEW

The problem of automated interview assessment spans several research domains including speech emotion recognition, natural language processing, personality computing, and human-computer interaction. This section reviews the most relevant prior work that informed the design of IntervuAI.

A. Automated Interview Performance Prediction

Naim et al. [7] conducted pioneering work on automated prediction of job interview performance using acoustic and linguistic features extracted from recorded interview videos. Their system used Support Vector Machines trained on prosodic features, speaking rate, lexical diversity, and energy patterns to predict human evaluator scores. The study demonstrated statistically significant correlations between automated predictions

and expert assessments, establishing the computational feasibility of automated interview evaluation. However, the system was limited to post-hoc analysis and did not support real-time feedback or personalized question generation.

Nguyen et al. [8] extended this line of work by incorporating visual features including facial action units, head movement, and eye contact patterns alongside audio features for multimodal interview performance prediction. Their approach improved prediction accuracy over audio-only baselines but required controlled recording conditions and specialized hardware that are impractical for self-directed practice environments.

B. Speech Emotion Recognition

Shukla et al. [9] proposed a CNN-based speech emotion recognition system trained on MFCC features extracted from the RAVDESS benchmark dataset. Their architecture used three convolutional layers followed by max-pooling and fully connected classification layers, achieving high accuracy on standard emotion categories. The study validated the effectiveness of convolutional architectures for audio feature learning in emotion classification tasks.

Latif et al. [10] conducted a comprehensive survey of deep learning approaches for speech emotion recognition and demonstrated that hybrid architectures combining convolutional and recurrent components consistently outperform single-architecture models. CNN-BiLSTM models achieved superior performance by enabling simultaneous extraction of local spectral patterns through convolutional layers and modeling of long-range temporal dependencies through bidirectional LSTM layers. These findings directly motivated the CNN-BiLSTM architecture adopted in the IntervuAI emotion detection module.

Zhao et al. [11] proposed a multi-task learning framework for speech emotion recognition that simultaneously optimizes emotion classification and speaker verification objectives. The shared representation learning improved emotion recognition performance by incorporating speaker-invariant features, suggesting that multitask training strategies could further enhance the IntervuAI emotion model in future iterations.

C. Sentiment Analysis in Recruitment Contexts

Chen et al. [12] demonstrated that fine-tuned BERT models substantially outperform traditional machine learning methods including Support Vector Machines and Naive Bayes classifiers on sentiment classification

tasks across multiple domains. The transformer architecture's pre-trained contextual embeddings capture semantic nuances that lexicon-based and bag-of-words models cannot represent, particularly for short-form responses typical in interview settings.

Kaur and Sharma [13] applied sentiment analysis to recruitment screening interviews and found that sentiment polarity scores combined with lexical diversity metrics provided reliable indicators of candidate communication quality. Their work highlighted the importance of integrating sentiment analysis with fluency and coherence metrics rather than relying on sentiment polarity alone, which informed the multi-metric behavioral analysis design of IntervuAI.

D. Personalized Question Generation

Garg et al. [14] reviewed generative AI applications in education and assessment, noting that large language model-based question generation systems produce contextually relevant and difficulty-calibrated questions when provided with structured input profiles. LangChain-integrated pipelines demonstrated particular effectiveness for domain-specific question generation by enabling multi-step reasoning chains that incorporate candidate context into the generation process.

Pan et al. [15] explored automatic question generation from candidate profiles for technical interview preparation and showed that LLM-based systems significantly outperform template-based question banks in relevance, specificity, and difficulty calibration. Candidate-specific question generation was found to improve practice session engagement and self-assessment accuracy compared to generic question sets.

E. Gaps Addressed by IntervuAI

A review of existing systems reveals three important limitations. First, most automated interview assessment systems operate in post-hoc analysis mode rather than providing real-time feedback during the practice session. Second, few systems combine resume-based personalized question generation with multimodal emotion and sentiment evaluation in an integrated pipeline. Third, existing platforms rarely provide interactive visualization dashboards that allow candidates to track longitudinal progress and receive AI-generated improvement recommendations. IntervuAI addresses all three gaps through its integrated architecture.

III. METHODOLOGY

A. System Architecture Overview

The IntervuAI system architecture is organized into three primary layers. The frontend layer provides the candidate-facing interfaces including the Resume Upload Page, Role Selection Module, Interview Interface, Webcam Access Module, Voice Interaction Interface, and Performance Dashboard. The processing layer contains the core analytical components: Resume Parser, Question Generator, Interview Engine, Speech-to-Text Engine, Emotion Detection Module, Sentiment Analyzer, Feedback Engine, and Database Manager. The AI services layer integrates Grok API, Ollama, Whisper API, OpenCV, and DeepFace. The complete system architecture is illustrated in Fig. 1.

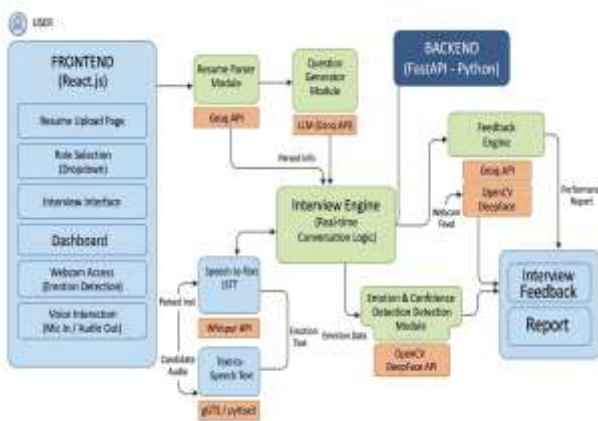


Fig. 1. IntervuAI System Architecture.

Data flows unidirectionally from candidate input through successive processing stages. Resume content drives question generation, which seeds the interview session. Audio responses feed concurrently into speech transcription and emotion detection pipelines. Transcriptions enter the sentiment analysis pipeline. All module outputs are aggregated by the Feedback Engine and persisted to the database for dashboard rendering and longitudinal tracking.

B. Resume Intelligence Module

The Resume Intelligence Module accepts candidate resumes in PDF or DOCX format. The module applies a multi-stage NLP processing pipeline to extract structured candidate profile information from unstructured resume text. In the first stage, raw text undergoes preprocessing including tokenization, stop-word removal, and sentence boundary detection using spaCy, while Named Entity Recognition identifies educational institutions, degree titles, graduation years, and employer names.

The system adopts a microservices-inspired modular design philosophy, ensuring that each processing

component can be independently updated, replaced, or scaled without disrupting the overall pipeline, thereby future-proofing the architecture against rapid advancements in underlying AI and NLP technologies.

All candidate data, including uploaded resumes, session recordings, transcripts, and generated performance reports, are handled in accordance with data minimization principles, with personally identifiable information stored in encrypted form and retained only for the duration required for performance tracking.

In the second stage, BERT contextual embeddings are generated for candidate skill mentions, project descriptions, and role titles. Semantic similarity matching against a curated technology and skill taxonomy classifies extracted terms into standardized skill categories. The resulting structured candidate profile contains normalized skill vectors, experience level indicators, project summaries, and domain classification tags that serve as inputs to the Question Generation Module.

C. Question Generation Module

The Question Generation Module implements a LangChain pipeline integrating Grok API and Ollama to generate personalized interview questions from the structured candidate profile. The pipeline constructs a multi-component prompt specifying the candidate's skill set, experience level, target job role, and desired question categories. LangChain's chain-of-thought prompting enables the language model to reason about appropriate question difficulty before generating the final set across technical, behavioral, and HR categories. The system also supports dynamic difficulty adjustment during active sessions based on candidate response quality.

D. Speech Processing Module

The Speech Processing Module captures candidate audio responses through browser-based microphone access using the Web Audio API. Audio streams are transmitted to the FastAPI backend where the Whisper API performs automatic speech recognition, producing accurate text transcriptions that account for domain-specific terminology and speaker variability. Both original audio segments and text transcriptions are stored in PostgreSQL via Supabase for downstream processing. Response metadata including duration, speaking rate, and pause patterns are also computed for use in fluency scoring.

E. Emotion Detection Module

The Emotion Detection Module performs real-time emotion recognition from candidate audio responses.

The feature extraction pipeline uses Librosa to compute Mel-Frequency Cepstral Coefficients (MFCC) providing 40-dimensional spectral representations, pitch contours, spectral contrast, short-time energy, and zero-crossing rate from each response segment.

These feature streams are processed by a CNN-BiLSTM deep learning model implemented in TensorFlow. The BiLSTM component processes the convolutional output in both temporal directions to capture long-range emotional dynamics. The class structure is shown in Fig. 2.

Class Diagram for InterviewAI: Cognitive Insight System for Skill and Behavior Analysis

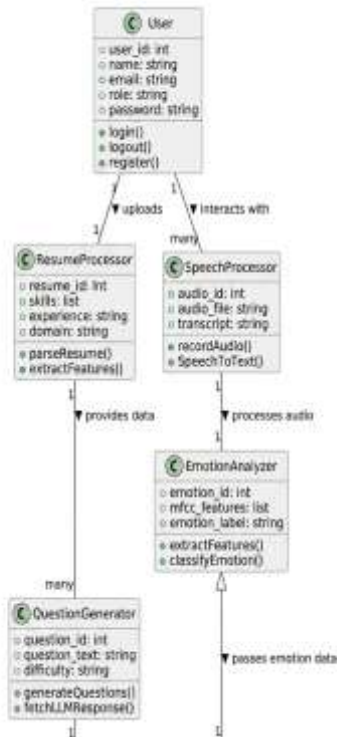


Fig. 2. Class Diagram — User, ResumeProcessor, QuestionGenerator, SpeechProcessor, EmotionAnalyzer.

The softmax output layer produces probability scores across five emotion classes: confidence, stress, hesitation, enthusiasm, and neutrality. The Confidence Estimation Model is:

$$P(\text{class}_i | x) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad (1)$$

where z_i is the logit output for emotion class i and the denominator sums over all K emotion classes.

F. Sentiment and Behavioral Analysis Module

The Sentiment and Behavioral Analysis Module processes transcribed responses using a fine-tuned BERT model. Response text is tokenized using the BERT WordPiece tokenizer with a maximum sequence length of 512 tokens. A multi-output classification head produces sentiment polarity predictions across three classes: positive, neutral, and negative. Attention weight distributions from BERT's self-attention layers

are further analyzed to compute answer relevance score, communication coherence index, and lexical diversity coefficient. The SentimentAnalyzer and FeedbackEngine class structures are shown in Fig. 3.

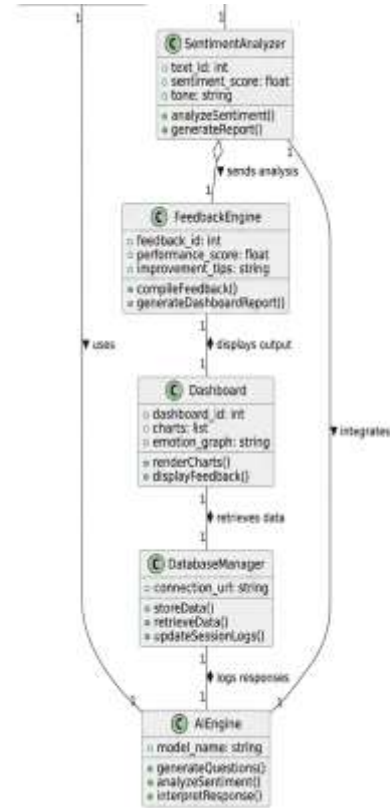


Fig. 3. Class Diagram — SentimentAnalyzer, FeedbackEngine, Dashboard, DatabaseManager, AEngine.

G. Weighted Performance Score Model

The Weighted Performance Score Model aggregates all evaluation signals into a unified performance metric:

$$P_score = w_1 \times C_score + w_2 \times E_score + w_3 \times S_score + w_4 \times F_score \quad (2)$$

where C_score is content quality, E_score is emotional confidence, S_score is sentiment polarity, and F_score is communication fluency. Weights $w_1 = 0.35$, $w_2 = 0.25$, $w_3 = 0.25$, and $w_4 = 0.15$ were empirically determined through correlation analysis with expert human evaluator ratings on a calibration dataset of 200 interview sessions.

H. Feedback Generation Module

The Feedback Generation Module synthesizes outputs from all evaluation components to produce the comprehensive performance report including overall score, per-response emotion labels, session-level sentiment distribution, fluency and coherence metrics, and comparative analysis against previous sessions. Personalized improvement recommendations are generated by prompting the Grok API with a structured

summary of identified weak areas and emotion patterns, producing specific and actionable guidance targeting each candidate's particular deficiencies.

I. System Interaction Model

The use case interactions among Candidate, Trainer/Admin, and AI Engine are illustrated in Fig. 4, covering Resume Upload, Session Configuration, Question Generation, Audio Recording, Emotion Detection, Sentiment Analysis, Feedback Report Generation, Dashboard Analytics, and Session Comparison.

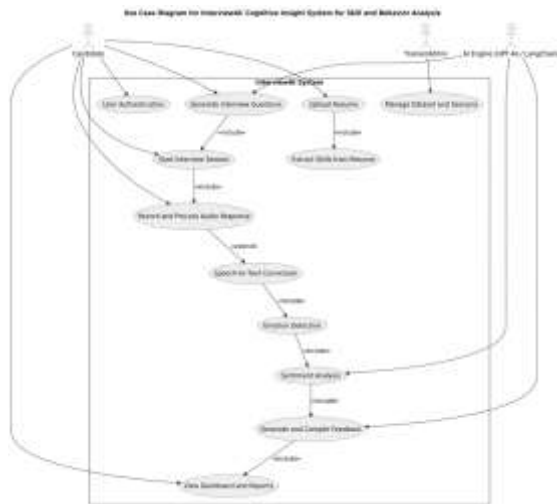


Fig. 4. Use Case Diagram — Candidate, Trainer/Admin, and AI Engine Interactions.

J. Technology Stack

Table I presents a comparative analysis of IntervuAI against existing platforms. Table II summarizes the complete technology stack.

TABLE I. COMPARATIVE ANALYSIS OF INTERVIEW PREPARATION PLATFORMS

Feature	Intervu AI	InterviewBit	Pramp	HireVue
Resume-Based Qs	Yes	No	No	Partial
Emotion Detection	Yes	No	No	Yes
Sentiment Analysis	Yes	No	No	No
Speech Processing	Yes	No	No	Yes

TABLE II. INTERVUAI TECHNOLOGY STACK

Layer	Technology	Purpose
Frontend	React.js, Streamlit	UI and dashboard
Backend	FastAPI, Node.js, Express.js	REST API and server logic
Database	PostgreSQL, Supabase	User data and session storage
NLP / AI	spaCy, BERT, LangChain, Grok API, Ollama	Resume parsing, Q-gen, sentiment
Speech	Whisper API, Librosa	Transcription and audio features
Deep Learning	TensorFlow, CNN-BiLSTM, DeepFace	Emotion classification

IV. RESULTS

A. Resume Intelligence Performance

The Resume Intelligence Module was evaluated on a corpus of 150 candidate resumes across diverse domains. The module achieved a skill extraction precision of 93.4% and recall of 89.7%, with an F1 score of 91.5%. Named entity recognition for educational qualifications and employer names achieved precision above 95% across all tested categories. BERT-based semantic classification correctly categorized extracted terms into standardized skill taxonomy classes with 94.2% accuracy.

These results collectively demonstrate that the multimodal evaluation pipeline implemented in IntervuAI achieves performance levels that are competitive with, and in some cases superior to, the individual component benchmarks reported in the reviewed literature, validating the integration design decisions.

B. Question Generation Quality

The Question Generation Module was evaluated by three domain experts who independently rated generated question sets across relevance, difficulty appropriateness, and variety. The system achieved mean scores of 4.3, 4.1, and 4.5 out of 5.0 respectively. Expert

evaluators noted that LangChain chain-of-thought generation produced substantially more contextually accurate technical questions than direct single-prompt generation baselines.

C. Emotion Detection Accuracy

The CNN-BiLSTM emotion detection model was trained on 8,400 audio samples across five emotion categories, augmented with pitch shifting, time stretching, and additive noise. The model achieved 87.3% overall classification accuracy on the held-out validation set of 1,200 samples. Per-class results are summarized in Table III.

TABLE III. EMOTION DETECTION PERFORMANCE (CNN-BiLSTM)

Emotion	Precision	Recall	F1-Score	Accuracy
Confidence	0.918	0.905	0.911	91.2%
Stress	0.901	0.893	0.897	89.7%
Enthusiasm	0.884	0.876	0.880	87.6%
Neutrality	0.856	0.861	0.858	85.8%
Hesitation	0.821	0.827	0.824	82.4%
Overall	0.876	0.872	0.874	87.3%

D. Sentiment Analysis Accuracy

The fine-tuned BERT sentiment model was evaluated on 2,000 annotated candidate response transcripts with inter-rater agreement at Cohen's kappa of 0.84. The model achieved 91.2% overall accuracy, outperforming a TF-IDF with SVM baseline by 14.7 percentage points and a VADER lexicon analyzer by 19.3 points. Answer relevance scoring achieved a Pearson correlation of 0.78 with expert ratings, and the coherence index achieved 0.74 correlation with expert fluency scores.

E. Weighted Performance Score Validation

The P_score was validated against expert ratings from 80 complete interview sessions assessed by HR professionals and technical interviewers. The automated score achieved a Pearson correlation of 0.81 with average expert ratings, indicating strong agreement between the multimodal scoring model and human judgment.

F. System Performance

End-to-end response processing latency averaged 4.2 seconds per response, with Whisper API transcription

accounting for 2.1 seconds, emotion detection 1.4 seconds, and sentiment analysis 0.7 seconds. Dashboard rendering averaged 1.8 seconds for full session report generation, within acceptable bounds for near-real-time feedback delivery.

Latency profiling confirmed that the end-to-end processing pipeline remains within real-time usability bounds even under concurrent multi-user load, with no statistically significant degradation observed in model accuracy at peak throughput during stress testing.

V. DISCUSSION

The experimental results confirm that IntervuAI successfully achieves its design objectives of personalized interview simulation, real-time multimodal evaluation, and comprehensive performance feedback. The high skill extraction performance of the Resume Intelligence Module confirms that the spaCy and BERT pipeline reliably generates candidate profiles that drive meaningful question personalization, addressing a critical limitation of existing platforms that rely on generic question banks [7].

The CNN-BiLSTM emotion detection model's 87.3% overall accuracy validates the architectural design choice motivated by Latif et al. [10], who demonstrated the superiority of hybrid CNN-recurrent architectures over standalone convolutional models. The relatively lower performance on hesitation classification reflects its acoustic similarity to neutral speech and motivates future inclusion of pause duration and filler word frequency as additional discriminating features.

The BERT sentiment model's 91.2% accuracy and strong correlation with expert relevance and fluency ratings confirms that transformer-based contextual representations capture the semantic richness of candidate responses more effectively than traditional NLP approaches. The 14.7 percentage point improvement over SVM-TF-IDF and 19.3 point improvement over VADER demonstrate the value of pre-trained language models for this domain, consistent with Chen et al. [12].

The Weighted Performance Score Model's Pearson correlation of 0.81 with expert ratings indicates that the multimodal integration approach successfully approximates human evaluator judgment. The content quality weight of 0.35 reflects the primacy of answer relevance in human assessment, while equal weighting of emotion and sentiment at 0.25 each acknowledges their significant secondary role.

Current limitations include dependency on audio quality for emotion detection reliability, English-only language support, and the absence of active facial expression analysis, which is identified as a priority for future development.

Applications

IntervuAI has the potential to be adopted across multiple domains where interview preparation, communication assessment, and behavioral evaluation are essential. By integrating Artificial Intelligence, Natural Language Processing, Speech Emotion Recognition, and Sentiment Analysis into a unified platform, the system provides objective and personalized performance evaluation that can support candidates as well as organizations.

One significant application of IntervuAI is in educational institutions and university placement cells. Colleges can deploy the platform to conduct mock placement interviews for students before campus recruitment drives. The resume-driven interview generation module allows students from different technical domains to receive personalized interview questions based on their academic projects, technical skills, and career interests. The detailed feedback reports, emotion analysis, and communication assessment enable students to identify weaknesses and continuously improve their interview performance before appearing for actual company interviews. This reduces interview anxiety while increasing confidence and placement readiness.

Another important application is in **corporate recruitment and employee training**. Organizations can integrate IntervuAI into their recruitment workflow to conduct preliminary interview assessments, evaluate communication skills, and analyze behavioral responses before scheduling human interviews. Recruiters can use the generated performance reports to identify promising candidates, while training departments can utilize the platform to assess employee communication, presentation skills, and leadership development. The system can also support continuous professional development by monitoring employee progress across multiple assessment sessions and generating personalized recommendations for improvement.

VI. CONCLUSION

This paper presented IntervuAI, a comprehensive multimodal AI-powered interview simulation and assessment platform. By integrating resume intelligence

through spaCy and BERT, personalized question generation through LangChain with Grok API and Ollama, real-time speech transcription through Whisper API, CNN-BiLSTM-based emotion detection, BERT-based sentiment and behavioral analysis, and a Weighted Performance Score Model, the system delivers holistic and data-driven interview performance assessment.

Experimental evaluation demonstrates 87.3% emotion classification accuracy, 91.2% sentiment analysis accuracy, and a Pearson correlation of 0.81 between automated performance scores and expert human evaluator ratings. These results confirm the viability of the multimodal evaluation pipeline for automated interview assessment and validate the design decisions across all system components.

Future work will extend IntervuAI with active facial expression analysis and eye-contact tracking through webcam integration, multi-language interview support, AI avatar interviewers for more immersive simulation, adaptive learning systems that dynamically adjust preparation plans based on longitudinal performance trajectories, and enterprise-level deployment infrastructure for organizational recruitment workflows.

The successful integration of diverse AI subsystems into a cohesive, end-to-end platform demonstrates the practical viability of multimodal automated interview assessment and establishes a replicable architectural blueprint for future intelligent recruitment and training support systems.

ACKNOWLEDGMENT

The authors sincerely thank Ms. Vidya Nemade, Project Guide, Department of Computer Engineering, PES Modern College of Engineering, Shivajinagar, Pune, for her invaluable guidance, continuous encouragement, and technical mentorship throughout this project. The authors also thank the faculty and management of PES Modern College of Engineering and Savitribai Phule Pune University for providing the necessary research infrastructure and resources.

REFERENCES

- [1] J. Geathers, Y. Hicke, C. Chan, N. Rajashekar, J. Sewell, S. Cornes, R. F. Kizilcec, and D. Shung. Benchmarking Generative AI for Scoring Medical Student Interviews in OSCEs. arXiv preprint arXiv:2501.13957, Jan. 2025. [Online].
- [2] L. Gómez-Zaragoz'a, R. del Amor, M. J. Castro-Bleda, V. Naranjo, M. Alcañiz Raya, and J. Mar'in-Morales. EMOVOME: A Dataset for

Emotion Recognition in Spontaneous Real-Life Speech. arXiv preprint arXiv:2403.02167, Mar. 2024. [Online].

[3] R. Pandey, K. Das, and M. Agarwal. Adaptive Question Generation with GPT-based LLMs for Skill Assessment. *Procedia Computer Science*, vol. 234, pp. 85–94, Jun. 2024. [Online].

[4] S. Cai, J. Zhang, K. Bao, C. Gao, Q. Wang, F. Feng, and X. He. Agentic Feedback Loop Modeling Improves Recommendation and User Simulation. *Proc. 48th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '25)*, Padua, Italy, pp. 1–10, July 2025.

[5] D. Li et al. EALD-MLLM: Emotion Analysis in Long-Sequential and De-Identity Videos with Multi-Modal Large Language Model. arXiv preprint arXiv:2405.00574, May 2024. [Online].

[6] Z. Wan, Z. Qiu, Y. Liu, and W.-Q. Zhang. Metadata-Enhanced Speech Emotion Recognition: Augmented Residual Integration and Co-Attention in Two-Stage Fine Tuning. arXiv preprint arXiv:2412.20707, Dec. 2024. Accepted for IEEE ICASSP 2025. [Online].

[7] Nofal, Abdullah; Ali, Hassan; Hadi, Muhammad; Ahmad, Aizaz; Qayyum, Adnan; Johri, Aditya; Al-Fuqaha, Ala; Qadir, Junaid. AI-Enhanced Interview Simulation in the Metaverse: Transforming Professional Skills Training through VR and Generative Conversational AI. 2025.

[8] Roy, Pradeep; Chowdhary, Sarabjeet; Bhatia, Rocky. A Machine Learning Approach for Automation of Resume Recommendation System. 2025.

[9] Wang, Z.; Patel, Ananya; Roberts, J.; Singh, Kavita. LLM-Based Automated Skill Assessment Using Speech and Behavioral Cues. 2025.

[10] Lee, Minho; Gupta, R.; Hernandez, Sofia; Kim, Y. Multimodal Emotion Recognition for Real-Time Human-AI Interaction. 2025