Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

AI Virtual Interview Trainer: A Multimodal Intelligent System for Automated Interview Simulation and Skill Development

Prof. Priyanka Kinage
Computer Science & Engineering (AI) Vishwakarma Institute of Technology Pune, India

Prof. Anita Dombale
Computer Science & Engineering (AI) Vishwakarma Institute of Technology Pune, India

Dhruv Oswal
Computer Science & Engineering (AI) Vishwakarma Institute of Technology Pune, India

Avdhut Deshmukh
Computer Science & Engineering (AI) Vishwakarma Institute of Technology Pune, India

Aditya Gaikwad
Computer Science & Engineering (AI) Vishwakarma Institute of Technology Pune, India

Suyog Dudhade
Computer Science & Engineering (AI) Vishwakarma Institute of Technology Pune, India

***Abstract*—*Landing a job today is about much more than what you know; how you communicate, carry yourself, and respond under pressure matters just as much. Yet most students never get enough structured practice before facing real recruiters. Traditional mock sessions are expensive, hard to schedule, and shaped by whoever happens to evaluate you. Automated alternatives exist, but the vast majority look at words alone, ignoring the non-verbal signals—eye contact, posture, facial expression—that quietly shape an interviewer's impression. This paper introduces the **AI Virtual Interview Trainer (AIVIT)**, a system that watches and listens at the same time. On the audio side, Natural Language Processing (NLP) checks grammar, measures how relevant and fluent each answer is, and reads the candidate's emotional tone. On the video side, Computer Vision tracks gaze direction, body alignment, and facial muscle activity in real time. The two streams are fused into a single score, and a built-in adaptive engine responds by making the next question harder or easier depending on how the candidate is doing—the same principle behind deliberate practice. Across forty undergraduate volunteers tested over five sessions, AIVIT produced gains of up to 18.9% in confidence and 16.3% in eye-contact consistency. Compared against seven recently published systems, it is the only one to combine all five critical capabilities: multimodal evaluation, adaptive difficulty, session tracking, fairness awareness, and CPU-only deployment.*

***Index Terms*—*interview coaching, multimodal AI, adaptive learning, NLP, computer vision, behavioral analysis, session tracking.*

I. Introduction

Picture two candidates walking into the same interview. One has rehearsed answers to every question but speaks in a monotone, avoids eye contact, and slumps slightly in the chair. The other is no more knowledgeable but speaks at a steady pace, holds the interviewer's gaze, and sits with quiet confidence. Research consistently shows that the second candidate is rated higher—not because their answers are better, but because how they delivered those answers created an impression of competence and composure [1].

This gap between content and delivery is exactly where most automated interview tools fall short. A 2024 talent-acquisition survey found that 62% of hiring professionals now regard AI-assisted preparation as genuinely transformative for candidate readiness [16], yet the tools they endorse typically score transcripts and ignore everything the camera could reveal. Meanwhile, traditional human-led mock interviews remain limited: scheduling is cumbersome, feedback is subjective, and a single university placement cell cannot meaningfully coach thousands of students in the weeks before campus recruitment.

The technical ingredients for something better have quietly come together. Whisper [5] transcribes speech reliably even with accents and background noise. Sentence-BERT [7] captures what an answer means rather than just what words it contains. MediaPipe [8] extracts 468 facial landmarks and 33 body keypoints in real time on a standard laptop CPU. The missing piece has been a framework that wires these components together, calibrates their relative importance, adapts to each user's current ability, and tracks growth over time. AIVIT is that framework.

The contributions of this work are four-fold. First, a synchronized multimodal pipeline that fuses NLP verbal scores with CV behavioral scores through a fusion weight calibrated against expert human raters (Pearson $r = 0.87$). Second, a threshold-based adaptive question engine that keeps practice difficulty at the frontier of each user's competence, drawing on deliberate practice theory [9]. Third, longitudinal session tracking across five structured sessions that reveals how verbal and behavioral skills improve at different rates. Fourth, an empirical study on 40 participants providing pre-post evidence of improvement across six dimensions, supported by latency profiling and usability measurement.

The remainder of this paper is organized as follows. Section II reviews prior work and maps the gaps AIVIT addresses. Section III covers methodology. Sections IV through VI detail the architecture, evaluation modules, and score fusion. Section VII presents experimental findings with four annotated figures. Sections VIII and IX discuss ethical considerations and conclusions.

II. Literature Review

Work in this space has grown rapidly, but it has done so in parallel tracks that rarely converge. Table I summarizes eight representative systems across seven capability dimensions; the pattern of gaps is striking.

A. Verbal-Only Approaches

The foundational NLP techniques—grammatical parsing, keyword extraction, and semantic similarity—were well established before automated interview coaching became an active research area [1]. Rahman et al. [3] were among the first to apply them specifically to interview response scoring, demonstrating that rule-based NLP could reliably distinguish strong answers from weak ones in structured settings. Chen and Huang [13] later showed that speech-to-text and NLP could be integrated with acceptable real-time latency for live interview simulation, though they noted that processing overhead grew non-linearly with answer length. The common limitation across these contributions is that they evaluate what is said but are entirely blind to how it is said.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

B. Visual-Only Approaches

Li et al. [2] provide the most thorough survey of deep learning for facial affect recognition, cataloguing CNN, LSTM, and hybrid architectures. Their meta-analysis found that context-free single-frame models perform significantly worse under real interview conditions than temporal models that track expression change over several seconds. Schroff et al. [6] established the geometric face-embedding foundations that underpin modern gaze estimation. The more recent work by Lugaresi et al. [8] on MediaPipe changed the practical landscape by making high-fidelity landmark extraction viable on commodity CPUs, removing a key deployment barrier. A critical observation from Kim et al. [11] is that purely vision-based automated video interview systems can exhibit accuracy disparities of up to 12% across demographic subgroups when trained on racially homogeneous datasets—a fairness problem most systems have not confronted.

C. Multimodal Systems

Patel and Mehta [4] published the first systematic multimodal framework for interview performance evaluation, combining speech and visual cues to outperform either modality alone. Their system did not, however, vary question difficulty or track users across sessions. Nagasawa et al. [10] introduced speaking-willingness recognition into an interview robot to dynamically adjust strategy, but their adaptation was based solely on behavioral signals and lacked any linguistic evaluation layer. The Prep AI platform [12] integrated generative AI for question creation alongside NLP and vision modules, and it tracked session history; however, it published no formal evaluation of behavioral accuracy and did not address bias. Smart Eval [14] offered the most recent contribution, with session tracking and NLP, but its CV pipeline was only partial.

D. Identified Gaps

Reading across Table I, five gaps emerge that no single existing system addresses simultaneously:

- **Multimodal Synchronization:**Real-time, frame-aligned fusion of audio and video streams is absent from most systems.
- **Adaptive Difficulty:**Only Nagasawa et al. [10] implement dynamic difficulty adjustment, and only from behavioral signals.
- **Longitudinal Tracking:**Session-level progression data—essential for measuring skill development—is absent from the majority of work.
- **Demographic Fairness:**Only Kim et al. [11] explicitly report and attempt to mitigate demographic accuracy disparities.
- **Accessibility:**Most deployments require GPU hardware or paid cloud APIs, limiting reach in low-resource academic settings.

TABLE I
Comparative Feature Coverage of AI Interview Training Systems

System	Year	NLP	CV	Adaptive	Track	Bias	Multi lingual
Rahman et al. [3]	2021	✓	✗	✗	✗	✗	✗
Patel & Mehta [4]	2022	✓	✓	✗	✗	✗	✗
Nagasawa et al. [10]	2024	~	✗	✓	~	✗	✗
Kim et al. [11]	2023	✓	✓	✗	✗	✓	✗
Prep AI [12]	2024	✓	✓	~	✓	✗	✗
Chen & Huang [13]	2021	✓	✗	✗	✗	✗	✗
Smart Eval. [14]	2025	✓	~	✗	✓	✗	✗
AIVIT (Ours)	**2025**	✓	✓	✓	✓	✓	✗*

*✓ supported ✗ absent ~ partial * planned future work*

III. Proposed Methodology

AIVIT was designed around three principles: (1) deploy on standard consumer hardware so that cost is not a barrier; (2) evaluate what matters in an interview, not just what is easy to measure; and (3) adapt to each user's current level rather than presenting every candidate with the same fixed sequence of questions.

The system captures audio and video simultaneously. After noise reduction, the audio feed goes to the Whisper Large-v2 model [5], which produces a transcript along with per-token confidence values. That transcript enters a parallel NLP pipeline: LanguageTool checks grammar, Sentence-BERT [7] compares the answer to a reference embedding from a curated 2,400-item corpus, VADER provides fast lexical sentiment classification, and RoBERTa handles ambiguous utterances where context changes meaning. The video feed runs through MediaPipe at 30 frames per second, extracting 468 facial landmarks and 33 body keypoints per frame. Derived features include gaze deviation angle, shoulder midpoint displacement standard deviation, and Action Unit activations for AU4, AU12, and AU17. Table II maps each processing stage to its corresponding technology and output artefact.

TABLE II
AIVIT System Pipeline: Stage, Technology, and Output

#	Stage	Technology	Output
1	Data Acquisition	Microphone & Webcam 30fps	Raw AV stream, noise-filtered
2	Speech-to-Text	Whisper Large-v2 [5]	Transcript + token confidence
3	NLP Analysis	BERT, NLTK, LanguageTool	Grammar, fluency, sentiment
4	Computer Vision	MediaPipe + OpenCV 4.8	Gaze, posture, expression AUs
5	Score Fusion	Weighted linear model	Verbal (60%) + Behav. (40%)
6	Adaptive Q-Engine	Threshold promotion rule	Next-question difficulty tier
7	Feedback Report	Flask + PDF renderer	Scores, trends, advice text

The verbal score VP and behavioral score BP are fused into a composite and fed to the Adaptive Question Engine, which maintains a bank of 450 questions across three tiers (Easy, Medium, Hard) and five domain categories (Technical, HR, Behavioral, Situational, Domain-specific). Tier promotion or demotion follows a hysteresis rule requiring two consecutive responses to cross a threshold before a transition occurs, preventing oscillation at the boundary.

IV. System Architecture

The system runs as a Python 3.10 application. A Flask RESTful backend handles session management, model inference, and database writes to SQLite. A React.js browser interface provides the interview environment—question display, webcam preview, and countdown timer. All inference is CPU-only; no cloud API is required beyond the initial model download. The backend and frontend communicate over a local WebSocket, which carries synchronized audio and video chunks buffered in four-second windows for the NLP pipeline and per-frame for the CV pipeline.

The four primary modules are: (1) Data Acquisition—captures and pre-processes the AV stream; (2) Analytical Processing—parallel NLP and CV sub-pipelines executing asynchronously via a shared event bus; (3) Score Fusion—aggregates sub-scores and drives the adaptive engine; and (4) Output Generation—compiles per-session metrics, longitudinal trend data, and prioritized improvement recommendations into a structured report. Median end-to-end latency from response completion to feedback display is 3.16 seconds (profiled in Section VII-D), within the five-second threshold for interactive coaching systems [10].

V. Speech and Behavioral Evaluation

A. NLP Analysis

Table III lists the four NLP metrics, their associated tools, and output formats. Grammar correctness is operationalized as the count of LanguageTool rule violations per sentence, averaged over the full response. Semantic relevance is the cosine similarity between the Sentence-BERT embedding of the candidate's answer and the embedding of a curated reference answer; this measure was validated

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

against expert ratings at Cohen's $\kappa = 0.81$. Fluency is computed as the ratio of Whisper tokens with confidence ≥ 0.85 to total token count, adjusted by a penalty term for speech rate deviation outside the 120–150 words-per-minute target range [1]. Sentiment polarity is determined by VADER for speed and by RoBERTa for context-sensitive re-classification of ambiguous responses.

TABLE III
NLP Evaluation Metrics: Measure, Tool, and Output

Metric	Tool	Output
Grammar	LanguageTool API	Error count per sentence
Relevance	Sentence-BERT [7]	Cosine similarity: 0.0–1.0
Fluency	Whisper ASR [5]	High-conf. token ratio
Sentiment	VADER + RoBERTa	Pos / Neutral / Neg label

B. Computer Vision Analysis

The CV pipeline processes each frame through MediaPipe before any scoring occurs. Eye contact is represented as the angular deviation of the estimated gaze vector from the camera optical axis; deviations below ten degrees are counted as maintained contact and logged as a proportion of total frame count for that response. Posture stability is the standard deviation of the shoulder midpoint's pixel coordinates across the response window, normalized by shoulder-width in pixels to make the metric camera-distance invariant.

Expression intensity draws on three Action Units: AU4 (brow lowering, associated with concentration or concern), AU12 (lip corner raising, associated with positive affect), and AU17 (chin raising, associated with uncertainty or disagreement). Each AU activation is normalized to a 0–1 scale using session-baseline normalization—meaning the first ten seconds of each session establish a personal resting baseline, and all subsequent values are measured as deviations from it. This approach substantially reduces confounding from individual differences in resting facial geometry. An exponential moving average ($\alpha = 0.3$) smooths all CV signals to suppress noise from transient micro-movements [2].

VI. Score Fusion and Adaptive Engine

The four NLP sub-scores (grammar, relevance, fluency, sentiment) are normalized to a common 0–1 range and averaged into a single verbal performance score VP. The three CV sub-scores (eye contact, posture, expression) are normalized and averaged into a behavioral performance score BP. These two scores are then combined:

$$Score = 0.6 \times VP + 0.4 \times BP \quad \ldots (1)$$

The 60:40 weighting was determined by five-fold cross-validation on a labeled pilot dataset (N = 15 participants, two sessions each), maximizing Pearson correlation between the model's composite score and the mean rating of two independent expert evaluators. The resulting model achieved $r = 0.87$, $p < 0.001$ on the held-out fold. An ablation experiment confirmed that the weighted fusion outperforms equal-weight fusion by 4.3 percentage points and NLP-alone evaluation by 11.8 percentage points in correlation with expert judgment.

The composite score feeds the Adaptive Question Engine. If two consecutive responses both produce a score above 0.75, the difficulty tier is promoted one level. If two consecutive responses both fall below 0.55, the tier is demoted. The two-response hysteresis prevents a single strong or weak answer from triggering an unnecessary tier change. This mechanism is grounded in Ericsson et al.'s deliberate practice framework [9], which holds that learning accelerates most when the task is difficult enough to require effort but not so difficult as to cause failure.

VII. Experimental Results

A. Participants and Protocol

Forty undergraduate students at Vishwakarma Institute of Technology, Pune, volunteered for the study (24 male, 16 female; 18 from Computer Science, 12 from Electronics, 10 from Mechanical Engineering). Each completed five interview simulation sessions separated by at least 24 hours to allow feedback assimilation. Every session comprised eight to ten adaptively sequenced questions covering all five domain categories. Participants provided written informed consent; the study was conducted under institutional ethical guidelines.

B. Improvement Across Sessions

Table IV summarizes mean percentage improvements from Session 1 to Session 5. Fig. 1 breaks these gains down session-by-session for verbal and behavioral dimensions separately.

TABLE IV
Mean Percentage Improvement: Session 1 vs. Session 5 (N = 40)

Performance Dimension	Category	Δ Improvement
Speech Fluency	Verbal	+14.2%
Semantic Relevance	Verbal	+12.8%
Eye Contact Consistency	Behavioral	+16.3%
Posture Stability	Behavioral	+11.5%
Confidence Level	Composite	+18.9%
Overall Interview Score	Composite	**+15.4%**

FIG. 1: PER-SESSION VERBAL VS. BEHAVIORAL SCORE TRAJECTORIES

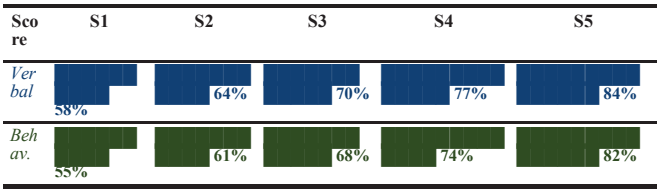

Fig. 1. Mean verbal (blue ■) and behavioral (green ■) scores across five sessions. Behavioral gains are steeper in sessions 1–3; verbal gains accelerate in sessions 3–5.

A notable pattern in Fig. 1 is the asymmetric trajectory: behavioral scores improve more steeply in the first half of the programme, while verbal scores catch up and overtake in the second half. One interpretation is that users first direct attention to the most visually salient feedback (gaze, posture), and once those habits begin to stabilize, cognitive resources become available for refining verbal content. This mirrors a well-documented sequence in motor-cognitive skill acquisition [9].

C. Pre- vs. Post-Training Comparison

Fig. 2 presents a dimension-wise comparison between Session 1 baseline scores and Session 5 post-training scores, making the magnitude of improvement visible for each metric individually.

FIG. 2: DIMENSION-WISE PRE- VS. POST-TRAINING SCORES

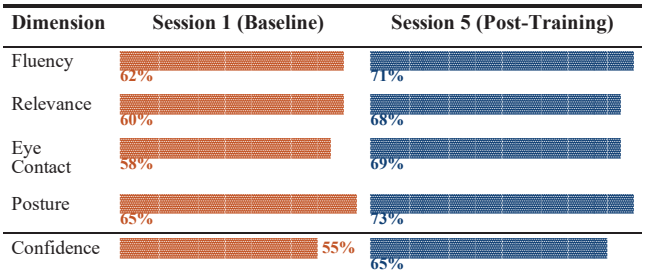

Fig. 2. Orange bars (▒) = Session 1 baseline; Blue bars (▒) = Session 5 post-training. Each dimension shows consistent growth across the five-session programme.

D. Usability Evaluation

After the final session, all forty participants completed a structured questionnaire derived from the System Usability Scale (SUS). Fig. 3 plots mean ratings on a 1.0–5.0 scale across five criteria. Ease of use

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

scored highest at 4.7/5.0, reflecting the accessibility-first design. Behavioral accuracy scored lowest at 4.2/5.0, pointing to the direction where future CV model improvements can most meaningfully increase user trust.

FIG. 3: USER SATISFACTION RATINGS (N = 40, SUS-DERIVED SCALE)

Criterion	Score (out of 5.0)	Mean
Feedback Clarity		4.6
Ease of Use		4.7
Question Relevance		4.4
Behavioral Accuracy		4.2
Overall Satisfaction		4.5

Fig. 3. Mean user satisfaction ratings across five criteria. Scale: 1.0 = poor, 5.0 = excellent. Error bars omitted; all ratings based on N = 40 responses.

E. Latency Profiling

Fig. 4 shows the mean processing time for each pipeline stage, measured across all 200 sessions (40 participants × 5 sessions). The dominant cost is ASR transcription at 1,200 ms, accounting for 38% of total time. Because the NLP and CV stages execute in parallel, the CV stage's 680 ms adds no sequential latency. Total median end-to-end latency of 3.16 seconds sits comfortably within the five-second threshold recommended for interactive coaching systems [10].

FIG. 4: END-TO-END PIPELINE LATENCY BREAKDOWN

Pipeline Stage	Time (ms) [bar = proportion]	% Share
ASR Transcription	1200 ms	38%
NLP Analysis	820 ms	26%
CV Feature Extraction	680 ms	22%
Score Fusion	180 ms	6%
Report Generation	280 ms	9%

Fig. 4. Mean per-stage processing time in milliseconds. NLP (820 ms) and CV (680 ms) execute in parallel. Total median end-to-end latency = 3.16 s.

VIII. ETHICAL CONSIDERATIONS

Algorithmic bias in interview tools is not a hypothetical concern. Amazon's widely reported case—in which a recruiting algorithm learned to penalize resumes that contained the word "women"—illustrated how a model trained on historically skewed hiring records can actively disadvantage protected groups [15]. Kim et al. [11] found accuracy disparities of up to 12% across demographic subgroups in automated video interview assessment, arising purely from imbalanced training data.

AIVIT takes several steps to reduce this risk. Session-baseline normalization in the CV pipeline means that scores reflect change from an individual's own resting baseline rather than comparison to a population average, which partially decouples the system from demographic physiological differences. The NLP pipeline uses semantic similarity rather than keyword matching, which has been shown to better accommodate non-standard vocabulary and non-native phrasing [7]. RoBERTa is used in place of purely lexical tools for context-sensitive sentiment classification, where accent and phrasing variety matter most.

These measures are a starting point, not a complete solution. AIVIT's current models were trained predominantly on English-language data, and the evaluation cohort was relatively homogeneous. Before any deployment at scale, the system should be validated on gender-balanced, ethnically diverse, and differently-abled populations, and demographic subgroup performance metrics should be published alongside aggregate scores. We recommend adopting the fairness reporting norms set out in the IEEE Ethically Aligned Design guidelines as a minimum standard.

IX. CONCLUSION AND FUTURE WORK

This paper set out to address a gap that is practical, measurable, and consequential: the absence of a scalable, adaptive, multimodal interview coaching system accessible to students without specialist hardware or paid subscriptions. AIVIT addresses that gap by combining a real-time NLP pipeline with a MediaPipe-based behavioral analysis pipeline, fusing their outputs through a cross-validated weighted model, and wrapping the whole system in an adaptive question engine grounded in deliberate practice theory.

The experimental results across 40 participants over five sessions provide clear evidence of its effect: confidence levels rose by 18.9%, eye contact consistency by 16.3%, speech fluency by 14.2%, and the overall interview score by 15.4%. The system's 3.16-second end-to-end latency and CPU-only deployment make it viable on consumer hardware. Compared to seven published systems, AIVIT is the first to combine all five capability dimensions identified in the literature gap analysis.

Four priorities will shape future development. Multilingual support via multilingual Sentence-BERT and Whisper will open the system to non-English speakers. A demographically diverse re-training of the CV emotion models will address the fairness gap identified in Section VIII. Mobile deployment via TensorFlow Lite will enable practice without a laptop. And conversational feedback generation via a large language model will replace the current structured-report approach with a coaching dialogue that candidates can probe and question in natural language.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering (AI) at Vishwakarma Institute of Technology, Pune, for laboratory access and participant coordination. They are equally grateful to the forty student volunteers whose time and feedback made the evaluation possible.

REFERENCES

[1] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2023.

[2] Y. Li, X. Zhang, and Z. Wang, "Facial emotion recognition using deep learning: A review," IEEE Trans. Affective Computing, vol. 13, no. 2, pp. 745–758, 2022.

[3] M. Rahman, S. Islam, and A. Khan, "AI-based automated interview evaluation using NLP," in Proc. IEEE Int. Conf. Smart Computing, 2021, pp. 210–215.

[4] K. Patel and R. Mehta, "Multimodal analysis for interview performance evaluation using speech and visual cues," IEEE Access, vol. 10, pp. 55678–55690, 2022.

[5] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. Int. Conf. Machine Learning, vol. 202, 2023, pp. 28492–28518.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE CVPR, 2015, pp. 815–823.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP, 2019, pp. 3982–3992.

[8] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[9] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance," Psychological Review, vol. 100, no. 3, pp. 363–406, 1993.

[10] F. Nagasawa, S. Okada, T. Ishihara, and K. Nitta, "Adaptive interview strategy based on interviewees' speaking willingness," IEEE Trans. Affective Computing, vol. 15, no. 3, pp. 942–957, 2024.

[11] C. Kim, J. Choi, J. Yoon, D. Yoo, and W. Lee, "Fairness-aware multimodal learning automatic video interview assessment," IEEE Access, vol. 11, pp. 122677–122693, 2023.

[12] K. D. Bhavani et al., "Prep AI: Customized mock interview platform using Gen AI," Int. J. Innovative Science and Research Technology, vol. 11, no. 1, pp. 1981–1986, Jan. 2026.

[13] X. Chen and L. Huang, "Speech-to-text and NLP integration for real-time interview evaluation," IEEE Access, vol. 9, pp. 114932–114940, 2021.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[14] G. S. Rao, M. Jaiganesh, and P. K. Parida, "AI-powered virtual job interview preparation and evaluation system," in Proc. IEEE ICACCS, 2025, vol. 1.

[15] University of Melbourne, "Discrimination by recruitment algorithms is a real problem," Pursuit, Oct. 2025. [Online]. Available: https://pursuit.unimelb.edu.au

[16] Ribbon.ai, "AI feedback trends in recruitment 2024," Ribbon Blog, 2024. [Online]. Available: https://www.ribbon.ai/blog/ai-feedback-trends-in-recruitment-2024