AI-Interviewer: A Multimodal Framework for Dynamic and Explainable Automated Job Interviews

Esraa M. Abdelhadi, Madiha S. Farouq, Farida K. Ali, Mohammad T. Omar, Muhammad Y. Abdelmoaty, Hania R. Mahmoud, Soha A. Ehssan

Department of Artificial Intelligence – Faculty of Computers and Artificial Intelligence

Helwan University, Cairo, Egypt

{esraaammar869, faridakh222, mohammadtarekomar, muhammaddyasserr, madihasaeidfarouq, haniarouby}@gmail.com

dr.soha@fci.helwan.edu.eg

Abstract—**This paper introduces AI-Interviewer, a fully automated platform designed to address the persistent challenges of inefficiency, human bias, and subjectivity in traditional job interviews. The system provides a scalable and fair solution for both recruiters and applicants by simulating a realistic, interactive interview experience. Our methodology integrates several state-of-the-art AI components into a unified pipeline. A Large Language Model enhanced with Retrieval-Augmented Generation (RAG) is used for dynamic, context-aware question generation tailored to the job description and candidate's curriculum vitae. During the session, the system captures and analyzes multimodal data, including real-time speech-to-text transcription, facial emotion recognition, and vocal emotion analysis. These inputs are processed by a novel judgment agent, which evaluates candidate responses against an AI-generated ideal answer, assigning qualitative scores and providing transparent, explainable reasoning. The system successfully delivers high-quality interview sessions and generates detailed performance reports, including emotion trends, structured feedback, and an AI-supported hire or reject recommendation. By enabling efficient, fair, and explainable assessments, the AI-Interviewer platform presents a significant advancement in recruitment technology.**

Keywords—Artificial Intelligence, Interview Automation, Multimodal Assessment, Emotion Analysis, Question Generation, Judgment Agent, Explainable AI.

I. INTRODUCTION

The recruitment process is a pivotal function in modern organizations, directly influencing performance, innovation, and workplace culture. However, traditional interview methods are frequently beset by significant challenges, including time-inefficiency, subjectivity, and the pervasive issue of human bias. Recruiters are often overwhelmed by the volume of applicants, leading to decision fatigue and inconsistent evaluation standards, which can inadvertently filter out qualified candidates and reduce diversity. On the other side of the hiring desk, job seekers often lack access to realistic practice environments, hindering their ability to prepare effectively and showcase their true potential.

While recent advancements in Artificial Intelligence (AI) have introduced automation into the hiring pipeline, many existing solutions address these problems in a fragmented manner. For instance, many AI-driven platforms rely on monomodal analysis, such as processing only textual or speech data, thereby failing to capture the rich, non-verbal cues like facial expressions and vocal tonality that are integral to human communication. Furthermore, a significant number of these systems employ static or pre-defined question banks, which lack the adaptability to probe a candidate's specific skills and experience in relation to the job requirements. This results in a generic and often superficial assessment. Finally, the "black-box" nature of many AI evaluation tools offers little transparency, eroding trust and providing no actionable feedback for candidate development.

To address these critical gaps, we present **AI-Interviewer**, a fully automated, multimodal interview platform designed to simulate a realistic, fair, and insightful evaluation experience. Our system leverages a suite of integrated AI components to create a comprehensive assessment framework. At its core is a dynamic question generation engine, powered by a Large Language Model (LLM), **LLaMA 3.3**, and enhanced with a Retrieval-Augmented Generation **(RAG)** pipeline. This enables the generation of context-aware questions tailored specifically to the candidate's CV and the job description.

During the interview, our system performs real-time multimodal analysis by capturing and interpreting the candidate's textual responses (via Whisper), vocal emotions (via a fine-tuned wav2vec 2.0 model), and facial expressions (via the DeepFace framework). The evaluation is conducted by a transparent AI Judgment Agent, based on the Mistral model, which provides not only a qualitative score but also a clear, human-readable rationale for its assessment. This dual-purpose platform serves as a powerful efficiency tool for recruiters and a valuable training tool for job seekers, offering them structured, personalized feedback to enhance their preparedness.

II. Related Work

In recent years, numerous studies have explored the application of artificial intelligence in job interviews and candidate evaluation. One such study, by Anumeha et al., proposes a multimodal system that combines audio, video, and textual data to assess interviewee behavior and performance using behavioral analytics techniques [1]. Similarly, the work of Surendra et al., investigates the use of generative AI models to simulate interview environments and evaluate candidates based on response coherence and intent [2].

Speech emotion recognition has also been a critical component in several AI-based interview systems. For example, the study by K. Venkataramanan and H. R. Rajamohan. focuses on using deep learning architectures for extracting emotional states from vocal cues, which can enhance the understanding of candidate behavior during interviews [3]. Another relevant contribution is the system developed by Patil and Ghorpade, which emphasizes the implementation of a rule-based and NLP-enhanced AI interviewer chatbot designed for HR and technical assessments [4].

A more structured approach to AI interviews was presented by B C Lee and B Y Kim, who developed a remote hiring tool based on AI-driven scoring mechanisms and structured questioning [5]. Additionally, Jadhav et al. introduced a multimodal AI framework that integrates emotion and behavioral analysis for deeper assessment of candidate personality traits [6].

These works collectively demonstrate the growing interest in automating interviews using AI techniques. However, they also reveal limitations related to dynamic adaptability, holistic feedback mechanisms, and integration into comprehensive platforms, which remain open areas for innovation.

III. PROPOSED MODEL AND SYSTEM ARCHITECTURE

The AI-Interviewer system is designed as a modular, end-to-end pipeline that automates the interview process from context acquisition to final assessment. The architecture orchestrates several specialized AI components to simulate an interactive, human-like interview and provide a comprehensive, data-driven evaluation. Fig. 1 illustrates the comprehensive architecture of our proposed system, which comprises four main stages: Input & Contextualization, Dynamic Question Generation, Multimodal Interview Interaction & Analysis, and AI-Powered Evaluation & Reporting.

A. Dynamic Question Generation Module

The foundation of a relevant interview is the quality of its questions. Our system moves beyond static question banks by employing a dynamic question generation module. This module leverages **LLaMA 3.3 (70B) model** integrated with a **Retrieval-Augmented Generation (RAG)** pipeline to produce questions that are both contextually relevant and non-repetitive.

The process is as follows:

1. ***Indexing:*** The system first ingests and parses the candidate's curriculum vitae (CV) and the job description, creating a vectorized knowledge base of the candidate's skills, experience, and the role's requirements.
2. ***Retrieval:*** At runtime, for each new question, the RAG system retrieves the most relevant text chunks from the indexed documents.
3. ***Generation:*** These retrieved chunks are then passed as context along with a structured prompt to the LLaMA 3.3 model. This grounds the model, enabling it to generate highly specific questions that probe the intersection of the candidate's background and the job's needs.
4. ***Ideal Answer Generation:*** Concurrently, for each question generated, the module also prompts the LLaMA 3.3 model to produce a high-quality "ideal answer." This answer serves as an objective benchmark for the subsequent evaluation stage, outlining the expected knowledge, clarity, and depth.

B. Multimodal Interaction and Analysis

To create an immersive and realistic interview experience, the system facilitates a voice- and video-based interaction, analyzing multiple data modalities in real-time.

1. **Text-to-Speech (TTS):** The AI interviewer's voice is synthesized using **Bark**, an advanced text-to-audio model. Bark was selected for its ability to produce highly natural and expressive speech, including non-verbal vocalizations such as pauses and variations in tone, which enhances the realism of the interaction.
2. **Speech-to-Text (STT):** The candidate's spoken responses are transcribed into text using OpenAI's **Whisper** model. Whisper is renowned for its state-of-the-art accuracy, robustness to diverse accents and background noise, and proficiency with conversational language, ensuring a reliable textual representation of the candidate's answers.
3. **Vocal Emotion Recognition:** To analyze the prosodic cues in the candidate's speech, we employ a **wav2vec 2.0 model fine-tuned** on a composite dataset of emotion-labeled corpora, including RAVDESS, TESS, and EMO-DB. This model detects emotional states such as happiness, anger, sadness, and neutrality from the user's vocal tone, providing insights into their confidence and disposition.

4. **Facial Emotion Analysis:** Complementing the vocal analysis, a facial emotion analysis module processes the candidate's video stream. The system leverages the **DeepFace framework**, which integrates multiple pre-trained deep learning models (e.g., VGG-Face) for high-accuracy emotion detection. To ensure real-time performance, the system analyzes video frames at a configurable sampling rate (e.g., every 15th frame), identifying dominant emotions such as happy, sad, angry, surprise, and neutral.

C. AI-Powered Judgment Agent

The evaluation of the candidate's performance is handled by a sophisticated Judgment Agent. This component is designed for fairness, consistency, and, critically, explainability. It uses the **Mistral-7B-Instruct-v0.3** model to evaluate the candidate's transcribed response against the pre-generated ideal answer.

The evaluation is based on four primary criteria:

- **Relevance:** How well the answer addresses the question.
- **Correctness:** The factual accuracy of the information provided.
- **Clarity:** The structure and coherence of the explanation.
- **Completeness:** The extent to which the answer covers all expected points.

For each response, the Judgment Agent outputs a qualitative score (e.g., Excellent, Good, Medium, Poor) and, most importantly, a **concise, human-readable rationale** justifying the score. This explainable AI (XAI) approach ensures transparency and provides actionable feedback by highlighting specific strengths and areas for improvement.

D. Final Report Generationn

Upon completion of the interview, the system synthesizes all collected data into a comprehensive performance report. This final stage is orchestrated by the **LLaMA 3.3** model, which processes the accumulated scores, rationales, and multimodal emotion data. The generated report includes:

- An overall assessment of the candidate's performance.
- A breakdown of question-wise scores with justifications from the Judgment Agent.
- A summary of detected emotional trends from both vocal and facial analysis.
- A transparent, AI-assisted **hire or reject recommendation**, grounded in the evidence collected throughout the interview.

IV. Results

To evaluate the performance and efficacy of the AI-Interviewer system, we conducted a series of experiments focusing on three key areas: the accuracy of the multimodal emotion recognition modules, the quality of the dynamic question generation, and the overall behavioral fidelity of the AI interviewer. Due to the absence of large-scale, gold-standard datasets for this domain, our evaluation combines quantitative metrics from controlled tests and qualitative analysis based on structured rubrics.

A. Experimental Setup

All experiments were conducted on a system with an Intel Core i7 CPU, 16 GB of RAM, and an NVIDIA GTX 1650 GPU. The software stack included Python 3.10, PyTorch 2.0, and HuggingFace Transformers. The emotion recognition models were fine-tuned on a composite dataset including RAVDESS, TESS, CREMA-D, and EMO-DB. The question generation module was developed using a custom dataset of 3,809 interview questions scraped from public sources and categorized across 26 unique job roles.

B. Accuracy of Multimodal Emotion Recognition

The performance of the emotion detection components is critical for providing nuanced behavioral insights.

- **Vocal Emotion Recognition:** The fine-tuned **wav2vec 2.0** model achieved an approximate classification accuracy of **85%** across seven emotion classes (happy, sad, angry, fear, disgust, surprise, neutral) on our held-out test set.
- **Facial Emotion Recognition:** The **DeepFace** framework was validated qualitatively. It demonstrated robust performance, consistently identifying dominant facial emotions across a range of lighting conditions and head orientations in real-time video streams.

These results confirm that the system's multimodal emotion analysis is sufficiently robust for providing consistent behavioral indicators.

C. Quality of Dynamic Question Generation

We evaluated multiple versions of the LLaMA model to optimize for relevance and diversity in question generation. The final model, **LLaMA 3.3 (70B)**, integrated with our RAG pipeline, demonstrated superior performance. As shown in our internal tests (detailed in Chapter 7.2.2), this model successfully **eliminated the repetition issue** observed in earlier versions. Furthermore, it delivered adaptive, seniority-aligned questions that were highly specific to the job role, with the Qdrant-based context injection significantly improving personalization and logical flow in multi-turn conversations.

D. Overall System Behavior and Fidelity

To assess the system's performance in a realistic scenario, a series of mock technical interviews were conducted. The transcripts were then evaluated by a higher-capability language model (**GPT-4**) against a 5-point Likert scale rubric covering key aspects of a human-like interview. The results are summarized in Table I.

Table I

Dimension	*Score (5)*	*Comments*
Question Relevance	4.9	Highly aligned with role; technically deep.
Contextual Responsiveness	4.7	Logically adapted to prior answers.
Evaluation and Scoring	4.8	Consistent and well-reasoned scoring.
Language Coherence	4.6	Formal and precise but lacked warmth.
Behavioral Realism	4.4	Professional tone, but emotionally neutral.

V. Conclusion

We presented **AI-Interviewer**, a novel framework designed to address the inherent challenges of bias, inefficiency, and subjectivity in traditional recruitment processes. Our work demonstrates that an integrated, multimodal AI system can provide a robust, fair, and effective alternative to conventional interview methods.

The core of our contribution lies in a synergistic architecture that combines dynamic, context-aware question generation using a **RAG-enhanced LLaMA 3.3 model** with comprehensive multimodal analysis of a candidate's speech, vocal tonality, and facial expressions. The subsequent evaluation, performed by an explainable **Mistral-based Judgment Agent**, ensures that assessments are not only consistent but also transparent, providing clear rationale for every score. Our experimental evaluation confirmed the system's high fidelity, achieving an aggregate performance score of **4.68 out of 5** in replicating technical interviews and successfully resolving key issues like question repetition and model overfitting that plagued earlier designs.

The significance of this work is twofold. First, it offers a scalable and efficient tool for recruiters, reducing manual overhead and mitigating human bias. Second, it serves as an invaluable training platform for job seekers, providing a realistic simulation environment with actionable, data-driven feedback. This dual-purpose design is a key achievement, creating a more equitable and developmental recruitment ecosystem.

While the results are promising, this work acknowledges limitations, primarily the need for large-scale validation with HR professionals and diverse demographic groups to ensure fairness and generalizability. Future work will focus on enhancing conversational realism to better assess behavioral and soft skills, expanding model capabilities across different languages, and pursuing deeper integration with existing Applicant Tracking Systems (ATS). In conclusion, the AI-Interviewer project lays a strong foundation for the next generation of recruitment technology, paving the way for a more intelligent, fair, and human-centric approach to hiring.

VI. Future Work

The promising results of the AI-Interviewer system lay the groundwork for several key areas of future research and development. Our efforts will be directed toward enhancing the system's robustness, scalability, and user-centric features.

1. **Transition to a Real-Time, Production-Ready Application**: Transition the system from its current prototype to a scalable, production-grade web application. This involves optimizing real-time performance on cloud infrastructure to create a secure, multi-tenant service for both organizations and individual users.
2. **Customizable, Constraint-Based Candidate Filtering**: Empower recruiters by implementing a constraint-based filtering mechanism. This feature will allow companies to define custom evaluation criteria and weights, enabling the AI to generate hiring recommendations that are precisely aligned with their specific needs.
3. **Expansion to Multilingual Support**: To address the demands of the global employment landscape, a crucial next step is to extend the platform's capabilities to multiple languages. This will involve fine-tuning the STT, TTS, and LLM components on multilingual datasets. A key challenge will be to ensure that the multimodal emotion analysis and scoring rubrics are culturally sensitive and perform equitably across different linguistic and cultural contexts.
4. **Continuous Model Improvement through Reinforcement Learning**: We aim to implement a continuous learning loop to progressively enhance the system's intelligence. The anonymized data generated from completed interviews, including candidate responses, evaluation scores, and final outcomes, will be used as a feedback signal. This data can be leveraged to further fine-tune the question generation and judgment models, potentially using reinforcement learning from human feedback (RLHF) techniques, to improve their accuracy, fairness, and conversational nuance over time.
5. **Dataset Expansion and Domain Generalization**: The performance of our system is fundamentally linked to the quality and diversity of its underlying data. Future work will focus on significantly expanding our interview question dataset to cover a broader range of job roles, industries, and seniority levels. We will also explore techniques for better domain generalization, enabling the system to perform effectively even for niche or newly emerging job functions with limited available data.

VII. References

[1] A. Agrawal, R. A. George, S. S. Ravi, S. Kamath and A. Kamar "Leveraging Multimodal Behavioral Analytics for Automated Job Interview Performance Assessment and Feedback," *arXiv preprint arXiv:2006.07909*, 2020. [Online]. Available: https://arxiv.org/pdf/2006.07909

[2] S. Mahajan, N. Sonawane, R. Bagal, S. Kulkarni, A. Suryawanshi and Y. Mhaisne, "Generative AI-Based Interview Simulation and Performance Analysis," *International Research Journal of Modernization in Engineering Technology and Science (IRJMETS)*, vol. 6, no. 5, May 2024. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper//issue_5_may_2024/56275/final/fin_irjmets1715666745.pdf

[3] K. Venkataramanan and H. R. Rajamohan, "Emotion Recognition from Speech," *arXiv preprint arXiv:1912.10458*, 2019. [Online]. Available: https://arxiv.org/pdf/1912.10458

[4] A. Patil and V. Ghorpade, "AI Interviewer Chatbot for Technical and HR Brilliance: A Tool for Upskilling Candidate," *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 1, pp. 234–239, Jan.–Apr. 2020. [Online]. Available: https://www.journal-dogorangsang.in/no_1_jan-april_20/57.pdf

[5] B C Lee and B Y Kim, "Development of an AI-Based Interview System for Remote Hiring," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 12, no. 3, pp. 537–544, Mar. 2021. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/66683706/IJARET_12_03_060-libre.pdf

[6] A. Jadhav, R. Ghodake, K. Muralidharan, G. T. Varma and V. Bharathi J., "AI-Based Multimodal Emotion and Behavior Analysis of Interviewee," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 7, no. 5, May 2023. [Online]. Available: https://www.researchgate.net/publication/370653388