International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 06 June-2026 | Impact Factor: 3.5

AI-Powered Virtual Interviewer with Real-Time Feedback

Prof. M.D. Ingle
Department of Computer Engineering
Jayawantrao Sawant College of Engineering
Pune, India

Pournima Mali
Department of Computer Engineering
Jayawantrao Sawant College of Engineering
Pune, India

Jagruti More
Department of Computer Engineering
JSPM's Jaywantrao Sawant College of Engineering
Pune, India

Anand Magar
Department of Computer Engineering
Jayawantrao Sawant College of Engineering
Pune, India

Kunal Patil
Department of Computer Engineering
JSPM's Jaywantrao Sawant College of Engineering
Pune, India

Abstract - The interview process is a critical gateway to professional employment, yet it remains heavily reliant on subjective human evaluation, scheduling constraints, and inconsistent assessment criteria. Candidates from diverse backgrounds often lack access to structured practice environments that can simulate real interview scenarios and deliver actionable, personalized feedback. This paper proposes an AI-Powered Virtual Interviewer system that addresses these limitations by delivering an end-to-end, automated interview experience with real-time, multi-dimensional feedback. The proposed system dynamically generates domain-specific questions using a fine-tuned large language model, transcribes candidate responses through an advanced automatic speech recognition module, evaluates answer quality through semantic similarity scoring, and simultaneously analyzes communication cues including speech fluency, sentiment, and facial expressions. A personalized post-interview feedback report is automatically generated and delivered to the candidate. This paper surveys five closely related research works in automated interview analysis, spoken dialogue systems, virtual coaching agents, and LLM-based question generation, identifies persistent gaps in existing approaches, and describes how the proposed system is designed to close them.

Keywords - Automated Interview System; Large Language Models; Automatic Speech Recognition; Real-Time Feedback; Sentiment Analysis; Facial Expression Analysis; Question Generation; Natural Language Processing; BERT; Whisper STT

I. INTRODUCTION

Recruitment and candidate assessment are foundational activities across every professional sector, yet the conventional interview process carries well-documented inefficiencies. Human interviewers are constrained by cognitive biases, schedule availability, and the inherent difficulty of consistently applying standardized evaluation rubrics across multiple candidates. Studies in organizational psychology have repeatedly demonstrated that unstructured interviews exhibit limited predictive validity for job performance, underscoring the need for more systematic and reproducible evaluation approaches. Simultaneously, candidates preparing for competitive job markets face a significant asymmetry of resources. While coaching academies and online mock-interview platforms exist, most provide either scripted question sets with no adaptive complexity, or generic feedback that fails to address individual communication patterns, content accuracy, and non-verbal cues holistically. This gap is particularly acute for students in tier-two and tier-three institutions in developing economies, where access to professional mentoring is limited.

The rapid maturation of artificial intelligence technologies, including large language models (LLMs), transformer-based natural language processing, and real-time computer vision, has created an unprecedented opportunity to automate and personalize the interview preparation experience. Conversational AI systems can now generate contextually appropriate questions, evaluate the semantic quality of unstructured spoken responses, and analyze paralinguistic signals such as speech rate, pause distribution, and facial affect, all within a single, unified computational pipeline.

This paper presents a survey grounding the design of an AI-Powered Virtual Interviewer system. The system is built around five core functional modules: (i) dynamic question generation using a fine-tuned GPT-class model, (ii) real-time speech transcription via Whisper v3, (iii) semantic answer quality evaluation using BERT-based scoring, (iv) multi-modal communication analysis encompassing sentiment and facial expression recognition, and (v) automated generation of a structured, personalized feedback report. The contributions of this survey are threefold: it reviews five closely related works that collectively motivate the system's design, presents a comparative analysis exposing persistent gaps in the state of the art, and describes the proposed architecture that closes those gaps.

The remainder of this paper is organized as follows. Section II reviews the five core related works that motivate and inform the proposed system. Section III presents a comparative analysis of these methods. Section IV identifies the research gaps that the proposed system is designed to address. Section V describes the system design and architecture. Section VI concludes with key findings and future directions.

II. LITERATURE REVIEW

A. Multimodal Interview Analysis Using Deep Learning

Hemamou et al. [1] introduced a multimodal neural framework for predicting interview outcomes by jointly processing video, audio, and textual modalities extracted from recorded candidate responses. Their architecture employed a hierarchical LSTM to model temporal dependencies across these

modalities and produced personality trait estimates aligned with the Big Five psychological model. The work demonstrated that audio-visual cues carry statistically significant predictive information beyond textual content alone. However, the system operates exclusively in an offline, post-hoc evaluation mode, requiring full video recordings as input. It provides no conversational capability, generates no adaptive questions in response to candidate answers, and delivers no real-time coaching feedback. These limitations directly motivate the proposed system's emphasis on live, session-aware interaction rather than retrospective analysis.

B. Predicting Interview Performance from Behavioral Cues

Naim et al. [2] investigated the automated prediction of interview performance scores using audio-visual behavioral signals, including speaking rate, eye gaze patterns, head nodding frequency, and smile intensity. Their regression-based models were trained on annotated interview datasets and achieved moderate correlation with human-assigned interview ratings. The study produced valuable insights into the behavioral markers that distinguish high-performing candidates from lower-performing peers. Nonetheless, the framework is constrained to evaluation and does not address the generative dimension of interviews, specifically the capacity to pose contextually adaptive questions, assess the semantic content of answers, or provide prescriptive feedback. The proposed AI Virtual Interviewer extends this line of inquiry by integrating behavioral signal analysis within a live, interactive pipeline.

C. Virtual Agent for Social Skill Training

Hoque et al. [3] developed MACH (My Automated Conversation coacH), a virtual human agent designed to train social and communication skills relevant to job interviews. MACH used real-time facial expression tracking, speech analysis, and a scripted rule-based dialogue engine to guide participants through simulated interview practice scenarios. Participants who trained with MACH showed measurable improvements in interpersonal skills as assessed by independent evaluators. Despite its innovation, MACH's dialogue management is fundamentally rule-based, making it incapable of generating novel or domain-adaptive questions. Its feedback mechanism relies on pre-authored templates rather than a contextual understanding of the specific content each candidate articulates. The proposed system supersedes this limitation by deploying a large language model for both dynamic question generation and content-aware feedback synthesis.

D. Spoken Dialogue Systems for Interview Domains

Chen and Jokinen [4] explored the application of statistical spoken dialogue systems to the interview domain, focusing on turn-taking management, dialogue state tracking, and response coherence in candidate-interviewer exchanges. Their work established formal models for multi-turn interview dialogue and demonstrated that dialogue history context significantly improves the relevance of follow-up questions. The study, however, operated in a simulated text-only environment and did not incorporate speech recognition, answer quality evaluation, or any feedback generation module. It also predates the transformer era and thus does not leverage the semantic richness afforded by modern pre-trained language models. The proposed system incorporates dialogue-aware question sequencing while extending it with voice interaction and semantic evaluation capabilities.

E. LLM-Based Question Generation for Domain-Specific Assessment

Guo et al. [5] examined the use of fine-tuned GPT-class models for automated question generation targeted at domain-specific knowledge assessments. Their approach combined retrieval-augmented generation with a supervised fine-tuning strategy to produce questions that are contextually grounded, non-repetitive, and calibrated to specified difficulty levels. Evaluation using BLEU and human judgment metrics showed high question quality scores exceeding prior rule-based and template-filling methods. The work, however, is scoped exclusively to the generation task and does not address how generated questions are delivered in a live session, how candidate responses are transcribed and evaluated, or how feedback is formulated from evaluation outcomes. The proposed system adopts the retrieval-augmented question generation paradigm established by this work and situates it within a complete end-to-end interview pipeline.

III. COMPARATIVE ANALYSIS OF EXISTING METHODS

Table I consolidates the five surveyed works across dimensions most relevant to the proposed system: task domain, core methodology, deployment platform, processing speed, reported accuracy, and whether a full end-to-end pipeline is provided.

TABLE I. COMPARATIVE SUMMARY OF RELATED WORK VS. AI VIRTUAL INTERVIEWER (PROPOSED)

Work	Task / Domain	Method	Accuracy
Hemamou et al. [1]	Automated interview analysis	Multimodal LSTM (video + audio + text)	High (valence/arousal)
Naim et al. [2]	Interview performance prediction	Audio-visual cues + ML regression	Moderate (r=0.48)
Hoque et al. [3]	Social skill training via virtual agent	Rule-based spoken dialogue + affective feedback	Not reported
Chen & Jokinen [4]	Spoken dialogue systems for interviews	Statistical dialogue management	Not reported
Guo et al. [5]	LLM-based question generation	GPT fine-tuning + retrieval	High (BLEU >0.7)
AI Virtual Interviewer (Proposed)	Question generation, speech eval, sentiment, feedback	LLM + Whisper STT + BERT sentiment + GPT scoring + FastAPI	Target >85%

a.

Several key observations emerge from this comparison. First, none of the surveyed works offers a complete end-to-end pipeline that spans question generation, live speech transcription, semantic answer evaluation, behavioral signal analysis, and automated feedback delivery within a single unified system. Second, the highest-performing analytical systems, such as Hemamou et al. [1] and Naim et al. [2], operate exclusively in offline mode and thus cannot support real-time candidate interaction. Third, while Hoque et al. [3] comes closest to an interactive experience, its rule-based dialogue engine fundamentally limits conversational depth and content adaptability. Fourth, LLM-based question generation as

demonstrated by Guo et al. [5] remains disconnected from evaluation and feedback stages. The proposed system is specifically architected to bridge all these dimensions into a coherent, deployable solution.

IV. RESEARCH GAP

The surveyed literature reveals five persistent and interrelated gaps that collectively motivate the design of the proposed AI Virtual Interviewer system.

A. Absence of a Unified Real-Time Interview Pipeline

Every surveyed work targets an isolated functional component: multimodal analysis [1], behavioral prediction [2], social skill coaching [3], dialogue management [4], or question generation [5]. No existing system integrates all these capabilities into a single real-time pipeline that a candidate can interact with through natural spoken language without human involvement. The proposed system directly addresses this fragmentation by unifying all stages, from question delivery through speech capture, transcription, semantic scoring, behavioral analysis, to feedback generation, within a single session-aware platform.

B. Lack of Semantic Answer Evaluation

Existing interactive systems, such as MACH [3], assess candidate performance primarily through surface-level behavioral signals such as smile intensity and eye contact, without evaluating the semantic content of spoken answers against domain-specific correctness criteria. This means a candidate could receive a positive communication score while providing factually incorrect or superficial responses. The proposed system addresses this gap by deploying a BERT-based semantic similarity module that scores each transcribed answer against a rubric derived from domain knowledge, providing content-grounded evaluation in addition to behavioral signal analysis.

C. Static and Non-Adaptive Question Strategies

Rule-based and template-driven systems like MACH [3] and early dialogue systems [4] are incapable of adapting question difficulty, domain focus, or follow-up specificity in response to candidate performance during the session. This rigidity limits the diagnostic value of the interaction and fails to replicate the adaptive probing that characterizes skilled human interviewers. The proposed system leverages a fine-tuned LLM with retrieval-augmented generation to dynamically adjust question complexity and relevance based on the candidate's prior responses and target domain, enabling genuinely adaptive interview progression.

D. No Structured, Personalized Feedback Delivery

None of the surveyed works provides automated generation of a structured, individualized feedback report as a session output. Hoque et al. [3] offer templated suggestions, while analytical systems [1][2] produce prediction scores without actionable guidance. The proposed system generates a comprehensive post-interview report covering content accuracy scores per question, communication quality metrics, sentiment trajectory, facial expression patterns, strengths identified, and prioritized improvement recommendations, all formatted for immediate candidate review.

E. Limited Robustness to Real-World Acoustic and Environmental Variability

Evaluations in works such as Naim et al. [2] and Hemamou et al. [1] are conducted under controlled laboratory or studio conditions that do not reflect the acoustic variability of home offices, university labs, or mobile environments where candidates realistically interact with preparation tools. The proposed system integrates WebRTC-based noise suppression, voice activity detection, and Whisper v3's multilingual robustness to maintain transcription and analysis accuracy across diverse environmental conditions, an aspect entirely absent from the surveyed literature.

V. SYSTEM DESIGN AND ARCHITECTURE

A. Architectural Overview

The proposed AI Virtual Interviewer follows a modular, layered architecture that integrates a web-based candidate interface, a real-time backend orchestration layer, an AI processing core, a persistent data store, and an automated reporting engine. This design separates concerns to enable independent scaling and maintenance of each functional component. The system is composed of five principal layers: the Web-Based Candidate Interface for session initiation, video capture, and response submission; the FastAPI Backend for real-time request routing and module coordination; the AI Core for question generation, transcription, semantic scoring, and behavioral analysis; the Database Layer for session state, question history, and evaluation records; and the Reporting Engine for automated feedback generation and delivery.

B. Candidate Interface Layer

The frontend is implemented as a React.js single-page application with WebRTC-based video and audio streaming. Upon session initiation, the candidate authenticates using OAuth 2.0 and selects a target domain such as software engineering, data science, or product management. The interface renders AI-generated questions in sequential order, captures the candidate's spoken responses through the browser microphone API, and streams audio chunks to the backend over a persistent WebSocket connection. Facial expression data is sampled at five-second intervals using a TensorFlow.js-based landmark detector running locally in the browser to minimize latency and preserve privacy.

C. Backend Orchestration Layer

The backend is built on FastAPI with Python 3.11 and serves as the central coordination unit of the system. It manages WebSocket sessions for continuous audio streaming, routes transcription requests to the Whisper STT module, dispatches transcribed text to the semantic scoring and sentiment analysis services, and aggregates evaluation scores across the full interview session. FastAPI is selected for its native asynchronous support and high throughput under concurrent candidate sessions.

D. AI Core

The AI processing layer encompasses four sub-modules. The question generation sub-module employs a fine-tuned GPT-4o model augmented with a retrieval mechanism that draws domain-specific context from a curated knowledge base, ensuring questions are technically grounded and dynamically varied. The speech transcription sub-module uses OpenAI Whisper v3 with WebRTC noise suppression applied as a

preprocessing step, achieving robust transcription accuracy across diverse acoustic environments. The semantic evaluation sub-module computes cosine similarity between the BERT embedding of the candidate's transcribed answer and a set of reference answer embeddings associated with the posed question, producing a content quality score on a normalized scale. The behavioral analysis sub-module independently processes textual sentiment using a fine-tuned RoBERTa model and facial sentiment using DeepFace landmark-based classification, producing a communication confidence score that complements the content score.

E. Technology Stack

TABLE II. AI VIRTUAL INTERVIEWER TECHNOLOGY STACK

Layer	Technology	Version	Function
User Interface	React.js / Flutter Web	React 18	Candidate-facing interview portal
Authentication	OAuth 2.0 / JWT	OAuth 2.0	Secure session management
Speech-to-Text	OpenAI Whisper	Whisper v3	Real-time speech transcription
Question Engine	GPT-4o (fine-tuned)	GPT-4o	Dynamic question generation
Answer Evaluation	BERT + Rubric Scorer	BERT-large	Semantic answer quality scoring
Sentiment Analysis	RoBERTa / DeepFace	RoBERTa-base	Facial & textual sentiment
Backend	FastAPI (Python)	Python 3.11	API orchestration
Database	PostgreSQL + Redis	PG 16 / Redis 7	Session storage & caching
Feedback Engine	GPT-4o + custom prompts	GPT-4o	Structured post-interview report
Deployment	Docker + AWS Lambda	Docker 24	Scalable cloud deployment

VI. CONCLUSION

This paper has presented a structured survey of the state of the art in automated interview analysis, interactive coaching agents, spoken dialogue systems, and LLM-based question generation, grounded in five closely related research works, and has identified the persistent gaps that collectively motivate the AI-Powered Virtual Interviewer system.

Multimodal behavioral analysis substantially enriches candidate evaluation beyond text alone. Hemamou et al. [1] and Naim et al. [2] demonstrate that audio-visual cues provide statistically significant predictive information, motivating the proposed system's integrated behavioral signal analysis module.

Interactive virtual agents can measurably improve candidate communication skills. Hoque et al. [3] validate the training efficacy of virtual interview agents, directly supporting the design of the proposed system's live conversational interface.

Semantic content evaluation is the critical missing capability in existing interactive systems. No surveyed interactive system evaluates the factual and conceptual quality of candidate responses, a gap that the proposed BERT-based semantic scoring module is specifically designed to close.

Dynamic, context-aware question generation is achievable through retrieval-augmented LLMs. Guo et al. [5] establish that fine-tuned GPT-class models can generate high-quality, domain-grounded assessment questions, validating the proposed system's question engine design.

Robustness to real-world acoustic variability remains an unaddressed challenge across the surveyed literature. The proposed system addresses this through Whisper v3 with integrated noise suppression and voice activity detection.

In summary, the AI-Powered Virtual Interviewer represents a significant advancement over existing isolated tools, consolidating question generation, real-time speech evaluation, behavioral analysis, and structured feedback delivery into a single, deployable platform. Future enhancements include multilingual interview support, domain-adaptive rubric learning through reinforcement feedback from professional interviewers, integration with job-description-aware question personalization, and on-device inference for privacy-sensitive deployment environments.

VII. ACKNOWLEDGMENT

The authors thank the Department of Computer Engineering at JSPM's Jaywantrao Sawant College of Engineering, Pune, for their support and guidance throughout this research.

VIII. REFERENCES

[1] L. Hemamou, G. Felhi, V. Vandenbussche, J.-C. Martin, and C. Clavel, "HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews," in Proc. AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 661–668, 2019.

[2] 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, Apr.–Jun. 2018, doi: 10.1109/TAFFC.2016.2614299.

[3] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "MACH: My Automated Conversation coacH," in Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), 2013,
pp. 697–706.

[4] L. Chen and K. Jokinen, "Spoken Dialogue Systems for Job Interview Training," in Proc. Workshop on Spoken Dialogue Systems Technology (IWSDS), 2011, pp. 1–10.

[5] Y. Guo, Z. Zhang, and S. Zhao, "Automated Interview Question Generation Using Retrieval-Augmented Large Language Models," arXiv preprint arXiv:2312.04345, 2023.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. Int. Conf. Machine Learning (ICML), vol. 202, 2023, pp. 28492–28518.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[8] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[9] S. Serengil and A. Ozpinar, "DeepFace: A Lightweight Face Recognition and Facial Attribute Analysis Framework for Python," in Proc. Innovations in Intelligent Systems and Applications Conference (ASYU), 2021, pp. 1–4.

[10] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020,pp. 1877–1901.