

A Survey of AI-Powered Multi-Agent Interview Platforms for Candidate Preparation and Intelligent Recruitment

Dr. Riyazahemed A. Jamadar¹, Suyash Zinjurke², Jayesh Nikumbh³, Deepak Shinde⁴

¹riyaz.jamadar@aissms.org ²zinjurke77h@gmail.com ³jayeshnikumbh65@gmail.com

⁴deep1904s@gmail.com

Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune, India

Abstract—The rapid integration of artificial intelligence into human resource management has catalyzed a fundamental transformation of conventional interview methodologies. Traditional recruitment processes are burdened by inherent subjectivity, limited scalability, resource constraints, and inconsistent assessment standards—challenges that automated, data-driven approaches are well-positioned to address. This survey presents a systematic and critical analysis of the emerging landscape of AI-powered interview platforms, focusing on three interconnected technological streams: multimodal behavioral analytics for holistic candidate assessment, large language model (LLM) applications for dynamic and adaptive question generation, and multi-agent architectures for intelligent orchestration of the interview process. By synthesizing findings from contemporary research spanning automated behavioral analysis, LLM fine-tuning for domain-specific interactions, and hierarchical agent coordination, this survey delineates the current state of capabilities, open technical challenges, ethical considerations, and promising future directions in intelligent recruitment systems.

Key Words: AI Interview Systems, Large Language Models, Multi-Agent Systems, Behavioral Analytics, Automated Recruitment, Adaptive Assessment, Multimodal Fusion.

I. INTRODUCTION

The digital transformation of talent acquisition represents one of the most consequential and rapidly evolving applications of artificial intelligence in organizational management. Interviews remain the cornerstone of recruitment, yet the traditional process—conducted entirely by human evaluators—is demonstrably susceptible to cognitive biases, inter-rater inconsistencies, scalability bottlenecks, and resource intensiveness [1]. As organizations seek to evaluate large candidate pools with greater efficiency and objectivity, AI-powered interview systems have emerged as a compelling paradigm capable of augmenting, or in specific contexts partially substituting, human evaluators with intelligent automation.

The evolution of automated interview technology can be traced across three broad generations. The first generation was characterized by rule-based resume screening and keyword-matching algorithms designed to filter candidates against static qualification criteria. The second generation introduced video-based asynchronous interviews with limited speech recognition and elementary prosodic feature extraction. The current, third generation represents a qualitative leap in capability: systems now incorporate sophisticated multimodal signal

processing, context-aware natural language understanding via large language models, and distributed multi-agent architectures that coordinate distinct phases of the interview pipeline with minimal human intervention.

Despite this momentum, the field remains fragmented. Multimodal behavioral analytics, LLM-based interaction management, and multi-agent system design have largely advanced as separate research threads, with limited integration across the three streams. Comprehensive survey treatments synthesizing these complementary perspectives remain scarce. This survey addresses that gap through a structured, cross-disciplinary review. The central contributions are as follows:

- A systematic synthesis of multimodal behavioral assessment techniques and their evolution in interview evaluation contexts.
- A comprehensive examination of LLM applications for interview question generation, response evaluation, and adaptive conversational management.
- A detailed review of multi-agent architectures and their suitability for managing complex, multi-phase interview workflows.

- An analysis of representative conceptual architectures integrating the above streams into end-to-end intelligent recruitment platforms.
- Identification of persistent research gaps, implementation challenges, and promising directions for future inquiry.

II. SURVEY METHODOLOGY

This survey was conducted through a structured review of contemporary research contributions related to AI-powered interview systems, multimodal behavioral analytics, large language models, and multi-agent recruitment architectures. Research papers were collected from widely recognized academic sources including IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and arXiv. The selection process prioritized publications focusing on automated interview assessment, multimodal candidate evaluation, conversational AI, recruitment automation, and intelligent orchestration systems published between 2013 and 2025. Both foundational studies and recent advancements were considered to capture the evolution of the field comprehensively.

The reviewed literature was subsequently categorized into three principal technological streams: (i) multimodal behavioral analytics, (ii) large language model integration, and (iii) multi-agent interview orchestration frameworks. This categorization enabled systematic comparative analysis of existing approaches, architectural patterns, implementation strategies, and emerging research challenges.

III. BACKGROUND AND TECHNOLOGICAL FOUNDATIONS

The design of modern AI interview systems rests on the convergence of several independently mature fields. Understanding these foundational technologies is essential for contextualizing current system capabilities and the open research problems that remain.

A. Multimodal Data Processing in Behavioral Assessment

Human behavior during an interview manifests across multiple simultaneous channels: the verbal content of spoken responses, paralinguistic acoustic cues embedded in speech, and nonverbal visual signals conveyed through facial expression, gaze, head movement, and body posture. Effective automated assessment requires capturing and interpreting all these channels in concert.

Visual analysis employs computer vision pipelines to extract behavioral cues from video recordings. Convolutional neural networks (CNNs) and, more recently, transformer-based vision encoders such as SigLIP2 have demonstrated strong capability in capturing semantic information about facial expression, eye contact patterns, and upper body dynamics [4]. Facial landmark detection underpins downstream features including head pose estimation and geometry-based emotion proxies.

Acoustic processing operates on two complementary levels. At the phonetic level, automatic speech recognition (ASR) converts spoken responses to text for downstream linguistic analysis. At the paralinguistic level, prosodic and spectral features—pitch contour, intensity modulation, speaking rate, voice quality metrics, and Mel-frequency cepstral coefficients (MFCCs)—carry information about emotional state, cognitive load, and communicative fluency [4]. Linguistic analysis further examines textual transcripts to assess vocabulary richness, syntactic complexity, and semantic coherence using transformer-based embedding models such as SFR-Mistral-Embedding, which substantially outperform traditional bag-of-words or TF-IDF representations on downstream tasks.

B. Large Language Models in Conversational Contexts

Large language models (LLMs), grounded in the transformer architecture, have transformed natural language processing by enabling few-shot and zero-shot generalization across an enormous range of tasks. In the interview domain, LLMs offer three primary functional capabilities. Dynamic question generation allows the system to formulate contextually appropriate follow-up questions based on candidate responses, specified interview dimensions, and role requirements. Fine-tuned LLMs substantially outperform template-based systems on both fluency and contextual relevance metrics [3]. Response evaluation extends beyond surface-level content analysis to assess technical accuracy, reasoning depth, and communicative clarity across multi-dimensional rubrics. Conversational management encompasses orchestration of natural interview flow, topic transitions, probing decisions, and acknowledgment generation [3].

C. Multi-Agent System Design Principles

The functional complexity of a comprehensive interview system—spanning question generation, response evaluation, behavioral analysis, and session management—makes monolithic architectures difficult

to maintain and extend. Multi-agent architectures distribute this complexity across specialized agents, each with a focused responsibility and defined communication interface. Key design principles relevant to interview platforms include role specialization, coordination protocols, and adaptive process management, which together motivate the three-agent architecture found at the core of several contemporary interview systems [3].

IV. LITERATURE REVIEW

A. Multimodal Behavioral Analytics for Interview Assessment

The foundational contribution in this stream is the work of Agrawal et al. [2], who proposed a multimodal analytical framework for automated job interview performance assessment using audio, video, and text transcripts from recorded mock interviews. The framework extracted prosodic features (pitch, intensity, MFCCs), facial landmarks and head pose parameters, and lexical features (speaking rate, vocabulary diversity, emotion scores) from the MIT Interview Dataset comprising 138 videos of 69 candidates. Random Forest classifiers were trained independently for each of nine behavioral labels. Results demonstrated high accuracy for objective metrics but more modest performance for subjective constructs such as authenticity, illuminating a fundamental tension between the measurability of behavioral signals and the complexity of underlying psychological constructs.

Building substantially on this foundation, Li et al. [4] introduced the 365 Framework—a comprehensive multimodal assessment approach that systematically explores three modalities (video, audio, and transcribed text), six candidate responses per session, and five evaluation dimensions (integrity, cooperation, social versatility, development orientation, and overall employability). The framework employed state-of-the-art modality-specific feature extractors: SigLIP2 for visual features, Emotion2Vec for acoustic emotion representations, and SFR-Mistral-Embedding for semantic text embeddings. A central architectural contribution is the Multimodal Shared Compression Multilayer Perceptron (MSCMLP), which projects features from all three modalities onto a shared latent manifold through shared basis vectors while retaining modality-specific activation patterns via independent projection layers. An ensemble of 32 independent regression heads further enhanced prediction robustness. This framework secured first place at the AVI Challenge

2025 benchmark, validating the efficacy of the proposed design.

Yi et al. [6] addressed multimodal personality assessment in structured online job interviews through a neural network incorporating prompt learning to handle missing audio modalities, combined with re-sampling methods and margin ranking loss to mitigate label imbalance. Their approach extended evaluation to eight criteria covering the entire interview session. This work underscores the practical challenge of modality dropout in real-world deployments and the importance of architectures that degrade gracefully under incomplete data stream conditions. CNN-based approaches have also been explored for emotion-focused interview evaluation. Manasvi et al. [5] demonstrated effective integration of CNNs for facial emotion recognition combined with recurrent architectures for temporal analysis of speech and body language, with real-time processing capabilities suitable for live interview deployment. Earlier foundational research by Nguyen et al. [10] established that upper body movement and facial cues provide meaningful predictive signal for personality trait ratings, validating the practical relevance of nonverbal channels.

B. Large Language Model Applications in Interview Systems

The seminal contribution in this stream is the interview system proposed by Quan et al. [3], which introduced both a novel dataset construction methodology and an LLM-based multi-agent architecture for end-to-end interview management. The dataset construction pipeline used a ChatGPT-based iterative seed enhancement algorithm to generate approximately 35,000 domain-specific prompt-question pairs across modules covering educational experience, research background, and personality assessment—addressing the critical scarcity of high-quality interview-domain training data. The system was implemented using ChatGLM3 as the base model and evaluated under three fine-tuning regimes: QLoRA, LoRA, and Freeze fine-tuning.

Younes et al. [7] advanced recruitment automation through a fine-tuning methodology designed for structured information extraction from candidate profiles. Their approach combined synthetic datasets with resumes parsed into standardized JSON format by a large-parameter LLM, enabling scalable training data augmentation without manual annotation. Fine-tuned variants substantially outperformed base LLMs on recruitment-specific benchmark metrics. The

standardized JSON output format addresses system integration challenges by providing a consistent schema for downstream applicant tracking and HRIS platforms. A large-scale empirical evaluation by Ansari et al. [8], analyzing data from over 300,000 AI-conducted interviews, revealed that ASR-LLM-TTS component selection decisions have substantial and measurable implications for overall system quality, providing actionable guidance for practitioners selecting technology stacks for deployment.

C. Multi-Agent Architectures for Interview Management

The three-agent framework of Quan et al. [3] remains the most thoroughly documented multi-agent interview architecture in the literature. The Role Agent determines the sequencing of primary interview dimensions and assigns corresponding roles to downstream agents using prompt-engineered instruction sets. The Environment Agent navigates to the appropriate secondary dimension context and manages contextual transitions between topics, deciding when sufficient information has been gathered based on a formal transition criterion. The Memory Agent evaluates whether each candidate response is adequate for progression or whether further probing is warranted through a binary decision function conditioned on the accumulated question-answer history. This architecture enables adaptive interview flows that respond to candidate behavior rather than following a rigid predetermined script.

Multimodal hierarchical attention neural networks for recruitment assessment [9] have demonstrated the utility of hierarchical attention mechanisms for integrating heterogeneous candidate data from resumes, social media profiles, and interview video. The hierarchical attention mechanism dynamically prioritizes and weights different data dimensions, supporting both interpretability and holistic evaluation. Research by Naim et al. [14] and Chen et al. [15] further established the value of combined verbal and nonverbal cues for automated interview scoring, providing empirical grounding for multimodal system design decisions.

V. SYSTEM ARCHITECTURE ANALYSIS

Modern AI interview platforms are distinguished by the sophistication of their architectural design—the structural choices that govern how system components interact, scale, and evolve over time. This section examines

prevalent architectural patterns and their implications for system quality attributes.

A. Multi-Layered Architecture Design

Contemporary systems universally adopt a layered architecture that cleanly separates presentation, application, and data concerns, as illustrated in Fig. 1. This approach supports modular development, independent horizontal scaling of compute-intensive components, and maintainable long-term evolution. The Presentation Layer provides role-differentiated interfaces for candidates and recruiters. Candidate interfaces encompass practice session management, real-time behavioral feedback displays, and performance analytics dashboards. Recruiter interfaces provide candidate review workflows, structured scoring summaries, comparative analytics, and scheduling tools.

The Application Layer serves as the intelligence core of the platform, integrating modules for authentication, AI-driven interview orchestration, behavioral analysis, and notification management. The Data Layer ensures secure storage and management of structured and unstructured interview data, including user profiles, transcripts, assessment scores, and video recordings.

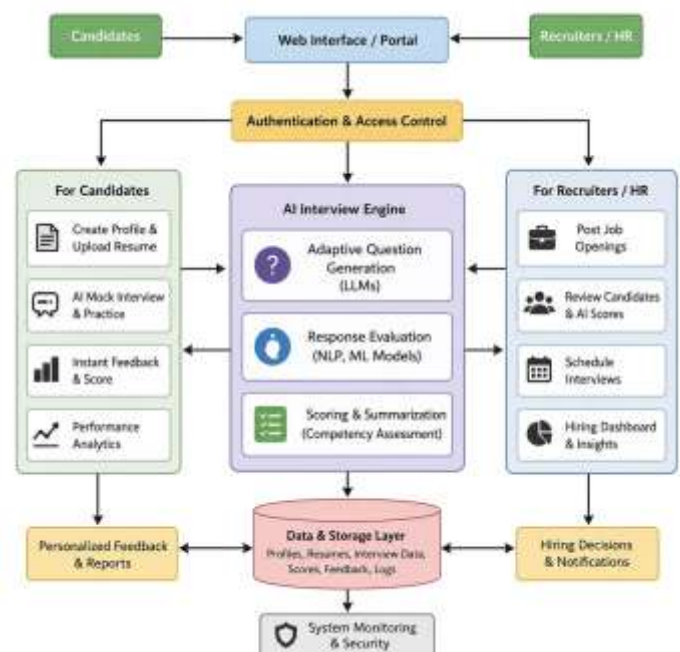


Fig. 1. Conceptual architecture of the AI-powered multi-agent interview platform illustrating candidate and recruiter interaction layers, adaptive interview engine, response evaluation modules, and centralized data management.

B. Dual-Perspective Service Design

A distinctive architectural consideration for platforms serving both candidates and recruiters is the dual-

perspective service design. Candidate-facing services must prioritize responsiveness, developmental feedback delivery, and privacy protection, while recruiter-facing services emphasize assessment reliability, auditability, and integration with existing HR workflow systems. These divergent requirements necessitate carefully designed API boundaries and access control policies that ensure data from one perspective is appropriately scoped before being surfaced to the other.

C. External Service Integration

Production-grade interview platforms extensively leverage external services for specialized capabilities. LLM APIs enable sophisticated question generation and response evaluation without the overhead of on-premise model hosting. Cloud speech processing services provide robust ASR across diverse accents and recording conditions. Scheduling and communication services automate session coordination. This service-oriented approach reduces time-to-deployment and allows rapid adoption of improved underlying models, but introduces dependencies on third-party uptime, API versioning stability, and data governance policies that must be carefully managed.

VI. TECHNOLOGICAL COMPONENTS AND IMPLEMENTATION

A. Convolutional Neural Networks for Facial Behavioral Analysis

Facial analysis pipelines in AI interview systems typically comprise sequential stages: face detection to localize the candidate within each video frame; landmark detection to identify anatomically defined facial keypoints enabling computation of geometric features associated with different emotional states; and CNN-based classification into discrete emotion categories or regression of continuous affect dimensions. Research has demonstrated meaningful accuracy in multi-class emotion recognition from interview video, with performance varying considerably depending on the target label's objective measurability [2], [11]. Normalization procedures accounting for skin tone variation, illumination changes, and head pose variation are critical for generalizing beyond controlled laboratory conditions to realistic deployment settings.

B. Speech and Acoustic Processing

Audio analysis in interview contexts spans two complementary tracks. The first is automatic speech recognition, which transcribes spoken responses for

downstream linguistic analysis; empirical evaluation at scale confirms that different ASR-LLM pairings produce substantially different interview quality metrics [8]. The second track is paralinguistic feature extraction, computing prosodic and spectral descriptors including fundamental frequency trajectories, intensity contours, MFCC vectors, zero-crossing rates, and spectral flux—collectively capturing the candidate's speaking style, emotional regulation, and communicative confidence independently of lexical content. Self-supervised acoustic representations such as Emotion2Vec learn universal emotion-related features from large unlabeled corpora, offering superior generalization compared to handcrafted feature engineering approaches [4].

C. Natural Language Processing and LLM Integration

LLM integration enables sophisticated analysis of candidate response content along multiple dimensions. Transformer-based embeddings encode the semantic content of transcribed responses into dense vector representations capturing meaning invariantly to surface-level paraphrase. Domain adaptation through fine-tuning is essential: base LLMs trained on general-purpose corpora tend to generate questions that lack the specificity, professional register, and dimension-alignment required for structured interviews [3]. The iterative seed enhancement algorithm provides an effective and scalable approach to constructing high-quality fine-tuning data by using an LLM to iteratively expand, critique, and deduplicate a small initial set of domain-specific examples. Prompt engineering plays a complementary role, enabling dynamic control of agent behavior through carefully crafted instruction templates without requiring model retraining.

D. Multimodal Fusion Architectures

Fusing features from heterogeneous modalities is both technically challenging and critically important for assessment quality. Three fusion strategies are prevalent in the literature. Early fusion concatenates features from all modalities before training a unified model—simple but susceptible to modality imbalance and requiring careful normalization. Model-level fusion (mid-fusion) employs modality-specific sub-networks producing intermediate representations subsequently combined via an integration module; the MSCMLP proposed by Li et al. [4] exemplifies this paradigm, where shared basis vectors facilitate cross-modal interaction while independent projection layers preserve modality-specific patterns. Late fusion trains separate modality-specific models and combines predictions through averaging or

learned aggregation, offering robustness to missing modalities but failing to capture cross-modal correlations. Table I summarizes the comparative characteristics of major surveyed system approaches.

TABLE I

Comparative Overview of AI Interview System Approaches

Approach	Key Strengths	Limitations	Best Use Case
Multimodal Behavioral Analytics	Comprehensive, objective, quantifiable assessment signals	High compute cost; requires large labelled datasets	High-stakes roles; leadership evaluation
LLM-Based Questioning	Adaptive, contextual, natural dialogue quality	Bias risk; domain gap; requires fine-tuning	Technical interviews; cultural fit
Multi-Agent Orchestration	Scalable, specialised, workflow-flexible	Coordination overhead; elevated system complexity	Enterprise pipelines; structured multi-phase evaluation
Traditional Rule-Based Automation	Simple, reliable, low deployment cost	Rigid, static; superficial candidate assessment	High-volume initial resume screening

VII. IMPLEMENTATION CHALLENGES AND LIMITATIONS

A. Technical Challenges

Multimodal data integration presents fundamental difficulties arising from differences in data characteristics, sampling rates, temporal alignment requirements, and modality-specific noise profiles. Synchronizing facial video features, acoustic features, and ASR-derived text features—each operating at different temporal granularities—requires careful pipeline engineering. Current fusion approaches also struggle with missing or corrupted data streams, common in real-world deployment where network instability or hardware failure may affect individual modalities independently.

Real-time processing constraints create an inherent tension between assessment thoroughness and system responsiveness. Comprehensive multimodal analysis, particularly when using large pre-trained feature extractors, demands substantial computational resources. For web-based platforms that must deliver immediate feedback, this creates significant engineering challenges around model quantization, batching strategy, and inference caching. Model generalization across diverse cultural contexts, industry domains, role types, and demographic groups remains an open problem. Training datasets are frequently sourced from narrow populations and controlled settings, limiting the ecological validity of deployed models. Publicly available annotated interview datasets remain small relative to the diversity of real-world candidate populations, creating a persistent data bottleneck for generalization research.

B. Ethical and Societal Challenges

Algorithmic bias and fairness constitute perhaps the most pressing concern in AI interview system deployment. Systems trained on historical interview data may encode and perpetuate the preferences of past human interviewers, who are themselves susceptible to demographic bias [12]. Facial analysis systems may perform differently across skin tones; acoustic models may disadvantage non-native speakers. Mitigating these risks requires diverse and representative training data, intersectional fairness auditing, and ongoing post-deployment monitoring.

Transparency and explainability challenges arise from the complexity of modern deep learning systems. When candidates receive unfavorable assessments, they have a reasonable expectation of understanding the basis for that evaluation; however, internal representations of transformer-based models and ensemble fusion architectures are not readily interpretable by non-technical stakeholders. Privacy and data protection requirements are particularly stringent for systems processing biometric data—including facial video and voice recordings—often classified as sensitive personal data under regulations such as GDPR. Furthermore, candidates evaluated by AI systems may experience elevated anxiety or perceive the process as impersonal, potentially undermining organizational employer brand. Research by Hoque et al. [13] underscores that candidates value specific, actionable developmental feedback, a principle that well-designed platforms should treat as a core functional requirement.

VIII. FUTURE RESEARCH DIRECTIONS

A. Personalized and Culturally Adaptive Assessment

Present systems predominantly apply uniform evaluation criteria and normative benchmarks across all candidates. Future work should develop personalized assessment frameworks that establish individual behavioral baselines and adapt evaluation criteria to account for cultural communication norms, linguistic diversity, neurodiversity, and role-specific competency requirements. Such frameworks require richer longitudinal datasets, causal modeling of the relationship between behavioral signals and downstream job performance, and culturally diverse annotator pools for ground-truth label generation.

B. Advanced Multimodal Fusion with Attention Mechanisms

The MSCMLP architecture represents a meaningful step toward principled cross-modal integration, but substantial opportunities remain for more sophisticated fusion approaches. Attention-based fusion mechanisms that dynamically weight modality contributions based on estimated per-instance reliability—downweighting a degraded audio channel in the presence of background noise, for example—could substantially improve robustness under realistic deployment conditions. Hierarchical fusion architectures integrating modalities at multiple temporal scales (frame-level, utterance-level, and session-level) may better capture temporal dynamics of behavioral expression that single-scale approaches miss.

C. Explainable AI for Interview Assessment

Developing interpretable assessment systems capable of articulating the basis for evaluations in terms meaningful to candidates and stakeholders is a critical research priority. Promising directions include attention visualization, concept-based explanations, counterfactual generation, and natural language explanation generation using LLMs grounded in model outputs. Any explanation mechanism must be validated for both faithfulness to the underlying model and comprehensibility to lay audiences with limited AI literacy.

D. Hybrid Human-AI Evaluation Frameworks

Rather than pursuing full automation, a promising direction involves the principled design of hybrid evaluation systems that leverage complementary strengths of human judgment and AI analytics. AI components could handle initial screening, standardized

assessment of objective behavioral metrics, and generation of structured candidate profiles, while human evaluators focus on final-stage holistic judgment, relationship establishment, and assessment of nuanced characteristics. Research is needed to define optimal handoff points and to design human-in-the-loop workflows that preserve efficiency benefits without sacrificing assessment quality or candidate experience.

E. Integration with Talent Lifecycle Management

The scope of AI interview systems need not be limited to the hiring decision. Future platforms could connect interview performance profiles to downstream talent management processes: onboarding program personalization, learning and development recommendations, team composition analytics, and longitudinal career development planning. Such integration would generate the longitudinal outcome data necessary to empirically validate and continuously improve assessment models against actual on-the-job performance measures.

IX. CONCLUSION

This survey has provided a systematic and critical review of the current landscape of AI-powered interview systems, examining the state of the art across three foundational technological streams—multimodal behavioral analytics, large language model applications, and multi-agent architectures—and synthesizing these perspectives within the broader context of intelligent, AI-driven interview ecosystems.

The literature consistently demonstrates that multimodal approaches substantially outperform unimodal baselines in candidate assessment, though the performance advantage is most pronounced for objective behavioral metrics and diminishes for complex psychological constructs. LLM fine-tuning for interview-specific domains produces dramatic improvements in question quality and contextual appropriateness, and multi-agent architectures provide the structural flexibility needed to manage end-to-end interview complexity. When integrated, these three streams offer a credible path toward interview platforms that are simultaneously more efficient, more consistent, and more informative than traditional human-only processes.

The field, however, faces substantial and interconnected challenges. The technical difficulty of real-time multimodal fusion at scale, the persistent problem of

algorithmic bias and limited model generalizability, and the ethical imperatives of transparency, explainability, and data protection collectively require sustained interdisciplinary effort. Progress on these fronts depends on collaboration between AI researchers, HR professionals, organizational psychologists, legal scholars, and the candidates whom these systems are designed to serve. The future of AI-powered interview systems lies not in the wholesale replacement of human judgment, but in the thoughtful augmentation of it—creating evaluation frameworks that harness the precision and consistency of artificial intelligence while preserving the empathy, contextual sensitivity, and ethical accountability that remain distinctively human contributions.

REFERENCES

- [1] A. I. Uffcutt and D. J. Woehr, "Further Analysis of Employment Interview Validity: A Quantitative Evaluation of Interviewer-Related Structuring Methods," *Journal of Organizational Behavior*, vol. 20, no. 4, pp. 549–560, 1999.
- [2] A. Agrawal, R. A. George, S. S. Ravi, S. Kamath, and A. K. M., "Leveraging Multimodal Behavioral Analytics for Automated Job Interview Performance Assessment and Feedback," *arXiv preprint arXiv:2006.07909*, 2020.
- [3] X. Quan, Y. Zhang, X. Xie, and Y. Wang, "An Interview System Based on Large Language Models and Multi-Agent Interactions," in *Proc. 20th IEEE Int. Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2024, pp. 1–8, doi: 10.1109/ICNC-FSKD64080.2024.10702317.
- [4] J. Li, Y. Wang, W. Qian, J. Hu, Z. Hu, R. Hong, and M. Wang, "Listening to the Unspoken: Exploring '365' Aspects of Multimodal Interview Performance Assessment," *arXiv preprint arXiv:2507.22676*, 2025.
- [5] H. Manasvi et al., "A Convolutional Neural Network Approach to Interview Simulation and Evaluation," *Int. Journal of Scientific & Applied Research in Technology*, vol. 11, no. 5, May 2025.
- [6] S. Yi and T. Yamasaki, "Enhancing Online Structured Job Interviews: Personality Assessment Using Multimodal Neural Networks," in *Proc. JSAI Annual Conf.*, 2024.
- [7] M. T. Younes, O. Walid, K. Shaban, A. Hamdi, and M. Hassan, "Augmented Fine-Tuned LLMs for Enhanced Recruitment Automation," *arXiv preprint arXiv:2509.06196*, 2024.
- [8] A. Ansari et al., "Evaluating Speech-to-Text $\times$ LLM $\times$ Text-to-Speech Combinations for AI Interview Systems," *arXiv preprint arXiv:2507.16835*, 2024.
- [9] "Multimodal Hierarchical Attention Neural Network for Recruitment," *Int. Journal of Advanced Research in Science and Technology*, vol. 13, no. 7, 2024.
- [10] L. S. Nguyen, A. Marcos-Ramiro, M. M. Romera, and D. Gatica-Perez, "Multimodal Analysis of Body Communication Cues in Employment Interviews," in *Proc. 15th ACM Int. Conf. on Multimodal Interaction (ICMI)*, 2013, pp. 437–444.
- [11] Z. Y. Huang et al., "A Study on Computer Vision for Facial Emotion Recognition," *Scientific Reports*, vol. 13, article 8953, May 2023.
- [12] A. Vinciarelli and G. Mohammadi, "A Survey of Personality Computing," *IEEE Trans. on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.
- [13] M. E. Hoque, M. Courgeon, J. Martin, B. Mutlu, and R. W. Picard, "MACH: My Automated Conversation Coach," in *Proc. ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp)*, 2016.
- [14] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated Prediction and Analysis of Job Interview Performance: The Role of What You Say and How You Say It," in *Proc. 11th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2015, pp. 1–6.
- [15] L. Chen et al., "Automatic Scoring of Monologue Video Interviews Using Multimodal Cues," in *Proc. INTERSPEECH*, 2016, pp. 32–36.