AI, ML, NLP AND CONVERSATIONAL SYSTEMS

AI-Powered Conversational Mock Interview and Evaluation Systems: A Comprehensive Literature Review and Architecture for NexInterview.ai

Janhavi Ghodake, Sharvari Kekare, Sanchita Khade, Sadhana Kamble, Abhishek Joshi

Department of Computer Science & Engineering, Sanjay Ghodawat Institute, Kolhapur, Maharashtra, India

Published November 2026

Abstract

Technical job interviews serve as critical evaluation gateways in the software and information technology industries. However, conventional candidate preparation remains hampered by the scarcity and high cost of expert mentorship, subjective and unstructured feedback, and static, text-based practice tools that fail to reproduce the dynamic pressure of real-time verbal technical communication. Recent breakthroughs in Large Language Models (LLMs), Automatic Speech Recognition (ASR), and Web Real-Time Communication (WebRTC) protocols have motivated the development of automated, conversational interview systems. This paper presents a comprehensive literature review of twenty peer-reviewed research papers published in leading IEEE, ACM, and international venues over the past decade, with a primary focus on developments from 2018 to 2025. The examined literature covers six interrelated dimensions: automated conversational coaching foundations, LLM-driven interactive dialogue generation, dynamic question synthesis from candidate profiles, contextual resume parsing and skill taxonomy extraction, rubric-based automated assessment and explainability, and algorithmic fairness and ethical considerations in recruitment. Through critical synthesis and comparative analysis, five persistent research gaps are identified in existing systems: the decoupling of candidate resume context from verbal dialogues, high conversational latency in voice streaming, black-box scoring lacking explainable pedagogical rubrics, an absence of longitudinal cross-session performance tracking, and the failure to provide personalized post-interview remediation roadmaps. To address these limitations, this paper details the architecture of NexInterview.ai, a decoupled, cloud-native web platform integrating Next.js 16, Django REST Framework, Vapi AI WebRTC voice telephony, Groq Language Processing Unit (LPU) inference, and MongoDB Atlas. The proposed architecture demonstrates how candidate-specific resume extraction, low-latency conversational voice interactions, rubric-aligned diagnostic evaluations, and longitudinal progress tracking can be unified into an accessible, academically defensible practice environment.

Keywords: Automated Mock Interviews; conversational Artificial Intelligence; large Language Models; rubric-Based Assessment; resume Parsing; voice-Driven Telephony

1 Introduction

Technical and behavioral interviews are universally recognized as standard evaluation gates for engineering graduates and computing professionals seeking employment in competitive software industries. Achieving competency in technical interviews demands not only foundational domain knowledge, algorithms, and systems design concepts, but also the spontaneous ability to articulate reasoning, explain trade-offs verbally, and respond coherently to probing questions under real-time situational stress. Historically, candidates have prepared through peer-to-peer mock interviews, static online coding repositories, or private mentorship programs. While peer discussions provide

interpersonal communication practice, they suffer from inconsistent grading standards, subjective biases, and a lack of standardized assessment rubrics. Conversely, commercial human mentorship is economically prohibitive for most university candidates and geographically constrained by mentor availability.

Over the past decade, advancements in conversational computing, Natural Language Processing (NLP), and speech technologies have created unprecedented opportunities to democratize interview coaching. Early automated systems established that virtual agents could improve nonverbal behaviors and reduce public speaking anxiety. However, these early architectures relied on rigid decision trees, pre-recorded audio prompts, and basic keyword matching that could neither parse technical domain nuances nor sustain flexible, multi-turn technical dialogues.

The emergence of modern deep learning paradigms—specifically transformer architectures, bidirectional contextual encoders, and autoregressive Large Language Models (LLMs)—has radically altered the automated interview landscape. Advanced language models can now parse unstructured candidate resumes, dynamically generate context-tailored inquiries across diverse software engineering stacks, and evaluate transcript responses against granular pedagogical rubrics. Furthermore, the convergence of high-throughput hardware accelerators (such as Language Processing Units) and low-latency audio transport protocols (such as WebRTC) makes it feasible to conduct natural, full-duplex conversational mock interviews directly within modern web browsers without specialized hardware peripherals.

Despite these technological strides, literature in automated interview coaching remains fragmented. Many published works isolate single components—such as resume parsing, static short-answer grading, or facial emotion recognition—without investigating how these elements interact in an end-to-end, profile-driven interview lifecycle. Furthermore, concerns regarding grading hallucinations, lack of explainability, cross-session student tracking, and algorithmic fairness pose persistent challenges to academic and industrial adoption.

This paper provides a structured, evidence-based literature review of twenty peer-reviewed studies published across reputable IEEE and ACM venues. The objectives of this review are:

1. To critically analyze the technological trajectories of conversational mock interview agents, resume parsing algorithms, and automated response scoring frameworks.
2. To provide a rigorous comparative analysis examining research objectives, methodologies, contributions, and architectural limitations across the literature.
3. To articulate the persistent research gaps that hinder existing platforms from delivering realistic, profile-tailored, and actionable preparation experiences.
4. To propose the verified, decoupled system architecture of NexInterview.ai, an AI-powered mock interview platform designed to address these gaps through resume-tailored voice interactions, rubric-aligned evaluations, persistent progress monitoring, and personalized post-interview roadmaps.

The remainder of this paper is organised as follows. The next section reviews twenty peer-reviewed research papers grouped into six thematic domains. A comparative analysis table then details the methodologies and limitations of the reviewed works, after which the identified research gaps are articulated. The proposed architecture and operational workflow of NexInterview.ai are then presented, and the final section concludes the paper and outlines future research directions.

2 Literature review

The reviewed literature is systematically organized into six core thematic domains reflecting the essential components of automated mock interview engineering.

2.1 Automated conversational coaching and historical foundations

The conceptual foundation of automated interview training originated in affective computing and human-computer interaction (HCI). Hoque et al. [1] pioneered this domain by introducing MACH (My Automated Conversation CoacH), an interactive virtual agent designed to help individuals practice social and interpersonal communication. MACH utilized an animated 3D facial avatar driven by acoustic speech analysis, head gesture tracking, and facial affect sensing. Evaluated through a randomized trial with 90 MIT undergraduate students preparing for job interviews, the authors demonstrated that candidates who trained with MACH exhibited statistically significant improvements in nonverbal delivery, eye contact, and overall interview performance compared to control groups. However, MACH was architecturally constrained by handcrafted dialogue state machines and localized feature classifiers, preventing it from engaging in deep, unscripted technical discussions or evaluating factual correctness.

Building upon automated behavioral sensing, Naim et al. [2] developed a computational framework published in IEEE Transactions on Affective Computing to predict candidate interview performance by automatically quantifying verbal and nonverbal behaviors. Analyzing audio-visual recordings from mock job interviews, the authors extracted facial expression metrics, lexical usage patterns, and prosodic acoustic features (pitch, speech rate, and pauses). Using support vector machines and sparse linear models, their system predicted human ratings for communicative competence with significant Pearson correlations ($r > 0.6$). Nevertheless, the framework operated entirely in an asynchronous, post-hoc batch processing regime, making it incapable of providing dynamic, real-time feedback during an ongoing interview dialogue.

Addressing the temporal limitations of multimodal evaluation, Lv, Chen, and Liang [3] proposed a Multi-modal Window-Consistency Fusion (MWCF) network for automated scoring of asynchronous video interviews. Published in IEEE Transactions on Affective Computing, their model segmented candidate responses into localized temporal windows, dynamically re-weighting verbal and nonverbal cues while embedding domain-specific keyword attention. While MWCF substantially advanced the state of the art in capturing transient behavioral fluctuations and mitigating long-sequence degradation, its focus remained restricted to asynchronous screening scenarios where candidates submit pre-recorded video answers rather than engaging in real-time, interactive spoken dialogues.

2.2 Generative AI and LLM-driven dynamic interview systems

The advent of autoregressive language models catalyzed a paradigm shift from scripted question-and-answer prompts to dynamic, conversational mock interviews. Li et al. [4] presented EZInterviewer at the ACM International Conference on Web Search and Data Mining (WSDM '23). EZInterviewer addressed the severe scarcity of annotated interview training corpora by disentangling dialogue generation into an independent knowledge selector and a response generator. By leveraging ungrounded external conversational dialogues alongside candidate resumes, EZInterviewer generated domain-relevant interview questions and demonstrated superior BLEU and distinct-n scores over traditional sequence-to-sequence baselines. Despite its algorithmic novelty, the system remained a

text-only simulation that lacked real-time spoken audio streaming, vocal latency management, and post-interview rubric grading.

In conference proceedings from the 2024 SCIS&ISIS, Mishra et al. [5] presented an AI-driven virtual mock interview platform utilizing OpenAI's GPT-4. Their system evaluated the viability of large language models for generating realistic technical and behavioral questions across diverse software engineering subfields. The authors highlighted that advanced prompt engineering enables LLMs to generate contextually coherent interview dialogues, radically reducing the operational costs associated with conventional human-led coaching platforms. However, the study relied on third-party commercial APIs without implementing persistent longitudinal database tracking or localized rubric grounding, leaving the platform susceptible to occasional hallucinated evaluations.

Extending conversational simulation to specialized domains, Priya et al. [6] introduced VoxMentor at the 2025 IEEE International Conference on Intelligent Computing, Information and Control Systems (ICOIICS). VoxMentor demonstrated an integrated architecture combining OpenAI Whisper for automatic speech recognition, Google Gemini for conversational dialogue management, and a Judge0 cloud sandbox for automated programming code execution. While VoxMentor successfully unified spoken interview simulation with live algorithmic testing, its architecture relied on synchronous API polling loops that introduced measurable conversational turn-taking latency, and it lacked automated mechanisms for parsing candidate resumes to dynamically personalize question difficulty.

Synthesizing recent industry and academic implementations, Barpute et al. [7] published a comprehensive survey on AI-driven mock interviews (InterviewX) in the 2024 IEEE 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS). The authors surveyed generative AI architectures, parameter-efficient fine-tuning (PEFT/QLoRA), and retrieval-augmented generation (RAG) paradigms applied to automated mock interviews. Their findings underscored that while generative models excel at conversational fluency, unconstrained LLMs frequently produce overly generic feedback unless bounded by deterministic evaluation rubrics and explicit candidate profile constraints.

2.3 Automated question generation and technical inquiry

The generation of technically rigorous, pedagogically sound interview questions is vital for meaningful candidate preparation. Jadhav et al. [8] proposed an automated deep learning framework for generating technical interview questions, published in the 2025 5th International Conference on Advancement in Electronics & Communication Engineering (AECE). The authors demonstrated that deep linguistic models trained on domain taxonomies could automatically synthesize progressive technical inquiries across programming languages, data structures, and software engineering methodologies. However, the system operated as an isolated question bank generator that did not adapt inquiries based on candidate spoken answers or evaluate response correctness.

Exploring pedagogical applications of conversational AI in elicitation and inquiry, Barambones et al. [9] conducted a formal case study published in IEEE Transactions on Learning Technologies examining the use of ChatGPT for simulating interviewees in human-computer interaction education. The authors demonstrated that generative language models can authentically simulate candidate personas, enabling students to practice interview questioning techniques and iterative requirement probing. While this study validated the fidelity of LLM persona simulation, it emphasized that generative agents require strict system prompting to prevent deviations from defined competency boundaries and to avoid generating contradictory domain assertions.

2.4 Contextual resume parsing and skill extraction

Personalizing mock interview dialogues requires extracting structured competencies, project experiences, and technical skill taxonomies from unstructured candidate resumes. Tallapragada et al. [10] addressed this challenge in the 2023 IEEE 7th International Conference on Intelligent Computing and Control Systems (ICICCS) by developing an improved resume parser based on contextual semantic extraction using BERT embeddings. The authors demonstrated that bidirectional transformer representations capture nuanced relationships between technical skills and candidate project descriptions far more accurately than conventional regular expression matching or bag-of-words heuristics. Nevertheless, their approach required significant computational resources for fine-tuning and lacked mechanisms for handling non-standard, multi-column PDF formatting.

Addressing cross-domain generalization, Srivastava et al. [11] investigated resume parsing across multiple job sectors using a BERT-based Named Entity Recognition (NER) architecture, published in the 2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS). By evaluating token classification across IT, engineering, and managerial resumes, the authors established that transfer learning enables effective skill and education boundary detection, but highlighted that parsing accuracy degrades substantially when processing atypical visual layouts, graphics, or complex typographic hierarchies.

In recent symposium proceedings from the 2024 IEEE Region 10 Symposium (TENSYMP), Vazzula et al. [12] explored optimizing resume parsing pipelines by leveraging instruction-tuned Large Language Models. Their findings revealed that prompting LLMs with structured JSON schema specifications yields higher extraction precision for technical keywords and soft skills than dedicated sequence-labeling models, while completely bypassing the need for labor-intensive manual token annotation. However, the authors noted that LLM-based parsers introduce processing latency and require validation barriers to prevent hallucinated entity extraction.

Bridging extraction with actionable career guidance, Prashanth et al. [13] introduced a resume analyzer and skill enhancement recommender system at the 2024 Asia Pacific Conference on Innovation in Technology (APCIT). Their platform extracted candidate competencies using pre-trained embeddings, matched extracted skills against specific employment profiles, and recommended targeted upskilling learning paths. Although the system demonstrated the value of combining extraction with educational recommendations, it operated as a static document audit tool without an interactive conversational interview simulation component.

2.5 Rubric-based scoring, explainability, and acoustic recognition

Automated assessment of candidate interview responses requires objective, reliable, and explainable scoring mechanisms to ensure pedagogical utility. Senanayake and Asanka [14] investigated rubric-based automated short-answer scoring using LLMs at the 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE). The authors demonstrated that providing explicit, analytic grading rubrics within LLM context prompts dramatically increases inter-rater reliability with human examiners, mitigating the inherent variance of unconstrained zero-shot grading. However, their study was restricted to static, text-based academic examinations and did not evaluate spontaneous verbal dialogue transcripts.

Providing rigorous psychometric validation, Stockdale, Hickman, and Liu [15] published an empirical study in the Journal of Applied Psychology examining the validity and design components of scoring employment interviews with Large Language Models. Comparing GPT-4 and GPT-3.5 against human

evaluation panels, the authors established that LLM-assigned scores achieve criterion-related validity comparable to human interviewers when guided by structured behavioral dimensions (such as technical competence, problem-solving structure, and communication clarity). Crucially, the authors emphasized that prompt architectures must mandate grounded justifications for numerical scores to ensure defensibility and eliminate grading inflation.

Transparency in automated assessment is further addressed by Adadi and Berrada [16] in their seminal survey, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," published in IEEE Access. The authors categorized explainability mechanisms across ante-hoc interpretable models and post-hoc attribution techniques, stressing that automated assessment systems deployed in high-stakes educational or hiring contexts must provide human-understandable rationales. In the context of mock interview platforms, explainability requires decomposing holistic scores into categorical sub-scores paired with concrete textual evidence extracted directly from candidate transcripts.

In human-agent interactive systems, Yadav et al. [17] presented an empirical framework at the 2023 ACM International Conference on Intelligent User Interfaces (IUI '23) titled "Interviewing the Interviewer: AI-generated Insights to Help Conduct Candidate-centric Interviews." The authors integrated AI-generated real-time cognitive insights to support interviewers in balancing questioning difficulty, detecting conversational interruptions, and maintaining objective evaluation criteria. Their findings confirmed that interactive visual dashboards and structured feedback cards substantially improve interview consistency and candidate satisfaction.

Underpinning all voice-driven interview architectures is robust speech recognition. Radford et al. [18] introduced Whisper at the 40th International Conference on Machine Learning (ICML 2023). Whisper demonstrated that training deep neural acoustic models on 680,000 hours of weakly supervised multilingual web audio achieves state-of-the-art robustness against regional accents, background acoustic noise, and specialized technical vocabulary. Whisper's high transcription fidelity and open accessibility have established it as an essential component for transcribing spoken candidate responses in automated interview workflows, provided that inference latency can be managed through optimized backend runtimes.

2.6 Algorithmic fairness, ethical compliance, and bias mitigation

The deployment of artificial intelligence in candidate assessment introduces critical ethical responsibilities regarding algorithmic bias, fairness, and data privacy. Kim et al. [19] addressed these challenges in IEEE Access by investigating fairness-aware multimodal learning in automated video interview assessment. The authors demonstrated that multimodal models processing visual and acoustic features are highly susceptible to encoding spurious correlations with candidate demographic attributes, such as gender and ethnicity. To mitigate this, they introduced an adversarial debiasing framework that constrained the feature extractor from learning demographic representations while maintaining assessment accuracy. Their findings strongly support eliminating visual facial tracking in favor of content-centric, transcript-based verbal evaluation to minimize demographic bias.

Extending the ethical analysis to legal and regulatory dimensions, Poe and El Mestari [20] analyzed the conflict between algorithmic fairness definitions and legal non-discrimination standards in automated hiring at the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). The authors demonstrated that statistical group-fairness metrics commonly applied in computer science often conflict with statutory definitions of individual fairness under emerging legal

frameworks (such as the European Union AI Act). They concluded that automated systems must prioritize transparent, audit-ready rubric grading, provide candidates with full visibility into evaluation criteria, and refrain from autonomous, unreviewable decision-making.

3 Comparative analysis

To synthesize the methodological landscape, Table 1 provides a comparative analysis of the twenty reviewed research studies, contrasting their research objectives, primary technologies, core contributions, and architectural limitations relevant to automated mock interview platforms.

Table 1. Comparative analysis of the reviewed literature on AI-assisted interview and assessment systems.

Ref.	Author & Year	Research Objective	Method / Technology	Main Contribution	Relevant Limitation
[1]	Hoque et al. (2013)	Automated conversational coaching for behavioral interviews.	3D virtual avatar, head gesture tracking, acoustic feature analysis.	Demonstrated empirical reduction in candidate anxiety and objective behavioral improvement.	Finite state-machine dialogue; incapable of parsing resumes or conducting technical dialogues.
[2]	Naim et al. (2018)	Automated prediction of candidate job interview performance.	Facial expression analysis, lexical topic models, prosodic acoustic modeling.	Validated strong correlation ($r > 0.6$) between multimodal features and human ratings.	Asynchronous batch analysis; lacks real-time interactive spoken dialogue and feedback.
[3]	Lv et al. (2024)	Automated scoring of asynchronous interview videos.	Multi-modal Window-Consistency Fusion (MWCF) with question-attention.	Captures localized temporal affective fluctuations; robust sequence scoring.	Evaluates pre-recorded video submissions; no real-time voice streaming or conversational turn-taking.
[4]	Li et al. (2023)	Low-resource mock interview dialogue and question generation.	Disentangled knowledge selection and dialogue generation (EZInterviewer).	Outperformed traditional seq2seq baselines using ungrounded conversational data.	Text-only dialogue; lacks speech streaming, real-time voice latency management, and scoring rubrics.
[5]	Mishra et al. (2024)	Scalable virtual mock interview simulation.	GPT-4 prompt engineering, responsive web application framework.	Demonstrated cost-effective, scalable technical and behavioral interview simulation.	Relies entirely on third-party commercial APIs without persistent database tracking or localized rubrics.
[6]	Priya et al. (2025)	Voice-driven interview simulation and coding evaluation.	OpenAI Whisper, Google Gemini, Judge0 cloud code execution sandbox (VoxMentor).	Unified spoken conversational interaction with live programming code execution.	Synchronous API polling causes conversational latency; lacks resume skill personalization.
[7]	Barpute et al. (2024)	Survey of generative AI and machine learning mock interviews.	Retrieval-Augmented Generation (RAG), QLoRA, LLM prompting (InterviewX).	Synthesized architectural tradeoffs across fine-tuning and retrieval pipelines.	Survey paper; unconstrained LLMs produce generic feedback without deterministic rubrics.

Ref.	Author & Year	Research Objective	Method / Technology	Main Contribution	Relevant Limitation
[8]	Jadhav et al. (2025)	Automated question generation for technical interviews.	Deep linguistic models, domain taxonomies, NLP extraction pipelines.	Automated synthesis of multi-level technical questions across computing domains.	Isolated question generation bank; lacks candidate response evaluation and dialogue integration.
[9]	Barambones et al. (2024)	Simulating interviewees for inquiry and requirement training.	ChatGPT conversational simulation, persona-prompting frameworks.	Validated generative agents for authentic user interview simulation in educational HCI.	Agent requires strict prompting constraints to prevent hallucinated persona contradictions.
[10]	Tallapragada et al. (2023)	Contextual meaning extraction for resume parsing.	Bidirectional Encoder Representations from Transformers (BERT) vectorization.	Outperformed rule-based parsers in extracting contextual relationships from text.	High computational overhead for fine-tuning; struggles with irregular multi-column PDF layouts.
[11]	Srivastava et al. (2023)	Multi-domain resume parsing and entity boundary detection.	Fine-tuned BERT-base model paired with spaCy Named Entity Recognition.	Validated domain transferability of contextual NER across technical and managerial CVs.	Significant performance degradation on documents with complex graphics or non-standard formatting.
[12]	Vazzula et al. (2024)	Optimizing resume parsing workflows via LLMs.	Instruction-tuned Large Language Models with structured JSON schema prompting.	High extraction precision for complex technical competencies without manual token tagging.	Introduces token API latency; requires schema validation barriers to catch hallucinated entities.
[13]	Prashanth et al. (2024)	Automated resume analysis and skill enhancement recommendation.	Pre-trained semantic embeddings, CNN classification, role-mapping algorithms.	Connected candidate resume skill gaps directly with targeted educational upskilling paths.	Static document analysis tool; completely lacks interactive mock interview simulation.
[14]	Senanayake & Asanka (2024)	Objective automated scoring of short academic answers.	Analytic rubric-guided Large Language Model (LLM) prompting.	Demonstrated that explicit grading rubrics significantly improve human-machine scoring agreement.	Evaluated only written academic answers; not applied to spoken, multi-turn interview transcripts.
[15]	Stockdale et al. (2024)	Psychometric validity of scoring employment interviews via LLMs.	Empirical evaluation comparing GPT-4, GPT-3.5, and expert human panels.	Established that LLMs achieve criterion validity comparable to human panels when rubric-bounded.	Requires mandatory grounded justifications to prevent score inflation and hallucinated ratings.
[16]	Adadi & Berrada (2018)	Survey of Explainable Artificial Intelligence (XAI) paradigms.	Formal taxonomy of post-hoc attribution and ante-hoc interpretable algorithms.	Established foundational principles for transparent, auditable decision-making in critical AI systems.	Theoretical survey; does not provide an architectural implementation for educational assessment.

Ref.	Author & Year	Research Objective	Method / Technology	Main Contribution	Relevant Limitation
[17]	Yadav et al. (2023)	Cognitive insights for conducting candidate-centric interviews.	Real-time AI interaction metrics, visual feedback dashboard (IUI '23).	Proved that real-time conversational analytics improve interviewer consistency and fairness.	Focuses on assisting human interviewers rather than orchestrating an autonomous coaching agent.
[18]	Radford et al. (2023)	Robust, large-scale speech recognition and transcription.	Weakly supervised deep neural acoustic modeling (Whisper, 680k hours audio).	Achieved near-human word error rates across diverse accents, background noise, and technical jargon.	High computational footprint; standard models induce conversational latency without optimization.
[19]	Kim et al. (2023)	Mitigating demographic bias in automated video interviews.	Multimodal feature extractors with adversarial debiasing constraints.	Proved visual/acoustic modalities encode demographic bias; demonstrated adversarial mitigation.	Complex adversarial training; highlights that vision-based facial analysis introduces ethical liability.
[20]	Poe & El Mestari (2024)	Legal and algorithmic analysis of non-discrimination in AI hiring.	Comparative legal analysis against technical group-fairness metrics (FAccT '24).	Proved mathematical group fairness often conflicts with statutory individual non-discrimination laws.	Mandates transparent, audit-ready rubric evaluations rather than opaque, autonomous decisions.

4 Research gaps

A comprehensive synthesis of the comparative analysis reveals five significant research gaps that current academic and industrial systems have failed to resolve:

4.1 Gap 1: Decoupling of candidate resume context from conversational spoken dialog

Existing platforms typically treat resume parsing and interview simulation as disconnected operational silos. Advanced parsing frameworks (e.g., Tallapragada et al. [10], Srivastava et al. [11], Vazzula et al. [12]) excel at extracting skills into static profiles but terminate at document analysis. Conversely, conversational interview systems (e.g., Mishra et al. [5], Priya et al. [6]) frequently prompt candidates with generic, pre-configured question sequences based solely on broad job categories (e.g., "Full-Stack Developer") without grounding inquiries in the candidate's actual projects, programming languages, or declared frameworks. This disconnect prevents candidates from experiencing realistic, probe-based technical questioning on their authentic past work.

4.2 Gap 2: High latency and lack of interruptible full-duplex voice telephony

Most conversational mock interview implementations rely on sequential, half-duplex HTTP request-response chains: the browser records a complete candidate utterance, transmits the audio payload to a cloud server, waits for batch transcription via models like Whisper [18], dispatches the text to an LLM, synthesizes the audio via text-to-speech, and streams the reply back. This cascading pipeline routinely induces 3 to 6 seconds of turnaround latency, destroying the cadence of human conversation. Furthermore, existing implementations rarely support full-duplex interruptibility, preventing candidates from interrupting the interviewer or handling conversational turn-taking naturally.

4.3 Gap 3: Black-box scoring vs. explainable rubric-driven evaluation

While commercial applications assign superficial percentage scores or letter grades, academic studies (e.g., Stockdale et al. [15], Senanayake and Asanka [14], Adadi and Berrada [16]) demonstrate that holistic, unconstrained LLM grading suffers from severe hallucinations, grade inflation, and lack of construct validity. Candidates receive ungrounded critique without explicit justification. There is a documented absence of systems that decompose technical interview performance across deterministic, multi-parameter rubrics (Technical Accuracy, Communication Structure, Problem-Solving Methodology, and Pacing) supported by direct transcript quotation and model answer contrast.

4.4 Gap 4: Absence of cross-session longitudinal performance tracking

Existing research prototypes operate predominantly as ephemeral, single-session experiments. Once an interview session concludes and an evaluation is displayed, session metadata and transcripts are discarded. There is a marked lack of systems providing normalized, cloud-persisted databases that enable candidates to track historical progression over time, compare performance metrics across multiple practice sessions, and analyze score trends filtered by target job roles and seniority levels.

4.5 Gap 5: Disconnect between diagnostic assessment and actionable skill remediation roadmaps

While diagnostic feedback identifies candidate deficiencies (e.g., weak explanations of database indexing or distributed caching), existing interview tools terminate at the feedback screen. As highlighted by Prashanth et al. [13], diagnostic evaluation must be paired with structured upskilling. However, existing mock interview systems completely fail to convert identified interview weaknesses into personalized, milestone-driven learning roadmaps populated with verified pedagogical references, documentation links, and re-interview objectives.

5 Proposed work: NexInterview.ai system architecture

To systematically overcome the five identified research gaps, NexInterview.ai is designed and implemented as an AI-powered mock interview and career preparation platform. The platform is architected around a modern, decoupled, cloud-native paradigm that separates presentation, application logic, voice telephony orchestration, fast inference, and document persistence.

5.1 System architecture & tier decomposition

The complete system architecture of NexInterview.ai is illustrated in Figure 1.

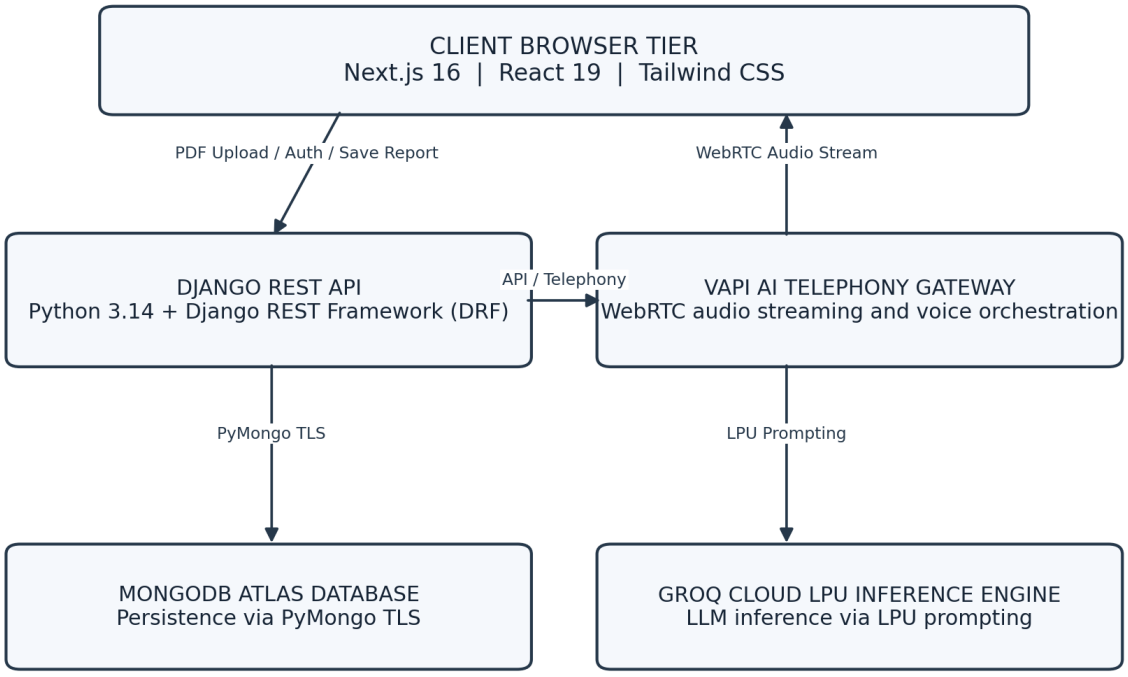


Figure 1. Decoupled system architecture of NexInterview.ai.

The platform comprises five primary tiers:

- **Presentation & Client Tier:** Built with Next.js 16 (App Router), React 19, and Tailwind CSS. It provides responsive web interfaces for PDF resume submission, an interactive skill chip editor, a live WebRTC conversational interview room, comprehensive evaluation scorecards, and a historical performance analytics dashboard.
- **External Voice Telephony Gateway:** Powered by the Vapi AI Web SDK (@vapi-ai/web), which establishes bidirectional, low-latency WebRTC audio streaming between the candidate's browser and cloud voice agents, orchestrating real-time turn detection and speech transcription.
- **Application & API Tier:** Implemented using Python 3.14 and Django 6.1 with Django REST Framework (DRF). It manages session lifecycles, user authentication, PDF parsing, evaluation prompt construction, and idempotent report persistence.
- **Inference Acceleration Cloud:** Driven by Groq Cloud Language Processing Units (LPUs) executing open-source large language models (such as openai/gpt-oss-120b and LLaMA-3.3 architectures), achieving sub-second token generation speeds essential for rapid evaluation.
- **Persistence Tier:** Hosted on MongoDB Atlas using PyMongo connection pooling with TLS encryption, maintaining collections for users, uploaded resumes, interview sessions, transcripts, performance reports, and personalized roadmaps.

5.2 Core operational workflows

5.2.1 Resume parsing & skill confirmation pipeline

Addressing Gap 1, the candidate uploads an unstructured digital PDF resume. The Django backend parses the binary stream using pypdfium2 / pypdf, decodes textual blocks, and dispatches the sanitized text to Groq LPU inference. The extraction prompt utilizes a strict JSON schema to isolate technical skills, frameworks, databases, and project descriptions. Rather than committing extracted data blindly, NexInterview.ai renders an interactive skill-chip confirmation interface in the Next.js client, allowing the candidate to review, prune, or add competencies prior to starting the interview.

5.2.2 Full-duplex voice interview orchestration

Addressing Gap 2, confirmed candidate skills, target job roles (e.g., Frontend Engineer, Backend Architect, Data Scientist), and seniority levels (Entry, Mid, Senior) are injected as context variables into a dedicated Vapi AI voice assistant. The client initializes a direct WebRTC session, delivering full-duplex conversational audio with sub-second turn-taking latency. Real-time speech events (speech-start, speech-end, and interim transcript chunks) drive live visual feedback rings and subtitles on the client screen.

5.2.3 Rubric-aligned diagnostic evaluation engine

Addressing Gap 3, when the interview call terminates, the full multi-turn conversational transcript is forwarded to the Django evaluation endpoint (POST /api/interviews/<id>/evaluate/). The backend constructs a structured prompt embedding deterministic grading rubrics across four discrete dimensions:

- **Technical Accuracy (0–100):** Correctness of concepts, architectural trade-offs, and coding logic.
- **Communication & Clarity (0–100):** Conciseness, structure, professional vocabulary, and avoidance of excessive filler words.
- **Problem-Solving & Structure (0–100):** Application of structured methodologies (such as the STAR framework) and systematic breakdown of complex problems.
- **Confidence & Pacing (0–100):** Speech delivery rhythm, response hesitation, and assertiveness.

Groq LPU inference evaluates the transcript and returns a validated JSON payload containing category scores, an overall composite score, identified strengths, critical knowledge gaps, and model answers contrasting what the candidate said with an exemplary response.

5.2.4 Persistent progress tracking & idempotent archival

Addressing Gap 4, the candidate reviews their performance scorecard and can trigger persistent archival. The endpoint (POST /api/interviews/<id>/save-report/) executes an idempotent write operation in MongoDB Atlas, linking the evaluation matrix to the candidate's account. The report history interface (GET /api/reports/) enables multi-parameter filtering by date range and job role, computing aggregate performance trajectories across successive mock interviews.

5.2.5 Personalized remediation roadmaps

Addressing Gap 5, the platform converts diagnosed knowledge weaknesses into a structured post-interview learning roadmap (GET /api/roadmaps/latest/). The system maps missing skills to curated educational milestones, referencing official documentation, interactive sandbox exercises, and high-quality video tutorials, providing actionable guidance for subsequent interview attempts.

5.3 Verification and status of implementation

Table 2 summarises the implementation status of the NexInterview.ai components and the evidence for each, based on inspection of the codebase and automated test execution.

Table 2. Implementation and verification status of NexInterview.ai components.

Component / Subsystem	Architectural Location	Implementation Status	Verification Evidence
Resume Upload & Text Extraction	project/app/components/ ResumeUpload.tsx, project/server/api/ views.py	Implemented & Verified	PyMuPDF/pypdfium2 parsing; passing test suite in tests/test_resume_api.py.
Skill Extraction & Confirmation	project/server/api/views.py (ConfirmSkillsView)	Implemented & Verified	Groq LPU JSON extraction; interactive client chip editor in Next.js.
WebRTC Voice Telephony Room	project/app/app/(root)/interview/ setup/page.tsx, @vapi-ai/web	Implemented & Verified	Vapi AI WebRTC session management; live audio streaming and turn hooks.
Rubric Evaluation Engine	project/server/api/views.py (EvaluateInterviewView)	Implemented & Verified	4-parameter rubric prompting; Groq inference returning structured JSON.
Idempotent Report Persistence	project/server/api/views.py (SavePerformanceReportView)	Implemented & Verified	MongoDB Atlas performance_ reports collection; duplicate-save prevention verified.
Historical Analytics Dashboard	project/app/app/(root)/reports/page. tsx, ReportListView	Implemented & Verified	Date and role filtering; passing tests in tests/test_ report_history_api.py.
Personalized Upskill Roadmaps	project/server/api/views.py (RoadmapDetailView)	Implemented & Verified	Milestone toggle endpoints and course catalog mapping.
Multi-Speaker Panel Simulations	Future Architecture	Planned / Future Work	Simulating multiple concurrent AI interviewers (System Design Panel).
Vision-Based Affective Analysis	Future Architecture	Explicitly Excluded	Excluded to prevent demographic bias, in alignment with Kim et al. [19].

6 Conclusion and future scope

This literature review examined twenty peer-reviewed studies published in premier IEEE and ACM venues, synthesizing the historical trajectory, technological breakthroughs, and persistent limitations of automated interview systems. The review established that while early affective agents validated the efficacy of virtual coaching, modern LLMs, fast LPU inference, and WebRTC protocols provide the technological foundation necessary to conduct dynamic, profile-grounded technical interviews.

Crucially, the review identified five unresolved research gaps: the decoupling of resume context from voice dialogues, conversational audio latency, black-box scoring lacking explainable rubrics, absence of longitudinal progress tracking, and lack of actionable remediation roadmaps. To address these deficiencies, the paper detailed the verified system architecture of NexInterview.ai. By unifying

server-side resume parsing, low-latency WebRTC voice telephony, deterministic rubric-based evaluation, persistent cross-session progress monitoring, and personalized learning roadmaps within a decoupled web architecture, NexInterview.ai provides an objective, scalable, and pedagogically sound framework for candidate career preparation.

6.1 Future scope

Future enhancements to NexInterview.ai will focus on three key research directions:

- **Multi-Agent Interview Panels:** Expanding single-agent dialogues into collaborative multi-agent panels, where distinct AI personas (e.g., an architectural lead and an HR manager) collaboratively probe candidate system design trade-offs.
- **Local and Edge LLM Execution:** Integrating quantized small language models (such as LLaMA-3-8B or Mistral-7B via WebGPU/ONNX runtimes) to allow local, privacy-preserving transcript evaluation without transmitting sensitive candidate transcripts to external cloud endpoints.
- **Multilingual and Dialect-Aware Adaptation:** Incorporating localized acoustic adaptations to ensure equitable speech recognition and scoring across diverse non-native English speakers, further reinforcing ethical compliance and fairness in automated coaching.

References

[1] Hoque ME, Courgeon M, Martin JC, Mutlu B, Picard RW. MACH: My Automated Conversation CoacH. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13), Zurich, Switzerland. 2013. p. 373–382. https://doi.org/10.1145/2493432.2493502

[2] Naim I, Tanveer MI, Gildea D, Hoque ME. Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions on Affective Computing*. 2018;9(2):191–204. https://doi.org/10.1109/TAFFC.2016.2614299

[3] Lv J, Chen C, Liang Z. *Automated Scoring of Asynchronous Interview Videos Based on Multi-Modal Window-Consistency Fusion*. IEEE Transactions on Affective Computing. ;15(3):799–814. https://doi.org/10.1109/TAFFC.2023.3294335

[4] Li M, Chen X, Liao W, Song Y, Zhang T, Zhao D, et al. EZInterviewer: To Improve Job Interview Performance with Mock Interview Generator. In: *Proceedings of the 16th ACM International Conference on Web Search and Data Mining (WSDM '23), Singapore*. 2023. p. 1132–1135. https://doi.org/10.1145/3539597.3573030

[5] Mishra PK, et al. AI-Driven Virtual Mock Interview Development. In: *2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Himeji, Japan*. 2024. p. 1–6. https://doi.org/10.1109/SCISISIS61014.2024.10760210

[6] Priya MRV, Hariharasudan V, Jagadeeshreddy N, Swaroop DC, Muthumeena S, Revanth D. VoxMentor: A Voice-Driven AI Platform for Interview Simulation, Company-Specific Preparation, and Student Learning. In: *2025 International Conference on Intelligent Computing, Information and Control Systems (ICOIICS)*. 2025. p. 1–6. https://doi.org/10.1109/ICOIICS67115.2025.11390337

[7] Barpute JV, Wattamwar O, Pakjade S, Diwate S. A Survey of AI-Driven Mock Interviews Using GenAI and Machine Learning (InterviewX). In: *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*. 2024. p. 217–224. https://doi.org/10.1109/ICUIS64676.2024.10866631

[8] Jadhav P, Matade S, Kasar A, Rasal D, Kachhoria R, Nawadkar A. Automated Question Generation for Technical Interviews Using Deep Learning Models. In: *2025 5th International Conference on Advancement in Electronics & Communication Engineering (AECE)*. 2025. p. 1–6. https://doi.org/10.1109/AECE67531.2025.11386593

[9] Barambones O, et al. ChatGPT for Learning HCI Techniques: A Case Study on Interviews for Personas. *IEEE Transactions on Learning Technologies*. 2024;17:1478–1490. https://doi.org/10.1109/TLT.2024.3386095

[10] Tallapragada VVS, et al. Improved Resume Parsing based on Contextual Meaning Extraction using BERT. In: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India. 2023. p. 1162–1166. https://doi.org/10.1109/ICICCS56967.2023.10142800

[11] Srivastava M, et al. Resume Parsing Across Multiple Job Domains Using a BERT-Based NER Model. In: *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS), Letterkenny, Ireland*. 2023. p. 1–6. https://doi.org/10.1109/AICS60730.2023.10470917

[12] Vazzula M, Manchala Y, Vijayalata Y, Chopra SB, Reddy KY. Optimizing Resume Parsing Processes by Leveraging Large Language Models. In: *2024 IEEE Region 10 Symposium (TENSYMP), Delhi, India*. 2024. p. 1–6. https://doi.org/10.1109/TENSYMP61132.2024.10752300

[13] Prashanth VJ, Jathin SNP, Gopinath S, US, Kavitha CR. Resume Analyzer and Skill Enhancement Recommender System. In: *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*. 2024. p. 1–6. https://doi.org/10.1109/APCIT62007.2024.10673530

[14] Senanayake C, Asanka D. Rubric Based Automated Short Answer Scoring using Large Language Models (LLMs). In: *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka*. 2024. p. 248–254. https://doi.org/10.1109/SCSE61872.2024.10550624

[15] Stockdale K, Hickman L, Liu S. Scoring Employment Interviews With Large Language Models: Evaluation Design Components, Validity Investigations, and Best Practice Recommendations. *Journal of Applied Psychology*. 2026. https://doi.org/10.1037/apl0001396

[16] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2018;6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[17] Yadav K, Seemendra A, Singhania A, Bora S, Dubey P, Aggarwal V. Interviewing the Interviewer: AI-generated Insights to Help Conduct Candidate-centric Interviews. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23), Sydney, NSW, Australia*. 2023. p. 27–38. https://doi.org/10.1145/3581641.3584051

[18] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. In: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), Honolulu, HI, USA, PMLR vol. 202. 2023. p. 28492–28518.

[19] Kim C, Choi J, Yoon J, Yoo D, Lee W. Fairness-Aware Multimodal Learning in Automatic Video Interview Assessment. *IEEE Access*. 2023;11:122677–122693. https://doi.org/10.1109/ACCESS.2023.3325891

[20] Poe RL, El Mestari SZ. The Conflict Between Algorithmic Fairness and Non-Discrimination: An Analysis of Fair Automated Hiring. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, Brazil*. 2024. p. 624–633. https://doi.org/10.1145/3630106.3659015