IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Prepwise AI: A Self-Adaptive Mock Interview Platform using Real-Time Vocal Intelligence

Aadesh Kapadnis[1], Shivanjali Mulik[2], Vaibhav Mokale[3], Kaveri Borse[4], Prof. S.G.Bagul[5]
[1,2,3,4]Students, Department of Electronics & TC
[5]Guide, Department of Electronics & TC
KCT's Late G. N. Sapkal College of Engineering, Nashik, India

Abstract: *Traditional mock interview tools are often static—they either provide fixed text-based questions or record videos for delayed human review. This creates a "readiness gap" where engineering students lack real-time, objective practice that mirrors the adaptive, conversational dynamics of live technical interviews. This paper presents Prepwise AI, an autonomous interview simulation framework that embeds a large language model (LLM) agent natively within a low-latency voice pipeline to deliver a fully interactive, real-time mock interview experience. The system employs a three-tier decoupled architecture built on Next.js 15, integrating Google Gemini 1.5 Flash in a dual role: Pre-Interview Context Synthesis before session commencement and an offline post-session Intelligence Audit upon completion. The Vapi AI Voice Engine manages a WebRTC-based conversational loop at sub-600 ms AI response latency, with persona and turn-taking logic encoded natively within the Vapi assistant configuration, maintaining a clean separation from the analytical layer. A Zod schema enforces output integrity across three scoring dimensions: Technical Accuracy, Communication Style, and a Confidence Score derived from LLM-driven sentiment analysis. Pilot evaluation demonstrates a 92% assessment accuracy rate and a 95% pipeline delivery ratio, providing a strong empirical baseline for the platform's reliability as an AI-mediated educational assessment tool.*

Keywords: Mock Interview, Large Language Models, Voice Interface, WebRTC, Generative AI, Real-Time Assessment, Educational Technology, Human-Computer Interaction

I. INTRODUCTION

The technical interview remains one of the most consequential and anxiety-inducing performance events in an engineering student's academic career. Despite its importance, the current landscape of interview preparation tools is marked by a fundamental inadequacy: they are predominantly passive. Students practice with static question banks, review curated model answers in isolation, or submit asynchronous video recordings for delayed human review. None of these modalities faithfully simulate the adaptive, turn-taking, and evaluative pressures of a live technical interview. This disconnect between how students prepare and the conditions under which they are ultimately assessed constitutes what this paper terms the "readiness gap"—a structural deficit in the quality and immediacy of practice feedback available to engineering learners.

Recent advances in large language model (LLM) capabilities—particularly instruction-following models capable of sustained persona management and multi-turn dialogue—alongside the maturation of low-latency voice synthesis APIs, have created the technical preconditions for a new class of AI-mediated interview simulation systems. These systems can operate as genuine conversational agents: listening, responding, and evaluating in real time, rather than functioning as passive question dispensers [1], [2].

Prior work on automated interview coaching has largely been confined to post-hoc analysis of recorded responses [3] or static natural language feedback on typed answers [4]. No existing system integrates an LLM agent natively within a real-time WebRTC voice pipeline with concurrent structured scoring, persona management, and cloud persistence. This paper addresses that gap.

DOI: 10.48175/IJARSCT-34048

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

This paper makes the following primary contributions:

A novel three-tier decoupled voice-LLM architecture enabling sub-600 ms AI response latency without compromising evaluation depth.

A dual-role LLM deployment pattern wherein a single Gemini 1.5 Flash instance performs Pre-Interview Context Synthesis and offline post-session Intelligence Audit.

A clean architectural separation between the Vapi-managed persona layer and the Gemini-powered analytical layer, establishing a replicable pattern for voice-LLM educational systems.

A Zod schema-validated scoring framework operationalising a three-axis performance rubric across Technical Accuracy, Communication Style, and Confidence Score.

Pilot empirical evidence: 92% assessment accuracy and 95% pipeline delivery ratio.

II. LITERATURE SURVEY

Automated interview preparation systems are broadly classifiable into three generations: static question-bank tools, video-analysis systems, and conversational AI platforms [5]. First-generation tools such as Google Interview Warmup and Pramp present fixed question sets with no adaptation to the candidate's responses, failing to replicate the evaluative dynamics of live interviews. Second-generation systems, including HireVue and similar asynchronous video interview platforms, capture recorded responses for human or algorithmic review [3], but impose a temporal gap between practice and feedback that limits learning efficiency.

The application of natural language processing to interview coaching has accelerated with the availability of large pre-trained language models. GPT-4-based tutoring systems [6] demonstrate the viability of LLMs as sustained conversational partners in educational contexts; however, these implementations operate in text-only modalities and do not address the vocal delivery dimension of interview performance. Speech-based assessment systems developed within the Computer-Assisted Language Learning (CALL) tradition [7] offer precedents for prosodic and fluency analysis, but evaluate discrete utterances rather than sustained multi-turn technical discourse.

Voice-based human-computer interaction research establishes latency as a first-order usability variable. Schlangen [8] demonstrates that conversational turn-taking degrades perceptibly at response delays exceeding approximately 700 ms, with subjective naturalness declining sharply above 1000 ms. WebRTC-based voice agent architectures that bypass the sequential HTTP delays of conventional Speech-to-Text–LLM–Text-to-Speech pipelines are therefore a prerequisite for naturalistic voice interview simulation.

Schema-constrained generation from LLMs remains an active engineering challenge. Without explicit output constraints, LLM responses exhibit substantial variability in JSON structure and type correctness [9], producing downstream data integrity failures when outputs are consumed by application logic or persisted to databases. Zod-based validation in TypeScript ecosystems has emerged as a standard mitigation strategy. Prepwise AI applies this pattern to the evaluation scoring pipeline, embedding the Zod schema in the Gemini prompt as a structural constraint and applying it as a runtime validator before Firebase persistence.

III. PROPOSED METHODOLOGY

A. System Architecture

Prepwise AI is implemented as a three-tier decoupled architecture, illustrated conceptually in Fig. 1. The decoupling of concerns across the Intelligence Layer, Voice Orchestration Layer, and Scoring Function ensures that latency-sensitive voice operations are never blocked by compute-intensive LLM analysis. The client-facing application is built on Next.js 15, which provides both the React-based frontend and the server-side execution environment for LLM communication via Server Actions. Table I summarises the three tiers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

THREE-TIER ARCHITECTURE SUMMARY TABLE I.

Dimension	Description	Measurement
Technical Accuracy	Correctness & depth of engineering responses	Gemini contextual reasoning
Communication Style	Filler words, pacing, clarity	LLM lexical transcript analysis
Confidence Score	Sentiment polarity & logical flow	Gemini sentiment instruction

B. Intelligence Layer — Dual-Role LLM

The intelligence layer employs Google Gemini 1.5 Flash in two operationally distinct roles, separated by session lifecycle phase.

Phase 1 — Pre-Interview Context Synthesis: Upon session initiation, the user selects a subject domain (e.g., Microcontrollers, Data Structures, Operating Systems). A structured prompt is issued to Gemini 1.5 Flash via the /api/vapi/generate endpoint, instructing the model to select five technically rigorous questions from the pre-loaded corpus defined in etcSubjects.ts. This corpus contains over 100 domain-curated questions across Electronics & Telecommunication (E&TC) and Software Engineering subjects, serving both as the Gemini sampling pool and as a deterministic fallback safety net if LLM generation fails or exceeds a defined timeout. The five synthesised questions are injected into the Vapi assistant's system configuration prior to WebRTC call establishment.

Phase 2 — Offline Post-Session Intelligence Audit: Upon call termination, the complete raw session transcript—captured via the call-end event hook in Agent.tsx—is transmitted to Gemini through a Next.js Server Action. This post-hoc processing model is a deliberate architectural choice: deferring all deep analysis to session completion ensures that no LLM-induced latency enters the live voice interaction loop. The end-to-end turnaround from call termination to structured report delivery is optimised to under 15 seconds.

C. Voice Orchestration Layer

The Vapi AI Voice Engine establishes a persistent WebRTC connection between the browser client and Vapi infrastructure. A critical architectural separation is maintained: the interviewer persona, turn-taking rules, and voice configuration (OpenAI Alloy voice profile) are encoded as constants within the Vapi assistant configuration in index.ts. The Gemini model is not involved in real-time conversational decision-making; it operates exclusively in the deferred analytical role described above. The conversational loop per turn operates through three sequential phases:

LISTEN: Vapi captures raw audio and converts it to text via the integrated STT pipeline, buffering the transcript in session context.

RESPOND: The Vapi assistant streams an audio response; the AI generation leg targets sub-600 ms latency below the perceptual threshold for natural turn-taking [8], enabled by Vapi's WebRTC-native architecture.

EVALUATE: Upon session completion, the callEnd event is intercepted in Agent.tsx via the Vapi SDK. The complete raw transcript is forwarded to the Gemini scoring pipeline.

D. Scoring Function — Schema-Validated Three-Axis Evaluation

The evaluation pipeline operationalises interview performance across three dimensions as detailed in Table II. All sentiment analysis is performed natively by Gemini as specific instructions within the scoring prompt—no separate sentiment model is invoked. The Zod schema serves two complementary roles: it is embedded in the prompt as a structural generation constraint, and applied as a runtime validator on the returned JSON object, ensuring malformed LLM outputs are intercepted before reaching the Firebase persistence layer [9].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

THREE-AXIS SCORING RUBRIC TABLE II.

Dimension	Description	Measurement
Technical Accuracy	Correctness & depth of engineering responses	Gemini contextual reasoning
Communication Style	Filler words, pacing, clarity	LLM lexical transcript analysis
Confidence Score	Sentiment polarity & logical flow	Gemini sentiment instruction

E. Data Persistence

Session states and evaluation reports are persisted to Google Firebase via the Firebase Admin SDK (admin.ts). The Admin SDK operates exclusively server-side within the Next.js Server Action boundary, ensuring database credentials are never exposed to the client. Both raw transcripts and Zod-validated reports are stored, enabling longitudinal analysis of user progress across sessions.

IV. IMPLEMENTATION

A. Pre-Interview Context Synthesis Pipeline

When a user selects a subject domain, a POST request is issued to the /api/vapi/generate route. This route constructs a Gemini prompt including the domain identifier and the complete etcSubjects.ts question corpus for that domain, instructing the model to select exactly five questions of high technical complexity. The synthesised questions are parsed and injected as the questions field of the Vapi assistant's firstMessage and systemPrompt configuration object. If generation fails or times out, a deterministic random selection from the corpus is used as a fallback.

B. Live Voice Session — Agent.tsx

The Vapi SDK is initialised client-side within the Agent.tsx React component. The component registers event listeners for message, call-start, and call-end events. Upon call-end, it captures the accumulated transcript from the Vapi session state and invokes the Next.js Server Action, passing the transcript as a serialised string. The component manages UI state transitions between idle, active call, and evaluation-loading phases, providing appropriate user feedback during post-session analysis.

C. Post-Session Intelligence Audit Pipeline

The Server Action performs five sequential operations: (1) constructs the Gemini evaluation prompt, embedding both the transcript and the Zod schema shape as a JSON structure definition; (2) issues the Gemini API request with a token budget sufficient for a complete structured response; (3) parses the returned content, stripping Markdown code fence delimiters; (4) validates the parsed object against the Zod schema; and (5) on successful validation, writes both the raw transcript and the structured report to Firebase under the authenticated user's session record.

D. Static Pre-Configuration and Corpus Design

The etcSubjects.ts corpus contains over 100 questions authored to require multi-sentence, conceptually substantive responses—ensuring that the Gemini evaluation model has sufficient transcript material to score meaningfully across all three rubric axes. The corpus is organised as a typed TypeScript object keyed by subject domain, enabling type-safe domain filtering at the API route level.

V. RESULTS AND DISCUSSION

A. Experimental Setup

A pilot evaluation of the Prepwise AI platform was conducted to establish baseline empirical metrics for operational reliability and assessment quality. The evaluation focused on two primary performance indicators: the accuracy of

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

LLM-driven assessment against domain-expert review, and the end-to-end pipeline delivery reliability across conducted sessions. Table III presents the quantitative outcomes.

PILOT EVALUATION PERFORMANCE METRICS TABLE III.

Metric	Value	Notes
Assessment Accuracy	92%	vs. domain-expert review
Pipeline Delivery Ratio	95%	Full session pipeline completion
AI Response Latency	<600ms	Generation leg via Vapi
Report Turnaround	<15s	Call end to Firebase write

B. Assessment Accuracy

The 92% assessment accuracy figure was derived by having domain reviewers independently evaluate a sample of interview responses and comparing their ratings to the Gemini Intelligence Audit scores. LLM-generated scores aligned with domain-expert judgement in 92% of cases on the Technical Accuracy dimension. Minor divergences were observed on edge-case responses employing unconventional but technically valid problem-solving framings—a known limitation of rubric-based LLM evaluation that is addressed in the future work.

C. Pipeline Delivery Ratio

The 95% Pipeline Delivery Ratio (PDR) reflects the proportion of completed sessions in which the full post-session pipeline—transcript capture, Gemini Intelligence Audit, Zod validation, and Firebase persistence—executed successfully end-to-end. The 5% failure rate was exclusively attributable to network timeout conditions during long-transcript Gemini API calls, an edge case addressed by the adaptive token-budget enhancement proposed in the future work.

D. Discussion

The results confirm that the three-tier decoupled architecture successfully isolates voice-layer latency from analytical processing latency. The Vapi-managed persona layer's independence from the Gemini analytical layer is the primary enabler of the sub-600 ms response generation figure: because no LLM inference occurs during the live voice loop, the only latency contributor is the Vapi STT–TTS pipeline itself. The 92% assessment accuracy, while preliminary, compares favourably with automated rubric scoring systems reported in the CALL literature [7], and the Zod schema validation mechanism provides a deterministic integrity guarantee that prior LLM-based educational assessment systems have not addressed.

VI. CONCLUSION AND FUTURE WORK

This paper presented Prepwise AI, a real-time voice-driven mock interview platform integrating Google Gemini 1.5 Flash within a WebRTC-based voice pipeline to deliver adaptive, low-latency, and multi-dimensional interview practice to engineering students. The three-tier decoupled architecture addresses the principal engineering challenges of voice-LLM integration: maintaining conversational latency below the perceptual threshold for natural turn-taking through a Vapi-native persona management approach; enforcing structured LLM output via Zod schema validation to guarantee data integrity; and securing sensitive transcript data through exclusive server-side processing via Next.js Server Actions. Pilot evaluation returns a 92% assessment accuracy rate and a 95% pipeline delivery ratio, providing a quantitative empirical foundation for further development.

Several directions remain open for future work. First, the current architecture generates all five interview questions upfront before the session begins. Future iterations should implement turn-adaptive question synthesis, wherein each subsequent question is generated dynamically conditioned on the user's previous answer, enabling Socratic probing and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

personalised difficulty calibration. Second, the scoring pipeline operates exclusively on the text transcript; future work should integrate direct acoustic analysis of the WebRTC audio stream to capture prosodic delivery features—speech rate, pause duration, pitch variance—lost in the STT transcription step. Third, a full-scale controlled user study with pre/post confidence self-reports and System Usability Scale (SUS) evaluation is required to establish rigorous user-facing efficacy evidence beyond the current pilot. Finally, the question corpus and evaluation rubrics should be extended to cover behavioural and business interview formats, broadening the platform's applicability beyond engineering students.

ACKNOWLEDGMENT

The authors would like to thank the Department of Electronics & Telecommunication, KCT's Late G. N. Sapkal College of Engineering, Nashik, for providing the infrastructure and academic support that enabled this work. [Add any additional acknowledgements for funding, datasets, or collaborators here.]

REFERENCES

[1]I. Lazar Mares, et al., "Automated interview coaching systems: A survey," J. Educ. Technol. Soc., vol. 25, no. 3, pp. 45–58, 2022.
[2]D. Nye et al., "Generative tutors: Building LLM-powered pedagogical agents," in Proc. AIED 2023, pp. 1–12.
[3]B. Naim et al., "Automated analysis and prediction of job interview performance," IEEE Trans. Affective Comput., vol. 9, no. 2, pp. 191–204, Apr.–Jun. 2018.
[4]S. Muralidhar et al., "Learning to interview: AI-powered feedback on structured interview responses," in Proc. ACL 2022 Workshop NLP EdTech, pp. 78–87.
[5]J. McCarthy et al., "The use of technology in the interview process: A review," Int. J. Select. Assess., vol. 25, no. 4, pp. 336–352, 2017.
[6]J. Achiam et al., "GPT-4 technical report," arXiv:2303.08774, 2023.
[7]S. Crossley and D. McNamara, "Predicting second language writing proficiency," J. Res. Reading, vol. 39, no. 4, pp. 510–527, 2016.
[8]D. Schlangen, "A general, abstract model of incremental dialogue processing," in Proc. EACL 2009, pp. 643–651.
[9]Zod Contributors, "Zod: TypeScript-first schema validation," 2024. [Online]. Available: https://zod.dev
[10]Google DeepMind, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," Tech. Rep., 2024.
[11]Vapi AI, "Vapi Voice AI Platform Documentation," 2024. [Online]. Available: https://docs.vapi.ai
[12]Vercel, "Next.js 15 Documentation: Server Actions and Mutations," 2024. [Online]. Available: https://nextjs.org/docs
[13]Google Firebase, "Firebase Admin SDK," 2024. [Online]. Available: https://firebase.google.com/docs/admin