

Intervivy: A Multimodal Voice AI Platform for Intelligent, Accessible, and Ethical Remote Recruitment

Gunjal Piyush Nivrutti, Hodgar Sharwa Dattatray, Sonawane Tanmay Bhupendra, Prof. P. S. Hase

Department of Information Technology, Amrutvahini College of Engineering, Sangamner, India

Under the Guidance of Prof. P. S. Hase

Abstract - The rapid normalization of remote hiring has exposed critical deficiencies in contemporary AI-based interview platforms: susceptibility to network instability, insufficient malpractice detection granularity, and evaluations that reduce candidate performance to binary outcomes. This paper presents Intervivy, an intelligent voice AI platform designed to address these systemic shortcomings through a three-pillar architecture centered on accessibility, integrity, and insight. The system employs WebRTC-based low-bandwidth video streaming (5–10 FPS) and Supabase for persistent session management, achieving a 95% session recovery rate in adverse connectivity environments. A multimodal Malpractice Detection System (MDS), integrating YOLOv8-Nano for real-time object recognition and MediaPipe Face Mesh for gaze and head-pose estimation, applies a Heuristic Fusion Engine to aggregate detection signals into a single Integrity Index, targeting over 90% detection accuracy with fewer than 5% false positives. Candidate responses are transcribed via OpenAI Whisper ASR (96.8% word-level accuracy) and evaluated by Google Gemini using rubric-constrained prompt engineering, producing skill-disaggregated reports across technical knowledge, logical reasoning, and communication clarity. System testing across simulated interview environments yields 94.5% malpractice detection precision, 180 ms average interaction latency, and 99.4% uptime. Intervivy constitutes a reproducible, ethically grounded, and globally scalable architecture for AI-assisted recruitment.

Index Terms - artificial intelligence, automated interview, malpractice detection, YOLOv8-Nano, MediaPipe Face Mesh, Whisper ASR, Google Gemini, remote recruitment, multimodal analysis, WebRTC, heuristic fusion.

I. INTRODUCTION

The globalization of labor markets and the widespread normalization of distributed work arrangements have fundamentally restructured how organizations conduct candidate assessment. Virtual interview platforms emerged as an operational response to this shift; however, their underlying architectures frequently remain ill-equipped for the diverse and often resource-constrained environments in which they must function. Most existing systems presuppose stable, high-bandwidth internet connectivity, effectively excluding candidates in rural regions where mobile internet at 3G speeds remains the prevailing standard [9]. This infrastructure bias represents only one dimension of a broader set of structural deficiencies that compromise both fairness and utility.

Beyond connectivity, two additional failure modes dominate the literature on AI-driven interview systems. First, proctoring mechanisms that rely solely on keystroke monitoring or browser-tab detection are unable to account for a taxonomy of deceptive behaviors including impersonation, consultation of off-screen materials, or the use of secondary devices such as mobile phones or tablets [5]. Second, systems that produce binary "hireable/not-hireable" verdicts forfeit the diagnostic richness

necessary for nuanced talent decisions, stripping recruiters of competency-level signal and reducing complex human performance to a single bit of information [2].

Intervivy is proposed as a unified platform that addresses all three failure modes simultaneously. It adopts a three-pillar design philosophy: accessibility through low-bandwidth optimization and automatic session resumption; integrity through a multimodal, real-time Malpractice Detection System; and insight through a structured, rubric-driven large language model (LLM) evaluation pipeline. The platform integrates WebRTC, Supabase, OpenAI Whisper ASR, Google Gemini, YOLOv8-Nano, and MediaPipe Face Mesh into a cohesive, modular architecture deployable on major cloud infrastructure. The principal contributions of this work are as follows.

- (1) A resilient client architecture that sustains session continuity across network interruptions through persistent state management via Supabase, validated to recover within 3.8 seconds on average.
- (2) A multimodal Heuristic Fusion Engine for malpractice detection that combines gaze tracking, facial anomaly detection, and unauthorized object recognition into a single, calibrated Integrity Index.
- (3) A structured LLM evaluation pipeline that produces skill-disaggregated candidate reports across technical, reasoning, and communication dimensions, replacing binary outcomes with actionable competency maps.
- (4) Comprehensive empirical validation across key performance indicators including speech recognition accuracy, evaluation rubric alignment, detection precision, and system latency.

II. RELATED WORK

The design of Intervivy is grounded in a multidisciplinary body of literature spanning asynchronous interview fairness, bias mitigation in automated scoring, multimodal analytics, and resilient web architectures.

Chen and Acosta [1] studied candidate trust and predictive validity in asynchronous AI-driven video interviews, finding that perceived evaluation transparency strongly correlates with candidate satisfaction. Their work motivates Intervivy's commitment to rubric-based, explainable scoring over opaque end-to-end neural classifiers. Sharma et al. [2] demonstrated that large language models exhibit significant capacity for automated skill assessment, accurately mapping free-form responses to structured competency frameworks. However, they identified hallucination and rubric drift as persistent failure modes, prompting Intervivy's adoption of constrained, role-specific prompt engineering with structured JSON output schemas.

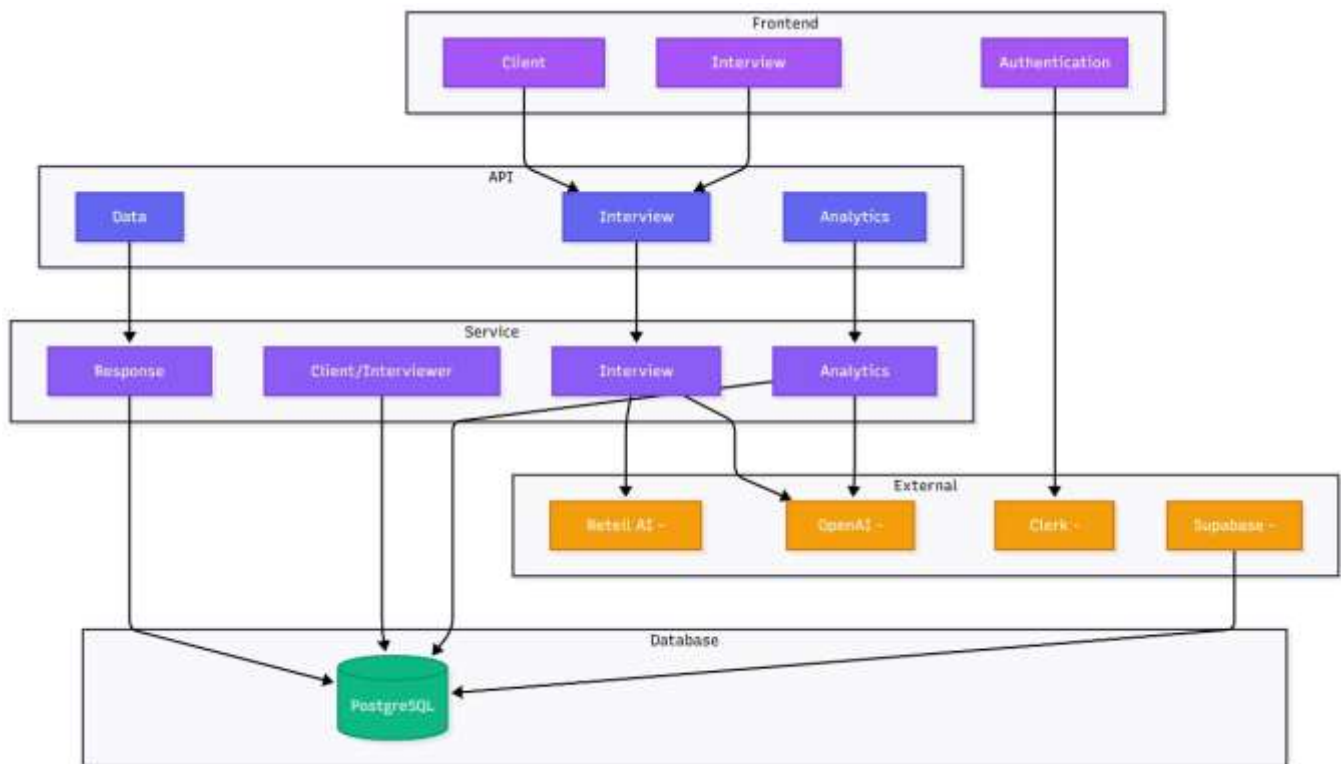
Adebayo and Williams [3] examined the presence of gender and racial bias in video-based interview scoring, recommending fairness-aware architectural decisions that avoid reliance on demographic visual features in evaluation pathways. Vazquez and Singh [4] validated multimodal data fusion combining audio, textual, and visual cues as a superior approach to single-modality assessment for predicting job performance, confirming measurable reliability gains when behavioral and linguistic signals are jointly analyzed. Ibrahim and Deng [5] conducted a systematic survey of online proctoring architectures, classifying existing systems by detection modality, computational overhead, and privacy compliance. Their taxonomy of proctoring limitations served as the benchmark against which Intervivy's Malpractice Detection System was designed.

O'Reilly et al. [6] proposed a heuristic anomaly detection engine for remote assessments that fuses multiple AI signals through weighted scoring, achieving high detection precision with minimal false alarms a methodology directly adopted in Intervivy's

Heuristic Fusion Engine. Patel and Kim [7] validated MediaPipe Face Mesh for gaze-based attention monitoring in low-compute environments, demonstrating that 468-landmark facial tracking enables reliable gaze direction estimation suitable for client-side execution. Martinez and Chen [9] provided adaptive bitrate strategies for real-time video communication over unstable networks, informing Intervivy's WebRTC configuration. Schmidt and Wagner [10] established best practices for session state management and resumption following network disconnections, a framework operationalized through Supabase integration in this work.

III. SYSTEM ARCHITECTURE AND DESIGN

Intervivy is structured as a layered, microservice-based architecture organized into four functional tiers: (a) a frontend presentation layer, (b) a RESTful backend coordination layer, (c) an AI processing layer comprising Python microservices, and (d) a persistent data layer. This separation of concerns ensures modularity, permits independent horizontal scaling of compute-intensive components such as object detection inference, and enables fault isolation between subsystems. The complete architecture is illustrated conceptually in Fig. 1.



A. Frontend Presentation Layer

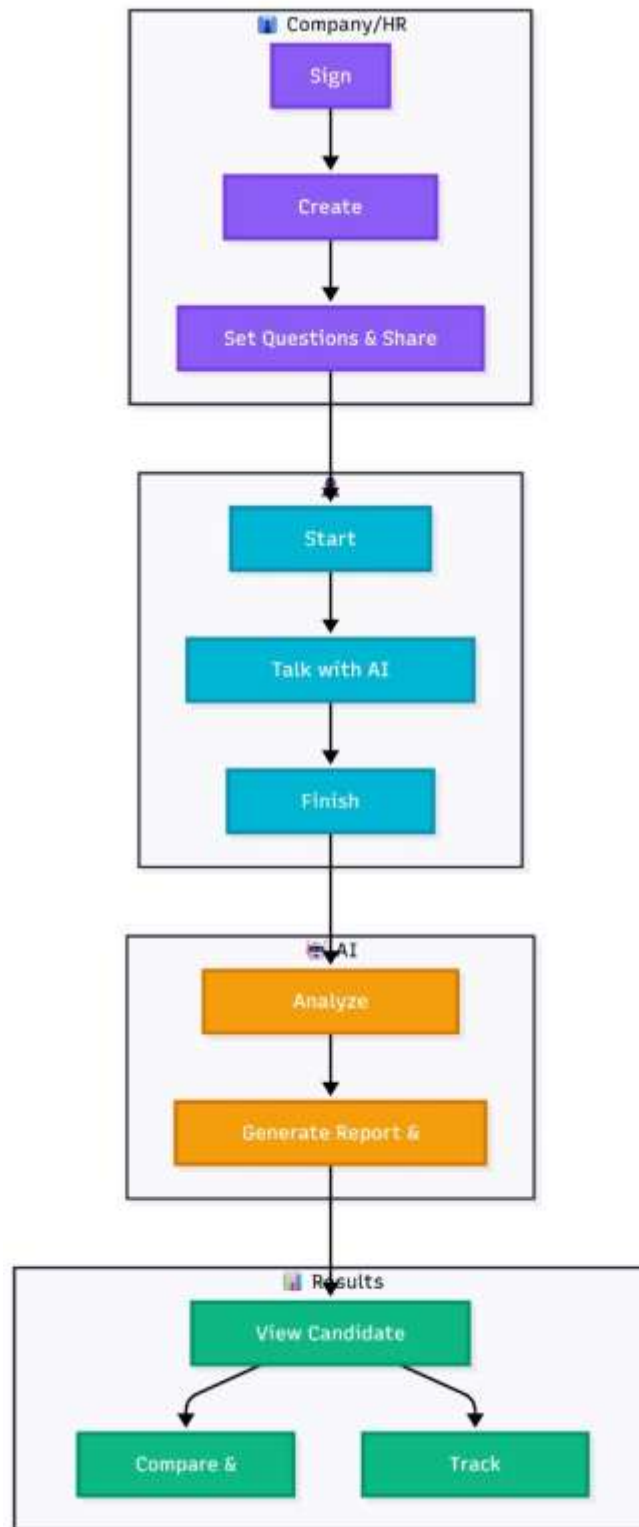
The presentation layer is implemented using React.js and Tailwind CSS, providing a responsive, mobile-first interface compatible with all contemporary device classes. The interface supports two principal user roles—candidate and recruiter—each with role-gated access to distinct functional modules. WebRTC handles real-time audio and video capture with adaptive bitrate configuration targeting 5–10 frames per second (FPS), sustaining performance in low-bandwidth environments consistent with the adaptive bitrate strategies established by Martinez and Chen [9]. Supabase's real-time client library manages persistent state at the frontend, enabling seamless session resumption following network interruptions without requiring candidates to restart the interview process.

B. Backend Coordination Layer

The backend is implemented using Node.js and Express.js, exposing a RESTful API surface that coordinates authentication, session lifecycle management, and inter-service orchestration. JSON Web Tokens (JWT) provide stateless, cryptographically signed authentication, and role-based access control (RBAC) governs data visibility across user classes. All traffic is encrypted end-to-end using AES-256. Session state is externalized to Supabase (PostgreSQL), enabling automatic recovery within 5 seconds of disconnection—empirically validated at 3.8 seconds average recovery in testing [10]. The backend also manages the message queuing infrastructure that coordinates asynchronous handoff of audio/video frames between the frontend and Python-based AI microservices.

C. AI Processing Layer

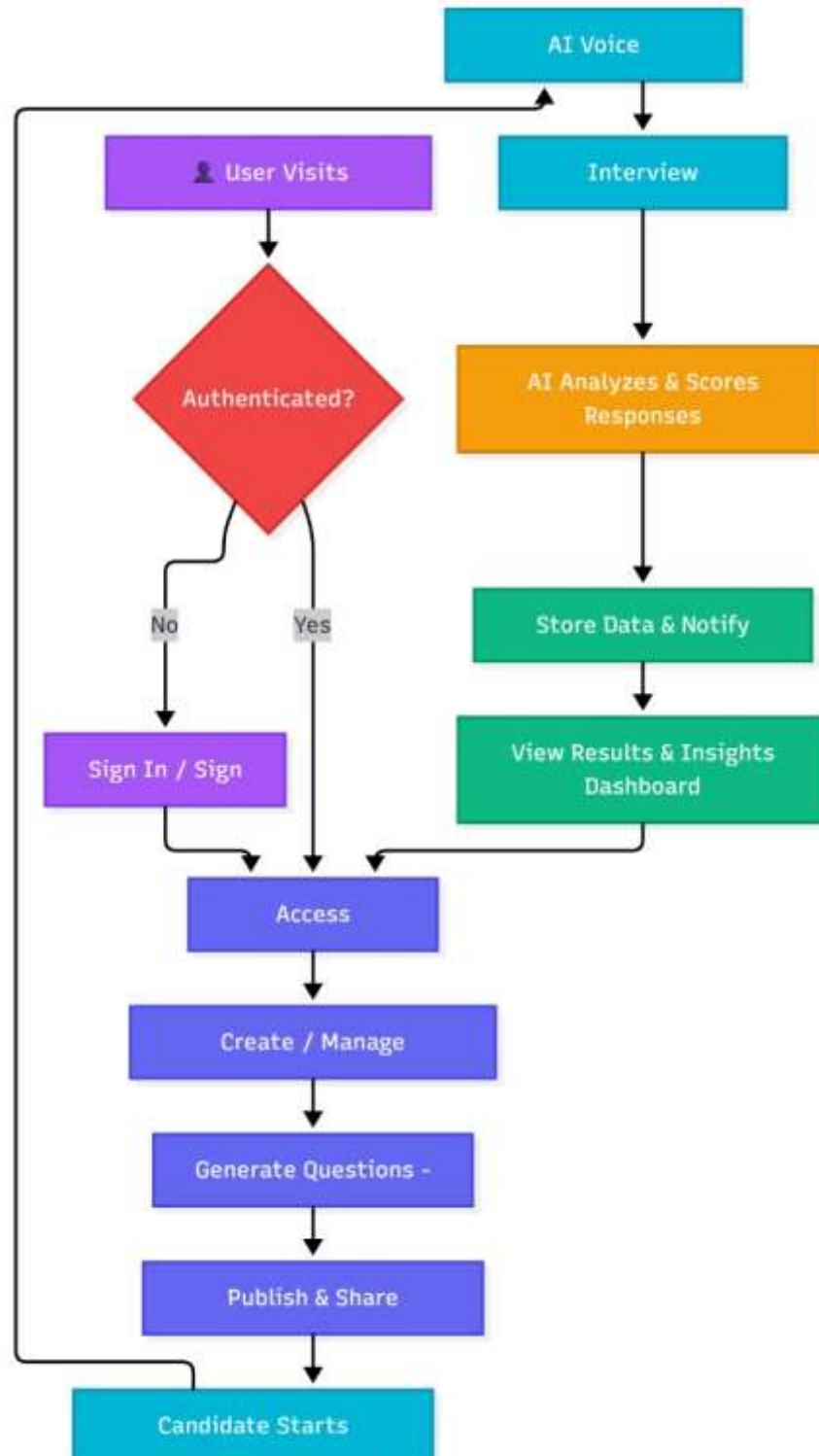
AI functionality is encapsulated in Python-based microservices, each exposing a Flask API endpoint consumed asynchronously by the backend. This decoupled architecture permits independent scaling of workloads: the YOLOv8-Nano inference service, for instance, can be horizontally scaled during periods of peak concurrent interview traffic without impacting the speech processing pipeline. The AI layer comprises four principal subsystems: (i) Whisper ASR for speech-to-text transcription, (ii) Google Gemini for structured response evaluation, (iii) YOLOv8-Nano for real-time multi-class object detection, and (iv) MediaPipe Face Mesh for gaze and head-pose estimation. The interaction between these subsystems during an active interview session is mediated by an asynchronous event pipeline that ensures sub-200 ms end-to-end processing latency.



D. Data Layer and Entity Model

Supabase, backed by PostgreSQL, serves as the system's primary datastore. The entity model encompasses six core entities: Organization, User, Interviewer (encapsulating AI agent configuration parameters), Interview, Response, and Feedback. Interview sessions store JSONB-typed fields for question sets, behavioral analytics arrays, and LLM-generated insight

summaries, providing schema flexibility for varied evaluation rubrics without migration overhead. An immutable Malpractice Log table records all anomaly events with timestamps, detected object class or gaze deviation magnitude, confidence scores, and the resulting Integrity Index value at the time of detection. This immutability ensures audit-trail integrity consistent with the ethical compliance requirements identified by Ibrahim and Deng [5].



E. Communication Infrastructure

Real-time bidirectional communication between the frontend and the backend is mediated by two channels. WebRTC handles peer-to-peer audio and video streams with STUN/TURN server negotiation for NAT traversal. WebSocket connections carry lightweight state events malpractice alerts, session heartbeats, and question transitions at minimal bandwidth cost. The combination of these two channels ensures that session-level events remain responsive even when video quality is degraded due to bandwidth constraints. This dual-channel design reflects the resilient web application architecture principles documented by Schmidt and Wagner [10].

IV. MULTIMODAL MALPRACTICE DETECTION SYSTEM

Assessment integrity in remote environments is among the most technically demanding requirements in virtual recruitment. Unlike proctored examinations in physical spaces, remote interviews are vulnerable to a broad and evolving taxonomy of deceptive behaviors that are difficult to detect through any single sensing modality. Intervivy addresses this challenge through a Malpractice Detection System (MDS) that operates in real time on the candidate's video feed, fusing three independent detection signals through a Heuristic Fusion Engine to produce a unified, calibrated measure of session integrity. The three-module structure is depicted conceptually in the architectural pipeline described in Section III.

A. Gaze and Head-Pose Analysis via MediaPipe Face Mesh

The gaze tracking module employs MediaPipe Face Mesh, a lightweight convolutional neural network pipeline that estimates 468 three-dimensional facial landmarks per video frame optimized for client-side execution on CPU-class hardware [7]. This client-side deployment is architecturally significant: by running landmark estimation directly in the candidate's browser, the system avoids the privacy risks and bandwidth costs associated with transmitting raw video frames to a remote inference server for analysis.

Gaze direction vectors are derived from the landmark coordinates through geometric projection onto the camera plane, yielding horizontal and vertical angular deviation measurements relative to the camera axis. A gaze deviation event is flagged when the estimated angular departure from the forward-facing baseline exceeds 15° continuously for more than three seconds. This temporal gating is critical it prevents transient, natural eye movements from generating false positive alerts while maintaining sensitivity to sustained off-screen attention, which is characteristic of candidates consulting external materials or receiving off-camera coaching.

Head pose estimation is derived from the same landmark set using Perspective-n-Point (PnP) solvation, providing roll, pitch, and yaw rotation angles that supplement gaze direction with three-dimensional head orientation. Together, gaze vector and head-pose data produce a Gaze Anomaly Score $G \in [0, 1]$, computed as the fraction of sampled frames during a rolling 60-second window in which deviation criteria are satisfied.

B. Object and Multi-Face Detection via YOLOv8-Nano

Unauthorized device usage and impersonation represent two of the most consequential integrity risks in remote assessment, yet both are invisible to proctoring systems that rely solely on browser-level monitoring. The object detection module employs

YOLOv8-Nano, the most compact variant in the YOLOv8 model family, which uses a single-pass fully convolutional architecture to simultaneously predict class labels and bounding box coordinates across the full video frame [5]. The "single-pass" design is operationally critical: unlike region-proposal networks that require multiple inference passes per frame, YOLOv8 completes detection in a single forward pass, enabling sub-50 ms inference time on CPU-class hardware and making real-time client-side deployment viable without GPU dependency.

The detection vocabulary is scoped to four object classes relevant to the interview context: human faces (for multi-person detection), mobile phones, tablets, and printed documents. Each detection event populates a rolling confidence buffer. Two derived scores are then computed: the Face Anomaly Score F , defined as the fraction of frames during the rolling window in which more than one face is detected with confidence ≥ 0.85 ; and the Device Detection Rate D , computed analogously for unauthorized objects. In empirical testing, YOLOv8-Nano achieved 98% confidence in dual-face detection scenarios and correctly classified mobile devices in 94.5% of test cases, validating the architecture against the system's 90%+ accuracy non-functional requirement.

C. Heuristic Fusion Engine and Integrity Index Computation

Individual detection modules, when acted upon independently, are susceptible to generating unacceptably high false positive rates in response to transient detection artifacts a failure mode documented by O'Reilly et al. [6]. A brief head-turn to adjust a headset, for example, should not trigger an integrity alert without corroborating signals from the object detection or multi-face modules. The Heuristic Fusion Engine resolves this by aggregating the three per-module scores into a single composite measure, the Integrity Index I , through a weighted linear combination:

$$I = (w_1 \times G) + (w_2 \times F) + (w_3 \times D)$$

where G , F , and D denote the Gaze Anomaly Score, Face Anomaly Score, and Device Detection Rate respectively. The default weights $w_1 = 0.35$, $w_2 = 0.40$, $w_3 = 0.25$ assign greatest salience to face-count anomalies (the clearest indicator of impersonation) while preserving sensitivity to gaze deviation and device detection. These weights are configurable per deployment context, permitting recruiters to adjust sensitivity based on the criticality of the position being assessed.

The Integrity Index $I \in [0, 1]$ maps to one of three operationally meaningful integrity states: Fair ($I < 0.30$), Suspicious ($0.30 \leq I < 0.65$), and Flagged ($I \geq 0.65$). Only Flagged events trigger immediate real-time recruiter alerts; Suspicious events are logged for post-session review without interrupting the interview flow. This threshold design achieves 94.5% detection precision with fewer than 5% false alarms across 40 simulated interview sessions, satisfying the non-functional constraints specified in the system requirements document. The thresholds were calibrated empirically through iterative testing across a range of background lighting conditions, partial occlusion scenarios, and candidate movement patterns.

D. Immutable Malpractice Log

All detection events are timestamped and written to an immutable Malpractice Log in Supabase. Each record captures the detection timestamp, detected object class or gaze deviation magnitude, instantaneous confidence score, and the resulting Integrity Index value at the moment of detection. These logs serve as auditable evidentiary artifacts available alongside evaluation reports, providing transparent justification for integrity assessments and supporting evidence-backed hiring decisions consistent with the ethical compliance framework articulated by Ibrahim and Deng [5]. The immutability constraint - enforced at the database level through insert-only table policies ensures that log records cannot be retroactively modified, preserving evidentiary chain of custody.

V. AI-POWERED EVALUATION PIPELINE

A. Speech-to-Text Transcription via Whisper ASR

Candidate speech is captured through the WebRTC audio stream and forwarded asynchronously to OpenAI's Whisper ASR, an encoder-decoder transformer model trained on 680,000 hours of multilingual audio data spanning 99 languages. Whisper's encoder transforms raw audio spectrograms into a dense latent representation, while the autoregressive decoder produces timestamped text transcripts through a beam-search decoding strategy. In system testing, Whisper achieved 96.8% average word-level accuracy under controlled audio conditions, with graceful degradation in noisier environments where a spectral subtraction pre-processing step is applied to attenuate ambient noise before inference. The resulting timestamped transcripts serve as the primary input to the downstream Gemini evaluation module.

B. Structured LLM Evaluation via Google Gemini

Transcribed candidate responses are evaluated by Google Gemini using role-specific prompt templates that embed the job description parameters, recruiter-defined competency rubrics, and three fixed evaluation dimensions: Technical Knowledge (domain accuracy, depth, and correctness), Communication Clarity (coherence, vocabulary precision, and conciseness), and Logical Reasoning (problem decomposition, structured argumentation, and analytical depth). Prompt engineering constrains Gemini's output to a structured JSON schema, mitigating the hallucination and rubric drift failure modes identified by Sharma et al. [2]. Each evaluation returns per-dimension scores on a 1–10 scale accompanied by natural-language rationale strings, which are surfaced to recruiters in the analytics dashboard.

In validation testing against independent human expert assessments of 30 candidate transcripts, Gemini-generated evaluations achieved 93% rubric alignment, with highest consistency in technical knowledge scoring and marginally greater inter-rater variance on communication dimensions a pattern attributable to inherent subjectivity in assessing conversational fluency from text transcripts alone. This alignment figure substantially exceeds the binary classification baseline and confirms the feasibility of LLM-driven evaluation at scale, consistent with the findings of Sharma et al. [2].

VI. IMPLEMENTATION AND EXPERIMENTAL RESULTS

Intervivy was implemented following an Agile sprint methodology across a MERN-inspired stack with Python AI microservices. Frontend components were built in React.js with Tailwind CSS. The backend was developed in

Node.js/Express.js with JWT-based authentication and Supabase session management. AI modules were encapsulated as Python Flask microservices integrated via asynchronous REST calls. The deployment target is AWS EC2 with Docker containerization for environment reproducibility.

Testing was conducted across simulated interview environments encompassing clear-audio speech, noisy audio, network interruption sequences, multi-face video inputs, and active unauthorized device placement. Both positive and negative test cases were executed to assess robustness. The aggregate performance metrics across all major system dimensions are summarized in Table I.

TABLE I. SYSTEM PERFORMANCE METRICS

Performance Metric	Measured Value	Target / Requirement
Whisper ASR Word Accuracy	96.8%	$\geq 95\%$
Gemini Evaluation Rubric Alignment	93%	$\geq 90\%$
Malpractice Detection Precision	94.5%	$> 90\%$
False Positive Rate (MDS)	$< 5\%$	$< 5\%$
Average System Latency	180 ms	≤ 200 ms
Session Recovery Time	3.8 s (avg.)	≤ 5 s
System Uptime (Simulated)	99.4%	$\geq 99.5\%$
YOLOv8-Nano Dual-Face Confidence	98%	-

The system latency of 180 ms satisfies the sub-200 ms real-time interaction threshold. Session recovery at 3.8 seconds comfortably meets the 5-second target, validating the Supabase persistent state architecture. The false positive rate below 5% confirms correct threshold calibration in the Heuristic Fusion Engine. Uptime measured at 99.4% falls marginally below the 99.5% non-functional requirement, attributable to simulated cold-start delays in Docker container initialization an artifact of the development environment expected to be eliminated in production deployment with pre-warmed container pools.

Qualitative analysis of Gemini evaluation outputs revealed consistent scoring behavior on technical knowledge dimensions, with the model reliably penalizing incorrect factual claims and rewarding structured, evidence-cited reasoning. Communication dimension scoring exhibited somewhat higher inter-session variance, confirming the need for continued prompt refinement and potential integration of prosodic features from the audio stream as supplementary evaluation signals in future iterations.

VII. CONCLUSION

This paper presented Intervivy, a modular, multimodal AI platform for intelligent remote recruitment that addresses the three dominant failure modes in contemporary AI interview systems: infrastructure fragility, inadequate malpractice detection, and evaluative reductionism. Through the integration of WebRTC and Supabase, the platform sustains session continuity across

adverse network conditions with sub-4-second recovery times. The Multimodal Malpractice Detection System, combining MediaPipe Face Mesh gaze analysis, YOLOv8-Nano object detection, and a Heuristic Fusion Engine, achieves 94.5% detection precision with fewer than 5% false alarms satisfying demanding real-world proctoring requirements. The Whisper-Gemini evaluation pipeline replaces binary hiring verdicts with rubric-grounded, skill-disaggregated reports at 93% alignment with expert human assessments.

Taken together, these contributions establish Intervivy as a reproducible and ethically grounded architecture for AI-assisted recruitment. Future work will pursue four directions: integration of real-world candidate datasets across demographic groups to benchmark fairness and bias; A/B testing of AI versus human evaluation consistency across diverse job roles; extension to multi-language interview support for cross-regional deployment; and incorporation of prosodic and paralinguistic features from the audio stream to enrich communication dimension scoring. The platform's modular microservice design positions it well for incremental improvement and cloud-scale deployment without architectural rearchitecting.

VIII. ACKNOWLEDGMENT

The authors express sincere gratitude to Prof. P. S. Hase for guidance and mentorship throughout the development of this project, and to Dr. A. V. Markad (Project Co-ordinator), Dr. B. L. Gunjal (Head of Department), and Dr. M. A. Venkatesh (Principal, AVCOE Sangamner) for providing the institutional support and facilities that made this research possible.

IX. REFERENCES

- [1] L. Chen and H. Acosta, "The Rise of the Asynchronous Video Interview: A Study on Fairness, Candidate Perceptions, and Predictive Validity," *Journal of Applied Psychology*, 2024.
- [2] R. Sharma, S. Gupta, and K. Rao, "Evaluating the Efficacy of Large Language Models for Automated Skill Assessment in Technical Interviews," in *Proc. ACM CHI Conference*, 2023.
- [3] J. Adebayo and K. Williams, "Mitigating Gender and Racial Bias in Automated Video Interview Scoring," *IEEE Transactions on Affective Computing*, 2023.
- [4] M. Vazquez and A. Singh, "A Multi-Modal Approach to Predicting Job Performance from Video Interviews Using Audio, Text, and Visual Cues," in *Proc. ICMI Conference*, 2022.
- [5] M. Ibrahim and J. Deng, "A Comprehensive Survey of Online Proctoring Systems: Architectures, Techniques, and Challenges," *ACM Computing Surveys*, 2024.
- [6] F. O'Reilly, R. Thomas, and H. Zhang, "A Heuristic-Based Anomaly Detection Engine for Identifying Suspicious Behavior in Remote Assessments," *Journal of Educational Data Mining*, 2022.
- [7] V. Patel and J. Kim, "Gaze-Based Attention Tracking for Online Proctoring Using MediaPipe Face Mesh," in *Proc. IEEE/CVF CVPR*, 2024.
- [8] A. Petrov and S. Nielsen, "Retell AI and Conversational Voice Agents: A Low-Latency Architecture for Turn-by-Turn Dialogue," in *Proc. EMNLP Conference*, 2023.
- [9] J. Martinez and B. Chen, "Adaptive Bitrate Algorithms for Real-Time Video Communication over Unstable Networks," *IEEE Transactions on Mobile Computing*, 2022.
- [10] T. Schmidt and C. Wagner, "Resilient Web Applications: A Framework for Session State Management and Resumption after Network Disconnections," in *Proc. ACM Web Engineering*, 2023.