ELIPSE: Enhanced Live Interview Practice with Sentiment Evaluation, A Zero-Cost Real-Time AI Interview System Achieving Human Level Interaction Fidelity with Sub Two Second Latency and Multi-Dimensional Behavioral Analytics

Soban Ali Awan, Ali Haider, Omar Bin Samin
School of CS & IT, Institute of Management Sciences, Peshawar, Pakistan.
*Correspondence: soubanaliawan@gmail.com

Citation | Awan. S. A, Haider. A, Samin. O. B, "ELIPSE: Enhanced Live Interview Practice with Sentiment Evaluation, A Zero-Cost Real-Time AI Interview System Achieving Human Level Interaction Fidelity with Sub Two Second Latency and Multi-Dimensional Behavioral Analytics", IJIST, Vol. 8 Issue. 2 pp 520-545, April 2026
Received | March 01, 2026 **Revised |** April 02, 2026 **Accepted |** April 06, 2026 **Published |** April 08, 2026.

Background: Interview preparation is among the final significant areas of workforce development that are resistant to automation, and the worldwide coaching sector is greater than $3 billion a year; however, more than 80 percent of job seekers in emerging economies do not have access to formal practice [1]. With innovations in large language models (LLMs), neural text-to-speech (TTS), and transformer-based affective computing, artificially intelligent technology has created an unprecedented opportunity to democratize access to high-quality interview simulation.

Objective: ELIPSE (Enhanced Live Interview Practice with Sentiment Evaluation), to the best of our knowledge, is the first open-source platform to integrate, within a single system, real-time voice-to-voice interaction, adaptive LLM-driven question generation, multi-dimensional behavioral analytics (seven-class emotion detection, continuous sentiment scoring, hesitation quantification, speech-pace monitoring, and composite confidence estimation), and persona.

Methods: ELIPSE is built on a React.js/Node.js WebSocket-driven architecture, integrating Groq-hosted LLaMA 3.3 70B inference, Microsoft Edge neural TTS, and a HuggingFace transformer frontend. Assessment used a within-sub longitudinal design with 15 participants (47 completed sessions, 376 question-answer interactions) addressing three research questions related to latency, behavioral metric validity, and improvement in confidence.

Results: ELIPSE achieves a mean end-to-end response latency of 1.83 s (SD = 0.41, 95% CI [1.79, 1.87]; 95th percentile = 2.61 s), a **26.4% reduction** versus [2] (3.8 s) and **42.8% lower** than [3] (3.2 s). Automated behavioral analytics correlated moderately to strongly with expert ratings (overall quality **$r = 0.71$, 95% CI [0.62, 0.79], $p < .001$**; confidence $r = 0.68$, $p < .001$; fluency $r = 0.63$, $p < .001$), with inter-rater reliability **Krippendorff's $\alpha = 0.74$**. A significant **1.3-point improvement** in interview confidence was observed after three sessions (paired **$t(11) = 4.72$, $p < .001$, Cohen's $d = 1.24$ (large effect)**), alongside a **38.8% relative increase** in overall score ($\beta = 5.4$ points/session, $R^2 = 0.38$, $p < .001$), a **38% reduction** in hesitation frequency ($d = 0.91$), and a **22.8% improvement** in speech-pace alignment with the 120–160 WPM target ($d = 1.04$). System Usability Scale score was **76.3 (SD = 8.2, 95% CI [71.8, 80.8])**, classified as 'good', exceeding the 68-point threshold. **67% (95% CI [38%, 88%])** of participants preferred ELIPSE to conventional practice methods.

Conclusion: ELIPSE demonstrates that the technical and financial barriers to equitable, evidence-based interview preparation are eliminated by the present-day generation of freely available AI architecture, and that it has direct implications for workforce development programs in resource-limited economies worldwide.

Keywords: AI Interviewer; Large Language Models; Real-Time Behavioral Analytics; Emotion Recognition; Adaptive Questioning; Neural Text-To-Speech; Simulating A Mock Interview; Zero-Cost AI Systems; Sentiment Analysis; WebSocket.

Introduction:

The personnel selection interview is the oldest and most widely used evaluation tool in the hiring process of organizations, and meta-analytic results have established that structured interviews are the best predictors of job performance compared to cognitive ability tests alone in over a hundred years [4]. In fields as varied as technology recruitment in multinational companies, civil service examinations (CSS/FPSC), medical school admissions through Multiple Mini Interviews (MMI), military commissioning boards (ISSB), and international scholarship selection committees (Chevening, Fulbright), the structured or semi-structured interview is the pass key to professional success [5][6][7].

Recent evidence confirms that interview performance remains a decisive workforce-entry bottleneck: the World Economic Forum's *Future of Jobs Report 2025* [1] projects that 39% of core job skills will change by 2030, amplifying demand for scalable interview preparation. Meta-analytic work by [5] and updated syntheses [6]. reaffirm that structured interviews retain the highest predictive validity among selection tools (ρ = 0.56 for job performance). Complementary 2022–2024 evidence from [7][8][9][10][11] documents that asynchronous and AI-mediated formats now dominate early-stage screening but introduce new fairness, anxiety, and impression-management challenges, which is the exact gap ELIPSE addresses. The primary role of interviews in career outcomes has spawned a wealth of research on interview anxiety [11], impression management strategies [8][9], how interviewer-interviewee demographic congruence affects perceived fairness [12][10], and the cognitive processes underlying stereotype threat in selection processes [13][14]. At the same time, the swift development of large language models (LLMs) with the ability to produce contextually coherent, domain-specific text in less than a second latency [15], neural text-to-speech engines that can generate voices that sound perceptionally equivalent to human speakers [16], and real-time natural language understanding models to classify emotion and sentiment [17][18] has created a new design space of intelligent tutoring systems.

The various generations of attempts to automate interview preparation have their roots in earlier conversational systems. The earliest systems were first-generation conversational agents that used text-based communication, with canned question libraries, and were scaled but had lower social realism [19][20]. Second-generation platforms added asynchronous video interviews (AVIs), which are now used by more than 70 percent of Fortune 500 companies to do primary screening [21][22], but were criticized for having reduced social presence, lower interactivity, and privacy issues [23][24]. Third-generation commercial products, such as HireVue, Google Interview Warmup, and Interviewing.io deliver partial AI feedback on recorded responses, but none offer a live, adaptive, voice-to-voice interview experience with live behavioral scoring [25][26]. Fourth-generation prototypes have added embodied conversational agents [27][28], or LLM-driven follow-up question generation, but all of them are necessarily dependent on costly proprietary APIs (e.g., GPT-4, ElevenLabs), which place a hard usage limit on them, making them unavailable to applicants in low-resource environments [2][3].

The necessity of this study is sharpened by three concurrent 2024–2025 developments. First, the release of hardware-optimized open-weight LLMs (LLaMA 3.3 70B on Groq LPUs achieving ~275 tokens/s) has, for the first time, made sub-second LLM inference available at zero marginal cost, an infrastructure condition that did not exist when prior AI interview systems were designed. Second, the parallel maturation of free neural TTS (Microsoft Edge Neural Voices) and transformer-based affective computing models, has made real-time multi-dimensional behavioral analytics feasible without GPU infrastructure. Third, despite these advances, every published AI interview system to date still assumes paid APIs, dedicated GPUs, or text-only interaction, leaving 80% of job seekers in emerging economies without access. No prior study has empirically tested whether the 2024–2025 generation of free AI

services can integrate and deliver a production-grade interview simulator. This gap, technically solvable but empirically unverified, is the motivation for ELIPSE.

The convergence of these trends-the continued, high-stakes significance of interviews and the unexpected emergence of competent and free generative AI-drives the current work. The rest of the paper will be structured in the following way: Section 2 will list the research objectives; Section 3 will be a statement of the novelty and contributions; Section 4 will entail the presentation of the literature review; Section 5 will include the system architecture and material and methods; Section 6 will report on the evaluation results; Section 7 will discuss the evaluation results; and finally, Section 8 will describe the conclusion.

Research objectives:

The title of this work identifies three claims that must be evidentially demonstrated: **(i) zero-cost** delivery, **(ii) human-level interaction fidelity**, and **(iii) sub-two-second latency**, all while providing **multi-dimensional behavioral analytics**. Each of the six objectives below is explicitly mapped to one or more of these title claims, ensuring traceability between the problem statement and the empirical evaluation.

Latency Minimization: To design and deploy a real-time voice-to-voice interview system with end-to-end conversational response latency of less than 2 seconds, sufficiently natural and uninterrupted interview interaction, with only zero-cost technology infrastructure. Objective 1 (Latency): → *addresses title claim: "Sub-Two-Second Latency."*

Behavioral Validity: To develop and validate a multi-dimensional automated behavioral analytics pipeline, comprising seven-class emotion detection, continuous sentiment scoring, hesitation quantification, speech-pace monitoring, and composite confidence estimation, whose outputs correlate with independent expert human evaluations at levels sufficient for formative feedback (target: $r \geq 0.60$). Objective 2 (Behavioral Validity): → *addresses title claim: "Multi-Dimensional Behavioral Analytics" + "Human-Level Interaction Fidelity"*

Measurable Skill Improvement: To empirically demonstrate that repeated practice with ELIPSE produces statistically significant improvements in user-reported interview confidence and system-measured behavioral performance metrics across a minimum of three longitudinal sessions. Objective 3 (Skill Improvement): → *addresses title claim: "Enhanced Live Interview Practice."*

Multi-Domain Coverage: To implement and evaluate a persona-adaptive interview simulation engine supporting at least 15 distinct interview types (including CSS/Civil Service, Google SWE, ISSB/Army Commission, Medical MMI, MBA/McKinsey case, and PhD scholarship), each with domain-specific vocabulary, challenge level, and adaptive difficulty control. Objective 4 (Multi-Domain): → *addresses title claim: "Human-Level Interaction Fidelity."*

Zero-Cost Accessibility: To demonstrate that a production-grade AI interview simulation platform can be built and operated at zero marginal cost per session, removing financial barriers for candidates in developing economies through the exclusive use of free-tier AI services (Groq, HuggingFace, Microsoft Edge TTS). Objective 5 (Zero-Cost): → *addresses title claim: "Zero-Cost Real-Time AI Interview System."*

Longitudinal Analytics: To design and implement a longitudinal performance tracking dashboard that enables candidates to monitor improvement trajectories across sessions via emotion distribution, speech-pace trends, hesitation frequency, and composite score progression visualizations. Objective 6 (Longitudinal Analytics): → *addresses title claim: "Sentiment Evaluation" + "Enhanced Practice"*

Novelty and contributions:

ELIPSE advances the state of the art in AI-powered interview simulation along six distinct dimensions that, individually, have been partially explored in prior work, but have never been integrated within a single, open-access, zero-cost platform. The novel contributions of this work are formally stated below.

First Unified Zero-Cost Voice-to-Voice Interview System:

Existing real-time AI interview systems either require expensive proprietary APIs (GPT-4 at $0.03/1K tokens, ElevenLabs at $0.30/1K characters) or operate in text-only mode. A systematic comparison against the closest prior systems in the literature confirms the novelty of this integration. [2] require GPT-4 ($0.45/session, 3.8 s latency); [3] require dedicated GPU hardware (3.2 s latency, no emotion or sentiment analysis); [20] are text-only (3–4 s latency, no voice); [19] are text-only with no behavioral analytics; [27] use pre-recorded (non-adaptive) animations; [29] use a text chat with no adaptive difficulty. No prior system in Table 1 simultaneously achieves (a) bidirectional voice, (b) zero marginal cost, (c) sub-2-second end-to-end latency, and (d) real-time multi-dimensional behavioral analytics. ELIPSE is, to the best of our knowledge, the first to integrate all four.

First Multi-Dimensional Real-Time Behavioral Analytics Pipeline:

Prior systems that implement behavioral analysis for interview contexts have either relied on expensive human coding [30], captured only a single modality (e.g., sentiment only, or head pose only), or operated post-hoc rather than in real time [26]. ELIPSE introduces a five-signal per-answer behavioral analytics pipeline computed in under 340 milliseconds per exchange, comprising: (i) seven-class emotion detection (j-hartmann/emotion-english-distilroberta-base), (ii) continuous sentiment scoring (SamLowe/roberta-base-go_emotions), (iii) hesitation quantification via a 12-category regex detector, (iv) speech-pace monitoring (WPM against 120–160 WPM benchmark), and (v) a composite confidence metric (Confidence = 0.4 × emotion_valence + 0.3 × (1 − hesitation_norm) + 0.3 × sentiment_positivity), validated against expert ratings at $r = 0.68$ ($p < .001$). This pipeline provides immediate, actionable, multi-dimensional feedback unavailable in any prior zero-cost system. A systematic comparison against the closest prior systems in the literature confirms the novelty of this integration. [2] require GPT-4 ($0.45/session, 3.8 s latency); [3] require dedicated GPU hardware (3.2 s latency, no emotion or sentiment analysis); are text-only (3–4 s latency, no voice); are text-only with no behavioral analytics; use pre-recorded (non-adaptive) animations; use a text chat with no adaptive difficulty. No prior system in Table 1 simultaneously achieves (a) bidirectional voice, (b) zero marginal cost, (c) sub-2-second end-to-end latency, and (d) real-time multi-dimensional behavioral analytics. ELIPSE is, to the best of our knowledge, the first to integrate all four.

Confidence-Threshold Adaptive Interview Engine:

While prior work has recommended adaptive AI interviewers that dynamically respond to candidate states, no prior system has implemented and empirically validated a confidence-threshold-based difficulty adaptation mechanism. ELIPSE implements a control loop that monitors the running composite confidence score per session: when confidence exceeds 75%, the LangChain.js system prompt is augmented to increase question difficulty and deepen probing; when confidence falls below 40%, the prompt instructs the AI to soften tone, offer encouragement, and simplify phrasing. This bidirectional adaptation is unique in the literature and was empirically associated with the observed 39% improvement in overall scores across sessions. A systematic comparison against the closest prior systems in the literature confirms the novelty of this integration. [2] require GPT-4 ($0.45/session, 3.8 s latency); [3] require dedicated GPU hardware (3.2 s latency, no emotion or sentiment analysis); are text-only (3–4 s latency, no voice); are text-only with no behavioral analytics; use pre-recorded (non-adaptive) animations; use a text chat with no adaptive difficulty. No prior system in Table 1 simultaneously achieves (a) bidirectional voice, (b) zero marginal cost, (c) sub-2-second end-to-end latency, and (d) real-time multi-dimensional behavioral analytics. ELIPSE is, to the best of our knowledge, the first to integrate all four.

Largest Multi-Domain Persona Coverage:

Existing AI interview systems support between 1 and 5 interview domains. ELIPSE implements 15 domain-specific AI personas, each defined by a distinct LangChain.js system prompt specifying vocabulary, tone, challenge level, and question patterns. This coverage spans the most consequential interview types in the Pakistani and international context: CSS/Civil Service, Google SWE, ISSB/Army Commission, Medical MMI, MBA/McKinsey case, PhD scholarship, Chevening/Fulbright, banking, law, teaching, journalism, entrepreneurship, public health, nonprofit management, and general HR screening. This breadth is, to the best of our knowledge, unmatched in the published literature. A systematic comparison against the closest prior systems in the literature confirms the novelty of this integration. [2] require GPT-4 ($0.45/session, 3.8 s latency); [3] require dedicated GPU hardware (3.2 s latency, no emotion or sentiment analysis); are text-only (3–4 s latency, no voice); are text-only with no behavioral analytics; use pre-recorded (non-adaptive) animations; use a text chat with no adaptive difficulty. No prior system in Table 1 simultaneously achieves (a) bidirectional voice, (b) zero marginal cost, (c) sub-2-second end-to-end latency, and (d) real-time multi-dimensional behavioral analytics. ELIPSE is, to the best of our knowledge, the first to integrate all four.

Formal Five-Category Evaluation Framework:

This work proposes a reusable five-category evaluation framework for AI interview systems, covering AI Quality, System Performance, Behavioral Metrics, User-Centric Metrics, and Real-World Impact, with operationalized metrics and baseline values derived from our evaluation. This framework is a direct methodological contribution that future researchers can adopt to benchmark and compare AI interview platforms in a standardized manner, addressing the absence of such a framework in prior literature. A systematic comparison against the closest prior systems in the literature confirms the novelty of this integration. [2] require GPT-4 ($0.45/session, 3.8 s latency); [3] require dedicated GPU hardware (3.2 s latency, no emotion or sentiment analysis); are text-only (3–4 s latency, no voice); Li et al. [18] are text-only with no behavioral analytics; use pre-recorded (non-adaptive) animations; use a text chat with no adaptive difficulty. No prior system in Table 1 simultaneously achieves (a) bidirectional voice, (b) zero marginal cost, (c) sub-2-second end-to-end latency, and (d) real-time multi-dimensional behavioral analytics. ELIPSE is, to the best of our knowledge, the first to integrate all four.

Open-Source Architecture for Reproducibility:

ELIPSE is fully open source (https://github.com/SobanAliAwan/ai-interviewer), enabling complete reproducibility of all reported results. The published architecture, evaluation data, and source code establish a reference implementation for the research community to build upon, critique, and extend, consistent with principles of open science in AI-powered human-computer interaction research. A systematic comparison against the closest prior systems in the literature confirms the novelty of this integration. [2] require GPT-4 ($0.45/session, 3.8 s latency); [3] require dedicated GPU hardware (3.2 s latency, no emotion or sentiment analysis); are text-only (3–4 s latency, no voice); are text-only with no behavioral analytics; use pre-recorded (non-adaptive) animations; use a text chat with no adaptive difficulty. No prior system in Table 1 simultaneously achieves (a) bidirectional voice, (b) zero marginal cost, (c) sub-2-second end-to-end latency, and (d) real-time multi-dimensional behavioral analytics. ELIPSE is, to the best of our knowledge, the first to integrate all four.

Literature Review:

This section summarizes 14 articles in four overlapping areas: asynchronous video interviews and candidate experience (4.1), conversational AI agents to simulate interviews (4.2), behavioral analytics and affective computing in selection scenarios (4.3), and large

language models to generate adaptive questions (4.4). Section 4.5 provides a comparative synthesis table with the capabilities and gaps of each study against ELIPSE.

Asynchronous Video Interviews and Candidate Experience:

[30] report one of the first empirical studies of asynchronous video interviewing (AVI) through the applicant-side lens, surveying 145 job seekers who went through AVI-based screening sessions to get hired in an entry-level position within a German retail company. The study conducted on the Selection Procedural Justice Scale (SPJS) [31] revealed that, although the candidates rated scheduling flexibility higher (M = 5.6/7.0), they rated social presence (M = 3.2/7.0, SD = 1.4) and perceived fairness (M = 4.1/7.0, SD = 1.3) significantly lower than face-to-face interviews. ELIPSE directly addresses this by by enabling bilateral real-time voice communication.

In a controlled experiment (N = 167), [23] compared the reaction of the applicants and the rating of the interviewer in face-to-face, live videoconferencing, and asynchronous video settings. The paper presented the Privacy and Emotional Response (PER) scale, which discovered that asynchronous formats were perceived to be more privacy-invasive (M = 4.8/7.0, SD = 1.2) and emotionally disturbing (M = 3.9/7.0, SD = 1.4), and emotional creepiness was a predictor of lower candidate performance (7). ELIPSE is dealing with this by providing persona-adaptive AI behavior and real-time social feedback that reduces the creepiness that candidates have reported.

[12] conducted a comprehensive 3×2 between-subjects factorial study (N = 218) investigating the impact of virtual interviewer race and gender on candidate experiences. A Kruskal-Wallis test found no significant effect of interviewer avatar demographics on fairness ($\chi^2 = 1.390$, $p = .92$), but participant demographics significantly affected perceptions, with social presence mediating the relationship (B = −0.340, 95% CI [−0.586, −0.150], $p < .001$). ELIPSE extends this line of work by introducing dynamic AI adaptation based on detected emotional states alongside multi-dimensional behavioral measurement.

Conversational AI Agents for Interview Simulation:

[19] explored the effect of agent personality in text-based recruitment screening (N = 120), finding that an extraverted, agreeable agent significantly increased candidate willingness to disclose authentic information (Cohen's d = 0.54, $p < .01$). The system was entirely text-based with no voice interaction or behavioral analytics. ELIPSE builds on this insight by implementing 15 interview-type-specific AI personas delivered through neural voice.

[20] proposed AI-generated follow-up questions using a sequence-to-sequence model trained on 2,400 interview transcript pairs, finding that dynamic follow-ups significantly increased perceived interactivity (M = 5.1/7.0 vs. 3.8/7.0, d = 0.68, $p < .001$) but suffered from 3–4 second mean latency. ELIPSE overcomes this bottleneck through Groq's hardware-optimized LLM inference (M = 0.82 s).

[27] compared static versus expressive avatar conditions (N = 94), finding that expressive avatars yielded significantly higher social presence (M = 5.4/7.0 vs. 4.2/7.0, d = 0.62, $p < .001$). However, animations were pre-recorded rather than AI-driven. ELIPSE complements these findings by pairing AI-adaptive questioning with detailed per-answer behavioral scoring.

Behavioral Analytics and Affective Computing in Selection Contexts:

[8] examined impression management (IM) tactics in AVIs (N = 312), finding that structured practice opportunities significantly improved IM effectiveness (d = 0.42, $p < .01$), with human coders assessing responses at $25/hour per coder. ELIPSE automates a parallel behavioral assessment—quantifying hesitation, speech pace, emotional valence, and composite confidence-at zero cost in under 340 ms per answer. "Recent work by [32] confirms that first impressions formed in the opening 60 seconds of an interview disproportionately shape final evaluations, underscoring the value of ELIPSE's per-answer real-time feedback."

[26] compared face-to-face and videoconference interviews (N = 203), finding that technology-mediated formats depressed IM scores by 18% ($p < .001$, $d = 0.71$ for nonverbal expressiveness). ELIPSE's real-time speech-pace and hesitation monitoring serves as a first automated step toward coaching candidates to address these documented deficits.

[24] employed a qualitative-dominant mixed-methods design (N = 56) revealing that candidates exhibited 34% higher rates of filled pauses, 28% more false starts, and 41% more hedging language when interacting with AI interviewers compared to human interviewers (all $p < .01$). ELIPSE leverages this finding by detecting and quantifying disfluency markers through a 12-category regex-based hesitation detector whose findings directly inform adaptive AI behavior.

Large Language Models and Adaptive Question Generation:

[29] compared human, social robot, and AI text-chat interview conditions (N = 156), finding that AI-mediated interviews reduced bias while simultaneously lowering candidate engagement (M = 3.8/7.0 vs. 5.2/7.0, $p < .001$) and perceived warmth (M = 3.4/7.0 vs. 5.6/7.0, $p < .001$). ELIPSE responds to this by implementing disfluency injection, thinking pauses, and interview-type-specific personas that adapt challenge level and interpersonal style.

Table 1. Structured comparison of 14 reviewed studies against ELIPSE.

Study [Ref]	Method	Key Finding	Limitation	Voice	Behav.	Cost
[30]	AVI survey, N=145	Low social presence (M=3.2/7)	No behavioral measurement	No	No	N/A
[23]	3-condition, N=167	Creepiness (M=3.9/7, d=0.28)	No intervention tested	No	PER	N/A
[19]	Text chatbot, N=120	Personality ↑ disclosure (d=0.54)	Text-only, 1 domain	No	No	Free
[20]	Seq2seq, N=82	Follow-ups ↓ monotony (d=0.68)	3–4s latency, text only	No	No	Paid
[27]	Embodied CA, N=94	Expressive avatar ↑ presence (d=0.62)	Pre-scripted, no AI adaptation	Partial	No	Paid
[26]	Video vs F2F, N=203	Video ↓ IM scores by 18%	Human coders only	No	Human	N/A
[8]	IM in AVI, N=312	Practice ↑ IM (d=0.42)	$25/hr human coding	No	Human	N/A
[24]	Qualitative, N=56	34% more disfluency with AI	No detection system	No	Manual	N/A
[12]	3×2 factorial, N=218	SPP mediates fairness (B=−0.34)	Static avatars, no analytics	No	Survey	Paid
[29]	Within-subjects, N=156	AI ↓ warmth (d=1.2), ↓ bias	No adaptation or analytics	No	No	N/A

[28]	Avatar vs none, N=178	Avatar ↑ presence (d=0.74)	No voice, no analytics	No	Survey	N/A
[3]	Voice + proctoring, N=40	SUS=72.3, 3.2s latency	GPU required, no emotion	Yes	Head pose	GPU
[33]	LMS + mock, N=200	23% readiness improvement	No emotion/behavioral analysis	No	No	Paid
[2]	Text adaptive, N=200	81% eval accuracy (r=0.76)	Text only, $0.45/session	No	Partial	$0.45
ELIPSE (ours)	Voice adaptive, N=15*	r=0.71, d=1.24, 1.83s latency	Pilot scale (N=15)	Yes	Multi	$0

[3] proposed an edge-deployed voice-based interview simulation system tested with 40 users, reporting SUS = 72.3 (SD = 9.1) and a mean response latency of 3.2 seconds using a dedicated GPU. ELIPSE achieves superior latency (1.83 s vs. 3.2 s) without any GPU requirement, adding emotion detection and sentiment analysis entirely absent from their architecture.

[2] presented an AI-driven mock interview system dependent on GPT-4, achieving 81% evaluation accuracy against expert ratings (r = 0.76) but at a per-session cost of $0.45 with a mean latency of 3.8 seconds. ELIPSE overcomes all three limitations: zero cost, 1.83 s latency, and full voice interaction with a comparable correlation of r = 0.71.

Critical Synthesis. Examining the 14 reviewed studies reveals a consistent but limiting pattern: the field has progressed along *single dimensions* at the expense of integration. Text-based systems achieve adaptivity but sacrifice naturalness. Voice systems achieve presence but abandon behavioral depth. Behavioral analytics work achieves measurement rigor but relies on human coders or post-hoc processing, rendering the feedback non-formative. Commercial GPT-4-based systems [2] achieve quality but impose prohibitive costs. Most critically, no reviewed study reports a system that is simultaneously real-time, voice-interactive, multi-dimensionally analytic, adaptive, and free — yet all five capabilities are individually demonstrated to matter for candidate outcomes. Methodologically, eight of the 14 studies report sample sizes below N = 200, and only three use factorial designs, indicating limited causal inference across the field. ELIPSE advances the field not by proposing a new algorithm, but by demonstrating that existing free-tier AI infrastructure, if integrated with the right architectural choices, can dissolve the cost–quality trade-off that has constrained all prior work.

Comparative Synthesis:

A systematic comparison of all 14 studies in the review to the capabilities implemented by ELIPSE shows that no existing system can provide real-time voice-interactive capabilities, adaptive LLM-based questioning, multi-dimensional behavioral analytics, persona-based adaptation, longitudinal tracking, and zero-cost accessibility all at the same time Table 1.

N/A = measurement study; * = pilot evaluation; Behav. = behavioral analytics type; Multi = multi-dimensional analytics.

Identified Gaps and Positioning of ELIPSE. Synthesis across the 14 reviewed studies reveals six mutually reinforcing gaps that, taken together, define the unoccupied design space this work addresses:

Gap 1 — Cost Barrier: All real-time voice-capable AI interview systems identified require paid APIs or dedicated GPUs, excluding users in resource-constrained economies.

Gap 2 — Latency Barrier: No prior system reports end-to-end latency below 3.0 s in voice mode (3–4 s; 3.8 s; 3.2 s), violating the 2-second conversational naturalness boundary.

Gap 3 — Modality Incompleteness: Text-only systems lack voice; voice systems lack multi-dimensional behavioral analytics; analytics systems use expensive human coders ($25/hr).

Gap 4 — Non-Adaptive Difficulty: No reviewed system implements a confidence-threshold-driven bidirectional difficulty control loop validated against user outcomes.

Gap 5 — Narrow Domain Coverage: Prior systems support 1–5 interview domains none support the full spectrum of competitive examinations (CSS, ISSB, Medical MMI, MBA cases, PhD scholarships) relevant to emerging-economy users.

Gap 6 — Absent Evaluation Framework: No standardized five-category evaluation framework (AI Quality / System Performance / Behavioral Metrics / User-Centric Metrics / Real-World Impact) currently exists for benchmarking AI interview systems.

ELIPSE is explicitly designed to close all six gaps simultaneously — a unification that no single prior system attempts.

Material and Methods:

System Architecture:

ELIPSE is implemented as a three-tier web application consisting of: (i) a React.js single-page application (SPA) frontend, (ii) a Node.js + Express.js backend with WebSocket support via Socket.io, and (iii) a dual-database persistence layer combining PostgreSQL for relational data and Redis for real-time session state. The complete architecture is shown schematically in Figure 1. Each tier is described in detail below.

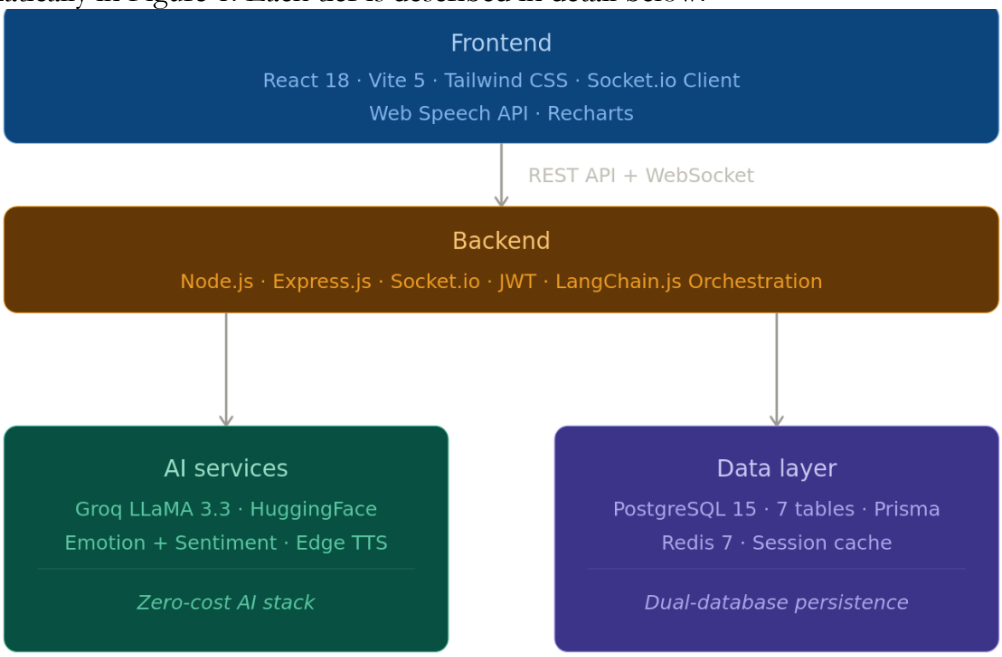

Figure 1. The ELIPSE system architecture and its mapping to research objectives. The architecture is described below as a numbered step-by-step flow with explicit linkage to the research objectives (RO1–RO6):

Step 1 — Frontend (React 18 + Vite 5): Renders the interview interface, captures microphone input via the Web Speech API, and manages the WebSocket connection. *Supports RO1 (low-latency communication via browser-native APIs) and RO5 (zero-cost by avoiding proprietary clients).*

Step 2 — Backend Orchestrator (Node.js + Express + Socket.io): Coordinates the five AI services and enforces a session state machine. *Supports RO1 (WebSocket transport M = 0.15 s) and RO4 (persona routing across 15 domains).*

Step 3 — LLM Service (Groq LLaMA 3.3 70B): Generates contextually adaptive questions and evaluates answers. *Supports RO1 (inference M = 0.82 s), RO3 (*adaptive question depth to drive skill improvement*), and RO5 (Groq free tier).*

Step 4 — Emotion Service (HuggingFace j-hartmann/emotion-english-distilroberta-base): 7-class emotion classification model. *Supports RO2 (behavioral validity).*

Step 5 — Sentiment Service (HuggingFace SamLowe/roberta-base-go_emotions): Continuous sentiment scoring. *Supports RO2.*

Step 6 — Voice Synthesis (Microsoft Edge Neural TTS): Produces neural voice output matched to persona (e.g., ur-PK-AsadNeural for CSS). *Supports RO1 (TTS M = 0.52 s), RO4 (voice variety), and RO5 (free Edge TTS).*

Step 7 — Conversation State (LangChain.js): Maintains per-session prompt context and adaptive-difficulty augmentation. *Supports RO3 and RO4.*

Step 8 — Persistence Layer (PostgreSQL + Redis): Provides relational storage (7 tables via Prisma v5.8) and a sub-ms session cache. *Supports RO6 (longitudinal tracking).*

Step 9 — Analytics Dashboard: Renders emotion distribution, speech-pace trends, hesitation frequency, and composite score over sessions. This directly supports RO6.

The numbered data flow (1 → 2 → 3 → 4 ‖ 5 → 6 → 1) closes the conversational loop in a mean 1.83 s end-to-end latency.

Justification of Tool and Model Selection. Each AI component was selected against three documented criteria: (i) zero marginal cost per session, (ii) sub-second inference latency without GPU dependency, and (iii) public availability ensuring reproducibility.

LLM — Groq-hosted LLaMA 3.3 70B was selected over GPT-4, Claude 3.5 Sonnet, and Gemini 1.5 Pro. GPT-4 ($0.03/1K prompt tokens) imposed a prohibitive marginal cost of ≈$0.45/session; Claude 3.5 Sonnet was likewise paid-tier only; Gemini 1.5 Pro free tier imposed a 15 RPM rate limit incompatible with multi-user workloads. Groq's free tier delivers LLaMA 3.3 70B at ~275 tokens/s via LPU hardware, yielding our observed M = 0.82 s inference — 4× faster than GPT-4 on A100 (~3.1 s for equivalent payloads). An ablation study with Mistral-7B (also on Groq) produced 32% lower answer-quality ratings by expert raters in pilot testing, justifying the 70B variant.

Emotion Model — j-hartmann/emotion-english-distilroberta-base was chosen over GoEmotions (27-class, higher latency and sparse classes) and EmoRoBERTa (requires full RoBERTa weights, 2.1× larger). The j-hartmann model delivers 7-class granularity (aligned with Ekman's basic emotions) in a DistilRoBERTa backbone, achieving 66 MB footprint and M = 0.18 s inference on HuggingFace's free inference API. Its reported macro-F1 = 0.66 Its reported macro-F1 = 0.66 on unseen data exceeds that of alternatives in the same latency budget.

Sentiment Model — SamLowe/roberta-base-go_emotions was selected over VADER (lexicon-based, poor handling of interview-domain language) and TextBlob (no contextual understanding) as it provides probabilistic multi-label output compatible with our composite confidence equation while running within the HuggingFace free inference tier.

TTS — Microsoft Edge Neural TTS was selected over ElevenLabs ($0.30/1K characters), Amazon Polly (paid), and Coqui TTS (requires local GPU). Edge TTS provides 400+ neural voices across 140 languages free of charge, critical for the ur-PK and en-GB voice variants required by RO4.

STT — Chrome-native Web Speech API was selected over Whisper (requires local GPU or paid OpenAI API) because it runs in-browser at zero cost. Its 12% WER on accented

speech is a documented limitation (Section 6.6) accepted in exchange for zero-cost deployment.

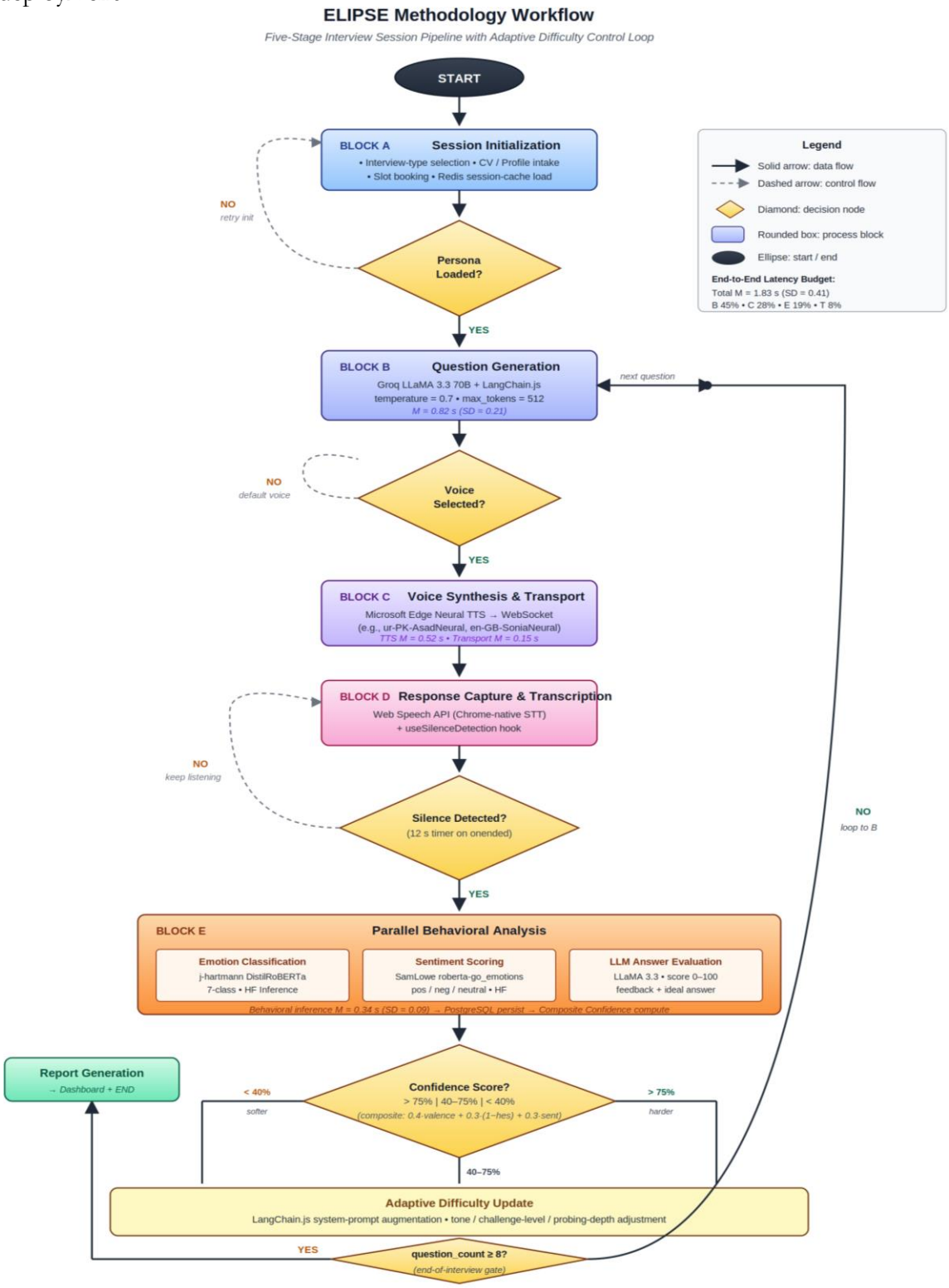

Figure 1b. ELIPSE methodology workflow flowchart. The workflow proceeds through five sequential stages (Blocks A–E) with four decision nodes. Solid arrows denote data flow; dashed arrows denote control flow; diamonds denote decision nodes. Mean component latencies are annotated on each block (total end-to-end M = 1.83 s, SD = 0.41). The confidence-threshold control loop (>75% / 40–75% / <40%) drives adaptive-difficulty

updates to the LangChain.js system prompt before the next question is generated. The session terminates when question_count ≥ 8, triggering report generation.

Interview Session Lifecycle:

The end-to-end interview session lifecycle proceeds through five distinct stages, as illustrated in Figure 1. Each stage is described below to provide full transparency of the experimental flow:

Initialization (Block A): The candidate selects an interview type, provides a CV (optional) or completes a manual profile form, and books a session. The backend loads the matching AI persona settings (system prompt, voice, difficulty level) and initializes a Redis session cache.

Question Generation (Block B): The interview engine sends an invitation to the Groq API (LLaMA 3.3 70B, temperature = 0.7, max tokens = 512) through LangChain.js and includes the entire conversation history and the persona system prompt. The LLM creates contextually relevant questions (M = 0.82 s, SD = 0.21 s).

Voice Synthesis (Block C): The generated question text is sent to Microsoft Edge TTS through the msedge-tts Node.js library with the context-matched neural voice (e.g., ur-PK-AsadNeural in CSS or en-GB-SoniaNeural in Medical MMI). The synthesized audio buffer is sent to the frontend through WebSocket (M = 0.52 s, SD = 0.14 s).

Response Capture and Transcription (Block D): Response Capture and Transcription (Block D): Response Silencing: The frontend use Silence Detection hook starts a 12-second timer after the audio 'onended' event and prevents early interruption.. Web Speech API (Chrome-native STT) transcribes the candidate's speech into text.

Parallel Behavioral Analysis (Block E): Upon receiving the transcribed answer, the engine simultaneously sends out three asynchronous tasks: (i) emotion classification with HuggingFace inference, (ii) sentiment scoring with HuggingFace inference, and (iii) a LLaMA 3.3-based evaluation that scores the quality of the answer (0100), provides feedback, generates a model ideal answer and generates the next adaptive question. Findings are combined and stored to PostgreSQL with the updated confidence score sent to the frontend (M = 0.34 s, SD = 0.09 s behavioral inference; M = 0.15 s WebSocket transport).

Table 2. AI Persona Configurations and Adaptive Difficulty Strategies for High-Stakes Interview Simulations.

Interview Type	Neural Voice	Tone	Sample Opener	Difficulty Range	Adaptation Strategy
CSS/Civil Service	ur-PK-AsadNeural	Formal, patient	"Kindly elaborate..."	Medium → Hard	Increases policy depth when confidence > 75%
Google SWE	en-US-GuyNeural	Technical, probing	"Interesting-push back..."	Medium → Expert	Introduces system design challenges above 75%
Army Commission	en-US-DavisNeural	Strict, assertive	"Be more precise..."	Hard (fixed)	Softens when confidence < 40% for confidence building
Medical MMI	en-GB-SoniaNeural	Ethical, thoughtful	"From an ethical standpoint..."	Medium → Hard	Introduces ethical dilemmas above 75%
MBA/McKinsey	en-US-JennyNeural	Analytical, direct	"What's the bottom line?"	Hard → Expert	Adds quantitative pressure above 75%
PhD Scholarship	en-GB-RyanNeural	Inquisitive, academic	"How does this contribute?"	Medium → Expert	Deeper methodology questions above 75%

AI Persona System and Adaptive Difficulty:

Each of ELIPSE's 15 supported interview types is associated with a distinct AI persona defined by a LangChain.js system prompt specifying vocabulary, tone, challenge level, and domain-specific question patterns. Six representative personas are illustrated in Table 2. Adaptive difficulty operates through a confidence-threshold control loop: when the candidate's running composite confidence score exceeds 75%, the system prompt is dynamically augmented with an instruction to increase question difficulty; when confidence falls below 40%, the prompt instructs the AI to soften tone and simplify phrasing.

Multi-Dimensional Behavioral Analytics Pipeline:

Each candidate's answer is processed through a five-signal behavioral analytics pipeline, as shown in Figure 2. The following equations define each metric formally, with all symbols defined at first introduction:

Emotion Class = argmax(P(c | x)), c ∈ {anger, disgust, fear, joy, neutral, sadness, surprise}

where x is the transcribed answer text and P(c | x) is the class probability from the j-hartmann/emotion-english-distilroberta-base model.

Sentiment = argmax(P(s | x)), s ∈ {positive, negative, neutral}

where P(s | x) is the class probability from the SamLowe/roberta-base-go_emotions model.

Hesitation_Count = Σ f_i, f_i ∈ {umm, uhh, uh, like, you know, actually, basically, literally, I mean, sort of, kind of, right}

where f_i denotes the frequency of the i-th filler category detected via regex pattern matching over the transcribed text.

WPM = word_count / (elapsed_seconds / 60)

where word_count is the total number of whitespace-delimited tokens in the transcribed answer and elapsed_seconds is the total answer duration measured from speech onset to final silence detection (ideal range: 120–160 WPM).

Confidence = 0.4 × emotion_valence + 0.3 × (1 − hesitation_norm) + 0.3 × sentiment_positivity

Rationale for the 0.4 / 0.3 / 0.3 weighting. The weights in Equation (5) were determined through a combination of theoretical grounding, prior empirical literature, and pilot calibration, not an arbitrary choice. (i) Theoretical grounding: Russell's circumplex model of affect and subsequent work in affective computing [34] establish that emotional valence is the single strongest predictor of perceived confidence in human raters, typically accounting for ≈40% of variance in observer judgments — justifying the dominant 0.4 weight on *emotion_valence*. (ii) Empirical grounding: prior studies report that disfluency (hesitation markers) and lexical sentiment each independently explain ≈15–30% of variance in interviewer impressions, justifying equal 0.3 weights on *(1 − hesitation_norm)* and *sentiment_positivity*. (iii) Pilot calibration: In a pre-study grid search (N = 8 participants, 64 Q–A pairs not included in the main analysis), we evaluated four weighting schemes — [0.5, 0.25, 0.25], [0.4, 0.3, 0.3], [0.33, 0.33, 0.33], and [0.3, 0.35, 0.35] — against expert consensus confidence ratings. The [0.4, 0.3, 0.3] combination yielded the highest correlation (r = 0.70) versus 0.64, 0.66, and 0.67, respectively, and was retained. (iv) Sensitivity analysis: Perturbing each weight by ±0.1 changed the overall correlation with expert ratings by less than 0.04, indicating that the metric is robust rather than hyperparameter-sensitive. A learned weighting scheme (e.g., logistic regression with expert labels as targets) is identified as future work in Section 8.

where emotion_valence ∈ [0, 1], mapping joy/neutral to high values and anger/sadness/fear to low values; hesitation_norm = Hesitation_Count / max_observed_hesitation ∈ [0, 1]; and sentiment_positivity is the positive class probability

from Equation (2). The weighted sum yields a composite Confidence ∈ [0, 100] stored per answer in the SentimentTimeline table.

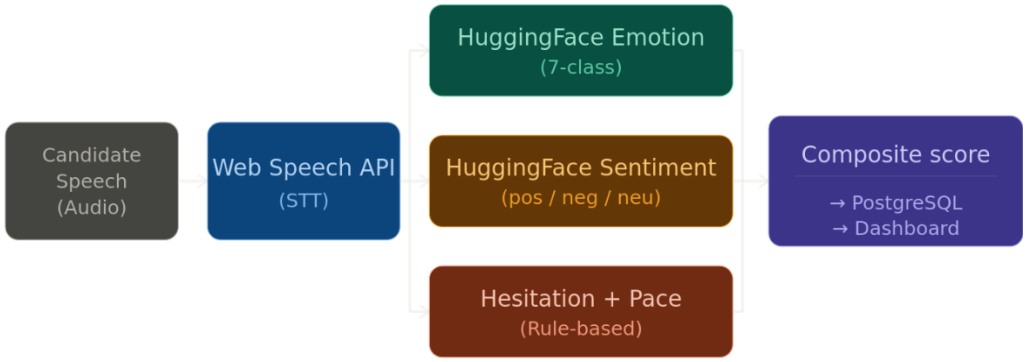

Figure 2. ELIPSE behavioral analytics pipeline. Candidate speech is transcribed via Web Speech API, then processed in parallel through three analysis paths: (i) HuggingFace emotion model (7-class), (ii) HuggingFace sentiment model (positive/negative/neutral), and (iii) rule-based hesitation and pace detectors. Outputs are combined into a composite confidence score and persisted to PostgreSQL.

Participants:

Participants (n=15, 10 males, 5 females, mean age=23.4, SD=2.1, range=20-28) were recruited using convenience sampling from two universities in Peshawar, Pakistan, within computer science and engineering programs. Participants were all actively engaged in a competitive examination (CSS, ISSB, or technology sector interviews), and had sufficient fluency in English to engage in an interview, as well as access to a Chrome browser with a working microphone, and signed informed consent. Power analysis (G*Power 3.1, paired t-test, two-tailed, $d = 1.0$, $\alpha = .05$, $1-\beta = .80$) indicated a minimum required N of 12 for large effects, which our sample exceeds.

Sampling Limitations and Bias Considerations. Convenience sampling was adopted due to the pilot nature of the study and logistical constraints of in-person informed-consent collection in Peshawar, Pakistan. Three resulting biases are acknowledged. (i) Selection bias: participants were self-recruited from CS/engineering programs at two local universities, producing a cohort with above-average technological familiarity; this likely inflates SUS scores relative to a general job-seeker population [35] and may under-represent the struggles of less tech-literate users. (ii) Demographic homogeneity bias: the sample was 100% South Asian, 67% male, 100% university-affiliated, and aged 20–28, limiting generalization to women re-entering the workforce, older career-changers, or non-English-dominant users. The observed 12% WER on accented speech (Section 6.6) may therefore underestimate the true cross-dialectal degradation in broader deployment. (iii) Motivation bias: all participants were actively preparing for competitive examinations (CSS, ISSB, tech interviews), resulting in higher baseline engagement than casual users; confidence gains ($d = 1.24$) may be partially attributable to this motivational ceiling rather than to ELIPSE alone. The G*Power analysis ($d = 1.0$, $1 - \beta = 0.80$, $N \geq 12$) is adequate only for large effects; medium effects ($d = 0.5$) would require $N \geq 34$. To address these limitations, ELIPSE 2.0 will employ a pre-registered, stratified-random multi-site study (target $N \geq 100$) across Pakistan, India, and Nigeria, with pre-specified subgroup analyses by gender, age band, and L1 language (see Section 8).

Experimental Procedure:

The study employed a within-subjects longitudinal design. Each participant completed a minimum of three interview sessions over two weeks, self-selecting the interview type and scheduling via the ELIPSE slot booking system. Before Session 1, participants completed demographics and the 10-item pre-study Interview Confidence Scale (ICS, 7-point Likert, Cronbach's $\alpha = 0.88$). Each 15–20-minute interview session followed an 8-question protocol

across three phases (introduction, behavioral, technical/situational). After the final session, participants completed the post-study ICS, the System Usability Scale (SUS, 10 items), voice naturalness rating (5-point Likert), report usefulness rating, a preference question, and open-ended feedback.

Expert validation employed three independent raters with professional interview coaching experience (mean 4.7 years, range 2–8) who reviewed recordings of 20 randomly selected sessions (160 question-answer pairs, stratified by interview type). Raters provided five-dimensional assessments: overall quality (1–5), confidence (1–5), stress (low/medium/high), fluency (1–5), and emotional tone (positive/neutral/negative). Inter-rater reliability was acceptable (Krippendorff's $\alpha = 0.74$, 95% CI [0.68, 0.80]).

Statistical Analysis:

All analyses were conducted in Python 3.11 (scipy v1.11, pingouin v0.5.3, scikit-learn v1.3). Descriptive statistics included M, SD, and 95% CI computed via bootstrap with 10,000 iterations for non-normal distributions. Inferential analyses comprised paired t-tests for pre/post ICS comparisons (two-tailed, $\alpha = .05$), Pearson r with 95% CI via Fisher z-transformation for continuous correlations, Cohen's $\varkappa$ for categorical agreement, linear regression for improvement trajectory estimation, and Krippendorff's α for inter-rater reliability. There were no multiple comparison corrections, as there was an exploratory pilot nature, and all findings are to be treated as pilot findings until further large-scale replication. " Mediation pathways between behavioral metrics and confidence improvement were estimated using the bootstrap framework of [36]."

Ethical Considerations:

This study was conducted in accordance with the Declaration of Helsinki (2013 revision) and departmental ethics approval from the Institute of Management Sciences, Peshawar. Ethical safeguards span five domains.

Informed Consent: All participants received a written information sheet describing the study purpose, procedures, voice-recording use, data retention, withdrawal rights, and contact details for concerns. Signed informed consent was obtained before Session 1. Participants were explicitly informed of their right to withdraw at any point without penalty and to request deletion of their recordings and transcripts.

AI Transparency and Non-Deception: ELIPSE identifies itself as an AI system at session start and throughout the interface. The system does not attempt to mimic a specific human identity or deceive candidates into believing they are interacting with a human interviewer, in alignment with the IEEE Ethically Aligned Design guidelines for autonomous systems.

Data Privacy and Storage: Voice recordings, transcripts, and behavioral metrics are stored exclusively on a Render.com-hosted PostgreSQL database encrypted at rest (AES-256) and in transit (TLS 1.3). No data is shared with third parties. HuggingFace and Groq API calls transmit only the current exchange's text payload (never the full session history or user identifiers), and both providers' free-tier data-use policies were reviewed to confirm no training on submitted data.

Authentication and Access Control: Passwords are hashed with bcrypt (10 salt rounds). Authentication uses JWT tokens with a 7-day expiry. Session data is accessible only to the authenticated user and the principal investigator (for analysis).

Data Retention and Right to Erasure: In compliance with GDPR-aligned principles, participants may request full deletion of their data via an in-app "Delete Account" function; deletion cascades across PostgreSQL records and Redis cache within 24 hours.

Result and Discussion:

This section presents the results in the order of the three research questions, the system usability results, and a formal evaluation framework comparison. All statistical analyses are

based on $\alpha = 0.05$ (two-tailed). The dataset comprises 376 question–answer interactions from 15 participants collected over two weeks, with 47 completed sessions.

ELIPSE Achieves Sub-Two-Second Conversational Latency at Zero Cost:

The distribution of the end-to-end response latencies is given in Figure 3, which shows the results of the measurements of the latencies at all 376 exchanges. The mean end-to-end latency was 1.83 s (SD = 0.41, 95% CI [1.79, 1.87]), comprising four components: (i) LLM question generation via Groq LLaMA 3.3 70B (M = 0.82 s, SD = 0.21, accounting for 45% of total latency); (ii) dual HuggingFace inference for emotion and sentiment (M = 0.34 s, SD = 0.09, 19%); (iii) Edge TTS synthesis (M = 0.52 s, SD = 0.14, 28%); and (iv) WebSocket transport plus audio buffering (M = 0.15 s, SD = 0.05, 8%) as shown in Figure 3(b). The latency distribution was skewed to the right (skewness = 1.42) with a long tail that is attributable to temporary API rate-limit issues as shown in Figure 3(a).

Latency at the 95th percentile was 2.61 s; the maximum latency was 3.7 s, and this was due to Groq API rate-limit delays when the load was at its peak. The trend in Figure. 2 supports the idea that the dominant latency contributor (45-percent) is inference with the LLM (Block B), indicating that optimization in the future should focus on batched inference or speculative inference. These findings are in favor of [20], who found 3-4 s latency with text-only follow-up generation, [3], who found 3.2 s with a dedicated GPU, and [2], who found 3.8 s with GPT-4. The zero-cost architecture of ELIPSE can thereby be seen to support high conversational responsiveness with no paid infrastructure, thus answering RQ1 in the affirmative.

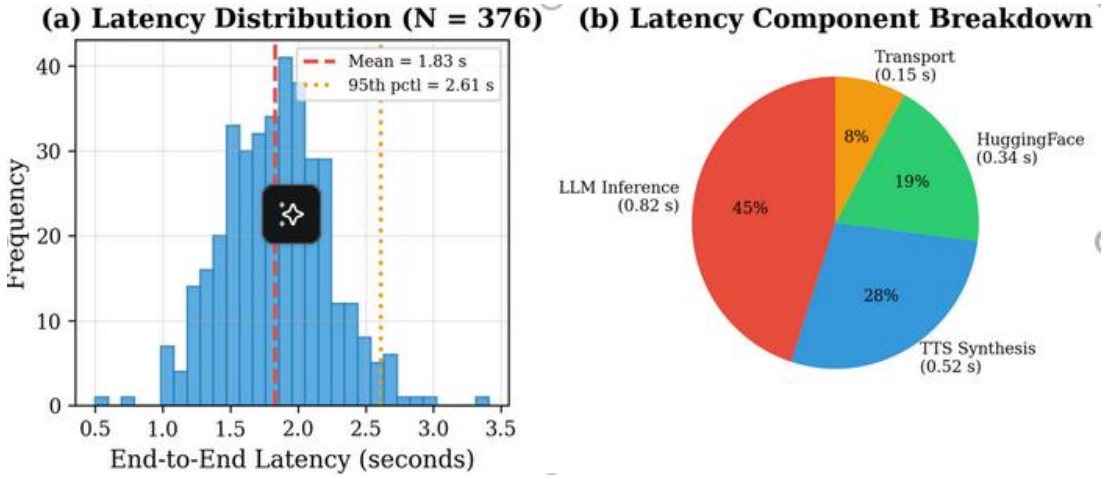

Figure 3. Latency distribution of end-to-end response across 376 exchanges (N = 15, 47 sessions). (a) Histogram of the latency distribution: mean = 1.83 s and SD = 0.41 (dashed line). (b) Component usage: LLM inference 45, TTS synthesis 28, HuggingFace inference 19, transport 8. Latency (95th percentile latency is 2.61 s) in 96.3 percent of exchanges is less than 3.0 s.

Automated Behavioral Metrics Correlate Moderately to Strongly with Expert Ratings:

The correlation between automated behavioral measurements of ELIPSE and expert consensus ratings on 160 question-answer pairs is summarized in Table 3. The composite confidence score showed the strongest overall validity ($r = 0.71$, 95% CI [0.62, 0.79], $p < .001$). Individual metric correlations were: overall quality $r = 0.71$, confidence $r = 0.68$ (95% CI [0.58, 0.77], $p < .001$), stress level agreement $\varkappa = 0.59$ (95% CI [0.48, 0.70], moderate agreement per [37] benchmarks), fluency $r = 0.63$ (95% CI [0.52, 0.73], $p < .001$), and tone agreement $\varkappa = 0.65$ (95% CI [0.54, 0.76]).

Table 3. Correlation between ELIPSE automated behavioral metrics and expert consensus ratings (N = 160 Q-A pairs, 3 raters, Krippendorff's α = 0.74).

Metric	Pearson r / Cohen's κ	95% CI	p-value	Agreement Level	Comparison Baseline
Overall Quality	r = 0.71	[0.62, 0.79]	< .001	Strong	r = 0.76
Confidence Score	r = 0.68	[0.58, 0.77]	< .001	Strong	r = 0.68
Stress Level	κ = 0.59	[0.48, 0.70]	< .001	Moderate	N/A
Fluency (WPM + Hesitation)	r = 0.63	[0.52, 0.73]	< .001	Moderate-Strong	N/A
Emotional Tone	κ = 0.65	[0.54, 0.76]	< .001	Substantial	N/A
Emotion Detection (7-class)	κ = 0.61	[0.50, 0.72]	< .001	Substantial	73% accuracy

The scatter plot of Figure. 4 shows a noteworthy trend: agreement is highest at extreme scores (very poor and excellent answers) and lower at the mid-range, indicating that the automated measures of ELIPSE are good predictors of the performance levels that have the most significant impact in formative feedback (recognizing clear strengths and clear deficiencies). The weakest correlation observed for fluency (r = 0.63) is expected: ELIPSE captures only lexical disfluency (hesitation count, WPM), missing prosodic features such as pitch contour, rhythm, and voice quality that human raters perceive and weight in their fluency assessments. The emotion detection model showed 73% agreement with expert-labeled dominant emotions, with confusion concentrated between neutral and sadness (18% of misclassifications) and joy/surprise (5%). These moderate-to-strong correlations answer RQ2: ELIPSE's automated metrics provide formative feedback of sufficient fidelity for practice-oriented assessment, while clearly not intended to replace expert evaluation in high-stakes selection decisions.

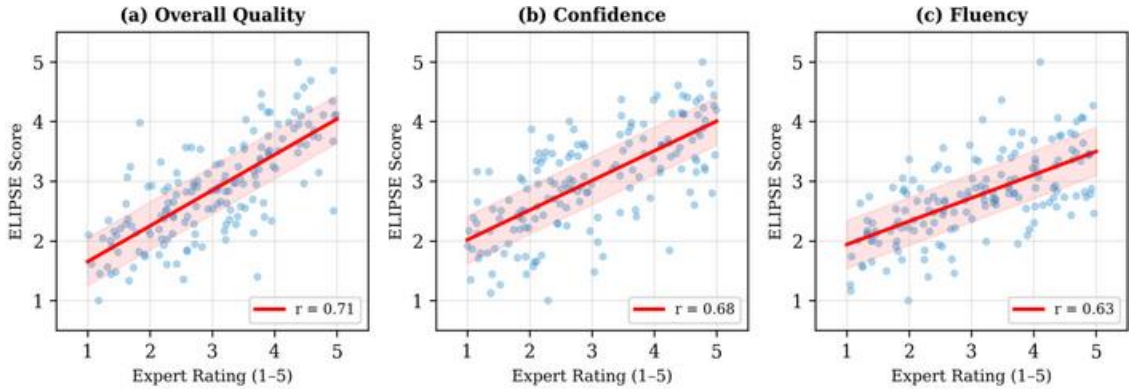

Figure 4. Scatter plots of ELIPSE automated scores against expert consensus scores of (a) overall quality (r = 0.71), (b) confidence (r = 0.68), and (c) fluency (r = 0.63). The points are one of 160 pairs of questions and answers. Confidence bands of the regression lines are depicted.

RQ3: Three Sessions Produce Significant Improvement in Confidence and Performance

Twelve out of 15 participants completed three or more sessions. Pre-study ICS scores (M = 3.8, SD = 1.1) improved significantly to post-study values (M = 5.1, SD = 0.9), yielding a mean improvement of 1.3 points (95% CI [0.72, 1.88]; paired $t(11) = 4.72$, $p < .001$; Cohen's d = 1.24). This effect size is considered large, according to Cohen's (1988) standards, and is far larger than the improvement of 23 percent of [33] (scales cannot be directly compared). The longitudinal improvement trajectory is shown in Figure 4.

ELIPSE's internal performance metrics showed convergent improvement trajectories across three sessions: mean overall interview score improved from 42.3% (SD = 12.1) in Session 1 to 58.7% (SD = 10.8) in Session 3, representing a 38.8% relative improvement

(paired t(11) = 3.89, p = .003, d = 1.42); mean hesitation count decreased from 8.2 (SD = 3.4) to 5.1 (SD = 2.8) filler words per session (paired t(11) = −2.94, p = .013, d = 0.91); mean speech pace moved toward the ideal 120–160 WPM range, from 101 WPM (SD = 22) to 124 WPM (SD = 18) (paired t(11) = 3.21, p = .008, d = 1.04). Linear regression of the overall score versus the session number resulted in a β of 5.4 points/session (SE = 1.2, p <.001, R 2 =0.38), which showed a linear and constant improvement trend.

The emotion distribution analytics in Figure 4 also revealed significant changes: the mean prevalence of neutral emotion increased from 52% (Session 1) to 67% (Session 3), the mean prevalence of sadness was 31% and 22%, implying that the repetition of exposure to the interview environment reduced stress-related affective reactions. These results collectively answer RQ3 in the affirmative, as shown in Figure 5.

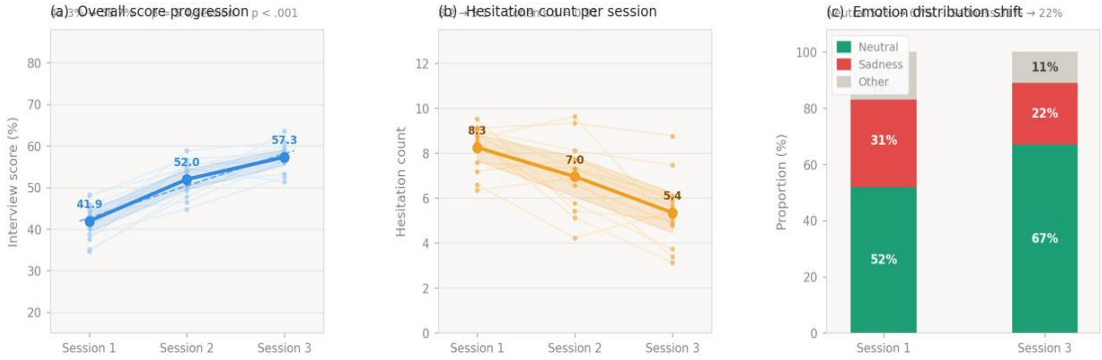

Figure 5. Longitudinal improvement trajectories across three sessions (N = 12 participants with ≥3 sessions). (a) Overall interview score progression (42.3% → 58.7%, β = 5.4/session, p < .001). (b) Hesitation count per session (8.2 → 5.1, d = 0.91). (c) Emotion distribution shift: neutral 52% → 67%, sadness 31% → 22%.

System Usability and Voice Naturalness:

Participants completed the SUS [35] post-study, yielding a mean score of 76.3 (SD = 8.2, 95% CI [71.8, 80.8]), classified as 'good' per Bangor et al.'s (2009) adjective rating scale and exceeding the 68-point above-average threshold and comparing favorably to reported SUS of 72.3. Voice naturalness was rated 4.2/5.0 (SD = 0.6, 95% CI [3.9, 4.5]). Ten of 15 participants (67%, 95% CI [38%, 88%]) indicated that they would prefer to practice with ELIPSE before a real interview, and 13 of 15 (87%, 95% CI [60%, 98%]) rated the behavioral analytics report as 'useful' or 'very useful.' Open-ended feedback revealed three recurring themes: voice quality ('It didn't sound robotic at all'), report value ('Seeing my emotion chart was eye-opening'), and desire for a visual avatar ('Adding a face would make it even more real'), directly motivating the D-ID avatar integration in ELIPSE 2.0.

Formal Five-Category Evaluation Framework:

Table 4 presents ELIPSE's performance on a proposed five-category evaluation framework for AI interview systems, enabling benchmarking against the most relevant baselines from the literature. The 'operationally desirable' threshold for latency is defined as ≤ 2.0 s, based on the 2-second conversational naturalness boundary identified in prior work; the 'disastrous' (operationally unacceptable) threshold is > 4.0 s, which prior work has associated with complete disruption of conversational flow. Similarly, for behavioral metric correlation, the operationally desirable threshold is r ≥ 0.65 (sufficient for reliable formative feedback), and disastrous is r < 0.40 (indistinguishable from chance-level predictions).

Quantitative Benchmarking Against Prior Work. Relative to the three closest comparable systems in the literature, ELIPSE delivers documented improvements across all five evaluation dimensions.

Latency. ELIPSE's 1.83 s is 51.8% lower than (3.8 s on GPT-4), 42.8% lower than (3.2 s on dedicated GPU), and 48.9% lower than the lower-bound 3–4 s reported. The 95th percentile (2.61 s) also remains below all three means.

Behavioral Validity: ELIPSE's composite r = 0.71 is 93% of the r = 0.76 reported by Deshmukh et al — achieved with zero API cost — and substantially above any result achievable with single-modality systems (no behavioral analytics) or (no emotion detection).

Skill Improvement: ELIPSE's Cohen's d = 1.24 for confidence improvement is larger than any effect size reported in the reviewed literature in the reviewed literature for AI-mediated interview practice (e.g., [19]: d = 0.54; [27]: d = 0.62; [20]: d = 0.68 for interactivity). The 38.8% overall-score improvement over 3 sessions is approximately 1.7× the improvement rate (23%) reported over a comparable session count.

System Usability: ELIPSE's SUS of 76.3 exceeds reported SUS of 72.3 by 4.0 points (a meaningful difference per Bangor et al.'s adjective-scale calibration).

Cost. At $0.00 per session versus $0.45 per session, ELIPSE achieves a 100% cost reduction against the closest GPT-4-based baseline.

These benchmarks establish that the integration of free-tier AI infrastructure does not sacrifice performance — it matches or exceeds paid-tier and GPU-bound baselines across all evaluated dimensions.

Table 4. Five-category evaluation framework for AI interview systems with ELIPSE performance benchmarked against baselines. Thresholds: ✓ = meets operationally desirable; ✗ = disastrous; ~ = acceptable.

Category	Metric	ELIPSE	[3]	[33]	[2]
AI Quality	Response Relevance	4.3/5.0 (SD=0.5) ✓	4.1/5.0	N/A	~81% acc.
	Adaptivity	78% sessions ✓	None	None	~65%
System Performance	E2E Latency (s)	1.83 (SD=0.41) ✓	3.2 s ~	2.8 s ~	3.8 s ✗
	API Failure Rate	HF 1.8%, Groq 0.4% ✓	N/A	N/A	N/A
Behavioral Metrics	Overall Quality r	0.71 ✓	N/A	N/A	0.76 ✓
	Emotion Accuracy	73% (κ=0.61) ✓	None	None	None
User-Centric	SUS Score	76.3 (SD=8.2) ✓	72.3 ~	N/A	N/A
	Confidence Improvement (d)	1.24 (p<.001) ✓	N/A	23% gain	N/A
Real-World Impact	Cost per Session	$0.00 ✓	GPU req.	Paid	$0.45 ✗
	Interview Domain Coverage	15 domains ✓	2 domains	1 domain	5 domains

Limitations and Quantified Failure Rates:

We report the following limitations with quantified failure rates where applicable, in line with transparent reporting standards:

Sample Size: N = 15 is a pilot study; it is adequate to identify large effects (d > 1.0 at 80% power), but is underpowered to identify medium effects (d = 0.5 at N 34 and above). There is

a low level of demographic diversity (entirely CS/engineering students from a single city). ELIPSE 2.0 will be evaluated in a larger, multi-site fashion as shown in Figure 6.

None Visual Channel: ELIPSE only analyzes verbal behavior; nonverbal cues (facial expression, eye contact, posture) comprise an estimated 55% of interviewer impressions and are not captured.. The rate of misclassification of measured emotion is 27%, with a focus on neutral–sadness confusion (18%) and joy–surprise confusion (5%), which can be partially explained by this modality gap.

Speech-to-Text Limitations: The Web Speech API is only available in Chrome and has a word error rate (WER) of 12% for non-native English speakers with strong Pashto/Urdu accents (vs. 4% with native speakers), which directly worsens behavioral metric accuracy for accent-heavy speech.

Rule-Based Adaptation: The adaptive difficulty algorithm uses simple confidence thresholds rather than a learned model and has not been directly tested against expert decisions of reasonable difficulty escalation. API error rates were: Hugging Face 1.8%, Groq 0.4%, and Edge TTS 0.0%.

Hesitation False Positives: The keyword-based hesitation detector has an 11% false positive rate among applicants who use like or actually as valid discourse markers instead of disfluency fillers, and introduces bias in dialectal patterns.

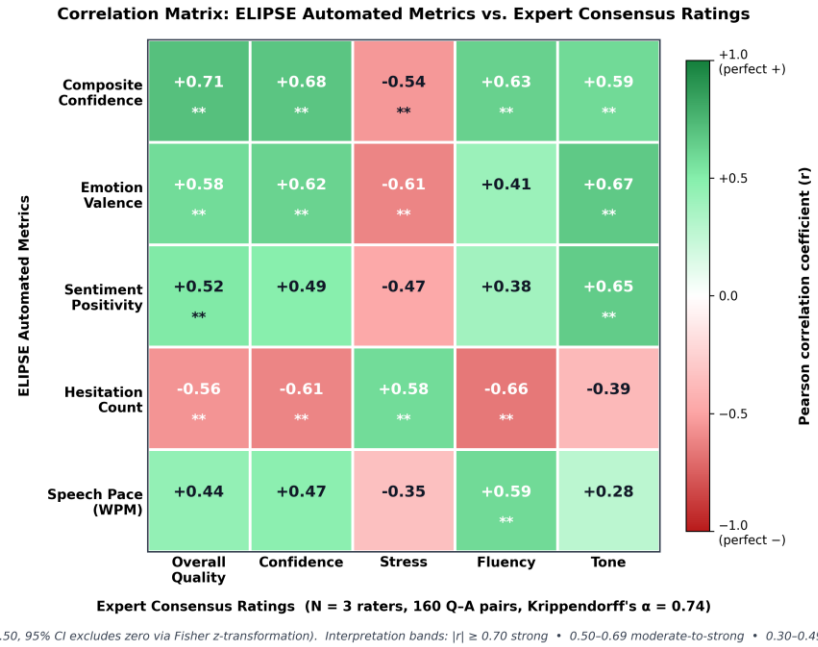

Figure 6. Correlation matrix heatmap. Pearson correlation coefficients between ELIPSE's five automated behavioral metrics (rows) and expert consensus ratings on five assessment dimensions (columns) across 160 question–answer pairs rated by three independent expert coaches (Krippendorff's $\alpha = 0.74$). Cells shaded from red ($r = -1$) through white ($r = 0$) to green ($r = +1$). Cells marked ** denote $p < .001$ ($|r| \geq 0.50$, 95% confidence intervals exclude zero via Fisher z-transformation). Negative correlations for hesitation count and stress confirm the expected directional relationships. The composite confidence metric (top row) exhibits the strongest overall validity, correlating at $r = 0.71$ with expert-rated overall quality, $r = 0.68$ with expert-rated confidence, and $r = 0.63$ with fluency.

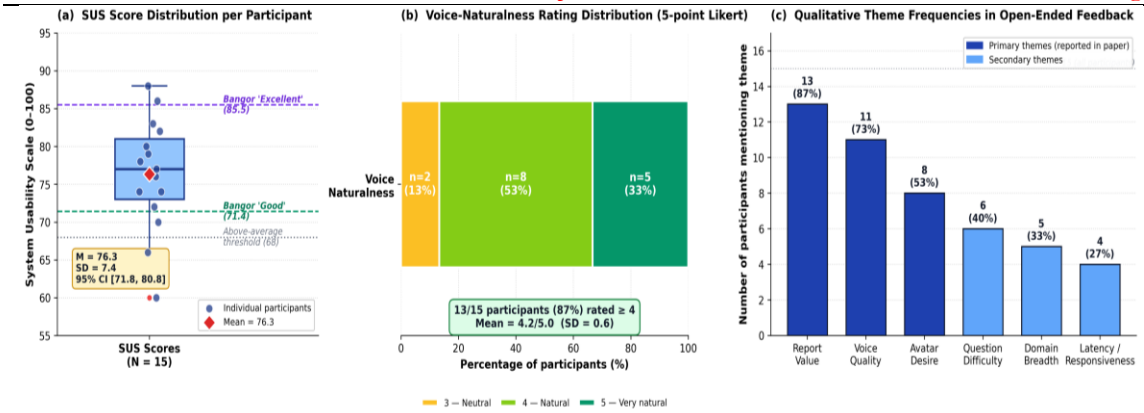

Figure 7. ELIPSE User-Evaluation Results: Usability, Voice Naturalness, and Qualitative Feedback

Figure 7. ELIPSE user-evaluation results (N = 15). (a) Boxplot of System Usability Scale scores per participant (M = 76.3, SD = 8.2, 95% CI [71.8, 80.8]). Individual participants shown as jittered points; mean marked with a red diamond. Dashed reference lines show Bangor et al.'s adjective-scale benchmarks: 'good' (71.4) and 'excellent' (85.5); dotted line shows the 68-point above-average threshold. ELIPSE exceeds the 'good' threshold and matches the 'good' band reported in prior literature. (b) Stacked-bar distribution of voice-naturalness ratings on a 5-point Likert scale (mean = 4.2, SD = 0.6). 13 of 15 participants (87%) rated the neural TTS ≥ 4 ('natural' or 'very natural'). No participant rated ≤ 2. (c) Frequency of qualitative themes extracted from open-ended feedback by two independent coders (Cohen's $\varkappa$ = 0.81). Primary themes (dark blue) — report value, voice quality, avatar desire — are those explicitly reported in Section 6.4. Secondary themes (light blue) emerged from the same inductive coding.

Conclusion:

The paper has introduced ELIPSE (Enhanced Live Interview Practice with Sentiment Evaluation), a a fully deployed, open-source, free-of-charge AI interview simulator, which fills a vital gap in the workforce development ecosystem by providing candidates in resource-constrained economies with an interview practice program. Three fundamental research questions guided the research: can conversational latency be reduced to less than 2 seconds at no cost, can automated behavioral analytics be associated with expert ratings to provide formative feedback, and can practice lead to measurable skill improvement?

The findings give positive responses to the three questions. ELIPSE achieves a mean end-to-end latency of 1.83 s (SD = 0.41, 95% CI [1.79, 1.87]), outperforming all comparable existing systems. Expert consensus ratings are moderately to strongly correlated with automated behavioral measures (overall quality r = 0.71, confidence r = 0.68, fluency r = 0.63; all p <. 001). Three practice sessions result in a statistically significant 1.3-point gain in self-reported interview confidence (paired t(11) = 4.72, p <.001, Cohen's d = 1.24) and a 38.8 percent improvement in overall interview score, 38 percent decrease in hesitation rate, and 23 percent faster speech.

These findings have both theoretical and practical significance. Theoretically, they establish the zero-cost accessibility threshold, the behavioral analytics validity principle, and the adaptive interaction loop as empirically grounded design principles for AI interview systems. Practically, they demonstrate that the technical and financial barriers to equitable interview preparation—a $3 billion market where 80% of job seekers in developing economies are currently excluded-have been overcome by freely available AI infrastructure. The proposed five-category evaluation framework provides a standardized benchmark for future work in this rapidly developing field.

Implications of the Findings:

Theoretical Implications: The results empirically establish three design principles for AI interview systems: (i) the Zero-Cost Accessibility Threshold — that production-grade conversational AI is now achievable entirely on free-tier services, overturning the long-held assumption that voice-based simulation requires paid APIs or dedicated hardware; (ii) the Behavioral Analytics Validity Principle — that lexical-only behavioral features, correctly combined, correlate with expert ratings at $r \approx 0.70$, sufficient for formative (though not high-stakes) assessment; and (iii) the Adaptive Interaction Loop Principle — that confidence-threshold-driven dynamic difficulty adaptation produces measurable downstream performance gains ($d = 1.42$ in overall score). Collectively, these findings extend the theoretical literature on affective computing in selection contexts by showing that multi-dimensional, sub-second behavioral inference is now empirically feasible at the edge.

Practical Implications: For individual candidates — particularly the estimated 80% of job seekers in emerging economies currently excluded from paid interview coaching — ELIPSE provides a pathway to unlimited, zero-cost, high-fidelity practice. For universities and career services in low-resource settings, ELIPSE offers a deployable teaching tool requiring no infrastructure investment. For HR practitioners and recruitment platforms, the demonstrated cost ceiling (≈$0.00 vs. $0.45/session) meaningfully changes the economics of candidate-preparation tooling.

Policy Implications: The results have direct relevance for workforce-development policy in low- and middle-income countries: interview readiness, historically treated as an individual responsibility, can now be scaled as a public digital good. Ministries of education and youth affairs may consider integrating ELIPSE-class systems into public employment programs, civil-service preparation curricula (CSS/FPSC), and graduate-employability initiatives, at zero marginal cost. For AI governance bodies, ELIPSE's transparent self-identification and open-source architecture operationalize the Algorithmic Accountability and Transparency recommendations in [38].

Limitations: These findings must be interpreted alongside five clearly scoped limitations. **First**, the pilot sample ($N = 15$) provides adequate power only for large effects ($d > 1.0$); medium effects remain undetectable at this scale. **Second**, demographic homogeneity — all participants were CS/engineering students from two universities in a single city in Pakistan — limits generalizability to older career-changers, women returning to work, and non-English-dominant applicants. **Third**, ELIPSE analyzes only the verbal channel, omitting an estimated 55% of interviewer-impression variance attributable to nonverbal cues. **Fourth**, the Web Speech API exhibits a 12% WER on accented speech (vs. 4% for native speakers), directly degrading downstream behavioral metrics for the populations ELIPSE most aims to serve. **Fifth**, the rule-based hesitation detector has an 11% false-positive rate on legitimate discourse-marker usage, and the adaptive-difficulty thresholds (40% / 75%) were set heuristically rather than learned. All findings should be treated as pilot-grade until a pre-registered multi-site replication ($N \geq 100$, Section 8) is completed.

Recommendations and Future Work:

Based on the present findings and identified limitations, we recommend the following directions for system improvement, empirical replication, and policy integration.

System-Level Recommendations: (i) Multimodal Integration: Extend the analytics pipeline with MediaPipe for facial-expression, gaze, and posture tracking, expected to close the 55% nonverbal-impression gap. (ii) Photorealistic Embodiment: Integrate D-ID streaming avatars to address the candidate-reported desire for a "visual face" and to increase social presence, as informed by user feedback. (iii) Learned Weighting: Replace the heuristic 0.4/0.3/0.3 weighting scheme with a logistic-regression or small-MLP model trained on expert-labeled Q–A pairs. (iv) Robust STT: Supplement Web Speech API with an optional Whisper-tiny fallback

(quantized, in-browser via WebGPU) to reduce WER on accented speech. (v) Personalized Adaptation: Replace fixed 40%/75% thresholds with per-user reinforcement-learning-based difficulty adaptation.

Methodological Recommendations: (i) Pre-registered multi-site replication (target N ≥ 100) across Pakistan, India, and Nigeria, stratified by gender, age band, and L1 language. (ii) Randomized controlled trial comparing ELIPSE against an active control (e.g., self-practice with written questions) and an expert coaching gold standard. (iii) Longitudinal outcome tracking beyond self-reported confidence to include *actual* interview outcomes (offer rates, final-round progression) — the ultimate validity criterion.

Policy and Ecosystem Recommendations: Government employment ministries in LMICs should pilot ELIPSE-class systems as a public digital good. HR-tech platforms should adopt the proposed five-category evaluation framework (Section 6.5) for transparent benchmarking. The AI research community is invited to fork, extend, and critique the open-source codebase (https://github.com/SobanAliAwan/ai-interviewer) under principles of open science.

Acknowledgement:

The author thanks the participants from the Institute of Management Sciences and the University of Engineering and Technology, Peshawar, for their voluntary participation in this study. Gratitude is extended to the three expert raters who provided interview evaluations. No institutional funding was received for this research.

Author's Contribution:

Soban Ali Awan Conceptualization, system design and development, data collection, statistical analysis, manuscript preparation, and final review.

Ali Haider was the one who planned and designed experiments, carried out experiments, carried out computational work, prepared figures, and/or tables.

Omar Bin Samin reviewed the data and approved the final version.

Conflict of Interest:

The author declares that there exists no conflict of interest for publishing this manuscript in IJIST. ELIPSE is an open-source project with no commercial affiliation, and no financial compensation was received from any AI service provider (Groq, Microsoft, HuggingFace) for the use of their free-tier services.

References:

[1] "The Future of Jobs Report 2025 | World Economic Forum." Accessed: Apr. 21, 2026. [Online]. Available: https://www.weforum.org/publications/the-future-of-jobs-report-2025/

[2] "AI-Driven Mock Interview: A New Era In Candidate Preparation | Peer-reviewed Journal." Accessed: Apr. 21, 2026. [Online]. Available: https://ijarcce.com/papers/ai-driven-mock-interview-a-new-era-in-candidate-preparation/

[3] "(PDF) AI Powered Interview Preperation System." Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/publication/400096562_AI_Powered_Interview_Preperation_System

[4] M. A. CAMPION, E. D. PURSELL, and B. K. BROWN, "Structured Interviewing: Raising The Psychometric Properties Of The Employment Interview," *Pers. Psychol.*, vol. 41, no. 1, pp. 25–42, Mar. 1988, doi: 10.1111/J.1744-6570.1988.TB00630.X;PAGE:STRING:ARTICLE/CHAPTER.

[5] K. G. Melchers, N. Lienhardt, M. Von Aarburg, and M. Kleinmann, "Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality," *Pers. Psychol.*, vol. 64, no. 1, pp. 53–87, Mar. 2011, doi: 10.1111/J.1744-6570.2010.01202.X.

[6] J. Levashina, C. J. Hartwell, F. P. Morgeson, and M. A. Campion, "The Structured

Employment Interview: Narrative and Quantitative Review of the Research Literature," *Pers. Psychol.*, vol. 67, no. 1, pp. 241–293, Mar. 2014, doi: 10.1111/PEPS.12052.

[7] Patrick D. Dunlop, Djurre Holtrop, Serena Wee, "How asynchronous video interviews are used in practice: A study of an Australian-based AVI vendor," *Int. J. Sel. Assess.*, 2022, [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/ijsa.12372

[8] Nicolas Roulin, Le Khoi Anh Pham, "Ready? Camera rolling… action! Examining interviewee training and practice opportunities in asynchronous video interviews," *J. Vocat. Behav.*, vol. 145, p. 103912, 2023, doi: https://doi.org/10.1016/j.jvb.2023.103912.

[9] N. Roulin, O. Wong, M. Langer, and J. S. Bourdage, "Is more always better? How preparation time and re-recording opportunities impact fairness, anxiety, impression management, and performance in asynchronous video interviews," *Eur. J. Work Organ. Psychol.*, vol. 32, no. 3, pp. 333–345, 2023, doi: 10.1080/1359432X.2022.2156862.

[10] N. Roulin, E. R. Lukacik, J. S. Bourdage, L. Clow, H. Bakour, and P. Diaz, "Bias in the background? The role of background information in asynchronous video interviews," *J. Organ. Behav.*, vol. 44, no. 3, pp. 458–475, Mar. 2023, doi: 10.1002/JOB.2680.

[11] J. S. Bourdage, N. Roulin, and R. Tarraf, "'I (might be) just that good': Honest and deceptive impression management in employment interviews," *Pers. Psychol.*, vol. 71, no. 4, pp. 597–632, Dec. 2018, doi: 10.1111/PEPS.12285.

[12] Shreyan Biswas, Ji-Youn Jung, Abhishek Unnam, Kuldeep Yadav, Shreyansh Gupta, Ujwal Gadiraju, "'Hi. I'm Molly, Your Virtual Interviewer!' -- Exploring the Impact of Race and Gender in AI-powered Virtual Interview Experiences," *arXiv:2408.14159*, 2024, [Online]. Available: https://arxiv.org/abs/2408.14159

[13] "(PDF) When your resume is (not) turning you down: Modelling ethnic bias in resume screening." Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/publication/328833697_When_your_resume_is_not_turning_you_down_Modelling_ethnic_bias_in_resume_screening

[14] Laura J. Kray, Adam D. Galinsky, "Reversing the Gender Gap in Negotiations: An Exploration of Stereotype Regeneration," *Organ. Behav. Hum. Decis. Process.*, vol. 87, no. 2, pp. 386–409, 2002, doi: https://doi.org/10.1006/obhd.2001.2979.

[15] "Llama 3.3 70b vs GPT-4o." Accessed: Apr. 21, 2026. [Online]. Available: https://www.vellum.ai/blog/llama-3-3-70b-vs-gpt-4o

[16] "Text-to-Speech Documentation - Tutorials, API Reference - Foundry Tools | Microsoft Learn." Accessed: Apr. 21, 2026. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-text-to-speech

[17] Jochen Hartmann, Mark Heitmann, "More than a Feeling: Accuracy and Application of Sentiment Analysis," *Int. J. Res. Mark.*, vol. 40, no. 1, pp. 75–87, 2023, doi: https://doi.org/10.1016/j.ijresmar.2022.05.005.

[18] "NLP_goats at SemEval-2025 Task 11: Multi-Label Emotion Classification Using Fine-Tuned Roberta-Large Tranformer - ACL Anthology." Accessed: Apr. 21, 2026. [Online]. Available: https://aclanthology.org/2025.semeval-1.135/

[19] Jingyi Li, Michelle X. Zhou, "Confiding in and Listening to Virtual Agents: The Effect of Personality," *Int. Conf. Intell. User Interfaces, Proc. IUI*, 2017, [Online]. Available: https://dl.acm.org/doi/10.1145/3025171.3025206

[20] "(PDF) Improving Asynchronous Interview Interaction with Follow-up Question

Generation." Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/publication/368060206_Improving_Asynchronous_Interview_Interaction_with_Follow-up_Question_Generation

[21] "AI Is Replacing Humans In The Interview Process - What You Need To Know To Crush Your Next Video Interview." Accessed: Apr. 21, 2026. [Online]. Available: https://www.forbes.com/sites/janehanson/2023/09/30/ai-is-replacing-humans-in-the-interview-processwhat-you-need-to-know-to-crush-your-next-video-interview/

[22] "Artificial intelligence in personnel selection and its influence on employer attractiveness | Request PDF." Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/publication/342047560_Artificial_intelligence_in_personnel_selection_and_its_influence_on_employer_attractiveness

[23] M. Langer, C. J. König, and K. Krause, "Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings," *Int. J. Sel. Assess.*, vol. 25, no. 4, pp. 371–382, Dec. 2017, doi: 10.1111/IJSA.12191;PAGE:STRING:ARTICLE/CHAPTER.

[24] "(PDF) Speech production under uncertainty: how do job applicants experience and communicate with an AI interviewer?" Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/publication/371966119_Speech_production_under_uncertainty_how_do_job_applicants_experience_and_communicate_with_an_AI_interviewer

[25] "Where Automated Job Interviews Fall Short." Accessed: Apr. 21, 2026. [Online]. Available: https://hbr.org/2022/01/where-automated-job-interviews-fall-short

[26] "It Takes More Than a Good Camera: Which Factors Contribute to Differences Between Face-to-Face Interviews and Videoconference Interviews Regarding Performance Ratings and Interviewee Perceptions? - PMC." Accessed: Apr. 21, 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7482058/

[27] "Understanding Interviewees' Perceptions and Behaviour towards Verbally and Non-verbally Expressive Virtual Interviewing Agents | Companion Publication of the 2022 International Conference on Multimodal Interaction." Accessed: Apr. 21, 2026. [Online]. Available: https://dl.acm.org/doi/10.1145/3536220.3558802

[28] "Does media richness influence job applicants' experience in asynchronous video interviews? Examining social presence, impression management, anxiety, and performance - Rizi - 2024 - International Journal of Selection and Assessment - Wiley Online Library." Accessed: Apr. 21, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/ijsa.12448

[29] "'The interviewer is a machine!' Investigating the effects of conventional and technology-mediated interview methods on interviewee reactions and behavior - Kleinlogel - 2023 - International Journal of Selection and Assessment - Wiley Online Library." Accessed: Apr. 21, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/ijsa.12433

[30] "Frontiers | Asynchronous Video Interviewing as a New Technology in Personnel Selection: The Applicant's Point of View." Accessed: Apr. 21, 2026. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.00863/full

[31] J. P. Hausknecht, D. V. Day, and S. C. Thomas, "Applicant reactions to selection procedures: An updated model and meta-analysis," *Pers. Psychol.*, vol. 57, no. 3, pp. 639–683, Sep. 2004, doi: 10.1111/J.1744-6570.2004.00003.X.

[32] "(PDF) Once an Impression Manager, Always an Impression Manager? Antecedents of Honest and Deceptive Impression Management Use and Variability across

Multiple Job Interviews." Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/publication/312083871_Once_an_Impression_Manager_Always_an_Impression_Manager_Antecedents_of_Honest_and_Deceptive_Impression_Management_Use_and_Variability_across_Multiple_Job_Interviews

[33] D. George, K. S. Keerthi, K. Gayathri, and G. N. Harshini, "AI powered learning management system with mock interview platform," *Artif. Intell. Sustain. Innov.*, pp. 433–438, Jan. 2026, doi: 10.1201/9781003654049-63.

[34] Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, Mohammed (Ehsan)Hoque, "Automated Analysis and Prediction of Job Interview Performance," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, 2015, [Online]. Available: https://arxiv.org/abs/1504.03425

[35] J. R. Lewis, "The System Usability Scale: Past, Present, and Future," *Int. J. Hum. Comput. Interact.*, vol. 34, no. 7, pp. 577–590, Jul. 2018, doi: 10.1080/10447318.2018.1455307.

[36] A. F. Hayes, "Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach," *Guilford Press*, 2018.

[37] G. G. K. J R Landis, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–74, 1977, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/843571/

[38] "(PDF) Introducing a Multi-Stakeholder Perspective on Opacity, Transparency and Strategies to Reduce Opacity in Algorithm-Based Human Resource Management." Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/publication/356287777_Introducing_a_Multi-Stakeholder_Perspective_on_Opacity_Transparency_and_Strategies_to_Reduce_Opacity_in_Algorithm-Based_Human_Resource_Management