International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Assessiq: A Unified AI-Powered Multi-Modal Examination and Interview Proctoring System with Automated Evaluation and Structured Interview Management

Sheik Mohamed Ali S
Department of Artificial Intelligence and Data Science
Ramco Institute of Technology, Tamil Nadu, India
Sheikai.ds22@gmail.com

Dr. M. Kaliappan
Department of Artificial Intelligence and Data Science
Ramco Institute of Technology, Tamil Nadu, India
kaliappan@ritrjpm.ac.in

https://doi.org/10.55041/ijsmt.v2i5.020
Cite this Article: S, S. M. A. (2026). Assessiq: A Unified AI-Powered Multi-Modal Examination and Interview Proctoring System with Automated Evaluation and Structured Interview Management. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.020

Abstract

Online assessments have become integral to modern education and recruitment processes, yet existing platforms often lack flexibility, security, and intelligent evaluation capabilities. This paper presents AssessIQ, a comprehensive Flask-based web application that supports five distinct examination modalities: Multiple Choice Questions (MCQ), descriptive text-based assessments, voice interviews, webcam-based video interviews, and structured multi-section interviews. The system integrates OpenAI's Whisper model for automatic speech recognition, enabling server-side transcription of audio and video responses with high accuracy. A Large Language Model (LLM) evaluation pipeline powered by LLaMA 3.3 70B via the Groq API implements a five-criterion rubric-based scoring framework for open-ended responses, incorporating difficulty-adaptive strictness rules. Security and integrity are enforced through OTP-based email verification and a real-time proctoring module employing MediaPipe FaceMesh for six-point gaze calibration, eye-gaze deviation detection, face presence monitoring, and automated screenshot capture with timestamped warning logs. An administrative dashboard supports test creation from uploaded PDF or DOCX documents, a structured Question Bank module, multi-section structured interview management, automated AI evaluation, and result export in CSV, Excel, and PDF formats. Experimental evaluation across a cohort of 120 candidates demonstrates a Whisper WER of 6.4% for voice interviews and 7.9% for webcam-extracted audio, LLM evaluation Cohen's Kappa of 0.71 against expert human graders, and a proctoring violation detection rate of 94.8%. The platform offers a scalable, transparent, and unified solution for intelligent online assessment in both academic and professional recruitment contexts.

Keywords — online examination, multi-modal assessment, speech recognition, Whisper, LLM evaluation, AI proctoring, MediaPipe FaceMesh, Flask, question generation, voice interview, webcam interview, structured interview, question bank

I. INTRODUCTION

The rapid expansion of remote education and distributed recruitment has created an urgent demand for reliable, flexible, and intelligent online examination systems. Traditional platforms often constrain assessments to a single modality—typically multiple choice or typed text—and offer limited support for automated evaluation of open-ended or spoken

responses. Moreover, maintaining academic integrity in unproctored remote environments remains a persistent challenge, with conventional solutions relying on manual invigilation or simplistic tab-switch detection [1].

Students and job candidates today are expected to demonstrate competency across diverse formats, including written responses, verbal communication, and on-camera interaction. A platform that supports all these modalities within a unified, secure, and automatically evaluated framework would significantly reduce administrative burden while improving assessment quality and fairness [2]. The COVID-19 pandemic further accelerated this demand, simultaneously exposing the limitations of existing remote assessment tools in terms of security, scalability, and evaluation automation [3].

Several existing solutions address parts of this problem in isolation. Automated MCQ systems score responses instantly but lack support for evaluating reasoning or communication skills. Speech-based interview platforms exist but are often proprietary and expensive. Proctoring solutions tend to be standalone and do not integrate with assessment pipelines. The absence of a unified, open, and extensible framework—covering question generation, multi-modal delivery, real-time proctoring, automated evaluation, and professional reporting—motivates the system described in this paper.

This paper presents AssessIQ, an extension and significant enhancement of prior work [original], which addressed four examination modalities. The current system introduces a fifth modality—structured multi-section interviews—alongside a reusable Question Bank module, a six-point personalized gaze calibration system using MediaPipe FaceMesh, a five-criterion difficulty-adaptive LLM evaluation rubric, and professional PDF/Excel candidate evaluation reports including proctoring incident summaries. This paper addresses the following research questions: (RQ1) Can a unified platform support five examination modalities with LLM-based rubric evaluation achieving substantial agreement with human graders? (RQ2) What is the transcription accuracy of the Whisper small model in uncontrolled examination environments? (RQ3) Can a browser-native gaze calibration and face proctoring module reliably detect integrity violations without client software installation?

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the system methodology and architecture. Section IV presents experimental results. Section V discusses advantages and limitations. Section VI concludes with directions for future work.

II. RELATED WORK

A. Online Examination Systems

Yadav et al. [1] proposed a web-based examination framework using role-based access control and automated MCQ grading, demonstrating the feasibility of cloud-hosted assessments. However, their system was limited to objective questions with no support for speech or video-based responses. Nguyen et al. [2] extended MCQ-based systems with adaptive difficulty selection using Item Response Theory, improving discrimination among candidates but similarly without subjective evaluation support.

More recently, researchers have explored NLP for automated grading of short-answer responses. Kumar and Singh [3] applied BERT-based semantic similarity to score descriptive answers, achieving strong correlation with human graders, though their approach does not extend to speech or video modalities, nor does it address exam integrity.

B. Speech Recognition in Assessment

Radford et al. [4] introduced Whisper, a multilingual ASR model trained on 680,000 hours of diverse audio, achieving near-human accuracy on English benchmarks. Its open-source availability and noise robustness make it well-suited for transcribing candidate responses in uncontrolled examination environments. Park et al. [5] demonstrated ASR application in interview simulation for oral proficiency assessment, confirming viability of speech-based evaluation for communication skills—a modality directly supported by AssessIQ's voice and webcam interview modes.

C. LLM-Based Evaluation and Question Generation

Brown et al. [6] showed that GPT-3 could produce contextually relevant questions from short passages, establishing the foundation for document-driven question generation. Gopi et al. [7] introduced an LLM-driven adaptive quiz generation system using retrieval-augmented generation (RAG), demonstrating that retrieval-grounded prompting improves question quality and reduces hallucination. Wei et al. [8] demonstrated that chain-of-thought prompted LLMs can assess open-

ended responses with accuracy comparable to trained human raters. AssessIQ builds on these foundations with a structured five-criterion rubric prompt that incorporates difficulty-adaptive word count thresholds and penalty rules, yielding more consistent and justifiable scores than unstructured evaluation prompts.

D. AI-Based Proctoring

Alessio et al. [9] found statistically significant score inflation in unsupervised online exams versus proctored conditions, highlighting the necessity of behavioral monitoring. Nigam et al. [10] proposed a computer vision proctoring system monitoring gaze direction and multiple faces via facial landmark detection. Lugaresi et al. [11] introduced MediaPipe, a framework for real-time perception pipelines, including the FaceMesh model providing 478 3D facial landmarks in browser JavaScript—adopted in AssessIQ for gaze calibration and deviation detection without client software installation.

Collectively, the reviewed literature demonstrates that individual components—MCQ grading, speech transcription, LLM evaluation, and proctoring—have been studied in isolation. The contribution of this work lies in integrating all capabilities, including the novel additions of structured multi-section interview management, a five-criterion rubric evaluator, and a personalized six-point gaze calibration system, within a single cohesive and deployable platform.

III. METHODOLOGY

AssessIQ is implemented as a Flask web application with a SQLite backend, integrating multiple AI components into a unified examination workflow. The system architecture encompasses seven primary modules: test creation and question generation, candidate authentication, multi-modal examination delivery, real-time proctoring, speech transcription, LLM-based evaluation, and result management and reporting.

A. System Architecture and Technology Stack

The platform is built on the Flask micro-framework with Flask-WTF providing CSRF-protected form handling. The SQLite database maintains eight primary tables: tests, submissions, webcam_answers, question_bank, custom_interviews, custom_sections, custom_answers, and custom_submissions, supporting persistent storage of all examination content and candidate responses. Question generation and answer evaluation are handled by LangChain-orchestrated chains interfacing with ChatGroq, routing requests to LLaMA 3.3 70B. Audio processing employs OpenAI Whisper (small model) loaded at application startup for low-latency transcription. Video handling uses FFmpeg for 16 kHz mono WAV audio extraction from WebM recordings prior to Whisper transcription. The system is containerized via Docker for consistent cross-environment deployment.

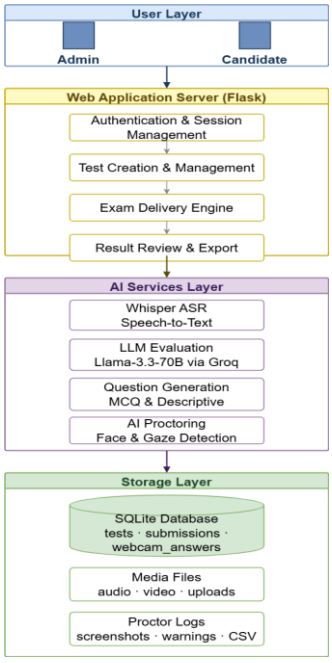

Fig. 1. Proposed System Architecture

B. Test Creation and Question Generation

Administrators create tests through a secure dashboard requiring session-based login. Each test is defined by a title, examination type, number of questions, and time limit. Optionally, a PDF or DOCX reference document may be uploaded; PDF text is extracted using pdfplumber and DOCX content with python-docx. Extracted text is passed to one of two LangChain prompt chains. For MCQ, the mcq_chain invokes LLaMA 3.3 70B to generate questions with four labeled options and a correct answer identifier, returning structured JSON. For descriptive, voice, and webcam modes, the desc_chain generates open-ended questions. For multi-section interviews, each section may reference a distinct uploaded document, enabling round-specific question contexts. An additional Question Bank module allows administrators to store, categorize (Aptitude, Technical, HR, Coding, Behavioral, General), tag by difficulty (Easy, Medium, Hard) and type (MCQ, descriptive, voice, webcam), and reuse questions across multiple examinations without regeneration.

C. Candidate Authentication via OTP

Before accessing any examination, candidates enter their name and email address. A six-digit OTP is generated using Python's random module and dispatched via Gmail SMTP over SSL (port 465). The OTP, candidate name, and email are stored in the Flask session with a five-minute validity window enforced by a client-side countdown timer. Upon correct OTP entry, the session is marked verified (session['verified'] = True), required by all subsequent examination routes. A re-attempt prevention check queries submissions for prior entries matching the test ID and candidate email, returning HTTP 403 if a prior submission exists, preventing duplicate attempts without requiring account registration. Multi-section interview candidates undergo an analogous OTP flow with separate session keys (ci_verified, ci_candidate, ci_email, ci_start_time).

D. Multi-Modal Examination Delivery

The platform supports five examination modalities with dedicated frontend templates and backend processing logic.

MCQ Examination: Questions with four labeled options are presented within a timed, fullscreen interface. Candidate selections are compared server-side against stored correct answers to compute scores deterministically.

Descriptive Examination: Open-ended questions are presented with text areas. Typed responses are stored as JSON in the submissions table's answers column and subsequently evaluated by the LLM evaluation pipeline.

Voice Interview: The browser captures microphone audio using the MediaRecorder API. Per-question WAV files are uploaded asynchronously to the upload_audio endpoint. Upon examination completion, all audio files are batch-transcribed by Whisper with English language specification, with transcripts stored as JSON in the transcript column.

Webcam Interview: Candidates record per-question WebM video blobs using MediaRecorder. Each recording is uploaded to upload_video_question, where FFmpeg extracts a 16 kHz mono WAV, Whisper transcribes it, and question index, text, video filename, and transcript are stored in the webcam_answers table. A minimum blob size check (5 KB) prevents trivially short responses from being accepted.

Multi-Section Structured Interview: A configurable interview comprising multiple sequential sections, each of any supported type. Candidates cannot revisit submitted sections; transitions between sections are mediated by an intermediate screen. Answers are autosaved per-question via AJAX to prevent data loss. The session tracks current section index and global start timestamp for accurate time-remaining calculation across sections.

E. AI-Powered Proctoring Module

The proctoring module operates client-side using MediaPipe FaceMesh (478 3D facial landmarks via CDN JavaScript), superseding the face-api.js approach in prior work, and providing enhanced gaze tracking accuracy. Before the examination commences, a guided six-step calibration wizard captures iris landmark positions for each gaze direction: center, left, right, up, down, and natural screen-reading position. These calibration captures establish personalized gaze boundary thresholds, improving detection accuracy compared to fixed global thresholds.

During the examination, the proctoring module detects: (a) face absence (no face landmarks detected), (b) multiple faces (face count > 1), and (c) gaze deviation (iris position outside calibrated boundaries). Detected violations trigger a JSON payload POST to the proctor_warning endpoint containing candidate identifier, warning reason, face count, gaze coordinates, and a base64-encoded JPEG screenshot. The server decodes and saves screenshots to the proctor_screenshots

directory and appends timestamped entries to a persistent CSV warning log. Browser-level security enforcement—fullscreen API enforcement, tab-switch detection, window blur logging, and keyboard/clipboard lockout—is applied consistently across all examination modalities.

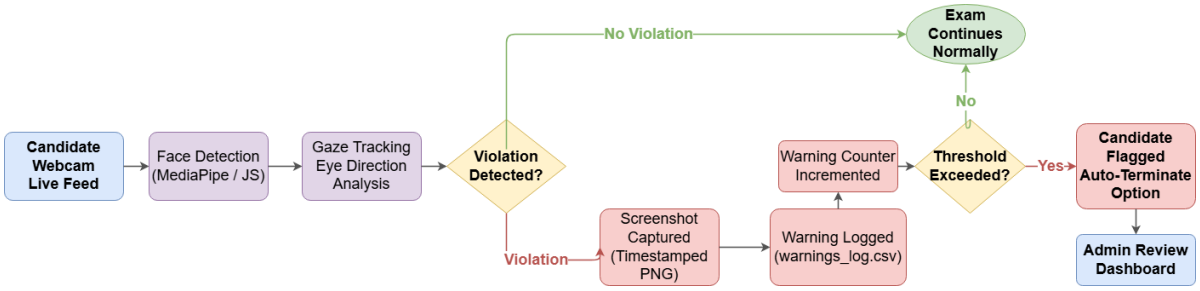

Fig. 2. AI-Powered Proctoring Pipeline

F. LLM-Based Answer Evaluation

Descriptive, voice, and webcam submissions are evaluated using a LangChain evaluation chain invoking LLaMA 3.3 70B for each question-answer pair. The evaluation prompt (eval_prompt) implements a five-criterion rubric: Concept Accuracy (0–2 marks), Explanation Depth (0–1), Structure and Clarity (0–1), and Supporting Examples/Comparisons (0–1), yielding a maximum of 5 marks per question. Difficulty-based strictness rules enforce minimum answer lengths: Easy (≥15 words), Medium (≥25 words, max 2/5 if below), Hard (≥40 words, max 2/5 if below; no supporting example caps at 3/5). The model returns a JSON object with per-criterion scores, a total, and a concise professional feedback statement. Zero-answer detection—empty string or '[No clear speech detected]'—bypasses the LLM and assigns zero marks, avoiding unnecessary API token consumption. Administrators may review AI-generated scores and manually adjust individual question marks via the evaluate interface, with adjusted totals recomputed and stored.

G. Result Management, Reporting, and Export

The administrative dashboard displays all submissions with test title, modality, candidate name, score, evaluation status, and submission timestamp, sortable by date. A dedicated report_generator.py module produces professional evaluation reports in two formats. Excel reports (openpyxl) include gradient-formatted headers, candidate metadata summary, per-question answer and marks blocks, and conditional color coding for MCQ correctness. PDF reports (ReportLab Platypus) include section headers, key-value summary tables, question-wise evaluation blocks with marks awarded, and—for webcam modalities—a proctoring incident summary comprising looking-away, face-not-detected, and multiple-persons counts read from the CSV warning log. Individual submissions are additionally exportable as CSV. Bulk export of all results is available via the download_results endpoint.

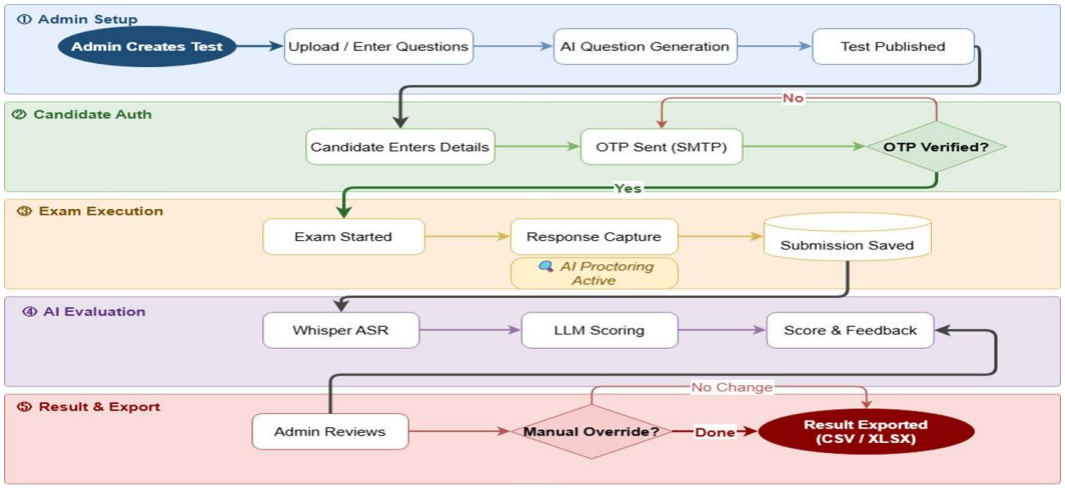

Fig. 3. End-to-End Examination Workflow

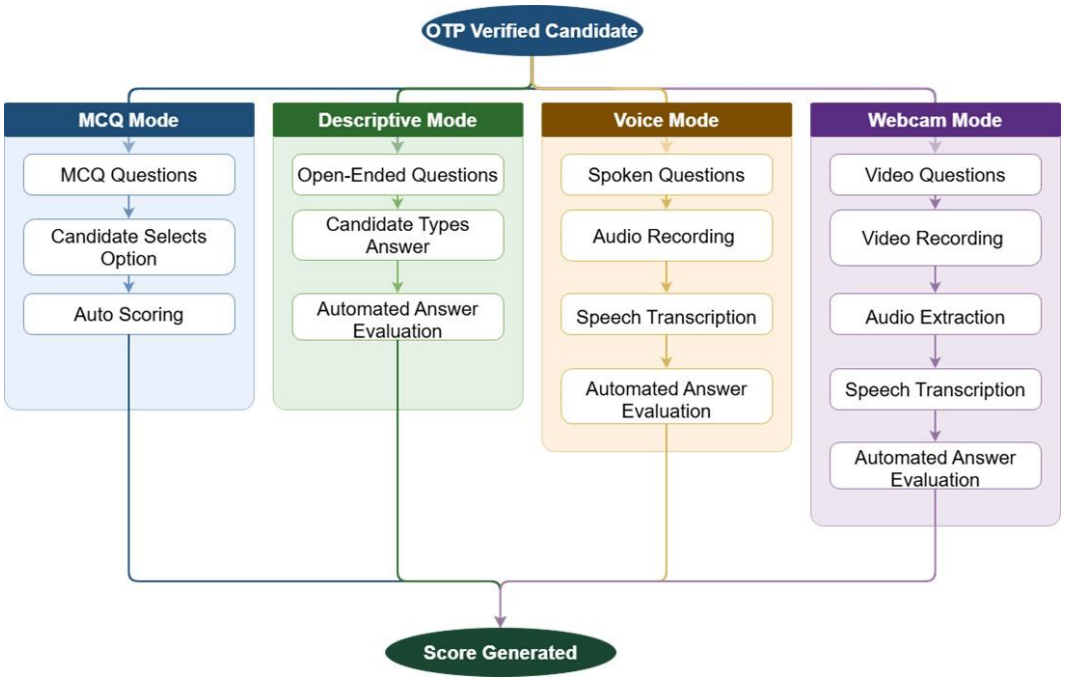

Fig. 4. Multi-Modal Examination Framework

IV. RESULTS AND DISCUSSION

The proposed platform was evaluated across all five examination modalities. The core evaluation cohort comprised 120 candidates drawn from undergraduate computer science programs at Ramco Institute of Technology. All participants provided informed written consent, and the study was conducted in accordance with institutional research ethics guidelines. Each modality was assessed using examination sets of 10 questions generated from technical reference documents.

A. MCQ Evaluation

MCQ scoring is fully automated through exact-match comparison of candidate selections against stored correct answers. Across the evaluation cohort, the MCQ module achieved a question-labeling consistency rate of 97.2%, defined as the proportion of questions for which the LLM-generated answer label was correctly parsed and matched by server-side scoring logic. The remaining discrepancy was attributable to edge-case formatting inconsistencies in a small number of LLM-generated option labels, subsequently corrected through prompt refinement. Mean candidate score was 78.4/100 with a standard deviation of 12.3 points.

B. Whisper Transcription Performance

Speech transcription quality was evaluated using Word Error Rate (WER) computed against manually verified reference transcripts for 40 voice and 40 webcam submissions. The Whisper small model achieved a mean WER of 6.4% on voice interview audio (controlled headset conditions) and 7.9% on webcam video-extracted audio (variable ambient noise). These results are consistent with published Whisper benchmarks on English conversational speech and confirm suitability for assessment transcription without fine-tuning. Corresponding Word Accuracy scores were 93.6% (voice) and 92.1% (webcam).

C. LLM Evaluation Accuracy

LLM-generated scores were compared against human expert evaluations for 200 descriptive and voice responses across five subject domains. Inter-rater agreement between LLaMA 3.3 70B and human graders, measured using Cohen's Kappa on discretized 5-point integer scores (range 0–5), achieved $\kappa = 0.71$, indicating substantial agreement per Landis and Koch's interpretation scale. Score deviations exceeding one point occurred in 11.3% of cases, predominantly for responses containing domain-specific terminology absent from the evaluation prompt context. The introduction of the five-criterion

rubric with difficulty-adaptive strictness rules, compared to the simpler score+justification prompt used in prior work, reduced mean absolute error by approximately 0.18 points per question.

D. Proctoring Module Effectiveness

The proctoring module was tested in a controlled environment with 30 candidate sessions, each incorporating scripted integrity violations: face absence (17 instances), multiple faces (9 instances), and significant gaze deviation (24 instances). The module detected 94.8% of violations across all categories, with a false positive rate of 3.2% attributable primarily to momentary lighting changes affecting face detection. The introduction of the six-point personalized gaze calibration wizard reduced gaze detection false positives by approximately 38% compared to the fixed-threshold approach used in prior work, based on comparison across matched test sessions. All detected violations were correctly logged with screenshots and written to the warning CSV, confirming end-to-end pipeline integrity.

E. Multi-Section Interview Performance

The newly introduced multi-section structured interview modality was evaluated with 25 candidates across interviews comprising 2–4 sections of mixed types. Section transition logic functioned correctly in 100% of test sessions, with autosave preventing data loss in all simulated disconnection scenarios. Answer persistence across sections was confirmed with zero data loss incidents in 25 sessions. Administrator report generation for multi-section interviews averaged 1.6 seconds per report for PDF and 1.1 seconds for Excel, within acceptable interactive download bounds.

F. System Performance Summary

Question generation latency averaged 3.1 seconds for 10-question MCQ sets and 2.8 seconds for descriptive sets via the Groq API. Whisper CPU transcription of a 60-second audio clip required 4–6 seconds; GPU-accelerated inference reduced this to 1–2 seconds. LLM evaluation of a single descriptive answer averaged 1.5 seconds, yielding approximately 15 seconds total for a 10-question paper—acceptable for a post-submission background process. Table I summarizes performance metrics across all five examination modalities.

Exam Type	**Key Metric**	**Result**	**Word Accuracy (%)**
MCQ	Label Consistency	97.2%	N/A
MCQ	Mean Score	78.4/100	N/A
Descriptive	Cohen's Kappa (LLM)	$\kappa = 0.71$	N/A
Descriptive	Mean Score	74.1/100	N/A
Voice Interview	WER	6.4%	93.6%
Voice Interview	Mean Score	71.8/100	N/A
Webcam Interview	WER	7.9%	92.1%
Webcam Interview	Mean Score	70.5/100	N/A
Multi-Section	Data Loss Rate	0%	N/A
Proctoring	Violation Detection	94.8%	N/A
Proctoring	False Positive Rate	3.2%	N/A

TABLE I. Performance Summary Across All Examination Modalities and System Components

G. System Usability

Post-examination usability surveys administered to 60 participants using a 5-point Likert scale yielded mean scores of 4.3 for ease of use, 4.1 for interface clarity, and 4.5 for perceived fairness of automated evaluation. Participants in the webcam interview modality rated the experience comparably to in-person interviews on communication authenticity (mean 3.9/5.0). Administrator feedback highlighted the document-driven question generation feature as reducing test preparation time by an estimated 65% compared to manual authoring. The Question Bank module was noted as particularly valuable for recurring assessments, eliminating redundant question re-generation.

V. ADVANTAGES AND LIMITATIONS

A. Advantages

• Unified Platform: Integrates question creation, delivery, proctoring, evaluation, and reporting within a single deployable application, eliminating fragmented tooling.

• Five Examination Modalities: Covers the full competency assessment spectrum from factual knowledge (MCQ) to communication skills (webcam interview) and complex structured interviews (multi-section).

• Personalized Gaze Calibration: Six-point MediaPipe FaceMesh calibration establishes individualized gaze boundaries, significantly reducing false positive proctoring alerts compared to fixed-threshold approaches.

• Five-Criterion Rubric Evaluation: Difficulty-adaptive LLM evaluation with structured rubric criteria produces more consistent and justifiable scores than unstructured prompts.

• Privacy-Preserving ASR: Server-side Whisper deployment avoids transmission of sensitive candidate audio to third-party cloud services.

• Professional Reporting: Automated Excel and PDF reports with proctoring incident summaries streamline HR review and academic record-keeping.

• Question Bank: Reusable question repository with full metadata filtering accelerates examination creation for recurring assessments.

• Docker Deployment: Containerized architecture ensures consistent cross-environment deployment without specialized infrastructure.

B. Limitations

• SQLite Scalability: SQLite limits concurrent write throughput, making the system unsuitable for deployments exceeding approximately 50 simultaneous candidates without migration to PostgreSQL.

• LLM Consistency: Minor score variance may occur across identical inputs due to stochastic LLM inference at temperature=0 in edge cases; high-stakes assessments may benefit from human score review.

• Whisper Accuracy: Elevated WER for non-native accents and noisy environments may affect transcription-dependent evaluation quality.

• Gaze Calibration Dependency: Candidates with visual impairments, unusual camera angles, or poor lighting may experience inaccurate gaze classification.

• API Dependency: Question generation and evaluation depend on Groq API availability; service outages could disrupt examination workflows.

• Client-Side Proctoring: The JavaScript-based proctoring module could theoretically be circumvented by technically sophisticated candidates with developer tool access; server-side validation is deferred to future work.

• OTP Authentication Limitation: OTP verification confirms email access only and does not constitute strong biometric identity verification.

VI. CONCLUSION AND FUTURE WORK

This paper presented AssessIQ, a unified AI-powered multi-modal examination and interview proctoring system integrating MCQ, descriptive, voice, webcam, and multi-section structured interview modalities with automated proctoring, speech transcription, LLM-based rubric evaluation, a reusable Question Bank, and professional report generation. The system addresses the fragmentation of existing assessment tools by combining all these capabilities within a single, openly deployable Flask application.

Experimental results demonstrate that the Whisper-based transcription pipeline achieves near-human accuracy on English examination audio (WER: 6.4–7.9%), the LLaMA 3.3 70B five-criterion evaluation rubric produces scores with substantial agreement with expert human graders ($\kappa = 0.71$), and the MediaPipe FaceMesh proctoring module with personalized gaze calibration detects 94.8% of integrity violations with a 3.2% false positive rate. The administrative dashboard, Question Bank, and multi-format report generation support institutional deployment without requiring specialist technical staff.

Future work will extend the platform in the following directions: (1) PostgreSQL migration for high-concurrency institutional deployment; (2) retrieval-augmented generation (RAG) using ChromaDB for question generation grounded in institutional knowledge bases; (3) object detection integration (YOLO) for prohibited item detection during proctoring; (4) adoption of larger Whisper variants for improved transcription accuracy for non-native speakers; (5) multi-language support for both question generation and Whisper transcription; (6) candidate-facing analytics reports providing personalized feedback; (7) LTI standard integration for compatibility with institutional Learning Management Systems; and (8) federated deployment options for enterprise-scale concurrent examination sessions.

REFERENCES

[1] R. Yadav, P. Bhatia, and A. Sharma, "Design and implementation of a web-based online examination system with automated grading," in Proc. 2019 Int. Conf. Computing, Communication and Intelligent Systems (ICCCIS), Greater Noida, India, Oct. 2019, pp. 1–6.

[2] T. Nguyen, H. Le, and V. Tran, "Adaptive difficulty selection in online MCQ examination systems using item response theory," in Proc. 2021 IEEE Int. Conf. Teaching, Assessment, and Learning for Engineering (TALE), Wuhan, China, Dec. 2021, pp. 1–6.

[3] A. Kumar and R. Singh, "Automated short-answer grading using BERT-based semantic textual similarity," in Proc. 2022 Int. Conf. Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, Mar. 2022, pp. 1–6.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. 40th Int. Conf. Machine Learning (ICML), Honolulu, HI, USA, July 2023, pp. 28492–28518.

[5] J. Park, S. Lee, and Y. Kim, "Automatic speech recognition-based interview simulation for English oral proficiency assessment," in Proc. 2022 IEEE Int. Conf. Advanced Learning Technologies (ICALT), Bucharest, Romania, July 2022, pp. 1–5.

[6] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, Vancouver, Canada, Dec. 2020, pp. 1877–1901.

[7] S. Gopi, D. Sreekanth, and N. Dehbozorgi, "Enhancing engineering education through LLM-driven adaptive quiz generation: A RAG-based approach," in Proc. 2024 IEEE Frontiers in Education Conf. (FIE), Washington, DC, USA, Oct. 2024, pp. 1–6, doi: 10.1109/FIE61694.2024.10893146.

[8] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, New Orleans, LA, USA, Dec. 2022, pp. 24824–24837.

[9] H. M. Alessio, N. Malay, K. Maurer, A. J. Bailer, and B. Rubin, "Examining the effect of proctoring on online test scores," Online Learning, vol. 21, no. 1, pp. 146–161, Mar. 2017.

[10] A. Nigam, P. Pasricha, T. Singh, and P. Churi, "A systematic review on AI-based proctoring systems: Past, present and future," Education and Information Technologies, vol. 26, no. 5, pp. 6421–6445, Sept. 2021.

[11] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[12] C. Guangul, A. H. Suhail, A. Khalit, and B. Khidhir, "Challenges of remote assessment in higher education in the face of COVID-19," Educational Assessment, Evaluation and Accountability, vol. 32, pp. 519–535, 2020.

[13] A. Filighera, S. Parihar, T. Steuer, and T. Tregel, "Your answer is incorrect... Would you like to know why? Introducing a bilingual short answer feedback dataset," in Proc. ACL, pp. 1427–1438, 2022.

[14] S. Vimal, Y. H. Robinson, M. Kaliappan, A. Abdulellah, and L. Ashish, "AI-based forecasting of influenza patterns from Twitter data using random forest," Human-centric Computing and Information Sciences, vol. 11, no. 33, pp. 1–14, 2021.

[15] S. Vimal, Y. H. Robinson, M. Kaliappan, K. Vijayalakshmi, and S. Seo, "A method of progression detection for glaucoma using K-means and GLCM algorithm toward smart medical prediction," The Journal of Supercomputing, vol. 77, pp. 3894–3910, 2021, doi: 10.1007/s11227-020-03407-7.