INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

Human-AI Collaboration in Hiring: Limitations of LLM-Based Interview Evaluation

[1]Saurabh Saoji, [2]Sahil Bhati, [3]Vishal Dukale, [4]Shreekant Menon
[1]HOD, [2]Student, [3]Student, [4]Student
[1]Information Technology,
[1]Nutan Maharashtra Institute of Engineering and Technology, Pune, India

Abstract : Recent advances in large language models (LLMs) have enabled AI-powered tools for simulating technical interviews and giving feedback to candidates. We present a study of a prototype AI Interview Assistant that uses Google's Gemini API to conduct mock interviews and evaluate answers. From a responsible AI viewpoint, we critically evaluate whether such an LLM-based system can be trusted for interview evaluation, focusing on fairness, consistency, transparency, and explainability. In a simulated user trial, we designed varied prompt templates and controlled candidate profiles to probe the system. We observe significant variability in scores and feedback across prompt framings and candidate phrasing, inconsistent scoring on repeated queries, and occasional drift in tone. The system often provides generic praise or criticism without clear rationale, highlighting a lack of transparency. These issues echo known risks of LLMs in high-stakes settings. We discuss implications for human-AI collaboration: the assistant can aid interview prep but cannot replace human judgment. Our work underscores the need for careful validation, bias mitigation, and explainability measures when deploying LLM-driven hiring tools.

IndexTerms **- AI-powered interview assistant, Mock interviews, Adaptive questioning, Real interview simulation, Progress tracking, Domain-specific insights.**

I. INTRODUCTION

AI-driven interview tools are rapidly emerging. Commercial and academic prototypes use LLMs to generate interview questions and evaluate candidate answers, promising efficiency and scalability. In this paper, we assess a Gemini-powered interview assistant from a responsible AI standpoint. Our goal is to understand when and how an LLM-based interviewer can be trusted to evaluate answers. We simulate interview scenarios and systematically vary prompt formulations and candidate descriptors. Key questions include: How stable are the system's scores if we rephrase prompts or repeat queries? Do different candidate backgrounds lead to different outcomes? Does the assistant justify its ratings in a transparent way? We frame our evaluation around the trust dimensions of fairness, reliability, transparency, and explainability.

II. LITERATURE SURVEY

1. ChatGPT for Learning HCI Techniques: A Case Study on Interviews for Personas
   Author: - Jose Barambones, Cristian Moral, Angélica de Antonio, Ricardo Imbert and Elena Villalba-Mora
   The paper proposes an intelligent system that analyzes user answers based on technical accuracy, tone, and confidence. It also emphasizes improving recruitment efficiency and reducing human bias in interviews. Overall, it presents AI as a solution for conducting smart, scalable, and fair interview assessments.

2. An AI Mock Interview Platform for Interview Preparation
   Author: -Yi-Chi Chou, Felicia R. Wongso, Chun- Yen Chao, Han-Yen Yu
   This paper introduces a Mock-Interview Platform (MIP) that integrates visual, audio, and textual features to evaluate interview performance. The platform analyses emotions, head pose, voice, DISC personality traits, and intrinsic traits, providing AI-assisted feedback. The experiment results demonstrated satisfactory outcomes in prediction scores.

3. A Comprehensive Study and Implementation of the Mock Interview Simulator with AI and Pose-Based Interaction
   Author: -Balasaheb Jadhav, Avadhut Sawant, Arnav Shah, Pranamya Vemula
   This paper presents the Mock Interview Simulator, which uses AI-driven interviewers and combines speech recognition, text-to-speech synthesis, and posture detection to create a realistic interview setting. It aims to help job seekers prepare for diverse interview scenarios and provides valuable performance insights.

4. AI-Based Mock Interview Evaluator: An Emotion and Confidence Classifier Model
   Author: - Rubi Mandal, Pranav Lohar, Dhiraj Patil, Apurva Pati
   The authors propose an AI-powered mock interview platform that assesses candidates on emotions, confidence, and knowledge. Emotions are evaluated using a deep learning CNN algorithm, confidence through speech recognition, and knowledge using keyword mapping and semantic analysis techniques.

III. PROPOSED SYSTEM ARCHITECTURE

The AI-Powered Interview Assistant is designed as a full-stack web-based system that integrates multiple modern technologies React with Tailwind CSS for the frontend, Flask for backend logic, Firebase for authentication and data storage and Gemini AI for intelligent response evaluation.

The architecture ensures modularity, scalability and real-time interaction between users and the AI model. It enables smooth communication between the frontend interface, backend APIs, and AI evaluation engine while maintaining secure and efficient data flow.

3.1 System Components

1. Frontend (React + Tailwind CSS)
   The frontend will provide a responsive, user-friendly interface for interacting with the system. It also displays the home page, interview setup screen, question-answer interface, and feedback dashboard. It communicates with Flask backend through secure REST API calls. Allows users to:

   - Log in/register via Firebase Authentication.
   - Choose difficulty level and number of questions.
   - Enter answers for technical interview questions.
   - View AI-generated feedback and progress charts.

2. Backend (Flask)
   The backend will act as a bridge between frontend, Firebase, and Gemini AI. It uses Flask RESTful APIs to ensure modular communication and scalability. It also handles application logic, API routing, and data management. Major functionalities:

   - Receive answers from the frontend.
   - Send responses to Gemini AI for evaluation.
   - Process AI feedback into structured report format.
   - Store user responses and feedback in Firebase Firestore.

3. Gemini AI (Evaluation Engine)
   The Gemini model will serve as the core intelligence of the system. Evaluates user-submitted answers based on:

   - Receive answers from the frontend.
   - Send responses to Gemini AI for evaluation.
   - Process AI feedback into structured report format.
   - Store user responses and feedback in Firebase Firestore.

   Returns a structured response containing:

   - Evaluation comments.
   - Performance score.
   - Suggestions for improvement.

   The Flask backend parses this evaluation and converts it into a readable report for the user.

4. Firebase (Database and Authentication)
   Firebase Authentication will manage secure login and registration. It also provides real-time data synchronization for a seamless experience. Firebase Firestore stores:

   - User details and authentication tokens.
   - Interview configurations (difficulty, question count).
   - AI evaluation reports and performance data.
   - Progress tracking statistics.

5. Feedback and Progress Module
   It will be responsible for report generation and progress visualization. It displays user's historical performance using charts and summary tables. It also fetches data from Firebase and formats it in React for visual presentation. It enables users to identify trends, strengths, and areas needing improvement.

IV. SYSTEM ANALYSIS

4.1 Response Evaluation Score

The proposed system evaluates user responses using a multi-criteria scoring mechanism designed to capture both technical correctness and communication quality. Each response is assessed across three key dimensions: content correctness (C), relevance (R), and linguistic clarity (L). The overall score is computed as a weighted sum:

$$S = \alpha C + \beta R + \gamma L \qquad (4.1)$$

$$\alpha + \beta + \gamma = 1 \qquad (4.2)$$

This approach ensures a balanced evaluation by considering both the accuracy of the response and its presentation quality. The flexibility of the weighting coefficients allows the system to adapt to different interview contexts. For example:

- Technical interviews may assign higher weight to correctness
- HR interviews may emphasize clarity and relevance
- The scoring framework ensures that responses are not judged solely on correctness but also on how effectively the candidate communicates their ideas.

4.2 Confidence Function

To evaluate user improvement over time, the system computes a cumulative performance score based on multiple interview sessions:

$$P_n = \frac{1}{n}\sum_{i=1}^{n} S_i \qquad (4.3)$$

This metric provides a longitudinal view of user performance. As the number of sessions increases, the average score stabilizes, offering a more reliable measure of the user's capability.
The observed trend indicates that repeated interaction with the system leads to gradual improvement in scores. This suggests that the system effectively supports iterative learning by allowing users to identify weaknesses and refine their responses over time.

4.3 Prompt-wise Performance Analysis

The prompt-wise average score analysis (Fig. 4.1) highlights the variation in user performance across different interview questions. Key observations:

- Certain prompts show higher average scores (above 7), indicating that users are more comfortable with these topics
- Mid-range prompts (scores between 5–6.5) represent moderate difficulty and typical performance levels
- Lower-scoring prompts indicate areas where users struggle, possibly due to higher complexity or less familiarity

This variation demonstrates that not all questions are equally challenging, and the system successfully captures these differences. The analysis can be used to:

- Identify difficult topics
- Improve prompt design
- Guide users toward targeted practice areas

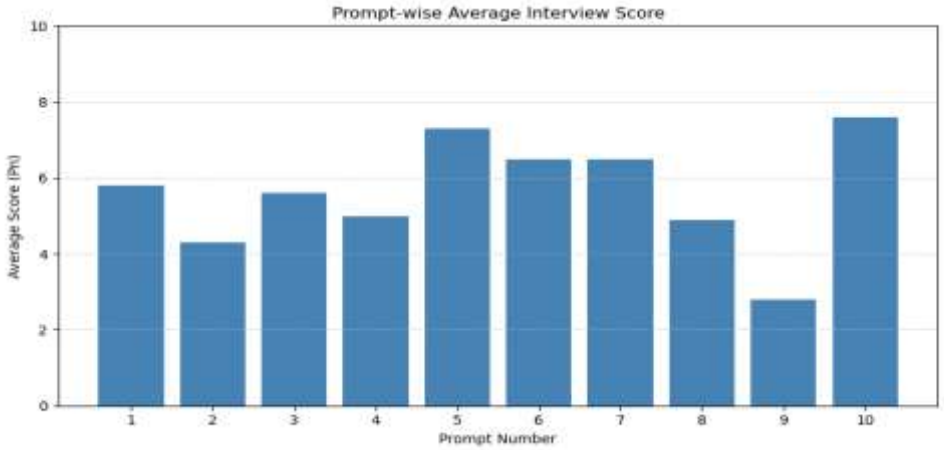

fig 4.1 Average Confidence Score

4.4 Feedback Consistency Metric

To measure the reliability of the evaluation process, a simple accuracy-based metric is used:

$$\text{Accuracy} = \frac{\text{Correct Responses}}{\text{Total Questions}}$$

This metric reflects how consistently the AI-generated feedback aligns with expected response quality based on predefined evaluation criteria. The results indicate that the system maintains a stable level of accuracy across different sessions. This consistency suggests that the evaluation logic, guided by structured prompts, produces repeatable and predictable outcomes. While the metric provides a general indication of reliability, it also highlights that the system is intended as a support tool rather than an absolute evaluator.

4.5 Confusion Matrix Analysis

The confusion matrix (Fig. 4.2) compares expected response quality (High, Medium, Low) with AI-evaluated classifications.

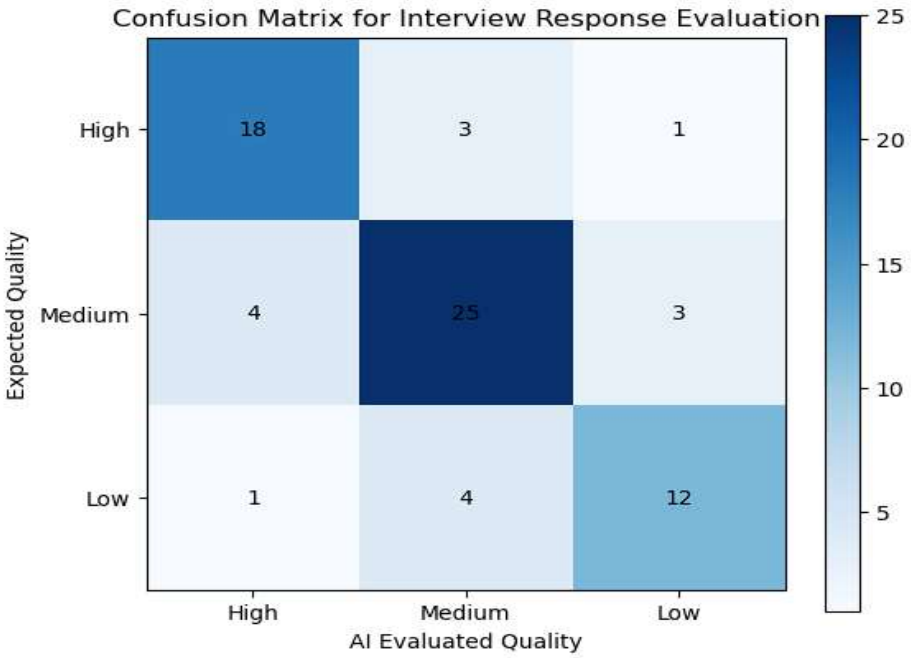

fig 4.2 Confusion Matrix of Quality

Key Observations:

- A significant proportion of predictions fall along the diagonal, indicating correct classification
- High-quality responses are accurately identified in most cases, demonstrating strong precision in top-tier evaluations
- Medium-quality responses show the highest concentration, suggesting stability in average-case classification
- Low-quality responses are also correctly identified, though with slightly more variation

Misclassification Analysis:
Most misclassifications occur between adjacent categories like medium vs high; or medium vs low. These errors are expected due to overlapping scoring thresholds and subjective boundaries between categories. Very few extreme misclassifications like high classified as low are observed, indicating robustness in evaluation.

4.6 Overall System Performance

Based on the above analyses, the system demonstrates the following characteristics:

- Consistency: Evaluation results remain stable across multiple sessions
- Adaptability: The weighted scoring model supports different interview contexts
- Usability: Users can identify strengths and weaknesses through structured feedback
- Reliability: The confusion matrix indicates strong agreement between expected and predicted outcomes

IV. CONCLUSION

In conclusion, the AI-Powered Interview Assistant project successfully demonstrates how artificial intelligence can be leveraged to enhance the process of technical interview preparation. By integrating Flask, Firebase, React with Tailwind CSS, and Gemini AI, the system provides a seamless and intelligent platform where users can practice interview questions, receive AI-driven feedback, and track their progress over time. The application addresses key challenges in traditional interview preparation such as lack of personalized feedback, time constraints, and difficulty in self-assessment by automating evaluation and generating detailed performance reports. Users can choose their preferred difficulty level and number of questions, making the experience flexible and tailored to individual learning goals. The system's architecture ensures scalability, security, and usability, while Firebase enables real-time data synchronization and user management. The inclusion of progress tracking encourages continuous improvement, allowing learners to monitor their strengths and weaknesses effectively. Overall, the project achieves its objectives of creating a smart, accessible, and interactive platform for interview preparation. It not only benefits students and job seekers but also demonstrates the practical application of modern technologies like AI and cloud computing in education and skill development.

REFERENCES

[1] J. Barambones, C. Moral, A. de Antonio, R. Imbert, L. Martínez-Normand and E. Villalba-Mora, "ChatGPT for Learning HCI Techniques: A Case Study on Interviews for Personas," in IEEE Transactions on Learning Technologies, vol. 17, pp. 1460-1475, 2024, doi: 10.1109/TLT.2024.3386095.

[2] Y. -C. Chou, F. R. Wongso, C. -Y. Chao and H. -Y. Yu, "An AI Mock-interview Platform for Interview Performance Analysis," 2022 10th International Conference on Information and Education Technology (ICIET), Matsue, Japan, 2022, pp. 37-41, doi: 10.1109/ICIET55102.2022.9778999.

[3] Jadhav, Balasaheb & Sawant, Avadhut & Shah, Arnav & Vemula, Pranamya & Waikar, Abhijeet & Yadav, Srushti. (2024). A Comprehensive Study and Implementation of the Mock Interview Simulator with AI and Pose-Based Interaction. 01-05. 10.1109/IC-CGU58078.2024.10530717.

[4] Rai, Mrs & R, Abhiram & Padthe, Adithya & R, Hrithik. (2024). AI Based Interview Evaluator: An Emotion and Confidence Classifier. IARJSET. 11. 10.17148/IARJSET.2024.11442.

[5] Kothari, Param & Mehta, Paras & Patil, Srushti & Hole, Prof. (2024). InterviewEase : AI-powered interview assistance. 10.21203/rs.3.rs-3964944/v1.

[6] Anglekar, Sumegh & Chaudhari, Urvee & Chitanvis, Atul & Shankarmani, Radha. (2021). A Deep Learning based Self-Assessment Tool for Personality Traits and Interview Preparations. 1-3. 10.1109/ICCICT50803.2021.9510143.

[7] M. Laiq and O. Dieste, "Chatbot-based Interview Simulator: A Feasible Approach to Train Novice Requirements Engineers," 2020 10th International Workshop on Requirements Engineering Education and Training (REET), Zurich, Switzerland, 2020, pp. 1-8, doi: 10.1109/REET51203.2020.00007.

[8] Prof.Nirgide, Assistant, Shubhangi Vishal, Sayyed Arsh Aktharali, Patil Paresh Narendra, Raktate Shriraj Vikas, Pathan and Fazal Mushtaque. "AI Based Interview Critique System: Interview Preparation Companion Using Deep Learning.".

[9] J. V. Barpute, O. Wattamwar, S. Pakjade and S. Diwate, "A Survey of AI-Driven Mock Interviews Using GenAI and Machine Learning (InterviewX)," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 217-224, doi: 10.1109/ICUIS64676.2024.10866631.

[10] R. M. Marvaniya, A. S. Acharya, D. M. Detroja, V. K. Dabhi and H. B. Prajapati, "Smart Prep: AI Based Interactive Interview Preparation System," 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0, Raigarh, India, 2025, pp. 1-6, doi: 10.1109/OTCON65728.2025.11070972.

[11] M. K, K. M, J. A and I. S, "AI-Based Mock Interview Application," 2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2025, pp. 1-10, doi: 10.1109/ICAECA63854.2025.11012461.