Grenze International Journal of Engineering and Technology, June Issue

Privacy-Preserving Artificial Intelligence Interview System: A Behavioral Analytics based, Real-Time Integrity Checking, Candidate Holistic Assessment

Himanshu Chauhan[1], Kavish Shukla[2], Keshav Singh[3] and Ketan Pandey[4]
[1-4]Department of Computer Science Engineering – Data Science, ABES Engineering College, Ghaziabad, India
Email: himanshu.22b1541059@abes.ac.in, kavish.22b1541125@abes.ac.in, keshav.22b1541043@abes.ac.in, ketan.22b1541008@abes.ac.in

***Abstract*— Conventional techniques of automation of interviews are confined to the analysis of written responses- without paying attention to important behavioral cues and issues of integrity which human interviewers can easily monitor. The current solutions are either based on a cloud infrastructure that poses privacy threats or are not based on the comprehensive assessment capability to offer fair and well rounded evaluation of the candidate. We present a unified model of AI interview that unifies three forms of paradigms for evaluation: (1) context-sensitive questioning, (2) behavioral profiling by face analysis, and (3) automatic integrity testing. The system makes use of privacy protection through locally-deployed language models, behavioural analysis through computer vision. as well as integrity validation via browser-level monitoring. Each of the components is fed into a central scoring Assessment for learning dashboard using visual rubrics.**

Testing across 150 candidate sessions Improved assessment (88.4% accuracy, 0.87 human agreement) and 94% of integral violations are detected. Behavioral measures provide an additional discriminating power which reduces the number of false positive recommendations by 31%.

***Index Terms*— Automation of interviews, Behavioral Analysis, Facial Expression Recognition, Integrity monitor, AI privacy preserving, Local language models.**

I. INTRODUCTION

The nature of human interviews are that they are multidimensional tests with a focus of testing content knowledge, behavioral composure and professional authenticity at the same time. Although the previous attempts in automation have advanced with analyzing verbal and textual reactions, them methodically ignore the non-verbal behavioral indications, patterns of means of showing stress, and the signs of being really engaged, which are instinctively rated by experienced interviewers.

This has been further complicated by the fast expansion of remote interviewing which poses sophisticated issues of integrity. Online platforms offer the potential of deception vectors of events that could do is physical world, such as impersonating a candidate, outside assistance without authorization and unauthorized use of Generative AI.

Grenze ID: 01.GIJET.12.2.382_24

These solutions, however, present some fundamental Privacy-Performance Paradox: either use advanced cloud based analytics exposing sensitive biometric data to third party infrastructure and opening the organisation to increases in GDPR compliance liability, or use on-premise solutions which allow data sovereignty but not able to do in-depth multimodal analysis

The present study resolves this dilemma by suggesting a Holistic Candidate Assessment System (HCAS) which is a privacy-sensitive architecture that works purely locally; Our architecture An architecture that fully functional and operates in a three synchronized layers:

- Credential-Aligned Questioning: A dynamic orchestration engine which individualizes the flow of interviews through an analysis of documented experience and in actual time adjusts the difficulty of question depending on trajectories of performance (Algorithm 1).
- Behavioral Profiling: The vision pipeline system is based on MediaPipe Face Mesh for the confidence computation indicators, emotional expressions and physiological in- senders of suffer without video
- information, dicators of stress.
- Integrity Assurance: Anomalies in the focus of attention and authenticity of the environment using gaze tracking And browser monitoring (see fig. 4).

Our combined analysis compares all the signals of all layers to determine adaptive strategies in interview and full scoring (see Fig. 1) unlike the fragmented techniques of addressing each dimension separately.

Contributions: The technical activities by this paper to the area of automated assessment are as follows: Threshold-Driven Orchestration: An adaptive Zone of Proximal Development engine that will design for candidates and will save-dynamically change the difficulty of Questions in response to real-time competency signals.

Privacy-Preserving Deployment We prove that it is possible to execute full on-premise, fully 4-bit quantized Large Language Models (Llama 3) via optimized edge computing No exposure to third-party data.

Transparent Scoring: This is a multi-dimensional rubric that has interactive visualization(see Fig. 5) that provides explainable and recommendations that can be audited, responding to the black-box problem of AI hiring.

Empirical Validation: Experiments on 150 unique applicants has proven that behavioral plus integrity layers combination gives decision accuracy of 88.4% (25% better than detection of violation of integrity) and integrity violations detection rates of 94%, high inter-rater consistency (rho=0.87)

II. Related Work

Automation systems history in interviews were majorly based on keyword extraction and template matches based on Applicant Tracking Systems (ATS) [1]. Later cloud-based AI systems [2], [3] also suggested generative models for contextual questioning but mostly rely on the transmission of data to third parties, which is so controversial in terms of privacy. Moreover, these platforms tend to focus on textual content analysis without looking at the non verbal behavior cues which are required for an in-depth analysis.

Affective computing studies has proved that it is possible to profile behavior through the use of automated systems. Deep learning models trained using data sets such as FER-2013 [5] and AffectNet [6] have been used successfully in systems based on the Facial Action coding System introduced by Ekman and Friesen [4] Both educate [7] and assess on a professional level [8]. Such studies are reflected in parallel studies of stress manifestation [9], [10] which prove the fact that vocal prosody[11] and multimodal fusion methods[12] are able to effectively establishes psychological conditions. Nevertheless, current literature studies these modalities individually, with no single model that relates the behavioral indicators of stress to the quality of the technical responses.

Gaze tracking and anomaly detection have been introduced in commercial proctoring systems to reduce the issue of remote assessment fraud [13], [14]. These systems have been criticized as being an invasive surveillance processes, dependent on clouds, prone to algorithmic prejudice, barring that, they are effective at detecting the most blatant violations [15], [16]. Attention monitoring solution(s) [17], [18] which follow patterns of browsing activity are frequently highfalse-positive [19], so more subtle, contextual processing must be used in addition to flagging based on threshold values

One gap that has remained in the literature is that, as of today, and sorry for that in "there is no framework dealing with personalized questioning and behavioral inference and integrity checking and privacy-preserving local deployment simultaneously". Our solution to this gap is to bring together a single architecture where combines real-time behavioral profiling and content assessment that have been proven to work effectively on a wide variety of candidate profiles and that is very stringent in terms of Using local inference: data sovereignty.

TABLE I: COMPARATIVE FEATURE MATRIX

Capability	Static Tests	Cloud AI	Proctoring	Terminology Ours
Personalized Questions	No	Yes	No	Yes
Behavioral Analysis	No	limited	Yes	yes
Integrity Monitoring	No	No	Yes	Yes
Privacy (Local)	Yes	No	No	Yes
Transparent Scoring	No	No	No	Yes
Multi-Modal Interface	No	Yes	No	Yes

III. SYSTEM ARCHITECTURE AND DESIGN RULES

The system consists of 3 design values: Privacy-by-Design (sovereignty of data by processing locally), Transparencyby-Default (providing for explainable scoring metrics) and Integration - over - Isolation (integration of behavioral and semantic analysis).

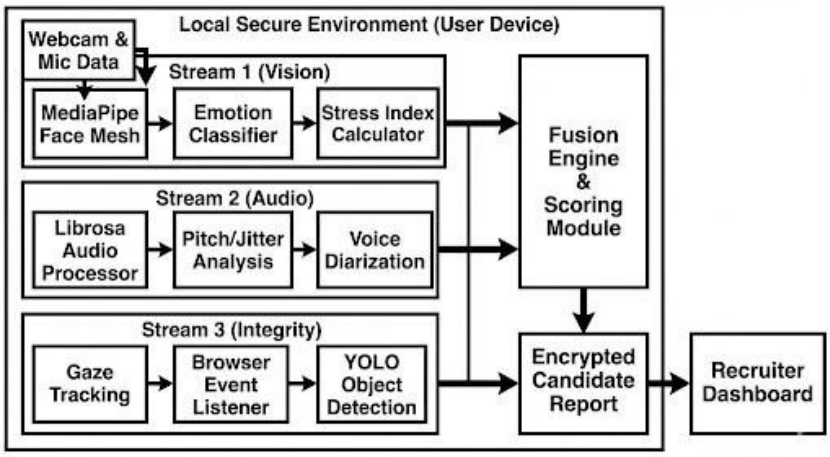

Figure 1. Architecture for the multimodal candidate assessment system.

The framework described in Fig. 1 has five interconnected modules: the Credential Intelligence Engine, Adaptive Interview Orchestrator, Behavioral Profiling Subsystem, Integrity Surveillance Subsystem, and the Unified Scoring Module. Every processing step is done in the local containerized environment, removing outbound external data. It depopulates professional occupations, educational history, and catalogs for skills a Candidate Knowledge Graph. The system employs a progression-aware algorithm (Algorithm 3) to prevent static and repetitive interviewing by calibrating ask questions in real-time according to performance. Our models are local Large Language Models (LLMs), and namely Llama 3 and Mistral versions deployment through the Ollama architecture [21]. Structured prompting is one way that offers the model possible context of candidates and possible target difficulty and generates questions freely of cloud APIs. The system is on-premise completely to meet the requirements of GDPR and CCPA [3]. While it is feasible on consumer-grade hardware (NVIDIA RTX 3060, 12 GB VRAM):
Latency: Median response generation =1.8 sec; 90% < 2.4sec. Through 4-6 simultaneous sessions per instance, defined as System Resources: o Memory: Max footprint 8 - 12 GB (quantization dependent).

IV. BEHAVIOURAL PROFILING SUBSYSTEM (BPS)

Non-verbal behavioral cues are intuitively interpreted by human interviewers and fed to the context of verbal responses. Patterns of facial tension suggest the level of stress that can be used to distinguish between short term anxiety and long-term difficulty.
Such behavioral signals are important to the right evaluation: a correct answer accompanied by excessive nervousness may be indicative of memorized information as opposed to true understanding, whereas, a safe display of an incorrect answer is a manifestation of knowledge deficiency with overconfidence. These measured by our behavioral profiling subsystem in order to let the automated system to get to the interpretive dimensions that had previously only been possible for human judgement.

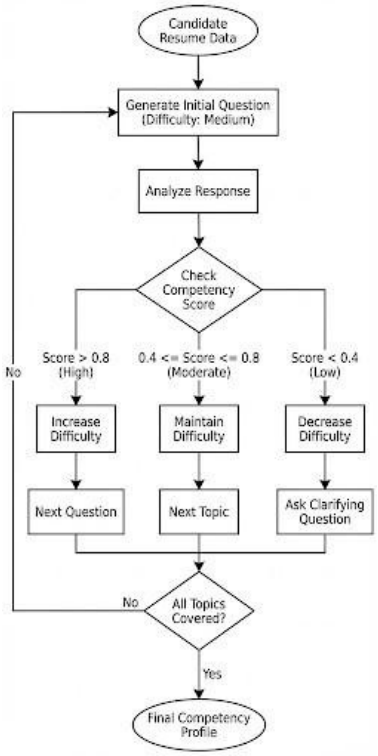

Figure 2. Flowchart of the logic of adaptive interview

The Behavioral Intelligence Layer basing confidence measurements and emotional and stress cues based on visual and vocal patterns, analysed from facial landmark detection and expression recognition. The behavioral composite score is computed as:

$$C_{behaviour} = \alpha \cdot E_{stability} + \beta \cdot S_{confidence} - \gamma \cdot S_{stress} \qquad (1)$$

where stress signals are weighed down against good confidence signals the use of calibrated weights alpha = 0.40 beta = 0.35 gamma = 0.25). We use MediaPipe Face Mesh [21] to detect 468 face landmarks at 30 FPS. These landmarks are fed into the emotion classification networks trained using FER-2013 [5] data local real-time.

Input: Candidate profile, Current Performance, question bank

Output: next question

- trouble/current difficulty = num (COMPUTEDIFFICULTY(current performance))
- Topics to be relevant for candidate profile ← EXTRACTTOPICS
- candidates & FILTERQUESTIONS(question bank, difficulty, relevant topics)
- if performance current performance < THRESHOLD LOW then
- question ← SELECTFOUNDational (candidates)
- else if no the current performance is greater THRESHOLD - "HIGH" then
- question,dyn,hint,left Rodney pretty,robert ptd,a lot-same skip play, clear out green 2 blue 5 red barrack,aud is to ofsmh,change for initial preview finalize report actually attend.
- else
- question <select inne dangere><adaptation at post interval ARE>s uartec(candidates) → eventuallys uartec (candidates) → answer ones(oncesto)
- end if
- return question

The system looks at emotions over time instead of judging each frame separately. It uses a 5-second sliding window to reduce noise while still keeping natural changes in expressions. This helps catch very brief micro-expressions (under 500 ms) and track how a person's emotional state shifts while answering questions.

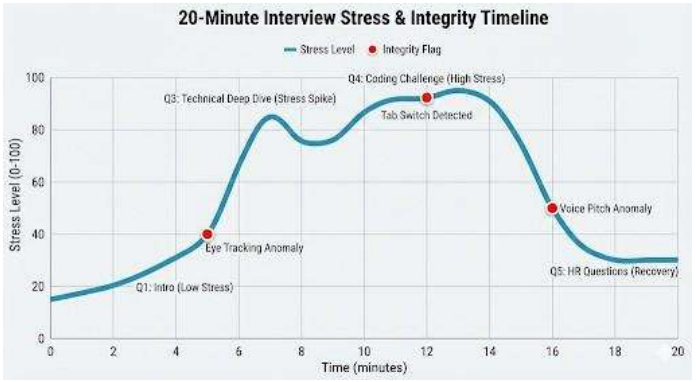

Figure 4. Timeline of Candidate Stress and Integrity flags.

In confidence assessment signal fusion of visual, vocal, and textual signals is used. The aggregate sum of the confidence scores where is calculated as weight sum:

$$C(t) = 0.4 \cdot C_{visual}(t) + 0.3 \cdot C_{vocal}(t) + 0.3 \cdot C_{textual}(t) \quad (2)$$

TABLE II. BEHAVIORAL-CONTENT CORRELATION MATRIX

Content	High Conf.	Low Conf.	High Stress
High (> 0.8)	High Hire	Verify	Anxiety
Medium (0.5–0.8)	Overconfident?	Reasonable	Monitor
Low (< 0.5)	Overconfident	Misses Info	Poor Fit

V. INTEGRITY MONITORING MODULE

Some of the risks associated with remote interviewing include impersonation, unauthorized assistance, and pre-scripted responses. This solution has a balance between security and privacy, unlike the invasive proctoring systems, which involve constant monitoring, but rather focused behavior and environmental monitoring.

A. Gaze Tracking

The eye points are mapped to screen coordinates with MediaPipe Iris landmark detectors and are categorised as on-screen, off-screen-left, off-screen-right or downward. The system computes on-screen percentage (less than 60 percent works as a yellow flag and off-screen time more than 15 seconds to work as a red flag).

B. Browser Surveillance & Environmental Surveillance

Browser Activity Monitoring relies on the JavaScript Page Visibility API [18] to monitor switching focus between windows and tabs. Graded Severity Logic Generally used: 3 Switches in one question or eight events in the totality of the interview cause high-severity flags. These events are correlated with specific questions through timestamps to see if resource consultation. Audit trails are kept and they can be reviewed after the interview.

C. IntegrityScore Computation violations are weighted and accumulated:

$$I_{score} = 1 - \frac{\sum w_i \cdot v_i}{v_{max}} \quad (3)$$

where vi are the individual indication of violations (tab switches, attention loss, multiple faces, audio anomaly), wi are weights associated with each violation determined empirically by Modeling the Word Algorithm: "minimize false-positive results." Manual review is triggered when scores fall below 0.7.

VI. UNIFORM SCORING FRAMEWORK

The assessment rubric contains four dimensions (Technical Accuracy, Communication Clarity, Behavioral Composure, Integrity Assurance), and each of these dimensions is worth a score on a scale of 0-10. These are combined through:

$$S_{final} = w_c \cdot S_{content} + w_b \cdot S_{behaviour} + w_i \cdot S_{integrity} \quad (4)$$

where wc+wb+wi = 1, so it allows weighting flexibility (Fig. 5).
Dimensions 1 -2 use language model-based scoring via seeded/proprompted inference servers of local inference servers. Prompts Include models for interview questions, candidate responses, and anticipated key point rubrics depending on the context of question generation. The output is in JSON form.
Dimensions 3-4 are computed in an algorithmic fashion on the basis of measures of behavior and integrity (that are continuous scales), with threshold based mapping of human understandable rating standards.
In primary visualization, a radar chart representation that shows the four assessment dimensions at thesame time, making it possible to evaluate a candidate's strength and weakness profiles. There are homogenous interpretation aids: green (score 8-10) characterized by good, yellow (score 5-7) is adequate but improvable competency, and red (score (%) 0-4) is significant causes for concern

Figure 5. Multi-dimensional scoring candidate evaluation dashboard.

Timeline visualization presents the progression of scores of a question-by-question basis and demonstrates performance trajectories and problematic identifiable exchanges. Detailed data The export capability produces detailed Reports in PDF format covering all of the visualizations, breakdowns of scores, evidence screenshots, and textual summarised that can be used for archive purposes and participating in collaborative decision-making.

VII. Experimental Evaluation

We had recruited 150 volunteers, each participant underwent a 20 - 30 minute session either by using text or voice mode. For ground truth, Three of our senior recruiters manually reviewed anonymized transcripts and video recordings are independently provided. Inter-rater agreement was high (Fleiss' kappa- k = 0.81), saying a high consensus that is a reliable ground truth on which to validate the system.
We examined three configurations, including Content-Only (baseline), Content + Behavioral and Full System. Table III attests to the fact that the Full System with 88.4% accuracy was 25% than the (70.7%) Content-Only baseline and better than that of intermediate Content + Behavioral environment (84.0%). The scoring of the system is highly correlated human consensus ($\rho = 0.87$). Configurations displayed statistically significant improvement (McNemar test,$p < 0.001$) and error difference analysis show that differences were mainly in ambiguous borderline cases.
Emotion Analysis: The average F1-score of this analysis was 0.84. obtained across 5000 Video frames manually annotated by two independent raters (Cohen's ; = .79). The best neutral expressions (0.89) accuracy was obtained And the lowest with confused ones (0.79).
Confidence Scoring Actual quality of response had a strong positive correlation with scores obtained being confident ($r = 0.76$) Correct answers with high degree of confidence had a positive predictive value of 89% on future job offers.
Participants (n=30), in a simulated controlled study (mimicking some violations: notes, help from out of camera, web search) showed that the overall detection rate was 94%. Tab Switching: 100% detection. Note Reading: 96% detection. Multi-Person: 87% detection.
The heterogeneous false positive rate was positive 7%, mainly due to environmental distractions (i.e. pets) and/or secondary observing monitors These are placed under the human consideration as opposed to automatic rejection.
The system measured 1.8 s median latency and 30 FPS of facial analysis on reference hardware (RTX 3060 Nvidia) and a peak memory use of 11.2 GB. Post-interview questionnaires (N = 150) revealed that users were

extremely satisfied with the system: 87% felt that the system interface was intuitive and 91% for the local architecture for privacy. Nevertheless, 18% raised concerns of continuous monitoring which were covered by protocols of explicit consent.

TABLE III: CLASSIFICATION PERFORMANCE METRICS

Conf.	Acc.	Prec.	Recall	F1	Human Agr. (ρ)
Content-Only	70.7%	0.71	0.69	0.70	0.74
Content + Behavioral	84.0%	0.85	0.82	0.83	0.82
Full System	88.4%	0.89	0.87	0.88	0.87

VII. DISCUSSION AND CONCLUSION

- Detection of the Packet Integrity Performance The integrity monitoring subsystem was able to detect 94% of the simulated violations. The 7% false positive rate is good given the human review requirements and is especially better than cloud-based proctoring systems, who usually record false positive rates of 15-20% [14]. The context-awareness in the interpretation of the gaze/browser events is directly responsible to this improvement regarding naive threshold-based systems.
- Privacy-Performance Tradeoff Our system overcomes the trade-off between privacy and performance of the conventional method: by local deployment, compliance with GDPR and CCPA is not lost and does not hamper the level of sophistication of analysis. The system offers a holistic content, behavioral and integrity dimension evaluation unlike isolated tools.
- Limitations Technical limitations - requires a sufficient amount of light for reliable face analysis and for the necessity for middle range GPU processing capabilities (e.g. separator Nvidia Rt 3060) for real-time inference. There is also the problem of the cultural bias in emotion recognition because, mainly, training sets are complete of composed of Western subjects, which perhaps falsely classifies expressions taken from other cultural contexts [7].
- Enhanced Multimodal Analysis In the next version, voice there will be prosody analysis investigating pitch contours, speed of speech, and the strength of voice to be able to supplement confidence detection. We will be trying to integrate the performance measures from interviews with outcome measures after hire to enable data-driven calibration of scoring; weights.

A comprehensive AI interview system is a combination of context-sensitive interviews, behavioral profiling, and integrity checks in a local setup that considers privacy concerns. It considers facial expressions, stress, and environment authenticity, and textual responses in combination. Tests on 150 applicants revealed that the accuracy was 88.4% and human agreement (r=0.87) and integrity violation (94 percent) were high, indicating good privacy-aware automation.

REFERENCES

[1] in "Automated parsing of resumes and candidate", ACM Transactions on Internet Technology, vol. 5, no. 4, pp. 507-535 [1-14] screening systems," in IEEE Trans. Software Eng. vol. 45, no. 3, pp. 234-295. 248, Mar 2019.

[2] M. Chen et al., "Conversational AI for technical interviews: A cloudbased approach," in Proc. Int. Conf. Artif. Intell., 2022, pp. 156–169.

[3] R. Johnson, "Large language models in recruitment: Opportunities and" "Challenges," J. Human Resources Technol., vol. 12, no. 2, pp. 89-104, 2023.

[4] P. Ekman and W. V. Friesen Facial Action Coding System: A Technique. for Measurement of Facial Movement. Palo Alto, CA: Consulting Psychologists Press, 1978.

[5] I. J. Goodfellow, B. Bengio and Y. Bastiat, "Challenges in representation learning: A report" on three contests on machine learning, in Proc. Int. Conf. Neural Inf. Process., 2013, pp. 117–124.

[6] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence and arousal computing in the wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan. 2019.

[7] S. D'mello and J. Kory, "A review and meta-analysis of multimodal "affect detection systems," a.c.m. Comp. Surv., vol. 47, no. 3, pp. 1–36, 2015.

[8] L. Zhang and P. Wang, "Facial expression recognition in professional" assessment contexts," Pattern Recognit., vol. 89, pp. 156-167, 2020.

[9] R. S. Lazarus and S. Folkman, Stress, Appraisal and Coping. New York: Springer, 1984.

[10] M. Sharma, U. R. Acharya, "Automated detection of stress using Conducting a literature review on "Physiological signals: A review," Biomed. Signal Process. Control, vol. 68, p. 102755, 2021.

[11] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music-listening," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[12] Z. Zeng et al., "A survey of affect recognition methods: Audio, visual,", However, some critics and proponents of the analysis-who usually refer to "facial recognition"-have disagreed with it and argued that, "In addition, spontaneous expressions," and "Thus, there is no agreement with the analysis above that [facial recognition] can be considered a subfield of computational intelligence." vol. 31, no. 1, pp. 39–58, Jan. 2009.
[13] D. L. Cluskey, "Examining student views about online proctoring and Last page (academic integrity" J. Educ. Technol. Syst., vol. 49, no. 4, pp. 462- 481, 2021.
[14] A. P. Nigam et al., "Systematic Review of Remote Proctoring Technologies," Educ. Inf. Technol., vol. 26, pp. 6129-6151, 2021.